# COMS21202: An Introduction to Doing Things with Data

Rui Ponte Costa
rui.costa@bristol.ac.uk

University of Bristol, Department of Computer Science
Bristol BS8 1UB, UK

January 21, 2020

# What is Data?

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)
  - Textual (emails, Web pages, medical records, ...)

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)
  - Textual (emails, Web pages, medical records, ...)
  - Visual (images, video, graphics, animations)

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)
  - Textual (emails, Web pages, medical records, ...)
  - Visual (images, video, graphics, animations)
  - Auditory (speech, audio)

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)
  - Textual (emails, Web pages, medical records, ...)
  - Visual (images, video, graphics, animations)
  - Auditory (speech, audio)
  - Signals (GPS signals, ...)

# What is Data?

- Data: Symbols, Patterns and Signals
  - Numeric (measurements, finances, ...)
  - Textual (emails, Web pages, medical records, ...)
  - Visual (images, video, graphics, animations)
  - Auditory (speech, audio)
  - Signals (GPS signals, ...)
  - Other... DNA sequence number

# This Unit

- This unit is about doing things with data... but not

# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching

# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)

# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)
  - sending

# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)
  - sending (Networking)

# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)
  - sending (Networking)
  - compressing or encrypting

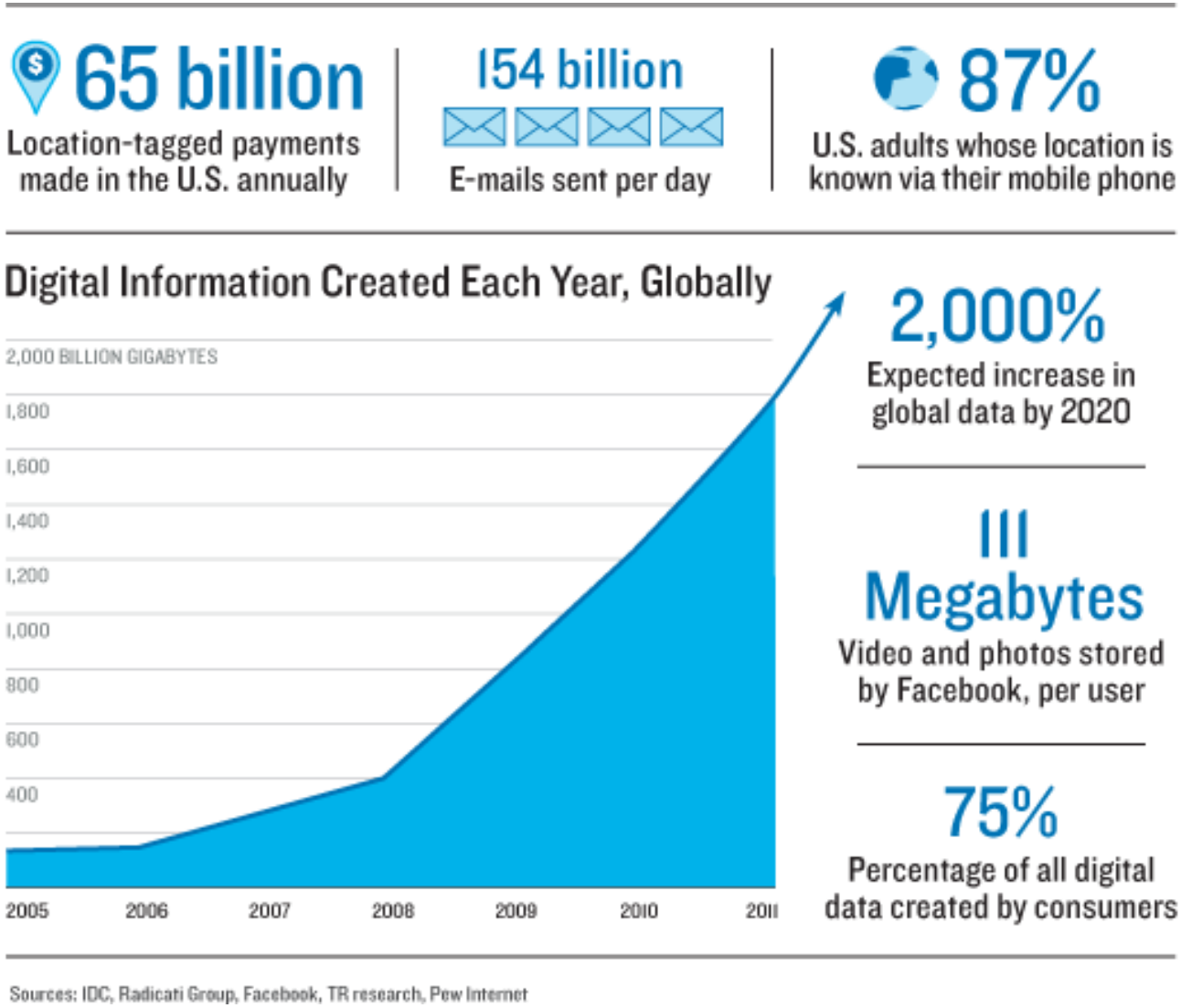
# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)
  - sending (Networking)
  - compressing or encrypting (Crypto I and Crypto II)

