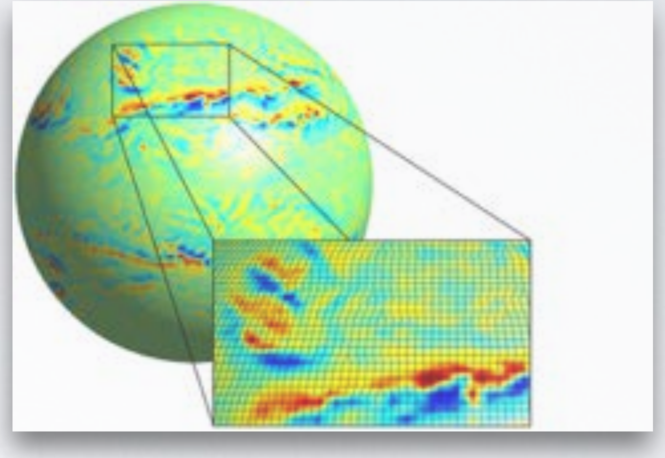
# HOW TO BECOME A DATA SCIENTIST

Jesse Steinweg-Woods, Ph.D.

# QUICK BIO

Ph.D. in Atmospheric Science from Texas A&M



Data Scientist at Argo Group Insurance (previous)

ARGO GROUP

Senior Data Scientist at tronc (Tribune Online Content)

tronc

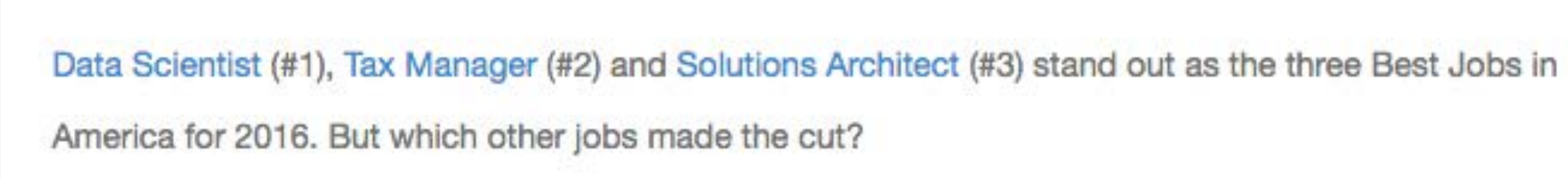
Los Angeles Times Chicago Tribune

# WHY BECOME A DATA SCIENTIST?

glassdoor®

# #1 Job (2016/2017)

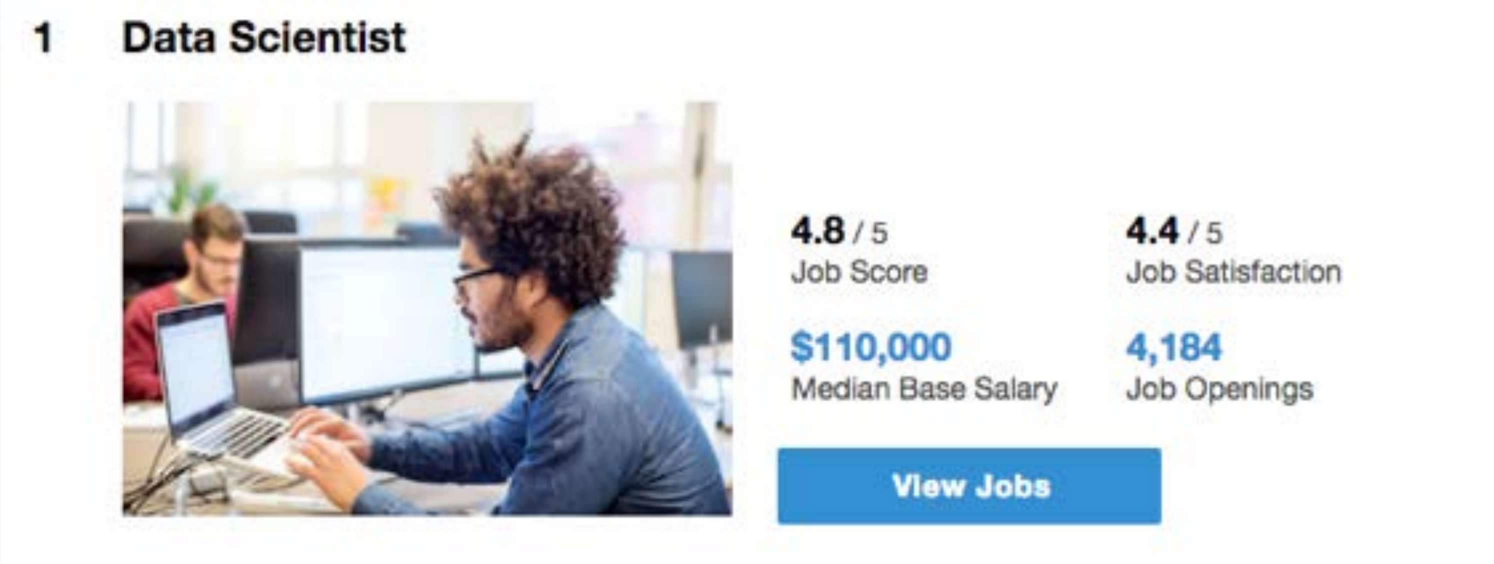
2016



Data Scientist (#1), Tax Manager (#2) and Solutions Architect (#3) stand out as the three Best Jobs in America for 2016. But which other jobs made the cut?

