Computer Science

CSC380: Principles of Data Science

Introduction and Course Overview

Kyoungseok Jang

Course instructors

Chicheng Zhang
chichengz@arizona.edu

Kyoungseok Jang
ksajks@arizona.edu

- Dr. Zhang will cover lectures before Feb. 28
- Dr. Jang will cover lectures after Mar. 2

Outline

- Data Science Introduction
  - What is data science?
  - Case studies

- Course Overview
  - Resources
  - Grading policy
  - What you will learn

COVID-19 Precautions

- Masks are not required but recommended.

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

Data Science Introduction

Data Science Job Market

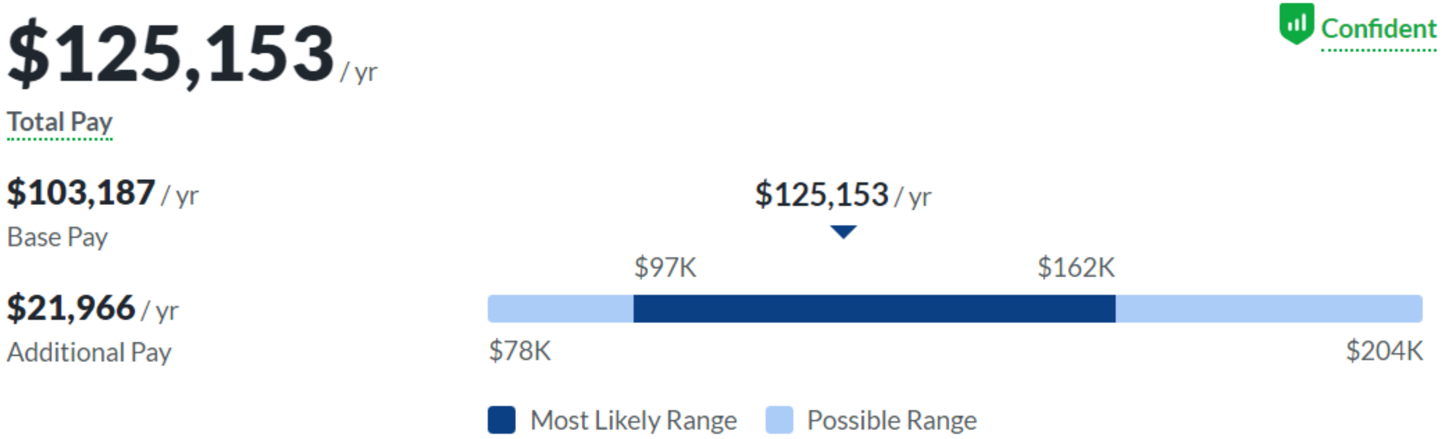
A search of "data scientist" jobs in the US (on 9/15/2022) shows…

Many job options available

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

2022's #3 best job in America, according to Glassdoor.com (2021's #2)

Lucrative pay (Glassdoor)

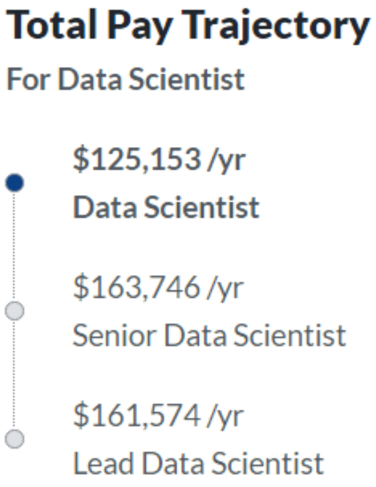


$125,153 / yr
Total Pay

$103,187 / yr
Base Pay

$21,966 / yr
Additional Pay

Confident

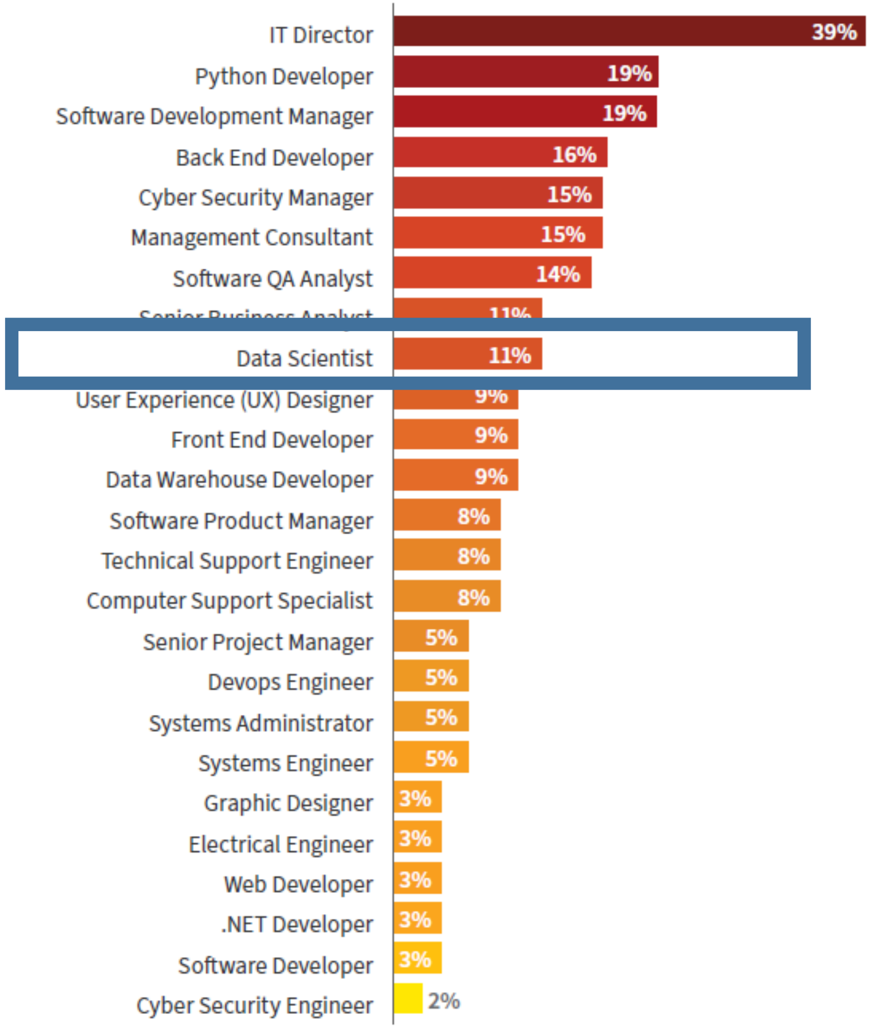


Total Pay Trajectory
For Data Scientist

- $125,153 /yr Data Scientist
- $163,746 /yr Senior Data Scientist
- $161,574 /yr Lead Data Scientist

Data Science Job Market

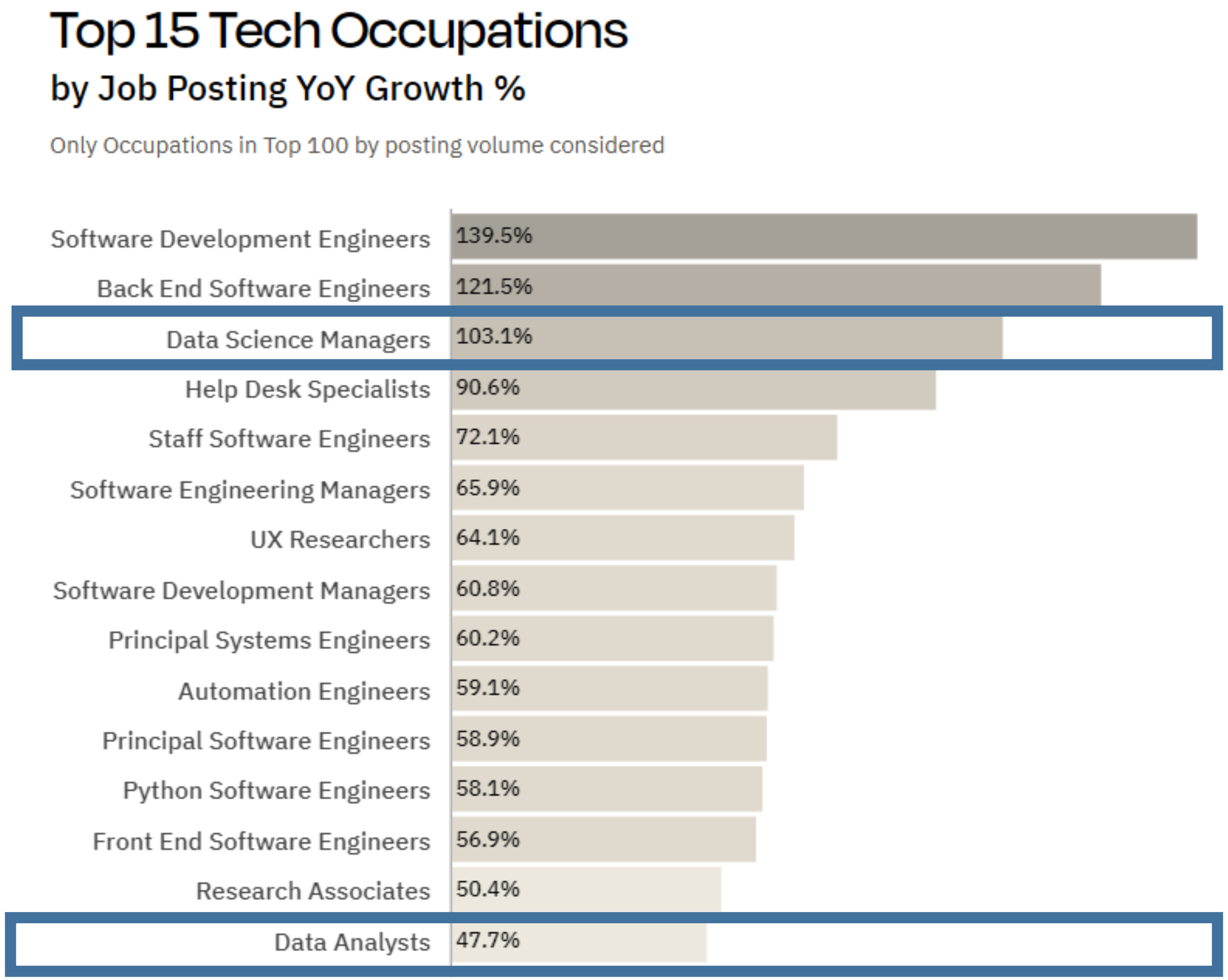
Among the top 10 fastest growing jobs in 2020



