COM SCI M148:
Introduction to Data Science

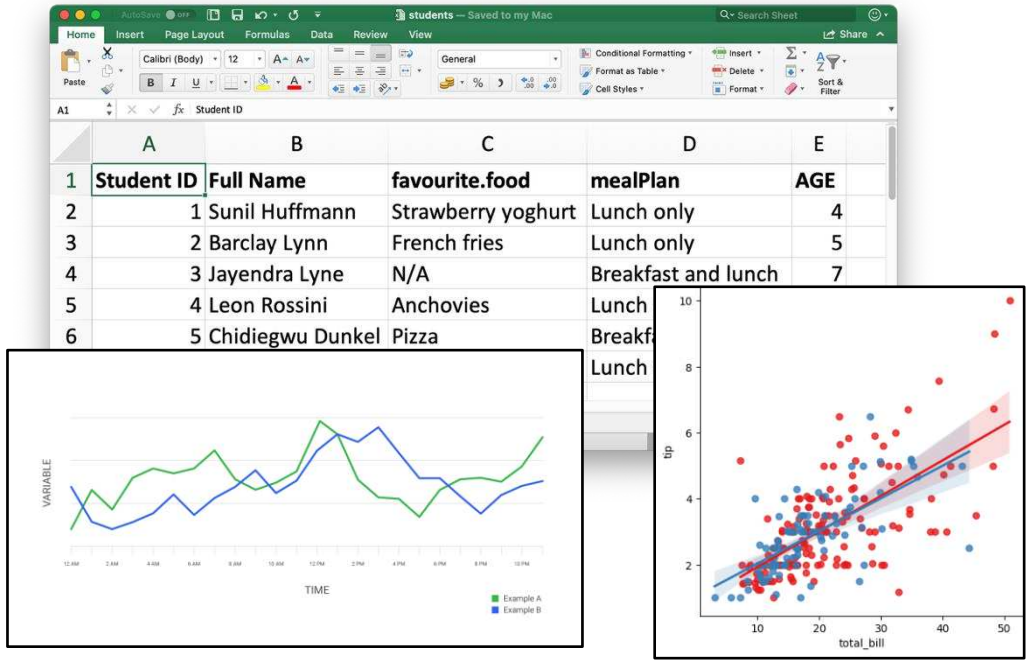
Prof. Nathan Szymanski
nszymanski@seas.ucla.edu
Department of Materials Science

UCLA

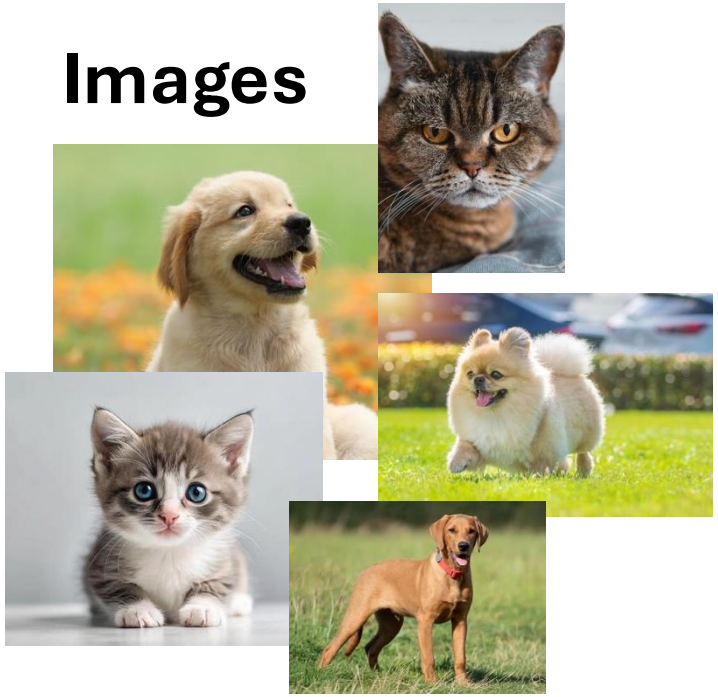
Samueli
School of Engineering

What is data…And why do we care about it?

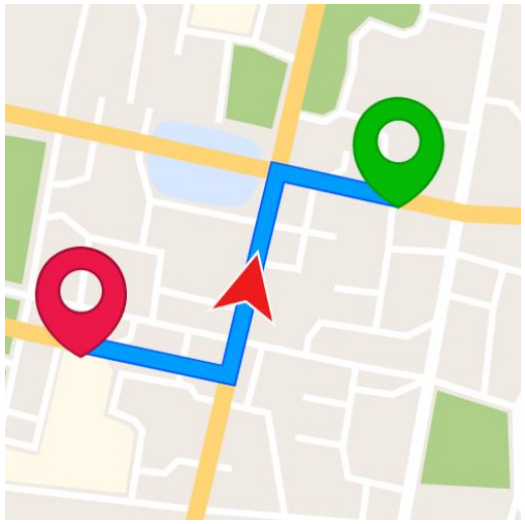
Spreadsheets & plots



Images



Graphs & maps



"Structured" data

Most data is unstructured



Natural language

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Audio

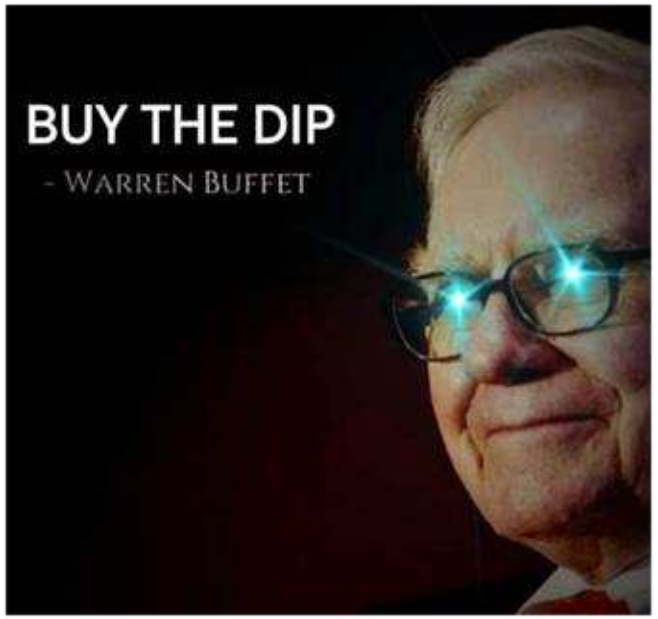


What is data…And why do we care about it?

Data → **Insights** → **Actions**



"It will keep going up!"



The role of the *data scientist*







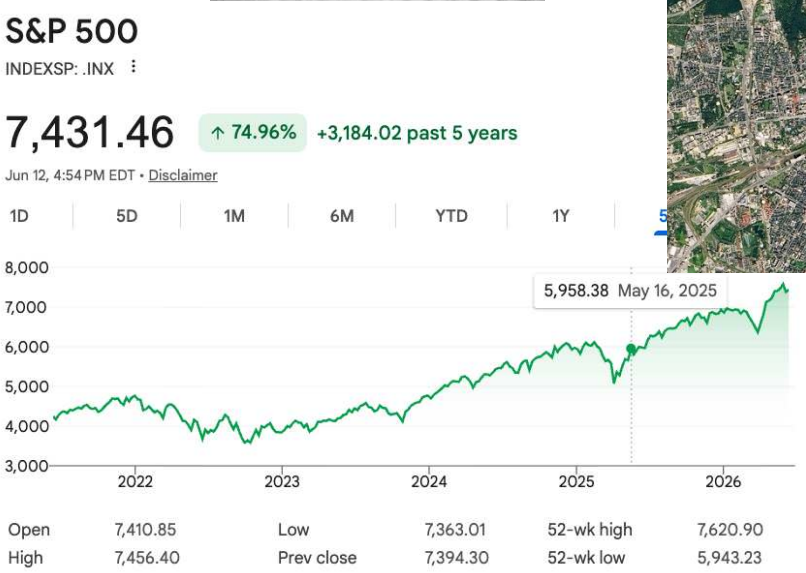




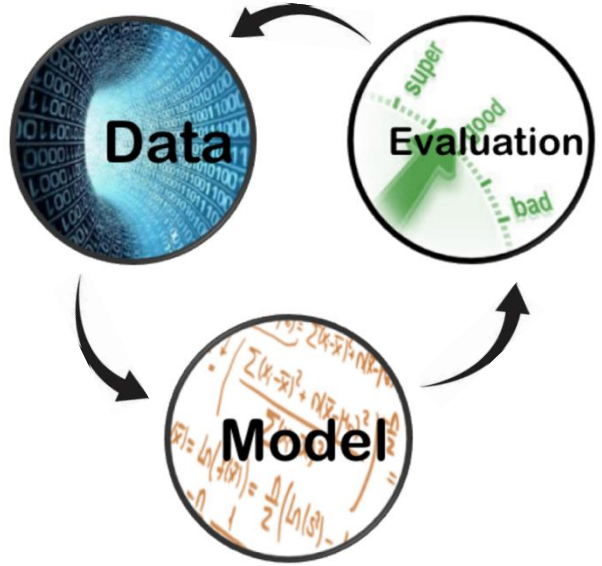
Ingest all the right data to make an **informed** decision



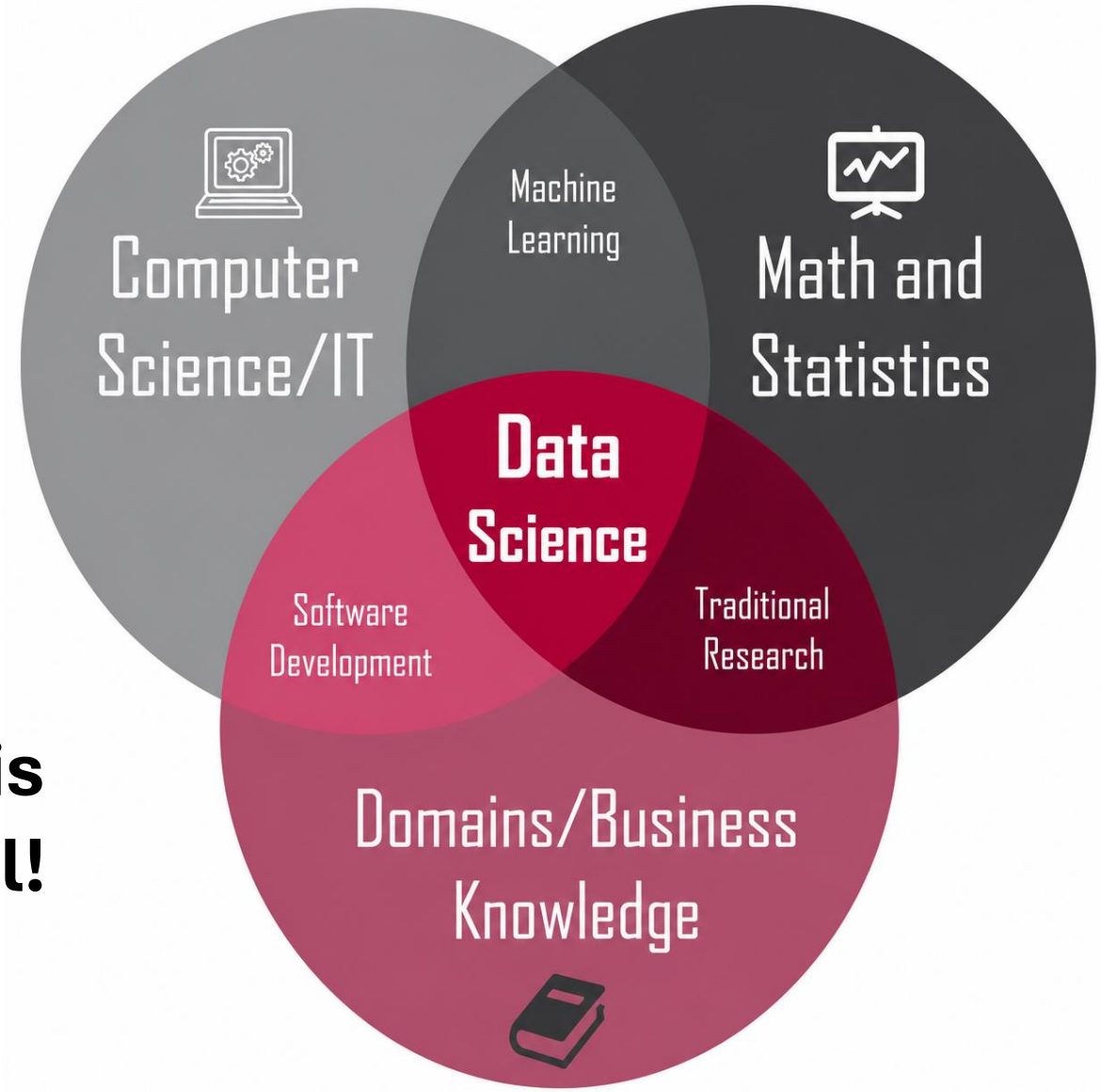
“There is a 72% chance this price will increase over the next month”





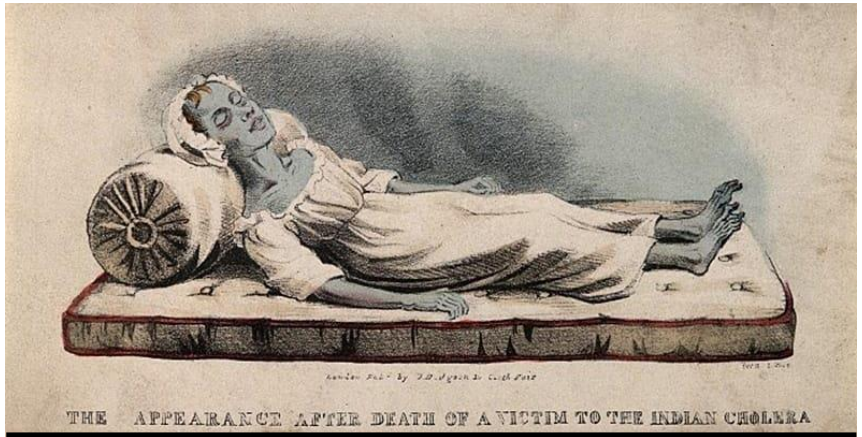


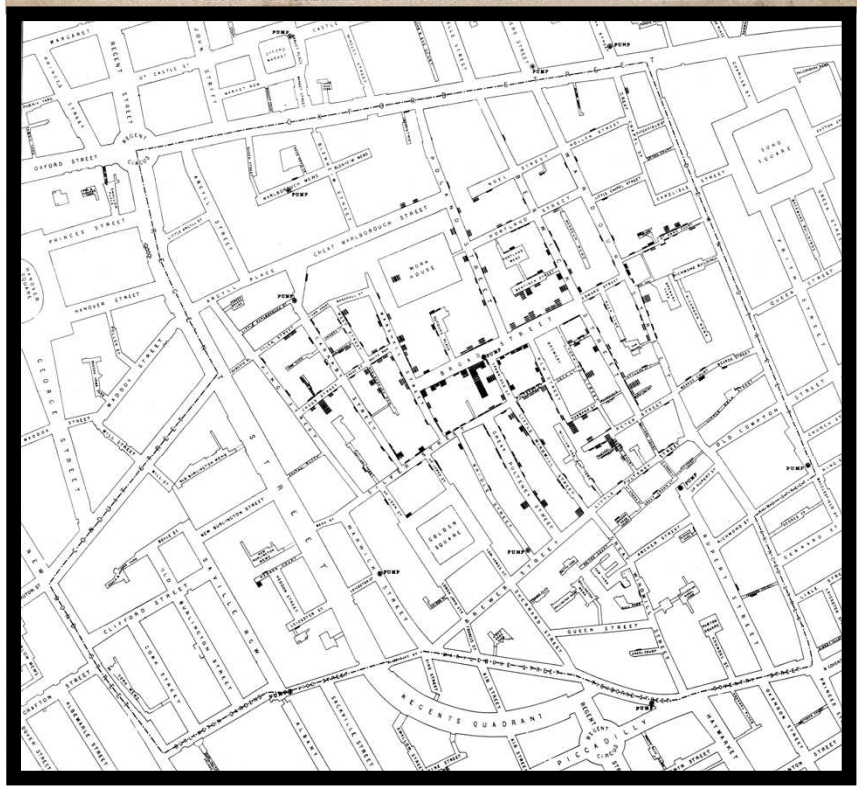
The role of the *data scientist*



Data science is highly intersectional!

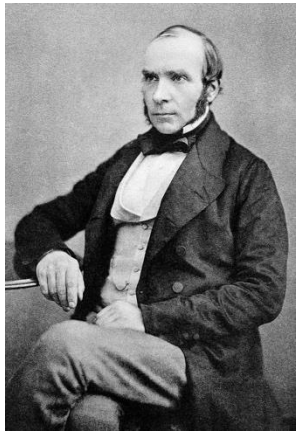
Data science has been around a long time





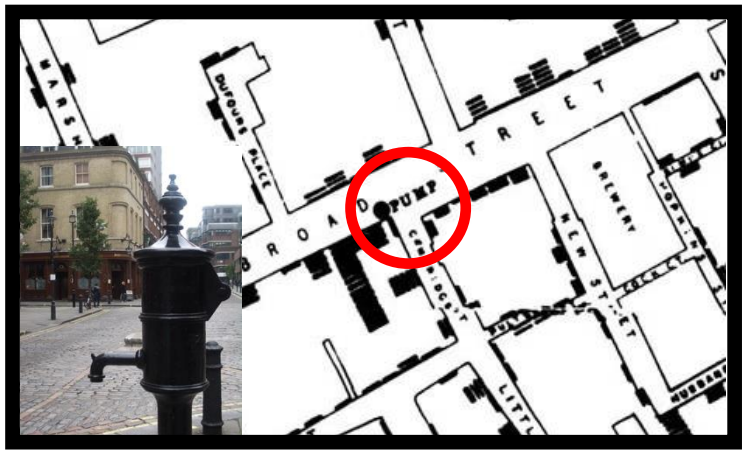
Early examples of data science:

- In 1854, there was a cholera epidemic in London
- A physician (John Snow) asked: "what's causing the outbreak?"
- He visualized the cases using bar plots on a map → the outbreak centered around just one water pump!
- After removing it, the outbreak ended ☺

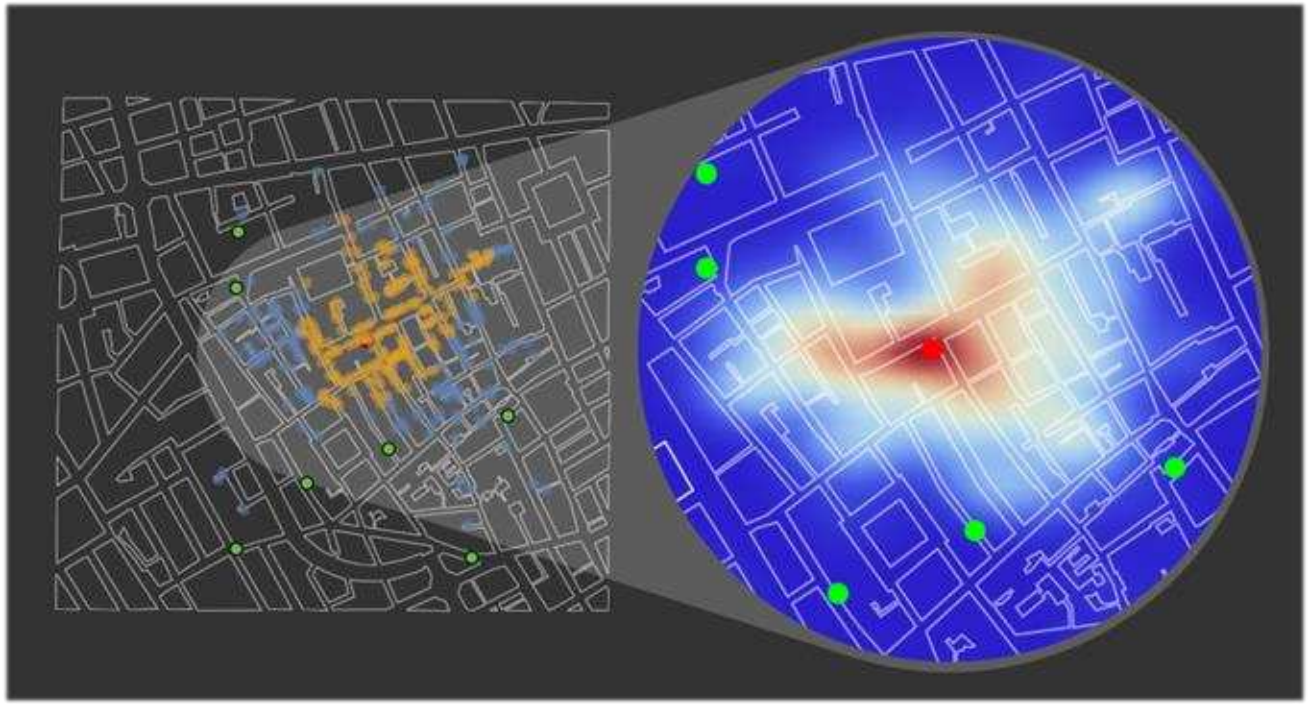


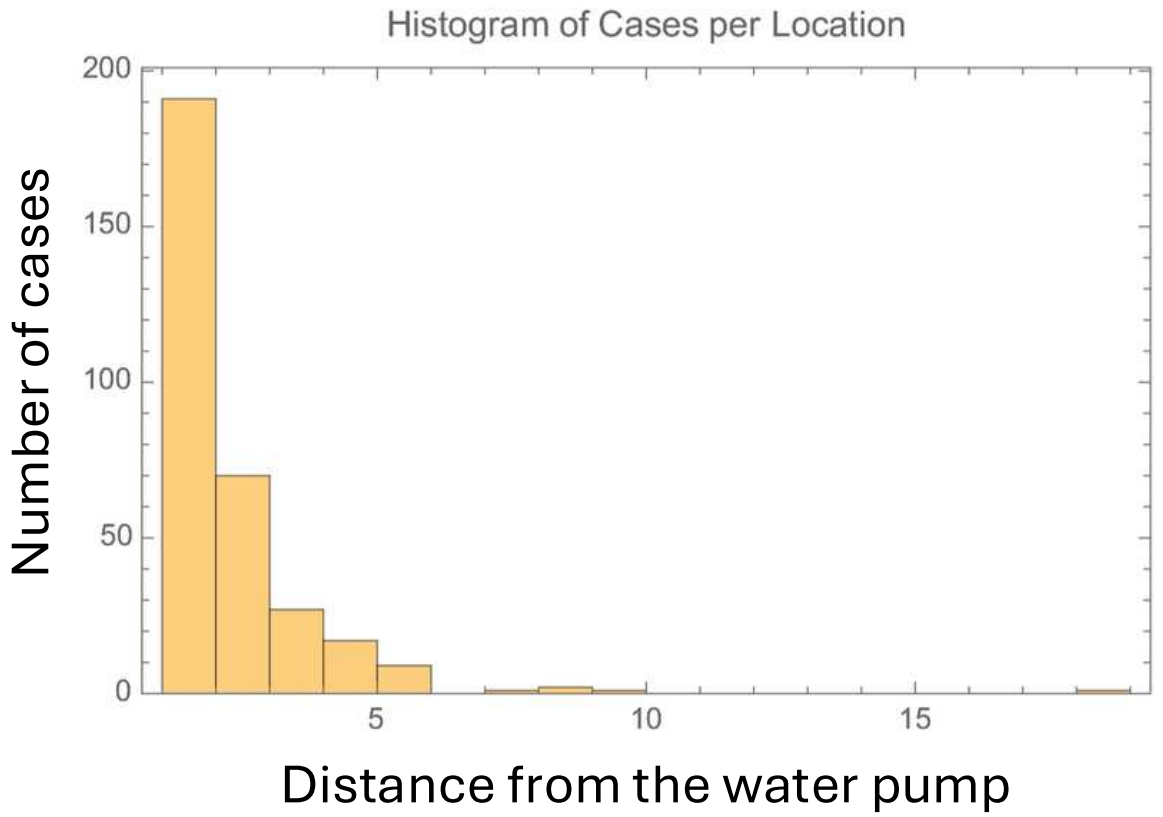


Not this one



Data science has been around a long time





...There are certainly better ways to **visualize the data**...
But John Snow did a nice job for his time

