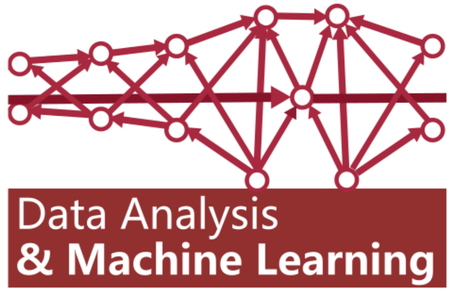


Data Analysis and Machine Learning 4

Week 1: Introduction, data modalities, variable types

Elliot J. Crowley, 16th January 2023

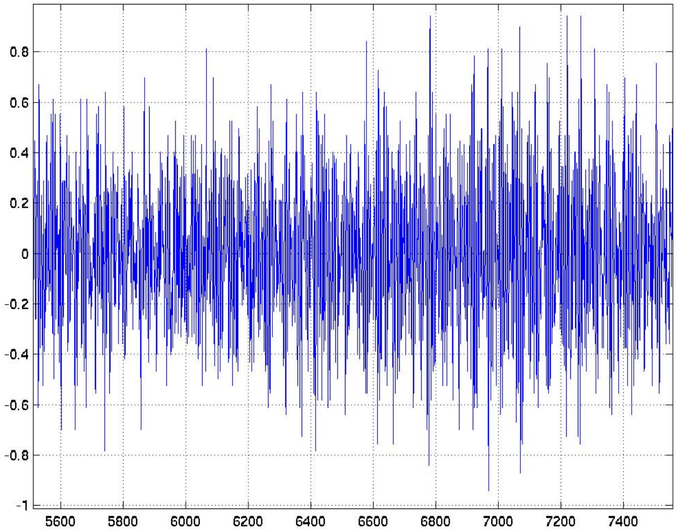
THE UNIVERSITY *of* EDINBURGH

What is data?

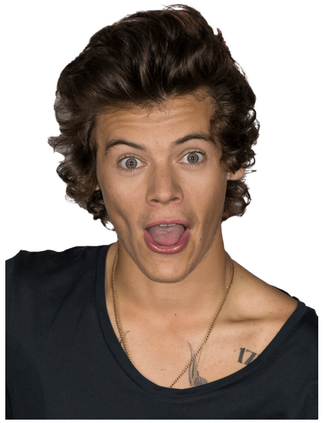
“information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer”

