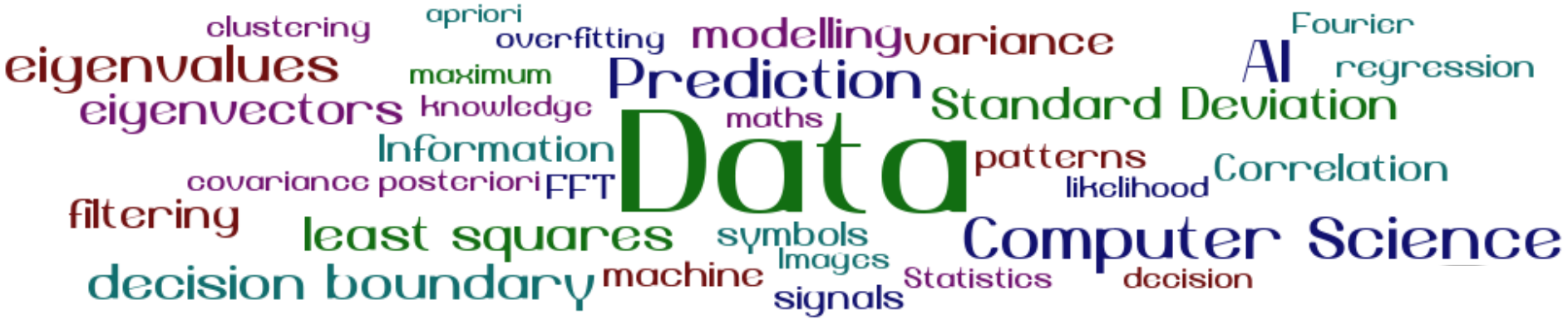
COMS20017 – Algorithms & Data



January 2025

Majid Mirmehdi

Lecture MM-01

What is Data?

- Data comes in many forms, e.g. text, symbols, patterns and signals!
- Data: *Structured and Unstructured*
  - Numeric (measurements, finance spreadsheets, ...)
  - Textual (emails, social media, web pages, medical records, ...)
  - Visual (images, video, graphics, animations)
  - Auditory (speech, audio)
  - Signals (GPS signals, accelerometer, heart rate, ...)
  - Many others...

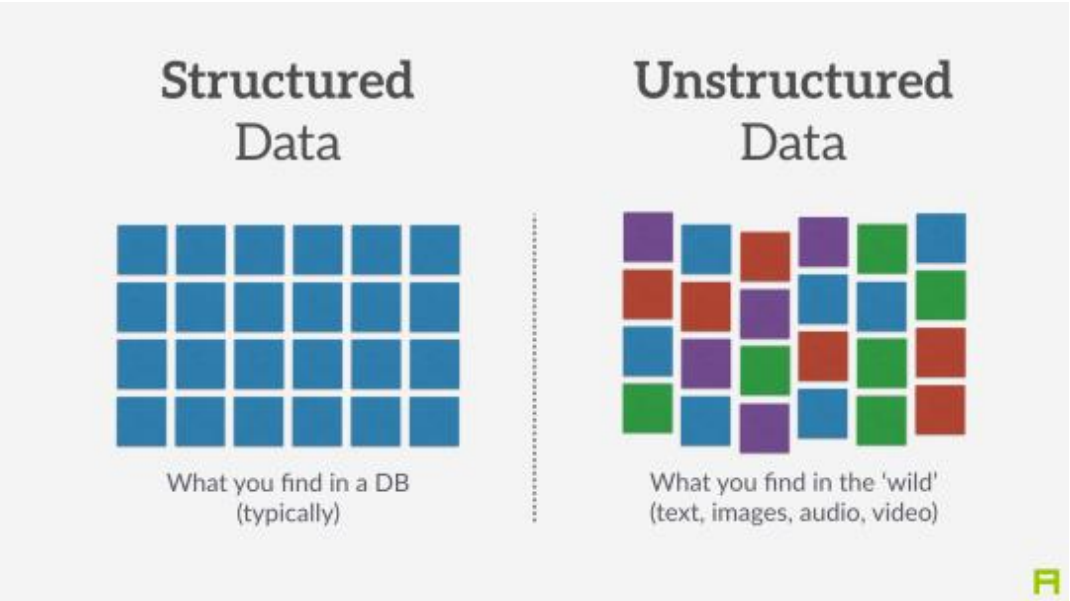


Image from Garrett Hollander

This Unit

- This unit is about doing things with data... *but* *not*
  - storing, shuffling, searching (Algorithms I & II)
  - sending (Computer Systems)
  - compressing or encrypting (Cryptology)

- This unit is about:
  - extracting knowledge from data
  - generating data and making predictions
  - making decisions based on data
  - Often referred to as:

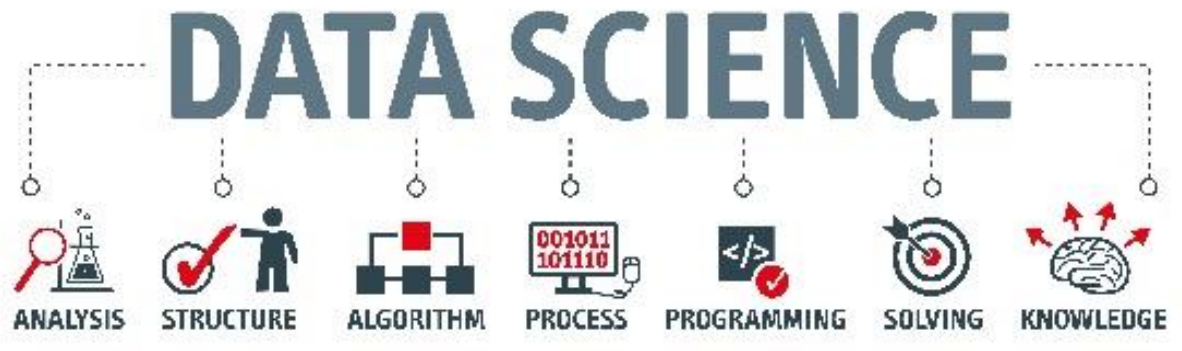


Image from https://www.datanami.com

Basic Statistics Concepts

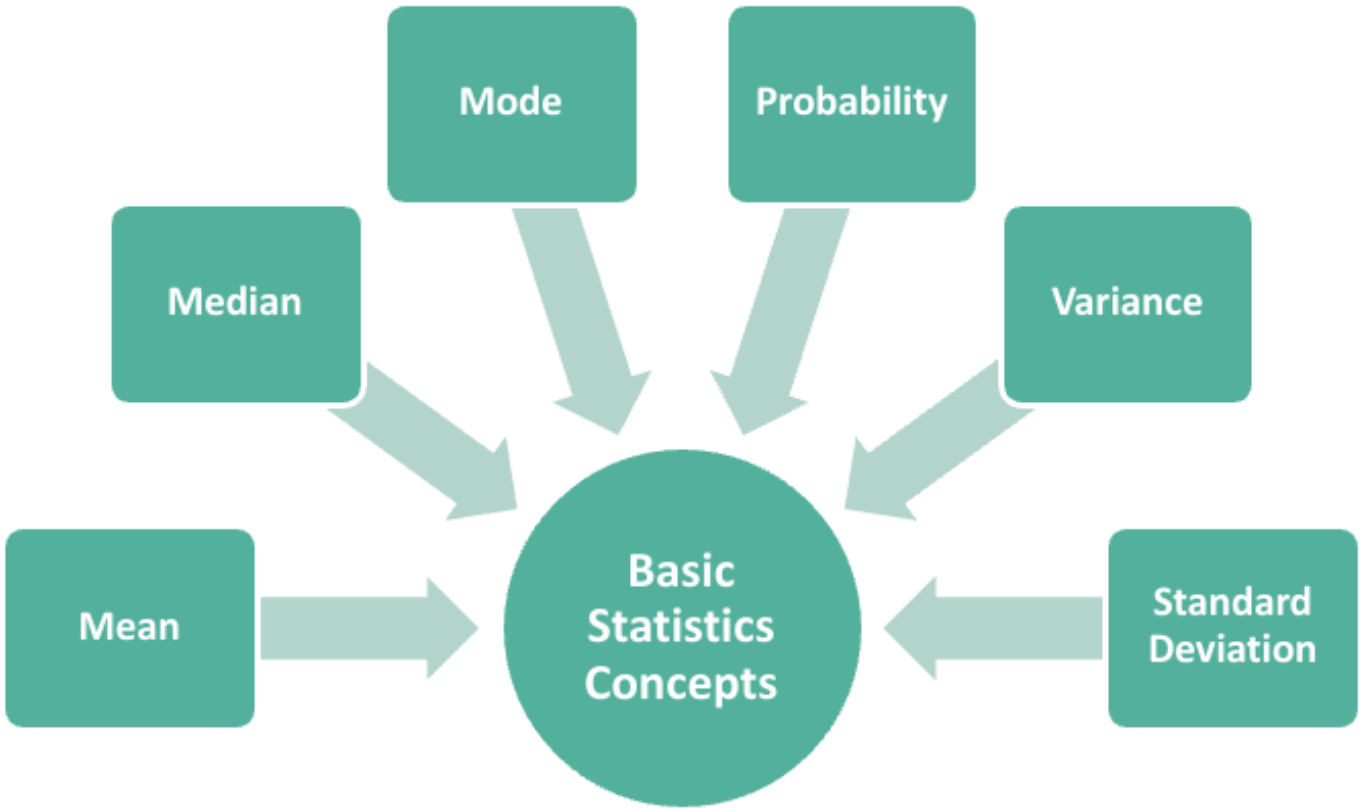


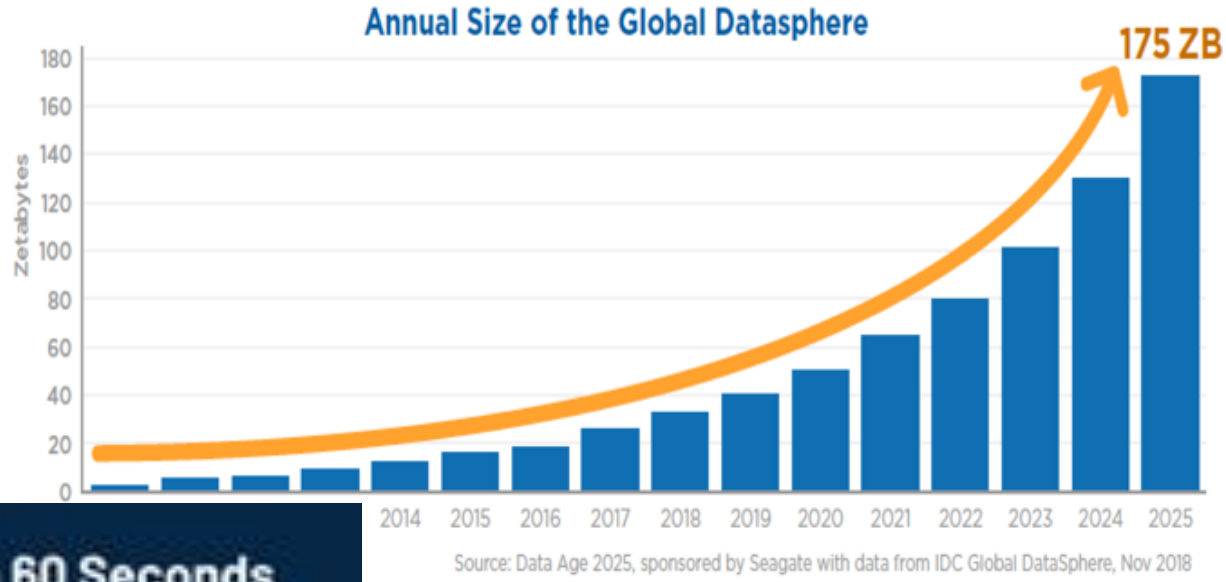
Image from https://www.wallstreetmojo.com/basic-statistics-concepts/

Same Basic Stats, Different Data!

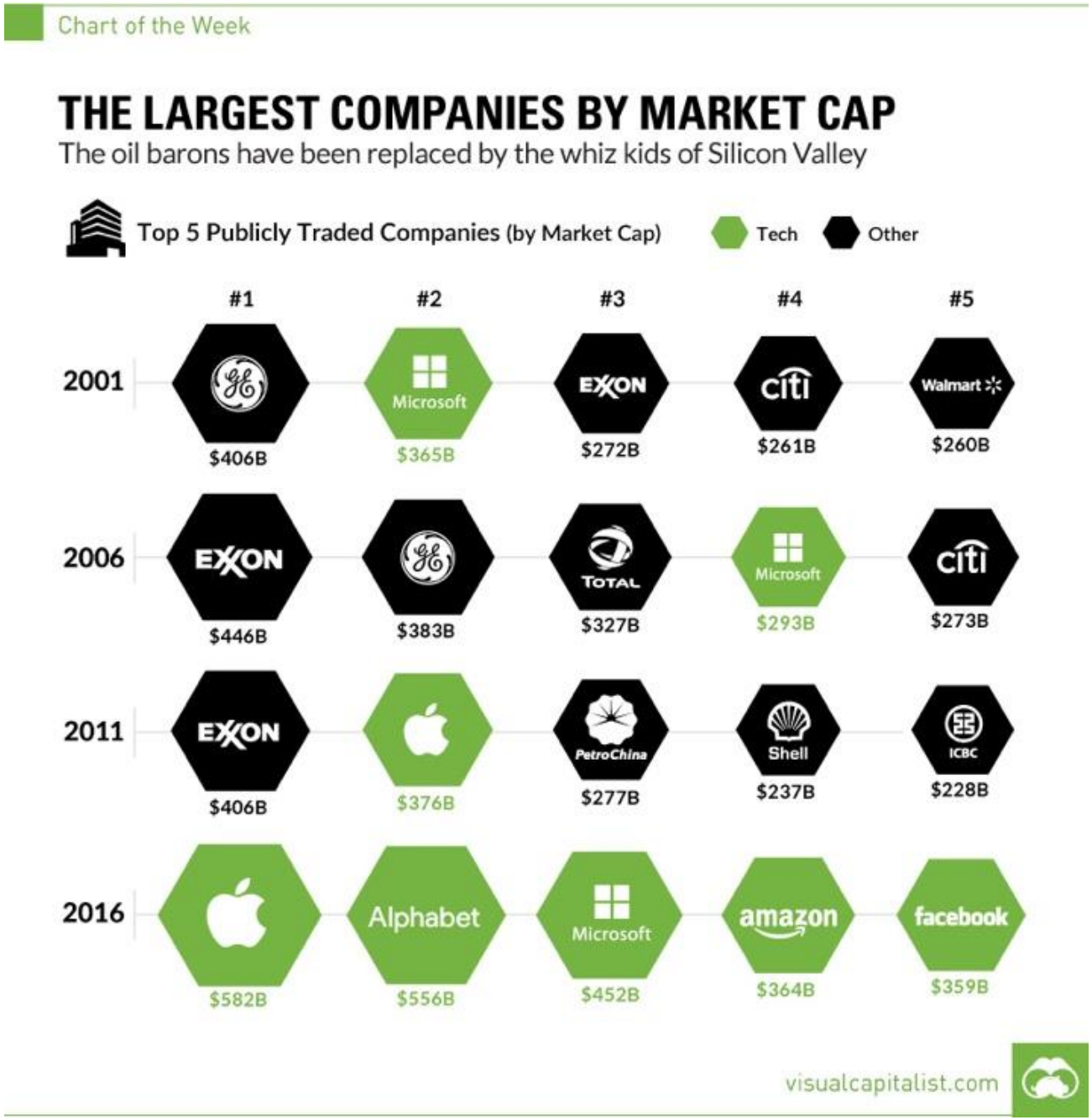
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Clip from https://www.autodesk.com/research/publications/same-stats-different-graphs

