Midterm Project

Midterm Project

- **Data Analysis Project**: Choose a dataset that is **relevant to your career goals** or your personal interest and propose an analysis that includes fitting at least a multilevel model.
- **Simulation based project**: Choose a phenomenon of interest that you can quantify, devise a simulation scheme to quantify the uncertainty in the process and validate your result.

Data Analysis Project

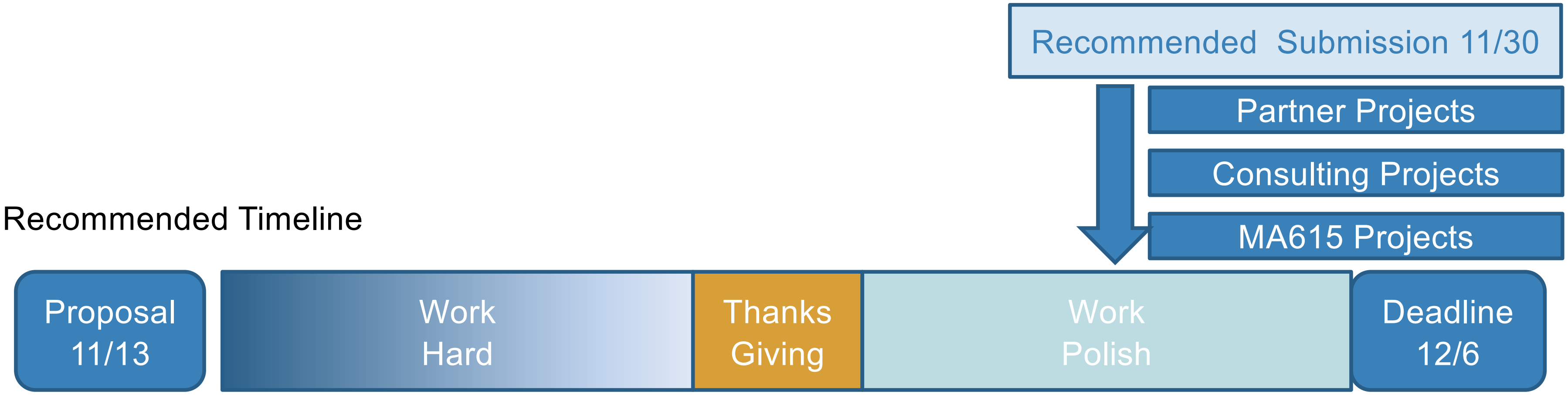
- Example data:
    - Tech:
        - Yelp Data challenge: https://www.yelp.com/dataset/challenge
        - AirBnB: http://tomslee.net/airbnb-data-collection-get-the-data
    - Consumer:
        - Customer Revenue prediction: https://www.kaggle.com/c/ga-customer-revenue-prediction
    - Medical:
        - Medicare, CDC: https://data.medicare.gov/data/. https://wonder.cdc.gov
    - Financial:
        - IMF: http://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42
        - Lending club: https://www.lendingclub.com/info/download-data.action
    - Music:
        - Million Songs: https://labrosa.ee.columbia.edu/millionsong/
    - etc. + extra bonus if you can combine different datasets.
- Criteria: A “LARGE” dataset with at least 10 groups that’s “Interesting”.

Simulation Project

- Criteria: you need to be able to quantify the outcome of interest and validate it using a real data that you can collect from various sources.
    - Analyze Boston: https://data.boston.gov/
    - MBTA API: https://www.mbta.com/developers/v3-api
    - Weather API: https://openweathermap.org/api

Timeline

- Nov 13th: Proposal Due
  - □ You are to chose from the two options below and propose a project that will help you showcase your skills to your future employee.
- Nov 30th: Recommended submission date
- Dec 6th : Final submission date (Mid



