Adventures in Statistics:
Encounters with Big Data

JEFF KNISLEY
EAST TENNESSEE STATE UNIVERSITY

MAA SESSION ON ADDING MODERN IDEAS
TO AN INTRODUCTORY STATISTICS COURSE

JOINT MATH MEETINGS, 2013

faculty.etsu.edu/knisleyj knisleyj@etsu.edu math.etsu.edu/netlogo

Outline of the Talk

- Randomization Test (Netlogo)
- Bootstrapping (Netlogo)
- Big Data!

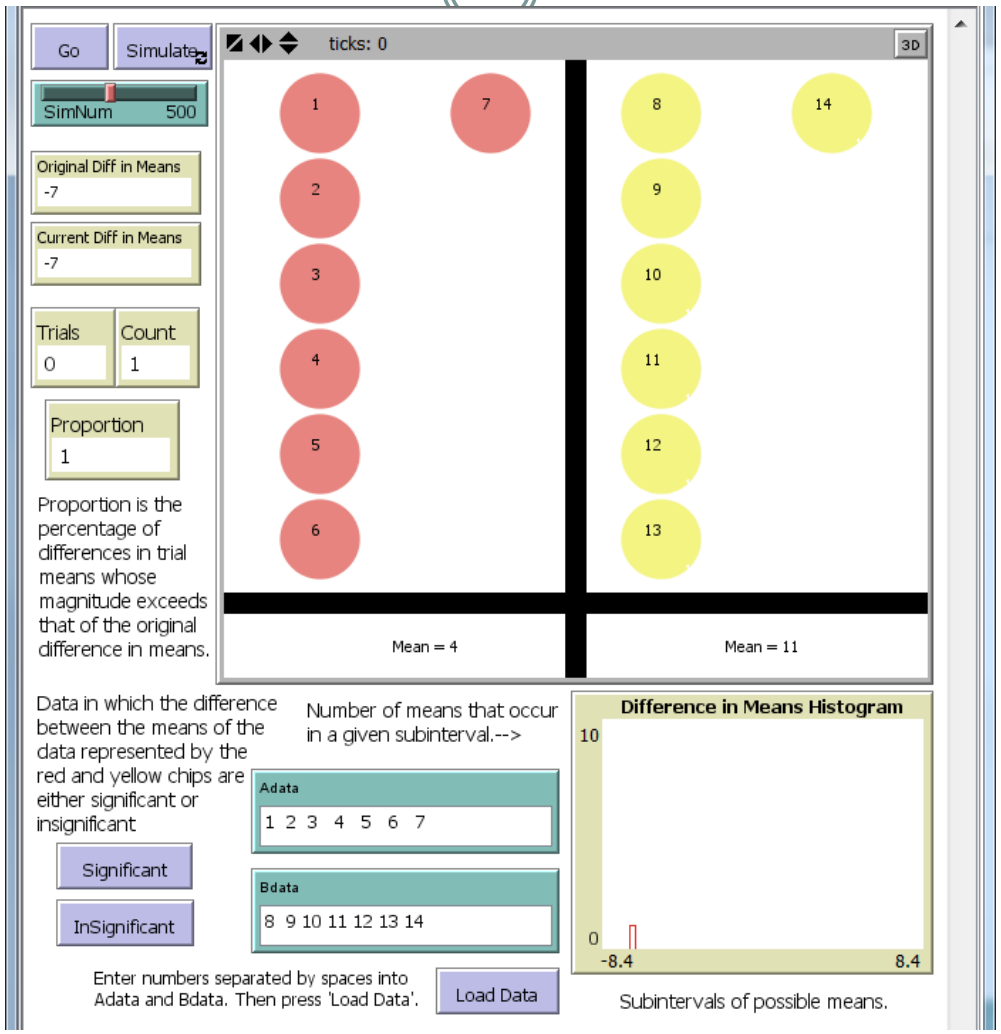
Note: Abstract a bit ambitious for an Introductory Course

