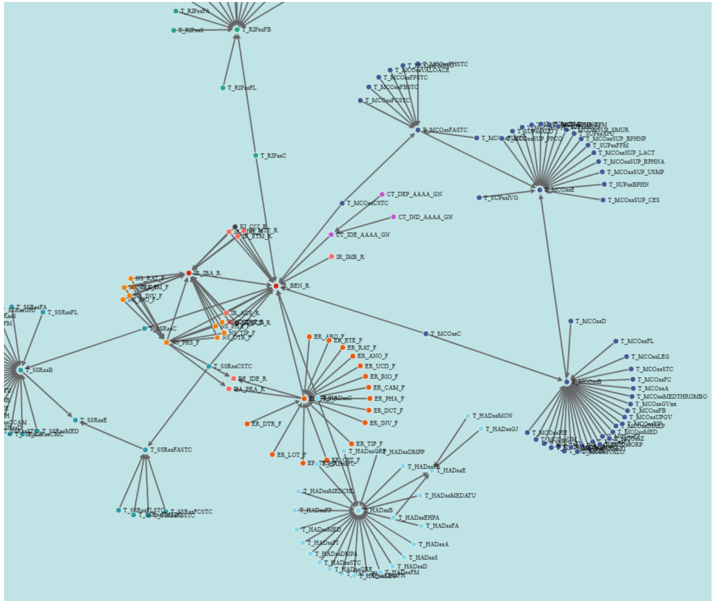
Event2vec, a python package for medical concept embeddings study

Initié avec:
Aude Leduc
Dinh Phong Nguyen
Albert Vuagnat

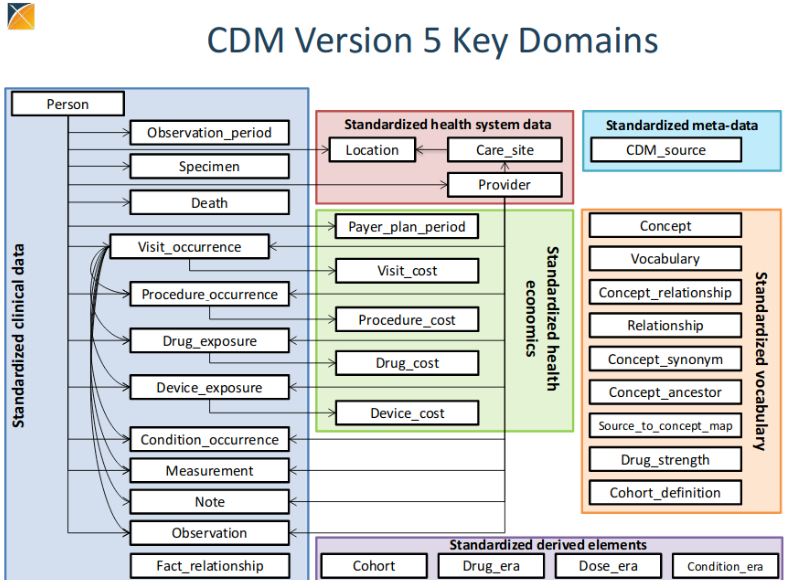
2023-03-30

Poursuivi avec:
Matthieu Doutreligne
Gaël Varoquaux (*superviser)*
Antoine Neuraz

