



---

# *Traffic Profile Prediction*

## **Collection, Cleansing, and Analysis of Traffic, Event, and Weather Data**

**David Vickers – Staff Engineer**

**Adam Van Horn – Research Analyst**

**NATMEC 2016**

**Innovations in Traffic Data Collection and Processing**



# *Overview*

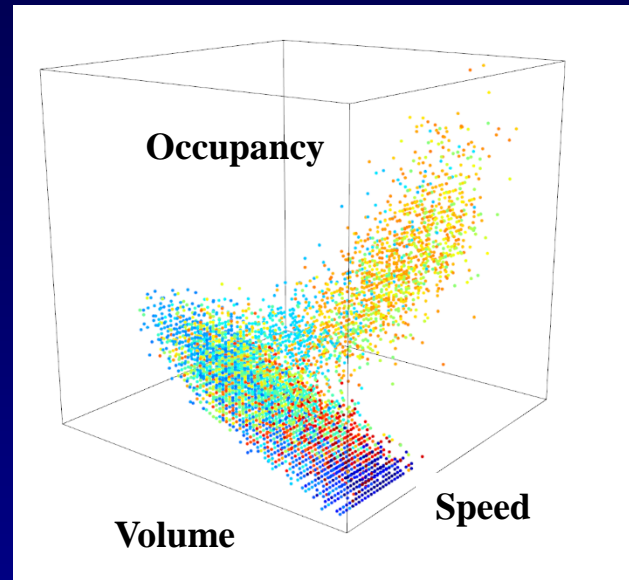
---

- **Types of Data**
- **Data Collection**
- **Data Cleansing**
- **Data Analysis**



# *Types of Data*

- **Roadway data**
  - Speed
  - Volume
  - Occupancy
  - Calculated Length
- **Event data**
  - Incidents from the traffic database
  - Sports Events
- **Condition data**
  - Weather
  - Holidays
  - Calendar

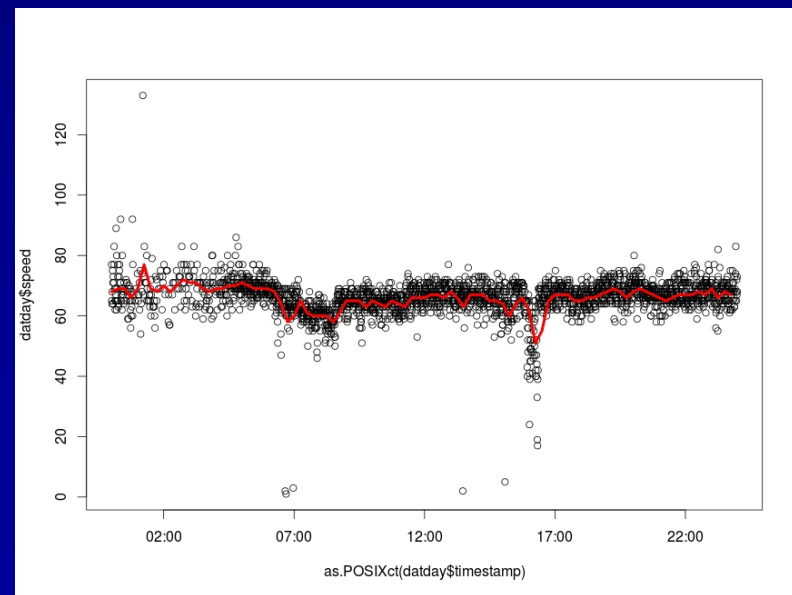




# Data Collection

- **Traffic data**
  - **Raw data files**
    - One file per day
    - All detectors from the district
    - 30 second or 5 minute data
  - **Averaged data**
    - Database table
    - 15 minute rolled up data
    - Selectable by lane, detector, etc.

