

# 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, 2019

# What is Data?



# 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
  - ▶ 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



**Megabytes**

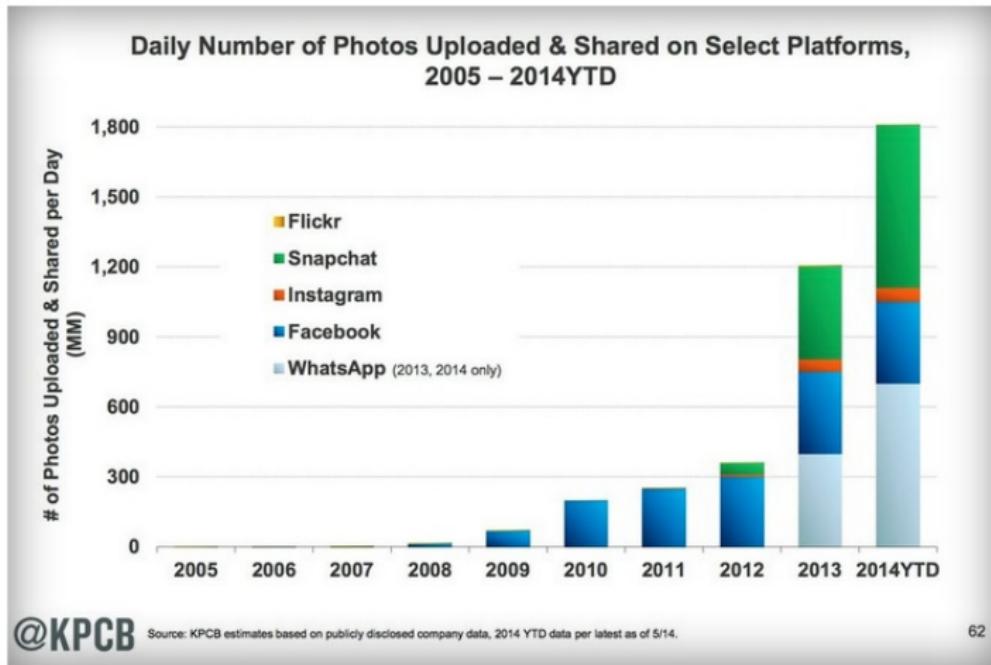
Video and photos stored  
by Facebook, per user

**75%**

Percentage of all digital  
data created by consumers

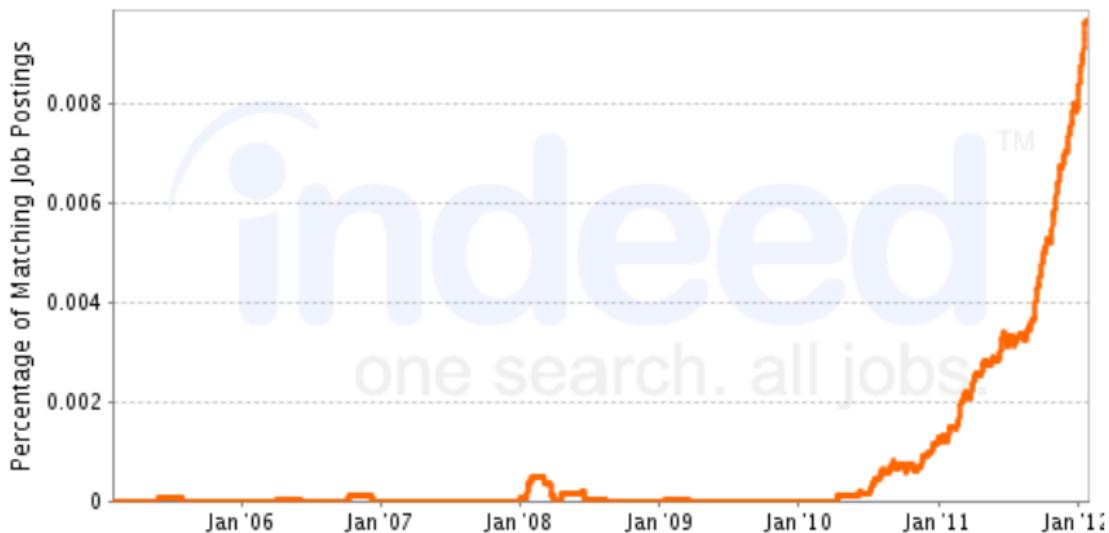
Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet

