Pivotal

A NEW PLATFORM FOR A NEW ERA

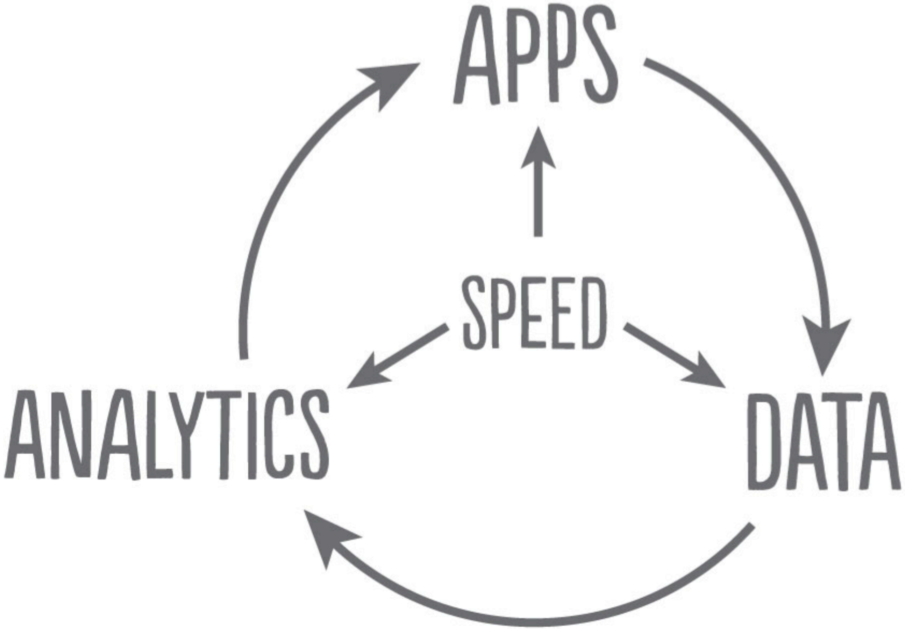
Driving the Future of Smart Cities
How to Beat the Traffic

Strata Santa Clara – February 13, 2013

Alexander Kagoshima, Data Scientist, @akagoshima
Noelle Sio, Senior Data Scientist, @noellesio
Ian Huston, Data Scientist, @ianhuston

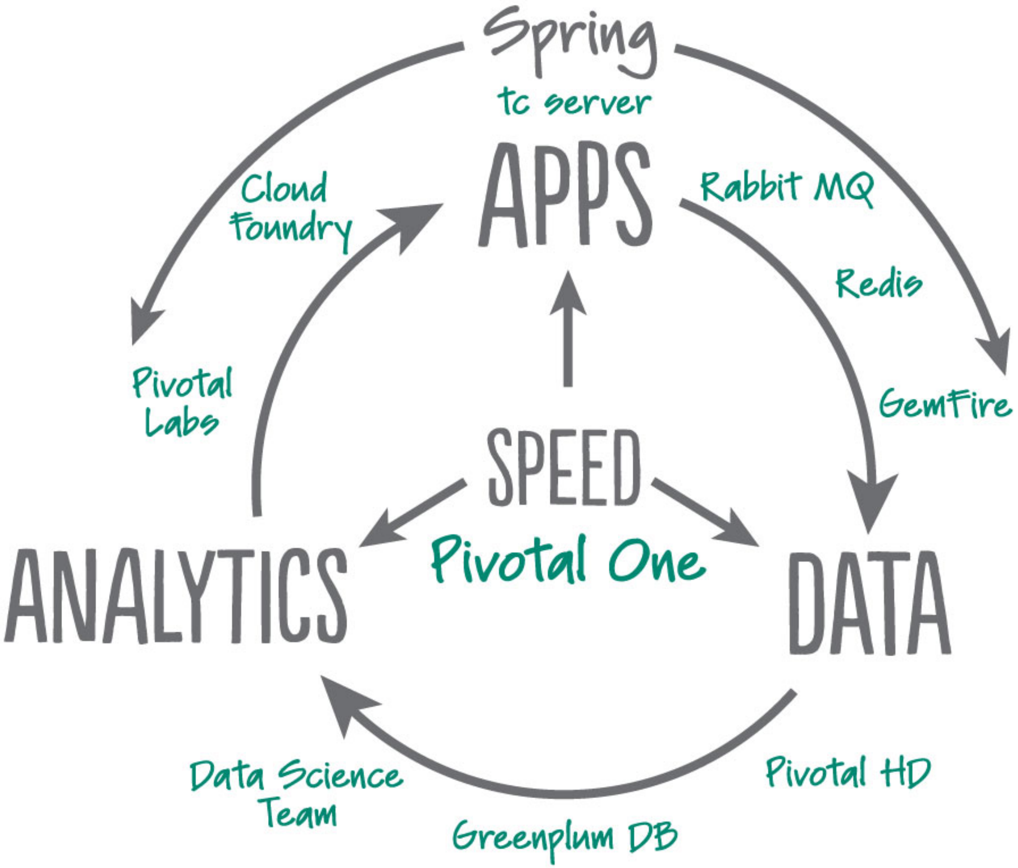
What Matters: Apps. Data. Analytics.

- ☑ *Apps power businesses, and those apps generate data*
- ☑ *Analytic insights from that data drive new app functionality, which in-turn drives new data*
- ☑ *The faster you can move around that cycle, the faster you learn, innovate & pull away from the competition*

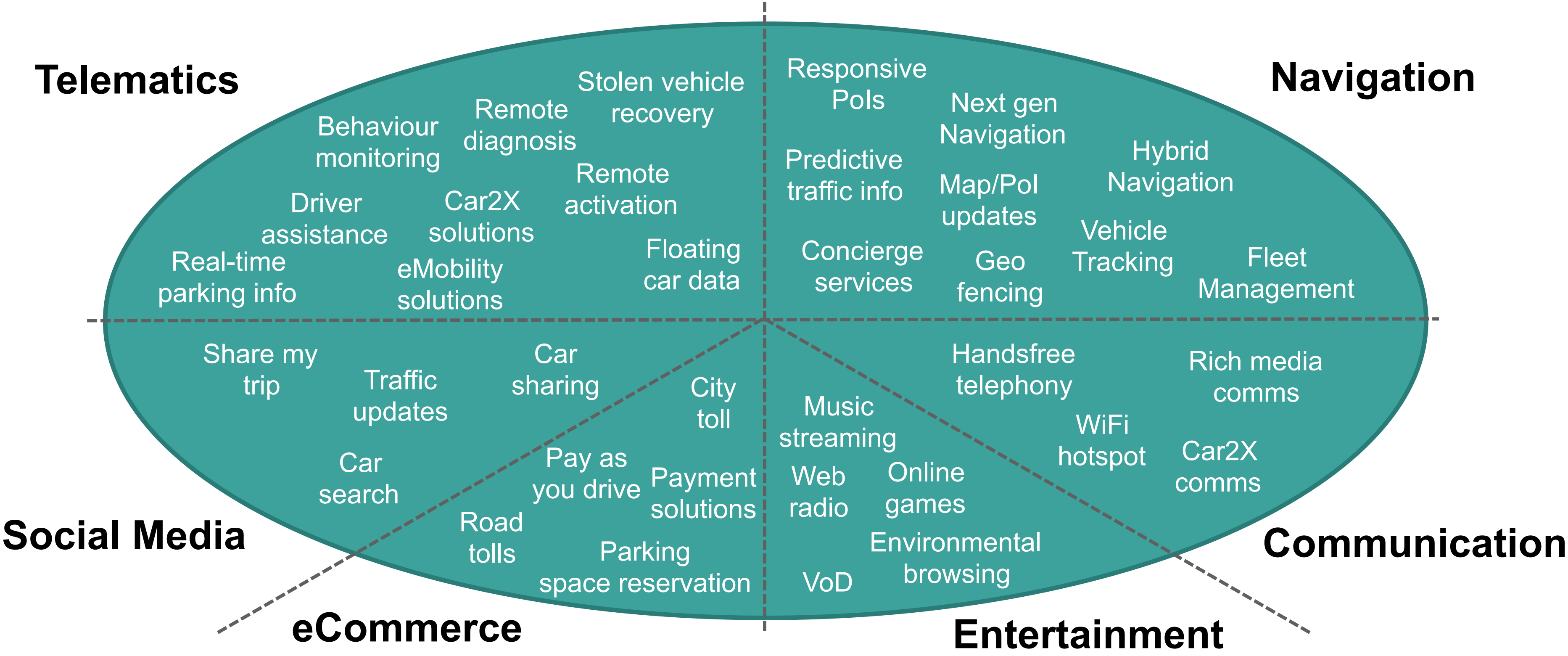


Pivotal's Opportunity

- *Uniquely positioned to help enterprises modernize each facet of this cycle today*
- *Comprehensive portfolio of products spanning Apps, Data & Analytics*
- *Converging these technologies into a coherent, next-gen Enterprise PaaS platform*



