Measuring Redundancy in Czech Electronic Health Records

Near-Duplicate Detection and Cluster Analysis

Krištof Anetta, Aleš Horák

xanetta@fi.muni.cz, hales@fi.muni.cz

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

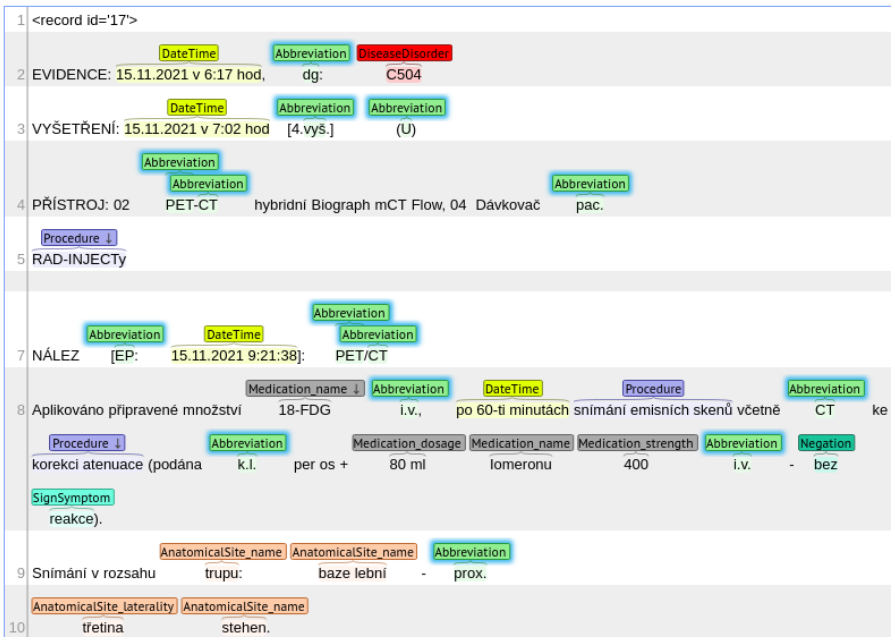
December 6, 2025

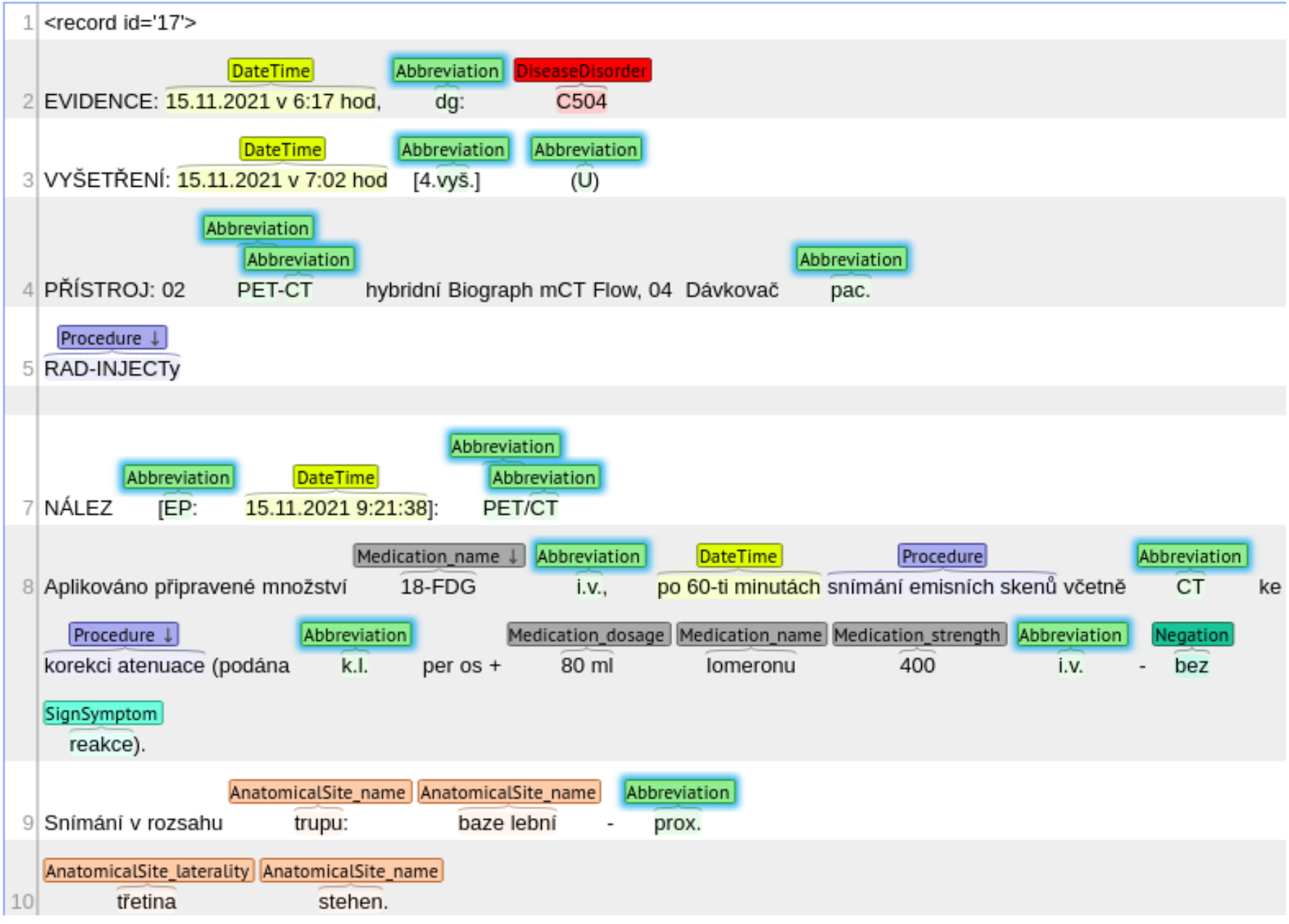
Introduction: Motivation

Key motivation of our research

Training the best possible medical LLMs for medical language understanding and other downstream medical application tasks.

