STAT 412/612 Class 1: Intro to R & Data Science

Kelsey Gonzalez

1/21/2021

Contents

1 Intro to R & Data Science

Learning Outcomes

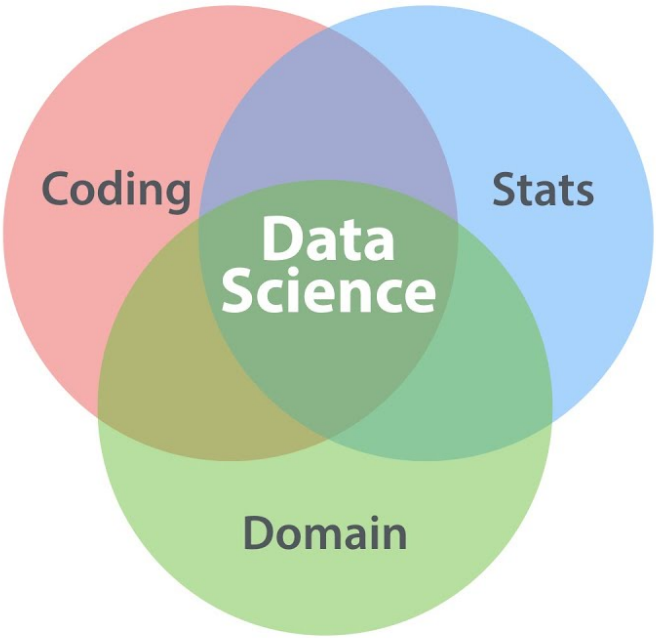
- Understand Data Science as the context for this course
- Understand R in the context of data science.

1.1 What is Data Science?

- Coined in 2001, "*Data science* is a discipline that incorporates varying degrees of Data Engineering, Scientific Method, Math, Statistics, Advanced Computing, Visualization, Hacker mindset, and Domain Expertise." Cleveland (2001)
- "Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes." Banton (2020)

- "Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data."
- **"Data Scientist: The Sexiest Job of the 21st Century"** Patil and Davenport (2012)

1.2 Components of Data Science



- Statistics
- Domain Knowledge
- Computation

1.3 Statistics

- Inferring general properties given data.
- Causal inference.
- Modeling (descriptive and predictive).
- Quantifying uncertainty.
- STAT 514, Statistical Methods, STAT 615 (Regression), STAT 627 (Machine Learning),
- Most of the STAT curriculum is applicable.

1.4 Domain Knowledge

- Expertise in an area of application
  - e.g., biology, psychology, economics, chemistry, etc..
- Allows you to understand data in context of the area and/or decision to be made.
- Lets you ask interesting questions.
- Lets you spot problems with existing analysis pipelines.
- Various "Tracks" in the data science program.

