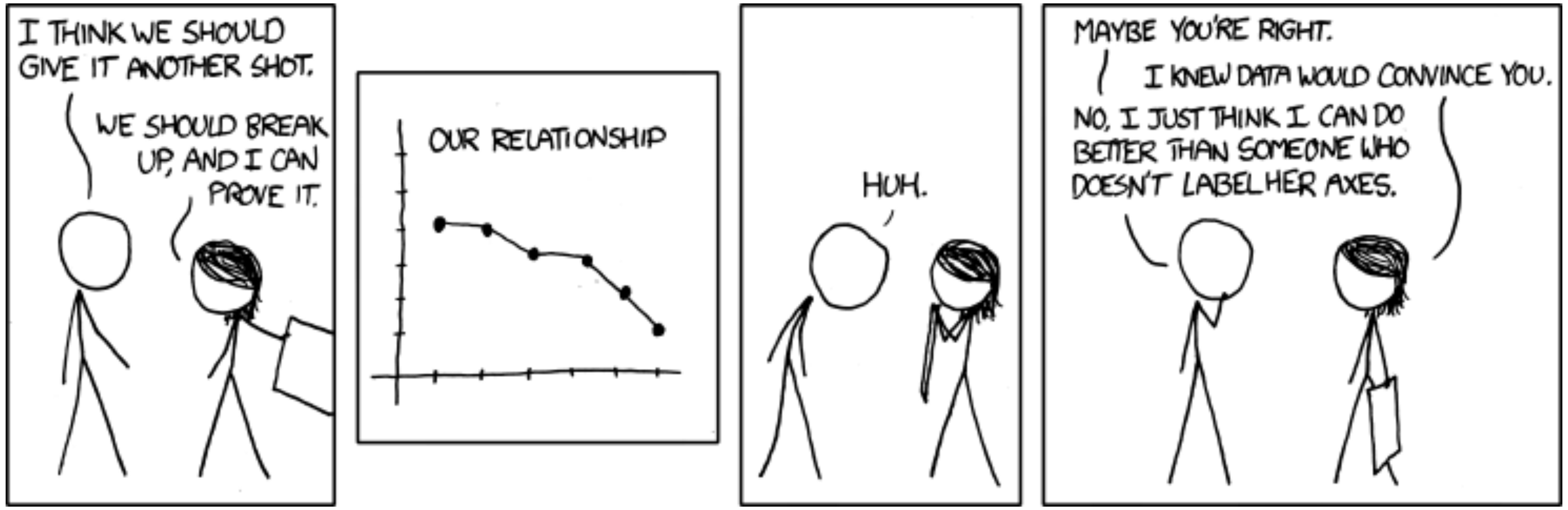
Introduction to Data Science
COMP 5360 / Math 4100

Kate Isaacs
u6044649@gcloud.utah.edu

Anna Little
little@math.utah.edu

THE UNIVERSITY OF UTAH



I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

[xkcd]

Project

It’s time to get serious about your project!

What you need:

- A team of 3
- An idea
- A dataset (that you actually can get!)
- http://datasciencecourse.net/2023/resources/

Project Phases

1. Announce your team and title (Friday, March 3)
2. Submit your project proposal (Friday, **March 17**)
3. Get/give peer feedback (**mandatory** in class on March 28)
4. Get written feedback from staff (by March 28)
5. Submit project milestone (Tuesday, **April 4**)
6. Get staff feedback (individual appointments, April 6-12)
7. Submit final project (Friday**, April 21**)
8. Project Awards (in class on April 25)

Note: Late policy does not apply; project items must be on time.

Project Requirements

Scope as agreed upon with Staff

Should contain:

- Data acquisition (scraping, API). Consider multiple datasets.
- Data cleanup and processing
- Exploratory Visualization
- Two different analysis methods (classification, regression, clustering, dimensionality reduction, NLP)
  - Evaluate alternative approaches for each one (e.g., compare two or more classification methods)
- Ethical considerations

You can skip one of these (except ethics), 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.

Ethical Considerations

Where in the process of your analysis were ethical decisions made? What were they?

Stakeholder analysis

- Who are the different “personas” relevant to your project?
- What are some incentives that may align or compete among these groups?

Is the data you collected biased or unbiased?

Are there certain groups that would be disproportionally affected by analysis or by the data?

Dont’s

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 Description & Acquisition

Ethical Considerations

Data Cleaning & Processing

Exploratory Analysis

Analysis Methodology

Project Schedule

Submit as PDF or Jupyter notebook to Canvas; **one per Group**

Milestone

Acquired, cleaned data

EDA

Sketches of your analysis methods

Submit zip file with Jupyter Notebook, data, other resources. **One per Group.**

Final Submission

Whole story in a notebook

Include interpretation!

