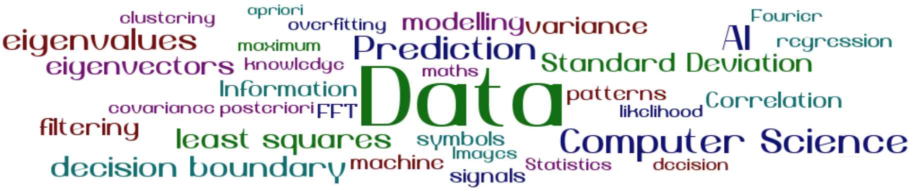
COMS20017 – Algorithms & Data



September 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...

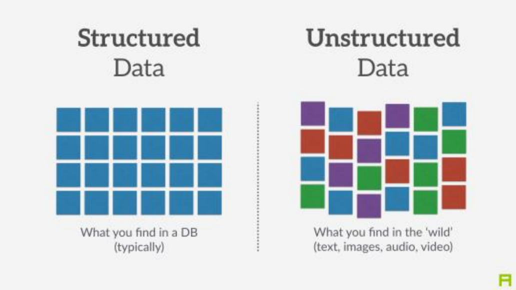

Structured Data
What you find in a DB (typically)
Unstructured Data
What you find in the 'wild' (text, images, audio, video)

Image from Garrett Hollander

This Unit

- This unit is about doing things with data... *but not*
  - storing, shuffling, searching (Algorithms)
  - 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:

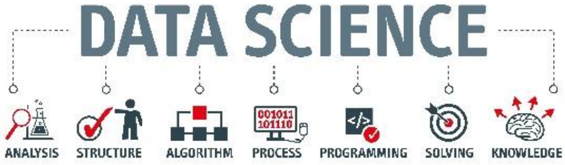


Image from https://www.datanami.com

Basic Statistics Concepts

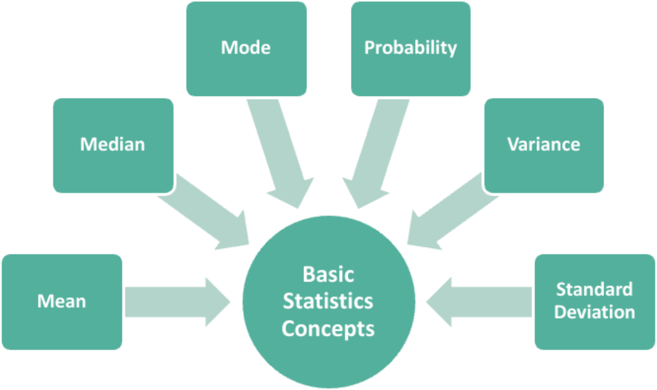
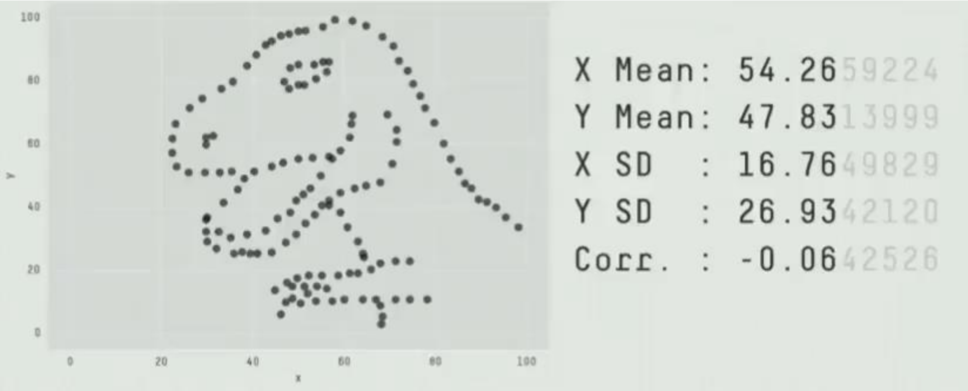


Image from https://www.wallstreetmojo.com/basic-statistics-concepts/

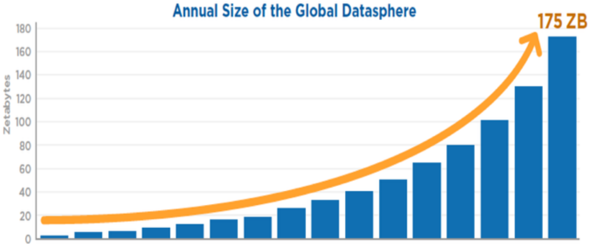
Same Basic Stats, Different Data!