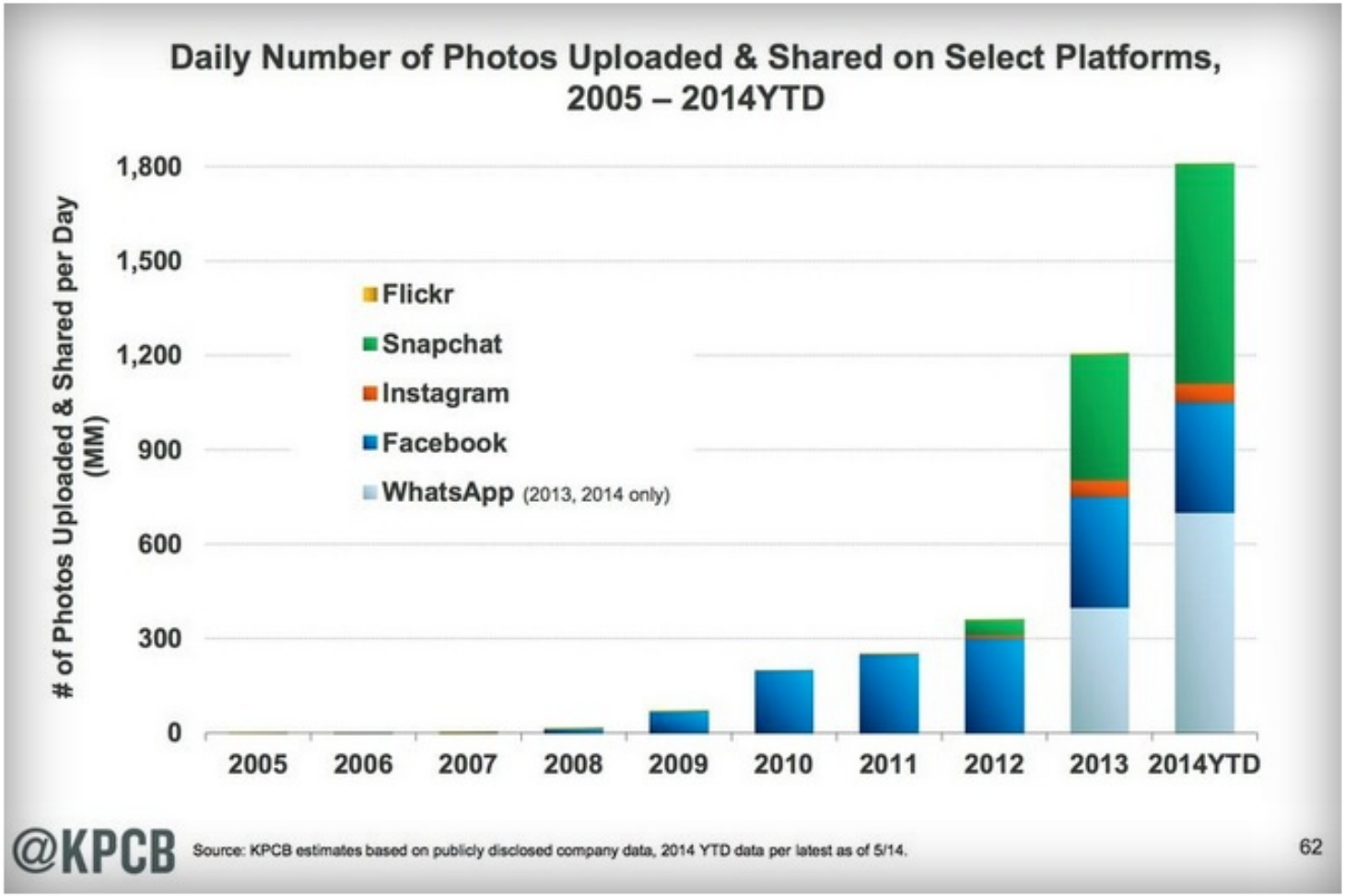
# This Unit

- This unit is about doing things with data... but not
  - storing, shuffling, searching (Data Structures and Algorithms)
  - sending (Networking)
  - compressing or encrypting (Crypto I and Crypto II)
- This unit is about:
  - extracting knowledge from data
  - generating data and making predictions
  - making decisions based on data
  - ... often referred to as: Data Science

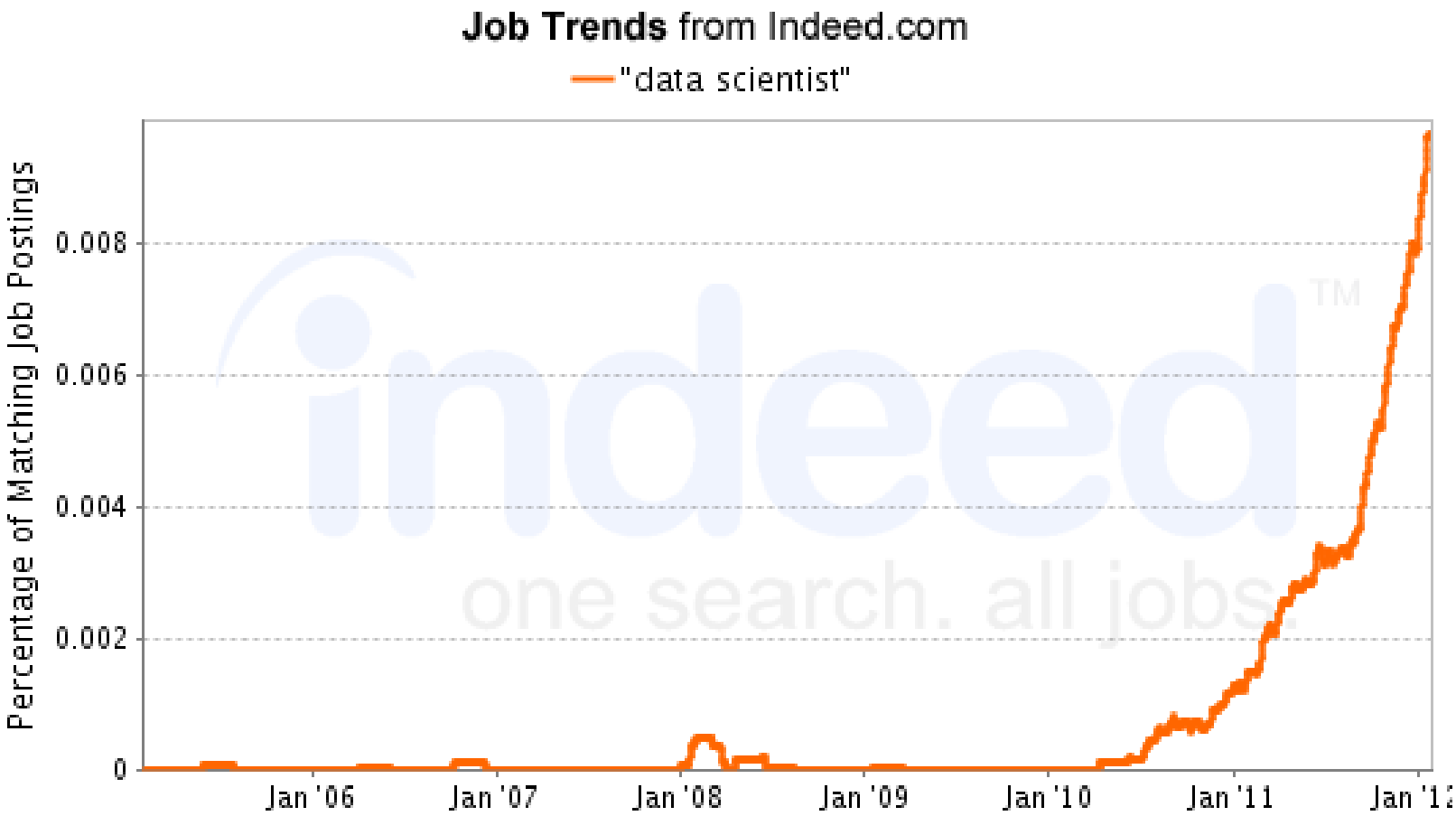
# This Unit


65 billion
Location-tagged payments made in the U.S. annually
154 billion
E-mails sent per day
87%
U.S. adults whose location is known via their mobile phone
Digital Information Created Each Year, Globally
2,000 BILLION GIGABYTES
1,800
1,600
1,400
1,200
1,000
800
600
400
2005
2006
2007
2008
2009
2010
2011
2,000%
Expected increase in global data by 2020
III Megabytes
Video and photos stored by Facebook, per user
75%
Percentage of all digital data created by consumers
Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet

# This Unit


Daily Number of Photos Uploaded & Shared on Select Platforms, 2005 – 2014YTD
# of Photos Uploaded & Shared per Day (MM)
1,800
1,500
1,200
900
600
300
0
Flickr
Snapchat
Instagram
Facebook
WhatsApp (2013, 2014 only)
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014YTD
@KPCB
Source: KPCB estimates based on publicly disclosed company data, 2014 YTD data per latest as of 5/14.
62

# This Unit


Job Trends from Indeed.com
"data scientist"
Percentage of Matching Job Postings
0.008
0.006
0.004
0.002
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
indeed™
one search. all jobs.

