Notes 01: Course Introduction

Taylor Arnold

30 August 2017

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

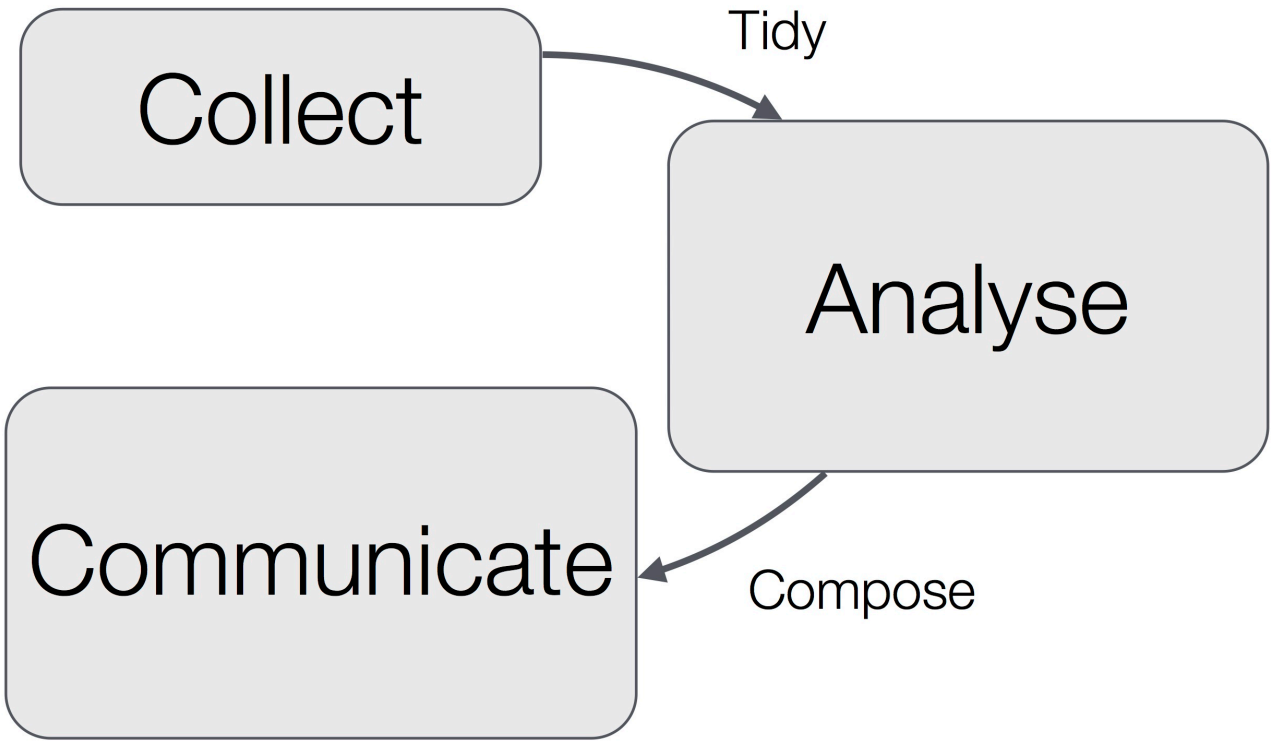
Steroids Probably Aren't Causing Baseball's Power Surge