https://www.glassdoor.com/blog/25-jobs-america-2016/

2017



1 Data Scientist

4.8 / 5 Job Score

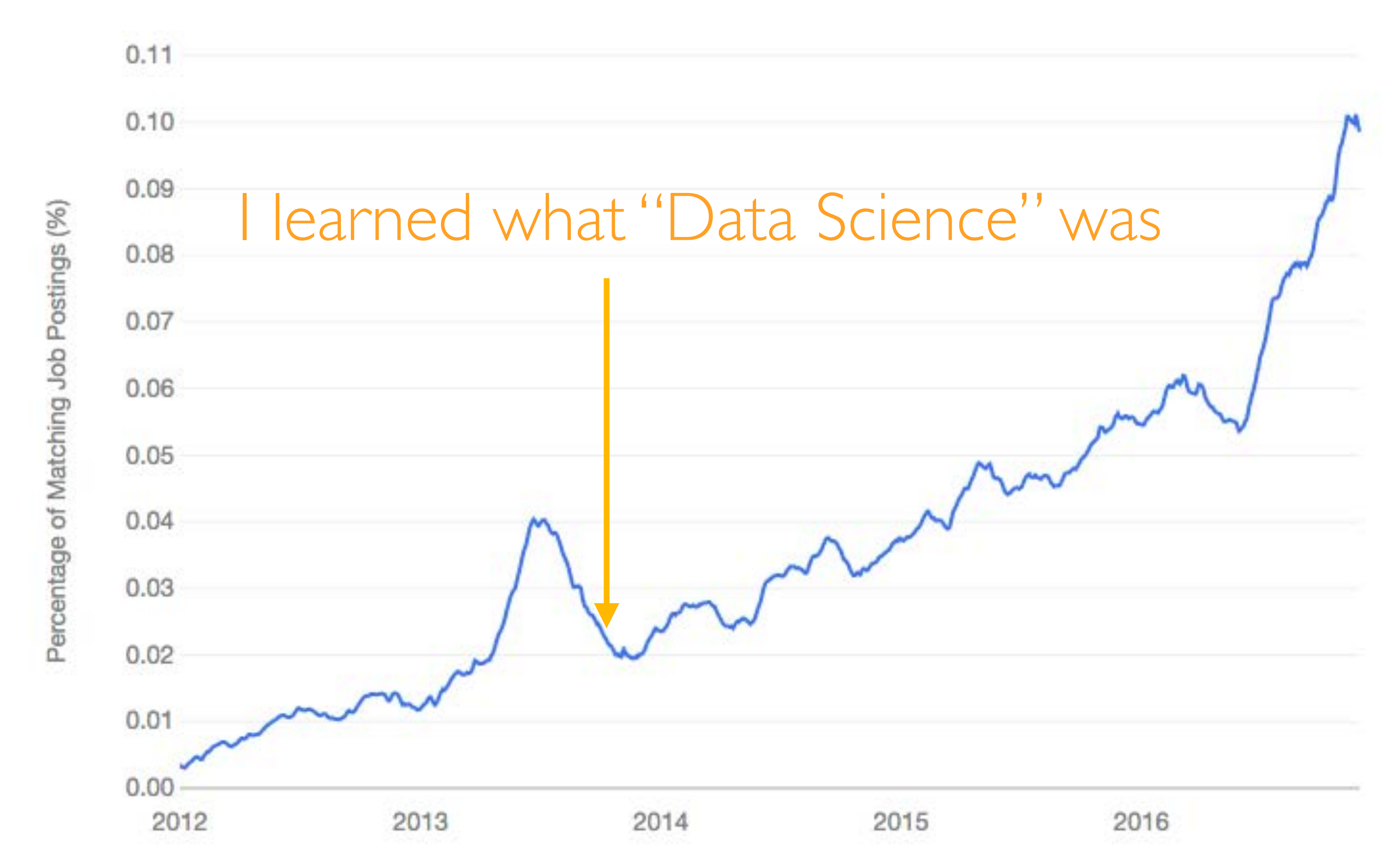
4.4 / 5 Job Satisfaction

$110,000 Median Base Salary

4,184 Job Openings

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

# INDEED JOB TRENDS FOR "DATA SCIENTIST"



I learned what "Data Science" was

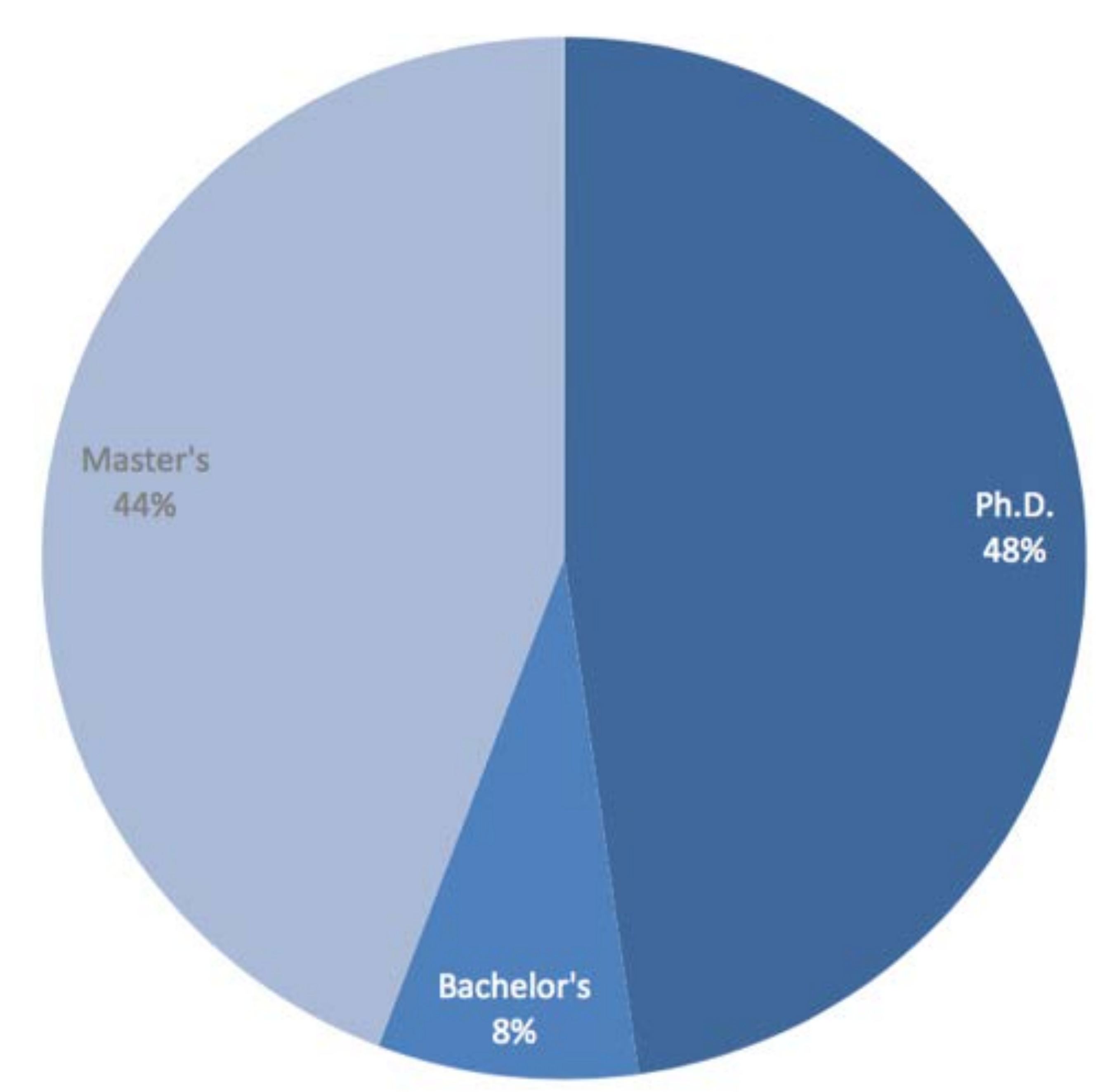
https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

# TRAINING TO BE A DATA SCIENTIST

eWEEK
What my friends think I do
What my mom thinks I do
What society thinks I do
Gradient vector
SELECT spending FROM db.users...
What my boss thinks I do
What I think I do
What I actually do

http://www.sintetia.com/wp-content/uploads/2014/05/Data-Scientist-What-I-really-do.png

