CSCI 3022
Intro to Data Science with Probability and Statistics

What is Data Science?

What is Data Science?

Seriously. What do **YOU** think it is?

What is Data Science?

Seriously. What do **YOU** think it is?

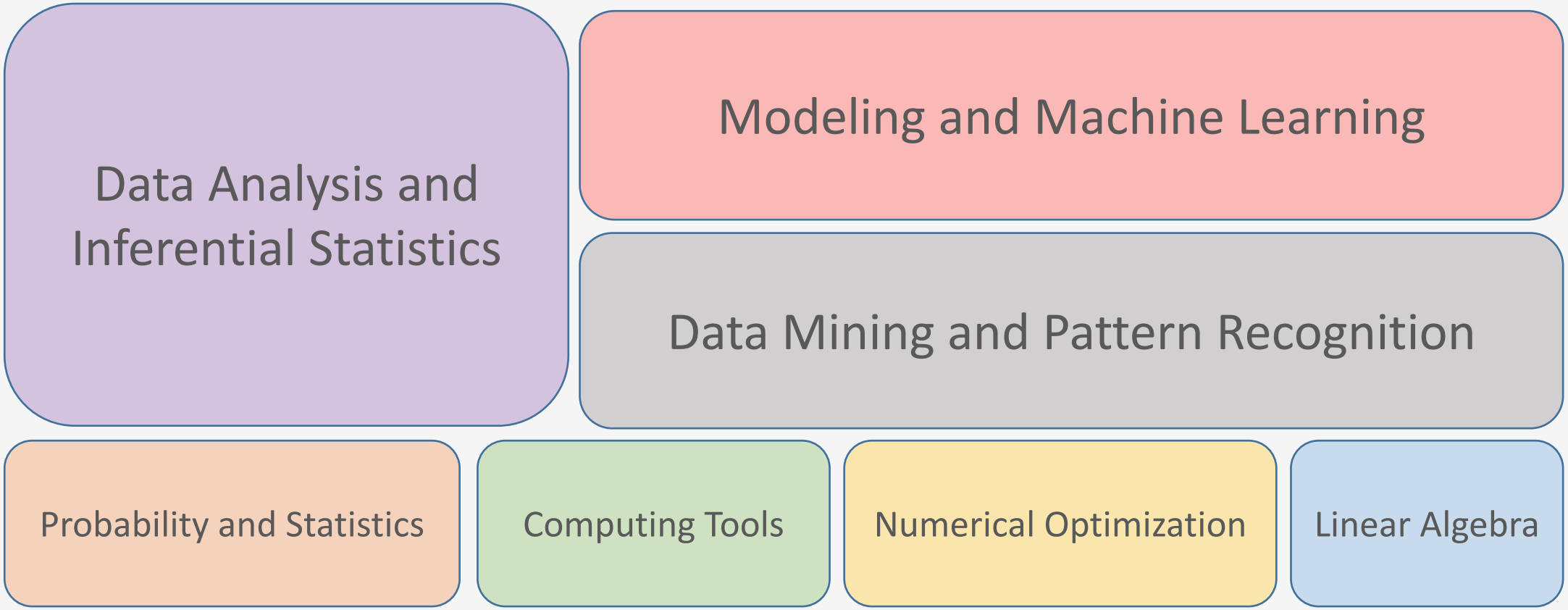
What is Data Science?

Data Analysis and Inferential Statistics

Modeling and Machine Learning

Data Mining and Pattern Recognition

What is Data Science?

