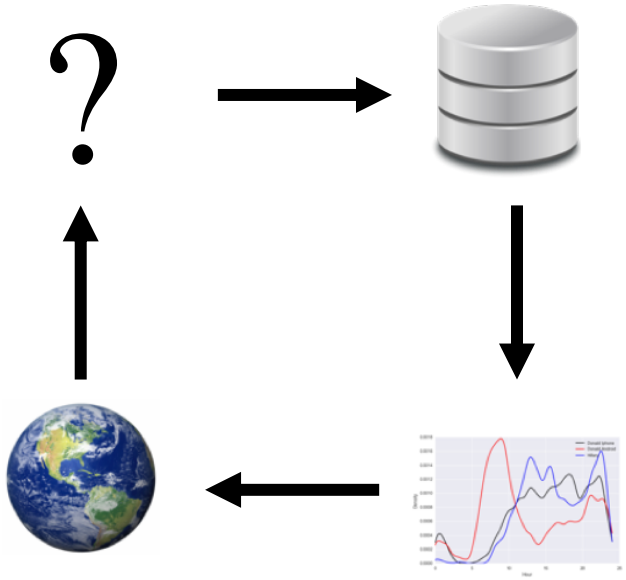
Data 100

Principles & Techniques of Data Science

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu



Questions for Today

- ***Who*** *am I?*
- ***Why*** *am I excited about Data Science?*
- **What** is Data Science?
- **What** is this class about?
- **Break**
- **Demo (who are you?)**

Slides from lecture available online at http://ds100.org/sp20

Joey Gonzalez

Joined EECS at UC Berkeley in 2016

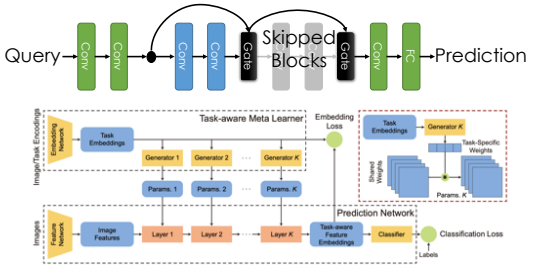
Co-director of the UC Berkeley RISE Lab

Research Area: *Machine Learning & Data Systems*

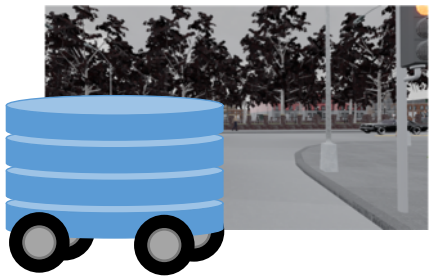
Distributed Data Systems

Systems for AI

Deep Learning



Autonomous Vehicles



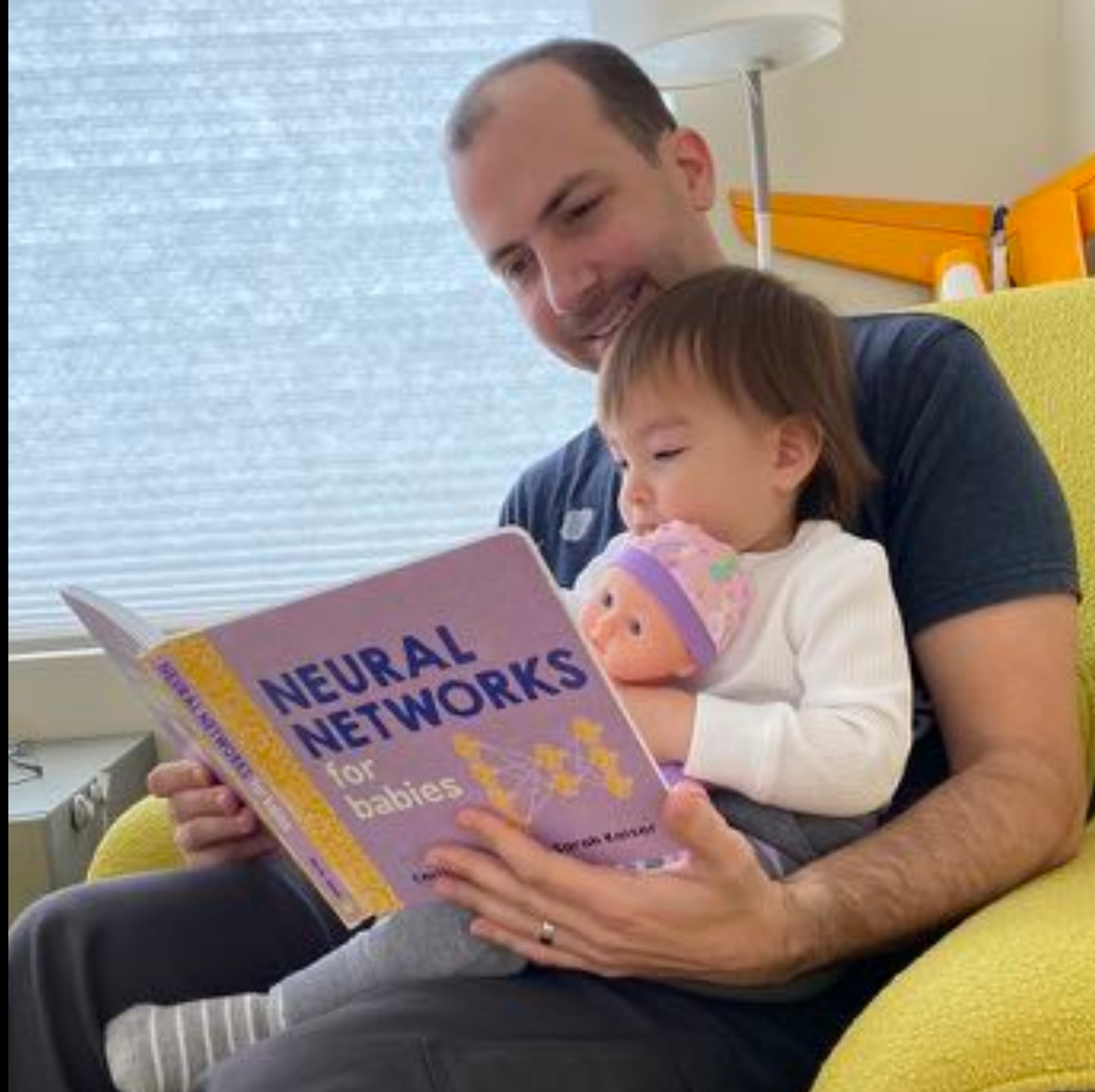
- Co-Founder of Turi Inc. (based on GraphLab)
  - Python tools for scalable data science → Acquired by Apple Inc.

Also…

Meta-learning with **Real** Neural Nets

I need to end office hours at 5:00 for daycare pickup.

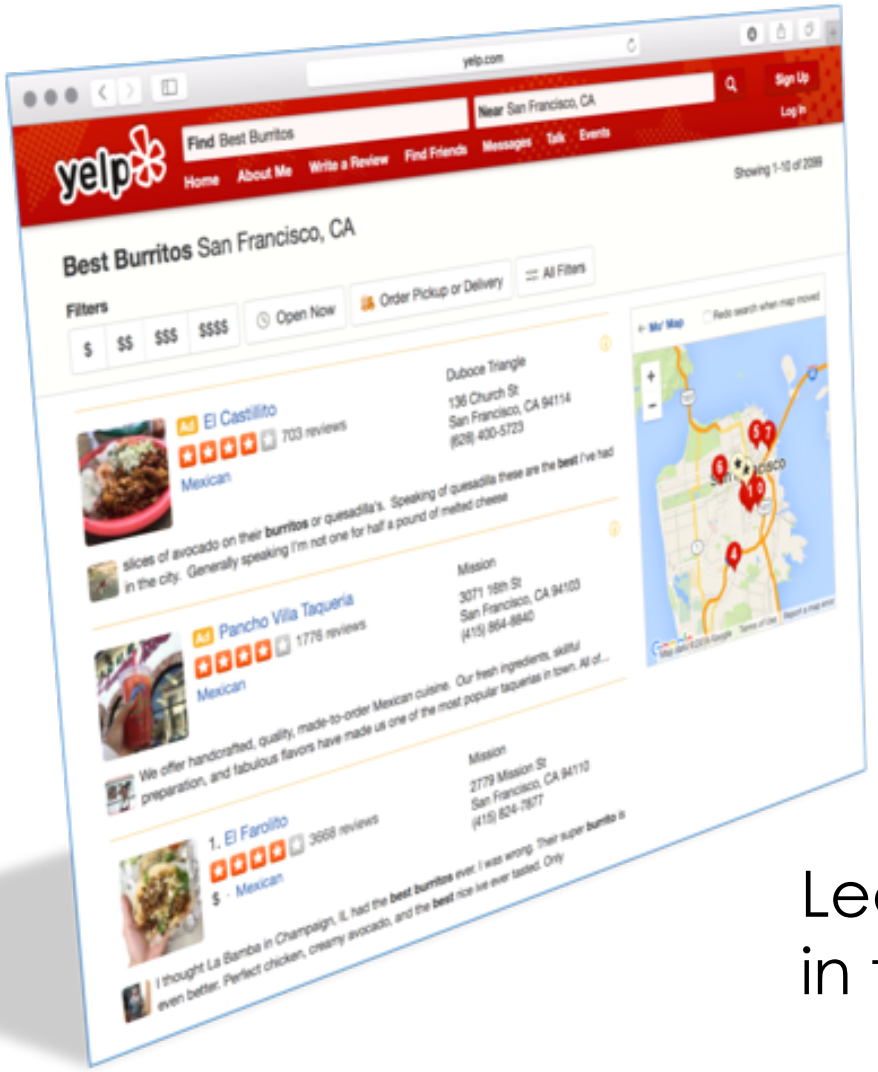
(If I forget, remind me!)



__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

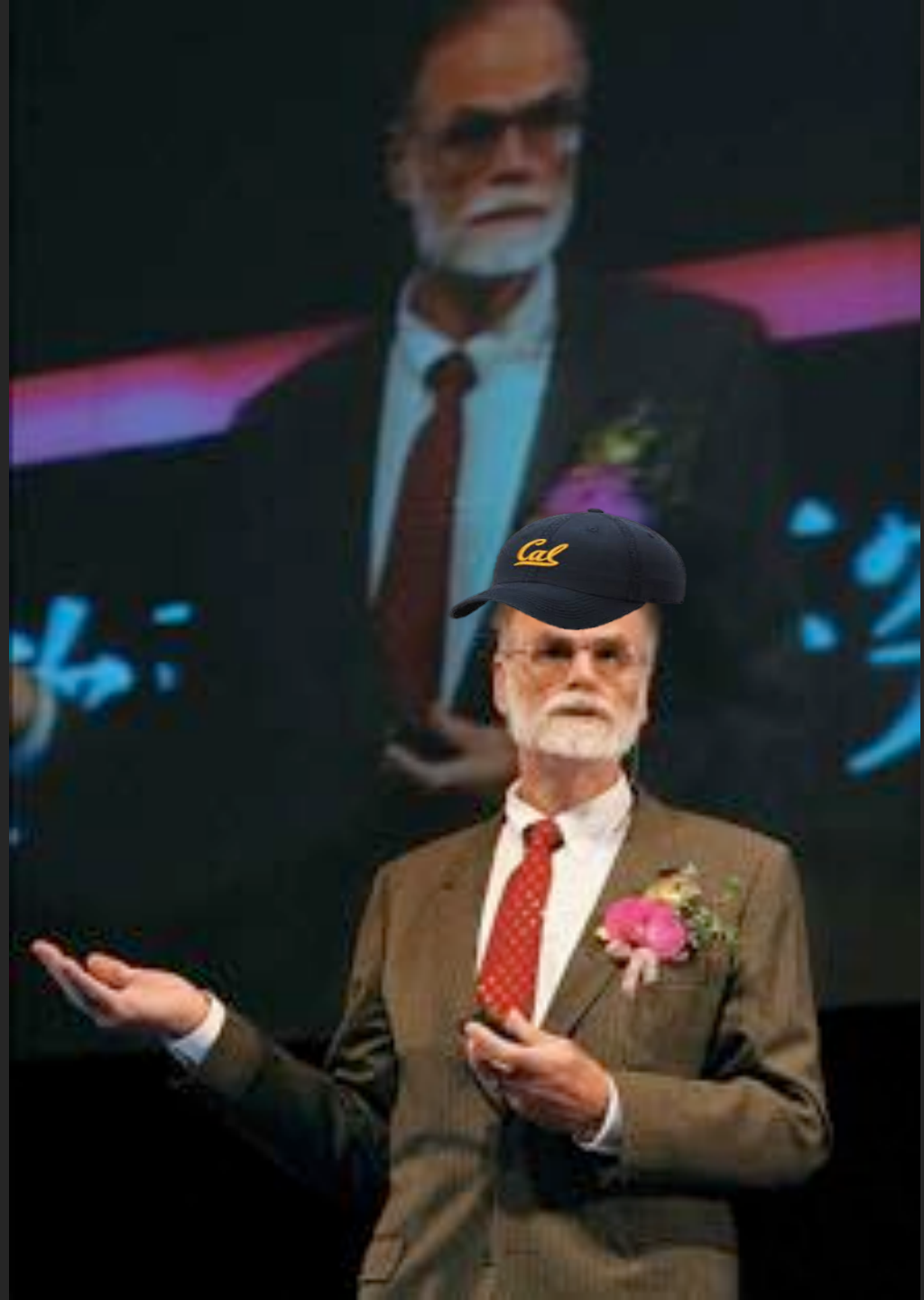
Learn about the dangers of eating
in the first project …