Source: Top Jobs in Dice Tech Q3 Report

Data Science Job Market

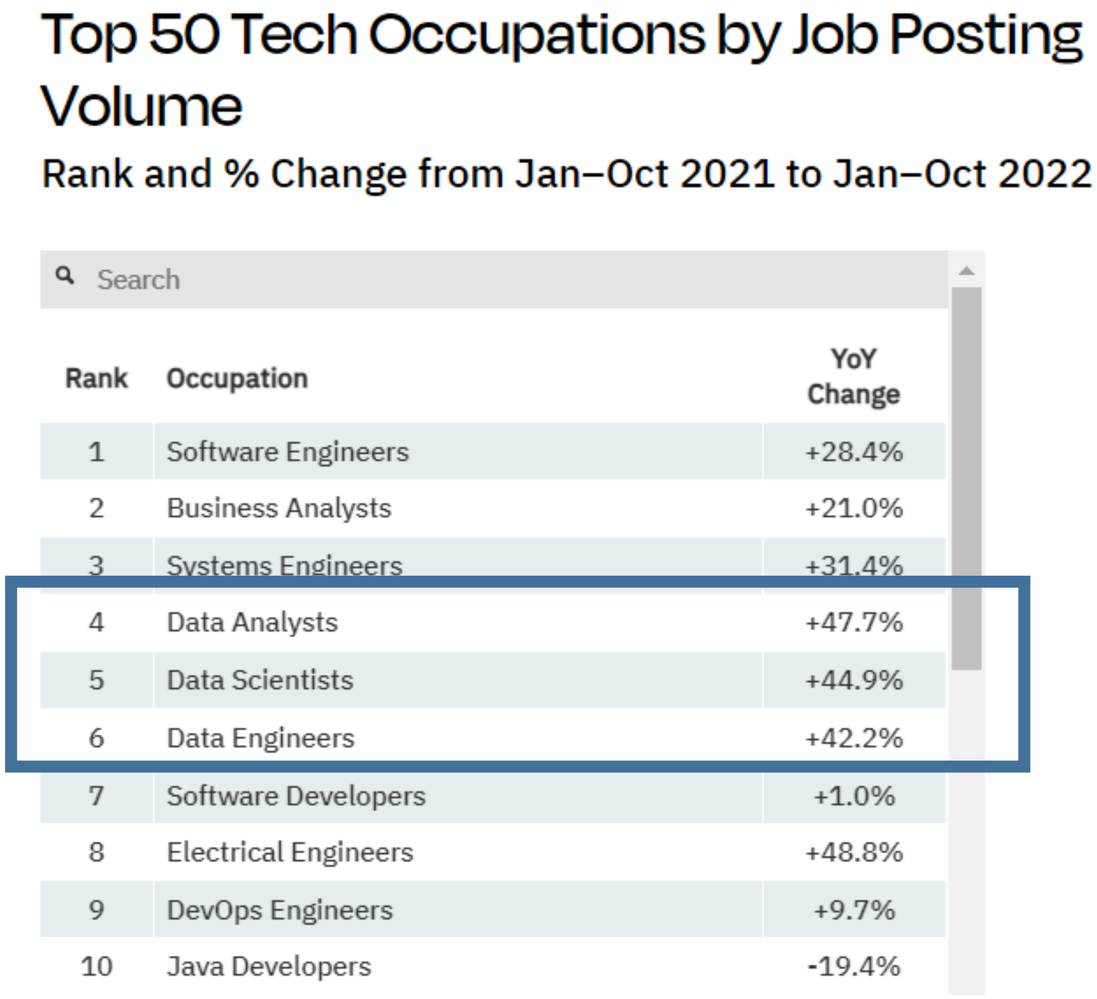
Now Data Science 'Manager' is top 3 fastest growing jobs in 2022



Top 15 Tech Occupations

by Job Posting YoY Growth %

Only Occupations in Top 100 by posting volume considered



Top 50 Tech Occupations by Job Posting Volume

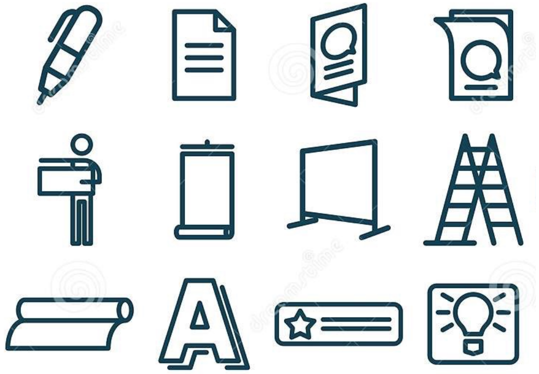
Rank and % Change from Jan–Oct 2021 to Jan–Oct 2022

Rank	Occupation	YoY Change
1	Software Engineers	+28.4%
2	Business Analysts	+21.0%
3	Systems Engineers	+31.4%
4	Data Analysts	+47.7%
5	Data Scientists	+44.9%
6	Data Engineers	+42.2%
7	Software Developers	+1.0%
8	Electrical Engineers	+48.8%
9	DevOps Engineers	+9.7%
10	Java Developers	-19.4%

What is “Data Science”?

Our Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon



Examples:

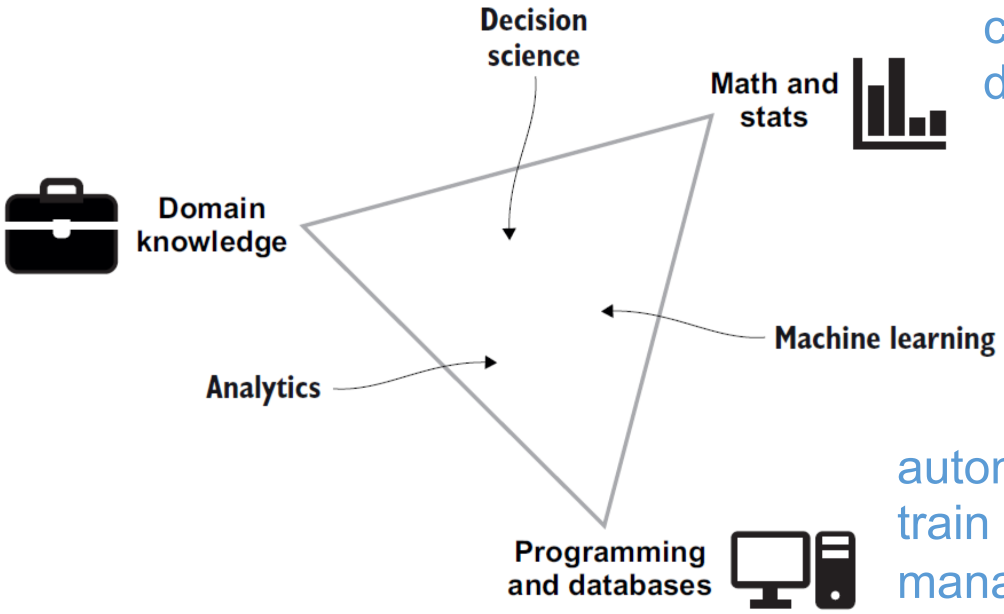
- Do people in college towns tend to buy more notebooks than people in other areas?
- Find out top-10 sales categories for each age group.
- Summarize product reviews w.r.t. product quality, customer service, etc.
- If we recommend pens to users from college town, how much will it increase our revenue?

What is “Data Science”?

Our Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon:
how customers behave

manufacturing:
how the process works



compute average/variance
design statistical models

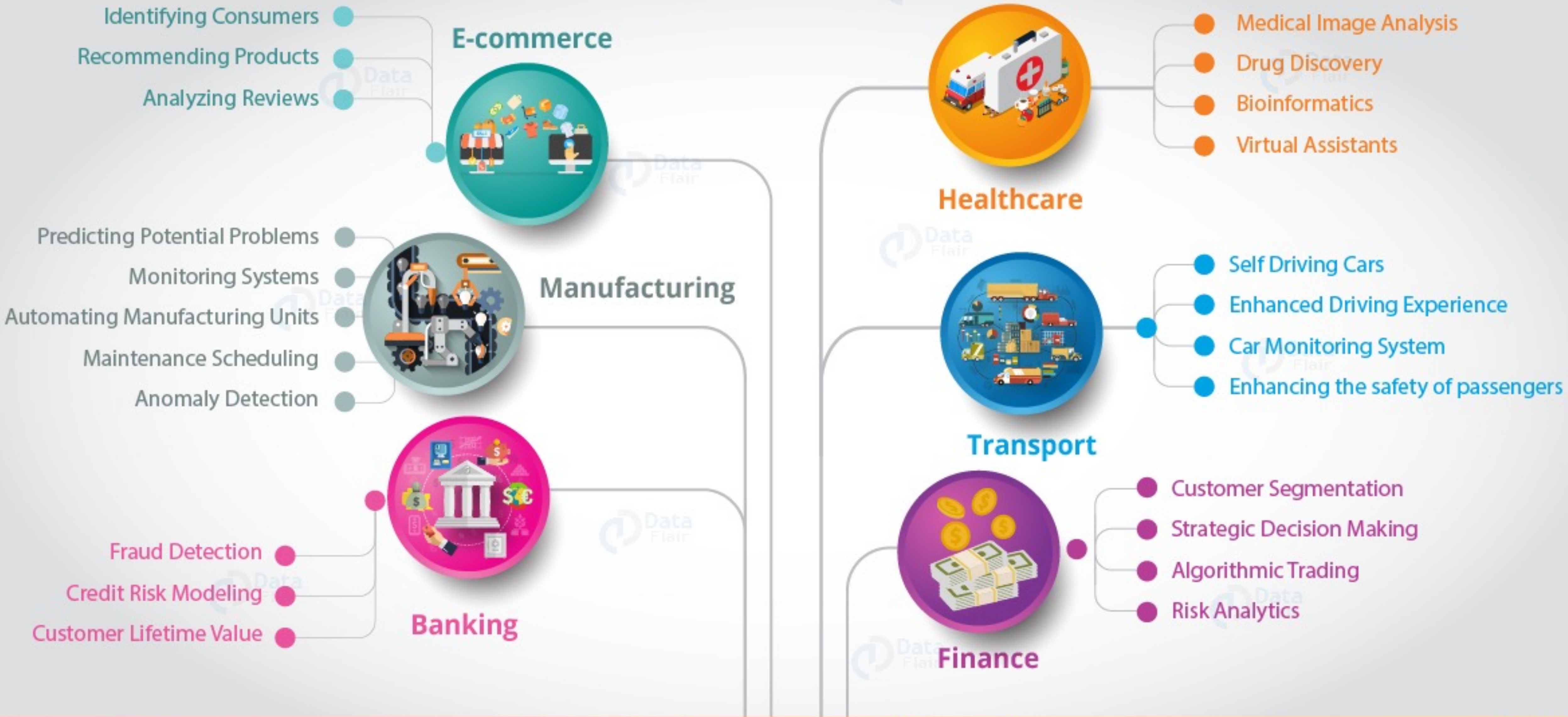
Caveat: this figure leaves out the important aspect of **visualization.**

automate analyses
train models
manage data (store, retrieve)

[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

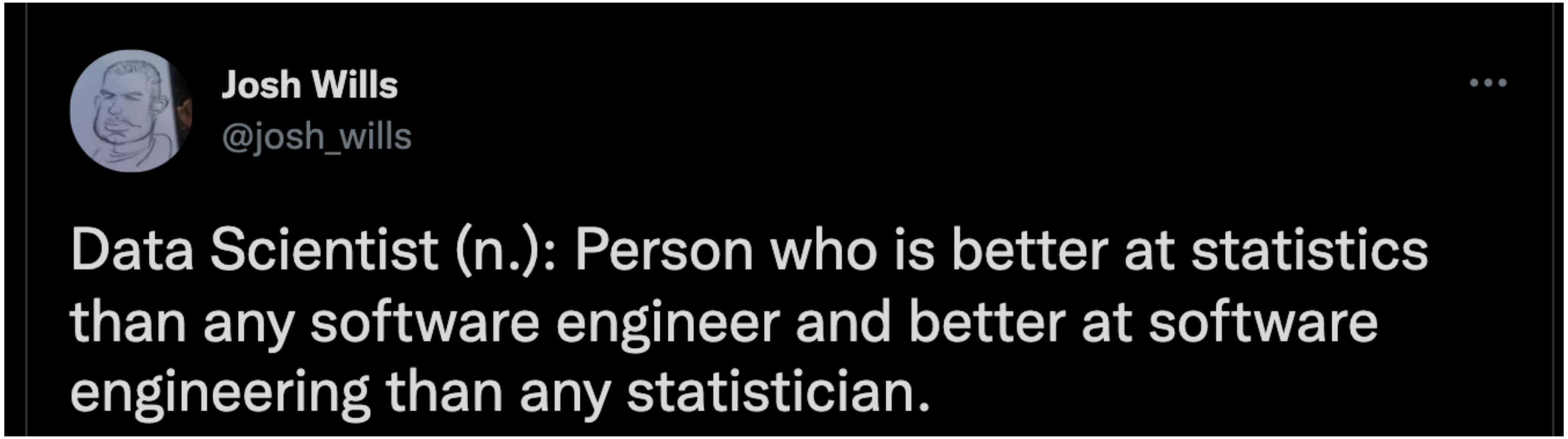
- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

Finance

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

Data Science Applications

Who is a Data Scientist?



