Leveraging big data: predicting traffic risk and providing early warning due to adverse weather conditions

Sreekumar E. Nampoothiri

University at Albany, State University of New York, sreekumare@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Library and Information Science Commons, Public Policy Commons, and the Urban Studies and Planning Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Dissertation Committee

Catherine T. Lawson, PhD (Chair)
Associate Professor, Department of Geography and Planning, University at Albany

Sue R. Faerman, PhD
Distinguished Teaching Professor Emerita, Rockefeller College of Public Affairs and Policy,
University at Albany

Eliot Rich, PhD
Associate Professor, Department of Information Systems and Business Analytics, School of
Business, University at Albany
ABSTRACT

The proliferation of big data has allowed researchers to delve deeper into data and gain better understandings within almost every field. In the fields of transportation planning and traffic management, past research has shown direct relationships among weather conditions and traffic speed, volume, and congestion. However, these studies have mostly relied on static data that were spatially and temporally sparse or collected on a specific roadway for a specific time period for research purposes. With the need to address the impacts of climate change, including an increasing number of extreme weather events as well as an increasing intensity of such events, big data provides researchers with an opportunity to understand the above relationships in more detail. The transportation sector had been one of the early fields to exploit the potentials of big data by developing applications that can detect the amount and speed of traffic on different roads and providing the information to users and facility managers. This research combines traffic and weather conditions data for three different types of road segments in the Capital Region of New York State to understand the link between these weather conditions and traffic speed that can ultimately be used to provide timely advance warning to travelers. The research demonstrates that big data indeed provide advantages over the simple datasets that are commonly available by allowing for more granular level assessments. In particular, it can provide insights into the most important times of day/year and types of roads for safety interventions. The findings are expected to improve traffic risk prediction during severe weather conditions, which would provide a new approach for early warning and impending deployment of automated vehicles. The research suggests a broad array of implications both for the scholarly literature and for practical applications of this type of analysis.

Key words: automated vehicles, big data, early warning, policy, traffic, weather
ACKNOWLEDGEMENTS

I would not have been able to complete this dissertation without several people. First and foremost, I am thankful to my advisor and Chair of the Dissertation Committee, Prof. Catherine Lawson. She believed in me more than I did. I have the opportunity to study and work under her during my Master’s studies at the University at Albany and that relationship continued even after graduating. The discussions we had over the years led me to the doctoral program in Information Science. The common interests we had led me to work with her and she was able to guide me in the right direction and spent hours discussing various aspects of the topic. She helped me gain access to the right data, which was one of the hindrances in my journey. I am indebted to her for this successful journey. I am eagerly looking forward to continuing to collaborate with her in the future.

I knew Prof. Sue Faerman for a long time and when I wanted to look for someone who can push me to the limits, ensure I complete the research, and produce good quality work, I did not have to think twice to approach her. Even with her busy schedule and the fact that she is a Professor Emerita, she was happy to join the Committee. I cannot thank her enough for forcing me to think harder to see the larger picture, complete the dissertation on time, produce high-quality research, tighten up all loose ends, and learn how to write a dissertation. Thank you for all the encouragement and timely reviews.

When I first started developing the research topic, Prof. Eliot encouraged me to think larger and think multi-disciplinary. Though I slowed down during the journey, he stayed with me and encouraged me to think like a researcher and not get bogged down on the details. Moving forward, that advice is going to be at the top of my mind. His critical comments at various junctures made me think about areas that I probably had at the back of my mind but never
brought out in my writing or analyses. Thank you for helping me making the dissertation more robust.

I am thankful to the late John Poorman who was the Director of the Capital District Transportation Committee (CDTC), my employer when I started my doctoral studies. John not only encouraged my research but provided flexibility at work to attend classes and provided agency funds to partially cover my fees and one of my conference attendances. I thank my then-colleagues at CDTC, Dave Jukins, Chris O’Neill, and Sandra Kinnicutt, for their support. I thank Angel Canales and Brian Menyuk at the New York State Department of Transportation for providing access to data. I thank the wonderful staff at AVAIL, University at Albany, for providing access to the NPMRDS system for collecting data. I am thankful to my current employer, Northern Virginia Transportation Authority (NVTA), and my wonderful colleagues in the larger Washington, DC metro region for making me think of various aspects in my research field.

I am thankful to Prof. Jagdish Gangolly, Prof. Jennifer Goodall, Prof. David Andersen, Prof. Deborah Andersen, Prof. George Richardson, Prof. Karl Rethemeyer, and Prof. Ray Bromley for their wisdom and support throughout the years. I am indebted to Tiffany Williams-Hart for assisting me in administrative matters so efficiently.

I am thankful to my in-laws Dr. Mahesh Chandra Tiwari and Shobha Tiwari to support my family in all possible ways over the years that gave me time and mental peace to carry out the research. I am thankful to my sons, Shreyas and Shourya, for enduring my quirks and absence from several family occasions. I have no words to express my indebtedness to my biggest supporter – my wife Asmita Tiwari. She did not view my journey from the sidelines but
actively ensured I focus on my tasks and never drop out. Just like Prof. Lawson, Asmita believed in me more than I did. Thank you to my best critic, friend, and partner in crime.

Finally, I want to thank my parents, Easwaran Nampoothiri and Syamala Devi, for their strong spirit to make a positive change, even if incremental, to the world and help people around. Without them instilling in me the values and courage to explore the world and areas of interest, I would not have been able to take the path I have traveled so far. The belief they have in me gives me the strength to go on.

The world has changed a lot from when I started my doctoral studies. The last year or so has tipped our world upside down. COVID-19 has made us realize how precarious our lives are in this world. I am grateful that I am able to support my family and complete the dissertation even in such a difficult time. I dedicate this work to the tireless health and essential workers across the world endangering their lives to make our lives go as smoothly as possible. Thank you!
TABLE OF CONTENTS

Abstract iii
Acknowledgements iv

Chapter 1. Introduction 1
  1.1. The Context and Relevance of the Research 1
  1.2. The Research Agenda 6
  1.3. Organization of the Dissertation 7

Chapter 2. Big Data and Data Science 8
  2.1. The Context 8
  2.2. What Is Big Data and Why Is It Important? 12
  2.3. Data Science and Data Analytics 15
  2.4. Big Data Applications 17
  2.5. Advantages and Challenges of Big Data 19
  2.6. Gaps in Research 19
  2.7. End of Theory or Theoretical End? 20
  2.8. Sense-making in the Big Data Context 23
  2.9. A Theoretical Framework of Big Data 26
  2.10. Conclusion 29

Chapter 3. Traffic Characteristics, Weather Conditions, and Big Data in Transportation 30
  3.1. Traffic Characteristics and Weather Conditions 30
  3.2. Methods Used in Traffic Research 35
  3.3. Big Data in Transportation 36
  3.4. Gaps in Research 43
  3.5. Conclusion 44

Chapter 4. Data Systems for Traffic and Weather 46
  4.1. Traffic Data 46
  4.2. Weather Data 51

Chapter 5. Research Questions, Methods, and Data 56
  5.1. Research Questions 56
5.2. Research Methods 56
5.3. Expected Results 57
5.4. Data Characteristics 57
5.5. Conclusion 62
Chapter 6. Analysis and Results 63
  6.1. Analysis and Results 63
Chapter 7. Discussion and Conclusions 74
  7.1. Introduction 74
  7.2. Implications for Literature 74
  7.3. Implications for Practice 80
  7.4. Limitations of the Research 88
  7.5. Directions for Future Research 90
  7.6. Conclusion 93
References 96

Appendices
Appendix 1. Syntax for Regression Analysis 118

List of Figures
Figure 1. Billion-dollar extreme weather events, 2000-2019 1
Figure 2. Projected congestion on National Highway System (NHS) roads in the US in 2020 2
Figure 3. Exponential growth in the amount of data 12
Figure 4. Traditional theory-based analysis 21
Figure 5. Big data theory-less analysis 22
Figure 6. Cynefin framework 25
Figure 7. Cynefin framework for big data environment 28
Figure 8. INRIX traffic maps 37
Figure 9. Mobile apps such as Waze, Moovit, and 511 38
Figure 10. Locations of MIST data-collection loop detectors 49
Figure 11. Mesonet stations in New York State 53
Figure 12. Location of study sites 58
Figure 13. Sample traffic data from counts

List of Tables

Table 1. Big data and traditional data
Table 2. Examples of organizations using big data mindset
Table 3. Research on traffic speed, volume, weather conditions, and incidents
Table 4. Sample traffic data from MIST dataset
Table 5. Sample traffic data from INRIX
Table 6. Sample weather condition data
Table 7. Dummy variables
Table 8. Correlation of speed in INRIX dataset and weather conditions for urban highway
Table 9. Regression results for INRIX data for urban highway
Table 10. Correlation of speed in MIST dataset and weather conditions for urban highway
Table 11. Regression results for MIST data for urban highway
Table 12. Coefficients of the model for INRIX data and MIST data for urban highway
Table 13. Percentage reduction in speeds – all road types; all day
Table 14. Percentage reduction in speeds – all road types; morning peak
Table 15. Percentage reduction in speeds – all road types; mid-day
Table 16. Percentage reduction in speeds – all road types; evening peak
Table 17. Percentage reduction in speeds – all road types; spring
Table 18. Percentage reduction in speeds – all road types; summer
Table 19. Percentage reduction in speeds – all road types; fall
Table 20. Percentage reduction in speeds – all road types; winter
Table 21. Comparison of speed reduction from previous and current research
Chapter 1. Introduction

1.1. The Context and Relevance of the Research

People across the world are being confronted with the negative effects of extreme weather events such as hurricanes, snowstorms, flooding, heat waves, and drought. The situation in United States is not different. Of the 273 billion-dollar weather and climate disasters that have occurred in the United States since 1980, 119 have taken place since 2010, and 14 of these in 2019 alone (NCEI, 2019). The extreme weather events in the United States since 1980 have resulted in more than 14,000 deaths and more than $1.7 trillion in damages (Schmeltz et al., 2015; Kishore et al., 2018; NCEI, 2019). Figure 1 provides a visual picture of the billion-dollar extreme weather events in the United States between 2000 and 2019. Global warming and associated climate change are expected to exacerbate the frequency and magnitude of such extreme events in the future (IPCC, 2014).

Figure 1. Billion-dollar extreme weather events in the United States, 2000-2019
Source: NCEI (2019) and Center for Climate and Energy Solutions (2019)
Traffic congestion and operations are major focus areas in the field of transportation planning and traffic management. The Texas Transportation Institute (TTI), INRIX, and TomTom, among others, have been tracking traffic congestion in the United States and other parts of the world for nearly two decades. According to TTI (2019), travelers in the United States wasted 8.8 billion hours on traffic congestion in 2017, burning an extra 3.3 billion gallons of fuel. TTI estimated the cost of this congestion to be about $166 billion in lost productivity and wasted energy. In addition, they estimate that “The average auto commuter spends 54 hours in congestion and wastes 21 gallons of fuel due to congestion at a cost of $1,010 in wasted time and fuel” (TTI, 2019; p. 1). Figure 2 presents a projection developed by the Federal Highway Administration in 2017 of the expected congestion on US roads in 2020. Research shows that about 15% of the total travel delay is due to weather factors (TRB, 2013). When you consider the increases in extreme weather events and impact of weather on traffic speeds and congestion, it is clear that climate change is going to have a seriously disruptive and costly impact on daily traffic patterns.

Figure 2. Projected congestion on National Highway System (NHS) roads in the US in 2020
Research has established direct relationships among weather conditions, traffic speed, and traffic congestion (Garber & Ehrhart, 2000; Qiu & Nixon, 2008; Strong et al., 2010; Rahman & Lownes, 2012; Andrey et al., 2013; Theofilatos, 2017; Sathiaraj et al., 2018). However, most prior studies were conducted using data collected over a short time period (in most cases, a few weeks) and lack granularity in terms of time of day and season of year. In general, traffic follows patterns associated with such activities as morning and evening commuting and holiday and summer travel. The different characteristics of these patterns—including volume of traffic, drivers’ familiarity with the route and road conditions, drivers’ urgency, etc.—make it important to provide drivers with strategies related to their specific driving conditions. Early warning can improve traffic flow, reduce potential incidents (accidents and disabled vehicles), and improve travel time (FHWA, 2003; Fayish & Jovanis, 2004; Knight et al., 2008).

One way of controlling the freeway traffic is by using Dynamic Speed Limits (DSL). This approach was introduced in the Europe in the 1970’s and 1980’s but has only received full attention in the United States in the past decade or so. DSL aims to provide mandatory as well as advisory traffic speed limits at different locations on a freeway depending on the downstream traffic flow, congestion, and incidents. The advance warning allows drivers to slow down smoothly, which can reduce the amount of stop-and-go traffic. The homogenization of traffic flow is expected to provide congestion relief (Smulders, 1990; Zackor, 1991), emission reduction (Stoelhorst, 2008; Soriguera et al., 2013), and reduction of incidents (Lee et al., 2006; Sultan et al., 2008) by increasing the road capacity and/or avoiding road capacity drop. Empirical research on DSL shows that there is potential for congestion relief, emission reduction, and reduction of incidents (Hegyi et al., 2005; van Nes et al., 2008; Heydecker & Addison, 2011; Strömgren & Lind, 2016), especially when combined with enforcement (Soriguera & Sala, 2014) and other
active traffic management strategies such as ramp metering (Torné et al., 2014). While researchers have focused heavily on congestion, traffic flow, capacity, throughput, lane change, and incidents, very little research has examined the effect of weather in the context of DSL and active traffic management strategies. Since other research shows a strong relationship between traffic and weather, a better understanding of weather impacts is very important to further the research and implementation of DSL strategies.

Transportation planners and traffic system operating agencies understand the importance of the connection between weather and traffic (TRB, 2014). For example, the Federal Highway Administration (FHWA) has been assisting state transportation agencies in developing weather surveying, monitoring, and prediction systems. Road Weather Information Systems (RWIS) consist of Environmental Sensor Stations (ESSs), a communication system for data transfer, and central systems to collect field data. ESSs are roadway stations with one or more fixed sensors measuring atmospheric (air temperature, visibility, precipitation type, etc.), pavement (temperature, condition, etc.) and/or water level conditions (nearby waterbodies, flooding, etc.). The data from RWIS are then used to develop forecasts, and to display or disseminate road weather information to operators (FHWA, 2017b). Many state departments of transportation (DOTs) are using this service for traveler information, speed management, road maintenance, and emergency management (FHWA, 2013). Iteris, Inc. is a private company that provides consulting services and produces sensors and other devices that record and predict traffic and weather conditions. Iteris has been working with transportation agencies to distribute a system of weather information to subscribers through its Transportation Information Center (TIC), a platform that comprises of website and apps (Iteris, 2019). In all the above cases, the focus is on pavement condition, and atmospheric weather conditions take a back seat.
Another fairly recent phenomenon that suggests the need for improved weather information systems is the deployment of connected and automated vehicles (CAVs) across the world. These vehicles need minute-by-minute and location-specific information for efficient operation. In return, they can provide roadside sensing systems with data about vehicle location, speed, and other factors that can help in determining an appropriate control strategy. Simulation studies show that CAVs and traffic management can mutually benefit from sharing data (Kattan et al., 2015; Grumert & Tapani, 2017). Automobile manufacturers have introduced several automated features that fall under the National Highway Traffic Safety Administration’s Level 1 (function-specific automation) through Level 3 (limited self-driving automation) of automation (Booz Allen Hamilton, 2016). Though most of the higher levels of automation beyond Level 3 are currently under pilot testing stages, there is an expectation that several of these functions will be in wide use soon. These features include adaptive cruise control, automatic braking, lane keeping, and parking assist. These functions rely largely on onboard sensors and cameras. Adverse weather conditions such as fog, rain, and snow can severely limit the functionality of these types of equipment. Research in this area is just beginning (Zhao et al., 2019).

Big data can play a significant role in furthering research in the above fields. The traditional ways of collecting data are cumbersome since they require manual or minimally-automated techniques and large amounts of time. This also means that the geographical coverage has been limited and very costly. The arrival of big data has changed the fields of data collection, processing, and research. In the field of transportation, this is amplified by the fact that there is a near-100 percent penetration of GPS-enabled mobile phones and other devices that transmit minute-by-minute location data. This has led to many route planning applications such as Google traffic that provide information about congestion, incidents, and alternate routes,
among others. It is important for researchers to utilize this new big data environment to explore relationships among traffic, incidents, and weather.

There is currently a large push by the Federal Highway Administration to make all transportation planning and infrastructure investment decisions based on data-supported performance measures and evaluation. The last two federal transportation acts, Moving Ahead for Progress in the 21st Century Act (MAP-21) and Fixing America's Surface Transportation Act (FAST), mandated the use of data for developing performance measures (TRB, 2013). As a follow up, the FHWA has made certain “big data” datasets available to planning and operating agencies and research institutions. State DOTs, metropolitan planning organizations (MPOs), and other regional agencies are currently developing performance measures for such variables as congestion, reliability, and safety, among others. As noted earlier, these factors have a direct relationship with weather conditions. FHWA has entered an agreement with data providers such as HERE and INRIX through the National Performance Management Research Data Set (NPMRDS) system to make big data available to the planners and operators such as MPOs and DOTs. A clear understanding of the relationship between traffic and weather is imperative to arrive at appropriate and realistic performance measures, which will be used for making decisions about the investment of scarce resources. In addition, it is important to understand the difference between using traditional data and big data to fully utilize the resources made available by the FHWA.

