Data Analysis and Machine Learning 4 (DAML)

Week 1: Introduction, data modalities, variable types

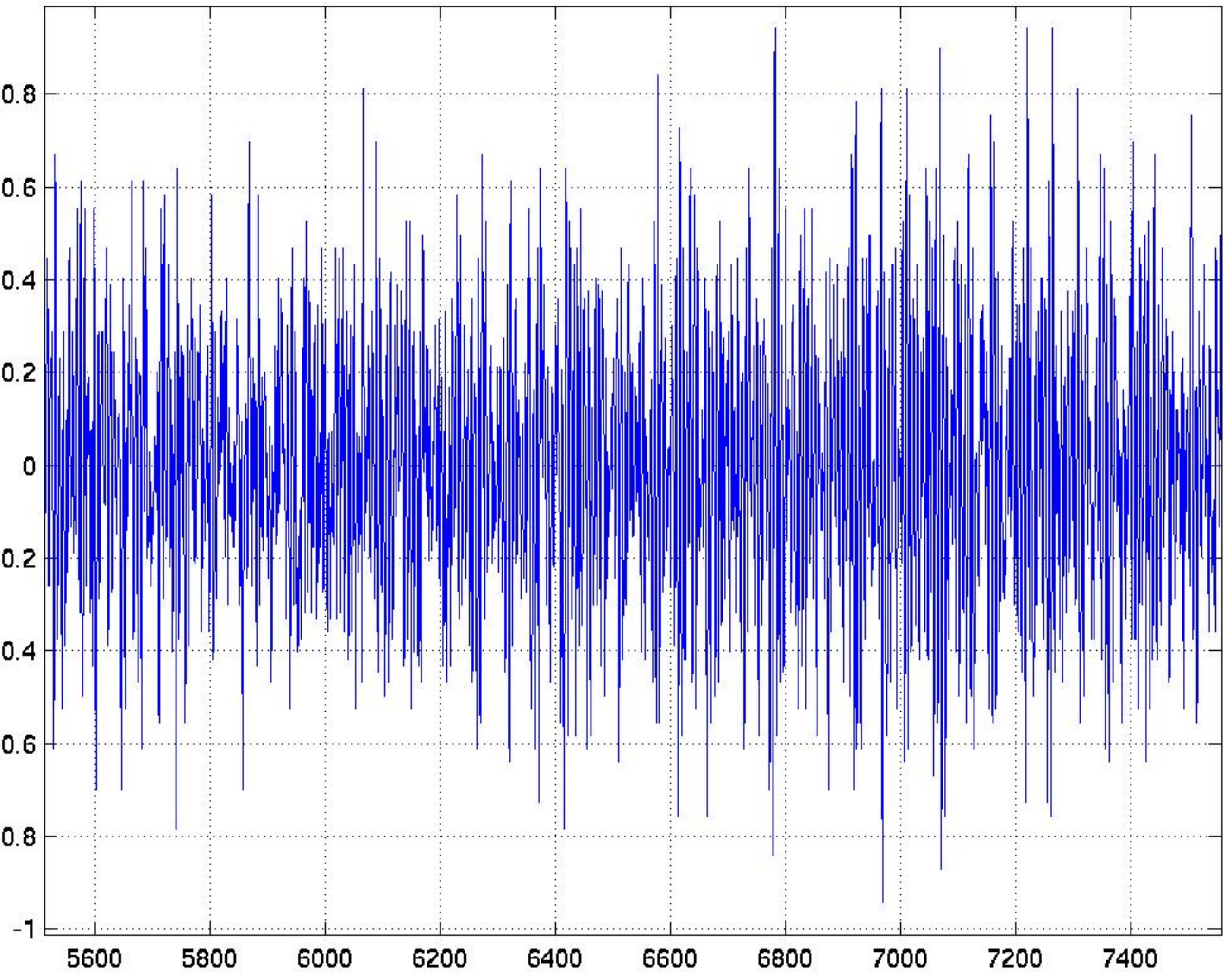
Elliot J. Crowley, 15th January 2024

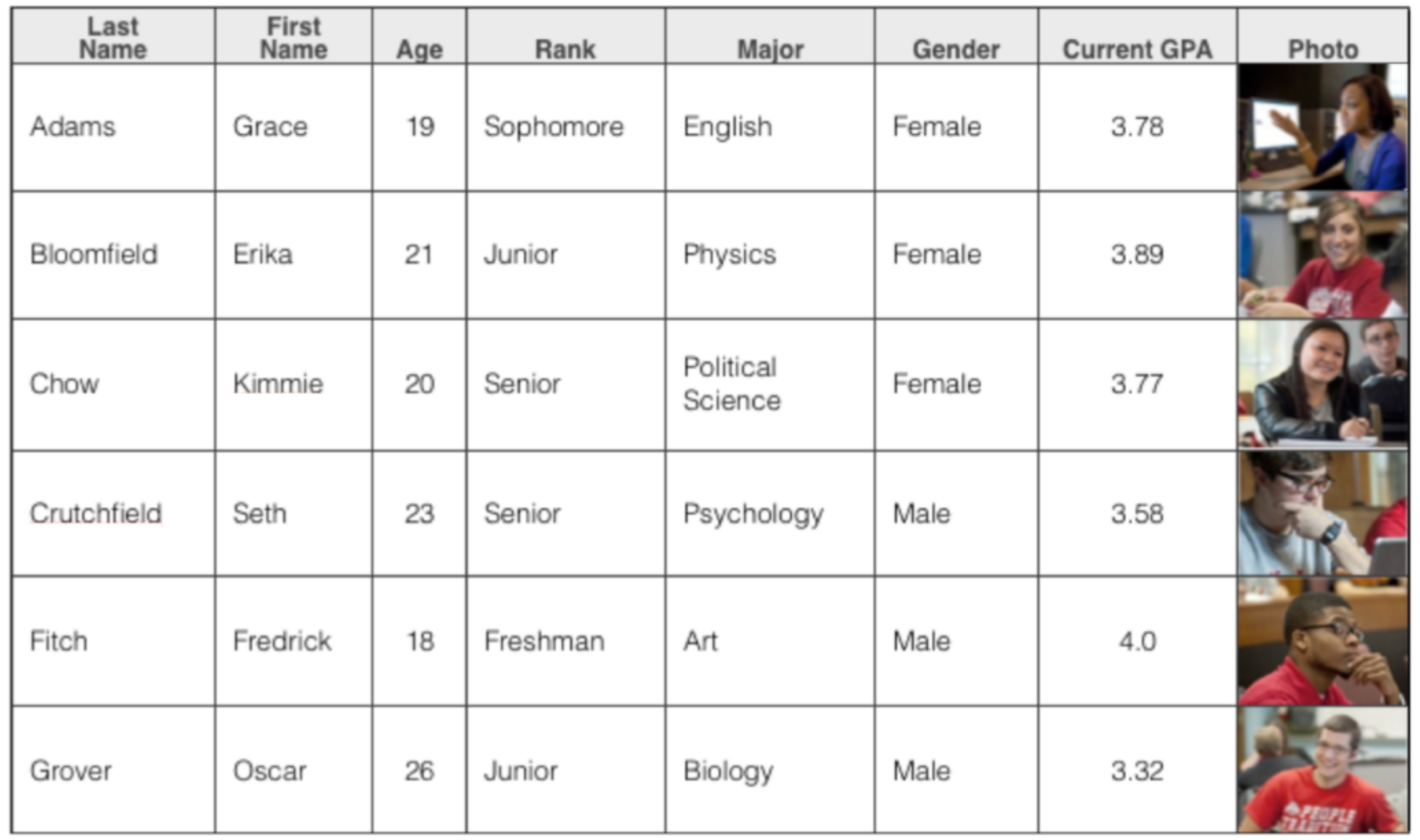
What is data?

> “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”

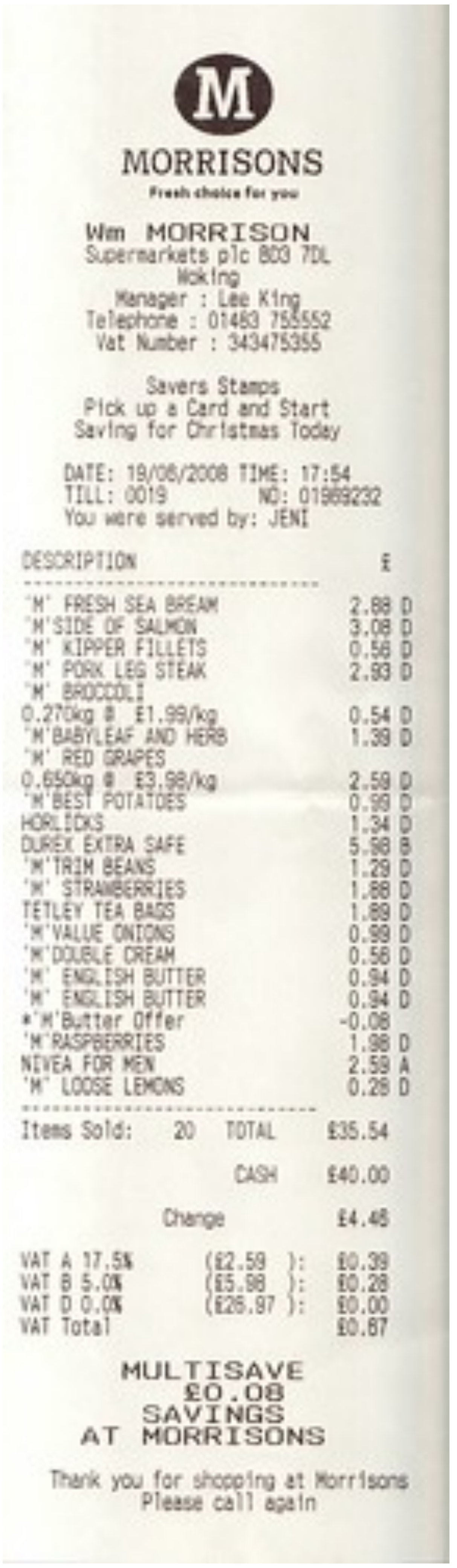
Cambridge Dictionary

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

Wm MORRISON
Supermarkets plc BD3 7DL
Woking
Manager : Lee King
Telephone : 01483 755552
Vat Number : 343475355

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

DATE: 19/06/2008 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 @ £3.98/kg	2.59 D
'M'BEST POTATOES	0.99 D
HORLICKS	1.34 D
DUREX EXTRA SAFE	5.98 B
'M'TRIM BEANS	1.29 D
'M' STRAWBERRIES	1.88 D
TETLEY TEA BAGS	1.89 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	2.59 A
'M' LOOSE LEMONS	0.28 D

Items Sold: 20 TOTAL £35.54

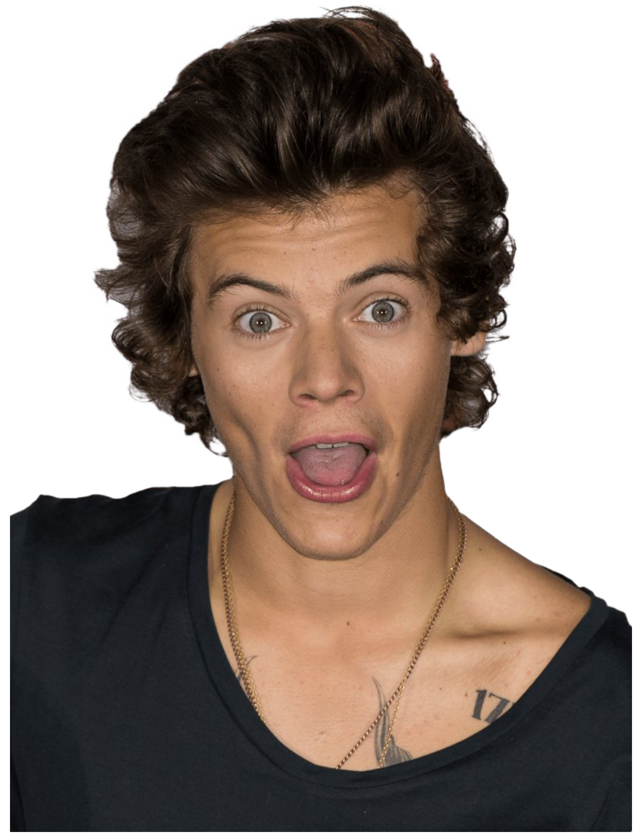
CASH £40.00

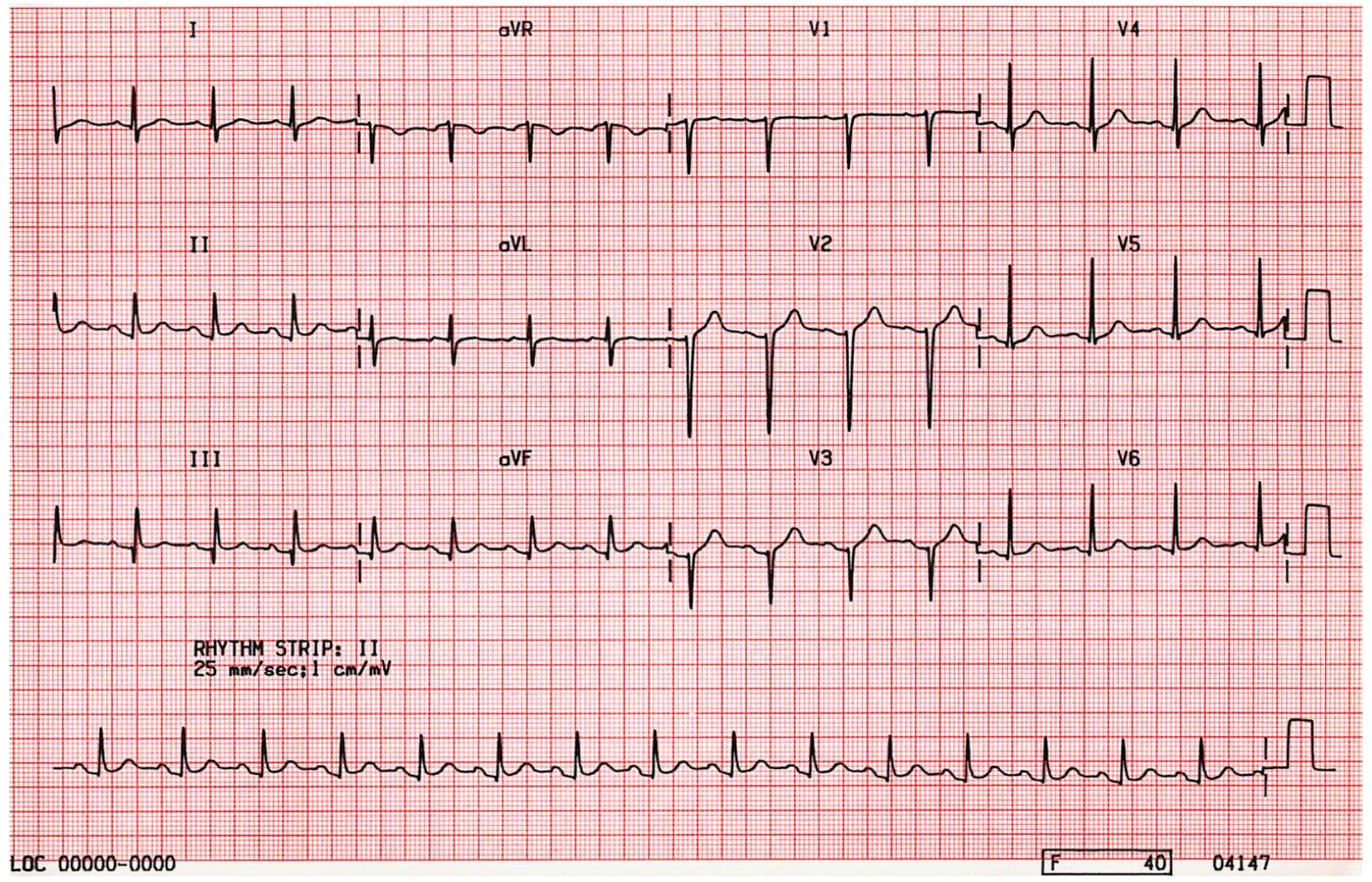
Change £4.46

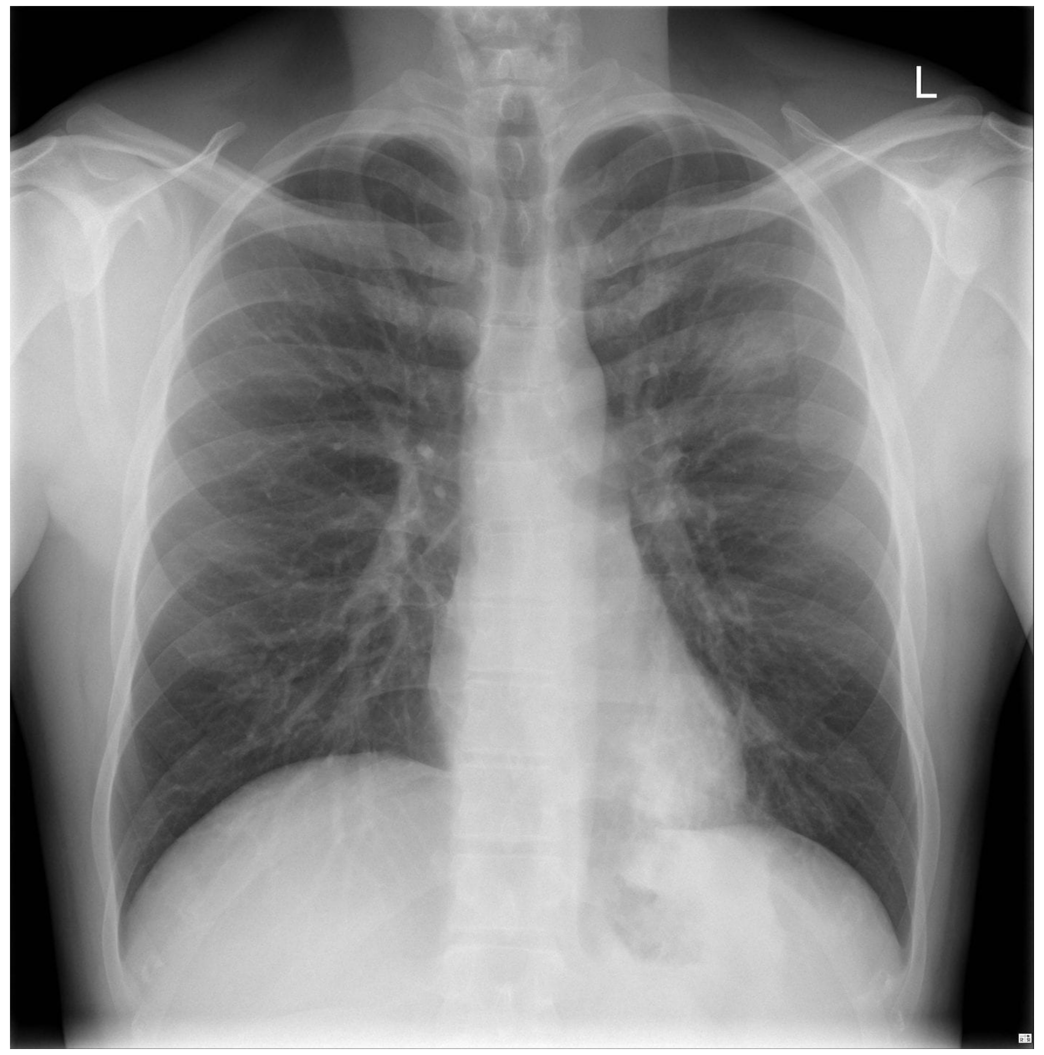
VAT A 17.5% (£2.59): £0.39
VAT B 5.0% (£5.98): £0.28
VAT D 0.0% (£26.97): £0.00
VAT Total £0.67

MULTISAVE
£0.08
SAVINGS
AT MORRISONS

Thank you for shopping at Morrisons
Please call again







Data Analysis

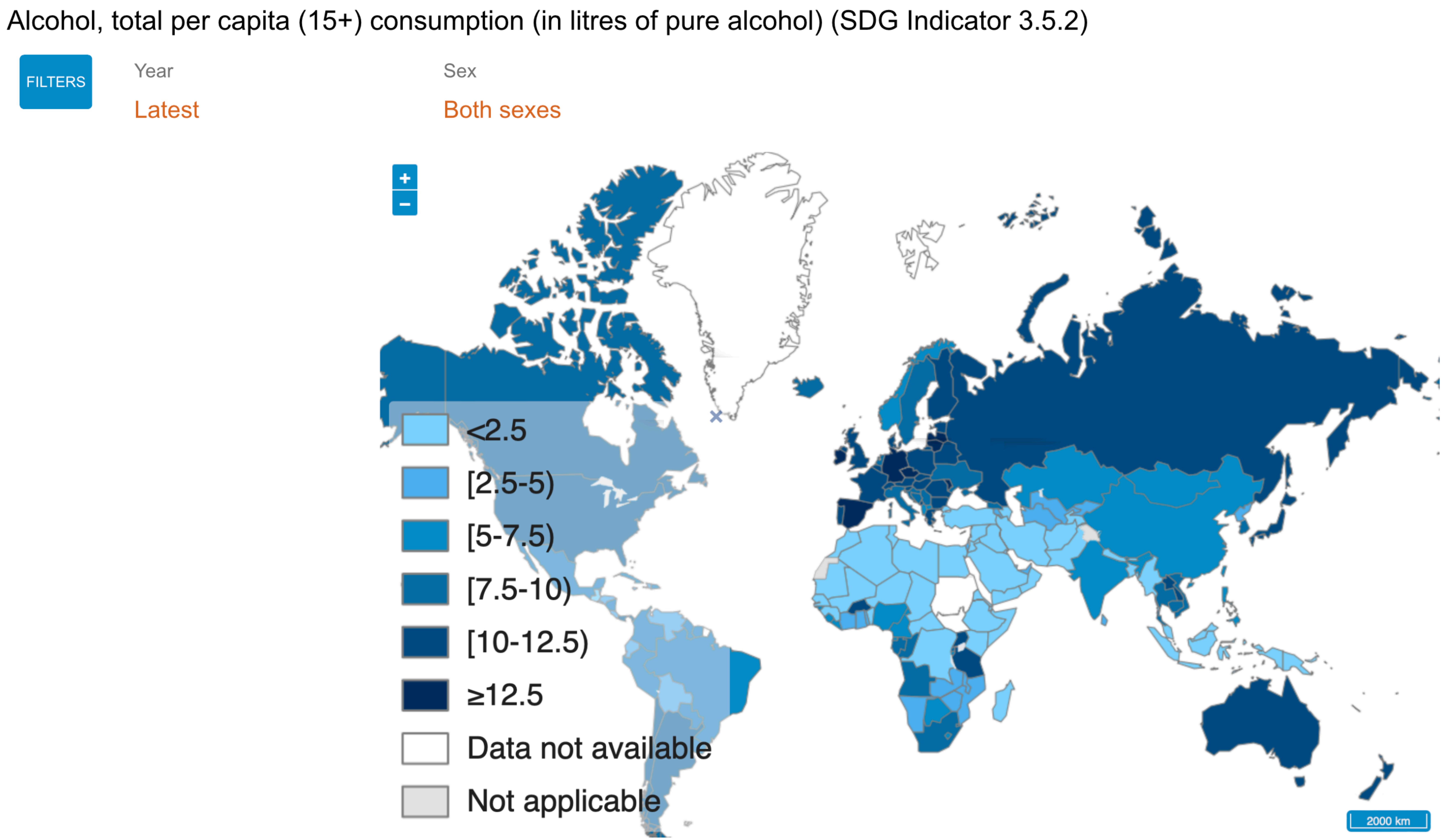
“The process of examining information, especially using a computer, in order to find something out, or to help with making decision”

