FINAL PROJECT KICKOFF!

Outline for today

- Logistics
- Development process
- Project Q&A & Advice

Logistics

- Schedule:
  - Hop-in tech dry run: December 6th
  - Presentation dry run: December 7th
  - Demo day: December 8th
  - Audience: employers, peers, family, mentors
  - Walk your friends/family through the presentation website beforehand!
- Length: 5 minutes (strict)
  - Private session rooms afterwards
- Make sure to fill out the google sheet posted by Kyla in the discord (finalize by week 11)
  - Descriptions will be sent out to potential employers attending demo day
  - Descriptions will appear in your session rooms on Hop-in
- **Speak to a mentor ASAP to get your project cleared**

Development process

Scoping your project

Essential

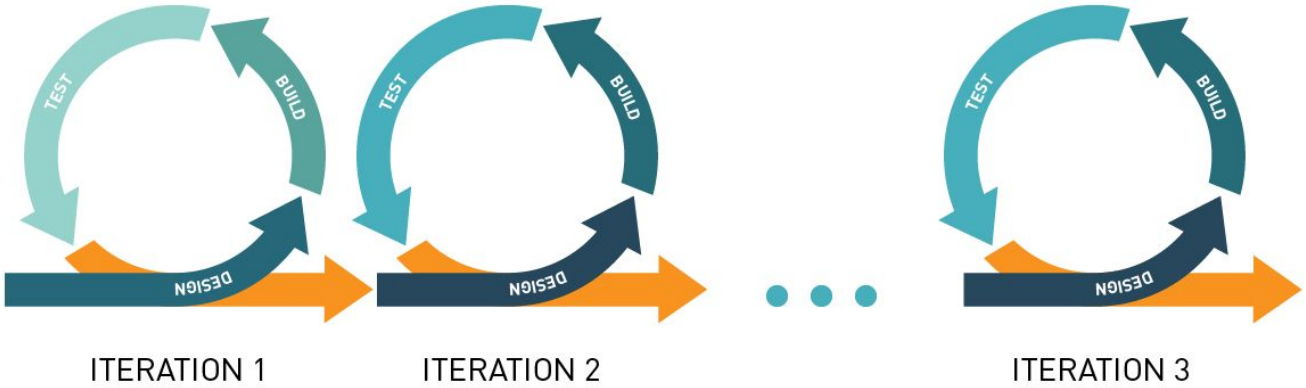
- What is the dataset, and is there one available?
- What are the inputs/outputs of the model (as *numbers*)?
- What do you want to do with your model (if anything)? Experiments? Applications?
- Is machine learning necessary?

Additional considerations

- Unless you want a big challenge, avoid having to learn something new (eg reinforcement learning, GANs)
- Is there prior work to help you out?
- Is it computationally expensive?

Iterative progress and difficulty

- Make a minimum viable product (MVP) early
- Dataset difficulty (eg simple or synthetic data before complex real world)
- Model complexity (eg linear autoregressive before LSTM)
- Task complexity (eg simplest solution before multifaceted)



Milestones

1. Research prior work
2. Acquire/preprocess/explore dataset
3. Design model
4. Train and evaluate model
5. **Version 1 of your capstone**
6. Perform experiments (if any)
7. Deploy model (or make figures)
8. Finish presentation

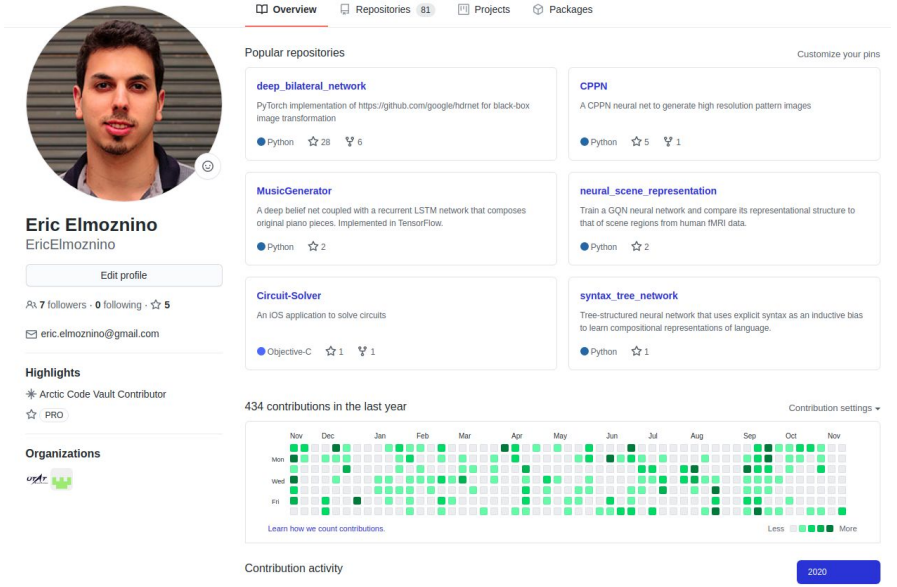
Each item must be:

a) Dated

b) Specific

Why use git?

