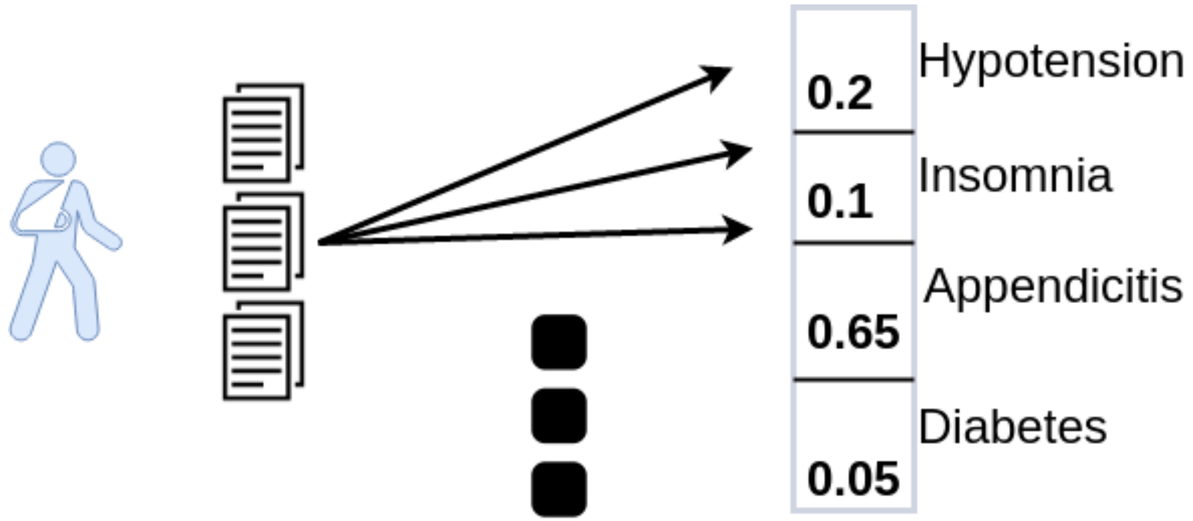
Autumn Rotation 2017

Guhan Venkataraman
Biomedical Informatics

RIVASLAB

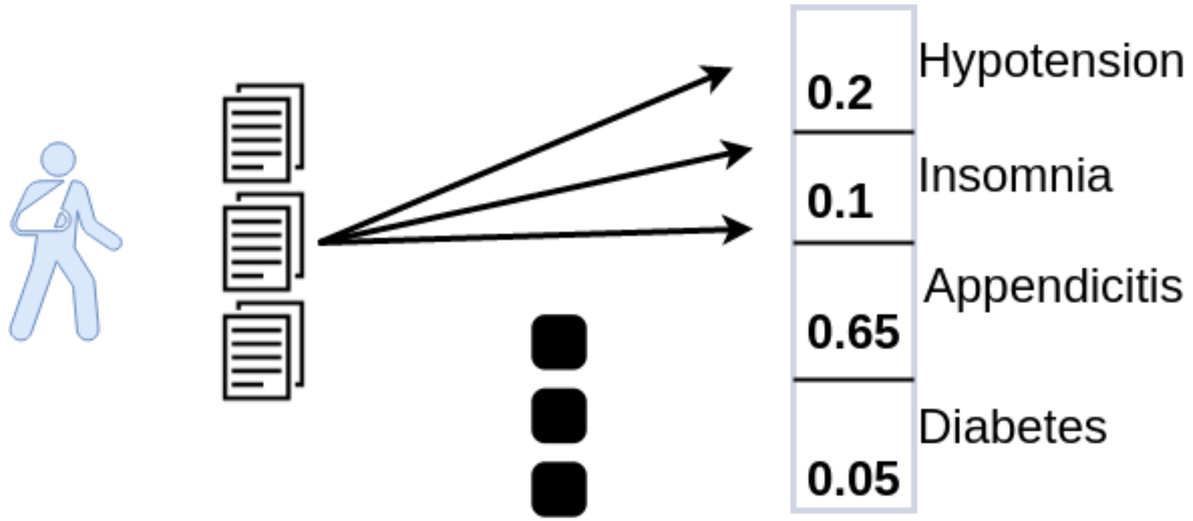
The CNT (Clinical Note Tagger) - Motivation

- Filling out structured data in EHR is a **drain** on resources
- Upwards of **$25 billion** spent on the tagging of clinical notes with ICD codes (hierarchical diagnosis/ intervention codes meant for billing, treatment, and clinical study purposes)

