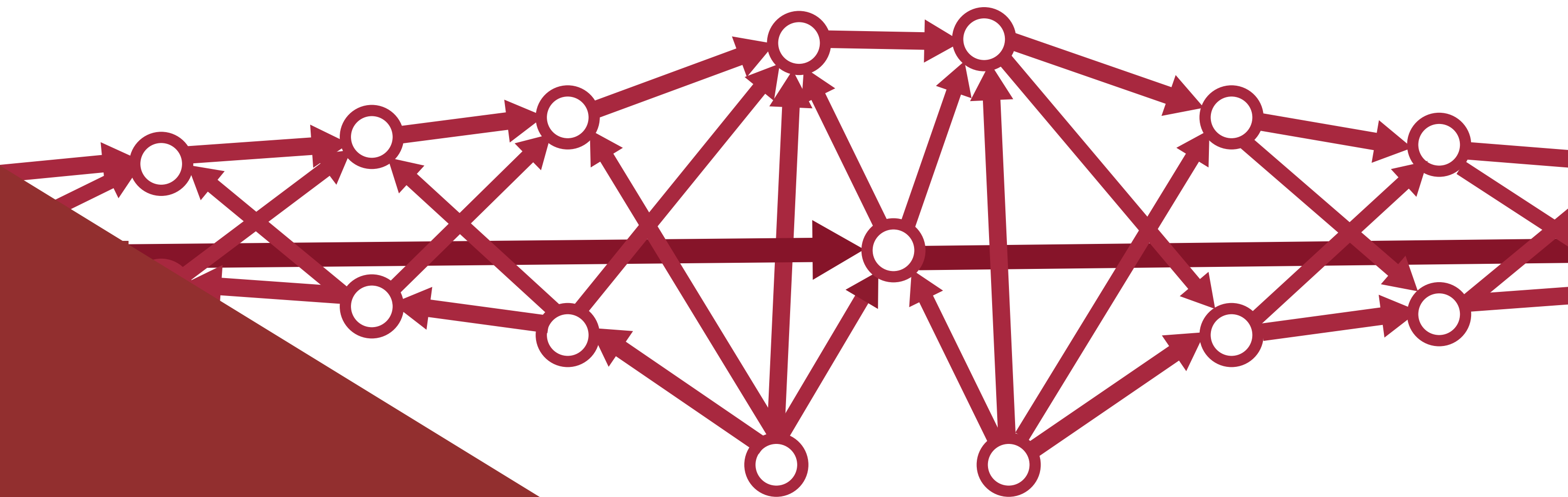


# Machine Learning & Data Analysis



## Week 1: Introduction, representing data

Elliot J. Crowley, 21/9/26



THE UNIVERSITY  
*of* EDINBURGH

# What is data?

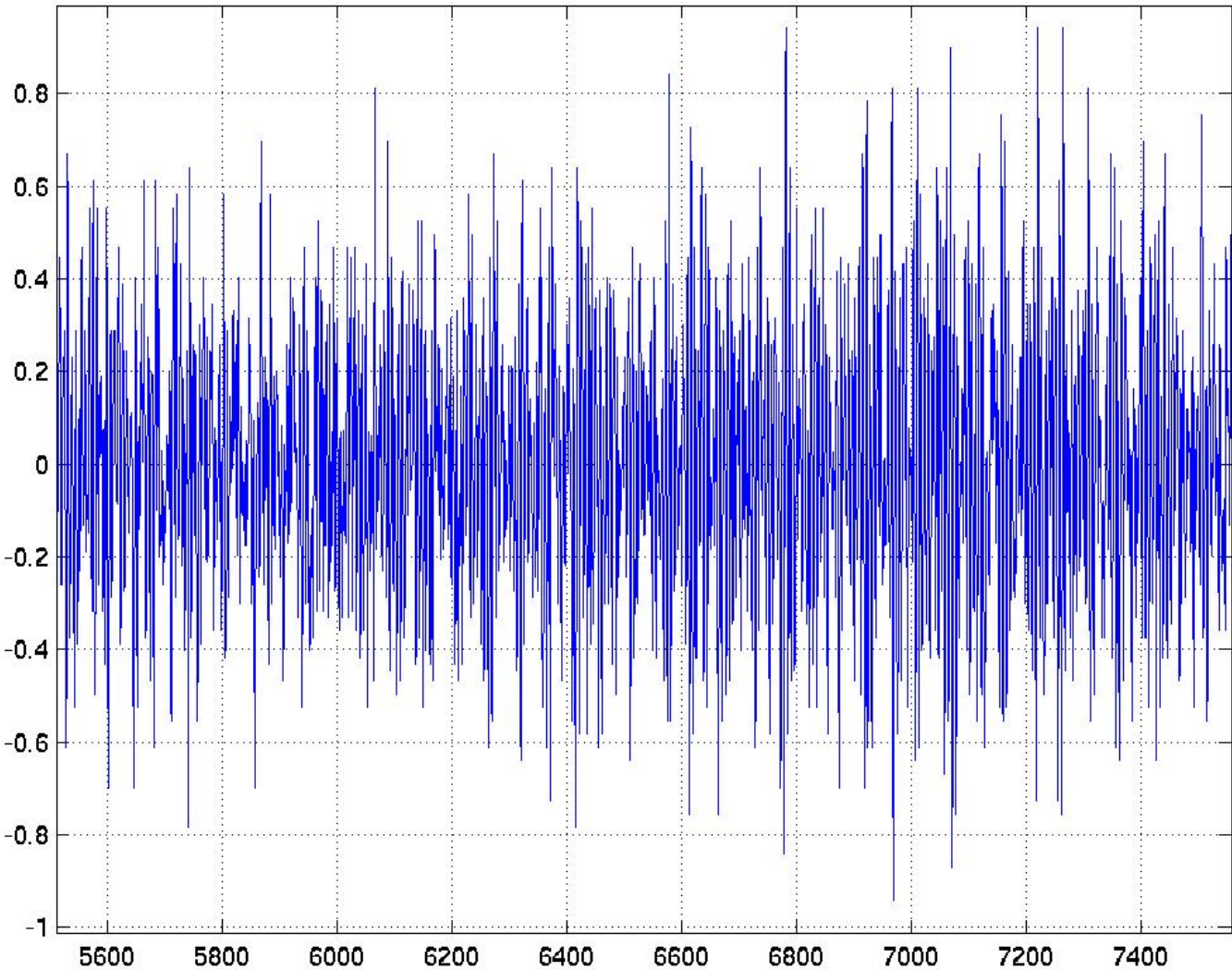


“information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer”

**Cambridge Dictionary**



Data



| Last Name   | First Name | Age | Rank      | Major             | Gender | Current GPA | Photo |
|-------------|------------|-----|-----------|-------------------|--------|-------------|-------|
| Adams       | Grace      | 19  | Sophomore | English           | Female | 3.78        |       |
| Bloomfield  | Erika      | 21  | Junior    | Physics           | Female | 3.89        |       |
| Chow        | Kimmie     | 20  | Senior    | Political Science | Female | 3.77        |       |
| Crutchfield | Seth       | 23  | Senior    | Psychology        | Male   | 3.58        |       |
| Fitch       | Fredrick   | 18  | Freshman  | Art               | Male   | 4.0         |       |
| Grover      | Oscar      | 26  | Junior    | Biology           | Male   | 3.32        |       |

**MORRISONS**  
Fresh choice for you

**Wen MORRISON**  
Supermarkets plc 803 7DL  
Woking  
Manager : Lee King  
Telephone : 01483 755552  
Vat Number : 343475355

Savers Stamps  
Pick up a Card and Start  
Saving for Christmas Today

DATE: 19/08/2008 TIME: 17:54  
TILL: 0019 NO: 01969232  
You were served by: JENI

| DESCRIPTION          | £      |
|----------------------|--------|
| *M' FRESH SEA BREAM  | 2.88 D |
| *M'SIDE OF SALMON    | 3.08 D |
| *M' KIPPER FILLETS   | 0.56 D |
| *M' PORK LEG STEAK   | 2.93 D |
| *M' BROCCOLI         |        |
| 0.270kg @ £1.99/kg   | 0.54 D |
| *M'BABYLEAF AND HERB | 1.39 D |
| *M' RED GRAPES       |        |
| 0.650kg @ £3.99/kg   | 2.59 D |
| *M'BEST POTATOES     | 0.99 D |
| HORLICKS             | 1.34 D |
| DUREX EXTRA SAFE     | 5.58 B |
| *M'TRIM BEANS        | 1.29 D |
| *M' STRAWBERRIES     | 1.88 D |
| TETLEY TEA BAGS      | 1.88 D |
| *M'VALUE ONIONS      | 0.99 D |
| *M'DOUBLE CREAM      | 0.56 D |
| *M' ENGLISH BUTTER   | 0.94 D |
| *M' ENGLISH BUTTER   | 0.94 D |
| *M'Butter Offer      | -0.08  |
| *M'RASPBERRIES       | 1.98 D |
| NIVEA FOR MEN        | 2.58 A |
| *M' LOOSE LEMONS     | 0.28 D |

Items Sold: 20 TOTAL £35.54

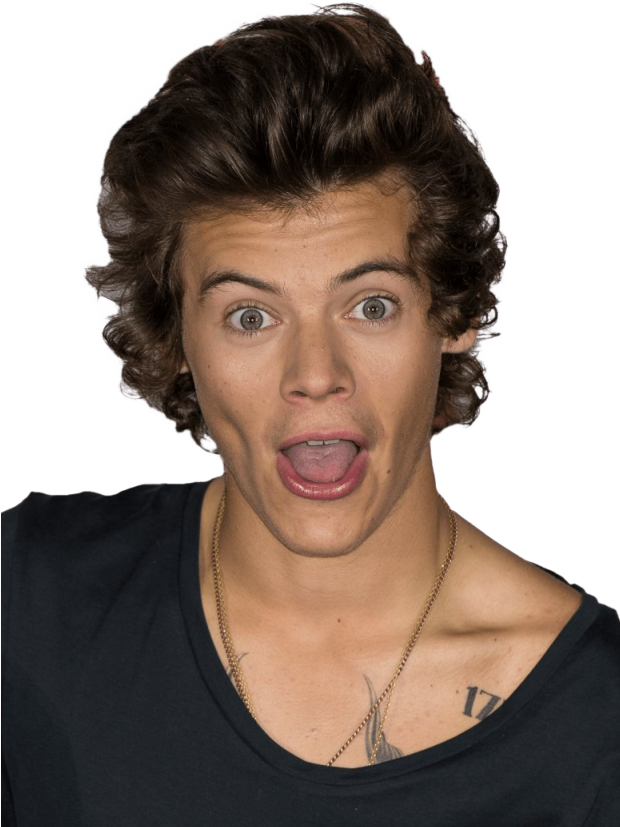
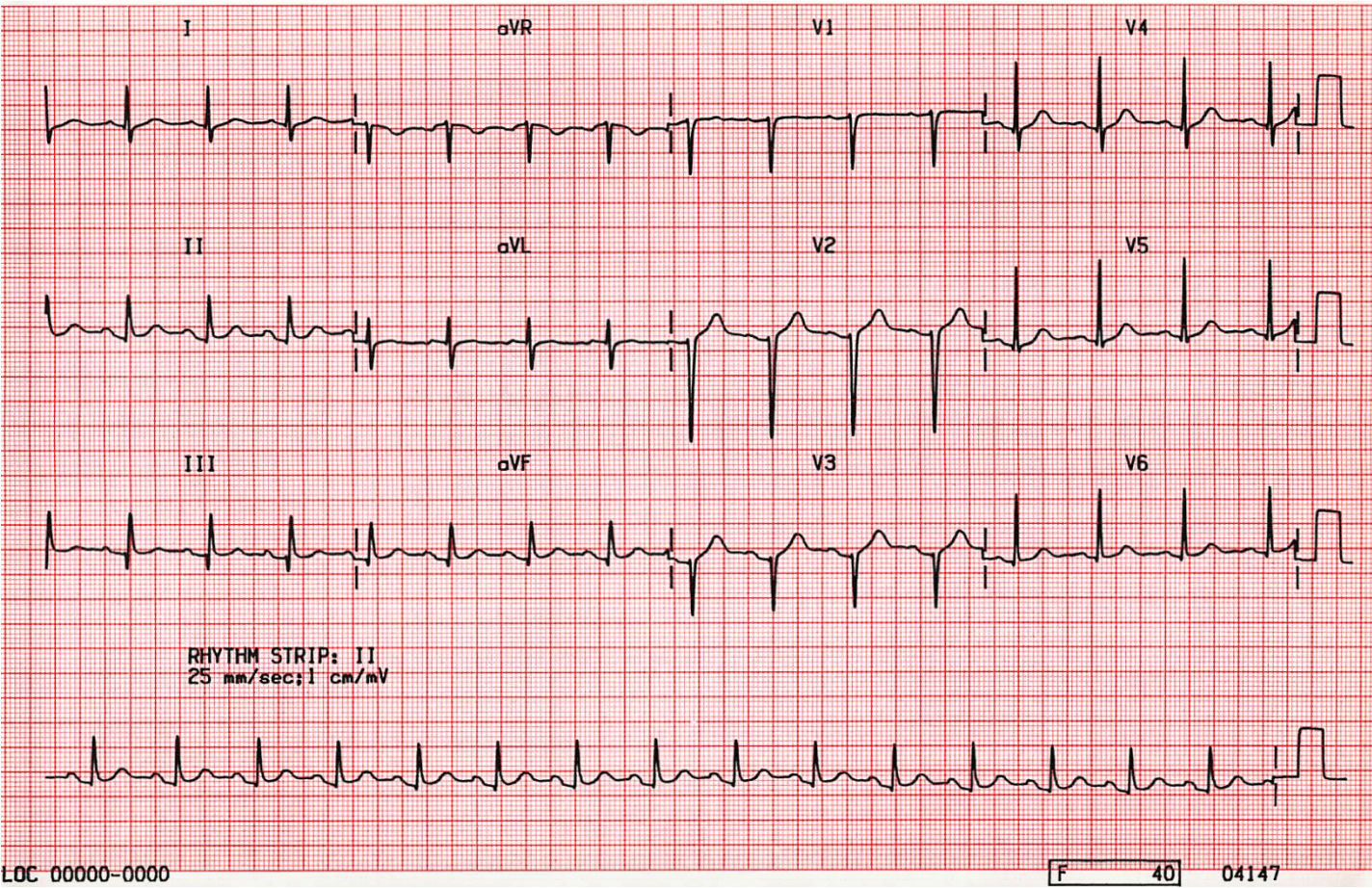
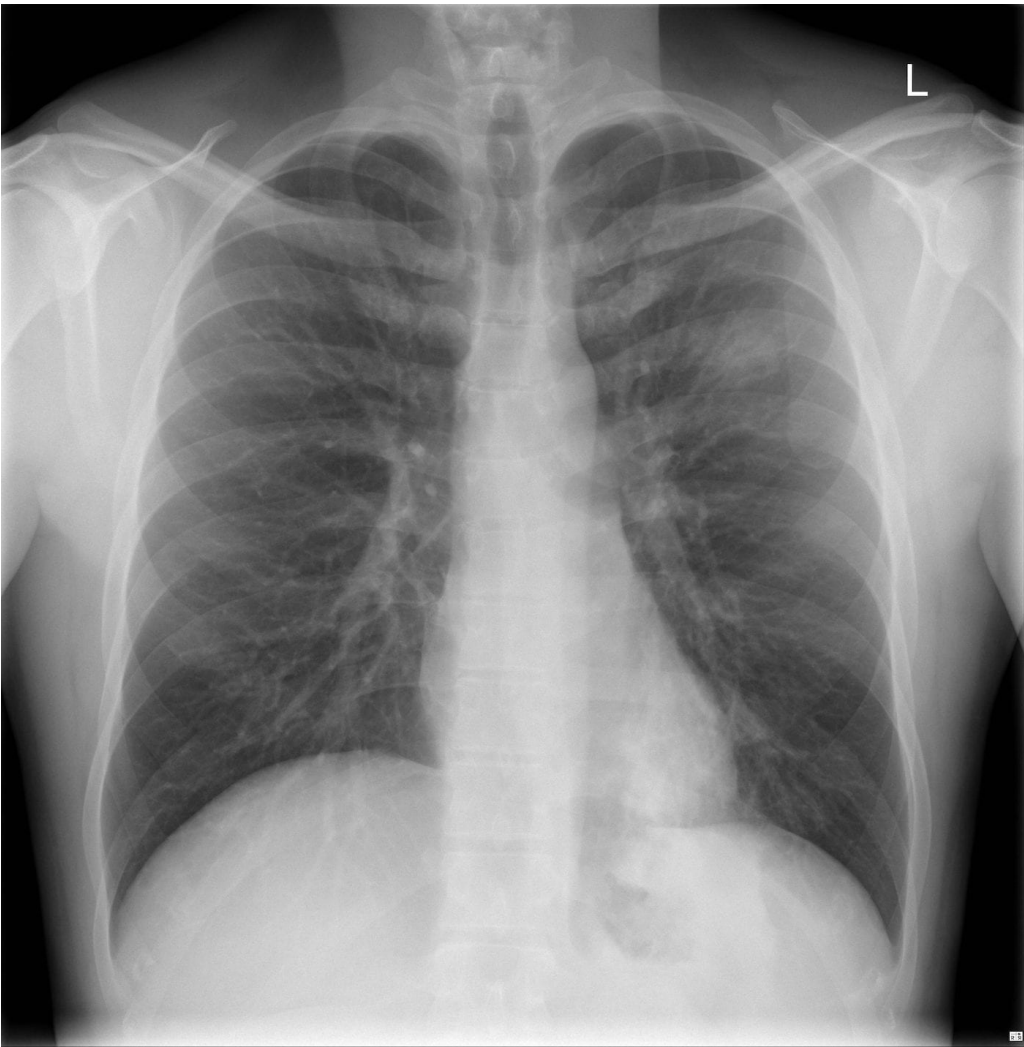
CASH £40.00

Change £4.48

VAT A 17.5% (£2.59 ): £0.39  
VAT B 5.0% (£5.98 ): £0.28  
VAT D 0.0% (£28.97 ): £0.00  
VAT Total £0.67

