Is a Data Scientist the New Quant?

Stuart Kozola
MathWorks
Data Science

- Facts or information used usually to calculate, analyze, or plan something
- Information that is produced or stored by a computer

Knowledge about or study of the natural world based upon facts learned through experiments and observation
- A particular area of scientific study (such as biology, physics, or chemistry)
- A subject that is formally studied in a college, university, etc.

Source: http://www.merriam-webster.com/dictionary/
Quant

- An expert at analyzing and managing quantitative data

Source: http://www.merriam-webster.com/dictionary/
What do Data Scientists do?

- Data Analysis
- Statistics
- Machine Learning
- Software Engineering
- Multivariable Calculus and Linear Algebra
- Big Data
- Data Munging
- Data Visualization and Communication

What do Quants do?

- Data Analysis
- Statistics
- Machine Learning
- Software Engineering
- Multivariable Stochastic Calculus, Linear Algebra, Mathematical Programming
- Big Data
- Data Munging
- Data Visualization and Communication
Is a Data Scientist the New Quant?

Source: http://indeed.com/trends
Data Science: 2 Major Trends

- Big Data
- Machine Learning
Big Data - Data Sources

- **Web**
  - Social Media
  - Web Forms
- **Businesses**
  - Transactions
  - Customers
- **Sensors**
  - Cameras
  - Accelerometers
  - GPS
  - Microphones
  - Weather stations

**File I/O**
- Text
- Spreadsheet
- XML
- CDF/HDF
- Image
- Audio
- Video
- Geospatial
- Web content

**Database Access**
- Financial Data
- ODBC
- JDBC
- HDFS (Hadoop)

**Hardware Access**
- Data acquisition
- Image capture
- GPU
- Lab instruments

**Communication Protocols**
- CAN (Controller Area Network)
- DDS (Data Distribution Service)
- OPC (OLE for Process Control)
- XCP (eXplicit Control Protocol)
Big Data

“Any collection of data sets so large and complex that it becomes difficult to process using … traditional data processing applications.”  

(Wikipedia)

- Traditional: local in-memory processing
- Laptop: 12 GB

\[ \text{---} \quad 8 \text{ GB} \]

\[ \gg a = \text{rand}(1E9,1); \quad \% \ 8 \text{ GB array} \]

- What about data processing?
Sensor Data
From 1 year of commercial flights in the US

20 terabytes of information per engine every hour
× twin-engine Boeing 737
× six-hour, cross-country flight from New York to Los Angeles
× # of commercial flights in the sky in the United States on any given day.

= 2,499,841,200 TB

2.5 ???
2.5 Zettabytes
Considerations for Big Data

- Data characteristics
  - Size, type and location of your data

- Compute platform
  - Single desktop machine or cluster

- Analysis Characteristics
  - Embarrassingly Parallel
  - Analyze sub-segments of data and aggregate results
  - Operate on entire dataset
New Big Data Capabilities in MATLAB

Memory and Data Access
- 64-bit processors
- Memory Mapped Variables
- Disk Variables
- Databases
- Datastores

Programming Constructs
- Streaming
- Block Processing
- Parallel-for loops
- GPU Arrays
- SPMD and Distributed Arrays
- MapReduce

Platforms
- Desktop (Multicore, GPU)
- Clusters
- Cloud Computing (MDCS on EC2)
- Hadoop
Big Data Challenges

- **Data Hygiene**
  - Data is dirty, it wasn’t collected with your use-case in mind

- **Data Munging**
  - Combining data from different sources
Data Hygiene

LET’S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH.

© marketoonist.com
Data Hygiene

Anomalies

Missing Data

Load (MW)

Data Munging

- Cleaning data that has errors, outliers, or duplicates

- Handling missing data
  - Discarding
  - Filtering
  - Imputation

- Merging and time-aligning data (might have different sample rates)

- Working with data in different domains
Domains

- Mathematical
- Time-series
- Signal
- Image & Video
- Acoustic
- Financial
- Geospatial
- Text
- Weather and Environmental
Machine Learning

Example: Human Activity Learning Using Mobile Phone Data

Machine learning uses **data** and produces a **program** to perform a **task**

**Task:** Human Activity Detection

**Standard Approach**

Hand Written Program

If $X_{acc} > 0.5$
then “SITTING”
If $Y_{acc} < 4$ and $Z_{acc} > 5$
then “STANDING”
...

**Formula or Equation**

$Y_{activity} = \beta_1 X_{acc} + \beta_2 Y_{acc} + \beta_3 Z_{acc} + ...$

**Machine Learning Approach**

**model:** Inputs $\rightarrow$ Outputs

$model = \langle Learning\ Algorithm\rangle(sensor\ _data,\ activity)$
Example: Human Activity Learning Using Mobile Phone Data

Objective: Train a classifier to classify human activity from sensor data

Data:

<table>
<thead>
<tr>
<th>Predictors</th>
<th>3-axial Accelerometer and Gyroscope data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>Activity: ![Activity Icons]</td>
</tr>
</tbody>
</table>

Approach:
- Extract features from raw sensor signals
- Train and compare classifiers
- Test results on new sensor data
Machine Learning is Everywhere

- Image Recognition
- Speech Recognition
- Stock Prediction
- Medical Diagnosis
- Data Analytics
- Robotics
- and more…
Overview – Machine Learning

Type of Learning

- Unsupervised Learning
  - Clustering
    - Group and interpret data based only on input data

- Supervised Learning
  - Classification
  - Regression
    - Develop predictive model based on both input and output data
Unsupervised Learning

Clustering

- k-Means, Fuzzy C-Means
- Hierarchical
- Neural Networks
- Gaussian Mixture
- Hidden Markov Model
Supervised Learning

Regression

- Neural Networks
- Decision Trees
- Ensemble Methods
- Non-linear Reg. (GLM, Logistic)
- Linear Regression

Classification

- Support Vector Machines
- Discriminant Analysis
- Naive Bayes
- Nearest Neighbor
Application: Retail / Supply Chain
Tesco: 2\textsuperscript{nd} largest retailer in the world

- How do promotions and weather affect food sales?
- Use historical data to develop a predictive model
- Validate model and incorporate into business systems
Off-highway vehicles

Automotive

Aeronautics

Clean Energy

Industrial Automation

Fleet Analytics

Process Analytics

Medical Devices

Health Monitoring

Condition Monitoring

Asset Analytics

Integrated Vehicle Health Management

Data

Oil & Gas
MathWorks Approach

- Don’t expect you to be an expert in everything
- Make it easy to analyze all types of data in any domain
- Integrated workflow
  - Access Data
  - Analysis
  - Deployment
- Flexible language for customizing
- One platform for multidisciplinary collaboration
Predictive Analytics Example
Energy Demand Forecasting

Forecast electricity demand for US power grids with live data from ISOs and weather stations using Neural Network models.

http://ec2-54-165-201-58.compute-1.amazonaws.com:8080/DemandForecastWeb/
Takeaways

- Quants – *original data scientists*?

- Data Science – Analytics across multiple domains
  - Big Data
  - Machine Learning

- Demand for data science skills is high

- Data is messy – getting the signal from the noise is often the biggest part
Moving Beyond the Hype to Practical Data Science

What you can learn to get ahead of the competition:

- **Stochastic Processes**
  - Learn from quantitative finance

- **From Static to Dynamic Systems**
  - Move beyond static relationships
  - System identification

- **Decision by Optimization**
  - Mathematical programming
  - Simulation based decision making