# ECSE-552 Final Project Software Architecture

Graham Smith
03/01/2022

# Source File Organization

- Python files
  - Use for core modules like defining model class or feature extraction
  - Bulk of code in this format as it’s easier to track diffs in GitHub
- Jupyter Notebook- Limited to setting up the execution environment
  - Download and install libraries
  - Connect to Google drive/external storage location
  - Initiate training data transfer and unzipping
  - Initiate training execution and display/save results
- GitHub stores:
  - Source code
  - Planning/architectural documents

# Network Issue w/ Data Stoarge

- Google Colab storage doesn’t persist → data needs to be uploaded each time
- Solution #1 – Store data on Gdrive and access Gdrive while training
  - Quota limits on per-user and per-file operation count and bandwidth quotas
  - Creates bottleneck where training loop could be stalled/waiting on data from network connection between Google Colab and Gdrive
- Better solution – Zip up data and transfer it all before training begins
  - More overhead upfront in terms of upload time
  - Data stored on drives associated with VM instance
  - Removes network connection from bottleneck
- Documented problem in Google Colab FAQ

# Usage of Google Drive

- Serve as a common reference point for everyone similar to GitHub
- Don’t make several copies across multiple Gdrives as maintenance becomes a pain
- Stores:
  - Training/validation/testing data (Pre-processed?)
  - Best model checkpoints from training?
  - Output graphs/logs/performance metrics