# Introduction to Data Science for Social Science

http://datascience.tntlab.org

Module 1

# Today's Agenda

- Syllabus and Course Website
- What is Data Science?
- DataCamp Demonstration
- R and R Studio Demonstration
- Programming Languages: A Decision I Made For You
- Course Organization

# Syllabus and Course Website

- http://datascience.tntlab.org

# What is Data Science?

- **Not really a thing.**

- Can refer to:
  - Data-based decision making
  - Exploratory data analysis
  - Predictive modeling
  - Computer-assisted data analysis
  - Programming

- Typical elements:
  - Programming
  - Emphasis on predictive modeling

# New Terms Often Refer to Things You Already Do

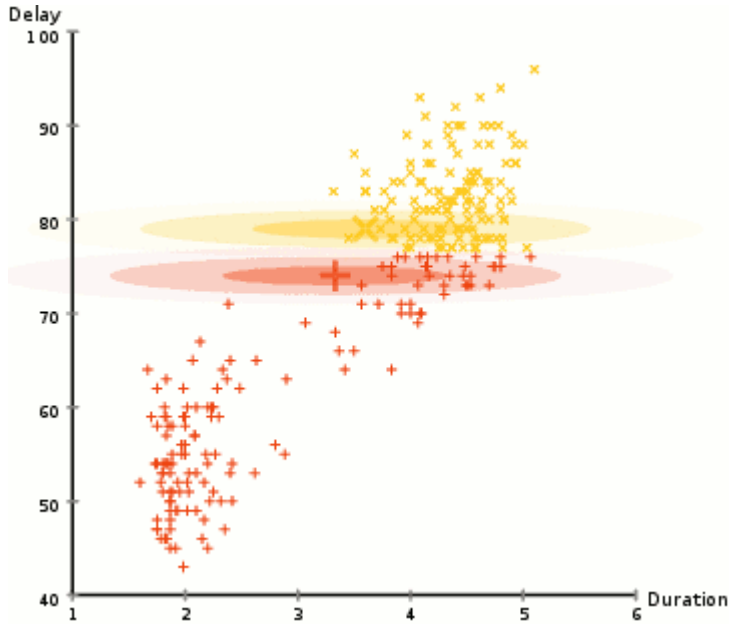
- Example: "Algorithm"
  - Refers to any step-by-step procedure that can be applied by a computer with a set input and output
  - Example algorithm:
    - **Input format:** one vector of integers
    - **Process:**
      1. Create two new variables, both with value zero (0), called *m* and *c*
      2. For each value in the input vector:
         1. Add that value to *m*
         2. Add 1 to *c*
      3. Divide *m* by *c* and return this value as output
    - **Output:** one floating-point number

# How is Data Science Different from Statistics?

- A lot of data science can be considered a subdiscipline of statistics, **computational statistics**. Consider "EM algorithms."

- In the job market, a lot of people hiring **data scientists** just want to hire something that can look at existing data and help them make more money.
  - Sometimes these people get saddled with basic data analyst jobs.

- Many data scientists get annoyed if you tell them they are actually statisticians.



# How Data Science is New

- I like to define **data science** as the **engineering** task of taking ambiguous, ill-defined, or unclean information, quantifying it, and drawing interesting, relevant, generalizable insights from it.
  - Many of the techniques developed for this purpose can be used to improve existing psychological processes and techniques.
  - Many of these techniques create opportunities to analyze data in ways that social scientists typically don't, allowing for better triangulation of theory.
- Hold on a minute: **Data engineering** is different still.
- "Data scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician" – Josh Wills

# Pause the Video and…

1. Create an account on datacamp.com.
2. Install R and R Studio on your system.
   - https://cran.r-project.org/bin/windows/base/
   - https://www.rstudio.com/products/rstudio/download/
3. Install Git

# R and R Studio Demonstration

- Basic R
  - "Base R" vs. Packages
- R Studio
  - CRAN (Comprehensive R Archive Network)
  - R vs R Studio
  - Panes
  - Workspaces
  - Keyboard Shortcuts
  - Built-in Command Help
  - Git
- "The R Mindset"

# A Programmer's Mindset

- You develop code; you don't "write" it.
  - A small quantity of code can take a while.
- Don't learn toolkits.  Learn the language.
- Think algorithmically, not procedurally. Inputs, processes, outputs.
- Don't repeat yourself.
- Use (and develop) abstractions.
- Create data pipelines, not standalone code.
- Build test code.

# Thinking Like a Programmer

- Remember our mean calculation algorithm:
  - **Input:** one vector of integers
  - **Process:**
    1. Create two new variables, both with value zero (0), called *m* and *c*
    2. For each value in the input vector:
       1. Add that value to *m*
       2. Add 1 to *c*
    3. Divide *m* by *c* and return this value as output
  - **Output:** one floating-point number

- Write an algorithm (right now!) to find the largest number in a set of numbers.

# A Decision I Made For You

- There are currently two major programming languages for statistical computing: R and Python

- R currently has much greater popularity with the social science crowd and Python with the computer science crowd; however, an applied data scientist will generally need both
  - R is better for statistical analyses and visualization (sort of)
  - Python is better at project-scale, for big data, and for handling text]
  - Neither is particularly good for "production"

  - For basic tasks, they have similar functionality

- If you end up alt-ac as a data scientist, you will probably, eventually, need to use both
  - Fortunately, these really are *languages*.