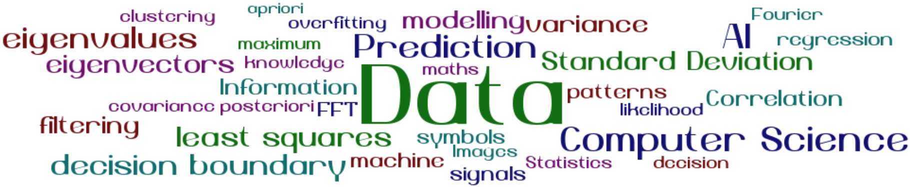
COMS20011 – Data-Driven Computer Science



January 2023

Majid Mirmehdi

Some slides in this lecture are adapted from those authored by **Dima Damen** and **Andrew Calway**

Lecture #1

COMS20011 Unit

- This is a “new” unit that started in the 2020-21 academic year
- Replaced the 20CP COMS20212 (SPS) unit
- Exam materials can be used for revision BUT...
- Use SPS materials with caution...depth, breadth & requirements may differ.

What is Data?

- Data comes in many forms, e.g. 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...

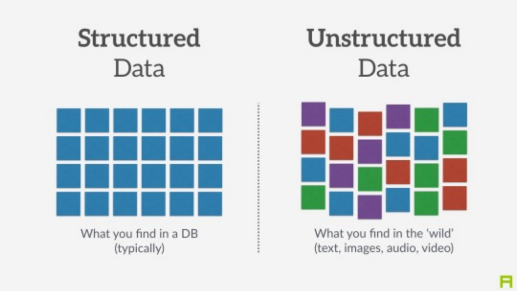


Image from Garrett Hollander

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:

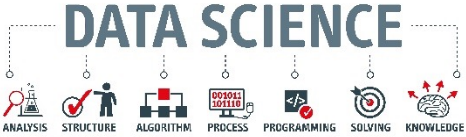
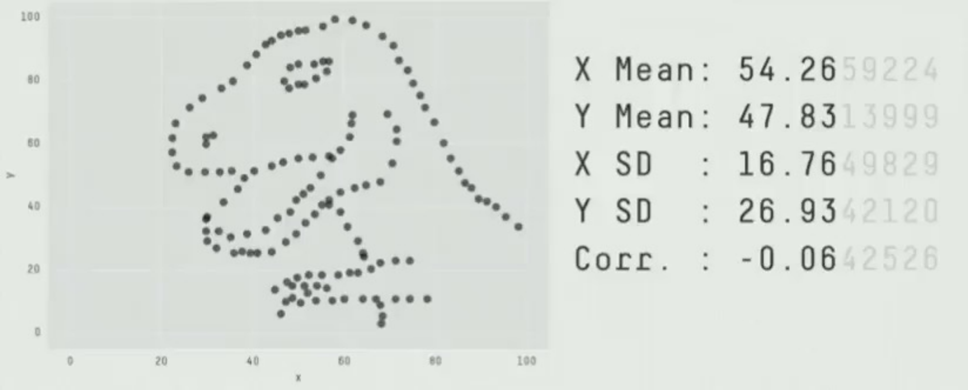


