Computer Science

CSC380: Principles of Data Science

Introduction and Course Overview

Prof. Jason Pacheco

TA: Enfa Rose George

TA: Saiful Islam Salim

Outline

- COVID-19 Precautions
- Data Science Introduction
- Course Overview

COVID-19 Precautions

- **Mask up in class**

- The vaccines are very safe and very effective

- Notify me if you fall ill and think it will impact coursework

If we are forced to go remote

- I will schedule Zoom lectures
- They will be accessible via D2L
- I will notify everyone by email

Data Science Job Market

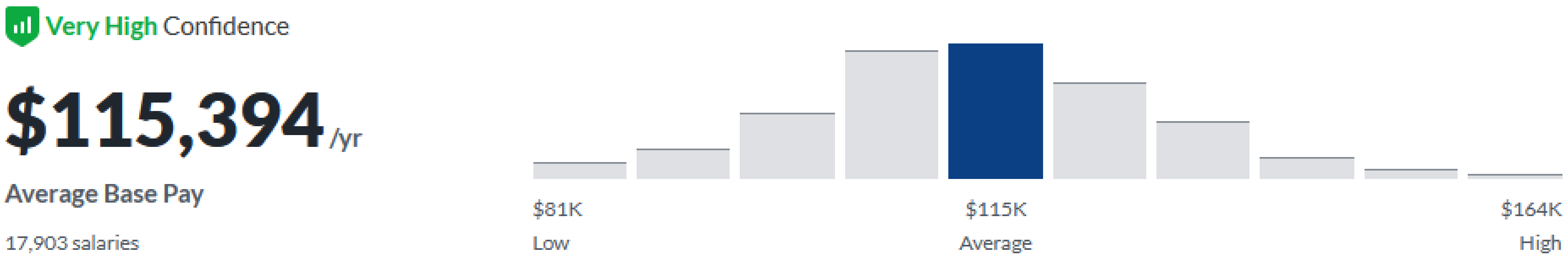
A search of "data scientist" jobs in the US (on 8/20/2021) shows…

Many job options available

- Indeed: 42,000+ jobs
- Glassdoor: 24,000+ jobs
- LinkedIn: 63,000+ jobs

2021's #2 best job in America, according to Glassdoor.com (after Java Developer)

Lucrative pay (Glassdoor)

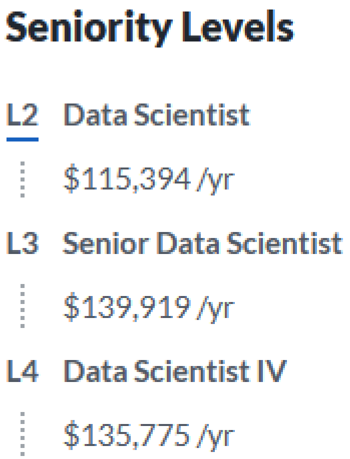


Very High Confidence

$115,394 /yr

Average Base Pay

17,903 salaries



Seniority Levels

L2 Data Scientist

$115,394 /yr

L3 Senior Data Scientist

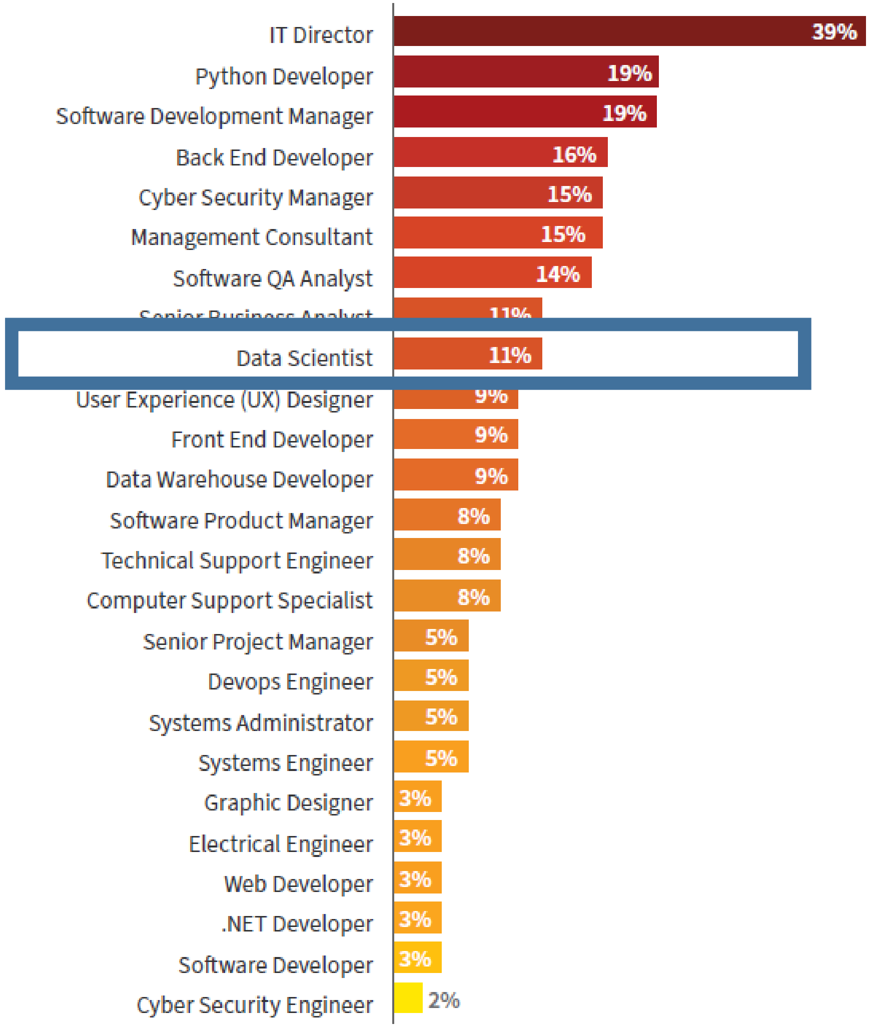
$139,919 /yr

L4 Data Scientist IV

$135,775 /yr

Data Science Job Market

Among the top 10 fastest growing jobs in 2020

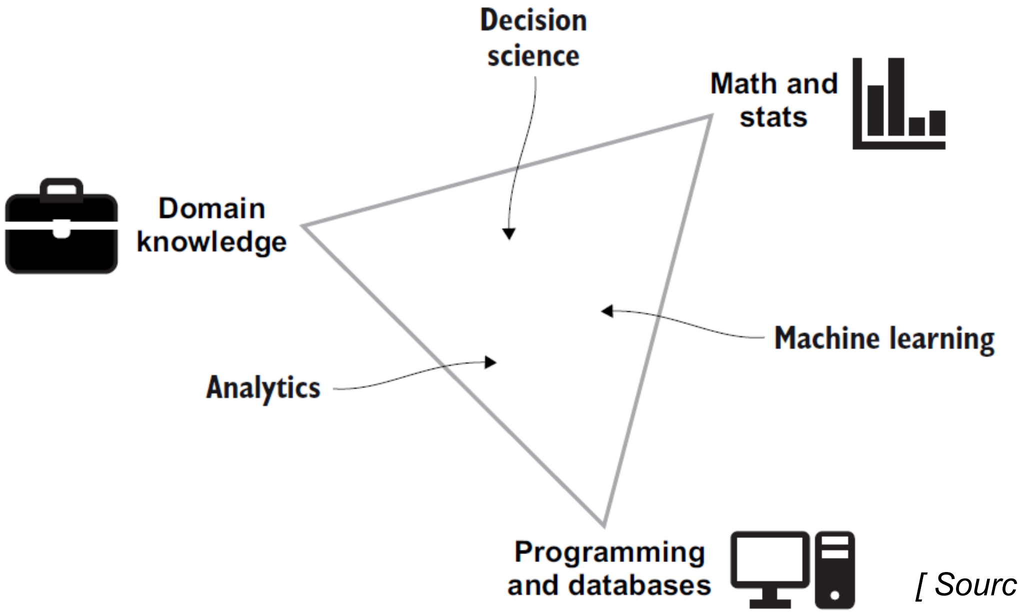


Source: Top Jobs in Dice Tech Q3 Report

Dice®

What is “Data Science”?

