COMS21202: An Introduction to Doing Things with Data

Dima Damen
Dima.Damen@bristol.ac.uk

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

January 22, 2016

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

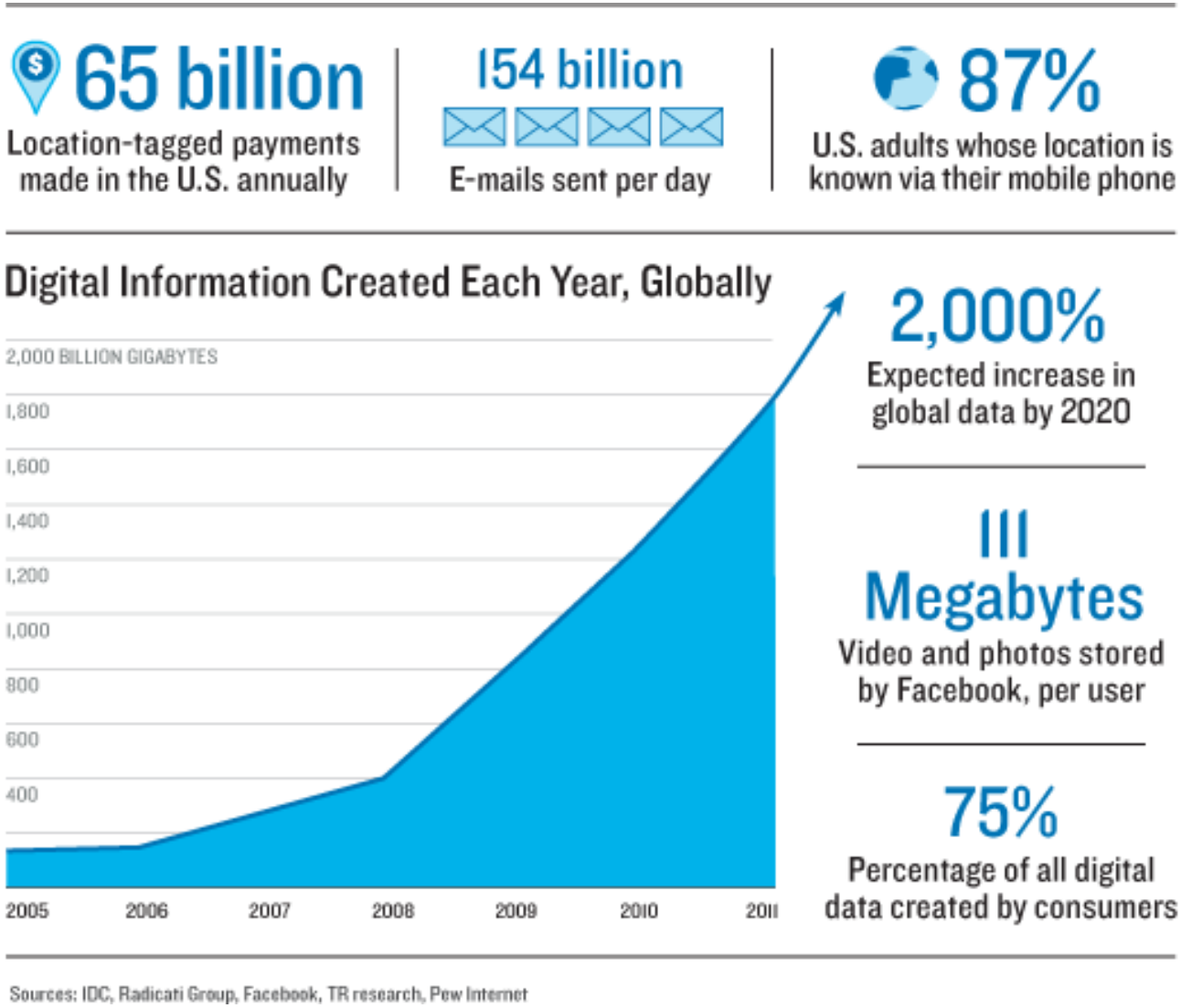
This Unit

- 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)