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

Data Science is transforming *Science*

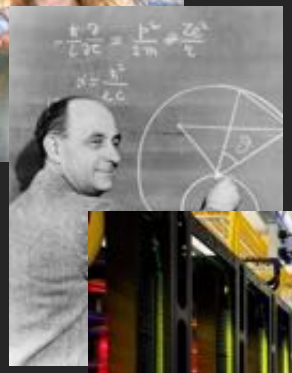
Jim Gray

Turing Award Winning Computer Scientist & Cal Alum. (1st CS PhD)



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

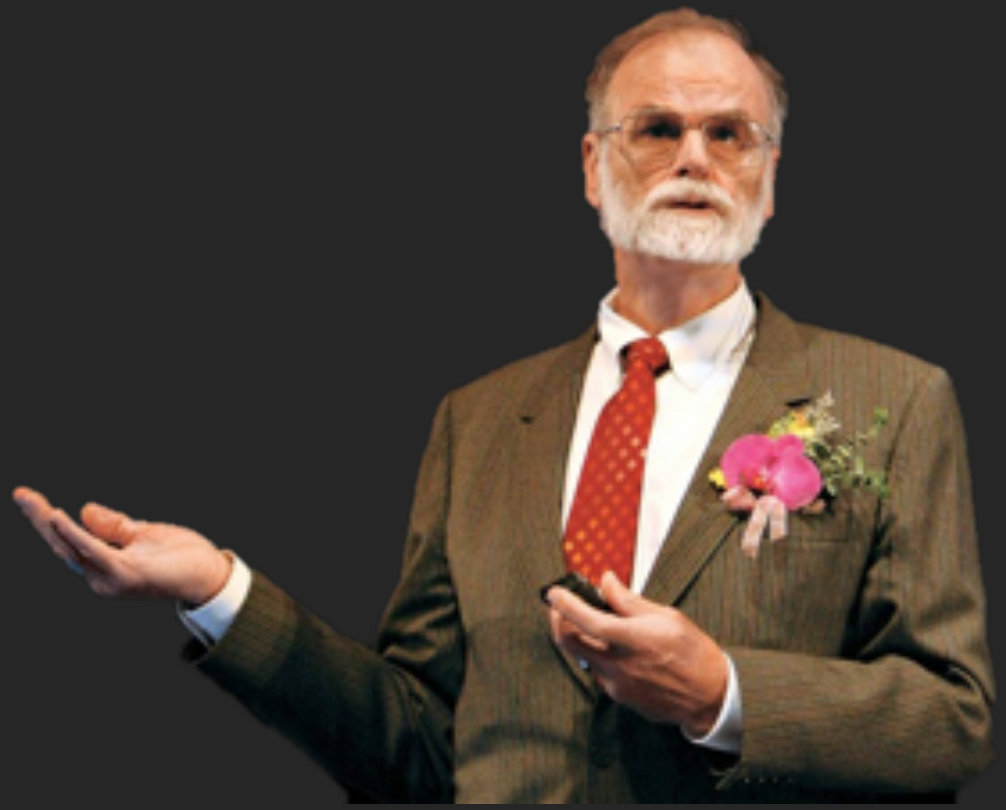
Experimental



Theoretical

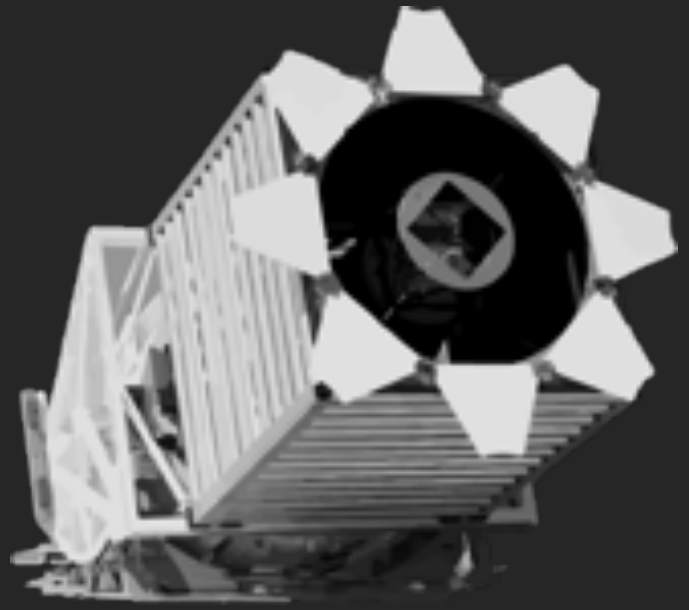
Simulation

Data Intensive



Jim Gray

Astronomy in the 4th Paradigm



Sloan Digital
Sky Survey (SDSS)

+

Database
Systems

Sky
Server

Technology Trends

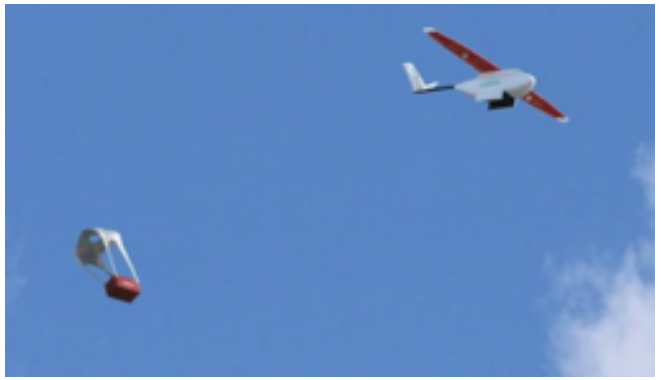
2020s ?

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

On the Threat of Artificial Intelligence

Stephen Hawking

There are more **immediate concerns.**

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

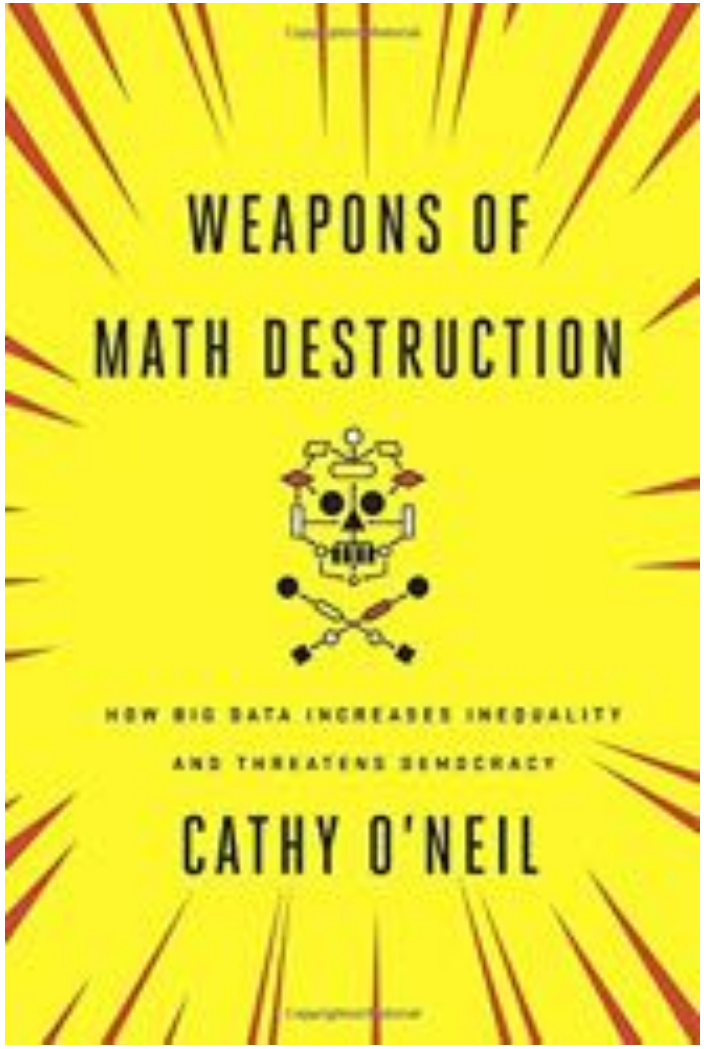
uart Russell
rofessor at
nia, Berkeley
ficial
Approach'.
cs professor
stitute of
he author of

NTELLIGENCE
E END
UMAN ERA

INAL
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 discuss 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