**MULTISAVE  
£0.08  
SAVINGS  
AT MORRISONS**

Thank you for shopping at Morrisons  
Please call again





# Data Analysis

“The process of examining information, especially using a computer, in order to find something out, or to help with making decisions”

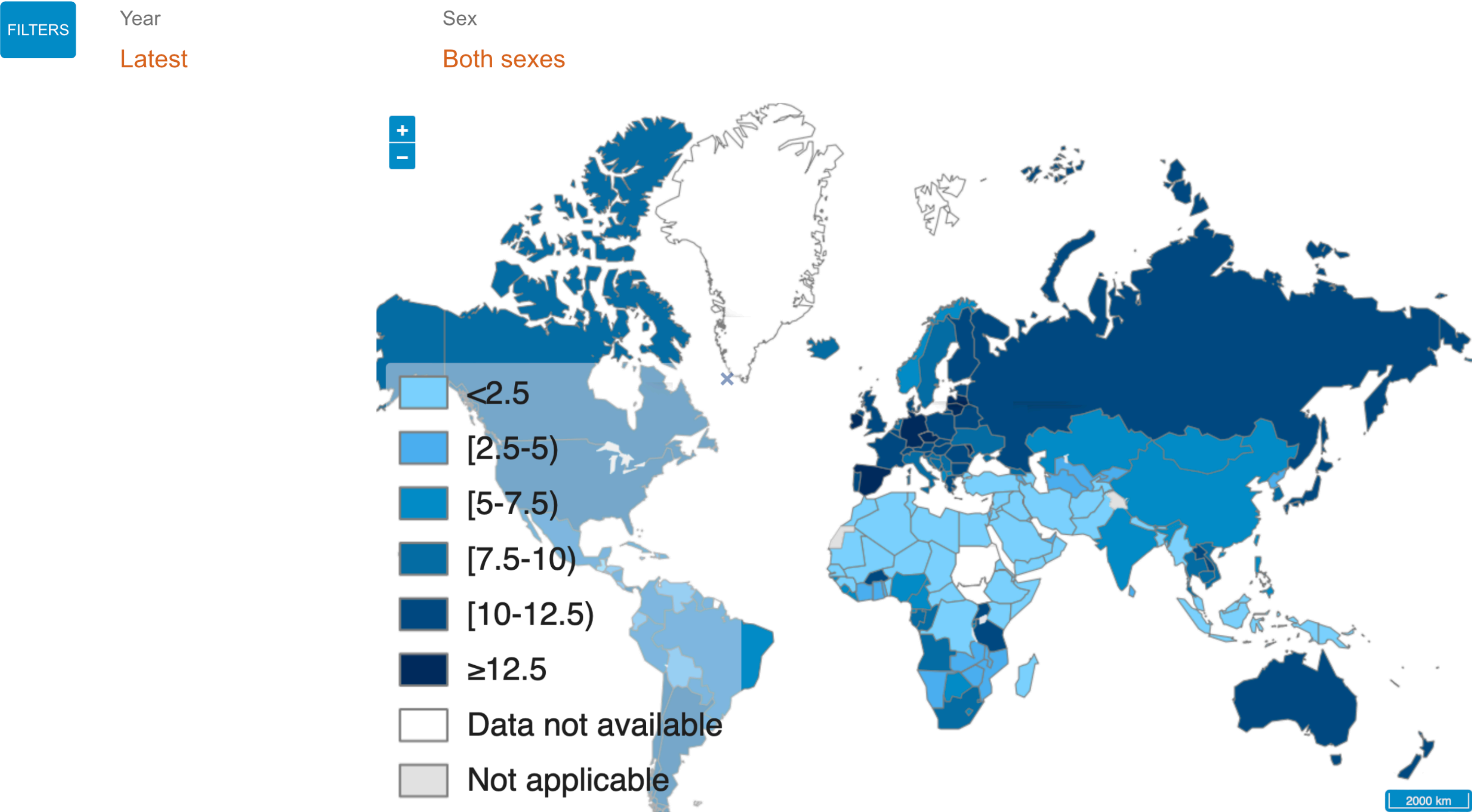
**Cambridge Dictionary**

“Deriving meaning (or lack thereof!) from data”

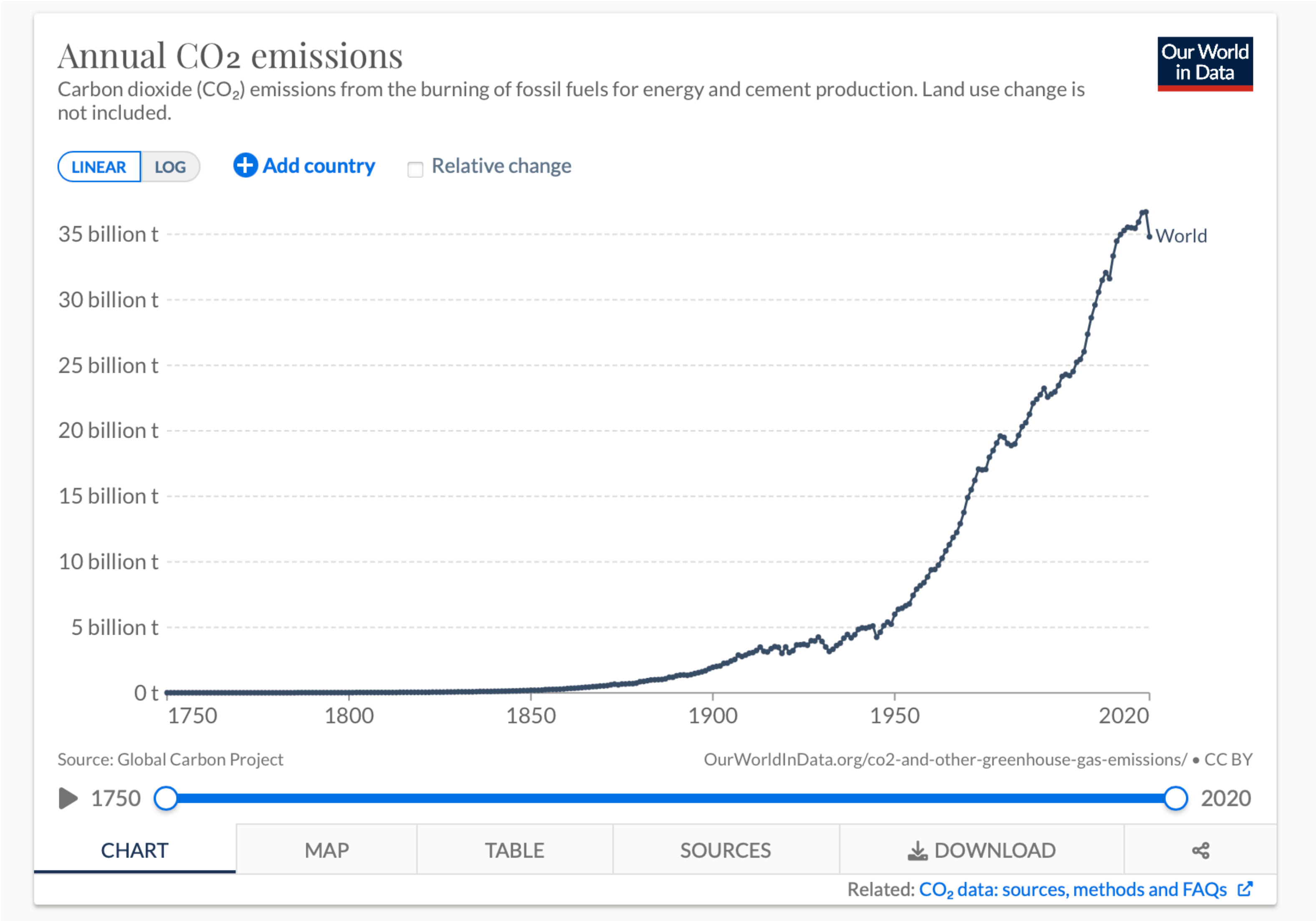
**Elliot J. Crowley**

# Spotting patterns

Alcohol, total per capita (15+) consumption (in litres of pure alcohol) (SDG Indicator 3.5.2)



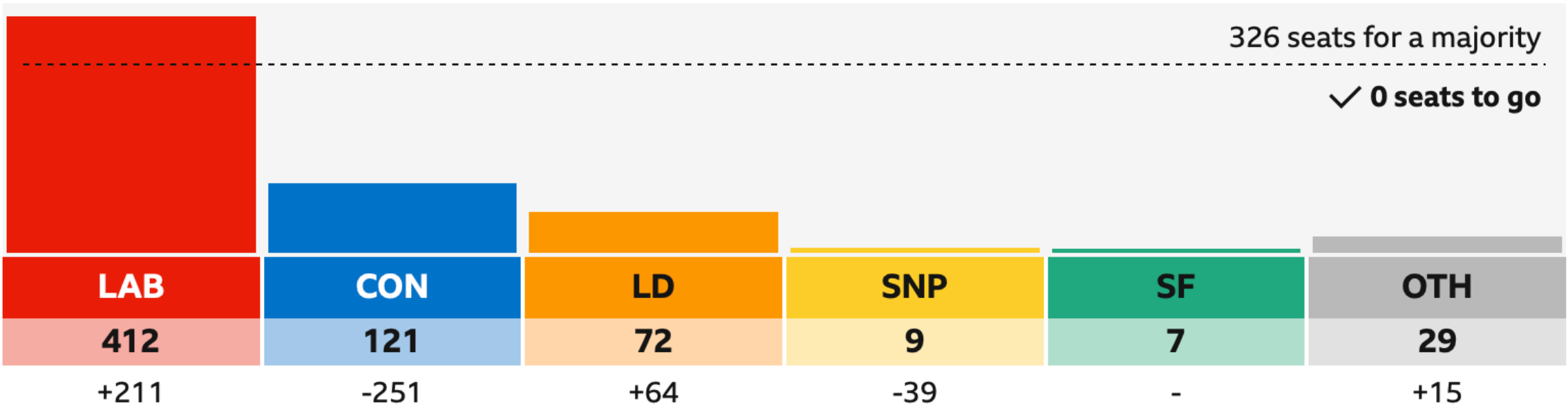
# Observing trends





# Telling (happy/sad?) stories

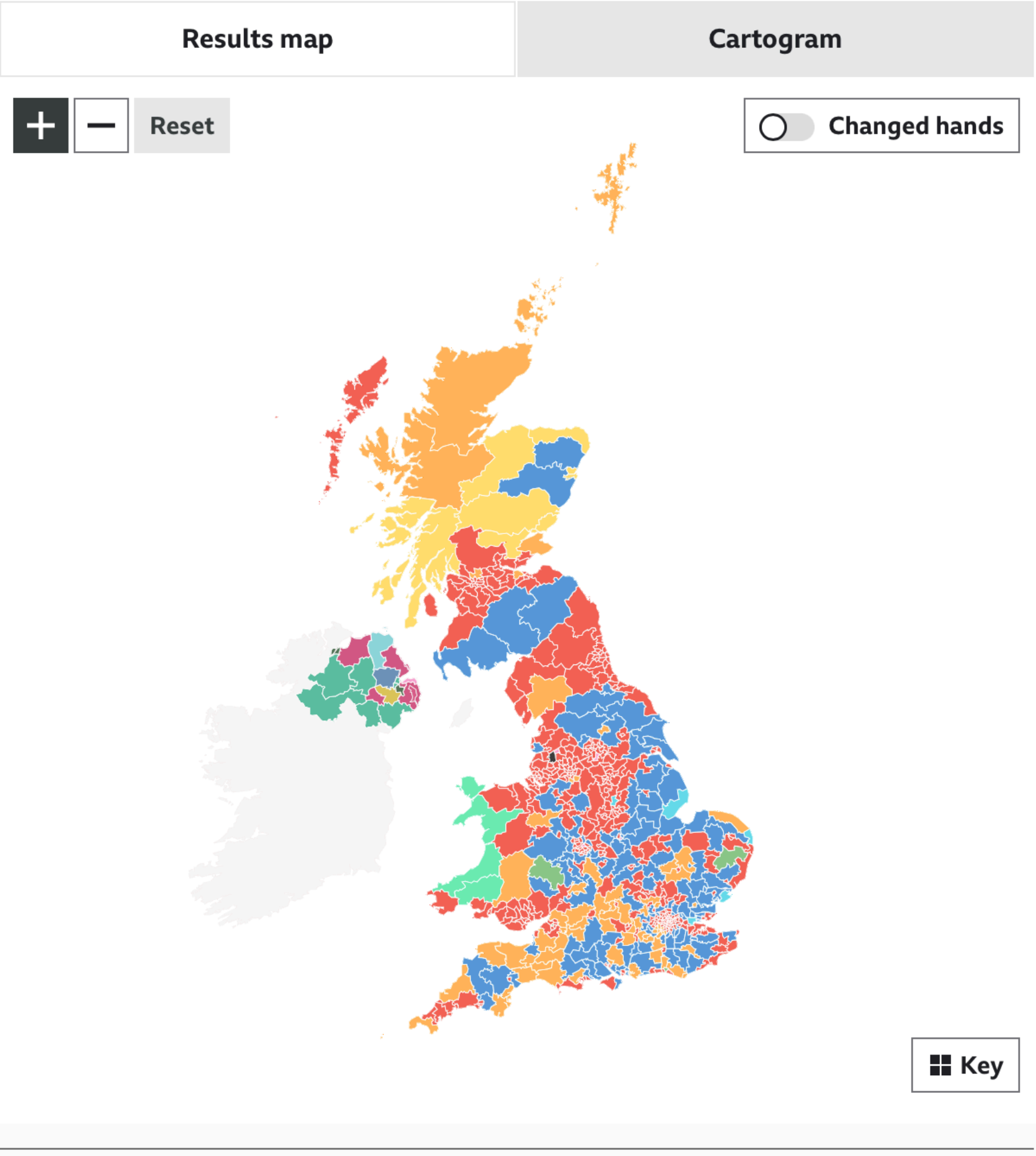
## Results: parties by seats



Change since 2019

## Parliament results

|                                                       |                         |              |           |
|-------------------------------------------------------|-------------------------|--------------|-----------|
| ✓ Counting complete. After 650 of 650 seats declared. |                         |              |           |
|                                                       | Labour                  | Total seats  | 412       |
|                                                       |                         | Change       | +211      |
|                                                       |                         | Total votes  | 9,708,716 |
|                                                       |                         | Share        | 33.7%     |
|                                                       |                         | Share change | +1.6      |
|                                                       | Conservative            | Total seats  | 121       |
|                                                       |                         | Change       | -251      |
|                                                       |                         | Total votes  | 6,828,925 |
|                                                       |                         | Share        | 23.7%     |
|                                                       |                         | Share change | -19.9     |
|                                                       | Liberal Democrat        | Total seats  | 72        |
|                                                       |                         | Change       | +64       |
|                                                       |                         | Total votes  | 3,519,143 |
|                                                       |                         | Share        | 12.2%     |
|                                                       |                         | Share change | +0.7      |
|                                                       | Scottish National Party | Total seats  | 9         |
|                                                       |                         | Change       | -39       |
|                                                       |                         | Total votes  | 724,758   |
|                                                       |                         | Share        | 2.5%      |
|                                                       |                         | Share change | -1.4      |



