Final Project Fun

Irene Chen

6.S897 / HST.956: Recitation 5

March 15, 2019

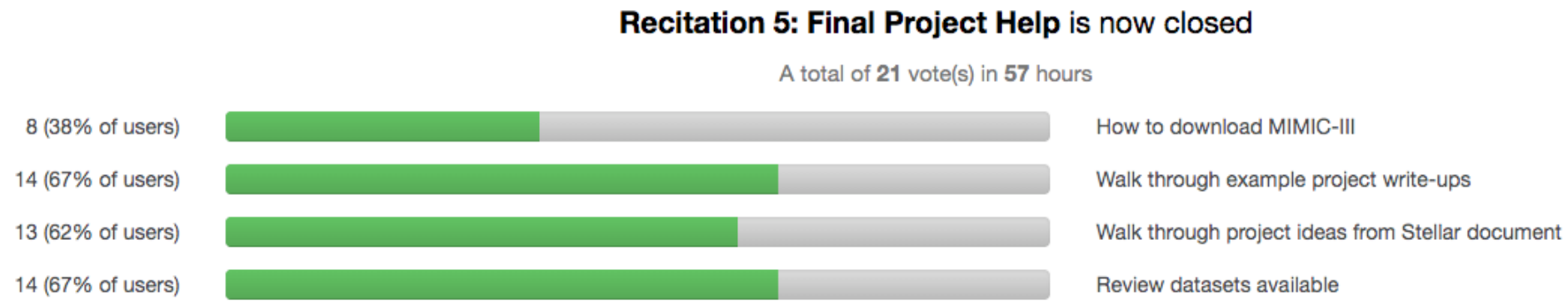
Final projects at a glance

- **Project proposals** due Thurs Mar 21 at 11:59pm
  - At most 3 pages, submit through Stellar
- **Project poster presentations** on Tues May 14, time TBD
- **Project reports** due Thurs May 16 at 11:59pm

- Groups of 3 (2 is possible with TA permission, definitely not 4)
- If interested in using IBM MarketScan dataset, email Willie and Irene

Agenda

1. How to download MIMIC-III
2. Review other datasets
3. Example project from proposal to write-up
4. Project ideas from Stellar document



How to install MIMIC-III?

How to install MIMIC-III?

1. Install Postgres: `brew install postgres`
2. Download CSVs: https://physionet.org/works/MIMICIIIClinicalDatabase/files/
3. Create empty database on Postgres
4. Create empty tables with `postgres_create_tables.sql` from https://github.com/MIT-LCP/mimic-code/tree/master/buildmimic/postgres
5. Load data from CSVs into empty tables with `postgres_load_data.sql` from https://github.com/MIT-LCP/mimic-code/tree/master/buildmimic/postgres

Unix/Mac: https://mimic.physionet.org/tutorials/install-mimic-locally-ubuntu/
Windows: https://mimic.physionet.org/tutorials/install-mimic-locally-windows/

Pop quiz!

You are working on the MIMIC-III dataset for your class project with your two partners. Can you use a public Github to coordinate work?

Pop quiz!

You are working on the MIMIC-III dataset for your class project with your two partners. Can you use a public Github to coordinate work?

Yes – as long as no data is posted publicly including in Jupyter Notebook cells.