Large observational structured databases

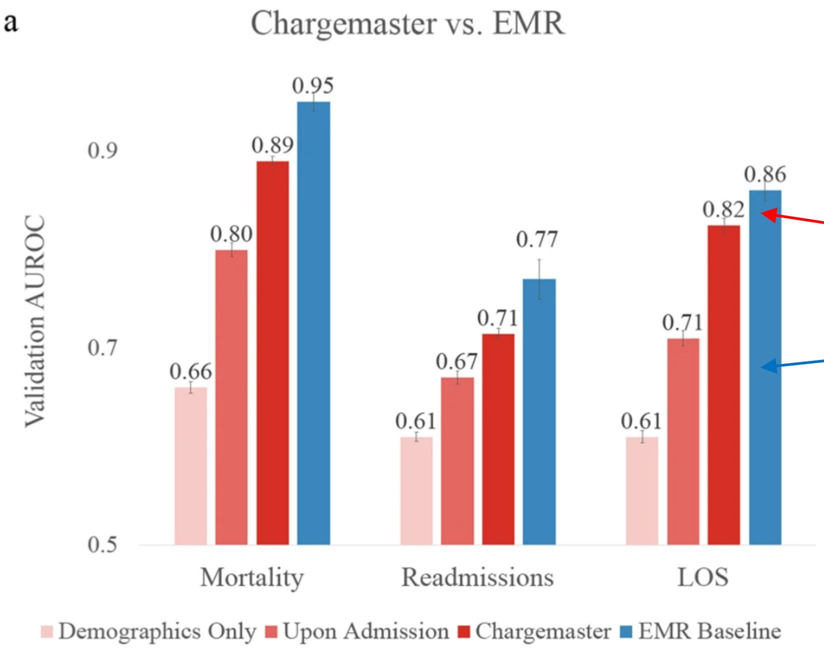


Medico-administrative database (claims) :
ex. SNDS
Care consumptions, reimbursements



Electronic Health Records (EHR/EMR):
ex. APHP data warehouse
Detailed clinical variables, medical reports, …

Despite the lack of precise endpoints, claims contain information



Compromise

Claims : Lot of patients (N >> 1)

vs

Cliniques : Lot of variables (D >> 1)

(Beaulieu-jones et al, 2021)

Performances of predictive models taking as inputs claims (chargemaster) or Electronic Medical Records (EMR)

Patient trajectories : timestamped collection of tokens

Multiple applications of ML in healthcare consider a **triplet event format**

(Rajkomar et al., 2018; Beam et al., 2019; Bacry et al., 2020; Chazard et al., 2022)

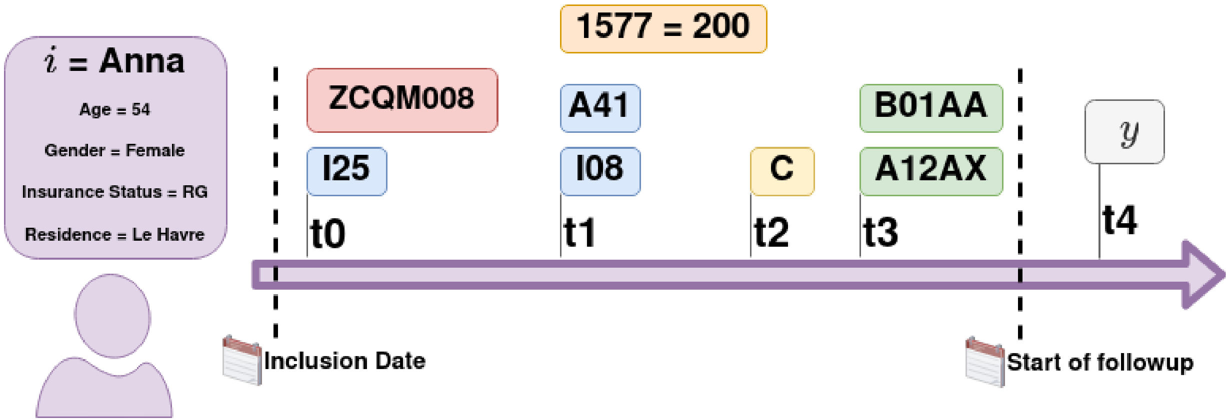
Advantages

- **Simple**
- **Sequential**
- **Comparability** of all type of healthcare information