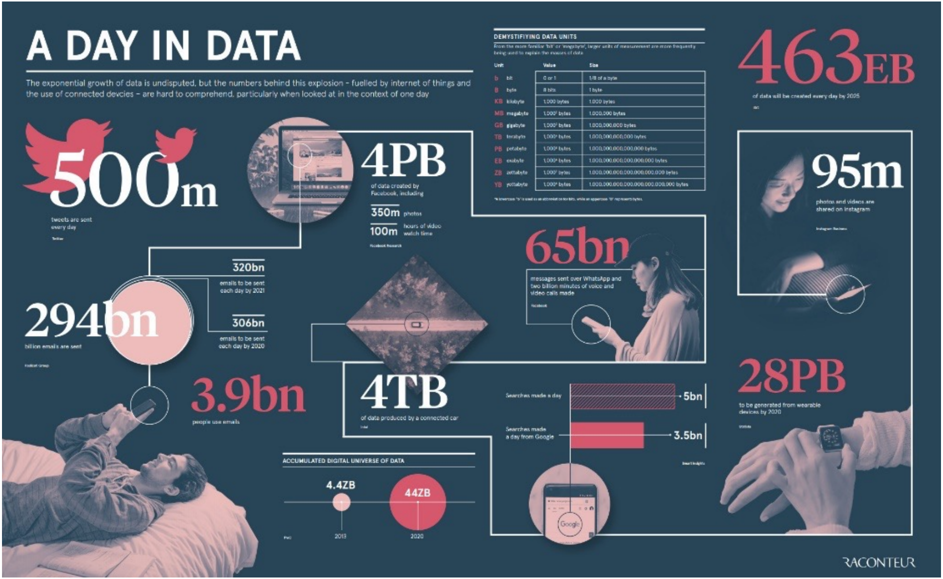
Image from https://www.datanami.com

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



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m tweets are sent every day

Twitter

4PB of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020

3.9bn people use emails

4TB of data produced by a connected car

Intel

65bn messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB of data will be created every day by 2025

IDC

95m photos and videos are shared on Instagram

Instagram Business

Searches made a day 5bn

Searches made a day from Google 3.5bn

Smart Insights

28PB to be generated from wearable devices by 2020

Statista

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB 2013

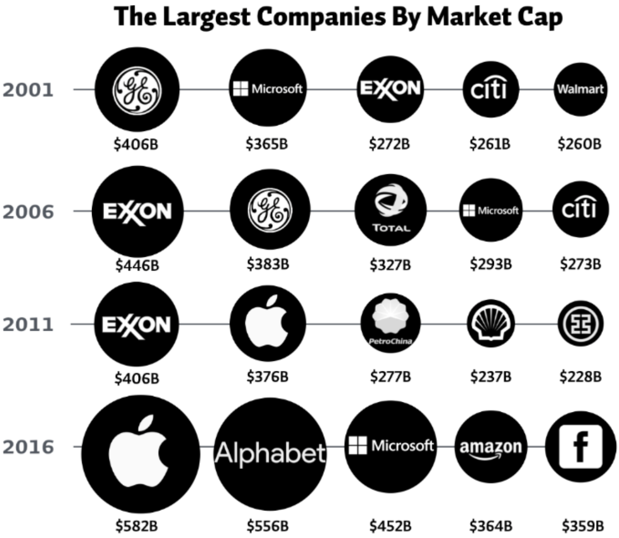
44ZB 2020

PwC

RACONTEUR

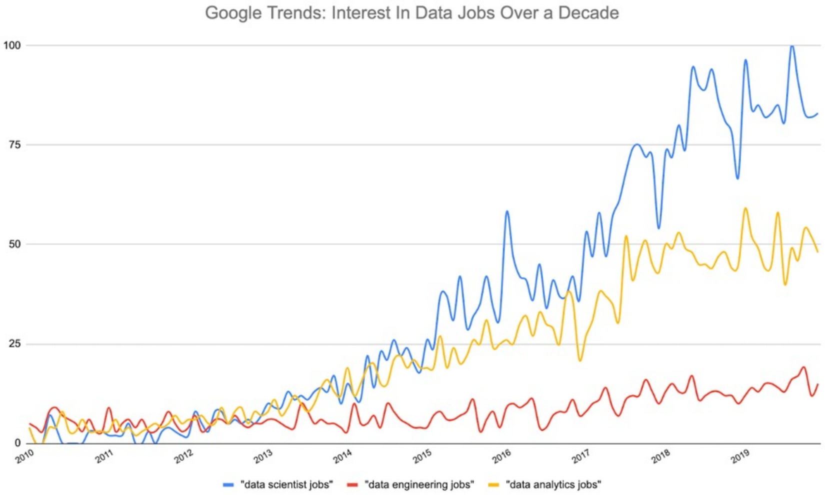
Image from Raconteur https://www.weforum.org