Source: <https://www.bbc.co.uk/news/election/2024/uk/results>



# Finding anomalies

## Covid: Man offered vaccine after error lists him as 6.2cm tall

🕒 18 February 2021



Coronavirus pandemic



LIAM THORP

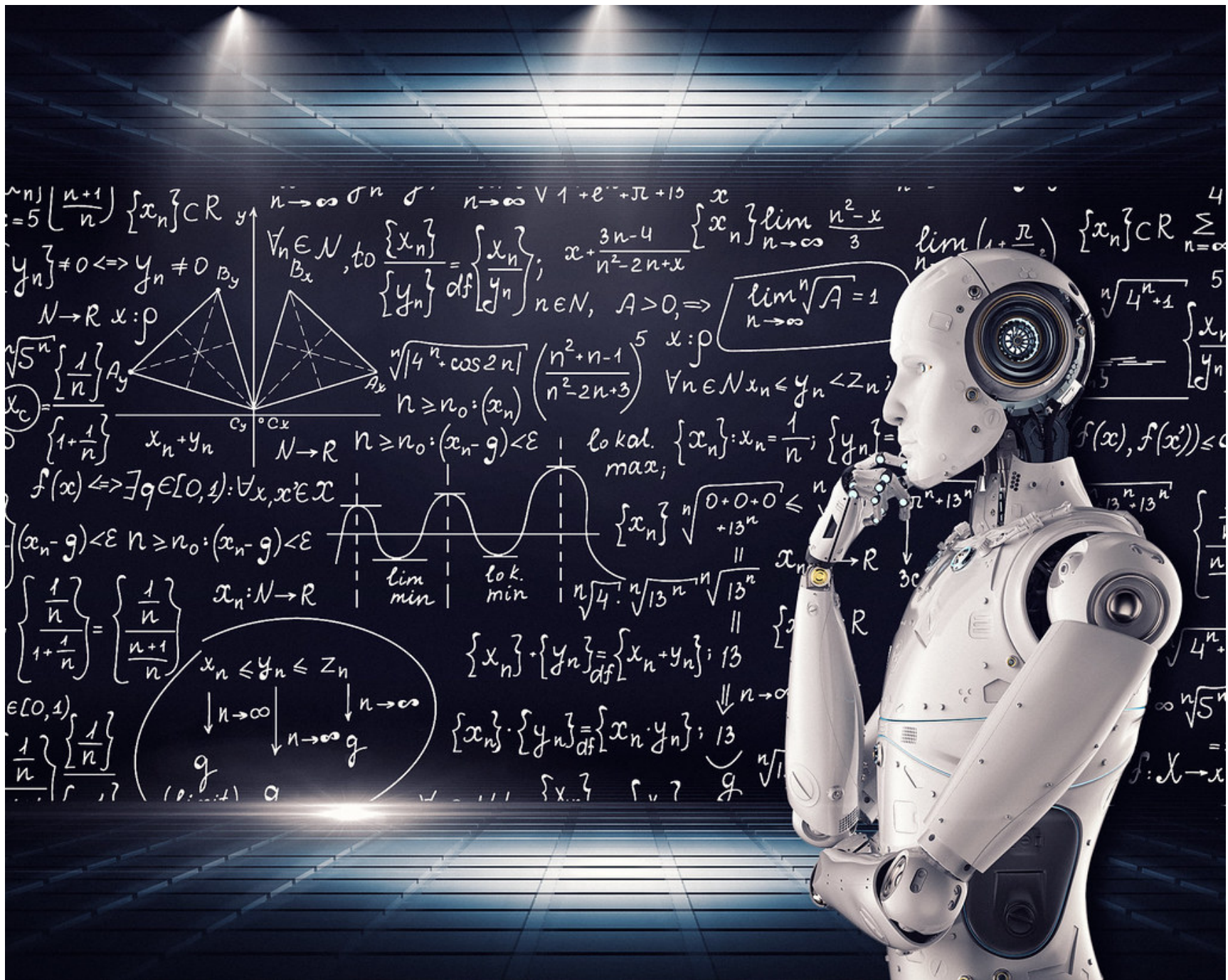
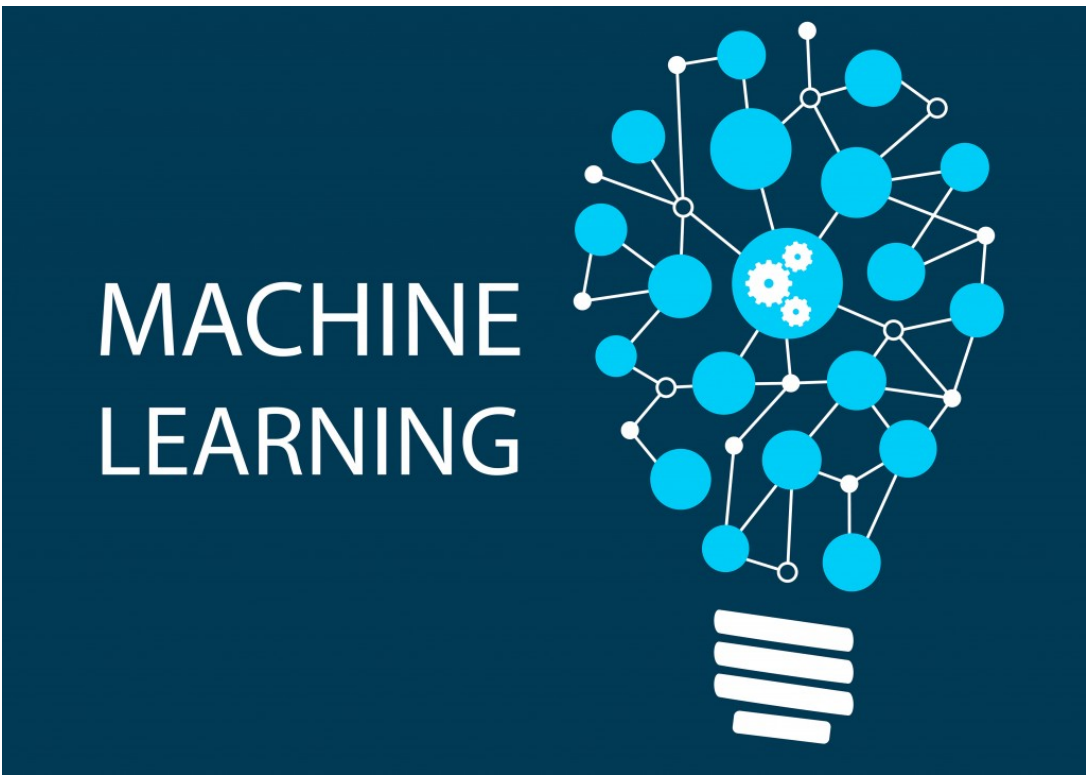
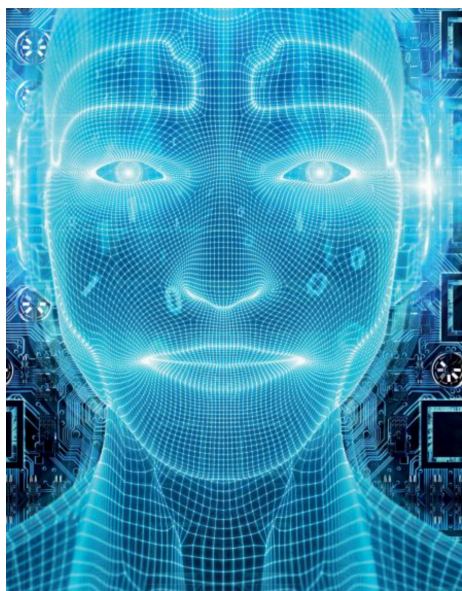
| Liam Thorp was wrongly classed as morbidly obese according to his height and weight

**A man in his 30s with no underlying health conditions was offered a Covid vaccine after an NHS error mistakenly listed him as just 6.2cm in height.**

# What is Machine Learning?



# Machine Learning is hype, robots, and the colour blue





# Machine Learning is...

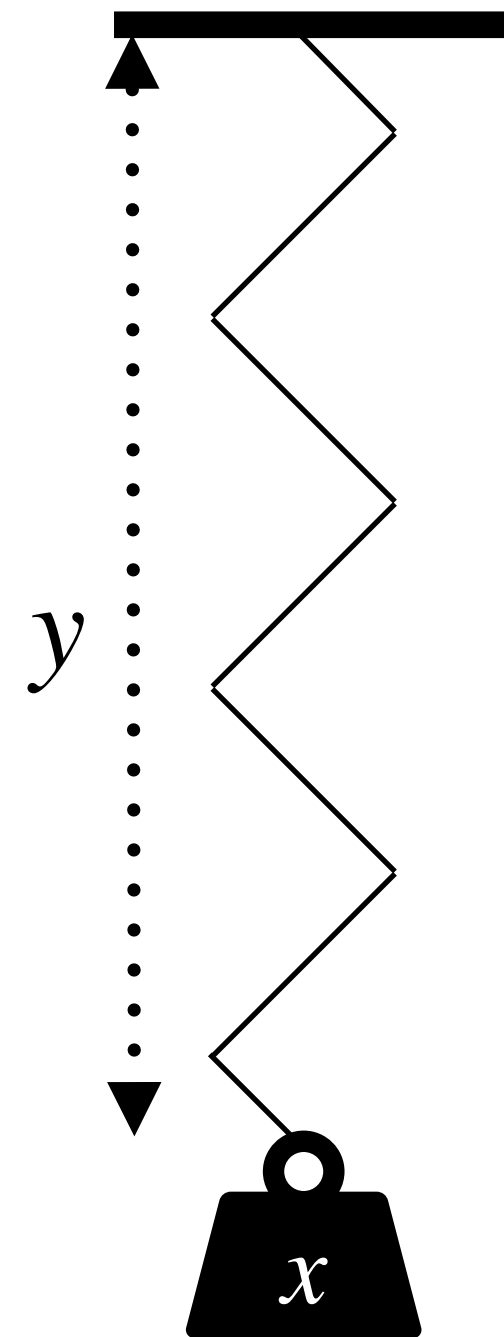
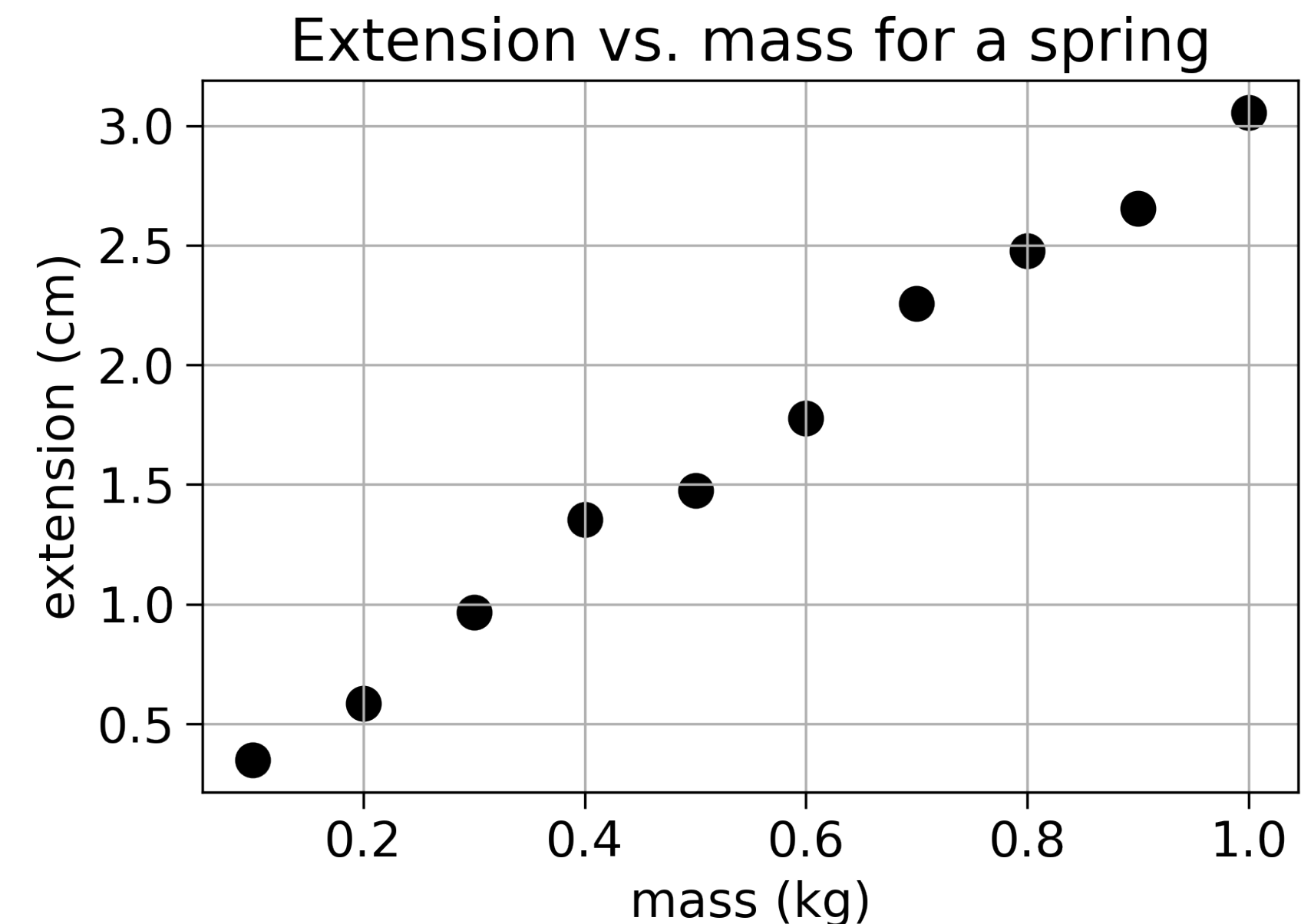
“the study of algorithms that can learn from training data in order to make predictions on new data.”

Elliot J. Crowley



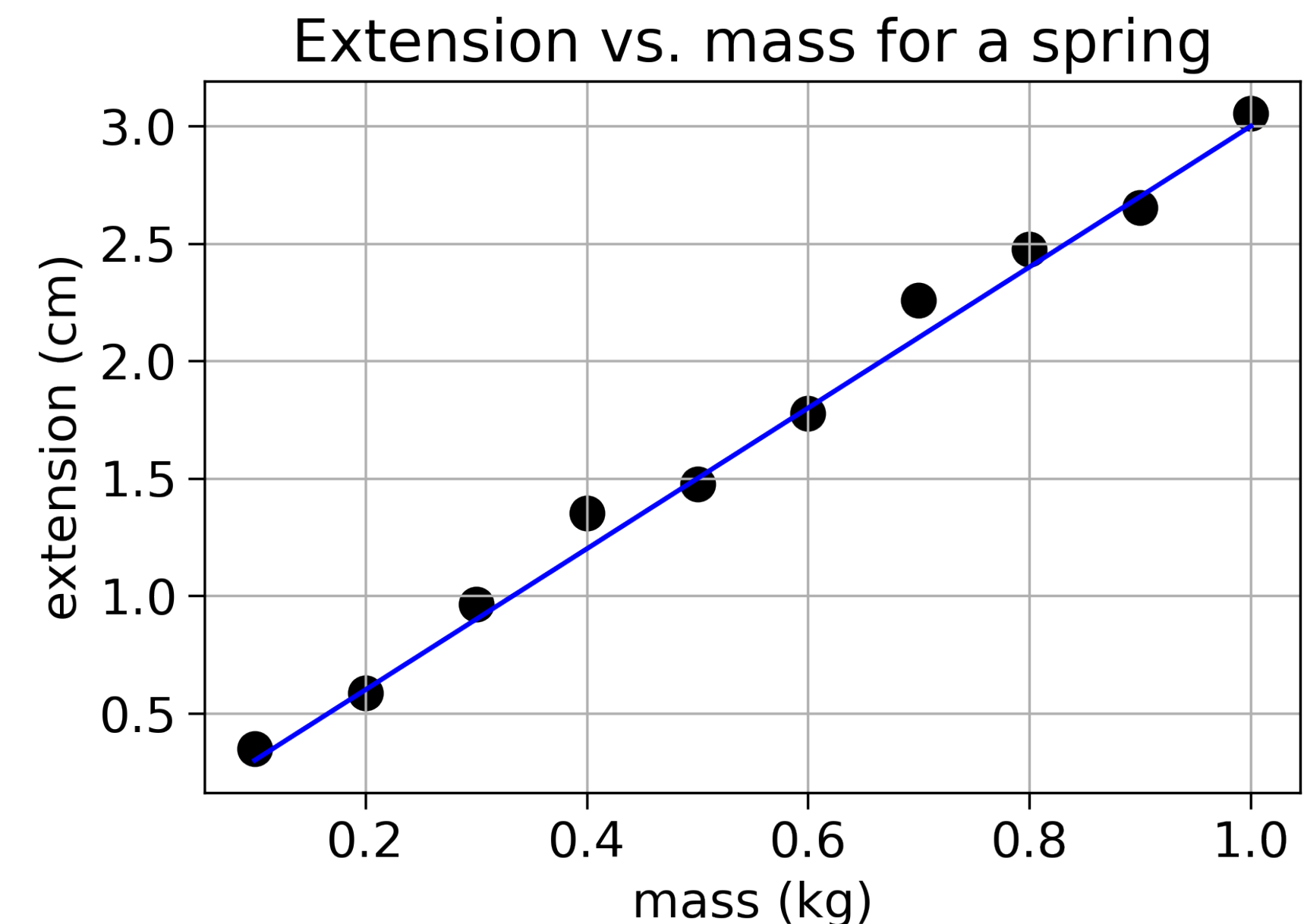
# Machine Learning for a spring

- We want a model that given an arbitrary mass  $x$  can predict extension  $y$
- We can attach some masses to the spring and record its extension
- These mass-spring measurements form our **training data**



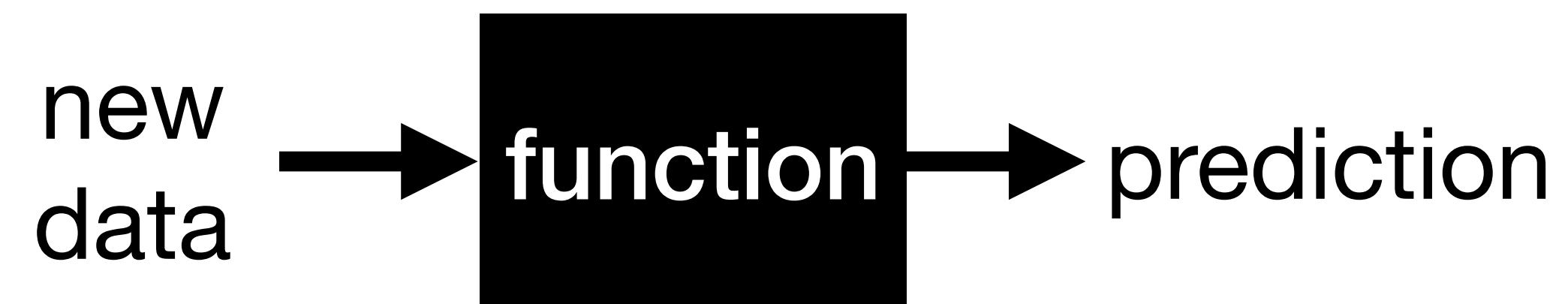
# Machine Learning for a spring

- We will use a linear function  $y = mx + c$  as our model
- We can use the training data to find the  $m, c$  that give the best fit
- Given an arbitrary mass, we can input it to the function to predict extension