- Main Goal of the Talk:
  - Students need non-programming, technology-driven, immersive experiences with large, complex data sets
  - R and Python are good for producing results, but focus is on syntax and programming, not statistical ideas/methods

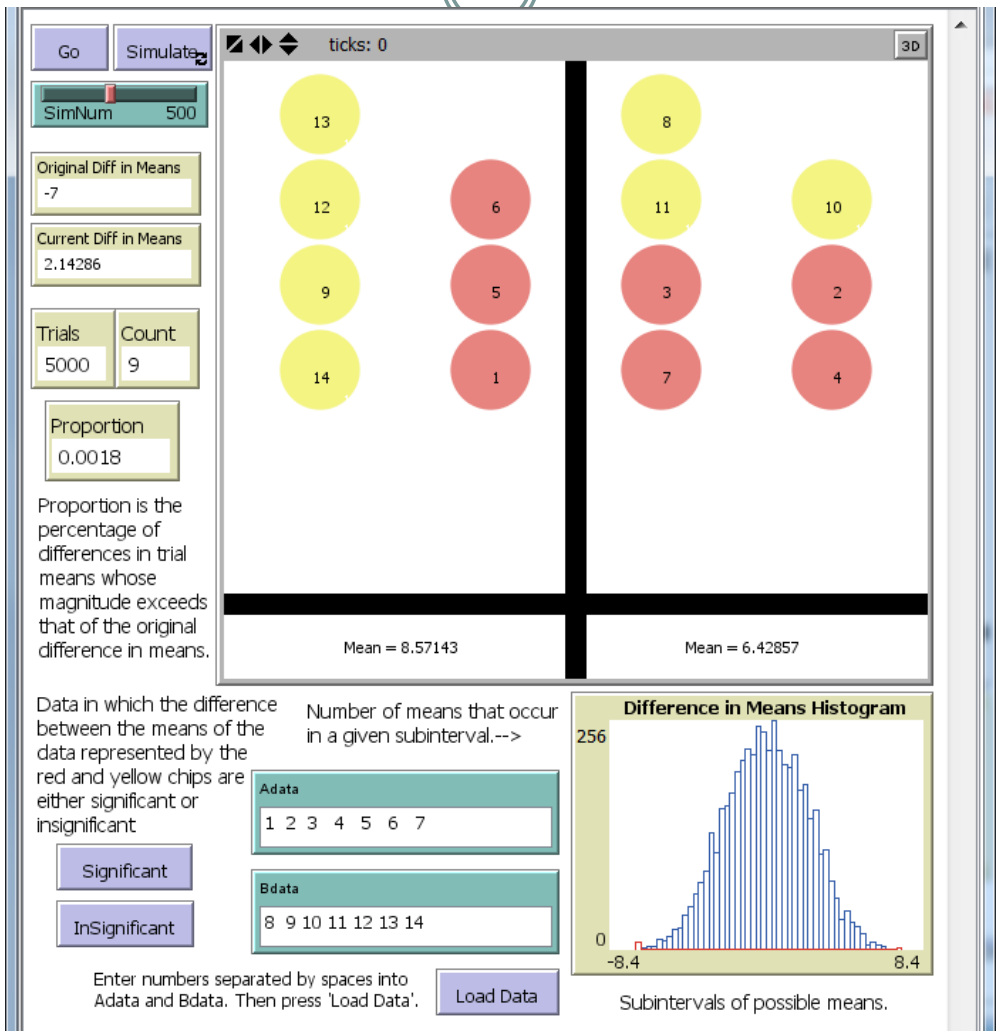
The Introductory Statistics Course

- Students with Traditional Backgrounds
  - Limited or no exposure to statistics, complex data, or computation
  - Calculator dependent. Little if any programming
- Goals
  - Experience with complex data
  - Initial foundation for programming and computation
- First Context:
  - Randomization: Red/Yellow Chips labeled with data. Test for significance in difference of the means of the two sets
  - Netlogo or Python to illustrate the “standard” classroom example and subsequently used as tool for analysis
  - Then move to R or more Python

