**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

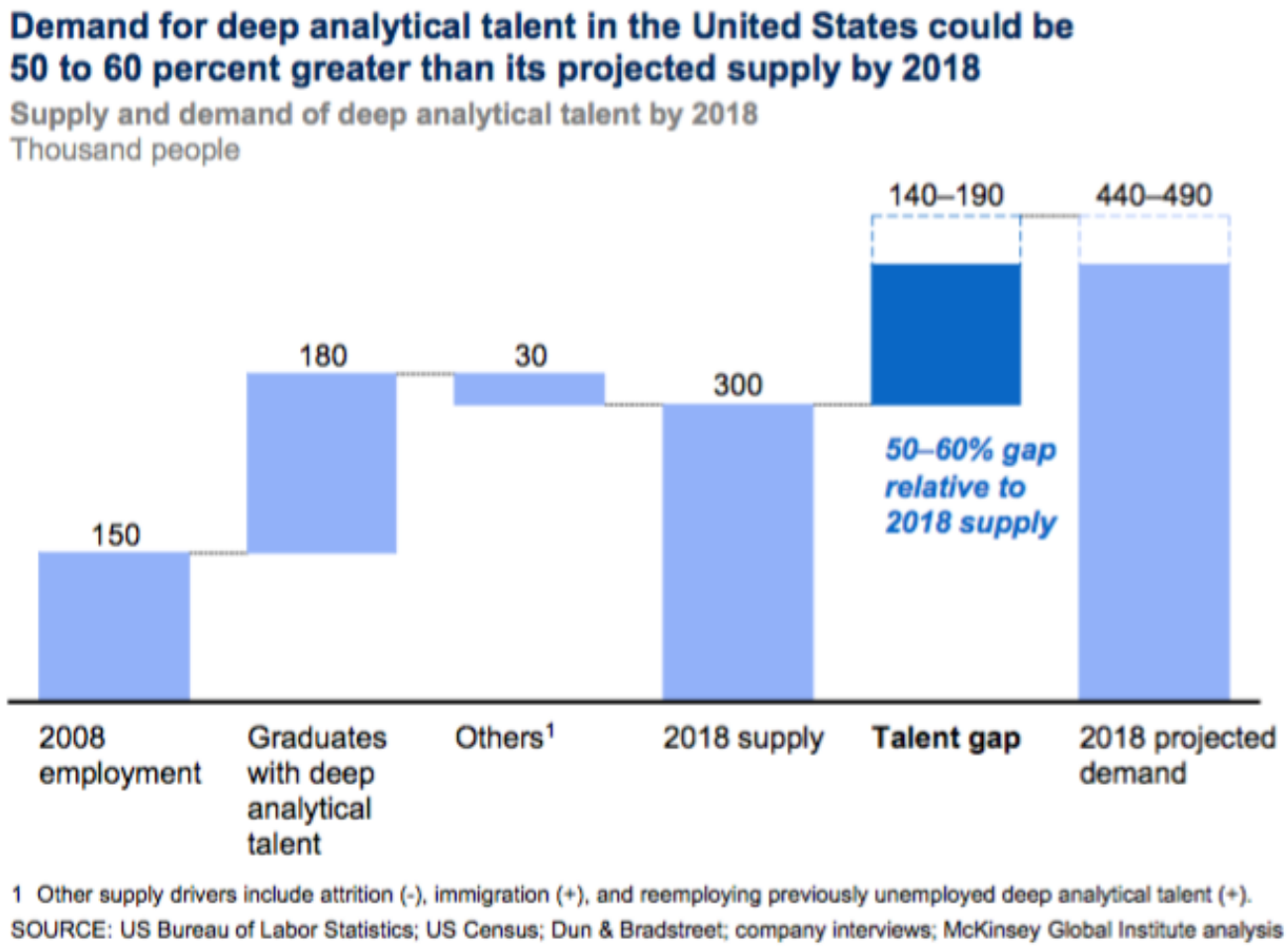
# What is data science?