# This Unit


Job Trends from Indeed.com
"Data Scientist"
Percentage of Matching Job Postings
0.04
0.02
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
Jan '14
Jan '15
indeed
one search. all jobs.

# This Unit is an introduction to.....




sources:dmnews.com, infinitdatum.com, code-n.org

# But it's not about the data, but the science

# But it's not about the data, but the science

## 'Like' curly fries on Facebook? Then you're clever

'Like' curly fries? Then there's a good chance you've got a high IQ, according to a Cambridge University project to discover what we unwittingly reveal about ourselves on Facebook.

311 50 0 4 365 Email

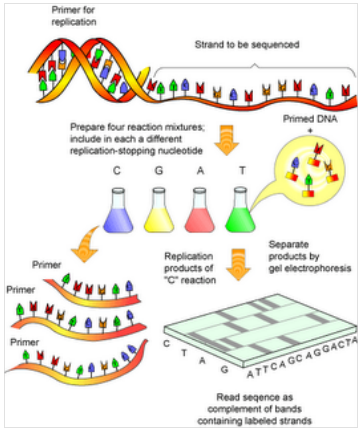
Curly Fries: Researchers at Cambridge's Psychometric Centre have joined forces with Microsoft to analyse more than nine million 'likes' on Facebook Photo: ALAMY

telegraph.co.uk

# This Unit

Why is it important for Computer Science?

