Lecture 1:
Introducing CS5702 - Modern Data

Prof. Martin Shepperd

martin.shepperd@brunel.ac.uk

Lecture Poll

Please use your mobile or computer to go to:

pollev.com/mshepperd

We will need it in a few minutes. Thanks!

Welcome

This is a great time to study data science, statistics and machine learning.

Agenda

1. Module overview
2. Teaching approach and resources
3. What is R and why do data scientists use it?
4. R basics
5. Getting help
6. Week 1 goals
7. Extension question and study

1. Module overview

Meet the team

— Professor Martin Shepperd (module leader)
— Professor Xiaohui (Hui) Liu (support lecturer)
— Ziyan (Leo) Fu (GTA)
— Yu Cao (GTA)
— Namir Oud (GTA)
— Matia Ghafourian (GTA)
— Jingzhong Fang (GTA)
— Nchongmaje Ndipenoch (GTA)

Lecture protocol

1. I will start at **1105 prompt**; please be ready
2. Be aware, lectures will be recorded
3. Feel free to ask **questions** as we go along or ...
4. ... ask during in a question gap (2-3 per lecture)
5. Be **considerate** of others (fellow students and me) and don't chat.

Thanks!

What is data?

Wisdom of crowds

Source: Kozalov and Valkov, https://

What's Modern Data about?

To provide an introduction to **data management** and **exploration**. ... appreciation of the richness and availability of different data sources ... **techniques, methods and processes** for modern data analysis.

— Study Guide

Data science is ...

Wisdom of crowds

Source: Kozalov and Valkov, https://

Data science is ...

Data science is an exciting discipline that allows you to turn **raw** data into **understanding**, **insight**, and **knowledge**.