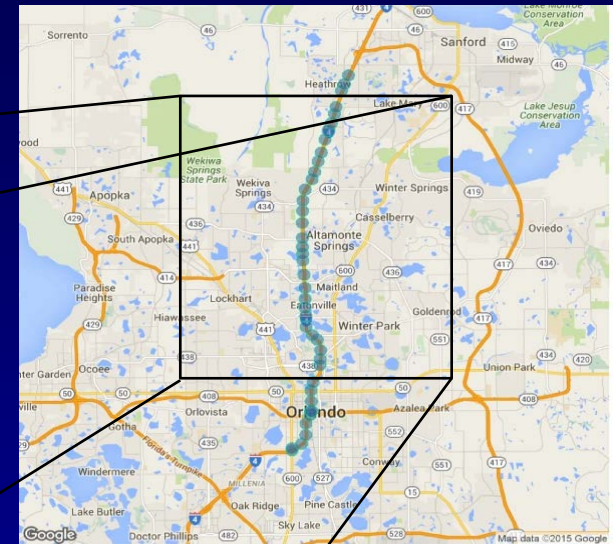
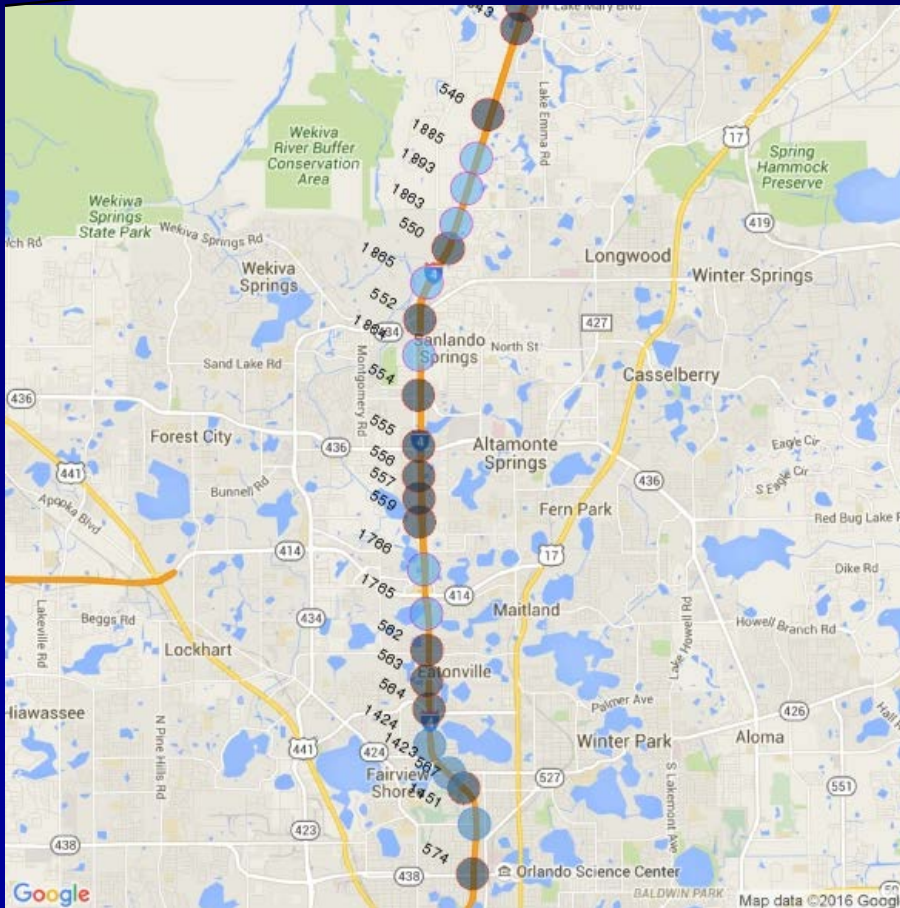
1	timestamp,	detector_id,	lane_id,	speed,	volume,	occupancy,	class_bin1,
2	11:58:57,	3081:tss:detector:MDX,	1322230:tss:lane:MDX,	55,	10,	20,	2,
3	11:58:57,	3081:tss:detector:MDX,	1322231:tss:lane:MDX,	58,	10,	20,	2,
4	11:58:57,	3081:tss:detector:MDX,	1322232:tss:lane:MDX,	69,	10,	20,	2,
5	11:58:57,	3081:tss:detector:MDX,	1322233:tss:lane:MDX,	67,	10,	20,	2,
6	11:59:27,	3064:tss:detector:MDX,	1322199:tss:lane:MDX,	67,	10,	20,	2,
7	11:59:27,	3064:tss:detector:MDX,	1322200:tss:lane:MDX,	61,	10,	20,	2,
8	11:59:27,	3064:tss:detector:MDX,	1322202:tss:lane:MDX,	66,	10,	20,	2,
9	11:59:27,	3064:tss:detector:MDX,	1322203:tss:lane:MDX,	56,	10,	20,	2,
10	11:59:25,	3061:tss:detector:MDX,	1322204:tss:lane:MDX,	64,	10,	20,	2,
11	11:59:25,	3061:tss:detector:MDX,	1322205:tss:lane:MDX,	59,	10,	20,	2,
12	11:59:25,	3061:tss:detector:MDX,	1322206:tss:lane:MDX,	51,	10,	20,	2,
13	11:59:25,	3061:tss:detector:MDX,	1322207:tss:lane:MDX,	54,	10,	20,	2,





# Data Collection

- **Focused analysis on I-4 Corridor FDOT D5**



## SunGuide Database

- **Detector Data**
- **Event Data**
  - **Accidents**
  - **Lane Closure**
  - **etc.**



# *Data Collection*

- **Event data**
  - **Final state or archive of transitions available**
  - **Data not consistently entered**
  - **Mix of causal and resulting data**
- **Weather data**
  - **Mix of formats available via APIs**
  - **Less data for historical weather**
  - **Most consistent is Airport data**
  - **Only approximately hourly**
  - **Not all airports are open 24 hours**
  - **Many variables available**
  - **Not all variables available all the time**



- External Data
- Generated or scraped
- Calendar
  - Day of the week
  - Day of the year
  - Moon phase
  - Holiday
  - School in session
- Sporting Events
  - From a nearby venue



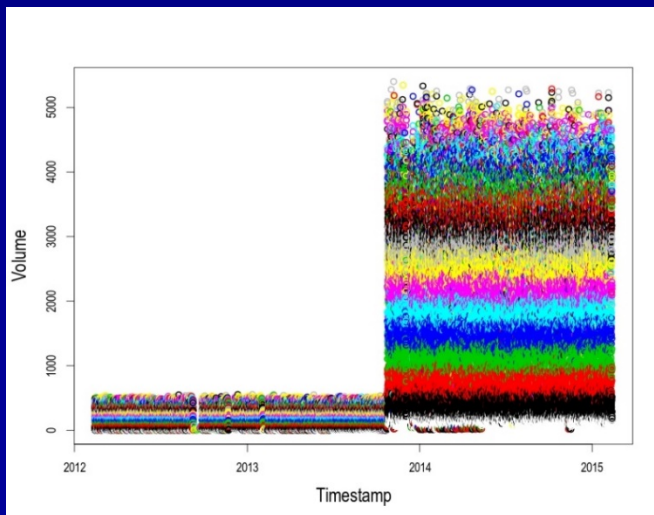
(Wrong Calendar?)



