# PS2

Julie Dawkins

February 1, 2024

## Introduction

- Measurement is a major tool of data scientists: to analyze data, we have to understand what measures are most useful, and what ways we can measure for a certain variable (eg, mean, standard deviation, etc.). This is essential for policy analysis and gathering insights.
- Statistical programming languages are also an essential tool of data scientists. Economists frequently use R and Python, and to a lesser degree Stata and Julia. These are scripted languages, rather than compiled, so they are more human-readable; they are built on other languages like C++.
- Web scraping is a major tool used to gather information from the Internet. It can be done through application program interfaces (APIs) or through downloading the HTML files and parsing their data. Many major websites use APIs, though sadly Twitter/X made theirs pricey :(. Parsing through text is also useful (since many websites don't use APIs) but may result in websites banning you since they are not sure why you are pinging their systems to gather information. This can be done through packages in R, Python, and Julia.
- Handling large data sets is its own challenge, since too much data can crash or stall your computer for hours at a time. This is where resilient distributed datasets (RDD) come in! You need a software like Hadoop or Spark, and with it, the system chops up your data and runs actions parallel. This is great also because any disruption in a computer cluster will not disrupt the whole system; it will transfer it to a new machine.

- Structured Query Language (SQL) is a common language that isn't necessary to handle super-large data, but it is helpful for transforming data into something more usable for a statistical language.
- Visualization is an incredibly useful tool for data scientists because it is essential to communicating results and quickly understanding data. R uses the library ggplot2(), Python uses matplotlib, Julia uses plots.jl, and Tableau is an interactive product for visualization.
- Finally, modeling is necessary to test theories, explain behavior, and predict behavior. Machine learning is helpful for prediction, but strong econometric modeling is necessary for explaining causality.