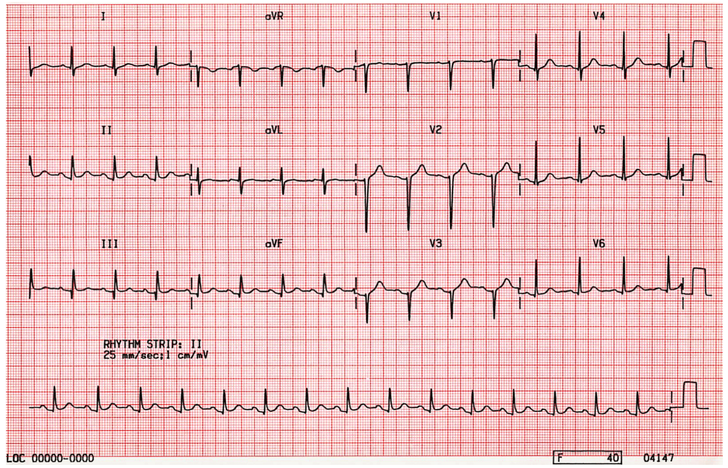
Cambridge Dictionary

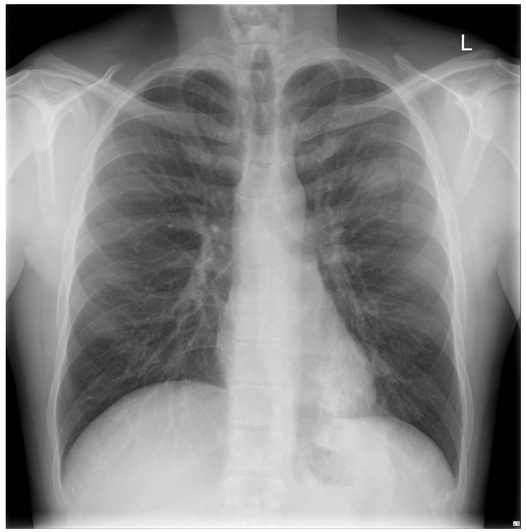
Data



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	







Data Analysis

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

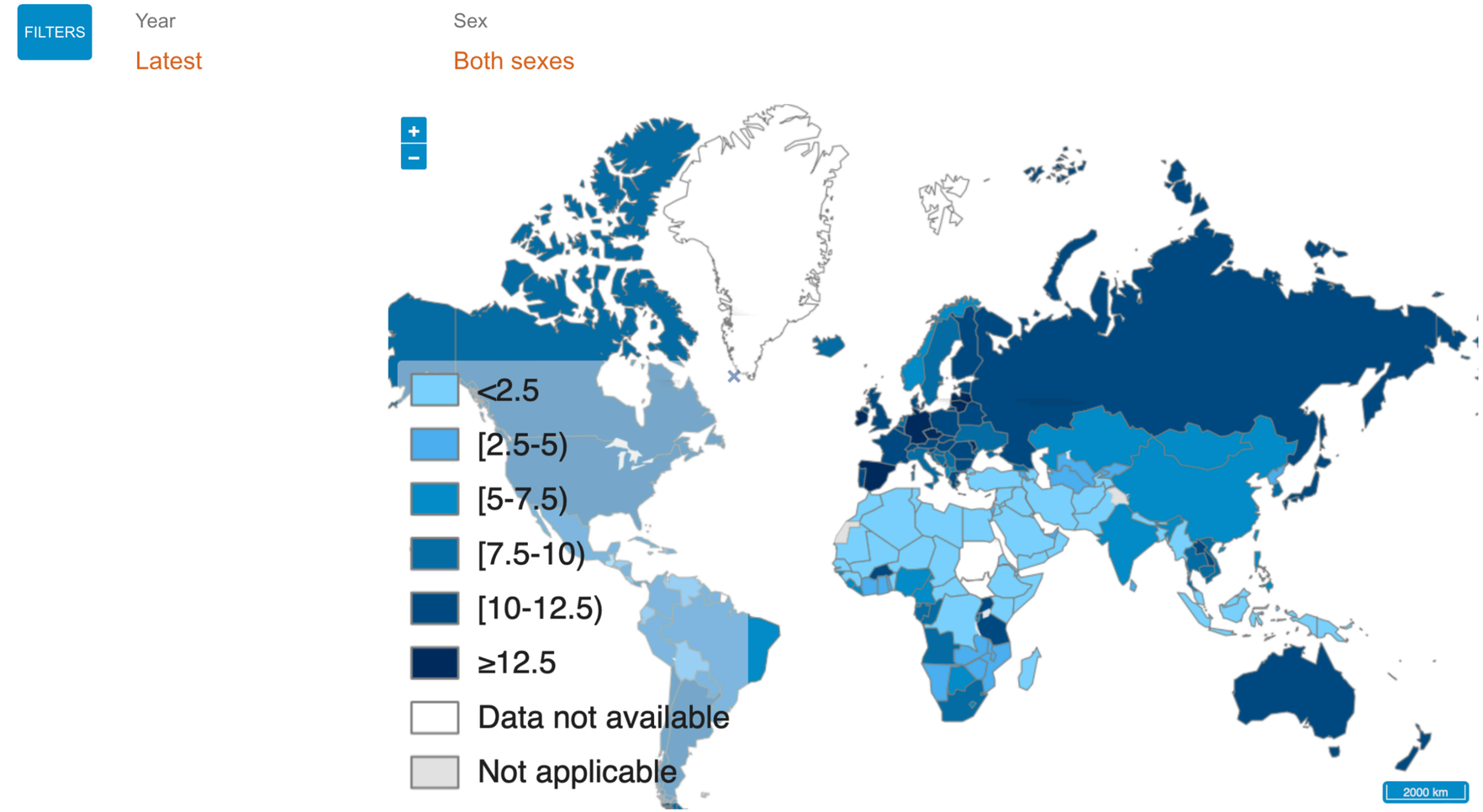
Cambridge Dictionary

“The natural application of data; you use it to do cool stuff!”

