Big Data for TIM – Big Opportunities, Big Challenges

Kelley Klaver Pecheux
AEM Corporation

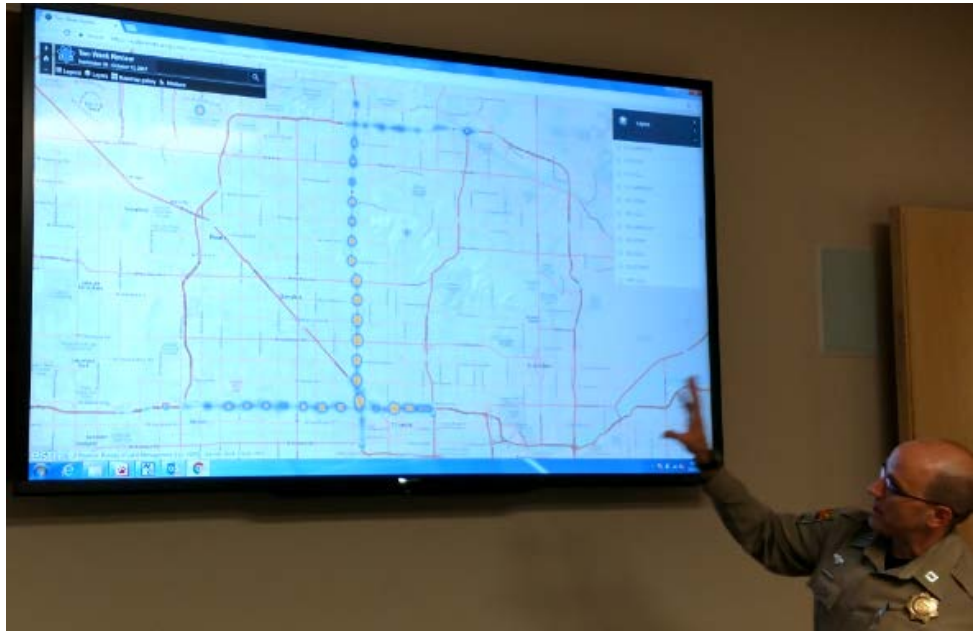
Penn State Traffic Engineering and Safety Conference
December 6, 2017

NCHRP 17-75 RESEARCH OBJECTIVE

- Develop guidelines that illuminate the concepts, opportunities, data sources, applications, analyses, options and challenges associated with the use of Big Data for TIM agencies and the opportunities for advancing the state of the practice.

OVERVIEW OF PRESENTATION

- State of the practice
  - TIM
  - Use of data for TIM
  - Big Data
- Potential Big Data opportunities for TIM
- TIM-relevant data source assessment
- Big Data challenges for TIM
- Summary and next steps

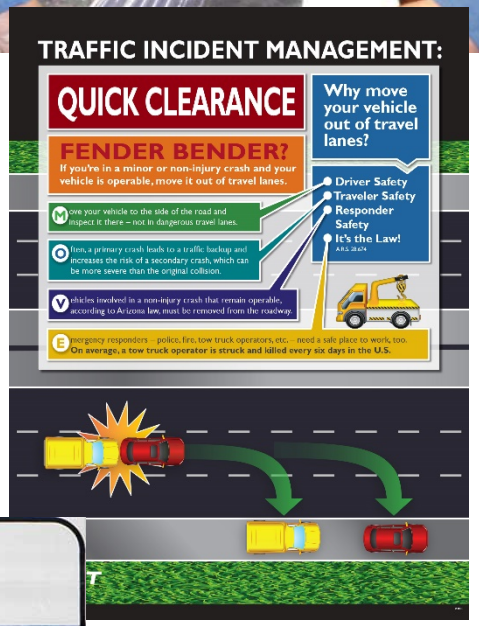


State of the Practice

STATE OF THE PRACTICE

TIM – Lots of progress

- Local, regional, and statewide TIM committees
- TIM legislation
- National TIM Responder Training
- TIM strategic plans
- Agency operating agreements
- Agency policies for safe and quick clearance



STATE OF THE PRACTICE

Use of Data for TIM

Performance Measurement/Management

- For most states, very basic data collection and analysis.
- Collecting data through TMCs, SSP programs, statewide crash report, ATMS-CAD integration (limited).
- Systems are usually not integrated or compatible.
- Lots of missing data due to manual collection, coverage (locations, hours)
- Quality generally low.