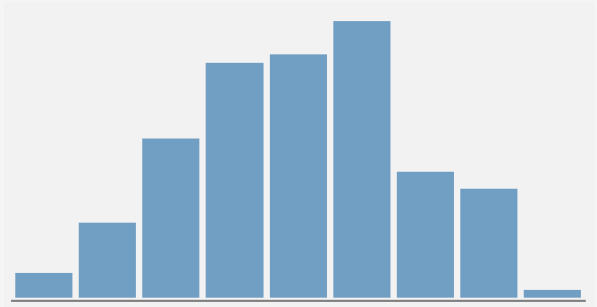


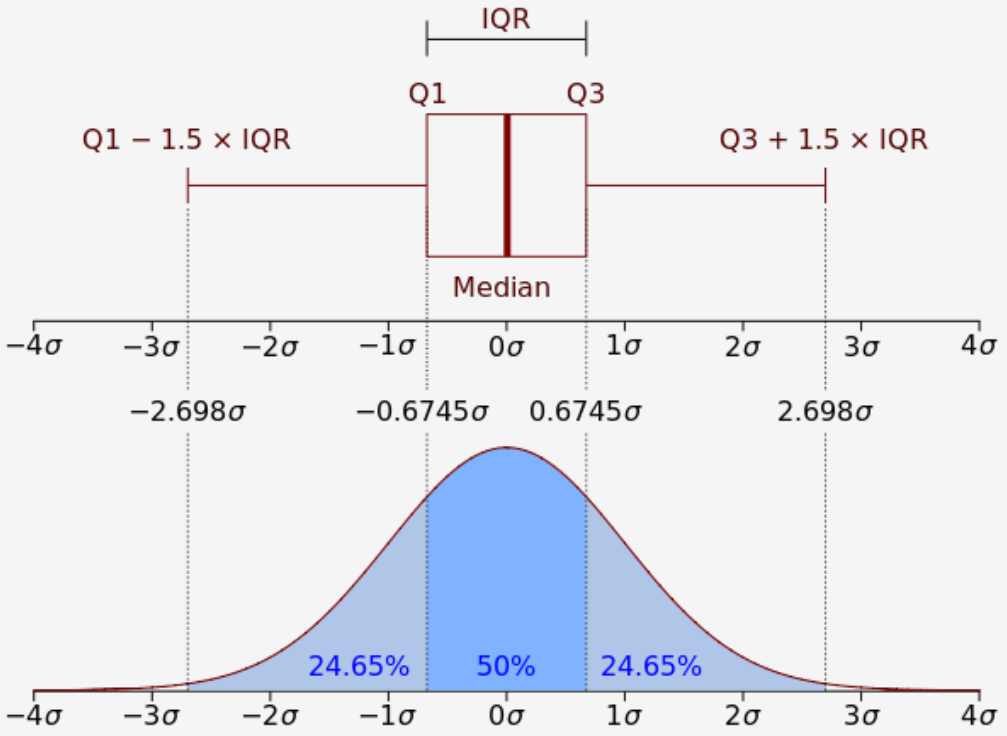
Course Topics

- Exploratory Data Analysis
- Cleaning, Munging, and Wrangling Data
- Probability Theory and Simulation
- Hypothesis Testing and Inferential Statistics
- Modeling and Classification

Exploratory Data Analysis

Dig into your data, compute simple statistics, draw pictures and look for **INSIGHT**





Data Cleaning and Wrangling

Some Data Scientists report they sped up to 80% of their time cleaning messy data sets

"This is not what I meant when I said 'we need better data cleansing!'"

Probability Theory and Simulation

Different probability distributions model different types of events

Probability Theory and Simulation

Different probability distributions model different types of events

The Binomial Distribution: The number of customers that will unsubscribe from your company's email-list as a function of how many advertisements you send them