- **Public**: Employers can see all the projects you've worked on
- **Versioned**: You will have a history and can roll back to old commits
- **Server deployment**: Just git pull to any new machine
- **Teamwork**: Everyone can work on their own copy and working versions to the master copy



Capstone Example on Github

1. https://github.com/GovindSuresh/reducing-bias-in-toxicity-classification
2. https://github.com/nomadtomas/sentweetment_analysizer
3. https://github.com/Greenford/billboard

Deployment and demo options

- Don’t *need* to deploy or demo anything; good figures can be just as good
  - If you have something like an image emotion classifier, a live demo can be cool
  - If you have something like a cancer classifier, just show us some performance metrics
- Consider whether your demo is too long or too risky (30 -45 seconds)
- Deployment options:
  - Flask API (might demo with Postman)
  - Flask app with UI (harder; need to write HTML)
  - AWS remote Flask app (if you need a faster machine and have money)
    - https://www.pythonanywhere.com/ (free)
    - https://www.heroku.com/ (free)
  - https://www.streamlit.io/ (good if you want UI, but learning a new library)
    - Heroku + Streamlit Tutorial

Code

- Define functions whenever possible (eg clean_data(df))
- Use pipelines for joint preprocessing, feature engineering, and model
- If using deep learning, generate training/validation curves as a function of epoch to see if your model is improving
- Save trained models and only retrain when needed
- Use .py files (.ipynb notebooks for EDA and rough work) so that you can split the project up into multiple files (eg data_cleaning.py, training.py, app.py)
- *Don’t show **any** code in your presentation*

Code quality: modularization

```
repo/
├── data
│   ├── raw_data.csv
│   └── preprocessed_data.csv
├── src
│   ├── modules
│   │   ├── data_preprocessing.py
│   │   └── modeling.py
│   │   └── figure_generation.py
│   ├── tests
│   │   ├── test_data_preprocessing.py
│   │   └── test_modeling.py
│   └── experiments.ipynb
├── output
│   ├── predictions.csv
│   ├── figures
│   │   ├── process_schematic.jpg
│   └── └── cluster_visualizations.jpg
└── README.md
```

```python
# data_preprocessing.py
...
def load_preprocessed_data():
    ...
    return X, y
...
```

```python
# experiments.ipynb
from modules.data_preprocessing import load_preprocessed_data
from modules.modeling import train_models
...
X, y = load_preprocessed_data()
best_model, cv_performance = train_models(X, y)
...
```

Compare to baselines

How good is your model? Contextualize it with a baseline

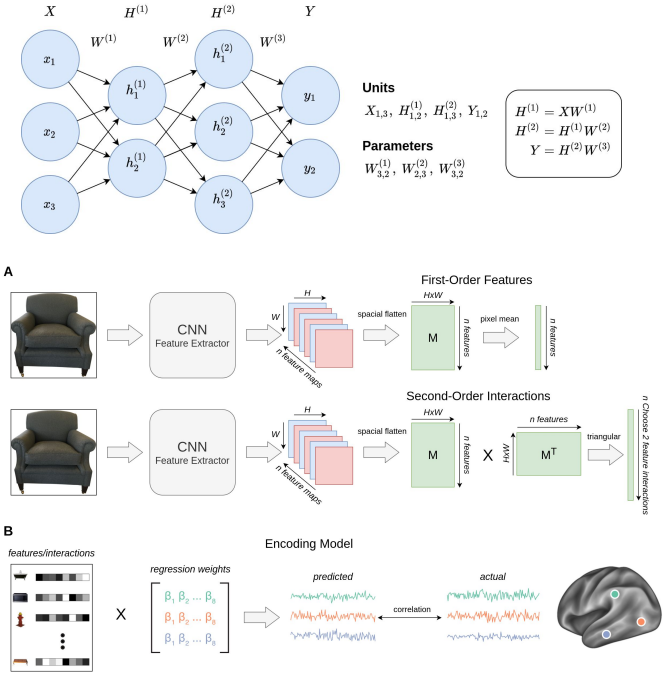
- For classification: always predicting the most frequent y value
  - Eg always predicting no-cancer if it is the most common class in the dataset
- For regression: always predicting the mean y value (implicit in r^2)
  - Eg always predicting the mean flight delay
- For forecasting: using a moving average model
  - Eg always predict the stock price tomorrow will be the average over the past week
- For any problem: hard-code a naive solution
  - Eg for hockey, always predict that the team with the better collective stats will win the game

