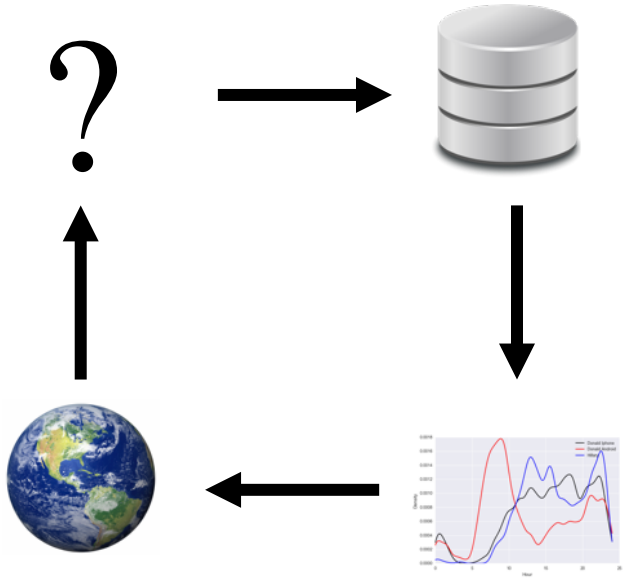
Data 100

Principles & Techniques of Data Science

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu



Questions for Today

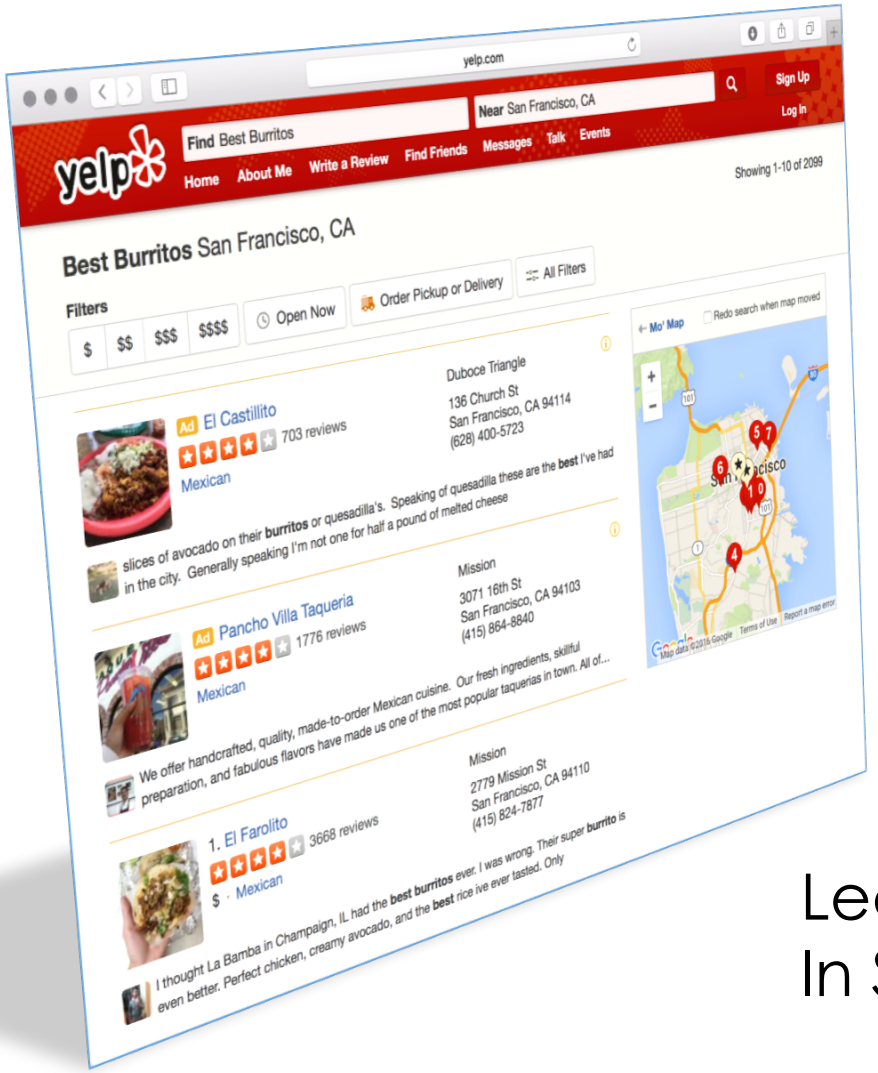
- ***Why*** *am I excited about Data Science?*
- **What** is Data Science?
- **Who** are we?
- **What** does it mean to be a data scientists today?
- **Break**
- **What** will I learn and **how?**
- **Demo (who are you?)!**

Slides from lecture available online at http://ds100.org/sp18

Why *am I excited about Data Science?*

Data is Changing the World

Where should I eat?



Where can I get the best burrito in SF?

Each ratings star added on a Yelp restaurant review translated to anywhere from a 5 percent to 9 percent effect on revenues.

-- Harvard Business School

Learn about eating the dangers of eating In SF in 2nd homework …

http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business

Data can help address climate change ...

DATA FOR CLIMATE ACTION

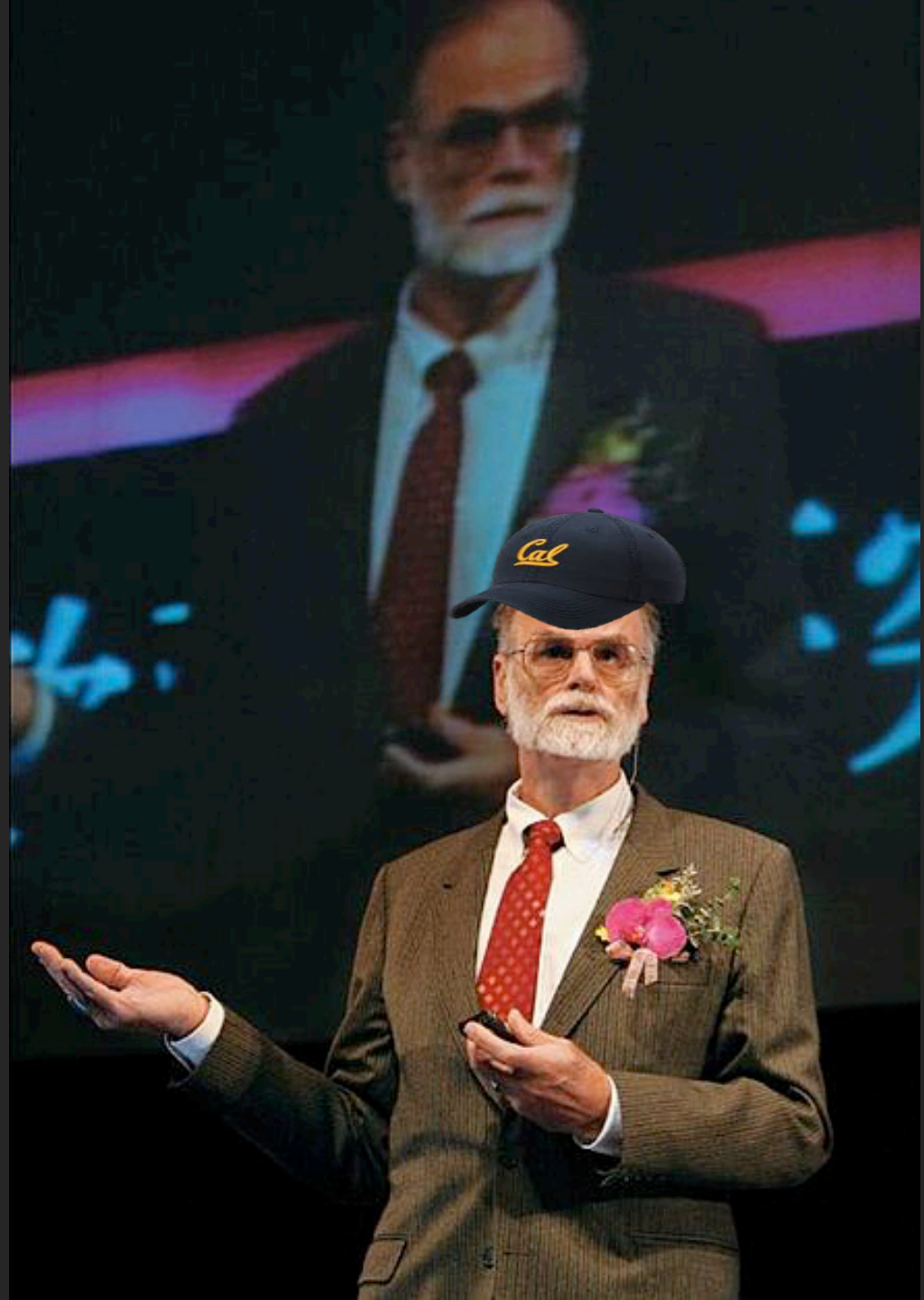
By tracking **sales data** on energy efficient appliances, data for climate action is helping **guide urban campaigns** to educate the general public and measure changes in purchasing behavior.

http://www.dataforclimateaction.org

Data Science is transfroming ***Science***

Jim Gray

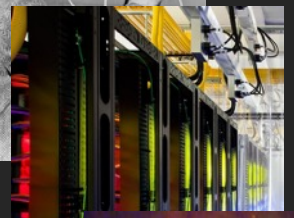
Turing Award Winning Computer Scientist & Cal Alum.



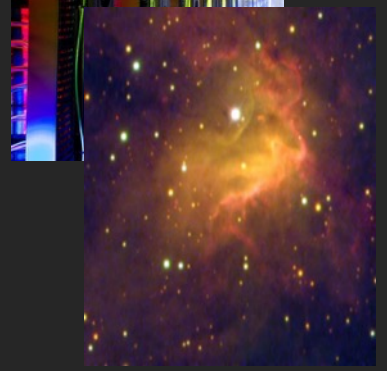
Introduced the idea of the **Fourth Paradigm** of Science

