Data Science - Global Plant Health MSc Course

Dan MacLean

November, 2023

Table of contents

About This Course

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

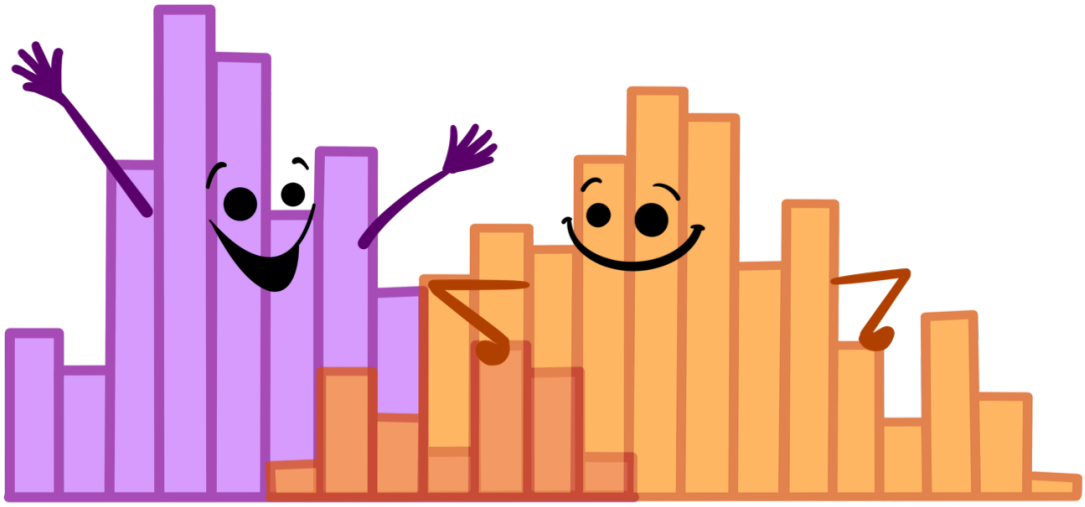

Figure 1: Artwork by AllisonHorst

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.

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.

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.

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.

