# Master thesis

19. May

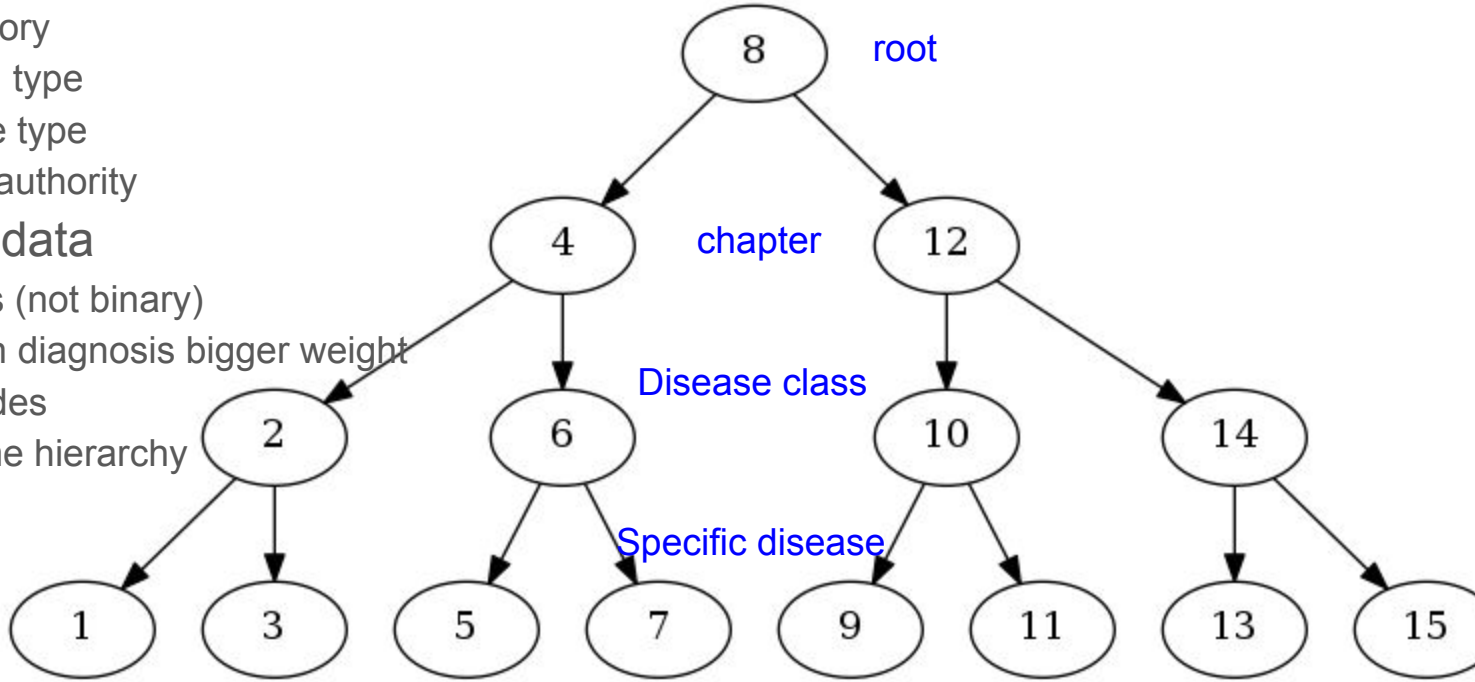
Jannik Gut

# Disease Ontology check

- All ICD-codes in Disease ontology collected
- Script written to count the amount of codes in BfS data, which are in Disease Ontology
  - Report % of
    - All diagnosis
    - Main diagnosis
    - Fully documented visits
  - Tested on very small subset
  - Get error “Permission denied” to BfS data often, but not always
    - Fix?
      - Already checked old code
      - Already checked Euler wiki

# BfS data

- Categorical data:
  - Gender
  - Age category
  - Admission type
  - Residence type
  - Referring authority
- Hierarchical data
  - ICD codes (not binary)
    - Main diagnosis bigger weight
  - CHOP codes
    - Same hierarchy



# Graph hierarchy embedding

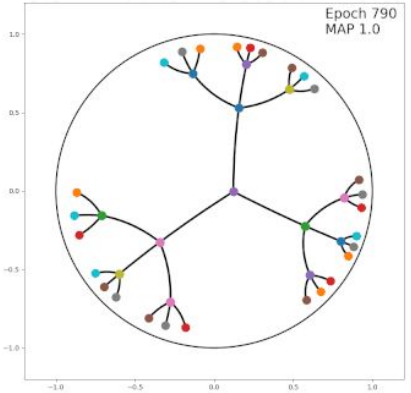
- Only one relation type
- Leafs (more than) half of the nodes have only one edge
- Add a two-hop relation during inference?
- Add a “not-related” relation during inference?
  - Should be covered by negative sampling.