1.2. The Research Agenda

This research compares and contrasts the results of analyses using traditional and big data in transportation. Using empirical data, this research explores the links among weather and
traffic and aims to fill the research gap associated with granularity of data in terms of time of day and time of year. In order to understand the links, a detailed literature research was conducted. The literature shows that past research has identified links with varying degrees. However, it is clear that there is a need to do more research at a granular level. An inquiry into the emerging field of big data shows that there is a lot of potential to use big data for its volume, velocity, and variety of data available in this area. The research looks at traffic and weather data for three locations in the Capital District of New York. Correlation analyses and Ordinary Least Square (OLS) regression were used as analytical methods.

1.3. Organization of the Dissertation

This dissertation is divided into seven chapters. Chapter 2 provides a review of the big data environment and a theoretical framework to understand the environment. Chapter 3 presents the current literature on traffic and weather characteristics and identifies gaps in research; it also presents the current use of big data in transportation. Chapter 4 provides a review of data systems that are currently available and those in development for traffic and weather conditions. This identifies the best data for this research. Chapter 5 presents the research questions and methods, and details the different characteristics of the data. Chapter 6 describes the analysis and results. The last chapter concludes the report by presenting the implications of the research, limitations, and directions for future research that would extend this research.
Chapter 2. Big Data and Data Science

2.1. The Context

The earliest known collected data are probably the population censuses in Babylonia and Egypt in the 4th century BCE. These included information about the number of people and livestock. The information was probably written down on a few clay tablets or stones. In contrast, it is estimated that by 2025, about 463,000,000,000,000,000,000 bytes (463 exabytes) of digital data will be created each day globally (Desjardins, 2019). The total amount of digital data grew by 200% annually between 2010 and 2018 and it is estimated to grow from 33ZB (Zeta Bytes) in 2018 to 175ZB in 2025 (IDC, 2018). This data explosion, together with a number of other factors, leads us to the often-used term, “big data.” Though everyone talks about big data, people have different meanings for this term, including a large amount of data, an all-encompassing term for complex data that need tools beyond traditional processing applications, the belief that more data will produce more insights and answers, and the attitude that combining data from multiple sources could lead to better decisions. The Internet of Things (IoT)—the digitally interconnected way of life where people, places, and activities are all connected through devices and the internet—accelerated the development of big data. The concept of big data is also related to understanding and applying data and insights from customers, partners and employees, which Frank et al. (2014) call Code Halos™.

Data, Information, and Knowledge

It is important to note the difference between data, information and knowledge. The word “data” is generally defined as observations (Davenport, 1997) and/or a set of discrete facts
(Davenport & Prusak, 1998) that could be analyzed. Data are essential for any understanding of a situation in a scientific manner. Alternatively, information is generally defined as data with relevance and purpose (Davenport, 1997), a message meant to change the receiver’s perception (Davenport & Prusak, 1998), and/or data vested with meaning (Choo et al., 2000). Knowledge is developed by going one step further and is defined as commitments and beliefs created from information (Nonaka & Takeuchi, 1995), the ability to assign meaning (Spek & Spijkervert, 1997), and/or experience, values, insights, and contextual information (Davenport & Prusak, 1998). Knowledge has been further divided into different categories, e.g., tacit vs. explicit (Nonaka, 1994; Choo, 1998); explicit vs. implicit; individual vs. collective (Spender, 1998), and so on.

Scientific research cannot be possible without knowledge and information created out of data. Therefore, for centuries, scientists and researchers collected data in various ways. However, in many cases, it was impossible to collect complete data about the whole population a researcher wanted to understand. For example, if physicians wanted to know more about a particular disease, they could not collect data from every person with the disease (“population”). This meant that the physician needed to collect data from a “sample” or subset of the “population.” Researchers derived statistical methods to collect and analyze data so that they could generate ‘information’ and further develop ‘knowledge’ that could then help them arrive at significantly relevant conclusions using the data from a sample that could be generalized to the population. The need for representative data led to the development of complex statistical methods in defining the population, sample size, sampling methods, extrapolation methods, and so on.
**Big Data**

The advent of computers and computing techniques made it easier for researchers to store and analyze larger datasets. However, the amount of data that was collected and analyzed was still very small and, most times, became irrelevant due to the passing of time or changed physical and social conditions. These datasets were mostly static, which means they were available or useful only for one particular study or use. In more recent times, the amount of data being collected, stored, and analyzed has increased exponentially. This processing of large amounts of data was made possible by rapid advances in data collection technologies, digitization of data, digital storage capabilities, and computing techniques. In addition, now data are available in dynamic form, which means they are available continuously and, in many situations, in real time. In addition, such data are being generated not only in scientific circles but in all aspects of life. Many individuals and businesses have understood the potential values of this data revolution and have used it or are using it to their advantage. Amazon is one of the most frequently mentioned examples in this sense. For example, Amazon began collecting data about the purchase of books (and later other objects) on its website to customize suggestions to its customers based not only on each individual’s preferences but also on other customers’ preferences, choices, and purchase habits.

**Data Science**

The ‘big data’ scenario has led to the development of many tools and techniques to analyze the huge amount of data available. These tools and techniques still use the basics of statistics to examine the relationships between and among different variables, causation of phenomenon, and prediction of potential actions. Some have suggested that the large amount of
data could potentially nullify the need for perfecting “sample” selection as the data are almost equal to the “population” itself (Myer-Schonberger & Cukier, 2013). Although this might be true for certain types of analyses, e.g., where you only need to know if a relationship exists, in other cases, it is still not possible to collect data from the population. In addition, in some situations you still want to understand the causation and statistical relationships do not assure causation. Thus, from a broader perspective, having data may not be enough. You need someone who can look at the larger picture and ask the right questions or bring new ideas that will allow for exploration or analysis of data. Many refer to this structured thought and analysis based on scientific principles as “data science” (Wu, C.F.J., 1997; Cleveland, 2001; Myer-Schonberger & Cukier, 2013). There are many professionals who essentially do the work of a data scientist but might be called other names in their profession, such as financial analysts.

**Data Visualization**

One main aspect of data science and big data is to visualize the data analyses (Unwin et al., 2008). The proverb “A picture is worth 1000 words” describes what is meant here in that visualization is more effective in conveying knowledge than are words. The knowledge that results from data analyses would be nearly inaccessible to even moderate-level experts if it was not visualized. Visualization can also give actionable insights. Thus, visualization can help experts, decision makers, and common people to quickly grasp the complexities and insights (Keim et al., 2013). Data visualization has almost become a profession unto itself. There are many tools and technologies today that allow for better visualization.
2.2. What Is Big Data and Why Is It Important?

In simple terms, big data involves the use of datasets that are large, complex, and difficult to process using traditional data processing methods (Cox & Ellsworth, 1997; Manyika et al., 2011; Snijders et al., 2012). Though this description captures the essence, the size of a dataset is very subjective. The amount of data that is considered small by one organization might be considered large by another that has considerably less capabilities to handle data. Many define big data in terms of three Vs – ‘volume,’ ‘variety,’ and ‘velocity’ (Berman, 2013; Zikopoulos et al., 2013). Volume refers to the amount of all types of data generated from different sources and stored in different formats. As seen in Figure 3, the digitization of almost every aspect of our life has generated enormous amounts of data. It is estimated that there are nearly as many bits of data in the digital universe as stars in our physical universe (IDC, 2014). Variety refers to the multiple types of structured and unstructured data collected through different means. In fact, one of the differences between traditional data and big data could be that traditional data were mostly structured and big data are mostly unstructured. The data could be available in static or dynamic (real time) forms. Again,
big data are mostly dynamic and continuously flow in, making it difficult for static data warehouses to handle. Velocity refers to the speed with which the data are being collected, transferred, and made available for analysis. According to Internet Live Stats (2020), every day there are nearly 400 million tweets sent on Twitter, 45 million photos uploaded on Instagram, 80 million posts made on Tumblr, 220 million calls made on Skype, 2.5 billion posts made on Facebook, 4 billion searches made on Google, 4 billion videos watched on YouTube, and 140 billion emails sent. Moreover, the amount of digital data is doubling every two years (IDC, 2014).

Zikopoulos et al. (2013) added a fourth V, ‘veracity,’ to their earlier typology of data descriptions. Veracity refers to the quality or trustworthiness of the data. The high volume of data from a variety of sources and topics coming at fast velocity results in both useful data and noise. Organizations need to be smart in understanding their needs and to focus on the useful data while removing noise. Even within that noise, there may be information that is useful for someone else. Living with the messiness of big data also means we may have to forgo exactness but embrace approximations that can bring us closer to the reality (Myer-Schonberger & Cukier, 2013). Hashem et al. (2015) also describe big data in terms of four Vs but their fourth V is ‘value,’ rather than veracity. Value refers to the process of discovering huge hidden values from large datasets that develop quickly in a variety of formats and contents with varying veracity. Reuse, recombination, and data extension can also increase the value of a given set of data (Myer-Schonberger & Cukier, 2013). This might be the most important aspect of big data.

Search engines, social media, and crowdsourcing have become huge generators of big data. Google uses its search engine in a variety of ways not only to provide accurate information for users but also to gain invaluable insights. For example, by analyzing the trends in searches
for flu and flu-related words, Google was able to predict geographical locations with high incidences of flu, which could then be used to deploy preventive activities (Ginsberg et al., 2009). The expansion of social media such as Facebook and Twitter has resulted in the generation of millions of pieces of data every minute that has high business value potential. Companies such as Google and INRIX use crowdsourcing from mobile phones and GPS units to collect data about traffic. They aggregate these individual pieces to make it into big data, process these data, and provide accurate information about traffic situations on highways and city streets. The exponential growth of genome data and the sharp drop in sequencing costs, which was made possible by big data, have transformed bioscience and biomedicine into data-driven sciences. Large retailers analyze their sales data linked to individuals through credit card and loyalty card numbers to predict future behavior. A case involving big data collection by Target, the US retailer, received much attention in the past decade. Essentially, Target upset a Minneapolis man by sending pregnancy-related coupons to his daughter. It turned out that Target predicted the girl’s pregnancy correctly by monitoring her shopping patterns and comparing that information to an enormous database detailing billions of dollars of sales.

As seen in Table 1, big data datasets are not just big in size, they also have particular characteristics that separate them from traditional (‘small’) data. Moreover, the concept of big data has a direct relationship with digitization, computing technology, and the internet, as well as with techniques and tools to manage and analyze it (Davenport, 2014). In that sense, the following definition of big data provides a good summary:

“Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale” (Hashem et al., 2015; p. 100).
Table 1. Big data and traditional data

<table>
<thead>
<tr>
<th></th>
<th>Big data</th>
<th>Traditional data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volume</strong></td>
<td>100s of terabytes or petabytes</td>
<td>Gigabytes to terabytes</td>
</tr>
<tr>
<td><strong>Variety</strong></td>
<td>Unstructured formats</td>
<td>Structured formats</td>
</tr>
<tr>
<td><strong>Velocity</strong></td>
<td>Continuous flow of data</td>
<td>Static data, updated periodically</td>
</tr>
<tr>
<td><strong>Veracity</strong></td>
<td>( n = \text{population}; \text{noise} )</td>
<td>( n = \text{sample} )</td>
</tr>
<tr>
<td><strong>Value</strong></td>
<td>Huge potential</td>
<td>Limited applications</td>
</tr>
<tr>
<td><strong>Analytics</strong></td>
<td>Machine learning</td>
<td>Hypothesis-based; statistical</td>
</tr>
</tbody>
</table>

2.3. **Data Science and Data Analytics**

Data science has emerged as one of the hottest new professions and academic disciplines in the first two decades of the 21st century. Many people using the term big data consider data in terms of the Vs discussed above. However, the use of data requires scientific and technical skills in order to deal with the data. Data science could be defined as the study of data for the generalizable extraction of knowledge from data. Data scientists need to employ structured thought and deep scientific analysis in order to create knowledge from data and information (Wu, C.F.J., 1997; Cleveland, 2001; Myer-Schonberger & Cukier, 2013).

One big difference noticed between big data and traditional data analytics is that there is a trend to move away from causational analysis to predictive analysis. Correlation among variables has gained more attention than causation. Many data scientists believe that they can identify patterns from data and predict what a person or company is going to do in certain situations.

“In a big data world, by contrast, we won’t have to be fixated on causality; instead, we can discover patterns and correlations in the data that offer us novel and invaluable insights… Big data is about what, not why” (Myer-Schonberger & Cukier, 2013; p. 13).
Predictive analytics based on pattern identification is used in fields such as marketing, financial services, insurance, and telecommunications, and involves assessing the likelihood that a similar unit in a different sample will exhibit the specific performance. Patterns in purchases of particular items in a store on a particular day or season, for example, could lead to displaying those items in prominent locations, providing coupons for the items, and other strategies to increase sales. Thus, some argue that identifying a correlation is enough to make decisions (Granville, 2014) and that, in these circumstances, understanding the causation is not necessary.

Correlation can tell you if two variables are somehow related or not. The correlation coefficient (r) shows the nature of the relationship between variables (strong or weak, positive or negative). The coefficient of determination (R²) is a measure of how much of the variation is ‘explained’ by the calculated relationship. Correlation alone, however, cannot guarantee that a causal relationship exists between the variables, even when a good correlation is identified between two variables.

“In order to state validly that ‘A is caused by B’, and that ‘an increase in B will cause a change in A’, it is first necessary to demonstrate that A and B do in fact change predictably in relation to one another, and further that the dependency is in fact A on B rather than B on A or on another unknown factor” (Brindle, 1994; p. 120).

A particular phenomenon could be caused by multiple factors. If you want to ‘predict’ any change in the phenomenon, you need to know the factors that cause most of that change. Any one particular factor can demonstrate high correlation with the phenomenon but predicting the future based on just that one factor could lead to erroneous predictions. Predictive analysis is best carried out once you have identified all the factors that cause the phenomenon and can use them in the predictive model based on statistical
methods. In the big data context, it is easy to examine correlations since there is a large amount of data for the different factors. Therefore, it is crucial to ask questions about the assumptions, methodological frameworks, and underlying biases embedded in the analysis of big data (Boyd & Crawford, 2012).

It is important to note that even if a researcher has conducted predictive analyses, causational analysis is not going to be redundant since, in many cases, the understanding about causation leads to value generation in terms of better public policies, business strategies, and/or cost savings. This suggests that there are two distinct styles of data analytics involved: causational analysis and predictive analysis. The first type will lead organizations to pursue big data applications that their business needs, rather than simply using what is available. The second type provides a basic understanding of the opportunities available from a data perspective.

2.4. Big Data Applications

There are many organizations that use big data applications in their business. For example, Google uses big data techniques, machine learning, and analytics to identify appropriate results for people’s searches. Google also uses such analytics to provide targeted advertisements. Amazon and Netflix use big data and data analytics to provide product recommendations based on users’ preferences and previous purchases. GE optimizes its service contracts and maintenance intervals based on a constant flow of data it acquires. As noted above, INRIX and Google provide information about traffic congestion on highways and roads based on crowdsourced data from mobile phones and GPS units. In addition to direct data collection through their apps, they also purchase data from companies (such as GPS providers
and cell phone providers) that hold this type of data. Smart cities in the future could be using numerous big data applications including smart traffic, smart infrastructure, and smart governance (Townsend, 2013). Facebook uses sentiment analysis to target advertisements. It is now a well-accepted fact that foreign powers used such features to influence the 2016 US presidential elections. Many companies like Kayak provide flight fares by comparing a combination of airline companies, routes, and prices. Retail giants Walmart and Target use their sales data for targeted marketing. PriceStats, an innovative product that initially was developed in an MIT class, provides information about inflation across the globe, which many banks and other financial institutions use. The United Nations has a framework called HunchWorks, which uses ‘digital smoke signals’ (social media reveling trends) to decide when to do detailed analyses of problems such as drought. Table 2 provides a handful of illustrations of how thousands of organizations use big data in their internal decisions and external business.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Search engine, advertisement, flu trends</td>
</tr>
<tr>
<td>Amazon</td>
<td>Product recommendations</td>
</tr>
<tr>
<td>Facebook</td>
<td>Sentiment analysis, advertisement</td>
</tr>
<tr>
<td>Netflix</td>
<td>Movie recommendations</td>
</tr>
<tr>
<td>GE</td>
<td>Optimization of service contracts and maintenance intervals</td>
</tr>
<tr>
<td>INRIX</td>
<td>Traffic congestion maps</td>
</tr>
<tr>
<td>Kayak</td>
<td>Flight fares</td>
</tr>
<tr>
<td>Walmart, Target, etc.</td>
<td>Marketing</td>
</tr>
<tr>
<td>PriceStats</td>
<td>Product prices and inflation</td>
</tr>
</tbody>
</table>
2.5. Advantages and Challenges of Big Data

Some advantages of using big data include value (monetary and other), quick insights (multiple iterations), quicker decision-making support, risk management (through tracking in real time), and cost savings. However, using big data also presents some challenges. These include issues regarding privacy (of people, organizations, behavior, communication, etc.), computing technology (need to accelerate abilities), analytical tools (diverse, flexible, and powerful to deal with large amount and type of data), data usage [finding meaning instead of just doing analytics is more important (Roehrig & Pring, 2013)], users (currently there are more statisticians and software programmers than subject matter experts), hype vs. reality (usefulness), and lack of expertise (gap between demand and supply).