Data science has been around a long time

H 1																
Li 7	Be 9											B 11	C 12	N 14	O 16	F 19
Na 23	Mg 1.2											Al 27	Si 28	P 31	S 32	Cl 35
K 39	Ca 40		Ti 48	V 51	Cr 52	Mn 55	Fe 56	Co 59	Ni 59	Cu 63	Zn 65			As 75	Se 78	Br 79
Rb 85	Sr 87	Y 88	Zr 90	Nb 94	Mo 96		Ru. 104	Rh 104	Pd 106	Ag 108	Cd 112	In 113	Sn 118	Sb 122	Te 125	I 127
Cs 133	Ba 137	La 180		Ta 182	W 186		Os 195	Ir 197	Pt 198	Au 199	Hg 200	Tl 204	Pb 207	Bi 210		

Missing elements

Di 138	Ce 140												Er 178			
	Th 231		U 240													

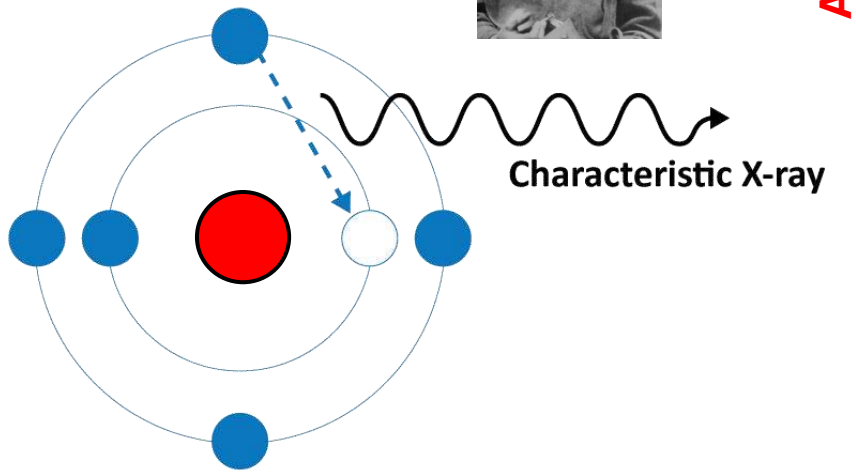
In the early 1900s, the periodic table was organized by relative mass – ***m***, not ***Z*** – and had many missing elements

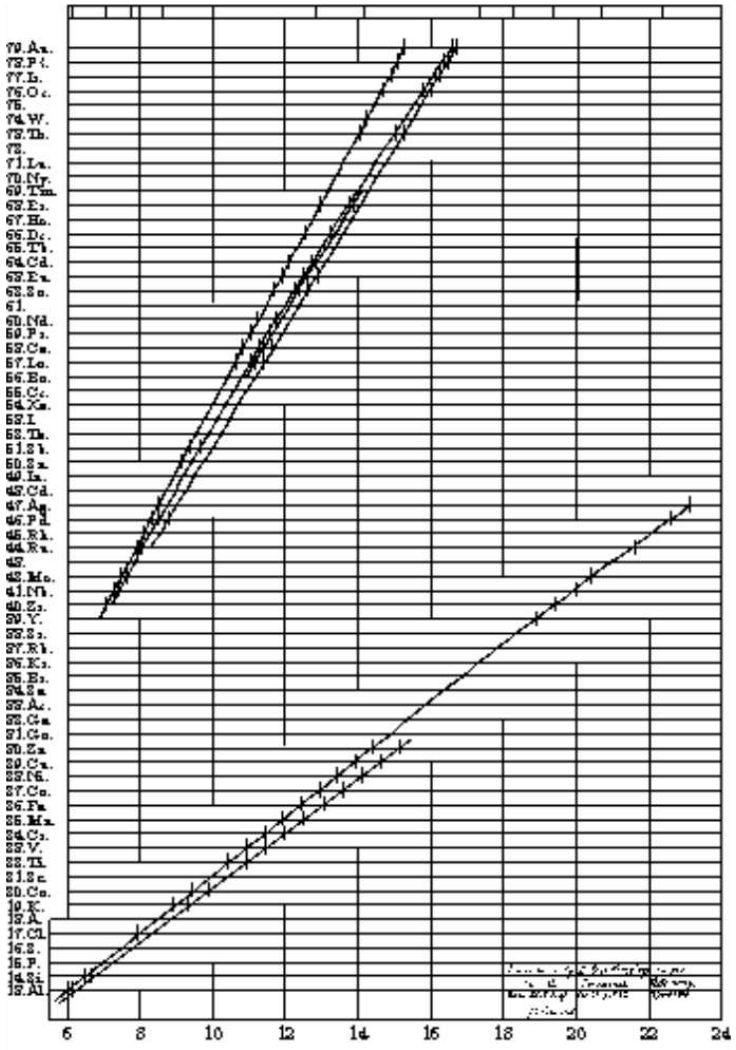


Henry Moseley (1913) found that X-rays emitted by atoms followed a trend:

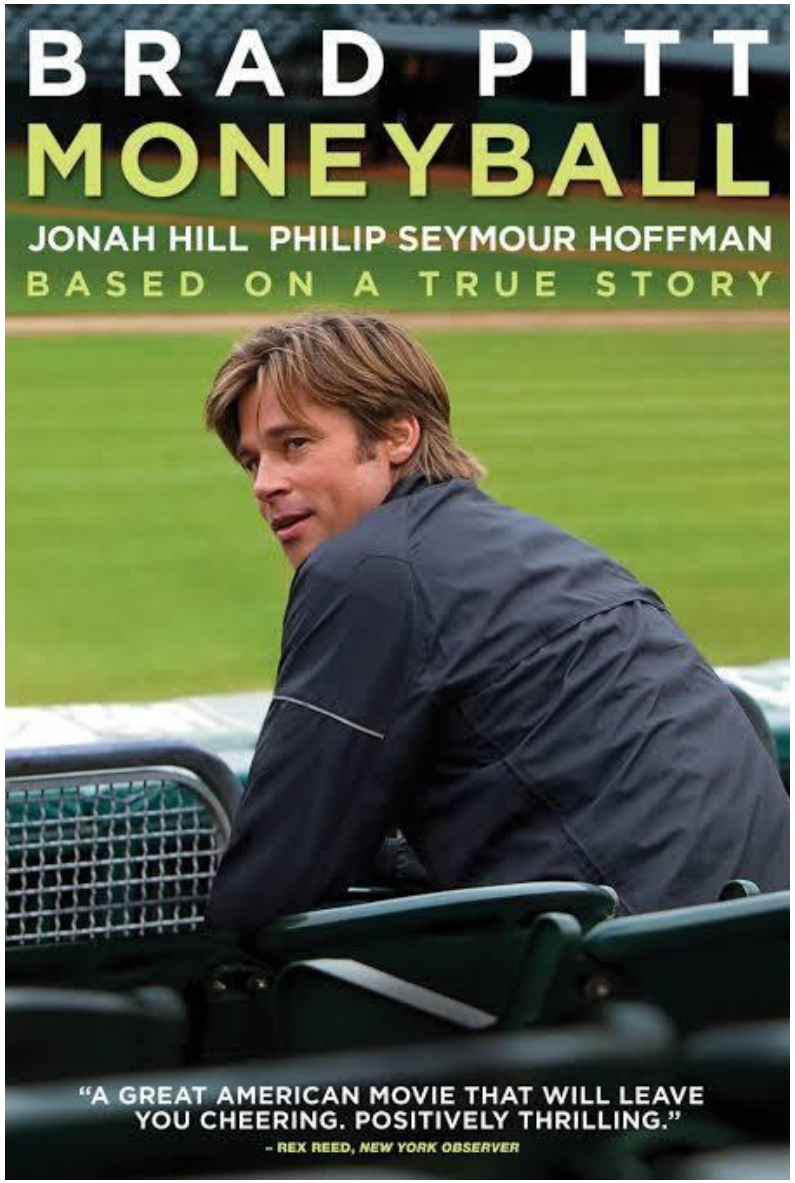
$$\sqrt{\nu} \propto Z$$

Z is the nuclear charge



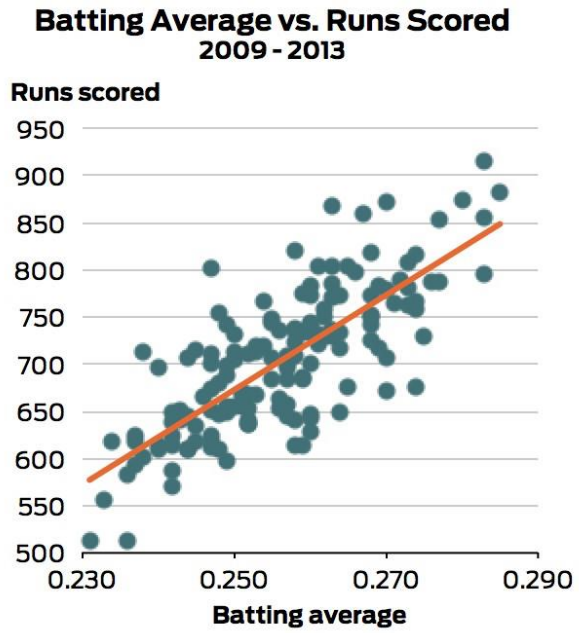


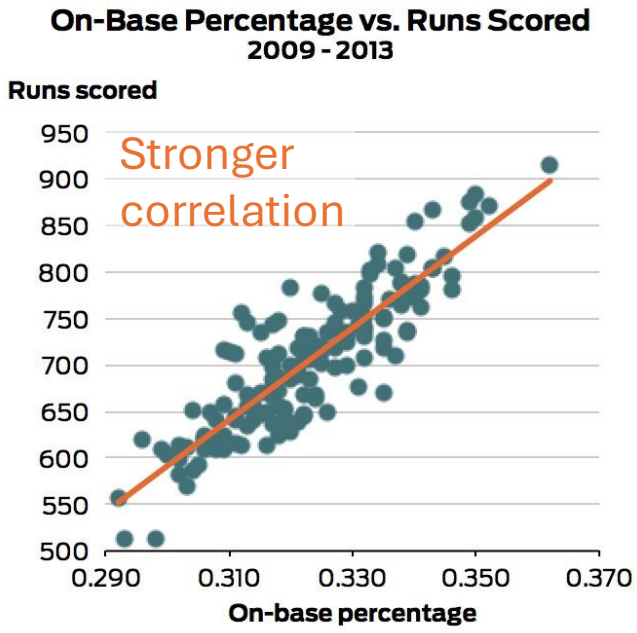
More recent examples of data science





- Oakland A's had one of the lowest budgets in the MLB
- Billy Beane used statistics to identify "cheap" players with good on-base %

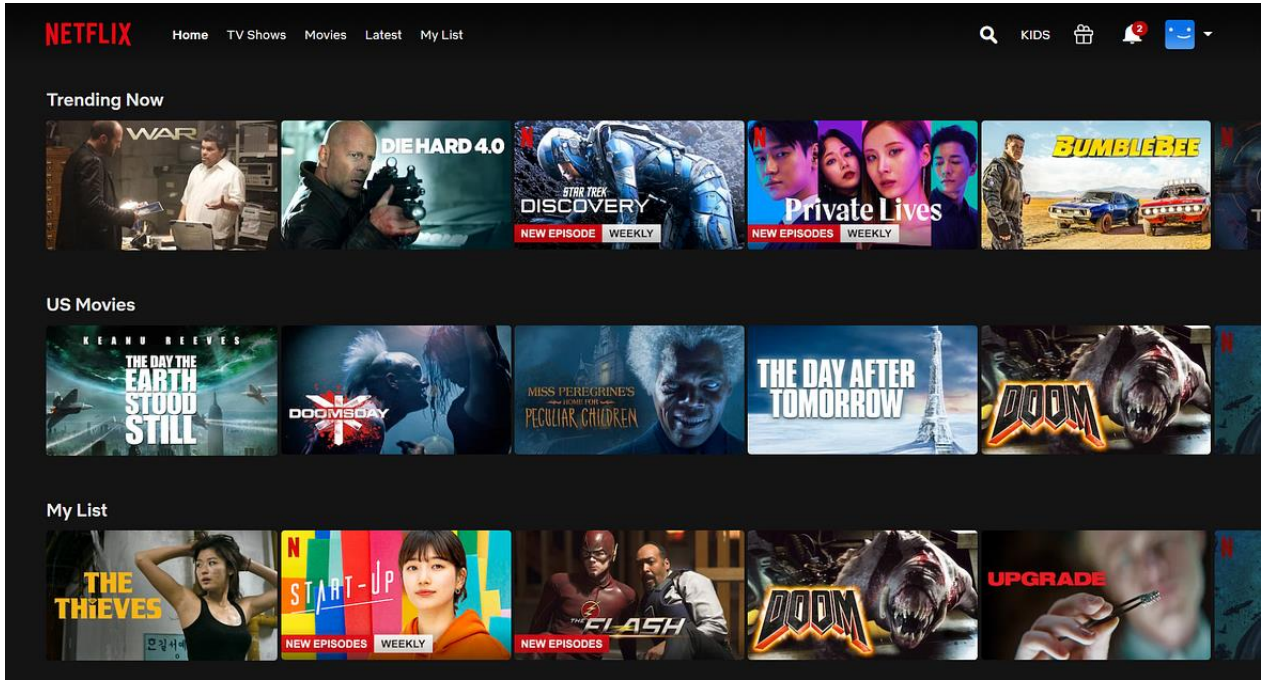




Result:
20-game win streak!

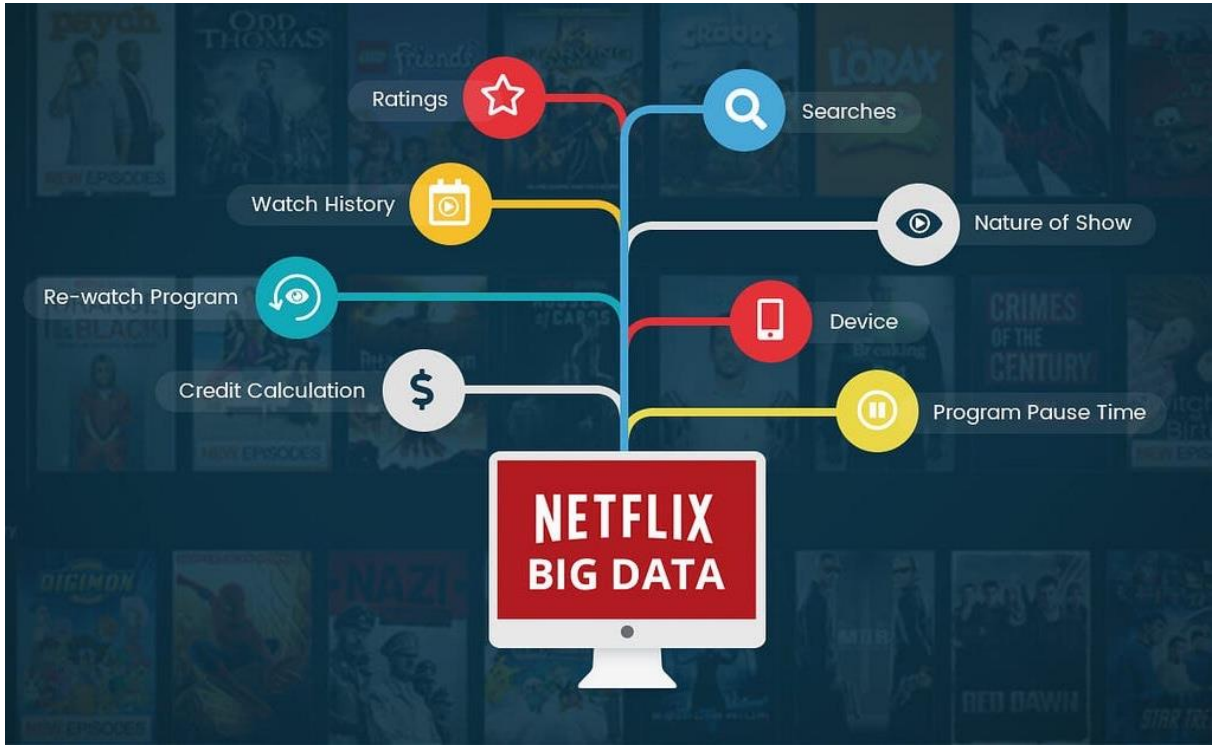
Today all teams have a major data analytics department

More recent examples of data science



Recommendation systems for Netflix, Spotify, YouTube, *etc.*

Recommended content accounts for **80% of all views/listens!** Big $$$



"People who liked this also liked…" (based on ratings, view time, etc.)

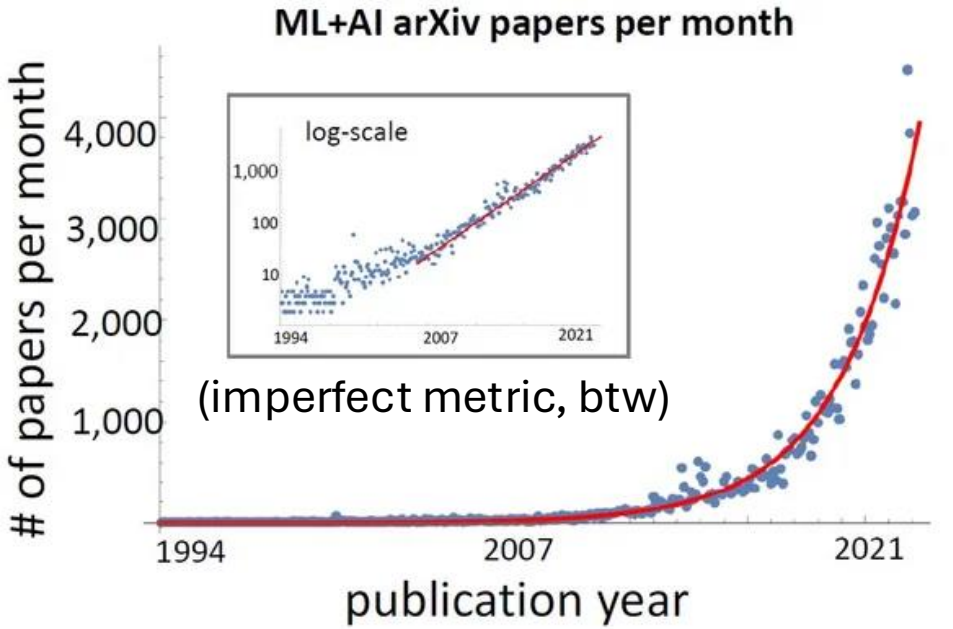
This is a highly complicated graph with up to *millions* of pieces of content

Enter: *Machine learning* and *AI*

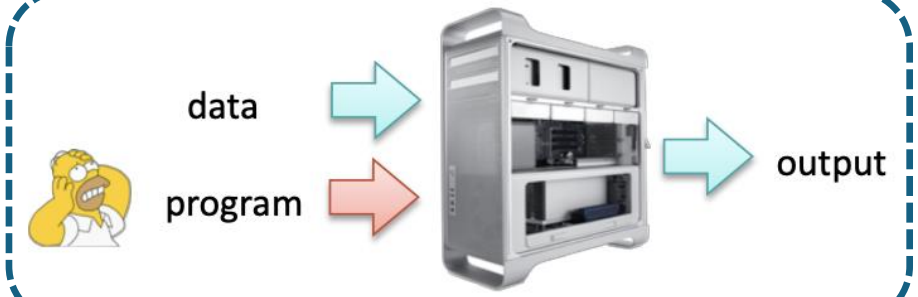
Some Studies in Machine Learning (1959)
Using the Game of Checkers

A. L. Samuel

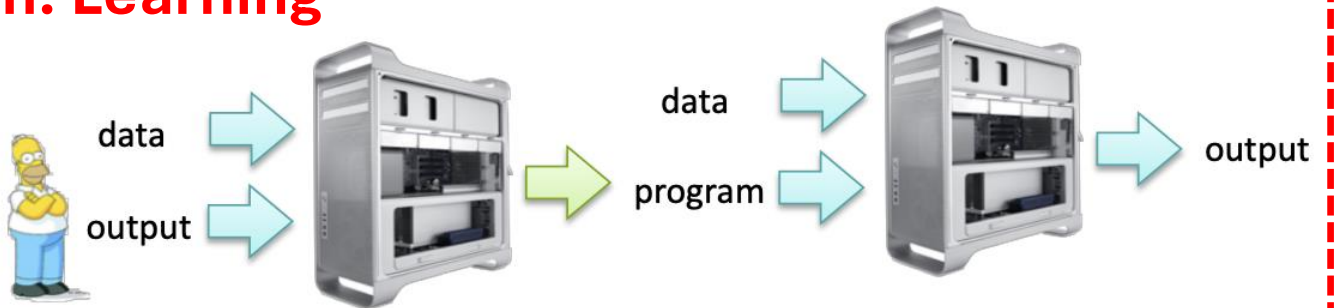
Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.



