A Data Science Course

Dan MacLean

2021-09-20

Contents

Chapter 1

About this course

Welcome to the handbook for the data science component of the TSL M.Sc in Global Plant Health.

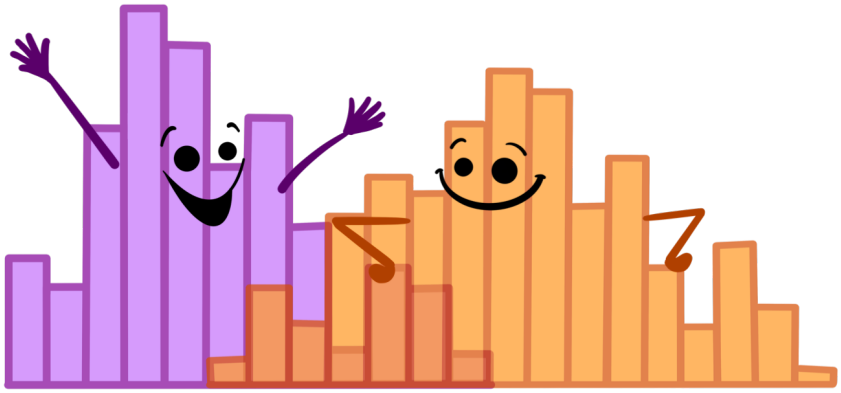

Figure 1.1: Artwork by AllisonHorst

In here you will find links to the written content for each of the separate topics covered as well as the descriptions of the course assessments and projects.

1.1 Course Delivery

Data Science is a set of practical research skills grounded in statistics and computer science. Learning a practical skill requires practice! So this course is not a lecture course, it is a 'flipped' classroom course with a very strong practical component. In a flipped classroom the onus is on the student to lead their

work and practice with the provided materials prior to contact and discussion time with the wider learning group and group mentors. To get the most out of this course you must read the relevant online material before you come to the contact sessions. Contact time will then be an opportunity to discuss the materials and any problems arising with the group and the teacher and to practice and problem solve with others. The aim is that by the end of the course you will have a strong practical grounding in applying data science approaches to research problems that will enhance the biology that you are doing.

1.2 Online materials

The rest of this handbook outlines the online materials provided to you for this course, broken down by the separate topics we will cover. The materials are a mixture of self-led tutorials and interactive challenges or problems to solve.

1.3 Installation of software and tools

If you are using a TSL laptop specifically given to you for the MSc then all the software and tools you will need should already be installed on that machine. Each topic presents installation instructions for those not accessing the materials as part of the MSc or for installation on other or personal machines should they be needed.

1.4 Course Topics

We will cover the following topics in data science

1. Introduction to Genomics
2. Data Exploration and Visualisation
3. Understanding Statistics With Linear Models
4. Introduction to Non-Frequentist Statistics
5. Introduction to Machine Learning
6. Beginning Programming
7. Literate Computation

1.5 Course Assessments

The culmination of all this learning and practice will be a written essay and a data analysis project on which you will be assessed and receive a grade. These will be set during the course and deadlines will be given to you explicitly for each. You will not be formally assessed on other projects, quizzes and challenges occurring through the course.

1.5.1 Deadlines

Important Assessment Dates

Assessment	Deadline	Return Date	Description
Formative 1 (Essay)	Wednesday, December 1, 2021	Wednesday, January 12, 2022	link
Formative 2 (Data)	Wednesday, December 15, 2021	Wednesday, January 26, 2022	link
Summative 1 (Essay)	Wednesday, January 12, 2022	Wednesday, February 9, 2022	link
Summative 2 (Essay)	Saturday, January 22, 2022	Tuesday, February 22, 2022	link

1.6 Schedule

Data Science Schedule 2021/2022

Date	Session	Location
Monday, October 4, 2021 13:00	Course Intro and Intro to Genomics	Room??
Monday, October 11, 2021 13:00	Intro to Genomics	Room??
Friday, October 29, 2021 13:00	Data Exploration and Visualistion	Room??
Friday, November 12, 2021 13:00	Understanding Statistics	Room??
Friday, November 26, 2021 13:00	Literate Computation	Room??
Friday, December 10, 2021 13:00	Beginning Programming	Room??
Friday, January 7, 2022 13:00	Non-Frequentist Stats	Room??
Friday, January 21, 2022 13:00	Machine Learning	Room??

Chapter 2

Why study data science?