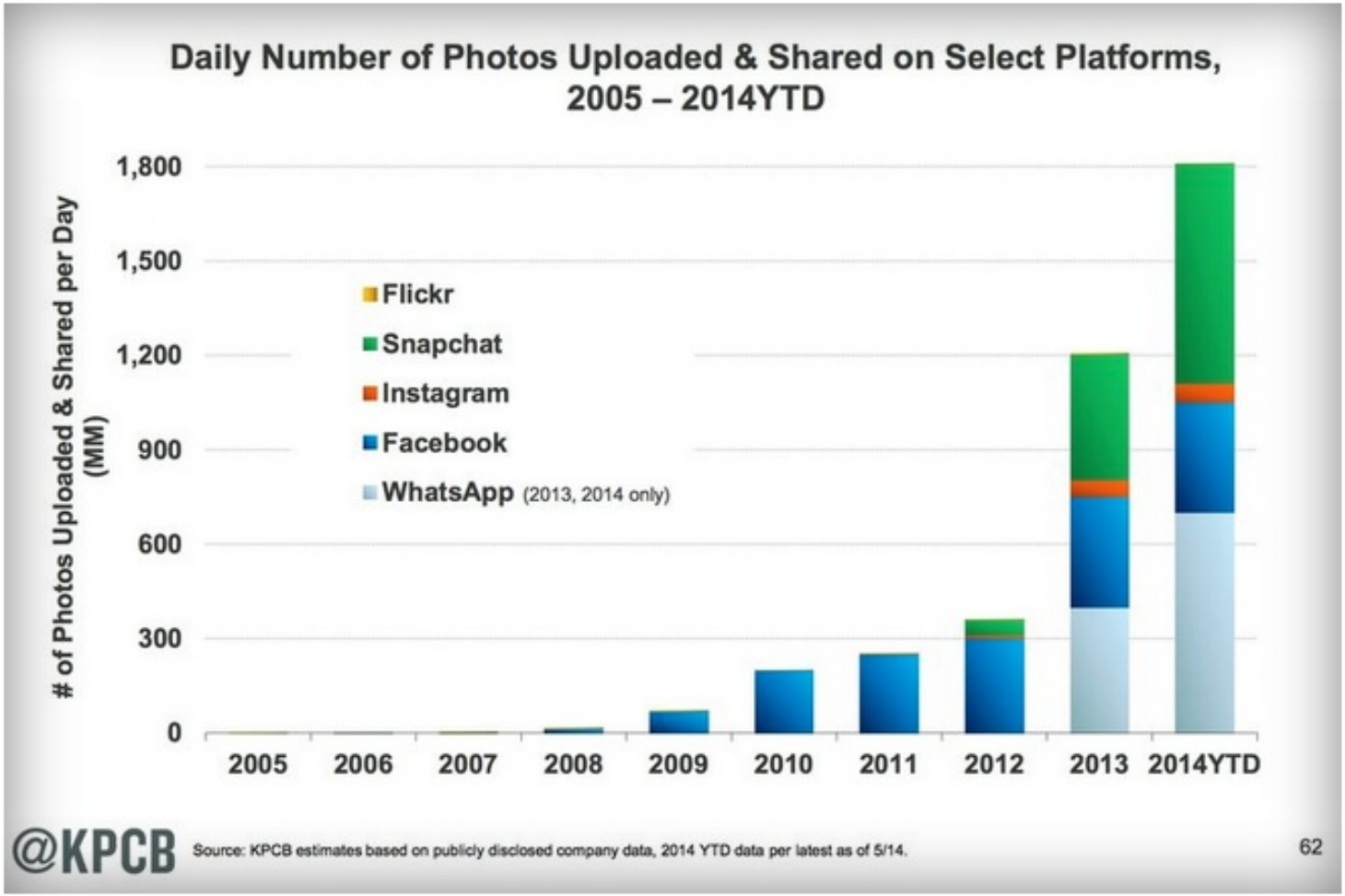
This Unit

- 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

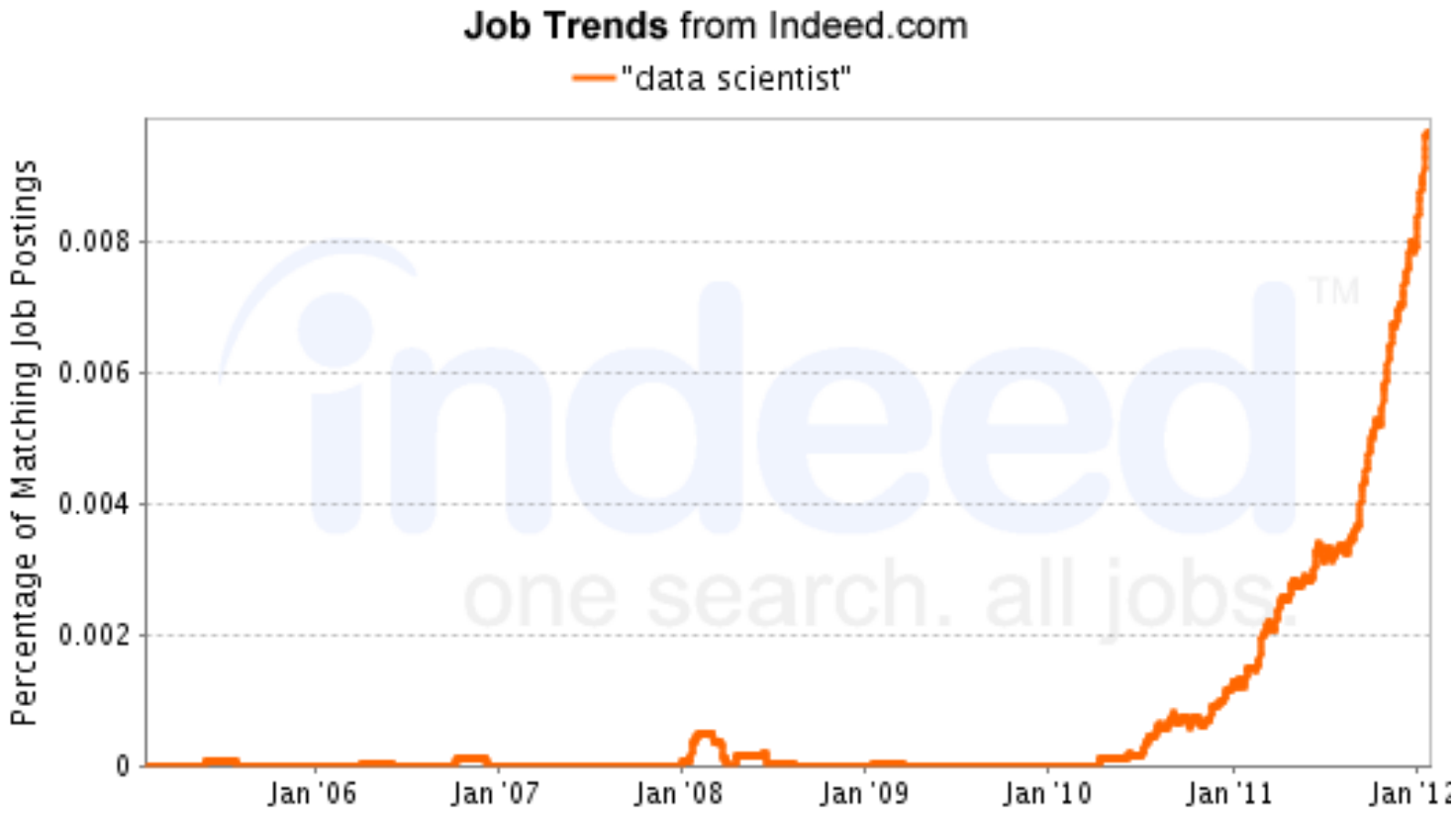
This Unit


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
111 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



source: topcoder.com/blog

This Unit

source: opensource.com

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

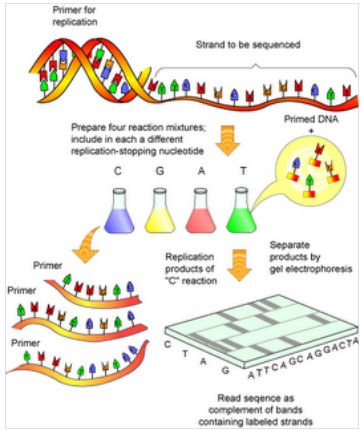
Curly Fries: Researchers at Cambridge's Psychometric Centre have joined forces with Microsoft to analyse more than nine million 'likes' on Facebook Photo: ALAMY

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.
- Hence, preparation for application units in years 3 and 4.