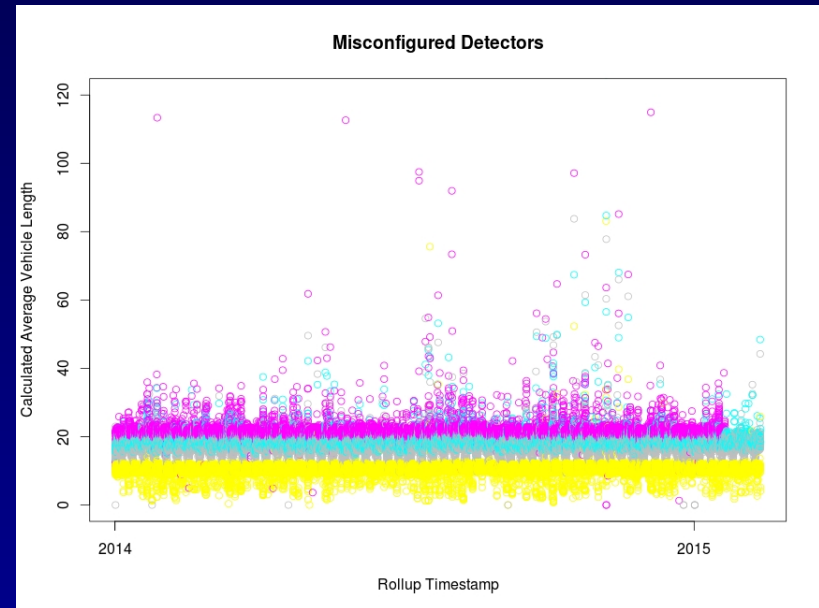


# Data Cleansing

- Missing data
- Transitions from 30 second to 5 minute data
- Inconsistent calculated length



Detector configuration change (Volume Data)



Calculated Average Vehicle Length for Single Detector

Vehicle Length (ft) =

$$\frac{\left( \text{Speed} * \text{Occupancy} * \frac{5280}{3600} * \Delta \text{Time} \right)}{\text{Volume} * 100}$$

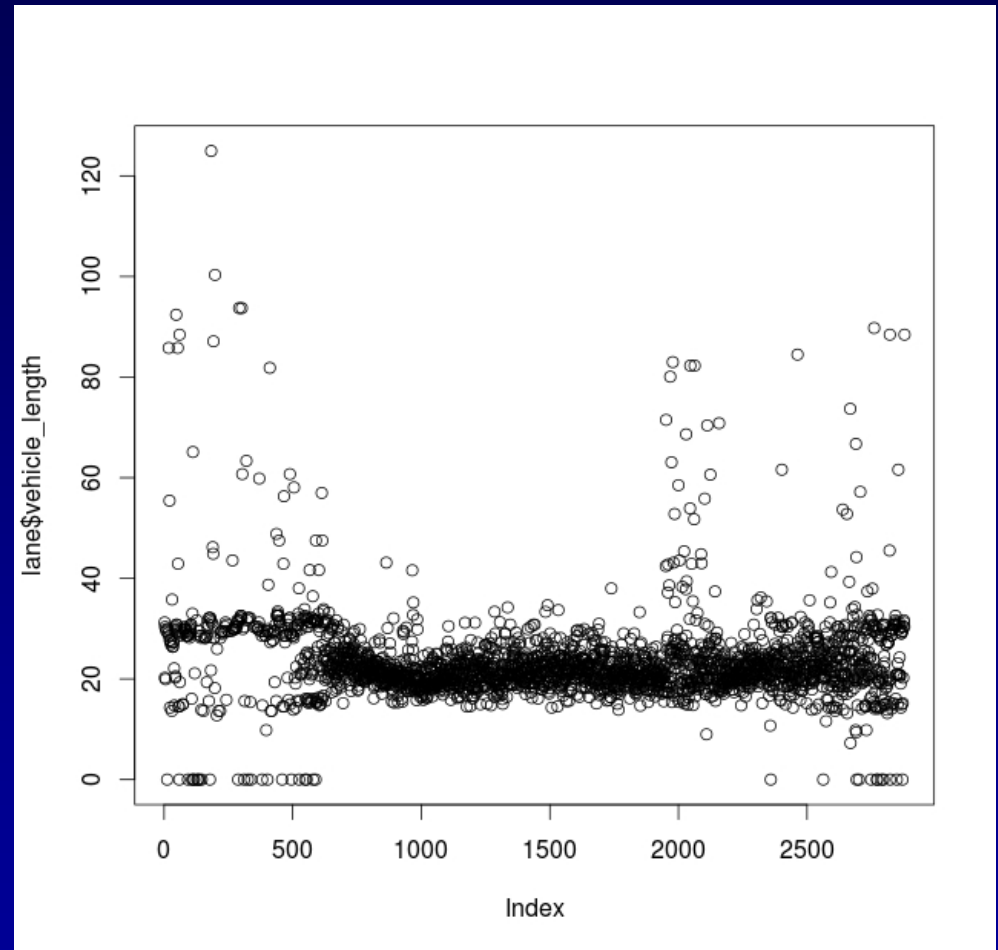
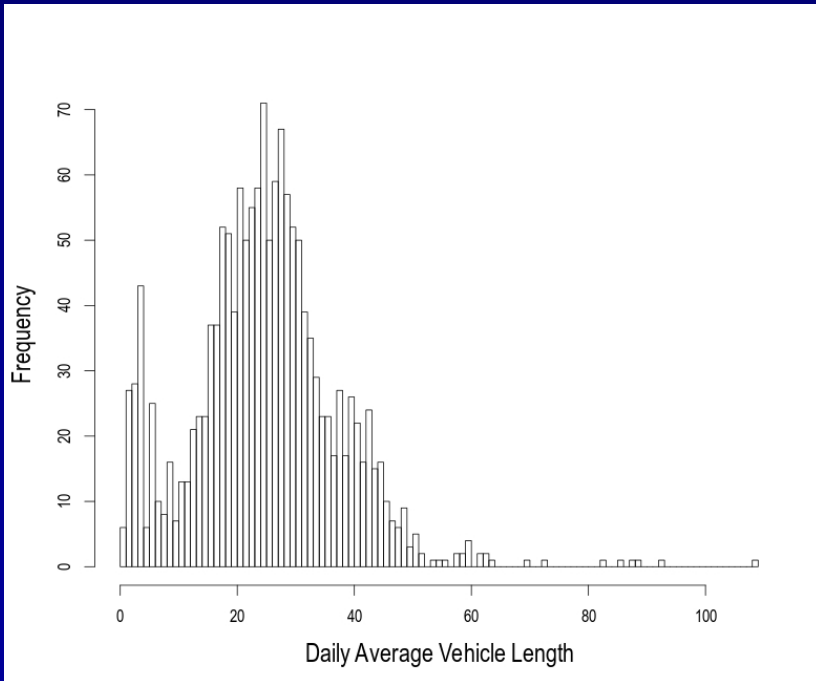
Volume \* 100