The Connected Car Drives Innovation

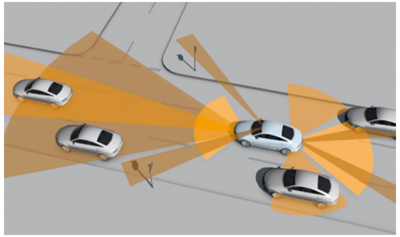


@noellesio, @gopivotal

Pivotal

Possible Data Science Use-Cases

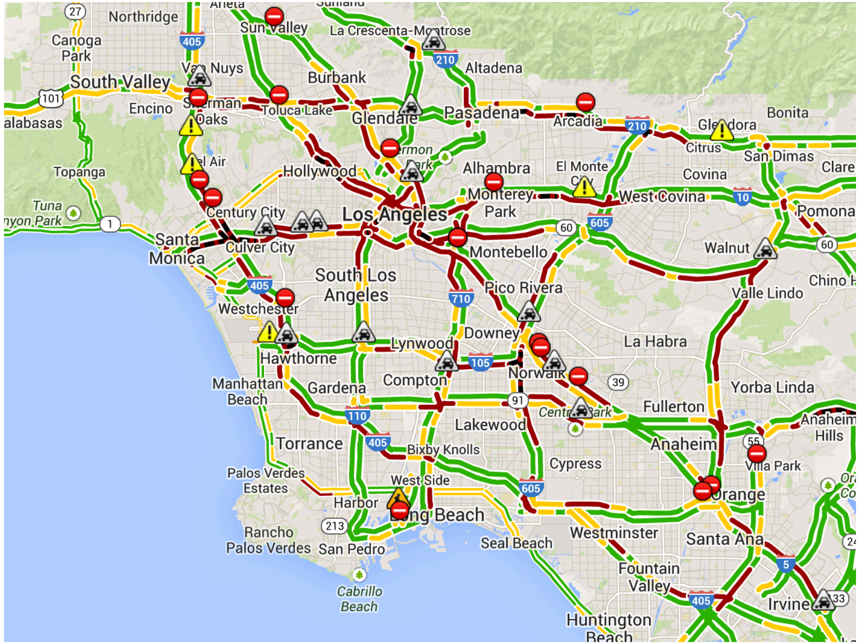
- **Predictive Car Maintenance**
  - More accurately predict part failure
  - Optimize part repair and replacement schedule
- **Leveraging Driving Behavior**
  - Useful to differentiate insurance pricing based on driving style
  - Optimize car design
- **Improving GPS Systems**
  - Establish baseline for traffic congestion
  - Gain a detailed view on traffic
  - Create more meaningful metrics for routing
- **Predictive Power for Assistance Systems**
  - Optimize fuel efficiency
  - Predict the future state of a car in the next 2 minutes (starts, stops, emergency braking)
- **Traffic Light Assistance**
  - Signal timing of traffic lights
  - Crowd sourcing of traffic signals

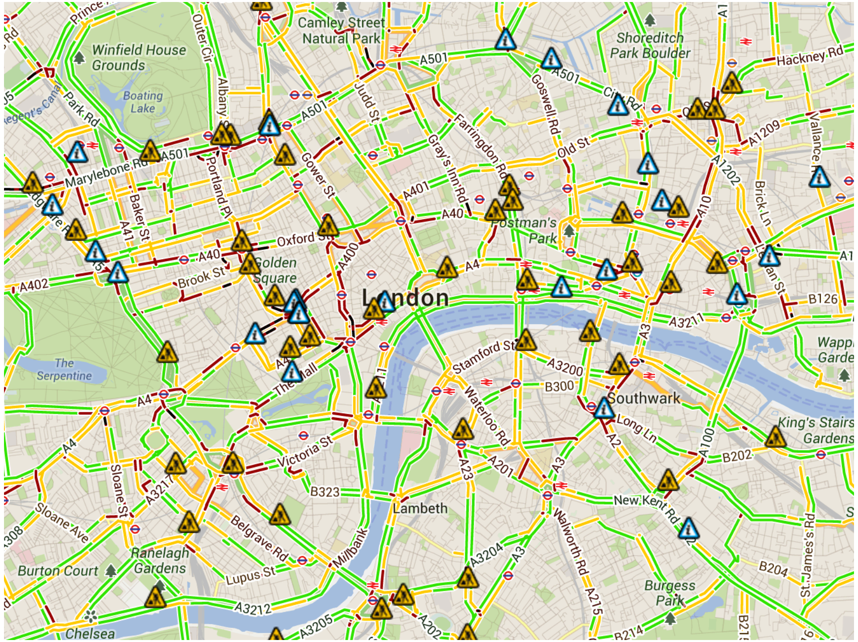


Pivotal

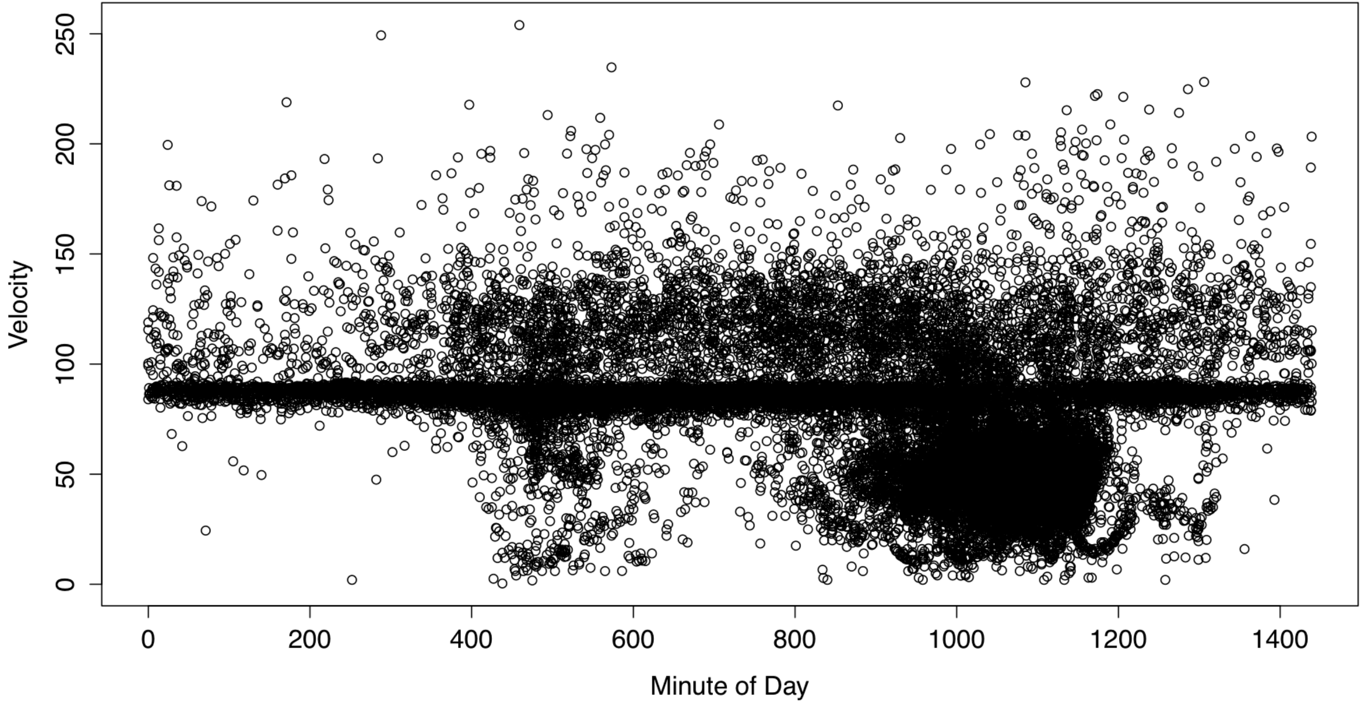
What does traffic data look like?

…like this?