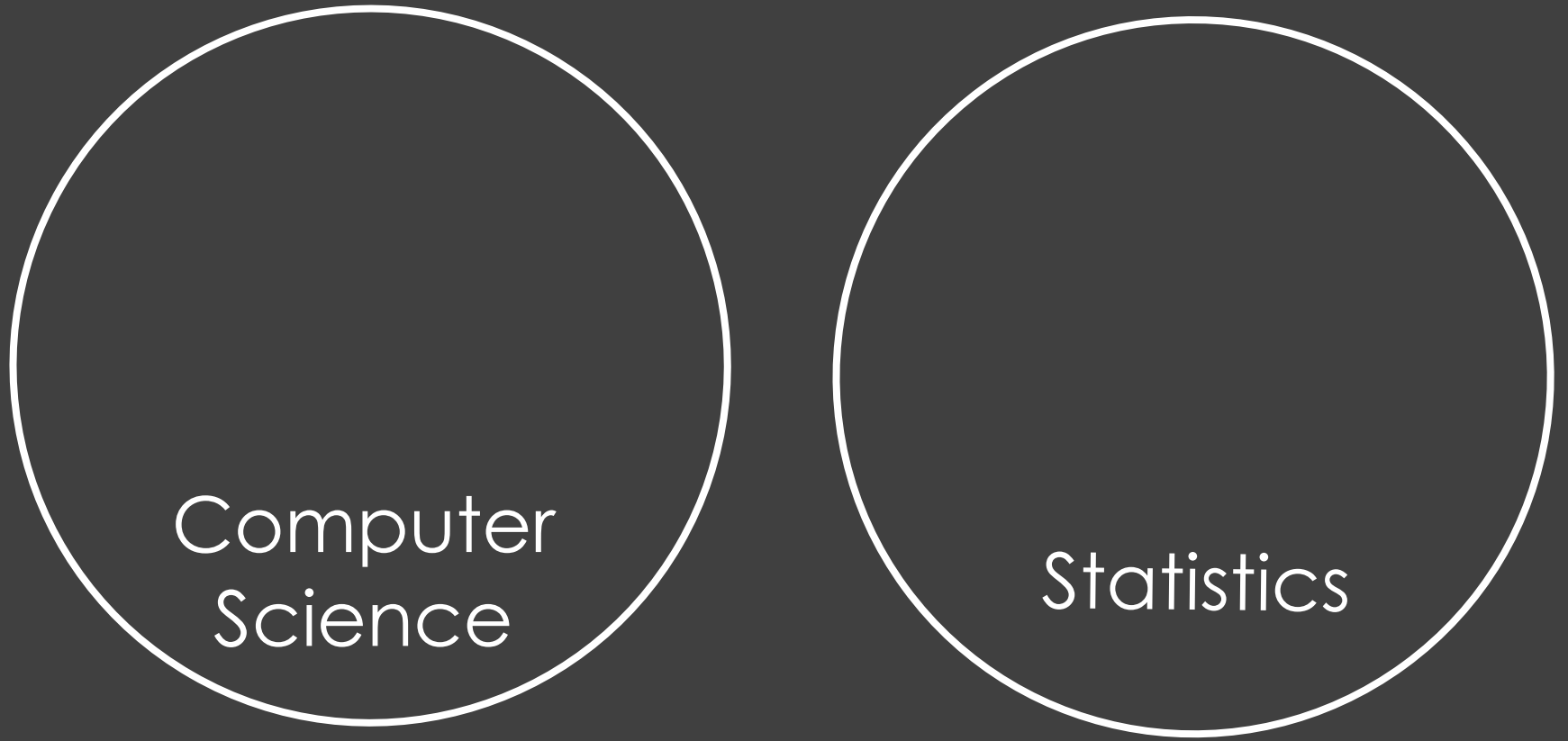
understand the world — **Science**

&

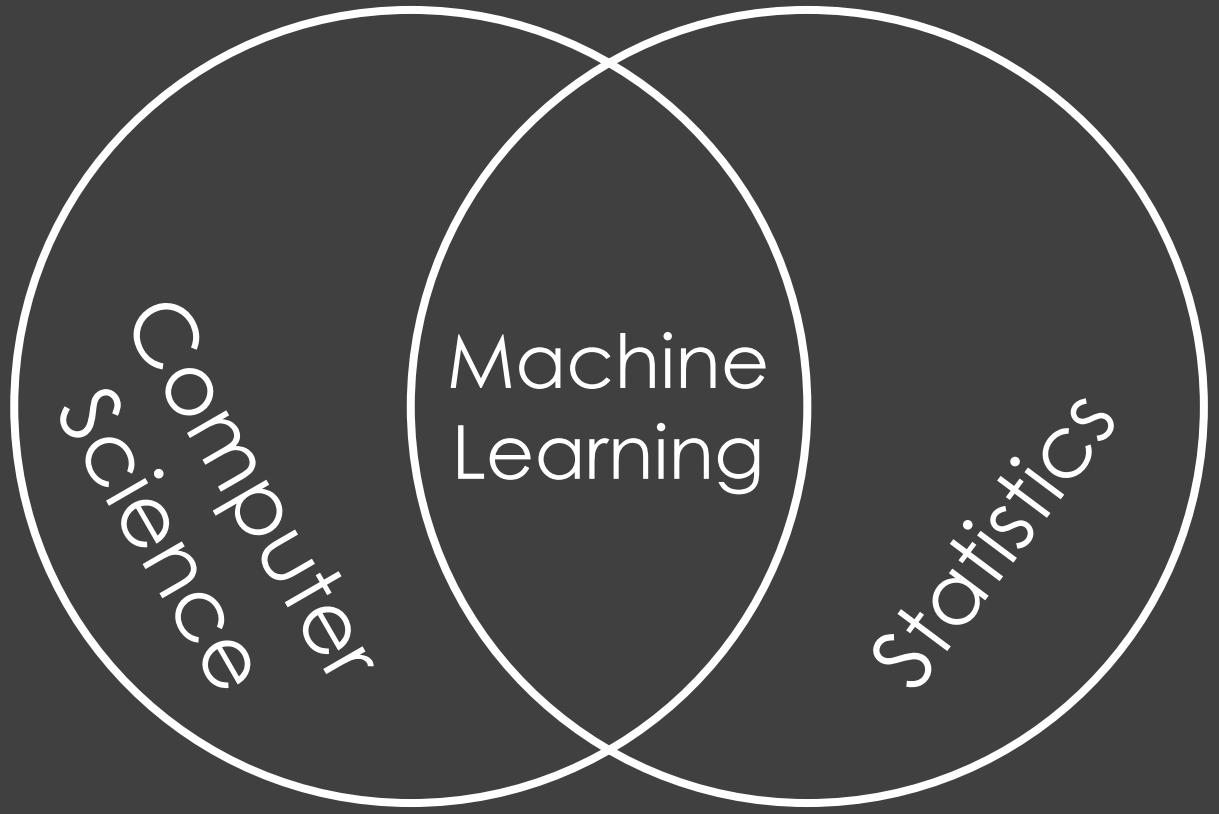
solve problems — **Engineering**

- *Data science is fundamentally interdisciplinary*

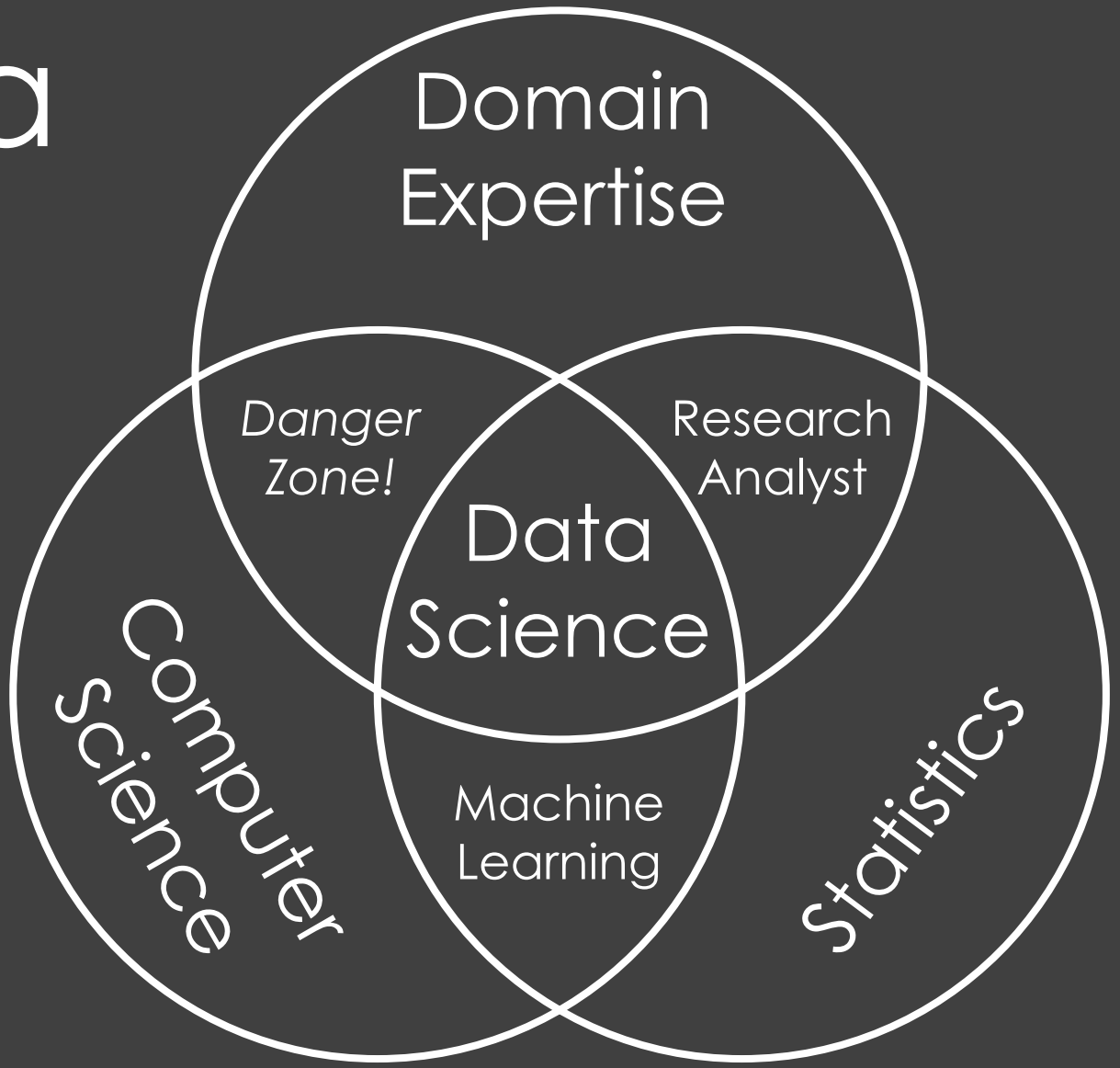
Skills of Data Science



Skills of Data Science



Skills of Data Science

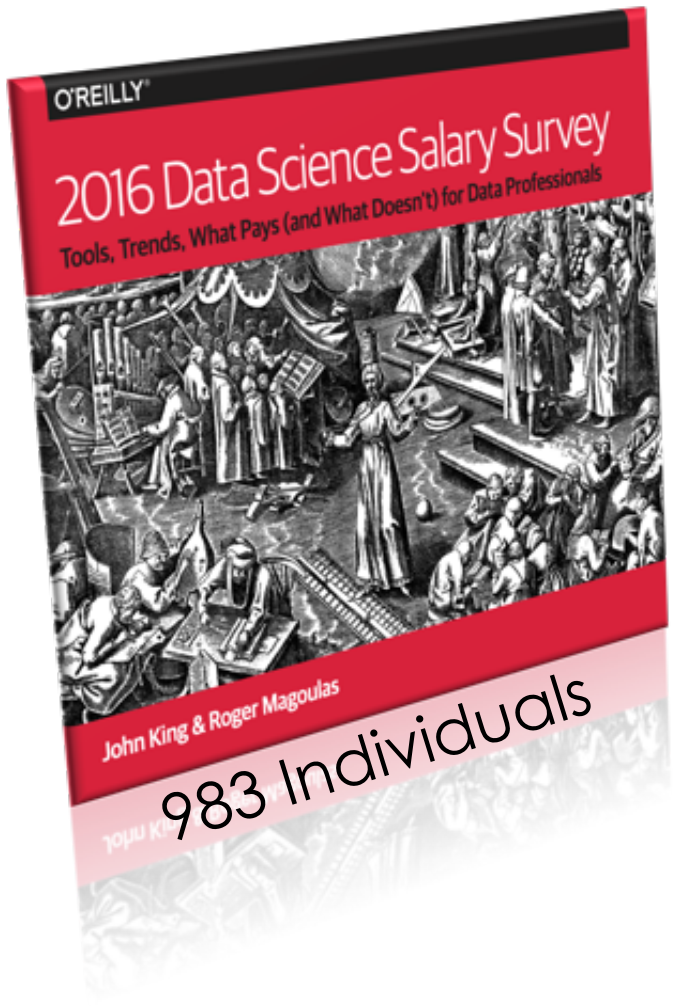


Drew Conway's Venn Diagram of Data Science

What does it mean to be a data scientist today?

How can we answer this question?

Data Science Surveys

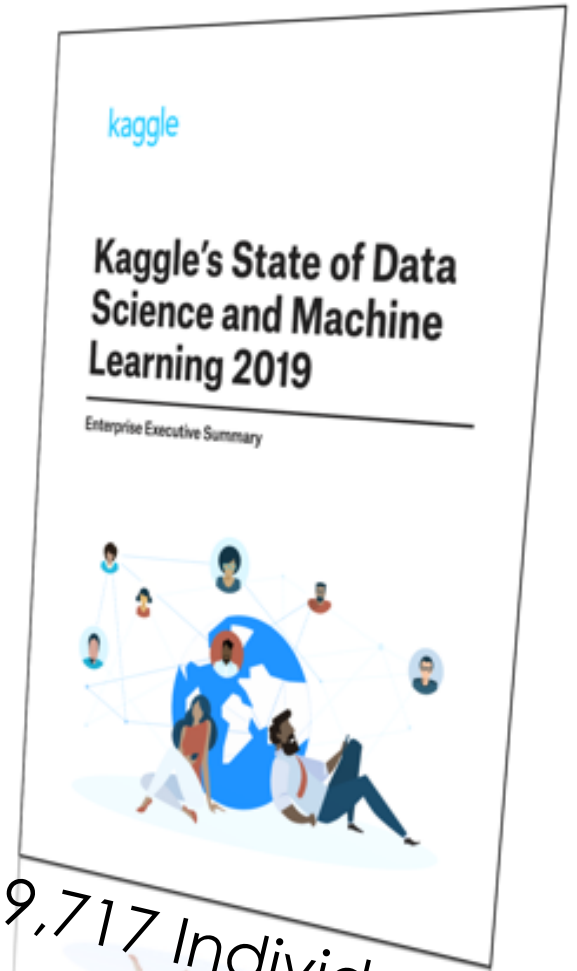


983 Individuals

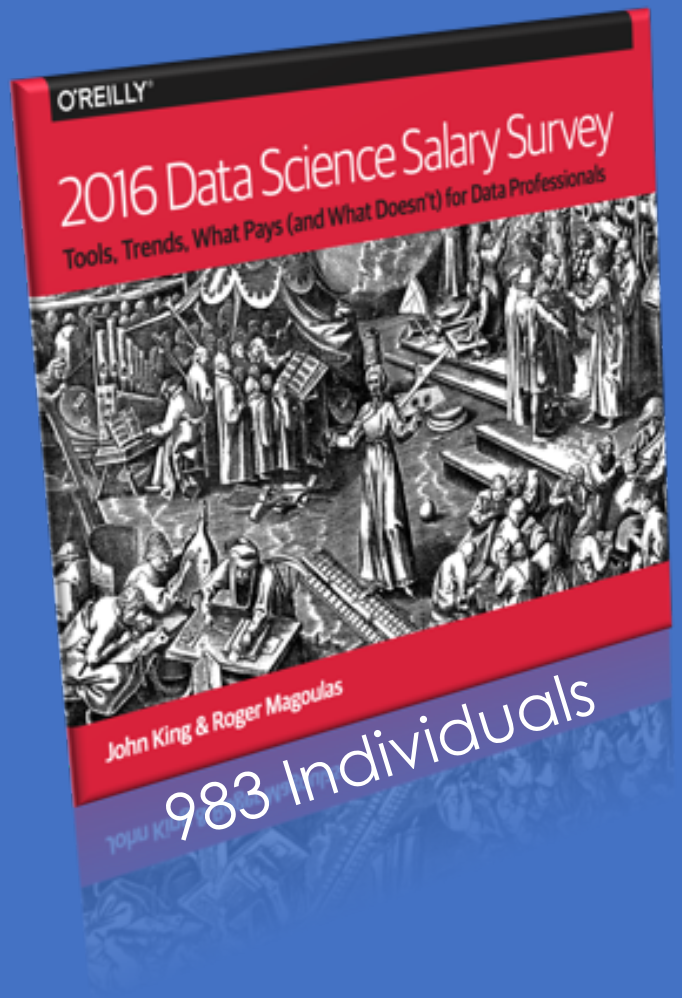
Asked people involved in data science events and competitions to complete an online survey

Self reported → Selection bias!

