DATA SCIENCE INTRODUCTION

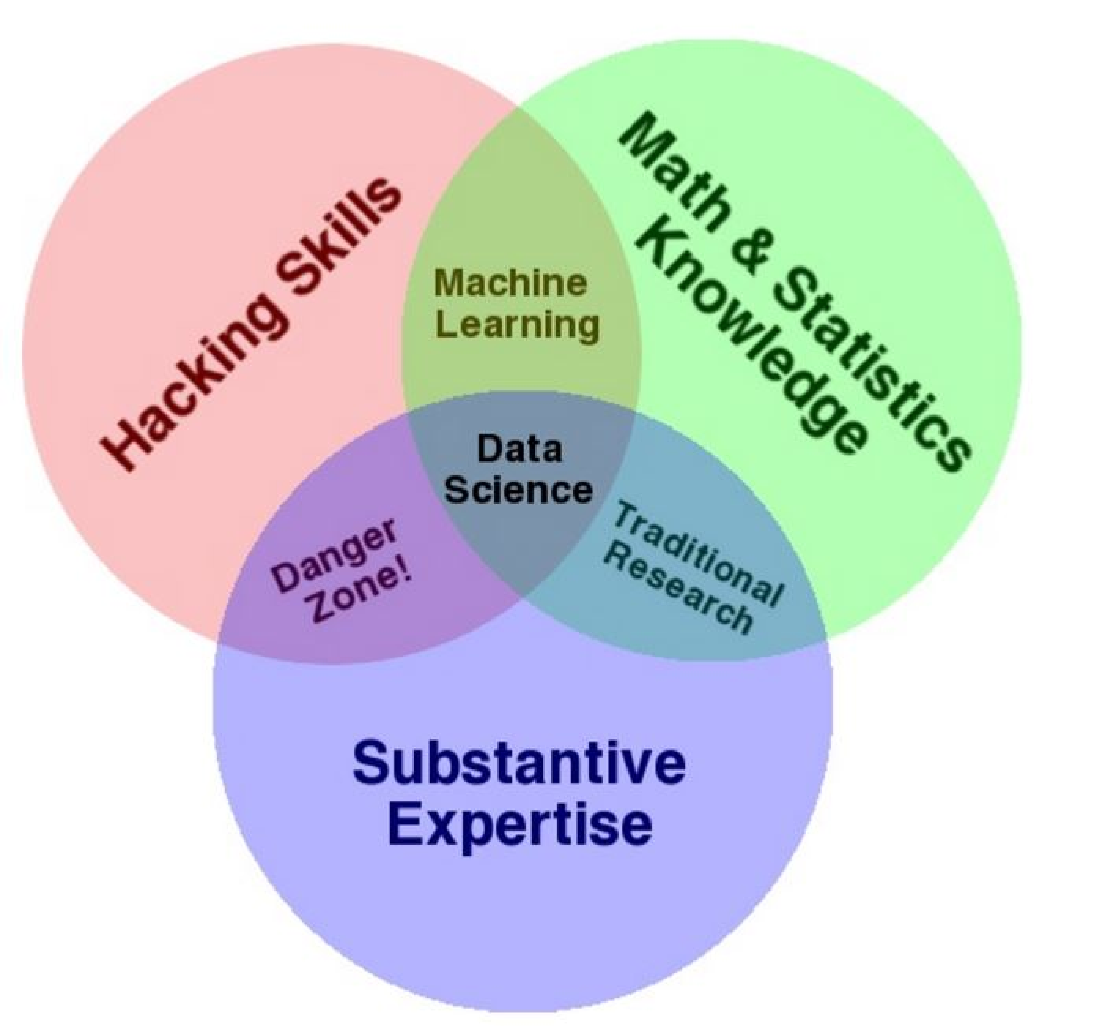
Data Science Introduction

In this session

- Venn diagram of data science
- What is data science?
- Data scientist
- Glassdoor best job in 2016 - 2017
- Data science job trend
- Data science job
- Data Scientist education levels
- Data science backgrounds
- Key topic to learn
- Learn Python library stack
- Go kaggle
- Get your degree
- Investigate the team
- Interview question type
- Take-home machine learning task
- Whiteboard coding
- Whiteboard SQL
- Bayes’ theorem
- Machine learning evaluation metrics
- Data Science job facts
- DS compared to ML engineer
- More information on Data Science