# This Unit



# This Unit

Job Trends from Indeed.com  
— "data scientist"



# 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

'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](http://telegraph.co.uk)

Dima Damen

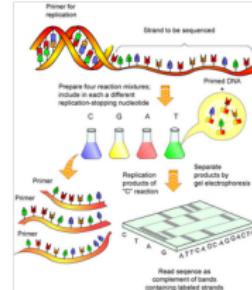
Dima.Damen@bristol.ac.uk

COMS21202: Intro

## 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 v



# Ex1. A Fishy Problem

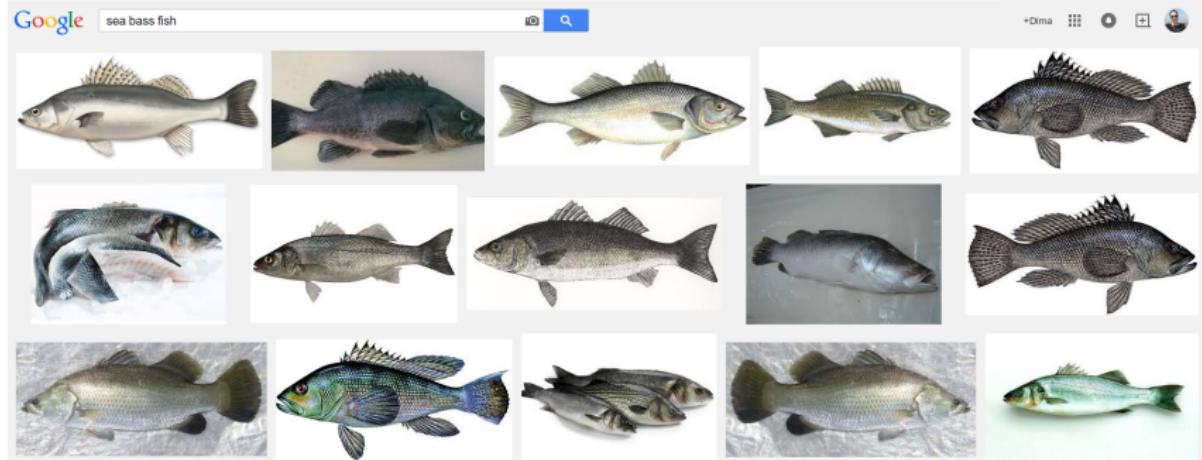


From: Pattern Classification by Duda, Hart and Stork