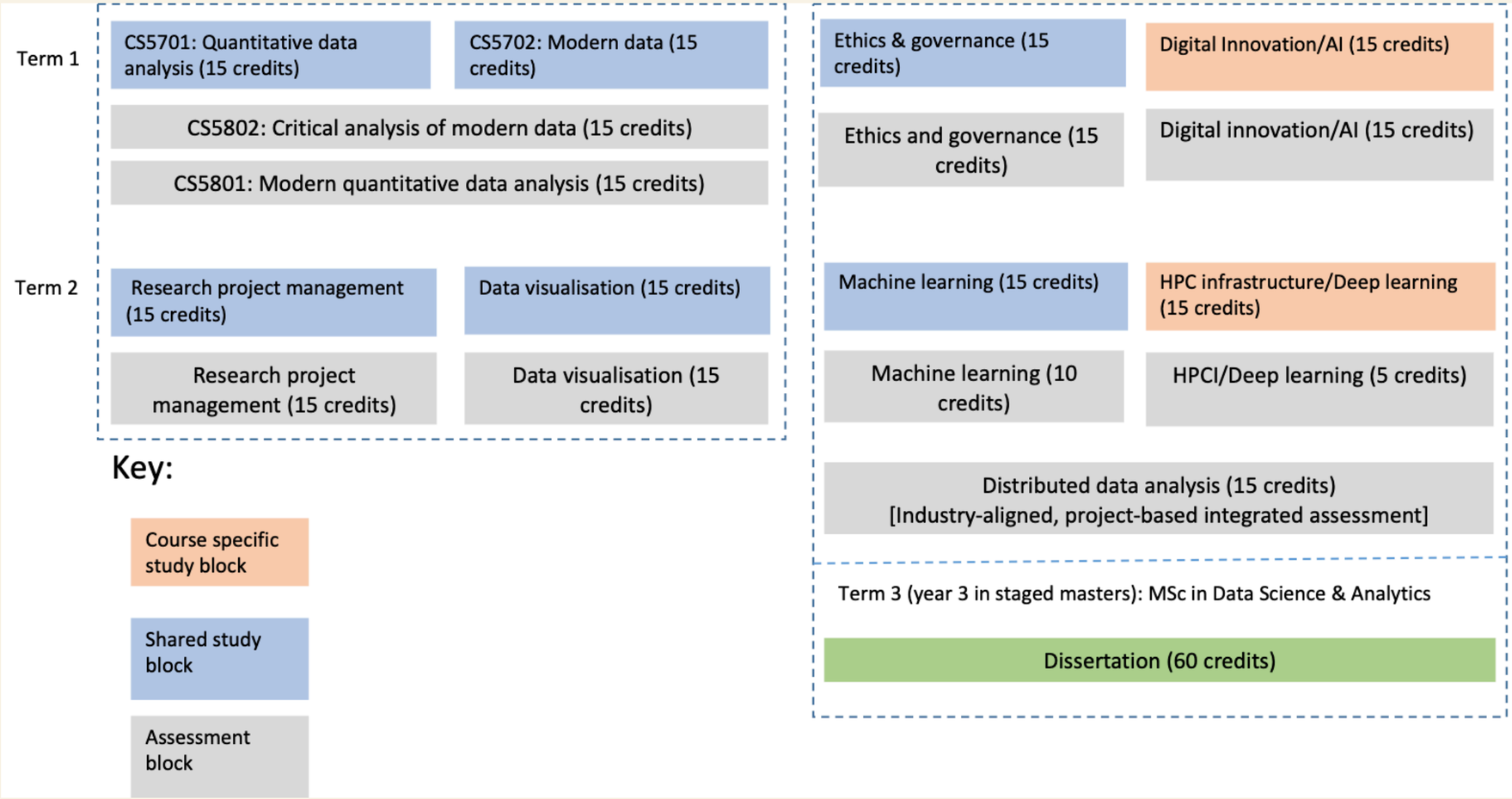
— Hadley Wickham and Garrett Grolemund

Opportunities

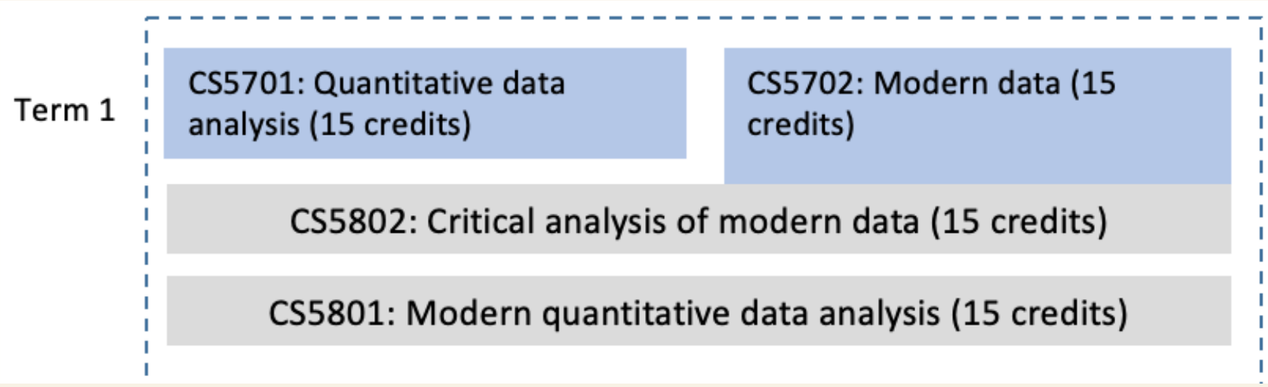
... there is more data and richer data available than ever before, coupled with more and more powerful analysis tools.

... incredible **opportunities** to collect, clean, merge, analyse and visualise data both for **good**, and for less good purposes.

