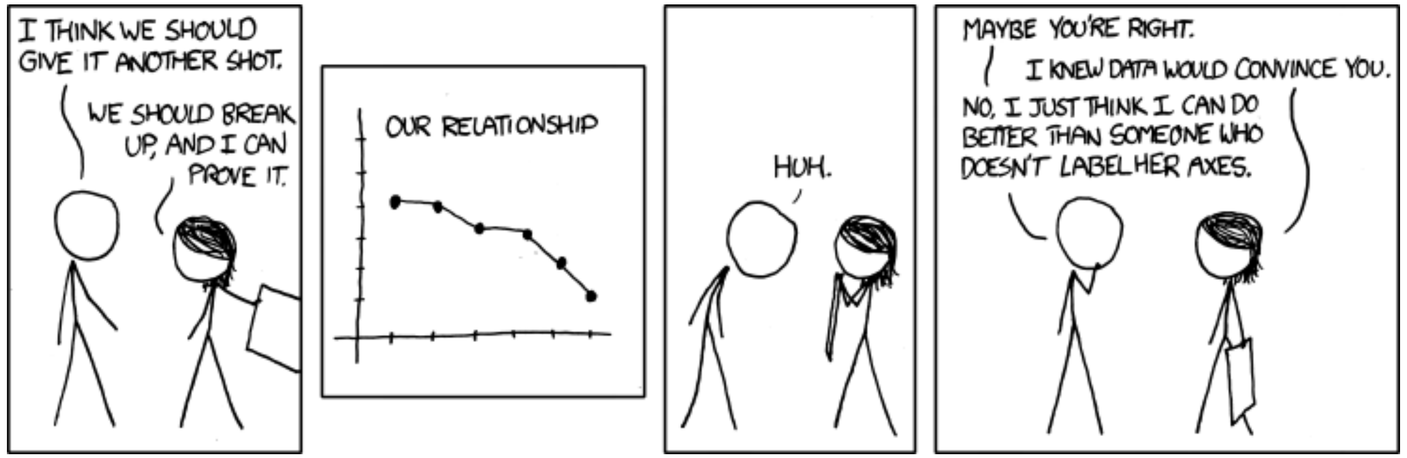
Introduction to Data Science
COMP 5360 / Math 4100

Alexander Lex
alex@sci.utah.edu

Braxton Osting
osting@math.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

Project

It’s time to start thinking about your project.

What you need:

- A team of 2-3
- An idea
- A dataset (that you actually can get!)
- http://datasciencecourse.net/2018/resources/

Project Phases

1. Announce your team and title (Tuesday, March 6)
2. Submit your project proposal (Tuesday, March 13)
3. Get feedback from staff (March 14-16; March 26-28)
4. Get/give peer feedback (in class on Thursday, March 15)
5. Submit project milestone (Sunday, April 1)
6. Get staff feedback (individual appointments, April 2-April 6)
7. Submit final project (Sunday, April 22)
8. Project Awards(in class on April 24)

Project Requirements

Scope as agreed upon with TAs

Should contain:

- Data acquisition (scraping, API)
- Data cleanup
- Exploratory Visualization
- Two different analysis methods (classification, regression, clustering, NLP)
  - Evaluate alternative approaches for each one (e.g., compare two or more classification methods)

You can skip one of these, but you have to make up in other areas

- E.g., if you work with clean & existing dataset, the analysis has to be more sophisticated

Be ambitious! Define your goals and categorize them:

- must have, nice to have, etc.

Don’ts

Don’t use a standard machine learning dataset (Kaggle, UCI ML Repository)

- These are pre-processed and only suitable for analysis, not for the whole DS process

Don’t pick a dataset where structured data is hard to extract

- E.g., text-only, relying on advanced NLP,
- extracting data from collection of PDFs,
- running your own survey (it’s hard to run a good survey)

Proposal Sections

Basic Info.

Background and Motivation

Project Objectives

Provide the primary questions you are trying to answer in your project.

Data

Data Processing

Exploratory Analysis

Analysis Methodology

Project Schedule

Submit as PDF or Jupyter notebook to Canvas

Milestone

Acquired, cleaned data

EDA

Sketches of your analysis methods

Submit zip file with Jupyter Notebook, data, other resources.

