85-309:
Statistical Concepts and Methods for Social and Behavioral Science

Spring 2020

Professor Dan Yurovsky

Why I love statistics

Undergrad in Computer Science at Carnegie Mellon

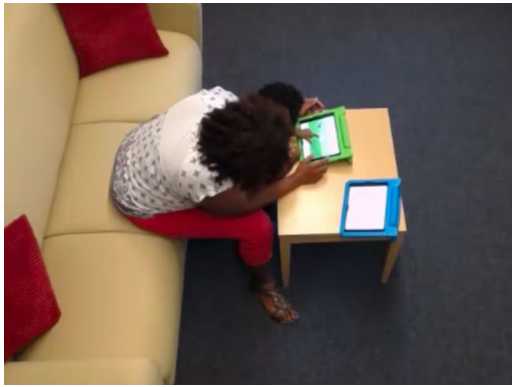
- Interested in AI and Machine Learning
  (basically applied statistics)

PhD in Cognitive Psychology at Indiana University

- Studied how infants learn language
  (basically applied statistics??)

Faculty back here at CMU

- Study how we communicate and learn from each-other
  (how change the statistics of our environment)
- Excited about using “big data” to understand how people learn and develop



https://callab.github.io/

Why you should love statistics too

1. Statistics are a way to cope with the absurd
2. Statistics are the connection between theory and the natural world
3. Statistics are an expression of liberty

Statistics is the Math of Existentialism

“Man stands face to face with the irrational.
He feels within him his longing for happiness and for reason.
The absurd is born of this confrontation between the human need and the unreasonable silence of the world.”

Albert Camus, The Myth of Sisyphus

To understand statistics is to embrace the absurd: *There is no certainty*, *only degrees of doubt*

Statistics connect scientific theories to the world

The artifacts of science are models

All models are wrong, but some are useful

George Box

Because there is no certainty, no model can be *True*.
Statistics is a set of tools for helping us to figure which ones are more useful.

Statistics are an expression of liberty

The fundamental premise of inferential statistics: You could be wrong!

The practice of statistics is *doubt* of authority

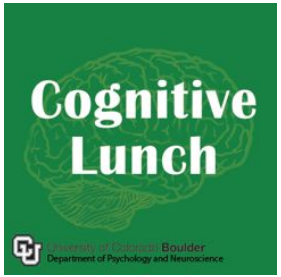
Ubi dubium ibi libertas

Thanks to John Kruschke

Goals for 85–31x/320/3/330/340 and 85–309

Goals for 85-309

A statistical story



A multi-scale approach to ambiguity reduction in word learning
A key question in language acquisition is how children and adults map words to their referents despite the ambiguity in naming events….

Denver 7 – The Denver Channel

Building a statistical model of flooding

Editor's note

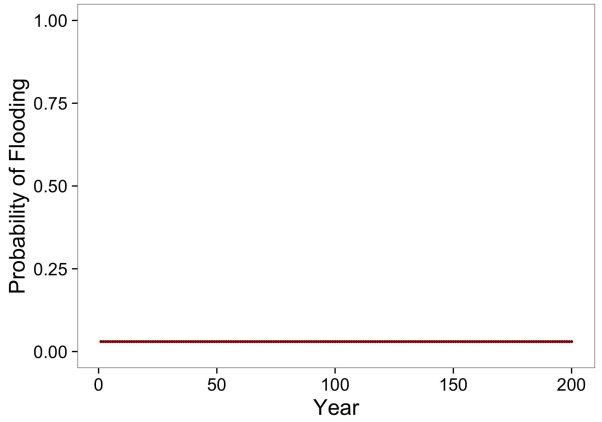
Boulder's 100-year flood: How to help, and how to talk about it

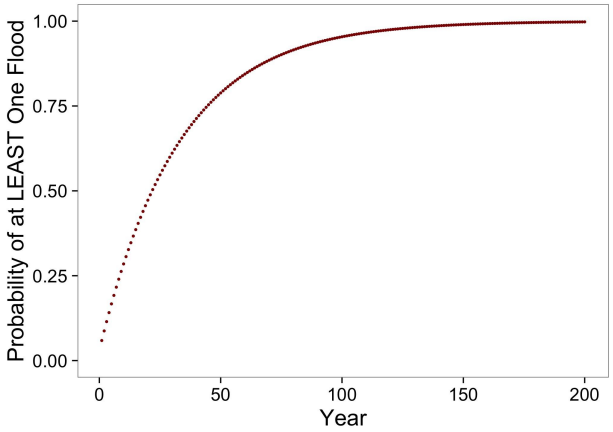
By Jenn Fields

fields@coloradodaily.com

Is the chance of flooding every year an **independent** event?