**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



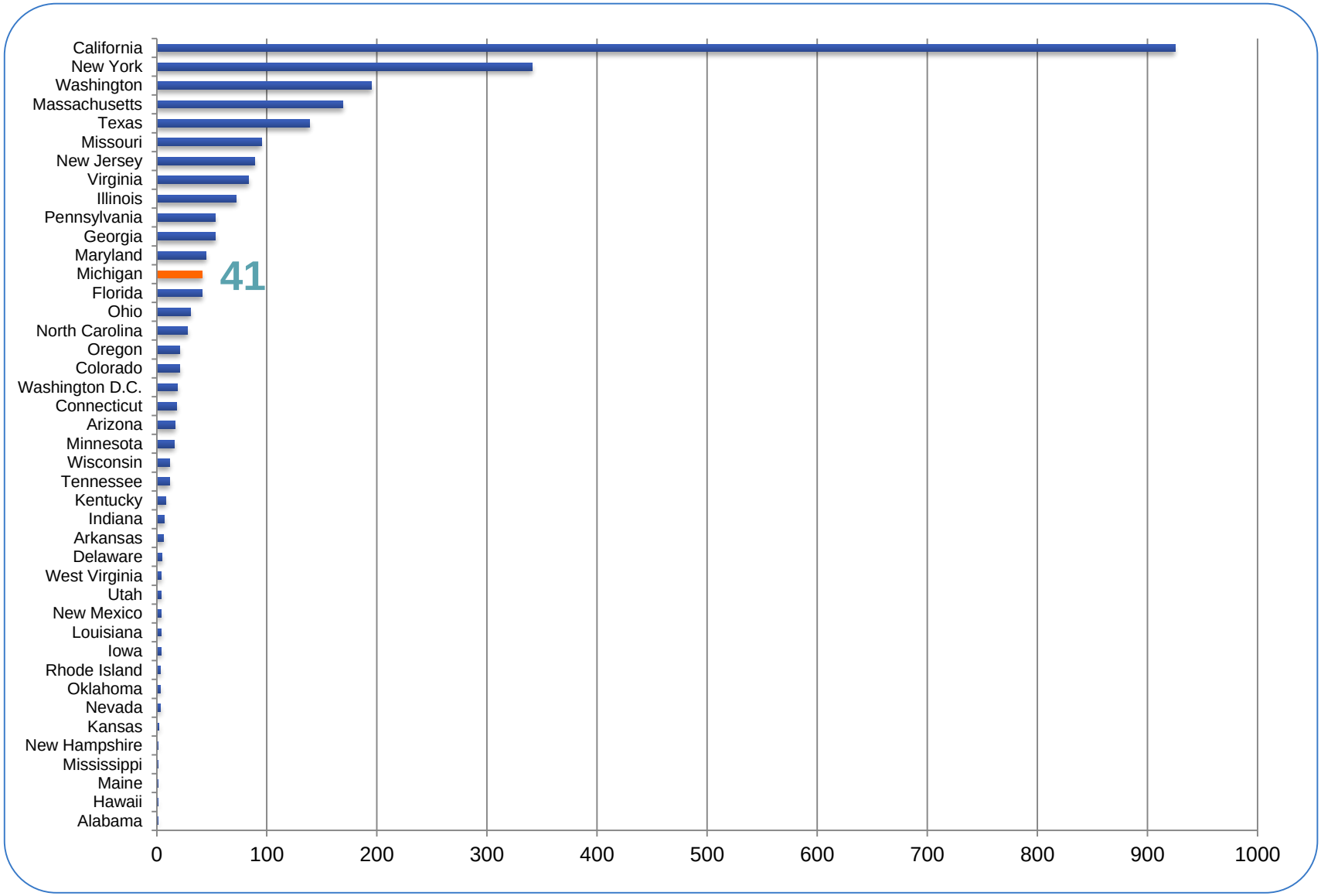

Source: McKinsey Global Institute, “Big data: the next frontier for innovation, competition, and productivity”, 2011