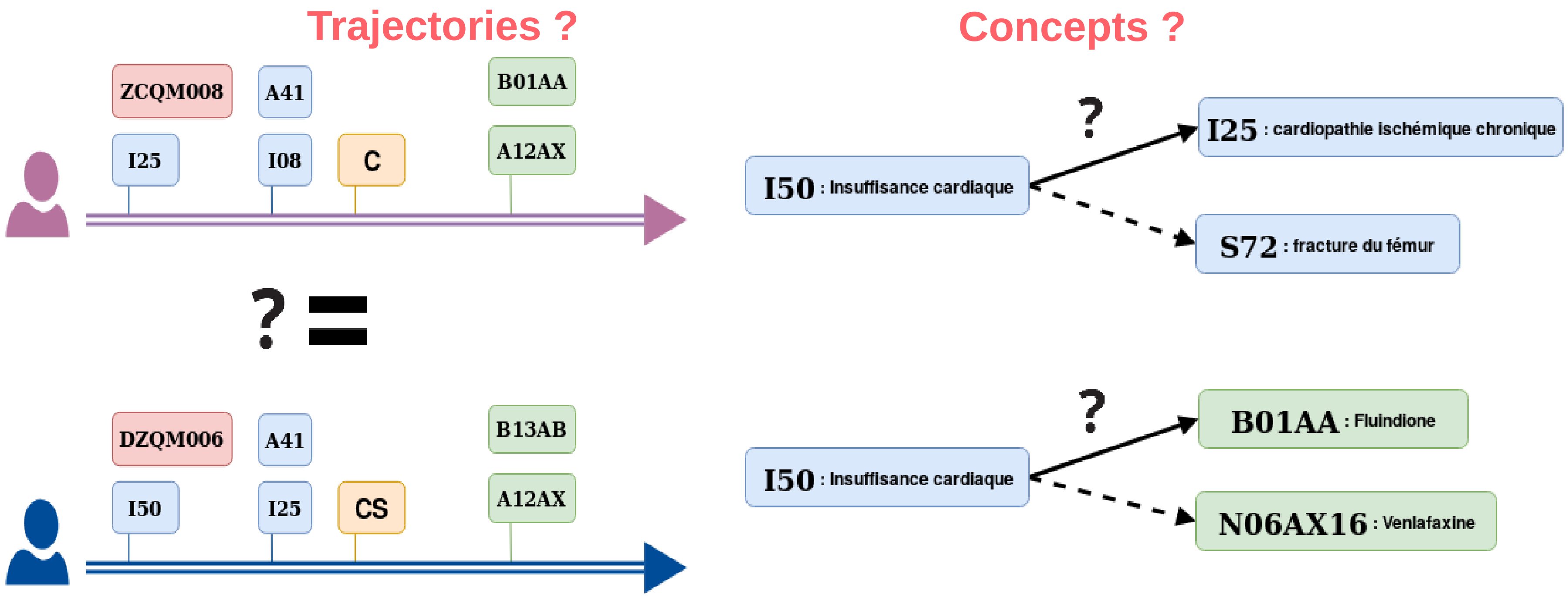
$$e = (i, t, c)$$

Difficulties

- **High cardinality** of codes
- Choices of **aggregation for statistical models**



Patient trajectories: How to derive proximity of



I. Context and motivations

II. Medical concept embeddings from structured events

III. Qualitative results

IV. Empirical evaluations 🚧

Medical embeddings of structured data, previous work

First concept representations algorithms
- **Tran et al., 2015:** nonnegative restricted bolzmann machines for suicide-prediction models
- **Miotto et al., 2016, deepatient:** Auto Encoder for 78 disease onsets prediction
- **Nguyen et al. 2016, deepr:** CNN for deep patient representation and unplanned readmission
- **Choi et al., 2016, med2vec:** MLP for visits and medical codes, for next visit billing codes prediction

Inclusion of time
- **Cai et al., 2019, CBOWA:** Build a time-aware context window, evaluate on clustering tasks
- **Beam et al., 2019, cui2vec:** Implement context aware svd-ppmi, evaluate on known associations detection
- **Xiang et al., 2019:** extend Beam's algorithm to fastText, applied to onset prediction of heart failure (w. LSTM)

Transformer-based models
- Rasmy et al., 2021, MedBert: Transformers for heart failure for diabetes patients (DHF) and pancreatic cancer prediction
- Solares et al., 2020, BEHRT: Transformers for 301 diseases predictions in future visits

A review paper with benchmarks

Solares et al., 2021, Transfer Learning in Electronic Health Records through Clinical Concept Embedding

Inspiration: back to basics: word2vec in NLP

Distributional hypothesis *(Firth, 1957)*: **Two words are close iif they appear in similar contexts:**
“You shall know a word by the company it keeps”

The **queen sits** on the **throne** and discusses with the king the problems of the ~~kingdom~~.