My Definition: *The process of using data to answer questions, extract knowledge, and predict future outcomes.*

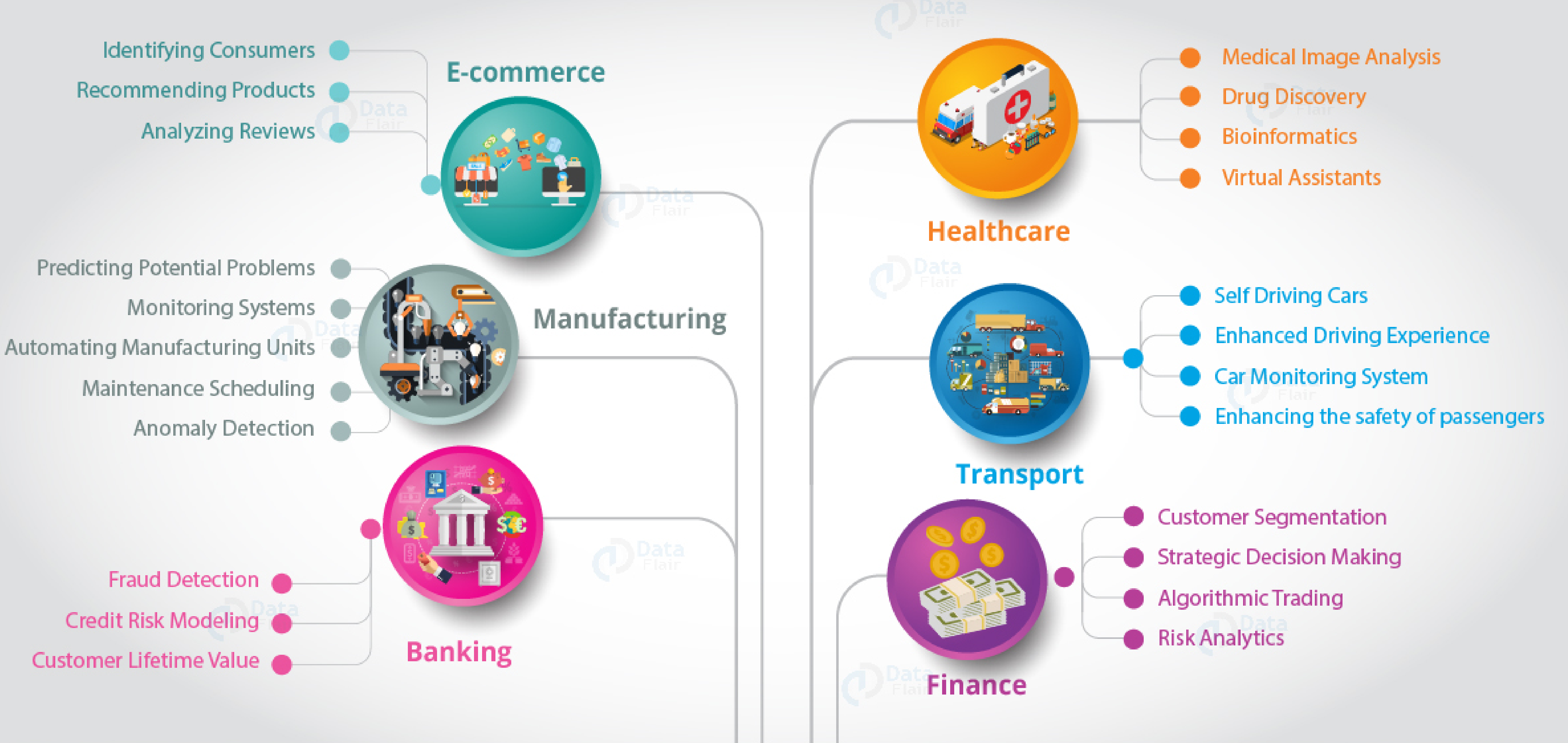


Caveat: I don’t *love* this figure since it leaves out **visualization.**

[Source: Robinson, E. and Nolis, J.]

Data Science Is:

- **Interdisciplinary**: Combines tools and techniques from Math / Statistics / CS
- **Exploratory**: Understanding data requires creative exploration and visualization
- **Applied Statistics & Probability** + extra stuff to handle, process, and visualize data



E-commerce

- Identifying Consumers
- Recommending Products
- Analyzing Reviews

Manufacturing

- Predicting Potential Problems
- Monitoring Systems
- Automating Manufacturing Units
- Maintenance Scheduling
- Anomaly Detection

Banking

- Fraud Detection
- Credit Risk Modeling
- Customer Lifetime Value

Healthcare

- Medical Image Analysis
- Drug Discovery
- Bioinformatics
- Virtual Assistants

Transport

- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

Finance

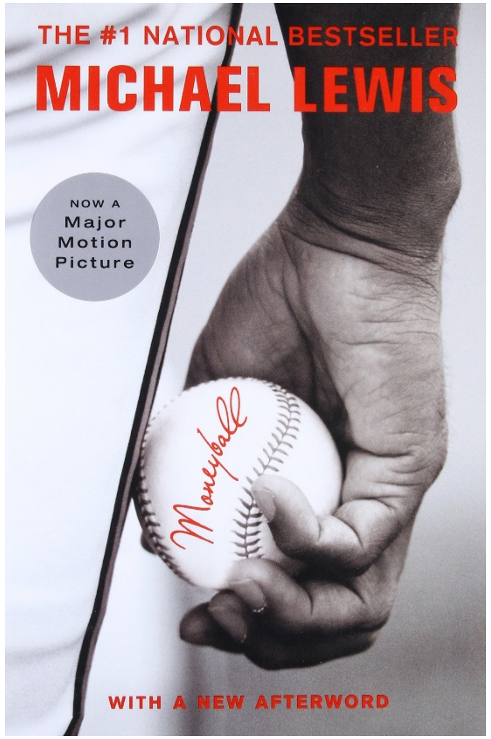
- Customer Segmentation
- Strategic Decision Making
- Algorithmic Trading
- Risk Analytics

Data Science Applications

Moneyball

Problem *How to assemble the best baseball team with a small budget?*

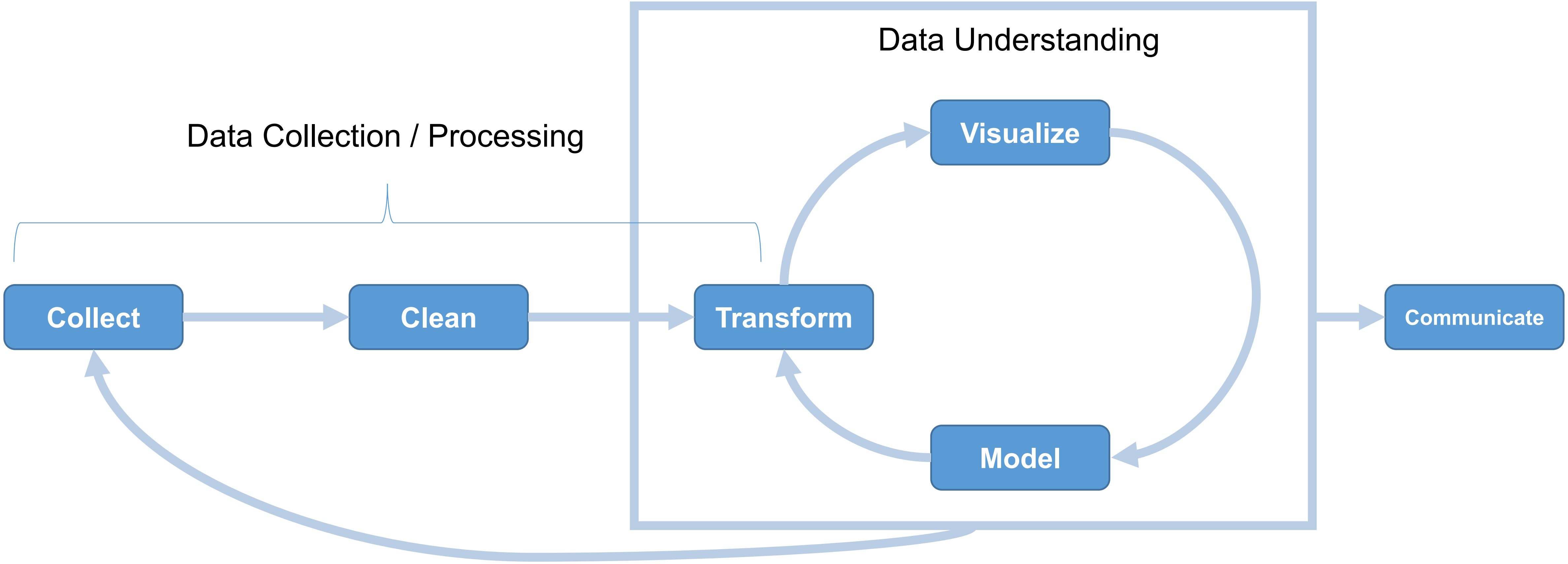
- 2002 Major League Baseball (MLB) draft
- Traditional team building relies on *scouts*
- Assumption: The collective wisdom of insiders is biased / flawed
- **SABRmetrics:** Data-driven and evidence-based approach to player quality evaluation
- *On-base %* and *Slugging %* are good indicators of offensive success
- Players with these “features” are cheaper compared to traditional statistics (stolen bases, runs batted in, batting average)



