So You Want to be a Data Scientist

Stephen Coggeshall
April 21, 2016
Outline

- What, why big data
- What is data science
- What are the necessary skills
- Description of the interview process
- Description of the typical work for data scientists
WHAT IS BIG DATA

Big Data is a term that describes large volumes of high velocity, complex and variable data that require advance techniques and technologies to enable the capture, storage, distribution, management and analysis of the information.
Data growth exhibits “Moore’s Law” doubling every 1 to 2 years.
Proliferation of Sensors (IOT)

DATA IS COLLECTED EVERYWHERE
Data Science

• Why - exponential growth of data and compute power

• What is data science?
  – The scientific discipline around typically large and difficult data sets to solve complex business problems

• It is applied in many industries, including
  – Financial services
  – Healthcare
  – Insurance
  – Telecomunication
  – Social networks
  – Gaming
  – Government
Examples of Data Science Applications

Risk Management
• Scores for credit, fraud, bankruptcy, actuarial…

Marketing
• Product offers, pricing optimization, customer analysis…

Financial Engineering
• Econometrics, derivative pricing, algorithmic trading…

Social Networks
• Network analysis, recommender systems, sentiment analysis…

Operations Optimization
• Production lines, chemical processes, manufacturing…

Transportation
• Traffic flow, smart grid, route/price optimization, logistics…

Search
• Online searches, document matching, link analysis…

Health
• Bioinformatics, drug discovery, image analysis…

…
Business Analytics Hierarchy

<table>
<thead>
<tr>
<th>Strategic</th>
<th>What new business opportunities do I have?</th>
<th>New products &amp; services around existing data, new data to acquire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescriptive</td>
<td>What should I do with my current businesses?</td>
<td>Optimization of actions, recommendations</td>
</tr>
<tr>
<td>Predictive</td>
<td>What will happen?</td>
<td>Forecasts, predictive models, machine learning</td>
</tr>
<tr>
<td>Explanatory</td>
<td>Why are things happening?</td>
<td>BI/BA analysis, slicing data different ways, diagnostics</td>
</tr>
<tr>
<td>Descriptive</td>
<td>What’s going on?</td>
<td>BI dashboards, reporting</td>
</tr>
</tbody>
</table>

Question

How and what
Desired Data Science Background and Skills

• Strong math and statistics
• Strong programming & software best practices
• Passion for data
• Machine learning methods
• Big data technologies (scalable file systems and computing architectures)
• Scientific exploration, discovery and testing discipline
• Curiosity, skepticism, detail oriented
• Desire to learn
Typical Data Scientist Work

• Lots of data manipulation, cleaning, exploration, analysis, discovery. Frequently deal with messy data
• Problem definition:
  – Fully understand business problem and goals
  – Understand details of the underlying data
  – Frame the problem: time series? predictive model? simulation? structure discovery?
  – What data is available? What are the inputs/outputs?
• Build preliminary models, iterate with business
• Finalize, implement
• Monitor, rebuild as needed
How to Get Into Data Science as a Career

• Solid foundation in basics: math, statistics, programming
• Buy a few books. Possibles:
  – Elements of Statistical Learning, Hastie, Tibshirani, Friedman
  – Pattern Recognition and Machine Learning, Bishop
  – Pattern Classification, Duda, Hart, Stork
  – Foundations of Predictive Analytics, Wu, Coggeshall
• Many excellent courses on Coursera. Chose the technical ones.
• Internship if possible, but not necessary.
• Now you know the basics and can converse. Interview for beginner position.
Description of the Interview Process

• Resumes selected for interviews
• Typically two phone interviews
  – Discuss resume, questions about ML, data, statistics, programming, math...
• In-house interview
  – Usually starts with a presentation, typically dissertation work
  – Series of individual and group interviews, ~45 mins
  – Same Q’s as phone interview, but in more detail
• The group meets next day, go around room,
  – Technically adept, smart, experience, culture fit?
Summary: Data Science is a Surging Field!

• Sensors everywhere, data collection everywhere, compute power growing exponentially

• Machine learning algorithms continue to grow in power and complexity
  – Rise in deep learning, convolutional neural nets, distributed/scalable machine learning

• Businesses relying more and more on data science

• The world is awash in data. Need people to transform data into intelligence into actions.