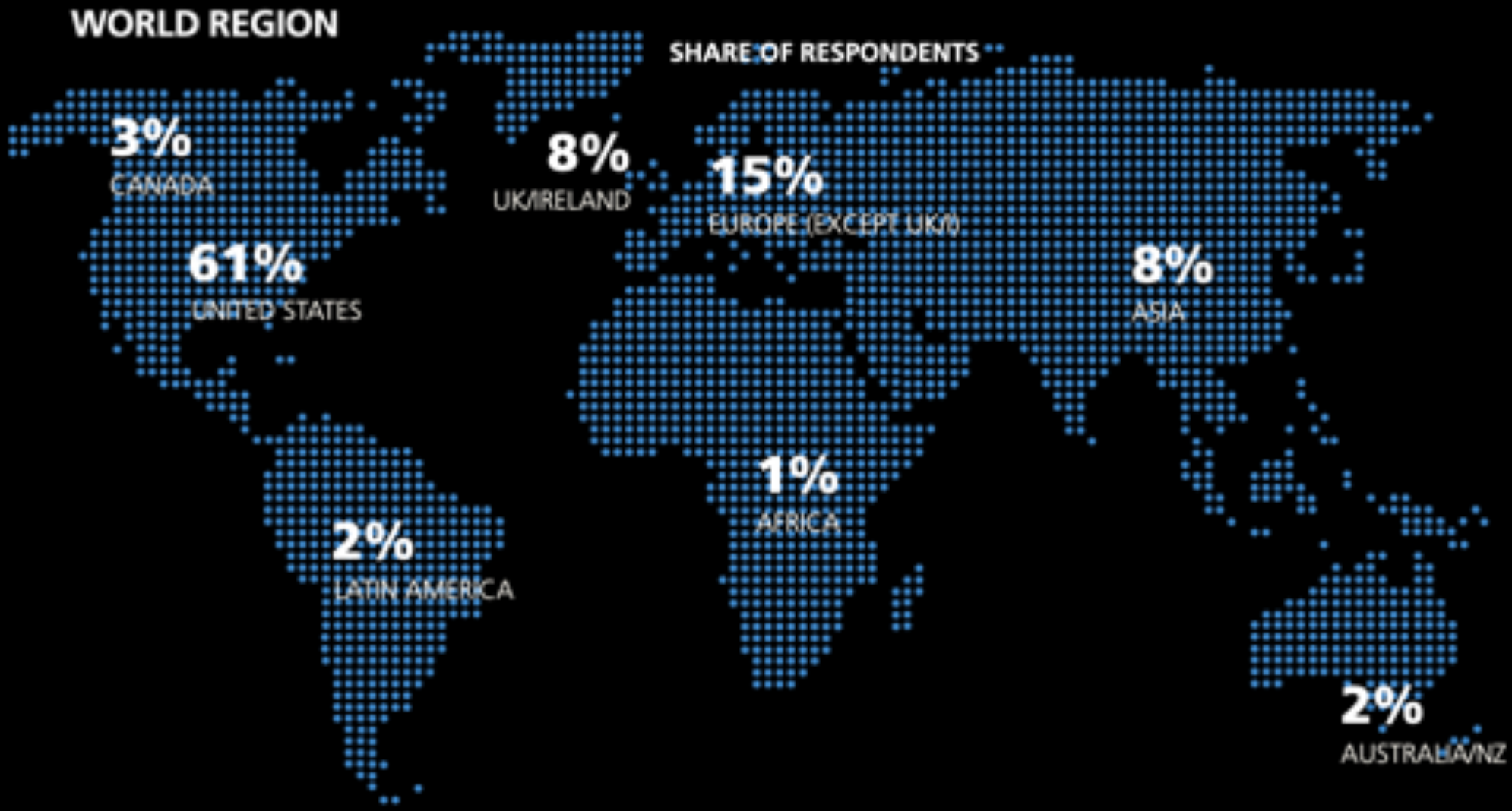
Still somewhat interesting …



19,717 Individuals

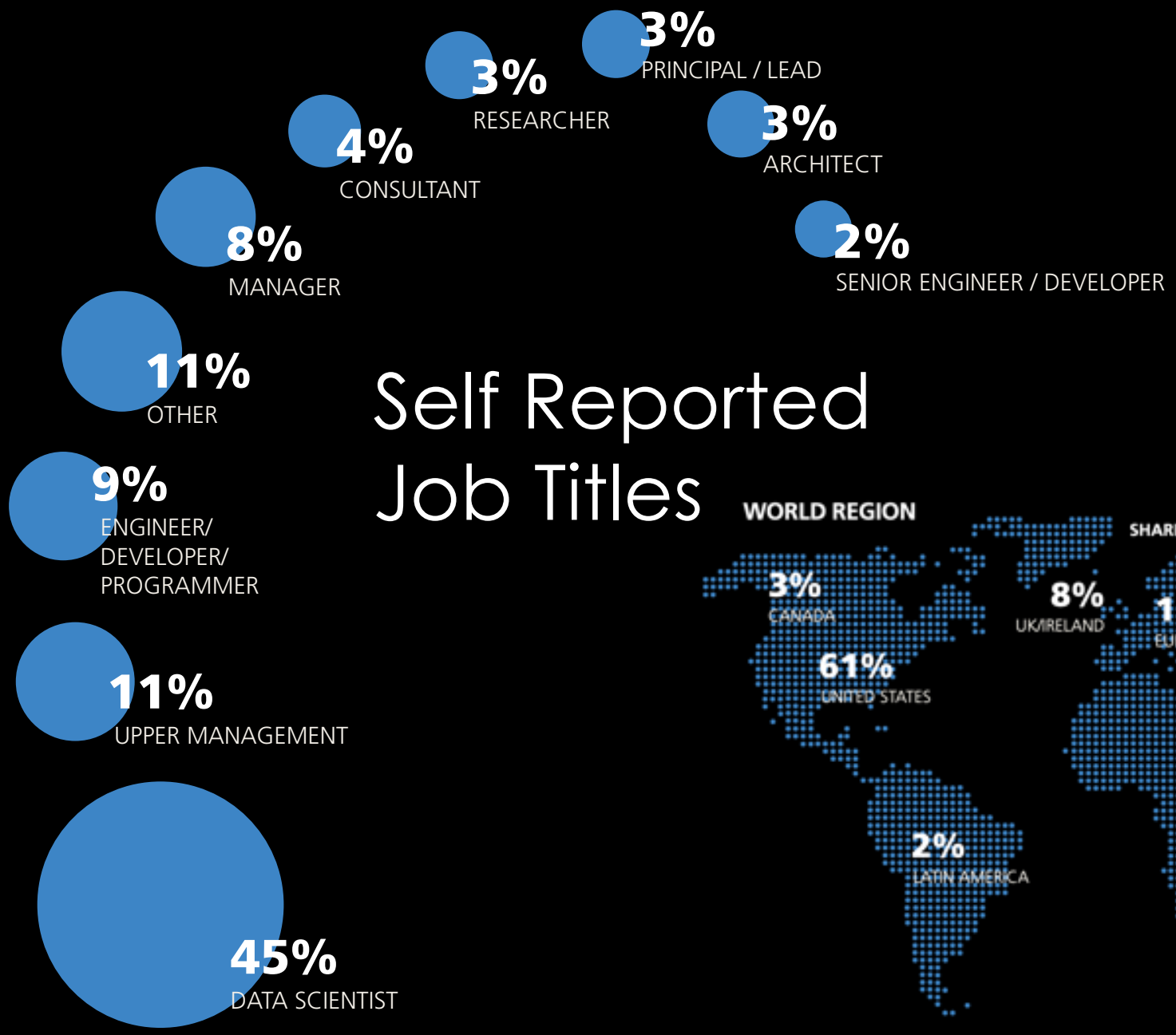


Demographics (2016)

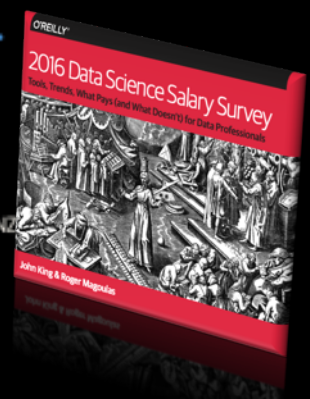


Survey selection bias …

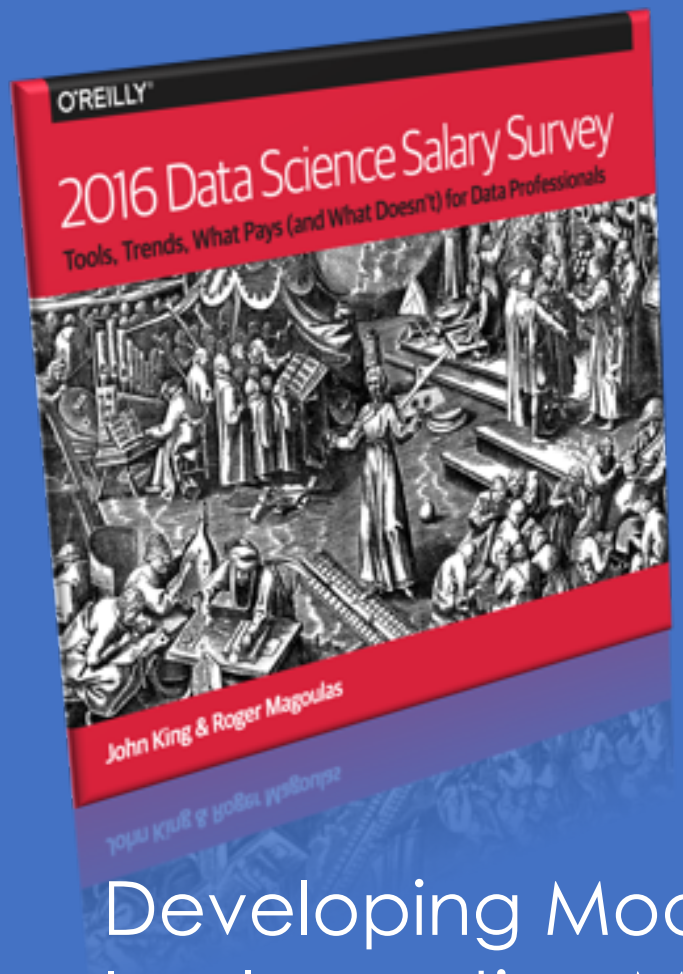
Self Reported Job Titles



Mix of Data Scientists, Management, and Engineering



Survey selection bias …



What do they do?

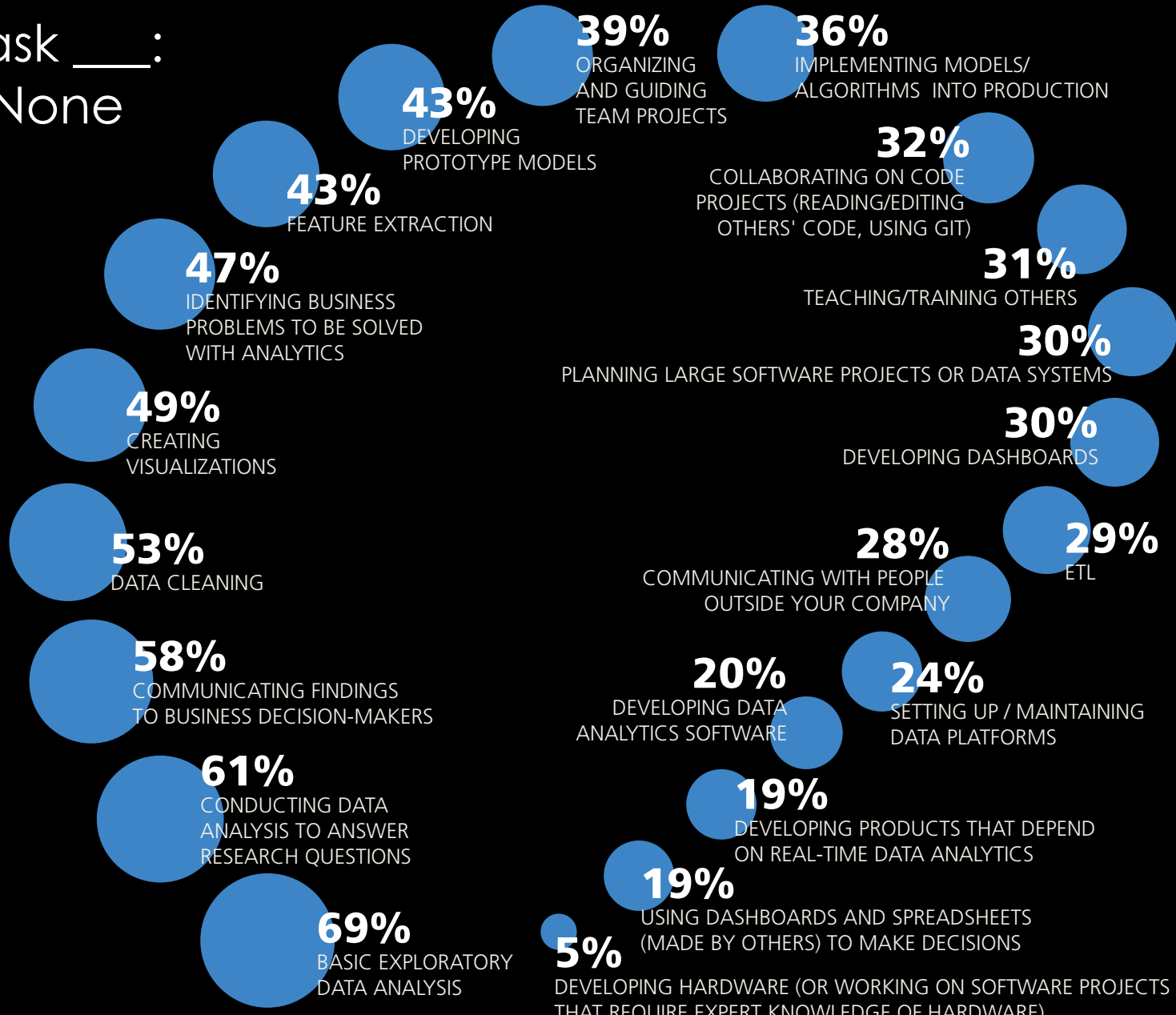
How involved are you in task ___:
(a) Major, (b) Minor, (c) None