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

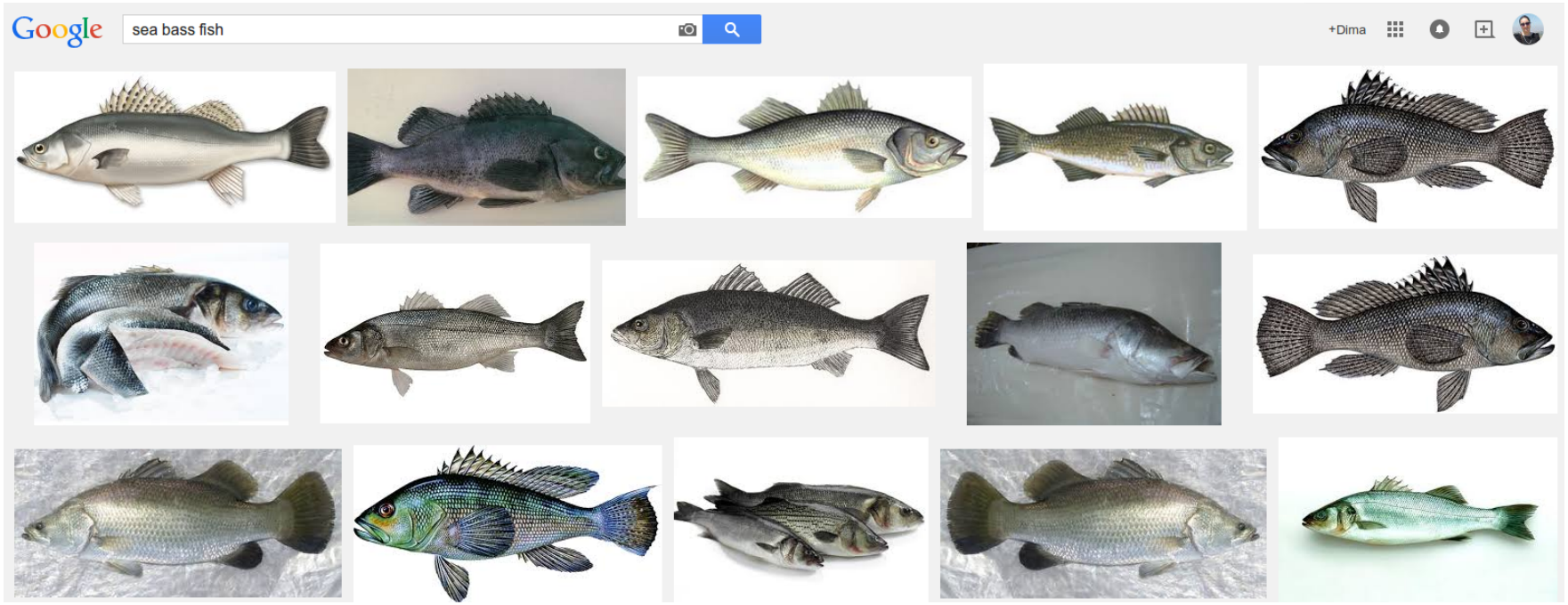
# Ex1. A Fishy Problem

From: Pattern Classification by Duda, Hart and Stork

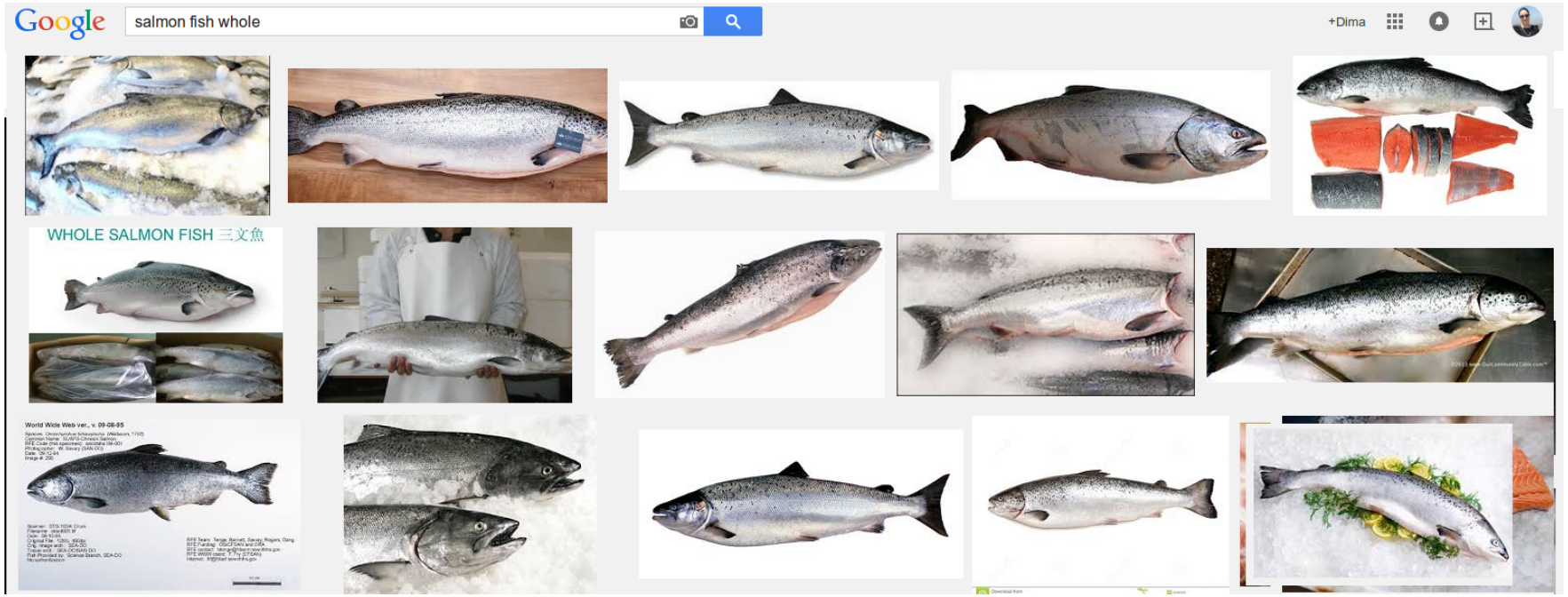
**Data:** images of fish

**Aim:** distinguish between sea bass and salmon

# Ex1. A Fishy Problem



# Ex1. A Fishy Problem



# Ex1. A Fishy Problem

Steps:

- Rui Ponte Costa
  - Majid Mirmehdi
- Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing
   - » Rui Ponte Costa
     - » Majid Mirmehdi
   - » Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing
2. Feature Selection

» Rui Ponte Costa

» Majid Mirmehdi

» Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing » Rui Ponte Costa
2. Feature Selection » Majid Mirmehdi
3. Classification » Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1] » Rui Ponte Costa
2. Feature Selection » Majid Mirmehdi
3. Classification » Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1] » Rui Ponte Costa
2. Feature Selection » Majid Mirmehdi
3. Classification [Unit - Part 2] » Laurence Aitchison **[unit director]**

# Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1] » Rui Ponte Costa
2. Feature Selection [Unit - Part 3] » Majid Mirmehdi
3. Classification [Unit - Part 2] » Laurence Aitchison **[unit director]**



# Fishing for a Solution

E.g.:

1. Pre-processing
2. Feature Selection
3. Classification

# Fishing for a Solution

E.g.:

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

# Fishing for a Solution

E.g.:

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

# Fishing for a Solution

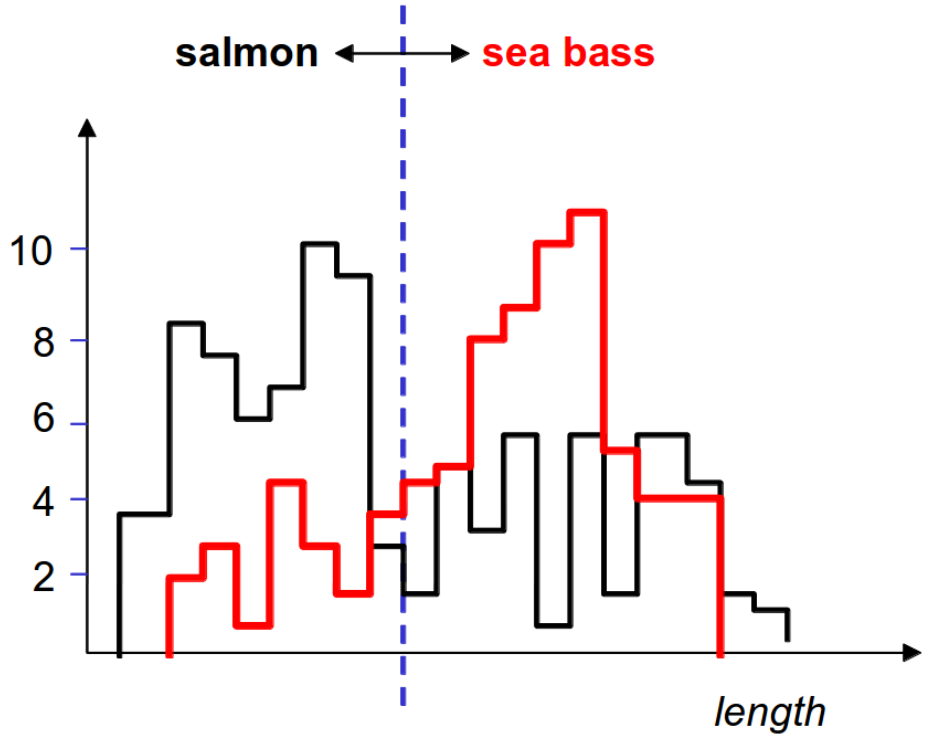
E.g.:

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

E.g.:

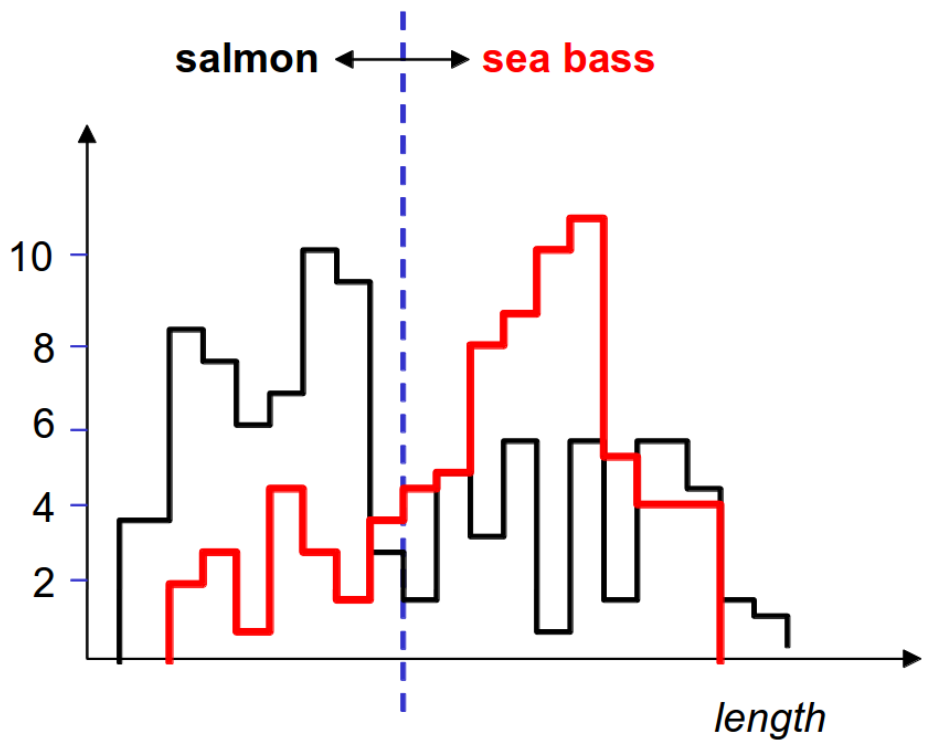
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. measure length
3. Classification e.g. find a threshold