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

Data Analysis

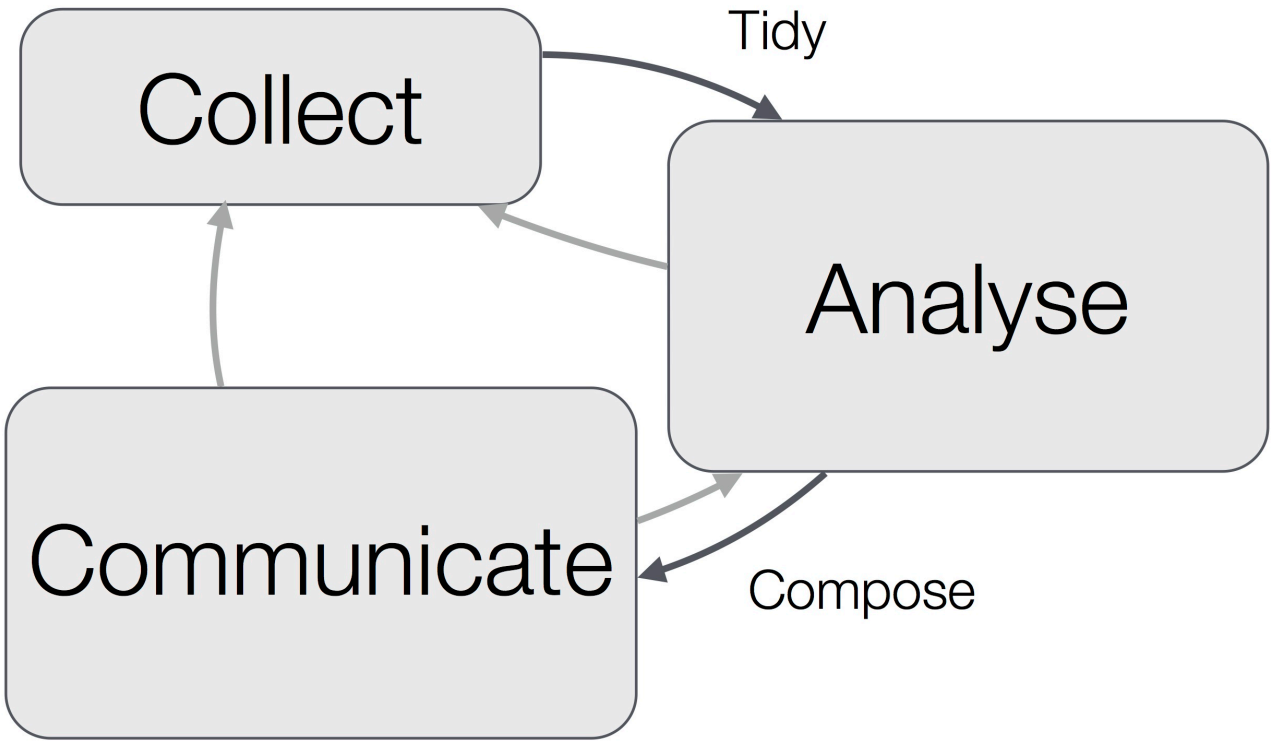
The **data** science process

starts with a question



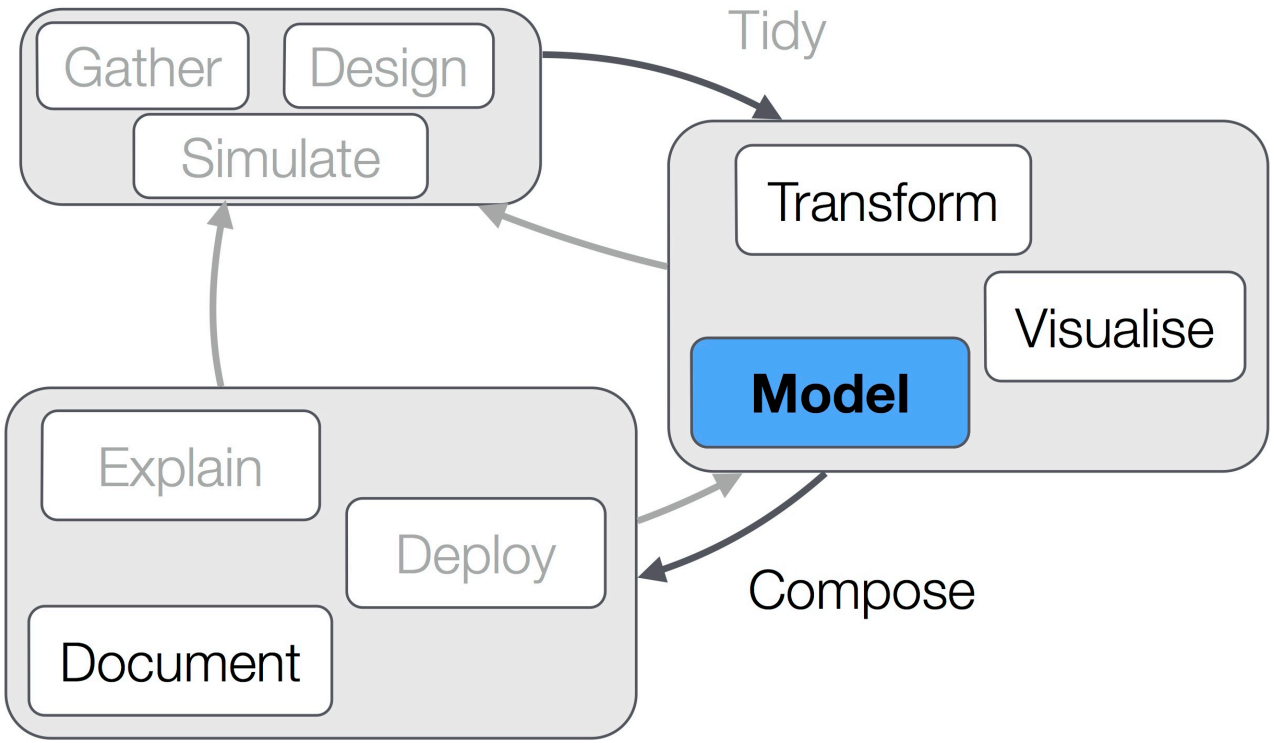
The **data** science process

starts with a question



The **data** science process

starts with a question



STAT 289

Naming

Technically, this class is **MATH 289**, but I'm going to call it **STAT 289** in these notes because it is really not a mathematics course.

Course Goal

The goal for this semester is to learn the skills to collect, analyse, and communicate results about data.

We will do a lot of programming in the course, but it is **not** a programming class.

Versus MATH 209

I will be teaching this class in a similar way that I teach MATH 209. However, we will move at a slightly faster pace and will cover the modelling aspects of data more thoroughly.

I'll assume that you have previously seen hypothesis testing and are at least roughly familiar with linear regression.

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 as either a letter grade or on a 4-point scale.

While occasionally possible to receive pluses / minuses or fractional points, these will usually be given a whole letter or number grade.

Grade Breakdown

There are three components to your final grade:

- Labs, 10%
- Participation, 10%
- Quizzes, 30% (drop lowest 2)
- Data Reports, 50%