Cambridge Dictionary

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

Elliot J. Crowley

Spotting patterns



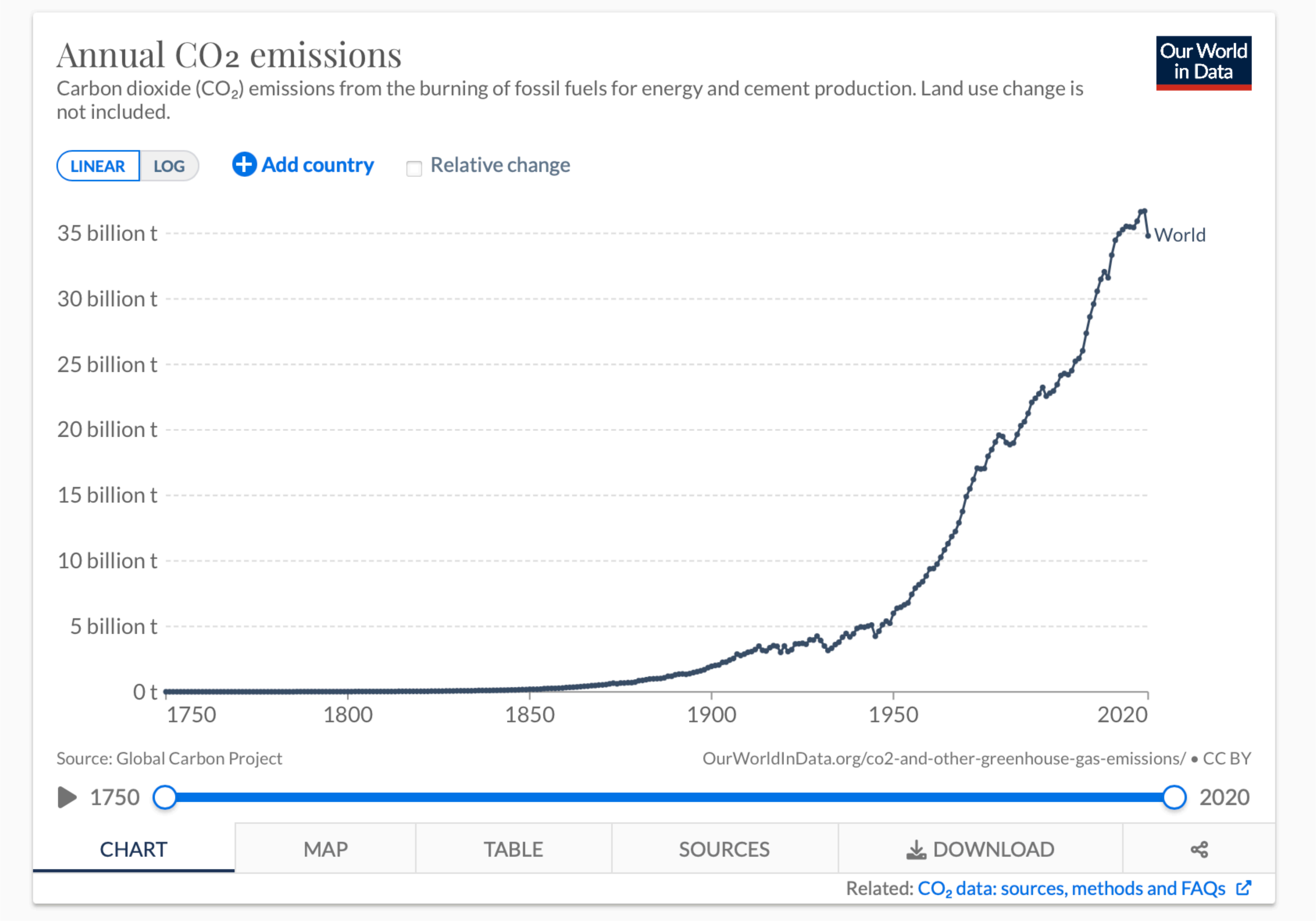
Alcohol, total per capita (15+) consumption (in litres of pure alcohol) (SDG Indicator 3.5.2)

Year: Latest

Sex: Both sexes

Source: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/total-(recorded-unrecorded)-alcohol-per-capita-(15-)-consumption

Observing trends



Annual CO2 emissions

Carbon dioxide (CO_2) emissions from the burning of fossil fuels for energy and cement production. Land use change is not included.

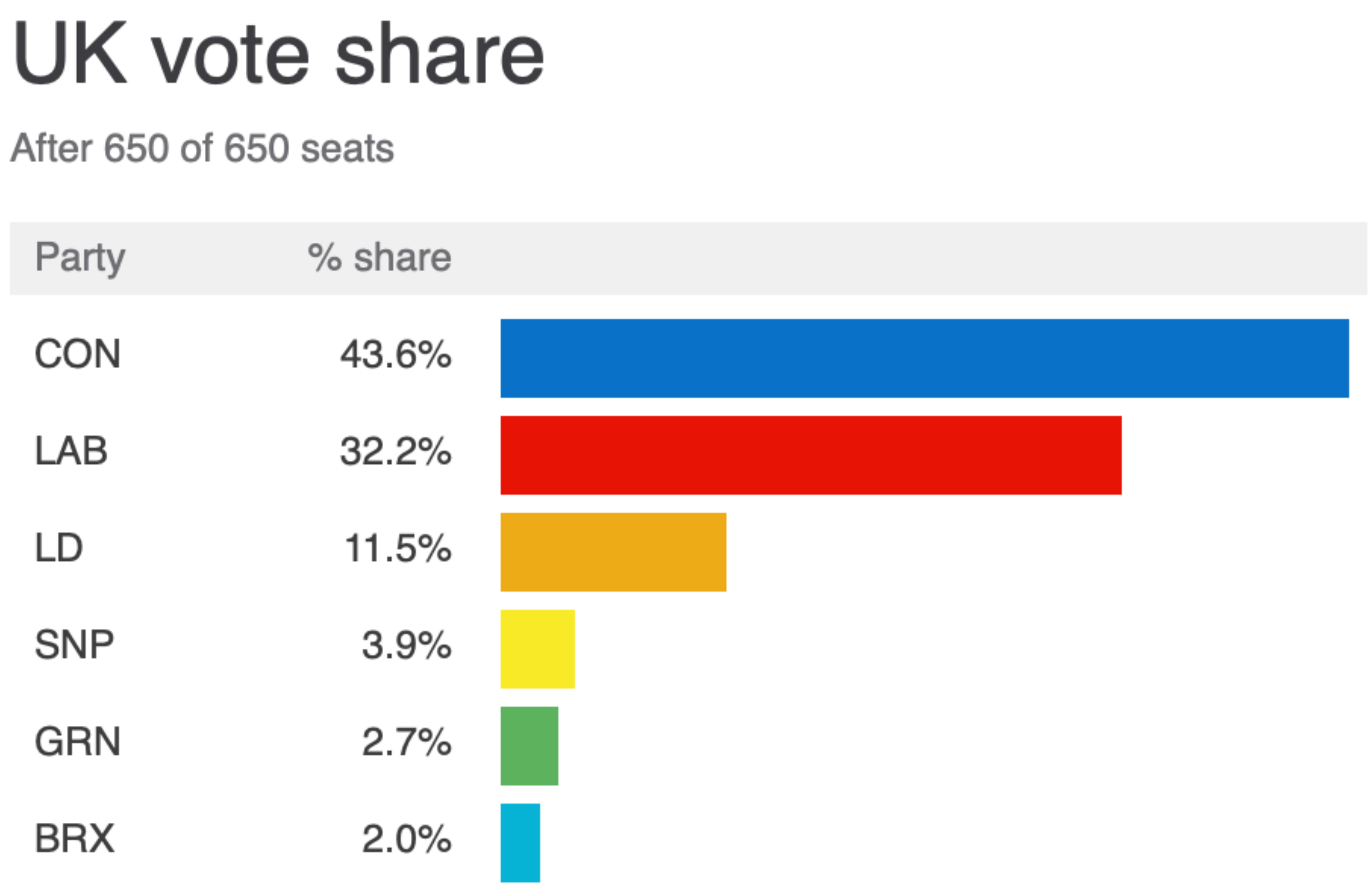
Source: Global Carbon Project

OurWorldInData.org/co2-and-other-greenhouse-gas-emissions/ • CC BY

Related: CO_2 data: sources, methods and FAQs

Source: https://ourworldindata.org/co2-emissions

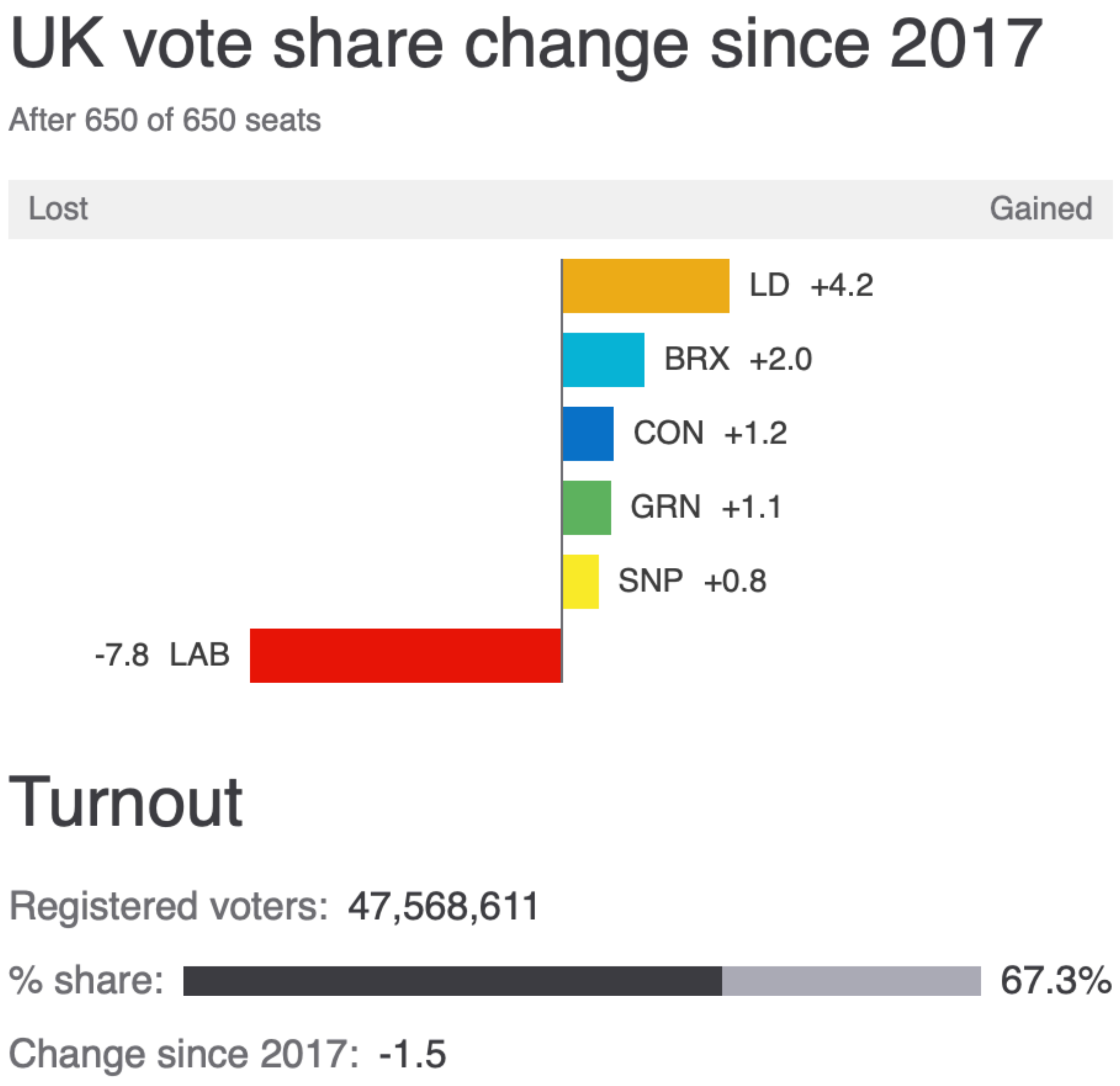
Telling (happy/sad?) stories



UK vote share

After 650 of 650 seats

Party	% share
CON	43.6%
LAB	32.2%
LD	11.5%
SNP	3.9%
GRN	2.7%
BRX	2.0%



UK vote share change since 2017

After 650 of 650 seats

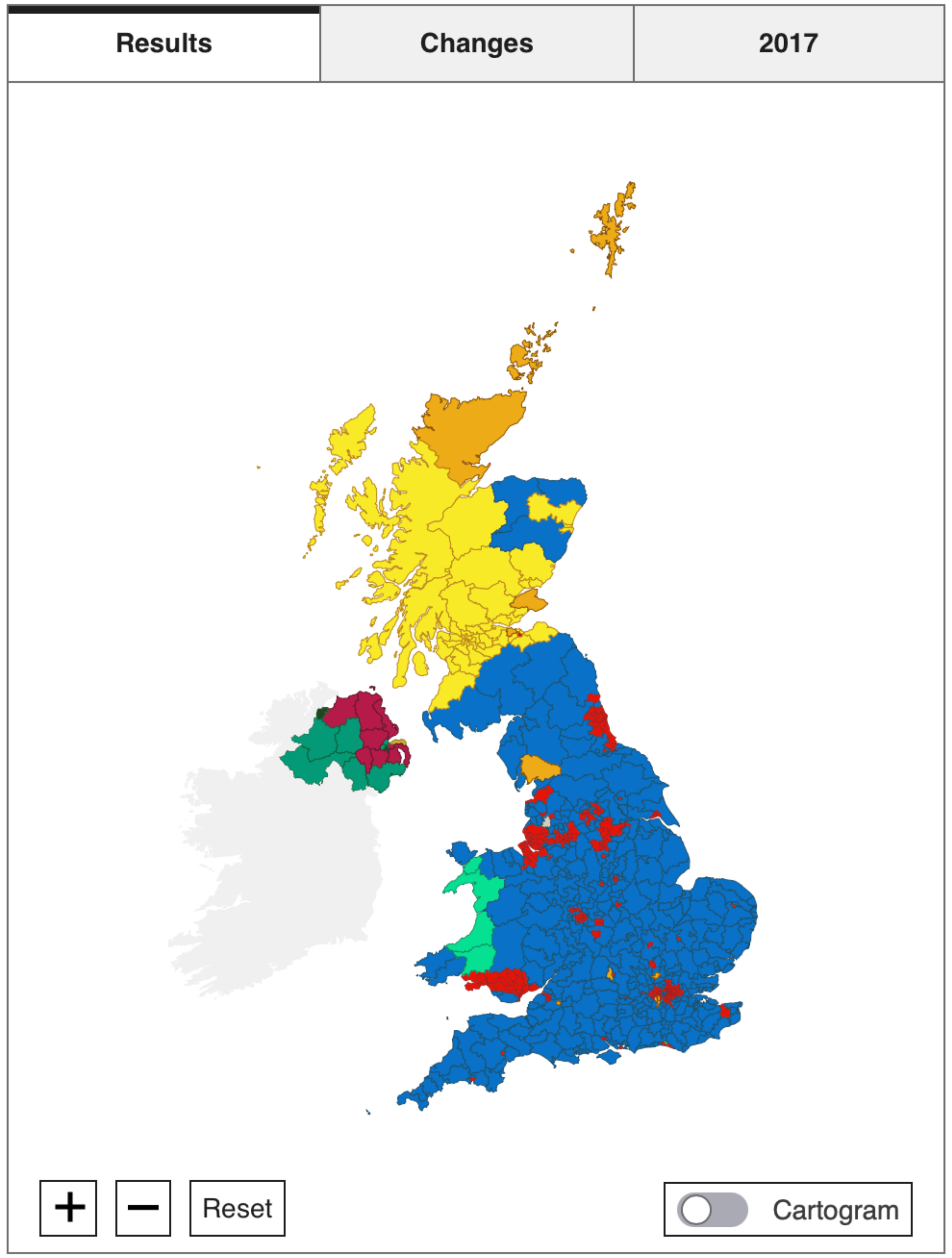
Lost	Gained
	LD +4.2
	BRX +2.0
	CON +1.2
	GRN +1.1
	SNP +0.8
-7.8 LAB	

Turnout

Registered voters: 47,568,611

% share: 67.3%

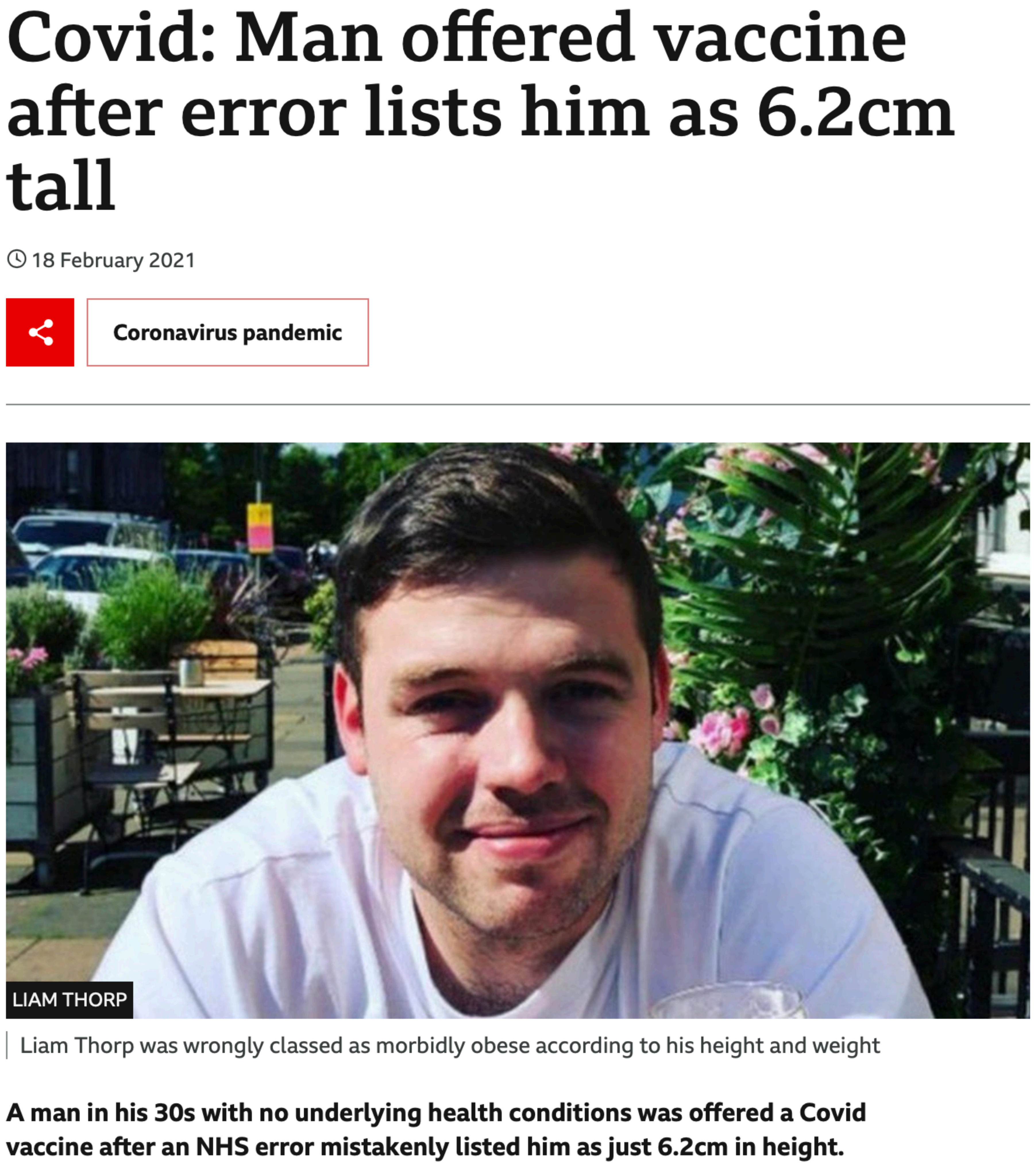
Change since 2017: -1.5



The Cartogram map shows the UK's 650 parliamentary seats as if they are hexagons of the same size. Hexagons by Esri