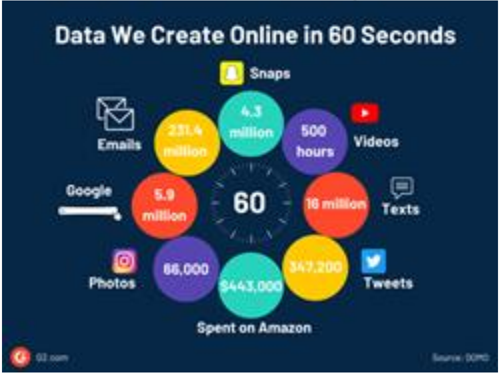
```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

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

Amount of Data!



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018



Data is the new Oil

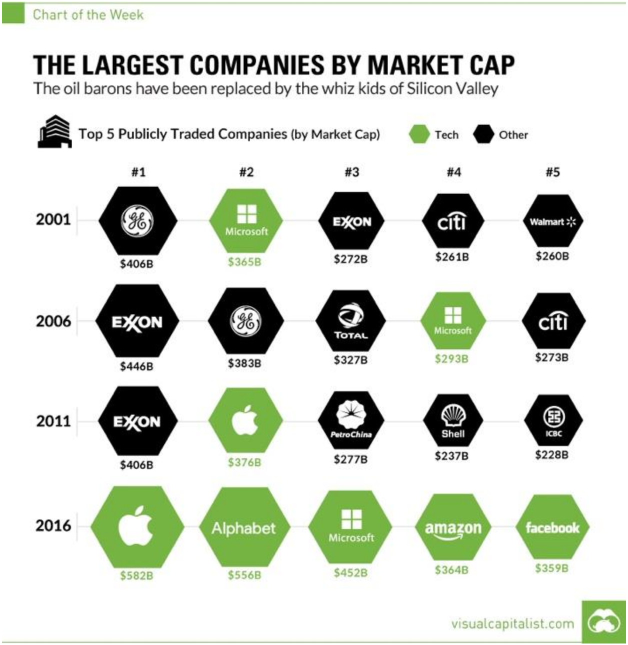


Chart of the Week

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

	#1	#2	#3	#4	#5
2001	GE $406B	Microsoft $365B	EXXON $272B	citi $261B	Walmart $260B
2006	EXXON $446B	GE $383B	TOTAL $327B	Microsoft $293B	citi $273B
2011	EXXON $406B	$376B	PetroChina $277B	Shell $237B	ICBC $228B
2016	$582B	Alphabet $556B	Microsoft $452B	amazon $364B	facebook $359B

visualcapitalist.com

Example Job Positions Involving Data

Data Analyst

+ Data retrieval

+ Spot trends and patterns

+ Visualise and report to others

Data Engineer

+ Design and maintain data management systems

+ Make data accessible to others

Data Scientist

+ Use ML techniques to derive insights

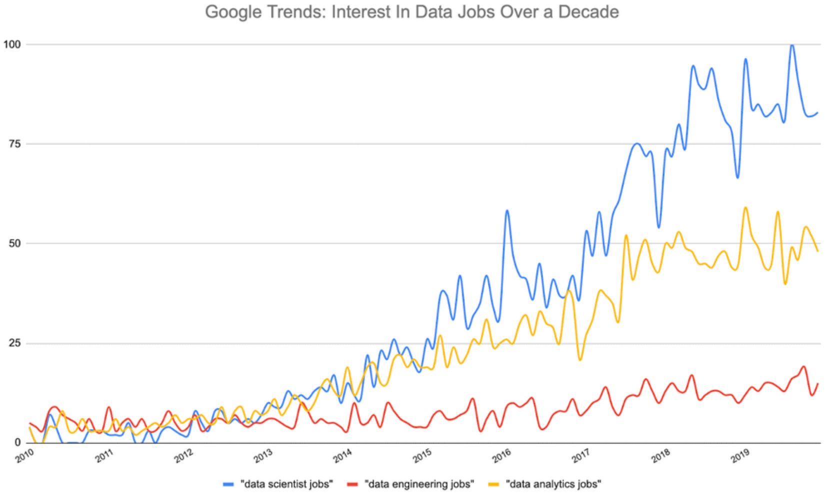
+ Make predictions on products, assets, etc. based on past data

ML Engineer

+ Design and implement ML methods