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.

Quizzes

There will be short quizzes given during the semester. These will be at the end of class on Tuesdays. All will consist of entirely objective questions, such as True/False, multiple choice, and matching.

Quizzes

There will be short quizzes given during the semester. These will be at the end of class on Tuesdays. All will consist of entirely objective questions, such as True/False, multiple choice, and matching.

You will be able to drop your two lowest grades; therefore, make-up quizzes will be offered only in extreme circumstances.

Data Reports

Data reports are short written documents that mix code, graphics, and prose to provide a comprehensive analysis of a data set. We will have a data report due approximately bi-weekly. 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.

Tentative Course Outline

- WEEK 01 - Introduction to R and RMarkdown
- WEEK 02 - Basic Graphics
- WEEK 03 - Variable Types
- WEEK 04 - Data Collection
- WEEK 05 - Data Manipulation
- WEEK 06 - Simple Linear Models
- WEEK 07 - Multivariate Linear Models
- WEEK 08 - Spatial Data
- WEEK 09 - Theories of Data Visualisation I
- WEEK 10 - Theories of Data Visualisation II
- WEEK 11 - Tidy Data
- WEEK 12 - Relational Data
- WEEK 13 - Strings and Dates
- WEEK 14 - Penalized Regression and Text Processing

Questions?

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