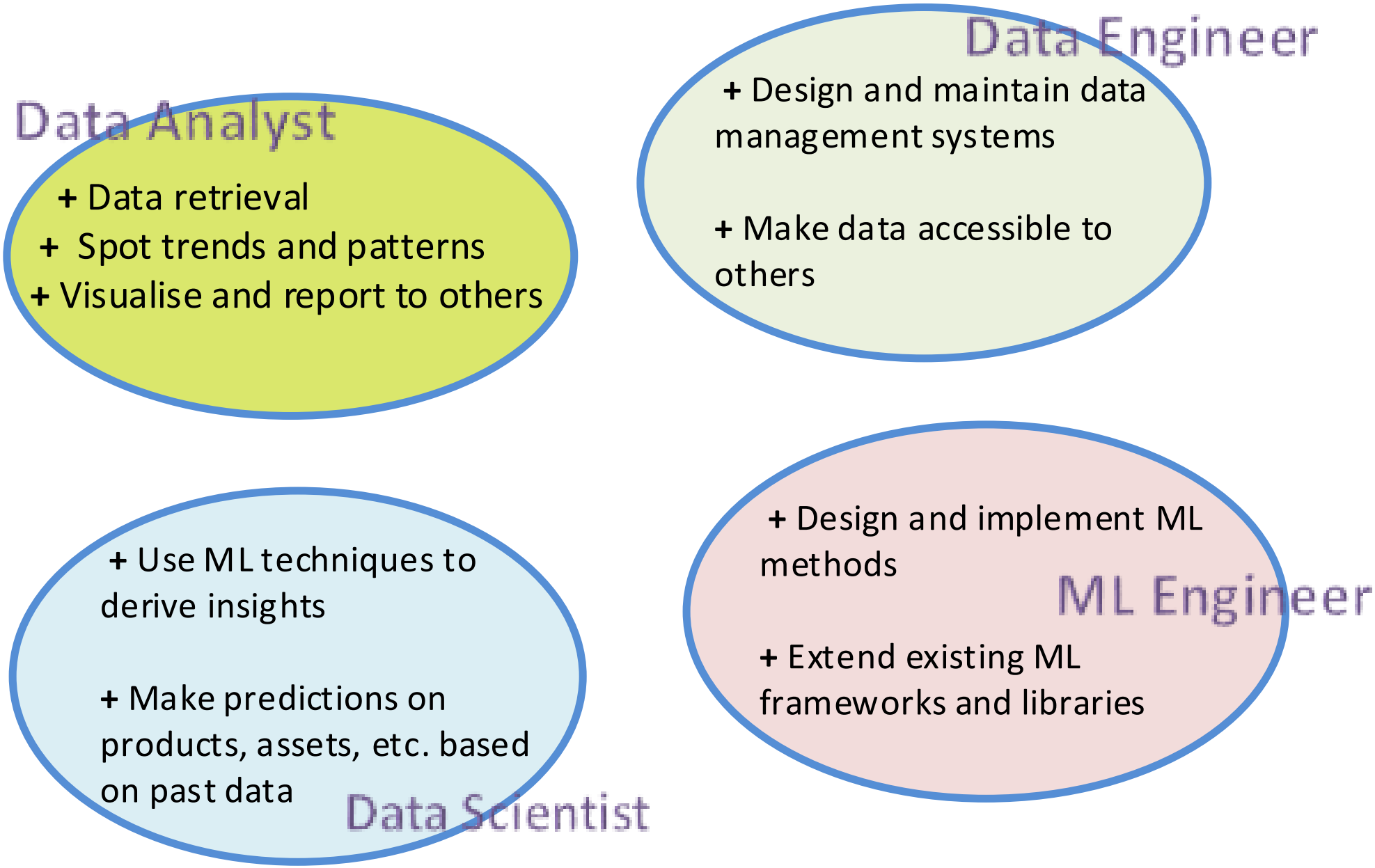
Amount of Data!



