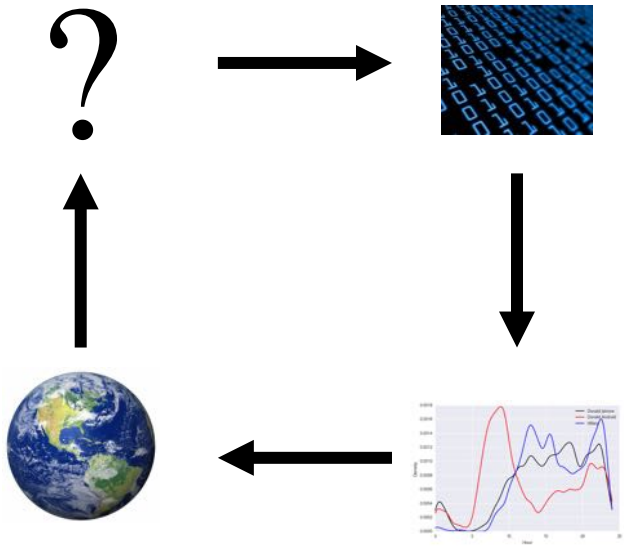
Data Science 100

Principles & Techniques of Data Science

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu



Enrolling & the Waitlist

We apologize for issues with the waitlist and CalCentral ☹

- I have notified 96 students that they have a spot
  - Several are having CalCentral issues
  - Several are no longer interested in taking the class
- At 11:00 AM I notified 25 additional students that they are at the top of the wait list
- If you are not in these groups please try to visit office hours or section and let us know you are still interested.

Questions for Today

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

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

__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

More about eating in SF on Thursday …

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

Green-lighting *House of Cards*

$100M: Buy House of Cards?

"We can look at consumer data and see what the appeal is for the director, for the stars and for similar dramas." -- *Steve Swasey* [Former VP @ Netflix]

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

PREMISE

We map reality on the ground.

Combines crowd-sourcing and machine learning
to better understand the developing world.

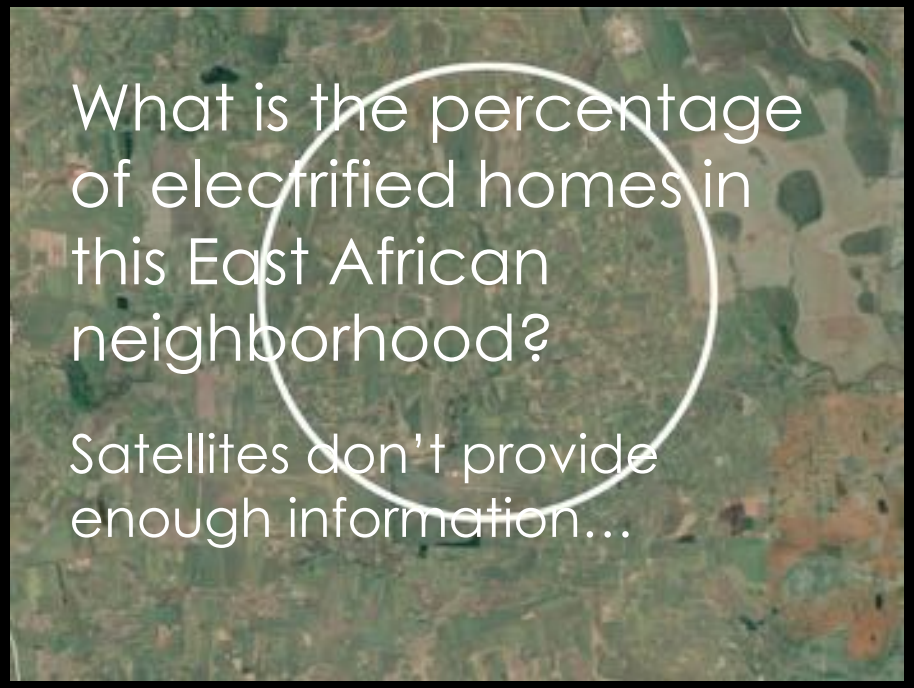
What is the percentage of electrified homes in this East African neighborhood?

Satellites don't provide enough information…

PREMISE

We map reality on the ground.

Combines crowd-sourcing and machine learning to better understand the developing world.



Pay villagers to take pictures of the sides of houses

PREMISE

We map reality on the ground.

Combines crowd-sourcing and machine learning to better understand the developing world.

Use machine learning and people to identify wiring on homes

PREMISE

We map reality on the ground.

Combines crowd-sourcing and machine learning to better understand the developing world.

Answer policy questions

PREMISE

We map reality on the ground.

Helping the Philippines deliver Universal Health Care by enforcing the tax on sales of alcohol and cigarettes.

Gapminder.org

Aggregating data to help understand healthcare, poverty, and education in the developing world.

Great example of data aggregation and visualization to tell a story

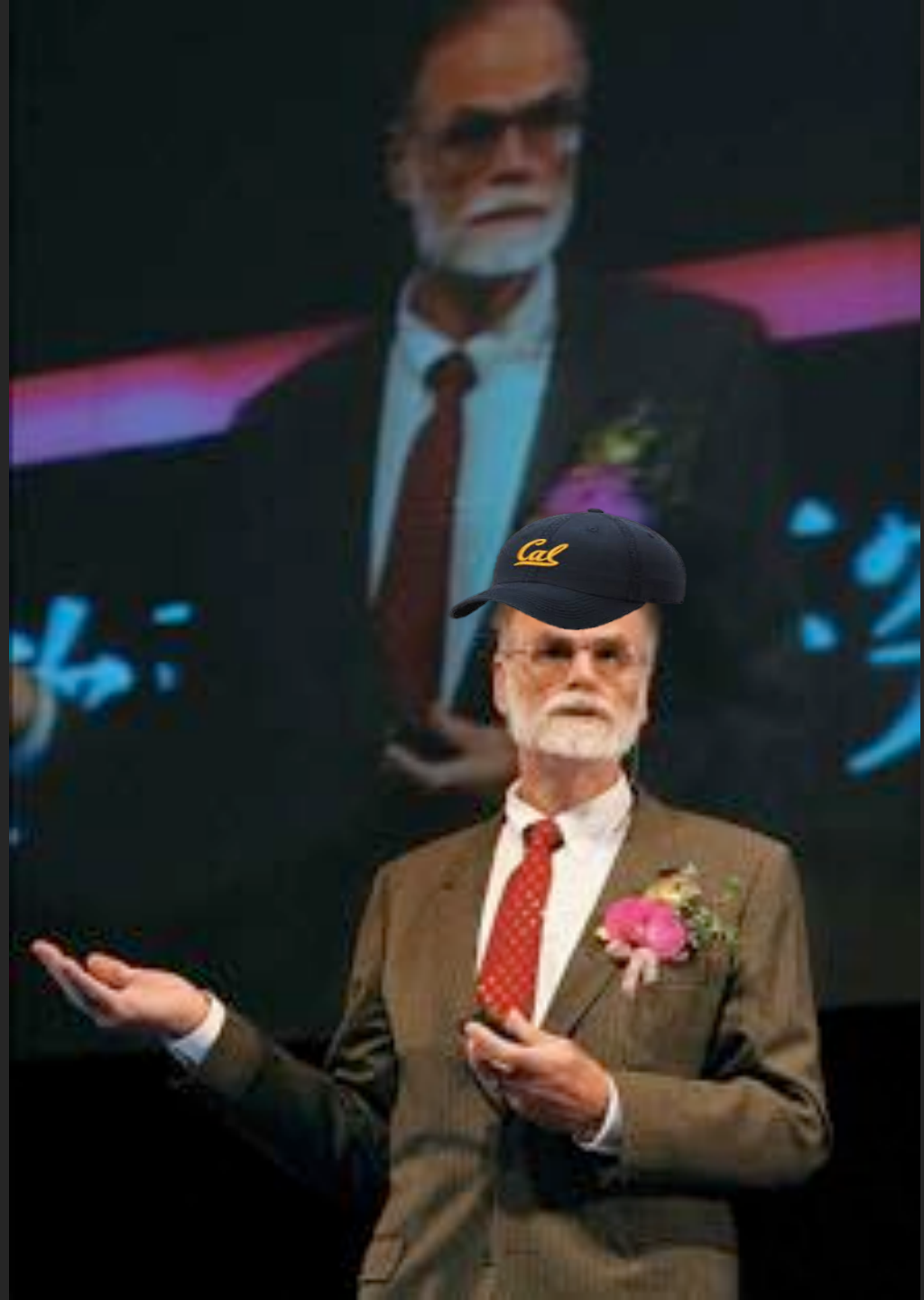
http://www.gapminder.org/world

https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen#t-198350

Data Science & *Science*

Jim Gray

Turing Award Winning Computer Scientist & Cal Alum.



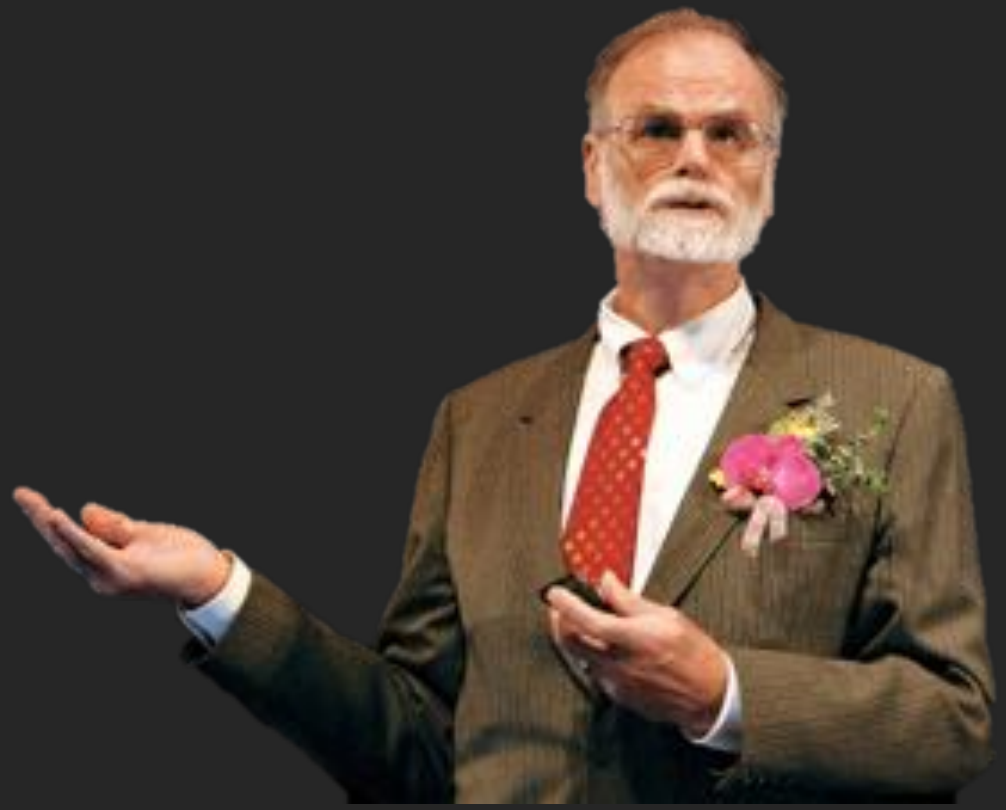
Introduced the idea of the **Fourth Paradigm** of Science

Experimental

Theoretical

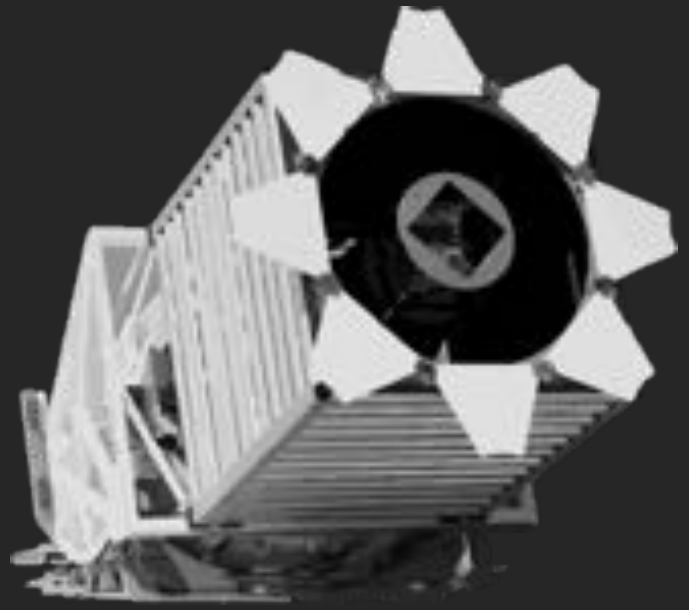
Simulation

Data Intensive



Jim Gray

Astronomy in the 4th Paradigm



Sloan Digital
Sky Survey (SDSS)

+

Database
Systems

Sky
Server

Science in the 4th Paradigm

SciServer

Astronomy

Connectomics

Cosmological Physics

Genomics

Oceanography

Technology Trends

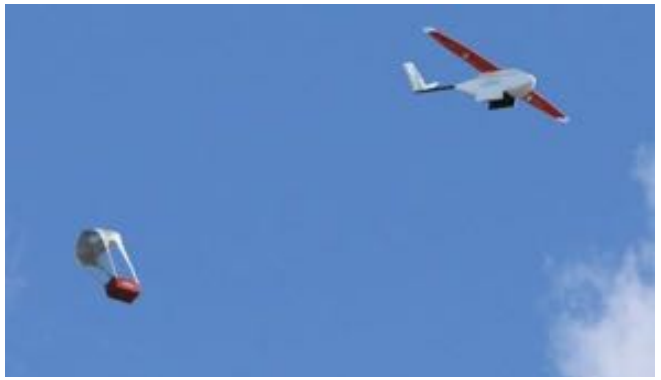
2020s ?

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 fa 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

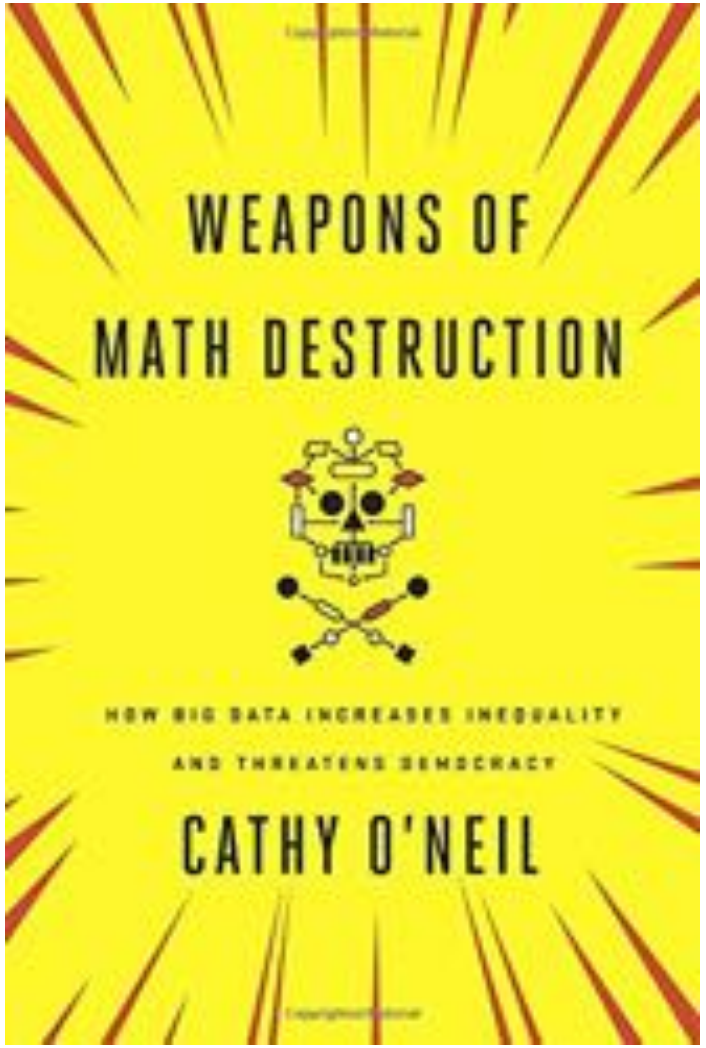
uart Russell
rofessor at
nia, Berkeley
ficial
Approach'.
cs professor
stitute of
he author of