- We
  - curate electronic health record (EHR) datasets
  - curate human annotation of medical concepts and develop machine annotation tools
  - train classifier models
- Having quality language data for training is essential




<record id='17'>
EVIDENCE: 15.11.2021 v 6:17 hod, dg: C504
VYŠETŘENÍ: 15.11.2021 v 7:02 hod [4.vyš.] (U)
PŘÍSTROJ: 02 PET-CT hybridní Biograph mCT Flow, 04 Dávkovač pac.
RAD-INJECTy
NÁLEZ [EP: 15.11.2021 9:21:38]: PET/CT
Aplikováno připravené množství 18-FDG i.v., po 60-ti minutách snímání emisních skenů včetně CT ke
korekci atenuace (podána k.l. per os + 80 ml Iomeronu 400 i.v. - bez
reakce).
Snímání v rozsahu trupu: baze lební - prox.
třetina stehen.

Introduction: EHRs and redundancy

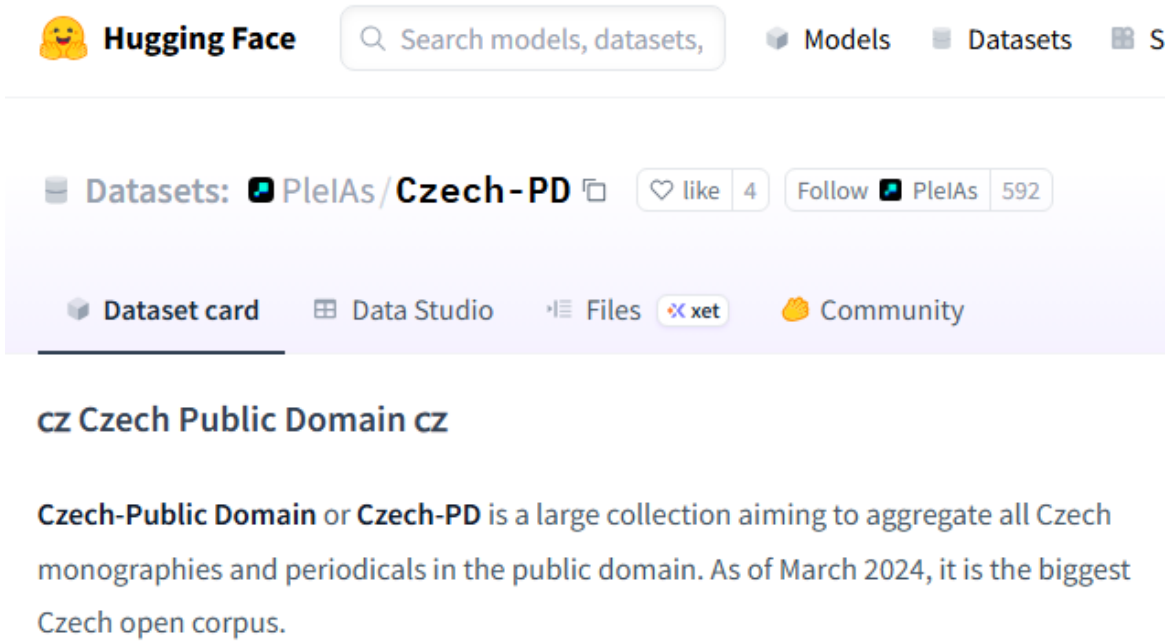
- Electronic health records (EHRs) contain extensive repetition arising from
  - templated structures
  - copy-paste practices
  - recurrent clinical phrasing
- This affects
  - the efficiency of data processing
  - downstream natural language processing applications

Introduction: Our EHR data

- CSEHR_corpus
  - 42 million words of oncology EHRs
  - 154,000+ individual records
  - Collected at Masaryk Memorial Cancer Institute in Brno
  - Substantial export reflecting very accurately the various practices and types of contemporary health documentation at a real medical institute

“Normal” language to compare: Czech-PD [5]