Probability Theory and Simulation

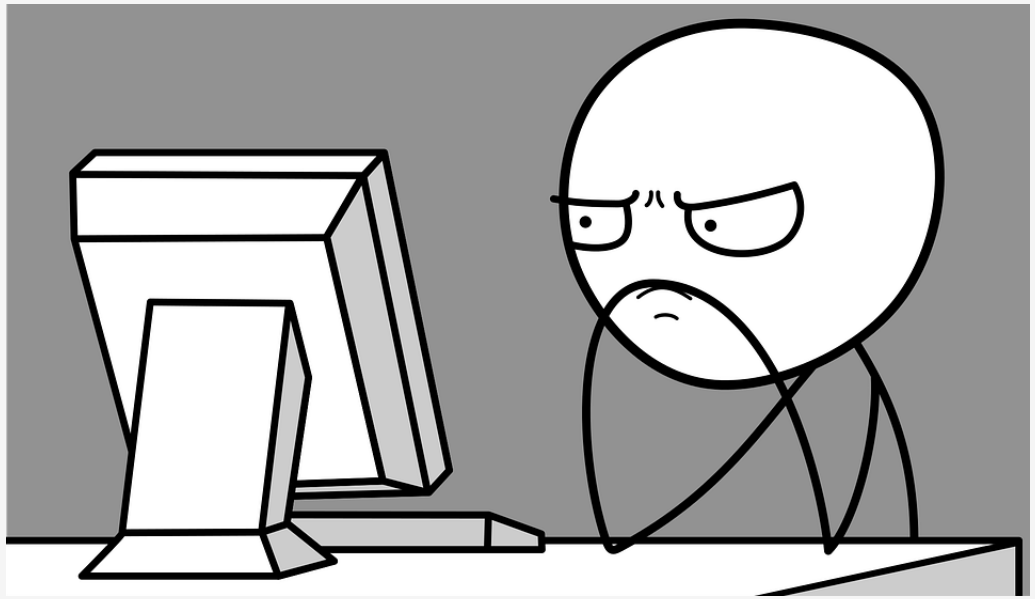
Different probability distributions model different types of events

The Poisson Distribution: The number of online customers that will visit your website over a particular time period, or the daily number of car crashes on a particular stretch of road

Probability Theory and Simulation

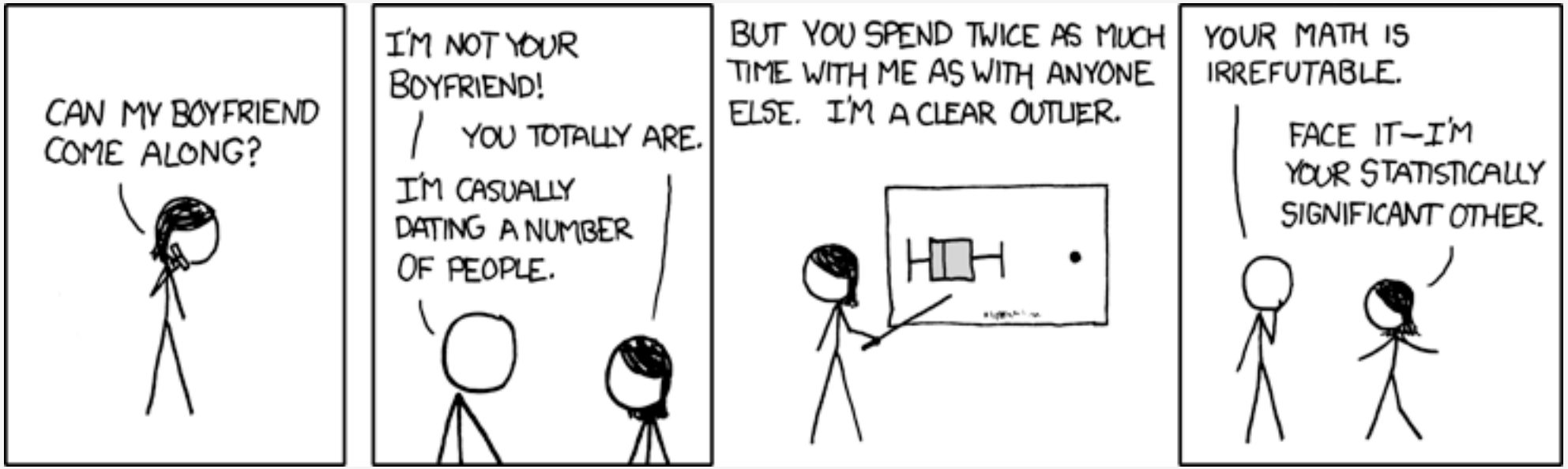
Different probability distributions model different types of events

The Exponential Distribution: The amount of time you can expect a compute node in a large-scale cluster to function before failure



Hypothesis Testing and Inferential Statistics

What is the data **really** telling us, and how **confident** should we be in our conclusions?