NTELLIGENCE
E END
UMAN ERA

EINAL
NTION
BARRAT

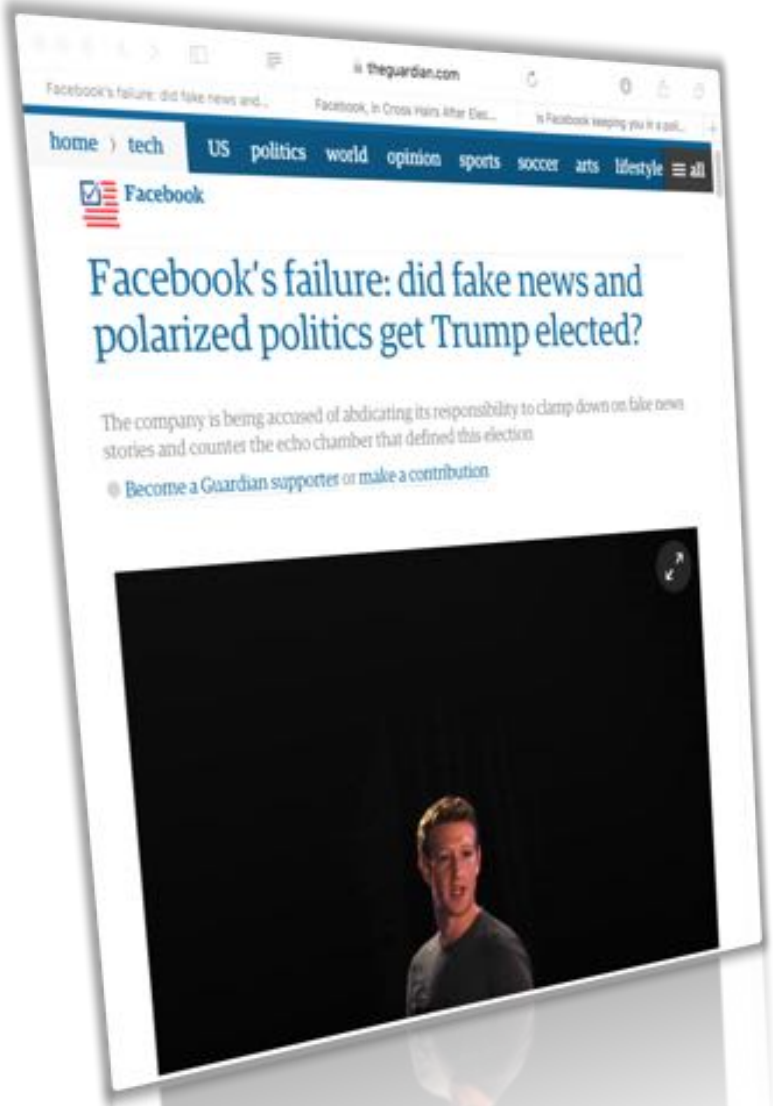
The Dark Side of Mathematical Models

- Obscuring complex decisions
  - Mortgage backed securities
  - Teaching scores & job advancement
- Reinforcing historical trends
  - Finding job candidates based on previous hiring data
  - Predictive models sending police to areas with frequent arrests
  - Denying jobs based on credit scores



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

Dangers of Targeting and Personalization



But … I am ***optimistic***

- Knowledge is empowering
- Data science offers **immense potential** to tackle 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 DS100!

Why are you excited about Data Science & DS100?

What are your concerns?

What is Data Science?

The recurring question across industry and academia.

Is Data Science …

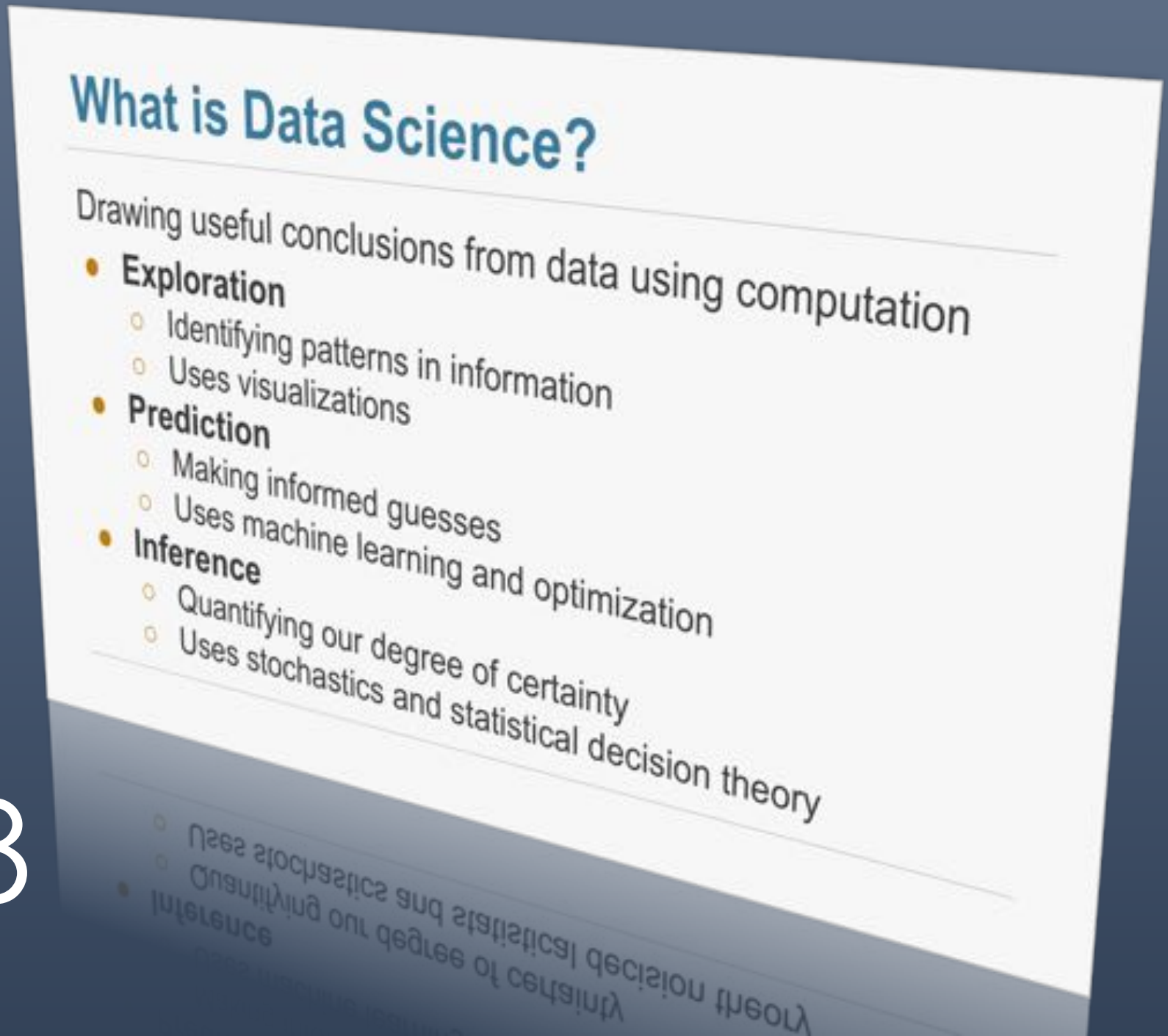
- Statistics?
  - Yes! Use data to infer properties of the world
- Machine Learning?
  - Yes! Use data to build algorithms that make predictions
- Computer Science?
  - Yes! Use computational thinking and abstraction to manage complexity
- Science, Art, and Engineering?
  - Yes! Combines the scientific method, creative thinking, and the ability to solve challenging problems …

How can we possibly teach all of this!??

Great definition …

I want something
more exciting!