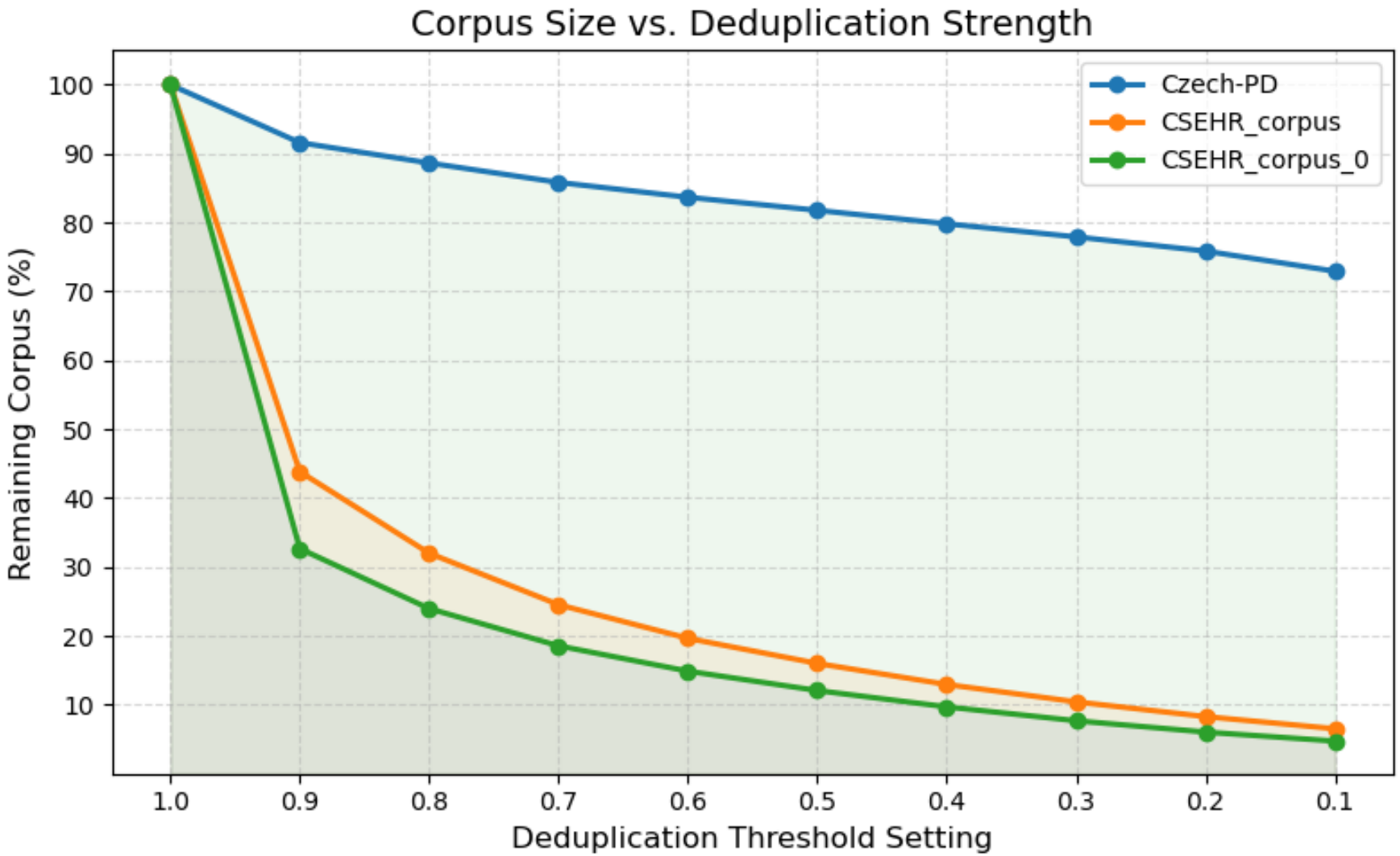
- Czech literary dataset
- OCR of public domain books
- 259 million words
- not deduplicated



Czech-PD example

Moravské národní POHÁDKY A POVĚSTI z okolí Htožnovského.
“Sbírka tato nechce podávati toliko myšlenky z národních pohádek a pověstí obrazností spisovatelek a spisovatelů nynějších rozmnožené a ozdobené, nebrž pouze obraznost národa samého, jak se ode dávna v pohádkách a pověstech jeví. Těch původních, nezměněných výtvorů ducha národního jest nám v celé rozpáhlosti a hlubokosti poznati třeba, abychom se přinčovali ryzému jazyku svému my, kteřížto jsme namnoze jazyky cizími se odnárodnili. A proto jsou slova a celé věty ryzostí a krásou mluvnickou vynikající a překvapující vyznačeny rozložitým písmem, aby neušly oku čtenářovu. Kterak a k čemu jich zužitkují spisovatelé naši, k tomu ovšem zřetele svého neobracujeme, ale přejeme, aby jich zužitkovali ve spisích svých ve prospěch národa. Svrchovaný jest čas, aby se tyto utěšené voničky či kytečky po zahrádkách lidu našeho sbíraly a tiskem jistému zapomenutí vyrvaly, jelikož nastala rozhodná doba přerody a obnovy národu uašemu, kdy tiskem a parou novější myšlenky do lidu našeho se hrnou…”