2.6. Gaps in Research

It seems that there is no underlying theory that explains or advances the big data environment. Such a theory or framework could help organizations make appropriate strategic changes to utilize big data potentials. Theories can also provide researchers necessary principles for scaling up inference and learning algorithms so that they can be used on a massive scale. Many researchers have already pointed out that there is a lack of theoretical underpinnings in big data (Snijders et al., 2012; Chen et al., 2014). Researchers have tried to develop theories about various components of the big data environment, particularly the computing world (Suciu, 2013; Chen et al., 2014). For example, many theories explain data manipulation as a mathematical or statistical problem. Theories that touch upon the data and information aspects can also be found in many subject matters (e.g. business and health). Another area of the data environment that has
a rich history of theorizing is communication. However, these theories remain in silos and are not well integrated across differing fields.

Integrating seemingly disparate datasets for new insights is one critical area of exploration in the big data environment. Value acquired from comprehensive utilization of multiple datasets is far higher than the sum value of using individual datasets. Data provenance is the process of data generation and evolution over time, particularly when multiple datasets are used. It is very important to understand how to integrate data provenance information featuring different standards used in different datasets (Chen et al., 2014). These multiple datasets might be in different areas of a single organization or across several (or many) different organizations. Understanding the potential of combining these datasets and accessing them is itself a difficult task. How this plays out in the big data environment needs to be explored further (Dhar, 2013; Wladawsky-Berger, 2014).

2.7. **End of Theory or Theoretical End?**

As noted above, many believe that the large volume of big data has rendered statistical sampling and theory-based analysis obsolete (Anderson, 2008; Myer-Schonberger & Cukier, 2013). On the other hand, there are researchers who look at big data as a complex system and suggest that any complex system needs theories to understand it better (Graham, 2012; West, 2013; Bar-Yam, 2015). This tussle between exploratory and hypothesis-based analytics is a central discussion point in big data.
Traditional Theory-based Model

As seen in Figure 4, the first step in traditional scientific research includes observing phenomena and describing and measuring different variables. These measurements are then analyzed to identify patterns, categories, and correlations. This knowledge is then combined with domain knowledge to develop a theory about the phenomena and causation. Once the theory is developed, it can be used to predict what might happen in the future. However, observations might include anomalies that do not conform with the theory. Further refinement of the model will allow for explaining the anomaly or suggesting conditions when anomalies can happen.

Figure 4. Traditional theory-based analysis
Source: Derived from Anderson, 2008; Myer-Schonberger & Cukier, 2013

Big-data Theory-less Model

However, a theory-less model in a big data environment only looks at the observations. It brings out the patterns by observing the data and predicts phenomena through correlations that
have been identified. However, this approach does not explain anomalies (see Figure 5). Google Flu Trends (GFT), the theory-free data-rich model, is an example. In 2009, GFT could predict flu outbreaks in the US much more quickly than the Centers for Disease Control and Prevention (CDC) just by analyzing Google search terms and locations. However, in 2014, GTF prediction overestimated flu outbreaks by a factor of two compared to actual CDC data. The reason could be search by healthy people or Google’s own search algorithms, which categorize search terms in a particular way.

Figure 5. Big data theory-less analysis
Source: Derived from Anderson, 2008; Myer-Schonberger & Cukier, 2013

The basic condition of ‘N = all,’ which is assumed for predictive analytics to become accurate, is not met in most cases. Facebook and Twitter usage data are increasingly being used to predict various trends. However, it should be noted that these users are only a fraction of the entire population. The prediction could be accurate for the user population, but this is only a subset of the entire population. Boston uses a smartphone app called Street Bump to detect
potholes (Harford, 2014). This crowdsourcing method allows for less manual checking of city streets. However, the dataset derived from this app favors the areas frequented by young and affluent users with app-enabled smartphones. Such sampling biases are overlooked in a theory-less environment. This points to an important aspect in the discussion of theory in big data. Theory-less analytics might only be appropriate for certain types of data or only useful in answering certain questions in certain domains. Theory is still critical in other areas. As researchers have pointed out, long-term predictive analytics require statistical models and theories (Siegel, 2013).

There is also a potential for spurious findings with the theory-less model. Since we are dealing with large volumes of data, the chance of finding some relationships among variables is very high. However, some of the relationships identified might be caused by a third variable, which might not have been directly identified in the analysis.

2.8. Sense-making in the Big Data Context

Sense-making is the process by which people give meaning to experience. Researchers have investigated sense-making in the information science context, developing theories underlying the ‘cognitive gap’ that individuals experience when attempting to make sense of observed data (Dervin, 1983, 1992, 1997; Russell et al., 1993). Today, large amounts of static and continuous data are collected through multiple sources and methods. As described earlier, these data need to be contextualized to make sense of their usefulness, thus creating ‘information.’ This information in context is what an analyst or system could examine in order to develop insights. This sense-making is important for individual researchers and organizations to not only understand their subjects’ (customers’) behavior but also to make sense of why they
do what they do. Sense-making brings together related observations to reveal interesting insights. Sense-making in big data involves making observations, making assertions about observations, using new observations to reverse earlier assertions, and drawing on the accumulated context for higher quality relevance detection and selection of a next best action (Jonas, 2012).

Insights or knowledge creation in big data should be looked at from a knowledge management perspective. The Cynefin framework, proposed by Snowden (2000), is a sense-making perspective that can inform us more about the ecosystem of big data.

**Cynefin Framework**

Cynefin, pronounced *kunev-in* (loosely translated as habitat), is a Welsh word that signifies the multiple factors in our environment and our experience that influence us in ways we can never understand (Snowden & Boon, 2007). As seen in Figure 6, the Cynefin framework has five domains—simple (direct cause-and-effect relationship that is visible to all), complicated (knowable but the relationship between cause and effect requires expert analysis), complex (relationship between cause and effect can only be perceived retrospectively), chaotic (cause and effect are not linearly linked), and disorder (not knowing which way of working is best)—reflecting the natural diversity, ambiguity and paradox within human communities. As things change, movement can occur between domains. The ideal direction of movement is from chaos to simple in a clockwise pattern, thus reducing disorder and bringing better understanding and the identification of a causal relationship. However, one should realize that the gradient from simple to chaos is very steep and there is a high cost associated with such a fall. Researchers have used the Cynefin framework to understand processes in their domains as diverse as
software development (Pelrine, 2011), agriculture (Shepherd et al., 2006), and politics (O’Neill, 2004).

The movement between domains occurs naturally over time. The natural drift in a clockwise direction is from the visibly unordered Chaotic, to Complex where the patterns of cause and effect are identified, to Complicated where the patterns of cause and effect are tested for reproducibility, to the visibly ordered Simple domain where the stabilized knowledge of cause and effect is harnessed as known solutions (Kazlauskas & Hasan, 2010a). This happens as people live together, and share mutual concerns and experience. Then, as ideas emerge, “convenience leads to stabilization and ordering of the ideas; tradition solidifies the ideas into ritual” (Kurtz & Snowden 2003; p. 479). Counterclockwise forces also can occur, including obsolescence and forgetfulness; the arrival of new challenges; and the curiosity and energy of new generations or outsiders who break the rules, question the current order of things or validity of established patterns, and thus radically shift the existing power and perspective (Kazlauskas & Hasan, 2010b).
2.9. A Theoretical Framework of Big Data

The big data context is essentially a wicked problem of knowledge management. The Cynefin framework is useful in understanding the different aspects of big data and data science contexts. This understanding is essential to extract the full potential of big data.

The big data context can be understood as the ‘chaotic’ domain in the Cynefin framework. The data explosion (volume, variety, velocity, veracity, value) has placed a tremendous amount of stress on traditional analytic systems that were not capable of handling this type of data. This forced researchers to develop better computing techniques and tools. Most of the discussion in the computing world, as described in an earlier section, tries to deal with this situation. A chaotic scene is impossible to comprehend. Techniques such as dimensionality reduction and distributed optimization have helped researchers to control the high speed and high volume of data, and allowed for analysis of big data. This is how the big data context is being moved from the ‘chaos’ domain to the ‘complex’ domain in the Cynefin framework. This movement is essential in transforming data (facts without context) into information (organized data within context). However, the chaotic situation can generate innovative ideas. Seeing opportunities, even without knowing the details of what and why (or causality), can lead to innovation. This is what Myer-Schonberger and Cukier (2013) called a ‘big data mindset.’

Even after the big data are managed and transferred to the ‘complex’ domain, the causal relationships might not be directly understood. The subject matter knowledge is still not interacting with the information gathered. Thus, theory and theory-based analysis are still not important. However, theory-less predictive analyses such as Amazon book suggestions and Google flu trend could be carried out. Data scientists should still be wary of spurious
relationships or omitted variable bias. Once the establishment of a new knowledge base is completed, information could move towards the ‘complicated’ domain.

In the ‘complicated’ domain, tacit knowledge from previous domains is already converted to explicit knowledge as theories have been developed or refined. This allows for theory-based analyses. Theory-based analysis can convert the information generated from ‘chaos’ and ‘complex’ domains to explicit knowledge (inferences and insights) through various traditional analytical methods. The movement from ‘complicated’ to ‘simple’ occurs through a transfer of established knowledge and insights. Visualization can play a big role in this transfer, especially when non-expert individuals are involved. The ‘simple’ domain is where descriptive statistics are used, and causal relationships are already understood; it is basically a matter of presenting those relationships. The established knowledge of cause and effect are utilized in everyday decision making. The big data environment under the Cynefin framework is summarized in Figure 7 below.
This framework for the big data environment has large implications for future research on big data, computing, organizational studies, and knowledge management, including the development of theories and techniques, policy changes based on correlation vs. causation, developing relationships between data processing and subject matter expertise, and new ways of collecting and managing data.
2.10. Conclusion

Digitization of our lives and advancements in computing power and technology have resulted in the big data phenomenon. What differentiate big data from traditional data are its volume, variety, velocity, veracity, and value. This new situation has sparked the development of new analytical techniques and tools. Data science is a term used to describe the collective domain of data collection, processing, analysis, and visualization. Many companies have already exploited the potentials of this new situation. However, the field of big data is still in its early stages. There are many challenges that researchers, practitioners, and policy makers in this field need to understand and overcome. These include topics such as data privacy, computing technology, analytical tools that can handle big data, data overload, involvement of SMEs, and available expertise to carry out analytics.

Research so far has examined various aspects of the big data phenomenon but there are still gaps. There is no broadly-accepted theoretical framework that can explain the big data context, how different stakeholders fit into it, or the relationships among different aspects of big data. One advantage of the big data environment is the potential for combining seemingly disparate datasets to derive insights. This is an area where research is needed. The impact of the big data context on organizational strategies/changes and data privacy are areas that need further attention. While there is debate surrounding theory-less analytics vs. theory-based analytics, it is important to understand the potentials and pitfalls of relying on correlation without considering causation. Since big data are clearly a complex system, this type of data needs to be understood in terms of complex systems theories and sense-making of the system for knowledge management. In that context, this chapter has provided a theoretical framework for understanding the big data context.
Researchers have identified connections among weather conditions, traffic incidents, traffic speeds, and congestion. This has led to Variable Message Signs (VMS) suggesting upstream drivers to reduce speeds when certain traffic volume/speed conditions are experienced downstream. Research shows that though drivers use caution and reduce speeds during extreme weather conditions, this does not necessarily result in a reduced number of crashes. A summary of these research findings is summarized in section 3.1 and Table 3. A review of the literature shows that prior research mostly used correlation analyses or regression analyses for understanding the relationship between traffic and weather conditions. A summary of other methods used in traffic-related research is summarized in section 3.2.

3.1. Traffic Characteristics and Weather Conditions

Severe Weather Conditions Lead to Reduced Travel Speeds and Volumes

Travel times and speeds are two elements of a transportation system that may be greatly affected by weather conditions (Koetse & Rietveld, 2009). Research also shows that both travel time increases and travel speed declines (although not in a linear fashion) with increasing intensity of snow and rain (Tsapakis et al., 2013). This suggests that the impact of rain and snow is a function of their intensity. Travel time increases sharply with heavy rain, light snow, and heavy snow, respectively, in that order. Although the estimates amongst studies are difficult to
compare in magnitude, the impact of rain and especially snow on traffic speed at congested links during rush hours appears to be significant (Knapp et al., 2000; Sabir et al., 2008; Theofilatos, 2017; Sathiaraj et al., 2018).

**Increased Speed Leads to Increased Crashes**

Research from as early as late the 1950s shows that increased speeds result in increased numbers of motor incidents (Haney & Weber, 1974; Aarts & van Schagen, 2006). The research also suggests that the relationship between pre-incident speed and incident involvement rate is a U-shaped curve. Speeds significantly higher or lower than that of the traffic stream have often been cited as associated with high incident rates. Studies have also found that speed variance decreases as the average speed increases (Garber & Gadiraju, 1989; Garber & Ehrhart, 2000). This could mean that not only increased speeds but vehicles moving at significantly different speeds from the average speed limits can cause incidents.

**Severe Weather Conditions Lead to Increased Crashes**

Driving occurs in a wide range of weather conditions including sunlight, rain, snow, and heat. Similarly, drivers encounter multiple conditions that influence the type of traffic they will experience such as congestion, road type, and speed of other drivers. Research suggests that drivers adapt to conditions in different ways though such adaptations are mostly insufficient to offset the additional risk posed by adverse weather conditions (Edwards, 2002; Maze et al., 2006; Unrau & Andrey, 2006; Kilpeläinen & Summala, 2007; Strong et al., 2010; Rahman & Lownes, 2012). The result is an increased rate of incidents under adverse weather conditions. Research clearly indicates that incident rates go up in severe weather conditions such as rain and
snow compared to clear conditions (Knapp et al., 2000; Khattak & Knapp, 2001; Keay & Simmonds, 2005; Qiu & Nixon, 2008; Andrey et al., 2013).

However, drivers do not necessarily become acclimatized to local weather patterns, which underscores the need to look at driver adaptations on shorter time scales with a view to identifying situations or driver groups where risks are particularly elevated (Andrey et al., 2013). In other words, even drivers who are frequently exposed to adverse weather conditions do not necessarily adapt more effectively to these situations than drivers who are occasionally confronted with similar conditions.

