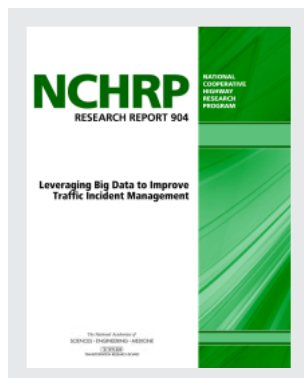


This PDF is available at <http://nap.edu/25604>

SHARE



Leveraging Big Data to Improve Traffic Incident Management (2019)

DETAILS

204 pages | 8.5 x 11 | PAPERBACK

ISBN 978-0-309-48071-0 | DOI 10.17226/25604

GET THIS BOOK

FIND RELATED TITLES

CONTRIBUTORS

Kelley Klaver Pecheux, Benjamin B. Pecheux, AEM Corporation: Grady Carrick, Enforcement Engineering, Inc.; National Cooperative Highway Research Program; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2019. *Leveraging Big Data to Improve Traffic Incident Management*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/25604>.

Visit the National Academies Press at NAP.edu and login or register to get:

- Access to free PDF downloads of thousands of scientific reports
- 10% off the price of print titles
- Email or social media notifications of new titles related to your interests
- Special offers and discounts



Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. ([Request Permission](#)) Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.

Copyright © National Academy of Sciences. All rights reserved.

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP RESEARCH REPORT 904

Leveraging Big Data to Improve Traffic Incident Management

Kelley Klaver Pecheux

Benjamin B. Pecheux

AEM CORPORATION

Herndon, VA

Grady Carrick

ENFORCEMENT ENGINEERING, INC.

Ponte Vedra, FL

Subscriber Categories

Highways • Operations and Traffic Management • Security and Emergencies

Research sponsored by the American Association of State Highway and Transportation Officials
in cooperation with the Federal Highway Administration



2019

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed, and implementable research is the most effective way to solve many problems facing state departments of transportation (DOTs) administrators and engineers. Often, highway problems are of local or regional interest and can best be studied by state DOTs individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation results in increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

Recognizing this need, the leadership of the American Association of State Highway and Transportation Officials (AASHTO) in 1962 initiated an objective national highway research program using modern scientific techniques—the National Cooperative Highway Research Program (NCHRP). NCHRP is supported on a continuing basis by funds from participating member states of AASHTO and receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

The Transportation Research Board (TRB) of the National Academies of Sciences, Engineering, and Medicine was requested by AASHTO to administer the research program because of TRB's recognized objectivity and understanding of modern research practices. TRB is uniquely suited for this purpose for many reasons: TRB maintains an extensive committee structure from which authorities on any highway transportation subject may be drawn; TRB possesses avenues of communications and cooperation with federal, state, and local governmental agencies, universities, and industry; TRB's relationship to the National Academies is an insurance of objectivity; and TRB maintains a full-time staff of specialists in highway transportation matters to bring the findings of research directly to those in a position to use them.

The program is developed on the basis of research needs identified by chief administrators and other staff of the highway and transportation departments, by committees of AASHTO, and by the Federal Highway Administration. Topics of the highest merit are selected by the AASHTO Special Committee on Research and Innovation (R&I), and each year R&I's recommendations are proposed to the AASHTO Board of Directors and the National Academies. Research projects to address these topics are defined by NCHRP, and qualified research agencies are selected from submitted proposals. Administration and surveillance of research contracts are the responsibilities of the National Academies and TRB.

The needs for highway research are many, and NCHRP can make significant contributions to solving highway transportation problems of mutual concern to many responsible groups. The program, however, is intended to complement, rather than to substitute for or duplicate, other highway research programs.

NCHRP RESEARCH REPORT 904

Project 17-75

ISSN 2572-3766 (Print)

ISSN 2572-3774 (Online)

ISBN 978-0-309-48071-0

Library of Congress Control Number 2019947990

© 2019 National Academy of Sciences. All rights reserved.

COPYRIGHT INFORMATION

Authors herein are responsible for the authenticity of their materials and for obtaining written permissions from publishers or persons who own the copyright to any previously published or copyrighted material used herein.

Cooperative Research Programs (CRP) grants permission to reproduce material in this publication for classroom and not-for-profit purposes. Permission is given with the understanding that none of the material will be used to imply TRB, AASHTO, FAA, FHWA, FMCSA, FRA, FTA, Office of the Assistant Secretary for Research and Technology, PHMSA, or TDC endorsement of a particular product, method, or practice. It is expected that those reproducing the material in this document for educational and not-for-profit uses will give appropriate acknowledgment of the source of any reprinted or reproduced material. For other uses of the material, request permission from CRP.

NOTICE

The research report was reviewed by the technical panel and accepted for publication according to procedures established and overseen by the Transportation Research Board and approved by the National Academies of Sciences, Engineering, and Medicine.

The opinions and conclusions expressed or implied in this report are those of the researchers who performed the research and are not necessarily those of the Transportation Research Board; the National Academies of Sciences, Engineering, and Medicine; or the program sponsors.

The Transportation Research Board; the National Academies of Sciences, Engineering, and Medicine; and the sponsors of the National Cooperative Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.

Published research reports of the

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

are available from

Transportation Research Board
Business Office
500 Fifth Street, NW
Washington, DC 20001

and can be ordered through the Internet by going to

<http://www.national-academies.org>

and then searching for TRB

Printed in the United States of America

The National Academies of **SCIENCES • ENGINEERING • MEDICINE**

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, non-governmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at www.national-academies.org.

The **Transportation Research Board** is one of seven major programs of the National Academies of Sciences, Engineering, and Medicine. The mission of the Transportation Research Board is to increase the benefits that transportation contributes to society by providing leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied committees, task forces, and panels annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

Learn more about the Transportation Research Board at www.TRB.org.

COOPERATIVE RESEARCH PROGRAMS

CRP STAFF FOR NCHRP RESEARCH REPORT 904

Christopher J. Hedges, *Director, Cooperative Research Programs*
Lori L. Sundstrom, *Deputy Director, Cooperative Research Programs*
William C. Rogers, *Senior Program Officer*
Jarrel McAfee, *Senior Program Assistant*
Eileen P. Delaney, *Director of Publications*
Natalie Barnes, *Associate Director of Publications*
Sharon Lamberton, *Editor*

NCHRP PROJECT 17-75 PANEL Field of Traffic—Area of Safety

Stephen W. Glascock, *Louisiana DOTD, Baton Rouge, LA*
Mara K. Campbell, *Jacobs, New Florence, MO*
Melissa L. Clark, *California DOT, Sacramento, CA*
Steve J. Cyra, *HNTB Corporation, Milwaukee, WI*
Edward Gincauskis, *Massachusetts DOT, Boston, MA*
Eric J. Hemphill, *North Texas Tollway Authority, Plano, TX*
Naveen Lamba, *Grant Thornton LLP, Alexandria, VA*
Eileen M. Singleton, *Baltimore Metropolitan Council, Baltimore, MD*
Pradeep Tiwari, *Phoenix, AZ*
Douglas Mark Tomlinson, *Pennsylvania DOT, Harrisburg, PA*
Paul Jodoin, *FHWA Liaison*
Victor T. Hom, *National Oceanic and Atmospheric Administration Liaison*
Richard A. Cunard, *TRB Liaison*



FOREWORD

By William C. Rogers

Staff Officer

Transportation Research Board

NCHRP Research Report 904 provides guidance for transportation and traffic incident management agencies to bring multiple comprehensive datasets (Big Data) together to derive useful information and relationships that could improve their efforts to reduce clearance times and increase highway safety. The ability to mine information on heretofore-unanticipated trends can provide significant opportunities for improving protocols, resource management, scene management, and real-time data sharing.

As the nation continues its migration toward Big Data, an overwhelming volume of data can be used to improve the current state of traffic incident management (TIM). Big Data is not just “a lot more data” than what was available before. It is a fundamental change in how data are collected, analyzed, and used to uncover trends and relationships. In general, recent advances in information technology (IT) have significantly increased data quantity, improved data quality, and enhanced data analytics. Recognizing the tremendous potential of Big Data applications both within and outside the realm of transportation, many agencies are faced with the challenge of using or even identifying the rich datasets that could be leveraged to enhance or improve TIM efforts. A need exists to develop a Big Data environment in which datasets from multiple sources can be managed and valued. The challenges are to discover the datasets, to merge them into a shared Big Data environment, to uncover important relationships, and to identify trends that may occur outside the traditional evaluation processes.

In NCHRP Project 17-75, “Leveraging Big Data to Improve Traffic Incident Management,” AEM Corporation was asked to develop guidelines that (1) describe current and emerging sources of Big Data that could improve TIM; (2) describe potential opportunities to leverage Big Data that could advance TIM state of the practice; (3) identify potential challenges (e.g., security, proprietary, or inter-operability issues) for TIM agencies seeking to leverage Big Data; and (4) develop a matrix of Big Data options for transportation and TIM agencies to use based on their current capabilities.

CONTENTS

1	Summary
6	Chapter 1 Introduction
7	1.1 Objective
7	1.2 Overview of Research and Organization of Report
9	Chapter 2 State of the Practice of TIM
9	2.1 State of the Practice
10	2.1.1 Establishment of Local, Regional, and Statewide TIM Committees
11	2.1.2 Implementation of TIM Legislation
11	2.1.3 Development and Implementation of National TIM Responder Training
11	2.1.4 Development of TIM Strategic Plans
12	2.1.5 Development and Implementation of Agency Operating Agreements
12	2.1.6 Implementation of Agency Policies for Safe and Quick Clearance
12	2.2 The Use of Data to Support TIM
12	2.2.1 TIM Performance Measurement and Management
14	2.2.2 Making the Business Case for TIM
15	2.3 Further Advancing the State of the Practice of TIM
16	Chapter 3 State of the Practice of Big Data
16	3.1 Big Data Definition
17	3.1.1 Volume
18	3.1.2 Variety
18	3.1.3 Velocity
19	3.1.4 Veracity
19	3.1.5 Value
19	3.2 The Move from Traditional Data Analysis to Big Data Analytics
19	3.2.1 Traditional Data Analysis
21	3.2.2 Hadoop: The Start of Big Data Tools
21	3.2.3 Current Big Data Tools
25	3.2.4 Big Data Architecture
26	3.2.5 Examples of Big Data Analytics
32	3.3 Big Data Applications in Transportation
32	3.3.1 Transportation Planning
33	3.3.2 Parking
33	3.3.3 Trucking
34	3.3.4 Public Transportation
34	3.3.5 Transportation Operations and ITSs
37	3.3.6 Emergency and Incident Management

39	Chapter 4 Big Data and TIM
40	4.1 Improve On-Scene Management Practices
42	4.2 Improve Resource Utilization and Management
44	4.3 Improve Safety
46	4.4 Enable Predictive TIM
48	4.5 Support Performance Measurement and Management
51	4.6 Support TIM Justification and Funding
53	4.7 Summary
55	Chapter 5 Assessment of Data Sources for TIM
55	5.1 Data Source Assessment Approach
56	5.1.1 Assessment Criteria
60	5.1.2 Data Maturity Assessment Approach
62	5.2 Findings
62	5.2.1 State Traffic Records Data
70	5.2.2 Transportation Data
75	5.2.3 Public Safety Data
77	5.2.4 Crowdsourced Data
81	5.2.5 Advanced Vehicle Systems Data
85	5.2.6 Aggregated Datasets
92	5.3 Summary
94	Chapter 6 Big Data Guidelines for TIM Agencies
96	6.1 Adopt a Deeper and Broader Perspective on Data Use
96	6.2 Collect More Data
98	6.3 Open and Share Data
99	6.3.1 Public Records Laws
99	6.3.2 Proprietary Data Formats
100	6.3.3 Contract Data Clauses
100	6.3.4 Benefits of Opening and Sharing Data
101	6.4 Use a Common Data Storage Environment
101	6.4.1 Data Silos
101	6.4.2 Data Virtualization
102	6.5 Adopt Cloud Technologies for the Storage and Retrieval of Data
103	6.5.1 Understand the Cost Savings of the Cloud
104	6.5.2 Understand Cloud Security
105	6.5.3 Recognize the Inherent Connection Between Big Data Analytics and the Cloud
106	6.6 Manage the Data Differently
106	6.6.1 Store the Data “As Is”
107	6.6.2 Maintain Data Accessibility
107	6.6.3 Structure the Data for Analysis
108	6.6.4 Ensure That Data Is Uniquely Identifiable
108	6.6.5 Sharing, Security, and Privacy
109	6.7 Process the Data
109	6.7.1 Process the Data Where It Is Located
110	6.7.2 Use Open-Source Software
112	6.7.3 Do Not Reinvent the Wheel
112	6.7.4 Understand the Ephemeral Nature of Big Data Analytics
113	6.8 Open and Share Outcomes and Products to Foster Data User Communities

114	Chapter 7	Summary and Next Steps
114	7.1	Summary of Findings
116	7.2	Next Steps
117	7.3	Suggestions and Priorities for Additional Related Research
118		Abbreviations
121		Glossary
125		References
132	Appendix A	Data Source Assessment Tables
181	Appendix B	Incident Response and Clearance Ontology (IRCO)

Note: Photographs, figures, and tables in this report may have been converted from color to grayscale for printing. The electronic version of the report (posted on the web at www.trb.org) retains the color versions.

SUMMARY

Leveraging Big Data to Improve Traffic Incident Management

The term *Big Data* represents a fundamental change in what data is collected and how it is collected, analyzed, and used to uncover trends and relationships. Big Data is not just about the volume of data, it also is about the velocity, variety, veracity, and value of data. The ability to merge multiple, diverse, and comprehensive datasets and then to mine the data to uncover or derive useful information on heretofore unknown or unanticipated trends and relationships could provide significant opportunities to advance the state of the practice of traffic incident management (TIM) policies, strategies, practices, and resource management.

Research Objectives

NCHRP Project 17-75, “Leveraging Big Data to Improve Traffic Incident Management,” had the following objectives: to conduct research to illuminate Big Data concepts, applications, and analyses; describe current and emerging sources of data that could improve TIM; describe potential opportunities for TIM agencies to leverage Big Data approaches; identify potential challenges associated with the use of Big Data; and develop guidelines to help advance the state of the practice for TIM agencies.

Research Approach

To meet the objectives of the project, the research approach included the following activities:

- Assess research, practices, and innovative approaches through a review of the literature.
- Organize and conduct a responder workshop to inform the development of an incident response and clearance ontology and to identify areas in which improvements to TIM are needed.
- Identify Big Data opportunities for TIM based on the current state of the practice and responder needs.
- Conduct a comprehensive assessment of a wide variety of TIM-relevant data sources to determine the openness, maturity, and readiness for Big Data applications.
- Create an incident response and clearance ontology.
- Develop guidelines that help to advance TIM agencies toward the application of Big Data.

Findings

State of the Practice of TIM and Big Data

The state of the practice of TIM shows significant advancement over the past decade, most notably through the development of regional and statewide TIM committees, the

2 Leveraging Big Data to Improve Traffic Incident Management

National Traffic Incident Management Responder Training Program, the implementation of TIM legislation, and the collection and analysis of TIM data for performance measurement. Among these advancements, however, the collection and use of TIM data by agencies have lagged. Recent guidance provided by TRB and the FHWA, as well as the ongoing FHWA “On-Ramp to Innovation: Every Day Counts” (EDC) initiative to improve the quantity and quality of TIM data, reflect national efforts to advance the collection and use of TIM data.

The findings from a review of the state of the practice in Big Data reinforce that Big Data is not new and indeed has been applied for nearly two decades by major technology companies. Big Data is characterized by the “five Vs”—volume, velocity, variety, veracity, and value—but it is not necessary for all datasets to possess all five qualities to be considered Big Data. Contrary to the relational database approach, Big Data analytics is not bound to a single set of tools to perform an analysis; rather, Big Data analytics encompass a wide variety of proprietary and open-source tools that can be customized and modified by users. These tools allow for the rapid transfer, processing, storage, and analysis of extremely large datasets. These tools have increased the ability to analyze divergent data, such as decades-old historical records and real-time streaming data, to derive value that previously could not be attained using traditional approaches that typically rely on relational databases.

Big Data applications in the field of transportation are more recent (having developed within the past few years) and include applications in areas such as planning, parking, trucking, public transportation, operations, and Intelligent Transportation Systems (ITSs). A significant gap exists between the current state of the practice in Big Data analytics and the current state of transportation agency applications of data for TIM. The research team identified a few TIM Big Data applications, but for the most part, these applications could be performed using relational databases. Generally, at the local and state levels, data is not collected at the volume needed to effectively use or apply Big Data approaches. Ways are available to expand on these initial approaches to Big Data for TIM, but the data must first be prepared, must be of a sufficient size and must cover a sufficient length of time to enable identification of meaningful patterns that yield value.

Big Data Opportunities for TIM

The application of Big Data technologies and analytics could further advance the state of the practice in TIM. To illustrate possible Big Data opportunities for TIM, the research discussed in this report contrasts traditional TIM data collection and analysis approaches with example Big Data applications for the same problem or research questions designed to:

- Improve scene management practice,
- Improve resource utilization and management,
- Improve safety,
- Enable predictive TIM,
- Support performance measurement and management, or
- Support TIM justification and funding.

Each example application describes the current practice, the potential for a Big Data approach, the differences in data needs and analytical approaches, and the possibilities and benefits afforded by Big Data. These example Big Data applications illustrate that Big Data approaches are not simply an improvement on current practices; rather, Big Data represents a radical change from traditional approaches—a complete paradigm shift—and many of the benefits of Big Data analytics will require aggregating data at least at the state level, if not at the national level.

Data Source Assessment

The research team conducted a comprehensive assessment of 31 TIM-relevant data sources organized across six data domains. The assessment included a description of each data source, its potential application for TIM, the costs of accessing the data, and challenges associated with the data sources. The data sources also were assessed using two different data maturity models, and the assessment included an overall evaluation of data readiness and openness.

The assessment findings confirmed that large gaps exist between the current state of TIM-relevant data and the application of this data for Big Data analytics. Although it may be tenable for agencies to merge a few existing datasets, developing and integrating most of the datasets will require major efforts. Building more detailed and integrated datasets will require the dedication of significant resources and expertise, and the application of non-traditional approaches. Challenges such as the lack of standards for data collection and storage, personally identifiable information (PII), legal restrictions, and agency culture and policies will limit the application of Big Data for TIM.

Existing TIM-relevant Big Data datasets (from sources like HERE Technologies, INRIX, and Waze) can provide a start to the use of Big Data, but these datasets lack the detail needed for effectively mining and understanding the nuances of incident response and TIM. Furthermore, even though traffic sensors and probes generate millions of data points every second, the relative infrequency of incidents (e.g., crashes) limits the application of Big Data to TIM unless the data is aggregated across multiple regions and organizations to increase its volume and variety. Finally, agencies must possess the willingness and openness to embrace the paradigm shift that is required to use Big Data. Continued unwillingness to open and share data or to utilize cloud infrastructure are basic factors that will limit the growth and application of Big Data within an organization.

Incident Response and Clearance Ontology

Although it may be possible to use implicit or existing relationships within data elements to perform simple Big Data analyses, more complex and insightful Big Data analyses require a more abstract and concise way to express the knowledge represented by the data. This can be done with an ontology. An ontology is a set of concepts and categories in a subject area or domain that show their properties and the relationships between them. NCHRP Project 17-75 included a first attempt at establishing an incident response and clearance ontology (IRCO), a formal naming and definition of the types, properties, and inter-relationships of the entities that exist in the TIM domain. The development of the IRCO was aided by a workshop attended by a broad range of incident responders who provided insights on the vocabulary, entities, and relationships associated with incident response and clearance. During the workshop, it was established that the TIM ontology should first focus on conceptualizing the response to an incident and how the response relates to the incident itself, as well as the incident environment and the personnel, actions, equipment, and response vehicles involved in the response.

The research team combined information gathered during the workshop with information from existing traffic incident-related ontologies identified in the literature to establish a basis for the IRCO. To capture the distributed and spatiotemporal nature of an incident response, as well as the various tasks performed by responders using various equipment, the Live OWL Documentation Environment (LODE) ontology was used. The LODE ontology allows an event to be described in time, in space, and in terms of who was involved

4 Leveraging Big Data to Improve Traffic Incident Management

during the event. The IRCO organizes all these details in terms of defined classes and super classes, various object properties, and various data properties.

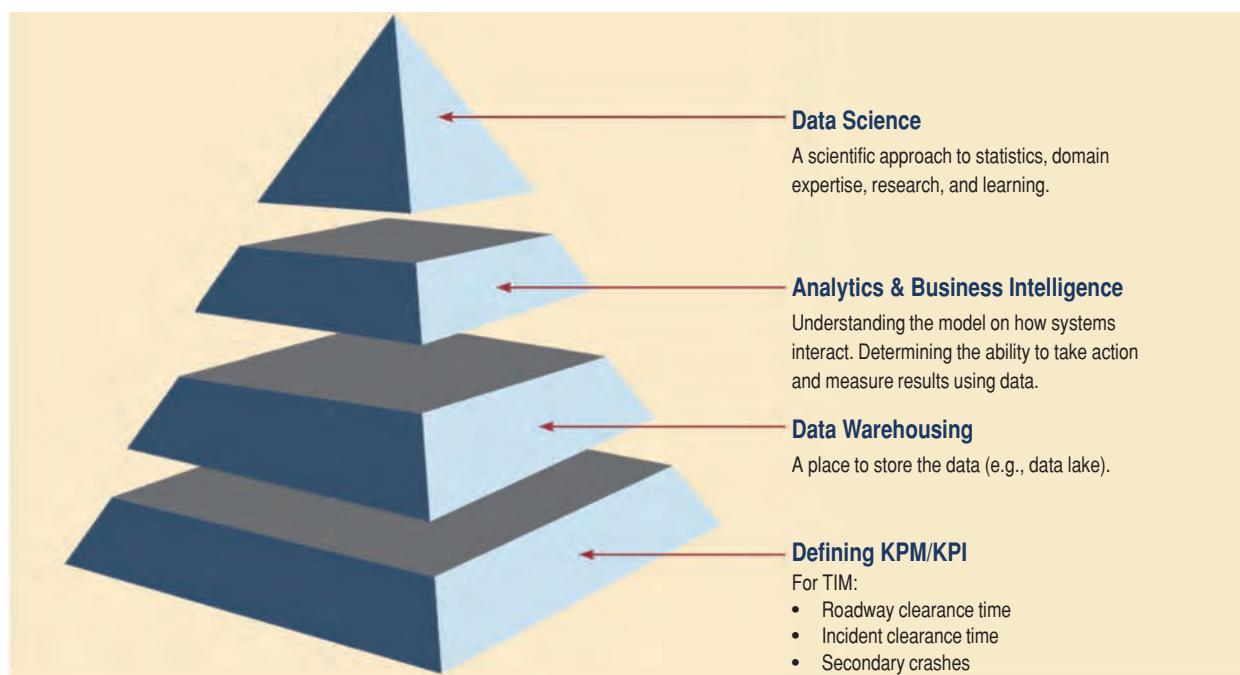
The IRCO attempts to show how the TIM-relevant datasets are related to each other. The IRCO helps analysts understand how to organize a Big Data data store (or Big Data *lake*) and data analysis system for TIM so that users can quickly understand and leverage the information that is available. A complete description of the ontology and a graphical representation of the resulting IRCO are provided in Appendix B of this report.

Big Data Guidelines for TIM Agencies

The Big Data pyramid (Figure S-1) illustrates four tiers associated with reaching the level of applying data science. These tiers include: (1) the foundational activity of defining key performance measures (KPMs) and key performance indicators (KPIs); (2) the development of a Big Data store in which to capture, store, manage, and analyze Big Data datasets; (3) the development and maintenance of analytics and business intelligence tools and processes; and (4) the achievement of a mature Big Data practice.

The research for NCHRP Project 17-75 suggests that the current state of the practice for TIM data collection, storage, and analysis is between the first and second tiers on the Big Data pyramid. At this point, very limited TIM data is being collected and shared among partner agencies, and a solid data lake as a foundation for the development of TIM business intelligence (the third tier of the Big Data pyramid) and TIM data science (the fourth/top tier of the Big Data pyramid) has yet to be built.

Based on the findings from this research, eight guidelines were developed to lay out the various changes that will be necessary for transportation and TIM agencies to develop a usable Big Data lake, implement agency-wide analytics and business intelligence, and



Source: Adapted from "Big Progress in Big Data" (Drow, Lange, and Laufer 2015)

Figure S-1. The Big Data pyramid.

pursue the development of an evolving data science environment beneficial to the entire agency. These guidelines will help position transportation and TIM agencies for Big Data.

The eight guidelines can be summarized as follows:

- Adopt a deeper and broader perspective on data use.
- Collect more data.
- Open and share data.
- Use a common data storage environment.
- Adopt cloud technologies for the storage and retrieval of data.
- Manage the data differently.
- Process the data.
- Open and share outcomes and products to foster data user communities.

These guidelines are further illuminated in Chapter 6 of this report.

Next Steps

The guidelines encourage agencies to begin putting research into practice by fully embracing low-cost, traditional good practices in data collection, cleaning, warehousing, and analysis with existing data sources. Agencies also are encouraged to concurrently identify opportunities to ready their organizations for Big Data. Opening and sharing data, both internally and externally, are critical cultural shifts that need to be embraced. An incremental approach is recommended that begins with developing the culture, policies, and expertise to improve the usability and increase the use of current data, as well as capturing opportunities to migrate from in-house servers to the cloud. These steps are the basis for positioning agencies to begin capitalizing on the opportunities afforded by Big Data.

The time is ripe for Big Data implementation. The technology is here, the tools are available, and the expertise exists to assist transportation agencies in both understanding and applying these technologies and tools to everyday questions and problems. Transportation agencies are encouraged to make the leap forward and begin to embrace the changes that will enable them to tackle Big Data. Even if—largely due to the pressures of organizational culture and a lack of data—transportation agencies have yet to fully accept and adopt the foundational principles of Big Data, the emergence of connected vehicle, traveler, and infrastructure data will soon drive this change. To avoid drowning in the imminent influx of data, and to capitalize on the wealth of information that can be derived from it, transportation agencies must ready themselves to use Big Data.

What are not yet readily available are effective strategies and techniques to break down the barriers (e.g., culture, legal, proprietary software) that impede transportation agencies from adopting Big Data practices. This is one area in which research can help agencies accelerate the adoption of Big Data for TIM. Once transportation and partner agencies have collected, opened, shared, and pooled enough (and varied) data in a cloud environment, further research can be conducted to explore the data using Big Data techniques to discover how it can help to improve specific components of TIM programs.



CHAPTER 1

Introduction

The term *Big Data* represents a fundamental change in what data is collected and how it is collected, analyzed, and used to uncover trends and relationships. The ability to merge multiple, diverse, and comprehensive datasets and then mine the data to uncover or derive useful information on heretofore unknown or unanticipated trends and relationships could provide significant opportunities to advance the state of the practice in TIM policies, strategies, practices, and resource management. For example Big Data could:

- Include non-traditional datasets to allow for the establishment of additional TIM performance measures, as well as the identification of TIM performance trends and the factors that impact performance;
- Bolster context-aware decision-making, such as aiding in resource allocation, on-scene actions, and scene management;
- Move TIM beyond intuitive operations, enabling prediction of when, where, and under what conditions problems are most likely to occur; and
- Help agencies build a more compelling and clear business case for their TIM programs, a fundamental step in securing continuous funding and supporting the long-term health and viability of these critical programs.

Big data involves a large *volume* of data, but it is not just about the volume of data. Four other “Vs” of Big Data also are important:

- *Velocity* refers to the speed/frequency at which the data is available.
- *Variety* refers to the diversity of the datasets available.
- *Veracity* refers to the legitimacy or trustworthiness of the data.
- *Value* refers to the worth of the data to its users.

Above all, Big Data must provide information that is of value to its users. One way to conceptualize the value of data, including Big Data, is through a value chain. Figure 1-1 illustrates a straightforward Big Data value chain from a presentation at a 2014 data symposium hosted by the Florida Department of Transportation (Florida DOT) (Kanniyappan and McQueen 2014). The figure represents the cascading benefits that can be derived from Big Data, ranging from data analysis to insight gains to better decision-making and, finally, to better design, planning, and operations. In their presentation, Kanniyappan and McQueen stated that, “Big Data may be as important to business and society as the Internet since more data leads to more accurate analysis.”

Not all the Vs need to be present for data to qualify as Big Data. In fact, for TIM, the most important of the Vs may be the variety of the available datasets. Given the multi-disciplinary nature of TIM and the wide range of responder organizations and roles—not to mention the roadway, environmental, and human factors involved in traffic incidents—having access to



Source: Kanniyappan and McQueen (2014)

Figure 1-1. The Big Data value chain.

diverse, associated datasets could be key in identifying ways to improve TIM. Many opportunities exist to improve TIM using Big Data, and numerous data sources exist from which to draw.

Traditional data sources for measuring and assessing TIM performance are transportation datasets like those generated at traffic management centers (TMCs) and by safety service patrol (SSP) programs. Some datasets generated by law enforcement, such as those associated with computer-aided dispatch (CAD) systems and crash reports, also are used to examine TIM performance. Yet-to-be-tapped data sources offer many additional opportunities to gain insights into where and how TIM could be improved, as well as when, where, and under what conditions traffic incidents are likely to occur so that the appropriate response can be pre-staged and/or immediately put into place. Moreover, leveraging Big Data that is associated with the connected and automated vehicle future may enable TIM operations that expedite incident detection and response while improving on-scene safety. Central to the concept of leveraging Big Data is the promise that analytics can illuminate critical actions that may result in significant improvements.

1.1 Objective

The objectives of NCHRP Project 17-75 were to conduct research to illuminate Big Data concepts, applications, and analyses; describe current and emerging sources of data that could improve TIM; describe potential opportunities for TIM agencies to leverage Big Data; identify potential challenges associated with the use of Big Data; and develop guidelines to help advance the state of the practice for TIM agencies.

1.2 Overview of Research and Organization of Report

To meet the objectives of this research project, an approach was laid out that included the following activities:

- Assess research, practices, and innovative approaches through a review of the literature.
- Organize and conduct a responder workshop to inform the development of an incident response and clearance ontology, and to identify areas in which improvements to TIM are needed.
- Identify Big Data opportunities for TIM based on the current state of the practice and responder needs.
- Conduct a comprehensive assessment of a wide variety of TIM-relevant data sources to determine the openness, maturity, and readiness for Big Data applications.
- Create an incident response and clearance ontology.
- Develop guidelines that help to advance TIM agencies toward the application of Big Data.

8 Leveraging Big Data to Improve Traffic Incident Management

This report describes the research approach in more detail and presents the findings associated with each of the research activities. The remaining chapters of the report are organized as follows:

- **Chapter 2: State of the Practice of TIM:** This chapter provides a high-level overview of the state of the practice in TIM procedures, training, data collection, and the use of data for measuring and monitoring TIM performance, and makes the business case for TIM. Examples are provided for agencies/organizations that are leaders in using data to assess and improve TIM.
- **Chapter 3: State of the Practice of Big Data:** This chapter provides a brief introduction to and history of Big Data, addresses the state of the practice, and provides a cross-industry overview of Big Data, including storage and analytics tools.
- **Chapter 4: Big Data and TIM:** This chapter explores Big Data opportunities for TIM by presenting specific examples that stem from applications representing the current state of the practice in TIM data collection and analysis. Each example begins with a summary of the traditional data collection and analysis approach. The summary is followed by presentation of a potential Big Data approach/opportunity to address the same problem or research question. Each example concludes with a discussion that contrasts the differing data needs and analytical approaches used in the traditional and Big Data approaches and highlights the possibilities and potential benefits afforded by Big Data.
- **Chapter 5: Assessment of Data Sources for TIM:** This chapter presents the approach and findings from a comprehensive assessment of 31 data sources in six categorized data domains that are relevant to TIM. The findings include a description of each data source, its potential application for TIM, the costs of accessing the data, and challenges associated with the data sources. The data sources also are assessed using two data maturity models, including an overall assessment of data readiness and openness. Detailed data assessment tables for each data source are presented in Appendix A.
- **Chapter 6: Big Data Guidelines for TIM Agencies:** This chapter presents the Big Data pyramid, a convenient visual guide for the application of data science. Based on the findings from this research, guidelines also are provided to support moving the TIM community (and state transportation agencies in general) toward the application of Big Data.
- **Chapter 7: Summary and Next Steps:** This chapter summarizes the findings of the research, sets forth potential next steps for the research findings, and addresses recommendations, needs, and priorities for additional related research.
- **Appendix A: Data Source Assessment Tables:** This appendix contains the detailed data assessment tables for 31 data sources, each including information on the organization that collects, maintains, and owns the data; how the data is collected; data structure; data size; data storage and management; data accessibility; data sensitivity; data openness; data challenges; and data costs.
- **Appendix B: Incident Response and Clearance Ontology (IRCO):** This appendix explains the concept of an ontology, and the value in creating ontologies. The approach and steps taken to develop an incident response and clearance ontology (IRCO) are established for application to TIM.

CHAPTER 2

State of the Practice of TIM

To improve TIM, attention is needed at all levels of TIM programs, including strategic, tactical, and support activities. Strategic activities focus on establishing TIM within the fabric of responder agencies through institutional structures, such as establishing a formal TIM performance measurement program and making the business case for TIM. Tactical TIM activities include operational efforts of incident response and include surveillance and detection, mobilization and response, scene management, and clearance and recovery. Support activities are typically those performed by practitioners who are not part of on-scene response and include communication, coordination, and management functions that enable incident responders to perform their jobs better and more efficiently. This chapter examines the state of the practice in TIM at all levels, as well as the current state of TIM data, and helps to set the stage for the broader objective of examining how Big Data might benefit TIM.

2.1 State of the Practice

The practice of TIM centers on the activities associated with traffic incident response, from incident detection and verification through recovery of the roadway to its normal operation. The foundation for the state of the practice in TIM can be traced to a series of publications that were created over the past decade. The use of this intellectual capital by state and local responder communities typically is spearheaded by the transportation agency. The leadership of the FHWA, along with state departments of transportation, responder agencies, industry, and academia, have created a national model for TIM.

Table 2-1 lists key information sources that provide strategies for effective TIM programs. FHWA's Traffic Incident & Events Management (TI&EM) Knowledge Management System (KMS), also called the *Traffic Incident Management Knowledgebase*, is another excellent source of TIM-related documents (FHWA 2017c).

In recent years, federal, state, and local institutions have driven significant change in the state of the practice in TIM through efforts to expand coordinated, multidisciplinary operations and to formalize TIM programs within the broader context of agency planning and operations. Improvements to individual and institutional effectiveness can be attributed in large part to:

- Establishment of local, regional, and statewide TIM committees,
- Implementation of TIM legislation,
- Development and implementation of a National TIM Responder Training Program,
- Development of local/statewide TIM strategic plans,
- Development and implementation of agency operating agreements, and
- Implementation of agency policies for safe quick clearance.

Table 2-1. Key information sources on TIM program elements and practices.

Title	Link
<i>Traffic Incident Management Handbook</i> (Owens et al. 2010)	http://www.ops.fhwa.dot.gov/eto_tim_pse/publications/timhandbook/tim_handbook.pdf
<i>Best Practices in Traffic Incident Management</i> (Carson 2010)	http://www.ops.fhwa.dot.gov/publications/fhwahop10050/fhwahop10050.pdf
<i>Field Operations Guide for Safety/Service Patrols</i> (Sparks, Schuh, and Smith 2009)	http://www.ops.fhwa.dot.gov/publications/fhwahop10014/fhwahop10014.pdf
<i>Traffic Incident Management in Hazardous Materials Spills in Incident Clearance</i> (Daniell 2009)	http://www.ops.fhwa.dot.gov/publications/fhwahop08058/fhwahop08058.pdf
<i>Traffic Control Concepts for Incident Clearance</i> (Birenbaum, Creel, and Wegmann 2009)	http://www.ops.fhwa.dot.gov/publications/fhwahop08057/fhwahop08057.pdf
<i>Federal Highway Administration Service Patrol Handbook</i> (Houston et al. 2008)	https://ops.fhwa.dot.gov/publications/fhwahop08031/ffsp_handbook.pdf
<i>Simplified Guide to the Incident Command System for Transportation Professionals</i> (Latonski and Ang-Olson 2006)	http://www.ops.fhwa.dot.gov/publications/ics_guide/ics_guide.pdf
<i>Alternate Route Handbook</i> (Dunn Engineering Associates 2006)	http://www.ops.fhwa.dot.gov/publications/ar_handbook/arh.pdf
<i>Traffic Incident Management Quick Clearance Laws: A National Review of Best Practices</i> (Carson 2008)	https://ops.fhwa.dot.gov/publications/fhwahop09005/quick_clear_laws.pdf
“Comprehensive Framework for Planning and Assessment of Traffic Incident Management Programs” (Jin et al. 2014)	https://journals.sagepub.com/toc/trra/2470/1
“A National Unified Goal for Traffic Incident Management (TIM): What Is it, and Why Is it Needed” (Corbin 2008)	https://www.pcb.its.dot.gov/t3/s080911/corbin.pdf
“Traffic Incident Management Cost Management and Cost Recovery: Executive Level Briefing” (Rensel et al. 2012)	https://ops.fhwa.dot.gov/eto_tim_pse/ppt/tim_cm_cr_exec_brief/tim_cm_cr_exec_brief.pdf
<i>Role of Transportation Management Centers in Emergency Operations: Guidebook</i> (Krechmer et al. 2012)	https://ops.fhwa.dot.gov/publications/fhwahop12050/fhwahop12050.pdf

2.1.1 Establishment of Local, Regional, and Statewide TIM Committees

The establishment of local, regional, and statewide TIM committees has led to better planning, coordination, and communications among TIM responder groups. Local TIM teams discuss tangible on-scene practices and operating considerations, and often debrief significant or problematic incidents. The local TIM team is a prominent part of Florida’s statewide strategy. In Florida, 25 individual TIM teams cover all urban and suburban areas and many rural areas of the state (Florida DOT TIM Teams 1996). The Traffic Incident Management for the Baltimore Region (TIMBR) Committee illustrates how regional stakeholders get together to plan and coordinate TIM activities as part of MPO planning efforts (Baltimore Metropolitan Council 2017).

Statewide committees support TIM at a broader, institutional level. Many state TIM groups include representatives from individual disciplines, generally represented by leadership of statewide associations or organizations for law enforcement, fire and rescue, transportation, and towing. Virginia provides an excellent example of an effective state TIM body (VA Exec. Order No. 58 [2013] and VA Exec. Order No. 15 [2015]).

In 2010, the governor of Virginia established the Virginia Traffic Incident Management Committee (since renamed the Virginia Statewide Traffic Incident Management [VASTIM])

Committee) and designated the state police and state transportation agencies to lead the effort. The VASTIM Committee has been instrumental in advancing TIM in the state (VA Exec. Order No. 58 [2013], VA Exec. Order No. 15 [2015]).

2.1.2 Implementation of TIM Legislation

TIM legislation has played an important role in advancing the state of the practice in TIM by promoting safety and quick clearance. Three principal TIM laws have been enacted to various degrees across the United States:

- “Driver Removal” laws require drivers involved in crashes to move their vehicles out of the roadway,
- “Authority Removal” laws give public officials the right to move cars and cargo at incidents, and
- “Move Over” laws require drivers to vacate the lane adjacent to emergency responders on multi-lane roadways or to slow down if they cannot safely move over or where there is only one directional lane of travel.

Move-over laws are present in every state, in the District of Columbia and in Puerto Rico. Although common, driver removal and authority removal laws are not found in every state (American Automobile Association 2017) (Carson 2008).

2.1.3 Development and Implementation of National TIM Responder Training

At the heart of national TIM progress is the National TIM Responder Training Program, which was developed under the second Strategic Highway Research Program (SHRP 2) and deployed beginning in the summer of 2012 as part of the second phase of the FHWA’s “On-Ramp to Innovation: Every Day Counts” (EDC-2) initiative (FHWA 2012). Through July 2018, more than 344,000 incident responders had attended multidisciplinary training that was developed “by responders, for responders.” The nine lessons in the training program have established a foundation for the way that traffic incidents are handled, covering important topics such as scene safety, vehicle positioning, incident command, and traffic control. The National TIM Responder Training Program has become the de-facto national standard for the state of the practice, and most preexisting state products have gone through equivalency reviews. An online version of the National TIM Responder Training Program is hosted by the National Highway Institute (FHWA n.d.-a).

2.1.4 Development of TIM Strategic Plans

Increasingly, TIM is being planned strategically in the form of guidance documents for program elements. Strategic plans at the local, regional, or state level stipulate the type of TIM activities, desired state, a time horizon, and a means to achieve objective. Strategic Highway Safety Plans (SHSPs), a type of state-level planning, are required by the FHWA for every state in which TIM has begun to find traction. Whether as a formal area of emphasis or a strategy, TIM and the components of TIM find the important nexus with safety in the SHSP.

Several states have made TIM a priority emphasis area; other states mention the value of TIM to support other safety or mobility goals like managing congestion, reducing aggressive driving, or promoting safety among vulnerable road users. The effectiveness of these plans is evidenced by the advanced state of the practice seen in states such as Florida, Oregon, and Maryland (Pecheux, Shah, and O’Donnell 2016). In addition, the FHWA notes that incorporating TIM

into the planning process is a good business practice and recommends it as a way of formalizing and institutionalizing TIM within an agency and the broader agency goals (Pecheux, Shah, and O'Donnell 2016).

2.1.5 Development and Implementation of Agency Operating Agreements

Public agencies are quite accustomed to formalizing relationships with each other to accomplish organizational objectives. Many state and local agencies have developed memoranda of understanding, operating policy statements, and agreements to work cooperatively, establish roles and responsibilities, and/or set forth targets for TIM performance. One of many examples is the Joint Operating Program Statement (JOPS) by the Washington State Patrol, Washington Fire Chiefs, and the Washington State Department of Transportation (Washington State DOT 2016a).

2.1.6 Implementation of Agency Policies for Safe and Quick Clearance

Policies for safe and quick clearance on traffic incidents have led to improvements in TIM performance and have advanced the state of the practice in TIM. In early 2011, a major policy revision in Arizona required police officers to move vehicles involved in incidents completely off the roadway (away from view) as quickly as possible. The Arizona Department of Public Safety (AZDPS) used data collected via the crash report on roadway and incident clearance times before and after this policy change to determine if the policy had an impact on TIM performance. The results showed significant reductions in both roadway and incident clearance times for non-injury and injury crashes (Pecheux 2016).

2.2 The Use of Data to Support TIM

Continued advancements in TIM will require more and better data to quantify improvements, justify funding, and guide future development of the practice. Reviewing the incident timeline sets the stage for data collection and opportunities for Big Data to improve TIM. The incident timeline is shown in Figure 2-1 (FHWA 2013b).

At any point along the timeline—detection, verification, response, roadway clearance, incident clearance, and return to normal flow—possibility exists for improvement. The identification of where these improvements could be made can be facilitated by the collection and analysis of data, including Big Data. The current state of the practice in the use of data for TIM primarily relates to TIM performance measurement and management and in making the business case for TIM programs.

2.2.1 TIM Performance Measurement and Management

TIM performance measurement and management is another way in which agencies are advancing the state of the practice of TIM. Performance measurement and management, which are becoming more and more important, require the collection of data.

In cooperation with 11 states as part of a Focus State Initiative, the FHWA has defined three national performance measures for TIM (Owens et al. 2009):

- **Roadway clearance time (RCT):** The time between the first recordable awareness of the incident by a responsible agency and the time that all lanes are available for traffic flow.

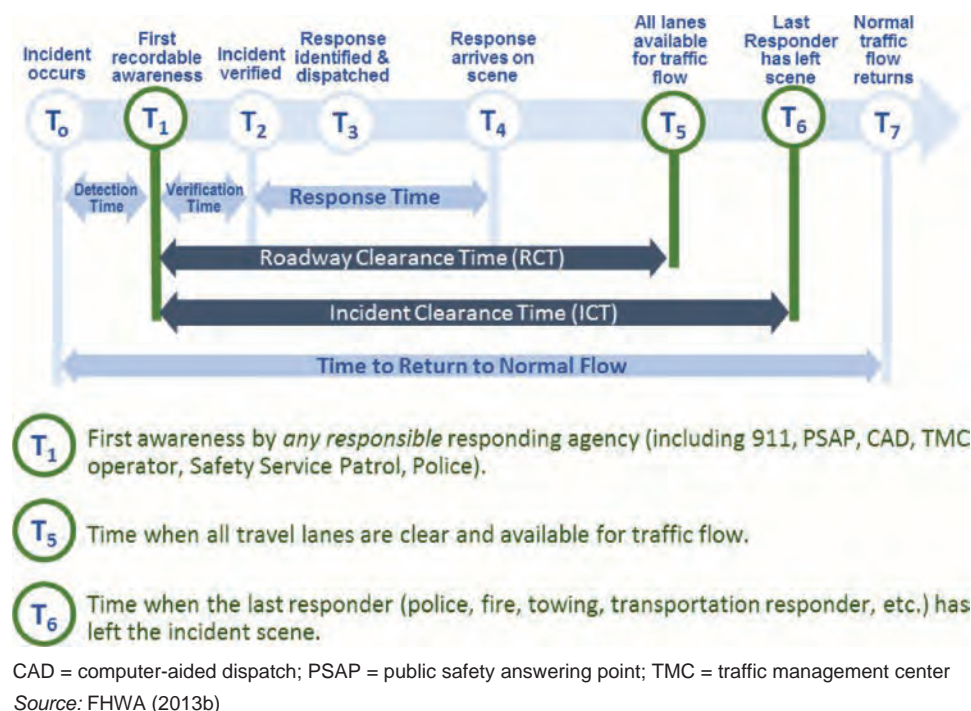


Figure 2-1. Incident timeline.

- **Incident clearance time (ICT):** The time between the first recordable awareness of the incident by a responsible agency and the time at which the last responder has left the scene.
- **Secondary crashes:** Subsequent crashes are crashes that occur at the scene of an original incident or in the queue, including crashes involving vehicles traveling in the opposite direction. (The FHWA has more recently suggested that a crash-to-crash relationship between the primary and secondary incident be used for simplicity).

The TIM timeline in Figure 2-1 shows the importance of T₁, T₅, and T₆; these points represent the data points needed to calculate RCT and ICT. Work performed under NCHRP Project 07-20, “Guidance for Implementation of Traffic Incident Management Performance Measurement,” (Pecheux, Brydia, and Holzbach 2014) and subsequent direction from the FHWA have provided guidance to agencies on TIM data collection, analysis, and performance measurement, including a comprehensive list of the types of data central to measuring TIM performance (Pecheux 2016). To date, the primary sources of data for TIM performance analysis have been transportation management centers (TMCs) and SSP programs. The state of the practice has expanded more recently to include the use of data from law enforcement crash reports and computer-aided dispatch (CAD) systems, as well as integration of data from various systems to improve the quality and quantity of data. Nonetheless, the current state of data collection and use by TIM programs varies significantly between states, between agencies/regions within states, and even within agencies. The following are highlights of the current and emerging TIM data practices from leading agencies/states:

- **Use of data from crash reports:** The AZDPS has both pioneered and championed the use of data in TIM. In 2010, AZDPS officers in the metropolitan Phoenix area began collecting additional data elements in conjunction with traffic crash investigations. Using electronic reporting software, the agency modified the officer input interface and underlying database to add RCT, ICT, and secondary crashes to the myriad data elements on the statewide reporting format. The program later expanded to AZDPS officers statewide and ultimately

led to changes in the statewide reporting format for all agencies in 2014 (Pecheux 2016). AZDPS actively measures and reports the TIM performance measures as part of its standard operating procedures.

- **Integration of TMC and SSP datasets:** The integration of TMC and SSP data increases the transportation dataset, as SSP operators handle a significant number of incidents as single responders, and many of their activities are self-initiated. The move from paper responder logs and voice communications to mobile computing platforms and smart devices to document times and activities has improved the quantity and quality of data available from SSP programs (Florida DOT 2011). In Washington State, incident response (IR) crew members enter incident data using laptop computers in their trucks to create an electronic incident report. After each shift, the data is uploaded to the Washington Incident Tracking System (WITS), a centralized statewide database (Washington State DOT 2016b). Operators at the Niagara International Transportation Technology Coalition (NITTEC) traffic operations center (TOC) in Buffalo, New York, can see both their incident entry screen and Buffalo's SSP activity log. The TOC data entry screen contains data elements for the entire incident timeline, as well as a checkbox for secondary crashes (Pecheux 2016).
- **Integration of TMC and CAD datasets:** A key step for improved data has been the integration of TMC and law enforcement computer-aided dispatch (CAD) systems. Integration enables data to be captured for a larger proportion of statewide incidents and, in particular, those outside TMC or SSP geographic and temporal operations. Minnesota, Wisconsin, and Virginia have integrated their CAD and TMC or advanced traffic management systems (ATMS) (Pecheux 2016). Other states (e.g., New Jersey) are implementing changes to facilitate integration (NJ TIM n.d.).
- **Use of crowdsourced data:** Crowdsourced mobile applications (e.g., Waze) and data consolidators (e.g., INRIX and HERE) are providing new data that agencies use to various degrees, such as in the National Performance Management Research Data Set (NPMRDS) (FHWA 2013a). These data sources offer opportunities to expand TIM practices beyond major urban freeways to suburban and rural freeways as well as major arterials. These data sources are now being applied by a few agencies for incident detection and are approaching the realm of Big Data, but this data is still on the cusp of being used for TIM performance measurement and management. Some states, including Florida and Massachusetts, have begun using Waze to supplement existing surveillance and detection systems. Other states, such as North Carolina and Iowa, use INRIX for analytics-based incident detection (Barichello and Knickerbocker 2017, Oerter 2010). Waze Connected Citizen Program data and 511 travel information system data are being integrated to enhance both datasets and to improve situational awareness for both TMC operators and Waze users (Smith 2016).
- **Use of unmanned aerial vehicle (UAV) technology:** Information specific to the location and nature of incidents enables a more effective response among fire and rescue, EMS, transportation, towing and recovery, hazardous materials, coroner, and other entities. Some agencies (e.g., in New York City and Toronto) are beginning to explore the potential of UAV technology to capture incident details for accident investigation before and during scene management (Durkin 2015).

2.2.2 Making the Business Case for TIM

Ultimately, to advance the state of the practice in TIM, TIM programs must be consistently supported and funded. The need for justification is no more evident than in the competition for agency funds, which occurs more and more often amidst dwindling agency operating budgets. TIM programs often are targeted for defunding because their value is not readily recognized. For example, SSP programs may be seen only from the lens of their motorist assistance function, rather than as a service that enhances safety, reduces roadway congestion, and mitigates the

likelihood of secondary crashes. The FHWA recently developed a guide for agencies on how to make the business case for TIM in which data plays a critical role—particularly data that documents the need for or benefits from a program’s activities in a way that can be balanced against its cost (Pechoux, Shah, and O’Donnell 2016). Few agencies maximize the use of data for this purpose. Given the growing availability of TIM data, however, the business case for TIM is one that can be more readily supported. Demonstrating the usefulness of data that supports decisions is a powerful technique for making the business case, as is evidenced by the following examples from Oregon and Maryland:

- In Oregon, maintenance crews were routinely tasked with supporting TIM functions, often at the expense of their other responsibilities. In one area of the state, the Oregon DOT used data to demonstrate the need for a dedicated incident responder to improve both traffic and maintenance operations. A maintenance position was sacrificed to create a position for a dedicated incident responder, with a positive outcome for both functions (Pechoux, Shah, and O’Donnell 2016).
- The Maryland State Highway Association analyzed data on incident clearance times to demonstrate the need for and value of expanding their TIM operations. By making the business case for TIM, the program secured funding for expanded SSP operations to include all major routes within the state and to modify the patrol hours for three of the TOCs from a 15-hour, 5-days-per-week operation to a 24-hour, 7-days-per-week operation (Pechoux, Shah, and O’Donnell 2016).

2.3 Further Advancing the State of the Practice of TIM

TIM committees, national responder training, legislation, quick-clearance policies, and operating agreements have advanced the state of the practice of TIM over the past decade. The resulting improvements in responder effectiveness, combined with emerging TIM data collection systems and processes, have positioned TIM to make another step forward in the coming years. Toward this end, the FHWA has undertaken an ambitious program to accelerate the nationwide implementation of TIM data collection and use by states. Running through calendar years 2017 and 2018, the fourth iteration of the FHWA’s Every Data Counts program (EDC-4) has been a 2-year effort to assist adopting states in gathering a greater quantity and quality of TIM data, focused on RCT, ICT, and secondary crashes. Thirty-five states worked to implement the EDC-4 TIM data innovation, and EDC-4 was successful at evolving the state of the practice in the collection and use of TIM data, as 20 states reported advancing at least one level during the 2-year period.

Although the recent and ongoing progress is promising, another step has yet to be taken from the current state of practice to apply Big Data analytics in TIM. The increased quantity and improved quality of TIM-related data shows promise that the application of Big Data can further advance the state of the practice by uncovering trends and relationships that lead to improvements in TIM strategic, tactical, and support activities. Big Data analytics have the potential to spur modifications to policies, procedures, and training, thereby improving the safety and effectiveness of incident responders, enabling them to perform their jobs better. The application of Big Data could advance the state of the practice in TIM performance management and provide ammunition to make a far more compelling business case for TIM programs and strategies. Advanced analytics could equip practitioners with information for decision support. If those analytics can be used in real time and are even predictive of traffic impact, responder actions, on-scene activities, traveler information, traffic management, and clearance strategies might be adjusted, leading to reductions in congestion and secondary crashes.



CHAPTER 3

State of the Practice of Big Data

By 1996, digital storage became a more cost-effective option for storing data than paper (SB 2016). Companies began switching to digital files to store and archive their data, which allowed for a large amount of digital data to become available and ready for analysis. At the same time, the number of Internet users started to grow, rapidly multiplying the generation of data at a rate that only increases each year, as is evidenced by the counters visible on www.InternetLiveStats.com (accessed in 2017). These two events mark the beginning of what is called Big Data. Moreover, the size of Big Data only continues to grow. In 2013, anything over 500 gigabytes (GB) was considered Big Data; now, Big Data involves terabytes (TB) of data.

3.1 Big Data Definition

As a popular term, *Big Data* often is used simply to mean a volume of data that is so massive it is difficult to process using traditional database and software techniques; however, the volume of data only tells part of the story. To better understand what constitutes Big Data, it is helpful to distinguish among three types of data:

- **Structured data:** The data in traditional relational databases follows a specific structure and format and resides in a fixed field within a record according to a database *schema*. Traditional relational databases require that the ingested data be processed through a process called ETL (extract-transform-load) to formally organize the data before it can be queried. The queries themselves are commonly written using a structured query language (SQL). Familiar names associated with relational databases include Oracle, MySQL, Microsoft SQL Server, and PostgreSQL.
- **Semi-structured data:** This is a form of structured data that does not conform to the formal structure of the data models associated with relational databases but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Extensible mark-up language (XML) and JavaScript Object Notation (JSON) are open standards for structuring semi-structured data.
- **Unstructured data:** This term usually refers to data that is not organized and that does not reside in a traditional relational database. Examples of unstructured data are e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages, and many other kinds of business documents.

Significantly, Big Data expands the scope of traditional relational databases to include data that is unstructured. Big Data storage and analytics take place in a *distributed computing environment*, meaning that these tasks are performed across multiple servers and/or in the cloud. An effective rule of thumb is that if the analysis can be run on a laptop or workstation, it is not Big Data.

Big Data datasets often are characterized using five attributes, referred to as the “five Vs”: volume, variety, velocity, veracity, and value.

3.1.1 Volume

Volume characterizes the main aspect of a Big Data dataset. In 2007, a manufacturer of data storage devices predicted that the size of the digital universe in 2010 would be close to 988 exabytes, and that it would grow by 57 percent every year (Gantz 2007). In 2010, Thomson Reuters estimated in its annual report that the world was “awash with data—800 exabytes and rising” (Thomson and Gloer n.d.). In 2013, IBM estimated that the world produced about 2.5 billion gigabytes (GB)—equivalent to 2.5 EB—each day and that 80 percent of that data was unstructured (IBM 2013). The Intelligence Community Comprehensive National Cybersecurity Initiative Data Center, which opened in Utah in 2014, is one of the largest data centers in the world, with an estimated storage capacity between 3 EB and 12 EB. The center occupies an area of about 1,500,000 million square feet and cost \$1.5 billion to build (Lima 2015).

Figure 3-1 shows a representation of the data size scale. Currently, Big Data is generally considered more than 1 TB; however, the size characterization of Big Data is continuously changing. For example, in 2013 the 90 petabytes (PB) of data stored by eBay was considered a large volume; by comparison, in 2017, Walmart was handling 200 billion rows of transactional data every few weeks, pulling in information from 200 streams of internal and external data, including meteorological data, economic data, Nielsen data, telecommunications data, social media data, gas prices, and local events databases, and processing 2.5 petabytes of data every hour (Tay 2013, Marr 2017).

Crash data and most TMC data are generated on a much smaller scale. Five years of crash data from Florida represents less than 50 megabytes (MB). A year’s worth of data in the National

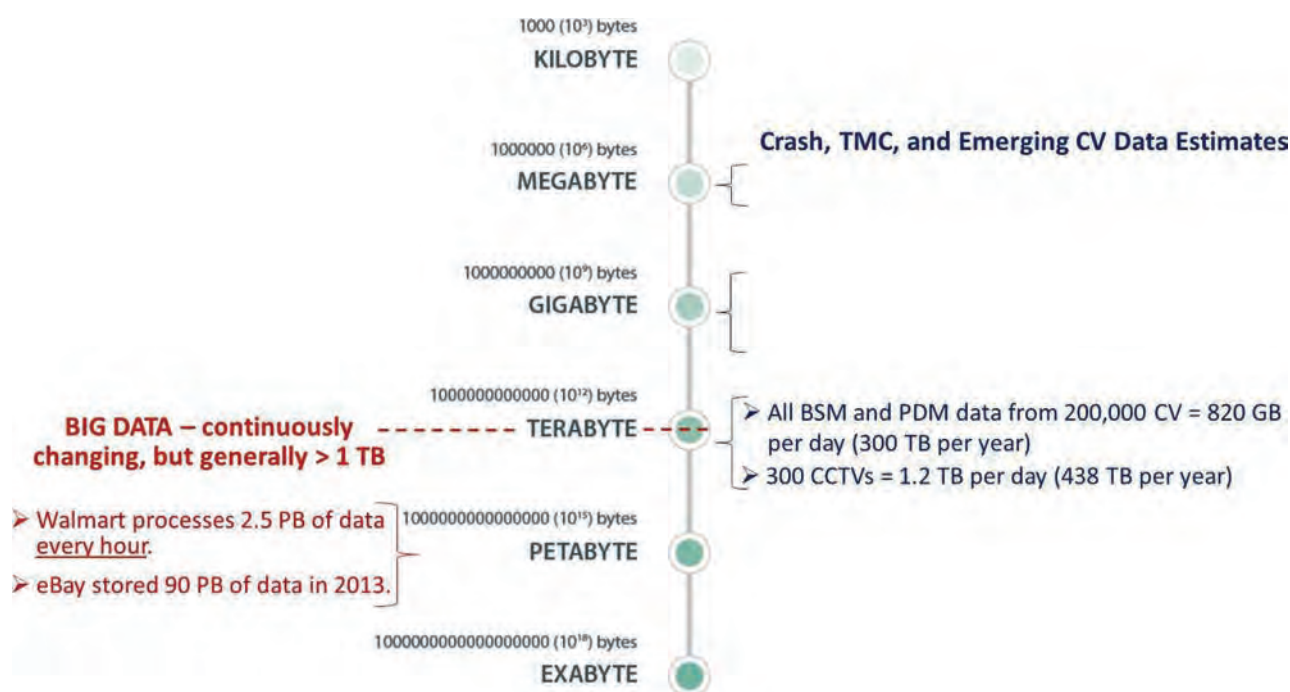


Figure 3-1. Data size scale with example dataset sizes.

For a sense of scale, consider that:

1 exabyte (EB)	=	1,000,000,000,000,000 bytes
1 petabyte (PB)	=	1,000,000,000,000,000 bytes
1 terabyte (TB)	=	1,000,000,000,000 bytes
1 gigabyte (GB)	=	1,000,000,000 bytes
1 megabyte (MB)	=	1,000,000 bytes
1 kilobyte (KB)	=	1,000 bytes

Emergency Medical Services Information System (NEMSIS)—consisting of 30.2 million records from 15,000 EMS agencies—makes up about 40 gigabytes (GB). On the other hand, the data generated by 300 TMC field devices currently is estimated at approximately 635 GB per year and, if stored, the data from 300 closed circuit television (CCTV) cameras would require hundreds of terabytes of storage each year (Gettman et al. 2017). Likewise, emerging connected vehicle data is expected to generate many terabytes of data per year.

3.1.2 Variety

Variety is one of the most interesting characteristics of Big Data datasets. As new information is created and older data is digitized, the diversity of data that can be processed and analyzed also is growing. Traditional data analysis, performed using relational databases or statistical software, only allowed for “table friendly” (i.e., structured) data to be processed. Some kinds of information, like that contained in a traditional bank statement (e.g., data, amount, balance, and time) can be expressed using data fields and can fit neatly in a relational database without extensive manipulation (i.e., ETL). Unstructured data (e.g., images or free text) is not table friendly. Without manual processing of the content, unstructured data can only be stored within tables as a series of unsearchable objects, and the ability of relational databases and statistical software to analyze such objects is limited.

