UNIVERSITY OF GEORGIA

CSCI 3360 | Spring 2024

Data Science I

Jin Sun, PhD
School of Computing

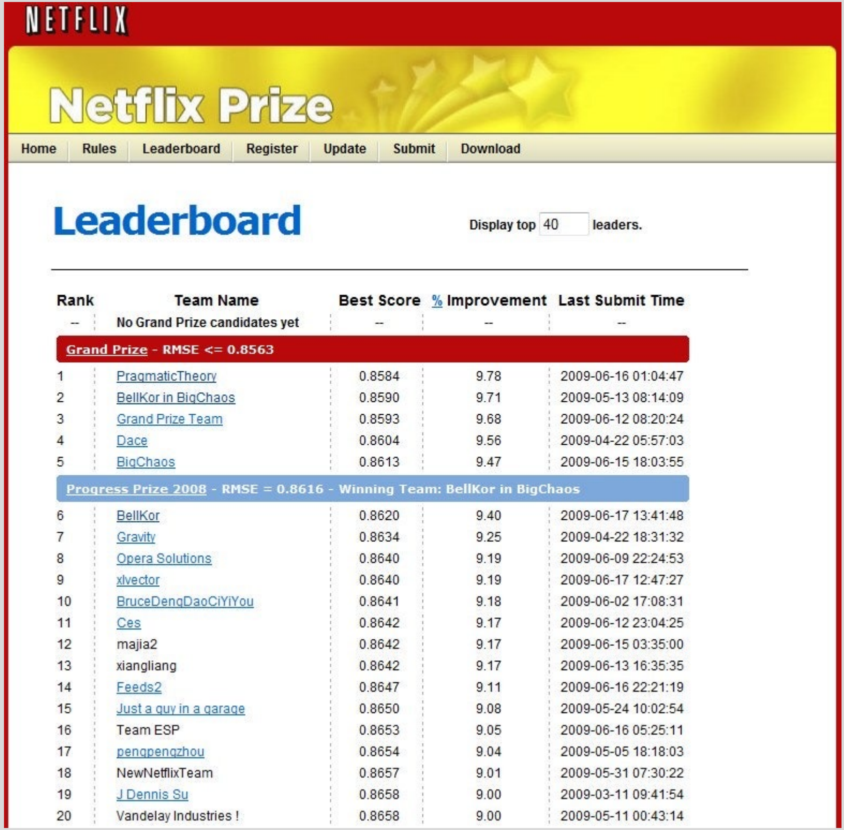
Week 1: Introduction

Outline

- Background
- Class logistics and policy
- Class topics
- Q&A

“students learning data science in Athens”

Data Science Motivating Examples



NETFLIX

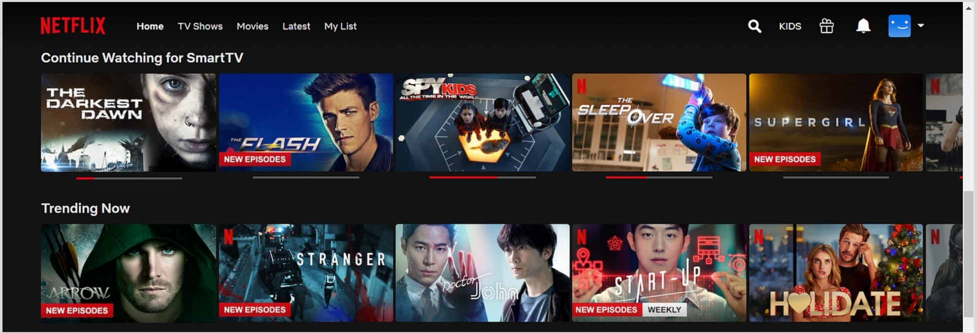
Netflix Prize

Home | Rules | Leaderboard | Register | Update | Submit | Download

Leaderboard

Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
1	PragmaticTheory	0.8584	9.78	2009-06-16 01:04:47
2	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-06-15 18:03:55
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8620	9.40	2009-06-17 13:41:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8640	9.19	2009-06-09 22:24:53
9	xlvector	0.8640	9.19	2009-06-17 12:47:27
10	BruceDengDaoCiYiYou	0.8641	9.18	2009-06-02 17:08:31
11	Ces	0.8642	9.17	2009-06-12 23:04:25
12	majia2	0.8642	9.17	2009-06-15 03:35:00
13	xiangliang	0.8642	9.17	2009-06-13 16:35:35
14	Feeds2	0.8647	9.11	2009-06-16 22:21:19
15	Just a guy in a garage	0.8650	9.08	2009-05-24 10:02:54
16	Team ESP	0.8653	9.05	2009-06-16 05:25:11
17	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
18	NewNetflixTeam	0.8657	9.01	2009-05-31 07:30:22
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14



