NCHRP 17-75: Leveraging Big Data to Improve TIM

NOCoE Webinar

December 13, 2018

Kelley Pecheux, Ph.D.

Benjamin Pecheux

Grady Carrick, Ph.D.

Overview of Presentation

- Objective of NCHRP 17-75
- Big Data Definition and Examples
- Overview of Big Data Guidelines for Transportation Agencies

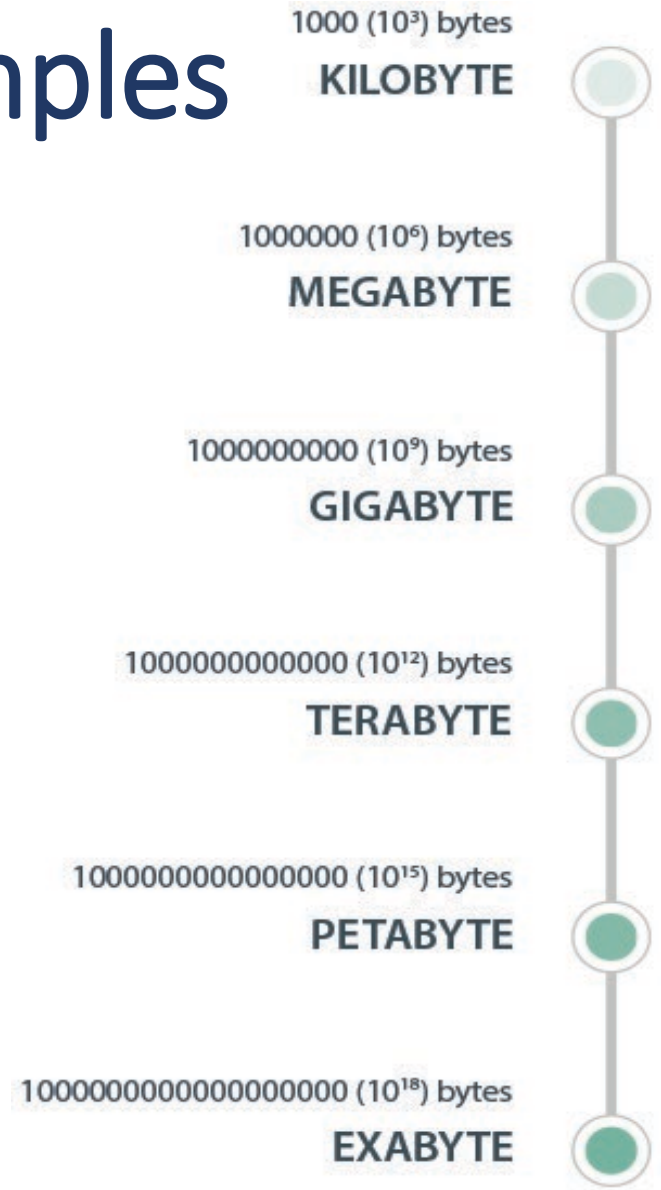
NCHRP 17-75 Objective

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.

What is Big Data?