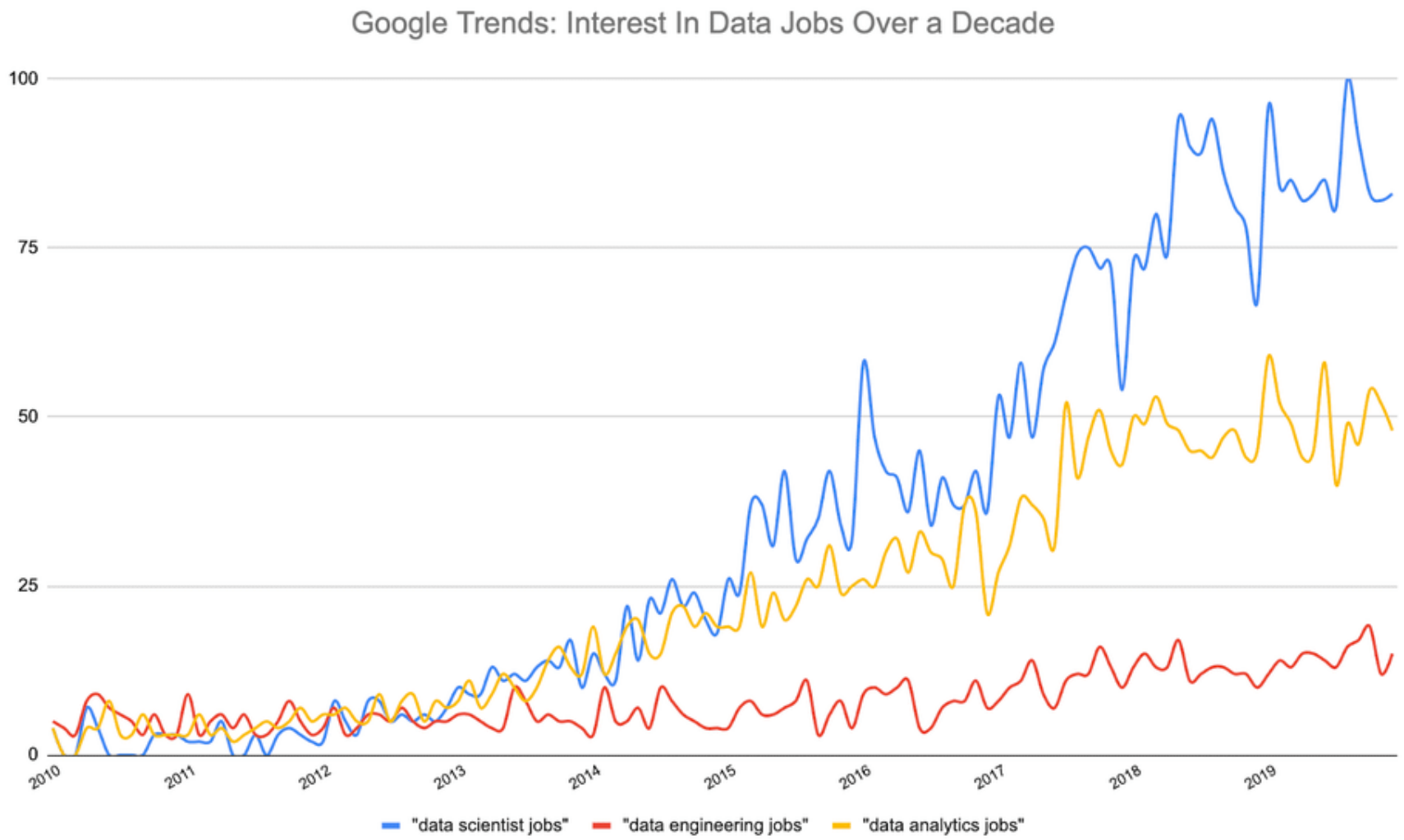
Data is the new Oil



Example Job Positions Involving Data



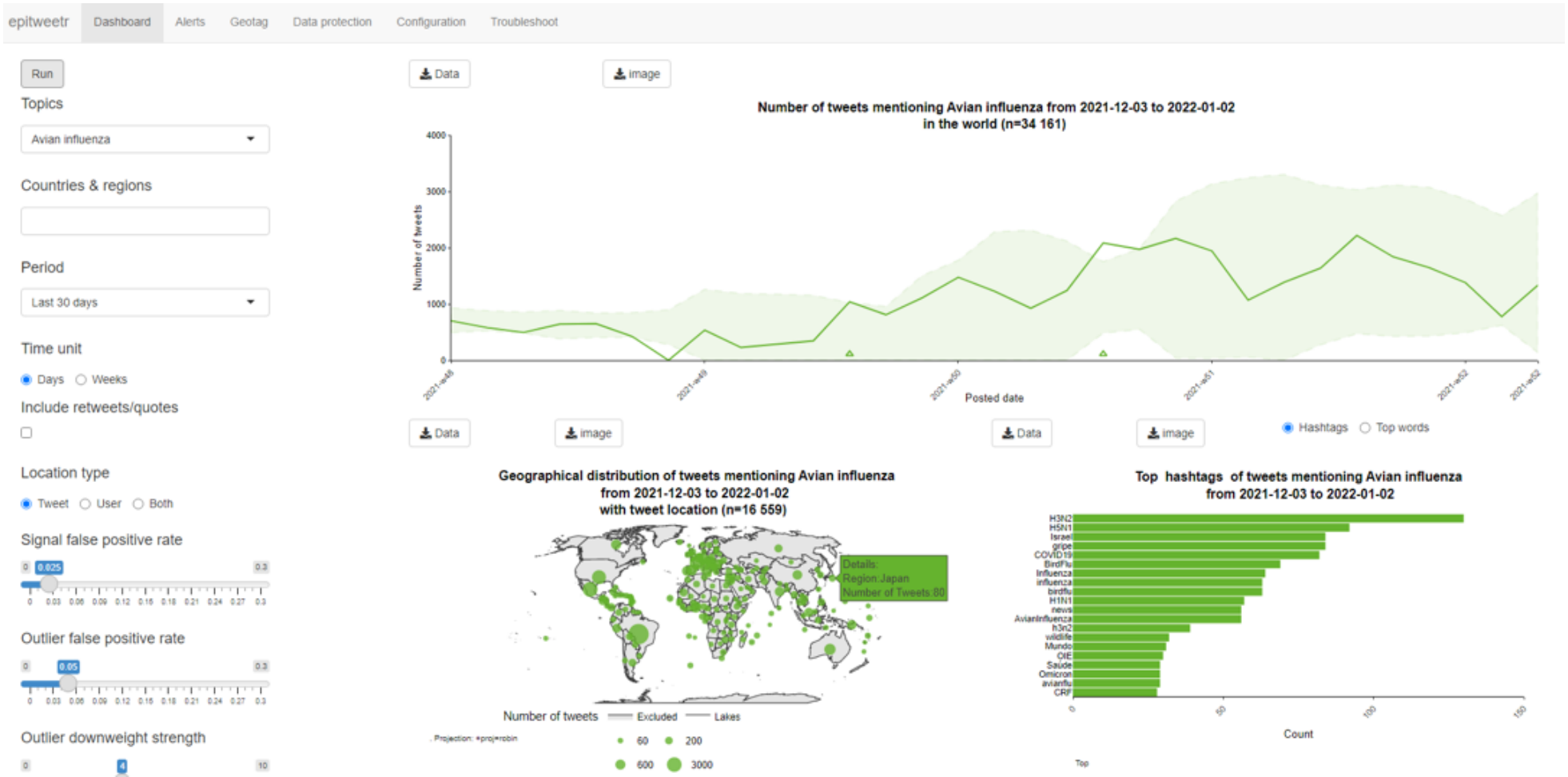
Data Science & Analytics



https://onlinedatasciencemasters.virginia.edu/blog/data-science-vs-data-engineering/

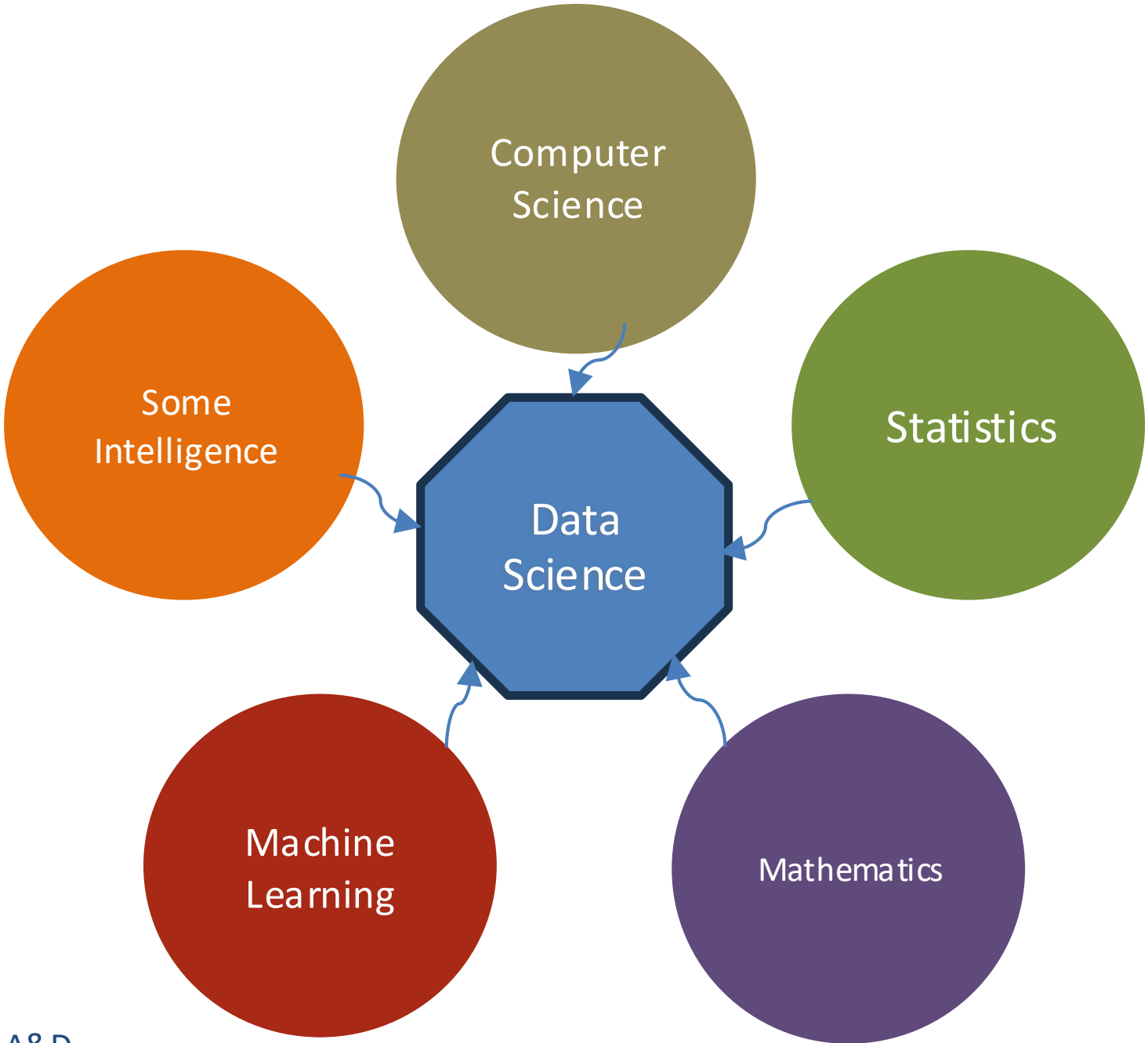
It’s not about the data – it’s about the science

Tracking and predicting [disease,mortality,floods,fires,fun etc.] by Twitter!



https://www.dicardiology.com/article/understanding-how-big-data-will-change-healthcare

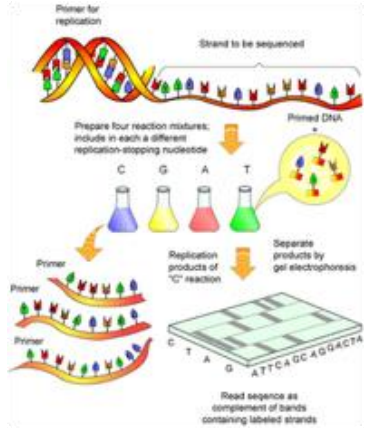
It's not about the data – it's about the science



This Unit

Why is it important for Computer Science?

- Fundamental to many related areas:
  - Artificial Intelligence, Machine Learning, Deep Learning
  - Image Processing and Pattern Recognition
  - Graphics, Animation and Virtual Reality
  - Computer Vision and Robotics
  - Speech and Audio Processing
  - With growing applications in: neuroscience, literature, agriculture, etc.
- Hence, preparation for units in years 3 and 4.



