Data Science in Action

Practical use cases that demonstrate how businesses generate value from data

21 March 2014 – SDS2014, Winterthur

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Introduction to Pivotal Data Labs

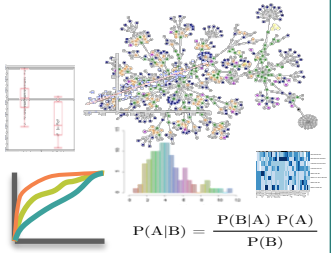
Our Team

- High caliber global team of machine learning experts from a wide variety of quantitative backgrounds
- Equally capable in coding & statistics

Our Tools

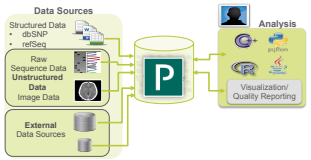
- Leading edge tools to implement machine learning collaboratively
- Have open-sourced several of our own tools for wide-spread use

Our Methods



- Parallelized a wide variety of machine learning algorithms for optimum performance on the Pivotal platform
- Agile, test-driven, customer focused

Our Process



- Analytical workflow aligned with business needs and optimized for speed
- Supports iterative and collaborative working

Our Experience

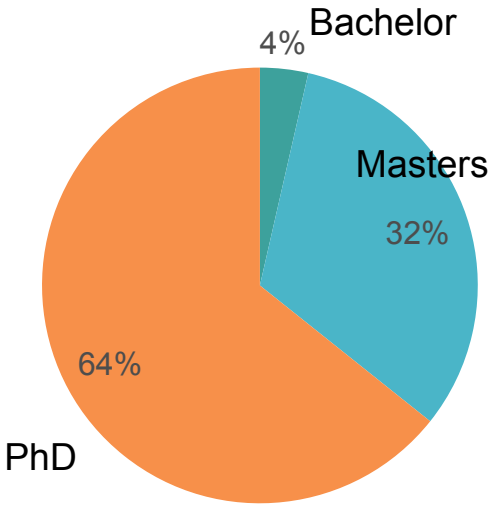
- More than 100 customer assignments carried out in the past 18 months
- Ensures quality and best practice in all our assignments

Pivotal

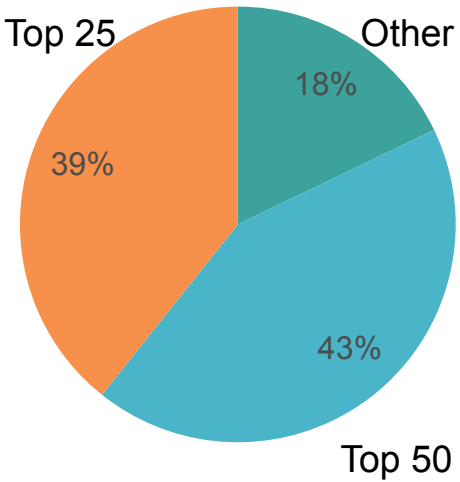
© Copyright 2014 Pivotal. All rights reserved.

Pivotal Data Science Team

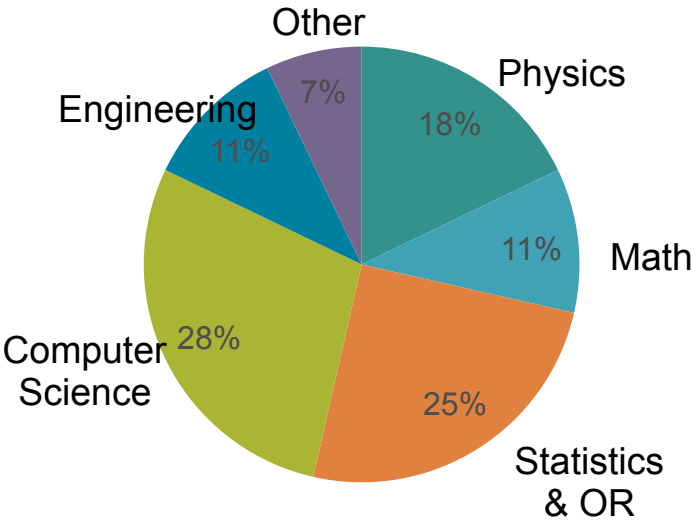
By Degree



By University



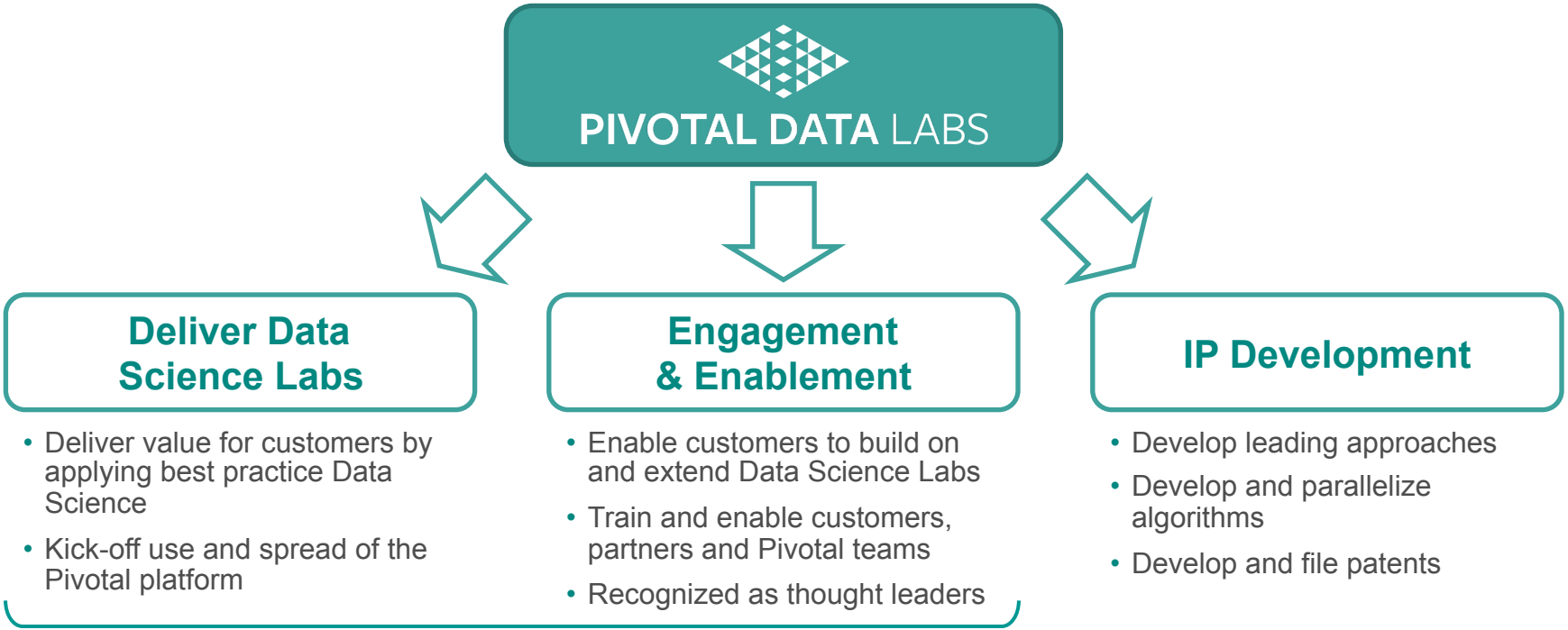
By Subject



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

What does the Pivotal Data Science team do?