Elliot J. Crowley

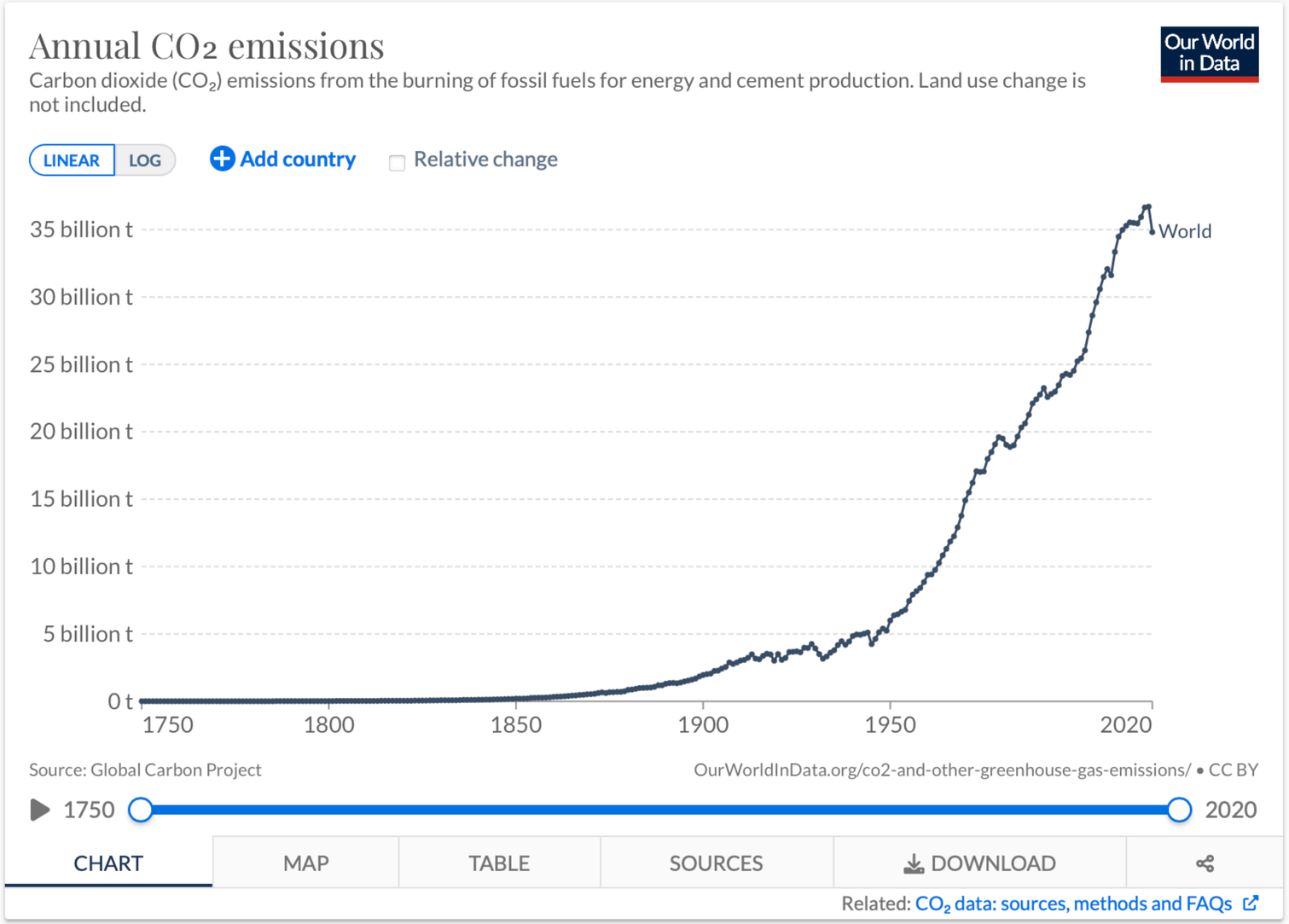
Spotting patterns

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



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

Observing trends



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

Telling (happy/sad?) stories

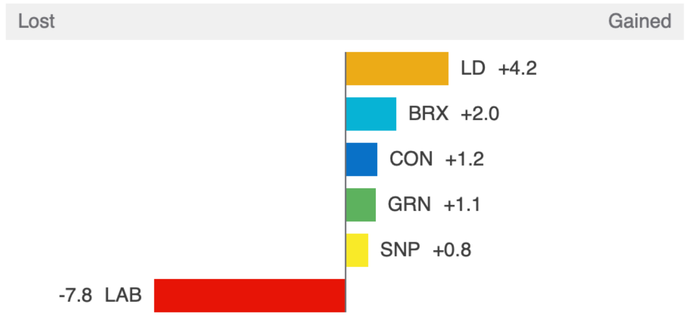
UK vote share

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

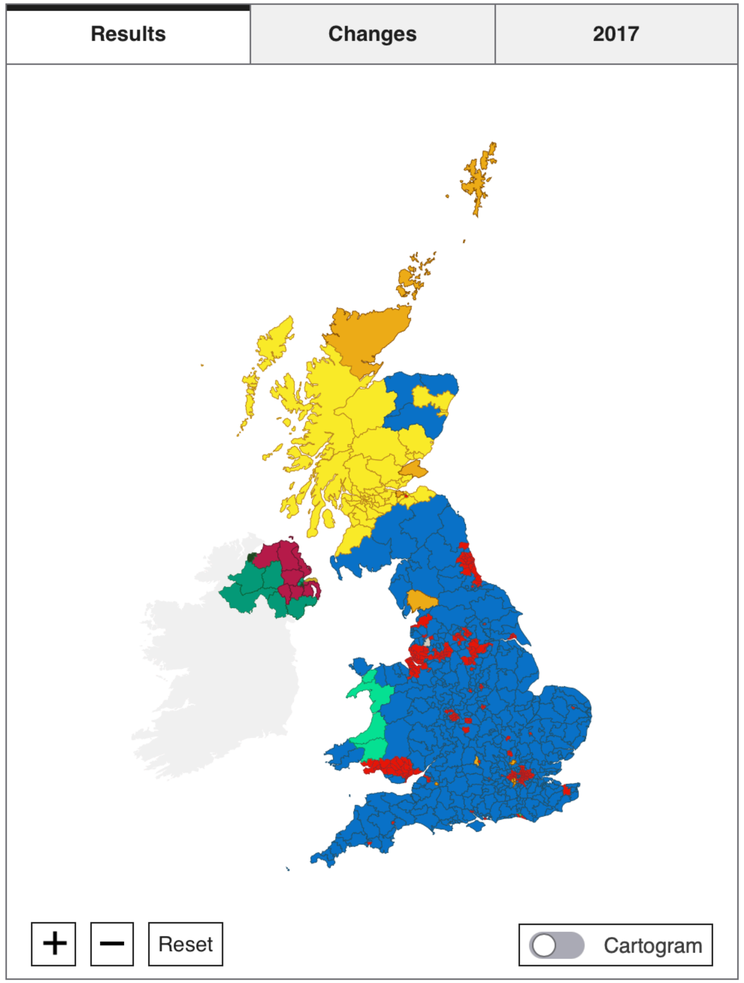


Turnout

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

Coronavirus pandemic

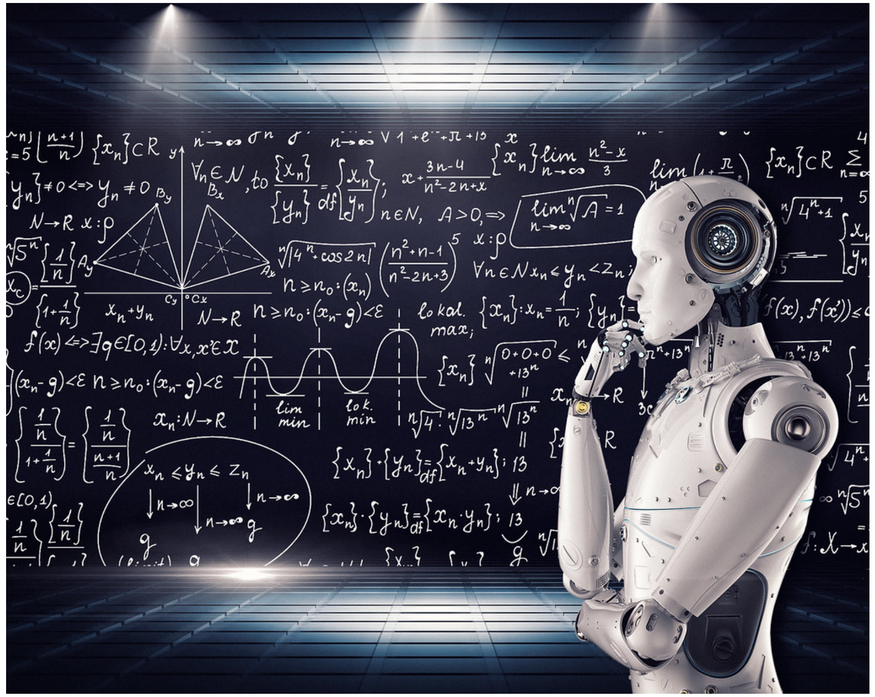
Liam Thorp was wrongly classed as morbidly obese according to his height and weight

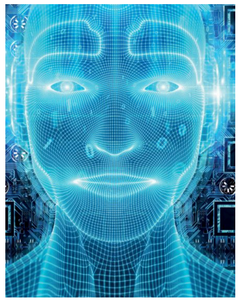
A man in his 30s with no underlying health conditions was offered a Covid vaccine after an NHS error mistakenly listed him as just 6.2cm in height.

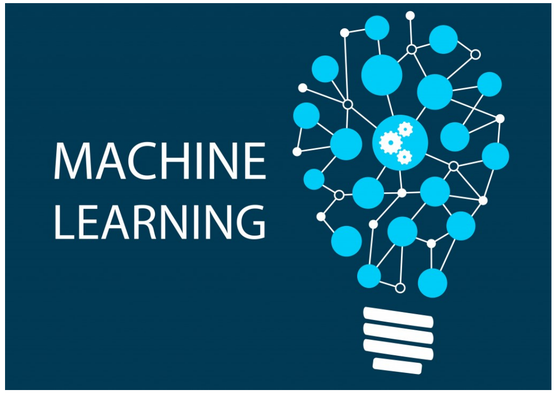
Source: https://www.bbc.co.uk/news/uk-england-merseyside-56111209

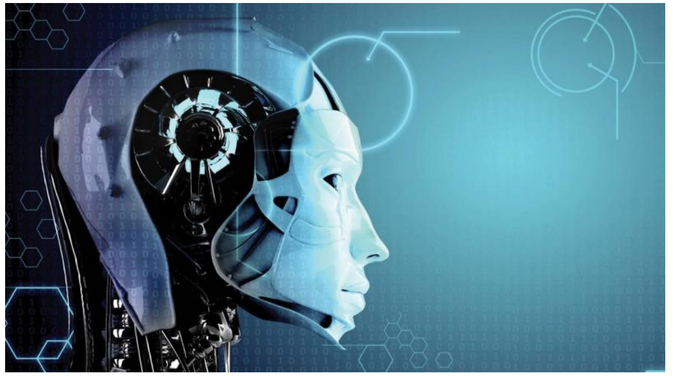
What is Machine Learning?

