

# COMS21202: An Introduction to Doing Things with Data

based on Dima Damen lecture notes

Rui Ponte Costa

[rui.costa@bristol.ac.uk](mailto:rui.costa@bristol.ac.uk)

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

January 22, 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

200

0

2005

2006

2007

2008

2009

2010

2011

**2,000%**

Expected increase in  
global data by 2020



**Megabytes**

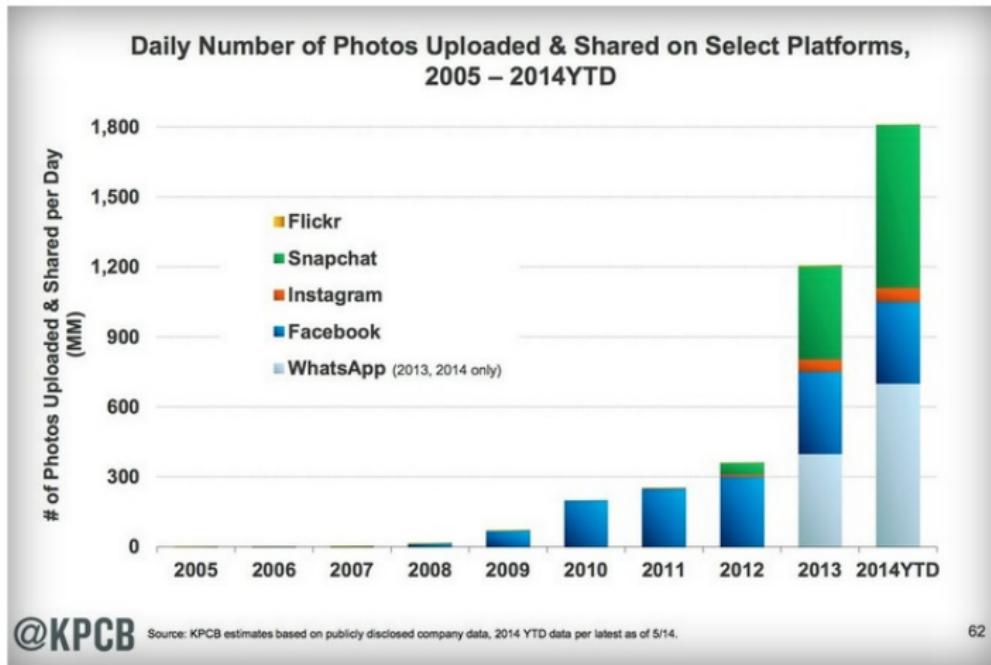
Video and photos stored  
by Facebook, per user

**75%**

Percentage of all digital  
data created by consumers

Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet

# This Unit



# This Unit



# This Unit



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



[sources.dmnnews.com](http://sources.dmnnews.com), [infinitdatum.com](http://infinitdatum.com), [code-n.org](http://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

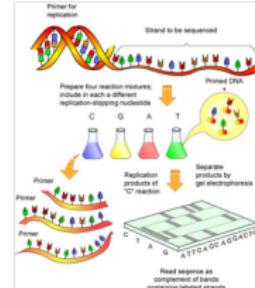


Curly Fries: Researchers at Cambridge's Psychometric Centre have joined forces with Microsoft to analyse more than nine million 'likes' on Facebook. Photo: ALAMY

