DATA MATTERS™

DATA SCIENCE SHORT COURSE SERIES

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

NC STATE

August 8-12, Hunt Library, Centennial Campus

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

On this page:

Course Descriptions

Introduction to Data Science

Instructor: Tom Carsey

Description

This course provides an introduction to data science, focusing on data about people. It will cover basic building blocks, key concepts, strengths and limitations, and the ethical issues that emerge in data science. Numerous examples will be discussed and sample code and data will be explored. This course will help equip participants from various disciplines and industries with a general understanding of data

Why Take This Course?

Data Science combines tools from information science, computer science, and statistics to collect, manage, analyze, and understand digital data. Modern data science pays particular interest to data regarding the social and economic attitudes and behaviors of people.

What Will Participants Learn?

This course will help equip participants from various disciplines and industries with a general understanding of data science terms, approaches, and strategies for effectively using data science.

Prerequisites and Requirements

None

Introduction to Information Visualization

Instructor: Ted Polley

Description

Participants will learn how to clean and structure data; see how freely available software can be used to create charts, maps, and graphs; and follow basic design suggestions to fine-tune the final presentation of visualizations for publication or reporting.

Why Take This Course?

Visualization is a growing area of interest for researchers in all disciplines. Visualizations can illuminate important trends in a data analysis project or help an audience engage emotionally with a research area. Many tools are available to produce visualizations, however, and it is not always clear which tool is best or how to structure data to work with the tool. This course will walk participants through a wide variety of data sources and chart types to help even beginners to visualization feel comfortable embarking on a new visualization project.

What Will Participants Learn?

The course will be organized into four major sections:

- Basic charts
- Static and web-based maps
- Network diagrams and hierarchical visualizations
- Graphic design for information visualization

The instructor will demonstrate several tools. These will likely include Excel, OpenRefine, QGIS, RAW, Sci2, Gephi, and R (though the course may adjust slightly to take advantage of any sudden changes in available technology). This course will have a hands-on component, but participants are not required to bring laptops if they prefer to just observe. The instructor will provide sample datasets for the hands-on sections. Participants are also welcome to bring their own data if they have specific questions.

This course will assume a basic understanding of spreadsheets as a way of storing and processing data. We may cover some tools that work with HTML (especially SVG) in advanced examples. Advanced examples will introduce students to the R programming language, emphasizing the importance of reproducibility. No prior experience in any programming language is necessary. Bringing a laptop is not required, but participants are encouraged to do so.

Introduction to Data Science Using R

Instructors: Chris Bail and Justin Post

Description

This course covers importing and exporting data, manipulating data or recoding variables, visualization and statistical analysis, and basic programming skills. The second section of the course covers data cleaning and coding, which can be somewhat complicated in R because it uses a variety of data formats that are not used within other languages. The third section covers basic descriptive analysis, including cross-tabs, histograms, and scatterplots, and basic linear regression models. The final section presents a brief introduction to programming in R, including “for” and “if” loops and vectorized commands.

Why Take This Course?

R has recently become the preferred computing and statistical analysis software for academic analysis because it offers unparalleled breadth of tools for virtually any model of interest to social scientists—and particularly those interested in so-called “big data.” Unfortunately R also has a steep learning curve because it is maintained by academics that have few career incentives to make it user friendly. Courses such as this one are therefore indispensable for obtaining a basic working knowledge of the language and learning how to navigate the complex web of information about R that is currently available online.

What Will Participants Learn?

This course is divided into four sections. The first section provides an overview of how to install R on your computer, import files, and interface with other software such as STATA, SPSS, and R. The second section of the course covers data cleaning and coding, which can be somewhat complicated in R because it uses a variety of data formats that are not used within other languages. The third section covers basic descriptive analysis, including cross-tabs, histograms, and scatterplots, and basic linear regression models. The final section presents a brief introduction to programming in R, including “for” and “if” loops and vectorized commands.

Prerequisites and Requirements

This course assumes no knowledge of computer programming, but basic familiarity with another statistical analysis software such as STATA, SPSS, or SAS will make the course easier to follow.

Note: In order to participate in the hands-on sections of the course, participants must bring their own laptop computer with enough space to install R and RStudio.

Working with Messy Data

Instructor: Brown Biggers

Description

When working with data, one thing is fairly certain: data is rarely in an optimized format. A misplaced space here, or an extra comma there, can mean the difference between two clicks, and two hours of work. In this course, we will work with ways to manipulate, interpret, and present data from webpages and text using the Anacondas distribution for Python version 2.7. Attendees will be expected to have done some introductory work with Python, but we will cover some basics for consistency. This class will also touch on regular expressions, and various imported libraries to extend Python functionality.