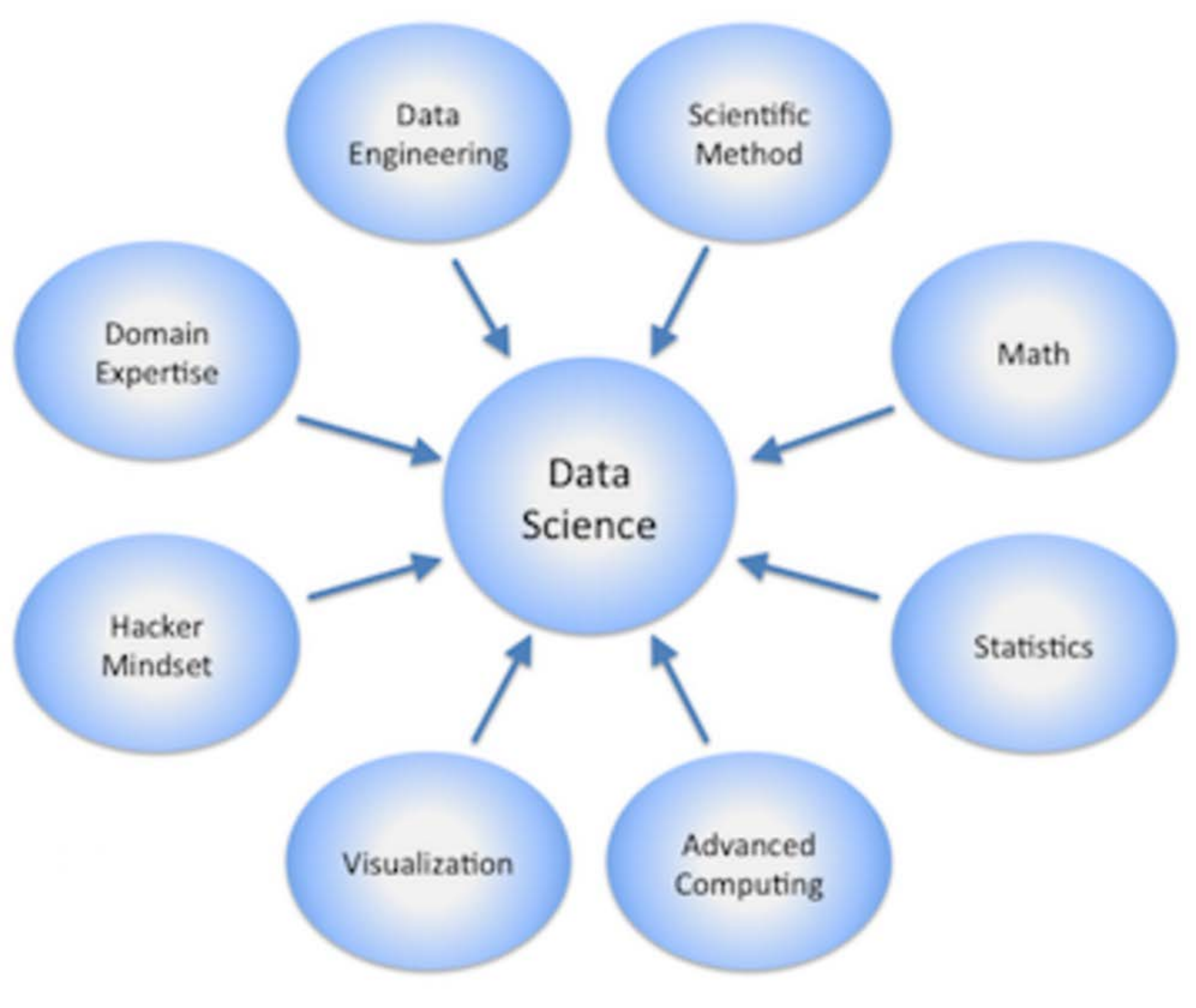
Data Science Introduction

What is Data Science



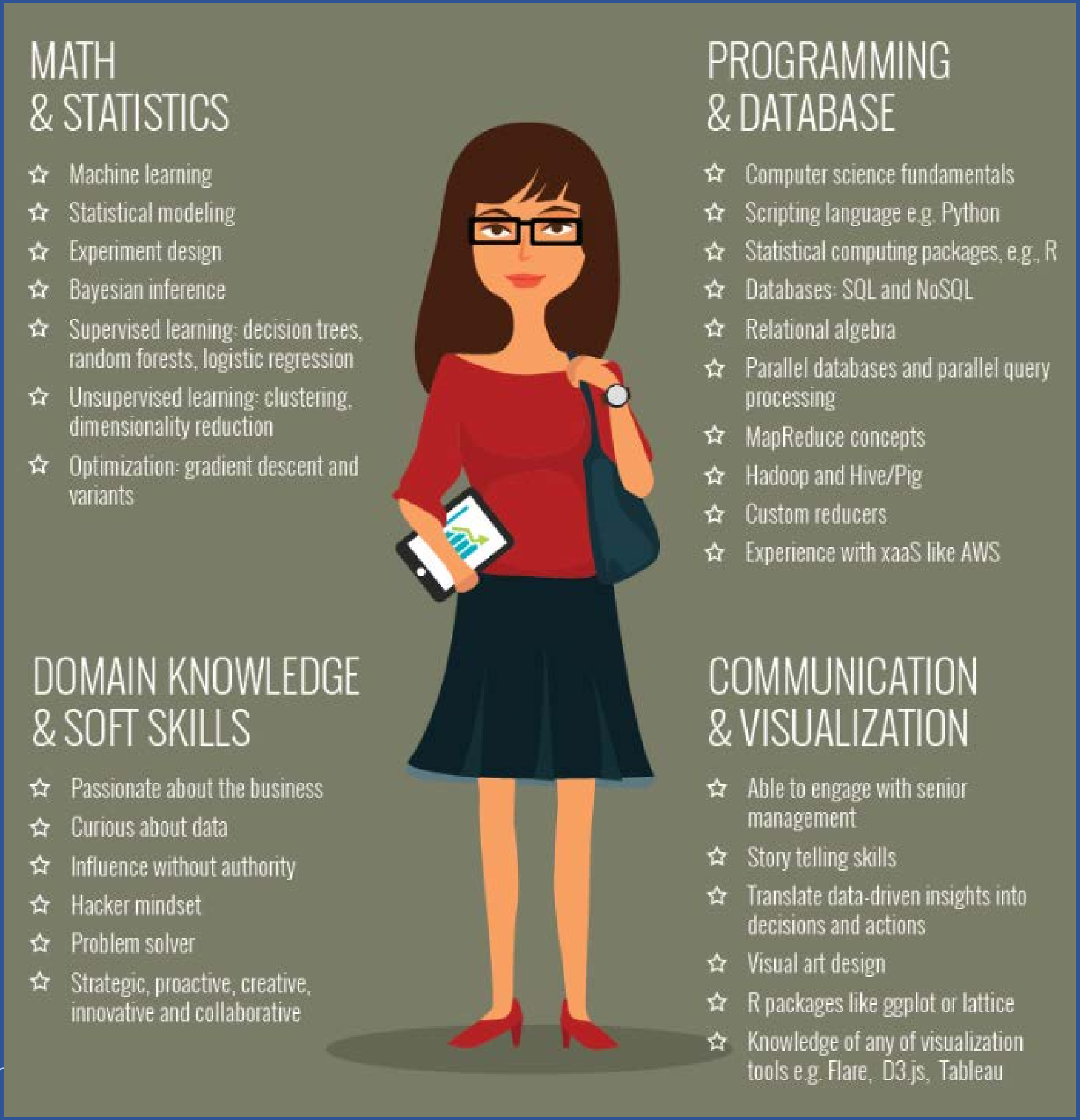
Data Science Introduction

What is data science?



Data Science Introduction

Data scientist



MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

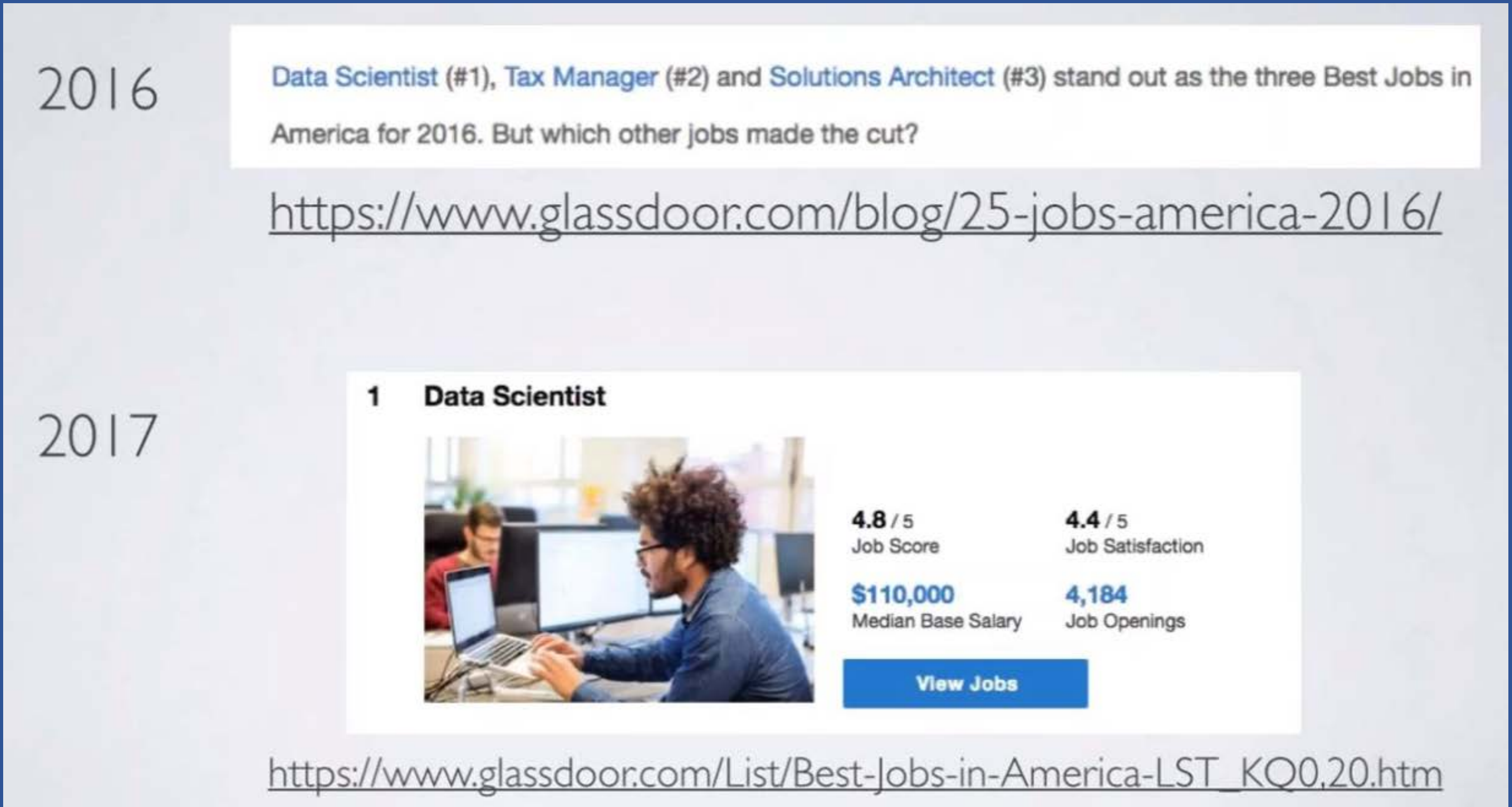
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Science Introduction