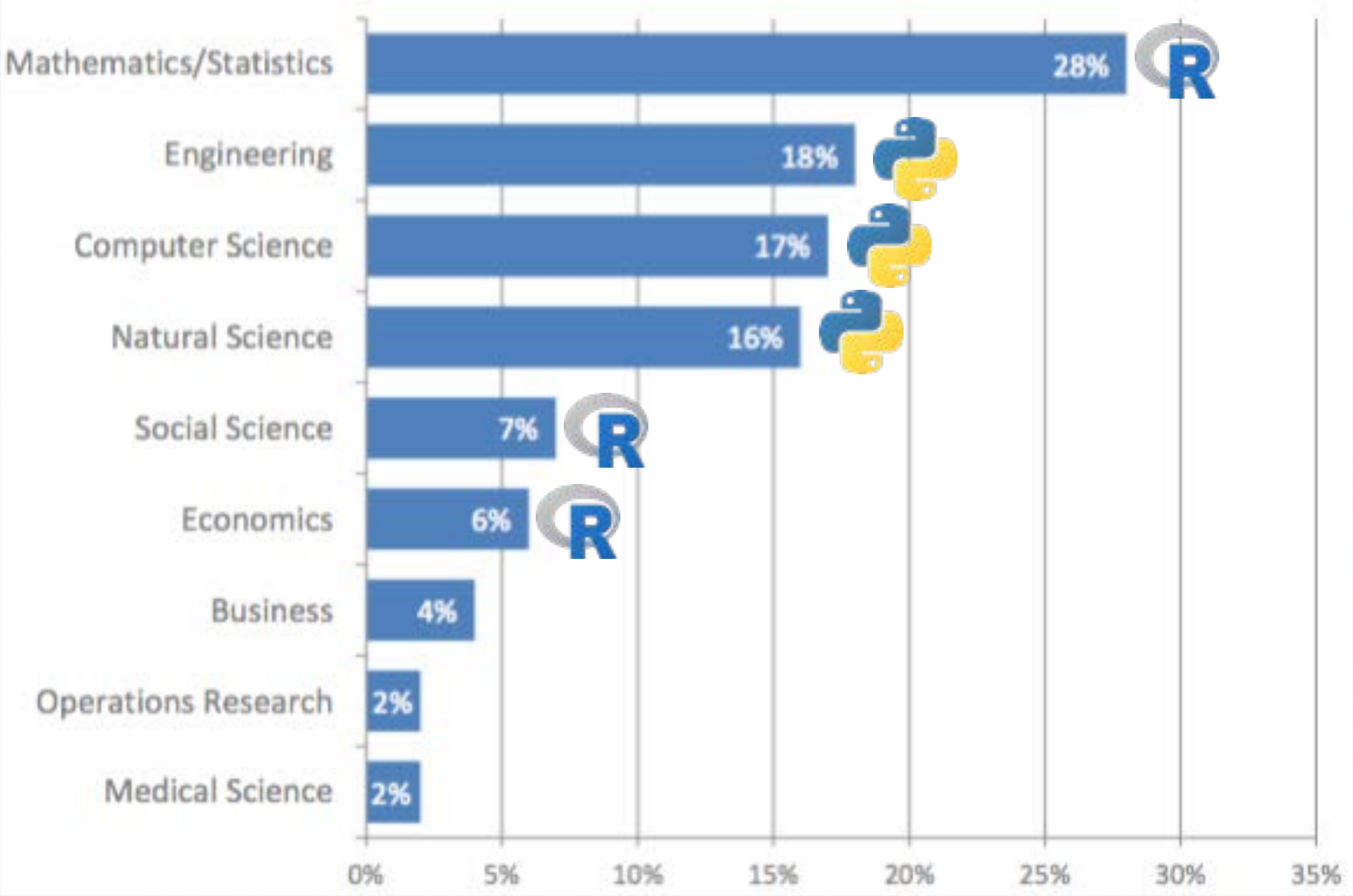
# DO YOU NEED A PH.D.?



Burtch Works 2016 Study, Data Scientist Education Levels

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

# POPULAR DATA SCIENCE BACKGROUNDS



Burtch Works 2016 Study, Data Scientist Backgrounds

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

R = R more common

🐍 = Python more common

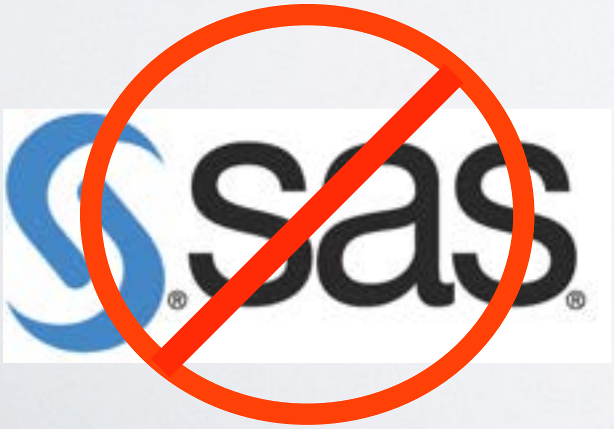
# KEY TOPICS TO LEARN

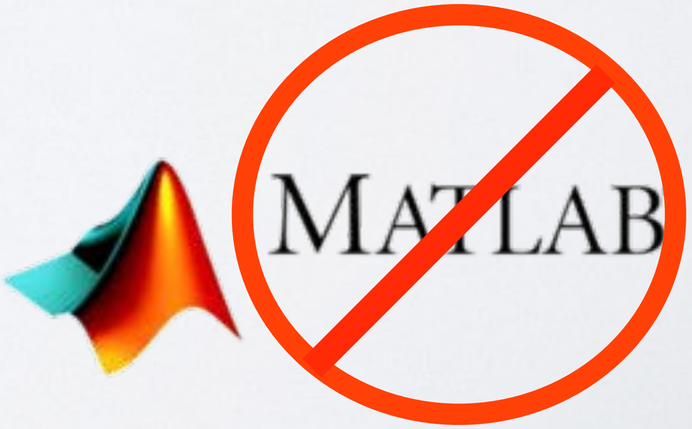
1. Pick an open-source language well-designed for Data Science

or

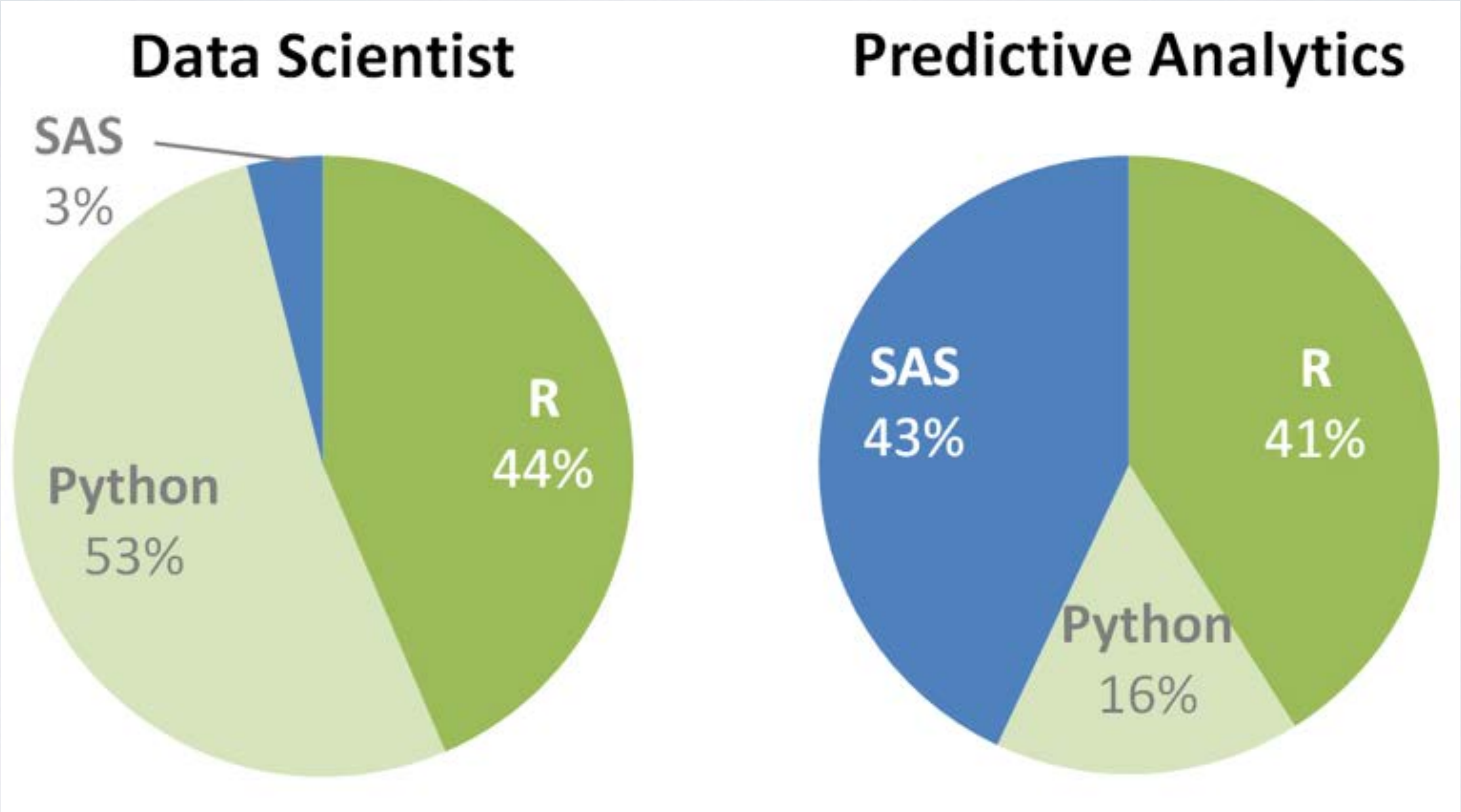
Python (my recommendation)

R (**if** you already know it well)





# TOOL USE BY POSITION



Burtch Works 2016 Tool Survey

http://www.burtchworks.com/2016/07/13/sas-r-python-survey-2016-tool-analytics-pros-prefer/

# PYTHON DATA STACK

SciPy

NumPy

matplotlib

IP[y]: IPython Interactive Computing

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikit learn

Learn these libraries well!