Prerequisites and Requirements

A basic familiarity of Python will be helpful. A good introduction can be found at http://learnpythonthehardway.org/book/

Open(ing) Data

Instructors: Sophia Lafferty-Hess, Thu-Mai Christian

Description

The benefits of making data open and accessible have been widely discussed within the academic and public policy communities. Sharing research data enables others to verify and build upon published results, supports transparency and accountability of research findings, increases the return on public investments in research, encourages new scientific innovations, and supports collaboration within and across disciplines. However, there are also some challenges related to opening up data to the broader community. This workshop will examine the opportunities and challenges of open access to data resources and some of the open-source mechanisms available to share research data. Specifically, participants will learn about:

- The open data access movement
- Data security considerations
- Protection of the confidentiality of research participants
- The process of anonymizing datasets
- Embargos and rights of first use
- Access restrictions
- Data ownership
- Data citation
- Other ethical questions related to data sharing and reuse

Prerequisites and Requirements

None

Programming in R

Instructors: Chris Bail and Justin Post

Description

This class provides students with an introduction to basic programming techniques in R, a program with stronger object-oriented programming facilities than most statistical computing languages. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R's popularity has increased substantially in recent years.

Why Take This Course?

This class will be useful to those who wish to restructure or clean unstructured data, collect new data in an automated fashion, or improve the speed of data analysis.

What Will Participants Learn?

Students will learn basic programming techniques such as functions, "for" loops, if/else statements, vectorized functions, and parallel computing techniques.

Prerequisites and Requirements

Basic familiarity with R syntax, objects (e.g. matrices, lists, data frames etc…).

Introduction to Machine Learning

Instructor: Ashok Krishnamurthy

Description

This course will introduce participants to a selection of the techniques used in Data Mining and Machine Learning in a hands-on, application-oriented way. Topics covered will include Data Exploration, Decision Trees, Clustering, Association Rules, Regression and Pattern Classification. The computing exercises will be based on the statistical programming language, R. At the end of the two days, you will be able to explore a data set, and determine which analysis method is appropriate for the data, and be able to use R packages to obtain results.

Why Take This Course?

The ready availability of digital data from numerous sources is a tremendous opportunity for businesses and scientists to obtain new insights and confirm hypotheses. Data mining provides the theoretical basis, algorithms and computational methods to manage, analyze and get information from the data. In the world of big data and data science, data mining is a fundamental tool for data insights.

What Will Participants Learn?

The course will be organized in the following major sections:

- Data exploration

- Decision trees
- Clustering
- Regression
- Classification

Each section will have an associated computer exercise. We will make extensive use of R and R packages in the computer exercises.

Prerequisites and Requirements

This course will assume a basic understanding of statistics (at the undergraduate level) and experience with R (at the level of the “Introduction to Data Science Using R” course).

Collecting, Classifying and Analyzing Textual Data

Instructor: Chris Bail

Description

This course explains how to collect, classify, and analyze text-based data from the internet or other digital sources using R. The course will cover screen-scraping, interfacing with Application Programming Interfaces (APIs), basic natural language processing such as topic models, and explain how these data can be incorporated into traditional social science models.

Why Take This Course?

Big data has become one of the most significant buzzwords in academic circles over the past few years, yet the study of how to use text as data crosses so many different academic disciplines, programming languages, and styles of communication that those who wish to enter this nascent field are quickly overwhelmed. This course will provide students with a panoramic perspective of the field and the programming skills necessary to navigate the rapidly growing wealth of information online about this subject.

What Will Participants Learn?

This course is divided into four segments. The first section will cover basic techniques for collecting text-based data from the internet such as screen scraping and writing code to extract data from application programming interfaces. The second section will explain how to clean and code text-based data using a variety of pre-processing techniques such as stemming. The third section will explain how to apply topic models and other natural language processing tools to sample data. The fourth and final section will discuss best practices for incorporating variables produced via these methods into conventional social science models such as regression or social network analysis.

Prerequisites and Requirements

This course assumes a basic working knowledge of the R language. Students with no knowledge of R might consider pairing this course with the “Introduction to Data Science in R” course that is also being offered early in the week.

Note: In order to participate in the hands-on sections of the course, participants must bring a laptop computer with enough space to install R and R Studio.

Data Curation: Managing Data throughout the Research Lifecycle

Instructors: Sophia Lafferty-Hess, Thu-Mai Christian

Description

This course will provide an introduction to data management best practices as well as demonstrations of digital curation tools including the Dataverse Network™ open source virtual archive platform.

Why Take This Course?

Today, a growing number of funding agencies and journals require researchers to share, archive, and plan for the management of their data. In 2013, an Office of Science and Technology Policy Memo highlighted the importance of providing open access to datasets and scholarly publications as a method of promoting innovation, accountability, transparency, and efficiency. As researchers and information professionals respond to these new requirements, data curation knowledge is necessary for the effective management, long-term preservation, and reuse of data.

What Will Participants Learn?

Participants will learn about:

- Diversity of data and their management needs across the research data lifecycle
- Impetus and importance of preserving and sharing data
- Processes required for preserving and sharing data
- Digital repository activities and assessment
- Role of advocacy and communication when discussing data management best practices

Prerequisites and Requirements

None