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

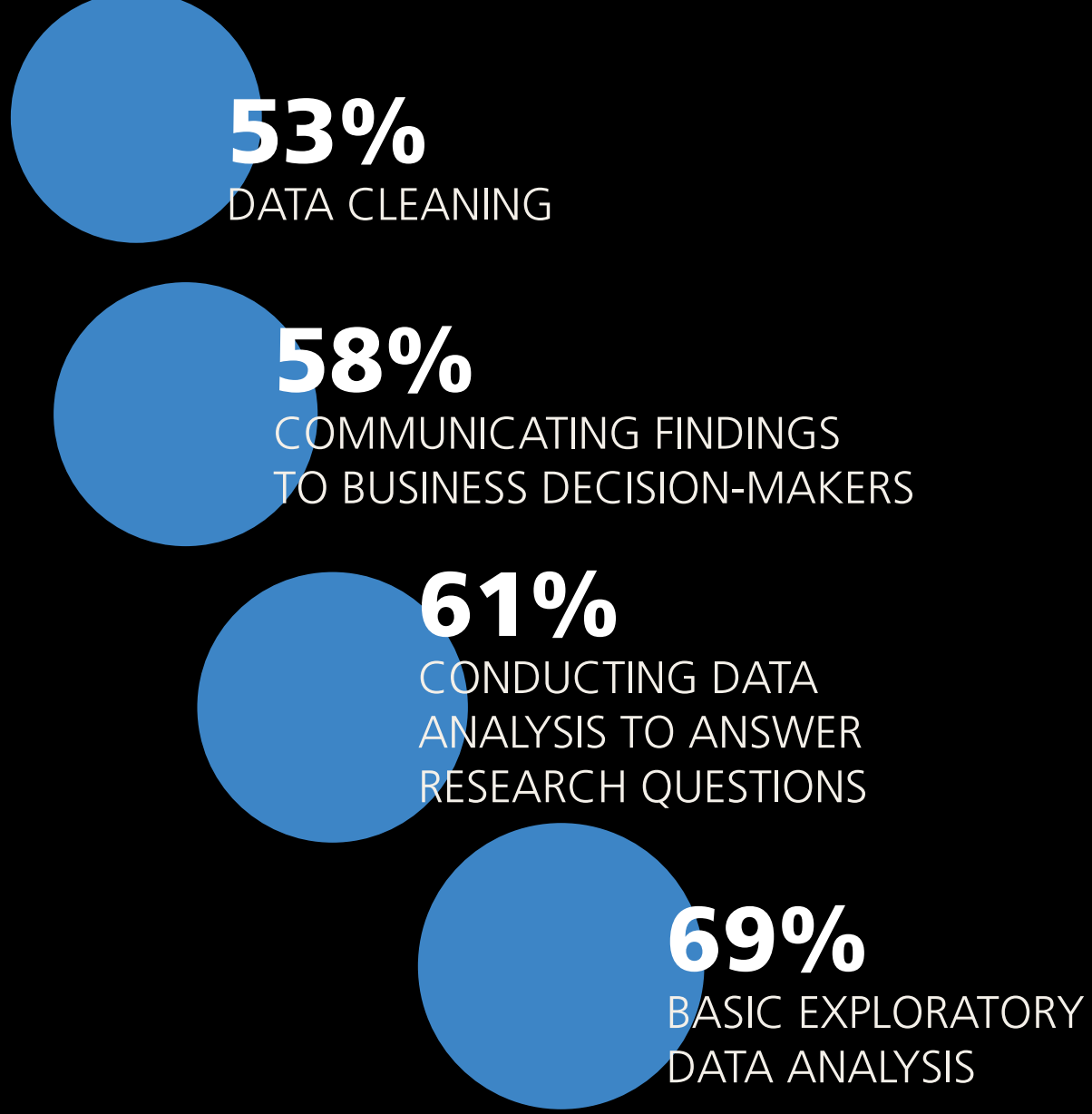


How involved are you in task ___:
(a) Major, (b) Minor, (c) None

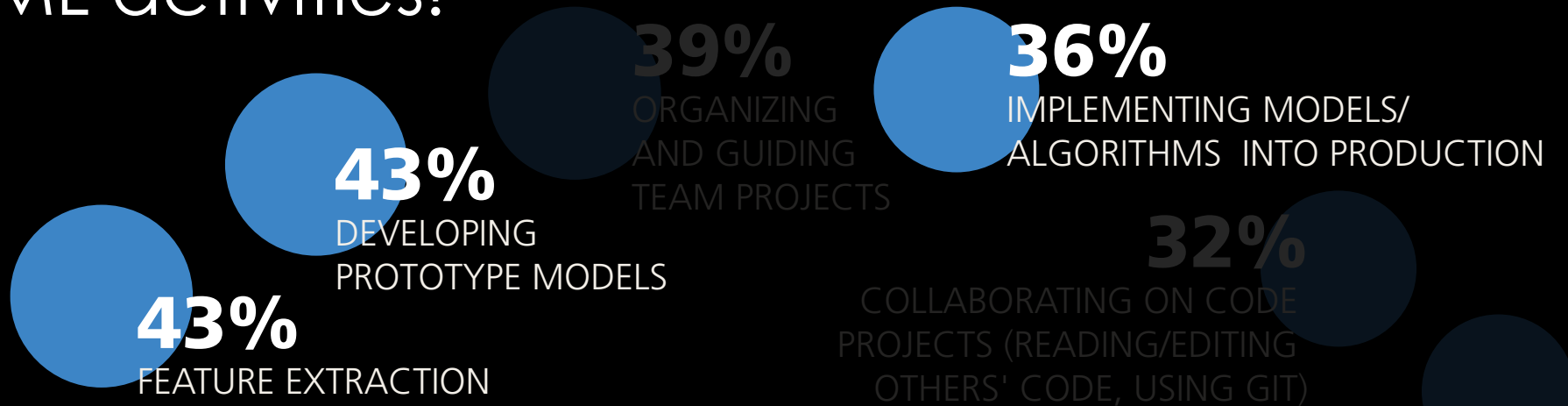
Are the top items surprising?

Data Cleaning ☹

Where are Modeling / Prediction?



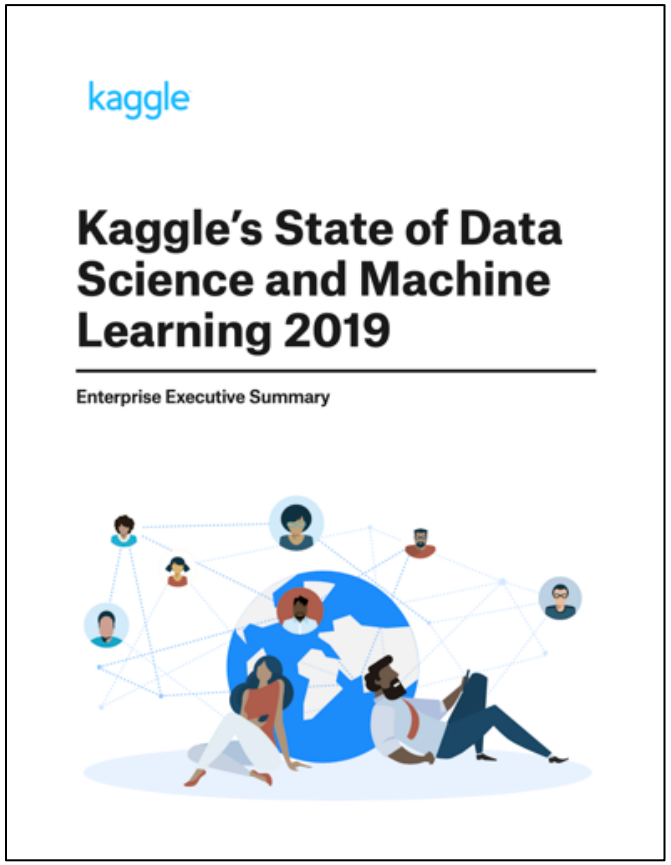
Less than half of respondents had major involvement in ML activities!



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

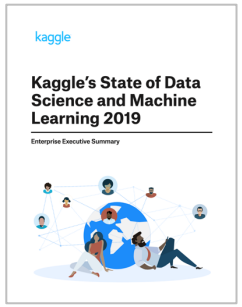
- major engagement → $7.4K boost

Kaggle Data Science and ML Survey



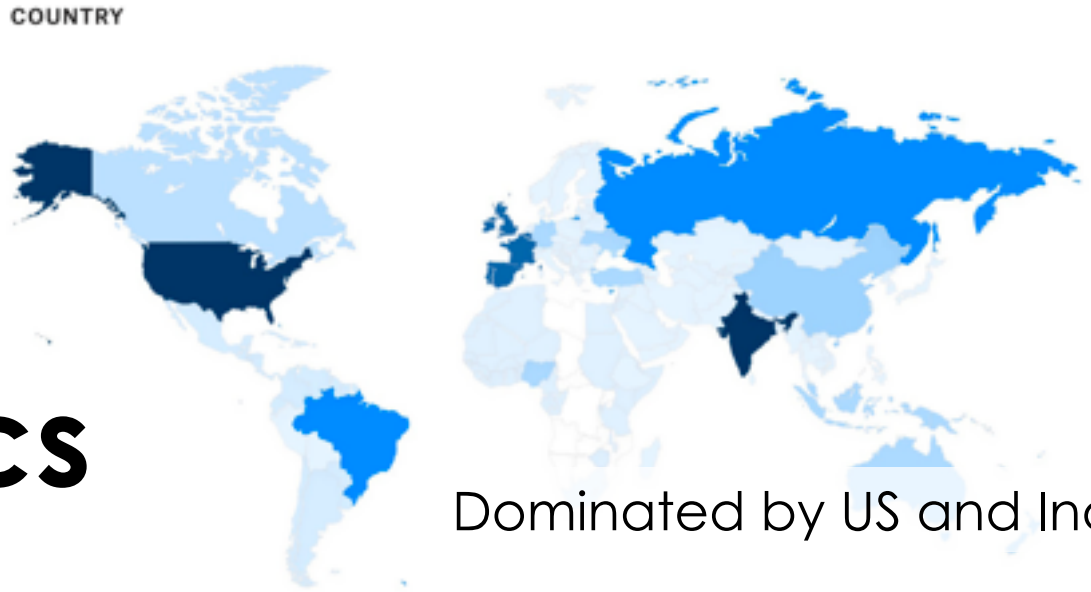
19,717 Individuals

- More recent: **2019**
- Over **19K participants**
- Survey Biases?
  - Focused on ML centric data science community.
- Slides focus on 21% with **Data Scientist** title
- For more details
  - **Raw Data:** https://www.kaggle.com/c/kaggle-survey-2019
  - **Executive Summary Report:** https://www.kaggle.com/kaggle-survey-2019

