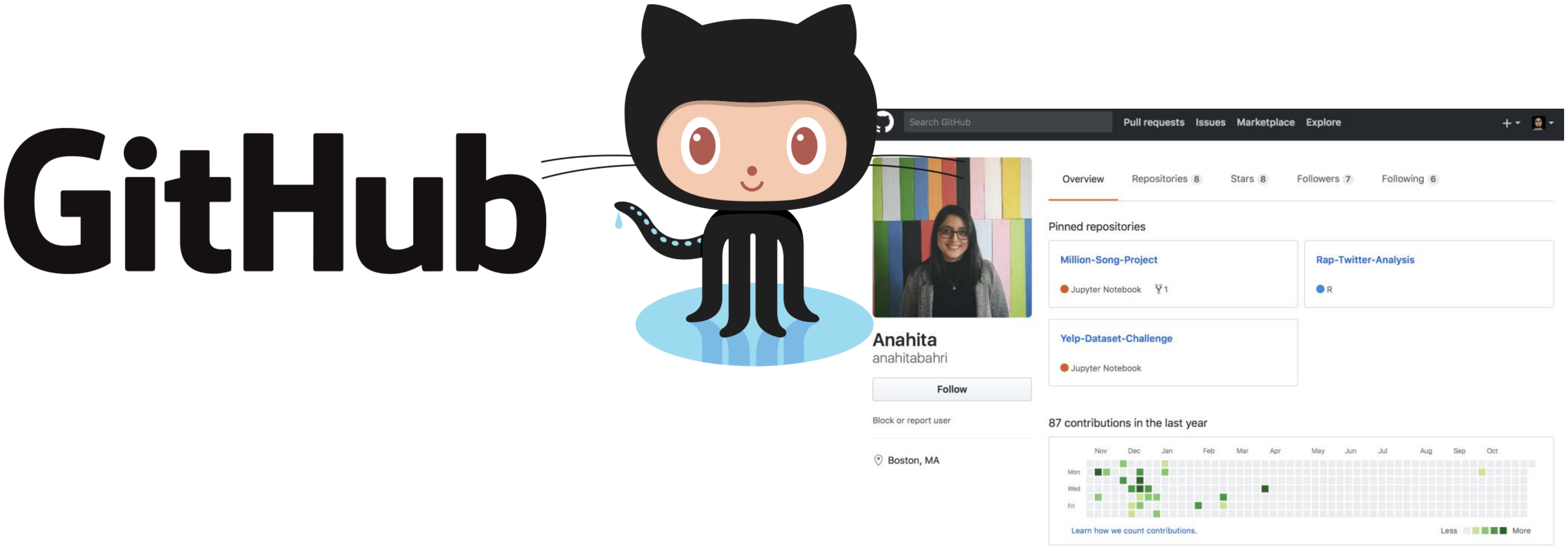
Midterm Project

Midterm Project

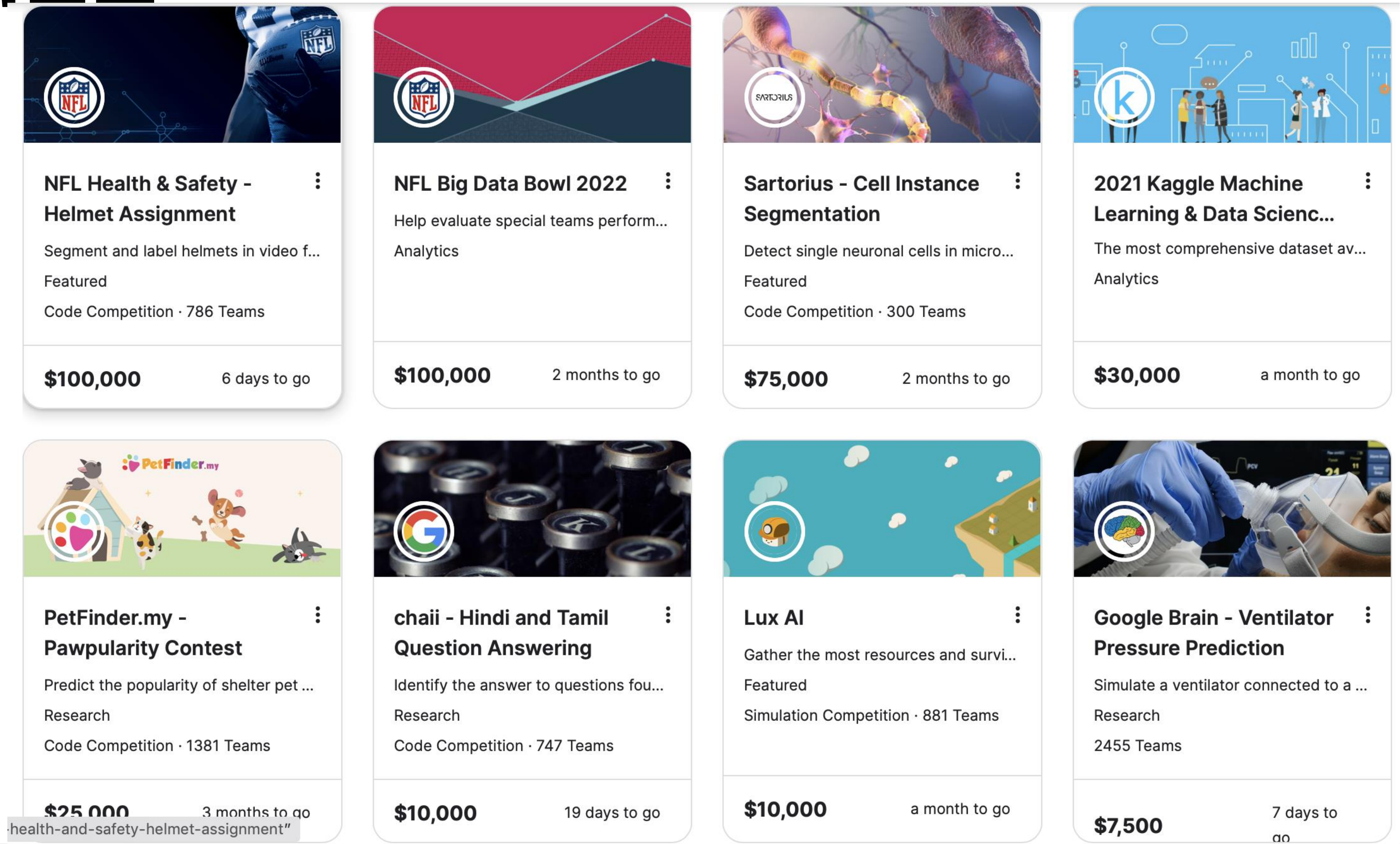
- Data Analysis Project: Choose a (large) dataset (or more) that is relevant for your career goals and display your capabilities for analyzing that includes fitting at least a multilevel model or doing causal inference.



Project Ideas

- Criteria:
  - A “LARGE” dataset with at least 10 groups that’s “Interesting” to compare using multilevel model.
  - Or a “LARGE” dataset that you can use causal inference technique to infer something new.
- Example data:
  - Tech:
    - Yelp Data challenge: https://www.yelp.com/dataset
    - AirBnB: http://insideairbnb.com/get-the-data.html
  - Consumer:
    - Customer Revenue prediction: https://www.kaggle.com/c/ga-customer-revenue-prediction
  - Medical:
    - Medicare, CDC: https://data.medicare.gov/data/. https://wonder.cdc.gov
  - Financial:
    - IMF: http://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42
    - Lending club: https://www.lendingclub.com/info/download-data.action
  - Music:
    - Million Songs: https://labrosa.ee.columbia.edu/millionsong/
  - etc. + extra bonus if you can combine different datasets.

KAGGLE



NFL Health & Safety - Helmet Assignment

Segment and label helmets in video f...

Featured

Code Competition · 786 Teams

$100,000 6 days to go

NFL Big Data Bowl 2022

Help evaluate special teams perform...

Analytics

$100,000 2 months to go

Sartorius - Cell Instance Segmentation

Detect single neuronal cells in micro...