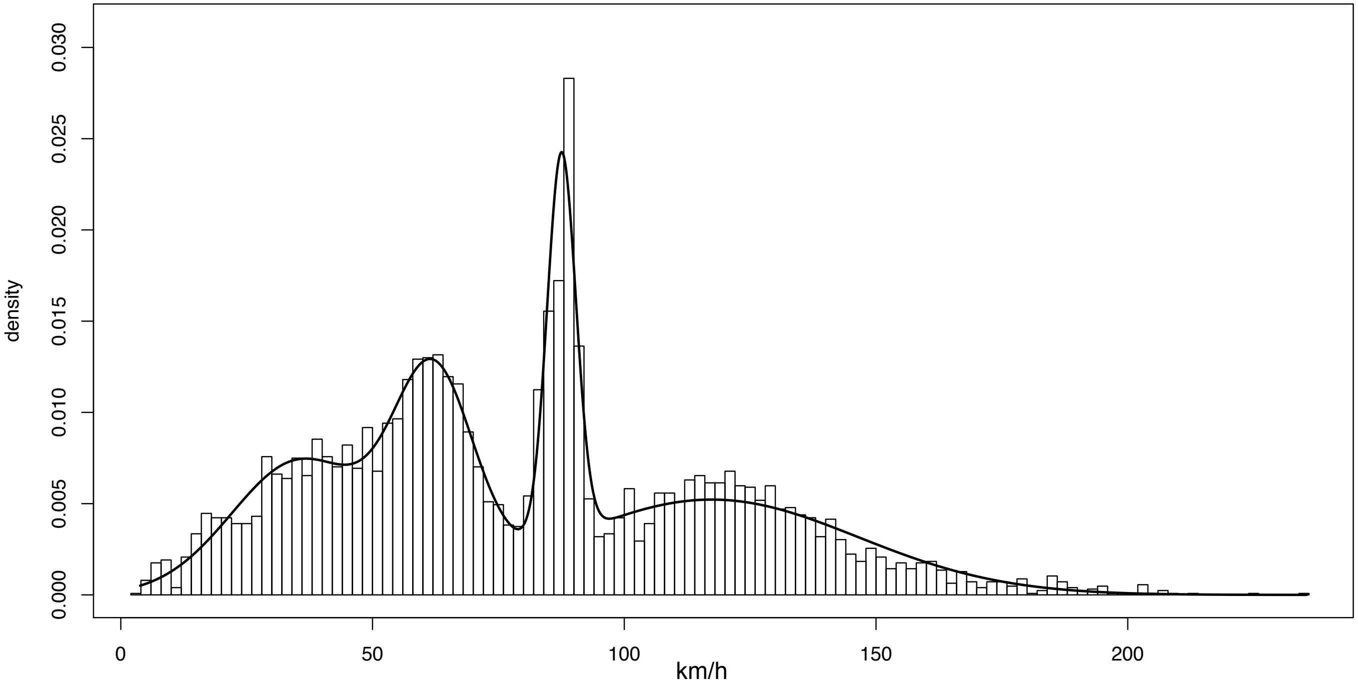

Northridge
Canoga Park
Sun Valley
La Crescenta-Montrose
Altadena
South Valley
Burbank
Encino
Toluca Lake
Glendale
Pasadena
Arcadia
Glendora
Bonita
Citrus
San Dimas
Topanga
Hollywood
Alhambra
El Monte
Covina
Monterey Park
West Covina
Pomona
Los Angeles
Century City
Santa Monica
Culver City
Montebello
Walnut
South Los Angeles
Pico Rivera
Valle Lindo
Westchester
Downey
La Habra
Lynwood
Hawthorne
Norwalk
Manhattan Beach
Gardena
Compton
Yorba Linda
Fullerton
Lakewood
Torrance
Bixby Knolls
Anaheim Hills
Anaheim
Cypress
Villa Park
Palos Verdes Estates
West Side
Harbor
Orange
Rancho Palos Verdes
San Pedro
Seal Beach
Westminster
Santa Ana
Fountain Valley
Cabrillo Beach
Irvine
Huntington Beach


Camley Street Natural Park
Shoreditch Park Boulder
Winfield House Grounds
Boating Lake
Hackney Rd
Marylebone Rd
Oxford St
Golden Square
Hoxton Park
London
The Serpentine
Stamford St
Southwark
King's Stairs Gardens
Victoria St
Lambeth
Sloane Ave
Ranelagh Gardens
Burton Court
Burgess Park
Chelsea

How fast are vehicles moving?

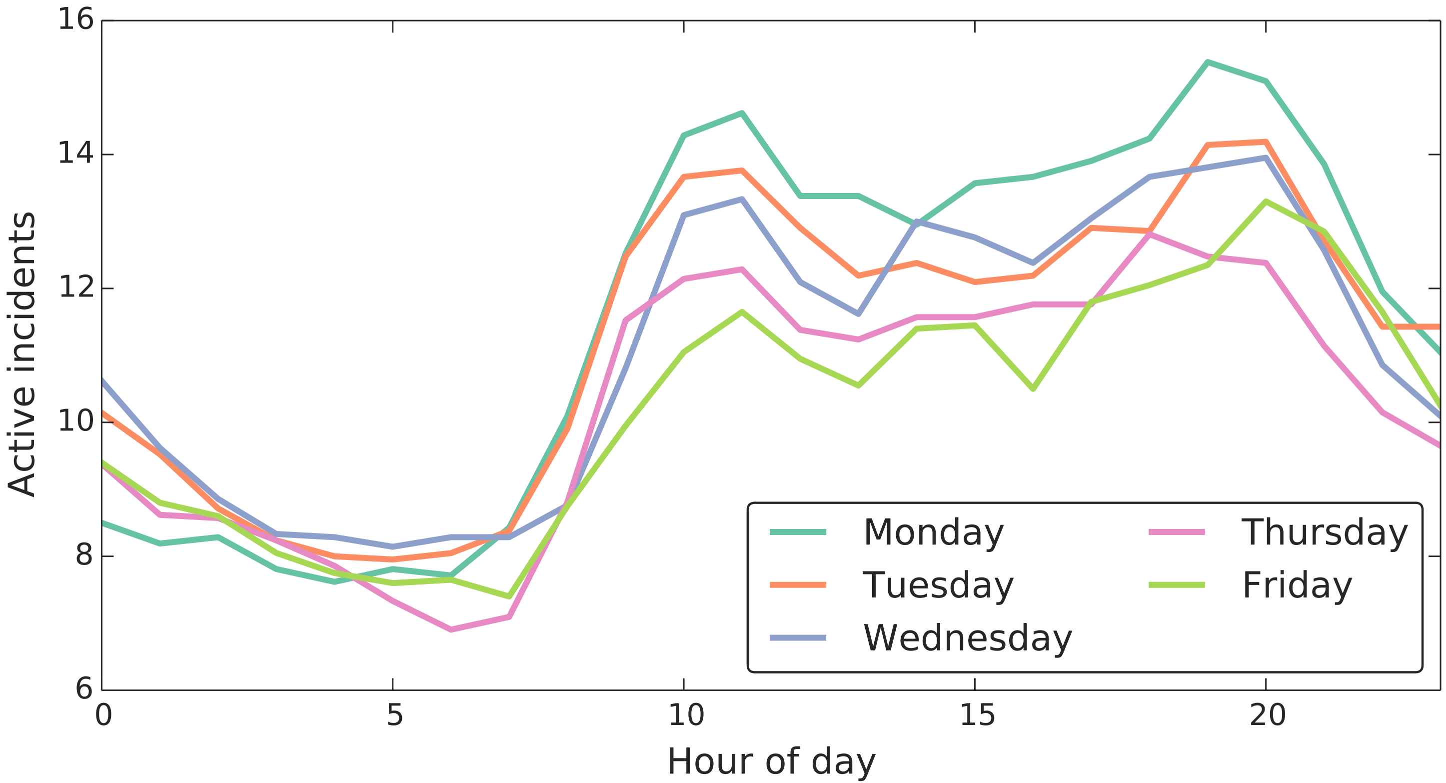


How fast are vehicles moving?



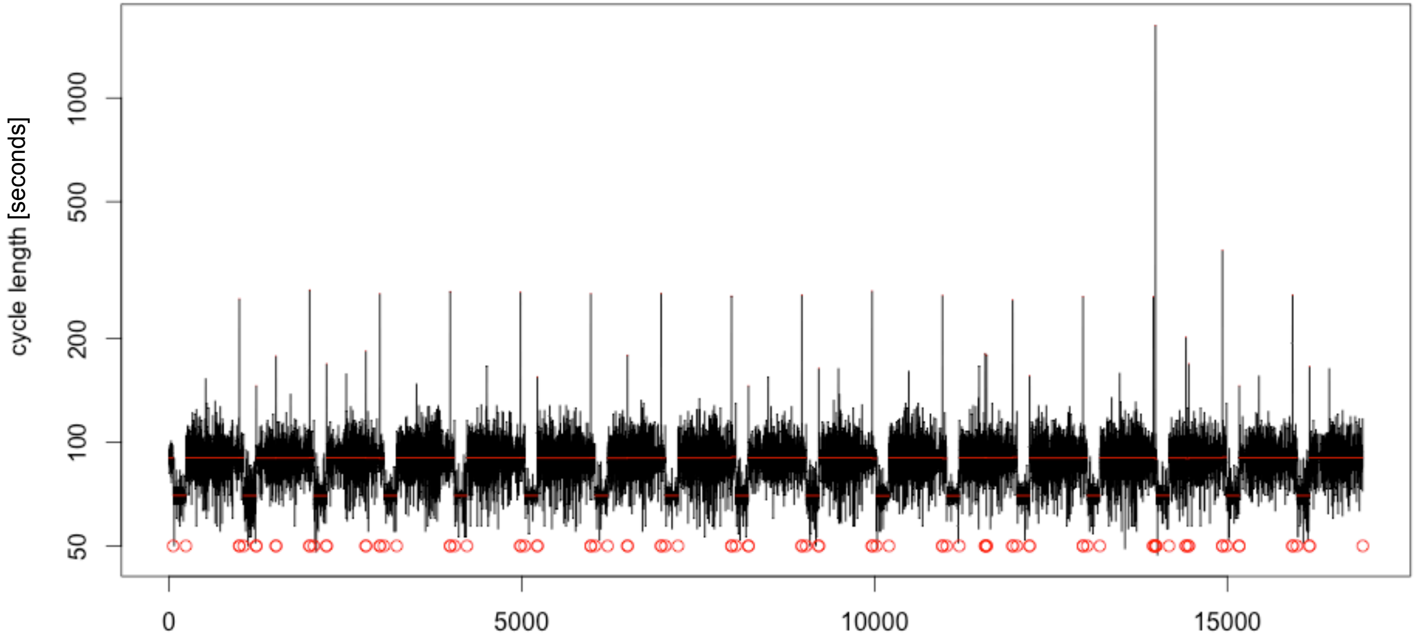
Pivotal™

When do disruptions happen?