*Data Scientists as **catalysts** for data-driven transformation*

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

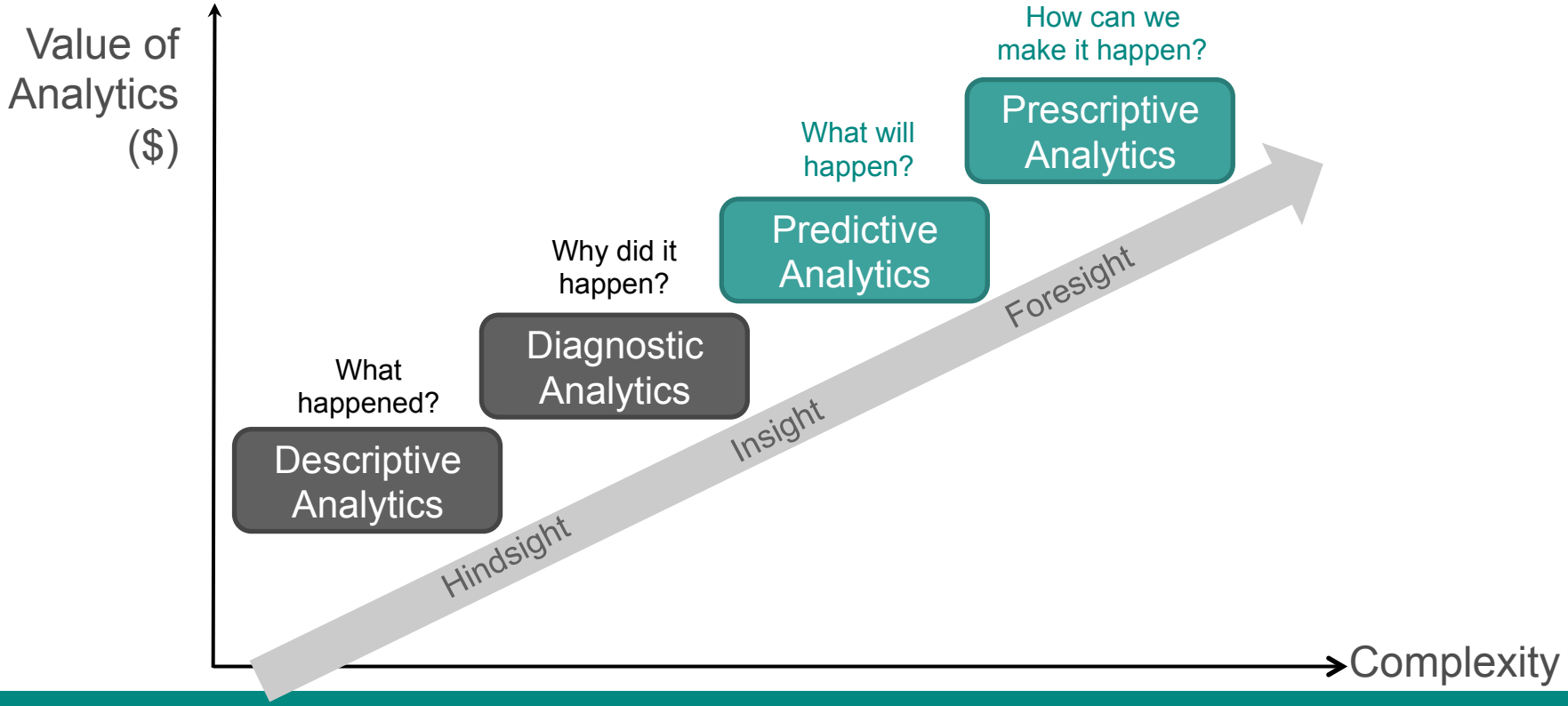
Data Science in Action

- **The Value of Data Science**
- The Practice of Data Science
- In-depth Use Case: Traffic Prediction
- Use Case Overviews
- Q&A

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

What do we mean with Data Science?



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Big Data & Data Science

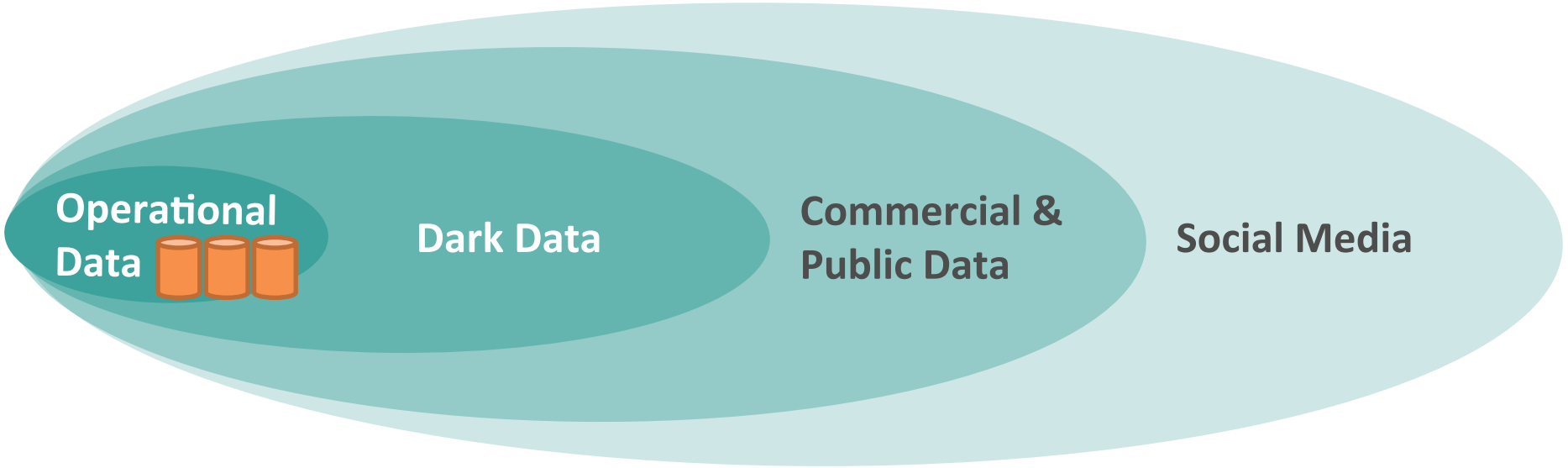
Decision = Data + Rules

"Big Data"

Data Science

© Copyright 2014 Pivotal. All rights reserved.

“Big Data”



© Copyright 2014 Pivotal. All rights reserved.

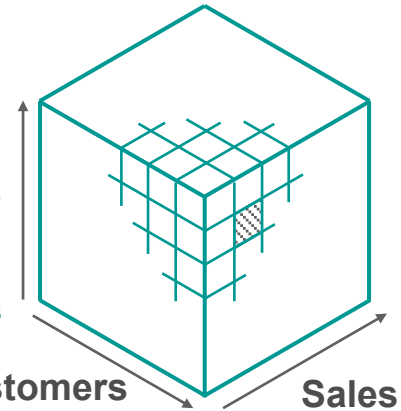
Combining data sources: Example

IPSQ (Quality) Owner: TS Production team Test flags from production line 1 year ~300GB	**APDM** Owner: TS Production team Full vehicle history including IPST (technical), IPSL (logistics), IPSQ test flags and all test results. 30 years ~TBs
FASTA Owner: Aftersales Dealership electronic tests Identifies early issues with cars >25TB	**IQS: Initial Quality Survey from JD Power** Owner: R&D Survey responses from new owners after 90 days for approx 1700 vehicles Few thousand lines ~MB
Social Data Owner: R&D Pulling 500MB per day from Twitter	**TQP** Owner: Supplier management PDFs of parts spec sheets ~ 500GB

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Generating value from data: Car configurator example

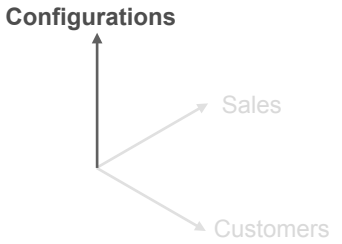


Configurations

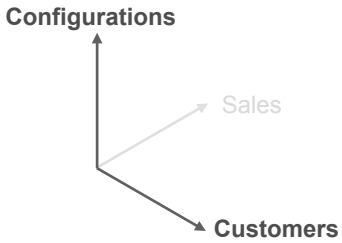
All car elements:
- Attribute frequencies (colors etc)
- Attribute combination frequencies

For instance:
- Browsing history
- Usage patterns
- Demographic insights

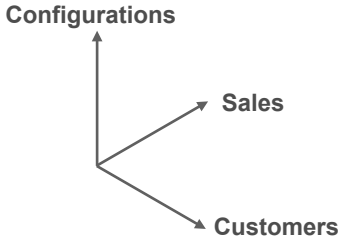
Ideally:
- Volumes
- Pricing
- By market
- Linkable to configurations