# 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

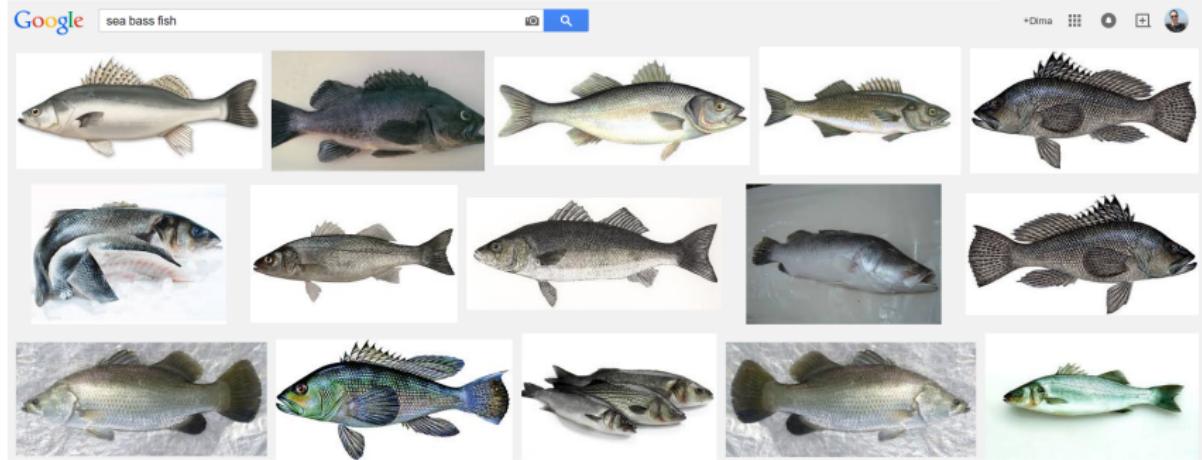


From: Pattern Classification by Duda, Hart and Stork

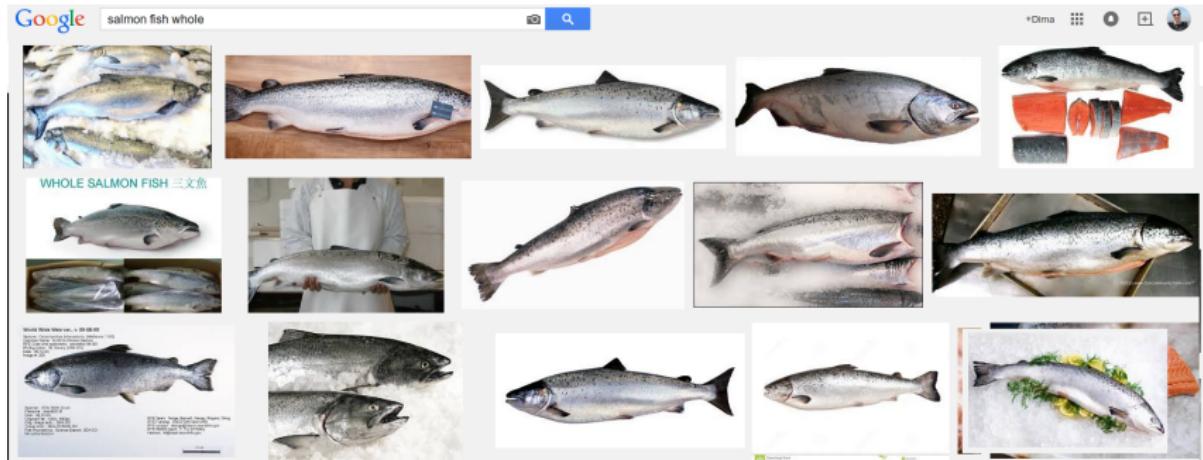
**Data:** images of fish

**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
  - » Rui Ponte Costa
2. Feature Selection
  - » 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

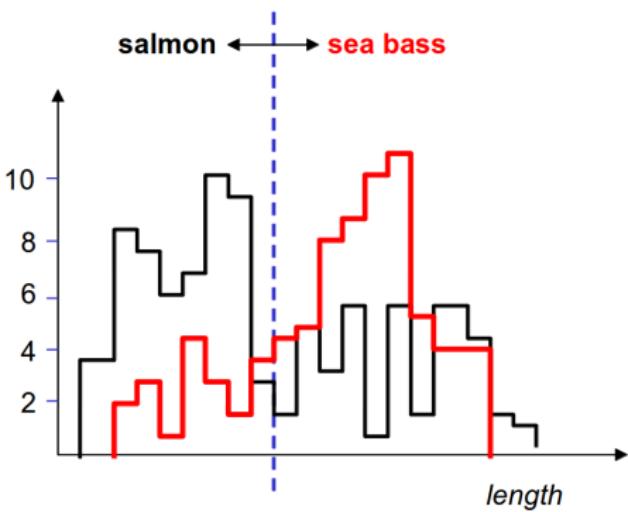
E.g.:

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.:

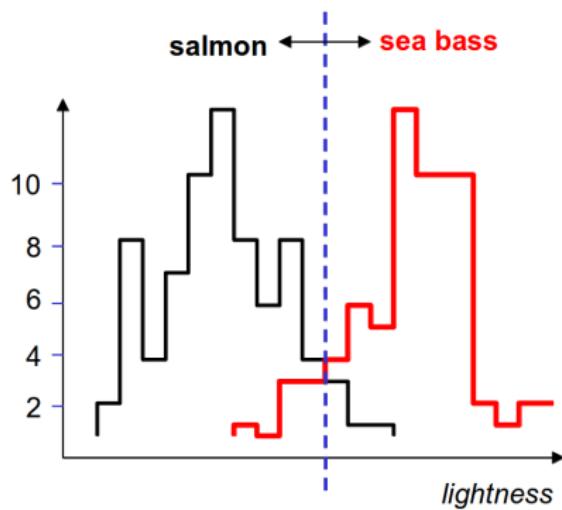
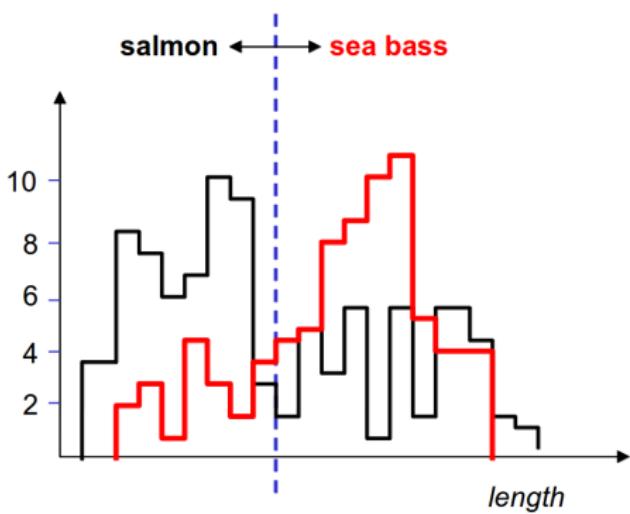
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. measure length
3. Classification e.g. find a threshold



# Fishing for a Solution

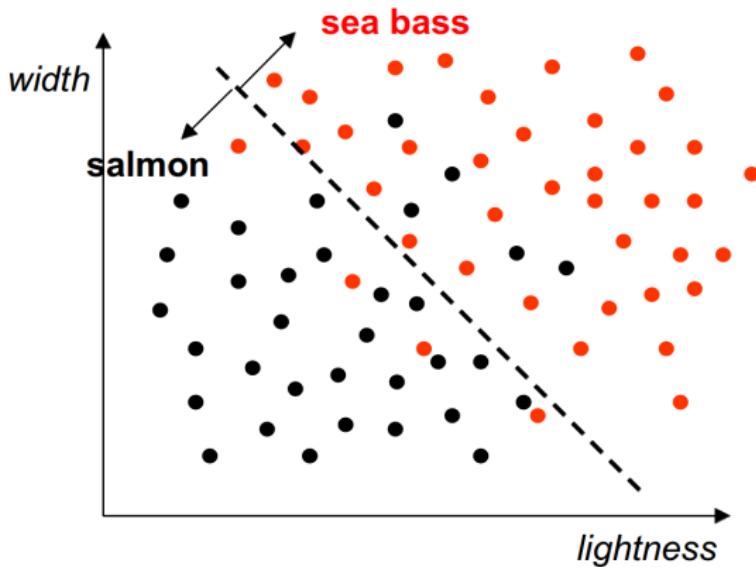
E.g.:

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 frequencies 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 frequencies in training data.