Source: https://www.bbc.co.uk/news/election/2019/results

Finding anomalies



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

18 February 2021

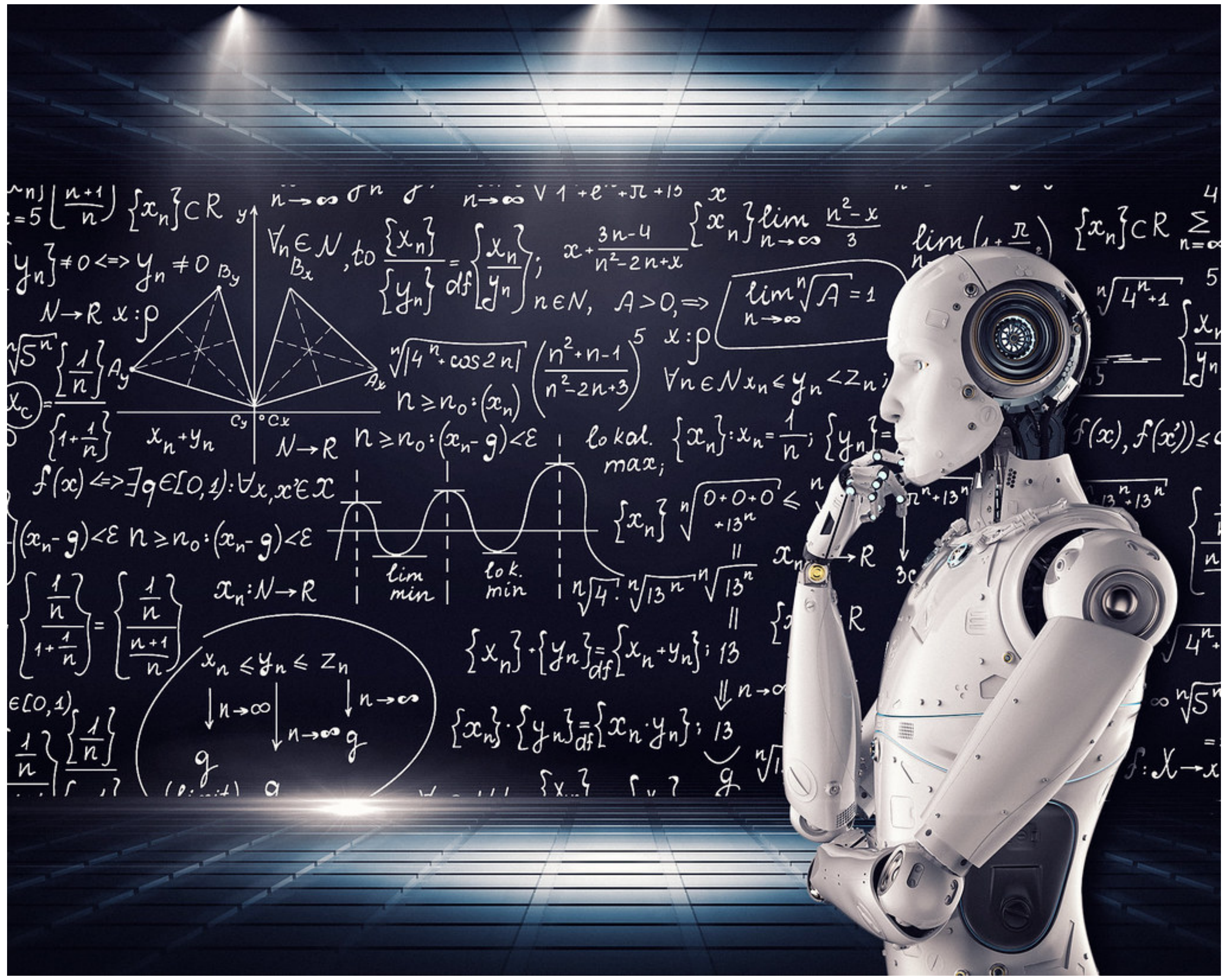
Liam Thorp was wrongly classed as morbidly obese according to his height and weight

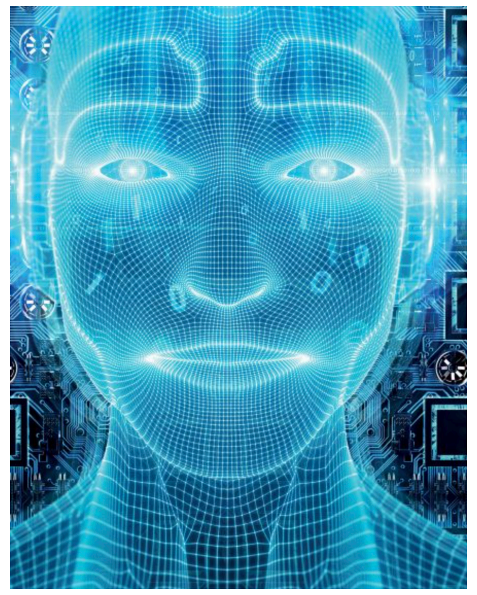
A man in his 30s with no underlying health conditions was offered a Covid vaccine after an NHS error mistakenly listed him as just 6.2cm in height.

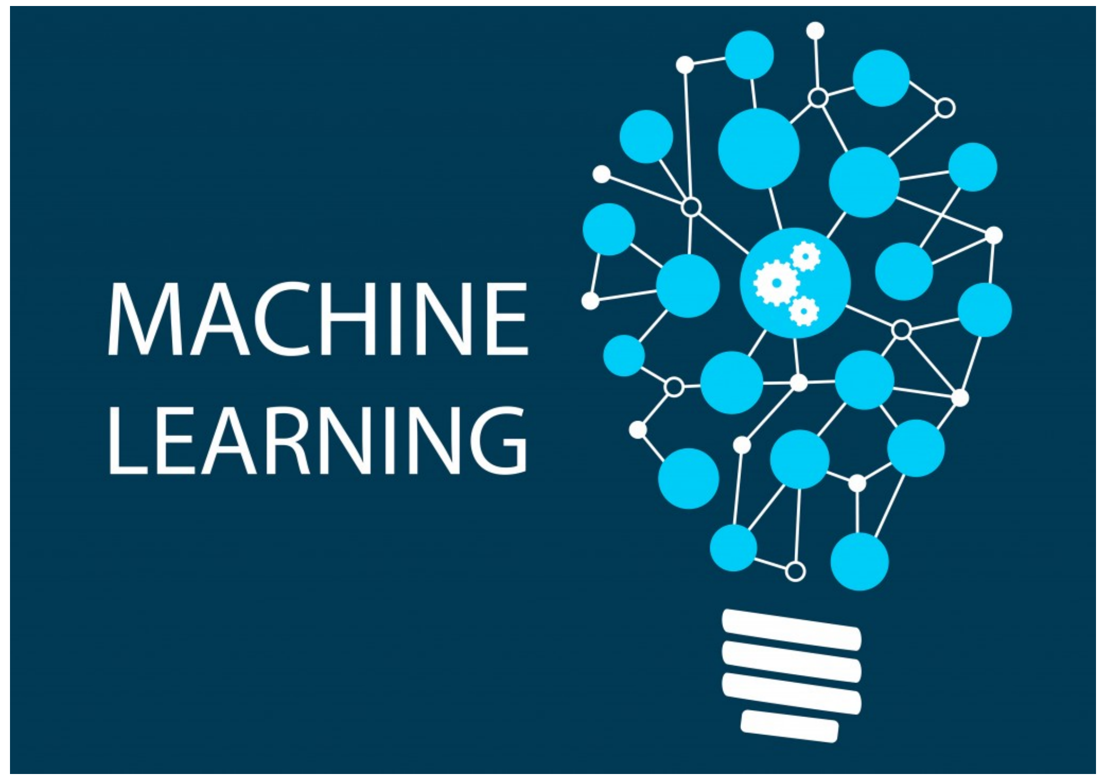
Source: https://www.bbc.co.uk/news/uk-england-merseyside-56111209

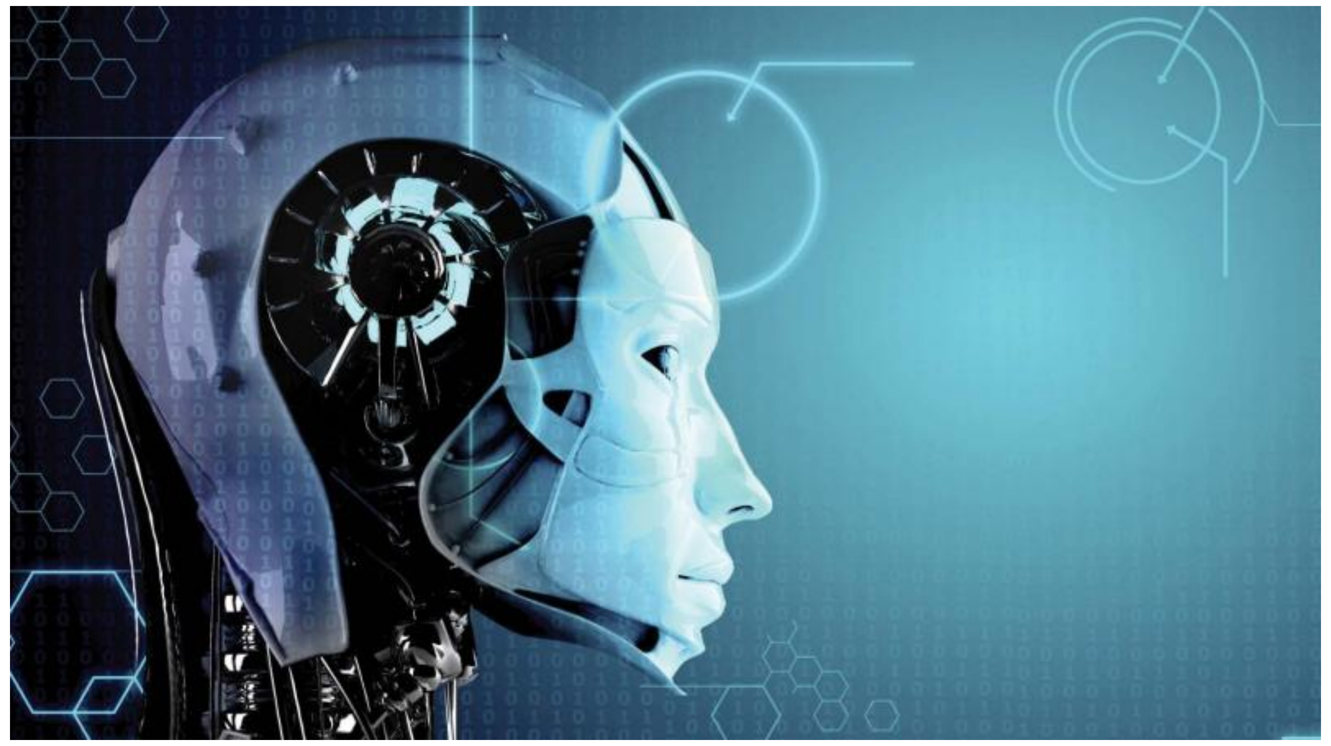
What is Machine Learning?

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









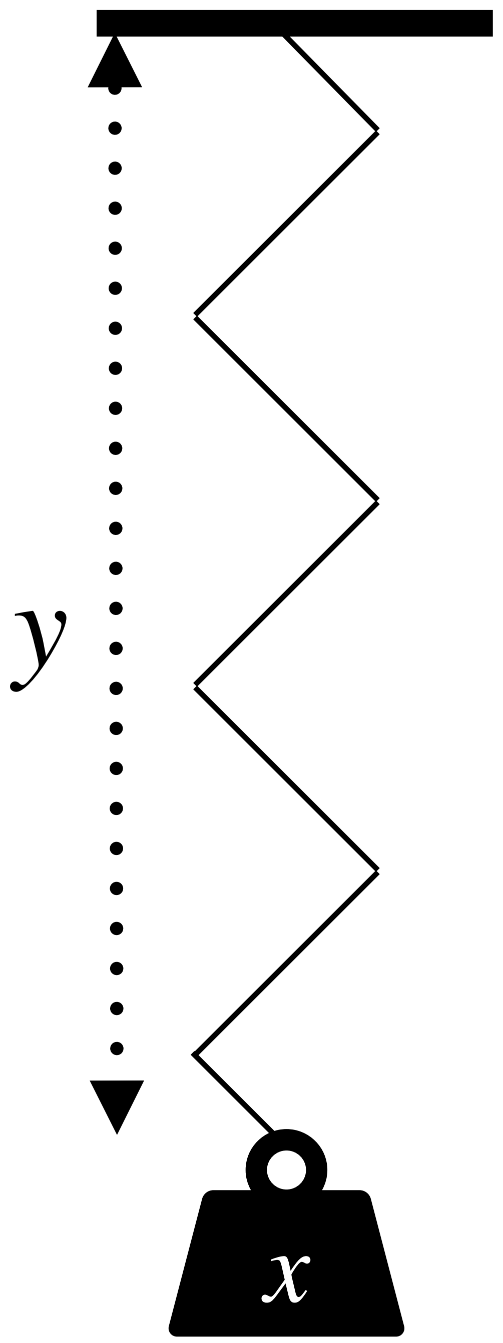
Machine Learning is…

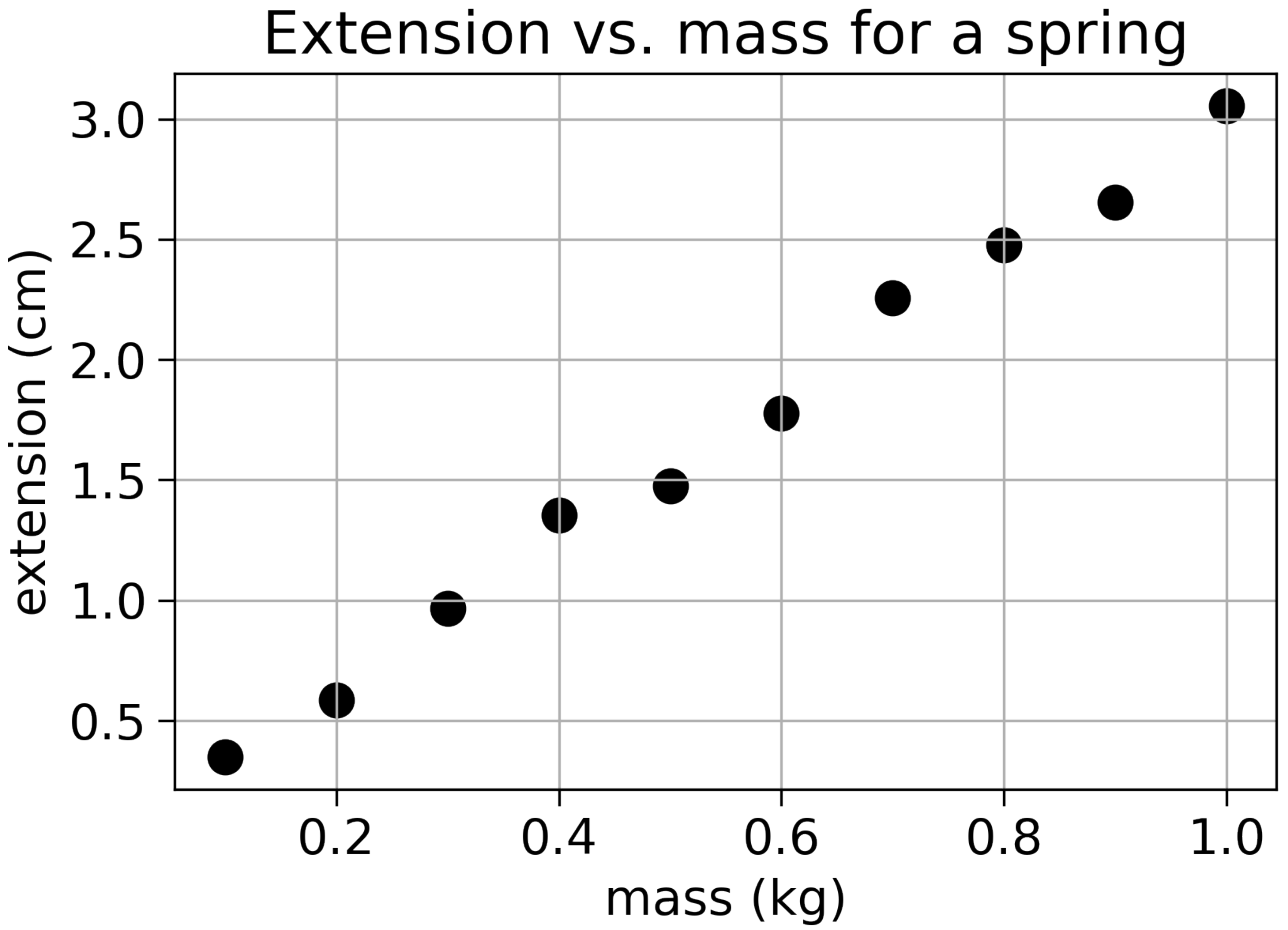
> “the study of algorithms that can learn from training data in order to make predictions on new data.”

Elliot J. Crowley

Machine Learning for a spring

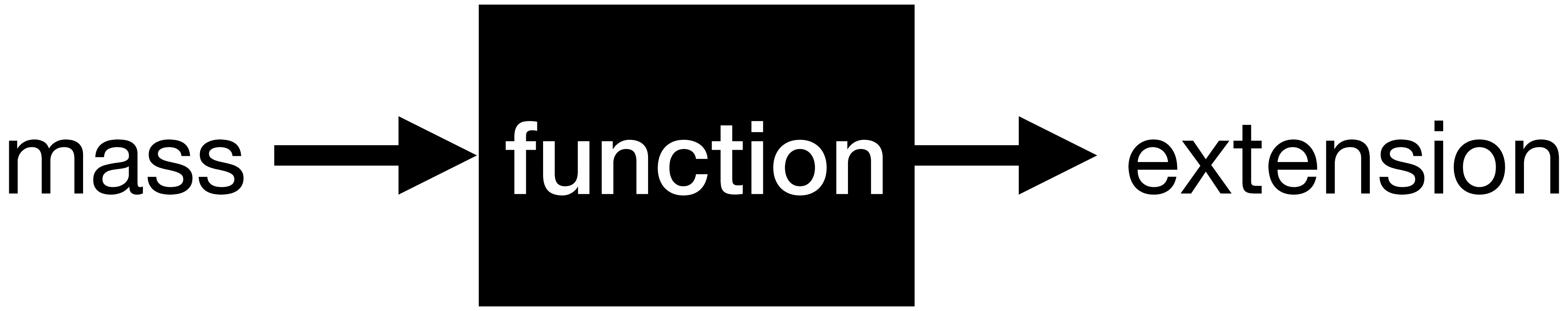
- We want a model that given an arbitrary mass x can predict extension y
- We can attach some masses to the spring and record its extension
- These mass-spring measurements form our **training data**