window = 2 x 5 words

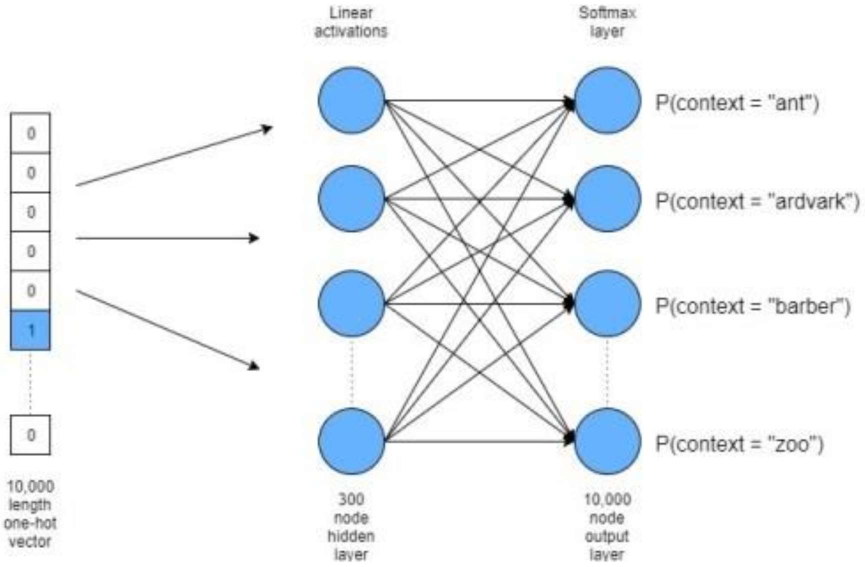
Proximity in the embedding space is forced by **proximity in the corpus**.

Focus on a context window approach

- **SGNS (word2vec):** Prediction of the context given a word thanks to a one-layer neural network

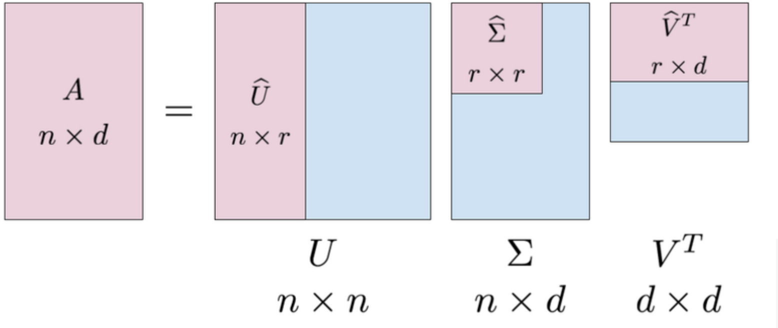
$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

positive example **negative example**

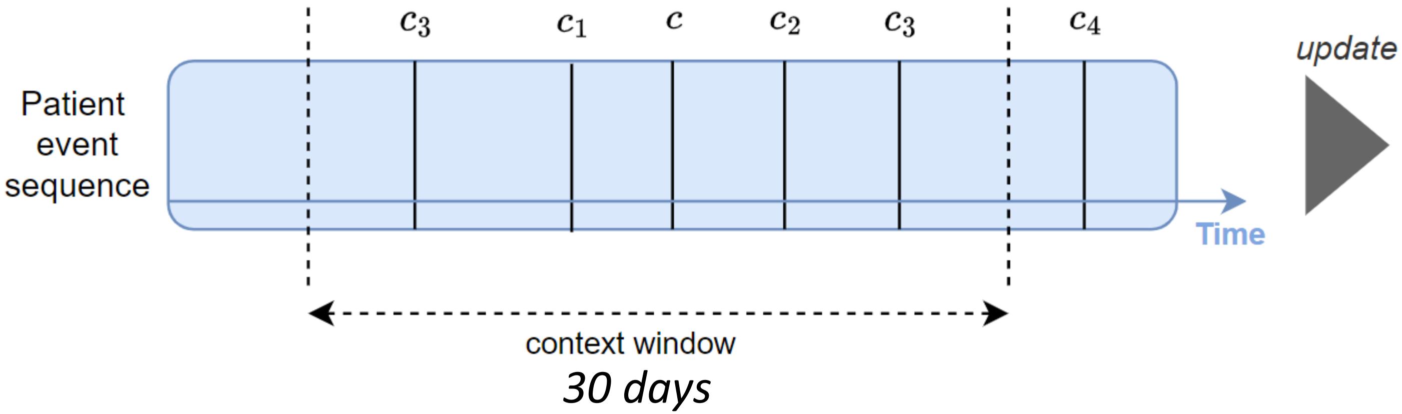


- **SVD-PPMI:** Singular vector decomposition of the transformed word co-occurrence matrix

$$\Phi(w) = \frac{1}{2}(U_d \cdot \sqrt{\Sigma_d} + (V_d \cdot \sqrt{\Sigma_d})$$



Adapting word2vec to patient trajectory *(Beam et al., 2019)*



$P(c, c_1) += 1$
$P(c, c_2) += 1$
$P(c, c_3) += 2$
$P(c, c_1) += 0$

Build a time dependant context for the co-occurrence matrix $P(c_i, c_j)$

Why concept embeddings could be interesting ?