Experimental

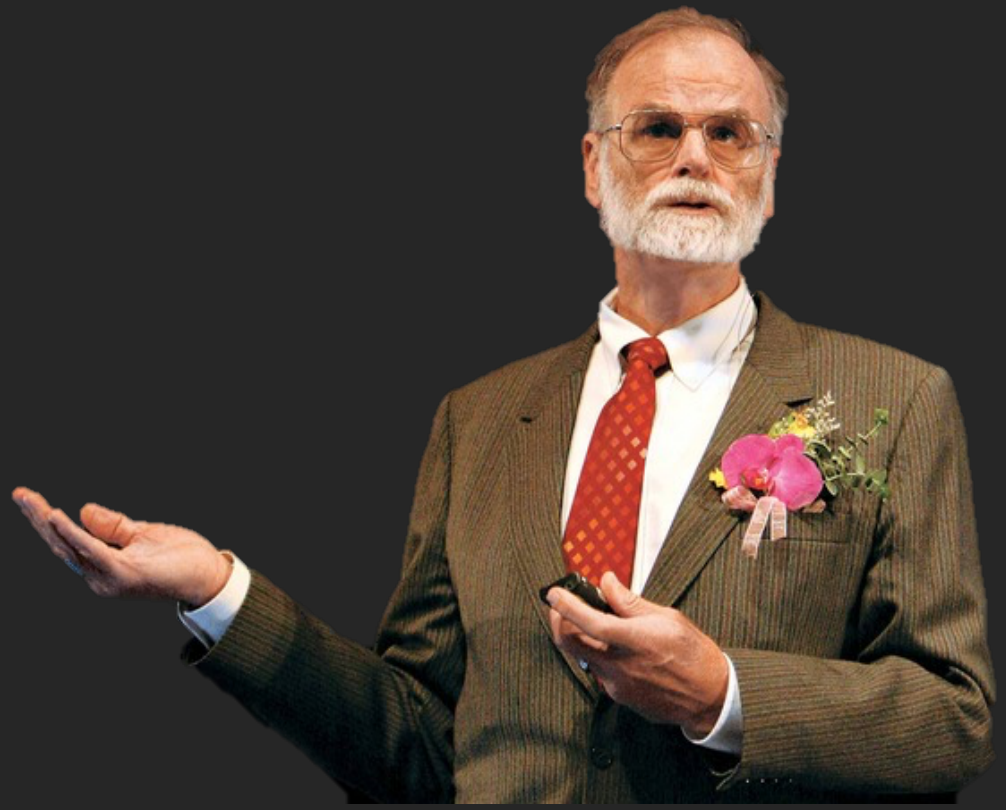
Theoretical



Simulation

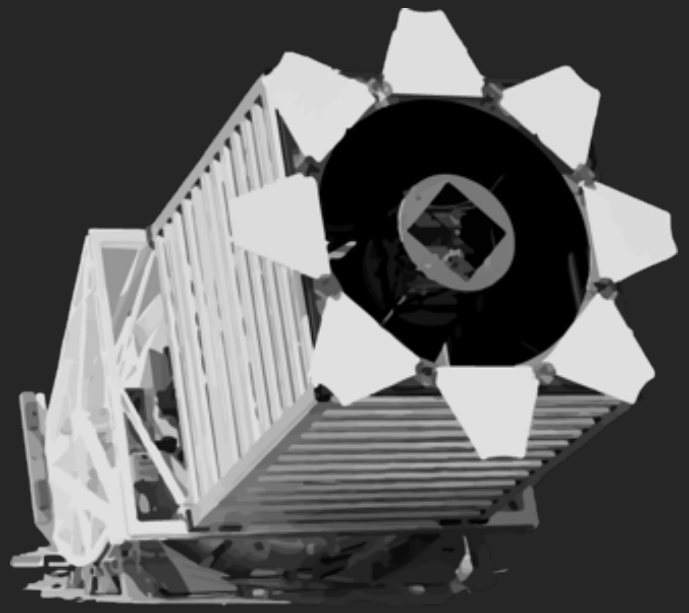


Data Intensive



Jim Gray

Astronomy in the 4th Paradigm



Sloan Digital
Sky Survey (SDSS)

+

Database
Systems

Sky
Server

Technology Trends

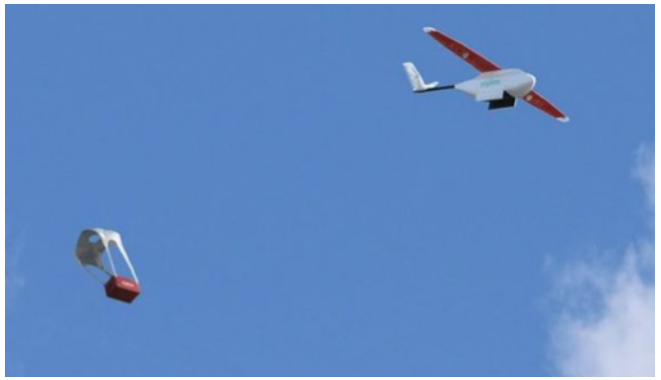
2020s ?

2010s Data Industry
- Collect and sell information

2000s Internet Industry
- Online retailers and services

1990s Software Industry
- Sold computer software

1980s Hardware Industry
- Sold computers



Real concern?

Slate

FUTURE TENSE | THE CITIZEN'S GUIDE TO THE FUTURE. | APRIL 28 2016 9:00 AM

FROM SLATE, NEW AMERICA, AND ASU

Killer Robots? Lost Jobs?

The threats that artificial intelligence researchers actually worry about.

364 239 28

14 On the Threat of Artificial Intelligence

Stephen Hawking

There are more **immediate concerns.**

The Off-Switch Game

Dylan Hadfield-Menell Anca Dragan Pieter Abbeel Stuart Russell
Department of Computer Science
University of California at Berkeley
Berkeley, CA 94709
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

Abstract

It is clear that one of the primary tools we can use to mitigate the potential risk from a misbehaving AI system is the ability to turn the system off. As the capabilities of AI systems improve, it is important to ensure that such systems do not adopt subgoals that prevent a human from switching them off. This is a challenge because many formulations of rational agents create strong incentives for self-preservation. This is not caused by a built-in instinct, but because a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead. Our goal is to study the incentives an agent has to allow itself to be switched off. We analyze a simple game between a human H and a robot R, where H can press R's off switch but R can disable the off switch. A traditional agent takes its reward function for granted: we show that such agents have an incentive to disable the off switch, except in the special case where H is perfectly rational. Our key insight is that for R to want to preserve its off switch, it needs to be uncertain

home › tech | US politics world opinion sports soccer arts lifestyle fas | all

Artificial intelligence (AI)

The rise of robots: forget evil AI - the real risk is far more insidious

It's far more likely that robots would inadvertently harm or frustrate humans while carrying out our orders than they would rise up against us

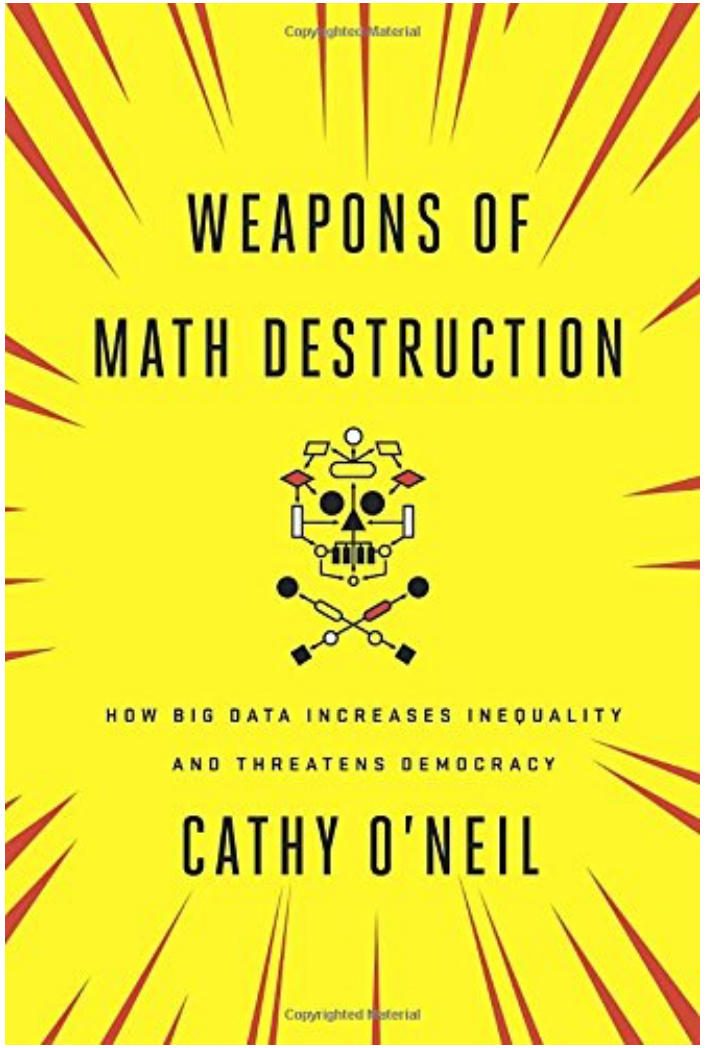
uart Russell
rofessor at
nia, Berkeley
ficial
Approach'.
cs professor
stitute of
he author of

NTELLIGENCE
HE END
UMAN ERA

EINAL
NTION
BARRAT

The Darker Side of Data Science

- Obscuring complex decisions
  - Mortgage backed securities → market crash
  - Teaching scores & job advancement
- Reinforcing historical trends and biases
  - Hiring based on previous hiring data
  - Recidivism and racially biased sentencing
  - Social media, news, and politics
- We will touch on the ethics of data science throughout the class



http://www.npr.org/2016/09/12/493654950/weapons-of-math-destruction-outlines-dangers-of-relying-on-data-analytics

But … I am ***optimistic***

- Knowledge is empowering
- Data science offers **immense potential** to address challenging problems facing society

- The future is in **your hands** and I believe

You will use your knowledge for good.

… I am thrilled to teach Data 100!

What is Data Science?

The recurring question across industry and academia.

My Definition for Data Science

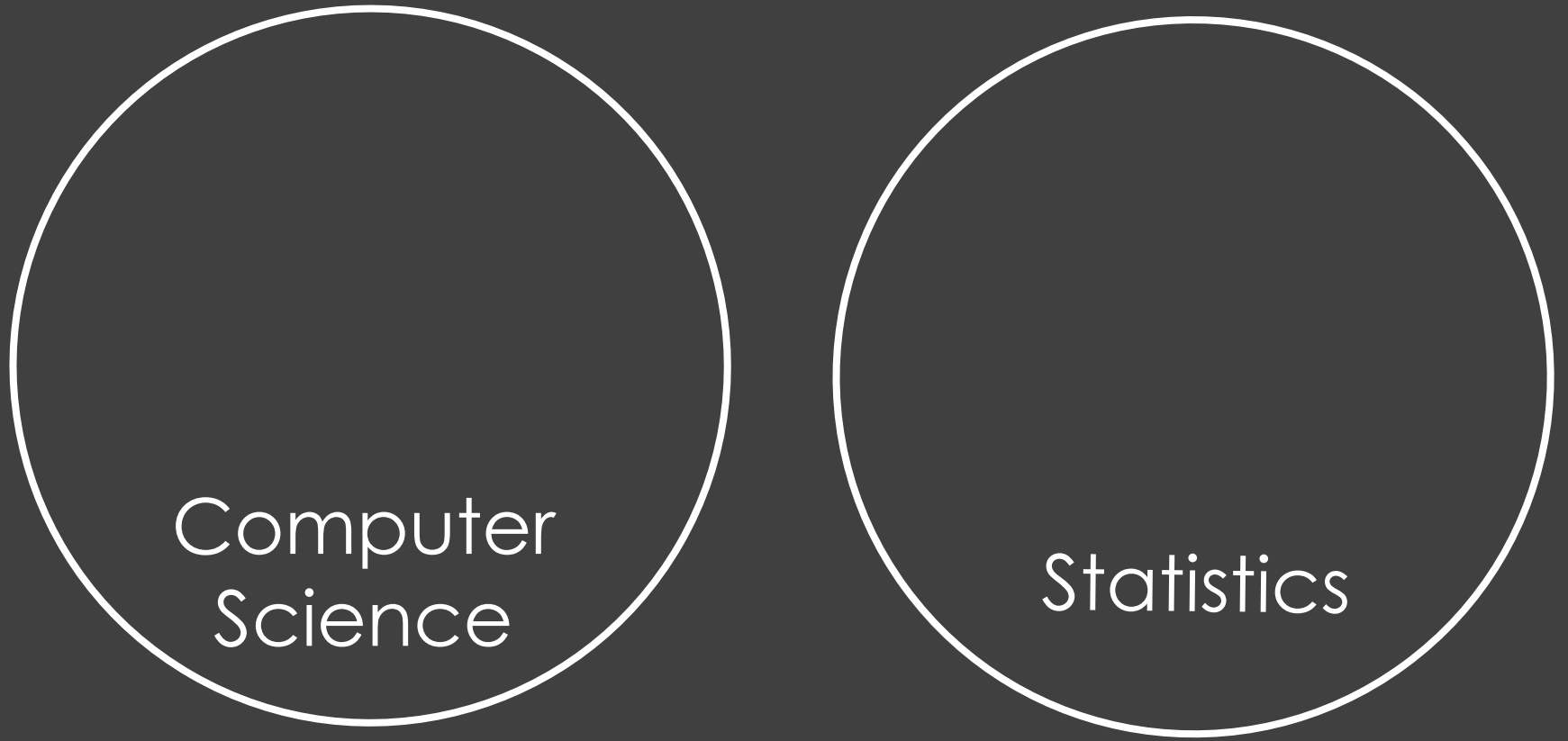
The application of **data centric**, **computational**, and **inferential thinking** to