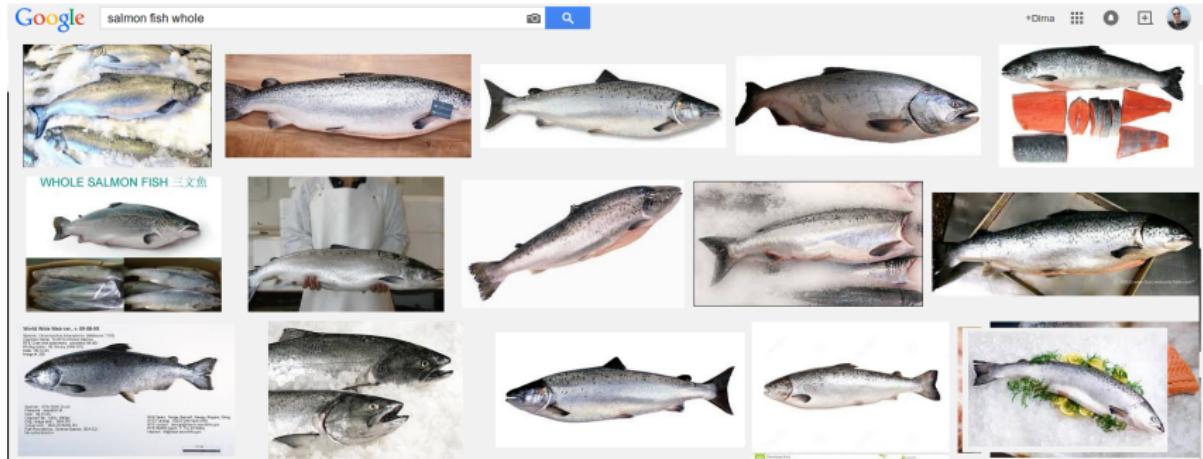
**Data:** images of fish

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

# Ex1. A Fishy Problem



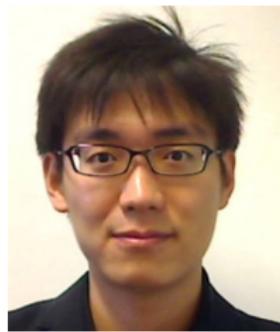
# Ex1. A Fishy Problem



# Ex1. A Fishy Problem

Steps:

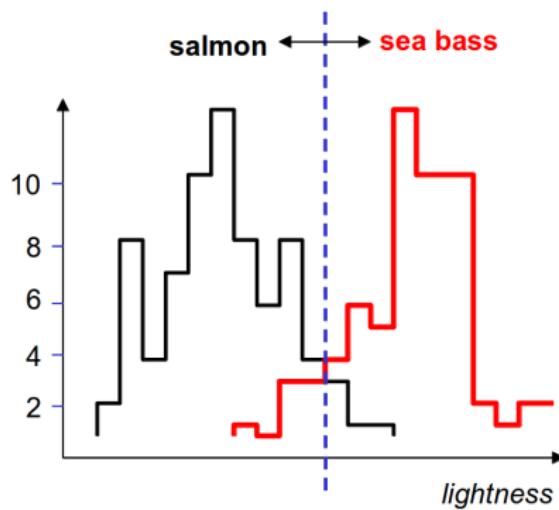
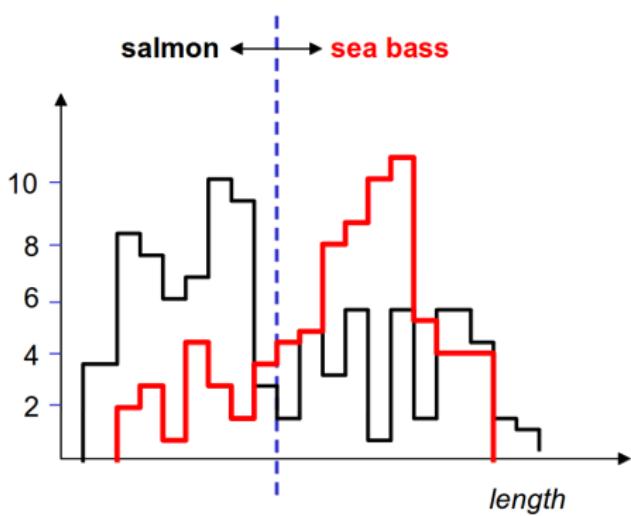
1. Pre-processing [Unit - Part 1]
2. Feature Selection [Unit - Part 3]
3. Classification [Unit - Part 2]



# Fishing for a Solution

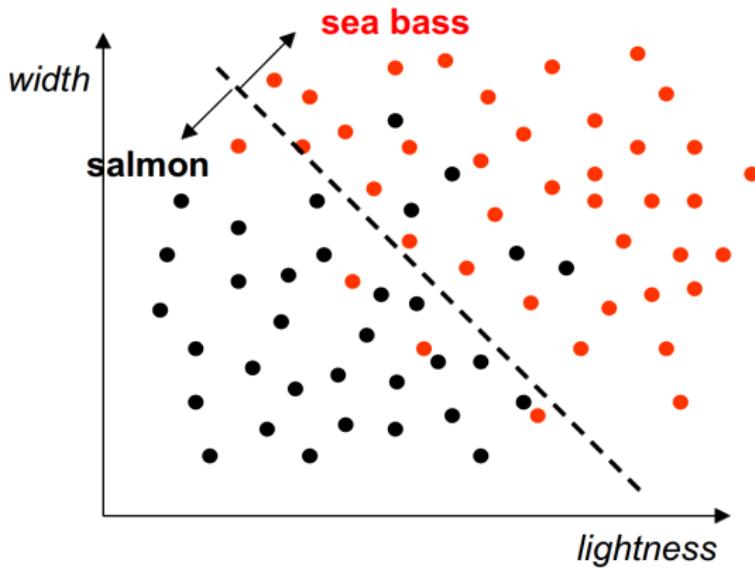
E.g.:

1. Pre-processing 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 **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 **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

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

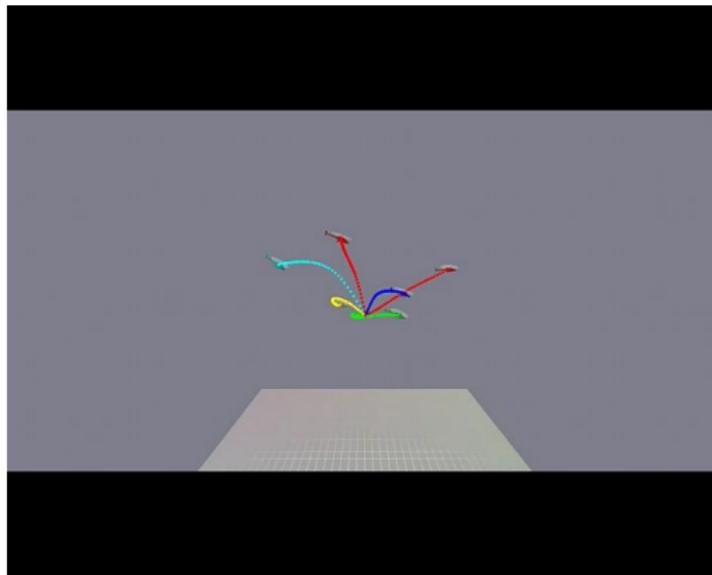
## Ex4. Automatic Helicopter



# Ex4. Automatic Helicopter

**Data:** expert demonstration

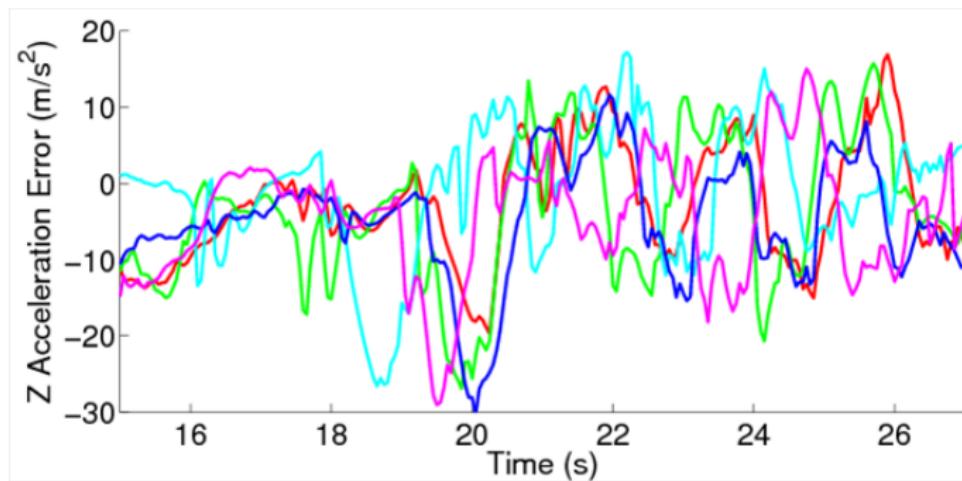
**Aim:** fly an autonomous helicopter



## Ex4. Automatic Helicopter

Steps:

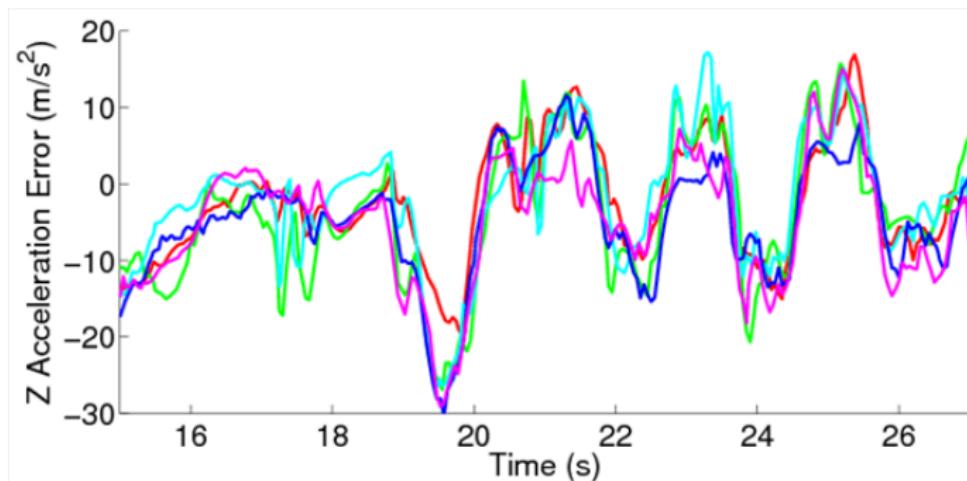
1. Pre-processing Align temporal sequences
2. Feature Selection
3. Model Building



