# Problem Set 2

Misael Serna

February 2021

## 1 Measurement

Measurement is a metric by which data is collected. It is important because there must be a way to determine what information means.

An example of this could be to answer the question of how dangerous a city is. We could choose to measure how many arrest are made weekly.

## 2 Statistical Programming Languages

There are many programming languages that can be used for statistics. Some of the most important are R, Python, Stata, SAS, and Julia.

What language you use is largely related to what you are most comfortable with. However, there are certain features of a language that may be more beneficial to the user's goals.

Programming languages are the actual "Tools" needed to gather data, manipulate it, and display it.

## 3 Visualization Tools

These tools are needed for displaying graphics within the previously mentioned programming languages. The ones we covered in class are ggplot2, matplotlib, and plots.jl.

## 4 Big Data Management Software

When dealing with large amounts of data special considerations must be given. This is because programming languages have policies on how to process data. This is typically stored and being used. For example if the data set is too large to be stored in a computers ram, R will not be able to process the data.

There are tactics that can be used to break down data to make it more manageable. However, in the long term it is much more practical to you RDD software to manage the data. This software breaks the data into chunks that it can then manage. In that way the user can manipulate the data set as though it was regular sized data.

## 5 Data Collection Tools

In the lectures we discussed web scraping. Web scrapping is term to describe collecting data from internet sources. There are two ways to do this. Using a sites API to download data, and to directly download HTML files and extracting information from them.

API's allow companies to safeguard their data so that it is not possible for their information to be easily gathered by a third party. The API is needed to have access to that information.

To avoid this, data can be extracted directly from the HTML. This has its own problems, but the previously described languages have packages that help in doing this.