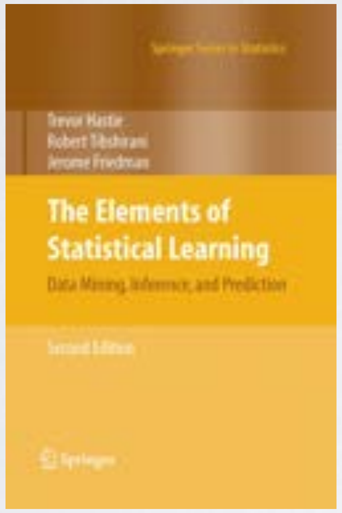
# 2. LEARN SQL

MySQL®

SQL Zoo: http://sqlzoo.net/

# 3. MACHINE LEARNING

Andrew Ng's Coursera Course:
https://www.coursera.org/learn/machine-learning

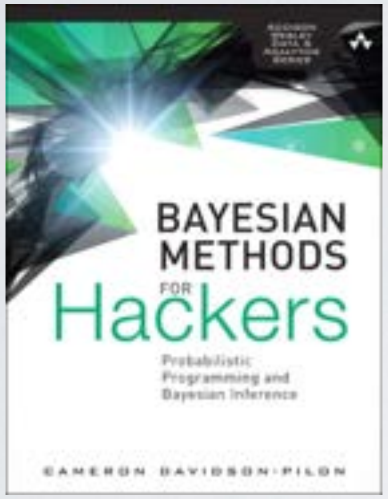

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Elements of Statistical Learning:
http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

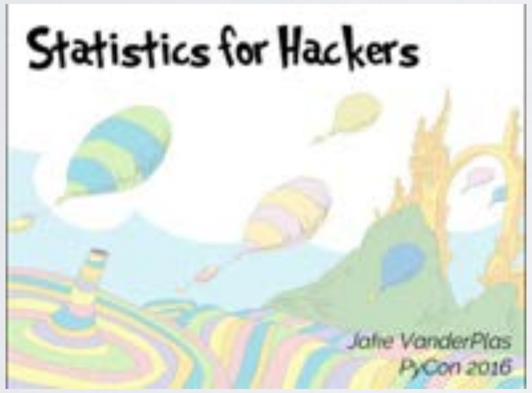
Scikit-Learn documentation:
http://scikit-learn.org/stable/documentation.html

# 4. BAYESIAN STATISTICS

Just the basics will be enough to start with …



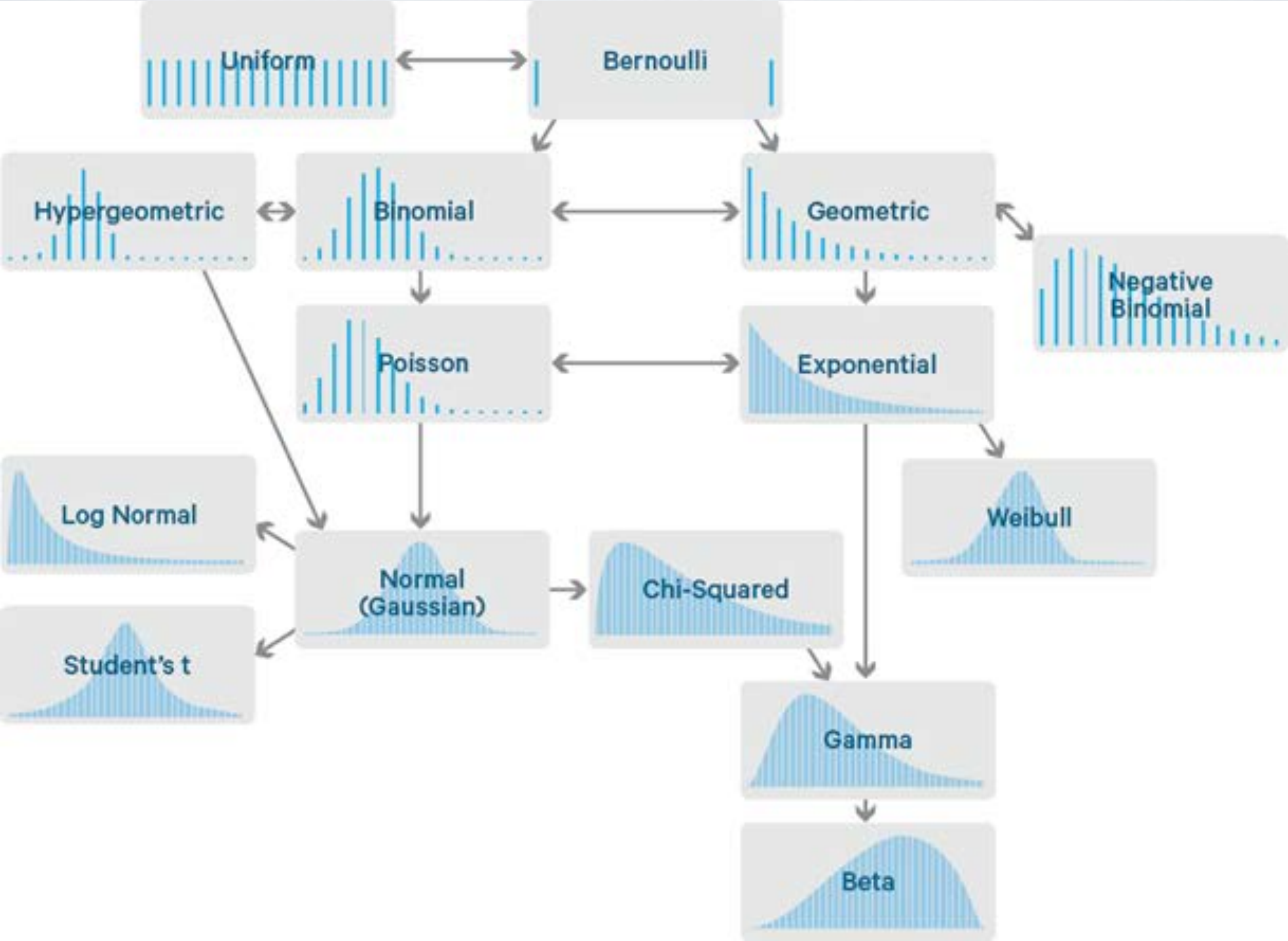
Bayesian Methods for Hackers:
https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers



Statistics for Hackers (talk by Jake VanderPlas):
https://speakerdeck.com/jakevdp/statistics-for-hackers