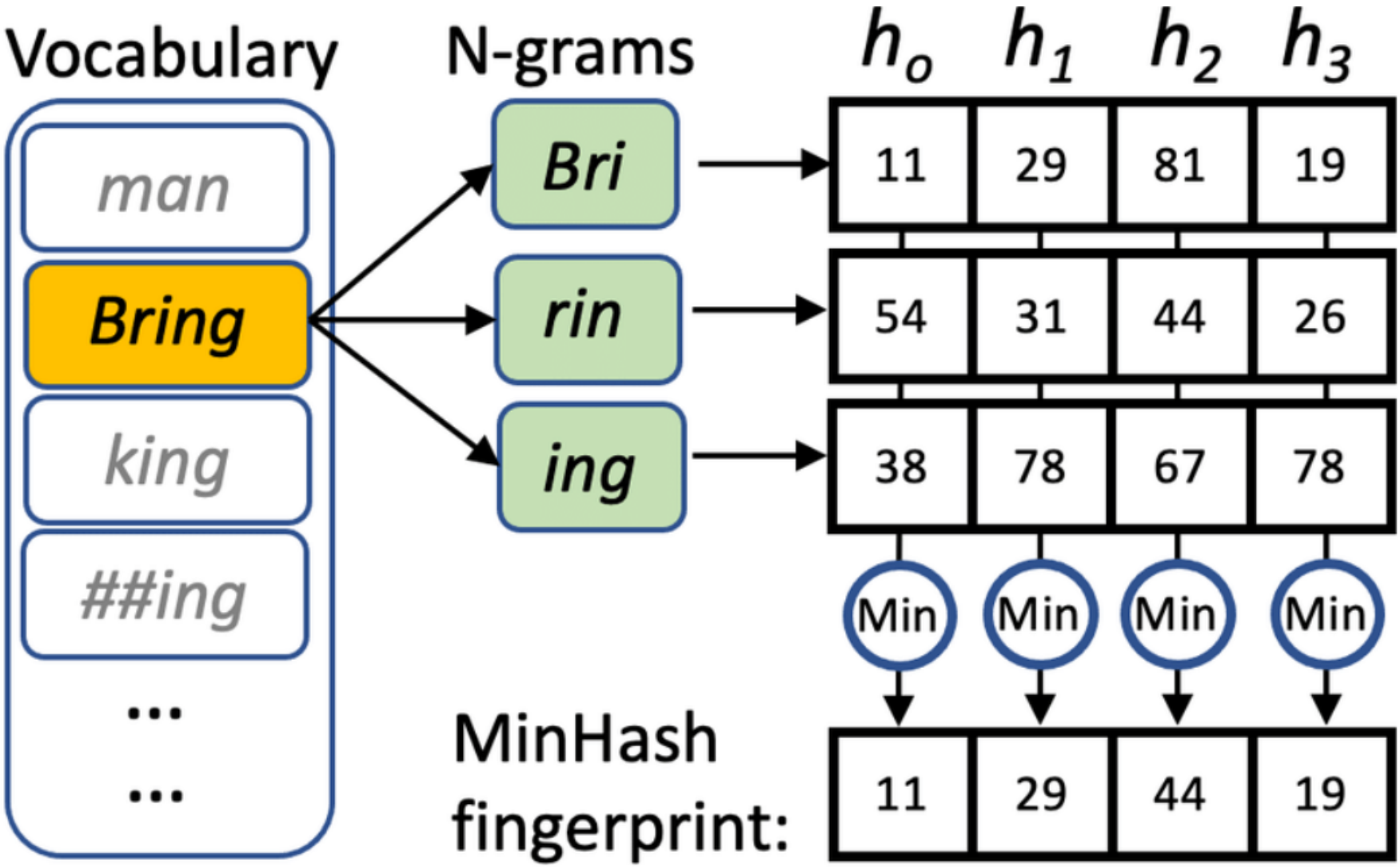
Deduplication with Onion: EHRs vs. Czech-PD



Clustering

- Clustering based on
  - MinHash [1]
    - approximation of the Jaccard similarity between two sets of tokens or shingles by representing each document through a compact signature of hash minima
  - Locality-Sensitive Hashing (LSH) [3]
    - partitioning of the signature space into hash buckets such that similar documents are likely to fall into the same bucket

MinHash illustration from Fusco et al. [2]



