Transportation Data Management and Analysis (TDMA)

An Introduction

AECOM

Southeast Florida FSUTMS Users Group

June 13, 2014

Manish Jain

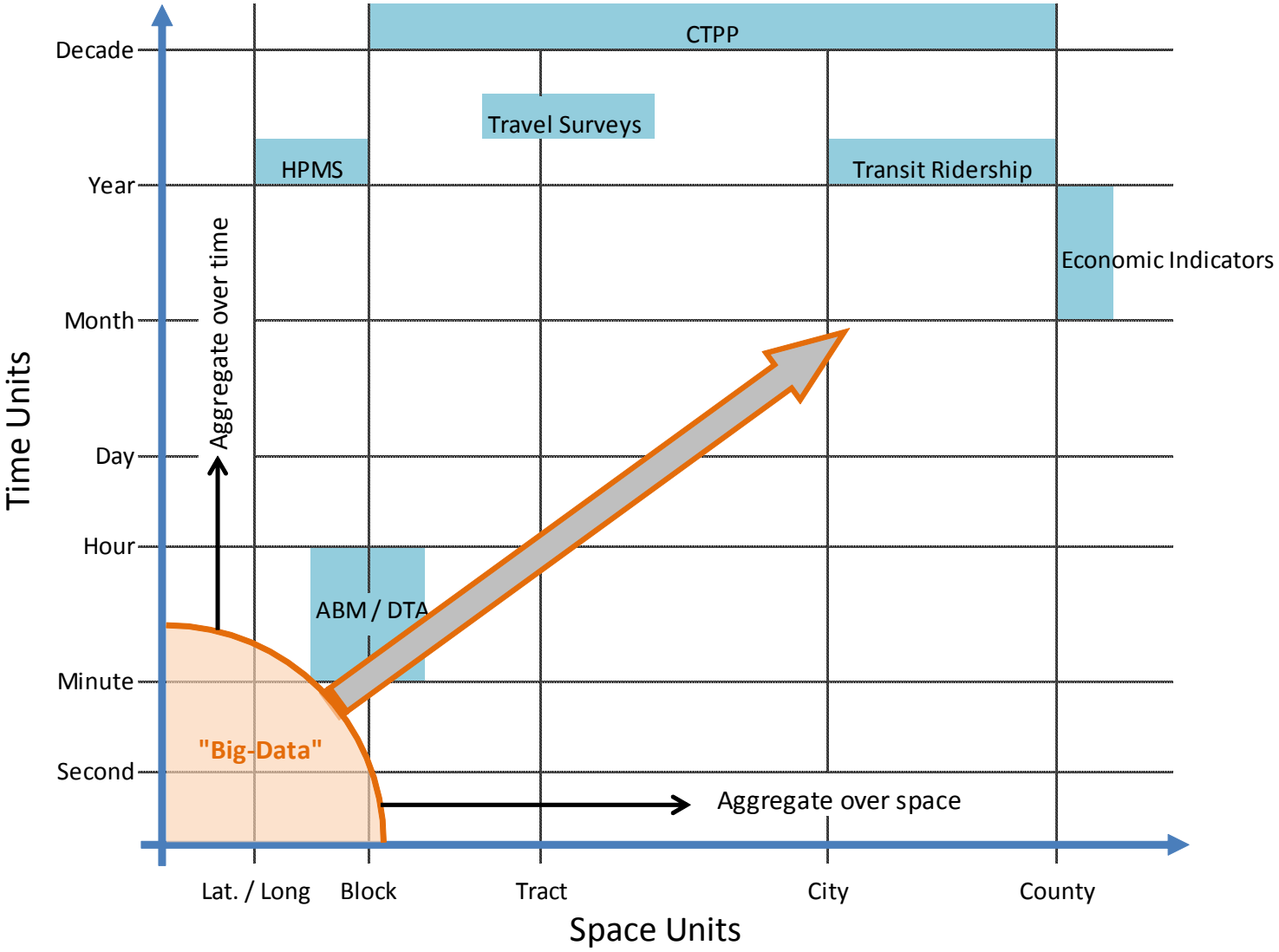
TDMA – An Introduction: Agenda

- Overview
- Select Datasets
- Data Processing Tools
- Examples
- Q&A

Transportation Data – Multiple Dimensions

- Modes
  - Highway, Arterials, Freight, Transit, etc.
- Devices
  - Roadway sensors/detectors, Transit AVL and APC, Bluetooth, GPS, etc.
- Sources
  - Arterial and Freeway Management Systems, Transit Operators, DOTs, TMCs, Freight, Third Party Data (INRIX/HERE), etc.
- Data Types
  - Traffic speed/volume/occupancy, incidents, transit passenger ONs/OFFs, Freight tonnage, etc.
- Miscellaneous
  - Activity Based Models, Dynamic Traffic Assignment, Specialized Transportation Services, etc.

Transportation "Big-Data"



