COSE471 Data Science 2021 Sp

Introduction and Course Overview

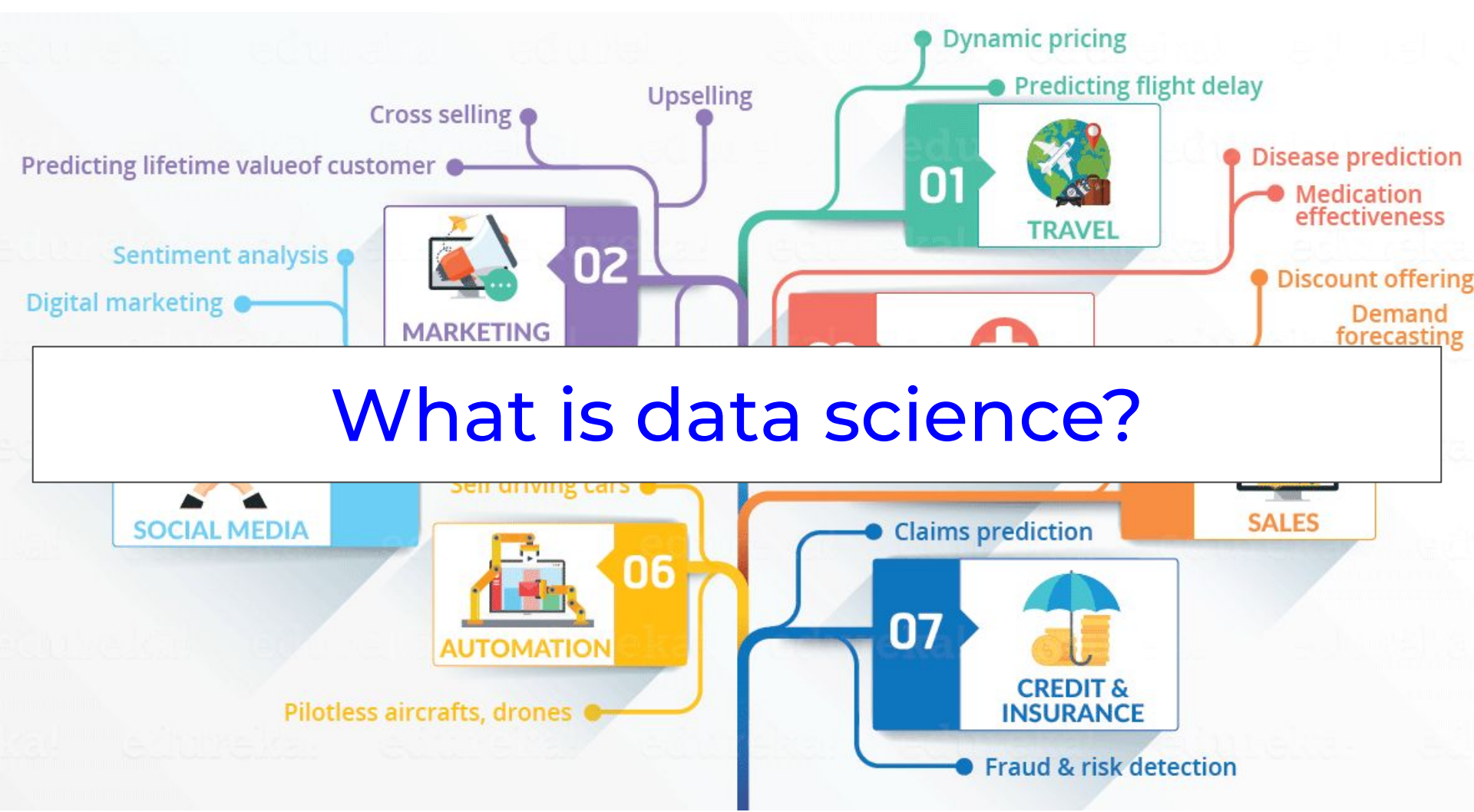
Department of Computer Science and Engineering

Korea University

* This material is adapted from Berkeley CS 100 (ds100.org) and may be copyrighted by them.

Outline

- What is data science?
- What will you learn in this class?
- Course overview
  - Lots of important details
  - Exams, homeworks, grading, formats, office hours, projects, TAs, etc.


Dynamic pricing
Predicting flight delay
Upselling
Cross selling
Predicting lifetime valueof customer
01
TRAVEL
Disease prediction
Medication effectiveness
Sentiment analysis
Digital marketing
02
MARKETING
Discount offering
Demand forecasting
What is data science?
SOCIAL MEDIA
Self driving cars
SALES
Claims prediction
06
AUTOMATION
07
CREDIT & INSURANCE
Pilotless aircrafts, drones
Fraud & risk detection

Data is changing the world

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

Data science is a fundamentally interdisciplinary field

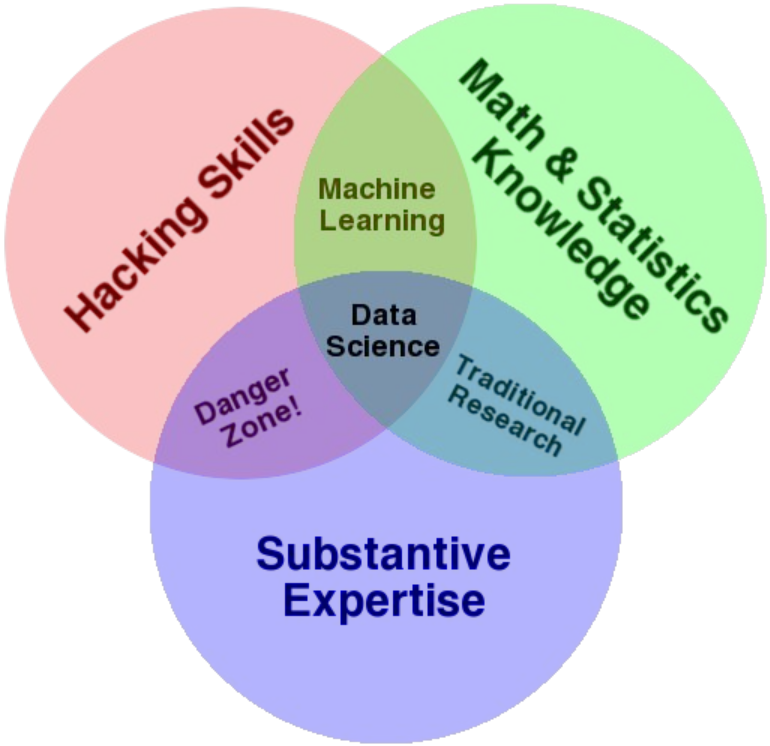
Joey Gonzalez

(UC Berkeley)

Data Science is the application of data centric, computational, and inferential thinking to:

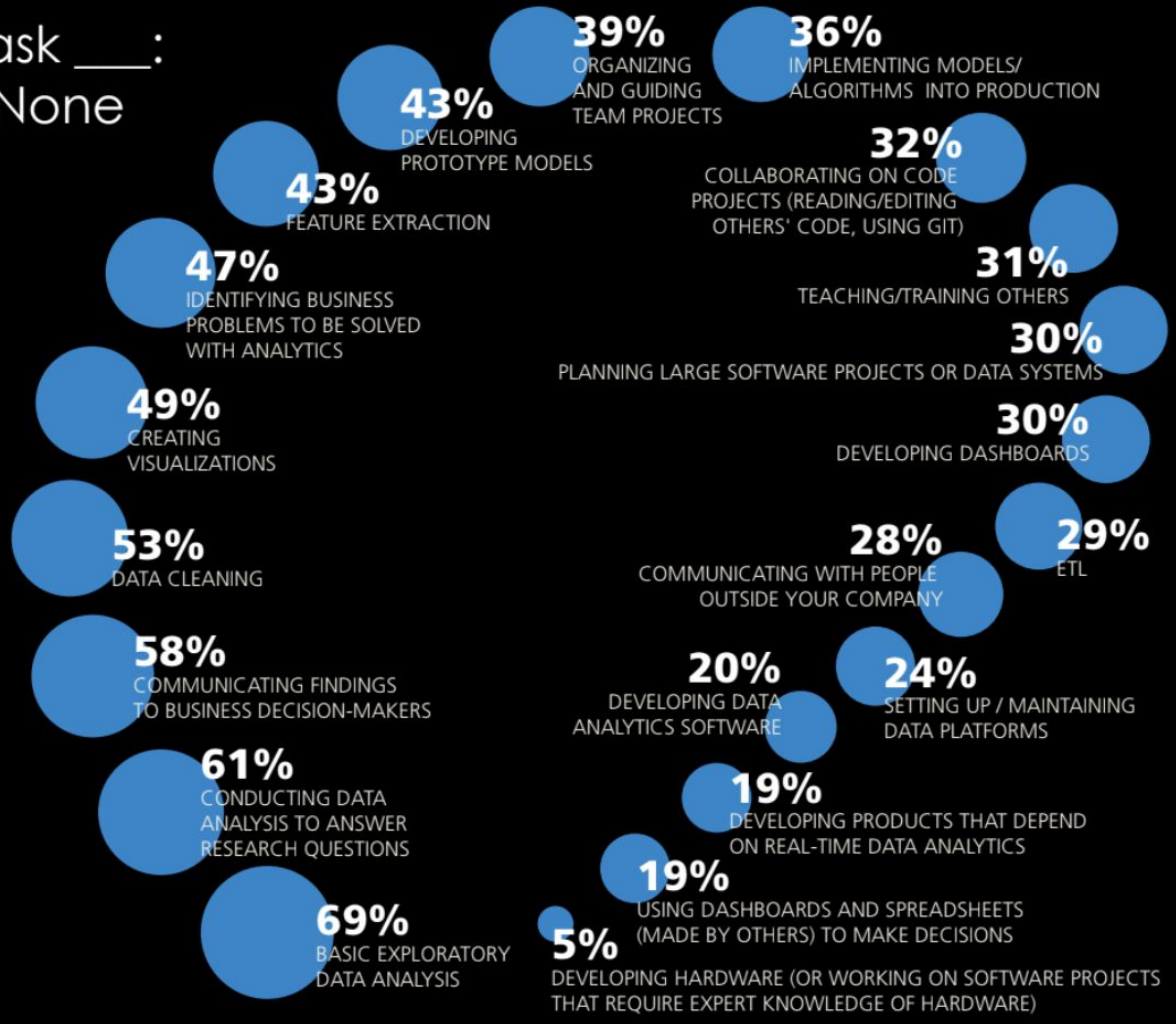
- Understand the world (science).
- Solve problems (engineering).

Data Science Venn Diagram



by Drew Conway in 2010 (link)

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



The tasks that data scientists say they work on regularly.

Self-reported. Based on the results of the 2016 Data Science Salary Survey.

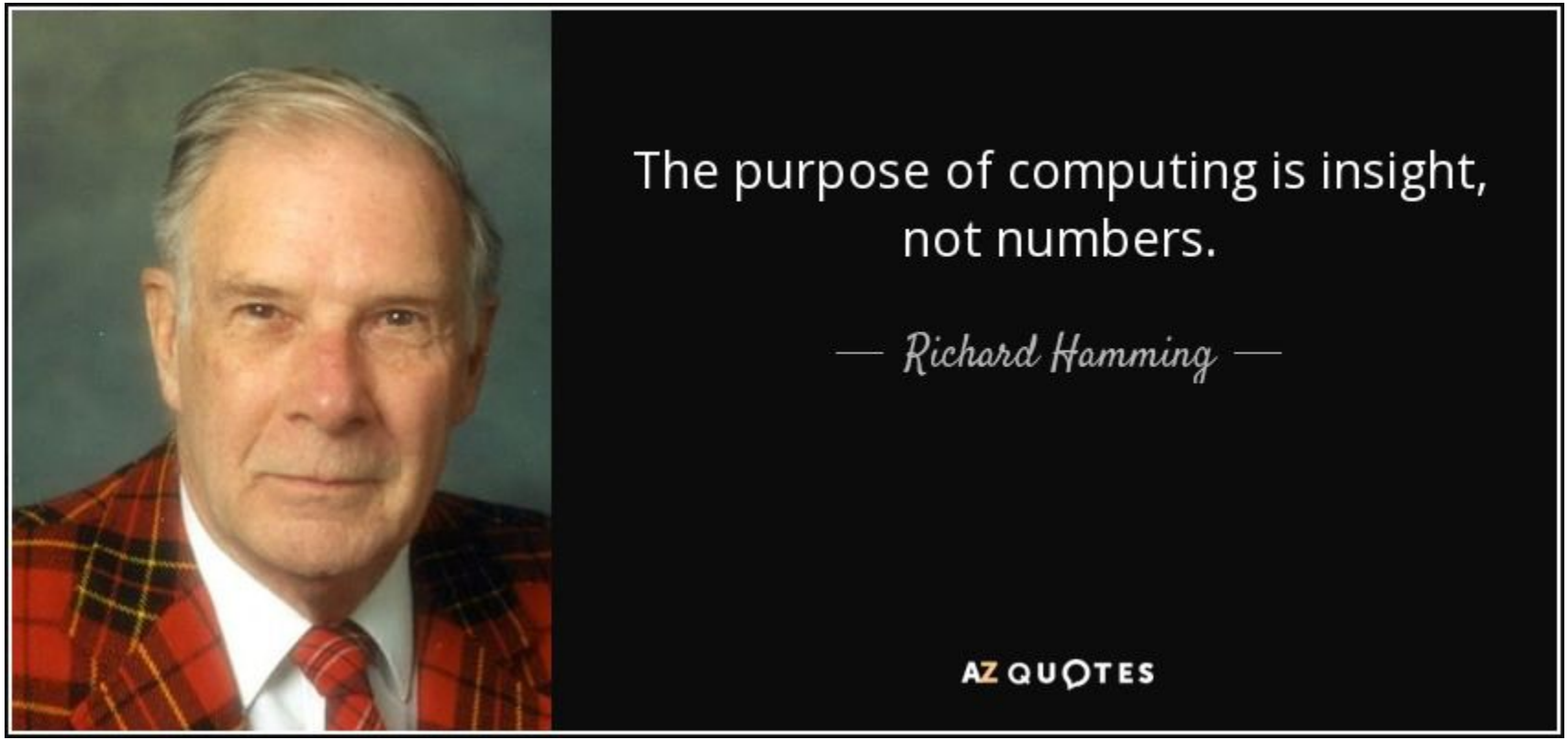
Insight

Good data analysis is not:

- Simple application of a statistics recipe.
- Simple application of statistical software.

There are many **tools** out there for data science, but they are merely tools.