Glassdoor best job in 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

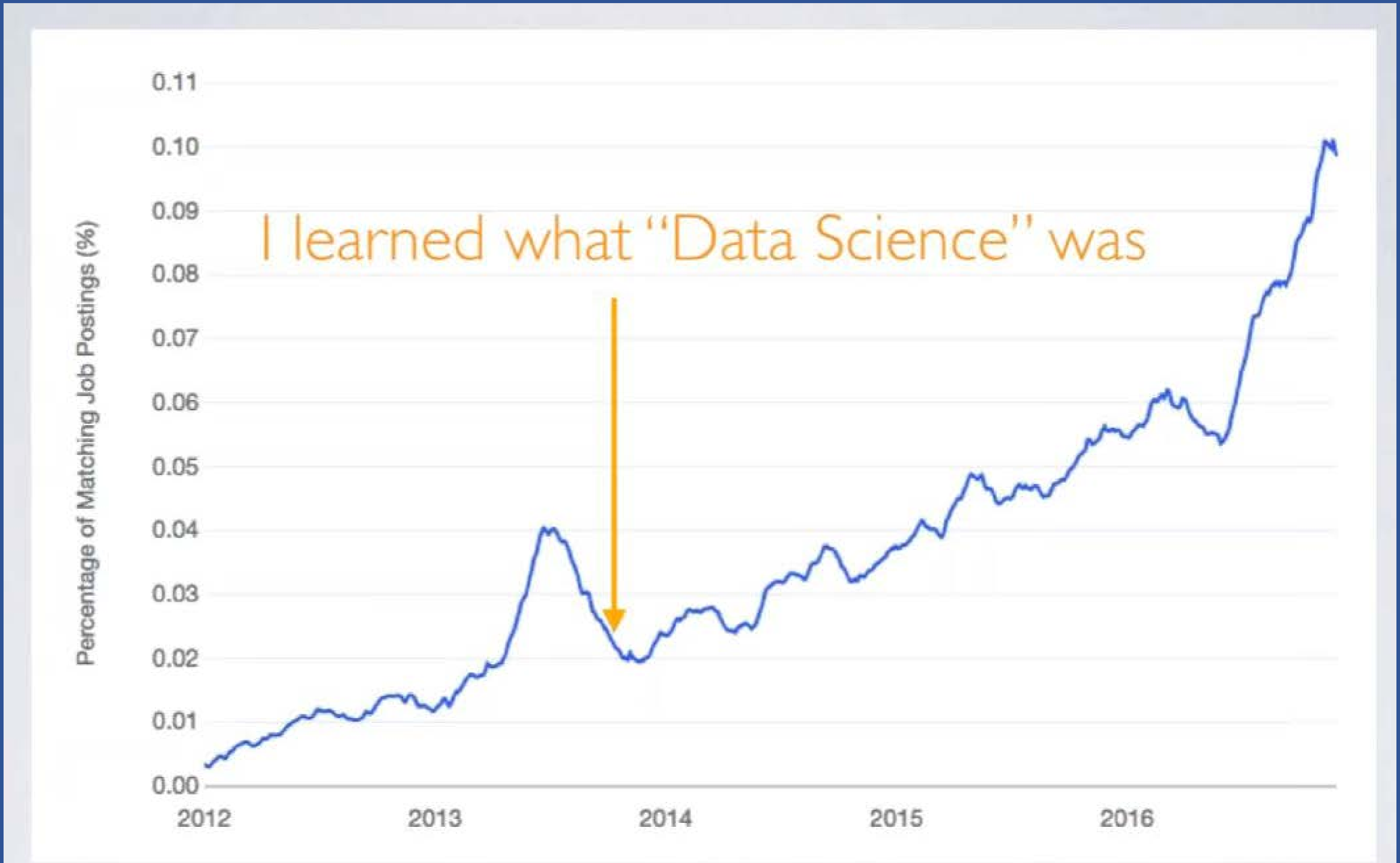
$110,000 Median Base Salary

4,184 Job Openings

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

Data Science Introduction

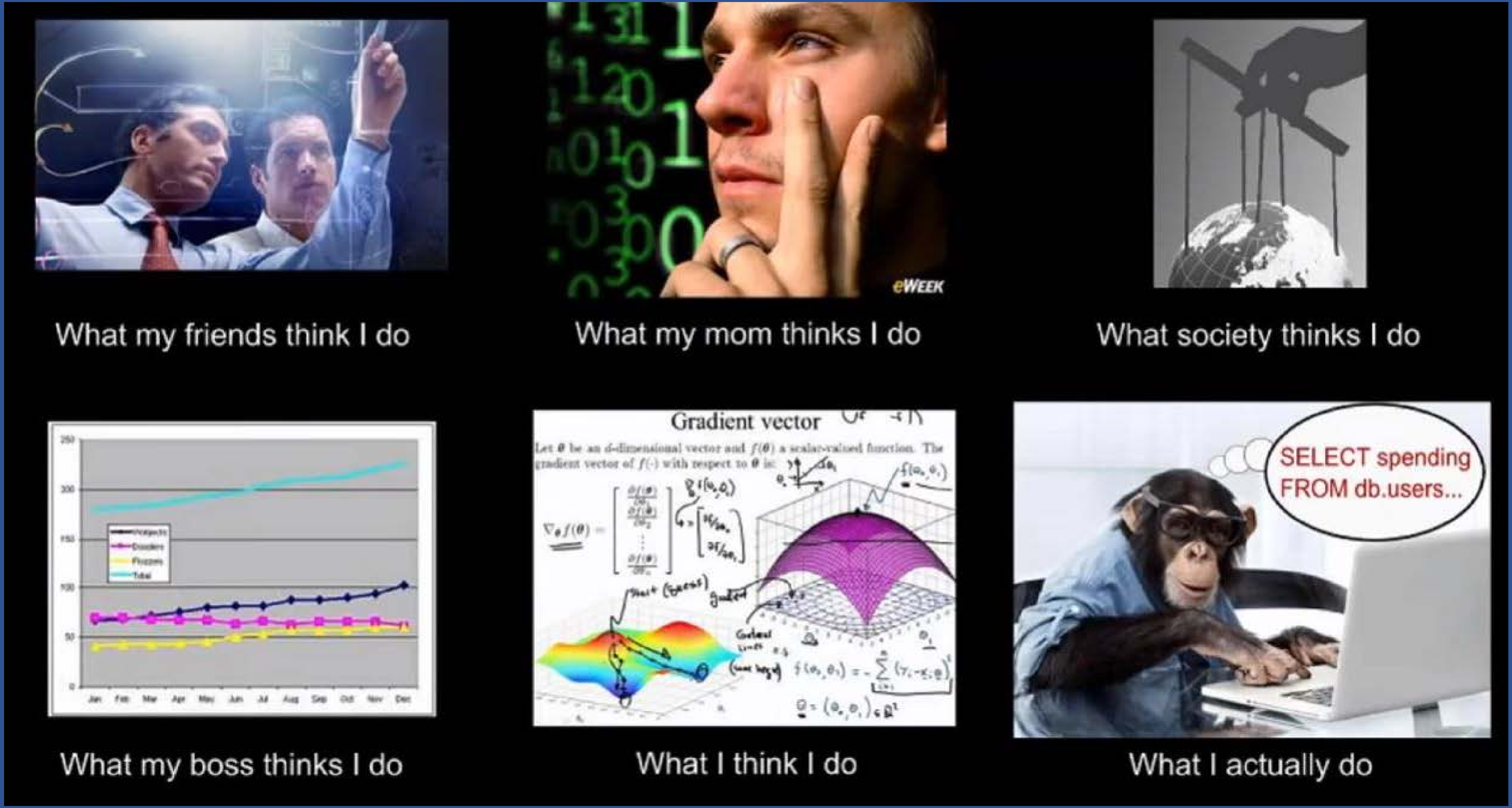
Data science job trend



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

Data Science Introduction

Data science job



What my friends think I do

What my mom thinks I do

What society thinks I do

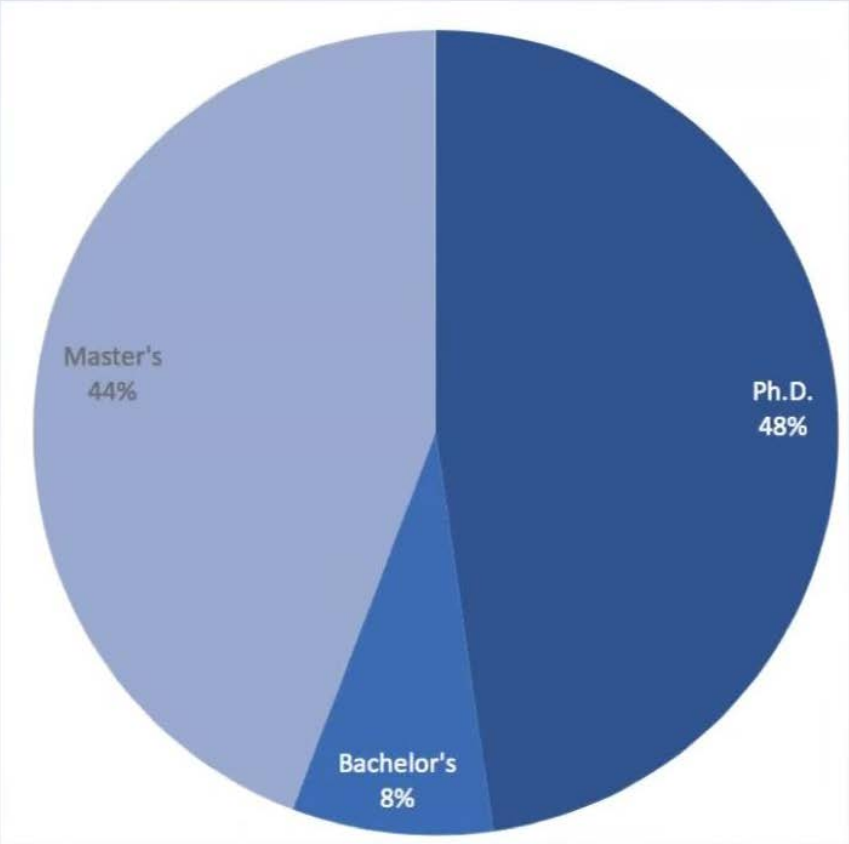
What my boss thinks I do

What I think I do

What I actually do

