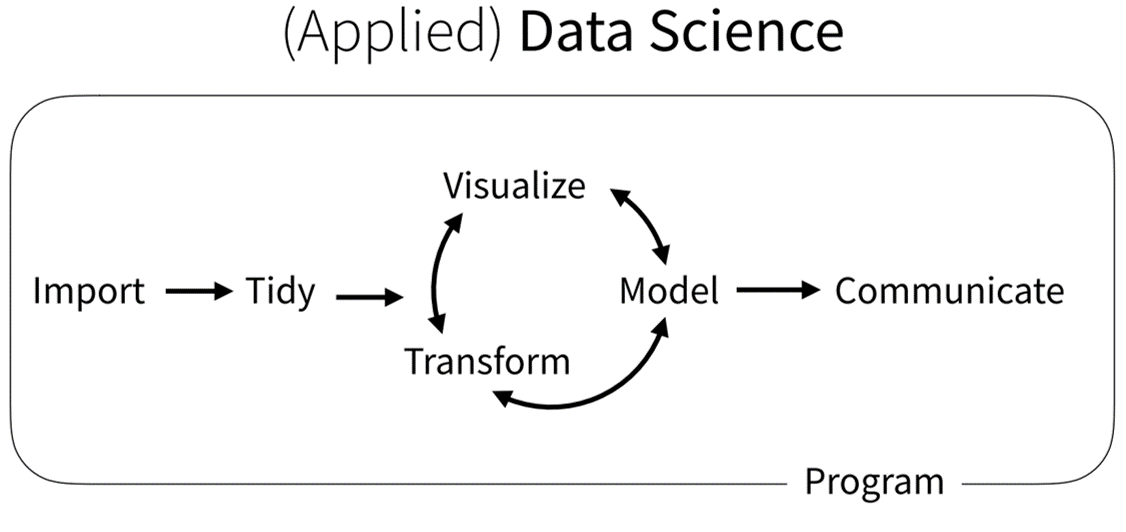
Thinking Quantitatively

## Learning Objectives

1. Understand basic summary statistics.
2. Understand common pitfalls in statistics.

[Lecture Link 1 - Basic Statistics](https://docs.google.com/presentation/d/1CKoyF90yze6sUtrmM4AUgc9aNC1NE3iftuwFU6SQ-J0/edit?usp=sharing)

[Lecture Link 2 - Pitfalls](https://docs.google.com/presentation/d/11UmBqtQeIYfPgR3SUs7F5PRljePmQsXb5UIhriQXFu0/edit?usp=sharing)



Modeling Workflow from Grolemund (2020) - [*The Tidyverse Cookbook*](https://rstudio-education.github.io/tidyverse-cookbook/how-to-use-this-book.html)

# Collecting Data

Data are individual pieces of quantitative (i.e. *numbers*) or qualitative (e.g. *categories*) information. For many scientific projects, data are produced from unique experiments conducted by individuals or groups. Increasingly, data are also produced, or collated, by large groups and released openly online. Neither process is better than the other and many scientists use both types of data to answer scientific questions.

We refer to “collecting data” as the process of getting data into a digital form that scientists will use for their own research project. If you want to know whether a mouse grows slower with a methylated gene, you might perform the experiment on mice with methylated and un-methylated genes, collect data on their weight at different time periods, and enter that into a spreadsheet. If you want to know the number of hospitalizations resulting from COVID-19 in a particular state, you probably won’t directly call hospitals and ask for the data. Instead, you would “collect” data from publically reported hospitalizations and enter those data into a spreadsheet. Students are sometimes confused by this, since we tend to describe science as only getting data directly from experiments or surveys, but both methods of collecting data are valid as long as the source of the data is trustworthy and can be verified.

# Gapminder

In this book, we rely heavily on data released by a group called *Gapminder*, a Swedish foundation that collates global public health data and releases them in a common format in .csv or .xlsx files at <gapminder.org>. They do other things as well, such as provide mesmerizing visualizations of data that will challenge your conventional wisdom. We’ll just use the data for this book, but if you need several days of glorious distraction, visit their website and check out some visualizations.

There are two ways to get data from Gapminder. First, you can download individual spreadsheets from their website here: <https://gapminder.org/data>. For example, to get data on the number of children born per woman in every country since 1800:

* 1. Go to gapminder: <https://gapminder.org/data>
  2. Click on *Babies per woman*
  3. Download the .csv or .xlsx file

Second, you can use the R package *gapminder* to directly load built-in data.

First install the package

#You only need to run the code below once. After it's installed, you don't need to install it again.  
install.packages("gapminder")

Next tell R that you want to use the package with the function *library()*

library(gapminder)