Big Data's Implications for Transportation Operations

An Exploration

www.its.dot.gov/index.htm

White Paper – December 19, 2014

FHWA-JPO-14-157

U.S. Department of Transportation

Produced by the John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
Intelligent Transportation Systems Joint Program Office

Notice

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.
The U.S. Government is not endorsing any manufacturers, products, or services cited herein and any trade name that may appear in the work has been included only because it is essential to the contents of the work.

Technical Report Documentation Page

1. Report No. FHWA-JPO-14-157	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Big Data's Implications for Transportation Operations: An Exploration	**5. Report Date** December 2014
	6. Performing Organization Code HW4YA1
7. Author(s) Matthew Burt, Matthew Cuddy, Michael Razo	**8. Performing Organization Report No.** VNTSC-FHWA-14-13
9. Performing Organization Name And Address U.S. Department of Transportation John A. Volpe National Transportation Systems Center 55 Broadway Cambridge, MA 02142-1093	**10. Work Unit No. (TRAIS)**
	11. Contract or Grant No.
12. Sponsoring Agency Name and Address U.S. Department of Transportation Intelligent Transportation Systems Joint Program Office 1200 New Jersey Ave. SE Washington, DC 20590	**13. Type of Report and Period Covered** White Paper
	14. Sponsoring Agency Code

15. Supplementary Notes
Program Manager: Dale Thompson

16. Abstract

The purpose of this white paper is 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. Department of Transportation (USDOT) Connected Vehicle Real-Time Data Capture and Management (DCM) Program. Big data is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that it exceeds the capabilities of traditional data management and analysis tools. This paper is not intended as a primer or "how to" on big data, per se, but rather is intended to explore the potential value of big data approaches in a future connected vehicle environment.

Big data is a process of knowledge generation that features the following approaches:

- Data capture that includes massive datasets encompassing all or most of the population being studied (as opposed to small samples); use of data from both purpose-specific and repurposed data collection; and utilization of crowdsourced and "electronic breadcrumb" data.
- Data management that features storage in decentralized and virtual locations (i.e., the cloud) and handles both structured and unstructured data.
- Data analysis that is often automated, with computers doing more of the work to find complex patterns among a large number of variables.

Big data approaches are needed to contend with the coming volume of connected vehicle and traveler data, to:

- Enable a wide range of new strategies that are expected to provide safety, mobility and environmental benefits, and
- Reduce the need for traditional data collection mechanisms (e.g., connected vehicle probes replacing traffic detectors).

This paper identifies two additional, broad areas where big data analytical approaches may be able to provide further value: 1) Transportation System Monitoring & Management; and 2) Traveler-Centered Transportation Strategies.

17. Key Words Big Data, Intelligent Transportation Systems (ITS), Connected Vehicle, Crowdsourcing, Data capture, Data management, Data analysis, Cloud computing	**18. Distribution Statement**

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 54	**22. Price**

Acknowledgements

The authors wish to thank the project sponsor, Dale Thompson of the Intelligent Transportation Systems Joint Program Office, for the support and leadership that made this white paper possible. The authors also acknowledge the outstanding contributions of other Volpe Center personnel in conducting research, contributing perspective, and assisting with document review and production.

Table of Contents

List of Tables

List of Figures

Executive Summary

The purpose of this white paper is 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. Department of Transportation (USDOT) Connected Vehicle Real-Time Data Capture and Management (DCM) Program. Big data – described briefly below and fully in Chapter 2 – is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that it exceeds the capabilities of traditional data management and analysis tools. This paper is not intended as a primer or "how to" on big data, per se, but rather is intended to explore the potential value of big data approaches in a future connected vehicle environment.

The big data approach can be applied to a wide range of data types and analytical challenges, and there is great potential for transportation stakeholders to introduce big data thinking into their current activities. This paper, however, which is intended to support the DCM Program, focuses on big data as it pertains to connected vehicle data, including data from travelers' mobile devices.

The primary intended audiences for this white paper are key DCM Program stakeholders, including the DCM Program Team, state and local transportation operators, and transportation operations researchers in academia. However, it is also intended to be useful to private sector DCM Program stakeholders, including those who may play a role in generating, capturing, managing or utilizing connected vehicle and connected traveler data.

The primary findings of this paper are summarized below, organized according to key questions that define the chapters in this paper.

What Is Big Data?

Big data is a process of knowledge generation that features the following approaches:

- Data Capture that includes massive datasets encompassing all or most of the population being studied (as opposed to small samples); use of data from both purpose-specific and repurposed data collection; and utilization of crowdsourced and "electronic breadcrumb" data.
- Data Management that features storage in decentralized and virtual locations (i.e., the cloud) and handles both structured and unstructured data.
- Data Analysis that is often automated, with computers doing more of the work to find complex patterns among a large number of variables.

What Could Big Data Mean for Transportation Operations in a Connected Vehicle Environment?

Big data approaches are needed to contend with the coming volume of connected vehicle and traveler data, to:

- Enable a wide range of new strategies that are expected to provide safety, mobility and environmental benefits, and
- Reduce the need for traditional data collection mechanisms (e.g., connected vehicle probes replacing traffic detectors).

This paper identifies two additional, broad areas where big data analytical approaches may be able to provide further value:

- Transportation System Monitoring & Management – This area focuses on the transportation system and utilizing a wide variety of data, including massive volumes of connected vehicle and connected traveler data, to paint a much richer picture of real-time conditions throughout the system and, critically, to enable much better predictions of impending conditions. Examples include using connected vehicle data on braking and lane changing to anticipate imminent traffic flow breakdowns or using vehicle, weather, and traveler behavior data to better predict schedule adherence breakdowns and the resulting bunching of transit buses. Improved prediction of traffic flow, transit schedule and other system breakdowns can lead to more proactive and effective responses on the part of transportation managers, which can improve safety and mobility.
- Traveler-Centered Transportation Strategies – This area focuses on the traveler and includes many privacy issues and questions about public versus private roles that would need to be explored before moving forward. This area could utilize a wide variety of connected traveler data to: 1) Construct detailed profiles describing the behaviors and inferring the preferences and priorities of individual travelers, 2) Utilize these profiles to develop highly personalized traveler information, travel demand management and other strategies – including those featuring non-monetary incentivization schemes, and 3) Implement these strategies by communicating directly with individual travelers in real-time, with highly context and location-specific information, utilizing the same handheld devices that are used to collect the traveler data that builds the profiles. Examples include sophisticated and dynamic ridesharing and transit ridership incentivization. These enhanced, highly-targeted strategies hold the potential to influence travelers' behavior much more effectively and on a larger scale than traditional approaches. They also produce safety, mobility and environmental benefits, such as by spreading demand across time and space to improve utilization of available system capacity.

What Does Big Data Mean Now in Transportation Operations?

Examples of big data approaches or techniques being used in current transportation operations are limited. A number of current transportation projects, such as the Integrated Corridor Management (ICM) systems in San Diego and Dallas and several Michigan Department of Transportation agency fleet connected vehicle projects, feature a number of approaches common to big data. Although very valuable projects, they do not, however, illustrate the full potential for big data, including that which can be realized by applying big data when connected vehicles are widely deployed across the country. Big data examples are somewhat more prevalent in the private sector. Although big data approaches are not yet common, these sorts of examples of big data in transportation suggest that a great deal of potential exists for applying data in the coming traveler data-rich connected vehicle environment.

How Could We Integrate Big Data Practices into Transportation Operations in a Connected Vehicle Environment?

Chapter 5 of this report summarizes and comments upon the recommendations from several recent USDOT and other studies regarding how big data approaches may be applied in transportation operations. We agree with most of those recommendations, including:

- The importance of resolving data ownership issues and the implications for roles.
- Investigating potential use of a third-party data broker (or multiple brokers) – an approach that may help address ownership and funding needs (the cost of capturing and managing data may be cost-prohibitive for government but profitable for the private sector).
- Development of data standards is an important step, especially so if transportation agencies are not collecting and managing the data themselves.
- Considering 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.
- Utilization of 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.

Next Steps

Recommended next steps for the DCM Program to promote consideration of big data in the connected vehicle environment – on the part of transportation operating agencies and within other (i.e., non-DCM) activities within the USDOT connected vehicle program – are as follows:

- Engage Stakeholders – Engage with a broad range of stakeholders that will include much more than the traffic and transit management communities to corroborate and disseminate the conclusions of this paper, particularly the value proposition for applying big data in transportation operations (see Chapter 3). Examples of the additional stakeholders that should be engaged include public and private, transportation and non-transportation, data analytic product and service providers, modelers, algorithm developers, and decision support system developers.
- Coordinate with Existing Activities – Ensure big data is appropriately considered in further DCM research, such as on potential approaches to manage the volume of connected vehicle data that will be captured, and in other USDOT connected vehicle activities, including the continued development and testing of connected vehicle applications and connected vehicle data policy research.
- Develop a Framework Defining Connected Vehicle Data Environment Roles – This framework would identify and evaluate options pertaining to the potential roles and

responsibilities for state, local and federal government and the private sector. In addition to incorporating input from various stakeholders, this effort would leverage USDOT connected vehicle policy work and the Connected Vehicle Reference Implementation Architecture (CVRIA).

- Develop Connected Vehicle Big Data Use Cases – This would entail identifying connected vehicle applications or use cases that incorporate big data analytical approaches (and the operational strategies that could derive from the knowledge gained through those approaches). Subsequent activities in this area could include development and demonstration of a selected, high-potential group of use cases or applications and drafting an institutional operational model for data capture and management.
- Investigate Big Data Resource Implications – Further investigate the potential cost and other resource implications of adopting big data approaches, based on the outcome of the use case investigation noted above. Costs should be addressed within the broader USDOT investigation of connected vehicle roles and responsibilities.

Chapter 1 Introduction

The purpose of this white paper is 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. Department of Transportation (USDOT) Connected Vehicle (CV) Real-Time Data Capture and Management (DCM) Program. This paper is not intended as a primer or “how to” on big data, per se, but rather is intended to explore how big data approaches may be applied in a future connected vehicle environment. As such, although the paper includes a number of references to specific tools and techniques to help characterize what big data is, it does not aim to provide a comprehensive list of technologies or practices or detailed instructions for applying them.

The primary intended audiences for this white paper are key DCM Program Stakeholders, including the DCM Program Team, state and local transportation operators, and transportation operations researchers in academia. However, it is also intended to be useful to private sector DCM Program stakeholders, including those who may play a role in generating, capturing, managing or utilizing connected vehicle and connected traveler data.

Big data is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that it exceeds the capabilities of traditional data management and analysis tools. Big data approaches can be applied in a wide range of transportation data uses. This paper, however, which is intended to support the DCM program, focuses on big data as it pertains to connected vehicle data, including data from travelers, e.g., handheld or “nomadic” devices. While this paper focuses on applying big data approaches to connected vehicle and connected traveler data, the broader discussion of what constitutes big data may also help illuminate how big data may be applied to other data types that are currently available.

White Paper Objectives

- Provide a clear and resonant definition of big data to support meaningful dialogue
- Identify the specific potential value of big data for transportation operations
- Summarize the current practice of big data in transportation operations
- Summarize and react to the current discourse on big data implications for the DCM Program
- Provide clear conclusions and recommendations for next steps

This investigation of big data is important for several reasons, including the fact that big data approaches will be helpful in taking advantage of the large volumes of connected vehicle and traveler data that are expected to become increasingly available over the next 20 years. With connected vehicle Basic Safety Messages generated every one-tenth of a second, a massive volume of data will be produced. Another reason for investigating big data is that additional benefits beyond simply coping with the volume of connected vehicle data are possible. This paper identifies two areas where we believe that big data may represent a paradigm shift and enable transformative benefits:

1. Transportation system monitoring and management featuring vastly improved situational awareness and prediction of problems before they manifest; and

2. A wide range of essentially new travel demand management, traveler information and other strategies that approach the individual traveler as a client and which leverage rich profiles of individual travelers' behavior and preferences to deliver highly personalized and effective recommendations and incentives that cumulatively enhance system performance.

Finally, this exploration of big data potential for connected vehicle and connected traveler data is intended to promote appropriate linkages and corroboration with other USDOT research, including the further development of connected vehicle applications and decision support system and Active Transportation Demand Management (ATDM) research.

The primary thrust of this paper is to explore whether and how big data approaches may provide value in the coming connected vehicle/traveler environment and identify some implications for adopting those approaches. This paper does not delve into the resource implications (infrastructure, workforce) associated with big data. However, the conception of big data and the potential value asserted here do provide some basis to draw some preliminary observations and those are included in Chapter 6 (Conclusions).