Adapted from ADMS presentation made by Dr. Genevieve Giuliano at USC

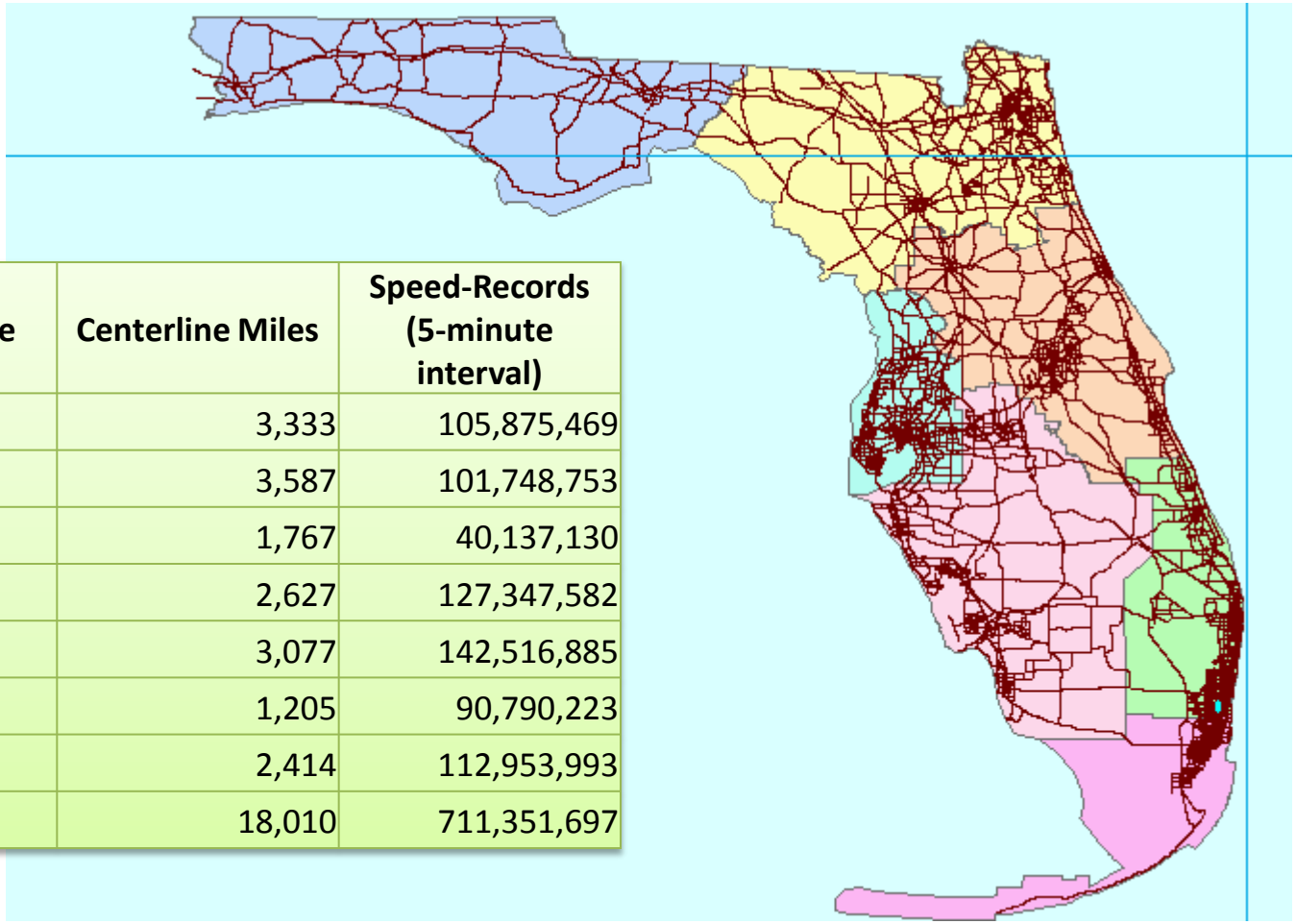
Select Datasets

Speed Dataset – INRIX

- Real-time speed data collected from nearly 100 million anonymous mobile phones, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices
- Temporal resolution: 1, 5, 15, 30 and 60 minutes
- Data elements: TMC, Speed, Travel Time, Reference Speed, Historical Speed, Confidence value
- Size: Current archive 19TB; growing @ 21GB/Day
- Geographic coverage: Freeways and arterials
- Traffic Message Channel (TMC) shape file for visualization

TMC	Date	Time	Speed	ReferenceSpeed	Average Speed	Score	TravelTimeMinutes	C_Value
110+04131	6/22/2009	5:45	41	46		30	0.10	99
110+04131	6/22/2009	5:50	41	46		30	0.10	100
110+04131	6/22/2009	5:55	41	46		30	0.10	100
110+04131	6/22/2009	6:00	41	46		30	0.10	100
110+04131	6/22/2009	6:05	46	46		10	0.09	

Speed Dataset: INRIX - Florida



District/State	Centerline Miles	Speed-Records (5-minute interval)
D1	3,333	105,875,469
D2	3,587	101,748,753
D3	1,767	40,137,130
D4	2,627	127,347,582
D5	3,077	142,516,885
D6	1,205	90,790,223
D7	2,414	112,953,993
Florida	18,010	711,351,697

Collected on 33,700 Traffic Message Channel (TMC) links during the period 7/1/2010 to 6/30/2011

Speed Dataset – NPMRDS / HERE

- National Performance Measure Research Data Set (NPMRDS), provided by HERE
- Similar probe data as INRIX but separates Cars and Trucks
- Temporal resolution: 5 minutes with no data imputations
- Data elements: TMC, Date, Epoch, Travel Time (All vehicles, passenger vehicles, freight trucks)
- Size: Monthly data files between 500MB and 4GB in size
- Geographic coverage: National Highway System
- TMC shape file for visualization
- Provided by FHWA to public agencies for free (HERE as vendor)

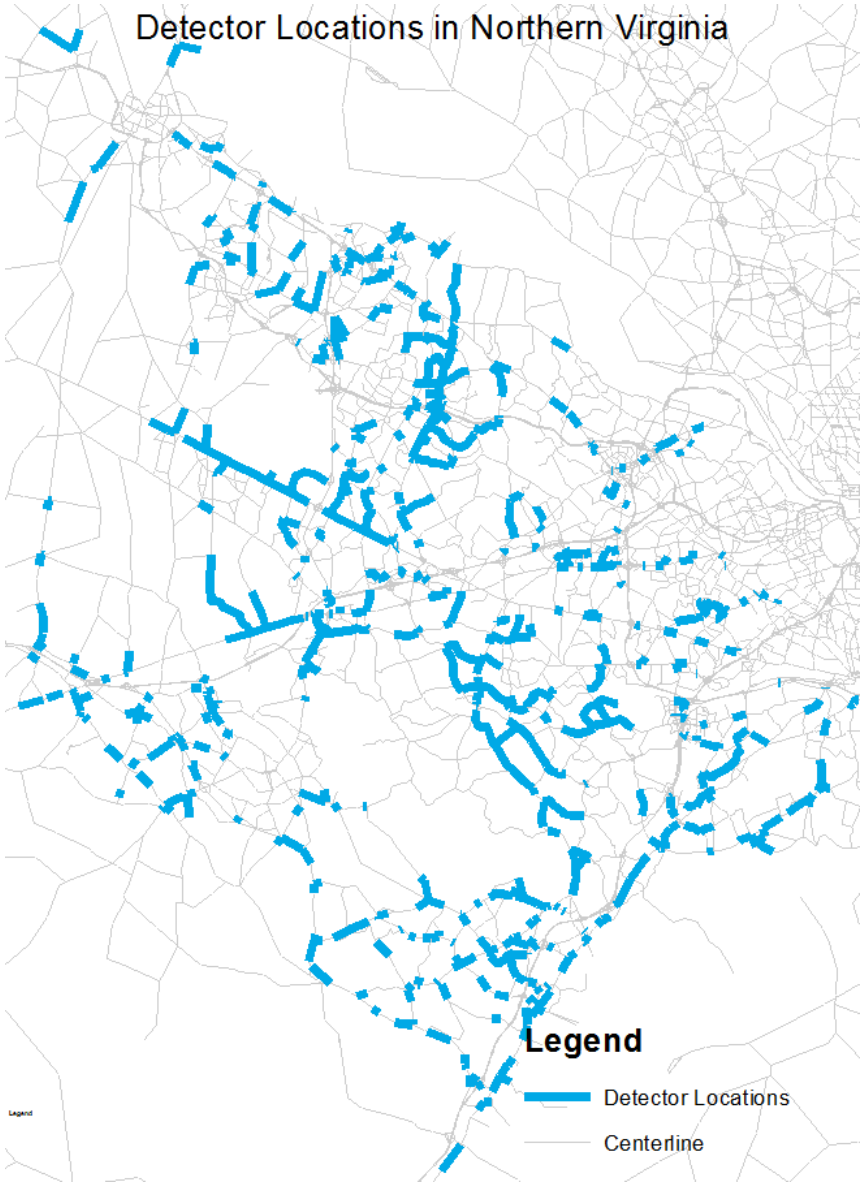
TMC	DATE	EPOCH	Travel_TIME_ALL_VEHICLES	Travel_TIME_PASSENGER_VEHICLES	Travel_TIME_FREIGHT_TRUCKS
115N04098	10052013	262	97	97	
115N04098	10012013	34	101	101	
115N04098	10012013	66	100		100
115N04098	10012013	98	112	111	126

