Week 1: R Basics

Data Science for Studying Language and the Mind

Katie Schuler

2024-08-27

Welcome

Follow along on the syllabus!

Paperwork

- When you arrive, complete this anonymous form: Who's in class
- You can also join the waitlist if you are not enrolled

Announcements

- The course is full and the room is full
- Ways to join:
    1. Watch for an opening (highest odds of getting in)
    2. Add your name to our waitlist

Course description

Data Sci for Lang & Mind is an entry-level course designed to teach basic principles of statistics and data science to students with little or no background in statistics or computer science. Students will learn to identify patterns in data using visualizations and descriptive statistics; make predictions from data using machine learning and optimization; and quantify the certainty of their predictions using statistical models. This course aims to help students build a foundation of critical thinking and computational skills that will allow them to work with data in all fields related to the study of the mind (e.g. linguistics, psychology, philosophy, cognitive science, neuroscience).

Prerequisites

There are *no prerequisites beyond high school algebra.* No prior programming or statistics experience is necessary, though you will still enjoy this course if you already have a little. Students who have taken several computer science or statistics classes should look for a more advanced course.

Teaching team

Instructor: Dr. Katie Schuler (she/her)

TAs:

- Brittany Zykoski
- Wesley Lincoln

About me, your instructor (Katie)

- You can call me Professor Schuler or Katie, whichever makes you more comfortable
- I live in Mt Airy with my husband and two kids (Dory, 2 and Joan, 6)
- At Penn I also have a research lab, the Child Language Lab and am on the Natural Science and Math Panel (a group focused on improving inclusive teaching in STEM at Penn).
- I'm a first-generation college student from Western NY. I worked 40 hours a week to put myself through college; I am still paying off my student loans.

My assumptions about you

You are an honest, kind, and hardworking student who wants to do well in and enjoy this class

. . .

You are very busy, and will sometimes have to prioritize other things above this class.

Course overview

Data science

Data science is about making decisions based on incomplete information.

. . .

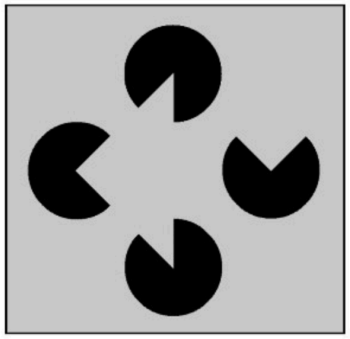

Figure 1: from Kok & de Lange (2014)

This concept is not new. Brains were built for doing this!

But we have new tools and lots more data!

. . .

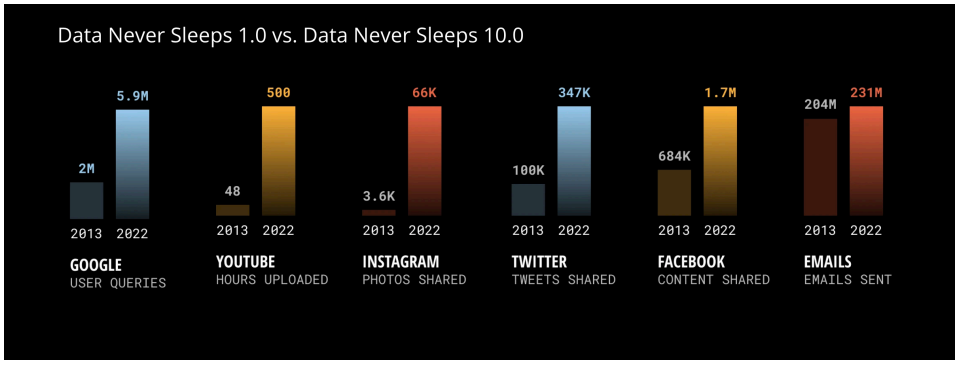

Figure 2: from https://web-assets.domo.com/miyagi/images/product/product-feature-22-data-never-sleeps-10.png

Data science workflow

The folks who wrote R for Data Science proposed the following data science workflow:

. . .

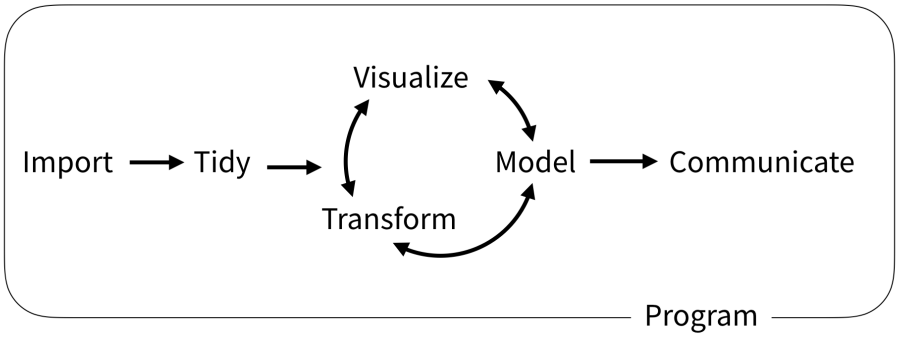

Figure 3: from R for Data Science

Overview of course

We will spend the first few weeks getting comfortable **programming in R**, including some useful skills for data science:

- R basics
- Data visualization
- Data wrangling (import, tidy, and transform)

Overview of course

Then, we will spend the next several weeks building a **foundation in basic statistics and model building**:

- Sampling distribution
- Hypothesis testing
- Model specification
- Model fitting
- Model accuracy
- Model reliability

Overview of course

Finally we will cover a selection of **more advanced topics** that are often applied in language and mind fields, with a focus on basic understanding:

- Classification
- Inference for regression
- Mixed-effect models

Syllabus, briefly

Each week will include two lectures and a lab:

- **Lectures** are on Tuesdays and Thursdays at 12pm and will be a mix of conceptual overviews and R tutorials. It is a good idea to bring your laptop so you can follow along and try stuff in R!
- **Labs** are on Thursday or Friday and will consist of (ungraded) practice problems and concept review with TAs. You may attend any lab section that works for your schedule.

Syllabus, briefly

There are 8 graded assessments:

- **6 Problem sets (40%)** in which you will be asked to apply your newly aquired R programming skills.
- **2 Midterm exams (60%)** in which you will be tested on your understanding of lecture concepts.

Syllabus, briefly

There are a few policies to take note of:

- Missed exams cannot be made up except in cases of genuine conflict or emergency (documentation and course action notice required). You may take the optional final exam to replace a missed or low scoring exam.
- You may request an extension on any problem set of up to 3 days. But extensions beyond 3 days will not be granted (because delying solutions will negative impact other students).
- You may submit any missed quiz or problem set by the end of the semester for half-credit (50%), even after solutions are posted.
- We will drop your lowest pset grade, but you must turn in all 6 assignements to be eligible.

Resources

In addition to our course website, we will use the following:

- google colab (r kernel) - for computing
- canvas- for posting grades
- gradescope - for submitting problem sets
- ed discussion - for announcements and questions

Wellness resources

Please consider using these Penn resources this semester:

- Weingarten Center for academic support and tutoring.
- Wellness at Penn for health and wellbeing.

Why R?

With many programming languages available for data science (e.g. R, Python, Julia, MATLAB), why use R?

- Built for stats, specifically
- Makes nice visualizations
- Lots of people are doing it, especially in academia
- Easier for beginners to understand
- Free and open source (though so are Python and Julia, MATLAB costs $)

Many ways to use R

- R Studio
- Jupyter
- VS Code
- and even simply the command line/terminal

Google Colab

- **Google Colab** is a cloud-based Jupyter notebook that allows you to write, execute, and share code like a google doc.
- We use Google Colab because it's simple and accessible to everyone. You can start programming right away, no setup required!

Secretly, R!

Google Colab officially supports Python, but secretly supports R (and Julia, too!)

- Update 2024: Google Colab now officially supports R!
- colab (r kernel)

Let's try it!

Google colab demo

Open a new R notebook:

- colab (r kernel) - use this link to start a new R notebook
- `File > New notebook` and then `Runtime > Change runtime type` to R

Cell types:

- `+ Code` - write and execute code
- `+ Text` - write text blocks in markdown

Left sidebar:

- `Table of contents` - outline from text headings
- `Find and replace` - find and/or replace
- `Files` - upload files to cloud session

Frequently used menu options:

- `File > Locate in Drive` - where in your Google Drive?
- `File > Save` - saves
- `File > Revision history` - history of changes you made
- `File > Download > Download .ipynb` - used to submit assignments!
- `File > Print` - prints
- `Runtime > Run all` - run all cells
- `Runtime > Run before` - run all cells before current active cell
- `Runtime > Restart and run all` - restart runtime, then run all

Frequently used keyboard shortcuts:

- `Cmd/Ctrl+S` - save
- `Cmd/Ctrl+Enter` - run focused cell
- `Cmd/Ctrl+Shift+A` - select all cells
- `Cmd/Ctrl+/` - comment/uncomment selection
- `Cmd/Ctrl+]` - increase indent
- `Cmd/Ctrl+[` - decrease indent

R Basics

We begin by defining some basic concepts:

Expressions

- **Expressions** are combinations of values, variables, operators, and functions that can be evaluated to produce a result. Expressions can be as simple as a single value or more complex involving calculations, comparisons, and function calls. They are the fundamental building blocks of programming.
  - `10` - a simple value expression that evaluates to `10`.
  - `x <- 10` - an expression that assigns the value of `10` to `x`.
  - `x + 10` - an expression that adds the value of `x` to `10`.
  - `a <- x + 10` - an expression that adds the value of `x` to `10` and assigns the result to the variable `a`

See you next time!

Looking forward to a great semester.