1.1 Framing Assumptions

The following assumptions have been used to provide focus for this white paper:

- **Traffic- and Transit-Centric Examples** – In order to keep the length of this paper manageable and to serve the objective of providing examples rather than an exhaustive vision, some of the discussion in this white paper focuses primarily on large, urban area traffic and transit operations. However, the overall DCM Program involves additional modes and environments, e.g., rural/intercity, international border crossings, freight, and emergency management.
- **Beyond Capture and Management** – As an activity of the Data Capture and Management Program, a focus on big data approaches related to capture and management is appropriate. However, because many aspects and benefits of big data pertain to analysis, this paper also considers analysis.
- **Focus on Transportation Operations** – This white paper focuses on transportation operations. However, there are a number of potentially compelling applications of connected vehicle and connected traveler big data approaches that are relevant to other, predominantly non-real-time, areas of transportation practice, including asset management and planning. References to these other areas may appear when they serve the purpose of communicating the value proposition of big data approaches.
- **Recognizing Privacy Challenges** – This white paper explores some transportation concepts that would require some system operator to collect and manage private data. For example, it could be valuable to create a transportation system user profile that includes a history of trips stored by day, time and mode; location of home and work; and locations of other frequent stops such as coffee shops or dry cleaners. We are not proposing that the Federal government have a role in managing such personal profiles, nor are we advocating that any business should do so. Rather, we are extrapolating from current customer profiling practices to sketch out a likely class of uses for transportation data.
- **Recognizing Data Availability Uncertainty** – There are a number of unanswered questions about exactly what data will be generated within a connected vehicle environment, who will have access and use rights to various data, and where and how agencies may choose to

collect and use connected vehicle generated data. This paper includes references to the broader connected vehicle community's efforts that are underway to address these questions, including efforts sponsored by the USDOT Connected Vehicle Policy Program, and includes recommendations to ensure that big data considerations are represented in those other deliberations. However, resolution of those questions is beyond the scope of this paper and to facilitate the exploration of the potential application of big data approaches, it is assumed in this paper that essentially all connected vehicle data is available to a big data practitioner.

1.2 Objectives of Remaining Chapters

The five chapters that constitute the remainder of this white paper, and the primary objectives of each, are as follows:

- **Chapter 2 – What is Big Data?** – The objective of this chapter is to describe big data in a clear manner that will resonate with transportation operations stakeholders, thereby providing a basis for a meaningful exploration and dialogue on the potential value of big data. A description is provided by identifying characteristics of big data and contrasting them with traditional approaches.
- **Chapter 3 – What Could Big Data Mean for Transportation Operations in a Connected Vehicle Environment?** – The objective of this chapter is to identify the potential value that could be derived from big data in surface transportation operations. It is asserted that big data approaches – particularly the technologies and techniques that support data capture and management – are needed to contend with the coming volume of connected vehicle and traveler data and that big data analytical approaches can provide significant further benefits.
- **Chapter 4 – What Does Big Data Mean Now in Transportation Operations?** – The objective of this chapter is to broadly characterize the extent to which big data approaches are being utilized in transportation operations applications by private businesses and public agencies. This is done through the examination of illustrative examples of businesses and leading-edge, agency-led transportation system management and connected vehicle and probe data projects.
- **Chapter 5 – How Could We Integrate Big Data Practices into Transportation Operations in a Connected Vehicle Environment?** – The objective of this chapter is to summarize and react to the major conclusions and recommendations of recent big data-related U.S. DOT and other connected vehicle community research and discourse. This chapter acknowledges that prior research has occurred and by incorporating the findings sets the stage to move forward in a coordinated fashion.
- **Chapter 6 – Conclusions and Next Steps** – The objective of this chapter is to summarize and distill the findings of the previous chapters into a limited number of clear takeaways for the reader and to recommend several next steps.

Chapter 2 What Is Big Data?

Chapter Objectives
- Describe big data - Establish a shared understanding of the concept to facilitate further dialogue

This chapter describes big data, focusing on the ways it differs from traditional data capture, management and analysis. The explanation here is intentionally application-neutral, fitting finance as well as it does transportation, to facilitate deeper consideration of how transportation operations could benefit from techniques being applied elsewhere.

A number of key technologies and techniques are referenced in this chapter and elsewhere in this paper, primarily to help explain what big data means. However, because this paper is not intended as a primer or “how to” on big data, per se, it does not provide a comprehensive list of the technologies or practices associated with big data or detailed instructions for applying them. Those readers interested in more information on various technologies and techniques related to big data are referred to Appendix A.

Big data is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that it exceeds the capabilities of traditional data management and analysis tools. Big data has often been discussed in terms of the 3 Vs: unprecedented volumes of data, with substantial variety in the types of data available, collected and analyzed at high velocity – in real time or near real-time. Some organizations have defined a fourth V, veracity, and even a fifth V, value. Regardless of whether one favors a three-, four- or even five-V definition of big data, the Vs alone are shorthand and leave a lot of questions unanswered. The following sections describe big data by contrasting it with traditional approaches for data capture, management and analysis. Table 2-1 provides a summary. Although the table lays out stark differences between traditional and big data approaches, in practice they lie on a continuum, with the two columns in the table staking out the endpoints. Many cases of analysis will land somewhere in the middle, looking like big data in some aspects and traditional analysis in others.

2.1.1 Capture

In data capture, there are three defining differences between big data approaches and traditional approaches.

First and most important, big data often involves measuring the behavior of nearly an entire population or system, whereas traditional analysis relies on statistical samples. While traditional analysis is often designed around the conditions that allow valid statistical inference about the characteristics of a population based on measurements on a small sample, big data-style analysis is built around the possibility of learning about systems by observing them in their entirety.

Table 2-1. Defining Characteristics of Traditional Versus Big Data Approaches

	Traditional Approaches	Big Data Approaches
Capture		
Statistical Sampling	**Yes** Small fractions of the populations are sampled	**No** Datasets encompass nearly all of a population
Experimental Design	**Critical** Guided by theories of causation and need for statistical validity	**Less Important** Data collected for other purposes often analyzed to address new questions (repurposed data)
Number of sources	**Limited** Data come from dedicated sampling/collection	**Very large** Crowdsourced data or “electronic breadcrumbs” – incidental, automatically or system-generated electronic records – often feed analyses
Management		
Storage	**Single Physical Location**	**Multiple Locations** Shared or virtual access (cloud)
Structure	**Structured** Data resides in fixed fields	**Structured and Unstructured** Some data not in fixed fields, e.g., video and text streams in addition to structured data
Analysis		
Analytical Approach	**Statistics** Traditional regression methods	**Data science** Pattern recognition and machine learning in addition to traditional statistics
Number of Variables	**Limited**	**Very large**
Processing	**Manual** People use specific tools and intuition	**Automated** Optimization routines create best-fit models

A second and related difference between traditional and big data approaches to data capture is that traditional approaches, because they feature collection of fractional samples, usually require much more careful planning to ensure that data are captured at the right place and time under the right conditions. Because a small number of observations must be leveraged to provide information about the entire population, those observations carry a lot of weight and must be chosen carefully. Often, preconceived notions of causation guide data collection so that what are anticipated to be the key variables are measured. With big data, knowledge is produced by observing much larger datasets, including the entire system/population; the question of which data to collect is eliminated (or at least deemphasized). An important byproduct of this fact is that data collected for one purpose can often be reused to answer other questions.

A third distinguishing feature of data capture for big data is that the data are often coming from multiple sources and are diverse. These data may include actively or passively crowdsourced data (that is, data collected by soliciting contributions from a large group of people, often via networked media such as the Internet) and/or electronic records of other activities incidental to the question at hand, also known as “electronic breadcrumbs.” This “multiple source” characteristic of big data is linked to the need for copious data about a population: both crowdsourcing and the “electronic breadcrumbs" are often less expensive than purpose-driven targeted data collection. It is also related to difference #2 above, that data can be repurposed more easily when working with big data.

2.1.2 Management

Data management for big data is distinctive principally in that it is usually decentralized as a matter of necessity. Big data capture often draws on decentralized data sources, as through crowdsourcing and/or "electronic breadcrumbs"; collection and consolidation of this data often requires distributed resources. Because of the size of the datasets involved, data handling and analysis is often impractical if undertaken only by dedicated servers, making cloud computing preferable. Cloud computing refers to networks of physically separate devices, connected via the Internet, that provide data sharing and management capabilities that would previously have been housed in a single computer. Finally, the diversity of the data that may be analyzed in big data approaches, including traditional structured databases and unstructured data such as video and text, lends itself to a diversity of data storage sites that need to be connected and collated to allow the analysis to proceed. In the context of transportation operations where various data in a region are collected and physically held by different jurisdictional and modal agencies or departments, big data necessarily implies data sharing among agencies. Big data techniques like cloud computing and federated data systems – an approach to providing data access by redirecting data requests to partner sites where the data actually resides – help address this challenge.

2.1.3 Analysis

When dealing with big data, datasets of interest often contain huge numbers of variables and vast numbers of data points. Traditional statistical methods are often insufficient and unsuitable for reliably gleaning information from such diverse datasets, and traditional data handling techniques, such as SQL-based databases, are often not up to the task of handling the volume. Instead, it is necessary to use a family of approaches referred to as "data science," which encompasses a number of fields including advanced statistics, signal processing, machine learning, and pattern recognition (Dhar 2013). To handle the volume of data, data science must be coupled with data handling techniques suited for large datasets such as distributed computing.

2.2 Big Data Examples

Big data analytics are in wide use across various industries.

Insurance Fraud Detection: General Electric (GE) receives fraudulent claims against its warranties on home appliances. The company's traditional method of detecting fraudulent claims is to compute 26 metrics for each claim (e.g., time to end of the warranty) and refer the claim to an auditor when multiple metrics falls outside acceptable ranges. One limitation of this claim-by-claim approach is that it is impossible to discern patterns emerging in the incoming claims. After instituting a big data approach looking across all aspects of all claims, GE estimated that it saved $5.1 million in the first year (Verma & Mani, 2014).

Big data characteristics:

- Capture: No sampling – considers all data
- Analysis: Data science – pattern recognition
- Analysis: Automated model development

Predicting Mechanical Failure: United Parcel Service (UPS) uses big data analytics to reduce its maintenance costs. Because on-the-road vehicle breakdowns tend to be expensive and disruptive to its operations, UPS used to replace certain parts on its trucks every two to three years. However, this led to the replacement of perfectly good parts; the simplistic maintenance plan was wasting money. Starting in the early 2000s, UPS began to use predictive analytics to identify those parts that were in fact nearing failure and in need of replacement. Equipping the vehicle undercarriage with an array of sensors, UPS identifies patterns in the sensor readings that corresponded with part failure. Armed with a fleet of sensor-equipped vehicles and knowledge of the patterns that presage failure, UPS is now able to predict part failures and replace parts only as needed (Mayer-Schönberger & Cukier 2013).

Big data characteristics:

- Capture: No sampling – considers all data
- Capture: Experimental design is a low priority – instrumented much of the undercarriage with inexpensive sensors rather than trying to guess where to put them
- Analysis: Data science – pattern recognition
- Analysis: Automated model development

Market Segmentation: Netflix estimates that 75 percent of its viewers' activity is guided by recommendations that it provides. Its recommendation engine is an algorithm fed by and based on a stream of disparate data. An essential part of that data is a video code catalog, wherein different movies and television shows are classified according to a wide range of categories, over 76,000 as of late 2013 (Madrigal 2014). User behavior is critical to the matching algorithm: the company tracks what a user watched, searched for, and rated, as well as browsing and scrolling behavior (Vanderbilt 2013).

Big data characteristics:

- Capture: No sampling – considers all data
- Capture: Number of sources is very large
- Capture: Experimental design is a low priority – analyzed all available behavior data rather than making assumptions about which data is most important
- Management: Virtual location
- Analysis: Data science – pattern recognition
- Analysis: Automated model development

2.3 Summary

Big data describes a new process of knowledge generation enabled by large, diverse datasets. Generally speaking, big data analysis relies on comprehensive and diverse data that more or less holistically describes a population or system, requires decentralized data management, and leads to the recognition of complex patterns. It has been used in other industries to detect rare events, predict system breakdown, and learn personal preferences, among other uses.

Chapter 3 What Could Big Data Mean for Transportation Operations in a Connected Vehicle Environment?

Chapter Objectives
• Identify the potential value of big data for transportation operations • Identify value in facilitating proposed connected vehicle applications • Identify value in enabling potentially transformative, new approaches • Illustrate how big data could be applied in a future, hypothetical connected vehicle environment

This chapter proposes some of the ways that big data approaches could provide value in transportation operations when connected vehicle and connected traveler data are available. It suggests that there are two primary areas where big data thinking and specific techniques could provide value and also presents an example showing big data applied within a future connected vehicle environment. The first area of proposed value pertains to enabling the sorts of connected vehicle applications that have been the subject of considerable planning on the part of USDOT, the American Association of State Highway Transportation Officials (AASHTO) and many other stakeholders. This first area primarily leverages data capture and management approaches that have been utilized successfully in big data activities, specifically, those approaches that will allow transportation operators to contend with the sheer volume of anticipated connected vehicle data. The second area focuses on the analytical value of big data approaches and looks beyond connected *vehicle* data to focus more heavily on the coming world of the connected *traveler*. The final section overlays big data on a regional information flow diagram for a hypothetical future connected vehicle environment.

3.1 Applying Big Data to Enable Connected Vehicle Applications

The first area in which big data-inspired approaches can provide benefit is in enabling the many connected vehicle applications that are anticipated. This section asserts that there are many anticipated connected vehicle applications expected to generate both societal benefits (e.g., reduced traffic fatalities) and direct benefits to operating agencies; and that the data generated by connected vehicles – including Basic Safety Messages generated every one-tenth of a second – and required for various applications will be sufficiently massive to require new, big data-inspired approaches to capture and management.