Traditional CS

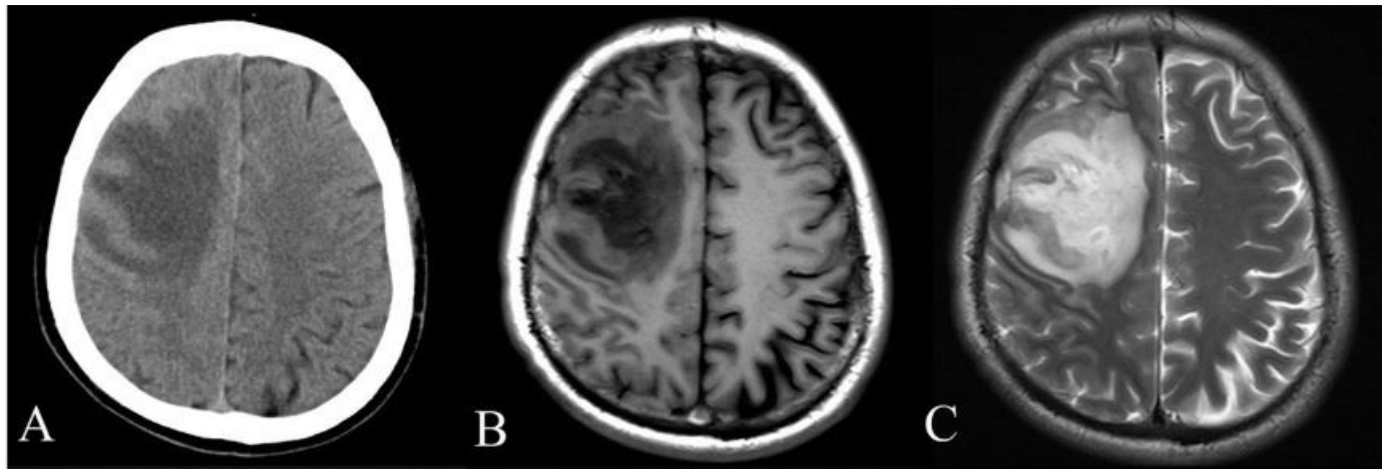


Mach. Learning



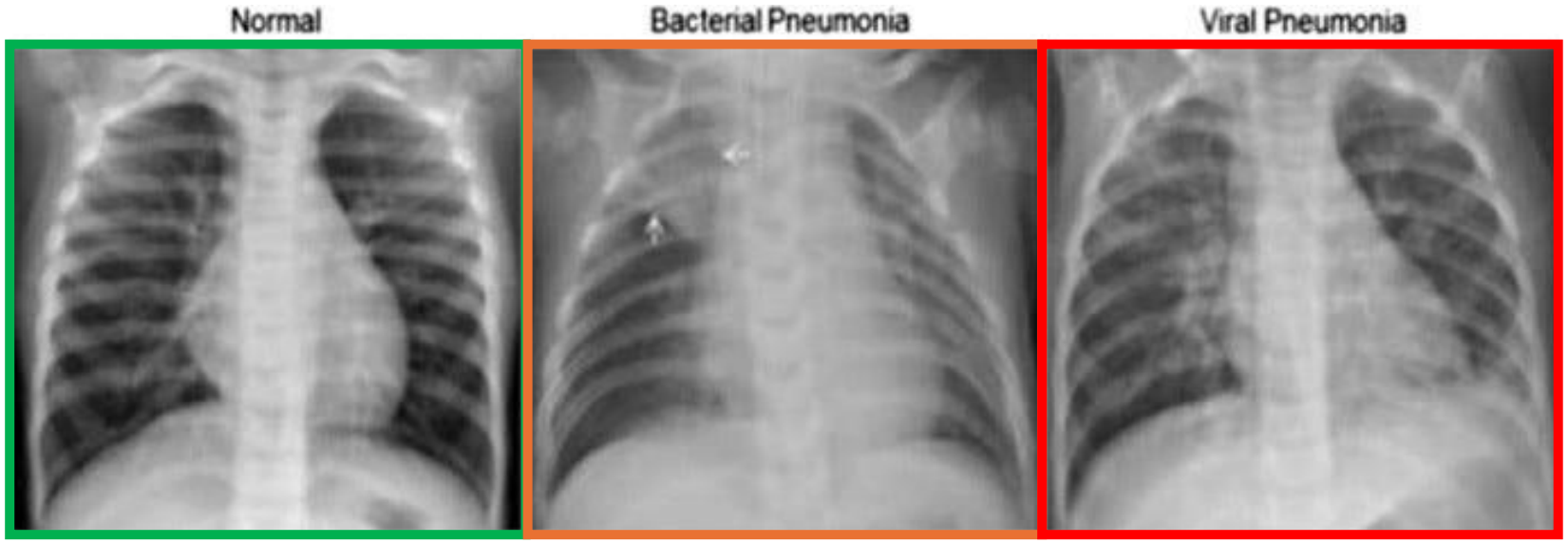
Krenn *et al.*, https://arxiv.org/abs/2210.00881 (2022)

Applications of Machine Learning (ML)

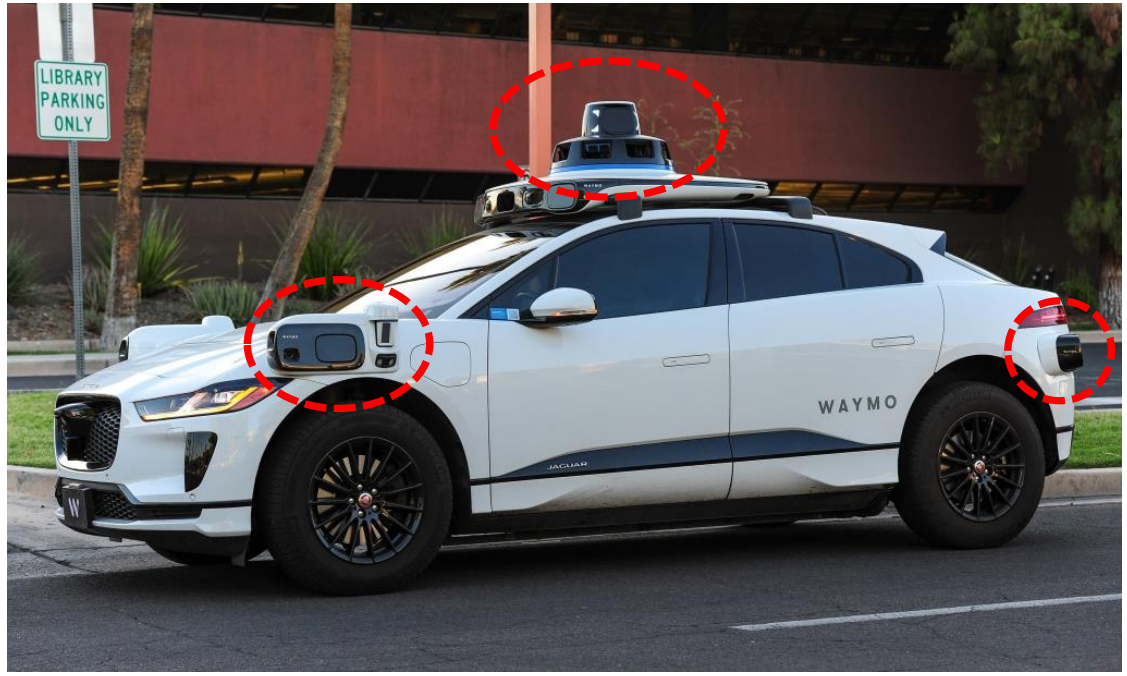


ML models are now widely used to do early detection of diseases from medical images (X-rays, CT scans, MRIs)

These models are trained on very **large datasets with expert labeling**



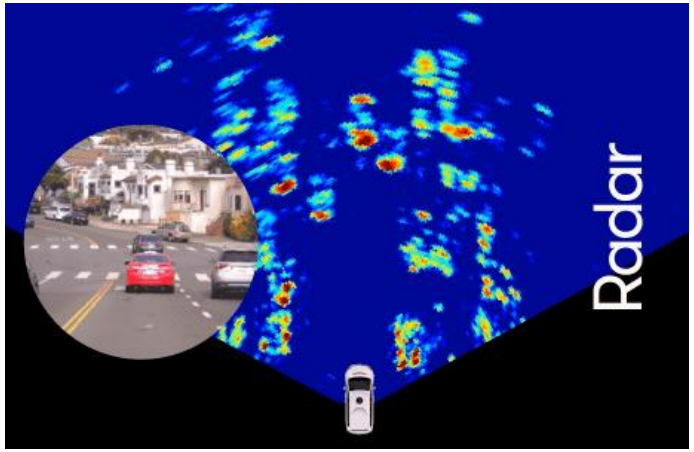
Applications of Machine Learning (ML)

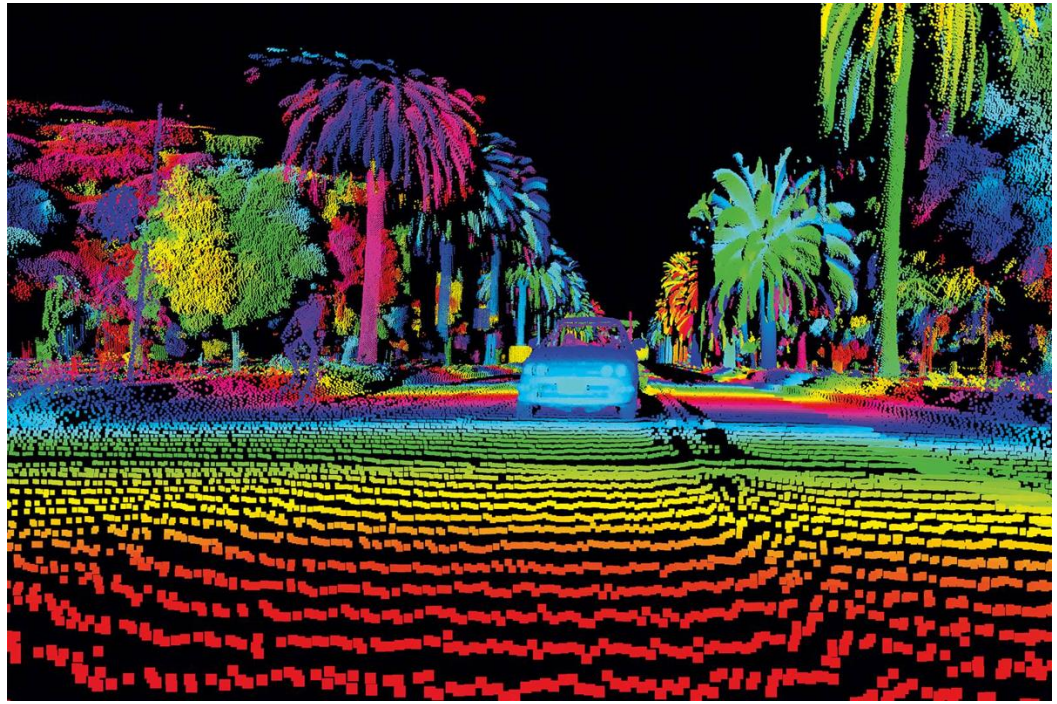


WAYMO

Self-driving vehicles:
Given input describing its surroundings, the car must decide what to do

Uses a combination of input data based on: Lidar, radar, and cameras

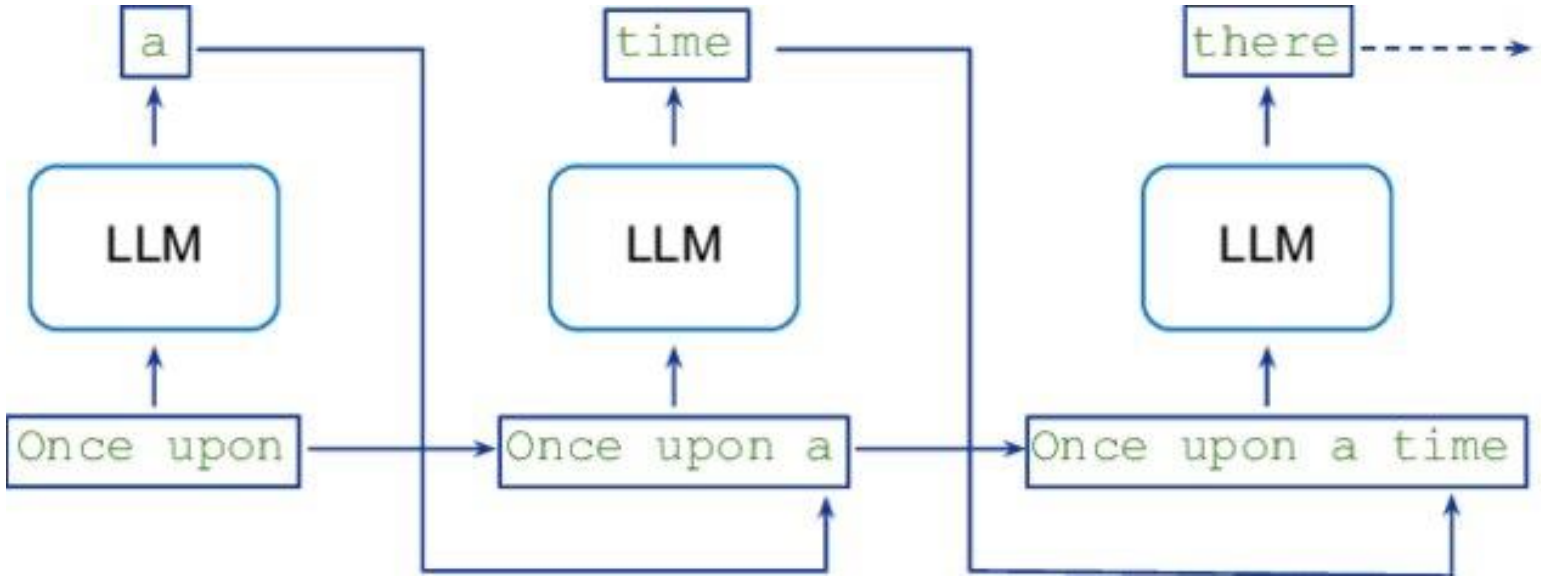




Applications of Machine Learning (ML)

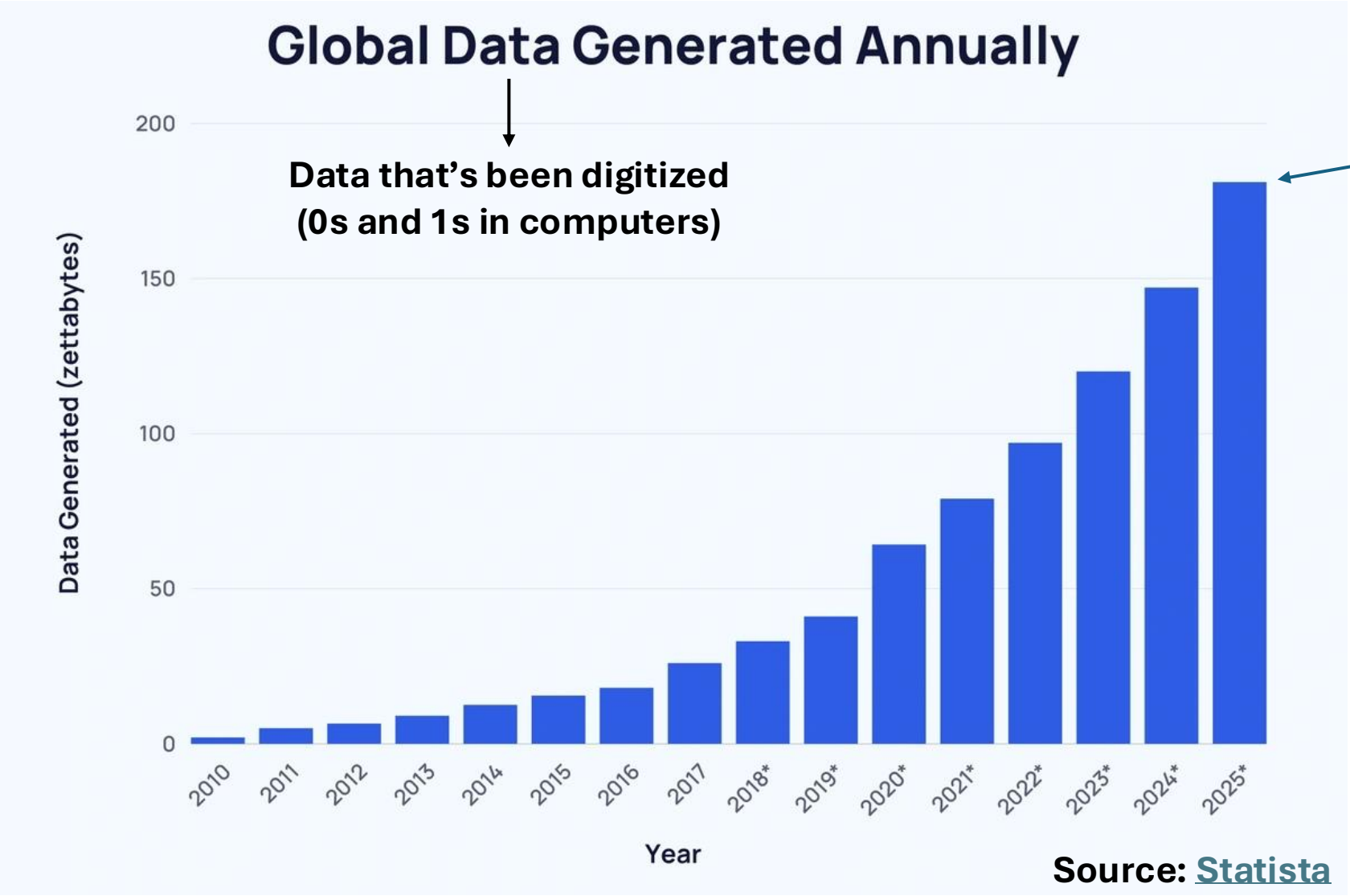
Claude ChatGPT Gemini

Large language models (LLMs) are essentially just **autoregressive models** trained on ***huge* amounts of text** (and lately, other data like images, *etc.*)



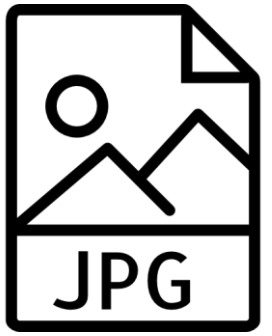
Given some words or phrases (tokens), the LLM predicts what text most likely comes next

There is *a lot* of data out there



Almost 200 zettabytes of *new* data in 1 year!

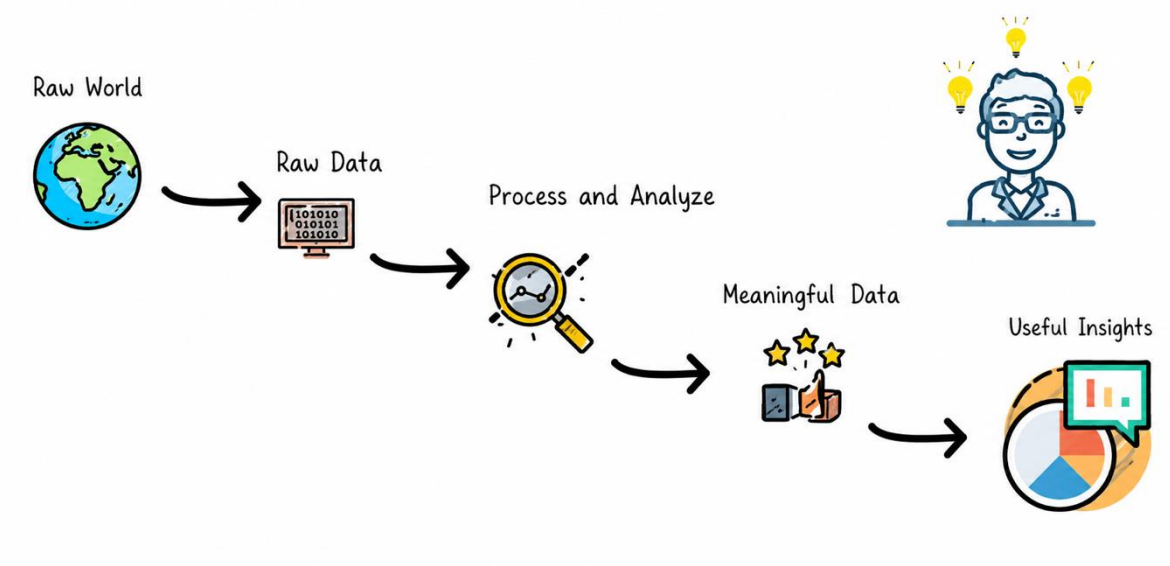
For reference:
1 zettabyte is equal to
1 billion (10^9) TB,
or 10^{15} Megabytes



A typical photo is 5-10 MB

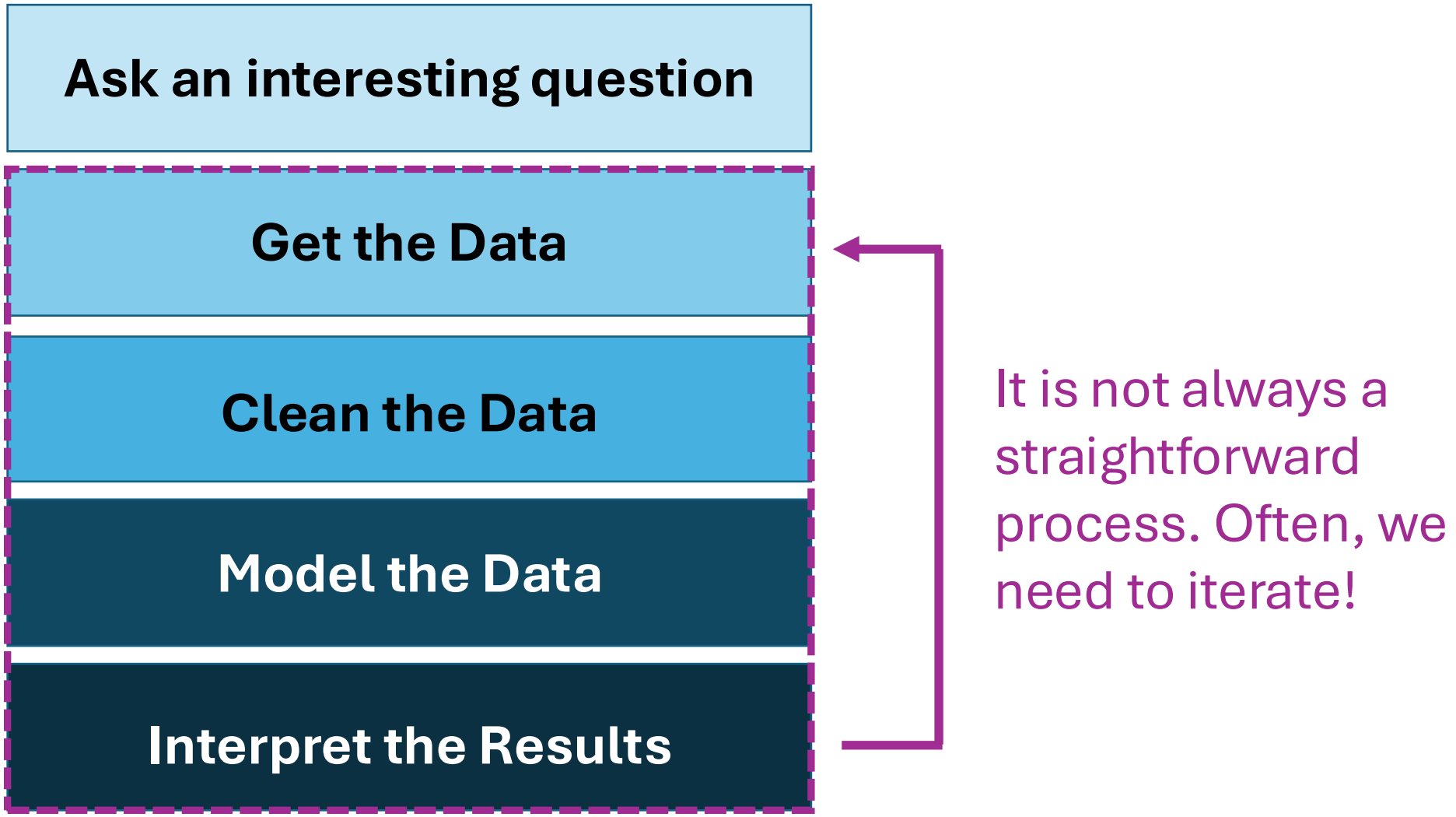
Data science is a step-by-step process

- Ask an interesting question
- Get the Data
- Clean the Data
- Model the Data
- Interpret the Results



Basically, everything you learn in this course will fall under one of these steps

Data science is a step-by-step process



Step 1: Define the problem

Ask an interesting question

Get the Data

Clean the Data

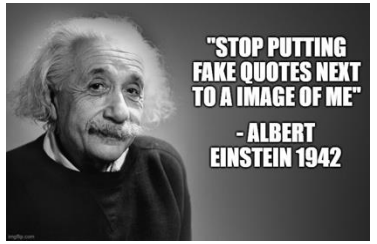
Model the Data

Interpret the Results

"If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions." - Albert Einstein*



*As with most Einstein quotes, this is unverified...



Every downstream step depends on asking the right question!