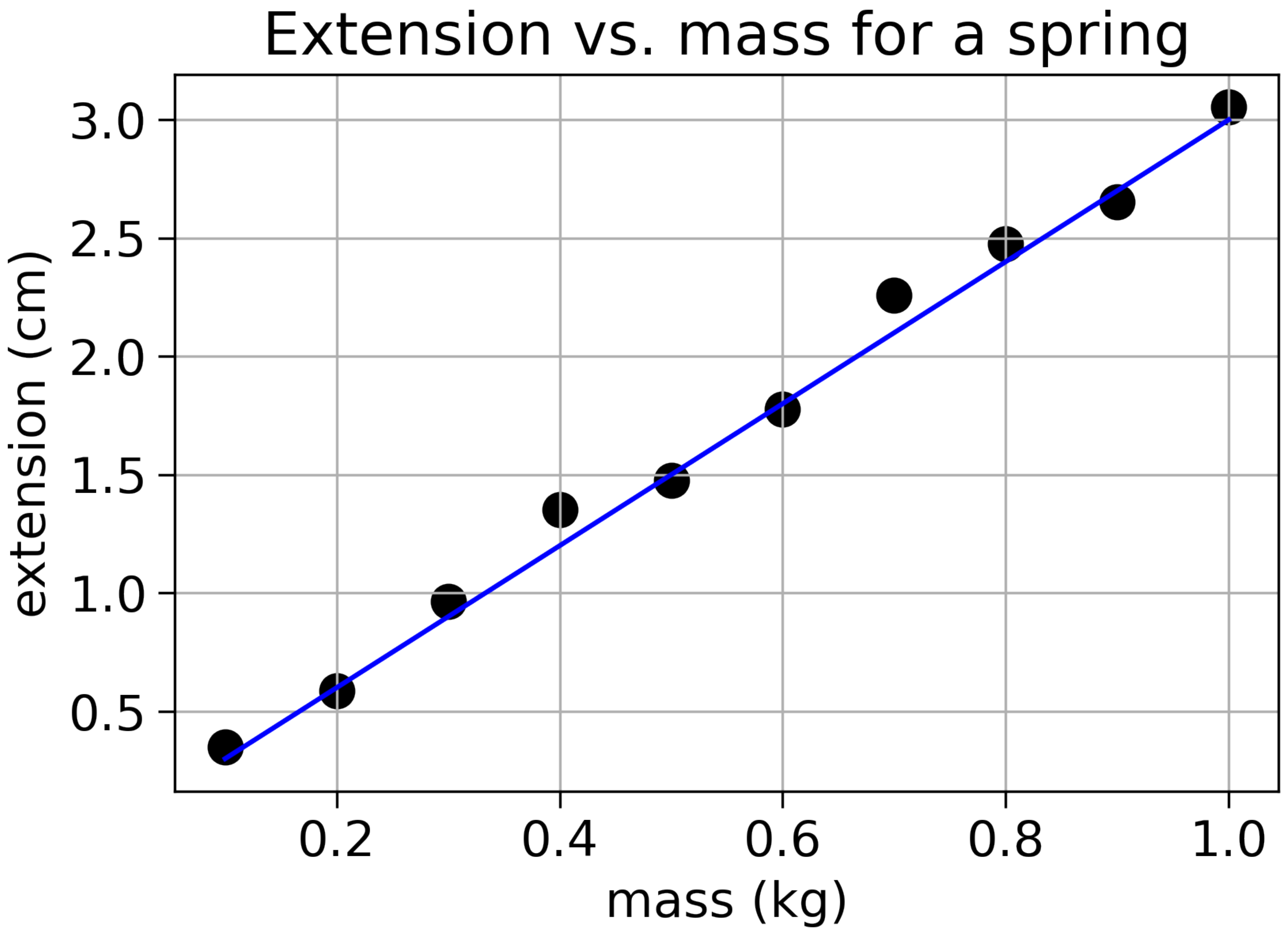
Machine Learning for a spring

- Will 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

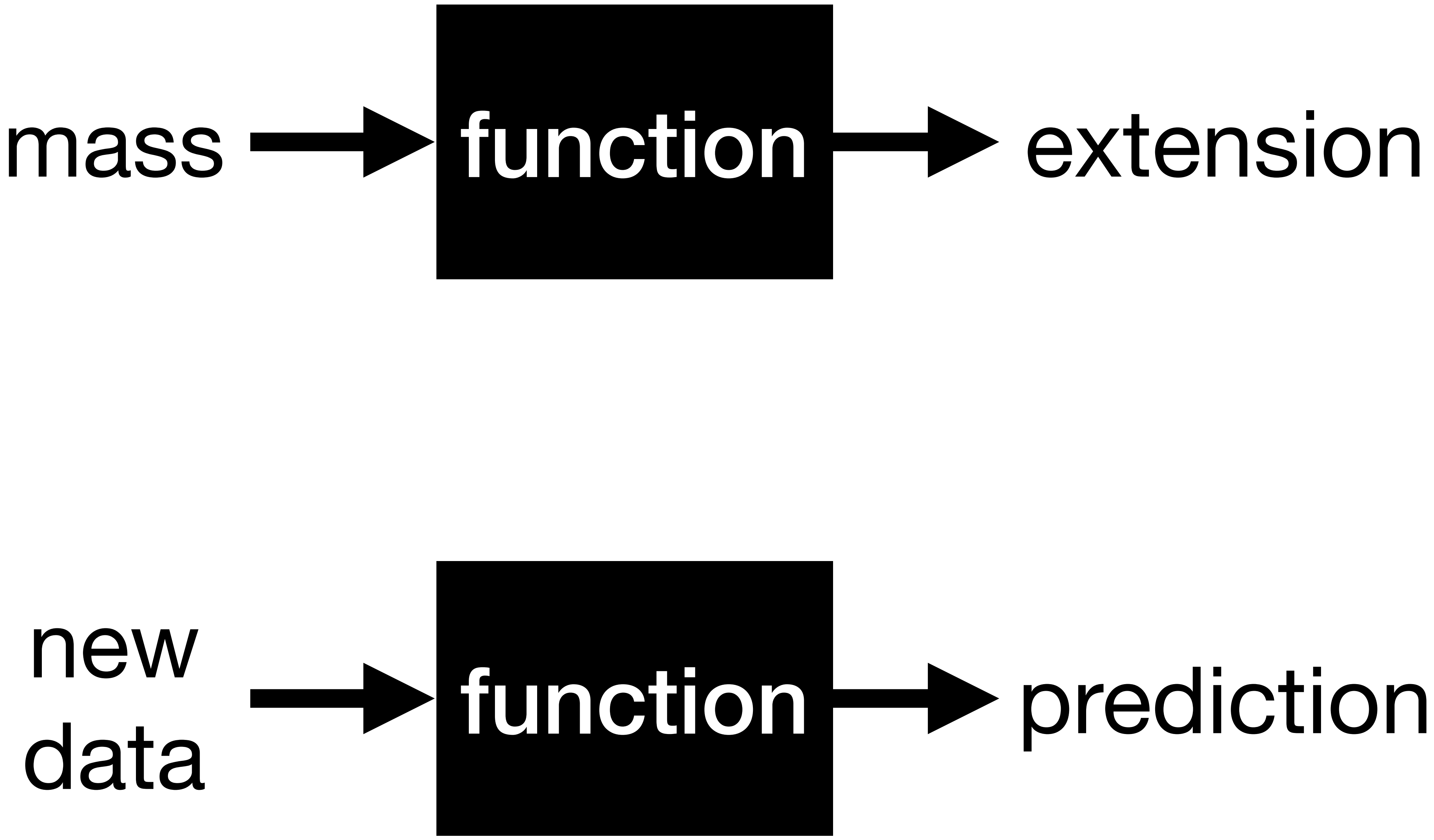

mass

function

extension



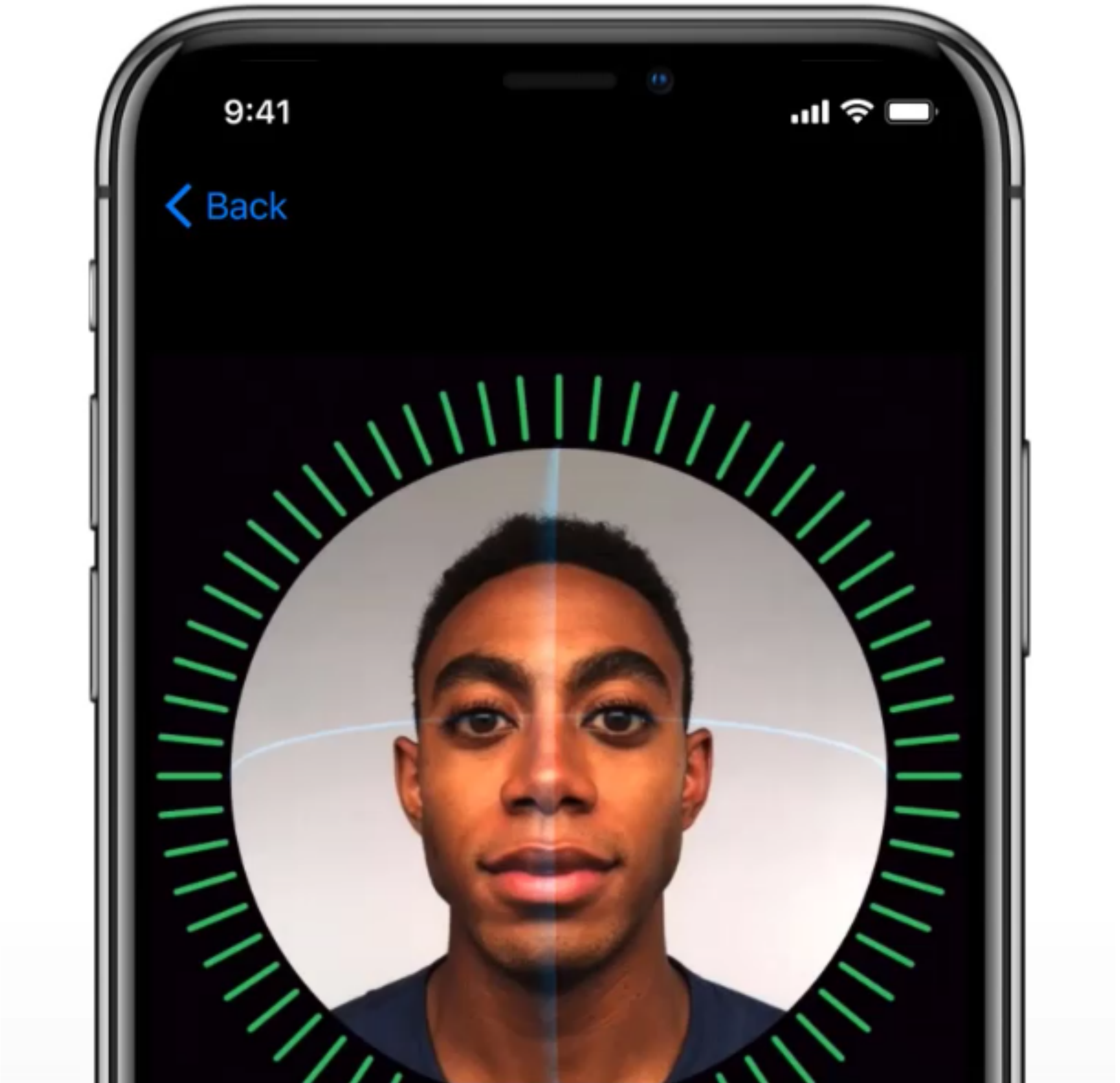
Is that it?

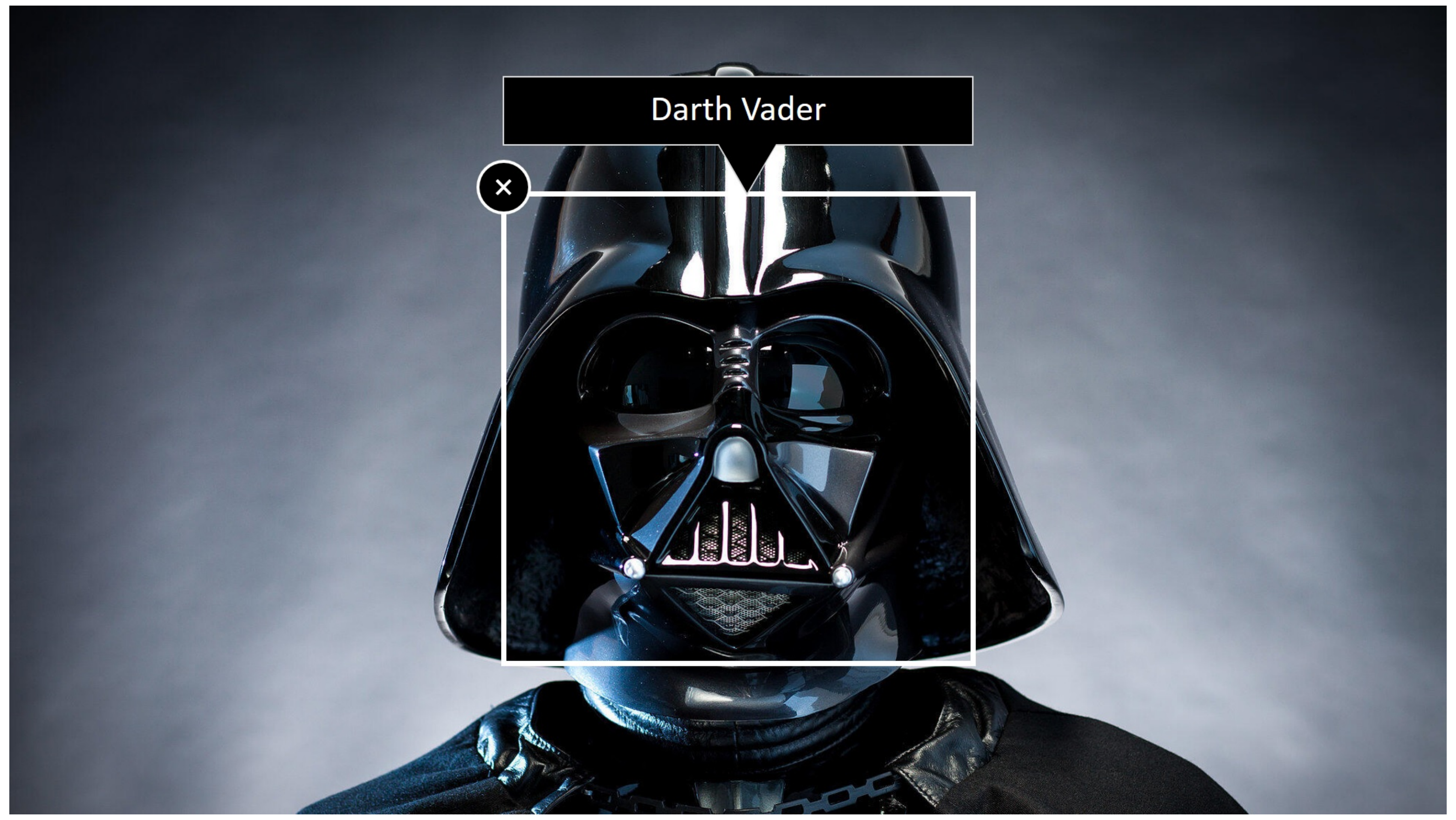


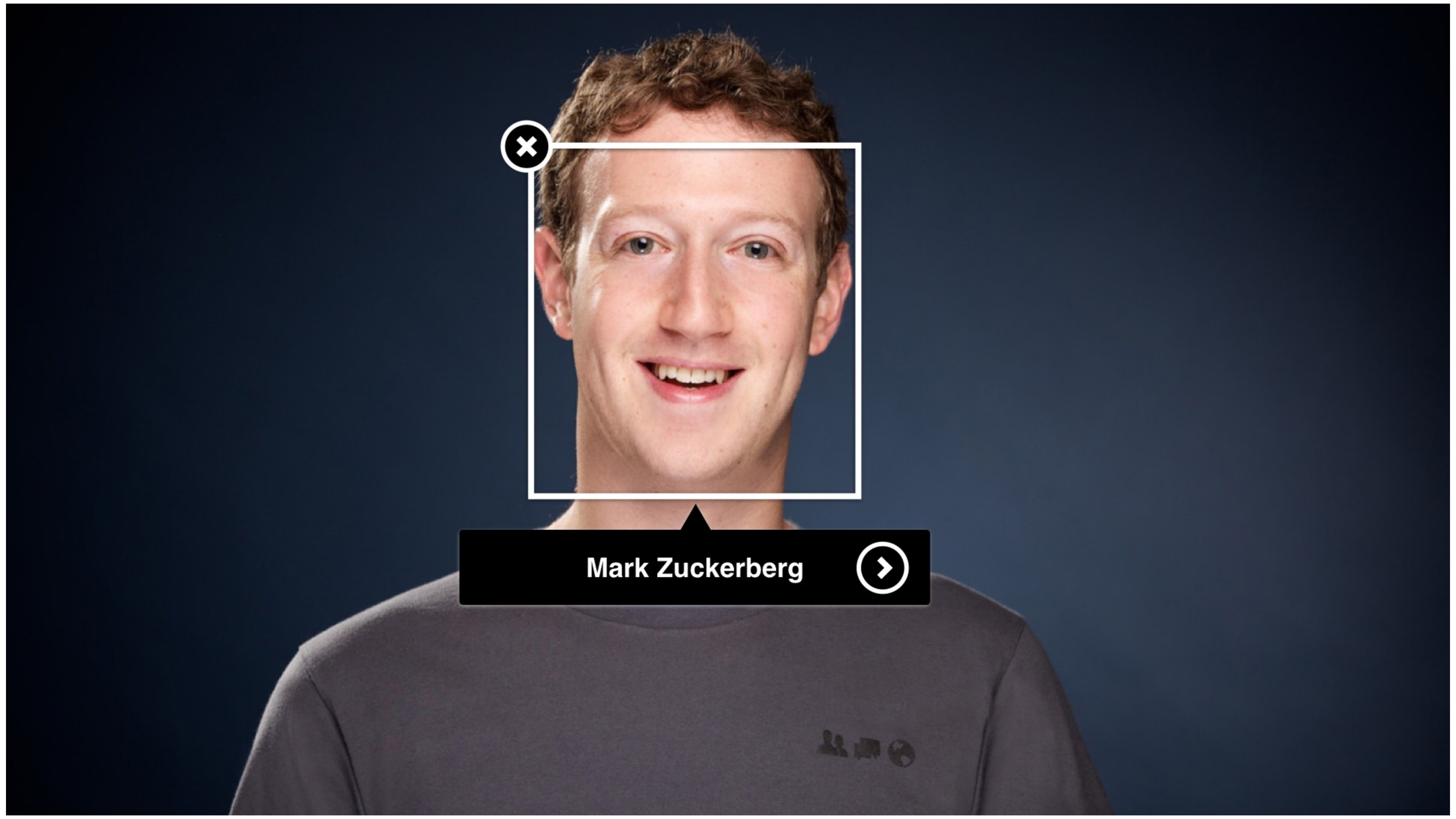
mass → function → extension

new data → function → prediction

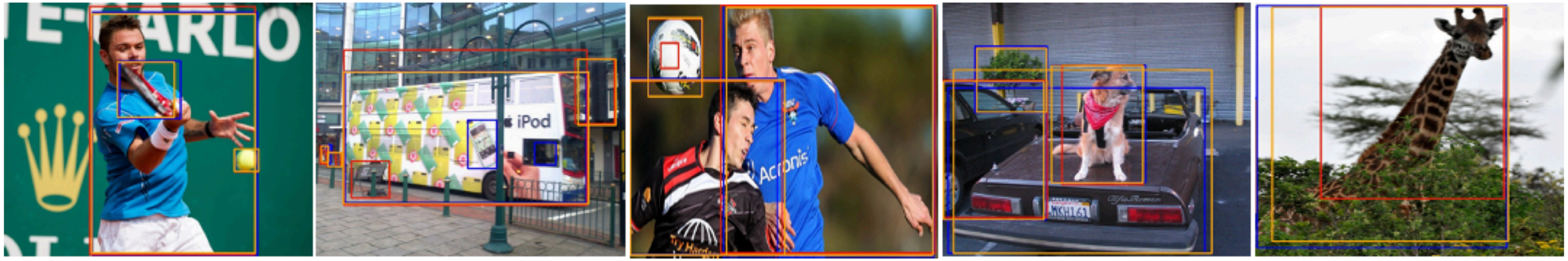
Face recognition

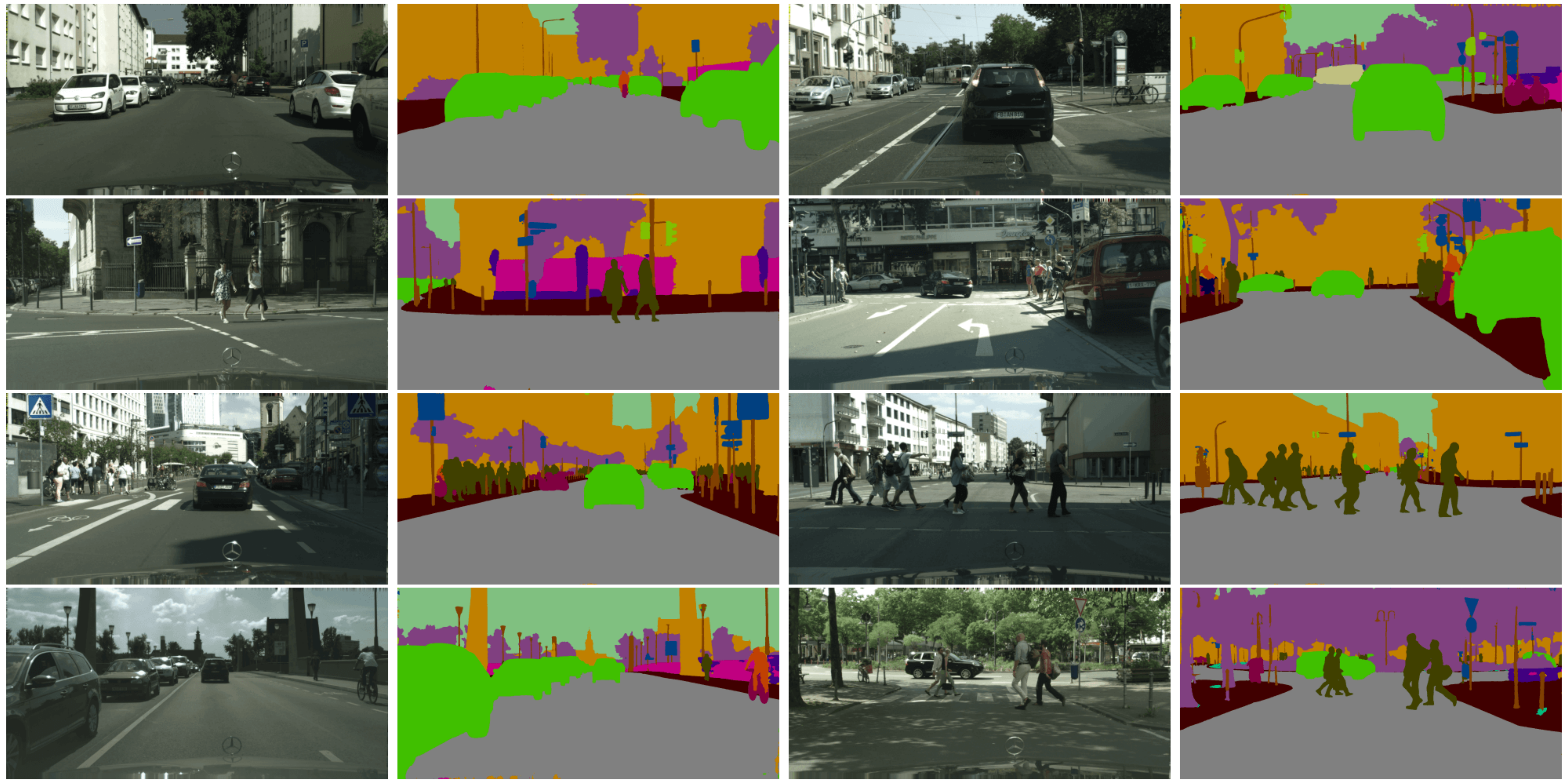






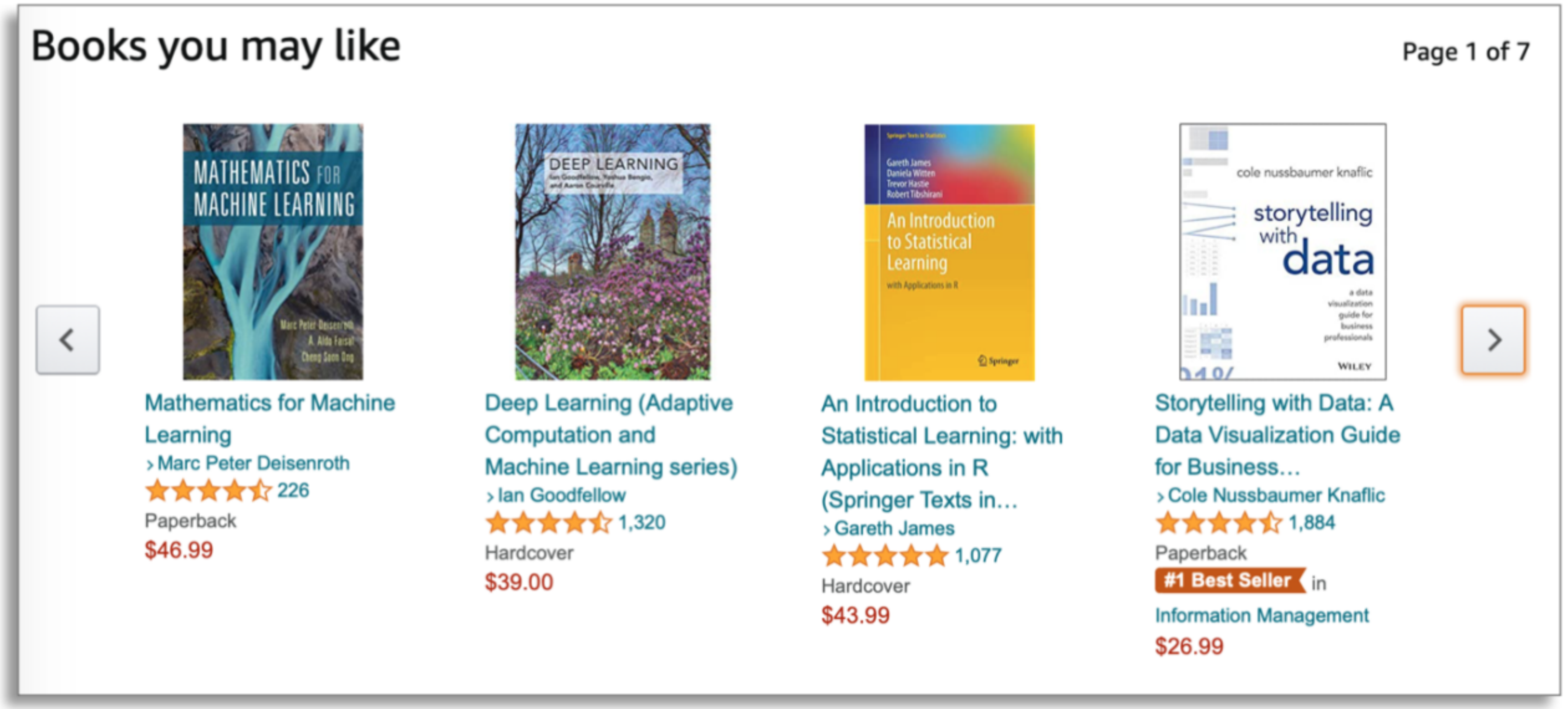
Detection and segmentation





Source: https://github.com/locuslab/mdeq/blob/master/images/segmentation_example.png

