Pizza will be served!
November 1st, 4:30pm, Seelye 109

Statistical and Data Science Program’s
Mini-Graduate School Fair

D. Betsy McCoach, PhD
University of Connecticut, Measurement, Evaluation, and Assessment

Stephanie Eckman, PhD, Smith '94
University of Maryland, Joint program in Survey Methodology

Jean Wu, PhD
Brown University, Biostatistics

Leontine Alkema, PhD
UMass Amherst, Biostatistics
UVM CS Graduate Program Recruiting Event
Friday November 3rd at noon (location TBD)
RSVP to Jordan on Slack
Announcements 3/3

- Known grading bug in A3 – will be fixed by tomorrow
Outline

• Final Project Overview
  – Big picture (recap)
  – Some possible datasets
  – Deliverables timeline

• Topic Brainstorm Activity

• EDA
Final project (recap)

- **Goal**: apply the ML techniques we’ve learned to solve a real-world problem you care about

**Final deliverable**: a poster (or interactive visualization) that will be demonstrated during our end-of-semester reception and a 2-page write up of the methods you used

- Example problems
The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available in both JSON and SQL files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

- 4,700,000 reviews
- 156,000 businesses
- 200,000 pictures
- 12 metropolitan areas

1,000,000 tips by 1,100,000 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 156,000 businesses
Photo Classification

Maybe you've heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?

Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user's relationships define their usage patterns? Where are the trend setters eating before it becomes popular?
Nangbéto Hydropower Dam in Togo

- Embankment dam on the Mono River
- Located in the Plateaux Region (2nd largest population in Togo)
- Supplies power, and as a secondary goal fish and water reserves
Past Projects

• Predicting Student Debt Upon College Graduation
• NYC 311 - What are some of the factors that determine the number of days spent to solve non-emergency issues in New York City?
• “What Are Your Odds?”: An Interactive Web Application to Visualize Health Outcomes
• Predicting Airbnb Listing Prices in the NYC Area
• …and many more!
Final Project Deliverables

- Nov. 8th - FP1: Data Appendix
- Nov. 27th – FP2: Initial Model
- Dec. 4th – FP3: Revised Model
- Dec 13th – FP4: Poster (Final Project Reception)
- Dec. 22nd - FP5: Final Write-Up
Step 1: Write a quick description of a data set you think would be interesting to explore at the top of the page, and write your 99 number at the bottom
Activity: real world problems

Step 2: Pass your description clockwise to the next person
Activity: real world problems

**Step 3:** Read the description of the dataset, and underneath the description, write a question you think someone might want to answer using it.
Activity: real world problems

Step 4: Fold over the top of the paper (leaving just your question visible), and pass it on. Now repeat!
To get credit for this activity

• **By the end of class**: write a quick Slack post about a potential final project topic

• Please include:
  - A description of the domain / dataset
  - The problem(s) you're trying to solve / question(s) you're trying to answer
  - The audience (who would care about your results?)
  - Where to find the data (if you know)

• Not 100% sure? Try a couple and get some feedback

• See a topic you like? Reply to the post and form a team!
Exploratory data analysis (EDA)

Our usual goal:
model some phenomenon using a dataset

Goal of EDA:
develop an understanding of a dataset

http://r4ds.had.co.nz/
Quick EDA walkthrough: diamonds

ggplot2
Useful starting questions

- **Q1:** what’s in my data?
- **How to find out:** `?str()`
Useful starting questions

• **Q2**: what type of *variation* occurs within my variables?

• **How to find out**: visualize the distribution

```
# Categorical
ggplot(data = diamonds) + geom_bar(mapping = aes(x = cut))

# Continuous
ggplot(data = diamonds) + geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
```
#protip 1: adjust bin size for more detail

```r
smaller <- diamonds %>%
  filter(carat < 3)

ggplot(data = smaller, mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.01)
```
#protip 2: visualizing multiple distributions

- Use `geom_freqpoly()` instead of `geom_histogram()`

```r
ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
  geom_freqpoly(binwidth = 0.1)
```
Useful starting questions

- **Q3**: am I missing any data?
- **How to find out**: `summary()`
Useful starting questions

- **Q3**: am I missing any data?
- **How to find out**: `summary()`
Useful starting questions

- **Q4**: what type of **covariation** occurs within my variables?
- **How to find out**: visualize the (relative) distribution

```r
ggplot(data = diamonds, mapping = aes(x = price)) + geom_freqpoly(mapping = aes(colour = cut), binwidth = 500)
```
Useful starting questions

- **Q4**: what type of **covariation** occurs within my variables?
- **How to find out**: visualize the (relative) distribution

```r
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_boxplot()
```
Your turn!

- Find some data and perform EDA

**By the end of class**: write a quick Slack post about a potential final project topic, including:
  - A description of the domain / dataset
  - The problem(s) you're trying to solve / question(s) you're trying to answer
  - The audience (who would care about your results?)
  - Where to find the data (if you know)

- See a topic you like? Reply to the post and form a team!
Coming up

- A6 due tonight
- FP1 released this afternoon