Basic Recommendation Engine



Advanced Recommendation Engine



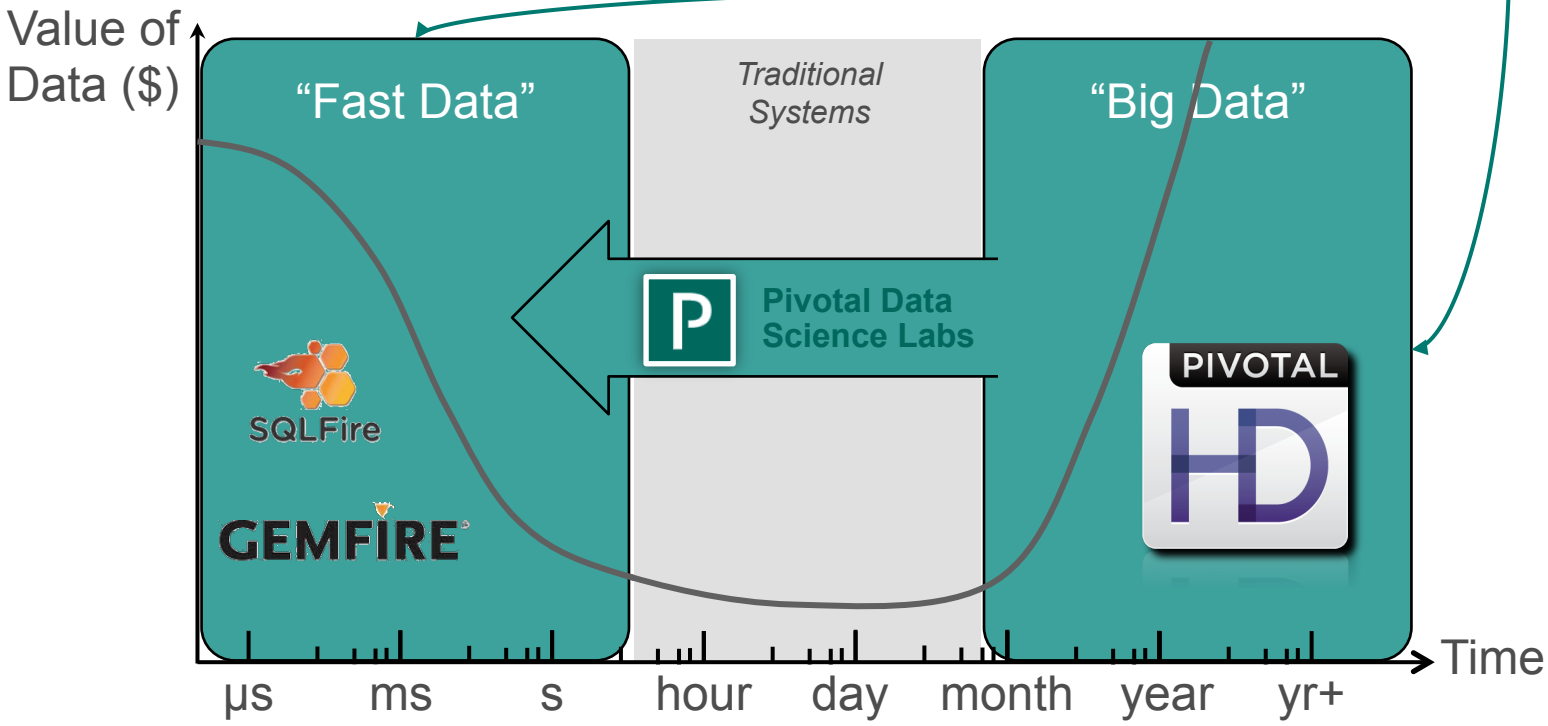
Yield Optimization Engine

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

The value of data over time

Pivotal



© Copyright 2014 Pivotal. All rights reserved.

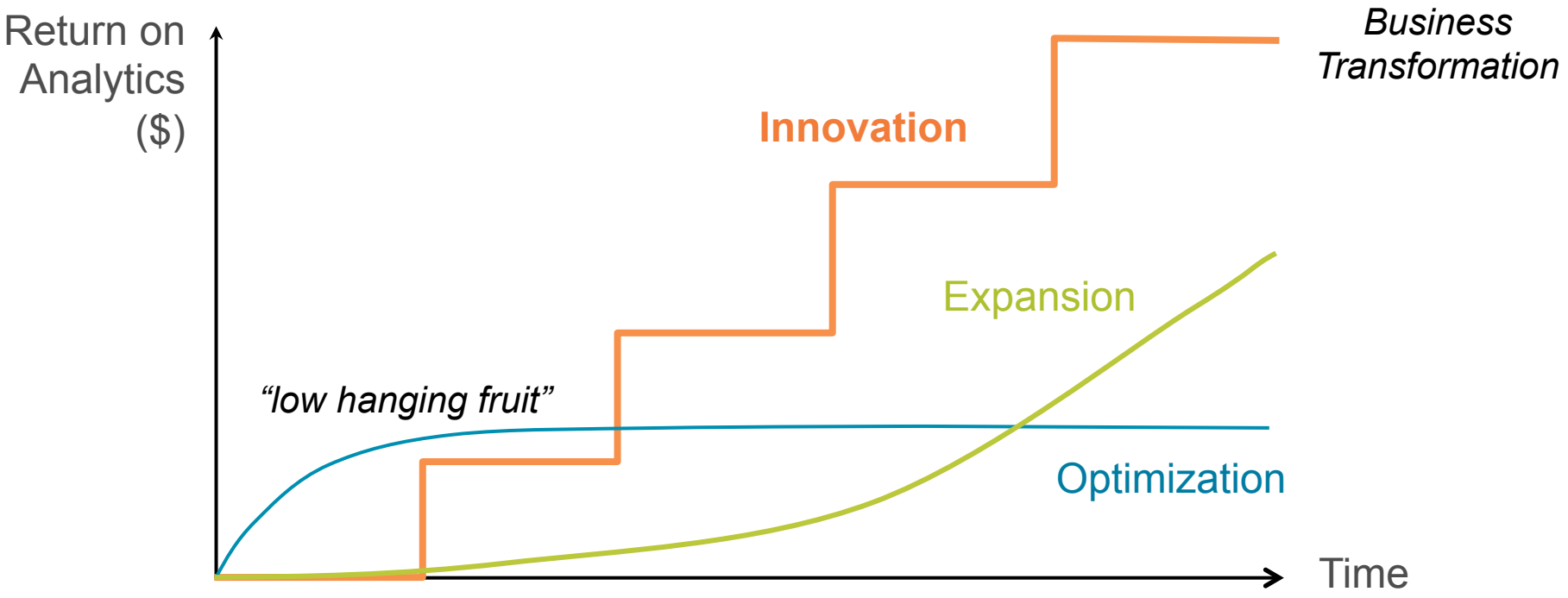
Data Science

- The use of statistical and **machine learning** techniques on **big multi-structured data** in a distributed computing environment to identify **correlations** and causal relationships, classify and predict events, identify patterns and anomalies, and infer probabilities, interest, and sentiment.
- In order to **drive automated low latency actions in response to events of interest**

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Why do Data Science?



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

What transforms businesses today?