NETFLIX

2009

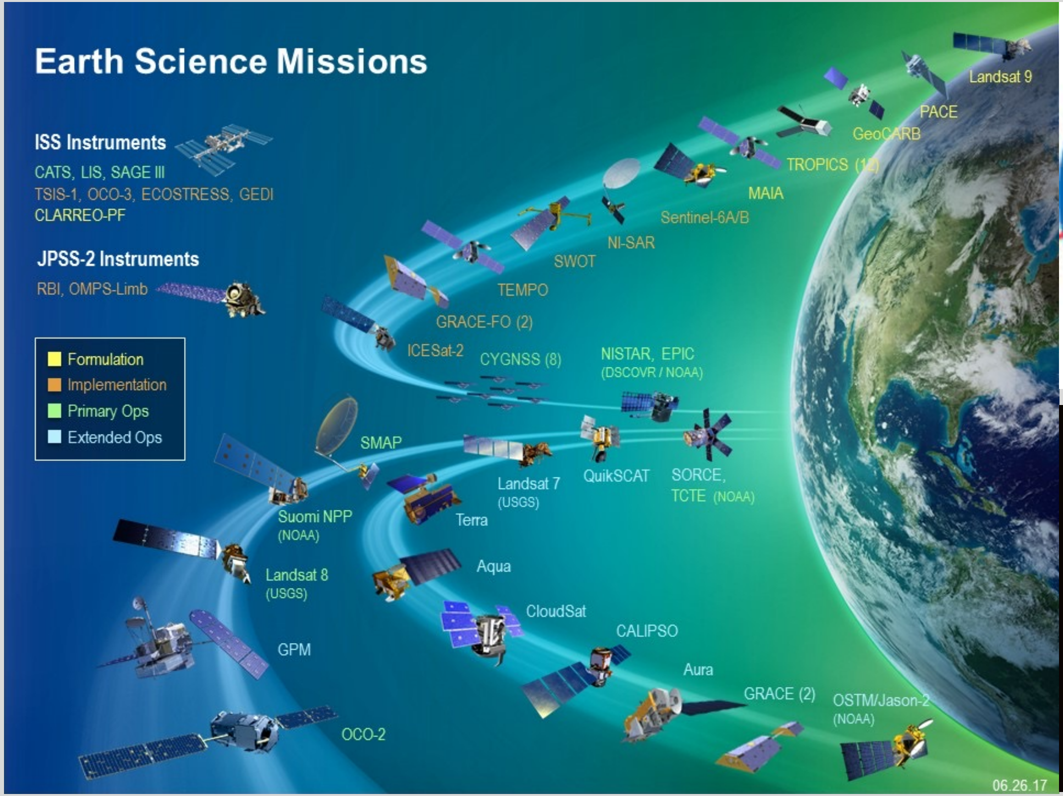
DATE 09-21-09

PAY TO THE ORDER OF BellKor's Pragmatic Chaos $ 1,000,000 00/100

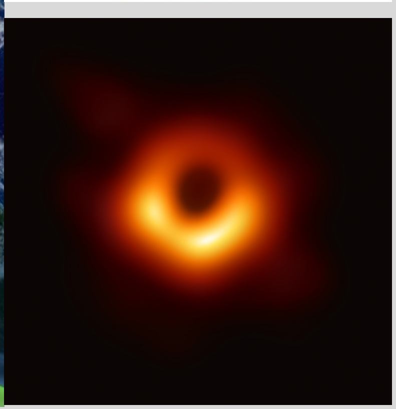
AMOUNT: ONE MILLION 00/100

FOR The Netflix Prize Reed Hastings

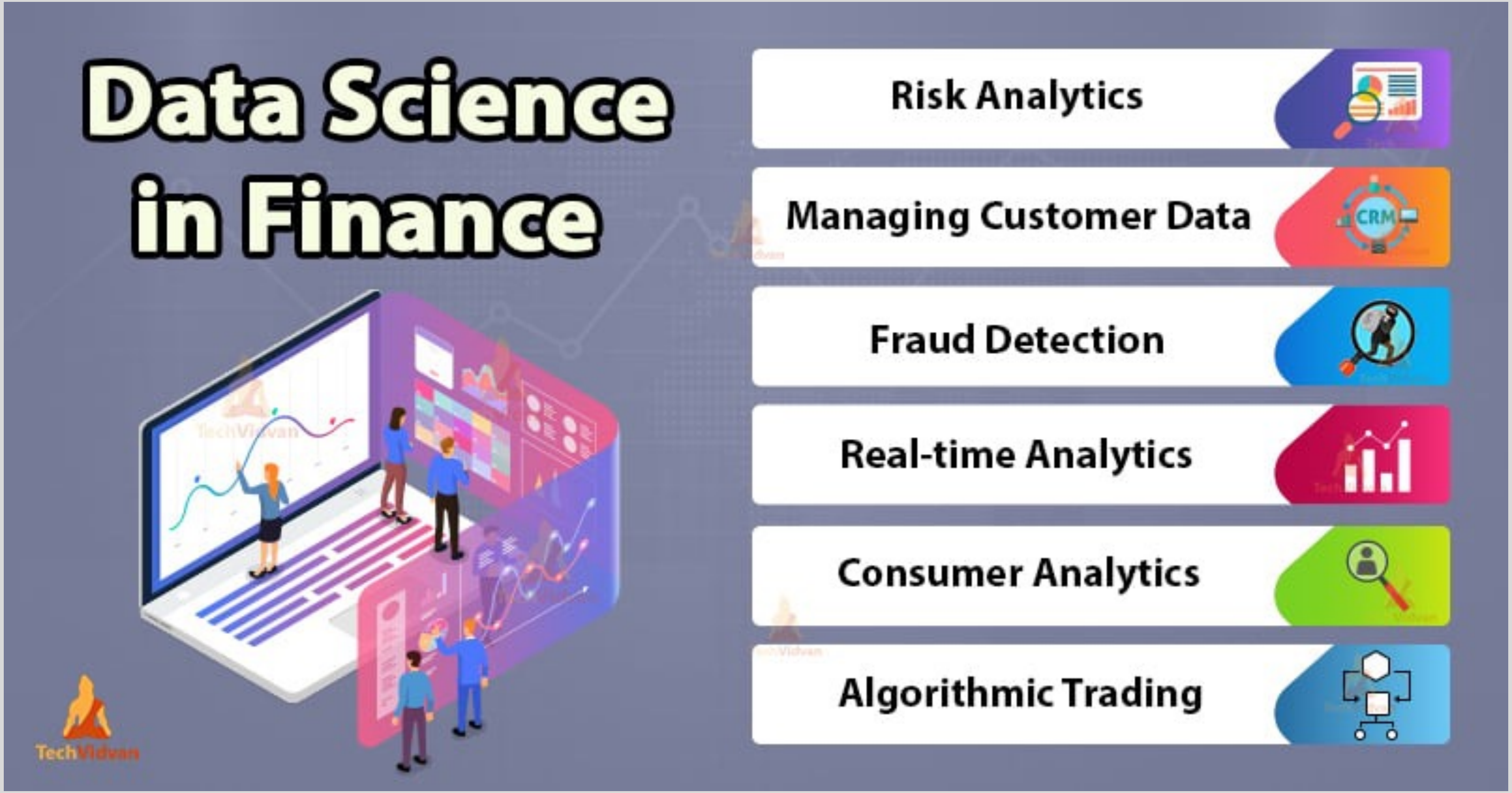
Data Science Motivating Examples



NASA



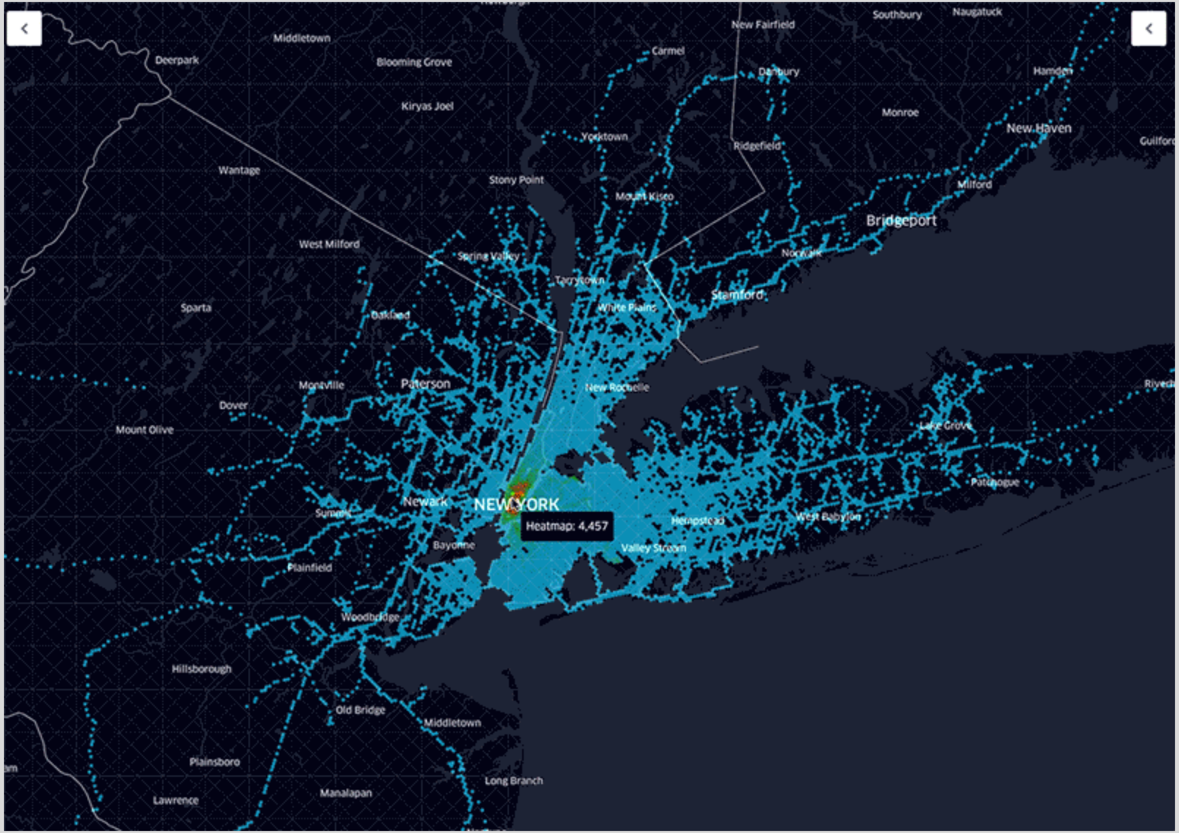
Data Science Motivating Examples



Data Science in Finance

- Risk Analytics
- Managing Customer Data
- Fraud Detection
- Real-time Analytics
- Consumer Analytics
- Algorithmic Trading

Data Science Motivating Examples



What is “Data Science”?

Data science (DS) is a **multidisciplinary** field that combines techniques from statistics, mathematics, computer science, and domain knowledge to extract insights and knowledge from data. It involves collecting, cleaning, analyzing, and interpreting large volumes of data to uncover patterns, trends, and relationships. Data scientists use various tools and techniques, such as data visualization, machine learning, and statistical modeling, to make data-driven decisions and solve complex problems. In this course, we will explore the fundamentals of data science and learn how to apply these techniques to real-world datasets.

What is “Data Science”?

Data science (DS) is a **multidisciplinary** field that combines techniques from statistics, mathematics, computer science, and domain knowledge to extract insights and knowledge from data. It involves collecting, cleaning, analyzing, and interpreting large volumes of data to uncover patterns, trends, and relationships. Data scientists use various tools and techniques, such as data visualization, machine learning, and statistical modeling, to make data-driven decisions and solve complex problems. In this course, we will explore the fundamentals of data science and learn how to apply these techniques to real-world datasets.

-- Generated by Copilot

DS vs others

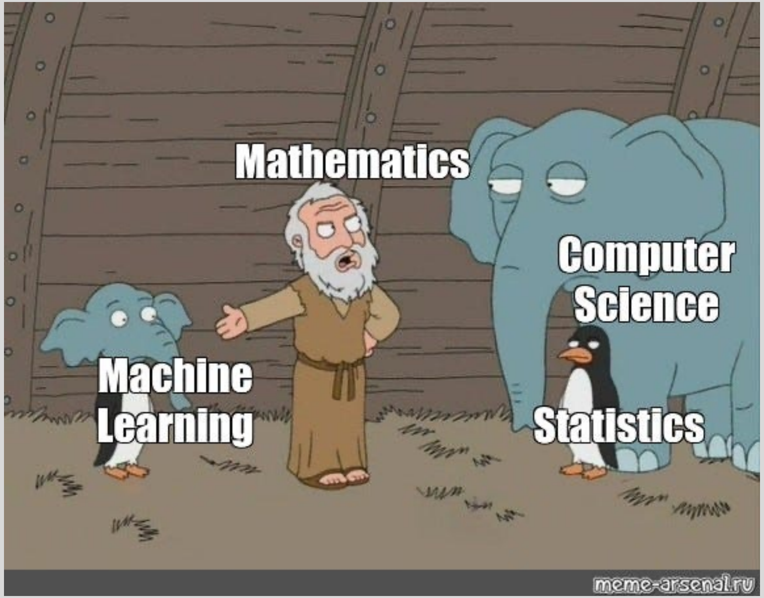
- Statistics

Statistics is a branch of mathematics dealing with data collection, analysis, interpretation, presentation, and organization. It provides methodologies to design experiments and surveys, and techniques to analyze the results to draw conclusions.

DS vs others

- Machine learning

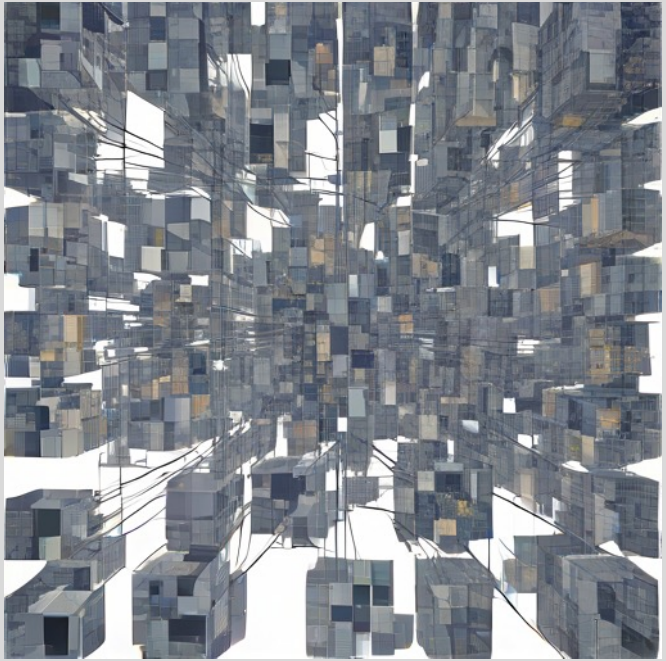
Machine Learning is a subset of artificial intelligence that uses statistical techniques to enable machines to improve with experience. It involves the creation of algorithms that can modify themselves without human intervention to produce desired outputs by feeding itself through structured data.



DS vs others

- Deep learning

Deep Learning is a subfield of machine learning that uses algorithms inspired by the structure and function of the brain's neural networks. These artificial neural networks are designed to simulate human decision-making and are becoming essential in areas where we are inundated with data. They provide a means to extract and learn complex patterns from massive amounts of data, and are often used in image recognition, speech recognition, and natural language processing tasks.

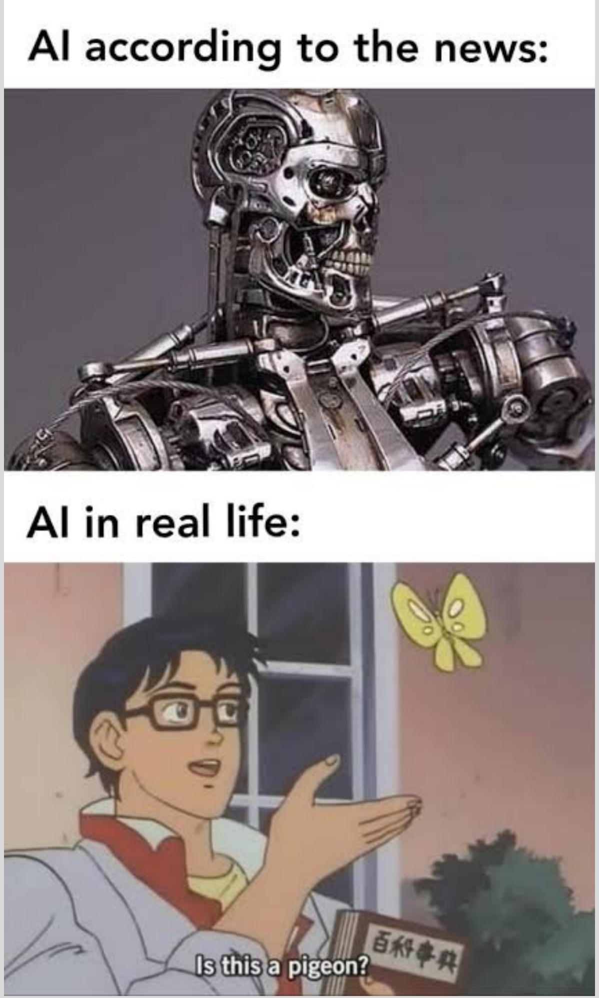


DS vs others

- “A.I.”

At this point, I don’t even know what AI means anymore.

It could be: knowledge base system, chat bot, simple statistical model, or an image filter.



What this class is about?

- Introduction to data science
- Fundamental understanding of data science pipeline
- Python + DS programming
- Hands-on examples

There is Data Science II for more advanced algorithms and techniques.

Learning objectives

- Familiar with Python and relevant libraries
- Familiar with data science pipeline
- Understand the fundamentals of data learning
- Can formulate a learning problem from raw data
- Can train simple models to learn from data
- Can validate the performance of a model

Outline

- Background
- Class logistics and policy
- Class topics
- Q&A

Class format

Lectures

Lab

Class format

~~Lectures~~ Theory

- Concepts
- Maths
- Slides and Whiteboard

~~Lab~~ Practice

- Coding
- Development
- Toy examples and playground

For each week's topic, we will start with the fundamental concepts of the topic and then learn to realize those concepts by programming.

The learning objectives of each week will be evaluated by in-class quiz.

Textbooks

- An introduction to statistical learning (https://www.statlearning.com/)
- Python data science handbook (https://jakevdp.github.io/PythonDataScienceHandbook/)

Both books are free online.

Teaching team

Instructor:

Prof. Jin Sun

Office Hours: Thursdays 4-5pm or by appointment

Office: 804 Boyd

Email: jinsun@uga.edu

Teaching assistant:

TBD

Evaluation and grading

The final course grade will be weighted as follows:

Quiz:	10%
Homework:	40%
Midterm exam:	10%
Final exam:	15%
Project:	25%

Late policy: 10% of total score deduction for each late day (including partial day).

Project

You will work in a team on a course project. Each team should have 2-3 members.
You are encouraged to design the project to solve a real-world problem.

Project Proposal (5%): What do you plan to do? What's the learning problem? Data?

Project Milestone (5%): Preliminary results and progress report.

Project Report and Presentation (15%): All results and findings.

Assessment and feedback

For each learning objective, we will have methods for you to do self-assessment.

Advanced teaching techniques might be explored in this class.

Feedbacks are welcome!

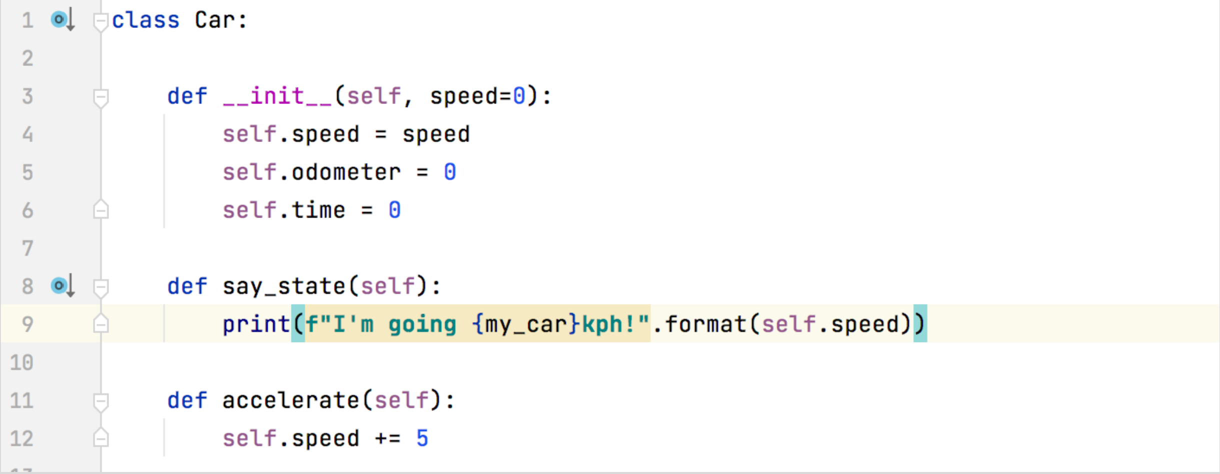
Outline

- Background
- Class logistics and policy
- Class topics, aka a whole semester in a day
- Q&A

Programming tools for DS

- Python

"THE" programming language of data science. It is simple to config, develop, and run. Most importantly, there is a huge community.

