Week 1: Introduction to the course!

Charlotte Hadley

Topics for today

1. Getting to know me
2. Getting to know you
3. Understanding the goals of the course
4. Why are we learning R?

Getting to know me

Charlotte Hadley

Please call me Charlotte or Charlie. My pronouns are she/her.

I don’t have a doctorate so it’s inaccurate to call me “Dr Hadley” or “Professor Hadley”.

If you absolutely must call be by a title your only option is “Miss Hadley” but I’d really prefer you didn’t.

What do I do?

I'm currently a full-time independent data science consultant and trainer through Visible Data Ltd.

Academic background

- 2015-2019: I worked at University of Oxford as a Research Support Officer and built the Interactive Data Network.
- 2010-2012: I began but ultimately quit a PhD in biomineralisation.
- 2006-2010: MPhys and BSc in physics from University of Leeds with a focus on biophysics

Industry background

- 2016-2022: I've been consulting and delivering training in industry both in-person and via LinkedIn Learning..
- 2012-2015: I was a senior consultant at Wolfram Research

Getting to know you

Can you tell me a little bit about yourselves?

- Your name
- Your pronouns
- Where you're from
- What have you studied before this Masters?

Introducing Etherpad

We’re going to use “Etherpads” during lectures and workshops so I can ask you questions and share code.

Here’s the link for today’s pad: bit.ly/eng7218_week-1_lecture-slides

How does this course work?

Course Structure

We have 11 weeks of teaching and each week has:

- an 2 hour lecture
  - In the lectures I will introduce topics and theory
  - Two hours is too long for most people to pay 100% attention.
    I'll insert breaks and experiment with other ways to break up the lectures.
- an 2 hour workshop
  - These workshops are **crucial**[1] for you to do well in the workshops.
  - The workshops will mix together guided and exploratory work.

[1] Of course, the real world means that 100% attendance is an unrealistic expectation. **Please** do get in touch with me if you miss workshops or lectures and I'll help as much as I can.

Laptops

This is a practical data science course - please think of the whole course as a lab.

Please **bring your laptop to every lecture and workshop**.

If this isn't practical for you **please** speak to me[1] and I will find a solution.

In today's workshop I will take you through all of the steps necessary to setup your machine to use R and RStudio.

Asking questions

> If you have a question during the lectures or workshops **please ask the question** when you think of them.
>
> There's no such thing as a 'pointless question' - particularly as in this course you'll be learning data science and using **3+** different programming languages[1].

If you want to ask me questions outside of our sessions please email me **charlotte.hadley@bcu.ac.uk**.

[1] You'll be learning R. But in order to use RMarkdown you will also need to use Markdown and YAML. You'll also likely end up using a little bit of HTML and CSS.

Course materials and website

I'd like to ask you **not** to read ahead in the lecture slides or workshops.

> This is because there are some exercises I'd like you to try in week N that are solved for you in week N+1.

As with all BCU modules you can find the lecture notes on Moodle.

However!

This course has a dedicated website (eng7218.netlify.app) that contains more materials than the Moodle page.

Course assessment

This module is 100% assessed with coursework that must be submitted **before 12:00 on Friday, 13 January 2023**.

Part of the coursework will require you to learn to use R and RMarkdown.

> I want to talk about the *goals* of this course before giving more details about how the assessment will work.

Understanding the goals of this course

So you know how to succeed

Course goals

I want you to succeed in this course.

I want you to succeed **after** this course in your career and/or research.

How to succeed in this course:

I want to break down each of these in turn:

- Feel confident in lectures
- Feel confident in workshops
- Feel confident in designing (and reading) data visualisations
- Feel confident in the assessment

Feel confident in lectures

For you to feel confident in the lectures I want you to:

- Ask questions if you feel lost or don’t understand something.
- Understand **why** something is being taught in the context of the module goals.

