Lecture 01: Course Introduction

Taylor Arnold

John Tukey

"The best thing about being a statistician is that you get to play in everyone's backyard."

About me

I am an applied statistician. A sampling of projects and datasets I have worked on include:

- cell phone telemetry
- emergency room patient flow
- finding holes in specialized medical coverage in rural US
- Canadian court case citations
- Olympic figure skating scoring
- auto insurance risk factors
- 170k documentary photographs from the 1930's
- treatment outcomes for open-angle glaucoma
- detecting radicalization from social media data
- financial warfare

Statistics?

Hans Rosling

Hans Rosling's 200 Countries, 200 Years, 4 Minutes:

- https://www.youtube.com/watch?v=jbkSRLYSojo

NYTimes

Gun homicides in New Zealand are about as common as deaths from falling from a ladder in the United States.

- http://nyti.ms/28yRifm

FiveThirtyEight

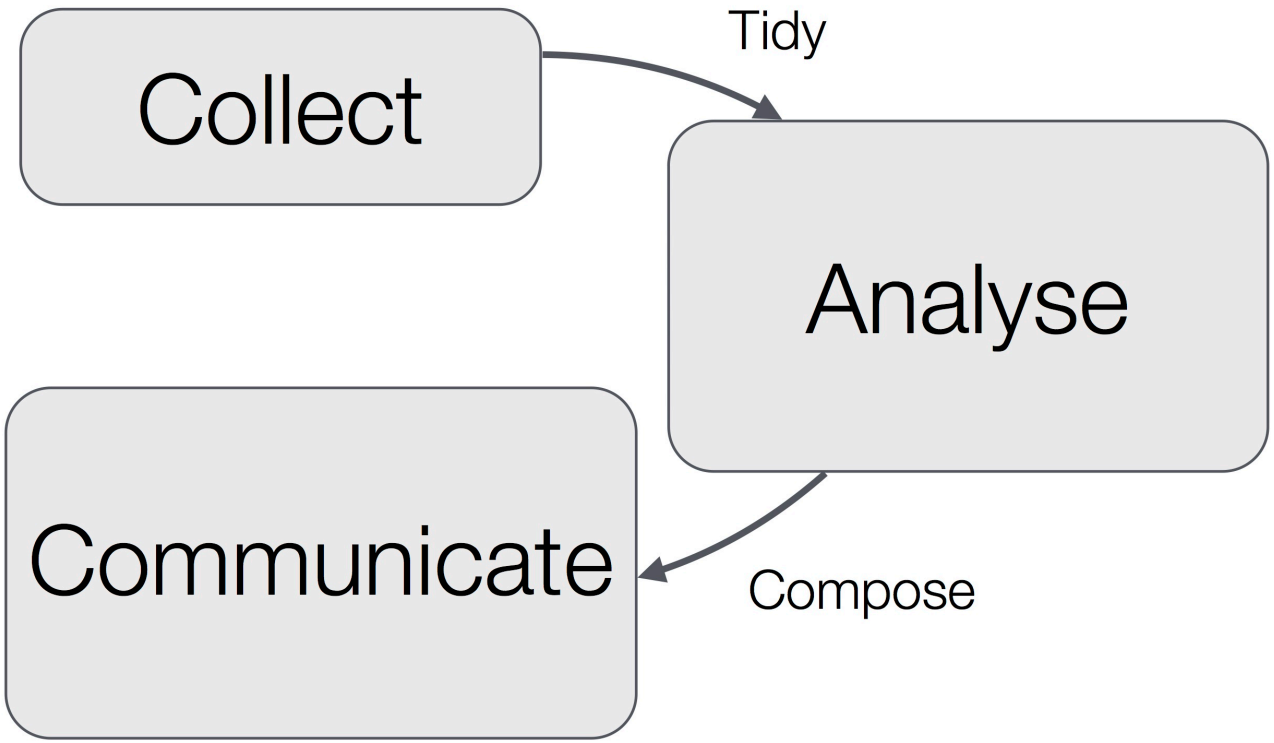
Steroids Probably Aren't Causing Baseball's Power Surge