- **They don’t do any of the important thinking!**


The purpose of computing is insight, not numbers.
— Richard Hamming —
AZ QUOTES

Example questions in data science

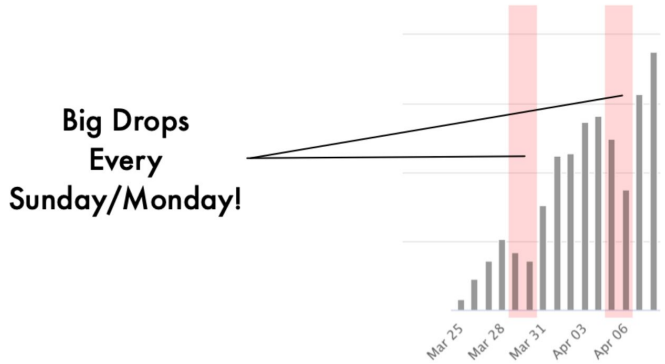
Some (broad) questions we might try to answer with data science:

- Is the world getting better or worse?
- Is the use of the COMPAS algorithm for prison sentencing fair?
- Who will win the 2020 presidential election in each state?
- Where should we put docking ports for our bikes?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?
- Will there be a third wave of COVID-19?

Data science drives policy and public understanding

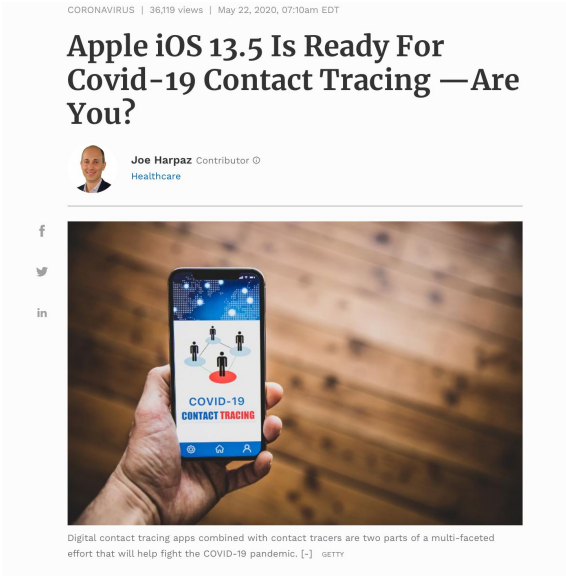
There are real-world implications of the work we do as data scientists.

Let's take a look at the daily numbers reported by the United Kingdom:



Daily Deaths due to COVID in the UK from https://www.worldometers.info/coronavirus/country/uk/

The problem is that this weekly cycle is fake. It's an artifact of how the data is collected and reported.



CORONAVIRUS | 36,119 views | May 22, 2020, 07:10am EDT

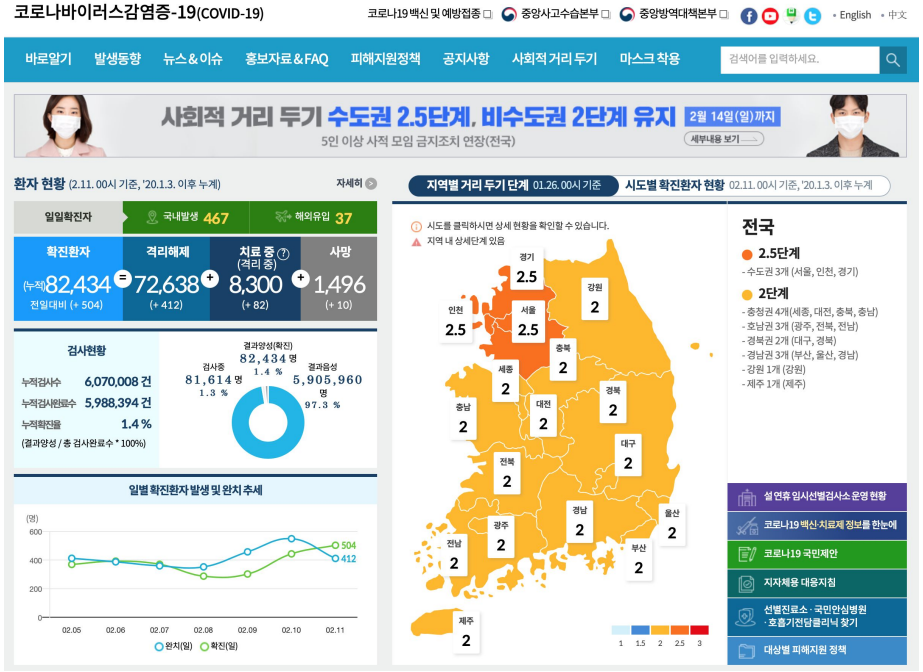
Apple iOS 13.5 Is Ready For Covid-19 Contact Tracing —Are You?

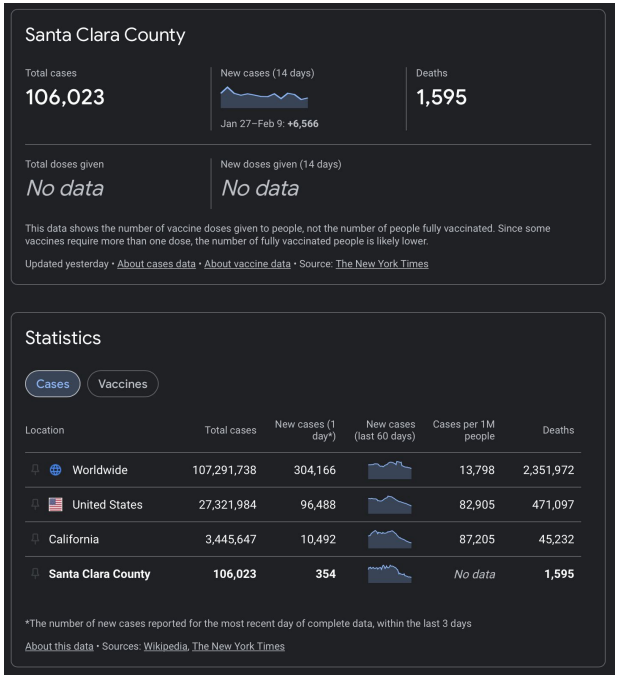
Joe Harpaz Contributor
Healthcare

Digital contact tracing apps combined with contact tracers are two parts of a multi-faceted effort that will help fight the COVID-19 pandemic. [-] GETTY

Data science drives policy and public understanding

There are real-world implications of the work we do as data scientists.





What will you learn in this class?

Course goals

Prepare

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

Enable

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

Empower

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

Syllabus (tentative)

Part I:

- Probability and Data Design
- Pandas Syntax
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Regular Expressions, Working with Text
- Visualization

◉ midterm (4/20: Wed)

Syllabus (tentative)

Part II:

- Modeling and Estimation
- Regression
- Feature Engineering
- Bias and Variance
- Cross-validation
- Regularization
- Gradient Descent
- Logistic Regression

Syllabus (tentative)

Part III:

- Decision Trees
- Principal Component Analysis
- Clustering
- Guest Lecture (Special topics in Data Science)

◉ final (6/16: Wed)

Course Logistics

Instructor and TAs

Jinkyu Kim
(instructor)

TBD
(TAs)

Instructor and TA will hold office hours and help create assignments and exams.

Online Live Lecture!

We will have

- **Online** Live Lectures
- **Online/Offline** Office Hours
- **Offline** Midterm/Final Exams

Lecture format

There are **two lectures per week**. Lectures will be **entirely a live online lecture**.