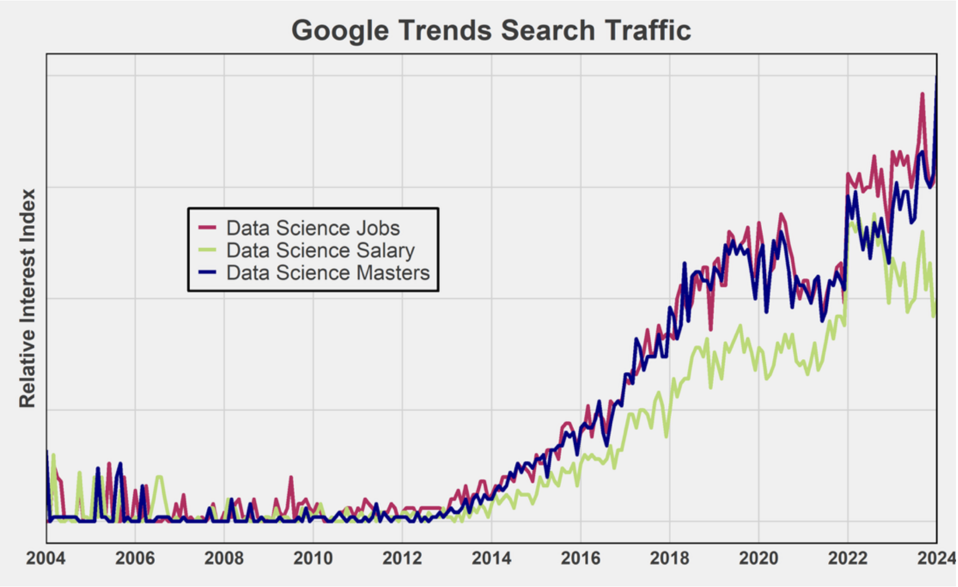
+ Extend existing ML frameworks and libraries

Data Science & Analytics



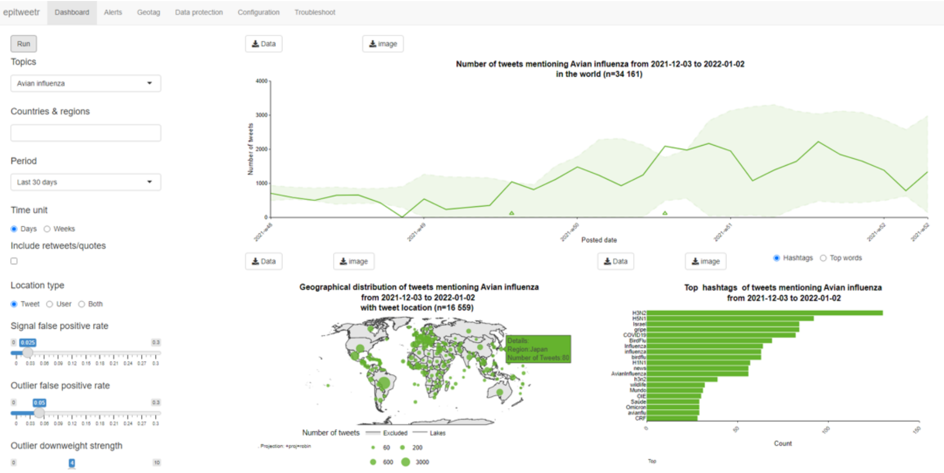
https://onlinedatasciencemasters.virginia.edu/blog/data-science-vs-data-engineering/

Data Science & Analytics



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

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



Topics: Avian influenza

Countries & regions:

Period: Last 30 days

Time unit: ● Days ○ Weeks

Include retweets/quotes: □

Location type: ● Tweet ○ User ○ Both

Signal false positive rate: 0.025 (0 – 0.3)

Outlier false positive rate: 0.05 (0 – 0.3)

Outlier downweight strength: 4 (0 – 10)

Number of tweets mentioning Avian influenza from 2021-12-03 to 2022-01-02
in the world (n=34 161)

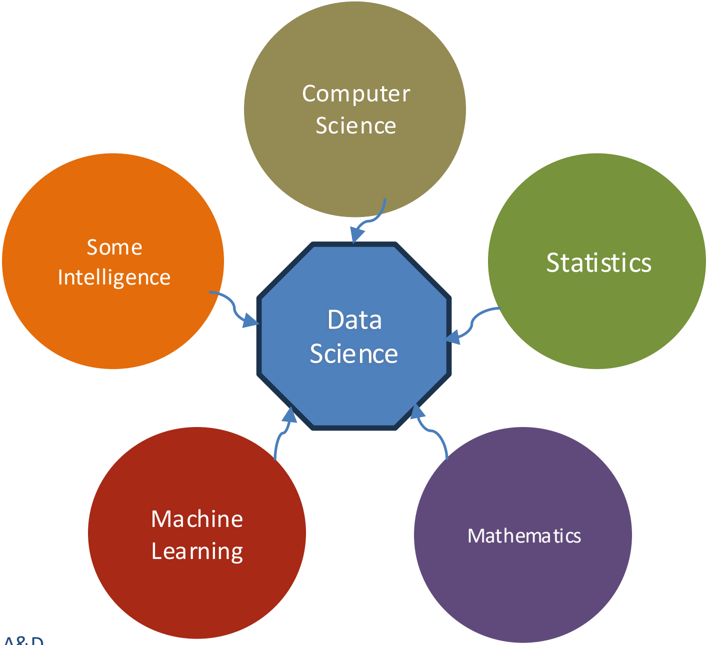
Geographical distribution of tweets mentioning Avian influenza
from 2021-12-03 to 2022-01-02
with tweet location (n=16 559)

● Hashtags ○ Top words

Top hashtags of tweets mentioning Avian influenza
from 2021-12-03 to 2022-01-02

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