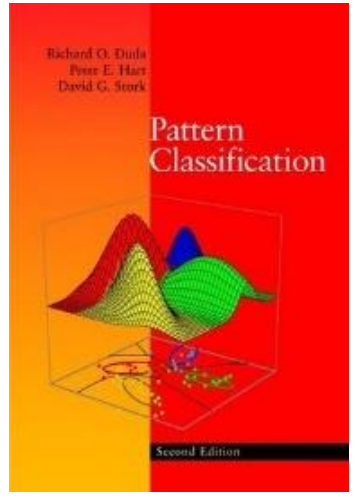
https://www.bris.ac.uk/unit-programme-catalogue/UnitDetails.jsa?unitCode=COMS20017

Ex1. A Fish Problem

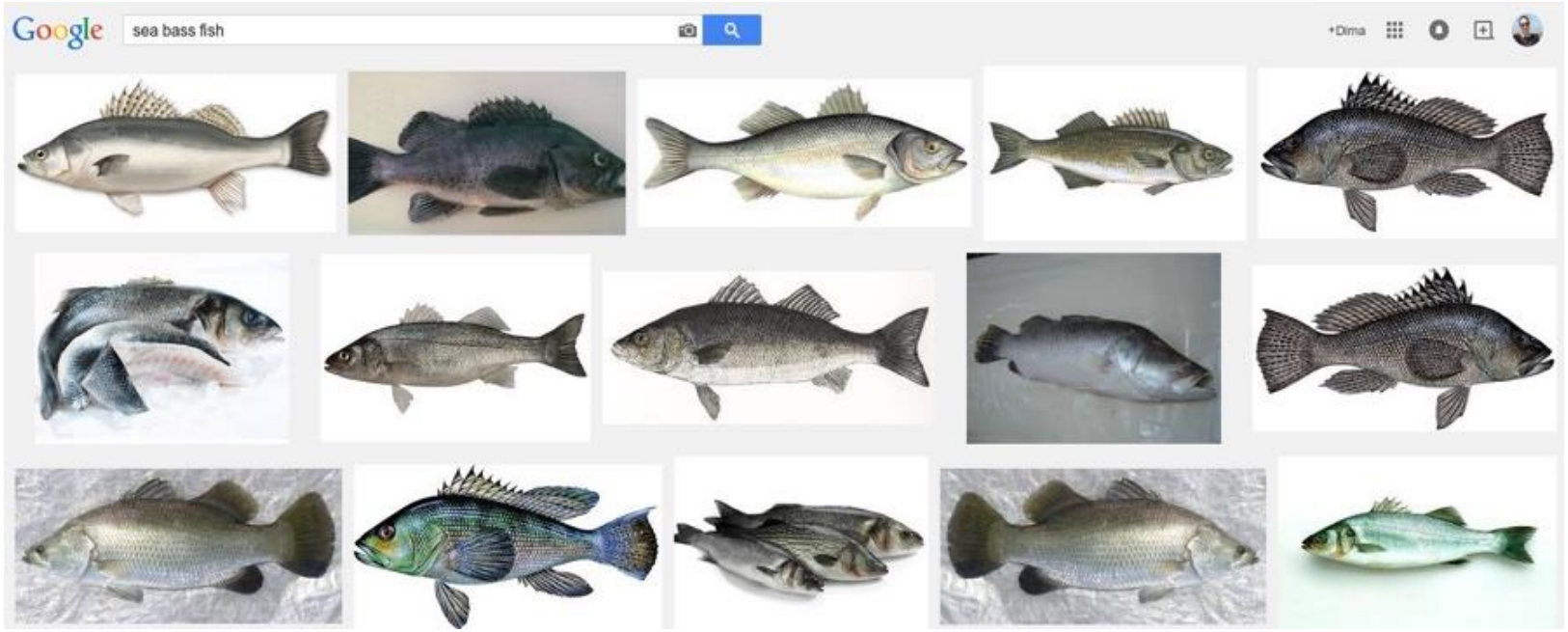
Data: images of fish

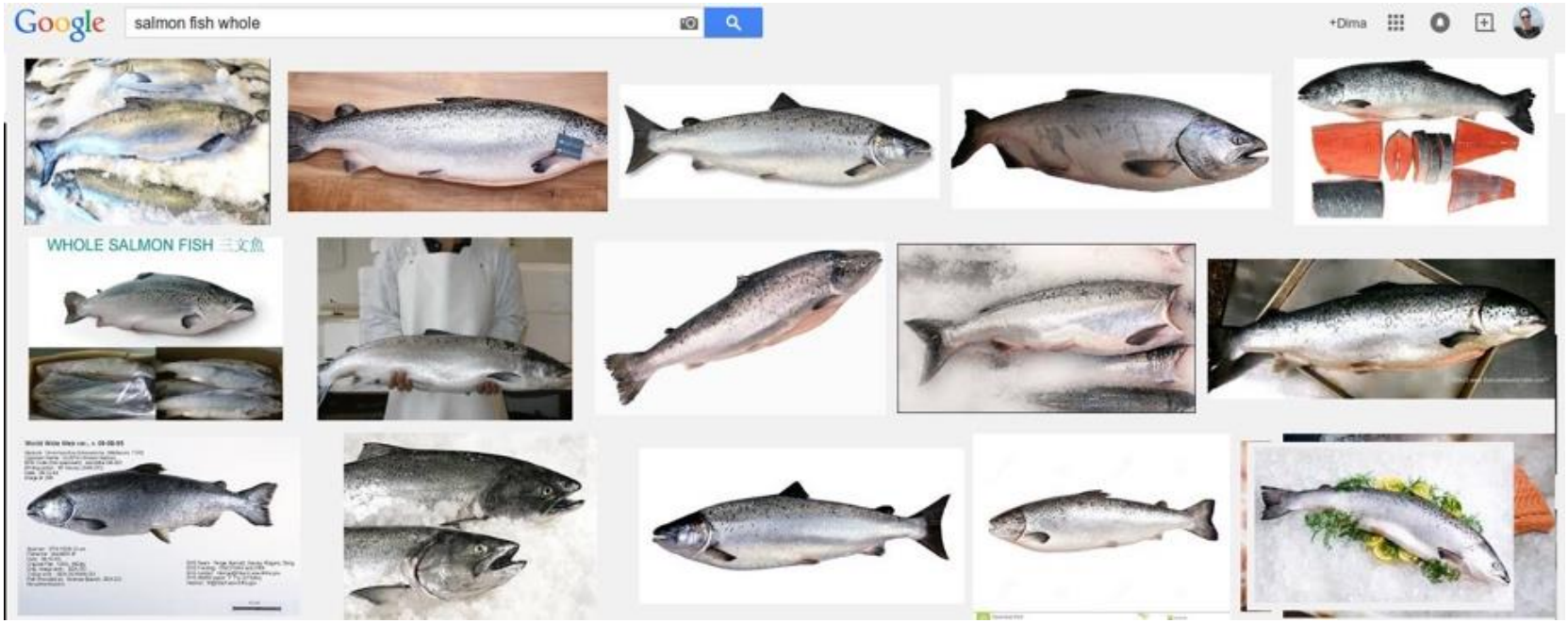
Aim: distinguish between sea bass and salmon

From: Pattern Classification by *Duda, Hart and Stork*, 2nd Edition, Wiley Interscience



Ex1. A Fishy Problem



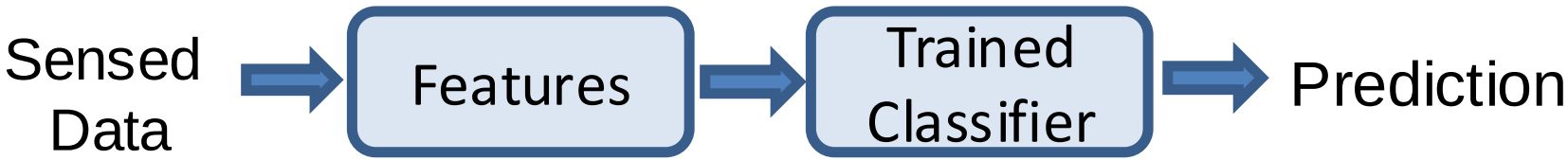


Features

They are the intrinsic traits, properties, or characteristics that tell one data/pattern/object apart from another.

Feature extraction and representation allows:

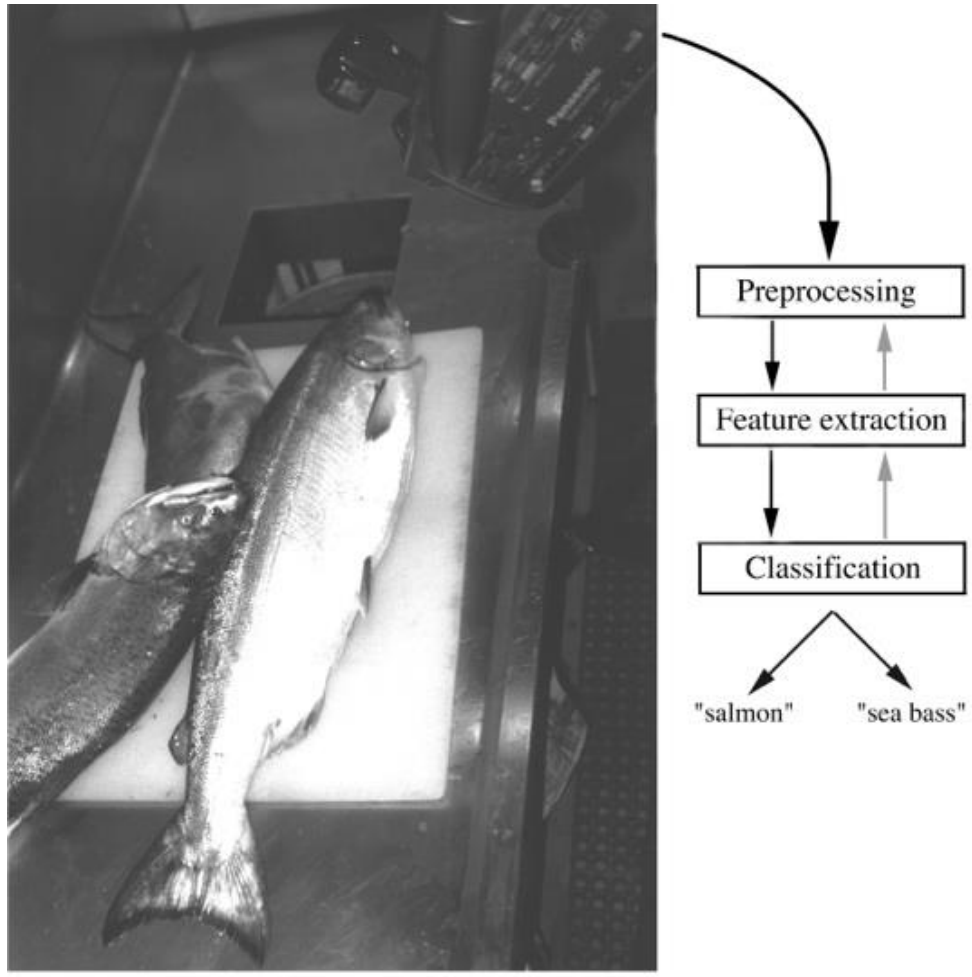
- Data reduction and abstraction
- Focus on relevant, distinguishing parts of data



Fishing for a Solution

Steps:

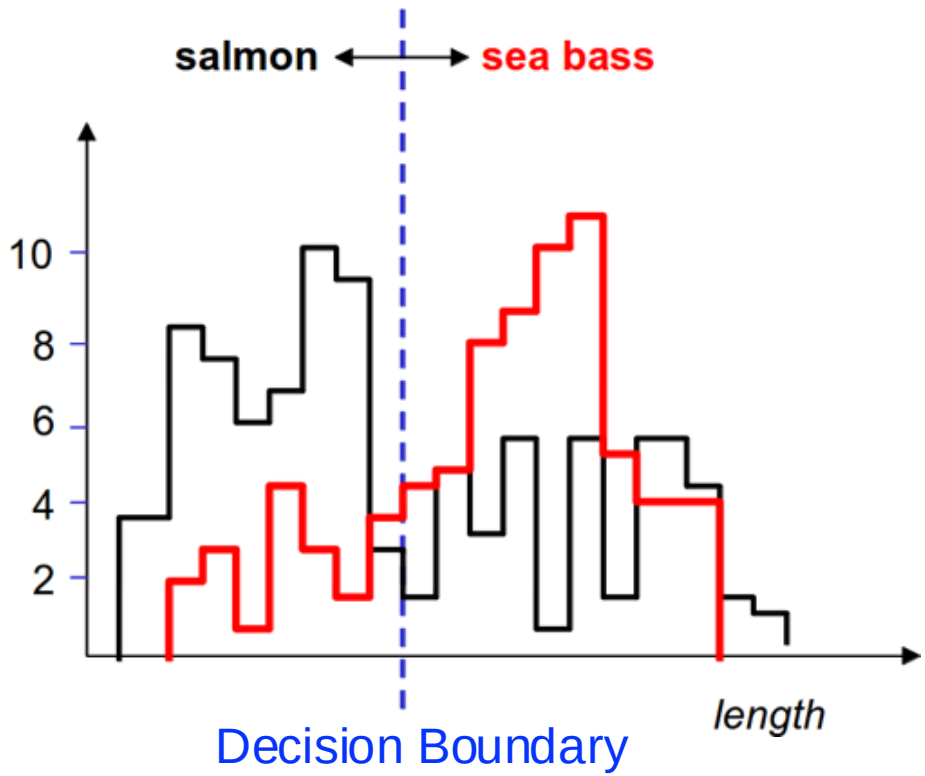
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. Measure length
3. Classification e.g. Find a threshold

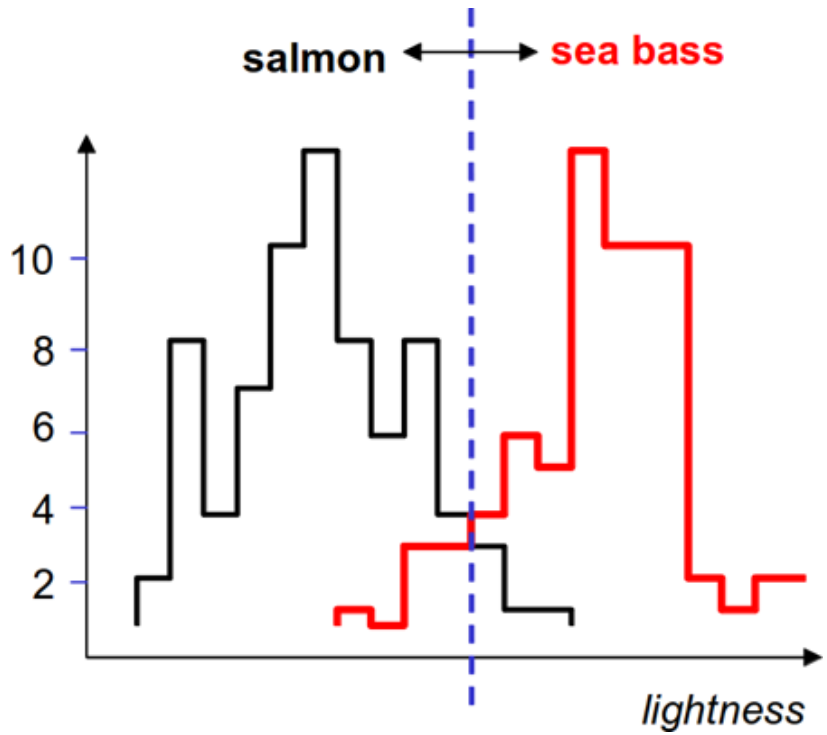


Fishing for a Solution

Steps:

1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. Measure length or lightness
3. Classification e.g. Find a threshold