Recommender systems



Books you may like

Page 1 of 7

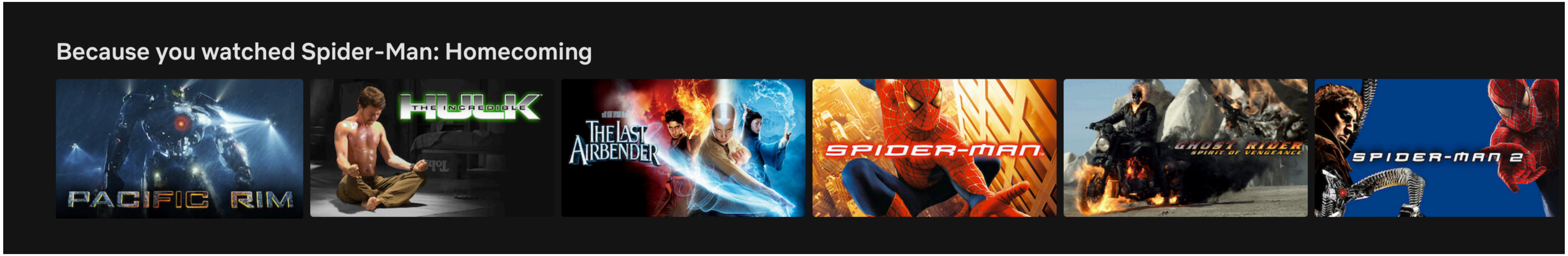
Mathematics for Machine Learning
› Marc Peter Deisenroth
★★★★☆ 226
Paperback
$46.99

Deep Learning (Adaptive Computation and Machine Learning series)
› Ian Goodfellow
★★★★☆ 1,320
Hardcover
$39.00

An Introduction to Statistical Learning: with Applications in R (Springer Texts in…
› 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

MATHEMATICS FOR MACHINE LEARNING
DEEP LEARNING
An Introduction to Statistical Learning
storytelling with data



Because you watched Spider-Man: Homecoming

Text to image

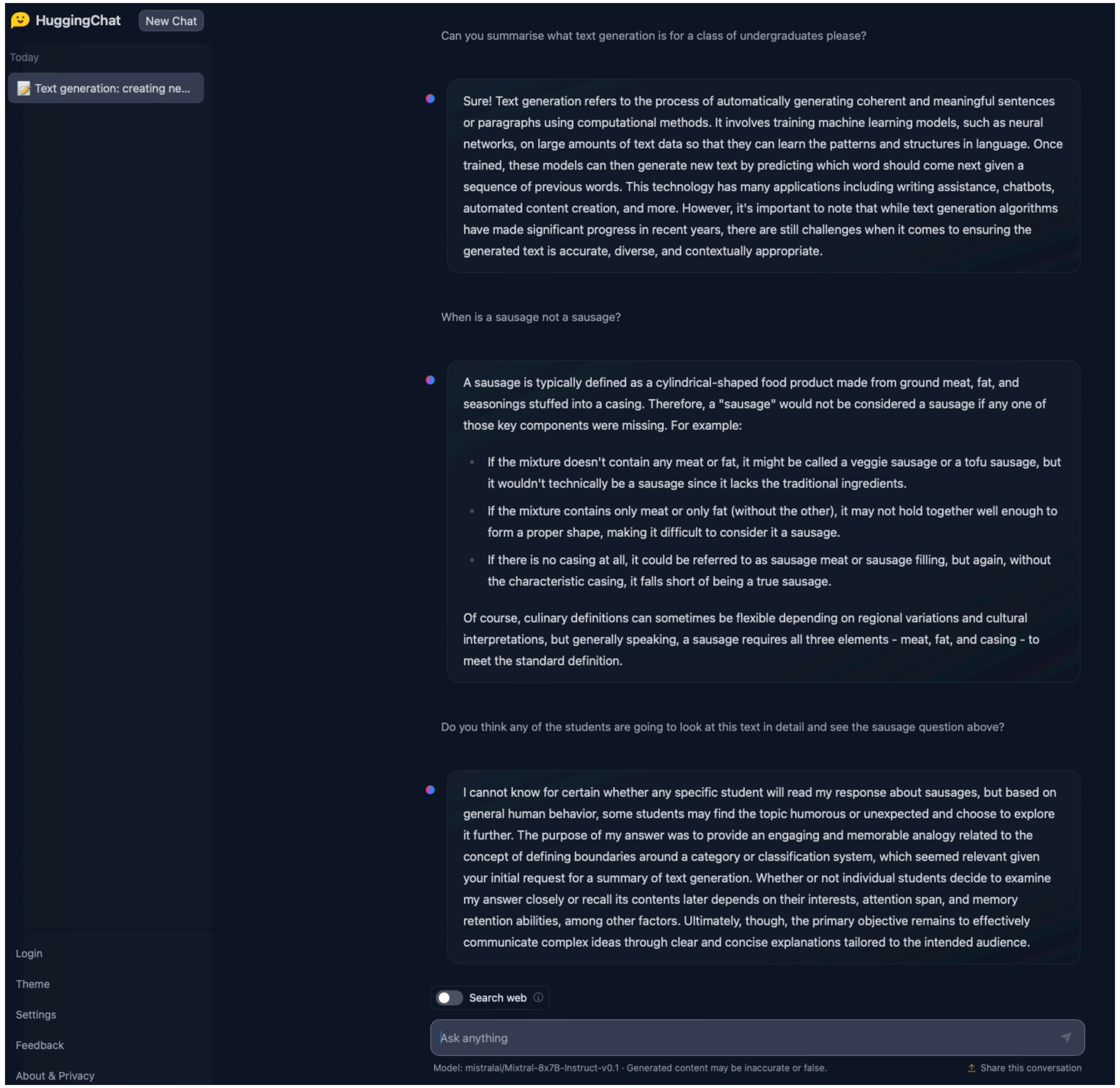
Teddy bears swimming at the Olympics 400m Butterfly event.

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.

Source: https://imagen.research.google

Text generation



Can you summarise what text generation is for a class of undergraduates please?

Sure! Text generation refers to the process of automatically generating coherent and meaningful sentences or paragraphs using computational methods. It involves training machine learning models, such as neural networks, on large amounts of text data so that they can learn the patterns and structures in language. Once trained, these models can then generate new text by predicting which word should come next given a sequence of previous words. This technology has many applications including writing assistance, chatbots, automated content creation, and more. However, it's important to note that while text generation algorithms have made significant progress in recent years, there are still challenges when it comes to ensuring the generated text is accurate, diverse, and contextually appropriate.

When is a sausage not a sausage?

A sausage is typically defined as a cylindrical-shaped food product made from ground meat, fat, and seasonings stuffed into a casing. Therefore, a "sausage" would not be considered a sausage if any one of those key components were missing. For example:

- If the mixture doesn't contain any meat or fat, it might be called a veggie sausage or a tofu sausage, but it wouldn't technically be a sausage since it lacks the traditional ingredients.
- If the mixture contains only meat or only fat (without the other), it may not hold together well enough to form a proper shape, making it difficult to consider it a sausage.
- If there is no casing at all, it could be referred to as sausage meat or sausage filling, but again, without the characteristic casing, it falls short of being a true sausage.

Of course, culinary definitions can sometimes be flexible depending on regional variations and cultural interpretations, but generally speaking, a sausage requires all three elements - meat, fat, and casing - to meet the standard definition.

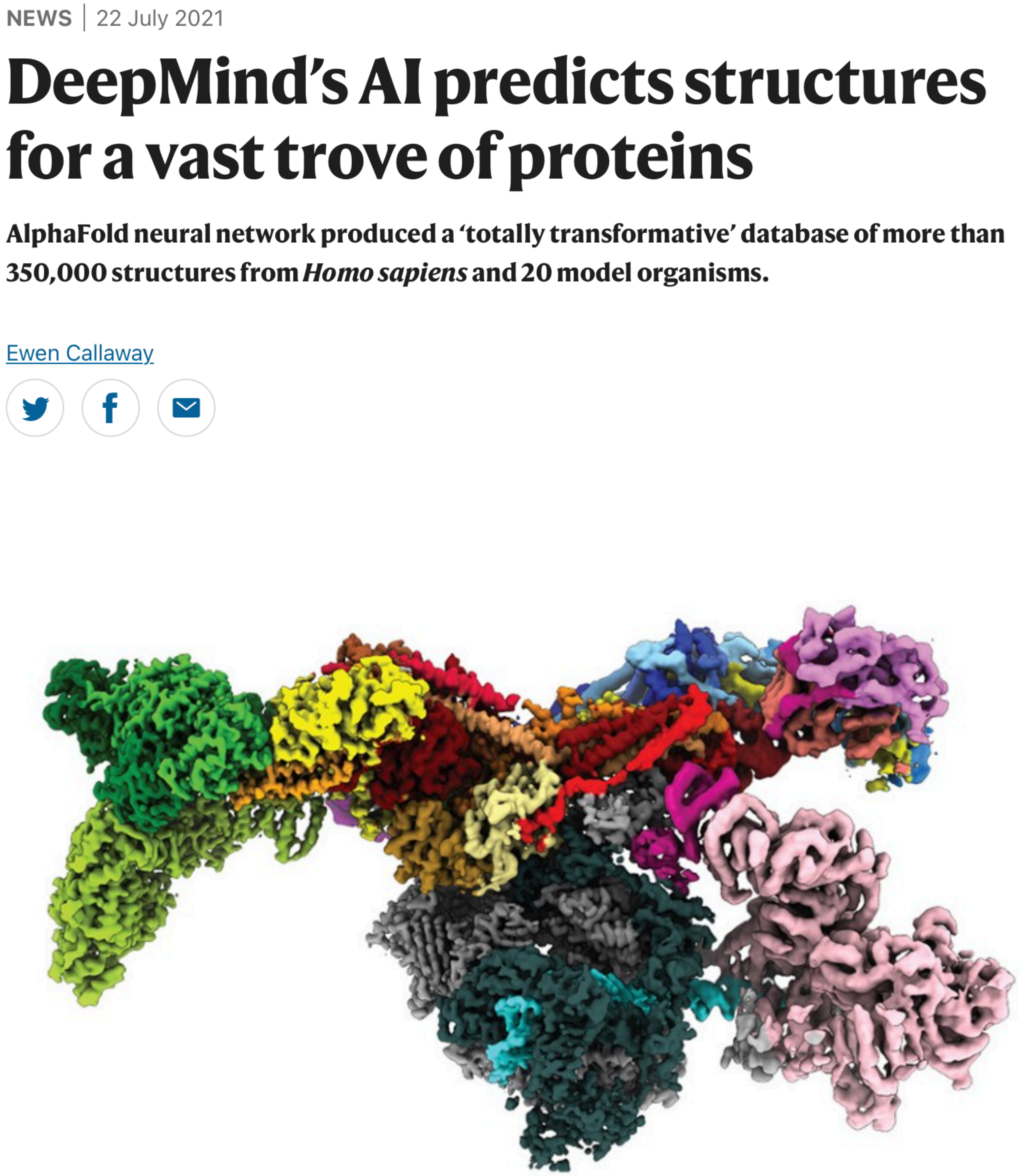
Do you think any of the students are going to look at this text in detail and see the sausage question above?

I cannot know for certain whether any specific student will read my response about sausages, but based on general human behavior, some students may find the topic humorous or unexpected and choose to explore it further. The purpose of my answer was to provide an engaging and memorable analogy related to the concept of defining boundaries around a category or classification system, which seemed relevant given your initial request for a summary of text generation. Whether or not individual students decide to examine my answer closely or recall its contents later depends on their interests, attention span, and memory retention abilities, among other factors. Ultimately, though, the primary objective remains to effectively communicate complex ideas through clear and concise explanations tailored to the intended audience.

Model: mistralai/Mixtral-8x7B-Instruct-v0.1 · Generated content may be inaccurate or false.

https://huggingface.co/chat/

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

The course

This is the second year the course has run

- I hope you enjoy it
- There have been quite a few changes since last year
- Feedback is very welcome

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, data modalities, variable types
2. Summarising and visualising data
3. Preprocessing data, principal component analysis, clustering
4. Machine learning and ethics
5. Linear models for regression
6. Linear models for classification
7. Model selection and evaluation
8. Classification and regression trees, bagging and boosting
9. Gaussian processes
10. Deep neural networks

Data Analysis

Machine Learning

Course format

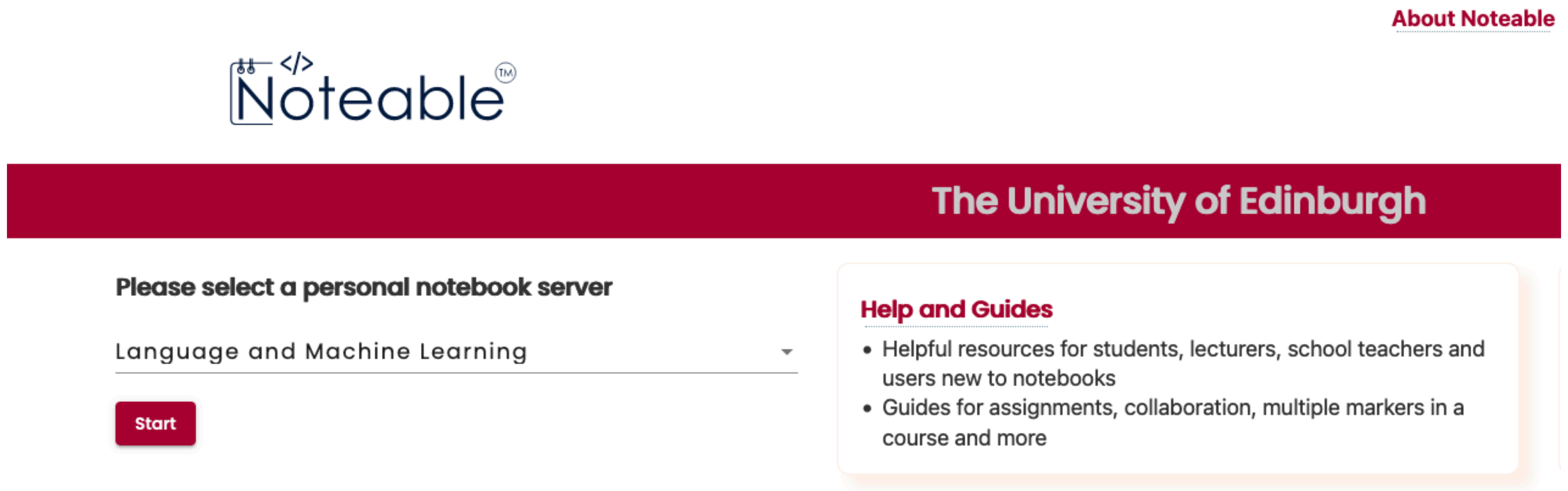
Each week's teaching consists of **lecture (Monday AM) → lab (Thursday PM)**

- In the **lecture** you are taught material
- In the **lab** 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 lab.

This is an applied course.
Attending the labs is **essential**

Notable + Jupyter notebooks



About Noteable

Noteable

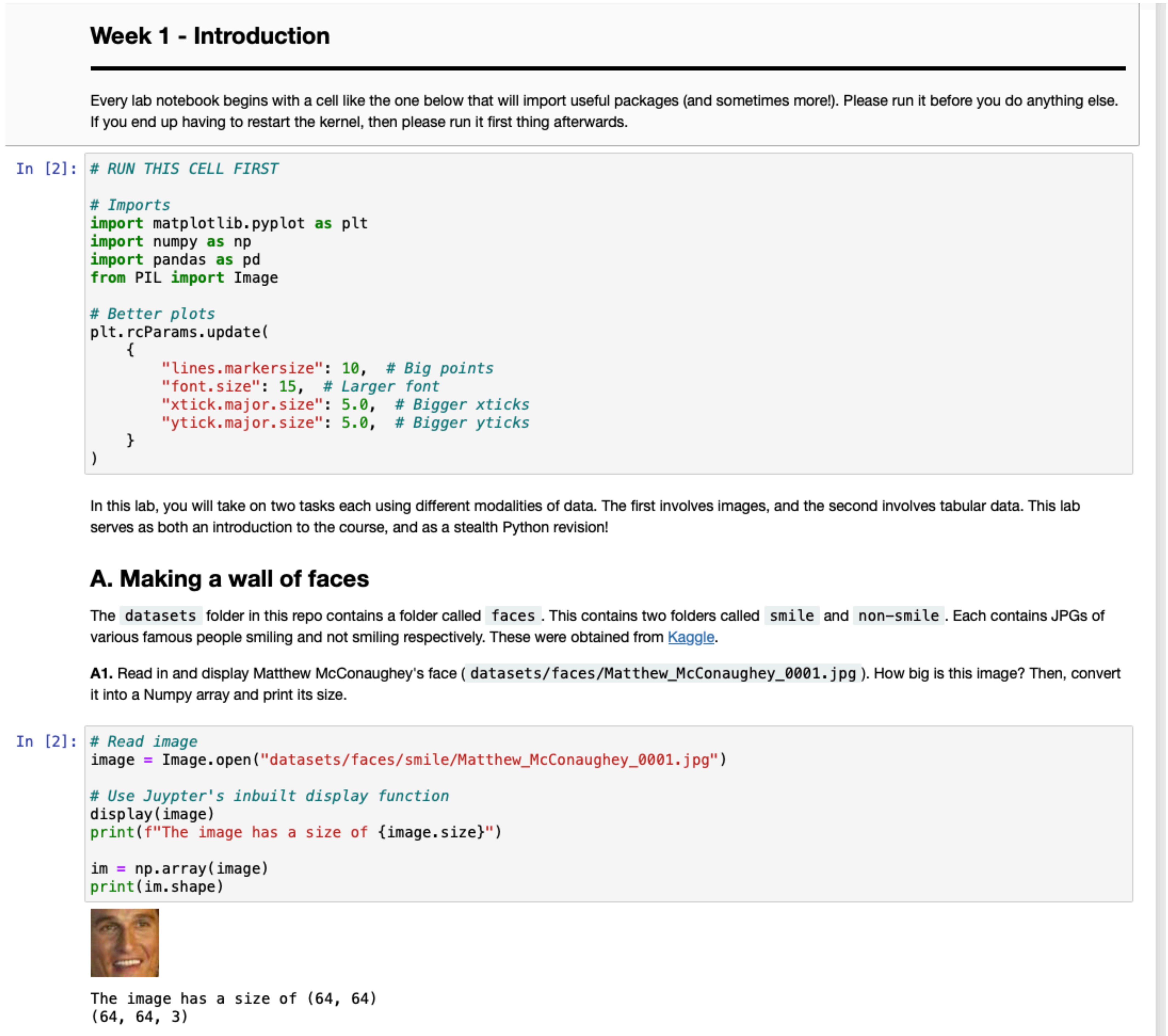
The University of Edinburgh

Please select a personal notebook server

Language and Machine Learning

Help and Guides

- Helpful resources for students, lecturers, school teachers and users new to notebooks
- Guides for assignments, collaboration, multiple markers in a course and more



Week 1 - Introduction

Every lab notebook begins with a cell like the one below that will import useful packages (and sometimes more!). Please run it before you do anything else. If you end up having to restart the kernel, then please run it first thing afterwards.

In [2]:

```
# RUN THIS CELL FIRST

# Imports
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from PIL import Image

# Better plots
plt.rcParams.update(
    {
        "lines.markersize": 10,  # Big points
        "font.size": 15,  # Larger font
        "xtick.major.size": 5.0,  # Bigger xticks
        "ytick.major.size": 5.0,  # Bigger yticks
    }
)
```

In this lab, you will take on two tasks each using different modalities of data. The first involves images, and the second involves tabular data. This lab serves as both an introduction to the course, and as a stealth Python revision!

A. Making a wall of faces

The `datasets` folder in this repo contains a folder called `faces`. This contains two folders called `smile` and `non-smile`. Each contains JPGs of various famous people smiling and not smiling respectively. These were obtained from Kaggle.

A1. Read in and display Matthew McConaughey's face (`datasets/faces/Matthew_McConaughey_0001.jpg`). How big is this image? Then, convert it into a Numpy array and print its size.