- http://53eig.ht/2aKodni

Data Analysis

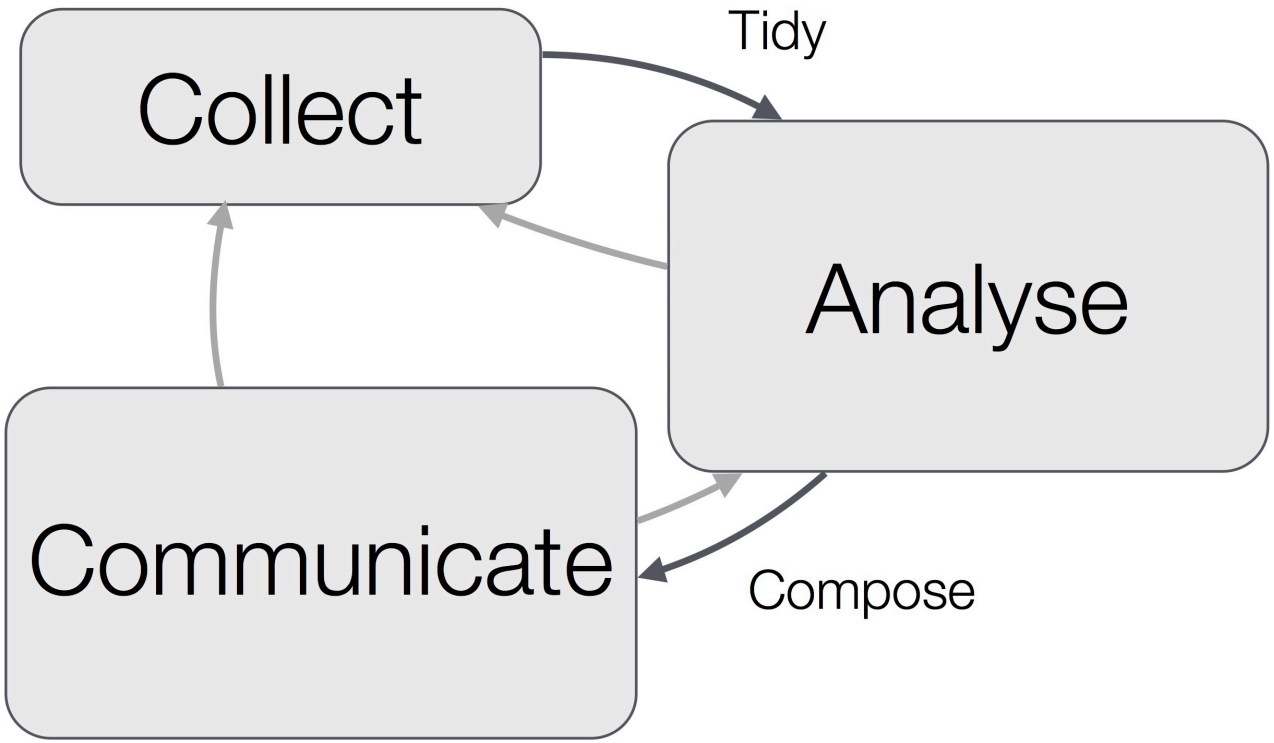
The **data** science process

starts with a question



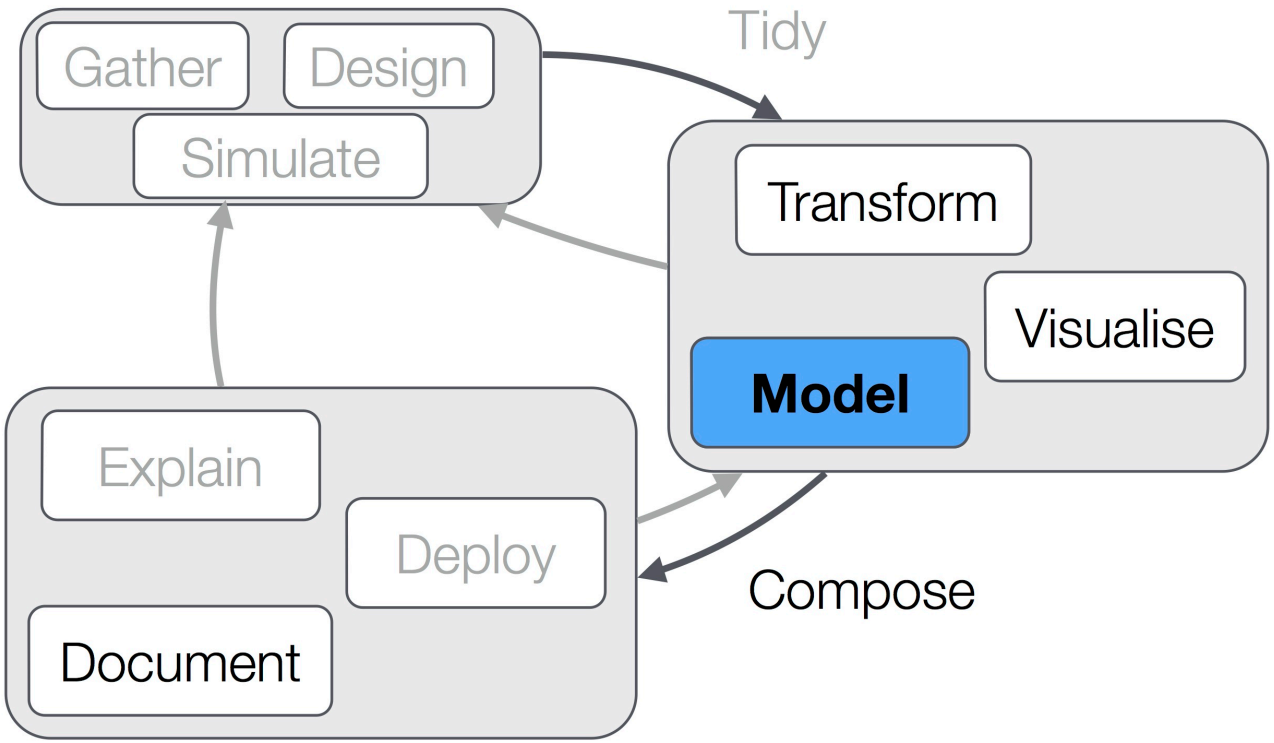
The **data** science process

starts with a question



The **data** science process

starts with a question



Syllabus

Computing

The focus of this course will be on applied statistics and data analysis over symbolic mathematics. To facilitate this, nearly every class assignment and exam will involve some form of computing. No prior programming experience is assumed or required.

R and RStudio

We will use the **R** programming environment throughout the semester. It is freely available for all major operating systems and is pre-installed on many campus computers.

Computers

I strongly recommend using your own machine for this course. The lab computers will be available as well, though I find them to be quite slow and poorly maintained. We will devote a substantial amount of the time learning how to work within the R programming framework.

Computers

I strongly recommend using your own machine for this course. The lab computers will be available as well, though I find them to be quite slow and poorly maintained. We will devote a substantial amount of the time learning how to work within the R programming framework.

Please bring your laptop to our next class meeting!

References

There is no required textbook for the course, but we will pull from a number of sources for material.

Course Website

All of the materials and assignments for the course will be posted on the class website:

https://statsmaths.github.io/stat289

GitHub

All of your work for this semester will be submitted through GitHub, the same platform that hosts our website. You'll need to set up a free account, which we will cover during our next class.