Randomization test in Netlogo



Randomization test in Netlogo



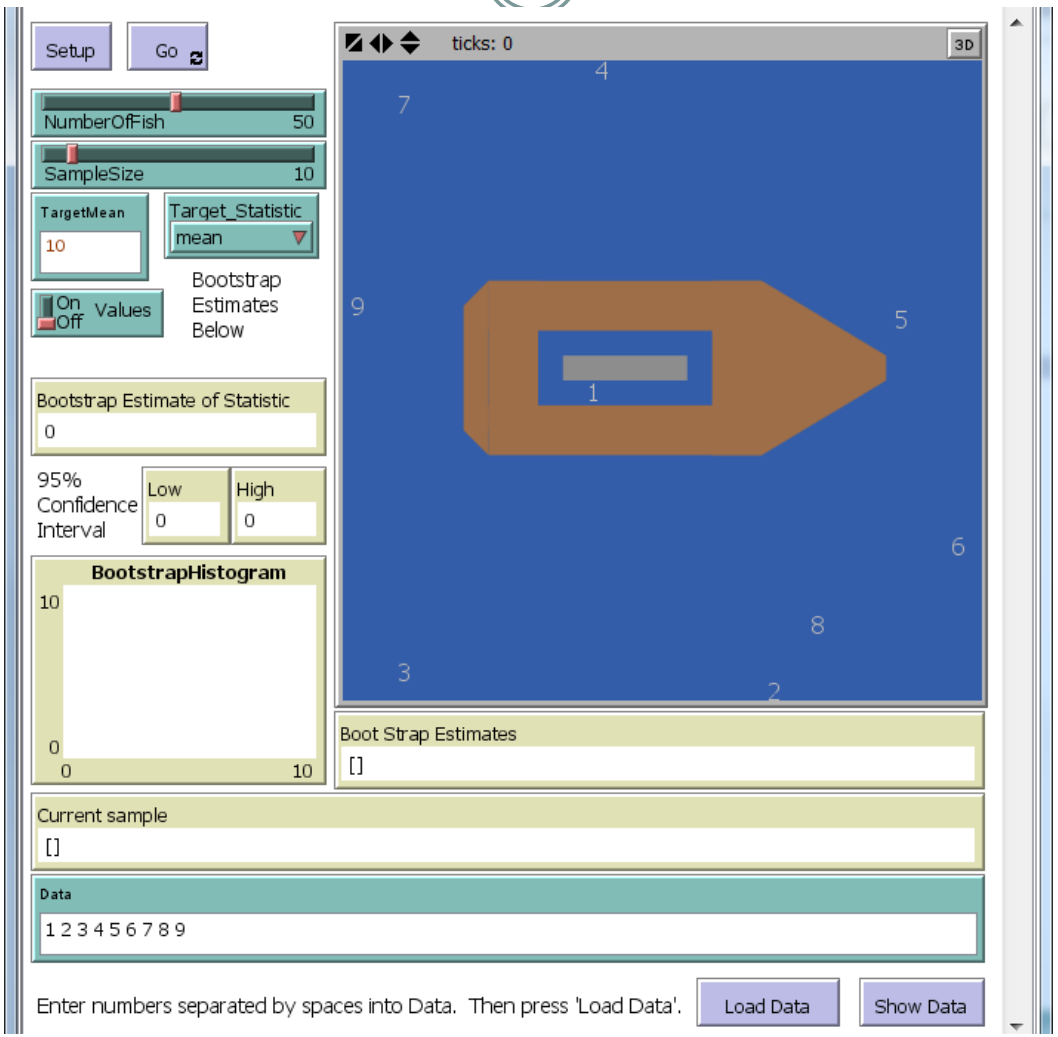
The Bootstrap

- Bootstrapping in statistics is a simple idea that nonetheless can be difficult for some to understand.
  - We start with a Netlogo applet that provides a visual illustration of the bootstrap process.
  - A lake has many sizes of fish, including possibly a few "whoppers"
    - Fish represent a "complex data set"
    - Any "whoppers" are outliers. The distribution is not normal.
- Bootstrapping as "catching fish" is obvious
  - As the fish pass under the boat, they are caught, their length is measured, and then they are released.
  - After a five tick delay to allow the fish population to redistribute, the process is repeated, possibly catching the same fish again
  - i.e., sampling with replacement.