Moneyball: Impact

- In 2002 Oakland A’s ($44M budget) were competitive to the New York Yankees ($125M budget)
- Toronto Blue Jays hired full-time sabermetric analysts
- 2020 season “masters of Moneyball” Tampa Bay Rays reached world series with the 3rd lowest salary of all MLB
- In 2019 Liverpool Football/Soccer adopted this approach to nearly win the title (they lost to Manchester)
- Brad Pitt got a paycheck out of it for the movie (7.6/10 IMDB)…

Data Science Workflow



[Adapted from: Grolemund and Wickham, 2018]

Election Forecasting: Disclaimer

This is a class about data science it is **not** a class about politics. We will discuss election forecasting **only** in the context of data science and we will **ignore politics**.

Election Forecasting

Problem *Who will win the 2020 US presidential election?*

Details

- There are 2 primary candidates Donald Trump & Joe Biden*
- The *incumbent* (Trump) is the sitting president
- There are 50 states, each has a number of *electors*
- Each elector has a vote in the *electoral college*
- Electors for each state vote for the majority vote in that state
- Maine and Nebraska use a district method
- The winner has the majority of 538 electors (typically 270 or more votes)

** Secondary candidates do not have a realistic chance of winning, but cannot be ignored since they affect votes for primary candidates*

Election Forecasting: The Model

FiveThirtyEight uses a proprietary statistical model based on…

Poll aggregation model

Weight accounts for poll sample size, timeliness, historical accuracy

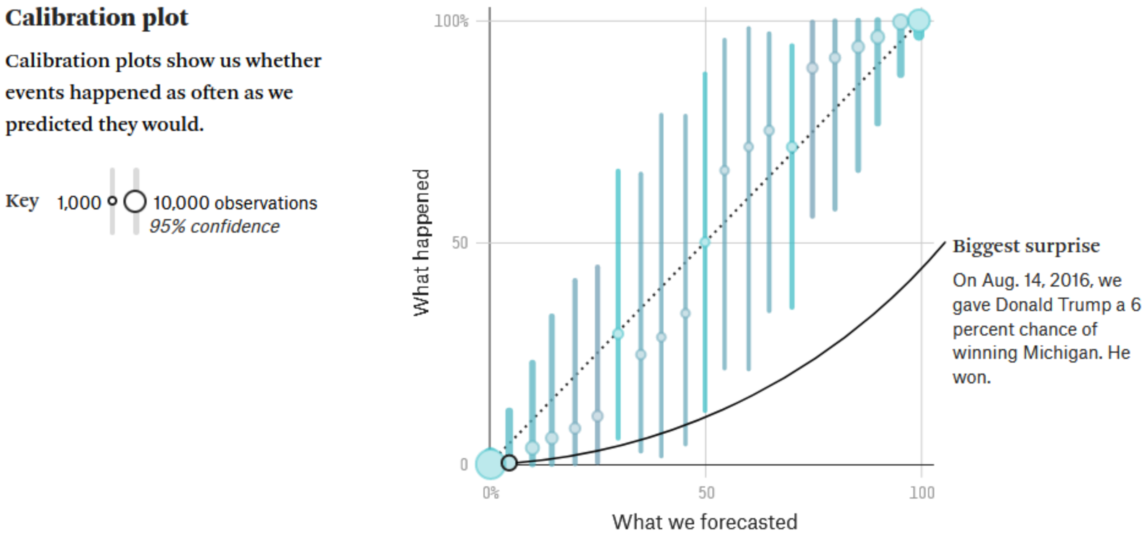
$$\text{prediction} = \sum_i \text{weight}_i \times \text{poll}_i + \text{random noise}$$

Additional model inputs

- States grouped by demographic subcategories
- Per capita income
- Age distribution of residents
- All features are *significant* to 85% level

Important properties of the model

- Predictive statements are *probabilistic*
- Assigns higher probability to extreme outliers
- Accounts for correlation among states / polls



Calibration plot

Calibration plots show us whether events happened as often as we predicted they would.

Key 1,000 10,000 observations *95% confidence*

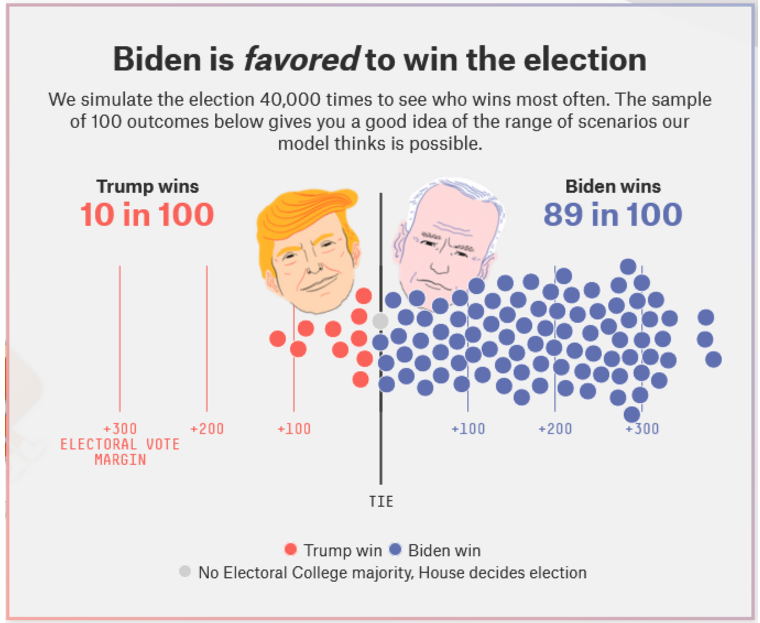
Biggest surprise

On Aug. 14, 2016, we gave Donald Trump a 6 percent chance of winning Michigan. He won.