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









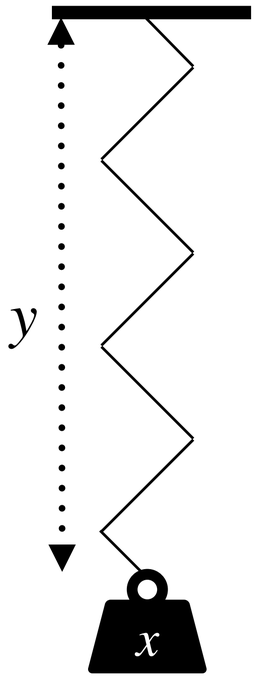
Machine Learning is…

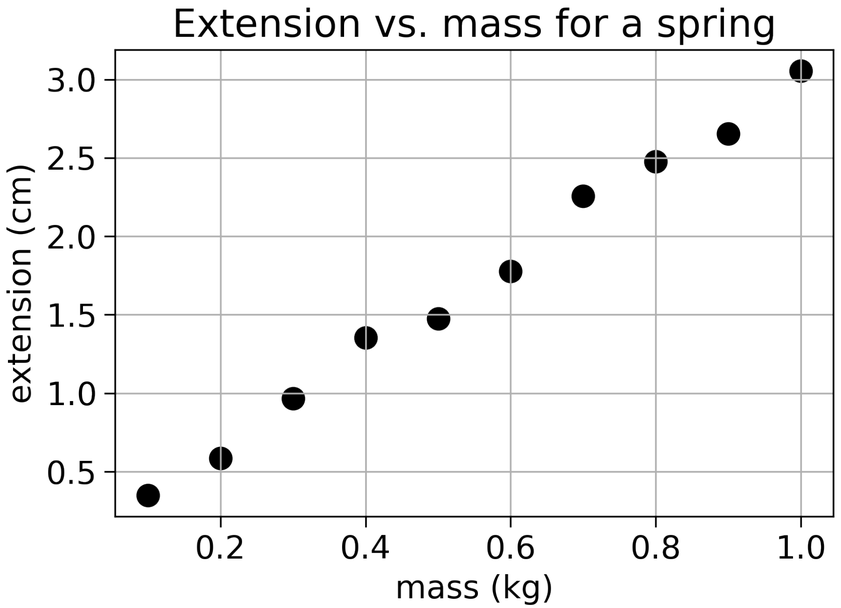
“the study of models that can learn from training data in order to make predictions on new data.”

Elliot J. Crowley

Machine Learning for a spring

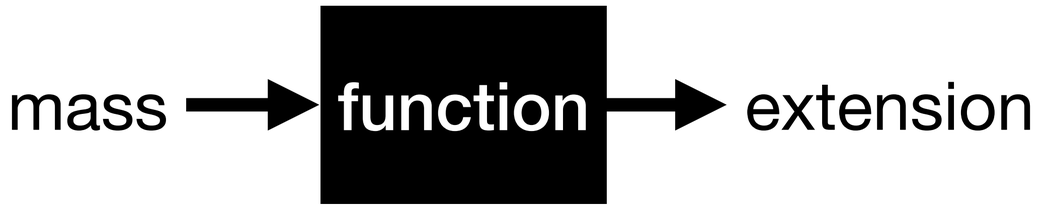
- We want a model that given an arbitrary mass x can predict extension y
- We can attach some masses to the spring and record its extension
- These mass-spring measurements form our **training data**