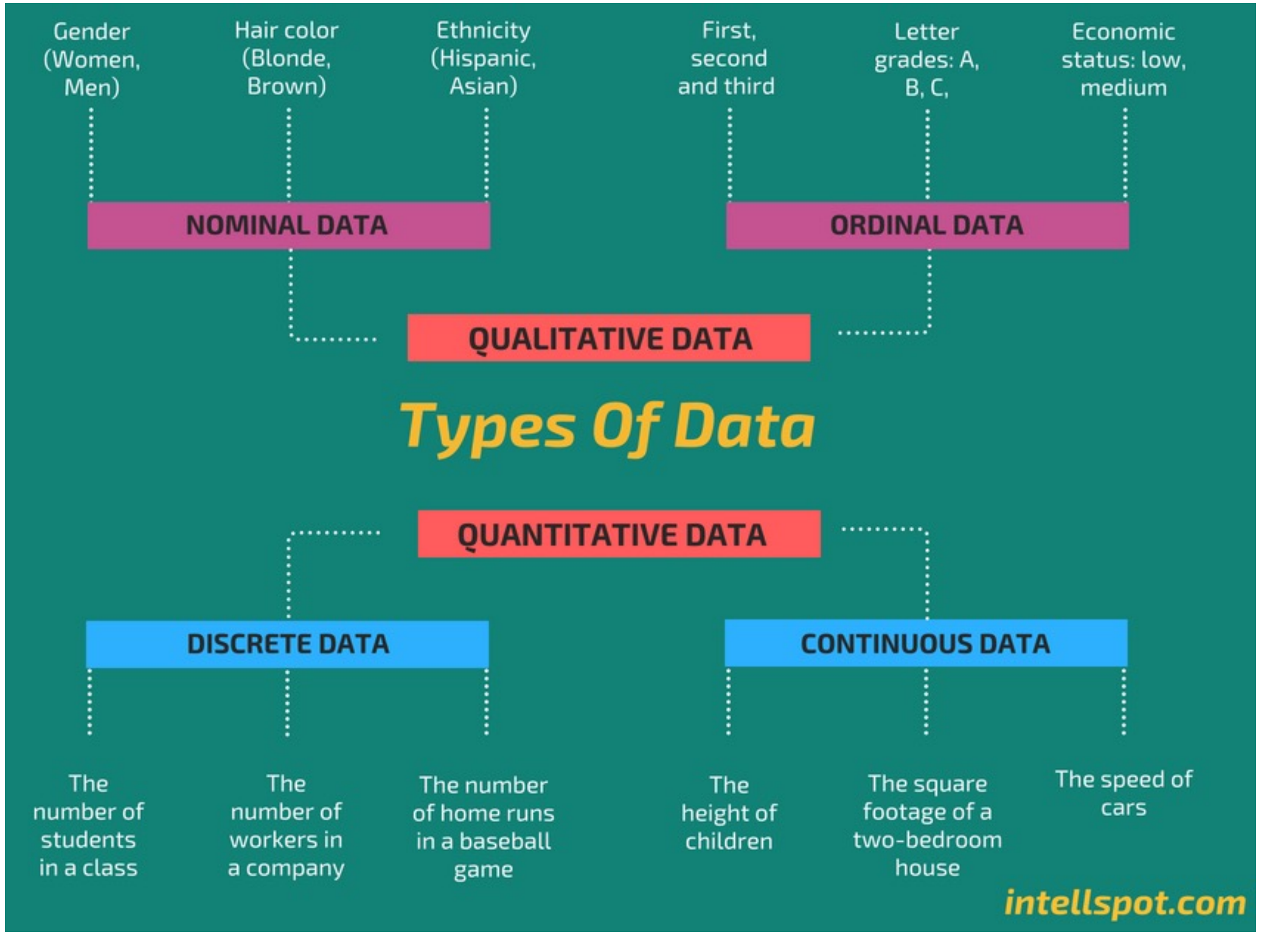
Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

So, you should hone your statistical skills and your value will increase in the job market!!

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

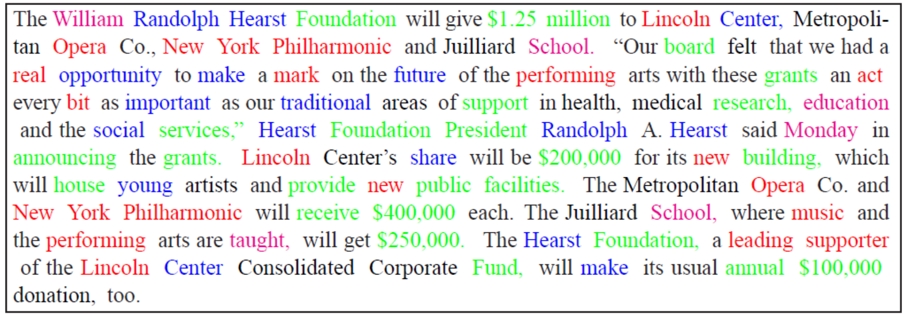
The height of children

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

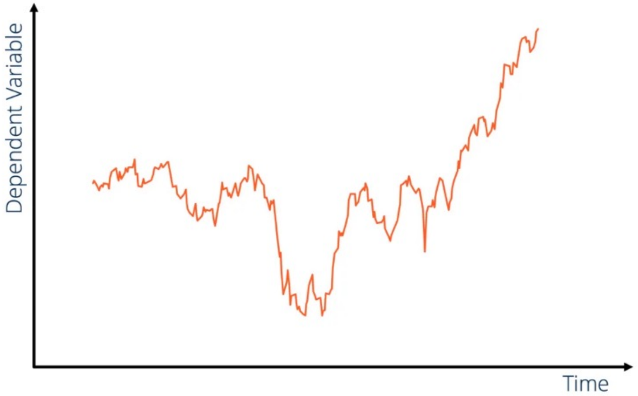
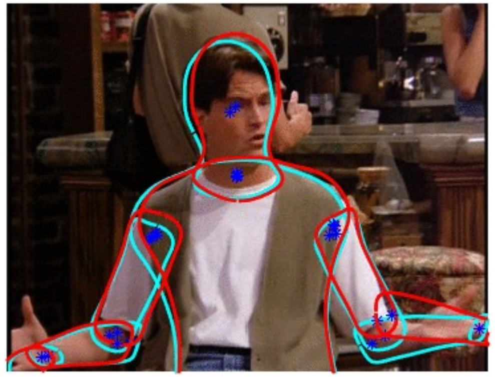
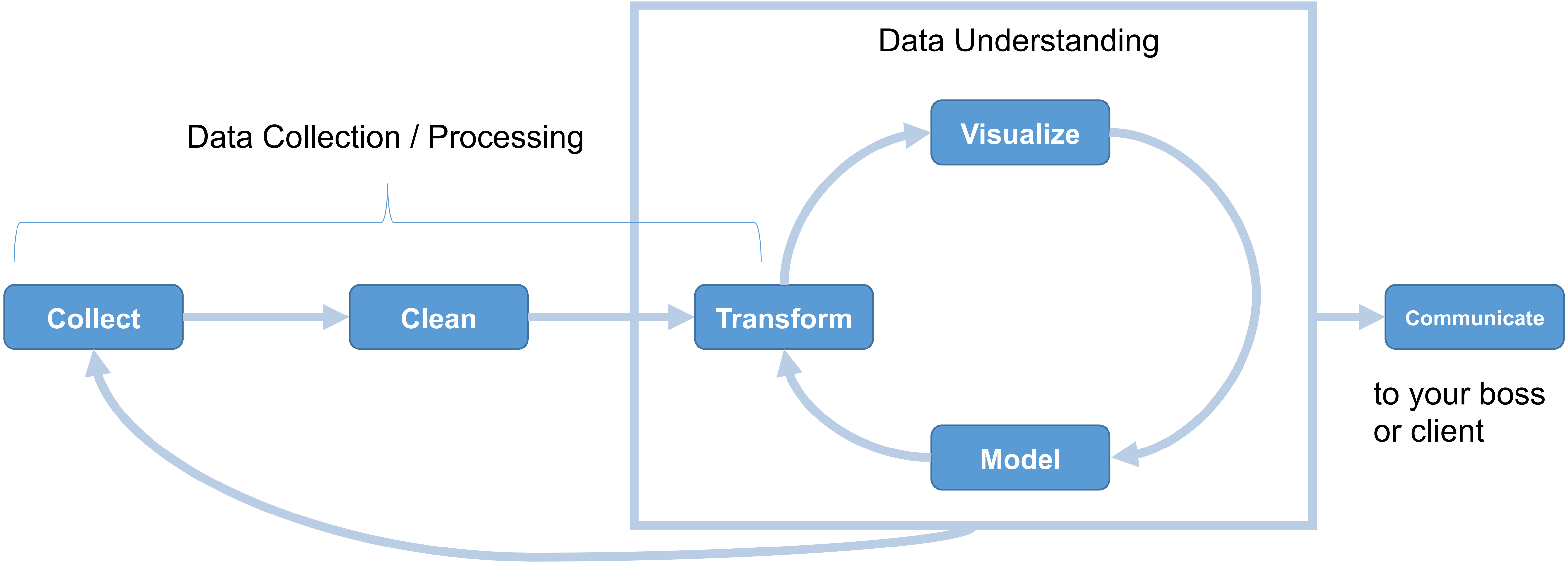


Image / Video



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

Data Science Workflow



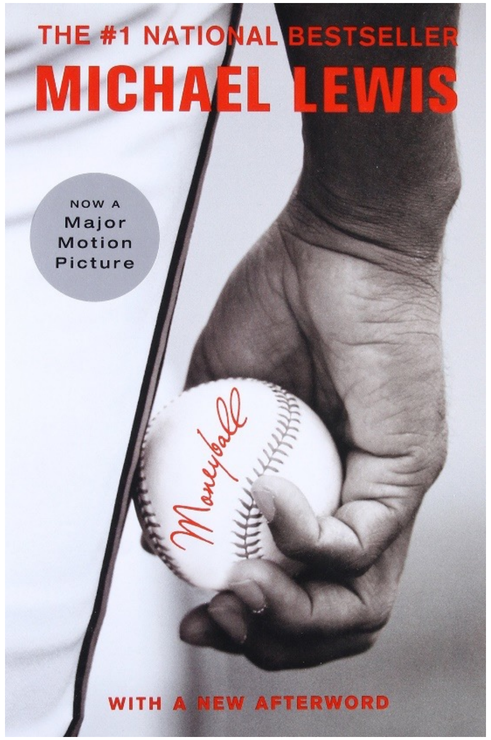
[Adapted from: Grolemund and Wickham, 2018]

Case Studies

Moneyball

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

- Story about the Oakland Athletics baseball team and its general manager **Billy Beane** for 2002 Major League Baseball (MLB) draft
- Traditional team building relies on *scouts* – but they are often biased and 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)



On-base %: how frequently a batter reaches base
Slugging %: the total number of bases a player records per at-bat

Moneyball: Impact

- In 2002 the Oakland Athletics ($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)

Election Forecasting: Disclaimer

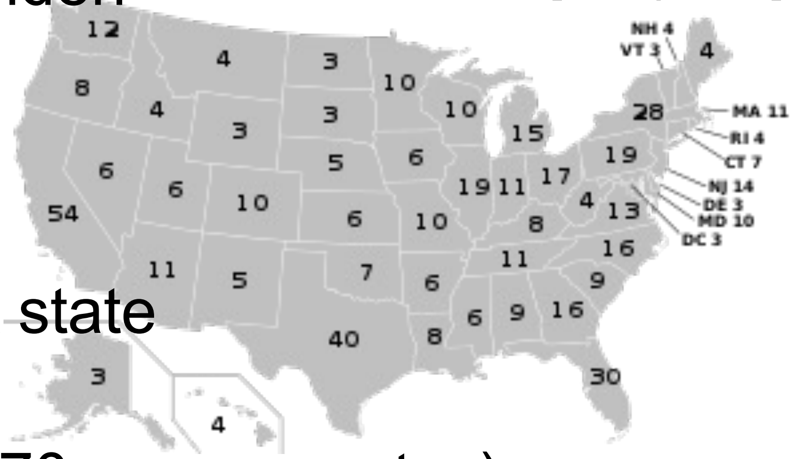
This is a class about data science it is **not** a class about politics. We will discuss election forecasting **only** in the context of data science and we will **ignore politics**.

