Academia ➔ Data Science

Nhung Ho
Why Data Science?
Why Data Science?

Analytics and Data Science Job Growth

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018:

<table>
<thead>
<tr>
<th>Year</th>
<th>Employment</th>
<th>Graduates</th>
<th>Others</th>
<th>2018 Supply</th>
<th>Talent Gap</th>
<th>2018 Projected Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>150</td>
<td>180</td>
<td>30</td>
<td>300</td>
<td>140-190</td>
<td>440-490</td>
</tr>
</tbody>
</table>

What is Data Science?

Consultant + Quant + Engineer

Consultant: Explores, identifies, advises
Quant: strategizes, codes
Engineer: product development

= Data Scientist
Data Science skills

Hacking Skills

Math & Statistics Knowledge

Substantive Expertise

Danger Zone!

Machine Learning

Data Science

Traditional Research
Data Science skills

- Computer Science
- Machine Learning
- Math and Statistics
- Unicorn
- Traditional Software
- Traditional Research
- Subject Matter Expertise
Types of Data Scientists: 
**Product Analytics**

What is it? 
Business analytics 
Skills? 
Curious, data explorers 
Statistics heavy, strong communication 
Who do they work for? 
Product managers, engineers 
Goals are to improve *internal* metrics
Types of Data Scientists: Product Analytics

Daily Active Users Breakdown

- Unique Page Views
- Post Viewers
- Liked a Post
- Commented on a Post
- Wall Posts
Types of Data Scientists: Predictive Modeling

What is it?
Risk, sales, optimization

Skills?
Machine learning
Statistics heavy

Who do they work for?
Goals are to optimize performance, reduce risk
Types of Data Scientists: Predictive Modeling

The line represents The Efficient Frontier, the optimal combination of risk and return.

Each dot represents a portfolio. Those closest to The Efficient Frontier have the potential to produce the greatest return with the lowest degree of risk.

EXPECTED RETURN

RISK/VOLATILITY
(STANDARD DEVIATION)
Types of Data Scientists: Data Products

What is it?
Product development
Skills?
Machine learning, statistics
Software engineering
Rapid prototyping
Who do they work for?
End user
# Types of Data Scientists: Data Products

<table>
<thead>
<tr>
<th>Name</th>
<th>Title/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pieter Dokkum</td>
<td><a href="mailto:pvdokkum@gmail.com">pvdokkum@gmail.com</a></td>
</tr>
<tr>
<td>Andrei Tokovinin</td>
<td>2nd Astronomer at Cerro Tololo Inter-American Observatory, Chile</td>
</tr>
<tr>
<td>Paul Goudroojj</td>
<td>2nd Senior Scientific Staff member at STScI, Baltimore, Maryland Area</td>
</tr>
<tr>
<td>Meg Schwamb</td>
<td><a href="mailto:megan.schwamb@gmail.com">megan.schwamb@gmail.com</a></td>
</tr>
<tr>
<td>Virgil Griffith</td>
<td>2nd Scientist &amp; Internet Man of Mystery, Greater Los Angeles Area</td>
</tr>
<tr>
<td>Beate Heinemann</td>
<td>2nd Professor of Physics at UC Berkeley and Lawrence Berkeley National Laboratory, Oakland, California</td>
</tr>
<tr>
<td>Nitya Kallivayalil</td>
<td><a href="mailto:nityak@gmail.com">nityak@gmail.com</a></td>
</tr>
<tr>
<td>Joseph Prang</td>
<td>2nd Analyst at Square, San Francisco, California</td>
</tr>
<tr>
<td>Hanh Huynh</td>
<td>2nd Premier Service Center Specialist, Salt Lake City, Utah</td>
</tr>
<tr>
<td>Ricardo Munoz</td>
<td><a href="mailto:kleinameise2@gmail.com">kleinameise2@gmail.com</a></td>
</tr>
</tbody>
</table>
Why are PhDs suited for this?

Experimental design
Statistics \textit{Fast learners}
Hypothesis driven \textit{Curious}
Hackers
Creative thinkers
Coders
Why are PhDs suited for this?

Data storytellers

Experimental design
Statistics
Fast learners
Hypothesis driven
Careers
Hackers
Creative thinkers
Coders
My transition

PhD 10%

90%
My transition

INSIGHT
DATA SCIENCE FELLOWS PROGRAM

Program Timeline

6 Weeks
Intro Week
Data Project
Company Visits

3-5 Weeks
Job Interviews
Insight Data Science

>30 companies

~20/500 applicants

100% placement

insightdatascience.com

July 28 session in NYC

deadline April 27
Who are these PhD data scientists?

- Phds+Postdocs
- Breakdown:
  - ~40% physics+astronomy
  - ~20% biological sciences
  - ~20% engineering
  - ~10% math+statistics
  - ~10% social sciences

% of Insight Fellows
What is the missing 10%?

• Machine Learning
  – Linear, logistic regression, random forest
• Databases
  – SQL, MongoDB
• Statistics
  – T-tests, p-value, Bayes’ theorem
• Computer Science fundamentals
  – Sorting algorithms, data structures, recursive programming, dynamic programming
Getting up to 100%

• **Machine Learning**
  – Andrew Ng’s Coursera Machine learning course

• **Databases**
  – Sqlzoo.net, Coursera Databases course

• **Statistics**
  – Khan Academy statistics course

• **Computer Science fundamentals**
  – Interactivepython.org

• **Kaggle competitions**!
Tools of the trade aka “stack”

- Python, python, python…
  - Pandas, numpy, scikitlearn,nltk, BeautifulSoup
- SQL, Hadoop, Hive, Pig
- Flask
- Javascript, Ruby on Rails
- No more Matlab or IDL, escape while you can!
My transition timeline

September 27: Defense
November 6: Apply to Insight

Jan 6  
Feb 21  
March 12  
Feb 24
Keep in touch!

www.astro.yale.edu/nth4
Linkedin.com/in/nhungho
Twitter.com/nhungkia
Email: nhung.t.ho@gmail.com