understand the world **&** *solve problems*

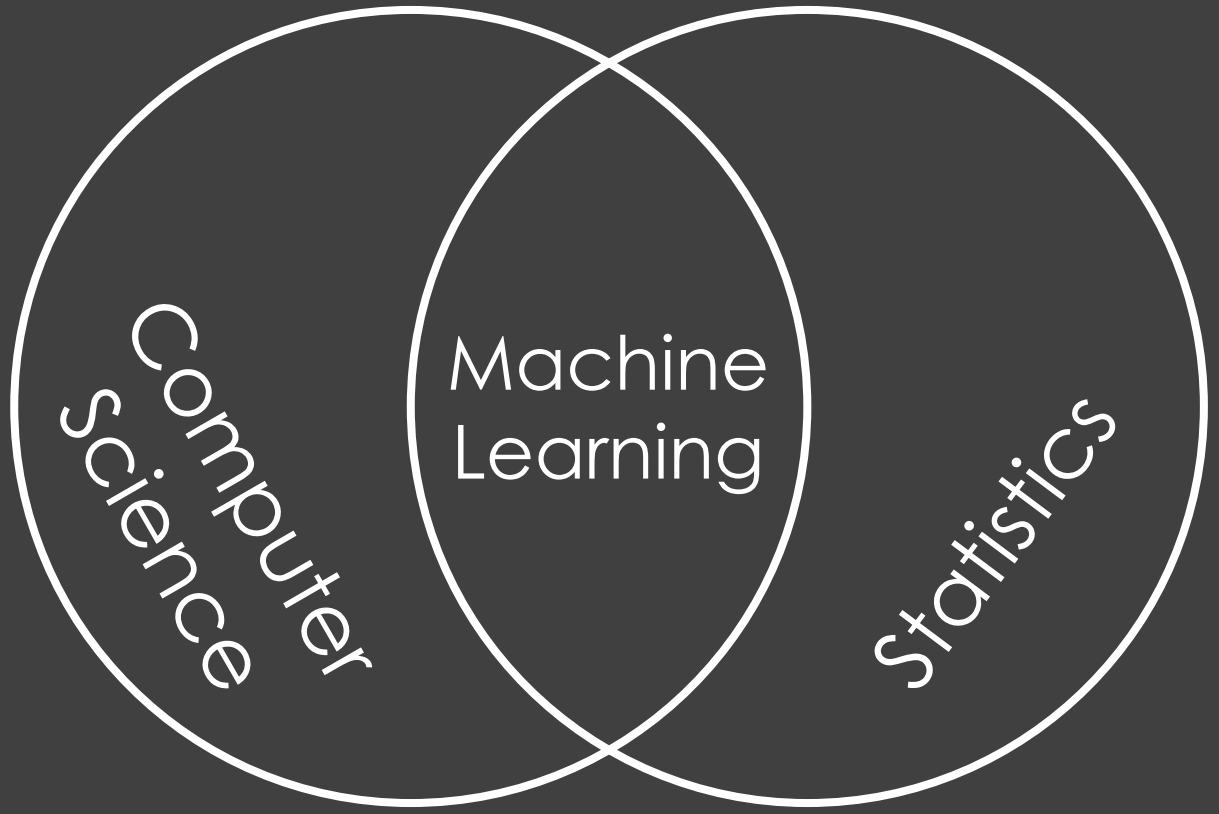
Science **Engineering**

➢ *Data science is fundamentally interdisciplinary*

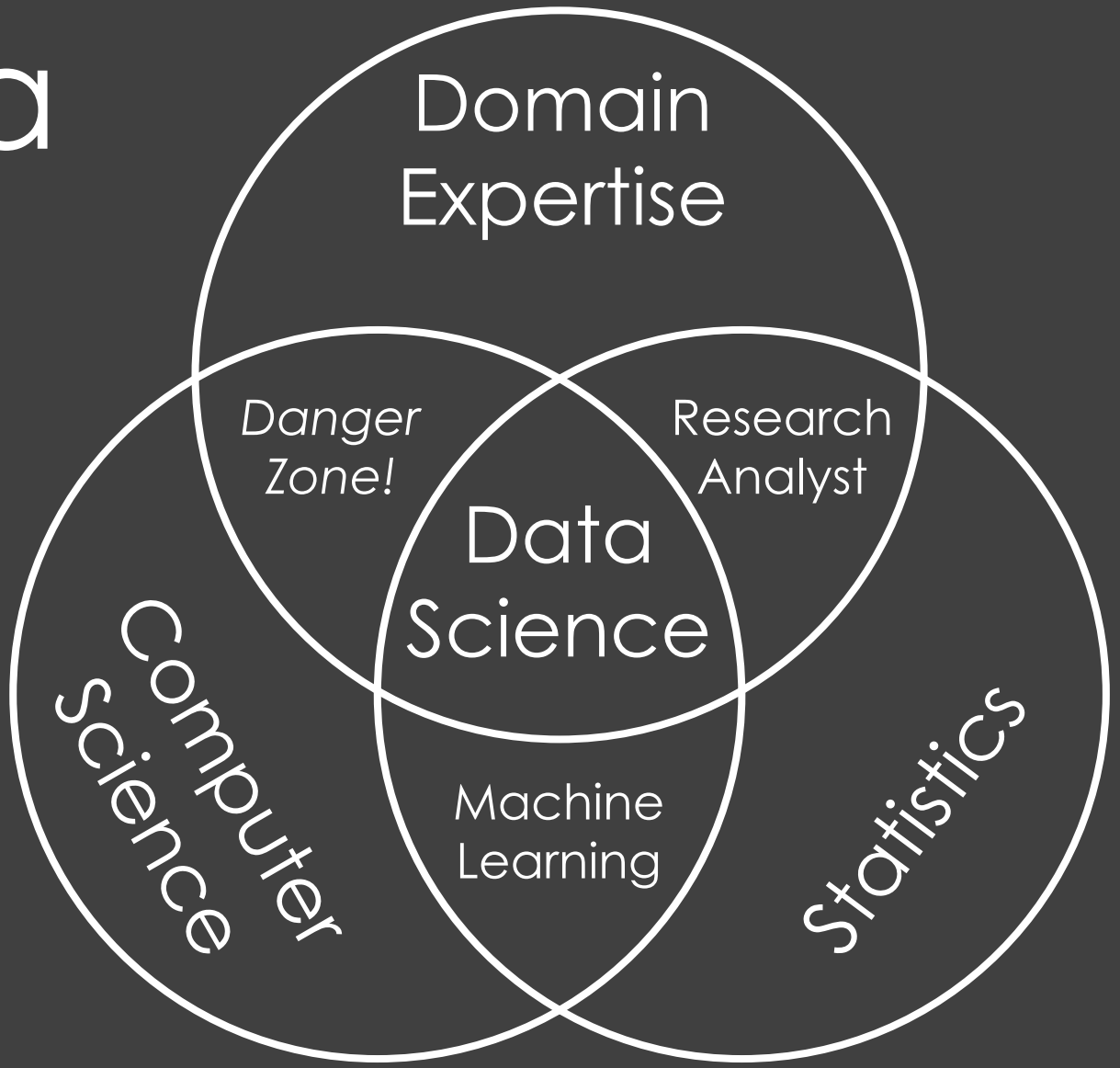
Skills of Data Science



Skills of Data Science



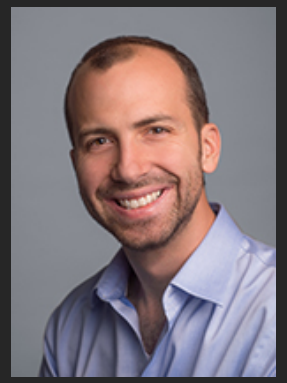
Skills of Data Science



Drew Conway's Venn Diagram of Data Science

Who are we?

The Data 100 Team

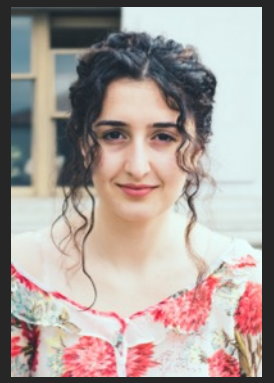

Joseph Gonzalez

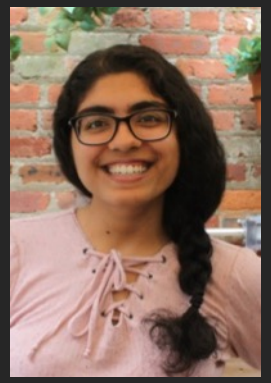

Fernando Perez


Aman Dhar


Andrew Do


Edward Fang

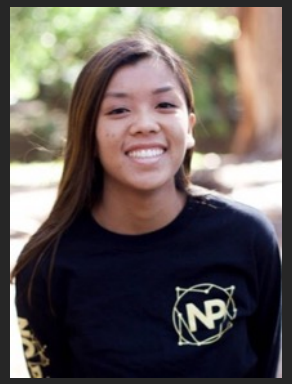

Manana Hakobyan


Sona Jeswani

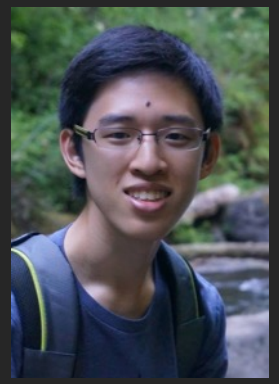

Biye Jiang

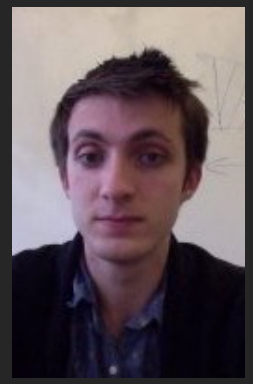

Joyce Lo


Simon Mo


Nhi Quach


Louis Rémus

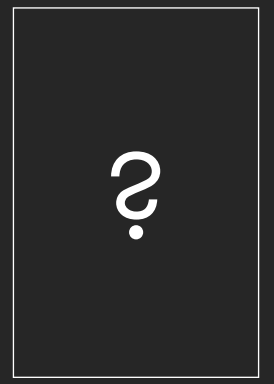

Caleb Siu


Jake Soloff


Weiwei Zhang


TBD


TBD


TBD

Joey Gonzalez

riselab
UC Berkeley

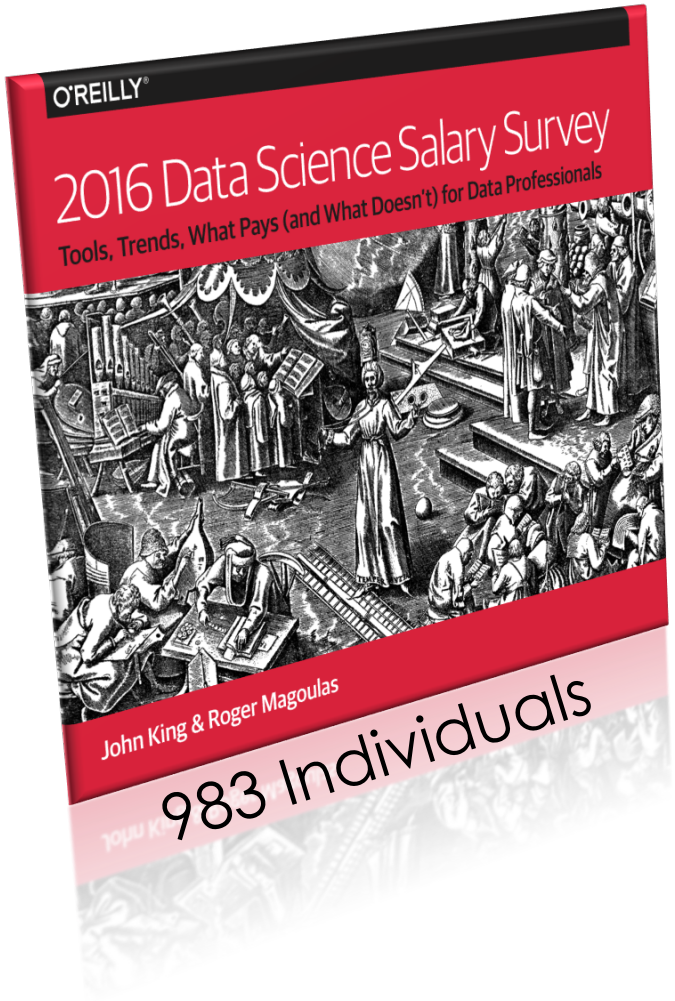
Joined EECS at UC Berkeley in 2016

Research Area: *Machine Learning & Data Systems*

- Study design of scalable systems for machine learning
  - **Algorithms:** designed parallel algorithms for statistical inference
  - **Abstractions:** introduced vertex programming & parameter server
  - **Systems:** developed GraphLab and parts of Apache Spark
- Co-Founder of Turi Inc.
  - Python tools for scalable data science
  - Acquired by Apple Inc. in 2016

What does it mean to be a data scientist today?