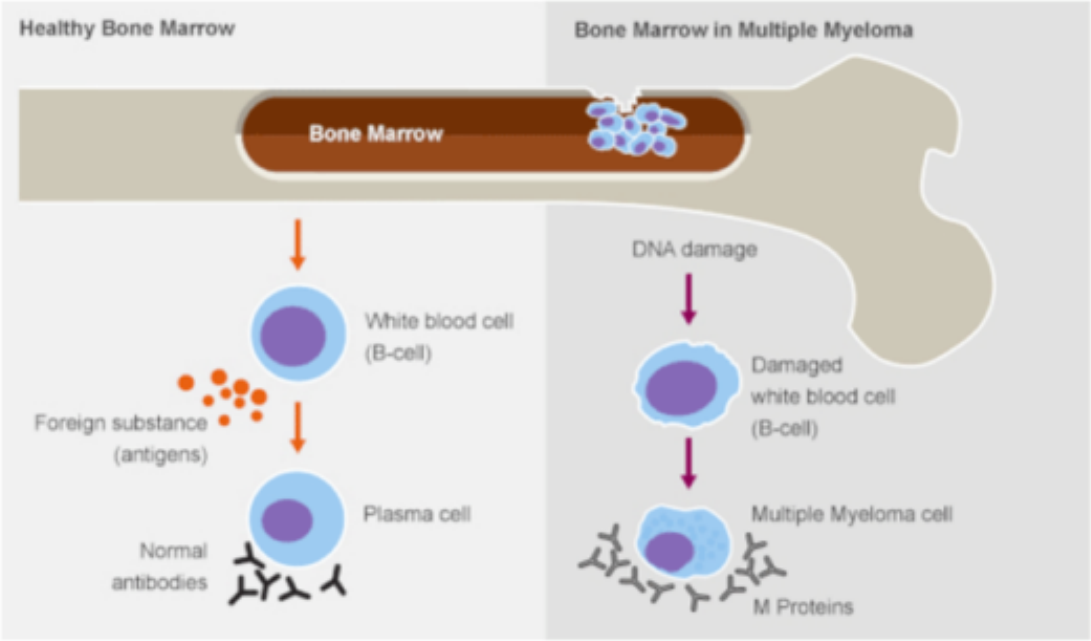
What datasets can we use?

IBM MarketScan

- Insurance claims: more complete because people want to be paid
- Full longitudinal clinical trajectory
- For more information, see Recitation 3 slides or HW2
- **Email Willie and Irene if you want to use this dataset!**

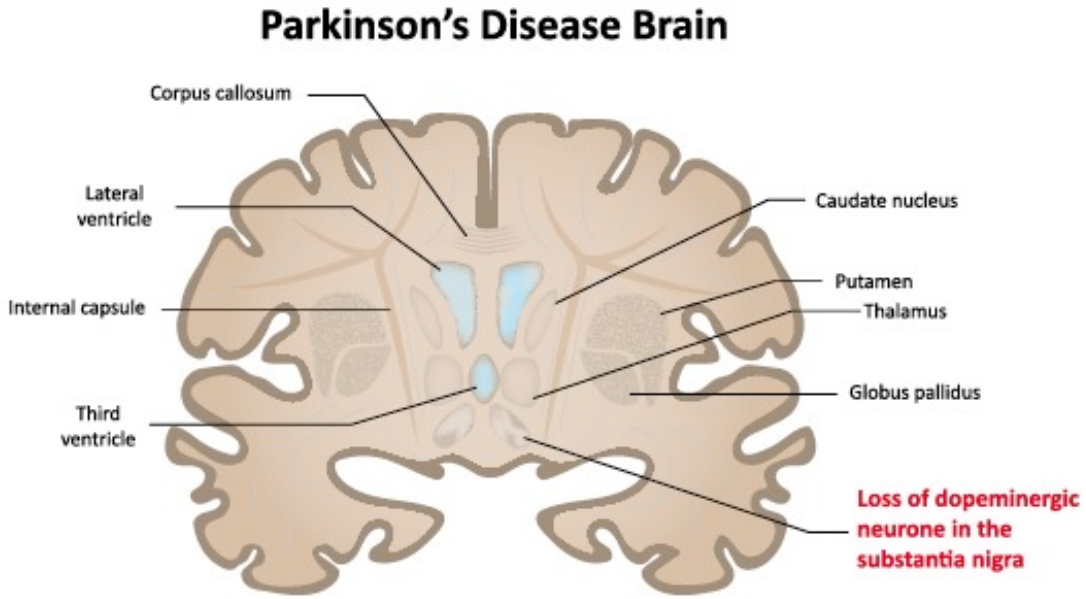
Multiple Myeloma Research Foundation CoMMpass

- Website: https://research.themmrf.org/
- Registry data of ~1200 patients from first diagnosis
- Genetic RNA-seq data, lab values, and patient progression over time



Parkinson's Progression Markers Initiative

- Website: http://www.ppmi-info.org/sion
- Dataset include healthy controls and patients with Parkinson's
- Includes patient progression of disease, for example early-stage and late-stage
- Ex: learn representation of MRI images



eICU Collaborative Research Database

- Website: https://eicu-crd.mit.edu/about/eicu/
- Critical care database across 200+ hospitals
- Includes labs, medications, patient outcomes standardized across hospitals
- Hospitals are anonymized, with only region number and teaching status provided

Other Datasets

- Kaggle cervical cancer screening
  - https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data
- Chest X-rays
  - https://physionet.org/physiobank/database/mimiccxr/
  - https://stanfordmlgroup.github.io/competitions/chexpert/
- Autism Sub-challenge
  - http://emotion-research.net/sigs/speech-sig/is13-compare
- More here: https://github.com/beamandrew/medical-data

What are you looking for in a final project?