- Digitization
- Internet of Things
- Pervasive Computing
- Pervasive Connectivity

Pivotal

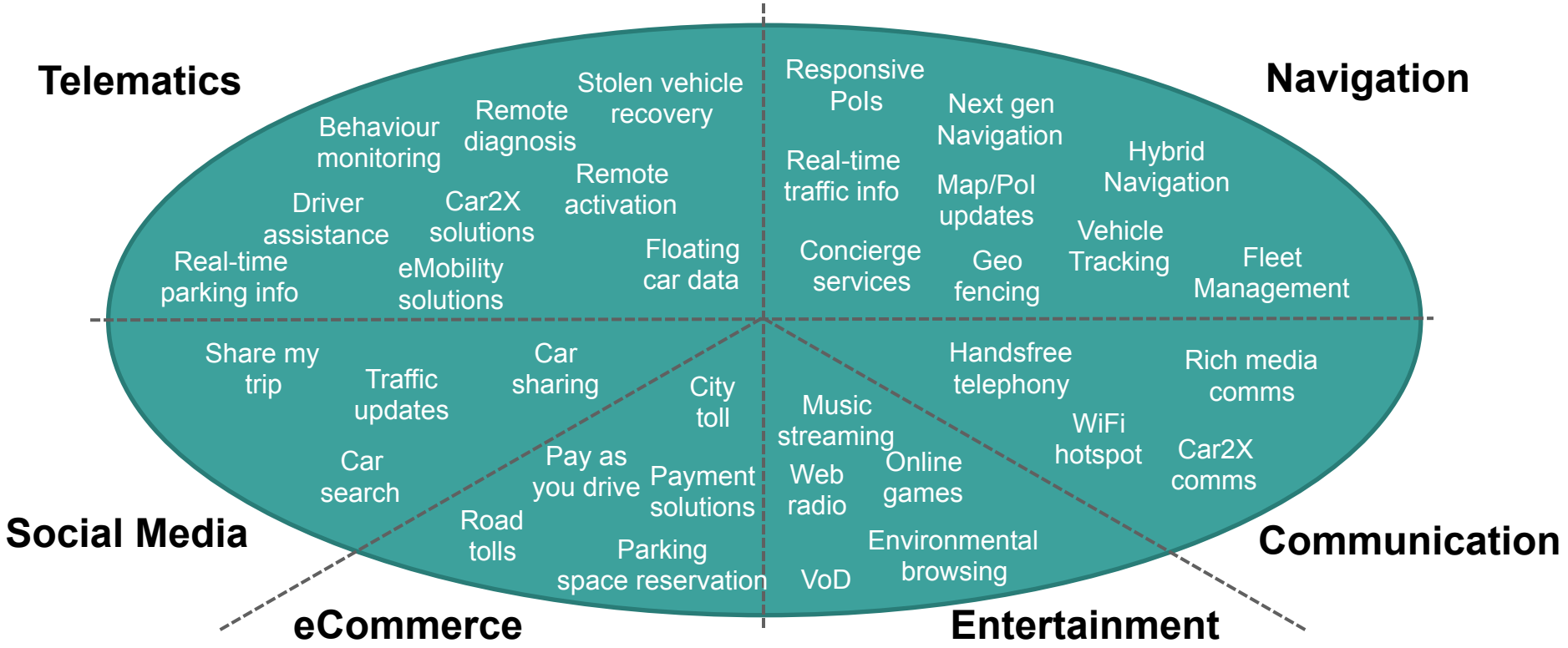
© Copyright 2014 Pivotal. All rights reserved.

Example: major paradigm shifts in automotive

Genesis	Mass Production	Modern Manufacturing	Platform Strategy	What's Next?
1885	*1908*	*1950s*	*1980s*	*2020*
				
Not a horse	**Mass availability** *"You can have any color of car, provided it's black"*	**Brand proliferation** *You can have any color*	**Globalization** *You can have anything anywhere*	**Connected, autonomous vehicles**

© Copyright 2014 Pivotal. All rights reserved.

The Connected Car drives innovation



© Copyright 2014 Pivotal. All rights reserved.

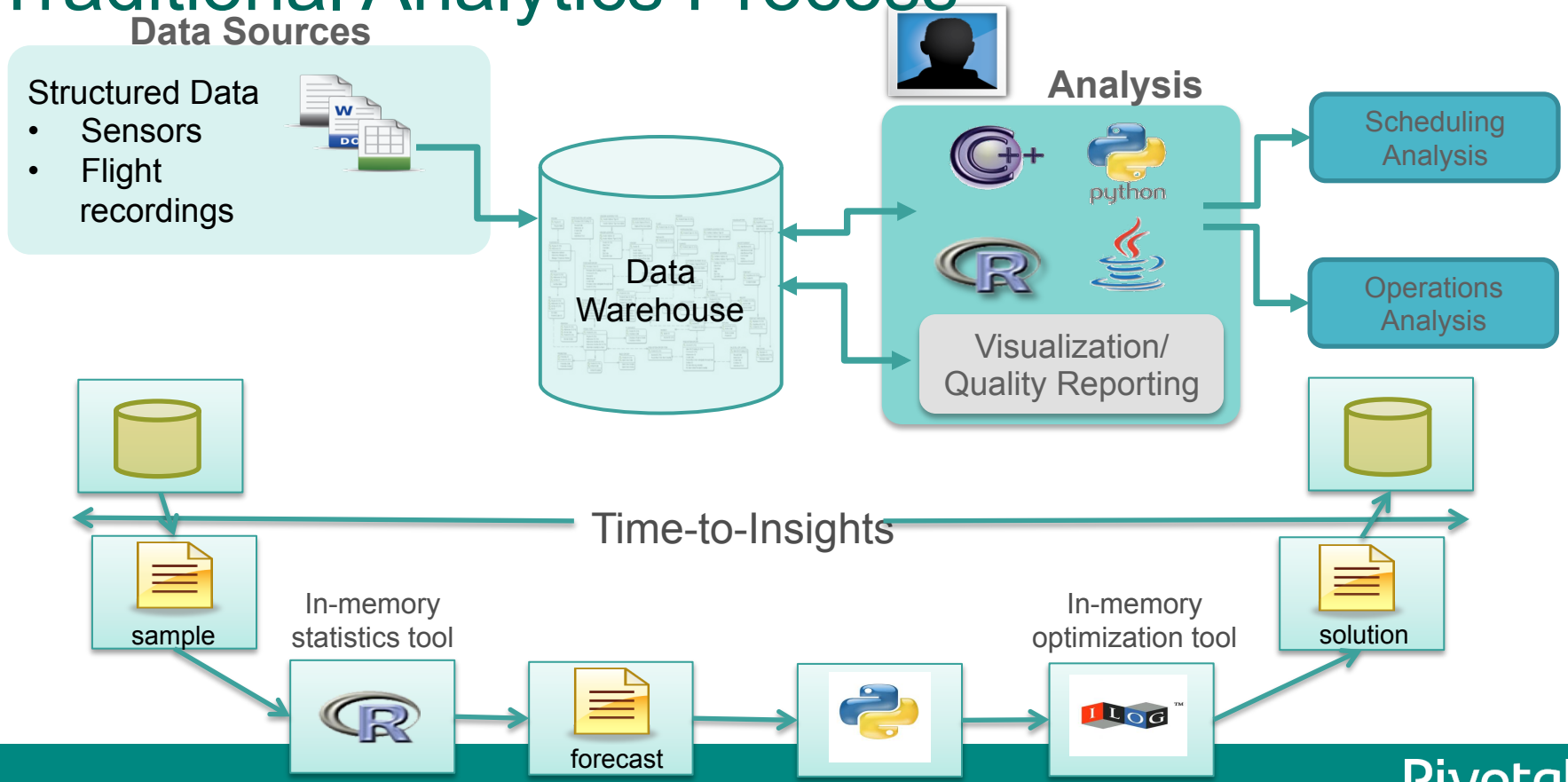
Data Science in Action

- The Value of Data Science
- **The Practice of Data Science**
- In-depth Use Case: Traffic Prediction
- Use Case Overviews
- Q&A