# 1 Data Science: the relevancy for PhDs

# Why are PhDs good candidates for DS?

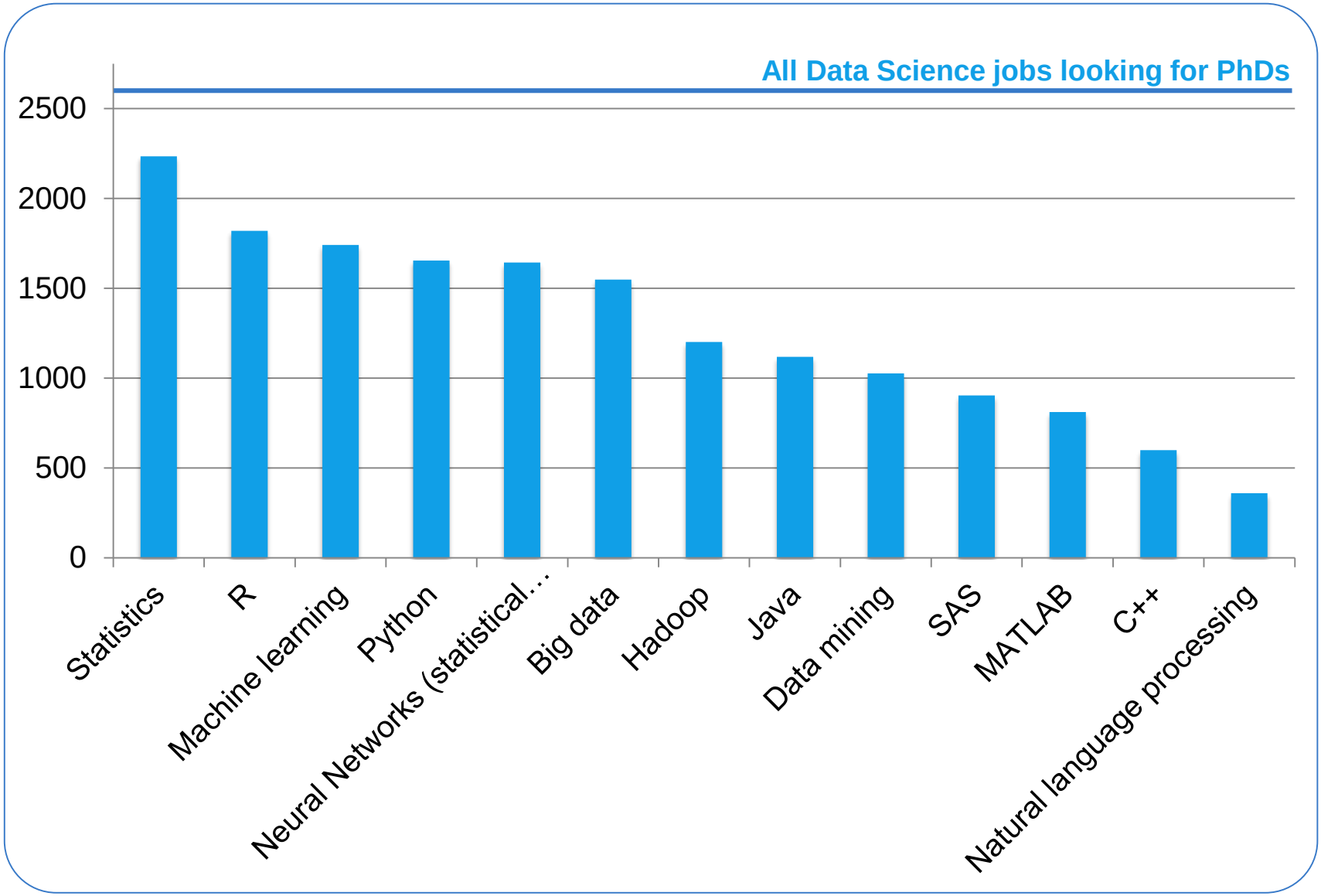
- Statistics is at the core of all of the work we do
- Many of us use scripting languages in our work (e.g. Matlab for cell tracking)
  - We all *could* use scripting in our work
- We have *time* to develop skills
- Degree suggests intelligence, grit, curiosity, technical mastery