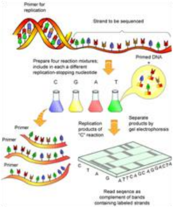
It’s not about the data – it’s about the science



This Unit

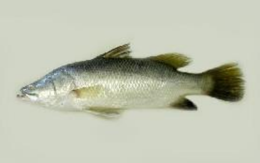
Why is it important for Computer Science?

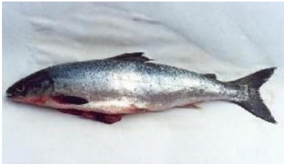
- Fundamental to many related areas:
  - Artificial Intelligence, Machine Learning, Deep Learning
  - Image Processing and Pattern Recognition
  - Graphics, Animation and Virtual Reality
  - Computer Vision and Robotics
  - Speech and Audio Processing
  - With growing applications in: neuroscience, literature,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


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 sequence as complement of bands containing labeled strands

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

Ex1. A Fish Problem

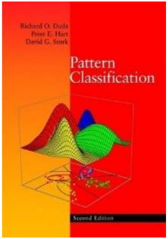




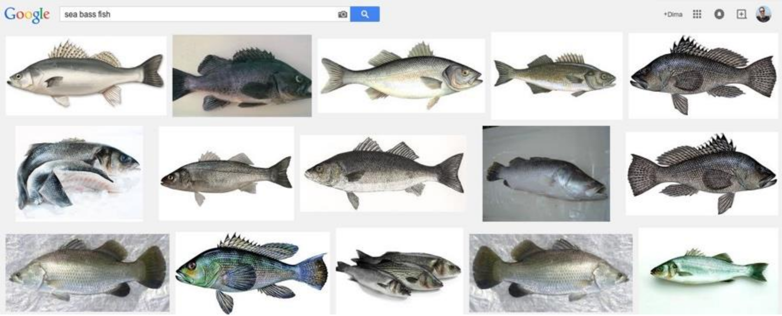
Data: images of fish

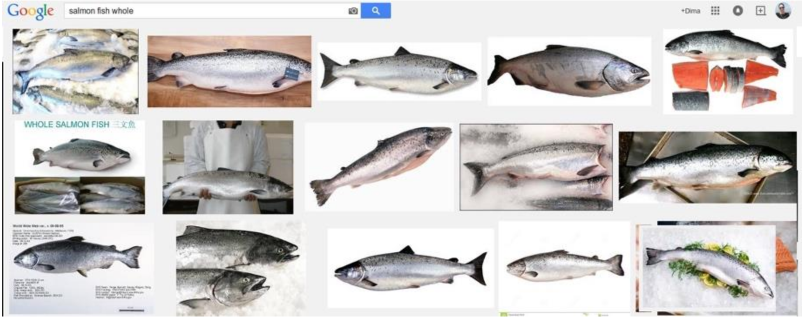
Aim: distinguish between sea bass and salmon

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



Ex1. A Fishy Problem


sea bass fish

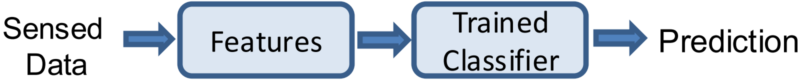

salmon fish whole