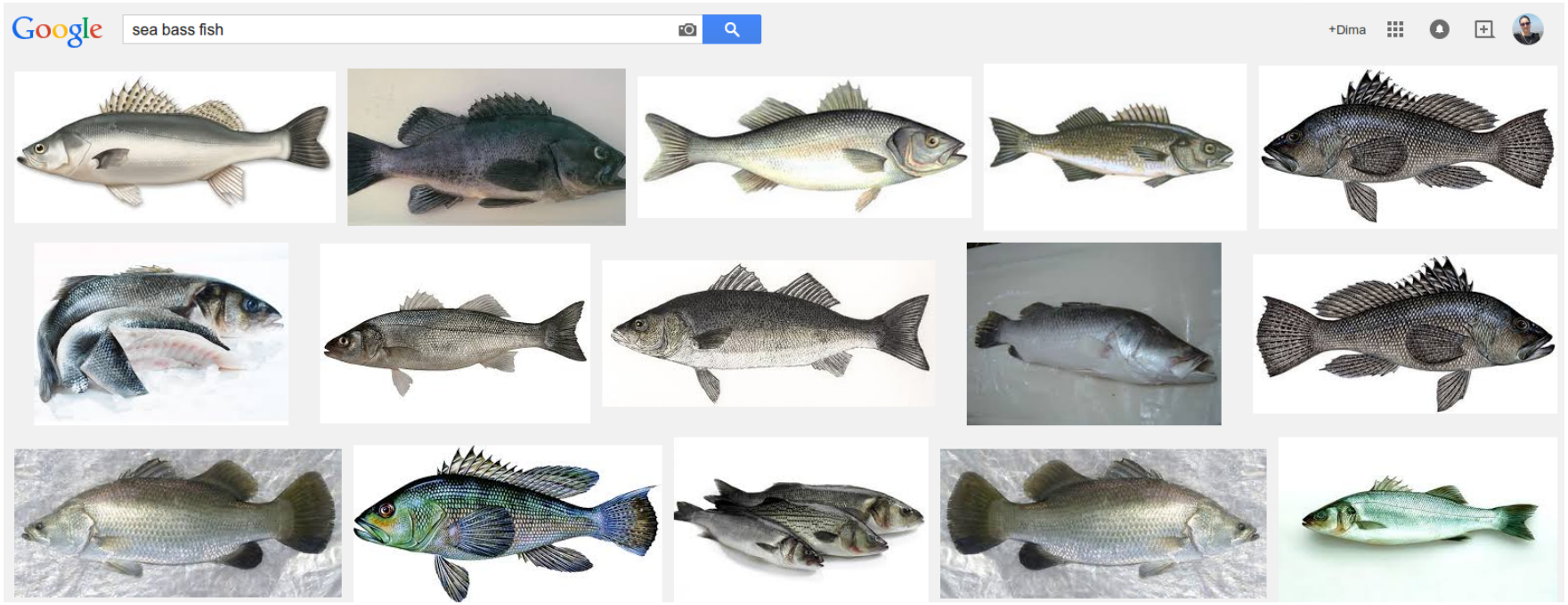
Ex1. A Fishy Problem

From: Pattern Classification by Duda, Hart and Stork

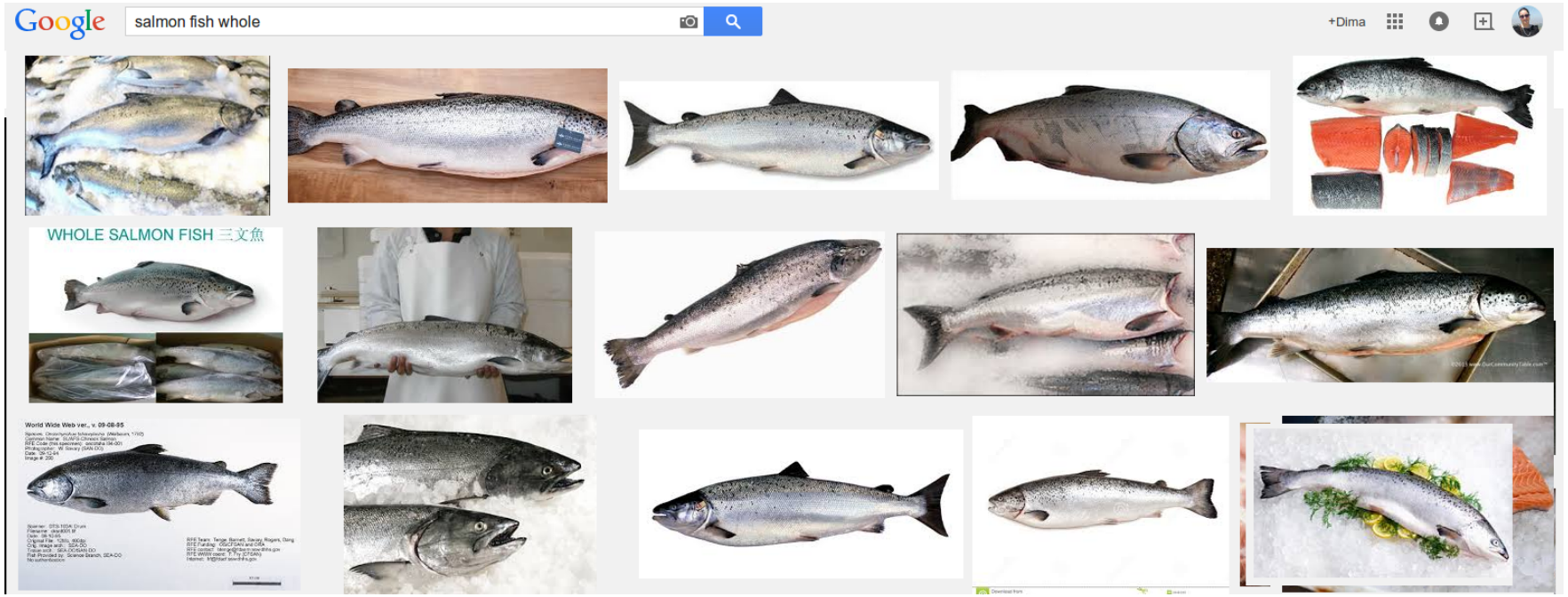
Data: images of fish

Aim: distinguish between sea bass and salmon

Ex1. A Fishy Problem



Ex1. A Fishy Problem



Ex1. A Fishy Problem

Steps:

Ex1. A Fishy Problem

Steps:

1. Pre-processing

Ex1. A Fishy Problem

Steps:

1. Pre-processing
2. Feature Selection

Ex1. A Fishy Problem

Steps:

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

Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1]
2. Feature Selection
3. Classification

Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1]
2. Feature Selection
3. Classification [Unit - Part 2]

Ex1. A Fishy Problem

Steps:

1. Pre-processing [Unit - Part 1]
2. Feature Selection [Unit - Part 3]
3. Classification [Unit - Part 2]



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

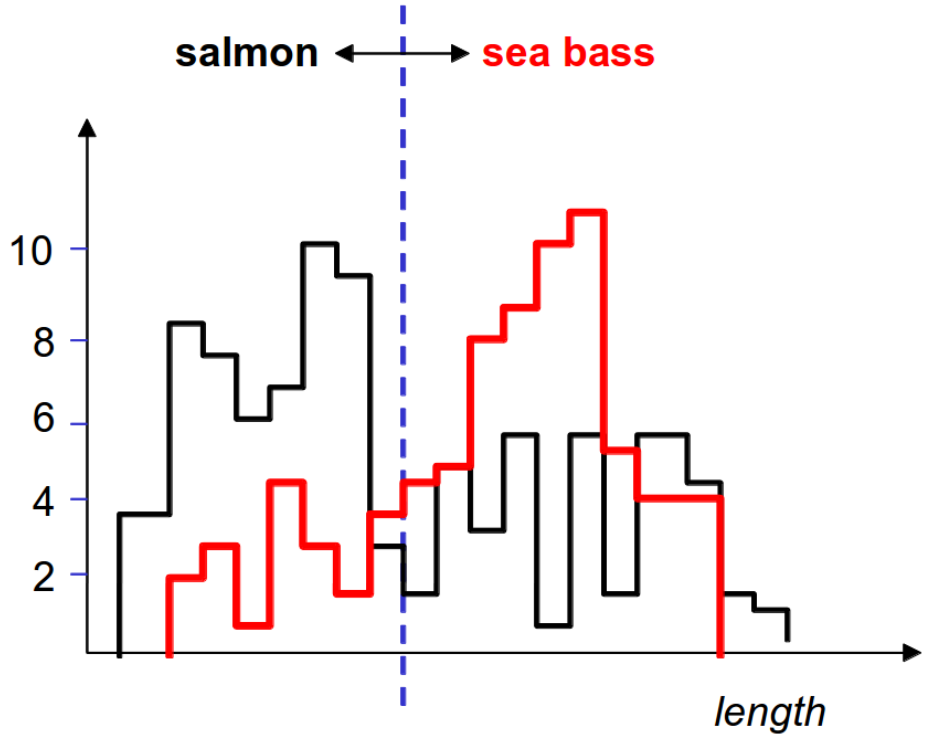
E.g.:

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.:

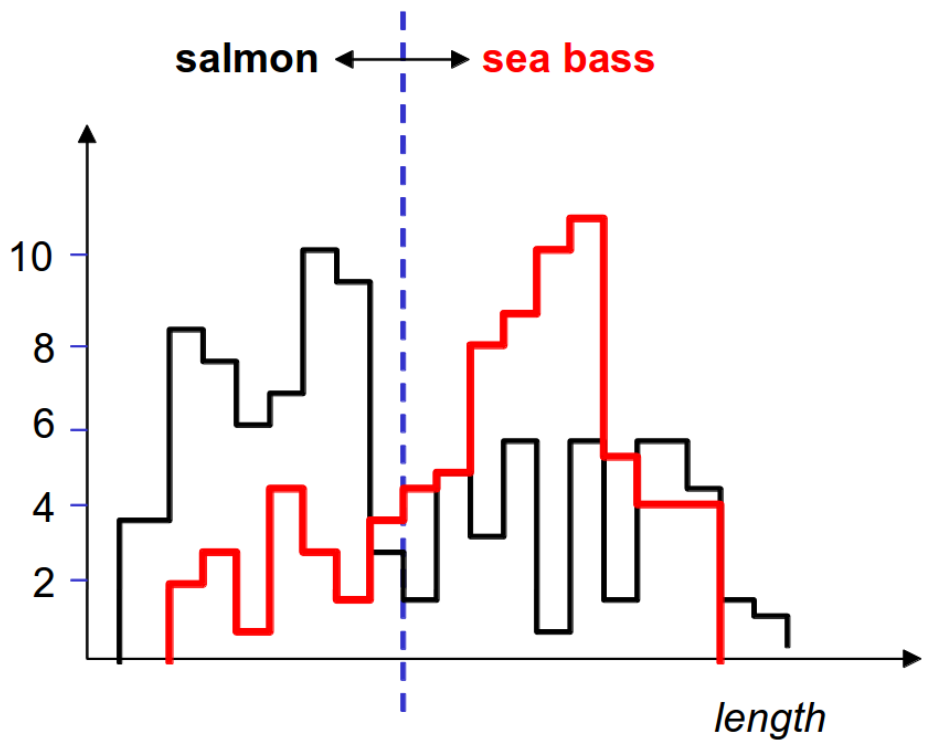
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. measure length
3. Classification e.g. find a threshold