Grades

All grades in this course will be given on as a letter grade. While occasionally possible to receive pluses / minuses or fractional points, these will usually be given a whole letter grade.

Grade Breakdown

There are three components to your final grade:

- Labs, 25%
- Data Reports, 75% (25% each)

Lab and Participation Grades

I expect most students to get full marks (A) for labs and participation. Students found to be delinquent in either will first receive a written warning, followed by an initial 50% reduction (C) in the respective grade, and finally a 100% reduction (F).

Final Grade

I want to make the grading extremely transparent, so your final grade will simply consist of taking your weighted numerical average using the following conversions:

- A: 4
- B: 3
- C: 2
- D: 1
- F: 0

And looking up your grade (rounded to the second digit) on the table provided in the syllabus.

Labs

Most class meetings, particularly in the first half of the semester, will have an interactive lab associated with it. These consist of a set of questions that must be answered with either small snippets of code or short descriptive answers. Your solutions must be uploaded to your GitHub page.

Data Reports

Data reports are short written documents that mix code, graphics, and prose to provide a comprehensive analysis of a data set. These must also be uploaded to GitHub. The format for these reports will be described in the first two weeks of the course.

Exams

This course has no exams, final or otherwise.

Attendance and Late Policy

You are expected to submit work on-time.

You should aim to attend all class meetings.

However I am fully aware that through the course of the semester various issues – illness, sports, and family emergencies – will prevent many of you from attending every class.