# Data Cleansing Calculated Lengths

- **Great disparities**
- **Consistent for some lanes**



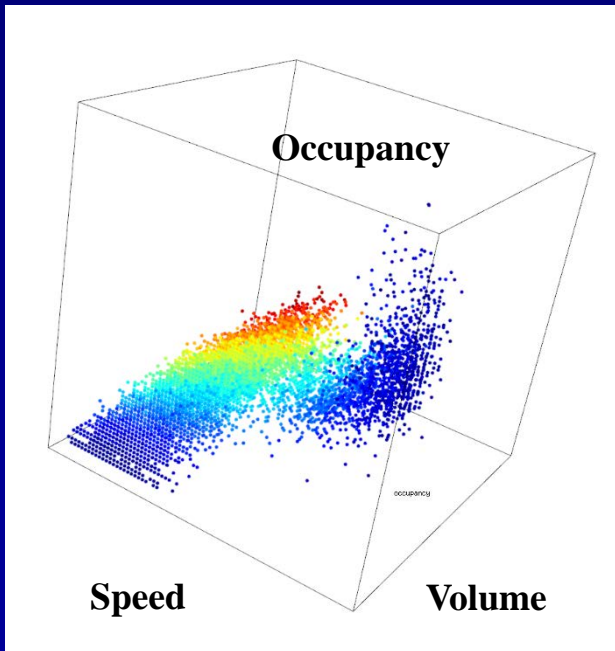
**Calculated vehicle length over time (one detector)**

