161505-1

Strategic Research

PROGRAM

Big Data Scan

September 2015

Texas A&M Transportation Institute

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
TTI/SRP/15/161505-1		

4. Title and Subtitle	5. Report Date
Big Data Scan	September 2015
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Robert Cuellar, Stacey G. Bricka, and Maarit M. Moran	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Texas A&M Transportation Institute College Station, Texas 77843-3135	11. Contract or Grant No. 10727

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Texas A&M Transportation Institute College Station, Texas 77843-3135	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by the State of Texas.
Project Title: Big Data Scan

16. Abstract

Research entities and private businesses are seeking to tap the information power within big data to create more effective decision making. To fully use the potential within these complex data sources, there is a need for transportation researchers to articulate the opportunities for integrating big data within decision support systems and to possess the analytic tools to process big data's information. Technology trends are creating big data within transportation. Data management demands can occur with data sets that are large, complex, or collected at such high speed that it becomes difficult to process this information by traditional data management tools.

This project explored focus areas where a research university can forge new statistical program methodologies to realize the benefits from big data resources. The project team prepared an inventory of ongoing research projects currently managing the analytics of large data sets. These projects were clustered into the focus areas of mobility, safety and operations, policy, and infrastructure. Interviews were conducted with researchers within these focus areas. Results show several opportunities involving big data analytics in areas such as energy and safety research. Potential projects cited included the forecasting of roadway deterioration by linking oil and gas activity data provided by state energy agencies with travel patterns on rural roads provided by state highway agencies. Other findings included the potential to consolidate large data files from multiple safety agencies into one database that would open opportunities for cross comparisons of cost, risk, and performance measurement analytics. Results show that a research university can promote data exploration, stimulate interest in big data analytics and build its portfolio in this area by investing in carefully selected research efforts to demonstrate the benefits of big data.

17. Key Words	18. Distribution Statement
Big Data, Data Analytics, Data Management, Statistics Programs, Cloud-Based Analysis, Transportation Research	No restrictions. This document is available to the public through NTIS: National Technical Information Service Alexandria, Virginia 22312 http://www.ntis.gov

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	57	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Big Data Scan

by

Robert Cuellar, P.E.
Senior Research Engineer
Texas A&M Transportation Institute

Stacey G. Bricka, Ph.D.
Research Scientist
Texas A&M Transportation Institute

and

Maarit M. Moran
Associate Transportation Researcher
Texas A&M Transportation Institute

Report TTI/SRP/15/161505-1
Project 161505-1
Project Title: Big Data Scan

September 2015

TEXAS A&M TRANSPORTATION INSTITUTE
College Station, Texas 77843-3135

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgments

The authors recognize that the support for this research was provided by the State of Texas. The research team also wishes to acknowledge the guidance provided by Ed Seymour, P.E., Ph.D., of the Texas A&M Transportation Institute in the design and execution of this study.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of this research project was to inventory and evaluate efforts across the Texas A&M Transportation Institute (TTI) focused on integrating big data analytics into the Institute's research portfolio, with the objective of identifying investment priorities in order to aid researchers as they pursue those opportunities. The research process involved inventorying projects that used large data, debriefing with researchers who indicated an interest in big data, conducting a best practices scan of related activities at peer university transportation research centers, and identifying foundational areas that could benefit from financial support to help TTI move toward leadership in the big data analytics area.

There are a couple of key points to keep in mind while reading this document. First, while it is natural to focus on the size of the data set as the definitional element in big data, it is more appropriate to consider where and how the data are analyzed when determining if research is big data. The power of big data analytics is in finding the needle in the haystack. It is in leveraging large volumes of varied traffic data sources in order to answer a key question that cannot be answered through the analysis of any one data source. Second, and equally important, big data analytics should be pursued as an additional tool in the TTI toolbox but not as the only tool. As TTI invests and advances its analytical capabilities to develop expertise in this area, big data analytics will complement, not replace, the current analytical tools, software, and use of large data.

The TTI project inventory (contained in Appendix) provides information on 23 existing projects, an additional proposal, and five ideas (offshoots of existing projects) that have been identified by TTI researchers as related to big data. TTI categorized the projects into major research areas at TTI such as policy (human behavior simulations, etc.), mobility, operations (travel simulation, safety operations, etc.), and infrastructure (geometric design, pavement maintenance, utilities, etc.). The results of this inventory reveal that most of TTI's big data projects are not yet large enough on their own to meet the widely accepted definition of big data based on volume, velocity, and variety. However, researchers are using ever-larger data sets, and the power of harnessing big data at TTI comes from the combination of those various data sets.

TTI's leadership in the application of transportation data to solve sponsor issues gives the Institute a natural entrance into the big data arena. Most of the projects listed in the TTI project inventory are candidates for big data analytics with the right expertise, software, and hardware resources. This research found that one element to be considered in building the big data analytics portfolio is how to hire and share data scientists and skilled analysts/programmers with the requisite software and analytical capabilities across divisions. In addition, once these experts are hired, the secondary challenge is the chicken-or-egg dilemma: wait to hire until funding is in place, or hire in advance and leverage the staff's talents and expertise to generate work. The key appears to be keeping these highly skilled scientists and analysts intellectually challenged while researchers work to leverage their expertise to bring in more relevant projects. Ideally, TTI should consider housing data scientists and programmers/analysts in a cost-recovery center or centralized program to give those researchers pursuing projects with big data analytics the potential to share this expertise. In turn, these centralized staff can then help to link and create

stronger working relationships across the programs, which would benefit all of TTI by providing a strong foundation for growing this new research area.

A second factor is TTI researchers' strong entrepreneurial spirit, which has successfully grown the agency and achieved high levels of diversity in its research portfolio. This research proposes the establishment and funding of a guiding committee that would provide a venue for the various groups to regularly communicate and jointly pursue opportunities.

In addition to needing to develop the right mix of staff and a forum for cross-divisional collaboration, TTI is limited in the software and hardware to support projects leveraging big data analytics. Brad Hoover, with Network and Information Systems (NIS), is conducting an independent assessment and forecast of hardware and software needs to support TTI's growth in this area. The project inventory developed in this project shows that the majority of software tools currently in use by TTI researchers are not designed for big data implementation, with the exception of a few researchers who use R and other programming languages. Statistics programs, spatial analysis tools, and a variety of SQL software were the most common types of software identified as being used. Both proprietary and open-source versions of software were used, and there was no consistency across all researchers.

With respect to hardware, most teams reported storing and accessing their data from servers centrally located at NIS. Although supported by high-speed connections, this remote access causes inefficiencies in data processing. Strong consideration should be given to investment in flexible cloud-based storage that can expand or contract based on project needs.

Discussions with industry experts suggest that TTI could also benefit by establishing a big data guidance committee. This committee, which would be comprised of researchers from across the Institute and would include one representative from the TTI Leadership Team, would focus on the broader issues of staff expertise, hardware, and software requirements, as well as on the marketing and pursuit of big data analytic applications. The committee would be most effective if funding were provided to cover time and travel to related workshops and conferences. A supplemental action to creation of a guidance committee would be the addition of a representative to the NIS advisory committee that promotes the interests of big data analytics.

As a leader in transportation research, TTI is well poised to lead the transportation industry into a big data analytics environment. Current sponsors are beginning to grapple with issues such as open data legislation, the need for data governance, and a fact-based approach to data privacy. The best way to get started in launching a big data analytics effort is to focus on a single-use case, identify an area of risk that requires a big data analytics solution, or maintain value of data assets (or any/all combinations of the three). Each approach requires committed funding to get started and then has the ability to convert to cost-sharing centers that would allow researchers to leverage this overhead investment into sponsored research. Parallel to these internal activities is the opportunity to partner with other Texas A&M University System members to leverage opportunities.

- A single-use case approach focuses on the development of one project and uses that project as a test case to identify what is needed to conduct big data analyses at a broader level. The Portland, Oregon, Regional Archive Listing (PORTAL), Regional Integrated

Transportation Information System (RITIS), and Urban Big Data Centre (UBDC) are all examples of specific projects that have led to broader and stronger data programs at their respective university. Using these as an example, TTI can leverage its partnership with McLane, Sensecorp, and Ayata to maximize the return on investment in such an effort.

- The second entrée into big data is to minimize risk. TTI is a respected leader in transportation research. With more than 600 employees researching a variety of cross-cutting topics in independent programs, the Institute faces risk in two ways: (a) ensuring consistency in reporting results and (b) maintaining value of data assets:
  - **Ensuring consistency in reporting.** Depending on the data used, it is possible to have multiple reports communicate what appears to be conflicting results of the same key transportation indicators. A very real example of this is with estimates of vehicles miles traveled (VMT). From a finance perspective, VMT is generated based on odometer estimates (all miles traveled). However, in the Texas Department of Transportation (TxDOT) traffic data, the reported VMT is only for VMT on roadways that are the responsibility of the state. As a result, very different estimates of VMT traveled in Bexar County can be generated from within TTI. While the standard research protocol (such as documenting the data source, the VMT calculation process, and other specifics) can help to minimize the probability that this would be seen as a conflict, it still requires time and attention by researchers to investigate conflicting results and remedy them (often after projects are closed out) and/or ensure that the reports being released document the apparent discrepancy so sponsors know this is not a mistake or a quality control issue.
  - **Maintaining value of data assets.** As indicated throughout this report, TTI currently has a decentralized data model. Each program/division houses and maintains its own data sets using its own systems and processes. Some units have hired data specialists with stronger programming and analysis skills; others have not. For each data project conducted independently, TTI faces the potential loss of value through missed opportunities to pool funds across projects and save money by efficiently collecting data that are of value to multiple projects, through loss of data knowledge and availability when a project ends, and through significant resource inefficiencies because of duplication of effort. Moving to a cloud-based storage system will help to provide a centralized location for these data sets (minimizing loss of data knowledge and availability) and open the door to improved communications regarding these assets.
- The A&M System is well positioned to launch a significant big data analytics initiative to reinforce its connection to the business world. By virtue of common membership, this provides TTI an opportunity to both contribute and join their efforts, as well as have access to resources to supplement TTI's expertise in attracting new sponsors. These efforts might include contributing toward Dr. Banks' initiative on ground-based autonomous systems, as well as supporting the innovative partnership between Mays Business School, the College of Science Department of Statistics, and SAS, which are positioned to create data scientists and thought leaders in big data predictive analytics.

TTI is well respected by transportation agencies, professionals, and industry experts alike. Its leadership and insights are highly valued across a multitude of topics, which gives TTI a strong platform to show leadership in big data analytics and access to big data opportunities. A historic challenge has been the lack of funding, which limits TTI in terms of having the right staff

resources, the right hardware, and access to shared software resources. TTI can address this challenge by establishing a big data guidance committee focused on encouraging and funding a big data initiative. TTI can promote data exploration, stimulate interest in big data analytics, and build its portfolio in this area by investing in carefully selected research efforts that can demonstrate the benefits of using big data analytics to create data mining efficiencies. This top-down approach is different from how TTI has grown new topic areas in the past, but the approach has the promise of launching a long and successful effort that will last far into the future.

1. INTRODUCTION

The term *big data* has emerged in recent years to characterize the creation and collection of a wealth of new sources and types of data. Research entities and private businesses are seeking to tap the information power within big data to support more effective decision making though the use of advanced analytical tools, software, and hardware. To fully utilize the potential within these complex data sources there is a need for new tools and skill sets in order to process big data's structured and unstructured data from disparate sources to answer new and more elusive research questions.

The opportunities associated with big data keep getting bigger. In 2013, there were 4.4 zettabytes (equal to a trillion gigabytes) of data in the digital universe. By 2020, it is estimated there will be 40 zettabytes (*1*). This dramatic growth is due in large part to technological advancements like the Internet and the spread of wirelessly connected devices. The sheer quantity and nature of big data has made finding data easy, while its dynamic and varied nature creates significant new challenges for storing, processing, integrating, and analyzing the data. Addressing these challenges can create opportunities to harness the value of big data.

The Texas A&M Transportation Institute (TTI) seeks to identify and solve transportation problems through research; to transfer technology and knowledge; and to develop diverse human resources to meet the transportation challenges of tomorrow. To that end, this project reviewed past and current TTI projects and practices in the context of how researchers currently deal with big data and where investment would allow TTI to leverage new opportunities and be recognized as a thought leader in the transportation industry. TTI researchers have experience in processing and analyzing voluminous data (in gigabytes) using traditional analytical tools (such as Microsoft Excel®, single processor programs, and SAS). They are on the cusp of processing and analyzing data of petabytes size by using non-traditional tools (such as multiprocessor algorithms and artificial intelligence). All indications are that change is on the way, and TTI will be playing a role in helping to manage that change.

2. PROJECT METHODOLOGY

This report reviews the emergence of big data analytics and its growing role in transportation research and planning; investigates existing projects at TTI that involve, or could benefit from, big data analytics; and identifies how technical and technological aspects of big data can open the door to new opportunities and sponsors for TTI. This research aimed to identify near-term opportunities for TTI to forge new methodologies and analytics that realize the benefits from big data resources.

This research involved four steps. First, the project team prepared an inventory of big data projects as defined by TTI's researchers in order to identify the extent to which TTI projects already leverage big data analytics. This produced a candidate list of current TTI projects that access large, complex data sets or integrate complex data analytics. Second, interviews with TTI researchers and the input of a TTI Executive Advisory team helped to identify and refine the opportunities and challenges for TTI researchers and the institution as it seeks to move fully into a big data analytics environment. This established a framework for identifying areas of significant impact based on current research and capabilities within TTI.