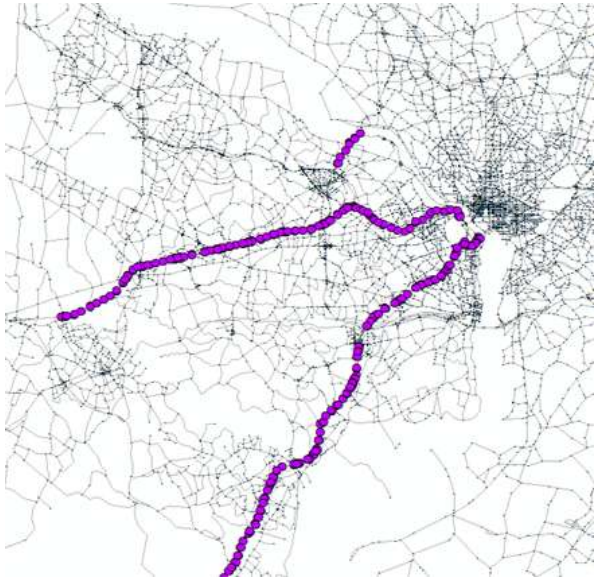
Count Dataset: VDOT ADMS

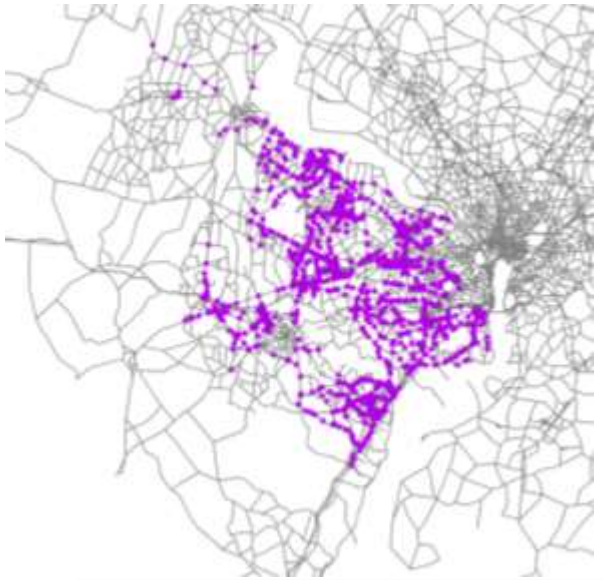
- VDOT's Northern Region Operations (NRO) has 1,337 signalized intersections with 15,765 detectors. Also has 1,319 detectors on freeways and ramps
- Volume, Occupancy and Speed data collected every 15-minutes
- Arterial data: 96 * 15,765 = 1,513,440 detector records/day
- Freeway data: 96 * 1,319 = 126,624 detector records / day
- Average data size per month: 2.4 GB

detector_id	data_date_time	ignore	Volume	occupancy	speed	ignore	status	data_period	ignore
300003511	10/7/2010 20:30	0	24	33	0	0	ONLINE	15	0
300003512	10/7/2010 20:30	0	24	1	0	0	ONLINE	15	0
300003513	10/7/2010 20:30	0	480	6	51	0	ONLINE	15	0
300003514	10/7/2010 20:30	0	608	9	44	0	ONLINE	15	0
300003515	10/7/2010 20:30	0	268	2	52	0	ONLINE	15	0