Three minute video that narrates project

Group Work

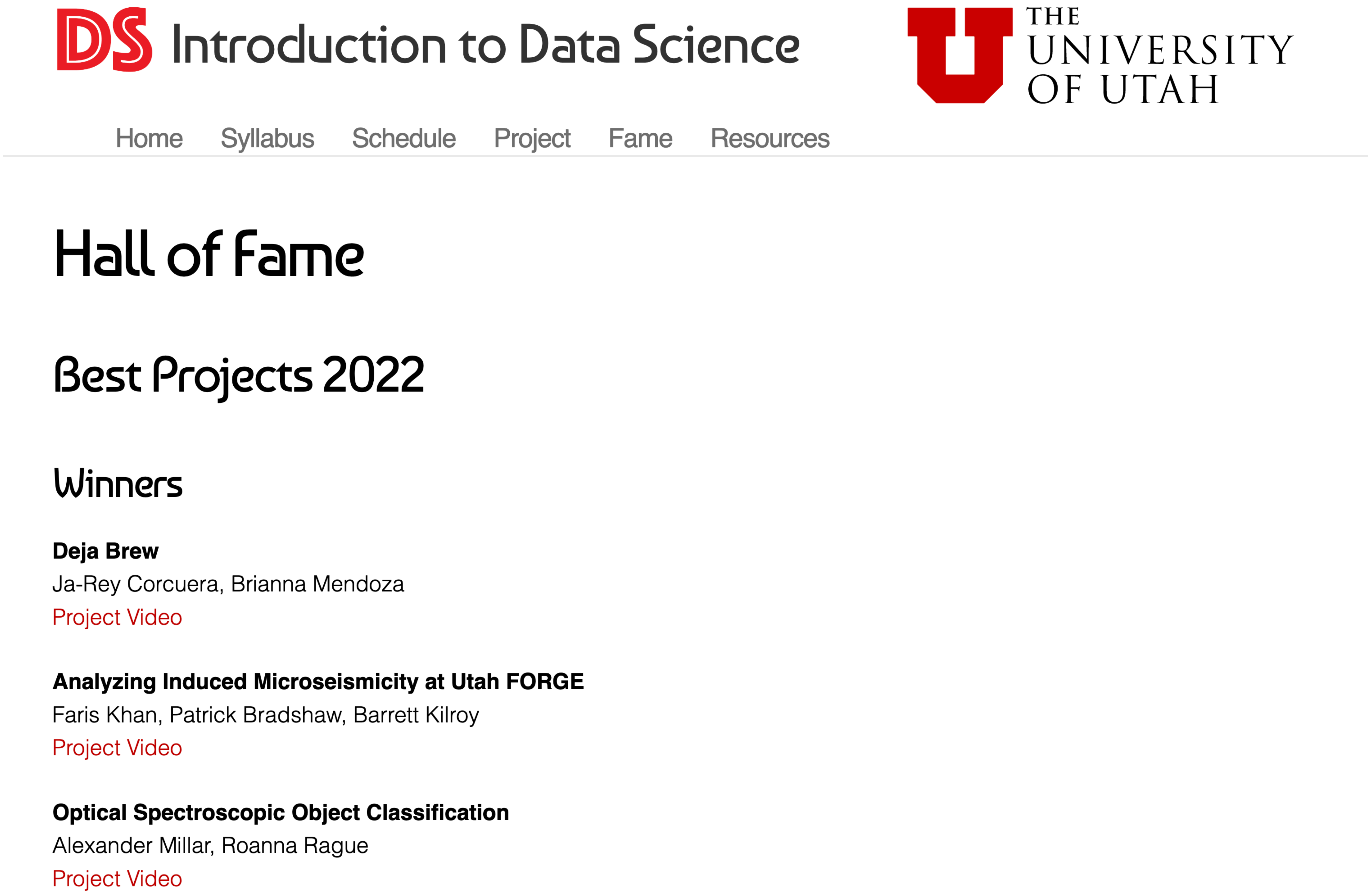
Be fair to your team-members

Stay within the schedule you agreed upon

Communicate immediately if there is a problem

Reach out to course staff if problem serious; do so before it’s too late.

Example Projects: Hall of Fame



DS Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Project Fame Resources

Hall of Fame

Best Projects 2022

Winners

Deja Brew
Ja-Rey Corcuera, Brianna Mendoza
Project Video

Analyzing Induced Microseismicity at Utah FORGE
Faris Khan, Patrick Bradshaw, Barrett Kilroy
Project Video

Optical Spectroscopic Object Classification
Alexander Millar, Roanna Rague
Project Video

http://datasciencecourse.net/2023/fame/