Module goals

- Demonstrate a systematic understanding of the principles and approaches in data science to be used in healthcare.
- Critically appraise the key considerations for using healthcare data including ethics, information governance and security issues relevant to health data science.
- Apply knowledge of the R language to read and wrangle healthcare datasets into the R environment for analysis.
- Design data visualisations and tables with the R language to communicate properties of datasets and the conclusions of data analyses.

Module goals

- Demonstrate a systematic understanding of the principles and approaches in data science to be used in healthcare.
- Critically appraise the key considerations for using healthcare data including ethics, information governance and security issues relevant to health data science.
- Apply knowledge of the R language to read and wrangle healthcare datasets into the R environment for analysis.
- Design data visualisations and tables with the R language to communicate properties of datasets and the conclusions of data analyses.

Module goals

- Demonstrate a systematic understanding of the principles and approaches in data science to be used in healthcare.
- Critically appraise the key considerations for using healthcare data including ethics, information governance and security issues relevant to health data science.
- Apply knowledge of the R language to read and wrangle healthcare datasets into the R environment for analysis.
- Design data visualisations and tables with the R language to communicate properties of datasets and the conclusions of data analyses.

Module goals

- Demonstrate a systematic understanding of the principles and approaches in data science to be used in healthcare.
- Critically appraise the key considerations for using healthcare data including ethics, information governance and security issues relevant to health data science.
- Apply knowledge of the R language to read and wrangle healthcare datasets into the R environment for analysis.
- Design data visualisations and tables with the R language to communicate properties of datasets and the conclusions of data analyses.

Feel confident in workshops

The workshops will run a little differently to the lectures.

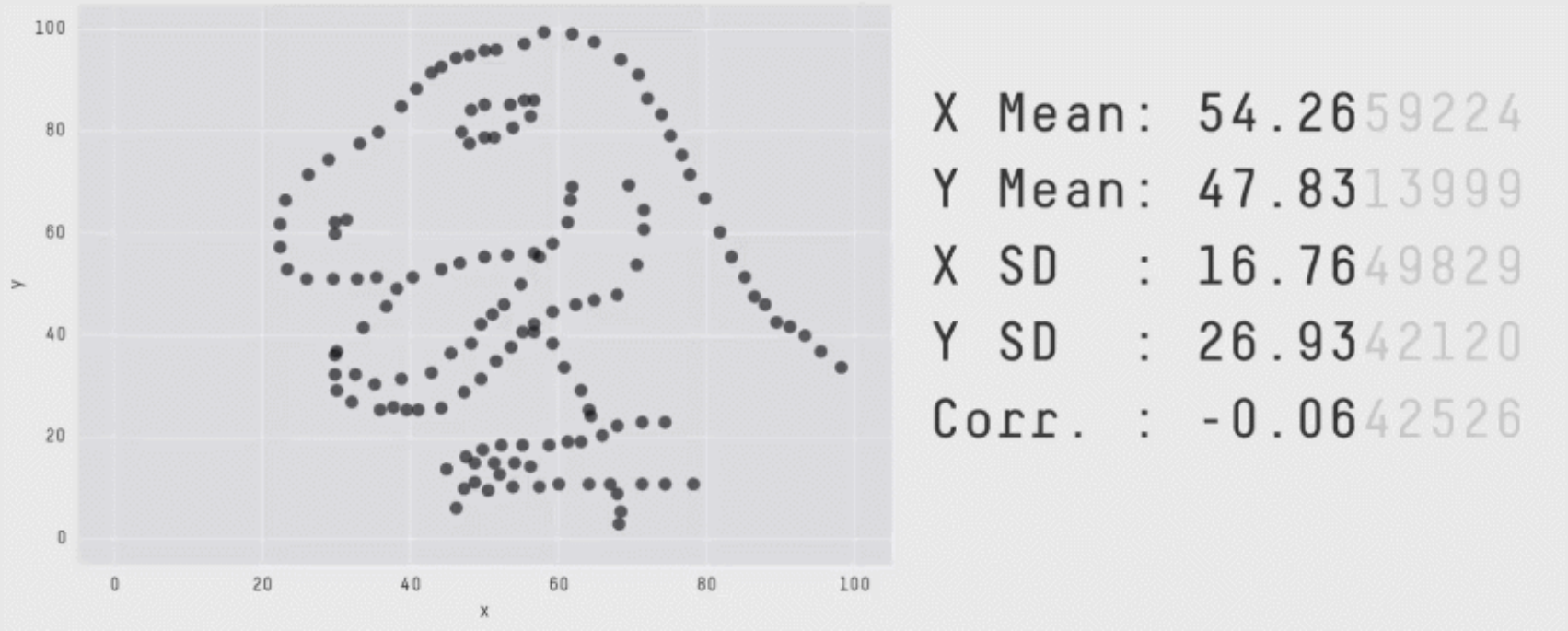
As we progress through the weeks we'll move away from you completing very specific tasks to more open eneded goals.