Big Data technologies do not require data to be neatly organized (structured) to be searched. Unstructured data like Twitter feeds, email content, audio files, MRI images, webpages, or web logs now can be processed directly as part of a Big Data query. No pre-processing is required, which greatly augments the amount of data that can be exploited. Consequently, virtually anything that can be captured and stored digitally can be analyzed and queried, even if the digital content does not include a meta model (i.e., a set of rules that defines a class of information and how to express it) that neatly defines it. Unstructured data is fundamental to Big Data, and one of the main goals in leveraging Big Data technology is to make sense of unstructured data.

3.1.3 Velocity

Velocity refers to the speed or frequency of data coming into Big Data datasets. Velocity adds another dimension to the increasing scale of Big Data datasets, particularly in regard to the complexity of processing this flow of data. When thinking about the frequency of text messages (technically, short message service [SMS] messages), social media status updates, or credit card swipes sent over the Internet daily, it is easy to have an appreciation for velocity. Not only do large amounts of unstructured data need to be processed rapidly, but that data is being augmented or modified constantly. Credit card fraud detection is a good example of a

rapidly changing Big Data dataset that needs to be processed quickly to catch suspicious transactions and deny payments.

3.1.4 Veracity

Veracity refers to the trustworthiness of the data in Big Data datasets. Traditional data analysis requires the raw data to go through an ETL process to be reformatted, cleaned, and purged of illogical, erroneous or outlying data. In contrast, Big Data datasets are all-inclusive; data is stored “as is” with minimal processing before being queried. Traditional data analysis through ETL could ensure that the data being queried is of high-quality and accuracy and can be trusted. Because Big Data datasets are all-inclusive, the quality, accuracy, and trustworthiness of the data is not guaranteed on query and therefore needs to be assessed by applying domain knowledge to the output to verify/validate the data or by exploring the data with separate queries for data validation. It is interesting to note that the newer the data, the less knowledgeable we are about it; as such, the trustworthiness of the data can only be derived from the patterns or trends observed in the data.

3.1.5 Value

Value denotes how Big Data datasets contribute to improving the status quo. Value involves determining a benefit and estimating the significance of that benefit across any conceivable circumstance. Value may be the most important of the five Vs, as investments in Big Data initiatives require a clear understanding of the benefits and associated costs. Before any attempt to collect or leverage Big Data, business cases need to be developed to assess the benefits and costs associated with data collection and analysis efforts.

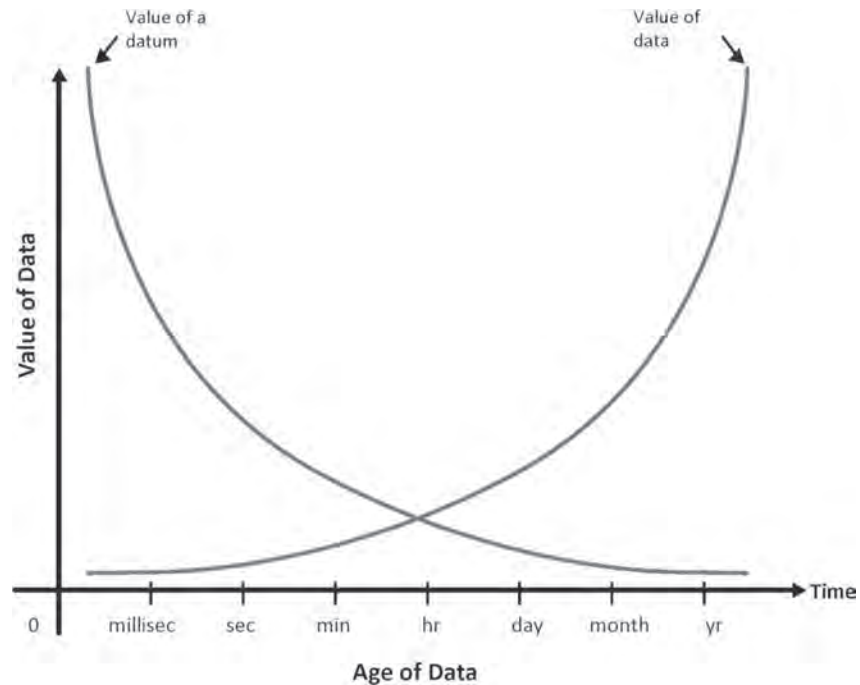
3.2 The Move from Traditional Data Analysis to Big Data Analytics

3.2.1 Traditional Data Analysis

Data analysis can be separated into two main categories: historical data analysis and real-time data analysis. Historical data analysis is the analysis of a large set of data collected over time to identify patterns or outliers. Traditionally, this type of analysis has been referred to as online analytical processing (OLAP). Typical applications of OLAP include business analytics such as reporting for sales, marketing, management reporting, business process management, budgeting and forecasting, and financial reporting. Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad hoc queries with a rapid execution time (Mailvaganam 2007).

Real-time data analysis is the analysis of a single datum within a few moments of its creation to assess its quality or react to its content. Traditionally, this type of analysis has been referred to as online transactional processing (OLTP). OLTP applications facilitate and manage transaction-oriented applications. The key goals of OLTP applications are availability, speed, concurrency, and recoverability (Oracle 1999). An example of real-time data analysis is the detection of fraudulent credit card transactions and the blocking of such transactions within seconds of their submission because the data does not appear to be in line with the previous purchases made by the account holder.

Figure 3-2 illustrates the value of both types of analysis. For OLTP, the value of a single datum is very high immediately after it has been created; a quick analysis can lead to immediate



Source: Adapted from VoltDB, Inc. (Stonebreaker and Jarr 2013)

Figure 3-2. Value of data.

corrective or augmentative action(s). As the datum ages, analysis supporting immediate actions is less valuable. Conversely, for OLAP, the value of a single datum is very low immediately after its creation. However, as a collection of data is created over time, the data accumulates into larger and more diverse datasets that can be analyzed effectively to reveal patterns and trends that inform and improve decision-making.

Before the advent of Big Data analytics, both OLTP and OLAP generally were performed using relational database management systems (RDBMSs). Although relational databases are a reliable way to store and search data, they tend to be strict. Consequently, the view of the world through the lens of a relational database is restricted. In addition, the schema used within a relational database is not easily changed, sometimes requiring months or years to modify. Relational databases were designed at a time when data did not change rapidly; therefore, they are not designed to handle change.

Although the use of relational databases has been satisfactory, the advent of larger, more complex, and more frequently changing datasets (i.e., datasets with greater volume, variety, and velocity) has rapidly increased the cost of developing and operating data stores using relational databases. These changes have led database architects and developers to seek less expensive, albeit more complex, alternatives to store and analyze new, large, and intricate datasets.

The shift from relational databases to Big Data began in the early 2000s, when online companies sought to index the content of the entire Internet to make it efficiently searchable. Even in those days, the Internet (essentially a very large dataset) held content so diverse it could not be organized into a relational database schema. Exponential growth and the very rapid pace of change in content and uses of the Internet contributed additional indexing challenges. Engineers faced four distinct issues in building a tool to complete the desired index, as follows:

1. The tool had to be schema-less (i.e., it could not be based on tables and columns).
2. The tool had to be durable (i.e., once written, data should never be lost).

3. The tool had to be capable of handling component failure (e.g., failure of the CPU, or memory, or of the network).
4. The tool had to be capable of automatically re-balancing its resources (e.g., allocating disk space consumption).

The solution was the development of Hadoop: an open-source, Java-based programming framework that supports the processing and storage of extremely large datasets in a distributed computing environment.

3.2.2 Hadoop: The Start of Big Data Tools

In the pursuit of an efficient way to index and search the Internet, efforts first focused on developing a file system capable of storing the data collected from the entire Internet. The file system had to run across multiple servers and be able to meet complex requirements. This indexing file system became known as the Hadoop Distributed File System (HDFS). Next, efforts focused on developing a rapid processing framework that could handle the data stored on the new file system in a fault-tolerant, distributed, and parallel fashion across all the servers. The processing framework became known as MapReduce. HDFS and MapReduce were then merged into a single product called Hadoop.

Hadoop makes it possible to run applications on systems with thousands of inexpensive servers (nodes) and to handle thousands of terabytes of data. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating in case of a node failure. This approach lowers the risk of catastrophic system failure and unexpected data loss, even if a significant number of nodes becomes inoperative. Hadoop quickly emerged as a foundation for Big Data processing tasks such as scientific analytics, business and sales planning, and processing enormous volumes of sensor data like that generated by the Internet of Things (IoT).

Since its initial release in 2011, Hadoop has been continuously developed and updated. Organizations have adopted it, modified it, and used it as a basis for new Big Data tools. Contrary to the approach taken with relational databases, the development of Hadoop did not culminate in a single specialized tool; rather, the indexing and processing framework has evolved as a series of specialized tools, each with distinct capabilities ranging from simple aggregation to complex text analysis and image analysis, and able to work on both historical and real-time datasets. Modified versions of Hadoop can now be found among the many cloud provider services. Turnkey services now available from many commercial providers can analyze extremely large and varied datasets, either historically or in real time, without incurring the large cost associated with building and maintaining separate server clusters.

3.2.3 Current Big Data Tools

Big Data analytics is not bound to a single set of tools to perform an analysis; rather, it encompasses a wide variety of proprietary and open-source tools that can be customized and modified by users. This section provides brief descriptions of the types of tools that compose the Big Data ecosystem.

3.2.3.1 Hadoop-Based Programming Frameworks

Based on the Hadoop software library created by the Apache Software Foundation, these programming frameworks allow for the distributed processing of large datasets across clusters of computers using a simple

Hadoop-based programming frameworks allow for the distributed processing of large datasets across clusters of computers using a simple programming model.

programming model. They can scale up from single servers to thousands of machines, each offering local computation and storage.

These frameworks are not databases. They store data, and users can pull data from them, but there are no queries involved. Data is stored on a distributed shared file system and then processed into a new dataset using a distributed processing framework such as MapReduce. The resulting dataset can then be retrieved by users. The data processing runs as a series of jobs, with each job essentially a separate Java application that goes into the data and pulls out information as needed.

This approach gives data analysts a lot of power and flexibility in comparison to the traditional SQL queries used with relational databases. Analysts can customize their jobs as needed, adding additional software such as text mining or image analysis software libraries, to process unstructured data like emails or photos. This flexibility also adds a lot of complexity to the data mining process. Customizing jobs to incorporate text mining, image analysis, or other software typically requires software programming knowledge, whereas executing SQL queries generally does not.

Hadoop-based programming frameworks are now being greatly modified to optimize their ability to manage their data and run concurrent jobs more efficiently. Many modifications have already been made to take advantage of memory storage instead of disk storage, as memory storage has become less expensive. These improvements have afforded the frameworks the ability to process very large datasets in batch (i.e., conduct historical analysis) and the ability to conduct real-time processing of large amounts of data flowing into the framework storage. Common Hadoop-based programming frameworks include Apache Hadoop, Apache Spark, Apache Storm, and AWS Elastic MapReduce.

3.2.3.2 NoSQL Databases

NoSQL databases are used to combine information from several sources into one comprehensive database and subsequently to run aggregation and filtering queries on the very large dataset.

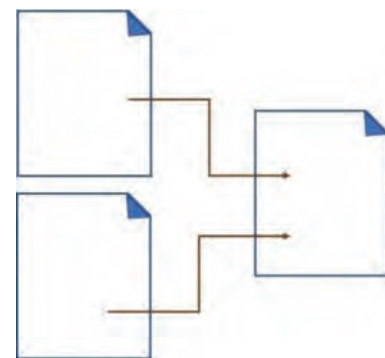
NoSQL databases are databases that began to be built in the early 2000s for large-scale database clustering in cloud and web applications. NoSQL databases are essentially Hadoop-based frameworks with an added interface to allow data to be queried. The query interface helps convert the query language for use in distributed jobs. The query layer works in combination with a query language.

NoSQL databases cannot offer the same consistency as relational databases and often are limited in their ability to run complex data analyses. Consequently, NoSQL databases are used more often for combining information from several sources into one comprehensive database and subsequently running aggregation and filtering queries on very large datasets. NoSQL databases do not require an established relational schema, but they often are used in combination with relational databases. Large-scale web organizations use NoSQL databases to focus on narrow operational goals and employ relational databases as add-ons when higher data consistency and data quality is necessary. Four types of NoSQL databases are:

- **Key-value databases:** Also called *key-value stores*, these databases implement a simple data model that pairs a unique key with an associated value. Because of their simplicity, key-value databases can lead to the development of extremely “performant” and highly scalable databases for session management and caching in web applications. (The word *performant* is a French word essentially meaning “able to perform at or above an expected level.” In software engineering, the term is commonly used to describe efficient and well-optimized software applications.) Implementations differ in the way they are oriented to

Key	Value
Incident 1	time, location, severity
Incident 2	time, location
Incident 3	time, severity, vehicles
Incident 4	time, district, responders

work with random access memory (RAM), solid-state drives, or disk drives. Examples of key-value databases include Aerospike, Berkeley DB, MemcacheDB, Redis and Riak.



- **Document-oriented databases:** Also called *document stores*, these databases focus on efficient storage, retrieval, and management of document-oriented information or semi-structured data, as well as on descriptions of that data in document format. They allow developers to create and update programs without the need to reference a master schema. Document-oriented databases often are used in combination with the scripting language JavaScript and its associated data interchange format, JSON, but XML and other data formats also are supported. Document-oriented databases often are used for content management and mobile application data handling. Examples of document-oriented databases include Couchbase Server, CouchDB, DocumentDB, MarkLogic, and MongoDB.

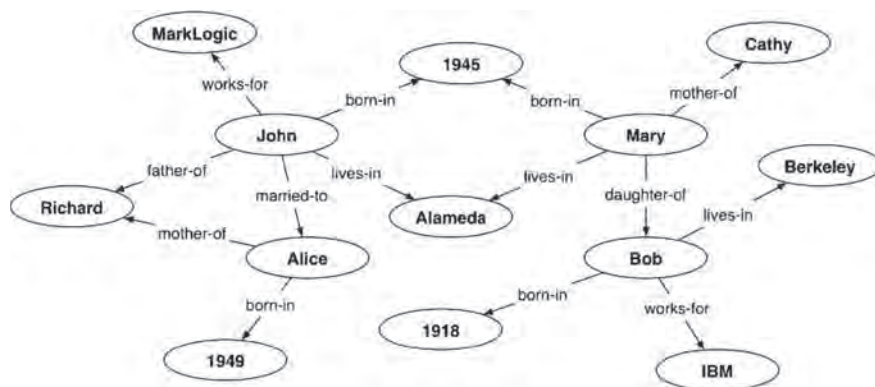
- **Wide-column stores:** These databases organize data tables as columns instead of rows. In wide-column stores, a column *family* consists of multiple rows. Each row can contain a different number of columns, and the columns (names and data types) do not have to match the columns in the other rows. Within its row, each column contains a name-value pair and a timestamp. Wide-column stores can be found in both SQL and NoSQL databases (Ian 2016). Wide-column stores can query large data volumes faster than conventional relational databases can. Wide-column data stores often are used for recommendation engines, catalogs, fraud detection, and other types of data processing. Examples of wide-column stores include Google BigTable, Cassandra, and HBase.

Incident 1	Time	Location	Severity
	Value	Value	Value
	Timestamp	Timestamp	Timestamp

Incident 2	Time	Location	
	Value	Value	
	Timestamp	Timestamp	

Incident 3	Time	Type	Vehicles Involved
	Value	Value	Value
	Timestamp	Timestamp	Timestamp

- **Graph stores:** Also called *graph databases*, graph stores organize data as nodes and edges, which represent connections between nodes. For example, a node could be a person, another node could be a specific name, and the edge (the connection or relationship between the two nodes) could be “has this name.” Because the graph system stores the relationship between nodes, it can support richer representations of data relationships. Also, unlike relational models that rely on strict schemas, the graph data model can evolve and adapt to data changes over time without requiring a complete redesign. Graph stores are applied in systems that must map relationships, such as reservation systems or customer relationship management systems. Examples of graph stores include AllegroGraph, IBM Graph, Neo4j, and Titan.



Source: <https://www.marklogic.com/blog/making-new-connections-ml-semantics/>

3.2.3.3 NewSQL Databases

NewSQL databases are basically clustered relational databases that are augmented with Hadoop-inspired, distributed, fault-tolerant architectures. The goal of NewSQL databases is

The goal of NewSQL databases is to deliver high availability and performance without sacrificing the robust consistency requirements and transaction capabilities found in relational databases.

to deliver high availability and performance to NoSQL databases without sacrificing the robust consistency requirements and transaction capabilities found in relational databases. NewSQL databases also support the standard relational database language, SQL, to access and modify their data. NewSQL databases are usually employed in applications within which many short database transactions accessing small amounts of indexed data are executed repetitively. These applications are typical of OLTP processing for activities such as shopping cart management or mobile phone tracking.

3.2.3.4 In-Memory/Graphics Processing Unit–Accelerated Databases

GPU-accelerated databases can perform queries sometimes hundreds of times faster than in-memory NewSQL databases and can search through billions of records in less than a second.

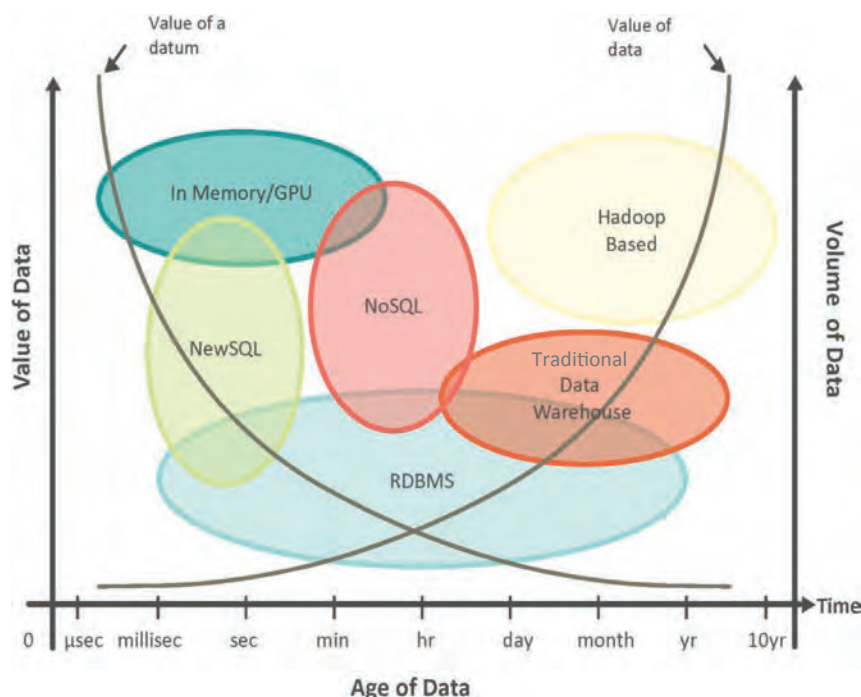
In-memory/graphics processing unit–accelerated databases (GPU-accelerated databases) are like NewSQL databases, but they use GPUs (microchips originally designed for video processing) instead of CPUs (central processing units) to perform query operations. GPU-accelerated databases can perform queries sometimes hundreds of times faster than in-memory NewSQL databases, and they can search through billions of records in less than a second. GPU-accelerated databases are still relatively new, but they are starting to generate interest because they often require fewer servers per cluster than NewSQL

databases, which offsets the additional cost of server GPUs. Examples of GPU-accelerated databases include Kinetica, MapD, and Blazing DB.

3.2.3.5 Summary

Many tools have been derived from the original Hadoop software. From leveraging in-memory and GPU-processing to incorporating relational database standards, the number and specificities of Big Data tools keep growing, but one convention seems to be common to all these tools: *schema-on-read*. The schema-on-read convention imposes a structure on raw data *after* it has been stored and as it is being read or queried. This approach contrasts with the *schema-on-write* convention—the foundation of relational databases—which imposes a structure *before* the data has been stored (i.e., ETL). The schema-on-read approach was not possible in earlier systems, as they did not have the capabilities required to handle less-structured data. As both hardware and software capabilities have increased, schema-on-read has now emerged as the main approach to organizing Big Data.

Figure 3-3 builds on the data value chart from Figure 3-2 by adding the volume processing capabilities (*y*-axis on the right) of traditional and new Big Data analytics tools. Looking at Figure 3-3, it is easy to identify how Big Data tools have brought significant improvement in handling the analysis of greater volumes of data, as well as in handling a wider range of data based on data age (from more rapidly changing, newer data to fixed, older data). Traditional RDBMSs running on a single server are limited in that they typically have difficulties ingesting and analyzing large amounts of data in real time, as compared to Big Data databases. Relational databases also have difficulties performing quick analyses on large datasets covering several years without pre-calculating and pre-aggregating the historical data (e.g., in a data cube). These limitations can be attributed to the limits of relational database server hardware (e.g., memory, network, CPU, storage). Additionally, relational databases are based on relational algebra, which creates strict models of how data can be stored and queried. Because they work on a schema-on-write basis, data entered into relational databases must be prepared and tailored to a template or database schema at the time of entry (e.g., using the ETL process) before any queries or analysis can occur. When users query the data in a traditional RDBMS, the data



Source: Adapted from VoltDB, Inc. (Stonebreaker and Jarr 2013)

Figure 3-3. Big Data exploitation of data value.

has already been organized into an easily manageable format that facilitates sorting, merging, aggregating, and calculating.

It should be noted that a traditional data warehouse—which is a complex analytical system composed of one or more relational databases—is not the same as a Big Data store. Traditional data warehouses were designed for historical analysis and deal with larger and more complex datasets by applying a “divide and conquer” approach, splitting the tasks of importing and organizing the data across multiple custom ETL processes and multiple domain-specific relational databases. Traditional data warehousing systems are very complex and difficult to maintain in the face of ever increasing and changing data.

3.2.4 Big Data Architecture

The architecture used to support the development of Big Data stores is called the Lambda architecture. The Lambda architecture is a data processing architecture designed to handle massive quantities of data by taking advantage of both batch-processing (i.e., historical data) and stream-processing (i.e., real-time data) methods. The Lambda architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data. The outputs of batch-processed historical data and real-time data streams may be joined before presentation of the data. The rise of the Lambda architecture has correlated with the growth of real-time analytics.

Figure 3-4 shows a graphical representation of the Lambda architecture. The left part of the chart shows the many data inputs (including geodata, sensor data, mobile data, logs, and so forth) entering the Big Data store through a common gateway. The data is then streamed to the historical data analysis system (shown in blue, in the top, shaded area) and the real-time data analysis system (shown in red in the bottom, shaded area). The historical data analysis system

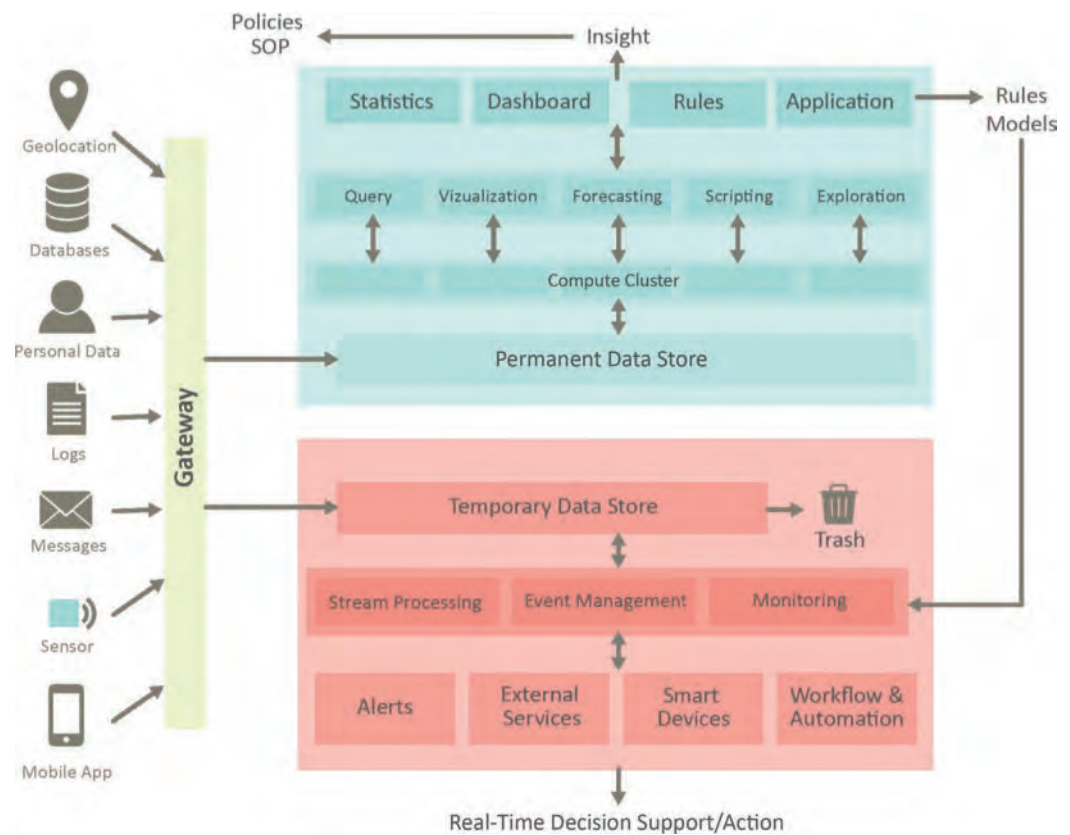


Figure 3-4. The Lambda architecture for a permanent data store.

is composed of a distributed storage system where data is archived indefinitely. A server cluster processes the stored data using various Big Data frameworks and databases to allow users to explore and query the data, create visualizations and dashboards, classify data, identify patterns or trends, and create rules or predictive models. The results of the queries, visualizations, or trends are then used to act on policies or standard operating procedures (SOP) or to revise strategic goals. The predictive models and rules are sent to the real-time data analysis layer to be tested and implemented.

The real-time data analysis system (shown in red) also is composed of a distributed storage system, but in this system streaming data is kept for a fixed period and then archived or discarded. A server cluster also processes the ever-changing data using real-time Big Data analysis tools to allow users to monitor the flowing data, detect anomalies, and predict upcoming events using the models and rules developed in the historical data analysis system. The results of the monitoring, detection, and prediction algorithms are then used to support real-time decisions/actions through email or mobile application alerts or by directly triggering actions on external workflow or devices.

3.2.5 Examples of Big Data Analytics

Although many traditional statistical techniques can be applied to Big Data analysis, newer techniques go beyond numbers to leverage text and image exploitation as well as machine learning. A key differentiator in Big Data analytics is the use of inductive statistics for pattern detection, generalizations, and predictions from large datasets with low information density by leveraging non-linear systems such as neural network models. Challenges with the application

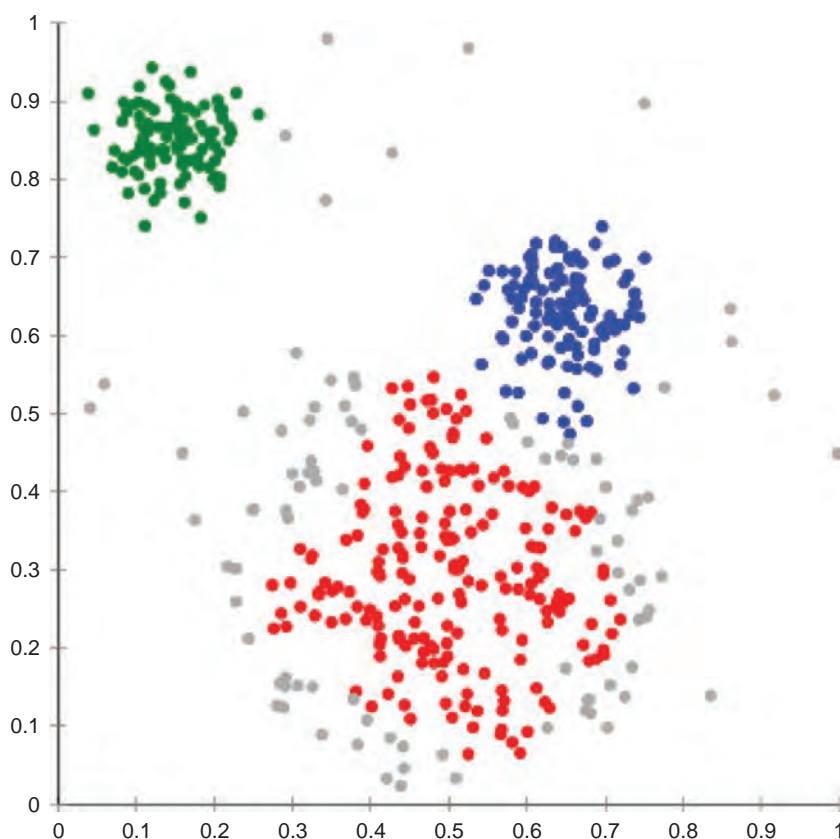
and use of Big Data analytics arise from the absence of theory to drive the analytics and critical judgment in interpreting the analytics. These shortcomings are of particular concern for evolving social systems.

This section describes a few examples of Big Data analytics, how they are performed, what kinds of results or insights they can provide, and what tools can be used to perform them.

3.2.5.1 Classification Using Clustering Analysis

Clustering analysis is a data mining task that consists of grouping a set of records in such a way that objects in the same group, called a *cluster*, are more like each other than they are to objects in other groups or clusters. Clustering analysis is a main task of exploratory data mining and a common technique for statistical data analysis, and is used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. The technique allows the categorization or grouping of records in datasets to uncover their natural organization or the natural affinities between records or groups of records. Figure 3-5 shows a generic example of a set of data points that have been grouped into three clusters—green (at top), blue (lower and toward the right) and red (roughly from the middle and extending to the bottom)—using a two-dimensional clustering analysis (Chire 2011). Note that not all points have been added to a cluster.

Many algorithms can perform clustering analysis. One of the most popular clustering algorithms, K-means, aims to partition a limited number of data points into a specified number



Source: Chire (2011), CC BY-SA 3.0

Figure 3-5. Example of two-dimensional clustering analysis visualization.

of clusters in which each observation belongs to the cluster with the nearest mean. Examples of software programs capable of performing clustering analysis include Apache Mahout, Apache Spark, and Revolution R Enterprise.

Example Application: The following scenario shows one way a clustering analysis could be used in a TIM application. A police department wants to make its presence in the field more efficient and decides to station its patrol cars so that patrol cars are near areas with high incident rates. A clustering analysis can be used to identify the best locations for the patrol cars, so that the right people and resources can be in position to respond to incidents more quickly.

3.2.5.2 Text Analysis

Text analysis, also called *text mining*, refers to techniques that extract information from textual data sources such as social network feeds, emails, blogs, online forums, survey responses, corporate documents, and news articles. Text analysis involves statistical analysis, computational linguistics, and machine learning. A popular Big Data text analysis performed on social media data is called *sentiment analysis* or *opinion mining*. Sentiment analysis is widely used in marketing and finance, and also in the political and social sciences. This type of text analysis analyzes social media messages that contain people’s opinions about “entities” such as products, organizations, or individuals, and about events such as traffic incidents. Figure 3-6 shows an example of one type of visualization that can be generated by text analysis: a word cloud that represents the most frequently occurring terms encountered in a set of recruiting documents. A word cloud is a visual representation of text data in which the font size or color of each word indicates its frequency or importance.

Many distinct text mining libraries exist, such as tm, NLTK, or GATE. Generally, their application to Big Data datasets occurs in three stages:

The first stage, called the *information retrieval* stage, involves the retrieval of plain text from semi-structured documents such as word-processing documents, social media posts, or even emails. This stage may include or be followed by *natural language processing*, which identifies grammatical, usage, or other features in the text to facilitate its use in computations and algorithms.

The second stage, called the *information extraction phase*, is the stage during which text mining libraries are used to mark up the text to identify meaning. During this phase, the *text corpus* (the entire text dataset) is augmented using metadata about the text. The metadata can



Figure 3-6. SSP word cloud example.

be information about the text (e.g., its author, title, date, edition) and/or information that has been extracted using the text mining libraries (e.g., all names or locations mentioned in the text).

The third stage, called the *data mining phase*, is when Big Data tools are used to perform analysis on the augmented text corpus to extract information and identify relationships between texts. The results of this third-stage analysis always reflect the preconceptions of those who created the metadata. Some examples of the types of analysis that can be performed on an augmented text corpus include:

- **Text categorization:** cataloguing texts into categories;
- **Text clustering:** clustering groups of automatically retrieved text into a list of meaningful categories;
- **Concept/entity extraction:** locating and classifying elements in text into predefined categories, such as persons, organizations, locations, monetary values, and so forth;
- **Granular taxonomies:** enabling organization or classification of information as a set of objects that can be displayed as a taxonomy;
- **Sentiment analysis:** identifying and extracting subjective information in source materials (e.g., emotions or beliefs);
- **Document summarization:** creating a shortened version of a text containing the most important elements; and
- **Entity relation modeling:** automated learning of relationships between data items.

Examples of software capable of performing text analysis include Apache Spark Machine Learning Library (MLlib) and Microsoft Azure Cognitive Services text analytics API.

Example Application: During the 2012 presidential election campaign, President Barack Obama's campaign team applied sentiment analysis to Twitter posts to identify swing voters and to spot the campaign discussion topics most likely to make these voters change their minds. The discussion topics identified were then used to create custom advertising for each of the identified swing voters (Issenberg 2012).

3.2.5.3 Image Analysis

Image analysis, also called image analytics, is a Big Data analysis performed on streamed (video) or archived image content. Image analysis involves using a variety of techniques to analyze, extract, and monitor meaningful information detected within images. This type of analysis is already being applied to closed-circuit television (CCTV) camera systems and video-sharing websites, primarily in relation to retail marketing and operations management. The images generated by CCTV cameras in retail outlets are extracted for business intelligence (Rice 2013). Algorithms allow retailers to measure the volume and movement patterns of customers in the store and to collect demographic information about customers—such as age, gender, and ethnicity—from video content. Valuable insights are then derived by correlating the extracted information with customer demographics to drive decisions about product placement, price, promotion, layout, and staffing.

Figure 3-7 shows an image that was analyzed using one of the Big Data image processing services (Amazon 2017). The table shows the list of terms detected by the image processing service, as well as the level of confidence.

For a long time, image analysis has been conducted using a process that converts color images to gray scale images; locates geometric shapes by means of edges, shades, or other defining features; and combines them to identify relevant image elements such as the location of a nose or eyes on a face. From the location and distance between these discovered features, an assessment can be made as to what the features together mean (e.g., the gender associated with

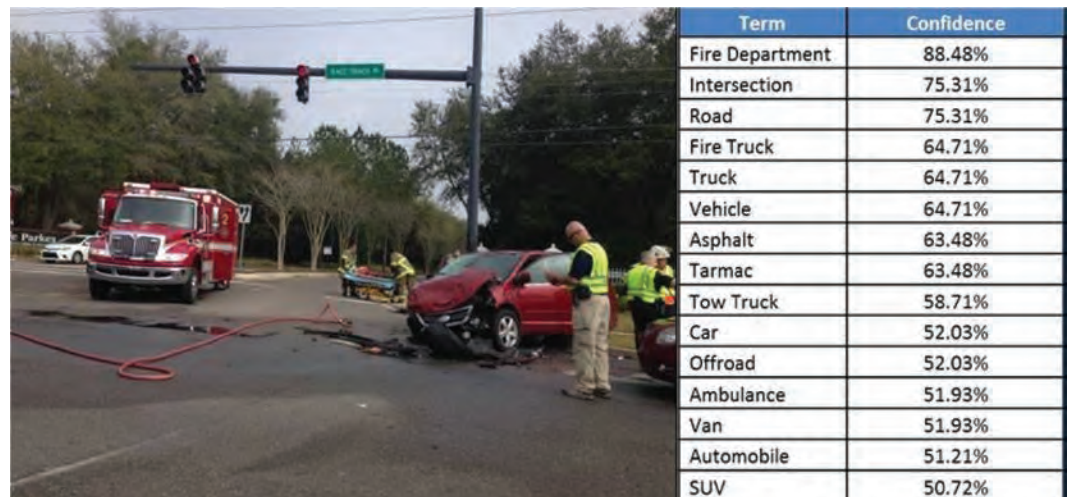


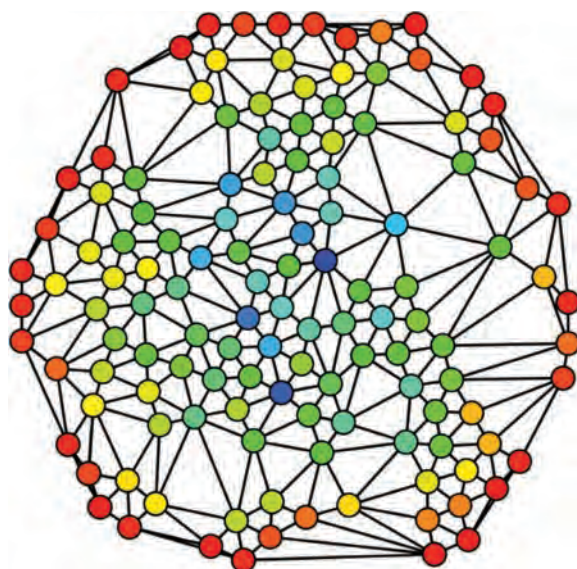
Figure 3-7. Traffic incident scene and associated image recognition results.

a face). Recently, important progress has been made in the field of neural networks to improve analysis performance. Neural networks are computational models that emulate the structure and functions of biological neural networks (brains). Neural networks are not new, but their application to image analysis has been rather unsuccessful in the past, mostly because of cost-prohibitive computing and a problem called *overfitting*. Overfitting occurs when a model (in this case, a neural network) tailors itself too closely to the data on which it has been trained and is not able to perform well on new data. Combined with the massive number of images generated by social media, advances in the design of neural network structures (often called *deep learning*) have allowed neural networks to become very effective at image analysis. Neural networks are now the basis of many Big Data image analysis software systems, whether custom or turnkey. Examples of software capable of performing image analysis include AWS Rekognition, Google Cloud Vision, and IBM Watson Visual Recognition.

Example Application: The public TV channel C-SPAN, which provides gavel-to-gavel proceedings of the U.S. House of Representatives, U.S. Senate, and other forums where public policy is discussed, debated, and decided, recently started to use a cloud-based Big Data image analysis service to tag its archived videos and associate each video frame with information such as who is speaking, who is on camera, and other details. The goal was to allow C-SPAN's video content to be easily indexed and made searchable. By performing image recognition analysis on more than 7,500 hours of video frame content, C-SPAN has been able to identify more than 97,000 entities, create a new database to store the newly indexed content, and allow its video archive to be searched much more effectively than before (Amazon Web Services n.d.).

3.2.5.4 Graph Analysis

Graph analysis techniques are derived from graph theory and are primarily based on the analysis of the structure of data and how data elements relate to each other. Social network analytics, for example, may use graph analysis to structure attributes of a social network and extract intelligence from the relationships among the participating entities. The structure of a social network can be modeled through a set of nodes and edges that represent participants and their relationships. The model can be visualized as a series of graphs composed of the nodes and the edges. The graphs can be mined to identify communities and influencers or to identify the shortest path between two individuals. This type of analysis is commonly found in social media and advertising, enterprises in which the insights gained can be leveraged in viral marketing to



Source: Rocchini (2007)

Figure 3-8. Example of graph node centrality (betweenness centrality) analysis.

enhance brand awareness and adoption. Figure 3-8 shows an example of the results of a graph analysis called *betweenness centrality*. This analysis identifies which nodes within a graph are the most connected (blue) and which are the least connected (red) (Rocchini 2007).

Many techniques are used to analyze graphs. The most popular technique is certainly the shortest-path calculation, often performed using the Dijkstra's algorithm, which calculates the shortest distance between two nodes of a graph. A real-life example of this technique is the calculation of driving directions between two locations used by any of several popular mobile applications. But graph analysis is much broader than shortest-path calculations. Four types of graph analysis are widely used:

1. **Path analysis:** This technique is used to determine the distances between nodes in a graph, and includes but is not limited to the shortest-path calculation. An obvious use case is route optimization that is particularly applicable to logistics, supply, and distribution chains, and to traffic optimization for "smart" cities.
2. **Connectivity analysis:** This technique can be applied to determine weaknesses in networks such as a utility power grid. It also enables comparisons of connectivity across networks.
3. **Community analysis:** This technique uses distance and density information to identify groups of people interacting with a social network. Community analysis can, for example, identify whether the interactions are transient, and it can predict if the network will grow.
4. **Centrality analysis:** This technique enables the identification of the nodes or edges that are the most connected to the rest of the graph. Centrality analysis makes it possible to find the most influential people in a social network or to identify the most frequently accessed web pages.

Although graph analysis techniques can be performed on small datasets, they often encounter problems at scale due to the nature of the algorithms used, the characteristics of the graph data, and the limitations of having commodity hardware clusters (i.e., the cloud) performing the analyses. These limitations often constrain graph analysis to approximate solutions rather than exact ones. Examples of software capable of performing graph analysis include Apache Spark GraphX, Titan, Neo4j, and Microsoft Azure Cosmos DB.

Example Application: Graph analysis often is used in fraud detection. In 2016, the International Consortium of Investigative Journalists (ICIJ) exposed highly connected networks of offshore tax structures used by the world's richest elites to circumvent their countries' offshore limitations. To uncover these networks, an ICIJ journalist used more than 11.5 million leaked documents (40 years of data totaling around 2.6 TB) to build a graph representing the connections between individuals and companies such as banks, law firms, and company incorporators found in the documents. The journalist then performed several analyses on the graph to identify the most central companies and individuals, eventually uncovering an entire network of 16,000 tax havens created by 500 banks hiding the money of 140 politicians in more than 50 countries (ICIJ 2016).

3.3 Big Data Applications in Transportation

Within the transportation industry, the concept of Big Data has become increasingly relevant over the past several years, particularly with the advancements in connected vehicle research and the availability of massive datasets. Big Data has been applied to public transportation, trucking/freight, logistics, planning, parking, rail, traffic operations, calibration and validation of traffic simulation models, asset management and maintenance, and even TIM.

This section presents a high-level overview of some of the Big Data approaches, findings, recommendations, and lessons learned, as presented in a wide range of publications most relevant to NCHRP Project 17-75.

3.3.1 Transportation Planning

Much focus is being placed on the application of Big Data in transportation planning. Of particular interest is the use of mobile phone data from telecommunication companies to identify travel patterns. Every time a mobile network subscriber uses the phone to make or receive a call, send or receive a text via SMS or an image via multimedia messaging service (MMS), or access the Internet, a record of that event is generated. These records are collectively termed "call detail records" (CDRs). Each record includes information about party identification, date, time, duration, and cell ID (antenna), which in turn has geolocation and antenna orientation (azimuth) (Lokanathan 2016). Dong et al. (2015) found that using CDR data from mobile communication carriers provides an opportunity to improve the analysis of complex travel patterns and behaviors for travel demand modeling to support transportation planning. Lokanathan et al. used 4 months of passive CDR data of voice calls for several million SIMs from a Sri Lankan mobile operator to explore to what degree the data could be used to create origin-destination (O/D) matrices that represent the flow of travelers between different geographic areas in the city of Colombo, Sri Lanka. The results illustrated that, despite some limitations, mobile network Big Data shows promise as a source of timely and relatively inexpensive insights for transportation planning in developing countries (Lokanathan 2016). Colak et al. (2014) developed a method to use passive CDR data as a low-cost option to improve transportation planning. The resulting trip matrices for Boston, Massachusetts, and Rio de Janeiro, Brazil, were comparable with existing information from local surveys in Boston and with existing OD matrices in Rio de Janeiro (Colak 2014). CDR data is inexpensive compared to active positioning data (e.g., global positioning systems, or GPS), but the data exists at the level of the active cells and is therefore less precise. In addition, not all mobile operators generate continuous active positioning data for all their subscribers, and even fewer operators store the data (Lokanathan 2016).

One benefit Big Data holds for transportation planners is the ability to track movements of vehicles and people on a scale never before imagined. Recent advances in crowd modeling systems have led to more focus on modeling complex locations; however, accurate data collection is one of the biggest limitations that crowd specialists face today (Alvarez 2015). Although researchers are exploring ways to track pedestrians, CDR data is not able to directly inform detailed analysis and understanding of movements at that level. New technology, like the fifth-generation (5G) mobile network, is needed to allow more detailed and more accurate tracking of mobile devices.

3.3.2 Parking

The parking industry has access to more data today than ever before, and the amount of data collected is growing both quickly and exponentially. Incredible amounts of data can be generated from a variety of sources, including space availability tools, meter and parking management systems, credit card and other electronic payment transactions, financial systems, and social media. For parking, the real value of Big Data comes when the data is compiled from all garages, meters, and parking spaces in a region (or the industry) and then that data is merged with data from local events (e.g., sporting events, festivals), holidays, weather patterns, and other drivers of customer activity. The analysis of this large amount of data allows insights to be gleaned into what drives demand peaks on a certain day of the week at a certain garage but not on other days or at other garages within the same vicinity. These insights can help garage operators refine their services and pricing to better meet the actual needs of customers who use their facilities at various times during the week, month, or year (Drow, Lange, and Laufer 2015).

3.3.3 Trucking

Trucking operations generate billions of pieces of information each day, including administrative data (e.g., human resources systems/driver histories), telematics data (e.g., position, speed, time, heading, fast acceleration, over-speed, hard cornering/braking), vehicle data from sensors (e.g., pressure monitoring systems, stability/control systems, refrigerated container monitoring, cargo status sensors), driver performance data, warehouse information, routing information, point of sale in the stores, driver interactions (e.g., enhanced messaging, navigation, re-routing), and fuel cards (e.g., vehicle identification or driver number, odometer reading, purchase number plus the date, time, location and total purchase). Big Data has been used to help fleets identify potential safety risks within their driver pools; provide detailed information on fuel consumption; determine which vehicles or components will need service based on performance metrics rather than a static schedule; provide insights on ways to improve customer service; issue alerts when preset thresholds or key performance indicators are exceeded; and develop scorecards showing multiple key performance indicators to show drivers how they are doing, how divisions are doing, how regions are doing, and so on. During the last 5 years, leading providers have developed a cloud platform that allows them to create and provide tools that simplify and automate activities from real-time operations to long term planning (Beach 2014).

Trucking companies use data to save money on fuel by using predictive modeling to select fuel-efficient trucks. One company depended on this data to help them make the right choice in selecting a new fleet of 50 trucks (a \$6 million decision). A predictive model was used to determine the actual fuel economy of the trucks being considered. The company combined data variables like driving behavior, fuel tank levels, load weight, road conditions, and much more. The details from the data provided executives with a clear picture of which trucks would provide the most fuel savings over time (Nemschoff 2014).

3.3.4 Public Transportation

Many city administrations recognize the value of using Big Data for public transportation, particularly for improving the management of bus fleets and optimizing maintenance and operations. In Sao Paulo, Brazil, Big Data collected in real time provides a more accurate picture of how many people ride the buses, which routes are on time, how drivers respond to changing conditions, and many other factors. The data helps to optimize operations by providing additional vehicles where demand warrants and by identifying which routes are the most efficient. Big Data analytics reduces the time needed to identify problems and make changes and with more accuracy and certainty (Delgado 2017).

Big Data played a big part in re-energizing London's transport network. Transport for London (TfL) collects data through ticketing systems, vehicle sensors, traffic signals, surveys, and social media. The use of prepaid travel cards, swiped to gain access to buses and trains, has enabled a huge amount of precise journey data to be collected. The data is anonymized and used to produce maps showing when and where people travel, giving a far more accurate overall picture and allowing more granular analysis at the level of individual journeys. TfL plans to increase the capacity for real-time analytics and work on integrating an even wider range of data sources to better plan services and inform customers (Marr 2015).

The New York City Transit Authority (NYCTA) developed a Big Data tool to assess the effects of planned service changes and unplanned disruptions and to support the monitoring of fast-changing patterns and trends in ridership behavior. The application combines data from the Metropolitan Transit Authority (MTA) bus automated vehicle location (AVL) system, an automated fare collection (AFC) system, the general transit feed specification (GTFS) schedule, and shapefile streams. (Shapefiles store information about the locations and attributes of geographical features.) The application is responsive to daily detours, special events, and weather-driven ridership. It also allows multiple days of route-level program output to be aggregated for schedule-making purposes, providing a significantly more representative understanding of typical passenger loads than was historically estimated using a few labor-intensive, on-board observations collected over a multi-year period (Zeng et al. 2015).

3.3.5 Transportation Operations and ITSs

The objective of a 2014 white paper by the U.S. DOT's ITS Joint Program Office was to expand the understanding of Big Data for transportation operations, the value it could provide, and the implications for the future direction of the U.S. DOT Connected Vehicle Real-Time Data Capture and Management (DCM) Program (Burt, Cuddy, and Razo 2014). The report summarizes recommendations and next steps from several recent U.S. DOT and other studies regarding how Big Data approaches may be applied in transportation operations. The white paper's recommendations and next steps included the following:

- Engage with a broad range of stakeholders (e.g., public and private, transportation and non-transportation, data analytic product and service providers, modelers, algorithm developers, and decision-support system developers) to disseminate the value proposition for applying Big Data in transportation operations;
- Develop a framework to identify and evaluate options pertaining to the potential roles and responsibilities for state, local, and federal government and the private sector;
- Resolve data ownership issues and the implications for roles;
- Investigate the potential use of a third-party data broker (or multiple brokers), which may help address ownership and funding needs (as the cost of capturing and managing data may be cost-prohibitive for government but profitable for the private sector);

- Develop data standards, especially if transportation agencies are not collecting and managing the data themselves;
- Consider approaches to reduce the volume of connected vehicle and traveler data so that it is more manageable while ensuring that all valuable data is collected;
- Utilize specific technologies and techniques like crowdsourcing, cloud computing, and federated database systems that have come to characterize the state-of-the-practice in Big Data and which will facilitate transportation operators or private sector data service providers in extracting value from connected-vehicle and traveler data;
- Develop connected vehicle Big Data use cases that incorporate Big Data analytics approaches and the operational strategies that could derive from the knowledge gained through those approaches; and
- Further investigate the potential cost and other resource implications of adopting Big Data approaches based on the outcome of the use-case investigation.

Shi and Abdel-Aty (2015) explored the viability of a proactive, real-time traffic monitoring strategy to evaluate operation and safety simultaneously. Data was obtained from a microwave vehicle detection system (MVDS) deployed along a 75-mile section of an Orlando expressway using a network of 275 detectors. Data mining using the random forest technique and Bayesian inference techniques were implemented to unveil, in real time, the effects of traffic dynamics on crash occurrence (Shi and Abdel-Aty 2015).

The Colorado Department of Transportation (Colorado DOT) is looking toward Big Data to solve growing everyday challenges. One challenge area involves winter weather. During snow events, hits on the public website can overload the internal servers, requiring that the CCTV cameras be temporarily turned off to accommodate the load. The Colorado DOT realizes that it cannot add servers to accommodate the relatively few days each year when this happens, and that a scalable Big Data architecture would allow them to expand or lower the system as needed. A second challenge area involves the amount of time that operators at Colorado DOT traffic operations centers spend manually entering data into the system. By moving toward Big Data, these manual activities can be automated using cloud-based systems, enhancing functionality and efficiency.

To support the Big Data approach, the Colorado DOT developed a white paper, “Integrating Big Data into Transportation Services” (Wiener and Braeckel 2016). The purpose of this paper was to provide an overview of current and future Big Data processing challenges at the state DOT as well as to present a set of candidate technologies that could be used to address such challenges. Based on the identified challenges, a list of Big Data needs for the Colorado DOT was developed, which included:

- Improving internal and external data sharing, including effective search and acquisition methods;
- Enhancing domain datasets such as AVL data, work zone data, and incident data by improving coverage, timeliness, and resolution;
- Integrating connected and autonomous vehicle data into Colorado DOT operations;
- Enhancing data analytics through improved capability, ease of application, and timeliness;
- Utilizing scalable and reliable computing and storage; and
- Handling high data volumes.

As a follow-on to the white paper, the Colorado DOT is working to implement the Data Analytics Intelligence System (DAISy), a Big Data platform that integrates a wide range of data sources (e.g., real-time video analytics, CAD, crowdsourced data, ATMS, safety patrol, AVL, vehicle probes, weather, connected vehicles, traffic signals, truck parking, CCTV, tolling, freight,

tunnel operations, chain stations, maintenance, and GIS shape files) in a cloud-based data lake. The project involves three phases:

- Phase 0, project documentation, was underway at time of the research for this project.
- Phase I, planned to begin by late 2018, involves the development and implementation of three use cases (one of which is advance incident detection) to prove the value of the Big Data approach. Phase I also will involve elicitation of stakeholder requirements and the development of a business case and functional/technical requirements for each of the three use cases.
- Phase II, anticipated for 2019, involves a full test of the DAISy system, including the testing of several more use cases. The data integration has already started and will continue throughout the three phases of the project.

The U.S. DOT recently published a report to provide agencies responsible for traffic management with an introduction to Big Data tools and technologies that can be used to aggregate, store, and analyze new forms of traveler-related data (i.e., connected travelers, connected vehicles, and connected infrastructure), and to identify ways these tools and technologies can be integrated into traffic management systems and TMCs (Gettman et al. 2017). Key contributions of the report include the following:

- Identification of how sharing data with other TMCs, systems, connected vehicles, travelers, and agency business processes or systems could affect the performance of a traffic management system or TMC;
- Identification of challenges and options for compiling, using, and sharing this emerging data;
- Presentation of potential use cases for integrating Big Data technology and tools into traffic management systems or TMCs;
- Identification of a national system architecture that illustrates the types of tools and interfaces that will be needed;
- Examples of the data processing and storage requirements for a typical agency when connected vehicle, traveler, and infrastructure data is being transferred to the TMC at significant levels; and
- Key questions to be addressed in developing a plan for leveraging the emerging data sources with Big Data tools and technologies.

The Big Data Europe (BDE) project seeks to develop an adaptable, easily deployable and usable solution that will allow interested user groups to extend their Big Data capabilities or to introduce Big Data technologies to their business processes. The project involves building a Big Data community and developing a Big Data Aggregator infrastructure that meets the requirements of users from the key societal sectors, minimizes the disruption to current workflows, and maximizes the opportunities to take advantage of the latest European research and technological developments, including multilingual data harvesting, data analytics, and data visualization (BDE n.d.).

Within the framework of the BDE project, ERTICO-ITS Europe organized a 2015 workshop on “Big Data for Smart, Green and Integrated Transport” (BDE and ERTICO-ITS Europe 2015). The workshop focused on the elicitation of requirements for Big Data management within the intelligent transportation domain. The workshop consisted of three sessions that were dedicated to data-centric initiatives in transportation, Big Data use cases in transportation, and technologies and tools used and envisioned. The workshop results indicated a clear need for Big Data solutions in transportation, and that the areas for Big Data application are diverse. Particularly significant and relevant outcomes/recommendations from this workshop included the following:

- Big Data in transport will lead to improved multi-source traffic and travel data availability and processing as well as to tools that improve multi-source traffic and travel data fusion. Combining big, open, and linked data will foster innovation and economic benefits.
- A future Big Data platform must allow real-time data analysis, which includes visualization tools that allow data mining and the visualization of analysis results, as well as automated coding and delivery of video data over cellular/Wi-Fi to cloud-based storage. The platform should use data structures that allow efficient data extraction and support open repositories with high-quality context information.
- If a common standard will be developed, it should be a non-discriminatory standard with open application programming interfaces (APIs).
- Policy-makers should provide clarity on the re-use rights of data.
- There should be a “free flow of data initiative” in the EU [European Union], and the EU should promote the use of open data.
- Current businesses are challenged by Big Data. An issue is to have the right mindset to make data available in the first place.
- Making data available also involves a risk factor. Public education/outreach is required to make people aware that data needs to be shared and that specific data is available and accessible.
- Transportation stakeholders need to contribute to the creation of large pools of well-documented and accessible road data (i.e., open and with known velocity, volume, and variety).

3.3.6 Emergency and Incident Management

The Rio de Janeiro Operations Center (ROC) is the first application of a citywide system to integrate all stages of crisis management from prediction, mitigation and preparation to immediate response. In traditional applications of top-down sensor networks, data from each department operates in isolation. The ROC’s approach to information exchange, on the other hand, is based on the understanding that overall communication channels are essential to getting the right data to the right place, which can make all the difference in an effective response. The ROC gathers data in real time through fixed sensors, video cameras, and GPS devices from 30 government departments and public agencies (including water, electricity, gas, trash collection and sanitation, weather, and traffic monitoring) in real time. Data fusion software collates the data using algorithms to identify patterns and trends, including where incidents are most likely to occur (International Transport Forum 2015).

The Waze Connected Citizens Program (CCP) brings cities and citizens together to identify what’s happening and where. The CCP promotes more efficient traffic monitoring by sharing crowdsourced incident reports from Waze users. Established as a two-way data share, Waze receives partner input such as feeds from road sensors, adds publicly available incident and road closure reports from the Waze traffic platform, and returns succinct, thorough overviews of current road conditions (Connected Citizens Program 2016).

Genesis PULSE is an example application that makes use of Waze crowdsourced Big Data to improve incident response. Genesis PULSE is a decision-support and situational awareness software solution that enhances existing CAD systems. As Waze users report traffic events, emergency call centers that are also Genesis PULSE customers can immediately see and pinpoint the incident in real time and use this information to effectively dispatch units. The results are increased situational awareness for dispatch personnel and administrative staff and decreased response times to incidents (GenCore Candeo, Ltd. 2017).

Continuous streams of video, traffic volume, speed, backups, weather, and more come into the Iowa State REACTOR lab from across the state every 20 seconds to 1 minute. Using the

data, researchers are developing the TIMELI system (Traffic Incident Management Enabled by Large-data Innovations), which will make use of emerging large-scale data analytics to reduce the number of incidents and improve incident detection. New traffic models, computer algorithms, computer display interfaces, and information visualizations will help operators make decisions and take actions (Iowa State University 2017).

Researchers at the University of California, Davis (UC Davis) Advanced Highway Maintenance and Construction Technology Research Center (AHMCT) are developing the third generation of the Responder system, which allows first responders to collect and share at-scene information quickly and efficiently. Unique features of Responder allow users to capture, annotate, and transmit images. Using GPS readings, the system automatically downloads local weather data, retrieves maps and aerial photos, and pinpoints the responder's location on the maps. Data includes CAL FIRE, InciWeb, CCTV camera images, Caltrans Chain Control, California Highway Patrol (CHP), daily and hourly forecasts, road information, Roadway Information System (RWIS), stream flow, changeable message signs (CMS), zone alerts, and zone forecasts (Clark et al. 2016).

Two caveats should be noted for these emergency and incident management examples:

1. Although the TIMELI system is a start to the application of Big Data in TIM, the amount of data generated in Iowa may not be sufficient to train algorithms that can be applied in other locations. Big Data tools are data hungry. As an example, the reason why Facebook's image recognition process is effective is not just because deep learning has been applied to the data; it is because Facebook was able to train the image recognition algorithms using hundreds of millions of images.
2. On its own, the Responder system in California likely is not a Big Data system, but rather a start that could be augmented to become a Big Data system.

Ways exist to expand on these initial approaches to Big Data for TIM, but first the data needs to be prepared, and enough historical data needs to be assembled to be able to find meaningful patterns.

CHAPTER 4

Big Data and TIM

As presented in Chapter 2, the state of the practice of TIM has advanced over the past decade through multiple approaches, including the development and implementation of the National TIM Responder Training Program, legislation, quick-clearance policies, TIM committees, and multi-agency operating agreements. The resulting improvements in responder safety and effectiveness, combined with the use of TIM-related data, have positioned TIM to make another step forward. Ongoing efforts, such as the FHWA's EDC ("Every Day Counts") TIM data innovation, are accelerating and advancing the implementation of TIM data collection and use among regional and state entities nationwide. Nonetheless, the current state of the practice in using data for TIM is limited. Moreover, these practices draw on traditional approaches to data collection and use, relying on engineering and decision-maker judgment, augmented by using quantitative analysis of limited data samples and often using subjective, manual, and resource-intensive strategies.

With the increased quantity and improved quality of TIM data, there is promise that the application of the Big Data technologies and analytics described in Chapter 3 can further advance the state of the practice in strategic, tactical, and support TIM activities. The ability to merge multiple, diverse, and comprehensive datasets and then to mine the data has the potential to improve TIM programs. The use of Big Data might afford opportunities to:

- Develop, evaluate, and refine TIM policies;
- Improve scene management practices;
- Improve resource utilization and management;
- Gain efficiencies with respect to the TIM timeline;
- Improve responder and public safety;
- Access and query data in real time to augment incident response actions;
- Enable predictive TIM;
- Support performance measurement and management; and
- Support TIM justification and funding.

At this point, three questions arise: How might Big Data be applied to TIM? What potential opportunities exist to leverage Big Data to improve TIM? What are the potential benefits of doing so? This chapter explores Big Data opportunities for TIM by presenting specific examples that stem from applications that represent the current state of the practice in TIM data collection and analysis. For each example, a summary of the traditional data collection and analysis approach is given. Then, a potential Big Data approach/opportunity to address the same problem or research question is presented, along with the benefits of the Big Data approach. The purpose of this discussion is to contrast the traditional approach with the Big Data approach, identify the differing data needs and analytical approaches, and discuss the possibilities and benefits afforded by Big Data.