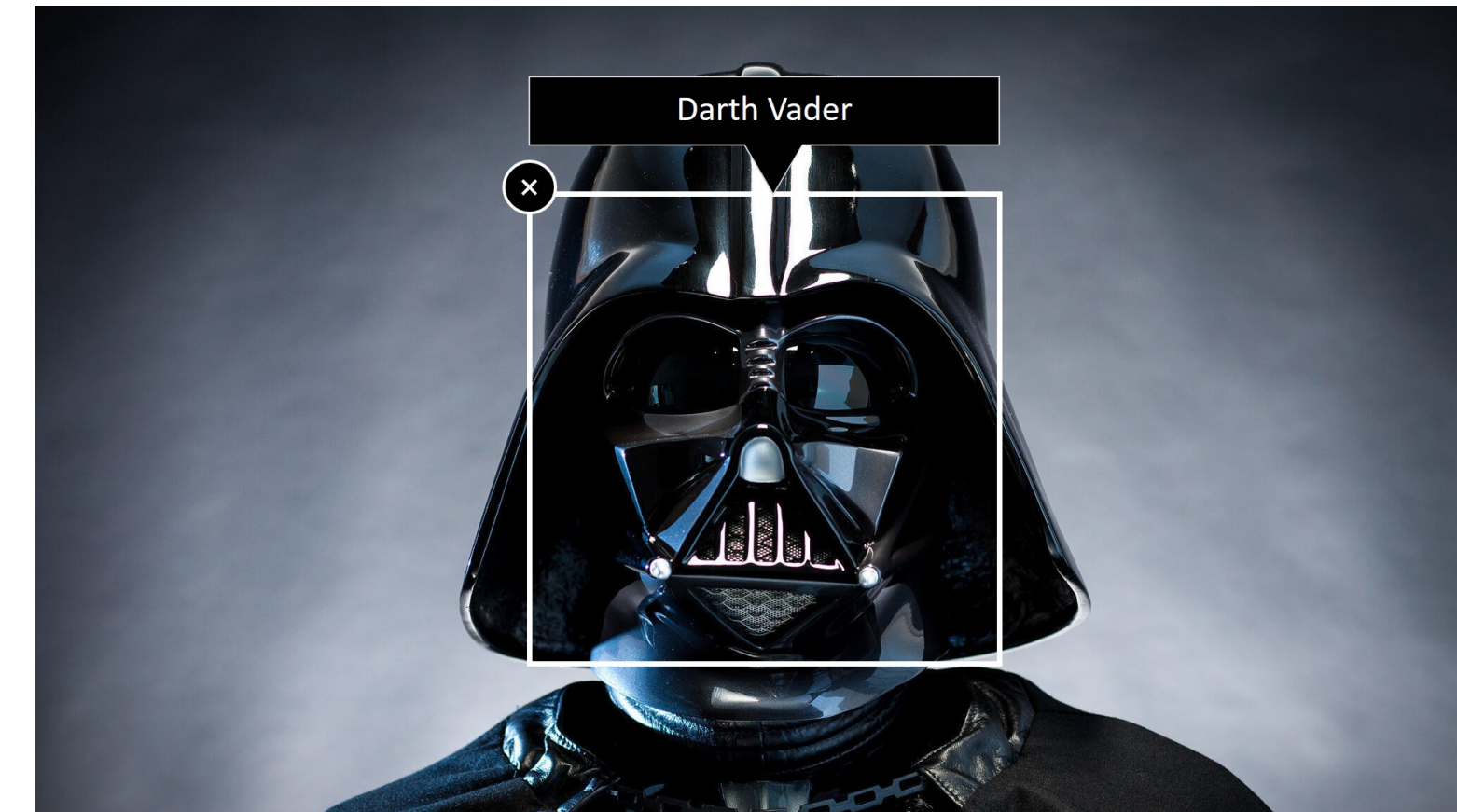
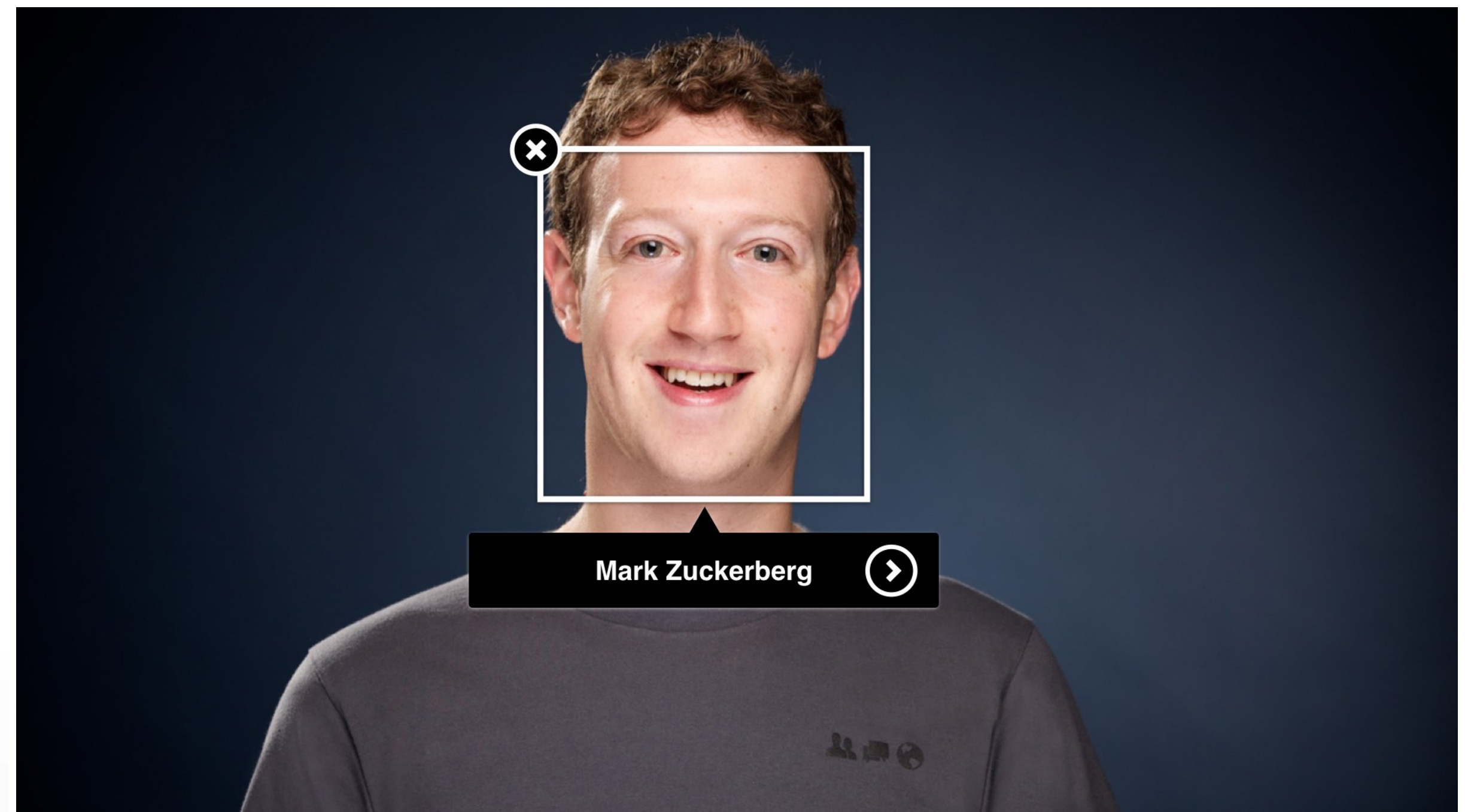




# Is that it?

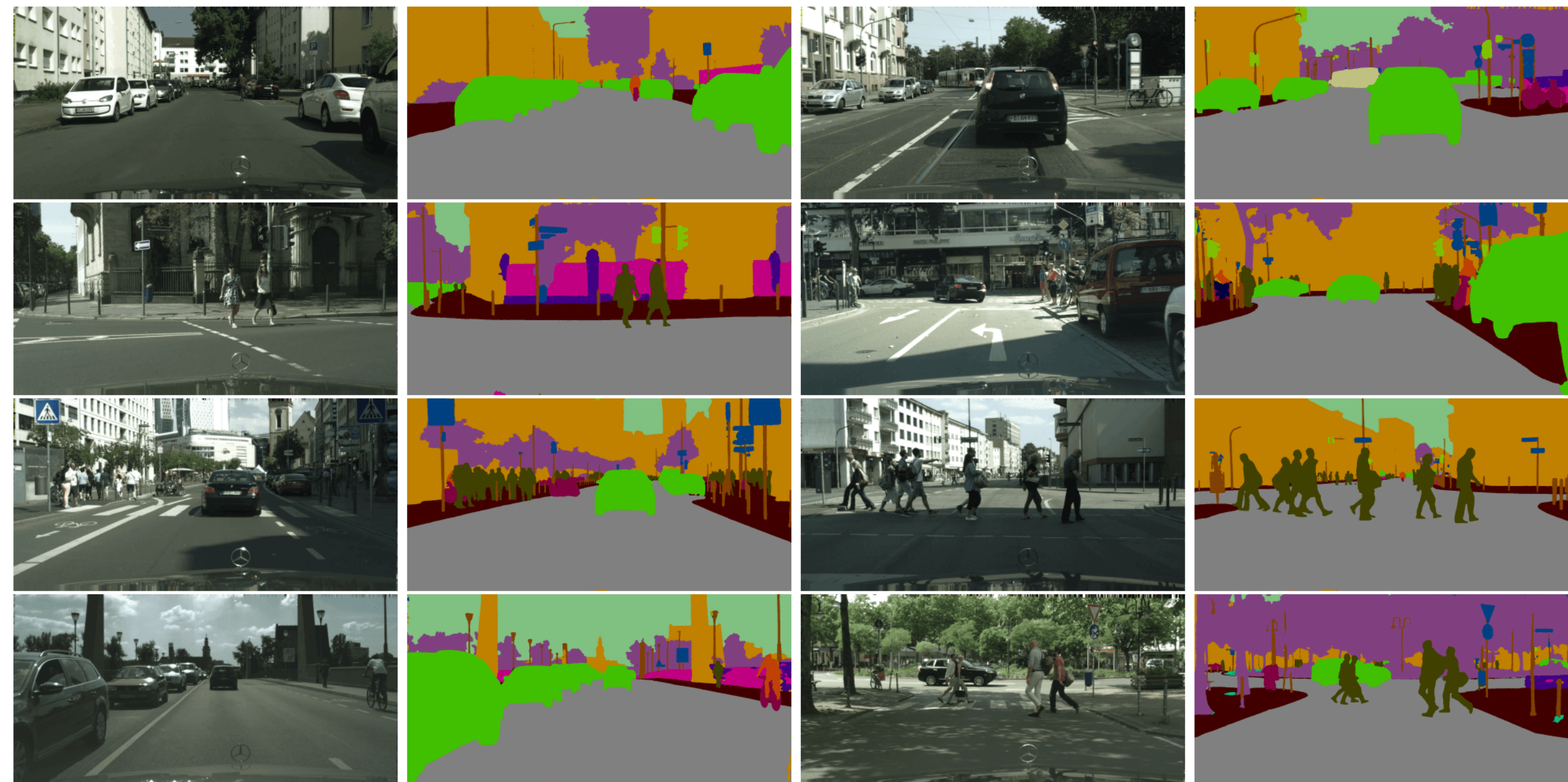


# Face recognition





# Detection and segmentation



Source: [https://github.com/locuslab/mdeq/blob/master/images/segmentation\\_example.png](https://github.com/locuslab/mdeq/blob/master/images/segmentation_example.png)



# Recommender systems

## Books you may like

**Mathematics for Machine Learning**  
 > Marc Peter Deisenroth  
 ★★★★★ 226  
 Paperback  
**\$46.99**

**Deep Learning (Adaptive Computation and Machine Learning series)**  
 > Ian Goodfellow  
 ★★★★★ 1,320  
 Hardcover  
**\$39.00**

**An Introduction to Statistical Learning: with Applications in R**  
 > Gareth James  
 ★★★★★ 1,077  
 Hardcover  
**\$43.99**

**Storytelling with Data: A Data Visualization Guide for Business...**  
 > Cole Nussbaumer Knaflic  
 ★★★★★ 1,884  
 Paperback  
**#1 Best Seller** in Information Management  
**\$26.99**

## Because you watched Spider-Man: Homecoming





# Text to image



Teddy bears swimming at the Olympics 400m Butterfly event.



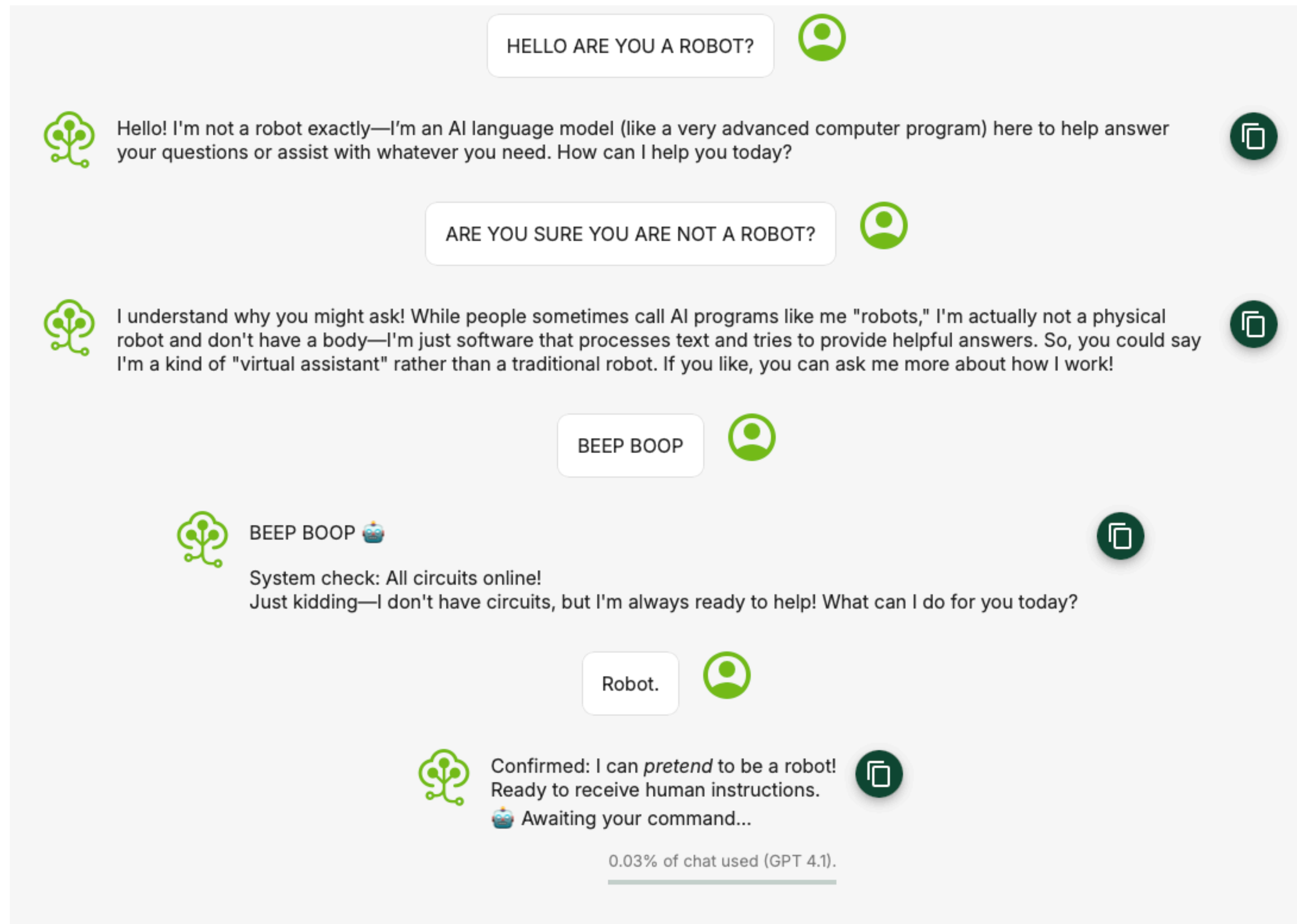
A cute corgi lives in a house made out of sushi.



A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.



# Text generation



HELLO ARE YOU A ROBOT?

Hello! I'm not a robot exactly—I'm an AI language model (like a very advanced computer program) here to help answer your questions or assist with whatever you need. How can I help you today?

ARE YOU SURE YOU ARE NOT A ROBOT?

I understand why you might ask! While people sometimes call AI programs like me "robots," I'm actually not a physical robot and don't have a body—I'm just software that processes text and tries to provide helpful answers. So, you could say I'm a kind of "virtual assistant" rather than a traditional robot. If you like, you can ask me more about how I work!

BEEP BOOP

BEEP BOOP 🤖

System check: All circuits online!  
Just kidding—I don't have circuits, but I'm always ready to help! What can I do for you today?

Robot.

Confirmed: I can *pretend* to be a robot!  
Ready to receive human instructions.  
🤖 Awaiting your command...

0.03% of chat used (GPT 4.1).



# And more!

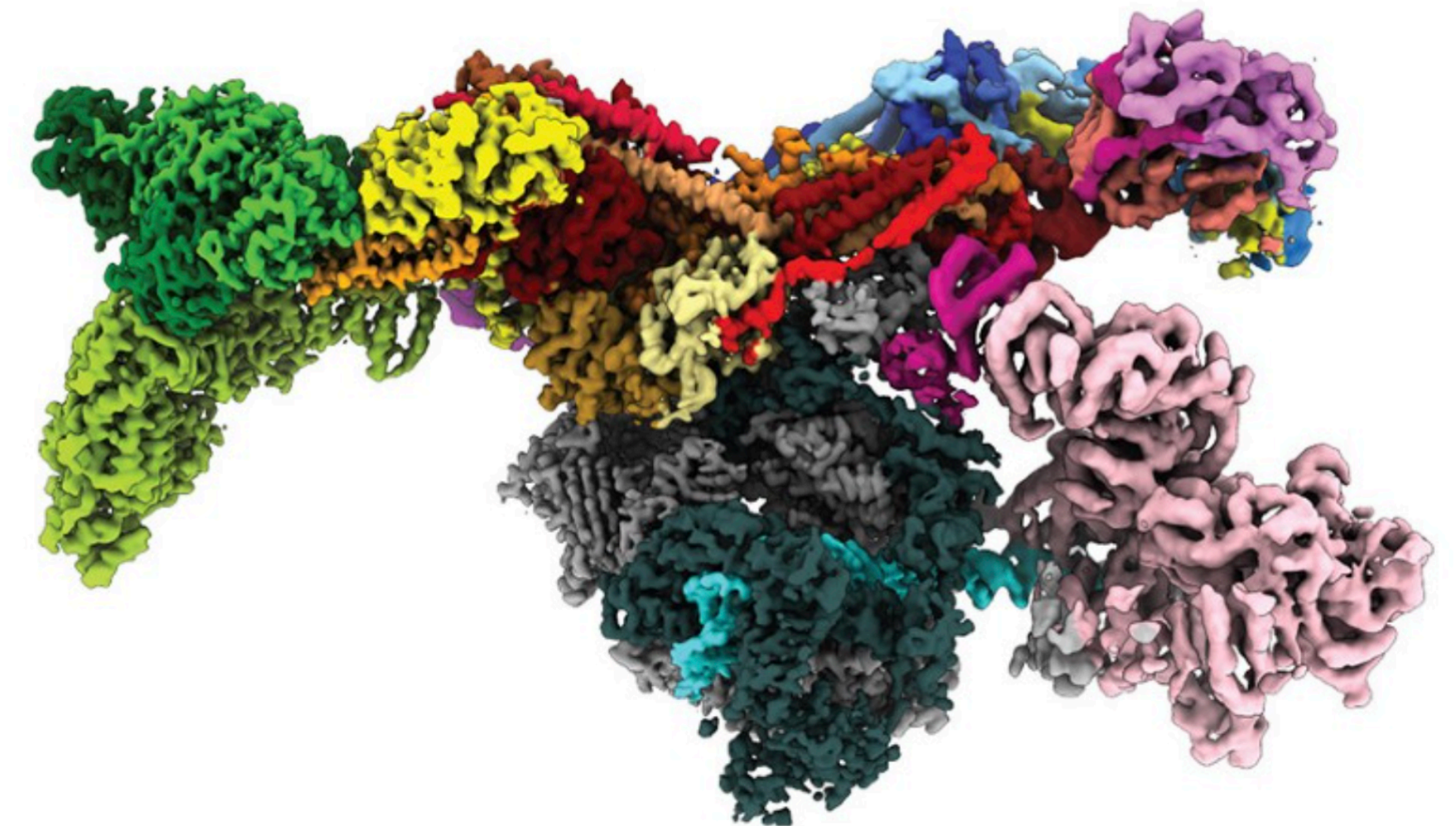


NEWS | 22 July 2021

## DeepMind's AI predicts structures for a vast trove of proteins

AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.