- Time and place
  - 9:00am - 10:15am, Mon/Wed
  - Room 202 Woojung Bldg. (+ Online)
- Teaching assistants
  - TBD

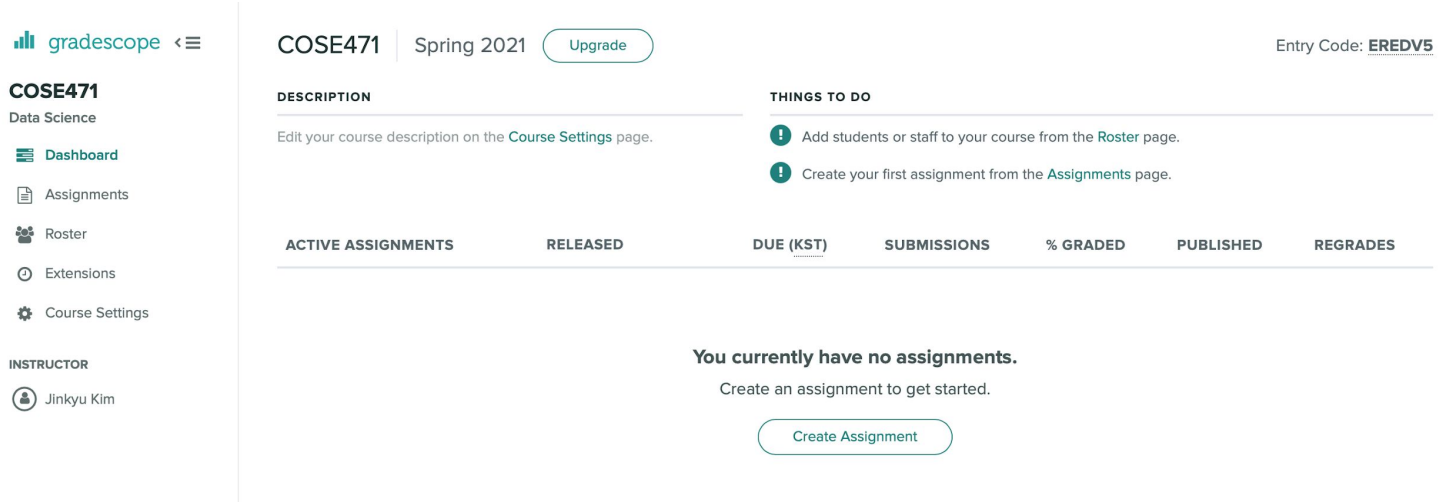
Blackboard

- Stay tuned! All announcements will be posted here.
- Where all lectures, assignments, and projects are posted.

Gradescope

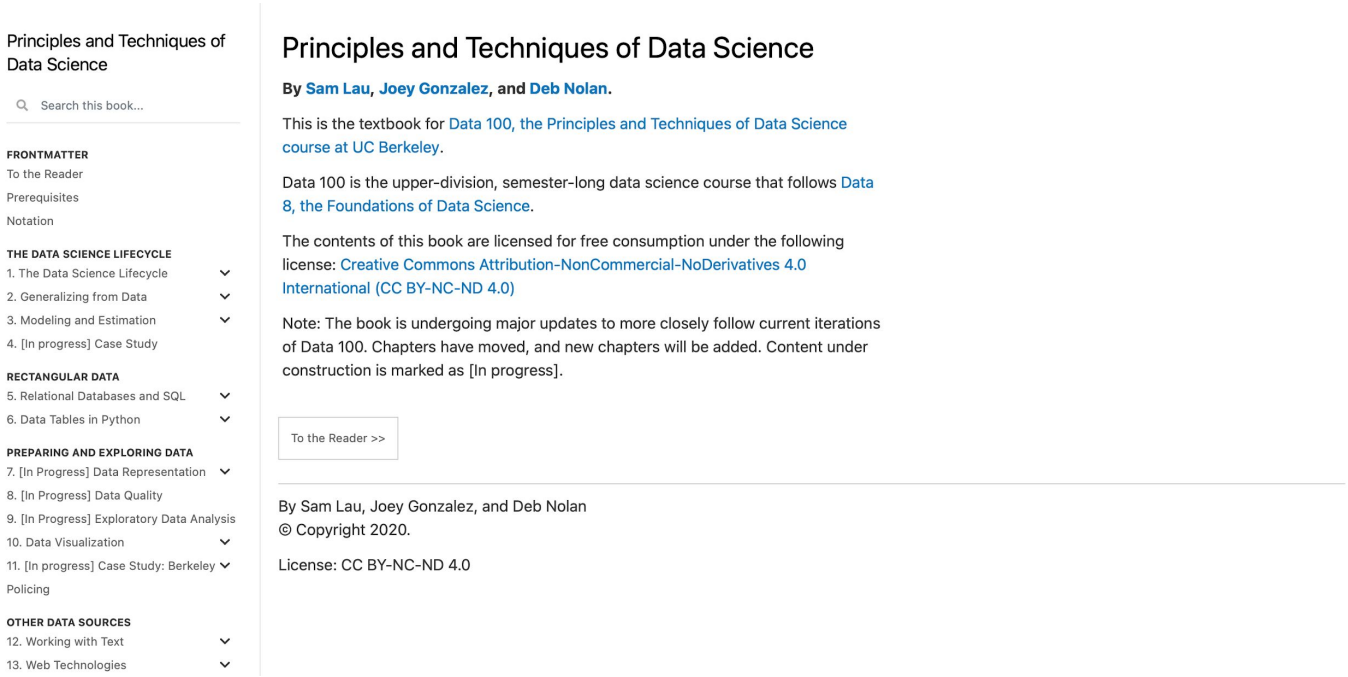
Gradescope (gradescope.com, by invitation only)

- Where all assignments are submitted, and where all of your grades in this course will live.
- Where your midterm and final exams are graded, and where all of your grades will live.
- If you are not invited by March 7, let us know!



Textbook

- Supplemental reading available at textbook.ds100.org (UC Berkeley CS100)



Principles and Techniques of Data Science

Search this book...

Principles and Techniques of Data Science

By Sam Lau, Joey Gonzalez, and Deb Nolan.

This is the textbook for Data 100, the Principles and Techniques of Data Science course at UC Berkeley.

Data 100 is the upper-division, semester-long data science course that follows Data 8, the Foundations of Data Science.

The contents of this book are licensed for free consumption under the following license: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

Note: The book is undergoing major updates to more closely follow current iterations of Data 100. Chapters have moved, and new chapters will be added. Content under construction is marked as [In progress].