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)



[Wikipedia]

** 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

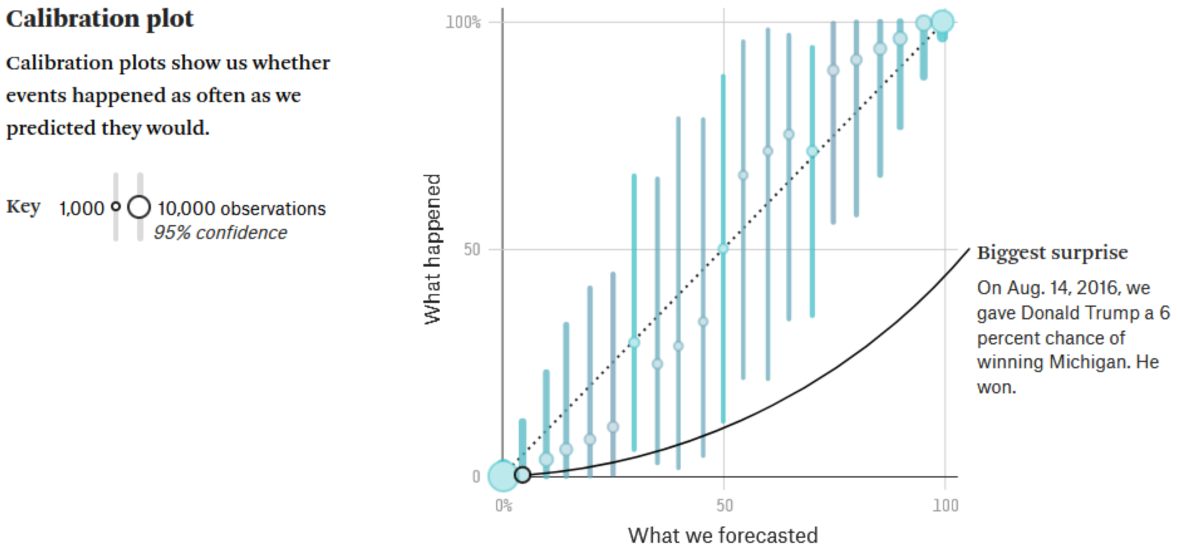
$$\text{prediction} = \sum_i \text{weight}_i \times \text{poll}_i + \text{random noise}$$

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.

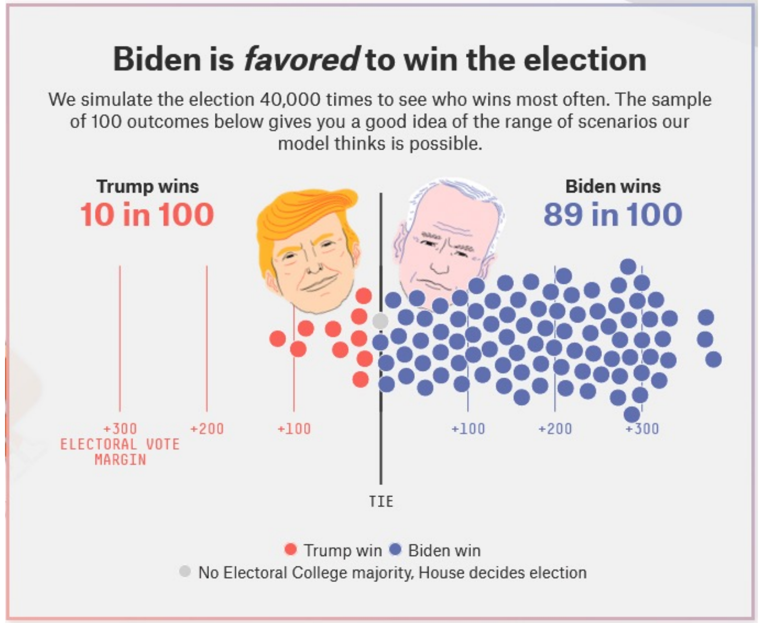
Biggest surprise

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

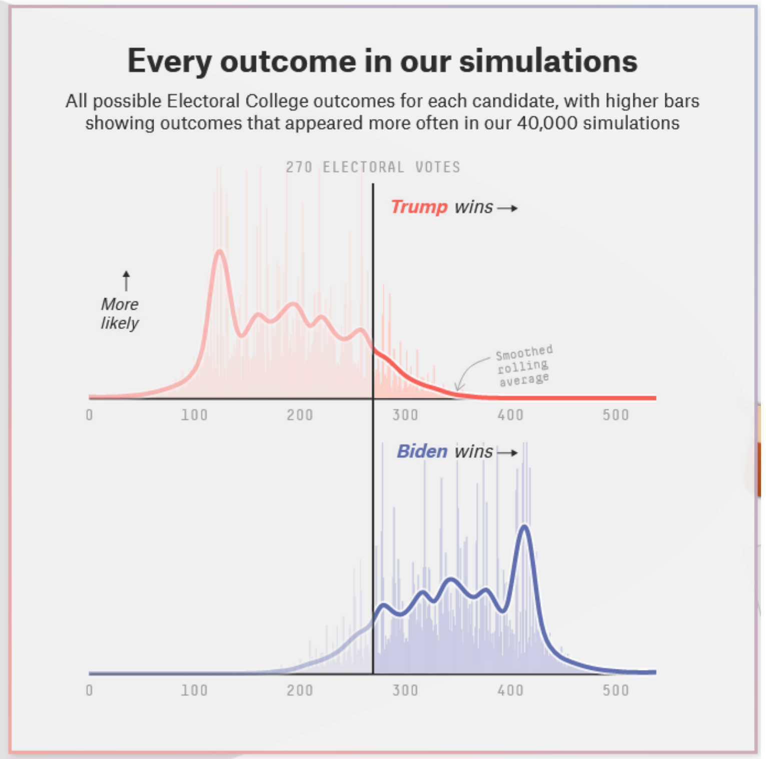
FiveThirtyEight

Election Forecasting: Visualizations

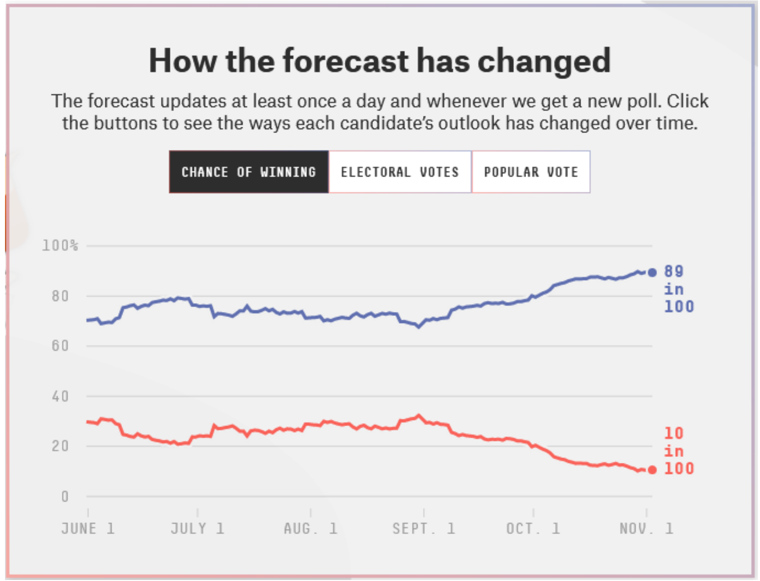
Generative (Bayesian) model allows simulation of random realizations…*


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

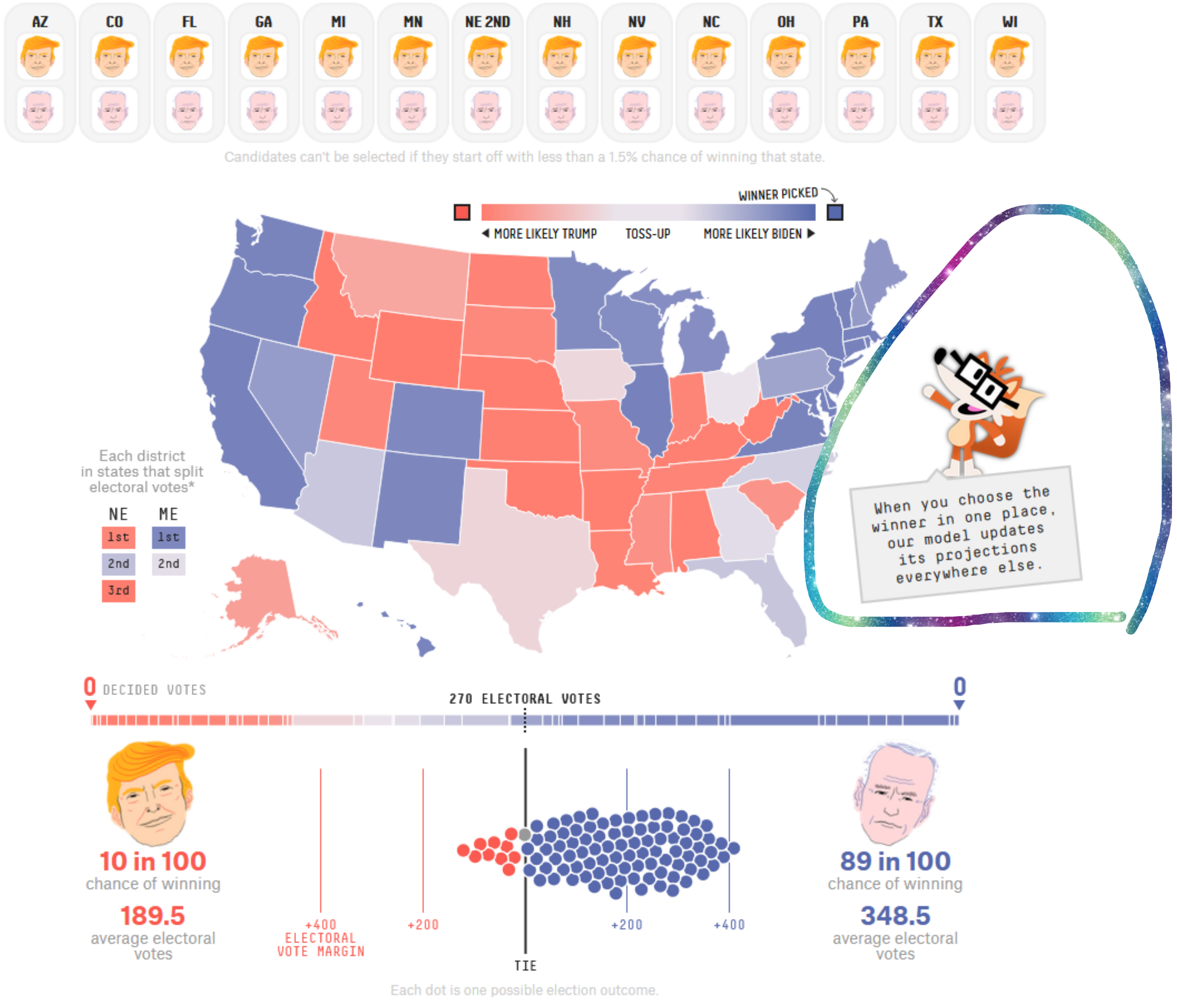
The forecast updates at least once a day and whenever we get a new poll. Click the buttons to see the ways each candidate's outlook has changed over time.