# 5. PROBABILITY DISTRIBUTIONS

## Know your common distributions and understand them



http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

# GETTING THE INTERVIEW

# PET PROJECTS

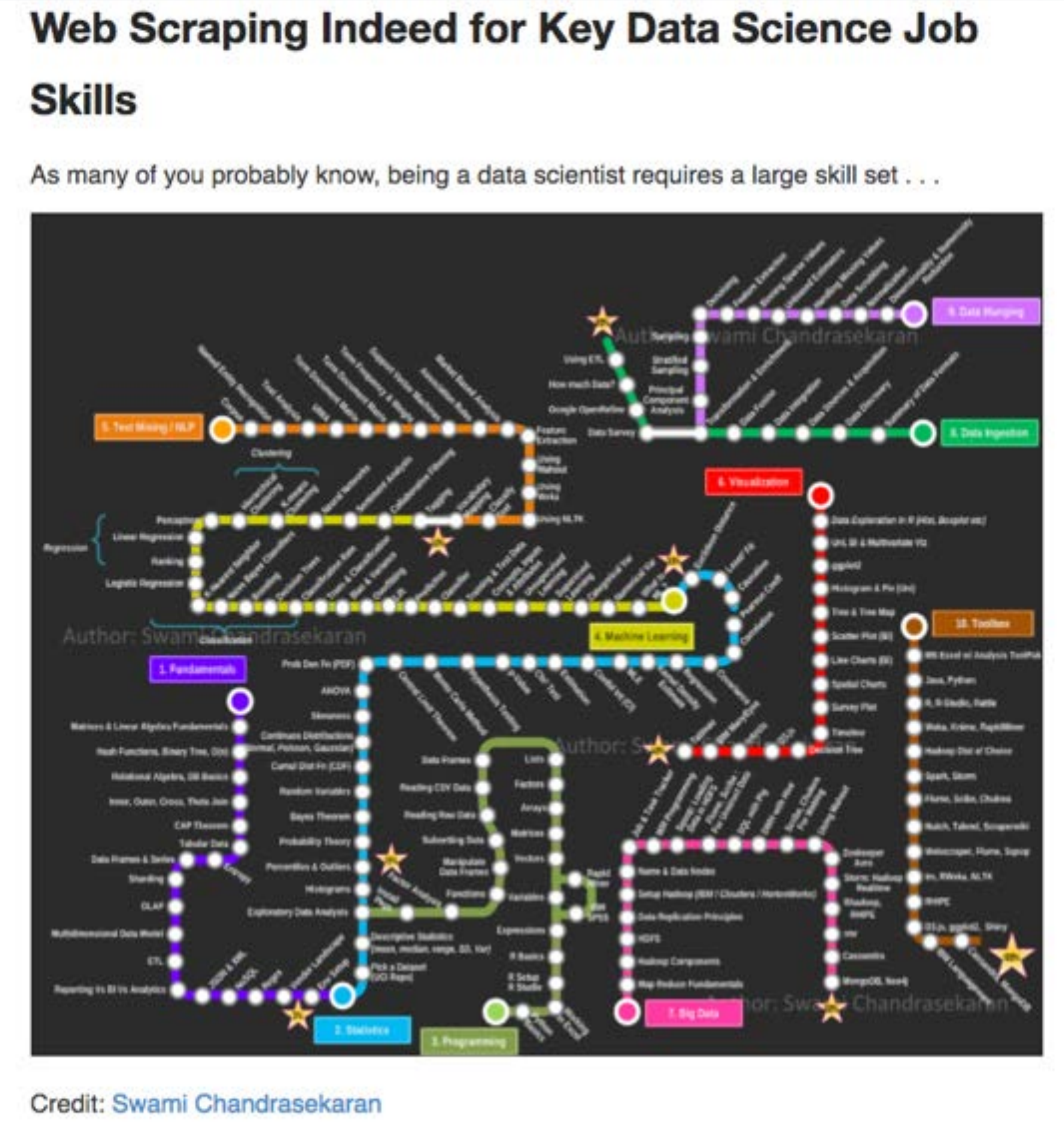
The best way to get better at data science is to **DO** data science.

**“Do a project you care about.
Make it good and share it.”**

- Monica Rogati, Data Science advisor

https://www.quora.com/How-can-I-become-a-data-scientist-1

# EXAMPLE PET PROJECTS: MY WEBSITE



## Web Scraping Indeed for Key Data Science Job Skills

As many of you probably know, being a data scientist requires a large skill set . . .

Credit: Swami Chandrasekaran

jessesw.com

# ABOUT kaggle

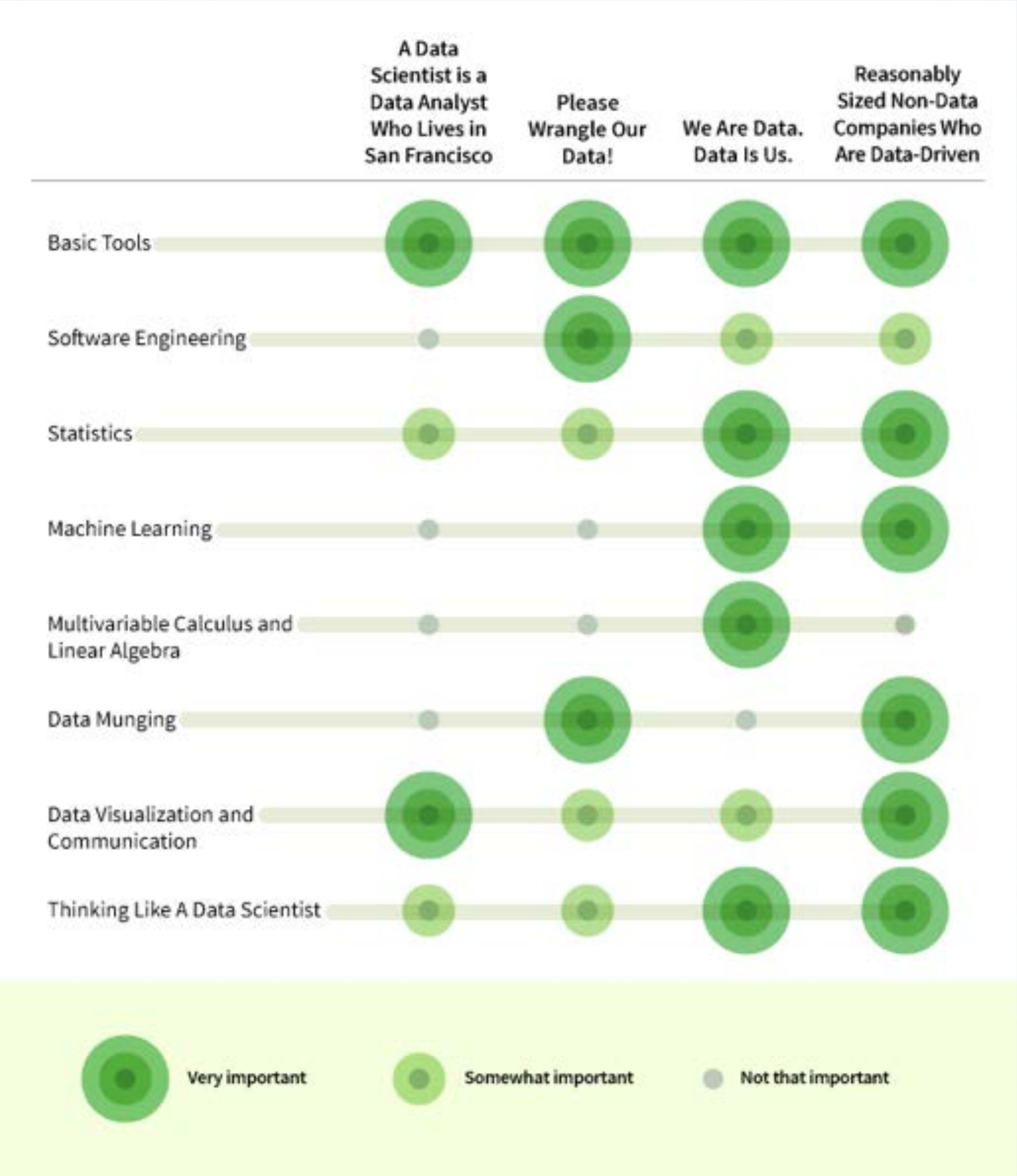
Great for practicing **machine learning**, not all of data science

Kaggle misses:

- Asking the right questions
- Decisions regarding data sources
- Which metrics to optimize
- Data Munging/Wrangling work

Pet Projects can cover **all** of these!