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


Sensed Data

Features

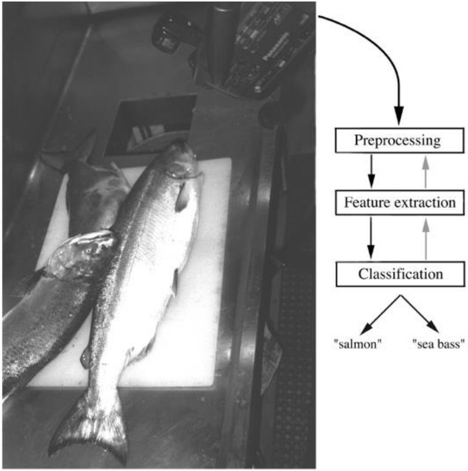
Trained Classifier

Prediction

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



Preprocessing

Feature extraction

Classification

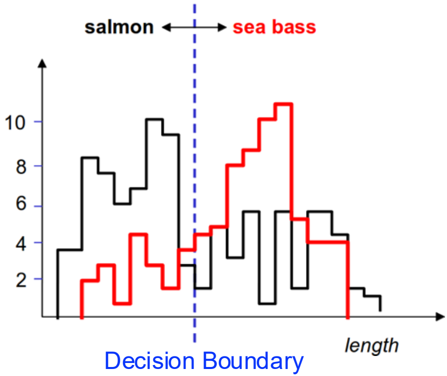
"salmon"

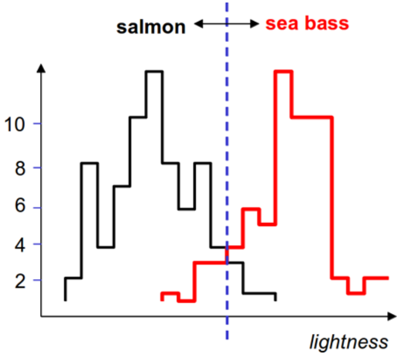
"sea bass"

Fishing for a Solution

Steps:

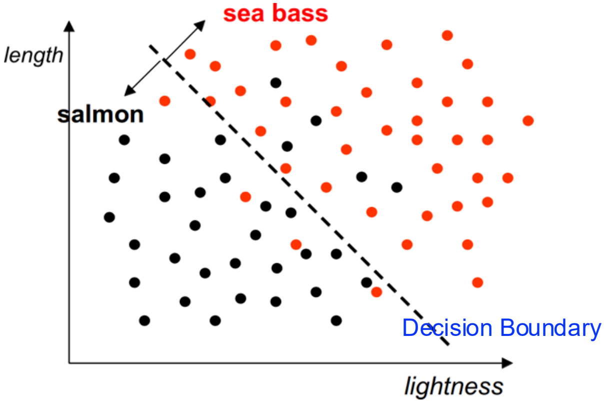
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. Measure length or lightness
3. Classification e.g. Find a threshold





Fishing for a Solution

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

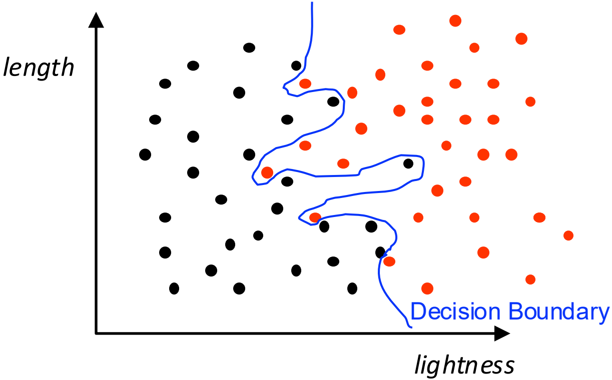


Scatterplot!

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

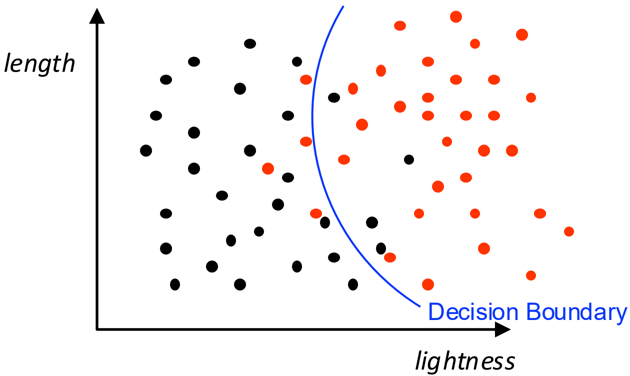
Fishing for a Solution

Complex decision model

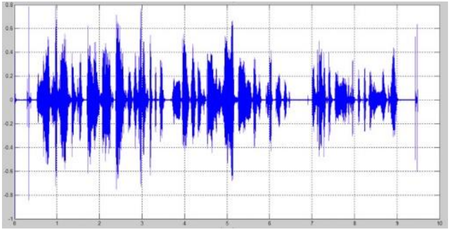


Fishing for a Solution

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

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

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

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