Featured

Code Competition · 300 Teams

$75,000 2 months to go

2021 Kaggle Machine Learning & Data Scienc...

The most comprehensive dataset av...

Analytics

$30,000 a month to go

PetFinder.my - Pawpularity Contest

Predict the popularity of shelter pet ...

Research

Code Competition · 1381 Teams

$25,000 3 months to go

chaii - Hindi and Tamil Question Answering

Identify the answer to questions fou...

Research

Code Competition · 747 Teams

$10,000 19 days to go

Lux AI

Gather the most resources and survi...

Featured

Simulation Competition · 881 Teams

$10,000 a month to go

Google Brain - Ventilator Pressure Prediction

Simulate a ventilator connected to a ...

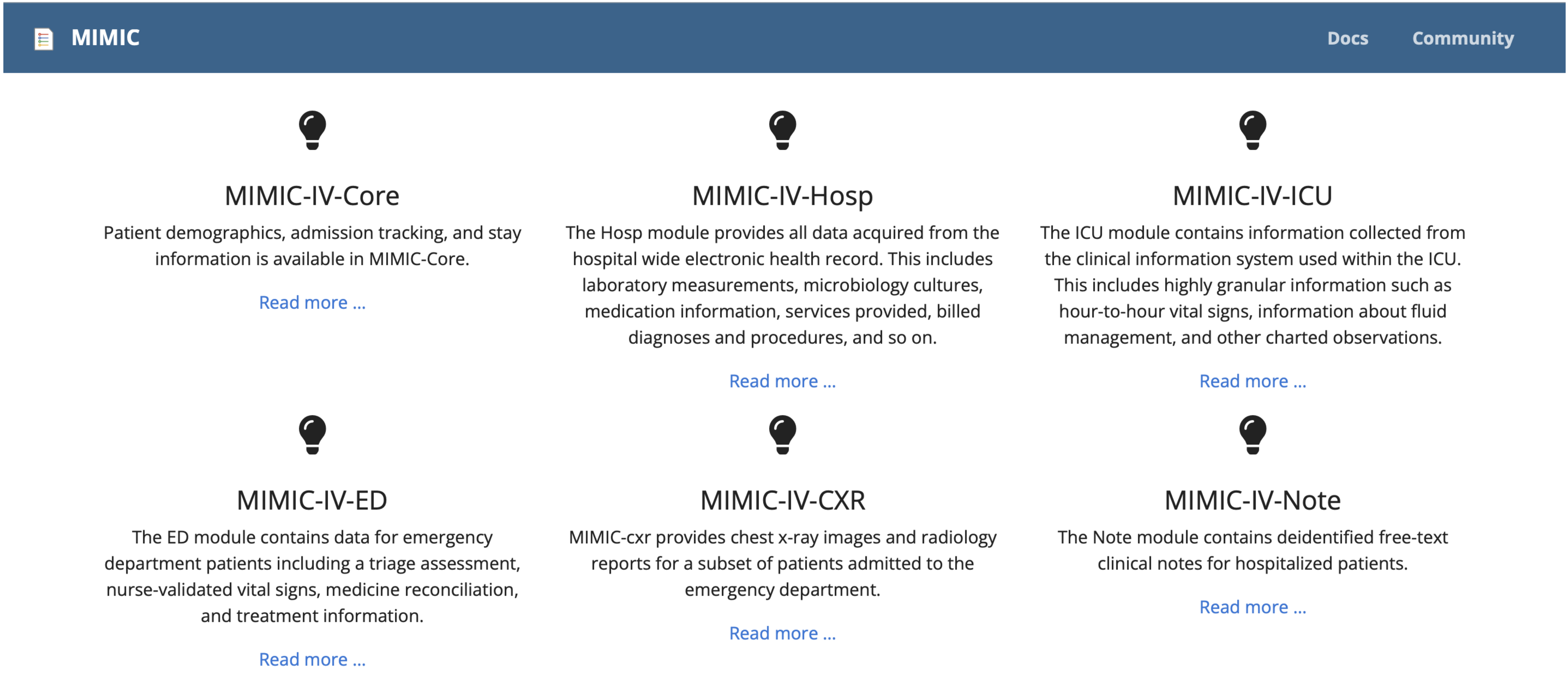
Research

2455 Teams

$7,500 7 days to go

-health-and-safety-helmet-assignment"

MIMIC



MIMIC-IV-Core

Patient demographics, admission tracking, and stay information is available in MIMIC-Core.