- Ask questions if you feel lost or don't understand something.
- Become confident in figuring out why your code doesn't work and problem solving it.
- Become confident in exploring new ideas, particularly different ways to explore and visualise data.

Feel confident in reading and designing data visualisations

Week 4 is meant to be when we focus on data visualisation.

However, we will start using data visualisations right from the beginning of the course.



Data visualisation produced by Matejka[1].

Feeling confident in the assessment

How the assessment is designed

There are two very different components to the assessment:

- the written component which requires you to explain concepts and critically analyse case studies.
    - You will use Microsoft Word (or your word processor of choice) for this part of the assessment.
- the coding component which requires you to write code to read, wrangle and visualise survey data.
    - You will use R and RMarkdown for this part of the assessment.

How the assessment is designed

There are 3 different sections in the assessment:

Section	Type of assessment
Part A) Open health data and anonymisation	Written component
Part B) Algorithms and health data	Written component
Part C) Analyse and visualize results from a health data survey	Coding component

The module website provides more details about these sections.

The colour coded timetable demonstrates which lectures and workshops will help you with each section.

Written component of the assessment

You will need to use *case studies* to demonstrate your understanding of concepts introduced in the lectures.

The course website's case studies section includes all the case studies we will cover in the course.

You are also strongly encouraged to research your own case studies (and please share them with the cohort!).

Coding component of the assessment

You will be making use of a tool called RMarkdown to answer the last part of the courswork

> Part C) Analyse and visualize results from a health data survey

RMarkdown allows you to write reports, presentations and even entire websites[1]. It's a very powerful tool that is widely used in industry and academia.

The thing that's powerful about it is that you can include (and run) R code in your documents.

[1] These slides and the entire module website is written with RMarkdown documents.

Coding component of the assessment

In the very first workshop I will thoroughly introduce you to R and RMarkdown.

In every subsequent workshop we will use RMarkdown.

You should have sufficient practice and expertise to answer the coding component of the coursework.

I will check in with you all about your confidence with the assessment in Week 8.

There is also a template GitHub repository that you can use for structuring your assessment.

How to succeed in this course:

Now I've covered these in more detail - do you have any questions?

- Feel confident in lectures
- Feel confident in workshops
- Feel confident in designing (and reading) data visualisations
- Feel confident in the assessment

Palate cleanser

Check out one of my favourite data visualisations ever: bit.ly/3QISech

How to succeed after this course:

Now I've covered these in more detail - do you have any questions?

- Understand there are data science careers in both academia and industry
- Practice reproducible research from now onwards
- Appreciate and make use of open data standards where possible
- Protect people by protecting data
- Require Fairness, Accountability and Transparency for algorithms

Data Science Careers

What do you folks want to do in the future?

What do we mean by data science?

There's lots of discourse about the difference between "data science jobs" and "data analysis jobs" but most of this is gate keeping.

For our purposes:

> We successfully do data science when we write reproducible code that reads and analyses code in such a way that we can others stories about the data.
>
> Data science might involve statistics, but it does not necessarily require it.

> Reproducible code means that other people can run the code we write on their machines.