Objectives

- Predictive and interpolation models *(cf. preceding review slide)*
- Treatment effects estimation thanks to G-formula *(Dorie et al., 2018, Wendling et al., 2018)*
- Vocabulary matching

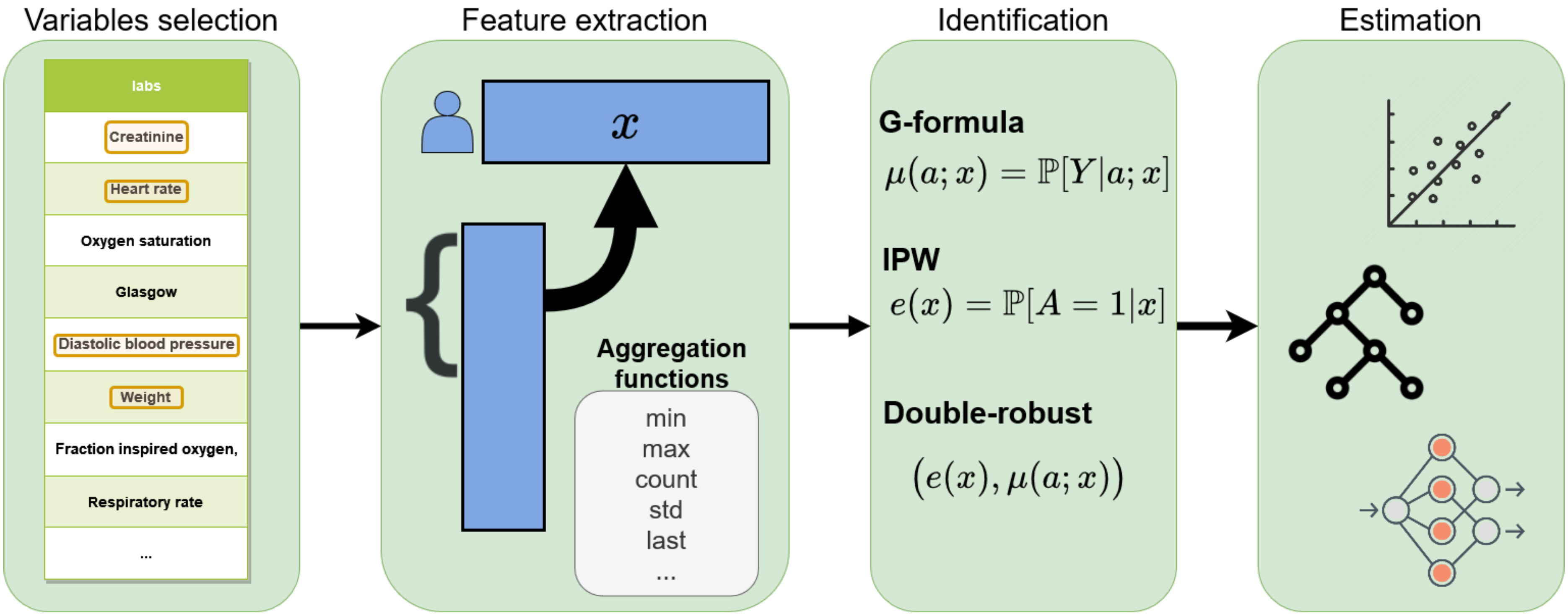
Advantages

- **Sharable** aggregated information
- Fewer **Hyper-parameter tuning**
- **Simple implementation** pandas + scipy
- **CPU only** easilly scalable w. distributed backend
- **No softmax** computation

Difficulties

- **Poor in-context** comprehension
- Different **choices of aggregation** for visit modelizations

Des choix multiples de modélisation



I. Context and motivations

II. Medical concept embeddings from structured events

III. Demo and qualitative results

IV. Empirical evaluations 🚧

Event2vec, a package to easily compute concept embeddings

A python package available on pypi

A pyspark version for big data (>500m rows)