**Increased Volume/Flow (Congestion) Does Not Necessarily Lead to Increased Crashes**

Some studies have indicated that a decrease in crash rates is associated with an increase in volume (Hall & Pendleton, 1990; Garber & Subramanyan, 2001). Similar to the relationship between speed and crashes, a U-shaped curve characterizes the relationship between traffic volume and crashes (Hakkert et al., 1996). This indicates that during low volume periods, such as those that occur during the early morning and late day hours, higher crash rates are observed than when the volume of traffic is greater. This might be attributed to the increased speed during low volume conditions and the increased crash rates during high speed travel. Research also tells us that crash rates are not dependent solely on any single factor, but on the complex interaction of multiple factors such as speed, volume, road type (highway vs. urban street), and geometric alignment such as turns and slopes. (Garber & Ehrhart, 2000).
Table 3. Research on traffic speed, volume, weather conditions, and incidents

<table>
<thead>
<tr>
<th>Source</th>
<th>Methods</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haney &amp; Weber (1974); Garber &amp; Gadiraju (1989); Garber &amp; Ehrhart (2000); Xu et al. (2013a); Yu &amp; Abel-Aty (2014); Theofilatos (2017)</td>
<td>Correlation; Multiple linear regression; Bayesian binary probit; Bayesian logistic.</td>
<td>The greater the vehicle speed deviation from the average speed, the greater the probability of crash. No or low R^2 values are reported.</td>
</tr>
<tr>
<td>Garber, N.J. &amp; Subramanyan, S. (2001); Xu et al. (2013a)</td>
<td>Correlation, multiple linear regression; Bayesian logistic regression.</td>
<td>Peak incident rates do not occur during peak flows.</td>
</tr>
<tr>
<td>Smith et al. (2004); Maze et al. (2006); Hranac et al. (2006); Wang et al. (2006); FHWA (2010); Tsapakis et al. (2013); Angel et al. (2014); Dehman &amp; Drakopoulos (2017); NHTSA (2018)</td>
<td>Means and percentages; Correlation; ANOVA; Multiple linear regression.</td>
<td>Light rain causes 2-10% reduction in speed. No or low R^2 values are reported.</td>
</tr>
<tr>
<td>Stern et al. (2003); Hranac et al (2006); FHWA (2010); Sathiaraj et al. (2018)</td>
<td>Means and percentages; Multiple linear regression</td>
<td>Heavy rain causes 4-20% reduction in speed.</td>
</tr>
<tr>
<td>Ibrahim &amp; Hall (1994); Kyte et al. (2001); Hranac et al. (2006); FHWA (2010); Tsapakis et al. (2013); Sathiaraj et al. (2018)</td>
<td>Descriptive statistics; Correlation; Multiple linear regression.</td>
<td>Light snow causes 5-15% reduction in speed. No or low R^2 values are reported.</td>
</tr>
<tr>
<td>Agarwal et al. (2005); Maze et al. (2006); Hranac et al. (2006); FHWA (2010); NHTSA (2018); Sathiaraj et al. (2018)</td>
<td>Means and percentages.</td>
<td>Heavy snow causes about 10-20% reduction in speed.</td>
</tr>
<tr>
<td>Knapp et al. (2000); Khattak and Knapp (2001); Qiu &amp; Nixon (2008); Andrey (2010); NHTSA (2018); Sathiaraj et al. (2018)</td>
<td>Matched-pair analysis; Poisson regression model; Binary logit model; Multiple linear regression.</td>
<td>Crash rates increased by 11 to 21 times during snow events. No R^2 values are reported.</td>
</tr>
<tr>
<td>Keay &amp; Simmonds (2005)</td>
<td>Matched-pair analysis</td>
<td>Length of time since the last rainfall increased the crash risk under wet conditions.</td>
</tr>
<tr>
<td>Keay &amp; Simmonds (2005); Qiu &amp; Nixon (2008)</td>
<td>Matched-pair analysis; Meta-analysis</td>
<td>Crash risk under wet conditions was larger than that under clear weather.</td>
</tr>
<tr>
<td>Edwards (2002); Maze et al. (2006); Unrau &amp; Andrey (2006); Kilpeläinen &amp; Summalä (2007); Strong et al. (2010); Rahman &amp; Lownes (2012)</td>
<td>Means and percentages; ANOVA; Logistic regression.</td>
<td>Usual driver adaptations are insufficient to offset the additional risk posed by adverse weather conditions.</td>
</tr>
<tr>
<td>Andrey et al. (2013)</td>
<td>Risk analysis; Descriptive statistics.</td>
<td>Drivers do not necessarily become acclimatized to local weather patterns.</td>
</tr>
</tbody>
</table>

Length of time since the last rainfall increased the crash risk under wet conditions. Crash risk under wet conditions was larger than that under clear weather. Usual driver adaptations are insufficient to offset the additional risk posed by adverse weather conditions. Drivers do not necessarily become acclimatized to local weather patterns.
Crash Prediction Models Linking Flow Characteristics and Weather Conditions Are Sparse

Researchers have only recently started to develop real-time crash prediction models. These models focus on estimating the likelihood of crash occurrence based on traffic flow characteristics such as volume, speed, standard deviation of speed, large speed differences between upstream and downstream, compression waves leading to abrupt changes in traffic flow, queue formation, and geometry (Oh et al., 2001; Hourdos et al., 2006; Golob et al., 2008; Christoforou et al., 2011; Hossain & Muromachi, 2011; Ahmed et al., 2012; Xu et al., 2013b; Li et al., 2014).

However, fewer studies have considered the impact of traffic flow characteristics on crash likelihood under different weather conditions (Golob & Recker, 2004; Ahmed et al., 2012; Xu et al., 2013a; Sathiaraj et al., 2018). The studies that have examined the impact of traffic flow characteristics on crash likelihood under different weather conditions have focused on rainy conditions. Findings from these studies, however, have not been consistent, and traffic flow characteristics contributing to crash risk were found to be different across different weather conditions.

Impact of Transition Weather Conditions on Crashes and Traffic Flow Are Not Studied

It is clear that the impact of traffic conditions on crashes, the impact of weather conditions on traffic flow as well as crashes, and prediction models for these situations are being studied and models are being developed. It is also documented that travelers take precautions during adverse weather conditions. However, it is reasonable to believe that the traffic conditions and crash probabilities can be affected by transition conditions of weather. A
transition period is one where weather conditions change from one type to another (e.g., thunderstorm, short period of low visibility on a clear day, etc.). This transition period could be short or long depending on the local weather patterns. Interestingly, transition conditions could lead to unsafe travel conditions that are more dangerous than actual severe weather due to the fact that travelers have not yet become aware of the situation and so have not adjusted their behavior.

3.2. Methods Used in Traffic Research

As noted in Table 3 above, most of the previous research has used regression analysis to study the relationships among traffic speed, incidents, and weather. One of the advantages of big data datasets is that they lend themselves to exploratory analysis. Two often-used methods for exploratory analyses are cluster analysis and factor analysis. In the transportation field, the research utilizing these methods include classification of road types based on traffic characteristics and pollutant emissions (Chen et al., 2008; Velázquez-Martínez et al., 2016), roads with different levels of congestion (Xu et al., 2012), road and pedestrian incident locations (Kim & Yamashita, 2007; Kumar & Toshniwal, 2016), cities by road congestion (Bian et al., 2016), traffic speed groups (Yang et al., 2017; Dogan et al., 2018), and travel time clusters for signal timing at intersections (Deb Nath et al., 2010), among others. Yuan et al. (2017) utilized cluster analysis to classify 23 factors influencing road incidents into 6 groups. Further, they utilized factor analysis to identify the relative importance of these groups. This study, however, only examined 166 samples. Though the data used by some researchers (e.g., Bian et al., 2016) could be considered big data, most had small sample sizes. Moreover, these studies focused on identifying patterns so that they can use those patterns for further deductive analyses such as
regression. In most of the above research, researchers used factor analysis for data reduction – combining several variables into a single factor so fewer variables can be entered into a regression equation (Bian et al., 2016; Yuan et al.; 2017); and cluster analysis for identifying patterns among the different cases such as cities with different congestion levels (Bian et al., 2016; Kumar & Toshniwal, 2016; Chen et al., 2018).

3.3. Big Data in Transportation

As noted above, the transportation sector had been one of the early adopters in using the potentials of big data. There are many data portals and apps compiling traffic and transit service data from different sources, including crowdsourced data, and providing composite pictures to travelers.

Early Uses

Early uses of big data in transportation were limited to providing static maps, but soon moved to GPS-based maps. As seen in Figure 8, GPS-based maps allowed people to generate route directions and travel without worrying about the details of routes and distances. However, this still did not include any information about the traffic conditions such as congestion and traffic incidents along the routes. In the late 2000s and early 2010s, public agencies, primarily departments of transportation, and private entities, such as INRIX and Google, started compiling traffic data from sensors embedded in the road and EZ-pass counters. This provided some information about the level of congestion. However, it was the crowdsourced speed data from standalone GPS units and/or cell phone GPS that showed the true potential of big data. Based on this information, Google maps and other apps were able to provide true assessments of traffic
conditions on roads and started suggesting alternate routes that would be faster. In addition to
direct data collection through their apps, these companies purchase data from other companies
(such as GPS providers and cell phone providers) that hold the data. There is, however, still a
delay in processing the data and bringing the true information to the user, which can be resolved
with faster communication and processing technologies.

Figure 8. INRIX traffic maps, as presented on smart phone, can give users detailed information
about current traffic conditions on roads.
Source: Author’s screen captures from INRIX App.

Recent Uses

Recently, there has been a flood of smart phone apps that provide traffic information such
as crowdsourced congestion and incident status, route choices, mode choices, General Transit
Feed Specification (GTFS) providing transit information, trip time reliability matrices, freight
door-to-door delivery times, pedestrian safety, real-time dashboards, trip origin-destination,
smart parking, variable road pricing, etc. Examples of such apps include Waze, Moovit, Metro
and Bus, 511, and iRide. There are also a multitude of apps created by transit agencies themselves. Some examples of big data application are provided below in Figure 9.

Figure 9. Mobile apps such as Waze, Moovit, and 511 have been pushing a large amount of traffic information to users; the source of data includes crowdsourced data from users themselves.

Source: Author’s screen captures from Waze, Moovit, 511, and MyTransitNYC Apps

Mobile phone penetration has been very high, even in underdeveloped countries. This suggests numerous opportunities to use mobile data for various services. Origin-Destination (OD) survey is essentially used for identifying a matrix of trips between different zones of a region (referred to as an OD matrix) and is used by transportation planners. Traditionally, getting this type of data in the context of an urban area has involved a difficult, expensive (millions of dollars), and time-consuming (years) survey process, and subsequent calibration using count data at strategic points on major roadways and transit routes was needed. The technical capacity required to achieve success is tremendous and, many times, quality control has been an issue. From their experience in Brazil’s Rio state, Mehndiratta and Alvim (2014) caution about the use of big data from mobile phones. They suggest that this type of data does
not provide insightful information on the underlying choices and motivations regarding individuals’ decisions, which are critical to designing a viable new investment. Even though the sample size is very large, other concerns include the fact that there are some elements of systemic self-selection and that segments that do not use mobile technology are left out. Also, in some places, data are available from only one mobile carrier and, if there are systematic preferences across different segments of the population for carriers, then those preferences could affect the accuracy of the final results.

Increasing air travel has resulted in growing airspace congestion. Brazil has recently introduced a system that harnesses GPS data to optimize the use of available airspace, enabling less separation between aircraft and shorter routes. The usual practice has been to line up planes preparing to land in an airborne queue. Under the new system, each plane is assigned its own flight path (Neumann, 2015). The distance, speed, and capabilities of each aircraft are processed in a way that results in the shortest flight path. Instead of the traditional way of queuing up on approach, planes can “curve in” much closer to the airport. Without the amount of different data, this would be impossible. The Rio de Janeiro Operations Centre uses mobile applications to warn citizens about heavy rain, strong wind, fog, energy shortages, traffic signal malfunctions, mudslides, fire, smoke and points of flooding. It also receives information from the public. On an average day, Rio’s transportation planners receive aggregated views from 110,000 drivers and reports on 60,000 traffic incidents (OECD/ITF, 2015). The crowdsourced data are overlaid with real-time information from various sensors and cameras and pushed out through apps such as Waze and Moovit. In the future, the city plans to start monitoring how cyclists move around the city using the cycling app Strava (Olsen, 2014).
As noted by Vlahogianni et al. (2015; p. 161),

“Big Data has been rapidly expanding into the transportation arena. However, the methods, models and algorithms that are used today in our domain to mine and explore data – think of estimation, prediction, validation of traffic and transportation theories and models – may not scale and/or perform well under these new conditions.”

This suggests that there is a need to rethink existing theories and models. Current research examines activity data collected from mobile phones’ call detail records, smart card systems, automatic passenger count systems, GPS, smart phone and vehicle location services, bike-sharing, and social media data. The researchers and practitioners then present big data applications such as travel demand estimation, non-work destination choice, transit travel experience, origin–destination estimation by trip purpose and time of day, willingness to travel by activity types, and traffic zoning (Wang et al., 2010; Jia et al., 2017; Li et al., 2018). Another set of research examines big data sources from GPS, Bluetooth readers, and loop detectors for traffic flow prediction, travel time prediction, and addressing GPS data requirements (Sunderrajan et al., 2016; Bertsimas et al., 2019). Current research also examines video, microwave vehicle detection system, GPS, and vehicle trajectory data to develop applications on proactive road safety analysis, traffic operations and safety monitoring, and calibration of traffic simulation model for safety assessment (Li et al., 2020).

Using GPS data collected from all taxis operating in the Chinese city of Harbin in 2013, Cui et al. (2016) developed a model that generated a set of indicators for measuring road transport performance between regions, and subsequently pinpointed the areas with serious
mismatch problems. Liu et al. (2016) looked at truck safety at rail crossings using location and traffic data to develop a method for benchmarking behaviors and evaluating control device improvements. Toole et al. (2015) developed a flexible, modular, and computationally efficient software system to integrate raw, massive data into estimates of travel demand and infrastructure performance. This system estimates multiple aspects of travel demand using call detail records from mobile phones in conjunction with open and crowdsourced geospatial data, census records, and surveys. It also brings together numerous existing and new algorithms to generate representative origin–destination matrices, directs trips through road networks constructed using open and crowdsourced data repositories, and performs analytics on the system’s output. This is then presented as an online, interactive visualization platform to communicate the results to researchers, policy makers, and the public.

DOTs in the United States are actively pushing for more apps to be developed in-house or by private entities through open data portals. Globally, institutions such as the World Bank are making the argument for keeping all data open so as to encourage creative solutions from private entities and individuals. Hackathons, competitions to create technology solutions within a short time span, are becoming increasingly popular. With the idea of smart cities currently coming to the fore, numerous big data applications including smart traffic, smart infrastructure, and smart governance could soon be the norm (Townsend, 2013). Smart cities are being instrumented with digital devices and infrastructure that produce ‘big data.’ Such data, smart city advocates argue, enable real-time analyses of city life and new modes of urban governance, and provide the raw material for envisioning and enacting more efficient, sustainable, competitive, productive, open and transparent cities (Kitchin, 2014). Smart cities will require embedded devices, such as sensors, actuators, and smartphones, leading to considerable business potential.
for the new era of the Internet of Things, in which all devices are capable of interconnecting and communicating with each other over the Internet. This integrated system will consist of smart home sensors, vehicular networking, weather and water sensors, smart parking sensors, and surveillance objects. The system implementation will consist of several different steps, beginning with data generation and moving to data collection, aggregation, filtration, classification, preprocessing, analysis and decision making. Rathore et al. (2016) propose such a system implemented using Hadoop with Spark, voltDB, Storm or S4 for real-time processing of the IoT data to generate results to establish the smart city. The advent of connected vehicles will bring much of this information into the forefront of transportation.

In any of these cases, one can see that the data being used focus primarily on traffic. The ideal benefit of the big data environment is beyond one type of data; it is the opportunity to combine seemingly disparate datasets such as traffic, weather, incidents, fare, etc. This has been found lacking in the current environment. Transport infrastructure involves complex networks with many participants. Moreover, the various players collect their own data and do not necessarily want to share their data, generally for privacy or business reasons. However, the essential benefits of understanding the traffic patterns would be useful for all. While sharing information, different players do not always have the same goals. Someone might want to reduce traffic at certain locations to reduce congestion, while others might want to increase traffic to generate business. Collectively finding a solution that makes every stakeholder a winner is not a simple task and requires a certain level of mutual trust that cannot be assumed. “Infrastructure in many cases is a natural monopoly. Governments therefore have an important role to play—in ensuring that operations are fair and cost-effective, and in creating a regulatory environment that allows data to be collected and used while protecting confidentiality and privacy” (Neuman, 2015).
3.4. Gaps in Research

In this era of big data and technological advancements in data collection/management and infrastructure, as well as the extensive application of Intelligent Transportation Systems (ITS), researchers can look at traffic and weather data by specific geography and conditions. While older studies used traffic data collected manually, mostly in 15-minute intervals, data are available minute by minute. Currently, there is a lack of research on the variation of impacts on more granular time scales. This is important since transportation planning and traffic management research and practice show that the traffic pattern, needs, and behaviors are different during different times of the day such as early morning, morning peak, mid-day, evening peak, and night. Big data provide the possibility of looking at these granular details.

Data are now available for all road segments throughout the day (through Bluetooth, cell phones and GPS devices), while it was previously based on specific locations at specific times (manual counts or loop detectors on the pavement). Most studies thus far have focused on freeways due to data availability and challenges on urban streets. Urban street segments are shorter in length, have lower operating speed, and experience more interrupted traffic due to driveways, pedestrians, and traffic signals. These factors make it difficult to pinpoint the reduction in speed in urban streets resulting from any one factor such as the weather condition (Tsapakis et al., 2013). GPS-based big data, however, make it possible to collect data on arterials and local roads with much more accuracy and for the whole day. This makes it possible to analyze the traffic patterns on arterials and local roads, which was not possible with traditional data.

Similarly, recent meteorological studies that examine weather satellite and radar technologies have substantially improved the data granularity (Heinselman & Torres, 2011;
Sutherland-Stacey et al., 2011). However, the utilization of such weather data in the transportation field remains somewhat limited. The potential of such types of analysis—involving a combination of datasets at such granular details—needs to be studied further.

Research shows that though severe weather conditions heavily impact traffic patterns, advance knowledge and precautionary actions can reduce the negative effects of severe weather conditions to some extent. However, there are no studies that look at the transition time between good weather conditions and severe weather conditions. The research thus far has not differentiated the transition time between severe conditions and between fair conditions and severe conditions, and the way travelers behave during these transitions. Further research should be conducted along these lines since it could be assumed that such transition times could be the most dangerous to travel safety. This could be the most critical situation for using advance notification that can reduce the impacts of these transitions, which suggests that we need better weather forecasting at a more geographically localized level and at shorter time intervals than is currently available since weather conditions can vary drastically within short distances and times. In other words, safe conditions can change to unsafe conditions within a short trip, suggesting the need for better weather forecasts and models.