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

and

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

A new Email is spam if

$$P(email\,|\,spam) > P(email\,|\,\neg spam) \qquad (3)$$

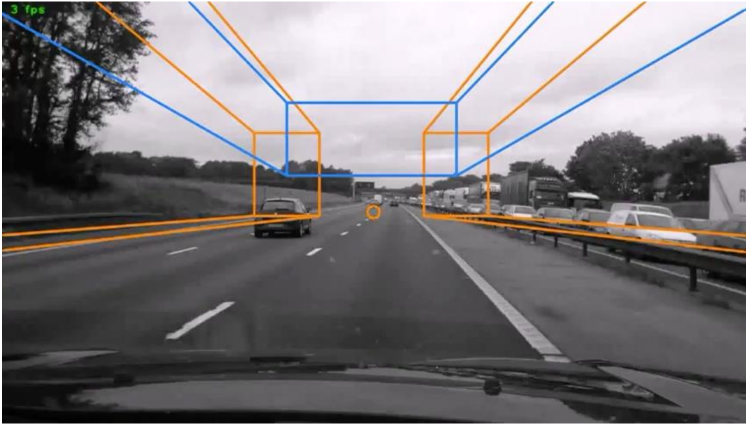
Image from https://www.kdnuggets.com/

Ex4.1 – Towards Autonomous Driving

Data: Video

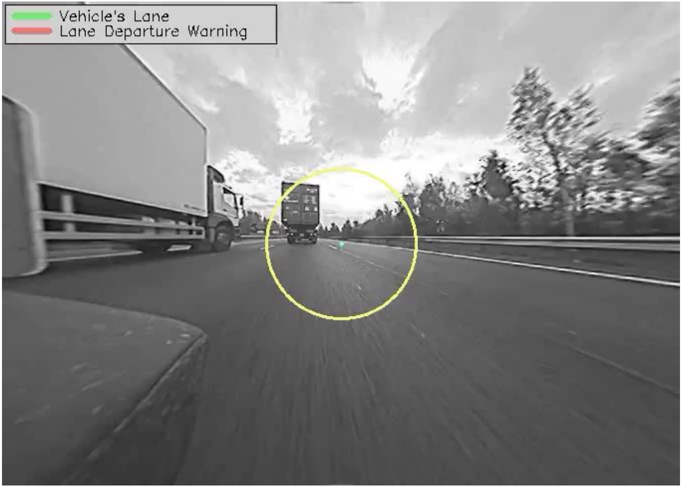
Aim: Determine knowledge from the road or inside the vehicle

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%)

Parts 1 & 3 – supported with Problem Sheets
Part 2 – supported with Problem Sheets and Labs

COMS20017 - Data

Lectures

Mondays 4pm in PHYS BLDG G42 POWELL

Thursdays 2pm in QUEENS BLDG 1.40 PUGSLEY

Unit pages: https://github.com/majidmirmehdi/COMS20017_DATA_25-26

Labs

Fridays 11:00 - 12:00 [by timetable]: Group 1

Fridays 12:00 - 13:00 [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!

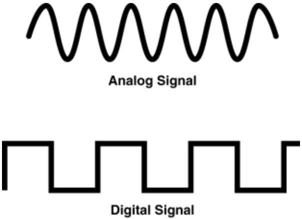
Very Welcome to Ask Questions…

You should **use the unit's 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



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