To the Reader >>

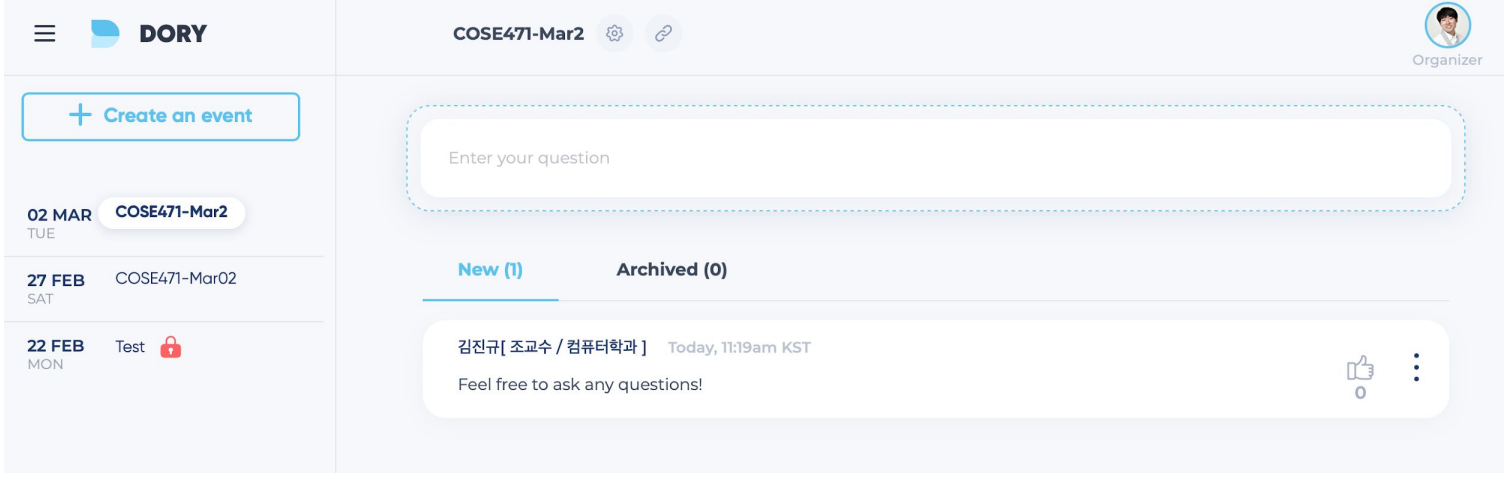
By Sam Lau, Joey Gonzalez, and Deb Nolan
© Copyright 2020.

License: CC BY-NC-ND 4.0

Dory

- Feel free to ask any questions on lectures; or to upvote
- We will always use the last 10 minutes to go over these questions.
- Today's Dory: https://www.dory.app/c/korea.ac.kr/a99bbb88_cose471-mar3

  (you will need korea.ac.kr credential)



Homeworks and projects

Homeworks are two-week-long assignments that are designed to help students develop an in-depth understanding of both the theoretical and practical aspects of ideas presented in lecture.

- Some homeworks will be on-paper written assignments
- The rest will be Jupyter notebooks.
- Homeworks and projects will be autograded; but some will be graded manually.

Projects are four-week-long assignments that integrate these ideas with real-world datasets.

Office hours

Zoom | Office hours (10-minute slots)

- Sign-up required
- Office hours will be announced soon
- These are led by TAs

Offline | Office hours (10-minute slots)

- Sign-up required
- Tue & Thu 1-2 pm Woojung Hall Rm. 202
- These are led by instructor

A typical two-week in the course

Monday	Tuesday	Wednesday	Thursday	Friday
Live lecture	Office Hours (Zoom/Offline)	Live lecture	Office Hours (Zoom/Offline)	
Homework released				
Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Live lecture	Office Hours (Zoom/Offline)	Live lecture	Office Hours (Zoom/Offline)	Homework Due

Exams

There will be in ***offline***

- Midterm:
  - April 21 (Wed)
  - 9:00 am - 10:15 am
  - closed book + 1 cheat sheet (both sides)
- Final
  - June 16 (Wed)
  - 9:00 - 10:15 am
  - closed book + 2 cheat sheet (both sides)

Note that time/date can be rescheduled by campus.

Alternate exams will only be given to students with a documented conflict. Please let instructor know if you have such a conflict.

Assignment submission

- Gradescope is the only place you will ever need to look for grades in this class.
  - Details of the assignment submission process will be announced soon.
  - Don’t wait to submit until the last minute!

- **Deadlines are firm at 11:59PM.**
  - You can submit projects/homeworks up to 5 days late without any penalty.
  - If you use more than 5 days, there will be 10% off per day.
    - Rounded up to the next day: 2 minutes late = 1 day late

Assignment submission and DSP accommodations

- Extensions are provided only to students with DSP accommodations, or in the case of exceptional circumstances (with instructor's approval).
    - If you have DSP accommodations, you will be received an email from us.

Grading

- Midterm (20%): closed book + 1 cheat sheet (both sides)
  - 9:00am on 4/21 (Wed)
- Final (30%): closed book + 2 cheat sheets (both sides)
  - 9:00am on 6/16 (Wed)
- Programming assignments (30%): done individually
  - assignments 1~7 (5% each, with 1 drop)
- Projects (20%): done individually
  - projects 1-2 (10% each)
- Class attendance (5+%)

Collaboration and academic dishonesty

Assignments

Data science is a collaborative activity! It is okay to discuss problems with friends.

- List their names at the top of your assignments. We provide a place to do this.
- You must write your solutions individually! Do not copy any other student's work.
- If we suspect that you have submitted plagiarized work, we will call you in for a meeting. If we then determine that plagiarism has occurred, we reserve the right to give you a negative full score (-100%) or lower on the assignments in question, along with reporting your offense to the University.

Exams

- Cheating on exams is a serious offense. We have methods of detecting cheating on exams – so **don't do it**!
- Students caught cheating on any exam will fail this course.

We are here to help!

We are here to help you, including

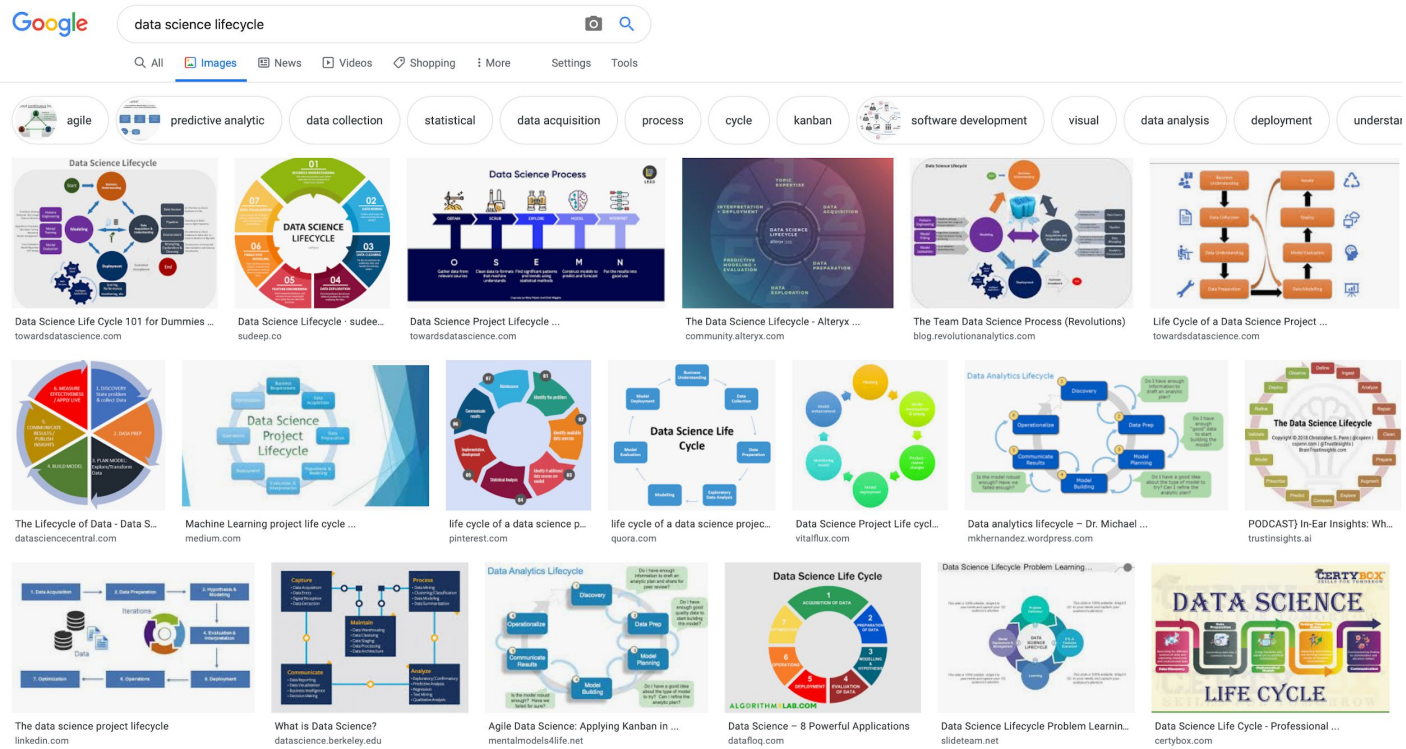
- Office hours
- Piazza (?)