3.1.1 Connected Vehicle Applications and their Benefits

Considerable effort has been expended by a broad range of stakeholders over the last decade or so identifying a large number of anticipated connected vehicle applications, including the data that will support them, and some of their benefits. Several key reports that discuss connected vehicle applications and their corresponding data include:

- AASHTO/USDOT Intelligent Transportation Systems-Joint Program Office "National Connected Vehicle Field Infrastructure Footprint Analysis" study, which has generated findings pertaining to dozens of applications and their applicability to nine Deployment Concepts (e.g., Rural Roadway, Urban Intersection) and six Deployment Scenarios (e.g., Urban, Multi-State Corridors).
- USDOT's "Connected Vehicle Reference Implementation Architecture" (CVRIA) website (Iteris 2014) lists over 50 proposed or potential connected vehicle applications across four categories (Environment, Mobility, Safety and Support). USDOT and their partners have developed Concepts of Operation (ConOps) for many of the applications. Sketch-level descriptions of over 40 applications are available on the USDOT website, and a list of these applications can be found in Appendix B.
- The USDOT has estimated that connected vehicle safety applications could potentially address about 4,503,000 police-reported (PR) or 81 percent of all-vehicle target crashes; 4,417,000 PR or 83 percent of all light-vehicle target crashes; and 272,000 PR or 72 percent of all heavy-truck target crashes annually (USDOT 2010).
- The USDOT report, "Benefits of Dynamic Mobility Applications, Preliminary Estimates from the Literature" estimated that full deployment of the set of mobility applications may be capable of eliminating more than one third of the travel time delay caused by congestion (USDOT 2012).
- National Cooperative Highway Research Program (NCHRP) Project 03-101, "Costs and Benefits of Public Sector Deployment of Vehicle-to-Infrastructure Technologies," identifies the following benefits of connected vehicle deployment for deploying agencies:

  - Crash response and cleanup cost avoidance due to reduced accidents from connected vehicle technology (references NHTSA/Volpe crash reduction estimates)
  - Work-zone accident reduction
  - Lower cost of pavement condition detection, i.e., fewer miles that agency personnel have to drive in the field to manually assess conditions
  - Reduced expenditures related to traffic monitoring and traveler information systems (e.g., authors assumed connected vehicle traffic probes would allow elimination of 20 percent of Michigan Test Bed traffic-detection equipment)
  - Reduced winter maintenance costs, e.g., better road weather data via connected vehicles can reduce the amount of plowing and salting done
  - Adaptive lighting, e.g., dimming the lights significantly when no vehicles are present, as derived from connected vehicle data
  - Reduction in traveler information costs, e.g., fewer Dynamic Message Sign (DMS) and 511 (phone and web) costs because agencies will be able to push information directly to connected vehicles (NCHRP 2014).

3.1.2 Connected Vehicle Data Volume

There are not yet robust projections of the volume of connected vehicle data that will be generated, although USDOT expects to further address this question in 2015 as additional functionality is introduced to the Southeast Michigan Connected Vehicle Test Bed. In the absence of accurate projections, two very broad and likely premature characterizations regarding the potential challenges associated with the volume of connected vehicle and traveler data are illustrative:

- In preliminary analysis, USDOT has estimated, under at least one possible scenario (which may not necessarily characterize a typical connected vehicle deployment), a data stream rate of between 10 and 27 petabytes per second of connected vehicle Basic Safety Message (BSM) data is possible, assuming full connected vehicle deployment (USDOT 2013). A petabyte is equal to 1,000 terabytes – enough to store the DNA of the entire population of the U.S. – and then clone them, twice; or more than four times the contents of the Library of Congress; or the storage capacity of 223,000 DVDs (Computer Weekly 2013).

- In an exercise cited in a 2011 USDOT real-time data capture and management state-of-the-practice scan, a case study was reported that projected that about 2 terabytes of data could be generated per day for pavement monitoring in an area the size of Washington, D.C. (USDOT 2011). This case study, which investigated a real potential pavement monitoring application on a test track, included parameters that may not be consistent with a pavement condition monitoring application that may emerge nationally.

3.2 Leveraging Big Data Analytics

This section proposes that there is a second, potentially vast area of big data benefit for transportation operations. This is above and beyond the value in applying data capture and management techniques and technologies like cloud computing and data federation (an approach to providing access to data by redirecting data requests to other, partner sites where the data actually resides) to help transportation personnel to manage the sheer volume of connected vehicle data.

The application of big data analytics to connected vehicle and connected traveler data could improve transportation system monitoring, which in turn could enhance the effectiveness of management strategies. This first area focuses primarily on the supply side of the transportation supply-demand equation and highlights a broad theme in big data: the ability to act more proactively—to better predict and mitigate adverse conditions before or as they materialize.

Connected vehicle and connected traveler data, coupled with big data analytical approaches, could enable a range of traveler information, travel demand management and other strategies to much better understand and effectively influence traveler behavior. This second area focuses more on the demand side of the transportation supply-demand equation and highlights another broad big data theme: dramatically improved ability to shape behavior based on much more comprehensive understanding of individual users/travelers.

3.2.1 Transportation System Monitoring and Management

Big Data Monitoring and Management Examples
• Predicting imminent traffic flow breakdown based on vehicle braking and lane changing • Predicting transit schedule adherence breakdown based on driver, vehicle, weather, traffic and traveler variables

Large-scale implementation of connected vehicles and connected travelers (i.e., capturing a wide range of data from mobile devices like smartphones) promises to provide a veritable ocean of new data for transportation operators. This possible future ocean of data includes a wide variety of vehicle and traveler probe data for essentially every important location throughout the entire multimodal transportation system, including freeways, arterial streets, buses, trains, sidewalks, parking lots, border crossings, etc. This will represent a dramatic, potentially paradigm-changing increase in data. Coverage will expand, such that essentially every important location and time period is covered – in itself an improvement over existing conditions that challenges our ability to fully imagine what could be leveraged. Maybe more revolutionary, there will also be a dramatic expansion of the kinds of data that will be available, such as real-time data on braking, acceleration and lane-changing behavior of vehicles on a given stretch of freeway.

New streams of real-time data from non-transportation sources may further enrich the picture, and point to new precursors and determinants of congestion and delay. What if currently available smart-home technology, which allows a house's energy management system to know which rooms are occupied (such as the bedroom versus the bathroom versus the kitchen, as a commuter heads out to work), could provide new ways to predict travel demand by time, mode and route?

By applying big data analytical approaches, this new world of real-time data (dramatically improved both in coverage and content) can potentially enable new or enhanced methods for predicting breakdowns in the transportation system. Consider incident detection, for example. Currently, the prevailing, real-world state of the practice is to infer the presence of traffic flow breakdowns based on data from fixed traffic detectors – essentially, to identify a problem once it's a problem. How might incident detection be enhanced by applying big data analytical approaches to the ocean of vehicle and traveler data? What if a big data-like exploration of all of the possible correlations between a wide range of conditions were to indicate that certain types of flow breakdowns could be accurately predicted based on the combination of 50 factors? Those factors might include "new" variables like the aforementioned vehicle lane-changing and braking data or even sun angle or day of the week or other variables that we would not necessarily expect to be correlated. How much could safety, mobility and environmental impact be enhanced if transportation operators could take action before traffic flow broke down?

Consider the example of transit operations and the perennial challenges of schedule adherence and bunching, where one or more vehicles fails to adhere to its schedule and maintain a headway. What if a big data analysis of massive volumes of connected vehicle and traveler data identified the constellation of conditions (bus driver behavior, transit vehicle performance, percentage of fares paid by cash versus transit pass, etc.) that precedes schedule adherence breakdowns?

In either of these cases (traffic flow or transit schedule adherence breakdown) what would the understanding of correlates or precipitates ultimately lead to in terms of prevention and response strategies? If we know that "scenario X" often precedes a breakdown, how much better might we intervene in a real-time or pre-planned fashion, in any of countless ways, to prevent that scenario from developing or to respond so much faster and effectively to it once it does develop? The point is that identifying patterns could represent only the beginning in a process that leads to much deeper understanding, including causation that could enable a far greater range of new or enhanced management strategies.

3.2.2 Traveler-Centered Transportation Management

In addition to enabling dramatically richer and more effective transportation system monitoring and management, smartphones coupled with big data techniques will likely allow a fundamentally different relationship between Transportation Management Centers (TMCs) (and other public and/or private entities) and transportation system users. The change in this relationship could come in two forms: 1) personalized transportation services, and 2) targeted incentives to influence users' transportation behavior.

Big Data Traveler-Centric Examples

- Transit ridership incentives based on rider locations and profiles
- Rideshare matching based on highly-personalized individual profiles built from observed behavior

First, targeted transportation services will likely become more prevalent. The industry of personal mobility providers is currently represented by on-demand taxi services such as Hailo, peer-to-peer ridesharing services such as Lyft, and corporate carsharing services such as Zipcar. Service providers of this type are likely to grow in market share as these trends progress:

- the market share of smartphones increases,
- consumer acceptance of transportation as a service grows (by virtue of the growing proportion of drivers who have grown up with mobile technology as well as natural age-independent adoption of the new practices), and
- policies mature around issues of insurance, industry regulation and taxation, and parking.

Second, system users' transportation decisions and preferences will become better understood and therefore prone to influence. One of the main application areas of big data is in market segmentation and targeted marketing. Amazon.com is among the early innovators in this area. When the company started, it employed a team of book reviewers to read books and develop recommendations so that the company could suggest what other books a customer might like to read, given that s/he had enjoyed a given book. At some point, the company tried an alternative approach based on big data: mathematically analyzing buying patterns of all of its customers to create a predictive recommendations engine. Quickly, the mathematical model vastly outperformed the panel of experts, and Amazon chose to use the model as its main source of recommendations (Mayer-Schönberger & Cukier 2013).

If and when smartphone-based applications are able to keep track of when, where and how people use the transportation system and link that to other factors, these software applications should be able to predict future travel choices, and potentially influence them. Consider these examples:

- Currently, in Montreal, the local transit agency partners with local retailers to increase ridership. By pairing real-time location of bus riders with a profile of their commercial activities, the agency is able to provide instant e-coupons as rewards for riding the bus. Retailers boost sales while the agency encourages transit ridership (Winterford 2013). The agency is targeting a 40 percent bump (Murphy 2014).

- In a hypothetical example, imagine that there was an app that could help a driver identify his or her perfect match for carpooling. The app would know the driver's home and work locations, typical departure times by day, and needs for childcare drop-offs and pick-ups. It would know the same information about carpooling candidates. To facilitate a personality match, the app would use an Amazon- or Match.com-style recommendation engine. For both people to be joined in a carpool, the app would also have a personal profile: age, number of children, recent books purchased, recent movies watched, political perspective, etc.

When transportation system users can be understood and treated more as customers, a world of possibilities for engaging them and influencing their behavior opens up. When we recognize the relationships between their transportation choices and their non-transportation activities, it becomes possible to, among other things, encourage them to make choices that improve overall system efficiency. This is exactly the case in Montreal, where the transit agency uses e-coupons to increase ridership and thereby reduce the number of drivers on the road. When we combine the possibility of influencing users' transportation choices with the likelihood of expanded options (following Lyft, Uber and Zipcar), we begin to see the potential for big data to create radical changes in transportation.

3.3 A Vision for the Application of Big Data Analytics in Transportation Operations

This section illustrates how and where we may expect big data analytics to have their greatest impacts on transportation operations. It relies on a description of a real-time multi-modal decision support system (RTMDSS) taken from a 2011 ITS JPO-produced concept of operations (SAIC & Delcan 2011). Figure 3-1, taken from this report, depicts data exchanges envisioned between various entities and the RTMDSS. The RTMDSS accepts data from and develops commands for roadway data collection systems, arterial management systems (AMS), freeway management systems (FMS), transit management systems, emergency management systems, connected vehicles, and third-party services that interface directly with the public, as indicated in the "Personal Info Access" box in the diagram.

