浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

ADS2 Lecture 1 Introduction

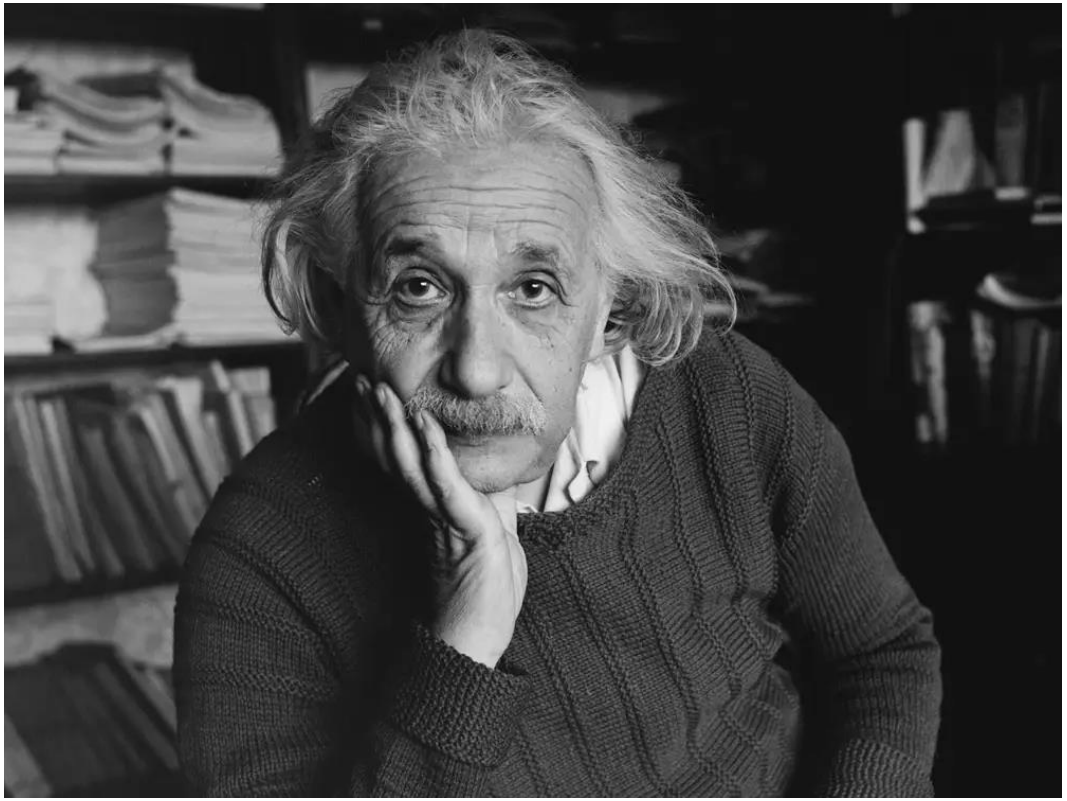
Dr Duncan MacGregor duncan.macgregor@ed.ac.uk

Semester 1, Week 1

2023-24

Well done for making it into year 2 of BMI

You're a scientist!



What now?!

We're going to turn you into a Data Scientist

Applied Data Science 2

Objectives for this week

- Learn about how this course is organised
- Think about what data science is, and how to do it
- Refresh your knowledge of R

What is data science?

Discuss: What is the difference between data and information?

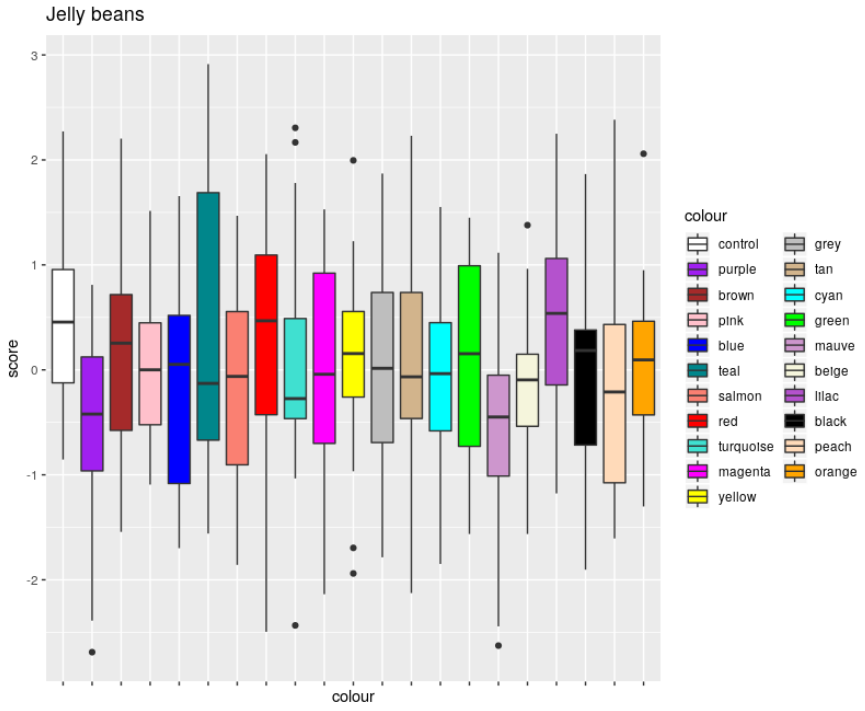
What will you be doing in ADS2?