Step 1: Define the problem

Ask an interesting question

Get the Data

Clean the Data

Model the Data

Interpret the Results

A good data science question is…

- **Specific:** it is clear what we are predicting or estimating
- **Measurable:** we can rigorously define success with a number
- **Answerable with data:** the necessary data exists or can be collected
- **Actionable:** someone can and will do something with the findings

Examples of good *vs.* poorly-framed questions

✗ Vague

a) "How can I make my business more profitable?"

b) "How can we reduce crime?"

c) "Why does iron discolor?"

✓ Well-framed

a) "How can I increase my profits by 20% in 6 months by varying the prices of items *X*, *Y*, and *Z*?"

b) "Which neighborhoods in LA have experienced the largest changes in car thefts over the past year?"

c) "What environmental factors most directly correlate with the rate of iron corrosion?"

Step 2: Collect the *right* data for the problem

Ask an interesting question

Get the Data

Clean the Data

Model the Data

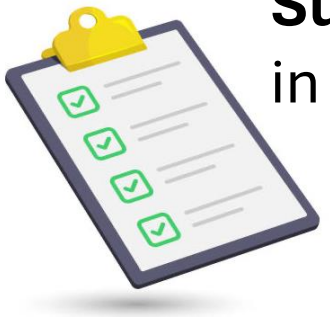
Interpret the Results

No matter what your initial question is, the *next* question should always be:

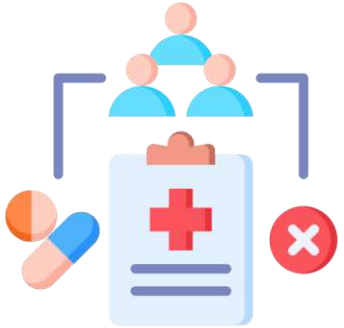
What data do I need to answer this question? And where can I get it?

Sources of data: Primary

Primary data is first-hand information you directly collect yourself



Surveys: common in social sciences

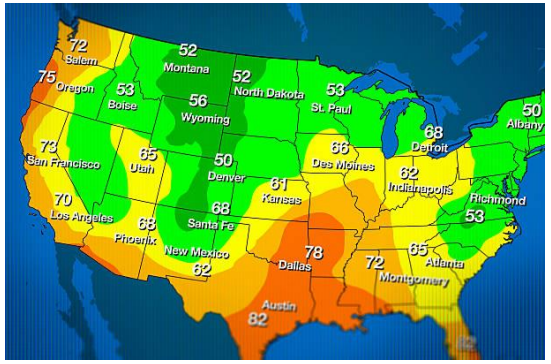


Clinical trials for medicine



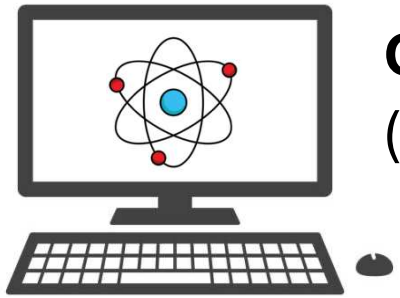


Physical measurements in the sciences





Sensor data (*e.g.*, temperature and satellite imagery)



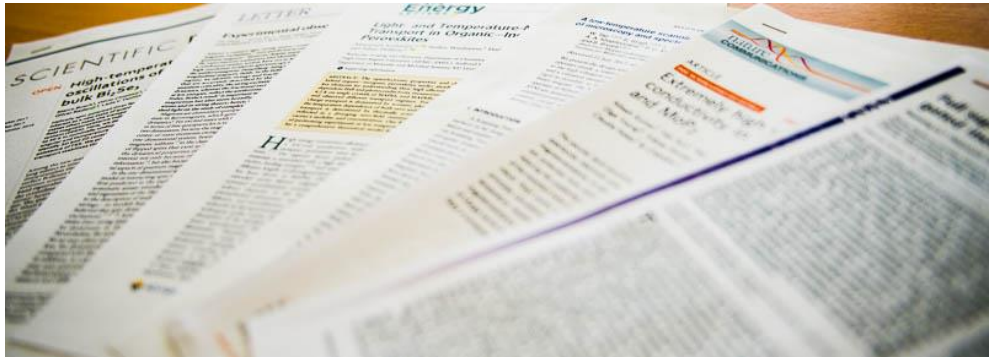
Computer simulations (*e.g.*, quantum chemistry)

Sources of data: Secondary

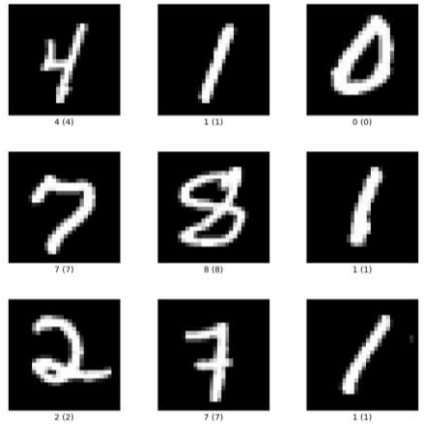
Secondary data is already-available information collected by someone else

Government datasets

Open scientific datasets

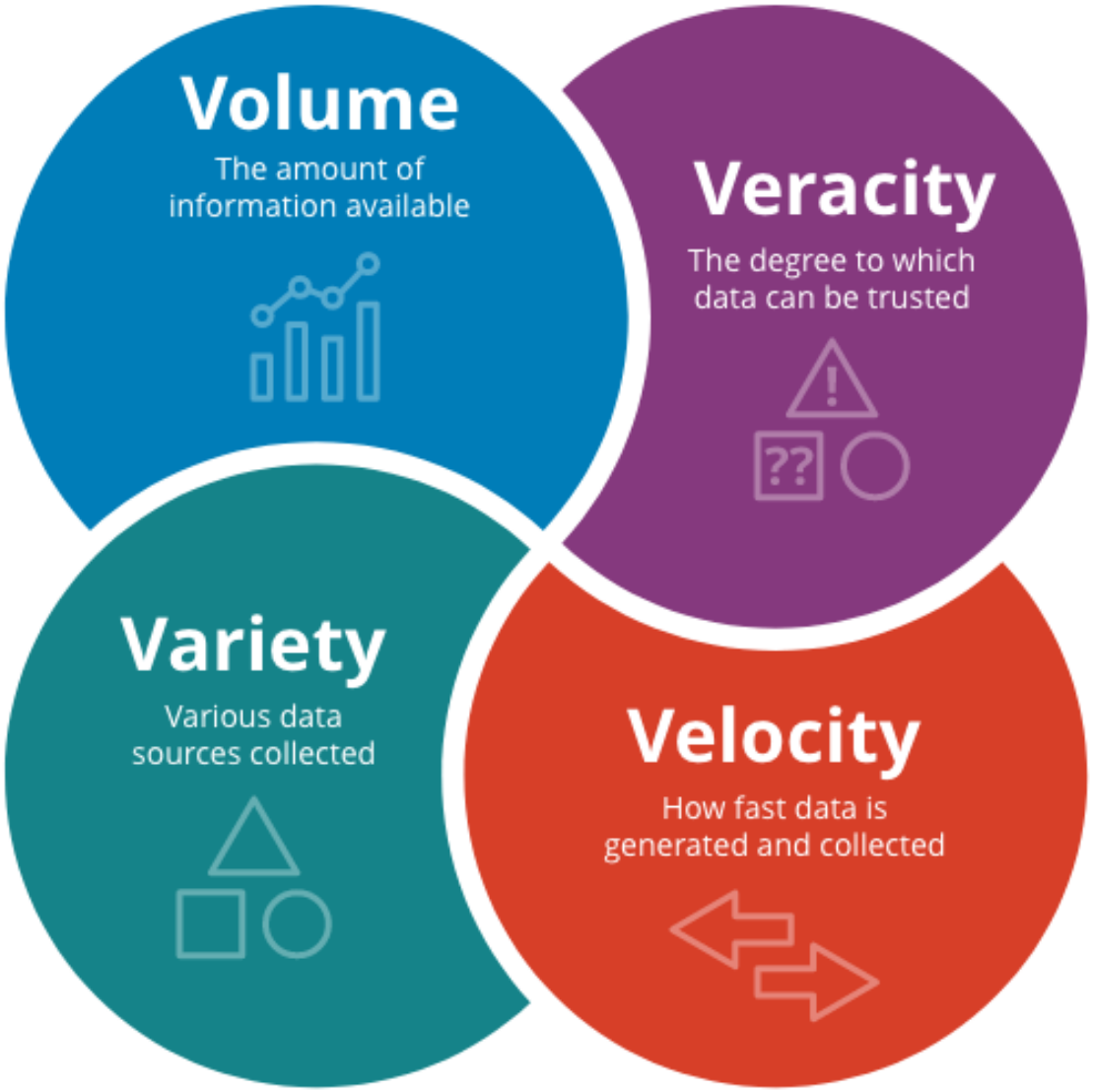


Scientific literature



Benchmark datasets specifically for ML

The 4 V's of big data

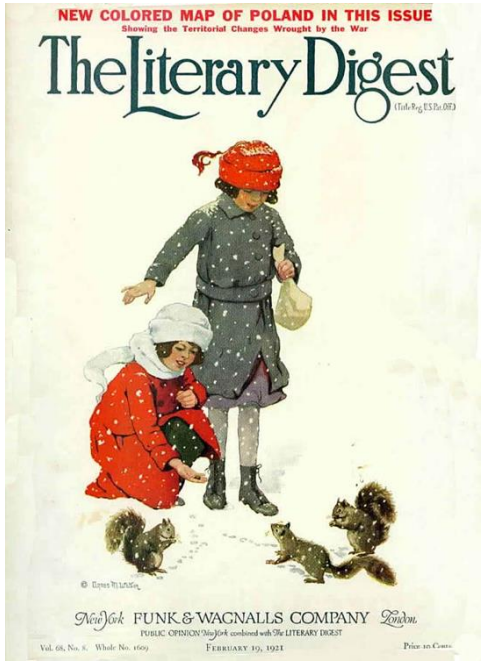


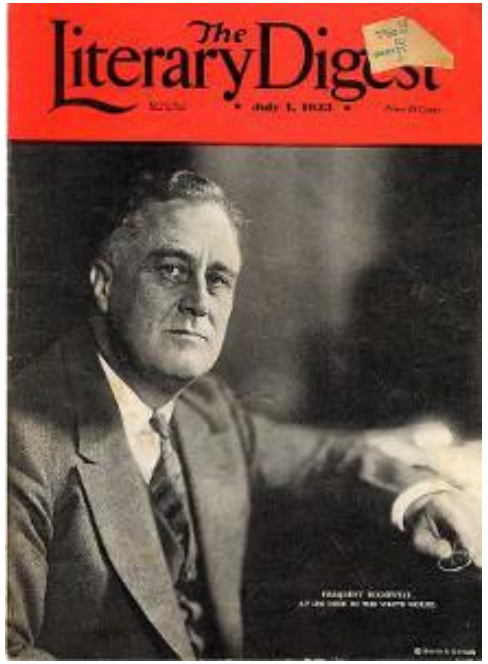
Typically, **primary data** has good veracity but limited volume and velocity

Whereas **secondary data** has much greater volume but its veracity will vary quite a bit from source to source…

Do you trust the dataset?
Is there collection error?
What about Bias?

Examples of bias in data collection





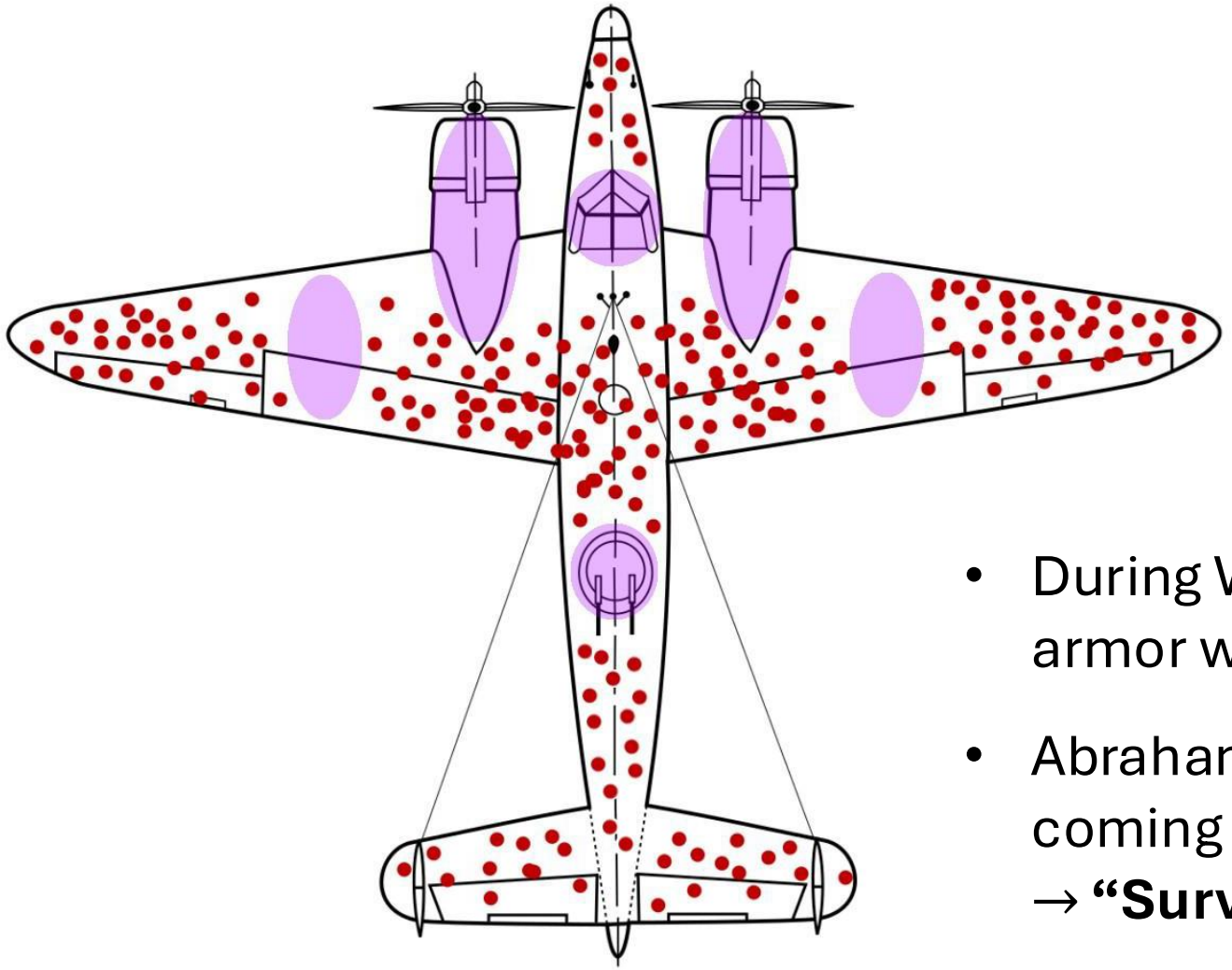
- In the 1936 election, the Literary Digest (incorrectly) predicted that Alf Landon would defeat Franklin D. Roosevelt in a landslide
- This was based on a poll conducted by surveying millions of people using **telephone directories** and **automobile registrations**…

During that time, following the Great Depression, people with cars and telephones were **disproportionately wealthy!**

Solution: Do a better job of data collection…Your sample needs to be representative of the general population



Examples of bias in data collection







- During World War 2, engineers initially added armor where the most bullet holes were observed
- Abraham Wald pointed out that this data was coming solely from airplanes that made it back…
  → **“Survivorship bias”**

Step 2: Collect the *right* data for the problem

Ask an interesting question

Get the Data

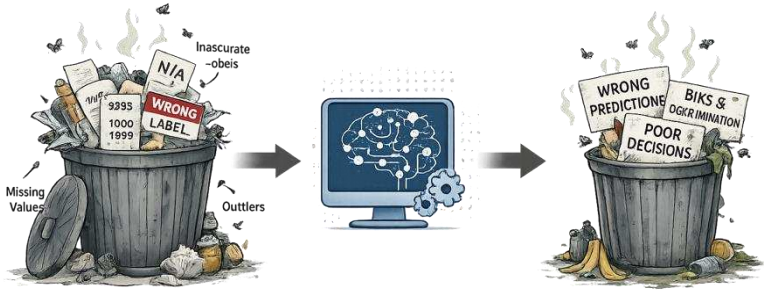
Clean the Data

Model the Data

Interpret the Results

Be deliberate with your data collection to **minimize bias** – and when you can't avoid it, at least account for it!

Remember…
"Garbage in = Garbage out"



(ironically, this AI-generated image is also kind of garbage)

A fancy model will not save you from bad data!

Step 3: Clean and curate the data

Ask an interesting question

Get the Data

Clean the Data

Model the Data

Interpret the Results

To avoid “garbage” – we should *clean* our data. Most real-world data is messy and imperfect!

Data curation is most of the job!

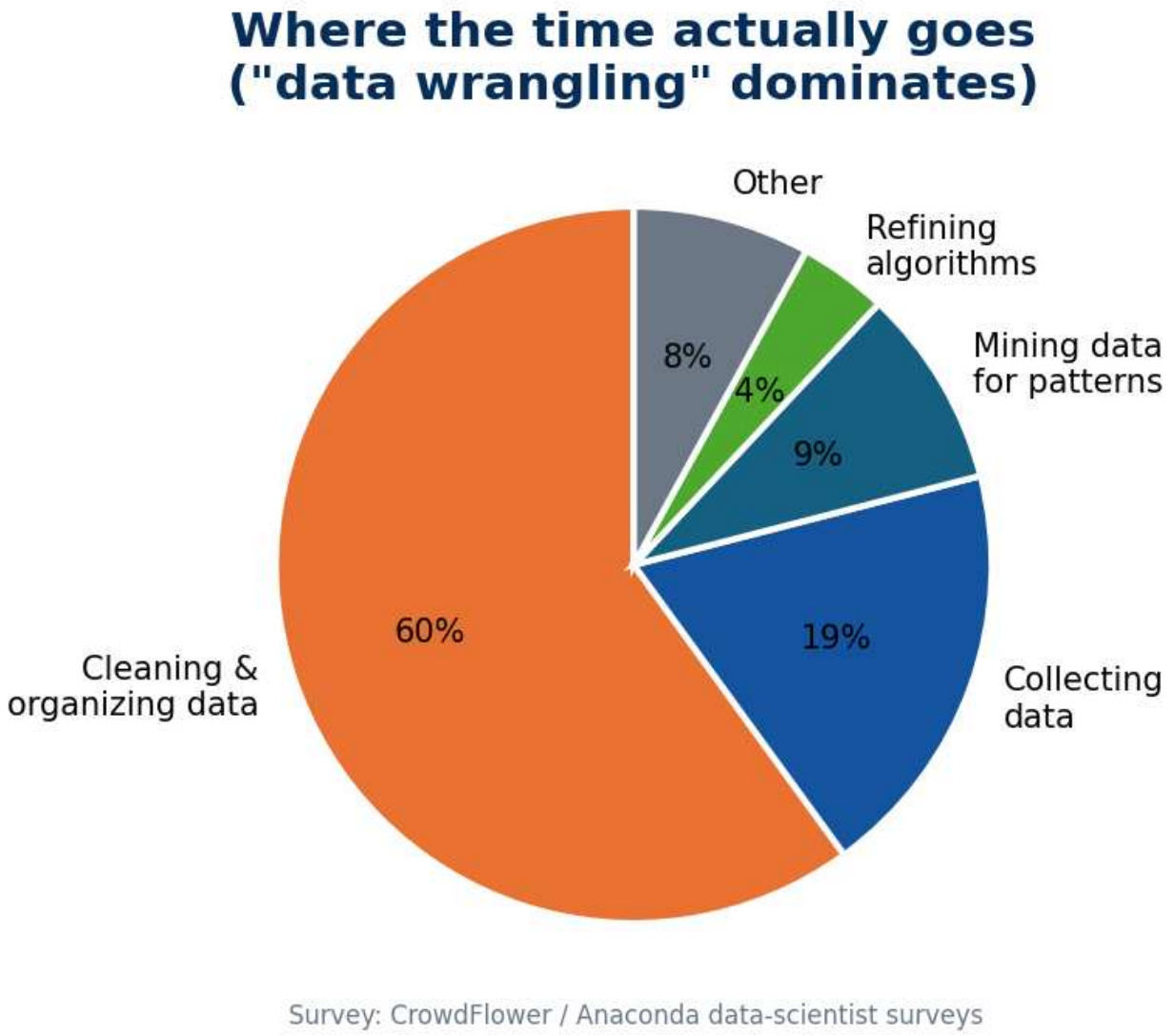
Ask an interesting question

Get the Data

Clean the Data

Model the Data

Interpret the Results



Data curation: Wrong/missing values

Company	Employee Name	Age	Time Spent (hours)
Apple	John S.	23	100
Apple	Evan B.	27	8
Apple	Emily B.	31	12
Google	Ava W.	—	7
Google	Noah A.	34	9



Incorrect measurement



Incorrect time



Incomplete data

Be wary of outliers…But don’t blindly remove them

In cases of missing values, you may remove the row entirely, in which case you lose the rest of the useful data, or ***impute*** by averaging

Data curation: Integration

Player Name	Team	Weight (lbs.)
P. Bardsley	Chelsea	150
D. McNeil	Chelsea	198
Adam Legzdins	Chelsea	170
Dan Agyei	Chelsea	168
David Luiz	Chelsea	192

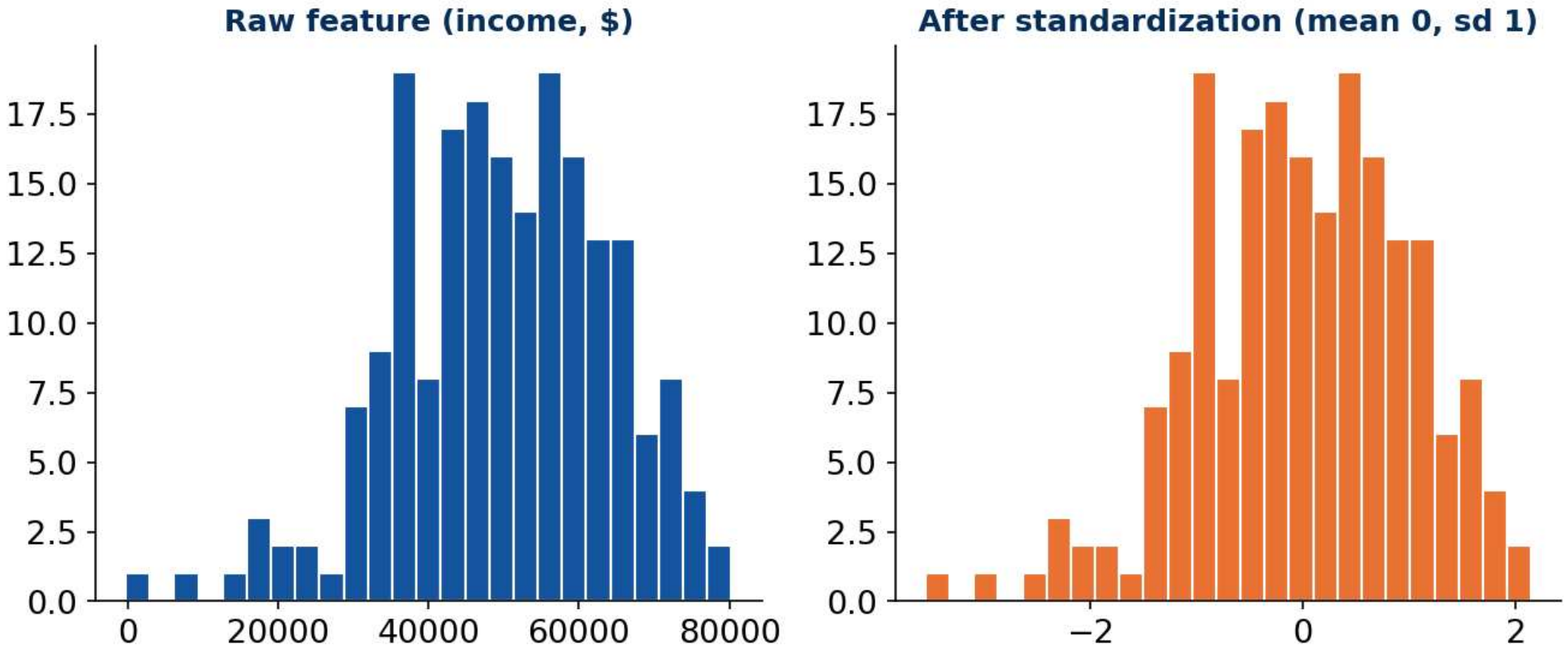
Source: X (in lbs.)