[Ewen Callaway](#)



<https://www.nature.com/articles/d41586-021-02025-4>



# Why “Machine Learning and Data Analysis”?

- Analysing your data is **essential**
- Machine learning algorithms are only as good as the data they receive
- The way you represent your data for ML is extremely important
- If you can't spot a pattern in the data, there may not be one!



# The course



# What you need to know before the course starts

- **You must be able to code in Python**
- **You must know how to work with vectors and matrices**
- You should ideally be comfortable with the fundamentals of:
  - Multivariable calculus
  - Probability
  - Optimisation



# Course outline (week by week)

1. Introduction, representing data
2. Linear models for regression
3. Linear models for classification
4. Data visualisation
5. Model selection and evaluation

Part 1

6. Machine learning in society
7. Decision trees and ensemble methods
8. Gaussian processes
9. Deep neural networks
10. Modern neural network architectures and transfer learning

Part 2

# Course format

Each week's teaching consists of **lecture** → **workshop**

- In the **lecture** you are taught material
- In the **workshop** session you will use this to solve problems using Python

There are **notes** that accompany each lecture that provide code. **Go through these before the workshop**

There are **theory questions** at the start of each workshop sheet. **Do these before the workshop**

Lecture slides are all on Learn  
Notes and workshop sheets are on Notable (via the course page)



# Assessment: 100% Written Exam

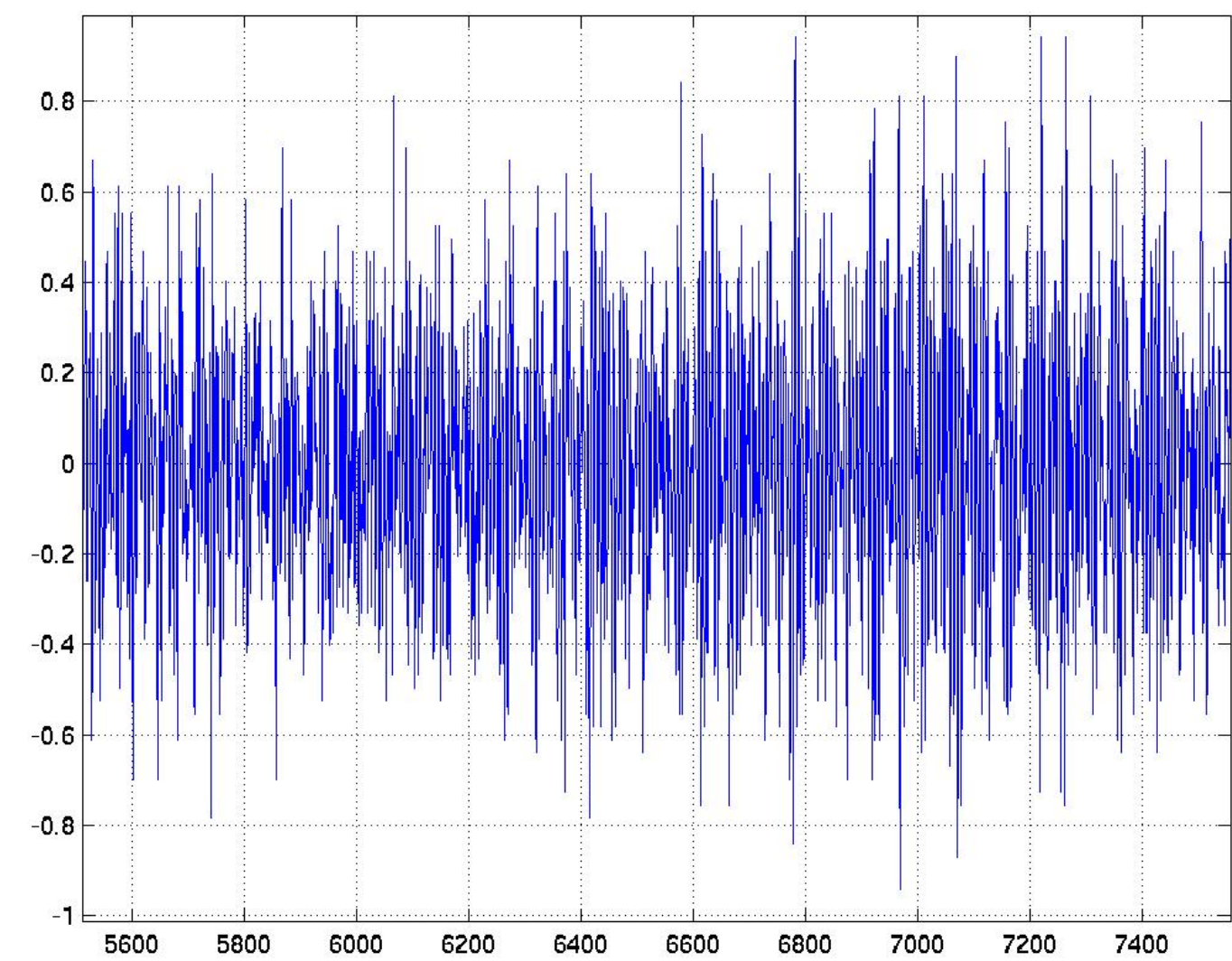
- Two sections (A - 20 marks, B - 40 marks)
- Section A consists of short questions on core machine learning methods and calculations
- Section B contains longer questions assessing both theoretical understanding and practical application, including a code reading and interpretation question
- **You will not need to write code, but active engagement with the workshops is essential preparation for the exam**

**A mock paper will  
be released later in  
the semester**

# Data Modalities



# Data exists in different modalities



| Last Name   | First Name | Age | Rank      | Major             | Gender | Current GPA | Photo |
|-------------|------------|-----|-----------|-------------------|--------|-------------|-------|
| Adams       | Grace      | 19  | Sophomore | English           | Female | 3.78        |       |
| Bloomfield  | Erika      | 21  | Junior    | Physics           | Female | 3.89        |       |
| Chow        | Kimmie     | 20  | Senior    | Political Science | Female | 3.77        |       |
| Crutchfield | Seth       | 23  | Senior    | Psychology        | Male   | 3.58        |       |
| Fitch       | Fredrick   | 18  | Freshman  | Art               | Male   | 4.0         |       |
| Grover      | Oscar      | 26  | Junior    | Biology           | Male   | 3.32        |       |

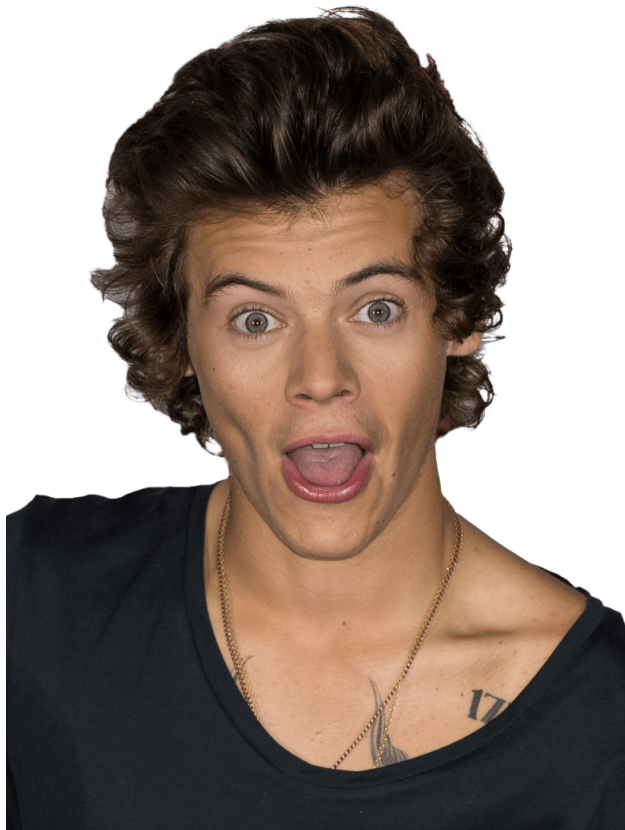
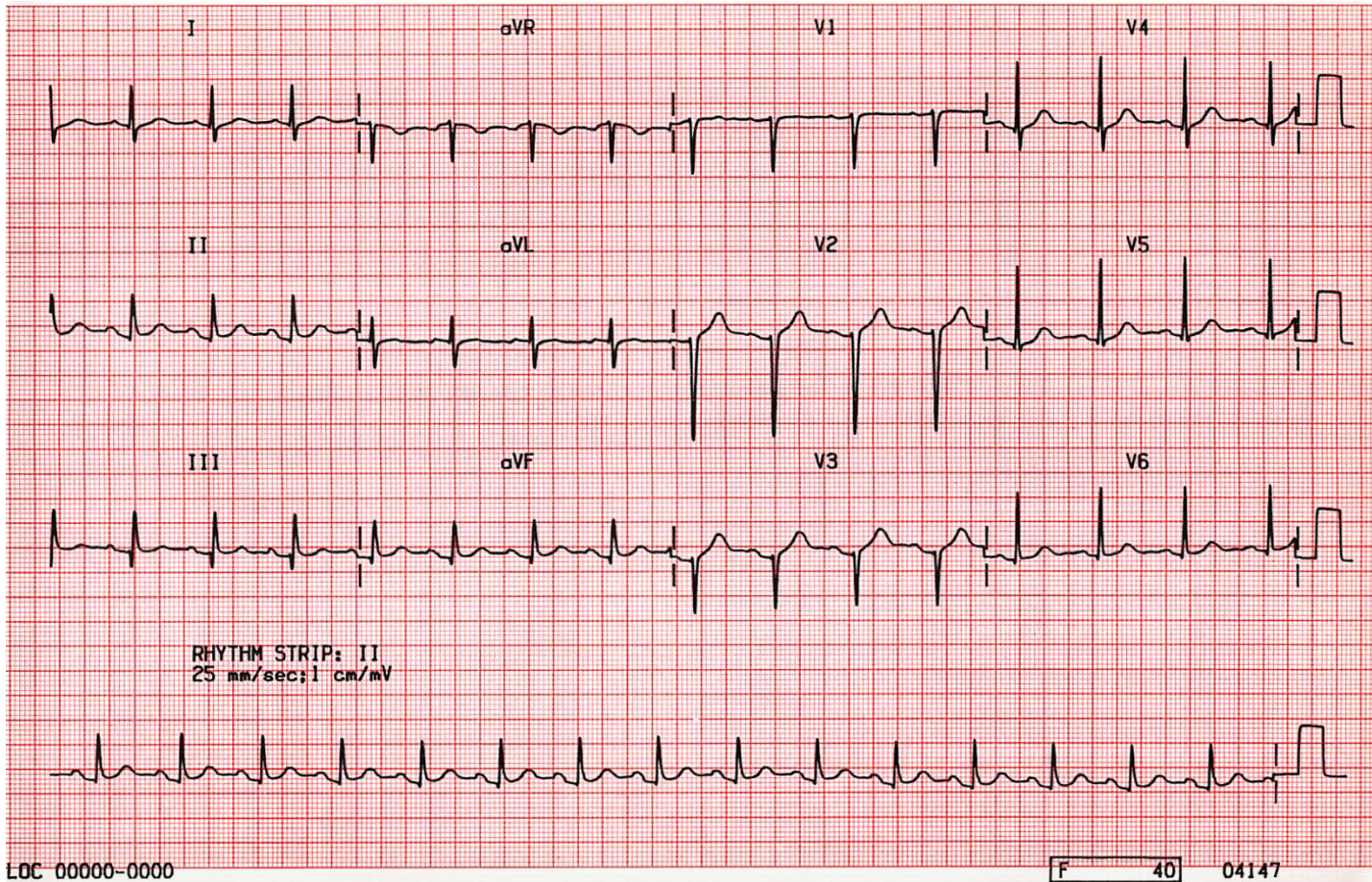
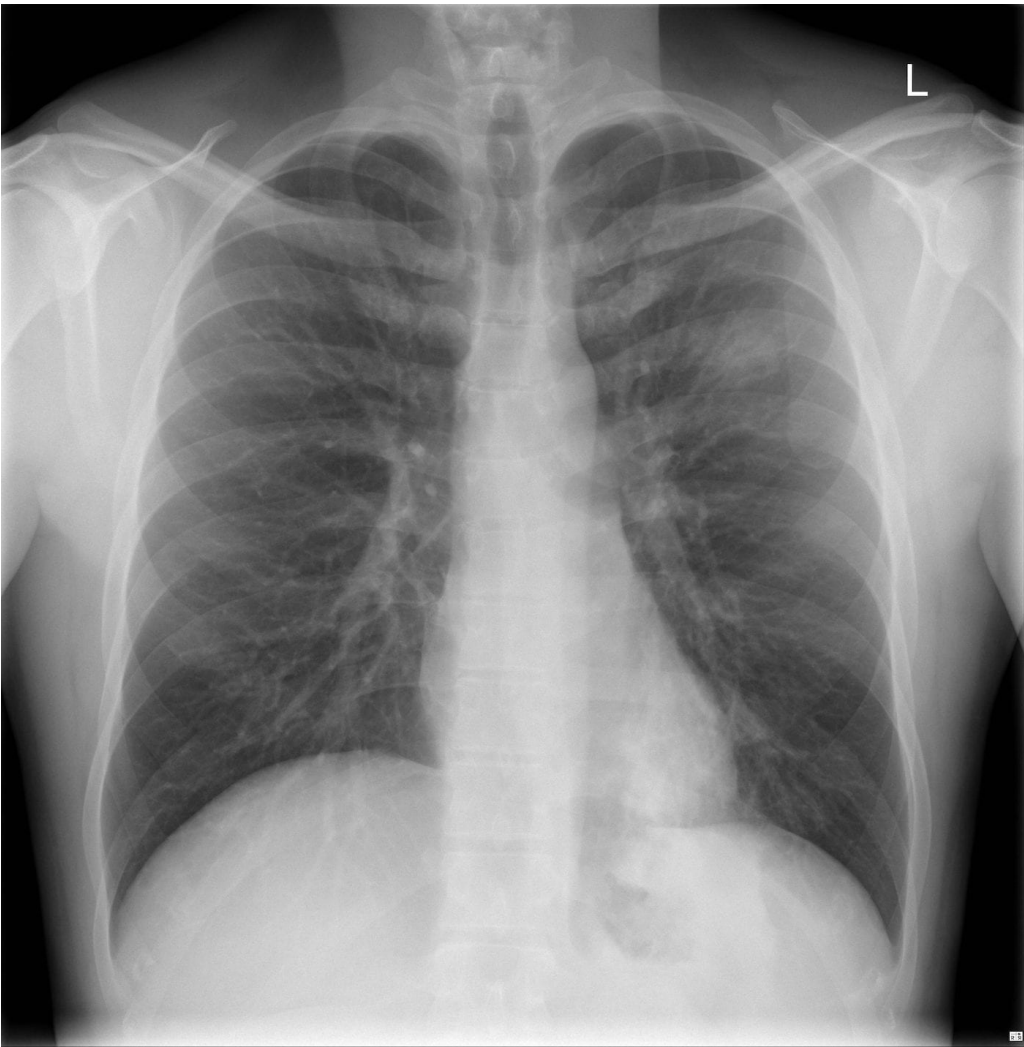
**MORRISONS**  
Fresh choice for you

**Wm MORRISON**  
Supermarkets plc B03 7DL  
Moking  
Manager : Lee King  
Telephone : 01483 755552  
Vat Number : 343475355

Savers Stamps  
Pick up a Card and Start  
Saving for Christmas Today

DATE: 19/08/2008 TIME: 17:54  
TILL: 0019 NO: 01969232  
You were served by: JENI

| DESCRIPTION                                                               | £      |
|---------------------------------------------------------------------------|--------|
| *M' FRESH SEA BREAM                                                       | 2.88 D |
| *M'SIDE OF SALMON                                                         | 3.08 D |
| *M' KIPPER FILLETS                                                        | 0.56 D |
| *M' PORK LEG STEAK                                                        | 2.93 D |
| *M' BROCCOLI                                                              |        |
| 0.270kg @ £1.99/kg                                                        | 0.54 D |
| *M' BABYLEAF AND HERB                                                     | 1.39 D |
| *M' RED GRAPES                                                            |        |
| 0.650kg @ £3.99/kg                                                        | 2.59 D |
| *M' BEST POTATOES                                                         | 0.99 D |
| HORLICKS                                                                  | 1.34 D |
| DUREX EXTRA SAFE                                                          | 5.68 B |
| *M' TRIM BEANS                                                            | 1.29 D |
| *M' STRAWBERRIES                                                          | 1.68 D |
| TETLEY TEA BAGS                                                           | 1.88 D |
| *M' VALUE ONIONS                                                          | 0.99 D |
| *M' DOUBLE CREAM                                                          | 0.56 D |
| *M' ENGLISH BUTTER                                                        | 0.94 D |
| *M' ENGLISH BUTTER                                                        | 0.94 D |
| *M' Butter Offer                                                          | -0.08  |
| *M' RASPBERRIES                                                           | 1.98 D |
| NIVEA FOR MEN                                                             | 2.58 A |
| *M' LOOSE LEMONS                                                          | 0.28 D |
| Items Sold: 20 TOTAL                                                      | £35.54 |
| CASH                                                                      | £40.00 |
| Change                                                                    | £4.46  |
| VAT A 17.5% (£2.59 ):                                                     | £0.39  |
| VAT B 5.0% (£5.98 ):                                                      | £0.28  |
| VAT D 0.0% (£26.97 ):                                                     | £0.00  |
| VAT Total                                                                 | £0.67  |
| <b>MULTISAVE</b><br><b>£0.08</b><br><b>SAVINGS</b><br><b>AT MORRISONS</b> |        |
| Thank you for shopping at Morrisons<br>Please call again                  |        |