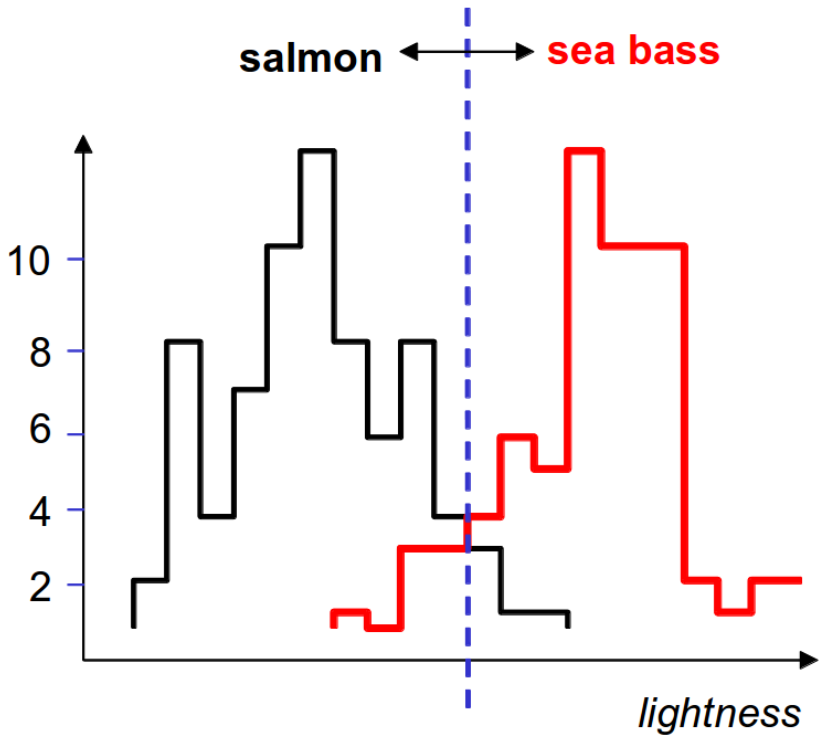


Fishing for a Solution

E.g.:

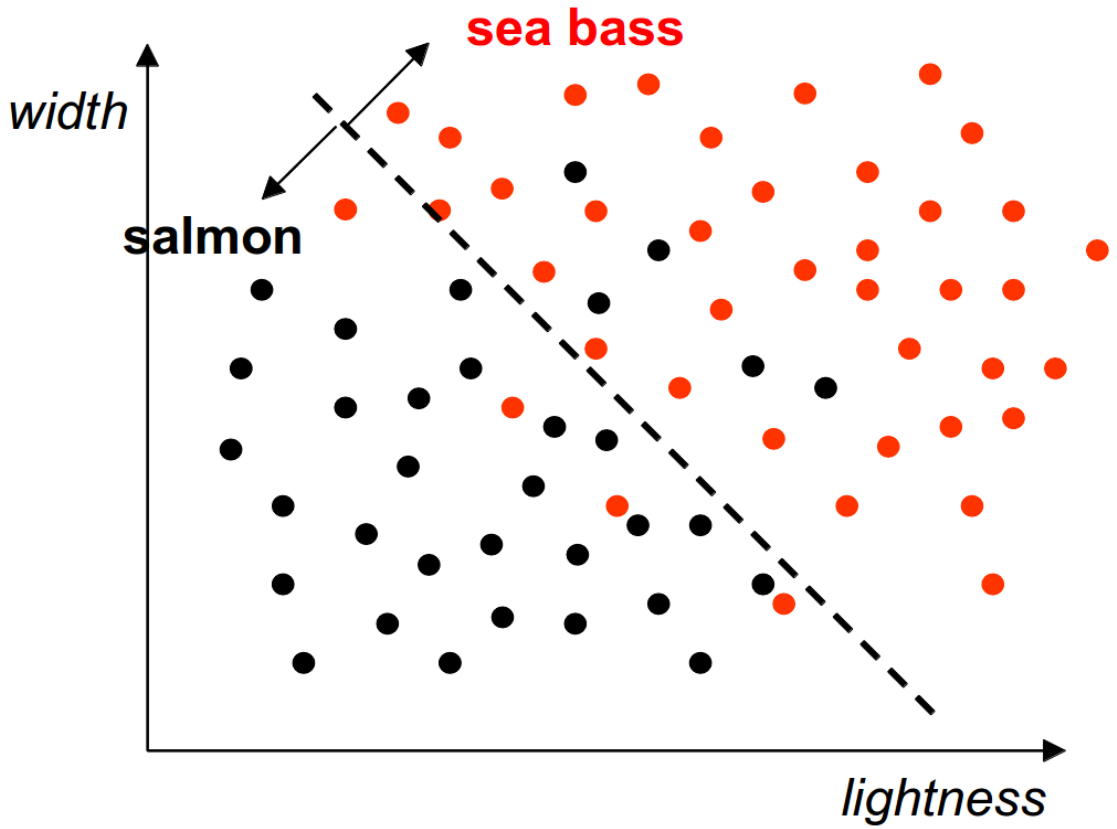
1. Pre-processing e.g. Rotate and align, Segment fish from background
2. Feature Selection e.g. measure length or brightness
3. Classification e.g. find a threshold





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

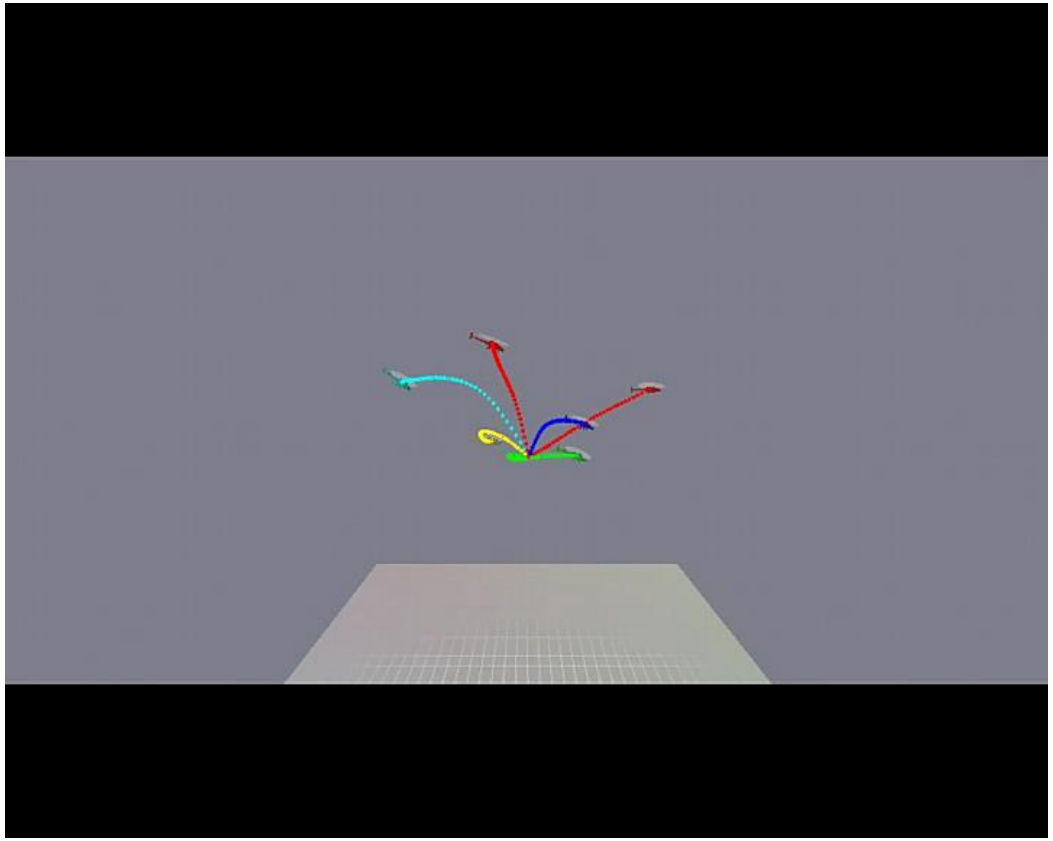
$$P(email|spam) > P(email|\neg spam) \quad (3)$$