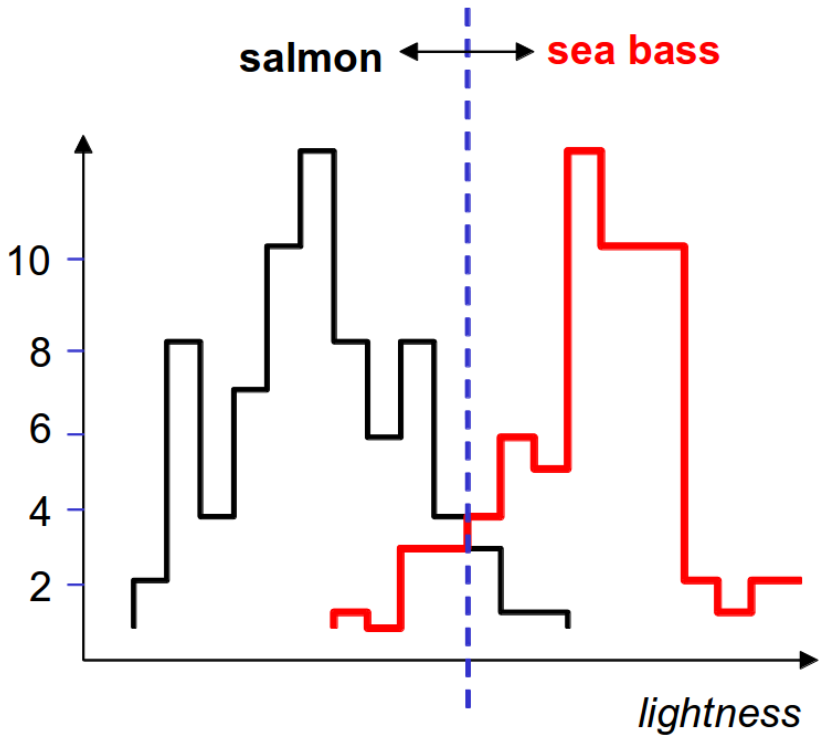


# Fishing for a Solution

E.g.:

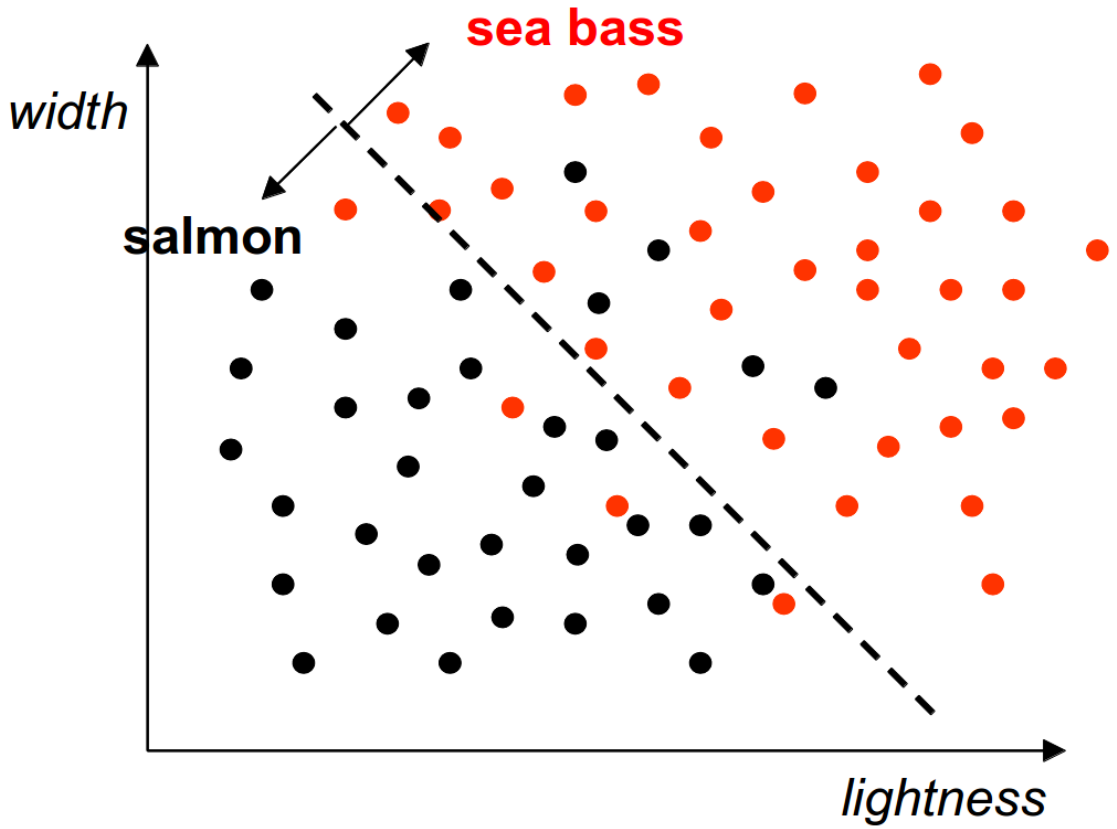
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. measure length or brightness
3. Classification e.g. find a threshold



# Fishing for a Solution

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



# Ex2. Speech Recognition

**Data:** analogue speech signals (time series numerical data)

**Aim:** convert audio into text

Steps:

1. Pre-processing
2. Feature Selection
3. Inference

# Ex2. Speech Recognition

**Data:** analogue speech signals (time series numerical data)

**Aim:** convert audio into text

Steps:

1. Pre-processing Digitisation
2. Feature Selection
3. Inference

# Ex2. Speech Recognition

**Data:** analogue speech signals (time series numerical data)

**Aim:** convert audio into text

Steps:

1. Pre-processing Digitisation
2. Feature Selection Wave amplitude
3. Inference

# Ex2. Speech Recognition

**Data:** analogue speech signals (time series numerical data)

**Aim:** convert audio into text

Steps:

1. Pre-processing Digitisation
2. Feature Selection Wave amplitude
3. Inference Hidden Markov Models and the Viterbi algorithm

# Ex3. Spam Filter

**Data:** email texts (text data)

**Aim:** determine whether the email is spam

Steps:

1. Pre-processing
2. Feature Selection
3. Classification

# Ex3. Spam Filter

**Data:** email texts (text data)

**Aim:** determine whether the email is spam

Steps:

1. Pre-processing Normalise words
2. Feature Selection
3. Classification

# Ex3. Spam Filter

**Data:** email texts (text data)

**Aim:** determine whether the email is spam

Steps:

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

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

# Ex3. Spam Filter

**Data:** email texts (text data)

**Aim:** determine whether the email is spam

Steps:

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 frequences in training data.