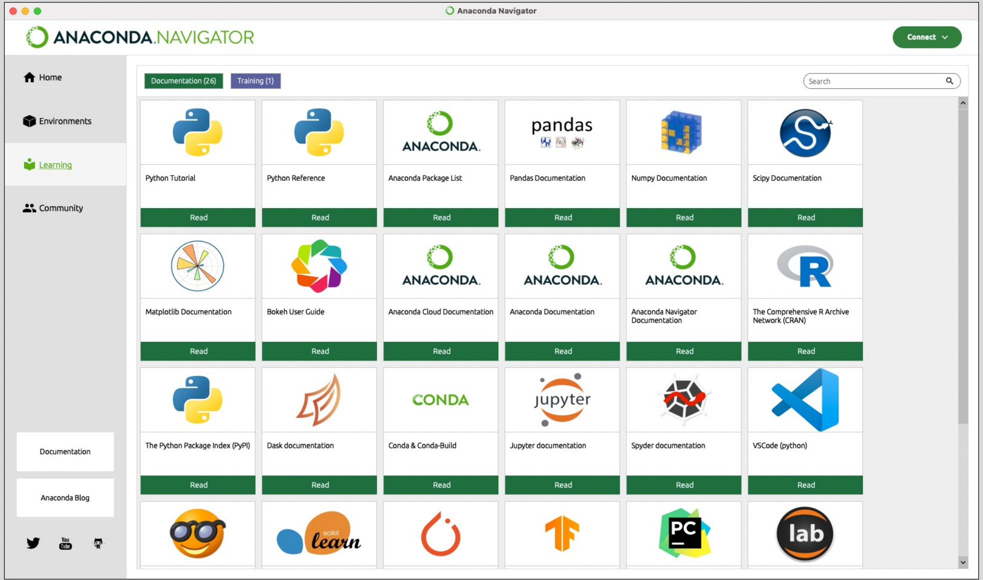


```python
class Car:

    def __init__(self, speed=0):
        self.speed = speed
        self.odometer = 0
        self.time = 0

    def say_state(self):
        print(f"I'm going {my_car}kph!".format(self.speed))

    def accelerate(self):
        self.speed += 5
```

Programming tools for DS

- Conda

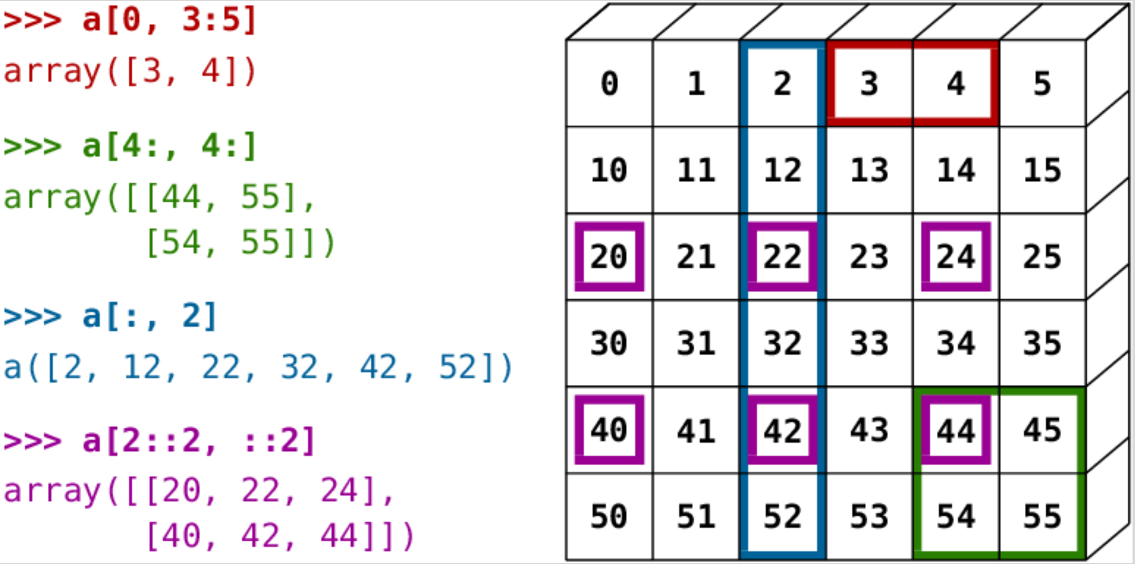
Package manager for Python. Make it easy for users to manage environments and packages. It is cross-platform.



Programming tools for DS

- Numpy

Python library with Matlab-like syntax for matrices and vectors operations. And much more.



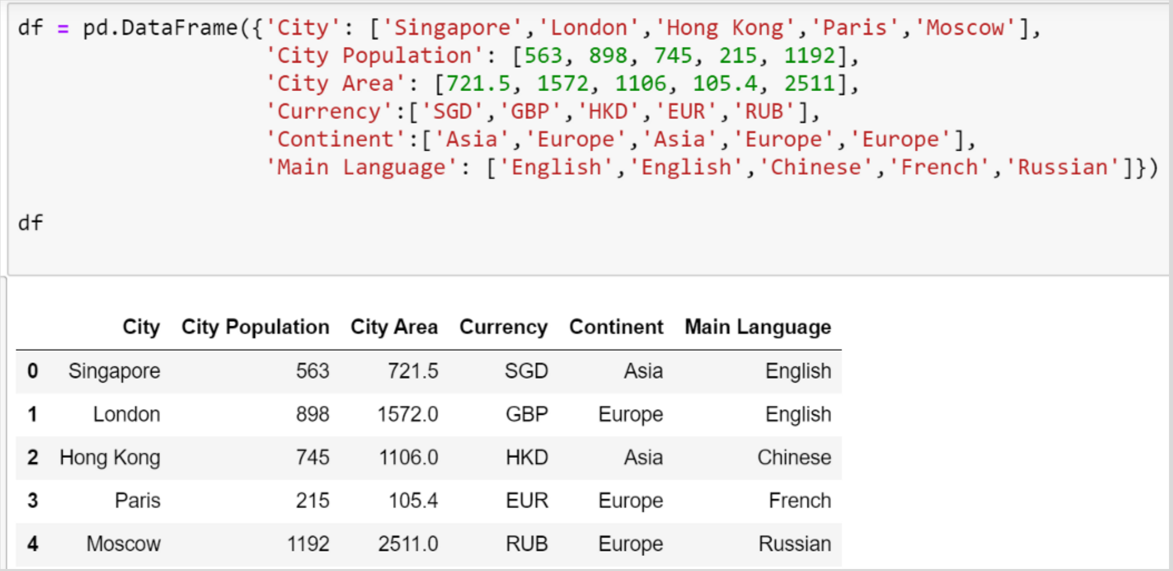
```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Programming tools for DS

- Pandas

Python library to handle data.



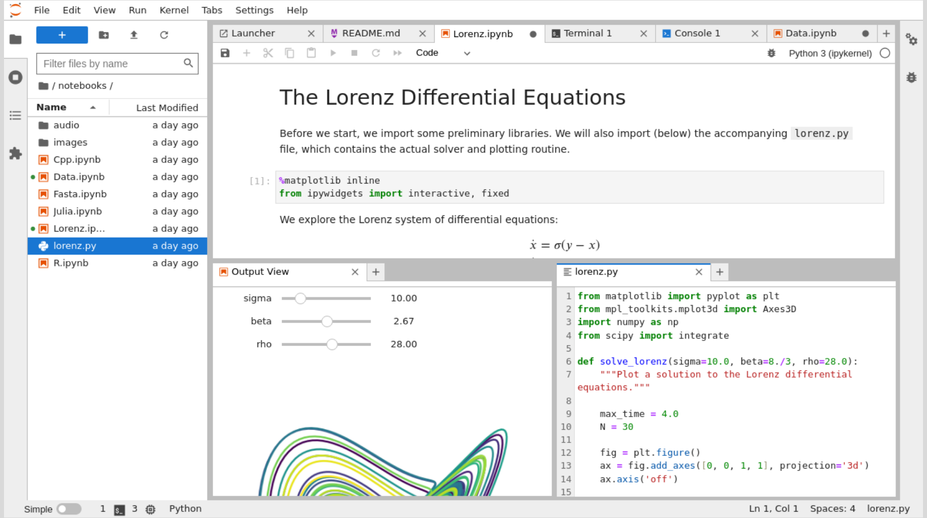
```
df = pd.DataFrame({'City': ['Singapore','London','Hong Kong','Paris','Moscow'],
                   'City Population': [563, 898, 745, 215, 1192],
                   'City Area': [721.5, 1572, 1106, 105.4, 2511],
                   'Currency':['SGD','GBP','HKD','EUR','RUB'],
                   'Continent':['Asia','Europe','Asia','Europe','Europe'],
                   'Main Language': ['English','English','Chinese','French','Russian']})

df
```

	City	City Population	City Area	Currency	Continent	Main Language
0	Singapore	563	721.5	SGD	Asia	English
1	London	898	1572.0	GBP	Europe	English
2	Hong Kong	745	1106.0	HKD	Asia	Chinese
3	Paris	215	105.4	EUR	Europe	French
4	Moscow	1192	2511.0	RUB	Europe	Russian

Programming tools for DS

- Jupyter notebook

Intuitive IDE. Nice interface and flexibility. You can run all coding part of this class on Jupyter notebook, or on Google Colab.



Name	Last Modified
audio	a day ago
images	a day ago
Cpp.ipynb	a day ago
Data.ipynb	a day ago
Fasta.ipynb	a day ago
Julia.ipynb	a day ago
Lorenz.ip...	a day ago
lorenz.py	a day ago
R.ipynb	a day ago

The Lorenz Differential Equations

Before we start, we import some preliminary libraries. We will also import (below) the accompanying `lorenz.py` file, which contains the actual solver and plotting routine.

```
[1]: %matplotlib inline
     from ipywidgets import interactive, fixed
```

We explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$

```
sigma    10.00
beta      2.67
rho      28.00
```

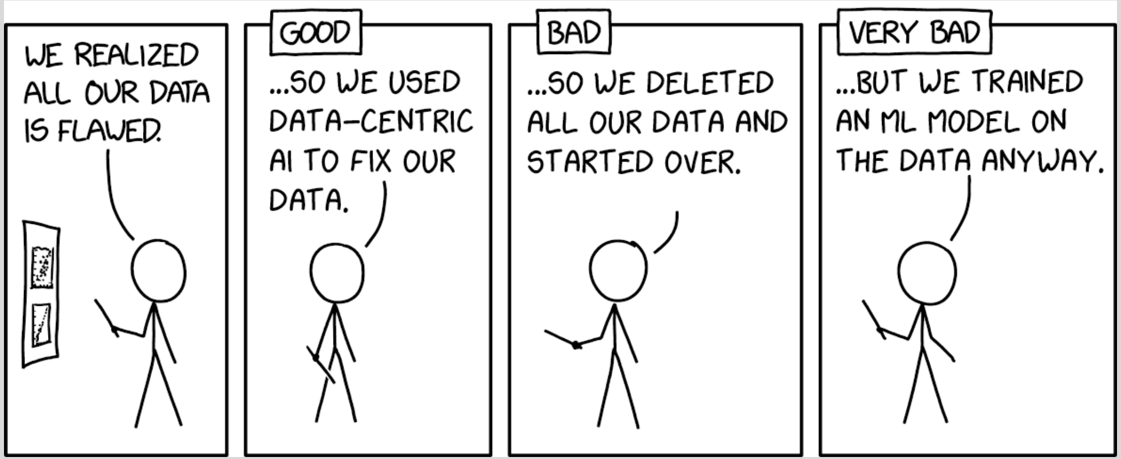
```
from matplotlib import pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
from scipy import integrate

def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""