Ex4. Automatic Helicopter

Ex4. Automatic Helicopter

Data: expert demonstration

Aim: fly an autonomous helicopter



Ex4. Automatic Helicopter

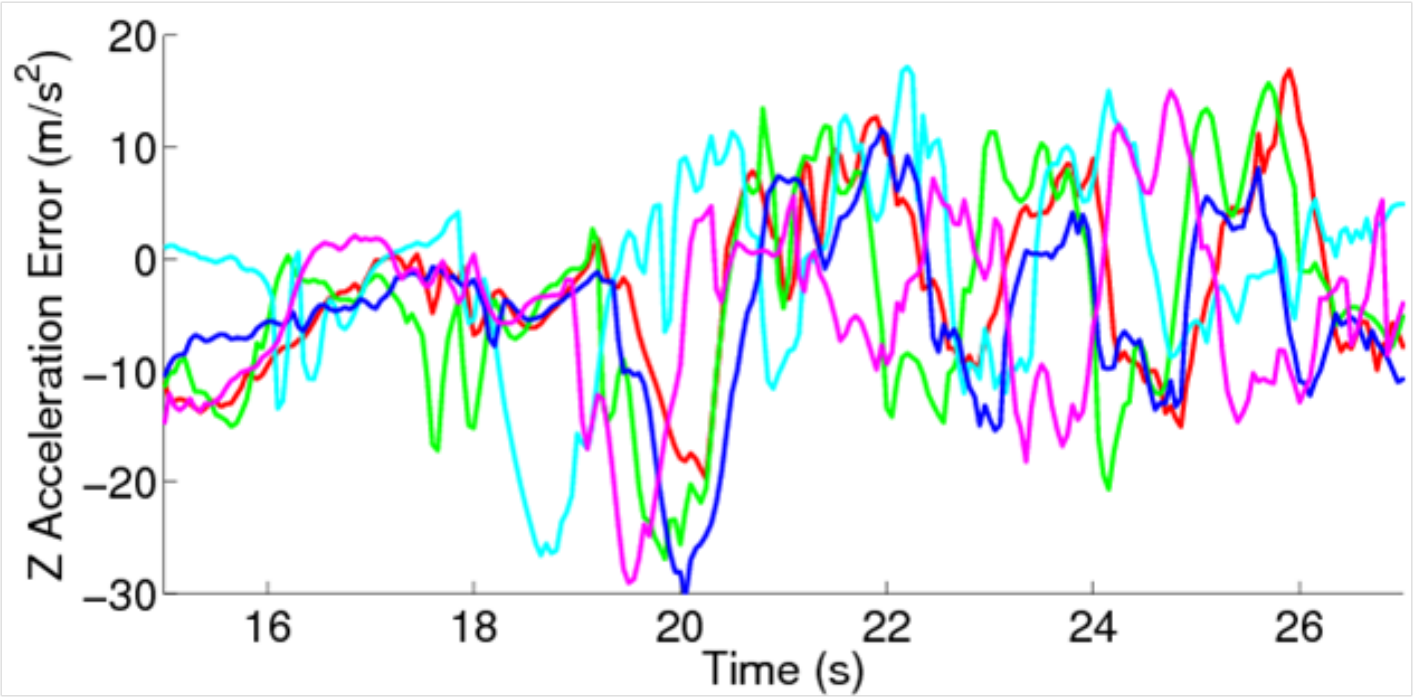
Steps:

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

Ex4. Automatic Helicopter

Steps:

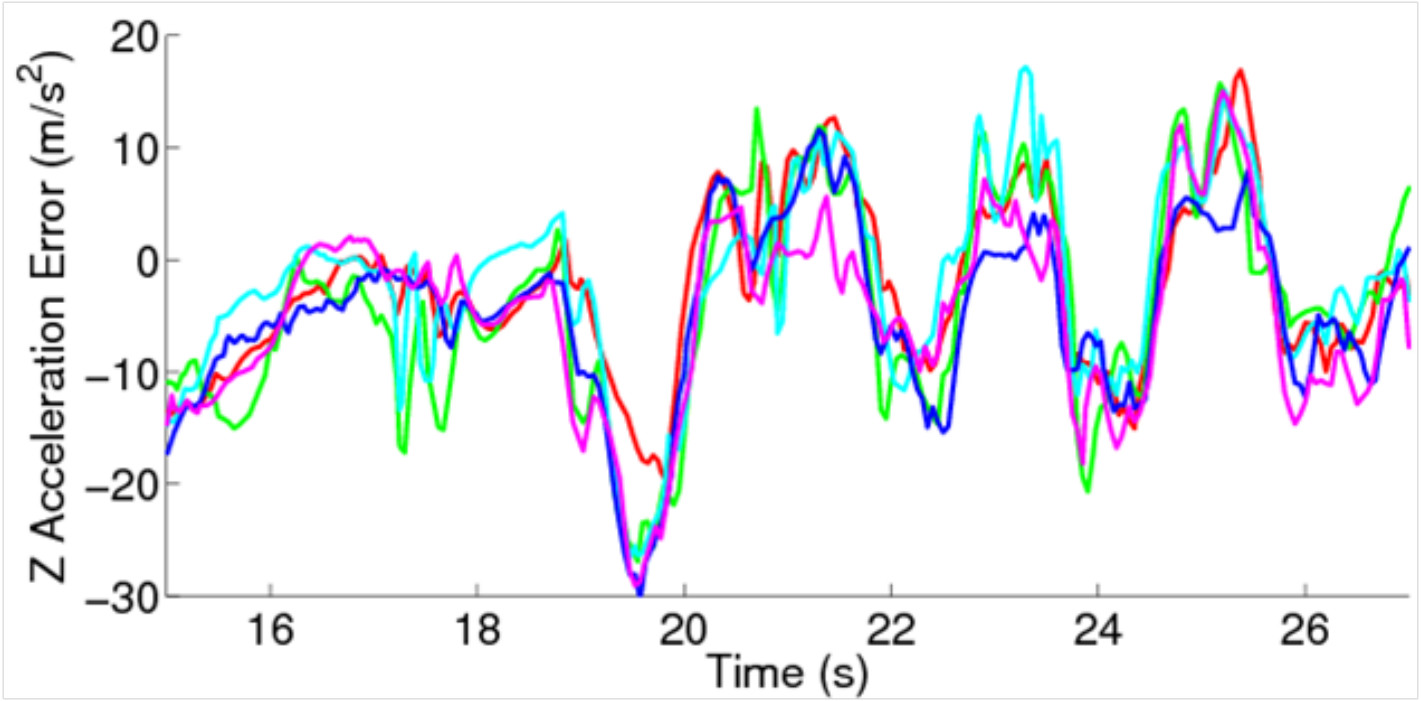
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



Ex4. Automatic Helicopter

Steps:

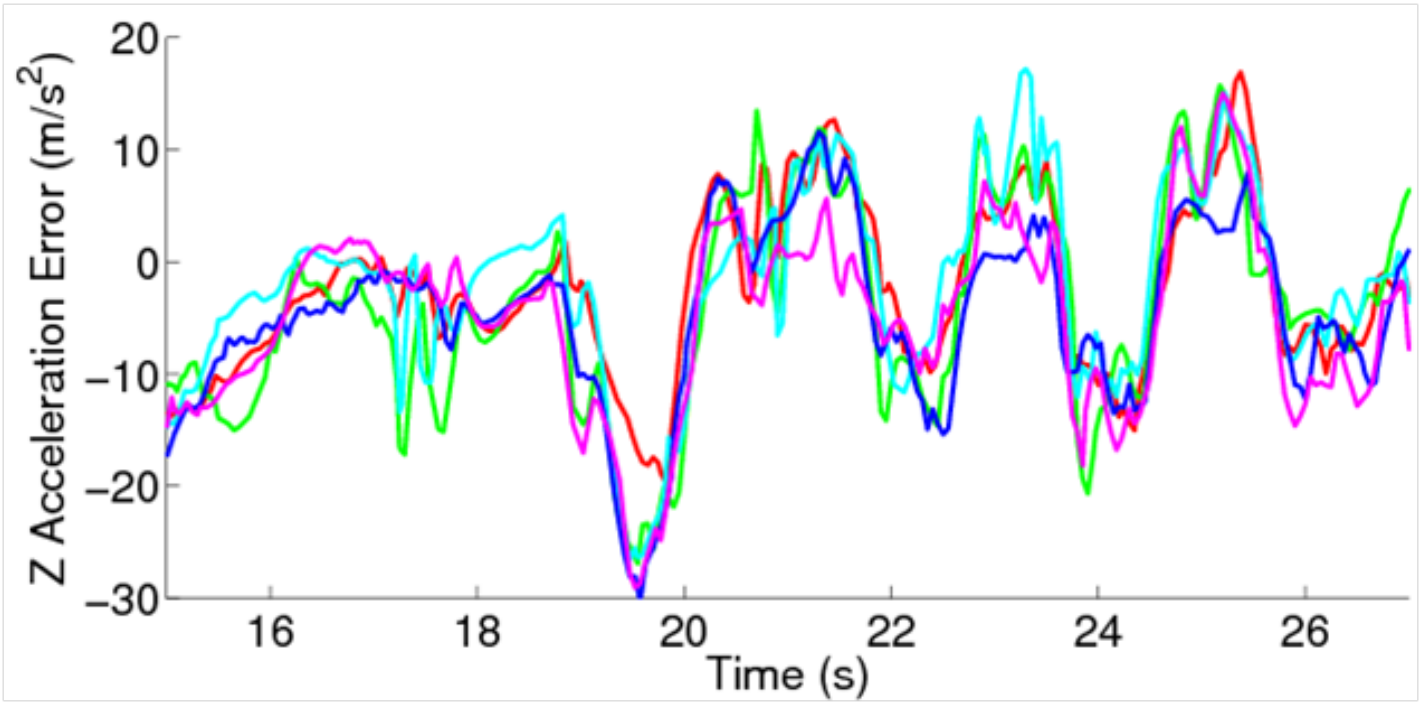
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



Ex4. Automatic Helicopter

Steps:

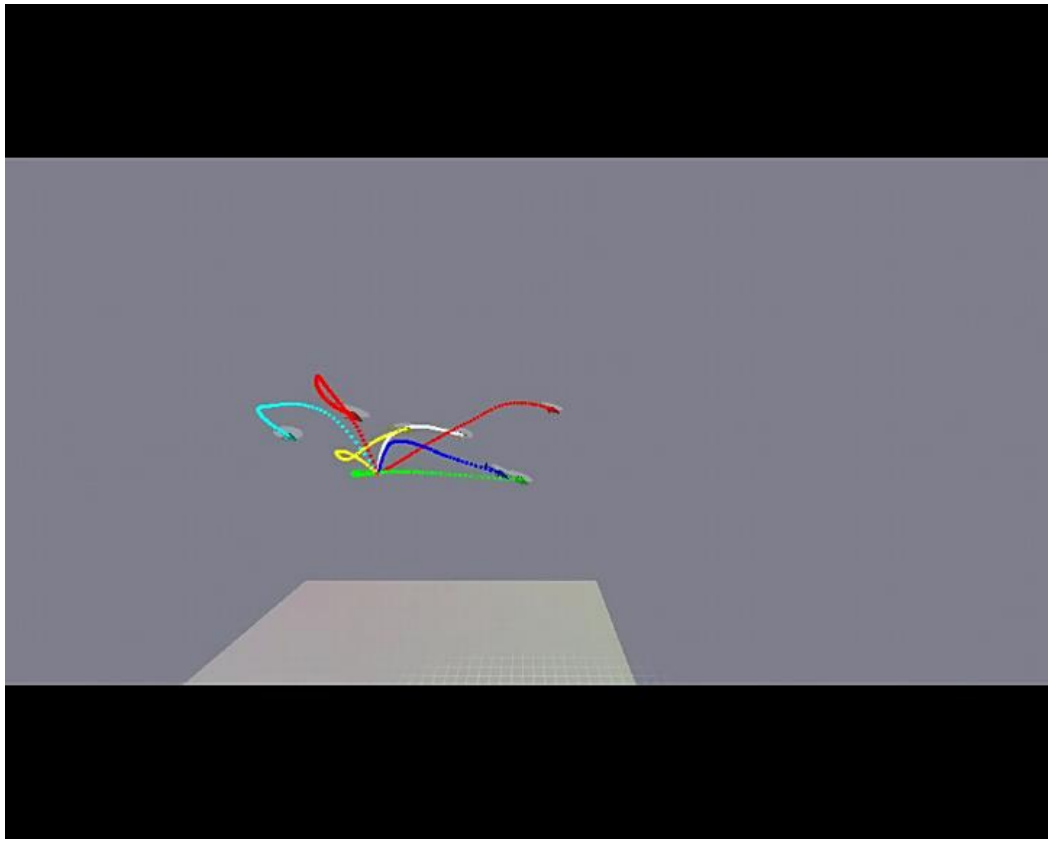
1. Pre-processing Align temporal sequences
2. Feature Selection control: acceleration, height, ...
3. Model Building



Ex4. Automatic Helicopter

Steps:

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



Unit Outline

Unit Outline

Weeks	*Monday Lecture*	*Thursday Lecture*	*Labs*	*Friday Lecture*	*Assessments*
13	Data, Data Modelling and Estimation (I)	Data, Data Modelling and Estimation (II)	Intro to Matlab/Jupiter Notebook I	Problem Class - Data Acquisition	-
14	Data, Data Modelling and Estimation (III)	Data, Data Modelling and Estimation (IV)	Intro to Matlab/Jupiter Notebook II	Problem Class - Deterministic Data Modelling	-
15	Data, Data Modelling and Estimation (V)	Data, Data Modelling and Estimation (VI)	Least Squares	Problem Class - Probabilistic Data Modelling	-
16	Test 1	Classification and Clustering I	Max. Likelihood	Classification and Clustering II	Test1, CW1 (set)
17	Classification and Clustering III	Problem Class - Clustering	Clustering - CW1a	Classification and Clustering IV	-
18	Classification and Clustering V	Classification and Clustering VI	Classification - CW1b	Problem Class - Classification	-
19	Test 2	Representation and Feature Extraction I	Extra (Optional) Lab	Representation and Feature Extraction II	Test 2, CW1 (deadline)
20	Representation and Feature Extraction III	Representation and Feature Extraction IV	-	Problem Class - Feature Extraction I	CW2 (set)
Easter Break					
21	Representation and Feature Extraction V	Representation and Feature Extraction VI	-	Representation and Feature Extraction VII	-
22	Representation and Feature Extraction VIII	Problem Class - Feature Extraction II	-	Drop-In Session	-
23	Test 3	Review	-	Review	Test 3, CW2 (deadline)
24	Review	Review	-	-	-

Assessments

- 3 x in-class tests (15%)
- One assignment in pairs (10%)
- One individual assignment (25%)
- Assessments are marked in the form of reports

Assessments

- 3 x in-class tests (15%)
- One assignment in pairs (10%)
- One individual assignment (25%)
- Assessments are marked in the form of reports - it's what you have understood about the data that matters

Assessments

- 3 x in-class tests (15%)
- One assignment in pairs (10%)
- One individual assignment (25%)
- Assessments are marked in the form of reports - it's what you have understood about the data that matters

- Exam (50%)

Assessments

- 3 x in-class tests (15%)
- One assignment in pairs (10%)
- One individual assignment (25%)
- Assessments are marked in the form of reports - it's what you have understood about the data that matters

- Exam (50%)

- Unit Averages
  - 2014/2015 Avg: 57
  - 2013/2014 Avg: 62

Labs

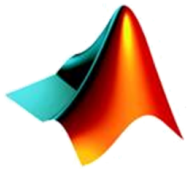
- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00

Labs

- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00
- **New** Lab Environment Options

Labs

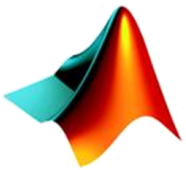
- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00
- **New** Lab Environment Options



MATLAB

Labs

- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00
- **New** Lab Environment Options



MATLAB

Labs

- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00
- **New** Lab Environment Options

- You can either
  - Choose Matlab - SPS's traditional choice
  - Choose Jupyter - because you know Python
  - Try out and choose (Weeks 13-14)

Labs

- CS Students (G400, G402, G403) Thursday 14:00-16:00
- CSE, CS&M, EngMath, Study Abroad (STAB, G160, G161, GG41, GGK1, GH46, J925) Friday 09:00-11:00
- **New** Lab Environment Options

- You can either
  - Choose Matlab - SPS's traditional choice
  - Choose Jupyter - because you know Python
  - Try out and choose (Weeks 13-14)
- **By Week 16 you should have:**
  - Decided on your environment
  - Decided on your lab partner (for CW1) from the same lab session
  - You have to work within your pairs for weeks 17-19

Tasks

- Next Lab (Week 13): Introduction to Matlab and/or Jupyter Notebook
- Sheet on unit web page

- Next Problem Class (Fri 12-1): Data Acquisition
- Prepare your answers in advance