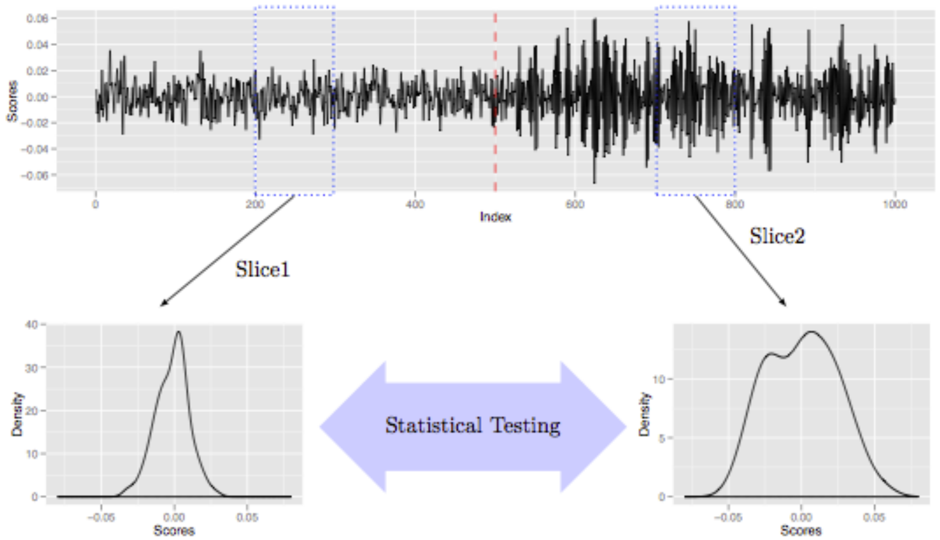
Pivotal

When will the light change?



Pivotal

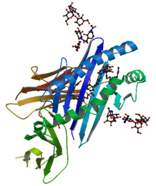
Taking Lessons From Other Disciplines

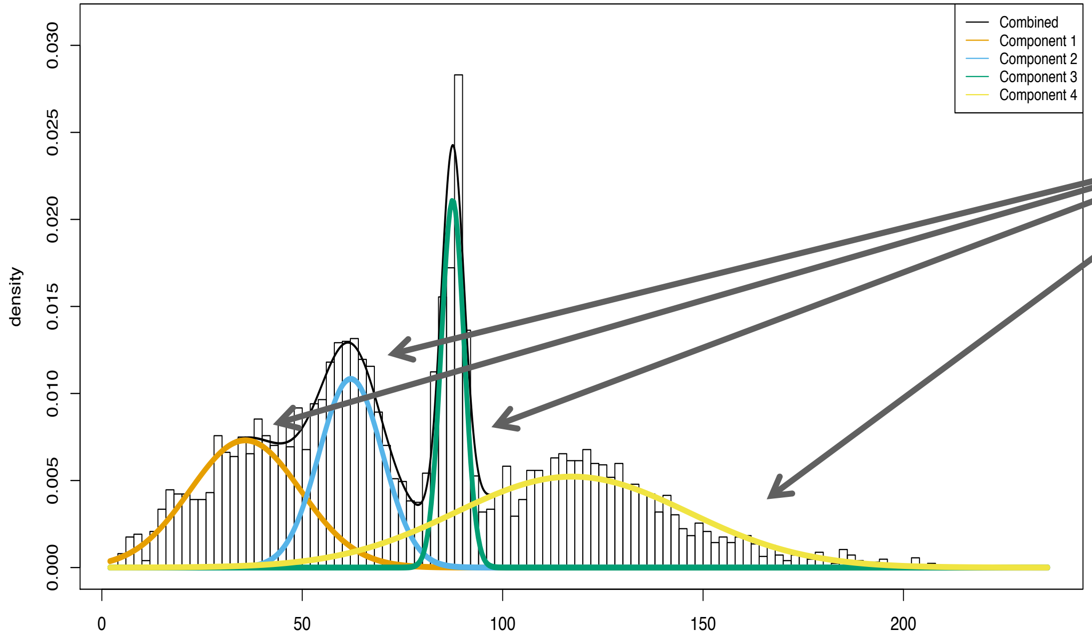


Change-Point Detection can be used to uncover **regimes in wind-turbine data**.

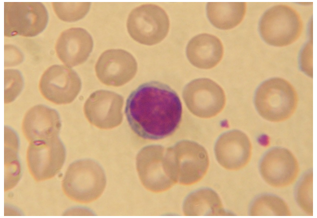
It can also be applied to uncover **regimes in traffic light switching** patterns.

Taking Lessons From Other Disciplines





Different **Cell** Populations



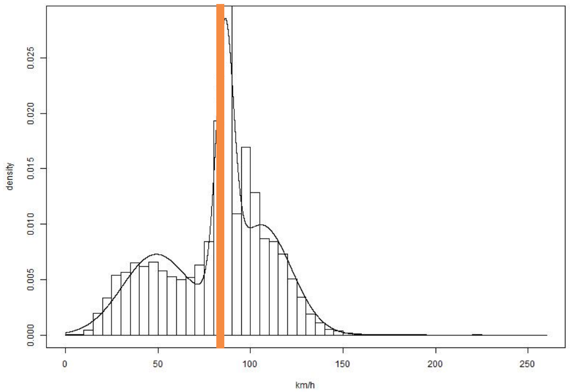
Different **Driving Conditions**

@noellesio, @gopivotal

Pivotal

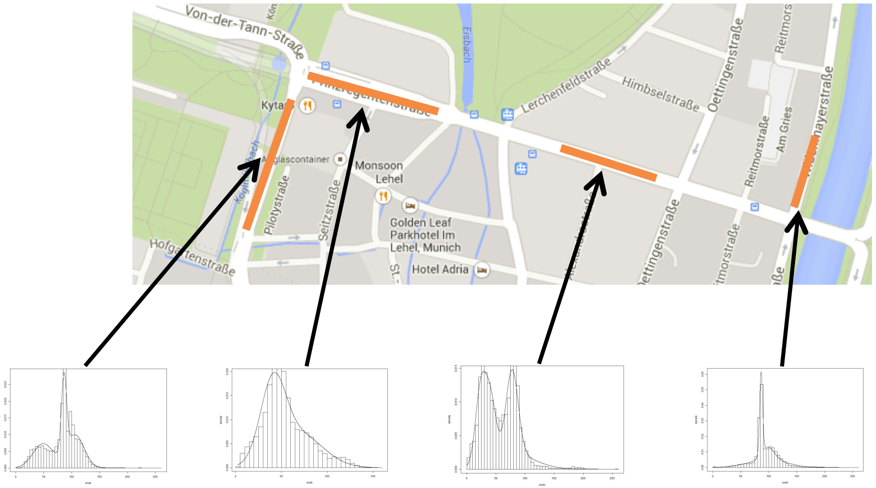
Understanding Traffic Flow

A dynamic, more detailed understanding of traffic is now possible. **Can we answer both '*What velocity?*' and '*Why?*'**



Context

- Current GPS systems are based on average velocity over street segments
- Real-time traffic information (e.g. Waze) <u>does not</u> deliver detailed view nor prediction



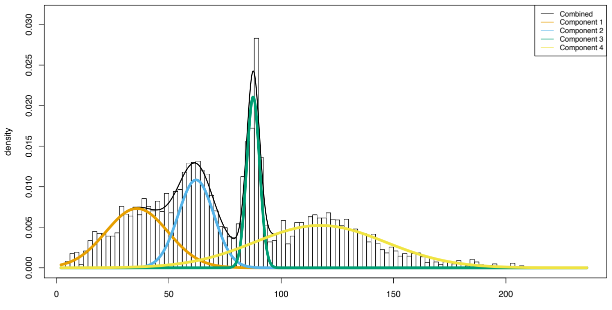
Pivotal

Our Approach – Multi-step algorithm

From our experience, **real-world data** often requires **multi-step procedures**

Step 1: Answer ‘What velocity?’

First find **distinct velocity groups**



Step 2: Answer ‘Why?’

Find **influencing effects**

?

?

?

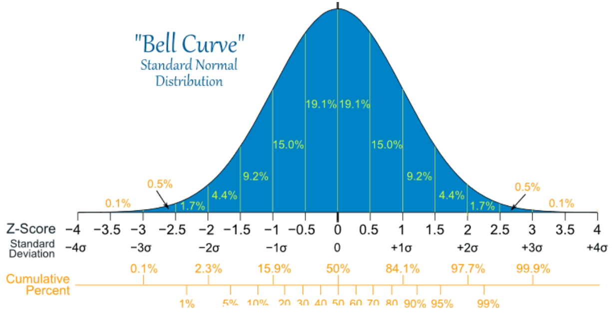
?

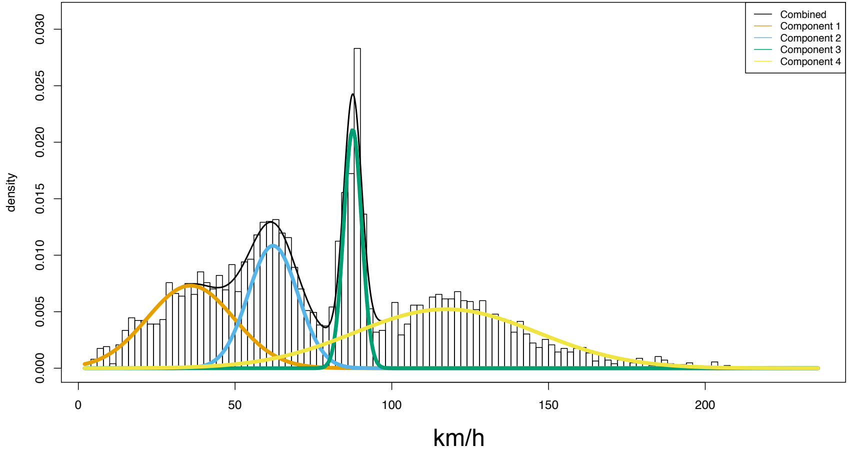
?