1.5 Computation – This class

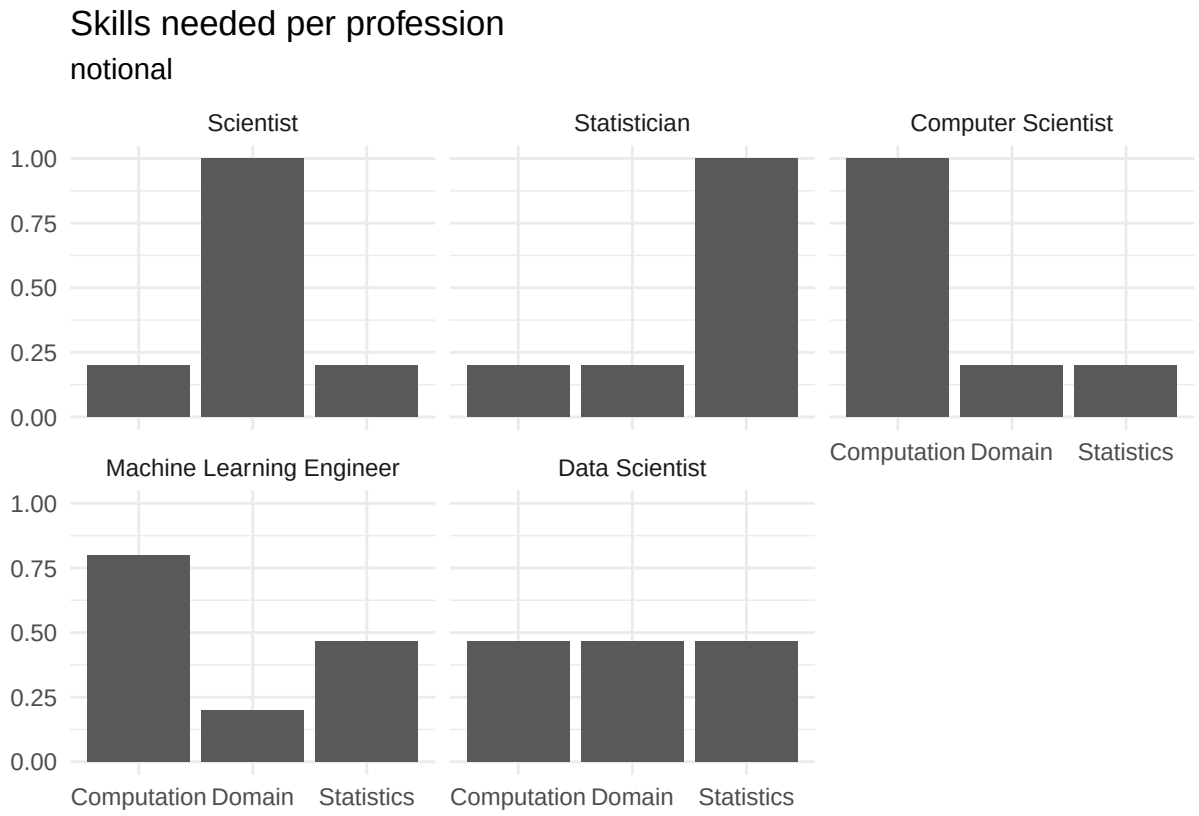
- Data import

- Data preparation
- Data exploration
- Data transformation
- Data visualization
- STAT 612 (R programming), STAT 613 (Data Science), most of the CS curriculum.

2 Various Professions

2.1 What makes a data scientist?

- People in diverse professions use these three skills to analyze data.
- Professions often differ by their level of expertise or interest in each skill.
- Data Science projects usually a "team activity"



2.2 Introductions

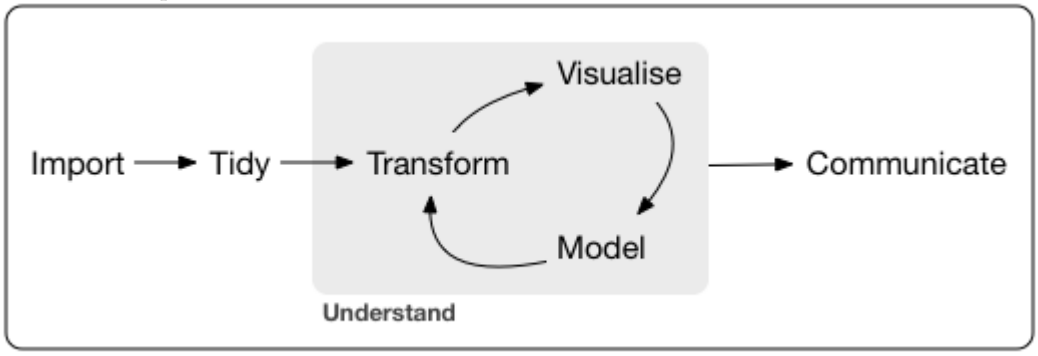
- We'll be doing group projects and you should form your groups of 2-4 people early in the semester.
- Let's take a break to introduce ourselves. Turn on your cameras and say Hi and one or two sentences about yourself and your goals for the course.

3 The Steps of an Analysis and R

3.1 Steps of a data analysis

- Before you start your data analysis and R

  - Something is happening in the world
  - Someone collects data
  - Someone asks a question



- R Time

3.2 Tools

- Many tools exist for these steps:
- General data tools: R, Python, Julia, Matlab, STATA, SAS
- Other tools: SQL (data import), git (version control), map/reduce software (for big data).
- Advantages and disadvantages to each.