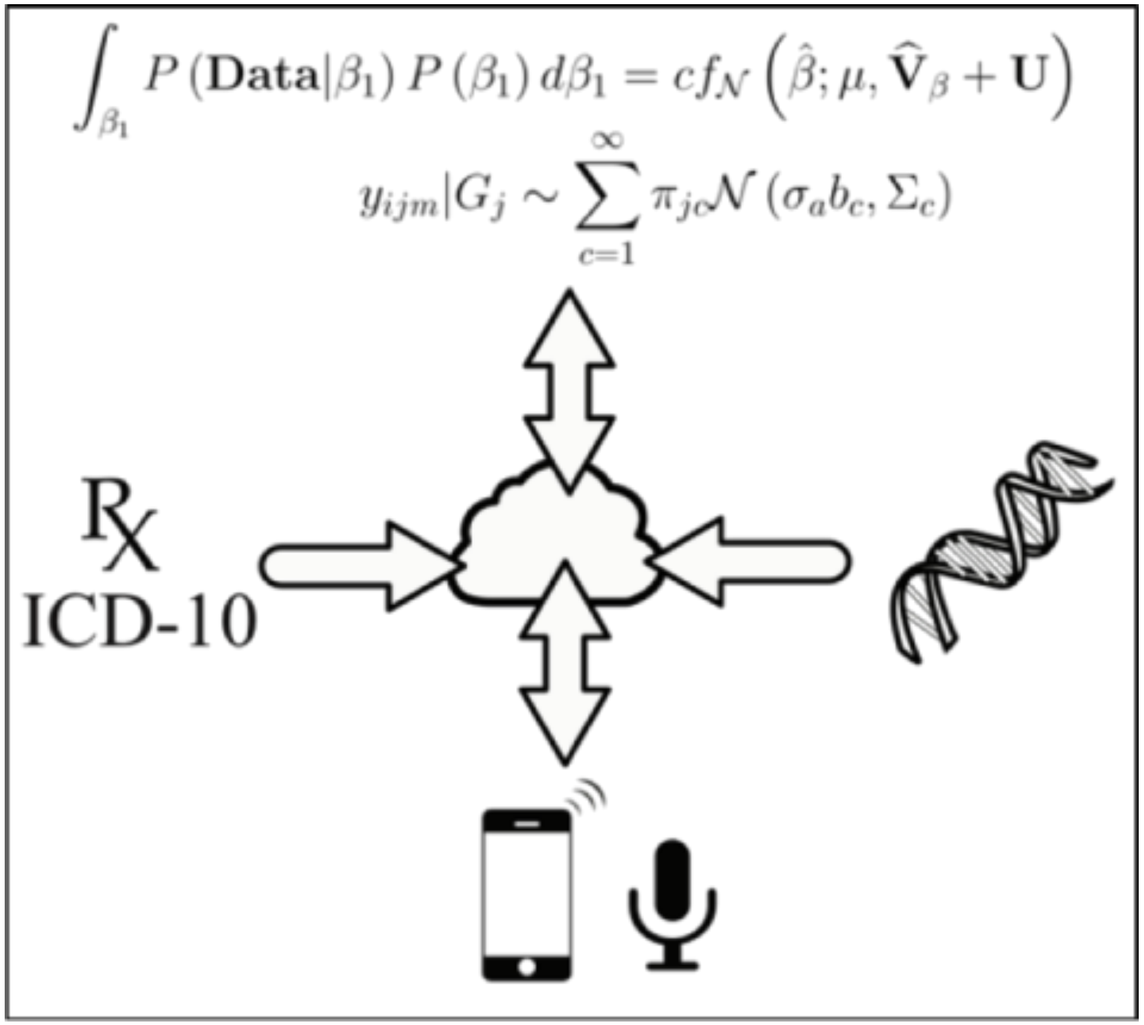


The CNT (Clinical Note Tagger) - Motivation

- **ICD codes** are **confusing,** but **correct assignment** is **critical** to patient outcome
- Goal of CNT: **Assign ICD-9 codes** to patients **using their clinical notes, reducing cognitive workload** and **standardizing billing practices** in the process



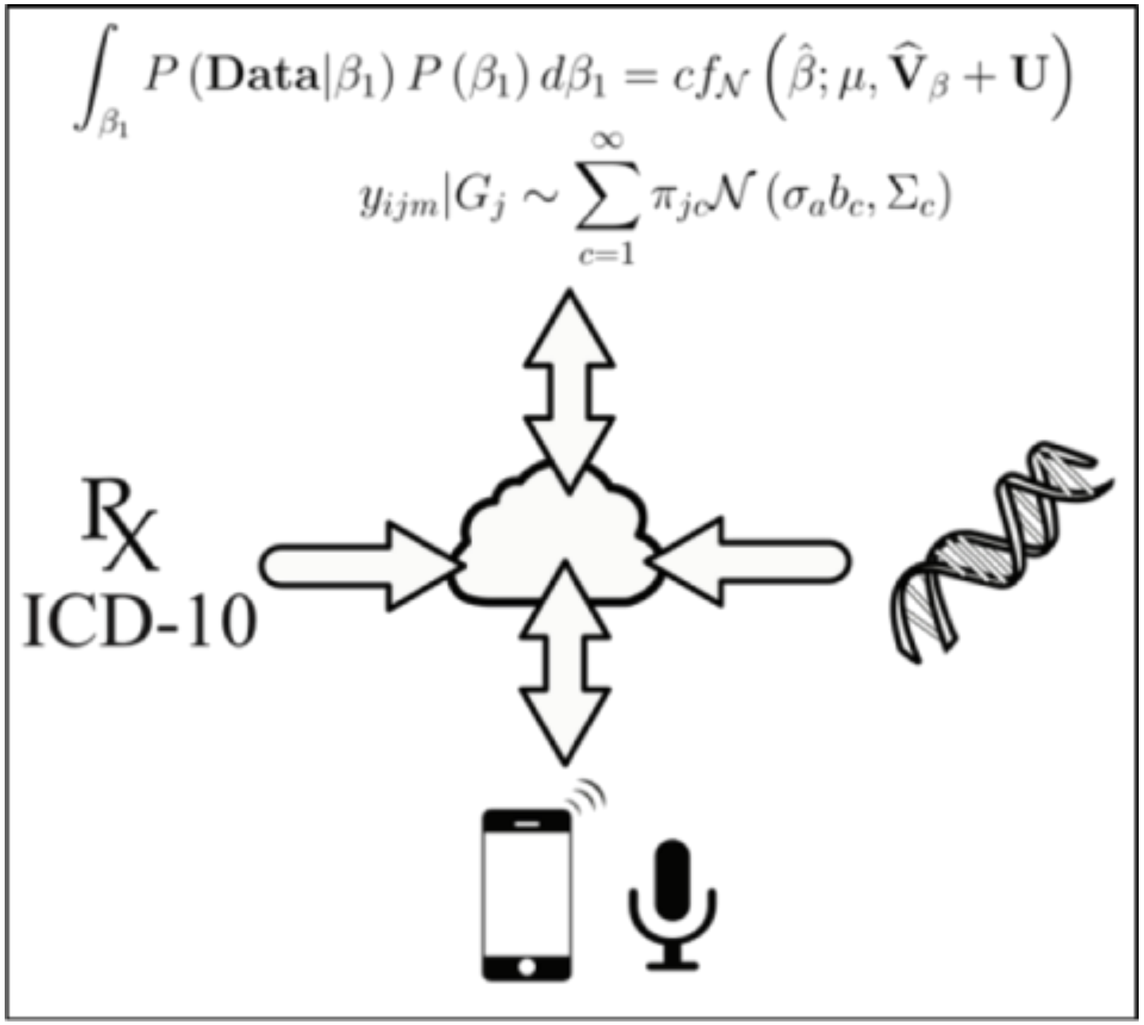
ALTUD: AnaLysis of Text/Speech for Understanding Diseases



- **Predicting ICD codes** shouldn't be the only way to **assign disease risk**
- With **surge** in **heterogeneous data** (longitudinal EHR including genetic, imaging, etc.), need a way to **store** it all and **make inferences**

RIVASLAB

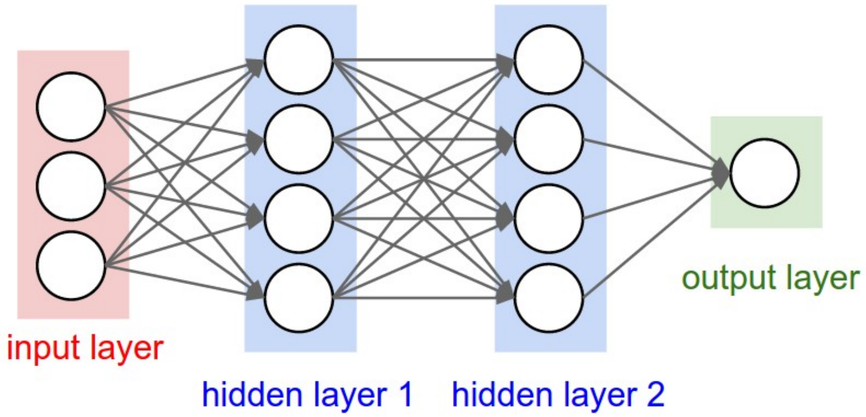
ALTUD: AnaLysis of Text/Speech for Understanding Diseases