Monetization of Benefits

- Due to lack of data, most states do not (cannot) monetize benefits.
- Challenging for even those states that do have some data (no baseline data, lack other necessary data) – questionable methodologies, ballpark estimates.

STATE OF THE PRACTICE

Big Data – What is Big Data?

- Extremely large datasets (generally national or international data sets):
  - Hundreds of millions of Facebook posts every day (posts, photos, videos, livestreams).
  - HERE – 80,000 data sources across 54 countries.
- Data that cannot be stored on even the biggest server available on the market.
- Data that can be analyzed using Big Data methodologies/tools (which do not perform well on “small” datasets) as opposed to traditional tools (RDBMS).
- Single approach (DOT, contractor) not enough to tackle data (takes many eyes).
- Data that must be processed using a shared-resources model (i.e., the cloud) - unless you can afford a “super computer.”

DATA SIZE EXAMPLES

Big Data

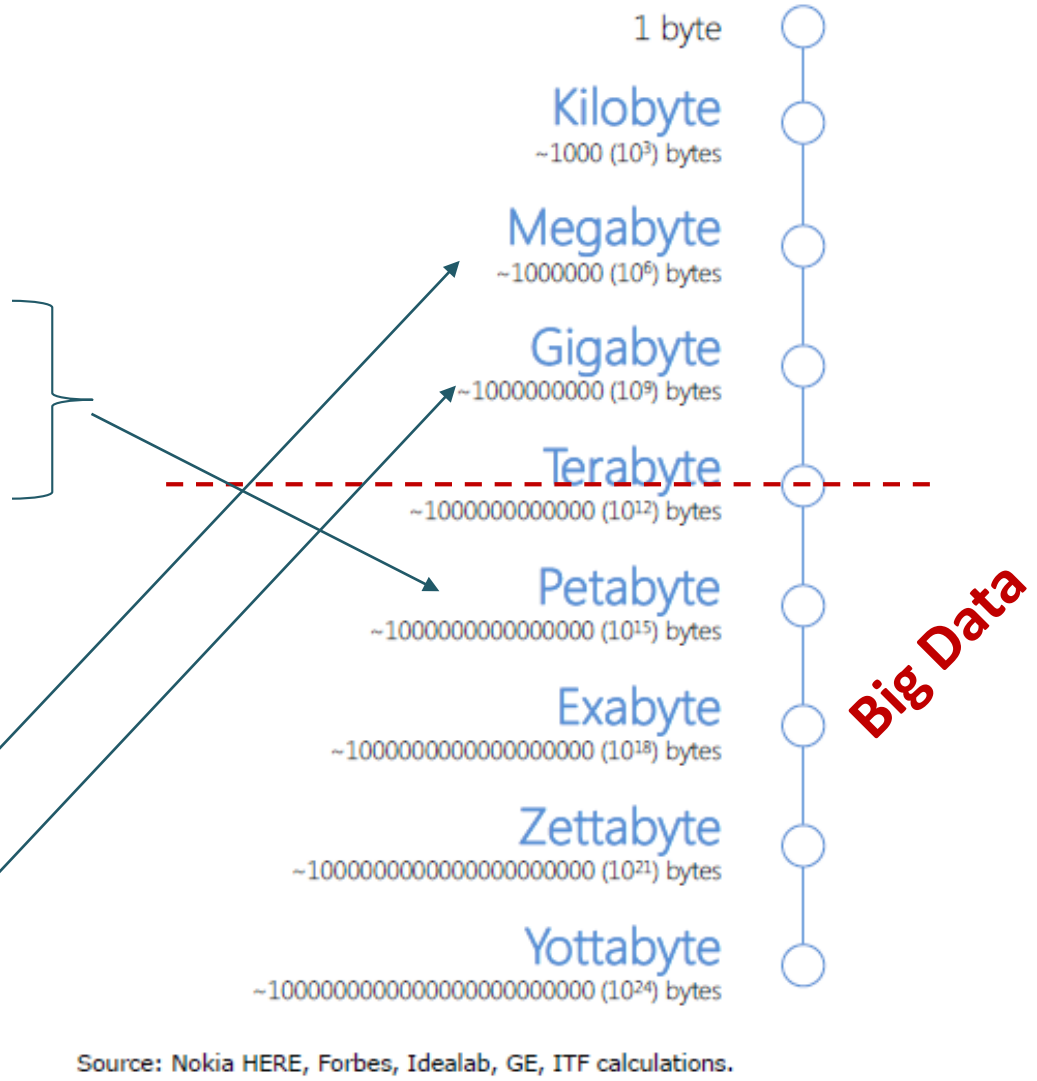
- Continuously changing but generally in excess of 1 terabyte.
- eBay uses two data warehouses at 7.5 petabytes.
- Walmart handles more than 1 million customer transactions every hour, estimated to contain more than 2.5 petabytes of data.