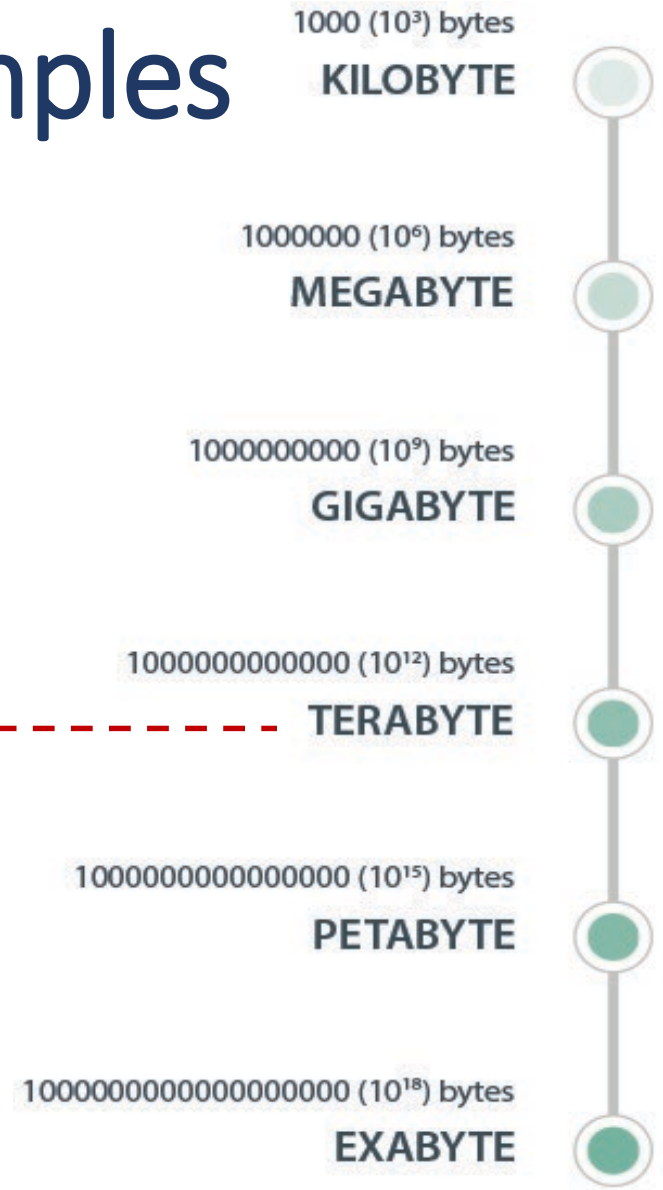
- Extremely large datasets – generally national or international data sets
- Diverse datasets – data sets that include different types of data from many sources
- Data that exceeds the capacity of the biggest server available on the market
- Data that is processed using a shared-resources model (i.e., the cloud) – cloud came out of the need for big data (it is too cost prohibitive to build the infrastructure in-house)
- Data that can be analyzed using big data technologies/ methodologies (there is so much information that different types of tools are needed)

Data Size Examples



Data Size Scale

Data Size Examples



BIG DATA – continuously changing, but generally > 1 TB

Data Size Scale

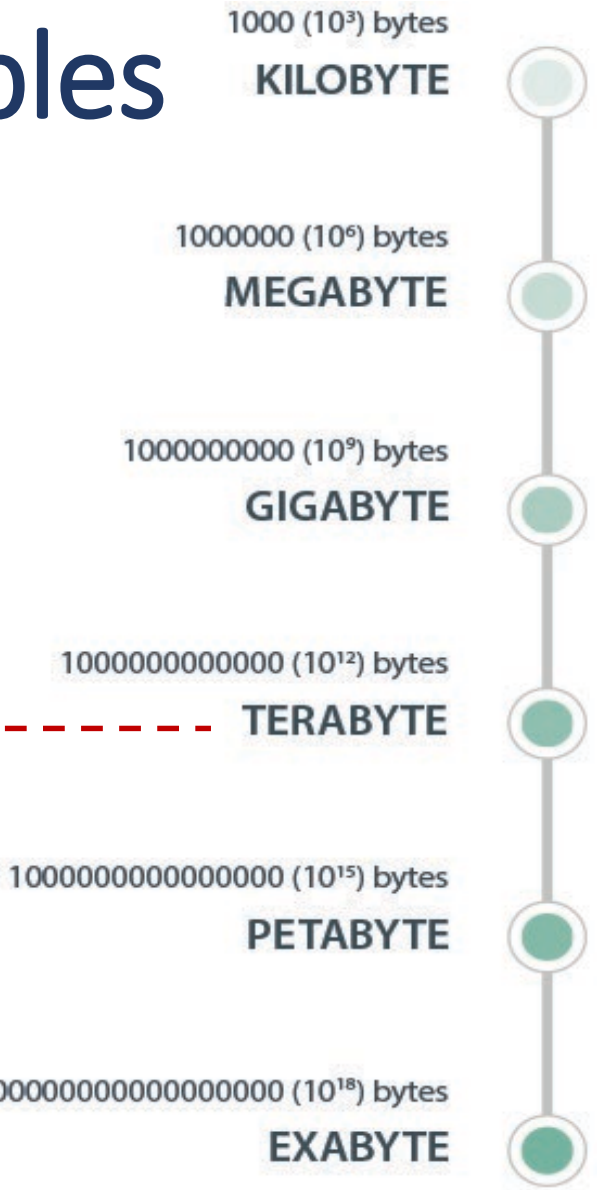
Data Size Examples

BIG DATA – continuously changing, but generally > 1 TB

- Walmart processes 2.5 PB of data every hour.
- eBay stored 90 PB of data in 2013.