Data 8



My Definition for Data Science

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

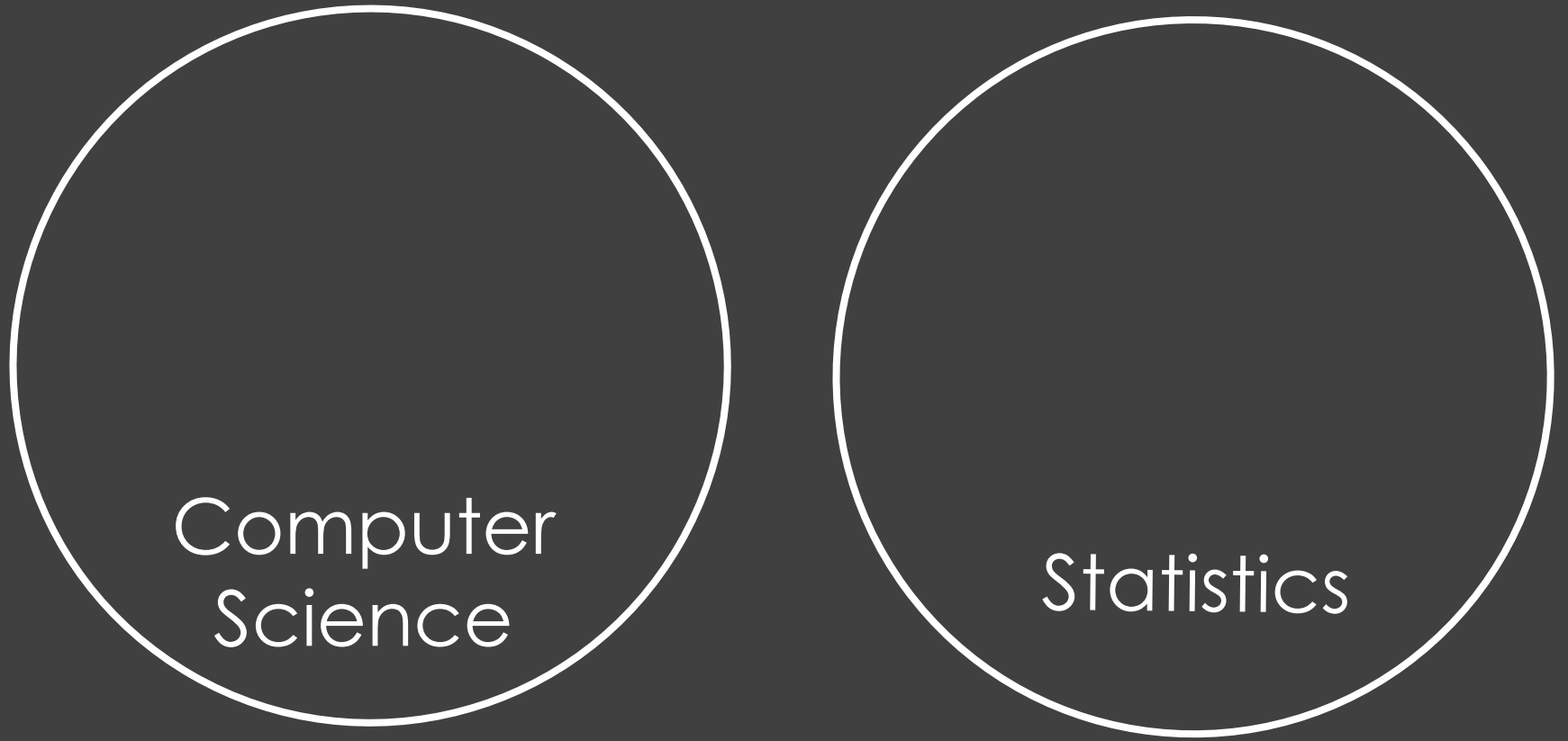
understand the world — **Science**

&

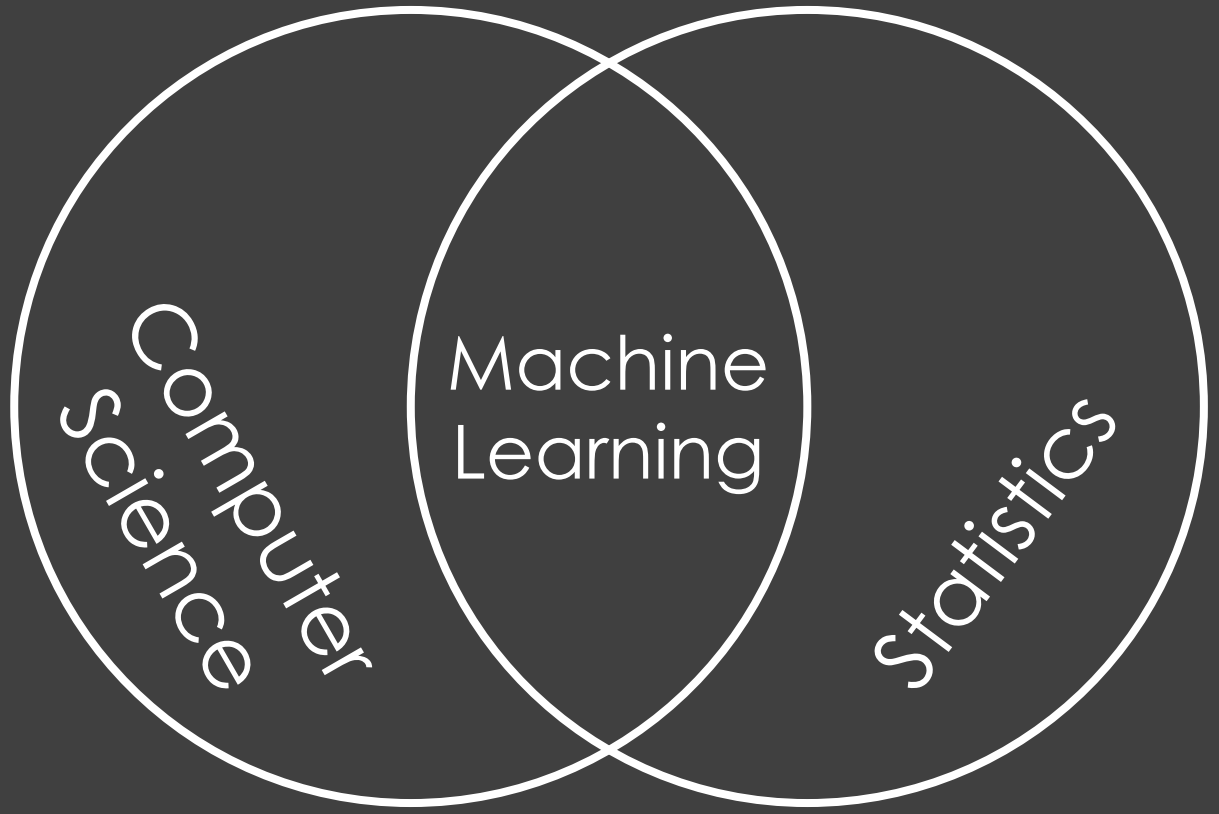
solve problems — **Engineering**

➢ *Data science is fundamentally interdisciplinary*

Skills of Data Science

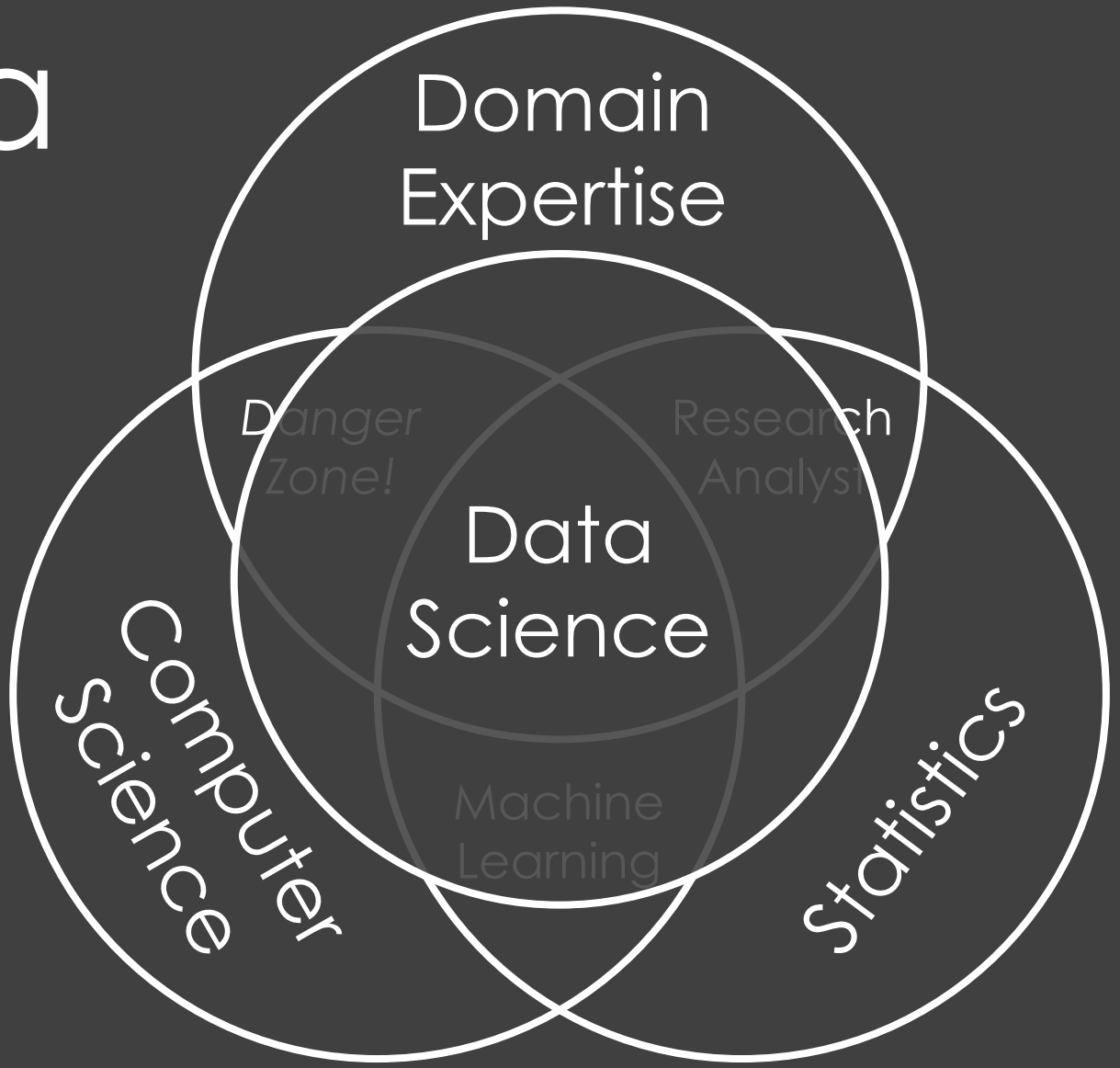


Skills of Data Science



Skills of Data Science

More ***Union*** than *Intersection*

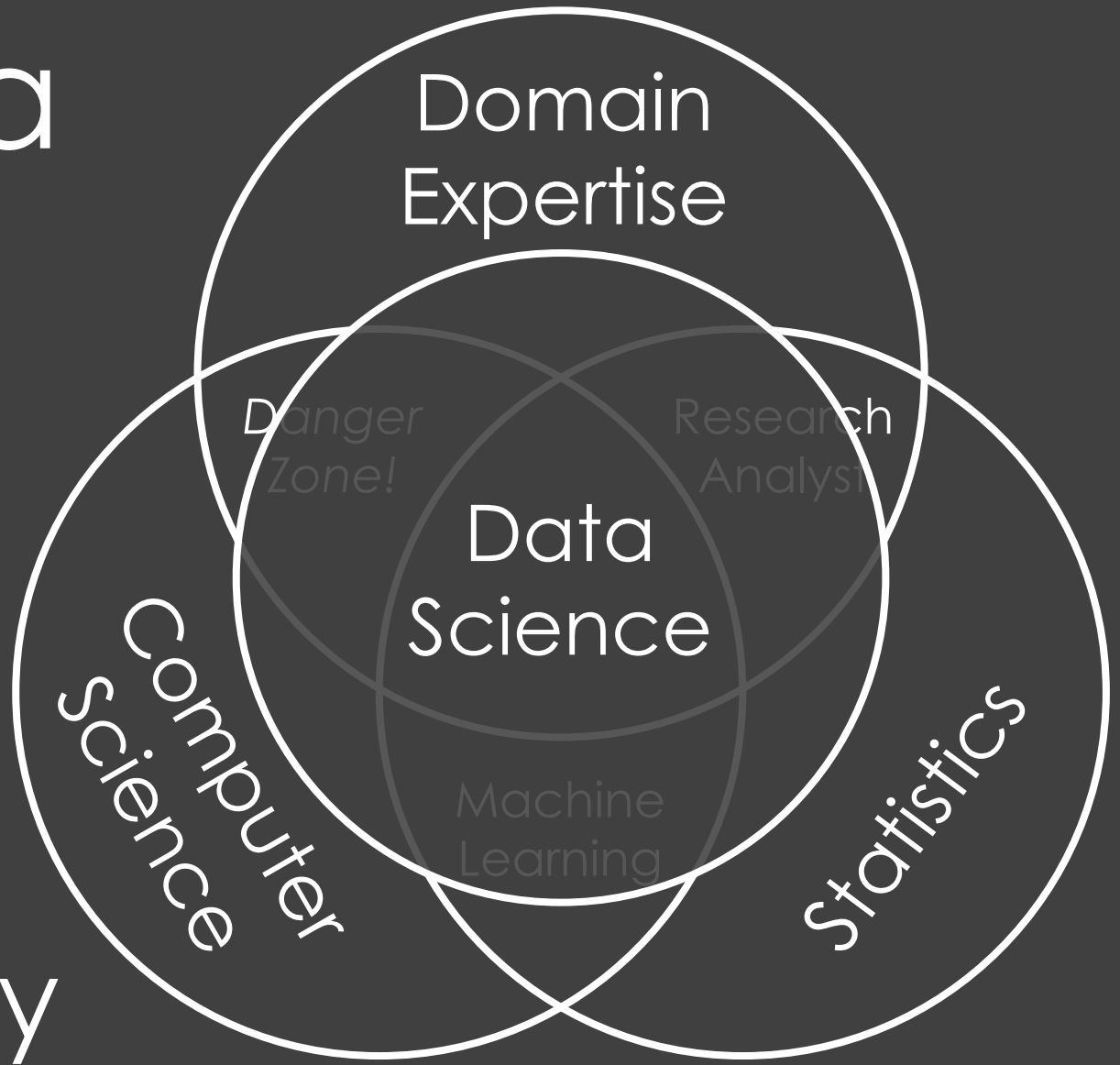


Drew Conway's Venn Diagram of Data Science

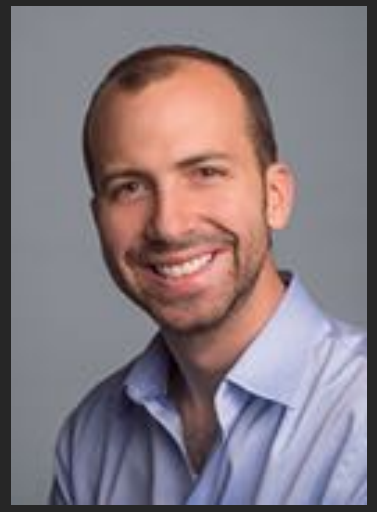
Skills of Data Science

More ***Union*** than *Intersection*

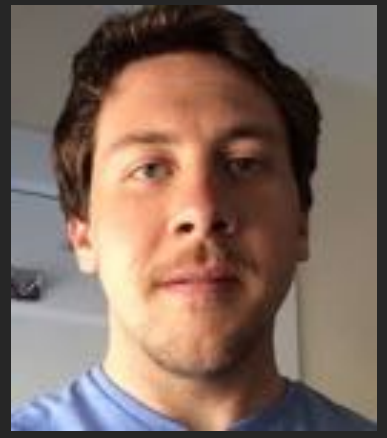
Data Science is interdisciplinary



Drew Conway's Venn Diagram of Data Science




DS100 Created and Taught by Faculty and TAs With Diverse Background & Perspectives

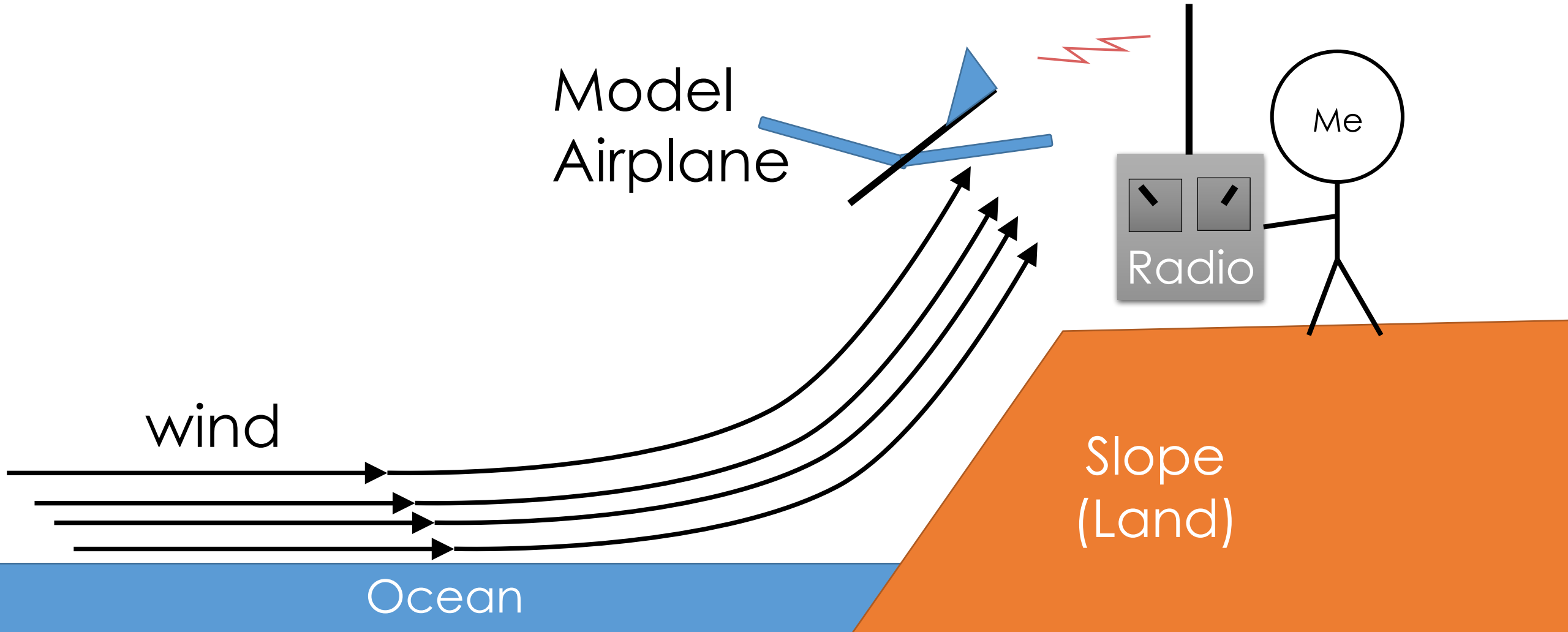
Joey Gonzalez

Joined EECS at UC Berkeley in 2016

Research Area: *Machine Learning & Data Systems*

- Study design of scalable systems for machine learning
  - **Algorithms:** Developed parallel algorithms for statistical inference
  - **Abstractions:** vertex centric programming abstraction
  - **Systems:** created the GraphLab and parts of Apache Spark
- Co-Founder of Turi Inc.
  - Python tools for scalable data science
  - sold to Apple Inc. in 2016
- Something interesting about me …

Slope Soaring "Drones"




Slope Soaring with the Berkeley Birds

Joe Hellerstein

Joined UC Berkeley in 1996

Research Area: *Database systems* + …

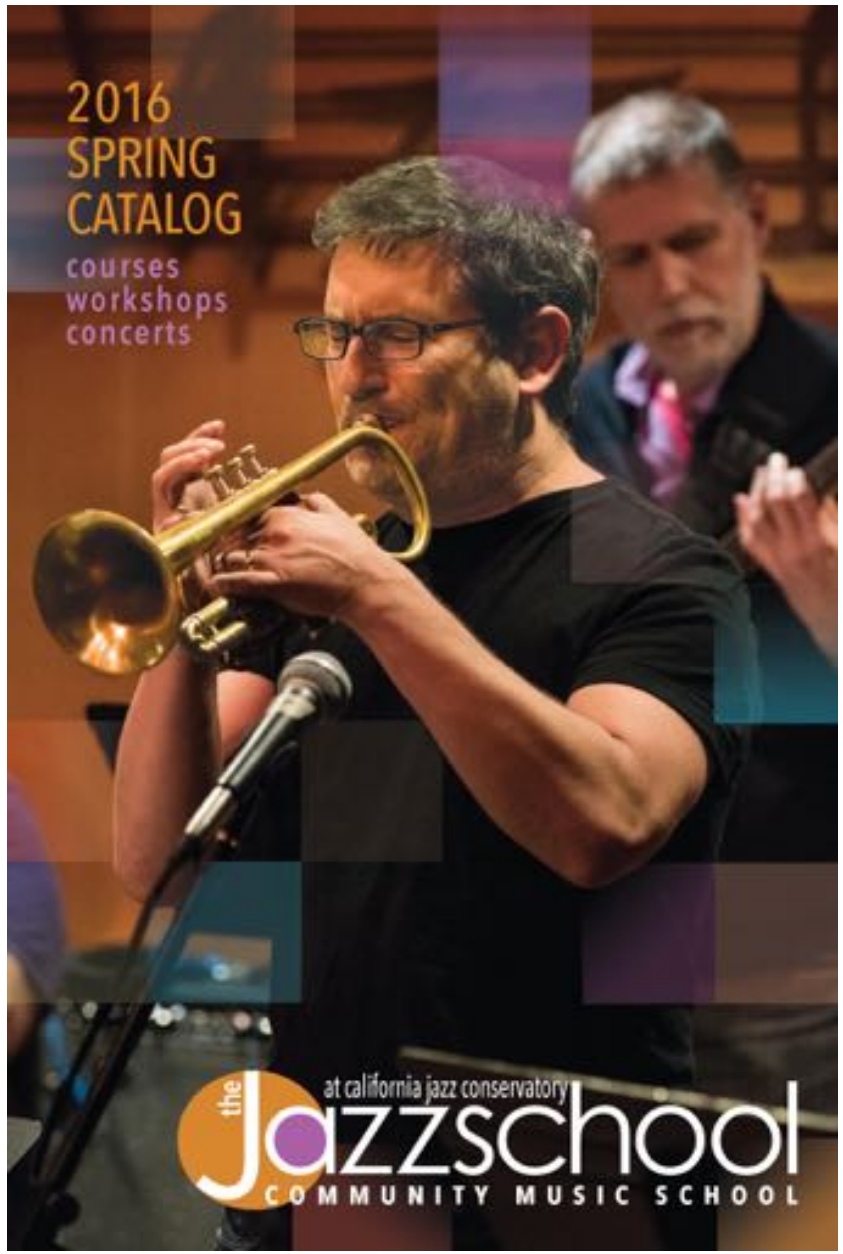
- Data-centric computer systems design
  - **Programming at scale:** Cloud (Bloom), ML (Apache MADlib)
  - **Visualization:** especially interactive visualizations
  - **Foundations:** systems internals & theory
- Co-Founder of Trifacta
  - Visual data wrangling for end users
- Something interesting about me …

Two Truths and a Lie

1. Taught Bill Gates how to scroll on a MacBook
2. Married Prof. Gonzalez' Ph.D. advisor
3. Advised President Obama on bank bailouts

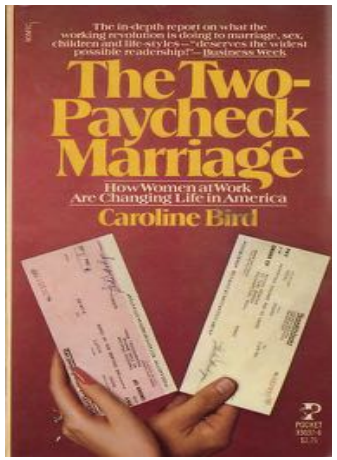
- The career that got away…
- Lucky to play with some real musicians…

  Vijay Iyer, Geoff Keezer, Joshua Redman, Anton Schwartz
  Carla Bley, Jane Ira Bloom, Lester Bowie, Benny Carter, Buck Clayton, Rosemary Clooney, Harry Connick, Jr., Harry "Sweets" Edison, Joe Henderson, Illinois Jacquet and Steve Swallow