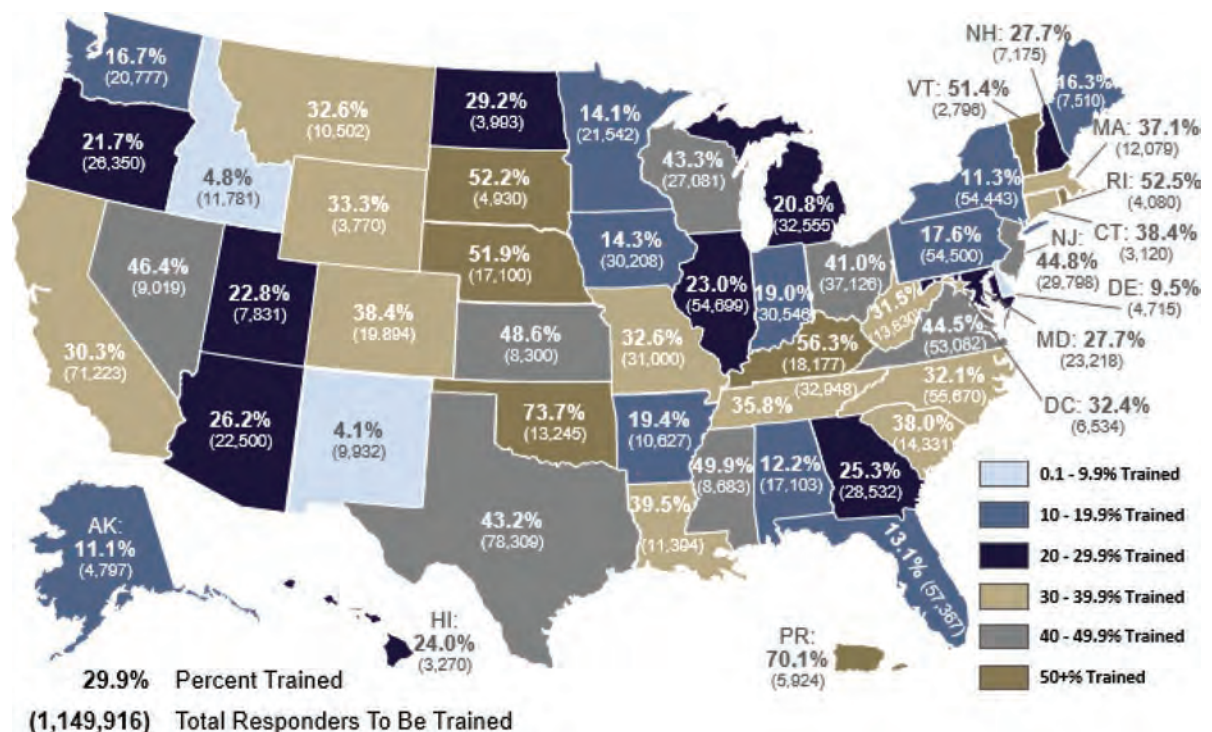
4.1 Improve On-Scene Management Practices

Big Data provides opportunities to examine existing TIM strategies and practices, and to consider how factors such as training, responder experience, and response discipline affect response efficiency (e.g., data could point to superior procedures among responder disciplines). Every traffic incident is distinct, being based on a unique combination of factors that include incident type and severity, location, the combination of individual responders on scene, weather conditions, and special events. A wide array of naturally occurring scenarios (i.e., variety), made available from multiple sources across the country, would enable a more robust exploration of the impacts of differing on-scene management strategies.

The National TIM Responder Training Program is the standard by which responders act. The 33 learning objectives of this program offer potential opportunities for improvement wherever data can inform or reinforce one or more of the learning objectives. For example, on-scene, real-time adjustments to responder actions (e.g., adjusting vehicle positions, scene lighting, temporary traffic control devices, and end-of-queue signage) could benefit traffic, safety, and travel time reliability. Protocols for various types of special circumstances like vehicle fires, HAZMAT, or hybrid and electric vehicles could aid responder safety. Development of prediction-assisted protocols or procedures, such as estimating the length of a queue during an incident, could enable responders to adjust traveler information and offer localized advance warning to prevent secondary crashes.

Example: Assessment of the National TIM Responder Training Program

In summer 2012, FHWA rolled out the National TIM Responder Training Program. As of July 2018, more than 344,000 responders had participated in the training nationwide (Figure 4-1).



FHWA wanted to assess how effective the training had been in reducing roadway and incident clearance times and secondary crashes.

Question

How effective has the National TIM Responder Training Program been in reducing roadway and incident clearance times and secondary crashes?

Traditional Approach

The U.S. DOT conducted a study to assess the effectiveness of the National TIM Responder Training Program (Einstein and Luna 2018). The evaluation focused on the effectiveness of the TIM training in three areas: (1) disseminating TIM concepts to a wide incident responder community, (2) changing/enhancing agency practices, and (3) improving TIM performance. The first two areas were evaluated using quantitative and qualitative measures of effectiveness such as the number of attendees and the number and proportion of disciplines at trainings, attendees' self-assessments of the value of training (through a post-course assessment), and changes in responder and agency practices with respect to on-scene traffic-incident practices and management (through interviews with responders). The third area (improving performance) was assessed using quantitative TIM performance measures calculated and crash report data collected from two areas: greater Phoenix, Arizona, and eastern Tennessee (Tennessee DOT Region 1). The analysis included 22,000 crashes from Phoenix and 6,400 crashes from eastern Tennessee, a relatively small number of crashes for a 4-year period (2012–2015). Aggregate performance measures (i.e., annual average clearance times) were used to show a decreasing trend in clearance times because disaggregate measures (i.e., clearance times by crash severity or number of vehicles involved in a crash) could not detect clear trends associated with the TIM training. The evaluation team noted concerns about missing data, erroneous data, and an inability to link TIM-trained responders to specific incidents.

Big Data Approach/Opportunity

A Big Data approach to assess the effectiveness of the National TIM Responder Training Program would collect and analyze data from the entire country. Data would be analyzed at the incident level (as opposed to the aggregate level) to identify:

- Quantitative trends between training and shifts in incident clearance duration, secondary crash frequency, and responder struck-by events;
- Areas in which significant improvements were absent;
- Successes; and
- What training or external factors contributed to the successes.

Additionally, the analysis of historic and current data could explore whether characteristics of the training and/or the percentage of responders trained affected responder on-scene behaviors that improved or reduced incident clearance and scene safety. Data of interest for the analysis would include crash data, CAD data, weather conditions, TIM programs and policies in place, and responders training data—responders trained and not trained, and associated information such as discipline, jurisdiction, age, years on the job, and so forth—training dates, training locations, training types, and trainers and associated information.

Time would be an important aspect of this analysis. Daily, responders receive training and new incidents occur across the United States. It takes time for the knowledge gained through training to translate to measurable benefits in the field (e.g., reductions in clearance times).

Big Data moves from analyses that are based on single snapshots in time to the ability to provide a continuous feedback loop as changes occur and new data is generated.

Therefore, rather than conduct a single geographically and temporally bound study that represents a single snapshot in time, the Big Data approach would be to perform the analysis regularly (e.g., weekly or monthly) to account for the arrival of new data and to identify trends and highlight outliers for further inspection. This is an important distinction between the data-weak traditional approach and the data-hungry Big Data approach. Rather than making decisions based on a few performance measures that are calculated once a year, analyses can be conducted continuously to monitor responder practices and adjust accordingly (e.g., through funding for training, training content, and training locations). Should one or more of the trainers, one or more of the learning objectives, or one or more of the training types be determined to be ineffective—either completely or under certain circumstances—the traditional approach will likely be too high level to detect these shortcomings, or it may take years to uncover the issues. For example, the U.S. DOT evaluation identified big differences in the numbers of responders trained across disciplines in Tennessee and Arizona (e.g., more fire and towing in Tennessee than in Arizona) and noted that average class size and mix of disciplines also may impact training effectiveness, but these differences could not be evaluated with the data at hand. With Big Data analytics, these differences, as well as negative trends and outliers, could be quickly detected and analyzed to identify and remedy training weaknesses or to expand on training strengths.

A retail analogy is provided by Walmart. A grocery team could not understand why sales had suddenly declined in a particular product category. Walmart's data scientists drilled into the data and quickly determined that pricing miscalculations had been made, leading to the products being listed at a higher price than they should have been in some regions. Big Data enables much faster and more accurate pinpointing and verification of problems caused by human error or miscalculation at the planning or execution stage of a particular business activity. If an organization cannot get insights until it has analyzed data for a month, a quarter, or even a year, it has lost sales, productivity, and efficiency within that time (Marr 2017).

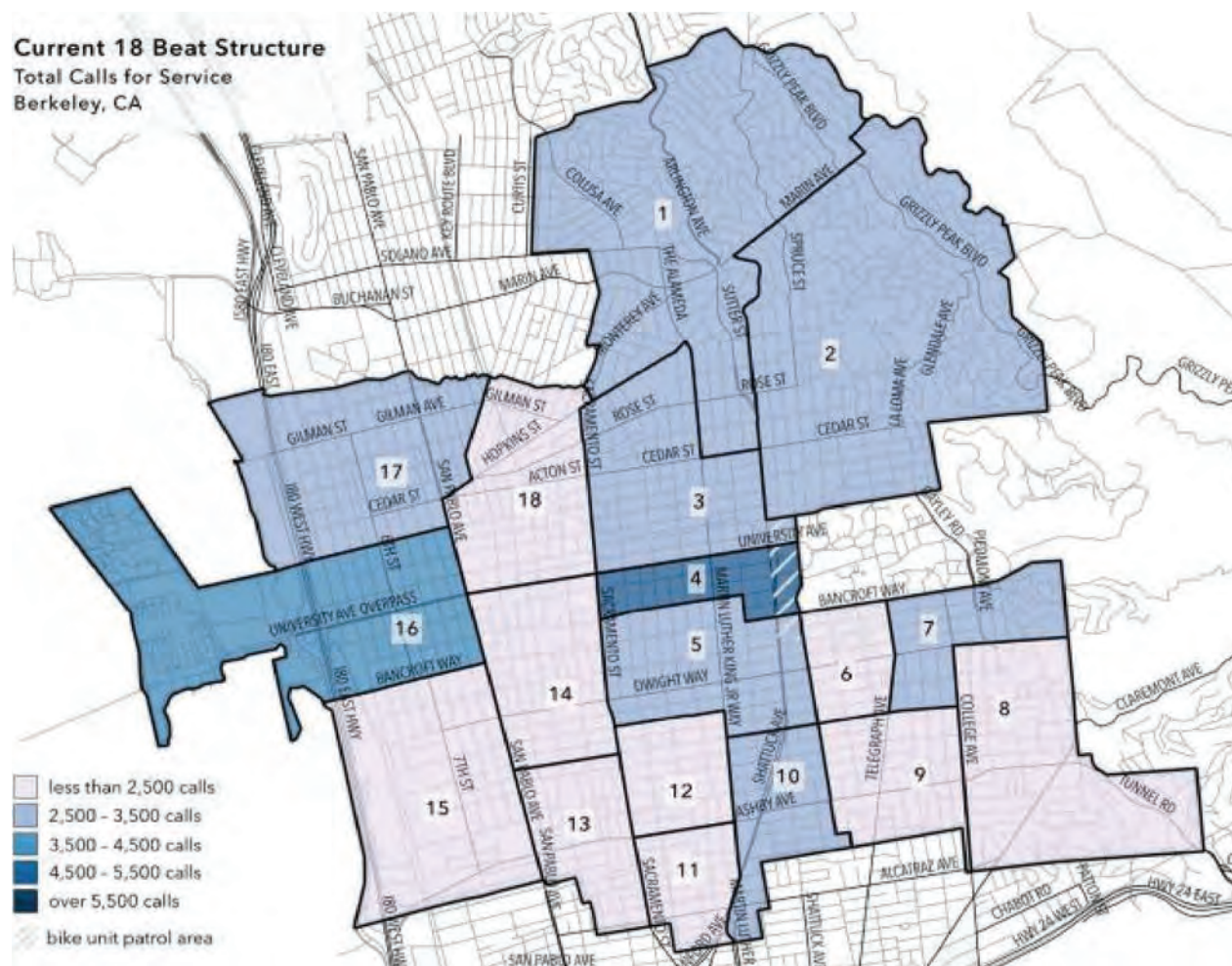
Through a low-level of granularity of data and low-cost, frequent analyses, Big Data can expose a more detailed and evolving picture of the incident response and training reality, helping to change policy and procedures as well as the mindset of “set and forget.” Continuous feedback, with data feeding the decision-making process, is necessary to remain efficient and effective.

4.2 Improve Resource Utilization and Management

Historical analysis of incident response could support huge advances in the deployment of TIM resources, including both personnel and equipment. Big Data analytics could help to optimize SSP routes by identifying the best days and hours of service based on weather, seasonality, and other factors to ensure that the appropriate number and type of resources are scheduled for a geographic expanse. Desired response time, circulation time, vehicle cost per mile, weather, special events, and estimated obligated/unobligated patrol time are factored in to determine staffing needs. Big Data analytics could assist in staging human and equipment resources for quicker and less costly responses.

Example: City of Berkeley, California, Police Patrol Beat Evaluation Study

The city of Berkeley, California, conducted a study to assess the existing beat structure and allocation of patrol staffing and to evaluate opportunities to improve the deployment of resources. Depicted in Figure 4-2, the city's existing system of 18 beats was based on



Source: City of Berkeley, California (Matrix 2014)

Figure 4-2. Existing beat structure and total calls for service.

20-year old crime trends, calls for service, and staffing levels, and needed to be updated to reflect existing conditions (Matrix 2014).

Question

What beat structure and allocation of patrol staffing will best improve the deployment of resources?

Traditional Approach

The approach to updating and improving the patrol beat structure was based on a qualitative and subjective assessment of data, as well as a quantitative assessment of four possible beat structures (16, 14, 11, and 4 beats) on calls for service, major crime, workload, geographical accountability, neighborhood integrity, and efficient travel. The methodology included interviews with the police chief and patrol staff; the collection of data to document workloads, costs, service levels, and operating practices; and town-hall style meetings. Analyses included a statistical analysis of call-for-service workloads and major crime throughout the city; a GIS assessment of the equity of various beat boundaries on call workload and crime; and the analysis of interview, survey, and town hall data (Matrix 2014).

Big Data Approach/Opportunity

The traditional approach applied to update the patrol beat structure was resource intensive (town hall meetings, interviews, surveys) and resulted in a limited (manageable) number of distinct options that were then assessed on seven criteria. The quantitative analysis and optimization resulted in one recommended patrol beat structure. The Big Data approach to this problem would offer advancements in multiple ways. The Big Data approach would leverage more advanced optimization methods, such as genetic and evolutionary algorithms, would add a wider set of data sources (e.g., weather, census, social media) to those used in the traditional approach, and would be applied at a more granular level (e.g., time of day, day of week, week of year, local events), going beyond the criteria that were capable of being optimized with the computing limitations of the traditional approach.

Big Data analytics could also offer an approach that automatically optimizes the number of beats and boundaries across the criteria. Furthermore, capitalizing on the computing power cost efficiencies available through Big Data analytics could allow for the simultaneous analysis of thousands of patrol beat structures as opposed to only four. Finally, Big Data could address issues of flexibility to better address future changes, and questions like the following:

- What comes after this study?
- How long will this patrol beat structure remain the “most” efficient?
- When will the resources be available to repeat the traditional study to represent the changing times?

Traditionally, working with little data at a high resolution, changes were hard to detect and the need to re-optimize was difficult to justify. The increases in data volume and data resolution have brought more light to the constantly changing and evolving world, and in many industries have exposed inefficiencies and gaps that can be corrected. The efficiencies of Big Data allow the analysis to be set up once and repeated over and over as new data becomes available. Big Data allows system changes to be identified quickly, enabling adjustments to be made far more frequently to maintain efficient beat structuring. In fact, Big Data analytics has the potential to take the patrol beat deployment decisions to the next level, moving from static beats to real-time dynamic beats that concentrate patrols in the areas with the greatest likelihood for need each day by factoring in variables such as weather, special events, and the mood of the population as expressed on social media.

4.3 Improve Safety

Big Data has the potential to dramatically advance safety through a better understanding of the characteristics of traffic incidents, and through improved responder situational awareness. When the traits that are most dangerous for responders and passing motorists are known, adjustments can be made to equipment and on-scene behavior to mitigate those dangers. Early warning systems (e.g., via an audible alert or a color-coded message on a mobile device or responder vehicle computer) might be developed to make on-scene personnel aware of varying degrees of danger associated with different combinations of incident conditions. Such analytics also could guide traveler information systems and safety messages provided via 511 or VMS systems, or other means to offer driver-customized, in-vehicle alert warnings of responder activity. Understanding emergency vehicle lighting and conspicuity through analytics has the potential to improve safety, particularly when better understanding is gained of how approaching motorists behave given those stimuli. Big Data also can be applied to identify the most effective frequencies, geographic areas, and content for responder training.

Example: Florida DOT “Move Over” Study

To mitigate the risk to responders at incident scenes, every state has implemented a law that requires drivers to move over or slow down when approaching a patrol vehicle that has stopped at the roadside. The Florida DOT conducted a study to determine the effectiveness of the “Move Over” law in Florida.

Question

How effective has the Move Over law been in mitigating risk to incident responders in Florida?

Traditional Approach

To determine the effectiveness of Florida’s Move Over law, the Florida DOT and the Florida Highway Patrol (FHP) supported a field study that involved the observation of right-lane vehicles passing staged police stops on three Florida freeways in differing parts of the state. Each staged stop involved the use of a civilian research vehicle, a marked police vehicle, video recording of passing traffic, and measurement of passing vehicle speeds using a laser speed measurement device (see Figure 4-3).

Differing patrol vehicle emergency lighting configurations—blue and red versus amber only—were tested (Carrick and Washburn 2012). This traditional field study approach provided results that were helpful in understanding how a convenience sample of 9,000 drivers reacted to a limited combination of emergency lighting configurations at a limited number of locations across the state of Florida. Notable concerns with this study are the secondary crash risk to researchers and law enforcement from remaining adjacent to high-speed traffic and the potential throughput loss along roadway facilities.

Big Data Approach/Opportunity

A Big Data approach to assess the effectiveness of the Move Over laws nationwide might involve a wide variety of naturally and constantly occurring data sources and the application of Big Data analytics to extract driver behaviors. Data of most interest to this study would include



Source: Grady Carrick (Carrick and Washburn 2012)

Figure 4-3. One of three staged stop sites.

police enforcement activities; vehicle telematics data (from passenger vehicles, commercial vehicles, and police fleet vehicles) including time of day, location, speed, lateral position, specification of response vehicles, active emergency lighting configurations at stops, and so forth; roadway inventory data like roadway classification, number of lanes, and horizontal and vertical curvature; and weather data. Using the data, speeds and compliance rates could be assessed for thousands of naturally occurring combinations of emergency lighting configurations, roadway types, vehicle types, locations, times of day, weather, and other factors (e.g., recent media campaigns) that influence compliance.

Big Data analytics might include:

Big Data moves from limited data samples meant to represent reality to leveraging the wide variety of data occurring naturally in the real world.

- A non-zero variance analysis to determine what factors impact behavior and compliance;
- A clustering analysis to group co-occurring factors into multiple scenarios/groups (e.g., compliance rates are high during non-peak periods on limited access highways with more than two lanes in each direction); and
- A classification of the uncovered groups/clusters.

The results would not only provide a detailed understanding of when, where, and under what conditions drivers comply or do not comply with the Move Over law, they could also help inform outreach, public education, and policy to further improve compliance. Instead of designing an experiment meant to represent reality based on a sample of data collected from a few locations and then extrapolating the results to other locations, the Big Data approach looks at actual behaviors occurring naturally across a wide area by leveraging the large volume of highly varied data available in the real world. Further, the Big Data approach also could be rerun as new data becomes available, making it easier to identify adjustments or corrections to policies, vehicle markings, emergency lighting systems, and other factors as needed.

4.4 Enable Predictive TIM

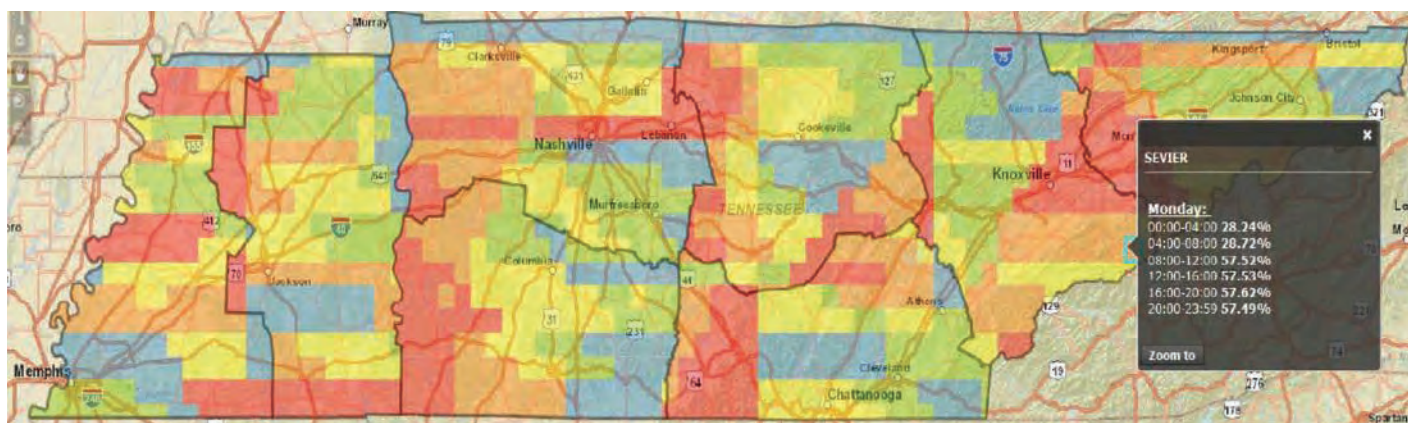
Big Data could be used to predict when, where, and under what conditions traffic incidents are most likely to occur so that the appropriate response can be pre-staged and/or more quickly deployed if necessary. Identifying the nature and causes of traffic crashes is fundamental to traffic safety analysis, and it precedes the implementation of countermeasures embodied in the “3Es”: engineering, education, and enforcement. For example, every TMC operator knows that when it rains there will be a spike in crash activity. Improved data integration and analytics has the potential to move the TMC observation beyond intuition and into the realm of predicting when and where problems are most likely to occur under specific and detailed conditions such as planned special events, periods of holiday travel, or even the daily rush hour. Using various types of data, and in particular detailed weather data, to uncover correlations and predict when and where to put resources is foundational to improving TIM planning and operations.

Example: Tennessee Highway Safety Office Predictive Analytics

Agencies face a continual challenge in allocating resources in the most cost-efficient and effective way possible. Tennessee’s Crash Reduction Analyzing Statistical History (C.R.A.S.H.) program uses software and data to perform analyses that inform the agency’s decisions.

Question

How can the state more efficiently allocate limited resources, deploying troopers to locations and at times with the greatest likelihood of crashes?



Source: Tennessee Department of Safety and Homeland Security (Freeze 2017)

Figure 4-4. THP C.R.A.S.H. software program—model results.

Traditional Approach

The C.R.A.S.H. program developed by the Tennessee Highway Patrol (THP) leverages data from every crash report filed in the state, from traffic citations, and includes data about weather and special events to analyze and predict when and where serious or fatal traffic crashes are most likely to occur. C.R.A.S.H. breaks Tennessee into 5-mile-by-6-mile sections and predicts traffic risks for each section in 4-hour increments every day. THP uses these analytics to more efficiently allocate limited resources by deploying troopers to locations and at times with the greatest likelihood of crashes. The models also help field supervisors design shift assignments, develop enforcement plans, and determine when and where to conduct grant-funded activities (Tennessee Department of Safety and Homeland Security 2017). The model results, an example of which is shown in Figure 4-4, have proven to be accurate about 70 percent of the time (Martinelli 2017).

Big Data Approach/Opportunity

The Big Data approach would be to move from predictive modeling of crashes using historical data to predicting crashes in real time for the purposes of reacting immediately to changes in the factors that are likely to lead to a crash. Instead of running the prediction models every day, the models might be run in parallel and continuously, using Big Data analytics in the cloud. Big Data predictive models would rely not only on historical data but also on real-time streaming data (such as speeds, volumes, occupancies, weather data, vehicle data, road weather conditions, data from social media, and events), which would be fed to the models in real time as it is generated and received to predict when and where there is a high probability for crashes. The outputs of such models could potentially feed real-time decision-support systems for active traffic management and dynamic resource allocation. One specific approach to this analysis would be the use of *deep learning* (a machine learning method), which allows complex relationships in large datasets to be captured efficiently. The more granular the data, the faster it changes, and a consequence of these fast changes is that the accuracy of deep learning models will start to drop as the existing relationships between the data begin to shift. The Big Data approach remedies this drawback by treating prediction models as short-lived and disposable. Big Data prediction approaches typically monitor the accuracy and performance of their current models in real time and develop new models as new data is added. Should a model stray from the existing level of performance and become less accurate, it can be discarded immediately and

The concepts of disposability and replaceability are inherent to Big Data infrastructure from hardware to software and models.

replaced by a newly developed model. The C.R.A.S.H. example illustrates the concepts of *disposability* and *replaceability*, which are inherent to Big Data infrastructure from hardware to software and models.

4.5 Support Performance Measurement and Management

Performance measurement and management are the ongoing processes undertaken in support of accomplishing the strategic objectives of a program. Performance measurement involves selecting quantitative performance measures to be tracked, setting performance targets, collecting and analyzing data in support of the performance measures, and ongoing monitoring and reporting of program accomplishments and areas that need improvement. Performance management goes further in that it involves active and continuous follow-up by program staff and managers to identify and implement specific strategies and tactics to improve efficiency and then to measure and report the outcomes of these strategies and tactics (i.e., did the strategy help to meet the performance targets?).

Performance measurement and management are data-driven processes. Without access to the appropriate data and analytics tools, performance measurement and management can be challenging, laborious, or downright impossible. The application of Big Data could support TIM agencies with their current performance measurement and management processes, and it could also expand the thinking and overall approach to the processes (e.g., through identification of additional, critical performance measures; identification of performance gaps or pitfalls; and identification of the actions necessary to improve performance).

Example: Oregon DOT Performance Management

In 2014, Oregon DOT management inquired why the mutual Oregon DOT/Oregon State Police (OSP) RCT goal of 90 minutes was exceeded in 1,088 incidents. The experience of the Oregon DOT elucidates the need for more data and more advanced analytics for TIM performance management.

Question

What factors contributed to 1,088 incidents exceeding the mutual Oregon DOT/OSP RCT goal of 90 minutes?

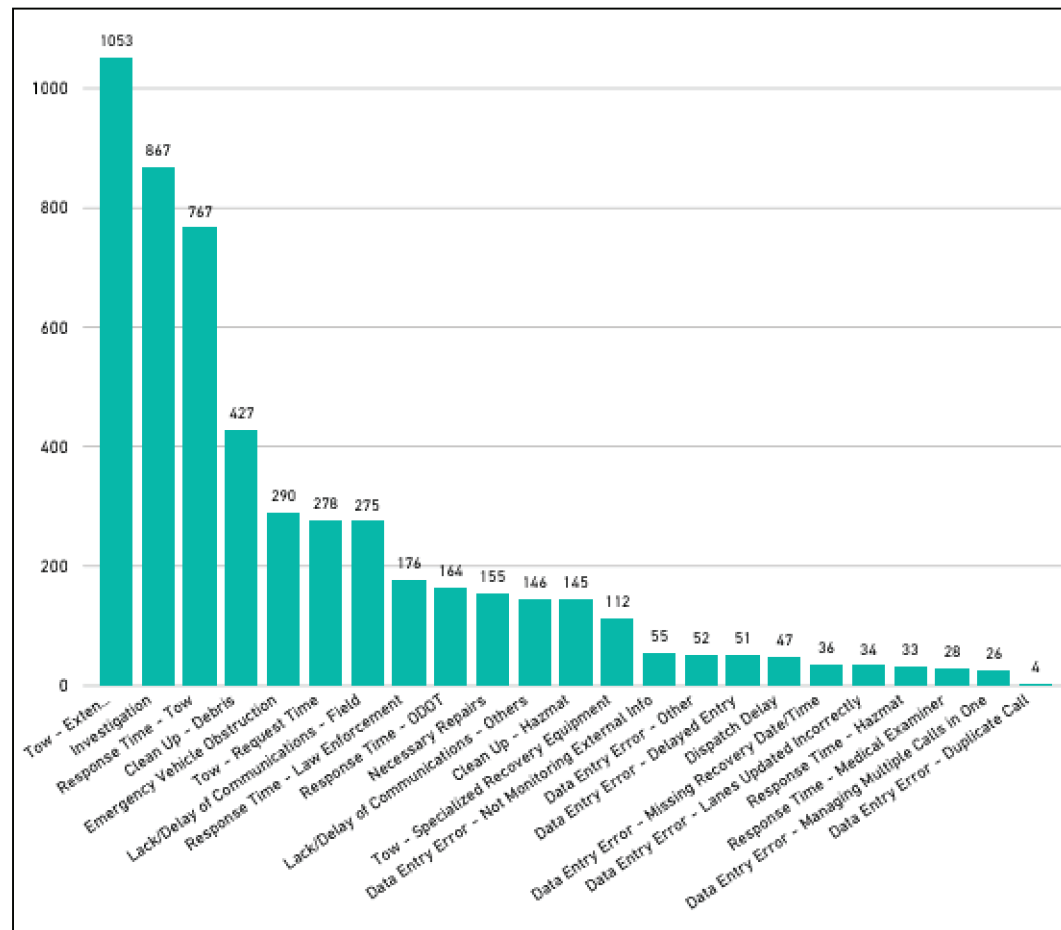
Traditional Approach

Using the data available at hand, the Oregon DOT determined an answer to this question by examining the problematic incidents “one at a time.” The approach to the analysis was to engage response partners to anecdotally create a list of factors known to generally contribute to longer clearance times, review each of the 1,088 incident reports, and categorize the incidents in relation to the list of factors. Following the analysis, specific actions were developed and implemented to address the most common causes of extended clearance times (Oregon DOT 2018). The Oregon DOT has since developed a process to communicate the causal factors for long clearance times directly from the field to the dispatch centers so that the data can be immediately entered into their system to drive an ongoing report (Figure 4-5).

Although this approach provides an excellent example of active performance management by the Oregon DOT, the analysis is based on a list of reasons for extended delays that was created based on subjective assessment (the anecdotal factors initially suggested by the response partners) rather than on tangible data.

Year	QTR	District	RTE
All	All	All	All

Assigned Cause	Count	% of Year
Tow - Extended Recovery	1053	20.2%
Investigation	867	16.6%
Response Time - Tow	767	14.7%
Clean Up - Debris	427	8.2%
Emergency Vehicle Obstruction	290	5.6%
Tow - Request Time	278	5.3%
Lack/Delay of Communications - Field	275	5.3%
Response Time - Law Enforcement	176	3.4%
Response Time - ODOT	164	3.1%
Necessary Repairs	155	3.0%
Lack/Delay of Communications - Others	146	2.8%
Clean Up - Hazmat	145	2.8%
Tow - Specialized Recovery Equipment	112	2.1%
Data Entry Error - Not Monitoring External Info	55	1.1%
Data Entry Error - Other	52	1.0%
Data Entry Error - Delayed Entry	51	1.0%
Dispatch Delay	47	0.9%
Data Entry Error - Missing Recovery Date/Time	36	0.7%
Data Entry Error - Lanes Updated Incorrectly	34	0.7%
Response Time - Hazmat	33	0.6%
Response Time - Medical Examiner	28	0.5%
Data Entry Error - Managing Multiple Calls in One	26	0.5%
Data Entry Error - Duplicate Call	4	0.1%
Total	5221	100.0%



Source: Oregon DOT (2018); used by permission

Figure 4-5. Incident clearance times exceeding 90 minutes.

Big Data Approach/Opportunity

A Big Data approach to this question would be to leverage a variety of data sources to automatically identify the factors (and combinations of factors) that lead to extended clearance times. At a minimum, a statewide analysis would be done; however, more insights could be drawn from multi-state or national data. Many of the conditions that lead to extended clearance times in Oregon are the same conditions that lead to extended clearance times in other states (as is suggested by the anecdotal factors developed by the Oregon DOT and partners shown in Figure 4-5). Relevant data for the analysis would include crash data, CAD data (timestamps of every notification, arrival, and departure from the incident scene), injury surveillance data, roadway data, weather data, and social media data. The analysis would be conducted at the incident level, which means that the clearance time and details of every incident would be compared against all others.

Big Data helps to reduce or eliminate the subjectivity, judgment, and bias often found in manual, qualitative, and human-driven analysis processes.

A graph analysis could be conducted, yielding results like those represented in Figure 4-6. To conduct such an analysis, data relevant to each incident and its response would be plotted to create a representative graph. The structure of the graphs would be based on a semantic graph ontology (a commonly shared vision of a domain). Each graph

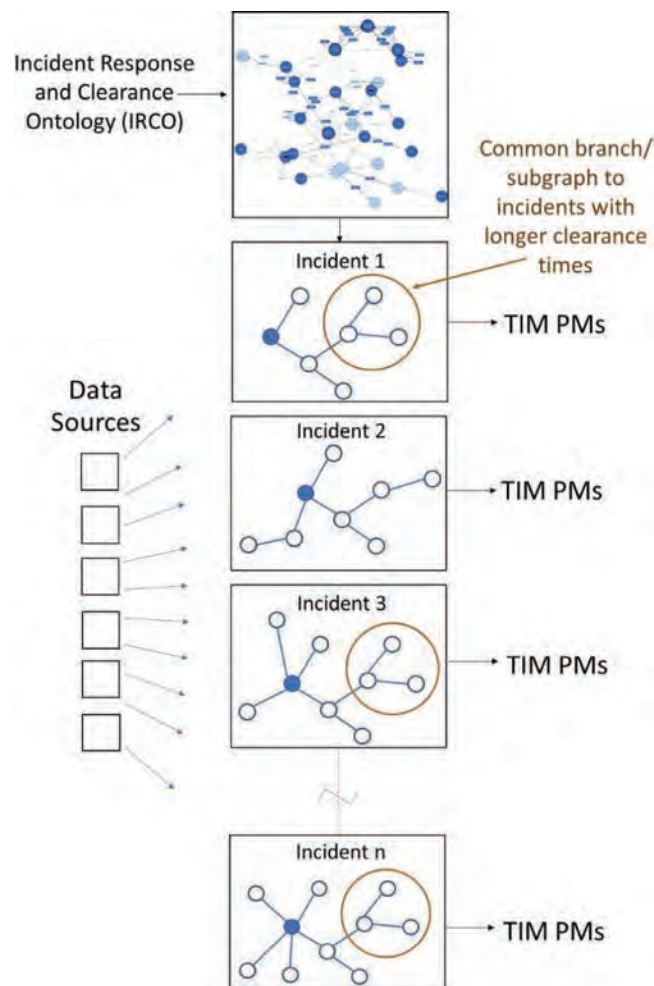


Figure 4-6. Representation of graph analytics for TIM performance management.

would provide a complete description of the incident/response (e.g., type, location, vehicles, injuries, responders, actions taken, timestamps). The incident graphs would be loaded into a graph database for analysis. Then, using graph analytics (e.g., graph similarity and subgraph matching algorithms), a portion of the incident graphs that are common to the incidents with long clearance times would be identified and extracted. The extracted subgraphs (consisting of branches, nodes, and values) would then be reviewed and classified to identify the various patterns (e.g., late tow truck arrival), triggers (e.g., heavy congestion), and thresholds (e.g., less than two responders on the scene) that are common to incident responses with extended clearance times. This approach would likely offer additional, even unexpected, insights into the causes for extended clearance times.

An initial incident response and clearance ontology (IRCO) was developed as part of NCHRP Project 17-75 and is presented in Appendix B to this report. The IRCO could be used to structure the graphs with the data available for this type of analysis.

4.6 Support TIM Justification and Funding

Response agencies, particularly public agencies, often are mission driven and task oriented at the expense of meticulous documentation of activities that serve to justify continued funding. Another opportunity for the application of Big Data for TIM is to build a collection of information that documents activities, program costs, and program outcomes to make an accurate and compelling business case for TIM. One of the most fundamental ways to improve the effectiveness of TIM is to ensure a dedicated and right-size funding stream. Historically, agencies have relied on TIM conventions, or “rules of thumb” (e.g., that 20 percent of incidents are secondary in nature, that each minute of blockage requires 4 minutes to recover) because this was the best or the only data available. Having data that helps make the business case for TIM increases the potential for securing TIM funding. Being able to demonstrate quantitatively the impacts of incidents on safety (e.g., secondary crashes), mobility (e.g., number of people stuck in incident-related congestion), the environment (e.g., air quality and fuel waste), and the economy (e.g., freight movement) will help to promote continued or increased funding for TIM programs.

Example: FHWA TIM Benefit-Cost (TIM-BC) Tool

The FHWA has developed a web-based TIM Benefit-Cost (TIM-BC) tool that assists TIM programs in determining the benefit-cost ratio for certain TIM activities (FHWA 2017b). The tool evaluates SSP, TIM laws, towing arrangements, training, dispatch colocation, and the establishment of TIM taskforces to quantify their benefits. The TIM-BC tool relies significantly on user inputs and/or default values for factors like average incident duration, average incident delay savings, and compliance rates, and is based on regression analyses from samples of data that are used to estimate the benefits of the TIM strategies and extrapolate them to other areas (Figure 4-7). The use of a Big Data approach and Big Data infrastructure could enhance the TIM-BC tool.

Question

How could a Big Data approach enhance the TIM-BC tool?

Traditional Approach

Following a traditional approach, development of the TIM-BC tool used simulations to generate data, which was subsequently used to develop and calibrate regression models. The simulations were needed because the necessary data was not available to develop the models

Segment:

Segment 1

Interstate 99 NB

Huntsville, AL

Roadway Geometry

SEGMENT LENGTH IN MILES: 15

NUMBER OF RAMPS: 2

NUMBER OF TRAFFIC LANES BY DIRECTION: 2

GENERAL TERRAIN: Rolling Hills

HORIZONTAL CURVATURE: Mild Curves

Calculate Ratio Reset Information

SSP Program Information

OPERATION TIME:

☒ AM Peak
☐ PM Peak
☐ Weekday Off Peak
☐ Weekend

INCIDENT DURATION:

Choose how to enter savings:

Average Duration By Lane Blockage

ENTER AVERAGE DURATION SAVINGS: (Minutes) 5

Traffic Information

POSTED MAINLANE SPEED LIMIT (MPH): 65

Time	Traffic Volume (VEH/H/Lane)	Truck Percentage (0-100)
AM PEAK	2700	12

Incident Information

AM Peak

Incident Blockage Severity	Average Incident Duration (Minutes)	Number of Managed Incidents
Shoulder Blockage	15	375
One Lane Blockage	25	150

PERCENTAGE OF ESTIMATED SECONDARY INCIDENTS (enter as 0-100): 4

Source: FHWA (Ma and Lochrane 2015)

Figure 4-7. TIM-BC tool SSP program inputs.

directly. Moreover, as was stated in the January 2016 report, all possible incident simulation combinations (of number of lanes, grade, free-flow speed, traffic volume and composition, number of lanes blocked, and ranges of incident duration) were not replicated (Ma et al. 2016). With the professional workstation used for the analysis, it would have required 16 years to conduct the 740,880 possible runs (three runs of each of 246,960 simulation combinations), not including the time to process the output. Consequently, only about 1,319 “representative” simulation combinations were replicated and used to develop and calibrate the regression models (Ma et al. 2016).

Big Data Approach/Opportunity

In only a bit more time than it would take to run a single simulation on a professional workstation, cloud/Big Data infrastructure would allow each of the 740,880 runs to be run in parallel, and the resulting models could be collated into a single Big Data database. Furthermore, in less than 1 second, Big Data querying/matching engines could be leveraged to efficiently match one of the hundreds of thousands of models in that database to user inputs from web interface tools. In other words, the use of Big Data infrastructure would require fewer assumptions and would result in a much more complete tool.

Given the computational time needed to run the models, the application of a Big Data computing platform could offer efficiencies as compared to the traditional simulations and modeling approach, even if the data for the development of the regression models had to rely on simulations. However, a true Big Data approach would leverage actual data to develop the regression models, rather than running hundreds of thousands of simulations to generate the necessary data. Multi-state or nationwide crash, CAD, roadway, and traffic data, as well as information on SSP programs, laws, levels of TIM training, towing arrangements, dispatch colocation, and TIM taskforces could be leveraged to determine if, where, and when these TIM strategies are effective; what factors impact success (e.g., geographic factors, implementation methods, socio-demographic factors); and what strategic, tactical, and support activities might be employed to improve the probability of success. With this Big Data approach, analysts could reduce reliance on models that typically include expert assumptions and theoretical relationships and, instead, shift to empirical evidence and analytics derived from the entire population of incidents.

Big Data computing power applies brute force analysis that allows for hundreds of thousands of parallel analyses in seconds.

4.7 Summary

This chapter has presented a range of example Big Data opportunities for TIM. The examples presented are by no means exhaustive; rather, they provide a glimpse into potential opportunities to improve TIM using Big Data approaches. Although the example Big Data opportunities are presented in contrast to the more traditional approaches to data collection and analysis, it is important to note that there is nothing inherently wrong with the traditional approaches. Rather, the examples illustrate that the Big Data approach is not simply an improvement on current practices, but instead a radical change from traditional approaches. Big Data represents a paradigm shift that goes beyond data collection and analysis practices to include different data storage, management, and security approaches; different approaches to financing and procuring IT services; and different approaches to development of skills among employees. The shift to Big Data will directly affect the fashion, speed, and frequency with which all businesses, including TIM, are conducted.

When experiments are designed, conscious or unconscious biases can be introduced. When a model is built, assumptions and simplifications typically are made, and when data samples

are collected and analyzed, results can be extrapolated to areas where they might not apply. Experiments, models, and data collection and analysis methods often are driven by budget limitations or by the limitations of the data, software, or computing capabilities at hand.

Many limitations can be overcome through the application of Big Data approaches. Assuming the necessary volume of data is available, Big Data computing power and techniques can allow the data to be leveraged without overwhelming the data analyst. Big Data also allows for a level of granularity (e.g., in data, time, combinations of factors, number of simulations) that traditional approaches cannot come close to meeting.

As evidenced by the examples, potential Big Data applications for TIM range far and wide, particularly compared to what can be done using traditional analytics. Yet, in most circumstances, the significant volume, variety, velocity, and veracity of data is needed to support Big Data analytics, and much of this data is not currently readily available. Moreover, given that incidents are infrequent events (and desirably so), TIM is at a disadvantage from a volume perspective. In counterpoint to limited volume, however, the multi-disciplinary aspect of TIM leads to a variety of data associated with incidents that could benefit from the application of Big Data analytics.

The next chapter presents a comprehensive assessment of selected datasets relevant to TIM. This assessment will help to better understand the maturity and readiness of these datasets to support Big Data analytics to improve TIM.

Assessment of Data Sources for TIM

TIM professionals are at the cusp of harnessing the potential of data to strengthen understanding of program operations and performance. Big Data has the potential to enhance exponentially both the breadth and depth of understanding of policies, strategies, and practices leading to more efficient, effective, and institutionalized programs. This chapter identifies, describes, and assesses current and emerging data sources that might be mined to support TIM program planning and operations and to ultimately advance the state of the practice in TIM.

5.1 Data Source Assessment Approach

The research team's approach to the data source assessment included the following activities:

- Develop initial list of data sources;
- Develop assessment criteria;
- Conduct research on data sources; and
- Identify and apply data maturity assessment model(s).

An initial list of data sources was developed based on the expertise of the research team, the findings from the state-of-the-practice review, and input from a variety of TIM responders and the NCHRP project panel. Next, a list of assessment criteria was developed that would provide a range of information about each source. With the list of data sources and the assessment criteria, research was then conducted to populate a data source assessment table for each data source. Information for the assessment was gathered from Internet and literature searches and reviews, and from interviews with the following data owners:

- The American Association of Motor Vehicle Administrators (AAMVA), for driver, vehicle, and commercial vehicle driver data;
- The Arizona Professional Towing and Recovery Association, for towing data;
- The University of Utah, regarding the National EMS Information System;
- The FMCSA, for motor carrier management information system data;
- The Florida Department of Emergency Management (FDEM), for emergency management data;
- The Florida Department of Highway Safety and Motor Vehicles, for citation/adjudication, crash, and licensing data;
- The Florida DOT, for roadway inventory, safety service patrol, traffic, weigh station, 511 system, and tolling data;
- The Florida Highway Patrol (FHP), for computer-aided dispatch data and crash data;
- HERE North America, LLC (a division of HERE Technologies), for vehicle probe speed data;
- The Nassau County Sheriff's Office, Nassau County, Florida, for 911 data and video data;

- The National Association of State Emergency Medical Services Officials (NASEMSO), for EMS data;
- The NHTSA, for crash data, including fatal data from the nationwide Fatality Analysis Reporting System (FARS);
- Southern Towing, in Jacksonville, FL, for safety service patrol data;
- The Sunshine State Towing Association (SSTA), for towing data;
- The Utah Department of Transportation (Utah DOT), for road weather data; and
- The Wisconsin Department of Transportation (Wisconsin DOT), for video data.

The research process uncovered new sources, and some of the data sources were merged, re-grouped, or eliminated due to the nature and/or relationships of the data sources. The result was a list of 31 data sources grouped into the following six data domains:

- State traffic records data,
- Transportation data,
- Public safety data,
- Crowdsourced data,
- Advanced vehicle systems data, and
- Aggregated datasets.

5.1.1 Assessment Criteria

The criteria that were applied to assess each data source, together with examples for each criterion, are shown in Table 5-1. Although most of these criteria are relatively self-explanatory,

Table 5-1. Data source assessment criteria.

Data Source Assessment Criteria	
Description of Data	A brief synopsis of the data
Organization that Collects, Maintains, and Owns the Data	Examples: State DOT, public safety agency, private vendor
How the Data Is Collected	Examples: Manually, via sensors, via video cameras, auto-populated, probes/crowdsourced
Data Structure	Examples: Unstructured (free text), semi-structured (XML, CSV, JSON, Excel), structured (SQL database)
Data Size	Examples: Megabytes (MB) for spreadsheets or PDFs, gigabytes (GB) for relational databases, terabytes (TB) for large relational databases, petabytes (PB) for NoSQL databases
Data Storage and Management	Examples: Office maintaining a spreadsheet, local (city/county) database, state database, national data store, in-house, cloud, third-party, length of archive
Data Accessibility	Examples: Call or email to request a data dump (disk), file transfer protocol (FTP), web services
Data Sensitivity	Examples: Yes/No, presence of personally identifiable information (PII), other sensitive or security related issues
Data Openness	Examples: Open, shared, closed (see 5.1.1.1 <i>Data Openness</i> in this chapter)
Data Challenges	Examples: Data silos, lack of standards, privacy, security, legal, interoperability (see 5.1.1.2 <i>Data Challenges</i> in this chapter)
Data Costs	Examples: Publicly available and free, one-time fee, subscription based, pay-as-you-go (see 5.1.1.3 <i>Data Costs</i> in this chapter)

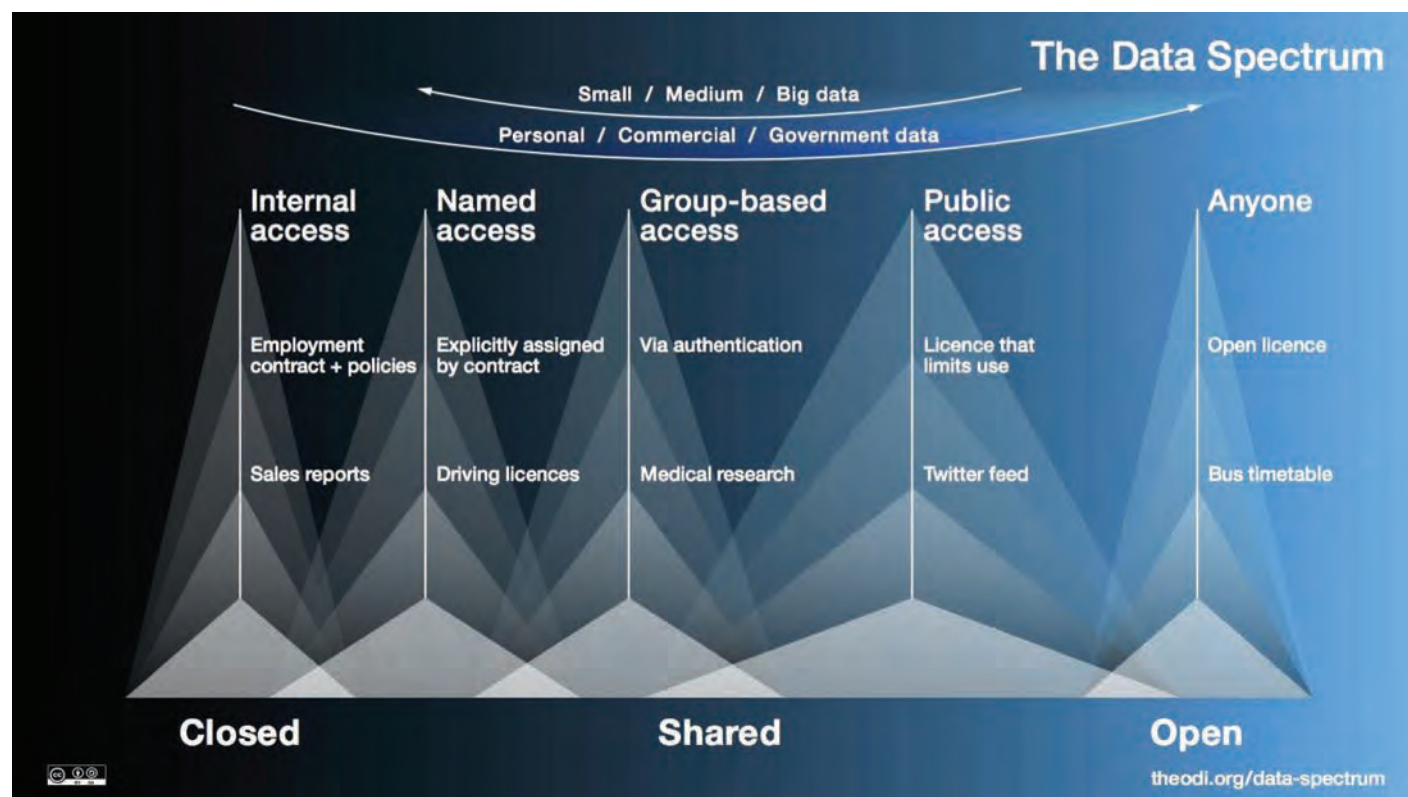
the data openness, data challenges, and data costs categories warrant more explanation and are discussed further in the text that follows Table 5-1.

5.1.1.1 Data Openness

The openness of data typically is assessed on three factors: availability and access, re-use and redistribution, and universal participation. For data to be considered open, the following conditions must apply (Open Knowledge Foundation n.d.-b):

- **Availability and access:** The data must be available as a whole in full granularity, at no more than a reasonable reproduction cost, and preferably by download over the Internet. The data must also be available in a convenient and modifiable form.
- **Re-use and redistribution:** The data must be provided under terms that permit re-use and redistribution, including the intermixing with other datasets.
- **Universal participation:** Anyone must be able to use, re-use, and redistribute the data. There should be no discrimination against fields of endeavor or against persons or groups.

In other words, “open data and content can be freely used, modified, and shared by anyone for any purpose” (Open Knowledge Foundation n.d.-a). At the other end of the spectrum, data that is considered *closed* “can only be accessed by its subject, owner, or holder” (Broad 2015). Somewhere in the middle is data that is *shared*. Shared data includes data with named access (e.g., data that is shared only with named people or organizations), data with attribute-based access (e.g., data that is made available to specific groups, such as public agencies or university students, who meet specific criteria), and public access data (e.g., data that is available to anyone but under terms and conditions that are not considered to be completely open) (Open Data Institute n.d.). Figure 5-1 illustrates the spectrum of data from closed to open.



Source: Open Data Institute (n.d.)

Figure 5-1. The data spectrum.

The openness of data is important for several reasons. For example, data openness allows for:

- **The interoperability of datasets:** Without interoperability, merging disparate datasets is very challenging, and without the ability to merge disparate datasets, it is impossible to discover relationships (correlations) between them, which is one of the primary goals of—and is essential to—Big Data analytics.
- **More people to use the data:** Openness improves the quality of the data and makes it more useful because, as more people explore and use the data, (1) more flaws are discovered and corrected and (2) the chances increase of discovering valuable insights from the vast and complex datasets.
- **Better data preservation:** Although digital data storage devices can keep data for a long time, they still decay and eventually fail, leading to data losses. Given the sheer number of devices involved, data storage device failures are even more frequent when storing Big Data datasets. Open data can more easily be stored in multiple locations and duplicated across many more storage devices, however, thus reducing the chances of data loss.

Closed data limits who can access the data, what data can be accessed, how it can be accessed, and what the data can be used for. From a Big Data perspective, data that is closed is limiting, as it may not be able to be joined with other datasets, be read by common Big Data analysis software, or be searchable and minable by a broader set of people.

Open data is essential to Big Data analytics; however, opening data involves a balancing act between maximizing the value that can be derived from opening the data and minimizing the privacy or security risks of doing so.

Although opening data provides many benefits, it also can expose sensitive data and increase privacy and security risks. In the private sector, opening data carries the additional risk of losing a competitive advantage. Therefore, opening data involves a balancing act between maximizing the value that can be derived from opening the data and minimizing the privacy, security, or business risks associated with doing so.

5.1.1.2 Data Challenges

Although Big Data approaches offer a host of benefits, several challenges are associated with Big Data, from accessing datasets to the data elements within the datasets to the use and analysis of the data. The list that follows is by no means exhaustive, but it offers a brief discussion of some of the most common challenges inherent in many datasets:

- **Data silos:** Every agency collects and stores data on some level. Often, however, the data is isolated within one or more business units and is not shared or integrated with data from the rest of the organization. Data silos often arise naturally; if institutional coordination has not been emphasized, for example, organizational units may have developed differing goals, priorities, responsibilities, and isolated datasets. The lack of coordination makes it harder to integrate these diverse datasets into the kinds of large, comprehensive datasets needed for Big Data analytics. The challenge of data silos is further complicated across organizations, agencies, and states.
- **Interoperability:** Accessibility and usability have a technical aspect that can be problematic when sharing or integrating data. The differing technical standards used for communication, storage, and retrieval of various datasets across and between organizations can increase the difficulty of merging disparate data and creating and maintaining comprehensive datasets.
- **Public records laws:** Given that many public agencies are bound by public records laws, agencies must be careful not to imperil third-party or private data. Although public records are records of public business, they are not necessarily available without restriction. Each level of government has policies and regulations that direct the availability of information

contained in public records. A common restriction is that data about a person is not normally available to others. In the United States, access to national public records is guided by the Freedom of Information Act (FOIA). All U.S. states also have some form of FOIA legislation, but the accessibility of public records varies across states. In some states, it is easy to request and receive documents; in other states, many exemptions and restricted document categories complicate and reduce access. Requests for access to records pursuant to FOIA may be refused if the information requested is subject to exemption; alternatively, some information may be redacted.

- **Security:** Particularly in electronic transmission between entities, sharing data always carries the risk that the data will be stolen, compromised, corrupted, or infected. The risks associated with data security can lead agencies to be unwilling to share their data or to accept data from others, which limits the aggregation of data for Big Data analytics.
- **Privacy:** Data privacy concerns are associated with the storage of data that contains personal details and *personally identifiable information* (PII). Some kinds of data are bound by privacy laws that restrict the use and distribution of the data (e.g., HIPAA, the Health Insurance Portability and Accountability Act of 1996). The ability to merge datasets that contain private data—particularly data covered by legal restrictions—is therefore limited. Having data that is only available at a lower resolution (because certain details or elements are removed for privacy reasons or by not having access to the data at all) can limit the possibility of analysis. Furthermore, given the need to work with or bypass the security measures that are used to protect the private data in each dataset, attempts to merge such datasets are fraught with risks to the safety, integrity, and completeness of the resulting information.
- **Proprietary data:** Some data can be considered *intellectual property* (i.e., its use may be restricted on the basis of its value as a trade secret or trademark, or under a copyright or patent). Such proprietary data can be the basis for a competitive advantage in business and therefore can be restricted and most (but not all) such proprietary data is generated in the private sector. Agreements for sharing proprietary data often are negotiated at the end-product level (applying to visualization tools, web tools, reports, and so forth), not at the raw data level. Sensitivity to privately owned information is required.
- **Retention:** The retention period for records can be another obstacle to building the large historical datasets needed for Big Data analytics. The agency policies or laws dictating how long data will be kept may not yet reflect the extensive data archiving needs associated with Big Data.
- **Emerging forms of data:** The technical and legal foundations for handling some kinds of data are new, and may have unique characteristics. Data associated with connected vehicles, autonomous vehicles, GPS, and photographic/surveillance using drones are examples of such emerging forms of data. According to a study conducted by the RAND Corporation, data ownership and privacy issues related to autonomous vehicle communications remain unsettled, and this is an important policy gap that needs to be addressed (Anderson et al. 2016).
- **Technical analysis expertise:** Big Data analysts and experts are in very high demand and tend to gravitate toward companies with existing Big Data expertise, that own large datasets, and that pay well above market. The result is a shortage of individuals with the expertise and desire to undertake such an endeavor in government agencies, among government contractors, and even at universities.
- **Inherent rarity and variability of traffic incidents:** Likely one of the biggest challenges for the application of Big Data to TIM is the fact that traffic incidents are rare, and no two traffic incidents are exactly alike. This inherent rarity presents challenges in developing sufficiently complete historical traffic incident datasets capable of characterizing both incidents and the associated responses well enough to effectively identify patterns and trends that can lead to improvements in traffic incident response.

5.1.1.3 Data Costs

The costs associated with obtaining, preparing, and using data can be divided into five cost categories; specifically, the cost of:

1. Acquiring the data,
2. Storing the data,
3. Securing the data,
4. Managing the data, and
5. Using the data.

Each cost category can be further divided, depending on the way the data is offered and the amount of work or infrastructure needed to acquire, store, secure, manage, and use it. As a result, the overall cost of using the data can sometimes be quite significant, even if the raw data is made available at no cost.

Weather data provides a helpful example, as follows:

- **National Oceanic and Atmospheric Administration (NOAA) weather data:** Raw weather data is available from NOAA free of charge. The data provided may not be readily usable, however: the datasets are in a scientific file format, they are large, and they change every few hours or days. To make the data useful, therefore, costs typically must be incurred to do the following:
 - Convert the data from the original format to a comma-separated value (CSV) or JSON file format for more effective data mining;
 - Create, operate, and maintain the infrastructure necessary to securely store and query the data (although this cost can be lowered significantly by using cloud services, if allowed); and
 - Update and maintain the data as new files become available.

Taken individually or together, the costs associated with making the data useful typically are not negligible and may be significant.
- **Commercial online weather data service:** A typical cloud-based weather data service would collect weather data from multiple weather agencies around the world (including NOAA), manage and maintain the data, and offer various historical and predictive real-time online services for a fee. Such online services rarely share or offer the entire dataset via download; rather, with each request, users obtain and see only some of the data. Essentially, the commercial service sells the ability to access and search a maintained weather dataset without taking on the costs of developing the necessary infrastructure and personnel to do so independently. Because its platform is designed to serve millions of requests per minute, the service company's own investment in these costs can be spread over many subscribers, which greatly lowers the cost of any single request.

Depending on the purpose, scale, and nature of the desired data analysis (real-time or historical), the existing data storage infrastructure, and the existing in-house data analysis tools and expertise, one approach to acquiring the data may be less costly than another. In many cases, however, the economies of scale offered by commercial online data services may be persuasive in comparison to the full cost of acquiring, storing, and maintaining data in-house.