Data Size Scale

Data Size Examples



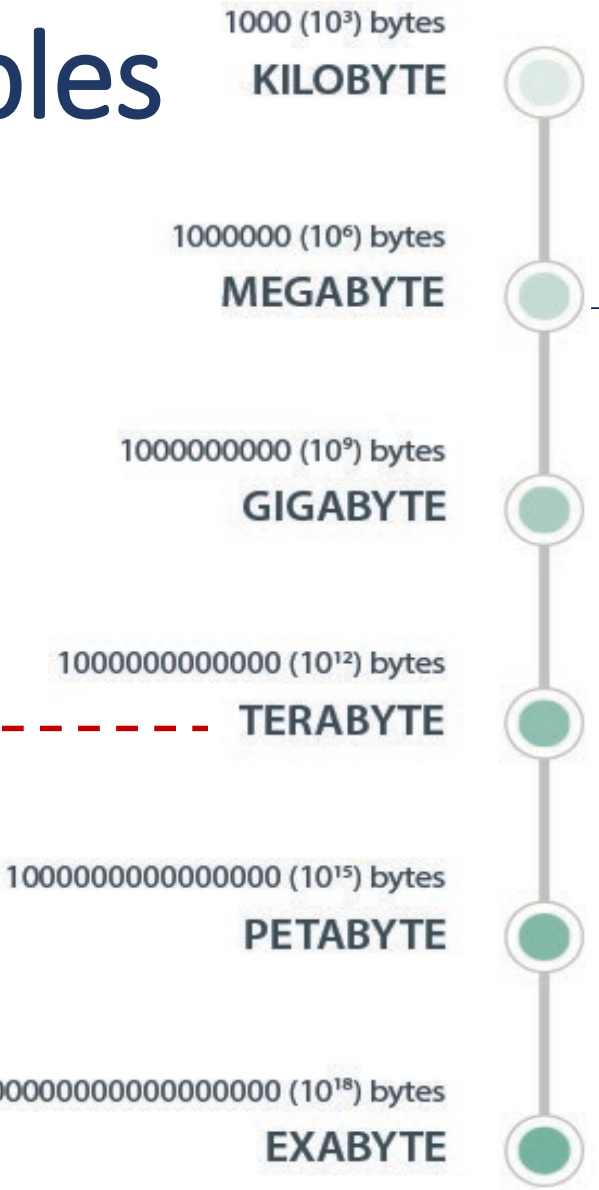
BIG DATA – continuously changing, but generally > 1 TB

- Walmart processes 2.5 PB of data every hour.
- eBay stored 90 PB of data in 2013.