Count Dataset: VDOT ADMS

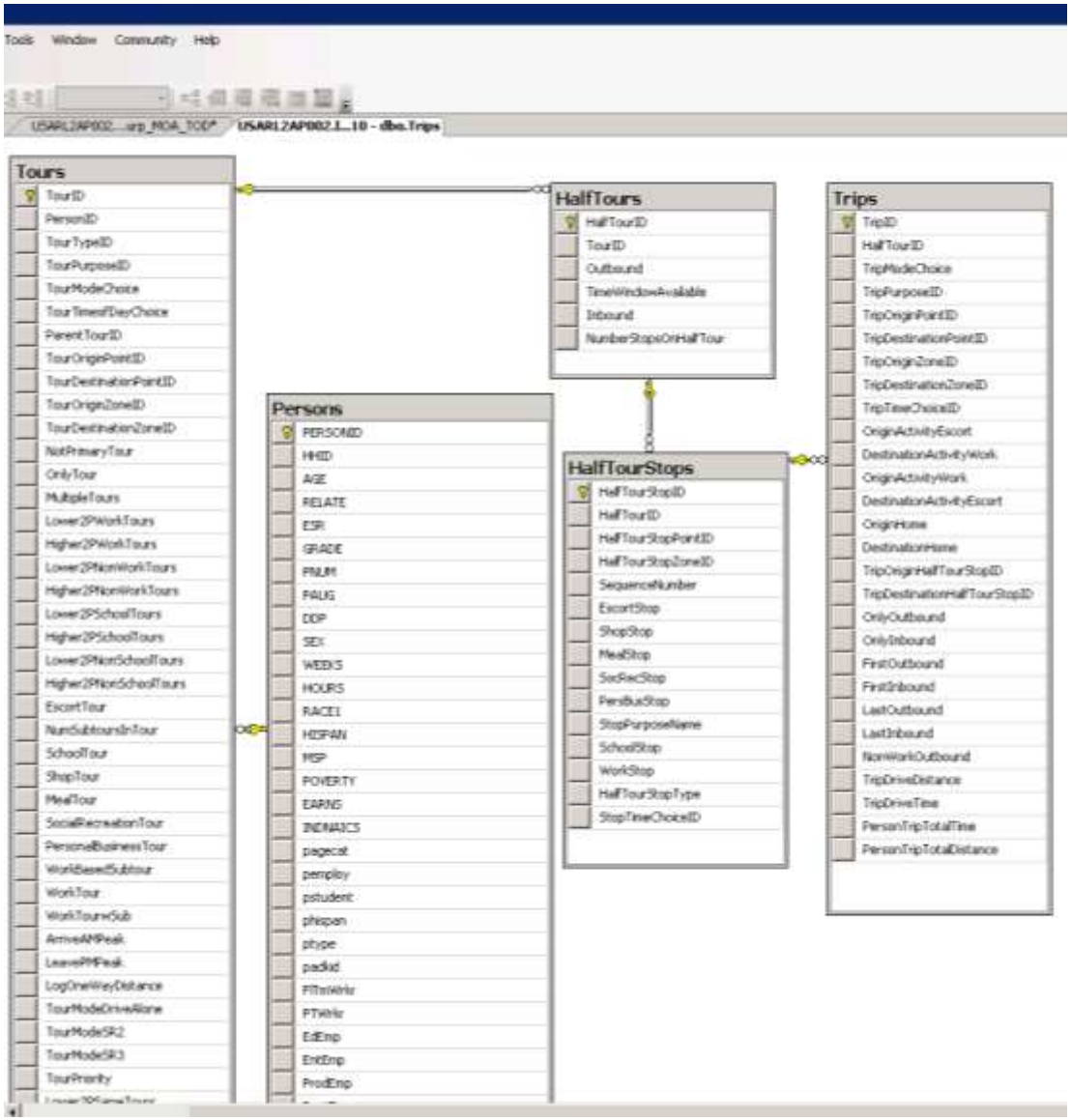






Miscellaneous Datasets: ABM

- Activity Based Model (ABMs) output tours and trips at record level
  - SERPM: Over 9 million tours and 21 million trips (Base year)
  - FOCUS: Over 4 million tours and 11 million trips (Year 2010)
- Typically model outputs stored in series of tables
  - Tours, Trips, Half-tours, Household data, Person data, etc.



SERPM ABM Database

Table	Table Name	Description
Synthetic Households	SYNHH	Input synthesized households from the PopSyn.
Synthetic Persons	SYNPERSON	Input synthesized persons from the PopSyn.
TAZ Data	TAZ	Input TAZ data.
MGRA Data	MGRA	Input MGRA (MAZ) data such as employment, etc.
TAP Data	TAP	Input TAP data.
MGRA to TAP Data	MGRATOTAP	MGRA to TAP distances, etc.
MGRA to Stop Data	MGRATOSTOP	MGRA to all transit stops distances, etc.
MGRA to MGRA Data	MGRATOMGRA	MGRA to MGRA distances, etc.
TAZ to TAP Data	TAZTOTAP	TAZ to TAP distances, etc.
Accessibilities	ACCESSIBILITIES	Model results for accessibilities.
Household Data	HHDATA	Model results for household level choice models.
Person Data	PERSONDATA	Model results for person level choice models.
Work and School Location	WSLOCATION	Model results for usual work and school location choice models.
Individual Tours	INDIVTOUR	Modeled individual tours.
Joint Tours	JOINTTOUR	Modeled joint tours.
Individual Trips	INDIVTRIP	Modeled individual trips.
Joint Trips	JOINTTRIP	Modeled joint trips.
CBD Vehicle Trips	CBDVEHICLES	Number of vehicle trips to CBD by time period.
All microsimulated trips	TRIP	All microsimulated trips.
District/County Definitions	DISTRICTDEFINITIONS	Mapping of TAZs to counties and districts.
PECAS Occupations	PECASCODES	Mapping of PECAS codes to occupations.

Miscellaneous Datasets (cont'd)

- Freight Flow Data
  - Transearch commodity flow data in series of tables. Requires relational database to query and summarize information
  - Surface Transportation Boards Way Bill Data in flat ASCII files
- Transit
  - APC data at each stop for each transit trip. Includes ONs, OFFs, Arrival Time, Departure Time, etc.
  - AVL data
  - Fare gate to fare gate data for each passenger
  - Specialized Transportation Service Data: Includes passenger information, trip purpose, boarding location, alighting location, trip time, fare, etc.