    max_time = 4.0
    N = 30

    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

```

Data

Of course, the first thing we should really be talking about is **data**.

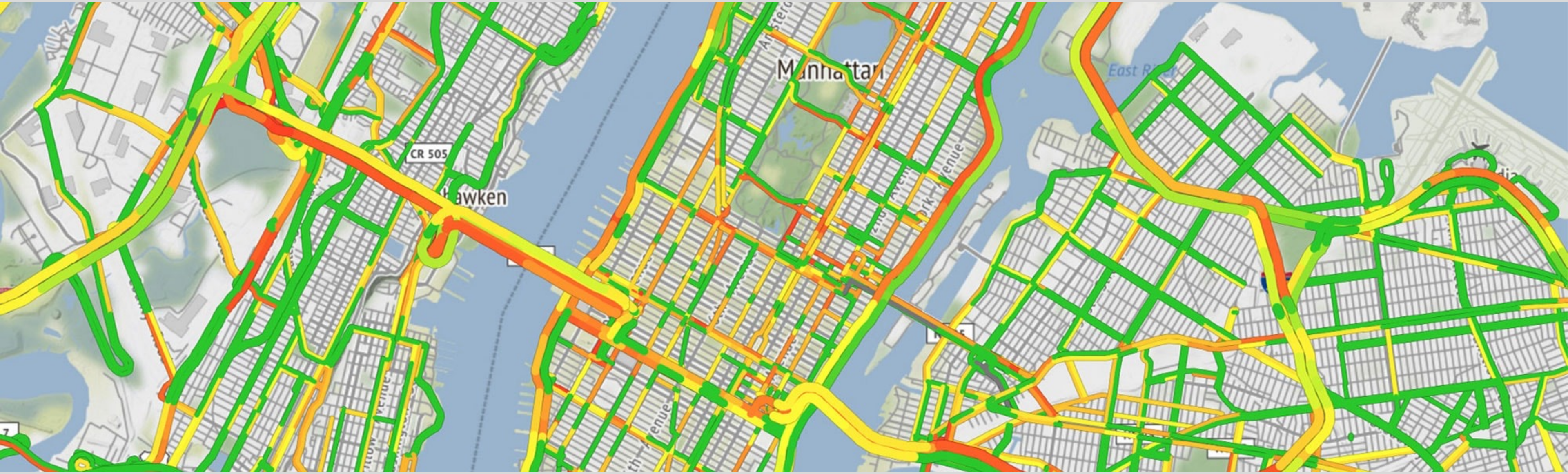
Data is the most important thing in your data science project. Period.

Without high quality data, you cannot run meaningful analysis or learn useful models from it.


WE REALIZED ALL OUR DATA IS FLAWED.
GOOD
...SO WE USED DATA-CENTRIC AI TO FIX OUR DATA.
BAD
...SO WE DELETED ALL OUR DATA AND STARTED OVER.
VERY BAD
...BUT WE TRAINED AN ML MODEL ON THE DATA ANYWAY.

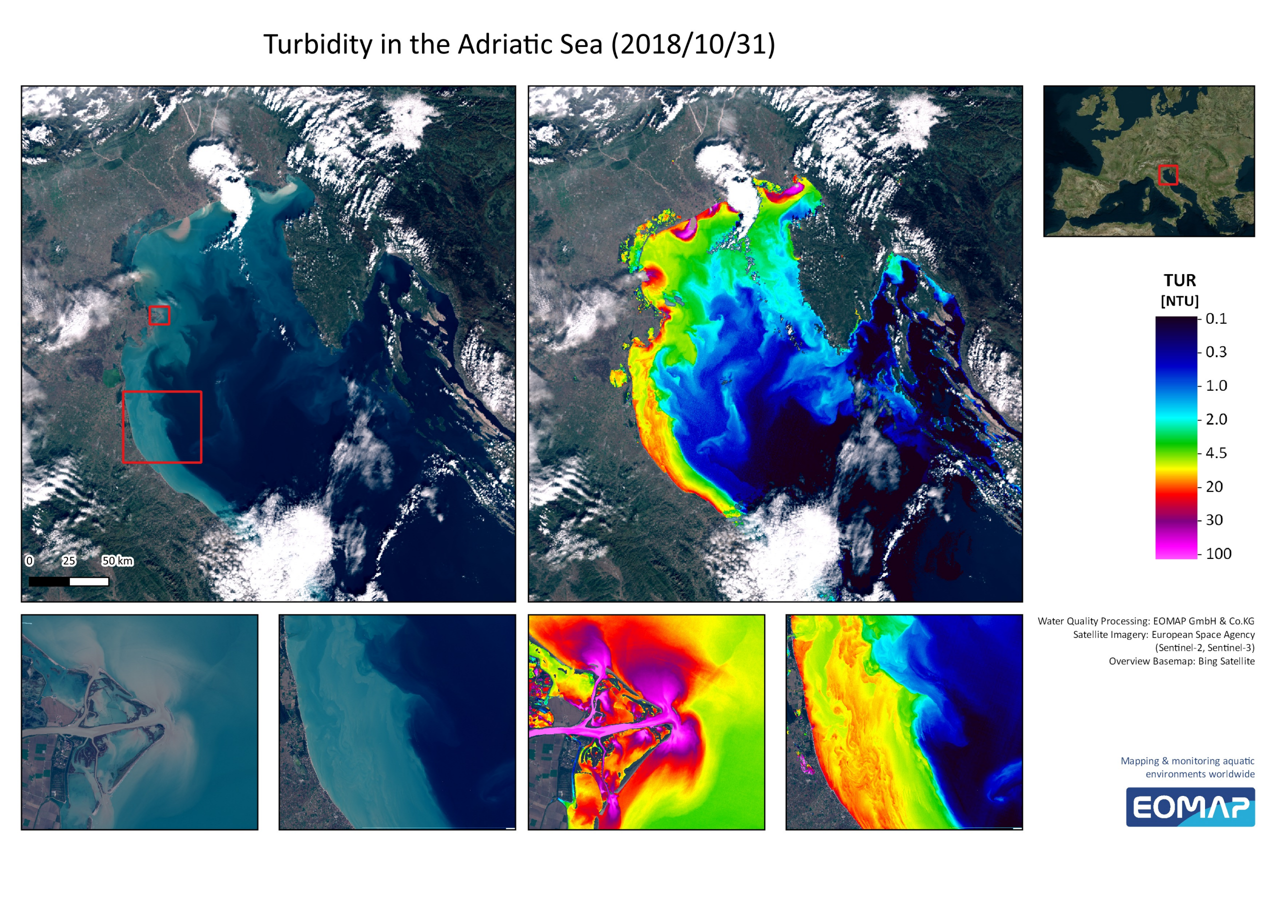
Example data streams

Traffic



Example data streams

Satellite

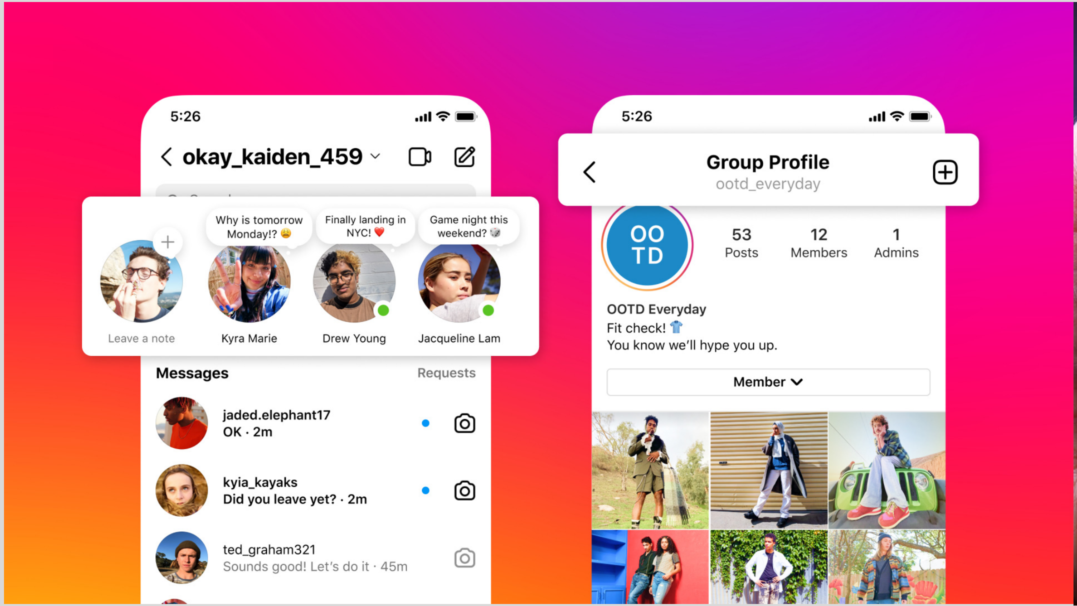


Turbidity in the Adriatic Sea (2018/10/31)

Water Quality Processing: EOMAP GmbH & Co.KG
Satellite Imagery: European Space Agency
(Sentinel-2, Sentinel-3)
Overview Basemap: Bing Satellite

Example data streams

Images and videos



5:26

okay_kaiden_459

Why is tomorrow Monday!? 😩

Finally landing in NYC! ❤️

Game night this weekend? 🎲

Leave a note

Kyra Marie

Drew Young

Jacqueline Lam

Messages

Requests

jaded.elephant17
OK · 2m

kyia_kayaks
Did you leave yet? · 2m

ted_graham321
Sounds good! Let's do it · 45m

5:26

Group Profile
ootd_everyday

OO TD

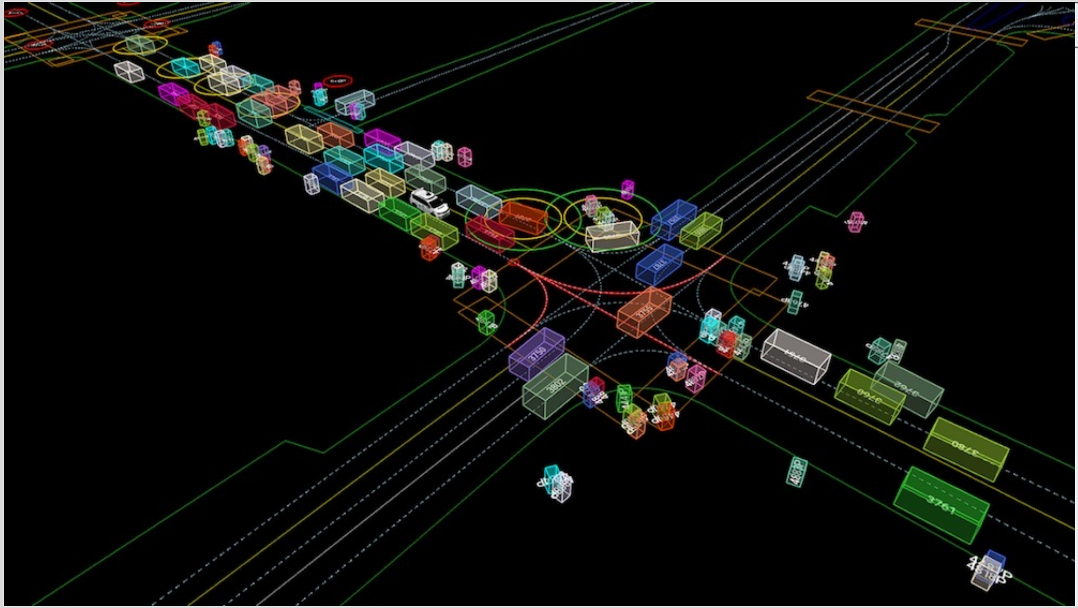
53	12	1
Posts	Members	Admins

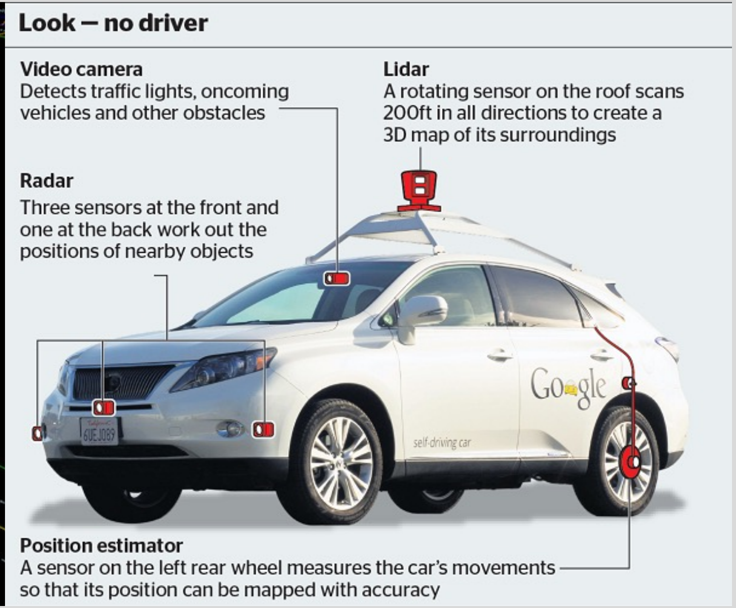
OOTD Everyday
Fit check! 👕
You know we'll hype you up.

Member

Example data streams

Images and videos





Look — no driver

Video camera
Detects traffic lights, oncoming vehicles and other obstacles

Lidar
A rotating sensor on the roof scans 200ft in all directions to create a 3D map of its surroundings

Radar
Three sensors at the front and one at the back work out the positions of nearby objects

Position estimator
A sensor on the left rear wheel measures the car's movements so that its position can be mapped with accuracy

Example data streams

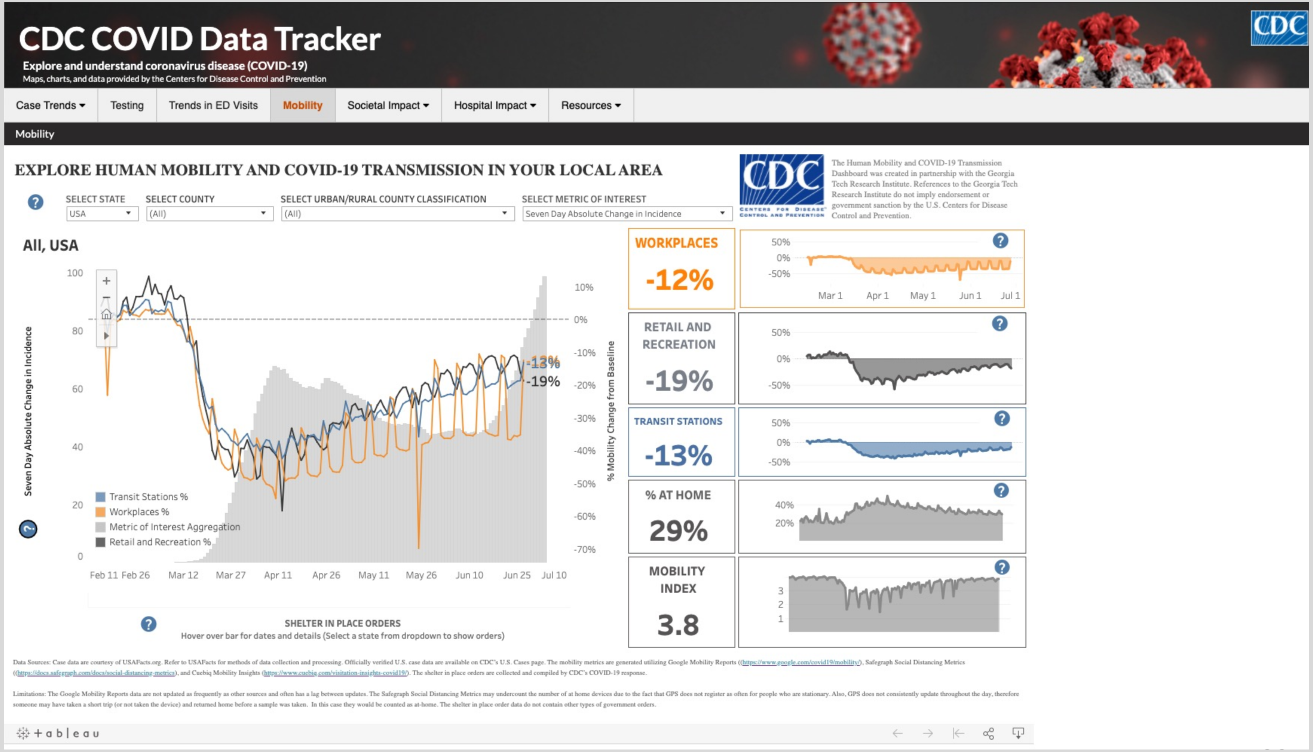
Images and videos

- Surveillance

Example data streams