?

Find Velocity Groups

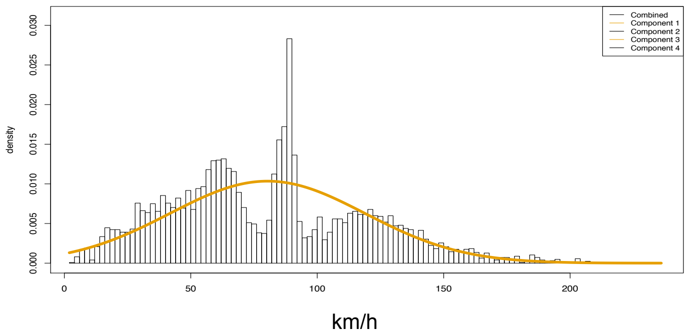
- Velocity distributions can be **fit well** with Gaussians
- An ‘overlay’ of multiple Gaussians is called **Gaussian Mixture Model**
- GMM fitting of the velocity distribution is done by **Expectation-Maximization** algorithm
- Shapes and positions of Gaussians **determine velocity groups**

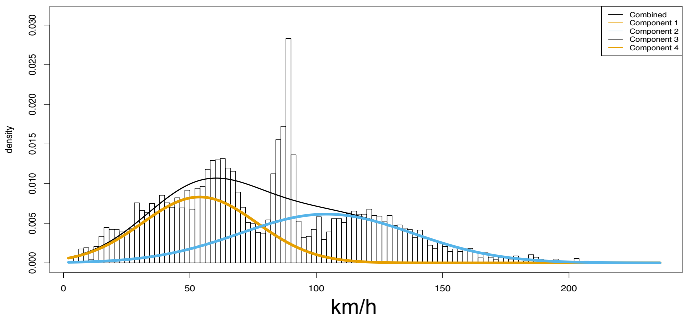


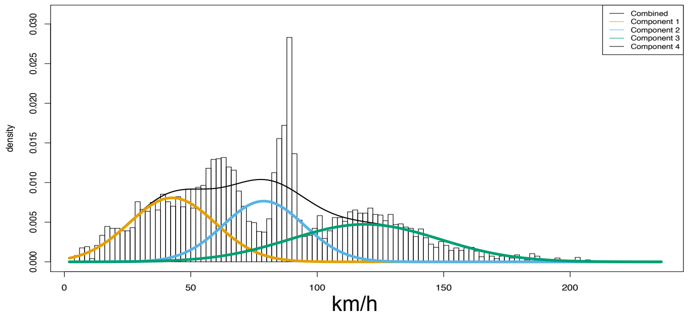


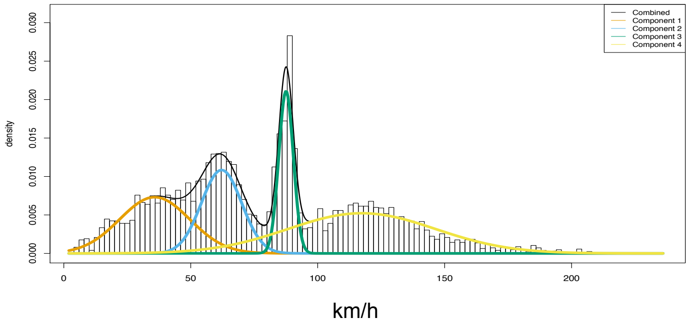
Pivotal

Gaussian Mixture Model





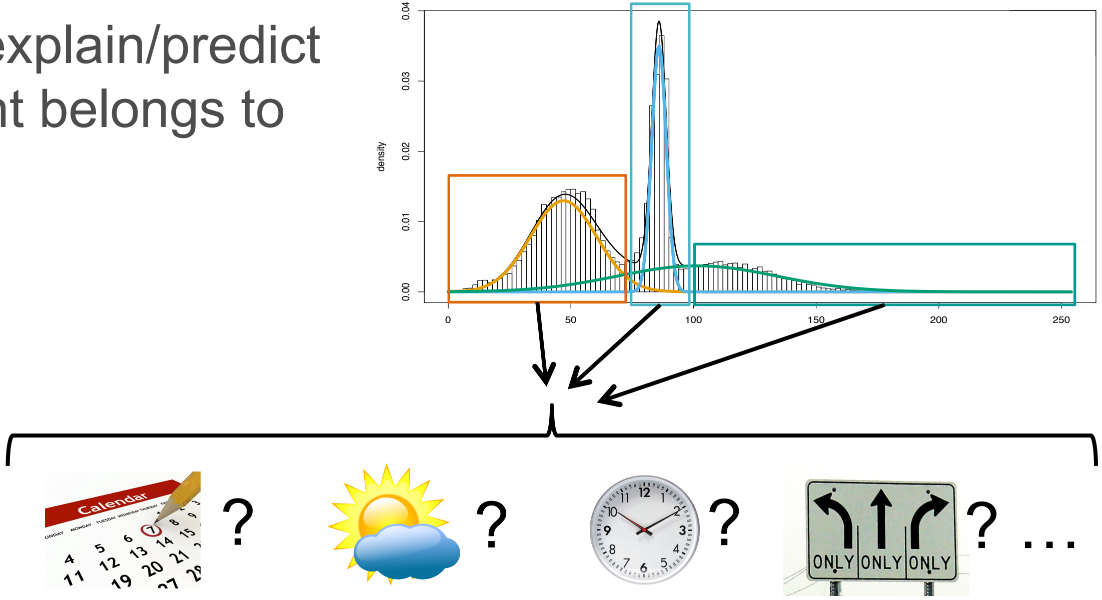




@akagoshima, @gopivotal

Predict Gaussians

- The second step seeks to explain/predict which Gaussian a data point belongs to
  - ➢ Classification task!



- Features for classification:
  - Time of day, day of week
  - Weather
  - Direction
  - Special Events
  - …

White and Black Boxes

- Analyze correlations between **features** of a data point and its **assignment** to a Gaussian
- From a Machine Learning point of view, this is **classification**

- Generate an *interpretable* model description

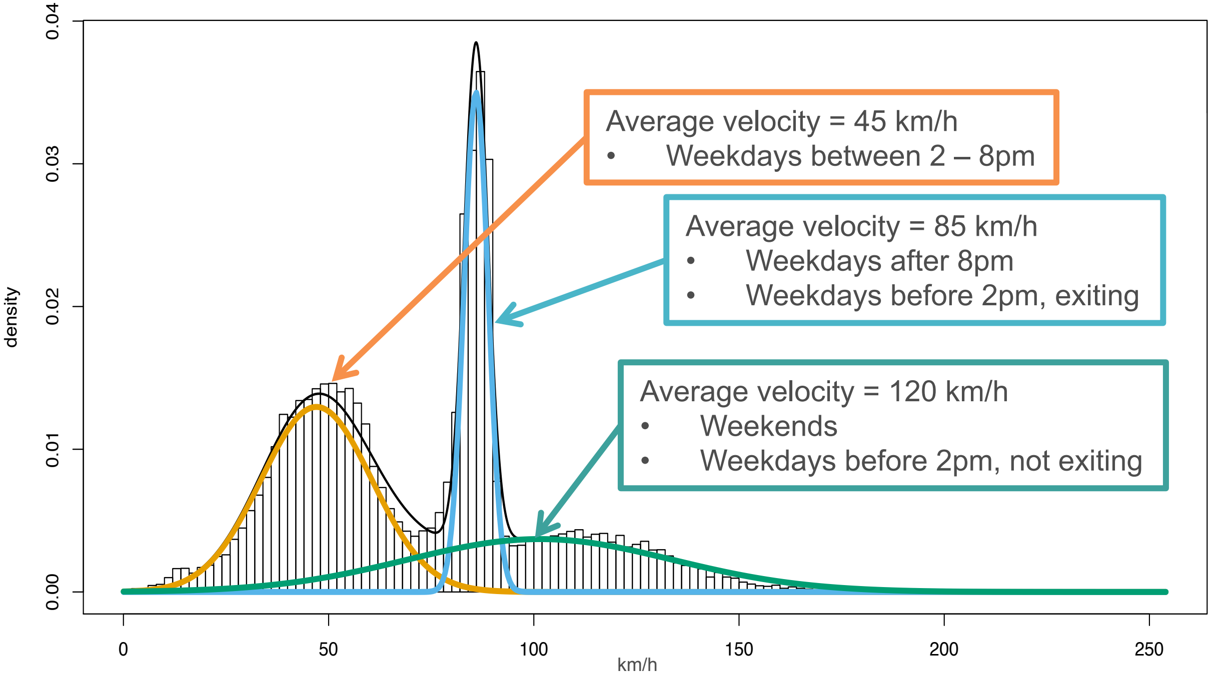
➢ **Explanation** of behavior

- Can capture *more complex correlations*

➢ **Prediction** of Gaussian assignment

Pivotal™

Putting it all together…



Two-Step algorithm: GMM + Classification

- Identified multiple velocity profiles for <u>every</u> road segment
- Intuitive and easily interpretable results
- Highly scalable for more features and data