- This task is rather uniform for a tree
  - “Sisters” could be have very similar representation
  - GIN encoding probably more expressive than strict TransE
    - Testing makes smarter, but first plan

# SubGNN

- Three channels
  - Neighbourhood
    - Weakness of SubGNN
  - Position in graph
    - Total position not interesting for all codes at bottom of same tree
    - Relative position pretty interesting
  - Structure
    - Not interesting if all components just nodes
    - Special component from main diagnosis up to root?
      - Implicitly boosts relative position in graph.
- New channel are categorical features
  - At first regular DNN
- **Exact prediction still not sure**
- Loss is length until correct node in CHOP hierarchy
  - Maybe different stages, first training stage only expect prediction at depth 1

# Other embedding types




- Strictly hierarchical → hyperbolic embeddings
  - General paper
  - ICD embeddings already exist
  - HyperCore for ICD codes (next slide)
- NLP-type techniques on ICD-codes
  - word2Vec → these Embeddings
  - BERT → Med-BERT (trained for a week on a 30 Mil. cohort)
- **If NLP techniques work well, why not try (hyperbolic) neural machine translation?**
  - From (Med-BERT) ICD to CHOP.

Table 6: Results of retrained CNNs on MIMIC-III with top-50 and full ICD codes. **Bold** text denotes the new state of the art obtained by an old model.

| Model | MIMIC-III Top-50 Codes | | | | | MIMIC-III Full Codes | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC-ROC | | F1 | | P@5 | AUC-ROC | | F1 | | P@8 | P@15 |
| | Macro | Micro | Macro | Micro | | Macro | Micro | Macro | Micro | | |
| CNN (Kim, 2014) | 87.6 | 90.7 | 57.6 | 62.5 | 62.0 | 80.6 | 96.9 | 4.2 | 41.9 | 58.1 | 44.3 |
| CAML (Mullenbach et al., 2018) | 87.5 | 90.9 | 53.2 | 61.4 | 60.9 | 89.5 | 98.6 | 8.8 | 53.9 | 70.9 | 44.5 |
| MultiResCNN (Li and Yu, 2020) | 89.9 | 92.8 | 60.6 | 67.0 | 64.1 | 91.0 | 98.6 | 8.5 | 55.2 | 73.4 | 58.4 |
| HyperCore (Cao et al., 2020) | 89.5 | 92.9 | 60.9 | 66.3 | 63.2 | 93.0 | 98.9 | 9.0 | 55.1 | 72.2 | 57.9 |
| CNN (retrained) | **90.8** | **93.1** | **62.4** | **67.1** | 64.0 | 85.0 | 97.4 | 5.9 | 36.5 | 49.0 | 39.4 |

# HyperCore

- Annotate ICD code from clinical notes (NLP)
  - Multilabel prediction
- Focusses on
  - Hierarchical structure → hyperbolic distance score
  - Co-occurrence → co-graph embedding
  - Project the embeddings with a matrix each and add them together
  - Each code then has a score
-
- Directly related codes attract each other in the hyperbolic space and negative samples repel each other.
- Clinical notes get encoded with CNNs to give code-wise document encodings.
  - The projection to the hyperbolic space is done with two MLPs
    - Length of the vector (bigger is further from root)
    - Direction of the vector (which kind of diagnosis)
-
- Co-graph has edge weights between codes is proportional to the amount of times the codes were seen together. Afterwards run normal GCN.

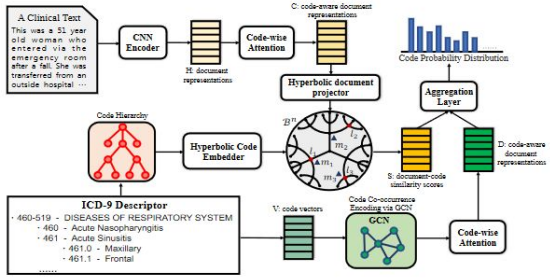



Figure 3: The architecture of **Hyper**bolic and **Co**-graph **R**epresentation method (HyperCore). In the Poincaré ball $\mathcal{B}^n$, we show the embeded code hierarchy (i.e., tree-like hierarchical structure). The dots $l_i$ $(i = 1, 2, 3)$ on the tree-like hierarchical structure and triangles $m_i$ $(i = 1, 2, 3)$ in the Poincaré ball denote hyperbolic code embeddings and hyperbolic document representations, respectively.

# Next steps

- Finally get Disease Ontology comparison done
  - Not too optimistic, honestly
- Think about how to predict codes for SubGNN model
- Read up on Neural Machine Translation
  - What is needed?
    - Do we have that for our problem?
- Report back on Friday evening