5.1.2 Data Maturity Assessment Approach

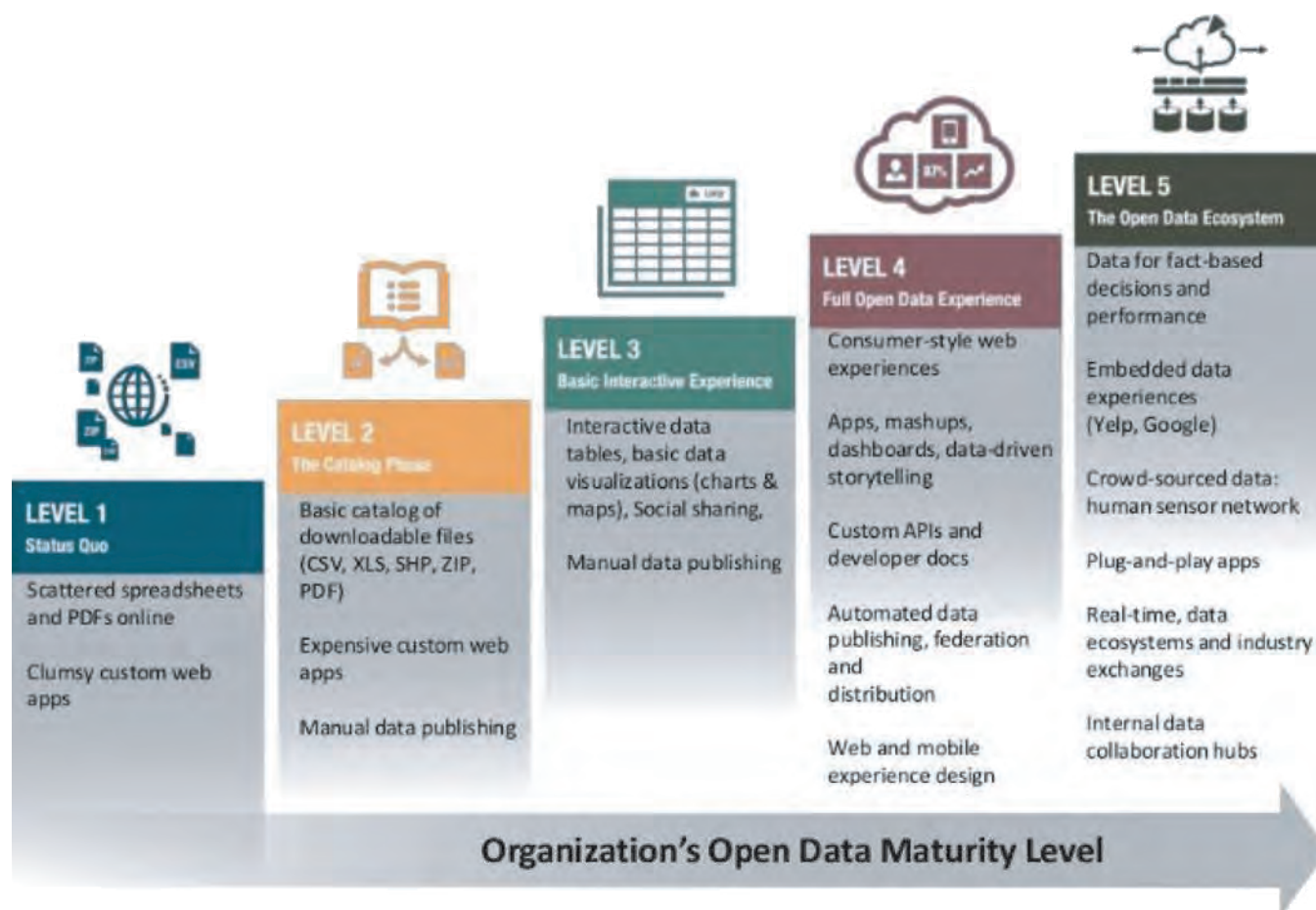
Following completion of the data assessment tables, the research team rated the maturity of the data sources using two different data maturity models: the Socrata Open Data Maturity Model and the Center for Data Science and Public Policy at the University of Chicago Data Maturity Framework. The Socrata Open Data Maturity Model provides a quick and simple way to classify data maturity in terms of a single level (1, 2, 3, 4, or 5), whereas the University of

Chicago Data Maturity Framework offers a more-involved assessment of data maturity, called data *readiness*, based on multiple criteria and multiple maturity levels, with no single qualitative or quantitative output. The use of both models provides a more comprehensive look at the maturity of each data source.

The Socrata Open Data Maturity Model is shown in Figure 5-2 (Socrata, Inc. 2014). The various levels emphasize an approach of open data *curation*. Data curation is the management of data throughout its lifecycle, from its creation and initial storage to the time when it is archived for posterity or deleted as obsolete. The main purpose of data curation is to ensure that the data is reliably retrievable for future research purposes or reuse. The Socrata Open Data Maturity Model categorizes the various stages of open data curation from unorganized and inaccessible (Level 1) to fully collaborative, interactive, shareable and augmentable (Level 5).

The University of Chicago Data Maturity Framework was developed at the university's Center for Data Science and Public Policy based on conversations and work with dozens of organizations regarding their data, their organizational culture, and their ability to act on any insights coming out of projects (University of Chicago n.d.). The framework consists of the following elements:

- Data Maturity Framework Questionnaire,
- Data and Tech Readiness Scorecard, and
- Organizational Readiness Scorecard.



Source: Socrata, Inc. (2014)

Figure 5-2. Socrata open data maturity model.

The questionnaire and scorecards were developed to help non-profits, government agencies, and other groups evaluate their data maturity and identify what they need to do to move forward with a successful data-driven project (Haynes 2015).

Using the Data and Tech Readiness Scorecard (Figure 5-3), the research team used the data readiness criteria from the Data Maturity Framework Questionnaire (listed in Table 5-2) to assess the readiness of each of 31 data sources.

5.2 Findings

This section presents the findings from the research team’s assessment of the 31 data sources classified within six data domains and summarized in Figure 5-4.

For each data domain, the research team’s findings are presented as follows:

1. The data sources within the domain are introduced and described.
2. A high-level summary of the data sources briefly discusses what can be found in the detailed assessment tables in Appendix A of this report.
3. Costs are addressed, and challenges are discussed.

For each data source, the subjective maturity assessment/rating results based on the Data Maturity Framework Questionnaire’s data readiness questions and the Socrata Data Maturity Model assessment are presented. The Socrata Data Maturity Model assessment results are presented using the following icons:



Whereas this chapter provides a summary assessment of the 31 data sources within the six domains, Appendix A provides a comprehensive inventory of all 31 sources in tabular form.

5.2.1 State Traffic Records Data

5.2.1.1 Description of Sources

The NHTSA has been instrumental in working with states to develop the processes that govern the collection, management, and analysis of state traffic records data. Traffic records are foundational to highway driving and the fiduciary role that states have in managing driver, vehicle, and related data. Functionally, a traffic records system includes the collection, management, and analysis of traffic safety data and comprises six core data systems—crash, driver, vehicle, roadway, citation and adjudication, and injury surveillance. High-quality state traffic records data is critical to effective safety programing, operational management, and strategic planning. NHTSA states that, “Every state—in cooperation with its local, regional, and federal partners—should maintain a traffic records system that supports the data-driven, science-based decision-making necessary to identify problems; develop, deploy, and evaluate countermeasures; and efficiently allocate resources” (NHTSA 2012).

Within the traffic records data domain, six core data sources were assessed:

- **Crash data:** Crash data, typically gathered by law enforcement, documents the characteristics of a crash and provides the who, what, when, where, how, and why about each incident. The

Data Maturity Framework Data and Tech Readiness Scorecard					
Category	Area	Lagging	Basic	Advanced	Leading
How Is Data Stored	Accessibility	Only accessible within the application where it is collected	Can be accessible outside the application but proprietary format, requiring specialized analysis software	All machine readable in standard open format (CSV, JSON, XML, database)	All machine readable in standard open format and available through an API
	Storage	Paper	PDFs or Images	Text Files	Databases
	Integration	Data sits in the source systems	Data is exported occasionally and integrated in ad hoc manner	Central data warehouse - realtime aggregation and linking (Automatic)	External data also integrated
What Is Collected?	Relevance and Sufficiency	The data you are collecting on subjects of interest is irrelevant to the problem you want to solve: ie you want to predict which students need extra support to graduate on-time but don't have data on graduation outcomes	Some of the data you have is relevant, but it is insufficient because key fields are missing, ie no data on academic behavior or attendance history, etc.	You have data that is helpful and relevant for solving the problem but not sufficient to solve it well, ie you have yearly academic and demographic information but are missing extra-curricular activities, or interventions they were targeted with	You have all the relevant data about all the entities being analyzed and it's sufficient to solve the problem you are tackling
	Quality	Missing rows (people/address level entities missing in the data)	Missing columns (variables missing)	No missing data but errors in data collection such as typos	No missing data and no errors in data collection
	Collection Frequency	Once and never again	yearly	frequently	realtime
	Granularity	City level aggregates	Zipcode/Block level aggregates	Individual level (person or address) data	Incident/Event level data
	History	No History Kept - old data is deleted	Historical data is stored but updates overwrite existing data	Historical data is stored and new data gets appended with timestamp, preserving old values	All history is kept and new data schema gets mapped to old schema so older data can be used
Other	Privacy	No privacy policy in place	no PII can be used for anything	ad-hoc approval process in place that allows selected PII data to be used for selected/approved projects	Software defined/controlled privacy protection that allows analytics to be done while preserving privacy based on predefined policies
	Documentation	no digital documentation or metadata: data exists but field descriptions or coded variables are not documented	data dictionary exists (variables and categories defined)	data dictionary plus full metadata available (including conditions under which the data were captured)	data dictionary plus full metadata available including collection assumptions, what's not collected, and potential biases

Source: University of Chicago (2017)

Figure 5-3. Data maturity framework: data and tech readiness scorecard.

Table 5-2. Data maturity framework questionnaire: data readiness questions.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				
Storage				
Integration				
Relevance and Sufficiency				
Quality				
Collection Frequency				
Granularity				
History				
Privacy				
Documentation				

Source: University of Chicago (n.d.)

Model Minimum Uniform Crash Criteria (MMUCC) is a voluntary data collection guideline. The MMUCC guideline identifies a minimum, standardized set of motor vehicle crash data elements and their attributes that states should consider including in a state crash data system. The MMUCC 5th Edition contains 115 data elements (U.S. DOT 2017b).

- **Vehicle data:** Vehicle data encompasses an inventory of data that enables the titling and registration of each vehicle under a state's jurisdiction to ensure that a descriptive record is maintained and made accessible for each vehicle and vehicle owner operating on public roadways. Vehicle information includes identification and ownership data for vehicles registered in the state and out-of-state vehicles that are involved in crashes within the state's boundaries. Although data elements vary between jurisdictions and are sometimes defined differently, data elements generally include the following:
 - Issuing agency,
 - Plate type,
 - Vehicle year,
 - Body style,

**Figure 5-4. Data domains and data sources assessed.**

- Vehicle weight, and
- Vehicle identification number (VIN), and
- Name of vehicle owner.
- **Driver data:** Driver data is used to maintain driver identities, histories, and licensing information for all records in the system. The driver data system ensures that each person licensed to drive has one identity, one license to drive, and one record. For each licensed driver, driver data generally includes the following:
 - Name,
 - Birth date,
 - License number,
 - Issuing state,
 - License type, and
 - Number of violations and points.
- **Roadway data:** Roadway data is composed of data collected by the state (state-maintained roadways and, in some cases, local roadways), as well as data from local sources such as county and municipal public works agencies and metropolitan planning organizations (MPOs). The Model Inventory of Roadway Elements (MIRE) is a recommended listing of roadway inventory and traffic elements critical to safety and is the major guideline pertaining to the roadway system. MIRE Version 1.0 is made up of 202 elements, of which 38 elements have been identified as Fundamental Data Elements (FDEs).
- **Citation and adjudication data:** Citation and adjudication databases maintain information about citations, arrests, and dispositions from delivery of citation through adjudication. The process is highly localized in data management. In most states, following local adjudication, the data is delivered to a state entity for driver's license reporting functions.
- **Injury surveillance data:** Injury surveillance data typically incorporates information about pre-hospital emergency medical services (EMS), trauma registry, emergency department, hospital discharge, rehabilitation, payer-related details, and mortality (e.g., death certificates, autopsies, and coroner and medical examiner reports). Given the numerous files and datasets that make up the injury surveillance system, there are a correspondingly large number of data standards and applicable guidelines for data collection.

5.2.1.2 Summary of Findings

NHTSA and its state-level partners have created a framework for systematically collecting and cataloging relevant traffic records (NHTSA 2012). Because diverse agencies handle traffic records data at the state level, each state has a Traffic Records Coordinating Committee (TRCC) made up of representative data collectors, managers, and users drawn from each of the core traffic records system areas. TRCC members also may include users of integrated datasets, which are created when various types of data from component systems are linked. TRCCs promote quality, accuracy, uniformity, and utility of data, but the committees themselves are not repositories for data.

Traffic crash report data is a high-value, high-quality set of data, particularly for evaluating historical characteristics and trends. Data elements in crash reports are driven by the MMUCC. Many states have begun to use crash reporting systems to document important data elements such as clearance times and secondary crashes for TIM performance analysis.

The role of government in regulating the licensing of drivers and the registration of vehicles is critical to roadway safety. At the state level, drivers are required to be licensed to operate a vehicle, and motor vehicles are required to be titled and registered. Licenses, titles, and registrations must be renewed when a significant change occurs (e.g., a driver moves from one state to another or a vehicle changes ownership) or on a regular basis as defined by the state. These license and registration processes generate data, which is collected and maintained by

state departments of motor vehicles or their equivalents. Large trucks and other commercial motor vehicles are an important subset of licensing and registration systems. For these vehicles and drivers, state systems are augmented by a pointer index that allows for expedited communication between state licensing authorities.

An important part of the driving record is the recording of crashes and traffic citations. Data about traffic citations issued by law enforcement and about cases adjudicated in the citation and adjudication system is fed by local and state court systems into the drivers license system.

State agencies share varying amounts of information with the American Association of Motor Vehicle Administrators (AAMVA). AAMVA develops and maintains many information systems that facilitate the electronic exchange of driver, vehicle, and identity information between organizations. Driver and motor vehicle systems are mature and, to some extent, standardized across states. The data is readily available to law enforcement incident responders via in-vehicle computer systems or via radio contact with dispatch. The data has the potential to augment TIM efforts, for example, when assessing the size and type of vehicle for towing requests.

Roadway inventory and asset management databases are used to collect and maintain data about a state's roadways, including all signs, signals, markings, and geometric and roadside characteristics. When combined with crash and other data, roadway data has the potential to reveal engineering and other issues associated with incidents and incident clearance.

The final type of data that makes up the state traffic records domain is injury surveillance data, which typically is created by EMS professionals who respond to crash scenes to aid the injured. EMS *run reports* form the basis for injury surveillance, but often that basic information is augmented by hospital data and, in the case of a fatality, by medical examiner or coroner data. Most EMS data is collected according to the National Emergency Medical Services Information System (NEMSIS) standard. EMS agencies collect the data at the local level and send the data to a state-level database. A subset of the data is then sent from the states to the NEMSIS national repository, which is maintained by the NEMSIS Technical Assistance Center (TAC) at the University of Utah. NEMSIS data (from local, state, and national databases) is an untapped source of data for TIM analysis.

In general, state traffic records data is a public asset and made available at little to no cost; however, public records laws and privacy issues may limit the availability of the data (or some data elements), which could result in a loss of data upon integration. Other challenges and limitations that are associated with the use of state traffic records data for Big Data analytics for TIM include the following:

- Because the MMUCC is voluntary, states use varying formats and names for data elements and attributes, and may combine (or split) MMUCC elements and attributes (U.S. DOT 2017b). These variations can make it difficult to compare, merge, or share crash data among states, between state and federal datasets, and, in some cases, even between different agencies within a state.
- Within any given state, many agencies utilize electronic crash-reporting systems, which can result in more complete and exploitable data; however, some agencies still use paper crash reports, which result in data that is less precise (vague time or location) or of lesser quality (missing fields, wrong categories, etc.). The latter approach also can delay the upload of crash reports into a local or state database as state or local personnel perform additional inquiries to obtain more precise or correct data. It should be noted that errors can occur in data accuracy or completeness in electronic crash data systems.
- State traffic records data, or data elements therein, may not be accessible due to PII and other restrictions like state laws that protect driver information.

- Disparities in the formats and names for data elements and attributes sometimes make it difficult for officials in one jurisdiction to interpret data elements that appear on the vehicle registration documents of another jurisdiction.
- Challenges in accessing the data in bulk or raw format may limit the usefulness for Big Data analytics. For administrative purposes, some traffic records data can be shared between states, but it rarely moves outside of “official purposes” because of the presence of PII or state laws that protect the information.
- Roadway inventory systems within and across the states range widely in maturity level, from simple spreadsheets to sophisticated web services, and this variation has an impact on the quality, timeliness, and accuracy of the data. Many agencies may not have a web portal or FTP site, which means that large datasets must be delivered via disc or mail. Some agencies only use basic file-sharing systems to store their data, and these systems lack the data management structure to easily find, retrieve, and format requested data quickly. Following a request, it is not uncommon to have to wait several days to receive data.
- The collection and management of roadway data may be distributed across agency districts, with the result that it is not routinely managed, updated, and maintained in a consistent fashion. Depending on budget and staff availability, each district may manage its roadway data differently. The result may be the storage of roadway data across various internal legacy systems with diverse structures and formats, which could make it very difficult to access and mine the data. The accuracy of roadway data also can be affected, as some agencies or districts may not have the resources to update records as soon as an asset is replaced or upgraded. Consequently, stale roadway data may remain in the dataset for weeks or months after asset work has been performed; worse, the dataset can hold data that incompletely reflects roadway assets.
- The NEMSIS location data available at the state and national level is limited to the zip code level. This limitation could greatly limit data analytics, as the resolution would be too low for meaningful analysis.

Even with these challenges, state traffic-record databases present a relatively easy starting point for creating TIM-relevant Big Data datasets from state data. The NHTSA has already established the MMUCC standard for state crash data and provides states with MMUCC mapping tools. NHTSA also offers associated technical assistance (e.g., the NHTSA Traffic Records GO Team program) to improve traffic records data collection, management, and analysis capabilities and to examine the quality of a state’s crash data, and provides specific recommendations to improve the quality, management, and use of that data to support safety decisions. As part of its roadway safety data program, the FHWA Office of Safety has established the Model Inventory of Roadway Elements (MIRE) to help transportation agencies improve their roadway and traffic data inventories (FHWA n.d.-b). In addition, the NHTSA’s 2012 Traffic Records Program Assessment Advisory gives states information on the contents, capabilities, and data quality of an effective traffic records system by describing an ideal system that supports high-quality decisions and leads to cost-effective improvements in highway and traffic safety. The NHTSA Advisory outlines a comprehensive approach for assessing the systems and processes that govern the collection, management, and analysis of traffic records data (NHTSA 2012). By using the MMUCC, MIRE, NEMSIS, and NHTSA Advisory as guides for creating more uniform databases, more state datasets could be combined and integrated into detailed and reliable datasets that could provide a solid foundation for TIM Big Data analysis.

The next set of tables (Tables 5-3 through 5-8) show the subjective readiness ratings given to each data type of the state traffic records data sources. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the data sources can be found in Appendix A, Tables A-1 through A-6.

Table 5-3. Crash data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	
Storage				√
Integration	√			
Relevance and Sufficiency		√	√	
Quality		√	√	
Collection Frequency	√	√		
Granularity				√
History			√	
Privacy			√	
Documentation		√	√	

**Table 5-4. Vehicle data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	√
Collection Frequency	√			
Granularity				√
History			√	
Privacy			√	
Documentation			√	

**Table 5-5. Driver data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage			√	√
Integration		√	√	
Relevance and Sufficiency			√	
Quality		√	√	
Collection Frequency		√		
Granularity			√	
History			√	
Privacy			√	
Documentation		√	√	



Table 5-6. Roadway data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	
Quality		√	√	√
Collection Frequency		√	√	
Granularity			√	
History		√	√	
Privacy	√			
Documentation			√	√

**Table 5-7. Citations and adjudication data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	√
Collection Frequency			√	√
Granularity				√
History			√	
Privacy			√	
Documentation			√	

**Table 5-8. Injury surveillance data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	√
Collection Frequency			√	√
Granularity			√	√
History			√	
Privacy			√	√
Documentation			√	



5.2.2 Transportation Data

5.2.2.1 Description of Sources

Within the transportation data domain, the following six data sources were assessed:

- **Traffic sensor data:** A suite of in-roadway or over-roadway sensors provides the mainstay for transportation agencies to plan for and operate the road network. Sensors include inductive loop detectors, magnetic sensors and detectors, video image processors, microwave radar sensors, laser radars, passive infrared and passive acoustic array sensors, and ultrasonic sensors, plus combinations of these sensor technologies. Certain detectors give direct information concerning vehicle passage and presence, whereas other traffic flow parameters, such as density and speed, are inferred from algorithms that interpret or analyze the measured data.
- **Traffic digital video data:** Digital video is a representation of moving visual images in the form of encoded digital data. Digital video data is collected by transportation agencies through closed-circuit television (CCTV) cameras (video surveillance), video detection, and automatic license and plate reader/recognition (ALPR) systems. Transportation agencies use CCTV cameras on highways and at ramp locations and intersections to monitor traffic from a central location. Video detection devices capture video images of traffic and analyze the information using algorithms for traffic management (e.g., traffic signal control). ALPR systems identify vehicles passing fixed locations using cameras that read the license plates.
- **Safety service patrol/incident response program data:** This data is collected by safety service patrol (SSP) or incident response (IR) staff present at the scene of an incident, which generally includes location of the incident, arrival and departure times, and assistance provided. Depending on the program, data may be collected by SSP/IR operators manually using simple paper forms or logs or electronically via laptops, tablets, or mobile phones, and may be communicated (e.g., via radio) back to a central location such as a TMC.
- **Road weather data:** Road weather data consists of precise, relevant, and timely weather information and its effects on the road (BTS 2011). Road weather data collected at roadway locations can include atmospheric, pavement, and water level conditions. Atmospheric data includes air temperature and humidity, visibility distance, wind speed and direction, precipitation type and rate, tornado or waterspout occurrence, lightning, storm cell location and track, as well as air quality. Pavement data includes pavement temperature, pavement freeze point, pavement condition (e.g., wet, icy, or flooded), pavement chemical concentration, and subsurface conditions (e.g., soil temperature). Water level data includes tide levels (e.g., high or low tide or hurricane storm surge) as well as stream, river, and lake levels near roads (FHWA 2017a).
- **Traveler information (511 system) data:** Acquiring, analyzing, and communicating information to inform and guide surface transportation travelers, 511 system data can include general traffic (congestion and speeds) and weather conditions, as well as the location of incidents, work zones, roadway closures, and planned special events. Data sources to 511 systems generally include the state DOT, highway patrol and police departments, transit agencies, and sometimes local jurisdictions and private companies.
- **Toll data:** Toll data, collected via electronic toll collection technology, includes the number of vehicles passing through toll gates, vehicle identification, automated vehicle classification, transaction processing, and violation enforcement data.

5.2.2.2 Summary of Findings

One of the most recognizable data domains with potential for application to TIM is that which is created and housed by transportation agencies. Transportation agencies collect, own, store, and manage a variety of datasets. Intelligent Transportation System (ITS) devices in the field, consisting of sensors and CCTV cameras, generate data about operations. These data often

converge in TMCs, where software systems like advanced traffic management systems (ATMSs) combine the data and store it in relational databases. Programs such as SSPs collect data related to the response activities associated with roadway incidents and crashes. Most SSP data is still collected using paper forms that are later entered into a database or spreadsheet or by a TMC operator in radio communication with incident responders. More modern ways of collecting service patrol data are becoming more prevalent. These systems, such as CAD systems or mobile phone/tablet applications, capture data at the scene using a more structured and strict data-collection process. Ultimately, much of the transportation data is packaged, along with other data, for real-time consumption by road users in the form of traveler information via 511 and similar systems.

Data collected by transportation agencies is most frequently used/analyzed for the maintenance, operation, and safety of the roadways. Increasingly, transportation data is being used for the analysis of performance. Although analyses of the datasets typically are conducted separately for specific purposes (e.g., safety analysis, operational analysis), Big Data offers opportunities to combine data sources to gain further insights and identify unforeseen trends about the operations and safety of roadways.

According to the FHWA's Road Weather Management Program (RWMP), weather plays a role in 24 percent of all crashes, having resulted in more than 7,100 deaths and more than 629,000 injuries over a 13-year period (BTS 2011). Understanding the safety implications of weather (*road weather* in the transportation world), most state DOTs operate road weather information systems (RWIS). RWIS collect and monitor weather data via environmental sensor station (ESS) equipment installed along roadways. Some RWIS programs also have expanded to use weather sensors affixed to AVL-enabled fleet vehicles to collect road weather and response data such as the salt spread rate and pavement temperature during operation.

Because transportation agencies own the data generated by DOT-owned systems, they have significant control over what data is collected and how the data is collected. Transportation agencies typically will share the data with other public agencies, and even with private agencies that have a legitimate use for the data. The data typically is characterized as public domain data and provided at no cost. Video data typically is available for free to the public (at low resolution) or to other agencies and institutions (at high resolution). Even when compressed, however, video and image data files require large storage capabilities. Consequently, the cost associated with the on-site storage and retention of video and images can be significant. The amount and quality of data stored, compression ratios, image size, and retention period are factors that impact operational cost. Cloud storage services typically are used to store video and images as they offer the most economical storage solution and allow video to be stored without degrading its quality; however, cloud storage is rarely used by TMCs. Some of the obstacles that currently prevent greater use of cloud storage are discussed in Chapter 6 of this report.

Although transportation data is extensive, limitations and challenges impact agencies' ability to leverage the data for Big Data analysis of TIM:

- Instrumentation of roadways with sensors, cameras, and ESS usually is geographically limited to the roadways and locations of most interest or concern. These locations include areas with significant congestion or weather-related issues along interstate highways, state highways, and sometimes (but much less frequently) major urban arterials. As such, TMCs and SSPs generally operate only in urban areas and sometimes have limited hours of operation. The result is large gaps in data across most states, limiting the potential for TIM analysis.
- From a systems perspective, legacy systems do not always integrate easily with other systems. In addition, the proprietary nature of many transportation systems limits what and how the data is collected, as well as the integration with data from other systems.

- TMCs are currently challenged with assimilating data from a variety of sources and deriving measures of traffic management performance. Big Data makes more data available to calculate meaningful measures, but the proliferation of Big Data also increases the demand for detailed reporting, thus increasing the challenges (Gettman et al. 2017).
- Variations, imprecision, and/or absence of location data within or across datasets can result in challenges to data use. For example, in datasets from one state agency, metadata cited 30+ formats related to location, ranging from latitude and longitude to mile markers to street names.
- The quality of traffic and RWIS sensor data depends greatly on the ability of the transportation agency to maintain its equipment regularly and to recalibrate or replace defective or drifting sensors swiftly. Without prompt and efficient maintenance, sensors can start to report erroneous values or report no data at all, slowly introducing gaps and biases in datasets that can be difficult to circumvent when performing analysis.
- Most TMC videos or images are not stored or archived. When video data is stored, it typically is stored and maintained only for a brief period, then purged to make room for newer video. This practice greatly limits the quantity of video content available for mining.
- Although some transportation video data is high definition, some video data remains at low definition, which affects the ability to efficiently analyze video feeds using automated video analysis software.
- Video collection is not uniform across space, time, and quality, which results in video/image datasets that are sparse, non-uniform, and unevenly distributed, and makes it difficult to extract general trends or patterns. Specific challenges include the following:
 - Coverage areas for roadway cameras vary, and existing camera views do not always provide complete coverage for all parts of the highway;
 - Equipment failures (e.g., of field cameras, communications networks, and recording systems) can increase the lack of coverage, particularly if maintenance to the cameras is not performed in a timely manner;
 - Weather conditions like snow and rain can degrade the quality of the video collected, in some cases making it impossible to extract metadata; and
 - Video container and compression standards vary widely across equipment types and manufacturers. These standards often are proprietary, with the result that the video cannot be converted easily to a common standard without losing some video data integrity.
- SSP data collected from paper forms or by radio communication and entered into spreadsheets or simple applications often lacks precise location data and can be of lower quality due to the inability to correct for misspelled words, non-existent categories, non-standardized abbreviations, and custom narratives. Complex analysis often is needed to correct or standardize lower quality or “fuzzy” content, and even with additional complex analysis, the resulting content may lack information precision and be less valuable. The current management of SSP data files (except for database systems) also may lead to difficulty ingesting and analyzing content. Spreadsheet files, for example, are often manually collected and stored in shared network folders. As data file formats evolve and improve (e.g., by adding new columns or refining the category names used to describe service patrol responses), the formats in the newer spreadsheet files can quickly cease to match the formats of previously created files. Unless a serious, sustained effort is made to routinely and continuously update all prior files, the content across files quickly becomes non-uniform and difficult to analyze without cleaning. In some cases, it can be impossible to retrofit older data files to match a newer data format because the historical data lacks the precision required by the new format.
- Environmental sensor stations (ESS) need to be monitored and maintained to counter sensor failure and sensor drift. Gaps in monitoring and maintenance can lead to some data quality issues (e.g., missing or erroneous data). To circumvent this problem, data aggregators perform quality checks and more advanced data verification and corrections on data made available

through the associated systems, such as NOAA’s Meteorological Assimilation Data Ingest System (MADIS) and the FHWA Weather Data Environment (WxDE).

- The nation’s 511 systems are designed to quickly broadcast traffic and transit event information to travelers, but they are not designed to store that data or even structure and organize it for later retrieval or searches. For analysis over time, the 511 data would need to be stored on a different system. Some data elements such as location, timestamps, and 511 event type, lend themselves easily to analysis, but data elements containing free text, such as event descriptions, are more challenging to mine and organize. These more challenging data elements will require more advanced text analysis to extract valuable keywords and topics essential to further analysis.
- Toll data may be difficult to obtain, both because of the sensitivity of the data and because of the possibility of private party ownership. The data structure is simple, and toll data should be easily reusable for Big Data analysis; however, automatic detection of vehicles at toll gates is known to be error prone, particularly when using ALPR (Laroca et al. 2018). Although data quality may be an issue when performing data analysis that requires vehicle identification (e.g., toll calculation or speed checking), TIM data analysis may not require identification of vehicles and therefore may not be affected by this issue.

The next set of tables (Tables 5-9 through 5-14) show the readiness of each data source. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the transportation data sources can be found in Appendix A, Tables A-7 through A-12.

Table 5-9. Traffic sensor data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage			√	√
Integration		√	√	
Relevance and Sufficiency			√	
Quality		√	√	
Collection Frequency		√	√	
Granularity		√	√	
History		√	√	
Privacy	N/A			
Documentation		√	√	



Table 5-10. Traffic video data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility	√	√		
Storage			√	√
Integration	√	√		
Relevance and Sufficiency	√	√		
Quality		√	√	
Collection Frequency	√	√		
Granularity	√	√		
History	√	√		
Privacy		√		
Documentation		√		



Table 5-11. SSP/IRP data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage			√	√
Integration	√	√	√	
Relevance and Sufficiency			√	
Quality		√		
Collection Frequency			√	√
Granularity			√	√
History			√	
Privacy			√	
Documentation	√	√	√	

**Table 5-12. Road weather data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage			√	√
Integration		√	√	
Relevance and Sufficiency		√	√	
Quality		√	√	
Collection Frequency		√	√	
Granularity		√	√	
History		√	√	
Privacy	N/A			
Documentation		√	√	

**Table 5-13. 511 system data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage		√	√	
Integration		√	√	
Relevance and Sufficiency	√	√		
Quality	√	√		
Collection Frequency		√	√	
Granularity		√	√	√
History		√	√	
Privacy		√		
Documentation		√	√	



Table 5-14. Toll data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	
Collection Frequency				√
Granularity				√
History			√	
Privacy		√		
Documentation		√		



5.2.3 Public Safety Data

5.2.3.1 Description of Sources

As the primary point of contact for the public via the 911 calling system, and with their recognizable role as “first responders,” public safety agencies are a critical part of TIM and generate valuable data. Public safety agencies generally are recognized to be law enforcement, fire and rescue, and emergency medical services (EMS). Although they are private enterprises, towing companies are authorized agents of law enforcement agencies. As such, towing is an important ally of law enforcement. Moreover, as the primary point of contact for the public via 911, and recognizable role as “first responders,” public safety agencies are a critical part of TIM and generate valuable data.

Within the public safety data domain, the research team assessed the following four data sources:

- **Law enforcement, fire and rescue, and EMS CAD system data:** CAD is a suite of software used to initiate public safety calls for service, to dispatch responders, and to facilitate and maintain communications and the status of responders in the field. CAD functions include the following:
 - Personnel log on/log off (with timestamps);
 - Incident generation and archiving, including generation of incident case numbers;
 - Assignment of field personnel to incidents;
 - Logging of updates; and
 - Timestamping for every action taken by the dispatcher.
- **Emergency communications center (ECC)/911 call center/public safety answering point (PSAP) data:** Data collected at ECCs via CAD systems is like the data collected by law enforcement and fire and rescue CAD systems, and many ECCs are even housed by state police or transportation management centers.
- **Digital video data:** Public safety agencies use various types of digital video technologies, including CCTV, ALPR, dashboard cameras, and wearable cameras. ALPR is used to capture license plate numbers and compare them to one or more databases of vehicles of interest and alert authorities when a vehicle of interest has been observed. Dashboard cameras and/or wearable cameras are used to monitor traffic stops and other enforcement activities. Basic dashboard cameras are video cameras with built-in or removable storage media that constantly record. More advanced dashboard cameras can have audio recording, GPS logging, speed sensors, accelerometers, and uninterrupted power supply capabilities. Body cameras vary and range from small, low-resolution models to high-definition models.

- **Towing and recovery data:** Towing and recovery data includes a catalog of calls for service and various timestamps associated with the response.

5.2.3.2 Summary of Findings

Data from public safety agencies represents information collected by and from a significant number of incident responders—particularly for incidents that require an official report or documentation by statute, for insurance company purposes, or in case of potential litigation. Public safety data is collected, owned, and managed by tens of thousands of public safety agencies across the United States. Many incidents begin with a call to the 911 system, which is operated in the public safety arena.

Because public safety agencies are TIM partners and almost always place responders on the scene of traffic crashes (as well as many non-crash traffic incidents), they provide very complete coverage of data collection for incidents, offering good potential for analytics. Public safety agencies typically use CAD to record information about the activities of employees, as well as the associated times of these activities. CAD systems can therefore be useful sources for timestamp data, particularly the time of first awareness of an incident, as well as the times of response, arrival, and departure of responders from incident scenes.

Because towing companies are important partners in TIM, their participation in quality data collection and efficient data exchange also could contribute significantly to improved TIM through data analytics, particularly in knowing which vehicles are on scene and at what times.

ECCs are an overlooked national resource that could provide critical information to the many public safety, public service, and homeland security disciplines that seek real-time information. According to an Association of Public-Safety Communications Officials (APCO) international report, “There is no better information set for real-time situational awareness for public safety than that found in ECC CAD systems” (APCO International 2010).

Electronic reporting and the use of technologies like in-vehicle computers and AVL have streamlined the collection and transmission of data from an incident scene back to ECC/CAD systems, although voice communications via radio remain a primary method of data collection/transmission in many jurisdictions. Advances in vehicle-mounted and wearable camera systems are creating new sets of data that hold potential for TIM and analytics.

Because public safety data is in the public domain, the data that can be shared can usually be shared at little to no cost; however, public records laws and privacy issues associated with sensitive information and PII create barriers to sharing the data. Because the data created by public safety agencies typically is not owned or managed by transportation agencies, transportation agencies have little control over obtaining and using the data. Other challenges and limitations associated with leveraging public safety data for Big Data analytics for TIM include the following:

- Some prominent CAD standards from a national organization are being implemented, but there is no national standard or regulatory authority. Consequently, among the 6,000+ PSAPs nationwide, only a few have implemented standards that enable operational or data analytics assessments. For example, *10 codes*—brevity codes used in voice communications (e.g., “10-4,” meaning “affirmative” or “OK”)—can vary from agency to agency. Missing or incomplete/low-quality records (e.g., record of the arrival of a responder on the scene but no record of departure) are not uncommon. These factors render the integration and analysis of CAD/PSAP data more challenging and costly. In addition, institutional and legal barriers limit agencies’ ability to tap these data sources in some locations.
- CAD data is recorded using an event database format (i.e., each row is an event that combines a single action, such as “responder arrived” or “responder departed,” with a single timestamp).

This organization can be ideal for data collection, but it can complicate data extraction and analysis because the data typically sought after may be distributed across more than one record (time on scene, number of responders on the scene).

- Partial or redacted datasets often are publicly available, but the additional analytical value that will be found in complete datasets may be difficult to access; access to the full dataset may be challenging due to local and state laws and restrictions.
- Many individual towing companies still do not maintain any data, and some maintain only limited data using paper logs or spreadsheets. In-house systems rarely go outside of the business. Ultimately, however, the biggest obstacle to acquiring towing and recovery data is the intellectual property and competitive value that it holds for the business owner. Cloud-based towing management software leverages the capabilities of mobile devices. Such cloud-based applications hold the potential to greatly increase the amount and quality of data that can be collected by towing companies by offering low-cost ways to manage towing businesses, even for small companies. The downside is that the applications are designed around the needs of private businesses (i.e., insurance companies), the data is private and of competitive value, and accessing it has proven prohibitively expensive.

Despite these challenges, agencies that have been able to integrate CAD data with transportation data at the TMC level have realized improved datasets. Incorporating additional data elements that typically are not included in transportation datasets (e.g., times of responders arriving at and leaving event scenes and the presence and types of injuries, if any) could provide new insights, such as how response times and injuries impact incident clearance.

The next set of tables (Tables 5-15 through 5-18) show the research teams' evaluation of the readiness of each data source. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the public safety data sources can be found in Appendix A, Tables A-13 through A-16.

5.2.4 Crowdsourced Data

5.2.4.1 Description of Sources

In transportation, the collection and use of crowdsourced data is becoming both more feasible and more useful. Typical crowdsourced data used by transportation agencies includes data from:

- Social media platforms (e.g., Twitter, Waze) in which data is collected automatically in the course of consumers' use of the apps.

Table 5-15. Law enforcement, fire and rescue, and EMS CAD system data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage				√
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	
Collection Frequency				√
Granularity				√
History			√	
Privacy			√	
Documentation		√	√	



Table 5-16. ECC/911 call center/PSAP data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage				√
Integration			√	√
Relevance and Sufficiency			√	
Quality		√	√	
Collection Frequency				√
Granularity			√	
History				√
Privacy			√	
Documentation			√	

**Table 5-17. Digital video data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	
Integration				√
Relevance and Sufficiency			√	
Quality		√	√	
Collection Frequency				√
Granularity		√	√	
History		√	√	
Privacy			√	
Documentation		√	√	

**Table 5-18. Towing and recovery data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	√
Storage		√	√	√
Integration		√	√	
Relevance and Sufficiency		√	√	
Quality		√		
Collection Frequency			√	
Granularity				√
History		√	√	
Privacy		√	√	
Documentation		√		



- Third-party vehicle probe data providers (e.g., HERE Technologies, INRIX) in which anonymous GPS-based data is automatically collected from vehicle fleets, consumer smart phones, and others including road sensors and toll tags.

Specially developed mobile apps (e.g., Utah DOT Citizen Reporting app) in which agencies enlist a digital community to provide specific information—such as traffic conditions, crashes, or road weather issues—from geographic areas not readily accessible to the agency.

The research team selected and assessed two widely used apps—Waze and Twitter—as being most relevant currently for generating data for TIM analysis. Note: Data other than vehicle probe data is collected and aggregated by HERE Technologies and INRIX, and is addressed in this chapter under “Aggregated Datasets.”

- **Waze data:** Data generated by users of the Waze community-based navigation mobile app includes real-time road information data, such as crashes, construction, police presence, road hazards, and traffic jams, along with confirmation of this information by other Waze users through “thumbs-up” or “thumbs-down” responses or through detailed messages. Additionally, Waze automatically records the speed at which users’ vehicles travel on the roadways.
- **Twitter data:** Data generated by users of the Twitter app includes the text of each tweet (up to a 144-character stream), an associated timestamp, and possible attachments (e.g., photos or videos). When users allow Twitter to share their locations, tweet locations (latitude, longitude) also are captured.

5.2.4.2 Summary of Findings

Crowdsourced data generally is collected, managed, and owned by private vendors (usually the companies that own the applications), although transportation agencies also are creating apps to directly collect crowdsourced data specific to their needs. The type of data that is collected automatically through devices (e.g., cell phones, navigation systems, or Bluetooth devices) consists largely of location data, as well as vehicle speeds and travel times. This data can be used to determine the locations of slow or stopped traffic, the locations of traffic incidents, and even the location of the back of the queue associated with a particular incident.

Crowdsourced and social media data collected via user input into mobile applications typically consists of a small amount of text, feedback to pre-established questions, validation of existing information, a rating of information published by another user, or even corrections to a map. Crowdsourced and social media data from user input can be used to assess crowd sentiments (e.g., through content analysis of Tweets), as well as the occurrence of traffic incidents and incident details (e.g., through Waze data). For TIM, crowdsourced and social media location data can be very valuable (e.g., to identify in real time the location of an incident). Location information is collected automatically on Waze but is optional on Twitter.

Crowdsourced and social media data offer many advantages. It can be collected anytime, anywhere; it does not require a costly physical infrastructure for data collection (e.g., sensors); and it offers the potential of near ubiquitous coverage, depending on the penetration of probes and/or the app user base. Some state transportation agencies are already testing and using crowd-sourced data for improving TIM, particularly for early detection (even before 911 calls) and verification of incidents. These datasets also could provide incident details, as well as data from rural and remote areas.

Twitter data is free to access and analyze, and the Waze Connected Citizen Program (CCP) is a partnership that allows the sharing of specific Waze data with public agencies for free. Waze does not, however, share or sell its raw data for use or analysis.

Open questions remain about the use of crowdsourced and social media data. Some of these questions, posed in the 2015 BDE and ERTICO-ITS Europe workshop report on *Smart, Green, and Integrated Transport*, can be paraphrased as follows:

- How can data users best decide which crowdsourced and social media data is reliable and to what degree?
- To what degree does the elective publication of social media sources give data users open rights to use information obtained from these sources?
- How can data sources or data users identify and prevent spoof outflows from malicious users?
- How can data sources and data users encourage beneficial services while discouraging inappropriate use of social media apps?

Depending on the dataset and the location, challenges and limitations associated with leveraging crowdsourced and social media data for Big Data analytics for TIM can include the following:

- Crowdsourced and social media data requiring user input can lack quality (e.g., app users click the wrong button, inaccurate perceptions lead to inaccurate descriptions of what is happening or what is reported).
- Free text is subject to errors (e.g., misspellings).
- The data reliability can vary tremendously by location, time, and service; therefore, its use can complicate analysis.
- Waze data-sharing policies do not allow users to fully access and exploit the data.
- Multiple rural states (e.g., Montana, Wyoming) have noted a lower usage of Waze, which results in less data and potentially less reliable data.
- Waze provides a reliability/confidence index with alert reports; however, these indices may not be of sufficient quality to satisfy the needs of transportation agencies.
- Challenges are associated with understanding which data analysts should use (i.e., without access to the raw data, users must rely on what Waze has extracted and shared, with the result that analyses may lack clarity on the accuracy of an event).
- The volume of streaming data necessary to monitor incidents can be challenging. (The phrase, “drinking from the fire hose” comes to mind). For example, to monitor for TIM relevant information or events by processing a live Twitter stream, the text of each tweet needs to be parsed, analyzed using text mining, correlated with similar tweets, and counted to establish the location and veracity of a detected event. This process is difficult to achieve in real time, particularly considering the number of irrelevant tweets, the possibility of not having enough relevant tweets, and the use of differing vocabulary to describe the same event. To provide accurate analysis, it is both challenging and important to have a lot of verified data (e.g., tweets that are and are not connected to the incidents).
- Lack of location data from crowdsources may make it difficult to leverage the content, especially in real-time analysis situations. For example, the International Transport Forum (2015) estimated the number of tweets that are geolocated at only 1 percent. This lack of location data can make it difficult to use tweets to detect the occurrence of roadway events such as incidents or free flow recovery.
- Twitter uses hashtags to qualify and categorize the free text content of tweets. Twitter users can create hashtags and use them within their messages, but the platform imposes no controls over how hashtags are formatted and used. Although some simple hashtags (e.g., “#accident”) exist, they are too general to allow tweets to be filtered to extract relevant TIM content.

Tables 5-19 and 5-20 show the research team’s assessment of the readiness of Waze and Twitter data, respectively. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the crowdsourced/social media data sources can be found in Appendix A, Tables A-17 and A-18.

Table 5-19. Waze data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				✓
Storage				✓
Integration				✓
Relevance and Sufficiency				✓
Quality			✓	
Collection Frequency				✓
Granularity				✓
History				✓
Privacy			✓	
Documentation				✓

**Table 5-20. Twitter data readiness.**

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				✓
Storage				✓
Integration				✓
Relevance and Sufficiency			✓	
Quality				✓
Collection Frequency				✓
Granularity			✓	
History				✓
Privacy			✓	
Documentation				✓



5.2.5 Advanced Vehicle Systems Data

5.2.5.1 Description of Sources

Advanced vehicle systems are the norm in modern automobile manufacturing. These systems record, share, and ingest information in a variety of ways, for a variety of purposes. Within the advanced vehicle systems data domain, four data sources were assessed:

- **Automated vehicle location (AVL) system data:** AVL is a means for automatically determining and transmitting the geographic location of a vehicle. AVL is used to manage vehicle fleets, such as service vehicles, public transportation vehicles, emergency vehicles, and commercial vehicles. AVL data includes real-time temporal and geospatial data (polled every few seconds), as well as vehicle logs (e.g., vehicle number, operator ID, route, direction, and arrival and departure times).
- **Event data recorder (EDR) data:** An EDR is a digital recording device that records data associated with a vehicle before and during a crash. As of 2006, an estimated 92 percent of new passenger vehicles had EDRs. In 2013 and later models, EDRs are required to record specific data in a standard format to make retrieving the information easier. A NHTSA regulation passed in 2012 provides that if a vehicle has an EDR, it must track 15 specific data elements, including speed, steering, braking, acceleration, seatbelt use, and—in the event of a crash—force of impact and whether airbags are deployed.
- **Vehicle telematics system data:** *Telematics* refers to the transfer of data to and from a vehicle. Vehicle telematics systems combine a GPS system with on-board sensors and diagnostics to

record speed, engine throttle, braking, ignition cycle, whether the driver was using a safety belt, airbag deployment, and the physics of crash events, including crash speed, change in forward crash speed, maximum change in forward crash speed, time from beginning of the crash event at which the maximum change in forward crash speed occurs, the number of crash events, the time between crash events, and whether the device completed recording. Unlike EDRs, which collect and store a few seconds of data immediately before and after a crash event, telematics systems continuously record all types of second-by-second data about vehicles and driver behavior, sometimes for years at a time. Telematic technologies collect raw vehicle data and overlay this information with GIS mapping data (e.g., road type and speed limits). The data is then “broadcast” via data links like Wi-Fi, GPS, Bluetooth, 3-axis accelerometers, and mobile broadband communications to auto manufacturers, fleet owners, and insurance companies (Klieman & Lyons 2014).

- **Automated and connected vehicle, connected traveler, and connected infrastructure data:** Automated vehicles are those in which at least some aspect of a safety-critical control function (e.g., steering, throttle, or braking) occurs without direct driver input. Connected vehicles are vehicles that use communication technologies to communicate with the driver, other vehicles on the road (V2V), roadside infrastructure (V2I), and the cloud (V2C) (Center for Advanced Automotive Technology n.d.). Automated and connected vehicle data is collected via micro-processors and dozens of sensors, including telematics and driver behavior data-collection systems on board the vehicles. Forward and side radar sensors, sonar, GPS, LIDAR, cameras, and monitoring systems will generate increasing amounts of data as connected and automated vehicles become more prevalent. The data is captured and recorded by the system and stored in on-board or cloud-based systems. A connected traveler is one that uses a mobile device that generates and transmits status data, including the traveler’s location, trip characteristics (e.g., speed), and mode and status (e.g., riding in a car, riding on transit, walking, biking) (Gettman et al. 2017). Connected infrastructure includes traditional ITS devices, such as traffic signals, ramp meters, CCTV, and RWIS and may eventually evolve to include standard Internet-of-Things (IoT) protocols as IoT technologies continue to mature (Gettman et al. 2017).

5.2.5.2 Summary of Findings

Vehicle technology has evolved in recent decades to encompass the monitoring and collection of data inside and outside the vehicle. AVL systems, which track the location of fleet-equipped vehicles, have the potential to benefit closest-unit dispatch and optimized-route assignment to incident scenes, and indicate which vehicles are on scene and at what times. Detailed raw temporal and spatial AVL data must be uploaded from the on-board computer to the central computer. Although older systems require manual intervention to upload the data, newer systems usually include an automatic high-speed communication device through which data is uploaded daily (e.g., when vehicles are fueled).

EDRs function like the black boxes used in aircraft in that they record a variety of information about the systems and operations of an individual vehicle. The data is contained within the EDR, and it must be downloaded with a specialized data-retrieval toolkit.

The use of onboard systems that automatically collect and communicate data to and from vehicles generally is termed *telematics*. Automated and connected vehicle technologies are the intelligent use of the information exchanged between the vehicle and the roadway or between multiple vehicles.

As potential sources of data, overlap certainly occurs among the systems in the advanced vehicle domain. EDR data has the potential to help with the understanding of the relationship between the vehicle, driver, and environment, the trilogy of crash causation. Telematics data holds greater promise for TIM applications, as telematics data is continuously recorded

over long periods of time and can be communicated in real time. In addition, as the cost of enabling mobile broadband communications has fallen, more automakers have been embedding telematics in vehicles. An estimated 70 percent of vehicles built since 2011 include some form of telematics system (Klieman & Lyons 2014).

In 2015, the International Transport Forum (ITF) concluded that safety improvements can be accelerated through the specification and harmonization of a limited set of safety-related vehicle data elements (International Transport Forum 2015). Specifically, the multinational organization concluded that technologies such as EDRs can provide post-crash data well suited for improving emergency services and forensic investigations, and if this vehicle-related data is shared in a common format, it could be used to enhance road safety. The ITF recommends that further work be pursued to identify a core set of safety-related data elements to be publicly shared and to ensure the encryption protocols necessary to secure data that could compromise privacy (International Transport Forum 2015).

Beyond vehicle-mounted EDRs and communication of vehicle and driver data via telematics, the fields of automated and connected vehicle technologies are largely emerging as data sources. Automated and connected technologies use cloud services to share information, and these data hold promise to be a good source of data for Big Data analytics.

Challenges and limitations associated with leveraging advanced vehicles systems data for Big Data analytics for TIM include the following:

- Older AVL systems rely on manual procedures to extract data (e.g., exchanging data cards or attaching an upload device), which adds a logistical complication to obtaining the data.
- AVL data typically is stored by the fleet owner and is rarely shared outside of the organization. AVL data accessibility for real-time analysis beyond the owner agency is currently limited, and the cost of obtaining this type of data is unknown.
- Although almost all vehicles now have some form of EDR, the current technology for data collection and storage, in conjunction with data privacy issues, limits the ability to aggregate and use EDR data. The use of telematics data, particularly the aggregation of the data, presents similar privacy challenges for consumers, the courts, law enforcement, automakers, insurers, and the telematics industry. Specific state laws and regulations vary, but EDR data is generally considered to belong to the vehicle owner, which means the owner's consent typically is required before the data can be obtained and used. In the absence of such consent, the data can only be obtained through a court order. Data ownership and privacy issues concerning automated and connected vehicle data are critical and largely unresolved issues.
- Each automaker and insurer uses a proprietary telemetry or usage-based insurance (UBI) program, which further impedes data sharing.
- Nelson (2016) has reported that autonomous vehicles are expected to generate and consume roughly 40 TB of data per vehicle for every 8 hours of driving, which creates challenges for data storage, management, and analysis.

On-board telematics devices that use the driver's mobile phone—examples include SnapShot® from Progressive insurance and the Automatic dashboard adapter and app by Automatic Labs™—collect some of the data collected by vehicles' EDRs and on-board sensors and stream it to large data stores where the data is analyzed. In the case of SnapShot® and similar applications available from other insurance companies, the primary function of the analysis is used to optimize the insurer's risks. In the case of Automatic, the adapter and smart-phone app work in combination to connect user-subscribers to a suite of services. These third-party devices require that a user agreement be signed by the primary owner or driver allowing the third party to collect and use the vehicle data, effectively circumventing the data privacy issue. The datasets created by such third parties may provide agencies an alternative way to access EDR/telematics

data, either partially or fully, without having to collect it one vehicle at a time. Similarly, telematics system user agreements may allow for the data to be reused or sold to entities other than the telematics system owner and/or the driver.

The next set of tables (Tables 5-21 through 5-24) show the research team's assessment of the readiness of the advanced vehicle systems data sources. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the advanced vehicle systems data sources can be found in Appendix A, Tables A-19 through A-22.

Table 5-21. AVL data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	√
Quality			√	
Collection Frequency				√
Granularity				√
History			√	
Privacy			√	
Documentation			√	



Table 5-22. EDR data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√		
Storage			√	
Integration		√		
Relevance and Sufficiency			√	
Quality			√	
Collection Frequency				√
Granularity				√
History		√		
Privacy			√	
Documentation		√	√	



Table 5-23. Vehicle telematics data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage			√	√
Integration			√	√
Relevance and Sufficiency			√	
Quality			√	
Collection Frequency				√
Granularity				√
History			√	
Privacy			√	
Documentation			√	



Table 5-24. Advanced and connected vehicle, traveler, and infrastructure data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				√
Storage				√
Integration				√
Relevance and Sufficiency				√
Quality			√	√
Collection Frequency				√
Granularity				√
History			√	√
Privacy			√	
Documentation			√	√



5.2.6 Aggregated Datasets

5.2.6.1 Description of Sources

Aggregated datasets are created when a source collects (aggregates) data that has originated from other sources for the purposes of adding value to the data. Within the aggregated datasets domain, the following data sources were assessed:

- **Regional Integrated Transportation Information System (RITIS):** RITIS is an automated data sharing, dissemination, and archiving system that was developed by and is maintained by the University of Maryland Center for Advanced Transportation Technology Laboratory (CATT Lab). RITIS data includes, but is not limited to, third-party probe data, DOT ATMS data, road weather data, virtual weigh station data, transit data, and parking spaces available. Not all types of data are available from all the locations providing data.
- **National Performance Management Research Data Set (NPMRDS):** Accessible to agencies with RITIS accounts, the FHWA's NPMRDS provides vehicle probe-based travel time data for passenger automobiles and trucks. The real-time probe data is collected from a variety of sources that include mobile devices, connected vehicles, portable navigation devices, and commercial fleets and sensors. The dataset includes historical average travel times in 5-minute increments daily covering the National Highway System (NHS).
- **Meteorological Assimilation Data Ingest System (MADIS) and MADIS Meteorological Surface Integrated Mesonet, from NOAA:** A meteorological observational database and data delivery system, MADIS runs operationally at the National Weather Service (NWS) National Centers for Environmental Prediction (NCEP) Central Operations. MADIS subscribers have access to an integrated, reliable, and easy-to-use database containing real-time and archived observational datasets. Also available are real-time gridded surface analyses. The surface analyses grids assimilate all the MADIS surface datasets, including the high-density Meteorological Surface Integrated Mesonet data. The MADIS Integrated Mesonet is a unique collection of thousands of mesonet stations from local, state, and federal agencies and private firms that help provide a finer density, higher frequency observational database for use by the greater meteorological community. The numerous data elements include atmospheric conditions (e.g., temperature, wind, precipitation, and pressure), visibility, nearby storms, and sunrise and sunset (NOAA 2016).
- **Third-party web service weather data:** Weather data available from web-based third-party data-as-a-service (DaaS) providers includes historical and forecast meteorological data and weather forecast data from various public and private weather data sources across the globe. Data elements include temperature, wind, precipitation probability, pressure, visibility, wind

speed, wind direction, cloud cover, visibility index, humidity, and other weather details, as well as ancillary data elements such as nearby storms, moon phase, sunrise, and sunset derived from multiple national and international meteorological data sources.

- **National Fire Incident Reporting System (NFIRS) data:** NFIRS is the standard national reporting system used by U.S. fire departments to report fires and other incidents to which they respond and to maintain records of these incidents in a uniform manner. Updated annually, NFIRS is the world's largest national database of fire incident information.
- **National Emergency Medical Services Information System (NEMSIS) data:** NEMSIS is a national repository of standardized EMS data elements from 49 states and 2 territories. Incident response data is collected by individual EMS agencies using NEMSIS-compliant software that electronically transmits the data to a state database. A subset of the data is then electronically transmitted from the agency databases to the national NEMSIS repository.
- **Motor Carrier Management Information System (MCMIS) data:** MCMIS is a computerized system whereby the FMCSA maintains a comprehensive record of the safety performance of the commercial motor carriers that are subject to the Federal Motor Carrier Safety Regulations (FMCSR) or Hazardous Materials Regulations (HMR). The data includes data elements on registration, crashes, inspections, and reviews.
- **HERE data:** HERE Technologies aggregates and analyzes traffic data from a broad range of sources, including “the world's largest compilation of both commercial and consumer probe data, the world's largest fixed proprietary sensor network, publicly available event-based data, and billions of historical traffic records” (Younas 2013). HERE Technologies also combines “20 billion real-time GPS probe points a month with historical information and search queries to learn where people are travelling and what the conditions are like” (Bonetti 2013; Younas 2013). The company asserts that almost half of all the data is less than 1 minute old, and more than three-quarters is less than 5 minutes old (Bonetti 2013). The data is provided to customer agencies through software-as-a-service (SaaS) and DaaS solutions.
- **INRIX data:** INRIX collects massive amounts of information about roadway speeds and vehicle counts from over 300 million real-time anonymous mobile phones, connected cars, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices. This data is enriched with event data such as traffic incidents, weather forecasts, special events, school schedules, parking occupancy, road construction, and more. INRIX provides the data to its customers through SaaS and DaaS solutions.

5.2.6.2 Summary of Findings

Aggregated datasets can be analyzed and compared across numerous geographic expanses and/or agencies. Some aggregated datasets could potentially function as “one-stop shops” for many types of data, assuming the data can be broadly accessed or downloaded and merged with other datasets. The data from some aggregated datasets may be available for download. In other cases, the proprietary nature of the data may mean it is not available for download. These cases usually involve private-sector or third-party companies that have built a data lake with valuable information. Such companies may offer a limited set of data services but not make their data available for download.

The cost of obtaining aggregated datasets varies greatly. Public release datasets (e.g., datasets from MADIS, NFIRS, NEMSIS, or MCMIS) may be available for free. Data from the NPMRDS is shared for free with state transportation agencies and MPOs, but it is not made available to other organizations or entities. Customized extracts from datasets like MADIS or MCMIS may be obtained at minimal costs. Pay-as-you-go solutions like third-party weather data and some other DaaS solutions can be relatively inexpensive. Finally, expensive, data purchasing options are available from private-sector data aggregators.

This section summarizes the research team’s assessment of the various aggregated datasets that have been described. For ease of reading, the summaries have been grouped as follows:

- RITIS and NPMRDS datasets,
- Weather datasets,
- Standardized public safety datasets,
- MCMIS dataset, and
- Private data aggregator datasets.

RITIS and NPMRDS datasets. RITIS was developed and is maintained by the University of Maryland Center for Advanced Transportation Technology Laboratory (CATT Lab). RITIS collects data from states, cities, and private companies on either a one-time basis (with limited geography and temporal coverage) or, for some data sources, on a recurring basis. RITIS also is the portal through which account holders can access the NPMRDS dataset, which was commissioned by FHWA and currently is provided by INRIX. Although RITIS provides advanced analysis and visualization tools using the data, challenges and limitations associated with leveraging RITIS and NPMRDS data for Big Data analytics for TIM include the following:

- RITIS data is made available only to certain types of users (e.g., individuals working at federal, state, or local transportation agencies or MPOs, members of law enforcement, public safety, or military agencies, and researchers or consultants “working on projects for a government partner”), which restricts broad-based access by private companies, contractors, and universities (CATT Lab 2015);
- RITIS data-sharing policies do not allow registered users to fully access and exploit the data;
- Although RITIS contains data from a wide array of data sources, no public documentation is provided as to what data sources are available from what locations and what data elements are included in the various data sources; and
- RITIS does not provide information or metrics about data availability, quality, and usability (with the exception of the NPMRDS data obtained through the NPMRDS Coverage Map).

Tables 5-25 and 5-26 show the research team’s assessment of the readiness of the RITIS and NPMRDS datasets, respectively. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the RITIS and NPMRDS datasets can be found in Appendix A, Table A-23 and Table A-24, respectively.

Weather datasets. RWIS are typical in transportation agencies, but an abundance of public, private, and non-profit organizations also collect, aggregate, and share weather data.

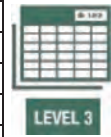
Table 5-25. RITIS data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility		√	√	
Storage			√	√
Integration			√	
Relevance and Sufficiency		√	√	
Quality		√	√	
Collection Frequency		√	√	
Granularity		√	√	√
History			√	√
Privacy	Unknown/not documented			
Documentation	Unknown/not documented			



Table 5-26. NPMRDS data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	
Storage			√	
Integration			√	
Relevance and Sufficiency			√	
Quality			√	√
Collection Frequency				√
Granularity			√	
History				√
Privacy	N/A			
Documentation	Not accessible			



Most notable is NOAA, which operates various weather databases (e.g., MADIS and the MADIS Integrated Mesonet). Many states share their RWIS data with these data systems. Weather data from federal or state agencies typically is offered at no charge, and even third-party aggregators regularly and frequently offer large amounts of data to users at very low costs. Big Data opportunities for TIM include the ability to determine more precisely the historical impacts of weather, environmental, and surface conditions on the cause of crashes, as well as the impacts of these conditions on incident clearance. The analysis of real-time weather and road weather data, including integrated forecast data, can help agencies better plan and execute incident response, clearance, and recovery as these activities unfold. Challenges and limitations associated with leveraging aggregated weather datasets for Big Data analytics for TIM include the following:

- The data in some of the datasets (e.g., MADIS) can become very messy in terms of format, content, and quality because of the diverse organizations that contribute to the dataset.
- Specific to MADIS, the NetCDF file format could be challenging to use for non-scientific staff because it requires the implementation of a dedicated API to access the data. NetCDF is used typically in scientific applications such as meteorological forecasting, not Big Data analysis. NetCDF is not a Big Data–friendly format and requires that the data be transformed into a simpler format to be processed.