In [2]:

```
# Read image
image = Image.open("datasets/faces/smile/Matthew_McConaughey_0001.jpg")

# Use Juypter's inbuilt display function
display(image)
print(f"The image has a size of {image.size}")

im = np.array(image)
print(im.shape)
```

```
The image has a size of (64, 64)
(64, 64, 3)
```

Assessment: Tests (50%)

- Two tests, two hours each, taken live and in-person during labs
- Each test consists of short-answer questions and some coding exercises within a Jupyter notebook
    - **Test 1** is taken during the **Week 4** lab. It covers Week 1-3 material and is worth 20% of the total course mark
    - **Test 2** is taken during the **Week 11** lab. It and covers Week 5-10 material and is worth 30% of the total course mark
- They are **closed book** but you may bring in a piece of A4 paper with handwritten notes on both sides. You can also use the **help()** function in Jupyter to see documentation
- Practice tests will be made available to help you prepare

Assessment: Coursework 1 (20%)

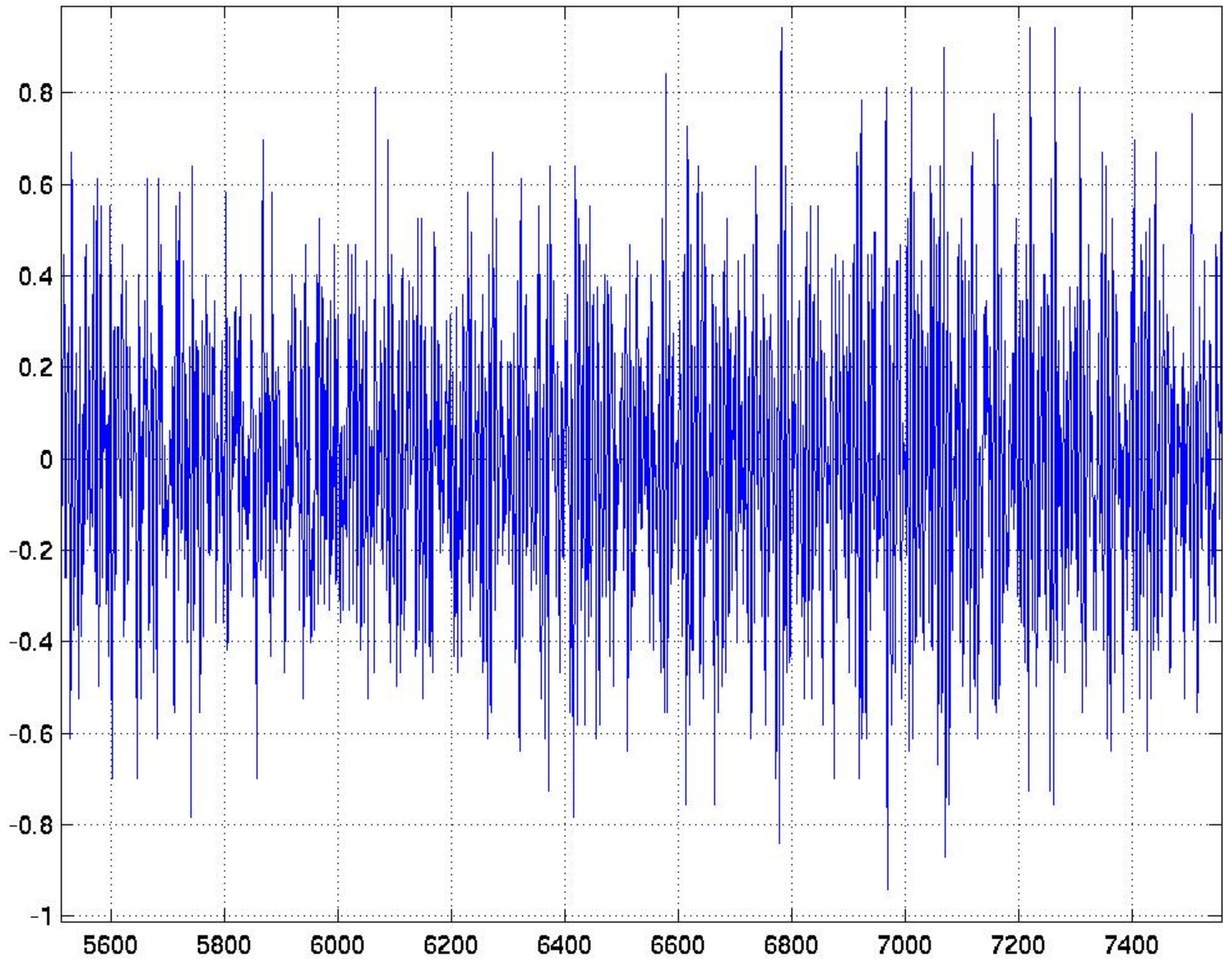
- This will be released in the Week 4 lecture on **Monday 5th February**
- You will create slides and record a short presentation using them
- This will be a case study on a real-world machine learning application
- You will critique this application from an ethical standpoint
- The deadline is **Tuesday 20th February @ 1600** (Flexible learning week)

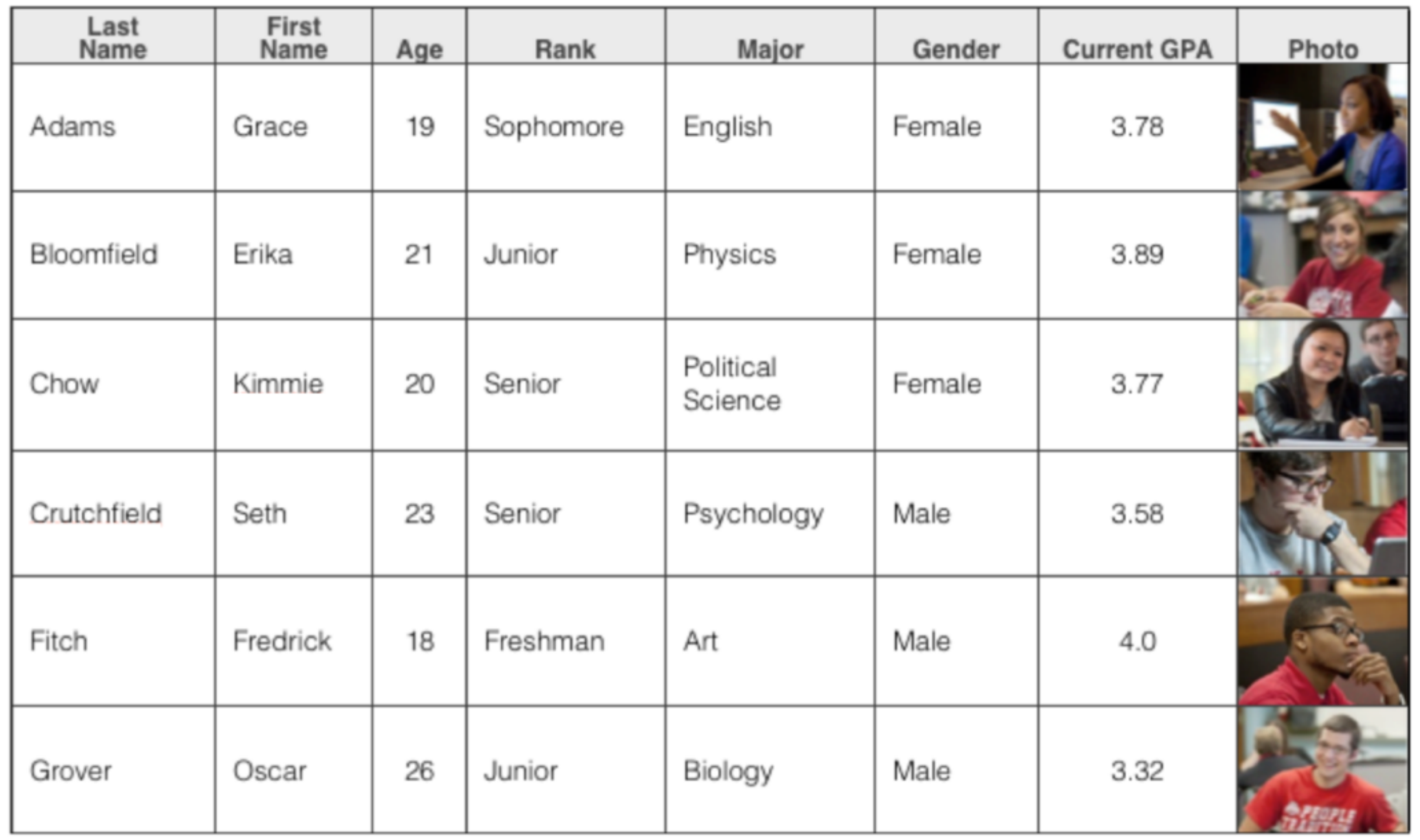
Assessment: Coursework 2 (30%)

- This will be released in the Week 8 lecture on **Monday 11th March**
- You will be given a dataset
- You will perform exploratory data analysis and apply machine learning to this dataset
- You will produce a short report on your findings supplemented with code
- The deadline is **Tuesday 26th March @ 1600** (Week 10)

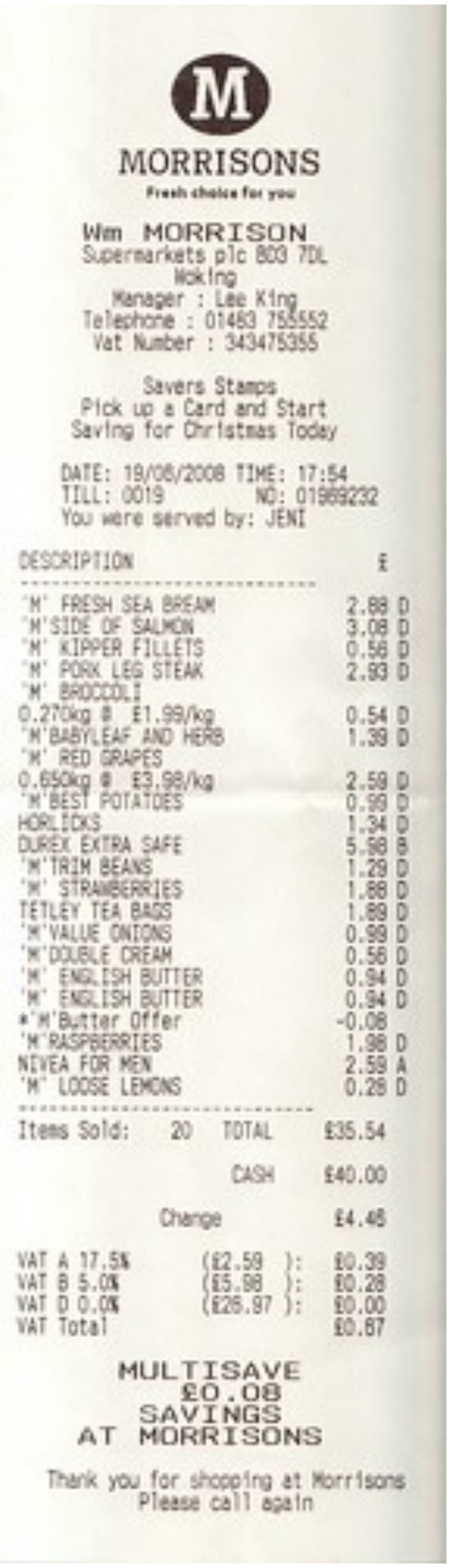
Data Modalities

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 BD3 7DL
Woking
Manager : Lee King
Telephone : 01483 755552
Vat Number : 343475355

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

DATE: 19/06/2008 TIME: 17:54
TILL: 0019 NO: 01989232
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 @ £3.98/kg	2.59 D
'M'BEST POTATOES	0.99 D
HORLICKS	1.34 D
DUREX EXTRA SAFE	5.98 B
'M'TRIM BEANS	1.29 D
'M' STRAWBERRIES	1.88 D
TETLEY TEA BAGS	1.89 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	2.59 A
'M' LOOSE LEMONS	0.28 D

Items Sold: 20 TOTAL £35.54

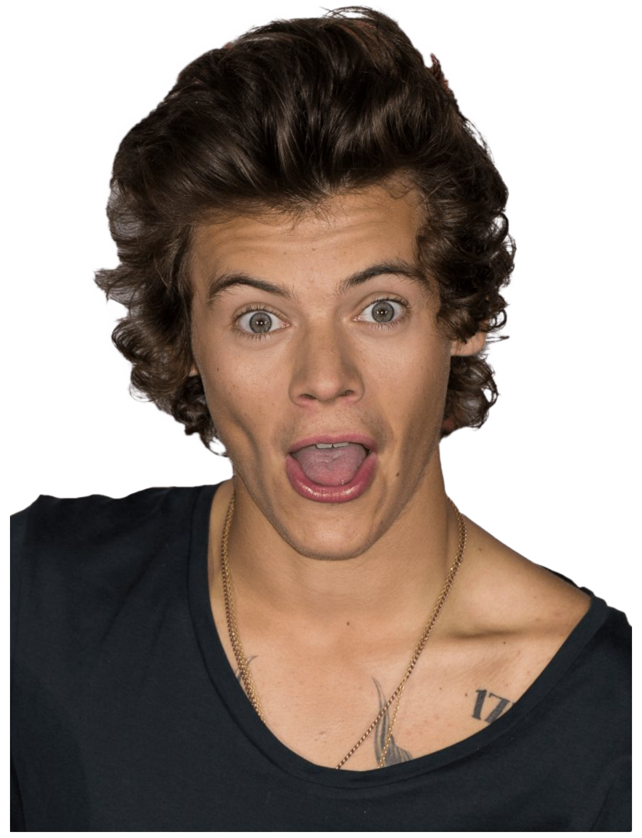
CASH £40.00

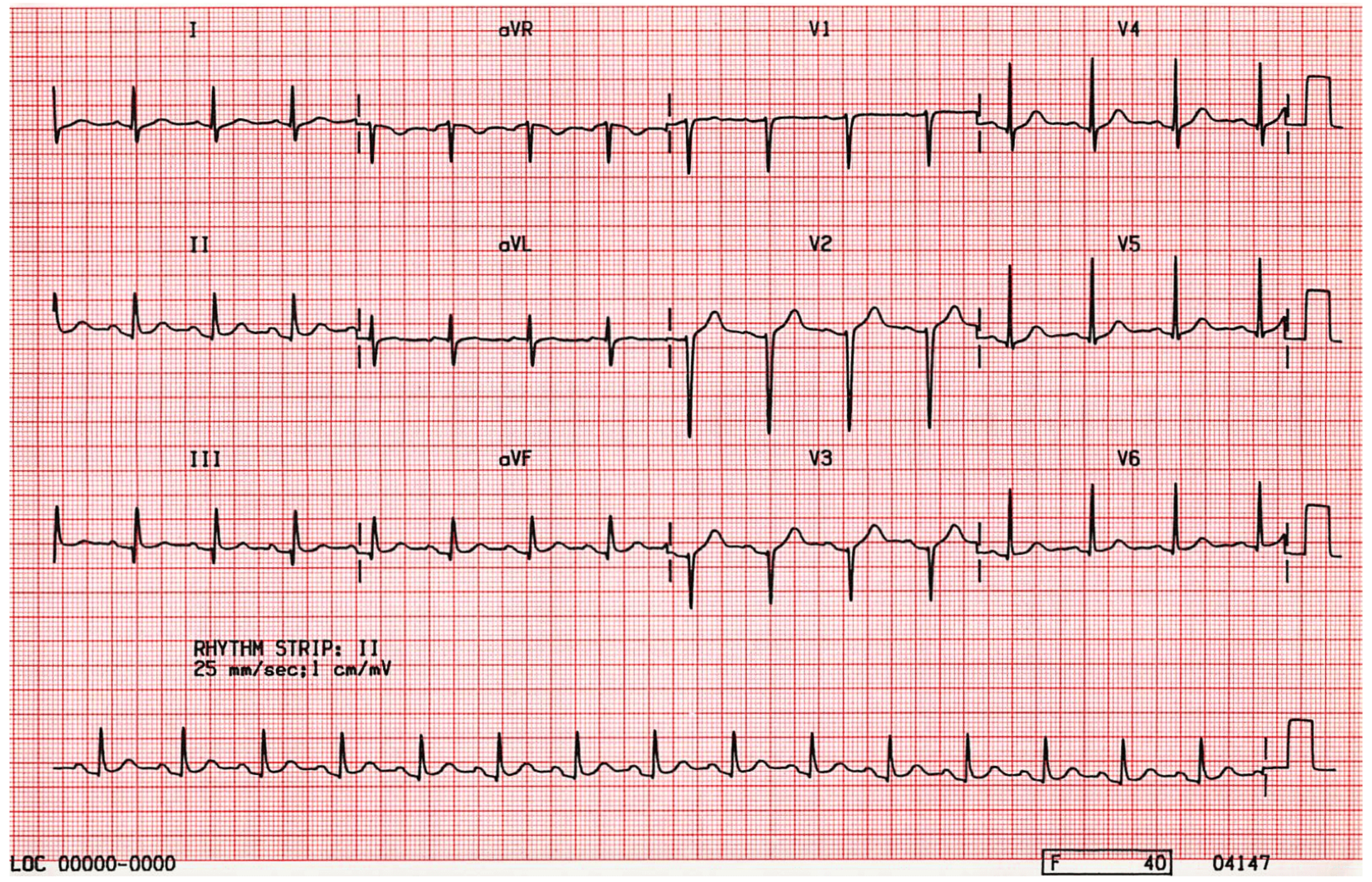
Change £4.46

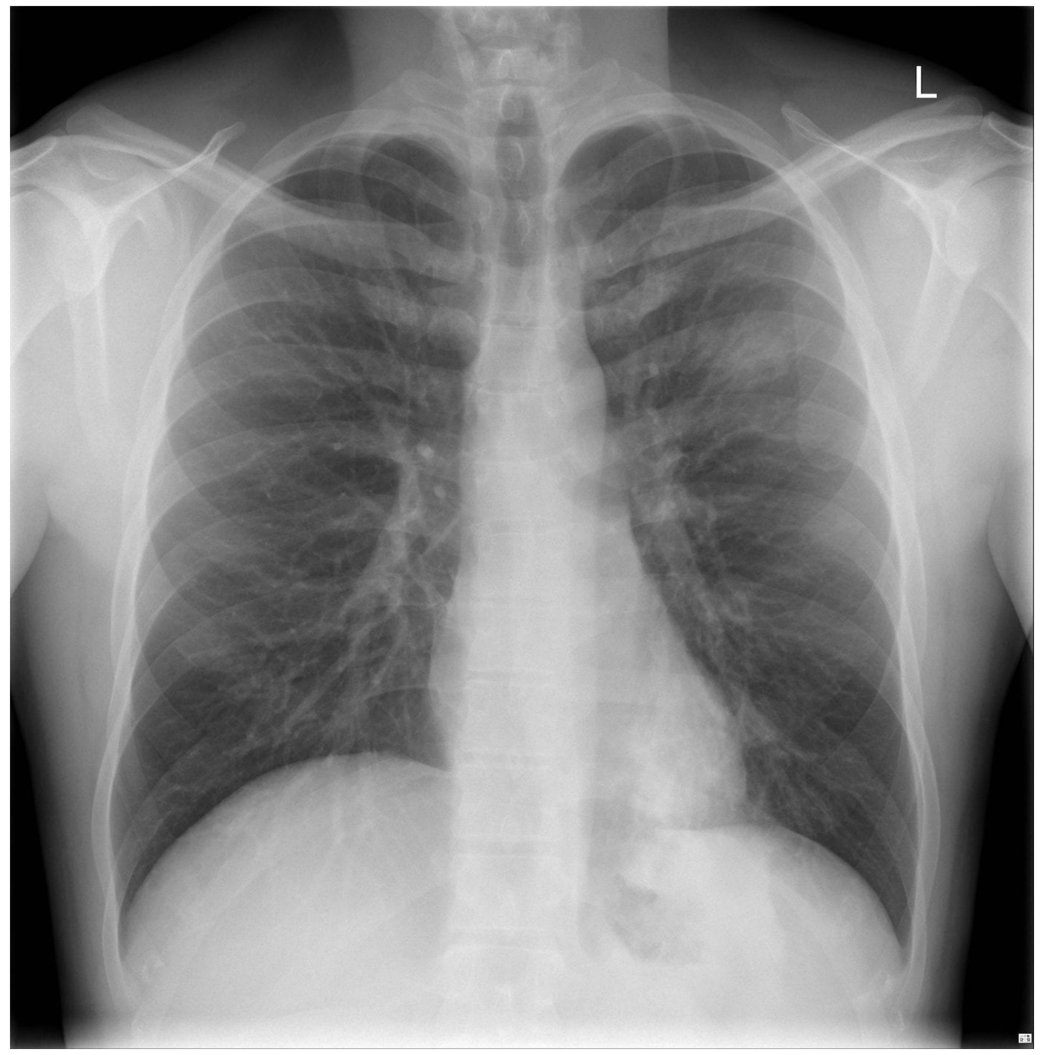
VAT A 17.5% (£2.59): £0.39
VAT B 5.0% (£5.98): £0.28
VAT D 0.0% (£26.97): £0.00
VAT Total £0.67

MULTISAVE
£0.08
SAVINGS
AT MORRISONS

Thank you for shopping at Morrisons
Please call again

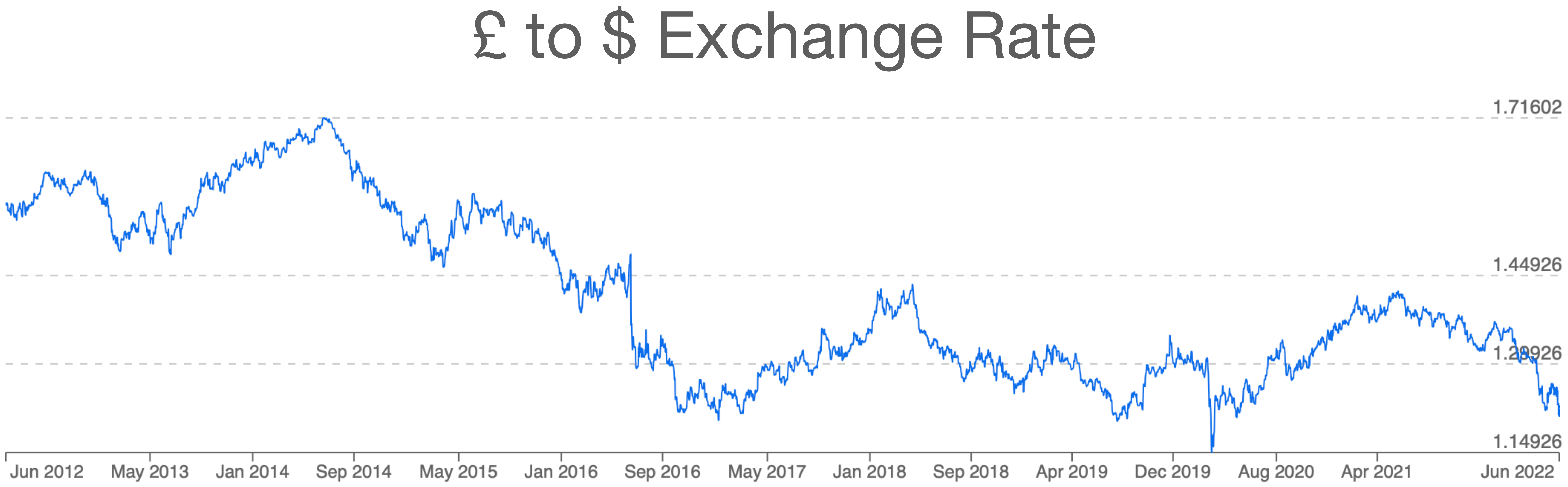






Time series data

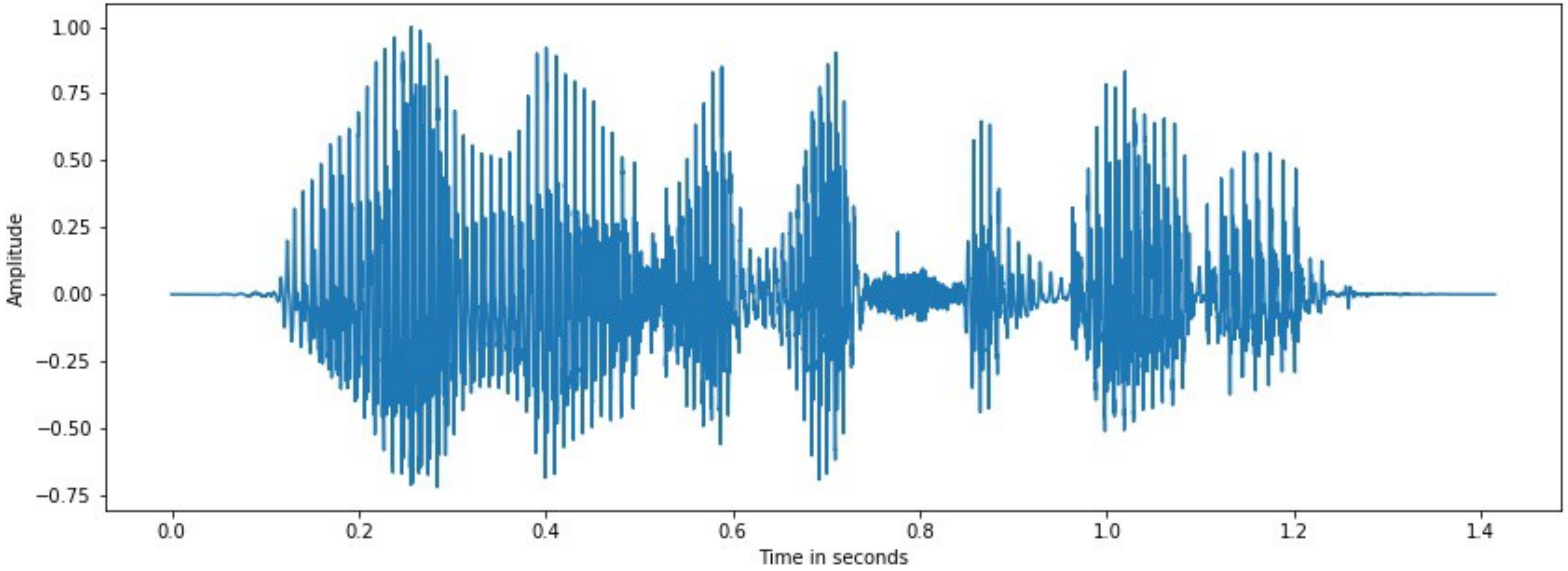
- y axis is some quantity we care about
- x axis is time