# PhD hiring for DS roles, by state



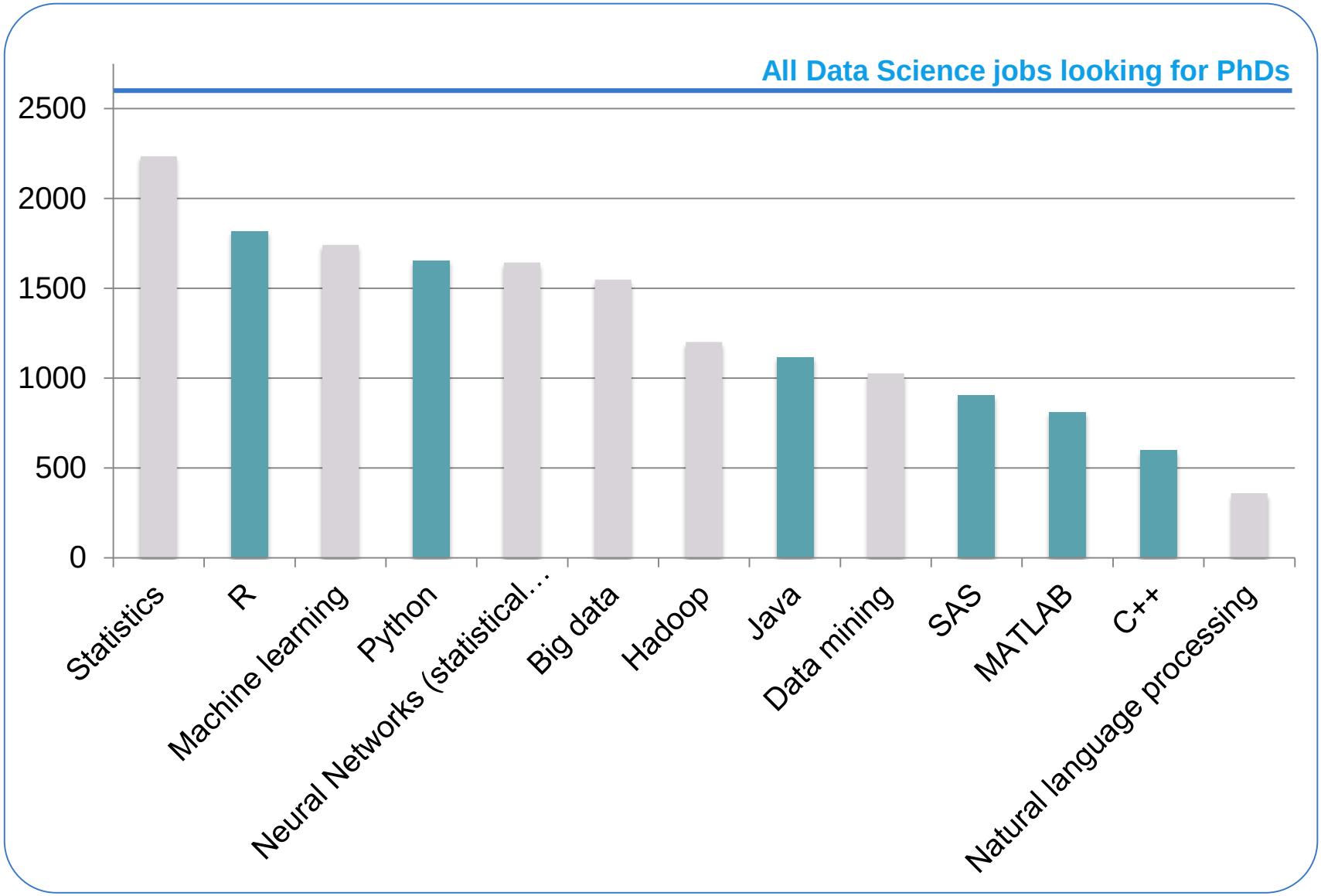

January 15, 2016 data

# Skills requested from PhDs for DS jobs




January 15, 2016 data

# Scripting languages requested from PhDs for DS jobs




January 15, 2016 data

# 3 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

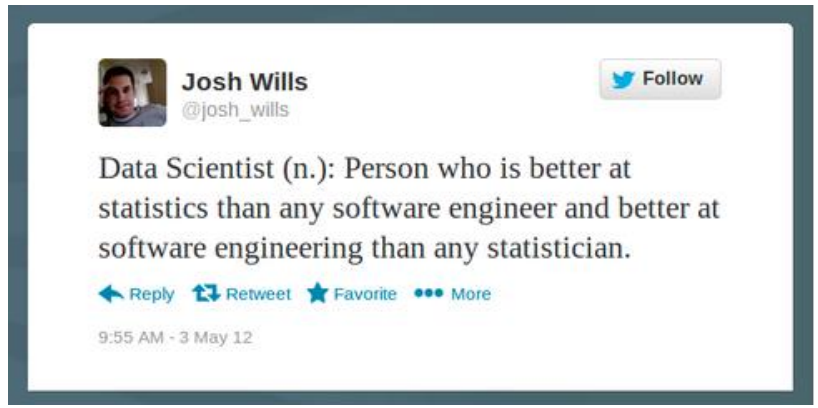
This is the role we are going to focus on