In step three, the data gathered in the previous steps were analyzed in terms of strengths, weaknesses, opportunities, and threats (SWOT). Ultimately, in step four, research results were used to evaluate areas of high benefit to TTI for migrating from traditional analysis techniques to big data analytics methods. These areas include:

- Current opportunities at TTI that involve, or could benefit from, big data applications and analytics.
- Potential sponsors for additional research identified through TxDOT, Transportation Research Board, and National Cooperative Highway Research Program avenues.
- Strategies to deploy non-traditional tools and methods to realize the benefits of big data at TTI.

3. DEFINING BIG DATA

The term *big data* characterizes the creation and collection of a wealth of new sources and types of data. The growth of big data is due in large part to technological advancements and the expanding network of wirelessly connected devices. Big data can be as difficult to define as it is to interpret. It comes from many sources—radio frequency identification chips, archives, business documents, weather stations, social media, sensor data, machine logs—and takes many different formats. What separates big data from traditional data lies in the way it is captured, stored, and, most importantly, analyzed. Big data is so large or complex that traditional data analysis techniques are often too slow or inadequate to process it. This can lead to a scenario where large data sets, often readily accessible, hold little value alone. Value is created and captured through the analysis, which may require new approaches designed specifically for the task of combining and combing multiple data sources.

The Three Vs

Big data's divergence from traditional data is often summarized in the three Vs: volume, variety, and velocity. Volume refers to the unprecedented large data sets being produced and distributed. Variety describes the broad number of sources and formats of the data being created; this spectrum is particularly true with transportation data. Velocity refers to the speed at which data are now being produced, often in real time (*2*). Big data can also be defined by a fourth V—veracity—or the degree of accuracy of data—a feature that can be harder to define or measure with big data (*3*).

Another definition of big data is "*data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures*" (*4*). It is not simply size that defines big data; its complexity creates a need to handle data in new ways and the potential to support business growth.

Technological Foundation

Recent technological advancement has led to expanded data access and availability and generated the tools and techniques necessary to analyze and draw value from that data. The growth of available data in the digital universe can be tied to factors such as the *Internet of Things*, a reference to the number of devices that are connected or interact wirelessly with the digital environment. In transportation, this growth is apparent in the proliferation of probe vehicles, dedicated short-range transmitters (DSRTs), global positioning system (GPS) tracking, and vehicle-to-infrastructure (V2I)/vehicle-to-vehicle (V2V) communication. Institutional shifts toward open data policies are another factor that has changed the face of data.

4. LITERATURE REVIEW—BIG DATA AND ITS APPLICATIONS IN TRANSPORTATION RESEARCH

This chapter provides a review of the literature on the emerging use of big data in transportation research (e.g., infrastructure, congestion, operations). Big data is still a new concept marked by rapid change; as such, this literature review focused on transportation data projects with particular emphasis on the following aspects:

- Data sources and software being used—identification of what tools are needed to use or transfer over to a new data source.
- Implementation of big data applications (in transportation context)—process of implementation, impetus, barriers, challenges, and opportunities.
- University centers with big data focus—steps to implementation and integrating big data into existing research.

Sources of Big Data In Transportation

Big data comes from a variety of sources, takes many different forms, and, despite the name, varies in physical size. Typical transportation big data sources tend to come from three main generators:

- Infrastructure data generators such as loop detectors, weather stations, and pavement monitors are a traditional transportation data source. New technology has led to innovations such as the integration of data streams into a single visual form (*5*) and advanced detection methods. Wireless Bluetooth detectors and lower-cost bicycle and pedestrian counters present alternatives to traditional infrastructure detectors (*6*,*7*). These tools offer the potential to provide more data at a lower cost as well as more adaptive and more flexible designs.
- Vehicle data generators have been growing in scale over the past decade, particularly in the context of V2V and V2I (*8*). In transit applications, the coupling of automated vehicle location data to automated passenger counting sensors has dramatically increased the amount of available data. Probe vehicles can replicate the output of infrastructure data generators and do so at a lower cost than fixed sensors (*9*). An industry has emerged, including companies such as INRIX, to sell probe-based travel data to public agencies (*10*).
- Mobile data, generated from smartphones and mobile devices, are the most recent and quickest growing segment of transportation big data and provide two key types of data. Ambient data, meaning GPS records of where a phone is located at all times, can be recorded and used without the mobile user being aware. Companies such as AirSage purchase this information from cell carriers stripped of personal data and repackage the data for transportation analyses. The second type of data is active location data collected from sensors contained within the phone; for the most part, these data come from apps or opt-in services available to the user as used by Google's GPS data service.

Each data generator has unique benefits and drawbacks. Traditional infrastructure data are generally more reliable and readily available. Newer technologies such as Bluetooth, GPS, and

Wi-Fi allow V2V and mobile data generators to replicate much of the more expensive traditional sensors data at a fraction of the cost. Data ownership and privacy issues can arise when data are being generated by and collected from system users as opposed to system infrastructure. However, valuable information and insights can be generated from the analysis of these data sets and through the combination of separate data streams.

Use of Big Data in Transportation

Big data is increasingly being incorporated into transportation research. As big data techniques and analytics evolve, organizations at the same time are looking for new ways to extract and generate value from these new data sources and tools. Most current research uses a combination of newer and preexisting technologies and existing sensors in the transportation network. Big data analysis and techniques are most commonly applied in two areas of transportation research—system management and user behavior.

Systems Operation and Management

Transportation system operation and management has always relied on large amounts of data to track traffic volumes and flows on the transportation network. Historically, transportation management systems have aggregated data into 5- or 15-minute bundles to reduce the size of the database. As a result, current operational tools are geared to these data sources. An emerging application of big data sources is the development of predictive models that build upon real-time traffic data sets to predict future patterns. Existing examples of this are found in real-time traffic displays and trip-planning tools for transit, while other big data sources offer the potential for more complex scenario planning for major events or contingencies (*2*). Generally, new technology and big data sources offer opportunities to redesign existing systems to be more effective.

Probe vehicles are a source of big data that can replicate and replace existing labor- and cost-intensive infrastructure data generators. Probe data may hold the potential to offer new analysis techniques unavailable with traditional sensors (*9*). Researchers are beginning to investigate the reliability and accuracy of purchased probe data from private firms who collect and sell this data to public agencies (*11*).

Bluetooth technology can detect and monitor traffic along arterial roads, and much of the current research is focused around sensor optimization. Research includes development of a methodology for analyzing measurement error (*12*), determining ideal placement of detectors (*13*), and testing coverage zone size effects on data accuracy (*6*). Carpenter, Fowler, and Adler looked at the use of Bluetooth data for the creation of origin-destination tables to study a complicated corridor with alternative route options (*14*).

Technology research on DSRC, V2V, and V2I includes efforts to incorporate additional data sets that expand on their traditional data collection. This includes the integration of weather and work zone data (*10*) and the expansion of cooperative adaptive cruise control systems (*8*).

Travel Behavior Strategies

Travel behavior uses for big data involve efforts to understand current travel and activity patterns and analyze this information to better forecast transportation needs, manage demand on the system, and inform behavior modification programs.

Research has analyzed how V2V, V2I, and DSRC technologies can be utilized to better understand and even influence travel behavior. Nowakowski et al. found that a real-time V2I system that alerts a driver to slower traffic flow on the roadway ahead produced smoother deceleration and a reduction in mean peak deceleration rates (*15*). Similar to an anti-crash alert, this system warns drivers with a soft safety sound indicator when they are approximately 60 seconds from reaching slower traffic.

Smartphone sensors also hold great potential as data generators for transportation research. Abdulazim et al. set forth a methodology for establishing an automated activity-travel survey using smartphones' internal sensors (*16*). Standard smartphones equipped with motion sensors, location sensors, and additional light, sound, and proximity sensors in their system combine location data, land-use data from open sources such as Foursquare, and pattern recognition software to identify mode type. This concept could enable seasonal analysis of travel patterns, reduce the amount of administrative efforts, and permit smartphone-based, behavior modification testing with real-time feedback. An example at TTI is the smartphone application being developed by the Human Factors Program.

Crowd sourcing data is another aspect of big data collection that is used to provide user information though communal participation. Apps such as Waze allow drivers to share real-time traffic conditions with other travelers (*17*). If this type of data can be captured and managed, it provides a large set of information that could be combined with existing data for innovative research, such as the smartphone activity-travel model using Foursquare data to determine land use. One strategy to address quality control of crowd-sourced data is the use of an expert user group, which acts at a level above the crowd-sourced data to ensure a higher level of data quality. Wikipedia typifies this model, allowing skilled or approved users to provide a cursory quality check for manipulation or errors before crowd-sourced data is accepted (*18*).

The basic underlying design is seen in the product or movie recommendation algorithms used by Amazon or Netflix, only it is applied to transportation to suggest or incentivize travel behaviors. Potential outcomes of this research include algorithms and interfaces that are user-specific, allowing for predictive enhancement or suggestions to alter behavior (*2*). As with systems operation and management, big data holds demonstrable value for traveler behavior research and potential influence points to modify behavior to improve the travel experience.

Opportunities

In addition to the expansion of transportation data sets themselves, development in big data computing and processing presents several opportunities for use in transportation research. Several of these are related to the growth of low-cost computing applications.

Cloud Computing

The growth of big data sources has led to innovative technological solutions in data management and analysis. Cloud computing is an alternative to traditional computing for the management, storage, and analysis of this data influx. Cloud computing, sometimes called the cloud, describes a new system in which data are not stored in a single location but shared among multiple remote databases. This is achieved through the establishment of a shared system allowing for the amount of data storage or computing capabilities to be adjusted to meet a user's need at a given time. It is analogous to how a water or electric utility allows users to pay and access only the amount of the utility they need to operate (*19*).

Programming Advancements (Working Smarter)

The ability to quickly analyze complex data sets is currently connected to advancements in computer science, such as MapReduce and other similar processes. MapReduce was introduced by Google in 2004 (*20*) and is a computation process that can process a large data set simultaneously utilizing multiple nodes (processors) in a cloud platform or in a local cluster environment. This decreases analysis time, though it varies with the size and access to the computing cloud (*21*).

Researchers can save time and increase capacity, as data processing tools are advancing toward real-time processing. Portland State University is developing two real-time processing programs in an attempt to allow for data analysis without archiving. NiagraST and latte are programs that allow real-time data streams to be aggregated almost instantly without having to first process and archive large amounts of sensor data (*22*).

Open-Source Software Applications

New transportation management applications have been developed using open-source software (OSS). OSS provides access to cost-effective non-proprietary programs for data management in lieu of traditional license-based software or programs. While the applications are open-source, OSS transportation data management developed from open-source applications can be used to create either an internal data management system with restricted public access or a final application that provides data to the public. While there is not a direct connection between OSS and big data, a large number of big data analytics algorithms (Apache Spark, Hadoop) and platforms are being built using OSS to spur its growth by allowing programmers to contribute in developing the technology.

The use of OSS has both advantages and challenges to be considered in any decision to develop a new application. OSS is generally non-proprietary, which means there is a lack of official support in development. This issue is often addressed through the creation of strong collaborative communities that have developed around specific programs. As such, many agencies model their transportation management applications on existing OSS applications to avoid the risks and costs of beginning from scratch. A challenge to the use of OSS in application development is a lack of staff trained in OSS development and management, combined with the often less user-friendly nature of OSS. The need for skilled staff or employee training must be considered against the other cost-saving benefits of OSS.

One example of the use of OSS in database and transportation management application creation is the Regional Integrated Transportation Information System (RITIS), a large-scale regional transportation integration portal developed at the University of Maryland's Center for Advanced Transportation Technology Laboratory (CATT Lab). It incorporates multiple agencies' data streams, including surface weather conditions, incidents, and traffic volumes and speeds, into a single interface and archive. The portal is built on a number of OSS, including Apache HTTP Server, PostgreSQL, and PHP. The decision to develop RITIS on an open-source system was in part a strategy to use cost-sensitive programs in order to keep barriers to entry as low as possible for smaller cooperating agencies (*23*).

The smart city concept, using digital technologies to broadly improve performance and reduce costs across city functions, is another example of public institutions capitalizing on big data and open-source applications. The City of Chicago made a recent push to create an open data portal for city data. The city decided to build its open-source portal through OSS in large part due to funding limitations. The city's chief data officer spent two years developing an open data portal on an OSS Linux-based system. The move was supported with events such as hack-a-thons and other government supported events encouraging the use and development of private applications based on this newly available city data. The national push for open government data, often referred to as Government 2.0, has been a philosophical shift in the way city data ownership is perceived and the benefits the city can gain from sharing this existing information with the public (*24*).

Alternative Research Model

New data sources and open data policies have led some researchers to change their entire approach. Traditionally, researchers start with a question or hypothesis and then search out the data necessary to test the hypothesis. Big data provides an alternative to this typical approach by allowing the creation of algorithms to uncover correlations in the data that can then be used as the basis of a hypothesis by researchers (*25*).

The power in this approach is that researchers can implement learning machines to identify and refine hidden correlations through the creation of large integrated data sets that were traditionally unnoticeable to the researcher. A willingness to accept a certain level of unknown correlation is also emerging with businesses looking to test if a model developed in this manner is superior to traditional models; the mindset is that the model does not need to be perfect but simply better than what presently exists (*25*).

Challenges

Ownership and Privacy Issues

Privacy issues often accompany big data due to the large amount of sensor recordings being collected unknown to users (*26*). This seems to be less of an issue over time as individuals see positive benefits arising from their willingness to provide their data for collection. Google's model of requiring users of their location-based services to opt-in to sharing their data is an example of this shift (*27*).

Data creation and ownership issues pose a challenge for new forms of data collection. The large and continuously growing number of sensors and wireless-connected devices are boosting the creation of data sets with potential uses for transportation research. Data created while driving can involve vehicle, mobile, and/or infrastructure data generators; working to integrate these three spheres of ownership can provide opportunities to apply big data techniques for innovative research.