Incident/TMC-Related Data

- 5 years of statewide crash data from Florida = < 500 MB
- 300 TMC field devices 5.8 MB per day
- 30.2 million records from 15K EMS agencies in US in 2015 = GB
- 300 CCTVs = 4 GB per day

Emerging CV Data Estimates

- All the BSM and PDM data from 200,000 CV = ~840 GB per day (300 TB per year)



Data Size Scale

STATE OF THE PRACTICE

Big Data - The Move From Traditional Data Analysis to Big Data Analytics

- Traditionally used tools (RDBMS, ETL, BI, statistics) to preprocess, store and analyze, and visualize data.
- Data is now too big, too diverse, and too messy for these tools to work.
- Big Data tools were specifically developed to process Big Data datasets; but no unique solution, rather an entire ecosystem.
- Big Data is saved first “as is,” then enriched, then analyzed and visualized.
- Continually evolving ecosystem.

STATE OF THE PRACTICE

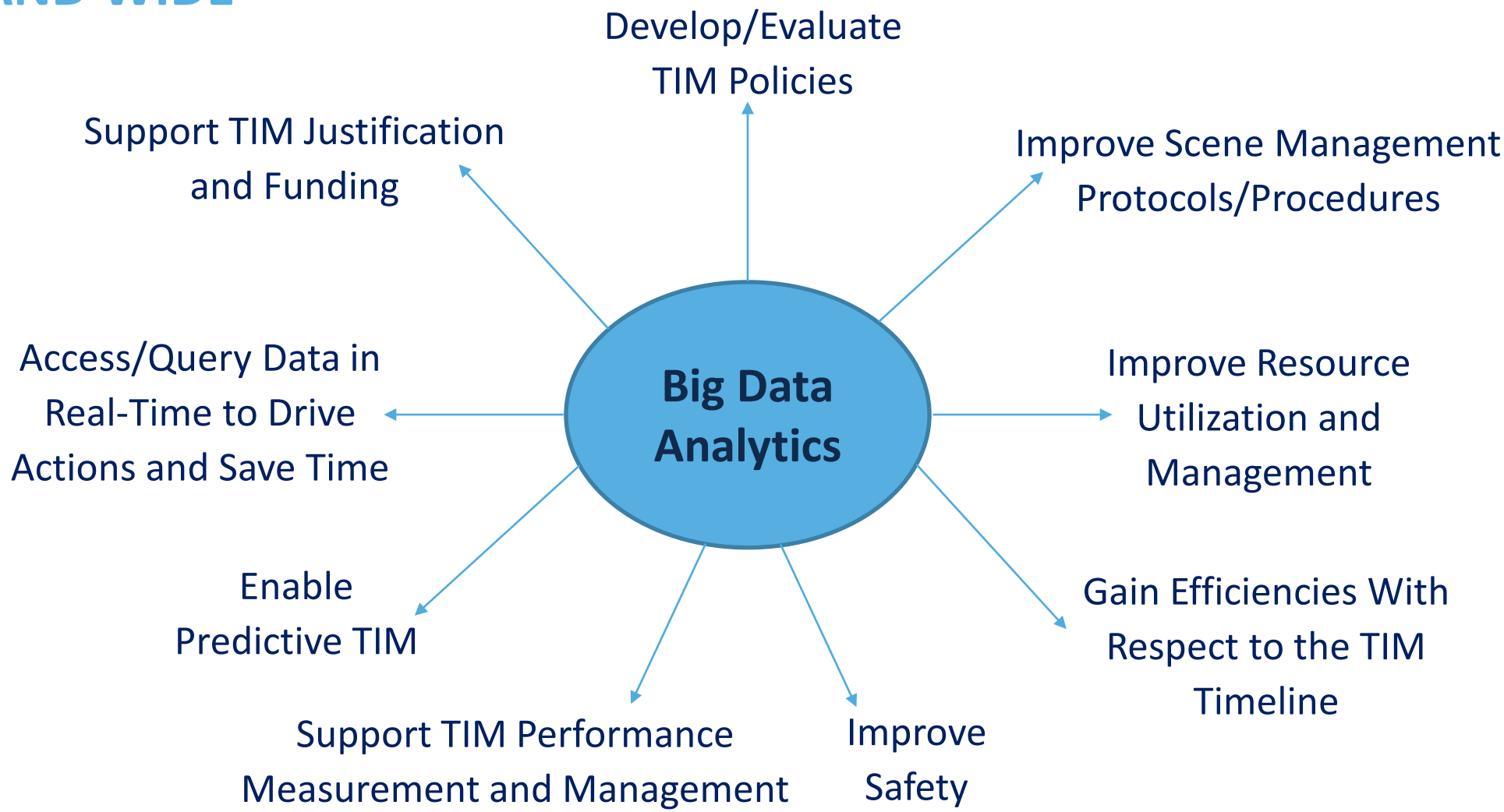
Big Data in Transportation – Not Native

- Transportation Planning:
  - Use of cellular data (e.g., call detail records).
  - Ability to track movements of vehicles and people at microscopic level.
  - Improve analysis of complex travel behaviors and patterns.
- Real-Time Traffic Conditions:
  - Use of cellular location and citizen reporting to collect real-time traffic conditions (speeds) and incidents (HERE, INRIX, WAZE).
  - Ability to visualize traffic conditions in real-time at a low resolution.