**Short and Long Vehicle Length Outliers**

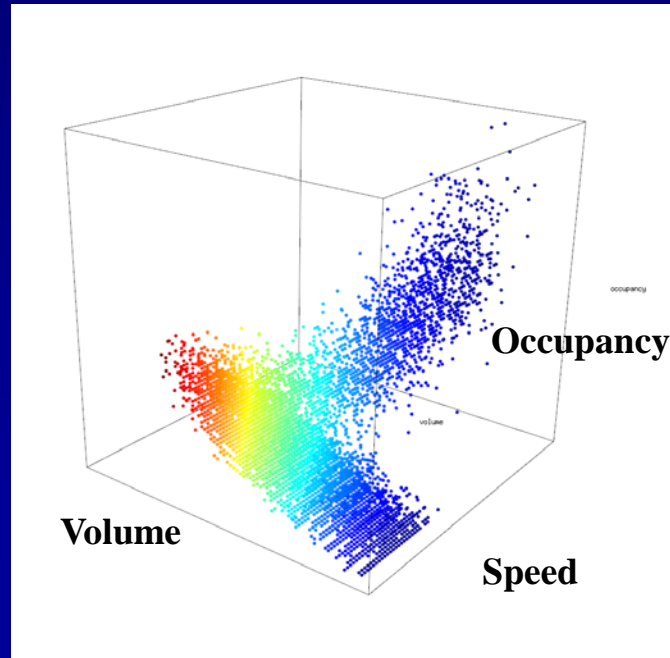


# Analytics - Visualizations

- **Speed-Volume-Occupancy curves**
- **Speed-time-distance with events**



**Speed-Volume-Occupancy-Throughput**



**Speed-Volume-Occupancy-Throughput**

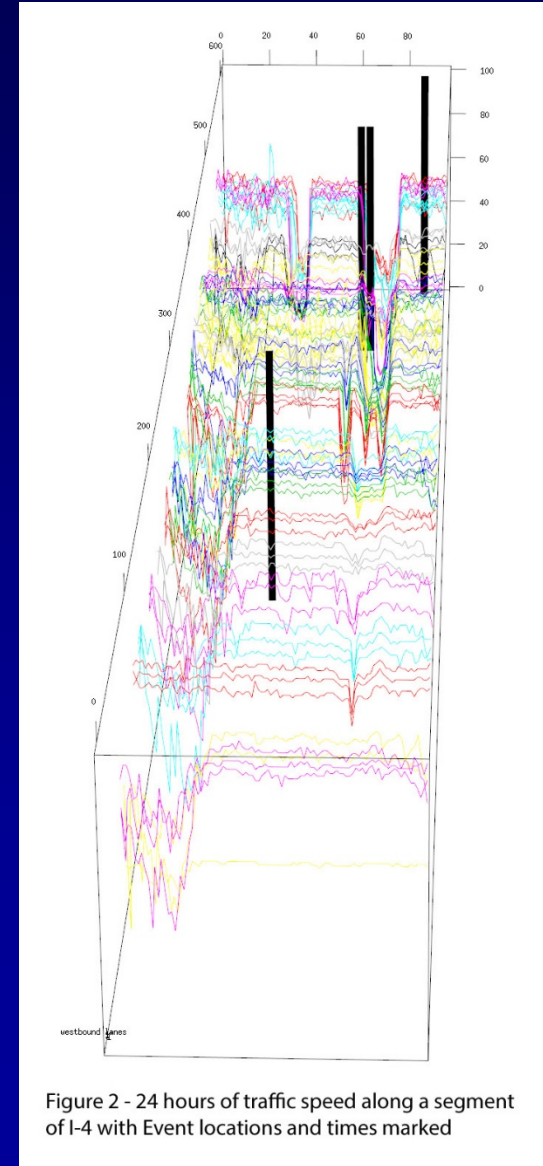
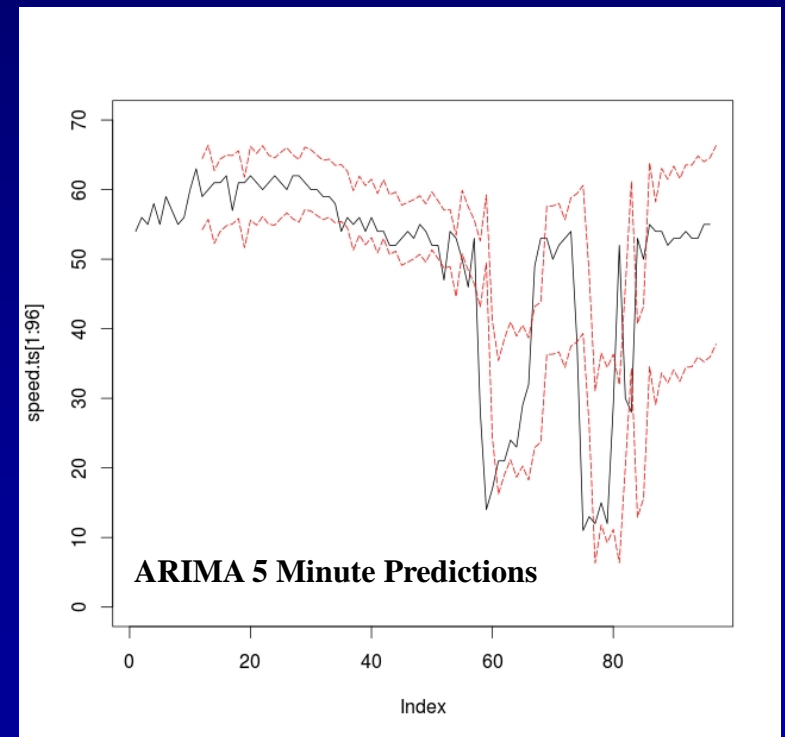
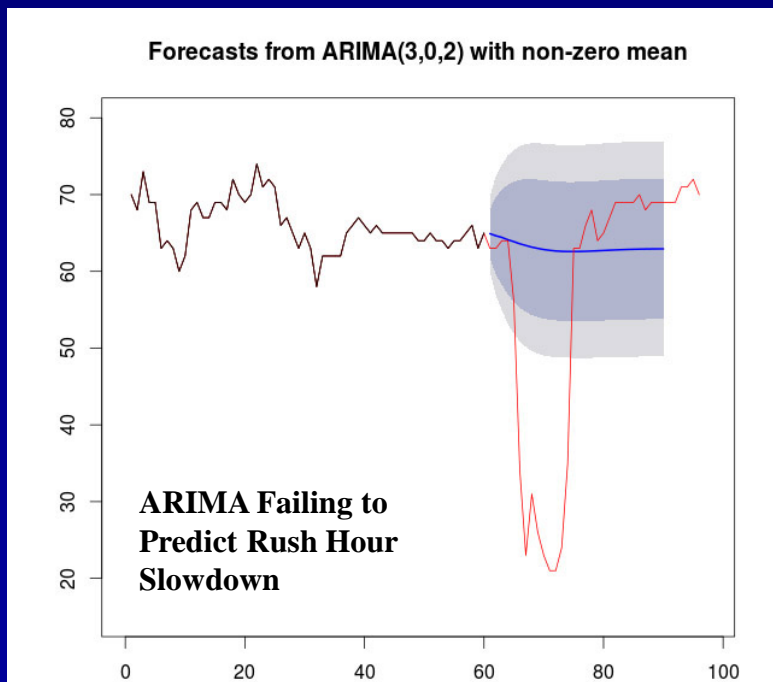


Figure 2 - 24 hours of traffic speed along a segment of I-4 with Event locations and times marked



# Analytics - ARIMA

- **AutoRegressive Integrated Moving-Average (ARIMA) model**
- **Very good steady-state predictions**
- **Lags at transitions**





# *Analytics – ARIMA Model Selection*

- **Evaluated models generated from several sensors on different days**
- **Evaluated their root mean square error on different sensors and days**
- **Selected the ARIMA (0,1,4) model as most general**