Read more ...

MIMIC-IV-Hosp

The Hosp module provides all data acquired from the hospital wide electronic health record. This includes laboratory measurements, microbiology cultures, medication information, services provided, billed diagnoses and procedures, and so on.

Read more ...

MIMIC-IV-ICU

The ICU module contains information collected from the clinical information system used within the ICU. This includes highly granular information such as hour-to-hour vital signs, information about fluid management, and other charted observations.

Read more ...

MIMIC-IV-ED

The ED module contains data for emergency department patients including a triage assessment, nurse-validated vital signs, medicine reconciliation, and treatment information.

Read more ...

MIMIC-IV-CXR

MIMIC-cxr provides chest x-ray images and radiology reports for a subset of patients admitted to the emergency department.

Read more ...

MIMIC-IV-Note

The Note module contains deidentified free-text clinical notes for hospitalized patients.

Read more ...

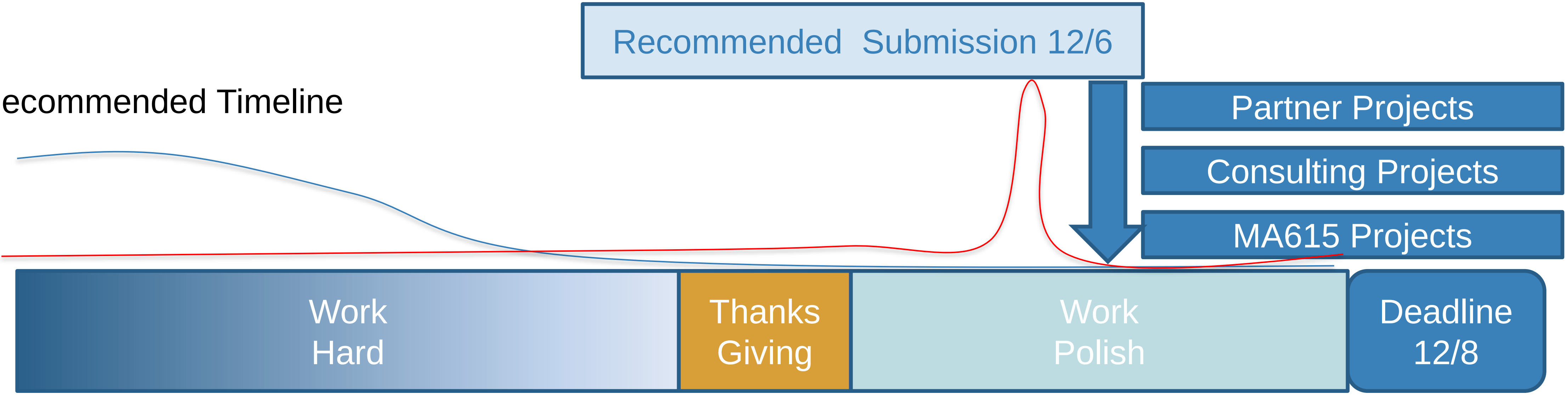
Timeline

- Dec 6nd : **Recommended** submission date
- Dec 8th : Final submission date

MON	TUE	WED	THR	FRI	SAT	SUN
10/31	*1*	*2*	*3*	*4*	*5*	*6*
7	*8*	*9*	*10*	*11*	*12*	*13*
14	*15*	*16*	*17*	*18*	*19*	*20*
21	*22*	*23*	*24*	*25*	*26*	*27*
28	*29*	*30*	*12/1*	*2*	*3*	*4*
5	*6*	*7*	*8*	*9*	*10*	*11*

