TSL Bioinformatics Training

Dan MacLean

2023-06-01

Table of contents

About

Welcome to the TSL Bioinformatics Training Website.

In here you will find links to the written content for each of the separate topics covered as well as the descriptions of the courses and a guide to helping you work out what you need.

> **!** Wait - I just want to use the HPC.
>
> That's great! To get an account, email George Deeks and he'll initiate that for you. There's a short induction with George that is compulsory - it aims to help you work out where everything is and acts as a refresher. If you're already capable at the command line it takes about 30 minutes.

> **!** Uh - I want to use the HPC but don't have command line experience yet.
>
> That's great too, seek out Alison MacFadyen and her Intro to Command Line course, this will give you the skills on working at the command line. Once you've got the basics you'll be good to use the HPC.

Course Delivery

Bioinformatics and Data Science comprise a set of practical research skills grounded in statistics and computer science. Learning a practical skill requires practice! So these courses are not lecture courses, they are a mixture of 'flipped' classroom courses and bootcamp style courses each with a very strong practical component. In this framework the onus is on the learner to lead their work and practice with the provided materials prior to contact and discussion time with the wider learning group and group mentors. 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 you will have a strong practical grounding in using bioinformatics and data science approaches to research problems that will enhance the biology that you are doing.

Online Materials

The rest of this site outlines the online materials provided, broken down by the separate topics we cover. The materials contain a mixture of self-led tutorials and interactive challenges or problems to solve.

Installation of software and tools

Most of the topics outline the stuff you will need to install. Sometimes this is fiddly and sometimes it just doesn't work (and often when it doesn't work it isn't because you did something wrong). If you don't feel confident or are having problems see George Deeks and Chris Rickett for help

Topics

The following topics are covered in our training

1. Introduction to Bioinformatics on the Command-Line and HPC
2. Introduction to Online Databases
3. Data Exploration and Visualisation
4. Understanding Statistics With Linear Models
5. Introduction to Non-Frequentist Statistics
6. Introduction to Machine Learning
7. Beginning Programming
8. Literate Computation

None of them are compulsory, all are open to everyone.

> **Cohorts**
>
> We provide our sessions to each TSL cohort separately - meaning that we generally host separate workshops for Students and Post Docs to reflect the different needs and interests of the different groups. If you're not sure which cohort you should go to, ask us. It may be that we need to organise something specific for you.

Upcoming sessions

Upcoming sessions

Session	Date	Cohort	Type	Start Ti
Data Exploration and Visualisation	2022-07-05	PD	Flipped Class Workshop Session	13
Data Exploration and Visualisation	2022-07-08	Stu	Flipped Class Workshop Session	13
Stats with Linear Models	2022-08-16	PD	Flipped Class Workshop Session	13
Stats with Linear Models	2022-08-18	Stu	Flipped Class Workshop Session	13
Other Topics	2022-09-13	PD	Flipped Class Workshop Session	13
Other Topics	2022-09-15	Stu	Flipped Class Workshop Session	13

How To Use These Courses

Session types

The topics are 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.