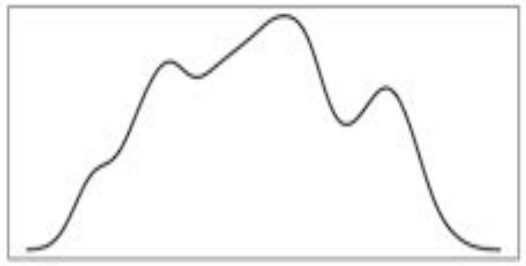
Deborah Nolan

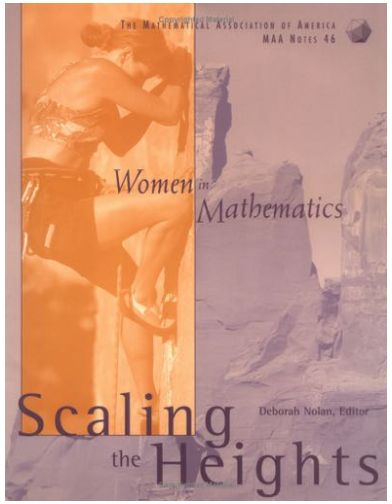
College: Summer Internship



First Job: Programmer –

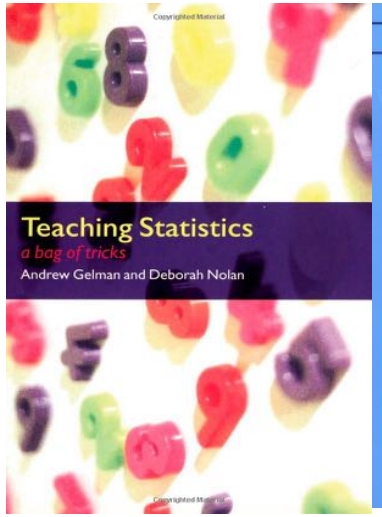
Early Research
Functional CLT



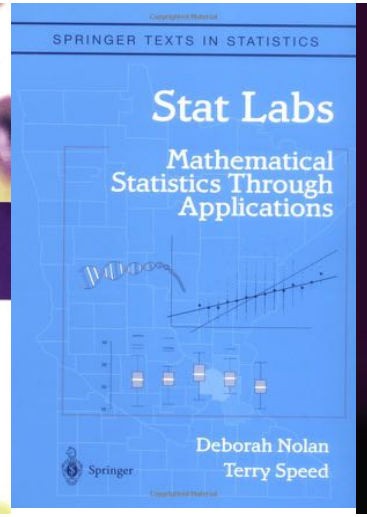


Summer Program for Women in math

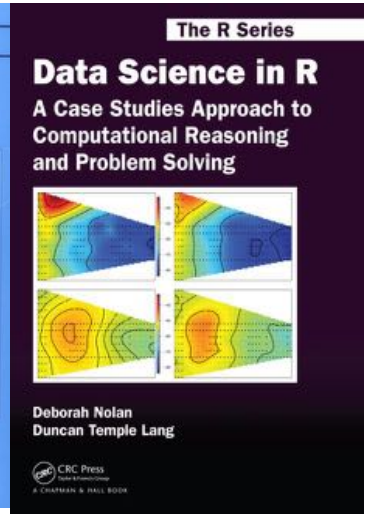
Teaching



Case Studies



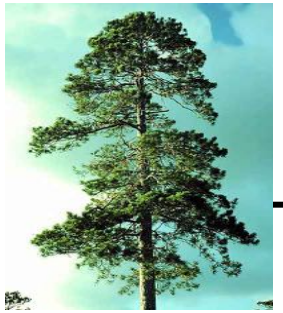
Data Science

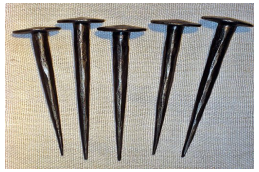


Project today –

- Prototype documents for
  - Reproducible research
  - Education

7 Friends +

100  +

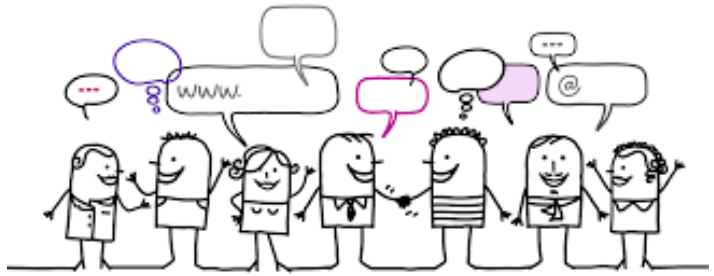
Hundreds of  +

Thousands of hours of 

Bin Yu, Professor of Stats and EECS

Started at Berkeley in 1993

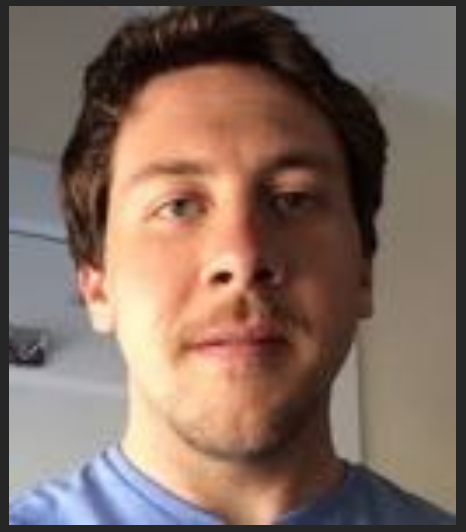
- **Research Areas:** Statistics, machine learning, causal inference, collaborative research in neuroscience, genomics, precision medicine, remote sensing, …
- **Approach**: Solving data problems via statistical and machine learning methods/algorithms, domain knowledge, and theory, while embedding students/postdocs in labs.
- *Other things that I love to do…*


www.


KEEP
CALM
I LOVE
being a
MOM

TAs

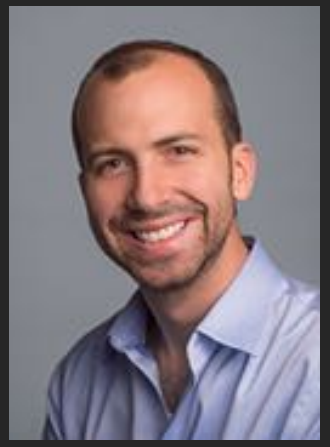
Andrew Do



Henry Milner

Sam Lau

Joseph Simonian




DS100 Created and Taught by Faculty and TAs
With Diverse Background & Perspectives

Challenge & Opportunity

Inconsistent terminology → Learn multiple terms

Disagree on issues → See multiple perspectives

Varying expertise → Learn from & with us

Intermission

5 Minute Break.

Ask a neighbor:

What is your name?

Emacs or Vim or …?

What do statisticians and pirates have in common?

Contemplate:

What are the ethics of data science?

Can data do harm?

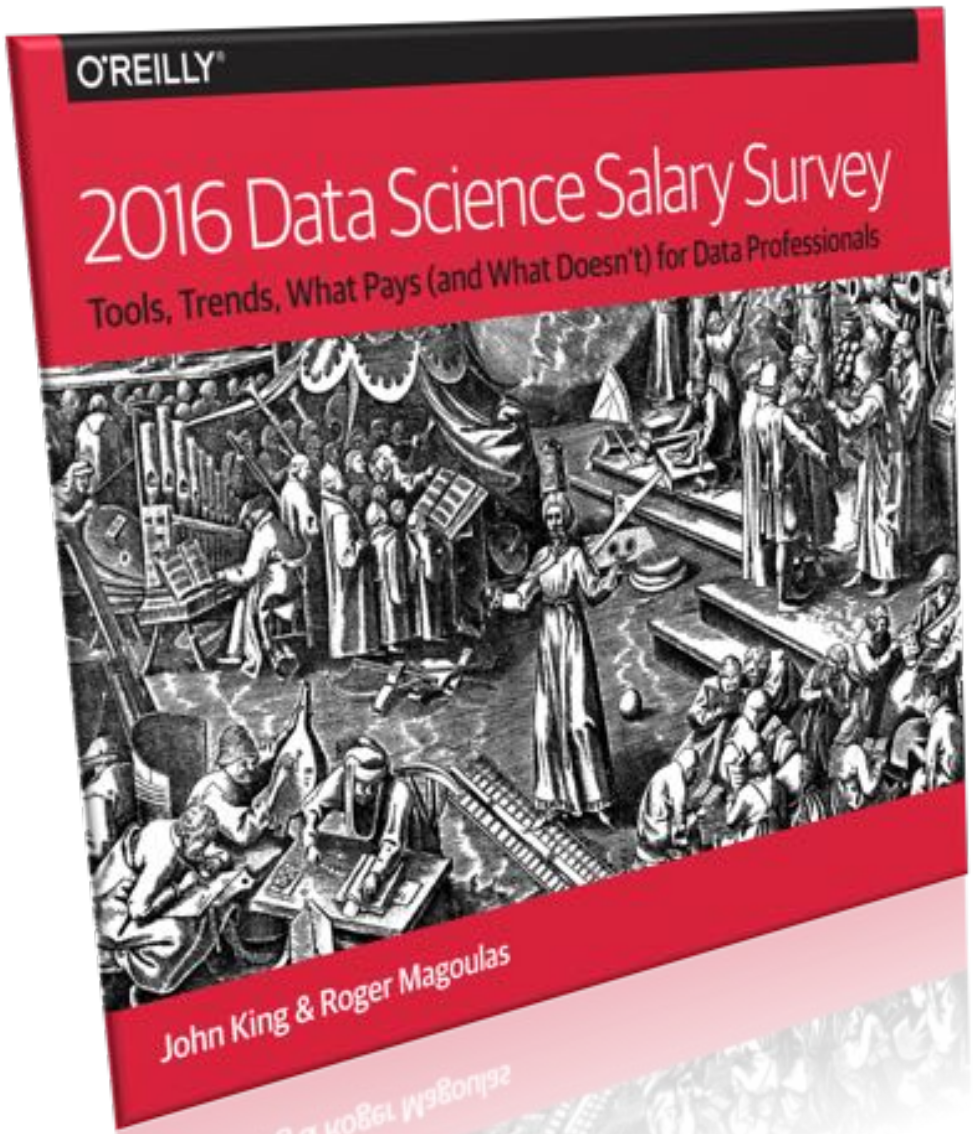
What do you want to get out of DS100?

Homework 1 will be Released Today

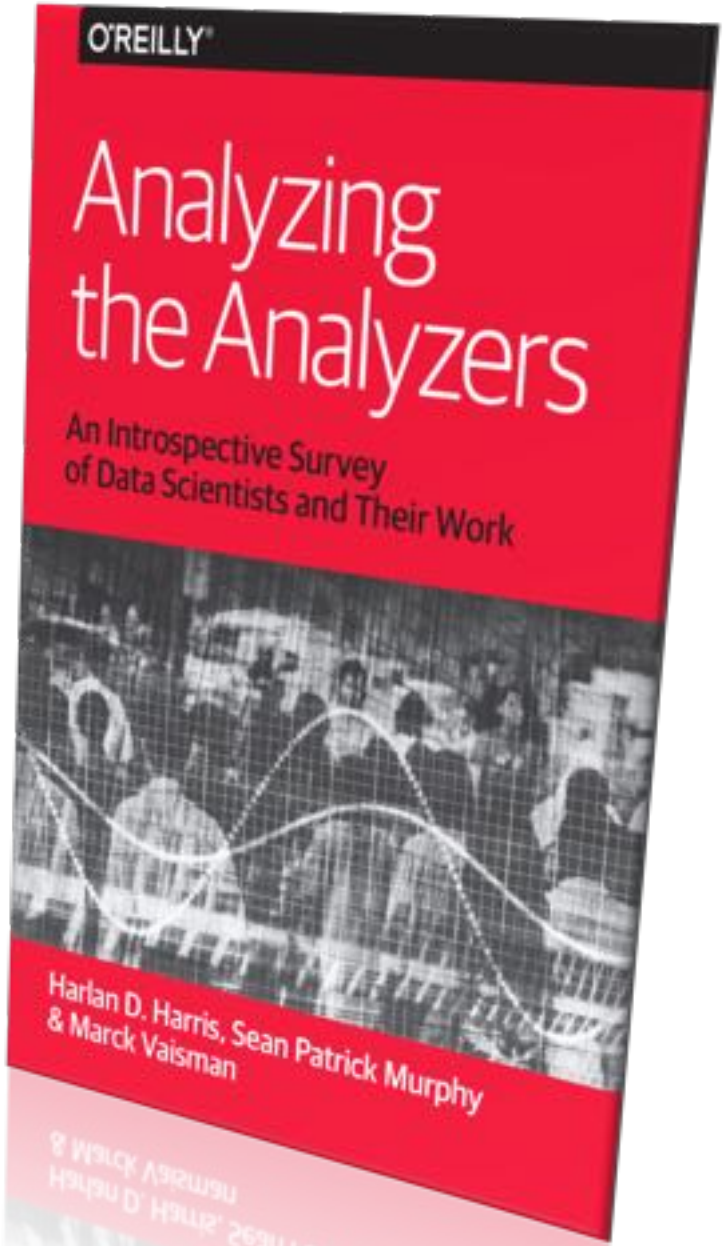
- Setting up your computer for data science
  - We want you to know how to start from scratch
- Warmup Homework:
  - Reviewing python & numpy
  - Fun prediction exercise
- Covered in **section** and **lab** this week
  - try to bring your computer …
- **Due next Tuesday at midnight (1 Week)**

What does it mean to be a data scientist today?

How can we answer this question?