We really want you to succeed in this class.

- These are particularly challenging times with everything going on in the world right now.
- *Feel free to reach out to any member of staff with any questions or concerns you have.*

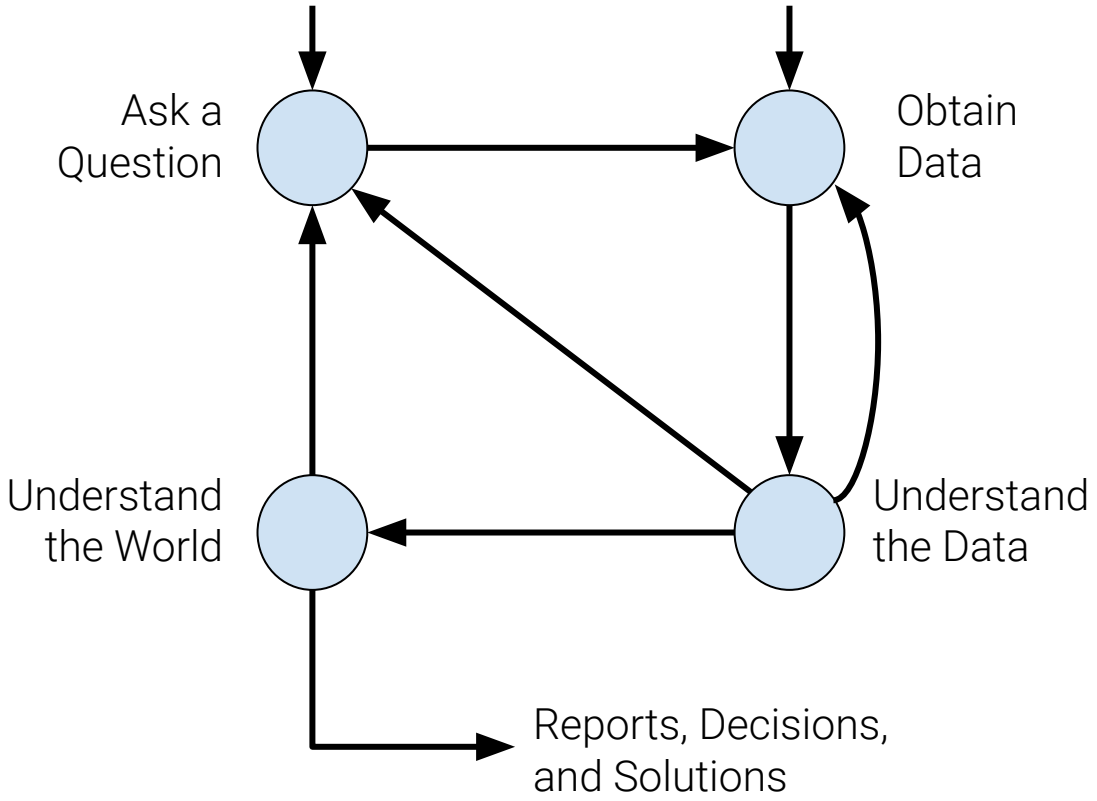
Welcome to COSE471!

Data Science Lifecycle



The “data science lifecycle” you will see out in the wild may be slightly different than the one we teach you, but the core ideas are all the same.

Data science lifecycle

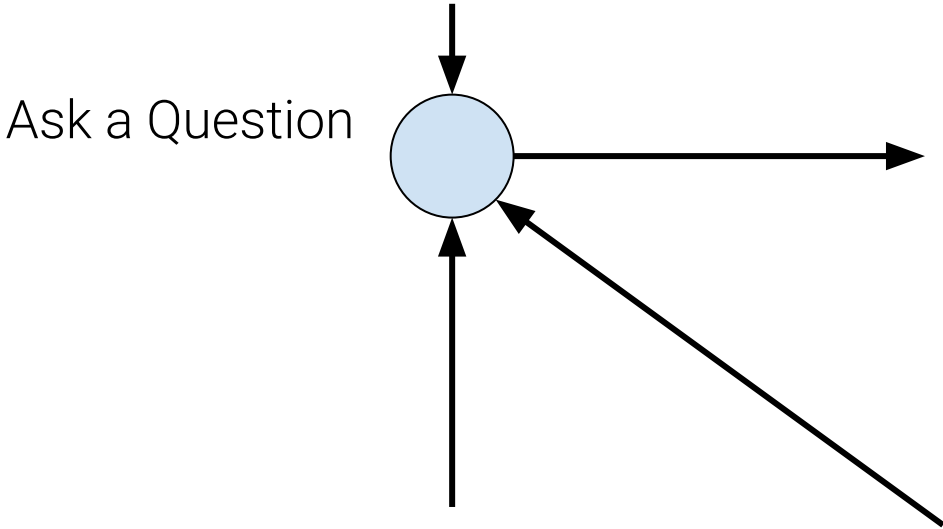


The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

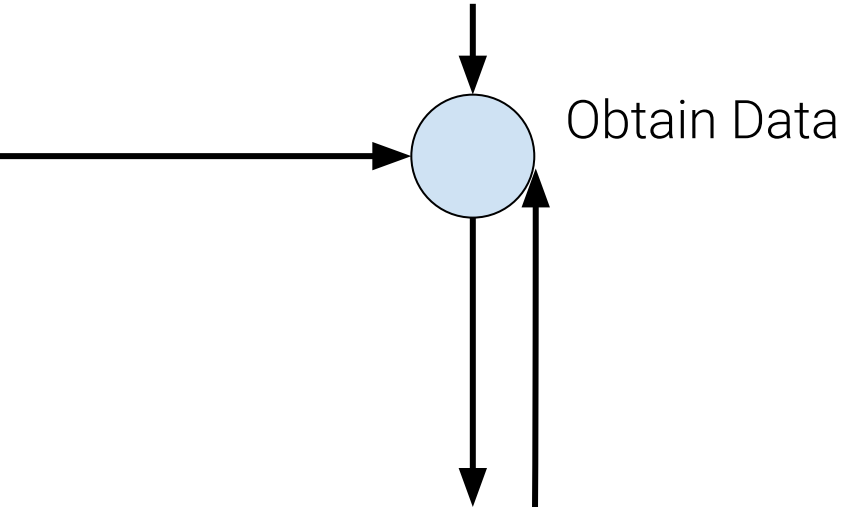
1. Question/Problem Formulation

- 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 for success?