All these datasets can become extremely large over time

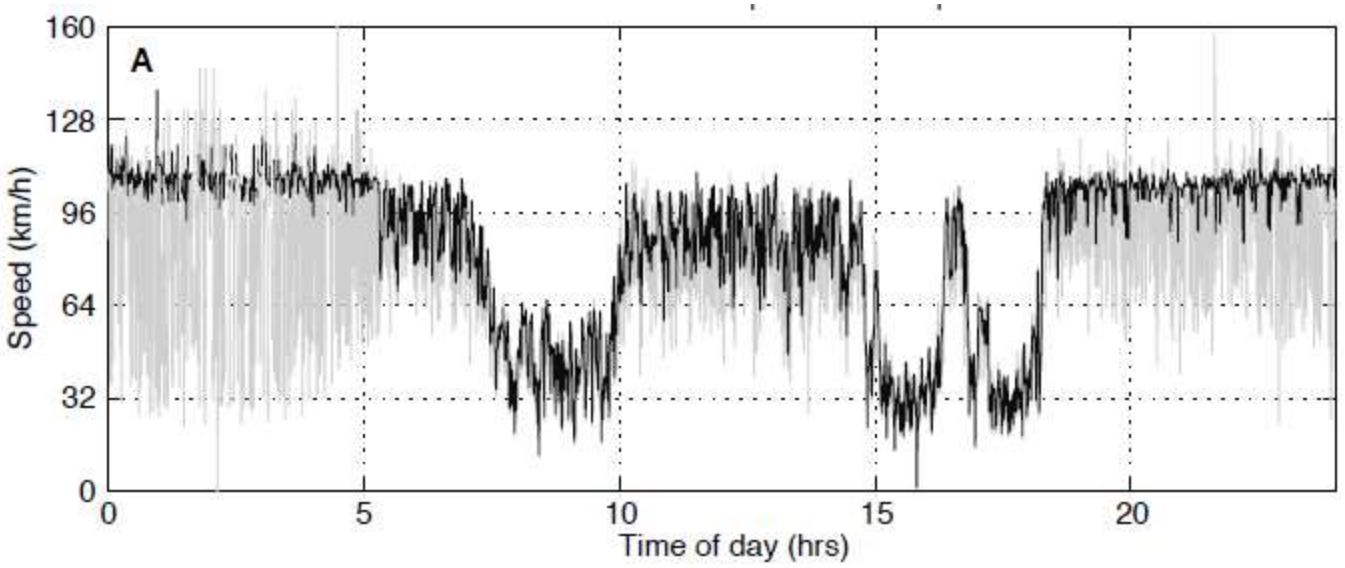
Data Processing and Examples

Data Processing

- Hardware and software requirements:
  - Excel can only handle up to a million records
  - Access has 2GB per table limit; will be quickly exceeded
  - Requires database and programming/scripting skills
  - GIS expertise required for mapping
    - TMCs in INRIX/HERE data have many to many relationship with shape file links
      - One LINK can reference many TMCs
      - One TMC can reference many TMCs
- Open Source tools and databases
  - MySQL, PostgreSQL, Python, R, TRANSIMS SysLib, etc.
- Commercial tools and databases
  - Microsoft SQL, Oracle, SPSS/SAS, C++, FORTRAN, Matlab, etc.

Data Visualization

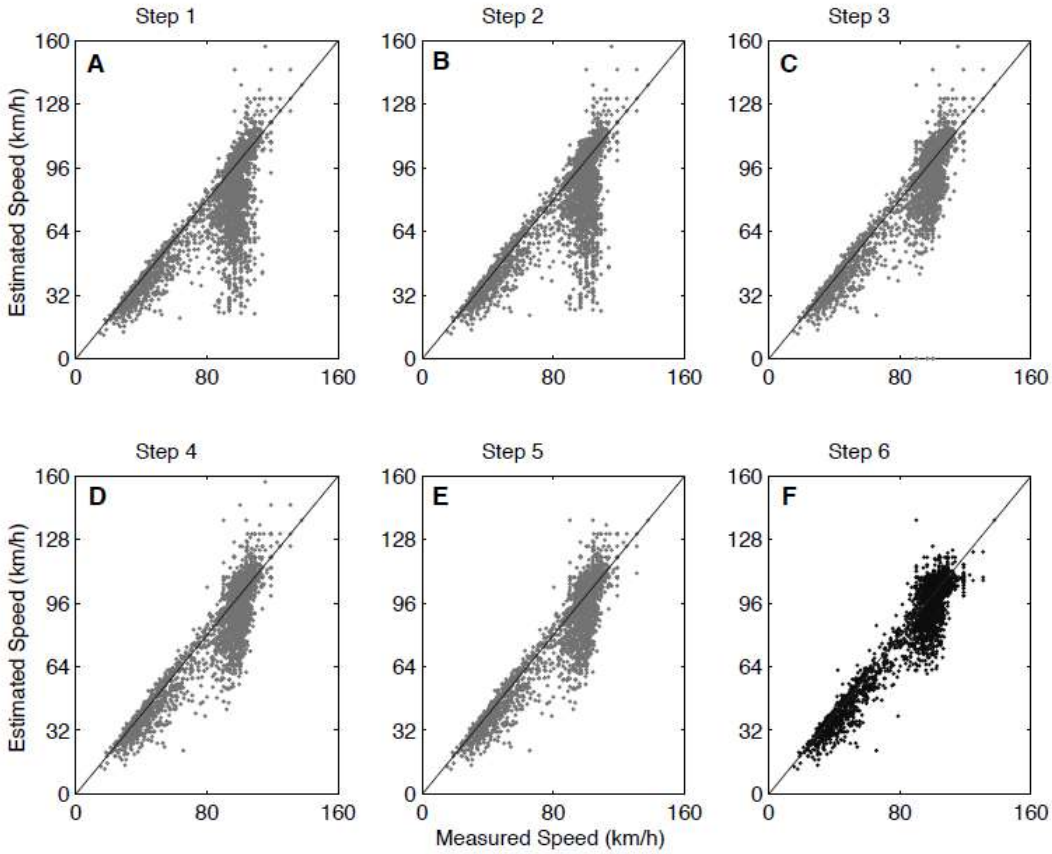
- Integrating INRIX/HERE data with GIS will require resolving relationships between TMCs and LINKS. Displaying data by direction could be challenging too!
  - Recommend managing the spatial data in a relational database system
  - Downside is that the spatial table would become huge
- Programming languages such as Matlab, R, and Python extensions provide powerful graphic capabilities



Data Filtering – An Example

Illustration of how data filtering and statistical analysis can be used to improve speed estimates:

- Data from detector station 4 on I-80 westbound at Berkeley Highway Laboratory (BHL), California
- Six step approach for data-filtering:
  - Step 1: Identify erroneous samples
  - Step 2: Raw estimate of speed
  - Step 3: Speed-Flow filter
  - Step 4: Speed-Occupancy filter
  - Step 5: Speed filter
  - Step 6: Moving median of three samples
- Used MATLAB for analysis

