Data Science 101
(Note: PhD focused)
About DOC (dropoutclub.org)

Our aspiration is to unite the global community of doctors, scientists and other biomedical professionals who seek to shape healthcare through innovative careers outside of traditional clinical and research tracks.

We focus on 3 specific objectives:

- Connect members with great opportunities that leverage their unique backgrounds and experience
- Help employers rapidly source talent with highly specific biomedical and business experience
- Facilitate the online and in-person exchange of ideas, insights and opportunities among our members

Ultimately we hope that this will help improve the healthcare system by placing those who understand the real content of healthcare in leadership positions.

Contact us at contact@dropoutclub.org
Contents

**Objective:** Understand the field of data science, your potential place in it, and what it will take to get there

1. Overview of data science
   - Definition and the broader market demand

2. Relevancy for PhDs
   - Why PhDs are a good fit for this career
   - The opportunities for PhDs in data science

3. Different roles in data science
   - What they are and how to recognize them
   - What to look for in job descriptions

4. What next?
1 Data Science: An overview
What is data science?

**Simple:** People who work on making data more useful
Let’s look an example of DS implementation

What is happening behind these recommendations?
There is great demand for data analysis expertise

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

Supply and demand of deep analytical talent by 2018

 Thousand people

<table>
<thead>
<tr>
<th>Year</th>
<th>Employment</th>
<th>Graduates with deep analytical talent</th>
<th>Others¹</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>

¹ Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

Source: McKinsey Global Institute, “Big data: the next frontier for innovation, competition, and productivity”, 2011
1 Data Science: the relevancy for PhDs
Why are PhDs good candidates for DS?

- Statistics is at the core of all of the work we do
- Many of us use scripting languages in our work (e.g. Matlab for cell tracking)
  - We all *could* use scripting in our work
- We have *time* to develop skills
- Degree suggests intelligence, grit, curiosity, technical mastery
PhD hiring for DS roles, by state

January 15, 2016 data

Copyright © 2016 DropOutClub LLC. All rights reserved.
Skills requested from PhDs for DS jobs

January 15, 2016 data
Scripting languages requested from PhDs for DS jobs

January 15, 2016 data

Copyright © 2016 DropOutClub LLC. All rights reserved.
What do I look for?
The two types of data scientists:
#1 Data Engineers

Who they are:
- Software developers that use data

What they do:
- Building tools that utilize data
- User-facing data implementation
- Wrote front-end code to connect the website to the data models created by someone else
- Wrote code to track clicks and send that data back to some db
The two types of data scientists: #2 Data Scientists

Who they are:

- Statisticians who code their analyses

What they do:

- Make sense of data
- People who clean, analyze, visualize, and experiment with data

This is the role we are going to focus on

---

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Tweet by Josh Wills (@josh_wills):

9:55 AM - 3 May 12
You need to know how to code
# How close to being competitive for a DS position am I?

## Level of competitiveness

<table>
<thead>
<tr>
<th>I use Excel for all analyses</th>
<th>I perform ANOVAs, co-factor out variables, etc in Prism and like software</th>
<th>I code simple analyses and graphs (in R or Python or Matlab)</th>
<th>I can build data pipelines and use a scripting language to handle all data cleaning and manipulation</th>
<th>I have implemented machine learning-based analyses on data</th>
<th>I have implemented multiple forms of machine learning on data</th>
</tr>
</thead>
</table>

12 months of working on it for 15 hours a week.

Lindsay
## Language in job descriptions to look for

<table>
<thead>
<tr>
<th>Data scientist terms</th>
<th>Data engineer terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “Analyst”</td>
<td>• “Programmer”</td>
</tr>
<tr>
<td>• “Statistician”</td>
<td>• “Engineer”</td>
</tr>
<tr>
<td>• “Analysis”</td>
<td>• “Data implementation”</td>
</tr>
<tr>
<td>• “Data insights”</td>
<td>• “Software”</td>
</tr>
<tr>
<td>• “Big data”</td>
<td>• “Back-end”</td>
</tr>
<tr>
<td>• “Traffic”</td>
<td>• “Recommendation systems”</td>
</tr>
<tr>
<td>• “Customer segmentation”</td>
<td>• “Computer science”</td>
</tr>
<tr>
<td>• “Work with developers”</td>
<td>• “database administration”</td>
</tr>
</tbody>
</table>
What is next?
If you’re not yet qualified...

<table>
<thead>
<tr>
<th>True beginner</th>
<th>Intermediate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>You don’t code, and do simple statistics.</strong></td>
<td><strong>You code okay and/or you know intermediate stats okay</strong></td>
</tr>
</tbody>
</table>

- Start using R right away, even for your simple analyses
- Enroll in the John Hopkins ‘Data Science Specialization’ course ($29)
- Find a patient friend who is much better than you

- You need a real project with real impact to truly improve your skills
  - Take your project’s data analysis to another level *(only do if you can significantly beef it up)*
  - A side project *(only do if you are truly passionate about it)*
  - Do more complex data analysis for someone else
  - Find a legit biostatistician
  - Use their help for this project
If you think you are already qualified…

<table>
<thead>
<tr>
<th>Find out if you’re qualified</th>
<th>Apply for data science bootcamps</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Change your LinkedIn blurb to say “Data scientist” instead of “Graduate student”</td>
<td>• There are TONS.</td>
</tr>
<tr>
<td>• Talk to data scientists to get a sense for where they were when the transitioned and ask for feedback</td>
<td>• Full time, part time, online</td>
</tr>
<tr>
<td>• Further test qualifications by applying for a few jobs – especially ones you are not absolutely gunning for</td>
<td>• Free and paid</td>
</tr>
<tr>
<td>• LinkedIn message a technical recruiting company (i.e. DS headhunters, ala BurtchWorks) your resume. Do they reach back out?</td>
<td>• 10-12 weeks</td>
</tr>
<tr>
<td></td>
<td>• Best resource to figure this out:</td>
</tr>
<tr>
<td></td>
<td>• <a href="http://www.skilledup.com/articles/list-data-science-bootcamps">http://www.skilledup.com/articles/list-data-science-bootcamps</a></td>
</tr>
<tr>
<td></td>
<td>• Yet-another-data-blog.blogspot.com</td>
</tr>
</tbody>
</table>