Getting into Data Science

Ivan Teo
Data Scientist, Stitch Fix
25 April 2019
My Journey

Theoretical and Computational Biophysics Group (Klaus Schulten Lab)

2009 - 2016

INSIGHT

https://www.insight datascience.com/

2016

CapitalOne

2017 - 2018

STITCH FIX

2019
Life in data science

Key areas
- Machine learning
- Experimentation
- Analytics

Business applications
- Forecasting
- Predictive servicing
- Risk management
- Ops optimization
- UI/UX
- Model effectiveness
- ROI
- User engagement
- Operational efficiency
Life in data science

Key areas
- Machine learning
- Experimentation
- Analytics
  - AB testing
  - Multi-armed bandits

Forecastsing
Predictive servicing
Risk management
Ops optimization

UI/UX
Model effectiveness

ROI
User engagement
Operational efficiency

Business applications

Amazon EC2

0.00342
0.0012
0.00549

Image of credit cards and financial website.

The Capital One® Platinum MasterCard®
Exclusively for Costco members.
Life in data science

Key areas
- Machine learning
- Experimentation
- Analytics

Business applications
- Forecasting
- Predictive servicing
- Risk management
- Ops optimization

- UI/UX
- Model effectiveness
- ROI
- User engagement
- Operational efficiency

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Credit score</th>
<th>Delinquencies</th>
<th>...</th>
<th>Credit limit</th>
<th>APR</th>
<th>Default in 2 mths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>675</td>
<td>0</td>
<td></td>
<td>5000</td>
<td>22.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>570</td>
<td>1</td>
<td></td>
<td>3000</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>720</td>
<td>0</td>
<td></td>
<td>4000</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>512</td>
<td>1</td>
<td></td>
<td>3000</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

“Customer A has 5% chance of defaulting within next 2 months”
Life in data science

Key areas
- Machine learning
  - Forecasting
  - Predictive servicing
  - Risk management
  - Ops optimization
- Experimentation
  - AB testing
  - Multi-armed bandits
- Analytics
  - UI/UX
    - Model effectiveness
  - ROI
    - User engagement
    - Operational efficiency

Business applications
- Life in data science
- Old Site Navigation, in use since 2015 Site Redesign
- New Site Navigation Concept; A/B test launch 5/30
Physicists in data science

Data scientists come from diverse educational backgrounds, but are disproportionately represented by graduate degrees in math/natural sciences.

\[ P(\text{math/science grad} \mid \text{DS}) \gg P(\text{math/science grad}) \]

Your background will serve you well:

\[ P(\text{DS} \mid \text{math/science grad}) \gg P(\text{DS}) \]
# Physicists in data science

<table>
<thead>
<tr>
<th>Hard skills:</th>
<th>Soft skills:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Python</td>
<td>- Communication</td>
</tr>
<tr>
<td>- SQL</td>
<td>- Schmoozing</td>
</tr>
<tr>
<td>- Stat 101</td>
<td>- Teamwork</td>
</tr>
<tr>
<td>- Machine learning basics</td>
<td>- Problem-solving</td>
</tr>
<tr>
<td>- R</td>
<td></td>
</tr>
<tr>
<td>- Hadoop</td>
<td></td>
</tr>
<tr>
<td>- Spark</td>
<td></td>
</tr>
<tr>
<td>- Cloud computing</td>
<td></td>
</tr>
</tbody>
</table>
# Physicists in data science

## Hard skills:
- Python
- SQL
- Stat 101
- Machine learning basics
- R
- Hadoop
- Spark
- Cloud computing

## Soft skills:
- Communication
- Schmoozing
- Teamwork
- Problem-solving

## Python & SQL:
- www.codeacademy.com
- www.hackerrank.com
- leetcode.com

## Machine learning:
- www.coursera.org/learn/machine-learning
- www.analyticsvidhya.com
- towardsdatascience.com
# Physicists in data science

<table>
<thead>
<tr>
<th>Hard skills:</th>
<th>Soft skills:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Python</td>
<td>- Communication</td>
</tr>
<tr>
<td>- SQL</td>
<td>- Schmoozing</td>
</tr>
<tr>
<td>- Stat 101</td>
<td>- Teamwork</td>
</tr>
<tr>
<td>- Machine learning basics</td>
<td>- Problem-solving</td>
</tr>
<tr>
<td>- R</td>
<td>-</td>
</tr>
<tr>
<td>- Hadoop</td>
<td>-</td>
</tr>
<tr>
<td>- Spark</td>
<td>-</td>
</tr>
<tr>
<td>- Cloud computing</td>
<td>-</td>
</tr>
</tbody>
</table>

Ever done any:
- Hypothesis testing?
- Dimensionality reduction, e.g. PCA?
- Clustering?
- Signal processing?
- Regression?
- etc …

Have you:
- Worked in collaboration with others?
- Given talks?
- Independently defined and solved problems?
- etc …
How to get in the game

Resume:
- Short and sweet
- Translate your experience (previous slide) into DS terms
- Quantify the impact of your work
- Talk to Career Office and recruiters about resume tips

Talking to people:
- For career advice, quality > quantity, ask people with multiple years of experience in leading companies
- For opportunities, anything goes
- Do not ask people for a job, unless they are actively recruiting
- If seeking advice, explicitly ask for it, e.g. don’t expect people to reach out just because you connected with them
- Expect to be ignored many times
- Maintain good relationships with recruiters
- Request mock interviews (important!!)

Signaling:
- Bootcamps
- Side projects
- Blogs
- Stat/ML courses

Applying to a role:
- Personalize your resume
- Cover letters are usually unnecessary
- Learn about the business
- Learn about how current employees feel (Glassdoor)
- Research typical compensation for the role and location (Comparably+Glassdoor+Google)