Big Data at UMTRI

Carol Flannagan, Ph.D.

Associate Research Scientist & Co-Director, Center for the Management of Information for Safe and Sustainable Transportation (CMISST)

September 11, 2013
What Is Big Data?
What Is Big Data?

Big Data is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand tools or traditional applications.

Wikipedia
What is Big Data?

What good is Big Data to a research organization?

1. Large samples are required to observe rare events (many observations)
2. Taking many different measurements is useful when you don’t know exactly what you’re looking for (many variables)
3. Greater density of data helps build better models (more detail)
What Is Big Data?

Three Dimensions of the Big Data Enterprise (Bin Yu)
Data
CMISST’s mission is to be a central repository of data, metadata, and knowledge about transportation data and analysis within and outside of UMTRI.

Become a member! (see cmisst.org or contact cacf@umich.edu)
Center for the Management of Information for Safe and Sustainable Transportation (CMISST)

Transportation Data Analysis

Data Dissemination

Methods Development

<http://www.cmisst.org>
Big Data at UMTRI

CRASH DATA

Crash datasets from 5-10 years of at least 20 states = ~20 million crashes

National samples of crash data—richer datasets, fewer observations
Big Data at UMTRI

DRIVING DATA
U.S. DOT Field Operational Tests Conducted by UMTRI

Light Vehicles – ACC
ICC FOT 130K mi, 108 drivers

Light Vehicles – Forward Crash/ACC
ACAS FOT 137K mi, 96 drivers

Light Vehicles – Lane Departure/Curve Speed
RDCW FOT 110K mi, 78 drivers

Light Vehicles & Heavy Trucks – Multiple Systems
IVBSS 804K mi 150+ drivers

Heavy Trucks - Rollover
RSA 480K mi, 23 drivers

RDCW FOT (UMTRI/Visteon)

ACAS FOT (GM/Delphi/UMTRI)

1990 2010
IVBSS Light Vehicles

- 16 vehicles each with four prototype crash warning systems
- 7 radars, 5 video streams, GPS, >500 other signals at 10 to 50 Hz
Safety Pilot Model Deployment
Safety Pilot Model Deployment Scope

• 1-year deployment began in 2012
• 2836 vehicles equipped
  – Passenger cars, trucks, and buses
  – Over 2500 vehicles with basic messages and GPS
  – 182 vehicles with full data acquisition system and warning systems
• 73 lane-miles of roadway instrumented with 25 roadside-equipment installations
• Over 200Tb (and counting…) of data including messages, video, warnings, driver response etc.
Big Data at UMTRI

ANTHROPOMETRY DATA

1970’s
Traditional methods

1990’s
FARO Arm

2010’s
VITUS XXL Scanner
UMTRI Data Sets

Adults

Elderly

Children

Soldiers

Reed, 2013
Big Data at UMTRI

SPATIAL DATA
CMISST Datasets and Linkages

Exposure Data

Population Demographics

Demographics, Travel exposure

Driving Datasets

Surrogates

Occupant

Other State Datasets

Location and Roadway

GIS Databases

Occupant

Vehicle ID Number

Vehicle Characteristics

Anthropometry Data
Computation
Big Data at UMTRI

We are in a constant battle with dataset size to find hardware and data structures that handle larger and larger datasets efficiently.

Solution 1: It’s only money…Buy more and faster storage and CPU

Solution 2: Be smarter…Build more efficient data structures for the data
Software tools also have to evolve

- Faculty, staff, and students all have to learn new tools to handle management and analysis of larger and more complex datasets

- For some datasets, we build custom tools
  - Metadata management system
  - Variable standardization on the fly
  - User search and access tools
Humans
Big Data at UMTRI

Big Data needs to be complemented by Big Judgment*!