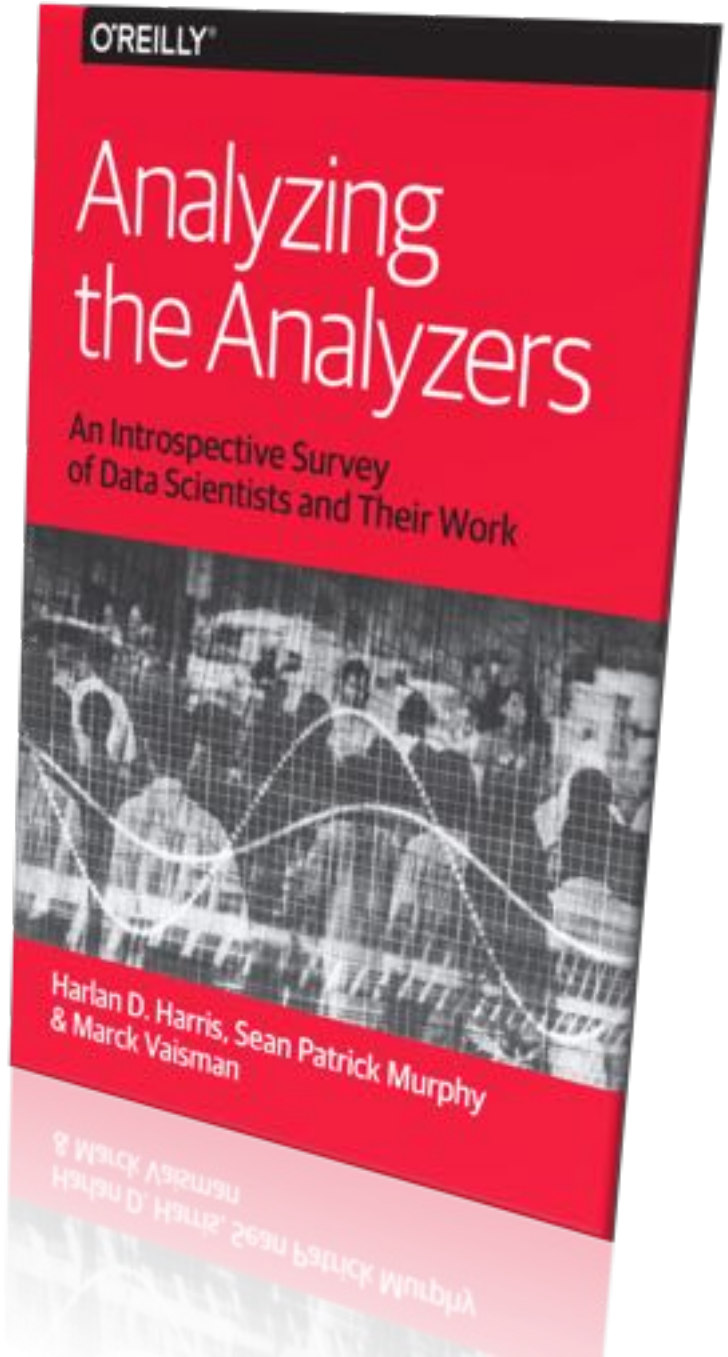
O'REILLY Surveys



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

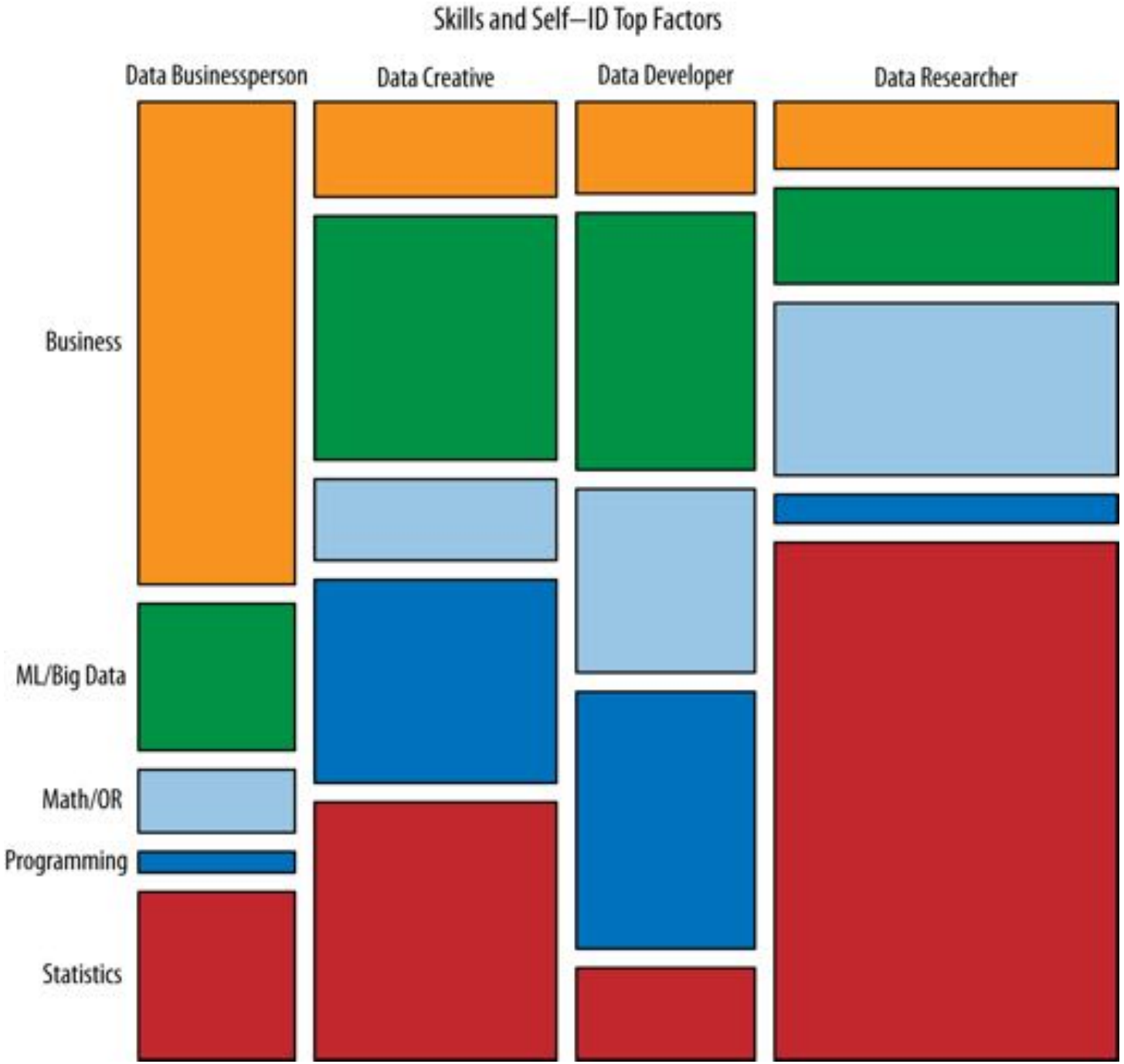
Analyzing the Analyzers

- Surveyed 250 people **in 2012**
  - Self reported as data scientists
  - Invited at data science meetups
- Asked to rank skills and activities
- Asked for job group:
  - Data Businessperson
  - Data Creative
  - Data Developer
  - Data Researcher



What are the skills that best define each group?

Data (Businessperson | Creative | Developer | Researcher)



Skill Patterns

- Different skill profiles
  - Business = Domain Knowledge.
- DS-100 focus
  - Data Creative / Developer

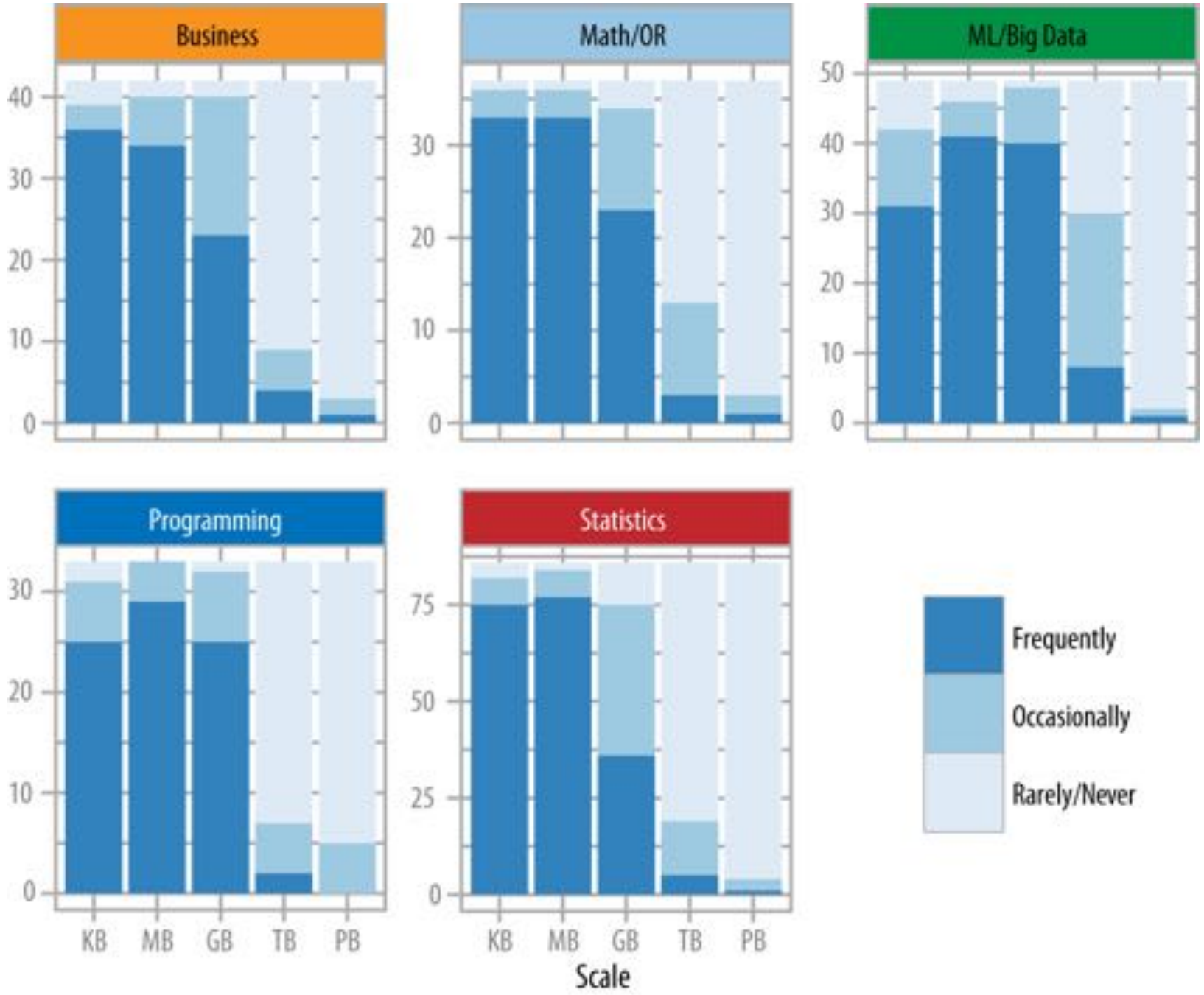
There is a lot of excitement around

Big Data

… how big is the data?

Data of Scale

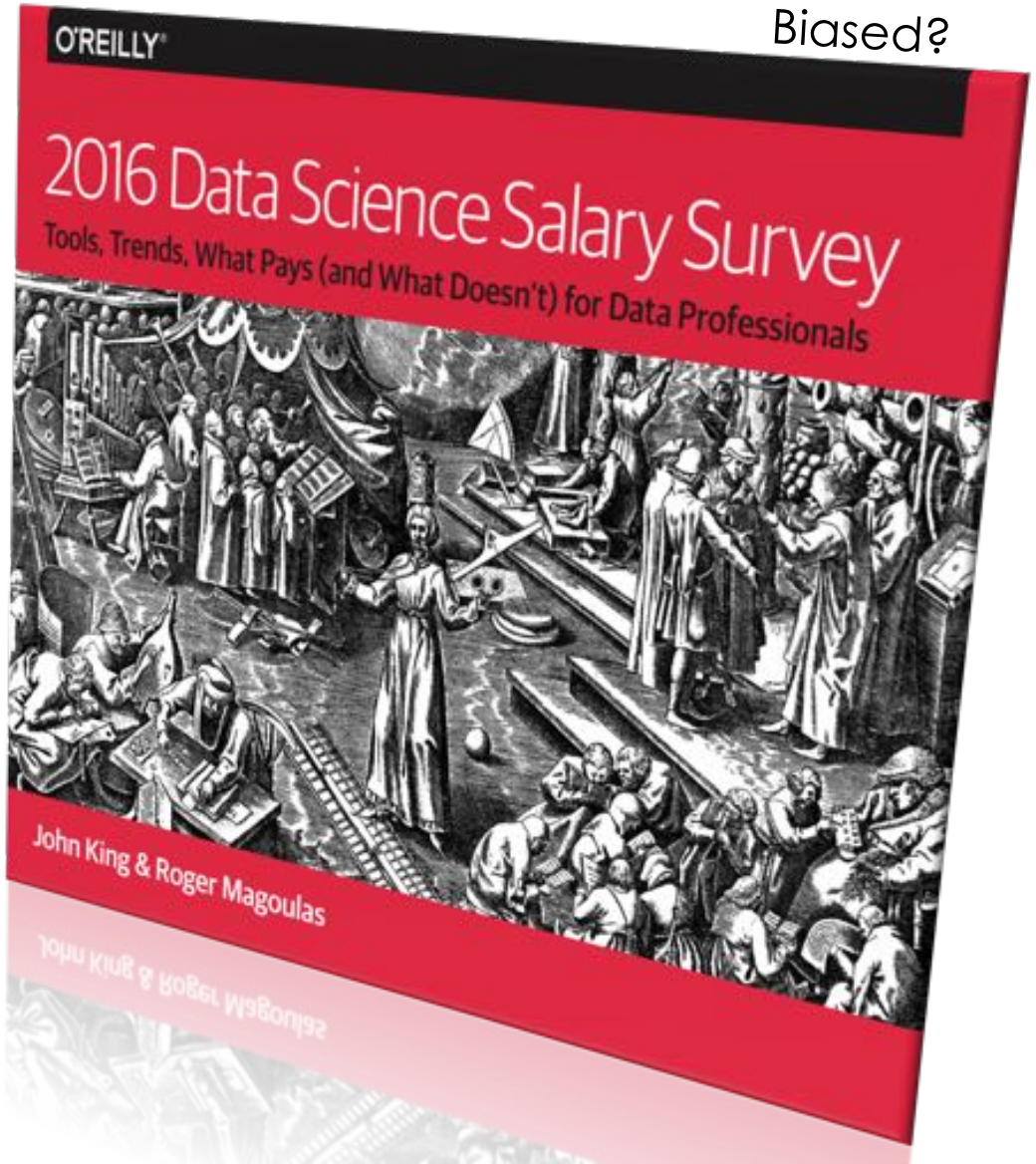
- Not usually big!
  - < 1GB
- Some TB to PB scale data
  - Frequently!



O'REILLY Data Science Salary Survey

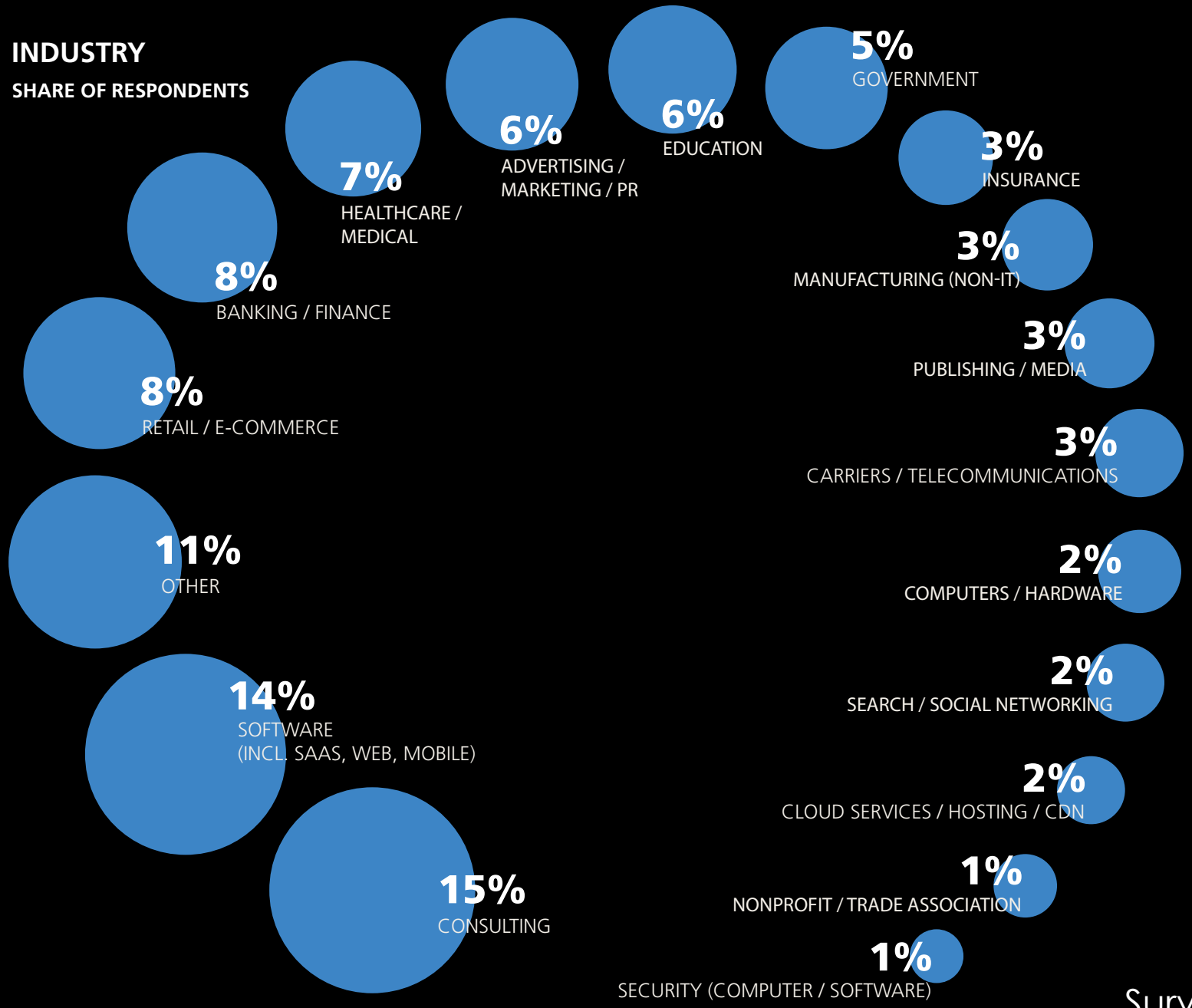
Biased?

- Conducted annually
- Anyone can take the survey
  - Promoted at O'Reilly Events
- The 2016 Survey Sample:
  - 64 Questions
  - **983 Respondents**
  - 45 Countries (61% US)
- Results …



Who took the survey?