3.5. Conclusion

Though the existing research has identified connections among weather conditions, traffic incidents, traffic speeds, and congestion, there is still a need to understand these at a more granular scale such as time of day and season of the year. Increased availability of big data has provided an opportunity to conduct such detailed analyses. A combination of traffic information and weather information can help transportation planners and operators manage traffic in a more
efficient and safe manner. This has become even more important with the introduction of autonomous vehicles, which are expected to be on roads within the next five years in many parts of the country.
Chapter 4. Data Systems for Traffic and Weather

There are several traffic detection systems and weather monitoring and prediction systems being used across the United States. There are pros and cons for each system. In order to use these data for predicting traffic conditions, the data should provide real-time high-resolution information in terms of frequency (up to minute-by-minute) and coverage (localized). The most important traffic and weather, as well as traffic detection, systems used today are described below.

4.1. Traffic Data

The traditional way of collecting and reporting traffic data is through traffic counts. Earlier ways of manually counting traffic at intersections or road segments during certain hours of the day reported traffic every 15 minutes during two- or three-hour peak periods such as morning and evening rush hours. Later, automatic traffic counters replaced manual counting for segment counts. These were sensors laid down on roads that counted vehicles by the vehicle type (auto vs. truck). These counts provided only the number of vehicles that passed through. Thus, these systems were useful for analyzing the amount of traffic that passed through that location during a particular timeframe but they did not provide any information regarding speed or other characteristics.

Recent technological developments have allowed for technologies that collect and disseminate traffic volume, speed, and other data on roads in near real time. These mostly include data collected and aggregated through mobile phones and GPS systems. A comparative analysis shows the pros and cons of different technologies.
INRIX Data

One of the most promising traffic data systems is provided by INRIX. INRIX is a global Software as a Service (SaaS) and Data as a Service (DaaS) company that provides real-time traffic information along with other types of transportation information. INRIX provides average travel time every 5 minutes, 24 hours a day, for the entire year. INRIX gathers real-time, predictive and historical data from more than 300 million sources, including local transportation authorities, commercial fleets, GPS, cell towers, INRIX mobile apps and cameras. The data cover National Highway System (NHS) roads and some additional critical connecting roads. INRIX data are direct observation data and not “imputed” data, meaning that you are getting observed traffic speeds and there may be times when there are no data. INRIX data and insights help transportation agencies in operations/system monitoring, work zone monitoring, incident detection/queue monitoring, bottleneck analyses, and congestion alerts.

National Performance Management Research Data Set (NPMRDS)

The National Performance Management Research Data Set (NPMRDS) provides average travel time every 5 minutes, 24 hours a day, for the entire year. This dataset is provided by the Federal Highway Administration to the state departments of transportation, metropolitan transportation organizations (MPOs), and their contractors for use under a non-disclosure agreement (FHWA, 2017c). NPMRDS provides vehicle probe-based travel time data for passenger autos and trucks. The contractor for providing the data was HERE Technologies from 2014 to 2016. From 2017 onward, the contractor is INRIX. The real-time probe data are collected from a variety of sources including mobile devices, connected autos, portable navigation devices, commercial fleets and sensors. NPMRDS includes historical average travel
times in five-minute increments on a daily basis for the National Highway System. The data are provided in two parts. The first part is a Traffic Message Channel (TMC) static file that contains data that do not change frequently. The second part includes travel times and identifies roadways geo-referenced to TMC location codes. The two datasets need to be joined through GIS-based software to provide the full picture.

The NPMRDS does not use imputed data; “null” records are included in the NPMRDS when probes are not present during a given five-minute period, which are referred to as epochs. The NPMRDS is not a real-time dataset. The NPMRDS is updated each month. It should be noted, however, that the data collection methods were sufficiently different between HERE and INRIX that comparison of the two datasets—2014-2016 and 2017 onwards—would be problematic. INRIX does have earlier data and additional coverage available for sale outside the NPMRDS contract.

**TomTom**

The TomTom dataset uses location information from its GPS users to derive speed data on highways and major arterials (TomTom, 2017). It provides minute-by-minute data. TomTom uses congestion as an interruption to traffic flow that a) affects at least a 1 km road stretch, b) results in vehicles exhibiting speeds lower than 56% of normal free-flow speed, and c) causes a delay of more than 90 seconds to a road user at that moment in time.

**Management Information System for Transportation (MIST)**

The New York State Department of Transportation (NYSDOT) collects traffic data through the Management Information System for Transportation (MIST) for a number of
locations on I-87 and I-90 as well as one location each on I-787 and Alternate Route 7 in New York State’s Capital Region (CDTC, 2007). This system provides traffic counts and speeds for every 15-minute interval throughout the year. In addition, this system also provides incident data (time of reporting to clearance time). This system was implemented in the Capital Region in 2000 and collects data from loops embedded in the pavement. As shown in Figure 10 below, the geographical coverage is very low, i.e., there are only 19 sites in the Capital Region. Also, electrical failures of the loop detectors lead to missing data at several times periods, days and/or lanes.

Figure 10. Locations of MIST data-collection loop detectors.
Source: CDTC (2007)
TRANSMIT

TRANSMIT is a system used for collecting traffic information from various vehicle sensor sources including EZ-Pass readers. It captures information such as aggregate data on average speeds, travel times, and the number of non-arriving vehicles (vehicles expected but not yet detected by the next reader downstream). This allows for efficient management of incidents and traffic.

Discussion

There are several data systems that provide various types of information about traffic flow including speed, volume, travel time, and delays. Some of these are actual measurements from the field while others are imputed based on historic trends or other methods. In general, all systems provide data for highways. As you move to arterial roads, however, the data availability differs from system to system and, even for systems that provide data, this is usually from a smaller sample. The systems differ in the time intervals they provide as well. While some datasets provide data at all speeds, others only provide data from roads that have speed limits above certain values. TMC segments also differ from dataset to dataset in terms of the definitions of distance, end points, etc. The format in which the data are made available also differ. Each dataset provides a different historic time frame. There are certain data gaps in all datasets. These factors could hinder easy compilation and comparison if one wants to combine the datasets. In general, the INRIX data appear to be the most used and the most robust in terms of time interval, data availability, and ease of manipulability. All these datasets are commercially available but the cost for an individual researcher might be high. The INRIX and NPMRDS data are available through special permission and agreements for researchers and would be useful.
4.2. Weather Data

There are several weather reporting systems in the United States. They differ in information provided, geographic coverage, and time intervals at which the data are updated. A comparative analysis shows the pros and cons of different technologies.

Automated Surface Observing System (ASOS) and Automated Weather Observation System (AWOS)

There are several types of automated weather stations in the United States. Two primary ones are the Automated Surface Observing System (ASOS) and the Automated Weather Observation System (AWOS). ASOS was developed as a joint effort between the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the Department of Defense (DOD). ASOS sensors provide hourly data on precipitation, light conditions, wind, pressure, and temperature at almost 1000 locations nationwide. Occasionally, ASOS provides more frequent data, especially if there is a big change at certain thresholds. AWOS units are mostly operated by FAA, although some state/local governments also operate a small number of these units. AWOS units can have a wide range of configurations but, in general, observe similar parameters as ASOS. The data collected through these systems are disseminated over radio frequency, internet, and automated phone service. The system uses state-of-the-art technology and can make fairly accurate observations, which are provided around-the-clock.

Though ASOS/AWOS are the primary weather detecting and reporting systems used in the United States, they have certain limitations. These systems are unable to report a variety of conditions such as patchy fog, smoke, volcanic eruptions, hail, multiple forms of precipitation at the same time, etc. In addition, the detectors can only detect conditions directly overhead, not
around in the horizon. For example, detectors cannot identify a storm moving into the area until it is directly on top of the area. For these reasons, NWS also manages a network of human observers to provide supplementary information.

There are only two ASOS/AWOS stations in the Capital Region. They are located at the Albany International Airport in Albany County and the Schenectady County Airport in Schenectady County. As one can easily understand, this means the coverage and accuracy of weather detection and prediction is very broad and location-specific variation cannot be accurately identified. Since ASOS primarily serves the aviation community, the sensors are physically located at airports. This means that they may not represent the best locations for the collection of climate data (or the best climate observing sites).

**New York State Mesonet**

The Mesonet was designed and developed by research scientists at the University at Albany-SUNY’s Atmospheric Sciences Research Center and Department of Atmospheric and Environmental Sciences and through collaboration with the Oklahoma Mesonet (NYSMESONET, 2017). Mesonet is a particular type of observing network, with weather stations spaced closely enough together to adequately sample “mesoscale” weather, which refers to weather phenomena that range in size from less than a mile to hundreds of miles long and last a few minutes to hours.

As seen in Figure 11, this system is still in the implementation stage with a final goal of about 125 monitoring stations, at least one station in each county of New York State. Of those stations, 17 will be enhanced with state-of-the-art vertical profiling instrumentation to capture vertical profiles of temperature, relative humidity, and “3D” wind speed as well as an accurate
measurement of boundary layer height and cloud base height. These vertical observations will, by themselves, provide a very detailed, three-dimensional “image” of the atmosphere at very high spatial and temporal resolutions. In addition, enhanced products will be created from these data and used as input to high-resolution Numerical Weather Prediction (NWP) models to improve severe weather forecasting.

New York State Mesonet will collect measurements of a number of surface and atmospheric variables, such as temperature, relative humidity, wind speed and direction, surface pressure, soil moisture, soil temperature, solar radiation, and precipitation amounts for rainfall and snow accumulation. Mesonet is expected to provide better data more frequently (5-minute intervals) in real time with an average spacing of 30 miles.

Figure 11. Mesonet stations in New York State
Source: NYSMESONET, 2017
There are seven New York State Mesonet locations in the four-county capital region. These are:

- Voorheesville in Albany County,
- Schaghticok, Schodack, and Stephentown in Rensselaer County,
- Schuylerville and Ballston Spa in Saratoga County, and
- Duanesburg in Schenectady County.

**Modern-Era Retrospective Analysis for Research and Applications (MERRA)**

MERRA has been hosted by Goddard Space Flight Center’s Global Modeling and Assimilation Office from its inception in 2008. MERRA integrates data from a variety of satellite systems into numerical models to recreate a synthetic data record of the weather. MERRA provides a reanalysis of discrete data, combining model fields with observations distributed irregularly in space and time into a spatially complete gridded meteorological dataset. About 120 observation types, including satellites, radiosondes, aircraft, balloons, ship and ocean buoys, and land surface observations, provide input for MERRA. Using advanced modeling techniques, the system then computes the probable weather patterns across the globe and can be accessed at regional scales for either a static point in time or a specified duration. Studies have observed that while the system is good for historic trend analyses and global forecasts, its use for real-time or near real-time observations at local levels is minimal (Schubert et al., 1993; Bosilovich et al., 2011; Reichle et al., 2011; Rienecker et al., 2011; Schubert et al., 2011).
Discussion

As mentioned earlier, for predicting traffic conditions, the ideal weather data will provide real-time, high-frequency, micro-level information. However, most of the data systems are only available at the meso level and are too geographically sparse. In addition, the data are not available in real time and at high frequency in terms of time intervals. This restricts the ability to predict real-time traffic conditions based on weather conditions. The accuracy of prediction models is also highly questionable. With these caveats, any analysis could use any one of these datasets since they all have some pros and some cons. The easy availability of ASOS/AWOS data makes it useful for researchers.

Combining the Datasets

The two types of datasets (weather and traffic) together provide a comprehensive picture of what is happening on our roadways. To date, not many researchers and practitioners have explored the possibility of combining these seemingly disparate datasets to generate potentially insightful inferences. From a preliminary look of the datasets, it is clear that these datasets could be combined using the date and time stamps provided. By combining the datasets, the data can be analyzed, and insights can be found that can help maintain smooth traffic flow without traffic incidents and disturbances before, during, and after severe weather events.
Chapter 5. Research Questions, Methods, and Data

Based on an understanding of big data and transportation, research questions and methods were developed and data were collected.

5.1. Research Questions

1. What are the relationships among traffic speed and weather conditions on different road types such as highways and arterials? How different are the results between the analyses using traditional and big data?

2. What are the relationships among traffic speed and weather conditions at different times of the day? How different are the results between the analyses using traditional and big data?

3. What are the relationships among traffic speed and weather conditions during different seasons such as winter and summer? How different are the results between the analyses using traditional and big data?

5.2. Research Methods

Based on the literature review and research questions, correlation analyses and Ordinary Least Squares (OLS) regression are used for this research. The outcome variable is the speed. Predictive variables are the current weather conditions. The weather conditions include the categories of fair, partly cloudy, cloudy, mostly cloudy, light rain, rain, fog, light snow, snow, and heavy snow. The ordinary least square model used is:

\[ \text{Speed} = \beta_0 + \beta_1 \text{Fair} + \beta_2 \text{Partly cloudy} + \ldots + \beta_n \text{Heavy snow} \]
As described in the next chapter, this analysis is done separately for the three locations at different timeframes and during different seasons. The analysis is carried out for four different timeframes – morning peak (7am to 9am), evening peak (4pm to 6pm), off peak (9am to 4pm), and all day. This analysis is expected to give different results for the different timeframes selected. Seasonal variations can be captured by slicing the data into four seasons and conducting separate analyses. For the three study sites, morning peak travel direction is southbound on highways and eastbound on the arterial and the reverse is true for the evening peak.

5.3. Expected Results

The analysis is expected to confirm that speeds go down from the average during adverse weather conditions, which would confirm existing research. This is expected to vary to different degrees as we change the timeframes, seasons, and location. The research is also expected to show that granularity in analysis is possible, thus confirming the value of big data. These new results will be evaluated to identify implications for transportation planning and operations professionals.

5.4. Data Characteristics

Data have been collected for three locations in the Capital Region of New York for the time period between April 3, 2017 and March 29, 2018. These locations provide a variety of traffic intensity, commute patterns, and geographical spread. More details are provided below.
Geography

The research looks at traffic and weather data for three locations in the Capital District of New York as shown on Figure 12. These sites are selected based on the level of development of the location (urban vs. rural), amount of traffic experienced (high, medium, low), and the nature of the road (highway vs arterial). These locations are:

1. Site 1: I-87 northbound between exits 4 and 5 (Town of Colonie). This location is a highway surrounded by a highly developed area (dense population and many businesses) with both morning and evening commuters, high daily local traffic, and high levels of reported congestion.

2. Site 2: I-87 northbound between exits 11 and 12 (Town of Malta). This location is a highway in a more rural area with less development, fewer daily commuters and less local traffic. It tends to have higher traffic speeds than Site 1, and low levels of reported congestion.

3. Site 3: Route 20 westbound between I-87 and Rapp Road/Crossgates Mall (Town of Guilderland). This location is an arterial road with a medium level of development and a medium level of commuting/daily traffic. The traffic speeds tend to be relatively low, and there is a fair amount of congestion during commuting hours.
Traffic Data

As seen in Figure 13, simple datasets include traffic counts collected on road segments that do not provide direct measurements of speed or travel time. One can assume the probable speeds based on engineering manuals. However, these are predicted speeds, not observed speeds. In addition, the count data were only collected for five to seven days per location. The frequency of data collection is once in five years at best. The latest data available for Site 1 is for June 2014, for Site 2 is May 2007, and for Site 3 is March 2009. Therefore, this dataset was not used for analyses.

Figure 13. Sample traffic data from counts
Source: NYSDOT (2019a)

The second set of data that was used for this research is the Management Information System for Transportation (MIST) data, which provides speeds, volume, and vehicle occupancy for traffic at select locations at 15-minute intervals (see Table 4 below). This dataset was collected through loop detectors embedded in the road pavement with separate detectors in each
lane; it is therefore more useful than the traffic count-based predicted data. This dataset is available for select locations including the first two locations (Sites 1 and 2) mentioned above. The geographical coverage, time interval, and missing data (due to electrical failure of the loops) make MIST data less valuable than the mobile-based data. However, this dataset is used for analysis at Site 1 and Site 2. MIST data can be accessed from New York State Department of Transportation.

Table 4. Sample traffic data from MIST dataset

<table>
<thead>
<tr>
<th>LINK_ID</th>
<th>DT</th>
<th>TIME</th>
<th>VOLUME</th>
<th>OCCUPANCY</th>
<th>SPEED</th>
</tr>
</thead>
<tbody>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>12:15:00 PM</td>
<td>3880</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>12:30:00 PM</td>
<td>3868</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>12:45:00 PM</td>
<td>3932</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>1:00:00 PM</td>
<td>3752</td>
<td>7</td>
<td>53</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>1:15:00 PM</td>
<td>4168</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>1:30:00 PM</td>
<td>4212</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>8005001</td>
<td>3/2/2017</td>
<td>1:45:00 PM</td>
<td>4044</td>
<td>7</td>
<td>54</td>
</tr>
</tbody>
</table>

Source: NYSDOT (2019b)

As seen in Table 5 below, mobile device-based data for the same time period were collected from the NPMRDS portal using INRIX data developed and maintained by the University at Albany Visualization and Informatics Lab (AVAIL) for the NYSDOT and the New York State Association of Metropolitan Planning Organizations (NYSAMPO). The data provide travel time by road segment at five-minute intervals on various highways and arterials. Each segment is called a TMC and the five-minute timeframes are called epochs. The travel time in seconds can be converted to speed (miles per hour) by dividing the length by travel time. These data are available for all three sites considered in this research.