FiveThirtyEight

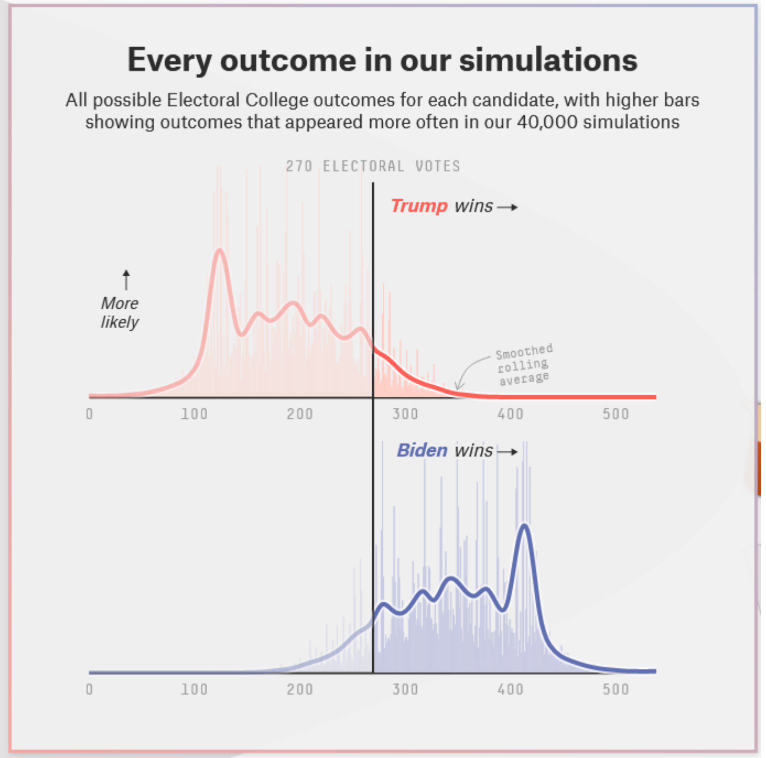
Election Forecasting: Visualizations

Generative (Bayesian) model allows simulation of random realizations…*



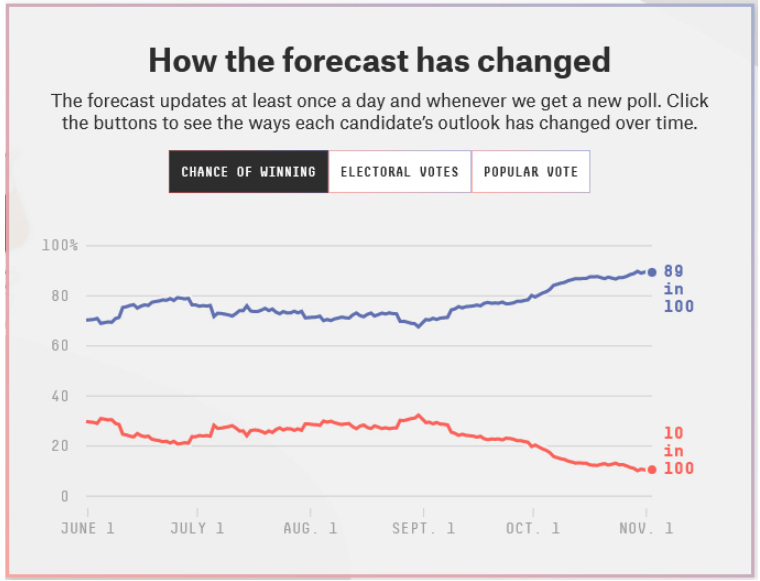
Biden is *favored* to win the election

We simulate the election 40,000 times to see who wins most often. The sample of 100 outcomes below gives you a good idea of the range of scenarios our model thinks is possible.



Every outcome in our simulations

All possible Electoral College outcomes for each candidate, with higher bars showing outcomes that appeared more often in our 40,000 simulations



How the forecast has changed

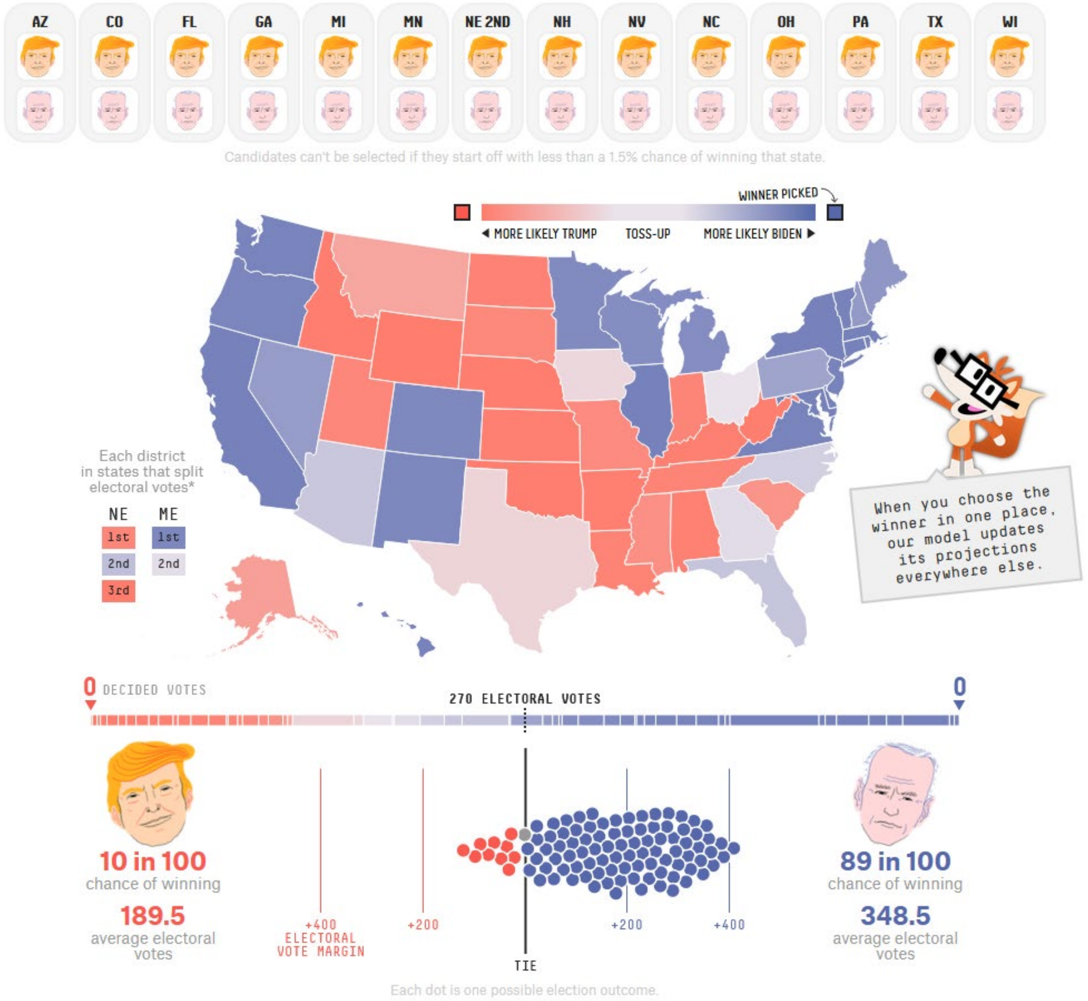
The forecast updates at least once a day and whenever we get a new poll. Click the buttons to see the ways each candidate's outlook has changed over time.

Click here to see visualizations

…visualizations targeted at communicating uncertainty about prediction.

Election Forecasting: Exploratory Analysis

Model also allows “what if” (e.g. counterfactual) analysis…



Candidates can't be selected if they start off with less than a 1.5% chance of winning that state.

Each district in states that split electoral votes*

When you choose the winner in one place, our model updates its projections everywhere else.

Each dot is one possible election outcome.

…this is a feature of model interpretability.