Who they are:

- Statisticians who code their analyses

What they do:

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



# You need to know how to code

```
movies=read.csv("~/Desktop/Data incubator/Project proposal/movies.csv")
graph=matrix(ncol=(dim(movies)[1]+1),nrow=length((-4):4))
colnames(graph)=c(as.character(movies$movie_name),"rel_month")
graph[1:dim(graph)[1],dim(graph)[2]]=paste((-4):4)
movies$ymd=as.POSIXct(paste(movies$year_released,
                                  movies$month_released,
                                  movies$day_released,sep="-"),
                            format="%Y-%m-%d",
                            tz="")

for (i in 1:dim(movies)[1]){
    for (j in 1:dim(graph)[1]){
      adj=j-4
      graph[j,i]=mean(
        comics$Estimated_sales[which(mapply(function(x) grepl(movies$superhero[i], x, fixed=TRUE),x=comics$Comic.book.Title)
                                     & comics$Year %in% year(movies$ymd[i])
                                     & comics$Month %in% (month(movies$ymd[i])+adj)
                                     & !(comics$X %in% "trade paperback"))])
    }
}
graph_melt=as.matrix(sapply(graph, as.numeric))
graph_melt=melt(graph,id=c("rel_month"))
colnames(graph_melt)=c("rel_month","movie","Estimated_sales")

ggplot(data=graph_melt[1:99,], aes(x=rel_month, y=as.numeric(as.character(Estimated_sales)), group=movie, colour=factor(movie))) +
  geom_line(size=.75) + geom_point() +
  scale_x_discrete(limits=c("-4","-3","-2","1","0","1","2","3","4")) +
  ylab("Estimated Sales (in Thousands)")

plot=matrix(ncol=(dim(graph)[2]),nrow=2)
colnames(plot)=colnames(graph)
for (i in 1:(dim(graph)[2])){
  plot[1,i]=mean(as.numeric(as.character(graph[1:4,i])))
  plot[2,i]=mean(as.numeric(as.character(graph[6:9,i])))
}
plot[1,12]=c("before")
plot[2,12]=c("after")

plot_melt=melt(plot,"rel_month")

plot_melt=melt(plot,id=c("rel_month"))
colnames(plot_melt)=c("rel_month","movie","Estimated_sales")
ggplot(data=plot_melt, aes(x=rel_month, y=as.numeric(as.character(Estimated_sales)), group=movie, colour=factor(movie))) +
  geom_line(size=.75) + geom_point() +
  scale_x_discrete(limits=c("1","2")) +
  ylab("Estimated Sales")
```

# How close to being competitive for a DS position am I?

*Level of competitiveness*

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

Lindsay

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

# Language in job descriptions to look for

## Data scientist terms

- “Analyst”
- “Statistician”
- “Analysis”
- “Data insights”
- “Big data”
- “Traffic”
- “Customer segmentation”
- “Work with developers”

## Data engineer terms

- “Programmer”
- “Engineer”
- “Data implementation”
- “Software”
- “Back-end”
- “Recommendation systems”
- “Computer science”
- “database administration”

# 4 What is next?

# If you're not yet qualified…

### True beginner

*You don't code, and do simple statistics.*

- 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

### Intermediate

*You code okay and/or you know intermediate stats okay*

- 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…

### Find out if you're qualified

- Change your LinkedIn blurb to say "Data scientist" instead of "Graduate student"
- Talk to data scientists to get a sense for where they were when the transitioned and ask for feedback
- Further test qualifications by applying for a few jobs – especially ones you are not absolutely gunning for
- LinkedIn message a technical recruiting company (i.e. DS headhunters, ala BurtchWorks) your resume. Do they reach back out?

### Apply for data science bootcamps

- There are TONS.
  - Full time, part time, online
  - Free and paid
  - 10-12 weeks
- Best resource to figure this out:
  - http://www.skilledup.com/articles/list-data-science-bootcamps
  - Yet-another-data-blog.blogspot.com