Speed (mph) = Length / (Travel Time in sec / (60 * 60))
As seen in Table 6, the archived ASOS/AWOS data for the Albany International Airport site can be accessed from the National Weather Service (NWS). This provides mostly hourly weather conditions but provides more granularity as and when conditions change. However, it should be noted that the data are not available at five-minute intervals as the traffic data are. This means the data need to be interpolated to match the traffic data. This must be understood as a limitation of the analysis. The conditions recorded are fair, partly cloudy, cloudy, light rain, rain, fog, light snow, snow, and heavy snow.

Table 6. Sample weather condition data

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Temp</th>
<th>Due Pt</th>
<th>Hum</th>
<th>Wind</th>
<th>Wind Speed</th>
<th>Wind Gust</th>
<th>Pressure</th>
<th>Precipitation</th>
<th>Preci Accu</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/3/2017</td>
<td>4:51 AM</td>
<td>33 F</td>
<td>27 F</td>
<td>78%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
</tr>
<tr>
<td>5:51 AM</td>
<td>33 F</td>
<td>78%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>6:51 AM</td>
<td>33 F</td>
<td>78%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>7:51 AM</td>
<td>36 F</td>
<td>79%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>8:51 AM</td>
<td>41 F</td>
<td>67%</td>
<td>SW</td>
<td>3 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>9:51 AM</td>
<td>45 F</td>
<td>58%</td>
<td>S</td>
<td>3 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>10:51 AM</td>
<td>49 F</td>
<td>50%</td>
<td>VAR</td>
<td>3 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>11:51 AM</td>
<td>53 F</td>
<td>43%</td>
<td>VAR</td>
<td>3 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>12:51 PM</td>
<td>56 F</td>
<td>40%</td>
<td>ESE</td>
<td>3 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.8 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>1:51 PM</td>
<td>57 F</td>
<td>39%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.7 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
<tr>
<td>2:51 PM</td>
<td>58 F</td>
<td>37%</td>
<td>CALM</td>
<td>0 mph</td>
<td>0 mph</td>
<td>0 mph</td>
<td>29.7 in</td>
<td>0.0 in</td>
<td>0.0 in</td>
<td>Cloudy</td>
<td></td>
</tr>
</tbody>
</table>

The data tables were organized to have time periods (5-minute or 15-minute interval) as rows and traffic speed and weather conditions as columns. Each time period is associated with one of the above weather conditions. In each row, the weather condition associated with that particular time period was given a value of one and the rest of the conditions were given values of zeros. See a sample data on Table 7 below.

Table 7. Dummy variables

<table>
<thead>
<tr>
<th>LINK_ID</th>
<th>Date</th>
<th>Time</th>
<th>Speed_mph</th>
<th>Fair</th>
<th>Partly Cloudy</th>
<th>Cloudy</th>
<th>Mostly Cloudy</th>
<th>Light Rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light Snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>8005001</td>
<td>4/3/2017</td>
<td>6:00:00 AM</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8005001</td>
<td>4/3/2017</td>
<td>6:15:00 AM</td>
<td>56</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8005001</td>
<td>4/3/2017</td>
<td>6:30:00 AM</td>
<td>56</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5.5. Conclusion

Correlation and regression analyses were selected as analytical methods to evaluate data and explore the research questions. Three sites that represent different levels of development, amount of traffic experienced, and the road type were selected. Data from two sources (MIST and INRIX) are used for two sites while due to non-availability of MIST data, only INRIX data is used for the third site. Weather data from NWS are used. Weather data are available only at one-hour intervals while other datasets are available at a more granular level. Thus, the same weather data for the hour are matched with each smaller interval of traffic data within that hour.
Chapter 6. Analysis and Results

As stated earlier, this research analyzes the data by using a linear regression model with traffic speed as the outcome variable and current weather conditions as predictive variables. This analysis was carried out separately for the three locations, different timeframes, and seasons using INRIX data and MIST data. The model employed is:

\[
\text{Speed} = \beta_0 + \beta_1 \text{ Fair} + \beta_2 \text{ Partly cloudy} + \ldots + \beta_n \text{ Heavy snow}
\]

Incident data were accessed via the NPMRDS tool website and was made available from TRANSCOM through a non-disclosure agreement. The NPMRDS tool matches the time stamps from incident data with those from the INRIX data. A quick look at the incident data shows that there were about 70 traffic incidents within the one-year analysis period in the segment analyzed. However, only 15 of them were found to have caused any significant delays. Therefore, the analysis did not separate out data for those timeframes.

6.1. Analyses and Results

As noted, the INRIX data are available in five-minute intervals, while the weather data are only available hourly, and so the weather conditions for each of the five-minute intervals within the hour are assumed to be the same for the entire hour. These data were used for correlation and regression analyses for the one-year time period. The regression model was developed by adding weather conditions to the model one by one. In the end, it turned out that certain conditions reduced the significance of the models. Therefore, the model was revised by adding one condition at a time but removing the latest condition if that condition reduced the
significance of the model or made other conditions lose their significance. The results are shown below in Tables 8 and 9.
Table 8. Correlation of speed in INRIX dataset and weather conditions for urban highway

<table>
<thead>
<tr>
<th>Speed (mph)</th>
<th>Fair</th>
<th>Partly cloudy</th>
<th>Cloudy</th>
<th>Mostly cloudy</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>.025**</td>
<td>.012*</td>
<td>.035**</td>
<td>-0.005</td>
<td>-0.078**</td>
<td>-0.048**</td>
<td>.064**</td>
<td>-0.027**</td>
<td>-0.083**</td>
<td>-0.068**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.000</td>
<td>0.035</td>
<td>0.000</td>
<td>0.348</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
<td>29952</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 9. Regression results for INRIX data for urban highway

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Constant</td>
<td>58.204</td>
<td>0.075</td>
<td></td>
<td>779.097</td>
<td>0.000</td>
</tr>
<tr>
<td>Light rain</td>
<td>-3.536</td>
<td>0.230</td>
<td>-0.087</td>
<td>-15.348</td>
<td>0.000</td>
</tr>
<tr>
<td>Rain</td>
<td>-7.865</td>
<td>0.863</td>
<td>-0.052</td>
<td>-9.111</td>
<td>0.000</td>
</tr>
<tr>
<td>Light snow</td>
<td>-1.823</td>
<td>0.260</td>
<td>-0.041</td>
<td>-7.018</td>
<td>0.000</td>
</tr>
<tr>
<td>Snow</td>
<td>-12.738</td>
<td>0.840</td>
<td>-0.089</td>
<td>-15.161</td>
<td>0.000</td>
</tr>
<tr>
<td>Heavy snow</td>
<td>-18.539</td>
<td>1.502</td>
<td>-0.071</td>
<td>-12.344</td>
<td>0.000</td>
</tr>
</tbody>
</table>

R squared: 0.023
F Statistic: 115.904
The results show that during heavy snow and snow, speeds come down considerably. During rain, speeds also come down considerably but to a lesser extent. Light rain and light snow result in minor reduction in speeds. These findings are similar to findings from prior research.

In this model, the coefficient for fog has a positive value, implying that speed goes up during fog conditions. This was not expected. A possible explanation for this result might be related to the time of day when fog is present (early morning) and the low traffic volume during that time frame. In order to test this, a dummy variable for early morning time (6am to 10am) was created. In fact, 84% of the fog conditions fell during this time frame. This is also the morning peak travel time, which means the traffic volume is relatively high; however, the location analyzed was I-87 Northbound between exits 2 and 4, which is in the off-peak direction, and so the volumes are expected to be low, potentially causing drivers to be less cautious.

It should be noted that the $R^2$ value is low for the model (0.023), and the variables considered predict only 2.3% of the variation in speed. This is understandable since there are other major factors such as traffic volume and geometry (flat and straight road) that are not considered in the model. However, the research goal was not to predict all the outcomes “precisely”; rather, it was to explore the relationship among the variables considered. Therefore, the model is still useful even with the low $R^2$ value. Moreover, the literature review showed that most researchers have reported similarly low $R^2$ values. Further, researchers in other fields, especially in public health, have confirmed the importance of low $R^2$ values, especially in situations where there are many other known causes (Kennedy, 2008; Filho et al., 2011; Hamilton et al., 2015; Kuwashima et al., 2015; Poitras et al., 2015; Weber et al., 2015). This is also consistent with the general experience with big data and data analytics where correlation is used more than causation.
The MIST data are available in 15-minute intervals. As with the analysis of the INRIX data, the weather data are available hourly, and so the weather conditions for the hour are considered the same for each 15-minute interval within that hour. These data were used for correlation and regression analyses. The regression model was developed by adding weather conditions to the model one by one. As was the case for the analysis of the INRIX data, certain conditions reduced the significance of the models. Therefore, the model was revised by adding one condition at a time but removing the latest condition if that condition reduced the significance of the model or made other conditions lose their significance. The conditions considered in the final model included light rain, rain, fog, light snow, snow, and heavy snow. The results of the analysis are shown below in Tables 10 and 11.
Table 10. Correlation of speed in MIST dataset and weather conditions for urban highway

<table>
<thead>
<tr>
<th>Speed (mph)</th>
<th>Fair</th>
<th>Partly cloudy</th>
<th>Cloudy</th>
<th>Mostly cloudy</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>0.011</td>
<td>0.015</td>
<td>.054**</td>
<td>-0.002</td>
<td>-.079**</td>
<td>-.063**</td>
<td>.091**</td>
<td>-.052**</td>
<td>-.114**</td>
<td>-0.086**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.277</td>
<td>0.140</td>
<td>0.000</td>
<td>0.848</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
<td>9773</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 11. Regression results for MIST data for urban highway

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
<th>R squared: 0.042</th>
<th>F statistic: 71.299</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>51.544</td>
<td>0.087</td>
<td></td>
<td>590.996</td>
<td>0.000</td>
<td>51.373</td>
<td>51.715</td>
</tr>
<tr>
<td>Light rain</td>
<td>-2.542</td>
<td>0.303</td>
<td>-0.084</td>
<td>-8.399</td>
<td>0.000</td>
<td>-3.135</td>
<td>-1.948</td>
</tr>
<tr>
<td>Rain</td>
<td>-7.523</td>
<td>1.132</td>
<td>-0.066</td>
<td>-6.647</td>
<td>0.000</td>
<td>-9.742</td>
<td>-5.305</td>
</tr>
<tr>
<td>Light snow</td>
<td>-2.002</td>
<td>0.341</td>
<td>-0.058</td>
<td>-5.867</td>
<td>0.000</td>
<td>-2.670</td>
<td>-1.333</td>
</tr>
<tr>
<td>Snow</td>
<td>-13.116</td>
<td>1.120</td>
<td>-0.116</td>
<td>-11.707</td>
<td>0.000</td>
<td>-15.312</td>
<td>-10.920</td>
</tr>
<tr>
<td>Heavy snow</td>
<td>-17.294</td>
<td>1.957</td>
<td>-0.088</td>
<td>-8.839</td>
<td>0.000</td>
<td>-21.130</td>
<td>-13.459</td>
</tr>
</tbody>
</table>
The results show that during heavy snow and snow, speeds come down considerably. During rain, the speeds also come down a great deal, but not as much as they do during heavy snow and snow. Light rain and light snow result in minor reduction in speeds. These findings are similar to the findings from previous research. Similar to the INRIX data analysis, a positive correlation is identified for fog conditions. Additionally, a low $R^2$ value (0.042) was observed in this analysis as well, i.e., only 4.2% of the variation can be explained with weather conditions. As explained earlier, the model is still useful for the purposes of the research.

A comparison of the results from the analysis using MIST data and the analysis using INRIX data shows that both datasets are very similar. While the INRIX dataset shows higher beta values for three weather conditions, the MIST dataset shows higher beta values for the remaining three weather conditions (see Table 12 below).

Table 12. Coefficients of the model for INRIX data and MIST data for urban highway

<table>
<thead>
<tr>
<th></th>
<th>INRIX</th>
<th></th>
<th>MIST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B$</td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Constant</td>
<td>57.873</td>
<td>57.745</td>
<td>58.002</td>
<td>51.544</td>
</tr>
<tr>
<td>Light rain</td>
<td>-3.205</td>
<td>-3.651</td>
<td>-2.760</td>
<td>-2.542</td>
</tr>
<tr>
<td>Fog</td>
<td>3.222</td>
<td>2.575</td>
<td>3.869</td>
<td>3.557</td>
</tr>
<tr>
<td>Light snow</td>
<td>-1.493</td>
<td>-1.996</td>
<td>-0.989</td>
<td>-2.002</td>
</tr>
</tbody>
</table>

Similar analyses were conducted for the other two sites to understand the relationships on rural highways and arterial roads. Note that the arterial location had only INRIX data and therefore, only one set of additional analyses was conducted for that location.
In addition, analyses were conducted to understand the speed and weather impacts during specific times of the day for the same three sites. Transportation planning practice focuses on peak travel time periods. The data were thus separated to differentiate among morning peak, evening peak, and mid-day hours. Morning peak (7 a.m.-9 a.m.) and evening peak (4 p.m.-6 p.m.) are the peak commuting periods and the high number of vehicles causes congestion and low speeds. However, the mid-day (11 a.m.-1 p.m.) traffic is expected to have lower volumes and higher speeds. These peak times were derived based on traffic counts from NYSDOT.

Similar to time periods within a day, seasons also are expected to show differences in the impact of weather conditions on speeds. Therefore, similar analyses were conducted for the four seasons (winter, spring, summer, fall). Since Albany has long winters, MitigateNY and NOAA Storm Events database were used to check for winter storms in the study area for the study time period. This verification pointed out that during the analysis periods, snowstorms were observed in December, January, February, and March. In order to take this into consideration, these four months were considered as winter months. April and May were considered as spring months, June, July, and August as summer months, and September, October, and November as fall months.

The results from the regression analyses were used to estimate speeds during adverse weather conditions. The coefficients were added/subtracted to/from the average speeds during fair weather. The percentage change in speeds were calculated and are presented below in Table 13 through Table 20.
Table 13. Percentage reduction in speeds during different weather conditions on different types of roads for the whole day

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-6%</td>
<td>-14%</td>
<td>-3%</td>
<td>-22%</td>
<td>-32%</td>
<td>0.023</td>
<td>18,950</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-6%</td>
<td>-15%</td>
<td>-5%</td>
<td>-26%</td>
<td>-34%</td>
<td>0.037</td>
<td>18,950</td>
<td>52</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-1%</td>
<td>-3%</td>
<td>-1%</td>
<td>-4%</td>
<td>-23%</td>
<td>0.107</td>
<td>14,000</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-3%</td>
<td>-6%</td>
<td>-4%</td>
<td>-28%</td>
<td>-8%</td>
<td>0.087</td>
<td>14,000</td>
<td>62</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-2%</td>
<td>-5%</td>
<td>-17%</td>
<td></td>
<td></td>
<td>0.002</td>
<td>2,150</td>
<td>25</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)

Table 14. Percentage reduction in speeds during different weather conditions on different types of roads during the morning peak period

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-5%</td>
<td>-4%</td>
<td></td>
<td></td>
<td>0.044</td>
<td>15,593</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-2%</td>
<td>-6%</td>
<td>-6%</td>
<td></td>
<td></td>
<td>0.289</td>
<td>15,593</td>
<td>55</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-1%</td>
<td>-5%</td>
<td>-3%</td>
<td></td>
<td></td>
<td>0.025</td>
<td>11,074</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-2%</td>
<td>-4%</td>
<td>-3%</td>
<td></td>
<td></td>
<td>0.066</td>
<td>11,074</td>
<td>63</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-2%</td>
<td>-31%</td>
<td>-6%</td>
<td>-2%</td>
<td></td>
<td>0.007</td>
<td>3,670</td>
<td>24</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)

Table 15. Percentage reduction in speeds during different weather conditions on different types of roads during the mid-day period

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-1%</td>
<td>-2%</td>
<td>-30%</td>
<td></td>
<td>0.080</td>
<td>17,721</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-2%</td>
<td>-1%</td>
<td>-1%</td>
<td>-33%</td>
<td></td>
<td>0.146</td>
<td>17,721</td>
<td>54</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-1%</td>
<td>-1%</td>
<td>-1%</td>
<td>-2%</td>
<td>-28%</td>
<td>0.243</td>
<td>12,175</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-2%</td>
<td>-2%</td>
<td>-2%</td>
<td>-36%</td>
<td></td>
<td>0.240</td>
<td>12,175</td>
<td>63</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td></td>
<td>-2%</td>
<td>-2%</td>
<td>-16%</td>
<td></td>
<td>0.004</td>
<td>3,145</td>
<td>24</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)
Table 16. Percentage reduction in speeds during different weather conditions on different types of roads during the evening peak period