Clustering: examples

```
--- Cluster 24 (3936 items) ---
[1] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [4.vyš.]  (U) EXAM TYPE: Břicho (posl. [[dat
on RS80A (2017)  NÁLEZ [EP: [[date]] [[time]]]...
[2] EVIDENCE: [[date]] v [[time]], dg: D059       VYŠETŘENÍ: [[date]] v [[time]] [2.vyš.]  (U) EXAM TYPE: Břicho PŘÍSTROJ: UZ
NÁLEZ [EP: [[date]] [[time]]]:  USG břicha  Já...
[3] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [7.vyš.]  (U) EXAM TYPE: Břicho (posl. [[dat
on RS80A (2016)  NÁLEZ [EP: [[date]] [[time]]]...
[4] EVIDENCE: [[date]] v [[time]], dg: C435       VYŠETŘENÍ: [[date]] v [[time]] [1.vyš.]  (U) EXAM TYPE: Břicho PŘÍSTROJ: UZ
  NÁLEZ [EP: [[date]] [[time]]]:     USG břich...
[5] EVIDENCE: [[date]] v [[time]], dg: C509       VYŠETŘENÍ: [[date]] v [[time]] [2.vyš.]  (U) EXAM TYPE: Břicho PŘÍSTROJ: UZ
  NÁLEZ [EP: [[date]] [[time]]]:  UZ jater + D...

--- Cluster 31 (3476 items) ---
[1] Předoperační anesteziologické vyšetření   Zákl. Dg.: LSK pro susp tumor ovaria Osobní anamn.: fumator, stp peritonitidě, s
y,  stp. resekci ovarii l. sin  - dermoid  201...
[2] Předoperační anesteziologické vyšetření  Zákl. Dg.: CA mammae l. dx. Osobní anamn.:fumator 10/d.,spirometrie - bez ventil
Operace:plastika močovodu,vazy kolene,CA bez k...
[3] Předoperační anesteziologické vyšetření  Zákl. Dg.:ca mammae dxt - [[date]] Mastectomia partialis l.dx., SNB axillae dext
mn.:migreny-opak hospit na neurologii, depresi...
[4] Předoperační anesteziologické vyšetření  Zákl. Dg.: ca mammae sin Osobní anamn.: fumator 5/d , hypertenze
k Farmakol.anamn.: Tritace, Amloratio Alergie:...
[5] Předoperační anesteziologické vyšetření  Zákl. Dg.: Ca mammae l.dx- stp. PME Osobní anamn.: VAS s vyhřezlou ploténkou v a
V heterozygot Operace: gynekologická GEU, salp...

--- Cluster 91 (3365 items) ---
[1]  Her-2 imunohistochemicky: Závěr: HER2 negativní (1+). Popis: Z parafinového materiálu byla provedena imunohistochemická
monoklonální protilátkou proti HER2 proteinu (...
[2]  Her-2 imunohistochemicky: Závěr: HER2 negativní (0) Popis: Z parafinového materiálu byla provedena imunohistochemická re
ilátkou proti HER2 proteinu (VENTANA HER2(4B5)...
[3]  Her-2 imunohistochemicky: Závěr: Výsledek IMH pomocí certifikovaného kitu: nebyla nalezena zvýšená exprese HER2 proteinu
riálu byla pomocí Ventana kitu provedena imuno...
[4]  Her-2 imunohistochemicky: Závěr: HER2  negativní 1+ Popis: Z parafinového materiálu byla provedena imunohistochemická re
noklonální protilátkou proti HER2 proteinu (VE...
[5]  Her-2 imunohistochemicky: Závěr: Nádor je Her-2 negativní (hodnoceno dle ASCO/CAP). Popis: Z parafinového materiálu byla
kce s certifikovanou primární monoklonální pro...
```

Clustering: examples

```
--- Cluster 4 (6623 items) ---
[1] EVIDENCE: [[date]] v [[time]], dg: Z123       VYŠETŘENÍ: [[date]] v [[time]] [7.vyš.]
[2] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [4.vyš.]
[3] EVIDENCE: [[date]] v [[time]], dg: C502       VYŠETŘENÍ: [[date]] v [[time]] [7.vyš.]
[4] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[5] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[6] EVIDENCE: [[date]] v [[time]], dg: C509       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[7] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [1.vyš.]
[8] EVIDENCE: [[date]] v [[time]], dg: C505       VYŠETŘENÍ: [[date]] v [[time]] [4.vyš.]
[9] EVIDENCE: [[date]] v [[time]], dg: D051       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[10] EVIDENCE: [[date]] v [[time]], dg: C504      VYŠETŘENÍ: [[date]] v [[time]] [7.vyš.]

--- Cluster 23 (5983 items) ---
[1] EVIDENCE: [[date]] v [[time]], dg: C503   OMKOCH    VYŠETŘENÍ: [[date]] v [[time]] [14.
[2] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [6.vyš.]
[3] EVIDENCE: [[date]] v [[time]], dg: C502       VYŠETŘENÍ: [[date]] v [[time]] [4.vyš.]
[4] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[5] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [1.vyš.]
[6] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [1.vyš.]
[7] EVIDENCE: [[date]] v [[time]], dg: C504       VYŠETŘENÍ: [[date]] v [[time]] [3.vyš.]
[8] EVIDENCE: [[date]] v [[time]], dg: C509   BCH    VYŠETŘENÍ: [[date]] v [[time]] [2.vyš.
[9] EVIDENCE: [[date]] v [[time]], dg: C509       VYŠETŘENÍ: [[date]] v [[time]] [4.vyš.]
[10] EVIDENCE: [[date]] v [[time]], dg: C504      VYŠETŘENÍ: [[date]] v [[time]] [1.vyš.]
```

