Stats 290: Computing for Data Science

John Chambers and Balasubramanian Narasimhan

Requirement

This course assumes you have access to a full-fledged laptop computer. We assume you can install software development tools and obtain access to command line. Devices like Chromebooks and Surface Pros are typically not suitable for this class.

Syllabus

A tentative syllabus follows not necessarily reflecting the order in which the topics will be introduced. As this is a course that caters to a wide audience, there may not be enough time to delve into all the topics. Therefore, an additional file, `schedule.html` is provided, which tries to be as specific as possible.

The statistical computing system R will serve as a focal point for most of the lectures.

We remind you that that this class assumes some familiarity with R and is certainly *not* an introduction to R class. Some programming maturity is expected and indeed required. There is a lot of coding in this class.

- **Overview** Data Science and Data Science Workflows.
- **Functional Programming** Functional programming and R; objects in R
- **R packages** Design, checks, publishing.
- **OOP** Object Oriented Computing in R, S4 Classes and Methods, Reference Classes, R6 and other languages.
- **Data** Data handling and manipulation, data formats, spreadsheets, abstractions to handling data in databases from R.
- **Graphics** Data Visualization: R graphics, ggplot2, networks
- **Tidyverse** The tidyverse ecosystem.
- **Serialization** Serialization of data and code, XML, Xschema, XSL, JSON, Google Protocol Buffers, web technologies.
- **Distributed Computing** Distributed computing, Optimization, Security and Privacy issues.
- **Interfaces** Intersystem interfaces: R and C,C++, Python, Java, etc.
- **Debugging** Debugging R, interactively in R and at the C level.
- **Efficiency** Large computations and large data; vectorizing; measuring efficiency.
- **Reproducibility** Tools and principles.
- **HPC** High Performance Computing, Cluster Computing, R facilities.
- **Web Interfaces** Web based interfaces, Shiny, libraries, publishing. Examples.
- **Special Topics** Packages for Deep Learning.

There will be one or two guest lectures which once again will affect what is covered.

Lecture notes

Lecture notes will be posted just before class on Canvas (https://canvas.stanford.edu). *Please be aware that sometimes we do make some edits after class, so be sure to check time-stamps.*

The lecture notes will be mostly R markdowns. The goal is for you to be able to reproduce *everything* the instructor does. This means there is an expectation that you actively participate and experiment with the code in these markdowns on your own. In particular, we expect you to render the R markdowns yourself; in that process, you will encounter important details that you might otherwise be unaware of.

Texts and References

There is no specific text for this class since the material is very varied. Recommended references are:

- *Software for Data Analysis* by John Chambers, Springer (2008).
- *Extending R* by John Chambers, CRC Press (2016).
- *R Graphics Cookbook* by Winston Chang, OReilly (2012).
- *Advanced R* (Second edition) by Hadley Wickham, CRC Press (2019).
- *R Packages* by Hadley Wickham, O'Reilly (2015).
- *R for Data Science* by Hadley Wickham and Garrett Grolemund, O'Reilly (2016).
- *Deep Learning with R* by Francois Chollet and J. J. Allaire.

  All of these are electronically available from the Stanford Libraries (https://searchworks.stanford.edu).

For the O'Reilly books, searchworks isn't always up to date. But you can access the book by first logging into a Stanford website using your sunet credentials, for example Axess (https://axess.stanford.edu), and then visiting the proquest website (http://proquest.safaribooksonline.com/). There, you can search for the title that you can bookmark it. Sometimes, there are restrictions on the number of concurrent users, in which case, one can just find these books online: nothing we use is not also available online.

In addition to this, there are numerous resources on the web, including R Blogs, Stackoverflow and CRAN manuals. *Google is your friend.*

Assignments and Projects

We remind all students that Stanford requires strict adherence to the Honor Code.

- All on-campus students are expected to attend class and to refrain from enrolling in any other class that meets at the same time.
- Assignments during the course, 4 in number (20 + 20 + 20 + 15) (75% total weight). Assignment 4 will include a reproducibility mini project that will be discussed in class.
- Final project: create an *interesting* R package *with a partner*.(25%) Only applies to those taking course for a letter grade!
- Please be aware that *Incomplete* grades will not be given in this class.

Note that with the proliferation of packages on CRAN and github, it is likely that someone else has already thought of your idea for a package. So the key is to do execute a chosen idea in the best fashion possible, using what is learned in the class.

Details on project requirements, some examples, and further details will be posted during the second week of class.

- There is no final exam for this course.

Just to reiterate: those taking the course for CR/NC are not expected to submit final projects. They should score at least 75% on the homeworks to get CR.

Delayed Submissions

Delayed homework/mini project will be only be accepted for a period of 24 hours past the deadline incurring a 10% penalty. The final project, however, is due as specified with no exception.

Special Considerations

Enrolled students with special needs should contact the instructor (`naras@stanford.edu`) without delay.

Office Hours and TAs

- Instructors: John M. Chambers (email: `jmc4@stanford.edu`) and Balasubramanian Narasimhan (email: `naras@stanford.edu`).
- Instructor (BN) office hours: Wednesdays 4:15pm to 5:30pm, Room 139 Sequoia Hall. For remote students, this will be done via [Zoom] (https://uit.stanford.edu/service/zoom (https://uit.stanford.edu/service/zoom)). It is your responsibility to ensure that set this up. Note that Stanford has a license for this software, so you can use the Stanford licensed client to avoid hassles.

  ZOOM Meeting ID for instructor (BN) is https://stanford.zoom.us/j/7039919024 (https://stanford.zoom.us/j/7039919024%5D).
- Our TAs are Stephen Bates (email: `stephen6@stanford.edu`), Kevin Fry (email: `kfry@stanford.edu`) and Isaac Gibbs (email: `igibbs@stanford.edu`). Their office hours will be posted soon. They will also be monitoring Piazza.

Help and Discussion

- We will use Canvas (https://canvas.stanford.edu) for class lecture slides and materials. Also for some announcements.
- We will use Piazza (https://piazza.com/stanford/winter2020/stats290/home) for discussions and questions. This will be activated after the first class.

SCPD Students

SCPD students should note that all official communications will be sent to your associated Stanford email address, not gmail or any other.