Course structure



Module structure



Week structure

Week	Lecture Topic	Lecturer	Labs
1	1. Module overview, motivating examples, the R ecosystem	MS	Joint with QDA: What is data science? What can we do with R? Basic R and descriptive stats.
2	2. The richness of data: structured and unstructured	MS	Analysing the module survey.
3	3. Engineering or hacking? Readable code and reproducible data analysis	MS	Finding and importing data. Simple visualisation of time series, smoothing.
4	4. Exploratory data analysis (EDA) and visualisation	MS	Exploring data: more complex questions (salary and gender analysis)
5	5. Data quality, cleaning and imputation	MS	Cleaning data set examples. Imputation with R packages.
6	6. Presenting data effectively	MS	Interactive data: R and shiny
7	7. Processing and analysing text	MS	Word clouds, etc
8	8. Data Science and Machine Learning: an introduction	XL	Predictive modelling and evaluation in R.
9	9. Data for good	XL	Coursework (CS5801) surgery
10-12	10. Exam (CS5802) revision	MS/XL	
13	Exam week	-	-

2. Teaching approach and resources

- Practical
- Lecture and lab
- Use R and RStudio
- Being organised
  - Decide on your personal to-do / checklist system **now**!
  - Keep up to date
  - Weekly checklists (as text files) can be grabbed from here

Learning Resources:

— Brightspace VLE

— The **"Modern Data"** interactive book

— Worksheets and linked files from GitHub

— Quizzes

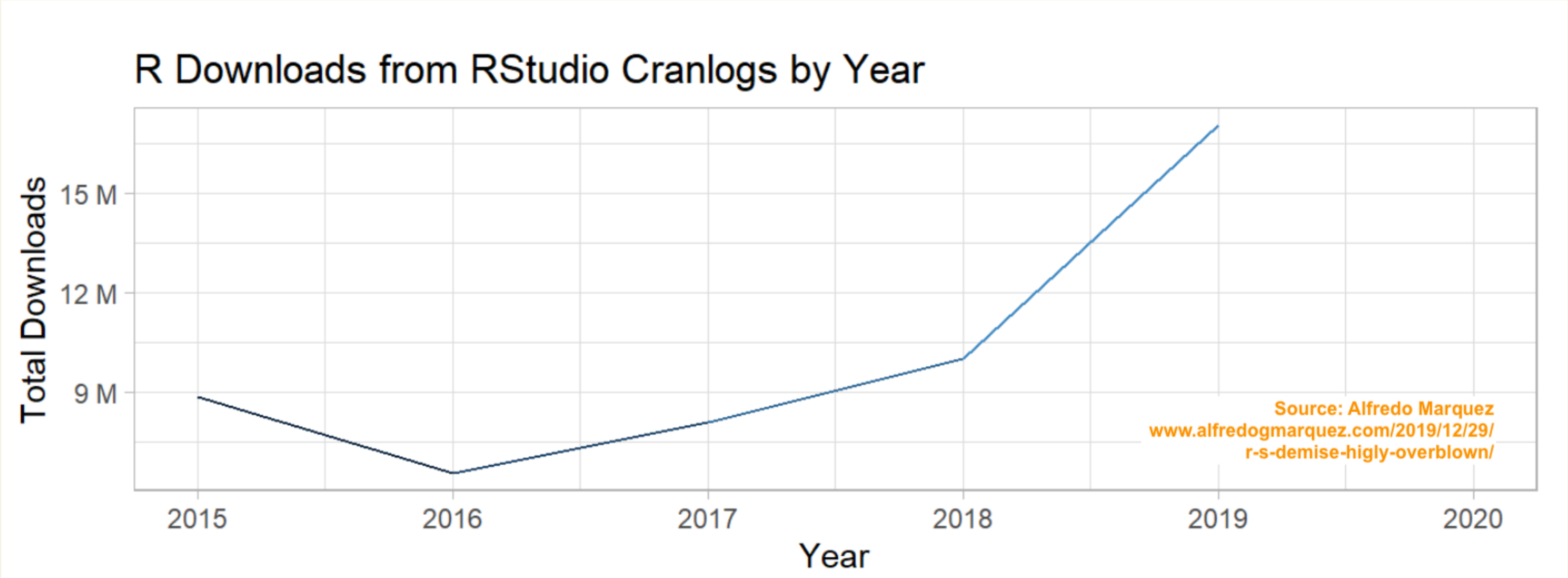
— Reading list and references (also your own research)

3. What is R and why do data scientists use it?

R is an open, purpose-designed, highly-extensible, statistical and data analysis programming language.