Clustering: examples

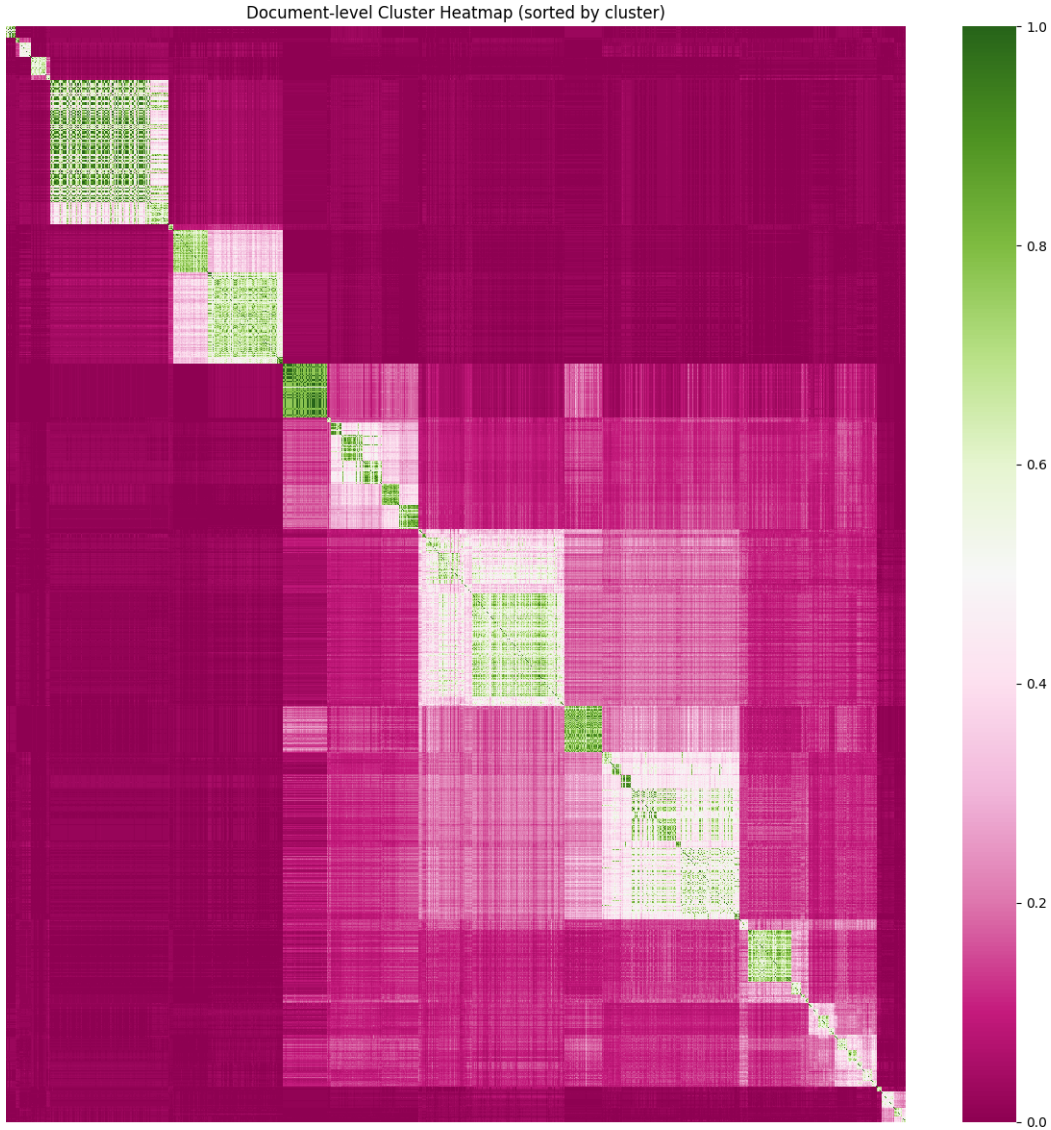
```
--- Cluster 31 (3476 items) ---
[1] Předoperační anesteziologické vyšetření   Zákl. Dg.: LSK pro susp tumor ovaria Osc
[2] Předoperační anesteziologické vyšetření  Zákl. Dg.: CA mammae l. dx. Osobní anamn.
[3] Předoperační anesteziologické vyšetření  Zákl. Dg.:ca mammae dxt - [[date]] Mastec
[4] Předoperační anesteziologické vyšetření  Zákl. Dg.: ca mammae sin Osobní anamn.: f
[5] Předoperační anesteziologické vyšetření  Zákl. Dg.: Ca mammae l.dx- stp. PME Osobr
[6] Předoperační anesteziologické vyšetření  Zákl. Dg.: Předoperační anesteziologické
[7] Předoperační anesteziologické vyšetření  Zákl. Dg.:  karcinom levé mamy Osobní ana
[8] Předoperační anesteziologické vyšetření  Zákl. Dg.Ca mammae l.sin.  Výkon: PME+ SN
[9] Předoperační anesteziologické vyšetření  Zákl. Dg.: Ca mammae l. sin. Osobní anamn
[10] Předoperační anesteziologické vyšetření  Zákl. Dg.: Ca mammae sin. Osobní anamn.:

--- Cluster 91 (3365 items) ---
[1]  Her-2 imunohistochemicky: Závěr: HER2 negativní (1+). Popis: Z parafinového mater
[2]  Her-2 imunohistochemicky: Závěr: HER2 negativní (0) Popis: Z parafinového materiá
[3]  Her-2 imunohistochemicky: Závěr: Výsledek IMH pomocí certifikovaného kitu: nebyla
[4]  Her-2 imunohistochemicky: Závěr: HER2  negativní 1+ Popis: Z parafinového materiá
[5]  Her-2 imunohistochemicky: Závěr: Nádor je Her-2 negativní (hodnoceno dle ASCO/CAP
[6]  Her-2 imunohistochemicky: Závěr: HER2 negativní 0 Popis: Z parafinového materiálu
[7]  Her-2 imunohistochemicky: Závěr: Nádor je Her-2 negativní (hodnoceno dle ASCO/CAP
[8]  Her-2 imunohistochemicky: Závěr: HER2 negativní (1+). Popis: Z parafinového mater
[9]  Her-2 imunohistochemicky: Závěr: Nádor je Her-2 negativní (hodnoceno dle ASCO/CAP
[10]  Her-2 imunohistochemicky: Závěr: HER2 nejednoznačné (2+) Popis: Z parafinového m
```