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-12%</td>
<td>-15%</td>
<td>-9%</td>
<td>-16%</td>
<td>0.028</td>
<td>28,018</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-14%</td>
<td>-17%</td>
<td>-5%</td>
<td>-10%</td>
<td>-16%</td>
<td>0.029</td>
<td>28,018</td>
<td>39</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-5%</td>
<td>-3%</td>
<td>-9%</td>
<td>-33%</td>
<td>0.106</td>
<td>22,449</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-5%</td>
<td>-7%</td>
<td>-2%</td>
<td>-12%</td>
<td>-23%</td>
<td>0.070</td>
<td>22,449</td>
<td>59</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-5%</td>
<td></td>
<td>-4%</td>
<td>-10%</td>
<td></td>
<td>0.003</td>
<td>4,502</td>
<td>23</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)

Table 17. Percentage reduction in speeds during different weather conditions on different types of roads during spring

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-6%</td>
<td>-18%</td>
<td>-2%</td>
<td>-3%</td>
<td></td>
<td>0.028</td>
<td>18,838</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-6%</td>
<td>-18%</td>
<td></td>
<td></td>
<td></td>
<td>0.041</td>
<td>18,838</td>
<td>52</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-1%</td>
<td>-1%</td>
<td>-2%</td>
<td></td>
<td></td>
<td>0.005</td>
<td>14,150</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-2%</td>
<td>-7%</td>
<td></td>
<td></td>
<td></td>
<td>0.043</td>
<td>14,150</td>
<td>63</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>N/A</td>
<td>24</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)

Table 18. Percentage reduction in speeds during different weather conditions on different types of roads during summer

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>R squared</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-6%</td>
<td>-19%</td>
<td></td>
<td></td>
<td></td>
<td>0.010</td>
<td>19,208</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-5%</td>
<td>-23%</td>
<td></td>
<td></td>
<td></td>
<td>0.013</td>
<td>19,208</td>
<td>51</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-1%</td>
<td>-6%</td>
<td>-1%</td>
<td></td>
<td></td>
<td>0.010</td>
<td>15,062</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>-3%</td>
<td>-5%</td>
<td></td>
<td></td>
<td></td>
<td>0.012</td>
<td>15,062</td>
<td>61</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-5%</td>
<td>-7%</td>
<td></td>
<td></td>
<td></td>
<td>0.003</td>
<td>N/A</td>
<td>24</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)
### Table 19. Percentage reduction in speeds during different weather conditions on different types of roads during fall

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>$R^2$</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-4%</td>
<td></td>
<td></td>
<td></td>
<td>0.002</td>
<td>22,077</td>
<td>56</td>
</tr>
<tr>
<td>MIST</td>
<td></td>
<td>-3%</td>
<td>-13%</td>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
<td>22,077</td>
<td>50</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-1%</td>
<td>-1%</td>
<td></td>
<td></td>
<td>0.009</td>
<td>14,155</td>
<td>70</td>
</tr>
<tr>
<td>MIST</td>
<td></td>
<td>-6%</td>
<td>-11%</td>
<td></td>
<td></td>
<td></td>
<td>0.056</td>
<td>N/A</td>
<td>62</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td></td>
<td>23</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)

### Table 20. Percentage reduction in speeds during different weather conditions on different types of roads during winter

<table>
<thead>
<tr>
<th>Source</th>
<th>Light rain</th>
<th>Rain</th>
<th>Fog</th>
<th>Light snow</th>
<th>Snow</th>
<th>Heavy Snow</th>
<th>$R^2$</th>
<th>Avg volume/hr.</th>
<th>Avg fair weather speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban highway</td>
<td>INRIX</td>
<td>-10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.079</td>
<td>17,589</td>
<td>60</td>
</tr>
<tr>
<td>MIST</td>
<td></td>
<td>-9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.104</td>
<td>17,589</td>
<td>53</td>
</tr>
<tr>
<td>Rural highway</td>
<td>INRIX</td>
<td>-2%</td>
<td>-4%</td>
<td></td>
<td></td>
<td></td>
<td>0.190</td>
<td>13,213</td>
<td>67</td>
</tr>
<tr>
<td>MIST</td>
<td></td>
<td>-2.9%</td>
<td>-2%</td>
<td></td>
<td></td>
<td></td>
<td>0.056</td>
<td>13,213</td>
<td>62</td>
</tr>
<tr>
<td>Arterial</td>
<td>INRIX</td>
<td>-4%</td>
<td>-18%</td>
<td></td>
<td></td>
<td></td>
<td>0.007</td>
<td>N/A</td>
<td>25</td>
</tr>
</tbody>
</table>

Note: For items in bold, the correlation between that weather condition and speed is significant at the 0.01 level (2-tailed)
Chapter 7. Discussion and Conclusions

7.1. Introduction

Findings from this research show that big data provide advantages over traditional data and combining datasets from different fields can provide richer findings. The big data datasets can provide deeper understandings of the relationship between weather conditions and traffic speed. The findings of prior research that showed the levels of impact different weather conditions have on traffic speed were confirmed. In addition, the impacts were further refined by road type, time of day, and season. In general, speeds on urban highways are more impacted than are speeds on rural highways and arterials. While snow affects speed at all times, rain affects speeds during morning and evening peak periods more than other times. Analyses based on different seasons show that during non-winter months, rain is the major cause of delay and urban highways are the most affected. Based on these findings, it is reasonable to argue that the most important road type to attend to is urban highways since this road type is affected by all types of adverse weather conditions at all time periods and during all four seasons. During snow/heavy snow, all road types need attention, especially during mid-day and evening. Special attention should be given to rural highways in the mid-day and evening and arterials in the morning. The implications of this research for literature and practice are discussed in sections 7.2 and 7.3, respectively. This is followed by a discussion of the limitations of this research in section 7.4; and directions for future research are provided in section 7.5.

7.2. Implications for Literature

The proliferation of big data in recent times has given researchers in different fields an opportunity to carry out analyses with greater spatial and temporal granularity than was possible
before with traditional data. However, it is imperative to understand the details of the big data context in order to make sure proper use of big data is achieved. The research presented here is built on a theoretical understanding of the big data environment and its application. The Cynefin framework (Snowden & Boon, 2007) proved to be a good framework for categorizing and understanding different aspects of big data. Thus, it allows researchers to make the complicated environment of big data less complex and to focus on the most essential aspects so that they can use basic analytical methods as well as more complex ones. The purpose of the research included use of the available big data to analyze relationships among weather conditions and traffic speeds that have been examined in previous research on this topic. The methodology and techniques used to analyze the data allowed for predictive analytics, and pointed to particular timeframes and locations for intervention. As indicated by the Cynefin framework, this analysis not only required expertise in data analytics techniques but also subject matter expertise to decide granularity with respect to location, time, and seasonal variations that are important from transportation planning and facility operating perspectives.

The research points to the importance of understanding the goal of the activity before jumping into the multitude of data available. Without such focus, the large amount of data could overwhelm the research. There are a number of data providers marketing their data products and one can easily get inundated with all types of data that may or may not be useful for the purpose of research. Setting the goal upfront allows the boundaries of the research to be defined so that the right type and amount of data can be acquired. Goal setting allowed this research to focus on the right data and to pick the most important components from the large datasets based on how useful they would be for the research. In addition, big data made it possible to bring together seemingly disparate fields and to understand the relationships between (and among) these fields.
Dhar (2013) and Wladawsky-Berger (2014) had suggested such research as one of the most important possibilities of big data. This research confirmed that such analyses are possible and can enrich the research on the field of big data.

One of the major discussion points among researchers in the big data environment is the relative importance of predictive power (Graham, 2012; West, 2013; Bar-Yam, 2015) versus correlation identification (Anderson, 2008; Myer-Schonberger & Cukier, 2013). This research shows that both have their place in the research. Correlation analysis was used to identify the weather types that influenced speeds and regression analysis was utilized to understand the intensity of impacts of each weather type.

The increased impacts of worsening weather conditions associated with climate change make it urgent to achieve a better understanding of the relationships between weather and various other fields. Transportation planning and road traffic management are two fields that require a better understanding of the impacts of variations in weather. Prior literature suggested that understanding the impact of adverse weather conditions is critical in terms of reducing traffic incidents, fatalities, and congestion; fuel consumption; environmental pollution; and associated economic costs (Rahman & Lownes, 2012; Andrey et al., 2013; Tsapakis et al., 2013; Sathiaraj et al., 2018). The majority of previous studies relied upon traditional data, or required data collection efforts for a specific study. Today, researchers have new forms of big data available for similar studies. Current literature suggests that use of big data datasets in transportation has been increasing but there is still a gap in understanding the relationship between traffic speeds and weather conditions.

In this context, this research tested the impact of adverse weather conditions on traffic speeds using different datasets, thus demonstrating an empirical test of the big data environment
and the larger field of knowledge in big data, transportation, and emergency management. Different big data datasets in transportation and weather were evaluated to determine whether big data sources offered resources with superior performance. The Capital Region in New York State provided the opportunity to get access to the latest big data available in transportation and, therefore, was selected as the study location. Two datasets, MIST and INRIX, were selected because they both have more granularity than traditional data sources, but these datasets were collected using different methodologies. The research shows that though these two datasets are comparable, INRIX data are available at 5-minute intervals whereas MIST data are available only at 15-minute intervals. For real-time prediction, such as for automated driving, such granular data become much more valuable. Focusing on understanding the effect of individual weather conditions for traffic operations purposes, rather than on predicting speeds, this research used OLS regression analysis. In order to understand variability, three different locations in the Capital Region that had differing characteristics in terms of nearby development (urban vs. rural) and road types (highways and arterials) were selected. Further analyses were conducted to understand differences due to time of day and seasonality. As expected, the analyses confirmed previous findings that speeds, on average, are reduced during adverse weather conditions. More specifically, the analyses found variation in impacts based on different timeframes, seasons, and locations.

The analyses for highways provided more or less similar results as previous research. For all road types, speed reductions occur during periods of snow (Tsapakis et al., 2013; NHTSA, 2018; Sathiaraj et al., 2018). While speeds on both highway types are affected by heavy snow, speeds on the arterial are not affected by heavy snow. A possible explanation for this might be that highway travelers may feel the safety risk more than those on the arterial and
therefore reduce their speeds. In addition, not many people travel during heavy snow.

Interestingly, traffic speeds on the urban highway are reduced during periods of rain while speeds on other road types are not. Similar to previous research (Smith et al., 2004; FHWA, 2010; Dehman & Drakopoulos, 2017; Sathiaraj et al., 2018), the speeds are reduced by 2%-10% on different roadways during light rain and by 5%-10% during heavy rain. However, speeds on urban highways are more impacted than are speeds on rural highways and arterials. Speeds are observed to be reduced only by 2%-5% on most roadways during light snow. This is contrary to previous research (FHWA, 2010; Tsapakis et al., 2013; Sathiaraj et al., 2018) that showed 5%-15% reduction during such conditions. However, it is noted that a 12% reduction was observed on rural highways. Therefore, overall, the results could be considered comparable. Speeds are observed to be reduced by 10%-36% during snow/heavy snow. Previous research (FHWA, 2010; Sathiaraj et al., 2018) shows the reduction to be only 10%-20%.

Though all adverse weather conditions affect traffic at all times, rain and heavy rain affect the morning peak traffic on arterials (31% reduction in speeds) and evening peak on urban highways (up to 17% reduction in speeds) much more than all other times. Traffic speeds on urban and rural highways are affected by light snow at all times but speeds on rural highways are affected more in the evening (up to 12% reduction in speeds). Snow/heavy snow conditions consistently reduce traffic speeds but have a higher impact during mid-day (up to 36% reduction in speeds) and evening (up to 33% reduction in speeds). All these findings show that the weather impacts different road types differently at different times of day.

As expected, traffic speeds on all roadways are reduced during periods of snow conditions during winter (up to 35% reduction in speeds). During other seasons, rain is the main cause of speed reduction, and reduces speeds up to 23%. This is very prominent on urban
highways compared to other road types. Combining these findings with the time-of-day findings suggest that mid-day and evenings on highways are the most critical segments that need special attention.

The comparison given in Table 21 below shows that the results of analyses using big data are similar to the results of analyses using traditional data. However, the comparison also shows that the granularity in big data allows for the comparisons of specific road types during specific time periods and seasons, which, in turn, would allow for more information about where interventions might be focused.

Table 21. Comparison of speed reduction from previous and current research

<table>
<thead>
<tr>
<th>Weather condition</th>
<th>Previous research</th>
<th>Current research</th>
<th>Additional observations from current research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light rain</td>
<td>2% - 10%</td>
<td>2% - 10%</td>
<td>Larger reduction on urban highways; evening peak – 5% - 15%</td>
</tr>
<tr>
<td></td>
<td>(Dehman &amp; Drakopoulos, 2017; NHTSA, 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy rain</td>
<td>4% - 20%</td>
<td>2% - 20%</td>
<td>Larger reduction on urban highways; summer on urban highways up to 23%</td>
</tr>
<tr>
<td></td>
<td>(NHTSA, 2018; Sathiaraj et al., 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Light snow</td>
<td>5% - 15%</td>
<td>2% - 12%</td>
<td>Larger reduction on rural highways in the evening peak – 12%</td>
</tr>
<tr>
<td></td>
<td>(Tsapakis et al., 2013; Sathiaraj et al., 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow/heavy snow</td>
<td>10% - 20%</td>
<td>10% - 36%</td>
<td>Larger reduction on urban highways (up to 34%); mid-day up to 36% on rural highways</td>
</tr>
<tr>
<td></td>
<td>(NHTSA, 2018; Sathiaraj et al., 2018)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on these findings, it is reasonable to argue that the most important road type to attend to is urban highways since this road type is affected by all types of adverse weather conditions at all time periods and during all four seasons. During snow/heavy snow, all road types need attention, especially during mid-day and evening. Special attention should be given to rural highways in the mid-day and evening and arterials in the morning.
The results show that big data can provide a more granular and nuanced understanding than was previously possible with traditional data. Analysis of different road types beyond highways is possible. Even within a particular road type, multiple locations might provide different results and the big data can provide that type of detail since the data are available from all locations whereas the traditional data systems were available only from a limited number of locations.

This finding is very important for management of road operations during adverse weather conditions. Currently, general warnings are provided across the board irrespective of weather type, road type, and time of day. The findings show that this could be customized to the specific situation. The results, when compared to previous research, show that the big data datasets with more granularity can be highly valuable in directing scarce financial, equipment, and staff resources to specific time periods and road types during different times of the year. As indicated in the literature, the cost of poorly managed traffic that results in incidents and congestion is very high.

7.3. Implications for Practice

The research showed that big data datasets are better than traditional and other advanced datasets in terms of spatial and temporal coverage, granularity, and cost. Traditional data collection approaches included temporary data collection using special tools on some high traffic locations only. The data collection frequency used to be once in two or more years. Big data approaches allow collection of data every day from all road locations. In addition, big data approaches provide 5-minute (e.g., INRIX), rather than 15-minute (e.g., MIST) or less frequent, interval data provided by traditional methods. Moreover, collection of data through traditional
approaches requires specialized equipment, electric connections, and the relay of data to a central processing center. This is not only costly but prone to missing data due to equipment and network failures. Moreover, collection of big data datasets such as INRIX uses mobile phones and GPS devices and provides much larger coverage spatially. Disaggregated data have become more important in the wake of the COVID-19 pandemic since traditional traffic patterns have been interrupted by the pandemic. Practitioners trying to figure out the traffic impacts will want to understand the nuances in spatial and temporal differences at the most granular level possible. Unit cost of data procurement is also expected to be lower since the data coverage has expanded.

The research also showed the need and potential methods to combine datasets from different fields (e.g., weather prediction and traffic). Cross-field collaboration and analyses are currently rare, particularly due to the tendency for professionals and agencies to work in silos. The availability of big data and advanced data analytics tools makes it possible to combine datasets with data collected from different fields (e.g., police records of incidents) and at different levels (e.g., time periods).

The research shows that more disaggregated weather data are required. For example, currently, weather data are only available for one site in the entire Capital Region. However, weather conditions can be drastically different from one locality to another and can change within a short period of time. Therefore, decision-makers at weather monitoring agencies should promote and invest in improving and expanding equipment that can capture more granular data and larger spatial coverage. The best value of big data can be derived only when both the traffic and weather data are in the same timeframe. In addition, current weather reporting systems such as AWOS do not provide road temperatures, which can predict the potential for ice and slippery conditions that have a strong influence on speeds and safety. The weather monitoring
community could deploy additional equipment at specific locations for cost-effective spatial coverage and calibrate such equipment to collect data focused on additional important weather factors. The research points to the need for increased and improved data observations in several locations and during specific timeframes. Big data providers could use this knowledge to either collect data from more devices or combine data from multiple sources to improve the quality and quantity.