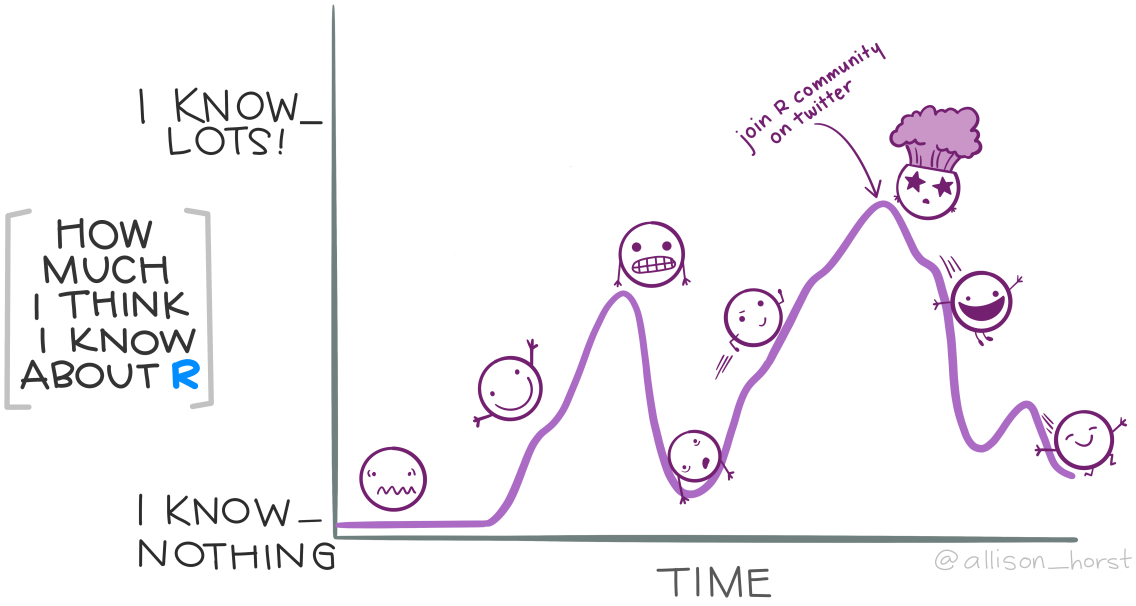
3.3 R

- R is a statistical programming (or scripting) language.
- You write code (a series of functions) to perform some task.
- R can be used to perform **all** of the tasks of a data analysis.
- R is built around the idea of *packages*: like apps
  - Packages are sets of functions designed to work together to accomplish a specific set of tasks
  - There are thousands of packages and you can install any one with a simple function `install_packages()`