Crash, TMC, and Emerging CV Data Estimates

Data Size Scale

Data Size Examples



Crash, TMC, and Emerging CV Data Estimates

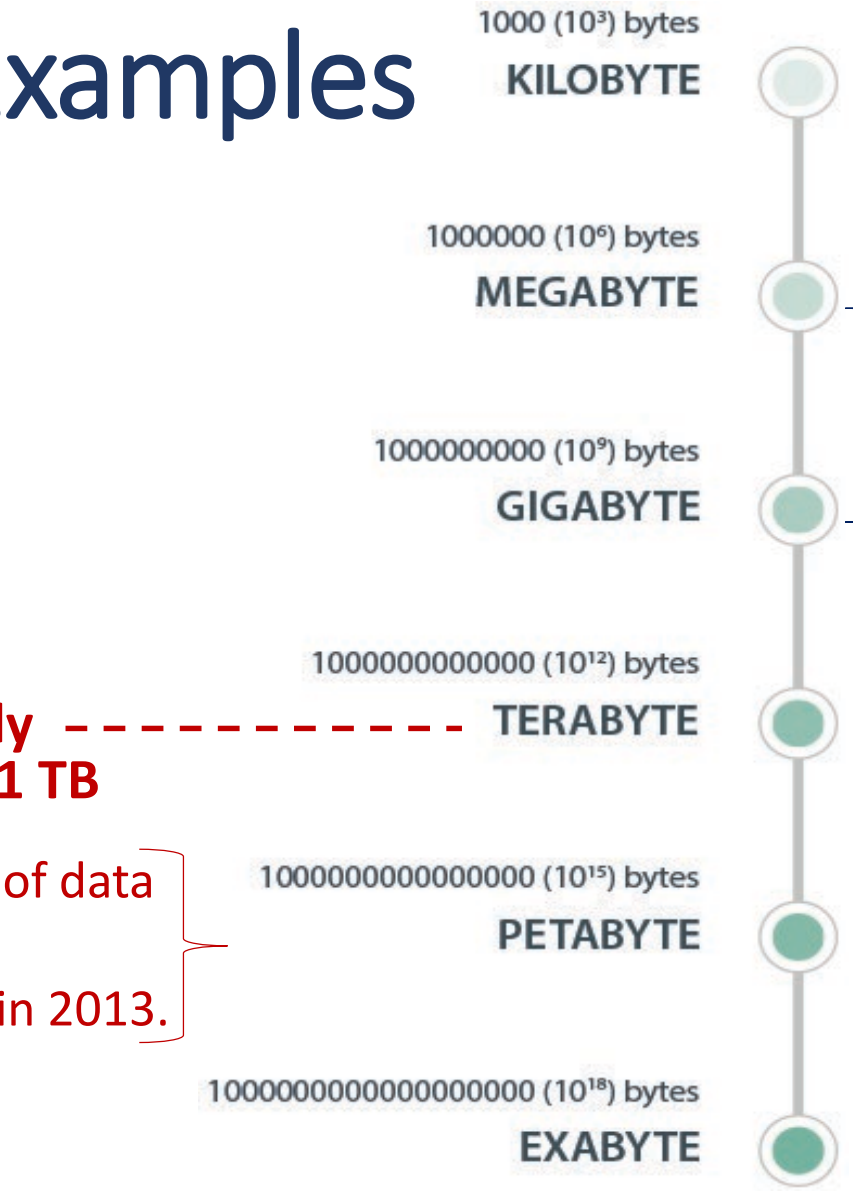
- 5 years of statewide crash data from Florida = < 500 MB

BIG DATA – continuously changing, but generally > 1 TB

- Walmart processes 2.5 PB of data every hour.
- eBay stored 90 PB of data in 2013.

Data Size Scale

Data Size Examples



Crash, TMC, and Emerging CV Data Estimates

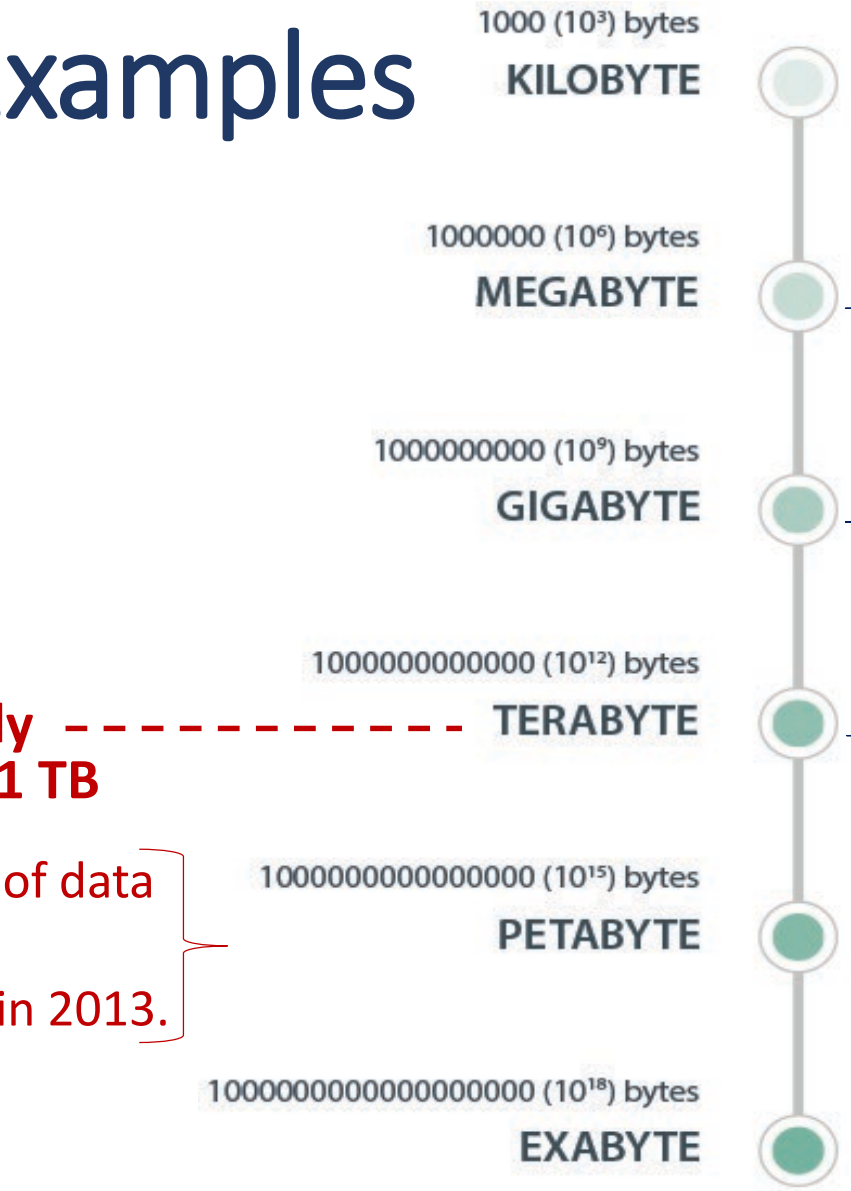
- 5 years of statewide crash data from Florida = < 500 MB
- 30.2 mill records from 15K US EMS agencies from 2015 = 40 GB
- 300 TMC field devices = 635 GB per year