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.

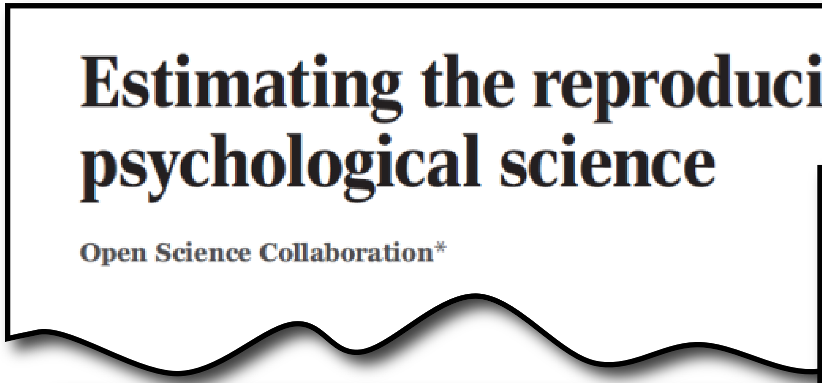
Each district in states that split electoral votes*

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

Each dot is one possible election outcome.

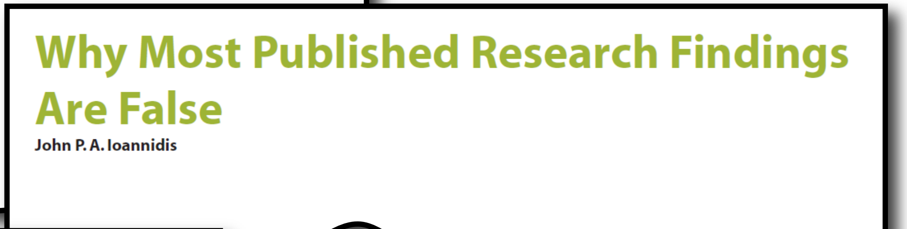
…this is a feature of model interpretability.

Bad Data Science & Statistics



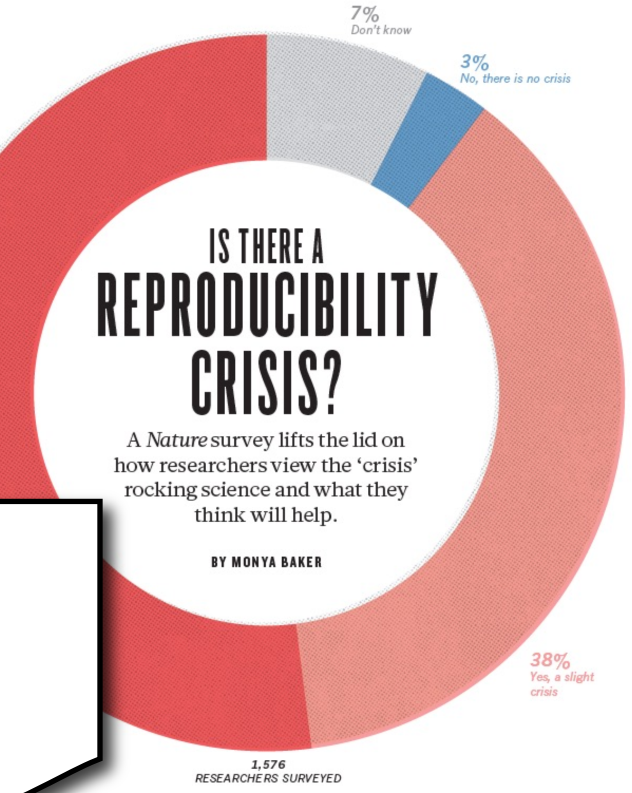
Estimating the reproducibility of psychological science

Open Science Collaboration*



Why Most Published Research Findings Are False

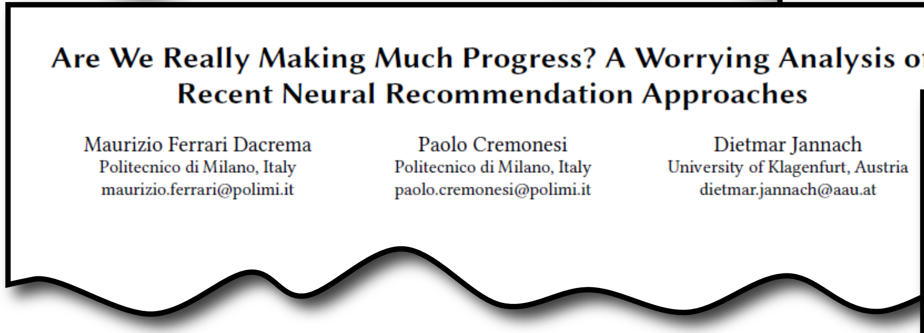
John P. A. Ioannidis



IS THERE A REPRODUCIBILITY CRISIS?

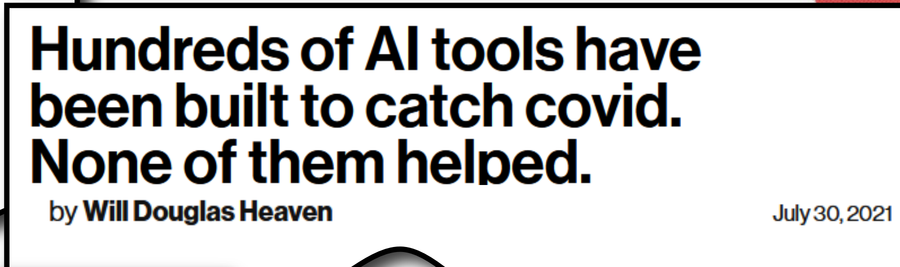
A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER



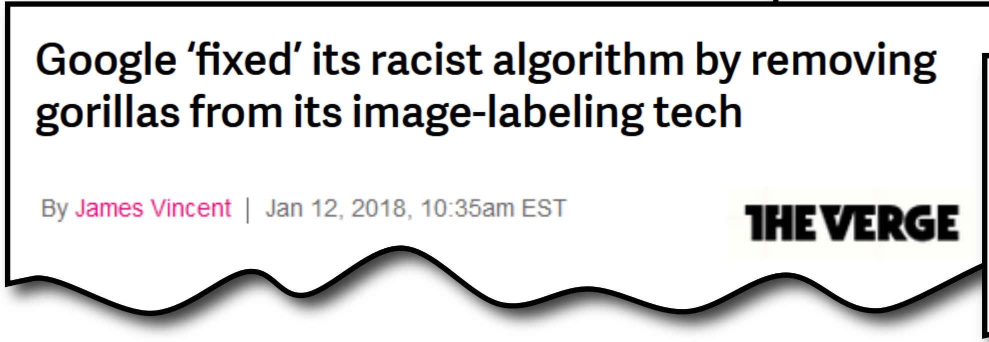
Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema	Paolo Cremonesi	Dietmar Jannach
Politecnico di Milano, Italy	Politecnico di Milano, Italy	University of Klagenfurt, Austria
maurizio.ferrari@polimi.it	paolo.cremonesi@polimi.it	dietmar.jannach@aau.at



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

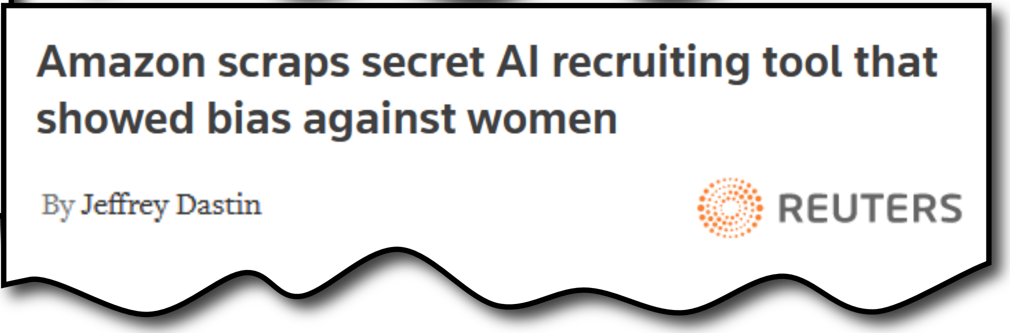
by **Will Douglas Heaven** July 30, 2021



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

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

julia

↑ **getting popular nowadays**

Python Packages Covered

NumPy

scikit learn

matplotlib

pandas

Other Useful Python Packages

TensorFlow

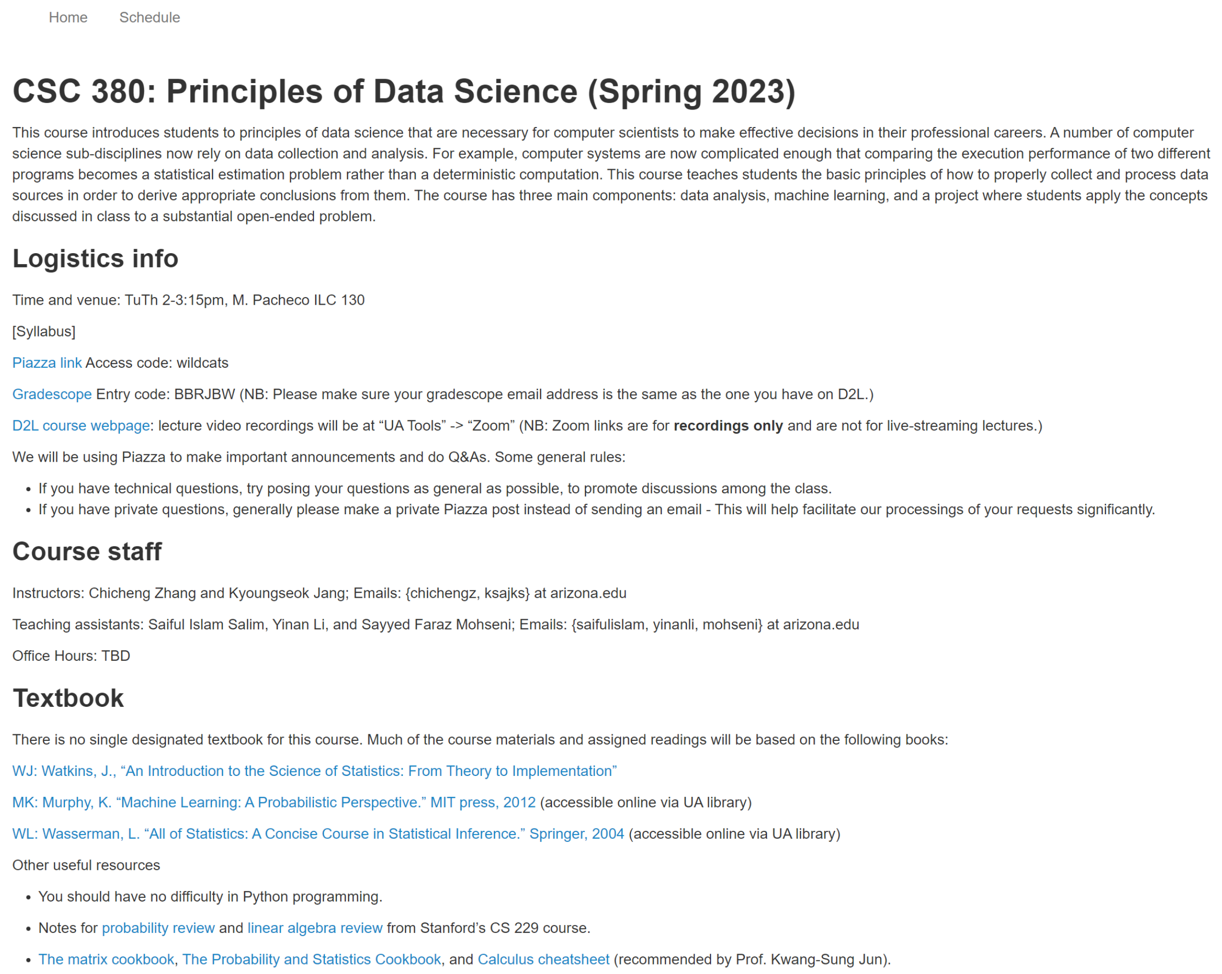
PyTorch

Altair

Course Overview

Course Overview: Resources

https://zcc1307.github.io/csc380-sp23/index.html



CSC 380: Principles of Data Science (Spring 2023)