Quality of Data

Reliability of data can be an issue with many big data sources. With such extensive and quickly growing data sets, traditional quality checking cannot always keep up with data creation. Commonly, algorithms will be developed to attempt to reduce the impact of errors in large data sets (*28*). The larger issue lies in structural gaps in data collection that may overlook or underrepresent reality in the end data sets, creating a bias in the data. This issue often arises among populations with low or no digital presences. Those populations who do not, or cannot, participate in the data collection process due to the distribution of connected devises are at risk of being systematically excluded from big data analyses, especially those based on personal devices. This issue extends beyond big data, but it may be exacerbated by the often unstructured format of new big data sources.

A number of new techniques and sensors are emerging in the wearables marketplace. Because they take different approaches, they may produce different results measuring the same or similar responses. This big data will need to be resolved and is a new challenge for researchers affiliated with the use of big data. One study examined measurement error reduction techniques in new arterial Bluetooth data collection through modeling various arterial placements, speeds, and through lanes (*12*). Portland State's interactive dashboard—the Portland, Oregon, Regional Archive Listing (PORTAL)—addresses data quality issues by automatically eliminating data that falls outside of pre-established ranges for probable results. Another common quality assurance technique is to have sensors trigger an alert if they are out of sync with nearby sensors or if they stop reporting. This allows for replacement scheduling to be managed at a system level. (*29*).

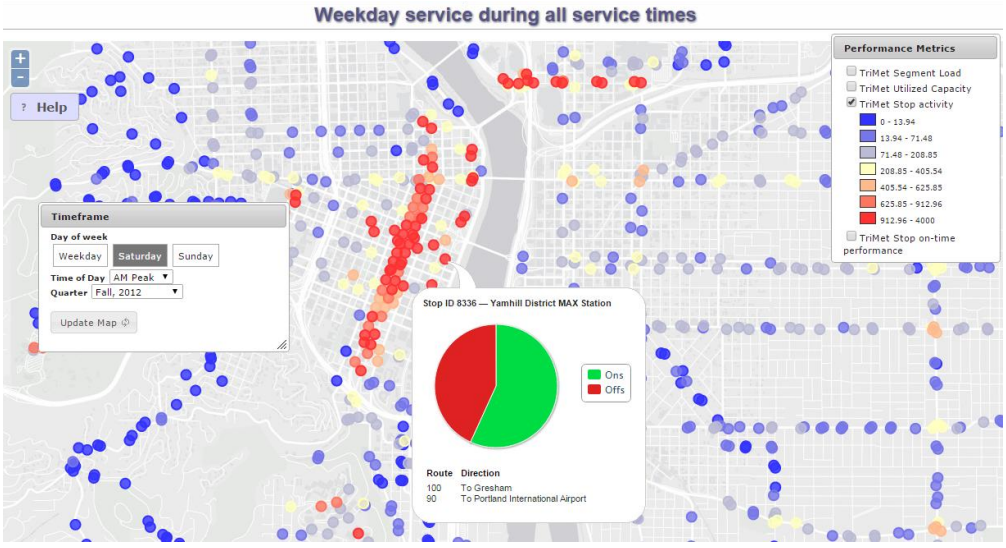
Data Integration Case Studies

Integrating a large variety of data is both a technical and an institutional challenge for big data users. Traditionally, transportation data have been captured and managed by individual agencies, often with limited communication with other city departments. This issue of institutional centers of excellence is a challenge that is being faced on multiple levels of government. The ability for agencies to make the institutional shift from a position considering all data as holding value and moving toward the big data position of open data with value being created through analysis requires much cooperation to be successful.

Portland, Oregon, Regional Archive Listing

An early example of data integration is PORTAL (*30*). PORTAL is an interactive dashboard portal developed by Portland State University and multiple transportation agencies including the City of Portland, the Oregon Department of Transportation, the Washington State Department of Transportation, Metro, and TriMet. It integrates data from each agency into an interactive map and dashboard that displays real-time information; it also archives data for future research.

PORTAL was developed as a key access point for transportation data around the Portland Metro region. Before its implementation, data collected from individual sensors were either stored within one agency or not archived at all. As shown in Figure 1, PORTAL was created with the intent to connect agencies and allow for data to be used for more than their initial purpose, with the belief that the data are too valuable to be used only once.



Source: http://portal.its.pdx.edu/.

Figure 1: Bus Stop Service Visualization Tool.

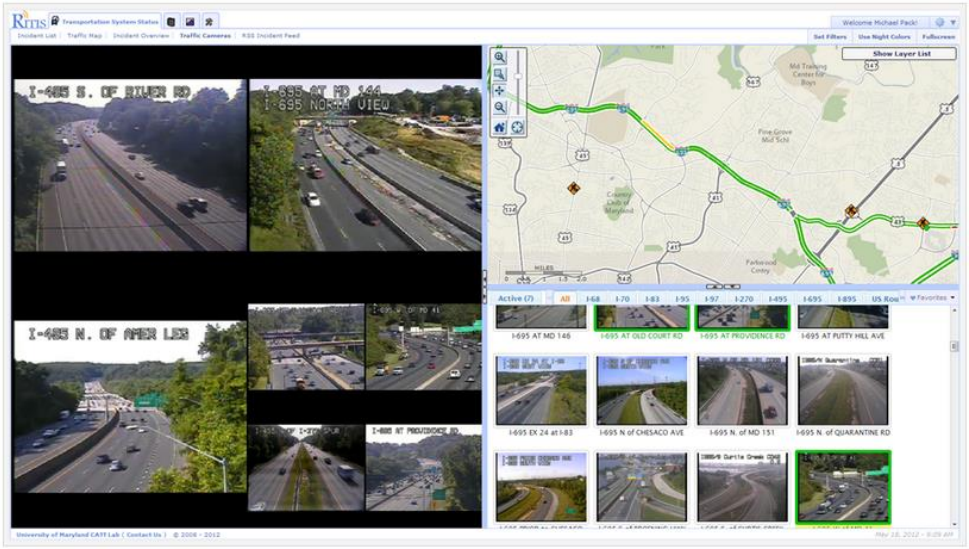
A challenge to implementing PORTAL at Portland State was generating agency cooperation and openness to share individual data sources. Cooperation required commitment and understanding by each agency that open data policies and secondary uses for sensor data would create new opportunities and better understanding of previously unseen patterns or innovations unachievable when the data were stored within each agency. The project benefited from the fact that TriMet, Portland's transit agency, had been on the cutting edge of this OSS philosophy, working with Google in 2005 to create general transit feed specification (GTFS) open-source data for transit data.

Integrating multiple sensor and data collection types into a single usable interface was another challenge for PORTAL. This data integration was achieved through the slow development of the PORTAL system. Starting with highway loop detectors, data collection and storage were constructed to store and display each sensor reading at the most effective visualization level. Once that layer was developed, transit, weather, and arterial Bluetooth sensors were given the same treatment, resulting in an integrated regional data archive and interface. The program is currently working to incorporate bike and pedestrian counts through signal delays and bike-share pilot data (*31*).

PORTAL also emphasizes the use of data visualization in understanding and applying the information collected from these various data sources. This emphasis on allowing deeper understanding and analysis of data is supported by open data elements allowing select data to be downloaded by the public, with no pre-authorization required.

Regional Integrated Transportation Information System

As mentioned earlier, RITIS is a large-scale real-time data feed providing participating agencies the ability to view transportation, weather, and related emergency information through a single service. The system was developed in 2006 at the University of Maryland's CATT Lab at the request of the National Capital Region Transportation Planning Board, the metropolitan planning organization for the Washington, D.C., region.



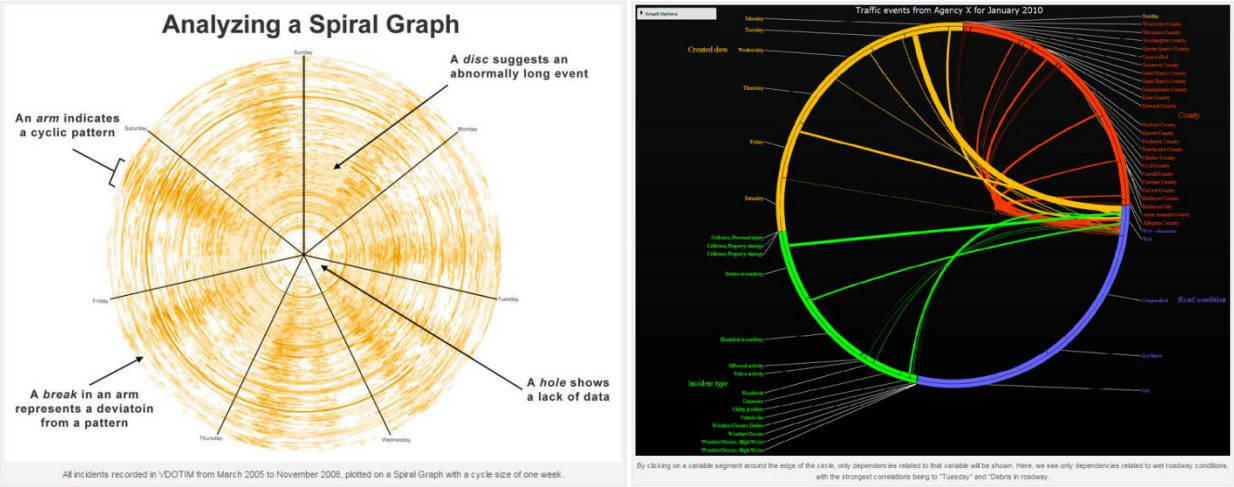
Source: http://www.cattlab.umd.edu/?portfolio=ritis.

Figure 2: RITIS Example Dashboard.

As shown in Figure 2, participating agencies self-determine what information they are willing to share with the system as a control for encouraging overall agency cooperation. This is complemented with a closed system that protects privacy by limiting access of real-time information and archived data downloading to pre-approved participants. The general public and private entities are not allowed access to RITIS information.

The initial purpose of RITIS was to allow the multiple agencies operating in the Washington, DC, region to have access to all available data. RITIS has added new data streams over time and integrated data archiving to allow for historical correlation analysis between previously disparate data. The integration of CLARUS (surface transportation weather data) now allows for correlation analysis between weather conditions and the RITIS archived data. CATT Lab uses this archived data to develop innovative analytic tools, including data visualization tools presenting correlation matching and temporally based analyses, as shown in Figure 3. There have now been over 15 tools developed, most using archived RITIS data.

RITIS and CATT Lab partnered with the private sector in the development of their analytic tools and sensor integration into RITIS. Recently, the University of Maryland's Industrial Partnership Program awarded Traffax Inc., a company specializing in freeway, arterial, and pedestrian Bluetooth tracking, $138,600 to partner with RITIS to integrate their data.



Source: http://www.cattlab.umd.edu/?page_id=17.

Figure 3: RITIS Analytic Tools Visualization Examples.

Urban Big Data Centre

The Urban Big Data Centre (UBDC) at the University of Glasgow focuses on the use of big data strategies to address urban issues, which also creates research opportunities in data management and analytic methodologies. One example of UBDC's research is the Integrated Multimedia City Data study to develop a data set including a range of urban data that includes travel time survey information, location-based blog posts, and television recordings. The methodology of this project, focusing on multimedia database storage, allows urban research into social exclusion, population movements' effects on economic development, and housing and environment, and also allows the formulation of new social/economic/environmental indicators.

Main challenges to the creation of the UBDC include data access, management, and protection. With an end goal of having all available data accessible to approved users, UBDC has created a three-tiered data storage system to protect privacy and to ensure participating agencies' data sets will remain secure. Based on the level of personal information, every data set is assigned to one of the three tiers:

- Open data has minimal user restrictions and is stored on site in UBDC's 700 terabyte data system. Generally theses data sets will be environmental, transportation, or property data that is either public information or contains no personal information.
- Safeguarded data are data that were provided to UBDC on agreement that research will need approval from the data owner; this is done through a research application that must be approved before data access is granted.
- Controlled data, or data that have personal information attached, such as education, health care, or public service usage data, are available for research. Potential researchers must go through training and apply for approval to gain access and all controlled data are destroyed after research is completed. Unlike the open and safeguarded data, controlled data are not stored on the UBDC system. This is achieved through the use of eDRIS (Electronic Data Research and Innovation Service), a service that allows the data owner

to directly share information with the approved researcher, further increasing the level of protection of the data.

Though housed in the University of Glasgow, UBDC has a strong commitment to partnerships with outside researchers. Currently, partnerships exist with several universities in the United Kingdom and with the University of Illinois–Chicago in the United States. Additionally, the center's structure allows for individual researchers to apply for access to data sets that would be inaccessible without the university's strong existing relationship with the data owners.

Case Study Summary

While each case study approaches data integration in a distinct way, a few commonalities emerge. Integration of existing data streams, the consideration of privacy and security during system creation, and a slow development process allow for both a well-functioning system and the flexibility to improve tools and analysis level over time.

The use of existing data sources builds upon the principle that data can be used for more than their primary purpose. The development of archives and data portals allows for subsequent downstream application, with the potential to generate additional value to the organization.

The case studies also demonstrate possible solutions for data privacy issues. A tiered data system as illustrated in the UBDC allows broader access to data streams that contain no identifiable attributes and more restricted access for those data that may hold risk. A tiered system also gives the original owners of the data the ability to help control access, which may alleviate concerns about a shared data environment.

Finally, the use of open-source applications allow for staged development that can adapt as the system incorporates new data sources or new analytical tools. Several case studies use open-source applications to develop data management systems that rely on skilled programmers using existing OSS applications.

5. Project Inventory—Big Data at TTI

As a starting point for evaluating big data opportunities at TTI, the research team developed an inventory of existing or proposed projects at TTI that use, or could benefit from the use of, big data analytics. Principal investigators from across the Institute contributed details regarding projects they felt relevant to the study. These results were used to inform the identification of research topic areas with potential applications of big data, as well as shape the design of the internal TTI interviews discussed in the next chapter.