Data is the new Oil



The Largest Companies By Market Cap

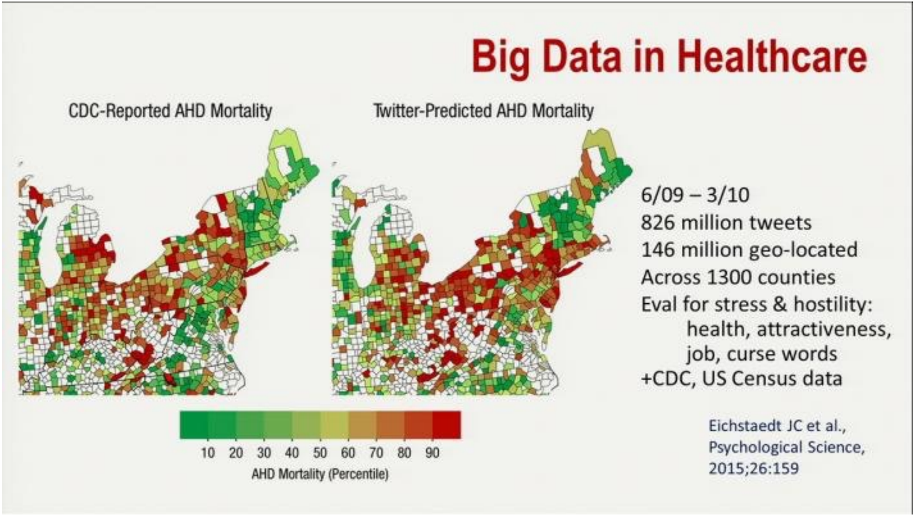
Data Science & Analytics



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

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

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



Big Data in Healthcare

CDC-Reported AHD Mortality
Twitter-Predicted AHD Mortality
10 20 30 40 50 60 70 80 90
AHD Mortality (Percentile)

6/09 – 3/10
826 million tweets
146 million geo-located
Across 1300 counties
Eval for stress & hostility:
health, attractiveness,
job, curse words
+CDC, US Census data

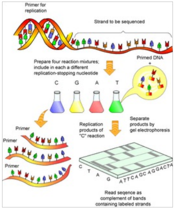
Eichstaedt JC et al.,
Psychological Science,
2015;26:159

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

This Unit

Why is it important for Computer Science?

- Fundamental to many application 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.

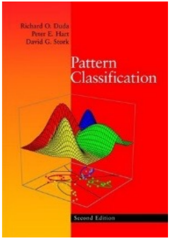

Primer for replication
Strand to be sequenced
Primed DNA
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
C G A T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C T A G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