Better Travel Time Prediction

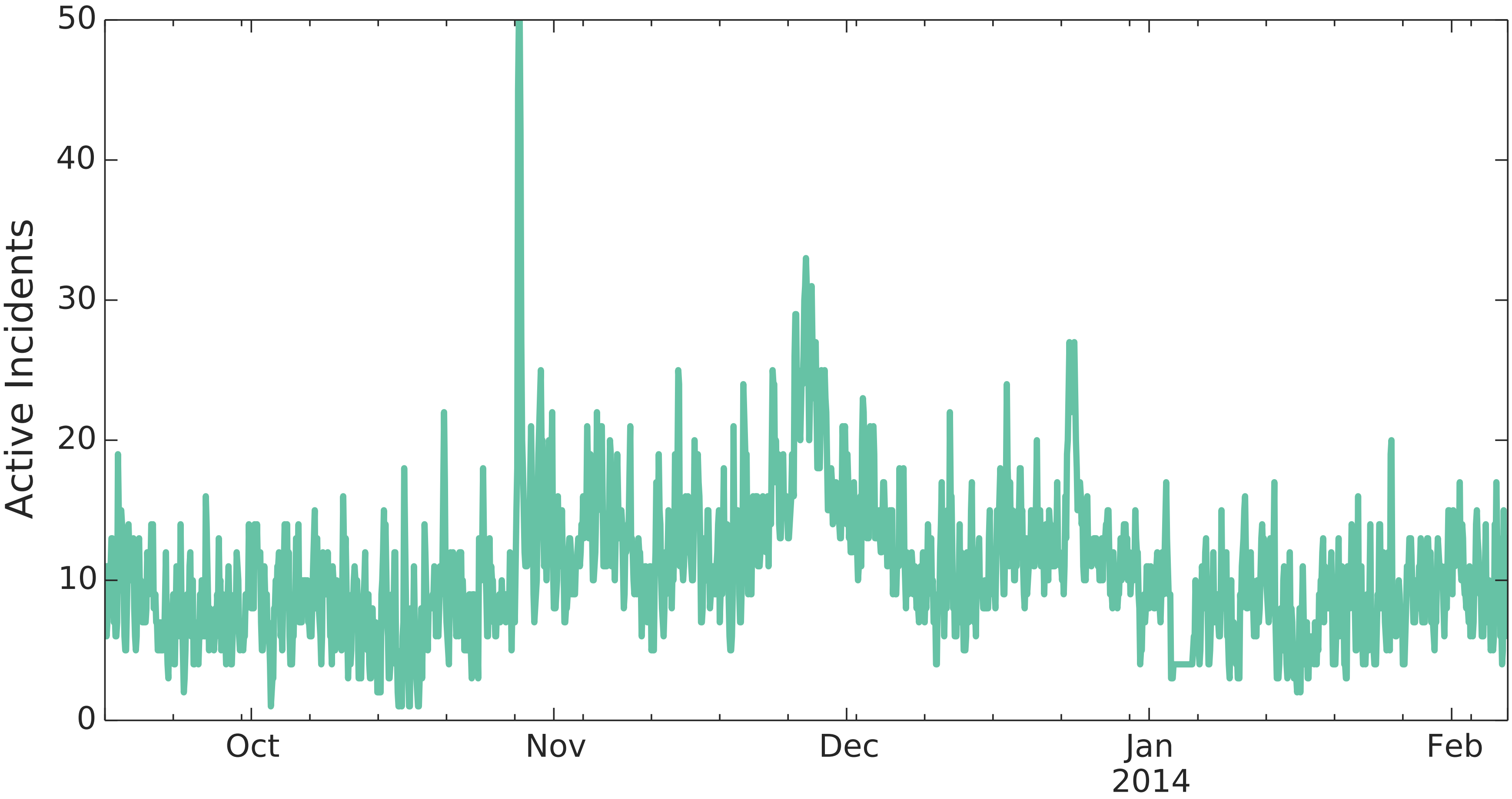
- **Traffic profiles emerged from data**
  - Without using metadata, we uncovered road segment traffic patterns
- **Identified Bias Effects**
  - Inferring the *impact of turns* and *day of week* on velocity
  - Able to predict *rush hour* by day and time by road segment
- **Traffic Light Patterns**
  - Infer *public transportation* effects on traffic
  - Automatically determine different *switching patterns*

London Road Traffic Disruptions

Can we predict when unexpected incidents will end?

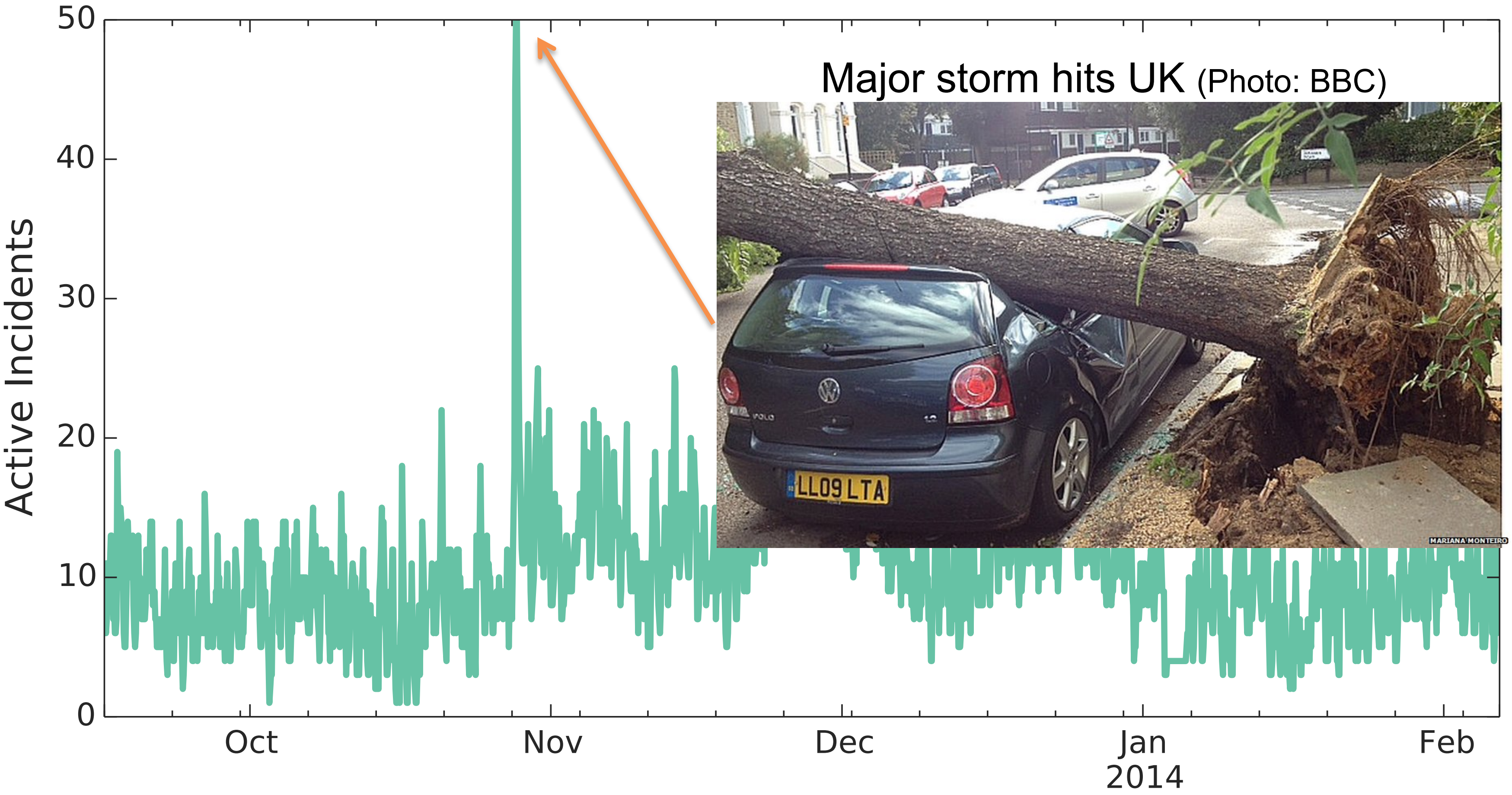
Publicly available data:

- Transport for London traffic feed (refreshed every 5 minutes)
- Weather Underground reports