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
```

Sections 1,2,3 and 7 are considered the core parts and each has guided study parts that we will use. Sections 4,5 and 6 are considered extension parts. The core will be covered first and the extension parts can be covered as we decide in the class dependent on preference and interest.

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.

Deadlines

~~Here are the course deadlines for the 2023/2024 year.~~

Deadlines not yet decided.

Schedule 2022/2023

Data Science Schedule 2022/2023

Date	Session	Location
2023-11-23 10:00:00	Course Intro and Intro to Genomics	MSc Suite
2022-11-30 10:00:00	Intro to Genomics	MSc Suite
2023-12-07 10:00:00	Intro to Genomics and Databases	MSc Suite
2023-12-12 10:00:00	Data Exploration and Visualisation Intro	MSc Suite
2023-12-15 14:00:00	Data Exploration and Visualisation Recap	MSc Suite
2024-01-09 10:00:00	Understanding Statistics Intro	MSc Suite
2024-01-12 14:00:00	Understanding Statistics Recap	MSc Suite
2024-01-16 10:00:00	More Understanding Stats/Literate Computation Intro	MSc Suite
2024-01-19 14:00:00	More Understanding Stats/Literate Computation Recap	MSc Suite
2024-01-23 10:00:00	Non-Frequentist Stats Intro	MSc Suite
2024-01-26 14:00:00	Non-Frequentist Stats Recap	MSc Suite
2024-01-30 10:00:00	Flexible Session Intro	MSc Suite
2024-02-02 14:00:00	Flexible Session Recap	MSc Suite

Why Care About 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

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.

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.

Figure 2: Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)

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.

How to Learn With This Course

Session types

The topics in this course will be delivered in one of three ways:

1. Bootcamp
2. Flipped classroom
3. Online tutorial

Each type of class has a different emphasis and approach regardless of the topic under study.

Bootcamp

The bootcamp style class is an instructor led practical session. You will be expected to follow an instructor who is teaching and live coding from the front of the class. This is not a lecture where you take notes - its a computer class where you work concurrently with the instructor as they exemplify the skills you will need to learn to use. Much more like a cookery class than a university lecture, the session proceeds with the instructor demonstrating how to do a specific computer-based skill and you try it out immediately. It is very, *very* useful to take the opportunity to read the materials for bootcamps *before* the session. Learning from this style requires you to be alert and engaged with the material in the session, not least you should put effort into joining in and being part of the group. Be prepared to make mistakes and be at peace with not knowing yet and being on a journey of learning. Sitting back and hoping to catch up with it later in your own time will put you behind and waste the opportunity of the session.

Flipped classroom

The 'flipped' classroom style here is one in which the reading is done by the student *before* the session. This gives time for reflection and understanding of the concepts such that they can be applied to solve the problem sets that come up in the session. The instructor will *not* conduct a comprehensive overview of the materials during the session, instead they will concentrate on issues arising in the application of the concepts. If you have not done the reading before the session then you will be behind from the start of the session and you will miss out on the

