Problem Set 2: Data Science Toolbox

Ana Gallart

February 5, 2023

- Measurement

  Measurements are used to construct insights and policies in quantitative research.

- Statistical programming languages

  We will use R, Python, and Julia in this class. Other languages used for statistical analysis include Stata, SAS, SPSS, Matlab, and JavaScipt. The language that will best suit your needs may vary depending on your field.

- Web Scraping

  Web scraping is the ability to use the publicly accessible data from web pages to inform a data scientist's objective.

  - APIs: Application program interfaces are like rules you have to follow to gt the data from web pages. Companies use them to guard their data.
  - Parsing: Downloading the HTML code is the way to web scrape web pages that don't have APIs. It is risky for websites that have APIs because overdoing this could provoke them to block your access.

- RDDs

  Resilient Distributed Data sets (RDDs) solve the issue of handling large data sets that may not fit on a hard drive. Using a software, huge data sets can be chopped into manageable chunks, executing actions in parallel on the chunks of data. RDDs are build to withstand disruptions in the cluster of computers, allowing for data to easily be transferred to another machine in the case of machine failure.

- SQL

  Structured Query Language (SQL) is a relational software that makes data more usable.

- Visualization

  Visualization allows people to see data in different ways that help analysis.

  - ggplot2 package in R
  - matplotlib in Python
  - Plots.jl in Julia
  - Tableau: software company with interactive data visualization products

- Modeling

  After collecting and cleaning your data, it's time to do statistical modeling. With the following goals- Use data to:

  - Test theories
  - Predict behavior
  - Explain behavior (requires causality)