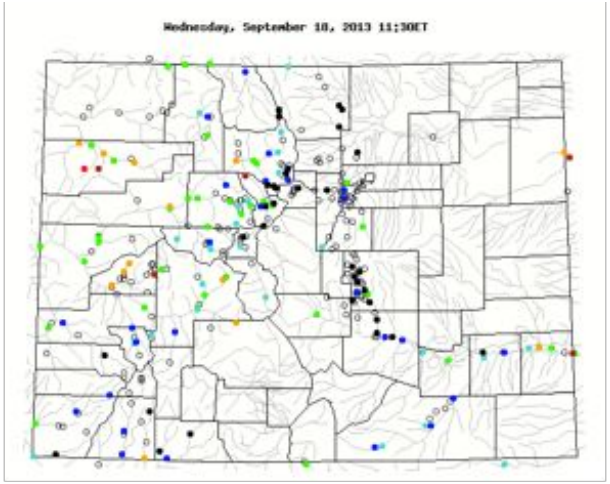
Every year you flip a coin, if it's heads you get a flood.
Only the coin is weighted, and tails happens 97/100 flips.

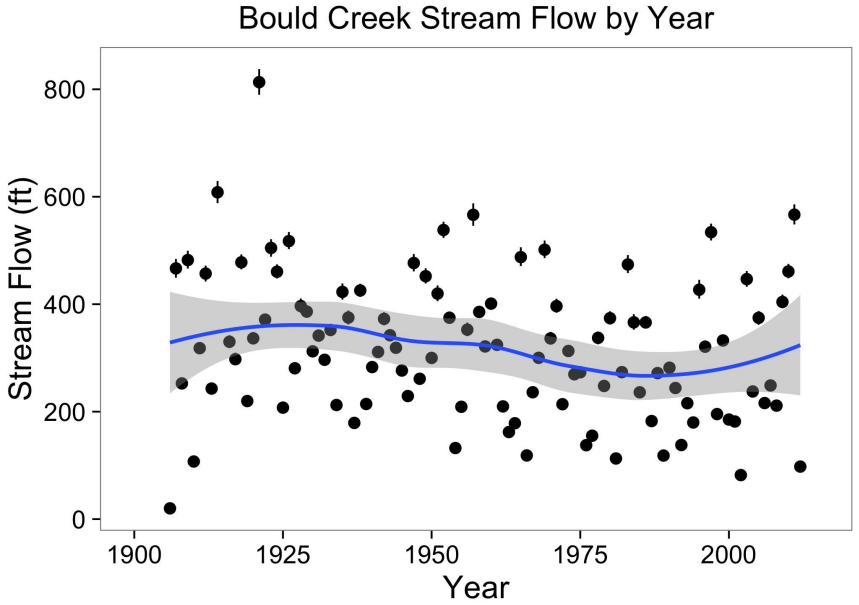




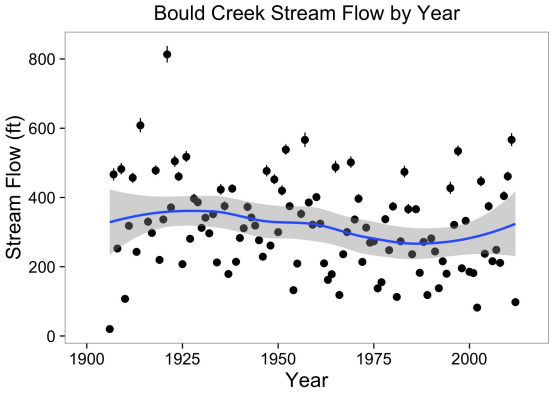
Let's get some data to answer the question