This course introduces students to principles of data science that are necessary for computer scientists to make effective decisions in their professional careers. A number of computer science sub-disciplines now rely on data collection and analysis. For example, computer systems are now complicated enough that comparing the execution performance of two different programs becomes a statistical estimation problem rather than a deterministic computation. This course teaches students the basic principles of how to properly collect and process data sources in order to derive appropriate conclusions from them. The course has three main components: data analysis, machine learning, and a project where students apply the concepts discussed in class to a substantial open-ended problem.

Logistics info

Time and venue: TuTh 2-3:15pm, M. Pacheco ILC 130

[Syllabus]

Piazza link Access code: wildcats

Gradescope Entry code: BBRJBW (NB: Please make sure your gradescope email address is the same as the one you have on D2L.)

D2L course webpage: lecture video recordings will be at “UA Tools” -> “Zoom” (NB: Zoom links are for **recordings only** and are not for live-streaming lectures.)

We will be using Piazza to make important announcements and do Q&As. Some general rules:

- If you have technical questions, try posing your questions as general as possible, to promote discussions among the class.
- If you have private questions, generally please make a private Piazza post instead of sending an email - This will help facilitate our processings of your requests significantly.

Course staff

Instructors: Chicheng Zhang and Kyoungseok Jang; Emails: {chichengz, ksajks} at arizona.edu

Teaching assistants: Saiful Islam Salim, Yinan Li, and Sayyed Faraz Mohseni; Emails: {saifulislam, yinanli, mohseni} at arizona.edu

Office Hours: TBD

Textbook

There is no single designated textbook for this course. Much of the course materials and assigned readings will be based on the following books:

WJ: Watkins, J., “An Introduction to the Science of Statistics: From Theory to Implementation”

MK: Murphy, K. “Machine Learning: A Probabilistic Perspective.” MIT press, 2012 (accessible online via UA library)

WL: Wasserman, L. “All of Statistics: A Concise Course in Statistical Inference.” Springer, 2004 (accessible online via UA library)

Other useful resources

- You should have no difficulty in Python programming.
- Notes for probability review and linear algebra review from Stanford’s CS 229 course.
- The matrix cookbook, The Probability and Statistics Cookbook, and Calculus cheatsheet (recommended by Prof. Kwang-Sung Jun).

Specific resources

- gradescope for assignment submission
- Piazza for discussions and Q&A.
- Readings and electronic textbooks
- Lecture slides (posted after class)

Every lecture accompanied by reading

- We may have a few “assigned reading check” quizzes throughout the semester

Attendance is required

Recordings will be available **after the class**.

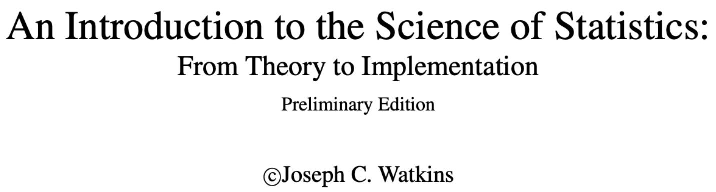
Textbooks

- No single designated textbook for this course.
- Much of the course materials and assigned readings will be based on:

Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation" (https://www.math.arizona.edu/~jwatkins/statbook.pdf)

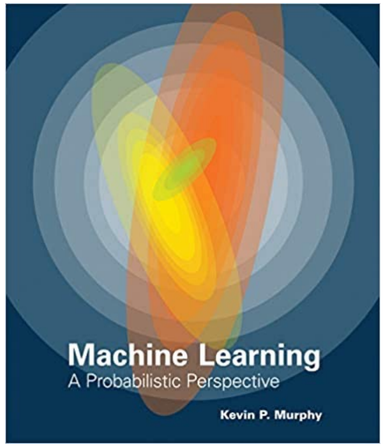
Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012 (UA Library)

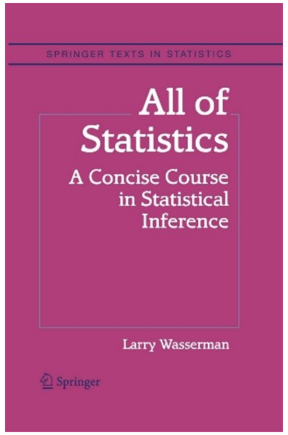
WL: Wasserman, L. “All of Statistics: A Concise Course in Statistical Inference.” Springer, 2004 (UA Library)



An Introduction to the Science of Statistics:
From Theory to Implementation
Preliminary Edition

©Joseph C. Watkins




SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer

Course TA

Your friendly course TAs…

Saiful Islam Salim
saifulislam@arizona.edu

Yinan Li
yinanli@arizona.edu

Sayyed Faraz Mohseni
mohseni@arizona.edu

Expected Skills

- This class will use a fair amount of **math**
  - Probability and Statistics
  - Some Linear Algebra
  - These are not required background for the course, but you will learn key concepts in the class.
- 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.*

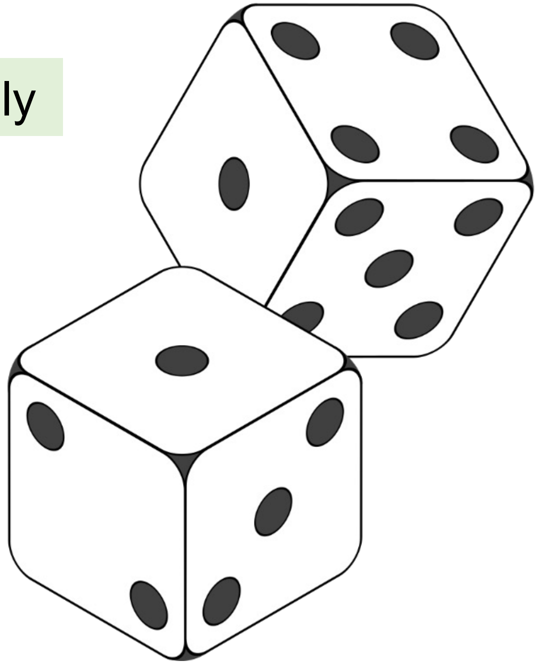
Probability and Statistics	Data Handling and Visualization	Machine Learning
Random events / variables, distributions / densities, moments, descriptive stats, estimation	Reading & cleaning, transformation & preprocessing, visualization	Predictive models, supervised learning, unsupervised learning, model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

↑ more on this in CSC 480/580

Probability and Statistics

Suppose we roll <u>two fair dices</u>…

fair die: each side is equally likely



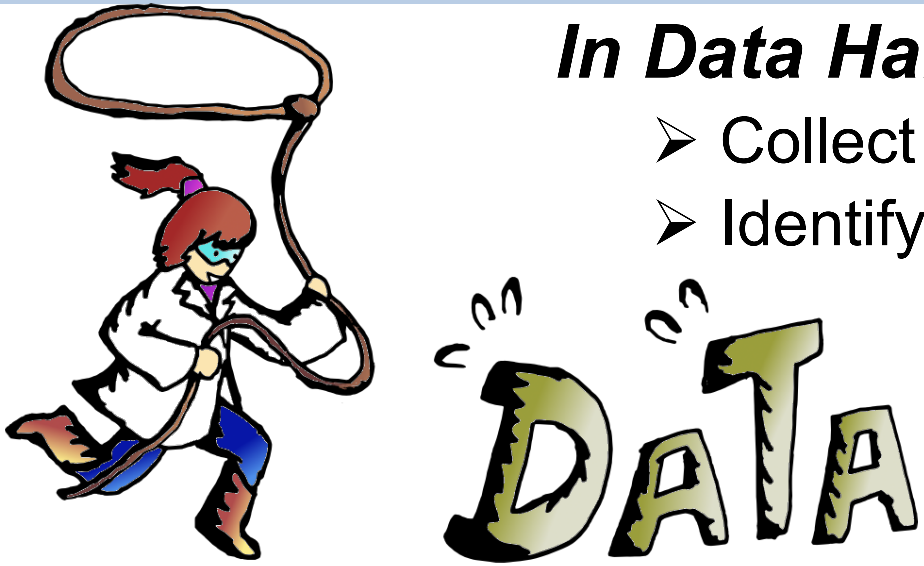
- ➢ What are the possible outcomes?
- ➢ What is the *probability* of rolling **even** numbers?

*… this is a **random trial** 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



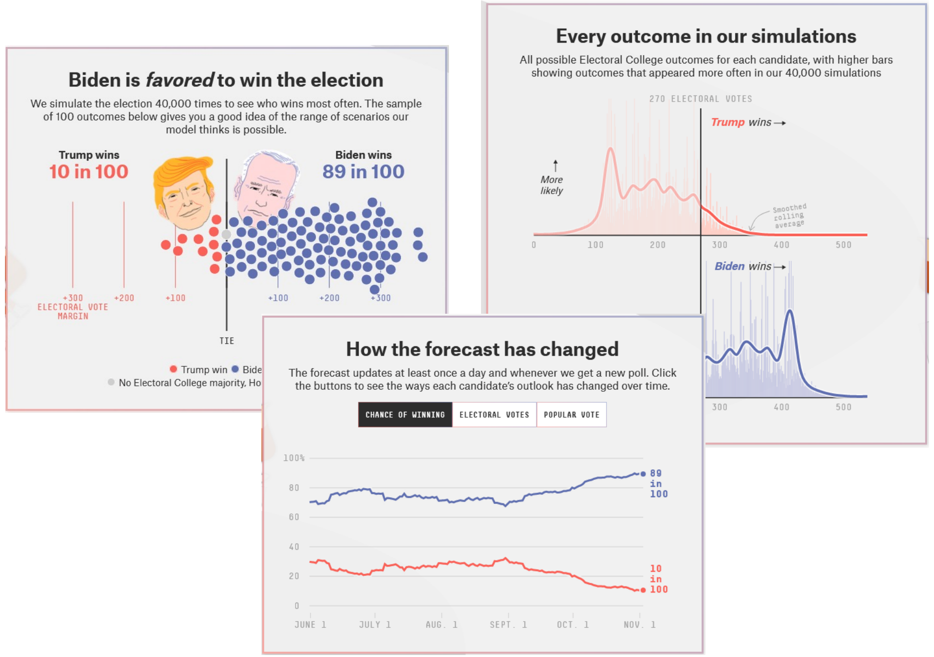
[Image Source: Code A Star]

In Data Handling we will learn to…

- ➢ Collect data
- ➢ Identify and avoid biased population samples
- ➢ Clean data and correct errors
- ➢ Transform and preprocess data (***wrangling***)