The TTI Big Data Research Discussion Group served as a lead source for the inventory of big data projects and researchers. Participation in this group was assumed as an indication of experience or interest in big data projects and tools. Additional communication with TTI researchers, the input of an advisory team, and the distribution of a request for information on big data projects expanded this research to the greater TTI community. This team of TTI leaders served as advisors throughout the research on project scoping and staff development. Finally, in-depth interviews were held with select groups of TTI researchers and are documented in detail in the subsequent chapter of this report.

Project Inventory

Outreach to TTI researchers and a review of major projects formed the basis of the project inventory. Information was gathered to identify projects that access large, complex data sets or integrate complex data analytics. The objective was to identify current approaches used within TTI for data gathering, analysis, and supporting tools and analysis for projects utilizing data identified by TTI principal investigators as being big data. *Ultimately, the inventory was designed to help researchers make the link between what they are currently doing and what more they could be doing with the right support.* Information was collected from September to October 2014. The full project inventory can be found in Appendix C of this report.

The inventory was designed to achieve several goals. The first goal was to acquire basic project information, purpose, scope, sponsor, time frame, and data sources for TTI projects. In collaboration with Brad Hoover at NIS, who was doing a concurrent review of software and hardware needs across TTI, researchers were asked to provide information on software uses and needs—analytical tools used and needed. The inventory also included some open-ended questions about limitations and opportunities for further research or analysis. The information requested and documented in the inventory is summarized in Table 1.

Table 1: Information Requested in Project Inventory.

Project Information	Activity (project, proposal, idea)
	Name of PI
	Project name/topic
	Sponsor
	Short scope summary
	Timeframe for working with this data (already doing it, < 1 year, etc.)
Data Needs and Software Tools	Data source
	Data file size, file format
	Analytic software tools (R, SAS, MySQL, etc.)
	Other comments
Opportunities and Limitations	What (else) could you do with this data if you had the funds and the software?
	Would you need help in identifying the appropriate analytic tools that need to be developed/located? (Yes/No)
	What's keeping you from fully leveraging this data?

Inventory Results

Over 30 responses from TTI researchers were received, providing information on 25 projects and an additional six proposals and ideas, some offshoots of existing projects. As shown in Table 2, the projects were categorized into major research areas at TTI including what is categorized as *policy* (human behavior simulations, etc.), *mobility*, *safety and operations* (travel simulation, safety operations, etc.), and *infrastructure* (geometric design, pavement maintenance, utilities, etc.). Although many projects do not fit neatly into one category, this characterization was used to support the design of the follow-up interviews and to analyze the project inventory. The remainder of this chapter presents the results and analysis of the project inventory.

The inventory also demonstrated a broad range of sponsors and partnerships in existing research projects. The most common sponsor was TxDOT, but other departments of transportation (DOTs), federal agencies, and other government or university partners were also noted. Several researchers reported using cellphone, GPS, and light detection and ranging (LiDAR) data. Many different researchers use Texas and TxDOT databases.

Table 2: Summary of Project Inventory Results.

Project Type	Number
Project	25
Proposal	1
Idea	5
	31

Project Focus Area	Number
Policy	1
Mobility	18
Safety & Operations	9
Infrastructure	3
Total	31

Project Sponsor	Number
Texas DOT	12
Other DOT	3
FHWA	3
MPO	1
Texas Legislature	2
Other Sponsors	6
Total	**27**

Software Inventory

TTI researchers used a range of different software. With the exception of R and programming languages, the majority of software programs used at TTI are not designed for big data implementation. Statistics programs, spatial analysis tools, and a variety of SQL software were the most common types of software identified. Both proprietary and open-source versions of software were used, and there was no consistency across all researchers in the selection of one software for one type of use. Current software programs used by researchers (based on project inventory, collected fall 2014) are summarized below.

- Spatial Analysis:
  - ArcGIS.
  - ArcInfo.
- Statistics:
  - SAS.
  - R.
- Database and Data Processing:
  - Excel.
  - Access.
  - Oracle.
  - Azure SQL (Microsoft, cloud-based).
  - PostgreSQL (open-source database system).

    - PostGIS (spatial support for database software).
    - MySQL (open source).
- Programming:
    - Python.
    - SQL.
    - PHP.
    - C++.
    - SSAS (Microsoft SQL Server Analysis Services, online analytical processing tool).
    - SSIS (data migration and integration component).
- Simulation/Modeling:
    - LS-OPT (simulation design).
    - LS-PREPOST (processing for LS-DYNA).
    - Vissim (microscopic traffic simulation).
    - DynusT.
    - HyperStudy.
    - HyperView.
- Other:
    - Cube Analyst (OD matrices).
    - INRIX data analytics tool.
    - TTI custom web-based tool.

6. INTERVIEWS—RESULTS OF TTI BIG DATA SCAN

Based on the inventory and a preliminary review of the TTI technical and research activities, group interviews within several major research areas at TTI were undertaken. The following topic areas were identified:

- Mobility.
- Safety and Operations.
- Operations and Energy Sector.
- Modeling.

The interviews were intended to review the potential in each area to develop a big data demonstration project, identify potential partnerships or collaborations, and, ultimately, demonstrate the benefits of big data tools and methods to TTI's researchers, sponsors, and partners. Interviews were conducted with researchers at TTI's College Station headquarters and in the Houston, San Antonio, El Paso, and Austin urban offices.

Summary of Interviews

Key conclusions that emerged based on interviews with key TTI researchers are highlighted below. (Detailed meeting summaries are provided in Appendix B of this report):

- Researchers provided a variety of ideas concerning what big data means to them. Some researchers defined big data as data sets that are too large to be processed by traditional programs like Microsoft Excel® on a standard desktop PC. Other researchers suggested it is the speed and volume (not the size of the data set) that matter. Some researchers considered the term *big data* to be simply a clever marketing term with little practical meaning, while others saw potential in TTI performing large data set processing and analysis. Finally, some researchers noted that adopting a standard big data definition for TTI could be a next step, while others disagreed that TTI should spend time or effort on a common definition.
- Researchers noted several opportunities for TTI involving big data analytics, especially in areas such as energy, transportation project planning, and safety research. Most TTI researchers believed there are research opportunities in the big data analytics space. For example, researchers in the San Antonio office noted several ongoing projects linking oil and gas activity data provided by the Texas Railroad Commission with travel patterns, noting that there is currently no easy way to compare these two data sets. Several researchers also noted that project planning and administration usually involve thousands of documents that can take up several megabytes of information. Furthermore, researchers noted that having the ability to capture and consolidate this information into one database could help open up several new opportunities in answering critical cost, risk, and performance questions. Finally, several researchers noted that future opportunities in safety research, especially in emerging areas such as autonomous vehicle technology, could provide TTI with a strong leadership position.
- Researchers perceived greater opportunities for leveraging big data information and analysis capabilities across TTI through a centralized portal maintained by NIS. Nearly

all researchers interviewed for this study noted that TTI researchers could do a better job coordinating and sharing information with each other. Because of TTI's reliance on project sponsors and limitations on agency-wide funding, progress is often limited to what can be accomplished within its decentralized environment of individual projects in different divisions across the Institute. This reliance on project funding also means that when the project funding ends, there are no resources to nurture the data or grow all data into a larger combined data set. For example, sometimes research project contracts only authorize funding for collecting data sets to be used for the purposes of that project. Once the project is finished and the project account is closed, the researcher tends to discard the data. This creates a storehouse of information that remains with the original project and limits the value of the data to only the original project. Therefore, several researchers suggested one way this storehouse effect could be addressed is through the creation of a centralized portal maintained by NIS. This centralized location could be the start of also co-locating shared expertise in big data analytics and software, as well as providing a stronger understanding of the hardware needs associated with maintaining and analyzing big data. The recent development of a cost center for the Current Research Information System (CRIS) database is an example of how this effort could begin.

- Finally, researchers noted that balancing Institutional Review Board (IRB) and other data privacy requirements while ensuring that accessing data does not become overly burdensome will be a key challenge to overcome. Several researchers noted that data privacy and security will continue to be a challenge. In some research areas, such as safety research, strict IRB/human subject research requirements must be followed. However, currently large data sets (which usually go through encryption and other protocols when accessed on cloud infrastructure) can be time consuming to access via remote computing infrastructure. These issues will need to be resolved in order to make progress. Examples as noted in the three university-based case studies in this report can serve to help find a way to balance these competing requirements on access to the data.

As noted previously, while most TTI researchers identified key big data challenges that will need to be overcome, nearly all researchers still perceived big data analytics as an area has great potential and research opportunities for the agency.

7. BIG DATA AT TTI SWOT ANALYSIS

The identification of the most effective uses of big data (sources, tools, or analytics) included the development of an action plan to engage TTI researchers, private partners, and public agencies in the use of non-traditional data management applications and tools that would demonstrate the benefits of big data tools and methods in TTI's research program. The identification of partners and projects for the big data demonstrations included consideration of TTI's favorable position in test bed development and opportunities to leverage existing cooperative relationships with logistics companies.

Researchers analyzed the strengths, weaknesses, opportunities, and threats (SWOT) gleaned from the previous analysis steps to help identify the elements of an effective action plan. Following is a summary of a SWOT analysis seeking to identify the best path forward for TTI.

Strengths

- A wealth of data sets.
- Expertise in data quality control.
- Expertise in transportation data sources.
- Existing and long-standing partnerships with public and private data providers.

Weaknesses

- Limited use/knowledge of advanced processing tools (most people use Excel, Access, and some statistical programs, not all of which are viable for manipulating big data).
- Limited staffing with the programming and analysis expertise to analyze and manipulate big data.
- Institutional hurdles including stringent privacy protection and project-based funding and charge system (restricts time and funding of researchers, and does not allow researchers to fully realize the full value or potential of data collected on individual projects where funding has ended or across projects, where the combining of multifaceted data sets has the most power in big data analytics).
- Lack of funding to develop a centralized data management and analysis system or to share resources (staff and software), causing programs to develop their own systems (duplicating use of resources across the Institute and creating storehouses of information that remain with the original project).

Opportunities

- Supercomputing facilities at Texas A&M University.
- Access to open-source and highly sophisticated tools such as R, SAS, and multiprocessor cluster environment.
- Researchers can be trained and begin using programming tools and open-source software on current projects.

- Presence of common themes among data collected and used by TTI—congestion, physical roadway attributes, safety data, weather, etc. This suggests that they can be combined at a higher level for powerful processing opportunities.
- The topic of big data analytics across Texas A&M provides multiple opportunities for TTI. Two key areas include:
    - Recent launch of a new Master of Science degree in analytics offered by the Texas A&M Statistics Department in partnership with the Mays School of Business. Aided in part by a large donation from the software company SAS, this degree was created with the specific intent to equip students to navigate the explosion of big data. It is intentionally designed to accommodate both traditional students as well as working professionals. In addition to providing a source of student workers, who traditionally have been groomed for full-time research positions, this program offers TTI researchers the opportunity to advance their own skills in this area as well as partner with the expertise of the Statistics Department in pursuing research opportunities.
    - Dr. Banks is leading the development of a white paper that is part of a university-wide initiative. This paper has the working title "Ground-Based Autonomous Systems." It will include drones in a support role to roadway-based transportation.

Threats

- TTI lacks a business plan that assesses how effective it might be in the big data analytics marketplace.
- Multiple researchers may be developing tools to process similar data in different projects. This is inefficient and creates a fragmented approach to growth in big data analytics.
- Many researchers with PhDs and years of experience are spending too much time processing data, which data scientists or analysts with good programming skills can perform in less time. This limits researchers' opportunity to focus on the analytics and reasoning instead of data processing that could be done by trained data scientists and analysts.
- Data growth is faster than storage capacity and processor growth. Storage use is growing by 2 terabytes per month (Brad Hoover, email correspondence, November 11, 2014). In addition, this growth comes during the life of a project, and then the data stay in place after the project ends, without any type of compression or formal archiving. This not only results in data being lost or forgotten but also creates false pressure on the NIS storage capacity and processing needs.

8. DEPLOYMENT STRATEGIES

The research outlined in this report supported the development of possible plans for mobilizing big data techniques (analytics and tools) within existing TTI programs and projects. Specifically, the SWOT analysis in the previous section provided the basis for the design of these deployment concepts. Furthermore, discussions with industry representatives suggest that TTI could benefit by establishing a Big Data Governance Committee to guide researchers toward the utilization of big data analytics.

In this chapter, three example projects are included to illustrate areas where the TTI Big Data Guidance Committee could demonstrate how big data analytics can be deployed at TTI.

Example 1: National Centers Match Proposal

The following is an example of a proposal that was submitted for consideration in the Spring 2015 National Center's Match call for project ideas, and although it was not originally selected, it was ultimately funded:

Proposal Title: Establishing Relationships with New/Different Commercial Big Data Providers

If the past few years are any indication, much of the mobile device data useful for transportation analysis will come from companies in the location-based services marketplace. The goal of public agencies collecting usable data directly from connected vehicles is admirable but still many years in the future (if ever fully implemented). TTI researchers recognize this increasing role of commercial data providers (e.g., INRIX, HERE/Nokia, TomTom) and have established excellent relationships with several companies while implementing their data into public-sector transportation analyses.

There has also been growing transportation industry interest in utilizing big data analytics for operational improvement evaluations associated with variable speed limits, vehicle incident tracking through usage of streaming data, and refinement of heat maps and other visualization tools for analysis of big data findings.