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

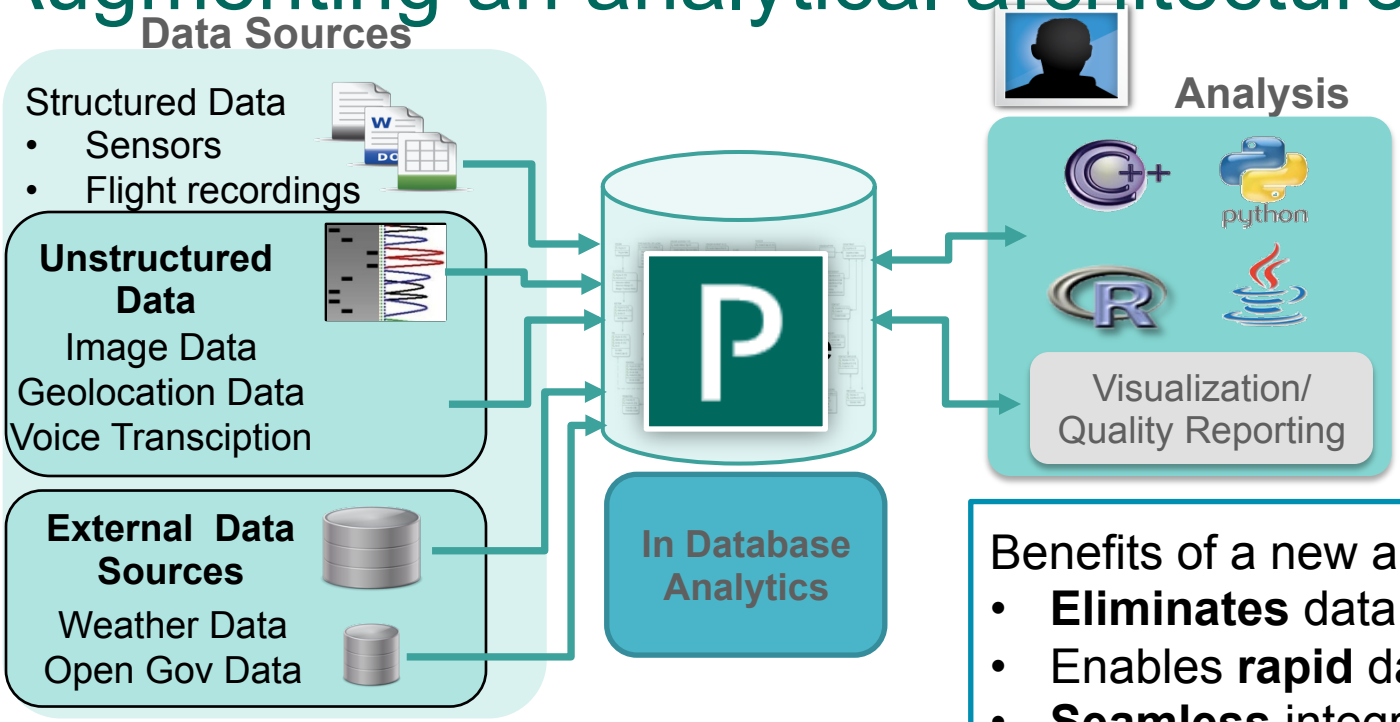
Traditional Analytics Process



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Augmenting an analytical architecture



Benefits of a new architecture:

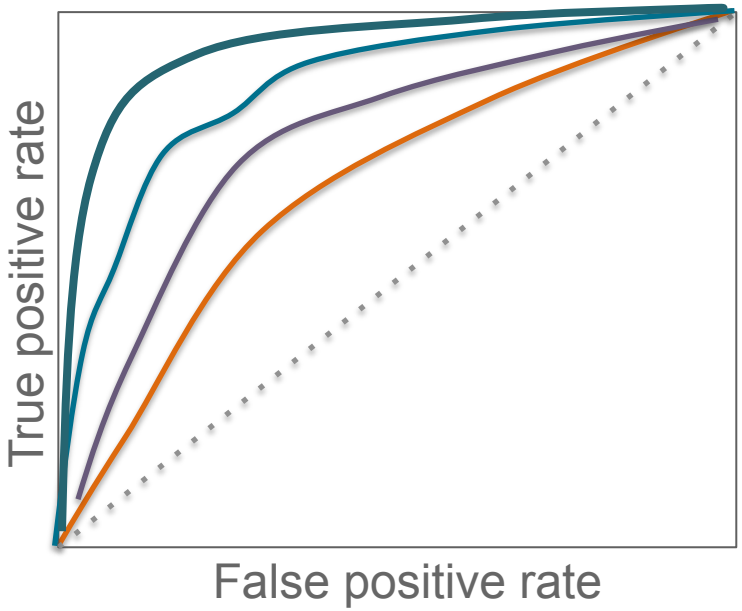
- **Eliminates** data movement
- Enables **rapid** data re-processing
- **Seamless** integration of additional external resources into analyses

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Machine Learning and Big Data

Getting the whole picture improves predictive power



- More data from different sources
- Provides a more complete view
- Improves statistics and inference

Sensor Data

Sensor Data + Flight Records

Sensor Data + Flight Records + Geoloc & Weather

Sensor Data + Flight Records + Geoloc & Weather + Imaging

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

The main types of use cases in practice

Data Mining

- Categorize types (segmentation)
- Categorize behaviors/usage
- Identify co-occurrences and associations
- Identify anomalies
- Identify attitudes
- Resolve entities

Predicting Behavior

- Predict churn likelihood
- Predict cross/up sales potential
- Predict fraud/waste/abuse likelihood
- Predict performance
- Predict reliability/quality
- Make a “recommendation”

Optimization

- Optimize processes
- Optimize process parameters
- Optimize asset allocation

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

PIVOTAL DATA SCIENCE TOOLKIT

1 Find Data

Platforms
- Greenplum DB
- Pivotal HD
- Hadoop (other)
- SAS HPA
- AWS

2 Write Code

Editing Tools
- Vi/Vim
- Emacs
- Smultron
- TextWrangler
- Eclipse
- Notepad++
- IPython
- Sublime

Languages
- SQL
- Bash scripting
- C
- C++
- C#
- Java
- Python
- R

3 Run Code

Interfaces
- pgAdminIII
- psql
- psycopg2
- Terminal
- Cygwin
- Putty
- Winscp

4 Write Code for Big Data

In-Database
- SQL
- PL/Python
- PL/Java
- PL/R
- PL/pgSQL

Hadoop
- Pig
- Hive
- Java
- HAWQ

5 Implement Algorithms

Libraries
- MADlib

Java
- Mahout

R
- (Too many to list!)

Text
- OpenNLP
- NLTK
- GPText

C++
- opencv

Python
- numpy
- scipy
- scikit-learn
- Pandas

Programs
- Alpine Miner
- Rstudio
- MATLAB
- SAS
- Stata

6 Show Results

Visualization
- python-matplotlib
- python-networkx
- D3.js
- Tableau
- GraphViz
- Gephi
- R (ggplot2, lattice, shiny)
- Excel

7 Collaborate

Sharing Tools
- Chorus
- Confluence
- Socialcast
- Github
- Google Drive & Hangouts

A large and varied tool box!

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

As Data Scientists, what do we want?

Infrastructure Independent	Fast & Scalable	Schema Free	Real Time	Easy to Use
• Open source • PaaS	• In-database analytics • MPP	• Hadoop	• In-memory data grids embedded into the platform	• SQL, not Java • Faster than Hive

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Data Science in Action

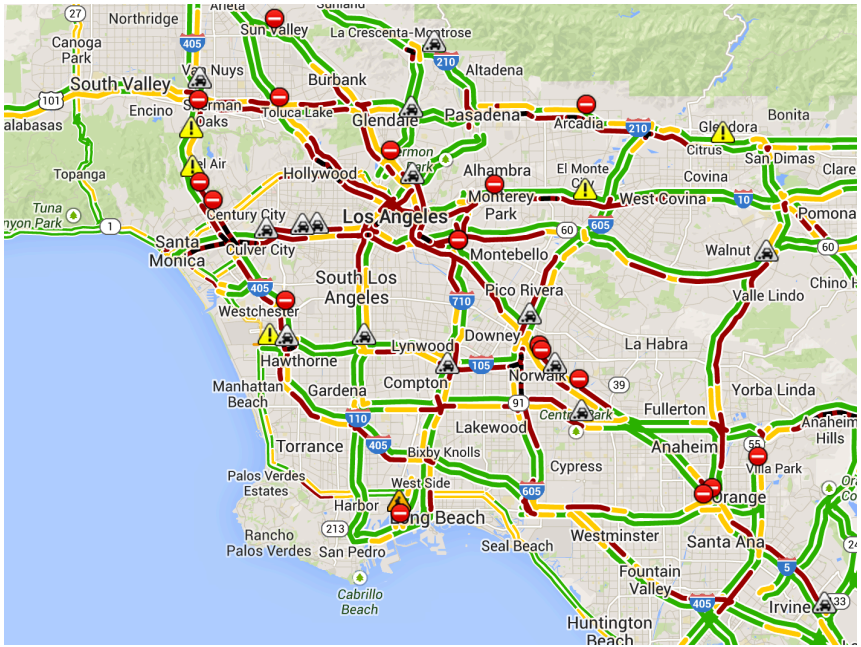
- The Value of Data Science
- The Practice of Data Science
- **In-depth Use Case: Traffic Prediction**
- Use Case Overviews
- Q&A