Quick start and step by step guides: https://straymat.gitlab.io/event2vec/tutorials/_0_tuto_event2vec.html

Load events

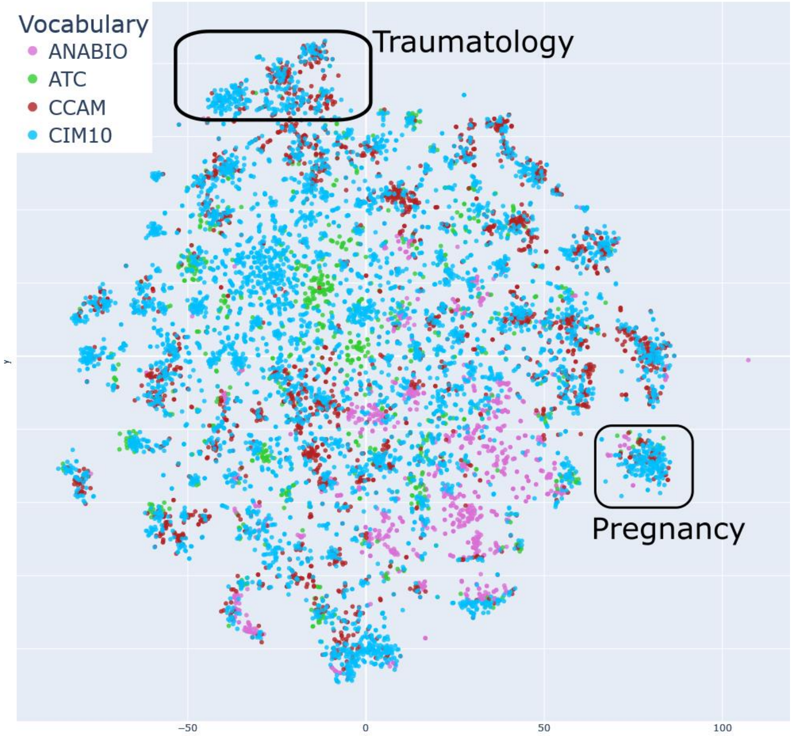
	person_id	start	event_source_concept_id
0	1	2018-11-08 19:24:15	CIM10:N182
4	1	2018-12-20 19:24:15	CCAM:JVJB01
8	2	1993-01-26 07:22:42	CIM10:E12
12	3	2009-04-25 10:14:21	CIM10:N182
9	2	2020-01-26 07:22:42	CIM10:E12

Build embeddings

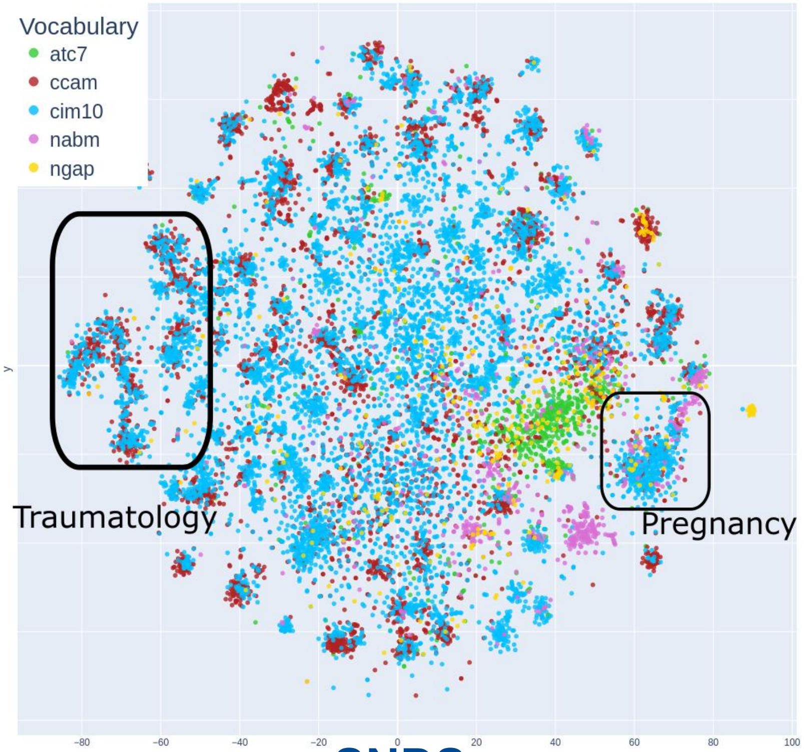
```python
alpha = 0.75
k = 1
d = 3

embeddings = event2vec(
    events=events,
    output_dir=output_dir,
    colname_concept="event_source_concept_id",
    window_orientation="center",
    window_radius_in_days=30,
    d=d,
    smoothing_factor=alpha,
    k=k,
    backend="pandas",
)
```