In this project, the TTI research team will perform the following tasks:

1. Conduct ecosystem market assessment—TTI will identify the leading companies in the connected car and connected trip market through online searches and coordination with the Accelerate Center.
2. Initiate contact and gauge initial interest—TTI will contact targeted companies and gauge each company's interest in demonstrating big data analysis techniques in research efforts. Teradata in the field of variable speed limit and Sensecorp in the area of vehicle incident tracking have shown interest in assisting TTI.
3. Establish business model for interested companies—For those companies that indicate interest in reselling their data to research agencies or government customers, TTI will follow up with more detailed discussion of government data needs, capital equipment/server requirements, pricing range, and licensing models. Face-to-face meetings may be most productive at this stage.

Example 2: Internal Proposal

Proposal Title: Data Management and Big Data—Internal Opportunities

This scope of work addresses creating a culture of big data within an enterprise leading to consistency, operating leverage, enhanced awareness, and synergy. In other words, creating a culture of leveraging big data INTERNALLY drives an enterprise strategy that enhances value both for the benefit of the organization and for its customers and stakeholders. This approach has as its core the operational construct to eat what you kill—if it is good enough for our customers to pay us, we should live by it as well.

This scope defines the operating plan for research projects to create and implement a culture of big data. Here are the steps for a candidate big data proposal to be viewed by the TTI governance committee as a priority for funding:

1. Defining enterprise data accessible across TTI—This step creates a framework for sharing applicable information in a consistent format that is easily accessible, is thoroughly indexed, mitigates reinventing the wheel, and facilitates the reuse of those same data for analysis and other projects to identify relationships, trends, and other actionable events that improve outcomes.
2. Establishing rules and access for proprietary data—Respecting proprietary information remains the critical objective in all TTI research and related activities. Definition needs to be established on how to provide permissible awareness while not compromising the integrity of confidentiality.
3. Eliminating stove pipes—All data regardless of restrictions or other concerns should be centrally managed for access. This will require policies and standards consistent with the highest levels of integrity and security appropriate for applicable sensitivity. Disaster recovery and backup systems are part of this process.
4. Moving to the cloud—With the foregoing defined, TTI can move all its storage and servers to the cloud. This will help ensure TTI can deliver cloud-based services and research, and demonstrate the necessary security, reliability, and availability at the highest commercial standards. This will also reduce investment needed in onsite infrastructure and assist in evaluation of new tools and processes expected by stakeholders.

Specific benefits will also be realized across TTI, including but not limited to:

- Marketing leverage—properly capturing enterprise data will enhance marketing and communicating the scale and scope of TTI's involvement, and market leadership, in various areas (e.g., connected and automated vehicles).
- Promotion of TTI as a cloud-based enterprise communicates that TTI is current with the Internet of Things and stays abreast of developments for the benefit of its stakeholders.
- TTI researchers should expect to increasingly be asked about data management, intellectual property addressability, security, etc., and having the organizational and enterprise skills will enhance the value of TTI's work for its clients.

In conclusion, the first step is for principal investigators to work with the TTI governance committee in defining enterprise data and creating a vision of how all TTI activities can be accessed and mined.

Example 3: Data Management and Big Data—External Opportunities

With the continuing evolution of and toward big data and the Internet of Things, TTI's clients and customers are increasingly expecting solutions that include expertise and capability to implement current technology and processes. As a standard set of deliverables, TTI should offer cloud-based services to its clients that include the same level of security, reliability, maintainability, and availability as TTI uses internally. In other words, we eat what we [kill/cook].

For a candidate big data proposal to be viewed by the TTI governance committee as a priority for funding, the TTI principal investigator should define a solution environment that utilizes a cloud architecture to help its external clients in a more cost-effective and data-leveraged way. The TTI Cloud will consist of a virtual private network that clients can connect to in order to review their data and peruse other enterprise data that are generally publicly available. That is one of the values of being a TTI client—access to the TTI Cloud.

An emerging trend in transportation reflects increased IP addressable devices over a network. This trend is accelerating with ipv6 deployments, connected and automated vehicles, and the need for better and improved maintainability and predictive maintenance of transportation assets.

Public-sector clients cannot take on the additional operating expense and IT burden of managing such an explosion of connectivity, while private-sector clients are seeking ways to access publicly available information without violating privacy rights. The solution is defining the TTI Cloud as a toolkit to take to market for TTI and its customers.

TTI needs a plan to develop the products, solutions, and offerings to make this a reality, including public/private partnerships and potentially a business plan.

REFERENCES

1 IDC. The Digital Universe Is Huge—And Growing Exponentially. 17. (emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf).

2 Volpe National Transportation Systems Center. "Big Data's Implication for Transportation Operations: An Exploration" U.S. Department of Transportation. March 2014. Intelligent Transportation Systems Joint Program Office.

3 IBM. The Four V's of Big Data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data

4 Dumbill, Edd. Defining Big Data. Forbes.com, May 7, 2014. http://www.forbes.com/sites/gilpress/2014/11/03/big-data-now-mainstream-in-large-companies-term-still-widely-disliked-a-new-survey-finds/2/.

5 Xie, G. and B. Hoeft. Freeway and Arterial System of Transportation Dashboard. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2271 Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 45–56.

6 Araghi, B, L. Christensen, R. Krishnan, J. Olesen, H. Lahrmann. Use of Low-Level Sensor Data to Improve the Accuracy of Bluetooth-Based Travel Time Estimation. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2338 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 29–34.

7 Ryus, P., Ferguson, E., Laustsen, K.M., Schneider, R.J., Proulx, F.R., Hull, T., Miranda-Moreno, L. NCHRP Report 797: Guidebook on Pedestrian and Bicycle Volume Data Collection. National Cooperative Highway Research Program, Transportation Research Board of the National Academies, Washington, D.C., 2014. http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=3159

8 Zohdy, I. and H. Rakha. Enhancing Roundabout Operation via Vehicle Connectivity. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2381 Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 91–100.

9 Brennan, T., S. Remias, G. Grimmer, D. Horton, E. Cox, D. Bullock. Probe Vehicle-Based Statewide Mobility Performance Measures for Decision Makers. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2338 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 78–90.

10 Maitipe, B., U. Ibrahim, M. Hayee, E. Kwon. Vehicle-to-Infrastructure and Vehicle-to-Vehicle Information System in Work Zones. In *Transportation Research Board: Journal*

of the Transportation Research Board, No. 2324 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 125–132.

11 Edwards, M. and M. Fontaine. Investigation of Travel Time Reliability in Work Zones with Private-Sector Data. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2272 Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 9–18.

12 Moghaddam, S., and B. Hellinga. Quantifying Measurement Error in Arterial Travel Times Measured by Bluetooth Detectors. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2395 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 111–122.

13 Kianfar, J. and P. Edara. Placement of Roadside Equipment in Connected Vehicle Environment for Travel Time Estimation. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2381 Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 20–27.

14 Carpenter, C. M. Fowler, T. Adler. Generating Route-Specific Origin-Destination Tables Using Bluetooth Technology. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2308 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 96–102.

15 Nowakowski, C., D. Vizzini, S. Gupta, R. Sengupta. Evaluation of Real-Time Freeway End-of-Queue Alerting System to Promote Driver Situational Awareness. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2324 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 37–43.

16 Abdulazim, T., H. Abdelgawad, K. Habib, B. Abdulhai. Using Smartphones and Sensor Technologies to Automate Collection of Travel Data. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2383 Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 44–52.

17 Waze Mobile. About Us. https://www.waze.com/about

18 Misra, A., A. Gooze, K. Watkins, M. Asad, C. Le Dantec. Crowdsourcing and Its Application to Transportation Data Collection and Management. . In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2414 Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 1–8.

19 Lei, H., T. Xing, J. Taylor, X. Zhou. Monitoring Travel Time Reliability from the Cloud. In *Transportation Research Board: Journal of the Transportation Research Board*, No. 2291 Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 35–43.

20 Dean, J, and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc. http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf

22 Tufte, K., J. Li, D. Maier, V. Papadimos, R. L. Bertini, and J. Rucker. Travel time estimation using NiagaraST and latte. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data—SIGMOD '07*, 2007, p. 1091.

23 Darter, M. T., K. S. Yen, and B. Ravani. Literature Review of National Developments in ATMS and Open-Source Software.* No. 65, 2011.

24 The University of Chicago. *Chicago: City of Big Data* 2013. Chicago Discovery Series. http://www.youtube.com/watch?v=PCiCUcCuOKg. Accessed November 21, 2014.

25 Bollier, D., and C. M. Firestone. *The Promise and Peril of Big Data*. 2010.

26 Houses of Parliament. Big and Open Data in Transport. *PostNote*. Number 472. July 2014.

27 Google, Inc. Privacy Policy. Last updated August 19, 2015. http://static.googleusercontent.com/media/www.google.com/en/us/intl/en/policies/privacy/google_privacy_policy_en.pdf

28 Bertini, R. L., S. Hansen, S. Matthews, A. Rodriguez, and A. Delcambre. PORTAL : Implementing a New Generation Archived Data User Service in Portland, Oregon. No. November, 2005, pp. 6–10.

29 Portland State University. Multimodal Data Set Clean-up for Portland Oregon Metropolitan Region - Data Set Description and Dictionary - Freeway Data. April 12, 2012. http://portal.its.pdx.edu/Portal/static/files/fhwa/Freeway%20Data%20Documentation.pdf

30 Portal. Portland State University. http://portal.its.pdx.edu/Portal/index.php/home

31 Tufte, K. Portal: Applications of New Technology to Transportation Data Archiving. Presentation at NATMEC Conference, Chicago, Illinois. July 1, 2014. http://onlinepubs.trb.org/onlinepubs/conferences/2014/NATMEC/Tufte.pdf

Appendix A – Interview Guide

NCP Big Data Project – interview outline

Assumes 2 hour meeting

1. **Welcome and introductions** (All) 15 minutes

Ask: Name, Program, Interest in Big Data

2. **Definitional**: Identify how each participant defines Big Data (Facilitator) 15 minute discussion depending on group size.

Ask: what does the term "big data" mean to you? Result should be an agreed upon definition reflecting the group's perspective. [Hypothesis is that this may vary across topic areas].

3. **Project overview and objectives** (Facilitator) 10 minutes
   - **Project Objective**: Identify near term opportunities for TTI to migrate from the use of traditional data analysis tools and to forge new methodologies and analytics to realize the benefits from Big Data resources.
   - **Categories of Big Data at TTI**
   - **Anticipated outcomes and benefit to TTI** (scope recap)

4. **Big Data at TTI – Projects** (Facilitator) 15 minute discussion
   - Summarize the project matrix – here are the "big data" projects taking place at TTI. Are we missing anything or a research area/perspective?
     - For projects represented in the interview: what is/are the key questions to be answered using Big Data? (key question)

5. **Big Data at TTI – Equipment needs** (Brad Hoover) 20 minute discussion
   - Data being used – what data, how selected, collection methods, storage methods, data set characteristics (size, format, variables of interest), software, hardware needs associated with housing the data, etc.

6. **Barriers and Opportunities / Lessons Learned** (Facilitator)– 30 minutes
   - What else could this data be used for? What other tools could have been used/useful?
   - How has this research benefitted TTI? What should other researchers know? What should the transportation research community know?

7. **Opportunity for attendees** to raise any other issues, questions, insights (15 minutes)

Big Picture evaluation to result from the individual interviews:

1. How does the research leverage the "power" of Big Data? Big data has been said to help find a needle in a haystack, and offers opportunities not present in traditional data. In other words, are we doing more than just replacing small data with big data?
2. What are the key lessons learned from our pioneer projects? What are the successes? What key themes can we draw?
3. Institutionally, what is helping and hindering our use of Big Data?

Appendix B – Topic Area Interview Summaries

Mobility Topic Area Interview Summary

This interview took place Tuesday, October 28th, 2014, facilitated by Rajat Rajbhandari in College Station and Maarit Moran via telephone. Meeting attendees included

- Ed Hard, Transportation Planning;
- Shawn Turner, Mobility Division;
- Kevin Balke, System Reliability Division;
- Praprut Songchitruska, System Reliability Division; and
- Brad Hoover, Network and Information Systems (NIS).

Definitional (Identify how each participant defines "Big Data"):

- The group considered the common definition based upon the three V's [volume, variety, velocity]. A fourth V has been added to that definition - veracity. Generally, big data for TTI is data that is bigger than we can handle ourselves; one example is nationwide cell phone data. Currently, big data is really only being used in collaborations with private providers - generally the researcher asks for a particular selection, they prepare the data and give it to us. Big data can also be considered "unprocessed, raw data." It can be real-time data. Another element of the definition discussed was the idea of "data fusion," or merging different data sets.
- "Big data" as a media or marketing term was also discussed. As such, the definition isn't that important, because what matters is whether we have to manage our resources differently. TTI researchers are dealing with new, large, or complicated data that "push the limits" of current data processing or computing capabilities. As such, solutions to better manage, store, archive data is certainly of interest.

Big Data at TTI – Projects

- Projects and interests cover many areas of mobility - interest in development of model inputs, trip tables, trip rates, cellular and passive GPS data, and ways to use these new forms of data to replace older methods for model inputs; performance measures, congestion modeling, and data as it applies to how we monitor alternative modes; connected vehicles, traffic management; data analytics projects including GPS data.
- Certain types of data have already led researchers to find solutions for data needs. Simulation data is one large source of data for some researchers but it is generated data so it is easier to control. Owning the data makes it is in managing and extracting data as you need it. Data storage and processing needs vary between users. Potential solutions include the Cloud, and there are researchers who utilize outside resources for high-performance computing for project work (example: use of supercomputer at Argonne Lab).
- The use of connected vehicle data as a source for other researchers was discussed. The consensus in the group was that in terms of applications for their research needs, USDOT-defined connected vehicle project data (particularly DSRC data) are not expected to be a major data source for mobility and congestion research. Nor is it expected to be a near-term application. There was more interest and potential for opportunity from GPS or cell data than CV. USDOT CV work is also moving towards cell and other types of communication data, however the tools and methods for accessing that data are not yet clear. A possible near term opportunity – team up with TXDOT to deploy CV, generating a resource that can be a selling point.