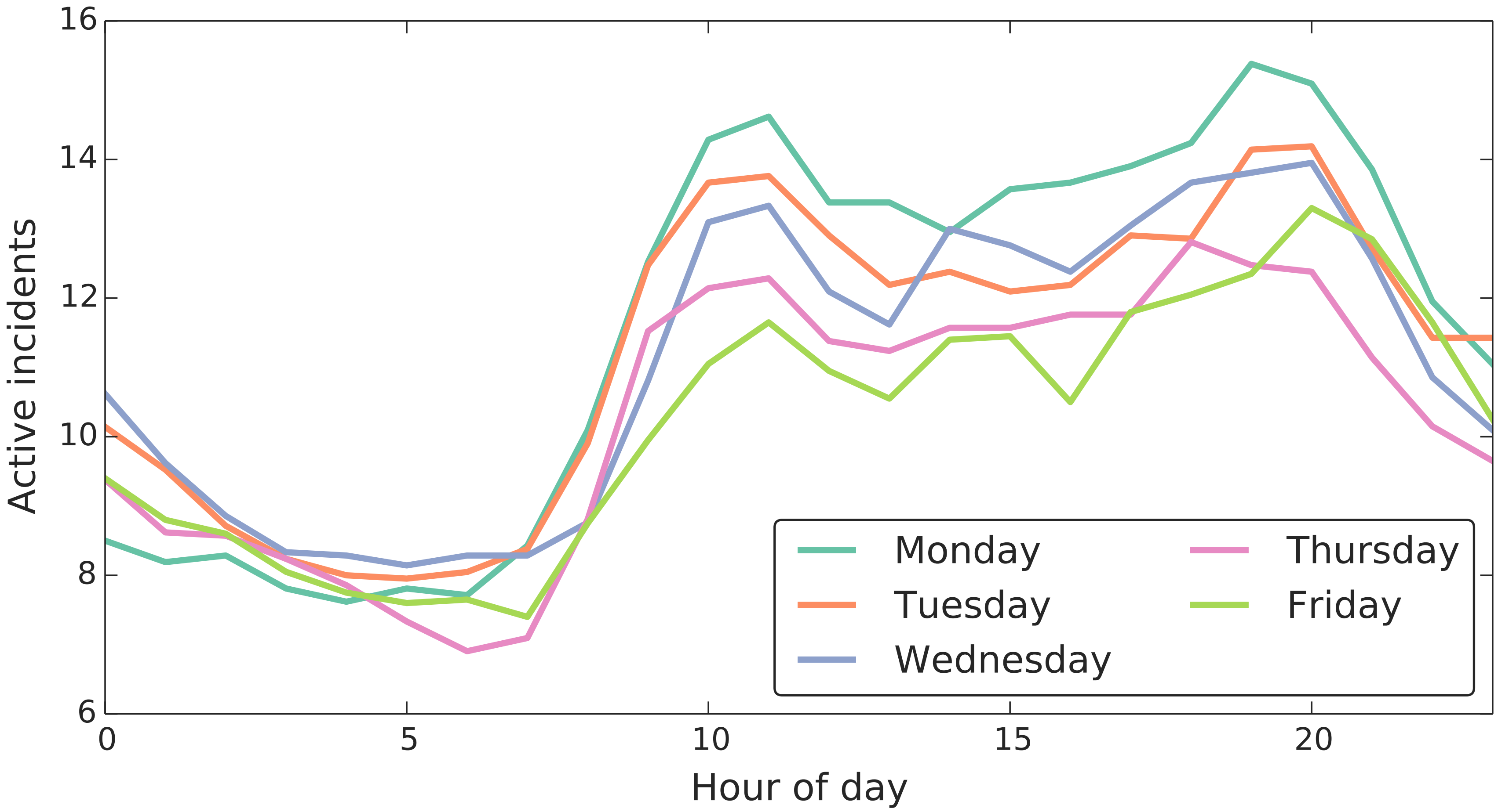
@ianhuston, @gopivotal

Pivotal



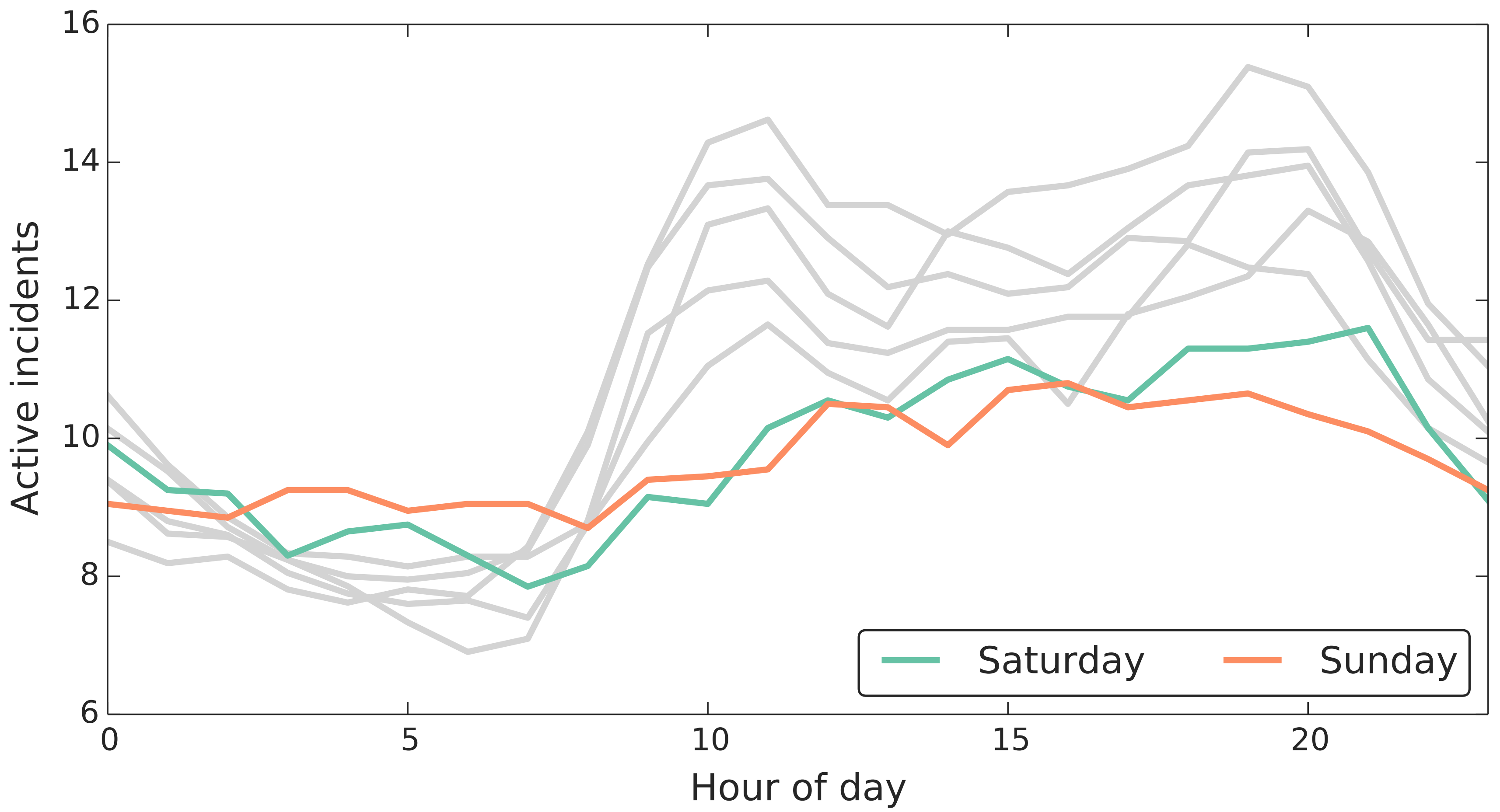
Major storm hits UK (Photo: BBC)
Active Incidents
0
10
20
30
40
50
Oct
Nov
Dec
Jan
2014
Feb
LL09 LTA
MARIANA MONTEIRO