Player Name	Team	Weight (kgs.)
Jamal Blackman	Chelsea	72
Ethan Ampadu	Chelsea	68
Billy Gilmour	Chelsea	73
Ike Ugbo	Chelsea	64.5
George McEachran	Chelsea	75

Source: Y (in kgs.)

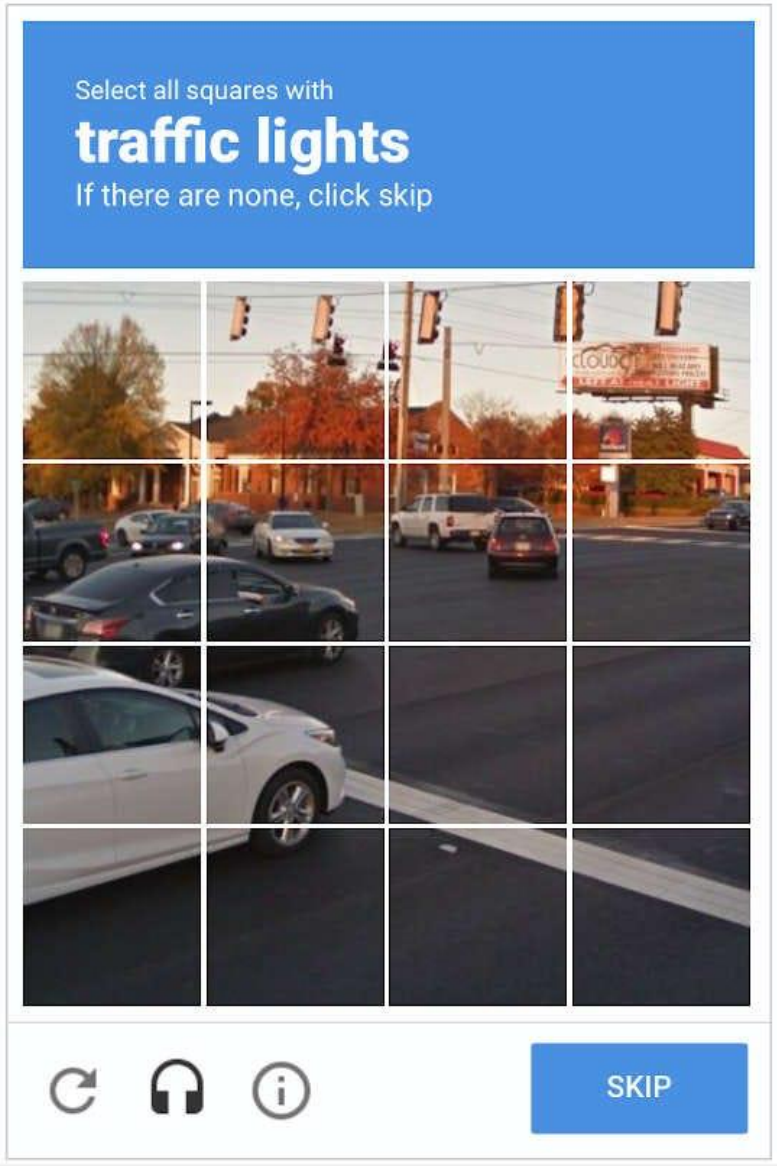
Data curation: Feature engineering & scaling

Numerical models often “play nicer” when dealing with **standardized numbers**



This is especially important when you have multiple features with different magnitudes

Data curation: Labeling



Some ML models need data to be **labeled…**
This can be very time-consuming!



Data curation: Labeling

The lack of large, well-labeled datasets was a major impediment to early ML



Advances like the ImageNet dataset (millions of annotated images!) helped solve this

Step 4: Build (and test) a model of the data

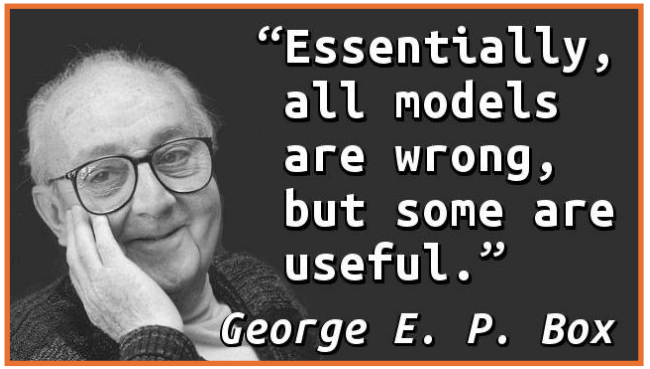
Ask an interesting question

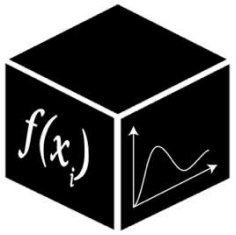
Get the Data

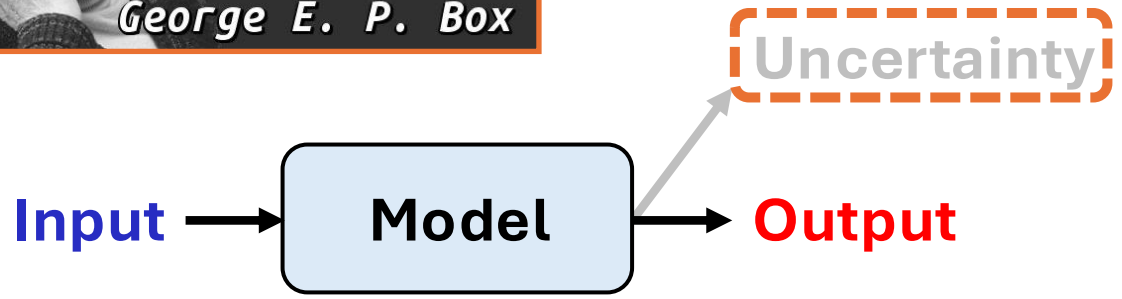
Clean the Data

Model the Data

Interpret the Results



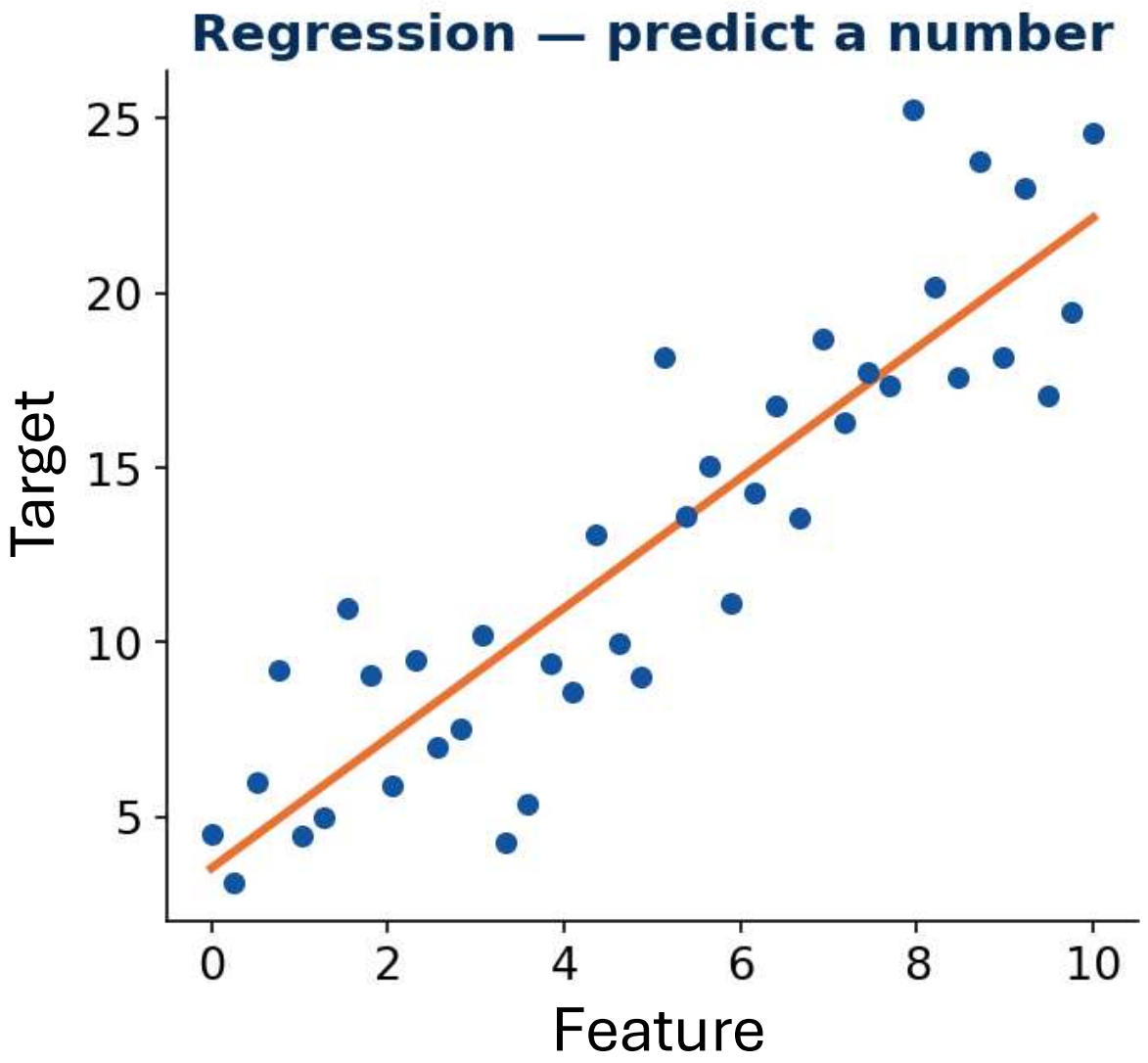




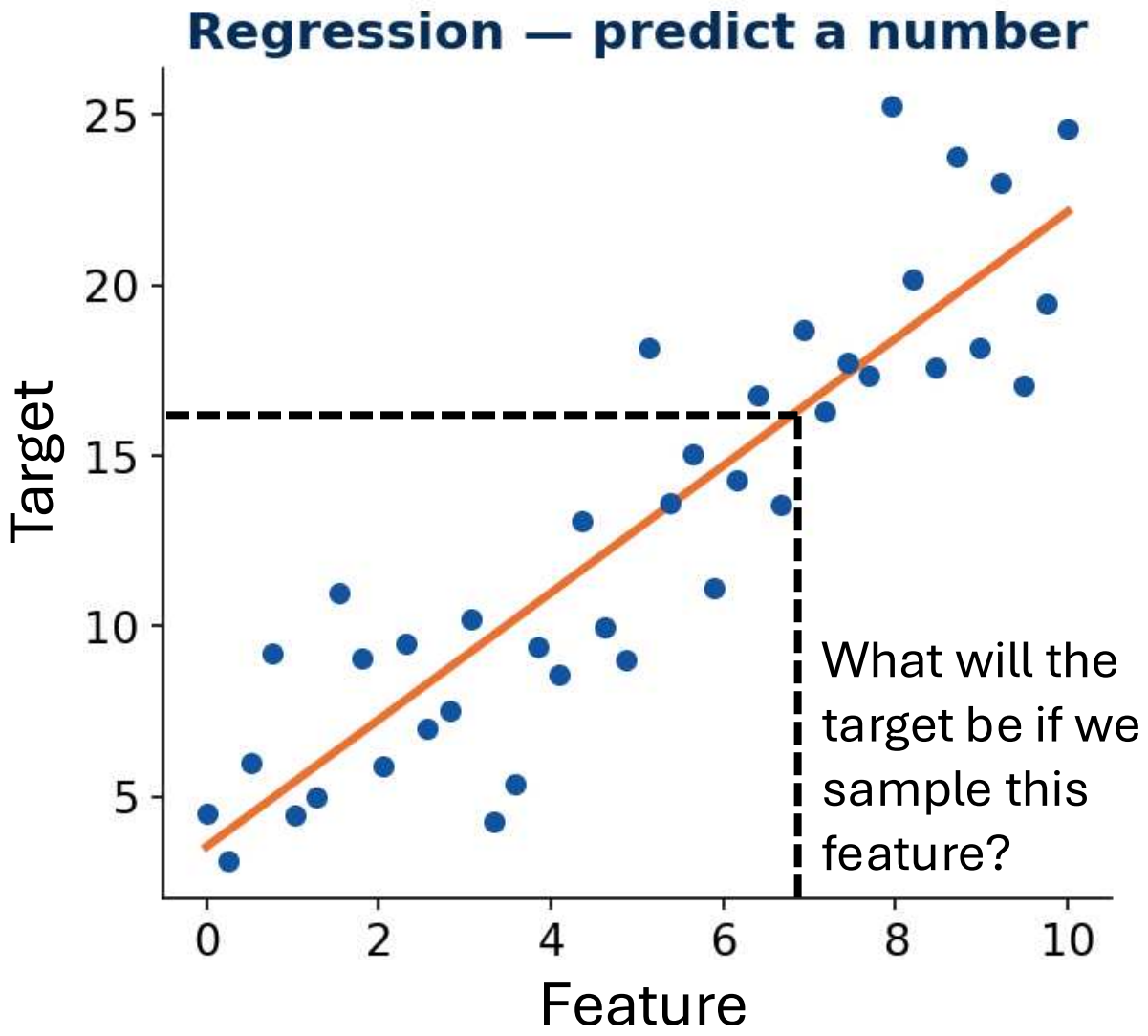
$$y = f(x)$$

The output is some *function* of the input; our goal is to (approximately) learn that function

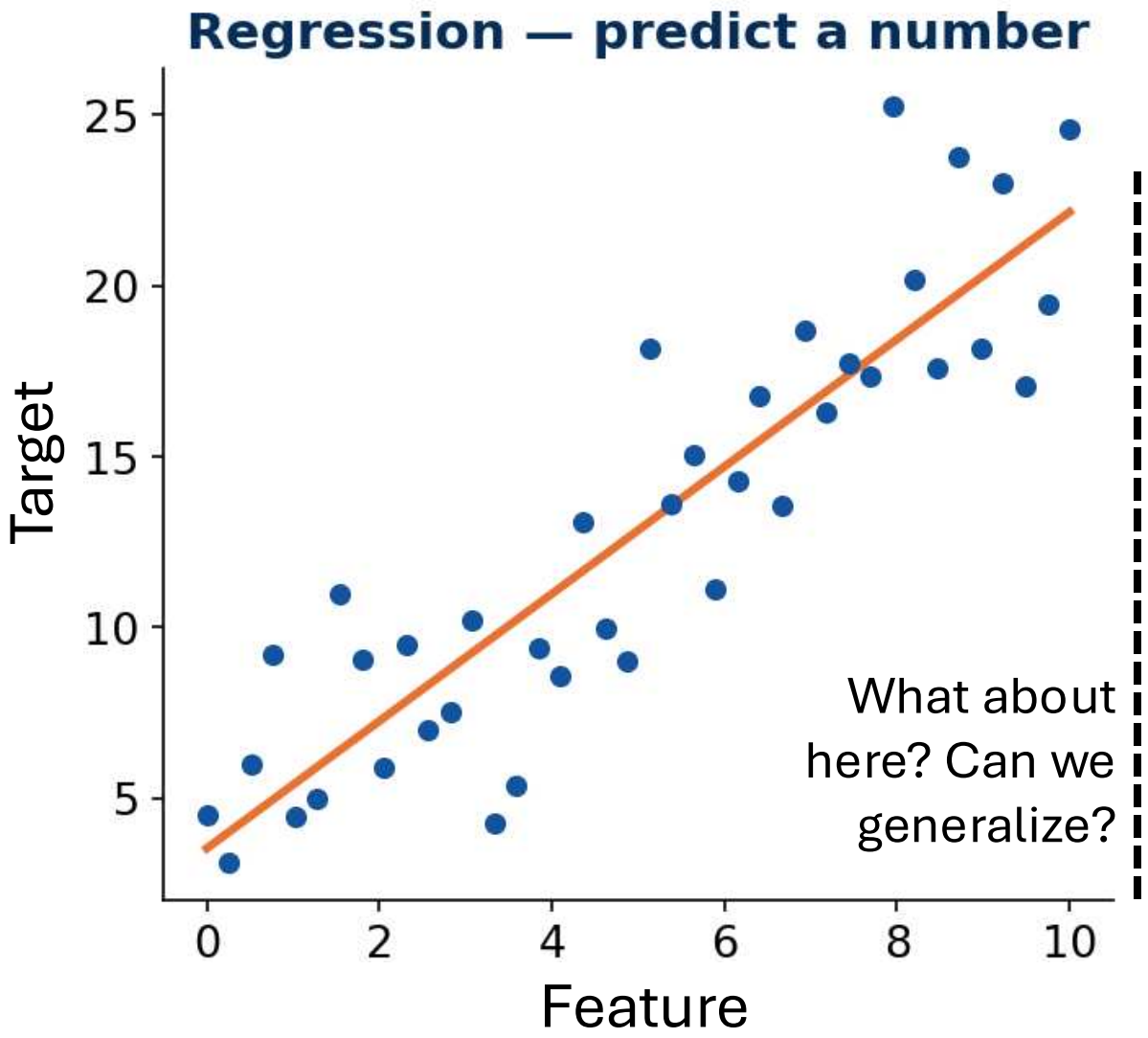
Regression *vs.* classification



Regression *vs.* classification



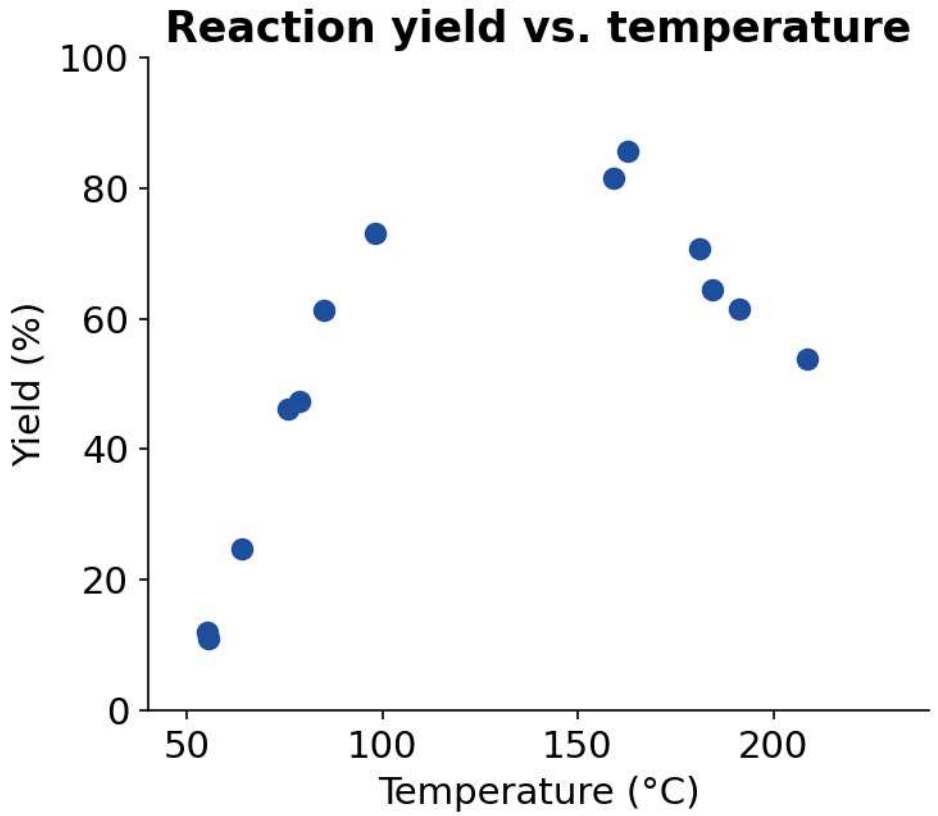
Regression *vs.* classification



An example of regression

I am trying to synthesize a new compound by heating reactants together. I have some preliminary data on yield (how much of my desired compound formed) *vs*. temperature. **Which temperature should I try next to maximize yield?**

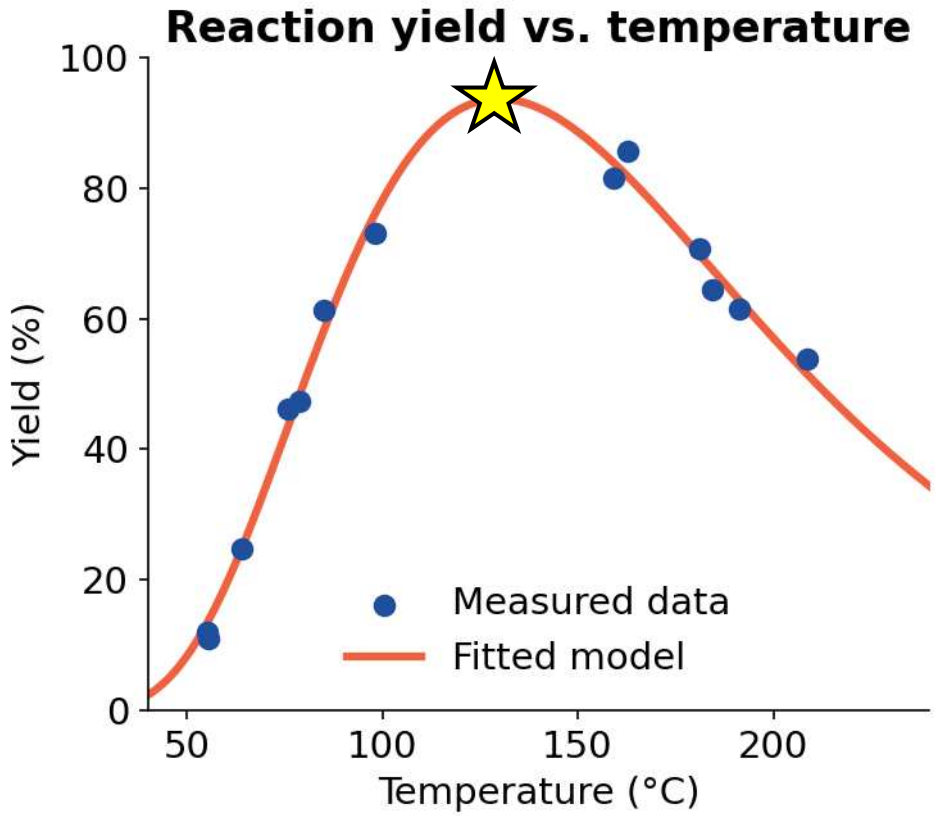




An example of regression

I am trying to synthesize a new compound by heating reactants together. I have some preliminary data on yield (how much of my desired compound formed) *vs*. temperature. **Which temperature should I try next to maximize yield?**