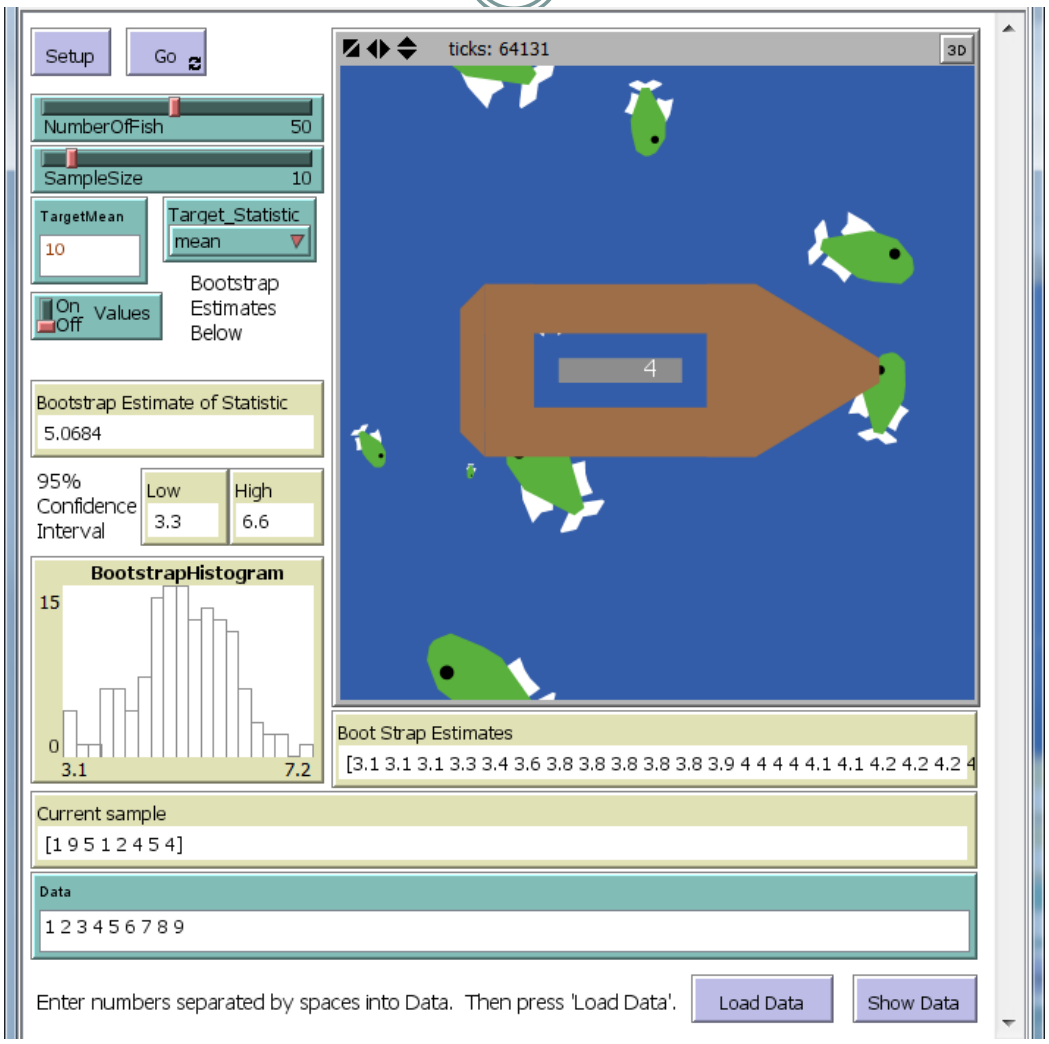
The Bootstrap

- Once a sufficiently large sample is obtained
  - The statistic (mean or median) is applied
  - The estimate is added to a list of BootStrapEstimates.
- The **BootstrapEstimates** list is an empirical distribution for the statistic
  - Allows confidence interval calculation, for example.
  - The length of a fish is a metaphor for any univariate data.
    - Switching on Values shows the numerical value in place of the fish.
    - Univariate data can also be loaded as space separated values and can also be produced via 'Show Data' in the Data box (negative numbers produce 'red' fish)..

Illustrating the Bootstrap



Illustrating the Bootstrap



What is 'Big Data'?

- Features of Big Data
  - Size: thousands, millions, billions, trillions, … of data points
  - **Complexity**: (much more important than size)
    - Small data sets are not complex
    - Not all large data sets are complex
    - Complexity = an underlying structure whose properties are also a goal in the data analysis
- Note: Due to Length, I have omitted discussing Python directly. We'll just look at its consequences

  (starting with the network on the next slide)

The (publicly accessible) ETSU.EDU Network

Colored by number of web-bot sits, from green (low) to red (hi).

April 19, 2012

[ARCHIVED CATALOG]