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

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

and

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

Email is spam if

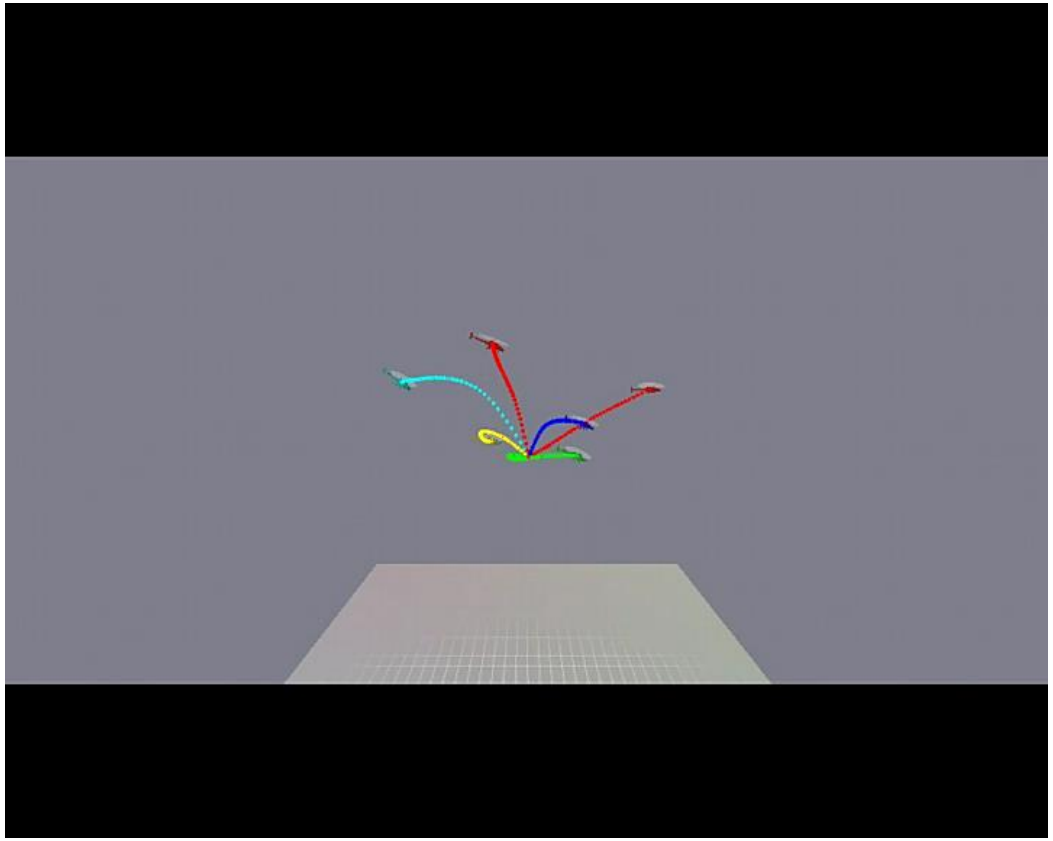
$$P(email|spam) > P(email|\neg spam) \quad (3)$$

# Ex4. Automatic Helicopter

# Ex4. Automatic Helicopter

**Data:** expert demonstration

**Aim:** fly an autonomous helicopter



# Ex4. Automatic Helicopter

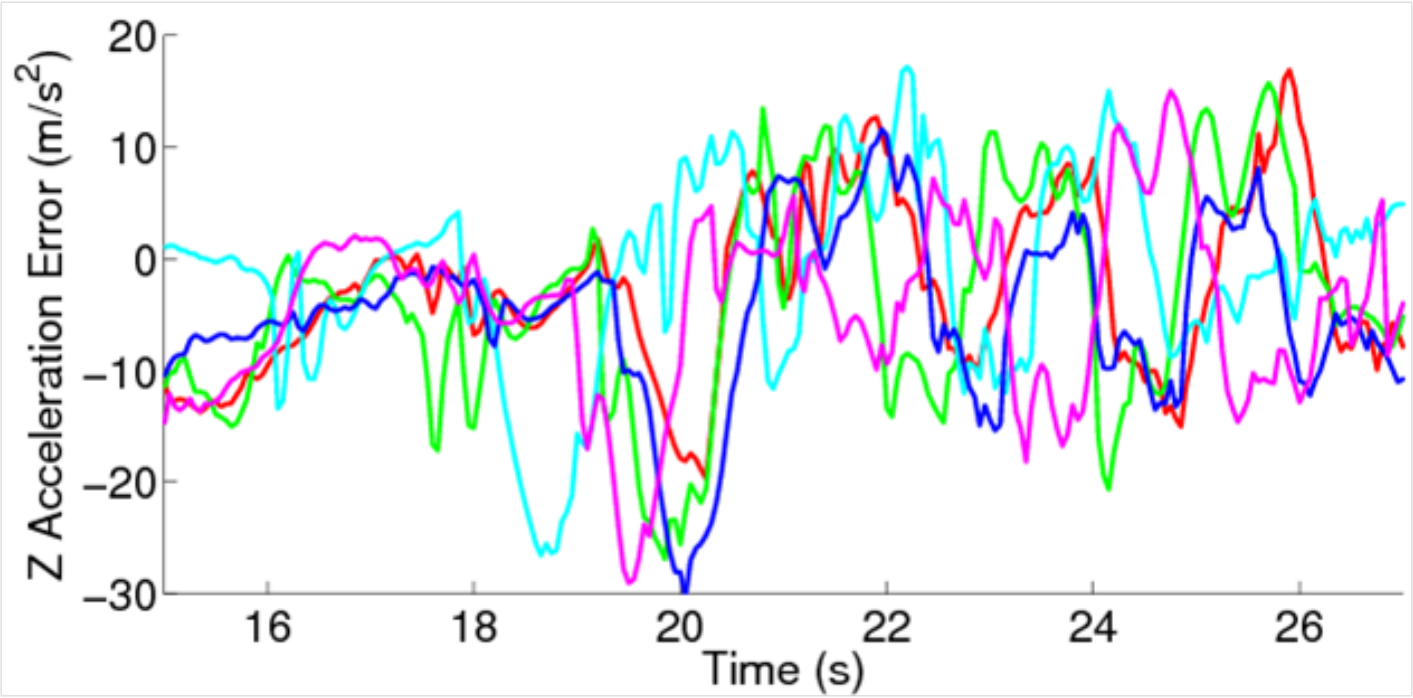
Steps:

1. Pre-processing
2. Feature Selection
3. Model Building

# Ex4. Automatic Helicopter

Steps:

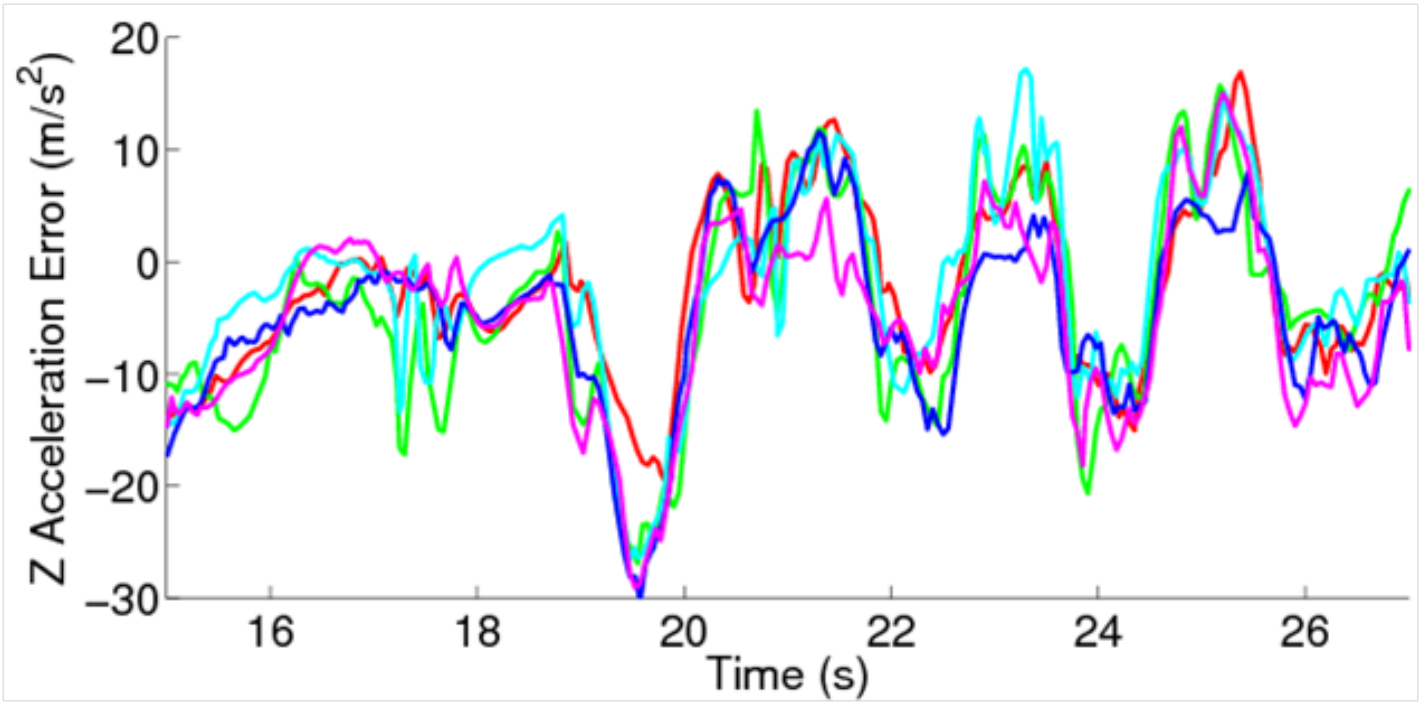
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



# Ex4. Automatic Helicopter

Steps:

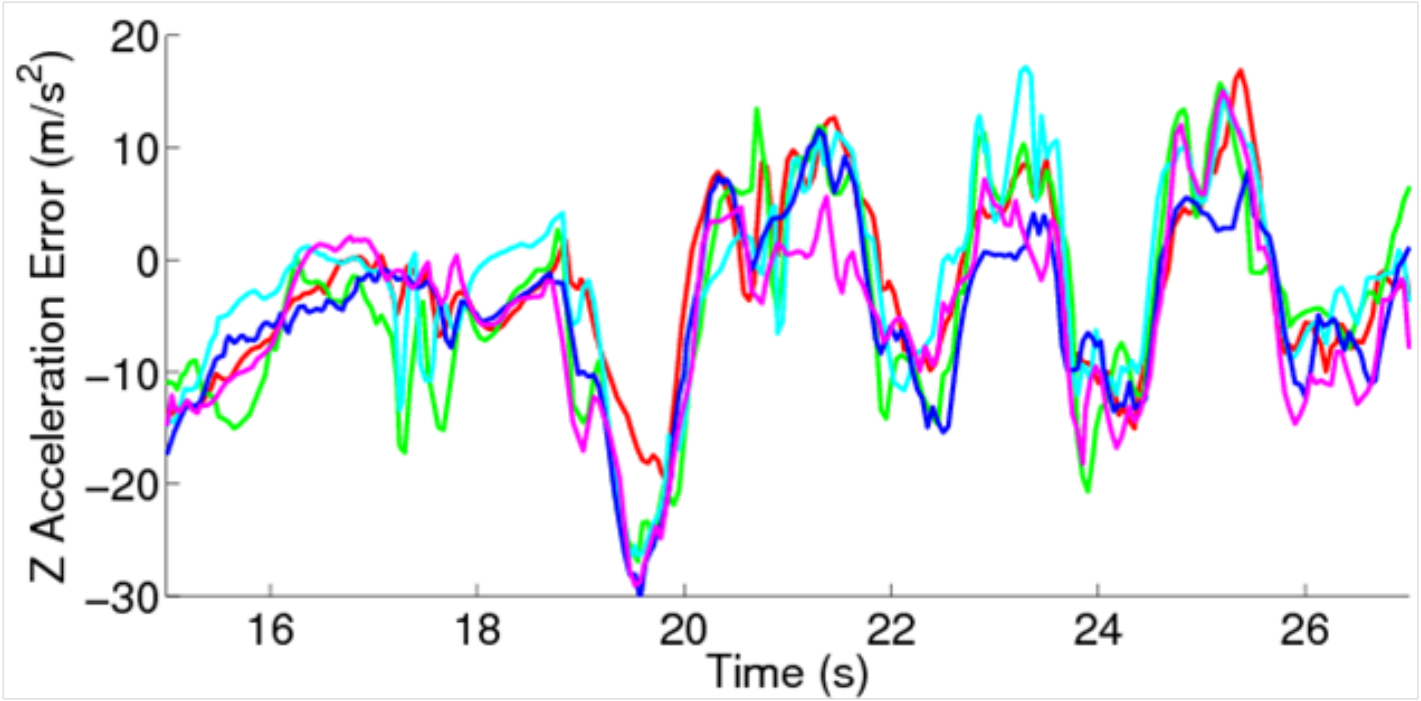
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