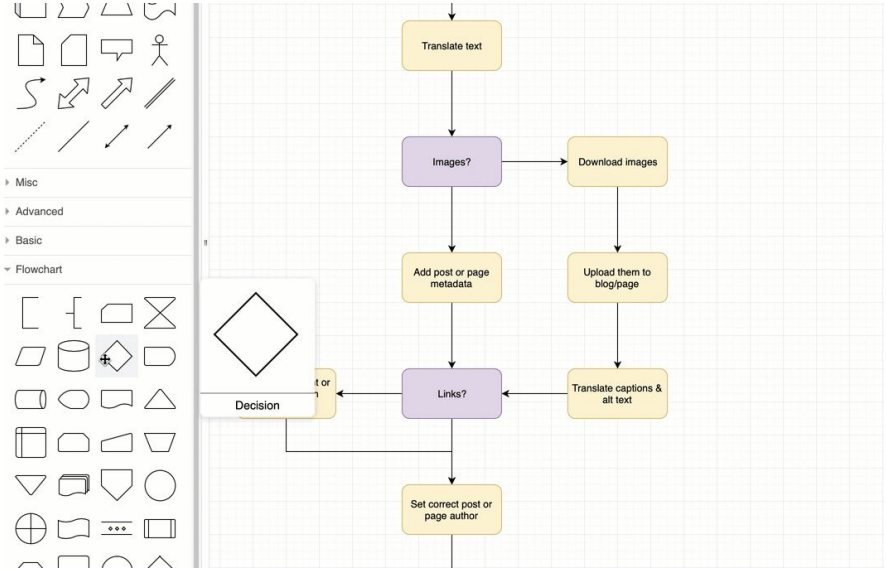
Presentation structure

- **Motivation**: What is the problem? Why is it important (either business, public good, or research perspective)?
- **Task**: Problem from a technical perspective. Description of the dataset, algorithm inputs/outputs, analyses done using model
- **Modeling**: *Important* aspects of your approach. How did you process the data or engineer features? What model did you use? Use schematics!
- **Results**: Visuals! Show metrics and experiments. Demo (if any)
- **Conclusions**: What worked? What didn't (and why)? How are we better off? Where could the project go next?

Figures: draw.io

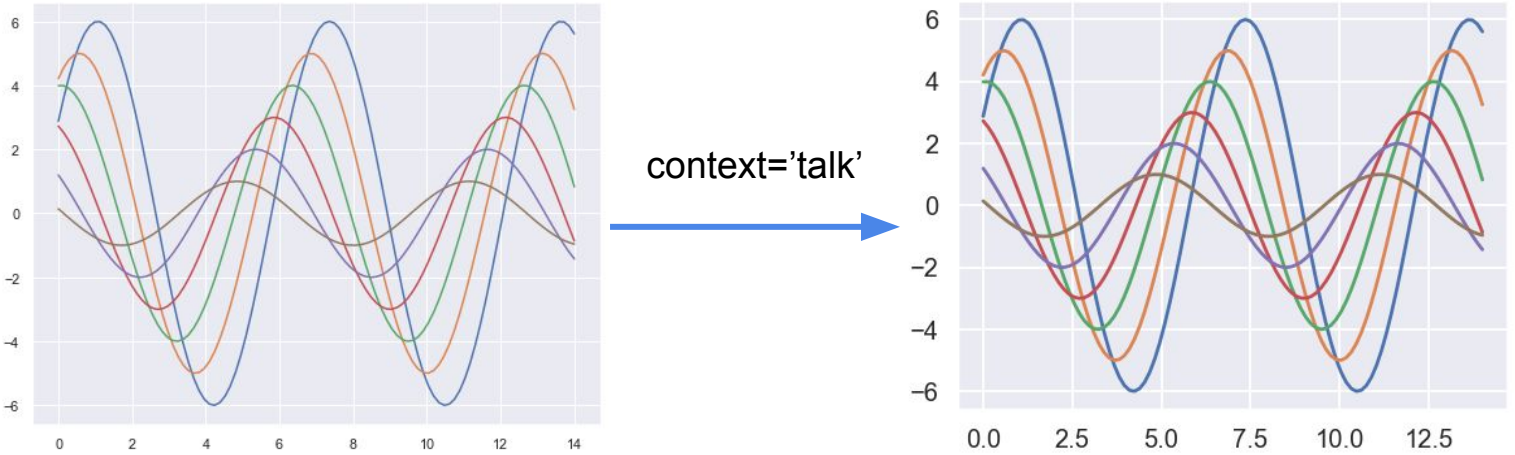
- Good for schematics, model diagrams, shapes, math typesetting, etc.





Figures: seaborn

```
# Right after importing seaborn (could also use 'whitegrid')
sns.set_theme(style='darkgrid', context='talk')
```



Activity (30 minutes)

- Get in pairs and spend 5 minutes each pitching your capstone idea to one another. (10 minutes)
  - What is your capstone?
  - What dataset will you be using?
  - Any potential challenges?
  - What is your MVP that you will be happy to achieve before demo day?
- **Everyone pitch their capstone idea in 1 minute (in main room)**
- It’s okay if no one has their finalized capstone.

Project Q&A & Advice