How can we answer this question?



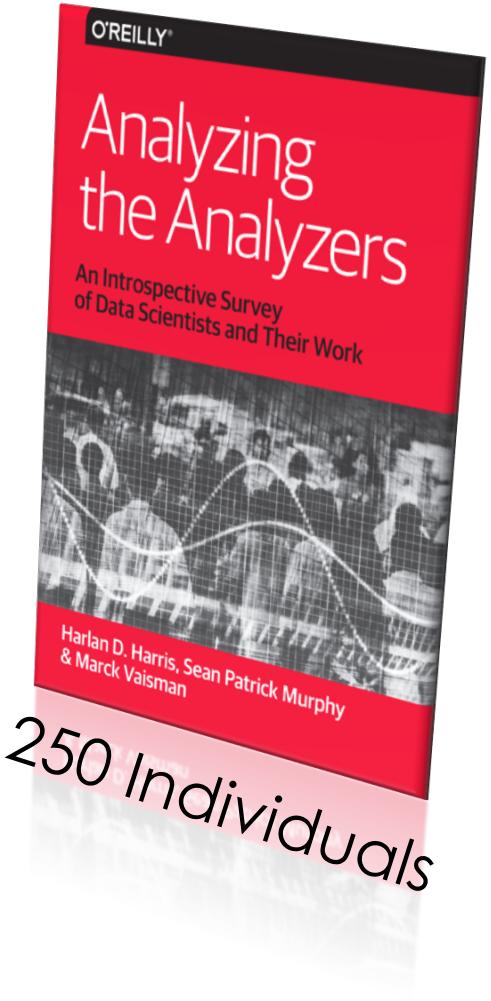
983 Individuals

O'REILLY Surveys

Asked people involved in data science events to complete an online survey

Self reported → Selection bias!

Still somewhat interesting …



250 Individuals

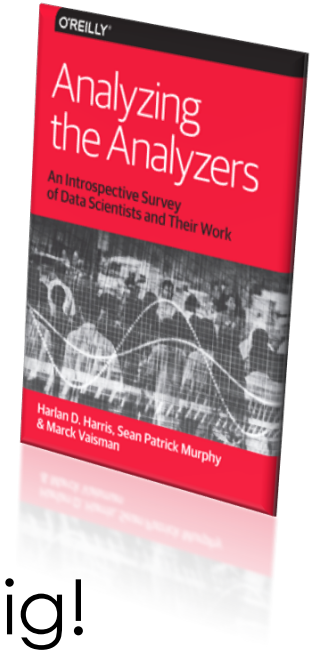
O'Reilly is a good source recent materials on data science.

There is a lot of excitement around

Big Data

… how big is the data?

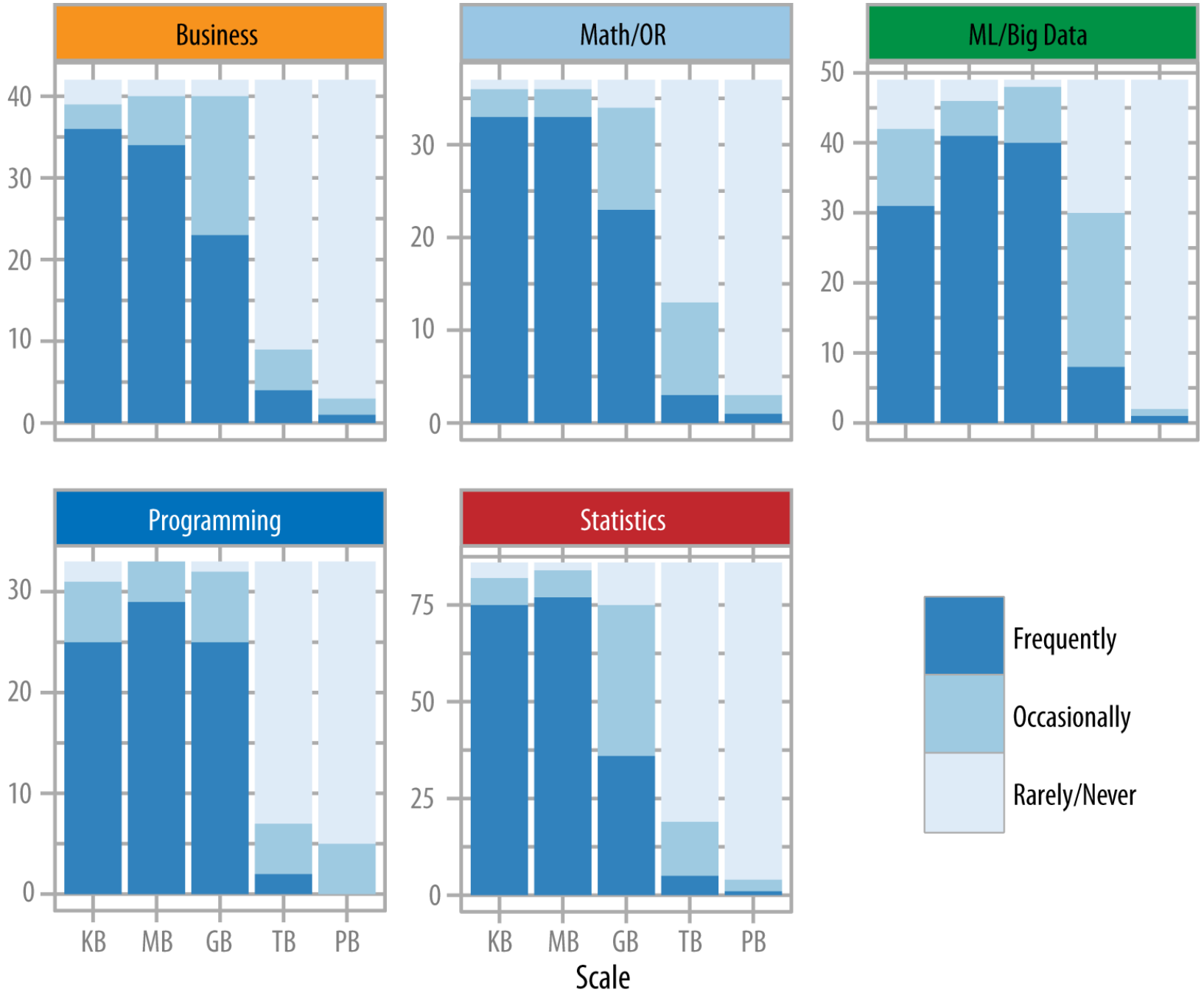
Scale of Data

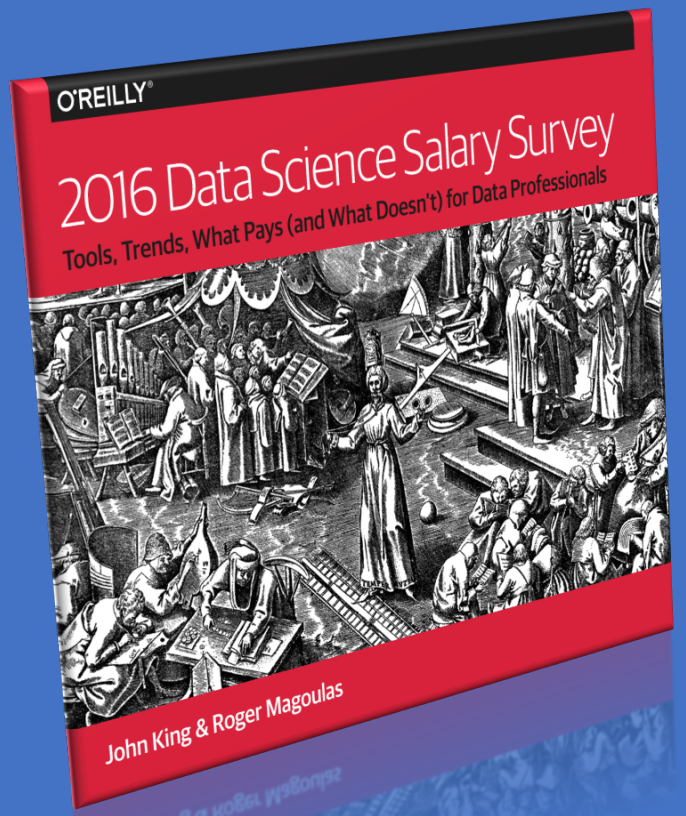


Not usually big!

< 1GB

However some data scientists frequently process TB to PB data sets.





What do they do?

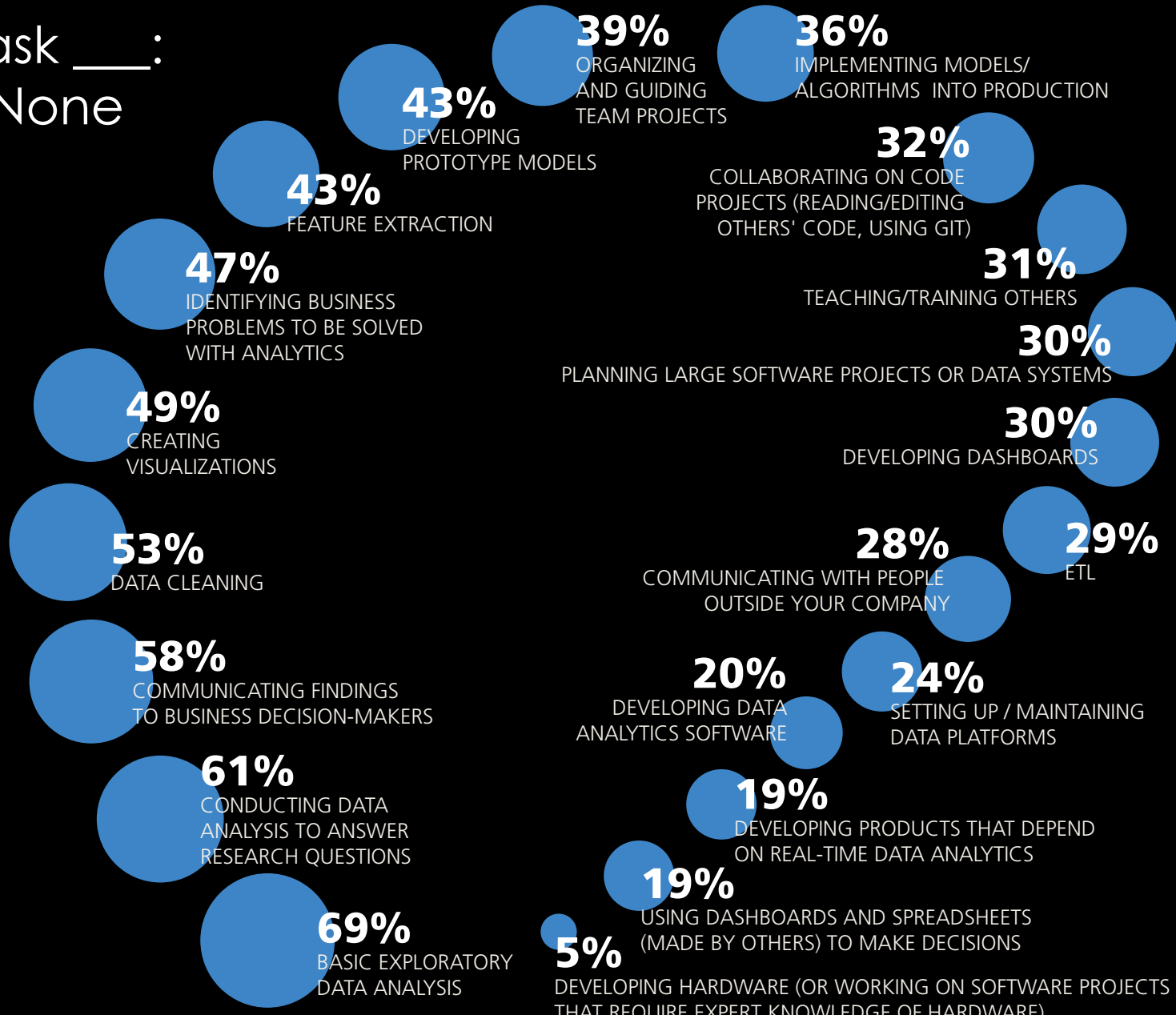
How involved are you in task ___:
(a) Major, (b) Minor, (c) None

