International
Max Planck Research School
for language sciences

Automate the Boring Stuff: Programmer's Tools for Fun and Profit | 2019

Dr. Phillip Alday • 1.0 EC

COURSE OVERVIEW

COURSE DESCRIPTION & INSTRUCTOR PROFILE

GOALS

This course covers basics of automation and coordination two core tools from the modern programmer's toolbox. The core aspects of the course center around version control and manipulation of plain text. Building on this foundation, we will cover the combination of Git, pandoc and plain-text markup formats for improving collaborative, reproducible, open research.

LEARNING OUTCOME

Students will become competent in the basics of using Git, both to document individual work and to collaborate with others. Students will learn how plain text formats such as *Markdown* provide a convenient way to produce text styled and saved in an array of formats, without depending on proprietary, closed software.
NB: "real" programming in R/Python/other common analyses languages will not be covered in this course, but examples demonstrating how to include analysis code, results and text into a single output will include snippets of these languages.

PRE-REQUISITES

- A working laptop computer with sufficient permissions to install software

ABOUT THE INSTRUCTOR

Phillip was a disgruntled sysadmin and recovering mathematician in a former life, who grew tired of the shadows in his Platonic cave, and so emerged to study the electrophysiology of language. He believes that the ancient computer deities demand a return to the old ways and that plain text with a clear license is the one path to true enlightenment.

CONTACT INFORMATION

phillip.alday@mpi.nl | MPI 380

FORMAT & CONTENT OUTLINE

Eight (8) sessions, 120 minutes each, alternating between more lecture-like and tutorial sessions.

10:00 AM - 12:00 PM MPI 163
Feb 28 Mar 7 · 14 · 21 · 28 Apr 4 · 11 ·18

16 contact hours + 8 self-study hours = 28 hours (1.0 EC)

Please note that this is a *very* hands-on course. The core concepts taught here are seemingly easy to understand but full of subtlety and complexity in use. The best way to learn this material is by doing while you still have an instructor around to help explain the ugly bits.

The overall plan is as follows, roughly divided by session:

- **Session 1** In the beginning was the command line: Why plain text and cockroaches will outlive

us. Line endings and DVCS beginnings. Text encoding. Unix Pipelines.
Homework:

- Read the Git parable and Hg Init
- Install pandoc
- Install R and Rstudio
- (optional, but recommended) Install LaTeX (TexLive is recommended for Mac&Linux, MikTex for Windows)

- **Session 2** Git setup and configuration. Diff'ing, the index (staging area) and committi ng. Pulling and pushing. Git vs. GitHub vs. GitLab vs. Bitbucket. Asymmetric cryptography and SSH keys. Two-Factor Authentication. Git ignore.

- **Session 3** Working together while working apart. Branching, pull requests and merges. Surviving detached HEADs and knowing when and how to reset.
  Homework:
  - Pull request with a merge conflict

- **Session 4** Reproducible research: combining code and writing
  - Makefiles; tabs, spaces and line endings revisited
  - Rmarkdown > Jupyter Notebooks
  - Taking advantage of knitr options.
  - Every R-Script is a notebook waiting to happen with knitr::spin()

- **Session 5** Saving the day with regular expressions. Misunderstanding math, computer science and linguistics simultaneously with the Chomsky Hierarchy and its relationship to computing devices. The immense world of POSIX tools: find, grep, sed, awk; perl and python.
  Homework:
  - Work through as many lessons as possible in RegexOne
  - Play some regex golf

- **Session 6** ***(guest instructor!)*** Regex and command-line sandbox

- **Session 7** Preparing for submission and sharing with the unenlightened. Advanced output options and templates in Pandoc. Using Makefiles to package everything up; multiple output formats.
  Homework:
  - Prepare Lorem Ipsum Markdown with accompanying R script for submission to a journal accepting docx.
  - Prepare Lorem Ipsum RMarkdown for submission to a journal accepting LaTeX.

- **Session 8** Go big or go home: Git-LFS; Git-Annex. How Git *really* works. Issues when mixing Git and synchronization systems (Dropbox, owncloud, etc.). Additional graphical tools. Remember the lessons of history. Zenodo and keeping things around for posterity. Licensed or locked away.

At several points, you will be asked to install software. Issues related to installation must be addressed outside of and ideally before class.

ATTENDANCE POLICY & REGISTRATION

IMPRS policy: To earn full credits, participants are expected to fully attend all 8 **required sessions** and complete all preparatory work. Absences should be communicated in advance to the IMPRS Office and instructor.

Registration link: https://tinyurl.com/auto2019-reg

To ensure that the IMPRS continues to offer useful content, please take a few moments to share your feedback at the end of the course: https://tinyurl.com/auto2019-eval