- ALTUD: a **database** in which clinical notes, genomes, and other clinically relevant data can **jointly** be **stored not far in the future**
  - **Aggregate risk** from a combination of risk models can be assigned to each patient, directing them to **gold standard diagnostic** and/or **preventive** options

RIVASLAB

CNT - Methods

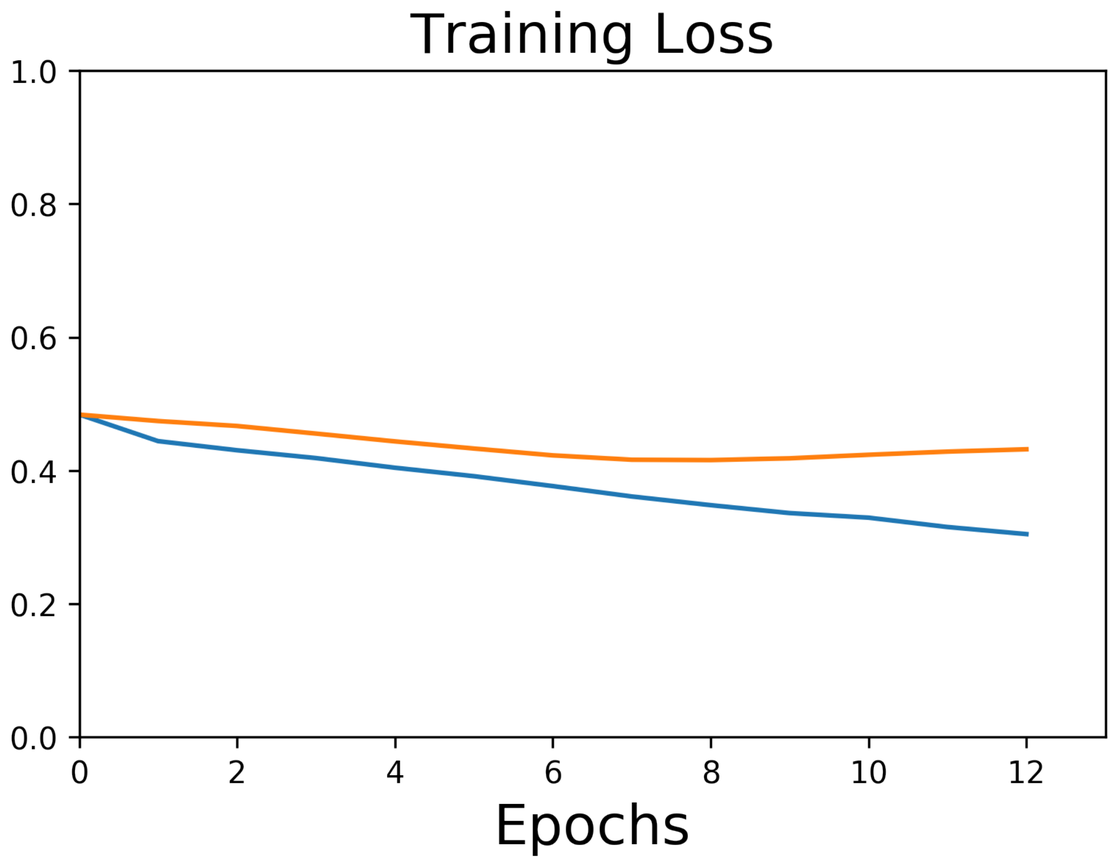


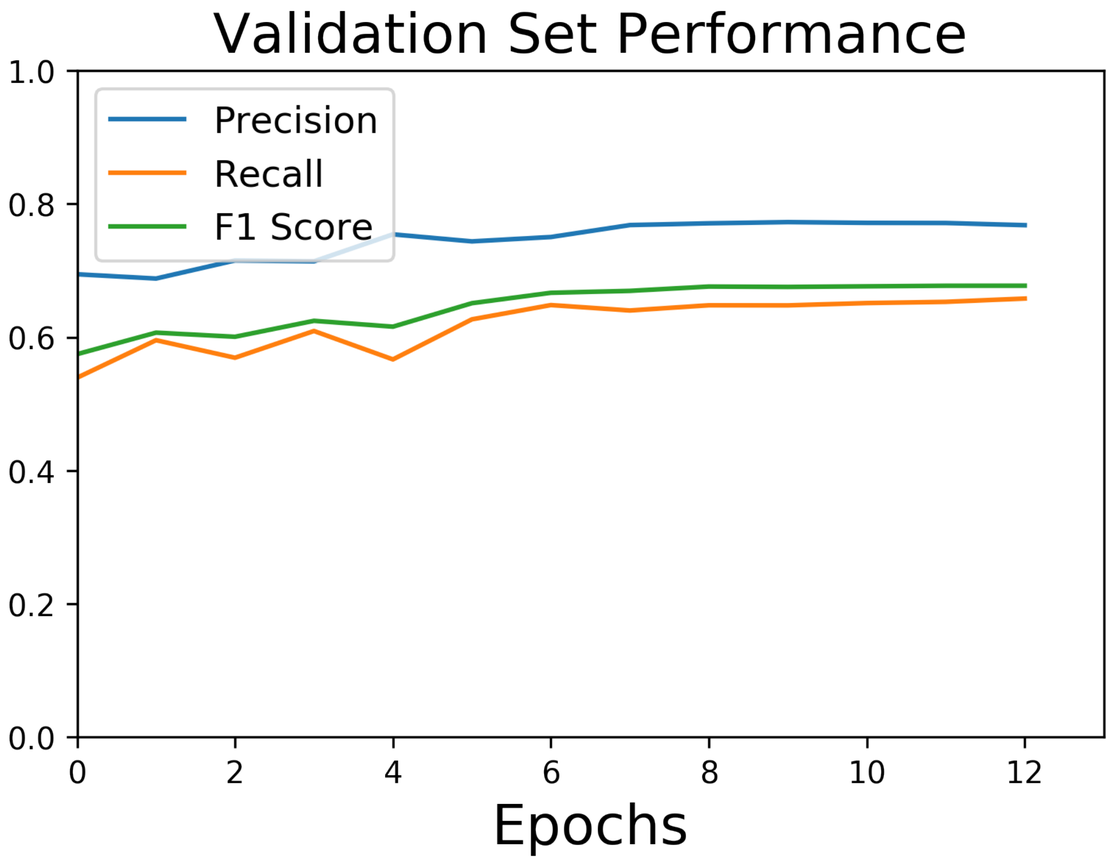
- Clinical notes, audio files **converted** to word vectors (i.e. GloVe, fastText)
- **LSTM** (type of RNN) used to **learn** the sequential nature of the notes
- **Trained** on truncated **notes** of a user-specified length
- Results yielded **accurate** prediction