# Ex4. Automatic Helicopter

Steps:

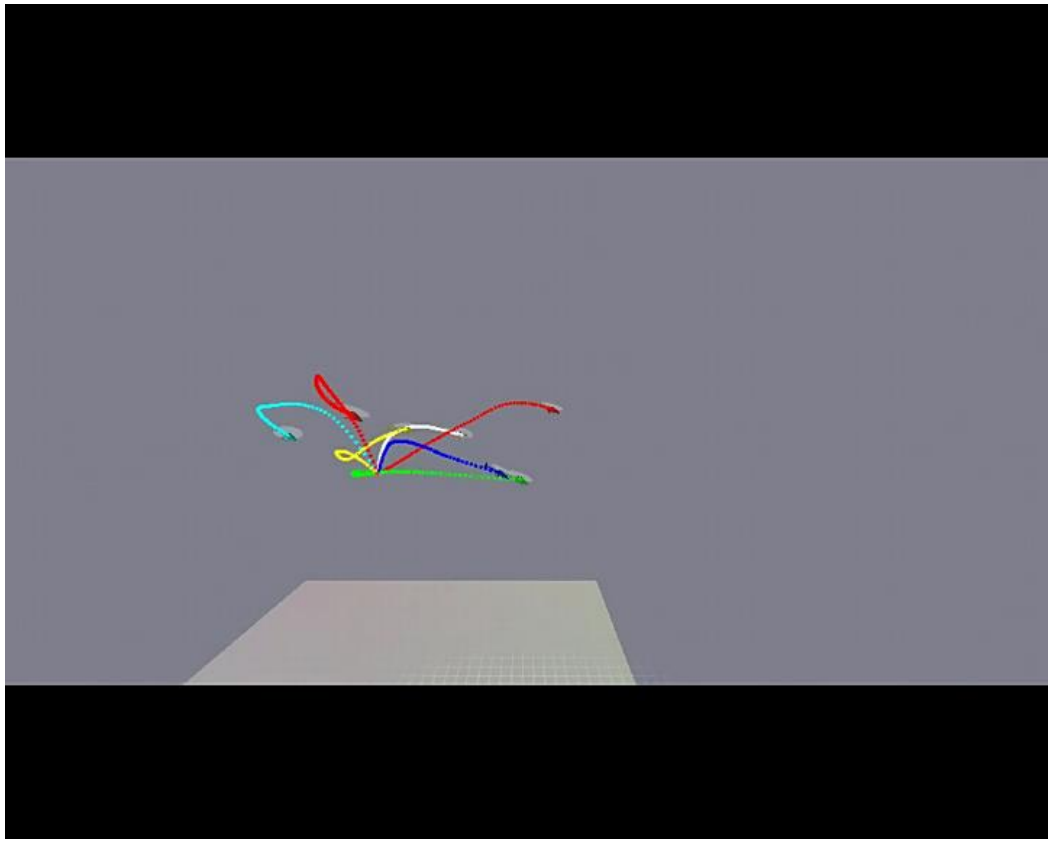
1. Pre-processing Align temporal sequences
2. Feature Selection control: acceleration, height, ...
3. Model Building



# Ex4. Automatic Helicopter

Steps:

1. Pre-processing Align temporal sequences
2. Feature Selection control: acceleration, height, ...
3. Model Building Bayesian model



# Unit Outline

https://uob-coms21202.github.io/COMS21202.github.io/

| *Weeks* | *Monday Lecture* | *Wednesday Lecture* | *Labs* | *Thursday Lecture* | *Assessments* |
|---|---|---|---|---|---|
| 13 | Data, Data Modelling and Estimation (I) | Data, Data Modelling and Estimation (II) | Intro to Jupiter Notebook I | Problem Class - Data Acquisition<br>answers | - |
| 14 | Data Modelling and Estimation (III) | Problem Class - Deterministic Data Modelling<br>answers | Intro to Jupiter Notebook II | Data, Data Modelling and Estimation (IV) | - |
| 15 | Data, Data Modelling and Estimation (V) | Problem Class - Probabilistic Data Modelling<br>answers | Least Squares | Review part I<br>answers | CW1 (set) |
| 16 | Classification I | Classification II | Maximum Likelihood | Clustering | - |
| 17 | Problem Class | Gaussian Mixture Methods | Fitting | Evaluation Methods | - |
| 18 | **Computer Science Explore Week** | | | | - |
| 19 | Problem class | Problem Class | Classification | Review part II | - |
| 20 | Features | Features | - | Features | - |
| 21 | Features | Problem Class | - | Features | - |
| **Easter Break** | | | | | |
| 22 | Features | Features | - | Problem Class | CW1 (deadline) |
| 23 | Review part I (Rui) | Review part II (Laurence) | - | Review Part III (Majid) | - |
| 24 | **Review week** | | | | |

# Assessments

- One course individual course work (40%)
- Assessment for course work is marked in the form of a report

# Assessments

- One course individual course work (40%)
- Assessment for course work is marked in the form of a report - it's what you have understood about the data that matters

# Assessments

- One course individual course work (40%)
- Assessment for course work is marked in the form of a report - it's what you have understood about the data that matters

- Exam (60%)

# Assessments

- One course individual course work (40%)
- Assessment for course work is marked in the form of a report - it's what you have understood about the data that matters
- Exam (60%)
- Unit Averages
  - 2016/2017 Avg: 60
  - 2015/2016 Avg: 56

# Labs

- Tuesdays 13:00 - 15:00 [by timetable]
- Thursday 09:00 - 11:00 [by timetable]

# Labs

- Tuesdays 13:00 - 15:00 [by timetable]
- Thursday 09:00 - 11:00 [by timetable]
- Lab Environment

# Labs

- Tuesdays 13:00 - 15:00 [by timetable]
- Thursday 09:00 - 11:00 [by timetable]
- Lab Environment


- **Lab Work:**
  - Do the labs in pairs during weeks 13-19
  - You can work in groups for CW but report has to be individual, weeks 15-22 [CW submission in Wk 22]

# Tasks

- Next Lab (Week 13): Introduction to Jupyter Notebook I
- Sheet on unit web page

- Next Problem Class (Thur 1-2): Data Acquisition
- Prepare your answers in advance [available online]