Private third-party weather data aggregators have begun to overcome some of these challenges by making the NOAA datasets easier to use, enriching the data with other data sources, and providing cost-effective web services/DaaS solutions at scale. Although the data from these third-party services cannot be downloaded in bulk (like it can from the NOAA databases), with a time and location for incidents, very detailed weather, environmental, and surface conditions for millions of incidents can be requested all at once (historically and in real time) and at a very low cost.

Tables 5-27 and 5-28 show the research team’s assessment of the readiness of the MADIS and third-party web services datasets, respectively. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the MADIS and third-party web service datasets can be found in Appendix A, Table A-25 and Table A-26, respectively.

Standardized public safety datasets. Standardized, aggregated, national datasets in the fire and EMS disciplines—specifically, NFIRS and NEMSIS—offer excellent models for standardization and aggregation of incident data collected at the local level and fed through state-level databases to national-level databases. Nevertheless, challenges and limitations associated with

Table 5-27. MADIS data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	√
Storage				√
Integration				√
Relevance and Sufficiency				√
Quality				√
Collection Frequency				√
Granularity				√
History				√
Privacy			√	
Documentation				√



leveraging these datasets for Big Data analytics for TIM remain. These challenges and limitations include the following:

- The NFIRS distributed dataset is not a complete dataset. It only contains fire and hazardous condition incidents (USFA 2017). The truncation of the dataset appears to be due to current data-size limitations in the storage and distribution system. These limitations are rather uncommon these days and denote either an obsolete system or obsolete data management practices, as the sharing of multi gigabyte files is now a commonplace occurrence.
- The NFIRS public data release files are published using the Dbase database file format (.dbf). Created in 1978 to be used with the MS-DOS operating system, this format is still common today on desktop-based database software, but it has had many iterations and variations. To be read, Dbase files require software capable of parsing the format's binary structure, which adds additional preparation work before the stored data can be exploited by typical Big Data tools. Alternative, Big Data-friendly formats (e.g., JSON, XML, TXT, or CSV) should be used instead, and many datasets that can be generated as .dbf files also can be generated using these formats.
- The U.S. Fire Administration (USFA) does not have a quality assurance system in place to check for codes that are not in the current data dictionary. As a result, the NFIRS public data release files contain invalid codes and may exhibit data inconsistencies that violate published documentation (FEMA 2011). In addition, because the NFIRS data is collected on a voluntary basis, sufficient data may not be available from some areas.

Table 5-28. Third-party web service weather data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				√
Storage				√
Integration				√
Relevance and Sufficiency			√	√
Quality				√
Collection Frequency				√
Granularity			√	√
History				√
Privacy			√	
Documentation				√



Table 5-29. NFIRS data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	
Storage			√	
Integration			√	
Relevance and Sufficiency			√	
Quality		√		
Collection Frequency			√	
Granularity			√	
History			√	
Privacy			√	
Documentation		√		



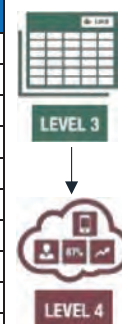
- The NEMSIS location data at the national level is limited to the zip code level, which could greatly limit data analytics, as this level of resolution would be too low for meaningful analysis. Data would need to be drawn from the local level, which significantly increases the effort needed to use the data for Big Data analyses of TIM.
- Aggregation of NEMSIS data due to data sensitivities limits the ability of users to fully access and exploit the data.

Tables 5-29 and 5-30 show the research team's assessment of the readiness of the NFIRS and NEMSIS datasets, respectively. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the NFIRS and NEMSIS datasets can be found in Appendix A, Tables A-27 and A-28.

MCMIS dataset. MCMIS contains information on the safety fitness of commercial motor carriers (trucks and buses) and hazardous material shippers. MCMIS data includes registration information for all motor carriers (e.g., U.S. DOT number, company name, address, contacts, number of vehicles, number of drivers, and other registration information); crash data for each commercial motor vehicle involved in a crash (e.g., U.S. DOT number, report number, crash date, severity of the crash [tow-away, injury, fatal] and vehicle data); data on roadside inspections conducted on motor carriers (e.g., U.S. DOT number, report number, inspection date, state, and vehicle and equipment information, and violations-related data); and information on reviews

Table 5-30. NEMSIS data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				√
Storage				√
Integration				√
Relevance and Sufficiency			√	
Quality				√
Collection Frequency				√
Granularity		√		
History			√	
Privacy			√	
Documentation			√	



or investigations conducted on motor carriers and other entities (e.g., U.S. DOT number, review date, review type, and safety rating). Although this data could provide value in Big Data analytics for TIM, a challenge or limitation is that the data is not available in raw format due to privacy and sensitivity concerns. The data may only be accessed through various extracts or reports (e.g., crash, census, inspection, safety profiles, or customized reports), which must be ordered for a small fee.

Table 5-31 shows the research team's assessment of the readiness of the MCMIS dataset. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment table for the MCMIS dataset can be found in Appendix A, Table A-29.

Private data aggregator datasets. The datasets available from HERE and INRIX may be the most advanced and comprehensive aggregated datasets relevant to transportation and TIM for Big Data analytics. HERE datasets aggregate and analyze road transportation data from more than 80,000 data sources covering over 180 countries. Most of the HERE datasets are real-time datasets designed to support real-time decision-making. Some of the HERE datasets are archived indefinitely to support some of the services HERE provides (e.g., mapping, visualization, and predictive services).

INRIX gathers real-time, predictive, and historical data from more than 300 million sources, including commercial fleets, GPS, cell towers, mobile devices and cameras. Speeds and vehicle counts covering more than 5 million miles of roadways worldwide are enriched with other data, including construction and road closures, real-time incidents, sporting and entertainment events, and hazardous road conditions precipitated by weather.

The primary challenge and limitation of using these datasets for Big Data analytics for TIM is that HERE and INRIX datasets are proprietary and cannot be accessed as a whole (in raw format). Rather, some of the data they contain is accessible through DaaS solutions or, in the case of INRIX, may be purchased as extracts via special requests, likely at a relatively steep price and still at a limited resolution.

Tables 5-32 and 5-33 summarize the research team's assessment of the readiness of the HERE and INRIX datasets. The maturity rating (based on the Socrata Maturity Model) is indicated by the icon(s) to the right of each table. The detailed data assessment tables for the HERE and INRIX datasets can be found in Appendix A, Tables A-30 and A-31.

Table 5-31. MCMIS readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility			√	
Storage			√	
Integration		√		
Relevance and Sufficiency		√		
Quality	√	√	√	
Collection Frequency		√	√	
Granularity			√	√
History			√	
Privacy			√	
Documentation		√		

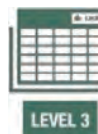




Table 5-32. HERE data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				✓
Storage				✓
Integration			✓	✓
Relevance and Sufficiency			✓	
Quality			✓	✓
Collection Frequency				✓
Granularity				✓
History				✓
Privacy			✓	
Documentation				✓


Table 5-33. INRIX data readiness.

Data Readiness	Lagging	Basic	Advanced	Leading
Accessibility				✓
Storage				✓
Integration			✓	✓
Relevance and Sufficiency			✓	
Quality			✓	
Collection Frequency				✓
Granularity				✓
History				✓
Privacy			✓	
Documentation				✓



5.3 Summary

This chapter has presented the research team's assessment of 31 data sources in six data domains. The data sources were assessed on several criteria and against two data maturity models. The purpose of the assessment was to bring to light the characteristics, practices, ease of accessibility, costs, and challenges associated with each of the data sources, particularly in relation to the potential use of the data for Big Data analytics for improving TIM.

The state of the practice encompasses datasets from sources that range from simple, scattered spreadsheets to relational databases, to turnkey services such as web services, APIs, GIS-based portals, and DaaS solutions that allow for data analysis and/or viewing of the data in graphic formats (e.g., on a map). None of the data sources reviewed (except for HERE, INRIX, and Waze) went beyond the use of relational databases, and relatively few of the data sources stored or managed the data in a way that could facilitate Big Data analytics. Even the more-advanced web services, APIs, and DaaS solutions were not ideal, because the proprietary nature of many of these services and systems did not lend itself to Big Data analytics. Because Big Data analytics makes use of a cluster of servers rather than individual workstations an environment is needed in which all the datasets can be stored.

The overall take-away from this assessment is that, even though a wide range of data sources could contribute to a better understanding of the trends, relationships, and dependencies associated with TIM operations and performance, existing challenges limit the immediate application of Big Data for TIM. Most of the data sources are not yet at a maturity level to support

Big Data analytics because these sources and datasets lack openness, completeness, quality, collection frequency, and/or granularity, or because they are inaccessible due to legal, privacy, and proprietary issues. More immediate applications for TIM may be feasible through the integration of state traffic records data at the state and national level; use and integration of nationwide probe data (e.g., data from systems like the NPMRDS, if made available, or purchased from third-party providers like HERE Technologies or INRIX, Inc.); integration of national weather data sources like MADIS or data from third-party weather services; and the use of social media or crowdsourced data like that available from Waze. Another opportunity, but one that would require a greater level of effort, would be to integrate public safety CAD data.

Moreover, at a state level, it is likely that integration of a variety of data sources would not constitute true Big Data because incidents are rare events (i.e., the volume of data is too small) and Big Data tools are data hungry. To build models for TIM response, it will be necessary to have a lot of data, which will likely require nationwide incident data.



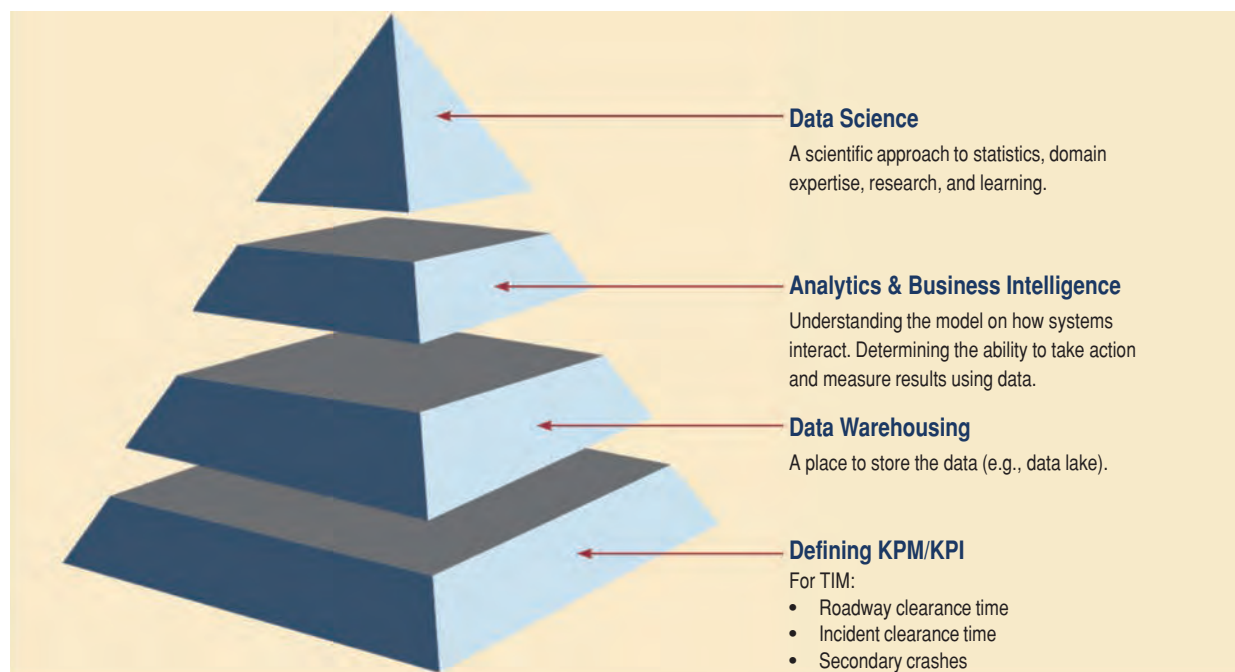
CHAPTER 6

Big Data Guidelines for TIM Agencies

Although most states understand the value of collecting and analyzing data to guide their business decisions, most fail to grasp the scale of the data, the expertise needed for Big Data analytics, and the significant shift away from traditional approaches (including approaches to data collection and analysis, data storage and management, and procurement of IT services) that would be required before the implementation of Big Data. Although a significant shift is required, few of the adjustments are technical in nature. Most Big Data tools these days are readily available, turnkey, and relatively inexpensive to deploy. In fact, the significant shift required relates more to the capability and willingness of humans and agencies to embrace and negotiate a new way of conducting business (i.e., collecting, storing, and sharing detailed data and embedding analyses of the data in everyday business processes).

The Big Data pyramid (Figure 6-1) illustrates the stages required to reach a level of applying data science, from the foundational activity of defining key performance measures (KPMs) and key performance indicators (KPIs) to the achievement of a mature Big Data practice at the top of the pyramid (Drow, Lange, and Laufer 2015). The stages shown in the pyramid are:

- **Defining KPMs/KPIs:** KPMs and KPIs are measurable values that demonstrate how effectively an organization or business domain is achieving key business objectives and targets. High-level KPMs/KPIs may focus on the overall performance of the agency, whereas low-level KPMs/KPIs may focus on department-specific processes such as operations, construction, or maintenance.
- **Data Warehousing:** This stage involves developing and maintaining an environment in which data created by the organization can be captured, stored, and managed to allow for the calculation of the KPMs/KPIs. Traditionally, data warehouses were designed using one or more relational databases, which stored cleaned and organized data; however, with the increasing volume and complexity of Big Data datasets, data warehouses have evolved to become large repositories of managed raw (uncleaned and unorganized) data on which analytics, business intelligence, and Big Data analytics can be performed. These repositories often are called *data lakes*.
- **Analytics and Business Intelligence:** With the data lake established, this stage involves developing and maintaining the analytics and business intelligence tools and processes needed to generate alerts, dashboards, reports, and other communications or interactive tools that allow agency personnel to (1) monitor KPMs/KPIs over time and across the agency, (2) be alerted when KPM/KPI thresholds are reached, and (3) investigate abnormal behaviors in KPMs/KPIs.
- **Data Science:** Having established one or more data lakes and developed the necessary analytics and business intelligence tools and processes, the topmost level of the pyramid consists of a data science environment that allows for many advanced data analysis tools and processes capable of (1) mining large amounts of unstructured data such as text, images,



Source: Adapted from "Big Progress in Big Data" (Drow, Lange, and Laufer 2015)

Figure 6-1. The Big Data pyramid.

and videos; (2) performing advanced statistics; (3) quantifying and classifying millions to billions of records; and (4) building prediction models to be assessed and used across the entire organization.

Based on the research conducted for this project and the information presented in Chapters 2 through 5, the current state of the practice for TIM data collection, storage, and analysis appears to be between the first and second tiers on the Big Data pyramid. At this point, very limited TIM data is being collected and shared amongst partner agencies, and a solid data lake has yet to be built as a foundation for the development of TIM business intelligence (the third tier of the Big Data pyramid) and TIM data science (the top tier of the pyramid). Accordingly, this chapter presents suggested guidelines that involve various changes that will be necessary for agencies to (1) develop a usable Big Data store (data lake), (2) implement agency-wide analytics and business intelligence, and (3) pursue the development of an evolving data science environment beneficial to the entire agency.

The guidelines are set forth to enable TIM agencies to position themselves for Big Data. Expressed at their highest level, the guidelines suggest that agencies prepare to:

- Adopt a deeper and broader perspective on data use;
- Collect more data;
- Open and share data;
- Use a common data storage environment;
- Adopt cloud technologies for the storage and retrieval of data;
- Manage the data differently;
- Process the data; and
- Open and share outcomes and products to foster data user communities.

The sections in this chapter provide more details, categorized as sub-guidelines within each of the high-level guidelines.

6.1 Adopt a Deeper and Broader Perspective on Data Use

Traditionally, many organizations have conducted business by relying on business intelligence (often reported on the basis of limited data), on expert opinions, and even on intuition. The functions of analysis and decision-making often have been limited to a relatively small number of high-level managers and executives. The structure of this traditional approach to conducting business ensures that the organization's vision, strategies, and operational decisions are shaped—and limited—by what is available to (and can be perceived, understood, and used effectively by) these individuals. It is an approach that no longer works in the context of Big Data. Big Data is too big, too complex, and too confusing to be tackled by a small set of individuals within an agency.

Transportation agencies are encouraged to develop Big Data within a collaborative environment.

Big Data enables many differing analyses to be performed on very large amounts of detailed business data in parallel, and at a relatively low-cost, by many individuals across the organization. A Big Data approach allows for the size and complexity of the data to be handled in a distributed fashion rather than a centralized one, enabling distributed decision-making across all levels of the organization. Compared to the traditional, centralized approach, which entrusts only a few

key individuals with analysis and decision-making, a distributed Big Data analytics approach takes advantage of the commoditization of data analysis to depersonalize decision-making. This approach enables members from the lowest level to the highest level of an organization to observe and react on their own to changes detected through the organization's large pool of data.

Although distributed decision-making across an entire organization would be beneficial, in the case of TIM, the benefit would be further enhanced if the Big Data approach was extended beyond the boundaries of the transportation agency to involve TIM partners such as law enforcement, fire, EMS, and towing companies. Ideally, transportation agencies could develop Big Data as a collaborative environment or ecosystem that gathers transportation employees, experts, contractors, consultants, other state and local employees, and members of universities to share, analyze, and visualize data to derive the most value from it. Only after such an environment is in place (i.e., multiple datasets are collected on a regular basis, shared, managed, and analyzed by many inside and outside the organization) can more advanced data analytics, such as deep learning, be developed to support efficient predictive, proactive, and real-time decision-making across the participating organizations.

Even with such an ecosystem in place, certain data-hungry, advanced analytics may not be able to be implemented at an agency level. These advanced analytics typically require hundreds of thousands to tens of millions of data points to develop effective models for medium to hard problems. Because traffic incidents are, by their nature, infrequent events, it is not likely that an agency or state on its own will be able to collect enough traffic incident data to satisfy the data needs of such advanced analytics. An even broader opening of the data environment to include traffic incident data from agencies across multiple states would be required. Ultimately, a shared nationwide dataset, collating detailed traffic incidents from multiple agencies, may be the ideal environment to apply advanced data analytics.

6.2 Collect More Data

The main tenet of Big Data is to identify and leverage patterns and behaviors within an organization or population by combing through large amounts of detailed data collected throughout the organization or population. The more detailed and extensive the data, the

better the chance of discovering patterns and behaviors that can be tracked, analyzed, predicted, and embedded into organizational decision-making processes. Without enough detailed data, however, Big Data analytics is not possible.

Although existing incident-related data may be sufficient for traditional decision-making, it is far from sufficient for transportation agencies to conduct Big Data analyses for TIM. The resolution with which the data is currently gathered is not sufficient to be able to perform Big Data analytics. Rather than attempting to summarize or aggregate data at collection, extensive and detailed data needs to be collected for every incident, including minor incidents. For example, instead of characterizing weather conditions using the MMUCC attributes (e.g., clear, cloudy, fog, smog, and smoke), detailed weather variables such as wind bearing, dew point, and cloud cover would be collected from the beginning to the end of each incident. In addition, data that is not currently collected, such as crowdsourced data and social media posts from the beginning to the end of an incident could be gathered and stored to provide additional data that might help detect incidents earlier and understand drivers' expectations and behaviors while stuck in traffic as the incident response unfolds.

Collecting data at this level of detail for every incident cannot be accomplished solely through traditional methods (i.e., using standard forms). TMCs and responders would be completely overwhelmed if they were required to collect such detailed data for every incident. Furthermore, responders do not have ready access to the detailed information (e.g., weather, roadway conditions, roadway characteristics) that would need to be associated with the incidents. Therefore, large detailed datasets need to be created by augmenting human-collected data with machine (sensor)-collected data and other external data sources to obtain a more complete and detailed description of incidents and their associated responses. For example, information about the responders involved, as well as their incident scene arrival and departure times, could be derived from AVL data logs rather than captured by a TMC operator or a law enforcement officer. Detailed weather data and detailed incident injury data could be derived from the information already collected by external data sources such as the NOAA MADIS dataset and the NEMSIS dataset. Thus, detailed weather and injury data for each incident could be collected by extracting data from each dataset surrounding the time and location of the incident without requiring human data entry. The most likely way transportation and TIM agencies will be able to build a data lake containing enough detailed data to leverage Big Data analytics is by integrating as many internal and external machine-collected and human-collected datasets as possible to establish sufficient volume and variety for Big Data analytics.

Transportation agencies can collect more data by augmenting internal datasets with external datasets.

Another challenge is that, although multiple existing datasets could be used to build a Big Data data lake for TIM, many of these existing datasets are not ready to be integrated into a single, minable data lake. Many of the datasets are siloed or are not accessible as a whole using a machine-friendly format. Data sharing and data use may be restricted by public record laws, proprietary storage solutions, the presence of sensitive information, or simply the fear of exposing potentially damaging information. Also, some of the data may not be complete enough or detailed enough to be used for Big Data analytics. These are all obstacles that will need to be remedied before the establishment of a solid foundation for TIM Big Data analytics.

The IRCO developed as part of this project is a first attempt to describe how TIM-relevant data elements in the various datasets relate to each other. As such, the IRCO can be used as a guide to how to integrate these various datasets and what in each dataset needs to be modified, augmented, and changed so that the relationship between the data elements can be exploited during analysis. The IRCO is presented and described in Appendix B.

Table 6-1. TIM-relevant datasets.

Dataset	Readiness	Challenges
State Traffic Records	High	Siloed, quality, legal
Social Media	High	Unstructured, quality, legal
Weather	High	Format, quality, completeness, accessibility
Nationwide Probe/Speed	High	Accessibility, quality, resolution, legal
NFIRS	High	Accessibility, resolution, completeness
NEMSIS	High	Accessibility, legal
AVL Data	Medium/High	Accessibility, quality, legal
Public Safety CAD	Medium	Unstructured, non-standard, completeness, quality, accessibility
MCMIS	Medium	Unstructured, quality
Safety Service Patrol	Medium	Accessibility, quality, completeness, legal
511 Data	Medium/low	Unstructured, completeness, quality
Telematics	Low	Quantity, quality, accessibility, legal
Traffic Sensor	Low	Accessibility, resolution, quantity, quality
Traffic Video	Low	Unstructured, accessibility, quantity, legal, quality
Public Safety Video	Low	Unstructured, accessibility, quantity, legal, quality
Toll	Low	Accessibility, legal

Table 6-1 lists TIM-relevant datasets that could be leveraged to build a data lake. For each dataset, the table provides the readiness for and associated challenges associated with integration of the data into a Big Data data lake.

When attempting to extract the most value from limited data using the traditional approach, the most difficult part of the analysis often is the selection/development of the software and tools. With Big Data, on the other hand, the data itself is the most difficult, most expensive, and most valuable part of the analysis. Without large amounts of detailed data, there are no Big Data analytics or predictions or classifications to support TIM decisions. Software that can analyze the necessary volumes of data is readily accessible, often inexpensive, and disposable, as new Big Data analytics solutions replace previous ones every 3 to 6 months. Therefore, at this stage, the first and foremost focus of Big Data for TIM is to ready and gather as many TIM-relevant datasets as possible to build a solid foundation for TIM Big Data analytics.

6.3 Open and Share Data

For Big Data analytics to work, “open” data must be available, meaning the following:

- The data must be available as a whole at no more than a reasonable reproduction cost;
- Users must be permitted to re-use, redistribute, and intermix the data with other datasets; and
- Ideally, the data should be available to any person, group, or field of endeavor.

The effectiveness of Big Data analytics depends intrinsically on the willingness of transportation agencies to open and share data, both internally and externally to partners.

One of the foundational aspects of Big Data analytics is the ability to explore and correlate a range of very large datasets to uncover unknown relations and patterns that could lead to an improvement in the state of the practice. If the data is shared in a previously aggregated or summarized form (as opposed to raw form), its value is tremendously diminished for Big Data analytics because it will lack the resolution needed to detect patterns and relationships. Similarly, the ability to leverage data for Big Data analytics can be compromised if the data is available in detail but in a format that is only accessible by using a specific software (the purchase or use of which involves a significant cost), because the cost of accessing the data may limit the scale at which it can be processed. Finally, some data can be available in detail using an accessible format, but its use and distribution may be restricted to select individuals or organizations. Here again, the value of Big Data analytics is significantly diminished, as the resources, skills, and interest needed to allow such analysis to be performed may not exist among the people or organizations that have the right to use this data.

The open aspect of Big Data functions in direct contradiction with traditional organizational views and culture about data. More often than not, detailed data is the sole property of a division or program, and only samples or summaries are shared with the rest of the organization or with external parties. This traditional approach persists for various reasons, which may include (1) resistance to loss of control over the data; (2) fear of exposing known or unknown poor performance or flaws, or (3) fear of potential lawsuits associated with data privacy concerns or potential security leaks. Nonetheless, without opening and sharing detailed data, there is no Big Data analytics. Big Data analytics is too large and complex to be the business of a single entity. By design, it focuses on allowing many entities the ability to explore many large and varied datasets rather than maximizing analytical value for a dedicated domain. Therefore, for Big Data analytics to be feasible, obstacles to the sharing and opening of datasets relevant to TIM need to be removed. The next section of this chapter describes three of the most common roadblocks to the opening and sharing of TIM-relevant datasets, and proposes possible solutions to remove or circumvent them.

6.3.1 Public Records Laws

Public records laws attempt to limit to the extent possible the legal risks encountered by agencies when sharing sensitive data such as PII. These laws are extremely restrictive and prohibitive to the point of limiting the storage, access, and processing of the data to specific physical buildings, as well as to specific systems and personnel. Although these hardline, over-sized solutions may be satisfactory from a legal standpoint, they reduce, and at times fully strip, the usability of data. To remedy this roadblock, alternative solutions need to be developed. One solution is to allow for the opening and sharing of a modified version of the original data, where sensitive data elements have been obfuscated or anonymized. Another solution may be to include legal disclaimers that protect agencies in the event of a data breach that occurs under the control of the data requester.

6.3.2 Proprietary Data Formats

Many widely used commercial software products use proprietary data formats that not only store the data created by users, but also make it difficult for users to export the stored data to another software. In other words, proprietary file formats attempt to lock users in so they must continue using a specific vendor's software. Because traditional data analysis uses relatively small amounts of data, this aspect of proprietary data file formats is not a huge obstacle. Most software provides more or less similar data analytics and visualizations, so the need seldom arises to move data from one software to another. Even when moving the data is an absolute must,

the cost and time needed to export or even recreate the data generally is not prohibitive. When dealing with Big Data analytics, however, the much larger size of the datasets involved and the constantly evolving variety of analyses and visualizations that can be performed mean that a Big Data dataset created using a proprietary file format significantly risks future accessibility and value. Indeed, converting the entire Big Data dataset to another format so it can be analyzed with other Big Data datasets will likely be cost and time prohibitive. While being of great benefit to the vendors, proprietary file formats also limit data analysis because no vendor can offer the full Big Data analytics domain, and most vendors are rather slow to adopt new analytical features when compared to open-source software supported by entire developer communities. There is also no certainty that a specific vendor will remain in business in the next few years. The Big Data world is fast-changing, and vendors and solutions come and go rapidly as new and faster analytic solutions are created. If the choice is made to adopt a Big Data solution using proprietary file formats, that choice incurs a potentially significant risk that the agency could be left with a lot of unusable data if the vendor goes out of business.

The only way for Big Data datasets to be merged and analyzed using a variety of constantly changing analytical software and solutions is to use non-proprietary and open file formats. These file formats do not hide the data they store, allowing human or machine users to easily retrieve the data to quickly re-use it. The research team suggests that open file formats be the only formats used to store data intended for use in Big Data analytics.

6.3.3 Contract Data Clauses

When transportation agencies, including TIM programs, outsource IT or data management to third parties, they also relinquish some control over the data. A third-party service provider may be unable or unwilling to reciprocally share the data being generated by transportation agencies, or to share information about the data (e.g., how it is organized, how it is managed, or its quality). Such restrictions may curtail data access in ways that preclude its use in a Big Data environment, and may even curtail data access entirely. For example, the data itself may be restricted, with only the results of analyses made available upon request (e.g., through a helpdesk service).

TIM agencies are cautioned that engaging in agreements with partners, vendors, or service providers that severely limit internal or external access to actual data or that attempt to share ownership of the data will impede the transition to Big Data analytics. Data is now the most valuable resource that organizations possess. Agencies are advised not to allow their data to be controlled or owned, even partially, by a third party.

6.3.4 Benefits of Opening and Sharing Data

Opening and sharing data allows datasets to be combined and analyzed to create new knowledge. Opening and sharing data also helps build a data culture across an organization by increasing transparency and accountability; helping develop trust, credibility, and reputation; promoting progress and innovation; and encouraging public education and community engagement.

The Utah DOT has recently started to develop an open data culture across the entire agency. Borrowing from the development of open data policies in the regional health care system, the Utah DOT has implemented in-house policies focused on fostering the opening and sharing of data by rewarding the publishing of data, whether good or bad, then working to improve its quality through monitoring and analysis (Applied Engineering Management Corp. and toXcel, LLC 2018).

6.4 Use a Common Data Storage Environment

A common data storage environment is vital for Big Data. In traditional data analysis, one or more datasets are imported into an analytic tool or platform like a relational database or a statistical software package and processed on the workstation or server where the analytical tool is installed. For Big Data datasets, this process is not feasible; the datasets are way too big to be easily moved in and out of storage without spending significant time and money. Also, traditional data analysis tools (except top-end tools that require the use of super computers) often are run on a single server, and even the server with the largest storage available on the market cannot store a Big Data dataset. Big Data datasets are so large that they need to be stored across multiple connected servers, called *clusters*. Unfortunately, most traditional analytical software does not work on server clusters, and the few that do are very expensive.

Transportation agencies can benefit by collocating datasets in a cloud environment.

To avoid having to invest in cost-prohibitive analytical solutions and having to spend large amounts of time to duplicate and move around large datasets, early Big Data ventures have adopted a different approach—never moving the data itself, but instead moving the data processing software to the data on each of the servers in the cluster. This premise is the foundation of cloud computing. All commercial and private cloud systems follow this principle, offering the ability to collocate data into a common storage environment with a series of development kits and tools to process it where it resides. Without collocation of datasets within the cloud, or a cloud-like common data storage environment that provides the ability to process data where it resides, there are no Big Data analytics.

6.4.1 Data Silos

Currently, most TIM agencies do not have a common storage system for the data they use. Rather, many data stores have been created within each agency (or each department or district within an agency). The hardware, software, and data management methods have varied across each implementation and been driven by organizational boundaries, available budgets, resources/skills, and contractor offerings. These kinds of data stores are commonly referred to as *data silos*. Storing and organizing data this way may have been sufficient for traditional data analysis and may have worked for years, but it will not allow for Big Data analytics on TIM agency data. For Big Data analytics to succeed, TIM agencies will need to extract data from each data silo and collocate all of it into common storage where the data can be processed “in situ.” An even better approach would be to bypass the need for extraction and store the data created in each department or district directly into the common storage, eliminating siloed data stores altogether. Common data storage has the potential to transform data analysis in an organization by providing a single repository for all the organization’s data (whether structured or unstructured, internal or external) and enabling analysts to mine all the organizational data that is currently scattered across a multitude of data stores.

6.4.2 Data Virtualization

Some IT vendors offer an alternative way of meeting the need for common data storage to use Big Data. Called *data virtualization*, this approach does not physically collocate datasets into a common storage environment. Instead, it links an organization’s various siloed data stores *without moving* the data, by providing a single “virtual” view of the data and allowing the data to be queried using distributed data processing across each of the individual data stores.

Data virtualization could easily allow for siloed datasets across an organization to be organized, managed, and queried without ever having to relocate the data into common physical storage; however, this approach has two main weaknesses. First, virtualized common data stores depend greatly on the performance and quality of the individual (siloed) data stores. Second, the ability of virtualized data stores to analyze data is limited because the hardware specifications and software capabilities of the data siloes may not permit the data processing tools to be moved to where the data resides in order to be run locally. To perform data analysis it would be necessary to copy the data from the silo into a temporary storage environment that is capable of running the data processing tool. The need to copy the data to run analyses essentially negates the benefits of the data virtualization.

Data virtualization solutions can be used to perform basic aggregation and filtering on organizational data to capture the trends of various KPIs and KPMs, but they are not suited for more advanced analytics such as classification, clustering, graph analytics, and machine learning. Data virtualization shows promise, but the concept is still new. Therefore, at this time, it is suggested that transportation and TIM agencies refrain from using this technology as they develop their common organizational data stores.

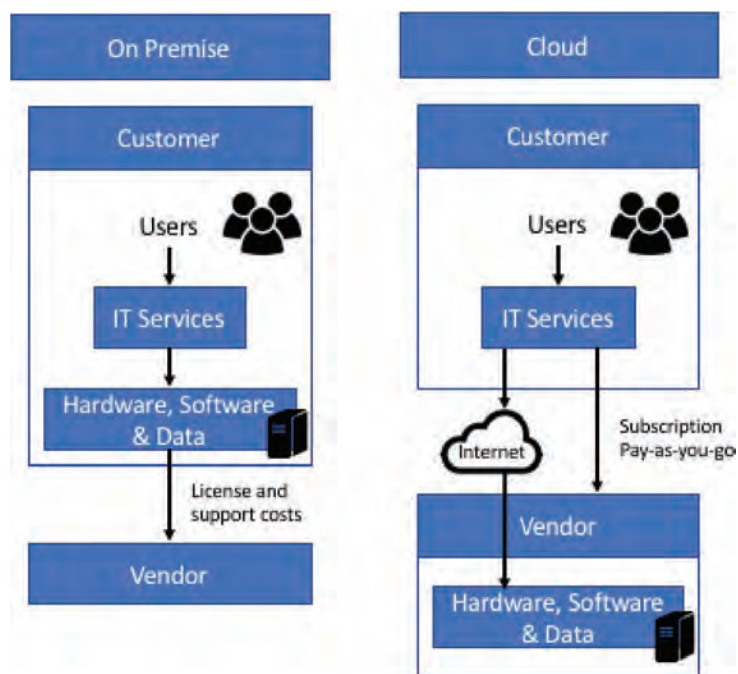
6.5 Adopt Cloud Technologies for the Storage and Retrieval of Data

Given their scalability, agility, affordability, redundancy, and protocols for safe sharing, cloud technologies can offer organizations substantial cost savings and improved security, which in many ways makes them an ideal fit for Big Data analytics.

Cloud technology is inherently linked to Big Data analytics. The cloud was born out of necessity when companies faced the enormous costs associated with the implementation and maintenance of on-premises infrastructure that could store and process Big Data datasets; however, cloud infrastructure is not just on-premises IT infrastructure relocated to a data center and made available as a service. The cloud represents a completely different type of IT infrastructure that is built entirely on relatively inexpensive and interchangeable *commodity hardware* and is designed to support the storage and processing of very large amounts of data for many users on a pay-as-you-go basis.

The rationale behind the use of cloud infrastructure is to increase IT efficiency and sustainability; reduce the risk of IT infrastructure obsolescence; benefit from scalable, flexible, and on-demand data storage and data analysis capabilities; and reduce IT infrastructure operations and maintenance time to a minimum by leasing a share of a huge IT infrastructure as opposed to owning it. Figure 6-2 shows a diagram representing the differences between on-premises and cloud architecture.

With the cloud, IT infrastructure is no longer defined primarily by the acquisition, installation, and maintenance of hardware and software; nor is it defined as the development and implementation of custom software solutions to support agency needs. Rather, the cloud enables agencies or companies to choose from among a series of services (e.g., data storage, data processing, business rules engines, messaging engines) on which to build their data processing workflows. Purchasers of cloud computing services eliminate the on-site need to obtain, maintain, or replace obsolete hardware, and to patch, maintain, and upgrade software. The company or agency is protected against sudden hardware failures and loss of data. Cloud services are redundant by design; service providers are able to quickly and automatically move to new hardware when failures occur and constantly maintain several copies of the data in parallel to ensure that no data is lost. Cloud services also can copy data and software to additional servers in real time to cope with demand surge, which means they can be operated and maintained to a defined



Source: NCHRP Research Report 865 (Applied Engineering Management Corp. and toXcel, LLC 2018)

Figure 6-2. On-premises versus cloud infrastructure.

level of service by the cloud service providers. As a result, the prime concern of an organization using cloud infrastructure is no longer to ensure the reliable and sustainable operation and maintenance of the IT infrastructure underlying its data workflows. Instead, the organization's focus can shift entirely to the design, operation, and maintenance of the many data workflows capable of improving business processes across the entire agency. In effect, using cloud-based services can enable an agency's IT management to switch from infrastructure administration to data storage, access, and processing administration.

6.5.1 Understand the Cost Savings of the Cloud

The emergence of cloud computing has made it easier to provide organizations newer and higher-capacity technology at a better cost. Cloud computing can reduce agencies' hardware- and software-related costs, and can make a wide array of applications available to any organization, big or small. Cloud computing minimizes the need for individual agencies or companies to purchase expensive hardware and yearly CPU software licenses. The costs of supporting the necessary IT infrastructure—now borne primarily by the service companies—are built into the prices the services charge to their users; however, these costs are spread across many more users, so each user's share of the cost is vastly reduced. Moreover, client organizations often are free to select and pay only for bundles of services targeted to their needs.

6.5.1.1 Scalability

A traditional approach to scaling up an existing IT infrastructure to increase processing power and storage space would require the addition of more physical servers and additional software licenses. The virtual nature of the cloud allows for unprecedented flexibility. Organizations can scale up or down to the desired level of processing power and storage space easily and quickly without having to add to or maintain the physical infrastructure.

In addition to growth-driven variations in processing power and storage, Big Data analytics adds a second layer of power and storage variability, as the analyses involved typically are not processed evenly over time. Dataset processing is rather irregular and includes large spikes driven by human decisions, environmental changes, or the obsolescence of data models, which can occur at any time. To handle such irregularities and accommodate peak data processing, an on-premises IT infrastructure would represent a significant IT infrastructure investment that would almost never be used at its full capacity. In contrast, cloud environments can scale up and down to adjust to surges and drops in data processing almost in real time. Organizations that use cloud-based services can maintain a much smaller on-site IT infrastructure while accessing (and paying for) the storage and processing strengths of the cloud on an as-needed basis.

6.5.1.2 Agility

Using the traditional approach, it can take weeks of setup and many days of troubleshooting to upgrade and transition from a legacy IT system to a newer IT system. Cloud computing services maintain a clear separation between data storage and data processing. Therefore, as new cloud data processing services become available, an organization can begin testing the new service on data within minutes while continuing to process the data with the current cloud services. The old and new systems can run in parallel. This facility also enables data stored in cloud infrastructure to be processed by many distinct and independent data workflows, satisfying the specific analytical needs of many groups within an organization (e.g., financial, operations, human resources), each evolving independently. As new requirements and business areas are created, new data workflows can be added without stopping, slowing, or affecting those already in place.

6.5.1.3 Affordability

Cloud computing can be beneficial for organizations that wish to use up-to-date technology while remaining on a budget. Before the cloud, companies used to invest huge sums of money selecting and setting up IT systems capable of satisfying all the needs of the organization, then spent large amounts of money for its upkeep until the system became obsolete, and then restarted the selection process for a new system, bearing the full cost of the IT system migration. Cloud environments constantly update their services and typically allow the new services to be tested at a reasonable cost. Given the ability to run cloud-based workflows in parallel, the new services can then be rapidly implemented into production with little to no downtime, effectively reducing migration costs from an entire system redesign to one of a data workflow redesign.

6.5.2 Understand Cloud Security

Files stored in reliable cloud services are some of the most secure files that an organization can have, provided the organization uses robust authentication and effective password policies. Cloud service companies all provide reliable and secure cloud services for consumer file storage and processing. Three important aspects of major cloud storage systems are redundancy, security, and safer sharing of data.

6.5.2.1 Redundancy

At any one time, cloud services typically store at least three copies of each piece of data, with each on different servers. If one copy is lost, another copy is immediately recreated on another server. For a file to be lost on a cloud system, all three copies would need to disappear at exactly the same moment (e.g., from the simultaneous failure of three separate hard drives on three different servers). Although this is extremely unlikely, at scale, when handling exabytes of data,

it does happen to a tiny fraction of the data. In the occurrence of such a rare failure, files generally can be recovered from server backups within a couple of days.

6.5.2.2 Security

Provided an organization effectively manages its credentialing process (ranging from passwords to involved authentication procedures), only authorized users can access the files it creates and stores on the cloud. Data that is stored on the cloud resides in files on compartmentalized virtual hard drives on servers that are located in remote, physically secure data centers. Access to these files is gained through highly secured, encrypted connections and can be restricted as desired to a larger or smaller set of authorized external machines. More often than not, the biggest security weaknesses of cloud systems are the weaknesses of the local machines (e.g., the laptops or workstations) being used to connect to them.

Although the federal government has established regulations and certifications such as the Federal Information Technology Acquisition Reform Act (FITARA) and Fed RAMP to ensure the security of cloud-based federal systems, it is important to note that state regulations are just now starting to take the cloud into consideration. Current state IT security regulations are mainly built around traditional IT assumptions that sometimes directly conflict with the adoption of cloud services by mandating, for example, that all state data be stored and processed within the premises of state buildings. Developing compliant and secure cloud-based systems for state agencies will not only be a matter of establishing and monitoring compliance with current laws, it also will be a matter of ongoing coordination as state laws and regulations adjust to fit the requirements of cloud services while maintaining their original intent.

6.5.2.3 Safer Sharing

Instead of sharing data using a physical storage medium like a thumb drive (or a hard drive for larger datasets), use of the cloud enables an organization to (1) grant real-time data access to certain people; (2) control what privileges approved users have with regard to the data (e.g., read, write, run analyses, generate reports); and (3) remove access immediately in case of problems. This managed access to the data minimizes the risk of corrupting data or infecting it with computer viruses, as can easily occur when data is copied using intermediate storage devices. Cloud storage services also have *versioning* systems that keep a history of each file, so that in the event of accidental or intentional corruption, deletion, or overwrite, the file can be recovered.

6.5.3 Recognize the Inherent Connection Between Big Data Analytics and the Cloud

The scalability, safety, and agility of cloud environments make them ideal for processing Big Data datasets. Cloud environments reduce the hardware- and software-related IT burden of organizations, allowing agencies to focus on their data. Many state DOTs have started to explore or use cloud services to reduce the cost of data storage (e.g., by using cloud-based word processing software that has built-in cloud back-up). That said, concerns related to outsourcing significant IT services and potentially sensitive data to a shared cloud-based IT infrastructure remain a barrier to cloud adoption by DOTs. Following current policies and regulations (which are based on a traditional IT approach), DOTs are likely to prefer to hire a contractor to host and manage a data center that is solely dedicated to the IT and data needs of the DOT rather than use a shared cloud environment. Unfortunately, this option does not suffice for Big Data analytics, for which the data storage and computing needs are simply too big to be funded by a single division or agency. To adopt Big Data analytics, transportation agencies—particularly TIM agencies—will need to adopt the use of the cloud environment.

Transportation and TIM agencies have two options for adopting cloud services:

- The first option is to use a commercial cloud service provider. This option is also the easiest to implement and would allow the transportation or TIM agency to benefit from an available, very large, and very flexible cloud-based infrastructure at a low cost. This option comes with the perceived risks of (1) storing agency-created data on infrastructure owned and maintained by an external party and (2) sharing the cloud services with other entities.
- The second option is for several transportation or TIM agencies to partner to build a private cloud. This option could offer more customization to the common needs and concerns of the agencies, but it would effectively limit the sharing of the cloud resources and services to the participating agencies and individuals within that community. The time and costs to create the new infrastructure, ensure adequate security, and migrate the various agencies' current data to the new, shared storage and processing system would be significant. The agencies also would retain all the costs of maintaining and continuing to update the infrastructure (both hardware and software).

A potential third option could bridge the first two options by combining the data storage of multiple agencies as in Option 2 and leveraging the use of commercially available cloud computing services as in Option 1. This option would still be significantly more expensive to implement, and it would not be able to scale as efficiently as the first option.

The research team advises that individual transportation and TIM agencies not attempt to build their own cloud infrastructure to support Big Data analytics. This approach will likely be cost prohibitive when compared to a commercial cloud or shared private cloud solution (and might even exceed the entire IT budget of the agency), and it will most likely never be able to achieve the required data processing capabilities within budget.

Big Data requires a different approach to data management. Transportation agencies are advised to store data “as is,” maintain access to data, structure the data for analysis, ensure that data is uniquely identifiable, and protect data without locking it down.

6.6 Manage the Data Differently

Big Data requires a different approach to data management. The collaborative nature of Big Data and the rapid pace of change of Big Data datasets and analysis tools are pushing data management away from strict control of data and software to a more flexible approach that supports collaborative and evolving analysis and focuses on data accessibility, sharing, and security; on metadata; and on real-time data quality monitoring.

6.6.1 Store the Data “As Is”

Data within the common data storage should not be modified from the way it was when it was collected. In other words, it should be stored “as is,” which is often referred to as storing “raw” or “unprocessed” data. This approach differs significantly from traditional data warehousing approaches, which first clean the data, then structure it according to a predesigned data model (i.e., schema), and then store it in a relational database.

Big Data datasets and analytics tools are rapidly changing and improving over time. Cleaning and organizing data according to a predefined data model is not ideal in this environment, as these steps may remove significant elements of the data that could be of interest in future analyses. Keeping the data in its raw state helps to prevent any loss of information and can facilitate future re-analysis and analytical reproducibility. As processing algorithms improve and computational power increases, new types of analyses will be able to take advantage of more granular variations in the data, outliers, and noise. If only cleaned and structured data has

been stored, these new analyses will not be possible. Storing the data in its raw format allows multiple analysts or researchers to perform differing analyses on the same data at the same time to confirm analytical results, assess the validity of statistical models, or directly compare findings across studies. For these reasons, data should be kept in raw format whenever possible (within technical limitations). In addition to being the simplest way to ensure transparency in analysis, having the data stored and archived in its original state gives a common point of reference for derivative analyses.

What constitutes raw data may vary depending on the type of data. Some data, such as video data, may not be able to be stored in a completely raw state. Raw video files typically are too huge to store economically; therefore, video files usually are minimally processed (compressed) to allow for storage. To the extent possible, transportation and TIM agencies are encouraged to store data in its purest form, and if derivations are required, they should be documented by archiving relevant code and intermediate datasets.

6.6.2 Maintain Data Accessibility

For effective use in Big Data analytics, the data that is placed in common storage also must be accessible to analysts. The formats used to publish or release the data (i.e., the digital basis on which the information is stored) matter when it comes to accessibility. Regardless of whether the source of the data is public or private, the data *format* can either be “open” or “closed.” An open format comes with specifications that the data is available to anyone and is free of charge, so that anyone can use the data in their own software with no limitations on re-use imposed by intellectual property rights. A closed format is a proprietary file format that (1) comes with the specification that the data is not publicly available, or (2) comes with specifications that make the data available for public use under certain limitations or conditions.

Data that has been released in a closed file format can cause significant obstacles to reusing the information encoded in it. For example, those who wish to use the information may need to buy the necessary proprietary software. Using data that has been stored using proprietary file formats can create dependence on third-party software or file-format license holders. Worse, it can mean that the data can only be read using certain software packages, which can prohibit Big Data analytics entirely. Open file formats permit data analysts and developers to produce multiple software packages and services without any limits or additional expenses and minimize the technical obstacles to reusing the data, which makes them perfectly suited to the nature of Big Data analytics. Consequently, for the purpose of conducting Big Data analyses, any data that is stored by transportation and TIM agencies into a shared common data storage environment should be stored using open (non-proprietary) file formats. Examples of open file formats include the CSV format, the JSON format, and the Apache Parquet file format.

6.6.3 Structure the Data for Analysis

Transportation and TIM agencies typically collect data on traffic incidents and responses through online or paper forms that are completed manually. The forms attempt to capture information to characterize and summarize each incident and response using multiple standardized and non-standardized fields (e.g., number of vehicles involved, injury level, weather conditions, number of lanes blocked). Often these records were developed to fulfill the needs of a specific domain, with the result that the data resides in independent data stores, in different formats, with little to no way to tie them together. Crash data and CAD data offer a good example: The combination of data elements in these two data sources could add value to the individual datasets, but often no common field (such as a record number) exists that can be

used to tie the two sources together. The lack of a common field makes it difficult to integrate the datasets.

Currently, to take full advantage of data placed in a common TIM data store, the data needs to be structured in a way that allows easy interpretation, use, and analysis. Typically, the data is structured such that each variable is set as a column, each observation is set as a row, and each type of observational unit is set as a table. Variations of this structure exist to meet the unique needs of the analyses to be conducted. More hierarchical ways of organizing the data, such as JSON and XML, also can be used.

A best practice for TIM data stored in common data storage is to annotate the data so that each file, as well as its content, provenance, and quality, can be identified and defined easily. This type of annotation is typically done using predefined organizational or nationwide standards by embedding data definitions directly within each file as metadata tags, or by creating metadata files associated with specific datasets.

Interoperability between datasets needs to be facilitated. This can be done by using variable names within each dataset that can be mapped to existing data standards. For example, the location of an incident record in an EMS database and the location of the same incident record in a state police CAD database could be expressed using a state-specific mile marker reference or by using the broader Census Bureau 2016 FIPS code and World Geodesic System 1984 (WGS84) reference system. These common referencing systems provide a more universally understandable way to describe location using latitude and longitude and county, city, and state codes. Used consistently across various datasets, such standards would facilitate data sharing across institutions, applications, and disciplines and would allow for these datasets to be merged and queried easily during analysis.

6.6.4 Ensure That Data Is Uniquely Identifiable

When dealing with Big Data datasets, it is often difficult to identify if specific data is accurate and genuine or if it has been corrupted (i.e., degraded, damaged, manipulated, or merely obsolete, having come from a neglected version of the dataset). To remedy this issue, common storage can use *cryptographic hashes*. Generated by an algorithm, a cryptographic hash is an alphanumeric string (e.g., SHA or MD5) that can take a “snapshot” of the data upon storage in the common data store. A cryptographic hash that uniquely identifies the data can be distributed across the dataset to ensure that the dataset has not been corrupted or manipulated. Given the volume of data in Big Data datasets, the likelihood of silent (undetected) data corruption is high. Consequently, it is suggested that methods like cryptographic hashes be used widely across data stores to ensure the sustainability of the collected datasets.

6.6.5 Sharing, Security, and Privacy

In datasets that contain information for which maintaining privacy is important, several methods can be put in place to protect data confidentiality without locking it down. These methods can involve both administrative (policy) steps and technical steps, as follows:

- Privacy protocols for the data can consider the various data stakeholders (e.g., funding agencies, human subjects or entities, collaborators). Both the National Science Foundation and National Institutes of Health have established data sharing policies and guidelines that can be used to develop privacy protocols that prevent sharing PII and that anonymize data on human subjects.
- Before distribution or sharing, sensitive data that is not required for analysis can be removed from the dataset.

- Because removing sensitive data can negatively affect the ability of the datasets to be mined in detail or merged with other datasets, alternative techniques to obfuscate sensitive data may be considered. Obfuscation methods like hashing techniques and encryption can anonymize personal information, but the methods used need to be sufficiently strong. In 2014, New York City officials shared publicly what they thought was anonymized data on cab drivers and over 173 million cab rides. However, the hashing method used was quickly recognized, and all 20 GB of data were de-anonymized in a matter of hours. To prevent this type of vulnerability, obfuscation methods should always be tested by a trusted third party before sharing the data, and the effectiveness of the method should be monitored over time (Goodin 2014).
- If the data itself allows identifiability, methods such as those used in the protection of medical datasets could be used. For example, sensitive datasets can be separated into two subsets: a reference dataset and a dataset containing changes against the reference dataset. The organization's policy may then specify that only the changed dataset is allowed to be shared, or it may specify that the data may never be shared but analysts are allowed to work on the changed dataset where it is stored. The latter option allows the organization to retain complete control over the data.

6.7 Process the Data

Because many TIM-relevant data sources have yet to achieve Big Data readiness, it is impossible to develop specific and detailed recommendations on how to approach the processing of TIM Big Data datasets. This section presents broader guidelines pertaining to cloud data processing.

Processing Big Data datasets is more challenging than processing smaller and more structured, traditional datasets. Traditional data processing algorithms typically require rapid access to any part of the dataset they process. Traditional data analysis software achieves this requirement by loading the entire dataset to process in computer memory (i.e., RAM) to be able to benefit from its speed. Unfortunately, no single server memory is large enough to hold an entire Big Data dataset. To be processed, Big Data datasets need to be split into smaller datasets and distributed across multiple servers, which means that algorithms used in traditional data processing software (e.g., linear regression, classification, and clustering) will not work on Big Data datasets. New algorithms capable of processing data scattered across multiple servers—in other words, algorithms designed for Big Data—need to be used. These algorithms often are more complex and more difficult to optimize than their traditional counterparts. Consequently, Big Data analyses need to be performed using Big Data analytics tools, and the data analysts using these tools need to be knowledgeable about their specificities and limitations.

Guidelines for processing TIM Big Data include:

- *Process the data where it is located,*
- *Use open source software,*
- *Do not reinvent the wheel, and*
- *Understand the ephemeral nature of Big Data analysis.*

6.7.1 Process the Data Where It Is Located

In the 1990s and early 2000s, it was typical to copy data to be analyzed to a new data store (e.g., a testing environment) where it could be sorted, filtered and optimized for data analysis and modeling using a specific data analytics tool. After analysis and testing, the resulting datasets and models were then moved (copied) back to the production environment where the data originated.

With Big Data analytics, quickly and easily copying or moving datasets is no longer an option. Big Data processing must be approached differently in that analyses must be run where the data

resides, without moving it, and the results typically are written to the same location. Consequently, Big Data analytics tools run directly on top of Big Data stores by moving the analytics tools through the data across multiple servers.

This data processing approach has resulted in dramatic increases in speed, quality, and usability, as well as a reduction in cost when considering the size of the datasets being processed. At the same time, this approach has introduced some difficulties. Like the data being analyzed, the analysis results are scattered across multiple servers and thus need to be accessed the same way (across multiple servers).

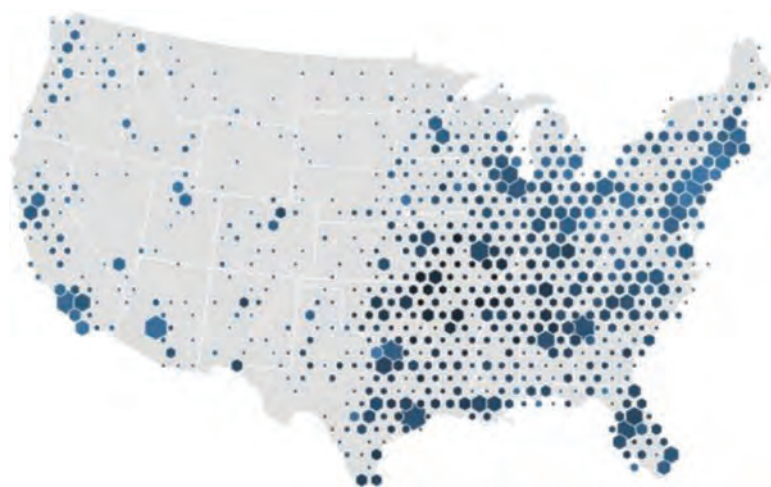
Given these access needs, Big Data post-processing tools also need to be able to access and work with large amounts of data distributed across multiple servers. Traditional visualizations like scatterplots and point maps lack the capacity to incorporate the volume of data points in Big Data results sets without turning them into unreadable charts or maps. New visualization tools, such as hexagonal bin maps and geographical heatmaps, have been designed to fit the needs of visualizing Big Data results sets (see Figure 6-3).

As a result, the research team suggests that transportation and TIM agencies developing Big Data analytics ensure that the tools they select are able to process the data where it resides, that the algorithms the tools support are designed to run on data scattered across multiple servers, and that the visualization and mapping tools being considered are capable of reading and rendering data across multiple servers.

6.7.2 Use Open-Source Software

In recent years, a quiet revolution has been taking place in the technology world. The popularity of open-source software has soared as more and more businesses have realized the value of moving away from the walled-in, proprietary technologies of old. It is no coincidence that this transformation has taken place in parallel with the explosion of interest in Big Data and analytics. The modular, fluid, and constantly evolving nature of open-source solutions is in sync with the needs of cutting-edge analytics projects for faster, more flexible, and potentially much more secure systems and platforms with which to implement them.

Open-source products are distributed under various open-source licenses. Open-source licenses grant the user the right to freely download and use products, and the products can also



Source: Bostock (2015)

Figure 6-3. Example of a hexagonal bin map.

Open-Source Can Mean Faster Fixes to Bugs and Vulnerabilities

Today it is commonly assumed that popular open-source projects are less likely than commercial closed-source software to include bugs and security vulnerabilities, and that bugs and vulnerabilities in open-source projects are likely to be found, fixed, and released faster than those in commercial software. Several conditions support this view:

- Popular open-source software typically will have many more eyes looking at it to find and fix problems. One argument used by opponents of open-source components has been that, because the code is open, it's easier for hackers to find security vulnerabilities and other weak points. The counterargument is that the same problems are likely to be discovered, faster, by "white hat" hackers, contributors (many open-source projects have hundreds or thousands of contributors), and users. Even if most open-source users are not reviewing the code when they first adopt it, they may do so if and when they encounter bugs, or when they want to modify the code to their needs.
- Open source projects typically fix vulnerabilities and release patches and new versions a lot faster. When a vulnerability in an open source project is reported—especially if it is a high-severity vulnerability—a fix often is released within a day or two. If the open-source software is developed by a commercial company, high visibility creates an urgency to fix issues, and may even lead to better code in the first place. In contrast, commercial vendors necessarily have longer update cycles.
- Realistically, nearly all commercial software now includes healthy chunks of open-source code. Modern commercial software developers do not reinvent the wheel; rather, they develop their own capabilities on top of open-source components, which often make up over 80% of the total lines of code. Thus, most commercial software is already susceptible to open source vulnerabilities. Unfortunately, many commercial vendors do not properly track and manage the security of their open-source components. As a result, fixes to bugs and vulnerabilities (including those that have been made to open-source components) can take a long time to make their way into the commercially released product. Commercial vendors may have fewer people working on a given project, and commercial vendors prioritize software updates based on commercial and financial considerations. Many commercial vendors still have release cycles of 6–12 months, so even after a vulnerability has been fixed, it may take months to release the fixed version to the market. Security researchers often complain that it can take months and even years for some vendors to address a vulnerability they have discovered. However long it takes to create and release a fix, customers remain exposed.

be modified, copied, and redistributed. Software developers can even strip out useful parts from one open-source project to use in their own products.

In the context of Big Data analytics, this approach allows software to be deployed, used, and modified at will across many servers, potentially at a much lower cost to the agency and with minimal, if any, restrictions. Open-source software can be scaled up to accommodate bursts of data-processing requests without having to request, pay for, and maintain additional licenses.

In comparison, if a proprietary software is used, the necessary flexibility would come at a significant cost, as software licenses would have to be purchased in advance to cover possible spikes and bursts in processing and future growth. By contrast, scaling up proprietary software means purchasing any necessary additional licenses—which, if overlooked, can lead to exorbitant penalties. Considering that most of the additional licenses would be used only partially, the costs to purchase them and the risk of penalties would be very difficult to justify.

Most new and emerging data management platforms have been developed in whole or in part based on open-source software (Paul 2008). The use of proprietary software by cloud customers is perceived by Big Data developers and data scientists as too risky when considering the potential for vendor lock-in, increasing fees, and the prospect of quick obsolescence.

Therefore, the research team suggests that transportation and TIM agencies adopt open-source software as a basis for their Big Data platforms. It is important to make sure that the chosen solutions are built on common architectures and possess effective, consistent commercial support. Alternatively, TIM agencies could use a cloud-based software as a service (SaaS) based on open-source software. These services currently are available from most cloud providers.

6.7.3 Do Not Reinvent the Wheel

Since the early days of Hadoop in 2001, significant focus has been given to software development to fulfill the growing needs of Big Data management and analytics. The software has progressively improved from bare-bones solutions requiring computer experts for installation and operation to turnkey cloud services that can be started with the click of a mouse. The developer communities behind this software are very active and continue to grow as new software tools and services are created. For this reason, when contemplating developing custom Big Data software solutions, transportation and TIM agencies are advised not to start from scratch. Before any development, analysts should investigate the possible existence of similar or partial solutions. More often than not, similar projects have already been started in one or more domains (e.g., healthcare, finance, advertising), and chances are that open-source software and developer communities are already supporting them. Thus, instead of attempting to develop solutions on their own, transportation and TIM agencies are encouraged to connect with these projects and communities to add their requirements, contribute to the code base, test the software with their own data, report performance and flaws, and expand them as needed. This approach will allow transportation and TIM agencies to benefit from the support of a much larger community of experts than they could gather in-house or through contracting, and could result in a significant reduction in development cost.