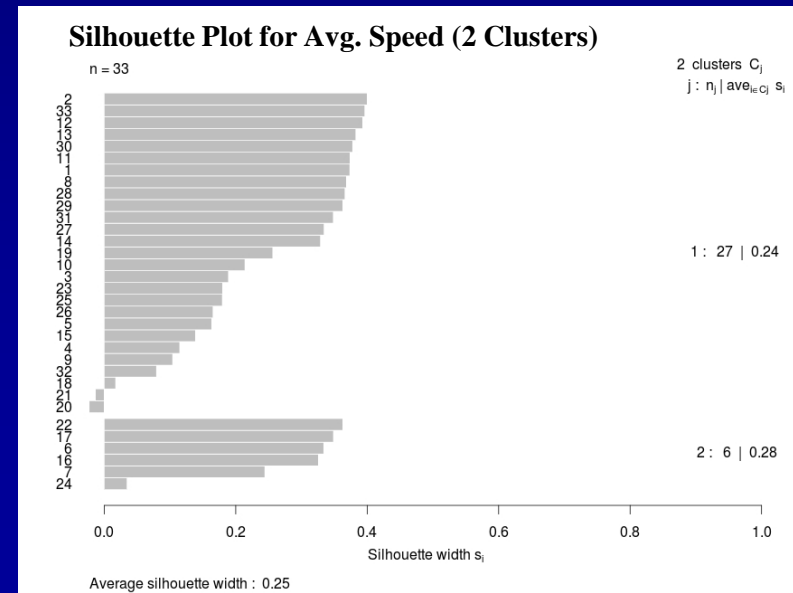
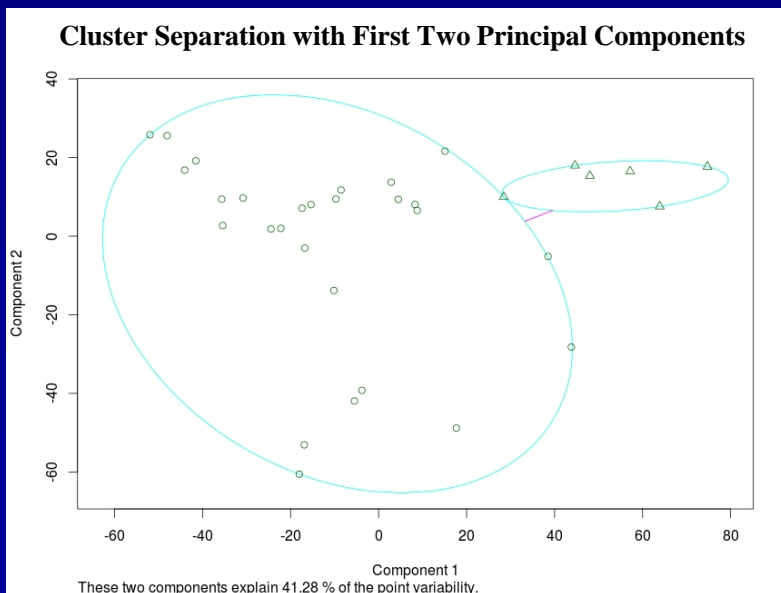
Evaluation of ARIMA Models

model	2784	2835	3469	3678	3920	4353	5454	5579	avg.
Arima(1,1,1)	3.435	4.787	5.445	6.661	3.799	3.959	5.515	3.263	4.608
Arima(0,1,0)	4.619	6.315	7.438	9.025	2.5	3.316	7.419	4.358	5.624
Arima(0,1,4)	3.465	4.804	5.529	6.749	3.513	3.718	5.593	3.297	4.584
Arima(5,1,0)	3.647	5.026	5.812	7.105	3.263	3.722	5.907	3.472	4.744



# Analytics - PAM

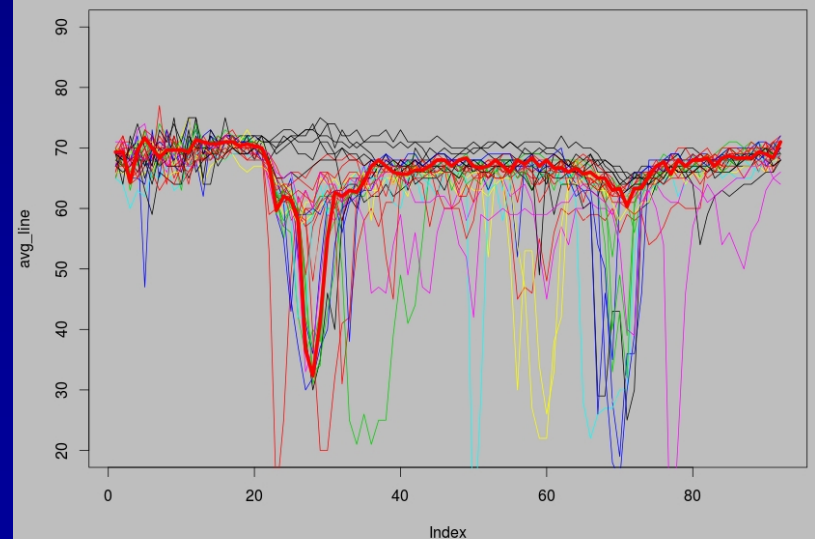
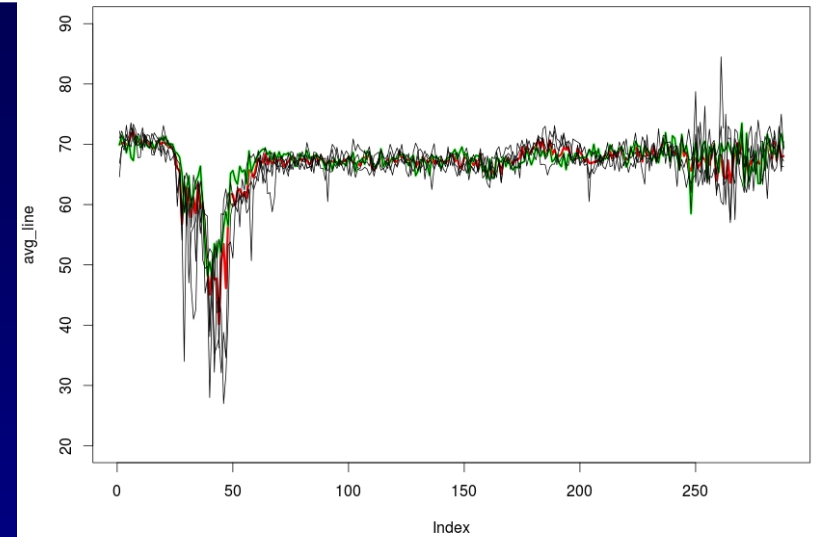
- **Partitioning Around Medoids**
  - **Distance-based clustering of curves**
  - **Medoid is the curve closest to the “center” of the cluster**
  - **The fit of the cluster is measured using the Silhouette**