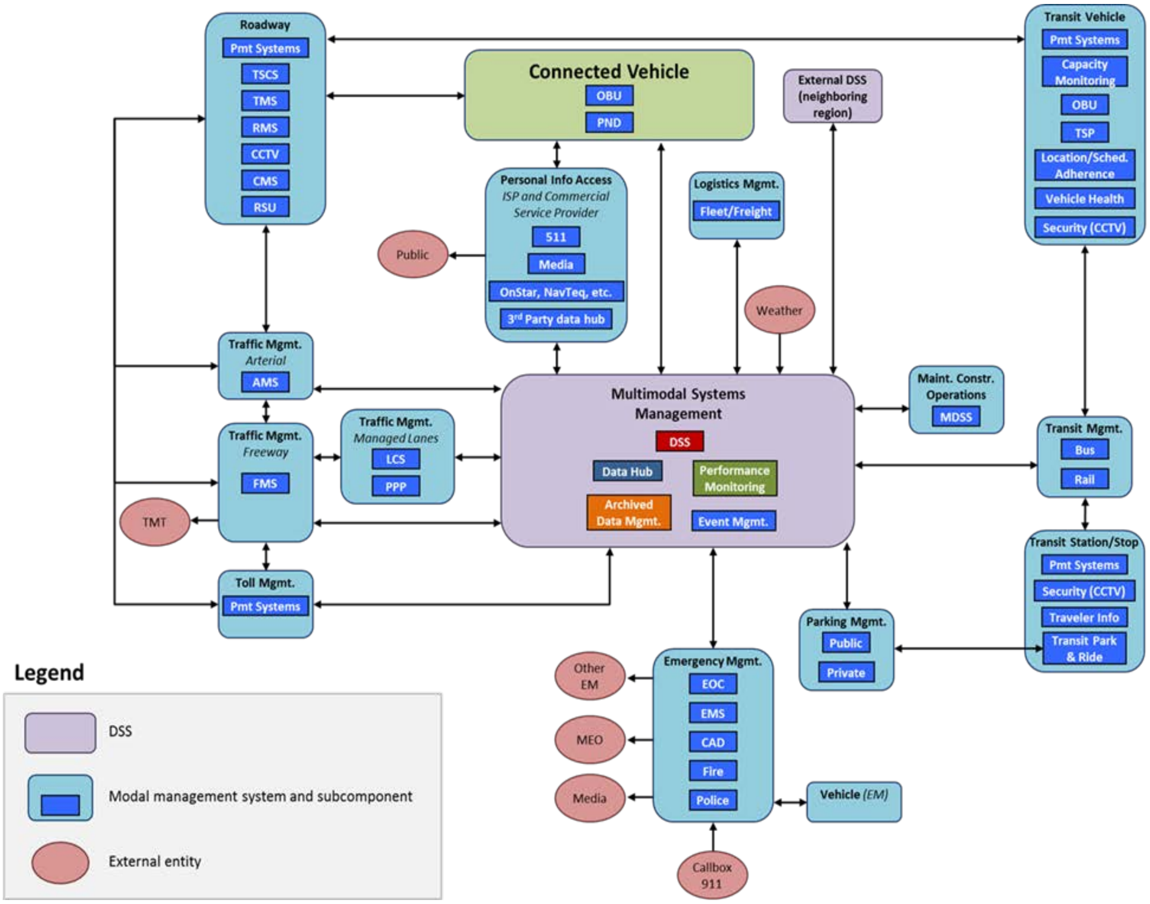

Image Source: USDOT/SAIC/Delcan

Figure 3-1. RTMDSS Generic System Information Flow Diagram

Figure 3-1 illustrates the relationships among a number of transportation operational functions. Big data will be relevant to many of these functions, and big data techniques and technologies can certainly aid in the extensive data sharing across entities that is represented in Figure 3-1. But we focus on a few areas that we believe best exemplify big data potential – areas that directly pertain to utilizing new data sources and communications channels to deliver enhanced system monitoring and management and highly personalized and targeted, traveler-centric transportation strategies. Figure 3-2 consists of the RTMDSS diagram (Figure 3-1) with the addition of overlaid annotations showing four numbered potential impact areas for big data analytics.

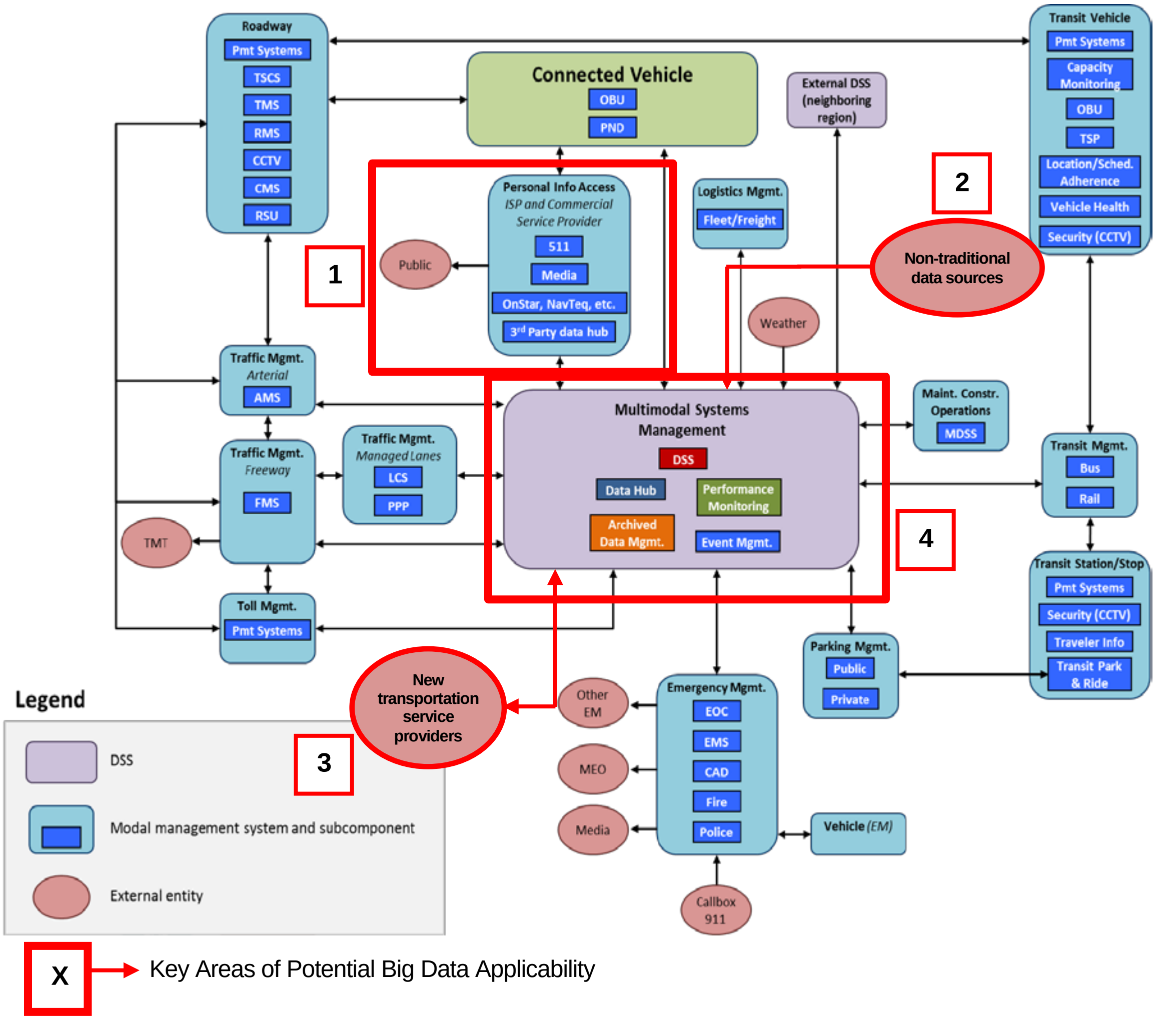

Image Source: USDOT/SAIC/Delcan/Volpe Center

Figure 3-2. RTMDSS Functional View Highlighting the Role of Big Data

Key Areas of Potential Impact for Big Data Analytics in RTMDSS Information Flow Scenario

1. Increased interaction between travelers/users and their personal devices

   Two-way exchanges of transportation-relevant information between users and their personal devices (smartphones and their successors) – that is, gathering as well as disseminating information through these devices – will become more frequent, richer, and more important to transportation decisions. For example, trip-planning smartphone apps commonly accept travel plan (itinerary) information. If and when big data scenarios such as personalized carpool ride-matching emerge, these apps may learn traveler preferences, link them to data streams about external events, both related to and apparently unrelated to transportation, and use this knowledge to foresee traveler needs and desires and thereby offer greater value to the traveler. At the same time, to the extent that these app-generated travel predictions can be anonymized and transmitted to the RTMDSS, multimodal system performance predictions and management should become more robust and effective.

2. Non-traditional data sources

   Some data not traditionally associated with transportation decisions are likely to become more important in predicting transportation system performance. This is partly driven by the fact that data are becoming available on more and more decisions and activities that people undertake. For example, currently available smart home technologies can know which rooms of the house are occupied at which times and the typical smart home system allows for some remote reporting and control capability. Following the UPS example cited in Chapter 2, we can imagine a time when people's patterns of movements in their homes can be linked to a particular departure time, maybe even by a particular mode, and perhaps even for a particular trip purpose or type of destination. It may eventually become possible to predict transportation system breakdowns by analyzing underlying patterns, just like UPS predicts part failures by analyzing part vibration patterns.

3. New transportation service providers

   Information technology, including but not limited to big data analytics, will allow companies or other entities to better understand travelers' transportation needs. Assuming privacy issues are properly addressed, such as through opt-in approaches or other means, this could be accomplished by 1) mining travelers' social media presence (as web-based e-mail providers do now by analyzing e-mail content), 2) measuring their movements by passive means (as when cell phone service providers use tower triangulation), or 3) by interacting with them directly through trip-planning apps. As these transportation needs are better understood, organizations will have increasing opportunities to provide targeted transportation services, as has begun with commercial carsharing services such as Zipcar and ridesharing services such as Avego. Increasing data will likely lead to more and different transportation service opportunities. This is not to say that we should expect a future full of hundreds of tiny service providers, as some market consolidation is inevitable. Rather, the point here is that there will likely be numerous transportation service providers working alongside personal vehicles and fixed-route transit to compose the multimodal transportation system.

4. Dramatically improved multimodal systems management

   Big data analytics should result in dramatic improvements in system performance predictions. More than anything else, this will result from the influx of more system performance data, thanks to connected vehicles and connected travelers, and new diverse and contextual data from non-transportation sources. Although this amounts to a change in effectiveness in a current activity rather than an entirely new function or activity, it warrants special attention because of its transformative potential.

3.4 Summary

Big data has the potential to improve transportation operations both incrementally and radically. Big data approaches to data capture and management analysis will accelerate and improve a range of ongoing and envisioned transportation operations applications. There is also real potential for dramatic benefits from big data analysis, particularly in the realms of transportation system monitoring and management and in traveler-centered transportation management. How much of this potential can be realized is an open question. One way to visualize the potential role of big data is to visualize what big data may enable within a future, hypothetical connected vehicle environment. At a high level, there are at least four ways in which big data can provide value in such an environment: 1) Increased and leveraged interactions between travelers/users and their personal devices; 2) Incorporation of non-traditional data sources; 3) Opportunities for new transportation service providers; and 4) Improved multimodal systems management.

Chapter 4 What Does Big Data Mean Now in Transportation Operations?

Chapter Objectives
• Broadly characterize the extent of current Big Data practice in transportation operations • Describe real-world examples of Big Data as defined in Chapter 2 and value as proposed in Chapter 3

Chapter 2 presented a definition of a fully-realized version of big data and Chapter 3 hypothesized how big data might provide substantial value in transportation operations. This chapter explores the extent to which big data approaches are currently being utilized in the practice of transportation operations. That is, do we find real-world examples of the theoretical conception of big data presented in Chapter 2 and are transportation operators and/or private entities utilizing big data to extract the sort of value hypothesized in Chapter 3? This chapter is organized into three sections, with the first focusing on private sector practices, the second focusing on public sector (transportation agency) practices, and the last extracting key points in a summary.

This assessment is based on a sketch-level review of current practice and is explicitly not exhaustive. Rather, we focus on a handful of examples that are intended to exemplify the state of the practice. Further, this assessment of the presence or absence of big data characteristics is not intended to judge or disparage the value of the practices nor their appropriateness to the purposes to which they have been designed and applied. Finally, in the case of the private-sector examples, the lack of available information on some of the key, proprietary details of the approaches somewhat limits our ability to fully judge the extent of big data characteristics.

4.1 Private Sector Activities

Numerous private-sector organizations offer big data-related transportation services and activities. Waze is a roadway navigation application for smartphones that relies on actively and passively crowdsourced data to recommend routes and estimate time of arrival. Its acquisition last year by Google signals that it may be a particularly advanced and promising smartphone navigation app. It features some aspects of big data:

- Massive data streams**:** As of June 2013, Waze was reported to have 50 million GPS-trace-generating users worldwide (Federman & Rosenthal 2013).
- Crowdsourced and diverse data: In addition to the probe vehicle GPS traces that they passively contribute, Waze users also report real-time gas prices, alerts on traffic incidents, and descriptions of local events such as county fairs. In 2012, 36 million drivers contributed 90 million reports (Shu 2013).
- Predictions using fused data: To estimate time of arrival and optimal route choice, Waze relies on historical average speeds on road segments for initial predictions and updates them with real-time GPS trace data (Waze 2014).

Waze has begun partnering with cities to provide traffic monitoring services. Its first partnership, with Rio de Janiero, was announced in July 2013, and the company plans to partner with at least nine others (Eisner 2014). As of February 2012, it also provided traffic-monitoring services to 23 television news programs (TechZulu 2013).

IBM partners with a number of cities worldwide, including Singapore and Stockholm, to provide traffic monitoring and predictions (Wakefield 2013). Their “Intelligent Transportation” product aggregates and analyzes multisource traffic data – both historical and near real time – to provide system monitoring information to the traffic-management center. The Traffic Prediction Tool produces network speed and volume forecasts up to 60 minutes into the future. Neither service requires unusually massive or diverse data streams, or data from probe vehicles. Connections to big data methods are not evident from the literature reviewed for this paper (Cotton 2013; Schaefer et al. 2011; Yelchuru et al. 2013), although IBM is likely using proprietary analytical methods that have not been publicly disclosed. Given our limited visibility into IBM’s approach, this solution seems to be an example of advanced but essentially traditional, rather than big data, transportation data capture, and management and analysis methods.

