Data Science in Action

Practical use cases that demonstrate how businesses generate value from data

21 March 2014 – SDS2014, Winterthur
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
Pivotal Data Science Team

By Degree
- PhD: 64%
- Masters: 32%
- Bachelor: 4%

By University
- Top 25: 39%
- Top 50: 43%
- Other: 18%

By Subject
- Statistics
- Computer Science: 28%
- Engineering: 18%
- Math: 11%
- Physics: 11%
- Other: 7%
- Engineering: 11%

© Copyright 2014 Pivotal. All rights reserved.
What does the Pivotal Data Science team do?

**Deliver Data Science Labs**
- Deliver value for customers by applying best practice Data Science
- Kick-off use and spread of the Pivotal platform

**Engagement & Enablement**
- Enable customers to build on and extend Data Science Labs
- Train and enable customers, partners and Pivotal teams
- Recognized as thought leaders

**IP Development**
- Develop leading approaches
- Develop and parallelize algorithms
- Develop and file patents

_Data Scientists as *catalysts* for data-driven transformation*_
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
What do we mean with Data Science?

- **Descriptive Analytics**
  - What happened?

- **Diagnostic Analytics**
  - Why did it happen?

- **Predictive Analytics**
  - What will happen?

- **Prescriptive Analytics**
  - How can we make it happen?

Value of Analytics ($)

Complexity
Big Data & Data Science

Decision = Data + Rules

“Big Data”

Data Science
“Big Data”

- Operational Data
- Dark Data
- Commercial & Public Data
- Social Media
## Combining data sources: Example

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Description</th>
</tr>
</thead>
</table>
| **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 |
Generating value from data: Car configurator example

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
The value of data over time

Value of Data ($)

“Fast Data”

“Big Data”

Traditional Systems

Time

Value of Data ($)

- µs
- ms
- s
- hour
- day
- month
- year
- yr+

“Fast Data”

“Big Data”

Pivotal Data Science Labs

SQLFire

GEMFIRE

© 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**
Why do Data Science?

Return on Analytics ($)

Time

Business Transformation

Innovation

Expansion

Optimization

“low hanging fruit”
What transforms businesses today?

- Digitization
- Internet of Things
- Pervasive Computing
- Pervasive Connectivity
Example: major paradigm shifts in automotive

<table>
<thead>
<tr>
<th>Genesis</th>
<th>Mass Production</th>
<th>Modern Manufacturing</th>
<th>Platform Strategy</th>
<th>What’s Next?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1885</td>
<td>1908</td>
<td>1950s</td>
<td>1980s</td>
<td>2020</td>
</tr>
</tbody>
</table>

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
The Connected Car drives innovation

Telematics
- Real-time parking info
- Behaviour monitoring
- Driver assistance
- eMobility solutions
- Car2X solutions
- Remote diagnosis
- Remote activation
- Stolen vehicle recovery
- Floating car data

Navigation
- Responsive Polis
- Real-time traffic info
- Concierge services
- Next gen Navigation
- Map/Pol updates
- Geo fencing
- Vehicle Tracking
- Hybrid Navigation
- Fleet Management
- Responsive PoIs
- Real-time traffic info
- Concierge services
- Next gen Navigation
- Map/Pol updates
- Geo fencing
- Vehicle Tracking
- Hybrid Navigation
- Fleet Management

Social Media
- Share my trip
- Traffic updates
- Car sharing
- City toll
- Pay as you drive
- Parking solutions
- Road tolls
- Parking space reservation

eCommerce
- City toll
- Pay as you drive
- Parking solutions
- Road tolls
- Parking space reservation

Communication
- Handsfree telephony
- WiFi hotspot
- Car2X comms
- Rich media comms
- VoD
- Environmental browsing
- Web radio
- Online games
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
Traditional Analytics Process

Data Sources

- Structured Data
  - Sensors
  - Flight recordings

Analysis

- Scheduling Analysis
- Operations Analysis

Data Warehouse

Visualization/Quality Reporting

Time-to-Insights

- In-memory statistics tool
- Sample
- Forecast
- In-memory optimization tool
- Solution
Augmenting an analytical architecture

Data Sources
- Structured Data
  - Sensors
  - Flight recordings
- Unstructured Data
  - Image Data
  - Geolocation Data
  - Voice Transcription
- External Data Sources
  - Weather Data
  - Open Gov Data

Analysis

Visualization/Quality Reporting

Benefits of a new architecture:
- **Eliminates** data movement
- Enables **rapid** data re-processing
- **Seamless** integration of additional external resources into analyses
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
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
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
     - Mahout
   - R
     - (Too many to list!)
   - Text
     - OpenNLP
     - NLTK
     - GPText
   - C++
     - opencv
   - Python
     - numpy
     - scipy
     - scikit-learn
     - Pandas
   - Programs
     - Alpine Miner
     - Rstudio
     - MATALAB
     - SAS
     - Stata

6. Show Results
   - Visualization
     - python-matplotlib
     - python-networkx
     - D3.js
     - Tableau
   - A large and varied tool box!

7. Collaborate
   - Sharing Tools
     - Chorus
     - Confluence
     - Socialcast
     - Github
     - Google Drive & Hangouts
As Data Scientists, what do we want?

<table>
<thead>
<tr>
<th>Infrastructure Independent</th>
<th>Fast &amp; Scalable</th>
<th>Schema Free</th>
<th>Real Time</th>
<th>Easy to Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Open source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• PaaS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• In-database analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• MPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Hadoop</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• In-memory data grids embedded into the platform</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• SQL, not Java</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Faster than Hive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
What does traffic data look like?
...like this?
…or this?

(Note: This is the least offensive topic cluster in our Twitter data!)
Velocity by Time of Day
Distribution of Velocity over Time
Velocity Distribution

Velocity Distribution

km/h

density

0 50 100 150 200

0.000 0.005 0.010 0.015 0.020 0.025 0.030

© 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
Gaussian Mixture Model
Decision Trees Example
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
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

BUILT FOR THE SPEED OF BUSINESS