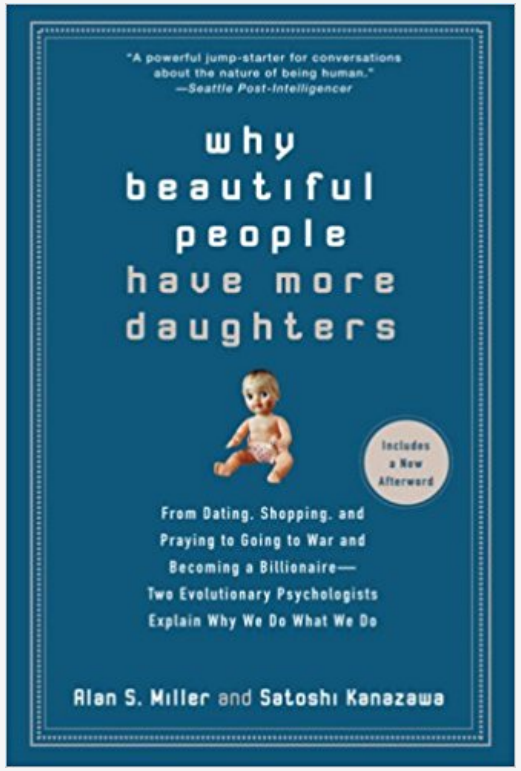


Hypothesis Testing and Inferential Statistics

What is the data **really** telling us, and how **confident** should we be in our conclusions?

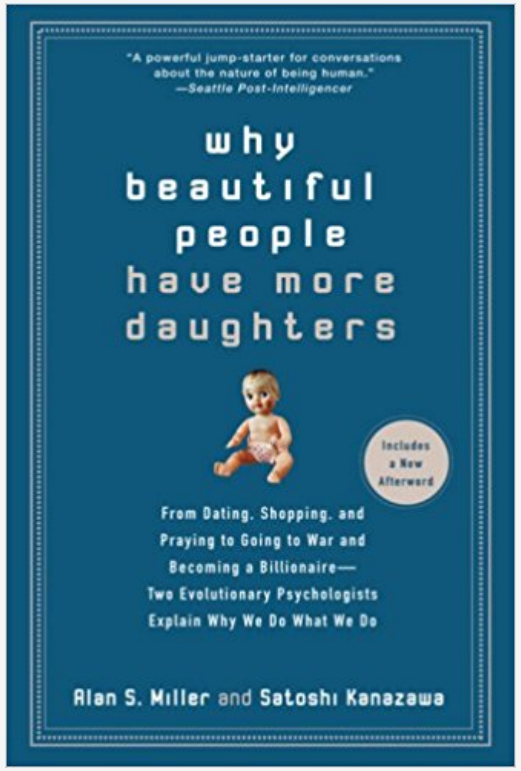
Hypothesis Testing and Inferential Statistics

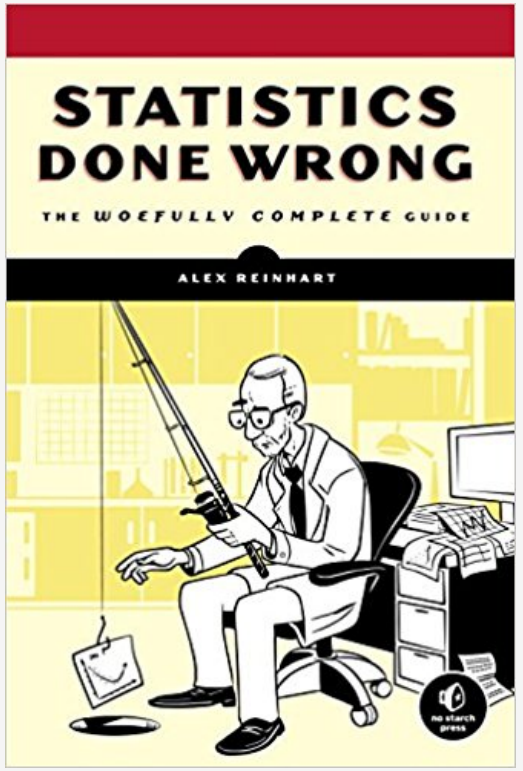
What is the data **really** telling us, and how **confident** should we be in our conclusions?