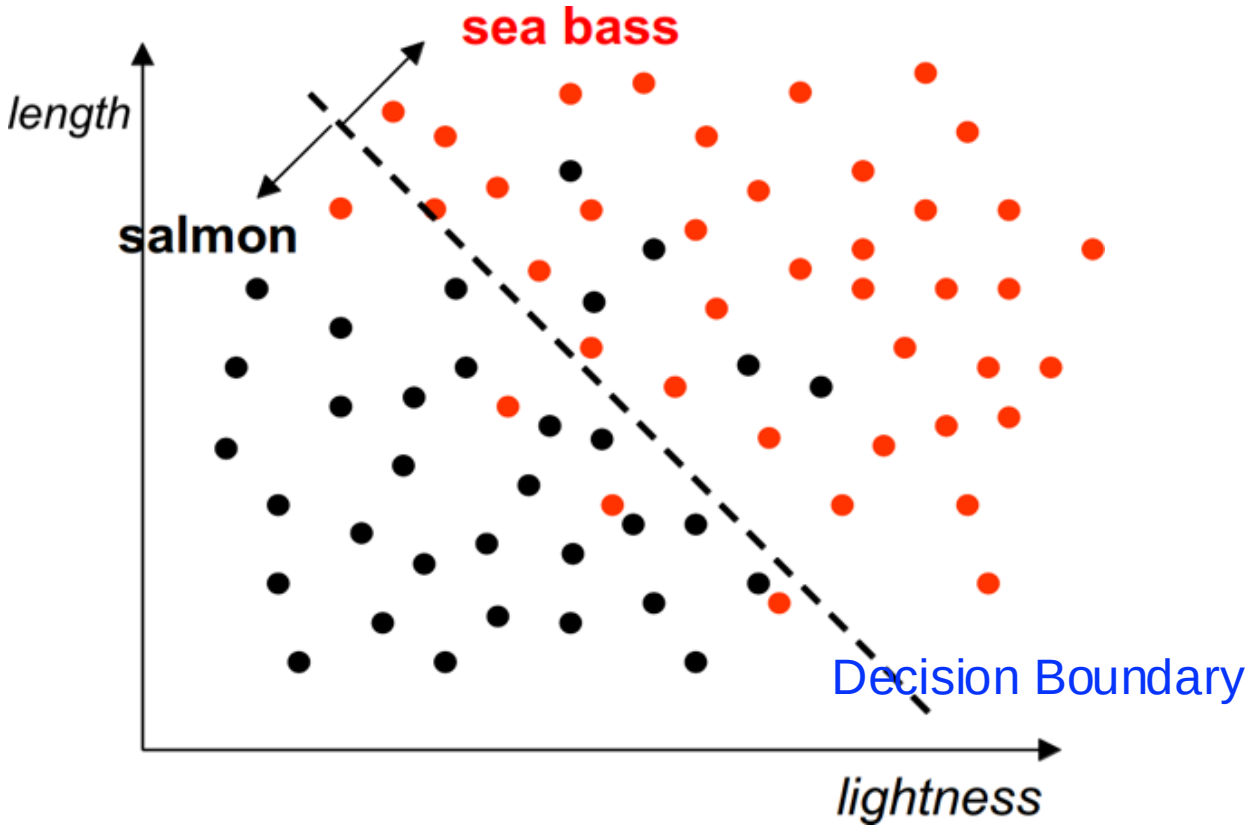




Fishing for a Solution

Multiple features could be selected, resulting in a multi-dimensional feature vector.

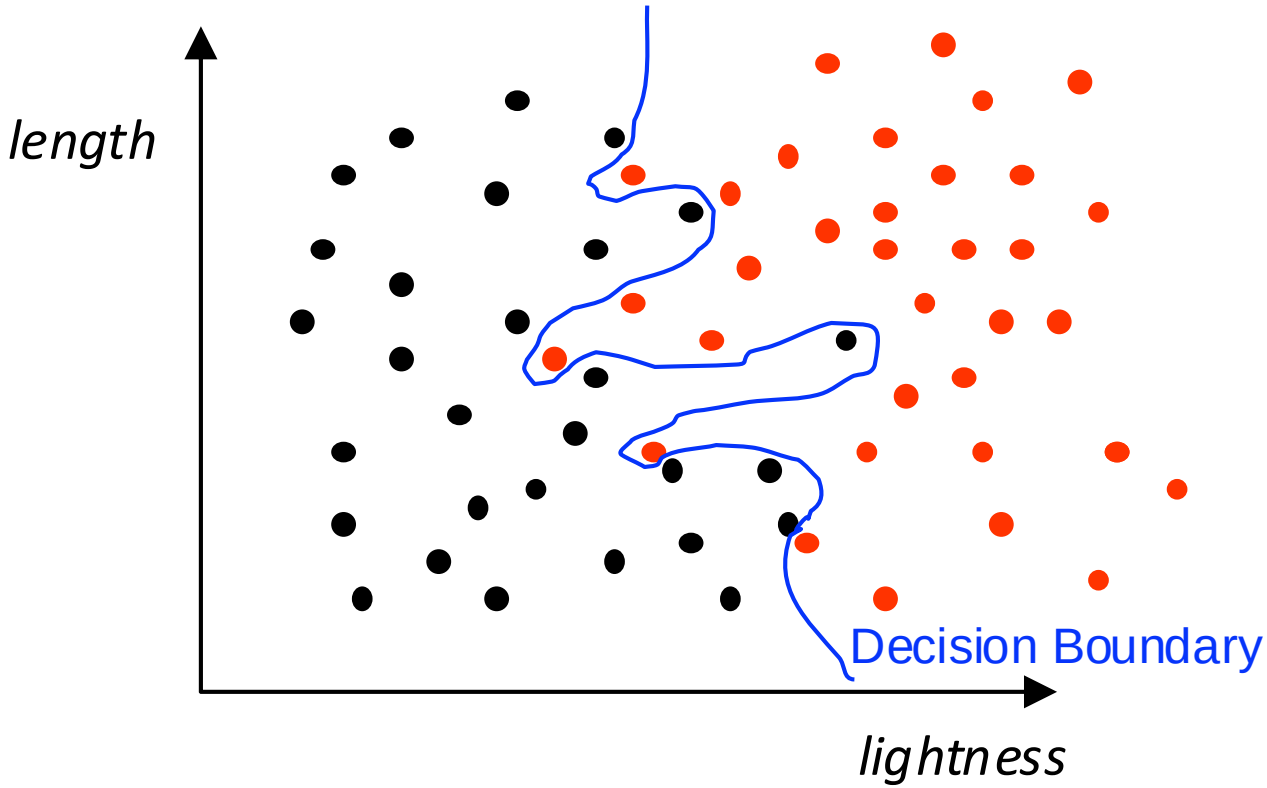
Scatterplot!



$$\textbf{Fish} \rightarrow \mathbf{x} = \{x_1, x_2\}$$

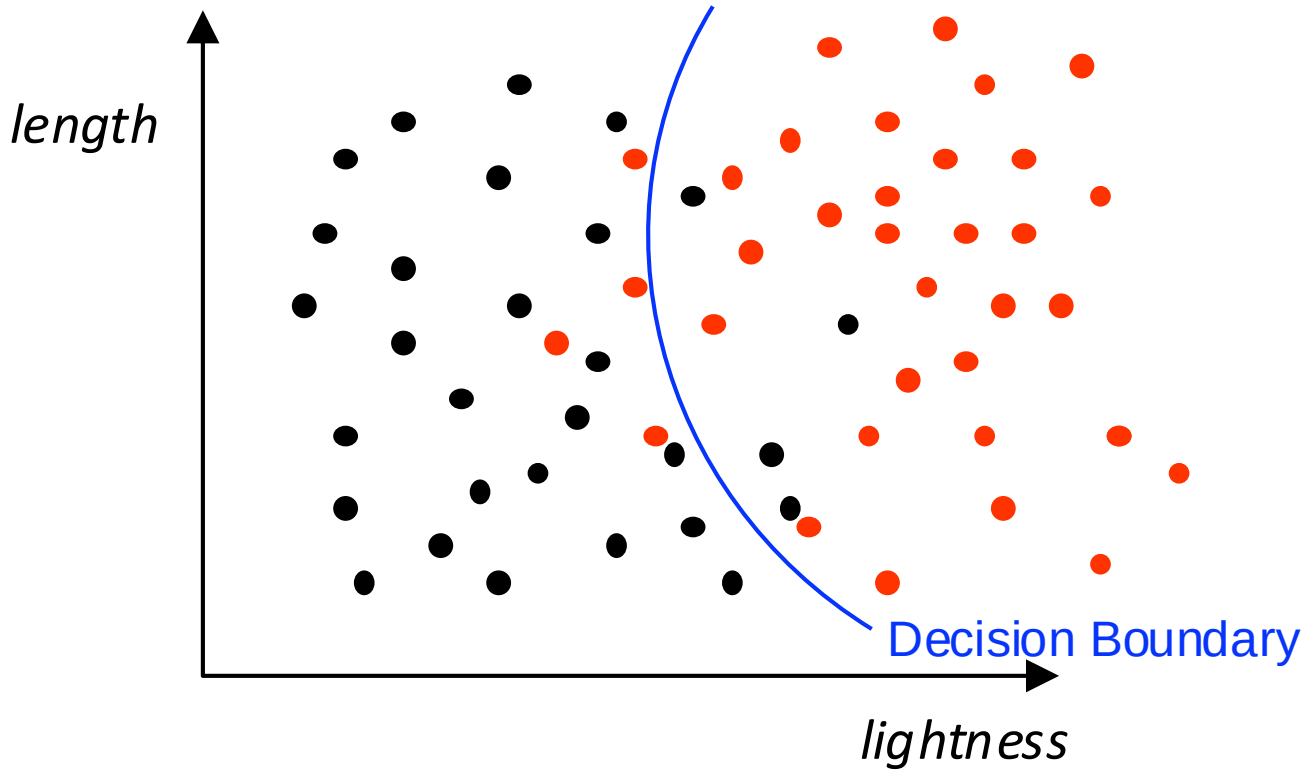
Fishing for a Solution

Complex decision model

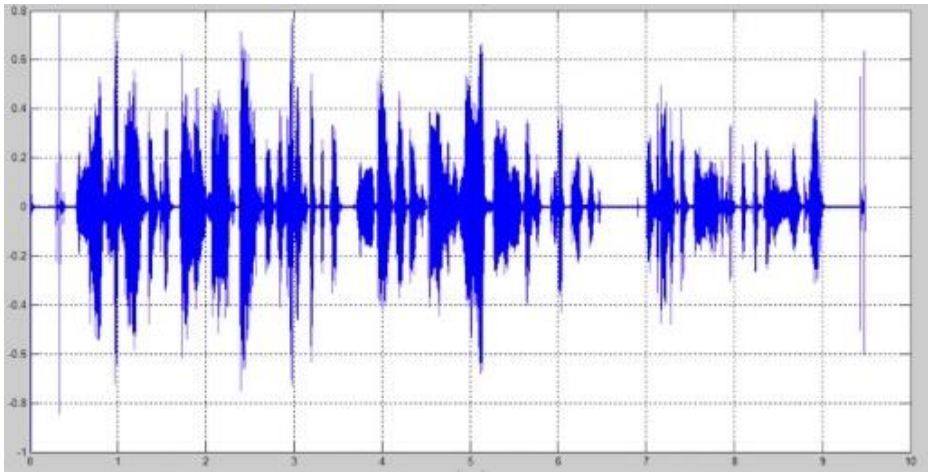


Fishing for a Solution

Optimal trade-off between performance and generalization



Ex2. Speech Recognition



Data: Analogue speech signals (time series numerical data)