CNT - Results (Oliver Bear Don’t Walk IV)

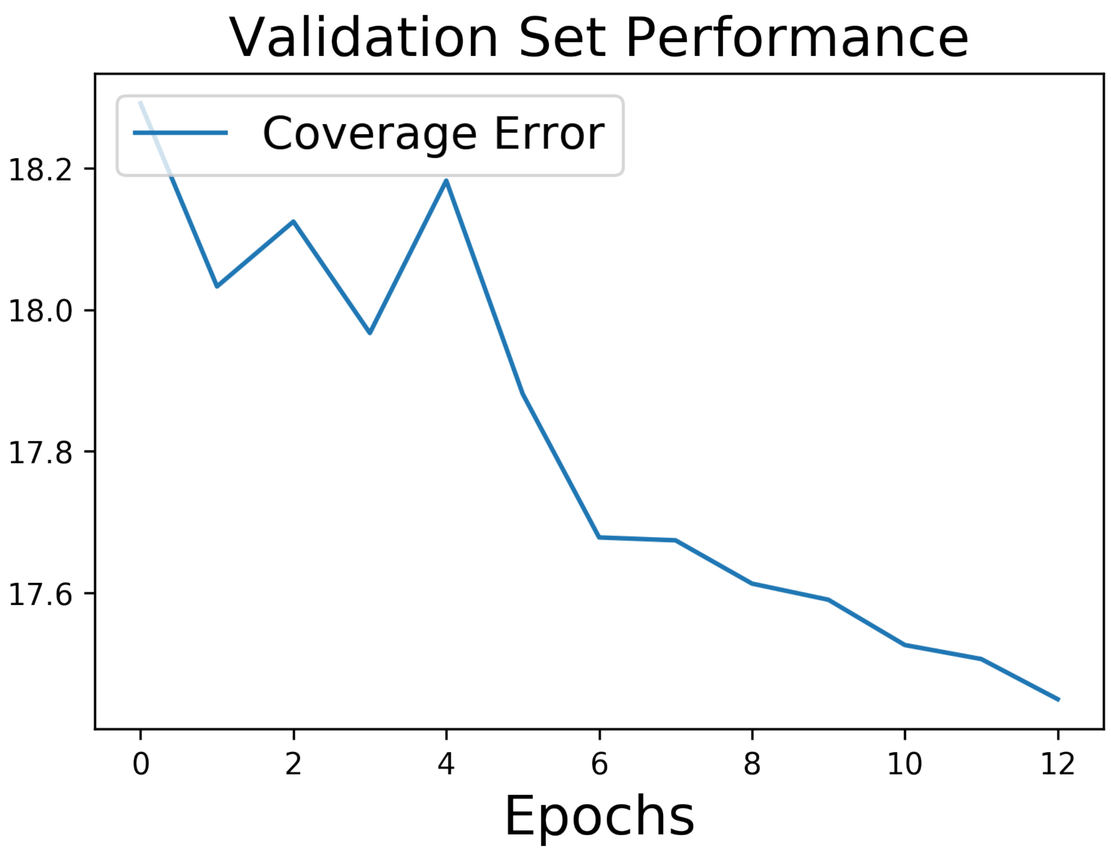
- **MIMIC** dataset
- LSTM trained on discharge summaries from Beth Israel Hospital + ICD-9 diagnoses from those same visits
- Without any curation/babysitting, does pretty well

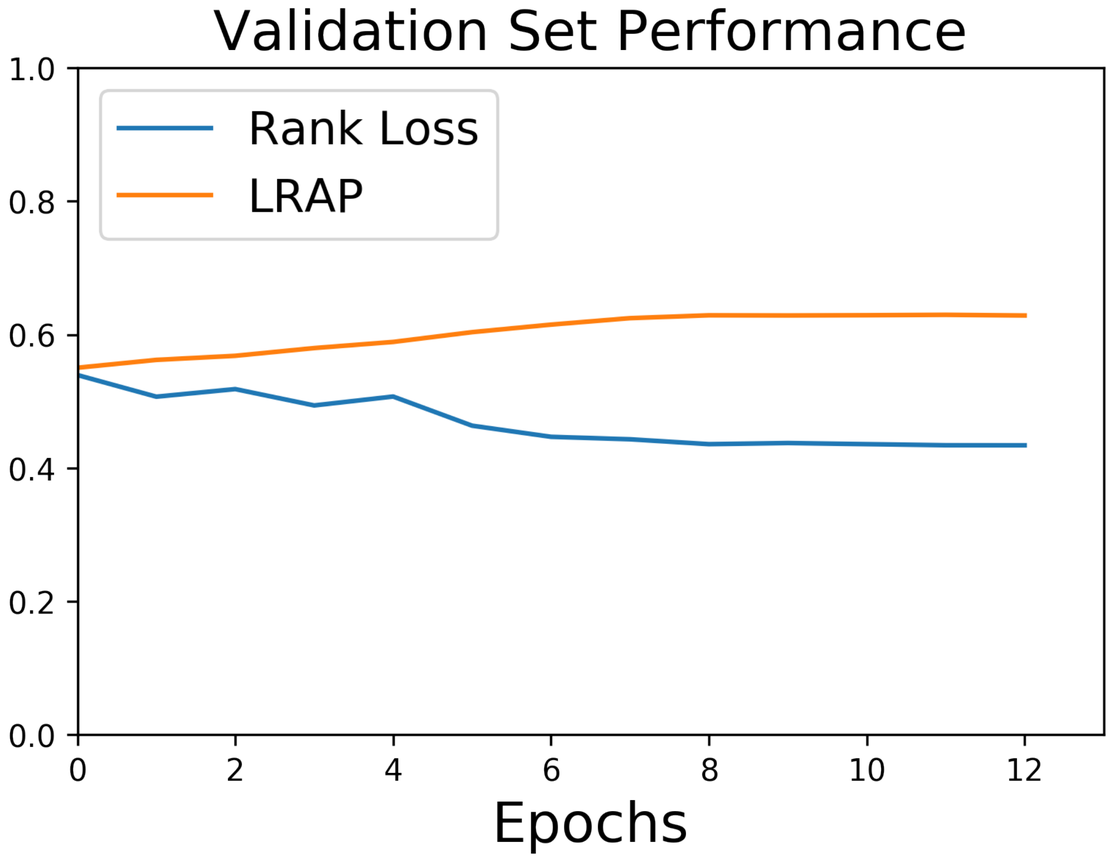
CNT - Results (Oliver Bear Don’t Walk IV)