HELP

[Add to Catalog] Print-Friendly Page

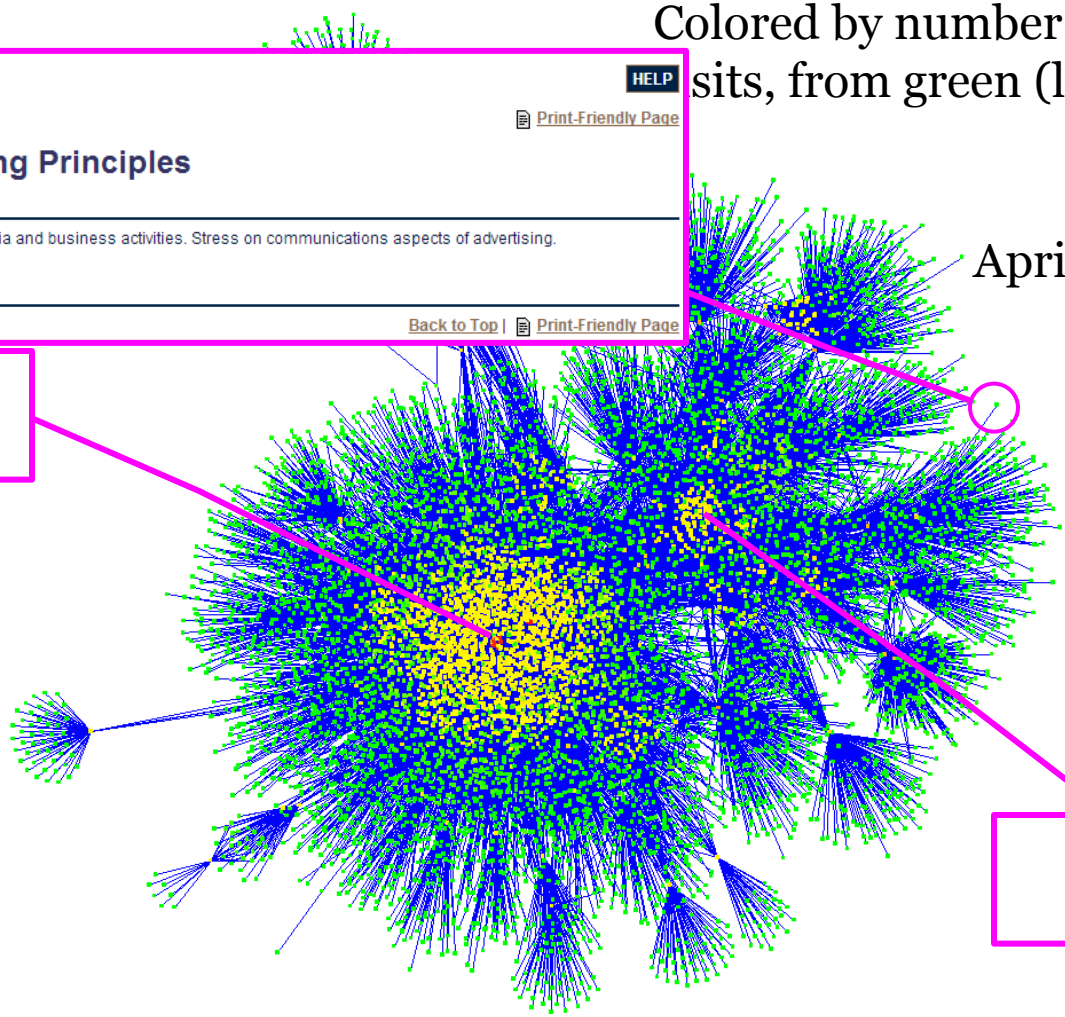
ADVR 3240 - Advertising Principles

(3 credits)

Advertising fundamentals in relation to the media and business activities. Stress on communications aspects of advertising.