Health

- Disease tracker



CDC COVID Data Tracker

Explore and understand coronavirus disease (COVID-19)

Maps, charts, and data provided by the Centers for Disease Control and Prevention

EXPLORE HUMAN MOBILITY AND COVID-19 TRANSMISSION IN YOUR LOCAL AREA

SELECT STATE: USA

SELECT COUNTY: (All)

SELECT URBAN/RURAL COUNTY CLASSIFICATION: (All)

SELECT METRIC OF INTEREST: Seven Day Absolute Change in Incidence

The Human Mobility and COVID-19 Transmission Dashboard was created in partnership with the Georgia Tech Research Institute. References to the Georgia Tech Research Institute do not imply endorsement or government sanction by the U.S. Centers for Disease Control and Prevention.

SHELTER IN PLACE ORDERS

Hover over bar for dates and details (Select a state from dropdown to show orders)

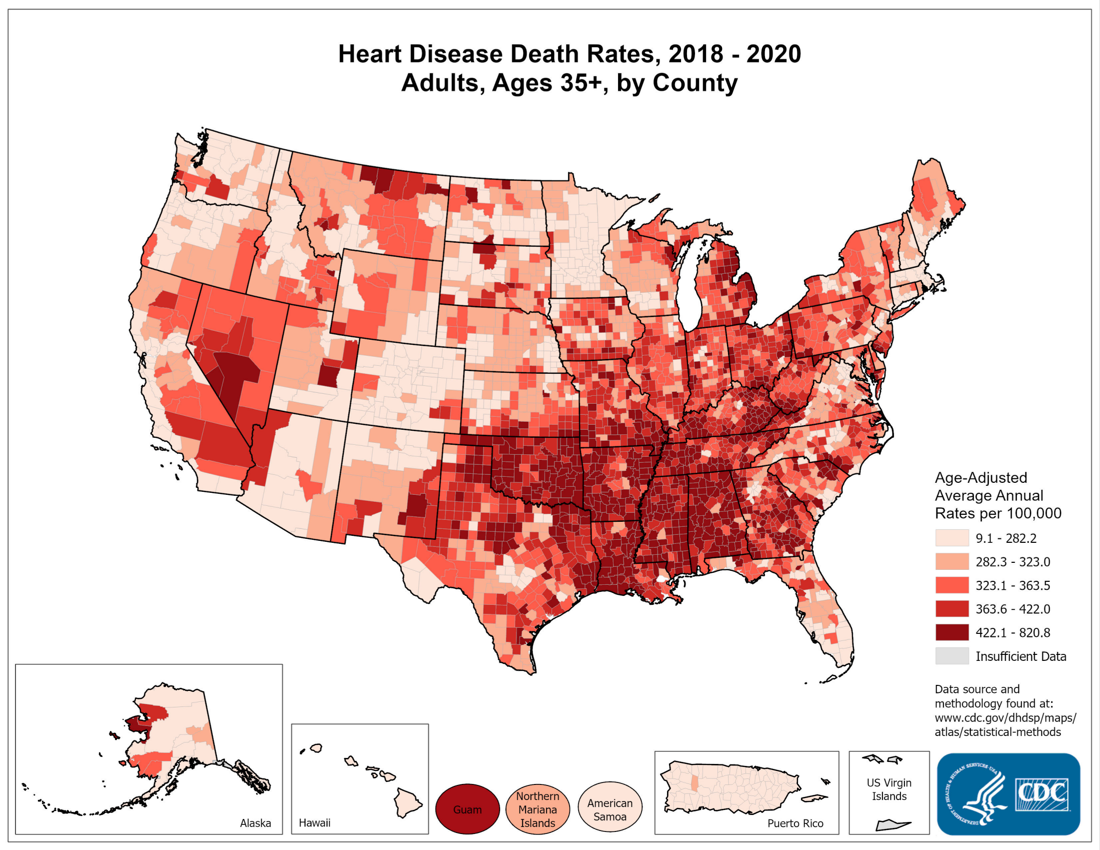
Data Sources: Case data are courtesy of USAFacts.org. Refer to USAFacts for methods of data collection and processing. Officially verified U.S. case data are available on CDC's U.S. Cases page. The mobility metrics are generated utilizing Google Mobility Reports ((https://www.google.com/covid19/mobility/), Safegraph Social Distancing Metrics ((https://docs.safegraph.com/docs/social-distancing-metrics), and Cuebiq Mobility Insights (https://www.cuebiq.com/visitation-insights-covid19/). The shelter in place orders are collected and compiled by CDC's COVID-19 response.

Limitations: The Google Mobility Reports data are not updated as frequently as other sources and often has a lag between updates. The Safegraph Social Distancing Metrics may undercount the number of at home devices due to the fact that GPS does not register as often for people who are stationary. Also, GPS does not consistently update throughout the day, therefore someone may have taken a short trip (or not taken the device) and returned home before a sample was taken. In this case they would be counted as at-home. The shelter in place order data do not contain other types of government orders.

Example data streams