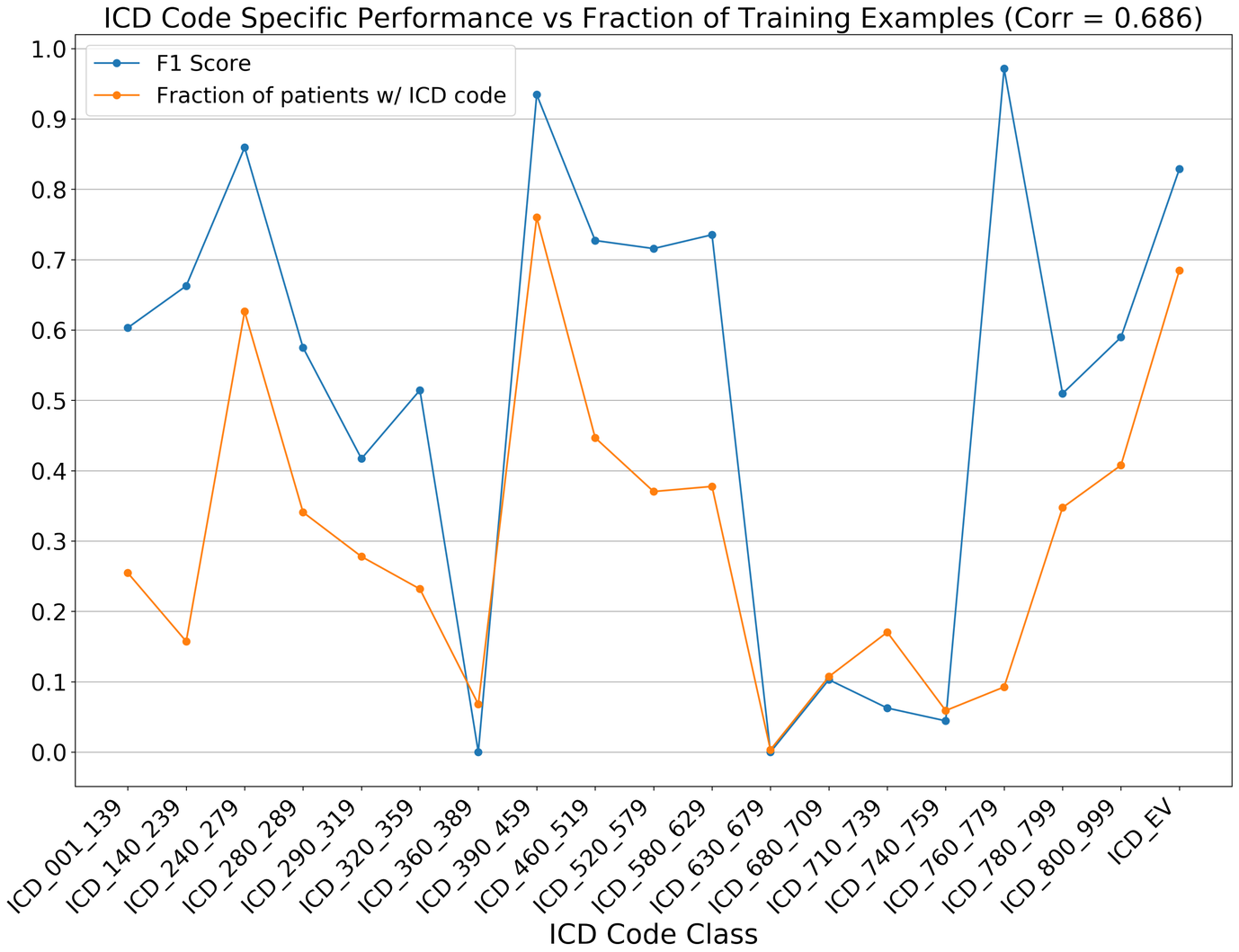
CNT - Results (Oliver Bear Don’t Walk IV)





RIVASLAB

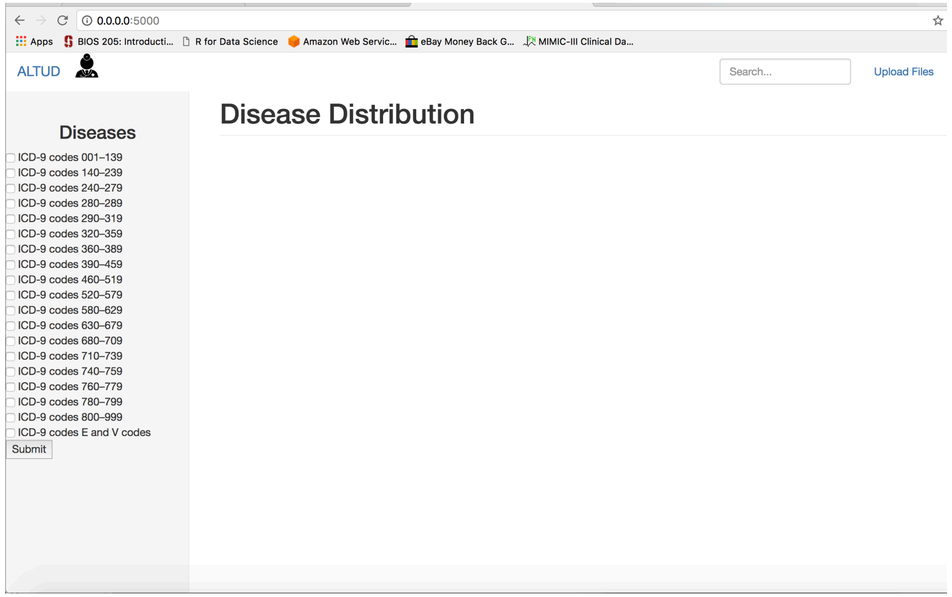
CNT - Results (Oliver Bear Don’t Walk IV)



RIVASLAB

ALTUD: Appearance and Functionality

- Biggest problem in science today is reproducibility
- Luckily, software engineers are really good at it
- Challenge 1: to make ALTUD platform universally accessible with the current modalities it supports (clinical notes and voice)
- Challenge 2: to make the clinicalNoteTagger intuitive to use on new datasets in order to plug into ALTUD and use out-of-box



RIVASLAB

ALTUD: Modifications

- To address challenge 1: Docker
  - VM alternative that sacrifices isolation for speed
  - Containers are light-weight and require fewer resources than full VMs
  - Fastest growing open-source platform
  - Perfect for our use case of making sure ALTUD dependencies are all taken care of, regardless of machine

```
docker build -t altud .; docker run -p 5000:5000 altud
```