# *Unperturbed Day Generation*

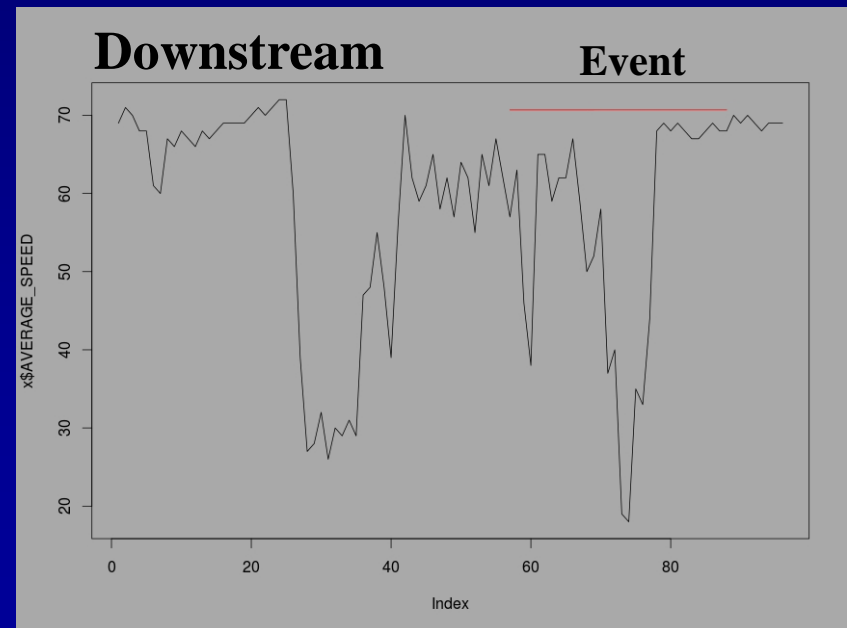
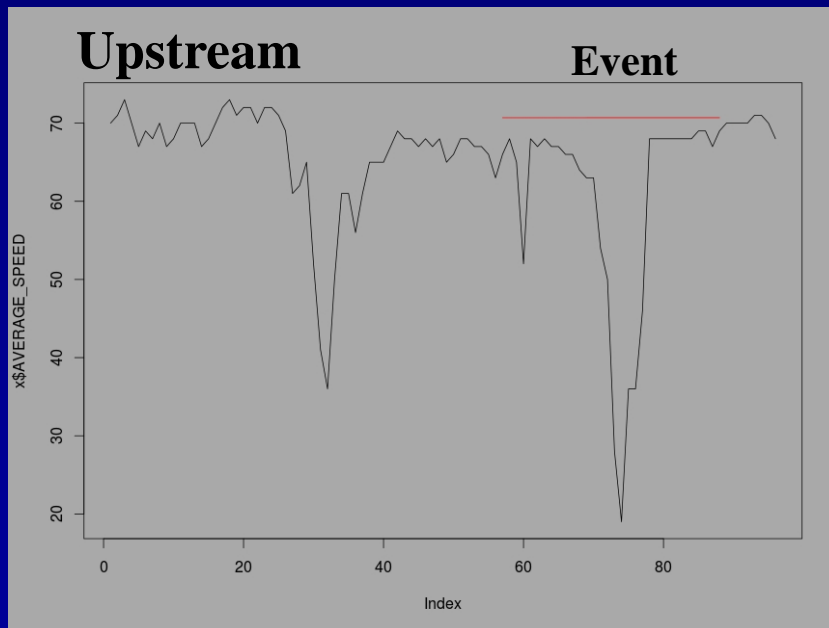
- **Used PAM to find most frequent cluster**
- **Used the curves within 5-10 percent of the distance of the medoid to generate unperturbed day**
- **Threw out most extreme day for each 15-minute interval**
- **Used the average of the remaining points to define the unperturbed day**