BIG DATA – continuously changing, but generally > 1 TB

- Walmart processes 2.5 PB of data every hour.
- eBay stored 90 PB of data in 2013.

Data Size Scale

Data Size Examples



Crash, TMC, and Emerging CV Data Estimates

- 5 years of statewide crash data from Florida = < 500 MB
- 30.2 mill records from 15K US EMS agencies from 2015 = 40 GB
- 300 TMC field devices = 635 GB per year
- All BSM and PDM data from 200,000 CV = 820 GB per day (300 TB/year)
- 300 CCTVs = 1.2 TB per day (438 TB/year)

BIG DATA – continuously changing, but generally > 1 TB

- Walmart processes 2.5 PB of data every hour.
- eBay stored 90 PB of data in 2013.

Data Size Scale

The Move From Traditional Data Analysis to Big Data Analytics

- Traditionally data is preprocessed using an ETL process to fit an RDBMS schema and then analyzed using statistical software and BI.
- Data is now too big, too diverse, and too messy for traditional tools to work.
- Big data, on the other hand, is stored “as is” (not preprocessed to fit a restrictive schema).
- Big data tools have been specifically developed to process big data datasets (and do not work well on small datasets).
- Rapidly and continuously evolving ecosystem.
- This move is disruptive and requires a complete paradigm shift in data collection, storage, processing, and analysis, as well as in the use of data for decision making.

aem

Traditional vs. Big Data Approaches

Traditional Approach – Slow Moving

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

Big Data Approach – Fast Moving

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

Big Data Opportunities and Challenges for TIM / Transportation Agencies

- Many opportunities for big data to improve/advance TIM (gain efficiencies, develop/evaluate policies, improve resource utilization and management, improve safety, enable predictive TIM, etc.)
- However, challenges need to be overcome for transportation agencies to take advantage of big data opportunities
  - Data silos
  - Interoperability
  - Privacy
  - Public records laws
  - Data retention
  - Proprietary data
  - Issues associated with emerging forms of data
  - Technical expertise
  - Security
  - Fear of the cloud

aem

Big Data Guidelines for Transportation Agencies

Overview of Guidelines

1. Adopt a deeper and broader perspective
2. Collect more data
3. Open and share data
4. Use a common data storage
5. Adopt cloud technologies for the storage and retrieval of data
6. Manage the data differently
7. Processing the data
8. Open and share outcomes and products to foster data user communities

1. Adopt a Deeper and Broader Perspective

Develop big data as a collaborative environment

- Trust in the data for decision-making – Move away from reactive decision-making based on limited data; intuition; and personal opinions, experience, and business understanding.
- Expand decision making beyond a "chosen few" – Big data is too big, too complex, and too confusing to be tackled by only by a small set of individuals within an agency.
- Enable members from the lowest level to the highest level of an organization to observe and react on their own to changes detected within the organization's large pool of data.
- Given the infrequent nature of traffic incidents, a shared nationwide dataset is likely the ideal environment to develop such advanced data analytics.