Demo

RIVASLAB

CNT: Modifications

Deep Learning AMI Amazon Linux - 3.3_Oct2017 - ami-78994d02

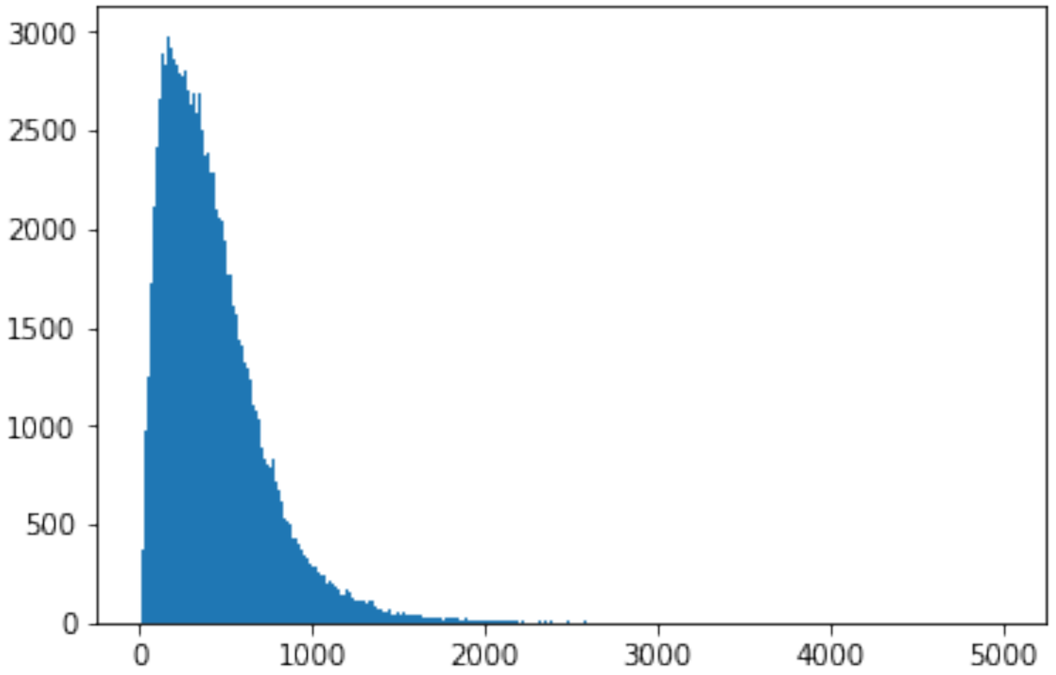
Deep Learning on Amazon Linux with MXNet, Tensorflow, Caffe, Theano, Torch, CNTK and Keras

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

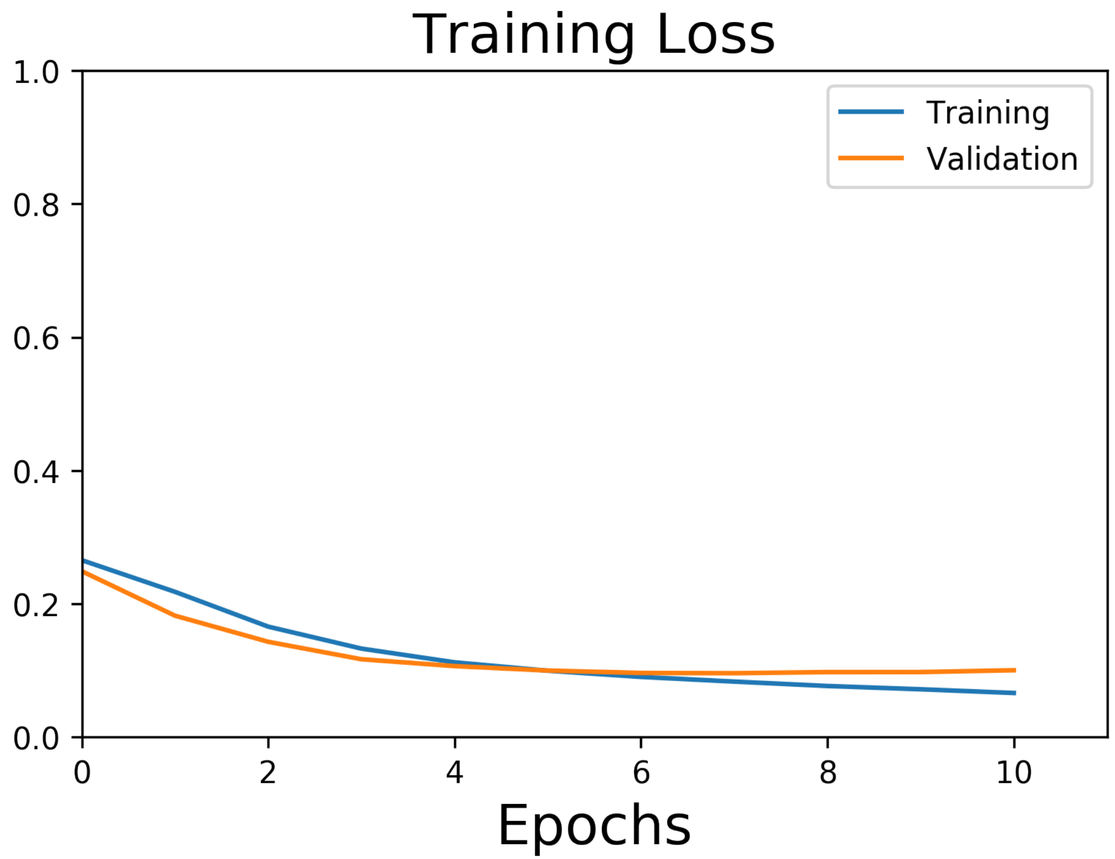
- To address challenge 2: Reconfiguration of the CNT
- Challenge: getting TensorFlow working on a GPU
- After many failed installation attempts, went for Amazon Deep Learning AMI (ami-78994d02)
- Had TensorFlow + GPU built in with all drivers pre-installed
- Was able to reproduce the plots Oliver had made for MIMIC data, and cleaned up the code significantly (made it better for first-time users via prompt interfaces, explanations, etc.)