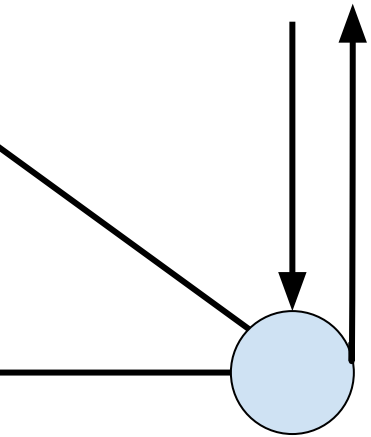


2. Data Acquisition and Cleaning

- 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?



3. Exploratory Data Analysis & Visualization

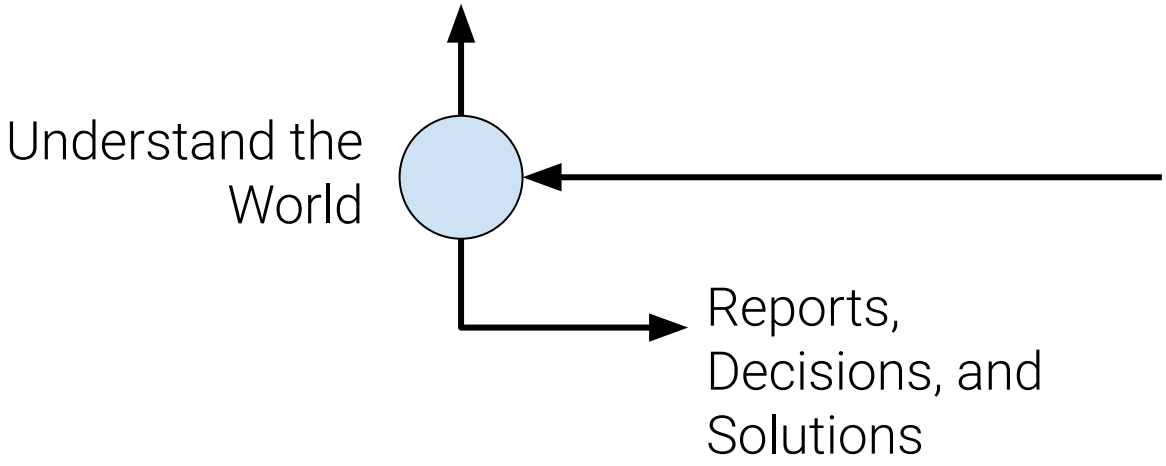


Understand the Data

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

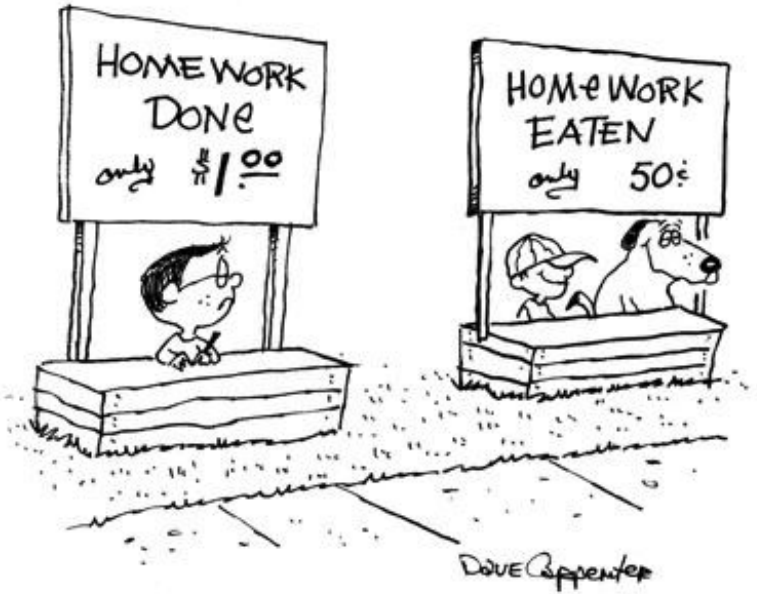
4. Prediction and Inference

- 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?



Homework #1

Due on Mar 19 (Fri) 11:59pm



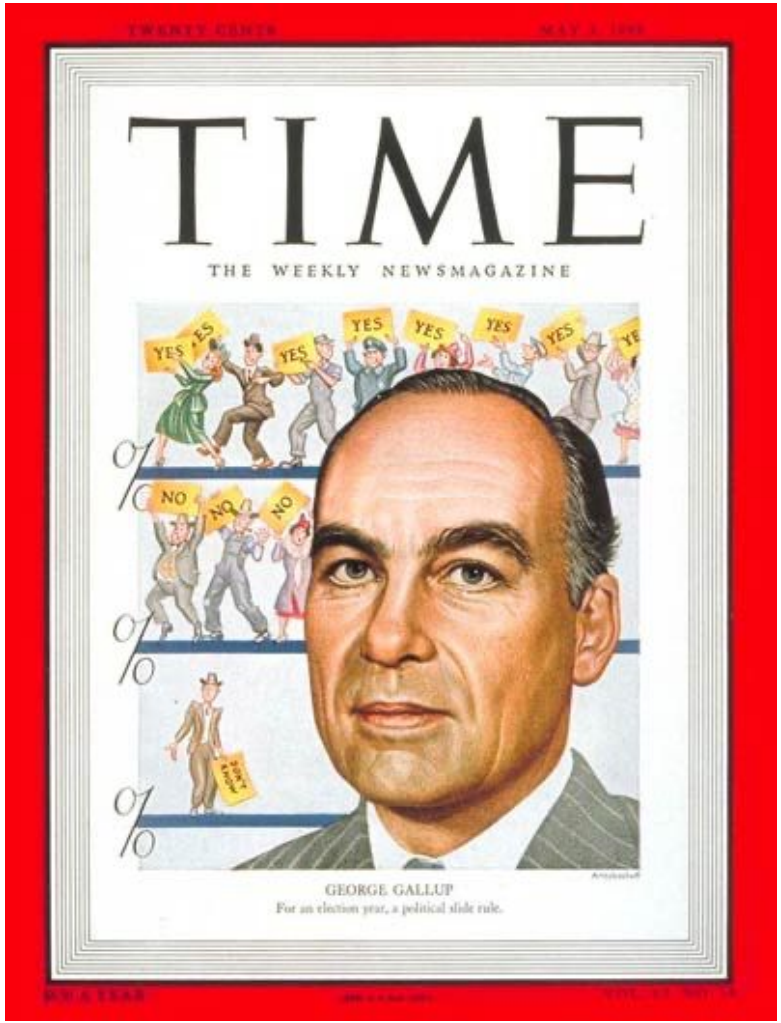
Dory

- https://www.dory.app/c/korea.ac.kr/a99bbb88_cose471-mar3

Next week

Probability and Data Design

- Why we need to sample in the first place
- What it means for our sample to be biased
- How to prevent these biases in our samples
- What exactly a sampling frame is, and why choosing a good one is important