6.7.4 Understand the Ephemeral Nature of Big Data Analytics

An important aspect of Big Data analytics is its ephemeral nature. The five Vs of Big Data have overwhelmed traditional hardware and pushed the adoption of what has come to be called disposable commodity hardware. Software in a Big Data environment also needs to be implemented in a “disposable” fashion. This is particularly relevant in the context of analytics and predictions, because the rapid changes occurring within the datasets can quickly lower the performance or quality of recently developed analytical components. To avoid this pitfall, it is best not to develop Big Data analytical solutions using a “set and forget” approach that assumes the analytical solution will be able to perform well for years to come. Instead, a more interactive approach to solution development needs to be adopted. This approach involves constantly monitoring the analytics results and redesigning the system as needed as soon as performance and quality begin to drop. The interactive approach is already being used in the commercial cloud industry. In online advertising, for example, machine learning models

predict the various ads that website visitors will be interested in seeing. Because the predictions lose accuracy within days or hours, the models are constantly retrained to maintain prediction accuracy over time.

6.8 Open and Share Outcomes and Products to Foster Data User Communities

Lastly, the research team suggests that TIM agencies open and share the results of their Big Data analyses. Unless sharing the data or analysis results would pose potential risks to privacy or security, the trends, patterns, models, visualizations, and outliers discovered through Big Data analytics can be shared directly with a broader community of agencies through common data storage. As the results are reviewed and analyses are recreated by other members of the community, better outcomes from these analyses will emerge as successes, flaws, errors, or previously undetected patterns. Previously unseen ways to leverage the data will more likely be discovered by a broad community than by a small set of experts involved in the development of the analysis. Ideally, not only data, but also analytical code, models, and visualizations would be shared.

Big Data datasets are becoming increasingly large and complex, and the recent adoption of connected vehicle and IoT technologies will only make for larger and more complex datasets. Without the adoption of a distributed approach to involve many “eyes” in mining this data, it is likely that many of the valuable patterns and correlations present in the data may go undetected. Transportation and TIM agencies are encouraged to support the development of data user communities drawn from government employees, government contractors, universities, the private sector, and citizens in order to form a continuously evolving collaborative environment that is able to maximize the value of its Big Data datasets.

Transportation agencies are encouraged to support the development of data user communities.



CHAPTER 7

Summary and Next Steps

This report has addressed the NCHRP Project 17-75 research objectives, which were to (1) describe and assess current and emerging sources of data that could improve TIM, (2) describe potential opportunities to leverage Big Data that could advance the TIM state of the practice, (3) identify potential challenges for TIM agencies to leverage Big Data, and (4) develop Big Data guidelines for TIM agencies. The sections in this chapter summarize the findings of the research, set forth potential next steps for the research findings, and address recommendations, needs, and priorities for additional related research.

7.1 Summary of Findings

The state of the practice in TIM shows significant advancement over the past decade, most notably through the development of regional and statewide TIM committees, the National TIM Responder Training Program efforts, the implementation of TIM legislation, and the collection and analysis of TIM data for performance measurement. Recent guidance, like that provided through efforts of the TRB and the FHWA—including FHWA’s ongoing “Every Day Counts” (EDC) initiative, now beginning its fifth round—reflects national efforts to advance the collection and use of TIM data.

The findings from a review of the state of the practice in Big Data reinforce awareness that:

- Big Data is not new; rather, Big Data technologies and techniques have been applied for nearly two decades by various companies;
- Although Big Data is characterized by the five Vs of volume, velocity, variety, veracity, and value, not all datasets need to possess all five of these qualities to be considered Big Data;
- Contrary to the relational database approach, Big Data analytics is not bound to a single set of tools to perform analyses; rather, Big Data analytics encompass a wide variety of proprietary and open-source tools that can be customized and modified by users;
- The tools used for Big Data analytics allow for the rapid transfer, processing, storage, and analysis of extremely large datasets, have increased the ability to analyze divergent data (e.g., decades-old historical records and real-time streaming data), and make it possible to derive value from data that cannot be attained using traditional data mining approaches that typically rely on relational databases.

Big Data applications in the field of transportation are more recent, having occurred within the past few years, and include applications in the areas of planning, parking, trucking, public transportation, operations, ITS, and other more niche areas. A significant gap exists between the current state of the practice in Big Data analytics (e.g., image recognition, graph analytics) and the state of DOT applications of data for TIM (e.g., manual use of Waze data for incident detection).

A few TIM Big Data applications were identified, but these were largely applications that could be performed using relational databases. Local data and state data generally are not collected at the volumes that make using or applying Big Data approaches practical. Ways are available to expand on these initial approaches to Big Data for TIM, but the data must first be prepared, must be of sufficient size, and must cover a sufficient length of time to identify meaningful patterns and yield value.

Big Data applications offer significant opportunities to improve TIM, as highlighted in Chapter 4 through contrasts made between traditional and Big Data approaches to common areas of TIM concerns. These example Big Data applications illustrate that, beyond offering improvement on current practices, the Big Data approach represents a radical change from traditional approaches. Big Data represents a paradigm shift that goes beyond data collection and analysis to include data storage, management, and security; the financial planning and procurement of IT services; the required skillsets of employees; and beyond. Opportunities to apply Big Data to TIM at a regional or state level are currently limited by the collection and availability of data and the capability maturity of analysts.

Although a few existing Big Data datasets (e.g., data available from HERE, INRIX, and Waze) might be immediately leveraged for TIM, these datasets alone lack the detail needed for effectively mining and understanding the nuances of incident response and TIM, and access to the raw data remains limited. Many of the benefits of Big Data analytics for TIM will require collecting and integrating more TIM-specific, detailed data (e.g., crash data or CAD data), at minimum at a state level if not at the national level, to establish sufficient volume and variety for uncovering relationships and insights. Discussing these opportunities now can help agencies identify the low-hanging fruit for Big Data in TIM, and will help agencies see the benefits of taking the next steps toward undertaking a TIM Big Data initiative.

The research identified many challenges and potential barriers that could impact the application of Big Data for TIM. At the forefront of these challenges are aspects of organizational culture—specifically, challenges that impede agencies' willingness and ability to embrace the paradigm shift that Big Data requires. Reluctance to open and share data, as well as impediments that stand in the way of using cloud infrastructure, are two central factors that will limit the growth and application of Big Data within an organization.

The application of Big Data also requires sensitivity to organizational capabilities. The level of technical expertise among existing TIM stakeholders at local, regional, and state agencies will likely vary widely, with the result that the skills and resources needed to close the gap between current data practices and Big Data practices may not be sufficient to comfortably and efficiently apply Big Data. Further, individuals who have Big Data expertise are in limited supply and in high demand, which may hamper agencies' ability to train or hire talent and purchase the requisite resources.

Fundamental to Big Data analytics is having access to large amounts of varied data. An assessment of 31 different data sources showed that a large gap exists between the current state of TIM-related data and the application of this data for Big Data analytics. Although merging a few datasets may be tenable for agencies, building the large, highly detailed, integrated datasets needed for Big Data will require significant resources, as well as the expertise to apply non-traditional approaches. Challenges such as the lack of standards for data collection and storage, PII, legal restrictions, and agency culture and policies will limit the application of Big Data for TIM. Furthermore, although millions of data points are generated every second by traffic sensors and probes, incidents are infrequent by nature and therefore relatively small in number. This limits the application of Big Data to TIM unless the data is aggregated across multiple regions and organizations to increase its volume and variety.

The research suggests that the current state of the practice for TIM data collection, storage, and analysis is between the first and second tiers on the Big Data pyramid. At this point, very limited TIM data is being collected and shared among partner agencies, and a solid data lake as a foundation for the development of TIM business intelligence and TIM data science has yet to be built. Therefore, based on the research findings, guidelines for transportation and TIM agencies were developed to lay out the various changes that will be necessary for agencies to develop a usable Big Data store (data lake), implement agency-wide analytics and business intelligence, and pursue the development of an evolving and beneficial data science environment. Expressed at their highest level, the guidelines suggest that agencies prepare to:

- Adopt a deeper and broader perspective;
- Collect more data;
- Open and share data;
- Use a common data storage;
- Adopt cloud technologies for the storage and retrieval of data;
- Manage the data differently;
- Process the data; and
- Open and share outcomes and products to foster data user communities.

By embracing these guidelines and the actions suggested to accompany them, agencies can address and overcome the challenges that limit the move toward the use of Big Data for TIM. Applying these guidelines will thus help position transportation and TIM agencies for Big Data.

7.2 Next Steps

Agencies are encouraged to begin following the guidelines and putting the research into practice by fully embracing low-cost, traditional best practices in data collection, cleaning, warehousing, and analysis with existing data sources. Agencies also are encouraged to concurrently identify opportunities to ready their organizations for Big Data. Opening and sharing data—both internally and externally—are critical cultural shifts that need to be embraced. An incremental approach is recommended that begins with developing the culture, policies, and expertise to improve the usability and increase the use of current data, and that captures opportunities to migrate from in-house servers to the cloud. These steps form the basis for positioning agencies to begin capitalizing on the opportunities afforded by Big Data.

Migrating research and guidelines from ideas into practice can begin by linking research results and outputs to related products and by engaging stakeholders. This endeavor can be enhanced by using the TIM Performance Measurement (TIM PM) Website (<http://nchrptimpm.timnetwork.org/>). A product of NCHRP Project 07-20, “Guidelines for the Implementation of TIM Performance Measurement,” TIM PM offers a natural location to obtain and share TIM information. NCHRP Project 07-20 was the first to offer a standardized set of TIM data elements, as well as a standardized way to organize these data elements in a database schema, so that TIM performance could be measured and analyzed. The website is a natural location at which to include information on Big Data as a next step in the application of data to improve TIM.

The FHWA’s TIM Capability Maturity Self-Assessment (TIM CMSA) (https://ops.fhwa.dot.gov/tsmoframeworktool/available_frameworks/traffic_incident.htm) tool is used by state and local TIM program managers to benchmark and evaluate TIM program strengths, weaknesses, successes, and areas for improvement on an annual basis, and to aid in the development of a targeted action plan for TIM. Data collection, integration, and sharing is a key part of the TIM CMSA. Given that Big Data is in everyone’s future, agencies should have an opportunity to

assess themselves on the foundational principles associated with readying their organizations for Big Data.

The EDC-2 National TIM Responder Training Program and the post-course assessment tool that was developed under SHRP 2 Project L32(C) offer other opportunities to incorporate effective guidance on the importance of good data collection and sharing practices and the understanding of how the data can help to improve TIM by informing decision-making, resource utilization/management, real-time TIM activities, and program funding decisions.

Finally, many states have statewide and/or regional TIM coalitions or committees that include participation and representation from the various TIM stakeholder disciplines. Regular coalition or committee meetings provide an opportunity for stakeholders to discuss TIM practices, share lessons learned, and discuss ways to improve TIM. These meetings also offer opportunities to introduce concepts that connect Big Data to TIM, embedding the knowledge at the responder level across the various disciplines. Receptiveness or interest may vary across responder communities, but the research for this project has made it clear that some responders recognize the importance of data and are willing to take pertinent information to their upper management. Presenting tangible examples of Big Data applications and outcomes specific to TIM operations can help spur interest and motivation to take action.

7.3 Suggestions and Priorities for Additional Related Research

Big Data is here, and transportation agencies are encouraged to begin embracing the changes required to tackle it. Traditional organizational cultures and lack of data may be holding back full acceptance and adoption of the foundational principles of Big Data, but the emergence of connected vehicle, traveler, and infrastructure data will soon be driving the change. To capitalize on the wealth of information that can be derived from these and other data sources—and to prevent system failures caused by data overload—transportation agencies must ready themselves for Big Data. The technology is here, the tools are available, and the expertise can be found to assist transportation agencies in both understanding and applying these technologies and tools to everyday questions and problems.

Effective strategies and techniques are needed to recognize and break down some of the barriers that still impede agencies' understanding and adoption of Big Data technologies. Additional research could help agencies find ways to overcome cultural barriers to opening and sharing data, and to resolve legal or proprietary concerns. Once transportation and partner agencies have collected, opened, shared, and pooled enough (and varied) data in a cloud environment, further research can then be conducted using Big Data techniques to discover how Big Data can help to improve specific components of TIM programs.



Abbreviations

AAMVA	American Association of Motor Vehicle Administrators
Arizona DOT	Arizona Department of Transportation
AFC	Automated Fare Collection
AHMCT	Advanced Highway Maintenance and Construction Technology Research Center
ALPR	Automatic License and Plate Reader/Recognition
APCO	Association of Public-Safety Communications Officials
App	Application
APTRA	Arizona Professional Towing and Recovery Association
ATMS	Advanced Traffic Management Systems
AVL	Automatic Vehicle Location
AWS	Amazon Web Services
AZDPS	Arizona Department of Public Safety
BDE	BigDataEurope
CAD	Computer-Aided Dispatch
CATT Lab	Center for Advanced Transportation Technology Laboratory
CCP	Connected Citizens Program
CCTV	Closed Circuit Television
Colorado DOT	Colorado Department of Transportation
CDR	Call Detail Records
CHP	California Highway Patrol
CMS	Changeable Message Sign
CMSA	Capability Maturity Self-Assessment
CPU	Central Processing Unit
C.R.A.S.H.	Crash Reduction Analyzing Statistical History
CSV	Comma-Separated Value
DaaS	Data-as-a-Service
DAISy	Data Analytics Intelligence System
DCM	Data Capture and Management
DOT	Department of Transportation
DSRC	Dedicated Short-Range Communications
ECC	Emergency Communications Center
EDC	Every Day Counts
EDR	Event Data Recorder
EMS	Emergency Medical Service(s)
ESS	Environmental Sensor Station
ETL	Extract-Transform-Load
FARS	Fatality Analysis Reporting System

FDE	Fundamental Data Elements
FDEM	Florida Department of Emergency Management
Florida DOT	Florida Department of Transportation
FHP	Florida Highway Patrol
FOIA	Freedom of Information Act
FTP	File Transfer Protocol
GPS	Global Positioning System
GPU	Graphics Processing Unit
GTFS	General Transit Feed Specification
HDFS	Hadoop Distributed File System
HIPAA	Health Insurance Portability and Accountability Act
ICIJ	International Consortium of Investigative Journalists
ICT	Incident Clearance Time
IoT	Internet of Things
IR	Incident Response
IRCO	Incident Response and Clearance Ontology
ITF	International Transport Forum
ITS	Intelligent Transportation Systems
JOPS	Joint Operations Policy Statement
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
KPM	Key Performance Measure
LODE	Live Owl Documentation Environment
MADIS	Meteorological Assimilation Data Ingest System
MAG	Maricopa Association of Governments
MCMIS	Motor Carrier Management Information System
MIRE	Model Inventory of Roadway Elements
MMUCC	Model Minimum Uniform Crash Criteria
MPO	Metropolitan Planning Organization
MTA	Metropolitan Transit Authority
MVDS	Microwave Vehicle Detection System
NASEMSO	National Association of State Emergency Medical Services Officials
NCEP	National Centers for Environmental Prediction
NCO	National Central Operations
NDS	Naturalistic Driving Study
NEMSIS	National Emergency Medical Services Information System
NFIRS	National Fire Incident Reporting System
NITTEC	Niagara International Transportation Technology Coalition
NOAA	National Oceanic and Atmospheric Administration
NPMRDS	National Performance Management Research Data Set
NWS	National Weather Service
NYCTA	New York City Transit Authority
O/D	Origin-Destination
ODI	Open Data Institute
Oregon DOT	Oregon Department of Transportation
OLAP	Online Analytical Processing
OLTP	Online Transactional Processing
OSP	Oregon State Police
PDR	Public Data Release
PII	Personally Identifiable Information
PSAP	Public Safety Answering Point

RCT	Roadway Clearance Time
RDBMS	Relational Database Management System
RFID	Radio Frequency Identification
RITIS	Regional Integrated Transportation Information System
ROC	Rio de Janeiro Operations Center
RWIS	Road Weather Information Systems
RWMP	Road Weather Management Program
SaaS	Software-as-a-Service
SHRP	Strategic Highway Research Program
SHSP	Strategic Highway Safety Plan
SOP	Standard Operating Procedures
SQL	Simple Query Language
SSP	Safety Service Patrol
TAC	Traffic Assistance Center
TfL	Transport for London
THP	Tennessee Highway Patrol
TIM	Traffic Incident Management
TIM-BC	Traffic Incident Management Benefit-Cost
TIMELI	Traffic Incident Management Enabled by Large-data Innovations
TIM PM	Traffic Incident Management Performance Measurement
TMC	Traffic Management Center
TOC	Traffic Operations Center
TRCC	Traffic Records Coordinating Committee
TPEG	Transport Protocol Experts Group
UAV	Unmanned Aerial Vehicle
UBI	Usage Based Insurance
Utah DOT	Utah Department of Transportation
V2C	Vehicle-to-Cloud
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
VASTIM	Virginia Statewide Traffic Incident Management Committee
VEACON	Vehicular Accident Ontology
VTTI	Virginia Tech Transportation Institute
W3C	World Wide Web Consortium
WITS	Washington Incident Tracking System
WMV	Windows Media Video
WxDE	Weather Data Environment
XML	Extensible Mark-up Language



Glossary

Batch Processing	<i>Batch processing</i> refers to a computer working automatically through a queue or batch of separate jobs or programs in a non-interactive manner.
Big Data	<i>Big Data</i> is data that traditional data management systems cannot manage due to its size and complexity.
Big Data Store	A <i>Big Data store</i> (or <i>data lake</i>) is a collection of repositories where very large raw datasets are stored and can be processed. A Big Data store differs from a traditional data warehouse, which is designed for historical analysis using relational databases.
Binning	<i>Binning</i> is the sorting of individual data into categories and representing data by their categories.
Cluster Analysis	<i>Cluster analysis</i> is the analysis of data to determine which groups of data are close together or similar to each other.
Crowdsourced Data	<i>Crowdsourced data</i> is data that is actively or passively collected from a very large number of individuals or organizations.
Cryptographic Hash	A <i>cryptographic hash</i> is the result of a computer process that converts a data input, such as a message, into a fixed-size alphanumerical sequence, preserving the uniqueness of the original data input while making it very difficult to convert it back to its original form.
Data Lake	A <i>data lake</i> is a very large collection of raw and unfiltered data that has not been altered from its original form before being stored. The term <i>data lake</i> is sometimes used synonymously with <i>Big Data store</i> .
Data Latency	<i>Data latency</i> is the time required for data to be stored or retrieved from a data store or database.
Data Maturity	<i>Data maturity</i> is the measure of how readily data can be used.
Data Model	A <i>data model</i> is an abstract model that organizes elements of data and standardizes their properties and how they relate to one another.
Data Store	A <i>data store</i> is a repository used for storing collections of data more complex than data tables.
Data Throughput	<i>Throughput</i> is the amount of data that can be moved safely through a data processing system.

Database Schema	A <i>database schema</i> is a type of data model used to organize data inside a relational database.
Distributed Computing	<i>Distributed computing</i> is a model in which components of a software system are shared among multiple computers to improve efficiency and performance.
Document-Oriented Database	A <i>document-oriented database</i> is a type of non-relational data store designed specifically for storing, retrieving, and managing document-oriented information such as crash records, loan applications, shopping carts, and so forth.
Extract-Transform-Load	ETL (<i>extract-transform-load</i>) is the process used to populate data into a relational database system, where raw data and unfiltered data are extracted from data sources, transformed into a usable format, and loaded into a final database.
Fault Tolerance	<i>Fault tolerance</i> is the property that enables a system to continue operating properly in the event of a failure.
GPU-Accelerated Database	A <i>GPU-accelerated database</i> is one that leverages a graphical processing unit (GPU) instead of a traditional central processing unit (CPU) to significantly increase its performance.
Graph Analysis	<i>Graph analysis</i> (also called <i>network analysis</i>) is a data analysis method that seeks to analyze data structured into a set of interconnected vertices and edges (a graph or a network). Graph analysis is commonly used in social media data analysis.
Graph Database	A <i>graph database</i> is a database that uses graph structures to represent, store, and query data.
Hadoop	<i>Hadoop</i> is an open-source distributed processing framework that manages data processing and storage for big data applications running on distributed commodity computer systems.
Key-Value Store	A <i>key-value store</i> (or key-value database) is one of the simplest forms of NoSQL databases designed to store and query data pairs expressed as keys and values.
Machine Learning	<i>Machine learning</i> is a subset of artificial intelligence that often uses statistical techniques to give computers the ability to “learn” with data without being explicitly programmed.
Mesonet	In meteorology (and climatology), a <i>mesonet</i> (mesoscale network) is a network of (typically) automated weather and environmental monitoring stations designed to observe mesoscale meteorological phenomena.
Metadata	<i>Metadata</i> is a set of data that describes and provides specific information about other data. Author, date created, date modified, and file size are examples of very basic document metadata. In word processing files, such basic metadata typically can be seen under “file properties.” Some types of metadata are generated automatically by software, and other types may be added to files as needed or desired. Collectively, the various types of metadata facilitate finding, organizing, identifying, using, archiving, and preserving digital resources.

NetCDF	<i>NetCDF</i> is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. NetCDF was originally developed by NASA and is now maintained by the University Corporation for Atmospheric Research.
Neural Network	A <i>neural network</i> is a form of machine learning that uses statistical techniques patterned after the operation of neurons in the human brain.
NewSQL	<i>NewSQL</i> is a class of modern relational database management systems (RDBMSs) that seek to provide the same scalable performance of NoSQL systems while maintaining some of the properties of a traditional database system.
NoSQL	<i>NoSQL</i> databases are non-RDBMSs that can accommodate a wide variety of data models, including key-value, document, columnar, and graph formats.
Ontology	In computer science, an <i>ontology</i> is a formal representation, formal naming, and definition of the categories, properties, and relations between concepts and data within a specific domain. An ontology is the data model used to organize data within a graph database.
Open Data	<i>Open data</i> is data that can be freely used, re-used, and redistributed by anyone, subject only and at most to the requirement to attribute and share alike.
Open-Source	The compound adjective “ <i>open-source</i> ” is used to describe software that people can freely copy, modify, and share because its design has been made publicly accessible. The term originated in the context of software development to designate a specific approach to creating computer programs.
Overfitting	<i>Overfitting</i> is a modeling error that occurs when a statistical model is too closely fit to a limited set of data points.
Patrol Beat	In police terminology, a <i>patrol beat</i> is the territory and time that a police officer patrols.
Relational Database	A <i>relational database</i> is a type of database organized as a collection of data items that are interconnected by pre-defined relationships (data schema). These data items are organized as a set of tables with columns and rows. Tables are used to hold information about the objects to be represented in the database. Each column in a table holds a certain kind of data, and a field stores the actual value of an attribute. In a relational database, data can be accessed in many ways without having to reorganize the database tables themselves.
Semi-structured Data	<i>Semi-structured data</i> is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. A good example of semi-structured data is an HTML page, in which text and images are structured using a hierarchy of tags.

Server Clusters	A <i>server cluster</i> , or <i>computer cluster</i> , is a set of connected computers that work together so that, in many respects, the cluster can be viewed as a single system. Computer clusters are used to increase performance and reliability when dealing with very large dataset processing.
Structured Data	<i>Structured data</i> is data that has been organized and formatted according to a specific data model.
Stream Processing	<i>Stream processing</i> , or <i>data stream processing</i> , is a type of data processing in which operations are performed on each individual datum sequentially as it becomes available. Stream processing processes data in real time as the data arrives. This approach contrasts with <i>batch processing</i> , in which data is first stored, then processed together in batches at regular intervals (for example, nightly).
Telematics	<i>Telematics</i> is a term that combines the words <i>telecommunications</i> and <i>informatics</i> to broadly describe the integrated use of communications and information technology to transmit, store, and receive information from telecommunications devices to remote objects over a network.
Unstructured Data	<i>Unstructured data</i> is data that is not organized in a pre-defined data model.
Value of Data	The <i>value of data</i> is the ability of data in a database to support business processes. Value is one of the five Vs of Big Data.
Variety of Data	The <i>variety of data</i> is the heterogeneity of data stored in a database. Variety is one of the five Vs of Big Data.
Velocity of Data	The <i>velocity of data</i> is the frequency with which new data is created in a database. Velocity is one of the five Vs of Big Data.
Veracity of Data	The <i>veracity of data</i> is the reliability of data in a database. Veracity is one of the five Vs of Big Data.
Volume of Data	The <i>volume of data</i> is the quantity of data that can be stored in a database. Volume is one of the five Vs of Big Data.
Wide Column Database	A <i>wide column database</i> is a type of NoSQL database that uses tables, rows, and columns to organize data, but unlike a relational database, allows for the names and format in each column to vary from row to row within the same table.



References

- Alvarez, P. 2015. "Big Data Helps Pedestrian Planning Take a Big Step Forward." *Transportation Professional*. Barrett, Byrd Associates, Tunbridge Wells, Kent, UK.
- Amazon. 2017. "Amazon Rekognition." AWS. Webpage: <https://aws.amazon.com/rekognition/> (accessed June 2017).
- Amazon Web Services. n.d. "C-SPAN Case Study." AWS Website: <https://aws.amazon.com/solutions/case-studies/cspan/> (accessed July 24, 2018).
- American Automobile Association. 2017. *AAA Digest of Motor Laws*. Webpage: <http://drivinglaws.aaa.com/tag/move-over-law/> (accessed May 15, 2017).
- Anderson, J. M., N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, and O. A. Oluwatola. 2016. *Autonomous Vehicle Technology*. Santa Monica, CA: RAND Corporation. Available online: http://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR443-2/RAND_RR443-2.pdf.
- APCO International. 2010. *High Priority Information Sharing Needs for Emergency Communications and First Responders*. Association of Public Safety Communications Officials, Daytona Beach, FL. Available online: <https://www.apcointl.org/doc/911-resources/375-high-priority-info-sharing-needs-for-emerg-comm-and-first-responders-final-pdf/file.html>.
- Applied Engineering Management Corp. and toXcel, LLC. 2018. *NCHRP Research Report 865: Guidance for Development and Management of Sustainable Enterprise Information Portals*. Transportation Research Board, Washington, D.C.
- Automotive Supply Chain. 2016. *Automotive Supply Chain*. Webpage: <http://automotivesupplychain.org/supply-chain/crown-commercial-service-appoints-ald-automotive-to-vehicle-telematics-framework-agreement/> (accessed July 2018).
- Baltimore Metropolitan Council. 2017. "Traffic Incident Management Committee" webpage: <https://baltometro.org/community/committees/traffic-incident-management-committee> (accessed June 2017; url updated May 2019).
- Barichello, K., and S. Knickerbocker. 2017. "Using INRIX Data in Iowa." *www.mtmug.org*. Webpage: http://www.mtmug.org/Presentations/rev_INRIX%20presentation%20MTMUG.pdf (accessed May 15, 2017).
- Barrachina, J., P. Garrido, M. Fogue, and F. J. Martinez. 2012. "VEACON: a Vehicular Accident Ontology Designed to Improve Safety on the Roads." *Journal of Network and Computer Applications*. Available online: https://www.researchgate.net/publication/236842047_Efficient_Regression_Testing_of_Ontology-Driven_Systems.
- BDE n.d. "Big Data Europe Empowering Communities with Data Technologies—Integrating Big Data, Software, and Communities for Addressing Europe's Societal Needs." *Big Data Europe*. Webpage: <https://drive.google.com/file/d/0By9UYHuCedbmjJKZF9YQkROcG8/view> (accessed June 10, 2017).
- BDE and ERTICO-ITS Europe. 2015. *Big Data Europe for Smart, Green and Integrated Transport*. First Workshop Report, ITS Conference (Bordeaux, Oct. 7, 2015).
- Beach, J. 2014. "Big Data & Trucking." *Trucking Info*. Webpage: <http://www.truckinginfo.com/article/story/2014/12/big-data-tracking.aspx> (accessed November 2016).
- Birenbaum, I., C. Creel, and S. Wegmann. 2009. *Traffic Control Concepts for Incident Clearance*. FHWA-HOP-08-057. Office of Transportation Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C. Available online: <https://ops.fhwa.dot.gov/publications/fhwahop08057/fhwahop08057.pdf>.
- Bonetti, P. 2013. "How to Really Outsmart Traffic." *HERE 360*. Webpage: <http://360.here.com/2013/07/09/how-to-really-outsmart-traffic/> (accessed June 2017).
- Bostock, M. 2015. "Bivariate Hexbin Map—Released Under the GNU General Public License, Version 3." *Mike Bostock's Blocks*. Webpage: <https://bl.ocks.org/mbostock/4330486> (accessed October 19, 2018).

- Branch, A. 2016. "Victims ID'd in Fatal Quassy Amusement Park Crash: Report." *Patch*. Webpage: <https://patch.com/connecticut/woodbury-middlebury/victims-idd-fatal-quassy-amusement-park-crash-report> (accessed July 2018).
- Brickley, D. n.d. "W3C Semantic Web Interest Group." W3C. Webpage: <https://www.w3.org/2003/01/geo/>.
- Broad, E. (2015) "Closed, Shared, Open Data: What's in a Name?" Blog (September 17, 2015): <http://oldsite.theodi.org/blog/closed-shared-open-data-whats-in-a-name>.
- Brooke, K., K. Dopart, T. Smith, A. Flannery. 2004. *NCHRP Report 520: Sharing Information Between Public Safety and Transportation Agencies for Traffic Incident Management*. Transportation Research Board of the National Academies, Washington, D.C.
- BTS. 2011. "Clarus." *Bureau of Transportation Statistics*. Webpage: https://ntl.bts.gov/lib/44000/44300/44374/FHWA-JPO-11-154_Clarus_Overview_final.pdf (accessed February 2017).
- Burt, M., M. Cuddy, and M. Razo. 2014. *Big Data's Implications for Transportation Operations: An Exploration*. FHWA-JPO-14-157. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Carrick, G., S. Srinivasan, and I. Bejleri. 2015. "Analysis and Characterization of Secondary Traffic Crashes in Florida." Presented at 94th Annual Meeting of the Transportation Research Board. Transportation Research Board, Washington, D.C.
- Carrick, G., K. Jermprapai, S. Srinivasan, and Y. Yin. 2017. "Development of Guidance for Safety Service Patrol Deployment Decisions in Florida." *Transportation Research Record: Journal of the Transportation Research Board*, 2660(1), Transportation Research Board of the National Academies, Washington, D.C., pp. 48-57. <https://doi.org/10.3141/2660-07>.
- Carrick, G., and S. Washburn. 2012. "The Move Over Law: Effect of Emergency Vehicle Lighting in Driver Compliance on Florida Freeways." *Transportation Research Record: Journal of the Transportation Research Board*, 2281. Transportation Research Board of the National Academies, Washington, D.C., pp. 1-7. <https://doi.org/10.3141/2281-01>.
- Carson, J. L. 2008. *Traffic Incident Management Quick Clearance Laws: A National Review of Best Practices*. FHWA-HOP-09-005. Office of Transportation Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Carson, J. L. 2010. *Best Practices in Traffic Incident Management*. FHWA-HOP-10-050. Office of Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- CATT Lab. 2015. "RITIS Platform, Features & Applications Overview." Webpage: <http://www.cattlab.umd.edu/files/RITIS%20Overview%20Book-2-2-15%20FINAL.pdf>.
- Center for Advanced Automotive Technology. n.d. "Connected and Automated Vehicles." *Center for Advanced Automotive Technology*. Webpage: http://autocaat.org/Technologies/Automated_and_Connected_Vehicles/ (accessed February 2017).
- Chire. 2011. "Single-linkage cluster analysis." Image, used under a Creative Commons Attribution-Share Alike 3.0 (CC BY-SA 3.0) Unported License." *Wikipedia*. Available online: https://en.wikipedia.org/wiki/Cluster_analysis#/media/File:SLINK-Gaussian-data.svg (accessed July 24, 2018).
- Clark, M., I. Turnbull, T.A. Lasky, and S. Donecker. 2016. "Responder System Allows Caltrans First Responders to Collect and Share At-Scene Information Quickly and Efficiently." *California Department of Transportation*. Webpage: http://www.dot.ca.gov/research/rural/docs/responder_phase_III_summary-2016-07-12.pdf (accessed March 2017).
- Colak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiretta, and M. Gonzalaz. 2014. "Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities." Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C.
- Connected Citizens Program. 2016. *Connected Citizens Program*. Webpage: https://wiki.waze.com/wiki/Connected_Citizens_Program (accessed June 2017).
- Corbin, J. 2008. "A National Unified Goal for Traffic Incident Management (TIM): What Is it, and Why Is it Needed." Presentation for ITS PCB T3 Webinar, September 11, 2008. National Traffic Incident Management Coalition, American Association of State Highway and Transportation Officials, Washington, D.C. Available online: <https://www.pcb.its.dot.gov/t3/s080911/corbin.pdf> (accessed June 2017).
- Cox, S., and C. Little. 2017. "Time Ontology in OWL." W3C. Webpage: <https://www.w3.org/TR/owl-time/> (accessed July 2017).
- Daniell, J. N. 2009. *Traffic Incident Management in Hazardous Materials Spills in Incident Clearance*. FHWA-HOP-08-058. Office of Transportation Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Dardailleur, D. 2012. "Draft Road Accident Ontology." *World Wide Web Consortium (W3C)*. Webpage: <http://www.w3.org/2012/06/rao.html#graph> (accessed May 2017).
- Delgado, R. 2017. "Big Data's Impact on Public Transportation. How Can Data Help Cities Build Transport Infrastructure?" *The Innovation Enterprise, Inc.* Webpage: <https://channels.theinnovationenterprise.com/articles/big-data-s-impact-on-public-transportation> (accessed May 2017).

- Dong, H., M. Wu, L. Jia, and X. Zhou. 2015. "Traffic Zone Division Based on Big Data from Mobile Phone Base Stations." *Transportation Research, Part C: Emerging Technologies*, Vol. 58, Part B. Elsevier, Ltd., London, UK, pp. 278-291.
- Drow, M., P. Lange, and B. Laufer. 2015. "Big Progress in Big Data." *International Parking Institute*. Webpage: <http://www.parking.org/2016/01/26/tpp-2015-02-big-progress-in-big-data> (accessed February 2016).
- Dunn Engineering Associates. 2006. *Alternate Route Handbook*. FHWA-HOP-06-092. HOTO-1, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Durkin, E. 2015. "NYC Explores Drone Use for Everything from Disaster Response to Traffic Surveillance." *New York Daily News* (November 23). Available online: <http://www.nydailynews.com/new-york/nyc-explores-drone-disaster-response-traffic-jams-article-1.2444023> (accessed June 2017).
- Einstein, N., and J. Luna. 2018. *SHRP2 Traffic Incident Management Responder Training Program: Final Report*. FHWA-HRT-18-038. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- FEMA. 2011. *National Fire Incident Reporting System, Version 5.0: Fire Data Analysis Guidelines and Issues*. U.S. Department of Homeland Security, Washington, D.C. Webpage: https://www.usfa.fema.gov/downloads/pdf/nfirs/nfirs_data_analysis_guidelines_issues.pdf.
- FHWA. n.d.-a. "National Traffic Incident Management Responder Training—Web Based." *National Highway Institute*. Webpage: https://www.nhi.fhwa.dot.gov/course-search?tab=0&key=133126&sf=0&course_no=133126A (accessed June 2017; url updated July 2019).
- FHWA. n.d.-b. *Roadway Safety Data Program*. Webpage: <https://safety.fhwa.dot.gov/rsdp/mire.aspx> (accessed June 2017).
- FHWA. 2012. *National TIM Responder Training Program*. Webpage: https://www.fhwa.dot.gov/goshrp2/Solutions/Reliability/L12_L32A_L32B/National_Traffic_Incident_Management_Responder_Training_Program (accessed June 2017).
- FHWA. 2013a. "National Performance Management Research Data Set (NPMRDS) Technical Frequently Asked Questions." *Freight Management and Operations*. Webpage: https://ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm (accessed June 2017).
- FHWA. 2013b. *National Traffic Incident Management Responder Training Program: Train-the-Trainer Guide*. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- FHWA. 2017a. "Surveillance, Monitoring, and Prediction" *Road Weather Management Program*. Webpage: https://ops.fhwa.dot.gov/weather/mitigating_impacts/surveillance.htm#esrw (accessed February 2017).
- FHWA. 2017b. "Traffic Incident Management Benefit-Cost (TIM-BC) Tool." *Federal Highway Administration/Software*. Webpage: <https://www.fhwa.dot.gov/software/research/operations/timbc/> (accessed July 24, 2018).
- FHWA. 2017c. *Traffic Incident Management Knowledgebase*. Webpage: https://ops.fhwa.dot.gov/eto_tim_pse/Preparedness/tim/knowledgebase/index.htm (accessed August 2018).
- Florida Department of Transportation. 1996. "TIM Teams." *Traffic Incident Management*. Webpage: <http://www.floridatim.com/Teams.htm> (accessed June 2017).
- Florida Department of Transportation. 2011. "Sunguide Disseminator." *Florida Department of Transportation*. Webpage: <http://www.fdot.gov/traffic/Newsletters/2011/2011-001-Jan.pdf> (accessed June 2017).
- Florida Department of Transportation. 2016. "NPMRDS." *Florida Department of Transportation*. Webpage: [http://www.fdot.gov/planning/statistics/multimodaldata/multimodal/National%20Performance%20Management%20Research%20Data%20Set%20\(NPMRDS\).pdf](http://www.fdot.gov/planning/statistics/multimodaldata/multimodal/National%20Performance%20Management%20Research%20Data%20Set%20(NPMRDS).pdf) (accessed April 2017).
- Freeze, B. 2017. "Predictive Analytics for TDOT HELP." Presentation at AASHTO STSMO Meeting (September 14, 2017, Rapid City, SD).
- Gantz, J.F. 2007. *The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010*. International Data Corporation, Framingham, MA. Available online: https://www.tobbb.org.tr/BilgiHizmetleri/Documents/Raporlar/Expanding_Digital_Universe_IDC_WhitePaper_022507.pdf.
- GenCore Candeo, Ltd. 2017. *Genesis PULSE in Partnership with Waze*. Webpage: <https://genesispulse.com/features/waze/> (accessed June 2017).
- Gettman D., A. Toppen, K. Hales, A. Voss, S. Engel, and D. El Azhari. 2017. *Integrating Emerging Data Sources into Operational Practice—Opportunities for Integration of Emerging Data for Traffic Management and TMCs*. FHWA-JPO-18-625 (Final Report). Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Goodin, D. 2014. "Poorly Anonymized Logs Reveal NYC Cab Drivers' Detailed Whereabouts." *Ars Technica*. Webpage: <https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/> (accessed March 1, 2018).
- Haynes, L. 2015. *Introducing the Data Maturity Framework*. Webpage: <https://dssg.uchicago.edu/2016/04/28/introducing-the-data-maturity-framework/> (accessed May 2017).
- Horridge, M. 2011. *A Practical Guide To Building OWL Ontologies, Using Protégé 4 and CO-ODE Tools, Edition 1.3*. technical, Manchester: The University of Manchester. Available online: http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf.

- Houston, N., C. Baldwin, A. Vann Easton, S. Cyra, M. Hustad, and K. Belmore. 2008. *Federal Highway Administration Service Patrol Handbook*. FHWA-HOP-08-031. HOTO-1, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Huddleston, G. 2016. "Cognitive Analytics Is Helping To Reduce Roadway Fatalities In Tennessee." *Forbes* (April 28, 2016). Available online: <https://www.forbes.com/sites/ibm/2016/04/28/cognitive-analytics-is-helping-to-reduce-roadway-fatalities-in-tennessee/#46d47fd07825> (accessed February 8, 2018).
- Ian. 2016. "What Is a Column Store Database?" *Database.guide* (June 23, 2016). Webpage: <http://database.guide/what-is-a-column-store-database> (accessed February 12, 2018).
- IBM. 2013. *What Will We Make of This Moment? 2013 IBM Annual Report*. IBM: Armonk, NY. Available online: https://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf.
- ICIJ. 2016. "Giant Leak of Offshore Financial Records Exposes Global Array of Crime and Corruption." *ICIJ Investigations: The Panama Papers*. Webpage: <https://panamapapers.icij.org/20160403-panama-papers-global-overview.html> (accessed June 2017).
- International Transport Forum. 2015. *Big Data and Transport: Understanding and Assessing Options*. Corporate Partnership Board Report. International Transport Forum at the OECD, Paris, France.
- internetlivestats.com. 2016. *Internet Live Stats*. Webpage: <http://www.internetlivestats.com/internet-users/> (accessed June 2017).
- Iowa State University. 2017. "Iowa State Engineers Dive Into Big Data to Develop Better System to Manage Traffic Incidents." *Iowa State University News Service* (March 22, 2017). Webpage: <http://www.news.iastate.edu/news/2017/03/22/timeli> (accessed June 2017).
- Issenberg, S. 2012. "How Obama Used Big Data to Rally Voters, Part 1: How President Obama's Campaign Used Big Data to Rally Individual Voters." *MIT Technology Review* (December 16, 2012). Webpage: <https://www.technologyreview.com/s/508836/how-obama-used-big-data-to-rally-voters-part-1/> (accessed July 24, 2018).
- Jin, X., M. S. Hossan, A. Gan, and D. Chen. 2014. "Comprehensive Framework for Planning and Assessment of Traffic Incident Management Programs." *Transportation Research Record*, Vol. 2470. Transportation Research Board of the National Academies, Washington, D.C., pp. 1–12.
- Kanniyappan, R., and B. McQueen. 2014. "What's the Big Deal about Big Data in Transportation?" Florida Department of Transportation, Data Symposium (Orlando, FL, October 23–24, 2014). Available Online: <http://www.fdot.gov/statistics/symposium/2014/bigdatatransport.pdf> (accessed November 2016).
- Kim, M., J. Cobb, M. J. Harrold, K. Malhotra, A. Orso, J. Saltz, S. B. Navathe, A. Post, and T. Kurc. 2012. *Efficient Regression Testing of Ontology-Driven Systems*. Minneapolis: International Symposium on Software Testing and Analysis. Available online: https://www.researchgate.net/publication/236842047_Efficient_Regression_Testing_of_Ontology-Driven_Systems.
- Klieman & Lyons. 2014. "Vehicle Telematics: A Useful Litigation Tool For Attorneys, A Boon To Insurers And The Privacy Concerns Big Data Raises For Us All." *Klieman & Lyons*. Webpage: <http://www.kliemanlyons.com/2014/09/vehicle-telematics-a-useful-litigation-tool-for-attorneys-a-boon-to-insurers-and-the-privacy-concerns-big-data-raises-for-us-all/> (accessed February 12, 2018).
- Krechmer, D., A. Samano III, P. Beer, and N. Boyd. 2012. *Role of Transportation Management Centers in Emergency Operations: Guidebook*. FHWA-HOP-12-050. Office of Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C. Available online: <https://ops.fhwa.dot.gov/publications/fhwahop12050/fhwahop12050.pdf>.
- Laroca, R., E. Severo, L. A. Zanolensi, L. S. Oliveira, G. Resende Goncalves, W. R. Schwartz, and D. Menotti. 2018. "A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector." *arXiv*. Webpage: <https://arxiv.org/pdf/1802.09567.pdf> (accessed July 24, 2018).
- Latonski, J., and J. Ang-Olson. 2006. *Simplified Guide to the Incident Command System for Transportation Professionals*. FHWA-HOP-06-004. Office of Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Lima, J. 2015. "Top 10 Biggest Data Centers from Around the World." *Computer Business Review*. Webpage: <http://www.cbonline.com/news/data-centre/top-10-biggest-data-centres-from-around-the-world-4545356/> (accessed June 2017).
- Lohmann, S., V. Link, E. Marbach, and S. Negru. 2014. "WebVOWL: Web-based Visualization of Ontologies." Presented at 19th International Conference on Knowledge Engineering and Knowledge Management (November 24–28, 2014, Linköping, Sweden). Available online: https://link.springer.com/chapter/10.1007%2F978-3-319-17966-7_21.
- Lokanathan, S., G. E. Kreindler, N. H. Nisansa de Silva, Y. Miyauchi, D. Dhananjaya, and R. Samarajiva. 2016. "The Potential of Mobile Network Big Data as a Tool in Colombo's Transportation and Urban Planning." *Information Technologies & International Development* [Special Issue], 12 (2), 63–73.
- Ma, J., and T. Lochrane. 2015. *User's Manual for the Traffic Incident Management Benefit-Cost (TIM-BC) Tool Version 1.0.0*. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

- Available online: <https://www.fhwa.dot.gov/publications/research/operations/15059/15059.pdf> (accessed August 2018).
- Ma, J., E. Miller-Hooks, M. Tariverdi, T. Lochrane, F. Zhou, D. Prentiss, K. Hudgins, P. Jodoin, Z. Huang, and M. Hailemariam. 2016. *User-Friendly Traffic Incident Management (TIM) Program Benefit-Cost Estimation Tool Version 1.2*. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C. Available online: <https://www.fhwa.dot.gov/publications/research/operations/16055/16055.pdf>.
- Mailvaganam, H. 2007. "Introduction to OLAP." *Data Warehousing Review*. Online: http://www.dwreview.com/OLAP/Introduction_OLAP.html (accessed June 2017).
- Marr, B. 2015. "How Big Data and The Internet of Things Improve Public Transport in London." *Forbes* (May 27, 2015). Available online: <http://www.forbes.com/sites/bernardmarr/2015/05/27/how-big-data-and-the-internet-of-things-improve-public-transport-in-london/#1b0ed2453ab3> (accessed November 2016).
- Marr, B. 2017. "Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud." *Forbes* (January 23, 2017). Available online: <https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/#7531c7e06c10> (accessed February 13, 2018).
- Martinelli, J. 2017. "Tennessee Sheriffs Will Receive Technology That Can Help Predict Fatal Crashes." *Nashville Public Radio* (October 11, 2017). Available online: <http://nashvillepublicradio.org/post/tennessee-sheriffs-will-receive-technology-can-help-predict-fatal-crashes#stream/0> (accessed February 8, 2018).
- Matrix. 2014. "Police Patrol Beat Evaluation Study." *City of Berkeley*. Webpage: [https://www.cityofberkeley.info/uploadedFiles/Police/Level_3_-_General/Berkeley%20Beat%20Structure%20Final%20Report%208-20-14\(1\).pdf](https://www.cityofberkeley.info/uploadedFiles/Police/Level_3_-_General/Berkeley%20Beat%20Structure%20Final%20Report%208-20-14(1).pdf) (accessed May 2017).
- Nelson, P. 2016. "Just One Autonomous Car Will Use 4,000 GB of Data/Day." *Network World* (December 7, 2016). Webpage: <http://www.networkworld.com/article/3147892/internet/one-autonomous-car-will-use-4000-gb-of-dataday.html> (accessed July 2017).
- Nemschoff, M. 2014. "Why the Transportation Industry Is Getting on Board with Big Data & Hadoop." *Map R Technologies, Inc.* (August 2014). Webpage: <https://www.mapr.com/blog/why-transportation-industry-getting-board-big-data-hadoop>.
- Neumann, C.-S. 2015. *Big Data Versus Big Congestion: Using Information to Improve Transport*. Available online: <http://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/big-data-versus-big-congestion-using-information-to-improve-transpo> (accessed June 2017).
- Ng, L. 2014. "Engage Users with Crowdsourced Translation." *OneSky*. Blog (September 22, 2014): <http://www.oneskyapp.com/blog/how-to-engage-users-with-crowdsourced-translation/> (accessed July 2018).
- NHTSA. 2012. *Traffic Records Program Assessment Advisory*. Technical Advisory. U.S. Department of Transportation, Washington, D.C. Available online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.
- NJ TIM. n.d. "New Jersey TIM Working Groups, TMC CAD [Computer Aided Dispatch] Integration." Webpage: <http://www.njtim.org/NJTIM/Group/GroupLogin>.
- NOAA. 2016. "MADIS." *National Oceanic and Atmospheric Administration*. Webpage: <https://madis.ncep.noaa.gov/> (accessed February 2017).
- ODPA. 2010. "Ontology:DOLCE+DnS Ultralite." *Ontology Design Patterns* (March 10). Association for Ontology Design & Patterns (ODPA). Webpage: http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite (accessed July 2017).
- Oerter, J. A. 2010. *North Carolina's Application of the INRIX Data*. North Carolina Department of Transportation. Available online: http://www.i95coalition.org/wp-content/uploads/2015/03/TRB_2010_Session_531-JoAnnOeter.pdf?5a9c76 (accessed May 15, 2017).
- Open Data Institute. n.d. "The Data Spectrum." *Open Data Institute*. Webpage: <https://theodi.org/data-spectrum> (accessed July 2017).
- Open Knowledge Foundation n.d.-a. "What Is Open Data?" *Open Data Handbook*. Webpage: <http://opendatahandbook.org/guide/en/what-is-open-data/> (accessed May 2017).
- Open Knowledge Foundation. n.d.-b. *Open Definition*. Online: <http://opendefinition.org/> (accessed May 2017).
- Oracle. 1999. "Oracle8i Designing and Tuning for Performance." *Oracle*. Webpage: http://docs.oracle.com/cd/A87860_01/doc/server.817/a76992/ch3_eval.htm#2680 (accessed June 2017).
- Oregon DOT. 2018. "Assigned Over 90 Minute Crash Causes." Workshop graphic (July 2018).
- Owens, N. D., A. H. Armstrong, C. Mitchell, and R. Brewster. 2009. *Federal Highway Administration Focus States Initiative: Traffic Incident Management Performance Measures: Final Report*. FHWA-HOP-10-010. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Owens, N. D., A. H. Armstrong, P. Sullivan, C. Mitchell, D. Newton, R. Brewster, and T. Trego. 2010. *Traffic Incident Management Handbook*. FHWA-HOP-10-013. Office of Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

- Paul, R. 2008. "Gartner: 80% of Commercial Apps to Use Open Source by 2012." *arstechnica.com* (February 5, 2008). Webpage: <https://arstechnica.com/information-technology/2008/02/gartner-80-percent-of-commercial-software-programs-will-include-open-source-by-2012/>.
- Pechoux, K. K. 2016. *Process for Establishing, Implementing, and Institutionalizing a Traffic Incident Management Performance Measurement Program*. FHWA-HOP-15-028. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Pechoux, K. K., R. E. Brydia, and J. Holzbach. 2014. "Guidance for Implementation of Traffic Incident Management Performance Measurement." *TIMPM* [NCHRP Project 07-20, "Guidance for Implementation of Traffic Incident Management Performance Measurement" website]. Webpage: <http://nchrptimpm.timnetwork.org/>.
- Pechoux, K. K., V. Shah, and C. O'Donnell. 2016. *Making the Business Case for Traffic Incident Management*. FHWA-HOP-16-084. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Petricek, T. 2010. "Parallel Programming in F# (III.): Aggregating data." *Thomas Petricek*. Blog (September 6, 2010): <http://tomasp.net/blog/fsharp-parallel-aggregate.aspx/> (accessed July 2018).
- Protégé. 2016. *Protégé*. Webpage: <http://protege.stanford.edu/> (accessed July 2017).
- Rensel, E., D. Lebo, B. Graves, K. Malarich, and C. Yorks. 2012. "Traffic Incident Management Cost Management and Cost Recovery: Executive-Level Briefing." Presentation. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C. Available online: https://ops.fhwa.dot.gov/eto_tim_pse/ppt/tim_cm_cr_exec_brief/tim_cm_cr_exec_brief.pdf (accessed June 2017).
- Rice, D. 2013. "3 Video Analytics Any Retailer Can Use to Boost Their Bottom Line." *SDM* (July 18, 2013). Webpage: <http://www.sdmag.com/articles/89461-video-analytics-any-retailer-can-use-to-boost-their-bottom-line> (accessed June 2017).
- Rocchini, C. 2007. "Graph Betweenness—Creative Commons Attribution-Share Alike 3.0 Unported License." *Wikipedia* (April 23, 2007). Webpage: https://commons.wikimedia.org/wiki/File:Graph_betweenness.svg (accessed July 24, 2018).
- SB. 2016. "Average Cost of Hard Drive Storage." *Statistic Brain Research Institute*. Webpage: <http://www.statisticbrain.com/average-cost-of-hard-drive-storage/> (accessed June 2017).
- Shaw, R. 2010. "LODE: An Ontology for Linking Open Descriptions of Events." *Linked Events*. Webpage: <http://linkedevents.org/ontology/> (accessed June 2017).
- Shi, Q. and M. Abdel-Aty. 2015. "Big Data Applications in Real-Time Traffic Operation and Safety Monitoring and Improvement on Urban Expressways." *Transportation Research Part C, Vol. 58*. Elsevier, Ltd., London, UK, pp. 380–394.
- Smith, A. 2016. "Wazing 511: How We Integrated Waze Connected Citizens Partnership Data (CCP) into Four State 511 Systems." Presented at 2016 National Rural ITS Conference (October 3, 2016, Chattanooga, TN). Available online: http://www.nationalruralitsconference.org/wp-content/uploads/2016/10/G2_Smith.pdf (accessed June 2017).
- Socrata, Inc. 2014. "Open Data Maturity Level." *www.socrata.com*. Webpage: https://www.slideshare.net/epsiplatform/big-data-session/6-Open_Data_Maturity_Model_6 (accessed May 2017).
- Sparks, J., J. Schuh, and A. Smith. 2009. *Field Operations Guide for Safety/Service Patrols*. FHWA-HOP-10-014. Office of Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.
- Stonebraker, M., and S. Jarr. 2013. "Navigating the Database Universe." Presentation at GOTOCon: AARHUS International Software Development Conference (January 29, 2013). Available online: <https://www.slideshare.net/BigDataCloud/big-data-cloud-meetup-jan-29-2013-mike-stonebraker-of-voltdb> (accessed February 2018).
- Tay, L. 2013. "Inside eBay's 90PB data warehouse." *IT News* (May 10, 2013). Webpage: <https://www.itnews.com.au/news/inside-ebays-90pb-data-warehouse-342615> (accessed February 12, 2018).
- Tennessee Department of Safety and Homeland Security. 2017. "Tennessee Highway Patrol Provides Predictive Analytics Software to All Tennessee Sheriffs' Offices." *Tennessee Department of Safety and Homeland Security*. Webpage: <https://www.tn.gov/safety/news/2017/10/10/tennessee-highway-patrol-provides-predictive-analytics-software-to-all-tennessee-sheriffs-offices.html> (accessed October 16, 2018).
- Thomson, D. and T. Gloer. n.d. "Annual Report 2010." *Thomson Reuters*. Available online: <http://archive.annual-report.thomsonreuters.com/2010/> (accessed June 2017).
- University of Chicago. n.d. "Data Maturity Framework Questionnaire." Center for Data Science and Public Policy. Webpage: http://dsapp.uchicago.edu/wp-content/uploads/2016/04/Data_Maturity_Framework_4.28.16.pdf (accessed May 2017).
- University of Chicago. 2017. "Data Maturity Framework." Center for Data Science and Public Policy. Webpage: <https://dsapp.uchicago.edu/home/resources/datamaturity/> (accessed May 2017).
- University of Manchester. 2009. *CO-ODE Project*. Webpage: <http://owl.cs.manchester.ac.uk/research/co-ode/> (accessed October 18, 2018).

- U.S. DOT. 2017a. "DOT|TRCC Data Systems and Attribute Icons." U.S. Department of Transportation (December 11, 2017). Webpage: <https://www.transportation.gov/government/traffic-records/dottrcc-data-systems-and-attribute-icons> (accessed July 2018).
- U.S. DOT. 2017b. *MMUCC*. Webpage: <https://www.transportation.gov/government/traffic-records/model-minimum-uniform-crash-criteria-mmucc-0> (accessed June 2017).
- USFA. 2017. *U.S. Fire Administration* (May). Webpage: https://www.usfa.fema.gov/data/statistics/order_download_data.html (accessed February 2017).
- VA Exec. Order No. 58. (February 4, 2013). *Commonwealth of Virginia, Office of the Governor*. Webpage: http://digitool1.lva.lib.va.us:8881/R/PRTKTK7BT14QPSA7A2524MRX85KUNDVUFDPD3LLU5ELGYBANSJV-02299?func=results-jump-full&set_entry=000065&set_number=818011&base=GEN01 http://digitool1.lva.lib.va.us:8881/R/PRTKTK7BT14QPSA7A2524MRX85KUNDVUFDPD3LLU5ELGYBANSJV-02299?func=results-jump-full&set_entry=000065&set_number=818011&base=GEN01 (accessed May 2019).
- VA Exec. Order No. 15. (April 24, 2015). *Commonwealth of Virginia, Office of the Governor*. Webpage: <https://governor.virginia.gov/media/3345/eo-15-updated-4222015ada.pdf> (accessed June 2017).
- Washington State Department of Transportation. 2016a. "Joint Operations Policy Statement." *Washington State Department of Transportation*. Webpage: <https://www.wsdot.wa.gov/NR/rdonlyres/F7F858A5-4246-4D34-8C24-5C91A0EB13CC/0/WSDOTWSPWFCJOPS.pdf> (accessed June 2017).
- Washington State Department of Transportation. 2016b. "Washington State." *TIMPM* [NCHRP Project 07-20, "Guidance for Implementation of Traffic Incident Management Performance Measurement" website]. Webpage: http://nchrptimpm.timnetwork.org/?page_id=1394 (accessed June 2017).
- Wiener, G., and A. Braeckel. 2016. *Integrating Big Data into Transportation Services*. White Paper. Colorado Department of Transportation, Denver, CO.
- Younas, M. 2013. "How to Really Outsmart Traffic." *Ibuymobile.co.uk* (July 9, 2013). Webpage: <http://ibuymobile.co.uk/reviews/2013/07/09/how-to-really-outsmart-traffic>.
- Zeng, Q., A. Reddy, A. Lu, and B. Levine. 2015. "Develop New York City Surface Transit Boarding and Alighting Ridership Daily Production Application Using Big Data." Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C.



APPENDIX A

Data Source Assessment Tables

Appendix A presents the detailed data assessment tables for 31 data sources. The criteria used to assess each data source are shown and described in Table 5-1 in Chapter 5 of this report. The data source assessments were qualitative, driven by the assessment criteria, and based on the information that was readily available for each source. For some of the data sources, interviews with data owners provided more detailed and specific information about the sources, allowing for a more complete understanding of the data and limitations. The data assessment is by no means exhaustive in terms of data sources or the information associated with each source. Some tables are more detailed than others, depending on the information available and/or whether the data is proprietary or business sensitive.

- A.1 STATE TRAFFIC RECORDS DATA SOURCES
 - Crash data (Table A-1)
 - Vehicle data (Table A-2)
 - Driver data (Table A-3)
 - Roadway data (Table A-4)
 - Citation and adjudication data (Table A-5)
 - Injury surveillance data (Table A-6)
- A.2 TRANSPORTATION DATA SOURCES
 - Traffic sensor data (Table A-7)
 - Traffic video data (Table A-8)
 - Freeway/safety service patrol and incident response program data (Table A-9)
 - 511 system data (Table A-10)
 - Road weather data (Table A-11)
 - Toll data (Table A-12)
- A.3 PUBLIC SAFETY DATA SOURCES
 - Law enforcement, fire and rescue, and EMS CAD system data (Table A-13)
 - Emergency communications system (ECC)/911 call center/public safety answering point (PSAP) data (Table A-14)
 - Video data (Table A-15)
 - Towing and recovery data (Table A-16)
- A.4 CROWDSOURCED/SOCIAL MEDIA DATA SOURCES
 - Waze data (Table A-17)
 - Twitter data (Table A-18)

- A.5 ADVANCED VEHICLE SYSTEMS DATA SOURCES
 - Automated vehicle location (AVL) system data (Table A-19)
 - Event data recorder (EDR) data (Table A-20)
 - Vehicle telematics data (Table A-21)
 - Automated and connected vehicle, traveler, and infrastructure data (Table A-22)
- A.6 AGGREGATED DATASETS
 - RITIS data assessment (Table A-23)
 - NPMRDS (Table A-24)
 - Meteorological Assimilation Data Ingest System (MADIS) and MADIS Integrated Mesonet (Table A-25)
 - Third-party web service weather data (Table A-26)
 - NFIRS data (Table A-27)
 - NEMSIS data (Table A-28)
 - MCMIS data (Table A-29)
 - HERE data (Table A-30)
 - INRIX data (Table A-31)

A.1 State Traffic Records Data Sources

Table A-1. Crash data.

Assessment Criteria	Assessment
Description of Data	Crash data includes detailed information about every reportable motor vehicle crash in a state, documents the characteristics of crashes, and provides the who, what, when, where, how, and why about each incident. ¹ Data elements include crash time, location, injury status, hazardous materials, motor carrier identification, roadway surface condition, total lanes in roadway, weather conditions, and other crash-specific data elements.
Who Collects, Maintains, and Owns the Data	Local, regional, and state law enforcement agencies collect the data via crash reports (either manually or electronically). Maintenance and ownership of the crash data varies among jurisdictions. Crash data is commonly aggregated at the state level.
How the Data Are Collected	Mostly electronically. When collected manually, paper reports are later keyed into electronic form. Data from multiple collection sources (paper and/or electronic) is then merged into a single database.
Data Structure	Structured and semi-structured. Each state has its own reporting system and storage system. The Model Minimum Uniform Crash Criteria (MMUCC) guideline is a minimum, standardized dataset for describing motor vehicle crashes and the vehicles, persons, and environment involved. ² The MMUCC contains 110 data elements, including 77 data elements to be collected at the scene; 10 data elements to be derived from the collected data; and 23 data elements to be obtained after linkage to driver history, injury, and roadway inventory data. MMUCC data is often exported in XML format.
Data Size, Storage, and Management	Gigabytes. Data is typically stored in relational databases maintained by local or statewide agencies. The database is kept in-house, archived in flat files, historical data is kept for several years (specific duration varies across agencies). Some crash data is aggregated at a national level like the Fatality Analysis Reporting System (FARS), which is maintained by the NHTSA to track all crashes involving a fatality.
Data Accessibility	Varies by agency. The closer the database schema is to the MMUCC, the easier the data can be understood and analyzed. Some agencies provide redacted public facing web-based information portals to query the data, while most states offer redacted large datasets that can be electronically downloaded.
Data Sensitivity	Personally identifiable information (PII) present in raw data; typically, redacted data is available for analysis.
Data Cost	Free, but some minor cost may be incurred to maintain data-sharing infrastructure.
Data Openness	Limited openness. Only redacted data is public. Access to non-redacted data needs to be granted by agency.
Data Challenges	Because the MMUCC is voluntary, states often use differing formats and names for data elements and attributes, or they may combine (or split) MMUCC elements and attributes. ² As a result, it can be very difficult to compare, merge, or share crash data among states, between state and federal datasets, and—in some cases—even between different agencies within a state. Although many agencies utilize electronic crash-reporting systems, which result in more complete and exploitable data, some agencies still use paper crash reports, which results in data that is less precise (vague time or location) or of lesser quality (e.g., missing fields, wrong categories). The latter can delay the upload of crash reports into a local or state database, as state or local personnel perform additional inquiries to obtain more precise or correct data.