Recommended Submission 11/30

Partner Projects

Consulting Projects

MA615 Projects

Recommended Timeline

Proposal 11/13

Work Hard

Thanks Giving

Work Polish

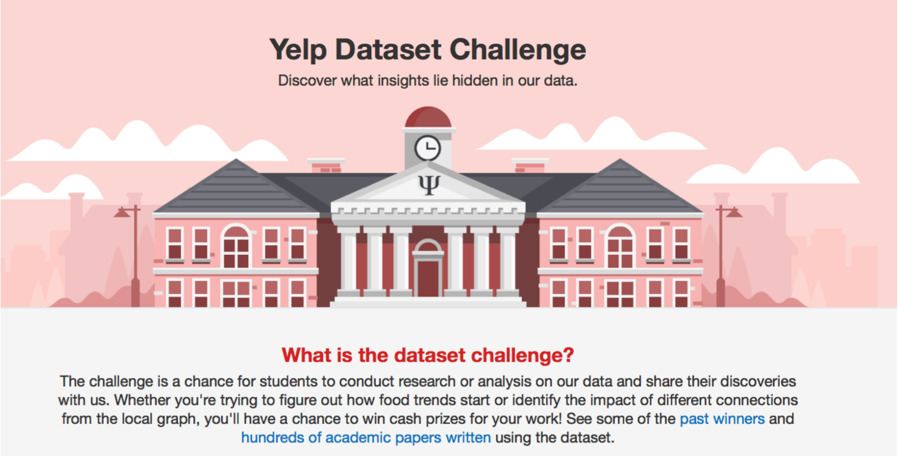
Deadline 12/6

Grading Rubric for the project

- (10) Overall Format: Can you confidently show it to a recruiter?
- (10) Novelty: New in some ways and interesting?
- (40) Accuracy: Model choice reasonable? Interpretations correct?
- (30) Validation: Detailed model checking to justify the result?
- (10) Discussions: Assessment of the result. Limitations? Future directions?
- (+10) Technical:
  - How did you deal with the big data challenge?
  - Did you integrate data from multiple sources?
- (∞) Passion

Yelp Data Challenge

- The 12th installment
  - □ https://www.yelp.com/dataset/challenge

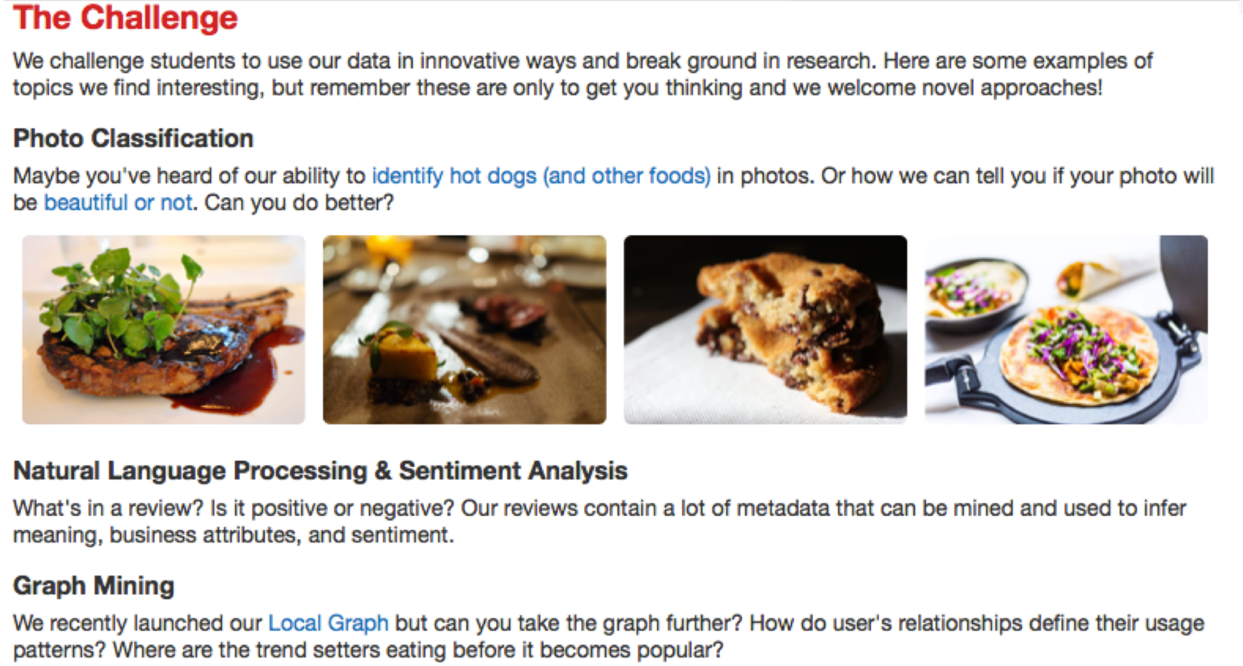


Yelp Dataset Challenge

Discover what insights lie hidden in our data.

What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the past winners and hundreds of academic papers written using the dataset.



The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

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?