Academic careers using data science

Researchers across **all divisions and departments** use data science:

- Visualising close reading in poetry (and elsewhere)
- Deciphering lost languages
- Crowd sourced projects on Zooniverse,
  - Transcribing weather data from logbooks.
  - Classifying baby noises to explore language development

Of course, data science is being used prolifically in the collection and analysis of healthcare data.

Traditional academic research job route

1. PhD.
   - In the UK these are usually funded for 3 years, but funding *might* be extendable
2. Several “post doc” positions.
   - “Post doc” positions are difficult to define but are almost always fixed term contracts for 2-3 years.
3. Lecturer positions
   - Permanent jobs
   - These might be anywhere from 100% research to 100% teaching

> However there are too many PhD vacancies with too few research positions.
>
> doi.org/10.1038/d41586-019-03439-x

Non-traditional academic research route

There are many non-traditional routes into academia.

Research Software Engineering (RSE) is an excellent non-traditional route for folks with a data science background.

The RSE community is responsible for designing, building and maintaining the code/software that underpins academic research

> This is important because code/software is not traditionally celebrated or considered in the academic publishing industry

The Society of Research Software Engineering provides resources and career opportunities.

There are **many** folks in academic research positions that do not have PhD and/or post docs.

Data science careers in industry

I highly recommend the Build a Career in Data Science book by Emily Robinson and Jacqueline Nolis. Both these authors are part of the R community.

- A lot of the advice from the book is available in a blogpost from Emily Robinson.

I also recommend this great thread from Jesse Mostipak from RStudio (and previously Kaggle).

There's lots of other great advice out there.

Practice reproducible ~~researeh~~ code from now onwards

Assume all code you write might be useful for someone else

I've mentioned *reproducibile code* several times and described it as code that other people can use.

It's actually quite difficult to make code reproducible half-way through a project - **always start with best practices**.

… reproducible code makes for an awesome portfolio

In week 2's lecture I'm going to introduce GitHub and recommend you use it as a portfolio for future job applications.

> You do **not** need to use GitHub in the assessment for this module.

Appreciate and make use of open data standards where possible

Open Data is good for everyone

We'll talk a lot about Open Data in Week 2.

I want to encourage you to consider using open data standards **where possible** as it can benefit:

- You.

> 25.36% (± 1.07%) higher citation impact [for articles linking to a data repository]
>
> 10.1371/journal.pone.0230416

- Other researchers
- Society

Protect people by protecting data

We'll be looking at data anonymisation in lots of detail in Week 6's lecture.

Whenever we're working with data about people (or groups) we **must** keep in mind protecting their identities.

> Privacy itself is valuable.
>
> We need to protect individuals and groups from harm that could result from private data is published.
>
> There are legal requirements for data protection, including GDPR in the UK.

In Week 2 when I talk about Open Data I will also mention minimum requirements for privacy.

Require Fairness, Accountability and Transparency for algorithms

In Week 7 we will look at the ethics in algorithms which requires us to consider 3 different concepts:

- Fairness: Is the *training data* behind the data fair (does it look at what we think it looks at)?.
- Accountability: What are the impacts and secondary consequences of applying an algorithm?
- Transparency: Understanding of how algorithms are used in decision making.

> If you help develop an algorithm you're intrinsically connected with how that algorithm is applied.[1]

[1] For clarity, I'm not saying that you are **responsible** for how the algorithm is applied. By ensuring documented fairness in algorithm development this will aide in the future accountability and transparency of the algorithm.

How to succeed after this course:

Now I've covered these in more detail - do you have any questions?

- Understand there are data science careers in both academia and industry
- Practice reproducible research from now onwards
- Appreciate and make use of open data standards where possible
- Protect people by protecting data
- Require Fairness, Accountability and Transparency for algorithms

Why are we learning R?

… why are we learning a programming language at all?

Programming vs GUI

You might expect me to bash all GUI based tools at the point, but there are incredibly powerful tools out there.

References

1. Matejka, J. & Fitzmaurice, G. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* 1290–1294 (Association for Computing Machinery, 2017). doi:10.1145/3025453.3025912.