Developing Models
Implementing ML Algorithms
Visualization

Exploratory Data Analysis (EDA)
Researching Questions
Writing Reports,

…

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

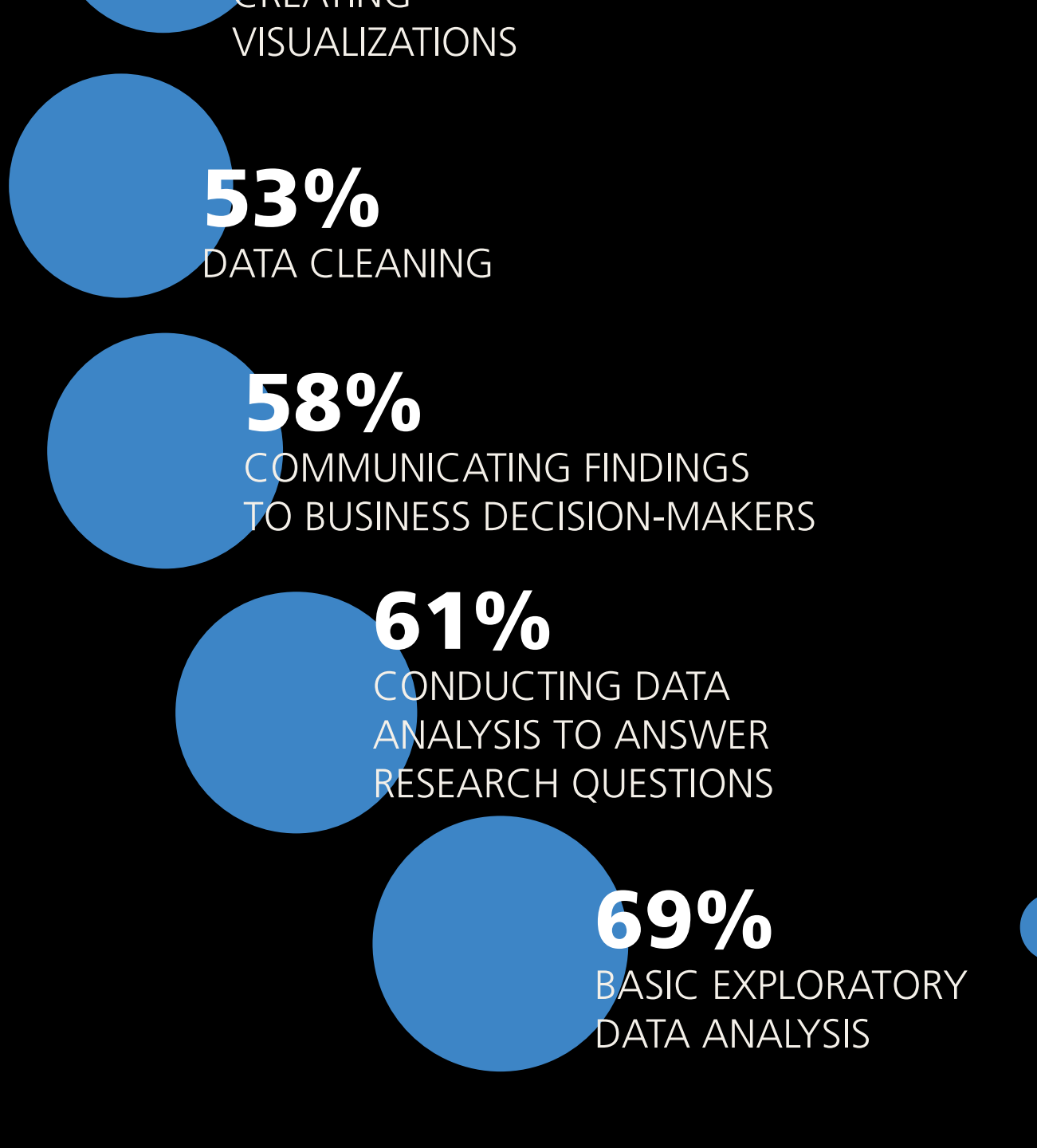


How involved are you in task ___:
(a) Major, (b) Minor, (c) None

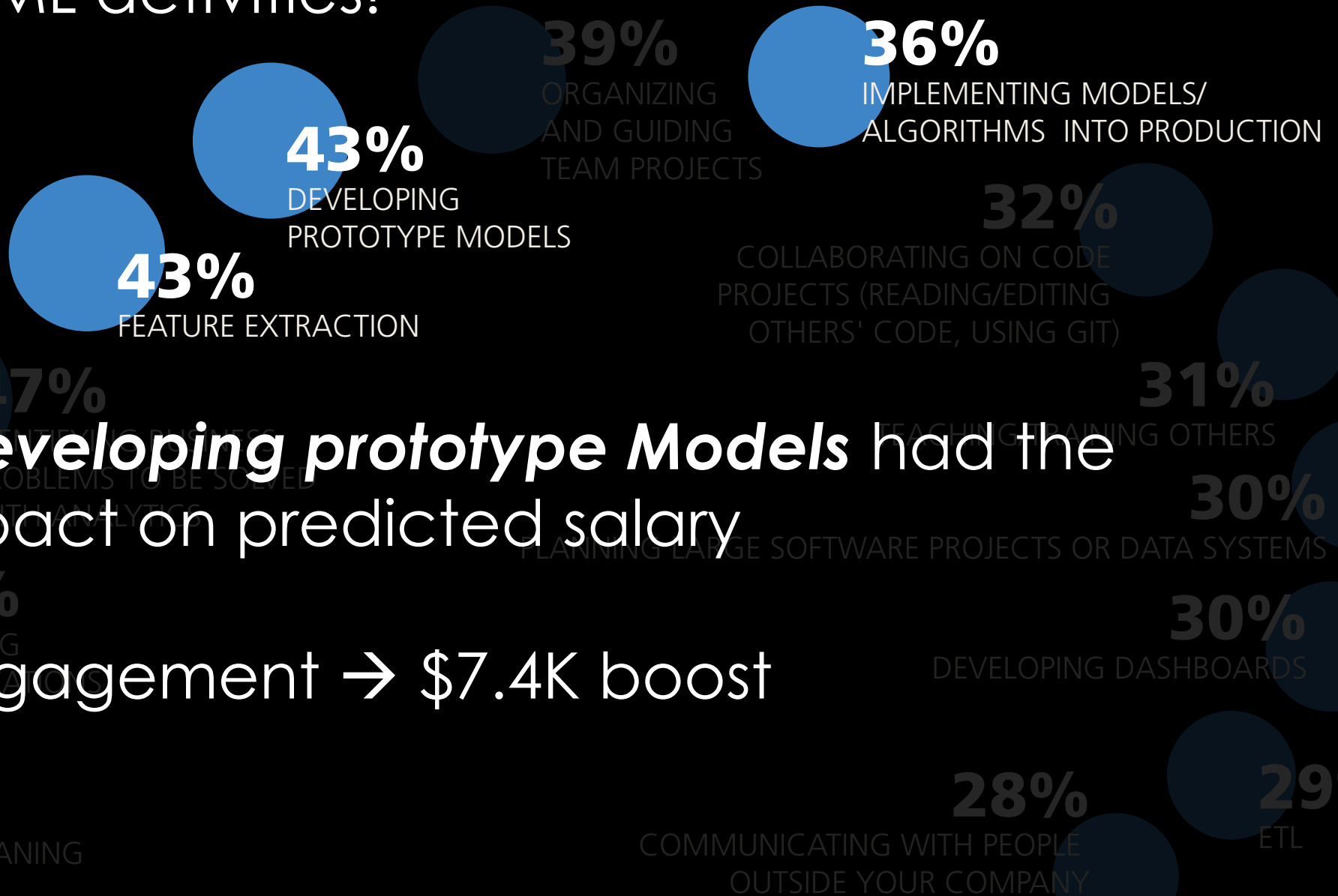
Are the top items surprising?

Data Cleaning ☹

Where are Modeling / Prediction?



Less than half of respondents had major involvement in ML activities!

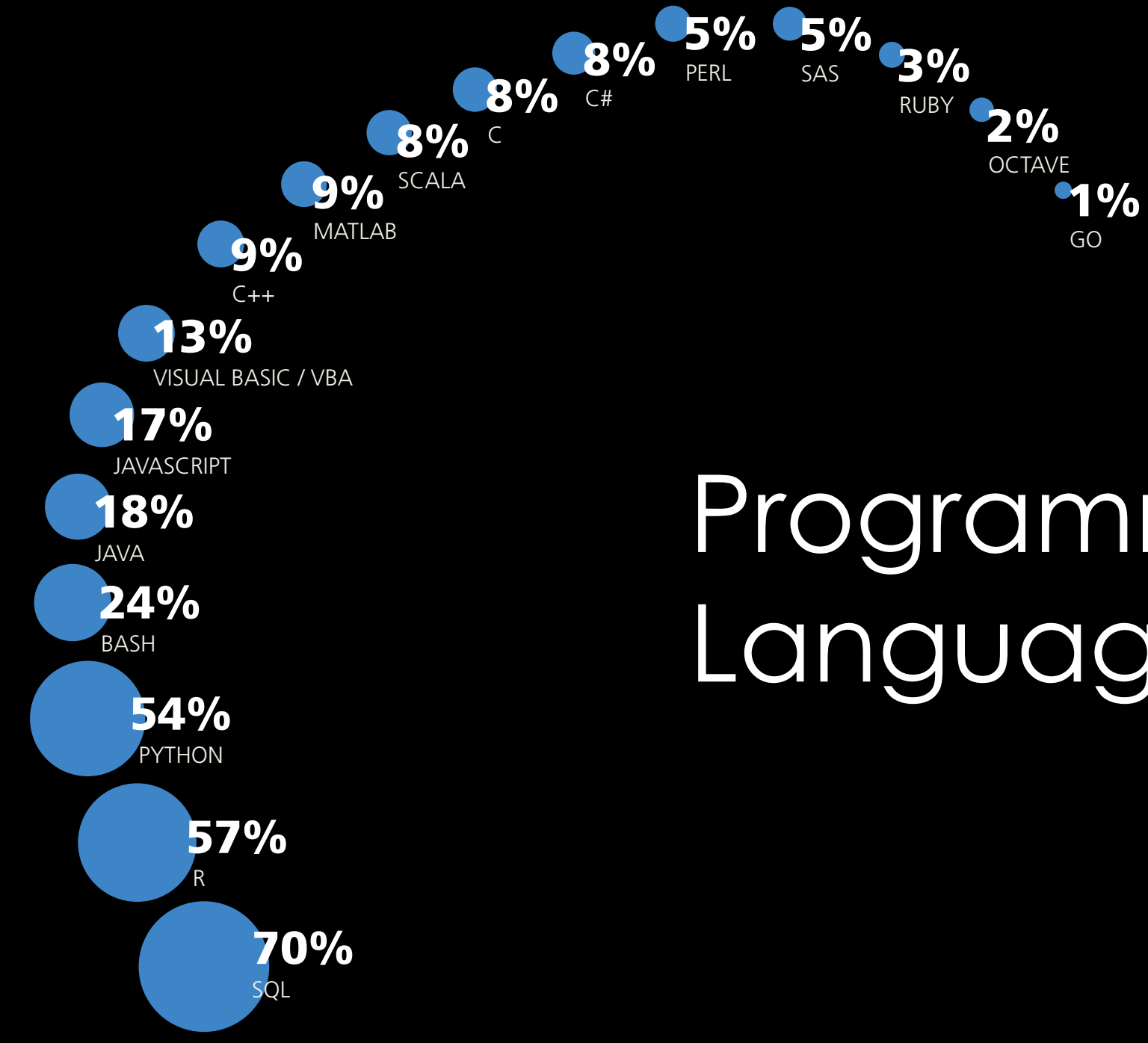


However, ***developing prototype Models*** had the greatest impact on predicted salary