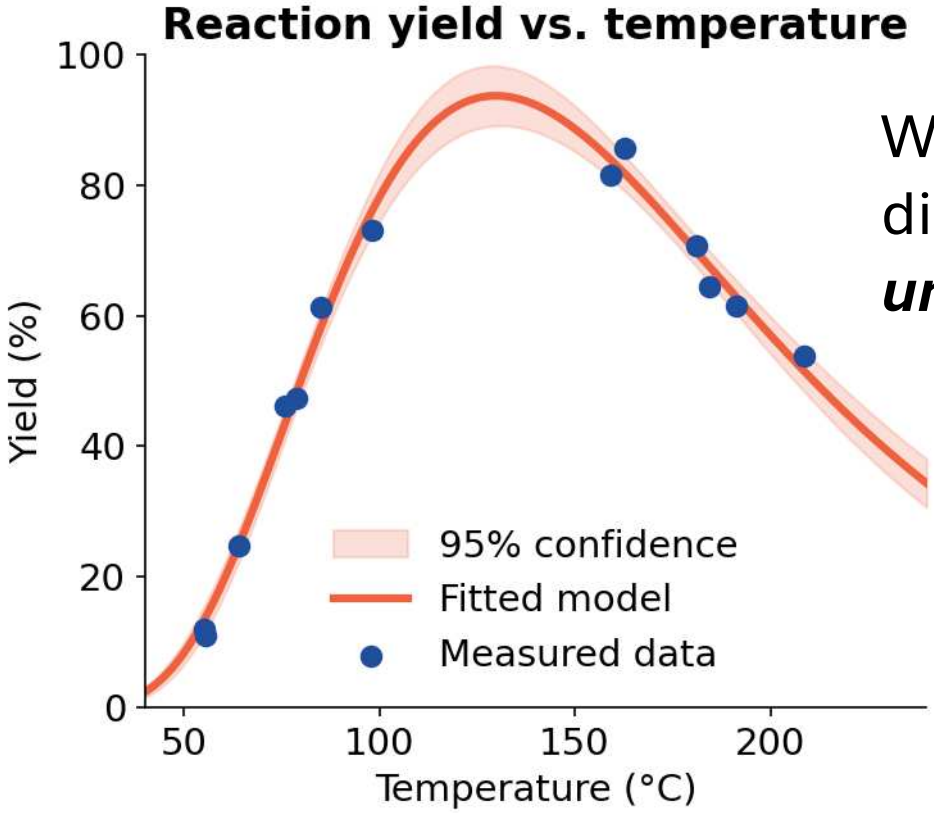




An example of regression

I am trying to synthesize a new compound by heating reactants together. I have some preliminary data on yield (how much of my desired compound formed) *vs*. temperature. **Which temperature should I try next to maximize yield?**

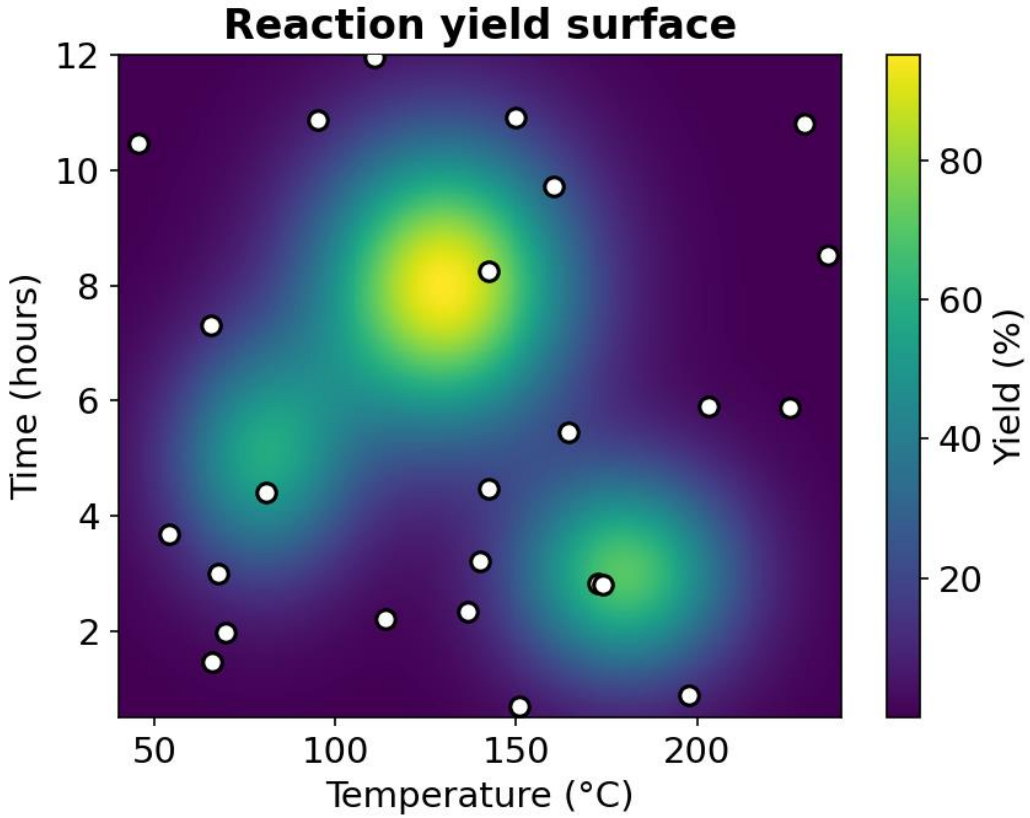


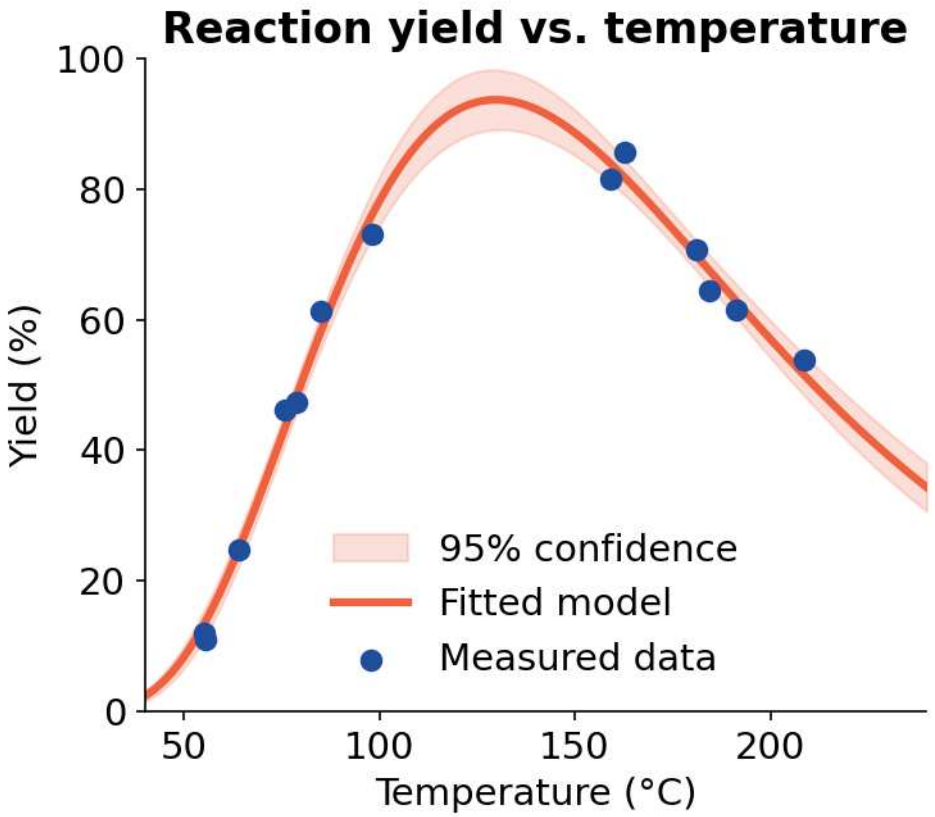


We'll also discuss ***uncertainty***

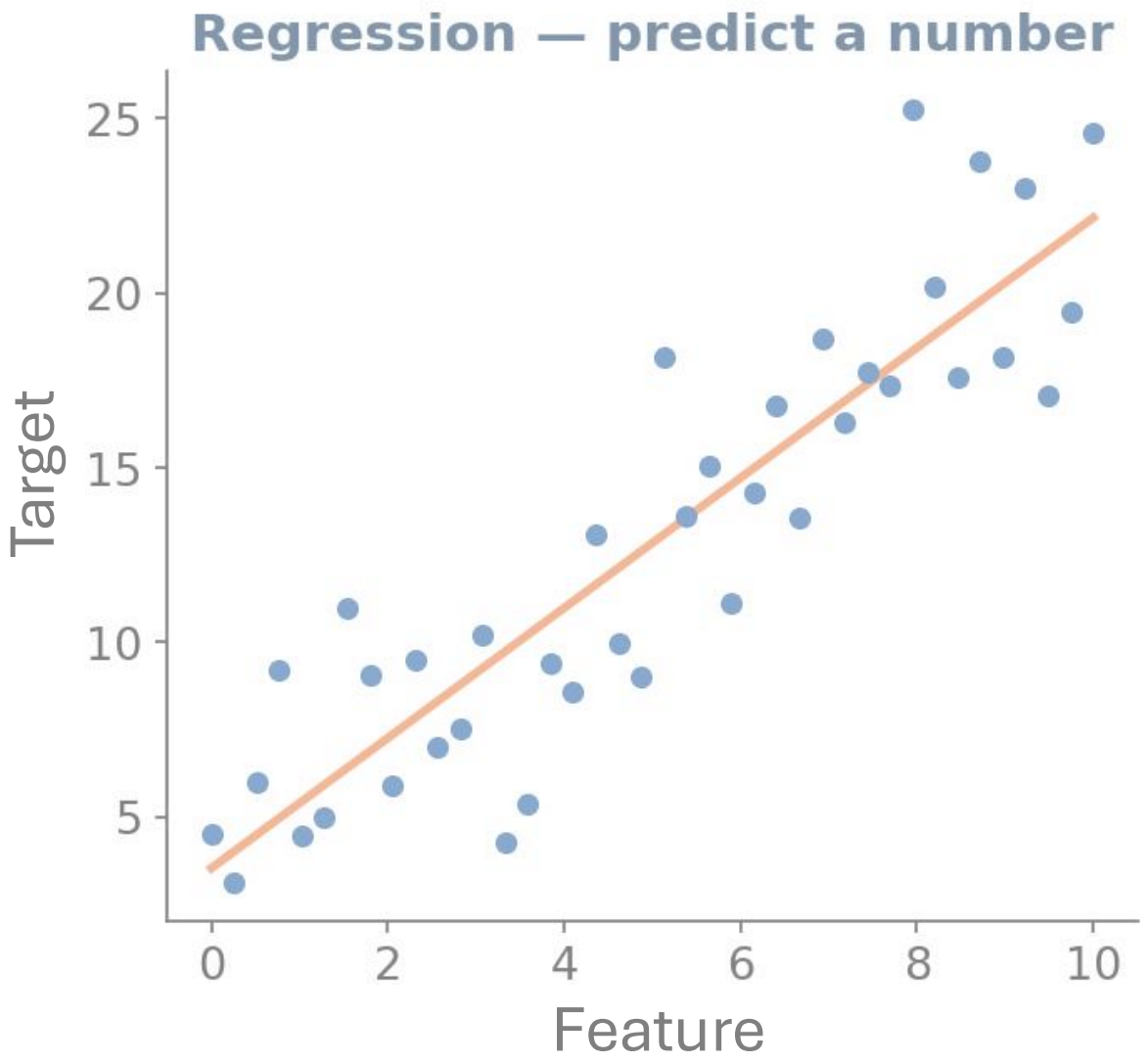
An example of regression

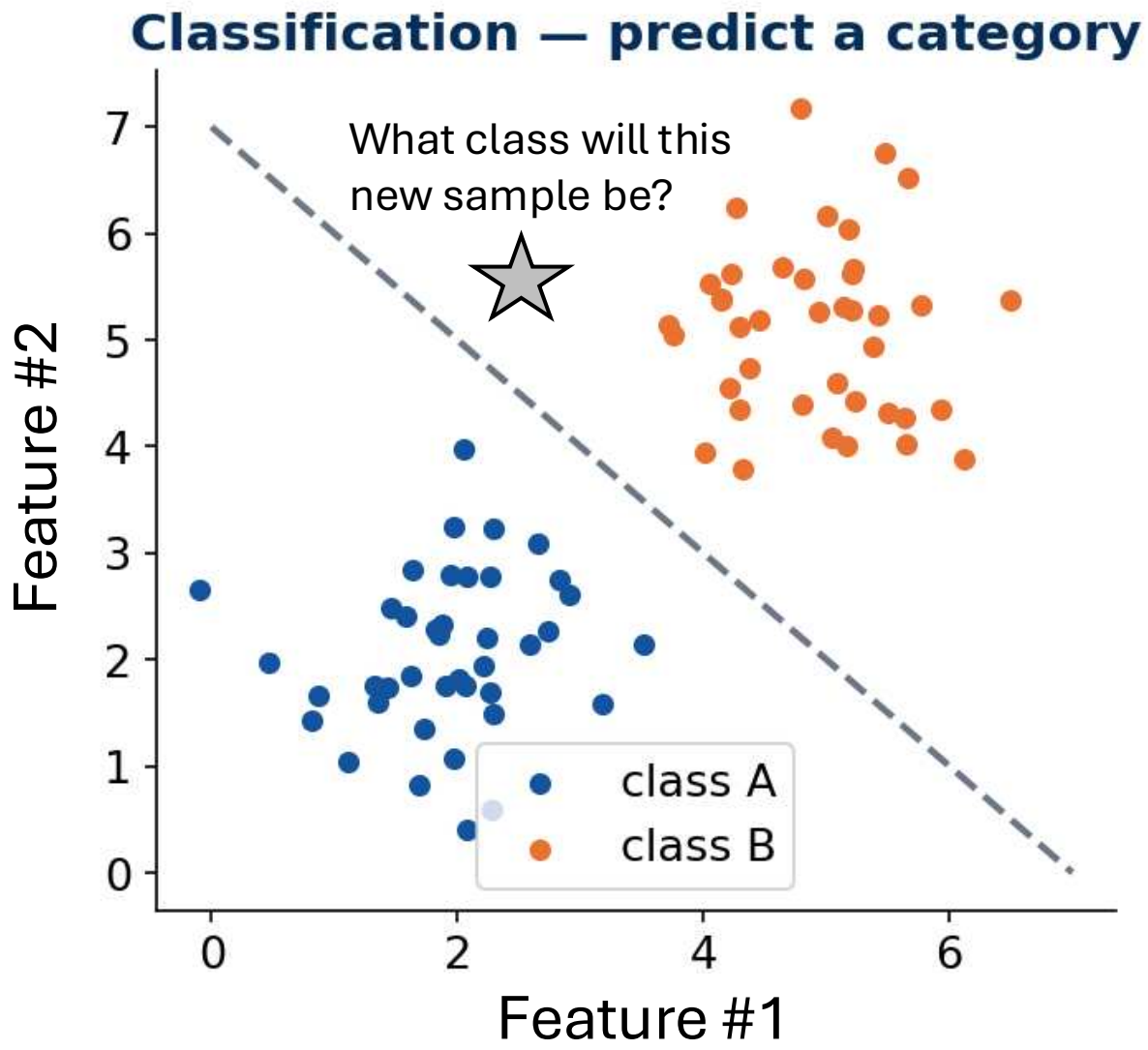
We'll also generalize to
multiple input variables





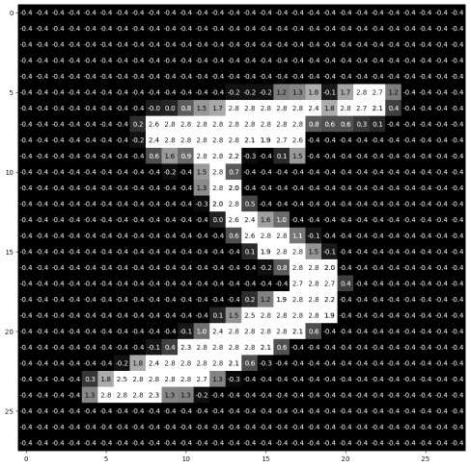
Regression *vs.* classification



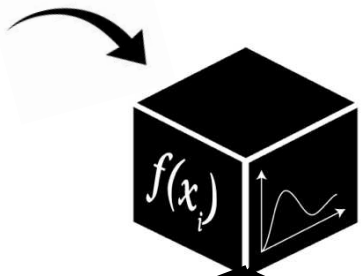


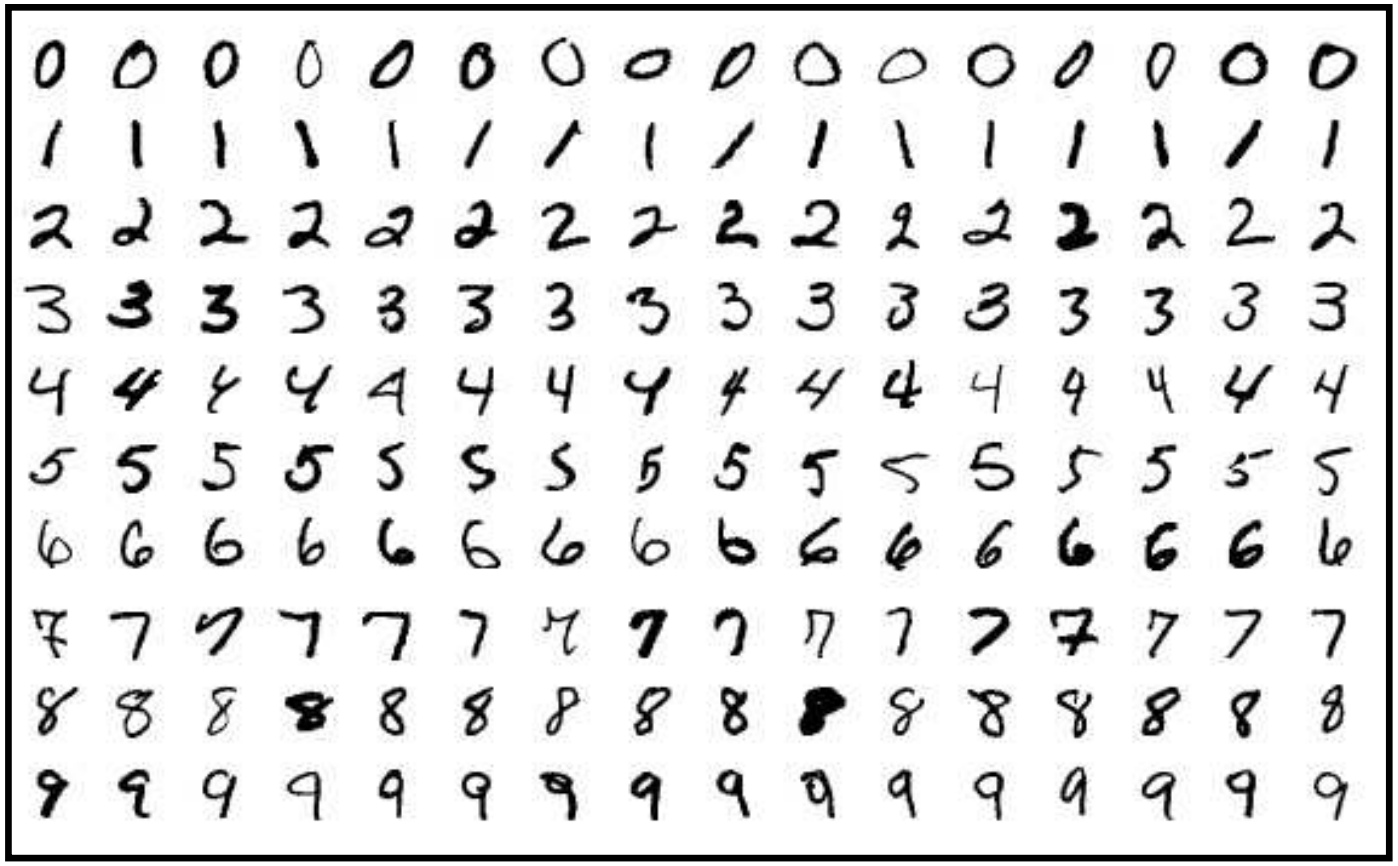
An example of classification

Can we train an ML model to **classify a handwritten digit** as **0, 1, 2, …, 9**?

