Summer Training Workshops
Data Science Institute
Wond’ry, Suite 300

Wednesday, June 5, 10a-12p

**Git and GitHub**
Software Carpentry Curriculum

Version control is the lab notebook of the digital world: it’s what professionals use to keep track of what they’ve done and to collaborate with other people. Every large software development project relies on it, and most programmers use it for their small jobs as well. And it isn’t just for software: books, papers, small data sets, and anything that changes over time or needs to be shared can and should be stored in a version control system.

Thursday, June 6, 10a-12p

**Data Organization in Spreadsheets**
Data Carpentry Curriculum

Good data organization is the foundation of any research project. Most researchers have data in spreadsheets, so it’s the place that many research projects start.

Typically we organize data in spreadsheets in ways that we as humans want to work with the data. However computers require data to be organized in particular ways. In order to use tools that make computation more efficient, such as programming languages like R or Python, we need to structure our data the way that computers need the data. Since this is where most research projects start, this is where we want to start too!

In this lesson, you will learn:
- Good data entry practices - formatting data tables in spreadsheets
- How to avoid common formatting mistakes
- Approaches for handling dates in spreadsheets
- Basic quality control and data manipulation in spreadsheets
- Exporting data from spreadsheets

In this lesson, however, you will not learn about data analysis with spreadsheets. This we'll be covering using tools such as R and Python that support reproducible work flows.
Tuesday, June 11, 9a-12p
Plotting and Programming in Python: Part 1
Software Carpentry Curriculum
The best way to learn how to program is to do something useful, so this introduction to Python is built
around a common scientific task: data analysis.

Wednesday, June 12, 9a-12p
Plotting and Programming in Python: Part 2
Software Carpentry Curriculum

Tuesday, June 18, 10a-12p
Data Analysis and Visualization with R: Part 1
R provides some of the best tools for data wrangling and data visualization. Combined with RStudio,
you have a powerful suite of tools that can handle small to large problems, and you’ll probably have fun
at the same time. We’ll cover the basics of the tidyverse, and introduce analytics and data science best
practices for reproducibility using notebooks.

Wednesday, June 19, 10a-12p
Data Analysis and Visualization with R: Part 2

Wednesday, June 26, 10a-12p
Introduction to Exploratory Data Analysis and Dashboarding using Tableau
Tableau is a tool widely used in industry to explore data and to build dashboards. It is fast, visual,
intuitive, and powerful. Although it does not lend itself to reproducible work flows, the benefits may
outweigh this shortcoming if you have a one-time dataset you need to explore quickly, or if you need to
produce a working dashboard.

Wednesday, July 10, 10a-12p
Machine Learning with H2O
Building predictive models is compute-intensive. H2O provides high-performance algorithms with a
simple-to-use front end that allows you to try out multiple approaches quickly and efficiently. In this
workshop, we’ll have start with a brief introduction to machine learning, and then try some hands-on
examples using H2O.

Wednesday, July 17, 10a-12p
Unix Shell
Software Carpentry Curriculum
The Unix shell has been around longer than most of its users have been alive. It has survived so long
because it’s a power tool that allows people to do complex things with just a few keystrokes. More
importantly, it helps them combine existing programs in new ways and automate repetitive tasks so they
aren’t typing the same things over and over again. Use of the shell is fundamental to using a wide range of
other powerful tools and computing resources (including “high-performance computing” supercomputers).
These lessons will start you on a path towards using these resources effectively.
Wednesday, July 24, 10a-12p

**Databases and SQL**
Data Carpentry Curriculum (Modified)
Three common options for storage are text files, spreadsheets, and databases. Text files are easiest to create, and work well with version control, but then we would have to build search and analysis tools ourselves. Spreadsheets are good for doing simple analyses, but they don’t handle large or complex data sets well. Databases, however, include powerful tools for search and analysis, and can handle large, complex data sets. These lessons will show how to use a database to explore data.

Wednesday, July 31, 10a-12p

**Participant’s Choice**

Please contact: datascience@vanderbilt.edu with any questions or concerns.