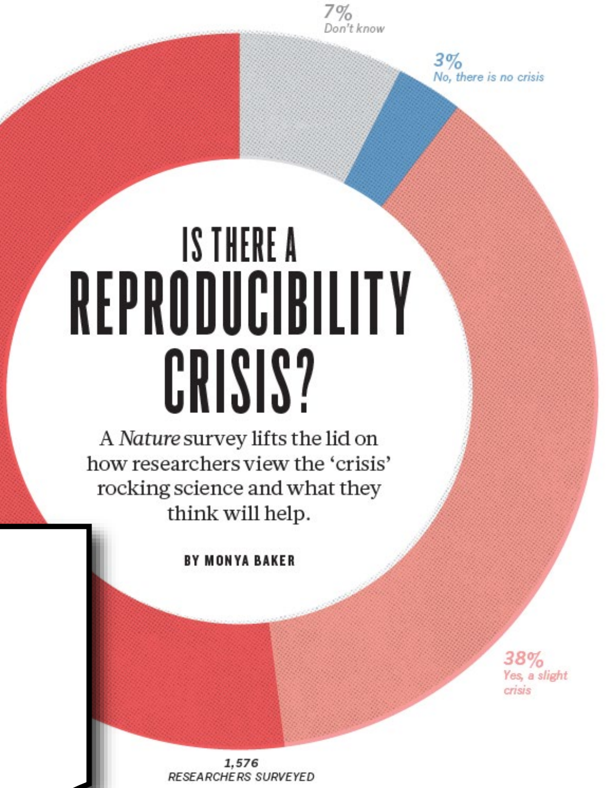
Bad Data Science & Statistics

Estimating the reproducibility of psychological science

Open Science Collaboration*

Why Most Published Research Findings Are False

John P. A. Ioannidis



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

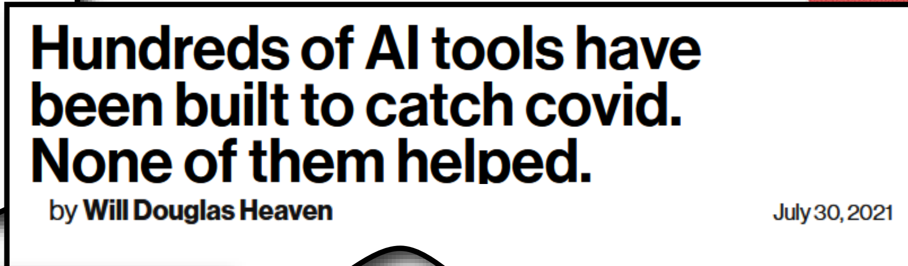
BY MONYA BAKER

Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema
Politecnico di Milano, Italy
maurizio.ferrari@polimi.it

Paolo Cremonesi
Politecnico di Milano, Italy
paolo.cremonesi@polimi.it

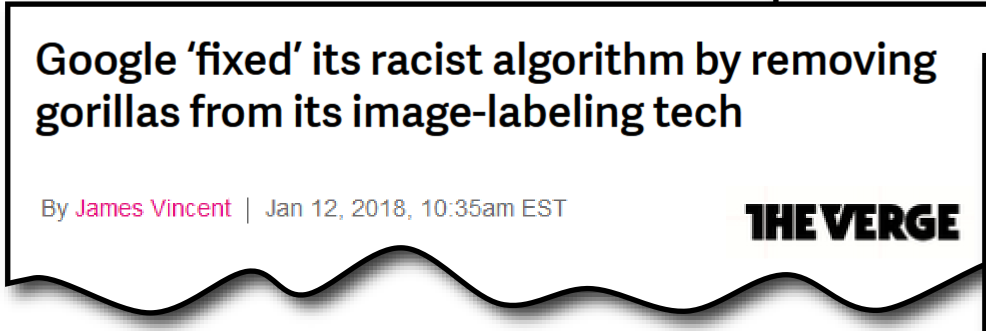
Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at



Hundreds of AI tools have been built to catch covid. None of them helped.

by **Will Douglas Heaven**

July 30, 2021



Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

By James Vincent | Jan 12, 2018, 10:35am EST

THE VERGE

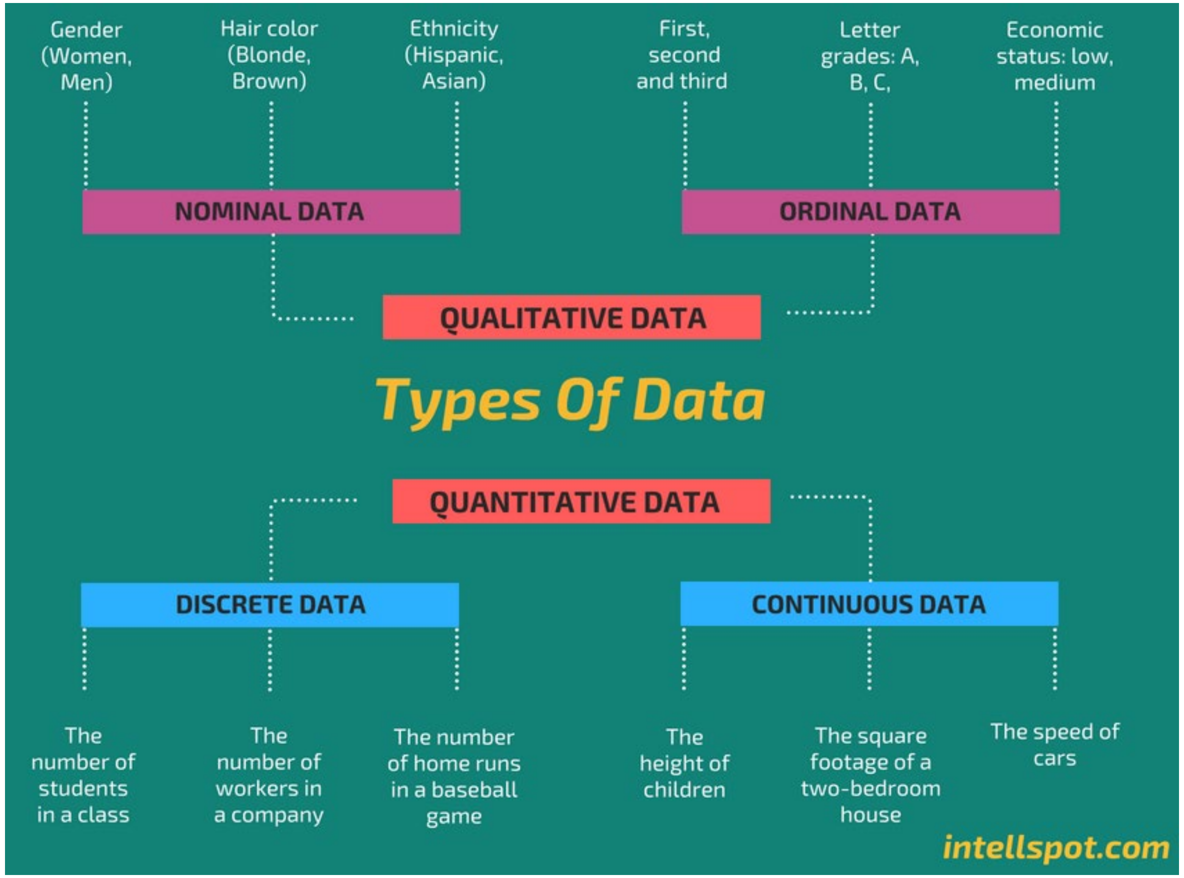
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

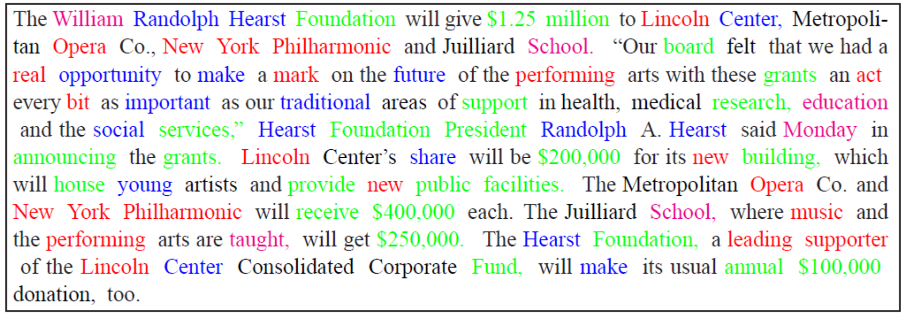
REUTERS

Types of Data

Data come in many forms, each requiring different approaches & models


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Natural Language



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Timeseries

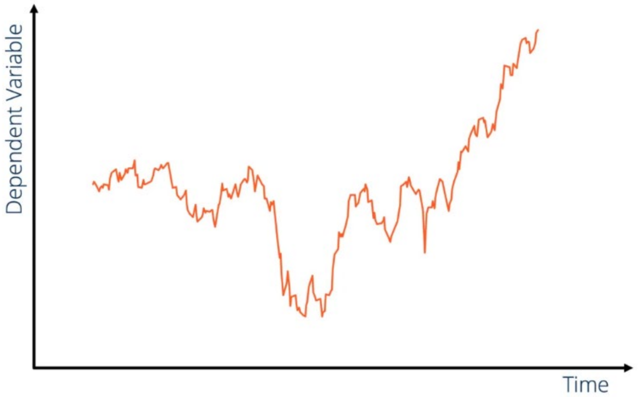
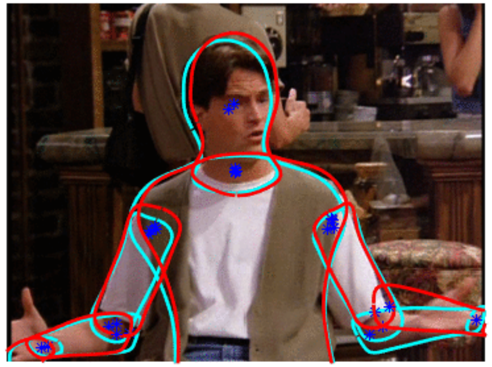