- The discussion focused on the fact that each researcher is focused on their own data applications and needs. Everyone individually processes the data for their unique applications. They are all different. Understanding the application requirements [for researchers] is critical.
- Non-disclosure agreements with data providers restrict us from secondary use of the data or use it for other projects.

Big Data at TTI – Equipment Needs:

- Currently, researchers are using SAS and R for analysis, and MS Excel and GIS more often to visualize data. R was noted as a useful tool, that is open source and constantly updating but not as user-friendly as products like SAS that incorporate a graphic user interface. It is easier for individuals with programing experience, and works with python for scripting language.
- The existing capabilities for data storage and usage were discussed. NIS is tracking use and TTI's needs are growing at 2 terrabytes a month in just storage. They get many requests for servers for processing. A&M is looking into research for big data computing. Need ability to turn on data processing ability, maybe a cloud can do it.
- Cloud capabilities were noted as a potentially valuable tool to investigate. Transferring to the cloud may a difficult transition. Safety/security is an issue but its something we know how to deal with now. Amazon Cloud meets the TAMU system audit requirements…. FISMA compliant? Confidentiality could be an issue, and possibly encryption.
- One question is do we have the expertise to use and take advantage of the tools? Education may be necessary. Changes may require having technical experts who understand the tools to work on that, while leaving researchers to do research. McLean advanced technologies may provide background expertise, but on the other hand add complexity to projects or require a lot of additional education.
- The idea of figuring out the cost of data processing and storage is an important piece of information going forward. It could even support a fee or line item for big data applications in proposals.

Barriers and Opportunities/Lessons Learned:

- For the applications in mobility and congestion, the data is coming from, and is expected to come from, private companies with access to individual travelers and the anonymity is already being addressed. In transportation at least, private sector providers seem to be ahead of the public sector on this front and are expected to continue to lead in this area. This is evidenced by government RFPs that reveal a lack of understanding of big data and unrealistic requests. It doesn't reflect GPS, passive cellular, etc actually provide.
- In cellular data, the US currently relies on one private provider with agreements with 2 cellular carriers. Will there be another one? Who will it be? We are currently dependent on these companies to buy their data without knowing their methods. Do we want [to collect our own] raw cellular data? What should our role be?
- Private companies are trying to create products to sell - someone else needs to measure, compare, and confirm the data - Rather than primary data providers, TTI could be a third party reviewer of these products that might be used by DOTs, others. Maybe TTI researchers can provide quality control or benchmarking and tracking the accuracy of data. This ties back into the different needs of different researchers – origin-destination data doesn't need to know the info at an individual car level, but for safety research that might be important.
- As researchers, we need to know the foundation of the data so we know what to ask for. Providers are not the experts, they simply collect and create the data but they don't understand

the information. We should be able to get data from multiple providers, in order to compare and understand how it is processed. How do we do this? TTI could have a role as an independent third-party "benchmarking" organization. Private providers provide a service, but they are not experienced or expert in the transportation data assessment and analysis that TTI can provide.

- Two projects were discussed that review private data sets with some technical comparisons and validation tests. Benchmarking processes have been around, like in real-time travel information. TTI did a defacto standard for travel time review, using independent re-identification using blue tooth for real-time travel time. Some private data companies have/do contract with TTI or other organizations to provide benchmarking.
- The storage of data was discussed. Researchers are storing a lot of data, but there is not always a particular future use in mind.
- If a data catalogue existed – an inventory of what TTI researchers are doing, what the focus is, who's using what tools – people would be interested in it. It was noted that datasets can become richer over time. Some researchers noted that they do go back and reprocess old data (not always successfully), and other researchers are required to actively archive their data regularly. This is costly and very time-consuming. It was also noted that storage is relatively cheap, but time is important to researchers. Another issue brought up was the potential loss of institutional knowledge in trying to create a centralized database – individual researchers archive lots of lots of travel data from across the state, keeping track of it for their own needs.
- Researchers are (in theory) interested in having data collected, archived, documented, and available for sharing within the organization but do not have the time or funding to do it themselves. It would have to be easy, and is expected to require money, staff time, expertise, communication, updating and constant updating. This likely does not apply to data with non-disclosure agreements or other restrictions on sharing. In the end, individual researchers do not have the time or funding to lead a open data catalogue – but they are happy to share and provide existing data. In the absence of this, "traditional" networking, communication among colleagues, and sharing of resources should not be forgotten.

Safety and Operations Topic Area Interview Summary

This interview took place Thursday, November 13th, 2014 from 3:00 – 4:30pm, facilitated by Nick Norboge, Maarit Moran, and Bob Cuellar. Participants included:

- Bryan Miller, Research and Implementation
- Kevin Balke, System Reliability Division
- Robert Wunderlich, Center for Transportation Safety

Definitional (What does "Big Data" mean to you?)

- Big data is not the same as large data, but the distinction is not in quantity, but how data is analyzed. Static data can be large, but if it's not changing then it is not so hard to work with. One example is real-time data - it can be challenging to manage—analyzing it on the fly as it changes (like a Twitter feed of new information). It can grow in dimensions.
- Complexity was also discussed as a feature of big data. Big data can be created out of small data; having to merge data from multiple sources is something that can always change and evolve over time. You may be looking at small elements over a long period of time. The combination of datasets is a major feature of "big data" (or just new data).
- Being able to visualize data over time and the tools to mine it appropriately really matter. Computers will eventually chug through the data, but it can take time, and that's as big of challenge as anything. Tools and methods are important.

Big Data at TTI – Projects

- Data needs vary if you have a continuing projects or one-time only. For some projects, we have a need to store and process lots of data but they are not recurring. Every once in a while we will have projects where we have to do data crunching and data manipulation—involved in Urban Partnership Agreement project, and we had large amounts of data we had to process.
- Exceptional item projects – School of Public Health, and Trauma Research Center at Memorial Hermann. Funding to create a linked dataset - trauma database (hospital and EMS records) and the CRIS database using probabilistic datasets. We're setting ourselves up to help compile the millions of records and analyze it for all sorts of safety studies.
- Lots of areas where we're ready or will be. We have a lot of data already; many projects have come out of collaboration – identifying a problem that needs a solution. Center for Transportation Safety is involved in many data projects, and many of these initiatives go beyond the Center. Collaboration with UMTRI, UH, TAMU ME, TAMU IE, etc. in a Toyota study that involves data collected 25 times per second. Naturalistic driving studies (ATLAS) that also generate a large amount of data.
- Traffic management data, including incidents, detector data. Data may range from large to "big" and the work includes archiving and quality control.
- Speed, occupancy data is collected and often combined with other databases. Connected vehicle data is another data source.
- CRIS (Construction Risk and Insurance Specialist) Data used to be based on data extracts—combination of common records with abstracts that have full information for every crash. TxDOT has gone away from extracts - which have a ton of personal identification information - and toward Microstrategy.

Big Data – Equipment Needs:

- Researchers may hire others who are more expert for big analysis. Otherwise use traditional tools like excel.
- IT issues can be a problem. Getting computers and networks to communicate together can be a very big challenge! IT security hurdles can lead to storing all data locally to avoid issues. Accessing networks, dealing with firewalls and security can also be an issue when you are trying to deal with partners. Lately, TxDOT wants us to use Microstrategy. We have to access their server—it can get bogged down!
    - Processing power and storage can sometimes become an issue
    - Simulation in Matlab are a big expense, researchers will buy computers if they need to.
- Cloud storage is one tool that can be used to circumvent IT issues. One researcher has started using Cloud services about 6 months ago, transitioning to Azure. An SQL server within the cloud is essentially a server that belongs to you that you can access, but is located physical elsewhere. Means there is nothing to buy or convert. UH has developed data processing system that allows processing within cloud, password protected.
    - Some researchers have had issues using the cloud, so there are some negatives. Used to be issues with state-owned data in a cloud, but that doesn't appear to be an issue any more. TTI employees end up firewalled from accessing Azure services, which is very odd. Not a TTI issue, it's a University issue. That's where we've had major challenges – cloud can be difficult with university rules. We've designed algorithms to use cloud, because cloud is where it's going.
- Privacy and security policies have challenges as well. Human Subject Data face major IRB issues and need to be worked through, especially with safety data. TTI has a sensitive information policy (personal identifiers), how you store it, who has access, etc. Ownership is also a tricky question. When you are manipulating data or combining datasets the ownership of the resulting data can also be a challenge.

Barriers and Opportunities/Lessons Learned:

- Expertise and techniques are important to using datasets. It's nice to know you have data available, and the opportunity is there to merge data from other sources and you don't know until you can search and find what you have. At the same time, you don't want to spend all your day calling others to find what other data is out there and available
- The idea of a consolidated TTI database was posed. In the past, NIS has tried to set something up. A solution, quite honestly, is to avoid the TTI network. Knowing who does what around TTI is a big challenge generally (not just a Big Data issue). There's talent in urban offices, in divisions, and Centers. We don't really know what those capabilities are. One thing I would suggest is to try to identify who's doing what, who's good at what, etc.
    - It would be helpful to understand what Big Data is and knowledge on how to visualize what the data say?
    - It would be helpful to know the links and bringing it all together better
    - If we knew what our data setup would be, data analysis, record keeping, etc. could be a very valuable thing to have
- What is TTI's role, can we leverage Big Data? We've always been in the business of processing Big Data; now it's getting more minute, relational, etc. We'll always be in the business of turning BD into useful information. The players are going to be different, developing the skill sets will be different, we're going to have a constant inflow of people who know the latest and greatest techniques

    - Chicken or egg problem: Do I start with a problem and then find data to solve it or do I gather all this data and then answer the problem at hand?
    - Are we going to get data from others or are we going to create it all on our own?
- IRB protocols, for human subject data generate some issues that must be worked out. For example with CRIS data—blanket IRB but will need to be in compliance with it - IRB Protocol requirements needed!

Operations and Energy Topic Area Interview Summary

This interview took place Tuesday, November 4th, 2014 from 2:00pm-4:00pm via internal TTI Polycom. The meeting was facilitated by Stacey Bricka and Bob Cuellar in Austin. Nick Norboge in College Station recorded the meeting proceedings. Meeting attendees included Sushant Sharma, Ioannis Tsapakis, Cesar Quiroga, and Edgar Kraus in the TTI San Antonio Castle Hills Office. "SAT" is used hereafter to refer to the TTI San Antonio Castle Hills office.

Welcome and Introductions:

- The SAT team works with a variety of big data sources, including GPS data for commercial vehicle enforcement, PMS data (oversize/overweight), and weigh-in-motion data. Data sources include TxDOT and the Railroad Commission, among others. Researchers in this TTI office have used large data sets for years in a variety of capacities and using different software and platforms.
- SAT discussed their interests in big data, with a desire to better understand the trends, statistical software, platforms, processes and tools that would help their do their jobs better. Of particularly interest is anything applicable for analyzing data from multiple different sources. A key question is defining big data (addressed in the next section).

Definitional (Identify how each participant defines "Big Data"):

- In general, for the SAT team, "big data" is any data set that requires a more to a different platform, software, and process.
    - Big data is information beyond what a typical personal computer can handle. Some data sets, especially data in "real-time," can't be handled on a personal computer. This is a common challenge in the SAT office.
        - Most basic data platforms that are easy to use (e.g., Microsoft Excel, Access) have a RAM capability of 50 MB or less. This can be slow and inefficient when processing large volumes of data.
    - Big data is data that cannot be processed using regular methods.
    - Big data is data that cannot be accommodated in typical software such as Excel or Access.
    - The example of Excel: One SAT team member defined "big data" as anything that cannot be handled or processed in Microsoft Excel. (The limits of Excel are familiar to most, as are its processing capabilities).
        - Datasets too large to store and manipulate in Excel creates issues in identifying what platform to use, how to process, and where/how the data are stored vs. processed.
    - The example of Access: For another researcher, the ability to manage and process data using Access, not Excel, was a determining factor in identifying "big data." Access is referred to here as a game changer because once datasets get larger than 1-2+ gigabytes in size, Access users must break out data sets into multiple sheets of information which can be time consuming and create the potential for error.
    - Above this threshold, Oracle or SQL server is required to handle the data in its most appropriate format. Both Oracle and SQL server have different processes and query types, which requires more effort, time, and expertise to perform.
- size alone is not be the only determining metric in big data needs but also consideration of the time it takes to process a data set vs. the time needed to extract an answer from the data.

- One example includes work on a project for the Policy Research Center (PRC) that involves plotting the location of oil and gas wells over the past few years. As long as analysis is limited to plotting locations, it is manageable and can be done via a desktop application. The moment that you involve production for each well by year, for example, it adds a completely different dimension that then requires a shift to a Oracle or SQL server package.