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

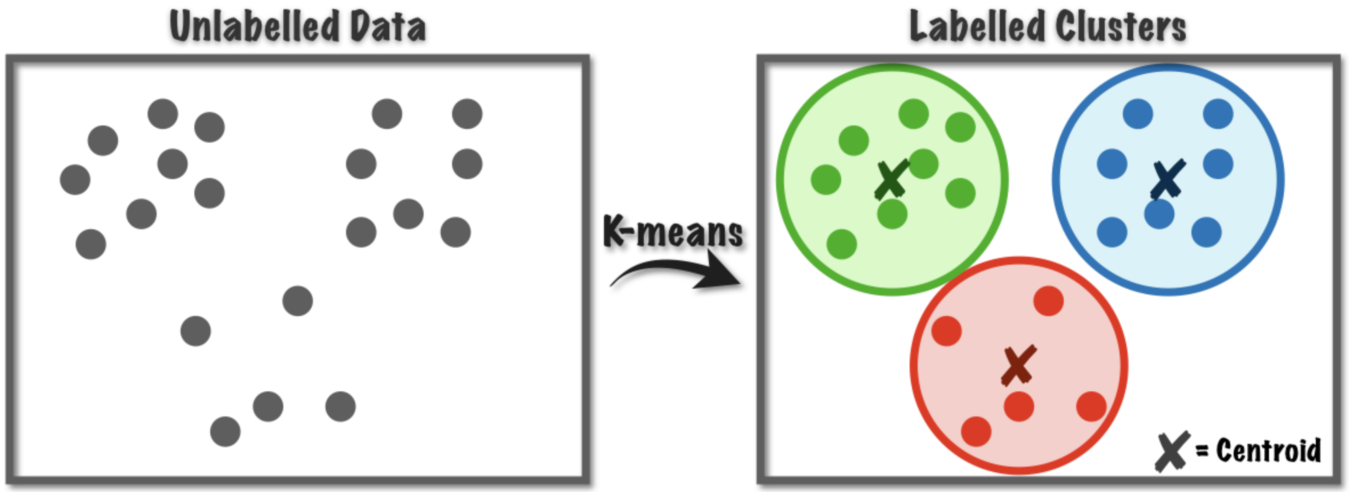
All possible Electoral College outcomes for each candidate, with higher bars showing outcomes that appeared more often in our 40,000 simulations

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 to 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]

Assignments / Exams / Grading

7 Homeworks + Midterm + Project + Final Exam

Homeworks

- Homeworks will be due in 8 days: e.g., out on Thursday, due on next Friday.
- You can do HWs individually or in pairs, but you must **contribute equally for each question** if working in pairs
- Grading will be available in 7 days excluding weekends/holidays.
- The HW with the lowest score will be dropped

Grading Breakdown

- Assignments: 36%
- Midterm: 20%
- Project: 14%
- Final Exam: 20%
- Participation: 10%

First assignment out next Thursday

Late Policy

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

No late submission policy

- Late submissions are not accepted, period.
- Strongly recommend that you plan to submit your work a day earlier.

Project

kaggle

- It is a previous Kaggle competition.
- A guided project. You will answer given questions, including some open questions.
- You will get a chance to try out various ML algorithms and get high accuracy.
- For top 10%, extra score (+2%).

Communication

- Announcements will be made via Piazza (please sign up)

- Homework submission: **gradescope** (see course website for the link)
  - Make sure your gradescope email address is the same as your D2L's

- **Piazza** (see course website for the link): we highly encourage that you ask and answer questions among yourselves.
  - We will chime in often.
  - You can also ask questions directly to us if it is personal.
  - Otherwise, please make the question as a public post so other students can benefit from it.

Office Hours

- Office hours will be held in person
- 1hr by the instructor, once a week.
- 1hr by each TA, once a week.
- The final office hour schedule will be announced at the end of this week.
- If you have a conflict with the schedule, let us know (Piazza)

Academic Integrity

Assignments are to be done independently,
unless explicitly marked as a collaborative homework.

If we or the TAs suspect you of having cheated

- You will be notified immediately
- We will have a conference where you can plead your case
- If we are not swayed then **<u>you will get an F grade</u>**, period.

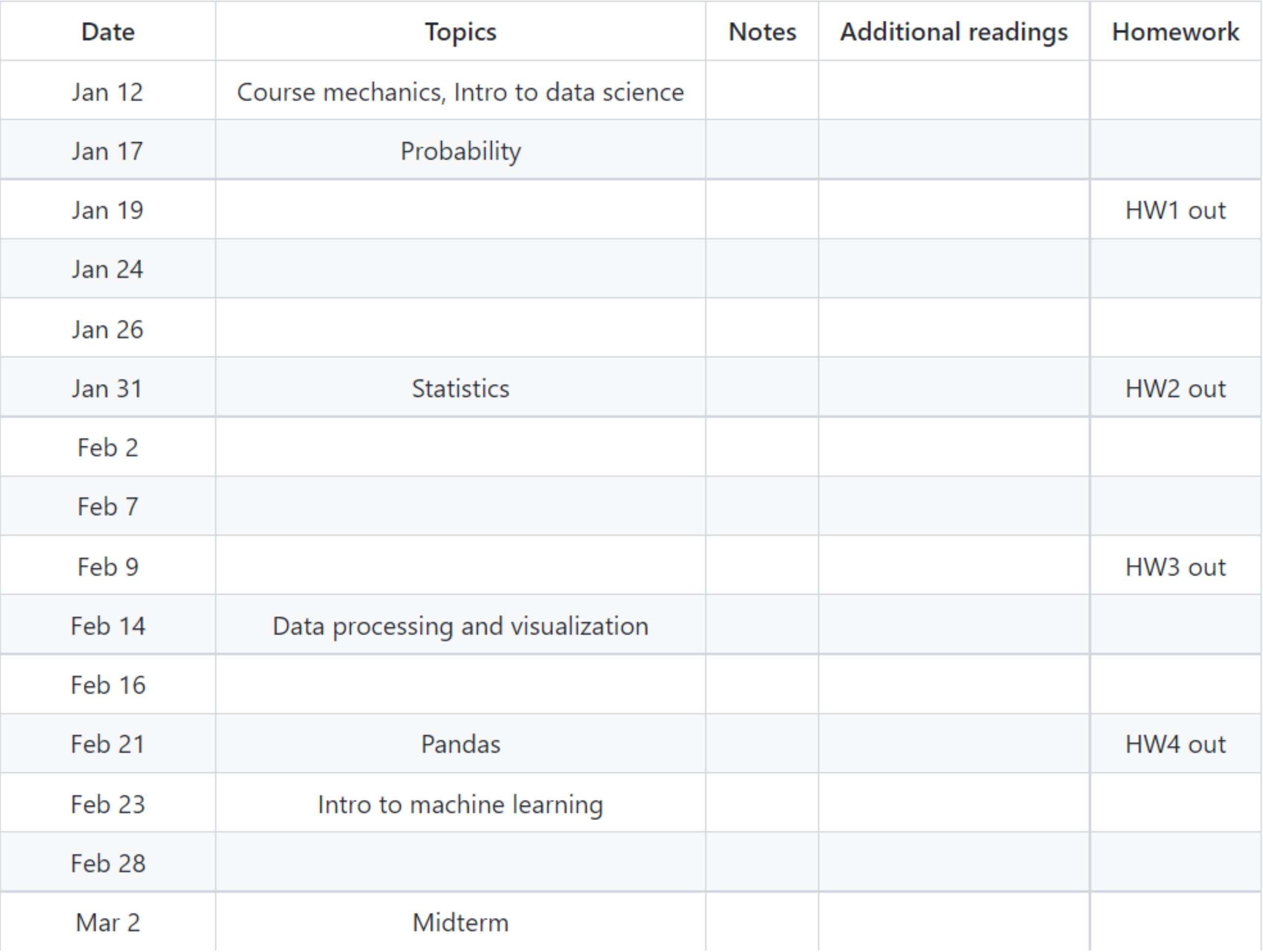
To avoid any unconscious cheating, you must write down **<u>who you have worked with</u>** and **<u>to what degree</u>** you got help, outside your group.

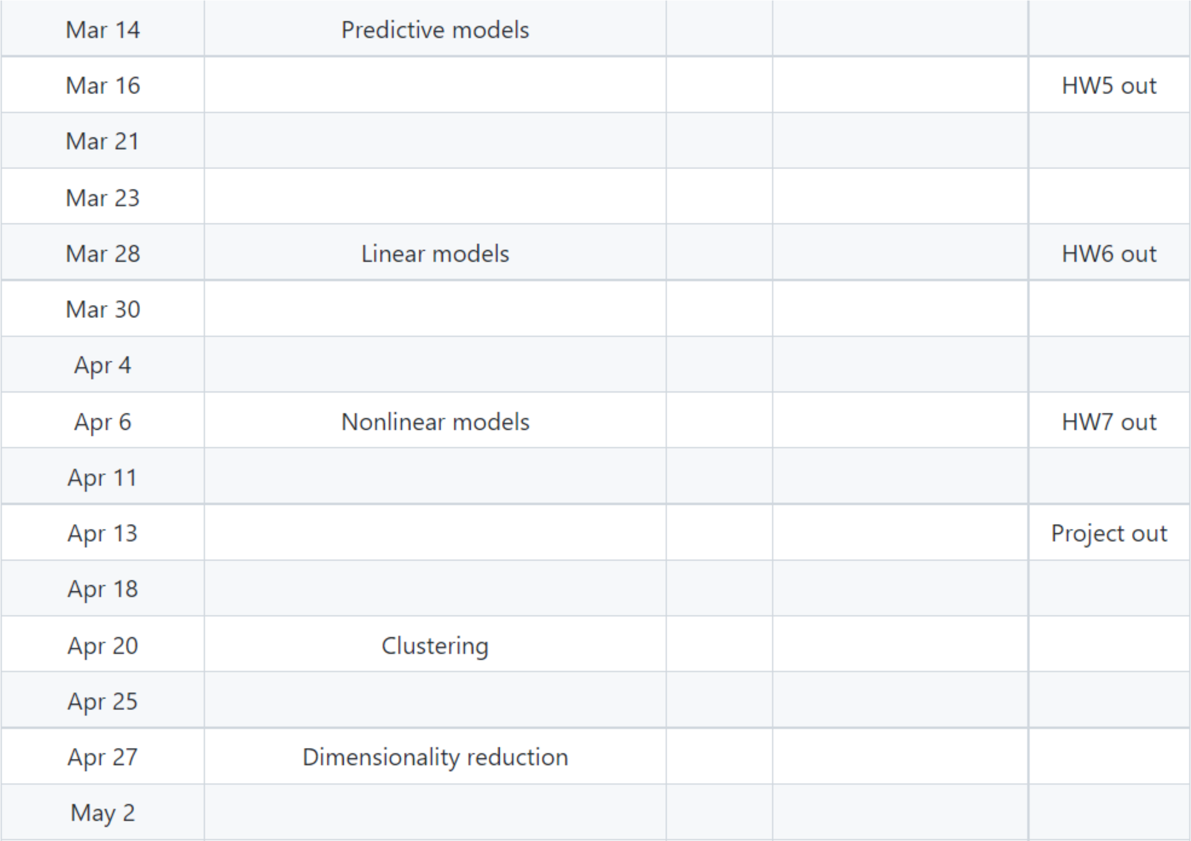
Bottom line: don’t cheat

Full Course Schedule (Tentative)