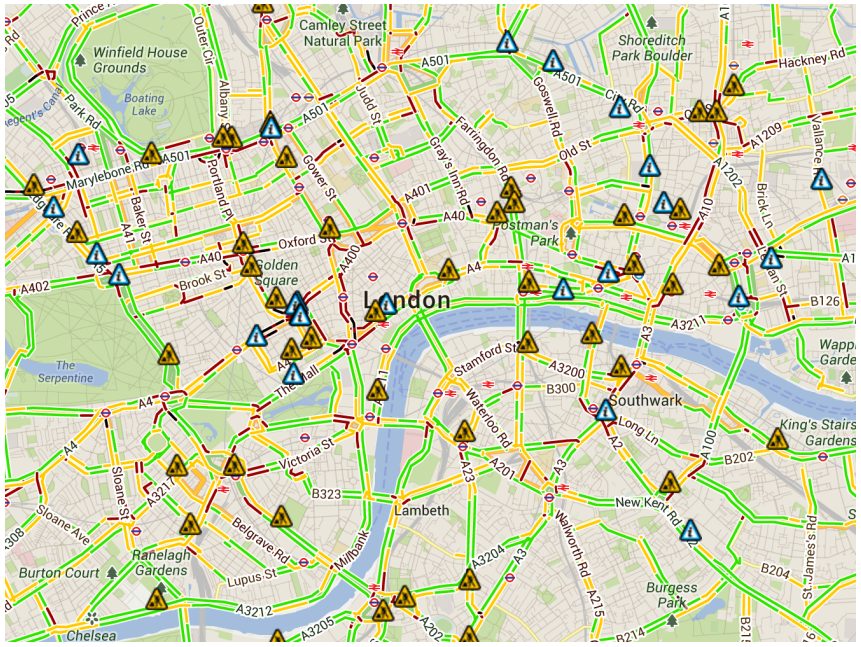
Pivotal

© Copyright 2014 Pivotal. All rights reserved.


What does traffic data look like?

...like this?





© Copyright 2014 Pivotal. All rights reserved.

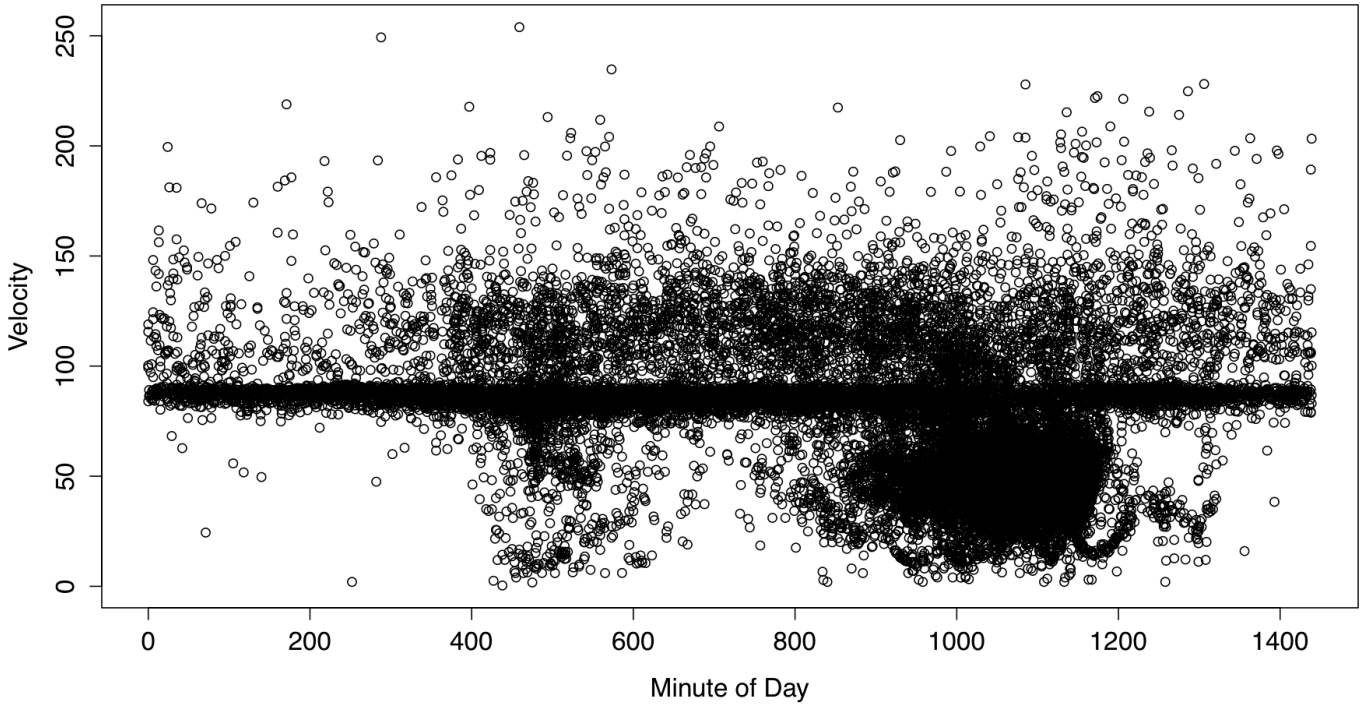
…or this?

(Note: This is the least offensive topic cluster in our Twitter data!)