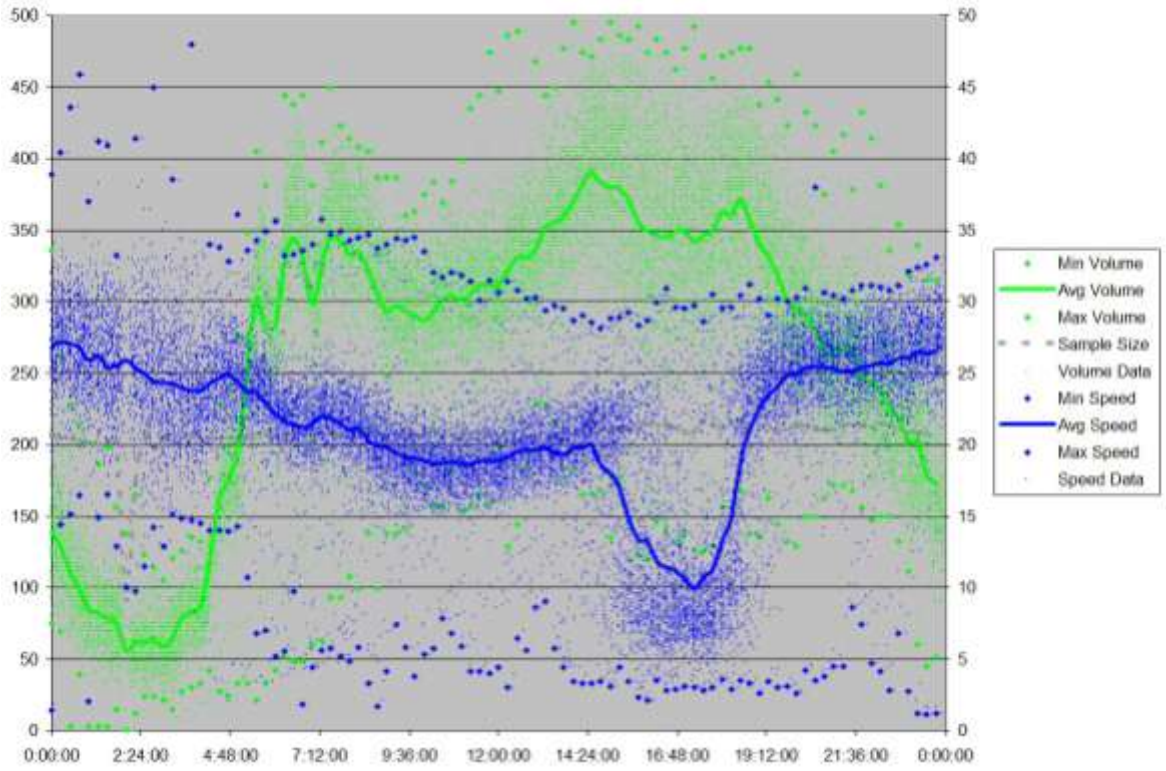


Plot showing progression of improvement in speed estimates. The plot A shows the estimates at the end of Step 1, B shows the estimates at the end of Step 2, C shows the estimates at the end of Step 3, D shows the estimates at end of Step 4, E at the end of Step 5 and F at the end of Step 6.

Reference: Jain, M and Coifman, B, "Improved Speed Estimates from Freeway Traffic Detectors", the Ohio State University.

Freeway Sensor Data – Example 1

- Data extraction and management in Python and SQLite3
- 800 sensors with about 700 operational at any point in time
- ~28 million records per year
- Plot shows data for:
  - Two month weekday (Mon-Fri) 15 minute averaged dataset
  - Single sensor on I-55, Chicago
  - Speed in meters/second



Courtesy: Dr. Hubert Ley, Argonne National Laboratory

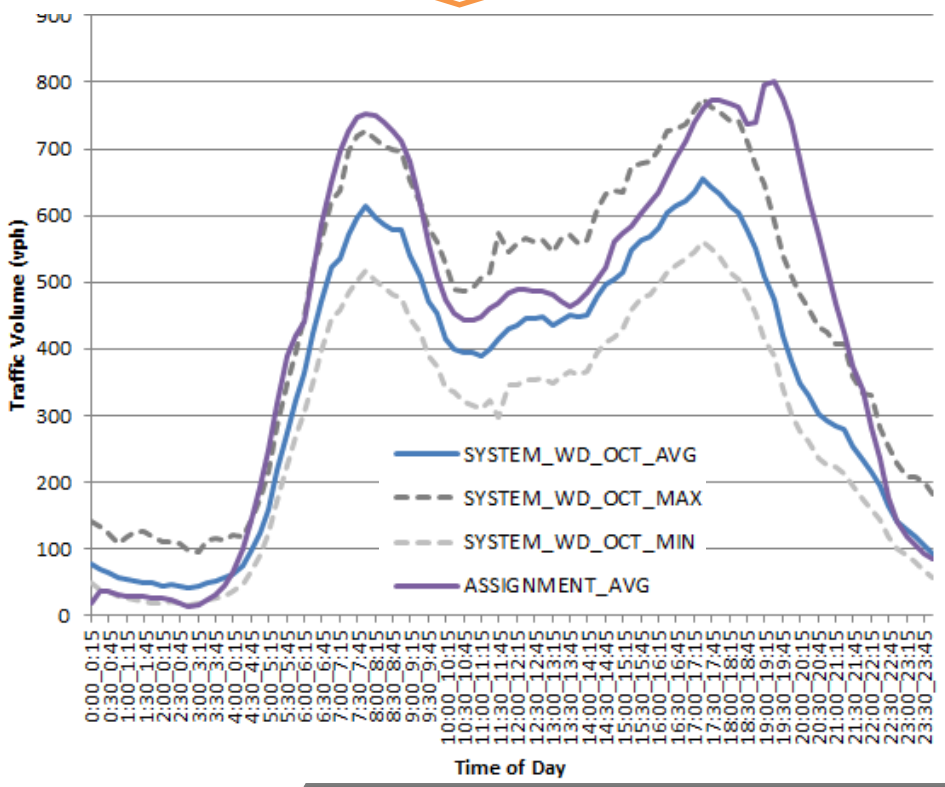
Count Data Summary – Example 2

- Data management and extraction using TRANSIMS SysLib CountSum program
- Customized filtering based on days / facility-types / detector-types / signal-types / station-types
- Outputs data in a variety of file-formats (dbase, binary, sqlite3, csv, etc.)
- Automatic or user guided tagging of detectors to links

• VDOT NRO Arterial Data; 1,513,440 detector records/day ; Multiple Days

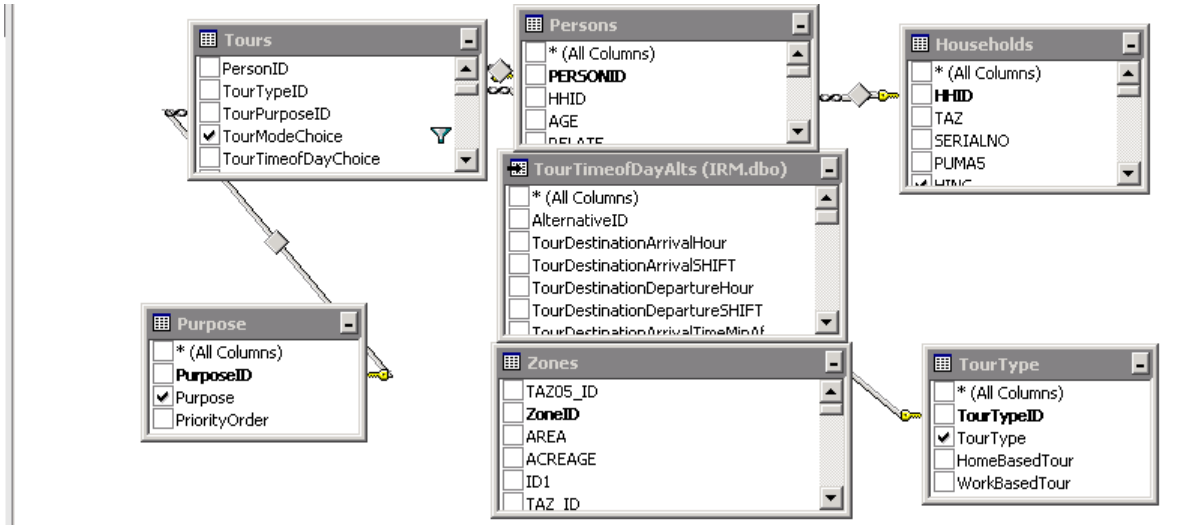
detector_id	data_date_time	ignore	Volume	occupancy	speed	ignore	status	data_period	ignore
300003511	10/7/2010 20:30	0	24	33	0	0	ONLINE	15	0
300003512	10/7/2010 20:30	0	24	1	0	0	ONLINE	15	0
300003513	10/7/2010 20:30	0	480	6	51	0	ONLINE	15	0
300003514	10/7/2010 20:30	0	608	9	44	0	ONLINE	15	0
300003515	10/7/2010 20:30	0	268	2	52	0	ONLINE	15	0