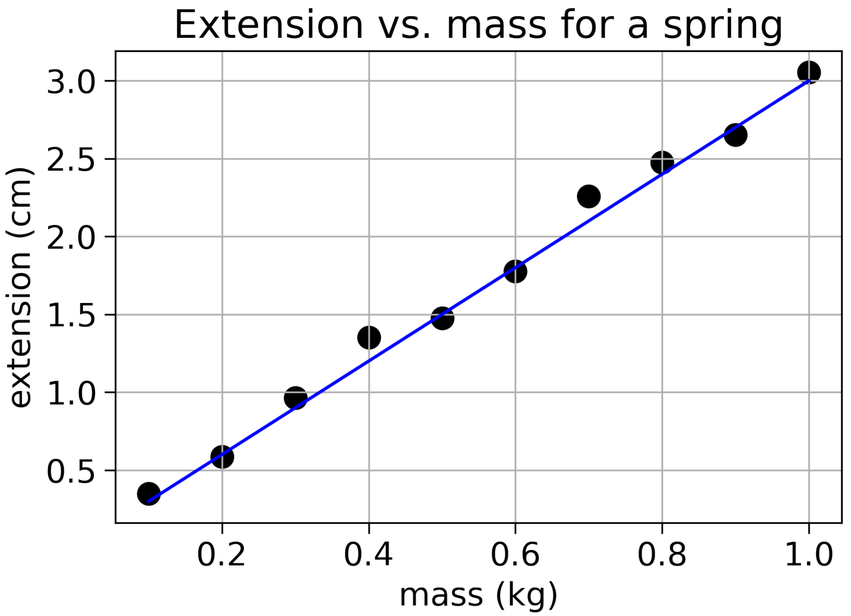




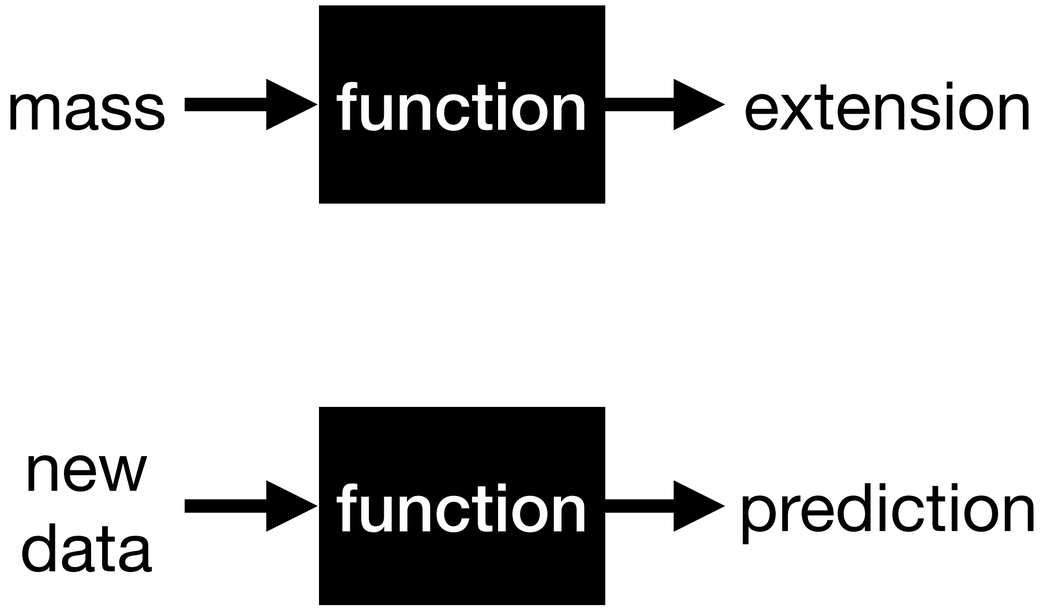
Machine Learning for a spring

- 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

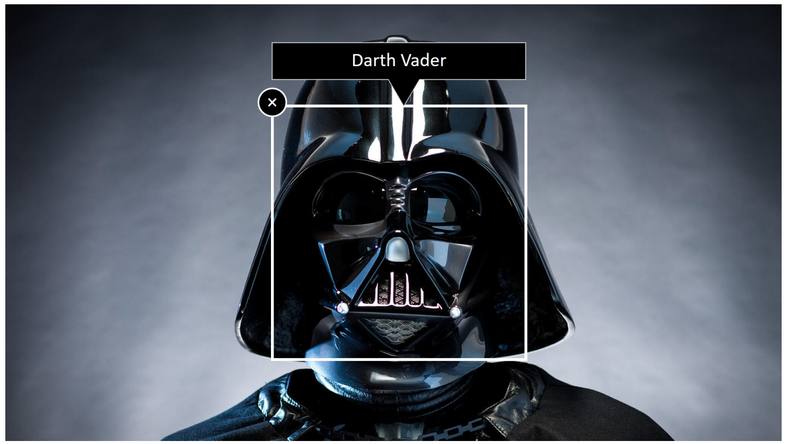


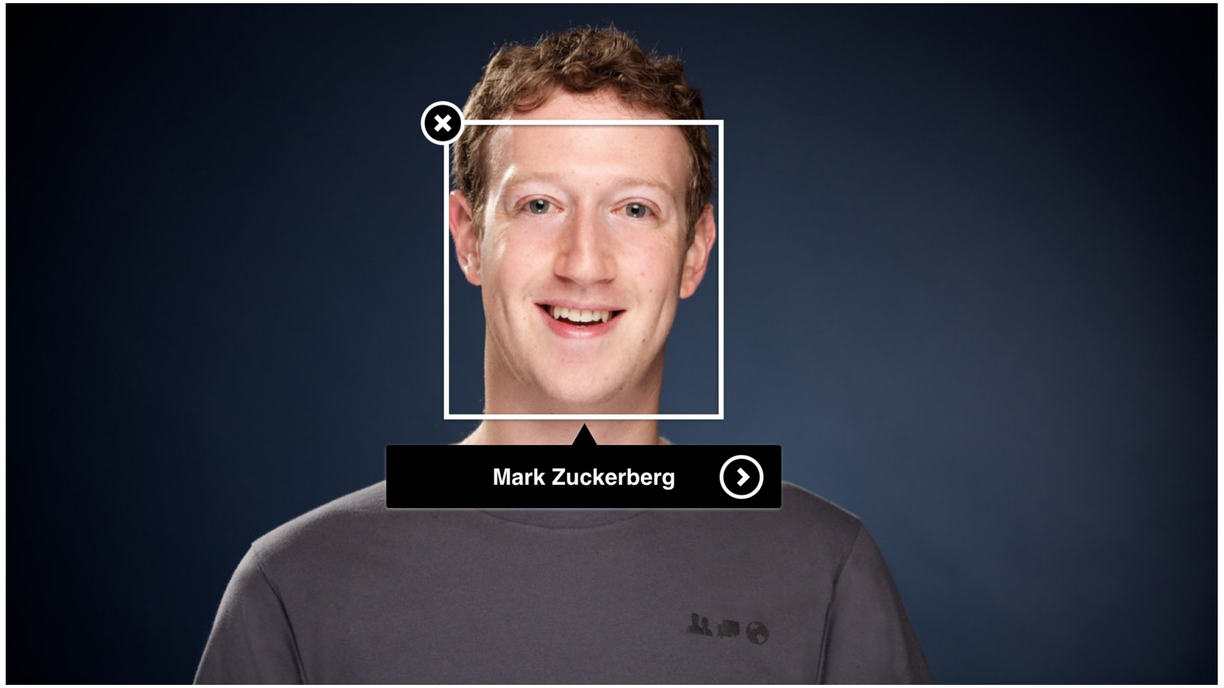


Is that it?

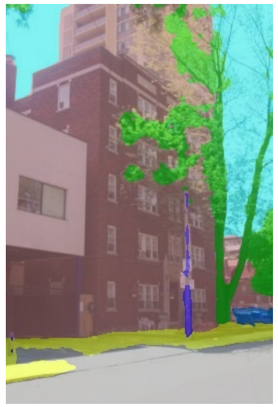


Face recognition

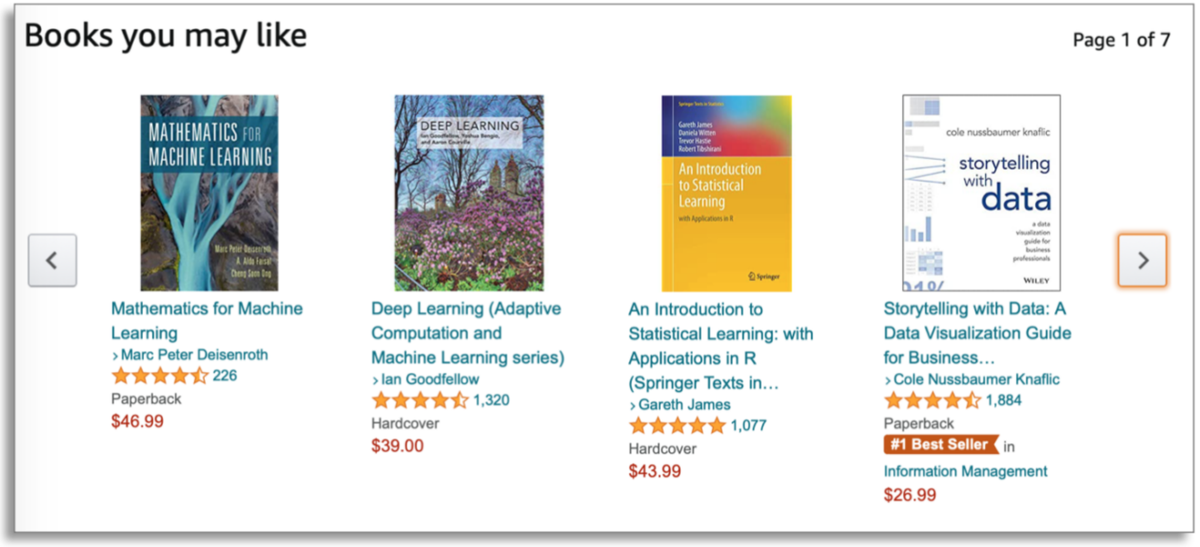


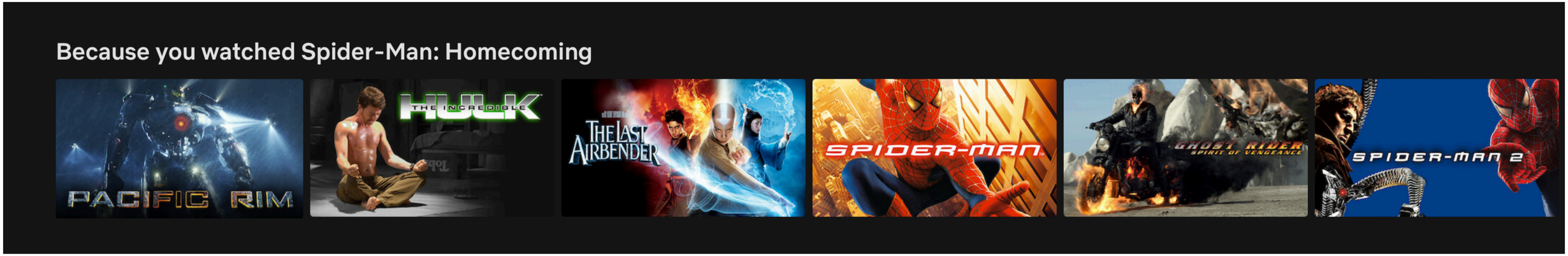


Detection and segmentation



Recommender systems





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

Tweet classifier

Classification

This is a basic prompt for detecting sentiment.