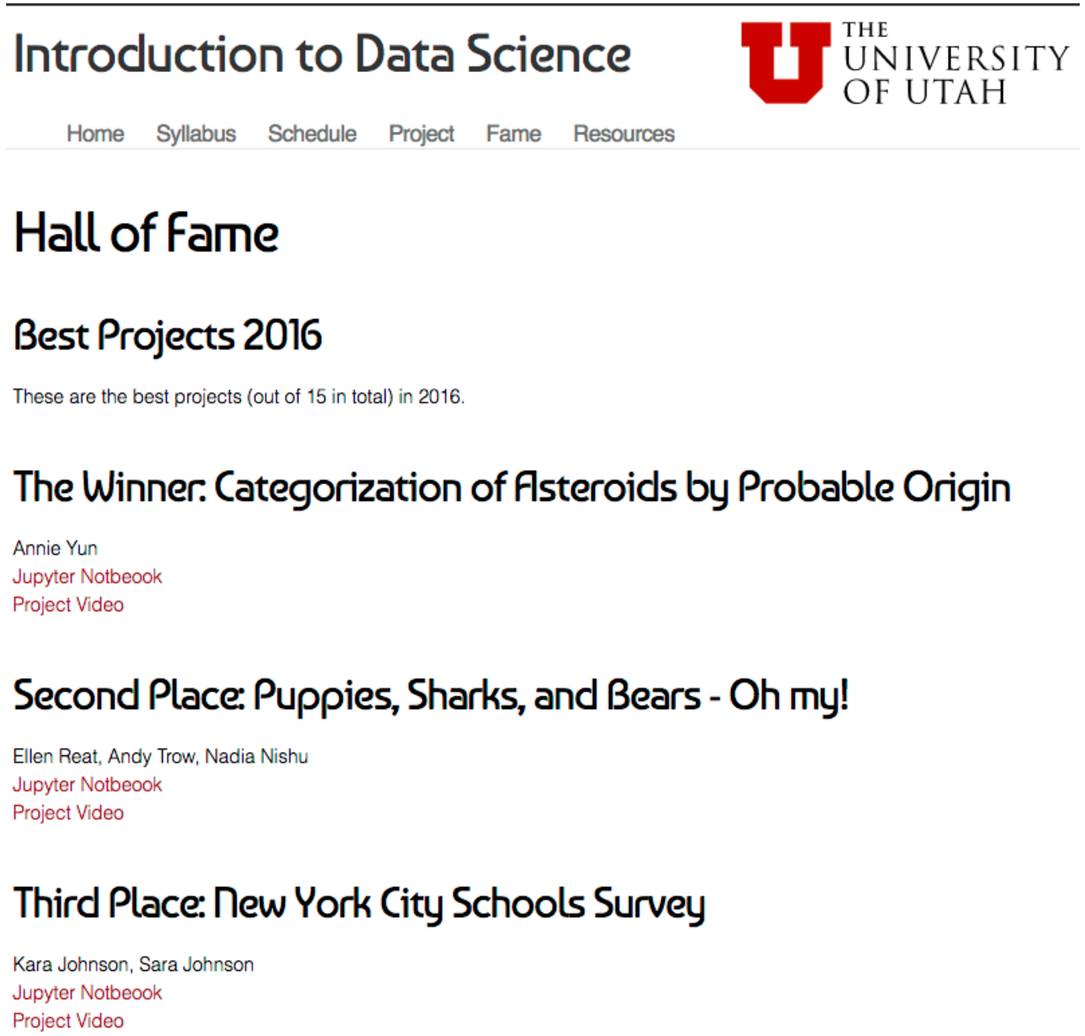
Final Submission

Whole story in a notebook

Include interpretation!

Three minute video that narrates project

Example Projects: Hall of Fame



Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Project Fame Resources

Hall of Fame

Best Projects 2016

These are the best projects (out of 15 in total) in 2016.

The Winner: Categorization of Asteroids by Probable Origin

Annie Yun
Jupyter Notbeook
Project Video

Second Place: Puppies, Sharks, and Bears - Oh my!

Ellen Reat, Andy Trow, Nadia Nishu
Jupyter Notbeook
Project Video

Third Place: New York City Schools Survey

Kara Johnson, Sara Johnson
Jupyter Notbeook
Project Video

http://datasciencecourse.net/2018/fame/