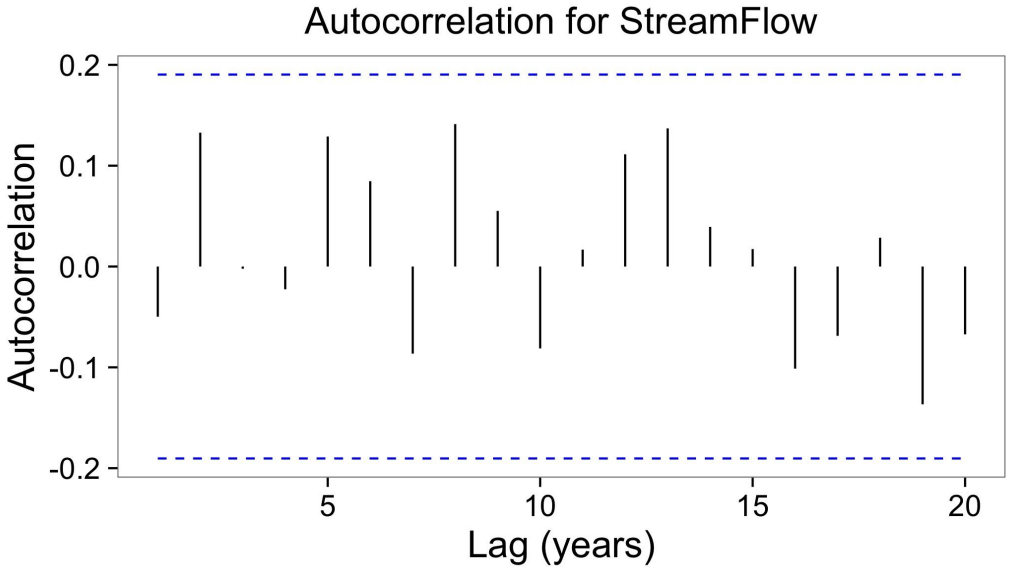
USGS
science for a changing world



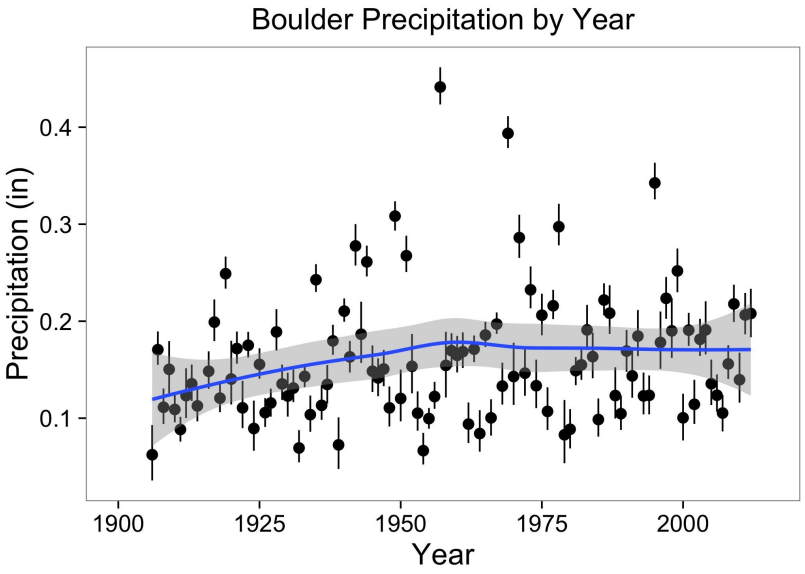


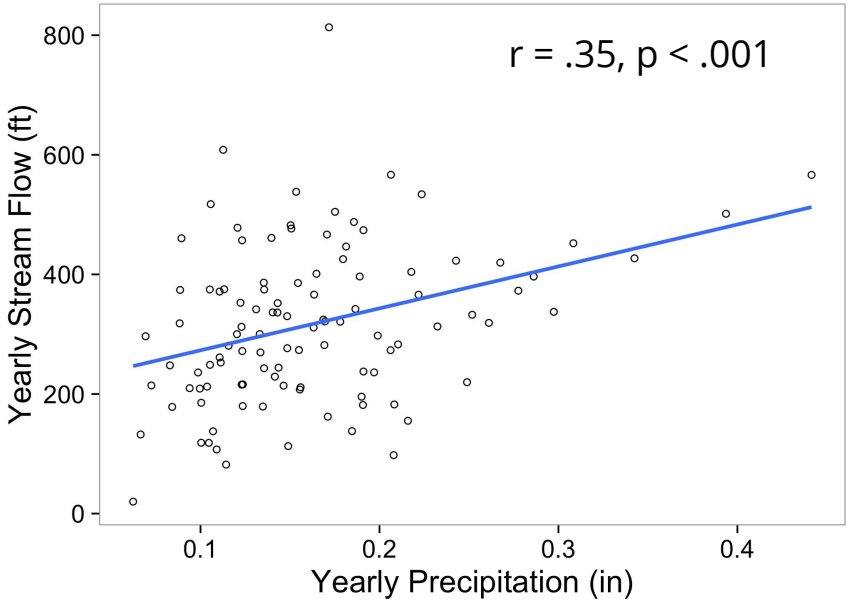
Autocorrelation: A way of testing for independence