# *Analytics - Events*

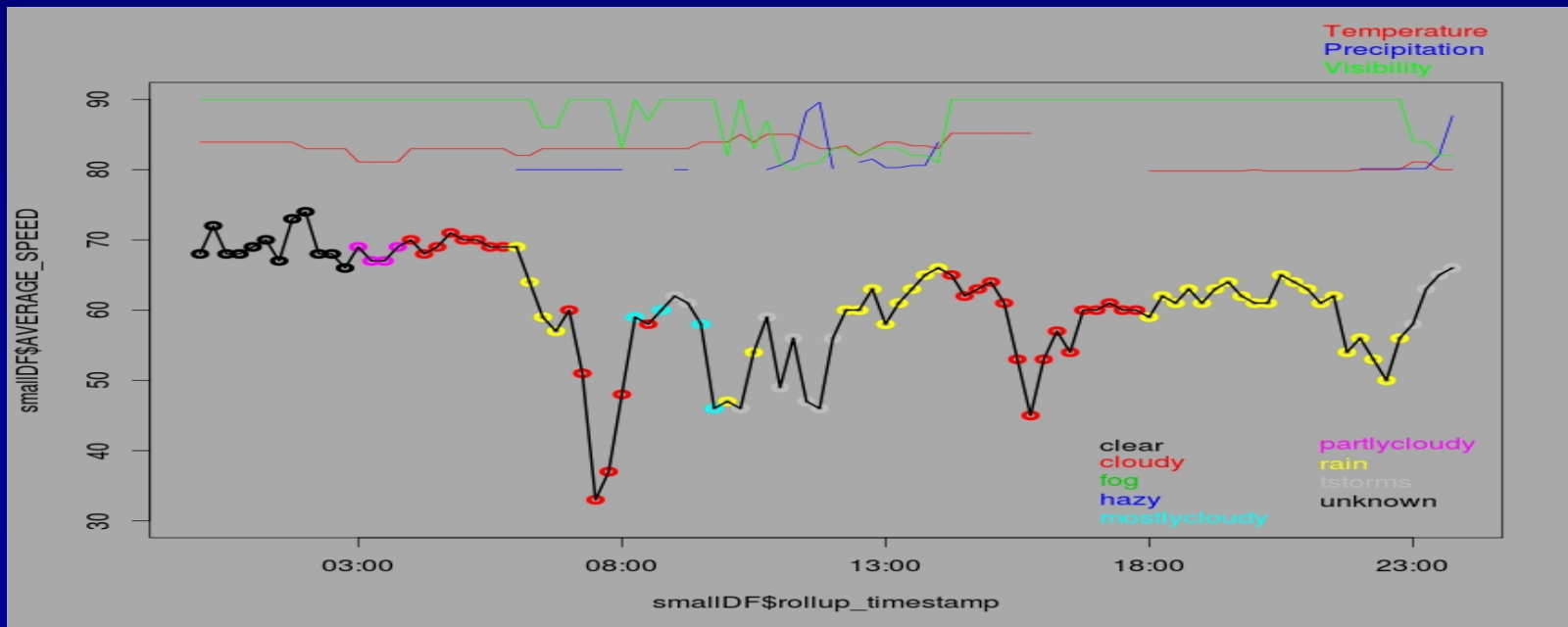
- **Events are not at detectors**
- **Event start and stop times are not well recorded**
- **Events may have impacts upstream and downstream of the event itself**





# Analytics - Weather

- Weather has measurable impacts on traffic predictions
- Real-time and future weather data may be more granular than historical data







## *Summary*

---

- **Data from a wide variety of sources is needed to predict traffic profiles**
- **Data for predicting traffic profiles frequently needs to be cleansed**
- **Algorithms for predicting traffic profiles must be able to deal with missing data**
- **The algorithm for predicting an unperturbed day provides a baseline for finding events**
- **Events, conditions, and weather have impacts on traffic profiles that must be taken into account to get good predictions**



# Technologies Used

**Spark** 1.5.2 Spark Master at spark://

URL: spark:  
 REST URL: spark (cluster mode)  
 Alive Workers: 2  
 Cores in use: 16 Total, 0 Used  
 Memory in use: 28.9 GB Total, 0.0 B Used  
 Applications: 0 Running, 3 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

**Workers**

Worker Id	Address	State	Cores	Memory
worker-20160325200439-		ALIVE	8 (0 Used)	14.4 GB (0.0 B Used)
worker-20160325200605-		ALIVE	8 (0 Used)	14.4 GB (0.0 B Used)

**Running Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20160327145138-0002	Create Rollup	24	1024.0 MB	2016/03/27 14:51:38	dataminer	FINISHED	33 min
app-20160325235534-0001	CSV -> Dataframe (parquet)	24	1024.0 MB	2016/03/25 23:55:34	dataminer	FINISHED	33 min
app-20160325200822-0000	CSV -> Dataframe (parquet)	24	1024.0 MB	2016/03/25 20:08:22	dataminer	FINISHED	33 min

**Completed Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20160327145138-0002	Create Rollup	24	1024.0 MB	2016/03/27 14:51:38	dataminer	FINISHED	33 min
app-20160325235534-0001	CSV -> Dataframe (parquet)	24	1024.0 MB	2016/03/25 23:55:34	dataminer	FINISHED	33 min
app-20160325200822-0000	CSV -> Dataframe (parquet)	24	1024.0 MB	2016/03/25 20:08:22	dataminer	FINISHED	33 min

<b>Spark SQL</b> (SQL)	<b>Spark Streaming</b> (Streaming)	<b>MLlib</b> (Machine learning)	<b>GraphX</b> (Graph computation)
<b>Spark</b> (General execution engine)			



**Hadoop** Overview Datanodes Snapshot Startup Progress Utilities

**Summary**

Security is off.  
 Safemode is off.  
 2876 files and directories, 5575 blocks = 8451 total filesystem object(s).  
 Heap Memory used 70.78 MB of 111.5 MB Heap Memory. Max Heap Memory is 889 MB.  
 Non Heap Memory used 42.12 MB of 61.69 MB Committed Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	14.44 TB
DFS Used:	774.63 GB
Non DFS Used:	1.71 TB
DFS Remaining:	11.98 TB
DFS Used%:	5.24%
DFS Remaining%:	82.94%
Block Pool Used:	774.63 GB
Block Pool Used%:	5.24%
DataNodes usages% (Min/Median/Max/stdDev):	3.33% / 6.38% / 7.12% / 1.57%
Live Nodes	4 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	2/9/2016, 2:14:14 PM





*Thank you!*

---

**Questions?**

**Contact Information**

**David Vickers:**

[david.vickers@swri.org](mailto:david.vickers@swri.org)

**Adam Van Horn:**

[avanhorn@swri.org](mailto:avanhorn@swri.org)

**SwRI ATMS:**

<http://www.swri.org/4org/d10/isd/default.htm>