Ridesharing (carpooling) apps are beginning to approach the sort of personalized, recommendation engine approach to transportation that was described in Chapter 3 as an example of potential big data application to transportation operations. MyRideBuddy is a dynamic, real-time ridesharing app available in Singapore. Using an “advanced real-time algorithm,” it recommends ride partner to potential drivers and vice-versa by matching origin and destination and also personal preferences. The service includes electronic cost sharing for the ride, so no cash changes hands. An app called Green Monkeys provides a similar service in France, and can help match carpoolers with unusual schedules semi-automatically (Murphy 2012). The two examples appear to offer sophisticated matching capabilities – akin to an automated carpooling bulletin board typical on college campuses years ago – but do not seem to include the automated preference learning that made Netflix and Amazon successful.

Companies are also seeking to measure travelers’ movements using cell phone data. For example, the technology used by AirSage relies on the position of cell phones relative to nearby cell towers, not on GPS or Bluetooth. The approach offers excellent coverage of the U.S. population, 91percent of which owns a cell phone as of 2013 (AirSage 2014). Furthermore, unlike smartphone-based approaches that rely on the user to leave on battery-draining GPS or Bluetooth transmission, AirSage’s approach captures the locations of cell phones whenever they are turned on, regardless of other user settings. However, cell phone location can be triangulated only to within 150 meters, which is much lower precision than GPS, and individual data is rarely available in real time, although aggregated data may be (Calabrese 2011).

The AirSage approach reflects at least one aspect of big data: it leverages massive data streams. AirSage collects 15 billion data points per day from 100 million cell phones. However, the inability to access individual travel behavior in real time (lacking the high “velocity” that typifies big data) has restrained its application to transportation operations. Instead, AirSage primarily markets its service for transportation planning.

4.2 Public Sector Activities

4.2.1 San Diego and Dallas Integrated Corridor Management

The Dallas and San Diego ICM projects focus on optimizing the use of all available multimodal infrastructure in a corridor by: 1) Efficiently distributing demand (via traveler information) physically and temporally across available capacity and, 2) Manipulating the available capacity by adjusting traffic-signal timing or pressing additional transit vehicles into service. Both the Dallas and San Diego systems are still evolving. Both sites aim to provide transportation operators with: 1) as full a picture of real-time transportation system conditions as possible, and 2) new Decision Support System tools for identifying appropriate and coordinated strategies to respond to planned and ad hoc system disruptions (e.g., construction, recurring congestion, and incident-related congestion). Given these dual objectives, data capture, management and analysis are essential functions.

Overall, both sites include a number of characteristics common to big data but do not yet employ a number of characteristics related to analytics which we argue most clearly differentiate big data from traditional practice. Both systems are collecting and integrating a large volume of data, but that data is still being collected using primarily fixed, infrastructure-based sensors for a single purpose (transportation management). In addition, the data is being shared and used by a limited number of people – all transportation agency partners – and the datasets are still managed using traditional data management tools. There is no crowdsourcing of data collection. To varying degrees, both sites utilize state-of-the-practice Decision Support Systems featuring macro- and/or microscopic analysis tools that perform a rudimentary form of pattern recognition. They both use an expert system that matches pre-defined response strategies to current conditions, presumably based on a limited number of performance metrics. These portions of the ICM systems are approaching, but not yet fully manifesting, the sort of data science approaches that we argue characterize the fullest embodiment of big data.

4.2.2 Michigan Department of Transportation

The Michigan Department of Transportation (MDOT) has been among the first state or local transportation agencies in the United States to explore the capture, management and utilization of connected vehicle data. The organization is currently pursuing three closely related projects in this area, all in partnership with the Federal Highway Administration:

- Data Use Analysis and Processing (DUAP)
- Integrated Mobile Observations (IMO)
- Weather Response Traffic Information (WxTINFO)

The initial phase of the DUAP project demonstrated that data from a small number of connected test vehicles could be converted to travel-time information. A second phase now in design considers broadly how connected vehicle data may support a wide range of potential business use cases at MDOT. The first use case to be identified is Pavement Conditions and an application in this area will be developed. Data sources include the vehicles operating in the large-scale, on-road USDOT Connected Vehicle Safety Pilot Model Deployment in Ann Arbor, Michigan. Data that has been preliminarily identified as relevant to the Pavement Conditions application includes traffic information (volume, mobility, vehicle weights), weather (precipitation, humidity, freeze/thaw cycles), vehicle accelerometry (X, Y, Z axis readings), and photos.

The IMO project will demonstrate the capability and value of capturing weather-related data (surface roughness and distress, surface temperature and atmospheric conditions, photos) via equipped MDOT snow plows and light-duty maintenance vehicles. The data is expected to be relevant to a wide range of applications, including real-time road quality monitoring, fleet monitoring and management, travel time and incident update information, and visibility monitoring.

The WxTINFO project utilizes the data collected through the IMO project and builds upon the DUAP infrastructure to generate and disseminate Motorist Advisory Warnings through MDOT's dynamic message signs (DMS) and MDOT's "MiDrive" statewide traveler information system. WxTINFO leverages the Regional Integrated Transportation Information System (RITIS) to conduct analysis and compute performance measures including user delay cost, congestion scans (i.e., visual representations of the temporal and physical extent of congestion), and bottleneck ranking. Operated by the University of Maryland Center for Advanced Transportation Technology Laboratory, RITIS is an automated data-sharing, dissemination, and archiving system that includes many performance measures, dashboards, and visual analytics tools that help agencies to gain situational awareness, measure performance, and communicate information between agencies and to the public (CATTLAB 2014).

A definitive assessment of the extent of big data in the MDOT projects is impossible, in part because the projects are still underway and some decisions that could take the projects deeper into or farther away from big data have not been made. However, based on the information at hand, it appears that the three MDOT projects collectively exhibit many aspects of big data but may not represent the fullest manifestation of the approach. Big data characteristics present in the projects include the collection of crowdsourced (probe) data to support a large number of potential uses, focusing mostly on pre-defined uses but also allowing for a wide variety of potential uses by various users, including data exploration using RITIS tools. Big data characteristics that appear absent from the projects include collection and consideration of a very large number of variables (the projects collect a limited number of variables). Also, it is not clear whether the analytical tools and processes used in the projects constitute the level of data science that we propose represent the epitome of big data.

4.3 Summary

There are currently few examples of big data in transportation operations that embody most of the distinguishing characteristics discussed in Chapter 2.0 or the spirit of the potential applications identified in Section 3.2. However, more examples are appearing all the time and any conclusions about the prevalence of big data in transportation operations must be adjusted frequently. Indeed, even during the time between the substantial completion of this white paper and its publication, a number of additional examples have been identified. These include a recent Portland Tri-Met (a public transportation agency) analysis of transit vehicle schedule adherence and bunching problems. Much as imagined in Section 3.2, the Tri-Met analysis considered non-traditional variables and concluded, somewhat unexpectedly, that individual driver behavior plays a greater role than even traffic conditions.

Big data in transportation operations is in a nascent phase. Between the public and private sectors there are few current, robust examples but there is an active dialogue—among agencies, researchers and current and erstwhile big data contractors—and increasing experimentation.

Chapter 5 How Could We Integrate Big Data Practices into Transportation Operations in a Connected Vehicle Environment?

Chapter Objectives

- Summarize findings of related research
- Interpret other findings in light of our framework (Chapters 2 and 3)

This chapter identifies themes among recommendations that have been put forward by other researchers for advancing and leveraging big data for transportation operations, and reviews them in the context of our definition of big data in Chapter 2 and observations on the potential value of big data for transportation operations in Chapter 3. After presenting findings from the source documents, this chapter offers a response to several of the key findings.

5.1 CVTA Activities

On June 18, 2013, the Connected Vehicle Trade Association (CVTA) held a Big Data Workshop, which was attended by automakers and suppliers, US DOT staff, academics, and other CVTA members (McCormick 2013). Participants addressed questions in four topic areas. Core questions and a sample of participants' answers are listed below. The results presented here illustrate the general state of the dialogue among some core stakeholders at an event that focused explicitly on the interface between connected vehicles and big data.

- An Overview of the Data: What data do you expect to be able to collect from vehicles?
  - Vehicle status: GPS location, speed, braking/acceleration events
  - Ancillary sensors: real-time weather, road condition
  - Driver/passenger: infotainment use patterns, customer satisfaction, demographic info (teenager, senior)
- Managing the Data: What are the critical unanswered questions in shaping a data management plan?
  - Who will manage and store vehicle-generated data?
  - Who will have access to it?
  - How will the data generate revenue enough to cover the costs of data collection?
- Shaping the Data: What data could be merged with vehicle data to create new knowledge?
  - Demographics
  - Purchasing patterns
  - Social networks
- Using the Knowledge: How would the knowledge be used?
  - Meet consumer desires

- Develop business leads
- Develop business intelligence

5.2 Real-Time DCM State of the Practice Assessment and Innovations Scan

"Real-Time Data Capture and Management State of the Practice Assessment and Innovations Scan," published by the ITS JPO in July 2011, includes a review of emerging data-management practices in seven sectors of the economy. The focus of the report was on innovations that are likely to meet the requirements of the Connected Vehicle Data Capture and Management Program and have the potential to be ready for the market by 2016. The report was released in three volumes, and contains innumerable findings and conclusions; only the most important and relevant ones are carried forward here.

1. Dynamic Interrogative Data Capture (DIDC) is necessary to reduce the volume of data that will be produced by the connected vehicle environment. DIDC is a term for the use of a strategy for sending given data elements based on internal triggers or external requests, rather than sending and storing a full, predefined set of variables at regular intervals. Although such strategies have been used in other industries, especially those with real-time data-driven decision making, they are not common in transportation data environments. That said, the SAE J2735 standard for Basic Safety Messages from connected vehicles provides a transportation-based example, in that it calls for certain data elements to be transmitted continuously and others only when a specific event occurs.
2. Crowdsourcing will be a natural mode of data capture in a connected vehicle environment.
3. Cloud computing and/or federated data systems will be necessary to store and provide access to the huge volume of data that will be produced by connected vehicles.
4. Data storage and access may be best handled by a third party. Storing and managing the volume of data to be generated by connected vehicles may be cost-prohibitive for the government, and setting up appropriate cost recovery models may be deemed outside the bounds of acceptable government activities. Assuming data privacy and security can be ensured, there may be a natural role for an outside organization to provide access to the data for a fee.

5.3 Big Data and Intelligent Transportation Systems

The ITS JPO/Noblis report, "Big Data and ITS," takes a preliminary look at how trends in big data can and should inform plans for the evolution of ITS technologies and practices. It outlines the types of transportation data that are used and needed, in addition to the sources of data: from infrastructure-based sensors, vehicle-based sensors, and from personal mobile devices such as smartphones. The report concludes that the need for infrastructure-based sensors will wane as vehicle-based sensors and mobile devices become more common, as they provide a lower-cost approach for collecting speed and travel time data. The report also notes that infrastructure-based sensors will be necessary for the foreseeable future to gather data such as vehicle counts, where mobile sensors would be effective only if nearly 100 percent of vehicles were equipped (McGurrin 2013).

The report includes a number of recommendations, some of which are based on its own analysis and others that come from a companion ITS JPO report, "ITS Horizons Scan" (McGurrin et al. 2013). All of the recommendations included in "Big Data and ITS" are listed here:

Role

- The Federal ITS program should focus on the facilitation of standards, with some potential effort on data collection as well.
- The Federal ITS program could function as a national clearinghouse, providing information about ongoing development efforts and management on innovation, as a way to coordinate and leverage research and development efforts and reduce duplicative efforts.
- The Federal ITS program does not need to focus on data privacy issues, as other industries will take the lead in this area.

Strategy

- Ensure that the appropriate resources are in place to store, manage, and analyze the big data coming into the USDOT's ITS program, such as the data coming from the Safety Pilot and upcoming connected vehicle field tests.
- Consider studying the policy and institutional issues, along with potential business models, for establishing connected vehicle data brokers.
- Identify candidate high-priority topic areas at the intersection of key transportation challenges and big data. Examples may include development of big data algorithms to predict system-level traffic phenomena, such as queues and shockwaves, and the use of predictive analytics to more effectively implement integrated corridor management and advanced travel and demand management techniques based on weather, demand, and incident patterns.
- Establish a dialogue with the public-sector transportation community to identify their interest in specific topics for capacity building and technical support relating to big data as well as related topics such as open data, cloud computing, and the Internet of Things. Incorporate the results into the Professional Capacity Building and other information exchange programs.

Tactics

- The fast-paced evolution of the field of big data in the private sector is largely incompatible with the current research and deployment model used in the government sector. Public sector R&D needs to change.
- The Federal ITS program could fund a broader array of research initiatives through the greater use of Broad Agency Announcements, with less dependence on single awardee research projects.