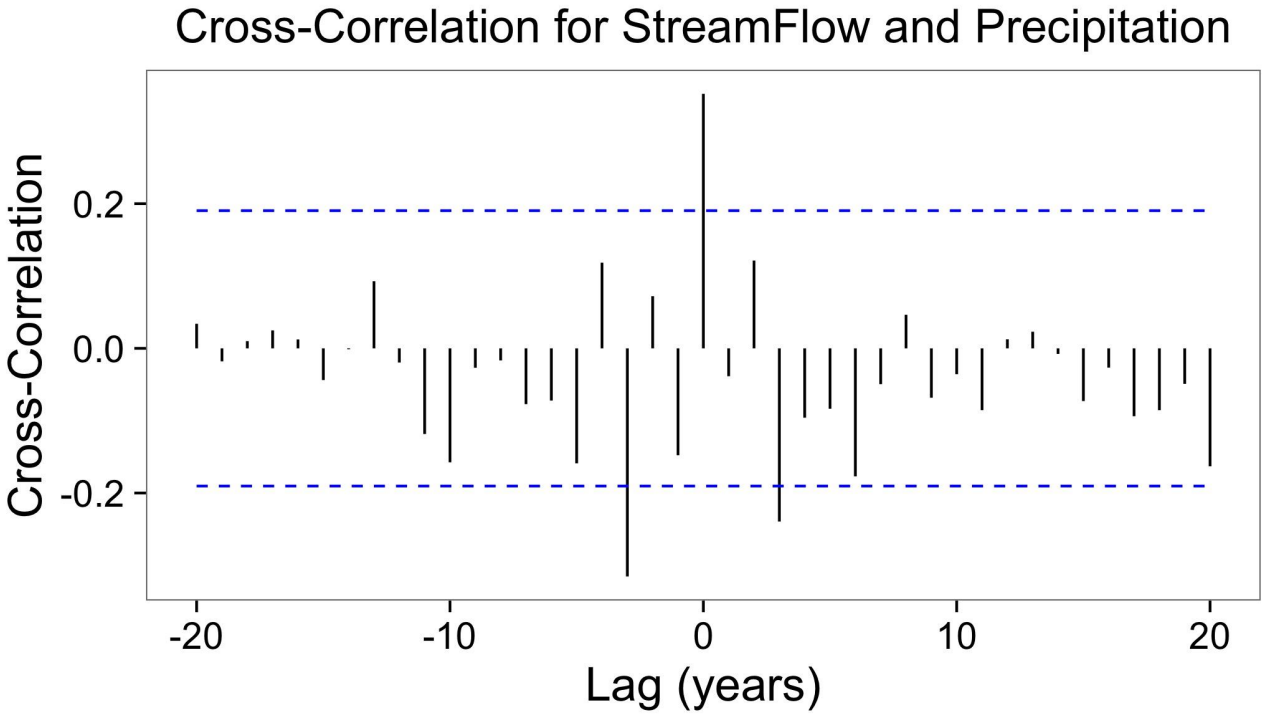


Trying to predict streamflow





Yearly precipitation *predicts* streamflow?