Pivotal



@ianhuston, @gopivotal

Pivotal

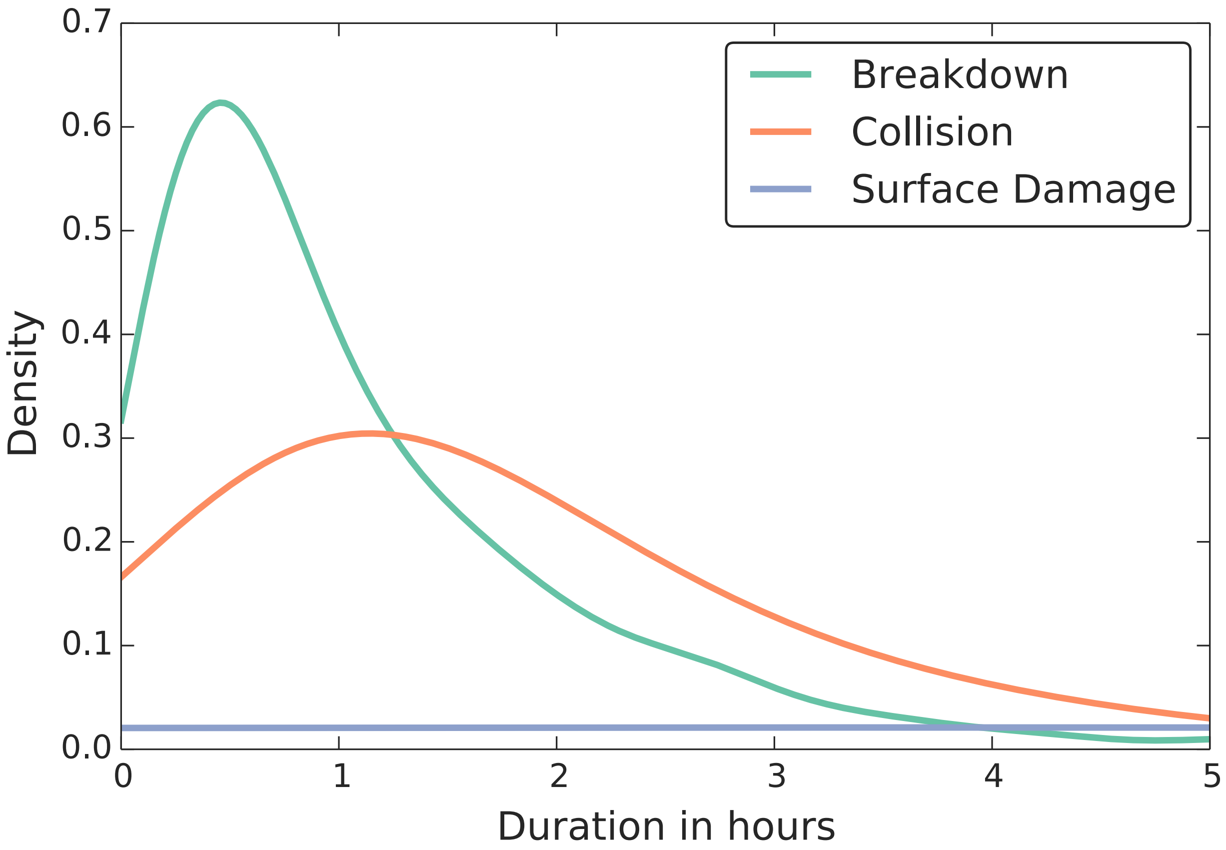


@ianhuston, @gopivotal

Pivotal

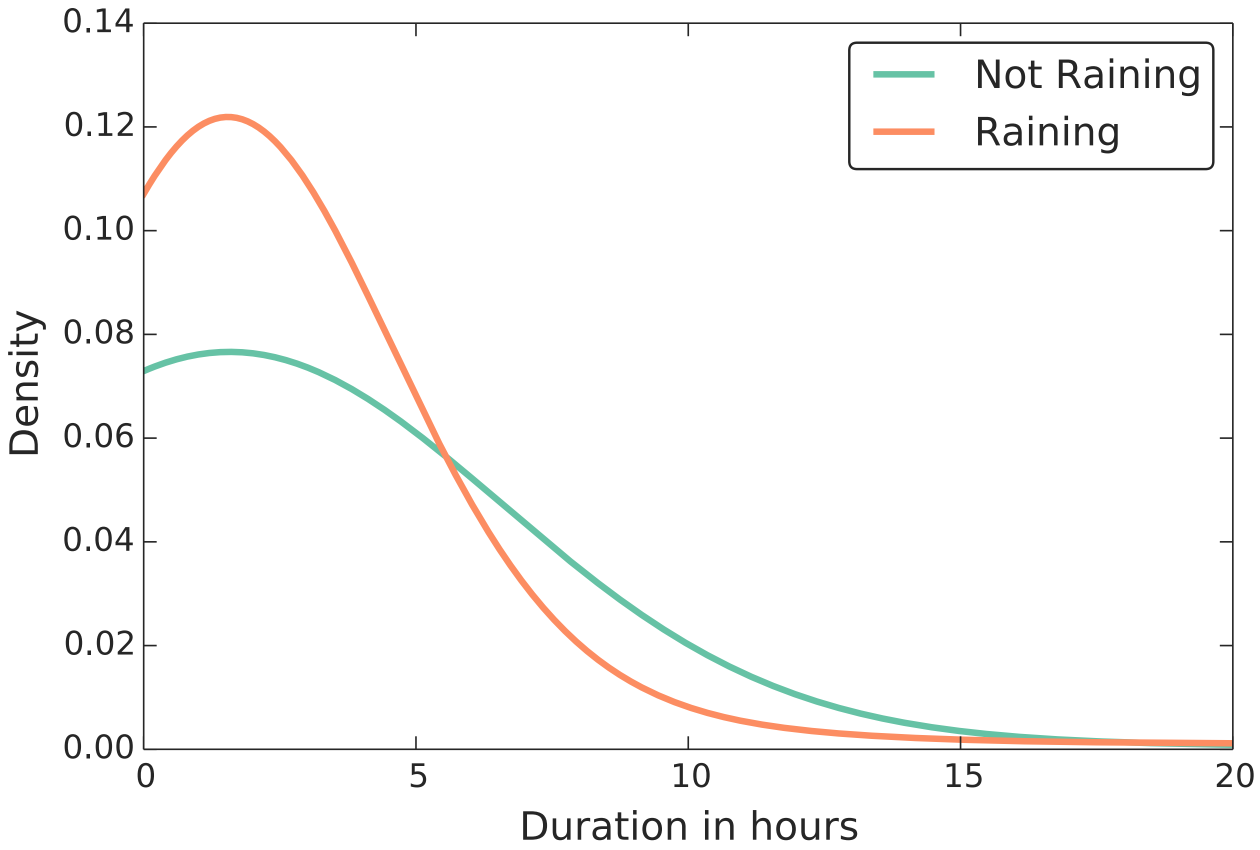
Durations are very different for different types of incident.

Mean duration for Surface Damage incidents is 107 hours!



Rain affects duration in a surprising way.

Incidents which start when it is raining finish **faster** than others.



Models

Linear Regression

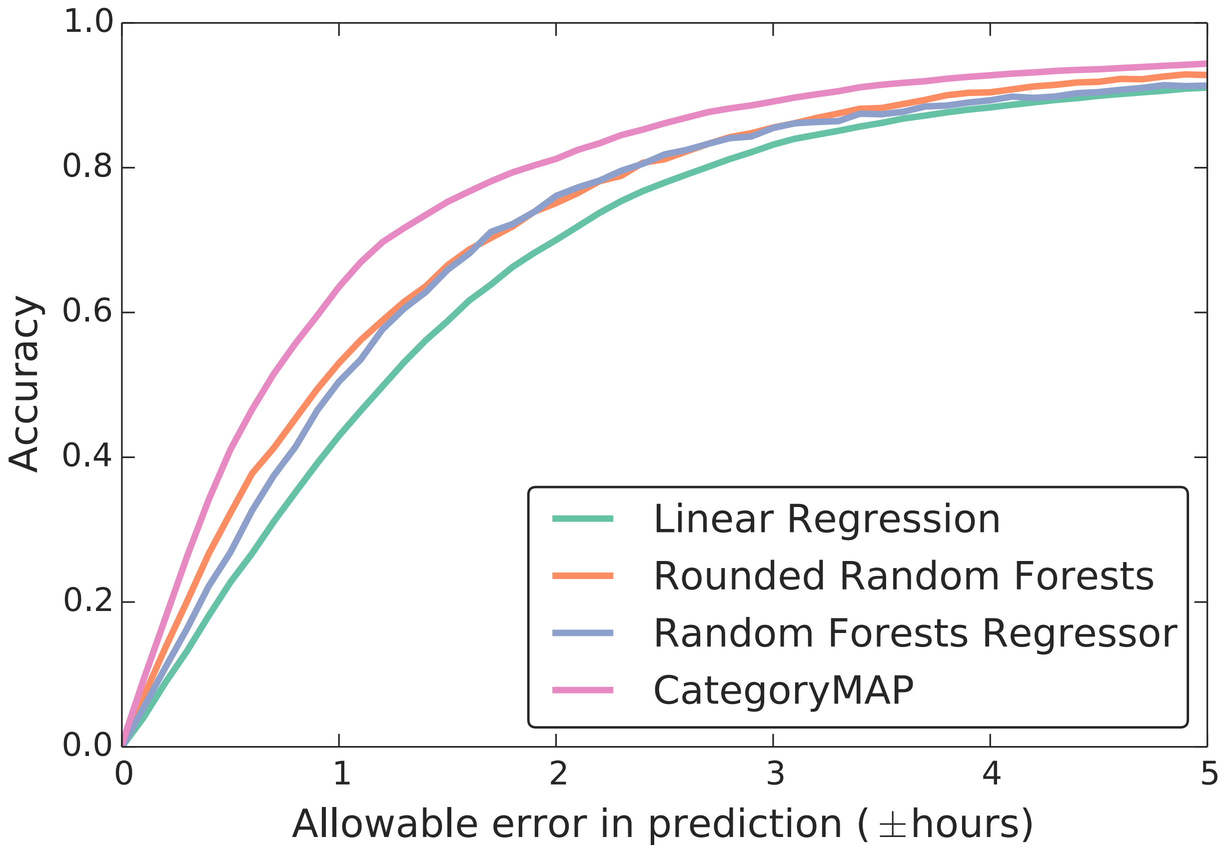
- Disruption reports & weather features

Random Forests

- Rounded categorical
- Regression

Category MAP

- Only use category of incident
- Maximum Likelihood estimate



Live Predictions

http://ds-demo-transport.cfapps.io

Using:

Pivotal™ Greenplum® Database

MADlib

Pivotal CF™

Active Incidents and Predicted Durations

Transport for London Data				Predictions	
Start Time	Location of Incident	# Streets Affected	Type of Incident	Total Duration	Time Remaining
Thursday, 6 Feb 14 20:40 Greenwich Mean Time	[A10] Turkey Street (EN1) (Enfield)	2	Flooding	**4.2 hours**	**2.7 hours**
Thursday, 6 Feb 14 21:00 Greenwich Mean Time	[A40] Western Avenue (HA4 ,UB5) (Hillingdon)	2	Flooding	**4.2 hours**	**3 hours**
Thursday, 6 Feb 14 20:44 Greenwich Mean Time	[A406] North Circular Road (E12) (Newham)	1	Fire	**2.2 hours**	**Up to 1 hour**
Thursday, 6 Feb 14 21:12 Greenwich Mean Time	[A205] Dulwich Common (SE22) (Southwark)	3	Collision	**1.1 hours**	**Up to 30 minutes**

Using 231209 reports about 9036 incidents since September 2013. Latest Update: Thursday, 6 Feb 14 21:50 Greenwich Mean Time

Summary

- Making use of a vibrant ecosystem of traffic data
- Innovative approaches needed to generate value from abundant and complex sources
- Connecting predictive models to traffic in the physical world is the future of smart cities

Pivotal™

Thank You!

Check out more of our Data Science use-cases at
www.goPivotal.com

Pivotal™

Pivotal

A NEW PLATFORM FOR A NEW ERA