Prompt

```
Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: "I loved the new Batman movie!"
Sentiment:
```

Sample response

```
Positive
```

Grammar correction

Transformation Generation

Corrects sentences into standard English.

Prompt

```
Correct this to standard English:

She no went to the market.
```

Sample response

```
She did not go to the market.
```

https://beta.openai.com/examples/

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 a brand new course!

- I hope you enjoy it
- There may be teething problems
- Feedback is very welcome

Course outline

- Data analysis (Weeks 1-3)
- Supervised learning and ethics (Week 4)
- Linear models (Weeks 5-7)
- Non-parametric and non-linear models (Weeks 8-10)

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 to success.

Assessment: Mini-tests (50%)

- There are three of these worth 16.66% each
  - **Test 1** is at the end of the **Week 3** lab and covers Weeks 1-3 material
  - **Test 2** is at the end of the **Week 7** lab and covers Weeks 5-7 material
  - **Test 3** is at the end of the **Week 10** lab and covers Weeks 8-10 material
- Each test consists of short-answer questions and some coding exercises
- They are open-book and must be taken live and in-person

Assessment: Coursework 1 (25%)

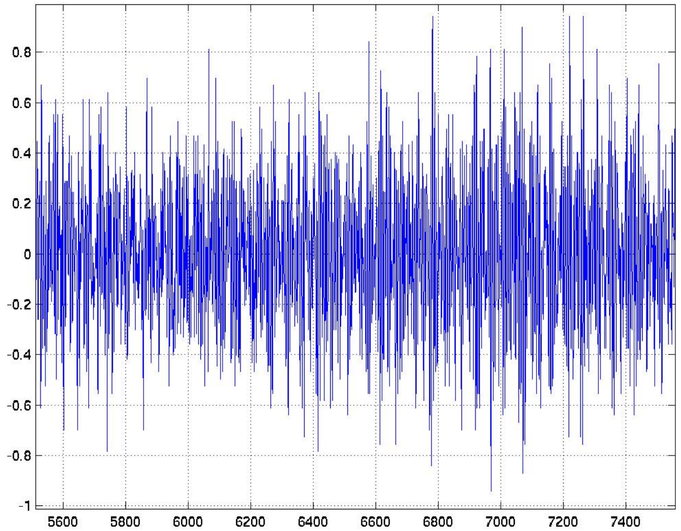
- This will be released in the Week 4 lecture on **Monday 6th February**
- You will create slides and record a short presentation using them
- This will be a case study on a real-world machine learning application
- You will critique this application from an ethical standpoint
- The deadline is **Tuesday 21st February @ 1600** (Flexible learning week)

Assessment: Coursework 2 (25%)

- This will be released in the Week 8 lecture on **Monday 13th 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 28th 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

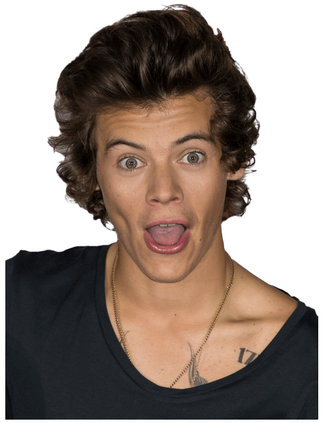
Savers Stamps
Pick up a Card and Start
Saving for Christmas Today

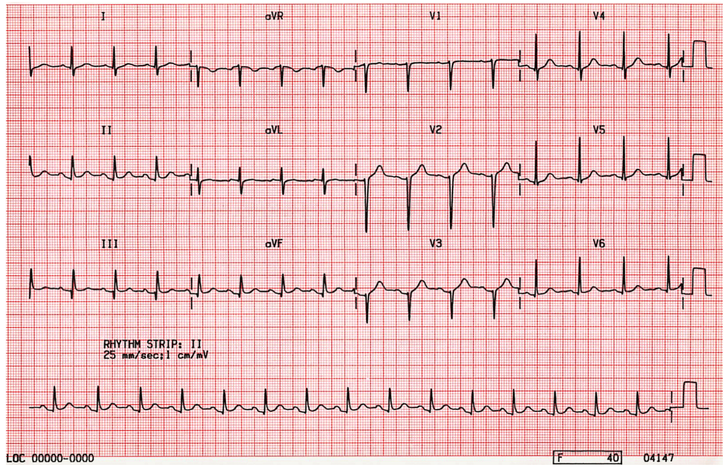
DATE: 19/06/2008 TIME: 17:54
TILL: 0019 NO: 01989232
You were served by: JENI

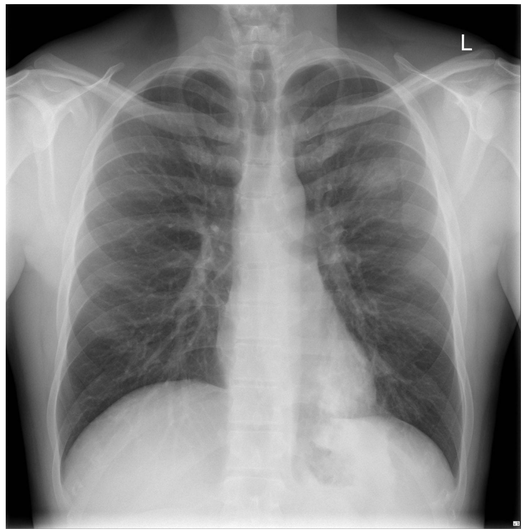
DESCRIPTION	£
'M' FRESH SEA BREAM	2.88 D
'M'SIDE OF SALMON	3.08 D
'M' KIPPER FILLETS	0.56 D
'M' PORK LEG STEAK	2.93 D
'M' BROCCOLI	
0.270kg @ £1.99/kg	0.54 D
'M'BABYLEAF AND HERB	1.39 D
'M' RED GRAPES	
0.650kg @ £3.99/kg	2.59 D
'M'BEST POTATOES	0.99 D
HORLICKS	1.34 D
DUREX EXTRA SAFE	5.98 B
'M'TRIM BEANS	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