Hypothesis Testing and Inferential Statistics

What is the data **really** telling us, and how **confident** should we be in our conclusions?





Modeling and Classification

A gentle foray into Machine Learning.

- Linear Regression
- Multiple Linear Regression
- Logistic Regression

The Plan

Goal: Fluency in the theoretical and computational aspects of data analysis

At the end of this course you'll be able to

1. Clean, munge, and **wrangle** data in Python and perform Exploratory Data Analysis
2. Draw **insight** from data by computing and interpreting classic summary statistics
3. Know the ins-and-outs of probability and how to use it to solve **real-world problems**
4. Perform statistical tests to determine if your conclusions are **real** or due to chance
5. Construct and analyze simple models to make predictions and **inferences** about data
6. Tell **compelling stories** about data using modern visualization and presentation tools

Course Logistics

Keep track of course webpages (Piazza and GitHub)

- Piazza: https://piazza.com/colorado/fall2017/csci3022
  - Send me private messages on Piazza, rather than emails
  - Address message specifically to me if necessary

- GitHub: https://github.com/chrisketelsen/csci3022
  - In-class work will be posted here, as well as homework
  - Good idea to clone repo and do a pull everyday before class
  - Good Git tutorial if you're unfamiliar: http://rogerdudler.github.io/git-guide/

Course Logistics

Course Work:

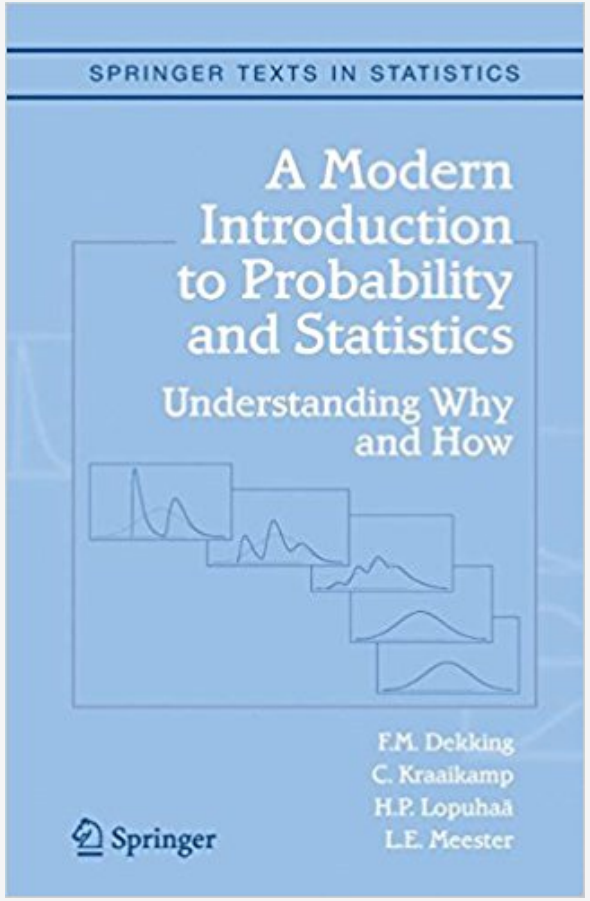
- Homework assignments every two weeks (35%)
  - Lowest homework score dropped
  - 3 total late days (1min – 23hr 59min late = 1 late day)
- Class Participation through tutorial problems and short Moodle Quizzes (5%)
- Midterm Exam (20%)
- Practicum (15%)
- Final Exam (25%)