R Advantages

- designed by statisticians
- powerful data handling, wrangling and storage capabilies
- flexible graphical facilities
- integrates with machine learning e.g., TensorFlow etc
- interactive dashboards
- large, open community
- easy integration with e.g., C, C++, FORTRAN
- widely used by researchers



Source: www.alfredogmarquez.com/2019/12/29/r-s-demise-higly-overblown/

4. R basics

As a **prerequisite** you should have completed the Getting Ready Chapter, in particular to have **installed** R and RStudio and run some simple R test examples.

R and variables

A variable is a named container for information and this information can be set, modified or referenced.

```
# This R code creates three different variables

numericVariable <- 10
stringVariable <- "Hello world!"
logicVariable <- TRUE
```

R infers the **data type** from what you assign. This is called implicit typing.

Data types

The data type is an attribute of a variable which tells the R interpreter how we intend to use the data.

— defines the operations that can be done

— the meaning of the data

— limits to values that can be stored, e.g., if the type is logical, only TRUE and FALSE

Simple data types in R

— numeric or floating point

— character or character string (if 2+ in length)

— logical (TRUE or FALSE)

Manipulating variables

```
# Initialise (or overwrite if it already exists) y to 5.3
y <- 5.3

# Multiply y by 13
y <- y * 13

# Display y
y
[1] 68.9
```

Useful complex data types

- **vector**: multiple instances of the same type
  - see Modern Data book
- **data frame**: multiple instances of different types
  - see Modern Data book

Creating and using vectors

— So far mainly focused on **atomic** variables.

— Often useful to store/analyse multiple instances e.g., the height of all the people in a sample.

— Use a vector of the same type of atomic variables

A simple vector in R

```
# The c() function *combines* elements into a vector
sampleHeights <- c(168, 176, 170)
is.vector(sampleHeights)
[1] TRUE

sampleHeights[2]
[1] 176
sampleHeights
[1] 168 176 170
```

A data frame in R

— Is a 2-dimensional structure

— A workhorse for the data analyst

— Multiple data types e.g., numeric and character

— Sometimes referred to as 'rectangular' data because each column is the same length (a special case of a List)

— Similar(ish) to a spreadsheet

A simple data frame in R

```
name <- c('Suzie Smith','Ahmed Abbas','Atilla The Hun')
salary <- c(21000, 23400, 26800)
employmentDF <- data.frame(name, salary)

# Show the top 6 rows of the dataframe
head(employmentDF)
            name salary
1    Suzie Smith  21000
2    Ahmed Abbas  23400
3 Atilla The Hun  26800
```

The View() function

```
# Note View has an upper case 'V'
View(employmentDF)
```

Variables

Observations

	name	salary
1	Suzie Smith	21000
2	Ahmed Abbas	23400
3	Atilla The Hun	26800

For more details on R basics

— Modern Data book

— also Kabacoff[1]

— the lab worksheets

[1] Kabacoff, R. (2015). R in Action: Data Analysis and Graphics With R (2nd ed.). Manning Publications.

5. Getting help ...

1. find/read the relevant cheatsheet
2. perspiration e.g., see this five step approach
3. talk it over with a fellow student
4. module **FAQs** on Brightspace
5. Stack overflow
6. ask a member of the course team

For more suggestions visit the subsection 0.2 "vi) Learn how to get help" in the Modern Data book.

6. Week 1 goals

By the end of this week you should:
[ ] completed the **Week 0** Getting Ready chapter
[ ] completed the **Week 1** Introduction chapter
[ ] have an appreciation of the background and development of R
[ ] understand the main components of the R ecosystem
[ ] be able to write, execute, save and organise simple R programs
[ ] to be confident in using RStudio for basic coding tasks

7. Extension activity

Read Provost and Fawcett[2] (it's only 8 pages) and determine what are the fundamental concepts of Data Science. Then rank them in order of importance.

[2] Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51--59. Access via Google Scholar

References

— Kabacoff, R. (2015). R in Action: Data Analysis and Graphics With R (2nd ed.). Manning Publications.

— Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51--59.

— Wickham, H., & Grolemund, G. (2018). R for data science. O'Reilly Media, Inc.