Data Science Introduction

Data Scientist education levels

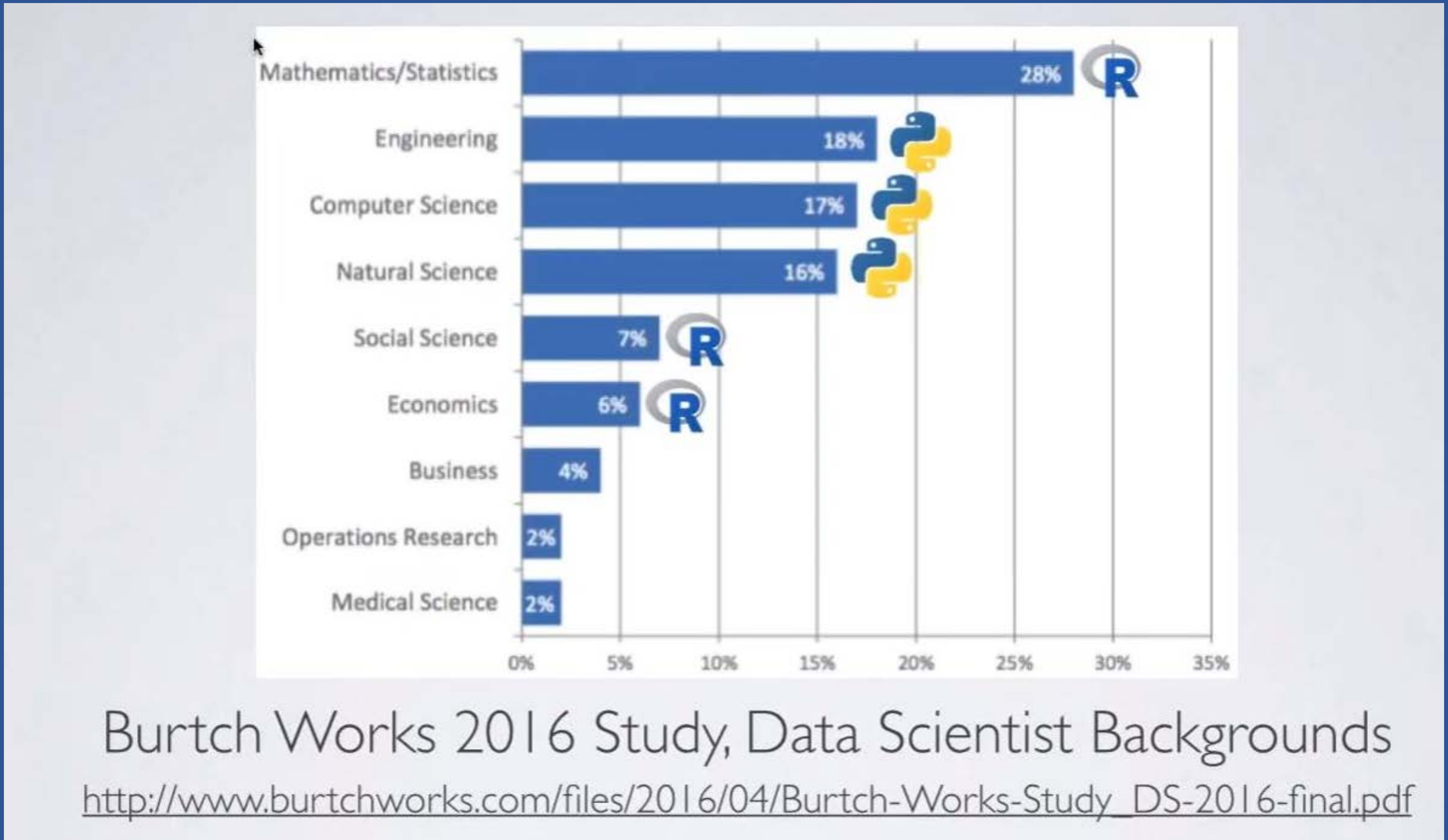


Burtch Works 2016 Study, Data Scientist Education Levels

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

Data Science Introduction

Data science backgrounds

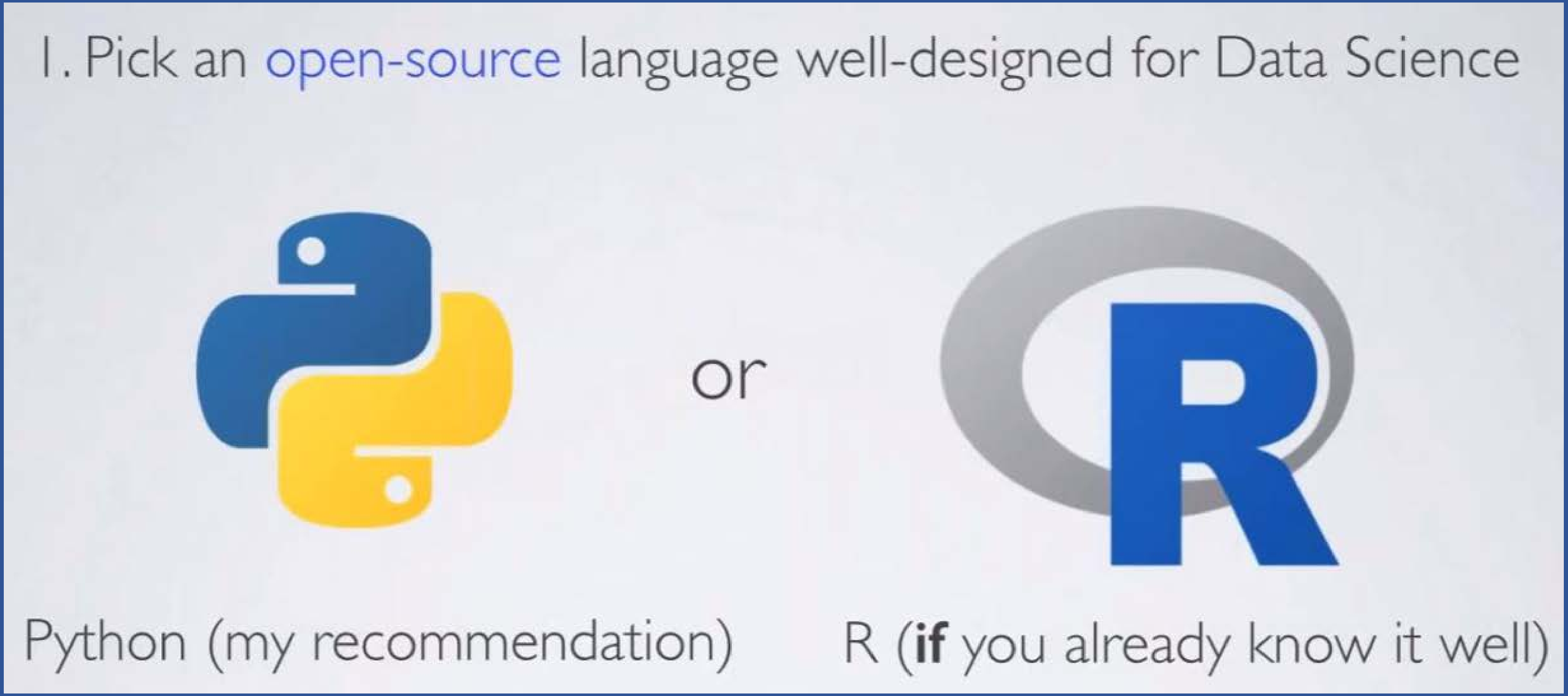


Burtch Works 2016 Study, Data Scientist Backgrounds

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

Data Science Introduction

Key topic to learn



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

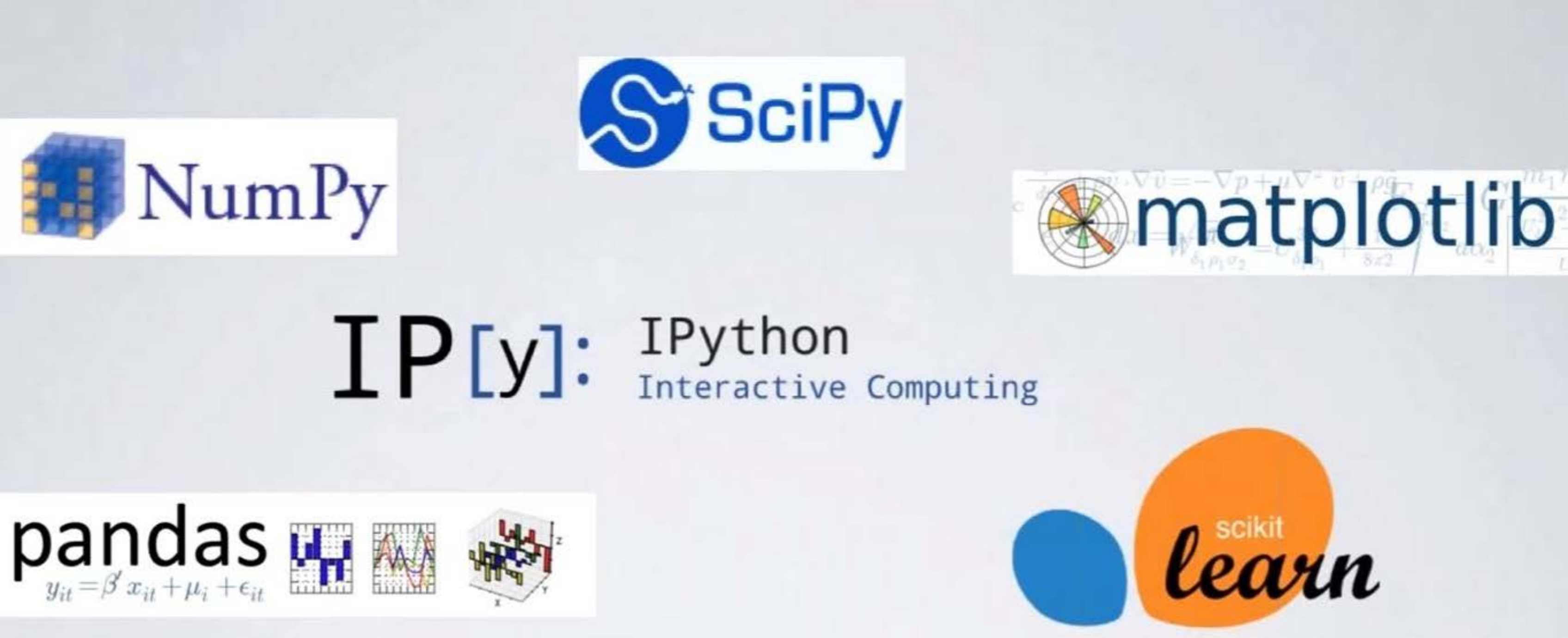
or

Python (my recommendation)