Tentatitve; We will constantly update the schedule page

March 7, 9: Spring Recess





Date	Topics	Notes	Additional readings	Homework
Jan 12	Course mechanics, Intro to data science			
Jan 17	Probability			
Jan 19				HW1 out
Jan 24				
Jan 26				
Jan 31	Statistics			HW2 out
Feb 2				
Feb 7				
Feb 9				HW3 out
Feb 14	Data processing and visualization			
Feb 16				
Feb 21	Pandas			HW4 out
Feb 23	Intro to machine learning			
Feb 28				
Mar 2	Midterm			
Mar 14	Predictive models			
Mar 16				HW5 out
Mar 21				
Mar 23				
Mar 28	Linear models			HW6 out
Mar 30				
Apr 4				
Apr 6	Nonlinear models			HW7 out
Apr 11				
Apr 13				Project out
Apr 18				
Apr 20	Clustering			
Apr 25				
Apr 27	Dimensionality reduction			
May 2				

Important Dates

- Jan 24: last date to self-withdraw without a ‘W’
- Mar 2: midterm
- Mar 28: last date to self-withdraw
  - $\geq$ 40% of your total grades will be available by then.
- Apr 13: final project out
- May 5: final project due
- May 8: final exam

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

We want to foster a comfortable and inclusive classroom experience

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

- Improper use of pronouns
- Microaggressions
- Miscellaneous statements / interactions

You can message us on Piazza or discuss in person

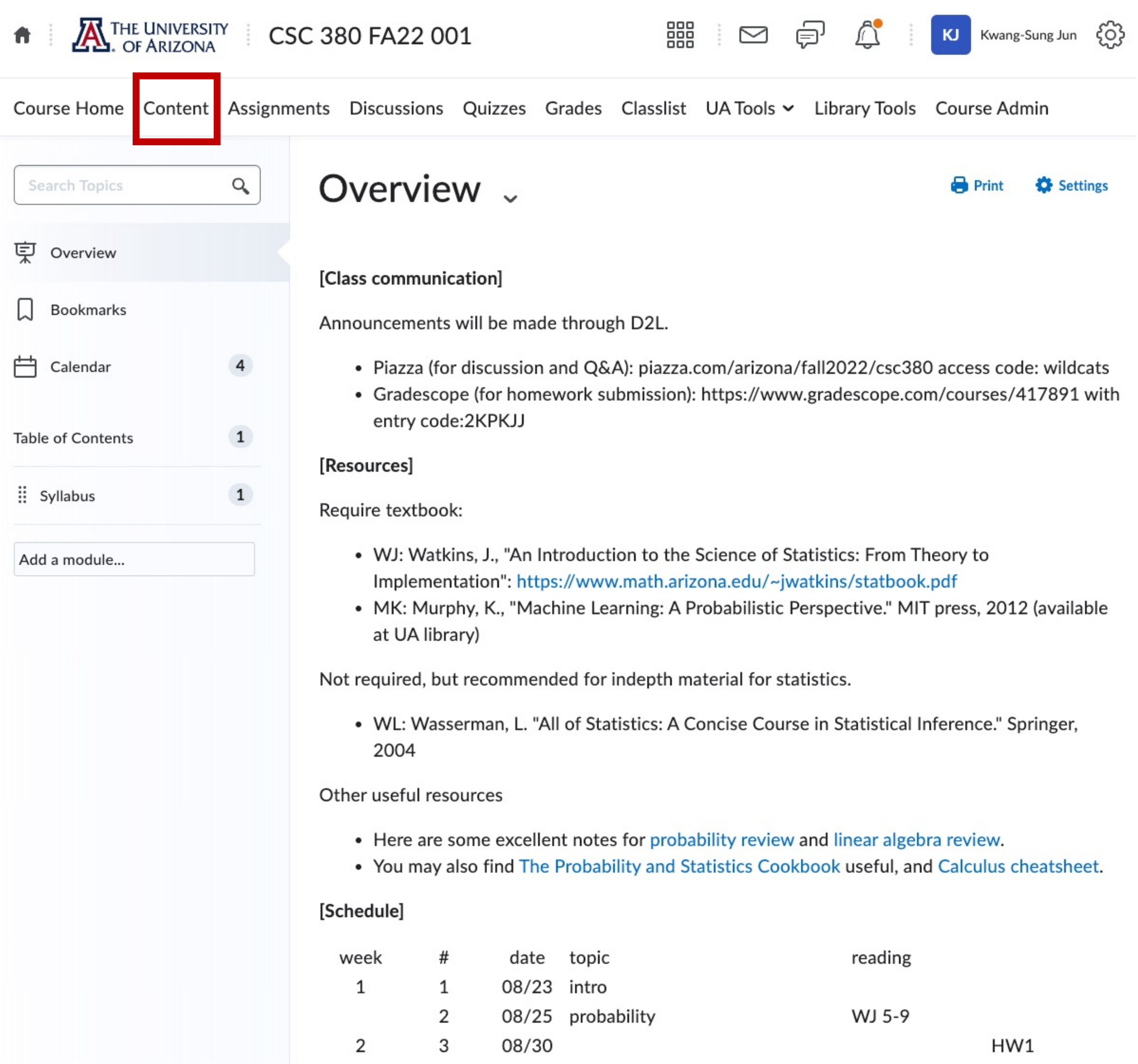
Reading Assignments

- Robinson and Nolis, "What is Data Science?” (link from course schedule page)

- ‘Probability and statistics cookbook’ is a good cheat sheet. Download it from http://statistics.zone/

Thank you

Course Overview: Resources



Overview

[Class communication]

Announcements will be made through D2L.

- Piazza (for discussion and Q&A): piazza.com/arizona/fall2022/csc380 access code: wildcats
- Gradescope (for homework submission): https://www.gradescope.com/courses/417891 with entry code:2KPKJJ

[Resources]

Require textbook:

- WJ: Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation": https://www.math.arizona.edu/~jwatkins/statbook.pdf
- MK: Murphy, K., "Machine Learning: A Probabilistic Perspective." MIT press, 2012 (available at UA library)

Not required, but recommended for indepth material for statistics.

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

Other useful resources

- Here are some excellent notes for probability review and linear algebra review.
- You may also find The Probability and Statistics Cookbook useful, and Calculus cheatsheet.

[Schedule]

week	#	date	topic	reading	
1	1	08/23	intro		
	2	08/25	probability	WJ 5-9	
2	3	08/30			HW1

Resources accessible on D2L

Specific resources

- gradescope for assignment submission
- Piazza for discussions and Q&A.
- Readings and electronic textbooks
- Lecture slides (posted after class)

Every lecture accompanied by reading

- We may have “assigned reading check” quizzes throughout the semester

Attendance is required

Recordings will be available **<u>after the class</u>**.

Homework

Let’s see your preferences.

- Collaborative? (say 3 people per group)

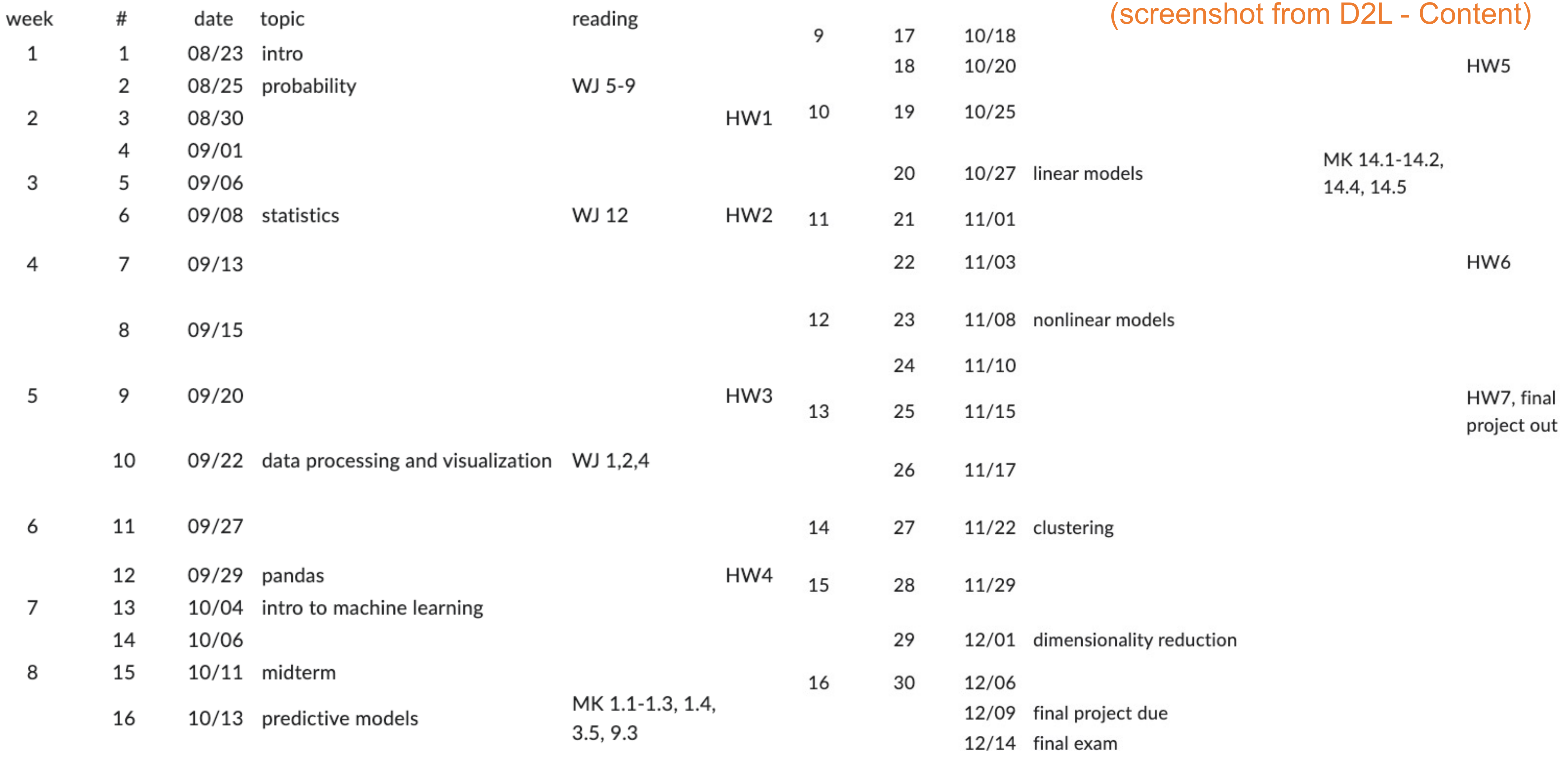
- Or individual?

(Even if it is collaborative, you will have to do you own homework, and you must understand your own answer. It just means that you can answers within your group)

D2L Walkthrough

- https://d2l.arizona.edu/d2l/home/1132174

Full Course Schedule (Tentative)



(screenshot from D2L - Content)

week	#	date	topic	reading	
1	1	08/23	intro		
	2	08/25	probability	WJ 5-9	
2	3	08/30			HW1
	4	09/01			
3	5	09/06			
	6	09/08	statistics	WJ 12	HW2
4	7	09/13			
	8	09/15			
5	9	09/20			HW3
	10	09/22	data processing and visualization	WJ 1,2,4	
6	11	09/27			
	12	09/29	pandas		HW4
7	13	10/04	intro to machine learning		
	14	10/06			
8	15	10/11	midterm		
	16	10/13	predictive models	MK 1.1-1.3, 1.4, 3.5, 9.3	
9	17	10/18			
	18	10/20			HW5
10	19	10/25			
	20	10/27	linear models	MK 14.1-14.2, 14.4, 14.5	
11	21	11/01			
	22	11/03			HW6
12	23	11/08	nonlinear models		
	24	11/10			
13	25	11/15			HW7, final project out
	26	11/17			
14	27	11/22	clustering		
15	28	11/29			
	29	12/01	dimensionality reduction		
16	30	12/06			
		12/09	final project due		
		12/14	final exam		