  - Ability to detect incidents early (before 911 call for non-major incidents).

Potential Big Data Opportunities for TIM

POTENTIAL BIG DATA OPPORTUNITIES FOR TIM

FAR AND WIDE



TIM Big Data Opportunity Examples

- Evaluate impacts of different scene management strategies (e.g., vehicle positioning, temporary traffic control device placement).
- Evaluate effectiveness of National TIM Responder Training (where it works and why).
- Optimize current and assess candidate SSP routes (days, hours, appropriate number and types of resources).
- Identify actions that reduce intervals along the TIM timeline.
- Better understand when, where, and why secondary crashes occur.
- Evaluate emergency vehicle lighting and conspicuity – better understand how approaching motorists behave given different stimuli.

Big data provides the granularity in the data and analytics that we don't have with traditional datasets and analytical tools.

TRADITIONAL VS. BIG DATA APPROACHES

Traditional Approach – Slow Moving

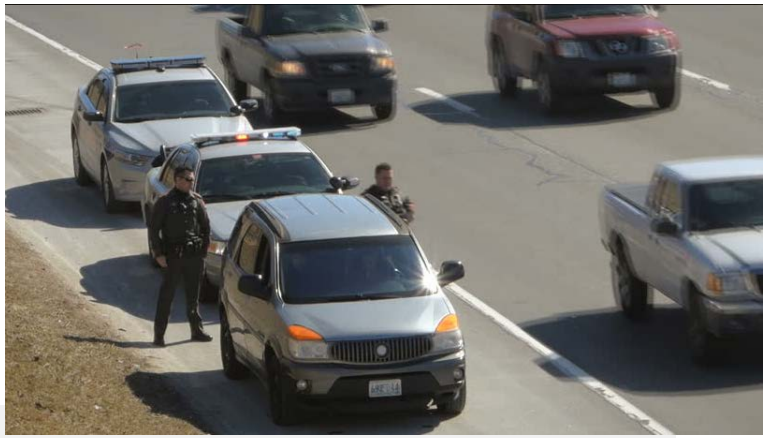
- Relies on people and processes.
- Based on samples of data / limited observations.
- Qualitative approaches, which can be subjective (e.g., interviews, manual review of crash reports).
- Can be manual, tedious, and resource intensive.
- Capture the data, run analysis, change standard operations/procedures.
- Do the study once, not usually repeated, results widely applied.