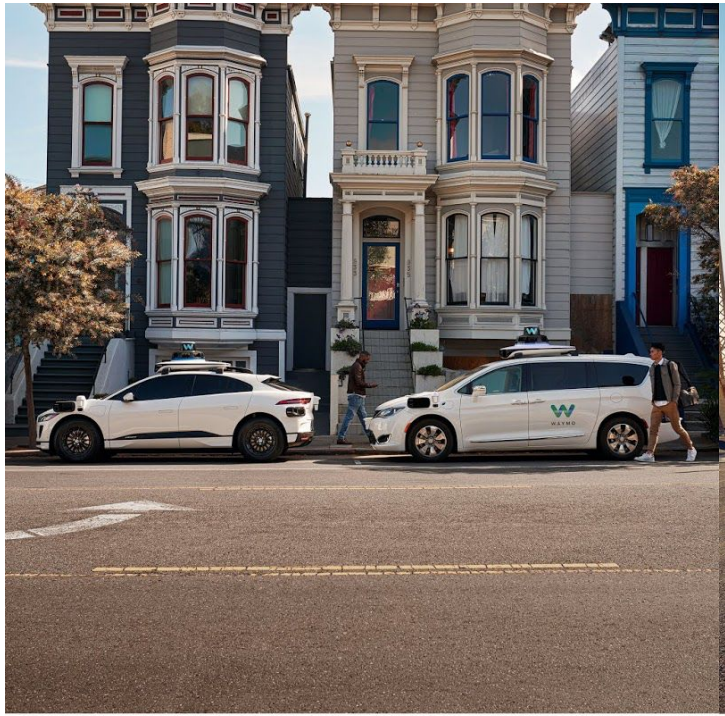


Introduction to Self-driving Cars

History (Waymo Journey)

This material may be copyrighted by Waymo LLC.

Building the World's Most Experienced Driver



WAYMO **ONE**

WAYMO **VIA**

Building the World's Most Experienced Driver

1,500+
monthly active riders

3X
tripled the number of weekly rides since January 2019

100,000+
number of total rides served since launching our rider programs in 2017

These numbers represent information for Waymo One, including the early rider program

Winter 2018
- Waymo One launches in Metro Phoenix

Spring 2019
- The Waymo Android app becomes available on Google Play
- Riders can now listen to their favorite song during their Waymo rides with the launch of our music feature

Summer / Fall 2019
- Fully driverless rides ramp up within Waymo's early rider program

Winter 2019
- The Waymo iOS app becomes available on the App Store
- Waymo One turns one! We're now serving over 1,500 riders

Building the World's Most Experienced Driver

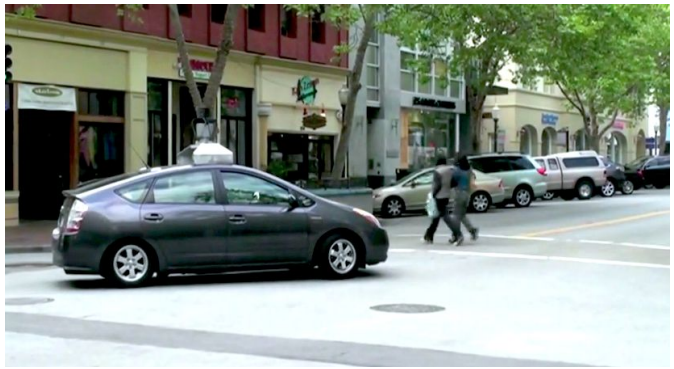
- **20 million+ miles** on public roads
- **15 billion+ miles** in simulation
- **25+ cities** across the USA

2009

The Google self-driving car project begins

2009

The Google self-driving car project begins



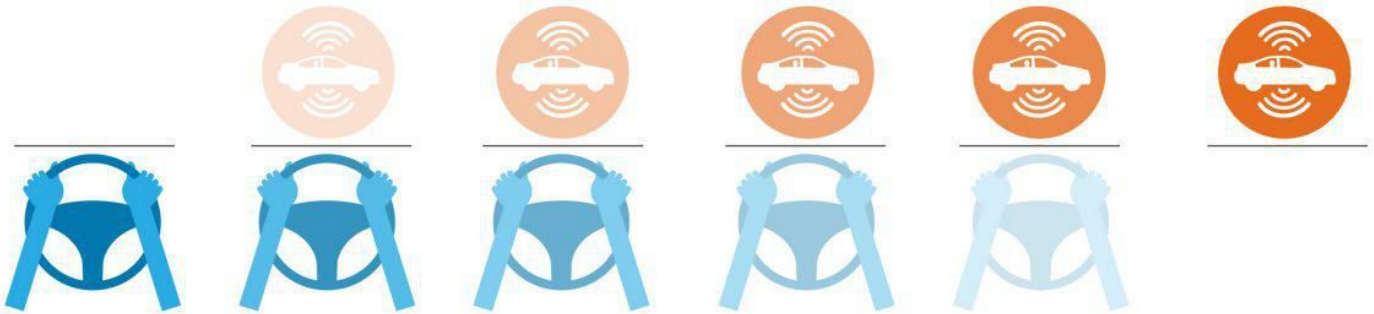
2009-2010

100-mile challenge

2013

Decided to focus on fully autonomous

Five Levels of Vehicle Autonomy



Level 0

No automation: the driver is in complete control of the vehicle at all times.

Level 1

Driver assistance: the vehicle can assist the driver or take control of either the vehicle's speed, through cruise control, or its lane position, through lane guidance.

Level 2

Occasional self-driving: the vehicle can take control of both the vehicle's speed and lane position in some situations, for example on limited-access freeways.

Level 3

Limited self-driving: the vehicle is in full control in some situations, monitors the road and traffic, and will inform the driver when he or she must take control.

Level 4

Full self-driving under certain conditions: the vehicle is in full control for the entire trip in these conditions, such as urban ride-sharing.

Level 5

Full self-driving under all conditions: the vehicle can operate without a human driver or occupants.

Source: SAE & NHTSA

2015

First fully autonomous ride on public roads

2017

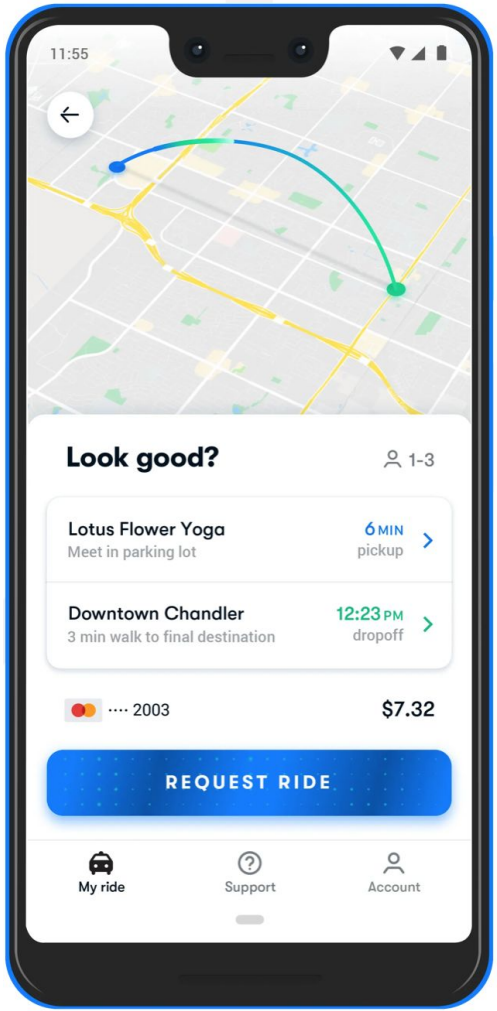
Welcomed members of the public to ride in self-driving vehicles for the first time

2019 Summer

Began matching riders with fully driverless rides

2020

Waymo ONE



Introduction to Autonomous Driving System?