Ex1. A Fishy Problem

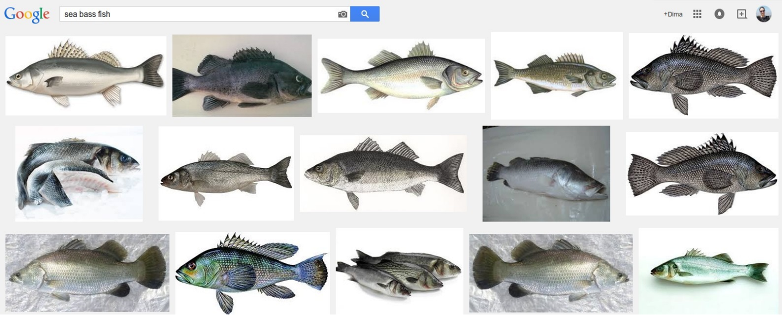
Data: images of fish

Aim: distinguish between sea bass and salmon

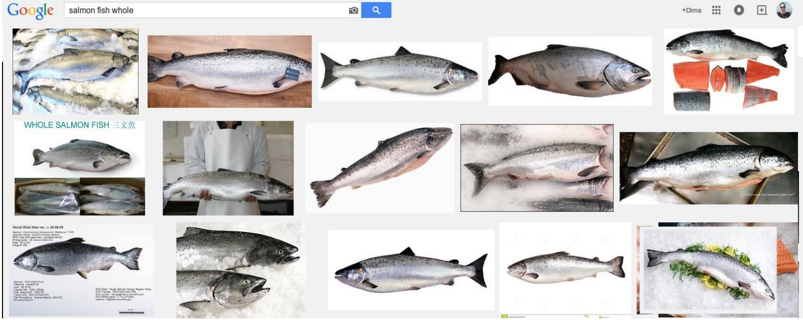
From: Pattern Classification by *Duda, Hart and Stork*, 2nd Edition, Wiley Interscience



Ex1. A Fishy Problem



sea bass fish

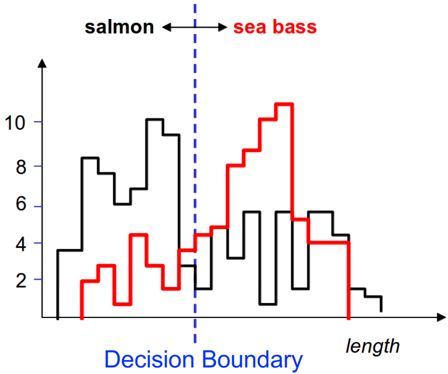


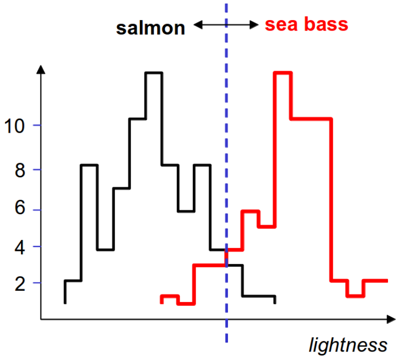
salmon fish whole

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

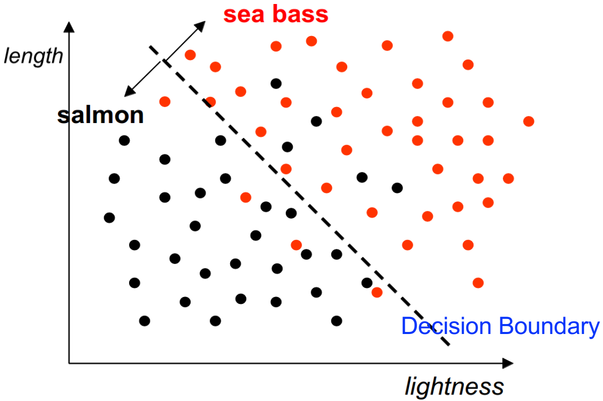




Fishing for a Solution

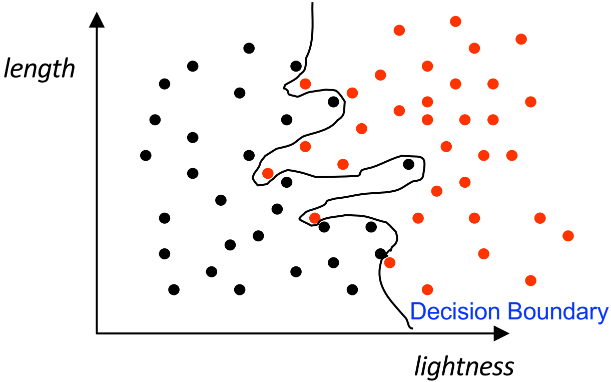
Multiple features could be selected, resulting in a multi-dimensional feature vector.

Scatterplot!