5.4 Emerging Transportation Technologies

“Developing Emerging Transportation Technologies in Texas” is the product of a task force that was assembled specifically to investigate several emerging technologies in transportation – connected vehicles, autonomous vehicles, electric systems, and cloud computing/crowdsourcing (Jin et al. 2013). For each technology, the task force briefly assessed the current state of practice, identified likely near-future developments, and made recommendations for facilitating and accommodating advancement. The report focuses heavily on connected vehicles and autonomous vehicles, but recognizes that cloud computing and crowdsourcing technologies are “well-developed in computer science, but [their] applications in transportation are still in the early development stage with much foreseeable potential.”

It adds that development and application of these technologies is likely to accelerate in the next 5 years. Finally, the document does not focus on specific applications or implementations, but makes several recommendations for the Texas Department of Transportation to support ongoing development and innovation. The most notable recommendation is the development of a data environment conducive to data sharing among both private and public entities.

Describing this proposed data environment, the report identifies five categories of transportation data:

- Travel demand
- Dynamic traffic conditions
- Traffic events (e.g., work zones, incidents, and cultural/sporting events)
- Environmental data (e.g., road conditions, weather, and air quality)
- Transportation infrastructure data (e.g., signage, signal timing, digital maps, and physical infrastructure condition)

The report recognizes that many data sources are underutilized due to the difficulty and/or expense of access for various entities. The task force recommends that the proposed data environment facilitate not only the dissemination of public-sector data, but also the sharing and selling of private-sector data.

The report also acknowledges that certain data ownership and licensing issues may need to be addressed at least in part by legislative or regulatory action.

5.5 Summary

Key conclusions drawn from across the four sources described in this chapter can be summarized as follows:

- Policy and Institutional Issues
    - There are significant unresolved questions pertaining to data ownership and management, including ownership and/or controlled access to various types of connected vehicle and connected traveler data.
    - A likely general strategy for data management that begins to speak to the ownership question would be to utilize one or more third-party “data brokers” rather than have Federal or state/local transportation agencies assume primary responsibility.
    - The Federal role may be in developing data standards.
- Technical Issues
    - As expected, there is definite interest on the part of the private sector in connected vehicle and traveler data. This interest can be used to support the kind of big data purposes seen in other industries, such as infotainment use patterns, customer satisfaction, and demographics to help meet customer desires and inform business development.
    - In terms of technologies and techniques, decentralized approaches are essentially a given, including crowdsourcing data collection and cloud-based and federated data management.
    - There may be a need to prioritize data collection, through DIDC or another approach.

Fundamentally, we agree with the policy-related conclusions above. The question of who manages access to what data will play a major role in determining what data is collected and how it is used. This question becomes more pressing when we consider the diversity of data producers and users implied by the discussion of crowdsourcing in Chapter 2 and personalized mobility services in Chapter 3. As to the data broker concept, it is likely that a private entity may have more flexibility in charging for access to recover costs and managing privacy. The private entity would also likely be more adaptable to changing needs of the potentially large, diverse communities of data providers and consumers mentioned in Chapters 2 and 3. If Federal entities are not directly involved in capturing or managing the data, developing and promulgating a standard that specifies the data needed to manage system-level mobility may be the best way to help ensure that data is available. The standard should provide the additional benefit of improving interoperability of mission-critical mobility services. It should be noted however that development and promulgation of a data standard does not ensure use.

The technical conclusions also resonate. Chapters 2 and 3 emphasize decentralized activity. The number and diversity of data providers and also data users should increase. Crowdsourcing and cloud computing are standard practice now, in any case. Data federation – leading data users to existing datasets stored elsewhere rather than consolidating all the data in a single location – is a logical way to avoid undue costs, especially when datasets get very large.

The technical conclusion on DIDC requires some explanation. Although DIDC is not a widely used term, the practice is common in closed, data-driven systems such as automobile control systems. Within these closed systems and in general, the underlying ideas make sense.

- There are costs to transmitting data within a system, and there may not be value in transmitting every piece of possible data. So, to the extent that it is expensive to transmit and manage the data, we need to identify priorities about what data is needed when.
- Some of that data will be necessary only at certain events or trigger points.
- It would be valuable to be able to change the data transmission and storage strategy.

The process of determining what data should be captured under what circumstances should include an in-depth consideration by a broad stakeholder group of big data analysis opportunities. Although this white paper has described big data as capturing, managing, and analyzing all or most of a system's data points, such an approach may not be feasible in every circumstance. Final decisions should seek to balance feasibility and other practical considerations against the risk of precluding potentially important future uses.

It appears that the policy and institutional recommendations outnumber and outweigh the technical recommendations in the sources we reviewed. If that is true, it may be because solving the policy challenges would unleash the technical know-how to solve the technical issues, too.

Chapter 6 Conclusions and Next Steps

This section identifies the major conclusions corresponding to each of the major purposes and objectives of the previous chapters, including those pertaining to the definition of "big data," the potential value of big data for transportation operations, and the state-of-the-practice of big data for transportation. Following the summary of conclusions, the implications of those conclusions are identified in terms of next steps and direction moving forward.

Chapter Objectives

- Identify and distill key conclusions
- Provide readers a short, clear list of primary takeaways
- Recommend next steps

6.1 Conclusions

The major conclusions of this white paper can be summarized as follows:

1. **A shared, comprehensive understanding of big data among transportation stakeholders should not be assumed, and we have provided a description that is intended to promote one.** This white paper describes big data as an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that they exceed the capabilities of traditional data management and analysis tools. Big data is described as often drawing on automatically system-generated data encompassing entire populations and utilizing techniques such as crowdsourcing, cloud computing, and automated pattern recognition analysis. Big data is contrasted with traditional approaches, acknowledging that there is a continuum and projects or approaches can manifest some, but not all, big data characteristics and somewhat, but not fully, leverage the potential value of big data. In summary, the description of big data provided here identifies the following characteristics that distinguish it from traditional approaches:

    - Data capture that includes massive datasets encompassing all or most of the population (as opposed to small samples); data collection to enable both pre-defined and a wide variety of un-defined analyses; and utilization of crowdsourced and "electronic breadcrumb" data.
    - Data Management that features storage in decentralized and virtual locations (i.e., cloud) and handles structured and unstructured data.
    - Data Analysis that is often more automated, with computers doing more of the work to find large and complex patterns among a massive number of variables that may not intuitively appear related.

2. **Big data has significant potential for transportation operations.** Because big data is very different than traditional transportation operations approaches and because it often includes discovering value through data exploration, it is impossible at this point to identify with great confidence all of the specific uses and benefits – as impossible as it would have

been to identify 20 years ago all that the Internet would enable. However, there appears to be potential in at least two broad areas:

- Transportation System Monitoring & Management – This area focuses on the transportation system and utilizing a wide variety of data – massive volumes of connected vehicle and connected traveler data – to paint a much richer picture of real-time conditions throughout the system and, critically, to enable much better predictions of impending conditions. Examples include using connected vehicle data on braking and lane changing to anticipate imminent traffic flow breakdowns or using vehicle, weather and traveler behavior data to better predict schedule adherence breakdowns and the resulting bunching of transit buses.
- Traveler-centered Transportation Strategies **–** This area focuses on the traveler and raises many privacy issues and questions about public versus private roles that would need to be explored before moving forward. This area could utilize a wide variety of connected traveler data to: 1) Construct detailed profiles describing the behaviors and inferring the preferences and priorities of individual travelers, 2) Utilize these profiles to develop highly personalized traveler information, travel demand management and other strategies – including monetary and/or non-monetary incentivization schemes, and 3) Implement these strategies by communicating directly with individual travelers in real-time, with highly context and location-specific information, utilizing the same handheld devices that are used to collect the traveler data that builds the profiles. Examples include very sophisticated and dynamic ridesharing and transit ridership incentivization.

3. **There is some big data work currently being done but not a lot that fully embodies the potential hypothesized in Chapter 3.** Some of the private sector approaches in particular exemplify several aspects of the potential approaches described in Chapter 3, such as Waze's use of crowdsourced data. These findings provide some validation of that potential but also indicate that much of it is still unrealized, especially since extensive connected vehicle data is not yet available or being incorporated. These findings suggest that additional exploration of big data implications for the DCM Program and elsewhere is warranted.

4. **Existing big data research activities may need to be broadened.** There has been some ITS JPO sponsored research on big data within the Connected Vehicle Program and some recommendations have been offered. For the most part, these recommendations reflect an intentional focus on data collection and management with less consideration of the implications of big data analysis. More explicit and deeper consideration of big data analysis is important because it can directly influence decisions about USDOT data capture and management-enabling activities. At a minimum, if DIDC includes selective data capture based on specific, pre-defined data needs, it is important that those needs be informed by the widest possible range of potential analyses – traditional and big data. Evolution of the ITS JPO data research program from the "Connected Vehicle Real-Time Data Capture and Management Program" to the "Connected Data Systems Program", and creation of the ITS JPO program category, "Enterprise Data," provide good mechanisms for increasing the focus on big data analytics (these are described in the ITS JPO "ITS Strategic Plan, 2015-2019": http://www.its.dot.gov/strategicplan).

5. **Understanding and capitalizing on big data requires broader, non-traditional and non-transportation stakeholder engagement.** There are few, if any, fully-realized examples of big data in transportation operations, and the people that know big data best are from outside the transportation operations community. Further, in a big data paradigm, roles and relationships are potentially less centralized, consolidated and governmental than traditional

approaches – concepts discussed more explicitly at the end of Chapter 5. Finally, many of the operational strategies that may be enabled through big data are not familiar to many traffic and transit managers and require, for example, greater depth of knowledge in statistics, machine learning, traveler demand management and the science of customer profiling and marketing.

6. **Costs and other resource implications are uncertain.** A transition to increased use of big data techniques for data capture, management and analysis would likely introduce additional costs or resource requirements – costs that may or may not be off-set by other cost savings or benefits generation – but it is not at all clear at this point. These issues were not explored in the recent USDOT studies that were reviewed and cannot be meaningfully addressed within this paper. It may be that some applications of big data approaches will require additional volumes or types of data, for example – a consideration that requires additional specification of big data uses cases before it can addressed. More likely still, using big data approaches may require some investment in in-house training and/or outside consulting services, such as those associated with mining data to identify patterns and profiles, developing predictive algorithms, and incorporating the algorithms into existing expert systems or Decision Support Systems. Those possibilities likewise depend on further elaboration of specific big data use cases as well as consideration of public versus private and state/local versus Federal roles and responsibilities in connected vehicle data capture overall. To guide programmatic decisions, any incremental costs would have to be weighed against savings due to the retirement of superseded technologies and any performance benefits resulting from big data implementation. The fact that some agencies have begun to outsource probe vehicle data capture, management and/or analysis to INRIX, IBM and others, suggests that outsourcing of some functions may provide a favorable value proposition. Nevertheless, any conclusions specific to the DCM Program will depend on further research on big data and on the broader question of connected vehicle data capture and management techniques and roles.

6.2 Next Steps

In this section, recommended next steps are identified in three broad areas:

1. Incorporating input from a diverse group of stakeholders, in particular reactions to the proposed value proposition from Chapter 3
2. Ensuring coordination as appropriate with related areas of research, such as the USDOT Connected Vehicle Policy program work on data policy
3. Additional new activities.

Next Steps

- Incorporate input from a diverse set of stakeholders
- Identify and link to related activities
- Consider next steps, including identification and prioritization of big-data use cases
- Develop and demonstrate selected use cases

6.2.1 Incorporating Input from a Diverse Set of Stakeholders

An appropriate initial step forward is to determine whether the definitions, concepts and potential benefit identified in this white paper resonate with transportation operations stakeholders. Is there agreement on the potential value of big data and what examples or uses cases best illustrate the potential?

It will be important to vet the findings of this paper with a broad range of stakeholders that will include much more than the traffic- and transit-management communities. Relevant stakeholders include:

- Transportation modelers, detection algorithm developers and other key players in the areas of prediction and Decision Support Systems
- Transportation operations managers (freeways; arterials; transit bus, rail, ferry; freight; borders)
- Travel Demand Management and Active Transportation and Demand Management practitioner communities
- Transportation planners
- Asset managers
- Transportation agency and private-sector mapping stakeholders
- Current practitioners of big data – managers and senior technical personnel from those industries and companies where the practices are most common and well developed
- Experts (public or private) in connected traveler and comparable person/customer information.

6.2.2 Identify and Link to Related Activities

This white paper references concepts that are likely under consideration and which impact and are impacted by other USDOT programs, including others within the Connected Vehicle Program such as Dynamic Mobility Applications, road weather and environmental programs. It will be important to identify these other activities and the extent to which they have reached corroborating or contradictory conclusions, the scope and status of their activities, and to map the overlap and potential gaps between what they are doing and the steps implied by the conclusions of this white paper, some of which are discussed in this concluding chapter. Emerging from this scan and consultation activity will be a clearer sense of how the conception of big data presented in this paper should be coordinated with or represented within other activities, including:

- The continued consideration of prior big data research recommendations such as DIDC;
- Connected vehicle applications development and testing, including remaining Dynamic Mobility Applications Program activities and the anticipated Connected Vehicle Pilot Deployments;
- Integrated Corridor Management activities, and other USDOT activities, related to Decision Support Systems;
- Connected Vehicle Program and other USDOT work in the area of person/traveler probe data collection and traveler information and travel demand management strategies; and
- Travel Demand Management research.

