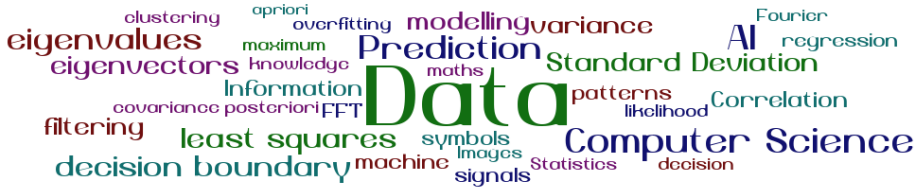


COMS20011 – Data-Driven Computer Science



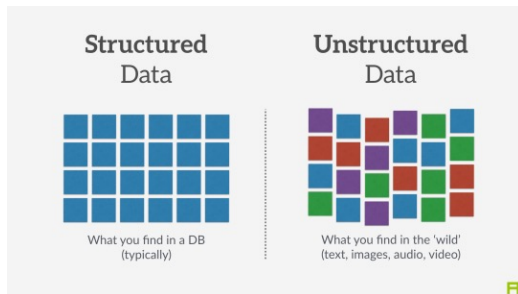
January 2024

Majid Mirmehdi

Lecture MM-01

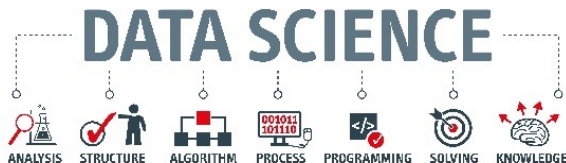
What is Data?

- 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...

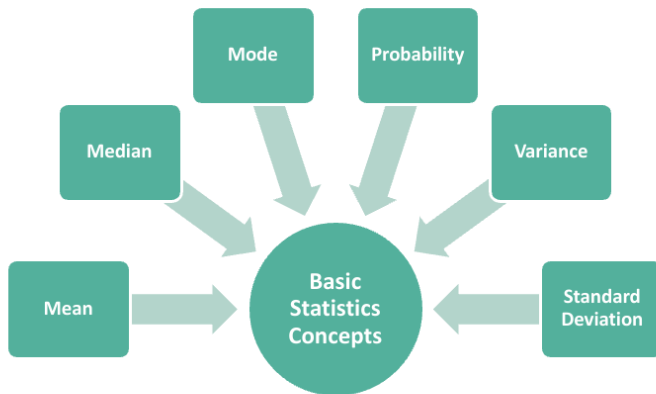


This Unit

- This unit is about doing things with data... *but not*
 - storing, shuffling, searching (Algorithms I & II)
 - sending (Computer Systems)
 - compressing or encrypting (Cryptography)
- This unit is about:
 - extracting knowledge from data
 - generating data and making predictions
 - making decisions based on data
 - Often referred to as:



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

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 devices - are hard to comprehend, particularly when looked at in the context of one day.

500m
tweets are sent every day

294bn
text messages are sent

320bn
emails to be sent each day by 2020

306bn
emails to be sent each day by 2020

3.9bn
people use emails

4PB

of data created by Facebook, including

350m photos
100m hours of video watch time

4TB

of data produced by a connected car

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

2015

2020

DEMYSTIFYING DATA UNITS

From the most familiar bit to the most complex, larger units of measurement are more frequently being used to represent the volume of data.

Unit	Value	Size
bit	0 or 1	1/8 of a byte
byte	8 bits	1 byte
kilobyte	1,024 bytes	1,024 bytes
megabyte	1,024 kilobytes	1,024,000 bytes
gigabyte	1,024 megabytes	1,024,000,000 bytes
terabyte	1,024 gigabytes	1,024,000,000,000 bytes
petabyte	1,024 terabytes	1,024,000,000,000,000 bytes
exabyte	1,024 petabytes	1,024,000,000,000,000,000 bytes
zettabyte	1,024 exabytes	1,024,000,000,000,000,000,000 bytes
yottabyte	1,024 zettabytes	1,024,000,000,000,000,000,000,000 bytes

*Note: 10 is used as a multiplier in binary, while 1,000 is used in decimal.

