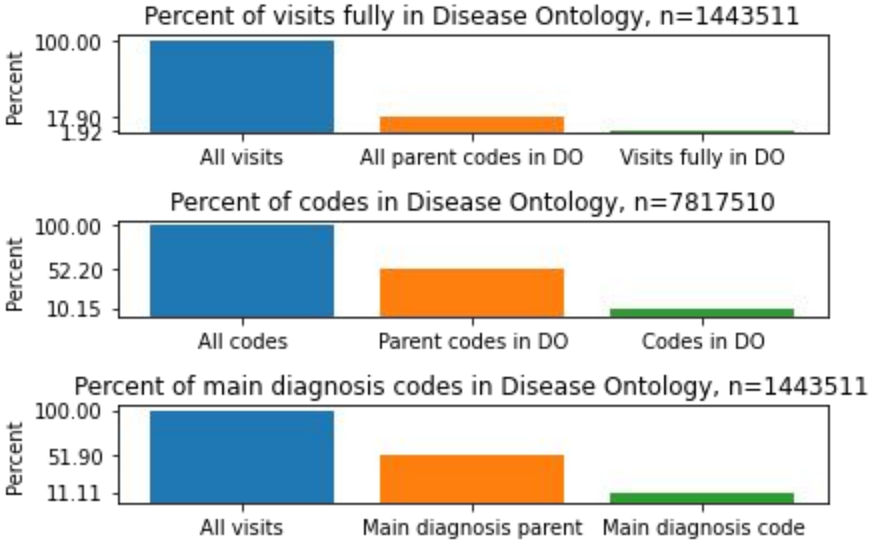
Master thesis

25. May
Jannik Gut

BfS codes in Disease Ontology

- Latest (2018) dataset
- Parent codes = second lowest level
- Repetitions counted
- Results
  - ~half the parent codes in DO
  - ~10% of full codes in DO
  - Only ~2% of visits fully in DO
  - Average visit has 5,4 codes
- Discussion
  - Codes in DO probably in similar category
  - Still, 10% of full codes and 2% of visits is thin



Thoughts on using SubGNN

- The hierarchy graph is in my opinion not interesting/informative/variable enough to warrant graph-only initialisation and processes
- SubGNN on this kind of graph feels shoehorned in, as (without changes) all components are just nodes, not more complex
  - Also the edges between the nodes are too symmetric
- The disease ontology alone is not more interesting than the ICD-hierarchy and I am not too optimistic about the amount of ICD codes in the disease ontology or SPOKE
- SPOKE has a lot of information, but I don’t know how much is relevant
- SPOKE is part of a different organisation, that we do not have on our servers, which makes many things more complicated (if even possible)
- **One step that helps towards the goal of CHOP prediction is to find good representations of ICD-codes, which can also be used for other tasks and, in my opinion has some room for research.**

ICD code embeddings

- Only co-occurrence based
  - Med-BERT/BEHRT
    - SOTA, also due to large cohort
    - Maybe hard to get, because of protected cohort
  - Word2Vec
    - Viable
    - Based on Gensim → we can do ourselves if needed
- Only hierarchy based
  - hyperE
  - (Note: hyperbolic ML still is in infancy, libraries are still WIP, but using hyperbolic embeddings with classical ML functions is viable)
- **I could not find a paper on how to combine them and am interested in doing just that.**

Existing Benchmarks

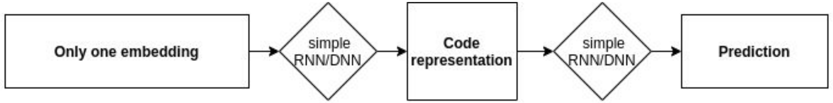
- From Med-BERT
  - Heart failure prediction in Cerner
  - Pancreatic cancer prediction in Cerner, Truven
- From BEHRT
  - Disease (code) prediction in next/future visit in UK CPRD
- From Word2Vec
  - Cluster evaluations in KPMAS
  - Logistic regression for unplanned hospital admission, patient mortality in KPMAS
- All(?) datasets are somewhat proprietary, but similar tests can be done on our data.
- *From Hyperbolic embeddings*
  - *Unplanned ICU readmission from MIMIC-III with auto-labelling*
  - *In-Hospital Mortality prediction from MIMIC-III with auto-labelling*