# Time series data

- $y$  axis is some quantity we care about
- $x$  axis is time

## £ to \$ Exchange Rate

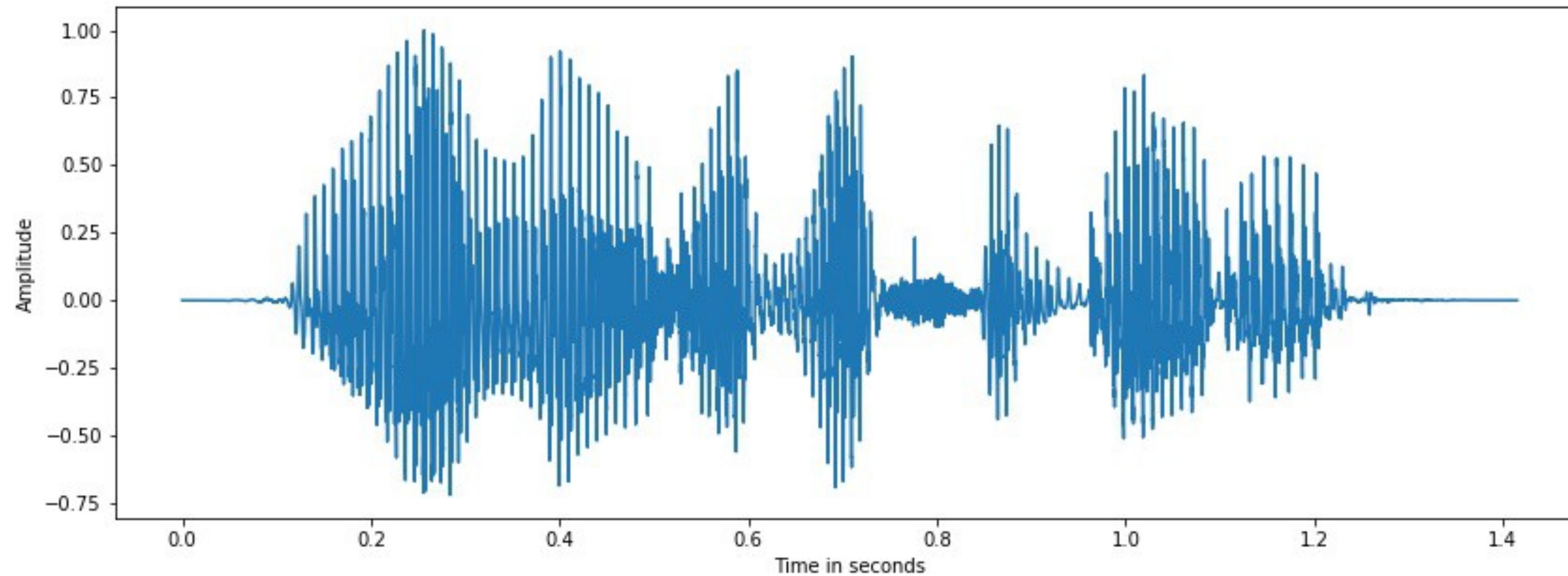


Source: <https://www.xe.com/currencycharts/?from=GBP&to=USD&view=10Y>



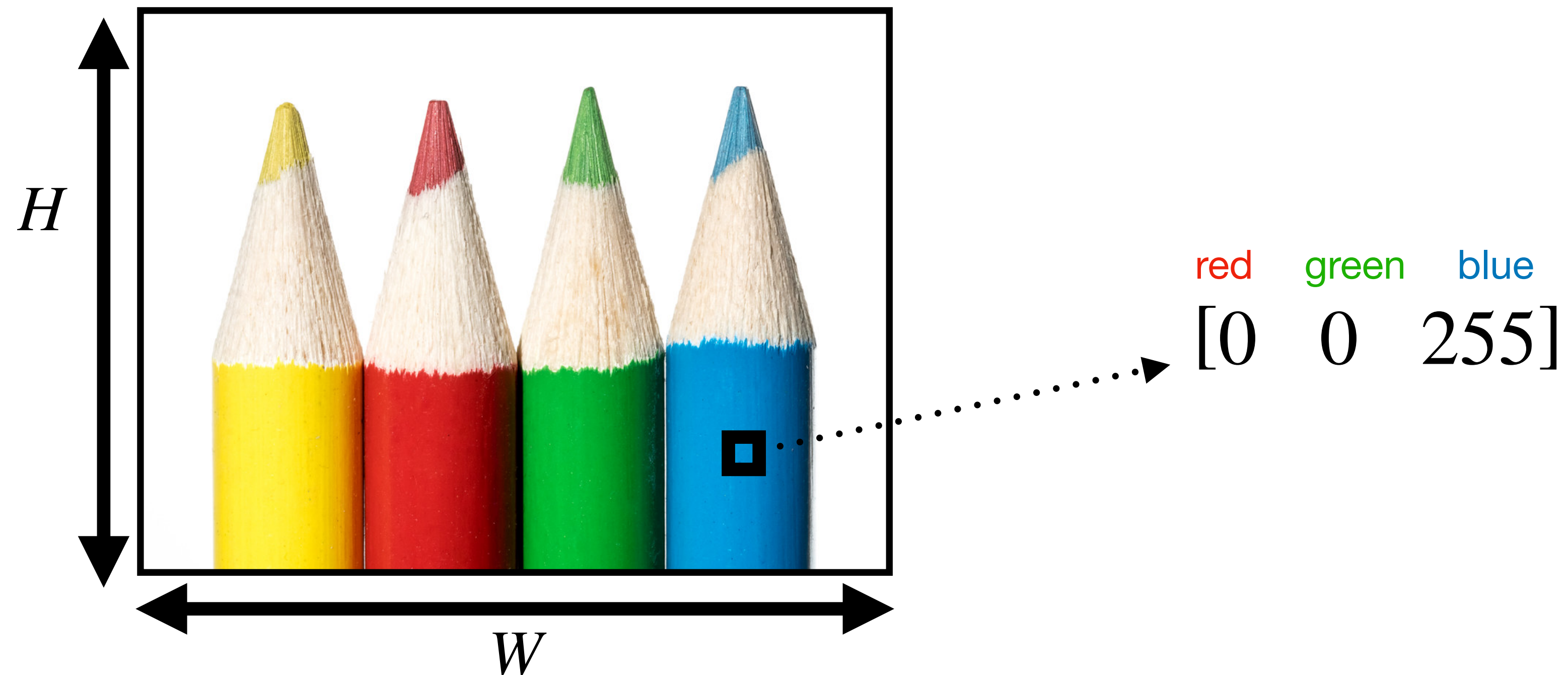
# Time series data

- For example, speech!



# Image data

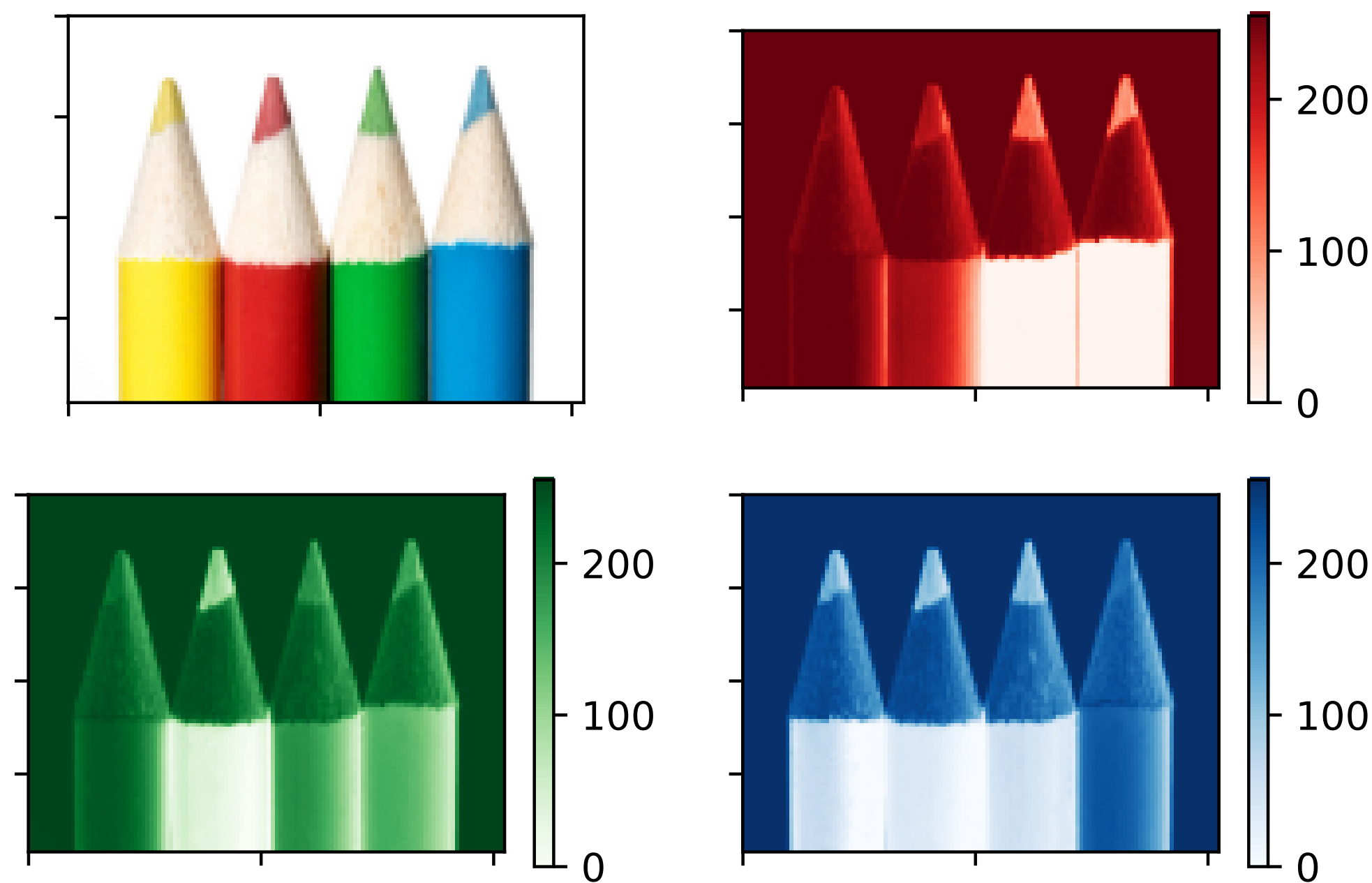
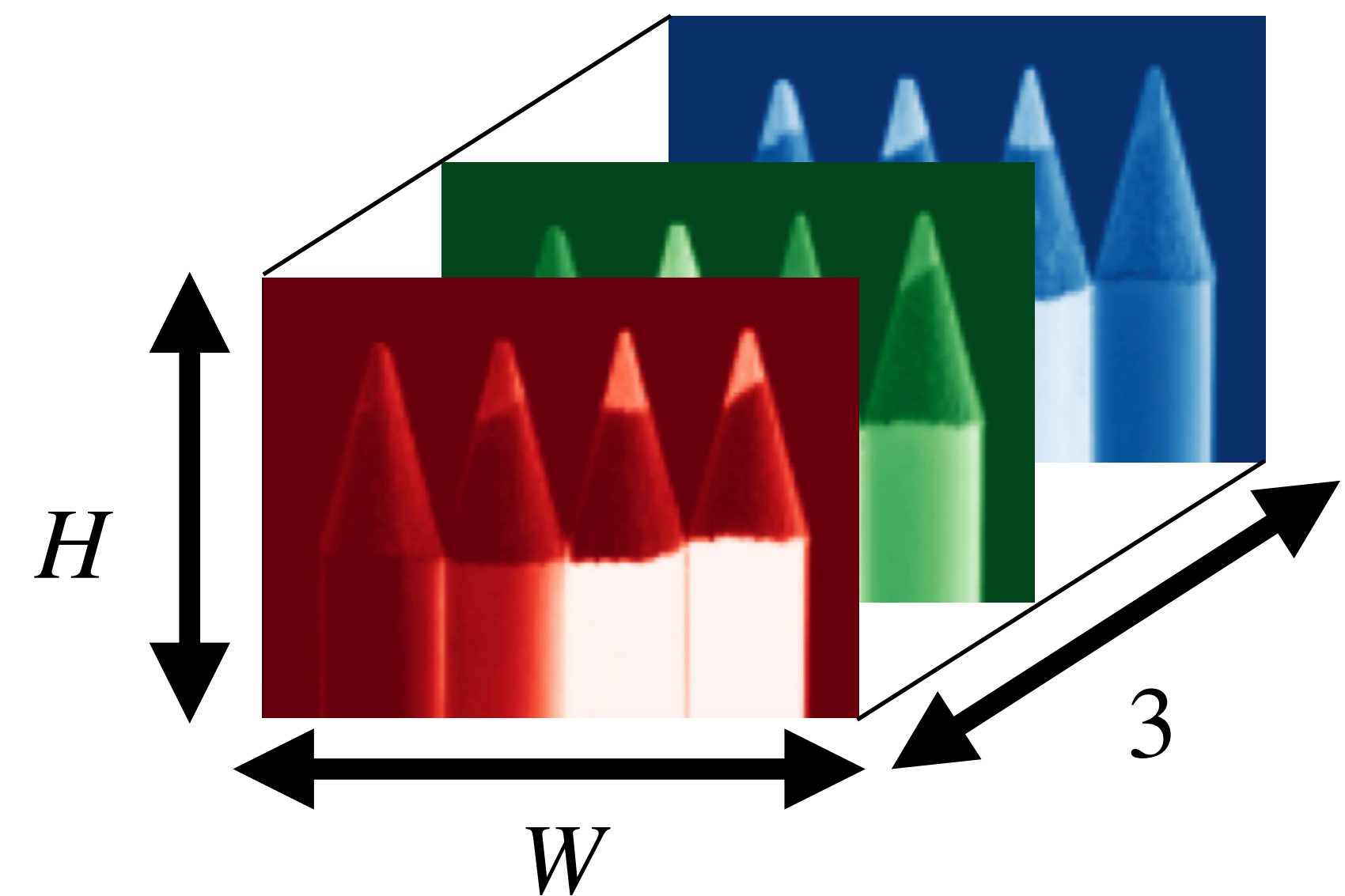
- An image is a rectangular array of  $H \times W$  pixels
- Each pixel consists of three numbers: the amount of **red**, **green**, and **blue**





# Image data

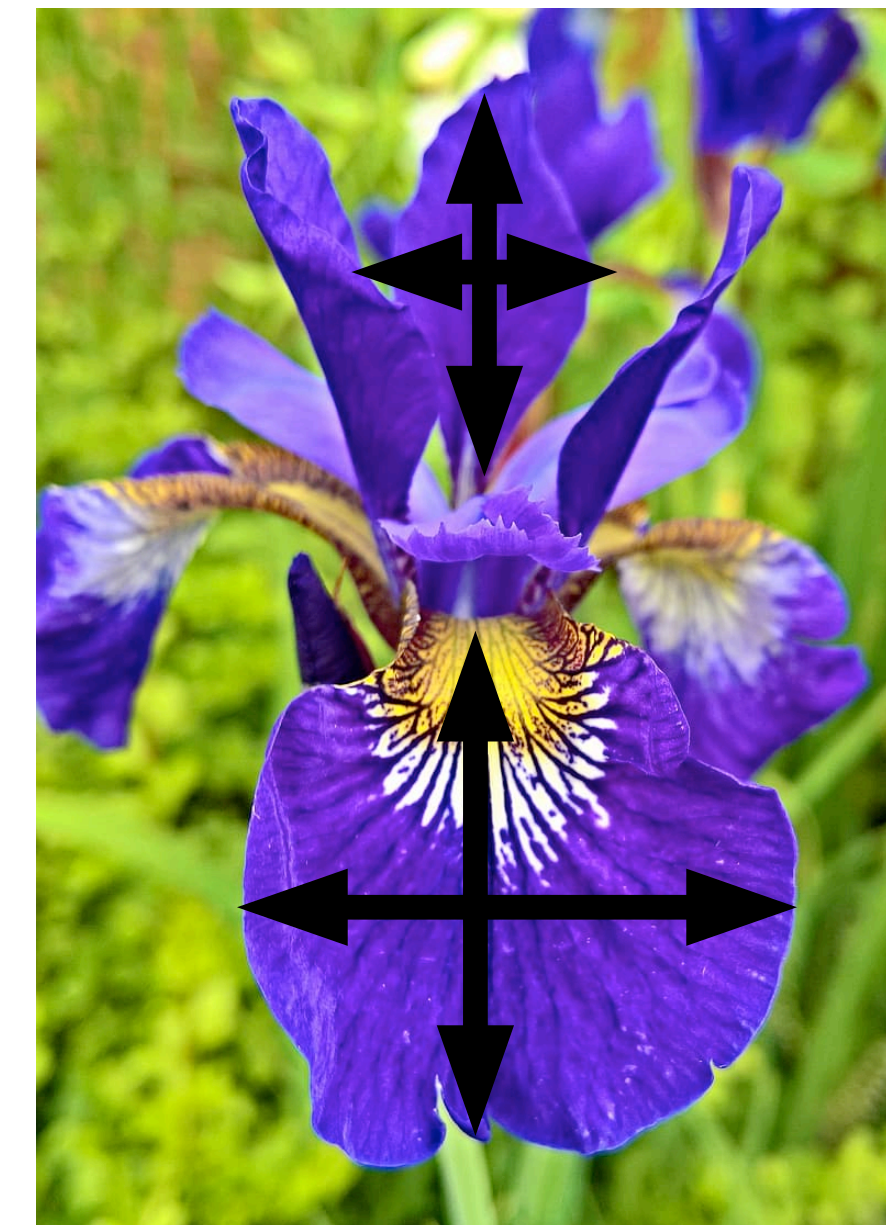
- This gives us a **red**, **green**, and **blue** 2D array
- These are stacked along the  $z$  axis to form a 3D array



# Tabular data

- Looks like a table with rows and columns
- Rows are objects and columns are attributes of those objects
- An example is the iris dataset of 150 flowers

|     | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) | species   |
|-----|-------------------|------------------|-------------------|------------------|-----------|
| 0   | 5.1               | 3.5              | 1.4               | 0.2              | setosa    |
| 1   | 4.9               | 3.0              | 1.4               | 0.2              | setosa    |
| 2   | 4.7               | 3.2              | 1.3               | 0.2              | setosa    |
| 3   | 4.6               | 3.1              | 1.5               | 0.2              | setosa    |
| 4   | 5.0               | 3.6              | 1.4               | 0.2              | setosa    |
| ... | ...               | ...              | ...               | ...              | ...       |
| 145 | 6.7               | 3.0              | 5.2               | 2.3              | virginica |
| 146 | 6.3               | 2.5              | 5.0               | 1.9              | virginica |
| 147 | 6.5               | 3.0              | 5.2               | 2.0              | virginica |
| 148 | 6.2               | 3.4              | 5.4               | 2.3              | virginica |
| 149 | 5.9               | 3.0              | 5.1               | 1.8              | virginica |





# Free-form data

- Largely unstructured and usually text
- Can (sometimes!) be hacked into e.g. tabular data

☆☆☆☆ 3/26/2017

Stopped by on a Sunday afternoon, not so crowded and we got a table outside right away. Service was not attentive, we had to go in to get waitstaff including ordering and paying the bill. Food was meh. Ordered the prosciutto scramble, arugula and fennel salad, and Caesar salad. Don't think our scramble came with prosciutto, and arugula salad was extremely sour and quite plain. Fried cauliflower was quite tasty.