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

Data Science Introduction

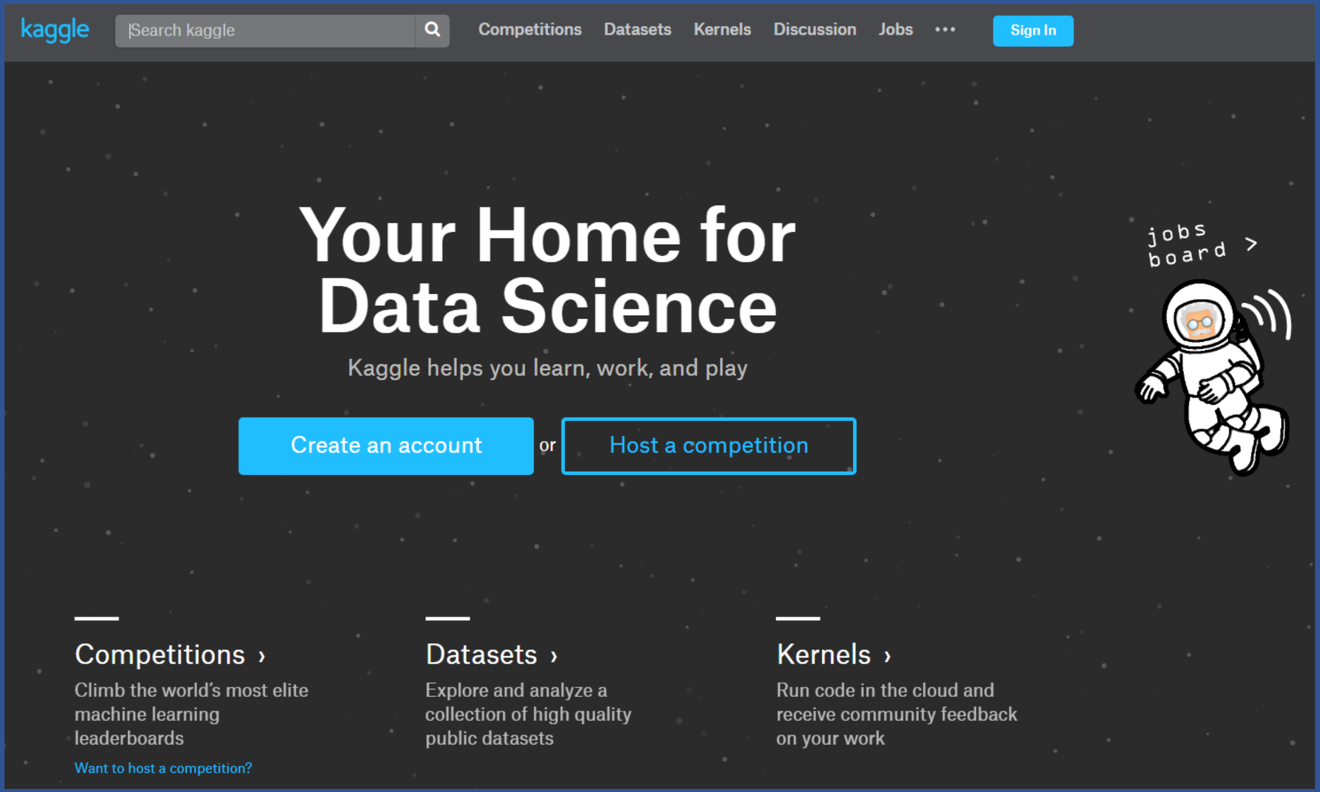
Learn Python library stack



Data Science Introduction

Go kaggle

- There are countless strategies that can be applied to any predictive modelling
- It is impossible to know at the outset which technique or analyst will be most effective
- Compete to produce the best models



Your Home for Data Science

Kaggle helps you learn, work, and play

Competitions ›

Climb the world's most elite machine learning leaderboards

Want to host a competition?

Datasets ›

Explore and analyze a collection of high quality public datasets

Kernels ›

Run code in the cloud and receive community feedback on your work

Data Science Introduction

Get your degree

Launch Your Career in Data Science

A nine-course introduction to data science, developed and taught by leading professors.

Johns Hopkins University (commonly referred to as **Johns Hopkins**, **JHU**, or simply **Hopkins**) is an American private research university in Baltimore, Maryland. Founded in 1876,

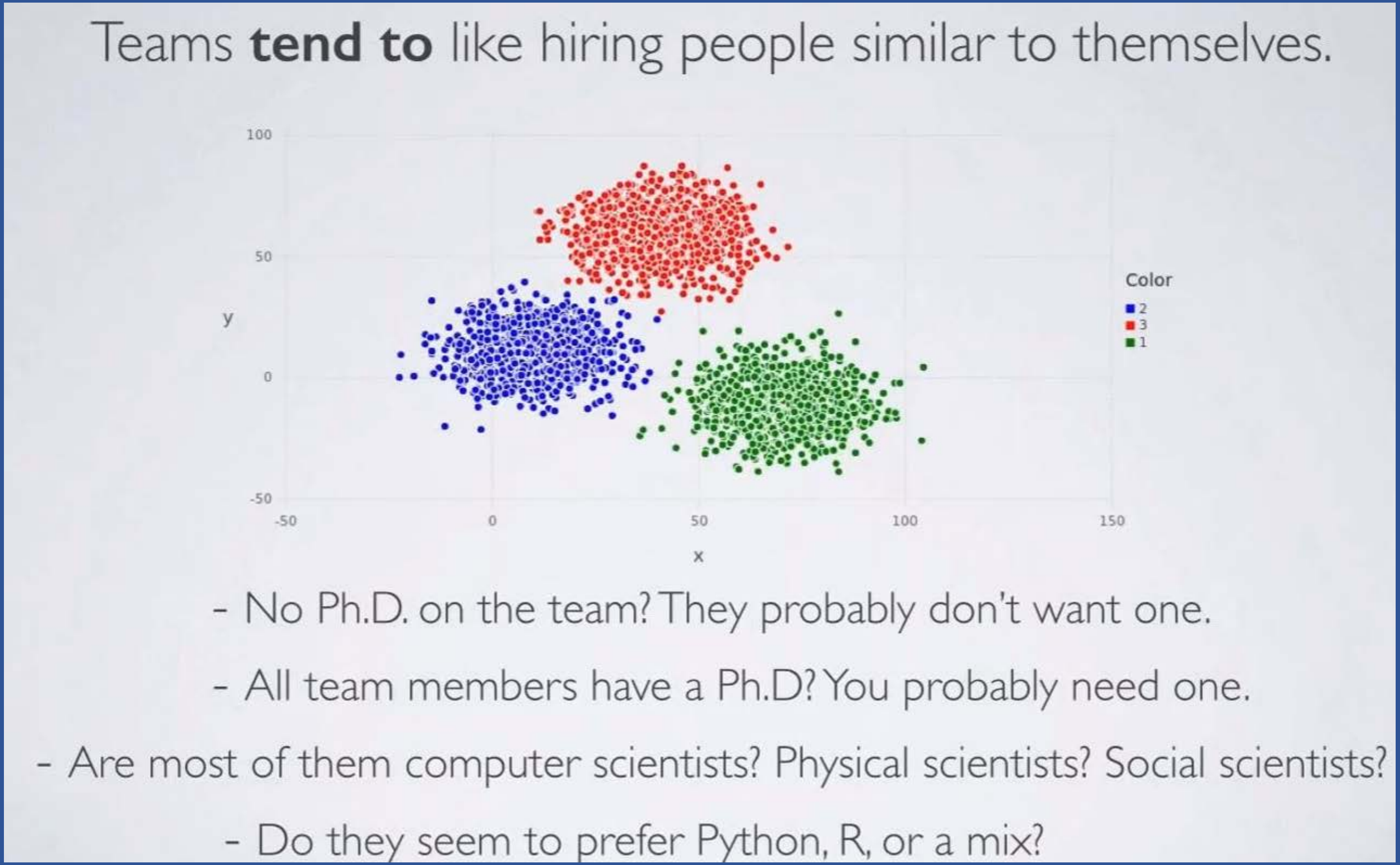
Ask the right questions, manipulate data sets, and create visualizations to communicate results.

This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.

Created by: JOHNS HOPKINS UNIVERSITY

Data Science Introduction

Investigate the teams

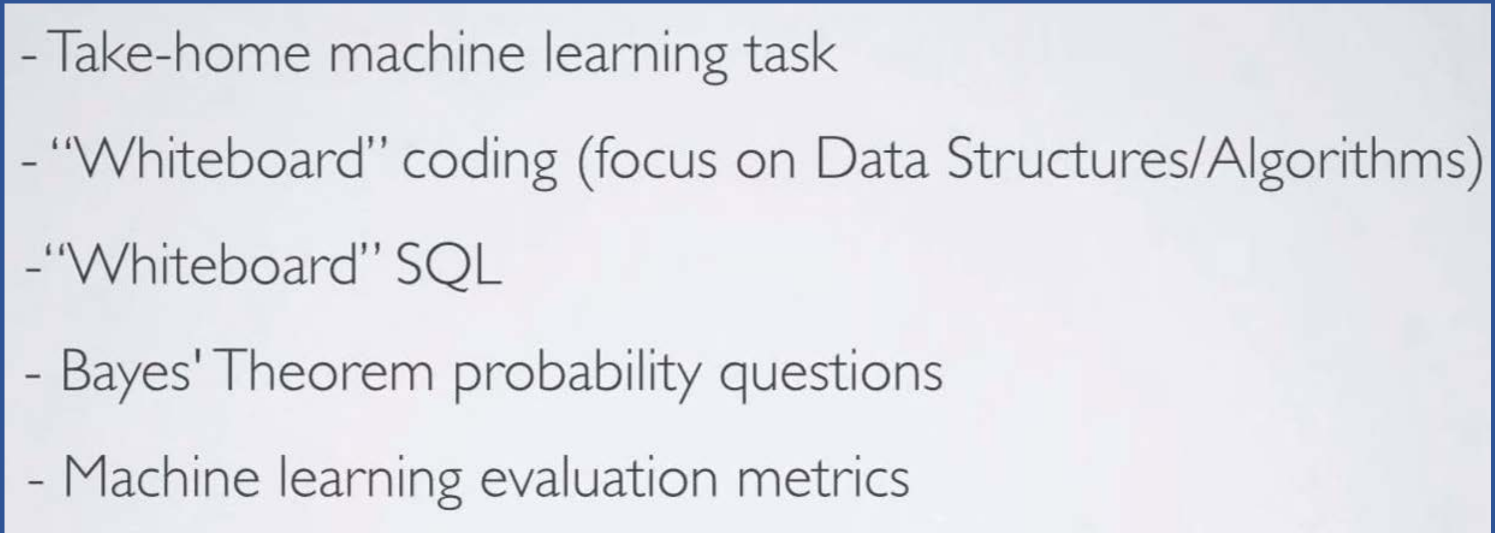


Teams **tend to** like hiring people similar to themselves.

- 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?