# WHICH JOBS TO APPLY TO



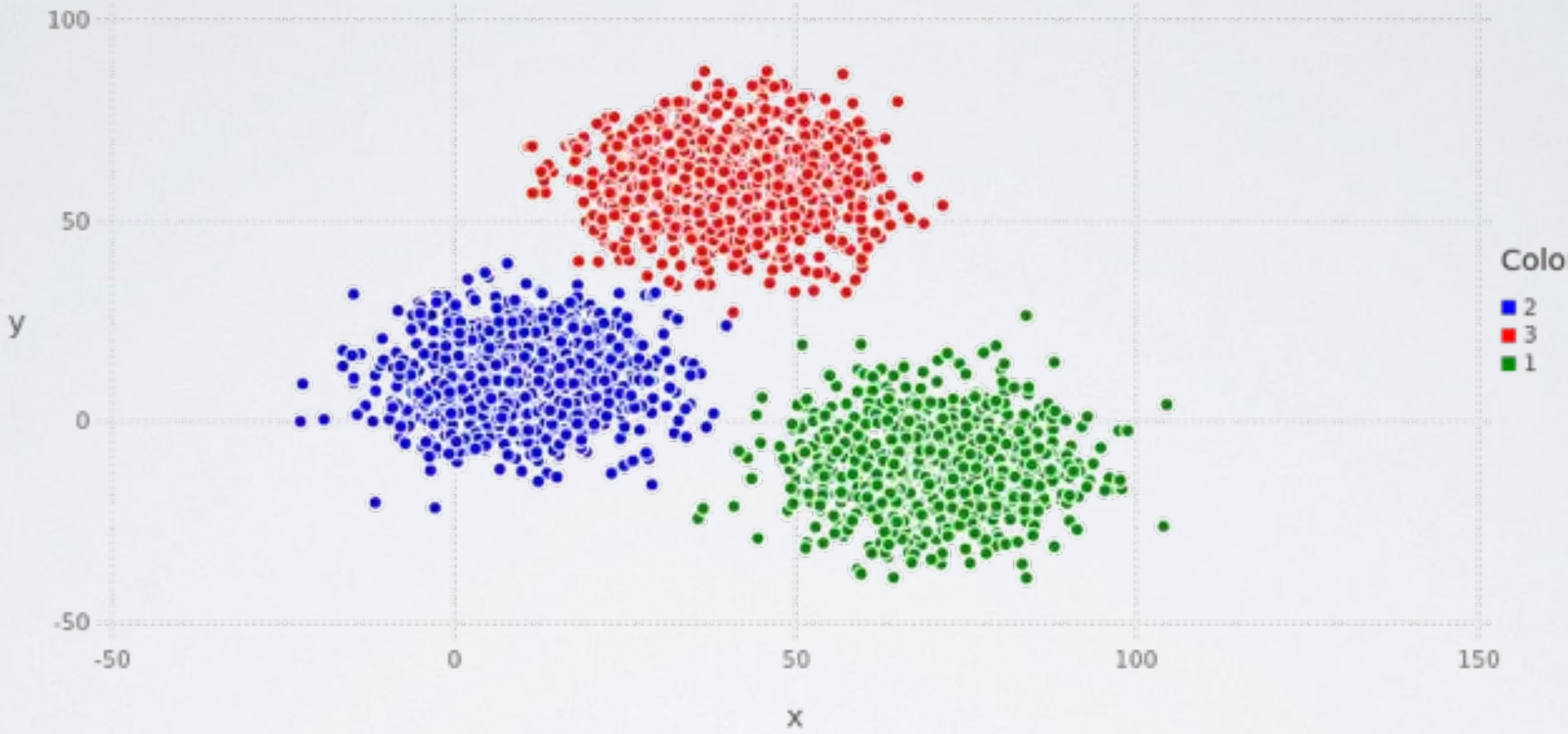
| | A Data Scientist is a Data Analyst Who Lives in San Francisco | Please Wrangle Our Data! | We Are Data. Data Is Us. | Reasonably Sized Non-Data Companies Who Are Data-Driven |
|---|---|---|---|---|
| Basic Tools | ⬤ | ⬤ | ⬤ | ⬤ |
| Software Engineering | • | ⬤ | ● | ● |
| Statistics | ● | ● | ⬤ | ⬤ |
| Machine Learning | • | • | ⬤ | ⬤ |
| Multivariable Calculus and Linear Algebra | • | • | ⬤ | • |
| Data Munging | • | ⬤ | • | ⬤ |
| Data Visualization and Communication | ⬤ | ● | ● | ⬤ |
| Thinking Like A Data Scientist | ● | ● | ⬤ | ⬤ |

⬤ Very important ● Somewhat important • Not that important

http://blog.udacity.com/2014/11/data-science-job-skills.html

# INVESTIGATE THE TEAM

Teams **tend to** like hiring people similar to themselves.



- No Ph.D. on the team? They probably don’t want one.
- All team members have a Ph.D? You probably need one.
- Are most of them computer scientists? Physical scientists? Social scientists?
- Do they seem to prefer Python, R, or a mix?