At first thought, you may imagine that data science is not going to be of much interest to you, your desire and interest is in biology, not computers and mathematics. That may be true but data science is actually going to be one of the main technical themes throughout your entire scientific career and it will help you truly master your science and to become an informed and healthily sceptical interpreter of scientific results. Adding data science as your super power will be of benefit in lots of ways. Learning data science now isn't a tax on your time so that you can get through this M.Sc, it's a valuable investment in your skill set that will help to give you an advantage that will pay definite dividends later. Getting on with data science is like getting on with Jedi training for biologists

2.1 All experimental science is data science sooner or later

"Data Science" is a new-ish term that describes an interdisciplinary approach to the practical application of statistics and computer science concepts and tools. The term data scientist is most often applied to people who analyse large amounts of high throughput data full time and have a research or industrial interest in that but the techniques of data science are useful to all researchers, data science is useful to biologists. There is no such thing as a purely lab- or field-based biologist, once our data are collected we must use it in a computer to perform analyses, draw conclusions and draw figures. All modern scientists are at some point a data scientist, irrespective of field and training. Knowledge of data science techniques will therefore underpin everything you do as a scientist at the most fundamental level, from experimental design, data collection, analysis and interpretation of results. A scientist that lacks data science skills will find it very difficult to progress beyond the most basic levels of expertise in experimental analysis.

Figure 2.1: Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)

2.2 Biology is getting bigger

All biological disciplines including genomics and genetics, microscopy, biochemistry and proteomics generate volumes of data that are too great to handle by a scientist armed only with a clipboard and spreadsheet, no matter how hard working they are. At the heart of data science are the techniques for dealing with the biggest data sets yet generated and learning just a little about them will stand you in good stead across your entire career.

2.3 Data science methods are 21st Century realisations of the scientific method

Reproducibility and clear description of process is the keystone of the scientific method, using point and click tools to carry out analyses often do nothing to help you reach this goal but many data science techniques are built with these qualities as a primary design principle and can help to reduce your workload by helping you to do things reproducibly over and again with ease. They can help you share and report on your research and data in a way that is transparent and clear to the reader. They can also help you make the data and analysis you have generated shareable and findable by others.

Chapter 3

Introduction to Genomics

Genomics is a big field with wide use within biology. On the Global Plant Health M.Sc you are going to learn about a lot of applications of genomics technologies and data science skills underpin *all* of those applications and analysis. So in this first, foundational topic we will look at how to run a basic genomics SNP calling pipeline on the Linux command line. This will include developing the basic skills to use the computer from the command line.

3.1 The materials

As a source for this we will follow the excellent Data Carpentry Genomics Curriculum https://datacarpentry.org/lessons/#genomics-workshop. We will do the following two lessons

1. Shell Genomics
2. Data Wrangling and Processing for Genomics
3. Genome, Transcriptome and Structural Databases

3.2 For you to do

Parts 1 (Shell Genomics) and 2 (Data Wrangling) will be full contact time instructor led courses. You should try reading the materials *before* the course dates - it will help a great deal.

Part 3 is mostly an interactive tutorial, please work through that *before* the scheduled discussion session in which we will go over as a group any questions and problems you may have had working through the materials.

Chapter 4

Data Exploration and Visualisation

Data exploration and visualisation is the first step in most analyses and the R statistical computing environment is a great tool to work with.

In this topic we will explore techniques in R that allow us answer research questions of our data in a way that follows consistent and straightforward principles. We will learn a data format for keeping our research data organised so that we can readily apply all sorts of useful tools to it. We will also learn a grammar of plots that will allow us to make informative and clear figures. By the end of this you will be comfortable with using R to summarise and work with data frames and create a wide range of plots.

4.1 The materials

For this topic we will use these courses

1. Using dplyr for data analysis
2. Using ggplot2 to producing quality plots

4.2 For you to do

These topics will be delivered flipped classroom-style. That means they are led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Figure 4.1: Artwork by @AllisonHorst

Chapter 5

Understanding Statistics With Linear Models

In this topic we will take a look at common statistical methods. Rather than launch into a long list of tests and conditions for use that seems arbitrary we will take advantage of the fact that they are all special cases of a simple tool called a linear model. We will learn about linear models and how to use them to carry out statistical inference. By the end of this topic you will have a clear understanding of how and why to use a linear model in R to carry out most test.

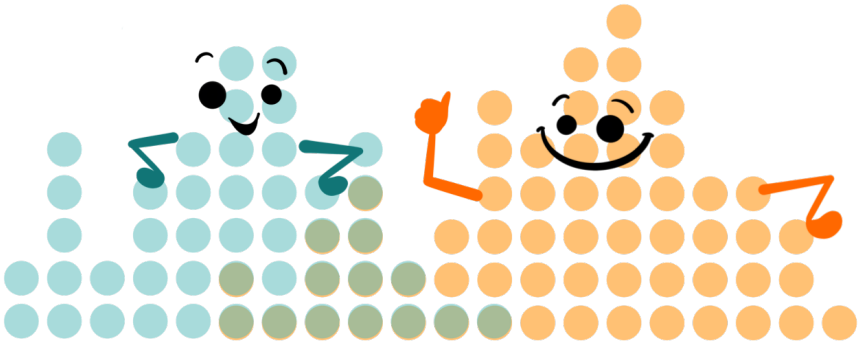

Figure 5.1: Artwork by @AllisonHorst

5.1 The materials

For this topic we will use this course

1. Understanding Statistical Thinking With Linear Models

5.2 For you to do

These topics will be delivered flipped classroom-style. That means they are led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Chapter 6