Overall a very mediocre place.

*Dracula* is a novel by [Bram Stoker](#), published in 1897. As an [epistolary novel](#), the narrative is related through letters, diary entries, and newspaper articles. It has no single protagonist, but opens with solicitor [Jonathan Harker](#) taking a business trip to stay at the castle of a [Transylvanian](#) noble, [Count Dracula](#). Harker escapes the castle after discovering that Dracula is a [vampire](#), and the Count moves to [England](#) and plagues the seaside town of [Whitby](#). A small group, led by [Abraham Van Helsing](#), hunt Dracula and, in the end, kill him.

*Dracula* was mostly written in the 1890s. Stoker produced over a hundred pages of notes for the novel, drawing extensively from [Transylvanian folklore](#) and [history](#). Some scholars have suggested that the character of Dracula was inspired by historical figures like the [Wallachian](#) prince [Vlad the Impaler](#) or the countess [Elizabeth Báthory](#), but there is widespread disagreement. Stoker's notes mention neither figure. He found the name *Dracula* in Whitby's public library while holidaying there, picking it because he thought it meant *devil* in [Romanian](#).

Following its publication, *Dracula* was positively received by reviewers who pointed to its effective use of horror. In contrast, reviewers who wrote negatively of the novel regarded it as excessively frightening. Comparisons to other works of [Gothic fiction](#) were common, including its structural similarity to [Wilkie Collins'](#) *The Woman in White* (1859). In the past century, *Dracula* has been situated as a piece of Gothic fiction. Modern scholars explore the novel within its historical context—the [Victorian era](#)—and discuss its depiction of gender roles, sexuality, and race.



**Elon Musk**   
@elonmusk

Follow

▼

Replying to [@JakeBlueatSM](#)

May be initiated not by the country leaders, but one of the AI's, if it decides that a prepemptive strike is most probable path to victory

11:36 PM - 3 Sep 2017

1,816 Retweets 5,926 Likes



 609  1.8K  5.9K 



# Nomenclature

- A **dataset** is a collection of **data points**
- A **data point** is a set of **elements**
- An **element** is a measurable or countable quantity



A dataset of **images**



An **image** is a set of **pixels**

A **pixel** measures the intensity of different colour(s)

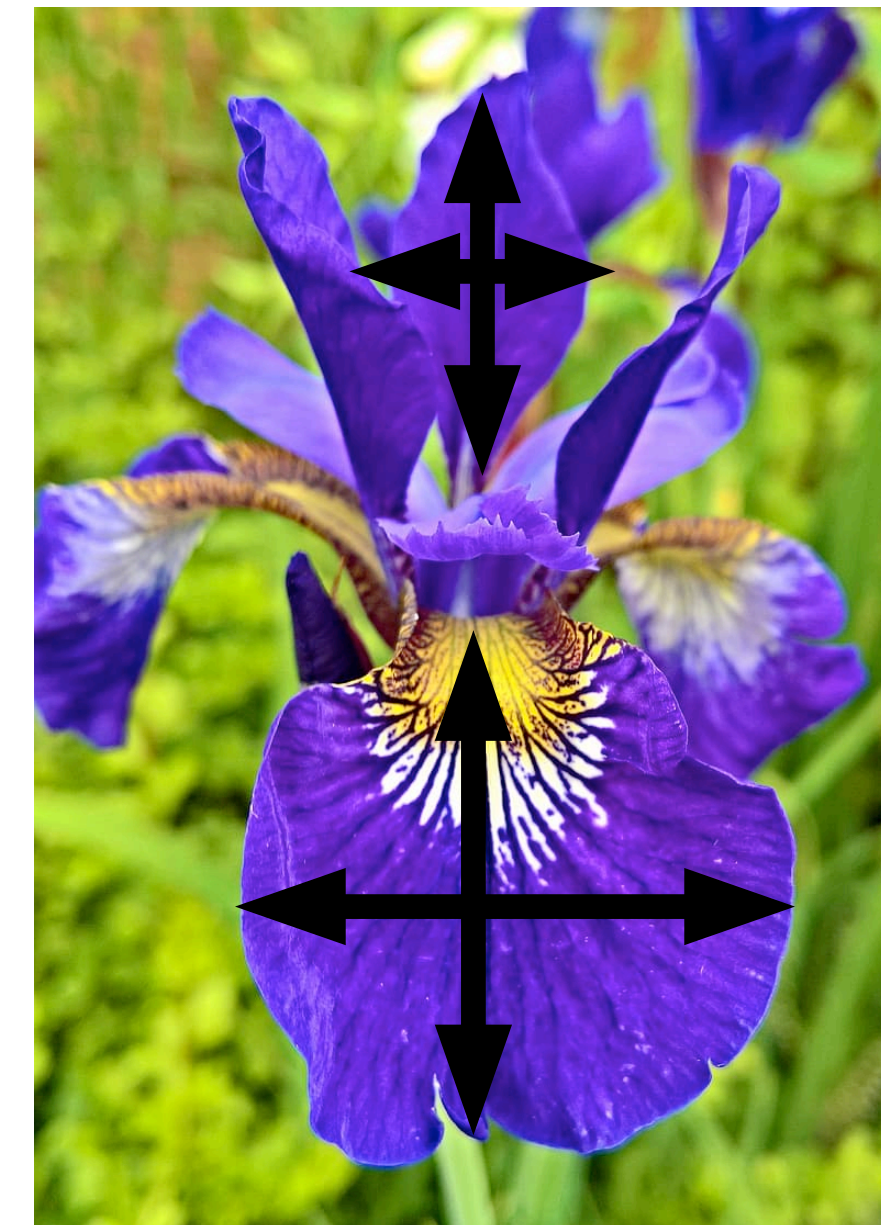


# Variable Types

# Tabular data (again!)

- A table is a dataset and its rows are data points
- Each data point is a set of elements which are measurements of some **attributes** or **features**
- The measurements for a given attribute **vary** across the dataset

|     | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) | species   |
|-----|-------------------|------------------|-------------------|------------------|-----------|
| 0   | 5.1               | 3.5              | 1.4               | 0.2              | setosa    |
| 1   | 4.9               | 3.0              | 1.4               | 0.2              | setosa    |
| 2   | 4.7               | 3.2              | 1.3               | 0.2              | setosa    |
| 3   | 4.6               | 3.1              | 1.5               | 0.2              | setosa    |
| 4   | 5.0               | 3.6              | 1.4               | 0.2              | setosa    |
| ... | ...               | ...              | ...               | ...              | ...       |
| 145 | 6.7               | 3.0              | 5.2               | 2.3              | virginica |
| 146 | 6.3               | 2.5              | 5.0               | 1.9              | virginica |
| 147 | 6.5               | 3.0              | 5.2               | 2.0              | virginica |
| 148 | 6.2               | 3.4              | 5.4               | 2.3              | virginica |
| 149 | 5.9               | 3.0              | 5.1               | 1.8              | virginica |





# Variables

- The measurements for a given attribute (/feature) **vary** across the dataset
- This means we can think of the attributes (/features) as **variables**
- There are different **types** of variables

|     | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) | species   |
|-----|-------------------|------------------|-------------------|------------------|-----------|
| 0   | 5.1               | 3.5              | 1.4               | 0.2              | setosa    |
| 1   | 4.9               | 3.0              | 1.4               | 0.2              | setosa    |
| 2   | 4.7               | 3.2              | 1.3               | 0.2              | setosa    |
| 3   | 4.6               | 3.1              | 1.5               | 0.2              | setosa    |
| 4   | 5.0               | 3.6              | 1.4               | 0.2              | setosa    |
| ... | ...               | ...              | ...               | ...              | ...       |
| 145 | 6.7               | 3.0              | 5.2               | 2.3              | virginica |
| 146 | 6.3               | 2.5              | 5.0               | 1.9              | virginica |
| 147 | 6.5               | 3.0              | 5.2               | 2.0              | virginica |
| 148 | 6.2               | 3.4              | 5.4               | 2.3              | virginica |
| 149 | 5.9               | 3.0              | 5.1               | 1.8              | virginica |

The measurements for the lengths and widths are positive real numbers

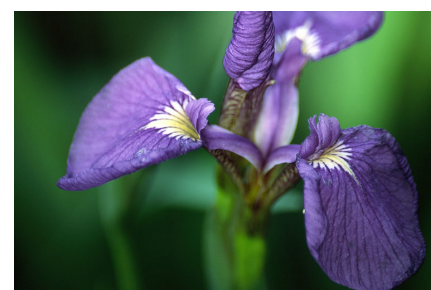
The measurements for species are “setosa”, “virginica”, or “versicolor”

# Categorical variables

Measurements of the variable correspond to descriptive categories

- For **nominal variables** the categories have no order
- For **ordinal variables** the categories are ordered (but don't fit on a number line)

iris species (nominal)



0  
setosa



1  
versicolor



2  
virginica



1  
primary



2  
secondary



3  
university

level of education (ordinal)



# Numerical variables

Measurements of the variable can be discrete or continuous

- For **discrete variables** they can only be integers
- For **continuous variables** they can be any real number (within a given range)



The number of times this man tosses this coin is discrete  
The length of a tie is continuous

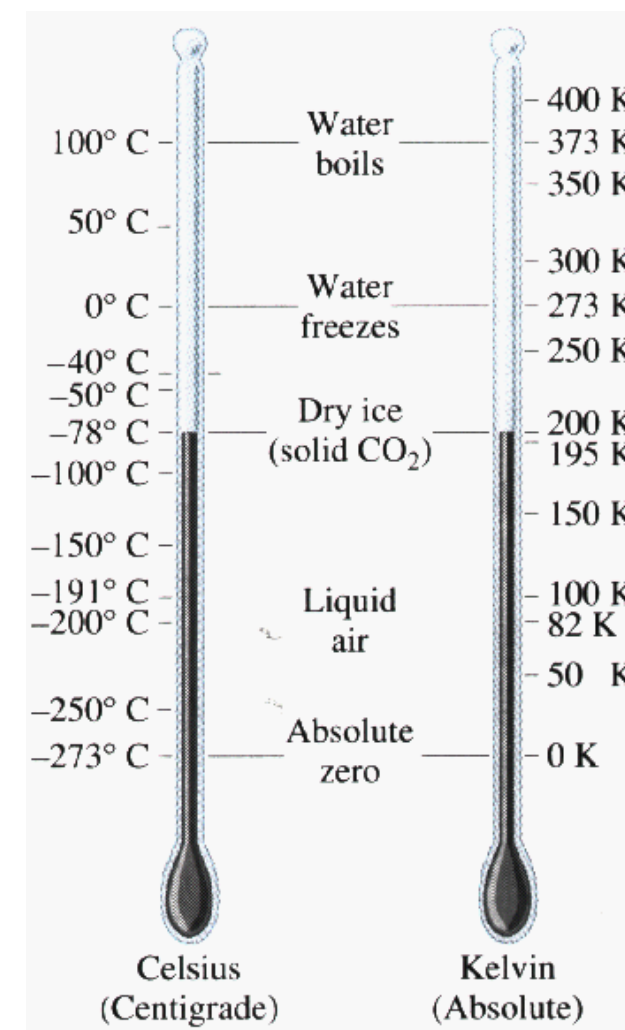
# Continuous variables

These can be further divided into interval and ratio

- For **interval variables** a zero measurement is just relative to something else
- For **ratio variables** zero is meaningful (i.e. the absence of something)

Temperature in Celsius is interval

Temperature in Kelvin is ratio



Ratios of (ahem) ratio variables  
are meaningful.  
10K is twice as hot as 5K

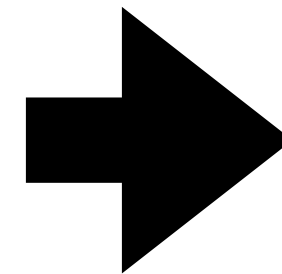


# Preprocessing

# Matrix inputs

- Most machine learning (ML) methods require a matrix input
- Our dataset must (usually) be represented by a matrix of real numeric values
- Discrete and continuous are both fine; we just pretend everything is continuous
- Given tabular data, we need to convert it into such a matrix

|   | Height (cm) | Age | Favourite colour |
|---|-------------|-----|------------------|
| 0 | 185         | 32  | blue             |
| 1 | 193         | 70  | red              |
| 2 | 147         | 77  | brown            |
| 3 | 163         | 26  | blue             |



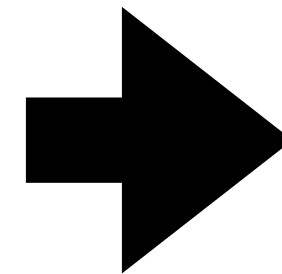
?



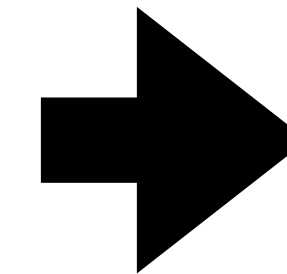
# Representing a dataset as a matrix

- We have tabular data with  $N$  data points (rows) and  $C$  features (cols)
- For now, we will drop features that don't correspond to numeric variables
- If there are now  $D$  features we can represent the dataset by an  $N \times D$  matrix

|   | Height (cm) | Age | Favourite colour |
|---|-------------|-----|------------------|
| 0 | 185         | 32  | blue             |
| 1 | 193         | 70  | red              |
| 2 | 147         | 77  | brown            |
| 3 | 163         | 26  | blue             |



|   | Height (cm) | Age |
|---|-------------|-----|
| 0 | 185         | 32  |
| 1 | 193         | 70  |
| 2 | 147         | 77  |
| 3 | 163         | 26  |

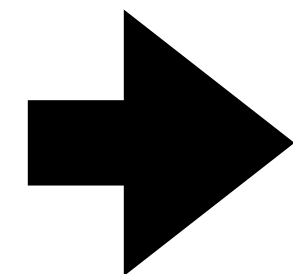


$$\mathbf{X} = \begin{bmatrix} 185 & 32 \\ 193 & 70 \\ 147 & 77 \\ 163 & 26 \end{bmatrix}$$
$$\mathbf{X} \in \mathbb{R}^{N \times D}$$