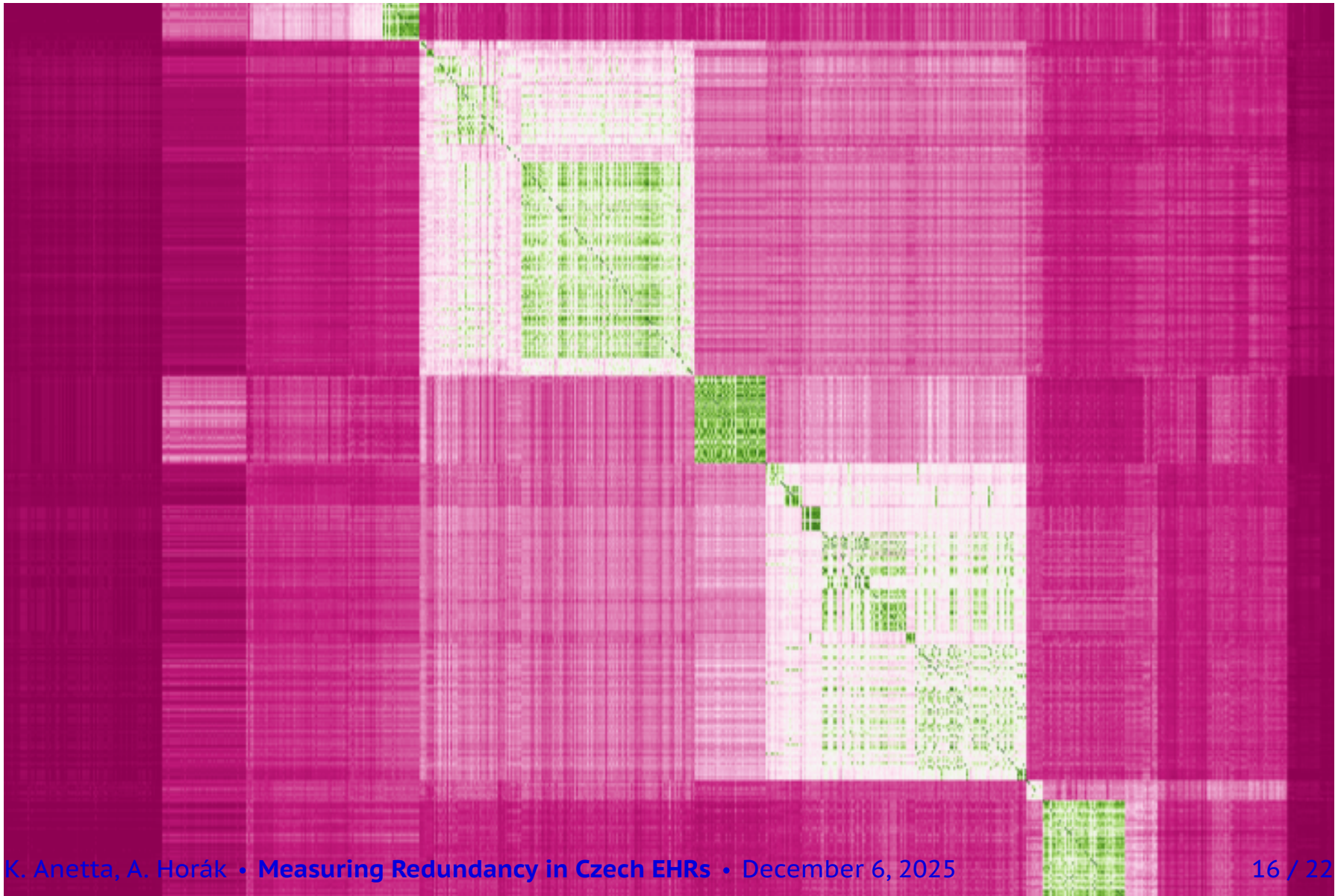
Visualizing superclusters

- In order to make the similarities between individual documents as palpable as possible, we
  - ran a greedy clustering algorithm with a Jaccard threshold of 0.5
  - extracted all clusters with more than 100 documents, totaling 58 clusters containing 27,039 of all 154,763 documents (17.47%)
  - arranged the clusters based on cluster-to-cluster centroid similarities
  - and, reducing clusters to ca. 10% of their respective sizes, plotted the 2,681 x 2,681 document similarity heatmap