Using statistics to understand the world

1. Come up with a hypothesis about the process that generates data

   “Flooding every year is an independent event like a coin flip”

2. Pose a prediction that would be made by this model

   “Knowing whether it flooded one year does not help you predict flooding the next year”

3. Find data to test this prediction (or at least an approximation) -- **Null Hypothesis Testing**

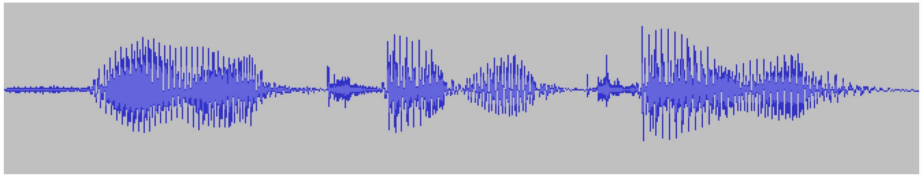
   “Boulder creek levels should be independent from year to year”

4. Ideally, pose an alternative model

   “Creek levels and rainfall are cyclical and have predictable periodicity”

5. Test this prediction

How do you know what words are?



He re ki tt y ki tt y

Word boundaries are not marked by silences! But we can hear them anyway

Thanks to Mike Frank

How do you know what words are?

bigoku vs. dobigo

Thanks to Julie Sedivy

Segmenting words by detecting dependence

olookwhatapprettybaby
whatapprettyshirt
ohlookatthehappybaby
itsprettylatealready
theresababycanyouseeit

If you just heard ty, you can't predict whether you will next hear ba

They are independent

If you just heard ba, you are very likely to next hear by

Thanks to Mike Frank

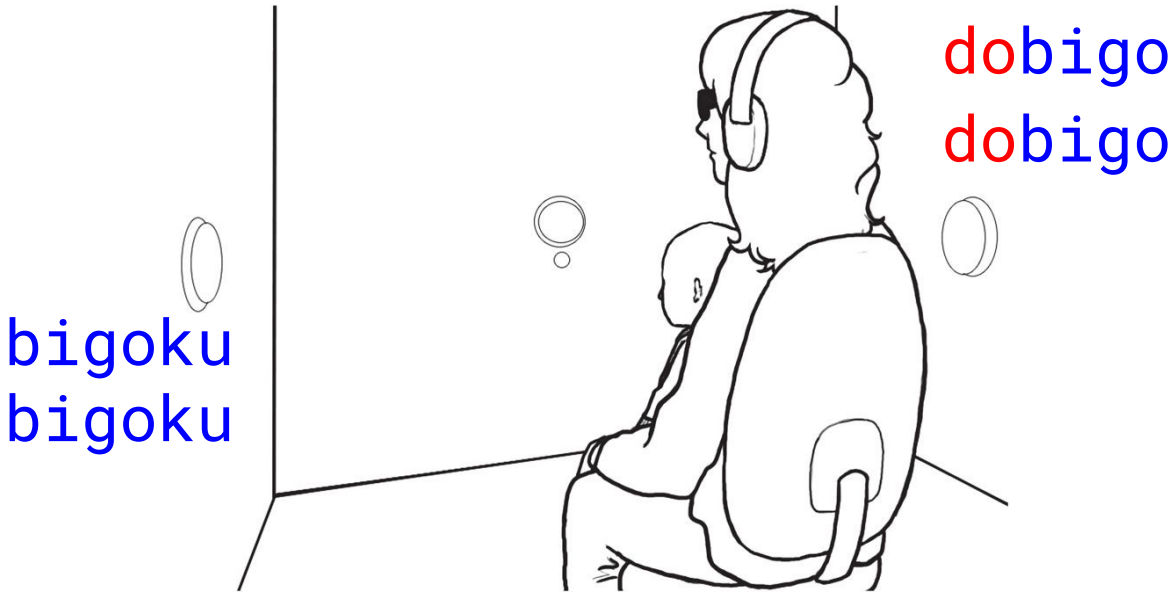
Segmenting words by detecting dependence

