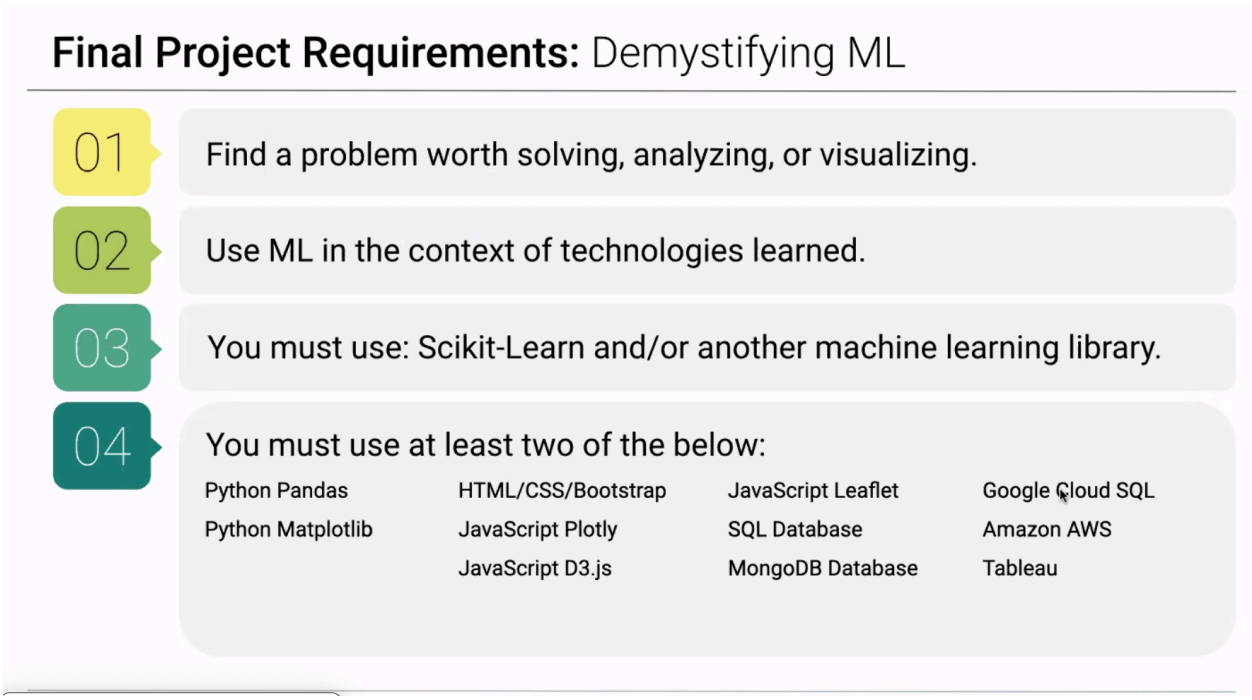
UCSD Data Science Bootcamp Final Project Proposal, 5/2/20

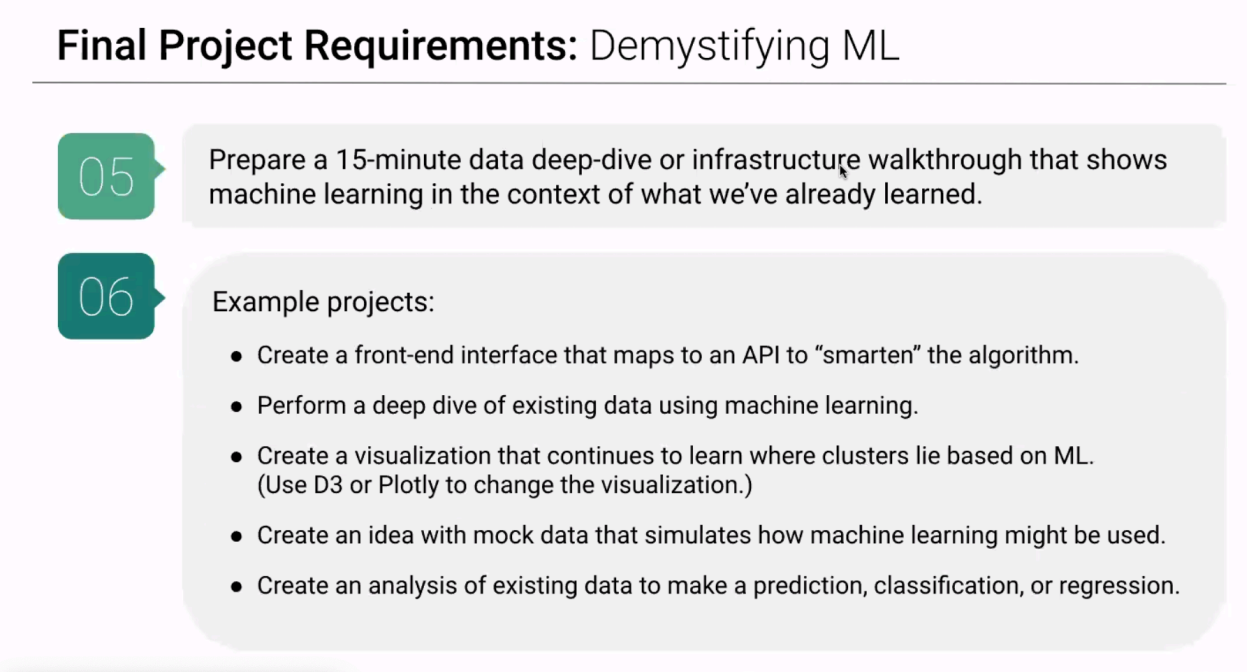
COVID-19 Machine Learning Analysis

Team Members:

- David Jaimes
- Grant Thompson
- Arundhati Chakraborty
- Alexis Perumal

Assignment





Project Objective

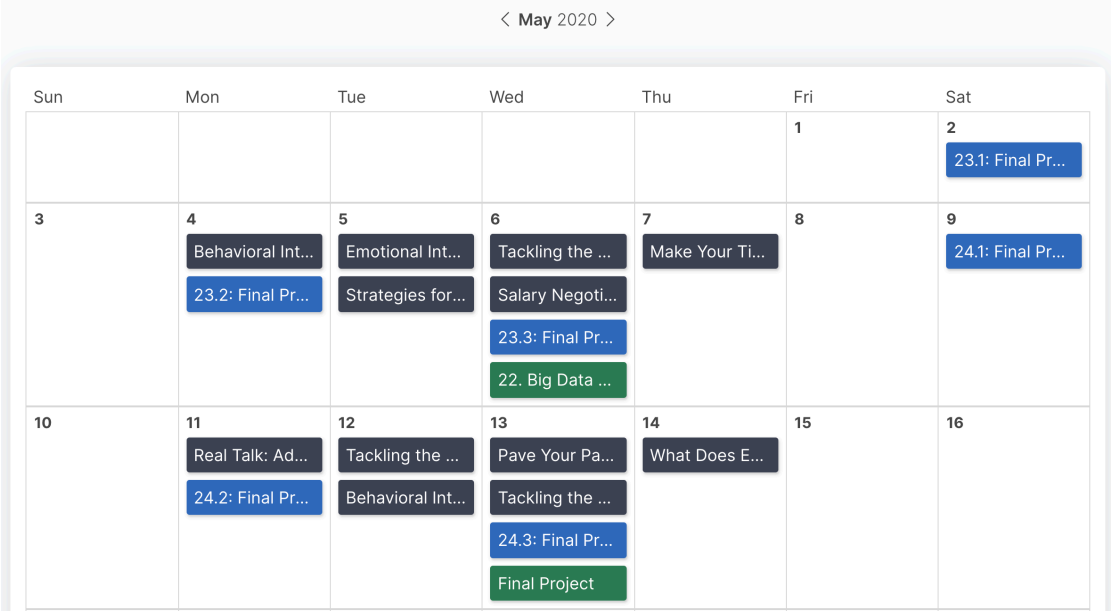
- Use ML and Data Science to better understand COVID-19 and if possible, help answer key questions about COVID such as:
    - Will a COVID-19 patient develop severe symptoms?
    - What markers correlate with a positive COVID-19 test?
    - Similarity of COVID-19 to other respiratory diseases at the DNA level?
    - Other COVID-19 questions listed in the Kaggle Uncover COVID-19 Challenge, sponsored by the Roche Data Science Coalition (RDSC), competition here:
        - Which populations are at risk of contracting COVID-19?
        - What is the incidence of infection with coronavirus among cancer patients?
        - Which patient populations pass away from COVID-19?
        - How is the implementation of existing strategies affecting the rates of COVID-19?
        - Plus many more questions listed.
    -

Project Description/Outline

- Find applicable datasets, e.g. Kaggle Einstein Hospital, Brazil, dataset, and explore them for suitability (completeness, quantity, etc.)
- Do data exploration including model building to determine what we can learn from them.
- Home in on the most promising data sets and where we think we can find meaningful results.
- Evaluate related work that has already been done and published in Kaggle.

- If we are able to achieve meaningful results on any of the RDSC/Kaggle questions, we may submit our response (batch 2 response submittal deadline 5/13/20, batch 3 6/3/20).
- Build a website illustrating our findings, showing our methods, and possibly providing interactivity for the user. (Example: Prior class team final project.)
- Possible Technologies:
    - ML Analysis: SciKit-Learn (Classifiers and/or regressors), Pandas, TensorFlow (neural networks), Jupyter Notebooks and/or Collab
    - Big Data: PySpark
    - Cloud: AWS RDS, S3, RDS Postgres, Flask/Fargate
    - Visualization: Tableau, Bootstrap, D3, Leaflet.js (Deep visualization is not the focus since we did that last project, but we'll add as time allows).

Rough Breakdown of Tasks



- Phase 1: Basic Data Exploration and ML Analysis, Sat., 5/2 – Wed.,5/6/20 – All
- Phase 2: Deeper ML Analysis and Web Stack – Wed., 5/6 - Sat., 5/9
    - ML
        - Select focus areas, specific team assignments, for deep analysis.
        - Conduct Analysis
        - Get Findings
    - Web Stack and Cloud Hosting
        - Build out a simple web stack (front end and backend)
        - Prototype out interactive web form
- Phase 3: Finalize Visualizations
    - Finalize Website, interactive form (blood results, other phenotypic data).
    - PowerPoint, animation
    - Demo

Useful Links

- Related Analysis
- Kaggle COVID-19 Data Source
- Kaggle Einstein Hospital, Brazil, dataset
- Team Project GitHub Repo
- Prior class example project
- Medium article on chest CT scans
- COVID Genomic Data
- Other Respiratory Genomic Data
- GISAID COVID-19 Database
- NCBI COVID-19
- SIR Infectious Disease Simulation
- Apple COVID-19 Page
- SciKit-Learn Documentation Page
- AWS Tutorial to build a serverless webapp in AWS with Fargate
- Tutorial on front end UI on top of Tensorflow