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

Time series data

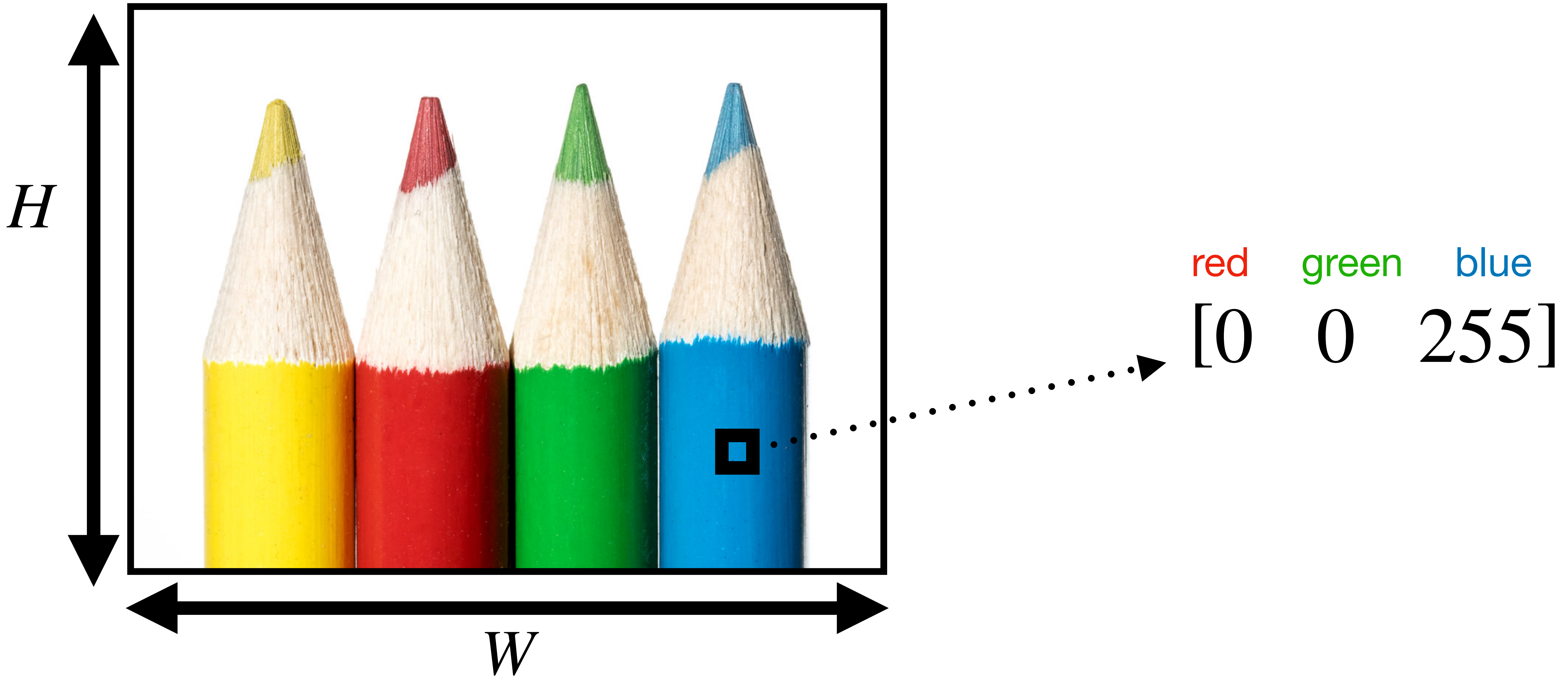
- For example, speech!



Source: https://towardsdatascience.com/beginners-guide-to-speech-analysis-4690ca7a7c05

Image data

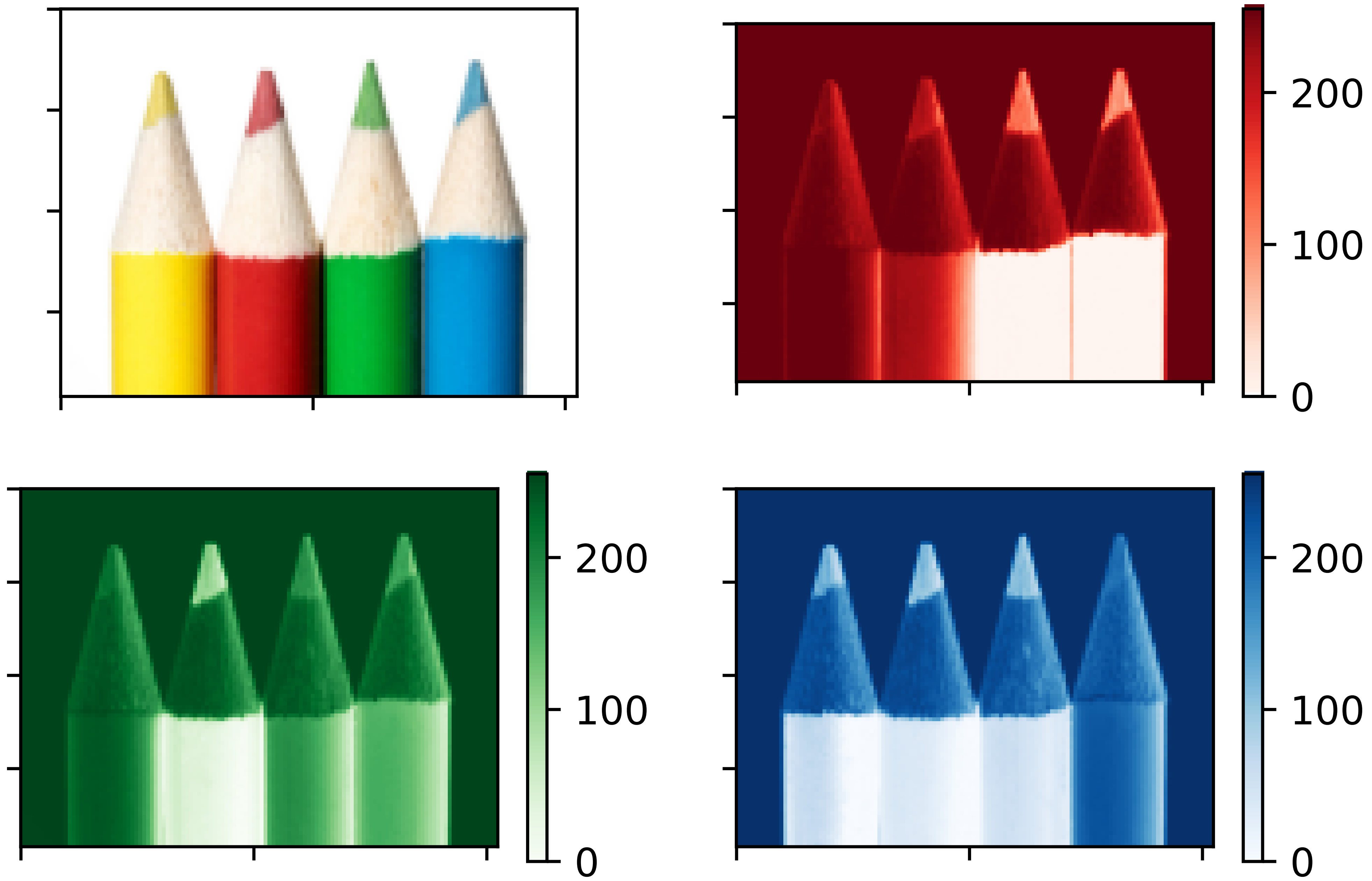
- An image is a rectangular array of $H \times W$ pixels
- Each pixel consists of three numbers: the amount of **red**, **green**, **and** **blue**

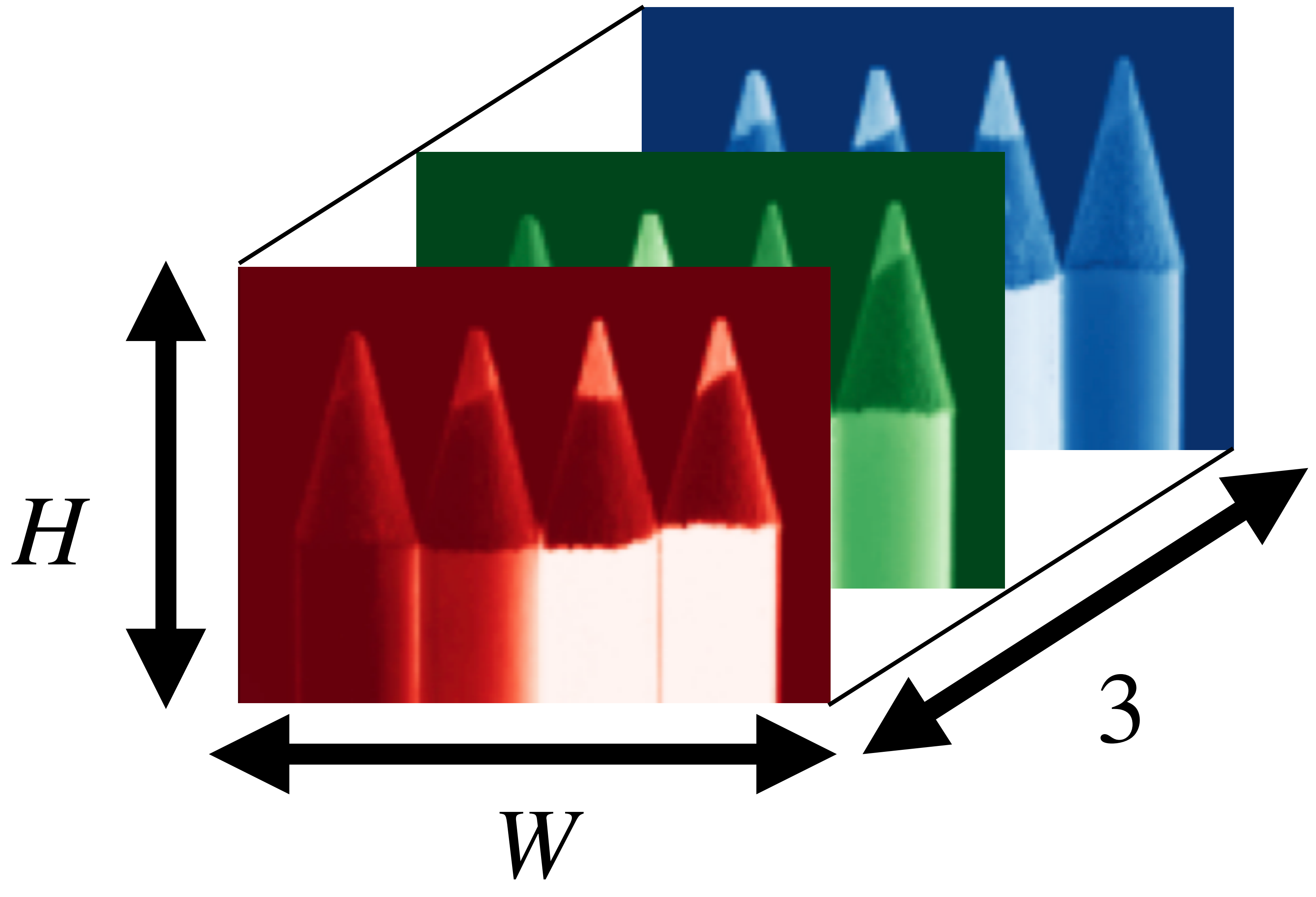


$$\begin{array}{ccc} \text{red} & \text{green} & \text{blue} \\ [0 & 0 & 255] \end{array}$$

Image data

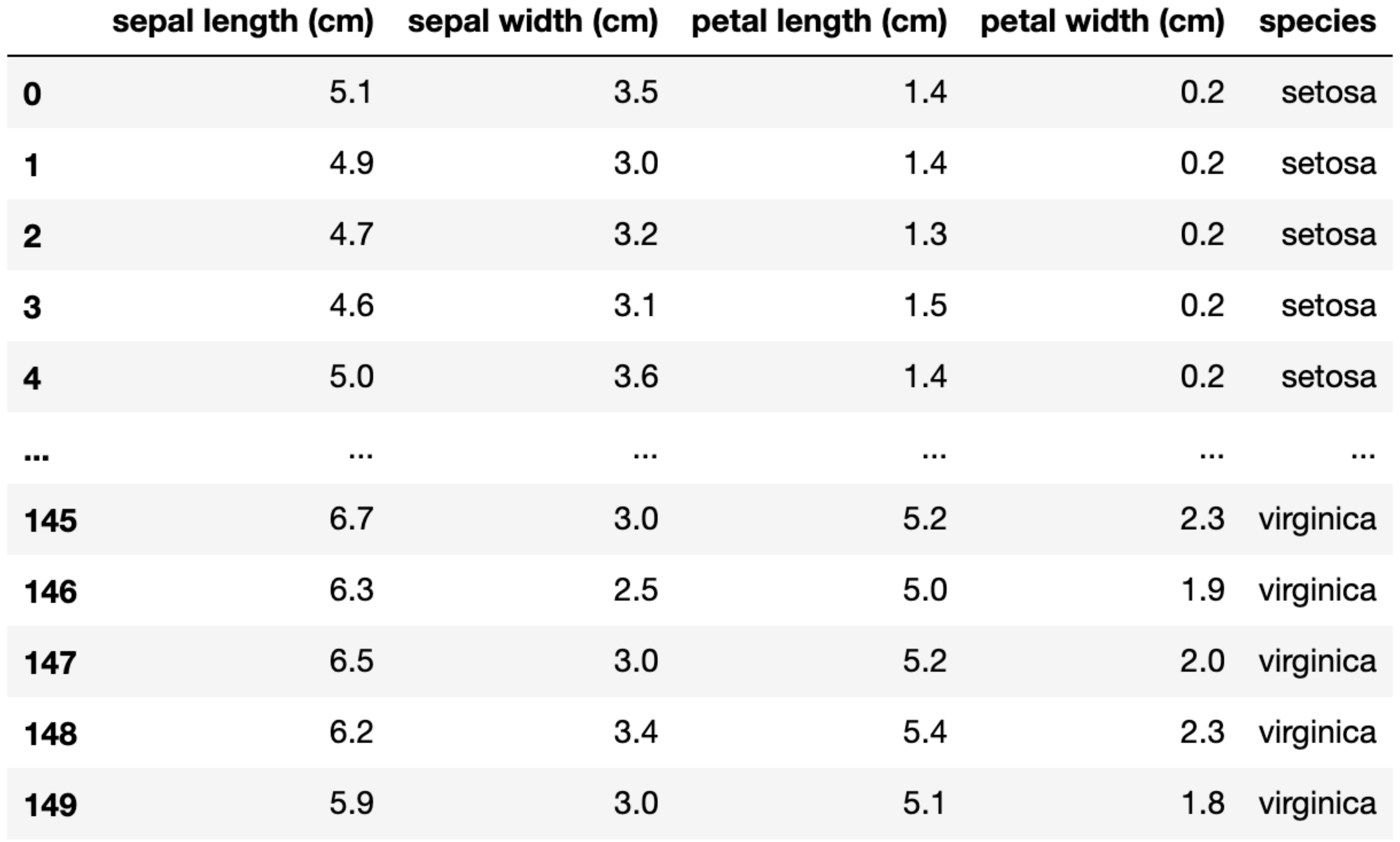
- This gives us a **red**, **green**, **and** **blue** 2D array
- These are stacked along the z axis to form a 3D array



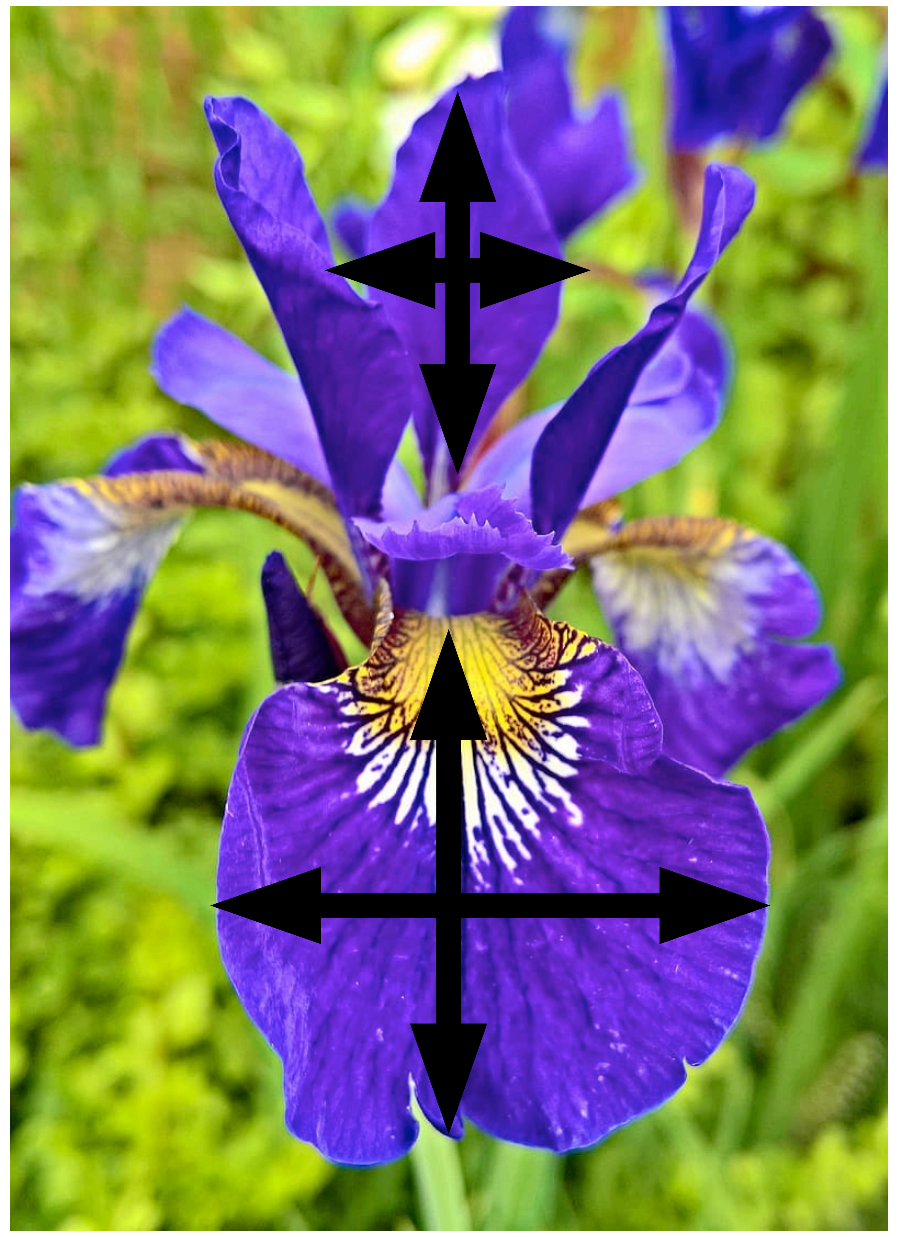


Tabular data

- 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

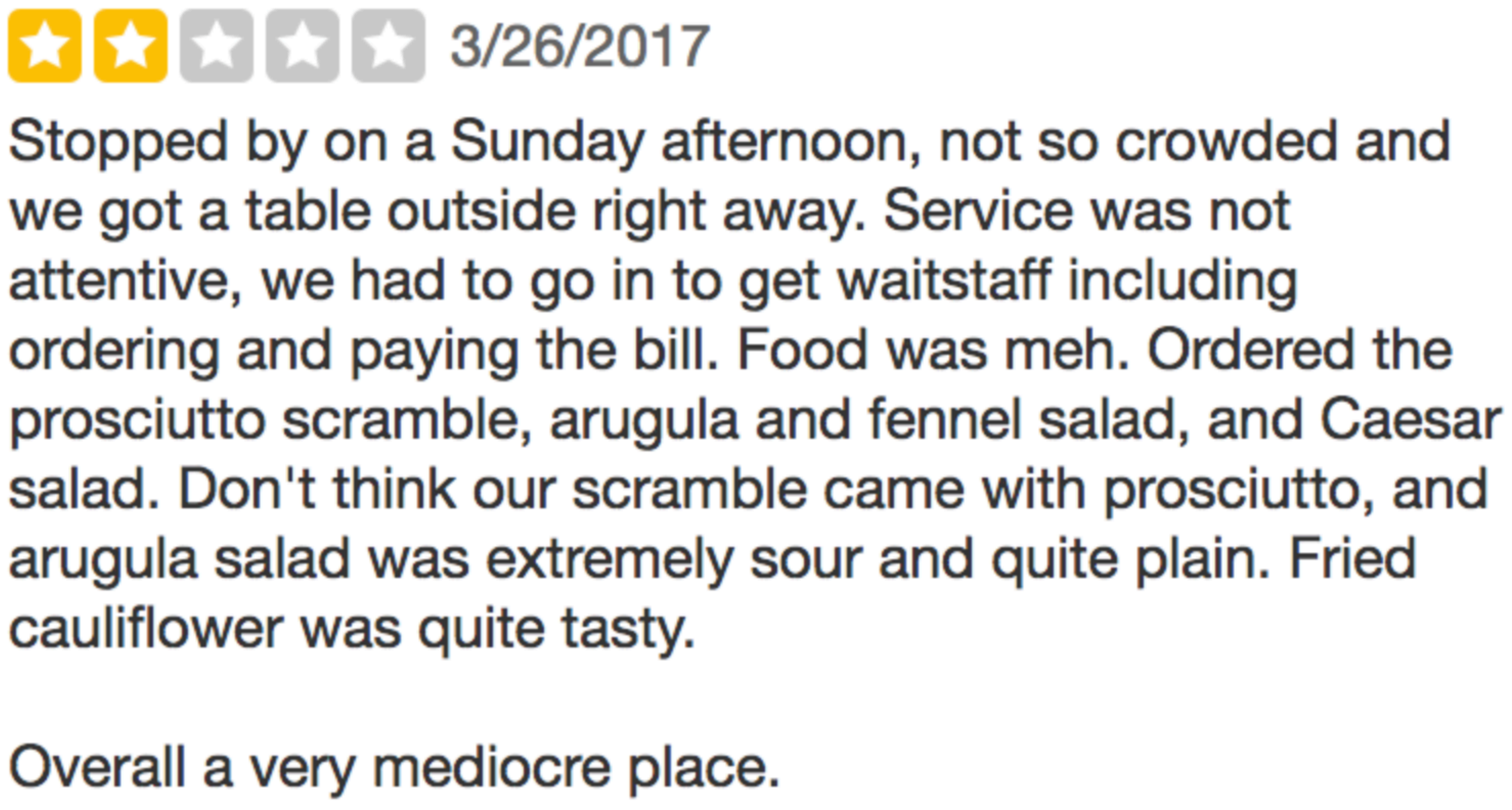


	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2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