In practise, data science is about statistics

You learned some statistics and programming in 1st year

Here you'll combine and **apply** these skills

Understand probability and statistics through running computer simulations



Learning Objectives for ADS2

After taking this course, you will be able to:

- Critically evaluate statistical representations in the scientific literature, as well as popular media
- Describe common methods for statistical inference and hypothesis testing, understand what data sets they can be applied to, and perform and interpret common hypothesis tests
- Understand the components of a dataset, handle and prepare raw data for further analysis, and display and describe datasets in meaningful ways, while considering ethical implications of data gathering, storage, analysis, and presentation
- Understand the probabilistic underpinnings of frequentist and Bayesian statistics
- Name and describe common Machine Learning methods and implement simple machine learning tasks

Learning Objectives for ADS2

After taking this course, you will be able to:

- Critically evaluate statistical representations in the scientific literature, as well as popular media
- Describe common methods for statistical inference and hypothesis testing, understand what data sets they can be applied to, and perform and interpret common hypothesis tests
- Understand the components of a dataset, handle and prepare raw data for further analysis, and display and describe datasets in meaningful ways, while considering ethical implications of data gathering, storage, analysis, and presentation
- Understand the probabilistic underpinnings of frequentist and Bayesian statistics
- Name and describe common Machine Learning methods and implement simple machine learning tasks

Discuss: Which of these do you have some experience with?

Course format

Weekly format

- 1 Lecture
- 1 Practical or Tutorial
- Weekly problem set (optional)
- Weekly student hour (optional)

Take note

- Some weeks are different (due to holidays)
- Rooms vary with content, please double-check!

Computers

You should bring your own computer. If you don't have one (or if it breaks down), contact course organisers as soon as possible.

Assessment

Summative (graded):

- Open-book timed coding challenge, semester 1 (30%)

  Semester 1 exam period

- Data analysis group project (30 %)

  Deadline: 12 April 2024, noon

- Open-book timed coding challenge, semester 2 (40 %)

  Semester 2 exam period

Formative (for practise):

- Weekly problem set: mixture of maths, coding and discussion questions; should take around 1 hour to complete. Not graded, but notes will be provided at the end of the week
- Practice open-book coding challenge, semester 1

Credits for ADS2 and your UoE/ZJE degrees

You require 55 ZJU credits* to progress to Y3.
You require 240 UoE credits* to progress to Y3.

* credits accumulated in Y1 and Y2

This course is worth

5 ZJU credits

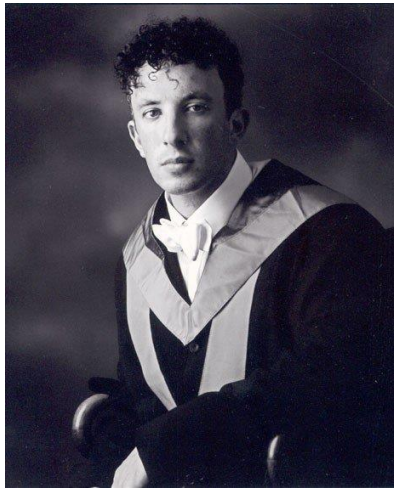
20 UoE credits

This course contributes 3.3% to the final ZJU GPA (3.6% for international students).

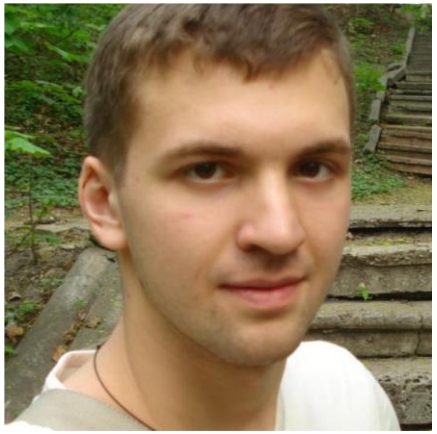
Your final UoE mark and classification is calculated from course marks in years 3 and 4. Marks for this course will still be included in your UoE degree transcript.