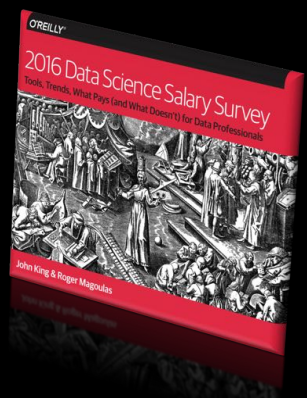
Sample bias …



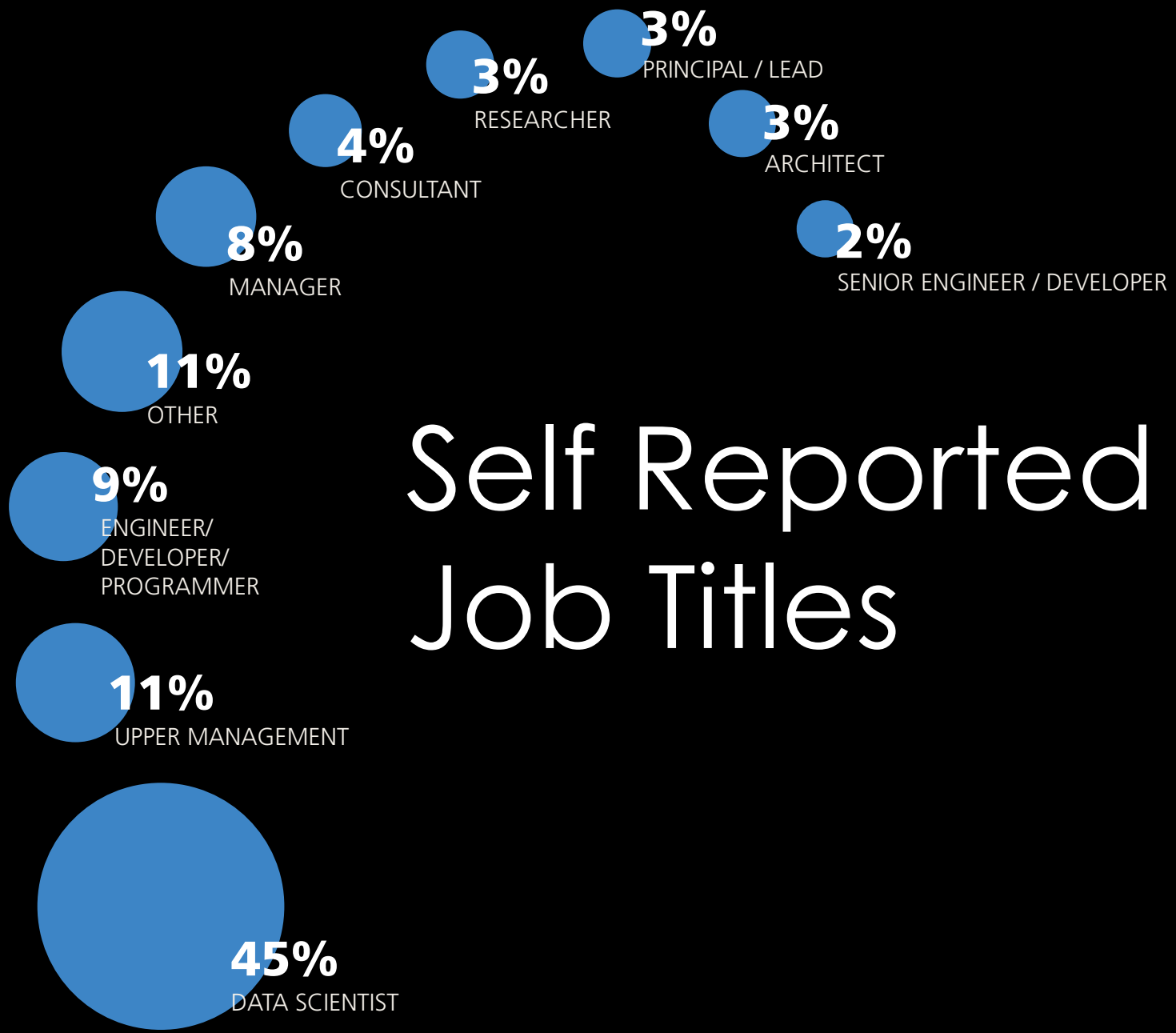
Breakdown over industries.

Top two are

- Consulting
- Software

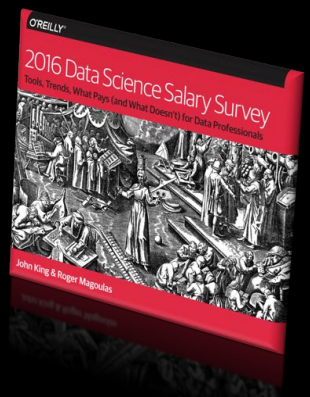


Survey selection bias …



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

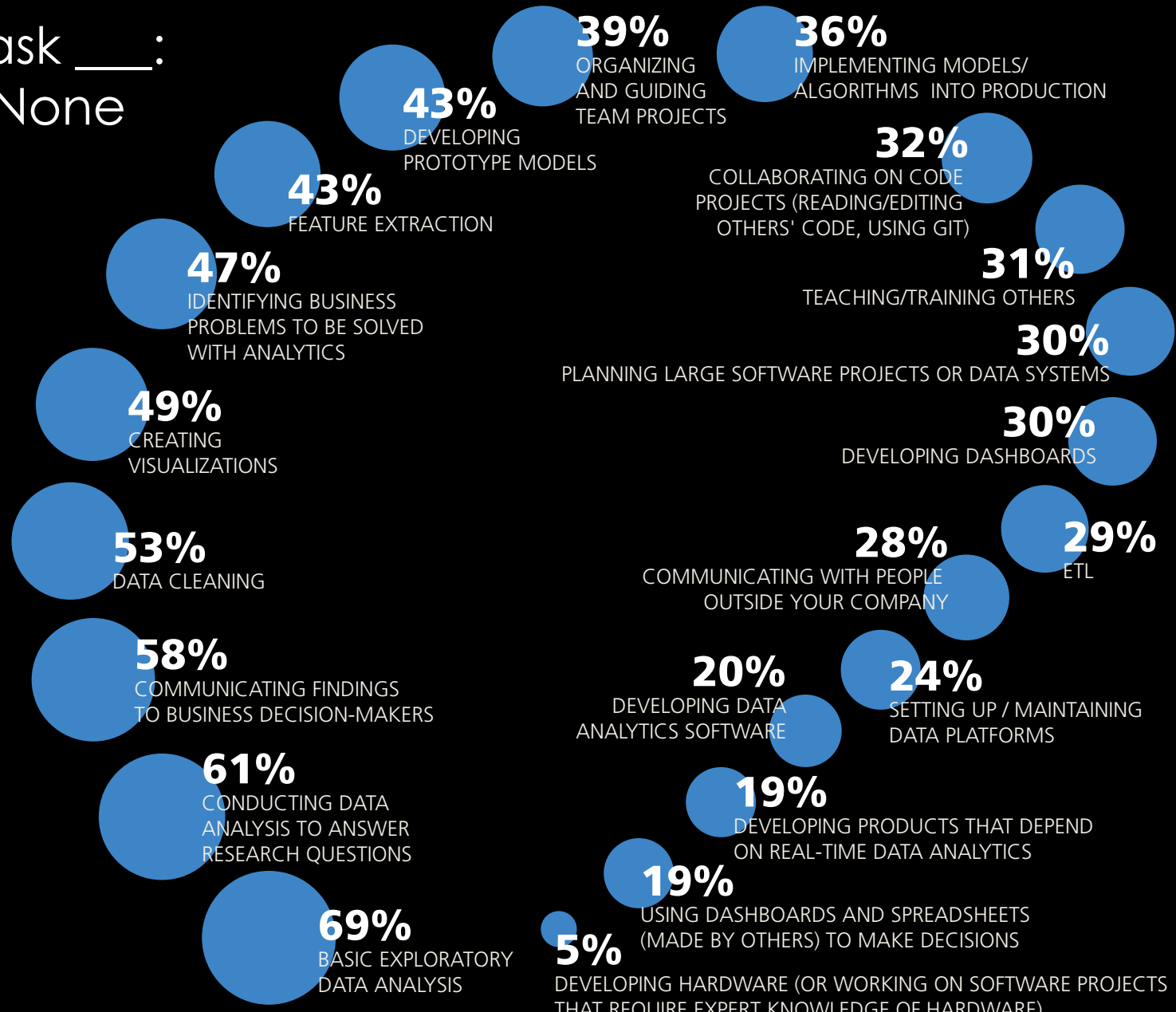
Developing Models
Implementing ML Algorithms
Visualization

Exploratory Data Analysis (EDA)
Researching Questions
Writing Reports,

…

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

Are the top items surprising?

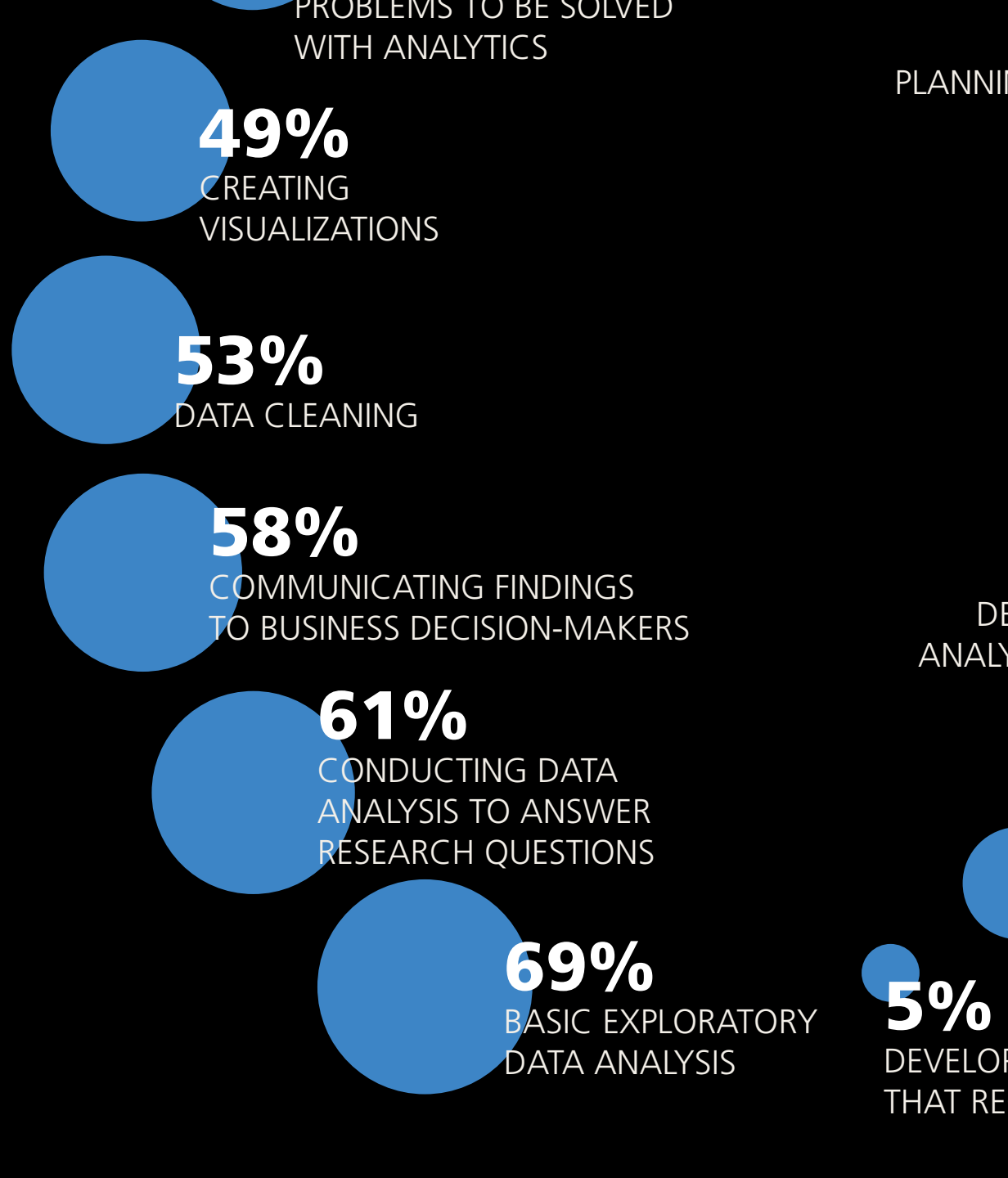


How involved are you in task ___:
(a) Major, (b) Minor, (c) None

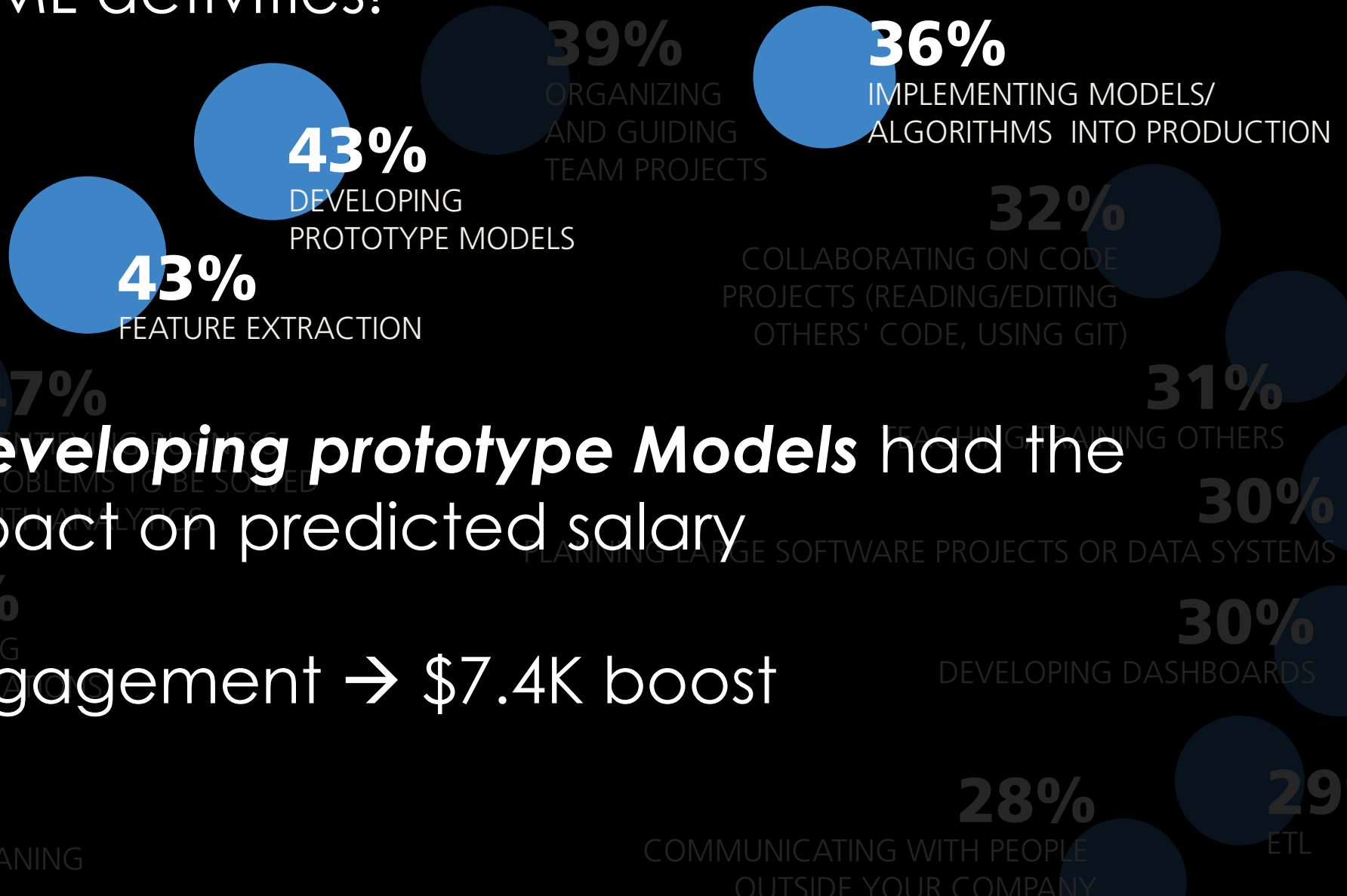
Are the top items surprising?