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

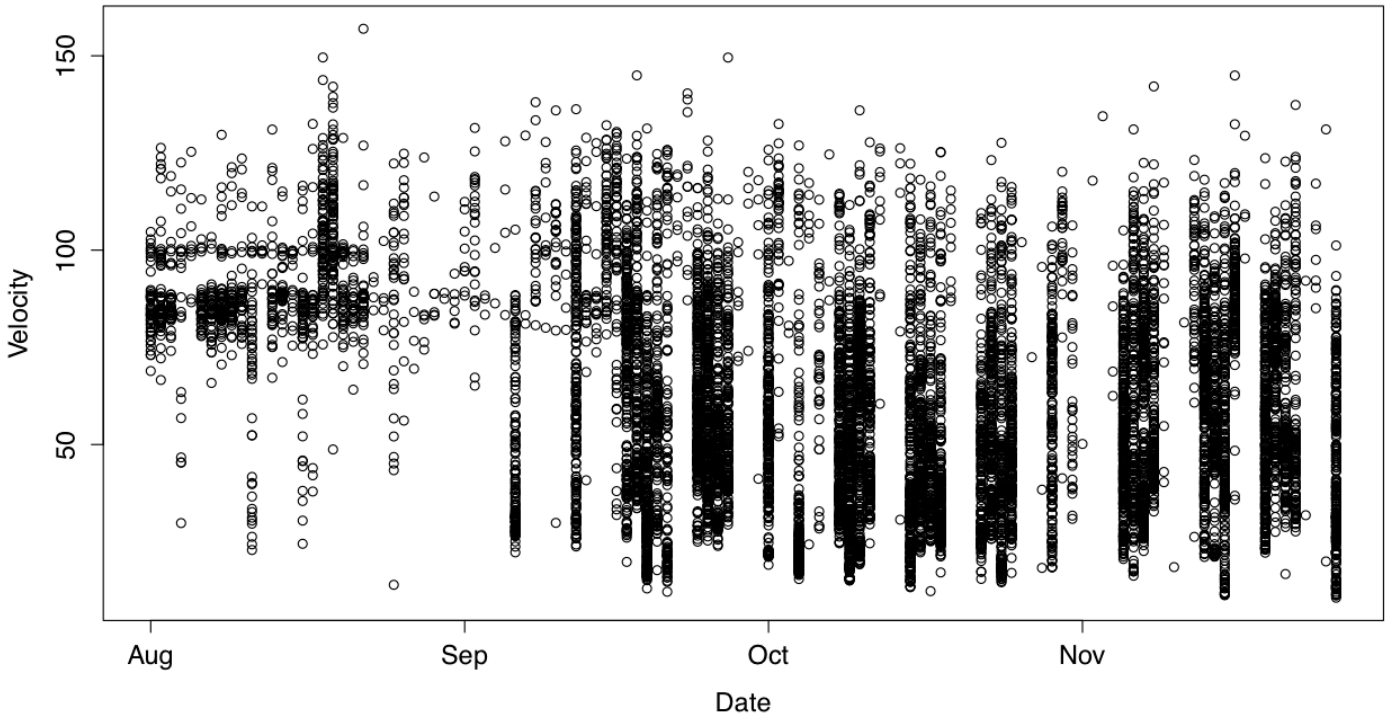
Velocity by Time of Day



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

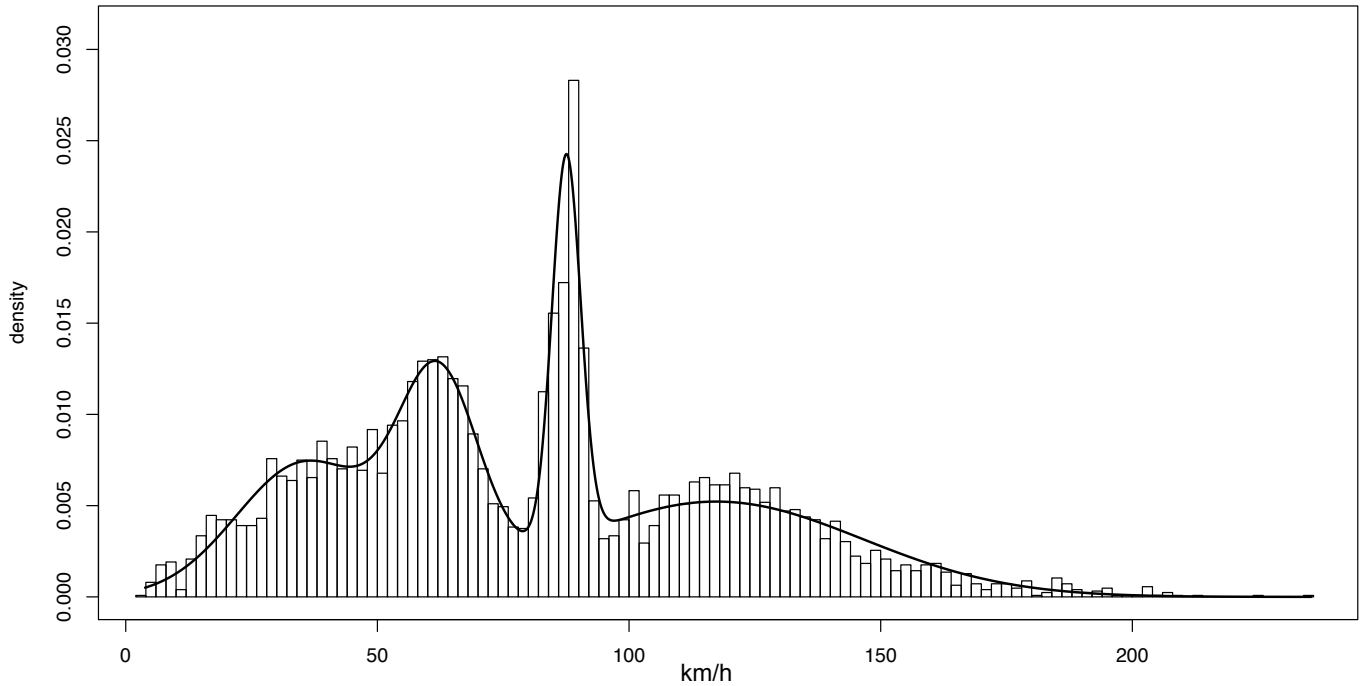
Distribution of Velocity over Time



Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Velocity Distribution

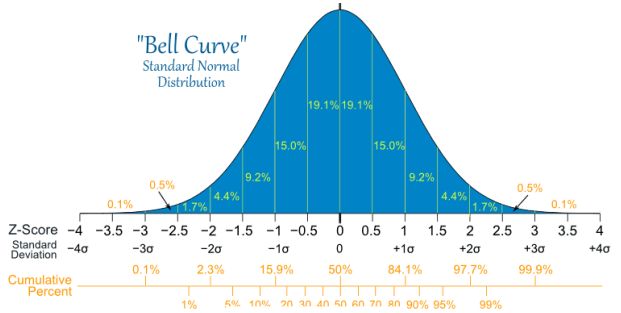


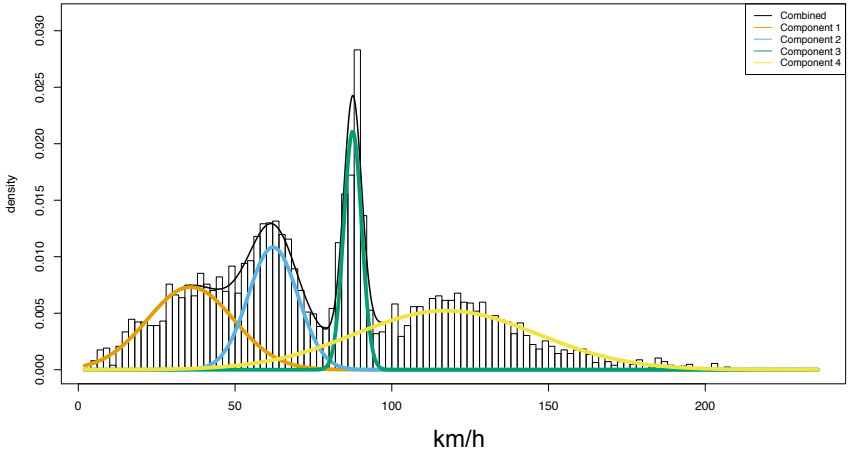
Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Find Velocity Groups

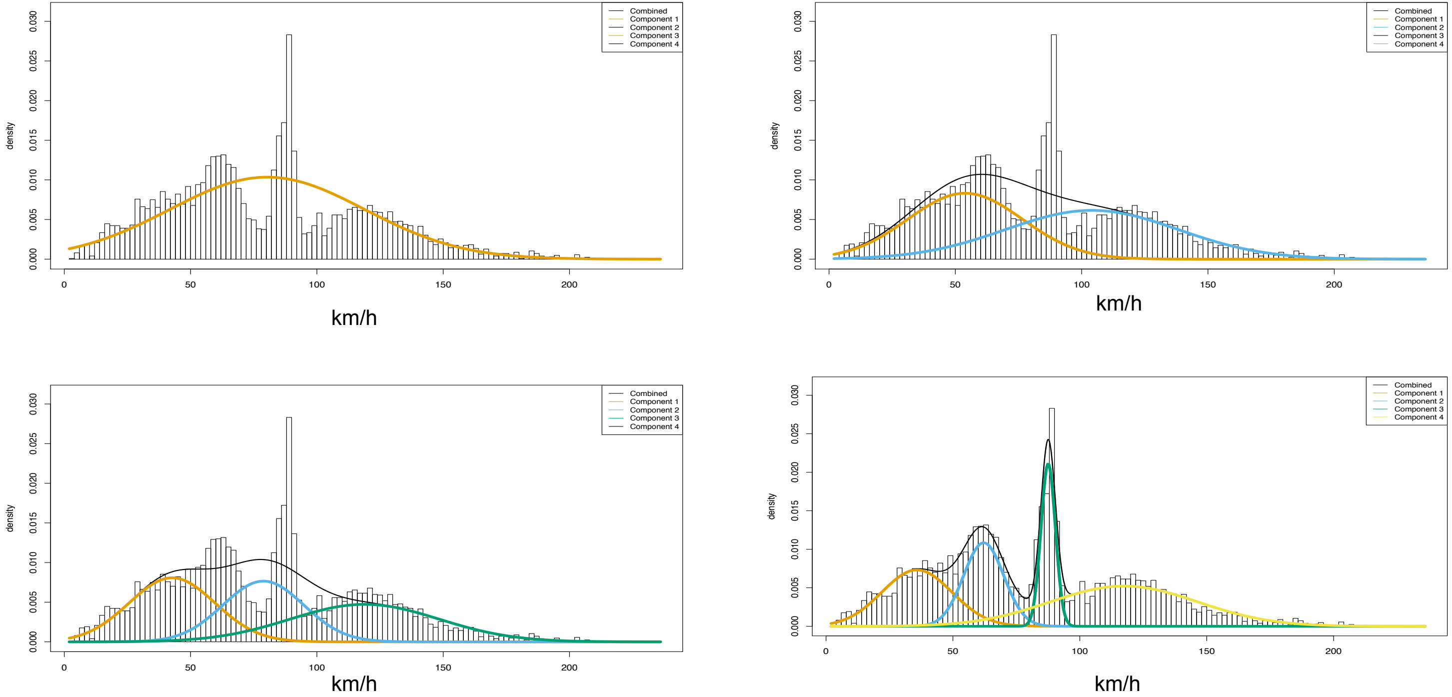
- Velocity distributions can be **fit well** with Gaussians
- An ‘overlay’ of multiple Gaussians is called **Gaussian Mixture Model**
- GMM fitting of the velocity distribution is done by **Expectation-Maximization** algorithm
- Shapes and positions of Gaussians **determine velocity groups**