Data Science Introduction

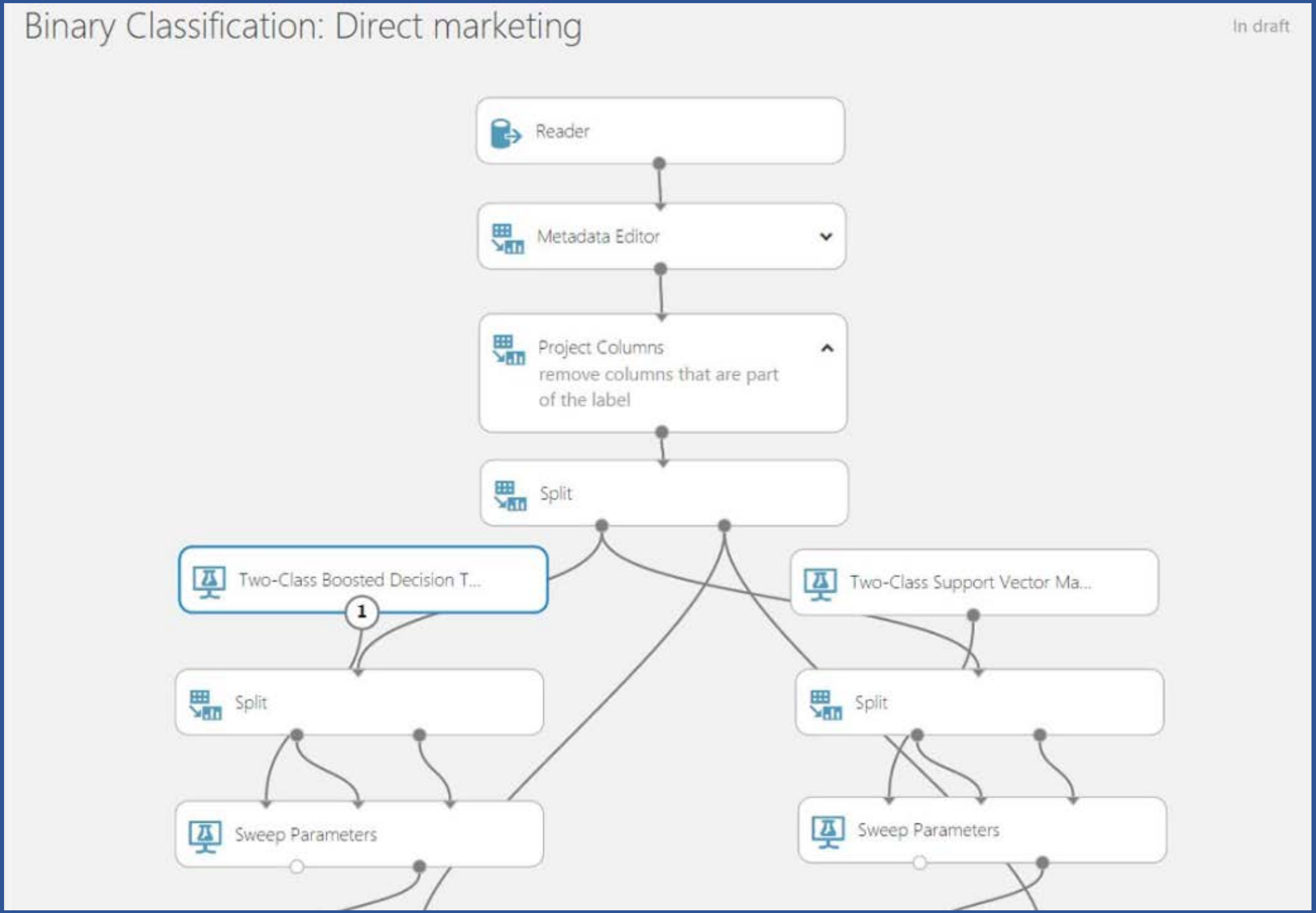
Interview question type



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

Data Science Introduction

Take-home machine learning task

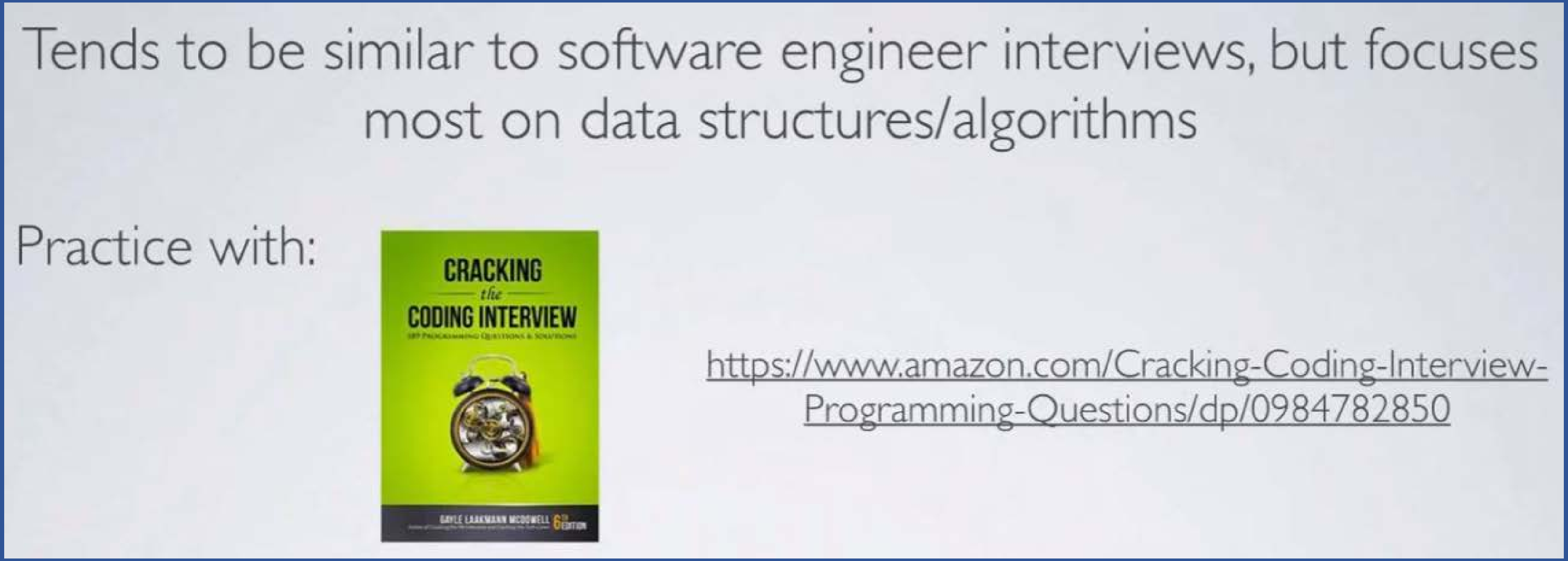


Binary Classification: Direct marketing

In draft
Reader
Metadata Editor
Project Columns
remove columns that are part of the label
Split
Two-Class Boosted Decision T...
1
Two-Class Support Vector Ma...
Split
Split
Sweep Parameters
Sweep Parameters

Data Science Introduction

Whiteboard coding



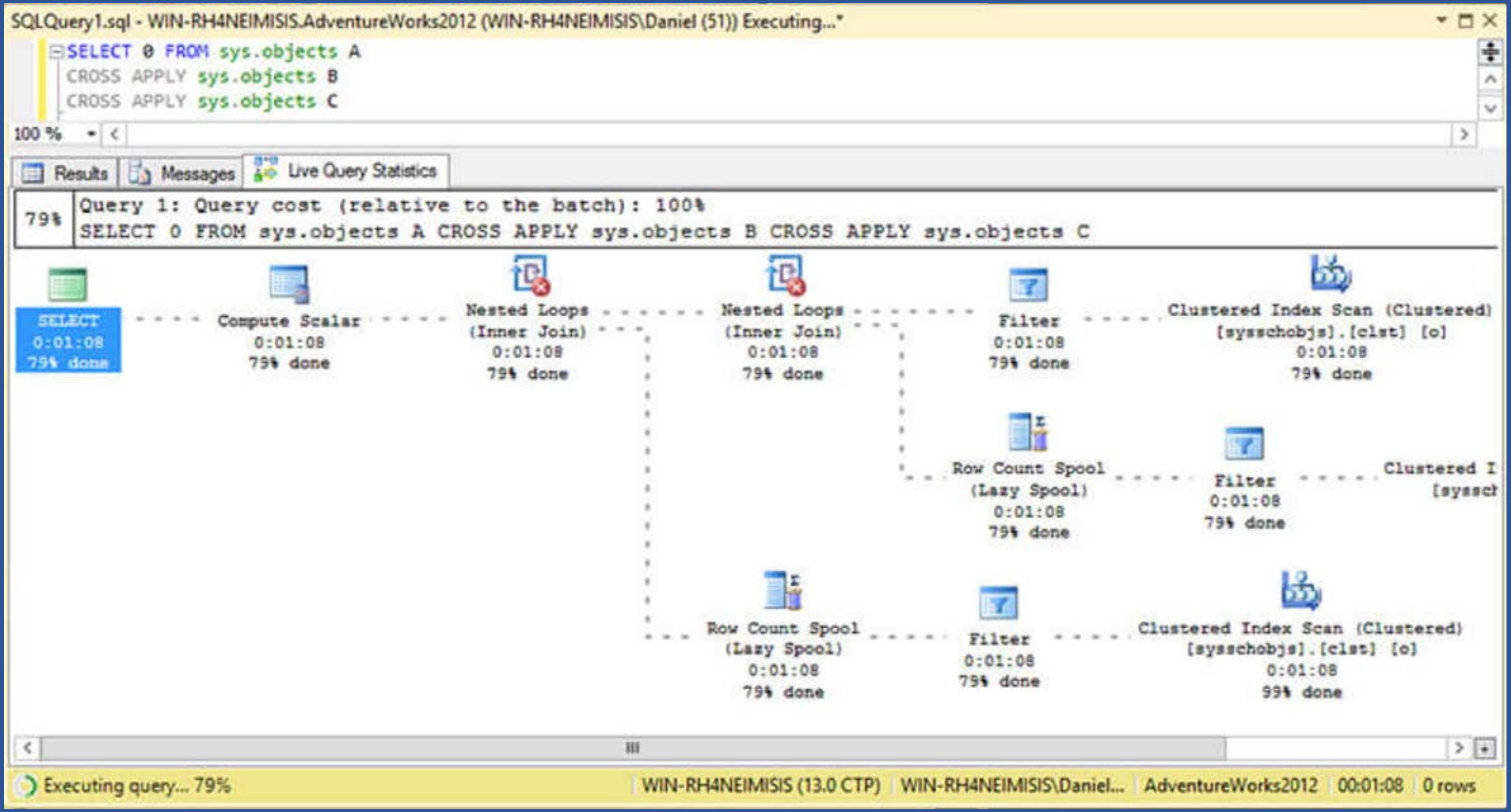
Tends to be similar to software engineer interviews, but focuses most on data structures/algorithms