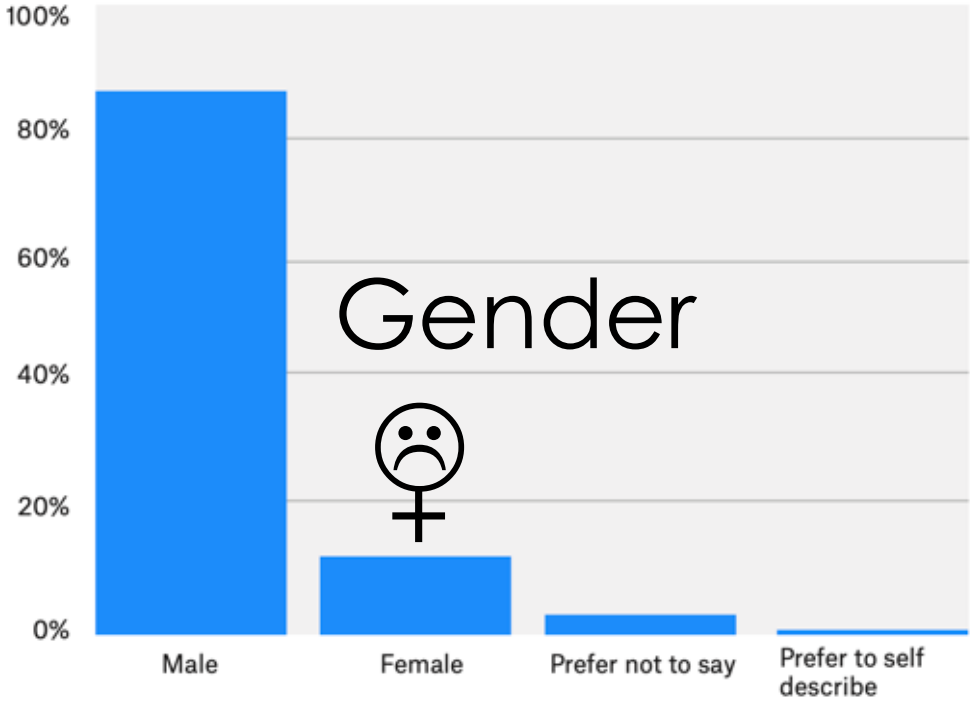


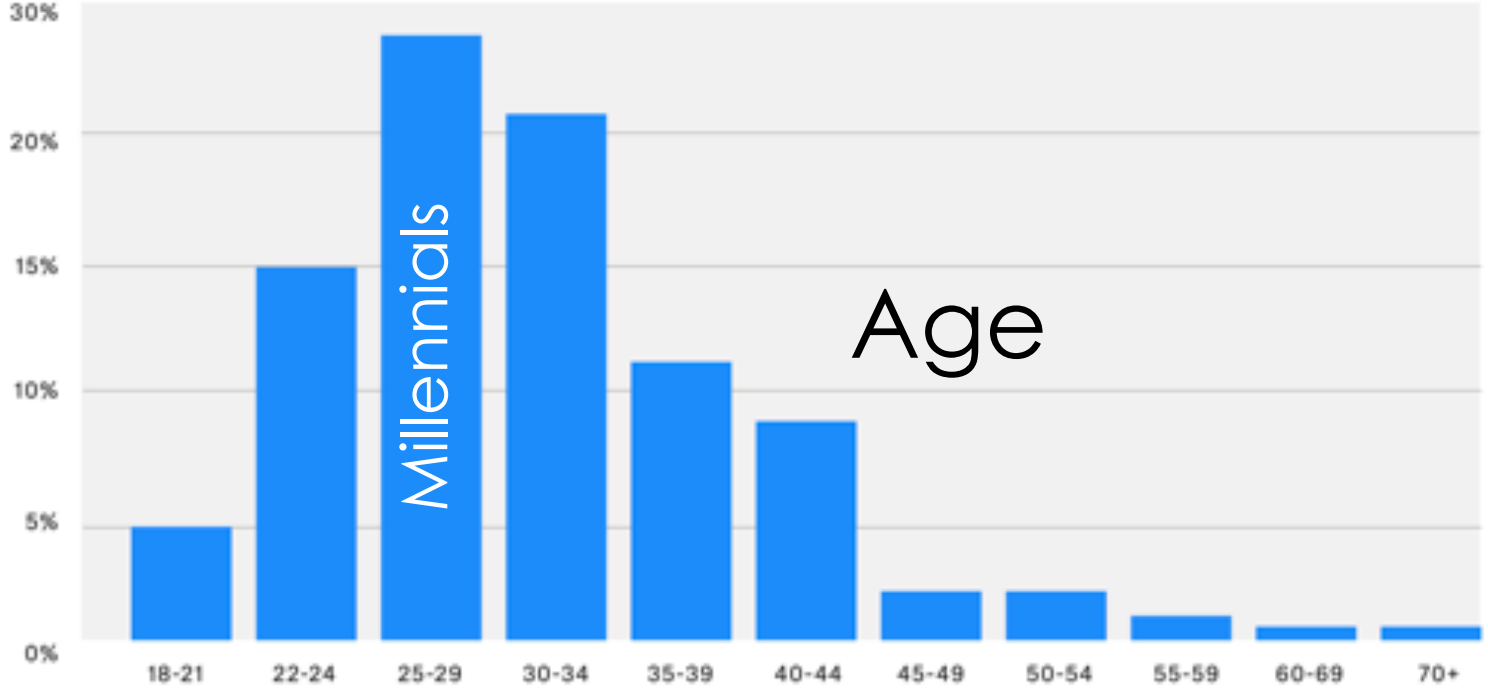
19,717 Individuals

Demographics

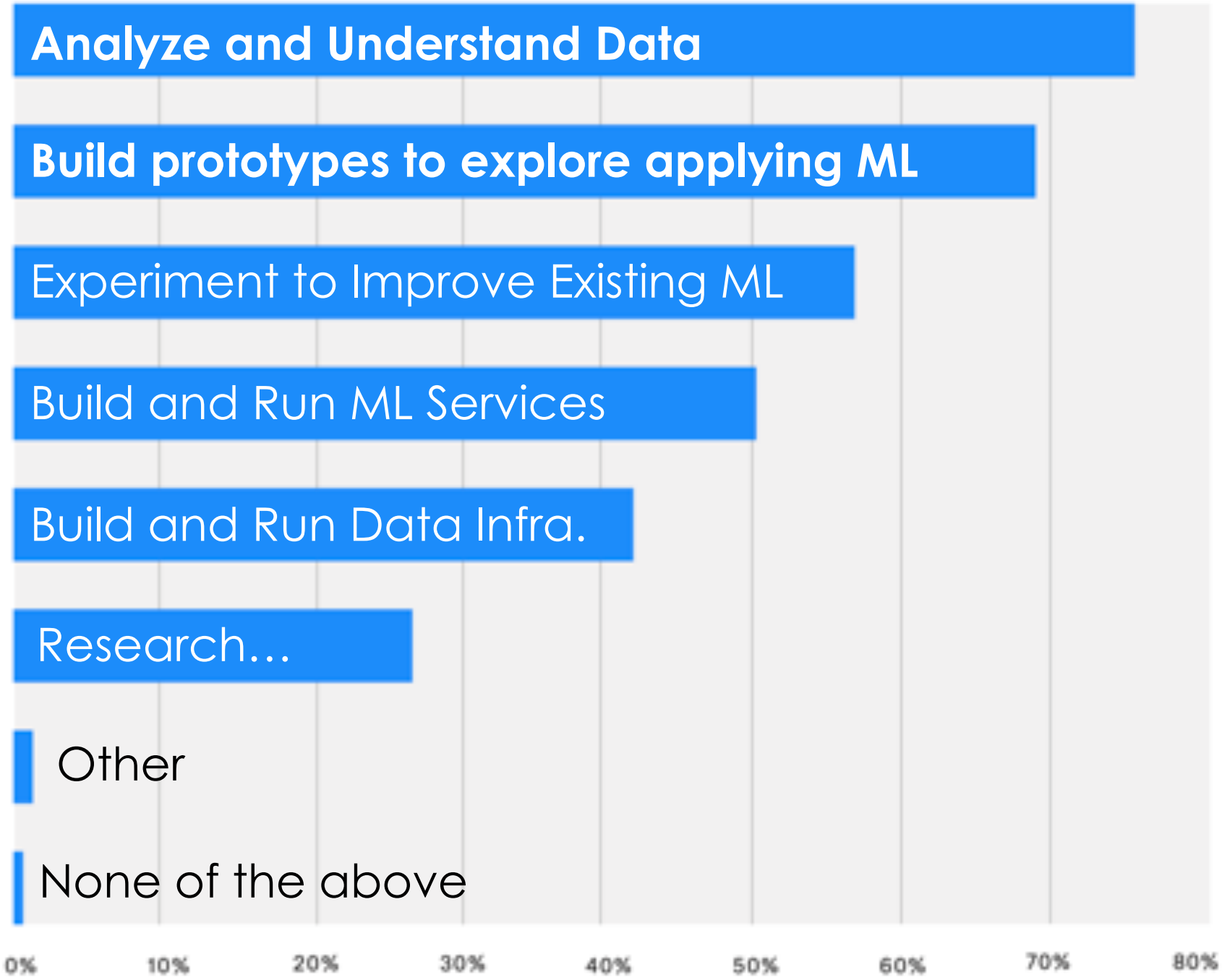


Dominated by US and India





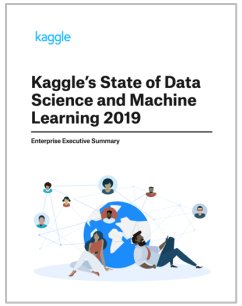
What do Data Scientists Do?

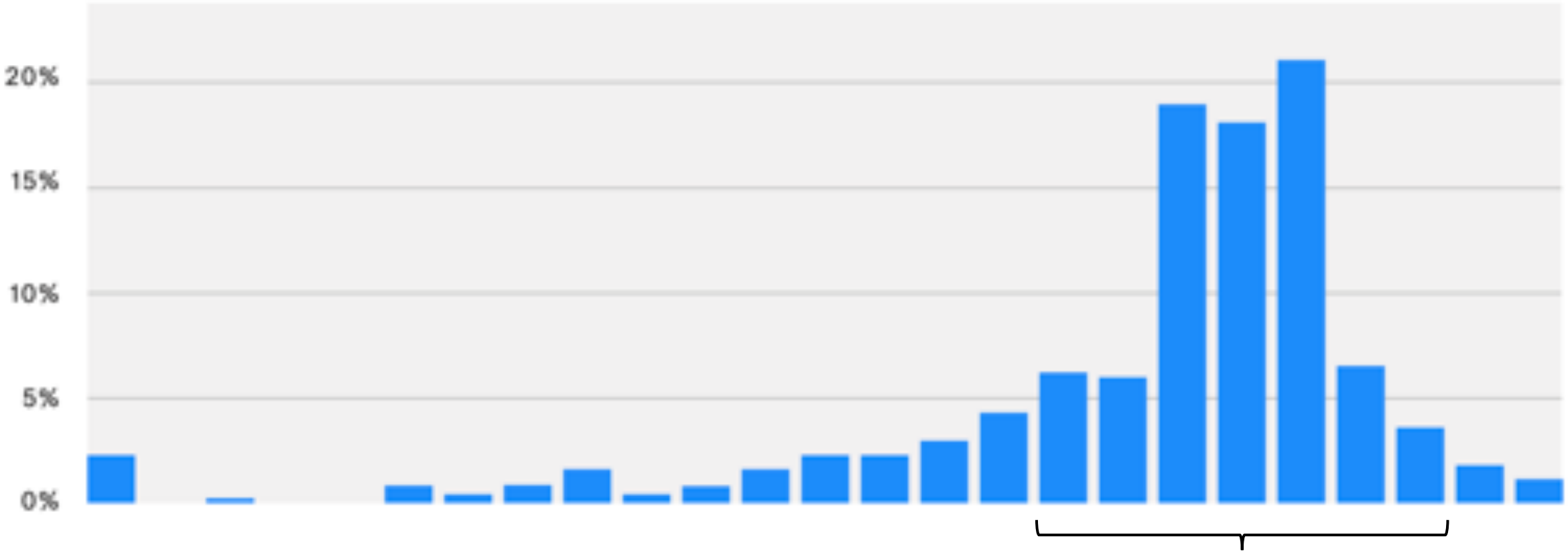


19,717 Individuals

US Salary Distribution

19,717 Individuals

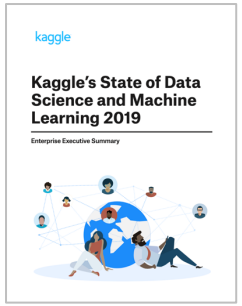


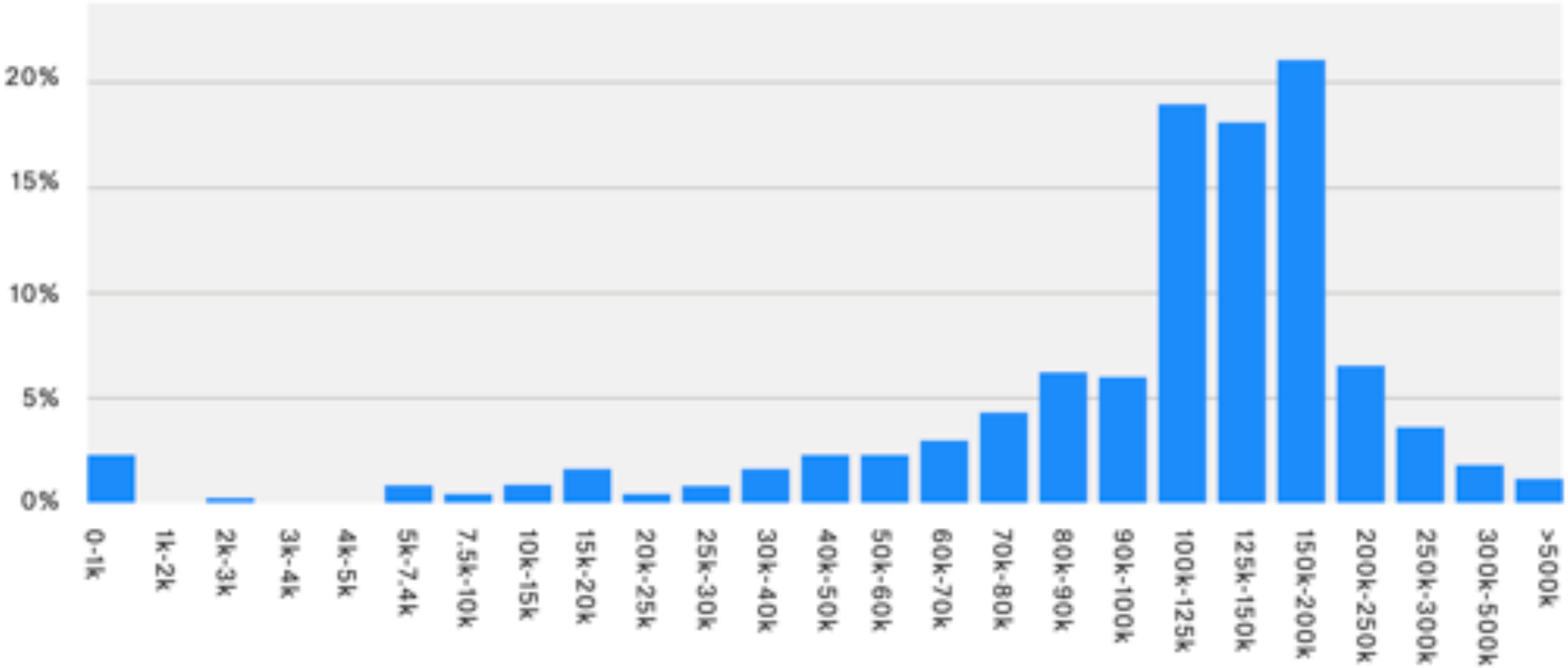


How much do they make?