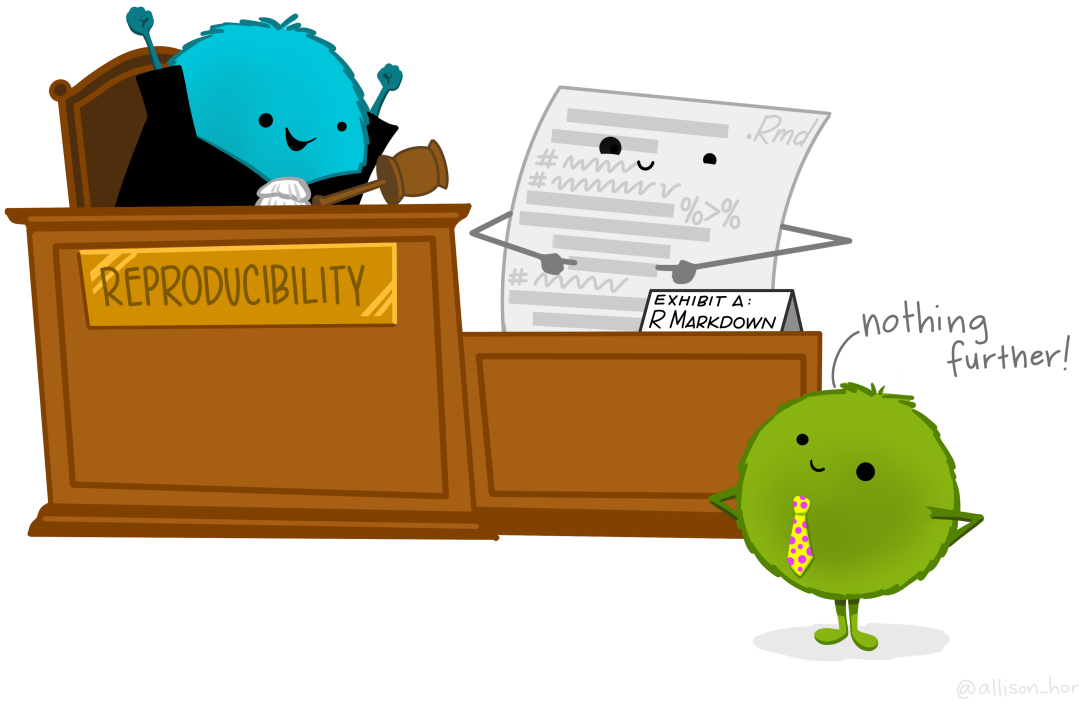
3.4 Motivation for R

1. It's free and open source.
   - You will always have access to R.
   - Not true for other software (Matlab, STATA, SAS).
2. It's widely used with a lot of community support.
   - If you need some special analysis, someone has often made an R package.

3. It's *relatively* easy (especially graphics and data wrangling).
    - "Evolution" driven by statisticians for local utility more than enterprise software



4. It enables reproducible research and analysis
    - Copying and pasting across spreadsheets can lead to mistakes - see excel mistake (**herndon2014does?**)
    - In R, you can essentially automate your analysis, reducing the chance for mistakes and making your analysis transparent to the wider research community as well as reproducible.



3.5 What about Python??

- Python is also a very good language for data science.
- As a more general computer language it can be used for developing broader applications.
- Computer scientists tend to prefer it because its design and syntax is more like a standard computer language.
  - Can make it harder to learn for a non-programmer.
- Main reason to use either tool is based on the use case and your collaborators.

3.6 Two main flavors of R Users

- There are two main flavors of R programmers: Base R users and tidyverse users.
- Base R is the default system - it's more general but not as intuitive or consistent as the tidyverse.
- tidyverse packages are much more convenient for the vast majority of tasks, as long as you drink the Koolaid.
  - They are not always the fastest but for many many uses and data sets they provide a convenient framework

4 This Class - See the Syllabus

4.1 Learning Outcomes

STAT 412/612 will *Develop your competence, creativity, and confidence as a data scientist working with R* so you can ...

1. Execute a regular process to execute reproducible research and analysis using R and R Studio and communicate the results and implications to others.

2. Install and use R packages for specific applications
3. Import data from a variety of external sources
4. Use tidyverse capabilities to transform data to support analysis in R
5. Use tidyverse graphical tools to visualize and understand data

6. Write basic R functions using control and data structures
7. Employ R functions to conduct statistical analysis and inference
8. Generate research or analytical reports and presentations using R Markdown and basic LaTeX capabilities.
9. Deliver an oral presentation describing your data science analysis to an audience .

4.2 Books and Resources:

- All material used in this course is free and online.
  - R for Data Science: https://r4ds.had.co.nz/ Grolemund and Wickham (2018)
  - Tidyverse Style Guide: https://style.tidyverse.org/ Wickham (2017)
  - RStudio Cheat Sheets: https://www.rstudio.com/resources/cheatsheets/

4.3 References

Banton, Caroline. 2020. "Inside Data Science and Its Applications." *Investopedia*. Investopedia. https://www.investopedia.com/terms/d/data-science.asp.

Cleveland, William S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review* 69 (1): 21–26.

Grolemund, Garrett, and Hadley Wickham. 2018. "R for Data Science."

Patil, THDJ, and T Davenport. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90 (10): 70–76.

Wickham, H. 2017. "The Tidyverse Style Guide."