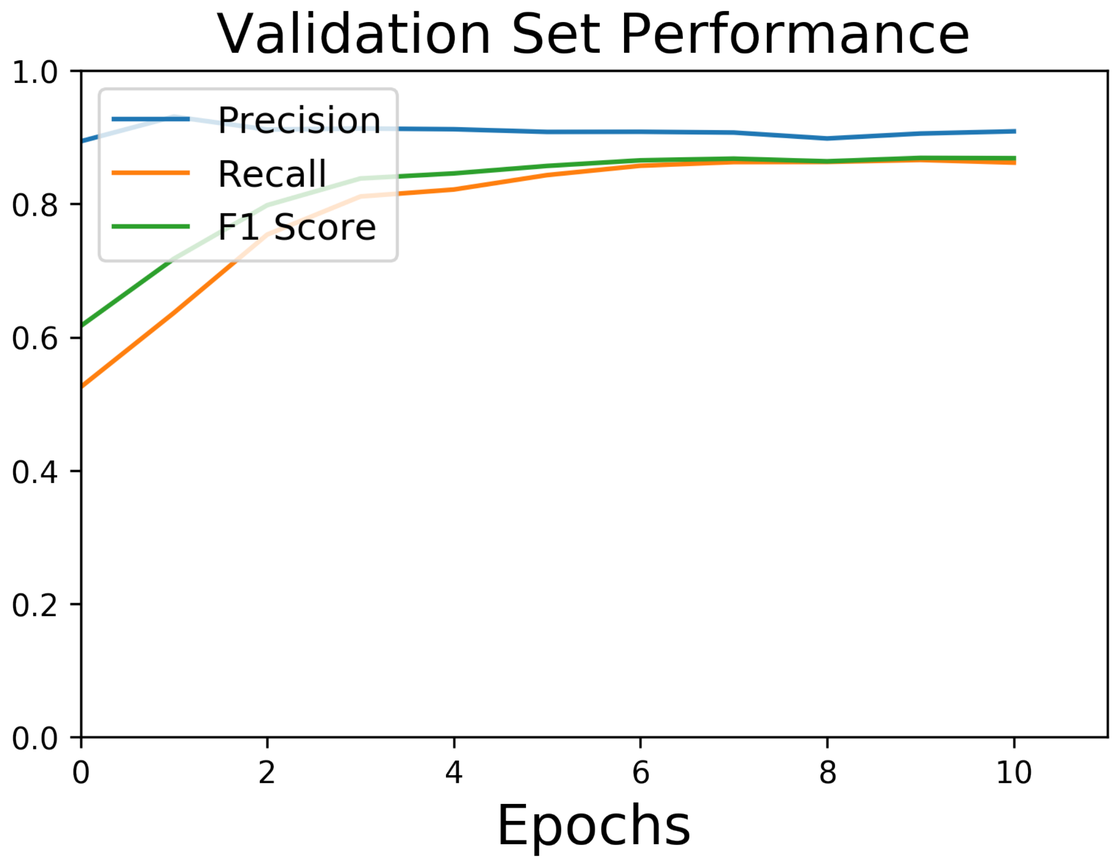
CNT: Extending Results to Veterinary Data from CSU

- Configured the tagger for the new data prepared by Ashley after discussing with Oliver
  - Changes: made tagger data-agnostic, as long as 1 column for hyphen-separated classes and 1 column with notes
  - Retained pre-trained word vectors (GloVe)
  - Set max note length at 1000 and trained
    - Happened to be same threshold as that set for MIMIC data
    - Captures most of the notes (see right)