Big Data at TTI – Projects:

- Researchers noted that TxDOT handles 80,000 miles of roadways. There is the potential to enhance this data with details on the roadway infrastructure, RHINO (Road Highway Inventory Network) data, intersection length features, crash data, operations data all in one location. Could this present a future opportunity for TTI? What if project data were also appended?
- Generally, anything that pertains to infrastructure can potentially be used as big data: imagery (aerial or satellite), LIDAR, as built (scanned images). LiDAR (light detection and ranging) is a remote sensing technology that emits intense, focused beams of light and measures the time it takes for those reflections to be detected by a sensor. It is typically used to measure the earth's surface but can provide useful information for transportation purposes. It is worth noting that LIDAR data is becoming more common at TTI.
  - LIDAR data is not as popular because it consumes a lot of resources on regular machines. Designers want to convert point cloud into solid objects to handle much more easily, but even that process takes a lot of computer resources. It really requires the right software to analyze efficiently.
- Another project is part of the Maintenance contract IAC, where researchers are analyzing Weigh-in-Motion (WIM) data. This dataset is not well documented or understood, which creates a challenge in itself.
- SAT researchers noted that a whole research topic area missing from the project inventory was project development and delivery. Developing projects requires a lot of files Such as construction/utility data, project development and delivery documents (often 500-1000 sheets of output), surveys, imagery, and design files. This data is relevant to other researchers as well (e.g. Environmental Quality Division, John Overman's work, etc.) and anyone who deals with construction or materials could benefit from project-based big data.
  - Projectwise is a platform from Bentley that is designed to help manage the design files. Early users of this software had issues, both technical and training challenges. In addition, a reliable network is critical for using multiple applications. Big data, bandwidth and connection issues can present problems for using this software.
- Researchers also noted that the civil engineering community generally is moving from 2-dimentional to 3-dimensional environment.
  - For design, engineers are increasingly using cross-disciplinary programs such as Bentley GEOPAK, Climate Interactive En-ROADS. Other programs such as Autodesk and Civil 3D. LiDAR comes with custom software.
  - 3D and Building information modeling (BIM) – there are a variety of 3D models and we need to optimize their design when trying to incorporate big data. SAT researchers also noted that the GIS component of 3D is presenting an added level of complexity.
    - There is a migration to 3D but current business process is not about 3D models.

Big Data at TTI – Equipment Needs:

- SAT explained their current hardware capabilities, noting that several months ago NIS worked with SAT to identify a better structure for the servers. The result is a server structure with 2 database servers and 2 development/web servers.
    - Of the database servers, one stores data for development purposes and the other is used for applications. Approved software on these database servers include Oracle, SQL server, and ARC SDE.
    - With respect to the development/web servers, a similar arrangement exists with one for data and the other for applications.
    - SAT manages operation and daily management but the 4 servers are housed in CS. The CS location is a result of DIR directive regarding centralized servers. The SAT team accesses the servers using remote desktop applications via 1 GB per second band width. This arrangement works well when it works, but sometimes the server response is too slow. SAT team members will often run a local copy of the data when the task is mission critical to overcome reliability issues.
- Researchers would like to explore the possibility of procuring Oracle Spatial, which is powerful and can perform spatial queries but is expensive.

Barriers and Opportunities/Lessons Learned (What other tools could have been used?):

- Cost. The Oracle Spatial software is powerful and can save the analyst time as well as boost analysis outcomes … but it is expensive
- SAT researchers noted that as an organization, TTI operates on the basis of "silos of excellence". We have multiple units that are referred to as "heavy data users" sometimes using the same datasets but each processing it separately and with very little interaction between the groups. The most common data sets are
    - CRIS data
    - RHINO data
    - HPMS data
    - WIM data
    - Railroad Commission data

  Developing formal databases in more consistent way would save time and money across all users. Efficiencies in scale could result from a central data processing unit rather than having urban offices around the state conducting their own data processing. For example, could Texas Railroad Commission data sets be integrated and shared with authorized users? Anything from the rail commission is untapped – barely scratching the surface. The commission has so many data sets so big and so obscure, this is a huge untapped potential
- What holds TTI back is that we have to "reinvent the wheel" with each group re-processing the data. , why not leverage the potential?
    - TTI obtains permission from TxDOT/RR Commission/etc. to make the data available internally.
    - NIS could set up process with permission restrictions.
    - The data could be stored with its documentation
    - The data could be combined to create one larger "master" database with a variety of indicators that would be available to researchers and sponsors alike.
    - Users could write query to extract the needed data yinto a working file, while preservingthe original data to maintain its integrity.

- In response to a question concerning how frequently data sets are used within the agency, researchers responded by noting that TTI operates on a project-by-project basis; therefore there is no funding to collect and maintain datasets. Many datasets are used for projects on a repetitive basis. A few are critical because used all the time by different groups, and if project ends today, the data could move to another project that starts in 6 months.
- SAT researchers also noted several possible advantages to working with cloud data infrastructure. SAT noted that they worked with NIS to optimize queries, not a lot of available resources to work on a daily basis.

  Cloud benefit is data administration. Whereas researchers in the past would have to spend time patching the servers, with cloud infrastructure this is no longer as important. Not just data access but also the ability to pre-process and pre-package data more quickly as well.
- Finally, in concluding the interview, SAT researchers provided the following important rhetorical question: How do you define transportation? Some look at logistical aspects (commodities), others pavements and infrastructure itself. Safety? This is likely to be important as opportunities in big data are defined for TTI in the future.

Modeling Topic Area Interview Summary

This interview took place Tuesday, November 4th from 1:00pm-2:00pm via Polycom. A follow-up meeting with Andy Mullins was conducted on Monday, November 10th from 1:00pm-2:00pm. Modeling topic area meeting attendees were Jeff Shelton, Byron Chigoy, and Andy Mullins.

Welcome and Introductions:

- Some of the data from modeling perspective, DTA runs, but the files can become very large quickly. Simulation based modeling and vehicle trajectories – path of every vehicle origination-destination pair, node sequence, and time stamp for each node. Run CAMPO for 24 hours, 9 million trips – 10 GIGS, every path of every vehicle generated. One file in one model. Entire model set is 30 GIGS. Another problem – have to run on the server in CS. So have to compress and transfer the files. NIS helped band width but transfers take a while. Now problems staying logged on.
- For meeting tomorrow with McLane, creating 300-400 mg/gig for a small AVI file. Where can this be stored? Desktop runs out, external problems can result, server can run out of room, and finally there are connectivity problems with the NIS server.
- Data cleaning challenges can result. For example, most data has to be post-processed (most model output is text files).
- Austin HTP operates in a similar manner – what can we learn from amorphous bits of info and what does that tell us about human behavior and travel patterns? Is it any better than traditional sampling methods? How to deal with it on a technological basis?
    - One of the important distinctions is IBM definition of big data – size (pedabytes), speed is velocity of intake. Researchers noted that they are used to analyzing data sets that are already cleaned (mostly) or cleaning and trying to make data sets clean. When dealing with large bits of data coming in, researchers currently don't have ability to clean, have to operate non-traditionally and accept error where it didn't exist in smaller files.
    - Postscripts software works well.
    - Technological basis: a lot of what we know about data for past 20 years has been focused on developing relational database systems and associated software. Relational database management systems are ill equipped to do analysis on big data. Google has pioneered structures, rack servers, chunk servers. Hadoop (Java package) goes in and tries to map reduce. Looks at data in very discrete bits and calculations. That takes time (drivers, how fast you can get at drivers, read write access on data pieces).

Definitional (Identify how each participant defines "Big Data"):

- Some researchers defined "big data" as data that must be processed beyond an Excel spreadsheet. Most data Houston works with is pulled from Houston Transtar servers or H-GAC servers and then aggregated/cleaned data is processed via Excel.
- One question that emerged is large data sets vs. extra-large sets. If over 10 GIG, large, over 100 GIG can be a very large dataset and can be difficult to process with a standard desktop computer.
- Velocity (speed of intake) is another important factor to consider. For example, how much getting in at any given time.
- Geography is also important (i.e., how many places are you getting data from, where pull data from – social media, different sources?).

- There is inefficiency within TTI that arises due to the hardware and processing time.

Big Data at TTI –Projects

- Researchers identified several projects that do or could employ better big data processing techniques.
- CAMPO, for example, has one scenario is 30 GIGs. The modeling data set are currently analyzing is over 300 GIGS. Researchers note that they are concerned the model will crash, so back-up for safety is essential.
- Predictive modeling of IH-35 construction related events. Can we take the Bluetooth real time data and do some type of predictive modeling? Similar to weather forecasting.
- Traffic – forecasts, 5 year forecasts, or real time (already stuck in traffic). How do make decisions about tomorrow? These are questions that could be addressed with better data analytics. Maybe not tomorrow but forecasting 15 minutes from now. Modeling results – there's a real need to answer this question. What is the travel time through I35 from Round Rock to downtown if take trucks off 35? What does the bridge operations look like on a typical day for cars and trucks?
- We could answer questions Today – too late, already in traffic. What if received info 15 minutes earlier? UTEP class time impact on traffic.
- Researchers also suggested possible opportunities for cross pollination within TTI. Take one source of data, for a lot of the art of data analysis, it is about stitching together the information to draw a conclusion. Hard to stitch, but if there was access to real time weather data and speed (e.g. Bluetooth) plus historic crash data and congestion implications, several new insights could be learned.
    - Demand files – matrix. Everything is text files are very large.
- INRIX data also presents opportunities for big data. Researchers noted that INRIX has 10 year crash record that is geospatially coded to networks. By running analyses and combining them, it could yield more powerful results. As an example of the tendency to cluster in "silos of excellence," crash researchers work on crashes, while speed researchers focus on speeds.
- Finally, several researchers identified a few problems with not coordinating better on research projects and sharing data that could be mutually beneficial. For example, some researchers don't know what other people are doing, which was considered a TTI problem, not just in the area of big data sharing. This problem also occurred with regard to proposals – researchers are looking for someone to fill a need, find out about common projects. Generally, the way information travels about what projects researchers are working on is simply through word of mouth. TTI day is informative and helpful in that regard.

Barriers and Opportunities/Lessons Learned (What other tools could have been used?):

- A lot of scenario analysis.
- Researchers noted that there's a need for more programmers: not engineers, not planners who became programmers but real nuts and bolts programmers that support larger combined offices. Different types of programmers – web /server / applications (e.g., Ruby, Python, java scripts, html applications, etc.) to the other side being C sharp, Java coding. Researchers generally know what they want to do and can see the programming structure in your mind, but don't know how to code it. There's a real need for programmers available to assist researchers in implementing their vision.
- Researchers also discussed what industries big data is used currently, noting that one place where big data gets used a lot is in credit scoring and credit analysis. However, the same

questions are not being asked in transportation. The same is true in the energy industry. This may be in part because construction is very slow in responding. Identifying these opportunities in the future could be beneficial to TTI.

- Researchers also noted that security is a big challenge. For example, when transferring things and uncompressing, Sophros kicks in and slows down the transfer process, which causes problems
- Major opportunities in improving disaggregate travel demand models, especially given the switch from households as a unit of analysis to the individual as a unit of analysis
- Furthermore, researchers noted that modeling is moving toward the why and away from just "this is what happened." The ability to tie in the why through survey or other analysis could help inform what we see in the Bluetooth travel data. Furthermore, there's a need to keep data in-house in order to perform longitudinal analysis. The researchers noted that when they get a new project, they only pull data from Houston TranStar or another agency. Sometimes they will get a project to compare 2013 travel survey data. Therefore, it would be nice to be able to go back to 2013 Bluetooth data and perform a paired comparison of that information. This could be of strategic importance to house (and own) our own data. In this regard, there are also opportunities to study effects of special events
- Finally, researchers noted that some researchers have excellent relationships with specific TxDOT divisions. For example, the Houston office has an excellent relationship with TxDOT's Travel Survey Collection Division. This relationship could be used to leverage future data opportunities throughout the agency