CountSum



ABM Data Extraction – Example 3

- Example SQL Query to extract transit tours by Income, Purpose, Mode of Access, Time of Day, and Attraction District
- Queries can be designed via a wizard and does not require extensive knowledge of SQL syntax.
- SERPM ABM provides some standard reporting queries; additional custom queries can be developed using T-SQL



Column	Alias	Table	Output	Sort Type	Sort Order	Filter	Or...	Or...
TourModeChoice		Tours	☑			IN ('Walk to ...		

```
SELECT    dbo.Tours.TourID, dbo.Tours.TourModeChoice, dbo.Households.HINC, dbo.Zones.DISTRICT, IRM.dbo.TourTimeofDayAlts.ArrivalTimeTransitPeriod,
          dbo.Purpose.Purpose, dbo.TourType.TourType
FROM      dbo.Households INNER JOIN
          dbo.Persons ON dbo.Households.HHID = dbo.Persons.HHID INNER JOIN
          dbo.Tours ON dbo.Persons.PERSONID = dbo.Tours.PersonID INNER JOIN
          dbo.TourType ON dbo.Tours.TourTypeID = dbo.TourType.TourTypeID INNER JOIN
          dbo.Purpose ON dbo.Tours.TourPurposeID = dbo.Purpose.PurposeID CROSS JOIN
          IRM.dbo.TourTimeofDayAlts CROSS JOIN
          dbo.Zones
WHERE     (dbo.Tours.TourModeChoice IN ('Walk to Transit', 'Drive to Transit'))
```

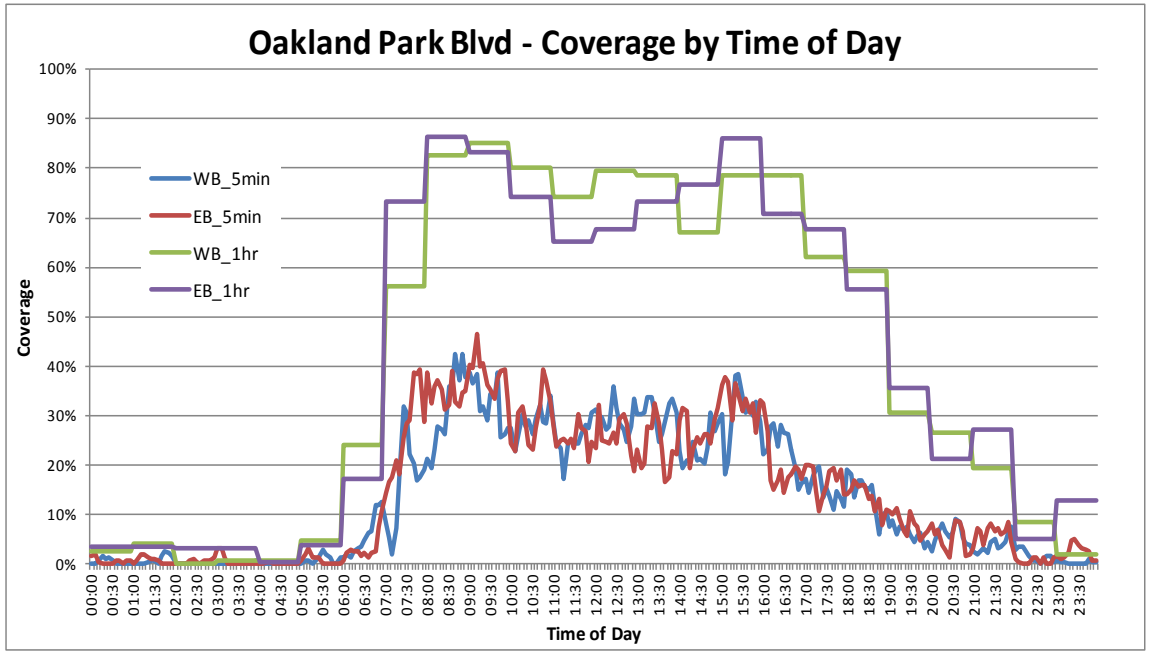
Results | Messages

	TourID	TourModeChoice	HINC	DISTRICT	ArrivalTimeTransitPeriod	Purpose	TourType
1	804125	Walk to Transit	12600	Rural	PM	School	Home--Based
2	804125	Walk to Transit	12600	Rural	PM	School	Home--Based
3	804125	Walk to Transit	12600	Rural	PM	School	Home--Based
4	804125	Walk to Transit	12600	Rural	PM	School	Home--Based
5	804125	Walk to Transit	12600	Rural	PM	School	Home--Based
6	804125	Walk to Transit	12600	Rural	PM	School	Home--Based

INRIX Data Coverage – Example 4

- 5-Minute INRIX Speed Data for D4 processed using Cube scripts to extract data for Oakland Park Blvd TMCs.
- Resultant data analyzed in Excel

TMC	SPEED	DATE	TIME
102+11717	36.00	10/05/2010	12:55
102+11717	26.00	10/05/2010	16:25
102+11717	26.00	10/05/2010	16:30
102+11717	41.00	10/05/2010	16:35
102+11717	39.00	10/05/2010	16:40
102+11717	33.00	10/05/2010	16:45
102+11717	35.00	10/06/2010	09:20
102+11717	39.00	10/06/2010	13:20
102+11717	27.00	10/06/2010	18:40
102+11717	29.00	10/06/2010	18:45
102+11717	29.00	10/06/2010	18:50