This research was possible because the big data in transportation were made available to the researchers through special permission from NYSDOT (which is the custodian of the data that was collected from travelers’ devices), collated by INRIX, purchased and distributed by FHWA, and packaged by AVAIL at the University at Albany. This shows the complexity of getting the data for research purposes. A more formal, accessible, and open data policy will be helpful for researchers without the need to go through special permissions each time they try to access data. The cost of such data procurement and distribution can be borne by various levels of government and academic institutions. In addition, the agencies with access to data could provide internships to students that will allow innovations. It is important to ensure data privacy when such wide sharing is made available.

The research points to the fact that, as suggested in the Cynefin framework for the big data environment, there is a need to use highly specialized tools and techniques to break down the big data from a ‘chaotic’ domain to ‘complex’ (relationship between cause and effect can only be perceived retrospectively) and ‘complicated’ (knowable but the relationship between cause and effect requires expert analysis) domains for better handling. In order to make this happen, highly skilled data scientists are required. Currently, there are only a handful of research teams such as AVAIL, the Center for Advanced Transportation Technology Laboratory at the
University of Maryland (CATT Lab), and the Texas Transportation Institute (TTI) that are equipped to provide such services.

The majority of transportation professionals are not data scientists and need user-friendly tools and interfaces to focus on answering the questions they are exploring. This also points to the fact that the big data environment needs a combination of data scientists and subject matter experts (SMEs). The Cynefin framework points to the need for a multi-disciplinary approach and this research confirmed the same. One implication of this for academic programs in transportation planning is that there is a need for more data analytics courses in transportation planning and other fields. The traditional courses do not generate professionals who can handle big data. In fact, conferences such as the annual conference of the Association of Metropolitan Planning Organizations (AMPO) have pointed to examples of transportation agencies hiring data scientists to deal with big data in recent years.

The research finding that the handling of such huge and disparate datasets requires a combination of computing and subject matter expertise can lead agencies to hire and sustain highly skilled professionals who are trained in cross-subject knowledge. For any such data analytics to be successful, decision-makers at agencies need to make appropriate changes in the human resources management policies to ensure they hire professionals with better analytical and multi-disciplinary skills. In addition to adding data scientists to the teams, agencies will need to establish multi-disciplinary coordination among different departments and agencies such as transportation, weather, emergency management, and information technology. These teams can enable a common understanding of the goals, data, tools, and techniques across knowledge domains and allow for efficient use of resources.
This research not only confirmed past research findings about the relationships among traffic, weather, and speeds, it also enhanced it by providing more granular results. Disaggregated data allowed for more detailed spatial and temporal understanding. The research provided coefficients that indicated the optimum speeds under different road and weather conditions that can improve traveler safety. Though the percentage reduction in speed can vary from one region to another, the method used proved to be sound and can be replicated elsewhere. Algorithms can be developed based on this research that can be generalized and used in other regions.

Though the R² values were generally low, in cases such as morning peak on the urban highway and mid-day on rural highway included in this study, weather conditions explain up to 28% of speeds. The goal of the research, however, was not to predict speeds, in general, but to understand the contribution weather conditions make to speeds. The analysis clearly illustrates the variation across weather conditions and road types, with statistically significant coefficients. It also points out that there are additional factors that contribute to speed variation. With the findings, roads can be made safer, based on the knowledge gained regarding these variations, with the recognition that, on average, drivers respond to weather conditions.

The research showed that better speed warning systems customized to specific times and locations can be developed and likely to be more useful than the current system of general speed warnings across the board. This will allow transportation planners and traffic operations managers to develop and deploy more accurate and timely warning systems. Traffic operations managers can operate managed lanes where variable speed limits can be applied and alert messages provided depending on the weather condition. The benefits are not only related to safety but also to reliability, energy efficiency, and reduced environmental impacts. Smooth
traffic patterns make travel and freight delivery times more predictable and reliable. This can have a substantial impact on freight supply-chain management, especially in the current environment of increased proliferation of e-commerce in recent years and the COVID-19 pandemic that has made e-commerce even more prevalent. Scarce resources can be strategically and more effectively utilized to get the maximum benefit. These benefits would grow exponentially as weather conditions deteriorate with climate change.

It is also clear that real-time analysis and information dissemination will become more critical in the future. In particular, connected and automated vehicles (CAVs) will need to take the longer delay time of all drivers into account in order to calibrate and implement the vehicle assistance systems that keep the vehicle speed under control and provide safe travel. Research and analyses similar to this could allow CAVs to automatically respond to weather conditions and reduce their speeds using artificial intelligence. Not only can such real-time analysis provide CAVs information about the appropriate driving behavior, the vehicles, in turn, can provide real-time data to data collection systems. Data from vehicles equipped with sensors to detect road surfaces’ freezing and snow accumulation conditions can then be fed back to the systems to be used in deciding responses. It should be noted that though the data are provided to researchers and practitioners in near-real-time, it is possible to provide data in real-time for CAVs and similar technologies for best utilization.

Decision-makers within related agencies could promote policies that incorporate weather impacts on road maintenance and design. This research indicated that heavy snow has more impact than light snow. Further exploration could provide potential levels of snow accumulation that may have different impacts. The road maintenance crew could use such analyses to keep road surface conditions under the threshold that can avoid traffic incidents. Transportation
planners and road designers could use this knowledge to better plan for and design new and improved facilities taking into consideration the weather effects. This could include decisions on the type of road surface (e.g., concrete vs asphalt), additional shoulder width, and unobstructed view.

Cost-effective data collection and analytical methods that are replicable are very important. These allow for continuous analyses and dissemination of findings as well as monitoring of effects over time without the need for a whole new set of analyses. However, these need support from decision-makers who could develop policies on data standards, data sharing, financial support, and human resources management.

In order to make standardized data collection and sharing more effective, federal and state governments will need to implement policies that give common and consistent directions for private and public organizations to develop innovative data collection/management tools and techniques while enabling users to have reliable and comparable data. Local governments are already developing regulatory policies for managing the proliferation of transportation network companies (TNCs) such as Uber and shared mobility devices (SMDs) such as bikeshare. Data policies could be added to these regulations or might be standalone policies that cover fields other than transportation. The federal policy to procure traffic big data and to share the data with state and metropolitan agencies is highly visionary. This not only allows for some standardization but also makes procurement cost-effective. In addition, such policies should be developed for interagency data sharing at federal, state, and local levels as well as among different levels of government, academic institutions, and researchers.

The cost of acquiring data is not small. However, research has shown that data can provide large benefits in terms of safety, reliability, and emission reduction. It can be safely
assumed that the cost of acquiring data is much less than the cost of human lives, property, negative effects on the economy, and environmental impacts that are caused by these weather conditions. Different levels of governments can join together in sharing this cost, which federal agencies have already started exploring (e.g., NPMRDS).

Road speeds are determined based on several factors including the type of road, the level of traffic, and location, among others. This research shows that the speeds should be variable under different weather conditions. In practice, average speeds are noted to be higher than the speed limits in many places across the country. Policies need to be developed to account for this so that agencies can reduce the speed limit below the desired speed limit to account for weather impacts.

Moving Ahead for Progress in the 21st Century (MAP-21) placed increased emphasis on performance management within the Federal-aid highway program and transit programs, and requires use of performance-based approaches in statewide and metropolitan transportation planning. Between 2016 and 2018, FHWA provided guidance on making performance measurement mandatory for transportation planning and programming. These measures include travel time reliability and safety among others. As this research showed, investing in data, tools, and techniques that enable better management of road traffic operations is key in improving the performance of the transportation system. In addition, many states have adopted Vision Zero policy for safety whereby municipalities are working to reduce the road incidents and bring fatalities to zero within a certain time period. Policies on data and weather systems will help municipalities better equip themselves in working towards Vision Zero.
7.4. Limitations of the Research

This research provided important findings that build on previous research and fill gaps in current fields of knowledge. There are, however, several limitations in this research that need to be understood. Due to the lack of availability of multiple years of data, the data were collected and analyzed for one year only. This potentially limited the ability to account for more variations in travel behavior. Since snow has the most impact on travel speeds, having more days when it was snowing in the dataset would have probably allowed for a better understanding of the effect of snow. Similarly, more of the other factors that influence traffic speed such as road construction and incidents could have been accounted for if the data involved multiple years.

The current analysis was carried out with data collected from three selected locations in the Capital Region in New York State. These included one location each from an urban highway, a rural highway, and an urban arterial road. More segments from all road types, including local streets, could have led to richer findings. The general weather pattern is different in different regions. This limits generalizability of the findings in terms of coefficients of impact to regions other than the Capital Region. A comparative analysis could have given a deeper understanding of similar and/or different weather patterns in different regions.

Though big data has more granular observations, it needs to be understood that if there are not enough observations made by the data collection devices, the data provided may cover more time but will not have many more observations than traditional data. This is particularly prominent on local streets and for late night and early morning timeframes.

In addition, data were collected from one direction of traffic only – northbound on highways and westbound on the arterial road. Analyzing both directions at the same location
could potentially give a better understanding of the influence of directionality in travel behavior. Most roads have a peak direction of travel – a direction in which the majority of travel occurs at different times, particularly in the morning and evening. At the two highway locations analyzed, the peak direction is southbound in the morning and northbound in the evening. On the arterial road, peak directions are eastbound in the morning and westbound in the evening. The selection of one direction only limited the possibility to differentiate the influence of peak travel direction. Another important factor that is missing in this research is the traffic volume. Higher traffic volumes make the road congested and reduce speeds irrespective of weather conditions. Including volume as one of the factors influencing speed could isolate weather impacts from congestion impacts.

Three road segments with more or less straight and flat geometry were used in this research. Speed limits, perception of safety, and speeds could be influenced by road geometry, which includes slopes, curves, and obstructed views (e.g., due to trees) as well as the presence of merge lanes, highway exits, and intersections (signalized and non-signalized). These were not taken into account in this research.

Incident data were not available at the level and ease that would have allowed them to be used for this research. MIST data used in this research did provide some of this type of data; however, these are entered manually into the system by staff monitoring the traffic and could be prone to error. Incident data from police reports and the Department of Motor Vehicles (DMV) could have given a more detailed understanding of different types of incidents in the study locations. Incident data might not only have provided an understanding of how they influenced traffic at different times of day and seasons but also the extent of the impact.
Weather data are only available at one-hour intervals and only for one location for the Capital Region. Weather could be different at different parts of the same metropolitan region at the same time. In addition, weather could change drastically in a short period of time. The data used cannot capture these variations. It is a limitation of the data available rather than a limitation of the research. More spatially and temporally granular weather data could have provided more detailed, and potentially different, models and results.

This research used the type of weather condition (e.g., snow) for analysis. There are several other weather conditions that might impact traffic speeds. These conditions, including amount of precipitation and road surface temperature, could affect travel speeds. The weather data available do not include factors such as the ice condition of the roads, which has large impact on travel behavior. This also is a drawback of the available data.

7.5. Directions for Future Research

The research findings, implications, and limitations point to several areas for future research. Literature on big data showed that there is a need for exploring integration of data from different fields to get the best value. Using OLS regression analysis, this research explored combining weather and traffic data to understand impacts of traffic speeds. There are other fields such as behavioral economics that can shed more light on the ways people respond to different conditions. Simulations can be developed to understand people’s behaviors related to various incentives or disincentives for different behaviors during adverse weather conditions. In fact, auto insurance companies have started offering discounts on premiums if vehicle owners install speed and break monitors in the vehicle. These devices monitor driving behavior in different driving conditions including weather and congestion. Once the technology is proved
robust, one can expect the premiums to go higher or lower based on the driver’s behavior. Similar behavioral aspects can be studied through simulations.

In addition, other multi-disciplinary perspectives need to be brought into the analyses. Traffic and many factors influencing it can be considered as a system with several feedback loops. As seen in the literature, traffic speed, volume, road geometry, incidents, and similar factors influence each other. This could be understood using a systems perspective. Are there balancing loops that limit potentials for change? Are there reinforcing loops that can be utilized to change behavior to achieve desirable outcomes?

One of the major advantages expected from big data is to give the possibility of a large sample size that makes N=all. However, there are limitations to this due to the lack of a number of observations in certain time periods (e.g., late night) and locations (e.g., local streets). Therefore, combining data from multiple sources such as INRIX, TomTom, and HERE can potentially increase the number of observations and give more coverage, spatially and temporally.

As a next step to the current research, algorithms reflecting the relationship between weather conditions and traffic speed can be developed and applied in real-time simulation studies. Pilot tests can be conducted by employing automated speed limits (ASL) on certain road segments and the impacts of these speed limits can be studied.

As mentioned above, one of the implications of this research is that it prompts governments to explore data sharing and cost sharing opportunities. The cost of acquiring and managing big data could be much lower than the savings from reduced incidents, fatalities, and other factors including environmental benefits. Such cost comparisons need to be carried out to develop policies at different levels of government in the use of big data.
Different organizations such as departments of transportation, departments of public works, and meteorological departments have a large role to play in traffic management. It is important to study their relationships and networks to improve collaboration and coordination of activities. This could include areas such as data sharing, common policies, managing traffic operations, and construction, among others.

For more generalizable results, data from multiple years can be used. The data for more years could remove any potential limiting factors such as road construction and economic downturns. In addition, data from multiple years can be used for longitudinal studies that can show any behavioral change over time. This study was conducted in the Capital Region of New York State. Within the region itself, the analysis could be expanded to compare more diverse locations, road types, and geometric conditions of the road segments. By increasing the coverage of data to both directions of traffic, the impact of peak travel direction can be accounted for.

Similar studies need to be conducted in other regions of the country to understand regional differences and compare results. In New York’s Capital Region, snow is identified as a major factor. Such comparative studies can show major factors in other regions (e.g., in the southern states of the United States) where weather patterns are totally different.

As and when available, detailed traffic incident data could be used to understand the weather and traffic conditions related to incidents. Removing data associated with traffic incidents could also potentially give a different picture. Similarly, removing data for the recurring traffic congestion from this dataset could remove the effect of general congestion and isolate the relationship for weather and traffic speeds. Future research could also look at potential reductions in incidents when measures such as variable speed limits are activated.
Current literature shows that some studies have used special weather observation equipment to collect data from multiple locations in a region. This can increase the number of samples available and test local variations in weather. This could be employed to understand the impacts better at different locations within a region. This type of equipment could be calibrated to provide data at more frequent intervals in order to allow for temporally granular analyses. Use of other weather factors such as amount of precipitation will also be possible in such situations. Sensors can be used to measure size of snow and ice cover and surface temperature on streets to understand their impacts. Such analyses using expanded data components could help account for different factors influencing traffic speed and increase the fit of regression analyses. This could potentially address the low $R^2$ identified in this and other previous research.

7.6. Conclusion

This dissertation explored the possibility of using big data in analyzing the relationships between adverse weather conditions and traffic characteristics. An understanding of what big data means and the multiple aspects of big data in a theoretical framework combined with the subject matter knowledge helped in framing the analytical methods in the right context. The literature provided the basis for an understanding of the relationships between weather conditions and traffic and for identifying gaps in the current literature. Further exploration of data systems available helped narrow down the potential analyses based on the best available datasets (MIST and INRIX). The potential of big data was explored by using more granular analyses than was used by previous research. This included different road types, times of day, and seasons of the year. Correlation and OLS regression analyses brought out the nuances of relationships between weather conditions and traffic with more granularity than was traditionally done and allowed the research to derive implications in greater detail.
The contributions of this research include highlighting the need and potential methods to combine datasets from different fields (e.g., weather prediction and traffic), a theoretical understanding of the big data environment and its applications, supporting and enhancing the literature on weather-traffic relationships, and providing directions for future research on this field. The practical implications for the profession include providing a better understanding of the potential for the analysis of big data datasets to provide insights on the weather impacts on traffic that can be used for better advance warning, the need for better weather reporting systems, and the potential for real-time analyses that can be used for upcoming deployment of technology such as automated vehicle systems.

As shown by this research, big data has made understanding nuances in various fields such as traffic management easier and more readily accomplished. Artificial intelligence and data science have made it possible to combine disparate sets of data from multiple sources to generate new insights that were not possible in the past. The empirical test carried out by this research demonstrates that simple analytical methods can be used to look at traffic and weather data to identify the desired safe speeds at different types of roads at different times of day and seasons depending on weather conditions. This allows for the traffic managers to provide advance warning, save lives, reduce injuries/property damage, manage congestion, reduce pollution, and save fuel and energy consumption, potentially at a fraction of the current cost. Once real-time data and analysis become a reality, algorithms based on such research can manage the traffic without continuous human intervention and costly infrastructure. Imagine the situation where vehicles collect and feed data to a repository that uses artificial intelligence to provide feedback to the vehicles to modify driving behavior to safely take people to their destinations and, in the process, saves huge public costs and reduces adverse impacts on climate
change thus becoming a self-sustained system. This research shows that such a future is not only possible; it is already here.


References


doi:10.1016/j.neucom.2015.08.100


http://www.nysmesonet.org/


Appendix 1. Syntax for Regression Analysis

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Speed_mph
/METHOD=ENTER lightrain rain fog lightsnow snow heavysnow
/RESIDUALS DURBIN HISTOGRAM(ZRESID).