¹ Traffic Records Program Assessment Advisory, NHTSA, U.S. Department of Transportation. Online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.

² Model Minimum Uniform Crash Criteria (MMUCC). Online: <https://www.transportation.gov/government/traffic-records/model-minimum-uniform-crash-criteria-mmucc-0>.

Table A-2 Vehicle data.

Assessment Criteria	Assessment
Description of Data	An inventory of data that enables the titling and registration of each vehicle under a state's jurisdiction to ensure that a descriptive record is maintained and made accessible for each vehicle and vehicle owner operating on public roadways. Vehicle information includes identification and ownership data for vehicles registered in the state, and out-of-state vehicles involved in crashes within the state's boundaries. Although data elements vary by jurisdiction and in element definitions, data elements generally include issuing agency; plate type; vehicle year, body style, weight, and identification number; and name of vehicle owner. ¹
Who Collects, Maintains, and Owns the Data	State-level government agency that administers vehicle registration and driver licensing (e.g., Department/Division/Office/Bureau of Motor Vehicles). The traditional department of motor vehicle (DMV) functions are handled by various agencies in different states (e.g., department of transportation, department of public safety, department of revenue, department of finance and administration, secretary of state, department of justice).
How the Data Are Collected	Electronically keyed at time of registration, automated license plate reader technology (ALPR), barcode/reader technology.
Data Structure	Structured.
Data Size, Storage, and Management	Gigabytes to terabytes. Data is stored in-house in relational database located in state agencies. Data is archived and maintained for multiple years (specific number of years varies from state to state).
Data Accessibility	Web services via criminal justice information networks and less-restrictive systems managed by the state licensing authority (limited to that state). The American Association of Motor Vehicle Administrators (AAMVA) maintains a pointer index system for commercial driver's license information called the National Motor Vehicle Title Information System (NMVTIS). ³ The NMVTIS also assists states and law enforcement in deterring and preventing title fraud and other crimes.
Data Sensitivity	Contains PII. The Driver's Privacy Protection Act (DPPA) of 1994 is a federal statute governing the privacy and disclosure of personal information gathered by state DMVs. ²
Data Costs	Free.
Data Openness	Data is not fully open due to personal information. Data can be shared, but usually is the basis of one inquiry/one record at a time. Personal information is protected by the DPPA.
Data Challenges	May not be accessible from the DMV due to PII and other restrictions. Disparities that sometimes make it difficult for officials in one jurisdiction to interpret data elements appearing on the vehicle registration document of another jurisdiction. ⁴

¹Traffic Records Program Assessment Advisory, National Highway Traffic Safety Administration, U.S. Department of Transportation. Online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.

²Drivers Privacy Protection Act (18 U.S.C. §2721 et. Seq.), Prohibition on Release and Use of Certain Personal Information from State Motor Vehicle Records. Online: <http://www.accessreports.com/statutes/DPPA1.htm>.

³American Association of Motor Vehicles Administrators, National Motor Vehicle Title Information System. Online: <http://www.aamva.org/nmvtis/> (accessed March 2017).

⁴Motor Vehicle Registration Document & Insurance Identification Best Practices Guide for Paper & Electronic Credentials, American Association of Motor Vehicle Administrators (August 2013). Online: <https://www.aamva.org/WorkArea/DownloadAsset.aspx?id=4437>.

Table A-3. Driver data.

Assessment Criteria	Assessment
Description of Data	Maintains driver identity, driving history, and license information for all records in the system. Contains information on each licensed driver, including name, birth date, license number, issuing state, license type, and historical driving record information (issuance, suspension, revocation, citations, crashes). The driver data system ensures that each person licensed to drive has one identity, one license to drive, and one record. ¹
Who Collects, Maintains, and Owns the Data	State-level government agency that administers vehicle registration and driver licensing (e.g., Department/Division/Office/Bureau of Motor Vehicles). The traditional DMV functions are handled by various agencies in different states (e.g., department of transportation, department of public safety, department of revenue, department of finance and administration, secretary of state, department of justice).
How the Data Are Collected	Electronically keyed, magnetic stripe, and barcode readers are three means of data collection. Typically, data also is reviewed physically for verification and updated through law enforcement or other means.
Data Structure	Structured.
Data Size, Storage, and Management	Gigabytes to terabytes. The data is stored in-house in relational databases located within the state agencies. Data is archived and maintained for multiple years (specific number of years varies from state to state).
Data Accessibility	<p>Each state has its own database. The information can be accessed via web services, criminal justice information networks, and less-restrictive systems managed by the state licensing authority (limited to that state) via FTP download. For example, Florida has a system called DAVID (Driver and Vehicle Information Database) that allows officers and courts to see driving records, all digital photos on file for drivers, and links to vehicles owned/registered.²</p> <p>States also share information with the American Association of Motor Vehicle Administrators (AAMVA). The AAMVA develops and maintains many information systems that facilitate the electronic exchange of driver, vehicle, and identity information between organizations (e.g., driver records, CDL skills testing, vehicle title, registration).³ For example, AAMVA maintains the Commercial Driver's License Information System (CDLIS), a nationwide computer system that enables state driver licensing agencies (SDLAs) to ensure that each commercial driver has only one driver's license and one complete driver record.</p> <p>Release of information is protected by the Drivers Privacy Protection Act (DPPA).⁴</p>
Data Sensitivity	Contains PII and, in some cases, legal privacy restrictions.
Data Costs	Free.
Data Openness	Limited openness, as the data contains PII and access needs to be requested.
Data Challenges	May not be accessible from the DMV due to PII and other restrictions like state laws protecting driver information.

¹Traffic Records Program Assessment Advisory, NHTSA, U.S. Department of Transportation. Online:

<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.

² DAVID, Florida Department of Highway Safety and Motor Vehicles. Online: <http://www.flhsmv.gov/courts/david/> (accessed March 2017).

³ American Association of Motor Vehicles Administrators, Application Services. Online: <http://www.aamva.org/Application-Services/> (accessed March 2017).

⁴ 18 U.S.C. 2721, U.S. Government Publishing Office. Online: <https://www.gpo.gov/fdsys/granule/USCODE-2011-title18/USCODE-2011-title18-partI-chap123-sec2721/content-detail.html> (accessed March 2017).

Table A-4. Roadway data.

Assessment Criteria	Assessment
Description of Data	Roadway datasets contain extensive information about roadway segments including roadway characteristics such as physical curvature, lane types and widths, pavement types, connected access roads, roadside descriptors, and interchange and ramp descriptors. Asset management datasets contains data relevant to the various equipment and facilities supporting roadways such as traffic signals, traffic signs, barriers, drainage, power stations, communications cables, etc. Roadway inventory data and asset inventory datasets are typically maintained in multiple separate databases. More advanced data management practices maintain these data in an integrated geospatial information systems (GIS) platform to allow assets and roadways to be easily located and mapped. The data itself ranges from tables to computer-aided design (CAD) drawings to geospatial vector data.
Who Collects, Maintains, and Owns the Data	State transportation agencies, county public works departments.
How the Data Are Collected	Manually, aerial images, linear referencing, GIS, cameras on vans, and backpacks. ³
Data Structure	Structured and semi-structured. The Model Inventory of Roadway Elements (MIRE) is a recommended listing of roadway inventory and traffic elements critical to safety management. ^{1,2} MIRE is intended as a guideline to help transportation agencies improve their roadway and traffic data inventories. It provides a basis for a standard of what can be considered a good/robust data inventory and helps agencies move toward the use of performance measures to assess data quality. The MIRE listing contains 202 data elements divided among three broad categories: (1) roadway segments, (2) roadway alignment, and (3) roadway junctions. The composition of MIRE was purposefully designed to link with supplemental databases, including: roadside fixed objects, signs, speed, automated enforcement devices, land use elements related to safety, bridge descriptors, and railroad grade-crossing descriptors.
Data Size, Storage, and Management	The size of the datasets varies among agencies (gigabytes to terabytes) and is directly correlated to the miles of the road network being inventoried, the level of detail being recorded for each roadway, the number of assets, and the level of details recorded for each asset. Dataset size further increases when maintaining detailed digitized CAD drawings and geospatial vector data files and linking these files to each roadway or asset record relevant data record. Storage of the datasets varies widely from agency to agency. Some agencies store roadway/asset data as spreadsheet files, some store scanned paper drawings or CAD files, and others store their data into fully integrated geospatial databases. Dataset management varies as well; typically, it is done by maintaining one or more file archives or databases in-house that contain multiple years of roadway and assets data. The file archive or database is updated periodically with new data to reflect the asset's maintenance or improvement history. Depending on the agency, the archives can be managed and centralized into a single location such as a GIS database or managed independently within each agency district.
Data Accessibility	Accessibility varies from agency to agency and can range from files and images mailed on disk or portable media to dedicated public web portals with searching and downloading capabilities.
Data Sensitivity	Sensitivity is dependent on the asset. For example, some asset data such as bridge CAD drawings and material information or models and versions of traffic signal management software, could be exploited by malicious individuals or groups.
Data Costs	Free.

(continued on next page)

Assessment Criteria	Assessment
Data Openness	Limited data openness to full openness, as some agencies do not publish this type of data to the public, whereas others maintain portals where the data can be easily searched, downloaded, and sometimes even visualized.
Data Challenges	<p>Data quality, delivery, timeliness, and accuracy vary widely across agencies. Many agencies may not have a web portal or FTP site, requiring that large datasets be delivered via disc or mail. Some agencies only use basic file-sharing systems to store their asset data, and these systems lack the data management structure to easily find, retrieve, and format requested asset data quickly. It is not uncommon to have to wait several days or weeks following a request to receive requested asset data.</p> <p>Asset data can also be distributed across agency districts and not routinely managed, updated, and maintained in a consistent fashion. Depending on budget and staff availability, each district may manage its asset data differently. The result may be the storage of asset data across various internal legacy systems with diverse structures and formats.⁴ This could make it very difficult to access and mine the asset data.</p> <p>The accuracy of the asset data also can be affected, as agencies or agency district resources may not have the resources to update assets records as soon as an asset is upgraded or replaced, resulting in stale asset data several weeks or months after asset work has been performed.</p>

¹ FHWA Roadway Safety Data Program. Online: <https://safety.fhwa.dot.gov/rsdp/mire.aspx>.

² Model Inventory of Roadway Elements VERSION 1.0, FHWA, U.S. Department of Transportation, October 2010. Online: https://safety.fhwa.dot.gov/tools/data_tools/mirereport/mirereport.pdf.

³ Khattak, A. J., J. E. Hummer, and H. A. Karimi. "New and Existing Roadway Inventory Data Acquisition Methods." *Journal of Transportation and Statistics, Vol 3, No 3, Paper 2*. Bureau of Transportation Statistics, U.S. Department of Transportation, Washington, D.C. Online: https://www.rita.dot.gov/bts/sites/rita.dot.gov/bts/files/publications/journal_of_transportation_and_statistics/volume_03_number_03/paper_02/index.html.

⁴ Asset Management Overview, FHWA, U.S. Department of Transportation (October 2010, December 2007). Online: https://www.fhwa.dot.gov/asset/if08008/assetmgmt_overview.pdf.

Table A-5. Citation and adjudication data.

Assessment Criteria	Assessment
Description of Data	Citation and adjudication databases maintain information about citations, arrests, and dispositions. The process is highly localized in data management from delivery of citation through adjudication. After the completion of local adjudication, the data will be delivered (in most states) to a state entity for driver's license reporting functions. Citation databases may contain information relevant to TIM, including occurrences of law enforcement activity along the roadside and potentially duration and type of activity.
Who Collects, Maintains, and Owns the Data	Law enforcement and parking enforcement are the primary point of data collection. Courts having jurisdiction coordinate with the state agency responsible for driver data.
How the Data Are Collected	Mostly electronic at point of collection. Paper documents are converted to electronic records at the court level.
Data Structure	Semi-structured and structured.
Data Size, Storage, and Management	Gigabytes to terabytes. State databases maintained in-house for multiple years.
Data Accessibility	FTP.
Data Sensitivity	PII.
Data Costs	Free.
Data Openness	Limited openness due to PII.
Data Challenges	May not be accessible from the DMV due to PII and other restrictions.

¹Traffic Records Program Assessment Advisory, NHTSA, U.S. Department of Transportation. Online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.

Table A-6. Injury surveillance data.

Assessment Criteria	Assessment
Description of Data	These surveillance systems typically incorporate pre-hospital emergency medical services (EMS), trauma registry, emergency department, hospital discharge, rehabilitation databases, payer-related databases, and mortality data (e.g., death certificates, autopsies, and coroner and medical examiner reports). The data from these various systems are used to track injury type, causation, severity, cost, and outcome. ¹
Who Collects, Maintains, and Owns the Data	EMS, hospitals (emergency departments, discharge, trauma registry), state vital records, medical examiner/coroner.
How the Data Are Collected	Given the numerous files and datasets that make up the injury surveillance system, a correspondingly large number of data standards and applicable guidelines exist for data collection. ¹ For example, EMS providers have been rapidly transitioning their paper records into electronic patient care reports (EPCRs) that are completed using laptop computers or tablets. ²
Data Structure	<p>Semi-structured and structured.</p> <p>The National Emergency Medical Services Information System (NEMSIS), developed through a collaborative effort with the EMS industry and originating from a memorandum of agreement among 52 states and territories, assigns specific definitions to 481 data elements identified as desirable to be collected on a national level for EMS. NEMSIS was developed to help states collect more standardized elements and eventually submit the data to a national EMS database.</p> <p>Administrative data files for emergency department visits and inpatient hospitalizations are based on the uniform billing code issued by the U.S. Department of Health and Human Services.¹</p> <p>The National Trauma Data Standard (NTDS), developed by the American College of Surgeons Committee on Trauma, provides data standards for trauma registry databases. Built on an XML schema shared with NEMSIS, the NTDS enables improved integration of EMS and trauma data.¹</p> <p>The U.S. Standard Certificates of Birth and Death and the Report of Fetal Death are the principal means of promoting uniformity in the data collected by the states. These documents are reviewed and revised approximately every 10 years through a process that includes broad input from data providers and users. The Centers for Disease Control and Prevention's National Center for Health Statistics provides guidance for cause of death coding based on ICD-10 standards.¹</p> <p>The AIS and the ISS are measures of injury severity. The AIS categorizes injury severity by body region and—when combined with crash data—can be used to describe injury patterns by crash configuration. The ISS provides a more comprehensive measure of injury severity when a patient has injuries to multiple body regions. The Glasgow Coma Scale is used to assess the neurologic state of a patient.¹</p>

Assessment Criteria	Assessment
Data Size, Storage, and Management	<p>Component databases—gigabytes.</p> <p>EMS providers, hospitals, state department of health, state databases, NEMSIS. In-house systems, maintained for multiple years.</p> <p>The EMS applications of today can sync up with monitoring equipment and computer-aided dispatch (CAD) systems to automatically populate data related to each assigned call. Providers can track and input the progress of a patient's vitals, automatically record medication dosage and times, capture and electronically save electrocardiograms (EKGs), and transmit that information to the awaiting hospital.</p> <p>One app can sometimes be utilized by the EMS user to manage all the information from a shift, from populating dispatch and patient information, to gathering and documenting current findings, to the transmission of a patient's records to a health-care facility.²</p>
Data Accessibility	<p>Ideally, data is made available for local and state agency use.</p> <p>FTP, data dump.</p>
Data Sensitivity	Contains PII. In addition to any applicable state statutes, state health-care data custodians must comply with the pertinent aspects of the Health Insurance Portability and Accountability Act of 1996 (HIPAA).
Data Costs	Potential cost; available through data-sharing agreements at no cost.
Data Openness	Limited openness due to PII.
Data Challenges	May not be accessible due to PII and other restrictions.

¹Traffic Records Program Assessment Advisory, NHTSA, U.S. Department of Transportation. Online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.

²Busa, M. Information-Sharing Applications & Technology for the Fire Service (July 30, 2013). Online: <http://www.firerescuemagazine.com/articles/print/volume-8/issue-9/technology/information-sharing-applications-technology-for-the-fire-service.html>.

A.2 TRANSPORTATION DATA SOURCES

Table A-7. Traffic sensor data.

Assessment Criteria	Assessment
Description of Data	Data from sensors, including inductive loop detectors, magnetic sensors and detectors, video image processors, microwave radar sensors, laser radars, passive infrared and passive acoustic array sensors, and ultrasonic sensors, plus combinations of sensor technologies. Certain detectors give direct information concerning vehicle passage and presence, while other traffic flow parameters such as density and speed are inferred from algorithms that interpret or analyze the measured data. Data elements collected include date, time, sensor ID, roadway ID, direction, annual average daily traffic (AADT), Truck AADT, volumes (vehicles/minute), speed, occupancy, vehicle classification.
Who Collects, Maintains, and Owns the Data	State, MPO, county, and city transportation agencies.
How the Data Are Collected	Automatically collected through the technology through sampling interval (e.g., 20 seconds, 60 seconds), or manually either by lane or for roadway sections.
Data Structure	Structured. ^{1,2}
Data Size, Storage, and Management	Gigabytes to terabytes, depending on the size of the area being monitored (e.g., regional, statewide). Typically stored as flat files or in relational databases. Data are typically aggregated at 1-, 5-, or 15-minute intervals for storage, analysis, and visualization.
Data Accessibility	Variable by entity ranging from aggregate data stored in CSV files by location on the premises to statewide web-accessible databases providing more granular data. Traffic sensor data is typically archived for several years, and many states have requirements on the history for storage. Examples of states and organizations that have developed data storage services includes, the Texas DOT, which has begun storing detailed traffic sensor data through its STARS II system; Caltrans, which uses its Performance Measurement System (PeMS) website to store more than 10 years of traffic sensor data; and the University of Maryland's CATT Lab, which consolidates traffic data into the Regional Integrated Transportation Information System (RITIS). The first two provide both visualizations and structured datasets to users, while RITIS focuses mostly on advanced visualizations of traffic sensor data for multiple states. Although visualizations and aggregated datasets are very valuable to human consumption, disaggregate, high-resolution data is essential to Big Data analysis. Raw traffic sensor data is often unavailable because the large volume of data can be costly to store, resulting in aggregation and storage of data from seconds of resolution to 5- or 15-minute averages, and even hourly or lower resolution for some organizations.
Data Sensitivity	None
Data Costs	Free/public. Data is most often offered at no cost online by state or regional organizations and is characterized as public domain data. A minor fee may be charged when requesting the data on a paper or disk format.
Data Openness	Limited data openness – Traffic sensor data is often shared with the public through high-level aggregation or visualizations such as maps, but rarely as raw data. More granular or raw traffic sensor data is typically not shared openly to the public, and its accessibility for federal, state, local, private, and public individuals is usually only granted upon request and after review of the intended use of the data.

Assessment Criteria	Assessment
Data Challenges	Institutional. Tied to the ability of the institution to be able to provide and manage access to raw traffic sensor data as well as its ability to ensure high traffic sensor data quality by monitoring sensor drift, performing recalibration on a regular basis, and maintaining precise sensor location information.

¹ Traffic Monitoring Guide (TMG), Federal Highway Administration, U.S. Department of Transportation (October 2016). Online: https://www.fhwa.dot.gov/policyinformation/tmguidetmg_fhwa_pl_17_003.pdf.

² AASHTO *Guidelines for Traffic Data Programs*, 2nd Ed. Online: https://bookstore.transportation.org/item_details.aspx?ID=1393.

Table A-8 Traffic digital video data.

Assessment Criteria	Assessment
Description of Data	<p>Digital video is a representation of moving visual images in the form of encoded digital data. Digital video data is collected by transportation agencies through closed-circuit television (CCTV) (video surveillance) cameras, video detection, and automatic license plate reader/recognition (ALPR) systems.</p> <ul style="list-style-type: none"> • CCTV systems use video cameras to transmit a signal to a specific place, on a limited set of monitors. Transportation agencies use CCTV cameras on highways, ramp locations, and intersections to monitor traffic from a central location such as a traffic management center (TMC). • Video detection devices capture video images of traffic and analyze the information using algorithms for traffic management (e.g., traffic signal control). • ALPR systems identify vehicles passing fixed locations using cameras that read the license plates. Such systems are widely used in electronic tolling applications.
Who Collects, Maintains, and Owns the Data	State and local transportation agencies, private toll operators, and parking lot managers.
How the Data Are Collected	<p>Video data is collected via various types of remote camera technologies, generally deployed at fixed locations but with selectable orientation connected to a centralized location. Collected video and related images are then viewed live from the central location and sometimes recorded and stored in video archive for various amounts of time, often dictated by law or budget.</p> <p>Automatic video recognition software such as ALPR systems located either on the camera or down the video stream can be added to automatically extract metadata from captured images within the video. This metadata is then associated with a specific camera and a specific timestamp or timeframe and saved to one or more databases for storage or sent as a message to alert authorities when a vehicle of interest has been observed.¹</p> <p>Watching live video images allows for the extraction of many relevant data elements; however, this approach to data processing is limited. Most modern approaches to capturing video embed metadata such as date, time, and location into video frames during capture using <i>exchangeable image file format</i> (EXIF) tags. This metadata can then be augmented using machine-learning software, which uses image processing algorithms to extract from each video frame additional metadata such as vehicle counts, estimated speeds, tag numbers of passing vehicles, vehicle type, vehicle orientation, and so forth. This metadata is then used to qualify and characterize the event recorded on video.</p>
Data Structure	Semi-structured.
Data Size, Storage, and Management	<p>Gigabytes to terabytes. Video data is notorious for its very large size. Common practice is to compress video data on capture and then transmit and store it in compressed form until it is viewed or used by automated recognition algorithms. Storage/recording of video images is largely a policy decision for transportation agencies. Three fundamental video recording approaches are used: (1) always (continuously record most feeds and retain them for a few days), (2) sometimes (initiate recording of individual feeds for specific events), and (3) never. Kuciemba and Swindler (2016) describe the benefits and limitations to each approach. Of 32 TMCs surveyed, five reported they recorded most feeds most of the time, 23 reported they recorded the videos only under limited circumstances, and four reported they never recorded videos.²</p>

Assessment Criteria	Assessment
Data Accessibility	<p>Accessibility varies widely among traffic video sources. Video is typically shared using a streaming method, which is commonly used to share video with media outlets and to some degree with the public via 511 and motorist-information websites using low resolution/quality video streaming. Transportation agencies also share in real time images extracted from video feeds with allied agencies like law enforcement, fire, and towing dispatch centers.</p> <p>Video streams and pictures also can be accessed by a restricted list of users from allied agencies using custom mobile or desktop applications. Alternatively, when stored or archived, TMC video can be provided upon request, which involves manual searches for the date, time, and location of the event requested. Most requests come from law enforcement. Video is typically copied onto a media store device and either picked up by or mailed to the requester.</p> <p>Although most traditional video data systems can store or archive and make the data accessible to a public or restricted audience, the data remains accessible at low resolution, which greatly limits its ability to be analyzed to provide value when machine processed. Digital and Internet Protocol (IP) camera systems offer an alternative that uses the Internet to transmit video to servers that can process the stream to add tags, clean the images, detect, and send alerts to interested parties directly using less communications bandwidth.</p>
Data Sensitivity	<p>When dealing with low-resolution video, generally the quality of the video is too low to allow sensitive information to be extracted. Low-quality video rarely depicts license plates and recognizable facial images; however, when dealing with high-definition video, sensitivity can increase greatly, as such information becomes visible and video processing can be performed automatically to detect sensitive information such as faces, license plates, location, and so forth.</p>
Data Costs	<p>Video is typically available for free to the public (at low resolution) or to other agencies and institutions (at high resolution). Video and image data files, even compressed, require large storage capabilities. Consequently, a non-negligible cost is associated with the retention of video and images. The amount and quality of data stored, compression ratios, image size, and retention period are factors that impact operational cost. Cloud storage services typically are used to store video and images because they offer the most economical storage solutions, allowing video to be stored without degrading its quality; but cloud storage is used rarely by TMCs.</p>
Data Openness	<p>Low-resolution video data from roadway CCTVs is usually open to the public. High-resolution video content is not usually accessible to the public; rather, it is made available only to requesting agencies on demand with a valid reason for obtaining the data (e.g., for a law enforcement investigation).</p>

(continued on next page)

Assessment Criteria	Assessment
Data Challenges	<p>In most cases, TMC video or images are not stored or archived. When stored, video data is only stored and maintained for a brief period; then it is purged to make room for newer video. This practice greatly limits the potential quantity of video content that could be mined. Also, video collection is not uniform across space, time, and quality:</p> <ul style="list-style-type: none"> • Coverage areas for roadway cameras varies; when present, camera views do not always provide complete coverage for all parts of the highway. • Equipment failures of field cameras, communications networks, and recording systems also can increase the lack of coverage when maintenance of cameras is not performed in a timely manner. • Weather conditions such as snow and rain can greatly affect the quality of the video collected, in some cases making it impossible to extract metadata. • Video container and compression standards vary widely between equipment and manufacturers. These standards often are proprietary and cannot be converted easily to a common standard without losing some video data integrity. <p>These challenges result in video/image datasets that are sparse, non-uniform, and unevenly distributed, making it difficult to extract general trends or patterns. Conversion of real-time video for large-scale distribution can be expensive and require considerable information technology (IT) infrastructure.</p>

¹ International Association of Chiefs of Police, About ALPR. Online: <http://www.iacp.org/ALPR-About>.

² Kuciemba, S., and K. Swindler, *Transportation Management Center Video Recording and Archiving Best General Practices*, U.S. Department of Transportation, Washington, D.C. (March 2016). Online: <https://ops.fhwa.dot.gov/publications/fhwahop16033/fhwahop16033.pdf>.

Table A-9. Safety service patrol and incident response program data.

Assessment Criteria	Assessment
Description of Data	<p>Data is collected by safety service patrol (SSP) program (often called freeway service patrol or incident response) staff that is present at the scene of an incident. Data collected generally includes time and location of incident, type of incident, arrival and departure times, responder and response vehicle identification, supplies expended (e.g., gas or a tire patch), and the assistance provided (e.g., refueling, repairing tire, blocking lane or calling tow vehicle) using either pre-established codes or keywords, or free text.</p> <p>Some SSP programs also request a response from the drivers/vehicles assisted in the form of a postcard survey or request to complete an online survey with structured and unstructured data. This data typically captures the quality and value of services provided.</p>
Who Collects, Maintains, and Owns the Data	State transportation agencies, metropolitan planning organizations (MPOs), transportation authorities.
How the Data Are Collected	Depending on the program, data is collected by the responder either manually (simple paper forms/logs), electronically (via laptops, tablets, mobile phones), and/or is communicated via radio back to a central location such as a TMC. ¹
Data Structure	Data structure varies based on the collection method. Data can range from free text on simple forms to standardized records in relational databases. Data is often integrated with TMC software and records management systems.
Data Size, Storage, and Management	Megabytes. Although service patrols and incident response programs respond to large numbers of incidents daily, most of these incidents are easy to mitigate and do not generate a large amount of information. Most SSP incidents can be described accurately in less than 10 data fields. Data archiving often is done in-house by maintaining spreadsheets for a period of time (e.g., a month or an entire year) and then integrated into TMC software systems as part of the system archiving process. Archiving duration varies greatly across agencies.
Data Accessibility	Accessibility varies by entity, ranging from CSV or Excel files to statewide web-accessible databases providing more detailed and organized data that can be searched easily. Free text fields often are used to capture the details of incident responses rather than a standardized taxonomy. Free text, while still providing valuable information, is more difficult to analyze. The presence of abbreviations, synonyms and orthographic mistakes in the text makes the use of advanced text analytics mandatory before valuable information can be extracted.
Data Sensitivity	May contain PII.
Data Costs	Service patrol data is most often offered at no cost by state or regional organizations upon request and acceptance. A minor fee may be charged to obtain the data on a paper or disk format.
Data Openness	Service patrol data has limited data openness, as in most cases the data is not publicly shared online, and a request including the intended use of the data needs to be made to the operating agency to obtain the data on portable media or through a file-sharing service such as FTP. SSP customer feedback is typically accessible only at the aggregate level.

(continued on next page)

Assessment Criteria	Assessment
Data Challenges	<p>Most service patrol data are still collected using paper forms that are later entered into a database or spreadsheet or by a TMC operator in radio communication with responders. More modern ways of collecting service patrol data are becoming more prevalent. These systems, such as computer-aided dispatch (CAD) systems or mobile phone/tablet applications, capture data at the scene using a more structured and strict data collection process.</p> <p>Data collected from paper forms or radio communication and subsequently entered in spreadsheets or simple applications often lacks precise location information and is of lower quality due to the inability to correct for misspelled words, non-existent categories, non-standardized abbreviations, and custom narratives. This lower quality requires complex analysis to correct content and attempt to standardize the “fuzzy” content; but even with additional complex analysis, the resulting content may lose information precision in the process and become less valuable.</p> <p>Additionally, the current data management of service patrol data files (except for database systems) may also lead to difficulty ingesting and analyzing content. Often, spreadsheet files are collected, stored into shared network folders, and managed manually. Data file formats evolve and improve on a regular basis with improvements such as adding new columns or changing the category name used to describe service patrol responses, but the new formats may not be retroactively applied to update previously created data files. This less-rigorous data management leads to content that is non-uniform and difficult to analyze without cleaning. In some cases, retrofitting a new data format in older data files is not possible, as the historical data is less precise than the new data format requires.</p>

¹ FHWA *Service Patrol Handbook*, U.S. Department of Transportation (November 2008). Online: https://ops.fhwa.dot.gov/publications/fhwahop08031/ffsp_handbook.pdf.

Table A-10. 511 system data.

Assessment Criteria	Assessment
Description of Data	Traveler information (511) systems acquire, analyze, and communicate information to assist surface transportation travelers. The 511 system data and information can include general traffic (congestion and speeds) and weather conditions, as well as the location of incidents, work zones, roadway closures, and planned special events. Data sources to 511 systems generally include the state DOT, the highway patrol and police departments, transit agencies, and sometimes local jurisdictions and private companies.
Who Collects, Maintains, and Owns the Data	Varies between public transportation agency, private companies, or combination of both.
How the Data Are Collected	Varies between manual, semi-automated, or automated.
Data Structure	Structured and semi-structured. ¹
Data Size, Storage, and Management	Megabytes to gigabytes depending on area covered. Storage and management is typically done on the premises using third-party systems. The 511 data systems are real-time information systems focused on delivering travel information to users in less than 3 seconds. Currently, no specific guidelines exist for how 511 data need to be stored or archived, and archiving practices vary widely across systems. ²
Data Accessibility	Mobile, web, web services, SMS, email and phone (text-to-speech).
Data Sensitivity	Not sensitive, except for homeland security concerns.
Data Costs	Free.
Data Openness	Aggregated data is open to the public via the 511 system. Raw data may be shared upon request via FTP or data dump.
Data Challenges	The 511 data are, first and foremost, real-time human readable information that is unstructured or semi-structured. Although 511 systems are designed to quickly broadcast traffic and transit event information to travelers, they are not designed to store that data or even structure and organize it for later retrieval or searches. The 511 data would need to be stored on a different system to be analyzed over time. Event data elements such as location, timestamps, and event type can be easily used for analysis, but data elements containing free text, such as event description, will be more challenging to mine and organize. These data elements will require more advanced text analysis to extract valuable keywords and topics essential to further analysis.

¹ Real-Time System Management Information Program Data Exchange Format Specification, Federal Highway Administration, U.S. Department of Transportation (August 2013). Online: <https://ops.fhwa.dot.gov/publications/fhwahop13047/fhwahop13047.pdf>.

² America's Travel Information Number, Implementation and Operational, Guidelines for 511 Services, Federal Highway Administration, U.S. Department of Transportation, Version 3 (September 2005). Online: https://ops.fhwa.dot.gov/511/resources/publications/511guide_ver3/511guide3.htm.

Table A-11. Road weather data.

Assessment Criteria	Assessment
Description of Data	<p>Road weather data is precise, facility-specific, and timely weather information as it pertains to the effects on the road.¹ Road weather data collected at roadway locations can include atmospheric, pavement, and water level conditions. Atmospheric data can include air temperature and humidity, visibility distance, wind speed and direction, precipitation type and rate, tornado or waterspout occurrence, lightning, storm cell location and track, as well as air quality data. Pavement data can include pavement temperature, pavement freeze point, pavement condition (e.g., wet, icy, flooded), pavement chemical concentration, and subsurface conditions (e.g., soil temperature). Water level data can include tide levels (e.g., hurricane storm surge) as well as stream, river, and lake levels near roads.²</p> <p>State agencies use different systems, and the development of the Clarus System was an attempt to standardize data across regions. Clarus is based on the premise that the integration of a wide variety of weather observing, forecasting, and data management systems, combined with robust and continuous data quality checking, could serve as the basis for timely, accurate, and reliable weather and road condition information.¹ Clarus provides targeted and route-specific road weather information. Clarus has become the RWIS (Road Weather Information System) of the MADIS (Meteorological Assimilation Data Ingest System) operated by the National Centers for Environmental Prediction (NCEP), a part of the National Weather Service (NWS). Clarus aggregates various weather data from all over the world.</p> <p>FHWA has developed a new research platform, the Weather Data Environment (WxDE).³ The WxDE incorporates much of the Clarus data and functionality, as well as various ways to augment station data using connected vehicle data and applications.</p>
How the Data Are Collected	<p>Managed by a state or local agency, a RWIS collects data generated by a group of environmental sensor stations (ESS) located in sensitive areas of an agency's road network. A communication network relays data from the stations to a central RWIS system where the stations' data is stored. The weather station data is then monitored by the RWIS and transmitted to automated warning systems, traffic management centers, emergency operations centers, and road maintenance facilities.</p> <p>Clarus is an RWIS data aggregator that relies on state and local agency ESS networks to collect pavement and meteorological data next to roadways nationwide. Clarus aggregates road weather data from more than 2,400 ESS owned by state transportation agencies.⁴</p> <p>The WxDE collects data in real time from both fixed ESS and mobile weather stations, such as automated vehicle location (AVL) systems and connected vehicle systems. In addition to collecting road weather data in a central location (like Clarus), WxDE provides additional enhancements by combining and correlating collected weather data and events data such as windshield wiper activation to further refine its data quality and value.³</p>
Data Structure	Semi-structured (CSV, XML, and NetCDF).

Assessment Criteria	Assessment
Data Size, Storage, and Management	Gigabytes to terabytes, depending on coverage and time window. RWIS data typically is stored in relational databases and archived in flat files. State and local RWIS data management and archiving policies vary across state and local agencies. MADIS data (which includes data from Clarus) is archived indefinitely by the NOAA National Environmental Satellite, Data, and Information Service (NESDIS) and is stored as files in either CSV or NetCDF format. WxDE is archived indefinitely.
Data Accessibility	State and local data typically is accessible through website maps, download, or FTP. MADIS offers an advanced website with maps, a download page to access its data, and an associated application programming interface (API) to access its data directly from other applications. MADIS offers several file formats, but some of its data is stored in what are called “NetCDF files.” NetCDF files are a common application-agnostic format used to store scientific data. NetCDF files require an API to be read, which is also provided on the site. MADIS grants several levels of access for its data, ranging from “public” to “NOAA only.” Most information from WxDE is available to all website visitors. Registered users are provided with some additional capabilities, such as creating data subscriptions and accessing data for which the original provider placed restrictions on its distribution by the WxDE. ³
Data Sensitivity	None.
Data Costs	Free.
Data Openness	Limited openness. Access is dataset-dependent. Some datasets are accessible to the public, others require user registration, and some are restricted to government users.
Data Challenges	ESS may not always be maintained or monitored to counter sensor failure and sensor drift, which can lead to data quality issues (e.g., missing data, erroneous data). To circumvent this problem, quality checks and more advanced data verification and correction are performed by aggregators such as MADIS and WxDE. The NetCDF file format could also be challenging to use for non-scientific staff because it requires the implementation of dedicated API to access the data.

¹ Bureau of Transportation Statistics. 2011. *Clarus*. Online: https://ntl.bts.gov/lib/44000/44300/44374/FHWA-JPO-11-154_Clarus_Overview_final.pdf (accessed February 2017).

² FHWA. 2017. “Surveillance, Monitoring, and Prediction.” Online: https://ops.fhwa.dot.gov/weather/mitigating_impacts/surveillance.htm#esrw (accessed February 2017).

³ Weather Data Environment, FHWA. Online: <https://wxde.fhwa.dot.gov/>.

⁴ National Environmental Sensor Station Map, Road Weather Management Program, FHWA (February 2017). Online: https://ops.fhwa.dot.gov/weather/mitigating_impacts/essmap.htm.

Table A-12. Toll data.

Assessment Criteria	Assessment
Description of Data	Toll data, collected via electronic toll collection technology, includes the number of vehicles passing through toll gates, vehicle identification (license plate), unique toll tag identifier, automated vehicle classification, transaction processing, violation enforcement, date/timestamp, and location information.
Who Collects, Maintains, and Owns the Data	State transportation agencies, tollway authorities.
How the Data Are Collected	Each time a vehicle crosses a toll gate, an active vehicle-mounted radio-frequency identification (RFID) tag communicates with an antenna at a toll gate via dedicated short-range communications (DSRC). During the communication, the RFID tag broadcasts a unique identifier that is recorded in the toll system database along with the time and location at the time of capture. Automatic license plate reader/recognition technology (ALPR) also is used in automated tolling. Cameras mounted on toll gates capture vehicles' license plate numbers using image recognition technology and store the numbers in the toll system database along with the time and location of the capture. ¹
Data Structure	Structured.
Data Size, Storage, and Management	Gigabytes to terabytes, depending on coverage area and time window. Data often is stored in-house and managed by the toll agency or third-party service provider. Data typically is stored in relational database systems.
Data Accessibility	Database dump files are delivered either through FTP or using portable media.
Data Sensitivity	Electronic toll collection data are considered very sensitive as they contain names, addresses, credit card information, vehicle description, and license plate number. This information poses a threat to the privacy of participants because the systems record when specific motor vehicles pass toll stations. From this information, one can infer the likely location of the vehicle's owner or primary driver at specific times.
Data Costs	Costs will depend on agreements established between toll operators and the agencies requesting the data. Many toll operators may provide at least some data to requesting agencies for free, but it is possible that some toll operators may impose fees if no provision for data access has been made before the request. Data describing the daily whereabouts of thousands and even hundreds of thousands of citizens is currently of high-value for the private sector.
Data Openness	Limited openness, mainly because of the high sensitivity of the data.
Data Challenges	Toll data may be difficult to obtain, both because of its sensitivity and because of the possibility of private-party ownership. Although the data structure is simple and toll data should be able to be reused easily for Big Data analysis, data quality can be an issue. Automatic detection of vehicles at toll gates is known to be error prone, particularly when using ALPR, which is known to have significant error rates. Although data quality may be an issue when performing data analysis that requires the identification of vehicles (e.g., toll calculation or speed checking), TIM data analysis may not require the need to identify vehicles and therefore may not be affected by this issue.

¹ Persad, K., C.M. Walton, S. Hussain. Toll Collection Technology and Best Practices, Texas Department of Transportation (January 2007). Online: https://ctr.utexas.edu/wp-content/uploads/pubs/0_5217_P1.pdf.

A.3 Public Safety Data

Table A-13. Law enforcement, fire and rescue, and EMS CAD system data.

Assessment Criteria	Assessment
Description of Data	<p>Law enforcement, fire and rescue, and EMS agencies use computer-aided dispatch (CAD) to initiate public safety calls for service, dispatch, and to facilitate and maintain communications and the status of responders in the field. CAD typically consists of a suite of software packages and modules that provide interfaces and services for call-takers, dispatchers, and field personnel. CAD includes:</p> <ul style="list-style-type: none"> • Logging on/off times of personnel. • Generating and archiving incidents. • Assigning field personnel to incidents. • Updating incidents and logging those updates. • Generating case numbers for incidents. • Timestamping every action taken by the dispatcher. <p>Relevant data elements include TIM timestamps, notification, dispatch, arrival/departure of agency responders, type of incident, disposition, and other incident details.¹</p>
Who Collects, Maintains, and Owns the Data	<p>Many of more than 12,000 Individual law enforcement agencies.</p> <p>Many of nearly 30,000 local fire departments.</p>
How the Data Are Collected	Human- and auto-populated using commercial CAD software and in-house systems.
Data Structure	Semi-structured to structured.
Data Size, Storage, and Management	<p>Megabytes (spreadsheets or PDFs) to gigabytes (relational databases) to terabytes (large Oracle databases). Data is typically managed and stored in-house at the local level or by third parties; maintained for several years.</p> <p>Data management and data interoperability procedures vary widely across the U.S.</p>
Data Accessibility	FTP and web download are typically available for single or limited incidents upon request; live public facing views share limited information; some CAD systems are integrated with TMCs.
Data Sensitivity	Most data are public record, but some that contain sensitive data fields, criminal investigation information, and criminal history information are not made available outside the collecting agency.
Data Costs	Free. Although some minor cost may be incurred to maintain data-sharing infrastructure.
Data Openness	Limited openness as full (filtered) data is available upon request only.
Data Challenges	<p>Time and cost to fill requests.</p> <p>Presence of sensitive data in data requested connection can complicate sharing as it would involve criminal justice systems.</p> <p>CAD data is recorded using an event database format, that is each row is an event combining an action such as “responder arrived” or “responder departed” with a timestamp. This data organization is ideal for collection but can complicate further data extraction and analysis as the data typically sought after is present in more than one record (time on scene, number of responders on the scene).</p>

¹ https://it.ojp.gov/documents/LEITSC_Law_Enforcement_CAD_Systems.pdf.

Table A-14. Emergency communication center (ECC)/911 call center/public safety answering point (PSAP) data.

Assessment Criteria	Assessment
Description of Data	Emergency communication centers (ECCs), also called <i>911 call centers</i> and <i>public safety answering points (PSAPs)</i> , are responsible for answering the 911 system for a geographic expanse following National Emergency Number Association (NENA) data standards.
Who Collects, Maintains, and Owns the Data	Approximately 6,500 locations across the United States serve as ECCs/PSAPs.
How the Data Are Collected	Incoming 911 calls are answered at the ECC/PSAP of the governmental agency that has jurisdiction over the caller's location. Location management via ANI (automatic number identification) and ALI (automatic location information) is the foundation of 911 data collection and call origination. With the location of the caller, based on telephone service provider information (landline or cellular), the call to 911 is first routed to the correct PSAP. When the 911 call arrives at the appropriate ECC/PSAP, it is answered by a specially trained operator or dispatcher. For landline calls, computer-aided dispatch (CAD) software uses the telephone number to retrieve and display the name, number, and location of the caller to the operator in near-real time. ¹ For wireless calls, the location is either handset based (GPS) or network based (towers). The integration of ANI/ALI functionality in modern CAD systems is common. The operator uses CAD software and interface to input information as described in Table A-13.
Data Structure	Structured and semi-structured (CSV, XML, RDF, JSON).
Data Size, Storage, and Management	Megabytes to gigabytes depending on coverage area and time frame. Data is managed and stored in-house at the local level or by third parties; maintained for several years. Typically, CAD systems use relational databases to store 911 data and flat file storage to archive it. How the data is managed and how interoperable it is varies widely across the United States. The Association of Public-Safety Communications Officials (APCO) and National Emergency Number Association (NENA) have jointly issued APCO/NENA ANS 1.107.1.2015, Standard for the Establishment of a Quality Assurance and Quality Improvement Program for Public Safety Answering Points, a voluntary standard that defines the recommended minimum components of a quality assurance/quality improvement (QA/QI) program within a public safety communications center. It recommends effective procedures for implementing the components of the QA/QI program to evaluate the performance of public safety communications personnel. ²
Data Accessibility	Data is typically accessible on request due to possible sensitivity in the data (criminal investigations, personal identification, comments). Redacted or partial 911 data can be found on https://www.data.gov from a variety of agencies (e.g., all police responses within the city of Seattle) and is refreshed at a variety of rates (e.g., 4 hours).
Data Sensitivity	Sometimes (criminal investigations, personal identification, comments).
Data Costs	Free.
Data Openness	Limited openness, as full data is available only upon request.

Assessment Criteria	Assessment
Data Challenges	<p>Some prominent standards from national organizations exist and are being implemented, but there is no national standard or regulatory authority. Consequently, among the 6,000+ PSAPs nationwide, only a few have implemented standards that enable operational or data analytics assessments. This can render the integration and analysis of 911 data more challenging and untenable from a time, cost, and resource perspective.</p> <p>Partial or redacted datasets are publicly available. Additional analytical value will be found in complete datasets, but access to the full dataset may be challenging due to local and state law restrictions.</p>

¹ https://en.wikipedia.org/wiki/Enhanced_9-1-1.

² APCO/NENA ANS 1.107.1. Standard for the Establishment of a Quality Assurance and Quality Improvement Program for Public Safety Answering Points (2015). Online: <https://www.apcointl.org/doc/911-resources/apco-standards/600-11071-2015-quality-assurance/file.html>.

Table A-15. Public safety digital video data.

Assessment Criteria	Assessment
Description of Data	As with transportation agencies, public safety agencies make use of various types of digital video technologies, including CCTV, ALPR, dashboard cameras, and wearable cameras. Public agencies use ALPR to capture license plate numbers and compare them to one or more databases of vehicles of interest and alert authorities when a vehicle of interest has been observed. ¹ Dashboard cameras and/or wearable cameras are used to monitor traffic stops and other enforcement activities. Basic dashboard cameras are video cameras with built-in or removable storage media that constantly record. More advanced dashboard cameras can have audio recording, GPS logging, speed sensors, accelerometers, and uninterrupted power supply capabilities. ² Body cameras range from small, low-resolution options to high-definition options.
Who Collects, Maintains, and Owns the Data	Public safety agencies.
How the Data Are Collected	Via various types of cameras. Video stream is either recorded in a continuous loop of a few hours on the camera device or streamed directly to a data center where it is recorded and archived.
Data Structure	Unstructured (video) and semi-structured (XML, JSON, CSV).
Data Size, Storage, and Management	<p>Terabytes. Like transportation agency highway cameras, video images from fixed roadway/venues are not always stored. Dash cams will record up to about 2 GB (about 6 hours) of video on a loop that refreshes continuously. Videos may be saved on a secure digital (SD) card or on an external drive, and typically download automatically to a server without human intervention.³ Body cameras use SD or microSD cards for storage. Depending on the model, they support anywhere from 4 GB to 120 GB of video storage and upload their video for storage automatically to a server without human intervention.⁴</p> <p>As an example, the Birmingham police initially purchased 5 TB of online storage to store the video from 319 body cameras. In just 2 months, the department used 1.5 TB of its allotment and was on track to exceed the 5 TB limit in about 6 months.⁵ Depending on the agency, either plain storage or media library software including metadata management is used.</p>
Data Accessibility	Data dump from server or device storage, upon request.
Data Sensitivity	Yes (faces, license plates, etc.).
Data Costs	A cost is incurred in the retention of video images. The amount and quality of data stored on storage media is subject to compression ratios, images stored per second, and image size, and the amount of data stored is affected by the retention period of the videos or images.
Data Openness	Not open. Sensitive and accessible on request only.

Assessment Criteria	Assessment
Data Challenges	<p>Dependency on wireless connection can be a technical obstacle.</p> <p>Institutional, technical, and legal – In most cases, video is stored or archived by law, but retention laws have not kept pace with video technology and greatly limit archiving. There are numerous legal restrictions regarding the acquisition, use, and storage of video images by law enforcement. Also, video that is not archived automatically from camera devices must be archived manually on a regular basis; failure to do so leads to the video data being overwritten and lost. These challenges greatly limit the quantity of video content that could be mined. Also, video data collection is not uniform across space, time, and quality:</p> <ul style="list-style-type: none"> • Equipment failures of cameras, communications networks, and recording systems can also increase the lack of coverage when maintenance is not able to remedy failure quickly. • Weather conditions can greatly affect the quality of the video collected making it impossible in some cases to extract metadata under conditions such as snow and rain. • Video resolution varies widely between devices and many devices are still recording video at low resolution which affects its ability to be processed effectively. • Video container and compression standards vary widely between equipment and manufacturers. These standards are often proprietary and cannot be converted easily to a common standard without losing some video data integrity. <p>These factors lead to video datasets that are sparse and non-uniform making it challenging to extract information or patterns from them.</p>

¹ <http://www.iacp.org/ALPR-About>.

² <https://www.lifewire.com/types-of-dash-cameras-534889>.

³ <http://www.randmcnally.com/support/faqs/what-is-the-recording-time-on-the-dash-cam-and-how-are-video-files-stored>.

⁴ <http://www.toptenreviews.com/electronics/photo-video/best-wearable-cameras/>.

⁵ <http://www.computerworld.com/article/2979627/cloud-storage/as-police-move-to-adopt-body-cams-storage-costs-set-to-skyrocket.html>.

Table A-16. Towing and recovery data.

Assessment Criteria	Assessment
Description of Data	Catalog of calls for service and various timestamps for response, such as dispatch, arrival, and departure times, as well as type of assistance, equipment, insurance, and financial transactions.
Who Collects, Maintains, and Owns the Data	Towing companies.
How the Data Are Collected	Data collection is typically manual or electronic. A few towing companies do not collect any data, relying on the state police or transportation dispatch for this data. Some towing companies utilize computer-aided dispatch (CAD) equipment coupled with touch screen mobile data terminals (MDTs) within each of the trucks. Electronic systems allow for the accurate mapping and recording of each dispatch and arrival time on all calls. Software programs allow for cloud-based management of dispatched jobs/trucks on a map in real time.
Data Structure	Semi-structured to fully structured.
Data Size, Storage, and Management	Megabytes. Private company database in-house or in the cloud.
Data Accessibility	Contact for data dump.
Data Sensitivity	Yes (financial transactions and company business practices).
Data Costs	Unknown. Data is private and may not be available for sale.
Data Openness	Not open. Data are proprietary to the towing and insurance entities.
Data Challenges	A predominance of individual providers still do not maintain any data at all or maintain only limited data through a paper log or spreadsheet. In-house systems rarely go outside of the business.

A.4 Crowdsourced/Social Media Data

Table A-17. Waze data.

Assessment Criteria	Assessment
Description of Data	<p>Data generated by users of the Waze community-based navigation mobile application, including real-time road information such as crashes, construction, police presence, road hazards, traffic jams, etc. Also captured is confirmation of this information by other Waze users through either a “thumbs-up” or “thumbs-down” response or through detailed messages. Additionally, Waze automatically records the speed at which users travel on the roadways and captures messages sent between users through the mobile app.</p> <p>Data elements relevant to TIM include incidents’ reported times, incident details (e.g., number/types of vehicles involved), incident clearance times, traveler sentiments, speeds.</p>
Who Collects, Maintains, and Owns the Data	Waze.
How the Data Are Collected	Road users report events using the Waze mobile application.
Data Structure	Semi-structured (CSV, JSON).
Data Size, Storage, and Management	Gigabytes to terabytes, depending on coverage and timeframe. A Waze event dataset (not including speed data) covering the entire nation from 2013 to 2016 contains about 120 million reports and has a size of about 120 GB. Waze data is managed on both the Amazon Web Services and the Google Cloud Platform clouds (since 2013) and Waze uses cloud file storage, NoSQL databases, and relational databases to manage its data. Waze data is archived indefinitely.
Data Accessibility	Only accessible through partnership with Waze. Waze data is shared through its Waze Connected Citizen Program, which provides either processed, cleaned data or web applications such as Waze Traffic View or third-party applications using Waze data such as Genesis PULSE (EMS support application).
Data Sensitivity	User information that Waze collects may be sensitive. Users agree to Waze’s use of the data (with PII) but not to sharing this information with other entities.
Data Costs	Not typically any cost, only a requirement for data sharing. Waze’s Connected Citizen Program seeks to improve the use of the data for the community. States can develop a partnership with Waze and share data to access Waze data. Private entities willing to become Waze partners may have to pay a cost to access the Waze data, but the cost of that access is not public.
Data Openness	Limited openness (partners only).
Data Challenges	<p>Waze data is a combination of both sensor data (speeds) and crowdsourced data (alerts or events), and as such does not contain perfect data. Although the error rate of location sensors on mobile phones is well known and can be circumvented using readings from other sensors in the vicinity, alerts sent by humans can be unreliable (e.g., pushing the wrong button, inaccuracies in what is happening/reporting). Free text is also used as part of Waze alert reports to provide additional details, which also allows for human error (e.g., misspellings, orthography). Waze does provide a way to assess the reliability and quality of its data by adding to its alert reports a reliability/confidence index ranging from 1–10. High-quality and highly reliable reports do not constitute most of the Waze alert reports, and some events/alerts may remain fuzzy or imprecise. Waze does not provide direct access to its raw data (e.g., how many people reported each incident, how many thumbs-up responses a report received), which may impair data users’ ability to assess the accuracy of Waze events/alerts.</p>

Table A-18. Twitter data.

Assessment Criteria	Assessment
Description of Data	<p>“Tweets” are generated by Twitter users using the Twitter app. Data includes tweet text (up to a 144-character stream), an associated timestamp, and possible attachments (e.g., photos, videos). When users allow Twitter to share their location, tweet locations (latitude, longitude) also are captured.</p> <p>Data elements relevant to TIM include incidents’ reported times, incident details (e.g., number/types of vehicles involved), incident photos or videos, incident clearance times, traveler sentiments.</p>
Who Collects, Maintains, and Owns the Data	Twitter.
How the Data Are Collected	Twitter collects, stores, and publishes all its users’ “Tweets” submitted using mobile phone, website, or IoT devices leveraging the Twitter API. Machine-submitted tweets can relate sensor readings or alerts, and it is not uncommon for software architects to leverage Twitter as a communications layer for their own software platform.
Data Structure	Semi-structured (CSV, JSON).
Data Size, Storage, and Management	<p>Terabytes. The data size of an average tweet is a few kilobytes, not counting attached media. Twitter manages and stores about 200 billion tweets a year, which is about 200 TB of data. Twitter manages its data using a custom developed and open-source data store including a large-scale, key-value store called Manhattan, a graph database called FlockDB, an open-source database called MySQL, as well as various storage and caching services.</p> <p>Tweets have been continuously archived by Twitter since 2006. Twitter provides a service (Twitter Archive) to allow its users to search and download its archive.</p>
Data Accessibility	Twitter possesses multiple APIs allowing developers to process the real-time stream of tweets, to search tweets by text, user, hashtag, location, date, and so forth. Third-party applications use the tweet stream to create additional data mining and visualization interfaces that can help augment (e.g., text mining, categorize, reverse geocoding) and visualize the raw Twitter data to help discover its content. These third-party services often require users to register and pay to search, analyze, and visualize the data. Examples of Twitter third-party applications include web services such as Tweepstmap, Twitonomy, and Mentionmap.
Data Sensitivity	PII, including name, user profile, and sometimes real-time user location (sensitive even if voluntarily published).
Data Costs	Twitter API is free with some limitations (e.g., how much at once, frequency). Costs occur when using third-party APIs or software to mine the Twitter dataset.
Data Openness	Open (tweets are public).

Assessment Criteria	Assessment
Data Challenges	<p>Two of the main challenges of using Twitter data are the large quantity of tweets generated every minute and the free text structure of its content (except for hashtags). When processing the Twitter data stream to monitor for TIM-relevant information or events, the text of each tweet would need to be parsed, analyzed using text mining, correlated with similar tweets, and counted to establish the location and veracity of a detected event. This analysis is challenging, as it needs to be done in real time, there may not be enough tweets describing the incident, and users are likely to use different vocabulary to describe the incident.</p> <p>Twitter uses hashtags to qualify and categorize the free text content of its tweets. Twitter users can create hashtags and use them when needed within their message. Some commonly used hashtags (e.g., <i>#accident</i>) exist, but these are too general to allow tweets to be filtered to extract relevant TIM content, and there is no control over how hashtags are used by Twitter users.</p> <p>Not all tweets are geolocated, which can make it difficult to use tweet text to detect the occurrence of roadway events such as incidents or the free-flow recovery.</p>

A.5 Advanced Vehicle Systems Data

Table A-19. Automated vehicle location (AVL) system data.

Assessment Criteria	Assessment
Description of Data	<p>AVL is a means for automatically determining and transmitting the geographic location of a vehicle with details that include date, time, address, longitude, and latitude. AVL is used to manage vehicle fleets, such as service vehicles, public transportation vehicles, emergency vehicles, and commercial vehicles. AVL data includes real-time temporal and geospatial data (polled every few seconds), as well as vehicle logs (e.g., vehicle number, operator ID, route, direction, arrival/departure times).</p> <p>Dispatchers can get a real-time snapshot of driver adherence to a route, provide customers with an estimated time of arrival, and communicate directly with drivers. Public safety agencies can use AVL technology to improve response times by dispatching the closest vehicles for emergencies.</p>
Who Collects, Maintains, and Owns the Data	Fleet owners (e.g., Safety Service Patrols, public safety agencies, transit agencies, towing companies).
How the Data Are Collected	A vehicle's position is located and tracked using a geographic positioning system (GPS) electronic device. The vehicle's position is either stored for later analysis or wirelessly communicated to the home base dispatch.
Data Structure	Semi-structured (CSV) or structured (SQL).
Data Size, Storage, and Management	<p>Gigabytes to terabytes, depending on geographic coverage and timeframe. Stored in-house or via a cloud-hosted service.</p> <p>As of 2017, an available GPS transmitting device cost less than \$20, was smaller than the size of a human thumb, was able to run 6 months or more between battery charges, and could communicate easily with smartphones.¹</p> <p>A transit system with about 200 vehicles will generate about 3,000,000 records annually. The leading GPS fleet management solutions should be able to retrieve historical data from any vehicle in a fleet as far back as when the vehicles were equipped with GPS tracking devices.²</p>
Data Accessibility	<p>For data owners, data must be uploaded from the on-board computer to the central computer. Newer systems usually include an automated, high-speed communication device through which data is uploaded daily (e.g., when vehicles are fueled). Older systems rely on manual intervention, such as exchanging data cards or attaching an upload device, which adds a logistical complication.³</p> <p>For non-owners, data may be obtained via FTP, data dump, or web services (if access is granted).</p>
Data Sensitivity	For some agencies, the AVL data may include residential data for personnel that operate the SSP or law enforcement vehicles. Data may require redaction before sharing.
Data Costs	For already-equipped vehicles, there should be no costs for obtaining data from publicly operated systems.
Data Openness	The data can be shared upon request, but it is generally not open.
Data Challenges	The absence of an effective upload mechanism can render an otherwise promising data collection system useless for off-line data analysis. ¹

¹ https://en.wikipedia.org/wiki/Automatic_vehicle_location.

² Malcolm, J. Automatic Vehicle Location Technology is Valuable for Fleets of All Sizes (October 7, 2014). Online: <https://www.hubs.com/power/explore/2014/09/automatic-vehicle-location-technology-is-valuable-for-fleets-of-all-sizes>.

³ Furth, P. G., B. Hemily, T. H. J. Muller, and J. G. Strathman. *TCRP Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management*. Transportation Research Board of the National Academies, Washington, D.C., 2006.

Table A-20. Event data recorder data.