# CONTACT THE TEAM LEAD

Sometimes it is helpful to email/message the lead data scientist

# GO TO Meetup EVENTS

Networking can help with job leads and allow you to learn new things

# INTERVIEW A LOT!

Finding a good fit takes **time!**

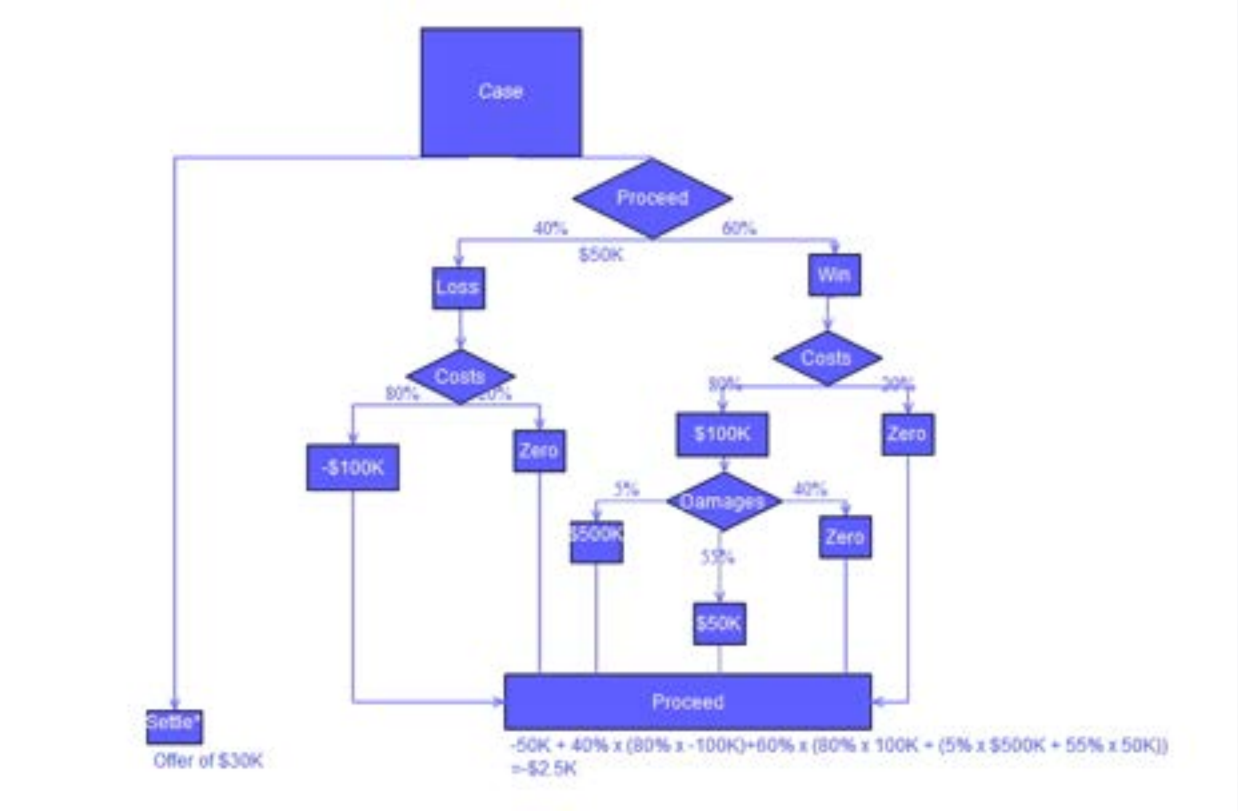
# INTERVIEW TIPS

Interview Question Types:

- Take-home machine learning task
- “Whiteboard” coding (focus on Data Structures/Algorithms)
- “Whiteboard” SQL
- Bayes' Theorem probability questions
- Machine learning evaluation metrics

# TAKE-HOME MACHINE LEARNING

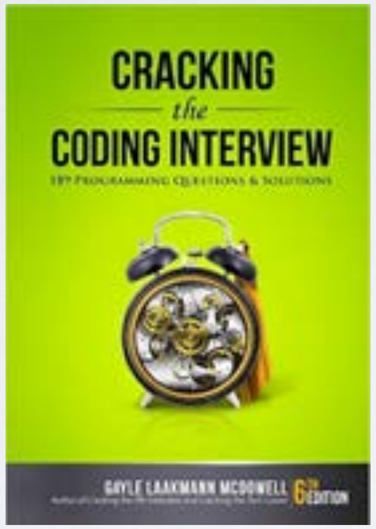
Practice with tree-based methods (random forests/gradient boosted trees)



# “WHITEBOARD” CODING

Tends to be similar to software engineer interviews, but focuses most on data structures/algorithms

Practice with:



https://www.amazon.com/Cracking-Coding-Interview-Programming-Questions/dp/0984782850

Interview Cake

https://www.interviewcake.com/

HackerRank

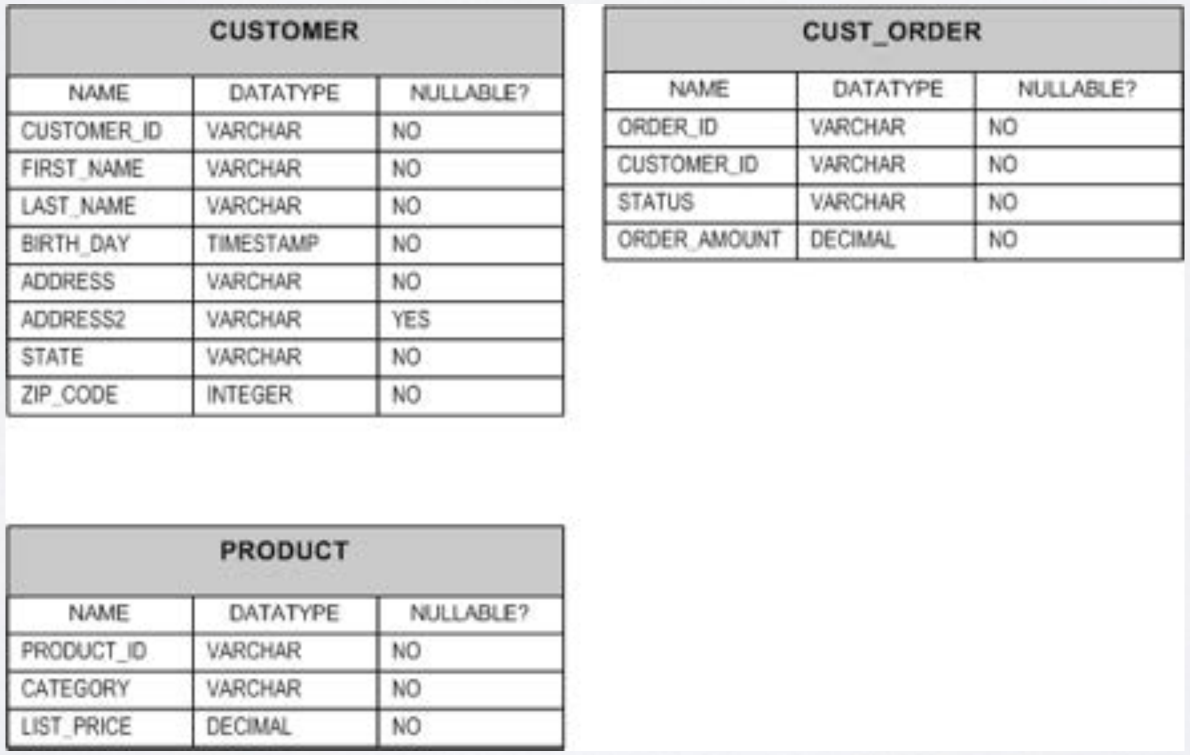
https://www.hackerrank.com/

https://projecteuler.net/

# “WHITEBOARD” SQL

These are easier if you have used SQL a fair amount

SQL Zoo is good review/practice



**CUSTOMER**

| NAME | DATATYPE | NULLABLE? |
|---|---|---|
| CUSTOMER_ID | VARCHAR | NO |
| FIRST_NAME | VARCHAR | NO |
| LAST_NAME | VARCHAR | NO |
| BIRTH_DAY | TIMESTAMP | NO |
| ADDRESS | VARCHAR | NO |
| ADDRESS2 | VARCHAR | YES |
| STATE | VARCHAR | NO |
| ZIP_CODE | INTEGER | NO |