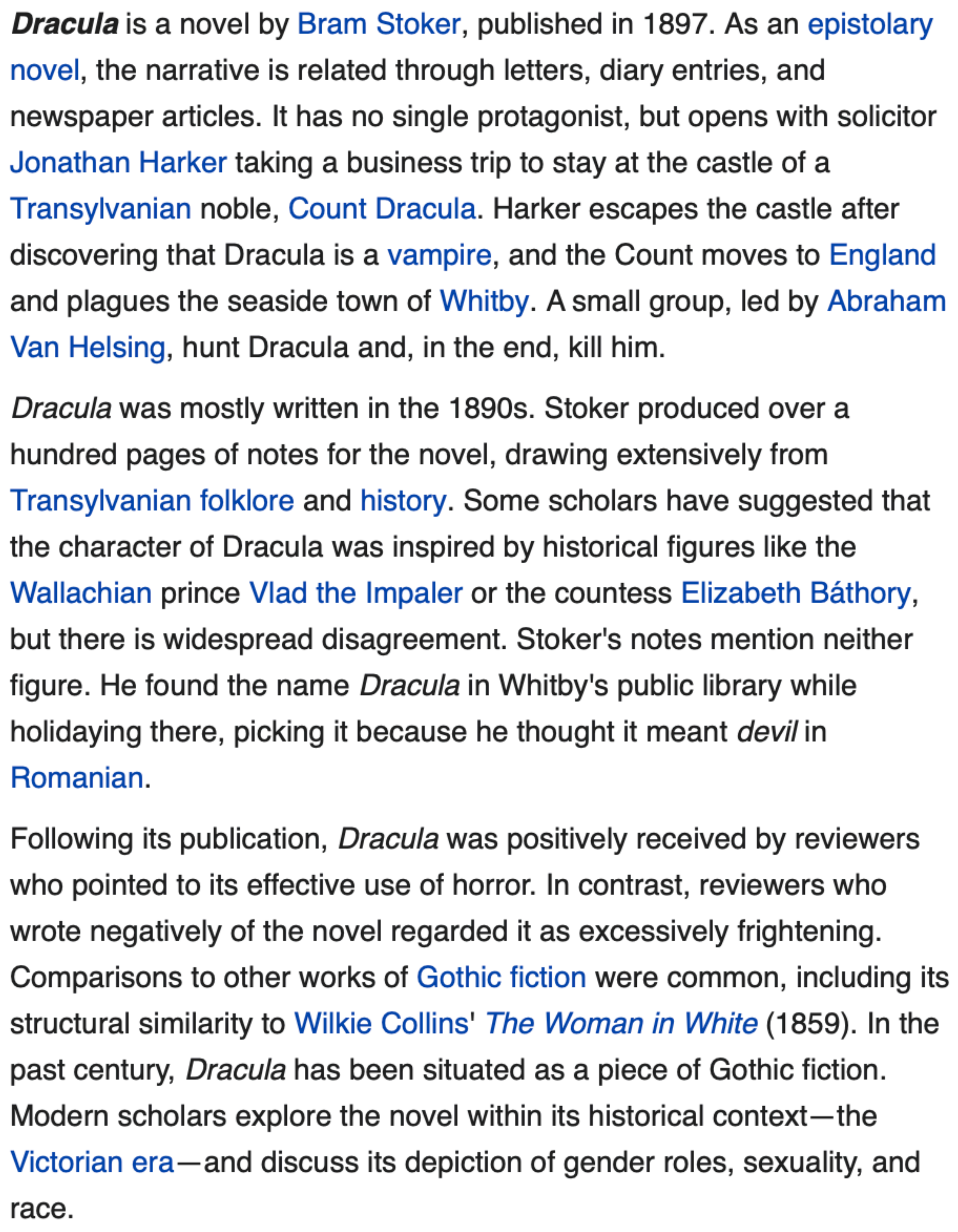
- Largely unstructured and usually text
- Can (sometimes!) be hacked into e.g. tabular data



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.

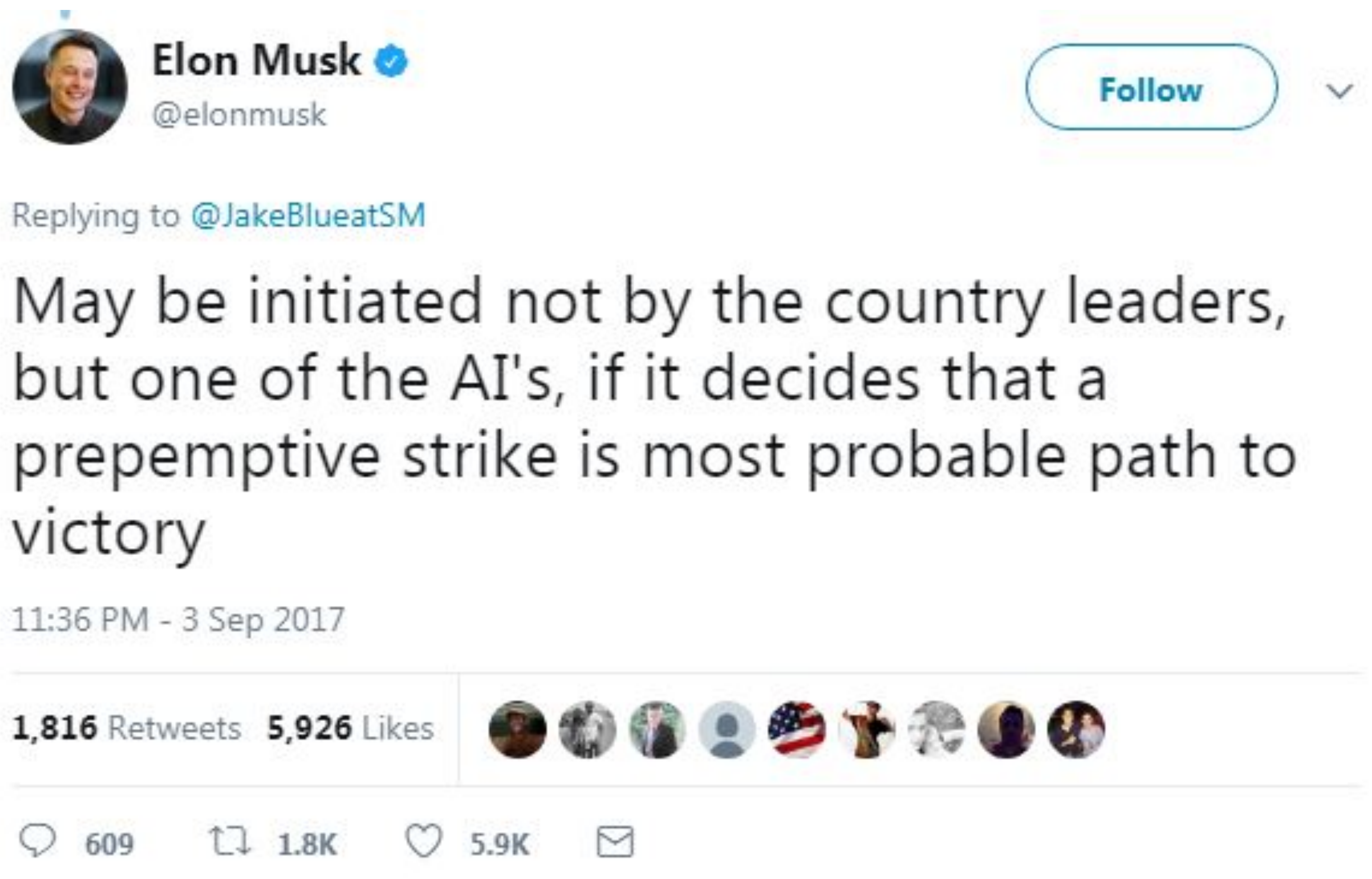
Overall a very mediocre place.



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

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

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**

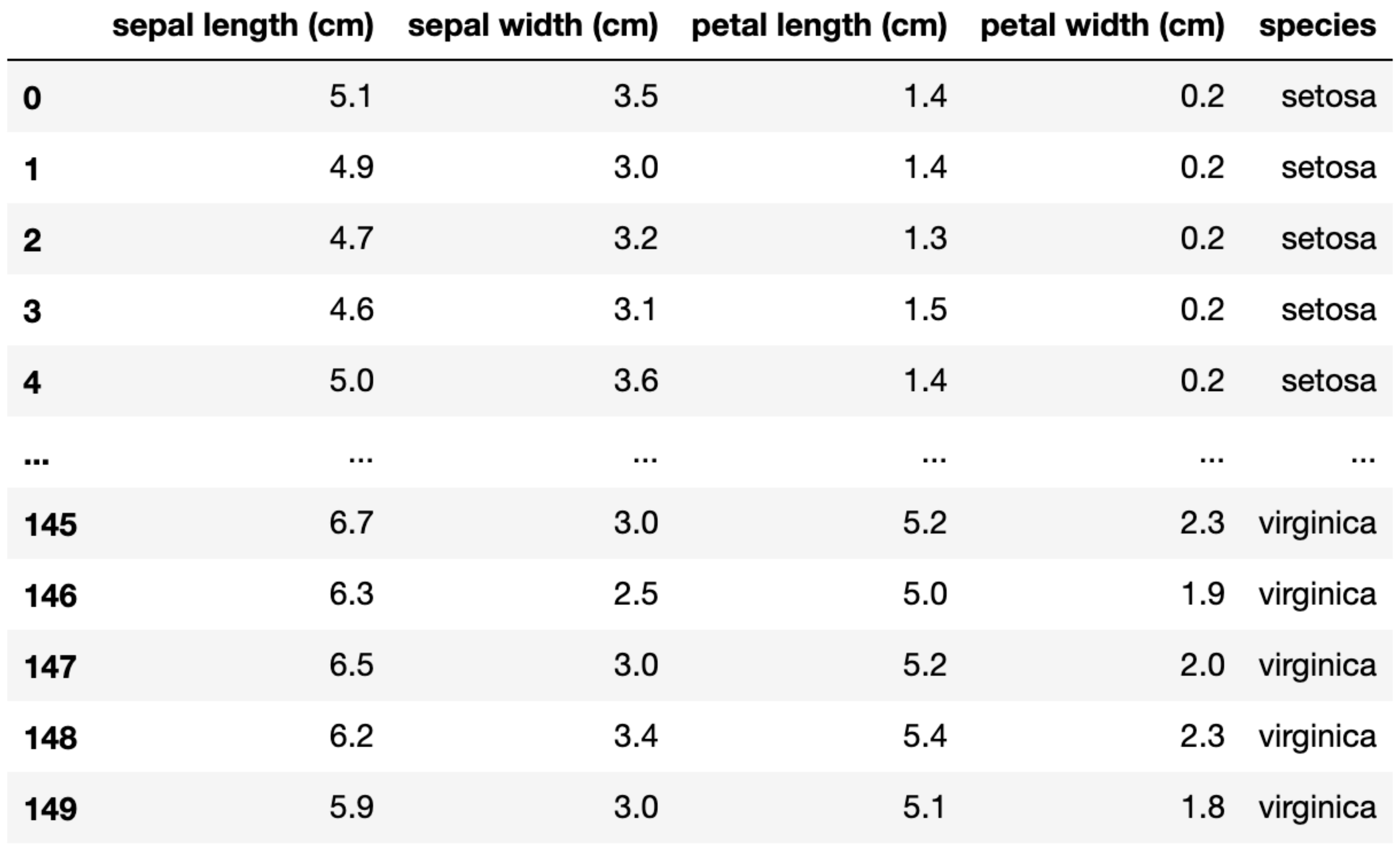
An **image** is a set of **pixels**

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

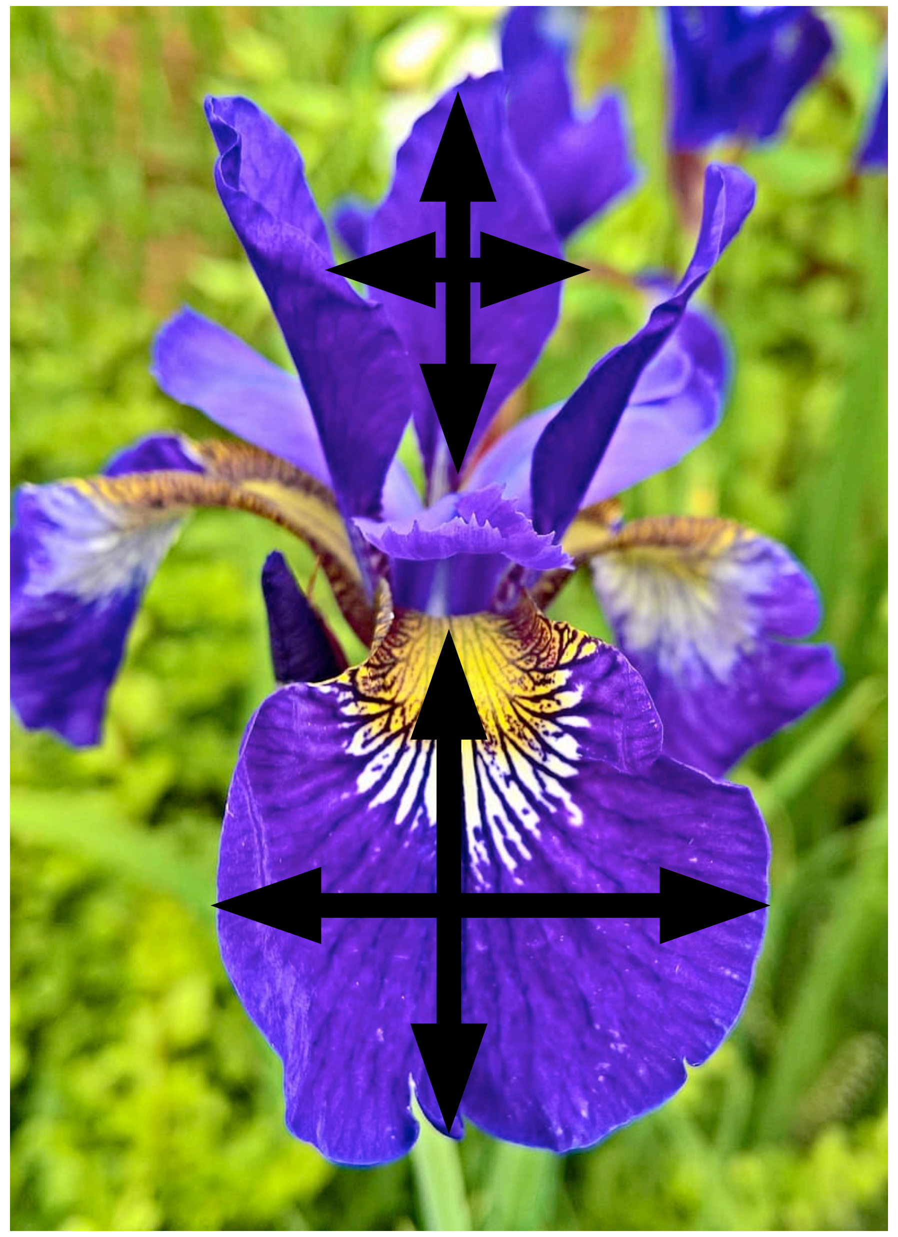
Variable Types

Tabular data (again!)

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

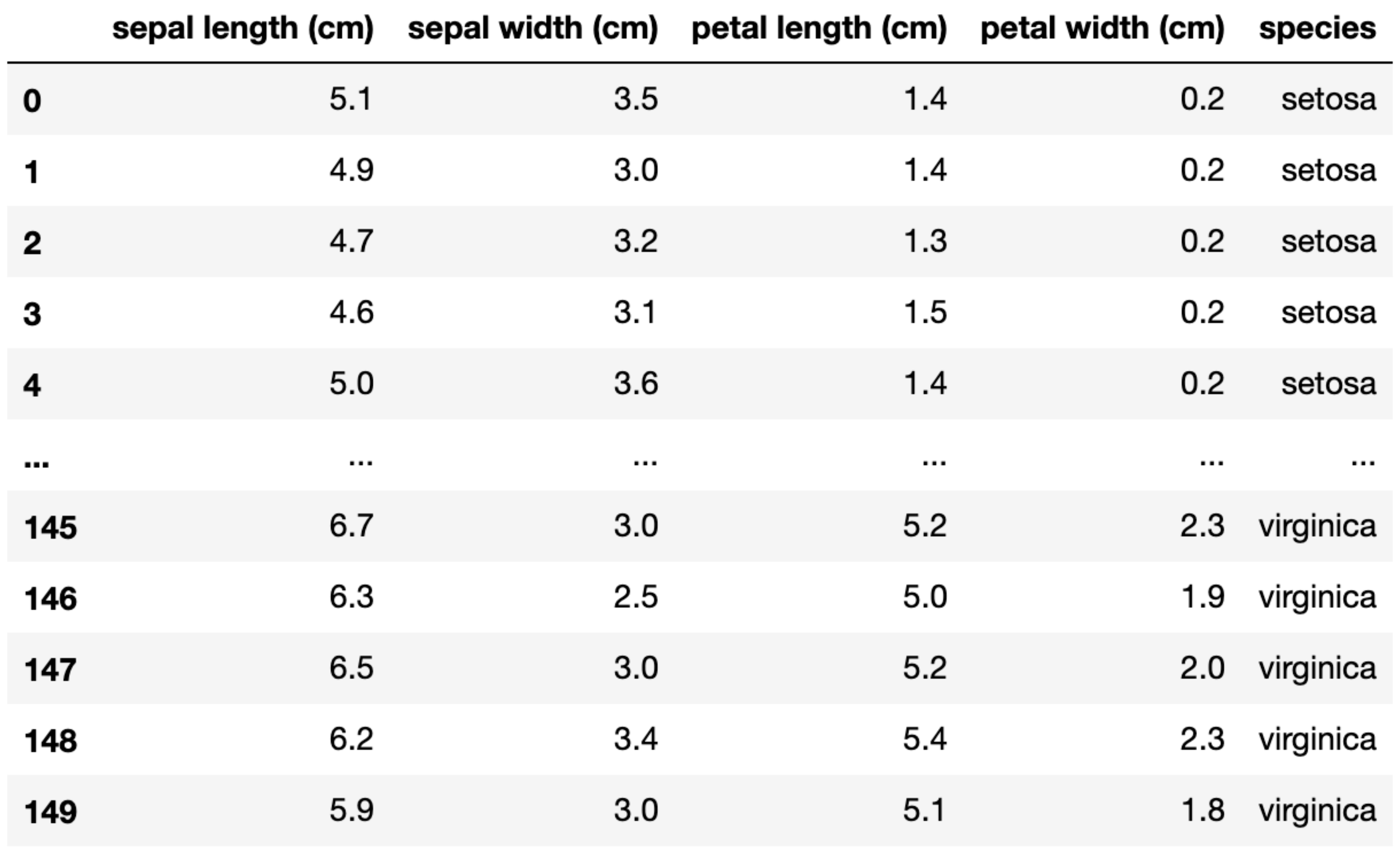


	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2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

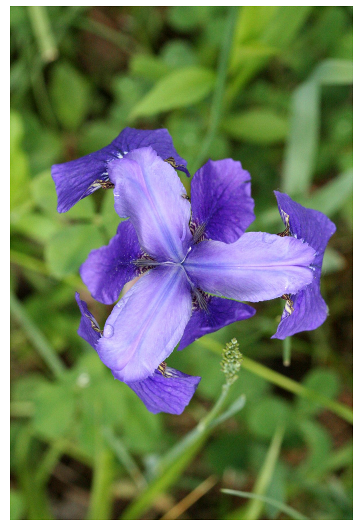
Measurements of the variable correspond to descriptive categories

- 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

level of education (ordinal)

1

primary

2

secondary

3

university

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

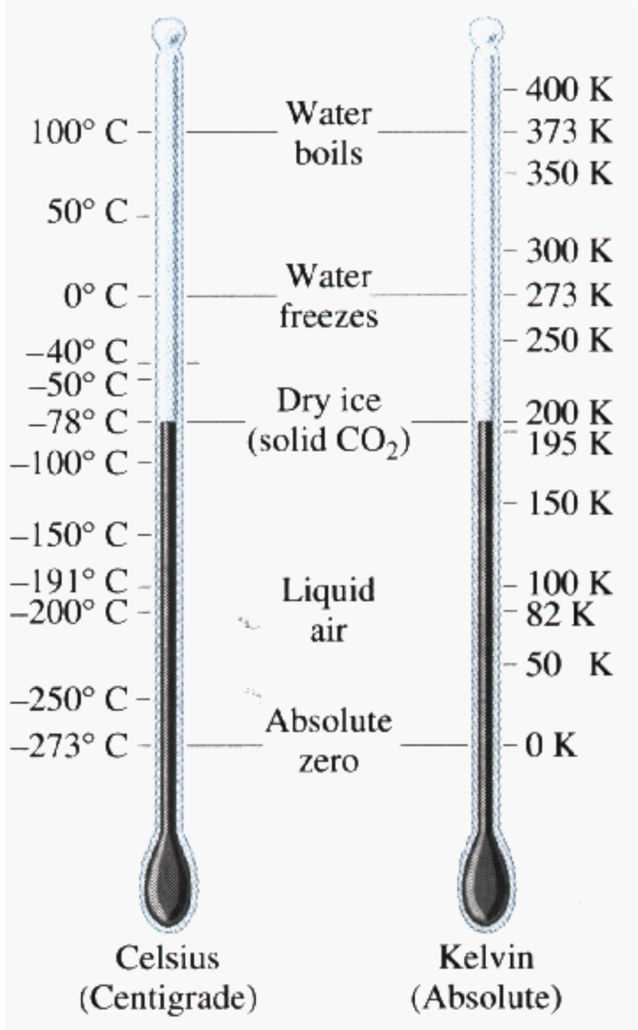
Continuous variables

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 Celcius is interval

Temperature in Kelvin is ratio



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

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