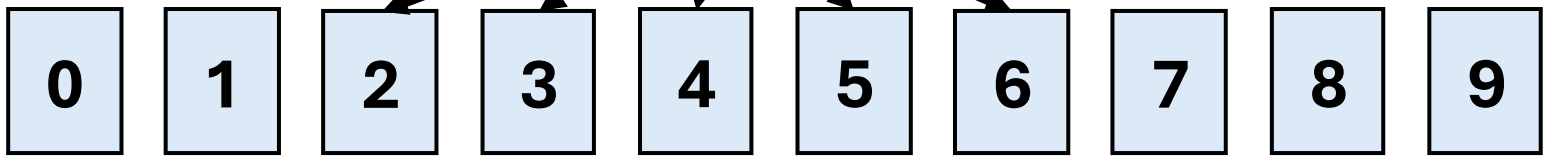


Input:
28×28 grid
of values



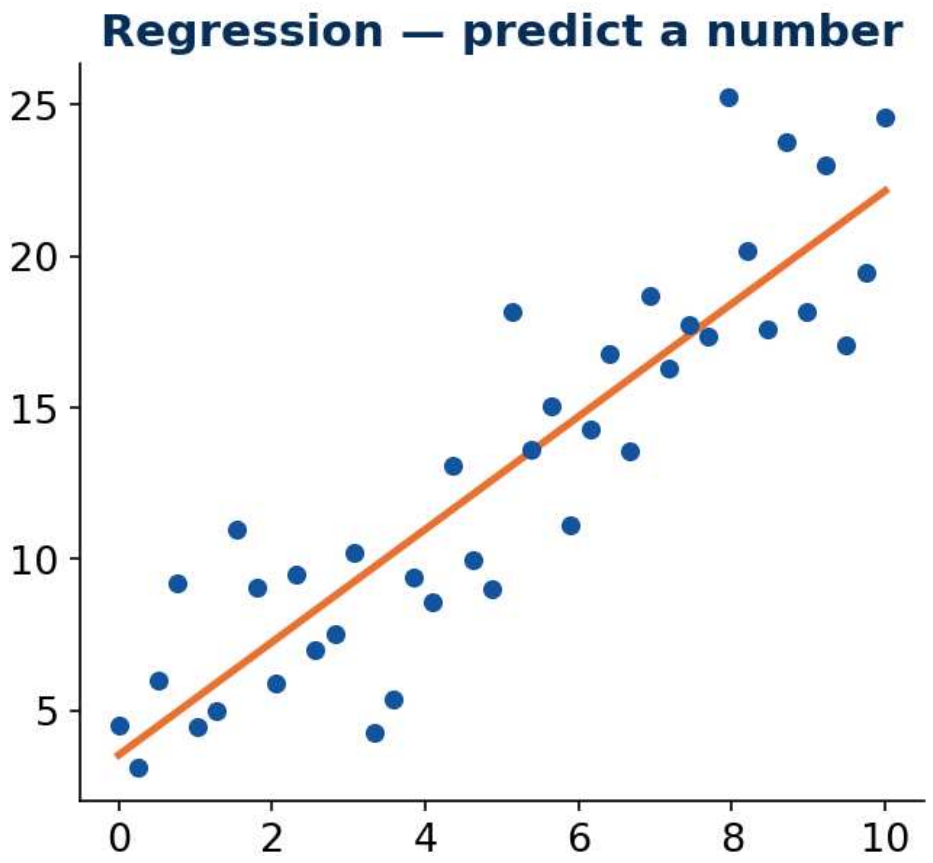


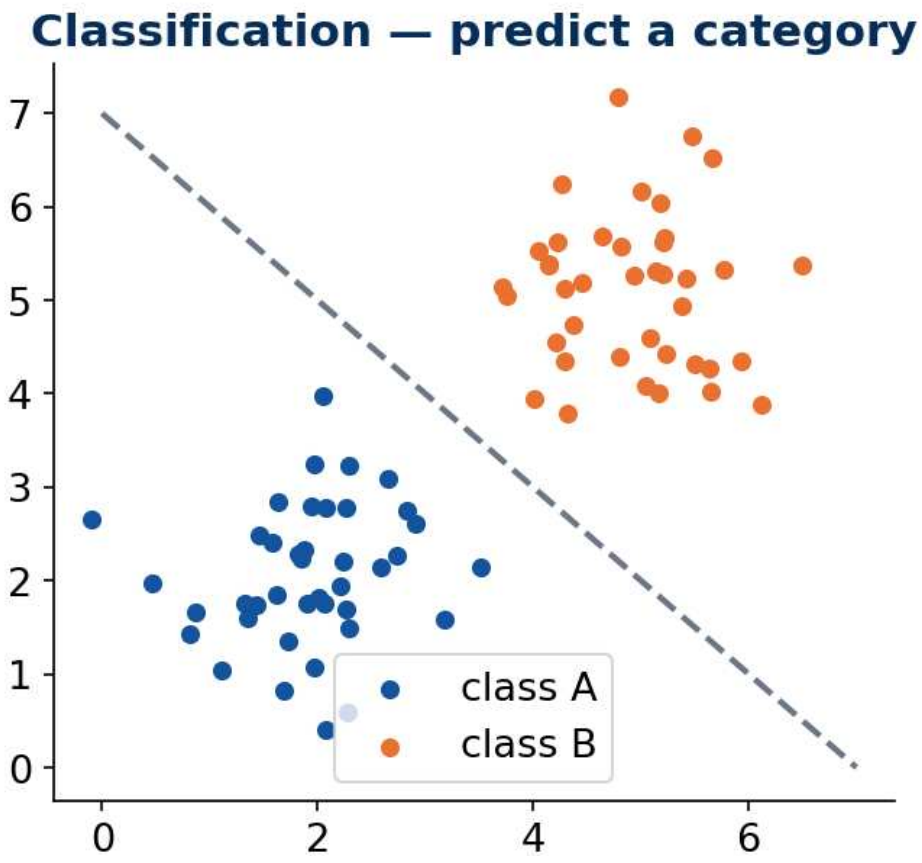
MNIST dataset



Surprisingly difficult,
prior to CNNs!

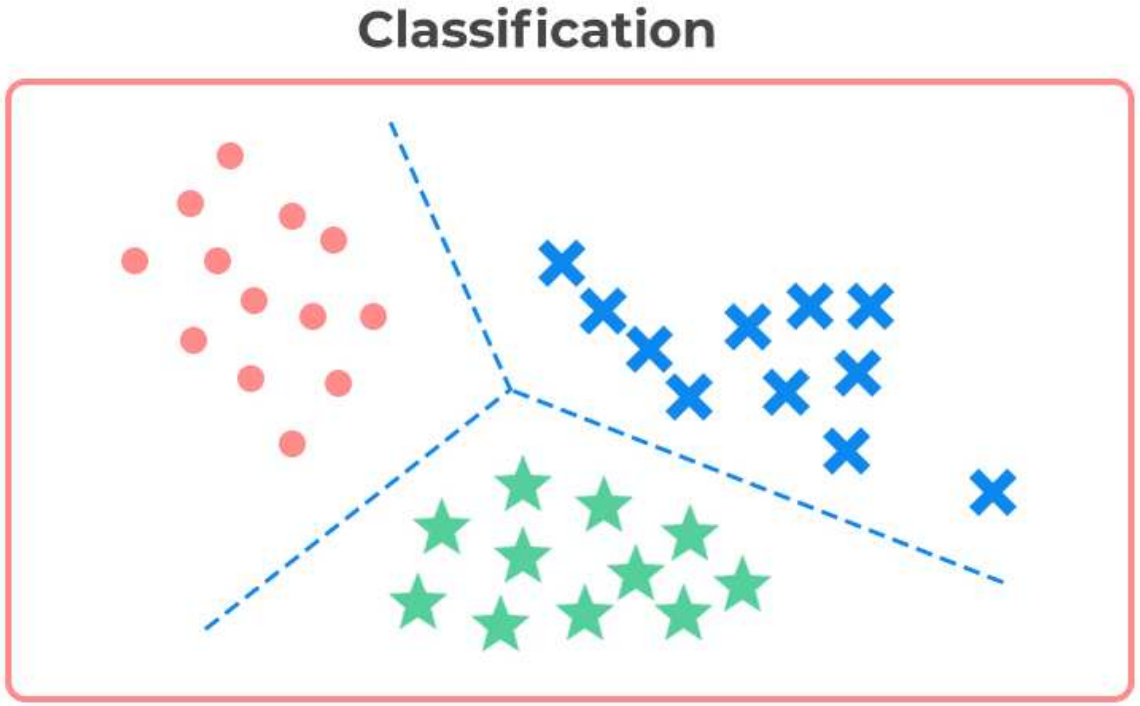
Supervised *vs.* unsupervised learning

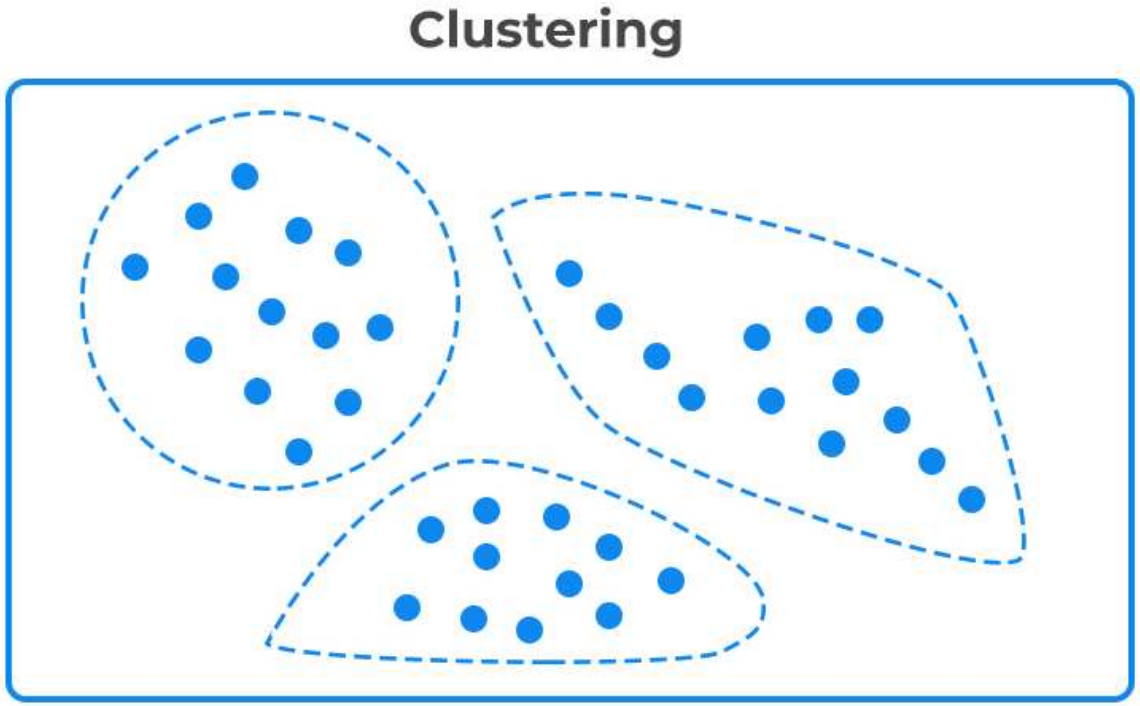




These are both examples of **supervised learning** because we have some **labeled data** to train on

Supervised *vs.* unsupervised learning

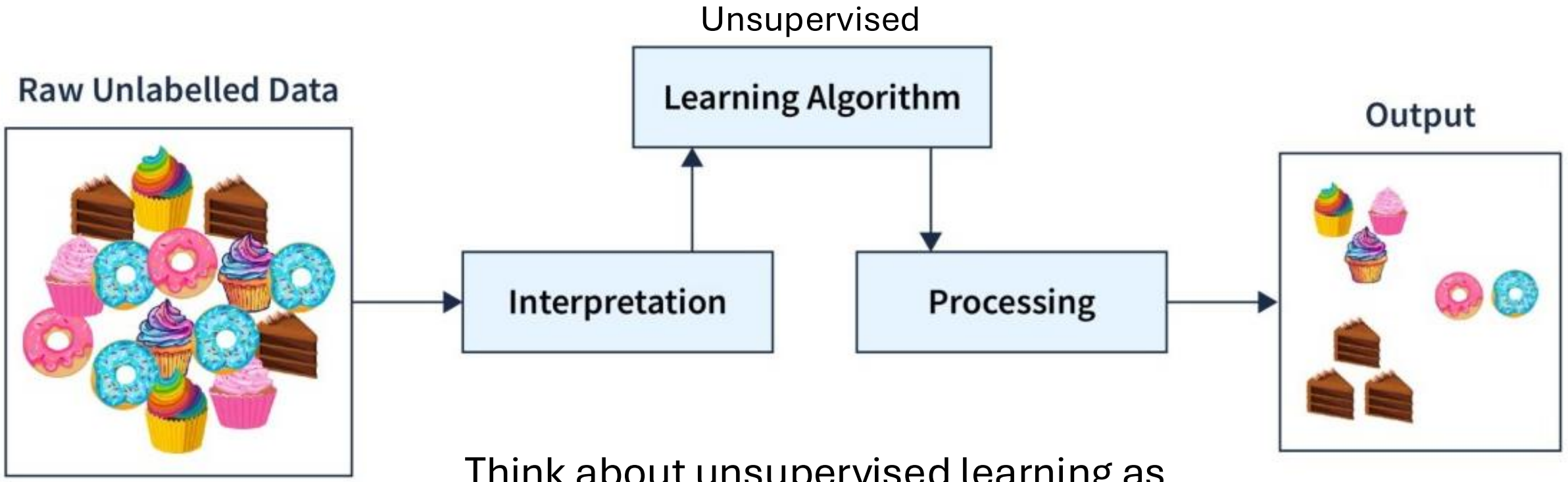




In this case, **we know** the class of each datapoint

In this case, **we do not know** the classes/labels but **we can still learn** from the structure of the data

Supervised *vs.* unsupervised learning



Think about unsupervised learning as **grouping related samples** together.
Humans can take a look after the fact.

Step 5: Interpret and visualize the results

Ask an interesting question

Get the Data

Clean the Data

Model the Data

Interpret the Results

How well does the model perform?

What conclusions can you draw from it?

How should you present these conclusions?

Evaluate performance…on what data?



Model learns the task

How good is the model on "new" data?

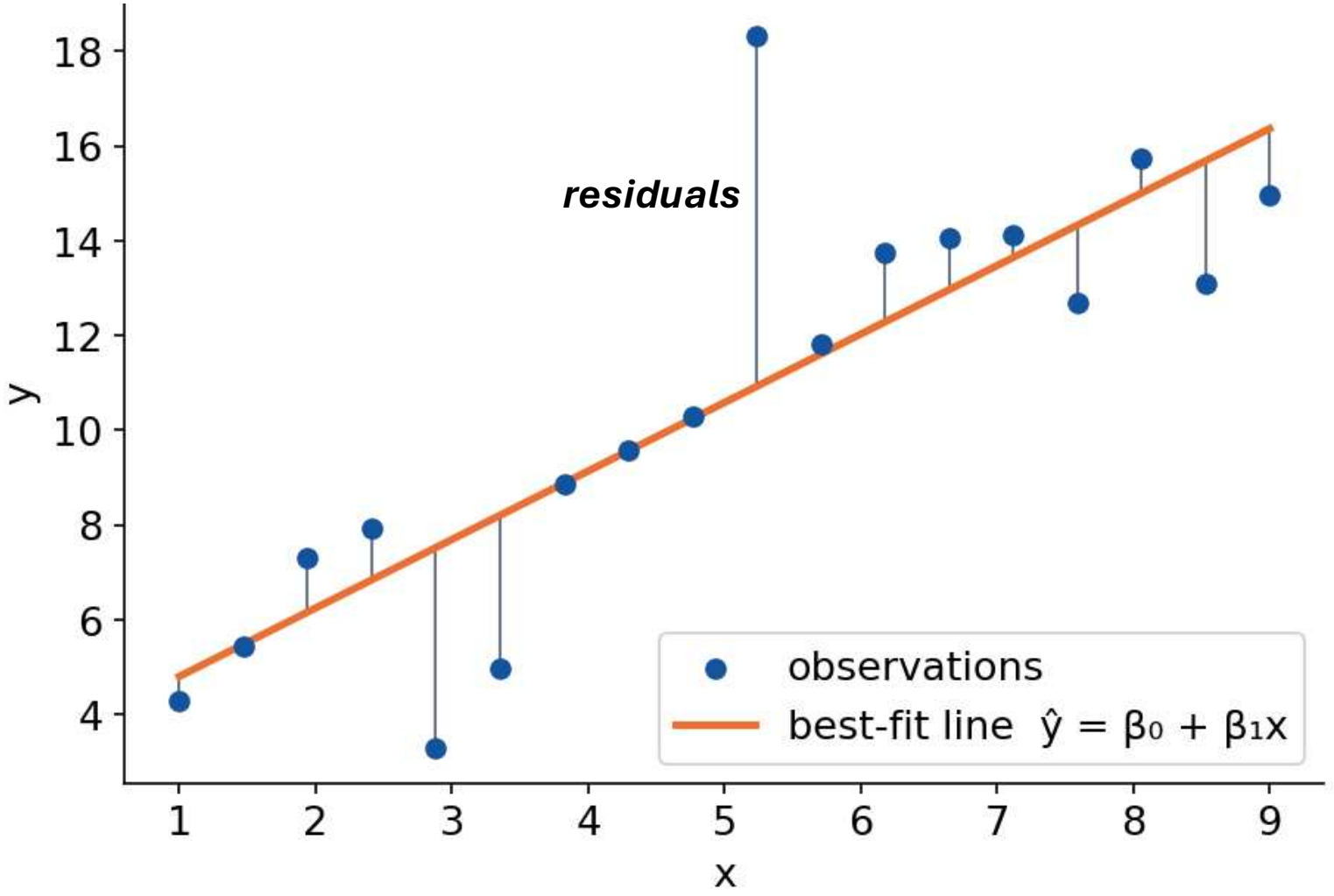
Like taking an exam

Performance metrics will be reported on the test dataset

This test dataset must be designed carefully…
We'll discuss this in detail later!

Choosing the right metrics: Regression

Metrics for regression models



Mean absolute error:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y^{i}_{\text{pred}} - y^{i}_{\text{actual}}\right|$$

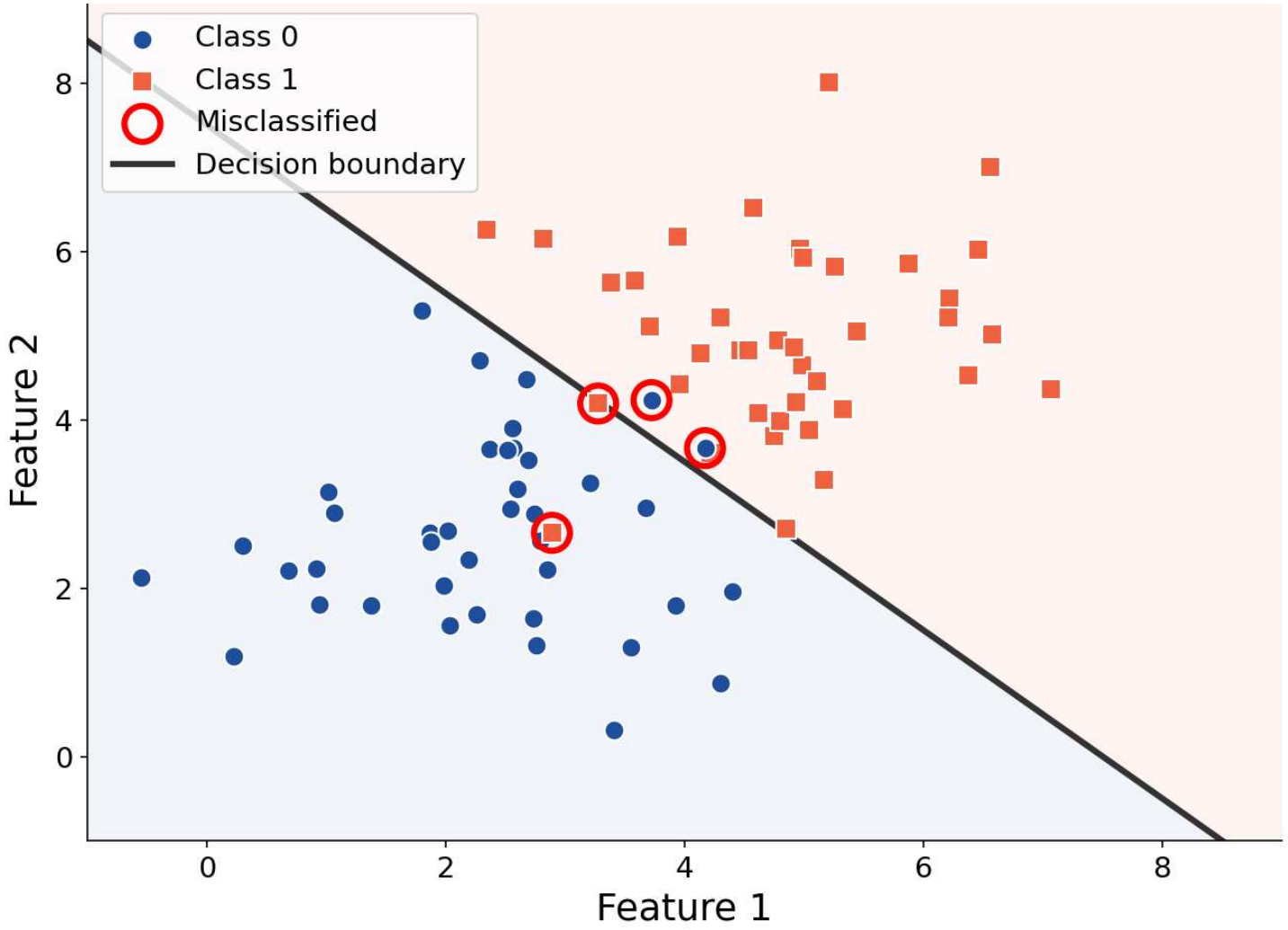
Mean squared error:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y^{i}_{\text{pred}} - y^{i}_{\text{actual}}\right|^{2}$$

MSE **penalizes outliers** more strongly than MAE

Choosing the right metrics: Classification

Metrics for classification models

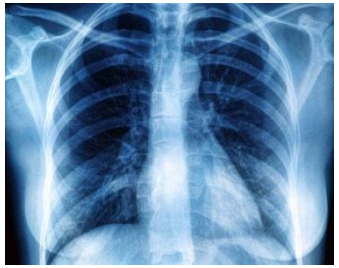


The simplest metric is **Accuracy**, defined as:

$$\frac{\text{Number of } \mathbf{correct} \text{ labels}}{\text{Number of } \mathbf{total} \text{ labels}}$$

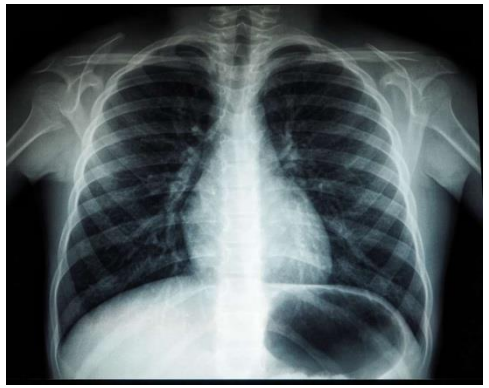
But is this really the *best* metric to use?

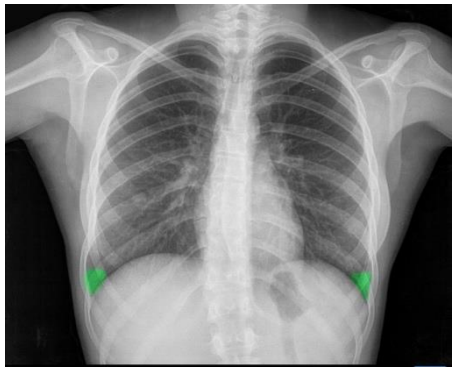
Accuracy can be misleading!













Does ***not*** have lung cancer (98% of cases)

You can get **98% accuracy** by just predicting **nobody has cancer**… Not a very useful model!