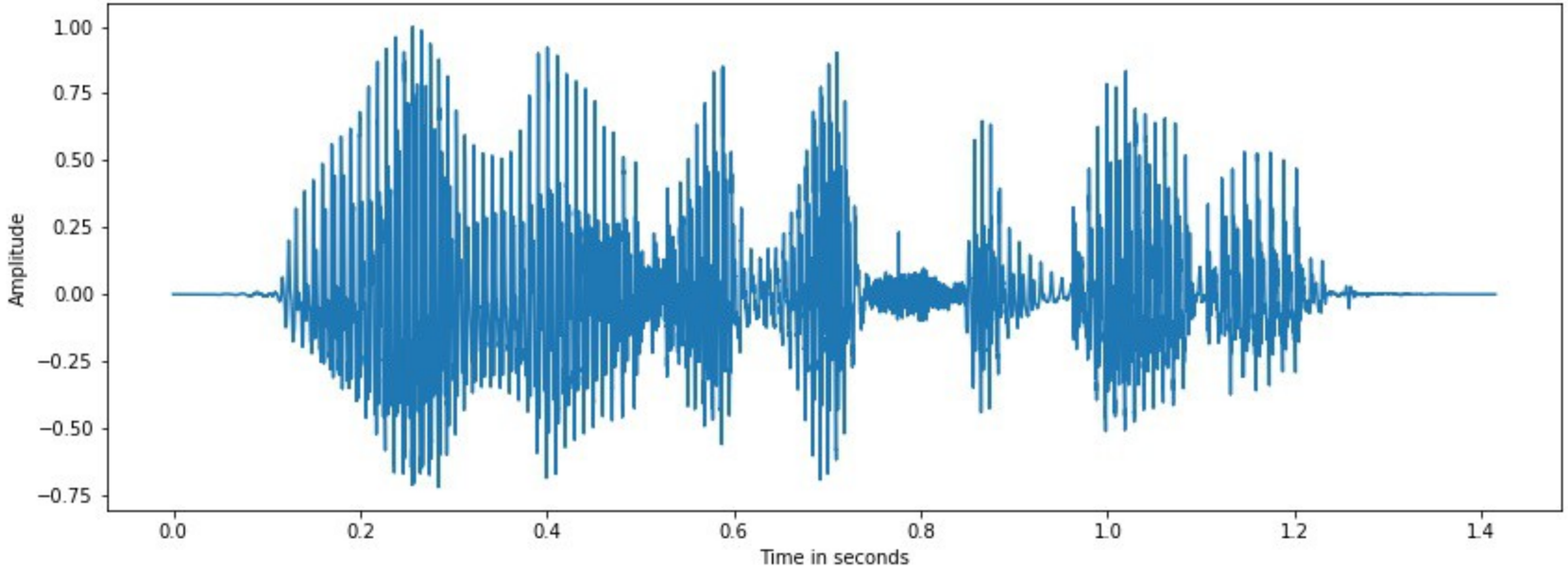
Time series data

- 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**

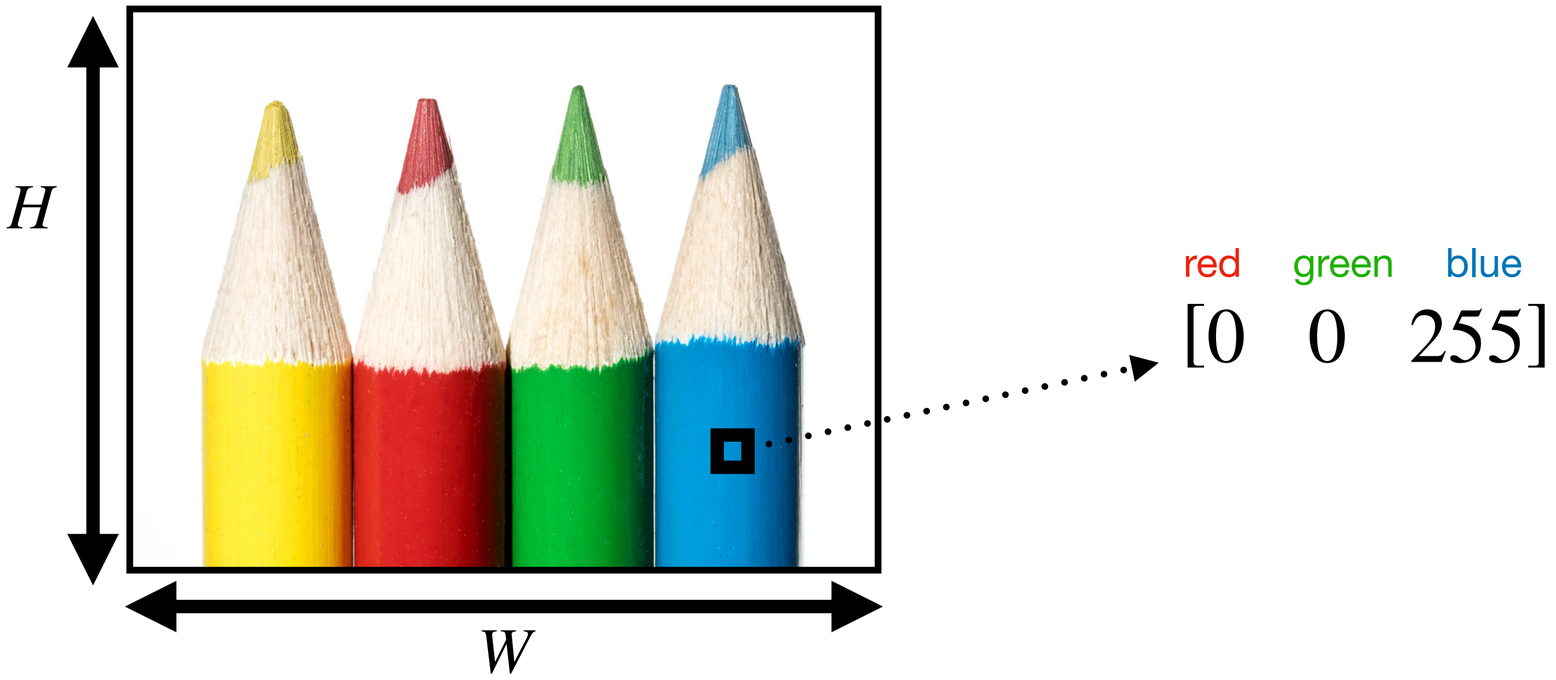
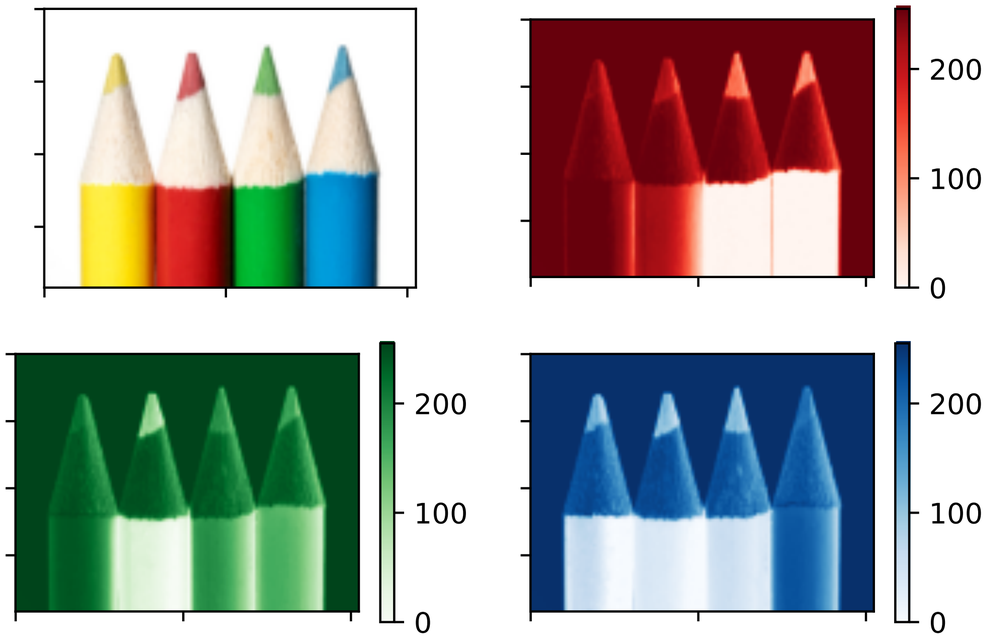
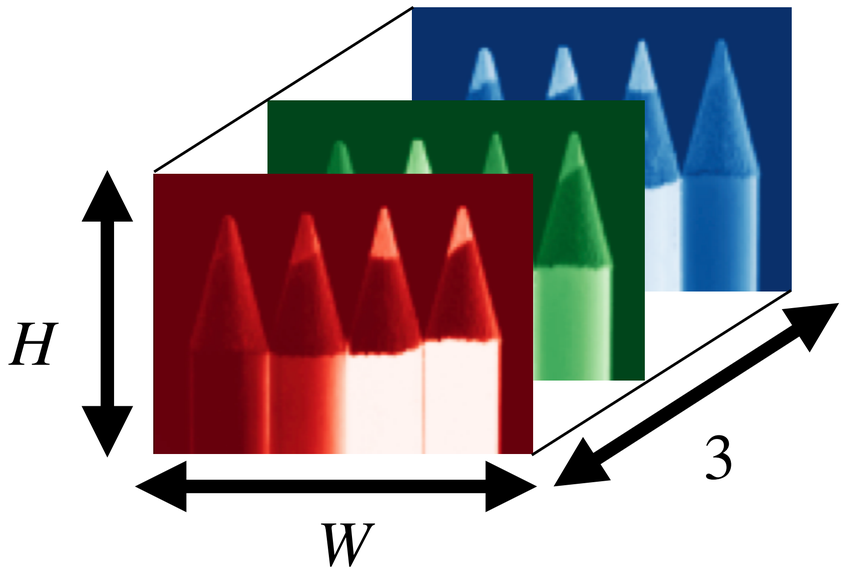