## Ex4. Automatic Helicopter

Steps:

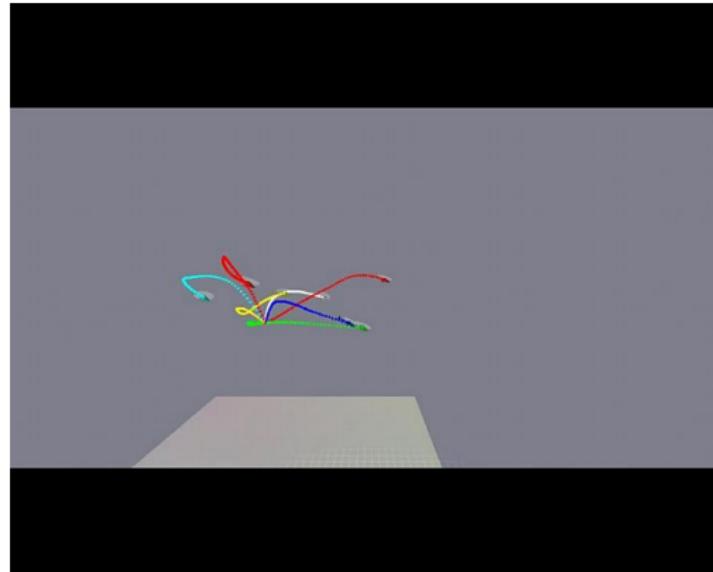
1. Pre-processing Align temporal sequences
2. Feature Selection control: acceleration, height, ...
3. Model Building



## Ex4. Automatic Helicopter

Steps:

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



# Unit Outline

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	-
14	Data, Data Modelling and Estimation (III)	Data Modelling and Estimation (IV)	Intro to Jupiter Notebook II	Problem Class - Deterministic Data Modelling	CW1 (set)
15	Data, Data Modelling and Estimation (V)	Data, Data Modelling and Estimation (VI)	Least Squares	Problem Class - Probabilistic Data Modelling	-
16	Review part I	Classification I	Maximum Likelihood	Classification II	-
17	Clustering I	Problem Class	Fitting	Clustering II	CW1 (deadline)
18	Computer Science Explore Week				-
19	Gaussian Mixture Methods	Evaluation Methods	Classification	Problem Class	CW2 (set)
20	Review Part II	Representation and Feature Extraction	Evaluating Classifiers	Representation and Feature Extraction	-
21	Representation and Feature Extraction	Representation and Feature Extraction	Feature Selection	Representation and Feature Extraction	-
22	Representation and Feature Extraction	Problem Class - Representation and Feature Extraction	-	Problem Class - Representation and Feature Extraction	-
Easter Break					
23	Representation and Feature Extraction	Review part III	-	Review Part I	CW2 (deadline)
24	Bank Holiday	-	-	-	

# Assessments

- ▶ One individual assignment (15%)
- ▶ One assignment in Pairs (25%)
- ▶ Assessment for CW2 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]
- ▶ Lab Environment



- ▶ **Lab Work:**

- ▶ You have to work individually for weeks 13-17 [Auto-Marked]
- ▶ You have to work within your pair for weeks 19-21 [CW submission in Wk 23]

# Tasks

- ▶ Next Lab (Week 13): Introduction to Jupyter Notebook I
  - ▶ Sheet on unit web page
- ▶ Next Problem Class (Thur 2-3): Data Acquisition
  - ▶ Prepare your answers in advance