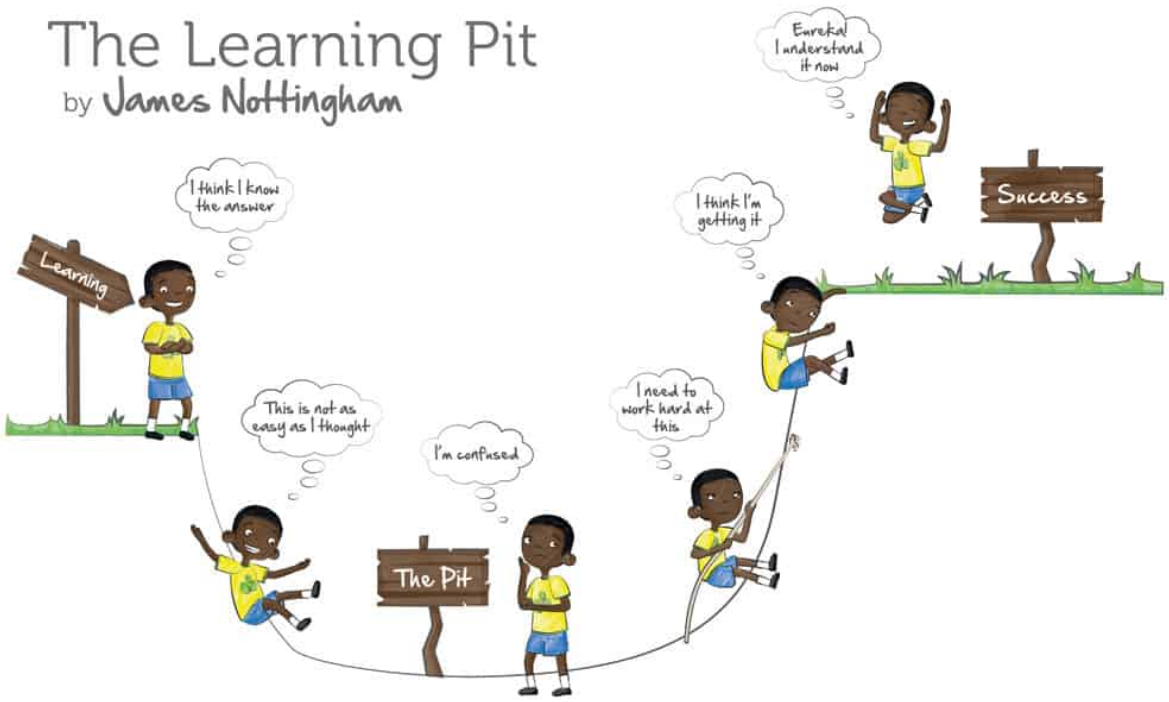

Figure 1: Artwork from James Nottingham - https://www.challenginglearning.com

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.

References

Part I

Topics

The topics we offer are selected to be of the most use to you practically in your research at TSL careers.

The selection ranges from getting to grips with the command line to getting practice in data science topics and

Which should I choose??

If you want to get started with command line to use the HPC or use the range of bioinformatics software then definitely look at the Intro to Command Line course. Otherwise the Data Exploration and Visualisation will be of great help in your day-to-day data analysis (especially that which comes out of command line programs). The fundamental topic is the Understanding Statistics with Linear Models which helps to show the linking thread between the statistical tools used in many data science and bioinformatics applications. The wider courses like Non-Frequentist Statistics and Machine Learning are excellent extensions that really power up the tool boxes and will let you get clearly at tricky data and big data sets.

They're all available all the time and the team are happy to discuss any aspect of them whenever, so feel free to try any and all.

1 Introduction to Command Line

Getting started with using a command line (or 'shell') is the fundamental step in bioinformatics. This interface is the 'real' interface to a computer and lets you work massively more efficiently and on very big scales. It is also the only way to use the large HPC cluster and the many bioinformatics programs on there. So you're going to need this skill.

1.1 The materials

This topic comprises the following courses. You may recognise some of it, as the first part follows the excellent Data Carpentry Shell Genomics Curriculum https://datacarpentry.org/lessons/#genomics-workshop

1. Shell Genomics
2. Data Wrangling and Processing for Genomics
3. Introduction to Installing HPC Software with Singularity

1.2 Delivery

As these topics are the first that people generally need to do we deliver them as instructor-led bootcamps. It is however *always* useful to read the materials provided before the course

1.3 Timetable

Much of this is done on-demand, so if a course session isn't listed in a timeframe that you're comfortable with then contact Clara Jégousse or George Deeks for information. In many instances we'll be able to sort a session soon.

2 Online Databases

There are a great number of excellent online resources in genomics, transcriptomics and structual biology. However, some of them can be a bit opaque to use. This self-led online tutorial will lead you through the fundamentals of using some of the most powerful - but confusing instances.

1. Genome, Transcriptome and Structural Databases

3 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.

Figure 3.1: Artwork by @AllisonHorst

3.1 The materials

For this topic we will use these courses

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

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 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.