65bn

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

Searches made a day **5bn**

Searches made a day from Google **3.5bn**

463EB

of data will be created every day by 2025

95m

photos and videos are shared on Instagram

28PB

to be generated from wearable devices by 2020

RACONTEUR

Data is the new Oil

Chart of the Week

THE LARGEST COMPANIES BY MARKET CAP

The oil barons have been replaced by the whiz kids of Silicon Valley



Top 5 Publicly Traded Companies (by Market Cap)



Tech



Other

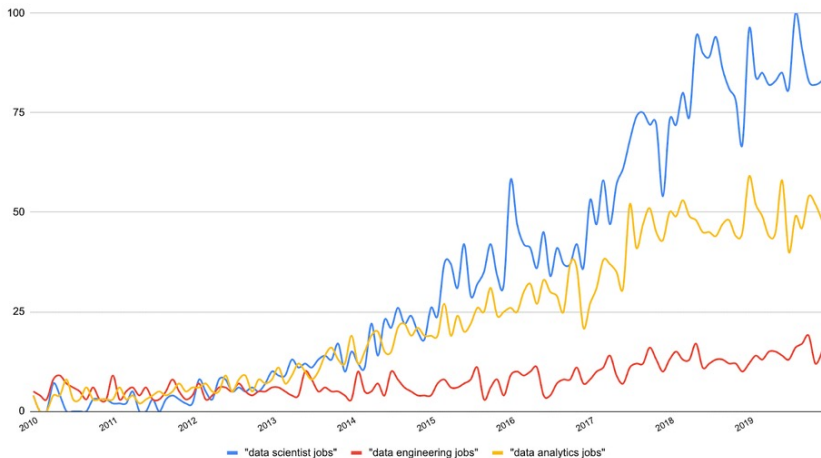


visualcapitalist.com



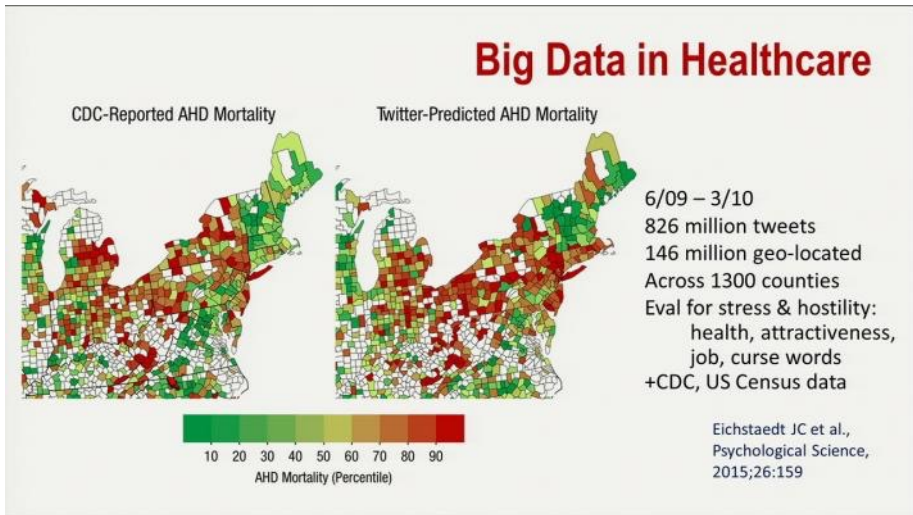
Data Science & Analytics

Google Trends: Interest In Data Jobs Over a Decade



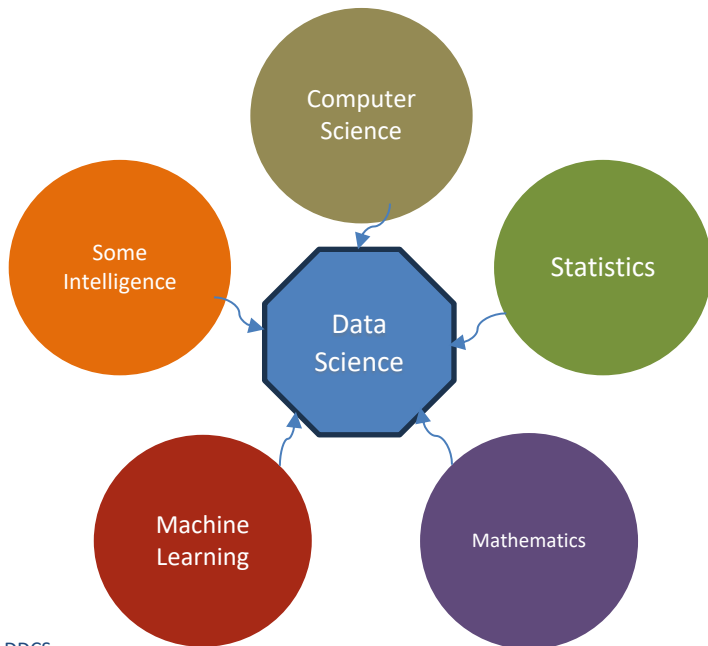
It's not about the data – it's about the science

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



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

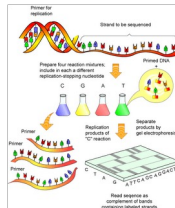
It's not about the data – it's about the science