Data Cleaning ☹

Where are Modeling / Prediction?



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

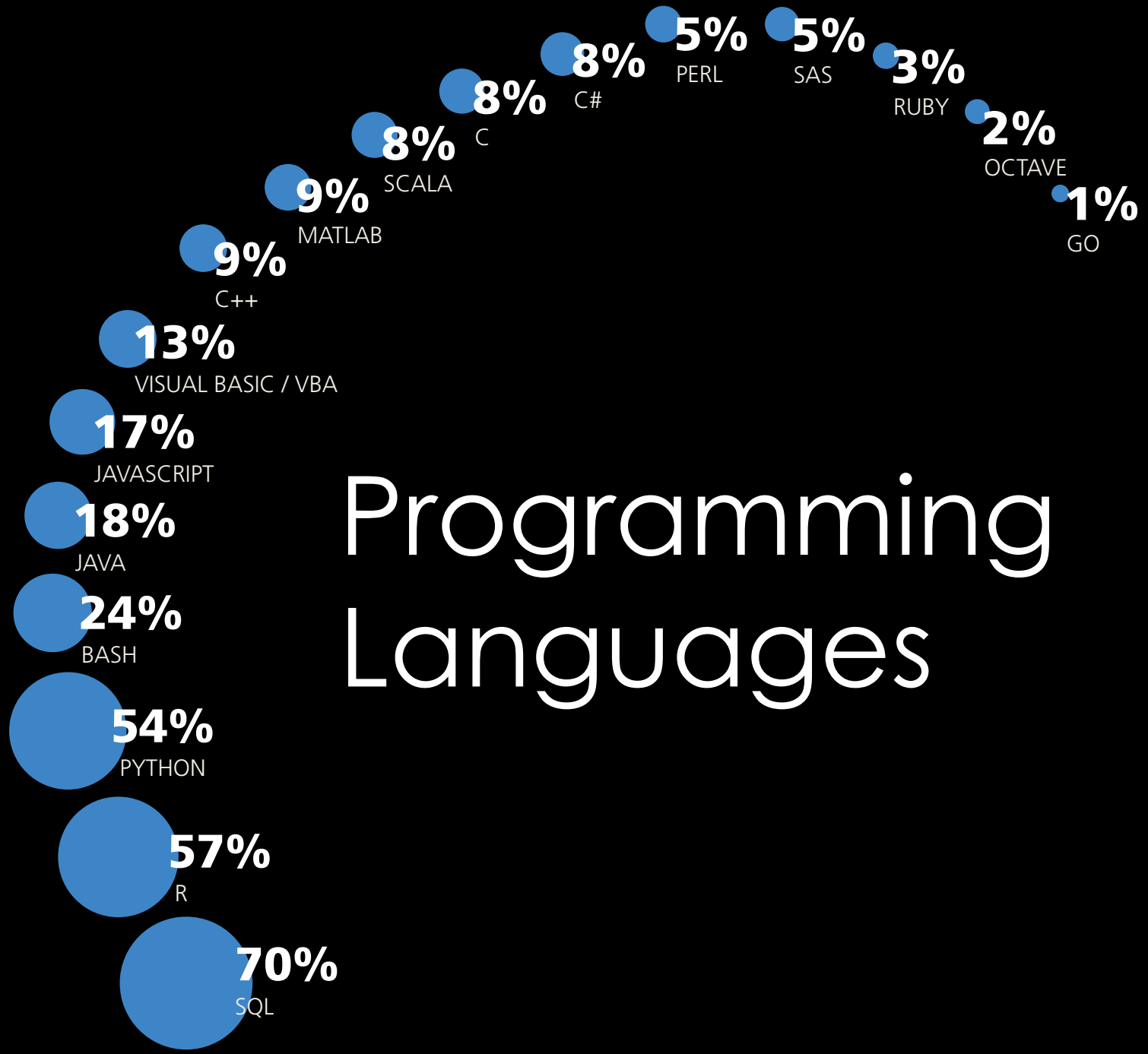


However, ***Developing prototype Models*** had the greatest impact on predicted salary

- major engagement → $7.4K boost

What tools do they use?

- Programming Languages
- Machine Learning
- Data Technologies



Programming Languages

SQL > R > Python

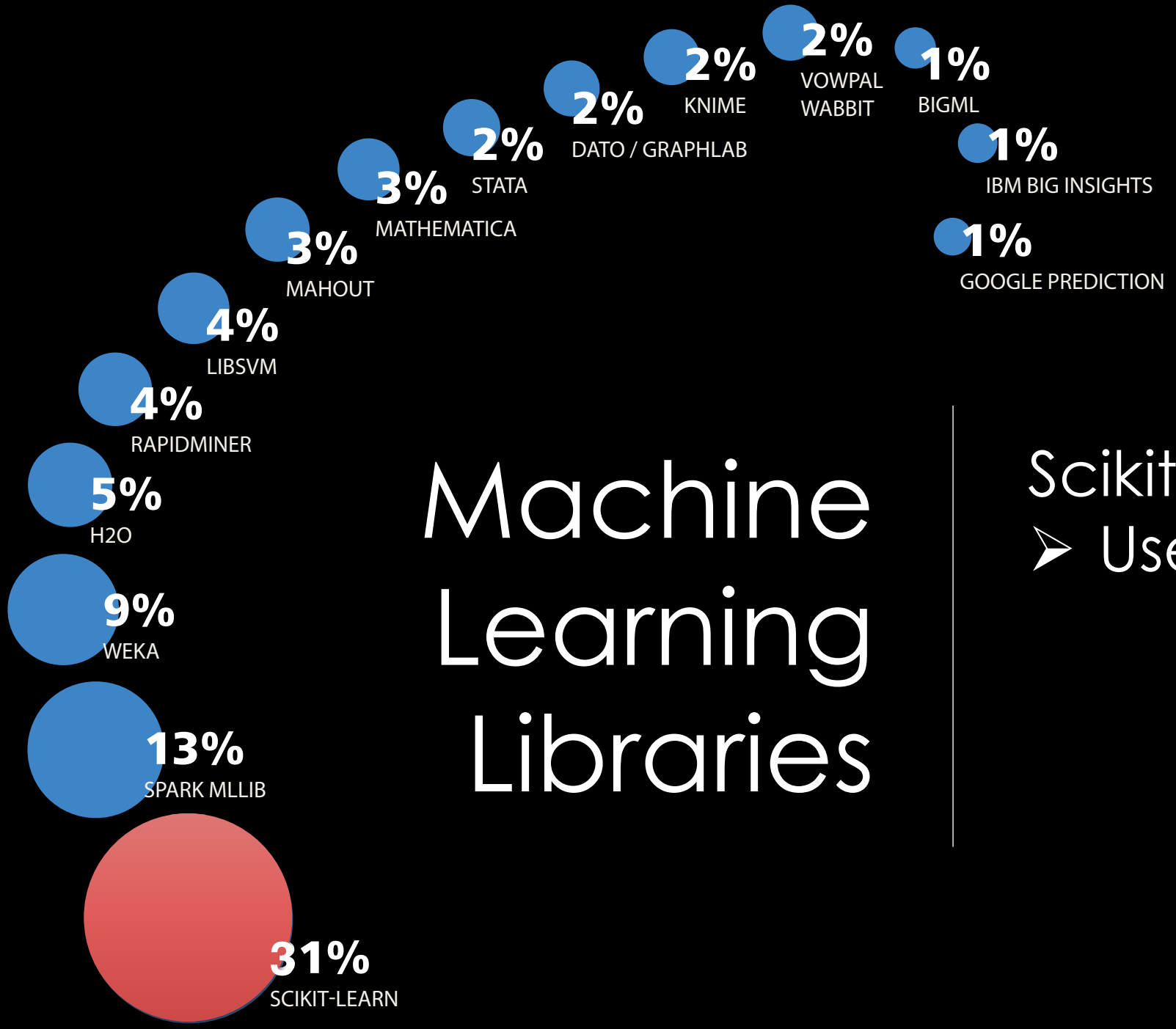
Cluster Analysis:

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

Python users had higher salaries.

Highest Paid?

- Scala



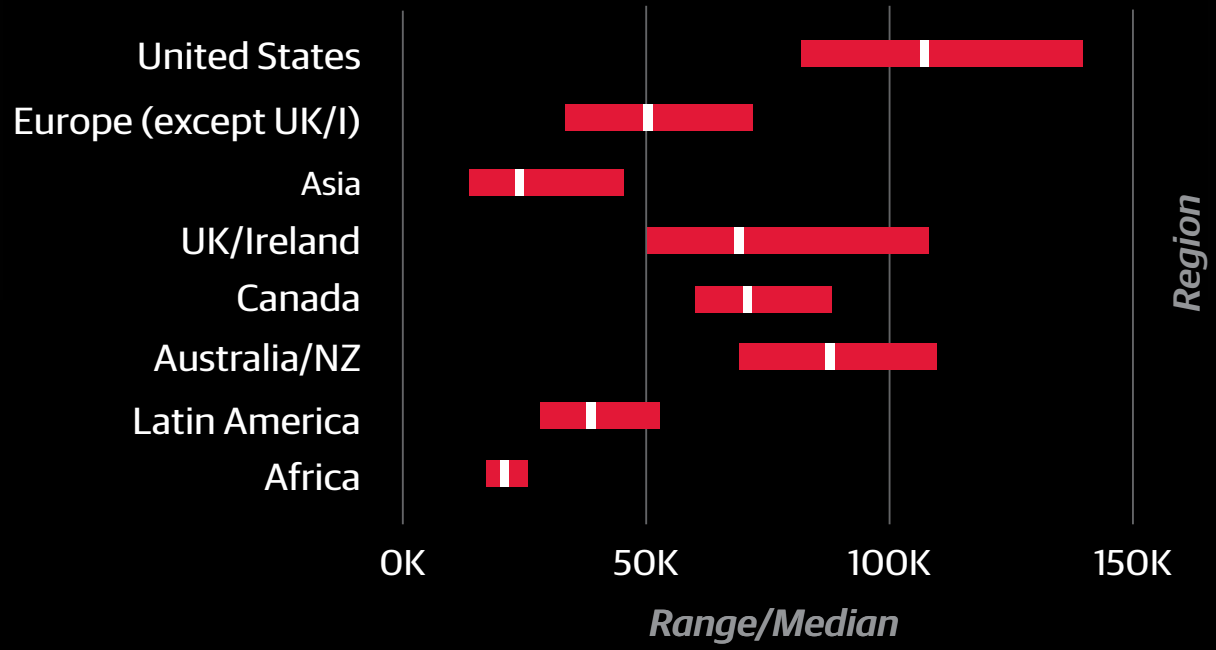
Machine Learning Libraries

Scikit-learn used most

- Used in DS100

How much are they paid?

SALARY MEDIAN AND IQRC (US DOLLARS)



Salary Depends on Location and Industry

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 (CS186), machine learning (CS189), and statistics (Stat-154), 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 tackle real-world problems

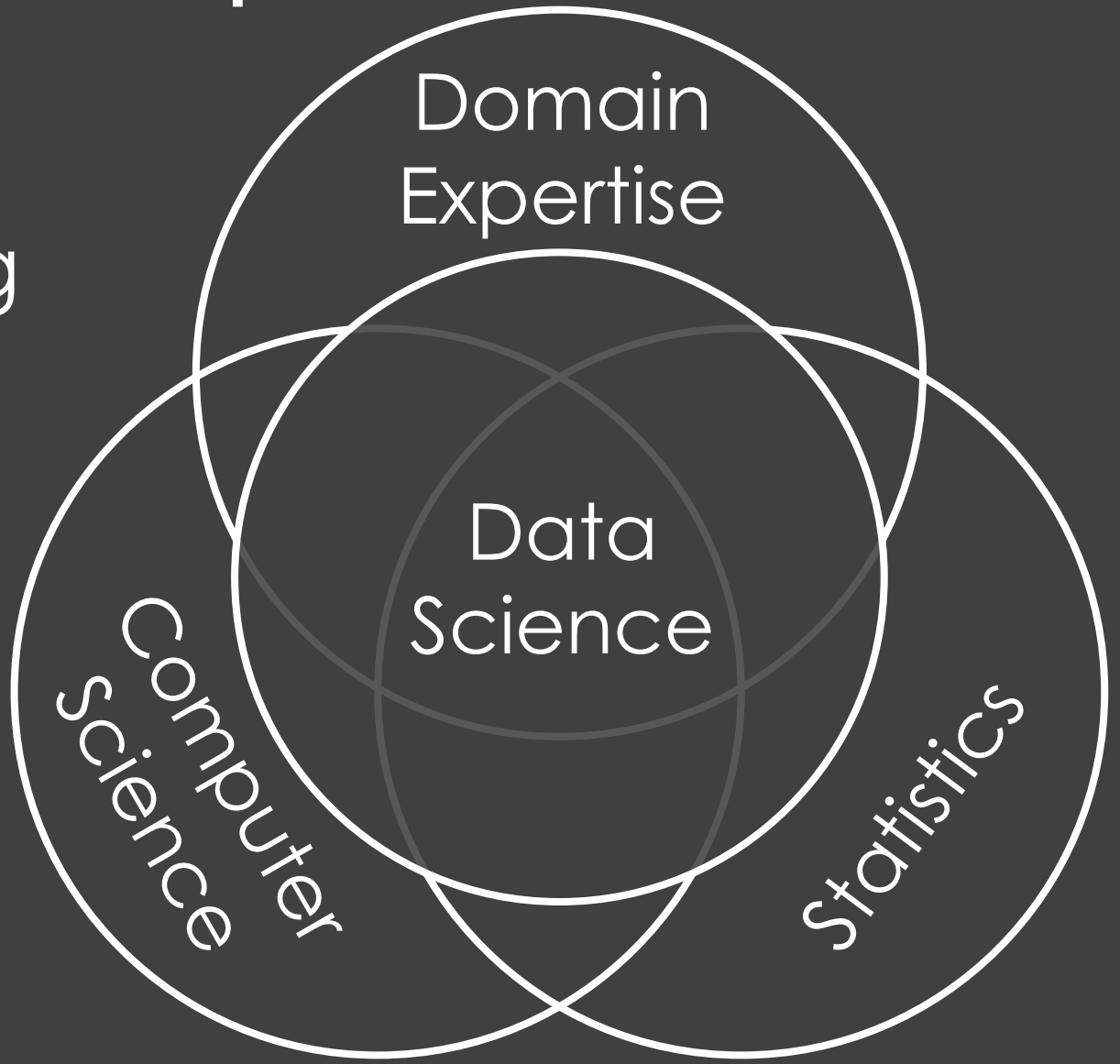
What will I learn?

Data Science Requires Many Skills

Can't cover everything in DS100.

Instead we cover

- Key Concepts
  - … some details
- Connections
- How to learn …



Big Concepts in Data Science

- Data preparation and representation
- Efficient data processing
- Question formulation and experimental design
- Exploratory data analysis
- Modeling, parameter estimation, and statistical inference
- Various prediction methods: generalized linear models, decision trees, neural networks, clustering, PCA, …
  - Overfitting, regularization, and cross validation

Principles Computer Science in Data Science

- Software Design & Debugging
  - How do we develop and maintain reliable & repeatable analysis?
- Abstraction and Algorithm Design
  - How do we break big problems into small problems?
- Computational Complexity
  - How do we tradeoff time and space to compute efficiently?
- Parallelism & Locality
  - How do we divide computation across resources?
- Others?

Principles Statistics in Data Science

- Experimental Design & Sampling
  - How do we collect data to accurately answer questions?
- Probability & Uncertainty
  - How do we quantify what we don't know?
- Modeling
  - How do we distill the essential structure of complex phenomena?
- Inference & Prediction
  - How do we use the known to reason about the unknown?
- Others?