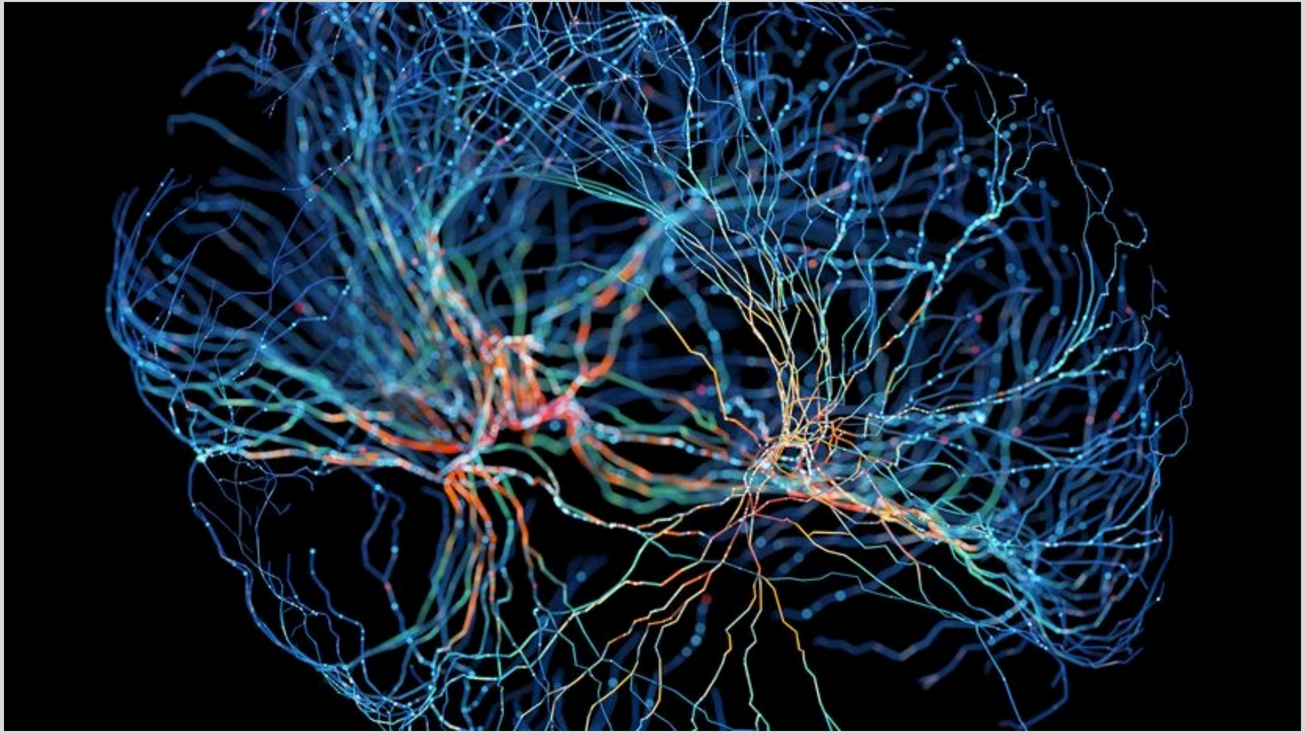
Health

- Disease tracker


Heart Disease Death Rates, 2018 - 2020
Adults, Ages 35+, by County
Age-Adjusted Average Annual Rates per 100,000
9.1 - 282.2
282.3 - 323.0
323.1 - 363.5
363.6 - 422.0
422.1 - 820.8
Insufficient Data
Data source and methodology found at: www.cdc.gov/dhdsp/maps/atlas/statistical-methods
Alaska
Hawaii
Guam
Northern Mariana Islands
American Samoa
Puerto Rico
US Virgin Islands
CDC

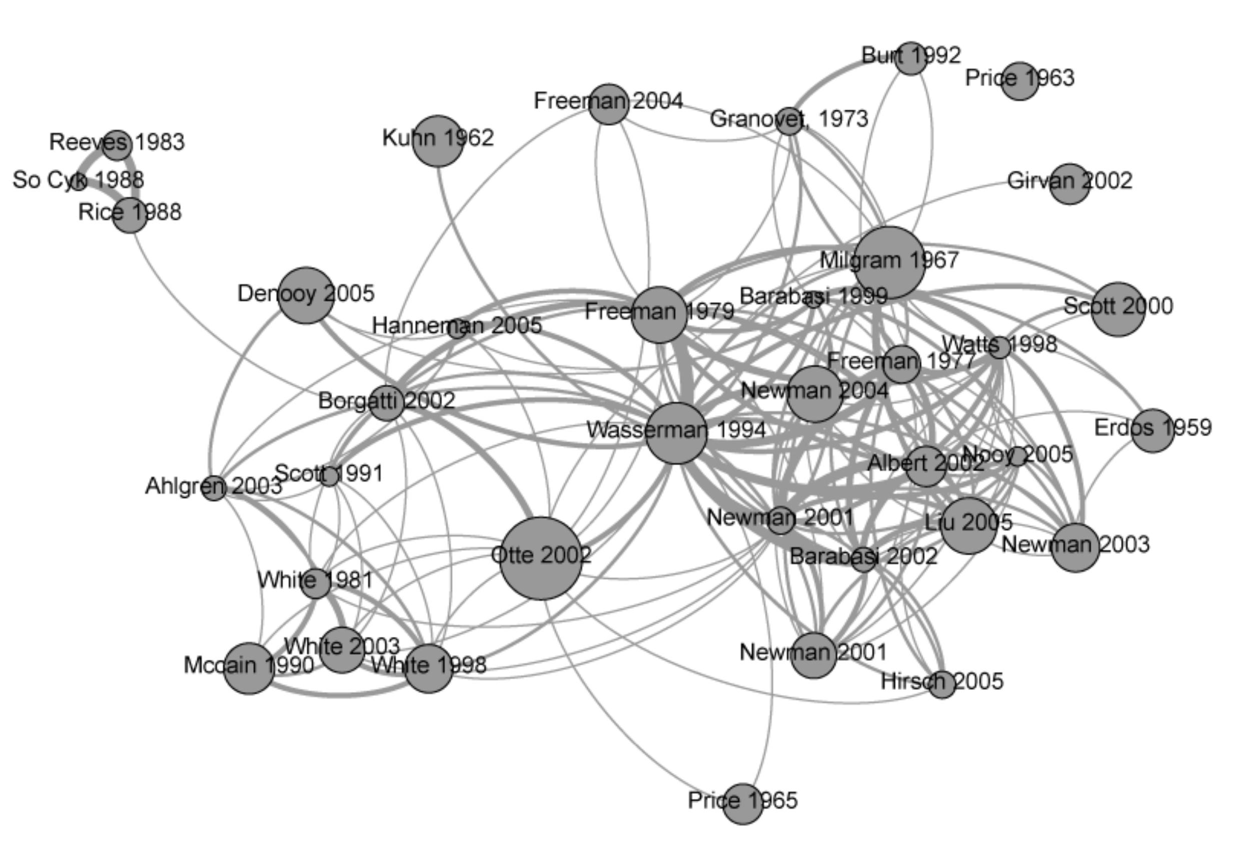
Example data streams

Brain



Example data streams

Scientific research



Data formats

Exact format of data depends on its domain of applications.

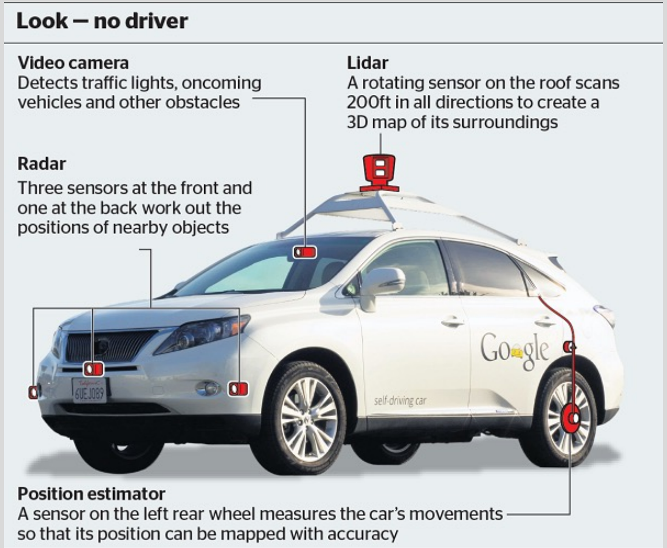
- Text strings
- Images
- Geo-tagged sensor readings
- Structured forms

Data formats

In most cases, raw data are in mixed formats.

For example, in self-driving cars, data contains:

- RGB images
- LiDAR
- GPS
- Vehicle speed and orientation
- etc



Look — no driver

Video camera
Detects traffic lights, oncoming vehicles and other obstacles

Lidar
A rotating sensor on the roof scans 200ft in all directions to create a 3D map of its surroundings

Radar
Three sensors at the front and one at the back work out the positions of nearby objects

Position estimator
A sensor on the left rear wheel measures the car's movements so that its position can be mapped with accuracy

(Almost) Infinite amount of data

In many scenarios, we have more data than we can handle.

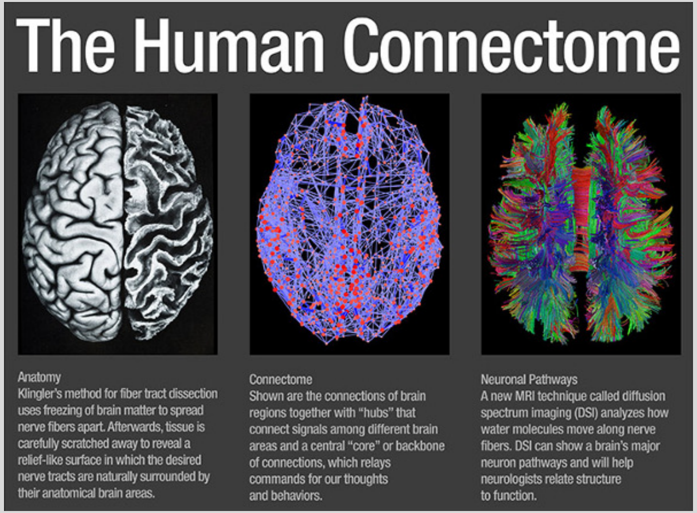
How much data YouTube generates?

"Every minute, over 500 hours (about 3 weeks) of video are uploaded to YouTube. According to Global Media Insight, that's equivalent to over 300,000 hours (about 34 years) of video every day. This represents around 30,000 hours of new video content being uploaded per hour." -- src

Not enough data

On the other hand, in many cases we simply don't have enough data or no data at all.

- Limited data: e.g., We cannot record the activity of every single neuron in human brain.



The Human Connectome

Anatomy