opportunity. Learning from this style is requires you to lead and take responsibility for the learning before class and to work to apply the theoretical knowledge in the practical session.

Online tutorial

The most traditional style of session will be the online tutorial in which you will work through an online guide with embedded quizzes and questions on a particular topic. The materials will be provided in the session. Learning from this style of session requires you to be thinking through the questions and reasoning to work out the answers with the materials provided.

Note-taking

The website of the course is connected to the online note-taking system hypothes.is. If you have an account there and are logged in you'll be able to add highlights, notes and comments that you will see in the web page itself. You can also make and join groups to share notes.

JONK

For all learning you should embrace JONK - 'Joy of Not Knowing' (Staricoff 2020). Although you are at least a Post-Grad there are still a lot of things and skills to learn and you will spend a lot of time confused. This is fine. And expected, and a necessary part of learning - if it was easy you wouldn't be learning anything. Over and over again you'll fall into the metaphorical 'Learning Pit' (Nottingham 2015) and through working at a problem sincerely with a view to understanding you'll reach real success - a good understanding of the topics you are learning about.

The dark side

Lots of people want to use data science and statistics so lots of other people publish blog posts and tutorials on those topics on the internet. A quick google will lead you to how-to's in briefer and shallower form than this course. Be wary of taking these as an answer, although they aren't wrong per se, the useful and particular ways of thinking about data science that this course is aimed at helping you understand are not so often developed in the quick online tutorials. They can help you with an immediate coding problem, it can take a very long time to develop broad and flexible understanding of data science through just looking stuff up on the internet and applying them until they seem to work. This dark side is quick, easy and seductive but won't bring you a good understanding. Take your time and work through the details and you'll emerge with a much more adaptable and applicable set of skills.

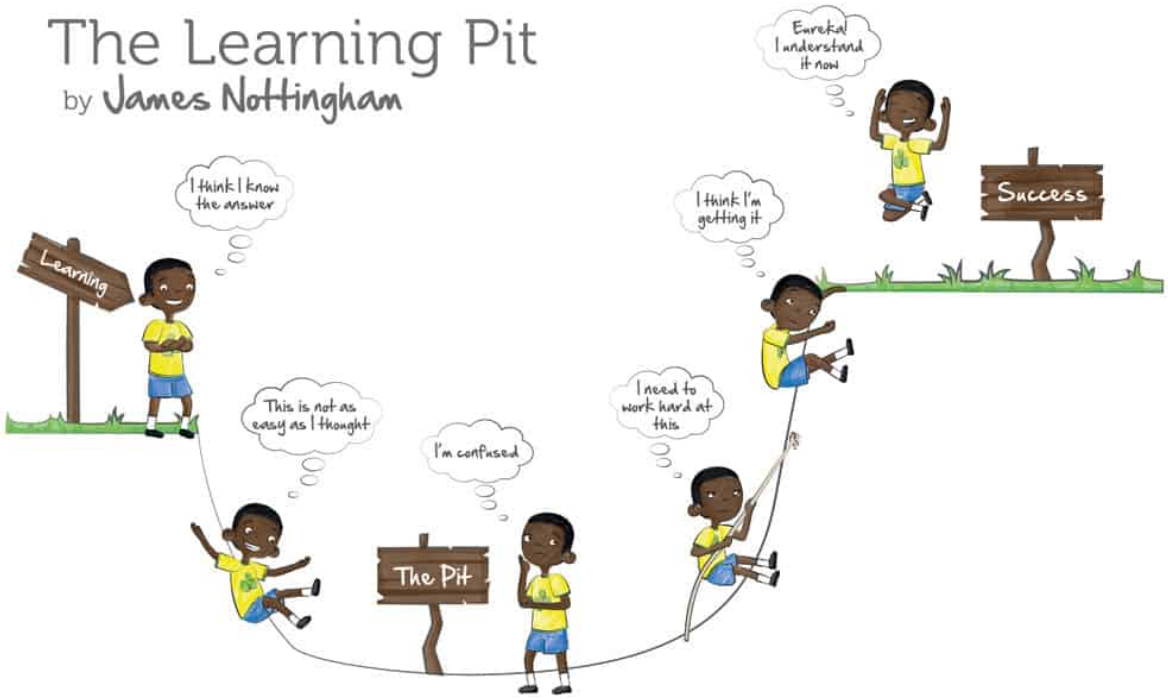

Figure 3: Artwork from James Nottingham - https://www.challenginglearning.com

Using ChatGPT and other 'bots

It is perfectly acceptable to use ChatGPT and other AI's to help you to work with code. AI's like that can't do 'thinking' but they can do 'parsing' so if you have a very specific question especially about syntax do feel free to use it. Remember that these sorts of tools are 'cliche machines' they generate an average answer, which isn't going to be helpful when you want to come up with an analysis strategy that is very tailored to an individual research question. It won't do a good job of telling you which test to use.

The twin learning challenge

One difficult thing about learning to solve data questions with code is that you have to learn two things at the same time; how to solve the problem in code (how to program) and how to actually correctly write the code (learn the syntax of your chosen language). Its perfectly possible to get the first thing right and falter with the second, the second being incredibly easy to get wrong. Syntax in computer languages is incredibly picky and requires an eye for detail. One misplaced letter or punctuation mark and the code fails. In order to help with this it can be useful to learn how to use some tools to help with syntax. Here's a short list.

1. *The Language Help System*