US Salary Distribution

19,717 Individuals





What are your *goals* for Data 100?

- 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

Course Information

The Data 100 Team

Web Resources

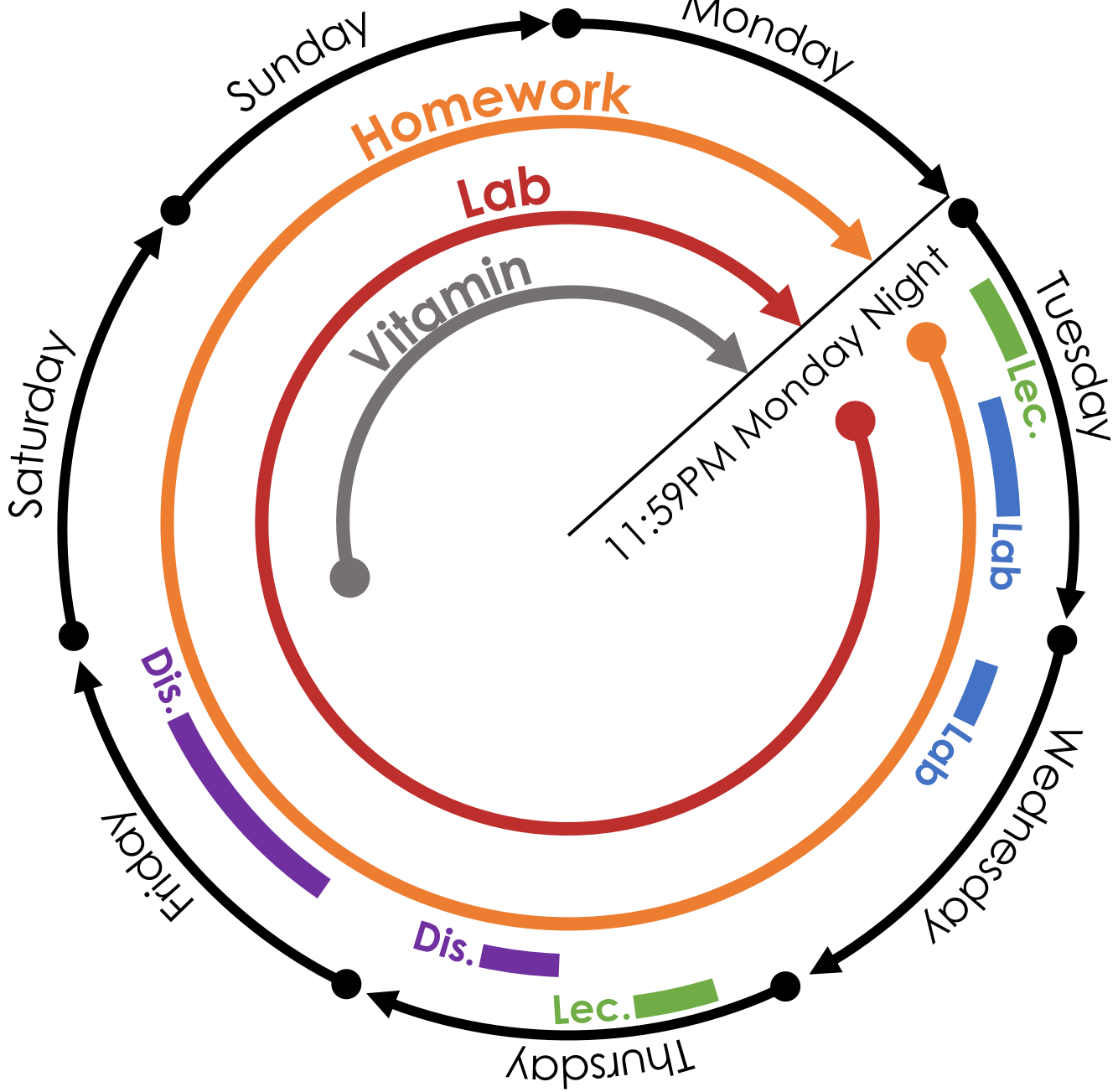
- The course website is: http://www.ds100.org/sp20/
  - Linked from http://ds100.org
- View the office hour schedule on the calendar here: http://www.ds100.org/sp20/calendar/.
- All communication will be through Piazza: https://piazza.com/berkeley/spring2020/data100
  - Used to provide help with **assignments** and **course concepts**.
  - If you have a private question, make a **private post**.
  - Please **do not post code publicly!**
- Online textbook: http://www.textbook.ds100.org/
  - We will provide additional study materials throughout semester

Prerequisites

- Official Prerequisites for this course:
  - Completion of Data 8.
  - Completion of CS61A or CS88.
  - Co-enrollment in EE16A or Math54
- These are **not strictly enforced** but we will not be teaching:
  - How to use Python
  - How to use Jupyter Notebooks
  - Basic Inference from Data 8
  - Basic Linear Algebra
- Homework 1 will help calibrate your background
  - Do Hw1 and skim the Data8 textbook: https://www.inferentialthinking.com

Weekly Cycle

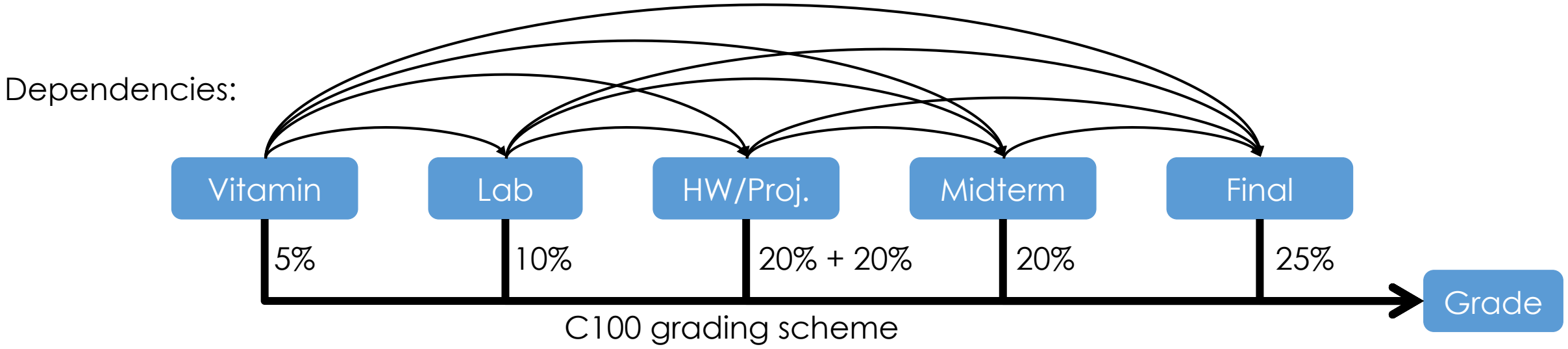
- Everything* is due on Monday nights.
- **Homework** covers big ideas and helps prepare for exams
  - **Projects:** 2-week homework
  - **Discussion** will often cover questions closely related to homework
- **Lab** prepares for HW
  - walk through part of lab
- **Vitamins** ensure you keep up with lecture (~10 minutes)



*unless there is a stated change in the schedule.

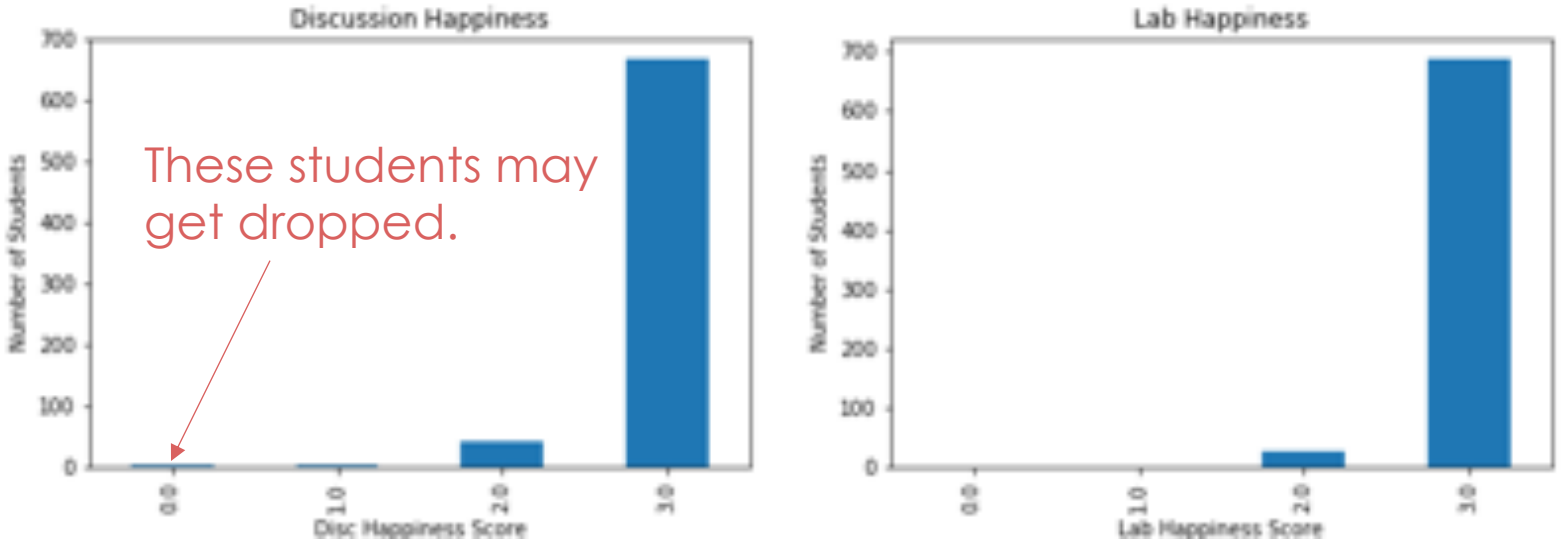
Grading Scheme

- See http://www.ds100.org/sp20/policies/ for details
  - Grading Schemes (C100 and C200)
  - Late Policy
- Goals for the current policy
  - We want you to **learn the material** and thus get a **good grade**



Section and Lab Times

- We will compute* an optimal assignments for lab & section
- Complete form: https://forms.gle/gupHuYomhizuf2B96
  - by midnight tonight!
- **Be honest** about your preferences and availability
  - If we are unable to find a time you may be forced to drop
  - You can **update** your preferences
- **Waitlisted** students should also fill in preferences



*Assignment Code

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** (☺)
- ***You must write your solutions individually***

Don't Cheat

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

Exams

- There is **one midterm** and **one final**
  - Midterm: **tentatively** (8:30PM --10:00PM) on Monday March 9th
  - Final will cover everything in the semester
- We are in the process of scheduling **makeup exams**
  - They will likely be on the **same day**
- Want to do well on the exams?
  - Watch or attend **lectures**
  - **Attend lab** and **discussion** and do the homework and projects
  - Review **previous exams**