- major engagement → $7.4K boost

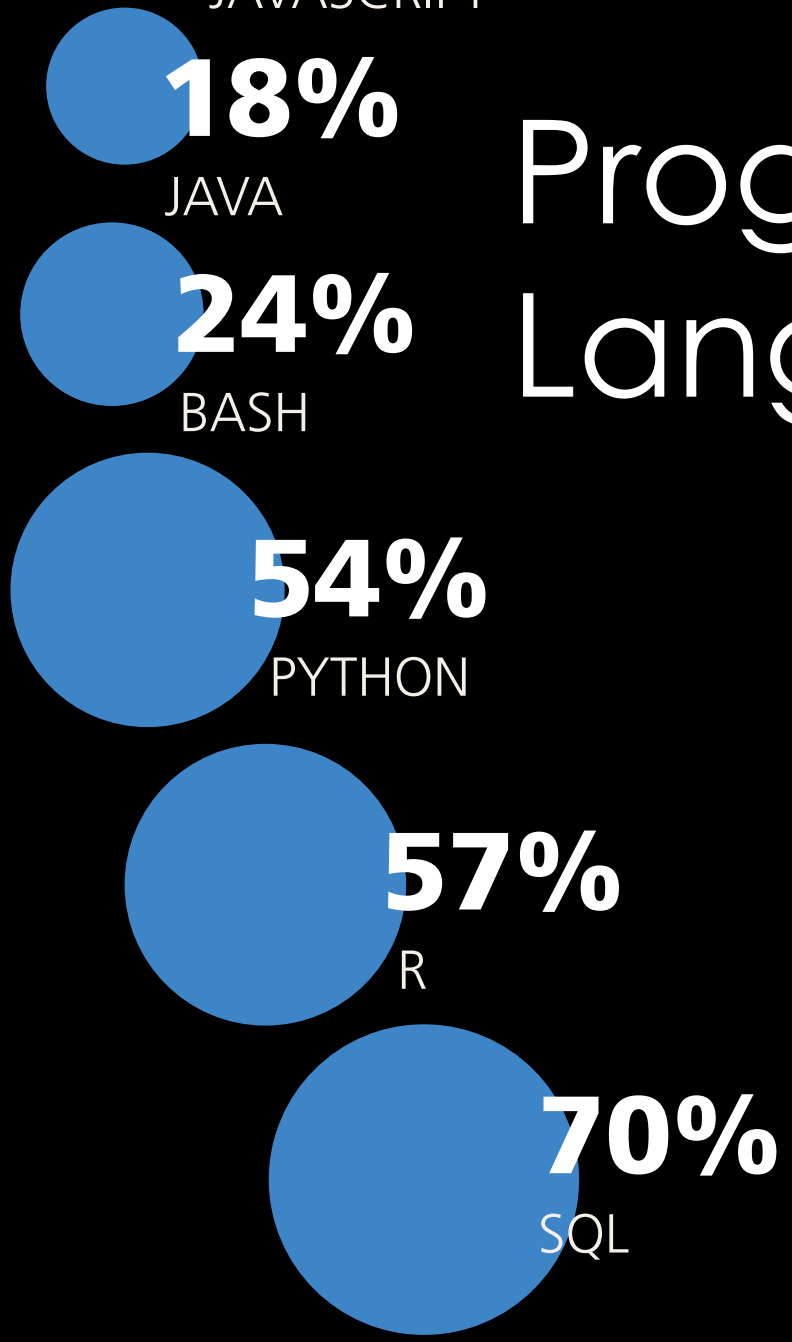
What tools do they use?

- Programming Languages
- Machine Learning



Programming Languages

JAVASCRIPT



Programming Languages

SQL > R > Python

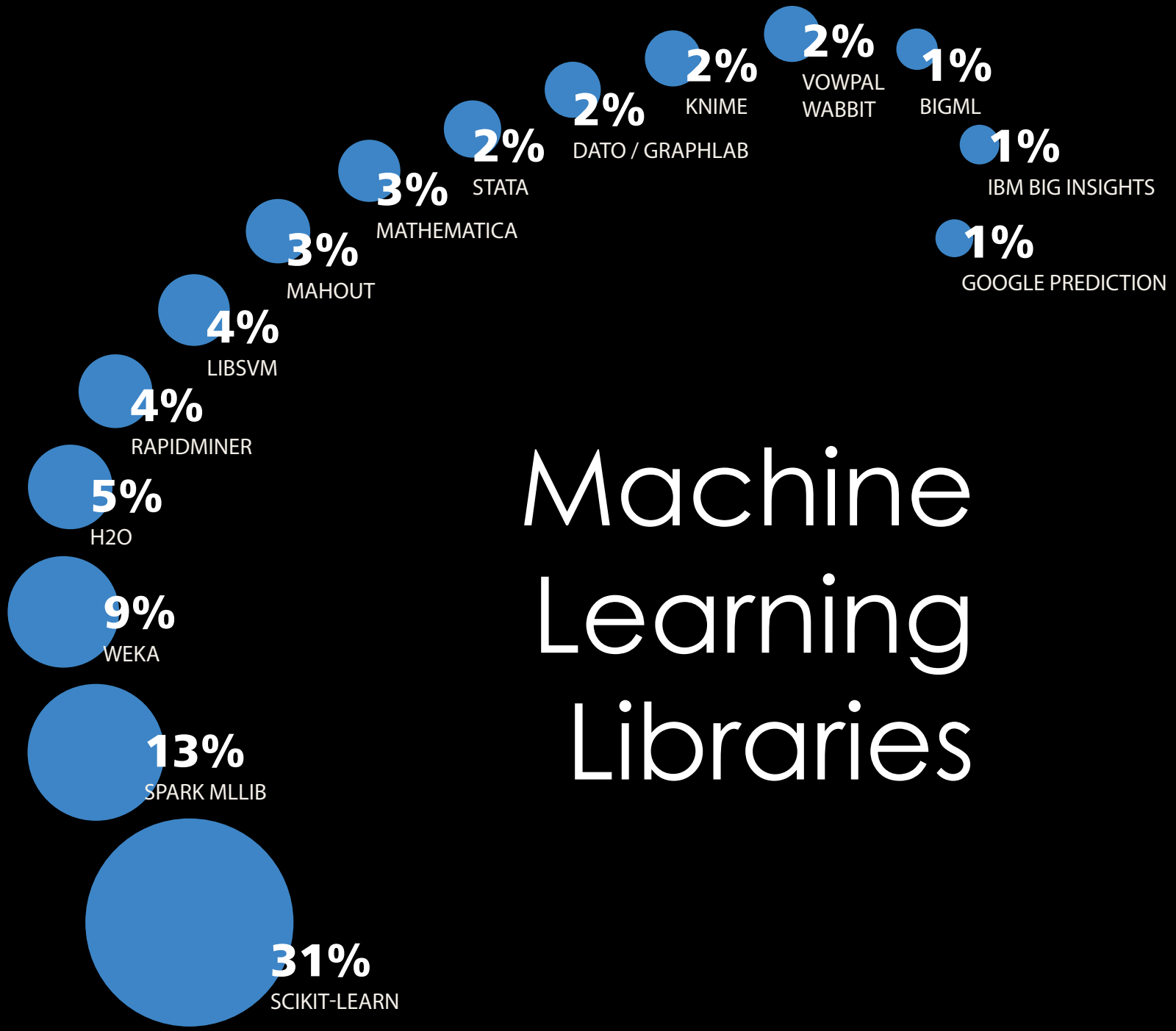
Cluster Analysis:

- Python > R: *data scientists*
- R > Python: *analysts*

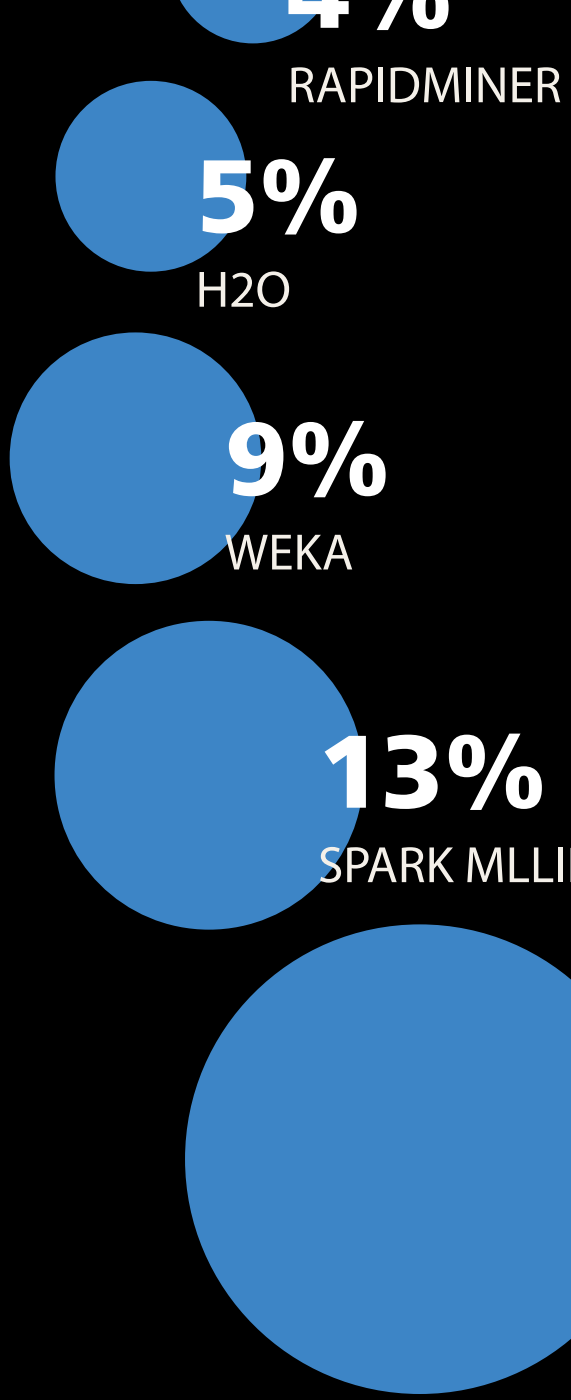
Python users had higher salaries.

Highest Paid?