Most languages have a help system, in R this is accessed in the 'help' pane of RStudio or by putting a `?` in front of a search term in the console, so `?print` will return entries in the help that have `print` in their title or description. (Note that `?how do I do a t-test` won't return much. It's a syntax based help, not a 'how to').

2. *ChatGPT*

   This is a great tool for checking syntax of general stuff, these bots have a good internal description of the syntax of well discussed topics, so you can double check your code for syntax errors by pasting it into ChatGPT. It might struggle a bit with less used packages and certainly will give you poor advice on how to actually create your analysis.

3. *Stack Overflow*

   A website for contributions on solving particular well-defined and specific (e.g 'how do I extract a column from a table') coding problems. If you have a problem with code then probably someone else has had the same issue asked about it and hopefully go an answer. Googling a coding problem will likely hit a Stack Overflow entry, so start with Google.

References

How To Succeed In This Course

How to make the most of the course

As this course is designed to be a practical one, you'll get the most out of it from practicing. The best way to do that is to come to the session prepared for the stuff that you need to practice. For the flipped classroom components especially, you should make sure that you have read the session materials linked from this site *before* the in-person session. This will mean that you will have had time to digest any new concepts and can concentrate on the application of them whilst doing the problem sets in class. Coming to the session to read and do the problem sets in one go will *not* be to your best advantage.

How to make the least of the course

A lot of people with biological training will have already gained some experience in statistical tests or linear models or other data science and bioinformatics. That is great and will be a big help. Some people with biological training who have gained experience of data science and statistical methods will have hated it and come to believe that the whole topic is not for them. Assuming that this course is just a restatement of prior material and that you already know it or that you will never know it are definite ways to miss the particular ways of thinking about data that this course will try to give you. Work with an open mind about what statistics and data science can bring to your science to avoid this.

How to work together

Working together is actively encouraged in the sessions, discussing the subjects, the particular problems and mapping out solutions to problems is a great way to come to a better understanding of them. Try to do this vocally with partners or groups, or try to sketch ideas out on paper before committing individual code into the computer.

How not to work together

Working together should not spill over into assessment submission, code and text submitted for those should be your own work, typed in by your own hand. Copy and Paste from the internet or a shared solution will very likely run foul of University plagiarism rules.

How to get a not-so-good mark

The assessment of this course is through an essay and a practical analysis problem. The sorts of problems you'll be asked to solve are discussed very widely across the internet, so when you're under time pressure you might be tempted to try a quick Google to see what those tutorials say and pull those methods into your work verbatim. Doing this not only runs the risk of University plagiarism rules, but also misses the point of much of the teaching. The course teaches some useful and particular ways of thinking about data science that are not so often developed in quick online tutorials. If you rely on the Google Copy and Paste strategy alone and don't show evidence of learning from the TSL course your examiners will likely interpret this as a missed learning outcome. Completely ignoring the materials in the extended section of the course (topics 4, 5 and 6) in your assessment will likely be interpreted as a missed learning outcome too.