Fishing for a Solution

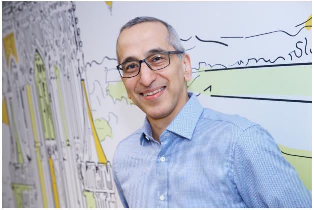
Complex decision model



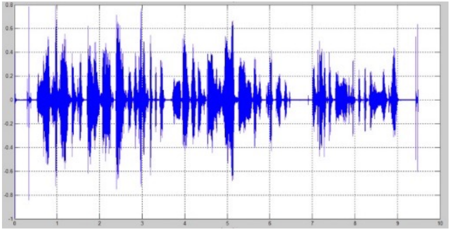
Typical Data Analysis Problem

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] → Laurence Aitchison **[UD]** (~50%)



Ex2. Speech Recognition



Data: Analogue speech signals (time series numerical data)

Aim: Convert audio into text (think Echo/Siri...)

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

Ex3. Spam Filter

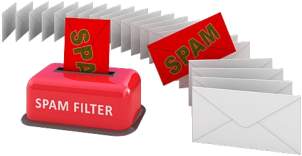


Image from https://www.kdnuggets.com/

Data: Email texts

Aim: Determine whether the email is spam

1. Pre-processing - Normalise words
2. Feature Selection - Presence of words
3. Classification - Naive Bayes classifier

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

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

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

and

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

A new Email is spam if

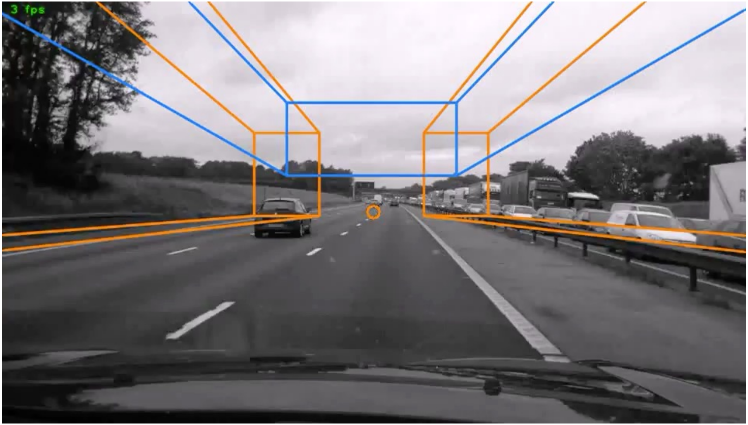
$$\underline{P(email \mid spam)} > \underline{P(email \mid \neg spam)} \qquad (3)$$

Ex4.1 – Towards Autonomous Driving

Data: Video

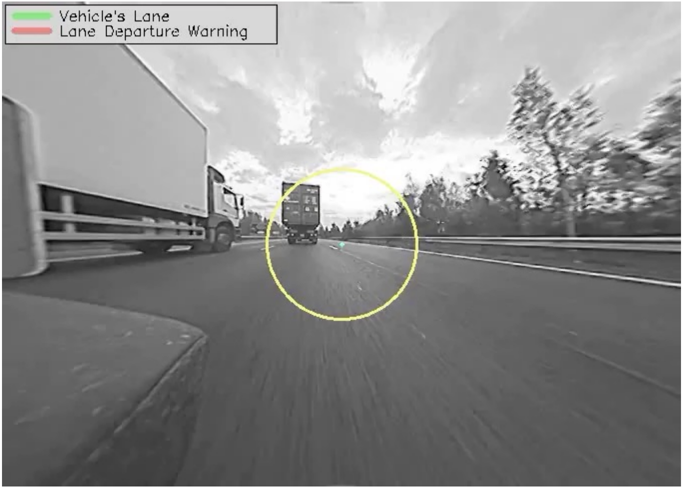
Aim: Determine knowledge from the road or inside the vehicle

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)

COMS20011 Unit

Lectures

- Mondays 14:00 - 14:50 – QUEENS PUGSLEY 1.40
- Thursdays 13:00 - 13:50 – CHEM LT1

Labs

- Thursdays 16:00 - 17:00 [by timetable]: Group 1
- Thursdays 17:00 - 18:00 [by timetable]: Group 2
- Lab Environment [Jupyter + Python]
- TA support in Teams: **grp-COMS20011_2022**
- Labs are essential for learning unit content!

Unit pages : https://github.com/LaurenceA/COMS20011_2022