Introduction to Non-Frequentist Statistics

In this topic we will take a look at alternatives to the standard stattesting tools and models and learn how to make inferences from our data without tests. We will learn to abandon the p-value as a final arbiter of statistical correctness and see that it is not the only statistic that matters. We will look at using Bayesian inference to make comparisons of hypotheses about our data By the end of this course you will be able to use and interpret standardised effect sizes and create confidence intervals and estimation plots in R. You will be able to use an interpret Bayesian versions of some common statistical tests

6.1 The materials

For this topic we will use these courses

1. Estimation Statistics
2. Bayesian Inference with Bayes Factors

6.2 For you to do

These topics will be delivered flipped classroom-style. That means they are led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Figure 6.1: Artwork by @AllisonHorst

Chapter 7

Introduction to Machine Learning

In this topic we will look at algorithms for performing machine learning as a way of classifying data into groups of similar or dissimilar items. We will use supervised and unsupervised methods in R. We will also study and use some deep learning methods. By the end of this course you will have an understanding of how and when to use classical machine learning in your research and an appreciation of what deep learning methods can achieve and their limitations.

7.1 The materials

Introduction to Machine Learning

7.2 For you to do

These topics will be delivered flipped classroom-style. That means they are led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Chapter 8

Beginning Programming

Programming is an essential but tricky tool to master for all data scientists. Writing code comes in many guises and in this course up to now we've been writing code for specific tasks with a clear and present idea of what we want to get done. General programming concepts are often harder to grasp because of their generality, so in this topic we will study the general concepts that will allow us to program for a general problem. We will introduce the concepts and learn to use the implementations of them in the Python general purpose programming language. At the end of this topic you will be aware of the most important concepts in programming and how to use them in Python.

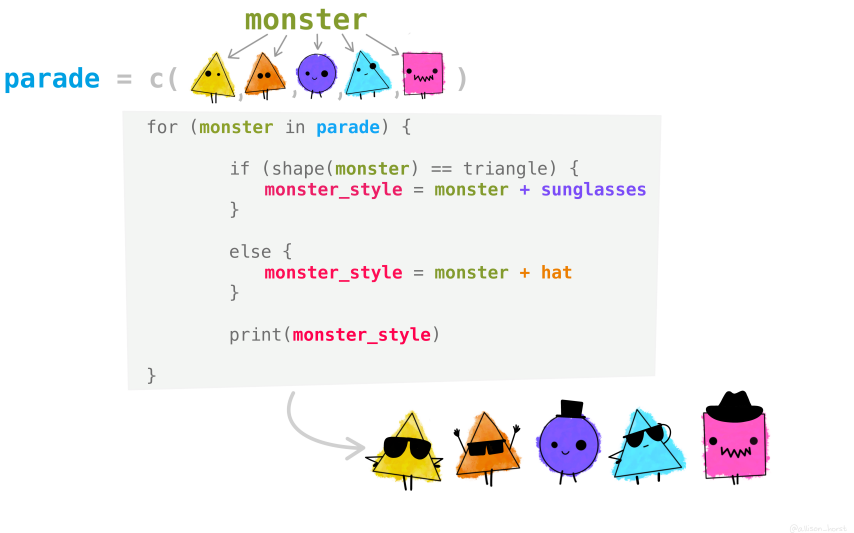

Figure 8.1: Artwork by @AllisonHorst

8.1 The materials

For this topic we will use these courses

1. A tao of programming
2. Beginning Programming with Python

8.2 For you to do

Part 1 (A tao of programming) will be delivered as a full contact instructor led-session. You should come to the scheduled session.

Part 2 (Beginning programming with Python) will be delivered flipped classroom-style. That means it is led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Chapter 9

Literate Computation

Sitting at a computer and typing commands is only of limited fun. Wrapping commands in scripts is a great way to re-do an analysis without having to type everything in again. But scripts are for computers to read, not humans, so understanding and interpreting an analysis rendered as code alone is not good for general understanding. In this course we'll look at tools and techniques for Literate Computation, a mixing of code, results and human language that results in easily understood executable and re-useable analysis documents. We will also look at tools and strategies for sharing and getting credit for your code and analysis. At the end of this topic you will have a good understanding of how to construct reproducible, reusable and readable analysis documents and how to share them online.

9.1 The materials

For this topic we will use this course:

1. Literate Computing

9.2 For you to do

These topics will be delivered flipped classroom-style. That means they are led by you and you have the responsibility for doing the reading prior to coming to the session. You will absolutely need to devote time and work through the materials on your own and bring any questions and queries that you have to the scheduled session for the topics.

Figure 9.1: Artwork by @AllisonHorst

Chapter 10

Acknowledgements

All of the artwork in this repo is by (\ ?)(https://twitter.com/allison_horst) and is licensed under a Creative Commons Attribution 4.0 International License.

Bibliography