Appendix C – Project Inventory

Focus Area (assigned)	Activity (Project, Proposal, Idea)	Name of PI	Project Name/Topic	Sponsor	Short Scope Summary	Data Source	Data file size, file format	Analytic software tools (R, SAS, MySQL etc..),	Other Comments	Timeframe for working with this data (already doing it, < 1 year… etc)	What (else) could you do with this data if you had the funds and the software?	Would you need help in identifying the appropriate analytic tools that need to be developed/located? (Yes/No)	What's keeping you from fully leveraging this data?
Infrastructure	Project	Sushant Sharma	*Identify routes used by os/ow trucks and improvement strategies for movement of os/ow routes and roads*	TXDOT IAC		TxPROS dataset (database of annual permits and single trip permit issued for OS/OW trucks by TxDMV)	2010-2013 GIS and oracle dump files 5 million records 35 GB	ArcGIS, Access,Oracle, MySQL server for querying	Data variables: Inspection Year, Report Number, Category, Commodity, Axle group Weights, City/State of Origins and Destinations, Inspection Time, Violation Description, Route Key, Citation Issued, Weight Violation, Driver Out of Service, Vehicle Out of Service, Latitude, Longitude, and Federal Code.	Already doing it	Developing clearance levels and routes required for efficient movement of OS/OW loads.		
Infrastructure	Project	Sushant Sharma	*Energy Sector Impact on Infrastructure*	TXDOT IAC		DPS inspection Dataset for Texas, WIM Dataset for all stations in Texas, Oil & Gas Well Location and Permits (RRC data), TPP Traffic Count Data, Crash Data (CRIS), Energy Sector Related Traffic Count Data, Roadway Highway Inventory Data (RHINO), Pavement Management Information System (PMIS) Data, Railway Data, TxPROS dataset- OS/OW truck routing	2010-2013, some datasets are as old as 2002 or 2007 onwards. DPS Inspection Data- 13GB; WIM Data 27 GB Other mentioned datasets are more or less within the same range.	MySQL Server and Oracle SQl for accessing and querying the datasets. Sometimes we use Access and Excel for quick processing. ArcGIS is used very often for visualization and processing too.	Huge number of data variables for each datasets. For example, just DPS data set has commodity and shipping information, such as origin, destination, contents, hazmat information, weight and equipment violations and descriptions, axle group weights, GVW, axle configuration etc.	Already doing it	There are many questions this dataset can answer. At the top of my head: Safety issues in Eagle Ford Shale Region. Damage Assessment due to Trucks (all types including OS/OW). Forecasting future of non renewable and renewable energy sectors and developing plan to reduce or manage pavement damage, safety in nearby regions and truck movement through counties or city roads.	Yes	Most of the time quick analysis is required for legislature and other critical meetings that requires careful analysis. However, processing time takes longer time because of specific queries to be developed and run. Sometime two different datasets needs to be joined. Data Modeling is challenging give the size of each dataset and amount of variables involved. Issues with importing text files as it being in no specific format.
Mobility	Proposal	Ed Hard	*Compare and Assess Long Distance Trips on major Texas corridors using Bluetooth and Cellular Data*	TXDOT IAC	Compare LD trips between BT and cell to help validate cell data as LD data source needed for SAM	Cell - AirSage, BT - TTI(Hou)	MBs and GBs, not big, but derived from big data	PostgreSQL, Postgis, Cube Analyst, ArcGIS	Proposed as 'special project' IAC-C in FY 15 or 16 under new/emerging data Task	FY 15 or FY 16		No	Non-disclosure agreement.
Mobility	Project	Bill Eisele	*Maryland State Highway Administration Freight Fluidity Measures*	Maryland SHA through Univ of MD	Implement a "freight fluidity" framework and approach in Maryland for freight decision-making.	Multiple; some include INRIX, ATRI, FAF, activity centers	May make use of national INRIX dataset from line 2 to access MD data	SAS, ArcInfo	Mobility and reliability measures, truck delay, costs of congestion	Already doing it	Expand this work to other modes, states, regions, US and globally; expand these analyses to TX!	No	Haven't started trying this yet, but I anticipate data acquisitions difficult due to proprietary concerns and jurisdictional issues
Mobility	Project	David Schrank, Tim Lomax, Bill Eisele	*Urban Mobility Report*	Currently unfunded; SWUTC for 2012 Report	Congestion rankings on 101 US urban areas	INRIX speed data; HPMS volume/inventory data; FAF commodity value	INRIX speed data in CSV format 73 GB; national shapefile in ArcInfo 2.6GB;	SAS, ArcInfo	Mobility and reliability measures, truck delay, costs of congestion	Already doing it	Note: we don't have funding for a 2015 "basic" report! We have multiple data analysis, method improvement and communication ideas if $ available.	No	Lack of time/money. We don't have secure funding to even produce a "basic" UMR in 2015.
Mobility	Project	David Schrank, Tim Lomax, Bill Eisele	*100 Most Congested Roadways List*	TXDOT IAC	Congestion rankings on 100 most congested roads in TX	INRIX speed data (or whomever wins the competitive RFP); TxDOT Roadway inventory	INRIX speed data in CSV format ~20 GB (2 files mixed-vehicle and truck-only); Texas shapefile in Arcinfo 275 MB;	SAS, ArcInfo	Mobility and reliability measures, truck delay, costs of congestion	Already doing it		No	
Mobility	Project	David Schrank, Tim Lomax, Bill Eisele	*Statewide Performance Measurement for MAP-21*	Within TxDOT IAC for Texas100	TTI responsible for assembling/computing all MAP-21 performance measures and targets	INRIX speed data (or whomever wins the competitive RFP); TxDOT Roadway inventory; pavement, bridge, safety info	Similar to 100 Most Congested Roadways Projects	SAS, Excel, ArcInfo	TTI "does the math" for congestion-related measures and obtains safety, bridge, pavement, from other TTI/TxDOT sources	Already doing it		No	
Mobility	Project	Ed Hard	*Tyler Cell, Bluetooth, and GPS External Survey Study*	TXDOT IAC	Develop External O&D trip data and matrices using BT, cell, and secondary GPD data and compare results between all technologies	Cell - AirSage, GPS - private sector, BT - TTI(Hou)	MBs and GBs, not big, but derived from big data	PostgreSQL, Postgis, Cube Analyst, R, ArcGIS	Depending on results, method/ study may be added to TxDOT Survey Program	Summer-Fall 2014		No	Confidentiality agreements. Need to write programs to link anonymized location-based data with attribute data
Mobility	Project	Ed Hard	*Analysis of Private-Sector GPS Data for OD Analysis*	TxDOT IAC	Analyzed the raw GPS data for OD analysis.	GPS - private sector	20GB/Day	R, Python		Fall 2013	Improve the algorithm, Expedite the processing time	No	Non-disclosure agreement.
Mobility	Project	Ed Hard, Praprut Songchitruksa (researcher)	*Using private-sector GPS data for O-D matrices, developing algorithm to scale-up datasets*		Mining private-sector GPS data for building OD matrices, rebuild trip chains from anonymized GPS data sources	private data	20GB/day	R, Python	current datasets (at TTI?) are too small to be "big data"	6 months	Improve the algorithm, Speed up the processing time	No	NDA in effect.
Mobility	Project	Juan Villa	*fluidity measures at border crossing*	TxDOT, Arizona DOT	develop performance measures for fluidity measures	border crossing time and volume data							INRIX covers Canada, not Mexico
Mobility	Project	Minh Le, Maarit Moran, Stacey Bricka	PRC - Congestion Footprint	PRC	Speed and travel time data used to test effect of state employee travel on congestion in Austin area	INRIX	Updated daily, web-based	INRIX Data Analytics tool, web-based		Current, 2014	Dataset provides a lot of information for various uses, this project timeline was too short for additional exploratory efforts. Also, study was limited by access to only 2-4 years of historical data.		Time
Mobility	Project	Rajat Rajbhandari	*DalTrans Annual Report*	TxDOT	Annual performance report of the transportation management system.	TxDOT ITS Data archived by TTI		SAS, excel	volume, speed, incident data	Annual Calendar Year Summary			additional funding
Mobility	Project	Shawn Turner	*Urban Congestion Report*	FHWA	Producing quarterly congestion reports for urban areas over 1M population	NPMRDS from FHWA	multi-GB each month. National shapefile in ArcInfo	SAS, Excel, ArcInfo		About to start	create reliability performance measures	No	
Mobility	Project	Tony Voigt	*TranStar Annual Report*	TxDOT	Annual performance report of the transportation management system.	TxDOT ITS Data archived by TTI.		SAS, excel	volume, speed, travel time, and incident data	Annual Calendar Year Summary			additional funding
Mobility	Project	Yatin Rathod, Stephen Ranft, Jason Crawford	*SH 161 Peak Hour Travel Lanes Project (Shoulder Running)*	TxDOT-IAC	*Comparing Bluetooth Travel time data, Traditional floating car travel time and Google Map traffic information*	Bluetooth data using portable units, manual travel time runs, Google map traffic captured images	> 1 million, txt files	MySQL, PHP	Converted Bluetooth data using Google Map API's location services data, converted the speeds data in graphic format using the color codes Google Maps is using.	Working on it since last year	Plot the Bluetooth O-D data to identify issues like queue jumping based on the readers location and vehicles detected at the location.	No	
Mobility	Project	Rafael Aldrete, Swapnil Samant, Jeff Shelton	*El Paso Transportation Data Warehouse*	CIITR	Archive regional transportation information e.g. Vehicle detector data, incident data, weather, border crossing times etc. Correlate this information for better traffic condition prediction	Vehicle detectors, social media, NOAA, City of El Paso website, RFID sensors, Bluetooth Sensors	variable	SSIS, SSAS		ongoing		yes	

Focus Area (assigned)	Activity (Project, Proposal, Idea)	Name of PI	Project Name/Topic	Sponsor	Short Scope Summary	Data Source	Data file size, file format	Analytic software tools (R, SAS, MySQL etc..),	Other Comments	Timeframe for working with this data (already doing it, < 1 year... etc)	What (else) could you do with this data if you had the funds and the software?	Would you need help in identifying the appropriate analytic tools that need to be developed/located? (Yes/No)	What's keeping you from fully leveraging this data?
Mobility	Project	Mullins, Andy	*Passive data in travel model development (research question: can passive cellular data sets be used to enhance the data input into the modeling process or used for calibration/validation applications.)*	TTI / National Centers Project	Establish benchmark of current uses of passive data in modeling, identify strategic research opportunities and topics for using passive data in current form and evaluating potential to obtain modeling data items	Lit review including 80+ MPOs, survey of practitioners and practitioner workshop						Yes	
Mobility	Idea	Byron Chigoy	*Data Review of Cell/GPS Data*	TxDOT	Review approx. 250 GB of point location data for the entire US and its suitability for integration into urban region external travel analysis	Cell/GPS	CSV; One minute files for one week	Python, PostgreSQL	n/a	n/a	Speed analysis; cut-through analysis, peak period analysis	Yes and No. The analytic tools are not the problem - getting adequate server infrastructure is. Would need dedicated racks and integration of open-source software using Linux or OSX rather than Windows based architecture.	funding
Mobility	Idea	Byron Chigoy, Tom Williams	*FHWA BAA problem statement and proposal*	FHWA	Evaluate using cell phone OD data to develop near term corridor and subregional forecasts	Cell/GPS	n/a	PostgreSQL, Postgis, Cube Analyst	Future opportunity to pursue	at least a year	n/a	n/a	n/a
Safety and Operations	Project	Minh Le and Bryan Miller	*Daltrans Detector Warehouse project.*	TxDOT (Dallas)	Roadway detector data is archived and QAQC checks are done on the fly to monitor system health and to calculate Performance Measures. One component is to integrate incident, weather, and maintenance data.	Daltrans (existing), Lone Star (future)	1 million records per day, 20 GB per year	SQLAzure, TTI custom web-based tool, Lone Star		Base database completed. Web-based tool in development.	Detectors could be configured to differentiate between cars and truck traffic with some minor calibration. Could also data mine incident and detector data to develop crash prediction tool based on prevailing conditions.		
Safety and Operations	Project	Akram Abu-Odeh (and others at Roadside safety division)	*simulations*	Multiple projects for states DOT, FHWA and private sponsors	predictive and validation simulations for different impact scenarios	Created from the simulation codes	multiple files ~100/ case. Each file around 150 ~250 MB, total around 50 GB/scenario	LS-OPT , HyperStudy, HyperView and LS-PREPOST	the solvers (data generators) are usually on HPC	we typically use 16 cores per job. However, we used 64, 128 cores before and 32 cores will be out typical lower end in the near future	perform DOE and optimization studies and machine learning	no	time and competition for resources
Infrastructure	Idea	David Ellis	*Developing performance metrics for public-private partnership projects*		In recent years, there's been a growing trend toward public-private partnerships as one method for enabling the construction, operations, and maintenance of transportation infrastructure. Having access to comprehensive dataset of these projects could lead to answering important questions in the future.	Private: Public Works Financing Major Project Database, Information Group P3 Database, World Bank Private Participation in Infrastructure, The Guardian PFI Contract Full List Database, European P3 Expertise Center, Electronic Municipal Market Access; Public: FHWA, State DOTs, RMAs (Texas), Regional/Local Tolling Authorities	n/a	n/a	This project may or may not be considered "big data" but it would nevertheless position TTI as the data source for this information. Potential partners for this effort include: Partnerships, British Columbia, Inc. and George Mason University's P3 Policy Center.	at least two years (long enough to fund a Ph.D. dissertation!)	Develop performance metrics for transportation projects	Yes	Funding! Access to private sector data sets
Safety and Operations	Project	Kevin Balke, Praprut Songchitruksa (researcher)	*prototype testing of speed harmonization and queue warning algorithm*	Battelle IDIQ	Develop and test the algorithms the speed harmonization and queue warning using connected vehicles	simulation data	10x per second, 2-3 GB's per hour	Vissim, Python		3 months		No	
Safety and Operations	Project	Srinivasa Sunkari, Praprut Songchitruksa (researcher)	*EAR Project, building platform for simulation of CV components*	FHWA	Develop simulation components for hardware-in-the-loop testing of connected vehicle applications	simulation data	TBD	Vissim, Python, C++	The challenge is processing speed within simulation at high-volume conditions.	9 months	Increase the simulation performance	No	
Safety and Operations	Project	Steve Venglar	San Antonio Inter-agency Contract (IAC)	TxDOT, San Antonio District	Process 2013 TxDOT TransGuide data; support IAC work if extracted data is reliable	2013 sensor data for traffic management centers (San Antonio)	35 GB, ".CSV" files	Lone Star	data: incidence logs, speed, traffic counts	Underway	TBD; analysis stops if data quality is poor	No	Data archiving not a top priority for Texas, historically. TTI expertise exists, data is hard to acquire
Safety and Operations	Idea	Rajbhandari, Carlson, Miles	*LiDAR data is used in automated vehicle, retroreflectivity, asset inventory*	Several		LiDAR data stream and archived data	Giga and terra bytes		Future opportunity	Have a data set for Riverside and a section of SH47	Plenty of research for TxDOT in retroreflectivity and asset inventory as well as automated vehicle testing	Yes, mostly in automated feature extraction	LiDAR equipment is expensive and we don't have the expertise to post process the data
Safety and Operations	Idea	Yatin Rathod	*Crash Analysis System*		Developing a crash analysis tool which correlates crash location, crash time with the weather data, pavement data and geometric features of the roadway	1)Crash data from CRIS or any other source which has the crash location in lat/long format. 2)Historical weather data at that time and location . 3) Geometric data (curves, curbs etc and pavement data from highway maintenance database	Possibly very large amount of data	Probably need BigQuery or similar software	This can be offered as a Software as a service (SaaS) where we develop and sell the service to other agencies.		Once we have all the data in a queryable format, sky is the limit :)	Yes	
Safety and Operations	Project	Jeff Shelton, Sharada Vadali, Valdez, Chandra	*Extreme Event - Critical Failure Economic Costs*	El Paso MPO		Simulation data and External data	24 hour simulated vehicle trajectory data - 10 GB	DynusT	Need classification of "big data".	Already doing it	Funds: More work on resiliency and the model	No	The temporal aspect is an advantage but need to understand this better.
Policy	Project	David Schrank, Tim Lomax	*TOSTADA- Tool using Stacked Data*	PRC/ Texas Legislature	A map layering tool to address various types of data in overlapping spatial view.	TxDOT/TTI congestion data, RhiNo traffic volumes, INRIS speed data, pavement, bridge data and truck commodity value data		GIS		Pilot completed 2014			
Safety and Operations	Project	Wunderlich	*Crash and hospital data correlation*	TTI									