**CUST_ORDER**

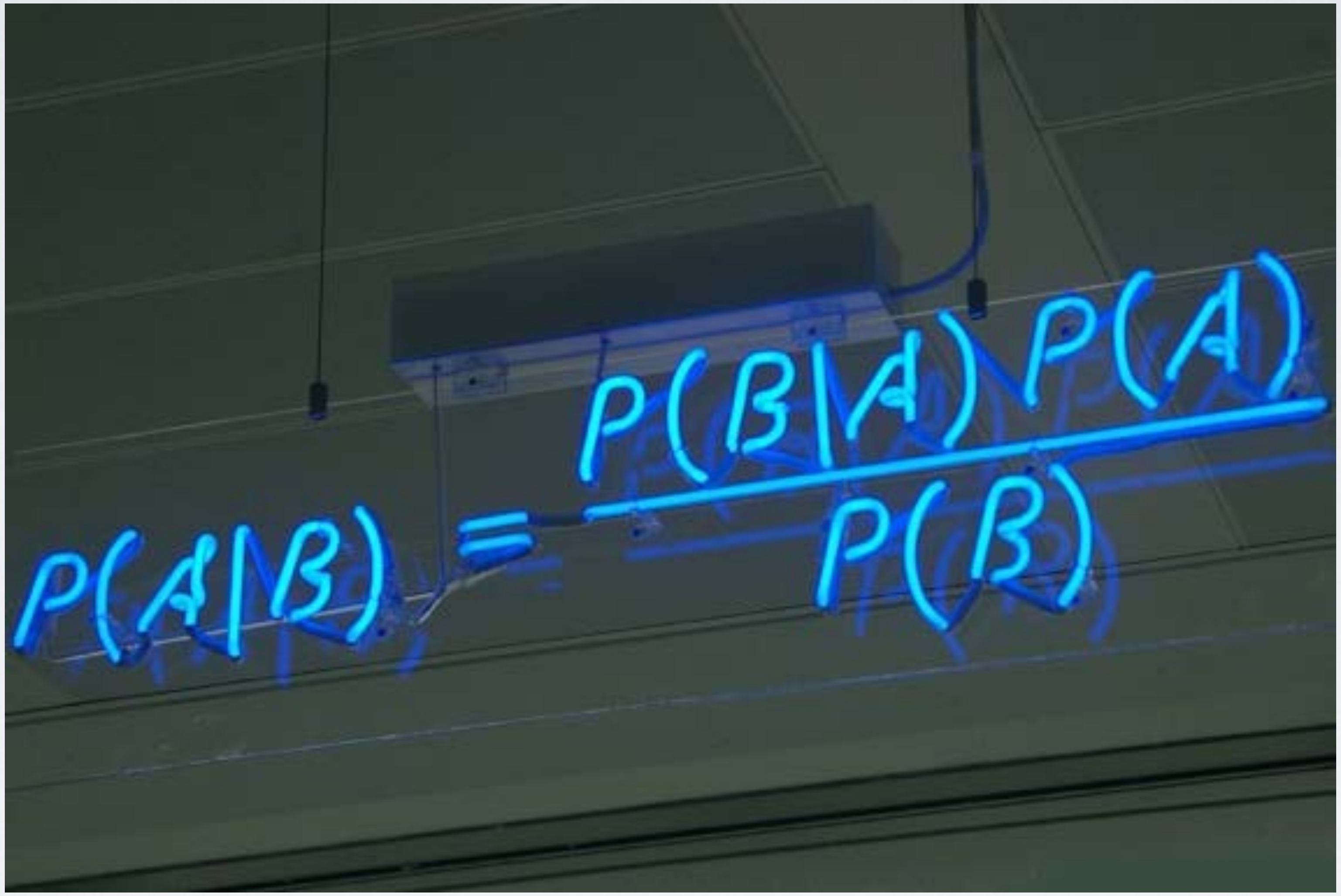
| NAME | DATATYPE | NULLABLE? |
|---|---|---|
| ORDER_ID | VARCHAR | NO |
| CUSTOMER_ID | VARCHAR | NO |
| STATUS | VARCHAR | NO |
| ORDER_AMOUNT | DECIMAL | NO |

**PRODUCT**

| NAME | DATATYPE | NULLABLE? |
|---|---|---|
| PRODUCT_ID | VARCHAR | NO |
| CATEGORY | VARCHAR | NO |
| LIST_PRICE | DECIMAL | NO |

# BAYES THEOREM

Just memorize this formula and understand its terms:

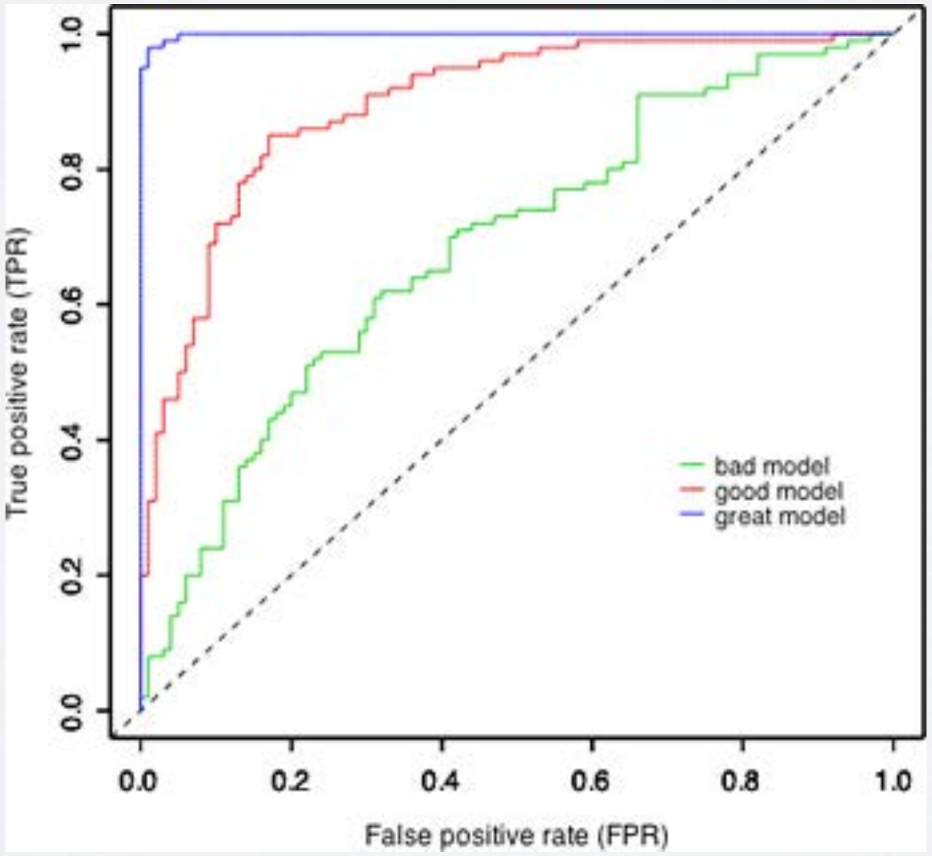


Sample problems at Glassdoor: https://www.glassdoor.com/Interview/data-scientist-interview-questions-SRCH_KO0,14.htm

# MACHINE LEARNING EVALUATION METRICS

Understand how to evaluate a model's performance:

- ROC curves
- cross-validation
- metrics for classification



# STUFF YOU PROBABLY DON'T NEED TO WORRY ABOUT (YET)

- Deep Learning (with exceptions for images/sound)
- Spark/Hadoop (most companies don't have necessary scale)
- Recommender systems (most companies don't need them)
- Advanced Natural Language Processing (know the basics)

# ALTERNATIVE FACTS OF DATA SCIENCE

# ALTERNATIVE FACT #1

Most of Data Science is fine-tuning models to get the highest performance possible

## REALITY:

You are going to spend most of your time cleaning/merging data

# ALTERNATIVE FACT #2

Big Data is EVERYWHERE! You will need Hadoop and Spark all the time to solve every problem!

## REALITY:

With exceptions, most problems can be handled on a single machine

# ALTERNATIVE FACT #3

Deep Learning solves EVERYTHING! Other methods are obsolete.

## REALITY:

You probably don't need it, unless you are working with images and want to maximize performance

# AUDIENCE QUESTIONS

# AUDIENCE QUESTIONS

How can a college fresher (say, studying in sophomore or final year) become a data scientist? What projects can they do? What skills should they focus on? How to start applying for jobs?

# AUDIENCE QUESTIONS

How can an experienced professional make a career shift into data science? Let's say, someone has 3 years of experience in Java, and they now want to become a data scientist. Or, let's say, someone knows Hive, Pig, Flume, Hadoop, what could be a natural career progression for them?

# AUDIENCE QUESTIONS

How is a Machine Learning Engineer different from a Data Scientist ?

# AUDIENCE QUESTIONS

How is a Statistician different from a Data Scientist?

# AUDIENCE QUESTIONS

How is a Data Engineer different from a Data Scientist ?

# AUDIENCE QUESTIONS

What is the relationship between data science and machine learning?

# AUDIENCE QUESTIONS

What are the most commonly used ML algorithms in industry today, so that students can master them first ?

# AUDIENCE QUESTIONS

What is the future of the Data Scientist job?
Will it survive after 5 - 10 years or get automated?

# AUDIENCE QUESTIONS

Can data science be used in building geological applications? If yes, what would be the starting point?

# GOOD LUCK!

@jmsteinw

jmsteinw@gmail.com

jessesw.com