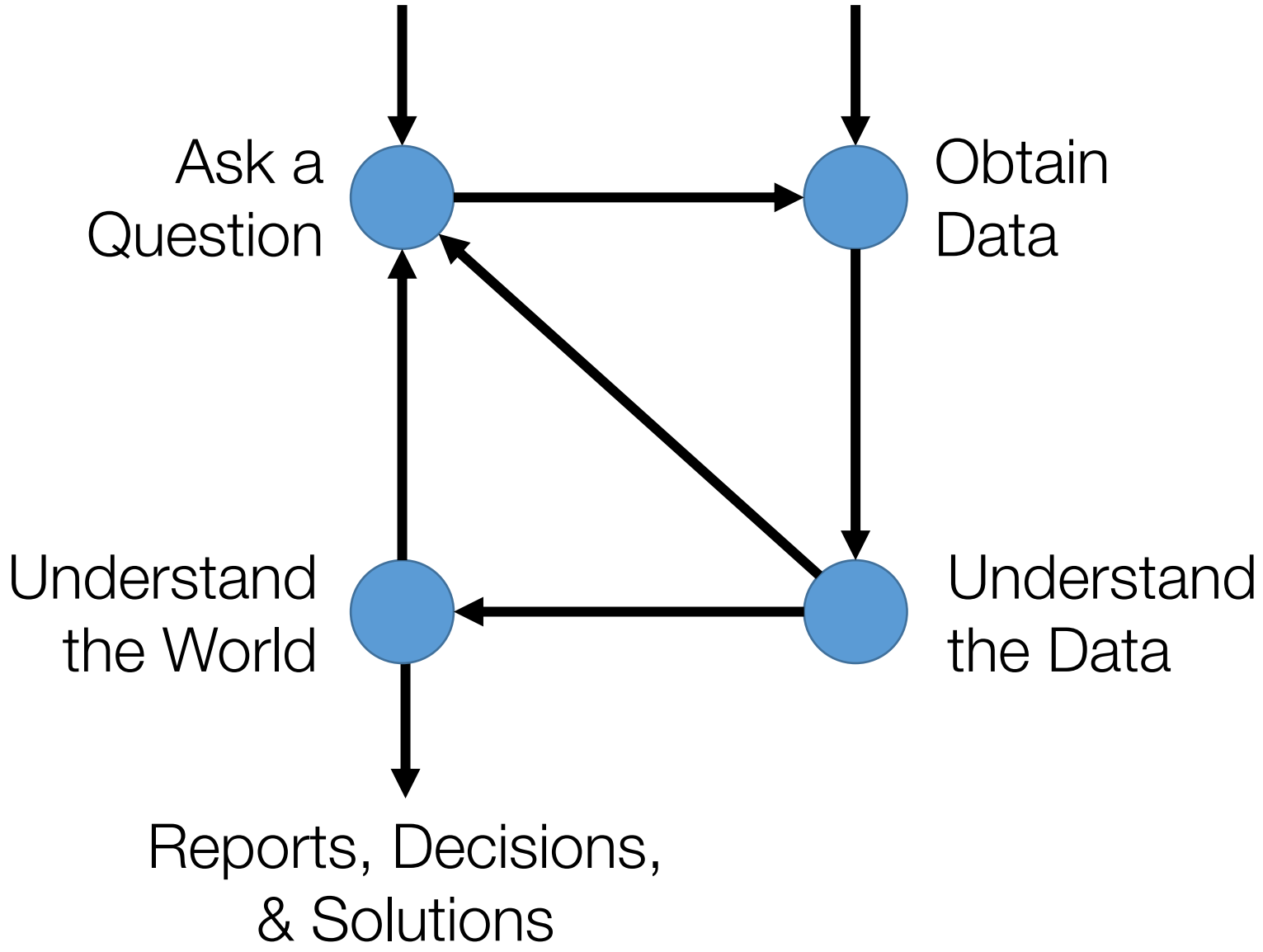
Big Class and the Wait List

- Not everyone will be able to attend lecture in person ☹
  - **Wheeler Lecture Hall: 744**
    - We cannot exceed occupancy → violates fire code
  - Lectures will be recorded and posted along with slides
- Current Enrollment Data:
  - **Undergraduate (C100):** 1105 slots, 192 on waitlist
  - **Graduate (C200):** 86 slots, 22 on waitlist
- "Will I get in?" Difficult to answer.
  - The rule of thumb is that **roughly 10% of students drop.**
  - All waitlisted students should do the assignments.
  - Concurrent enrollment → unlikely to get into class ☹
    - Consider taking Stat-131a instead

What will you learn in Data 100?

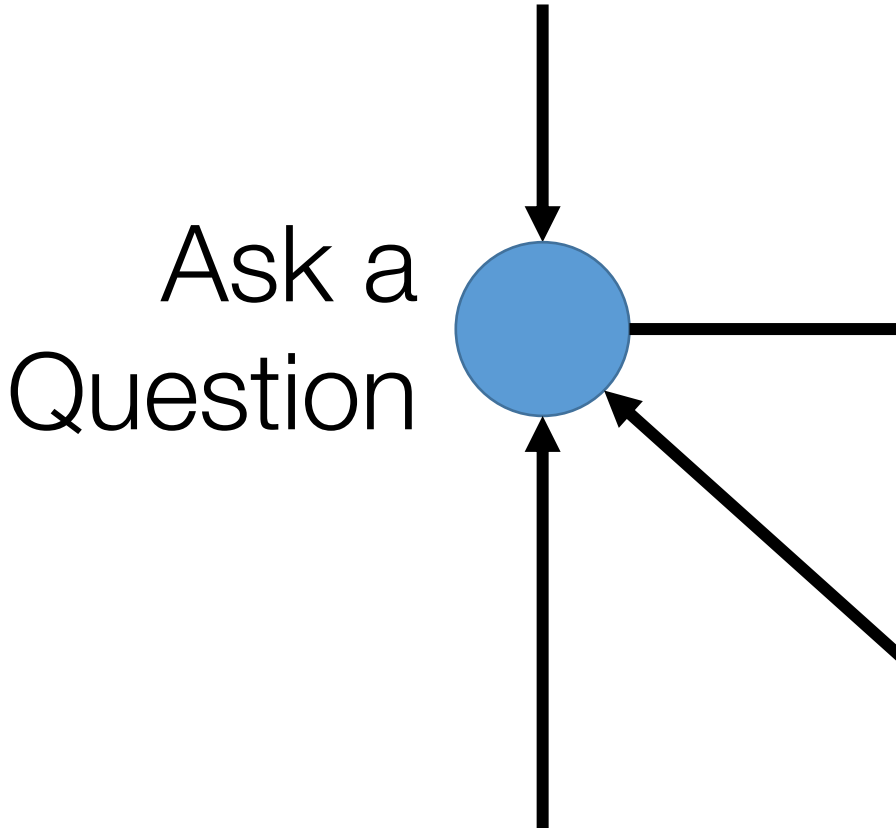
Data Science Lifecycle

High-level description of the data science workflow

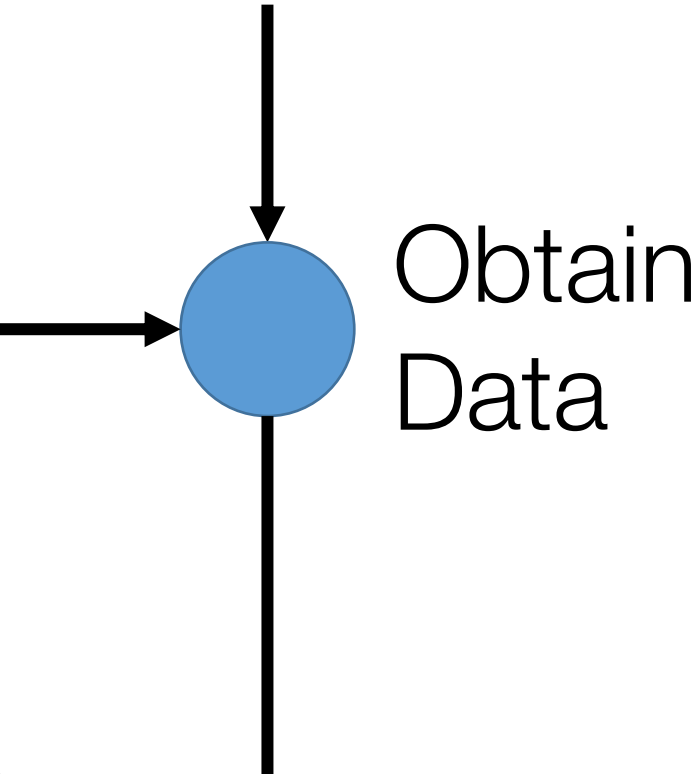


Question / Problem Formulation

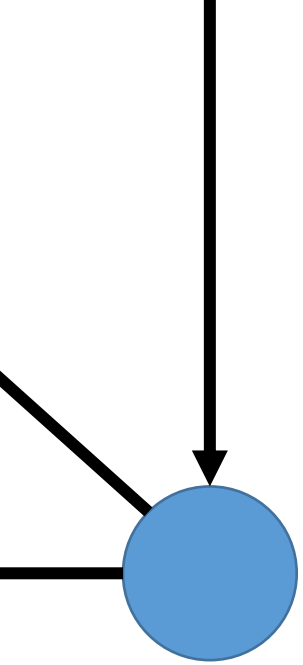
- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics of success?



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?

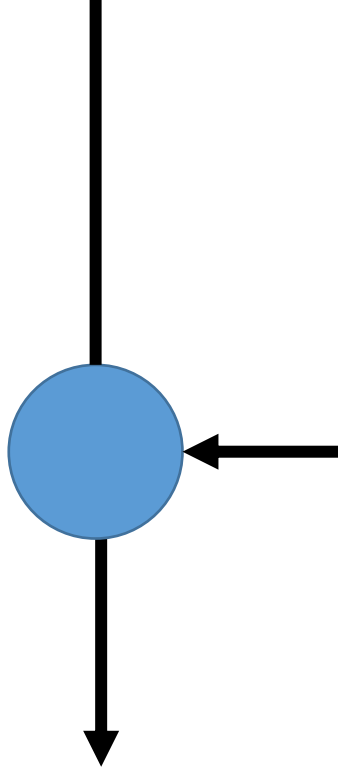


Understand the Data

- 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

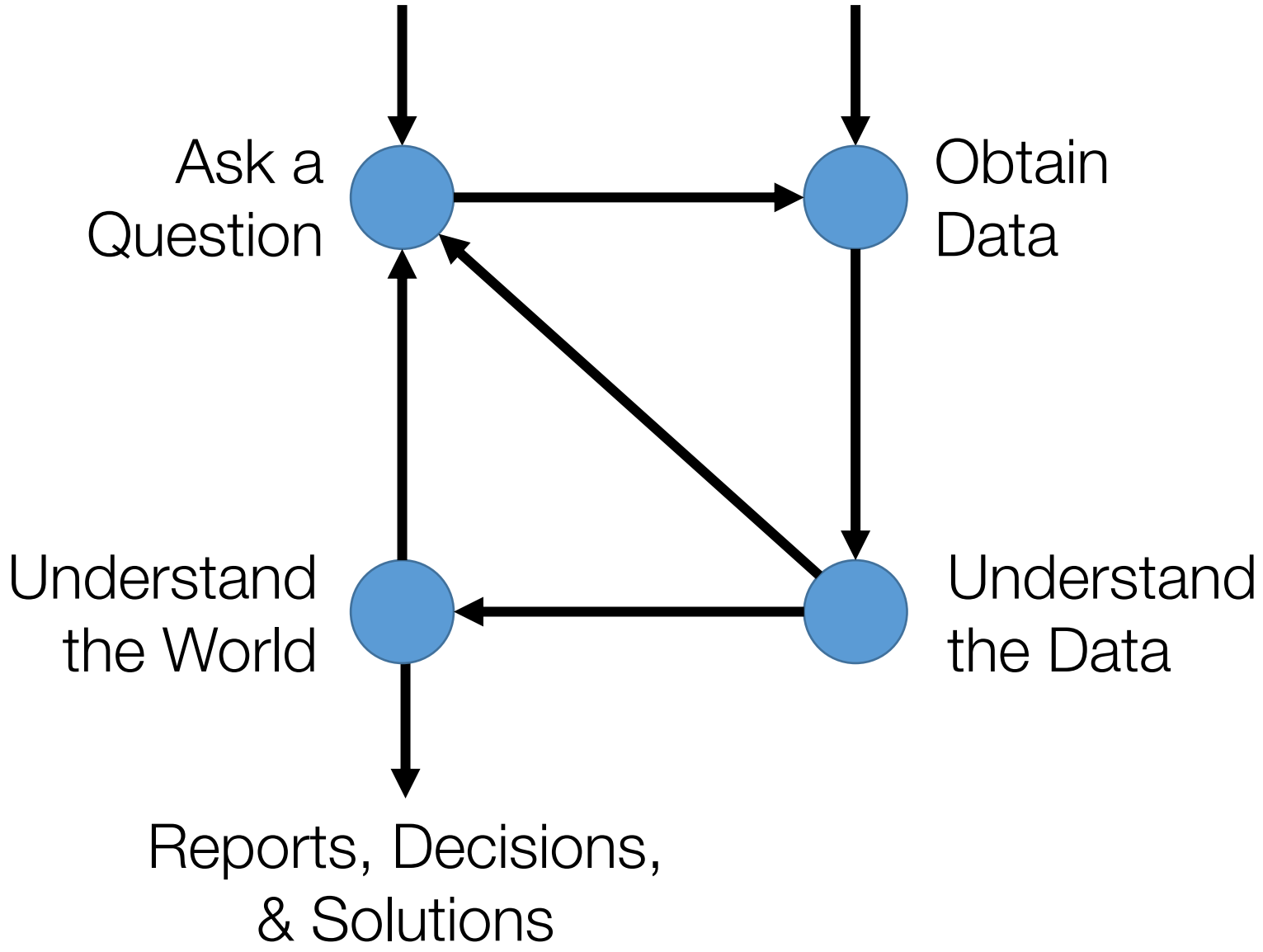
- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



Prediction and Inference

Data Science Lifecycle

High-level description of the data science workflow



Topics covered in Data 100

- Data collection and sampling
- Data cleaning and manipulation
- Regular Expressions
- SQL and Enterprise Data Management
- Exploratory Data Analysis & Visualization
- Model design & loss formulation
- Optimization and Gradient Descent
- Ideas from Probability and Statistics
- Least Squares Regression
- Logistic Regression
- Clustering
- Dimensionality Reduction
- Decision Trees
- Feature Engineering
- The Bias - Variance Tradeoff
- 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 technologies**
  - Jupyter notebooks, pandas, numpy, sqlite, matplotlib, altair?/plotly, pytorch, scikit-learn, …
- *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)

Intermission

5 Minute Break.

Ask a neighbor:

What is your name?

tabs or ***Spaces*** …?

What do statisticians and pirates have in common?

Entirely optional survey we will use for demo.

Pirates say