Particularly important will be to further investigate the extent to which big data approaches are or are not represented in the Dynamic Mobility Applications and the other (safety and environment) applications considered in the AASHTO Footprint Analysis and the USDOT CVRIA. The tentative conclusion of this white paper – that the paradigm is not as fully represented as it could be – can be verified through further investigation and dialogue.

One of the issues that should be addressed in this scan and mapping activity is: to what extent the DCM Program's focus on data capture and management – as opposed to data analysis – impacts the extent of DCM Program involvement or leadership in further exploring big data. Insomuch as understanding the intended users and uses of data directly drives decisions about capture and management, it seems that there is value in DCM Program participation, if not leadership, in any activities related to further elaboration of big data and its implications.

6.2.3 Looking Ahead to Potential Future Activities

Activities beyond those discussed earlier in this chapter are speculative. They are dependent on the outcomes of initial steps: 1) vetting this paper's conclusions and 2) determining whether and how considerations from this white paper are already being considered by others. Assuming that the conclusions of this white paper are found to be worth pursuing further and are not already being sufficiently addressed through other efforts, there are several next steps that should be considered. Although related and coordinated, these activities can be conceptually categorized into two main areas, one focusing on coordination with existing activities and one focusing on potential new DCM big data-centric activities:

Potential Future Activities

- Coordination with existing activities, including DIDC, CV infrastructure and data ownership and privacy policy development
- New activities to identify, prioritize, develop and demonstrate big data analytical paradigm use cases

- Coordination with Existing Activities – There are three areas where it could be useful to ensure a conduit and coordination between the further elaboration of the potential big data analytical paradigm and:
    - Any continuing consideration of the DIDC concept;
    - The AASHTO Footprint Analysis, FHWA Connected Vehicle Deployment Guidance, and other activities that include consideration of the design and operation of those aspects of the connected vehicle infrastructure (in-vehicle and roadside) that have data capture and management implications; and
    - The various Connected Vehicle Policy Program and related activities that are investigating data ownership and privacy issues and the implications for public versus private roles and responsibilities in the connected vehicle environment.
- Potential New Activities – To the extent to which promising big data analytical paradigm practices are not being fully considered through existing connected vehicle application development and testing activities, it would be appropriate to consider a new, complementary area of research, which could include the following steps:
    - Develop a data management roles framework for connected vehicle and connected traveler data that articulates the USDOT and other roles (this does not pre-suppose a particular role for USDOT), ensuring that the framework reflects the growing role of connected travelers and their mobile devices. This roles framework will be crucial for identifying opportunities for targeted DCM efforts.

  - Leverage the Connected Vehicle Policy Program to evaluate current public- and private-sector activities, roles, and partnerships in collecting, managing and analyzing data, focusing on connected traveler data.
  - Draft one or more preliminary concepts for the USDOT role and the roles of others in managing transportation operations data, drawing guidance from the Connected Vehicle Reference Implementation Architecture.
  - Draw on the big data use cases discussed below to test and revise the concept.

- Develop a list of potential big data analytical paradigm use cases describing what specific analysis would be done, what data and analysis tools would be needed, and how the knowledge gained would enhance transportation operations' effectiveness. This activity would need to draw upon input from a diverse group of stakeholders and crowdsourcing, in addition to more traditional methods such as interviews with selected stakeholders.
- Evaluate and prioritize the list of potential use cases based on a number of criteria that could include the magnitude of potential benefits/severity of the problem they would address.
- Identify the data standard development implications of the data required for the more important use cases.
- Assess data availability to support development of analysis tools and techniques for the highest-priority use cases. Is the data available now via any of the connected vehicle test beds; is it available in the RDE; is it available from the DMA small-scale prototype demonstrations; does it need to be built into the upcoming Connected Vehicle Pilots; does some of it need to be generated using some special, dedicated new activity?
- Based on the priority of the use cases and data availability, select a limited number of use cases to be developed and tested. Data would be obtained or generated, tools and techniques developed and exercised, and the predictions or other outputs of the tools tested. Key questions to be addressed through this development and testing work include those pertaining to data, data standards, data capture and management, algorithms, statistical testing, the cost-benefit of the new approaches relative to existing approaches, and the extent to which the new big data knowledge-generation approaches can lead to more effective transportation management strategies – that is, now that we understand it better, can we feasibly do anything to influence it?
- These activities could culminate with the identification of activities needed to support any specific transportation management strategies that are suggested by the improved understanding achieved through the big data analysis.
- Further investigate the potential cost and other resource implications of the adoption of big data approaches. This work will depend in part on the verification and elaboration of the utility of big data approaches discussed in other recommended next steps, including elaboration of big data use cases and the associated data and other requirements. The costs and resource requirements associated with big data are also a function of, and should be addressed within, the context of the broader question of connected vehicle environment roles and responsibilities that are being investigated in part through DCM Program activities and also through the USDOT Connected Vehicle Policy Program.

References

AirSage. (2014). The Future of Transportation Studies: A Comparative Review. http://www.airsage.com/Contact-Us/White-Paper/. Accessed February 12, 2014.

Calabrese, Francesco. (2011). Urban sensing using mobile phone network data: Ubicomp Tutorial. IBM Smarter Cities Technology Centre. http://research.microsoft.com/en-us/people/xingx/t05-urban-sensing.pdf. Accessed February 12, 2014.

Chui, Michael, Andy Miller, Roger Roberts. (2009). “Six ways to make Web 2.0 work,” McKinsey Quarterly (February).

ComputerWeekly. (2013). “What does a petabyte look like?” http://www.computerweekly.com/feature/What-does-a-petabyte-look-like. Accessed March 13, 2014.

Cortes, Corinna, Vladimir Vapnik. (1995). “Support-vector networks,” Machine Learning 20(3).

Cotton, Brian. (2013). Intelligent Urban Transportation: Predicting, Managing and Integrating Traffic Operations in Smarter Cities. Frost & Sullivan White Paper.

Dean, Jeffrey, Sanjay Ghemawat. (2004). MapReduce: Simplified data processing on large clusters. Sixth Symposium on Operating System Design and Implementation, San Francisco, CA.

Dhar, V. (2013). "Data science and prediction." Communications of the Association of Computing Machinery 56 (12): 64.

Eisnor, Di-Ann. (2014). There’s an App for That: Changing the Face of Urban Mobility. 93rd Annual Meeting of the Transportation Research Board, January 12-16, 2014.

Federman, Josef, Max J. Rosenthal. (2013). Waze sale signals new growth for Israeli high tech. http://news.yahoo.com/waze-sale-signals-growth-israeli-high-tech-174533585.html. Accessed February 11, 2014.

Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. (2003). “The Google file system.” 19th ACM Symposium on Operating Systems Principles, Lake George, NY.

Howe, Jeff. (2006). “The Rise of Crowdsourcing.” Wired. June 14, 2006.

Iteris. (2014). Connected Vehicle Reference Implementation Architecture: Applications. http://www.iteris.com/cvria/html/applications/applications.html. Accessed March 13, 2014.

Jin, Peter J., Dan Fagnant, Andrea Hall, C. Michael Walton, Jon Hockenyos, Mike Krusee. (2013). Developing Emerging Transportation Technologies in Texas. Center for Transportation Research: University of Texas at Austin. http://library.ctr.utexas.edu/ctr-publications/0-6803-1.pdf. Accessed June 4, 2014.

Madrigal, Alexis. (2014). How Netflix Reverse Engineered Hollywood. The Atlantic. http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/ Accessed January 16, 2014.

Mayer-Schönberger, Viktor, Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt. p. 59.

McCormick, Scott. (2013) Big Data Workshop White Paper. Connected Vehicle Trade Association.

McGurrin, Michael. (2013). Big Data and ITS. Noblis, Inc.

McGurrin, Michael, Amy Jacobi, Dawn Hardesty, William Ball, Mark Dunzo, H. Gilbert Miller, Stan Pietrowicz, Phil Tarnoff. (2013). ITS Horizon Scan: The Societal, Technical, and Environmental Trends That Will Influence ITS Research and Deployment. Noblis, Inc. Report No. FHWA-JPO-13-090

Murphy, Ken. (2014). Société de transport de Montréal (STM) Aims to Boost Ridership by 40% with a Mobile App. SAP insider. http://sapinsider.wispubs.com/Assets/Case-Studies/2014/January/STM. Accessed March 13, 2014.

Murphy, Samantha. (2012). How High-Tech Carpooling Saves Gas, Money and Time. Mashable. http://mashable.com/2012/05/09/urban-carpool/. Accessed July 23, 2013.

National Cooperative Highway Research Program. (2014). Project 03-101: Costs and Benefits of Public Sector Deployment of Vehicle-to-Infrastructure Technologies. SAIC.

SAIC & Delcan. (2011). Assessment of Emerging Opportunities for Real-Time, Multimodal Decision Support Systems in Transportation Operations: Concept of Operations Final Draft. RITA Report No. FHWA-JPO-10-058.

Schaefer, Steffen, Colin Harrison, Naveen Lamba, Vishwanath Srikanth. (2011). Smarter Cities Series: Understanding the IBM Approach to Traffic Management.

Shu, Catherine. (2013). Nav App Waze Says 36M Users Shared 900M Reports, While 65K Users Made 500M Map Edits. Techcrunch.com. February 6. http://techcrunch.com/2013/02/06/nav-app-waze-says-36m-users-shared-900m-reports-while-65k-users-made-500m-map-edits/. Accessed February 11, 2014.

TechZulu. (2013). Diann Eisnor of Waze | Source13 by Flurry Interview. http://www.youtube.com/watch?v=2VPffOLNne0. Accessed January 29, 2014.

Transportation Research Board. (2013). TRB Webinar: State of the Art Use of Probe Vehicle Data. http://www.trb.org/main/blurbs/169777.aspx. Accessed March 13, 2014.

University of Maryland Center for Advanced Transportation Technology Lab (CATTLAB). (2014). Regional Integrated Transportation Information System (RITIS). http://www.cattlab.umd.edu/?portfolio=ritis. Accessed June 4, 2014.

U.S. Department of Transportation. (2010). "Frequency of Target Crashes for IntelliDrive Safety Systems." USDOT/NHTSA/Volpe Center.

U.S. Department of Transportation. (2011). Real-Time Data Capture and Management State of the Practice Assessment and Innovations Scan; Task 3: Memo on Dynamic Interrogative Data Capture.

U.S. Department of Transportation. (2012). "Benefits of Dynamic Mobility Applications, Preliminary Estimates from the Literature – Final Report." USDOT, 1.

U.S. Department of Transportation. (2013). Some Observations on Probe Data in the V2V World: A Unified View of Shared Situation Data. U.S. DOT ITS Joint Program Office.

U.S. Department of Transportation: Intelligent Transportation Systems. (2014). Connected Vehicle Concepts of Operations (ConOps). http://www.its.dot.gov/connected_vehicle/connected_vehicle.htm. Accessed March 13, 2014.

Vanderbilt, Tom. (2013). The Algorithm Method: How Netflix Determines What You Watch. Wired, August. pp. 56-58.

Verma, Ruchi, Sathyan Ramakrishna Mani. (n.d.) "Using analytics for insurance fraud detection." http://www.infosys.com/FINsights/Documents/pdf/issue10/insurance-fraud-detection.pdf. Accessed January 14, 2014.

Wakefield, Jane. (2013). Tomorrow's cities: Do you want to live in a smart city? BBC: August 18. http://www.bbc.co.uk/news/technology-22538561. Accessed February 11, 2014.

Waze. (2014). Wiki: How Waze calculates routes. https://www.waze.com/wiki/How_Waze_calculates_routes. Accessed February 11, 2014.

Winterford, Brett. (2013). How Montreal transport skirted privacy laws. ITnews for Australian Business. http://www.itnews.com.au/BlogEntry/360480,how-montreal-transport-skirted-privacy-laws.aspx/0. Accessed October 15, 2013.

Yelchuru, Balaji, Sashank Singuluri, Swapnil Rajiwade. (2013). Active Transportation and Demand Management (ATDM) Foundational Research: Analysis, Modeling, and Simulation (AMS) Capabilities Assessment. Report No. FHWA-JPO-13-021.

APPENDIX A: Big Data Techniques and Technologies

The following list of big data techniques and technologies is excerpted from the McKinsey Global Institute report, “Big data: The next frontier for innovation, competition, and productivity,” which was published in June 2011. As of December 2014, the report was available here: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

Techniques for Analyzing Big Data

There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyze datasets. In this section, we provide a list of some categories of techniques applicable across a range of industries. This list is by no means exhaustive. Indeed, researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data. We note that not all of these techniques strictly require the use of big data—some of them can be applied effectively to smaller datasets (e.g., A/B testing, regression analysis). However, all of the techniques we list here can be applied to big data and, in general, larger and more diverse datasets can be used to generate more numerous and insightful results than smaller, less diverse ones.

A/B testing. A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control and treatment groups (see statistics). When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called “A/B/N” testing.

Association rule learning. A set of techniques for discovering interesting relationships – “association rules” – among variables in large databases. These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing (a commonly cited example is the discovery that many supermarket shoppers who buy diapers also tend to buy beer). Used for data mining.

Classification. A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. One application is the prediction of segment-specific customer behavior (e.g., buying decisions, churn rate, consumption rate) where there is a clear hypothesis or objective outcome. These techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning. Used for data mining.