Part I

Topics

The topics we cover in this course are selected to be of the most use to you practically in your future science-related careers. They are delivered with a focus on the student becoming an informed and critical user of the tools rather than a domain specialist. The emphasis is on breadth and on the practical use of the skills discussed.

1 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.

1.1 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, tailored to TSL.

1. Shell Genomics
2. Data Wrangling and Processing for Genomics

We will also use these materials

3. Genome, Transcriptome and Structural Databases

1.2 For you to do

Parts 1 (Shell Genomics) and 2 (Data Wrangling) will be full contact time instructor led bootcamp 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.

2 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.

2.1 The materials

For this topic we will use these courses

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

2.2 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 2.1: Artwork by @AllisonHorst

3 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 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 tests and comparisons.

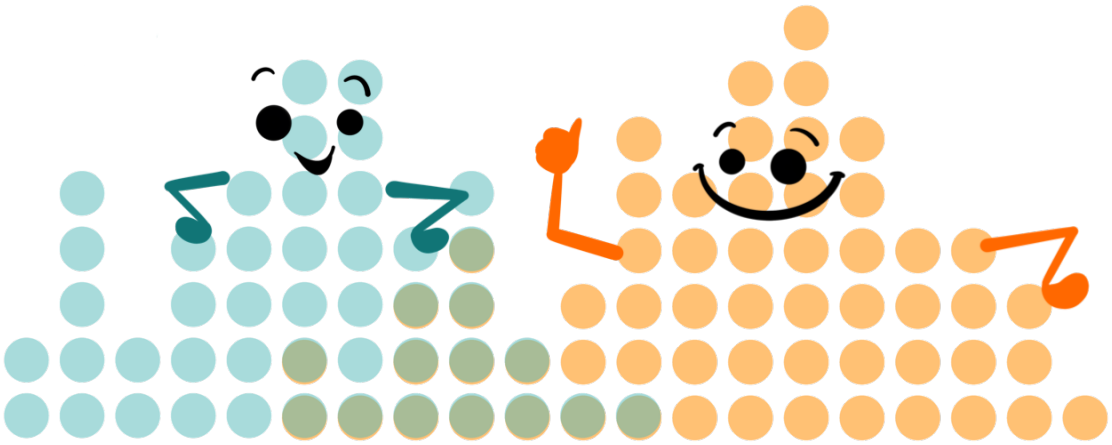

Figure 3.1: Artwork by @AllisonHorst

3.1 The materials

For this topic we will use this course

1. Understanding Statistical Thinking With Linear Models

3.2 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.

4 Introduction to Non-Frequentist Statistics

In this topic we will take a look at alternatives to the standard statistical testing 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 decision making 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 and interpret Bayesian versions of some common statistical tests

Figure 4.1: Artwork by @AllisonHorst

4.1 The materials

For this topic we will use these courses

1. Estimation Statistics
2. Bayesian Inference with Bayes Factors

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.

5 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.

5.1 The materials

For this topic we will use these courses

1. Introduction to Machine Learning

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.

6 Beginning Programming

Programming is an essential but tricky tool to master for all data scientists. 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 often seem less useful because of their generality, so in this topic we will study the general concepts that will allow us to program in most problem domains. We will introduce the central concepts of programming 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.

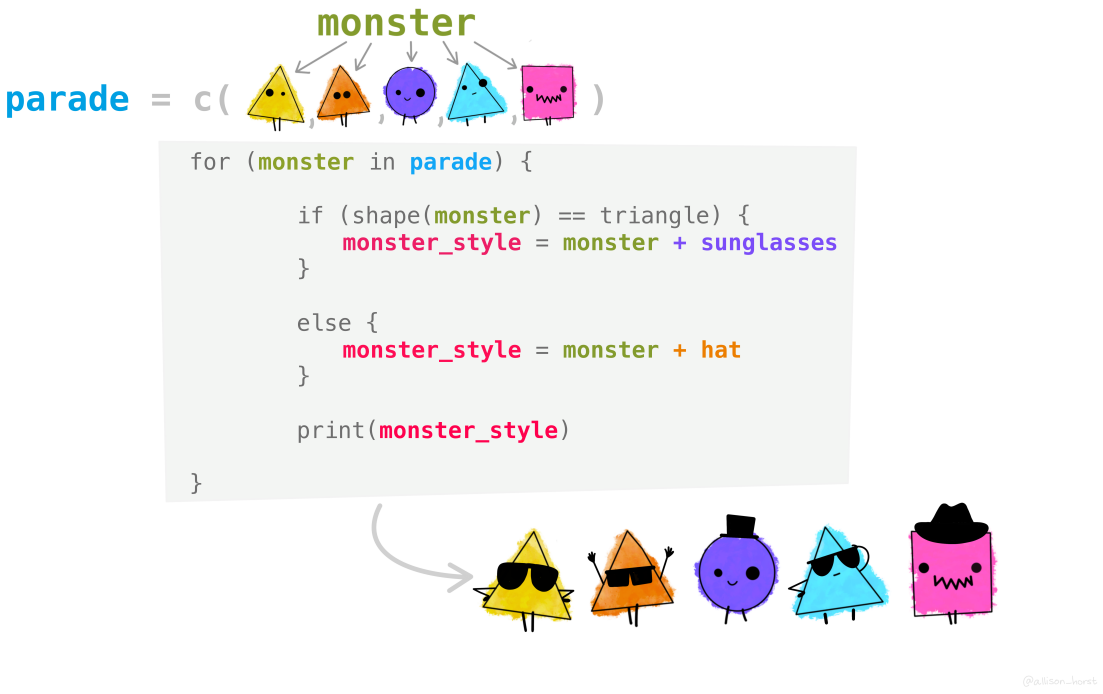

Figure 6.1: Artwork by @AllisonHorst

6.1 The materials

For this topic we will use these courses

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

6.2 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.

7 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.

Figure 7.1: Artwork by @AllisonHorst

7.1 The materials

For this topic we will use this course:

1. Literate Computing

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.

Nottingham, James A. 2015. *Challenging Learning*. Routledge.
Staricoff, Marcelo. 2020. *The Joy of Not Knowing*. Routledge.