- Scala



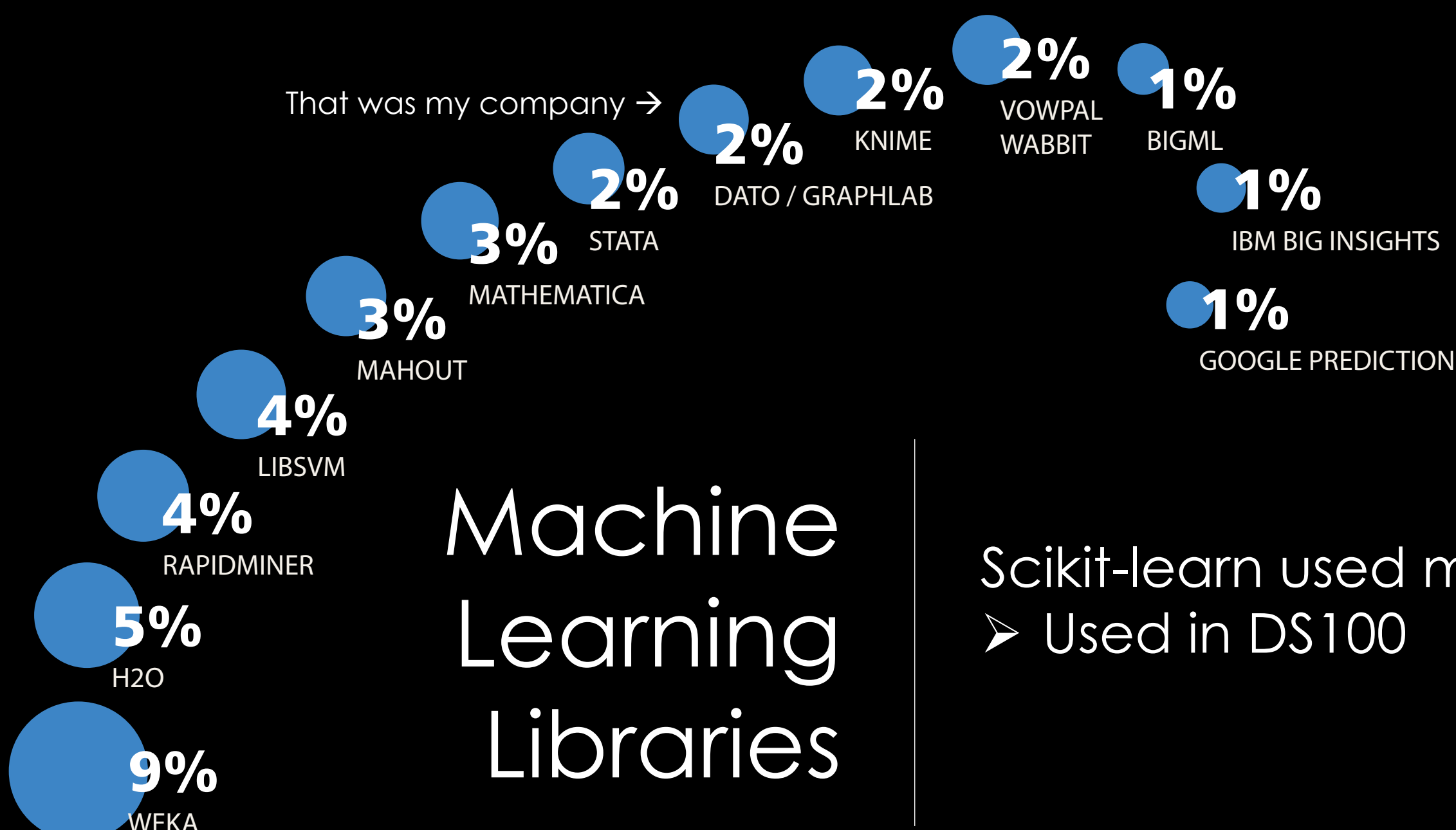
Machine Learning Libraries



Machine Learning Libraries

Scikit-learn used most

- Used in DS100

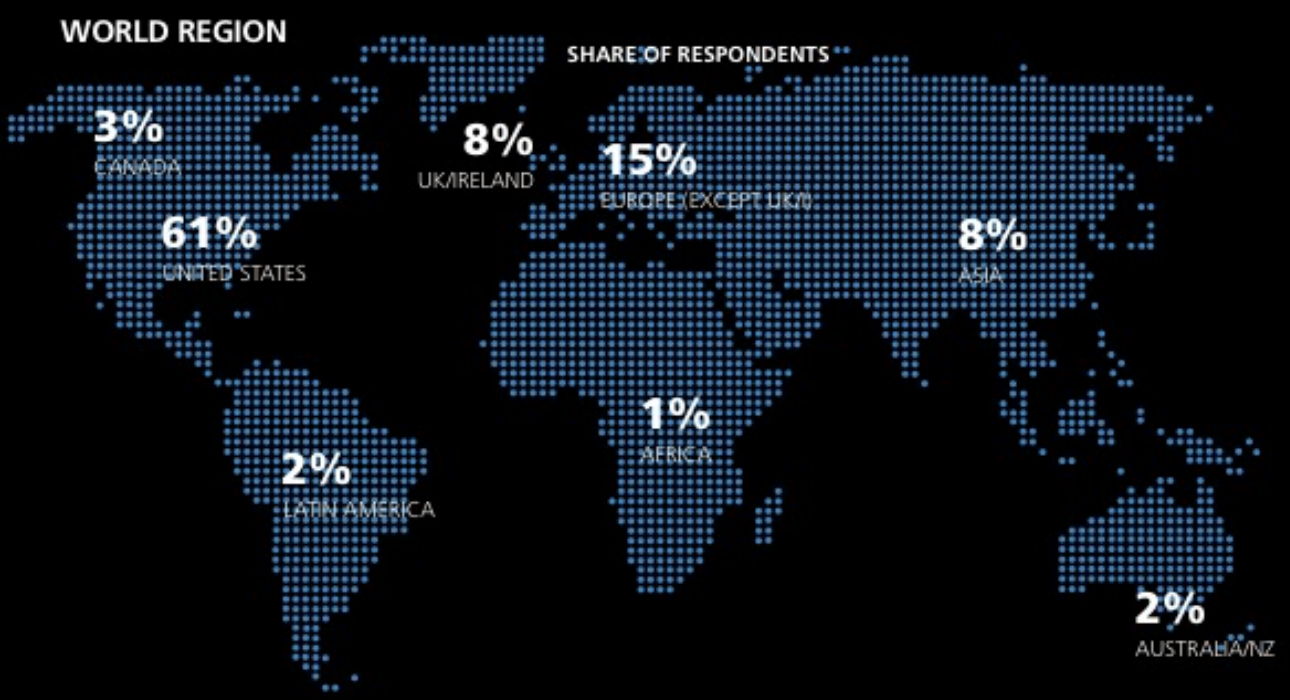


Machine Learning Libraries

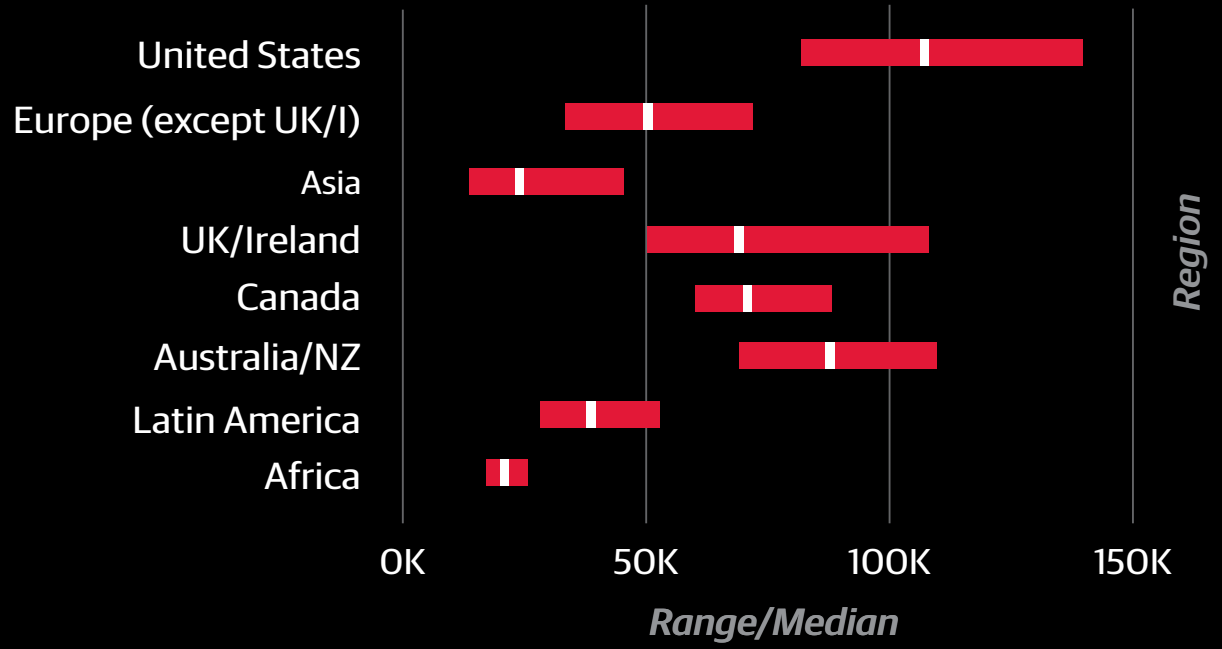
Scikit-learn used most

- Used in DS100

What is their annual income?



SALARY MEDIAN AND IQRC (US DOLLARS)



Salary Depends on Location and Industry

Intermission

5 Minute Break.

Ask a neighbor:

What is your name?

tabs or ***Spaces*** …?

What do statisticians and pirates have in common?

Contemplate:

What are the ethics of data science?

Can data do harm?

What do you want to get out of Data 100?

Pirates say

Important Administrative Reminders

- There will not be any labs or sections this week
- We will be computing optimal assignments for lab & section
  - complete the online section assignment poll
    - https://goo.gl/forms/YohOCvkrUia4zTeI2
- Signup for the DS100 Sp18 Piazza Page
  - https://piazza.com/berkeley/spring2018/ds100/home
- Homework 1 will go out next week and be due the following week.
  - You may start to setup your Python environment

What are your *goals* for DS100?

- What do you want to learn?
- How does this class fit into your future plans?

Our Goals

Prepare students for advanced Berkeley courses in data-management, machine learning, and statistics, by providing the necessary foundation and context

Enable students to start careers as data scientists by providing experience in working with ***real*** **data**, **tools**, and **techniques.**

Empower students to apply **computational** and **inferential thinking** to address real-world problems

What are the Prereqs. for Data 100

- Officially Listed Prerequisites:
  - Foundations in Data Science [Data8]
  - Computing [CS61a or CS88 or … E7]
  - Calculus and Linear Algebra [Math 54 or EE16a or Stat 88]
- We will not be enforcing prerequisites
  - … however you should be familiar with the material in these classes (especially Data8)
- Homework 1 will help verify your familiarity
  - Do Hw1 and skim the Data8 textbook: https://www.inferentialthinking.com

What will I learn?