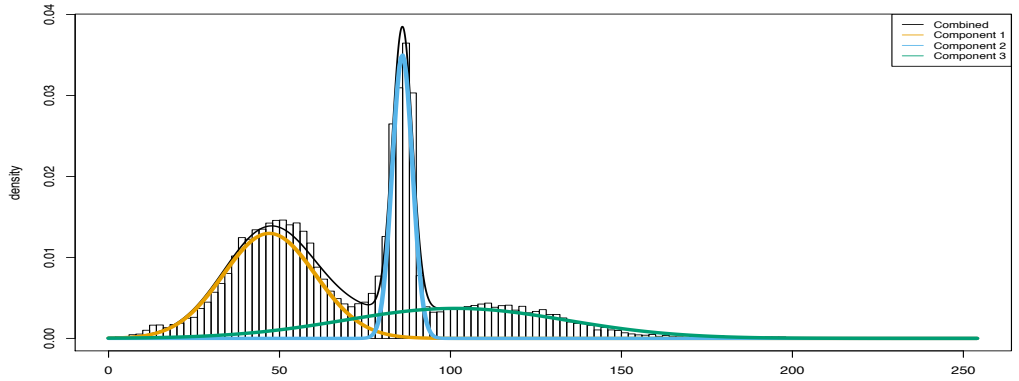
© Copyright 2014 Pivotal. All rights reserved.

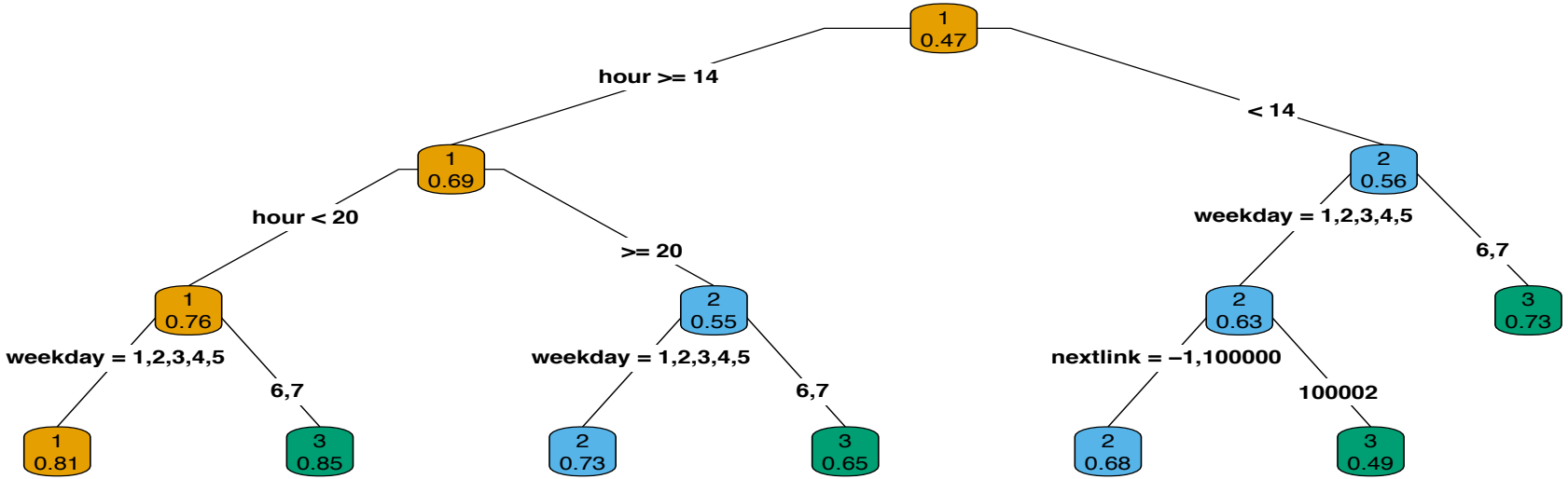
Gaussian Mixture Model



© Copyright 2014 Pivotal. All rights reserved.

Decision Trees
Example

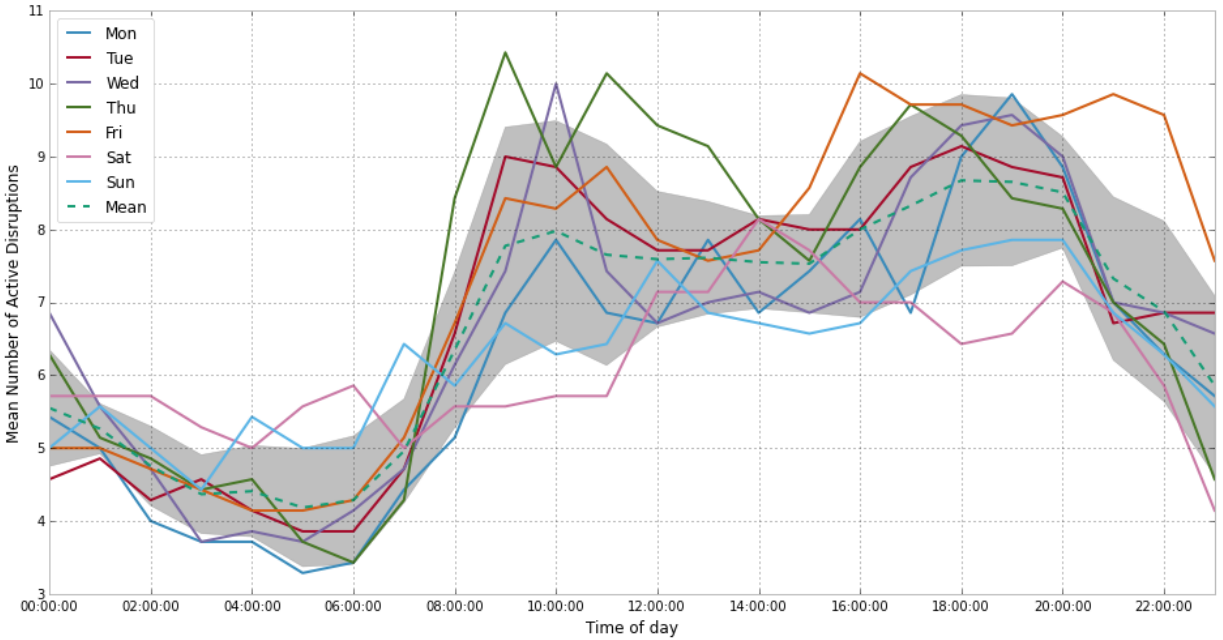




© Copyright 2014 Pivotal. All rights reserved.

Sneak Peek at our TfL Data Demo

- Used the freely accessible **TfL data** for a demo
- Shows # of **active disruptions** over different days **in London**



- ➢ Rush hour effects visible
- ➢ Nights are more quiet, but more disruptions on weekend nights

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

Data Science in Action

- The Value of Data Science
- The Practice of Data Science
- In-depth Use Case: Traffic Prediction
- **Use Case Overviews** – *not presented*
- Q&A

Pivotal

© Copyright 2014 Pivotal. All rights reserved.

BUILT FOR THE SPEED OF BUSINESS