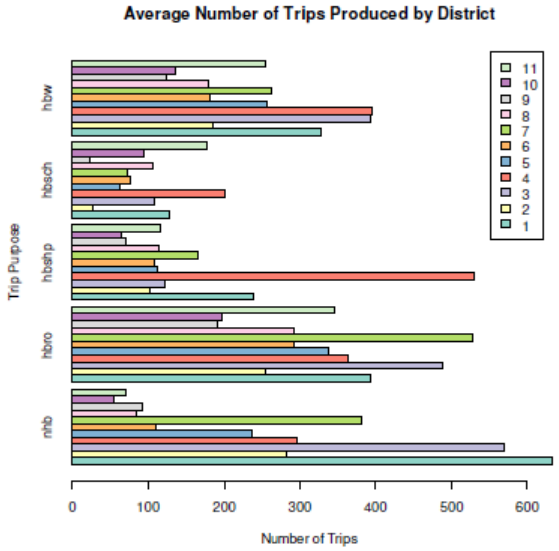


Census Data Extraction – Example 5

AECOM

Brian Gregor and Ben Stabler at Oregon DOT used R script to automate census data extraction:

- Script downloads Census 2000 county to county commuting data for any number of states
- Creates and saves the data on instate commutes
- Creates and saves an origin-destination matrix of the instate commutes
- Summarizes the instate commutes and saves the results




```
# Define function for retrieving a commuting file from the Census
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
get.commute <- function(state="OR", type="RES"){

      # Check for valid state and type codes
      state.abb <- c("AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",
"GA", "HI", "IA", "ID", "IL", "IN", "KS", "KY", "LA", "MA", "MD", "ME", "MI",
"MN", "MO", "MS", "MT", "NC", "ND", "NE", "NH", "NJ", "NM", "NV", "NY", "OH",
"OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX", "US", "UT", "VA", "VI", "WA",
"WI", "WV", "WY")
      types <- c("RES", "WRK")
      if(!(state %in% state.abb)) stop("Must use valid state abbreviation")
      if(!(type %in% types)) stop("Type must be RES or WRK")

      # Make the file name and the url to get the data from
      commute.file <- paste("2K", type, "CO_", state, ".txt", sep="")
      url.file <- paste("http://www.census.gov/population/cen2000/commuting/",
commute.file, sep="")

      # Connect to the Census 2000 county to county commute file by residence or
work place
      # For file data documentation see
http://www.census.gov/population/cen2000/commuting/coxcolayout.txt
      census.con <- url(url.file)

      # Read downloaded file into a data.frame
      # specify field widths (varies from Census documentation because a space
is located between each field
      census.width <- c(2,4,5,5,42,3,4,5,5,42,7)
      # set field names
      census.name <- c("res.state", "res.county", "res.msa", "res.pmsa",
"res.name", "wrk.state", "wrk.county", "wrk.msa", "wrk.pmsa", "wrk.name",
"count")
```

```
      # Summarize OD matrix in several ways
      commute.origins <- rowSums(commute.od)
      # compute computes by origin county
      commute.destinations <- colSums(commute.od)
      # compute commutes by destination county
      internal <- commute.od[row(commute.od)==col(commute.od)]
      # internal commutes are on the diagonal of of the od matrix
      internal.pct <- 100 * internal / commute.origins
      # compute the internal percentage
      internal.pct <- round(internal.pct, 1)
      # round the internal percentage to the first decimal place
      outflow <- commute.origins - internal
      # compute the number of commuters leaving the county
      inflow <- commute.destinations - internal
      # compute the number of commuters entering the county
```

Summary

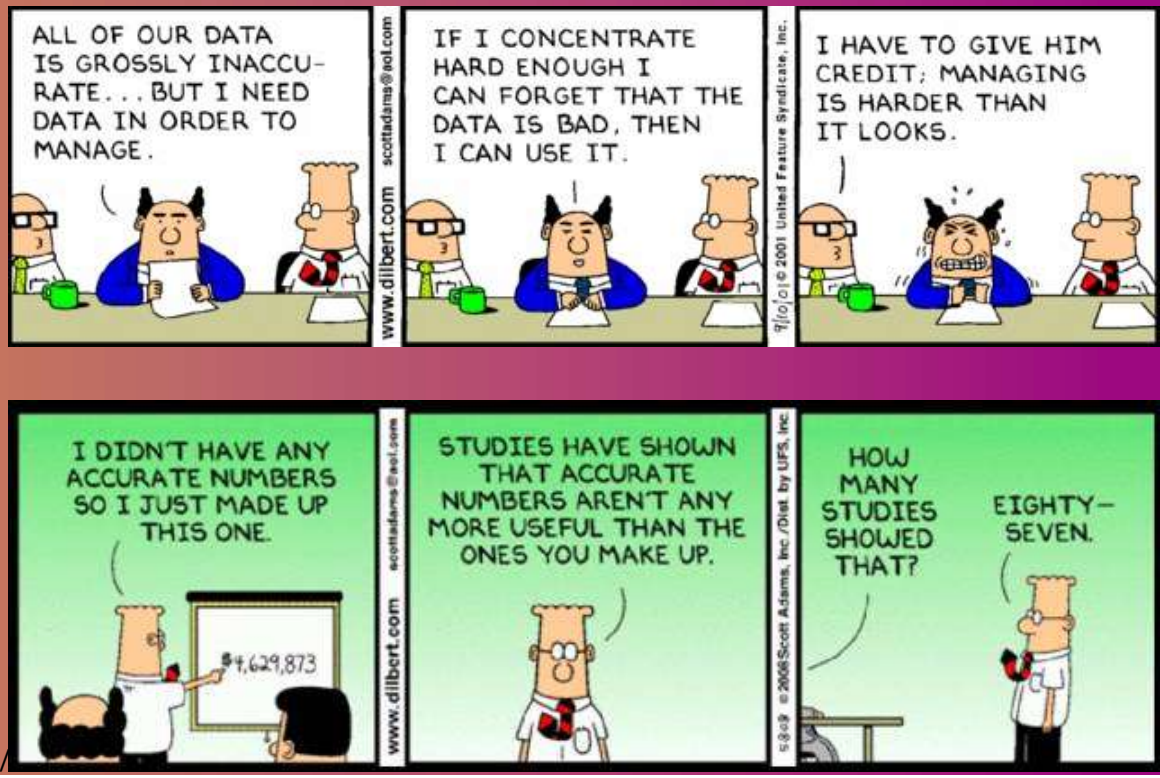
- Transportation Data archived from operations presents a tremendous opportunity for planning activities.
- Traditional data processing tools used by planners not capable of managing and analyzing “Big-Data”.
- Various data management and analysis tools exist and can be customized to agency needs.
- Three key phases for leveraging “Big-Data”
  i. Data Management
  ii. Data Analysis
  iii. Data Application
- Analysis and Application are next steps.

Questions?

Thank you!

Manish.Jain@AECOM.com

+1 703.340.3049



Courtesy: http://dilbert.com/strips/