New benchmarks

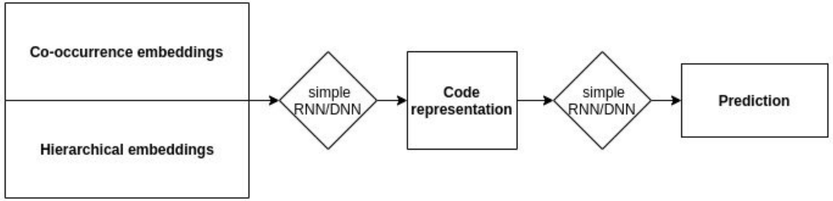
- **CHOP code prediction, a multi-label prediction task**
  - At least for the beginning, maybe change “simple” multi-label to neural machine translation, but topic for a later stage
- New ideas (needed)?
  - Data excursion may help

Modes

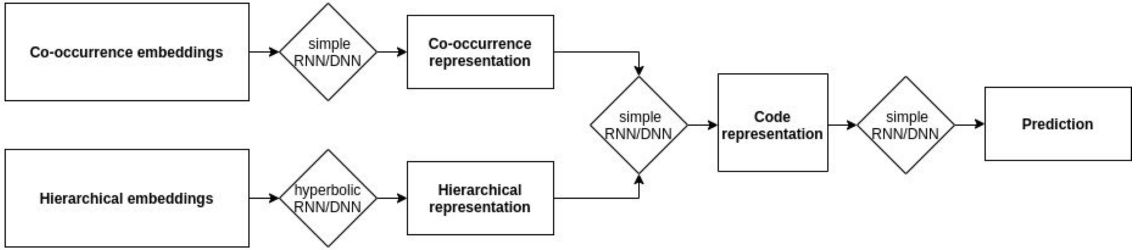
- Separated



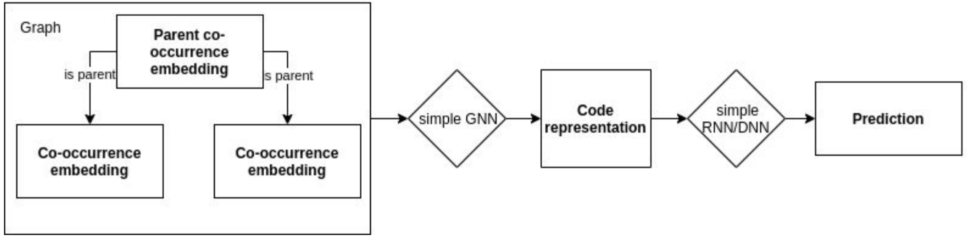
- Early fusion

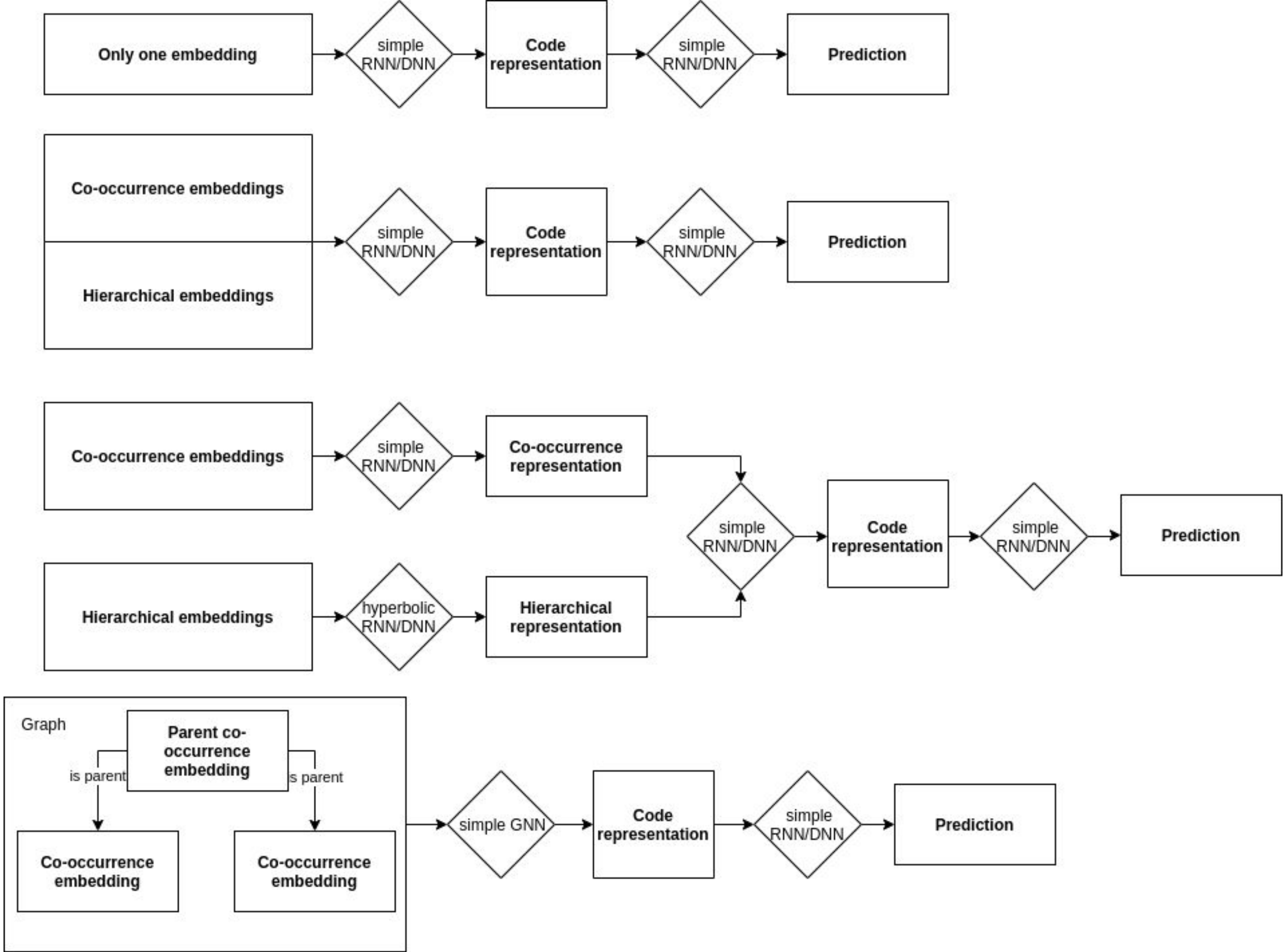


- Late(r) fusion



- Graph




Only one embedding
simple RNN/DNN
Code representation
simple RNN/DNN
Prediction
Co-occurrence embeddings
Hierarchical embeddings
simple RNN/DNN
Code representation
simple RNN/DNN
Prediction
Co-occurrence embeddings
simple RNN/DNN
Co-occurrence representation
Hierarchical embeddings
hyperbolic RNN/DNN
Hierarchical representation
simple RNN/DNN
Code representation
simple RNN/DNN
Prediction
Graph
Parent co-occurrence embedding
is parent
is parent
Co-occurrence embedding
Co-occurrence embedding
simple GNN
Code representation
simple RNN/DNN
Prediction

Next steps

- Mail ICD-BERT researchers?
- Read up on KAME or GRAM etc.
- Concrete concept even more
- Data excursion
  - Check which are ICD-9, which are ICD-10
  - Distribution
    - Codes
      - Also parent codes
    - “Normal” data
  - More data marshalling
  - Can I make marshalled pickles in my home directory on Leomed?