Practice with:

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

Data Science Introduction

Whiteboard SQL



SQLQuery1.sql - WIN-RH4NEIMISIS.AdventureWorks2012 (WIN-RH4NEIMISIS\Daniel (51)) Executing...*

```
SELECT 0 FROM sys.objects A
CROSS APPLY sys.objects B
CROSS APPLY sys.objects C
```

```
Query 1: Query cost (relative to the batch): 100%
SELECT 0 FROM sys.objects A CROSS APPLY sys.objects B CROSS APPLY sys.objects C
```

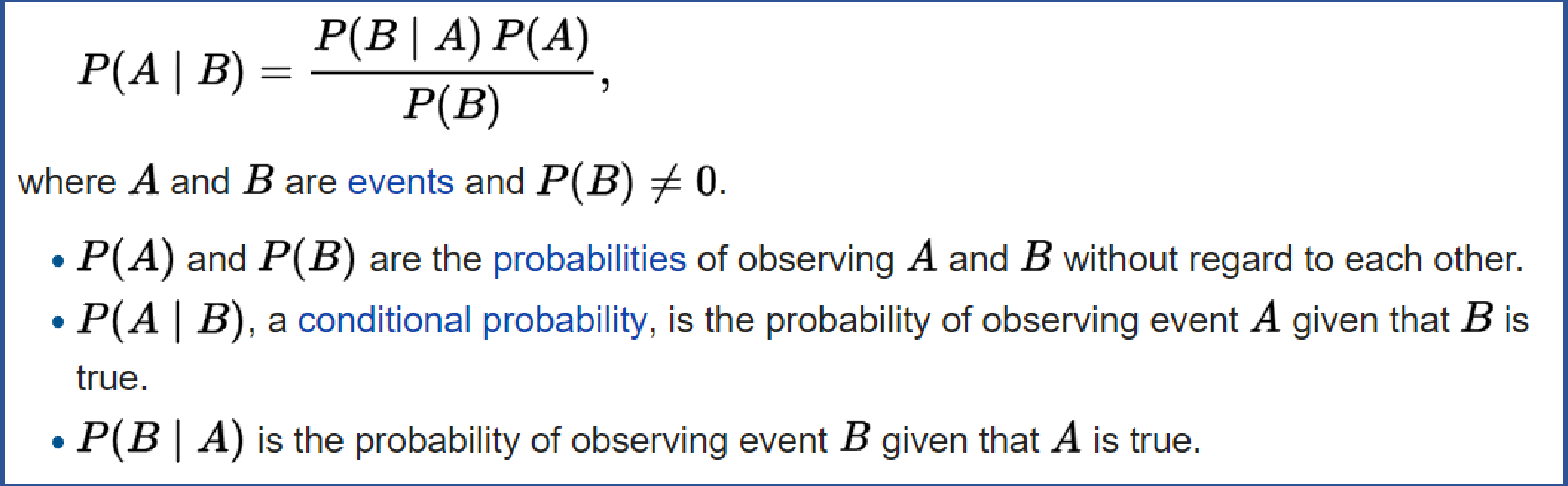
Executing query... 79% | WIN-RH4NEIMISIS (13.0 CTP) | WIN-RH4NEIMISIS\Daniel... | AdventureWorks2012 | 00:01:08 | 0 rows

100 %
Results
Messages
Live Query Statistics
79%
SELECT 0:01:08 79% done
Compute Scalar 0:01:08 79% done
Nested Loops (Inner Join) 0:01:08 79% done
Nested Loops (Inner Join) 0:01:08 79% done
Filter 0:01:08 79% done
Clustered Index Scan (Clustered) [sysschobjs].[clst] [o] 0:01:08 79% done
Row Count Spool (Lazy Spool) 0:01:08 79% done
Filter 0:01:08 79% done
Clustered I [sysscl
Row Count Spool (Lazy Spool) 0:01:08 79% done
Filter 0:01:08 79% done
Clustered Index Scan (Clustered) [sysschobjs].[clst] [o] 0:01:08 99% done

Data Science Introduction

Bayes’ theorem

- Memorize this formula



$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

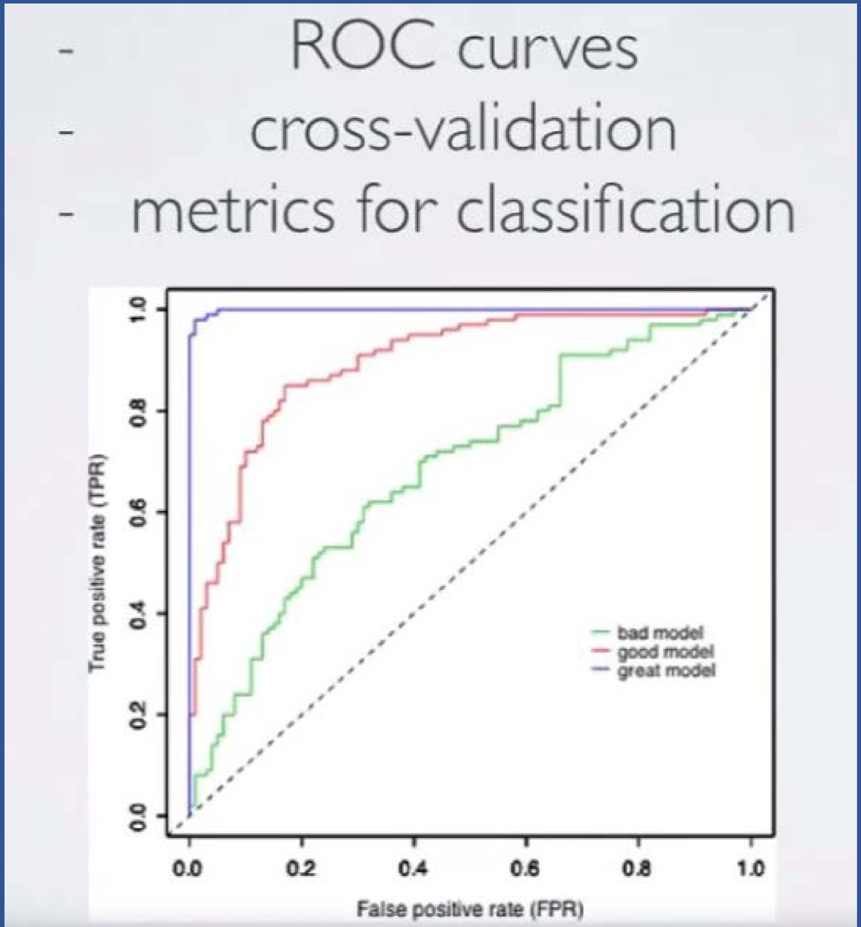
- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

- Understand terms

- Bayes’ theorem describes the probability of an event
- based on prior knowledge of conditions that might be related to the event
- For example, if cancer is related to age, age should be included as input parameter