Course staff

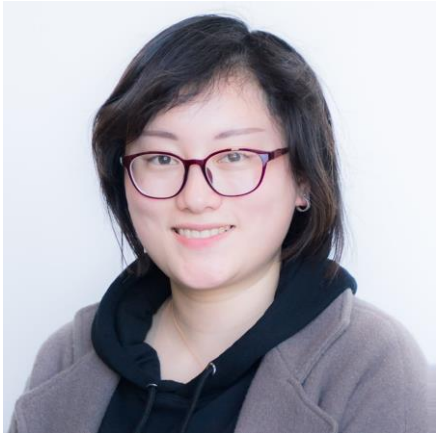
Course management



Dr Duncan MacGregor
Course Organiser
UoE



Dr Dmytro Shytikov
Deputy CO
ZJUE



Cheryl (Yanhui) Chen
Course Admin
ZJUE

Teaching staff

- Teaching staff from ZJE, ZJU and UoE
- Check weekly schedule and materials for details

Communication

Discuss

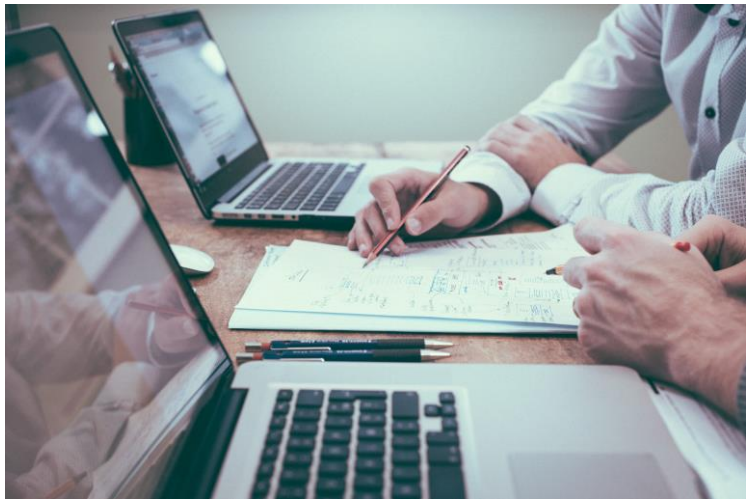
- Where do you get information about the course?
- Who can help you with problems?
- Who should you complain to?

Feedback from previous years

- We need a refresher on R from 1st year
  → Using the extra week this year to do a refresher on R today, and in the practical

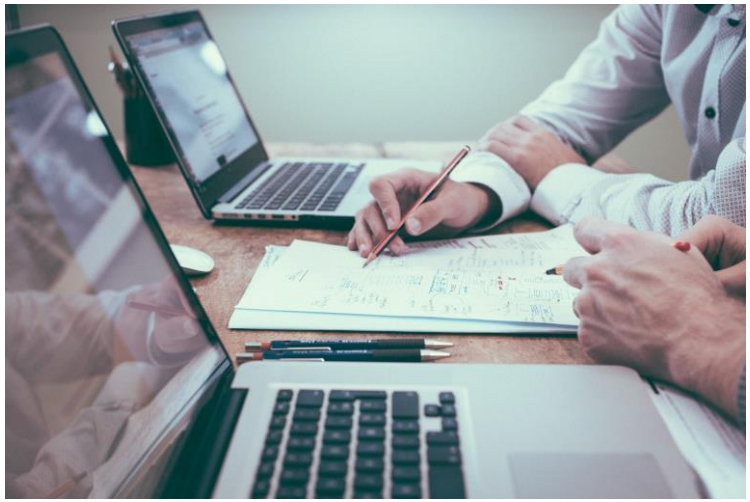
- You are not posting complete solutions to problem sets
  → Yes. Because…

Why don't we post complete solutions to problem sets?



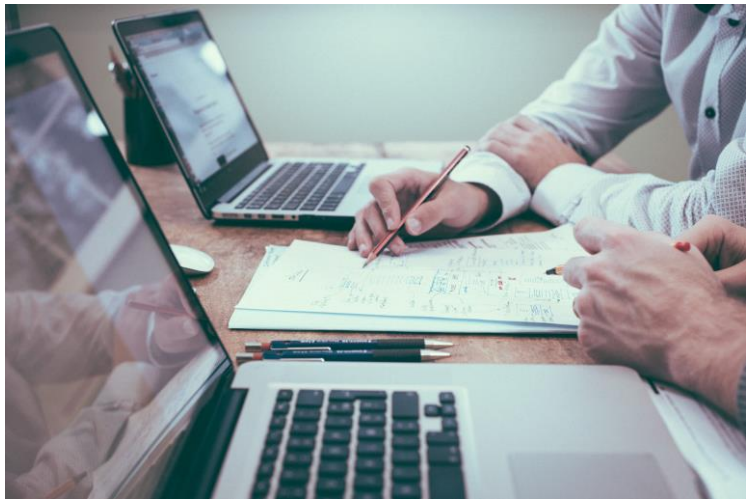
- Because the process is more important than the end result

Why don't we post complete solutions to problem sets?