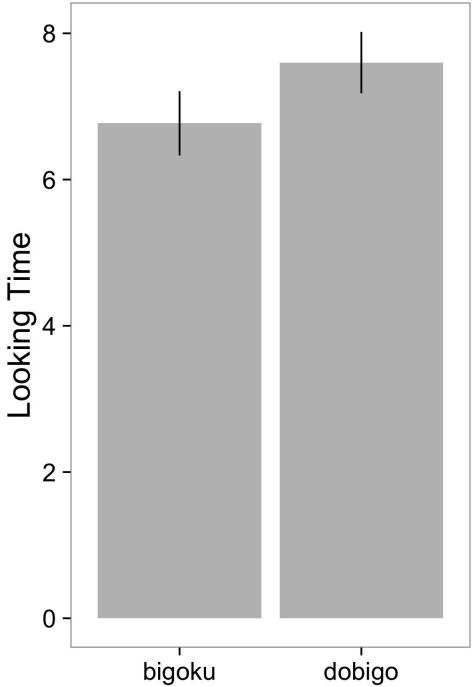
buladobigokudatibabuladotadupabigoku

Test: bigoku (word) vs. dobigo (partword)

Thanks to Mike Frank

Segmenting words by detecting dependence





Saffran, Aslin, & Newport (1996)

By the end of the quarter, you should be able to:

1. Understand how the way that data is collected affects what you can learn from it
2. Use statistical software to summarize this data numerically and visually
3. Build statistical models of the data. Understand which models are better and why
4. Make predictions about what kind of data you would expect to see in the future
5. Ask questions about the data, and make statistical inferences to answer them
6. Present these results in a transparent way to others
7. Understand the claims that others make from data and be able to critique them.

Course information

Teaching Team

Professor	Dr. Dan Yurovsky	yurovsky@cmu.edu
TA	Roderick Seow	yseow@andrew.cmu.edu

We want to help!

Come to our office hours, send us email, ask us questions!

Online Resources

Course Website:
https://dyurovsky.github.io/85309/

- Find syllabus, slides, etc.

Canvas:
https://www.cmu.edu/canvas/

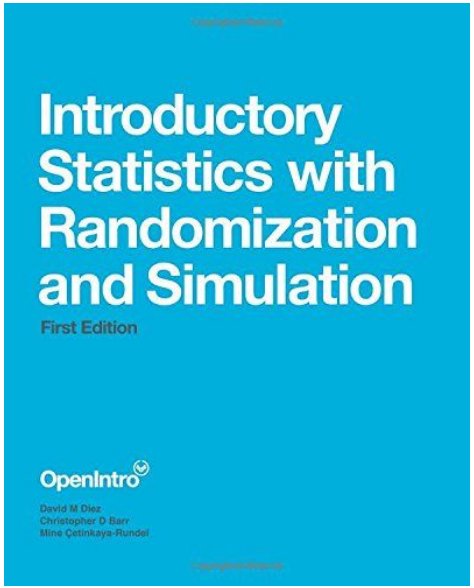
- Submit assignments

Piazza:
piazza.com/cmu/spring2020/85309/home

- Post and answer questions

Two parallel roads to the goal

Theory: Lectures and Textbook



Application: Labs and Project

RStudio®

Assessment and Grading

CAOS Pre and Post tests	5%
Quizzes	10%
Problem sets	20%
Labs	40%
Project	25%

Theory

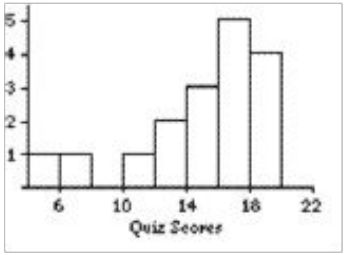
Application

Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) Test

Assessment Resource Tools for Improving Statistical Thinking

https://apps3.cehd.umn.edu/artist/caos.html

e.g. For this graphical display of Quiz Scores, which estimates of the mean and median are most plausible?



a. median = 13.0 and mean = 12.0

b. median = 14.0 and mean = 15.0

C. median = 16.0 and mean = 14.3

d. median = 16.5 and mean = 16.2

You will take a CAOS Pre and Post Test.

These will be graded for completion, not correctness.

Assessing your understanding of theory

Quizzes

There will be a **quiz** every wednesday at the start of lecture (except for this week). Quizzes are designed to give both you and your instructors rapid feedback about you understanding of the theory.

Your lowest grade will be dropped.

Problem Sets

There will be a **problem set** assigned for each of the first 5 units. These are designed to give you practice reasoning about the theory of statistics more deeply. You are encouraged to work together, **but must submit your own work**.

Assessing your understanding of application

Labs

Every friday, you will have a **lab** assignment. These are designed to give you practice applying the theoretical ideas you are learning to thinking about real data.