Big Data Approach – Fast Moving

- Relies on data.
- Based on large, expansive sources of (population) data.
- Many more opportunities to explore and analyze data – tens of thousands of variations of analyses.
- Analytics can be repeated over and over again as new data is available (refine the model in real-time).
- Reduces the subjectivity of analyses.
- Designed to be actionable right away.

FLORIDA SAFETY EXAMPLE

Objective – Assess the effectiveness of the Florida Move-Over Law



Traditional Approach

Methodology

- Involved a civilian research vehicle, a marked police vehicle, video recordings, and measurement of passing vehicle speeds.
- Observed 9000 right-lane vehicles passing staged police stops on selected Florida freeways – recorded speed and lane changing behavior.

Limitations

- Field experiment.
- Resource intensive.
- Risk exposure (“secondary” crashes).
- Limited observations (sample data) from limited roadways/corridors.
- Extrapolation of findings to other roadways/corridors across the state.

FLORIDA SAFETY EXAMPLE

Objective – Assess the effectiveness of the Florida Move-Over Law

Big Data Approach

Methodology

- Tap a wide variety of nationwide data sources relevant to the study.
- Use Big Data analytics (e.g., clustering) to assess compliance rates and to identify the factors that affect compliance.

Data Needs

- Telematics/AVL data (e.g., locations, speeds, lateral positions of vehicles, locations and warning system activity of response vehicles).
- Specification of response vehicles.
- Traffic volumes.
- Roadway inventory data.
- Weather data.
- State related info (driver outreach).

OREGON PERFORMANCE MEASUREMENT AND MANAGEMENT EXAMPLE

Objective – Address why there were 1,088 incidents where the mutual ODOT/OSP RCT goal of 90 minutes was exceeded (2014)

Traditional Approach

Methodology

- Engaged stakeholders.
- Developed anecdotal list of factors that contribute to longer clearance times.
- Manually reviewed individual crash reports and categorized incidents.
- Developed and implemented actions to address most common causes.

Limitations

- Manual
- Tedious
- Subjective
- Resource intensive
- Done once, not repeated often

OREGON PERFORMANCE MEASUREMENT AND MANAGEMENT EXAMPLE

Objective – Address why there were 1,088 incidents where the mutual ODOT/OSP RCT goal of 90 minutes was exceeded (2014)

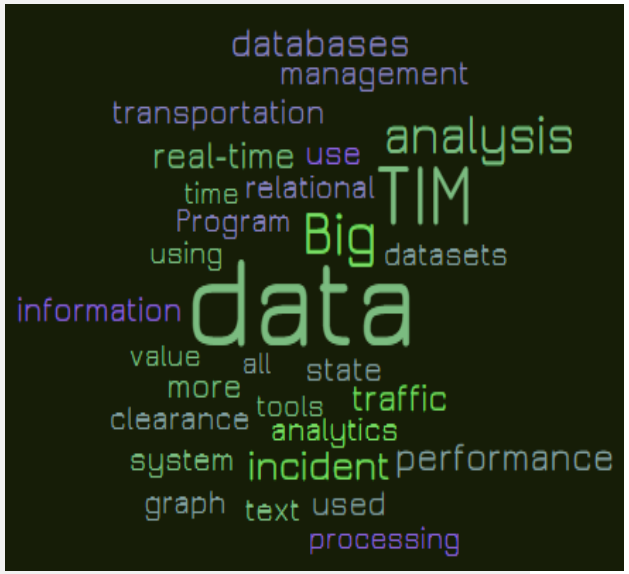
Big Data Approach

Methodology

- Text and cluster analysis on nationwide (or at least multi-state) crash data.
- Automatically identify causes based on data.
- Find correlations.
- Offer additional and likely unexpected insights into causes.

Data Needs

- Nationwide (multi-state) crash data
- CAD data
- Weather data
- Vehicle data
- Roadway inventory data



TIM-Relevant Data Source Assessment

30 DATA SOURCES ASSESSED

- **State Traffic Records Data** (crash, vehicle, driver, roadway, citation/ adjudication, injury surveillance).

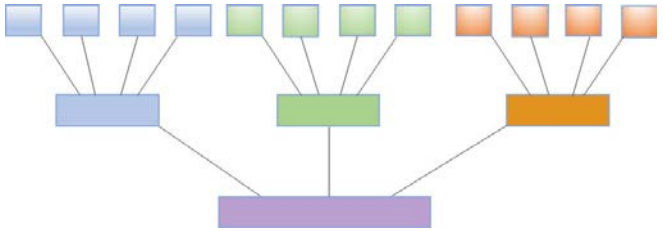
- **Transportation Data** (traffic sensors, video, safety service patrol, road weather, 511 system, toll).