Thanks Giving

Recommended Timeline



Recommended Submission 12/6

Partner Projects

Consulting Projects

MA615 Projects

Work Hard

Thanks Giving

Work Polish

Deadline 12/8

Grading Rubric for the project

- (10) Overall Format: Can you confidently show it to a recruiter?
  - □ Does it look professional? Is it written in proper language?
- (10) Novelty: New in some ways and interesting?
- (20) Accuracy: Model choice reasonable? Interpretations correct?
- (20) Validation: Detailed model checking to justify the result?
- (20) Discussions: Assessment of the result. Limitations? Future directions?
- (20) Technical:
  - □ How did you deal with the big data challenge?
  - □ Did you integrate data from multiple sources?

Final Submission Format

- Submit a nicely formatted PDF and a link to your GitHub repo to the blackboard.
- The main report should be at most 10 pages, including figures. (Appendix will not be counted)
    - Abstract (a paragraph): high-level summary of your work.
    - Introduction: Background and other information necessary to understand your work.
    - Method: What you did in some detail.
    - Result: What you found
    - Discussion: What you think this means and what are the next steps.
- Appendix: (not part of the page limit)
    - All the supporting results and details that may get in the way of your argument goes here.
    - Model checking details, figures that are not crucial,
- Supplement
    - Code, etc
- Please do not show raw R output in your report.

Learning from and working with others

- You may discuss your work with other people.
- But you are to each submit your own original work.
- We will check the originality of your work. You will get no point if
  - □ You have significant overlap in the content with someone from the class.
  - □ You have copied more than 20 percent of your text from some online source.
- Please note that if you are doing Kaggle competition, copying other people's work without attribution is not allowed.
- The screening will be done automatically. If your report is flagged as concerning and if you cannot prove your innocence, I will ask you to redo the work using other data from scratch.
- If you fail to submit this report, you will get an incomplete for the class.
- We have had challengers that copied from online work, TA, classmate, etc. All of them were caught. Please do not make me have to report you to the GRS.

- Public Datasets
  - https://github.com/awesomedata/awesome-public-datasets
  - https://github.com/apiad/datasets-list
  - https://github.com/datasets/openml-datasets/tree/master/data the same data as this list https://www.openml.org/search?type=data
  - https://archive.ics.uci.edu/ml/datasets.html
  - https://www.data.gov/
  - https://www.kaggle.com/datasets
  - http://datamob.org/
  - https://sites.google.com/a/drwren.com/wmd/details
  - https://data.cityofnewyork.us/data
  - http://snap.stanford.edu/data/index.html
  - http://aws.amazon.com/datasets/

- Event Data

  • GeoLife GPS Trajectories http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

  • Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors (MIT Project 2004) http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/

  • https://sites.google.com/a/drwren.com/wmd/home

  Frequent Itemset Mining Dataset Repository

  • http://fimi.ua.ac.be/data/

  Airline On-time Performance

  • http://www.eecs.wsu.edu/~yyao/StreamingGraphs.html

  • http://openflights.org/data.html

  Collection and Streaming of Graph Datasets

  • http://www.eecs.w su.edu/~yyao/StreamingGraphs.html

  Data Streams

  • http://www.quora.com/Where-can-I-find-public-or-free-real-time-or-streaming-data-sources

  • 3 hourly weather forecast and observational data - UK locations http://data.gov.uk/dataset/metoffice_uklocs3hr_fc

- New York Taxi Datasets
  - https://data.ny.gov/
  - TLC Trip Record Data

  http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

  http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

  - Another set published by chris whong http://chriswhong.com/open-data/foil_nyc_taxi/

  this data set is used for the DEBS 2015 challenge http://www.debs2015.org/call-grand-challenge.html

  - Another Description of this dataset http://publish.illinois.edu/dbwork/open-data/

  OpenStreet Map

  - http://wiki.openstreetmap.org/wiki/Planet.osm

  Dublin Bus GPS sample data from Dublin City Council (Insight Project)

  - https://data.gov.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

  Further data set at https://data.gov.ie

  Wikipedia Dump

  - Wikpedia Dump https://dumps.wikimedia.org/

  Amazon Review Data Downloader