Qualitative results: https://straymat.gitlab.io/event2vec/visualizations.html



APHP
(200K random patients)

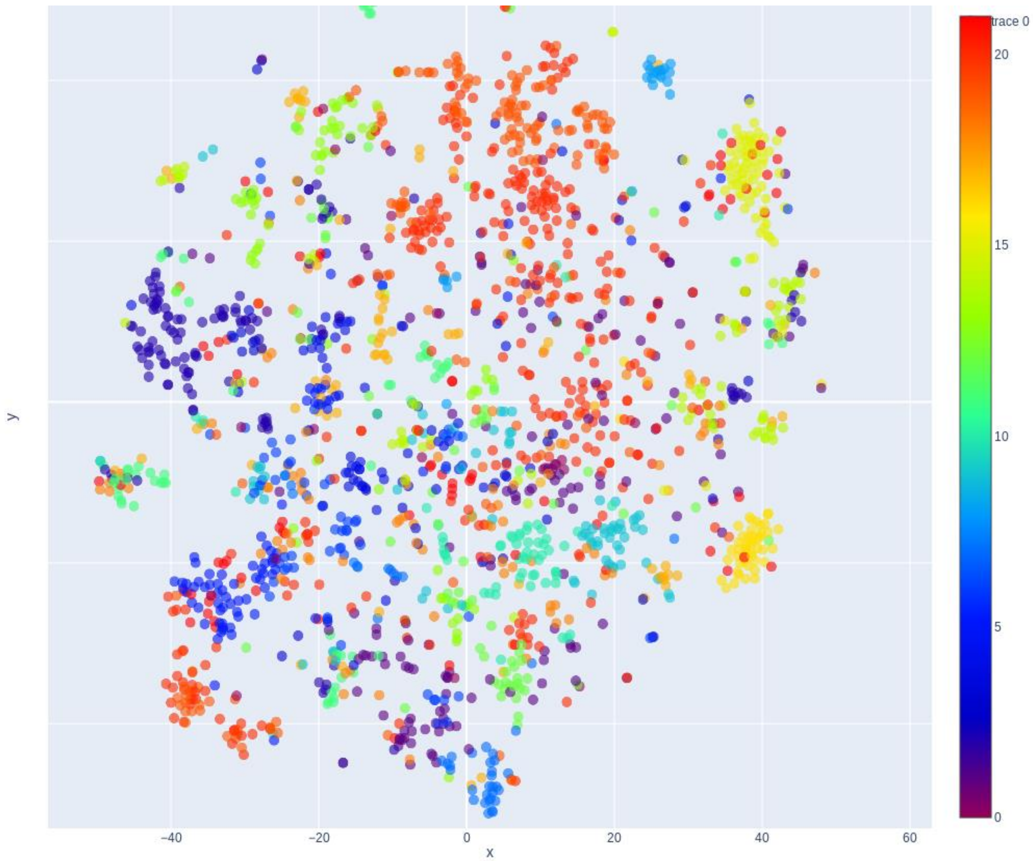


SNDS
(3M random patients)

Qualitative results, Hierarchy reconstruction

CIM10 billing diagnoses
Third level, r=30 jours

Colored by chapter



Extrinsic evaluation: Compare different models on a downstream task

Task: rehospitalization at 30 days, for plannification and outcome modeling (g-estimation)

Models = (featurizer, estimator):

Compared featurizers:

Count vectorizer (+SVD, D=30)