Domain Knowledge

- What are the key questions/problems in the domain?
- What is the context of the data?
  - What data is already available?
  - How and why was it collected?
  - What is the schema and limitations of the data?
  - How can more data be collected/obtained?
- What is the underlying process that generates the data?
  - causal structure, dependencies, …

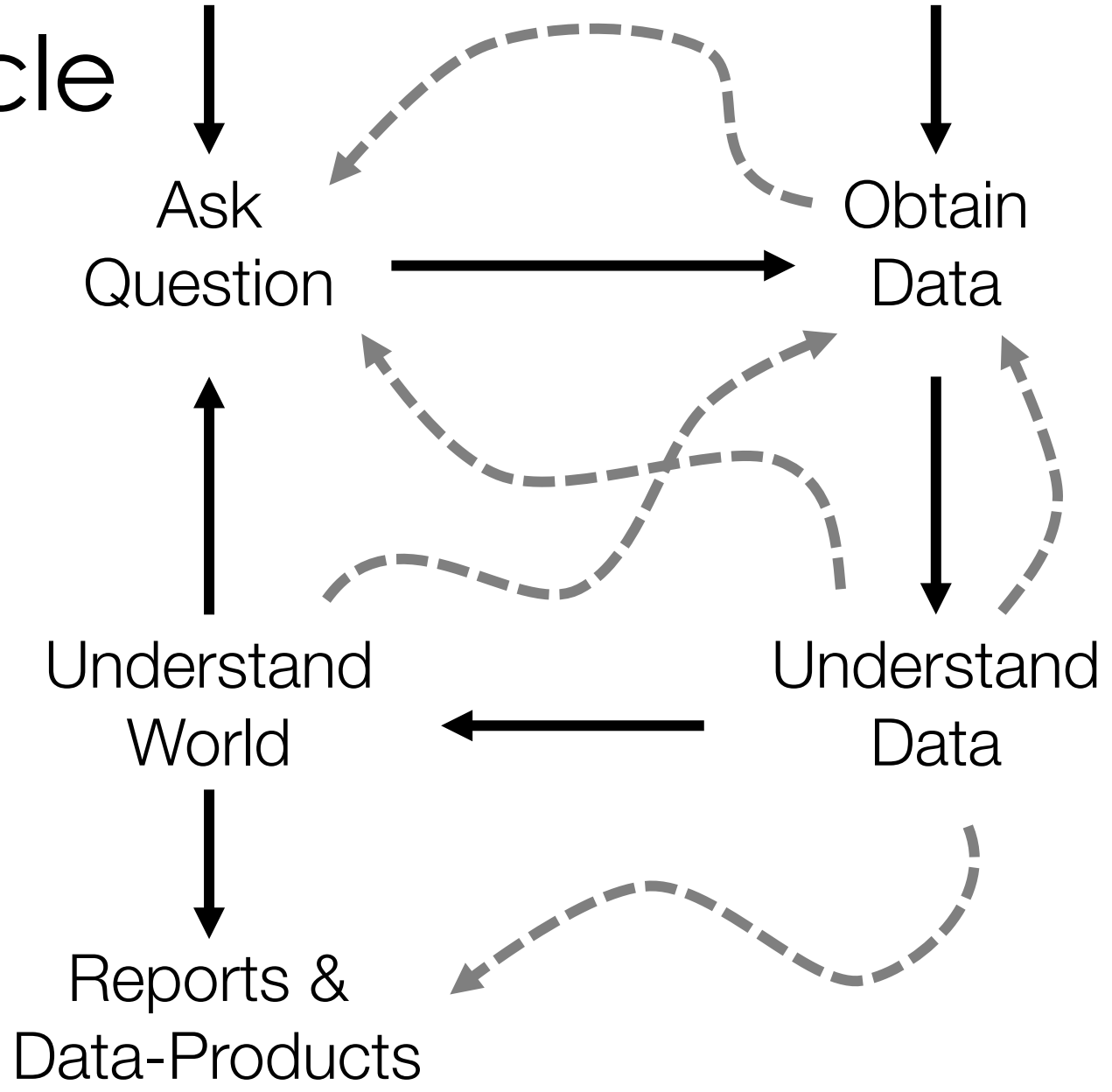
Data scientists must be inquisitive and learn new domains quickly …

Data Science Lifecycle

High-level description of the data science workflow

- Frame questions & design experiments
- Obtain and clean data
- Summarize and visualize data
- Inference and prediction

continuous process …



Working with 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***

Using Real Tools

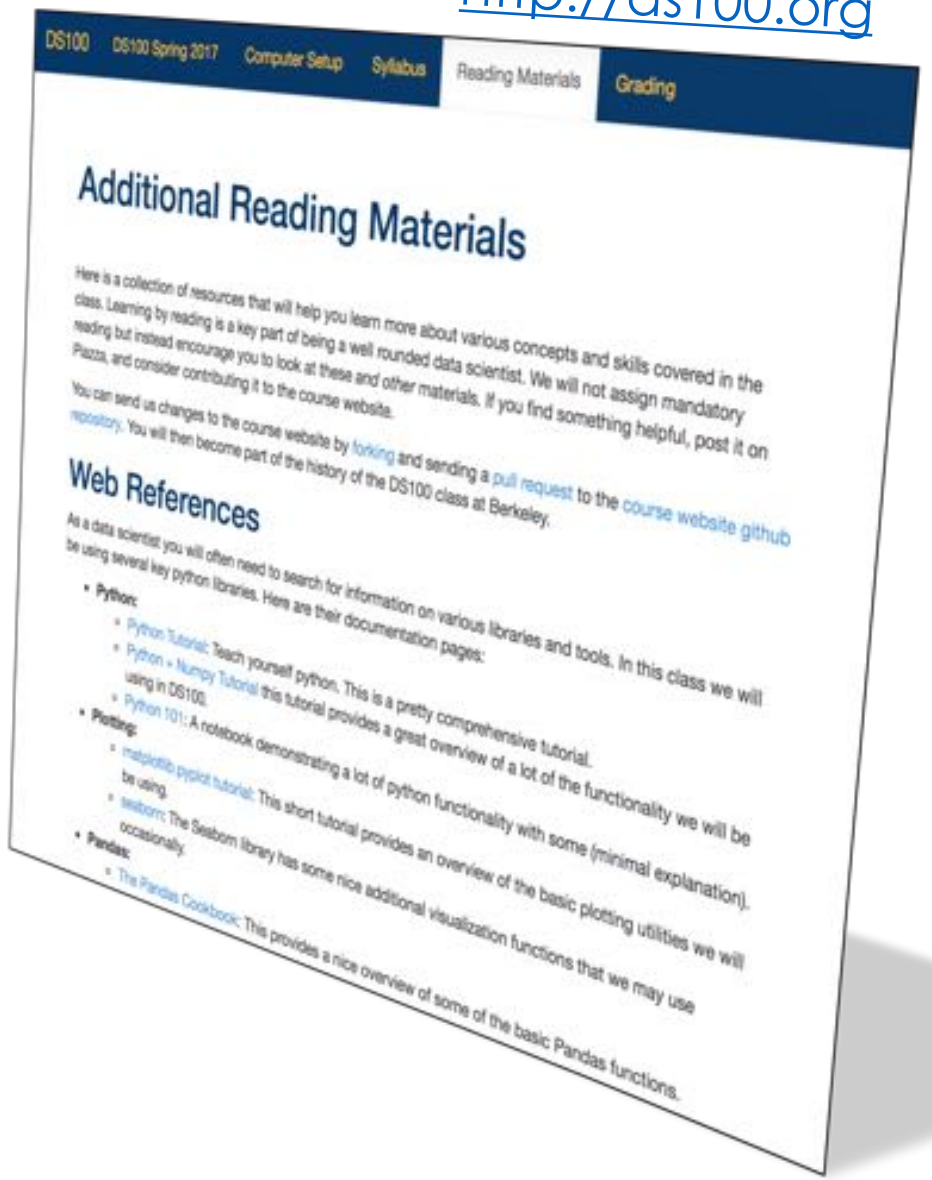
- Focus on Python programming language
- We will use various different technologies
  - Jupyter notebooks, pandas, numpy, matplotlib, SQL Server, github, Wrangler, plotly, tableau, Spark?, …
- *We **won't** teach you everything …*
  - You will learn to **read documentation**
  - You will learn to **teach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

Reading and Reference Materials

http://ds100.org

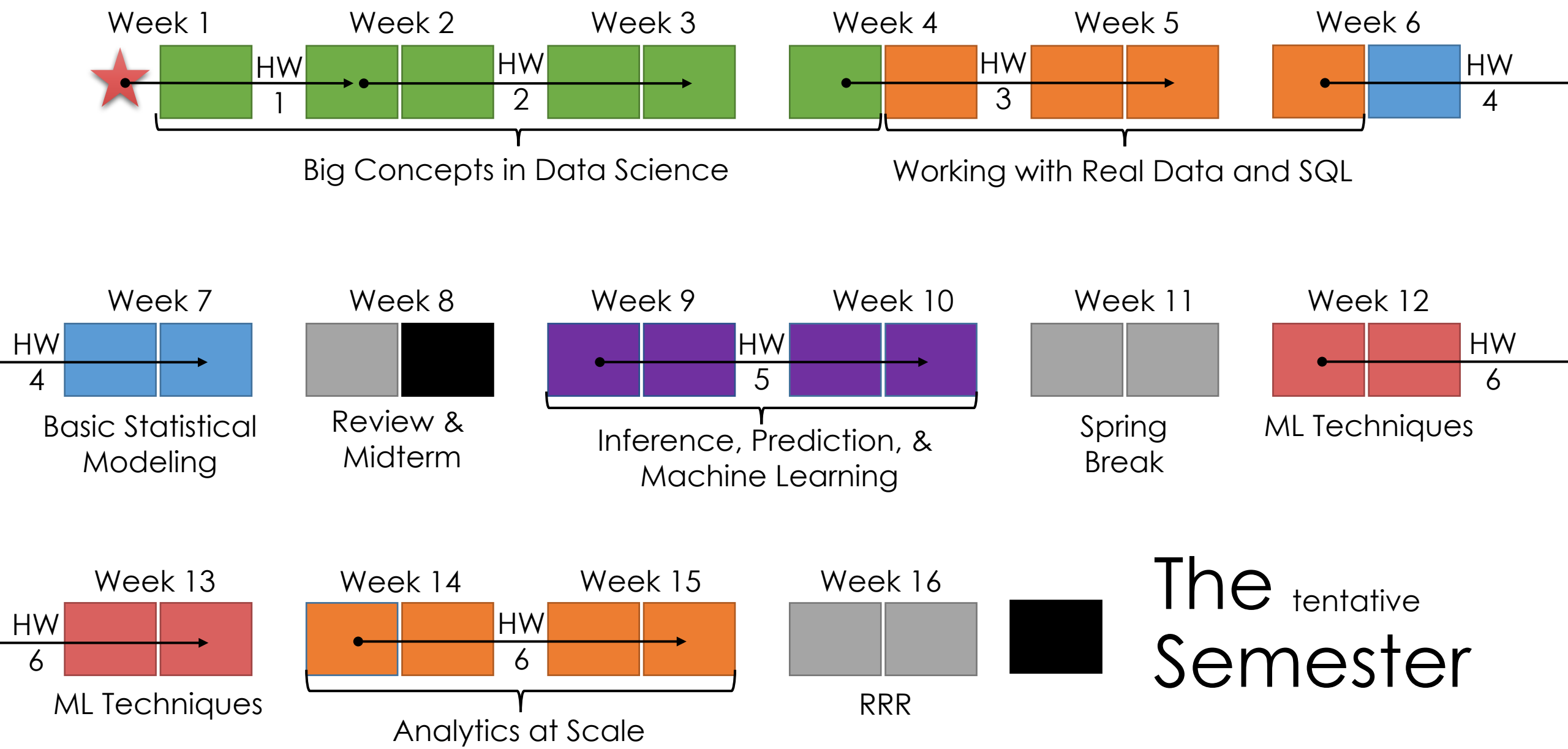
No single great book on data science

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



Several eBooks we will occasionally reference (optional)

- 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]



The tentative Semester

http://www.ds100.org/sp17/syllabus

Grades

[40%] 7 Homework: *be a data-scientist*

- 1 to 2 week long programming assignments

[13%] 13 Vitamins: *don't fall behind*

- Mini quizzes (1 per week of instruction)

[7%] 13 Labs: *improve computing skills*

- Completion graded

[15%] 1 Midterm: *checkpoint on progress*

- In class, healthy checkpoint

[25%] 1 Final

On Time Policy (don't be late)

- **5 days** of "slip-time" to be **used on homework** 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

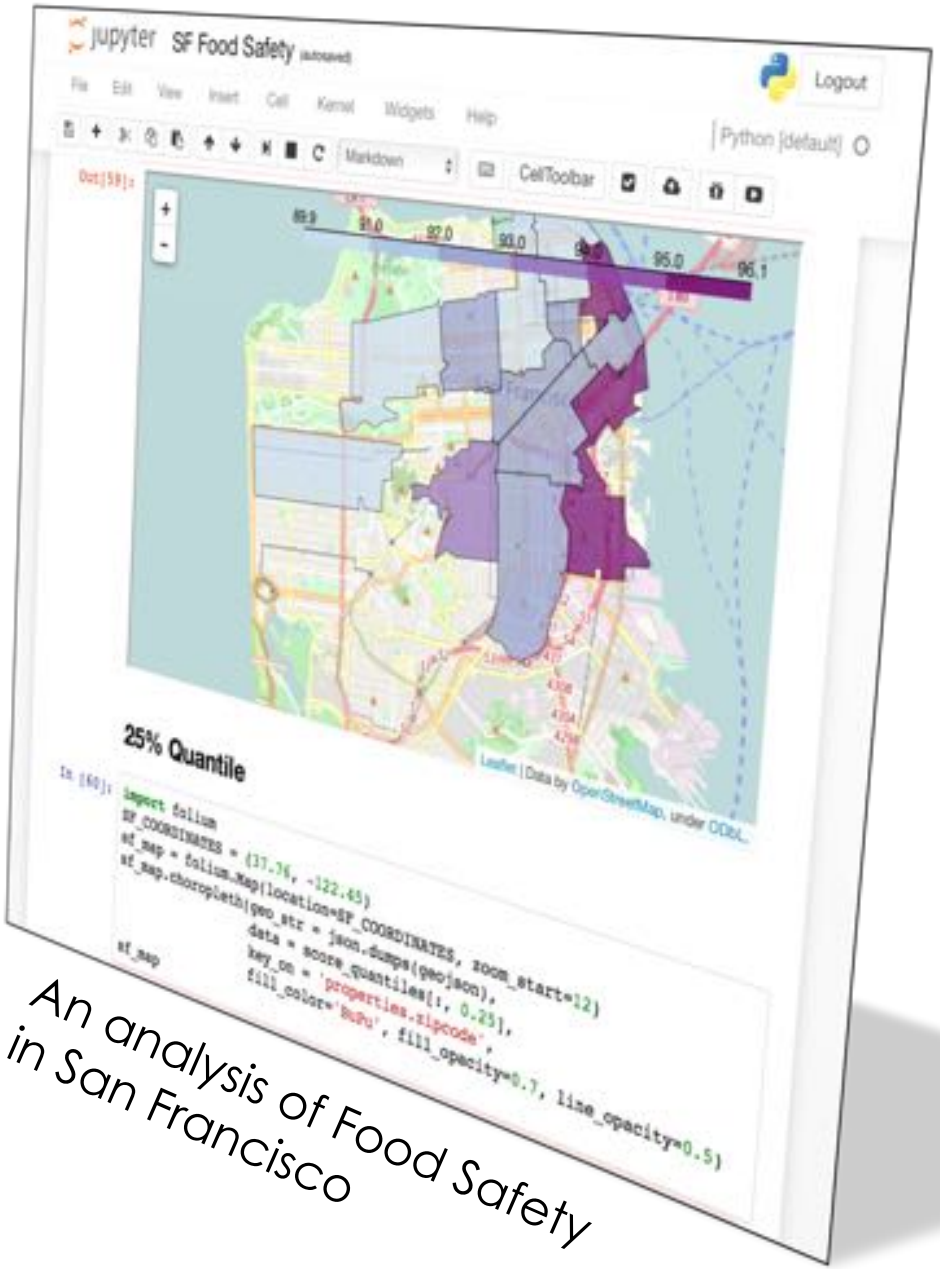
- Communication will be largely through Piazza
  - http://piazza.com/berkeley/spring2017/ds100

- We will also be updating the website with links to homework, lectures, and vitamins
  - http://www.ds100.org/sp17/

Today: Overview of DS100

Next Class

I will dig into the details by working through a real-world data science exercise using Python



An analysis of Food Safety in San Francisco