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=COMS20011>

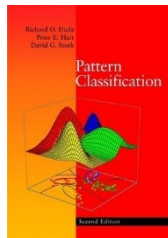
Ex1. A Fish Problem



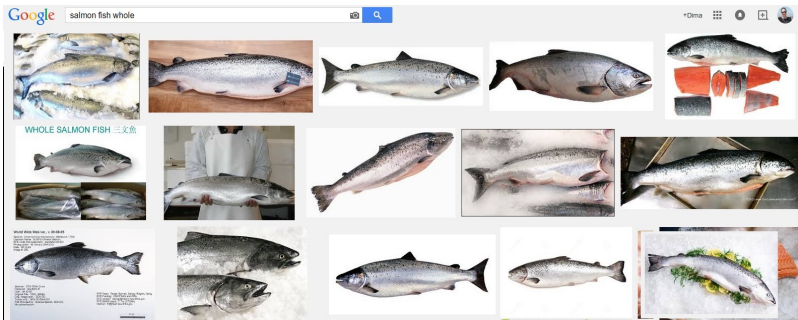
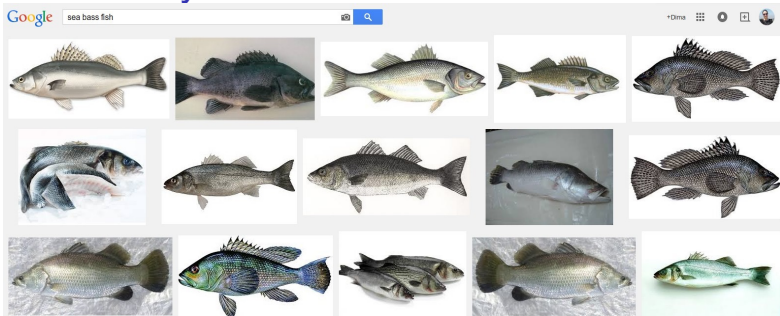
Data: images of fish

Aim: distinguish between sea bass and salmon

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



Ex1. A Fishy Problem

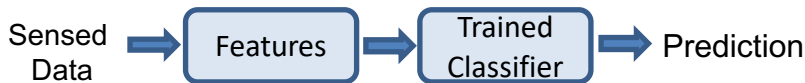


Features

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

Feature extraction and representation allows:

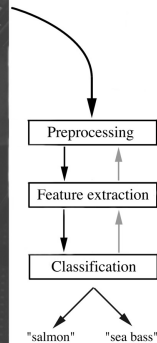
- Data reduction and abstraction
- Focus on relevant, distinguishing parts of data



Fishing for a Solution

Steps:

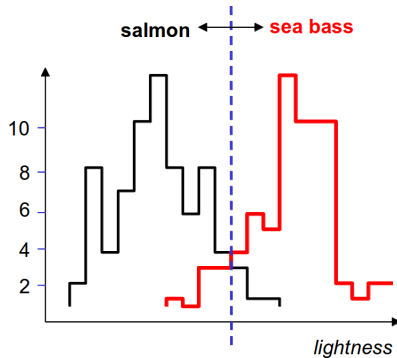
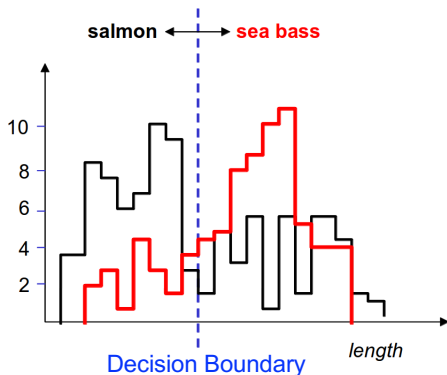
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. Measure length
3. Classification e.g. Find a threshold



Fishing for a Solution

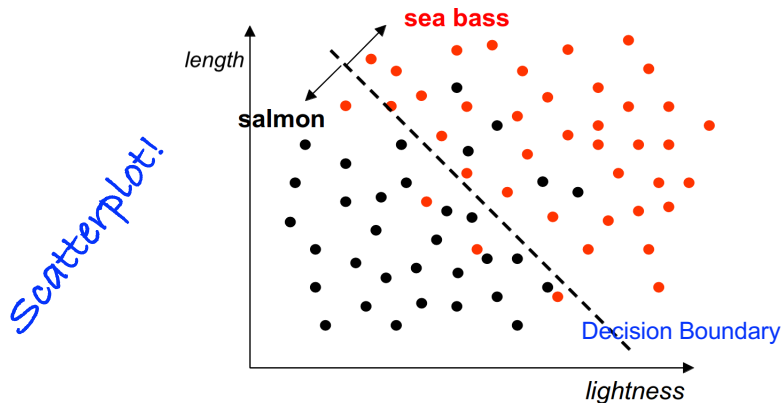
Steps:

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

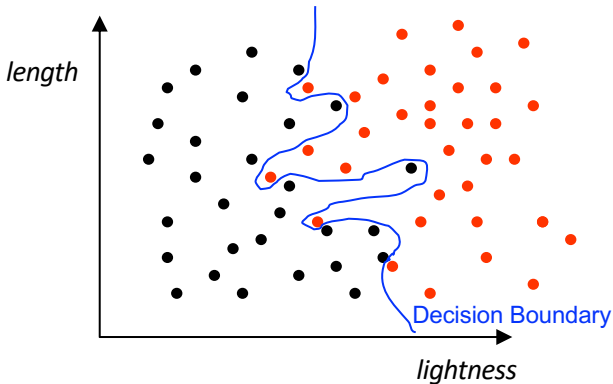
Multiple features could be selected, resulting in a multi-dimensional feature vector.



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

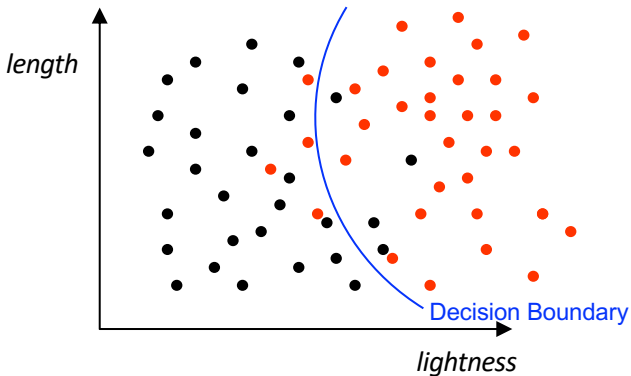
Fishing for a Solution

Complex decision model

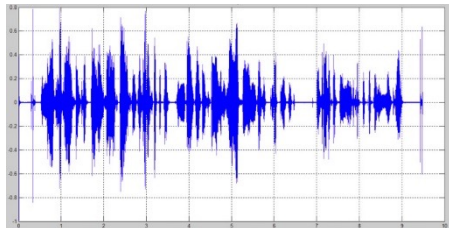


Fishing for a Solution

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**

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

Ex3. Spam Filter

Data: Texts of emails

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 $P(w_i | spam)$ and $P(w_i | \neg spam)$ from frequencies in training data.

For an Email that contains w_1, w_2, \dots, w_n of the subset of words, assume

$$P(email | spam) = P(w_1 | spam)P(w_2 | spam) \dots P(w_n | spam) \quad (1)$$

and

$$P(email | \neg spam) = P(w_1 | \neg spam)P(w_2 | \neg spam) \dots P(w_n | \neg spam) \quad (2)$$

A new Email is spam if

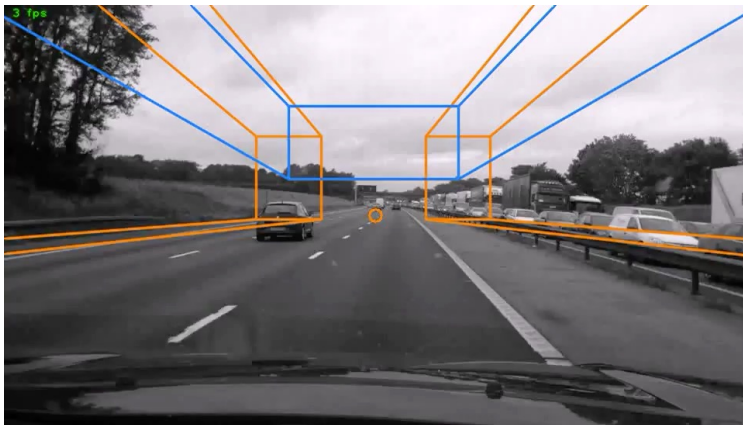
$$\underline{P(email | spam)} > \underline{P(email | \neg spam)} \quad (3)$$

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

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] → Charles Kind (~50%)



Lectures

- Thursdays 14:00 - 14:50 – QUEENS PUGSLEY 1.40
- Fridays 10:00 - 10:50 – QUEENS PUGSLEY 1.40

Unit pages: https://github.com/majidmirmehdi/COMS20011_23-24/

Labs

- Mondays 13:00 - 14:00 [by timetable]: Group 1
- Mondays 14:00 - 15:00 [by timetable]: Group 2
- Lab Environment [Jupyter + Python]
- TA support in Teams: **grp-COMS20011_2023**



Lectures and Labs are both essential for learning unit content!

Next lecture



Analog Signal



Digital Signal

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