[Add to Catalog] Back to Top | Print-Friendly Page

www.etsu.edu

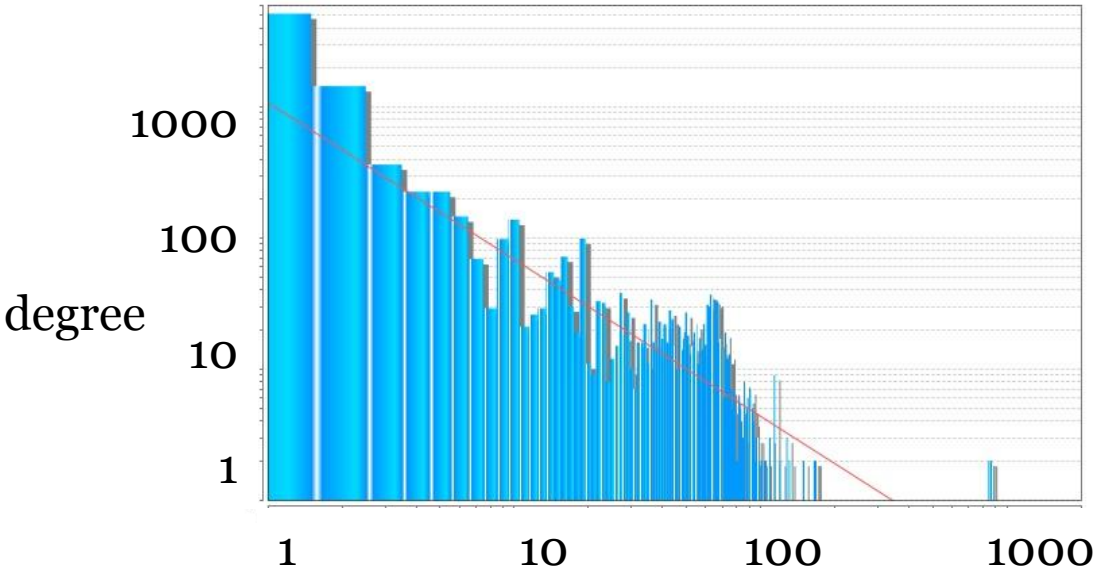


catalog

Webcrawler and network processing via Python

Big Data and the ETSU.EDU network

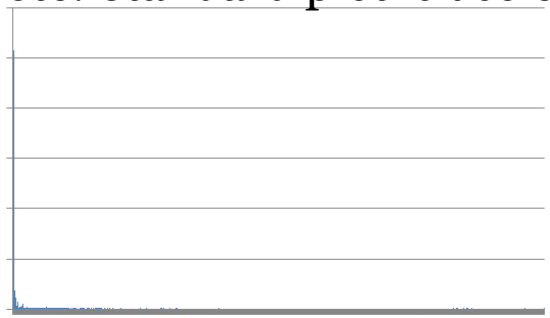
- The network is complex (just look at it!)
  - The complex structure is important
  - Data from such a network is consequently complex
- A Univariate Dataset
  - The degrees (number of pages each page is connected to)



Average degree is 10.94

Log-Log plot

Note: Standard plot is useless



Why Python (and R)?

- Big Data has a complex structure (e.g., networks)
  - Python and R, in particular, allow scientists to represent this structure without having to understand it a priori
  - Abstract concept of a *DataFrame* allows us to work with complex data in Python and R
    - Complex Data requires complex representations
    - Indexing is non-trivial
  - See for example, Python Pandas tutorial at http://www.youtube.com/watch?v=w26x-z-BdWQ (3 hours long but can Fast Forward quite often)
- Each Big Data project tends to require modifications to existing implementations