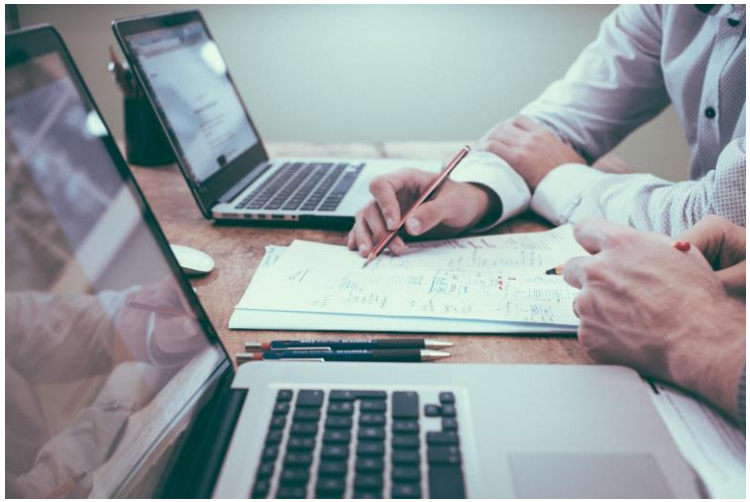
- Because the process is more important than the end result
- Because it's about learning to problem-solve, not about memorising the “correct” solution

Why don't we post complete solutions to problem sets?



- Because the process is more important than the end result
- Because it's about learning to problem-solve, not about memorising the “correct” solution
- Because there will be no correct answer sheets in your job

Why don't we post complete solutions to problem sets?



- Because the process is more important than the end result
- Because it's about learning to problem-solve, not about memorising the “correct” solution
- Because there will be no correct answer sheets in your job
- Because your solution may be better than mine!

Refresher on R

Why do we use R?

- quick to code
- popular - so lots of tools built for it
- good visualisation tools
- good stats tools

Refresher on R

Data types in R

- basic types – numeric, character, and logical (TRUE or FALSE)
- vectors – sequence or set of data of single type
- matrix – two-dimensional vector
- data frame – table that can contain data of multiple types

Refresher on R

```
> myvector = c(1,2,3,4,5)
> print(myvector)
[1] 1 2 3 4 5
> class(myvector)
[1] "numeric"
> myvec = myvector[1:3]
> print(myvec)
[1] 1 2 3
> myvecsquare = myvec * myvec
> print(myvecsquare)
[1] 1 4 9
>
```

Refresher on R

Data Frames

- data frames contain several vectors, which can be different types
- each vector can have a label

```
> codes = c(2,4,3,7,3,8,5,8,9,0)
> colours = c('red', 'blue', 'green', 'purple', 'white',
 'blue', 'red', 'green', 'black', 'orange')
> data = data.frame(codes, colours)
> data
   codes colours
1      2     red
2      4    blue
3      3   green
4      7  purple
5      3   white
6      8    blue
7      5     red
8      8   green
9      9   black
10     0  orange
```

Refresher on R

Data Frames

- data frames contain several vectors, which can be different types
- each vector can have a label

```
> data[data$colours == 'red' | data$colours == 'blue',]
  codes colours
1     2     red
2     4    blue
6     8    blue
7     5     red
>
```

Refresher on R

Data Frames

- data frames contain several vectors, which can be different types
- each vector can have a label

```
> data[data$colours == 'red' | data$colours == 'blue',]
  codes colours
1     2     red
2     4    blue
6     8    blue
7     5     red
>
```

- data frames are useful for packaging your data
- used for data import and for data plotting

Refresher on R

Data Import

- In ADS2 mostly use data stored in .csv files
- import using `read.csv()`

```
> data = read.csv("datafile.csv")
> head(data)
  X2   blue
1  4    red
2  5 orange
3  6 purple
4  7  black
5  3 yellow
6  2   grey
> data = read.csv("datafile.csv", header=FALSE)
> head(data)
  V1     V2
1  2   blue
2  4    red
3  5 orange
4  6 purple
5  7  black
6  3 yellow
```

Any questions?

- Talk to each other
- Use the discussion board!
- Email me duncan.macgregor@ed.ac.uk

Objectives for this week

- Learn about how this course is organised
- Think about what data science is, and how to do it
- Refresh your knowledge of R