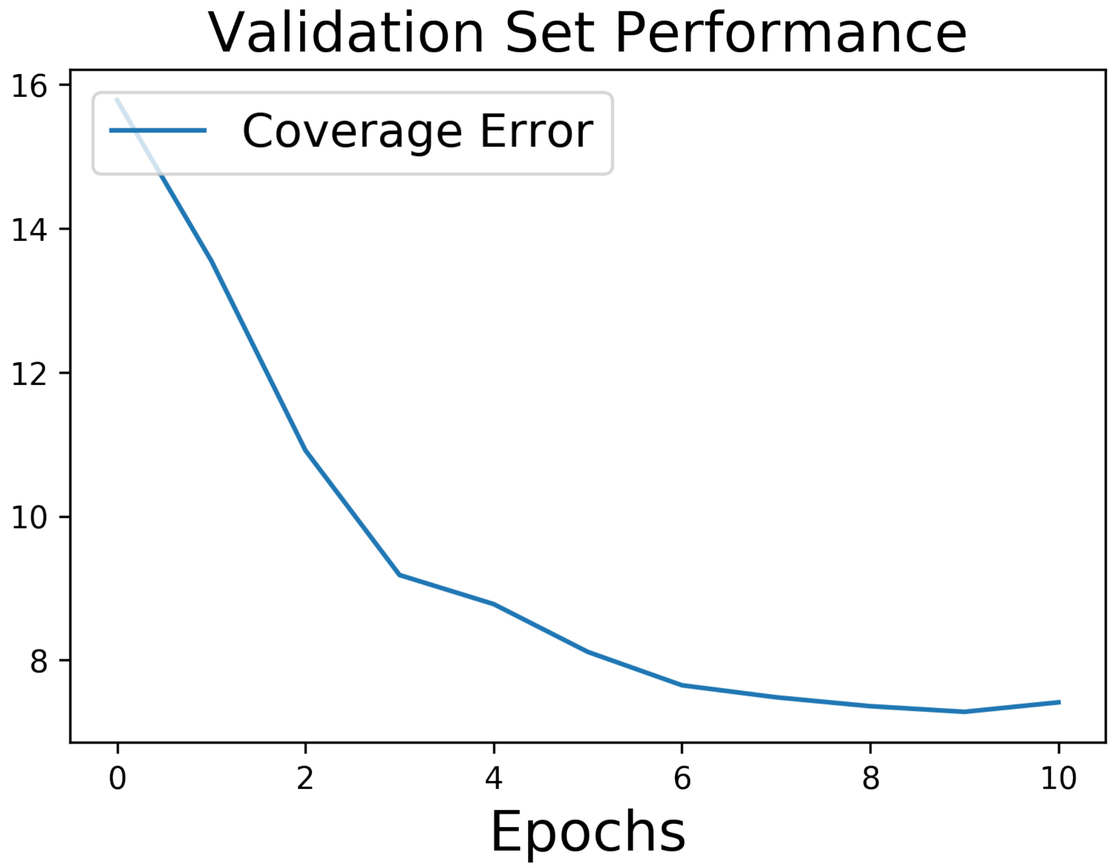


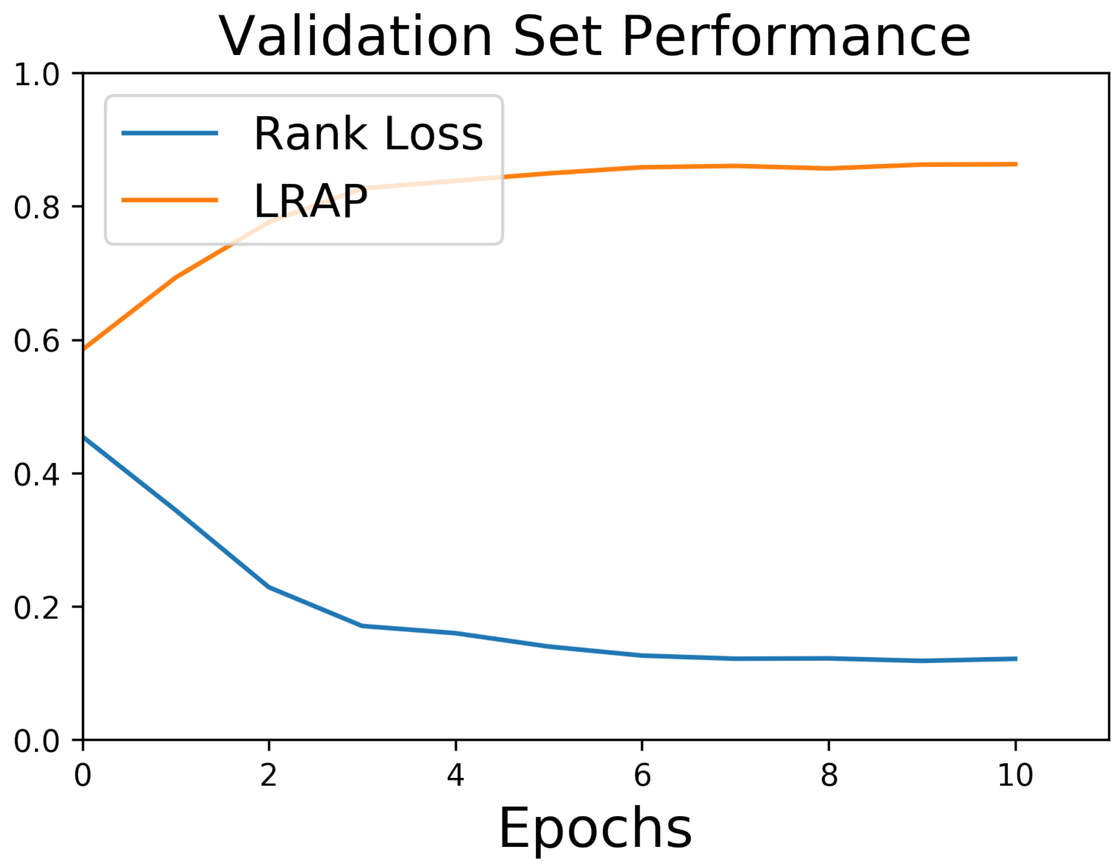
CNT - Results (SNOMED Veterinary Data - CSU)





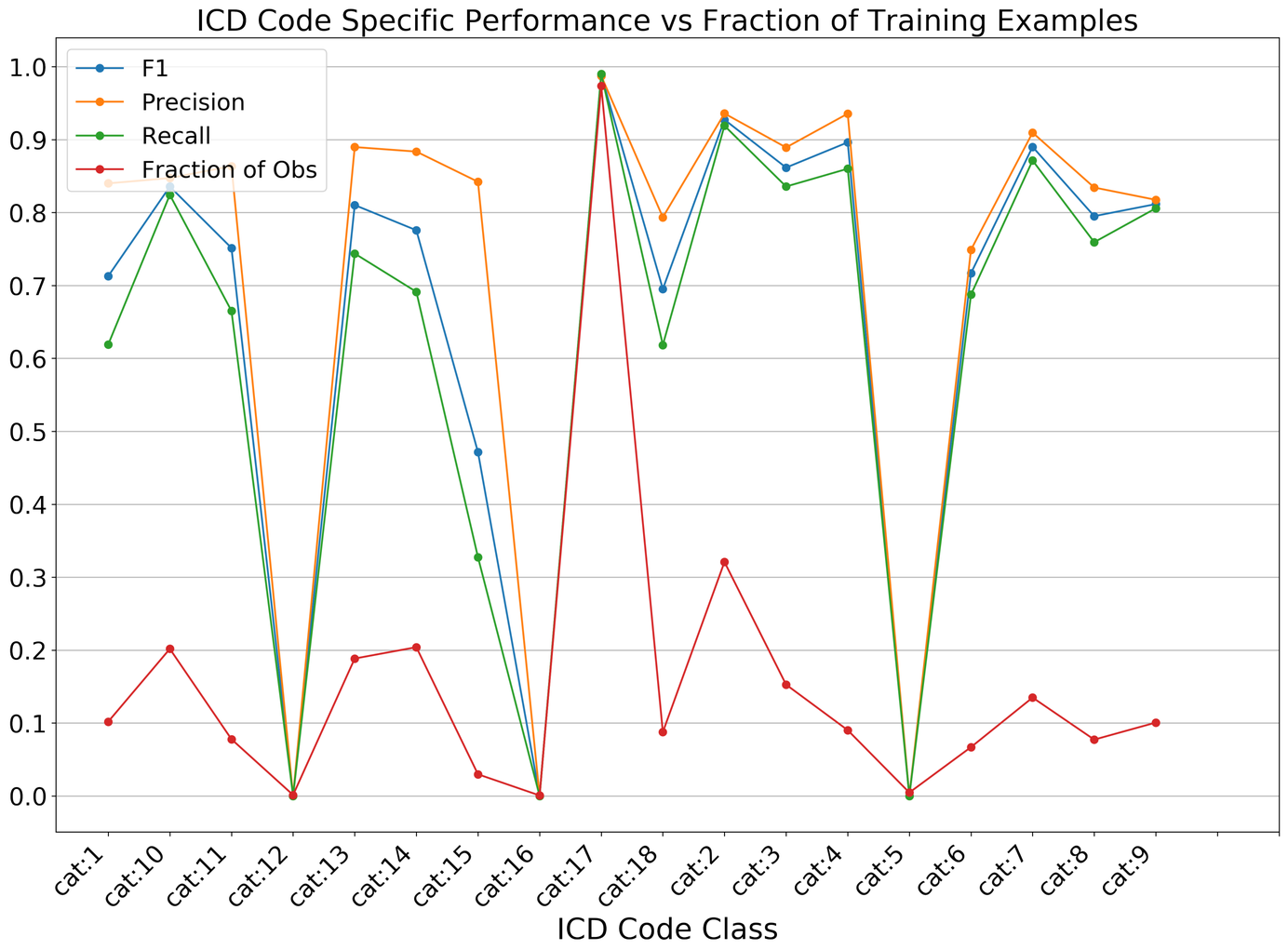
CNT - Results (SNOMED Veterinary Data - CSU)





RIVASLAB

CNT - Results (SNOMED Veterinary Data - CSU)



RIVASLAB

Acknowledgements

Rivas Lab

Dr. Manuel A. Rivas

Oliver Bear Don’t Walk IV

Chris DeBoever

Yosuke Tanigawa

Matthew Aguirre

Bustamante Lab

Ashley Zehnder

BMI Training Program

NLM Grant

RIVASLAB