Assessment Criteria	Assessment
Description of Data	An event data recorder (EDR) is a digital recording device that allows the monitoring and recording of telemetric data reflecting activities inside and outside of an automobile. An estimated 92 percent of new passenger vehicles had EDRs as of 2006. EDRs have been required in new vehicles since 2013 and are required to record data in a standard format to make its collection and processing easier. ¹ A NHTSA regulation passed in 2012 provides that if a vehicle has an EDR, it must track 15 specific data elements, including speed, steering, braking, acceleration, seatbelt use, and, in the event of a crash, force of impact and whether airbags deployed. ²
Who Collects, Maintains, and Owns the Data	Data resides in the EDR of individual vehicles. Enacted in December 2015, the federal Driver Privacy Act provides that information collected belongs to the owner or lessee of the vehicle. ³
How the Data Are Collected	EDRs collect event information from the in-vehicle network and from the vehicle GPS antenna. EDRs record event data in a continuous loop in a memory bank capable of storing a few minutes of data. Most EDRs are built into a vehicle's airbag control module and, upon a crash, are triggered to save the last 5 seconds of recorded information (e.g., airbag deployment, vehicle speed, engine throttle, and driver safety belt use) into a tamper-proof memory. ¹
Data Structure	Semi-structured.
Data Size, Storage, and Management	Megabytes. Data from the EDR is stored on stacked memory boards inside a crash-survivable memory unit. Most EDRs are programmed to record data in a continuous loop, writing over information again and again until the unit is triggered to save the data in the event of a crash. When a crash occurs, the device automatically saves up to 5 seconds of data representing the moments immediately before, during, and after an incident. ²
Data Accessibility	EDR data can be retrieved two ways: (1) via a connection to the vehicle's on-board diagnostics (OBD) port or (2) the EDR itself may be removed from the vehicle and the data retrieved directly. Downloading the data after a crash requires the use of a specialized data-retrieval tool kit that consists of hardware, software, and a special cable that plugs into the car's OBD port or the EDR itself. ⁴ The federal Driver Privacy Act of 2015 places limitations on data retrieval from EDRs. ³ Police, insurers, researchers, automakers, and others may gain access to the data with owner consent. Without consent, access may be obtained through a court order. For crashes that do not involve litigation, especially when police or insurers are interested in assessing fault, insurers may be able to access the EDRs in their policyholders' vehicles based on provisions in the insurance contract requiring policyholders to cooperate with the insurer. Some states prohibit insurance contracts from requiring policyholders to consent to access. ⁴
Data Sensitivity	EDR data characterizes driver behavior and as such can be used in court as evidence. Civil liberty and privacy groups have raised concerns about the implications of data recorders "spying" on drivers. ²
Data Costs	Crash data-retrieval kits cost between \$2,000 and \$10,000; however many law enforcement agencies have equipment or solicit vehicle dealerships for assistance.
Data Openness	The data is not open, as it requires custom equipment and the consent of the vehicle owner or a court order to be extracted from the EDR.

(continued on next page)

Assessment Criteria	Assessment
Data Challenges	Due to current technology, costs and data privacy issues associated with EDR data collection, and storage, EDR data cannot be collected and aggregated. Typically, EDR data must be downloaded one vehicle at a time after receiving the consent of the vehicle owner or a court order. Alternative ways to access EDR-like data have been created by third parties such as auto insurance companies. On-board telematics devices (e.g., SnapShot® from Progressive insurance or the Automatic dashboard adapter and app by Automatic Labs™) use the driver's mobile phone to obtain some of the data collected by the EDR, streaming it to large data stores where the data is analyzed to optimize insurance company risks. These third-party devices require a user agreement to be signed by the driver that allows the third-party to collect and use its vehicle data, effectively circumventing the data privacy issue. The datasets created by these third parties may be an alternative way to access EDR data partially or fully without having to collect it one vehicle at a time (see vehicle telematics systems data).

¹ Insurance Institute for Highway Safety, Highway Loss Data Institute, Event Data Recorders. Online: <http://www.iihs.org/iihs/topics/t/event-data-recorders/topicoverview> (accessed February 2017).

² Rafter, M. V. Decoding What's in Your Car's Black Box, Who Owns the Data and Who Can Tap It? (Edmunds, July 22, 2014). Online: <https://www.edmunds.com/car-technology/car-black-box-recorders-capture-crash-data.html>.

³ National Conference of State Legislatures, Privacy of Data from Event Data Recorders: State Statutes. Online: <http://www.ncsl.org/research/telecommunications-and-information-technology/privacy-of-data-from-event-data-recorders.aspx> (accessed February 2017).

⁴ Vehicle Telematics: A Useful Litigation Tool for Attorneys, A Boon to Insurers and the Privacy Concerns Big Data Raises for Us All. Klieman & Lyons (September 20). Online: <http://www.kliemanlyons.com/2014/09/vehicle-telematics-a-useful-litigation-tool-for-attorneys-a-boon-to-insurers-and-the-privacy-concerns-big-data-raises-for-us-all> (accessed March 2017).

Table A-21. Vehicle telematics systems data.

Assessment Criteria	Assessment
Description of Data	<p>Telematics is the transfer of data to and from a vehicle. Vehicle telematics systems combine a GPS system with on-board sensors and diagnostics to record speed, engine throttle, braking, ignition cycle, whether the driver was using a safety belt, airbag deployment, and the physics of crash events including crash speed, change in forward crash speed, maximum change in forward crash speed, time from beginning of crash event at which the maximum change in forward crash speed occurs, the number of crash events, the time between crash events and whether the device completed recording.¹</p> <p>Unlike Event Data Recorders (EDRs) that collect and store a few seconds of data immediately before and after a crash, telematic systems continuously record all types of second-by-second data about vehicles and driver behavior, sometimes for years at a time. Telematic technologies collect raw vehicle data and overlay this information with GIS mapping data (e.g., road type, speed limits). The data is then “broadcast” via data links such as Wi-Fi, GPS, Bluetooth, 3-axis accelerometers, and mobile broadband communications to auto manufacturers, fleet owners, and insurance companies. As the cost of enabling mobile broadband communications has fallen, automakers are increasingly embedding telematics in vehicles. Some form of telematics systems is now available in an estimated 70 percent of vehicles built since 2011.¹</p> <p>Advanced Automatic Crash Notification (AACN) is a component of telematics. The AACN Joint APCO/NENA Data Standardization Workgroup created the Vehicle Emergency Data Set (VEDS) to specifically address the need for an open standard format to be used for all providers and consumers of vehicle telematics information. VEDS is an XML-based data standard that provides useful and critical data elements and the schema set needed to facilitate an efficient emergency response to vehicular emergency incidents.²</p> <p>At the fringes, the term <i>telematics</i> also is used to describe “connected car” features in general, which include live weather, traffic and parking information on the dashboard, apps, voice-activated features, and social media integration.³</p>
Who Collects, Maintains, and Owns the Data	Auto manufacturers, telematics service providers (TSPs), insurance companies, and fleet owners.
How the Data Are Collected	<p>Data is collected by connecting to in-vehicle sensors using four distinct categories of telematics solutions—dongles, black boxes, embedded telematics, and smartphones:⁴</p> <ul style="list-style-type: none"> • Dongles are self-installed devices that are often provided by car insurers or may be purchased by the vehicle owner to monitor/record vehicle operation and/or driver behavior. • Black-box systems are professionally installed to monitor driving behavior and vehicle systems. • Embedded telematics are installed by some manufacturers and provide services such as remote diagnostics, navigation, and infotainment services. • Smartphones can work as stand-alone devices or be linked to vehicles’ systems (e.g., through Bluetooth) to transmit a variety of information to and from the car.
Data Structure	Raw data from telematics devices is in CSV format (semi-structured).
Data Size, Storage, and Management	Data is stored within the collection devices described above except where the devices interface with remote systems, call centers, and management systems. Some telematics systems, such as the ones deployed by auto insurance companies, store vehicle data in file storage, relational, or NoSQL databases for later analysis of the behavior of customers. Archiving of data varies depending on the data owner and chosen telematics solution.

(continued on next page)

Assessment Criteria	Assessment
Data Accessibility	Each automaker and insurer uses its own proprietary telemetry or usage-based insurance (UBI) programs to access and store telematics data. The telematics data can only be accessed via a court order.
Data Sensitivity	Telematics data, and especially the aggregation of the data, presents privacy challenges for consumers, the courts, law enforcement, automakers, insurers, and the telematics industry. Privacy settings and arrangements depend on the service. For example, BMW's ConnectedDrive may "collect and retain an electronic or other record" of a person's location or direction of travel at a given time. The OnStar® system by General Motors "complies with its legal obligation to court orders or subpoenas" but doesn't "share data with law enforcement absent a court order unless it is necessary to protect the safety of its customers or others." Ford has said that its Sync program doesn't track or transmit data continuously from a vehicle and that no data is transmitted from the vehicle without the customer's consent, indicating that "[l]ocation data is only shared with our partners when necessary to fulfill the services requested by the customer." ¹
Data Costs	N/A. Data can only be obtained through a court order.
Data Openness	Not open, as it requires a court order to be accessed.
Data Challenges	Typically, these systems are used with the consent of the vehicle owner and access to data is restricted to uses defined by the user/owner. Telematics system user agreements may allow for the collected data to be reused or sold to others than the telematics systems owner and the driver.

¹ Vehicle Telematics: A Useful Litigation Tool for Attorneys, A Boon to Insurers and the Privacy Concerns Big Data Raises for Us All, Klieman & Lyons (September 20). Online: <http://www.kliemanlyons.com/2014/09/vehicle-telematics-a-useful-litigation-tool-for-attorneys-a-boon-to-insurers-and-the-privacy-concerns-big-data-raises-for-us-all> (accessed March 2017).

² Association of Public-Safety Communications Officials, Comm Center & 911, AACN/VEDS Overview. Online: <https://www.apcointl.org/resources/telematics/aacn-and-veds.html> (accessed February 2017).

³ Carter, J. Telematics: What You Need to Know, TechRadar, June 27, 2012. Online: <http://www.techradar.com/news/car-tech/telematics-what-you-need-to-know-1087104> (accessed February 2017).

⁴ Karapiperis, D., B. Birnbaum, A. Brandenburg, S. Castagna, A. Greenberg, R. Harbage, A. Obersteadt. Usage-Based Insurance and Vehicle Telematics: Insurance Market and Regulatory Implications, National Association of Insurance Commissioners and the Center for Insurance Policy and Research (March 2015). Online: http://www.naic.org/documents/cipr_study_150324_usage_based_insurance_and_vehicle_telematics_study_series.pdf.

Table A-22. Automated and connected vehicle, traveler, and infrastructure data.

Assessment Criteria	Assessment
Description of Data	<p>Automated vehicles are those in which at least some aspect of a safety-critical control function (e.g., steering, throttle, or braking) occurs without direct driver input. Automated vehicles may be autonomous (i.e., use only vehicle sensors) or may be connected (i.e., use communications systems such as connected vehicle technology, in which cars and roadside infrastructure communicate wirelessly).¹ NHTSA has classified vehicle automation into six levels:²</p> <ul style="list-style-type: none"> • Level 0: The human driver does all the driving. • Level 1: An advanced driver assistance system (ADAS) on the vehicle can assist the human driver with either steering or braking/accelerating. • Level 2: An ADAS on the vehicle can control both steering and braking/accelerating under some circumstances. The human driver must continue to pay full attention and perform the rest of the driving task. • Level 3: An ADAS on the vehicle can perform all aspects of the driving task under some circumstances. In those circumstances, the human driver must be ready to take back control when the ADAS requests the human driver do so. In all other circumstances, the human driver performs the driving task. • Level 4: An ADAS on the vehicle can perform all driving tasks and monitor the driving environment in certain circumstances. The human need not pay attention in those circumstances. • Level 5: An ADAS on the vehicle can do all the driving in all circumstances. The human occupants are just passengers and need never be involved in driving. <p>Connected vehicles are vehicles that use any of a number of different communication technologies to communicate with the driver, other vehicles on the road (V2V), roadside infrastructure (V2I), and the cloud (V2C).³</p> <p>A connected traveler is one that uses a mobile device that generates and transmits status data via DSRC, Wi-Fi, Bluetooth, or cellular. Messages generated and distributed by connected travelers could include data representing the traveler's location, trip characteristics (e.g., speed), mode and status (e.g., riding in a car, riding on transit, walking, biking, etc.)(Gettman et al. 2017).</p> <p>DSRC technology generates, sends, and receives Basic Safety Messages (BSMs) to other vehicles and to roadside equipment (RSEs) at high frequency (10 times per second) and with very low latency (50 ms from transmission to receipt). A Probe Data Message (PDM) encapsulates a string of "snapshots" (a more comprehensive data element than the BSM) to provide vehicle trajectory information over a longer time frame than the local trajectories shared by the BSMs (Gettman et al. 2017).</p> <p>Connected infrastructure includes traditional ITS devices, such as traffic signals, ramp meters, CCTV, RWIS and may eventually evolve to include standard Internet-of-Things (IoT) protocols as IoT technologies continue to mature (Gettman et al. 2017).</p>
Who Collects, Maintains, and Owns the Data	<p>There is no clear property regime for ownership and control of such data. Thirty stakeholders, interviewed by RAND as part of the development of <i>Autonomous Vehicle Technology: A Guide for Policymakers</i>, were asked their opinion about who owned the data obtained by automated vehicles (AVs) as they move, gather, and transmit information. Not a single stakeholder was certain of the answer.²</p>

(continued on next page)

Assessment Criteria	Assessment
How the Data Are Collected	Data are collected via dozens of sensors that collect telematics, driver behavior, and environmental data. Sensors such as forward and side radar sensors, sonar, GPS, LiDAR, cameras, and monitoring systems generate AV and CV data. The amount of data generated is rather large and quickly exceeds the on-board data storage capacity; therefore, it is eventually stored in remote or cloud-based systems. AV and CV data can also be streamed directly to remote systems to be monitored in real time.
Data Structure	Semi-structured. ASN.1, XML, JSON, and CSV.
Data Size, Storage, and Management	Petabytes to zettabytes, depending on the number of vehicles collecting AV and CV data. It is estimated that connected vehicles may generate as much as 25 gigabytes per hour. It is assumed that not all this data will be stored and managed in its raw form, but at this scale cloud file storage and NoSQL databases will be required even for compressed or partial datasets. If all of the emerging data from connected vehicles, travelers, and infrastructure related to traffic operations is stored, the cumulative storage of a typical traffic management agency is estimated to be in the many thousands of terabytes by 2026 (Gettman et al. 2017).
Data Storage	Stakeholders interviewed in the RAND study identified policy questions concerning data use and legal issues (e.g., how long AV data should be maintained and by whom). ²
Data Accessibility	Stakeholders in the RAND study also raised the issue of whether data gathered, produced, or transmitted by AVs will be discoverable in legal proceedings. ² AV/CV aggregation and anonymization methods are being developed to facilitate accessibility.
Data Sensitivity	Some members of the AV industry are already working on how to anonymize vehicle data and aggregate it so that it does not reveal drivers' PII. One stakeholder identified privacy concerning AV data as a critical issue that needs immediate policy attention. Two stakeholders made a comparison to the information captured by EDRs currently installed in automobiles. ²
Data Costs	Unknown and may not be applicable depending on ultimate privacy policies.
Data Openness	Not open at this point.
Data Challenges	Data ownership and privacy issues related to AV communications remain unsettled and an important policy gap. ²

¹ Automated Vehicle Research, U.S. Department of Transportation. Online: https://www.its.dot.gov/automated_vehicle/ (accessed February 2017).

² Anderson, J. M., N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, O.A. Oluwatola. 2016. *Autonomous Vehicle Technology: A Guide for Policy Makers*. RAND Corporation, Santa Monica, CA. Online: http://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR443-2/RAND_RR443-2.pdf.

³ Center for Advanced Automotive Technology, *Connected and Automated Vehicles*. Online: http://autocaat.org/Technologies/Automated_and_Connected_Vehicles/ (accessed February 2017).

A.6 Aggregated Datasets

Table A-23. RITIS data assessment.

Assessment Criteria	Assessment
Description of Data	An automated traffic and emergency management data consolidation, sharing, dissemination, and archiving system. Data include, but are not limited to third-party probe data, DOT ATMS data, National Performance Management Research Data Set (NPMRDS) data, road weather data, CAD data, virtual weigh station data, transit data, and parking spaces available.
Who Collects, Maintains, and Owns the Data	University of Maryland CATT Lab and partners, including state DOTs, public safety agencies, transit agencies, and third-party data providers.
How the Data Are Collected	RITIS data feeds from transportation agencies, public safety agencies, transit agencies, and third-party data providers.
Data Structure	Structured (relational database, geospatial databases) and semi-structured (XML, JSON, GeoRSS).
Data Size, Storage, and Management	Gigabytes, possibly terabytes, depending on the dataset or coverage area. RITIS also collects geospatial and raster (image) data, which is bigger than typical events datasets. All data within RITIS is archived indefinitely.
Data Accessibility	<p>Public safety or DOT employees can register for an account to the RITIS platform by visiting https://www.ritis.org/register. Three types of feeds are available to users:¹</p> <ul style="list-style-type: none"> • The RITIS Filter Web Service, a polling web service, allows consumers to receive data in several different formats (XML, JSON, and GeoRSS). Provides data from the widest array of agency sources and allows consumers to filter data by source agencies and by specific fields. • The JMS Filter utilizes a real-time publish/subscribe model using a Java Messaging Service broker. Upon the initial connection, the subscriber receives a full inventory of devices or events followed by asynchronous incremental updates (from a limited number of data sources). • The XML Filter, an SSL [secure sockets layer] secured web page, provides a list of GZIP-ed XML files with a snapshot of current data. Data consumers poll the page at a set interval to pull the latest snapshot in the XML format (from a limited number of data sources). <p>Data within the RITIS archive can also be downloaded and/or exported so that users can perform their own, independent analyses.</p> <p>Generally, however, data are accessed through web tools that are designed for close inspection of defined events in space and time.</p>
Data Sensitivity	Accounts are not given to the general public or the private sector due to the sensitive nature of some of the data.
Data Costs	The University of Maryland CATT Lab makes the RITIS platform available to registered users for a fee, which depends on the services purchased.
Data Openness	Limited openness. RITIS was first and foremost designed to support the transportation side of emergency management (command center coordination) and as such does not share its data with the general public. The RITIS platform focuses on providing visualizations and user interfaces designed to support emergency management real-time decisions and in addition provides web services that can allow other applications to be integrated with RITIS. Users may be limited to viewing only their own data.

(continued on next page)

Assessment Criteria	Assessment
Data Challenges	<p>Although RITIS provides analysis tools and visualizations, its data-sharing limitations do not allow its users to fully exploit the data it collects. It is unclear if the data that RITIS stores is stored in individual databases or if it is stored in a single data repository where all its datasets can be explored at once. Many of the visualizations provided in the RITIS documentation are GIS based and allow the geospatial merger of distinct databases and datasets without fully integrating them.</p> <p>RITIS does not provide information about data coverage, quality, or usability. Its documentation provides examples of advanced tools and visualizations in various transportation management aspects. No indication is given as to how many of the RITIS users can run these analysis and visualizations using their own data. Although RITIS contains data from a wide array of data sources, it is unclear what data sources are available for different locations and what data elements are included in the various data sources (e.g., ATMS data varies widely agency to agency and sometimes even TMC to TMC within an agency).</p>

¹ RITIS Platform, Features & Applications Overview, CATT Laboratory, University of Maryland (2015). Online: <http://www.cattlab.umd.edu/files/RITIS%20Overview%20Book-2-2-15%20FINAL.pdf>.

Table A-24. National Performance Measures Research Data Set (NPMRDS).

Assessment Criteria	Assessment
Description of Data	The NPMRDS provides vehicle probe-based data for passenger automobiles and trucks. NPMRDS is a monthly archive of average travel times, reported every 5 minutes when data is available, on the National Highway System. Separate average travel times are included for “all traffic,” freight and passenger travel. ¹
Who Collects, Maintains, and Owns the Data	INRIX Traffic.
How the Data Are Collected	INRIX aggregates GPS probe data from a wide array of commercial vehicle fleets, connected cars, and mobile apps.
Data Structure	Semi-structured (CSV, shape files).
Data Size	Although the source data (INRIX) is Big Data, the size of the NPMRDS data files downloaded through RITIS’s “Massive Data Downloader” tool will depend on the size of the query (e.g., date range, number of roadways, etc.). Downloading the data from a website will eventually run into an upper limit to the size of the file than can be downloaded – e.g., client limitations, network bandwidth limitations (it could take 24 hours to download 100GB of data), limitations in the software handling the http transfer, storage capabilities of receiving desktop computer.
Data Storage and Management	Source data (INRIX) Before July 2017 the NPMRDS data was provided by HERE Technologies. Since July 2017, the data has been provided by INRIX through the CATT Lab’s RITIS system. A discontinuity in the data has been caused by the change in data providers. Agencies working with the dataset will have to adjust to a new kind of data/model (data doesn’t behave the same, doesn’t have same limitations).
Data Storage	Source data (INRIX)—Big Data infrastructure.
Data Accessibility	Available through the RITIS “Massive Data Downloader,” the official portal for all downloads of the NPMRDS. User must be a public agency and obtain a log-in to access the data. The Massive Data Downloader allows access to a sample of the data available to INRIX. The data is not accessible by a machine.
Data Sensitivity	None.
Data Costs	Free to states and MPOs.
Data Openness	Data are not open; only samples of data are available through the Massive Data Downloader; data are shared with state transportation agencies and MPOs only.
Data Challenges	Data cannot be used as a data source for Big Data (even though it’s based on a Big Data data source). The data cannot be accessed/custom-queried (the tool is designed for a pre-defined set of basic queries). Data has to be manually run on the RITIS system rather than put into a data lake for more in-depth analysis. Agencies would need to go directly to INRIX or a competitor, such as HERE Technologies to get the data for these purposes. Previously (when the data was provided by HERE) the data could be downloaded and put into a data repository for these purposes.

¹ National Operations Center of Excellence. Online: <https://transportationops.org/event/national-performance-management-research-data-set-npmrds-users-quarterly-technical-assistance>.

Table A-25. Meteorological Assimilation Data Ingest System (MADIS) and MADIS Integrated Mesonet—National Oceanic and Atmospheric Administration (NOAA).

Assessment Criteria	Assessment
Description of Data	The Meteorological Assimilation Data Ingest System (MADIS) is a meteorological observational database and data delivery system. MADIS runs operationally at the National Weather Service (NWS) National Centers for Environmental Prediction (NCEP) Central Operations (NCO). MADIS subscribers have access to an integrated, reliable, and easy-to-use database containing real-time and archived observational datasets. Also available are real-time gridded surface analyses. The surface analysis grids assimilate all the MADIS surface datasets, including the highly dense Integrated Mesonet data. ¹ The MADIS Integrated Mesonet is a unique collection of thousands of mesonet stations from local, state, and federal agencies, and private firms that help provide a finer density, higher frequency observational database for use by the greater meteorological community. ²
Who Collects, Maintains, and Owns the Data	NOAA.
How the Data Are Collected	MADIS ingests data from NOAA data sources and non-NOAA providers, decodes the data then encodes all the observational data into a common format with uniform observational units and timestamps. MADIS collects data from 33 state DOTs. All DOT observations are part of the MADIS Integrated Mesonet. ² MADIS also performs multiple data validation, checks, and cross correlations of nearby sensors data to maximize the quality of its dataset. MADIS data can be accessed raw or corrected. Many of the implementation details that arise in data ingest programs are automatically performed.
Data Structure	The MADIS is stored using NetCDF files, a scientific file format commonly used to store weather data.
Data Size, Storage, and Management ³	Gigabytes to terabytes. Daily totals for the government, research, and education Integrated Mesonet dataset—680 MB (compressed), 5.67 GB (uncompressed). ⁴ The data schedule is set by provider and ranges from every 5 minutes to once per day. Users can request data from July 2001 to the present. Quality checks are conducted, and the integrated datasets are stored along with a series of flags indicating the results of the various quality control checks.
Data Accessibility	<p>MADIS provides several methods for users to access the data. MADIS data is made available through using multiple data transfer protocols via the Internet: file transfer protocol (FTP), Unidata's Local Data Manager (LDM) software, web services using https, graphical displays.³ The web service API allows each user to specify station and observation types, as well as quality control choices, and domain and time boundaries. The provided MADIS web API and related utility programs allow easy access to MADIS observations without having to develop a program for reading NetCDF files.</p> <p>To access data, users must fill out a data application request. Some datasets are restricted by the provider. There are four distribution categories:⁵</p> <ul style="list-style-type: none"> • Distribution to government, research, and education organizations. • Sponsored access. • Public—full distribution. • Distribution to NOAA only. <p>Restrictions are based on the provider. Most of the datasets are available without restrictions.</p>
Data Sensitivity	No.

Assessment Criteria	Assessment
Data Costs	Free.
Data Openness	Limited data openness due to some restricted content and need for NetCDF format knowledge.
Data Challenges	The NetCDF file format could also be challenging to use for non-scientific staff as it requires the implementation of dedicated API to access the data. NetCDF typically is used in scientific applications such as meteorological forecasting, not in Big Data analysis. NetCDF is not a Big Data–friendly format and its data need to be transformed into a simpler, more Big Data–friendly format to be processed.

¹ Meteorological Assimilation Data Ingest System (MADIS), National Oceanic and Atmospheric Administration (June 16). Online: <https://madis-data.ncep.noaa.gov/> (accessed February 2017).

² Integrated Mesonet Data, National Oceanic and Atmospheric Administration (June 16). Online: https://madis.ncep.noaa.gov/madis_mesonet.shtml (accessed February 2017).

³ MADIS User Resources, National Oceanic and Atmospheric Administration (June 16). Online: https://madis.ncep.noaa.gov/user_resources.shtml (accessed February 2017).

⁴ MADIS Data Volume, National Oceanic and Atmospheric Administration (June 16). Online: https://madis.ncep.noaa.gov/madis_data_volume.shtml (accessed February 2017).

⁵ MADIS Dataset Restrictions, National Oceanic and Atmospheric Administration (June 16). https://madis.ncep.noaa.gov/madis_restrictions.shtml (accessed February 2017).

Table A-26. Third-party web service weather data.

Assessment Criteria	Assessment
Description of Data	Historical meteorological data and weather forecast data from various public and private weather data sources across the globe including temperature, wind, precipitation probability, pressure, visibility, wind speed, wind direction, cloud cover, visibility index, humidity, etc. as well as ancillary data elements such as nearby storms, moon phase, sunrise/set.
Who Collects, Maintains, and Owns the Data	Third-party real-time web service (e.g., Dark Sky).
How the Data Are Collected	Data is obtained from the datasets provided by multiple meteorological agencies from all over the world. Often mostly focused on U.S. and European datasets including MADIS and NEXRAD.
Data Structure	Semi-structured (JSON).
Data Size, Storage, and Management	Petabytes. Managed through Big Data database and cloud file storage. Data is updated. Data is typically updated every minute (Dark Sky).
Data Accessibility	Data is accessed through authenticated representational state transfer (REST)-based web service API. The API is not designed to support file downloads but can handle millions of requests at the same time. The API is used in the following way: A user sends forecast or weather data requests to the API specifying a time and location, and the API returns a very detailed historical reading or forecast for the next hours to days.
Data Sensitivity	No.
Data Costs	Low cost. Pay-as-you go. Low cost per transaction (e.g., \$0.10 per 1,000 requests). First 1,000 forecasts per day free.
Data Openness	Open.
Data Challenges	The primary drawback is that the data cannot be accessed as a whole; rather, existing datasets containing location and time need to be augmented using the API.

Table A-27 National Fire Incident Reporting System (NFIRS) Data

Assessment Criteria	Assessment
Description of Data	The National Fire Incident Reporting System (NFIRS) is the standard national reporting system used by U.S. fire departments to report fires and other incidents to which they respond and to maintain records of these incidents in a uniform manner. NFIRS is the world's largest, national, annual database of fire incident information. ¹ Data elements relevant to TIM include fire department, location, vehicle fire, arrival time, firefighter casualty, firefighter deaths, civilian deaths and injuries.
Who Collects, Maintains, and Owns the Data	Every U.S. state and the District of Columbia report NFIRS data. Although NFIRS participation is not mandatory at the national level, about 23,000 fire departments report in the NFIRS each year.
How the Data Are Collected	After responding to an incident, a fire department completes the appropriate NFIRS modules using NFIRS-compatible software programs. Each module collects a common set of information that describes the nature of the call, the actions firefighters took in response to the call, and the end results, including firefighter and civilian casualties and a property loss estimate. The fire department forwards its data to the state agency responsible for NFIRS data. The agency gathers data from all participating departments in the state and reports the compiled data to the U.S. Fire Administration (USFA). As part of the collection and compilation process, various validation tools are used to ensure the quality of the entered data.
Data Structure	Structured and semi-structured. The public data release (PDR) uses a relational database containing 20 tables. The NFIRS PDR data provided online is composed of 19 data tables (files) (modules) in Dbase database file format (.dbf) format. The same data is available from www.data.gov in flat file formats (TXT, CSV).
Data Size, Storage, and Management	Megabytes to gigabytes. Participating fire departments report about 22,000,000 incidents and 1,000,000 fires each year. The PDR contains more than 2 million incidents per year (gigabytes). Due to large file sizes, the files available in the NFIRS Public Data Release (PDR) consist only of fire and hazardous condition incidents. ²
Data Accessibility	PDR is provided online or on a CD-ROM, as a set of Dbase (.dbf) files or as a set zip file on www.data.gov
Data Sensitivity	No. No sensitive data is loaded in the PDR.
Data Costs	Each year the USFA compiles publicly released incidents collected by states during the previous calendar year into PDR that is made available free of charge. NFIRS software is available as free desktop and web-based applications from the USFA or as NFIRS standard-compliant products purchased from fire software vendors.
Data Openness	Open.

(continued on next page)

Assessment Criteria	Assessment
Data Challenges	<p>The USFA does not have a quality assurance system in place to check for codes that are not in the current data dictionary. Thus, the NFIRS PDR database contains invalid codes and may exhibit data inconsistencies that violate published documentation.³ Data is collected on a voluntary basis, so some areas may not have sufficient data.</p> <p>The distributed dataset is not a complete dataset. It only contains fire and hazardous condition incidents.⁴ The truncation of the dataset is apparently due to current data size limitations in the current storage and distribution system. This is rather uncommon these days, and it denotes either an obsolete system or obsolete data management practices, as the sharing of multi gigabytes files is a common occurrence today.</p> <p>The PDR dataset is published using the Dbase database file format (.dbf), which was created in 1978 to be used with the MS-DOS operating system. It is still common today on desktop-based database software but has had many iterations and variations. It requires software capable of parsing its binary structure to be read, which adds additional preparation work before it can be exploited by typical Big Data tools. JSON, XML, TXT, CSV should be used instead, as many databases capable of generating .dbf files can generate these Big Data–friendly formats as well.</p>

¹ <https://www.usfa.fema.gov/data/nfirs/about/index.html>.

² https://www.usfa.fema.gov/data/statistics/order_download_data.html.

³ https://www.usfa.fema.gov/downloads/pdf/nfirs/nfirs_data_analysis_guidelines_issues.pdf.

⁴ https://www.usfa.fema.gov/data/statistics/order_download_data.html.

Table A-28. National EMS Information System (NEMSIS) data.

Assessment Criteria	Assessment
Description of Data	NEMSIS is a national repository of standardized EMS data elements from 49 states and 2 territories. Data elements relevant to TIM include timestamps associated with the TIM timeline (e.g., notification, dispatch, arrival/departure of EMS responders), EMS agency, type of service requested, type of delay (e.g., dispatch, response, scene), chief complaint, alcohol/drug use, and procedures performed.
Who Collects, Maintains, and Owns the Data	EMS agencies collect the data at the local level. There are three tiers of ownership and maintenance of the data: local, state, and national. The data are collected by local agencies and therefore owned at the local level. States own and maintain their individual state-level NEMSIS databases based on data submitted by the local agencies. States then submit a subset of data to the NEMSIS national repository. The subset of data that is submitted to the national repository is owned by the nation and maintained by the NEMSIS Technical Assistance Center (TAC) at the University of Utah.
How the Data Are Collected	Data is collected by EMS personnel in the field and entered into a NEMSIS-compliant software program, which electronically transmits (via web services) the data to a state database. A subset of data is then electronically transmitted (via web services) from the state databases to the national NEMSIS repository. The data flows from the local level to the national level in just a few minutes.
Data Structure	Structured and semi-structured (XML).
Data Size, Storage, and Management	30.2 million records (gigabytes) were transmitted to NEMSIS in 2015. The national data repository is stored in-house at the NEMSIS TAC.
Data Accessibility	Although the data are not currently publicly available at the local level, they could be. 65 million records are available on the NEMSIS website. A full year of data can be obtained on a DVD from the NEMSIS TAC. NEMSIS TAC can release case-level data to researchers.
Data Sensitivity	Some data elements allow for the identity of the location (state) of the records (e.g., EMS agency, home zip code of patient, destination hospital). These data elements cannot be shared with the public.
Data Costs	The public-release dataset is available for free.
Data Openness	Limited openness because of lack of location resolution in aggregated datasets.
Data Challenges	Location data at the state and national level is limited to the zip code level, which could greatly limit data analytics because the resolution would be too low for meaningful analysis.

Table A-29. Motor Carrier Management Information System (MCMIS).

Assessment Criteria	Assessment
Description of Data	<p>The Motor Carrier Management Information system (MCMIS) is a computerized system whereby the FMCSA maintains a comprehensive record of the safety performance of the commercial motor carriers who are subject to the Federal Motor Carrier Safety Regulations (FMCSR) or Hazardous Materials Regulations (HMR).¹</p> <p>Records are maintained in four broad categories:</p> <ul style="list-style-type: none"> • Registration—Contains FMCSA registration data for all motor carriers (U.S. DOT#, company name, address, contacts, number of vehicles, number of drivers, and other registration information). • Crash—Contains data for each commercial motor vehicle involved in a crash (U.S. DOT#, report number, crash date, severity of the crash (tow-away, injury, fatal) and vehicle data, etc.). • Inspection—Contains data on roadside inspections conducted on motor carriers (U.S. DOT#, report number, inspection date, State, and vehicle and equipment information, and violations-related data, etc.). • Review—Contains information on reviews/investigations conducted on motor carriers and other entities (U.S. DOT#, review date, review type, safety rating, and so forth).
Who Collects, Maintains, and Owns the Data	State DOTs, state law enforcement, FMCSA.
How the Data Are Collected	Manual, electronic.
Data Structure	Structured.
Data Size, Storage, and Management	Terabytes, national data store.
Data Accessibility	Web service, web data files download, and requests via FMCSA data dissemination program.
Data Sensitivity	Contains PII.
Data Costs	<p>Some dataset downloads are free via: https://ai.fmcsa.dot.gov/SMS/Tools/Downloads.aspx.</p> <p>Customized extracts and reports via the data dissemination program incur fees (e.g., crash file extract \$36, personalized crash report \$33, inspection file extract \$70 per calendar year, and company safety profiles \$27.50 each with discounts for more profiles purchased).²</p>
Data Openness	Data is shared as reports; data are not open.
Data Challenges	Data is not available in raw format, only through specific reports.

¹ <https://ask.fmcsa.dot.gov/app/mcmiscatalog/mcmishome>.² https://ask.fmcsa.dot.gov/app/mcmiscatalog/c_chap3#crfe.

Table A-30. HERE data.

Assessment Criteria	Assessment
Description of Data	<p>HERE Technologies aggregates and analyzes road transportation data from more than 80,000 data sources covering more than 180 countries, including “the world’s largest compilation of both commercial and consumer probe data, the world’s largest fixed proprietary sensor network, publicly available event-based data and billions of historical traffic records,” weather, events data as well as panoramic imagery and LIDAR data from its own vehicle fleet.¹ HERE also relies on local source data and input from map users to generate constant daily map updates, such as real-time traffic, turn-by-turn directions, public transportation routes and information about local business and attractions. HERE combines “20 billion real-time GPS probe points a month with historical information and search queries to learn where people are traveling and what the conditions are like,” and per HERE, almost half of all the data is under one minute old and more than three-quarters is under five minutes old.¹</p> <p>Data relevant to TIM includes incident location (road segment), criticality, incident description, real-time traffic condition, three-dimensional (3D) visualization of incident surroundings (including roadway details), start/end times of incidents, construction data, venue data, weather data as well as estimated travel time to incident location and estimated traffic condition created by incident.</p>
Who Collects, Maintains, and Owns the Data	HERE Technologies
How the Data Are Collected	Cell phones, connected navigation systems, fixed proprietary sensors, Twitter, state and local DOT data, email alerts, HERE map application, HERE 3D footprint vehicle fleet, as well as local businesses and attractions.
Data Structure	Structured and semi-structured.
Data Size, Storage, and Management	Terabytes to petabytes. HERE data is stored and processed using various combinations of on-the-premises and cloud-hosted relational databases, NoSQL databases, file storage and compute clusters. Most of the HERE datasets are real-time datasets designed to support real-time decision-making. Some of the HERE datasets are archived indefinitely to support HERE services such as its mapping, visualization, and predictive services.
Data Accessibility	HERE data is accessible through multiple web services, ranging from mapping and visualization services, traffic analysis, traffic prediction, and APIs to mobile application software development kit and toolkits. Web services are accessible to the public and businesses for a monthly fee.
Data Sensitivity	No. Data is anonymized.
Data Costs	HERE data is available through a monthly subscription plan to both the public and businesses. The HERE plan cost varies from free (under 15,000 transactions per month) to \$500/month (150,000 transactions per month). Custom data plans are available for businesses requiring more transactions and services.
Data Openness	Limited openness. Accessed through web services.
Data Challenges	The primary drawback is that HERE data cannot be accessed as a whole (in raw format), but only through HERE web services.

¹ Here 360, How to Really Outsmart Traffic (July 9, 2013). Online: <http://360.here.com/2013/07/09/how-to-really-outsmart-traffic/> (accessed June 2017).

Table A-31. INRIX data.

Assessment Criteria	Assessment
Description of Data	INRIX collects massive amounts of information about roadway speeds and vehicle counts from more than 300 million real-time anonymous mobile phones, connected cars, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices. This data is enriched with event data (e.g., traffic incidents, weather forecasts, special events, school schedules, parking occupancy, road construction) to provide software-as-a-service (SaaS) and data-as-a-service (DaaS) solutions.
Who Collects, Maintains, and Owns the Data	INRIX Traffic.
How the Data Are Collected	Combination of a connected network of anonymized road sensors, devices, cars and drivers from more than 300 million sources, including commercial fleets, delivery and taxis, cameras as well as consumer vehicle data, parking data. This highly granular floating vehicle data is combined with traditional real-time traffic flow information as well as hundreds of market-specific criteria that affect traffic (e.g., construction and road closures, real-time incidents, sporting and entertainment events, weather forecasts, and school schedules).
Data Structure	Big Data infrastructure.
Data Size, Storage, and Management	500 TB of data analyzed daily. ¹ Cloud infrastructure for storage and management.
Data Accessibility	Raw data generally is not available. Access to the data is obtained through a variety of ways, including traffic tiles, a monitoring site, flexible APIs, and the Transport Protocol Experts Group (TPEG) Connect, which delivers traffic and travel information to connected vehicles and mobile devices over the Internet. ² Provides a comprehensive collection of historic speed and travel time data available in archival or profile formats. Available through a series of on-demand, cloud-based analytics suites that leverage INRIX data.
Data Sensitivity	The data is reportedly anonymized, so PII may be low or limited.
Data Cost	Unknown. Must contact INRIX for various pricing structures.
Data Openness	Not open. Proprietary.
Data Challenges	N/A

¹ <http://inrix.com/resources/inrix-traffic-brochure/>.² <http://inrix.com/products/traffic/>.



APPENDIX B

Incident Response and Clearance Ontology (IRCO)

B.1 What Is an Ontology?

Big Data in a vacuum is worthless. Big Data only has value when it is leveraged to drive decisions. Although it may be possible to use implicit or existing relationships within data elements to perform simple Big Data analyses, more complex and insightful Big Data analyses will require a more abstract and concise way to express the knowledge that the data represents—a vision or a structure characterizing what the data represents needs to be established. In computer science, this structure is known as an ontology.

An ontology is designed to establish a commonly shared vision of a domain. An ontology is a formal naming and definition of the types, properties, and inter-relationships of the entities that really or fundamentally exist in a domain. It is a grammar that, when applied to raw data, gives it an explicit meaning. It is a metaphoric pair of polarized glasses, allowing people to clarify raw data and reveal the information it contains universally. Before attempting to integrate any Big Data datasets and derive insight from them, it is essential to define what these data mean and the relationships that describe the specific context. In other words, it is essential to develop an ontology. This appendix describes the approach and steps taken to develop an incident response and clearance ontology (IRCO).

B.2 Development of the IRCO

Ontology development can be a challenging endeavor. There is no correct, prescribed development method; one or more viable alternatives always exist, and the best solution often depends on the application of that ontology. In addition, the process is discovery-based, iterative, and likely ongoing. For this reason, ontologies often are qualified as “incomplete” or “reductive” compared to the domain that they attempt to describe. A simple ontology that is reductive and does not cover every single observed case can still be used to map real data and display how data elements relate to each other. It also can reveal ways in which insight might be extracted or inferred from that data.

The development of the IRCO included the following steps:

- Determine the domain and scope of the ontology.
- Re-use existing ontologies to the extent possible.
- Enumerate important terms in the ontology.
- Define the classes and the class hierarchy.
- Define the properties and facets of each class.
- Create instances to test the ontology.

To assist with several of these steps, a workshop was conducted with first responders. The objectives of the workshop were two-fold: (1) gain insights on the vocabulary, entities, and relationships associated with incident response and clearance for the development of the ontology, and (2) identify opportunities to improve TIM through the application of Big Data. The morning session of the all-day workshop focused on the former objective, and the afternoon session of the workshop focused on the latter objective.

The workshop was conducted in Phoenix, Arizona, at the Arizona Department of Public Safety (AZDPS). Workshop attendees included members of the Aztec TIM Coalition, as well as subject matter experts from across the country. Specifically, the workshop included representation from the following agencies, organizations, and groups:

- Arizona State Troopers.
- AZDPS Dispatch Center Manager.
- Arizona Department of Transportation (Arizona DOT) Safety.
- Arizona DOT Emergency Management.
- Arizona DOT Traffic Records.
- Arizona DOT Data Systems.
- Arizona DOT ALERT.
- Maricopa County DOT REACT.
- Glendale Fire Department.
- Mesa Fire Department.
- Maricopa County DOT.
- Maricopa Association of Governments (MAG).
- Arizona Professional Towing and Recovery Association (APTRA).
- Phoenix Metro Towing.
- California Department of Transportation (Caltrans).
- Minnesota Department of Transportation (Minnesota DOT).
- City of Schertz, Texas, Fire Department.
- Florida Highway Patrol.
- FHWA Arizona Division Office.

The incident timeline was used to engage workshop participants in conversation about incident response and clearance. For each phase of the incident timeline, the group reviewed and discussed the following: Who? Does what? When? Where? How? And with what?

The next section of this appendix discusses the input from the workshop participants within the context of the ontology development steps.

B.2.1 Domain and Scope of the Ontology

During the workshop, the concept of an ontology was introduced to participants using a simple ontology called the *pizza ontology*. The pizza ontology is a well-known ontology in the semantic web community. It was developed for educational purposes by the University of Manchester (University of Manchester 2009). The workshop participants were then walked through the incident timeline and asked various questions to capture the various classes (e.g., vehicle, responder), class relationships, and data entities involved in an incident response.

The workshop approach did not produce all the information necessary for developing the ontology. As the workshop progressed through the incident timeline and the definition of the TIM ontology, it became clear that the IRCO designed following the workshop would be rather high-level and simple, and that further development and testing would be needed involving a large group of TIM professionals to develop a consensus on class names, class characteristics, and class relationships, as many options were mentioned by participants without clear, unanimous perspectives. Ideally, such an effort should be done using an ontology web development environment in which developers and testers can collaboratively design, test, publish, and maintain the IRCO.

Nonetheless, it was established during the workshop that the IRCO should first focus on conceptualizing the response to an incident and how the response relates to the incident itself (i.e., location, time of the day, vehicle, and occupants involved in the incident), as well as the incident environment (i.e., details of the roadway at the incident location, traffic conditions, weather conditions, and social media activities during the response), the personnel, actions, equipment, and response vehicles involved in the response.

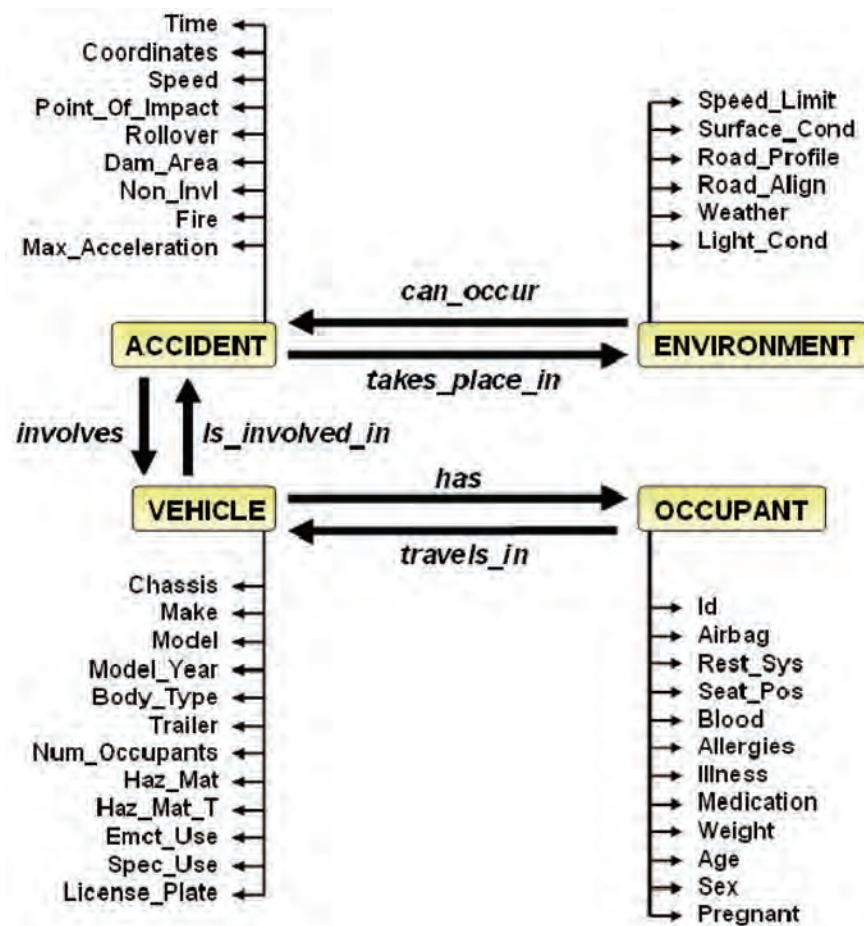
It also was established that the IRCO should be designed to provide answers to questions such as:

- What are the components of an incident response?
- Who is involved in an incident response?
- Where are the responders during an incident response?
- What do responders do during an incident response?
- How are traffic and weather conditions related to an incident response?
- How does responder training relate to an incident response?
- How do social media activities relate to an incident response?

B.2.2 Re-Use of External Ontologies in IRCO

Rather than design the IRCO from scratch, the IRCO was designed using components from existing ontologies. Information gathered during the workshop was combined with existing traffic incident–related ontologies to establish a basis for the IRCO. Several existing ontologies to describe a traffic incident were available, but most were presented as part of research papers rather than published in an ontology file format such as OWL (Ontology Web Language); therefore, several of the pre-existing ontologies could not be incorporated directly into the IRCO.

Of all the ontologies reviewed, a traffic incident ontology developed by Universitat Politècnica de València was chosen as a model for the IRCO. This ontology, the Vehicular Accident Ontology (VEACON), focuses on road safety (Barrachina et al. 2012). Figure B-1 shows a high-level representation of the VEACON ontology and the various data properties of the four main classes (accident, environment, vehicle and occupant).



Source: Barrachina et al. (2012)

Figure B-1. The VEACON ontology.

Although the VEACON ontology provides a good foundation for the description of an incident, it does not include any information about incident response. Therefore, to capture the distributed and spatiotemporal nature of an incident response, including the various tasks performed by responders using various tools, the LODE (Linking Open Descriptions of Events) ontology was used (Shaw 2010). The LODE ontology allows an event to be described in time, in space, and in terms of who was involved during the event. Figure B-2 shows a graphical representation of the LODE ontology. The LODE ontology also re-uses existing ontologies, such as (1) the DOLCE+DnS Untralite, a light-weight ontology for descriptions and situations (Ontology:DOLCE+DnS Untralite 2010); (2) the OWL-time ontology, a web ontology language used for temporal concepts (Cox and Little 2017); and (3) the World Wide Web Consortium (W3C) basic geospatial ontology aimed at describing the entities in space (Brickley n.d.).

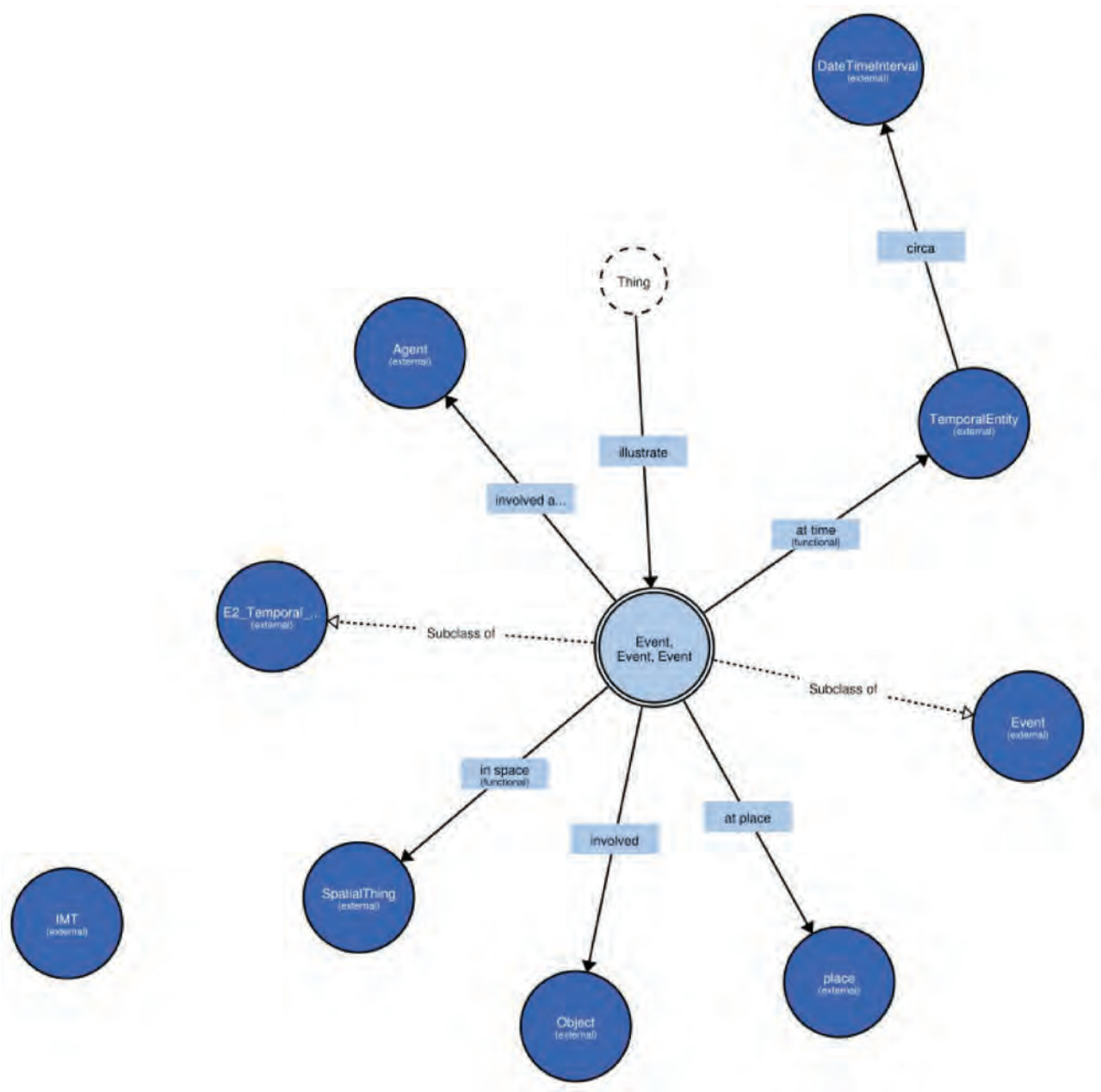


Figure B-2. A visualization of the LODE ontology created using the WebVOWL app.

These external ontologies were then imported into the open-source Protégé ontology development tool (Protégé 2016), and classes, object properties, and data properties were added to create the IRCO ontology. The next section lists each of the IRCO ontology components and their definitions.

B.2.3 IRCO Classes and Class Hierarchy

Table B-1 lists the various classes defined in the IRCO ontology, their super classes, and their definitions.

Table B-1. IRCO ontology classes.

Entity	Type	Superclass(es)	Comment
Agent	Class	lode:agent	lode:agent
Event	Class	Event E2_Temporal_Entity	Event E2_Temporal_Entity
Object	Class	lode:object	lode:object
Person	Class	'Spatial Thing' Agent dterms:Agent	a person
'Spatial Thing'	Class	Thing	geo:wgs84_pos:SpatialThing
TemporalEntity	Class	Thing	owl-time:TemporalEntity
foaf:media	Class	Thing	foaf:media
foaf:organization	Class	Thing	foaf:organization
action	Class	Event	action taken during an incident response
action_location	Class	'Spatial Thing'	action location details (latitude, longitude, fips, zipcode, etc.)
driver	Class	occupant	a vehicle driver involved in an incident
equipment	Class	incident_object	incident response equipment
incident	Class	Event	a traffic incident
incident_environment	Class	Thing	environment of an incident
incident_location	Class	'Spatial Thing'	incident location details (latitude, longitude, fips, zipcode, etc.)
incident_response	Class	Event	a response to a traffic incident
occupant	Class	Person incident_object Person schema:Person	an occupant of a vehicle involved in an incident
passenger	Class	occupant	a vehicle passenger involved in an incident
respondent	Class	Agent Person Person schema:Person	an incident respondent
respondent_organization	Class	foaf:organization	an incident respondent organization
respondent_vehicle	Class	equipment	a respondent vehicle involved in an incident response
roadway	Class	incident_environment	roadway details at the scene of an incident
severity	Class	Thing	the severity of an incident

tool	Class	equipment	a tool involved in an incident response
traffic_conditions	Class	incident_environment Event Event Event	traffic conditions events around the time and location of an incident
vehicle	Class	incident_object	a vehicle involved in an incident
weather_conditions	Class	incident_environment Event Event Event	weather conditions events around the time and location of an incident
tweet	Class	social_media Event Event Event	a tweet from social media website Twitter around the time and location of an incident
incident_time	Class	TemporalEntity	incident time details (start, end, duration, etc.)
media	Class	Thing	media such as image, video or sound
traffic_conditions_location	Class	'Spatial Thing'	traffic conditions details (latitude, longitude, fips, zipcode, etc.)
traffic_conditions_time	Class	TemporalEntity	traffic conditions events time details (start, end, duration, etc.)
response_performance	Class	Thing	the performance of an incident response
injury	Class	Thing	details about an individual's injuries
law	Class	Thing	law pertaining to incident responses
weather_conditions_time	Class	TemporalEntity	weather conditions time details (start, end, duration, etc.)
policies	Class	Thing	policy pertaining to incident responses
tweet_location	Class	'Spatial Thing'	tweet location details (latitude, longitude, fips, zipcode, etc.)
weather_conditions_location	Class	'Spatial Thing'	weather conditions location details (latitude, longitude, fips, zipcode, etc.)
social_media	Class	incident_environment	social media events around the time and location of an incident
standard_operation_procedure	Class	Thing	standard operation procedure pertaining to incident responses
tweet_time	Class	TemporalEntity	tweet time details (start, end, duration, etc.)
tweet_image	Class	foaf:media	an image attached to a tweet
incident_object	Class	Object	a passive entity involved in an incident
action_time	Class	TemporalEntity	action time details (start, end, duration, etc.)
respondent_training	Class	Thing	training received by an incident respondent

B.2.4 IRCO Object Properties

Table B-2 lists the various object properties defined in the IRCO ontology and their definition.

Table B-2. IRCO ontology object properties.

Entity	Type	Comment
involved	ObjectProperty	object or person involved in event
illustrate	ObjectProperty	media illustrate event
'at place'	ObjectProperty	occurred at place or location
'at time'	ObjectProperty	occurred at time or during time interval
'involved agent'	ObjectProperty	involved into event (active participant)
owl:topObjectProperty	ObjectProperty	involved into event (passive participant or object)
member	ObjectProperty	is a member of
hasAction	ObjectProperty	contain an action
hasEnvironment	ObjectProperty	happened during in an environment
hasOccupant	ObjectProperty	vehicle has occupant
hasParent	ObjectProperty	incident has parent incident
hasResponse	ObjectProperty	incident has response
hasSeverity	ObjectProperty	incident has severity
hasInjury	ObjectProperty	person has injury
isDerivedFrom	ObjectProperty	training is derived from
hasPerformance	ObjectProperty	incident response has performance
receivedTraining	ObjectProperty	respondent received a training
OccurredAfter	ObjectProperty	action occurred after another action

B.2.5 IRCO Data Properties

Table B-3 lists the various data properties defined in the IRCO ontology and their definitions.

Table B-3. IRCO ontology data properties.

Entity	Type	Comment
VIN	DataProperty	vehicle VIN (Vehicle Identification Number)
action_type	DataProperty	the type of an action
caused_delay	DataProperty	delay caused by incident and response
Description	DataProperty	description of an event
detection_time	DataProperty	detection time (TIM performance measure)
hazmat_involved	DataProperty	the presence of hazardous material in an incident
heavy_vehicle_involved	DataProperty	the presence of a heavy vehicle in an incident
incident_clearance_time	DataProperty	incident clearance time (TIM performance measure)
incident_identifier	DataProperty	incident identifier such as a call number
incident_response_cost	DataProperty	the cost of an incident response
incident_severity	DataProperty	the severity of an incident (major, minor, property damage only, etc.)
incident_type	DataProperty	the type of an incident (hazmat, injury, non-injury, fatality, etc.)
injury	DataProperty	the presence of injury in an incident
lane_involved_count	DataProperty	the number of lanes involved in the incident response
lane_involved_description	DataProperty	a description of the lanes involved in the incident response
license_plate	DataProperty	a vehicle license plate
make	DataProperty	the make of a vehicle
medical_condition	DataProperty	the medical condition of a vehicle occupant
model	DataProperty	the model of a vehicle
model_year	DataProperty	the model year of a vehicle
number_of_fatality	DataProperty	the number of fatalities in an incident
number_of_injury	DataProperty	the number of injuries in an incident
number_of_minor_injury	DataProperty	the number of minor injuries in an incident
number_of_serious_injury	DataProperty	the number of serious injuries in an incident
number_of_vehicle_involved	DataProperty	the number of vehicles involved in an incident

(continued on next page)

Table B-3. (Continued).

Entity	Type	Comment
property_damage	DataProperty	the presence of property damage in an incident
property_damage_cost	DataProperty	the cost of the property damage of an incident
response_time	DataProperty	response time TIM performance measure
roadway_clearance_time	DataProperty	roadway clearance time (TIM performance measure) the time to return to normal flow time (TIM performance measure)
roadway_direction	DataProperty	the direction of the roadway the incident occurred on
roadway_lighting_conditions	DataProperty	the lighting conditions of the roadway the incident occurred on
roadway_name	DataProperty	the name of the roadway the incident occurred on
roadway_surface_condition	DataProperty	the surface conditions of the roadway the incident occurred on
roadway_surface_temperature	DataProperty	the surface temperature of the roadway the incident occurred on
roadway_type	DataProperty	the type of the roadway the incident occurred on (rural road, highway, etc.)
source_name	DataProperty	the name of the source of the incident and response info
time_to_return_to_normal_flow	DataProperty	the time to return to normal flow time (TIM performance measure)
total_lane_at_scene	DataProperty	the total number of lanes at the scene of the incident
verification_time	DataProperty	Incident verification time (TIM performance measure)
weight	DataProperty	vehicle weight
workzone	DataProperty	the presence of a workzone in an incident
occupancy	DataProperty	traffic occupancy
deceased	DataProperty	if deceased
driver_license_number	DataProperty	driver license number
fatality	DataProperty	the presence of a fatality in an incident
speed	DataProperty	traffic speed
volume	DataProperty	traffic volume

Abbreviations and acronyms used without definitions in TRB publications:

A4A	Airlines for America
AAAE	American Association of Airport Executives
AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
ACI-NA	Airports Council International-North America
ACRP	Airport Cooperative Research Program
ADA	Americans with Disabilities Act
APTA	American Public Transportation Association
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
ATA	American Trucking Associations
CTAA	Community Transportation Association of America
CTBSSP	Commercial Truck and Bus Safety Synthesis Program
DHS	Department of Homeland Security
DOE	Department of Energy
EPA	Environmental Protection Agency
FAA	Federal Aviation Administration
FAST	Fixing America's Surface Transportation Act (2015)
FHWA	Federal Highway Administration
FMCSA	Federal Motor Carrier Safety Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
HMCRP	Hazardous Materials Cooperative Research Program
IEEE	Institute of Electrical and Electronics Engineers
ISTEA	Intermodal Surface Transportation Efficiency Act of 1991
ITE	Institute of Transportation Engineers
MAP-21	Moving Ahead for Progress in the 21st Century Act (2012)
NASA	National Aeronautics and Space Administration
NASAO	National Association of State Aviation Officials
NCFRP	National Cooperative Freight Research Program
NCHRP	National Cooperative Highway Research Program
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
PHMSA	Pipeline and Hazardous Materials Safety Administration
RITA	Research and Innovative Technology Administration
SAE	Society of Automotive Engineers
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (2005)
TCRP	Transit Cooperative Research Program
TDC	Transit Development Corporation
TEA-21	Transportation Equity Act for the 21st Century (1998)
TRB	Transportation Research Board
TSA	Transportation Security Administration
U.S. DOT	United States Department of Transportation

TRANSPORTATION RESEARCH BOARD
500 Fifth Street, NW
Washington, DC 20001

ADDRESS SERVICE REQUESTED

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies
of Sciences, Engineering, and Medicine for
independent, objective advice on issues that
affect people's lives worldwide.

www.national-academies.org

ISBN 978-0-309-48071-0



9 780309 480710

NON-PROFIT ORG.
U.S. POSTAGE
PAID
COLUMBIA, MD
PERMIT NO. 88