2. Collect More Data

- Focus less on software and tools (they are readily accessible).
- Focus on the data, which is the most difficult, expensive, and valuable part of the analysis.
- The more detailed and extensive the data, the better – without enough detailed data, big data techniques cannot be leveraged to inform decision-making.
- "Collect" more data by augmenting internal datasets with external datasets:
  - Existing incident-related data (e.g., TMC, crash) alone is far from enough data to conduct big data analyses for TIM, and the resolution of the data is not at a low enough level.
  - Create large, detailed datasets by augmenting human-collected data with machine/sensor-collected data (e.g., weather, AVL and other external data sources (e.g., crowdsourced) to obtain a more complete and detailed description of incidents and responses.

3. Readily Open and Share Data

Both internally and externally

- Common roadblocks to data sharing
  - Public records laws
  - Proprietary data formats
  - Contract data clauses
- Benefits of data sharing

Definition of Open Data

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

4. Use a Common Data Storage Environment

- Move away from multiple, diverse data stores (i.e., data silos) within the organization.
- Big data datasets are too big to be easily moved in and out of storage without spending significant time and money.
- Big data datasets are so large that they need to be stored across multiple connected servers, known as “clusters.”
- To avoid cost-prohibitive solutions, co-locate datasets in a cloud environment.
- Refrain from using “data virtualization” technology (the process of aggregating data from different sources to develop a single virtual view of data without knowing its exact storage location).

5. Adopt Cloud Technologies for the Storage and Retrieval of Data

- Understand the cost savings of the cloud
  - Scalability
  - Agility
  - Affordability
- The cloud is more secure than you think
  - Redundancy
  - Security
  - Safe sharing
- Cloud is ideal for big data analytics

Due to their scalability, agility, affordability, redundancy, and safe sharing, cloud technologies offer organizations substantial cost savings and improved security, and they are ideal for Big Data analytics.

6. Manage the Data Differently

Big data requires a different approach to data management

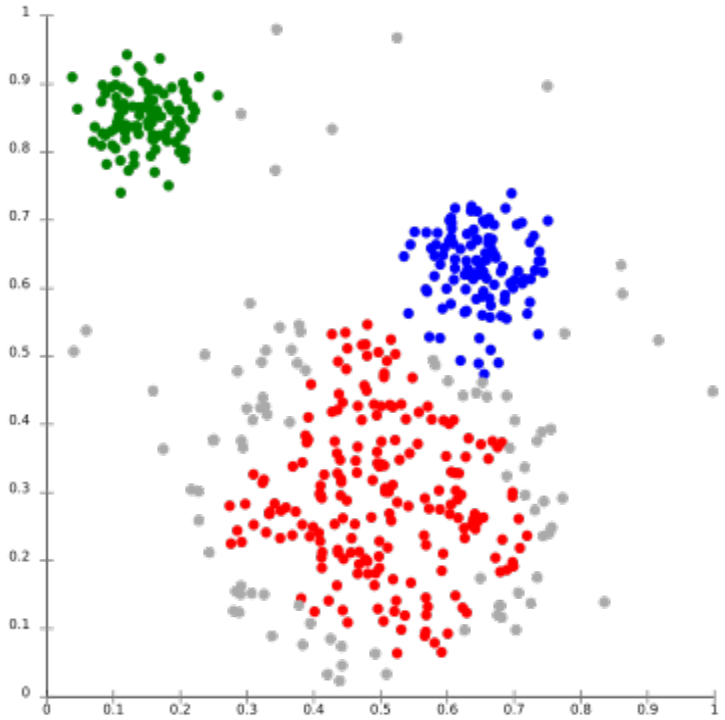
- Store the data as-is
- Maintain data accessibility
- Structure the data for analysis
- Ensure the data is uniquely identifiable
- Sharing, security, and privacy
- Protect data without locking it down

7. Process the Data Differently

- Process the data where it is located
- Use open source software
- Do not reinvent the wheel
- Understand the ephemeral nature of big data analytics

8. Open and Share Outcomes and Products to Foster Data User Communities

- Share trends, patterns, models, visualizations, and outliers discovered through big data analytics with a broader community directly through the common data storage
- Support the development of data user communities composed of government employees, government contractors, universities, the private sector, and citizens to form a continuously evolving collaborative environment able to maximize the value of its big data datasets.



Contact Information:

Kelley Klaver Pecheux
AEM Corporation
Kelley.Pecheux@aemcorp.com
(703) 350-8487

Benjamin Pecheux
AEM Corporation
Ben.Pecheux@aemcorp.com
(703) 989-4776

Grady Carrick
Enforcement Engineering, Inc.
gcarrick@enforcementengineering.com
(904) 705-8046