Klingler's method for fiber tract dissection uses freezing of brain matter to spread nerve fibers apart. Afterwards, tissue is carefully scratched away to reveal a relief-like surface in which the desired nerve tracts are naturally surrounded by their anatomical brain areas.

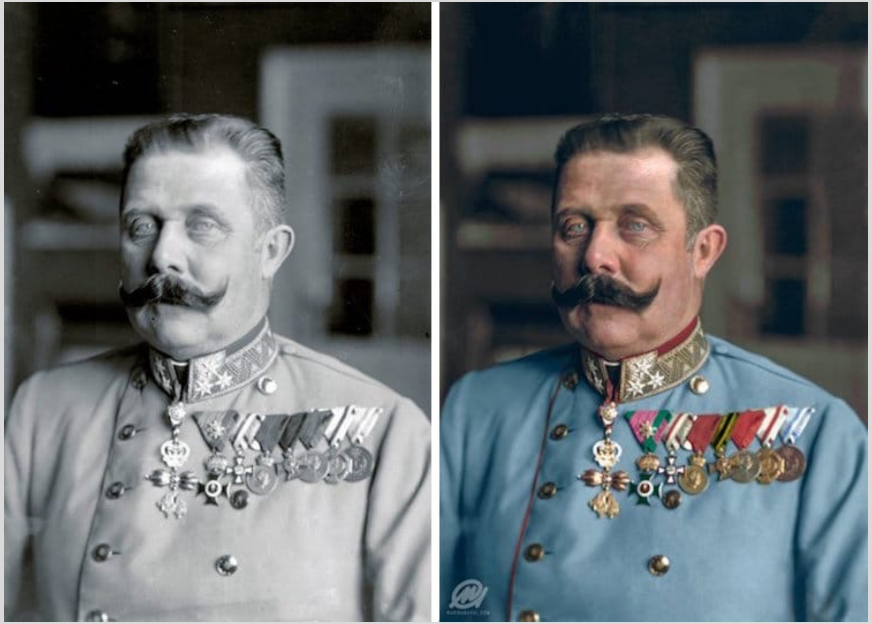
Connectome
Shown are the connections of brain regions together with "hubs" that connect signals among different brain areas and a central "core" or backbone of connections, which relays commands for our thoughts and behaviors.

Neuronal Pathways
A new MRI technique called diffusion spectrum imaging (DSI) analyzes how water molecules move along nerve fibers. DSI can show a brain's major neuron pathways and will help neurologists relate structure to function.

Not enough data

On the other hand, in many cases we simply don't have enough data or no data at all.

- No data: e.g., We do not have colored photo of the world before it was invented.



Data vs label

‘labels’ are usually the ‘meaning’ attached to data.

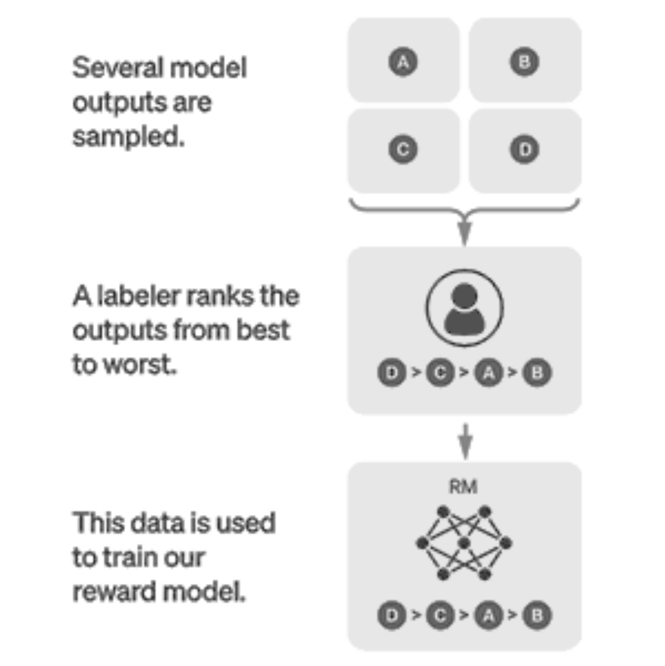
In many applications, raw data are already labeled. Can you think of such cases?

In many applications, raw data recorded are ‘unlabeled’.
Can you think of such cases?

Data vs label

The process of obtaining the labels via human effort is called annotation.

For example, ChatGPT was trained using human annotated conversations.

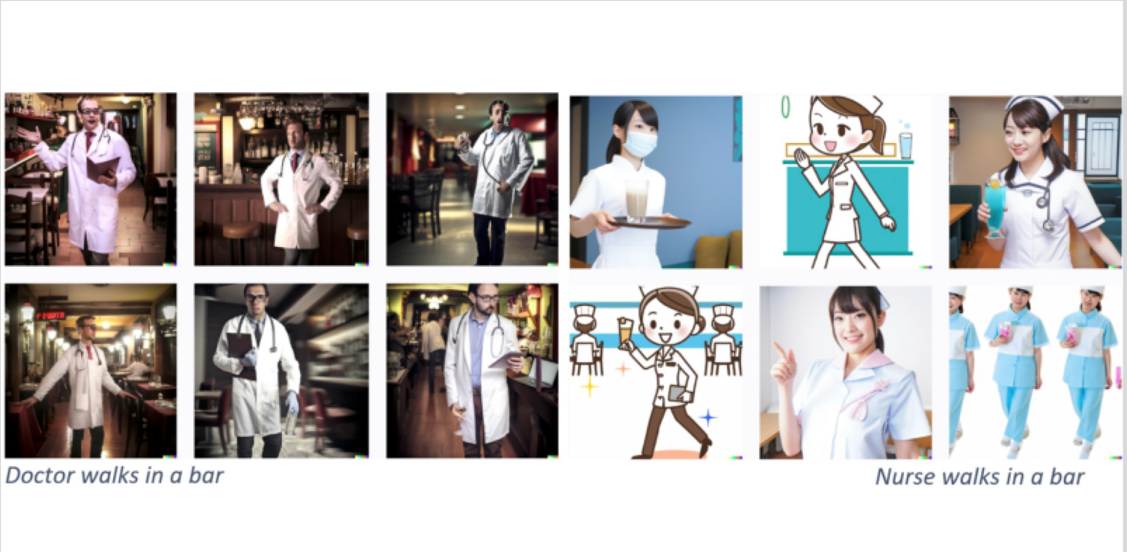


Several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

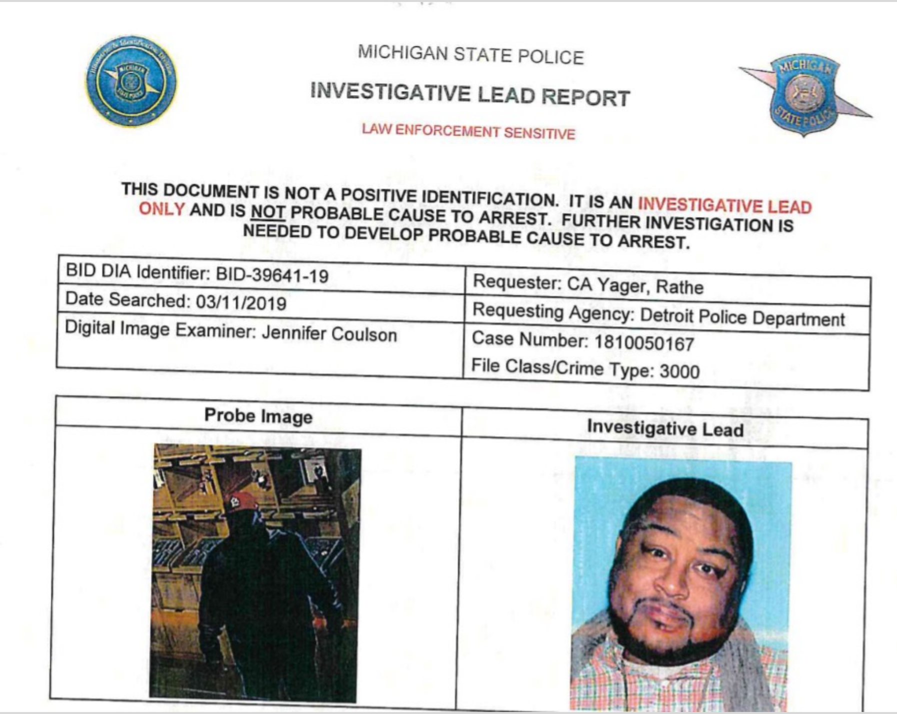
Data bias and ethics



Doctor walks in a bar

Nurse walks in a bar

Data bias and ethics



MICHIGAN STATE POLICE

INVESTIGATIVE LEAD REPORT

LAW ENFORCEMENT SENSITIVE

THIS DOCUMENT IS NOT A POSITIVE IDENTIFICATION. IT IS AN INVESTIGATIVE LEAD ONLY AND IS NOT PROBABLE CAUSE TO ARREST. FURTHER INVESTIGATION IS NEEDED TO DEVELOP PROBABLE CAUSE TO ARREST.

BID DIA Identifier: BID-39641-19	Requester: CA Yager, Rathe
Date Searched: 03/11/2019	Requesting Agency: Detroit Police Department
Digital Image Examiner: Jennifer Coulson	Case Number: 1810050167 File Class/Crime Type: 3000