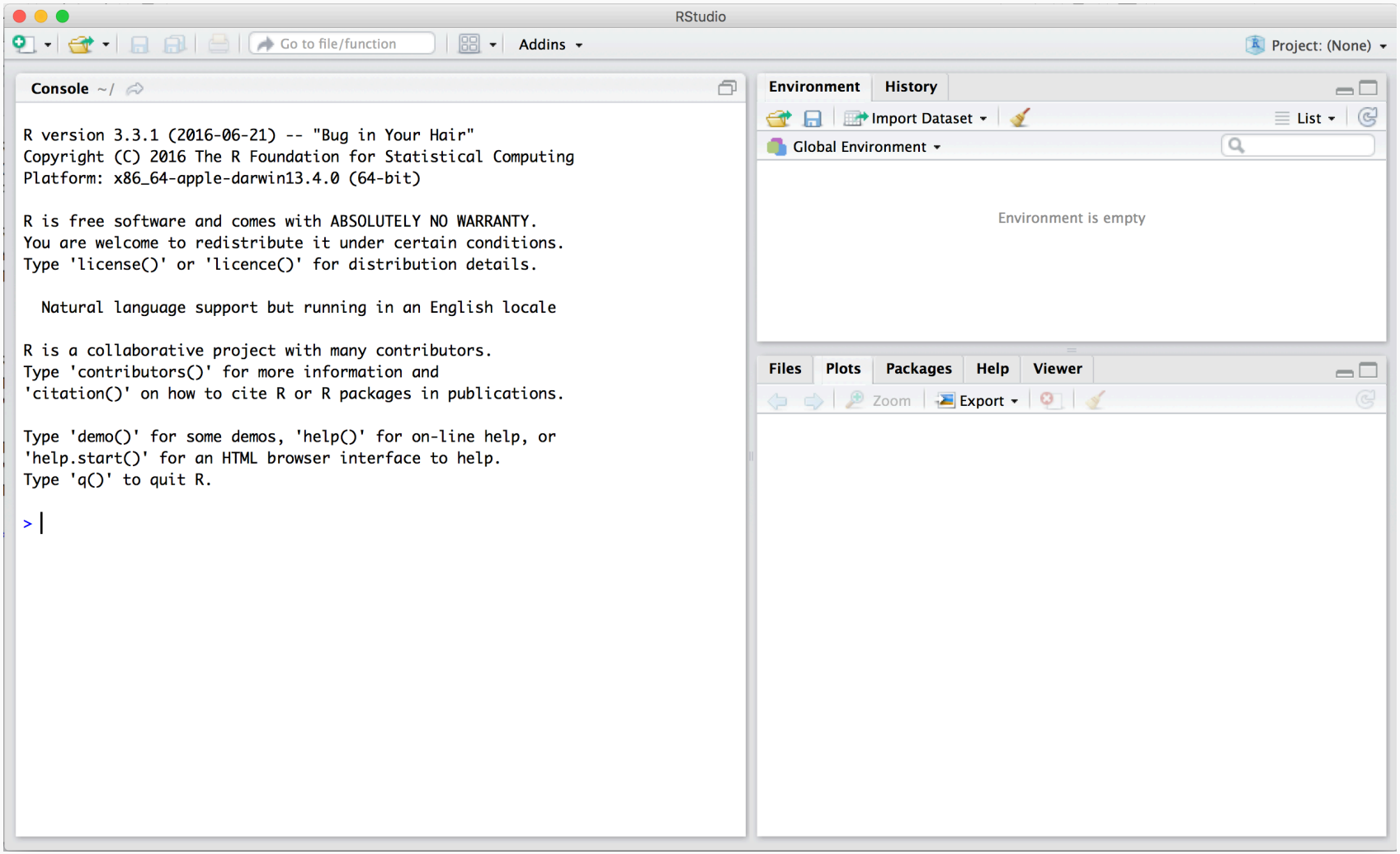
Please be respectful of my time, keep my informed of any issues as early as possible, and come prepared to most meetings of the course.

Questions?

A Bit of R

Launch RStudio

Go ahead and launch RStudio. You should see a window that looks like this:



RStudio Console

The panel on the left is where the action happens. It's called the *console*. Every time you launch RStudio, it will have the same text at the top of the console telling you the version of R that you're running. Below that information is the *prompt*. As its name suggests, this prompt is really a request, a request for a command.

Run a command

Type a simple algebraic expression in the prompt window, such as `1+1`. Hit enter and see the result appear.

Workspace

The panel in the upper right contains your *workspace* as well as a history of the commands that you've previously entered.

Plots and Output

Any plots that you generate will show up in the panel in the lower right corner. This is also where you can browse your files, access help, manage packages, etc.

Plots and Output

Type `example(plot)` and see what happens.

Summary

Questions for You

1. What are you interested in learning this semester?
2. What applications are you most excited about?
3. What's you background, formal or otherwise, in statistics?
4. Which of the following are you generally familiar with?
    - t-tests
    - linear regression
    - Python or R
    - GitHub

For next time

1. Bring a laptop to class so we can set it up
2. Create a free account on GitHub.com