- **Public Safety Data** (emergency responder CAD, 911/PSAP, digital video, towing/recovery).

- **Crowd-Sourced Data** (Waze, Twitter).

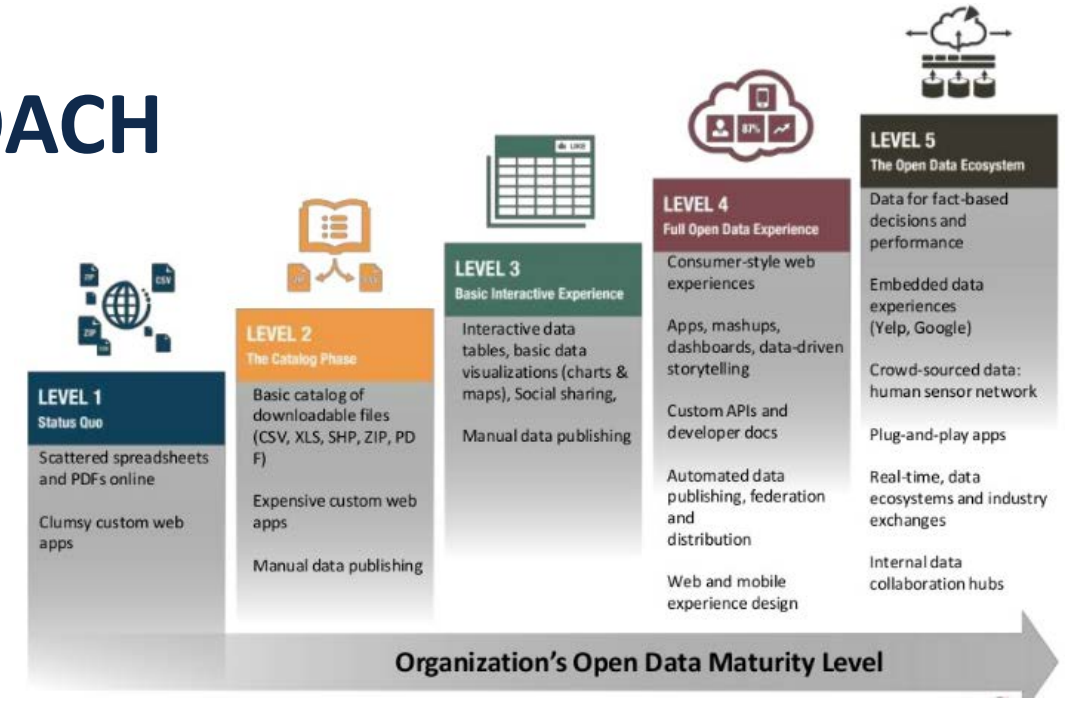
- **Advanced Vehicle Systems Data** (AVL, EDR, vehicle telematics, automated and connected vehicles).

- **Aggregated Data Sets** (HERE, INRIX, NEMSIS, NPMRDS, RITIS, MADIS, third-party weather service, MCMIS).



DATA SOURCE ASSESSMENT APPROACH

- Criteria:
  - Structure
  - Size
  - Storage and management
  - Accessibility
  - Sensitivity
  - Openness
  - Challenges
  - Costs
- Data Maturity Assessment Approach:
  - Open data maturity
  - Data readiness



Center for Data Science & Public Policy
THE UNIVERSITY OF CHICAGO

Data Maturity Framework
Data and Tech Readiness Scorecard

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

DATA SOURCE ASSESSMENT FINDINGS

- Lack of readiness and/or openness of data to support TIM Big Data analytics – extensive amount of work needed to ready data for use.
- Lack of volume of data at a statewide level to constitute Big Data.
- Big Data sources that might be leveraged for TIM at this time come with limitations:
  - Waze – noisy, dirty, false info, typical crowd-sourced data set – need to learn how to derive value from it and be supported by other datasets to be useful.
  - Speed data (HERE, INRIX) – *if* not aggregated – valuable at ends of TIM timeline (T_0 and T_7).
  - AVL data – not usually open/accessible (restricted).
  - Video – Has potential but currently offers limited value (usually low resolution, not archived, limited roadway coverage, video analytics in field not reliable).
  - Weather – *if* at low level geographically and timely (ancillary dataset - required but not valuable on its own).
- CV will eventually generate Big Data – but privacy/legal hurdles.