Big Data and the ETSU Network

- We can use the Netlogo experience to motivate the use of the 'bootstrap' package in R!

```
install.packages('bootstrap')
library(bootstrap)

Data =  read.csv('NetLogoAppExample.csv')
Result = bootstrap(Data$nlogoBootstrap,100,mean)
BootstrapEstimates = Result$thetastar
summary(BootstrapEstimates)
hist(BootstrapEstimates)

# Now with the ETSU Data
Data =  read.csv('ETSUDegreeSequence.csv')
Result = bootstrap(Data$degrees, 100, mean)
BootstrapEstimates = Result$thetastar
summary(BootstrapEstimates)
hist(BootstrapEstimates)
```

Note: R itself is not a good means of showing students what is going on with big data

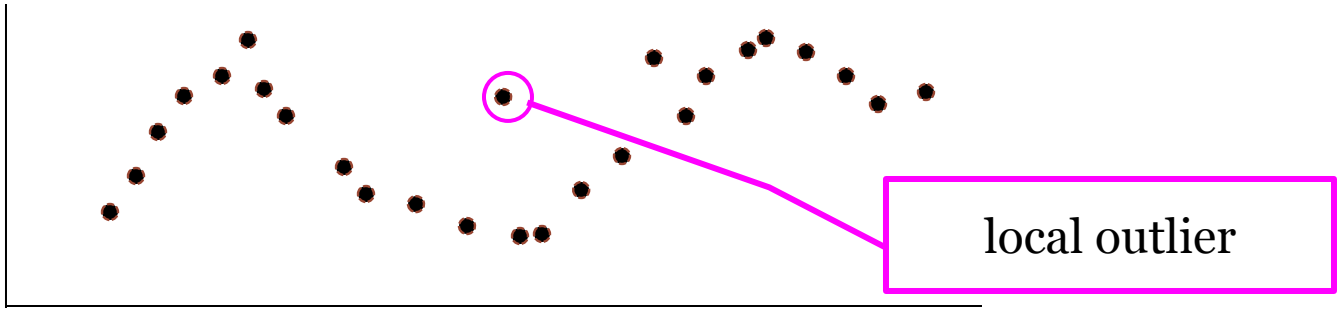
Must focus on syntax

Not obvious what is happening "under the hood"

R is great *once they know what is going on!*

Other Examples of (Big) Complex Data

- Network Analysis Examples
  - Social network analysis
  - Systems Biology
- Time Series examples
  - Information from multiple time series (ekg / eeg)
  - Local outliers in financial data (very common)



THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?