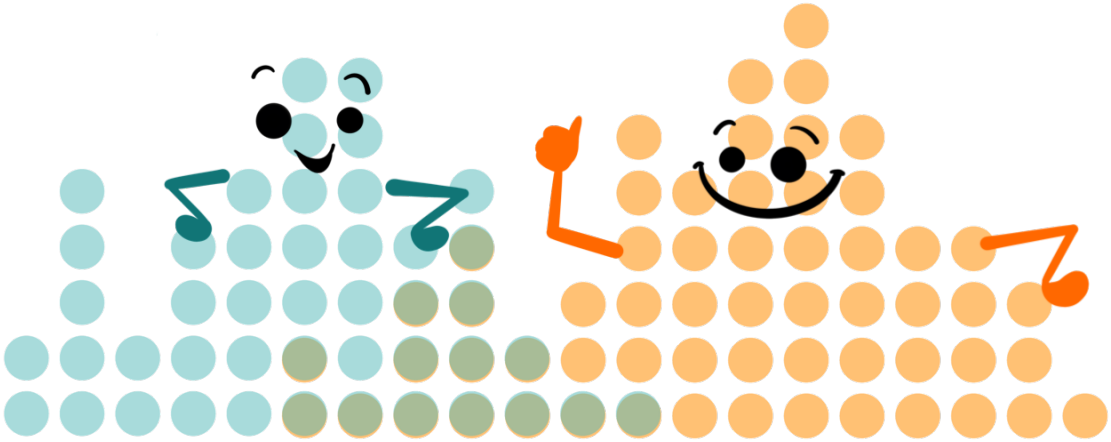

Figure 4.1: Artwork by @AllisonHorst

4.1 The materials

For this topic we will use this course

1. Understanding Statistical Thinking With Linear Models

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 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 5.1: Artwork by @AllisonHorst

5.1 The materials

For this topic we will use these courses

1. Estimation Statistics
2. Bayesian Inference with Bayes Factors

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 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.

6.1 The materials

For this topic we will use these courses

1. Introduction to Machine Learning

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.

7 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.

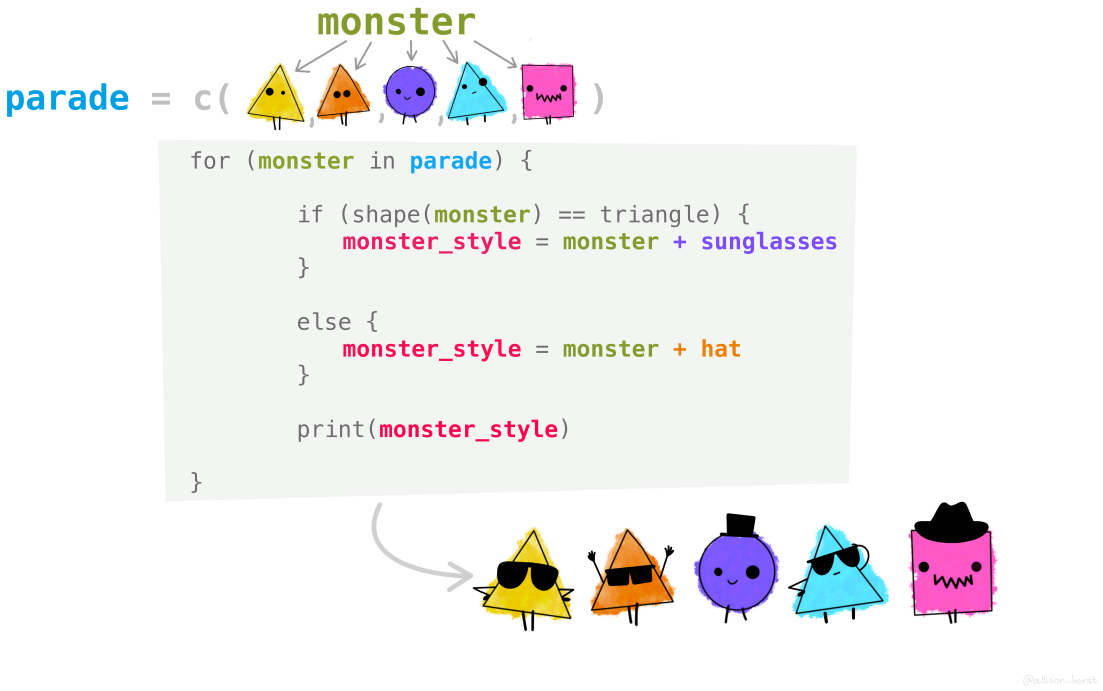

Figure 7.1: Artwork by @AllisonHorst

7.1 The materials

For this topic we will use these courses

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

7.2 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.

8 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 8.1: Artwork by @AllisonHorst

8.1 The materials

For this topic we will use this course:

1. Literate Computing

8.2 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.