Challenges

BIG DATA CHALLENGES FOR TIM AGENCIES

- Cultural, institutional, and policy barriers.
- Scattered data.
- Data accessibility.
- Data format.
- Quality, completeness, and usability of data.

BIG DATA CHALLENGES FOR TIM AGENCIES

Cultural, Institutional, and Policy Barriers

- Lack of data culture at leadership level (understanding that data is important)
  - Still operate in a world of limited data and slow changes with hierarchical structure (rely on people and processes).
- Hesitancy or resistance to sharing/opening data:
  - Don’t see the needs / benefits.
  - Fear of being judged poorly / reflected on badly.
- State-level policies against use of the cloud:
  - Misunderstanding at decision-maker level of difference between IT and data (data should be kept close to business processes, IT can be outsourced).
- Lack of technical expertise.

BIG DATA CHALLENGES FOR TIM AGENCIES (CONT'D)

Scattered Data

- Marked separation of business areas (with a DOT) – independent collection of data without thought of combining or sharing with others.
- Variety of TIM-related data = scattered data, owned/operated/restricted by different groups/organizations within and outside of the DOT (state and local public safety agencies).
- Huge challenge for building a data lake even when leveraging big data technologies designed to merge heterogeneous datasets.

BIG DATA CHALLENGES FOR TIM AGENCIES (CONT'D)

Data Accessibility

- Physical – designed for access to sample data by individuals (DVD, USB, FTP, web download) rather than giving access to large datasets by machines (cloud, data stream).
- Legal – public records laws restrict the use and distribution of some data (e.g., event data recorders).
- Unwillingness/inability of agencies to share data or to accept data from others:
  - Security (the threat that data could be stolen, compromised, corrupted, infected).
  - Privacy (sensitive, containing PII).
  - Proprietary (restricted because considered intellectual property or the basis for competitive advantage).

BIG DATA CHALLENGES FOR TIM AGENCIES (CONT'D)

Data Format

- Lack of consistency:
  - Discrepancies in ways to communicate the data (RSS, FTP and HTTP download, API and streaming).
  - Discrepancies in file format used (Excel, PDF, CSV, JSON, XML).
  - Discrepancies in the frequency with which data is aggregated and updated.
- Lack of standards (communication, storage, and retrieval of different datasets across and between organizations).
- Increases the difficulty of merging disparate data and creating and maintaining comprehensive data sets.

BIG DATA CHALLENGES FOR TIM AGENCIES (CONT'D)

Quality, Completeness, and Usability of Data

- Data not homogenous – all over the place in all three categories.
- Captured in a way that is not equivalent (TMC, crash reports, CAD).
- Retention – technical, storage, or policy affecting how long data can be stored or archived:
  - Most state DOTs do not retain video data for more than a few hours or days often because of the cost associated with its storage and the possible associated liability risk.
- Inherent rarity and variability in traffic incidents – presents challenges in developing sufficiently complete historical datasets capable of sufficiently characterizing both incidents and response.

Summary and Next Steps

SUMMARY

- Lots of improvements in TIM over the years using different approaches.
- Logical next step to take TIM to the next level is by making more effective use of data, particularly Big Data (i.e., more data and more advanced analytics).
- Current state of the practice in using data for TIM – very limited ("old" approach to data collection and use, many diverse and scattered data sets, huge data gaps).
- State of the practice in Big Data offers tremendous opportunities for TIM.
- TIM-relevant datasets not in a state of readiness and/or openness to be utilized for Big Data.
- Major challenges in bridging the gap between current state of the practice of TIM (DOTs) and current state of the practice in Big Data.
- Not a show-stopper, but a lot of work and changes required.

NEXT STEPS

- Develop Guidelines on Big Data for TIM Agencies (in process).
- Develop checklist of capabilities necessary for Big Data for TIM (DOTs).
- Project ends June 1, 2018.

Thank you

Kelley Klaver Pecheux, Ph.D.

AEM Corporation

Kelley.Pecheux@aemcorp.com