Aim: Convert audio into text (e.g. Alexa/Siri…)

1. Pre-processing Digitisation
2. Feature Selection Wave amplitude, frequencies
3. Inference Hidden Markov Models (Viterbi algorithm) or Deep learning

Ex3. Spam Filter

Data: Texts of emails

Aim: Determine whether the email is spam

1. Pre-processing - Normalise words (e.g. remove punctuation, find word roots)
2. Feature Selection - Presence of words
3. Classification - Naive Bayes classifier

Select subset of words w_i and determine $P(w_i \mid spam)$ and $P(w_i \mid \neg spam)$ from frequencies in training data.

For an Email that contains $w_1, w_2, .., w_n$ of the subset of words, assume

$$P(email \mid spam) = P(w_1 \mid spam) P(w_2 \mid spam) .. P(w_n \mid spam) \quad (1)$$

and

$$P(email \mid \neg spam) = P(w_1 \mid \neg spam) P(w_2 \mid \neg spam) .. P(w_n \mid \neg spam) \quad (2)$$

A new Email is spam if

$$P(email \mid spam) > P(email \mid \neg spam) \quad (3)$$

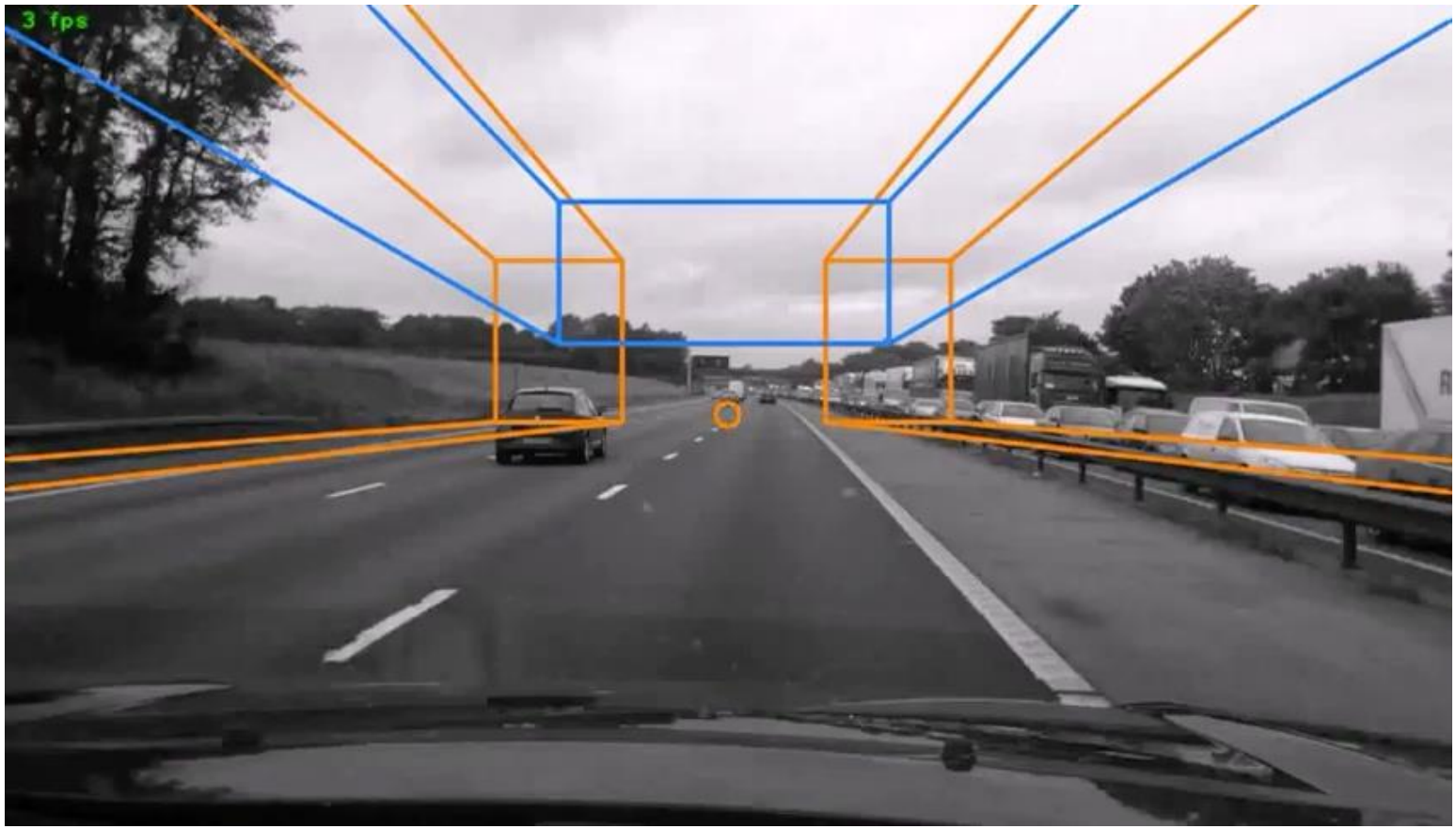
Image from https://www.kdnuggets.com/

Ex4.1 – Towards Autonomous Driving

Data: Video

Aim: Determine knowledge from the road or inside the vehicle

1. Pre-processing (Detect vanishing point)
2. Feature Selection (Use constraints to reduce number and dimensionality)
3. Recognition (Perspective transformations and OCR)



Ex4.2 – Towards Autonomous Driving

1. Pre-processing (Detect vanishing point)
2. Feature Selection (Straight lines)
3. Model Building (Detecting, predicting, decision making)

Ex4.3 – Towards Autonomous Driving

1. Pre-processing (Detect vanishing point)
2. Feature Selection (MSERs, Histogram of Gradients)
3. Classification (Support Vector Machines)

Ex4.4 – Towards Autonomous Driving

1. Pre-processing (Background subtraction)
2. Feature Selection (hand shapes)
3. Classification (Random Forest classifier)

COMS20017 - Data

Steps:

1. Pre-processing [Unit - Part 1] → Majid Mirmehdi (~10%)
2. Feature Selection [Unit - Part 3] → Majid Mirmehdi (~40%)
3. Modelling & Classification [Unit - Part 2] → Alin Achim (~50%)

COMS20017 - Data

Lectures (note exceptions listed on github page)

Mondays 3pm in PHYS BLDG G42 POWELL

Thursdays 2pm in PHYS BLDG G42 POWELL

Unit pages: https://github.com/majidmirmehdi/COMS20017_DATA_24-25/

Labs

Fridays 13:00 - 14:00 [by timetable]: Group 1

Fridays 14:00 - 15:00 [by timetable]: Group 2

Lab Environment [Jupyter + Python]

TA support in unit's Teams group

Lectures and Labs are both essential for learning unit content!

Very Welcome to Ask Questions…

You should use the **Teams channel** for raising queries on whatever aspects of the COMS20017 Data unit!

Queries will normally only be answered via email or via personal Teams messages, IF it is a personal question that cannot be shared.

Please post your query on the unit Teams channel for the benefit of others who may have the same query.

Next lecture

Analog Signal

Digital Signal

- **Data acquisition**
- **Data characteristics: distance measures**
- Data characteristics: summary statistics [*reminder*]
- Data normalisation and outliers