This means that results have to make sense, show stability, and be replicable…

*Harvard Business Review
Multidisciplinary domain experts:

Computer scientists
Programmers
Statisticians
Engineers
Psychologists
Social Scientists

Each discipline brings its own perspective and expertise and keeps the others honest.
We often need multiple disparate datasets to answer research questions
Benefits of Safety Content

Problem: How effective are different safety technologies in the field? Which are worth the cost??
Benefits of Safety Content

Data Needs:
1) Crash data with information about safety equipment on crash-involved vehicles

Data Challenges:
1) Little information about optional vehicle equipment is included in crash databases
2) Few vehicles have the newest technology, so very large datasets are needed
Benefits of Safety Content

Big Data Solution:

1) Manufacturers provide large sales datasets with safety content linked to Vehicle Identification Number (VIN)

2) UMTRI houses 12 state crash databases with full VIN included

3) Datasets are matched on VIN to search for crashes involving vehicles with and without new technologies
Benefits of Safety Content

In a recent study,

1.5 million VINs with safety content and
~10 million crashes (10 states over 5 years)
= 6,895 matches
Benefits of Safety Content

Sometimes 1.5 million isn’t enough...

We are launching a new program to pool VIN-linked safety content from many manufacturers to try to build a larger analysis database and get early estimates of field performance of the newest safety technologies.
Simulation is sometimes required to jump a gap between large datasets.
Benefits of Collision Mitigation
Braking (CMB) Systems in Heavy Trucks

Problem: What are the estimated benefits (in $$ or lives/injuries saved) of installing CMB systems in heavy trucks?

Data Needs (national level):

1) Data from heavy-truck-into-car crashes including pre-crash vehicle movements
2) Distribution of crash severities for those crashes
Benefits of Collision Mitigation Braking (CMB) Systems in Heavy Trucks

Data Challenges:

1) There are no data that include pre-crash kinematics (speed, deceleration, etc.) for truck-into-car crashes
Benefits of Collision Mitigation Braking (CMB) Systems in Heavy Trucks

Big Data Solution:

1) Use driving data to understand pre-crash movements; simulate crashes to produce baseline crash dataset

2) Use crash data to calibrate the overall distribution of simulated crashes
Benefits of Collision Mitigation Braking (CMB) Systems in Heavy Trucks

Developed simulation approach that uses:

- **Real-world truck-following-car kinematics from large driving database**
- Simulated crashes (simulate delayed driver response)
- Make sure results match crash-severity distributions from large crash database

**Initial Conditions:**
- Truck Speed = 40 mph
- Car Speed = 38 mph
- Range = 55 m

![Car Deceleration](image)
Weighting Baseline Cases to Match Crash Distribution

![Graph showing the distribution of crash speeds with and without weighting. The red line represents the weighted distribution, while the blue line represents the original distribution. The graph shows a peak near 12 mph in both cases, with the weighted distribution slightly shifted to the right.](image-url)
Creating Simulated Crashes

Result:

Baseline crash dataset with hundreds of thousands of simulated crashes with full detail on vehicle movements, representing the population of truck-into-car crashes in the U.S.
Big Data at UMTRI

MODELING

Rich statistical models can be built from rich datasets
Parametric Human Body Modeling

Problem: Crash dummies represent a one-size-fits-all approach to occupant protection in vehicles. People come in many different sizes and shapes, and we need to understand how body shape, size, and posture affects safety.
Parametric Human Body Modeling

Big Data Solution: Develop a statistical model of human bodies that represents a wide variety of body shapes and sizes as a function of a few parameters (e.g., stature, BMI, posture)

- Laser towers (4)
- Hand-held laser scanner to augment towers
- Red laser beam
- VITUS XXL Scanner
Parametric Human Body Modeling

Multiple subjects and postures
Surface Analysis

Output of regression model to predict seated body shape
(based on 338 scans from 126 men)

Median US Male
Stature 1755 mm
BMI 27.3 kg/m²

Stature 1600 mm

Stature 1900 mm

BMI 18 kg/m²

BMI 40 kg/m²
Posture and Body Shape

Output of regression model

Torso Recline

Random Simulated Men

Torso Flexion
Parametric Study on Obesity

BMI 25

BMI 40
One final thought:

Is Big always relative?
Maybe not—

- The new frontier in transportation safety is datasets too large for hand work (e.g., video or case review)
Thanks for your attention.

University of Michigan Transportation Research Institute (UMTRI)