Image / Video



The number of types is endless, these are just some examples

Programming Languages for Data Science

Python and R are both standard for data science these days

R

python™

We will use Python for this course since you should already know it

Python Packages Covered

NumPy

scikit learn

matplotlib

Other Useful Python Packages

TensorFlow

pandas

Altair

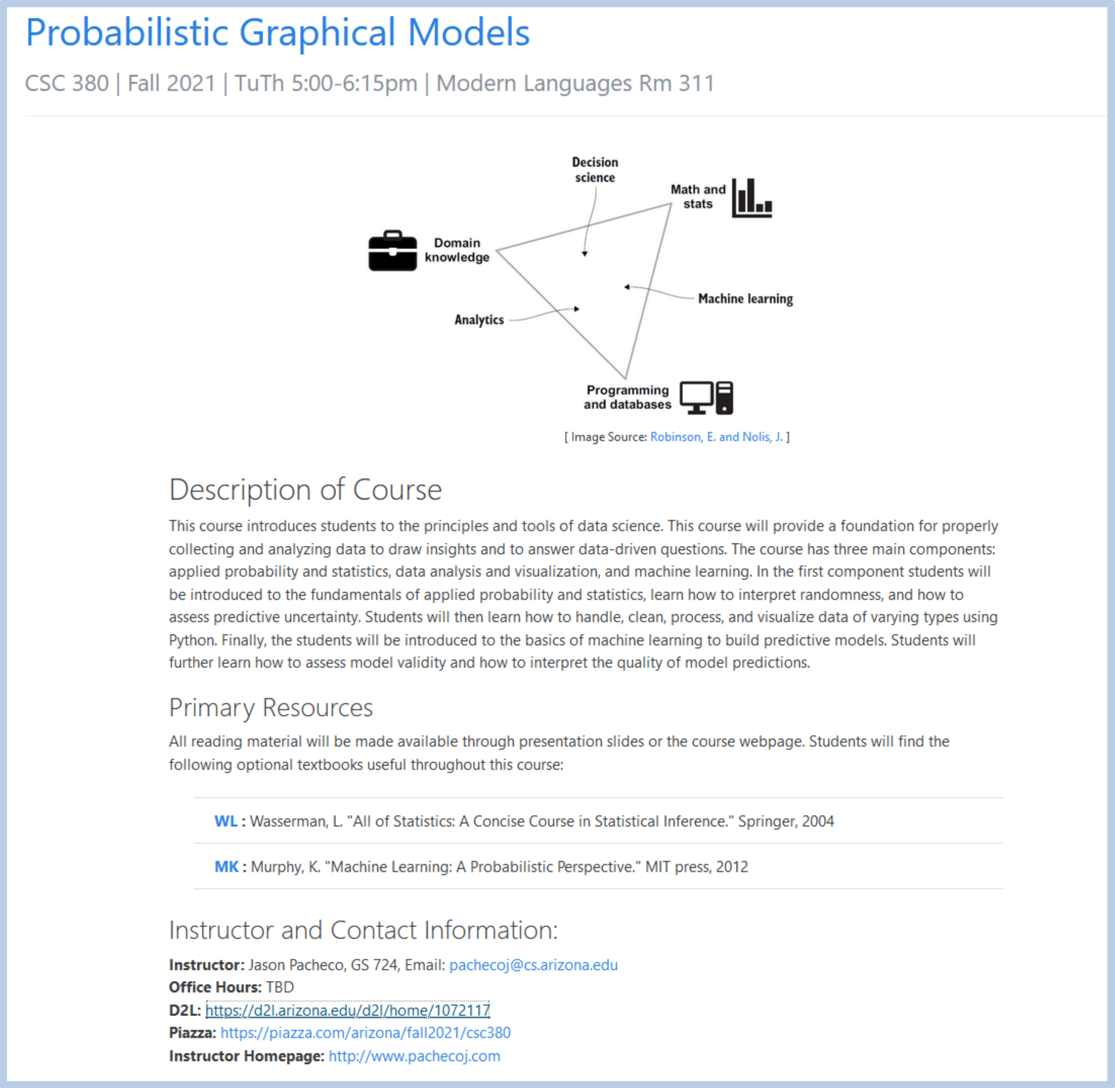
CYVERSE®

Transforming Science Through Data-Driven Discovery

- Created and primarily house here at **University of Arizona!**
- Now a full-fledged infrastructure for shared data science across many fields
- Cloud storage and computing computing resources
- Makes sharing data, code, etc. easy across large distributed teams
- Virtual machine environments for high performance computing
- Supports most standard languages / tools / etc.
- Can package code into containers and easily share with others

CyVerse is supported by the National Science Foundation under Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442

Course Overview: Resources



Probabilistic Graphical Models

CSC 380 | Fall 2021 | TuTh 5:00-6:15pm | Modern Languages Rm 311

[Image Source: Robinson, E. and Nolis, J.]

Description of Course

This course introduces students to the principles and tools of data science. This course will provide a foundation for properly collecting and analyzing data to draw insights and to answer data-driven questions. The course has three main components: applied probability and statistics, data analysis and visualization, and machine learning. In the first component students will be introduced to the fundamentals of applied probability and statistics, learn how to interpret randomness, and how to assess predictive uncertainty. Students will then learn how to handle, clean, process, and visualize data of varying types using Python. Finally, the students will be introduced to the basics of machine learning to build predictive models. Students will further learn how to assess model validity and how to interpret the quality of model predictions.

Primary Resources

All reading material will be made available through presentation slides or the course webpage. Students will find the following optional textbooks useful throughout this course:

WL : Wasserman, L. "All of Statistics: A Concise Course in Statistical Inference." Springer, 2004

MK : Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012

Instructor and Contact Information:

Instructor: Jason Pacheco, GS 724, Email: pachecoj@cs.arizona.edu
Office Hours: TBD
D2L: https://d2l.arizona.edu/d2l/home/1072117
Piazza: https://piazza.com/arizona/fall2021/csc380
Instructor Homepage: http://www.pachecoj.com

Resources accessible on course website
pachecoj.com/courses/csc380_fall21/

Specific resources

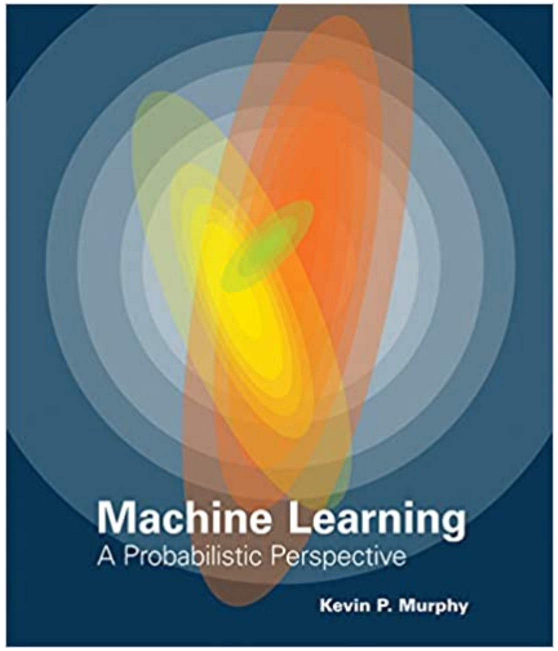
- D2L for assignment submission
- Piazza for **all** communication
- Readings and electronic textbooks
- Lecture slides (posted after class)

Every lecture accompanied by reading

- These will not be graded but are required
- Homeworks will incorporate material from readings
- Reading for today's lecture:

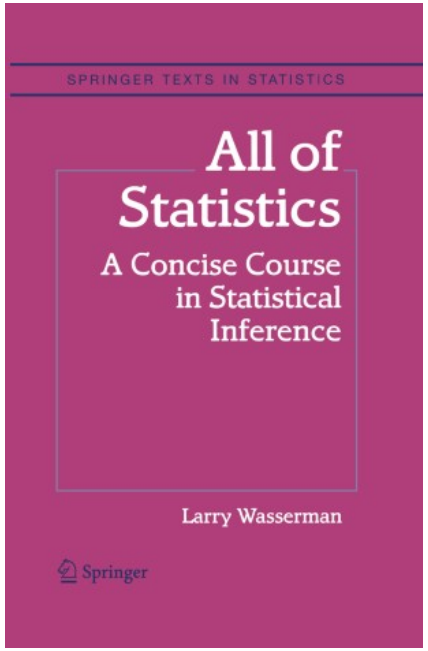
Robinson and Nolis, "What's Data Science?"