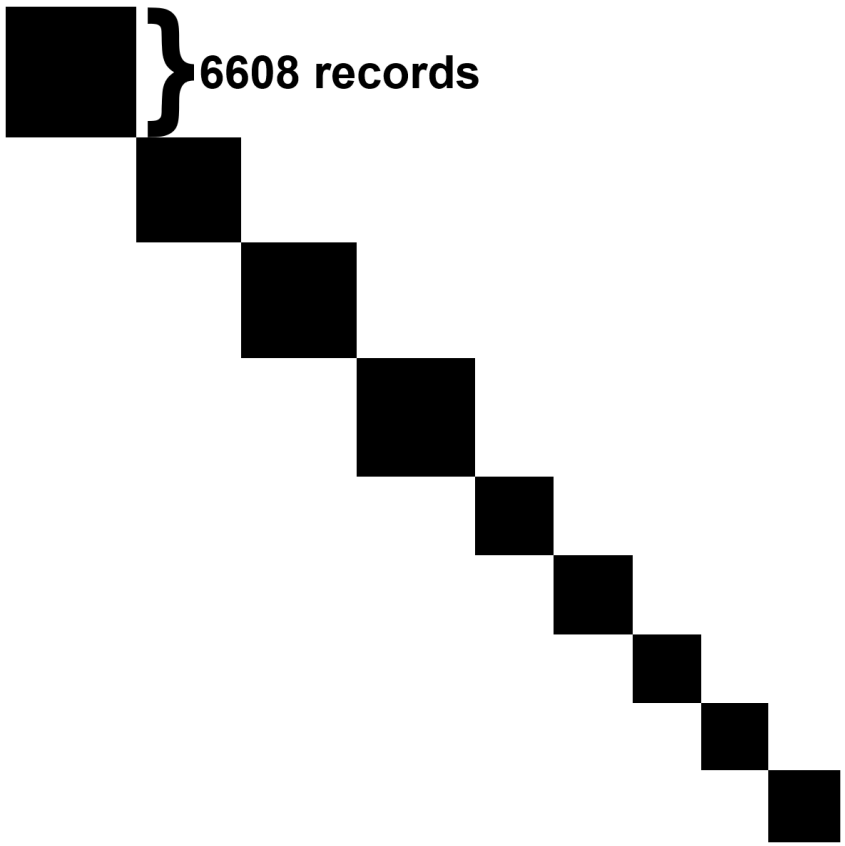
Visualizing superclusters (clusters with >100 docs)



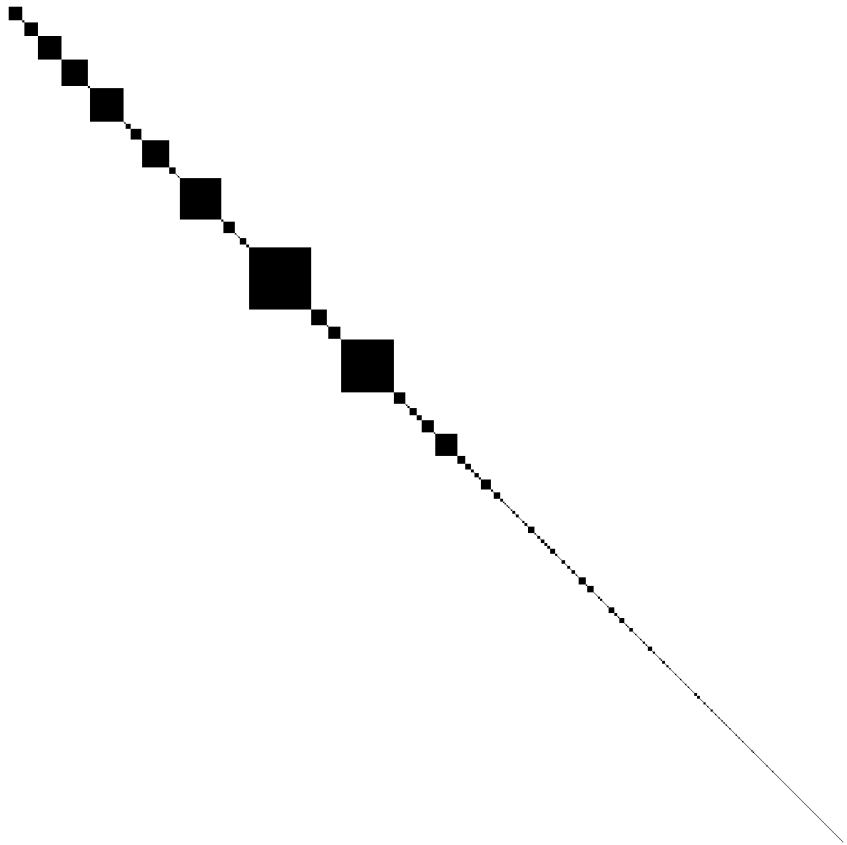
Visualizing superclusters: Close-up



Clustering granularity (25% most clustered)



(a) threshold 0.2



(b) threshold 0.5

Cluster-aware dataset pruning

- Keeping only 1 record from each cluster:
  - CSEHR_corpus_cluster50: after clustering with threshold 0.5, reduced to 82,059 documents and 56.7% of original size
  - CSEHR_corpus_cluster20: after clustering with threshold 0.2, reduced to 17,301 documents and 8.3% of original size.

Discussion and conclusion

- The deep patterns of redundancy in EHR datasets:
  - suggest that well-planned processing is necessary to make diverse, balanced, and efficient training data
  - provide further evidence in favor of using synthetic data for training (as in the case of GatorTronS [4])
- Next steps for our research:
  - Maximize the classification score using the available 42M dataset

Bibliography I

[1] Andrei Z Broder. “On the resemblance and containment of documents”. In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE. 1997, pp. 21–29.

[2] Francesco Fusco et al. “pNLP-mixer: An efficient all-MLP architecture for language”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 2023, pp. 53–60.

[3] Piotr Indyk and Rajeev Motwani. “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 1998, pp. 604–613.

Bibliography II

[4] Cheng Peng et al. “A study of generative large language model for medical research and healthcare”. In: *NPJ digital medicine* 6.1 (2023), p. 210.

[5] PleIAs. *Czech-PD: Czech Public Domain Corpus*. `https://huggingface.co/datasets/PleIAs/Czech-PD`. Accessed: 2025-11-04. 2024.

Thank you for your attention!

MUNI
FACULTY
OF INFORMATICS