

Scala for data science and machine learning

Level: Advanced

Duration: undefined hours

We are very happy to announce that Prof Darren Wilkinson is running a series of four courses for data science and statistics with Scala.

Course 2 will survey the Scala library ecosystem relevant to data science applications. Particular attention will be paid to Breeze, the Scala library for scientific computing and numerical linear algebra, and Smile, a library for data analysis and machine learning. We will look at reading and writing data, via internet connections and disk, using CSV and other formats. Data manipulation, visualisation/plotting, data summarisation, data analysis and model fitting will each be considered. Documentation libraries (mdoc) and testing frameworks (munit) will also be covered.



Course Outline

Course 2 will survey the Scala library ecosystem relevant to data science applications. Particular attention will be paid to Breeze, the Scala library for scientific computing and numerical linear algebra, and Smile, a library for data analysis and machine learning. We will look at reading and writing data, via internet connections and disk, using CSV and other formats. Data manipulation, visualisation/plotting, data summarisation, data analysis and model fitting will each be considered. Documentation libraries (mdoc) and testing frameworks (munit) will also be covered.

This suite of 4 half-day courses is aimed at statisticians and data scientists already familiar with a dynamic programming language (such as R, Python or Octave). Scala is a free modern, powerful, strongly-typed, functional programming language. It is fast and efficient, runs on the Java virtual machine (JVM), and is designed to easily exploit modern multi-core and distributed computing architectures. Scala is a favourite language for data engineering teams and others wanting to work with data at scale in an efficient, safe and timely fashion. For similar reasons, it is also very well suited to the development of robust data science, machine learning and statistical applications.

The courses can be taken independently, but do have pre-requisites which are detailed within the Prior Knowledge summaries. They will be delivered through a combination of lectures, live demos and hands-on practical sessions. The courses will be delivered by Prof Darren Wilkinson, a well-known expert in computational Bayesian statistics and a leading proponent of the use of strongly-typed FP languages (such as Scala) for scalable statistical computing. Participants will be expected to use their own laptops and to have a recent version of Java pre-installed. Other set-up instructions will be provided in advance to registered participants.

Prior Knowledge

Course 2 assumes a basic familiarity with programming in Scala broadly equivalent to that provided by Course 1 ([Introduction to Scala and functional programming](#)). Some familiarity with Scala 3 will be useful, but the course should be accessible to those with a background in Scala 2. It will be assumed that participants are already familiar with Sbt, and with writing simple Scala programs using an editor or IDE such as IntelliJ.

Contact

hello@jumpingrivers.com