Textbooks



Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012

[(UA Library)]



SPRINGER TEXTS IN STATISTICS

All of Statistics

A Concise Course in Statistical Inference

Larry Wasserman

Springer

Wasserman, L. “All of Statistics.” Springer, 2004

[(Springer)]

Additional readings on the course webpage

Course TAs

Your friendly course TAs…

Saiful Islam Salim

saifulislam@email.arizona.edu

Enfa Rose George

enfageorge@email.arizona.edu

Expected Skills

- This class will require a fair amount of math
  - Probability and Statistics (first few lectures)
  - Single-variable Calculus
  - Linear Algebra (two-lecture overview)

- This class will require a fair amount of **coding**
  - Reading in / cleaning / visualizing data
  - Simulating random processes
  - Training and evaluating machine learning models

- Early assignments will be mostly math, later will be coding

Course Overview

Course Objective *Introduction to basic concepts in data science and machine learning.*

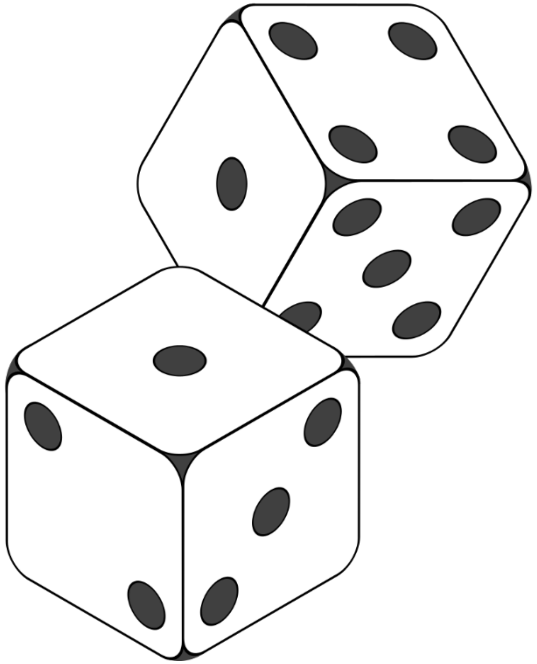
Probability and Statistics	Data Handling and Visualization	Machine Learning	Ethics and Fairness
Random events / variables, distributions / densities, moments, descriptive stats, estimation	Reading & cleaning, transformation & preprocessing, visualization	Predictive models, supervised learning, unsupervised learning, model checking	Data privacy, ethics, fairness

Probability and Statistics

Suppose we roll <u>two fair dice</u>…

- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?

*… this is an **experiment** or **random process**.*



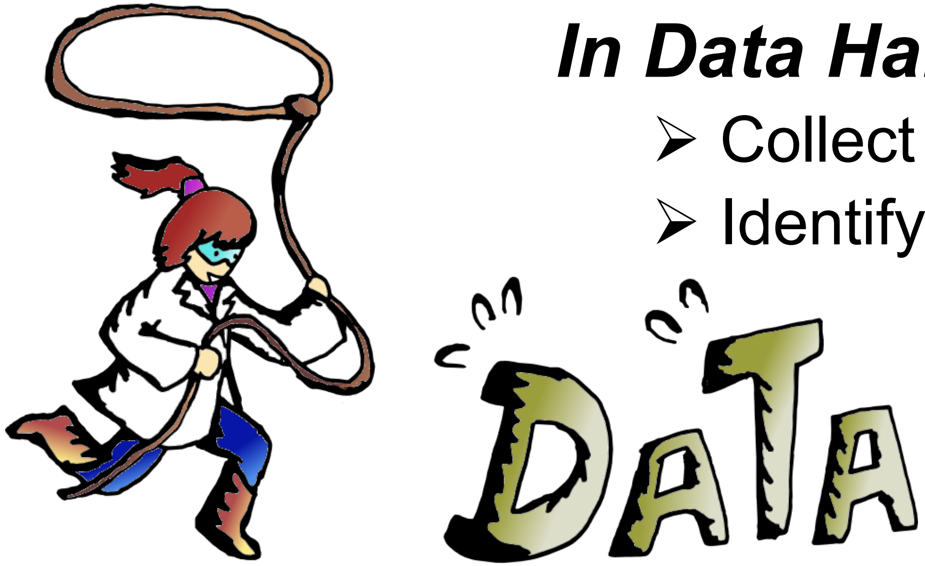
We will learn how to…

- Mathematically formulate outcomes and their probabilities?
- Describe characteristics of random processes
- Estimate unknown quantities (e.g. are the dice actually fair?)
- Characterize the uncertainty in random outcomes
- Identify and measure dependence among random quantities

Data Handling and Visualization

In Data Handling we will learn to…

- Collect data through population sampling
- Identify and avoid biased population samples

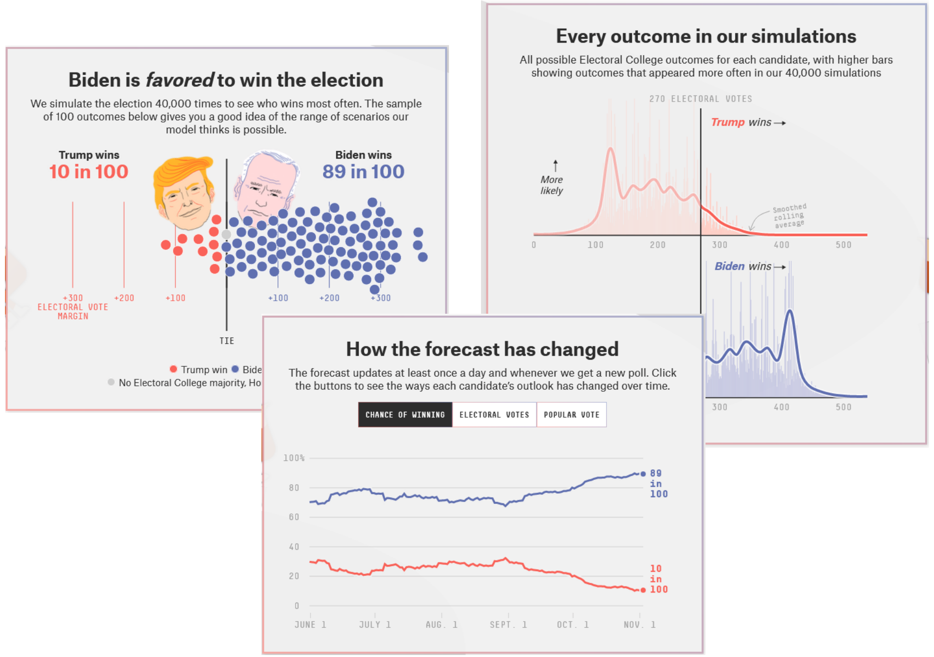


- Clean data and correct errors
- Transform and preprocess data (*wrangling*)

[Image Source: Code A Star]

In Data Visualization we will learn…

- Why visualization is important
- Exploratory data analysis
- Common forms of visualization
- Pitfalls and gotchas


Biden is *favored* to win the election

We simulate the election 40,000 times to see who wins most often. The sample of 100 outcomes below gives you a good idea of the range of scenarios our model thinks is possible.