Topics covered in Data 100

- Data collection and sampling
- Data cleaning and manipulation
- Regular Expressions
- SQL and Enterprise Data Management
- Xpath and web-scraping
- Exploratory Data Analysis & Visualization
- Hypothesis Testing & Confidence Int.
- Model design & loss formulation
- Batch and Stochastic Gradient Descent
- Ordinary Least Squares Regression
- Logistic Regression
- Feature Engineering
- The Bias - Variance Tradeoff & overfitting
- Regularization & Cross validation

We will use ***Real Data***

Homework, labs, and in class examples will build on real data:

- Twitter, Speeches, Scientific Data, Maps, Surveys, Images, …

The data will be:

- **messy** and you will have to clean it
- ***big(ish)*** and you will have to be a little clever to process it
- **complicated** and you will have to learn about the ***domain***

You will Learn How to Use Real Tools

- Focus on Python programming language
- We will use various different technologies
  - Jupyter notebooks, pandas, numpy, matplotlib, postgres, seaborn, scikit-learn, plotly, Dask, …
- *We **won't** teach you everything …*
  - You will learn to **read documentation**
  - You will learn to **teach yourself**
- ***BETA WARNING:*** Things will break …
  - You will learn how **to debug**
  - You will learn how **to get help** (on Piazza)

Reading and Reference Materials

No single great book (working on a Data 100 gitbook …)

- Lectures slides and screencasts will be available on online
- **Use online reference materials**

We will occasionally (in a few lectures) reference a few ebooks

- Joel Grus. "*Data Science from Scratch*" [eBook Link]
- Cathy O'Neil and Rachel Schutt. "*Doing Data Science*" [eBook Link]
- G. James, D. Witten, T. Hastie and R. Tibshirani. "*An Introduction to Statistical Learning.*" [pdf Link]
- Wes McKinney. "*Python for Data Analysis*" [pdf link]

Grades

[20%] 6 Homework assignments
(drop the lowest)

[10%] 2 Projects (multi-week homework's)

[10%] Labs (Graded on Completion)

[5%] Vitamins (weekly online quizzes)

[5%] In class participation

- Participate in at least 18 of the lectures for full credit.
- Using google forms or bcourses (**bring a browser**)

[20%] 1 Midterm (in class)

[30%] 1 Final

On Time Policy (don't be late)

- **5 days** of "slip-time" to be **used on homework/projects** for **unforeseen circumstances** (e.g., get sick or deadline conflicts)
- After you have used your slip-time budget
  - **20% per day for each late day**
- If you are having trouble finishing assignments on time let us know!

Collaboration Policy: ***Don't Cheat!***

- Data Science is a collaborative activity
- You may discuss problems with friends
  - List their names at the top of your assignments
  - We may periodically analyze the collaboration networks
- ***<u>You must write your solutions individually</u>***

<u>Don't Cheat</u>

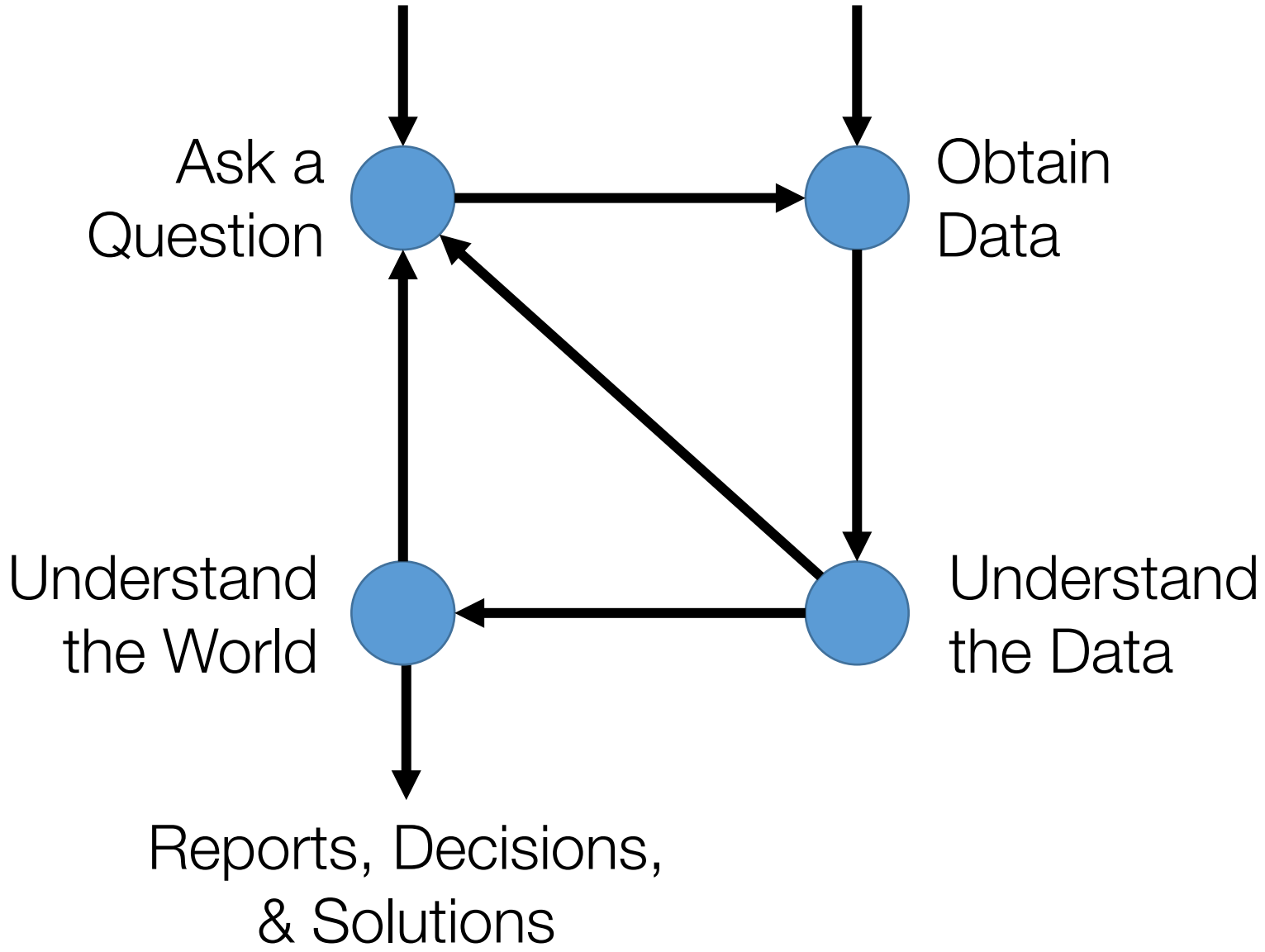
- Content in the homework and vitamins will be on the midterm and final
- If you are struggling let us know so we can help!

Staying Up to Date

- All communication will be through Piazza
  - https://piazza.com/berkeley/spring2018/ds100/home
  - If you have questions about assignments
    - Try commenting on the appropriate discussion
    - Do not share your code publicly
  - If you have private question → write a private post on Piazza
    - This will ensure a quick response
- We will also be updating the website with links to homework, lectures, and vitamins
  - http://www.ds100.org/sp18/

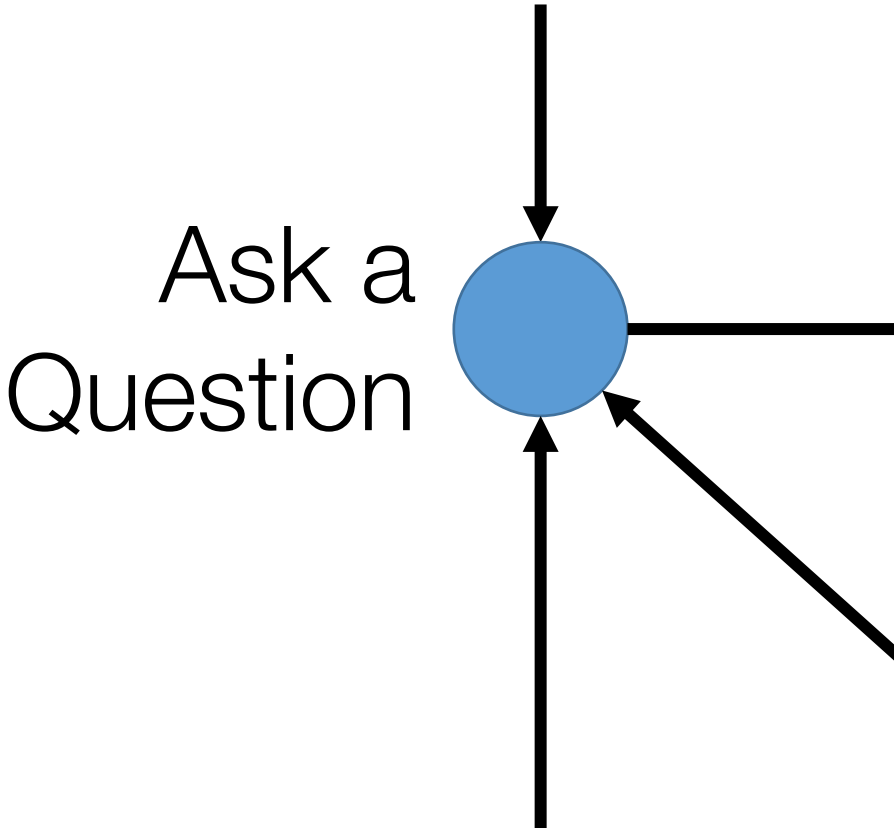
Data Science Lifecycle

High-level description of the data science workflow

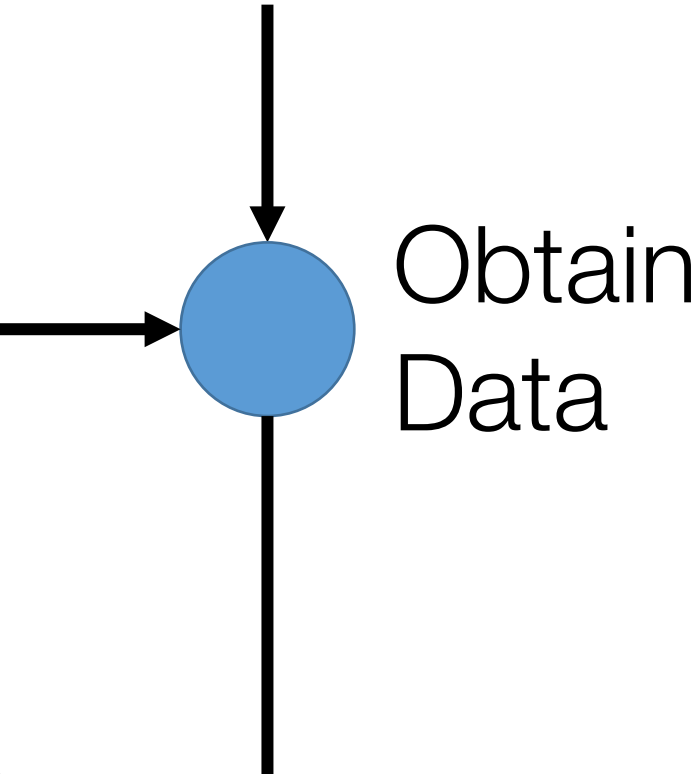

Ask a
Question
Obtain
Data
Understand
the World
Understand
the Data
Reports, Decisions,
& Solutions

Question / Problem Formulation

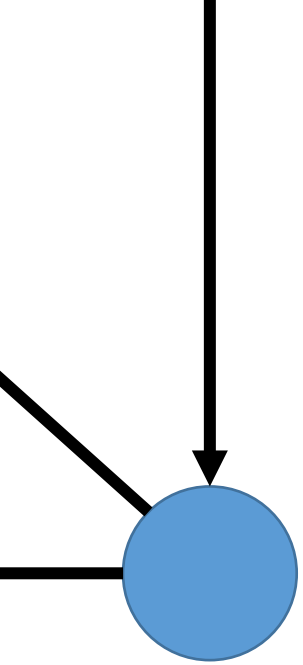
- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics of success?



Data Acquisition and Cleaning



- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?

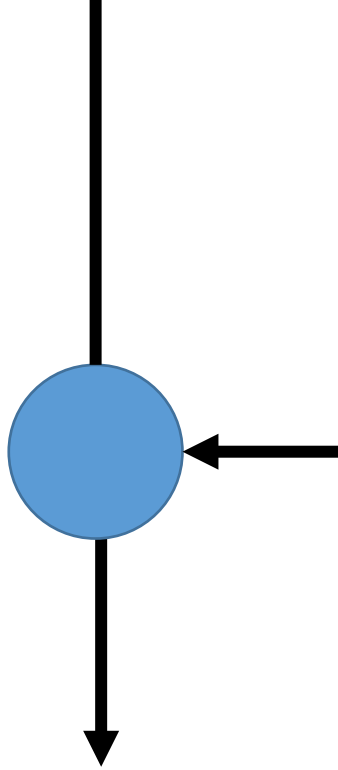


Understand the Data

- How is our data organized and what does it contain?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

Exploratory Data Analysis & Visualization

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



Predictions and Inference

Data Science Demo