Cluster analysis. A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning. Used for data mining.

Crowdsourcing. A technique for collecting data submitted by a large group of people or community (i.e., the "crowd") through an open call, usually through networked media such as the Web (Howe 2006). This is a type of mass collaboration and an instance of using Web 2.0 (Chui et al. 2009).

Data fusion and data integration. A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery. Data from social media, analyzed by natural language processing, can be combined with real-time sales data, in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behavior.

Data mining. A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behavior of customers.

Ensemble learning. Using multiple predictive models (each developed using statistics and/or machine learning) to obtain better predictive performance than could be obtained from any of the constituent models. This is a type of supervised learning.

Genetic algorithms. A technique used for optimization that is inspired by the process of natural evolution or "survival of the fittest." In this technique, potential solutions are encoded as "chromosomes" that can combine and mutate. These individual chromosomes are selected for survival within a modeled "environment" that determines the fitness or performance of each individual in the population. Often described as a type of "evolutionary algorithm," these algorithms are well-suited for solving nonlinear problems. Examples of applications include improving job scheduling in manufacturing and optimizing the performance of an investment portfolio.

Machine learning. A subspecialty of computer science (within a field historically called "artificial intelligence") concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Natural language processing is an example of machine learning.

Natural language processing (NLP). A set of techniques from a subspecialty of computer science (within a field historically called "artificial intelligence") and linguistics that uses computer algorithms to analyze human (natural) language. Many NLP techniques are types of machine learning. One application of NLP is using sentiment analysis on social media to determine how prospective customers are reacting to a branding campaign.

Neural networks. Computational models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find patterns in data. Neural networks are well-suited for finding nonlinear patterns. They can be used for pattern recognition and optimization. Some neural network applications involve supervised learning and others involve unsupervised learning. Examples of applications include identifying high-value customers that are at risk of leaving a particular company and identifying fraudulent insurance claims.

Network analysis. A set of techniques used to characterize relationships among discrete nodes in a graph or a network. In social network analysis, connections between individuals in a community or organization are analyzed, e.g., how information travels, or who has the most influence over whom. Examples of applications include identifying key opinion leaders to target for marketing, and identifying bottlenecks in enterprise information flows.

Optimization. A portfolio of numerical techniques used to redesign complex systems and processes to improve their performance according to one or more objective measures (e.g., cost, speed, or reliability). Examples of applications include improving operational processes such as scheduling, routing, and floor layout, and making strategic decisions such as product range strategy, linked investment analysis, and R&D portfolio strategy. Genetic algorithms are an example of an optimization technique.

Pattern recognition. A set of machine learning techniques that assign some sort of output value (or label) to a given input value (or instance) according to a specific algorithm. Classification techniques are an example.

Predictive modeling. A set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome. An example of an application in customer relationship management is the use of predictive models to estimate the likelihood that a customer will “churn” (i.e., change providers) or the likelihood that a customer can be cross-sold another product. Regression is one example of the many predictive modeling techniques.

Regression. A set of statistical techniques to determine how the value of the dependent variable changes when one or more independent variables are modified. Often used for forecasting or prediction. Examples of applications include forecasting sales volumes based on various market and economic variables or determining what measurable manufacturing parameters most influence customer satisfaction. Used for data mining.

Sentiment analysis. Application of natural language processing and other analytic techniques to identify and extract subjective information from source text material. Key aspects of these analyses include identifying the feature, aspect, or product about which a sentiment is being expressed, and determining the type, “polarity” (i.e., positive, negative, or neutral) and the degree and strength of the sentiment. Examples of applications include companies applying sentiment analysis to analyze social media (e.g., blogs, microblogs, and social networks) to determine how different customer segments and stakeholders are reacting to their products and actions.

Signal processing. A set of techniques from electrical engineering and applied mathematics originally developed to analyze discrete and continuous signals, i.e., representations of analog physical quantities (even if represented digitally) such as radio signals, sounds, and images. This category includes techniques from signal detection theory, which quantifies the ability to discern between signal and noise. Sample applications include modeling for time series analysis or implementing data fusion to determine a more precise reading by combining data from a set of less precise data sources (i.e., extracting the signal from the noise).

Spatial analysis. A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, e.g., addresses or latitude/longitude coordinates. Examples of applications include the incorporation of spatial data into spatial regressions (e.g., how is consumer willingness to purchase a product correlated with location?) or simulations (e.g., how would a manufacturing supply chain network perform with sites in different locations?).

Statistics. The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the “null hypothesis”), and what relationships between variables likely result from some kind of underlying causal relationship (i.e., that are “statistically significant”). Statistical techniques are also used to reduce the likelihood of Type I errors (“false positives”) and Type II errors (“false negatives”). An example of an application is A/B testing to determine what types of marketing material will most increase revenue (Ghemawat et al. 2003).

Supervised learning. The set of machine learning techniques that infer a function or relationship from a set of training data. Examples include classification and support vector machines. This is different from unsupervised learning (Cortes and Vapnik 1995).

Simulation. Modeling the behavior of complex systems, often used for forecasting, predicting and scenario planning. Monte Carlo simulations, for example, are a class of algorithms that rely on repeated random sampling, i.e., running thousands of simulations, each based on different assumptions. The result is a histogram that gives a probability distribution of outcomes. One application is assessing the likelihood of meeting financial targets given uncertainties about the success of various initiatives.

Time series analysis. Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data. Examples of time series analysis include the hourly value of a stock market index or the number of patients diagnosed with a given condition every day. Time series forecasting is the use of a model to predict future values of a time series based on known past values of the same or other series. Some of these techniques, e.g., structural modeling, decompose a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data. Examples of applications include forecasting sales figures, or predicting the number of people who will be diagnosed with an infectious disease.

Unsupervised learning. A set of machine learning techniques that finds hidden structure in unlabeled data. Cluster analysis is an example of unsupervised learning (in contrast to supervised learning).

Visualization. Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses.

Big Data Technologies

An ever-increasing number of technologies exist to aggregate, manipulate, manage, and analyze big data. We detail some of the more prominent technologies here, but this list is not exhaustive, especially as more technologies continue to be developed to support big data techniques.

Big Table. Proprietary distributed database system built on the Google File System. Inspiration for HBase.

Business intelligence (BI). A type of application software designed to report, analyze, and present data. BI tools are often used to read data that have been previously stored in a data warehouse or data mart. BI tools can also be used to create standard reports that are generated on a periodic basis, or to display information on real-time management dashboards, i.e., integrated displays of metrics that measure the performance of a system.

Cassandra. An open source (free) database management system designed to handle huge amounts of data on a distributed system. This system was originally developed at Facebook and is now managed as a project of the Apache Software foundation.

Cloud computing. A computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service through a network.

Data mart. Subset of a data warehouse, used to provide data to users usually through business intelligence tools.

Data warehouse. Specialized database optimized for reporting, often used for storing large amounts of structured data. Data is uploaded using ETL (extract, transform, and load) tools from operational data stores, and reports are often generated using business intelligence tools.

Distributed system. Multiple computers, communicating through a network, used to solve a common computational problem. The problem is divided into multiple tasks, each of which is solved by one or more computers working in parallel. Benefits of distributed systems include higher performance at a lower cost (i.e., because a cluster of lower-end computers can be less expensive than a single higher-end computer), higher reliability (i.e., because of a lack of a single point of failure), and more scalability (i.e., because increasing the power of a distributed system can be accomplished by simply adding more nodes rather than completely replacing a central computer).

Dynamo. Proprietary distributed data storage system developed by Amazon.

Extract, transform, and load (ETL). Software tools used to extract data from outside sources, transform them to fit operational needs, and load them into a database or data warehouse.

Google File System. Proprietary distributed file system developed by Google; part of the inspiration for Hadoop.

Hadoop. An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google's MapReduce and Google File System. It was originally developed at Yahoo! and is now managed as a project of the Apache Software Foundation.

HBase. An open source (free), distributed, non-relational database modeled on Google's Big Table. It was originally developed by Powerset and is now managed as a project of the Apache Software foundation as part of the Hadoop.

MapReduce. A software framework introduced by Google for processing huge datasets on certain kinds of problems on a distributed system (Dean & Ghemawat 2004). Also implemented in Hadoop.

Mashup. An application that uses and combines data presentation or functionality from two or more sources to create new services. These applications are often made available on the Web, and frequently use data accessed through open application programming interfaces or from open data sources.

Metadata. Data that describes the content and context of data files, e.g., means of creation, purpose, time and date of creation, and author.

Non-relational database. A database that does not store data in tables (rows and columns). (In contrast to relational database).

R. An open source (free) programming language and software environment for statistical computing and graphics. The R language has become a de facto standard among statisticians for developing statistical software and is widely used for statistical software development and data analysis. R is part of the GNU Project, a collaboration that supports open source projects.

Relational database. A database made up of a collection of tables (relations), i.e., data are stored in rows and columns. Relational database management systems (RDBMS) store a type of structured data. SQL is the most widely used language for managing relational databases (see item below).

Semi-structured data. Data that do not conform to fixed fields but contain tags and other markers to separate data elements. Examples of semi-structured data include XML or HTML-tagged text. Contrast with structured data and unstructured data.

SQL. Originally an acronym for structured query language, SQL is a computer language designed for managing data in relational databases. This technique includes the ability to insert, query, update, and delete data, as well as manage data schema (database structures) and control access to data in the database.

Stream processing. Technologies designed to process large real-time streams of event data. Stream processing enables applications such as algorithmic trading in financial services, RFID event processing applications, fraud detection, process monitoring, and location-based services in telecommunications. Also known as event stream processing.

Structured data. Data that reside in fixed fields. Examples of structured data include relational databases or data in spreadsheets. Contrast with semi-structured data and unstructured data.

Unstructured data. Data that do not reside in fixed fields. Examples include free-form text (e.g., books, articles, body of e-mail messages), untagged audio, image and video data. Contrast with structured data and semi-structured data.

Visualization. Technologies used for creating images, diagrams, or animations to communicate a message that are often used to synthesize the results of big data analyses (see the next section for examples).

APPENDIX B: Connected Vehicle Applications

The Connected Vehicle Reference Implementation Architecture is based on a set of applications that have been defined by various USDOT connected vehicle programs (Iteris 2014). The source for the application descriptions ranges from Concepts of Operations (ConOps), Requirements Specifications, or existing Standards and Architectures (DOT ITS 2014). Not every connected vehicle application conceived by public or private agencies is included here; this list is intended to be representative of the anticipated applications, in that it is the basis for USDOT planning.

Type	Group	Application Name
Environmental	AERIS/Sustainable Travel	Connected Eco-Driving
		Dynamic Eco-Routing
		Eco-Approach and Departure at Signalized Intersections
		Eco-Cooperative Adaptive Cruise Control
		Eco-Integrated Corridor Management Decision Support System
		Eco-Lanes Management
		Eco-Multimodal Real-Time Traveler Information
		Eco-Ramp Metering
		Eco-Smart Parking
		Eco-Speed Harmonization
		Eco-Traffic Signal Timing
		Eco-Transit Signal Priority
		Electric Charging Stations Management
		Low Emissions Zone Management
		Roadside Lighting
	Road Weather	Enhanced Maintenance Decision Support System
		Road Weather Advisories and Warnings for Motorists
		Road Weather Information and Routing Support for Emergency Responders
		Road Weather Information for Freight Carriers
		Road Weather Information for Public Transit
		Road Weather Information for Maintenance and Fleet Management System
		Variable Speed Limits for Weather-Responsive Traffic Management

Mobility	Border	Border Management Systems
	Commercial Vehicle Fleet Operations	Container Security
		Container/Chassis Operating Data
	Commercial Vehicle Roadside Operations	Smart Roadside Initiative
	Freight Advanced Traveler Information Systems	Freight Drayage Optimization
		Freight -Specific Dynamic Travel Planning
	Misc.	Ad Hoc Messages
	Planning & Performance Monitoring	Performance Monitoring and Planning
	Public Safety	Advanced Automatic Crash Notification Relay
		Emergency Communications and Evacuation
		Incident Scene Pre-Arrival Staging Guidance for Emergency Responders
		Incident Scene Work Zone Alerts for Drivers and Workers
	Traffic Network	Cooperative Adaptive Cruise Control
		Queue Warning
		Speed Harmonization
		Vehicle Data for Traffic Operations
	Traffic Signals	Emergency Vehicle Priority
		Freight Signal Priority
		Intelligent Traffic Signal System
		Pedestrian Mobility
		Transit Signal Priority
	Transit	Dynamic Ridesharing
		Dynamic Transit Operations
		Integrated Multimodal Electronic Payment
		Intermittent Bus Lanes
		Route ID for the Visually Impaired
		Smart Park and Ride System
		Transit Connection Protection
		Transit Stop Request
	Traveler Information	Advanced Traveler Information Systems
		Receive Parking Space Availability and Service Information
		Traveler Information Smart Parking

U.S. Department of Transportation
ITS Joint Program Office-HOIT
1200 New Jersey Avenue, SE
Washington, DC 20590

Toll-Free “Help Line” 866-367-7487
www.its.dot.gov

FHWA-JPO-14-157

U.S. Department of Transportation