Data Science Introduction

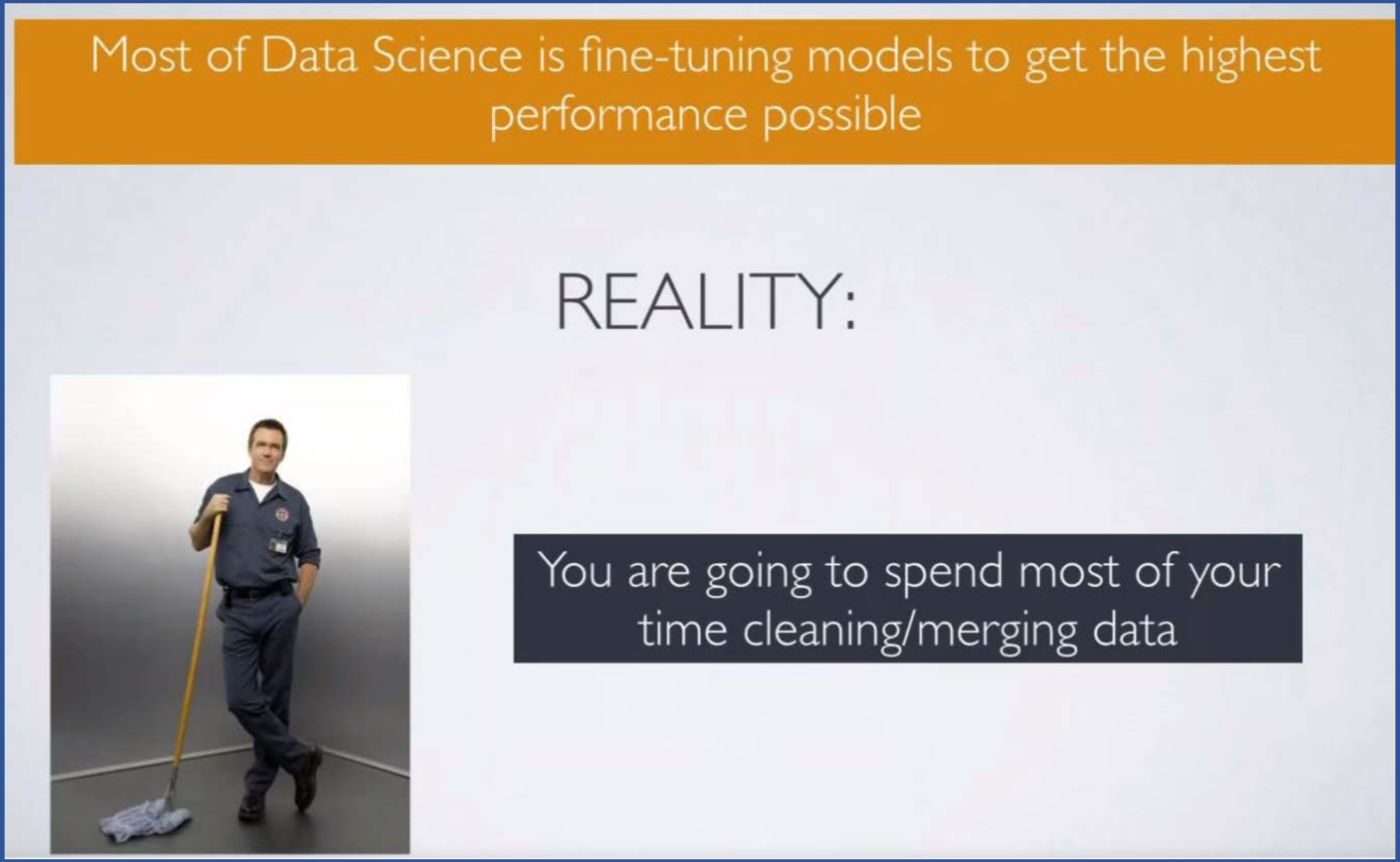
Machine learning evaluation metrics



- ROC curves
- cross-validation
- metrics for classification

Data Science Introduction

Data Science job fact #1



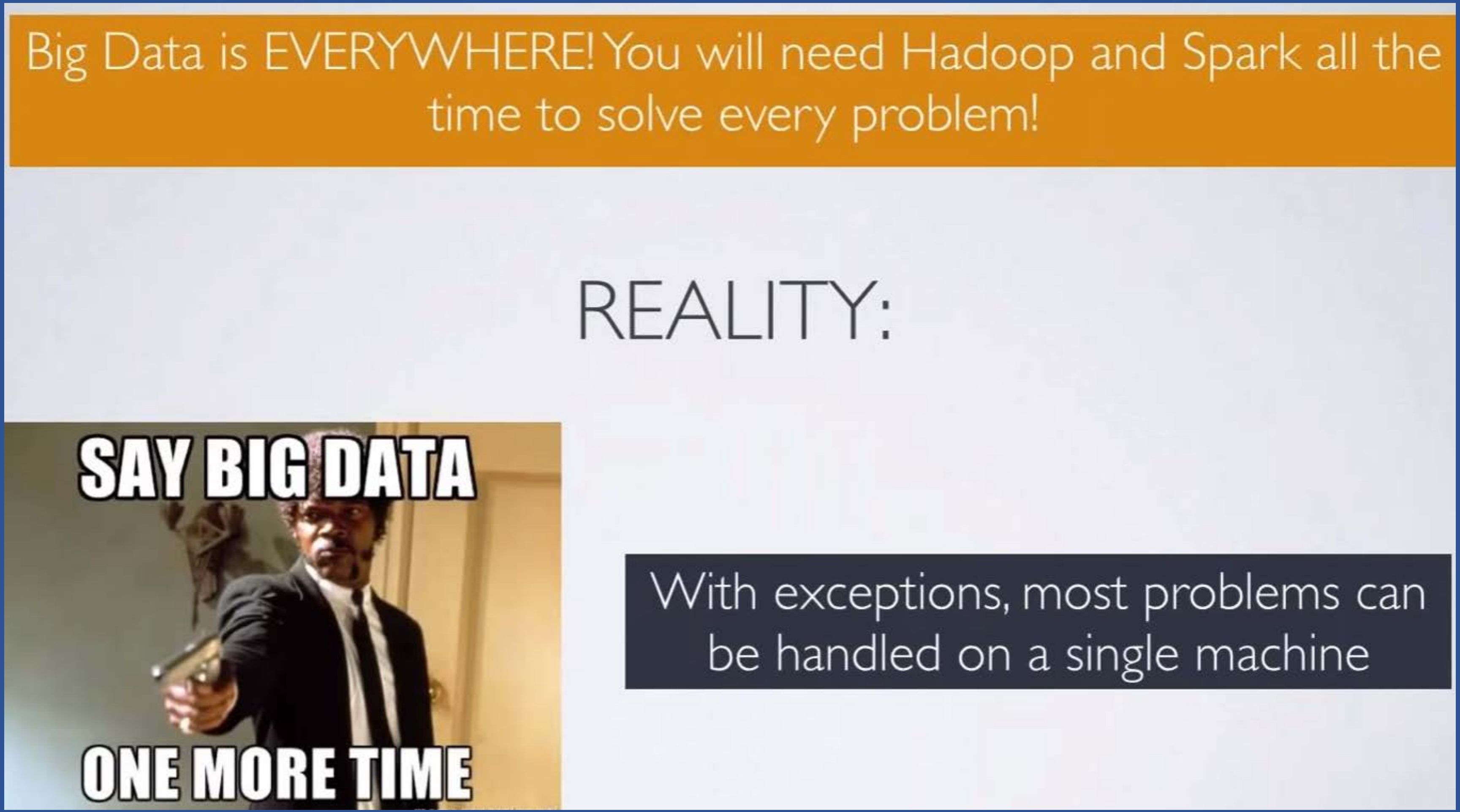
Most of Data Science is fine-tuning models to get the highest performance possible

REALITY:

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

Data Science Introduction

Data Science job fact #2



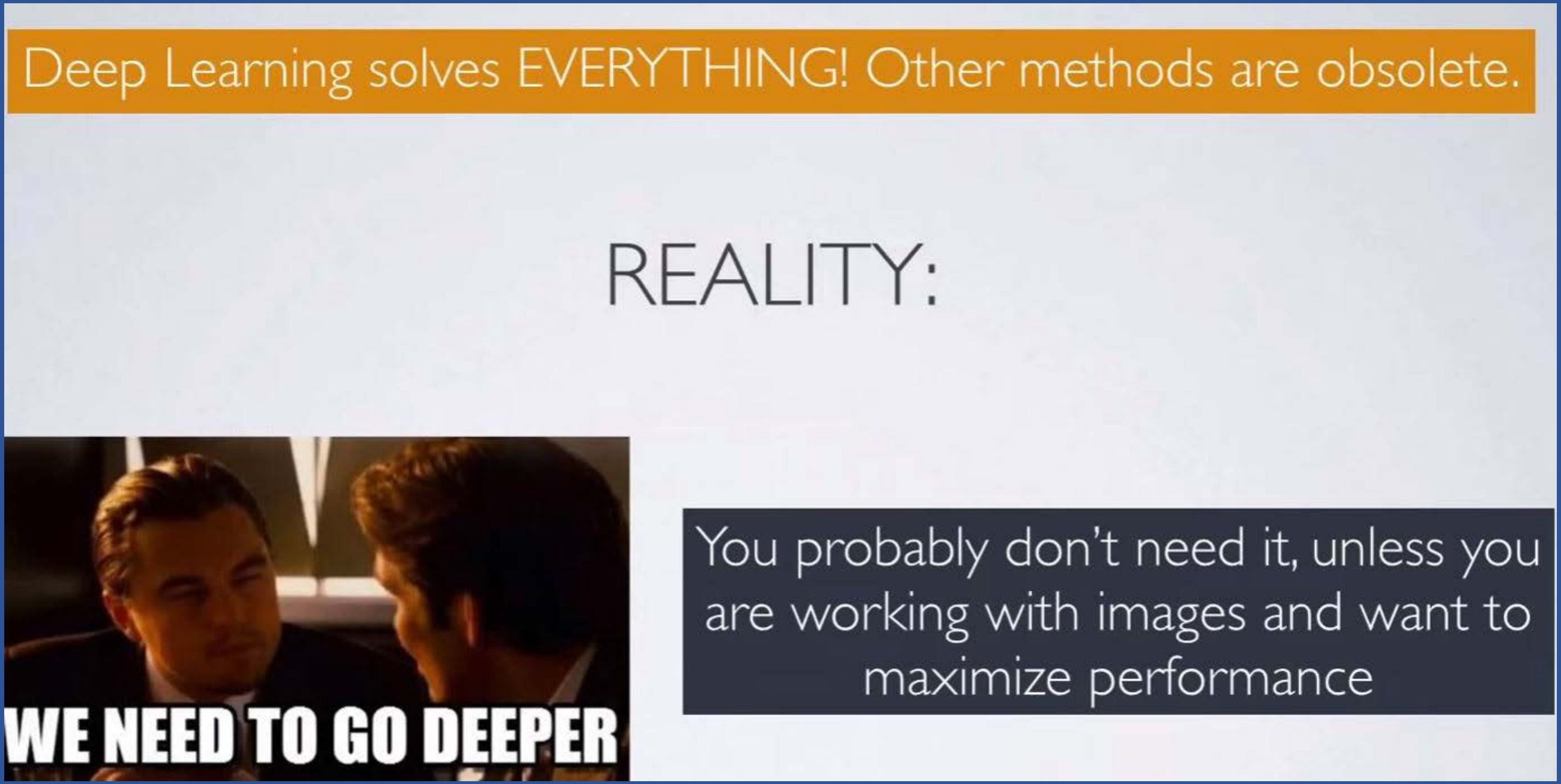
Big Data is EVERYWHERE! You will need Hadoop and Spark all the time to solve every problem!

REALITY:

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

Data Science Introduction

Data Science job 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

WE NEED TO GO DEEPER

Data Science Introduction

DS compared to ML engineer

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

Data Scientist

- Trained to be strong in Data
- R, Python, MATLAB
- Data treatment
- Evaluate ML algorithm
- Evaluate ML module

ML Engineer

- Trained to be strong in Coding
- C++, Java, C#
- Coding
- Change algorithm to code
- Create ML module

Data Science Introduction

More information

More information on Data Science

Doing Data Science by Cathy O'Neil, Rachel Schutt: Chapter 1. Introduction: What Is Data Science?
https://www.safaribooksonline.com/library/view/doing-data-science/9781449363871/ch01.html