Every outcome in our simulations

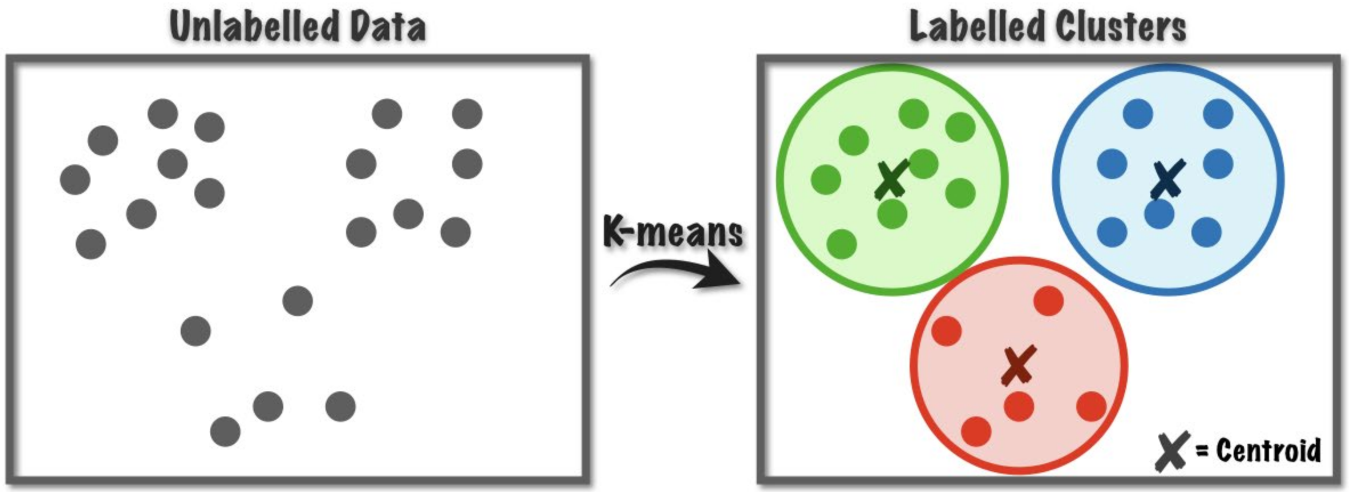
All possible Electoral College outcomes for each candidate, with higher bars showing outcomes that appeared more often in our 40,000 simulations

How the forecast has changed

The forecast updates at least once a day and whenever we get a new poll. Click the buttons to see the ways each candidate's outlook has changed over time.

Machine Learning

How do use data to learn underlying patterns and predict unknowns?



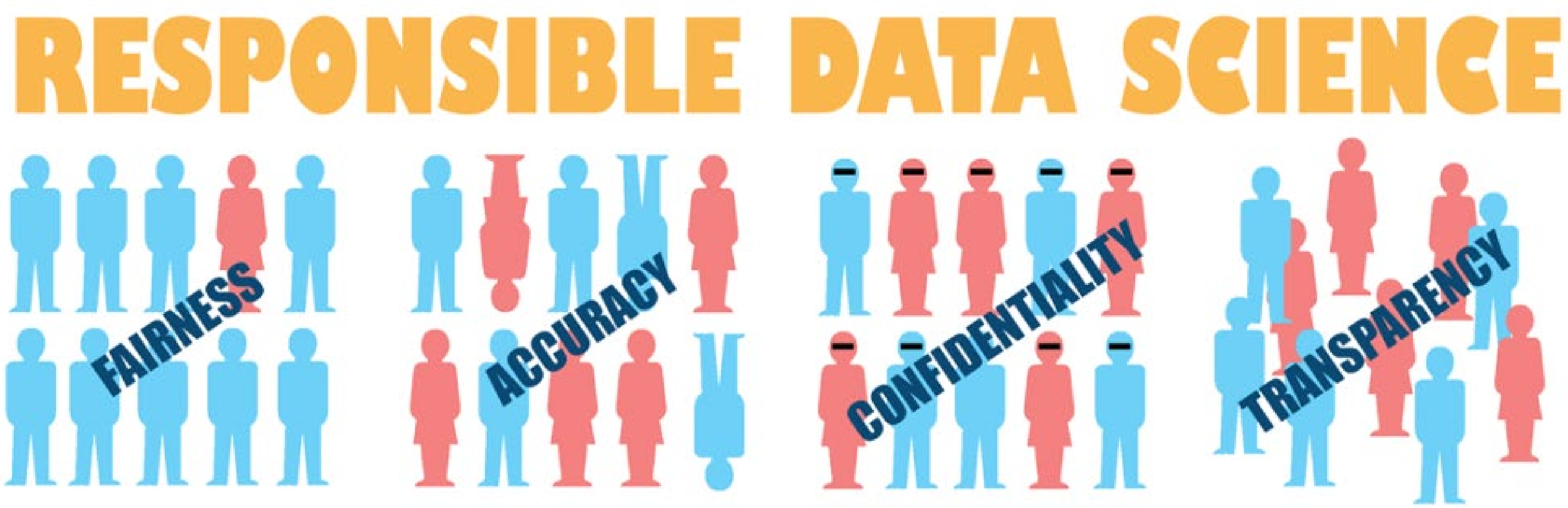
In Machine Learning we will learn…

- ➢ Principles of prediction
- ➢ Proper partitioning of training / validation / test data
- ➢ Unsupervised vs. supervised learning
- ➢ Linear and nonlinear models
- ➢

We will preface this section with a Linear Algebra primer

[Image Source: Towards Data Science]

Ethics and Fairness



In Ethics and Fairness we will learn…

- The principles of data privacy
- Identifying, measuring, and ensuring fairness
- Measuring accuracy through model validation and checking
- Transparent communication of findings and predictions

[Image Source: Responsible Data Science Consortium]

Assignments / Exams / Grading

9 Homeworks + Midterm + Final Exam

Homeworks

- Generally, you will have 1 week per assignment
- There will be an assignment nearly every week
- Assignment *typically* out and due on Thursdays
- Grades by one week after due date
- Some irregularity around holidays
- No assignment over Thanksgiving break

Grading Breakdown

- Homework: 60%
- Midterm: 20%
- Final: 20%

First assignment out one week from today

Attendance is ***strongly*** *encouraged, but not explicitly graded*

Late Policy

Late submissions impact other students, delay grading, and delay solutions

But sometimes we need a little extra time…

- **No more than 1** assignment **no more than 1** day late without penalty
- All subsequent late assignments will receive a zero score
- D2L will accept late assignments but they will be flagged

If you are struggling with time…

- Notify me (Piazza) at least 24hrs before the deadline
- Submit the best version of what you have by the deadline
- In general I **will not** grant extra time, and will grade what has been submitted

*If you submit **all** assignments on time, it may benefit your final grade*

Academic Integrity

Assignments are to be done independently…

If I or the TA suspects you of having cheated

- You will be notified immediately
- We will have a conference where you can plead your case
- If I am not swayed then you receive a zero for the assignment
- There is an appeals process if you are confident in your case

Bottom line don’t cheat

“*Office*” Hours

- Office hours will be held via Zoom, accessible via D2L
- I will hold two 1.5hr sessions each week
- The final office hour schedule will be announced next week
- If you have a conflict with the schedule, let me know (Piazza)

Mental Wellbeing

Some occasional stress / depression / anxiety is normal, but sometimes you may need extra help

- Non-emergency UA resources at Counseling & Psych Services Mon-Fri
  - Phone: 520-621-3334
  - Web: https://health.arizona.edu/counseling-psych-services

- Emergency resources in Tucson in this Google Doc

Inclusivity

I want to foster a comfortable and inclusive classroom experience

Please let me know if you feel excluded in any way, e.g.

- "Alice-and-Bob" style examples of material
- Improper use of pronouns
- Microagressions
- Miscellaneous statements / interactions

You can message me on Piazza or discuss in person