Data Science in Industry

Ellie Dobson
Pivotal
Network intrusion detection
networks in telecommunications
What has industry ever done for us?
Theory-driven approach

Data-driven approach

‘start with the system and work towards the data’

‘start with the data and work towards the system’
Complexity

Value

Descriptive Analytics

What happened?

Diagnostic Analytics

Why did it happen?

Predictive Analytics

What will happen?

Reactive Analytics

How can we make it happen?

What happened?

Why did it happen?

What will happen?

How can we make it happen?

Reporting

Model training

Real-time recommendation

Data marshalling

Machine learning

Optimization

Model prediction

Complexity
The value of data over time

Big Data

make insights from a large historical dataset...

Fast Data

... and use them to make split-second decisions on real time data

drive automated low latency actions in response to events of interest
drive automated low latency actions in response to events of interest
predictive maintenance
payment fraud
formula 1 racing
Big Data Complex Data

Operational Data

Dark Data

Commercial & Public Data

Social Data
### 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
     - Spark

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
   - **Sharing Tools**
     - GraphViz
     - Gephi
     - R (ggplot2, lattice, shiny)
     - Excel

7. **Collaborate**
   - **Sharing Tools**
     - Chorus
     - Confluence
     - Socialcast
     - Github
     - Google Drive & Hangouts
fashion analytics
video analysis
call centre analysis
What has industry ever done for us?
Thanks for listening