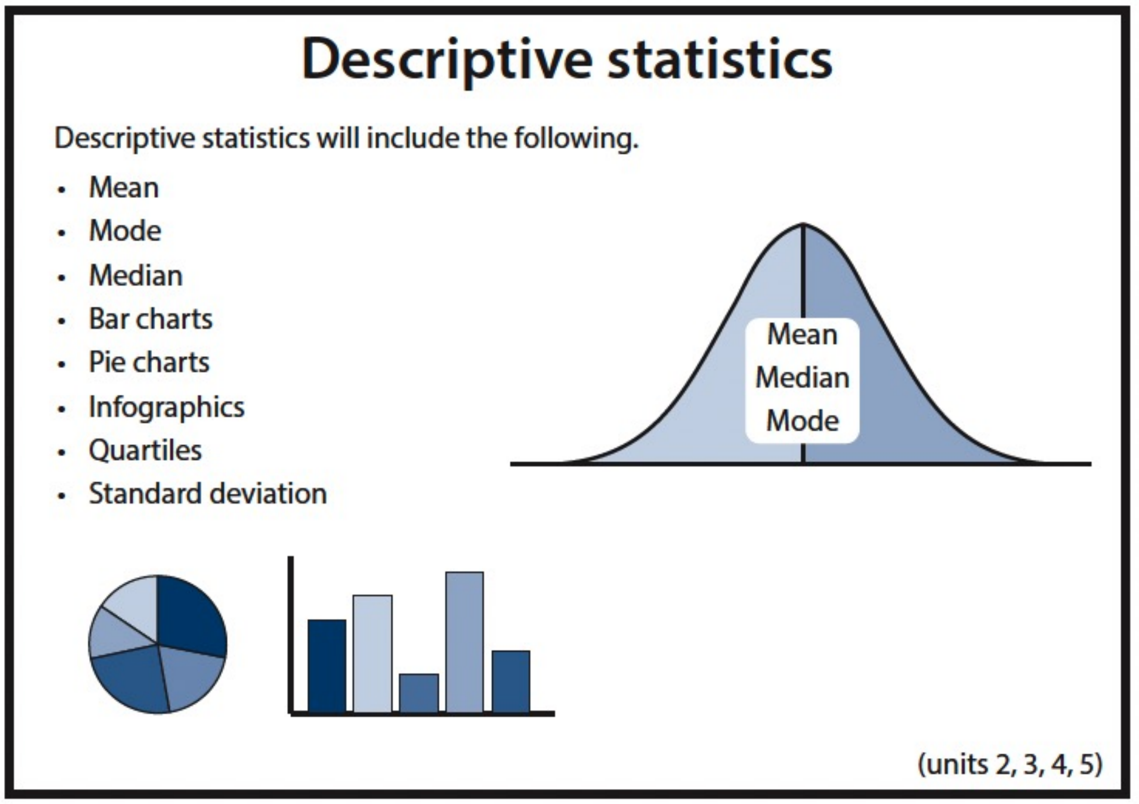
Probe Image	Investigative Lead

Understanding data

Visualization

Understanding data

Analysis



Descriptive statistics

Descriptive statistics will include the following.

- Mean
- Mode
- Median
- Bar charts
- Pie charts
- Infographics
- Quartiles
- Standard deviation

Mean
Median
Mode

(units 2, 3, 4, 5)

Learning from the data

What is ‘learning’?

“… learning is about predicting the future based on the past.”

-- CIML Book

Past: Training data

Fully labeled, unsupervised, semi-supervised

Classification, regression

Predicting: Model

Linear model, kernel method, nearest neighbor, decision trees, neural networks

Future: Testing data

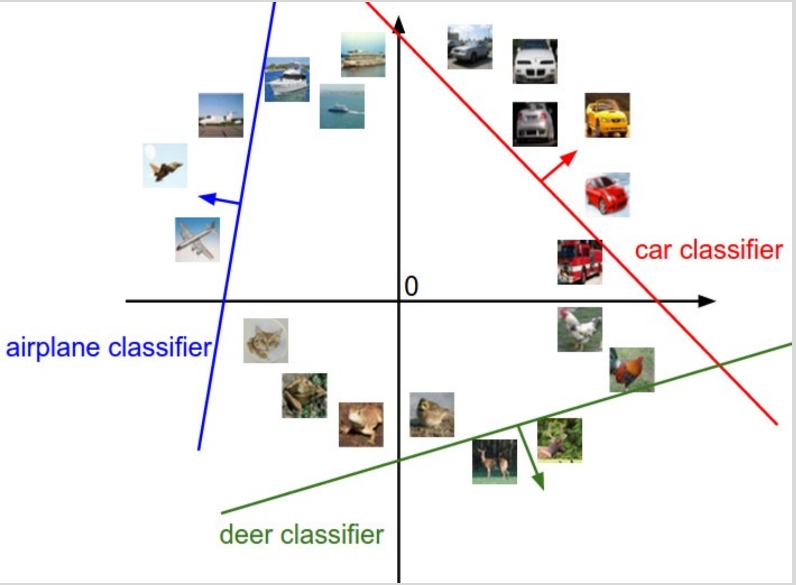
Out-of-distribution, domain shift, life-long learning

Evaluation Task-specific metrics, user studies

Learning from the data

- Classification

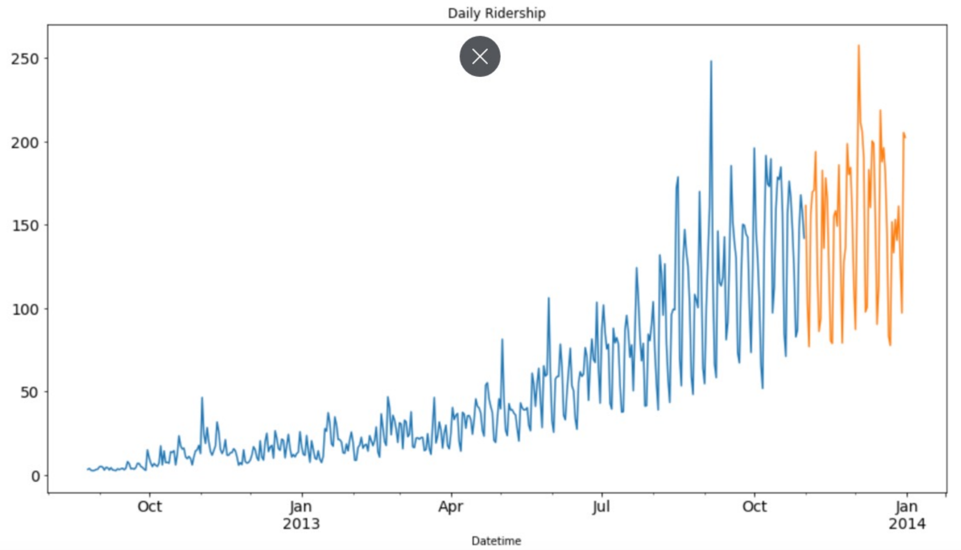
  e.g., rain or no rain? Cat or dog?



Learning from the data

- Regression

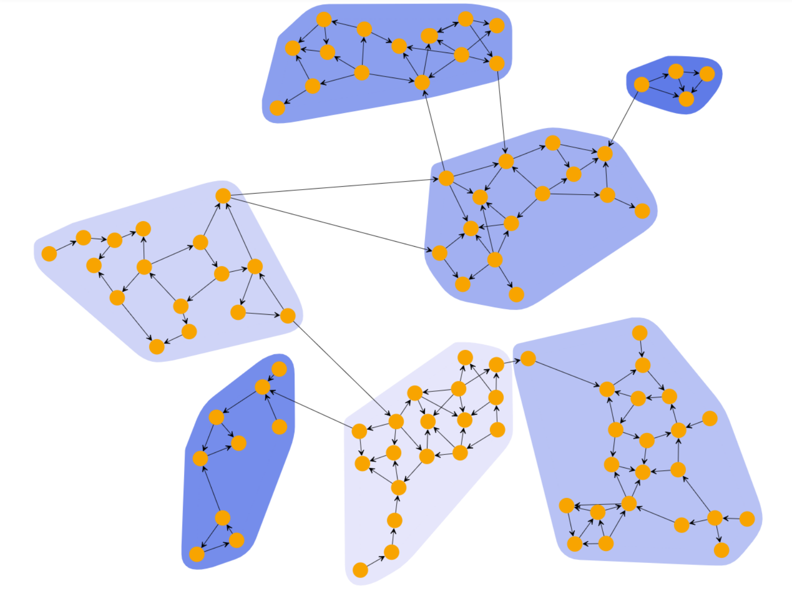
e.g., what is the stock price for TSLA tomorrow?



Learning from the data

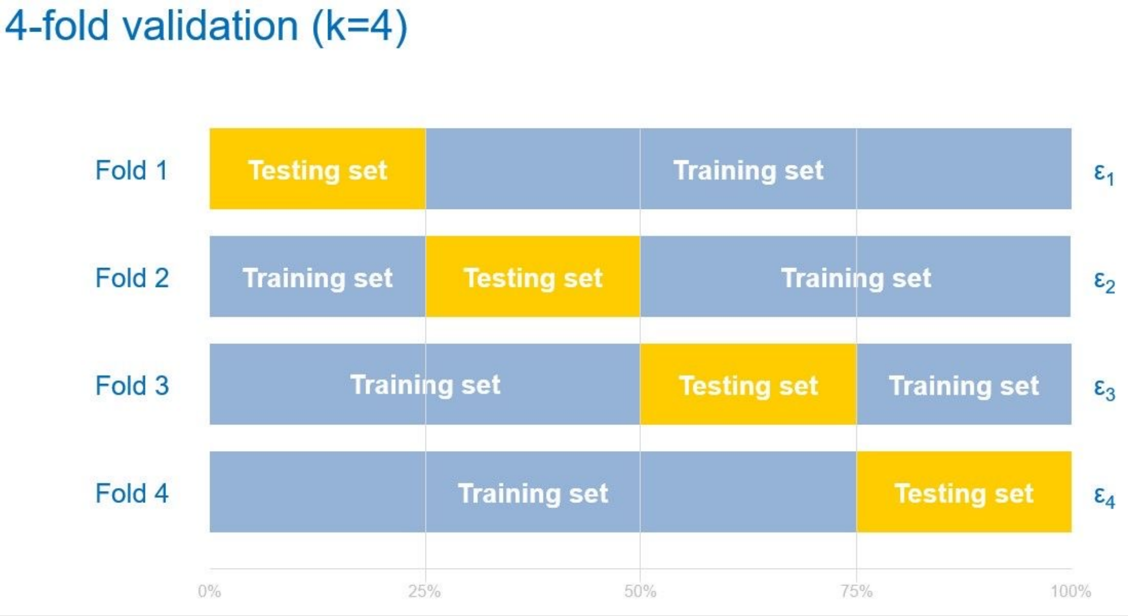
- Unsupervised learning

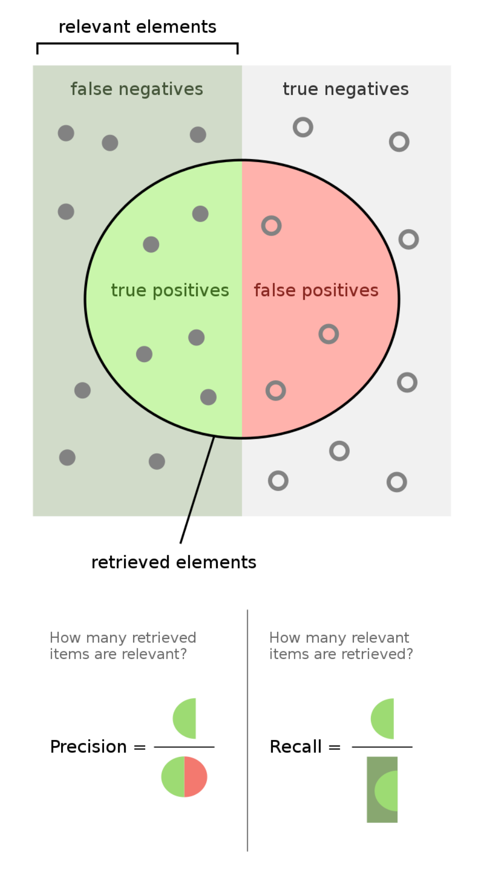
e.g., how many sub-communities exist in this social network?



Learning from the data

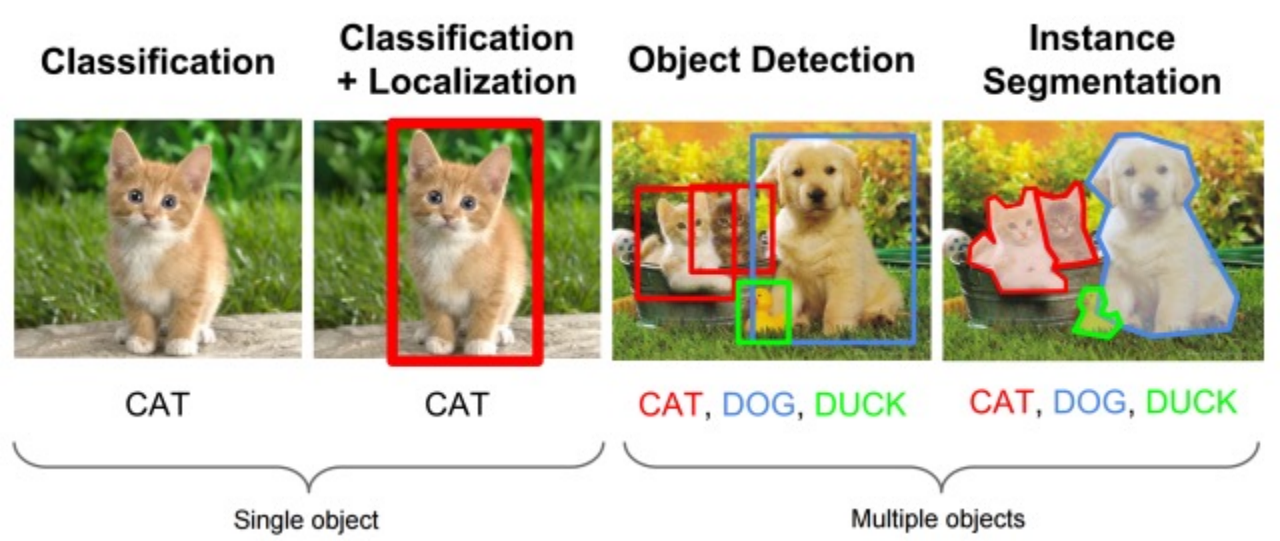
Evaluation and metrics





Applications and practices

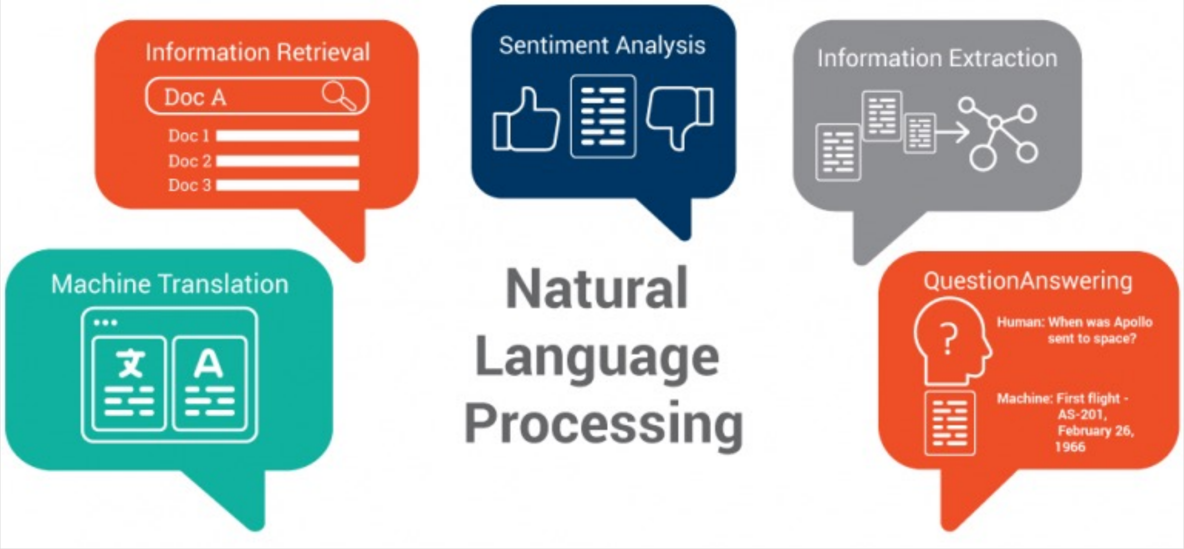
Vision



Cool things in computer vision now: text-to-image, 2d-to-3d, text-to-video, 3d reconstruction

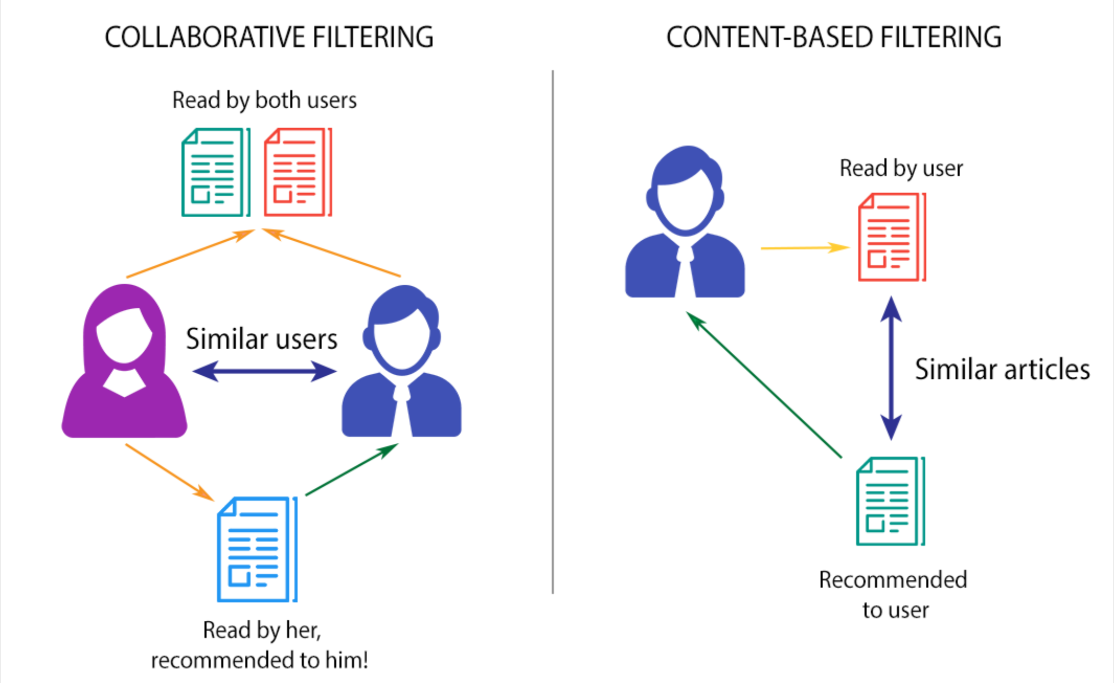
Applications and practices

Languages



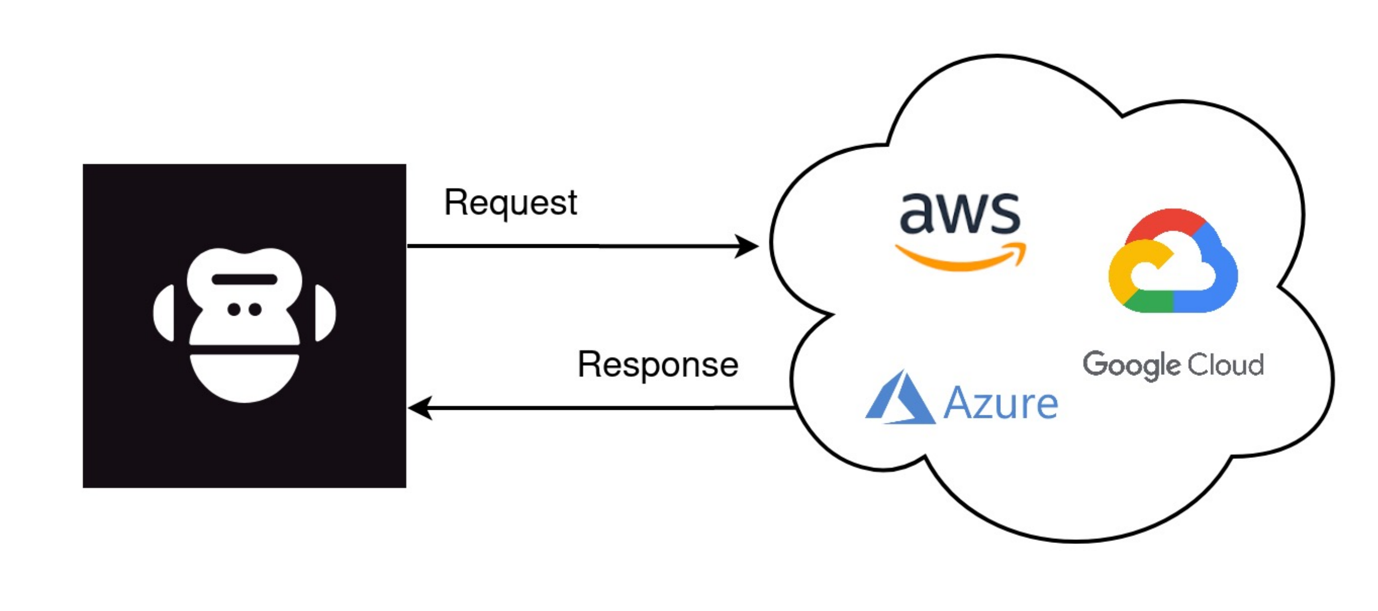
Applications and practices

Recommendation systems



Applications and practices

Real-world ready development



Outline

- Background
- Class logistics and policy
- Class topics
- Q&A

What’s next?

We will walk through example data science project pipelines and work on your first Kaggle challenge.

Start to think about your project and team!