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

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

and

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

Email is spam if

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

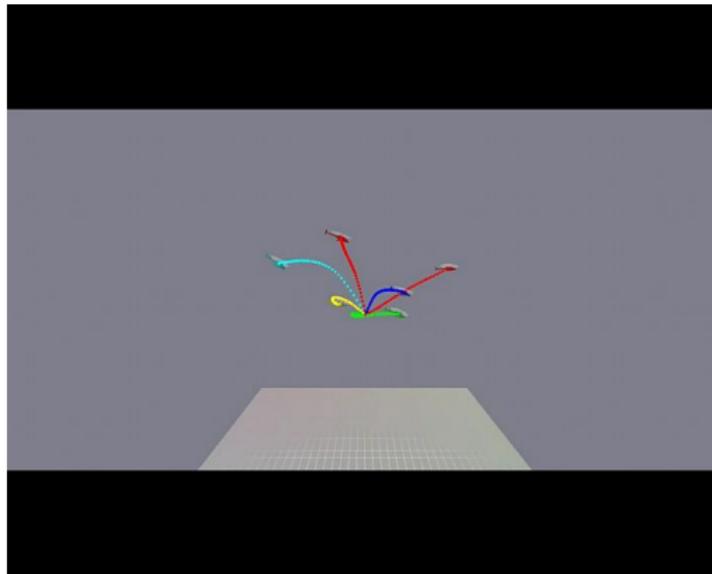
## Ex4. Automatic Helicopter



# Ex4. Automatic Helicopter

**Data:** expert demonstration

**Aim:** fly an autonomous helicopter



## Ex4. Automatic Helicopter

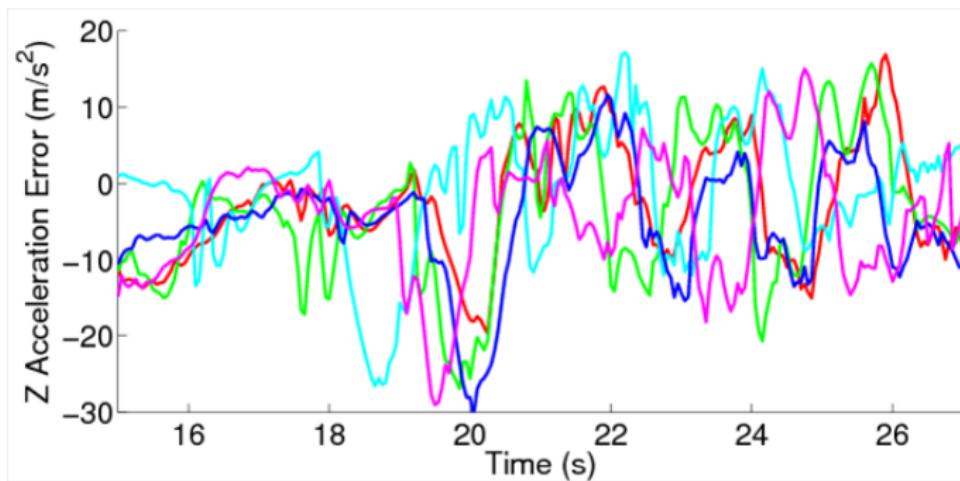
Steps:

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

## Ex4. Automatic Helicopter

Steps:

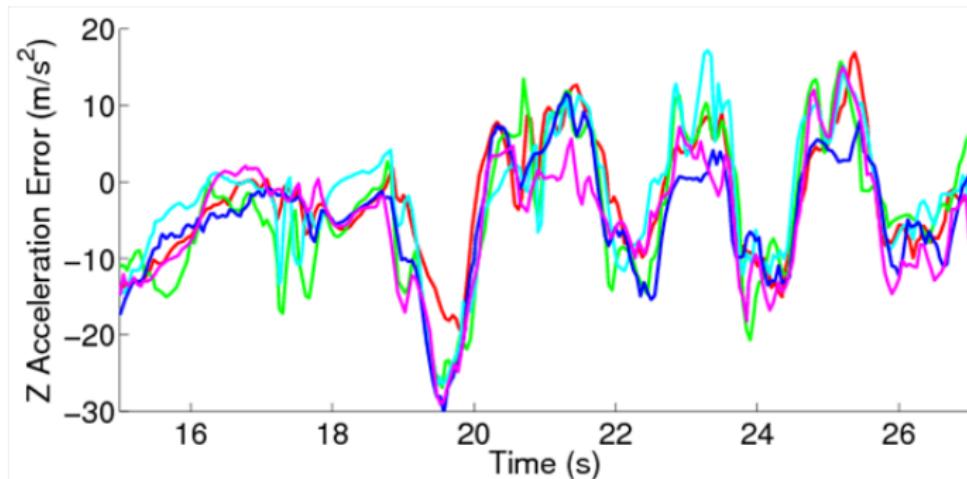
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



## Ex4. Automatic Helicopter

Steps:

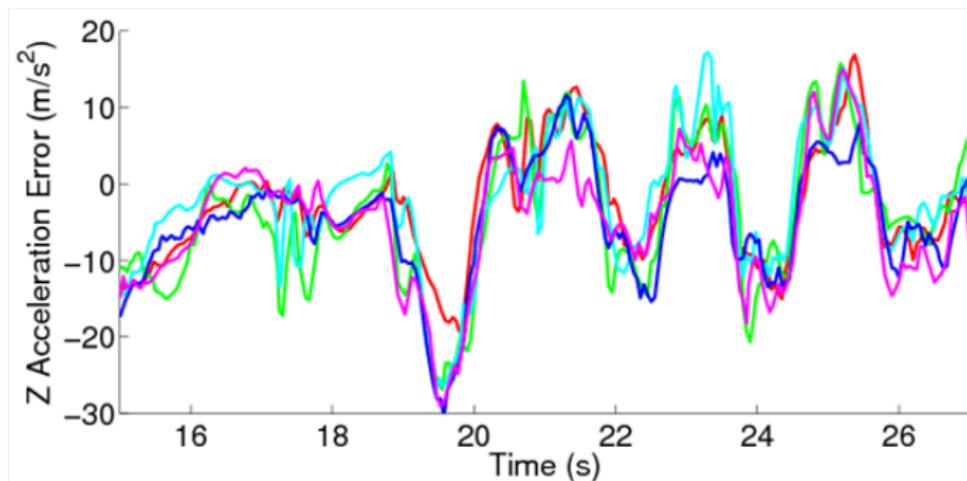
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



## Ex4. Automatic Helicopter

Steps:

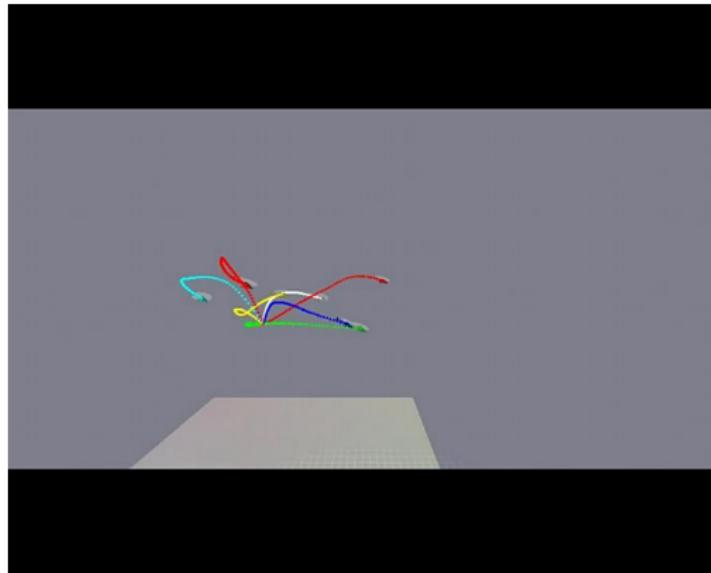
1. Pre-processing Align temporal sequences
2. Feature Selection control: acceleration, height, ...
3. Model Building



## 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	<a href="#">Data, Data Modelling and Estimation (I)</a>	Data, Data Modelling and Estimation (II)	<a href="#">Intro to Jupiter Notebook I</a>	Problem Class - Data Acquisition	-
14	Data Modelling and Estimation (III)	Problem Class - Deterministic Data Modelling	Intro to Jupiter Notebook II	Data, Data Modelling and Estimation (IV)	-
15	Data, Data Modelling and Estimation (V)	Problem Class - Probabilistic Data Modelling	Least Squares	Review part I	CW (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	<a href="#">Features</a>	<a href="#">Features</a>	-	<a href="#">Features</a>	-
21	<a href="#">Features</a>	<a href="#">Problem Class</a>	-	<a href="#">Features</a>	CW (deadline)
Easter Break					
22	<a href="#">Features</a>	<a href="#">Features</a>	-	Problem Class	-
23	Review part I (Rui)	Review part II (Laurence)	-	Review Part III (Majid)	-
24	Review week				

# Assessments

- ▶ One individual course work: report + code (40%)
- ▶ CW: You encouraged to discuss with others, but submissions are individual.
- ▶ Assessment for course work is marked in the form of a report

# Assessments

- ▶ One individual course work: report + code (40%)
- ▶ CW: You encouraged to discuss with others, but submissions are individual.
- ▶ 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 individual course work: report + code (40%)
- ▶ CW: You encouraged to discuss with others, but submissions are individual.
- ▶ 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 individual course work: report + code (40%)
- ▶ CW: You encouraged to discuss with others, but submissions are individual.
- ▶ 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
  - ▶ 2018/2019 Avg: 66
  - ▶ 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 15-21 [CW submission in Wk 21]

# Labs: Important!!

- ▶ Main source of 1:1 support will be from the TAs in the labs!
- ▶ Labs are essential for the coursework!
- ▶ Attendance will be taken.

# 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]