Example project write-up: Irene's from 2017

- How to model Heart Failure progression?
- **Original proposal**: Can we predict mortality and time until next visit?
- **Revised**: supervised and unsupervised approaches

Modeling Disease Progression for Congestive Heart Failure

Peniel Argaw, Irene Chen, Sebastian Gehrmann, Harlin Lee, Alisha Saxena

May 19, 2017

Abstract

Congestive heart failure (CHF) is a complex and chronic disease that is difficult to stage correctly. Modeling disease progression must span multiple organ systems and account for extraneous factors like unrelated health problems and generally messy data. In this paper, we explore several approaches to model CHF progression using data from Beth Israel Deaconess Medical Center. First, we use a patient's likelihood of death, time until death, and time until next hospital admission as proxies for their true CHF state. Second, we harness the abnormal lab test time-series data to learn disease progression without labels.

100 by recent count–are ubiquitous yet impractical [8]. It is therefore crucial to develop a better understanding of CHF itself through assessing the progression and the severity of CHF in a patient.

In this paper, we outline and implement a series of approaches to model disease progression using cardiology patient data from the Beth Israel Deaconess Medical Center (BIDMC). Our approaches can be summarized in two angles: first, we use a patient's likelihood of death, time until death, and time until next hospital admission as proxies for their true CHF state. Second, we harness the abnormal lab test time-series data to learn disease progression without labels through clustering and Markov jump processes.

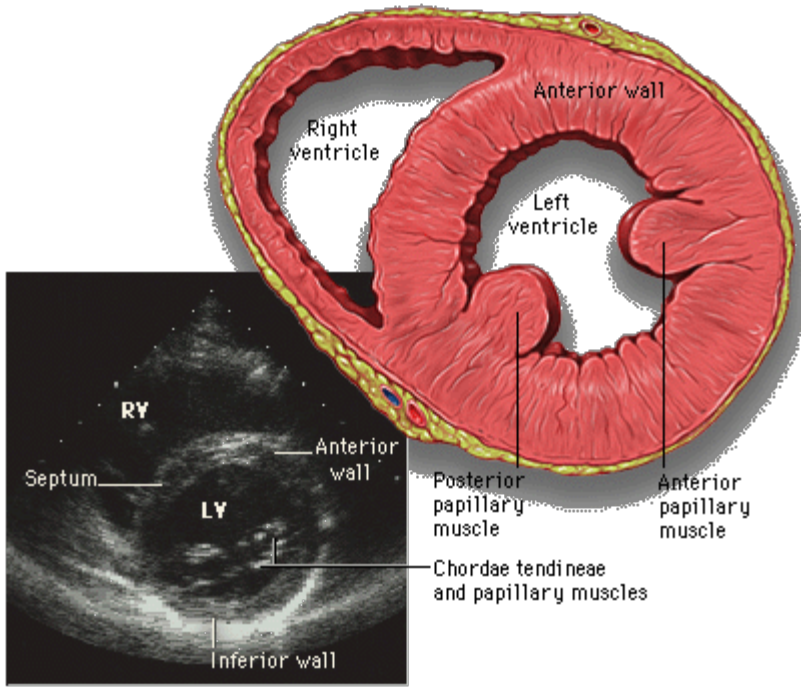
Heart Failure approaches

Supervised learning

- Predict mortality at a specific time, use regression scores to represent the HF state.
- Predict mortality as a multi-task learning problem.
- Predict time until next visit

Unsupervised learning

- Can we cluster disease trajectories while handling missing data?
- How well does [Wang et al, 2014] apply on this dataset?



What’s in that Stellar document?

A tale of two approaches

Clinical problems

- How can we better predict patient outcomes for a disease?
- How can we better understand (e.g. subtype, causal inference) a disease?
- What can we actually *do* in terms of interventions?

Machine learning methodology

- Data is imperfect: missing, not aligned, mislabeled, generally messy
- Privacy, interpretability, and fairness are all important
- We need generalizability of models across dataset shifts
- Deep learning has promise but needs to be evaluated rigorously