This is called ***class imbalance***

Does have lung cancer (2% of cases)

Metrics to account for class imbalance

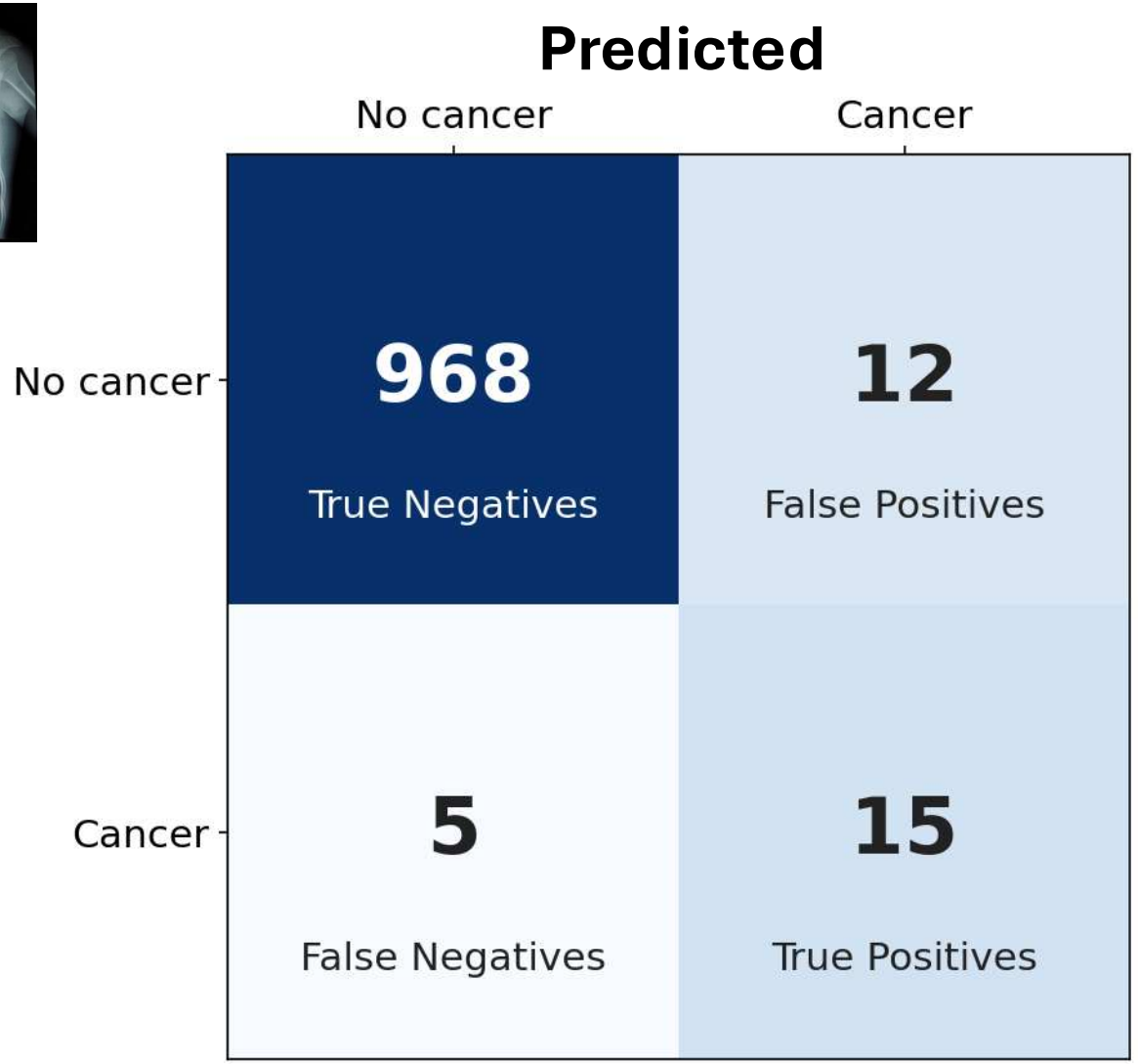
We can visualize model performance by using a **confusion matrix**



The number of true/false positives and negatives can also be quantified through:

Precision → false *positives* are costly
Recall → false *negatives* are costly
F_1-score → a balance of both

(…details to come…)



Step 5: Interpret and *visualize* the results

Ask an interesting question

Get the Data

Clean the Data

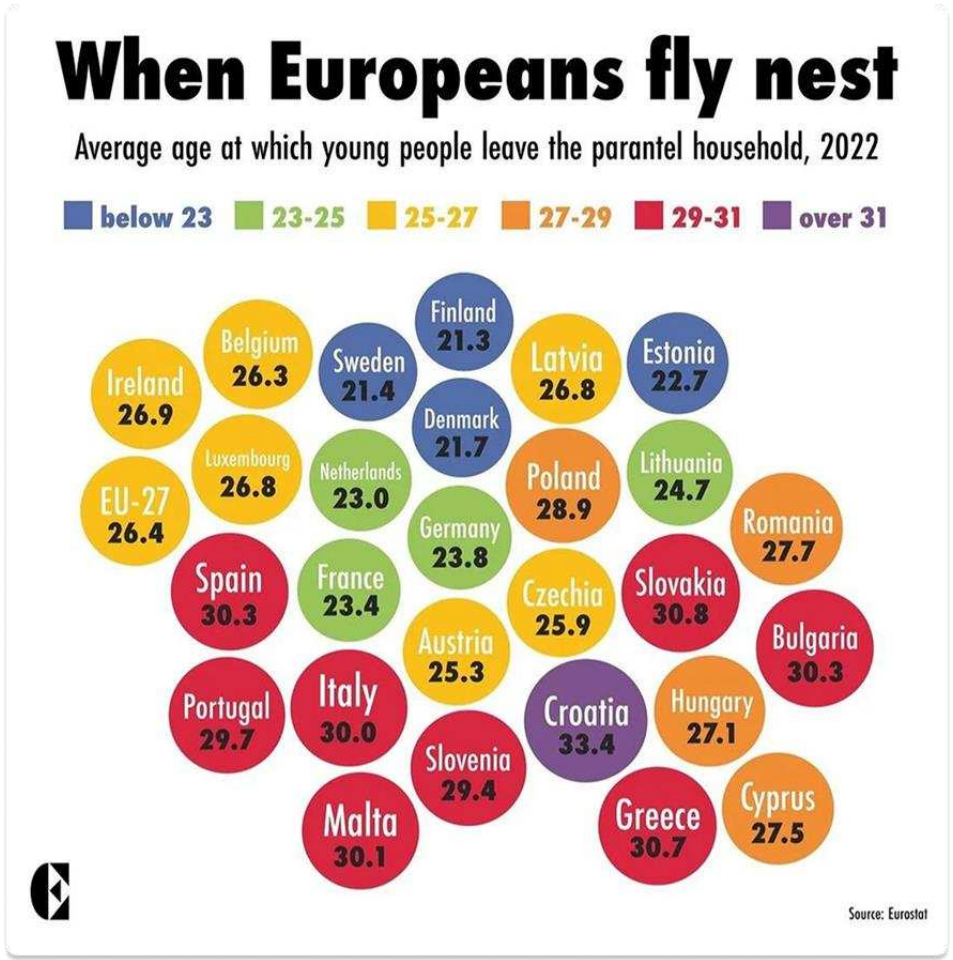
Model the Data

Interpret the Results

Your findings do not matter if you cannot effectively communicate them!

How well does the model perform?
What conclusions can you draw from it?
How should you present these conclusions?

Some examples of *bad* data visualization

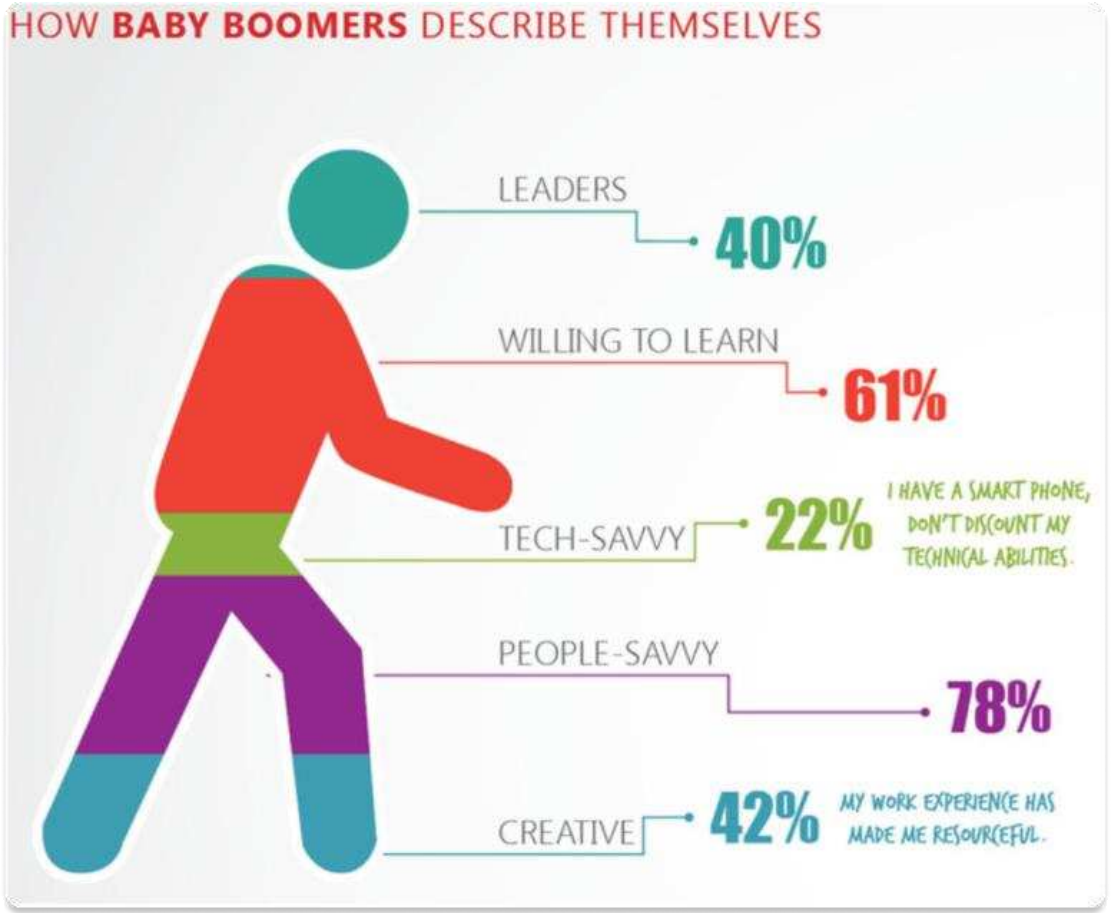


The numbers have no clear relation to the visual

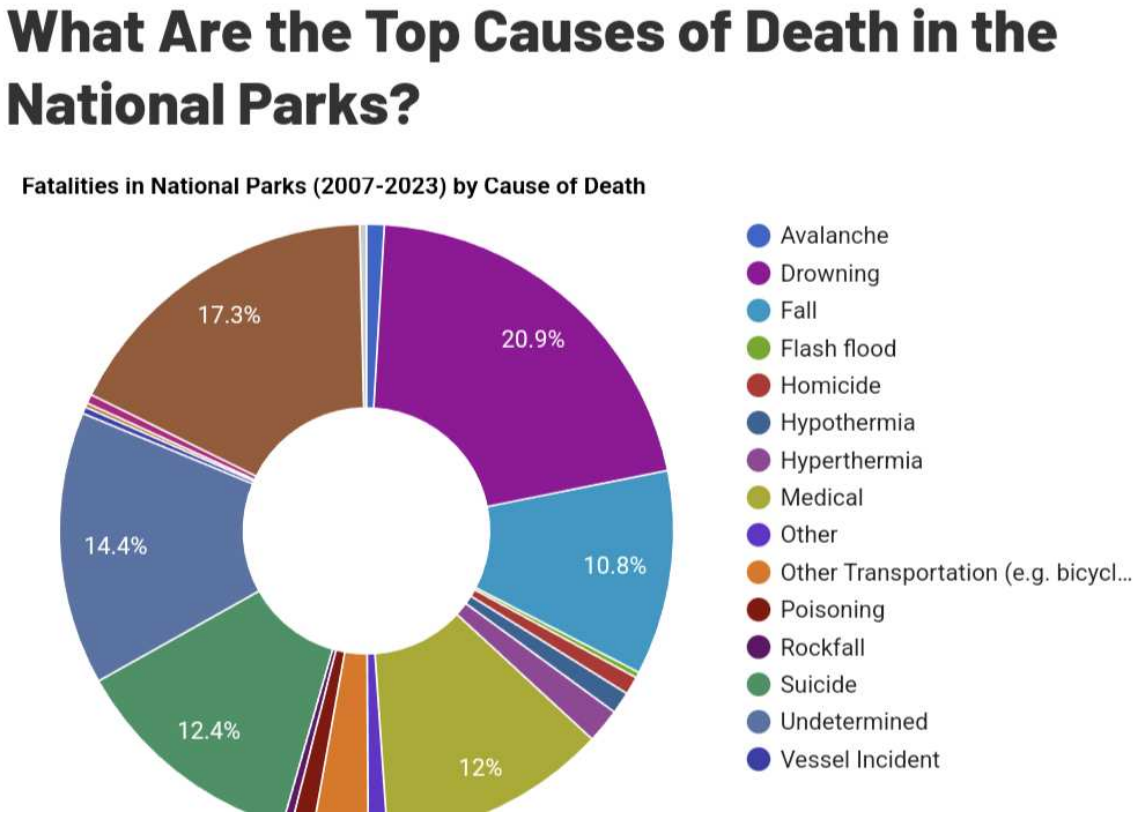


Can’t even see the bars in back…And what do the colors represent? What is the *y*-axis?

Some examples of *bad* data visualization



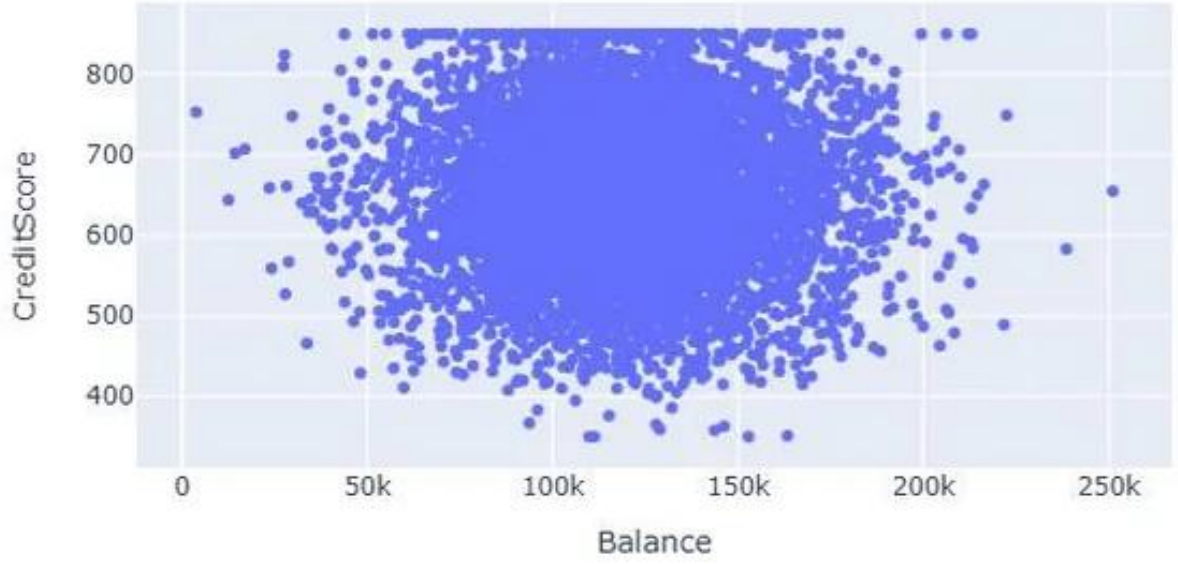
This oddly shaped “pie chart” is very non-intuitive



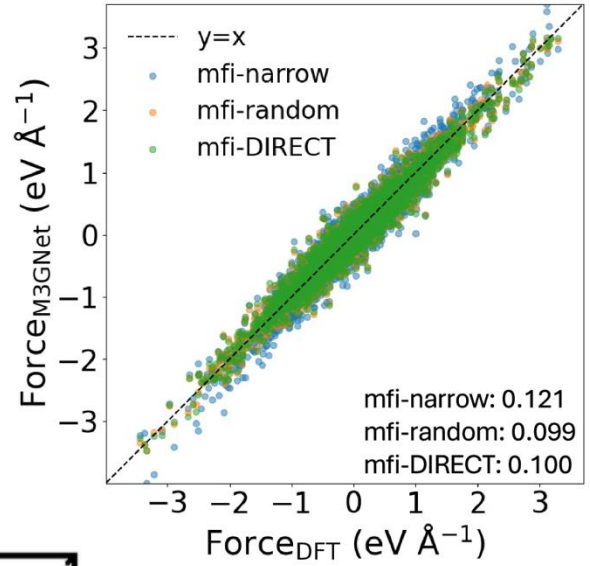
There are too many labels with no clear connections

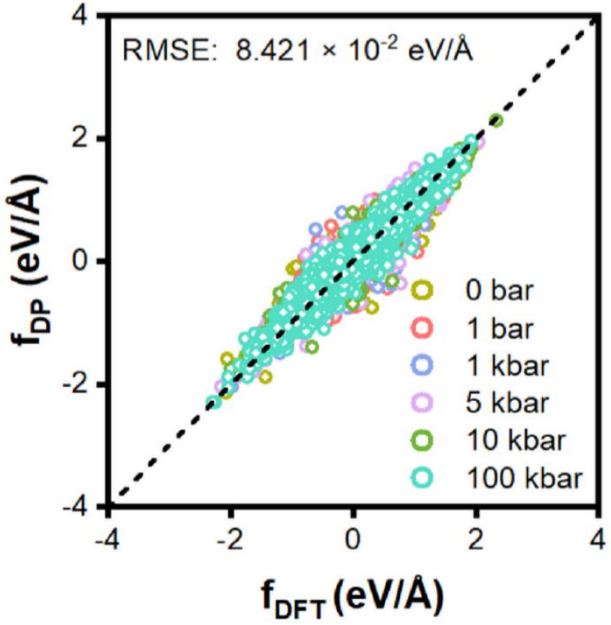
Some examples of *bad* data visualization

Too many points overlapping!
This one is unfortunately very common in scientific papers...



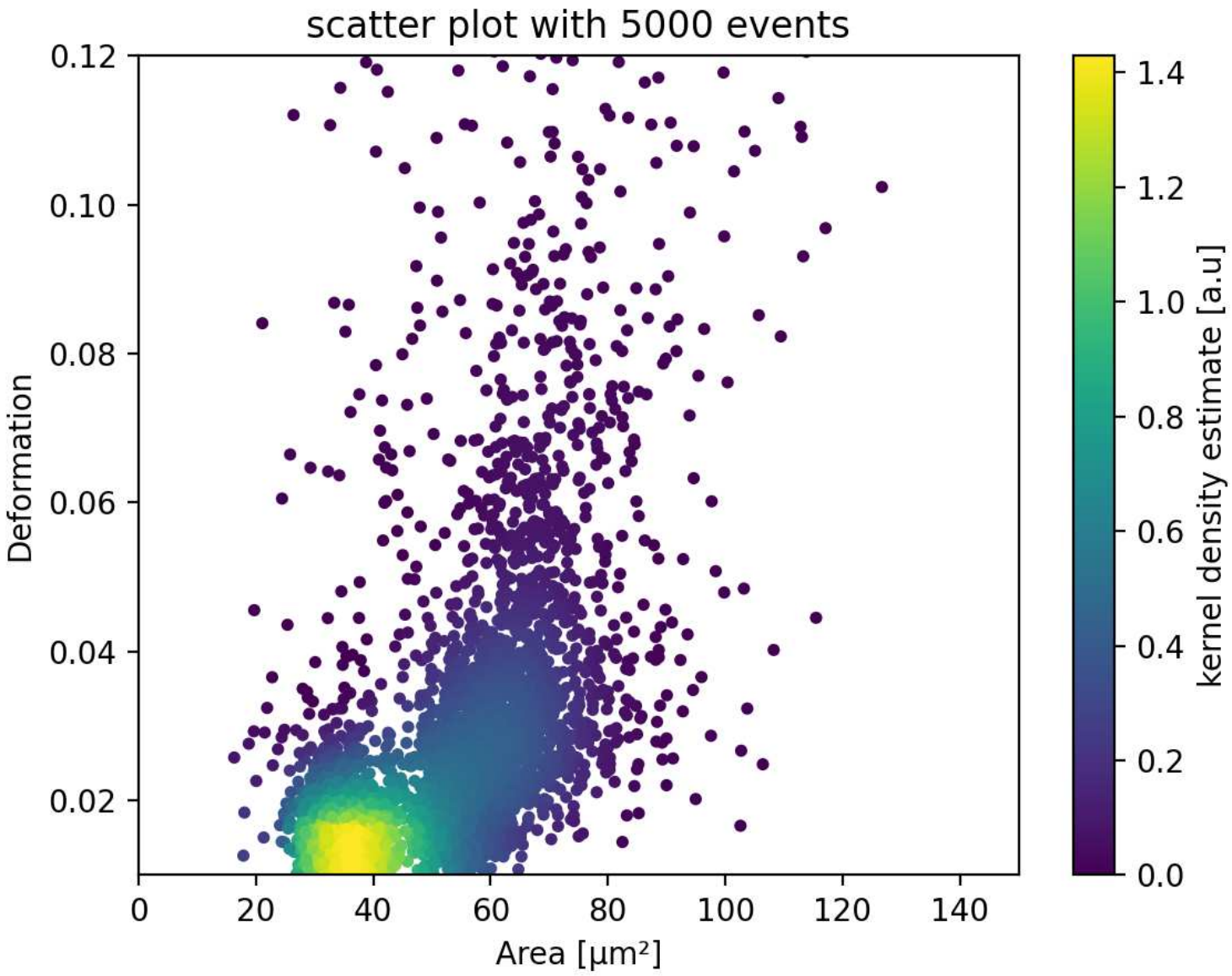
Which method is closest to the parity line? Unclear...





What effect does pressure have on this data? Again, unclear...

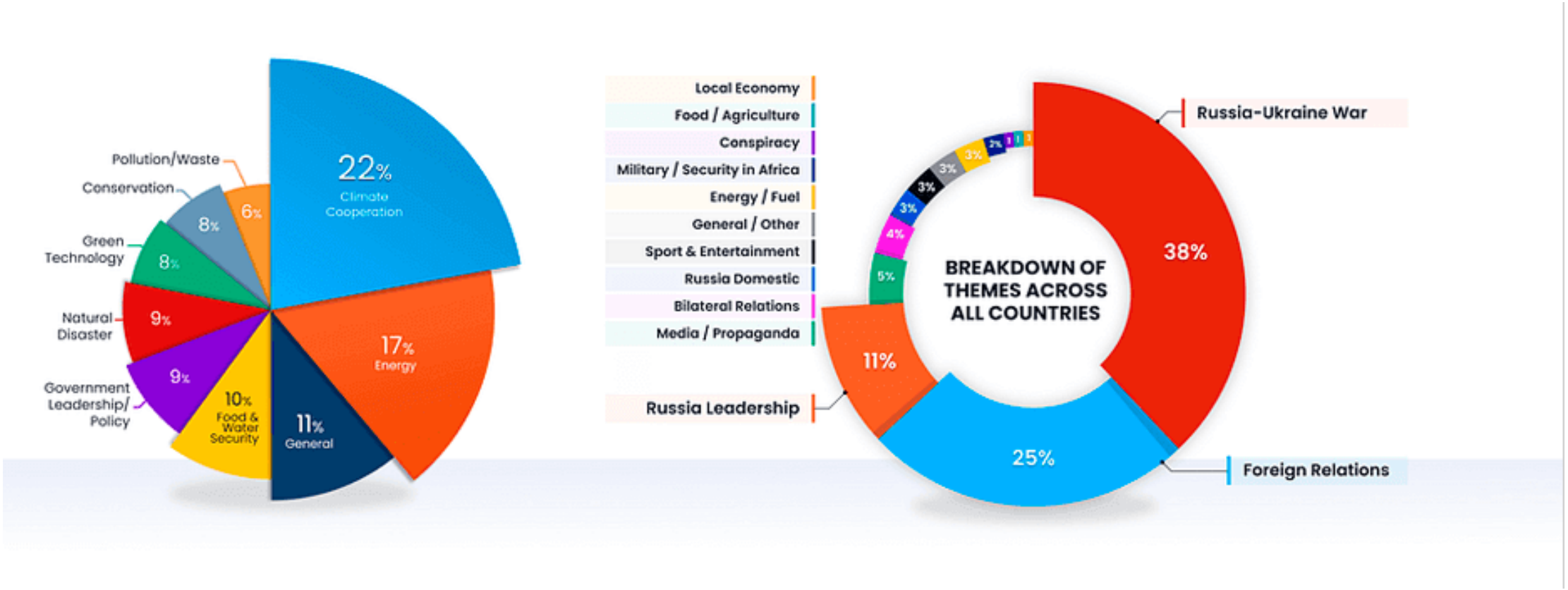
Some examples of *good* data visualization



This is the proper way to visualize scatter plots with high densities!

Point density is clearly represented by color

Some examples of *good* data visualization



- The wedge size is directly related to the percentage it represents
- Labels are placed near their corresponding slices
- The figure on the right is perhaps still a bit complicated…

Some examples of *good* data visualization



Use location when it makes sense

This visual is much more effective than a simple list/table of the data

Questions?