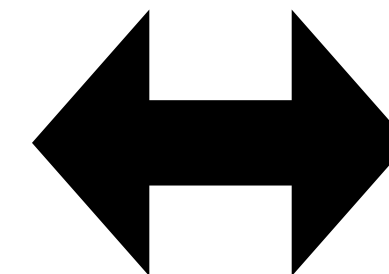
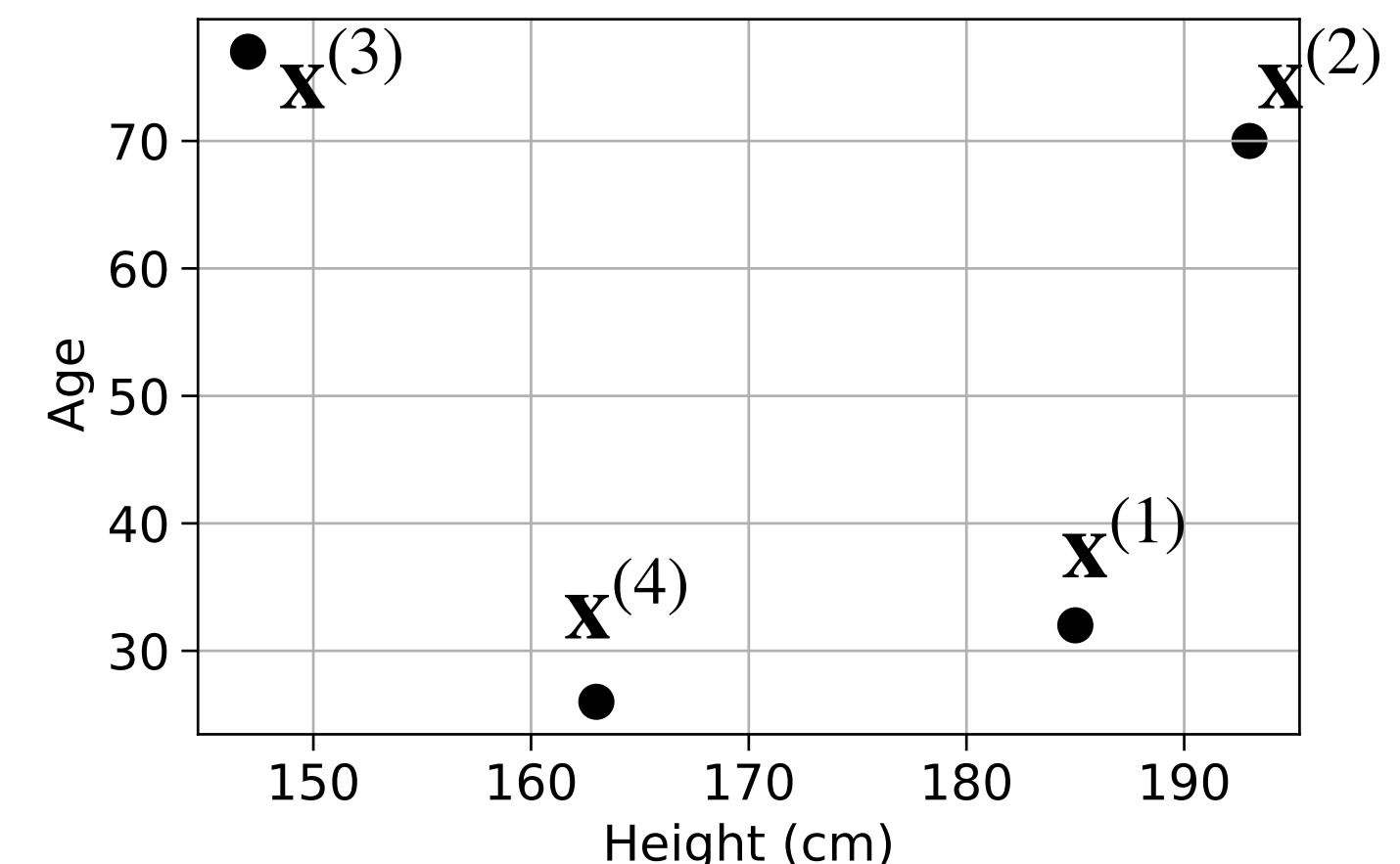
# Representing data points as vectors

- We are representing our dataset using a  $N \times D$  dataset matrix  $\mathbf{X}$
- Each row is a data point that lives in  $D$ -dimensional space
- Let's denote these as  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots, \mathbf{x}^{(N)}$  or  $\{\mathbf{x}^{(n)}\}_{n=1}^N$ . They are **vectors**

|   | Height (cm) | Age |
|---|-------------|-----|
| 0 | 185         | 32  |
| 1 | 193         | 70  |
| 2 | 147         | 77  |
| 3 | 163         | 26  |



$$\mathbf{X} = \begin{bmatrix} 185 & 32 \\ 193 & 70 \\ 147 & 77 \\ 163 & 26 \end{bmatrix}$$



$$\mathbf{X} \in \mathbb{R}^{N \times D}$$

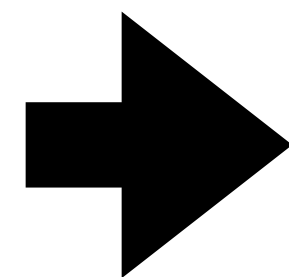
$$\mathbf{x} \in \mathbb{R}^D$$



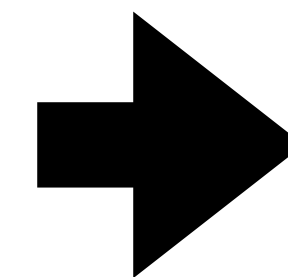
# What if we have a missing (or wrong!) value for a feature?

- Option 1: Remove the affected data point(/s)

|   | Height (cm) | Weight (kg) | Age |
|---|-------------|-------------|-----|
| 0 | 0.01        | 80.0        | 32  |
| 1 | 193.00      | NaN         | 70  |
| 2 | 147.00      | 62.0        | 77  |
| 3 | 163.00      | 72.0        | 26  |
| 4 | 150.00      | 64.0        | 21  |
| 5 | 185.00      | 74.0        | 52  |



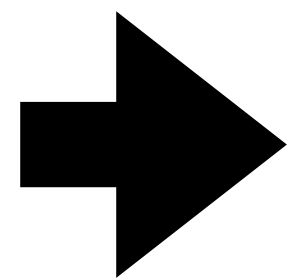
|              | Height (cm)       | Weight (kg)     | Age           |
|--------------|-------------------|-----------------|---------------|
| <del>0</del> | <del>0.01</del>   | <del>80.0</del> | <del>32</del> |
| <del>1</del> | <del>193.00</del> | <del>NaN</del>  | <del>70</del> |
| 2            | 147.00            | 62.0            | 77            |
| 3            | 163.00            | 72.0            | 26            |
| 4            | 150.00            | 64.0            | 21            |
| 5            | 185.00            | 74.0            | 52            |



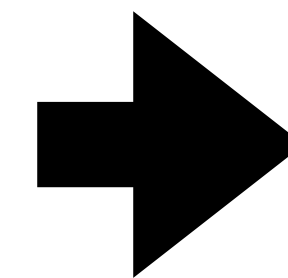
$$X = \begin{bmatrix} 147 & 62 & 77 \\ 163 & 72 & 26 \\ 150 & 64 & 21 \\ 185 & 74 & 52 \end{bmatrix}$$

- Option 2: Impute a value (e.g. the average for that feature)

|   | Height (cm) | Weight (kg) | Age |
|---|-------------|-------------|-----|
| 0 | 0.01        | 80.0        | 32  |
| 1 | 193.00      | NaN         | 70  |
| 2 | 147.00      | 62.0        | 77  |
| 3 | 163.00      | 72.0        | 26  |
| 4 | 150.00      | 64.0        | 21  |
| 5 | 185.00      | 74.0        | 52  |



|   | Height (cm) | Weight (kg) | Age |
|---|-------------|-------------|-----|
| 0 | 168         | 80.0        | 32  |
| 1 | 193.00      | 70          | 70  |
| 2 | 147.00      | 62.0        | 77  |
| 3 | 163.00      | 72.0        | 26  |
| 4 | 150.00      | 64.0        | 21  |
| 5 | 185.00      | 74.0        | 52  |

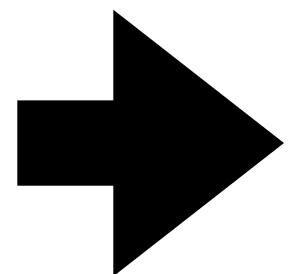


$$X = \begin{bmatrix} 168 & 80 & 32 \\ 193 & 70 & 70 \\ 147 & 62 & 77 \\ 163 & 72 & 26 \\ 150 & 64 & 21 \\ 185 & 74 & 52 \end{bmatrix}$$

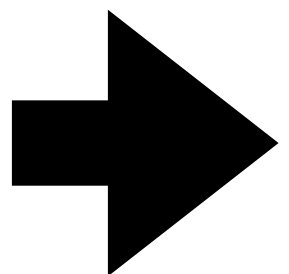
# What if we want to include categorical variables?

- **If ordinal** we can map to numbers that maintain order

|   | Height (cm) | Age | Highest qualification |
|---|-------------|-----|-----------------------|
| 0 | 185         | 32  | Bachelors             |
| 1 | 193         | 70  | PhD                   |
| 2 | 147         | 77  | Masters               |
| 3 | 163         | 26  | Bachelors             |



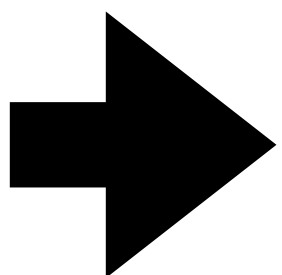
{‘Bachelors’: 1,  
‘Masters’: 2,  
‘PhD’: 3,}



$$X = \begin{bmatrix} 185 & 32 & 1 \\ 193 & 70 & 3 \\ 147 & 77 & 2 \\ 163 & 26 & 1 \end{bmatrix}$$

- **If nominal** we can create a **binary feature** for each category

|   | Height (cm) | Age | Favourite colour |
|---|-------------|-----|------------------|
| 0 | 185         | 32  | blue             |
| 1 | 193         | 70  | red              |
| 2 | 147         | 77  | brown            |
| 3 | 163         | 26  | blue             |



|   | Height (cm) | Age | Favourite colour_blue | Favourite colour_brown | Favourite colour_red |
|---|-------------|-----|-----------------------|------------------------|----------------------|
| 0 | 185         | 32  | 1                     | 0                      | 0                    |
| 1 | 193         | 70  | 0                     | 0                      | 1                    |
| 2 | 147         | 77  | 0                     | 1                      | 0                    |
| 3 | 163         | 26  | 1                     | 0                      | 0                    |



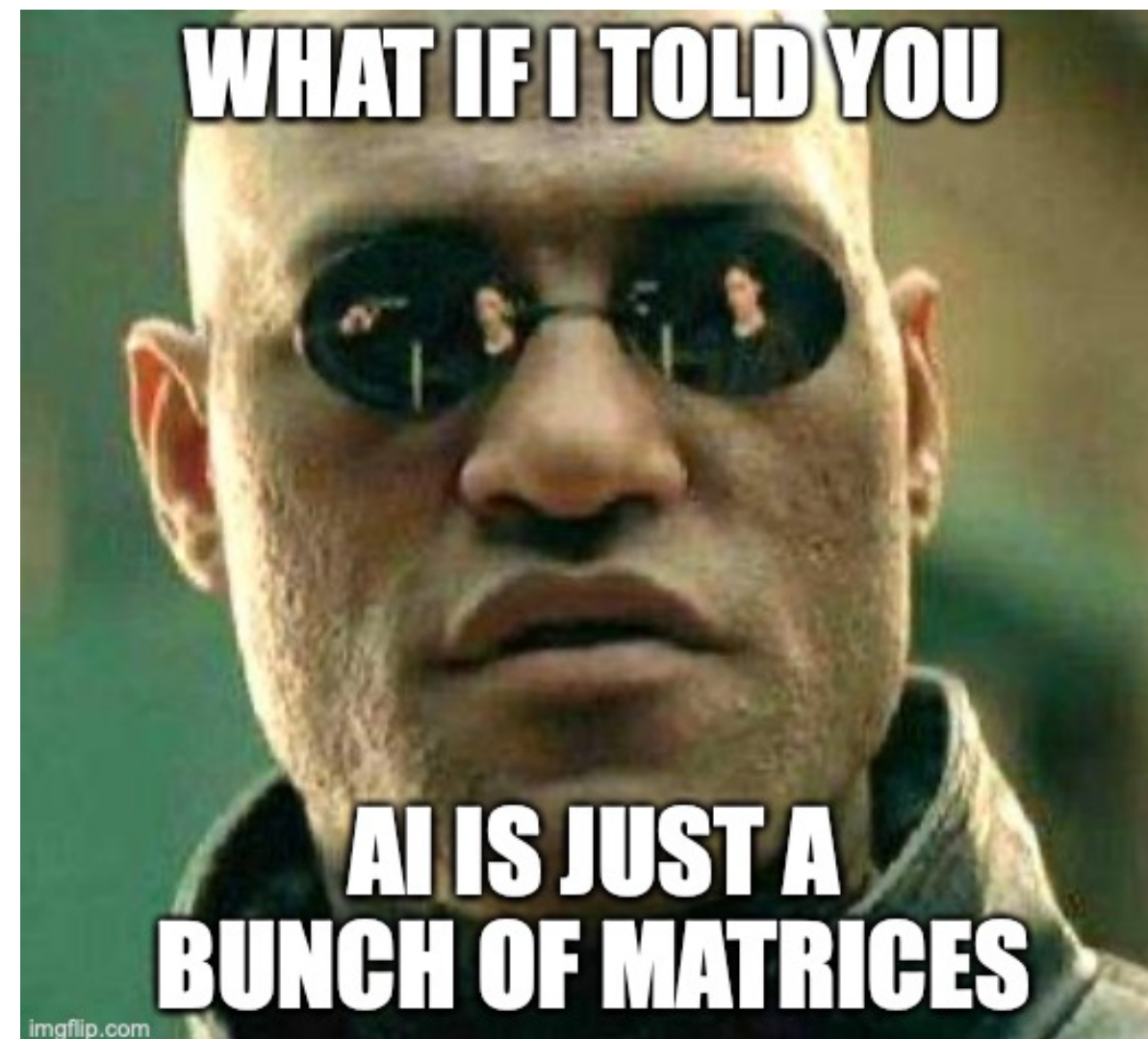
# Data points are column vectors

- It is standard with tabular data to have the rows as data points
- But in ML literature it is convention to denote all vectors including data points  $\mathbf{x}$  as **column vectors**
- It is also convention to represent a dataset as  $\mathbf{X} \in \mathbb{R}^{N \times D}$  (in the same way we just did) where the **rows** are those data points
- **Just be aware of this peculiarity!**

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \quad \mathbf{x}^{(n)} = \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_D^{(n)} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)\top} \\ \mathbf{x}^{(2)\top} \\ \mathbf{x}^{(3)\top} \\ \vdots \\ \mathbf{x}^{(N)\top} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_D^{(2)} \\ x_1^{(3)} & x_2^{(3)} & \dots & x_D^{(3)} \\ \dots & \dots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_D^{(N)} \end{bmatrix}$$

# Why vectors?

- We can now use the machinery of linear algebra for ML
- Matrices linearly transform vectors
- Computers are very good at matrix multiplication
- Neural networks consist of multiple matrices (See later in the course!)





# Can we represent other types of data as vectors?

- Yes! We can flatten or **vectorise** images



- We can represent text data as a histogram of word counts (a bag of words)  
e.g. [ # “I”, # “like”, # “sausage”, # “hate”]

I like sausage

$[1 \ 1 \ 1 \ 0]^T$

I hate sausage

$[1 \ 0 \ 1 \ 1]^T$

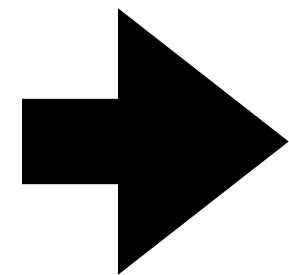
sausage sausage

$[0 \ 0 \ 2 \ 0]^T$

# Standardising your data

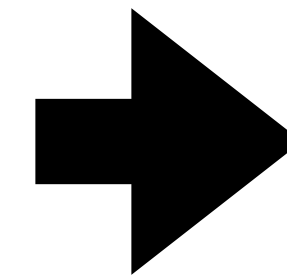
- Measurements of different features can have vastly different scales
- We want to compare features like-for-like and not let those with naturally large values dominate
- The solution is to **standardise** your data
- We want each column of  $\mathbf{X}$  to have a mean of 0 and an SD of 1

|   | Height (cm) | Age | Salary (£) |
|---|-------------|-----|------------|
| 0 | 190         | 44  | 25000      |
| 1 | 143         | 36  | 29000      |
| 2 | 152         | 20  | 100000     |
| 3 | 178         | 56  | 67000      |



$\mathbf{X} =$

$$\begin{bmatrix} 190 & 44 & 25000 \\ 143 & 36 & 29000 \\ 152 & 20 & 100000 \\ 178 & 56 & 67000 \end{bmatrix}$$



?



# Standardising your data

- We want each column of  $\mathbf{X}$  to have a mean of 0 and an SD of 1
- For each column, compute the mean and SD
- Then subtract the mean from each value and divide by SD
- **This is essential for many ML algorithms**

I will use  $\sum_n$  to mean  
“sum over all  $n$ ”

$$\mathbf{X}_{old} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_D^{(2)} \\ \dots & \dots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_D^{(N)} \end{bmatrix}$$

$$\mu_j = \frac{1}{N} \sum_n \mathbf{x}_j^{(n)}$$

$$\sigma_j^2 = \frac{1}{N} \sum_n (\mathbf{x}_j^{(n)} - \mu_j)^2$$

$$\mathbf{X}_{new} = \begin{bmatrix} \frac{x_1^{(1)} - \mu_1}{\sigma_1} & \frac{x_2^{(1)} - \mu_2}{\sigma_2} & \dots & \frac{x_D^{(1)} - \mu_D}{\sigma_D} \\ \frac{x_1^{(2)} - \mu_1}{\sigma_1} & \frac{x_2^{(2)} - \mu_2}{\sigma_2} & \dots & \frac{x_D^{(2)} - \mu_D}{\sigma_D} \\ \dots & \dots & \ddots & \vdots \\ \frac{x_1^{(N)} - \mu_1}{\sigma_1} & \frac{x_2^{(N)} - \mu_2}{\sigma_2} & \dots & \frac{x_D^{(N)} - \mu_D}{\sigma_D} \end{bmatrix}$$

# Normalising vs. standardising

- Nomenclature can vary but in this course standardising refers to scaling each variable to zero mean and unit variance
- We can do other forms of scaling e.g. divide each variable by its maximum value
- We will refer to other forms of scaling as normalising
- Generally, anything that gets different variables to similar ranges is fine; **just make sure you do it!**

If you have a bunch of binary variables you can just leave things alone!



# Summary

- We have considered different modalities of data e.g. tabular, image, freeform
- We have established the nomenclature for talking about data
- We have seen how attributes in tabular data can be treated as variables
- We have considered different variable types
- We have seen how to pre-process data into matrix form
- We have considered standardisation

# Questions



# Q1

Which of these statements are true?

1. A person's age group (e.g. child (0-12), teenager (13-19), adult (20-64), elderly (65+)) is an ordinal variable
2. The manufacturer of a car (e.g. Ford, Toyota, Honda) is an ordinal variable
3. Temperature in Celsius is an interval variable
4. A person's number of siblings is a ratio variable

# Q2

Consider a dataset, where each data point is a 3D bag of words vector where:

- The first element is the number of occurrences of the word "snap"
- The second element is the number of occurrences of the word "crackle"
- The third element is the number of occurrences of the word "pop"

Compute the Euclidean distance between the vectors corresponding to "snap crackle pop" and "pop crackle pop".



# Q3

Standardise the dataset matrix  $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$