These will likely be challenging, especially if they are your first exposure to programming. But we are here to help, and so is a sizeable chunk of the internet!

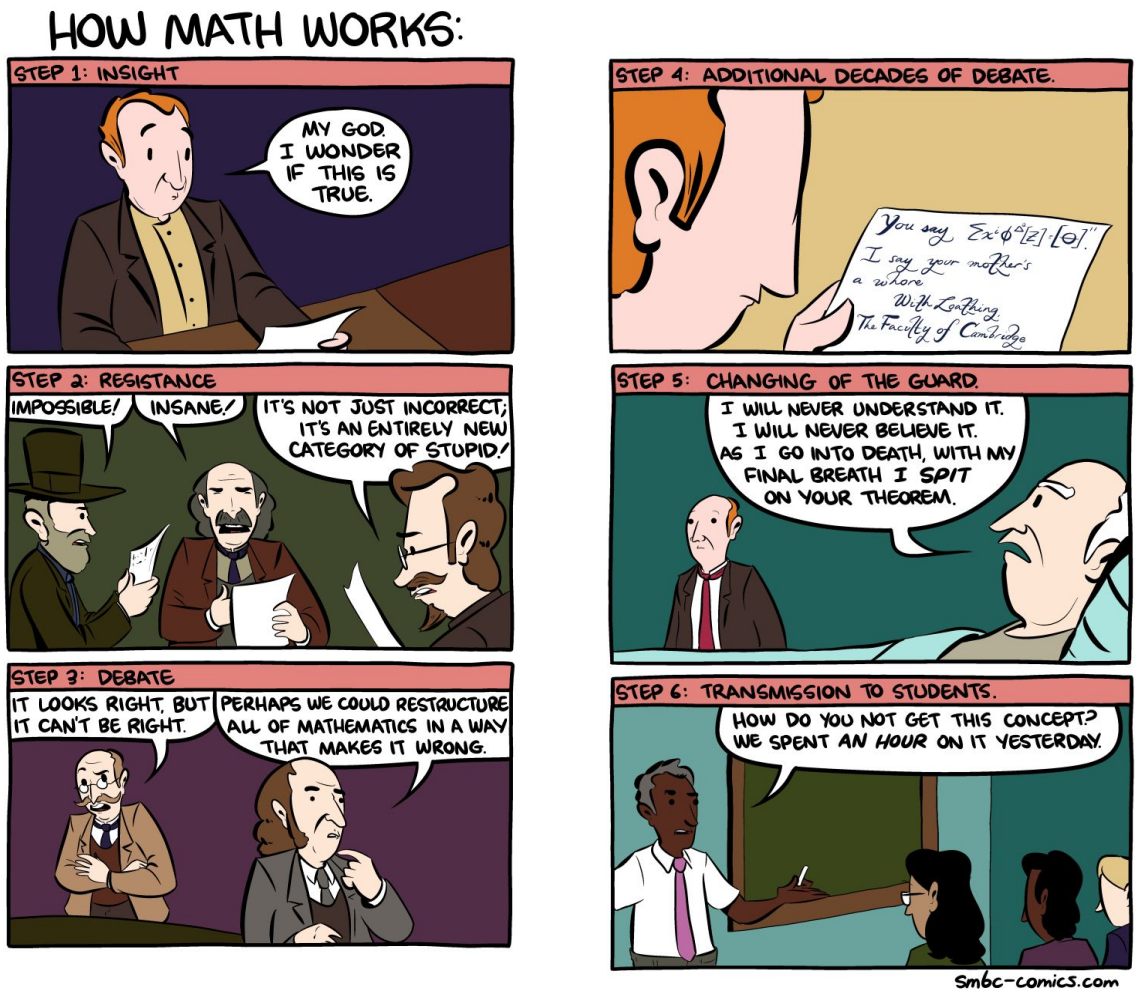
These skills are useful, transferable, and empowering. Seriously, you want to learn this!

Project

The capstone assessment for the class is a **final project**. You will be given a dataset, and your goal will be to show something interesting about it.
Think of this a larger, less structured lab assignment.

If you can do this, you (and we) will know that you really learned something!

The Curse of Knowledge



- These ideas are challenging
- If you don't understand them right away, don't worry!
- They took centuries to develop