Course Logistics

Collaboration Policy:

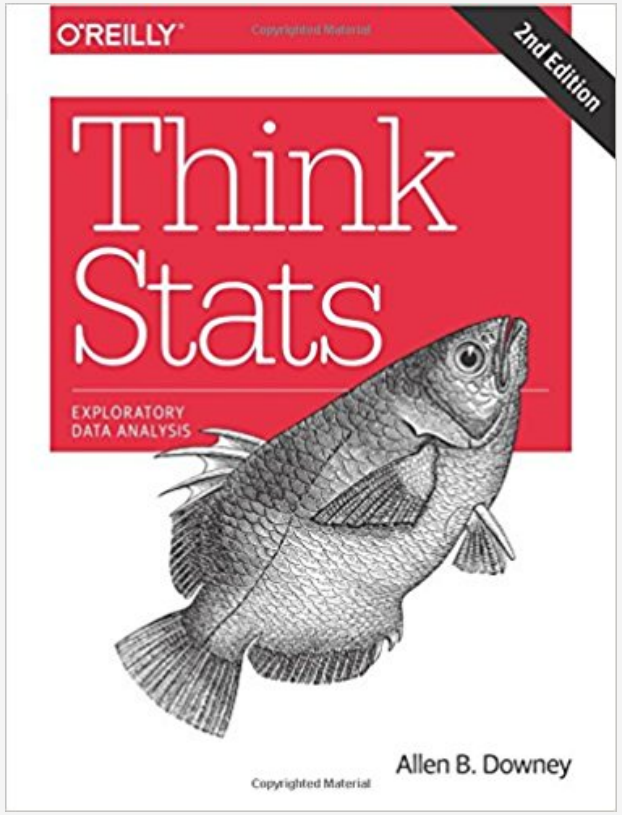
- Data Science is a collaborative field. Discuss problems with classmates and instructors.
- But do your own work. **Write solutions and CODE on your own**.
- Give **hints**, not solutions, on Piazza.
- Make repositories containing your homework **private** (GitHub, Azure)
- More info about collaboration on syllabus

Course Reading



- Good book with useful examples and exercises
- Doesn't force you to use R!
- Free PDF through CU library! (link on syllabus)
- Overly **mathy** sometimes
- Only responsible for what we cover in class
- Does things in slightly different order than us

Course Reading



- Supplemental Text on Data Analysis with Python
- Beware Python 2 vs Python 3 differences!
- Free PDF through publisher! (link on syllabus)
- Not mathy **enough** most of the time
- Won't really refer to it in class.
- Use for extra Python help.

Computing

- We will use Python 3 and in particular Numpy and Pandas
- Lot's of great data science libraries and decent plotting
- We'll exclusively work in Jupyter Notebooks
- Jupyter is ubiquitous DS collaboration and communication tool
- Easiest way to get **both** is **Anaconda Python 3.6**
- We **strongly recommend** you install local copy
- If not, you can use Microsoft Azure Notebooks
- Often work on problems in groups in class
- Bring a laptop or have a buddy with a laptop

About Me

- Starting 5th year as an instructor at CU (first 3 in APPM, last year in CS)
- Specialize in the **Mathy** courses (Discrete, Lin. Alg., Data Science, Machine Learning)
- Before CU, at Lawrence Livermore National Lab
- Before that, PhD in Applied Math at CU
- Before that, taught Philosophy at Washington State
- **Research**: Numerical Linear Algebra and **Stochastic Simulation**
- Please call me **Chris** or Dr. Ketelsen
- **Office Hours**: MW 2-3:30 in ECOT 731, or F 11-12pm by appointment

Let's Go to Work!

Let's start exploring some computing tools

Get out your laptop, or better yet, pair-up with someone else with a laptop