Image data

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

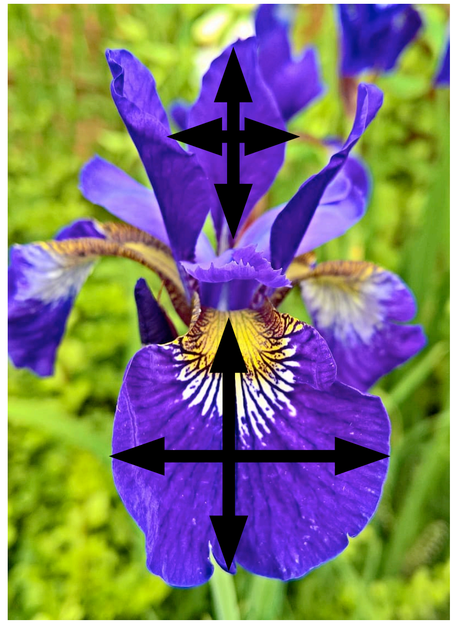




Tabular data

- 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



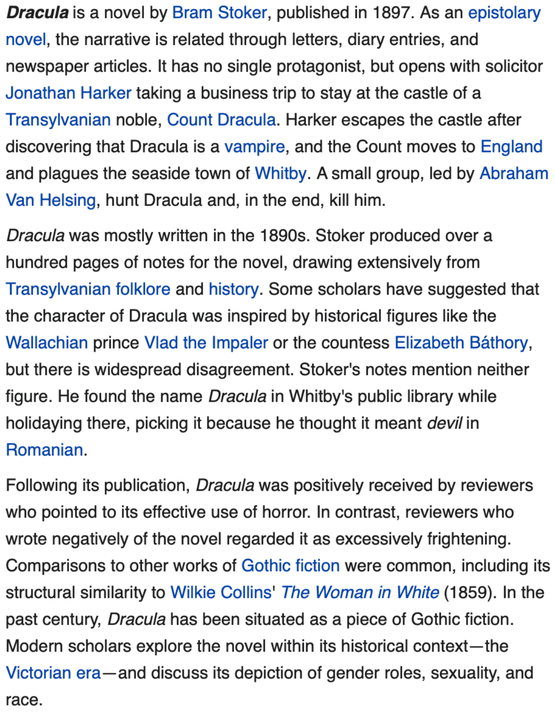
Free-form data

- 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 items (/ data points)**
- A **data item** is a set of **elements**
- An **element** is a measurable or countable quantity




A dataset of **images**

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

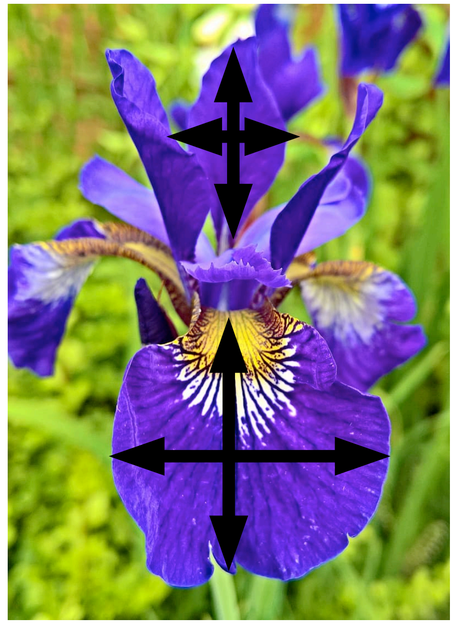
A **pixel** measures the intensity of different colour(s)

Variable Types

Tabular data (again!)

- A table is a dataset and its rows are data items
- Each data item is a set of elements which are measurements of attributes
- 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 **vary** across the dataset
- This means we can think of the attributes 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

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 whole numbers
- 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

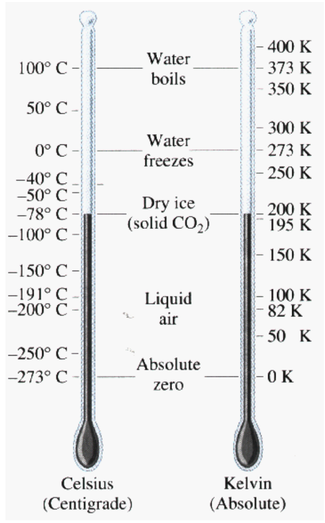
Continuous variables

These can be further divided into interval and ratio

- For **interval variables** there is no true zero measurement; It's relative
- For **ratio variables** zero has a clear meaning i.e. the absence of something

Temperature in Celcius is interval

Temperature in Kelvin is absolute



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