$[C, C_{decay}]$

Embeddings fit on train data

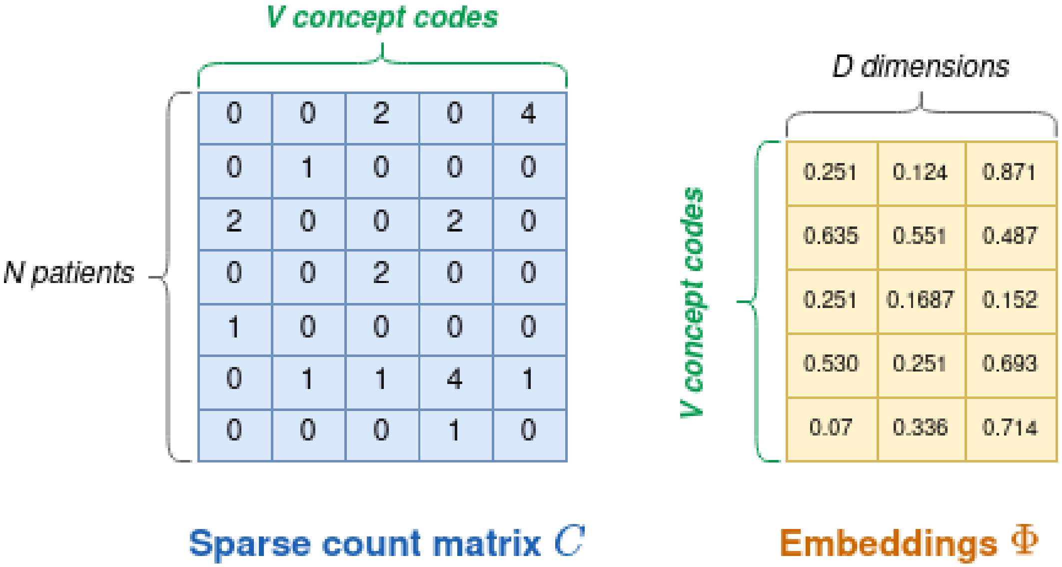
$[C \cdot \Phi_{train}, C_{decay} \cdot \Phi_{train}]$

SNDS Embeddings (+SVD, D=30)

$[C \cdot \Phi_{SNDS}, C_{decay} \cdot \Phi_{SNDS}]$

Compared estimators:

Random forests, Ridge classifier



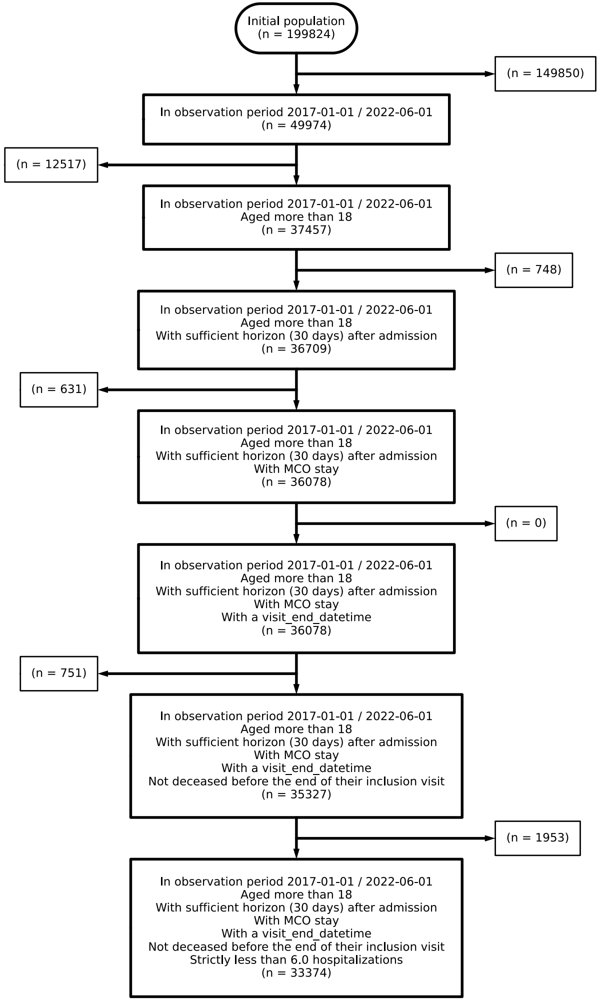
Population selection for prediction

Extraction of **200,000 random patients** from APHP EDS
With **complete hospitalization**
Study period 2017-2022 (stable Information System)
Sufficient horizon for followup (no right censure)
Exclusions: No children, not decesead during hospitalization

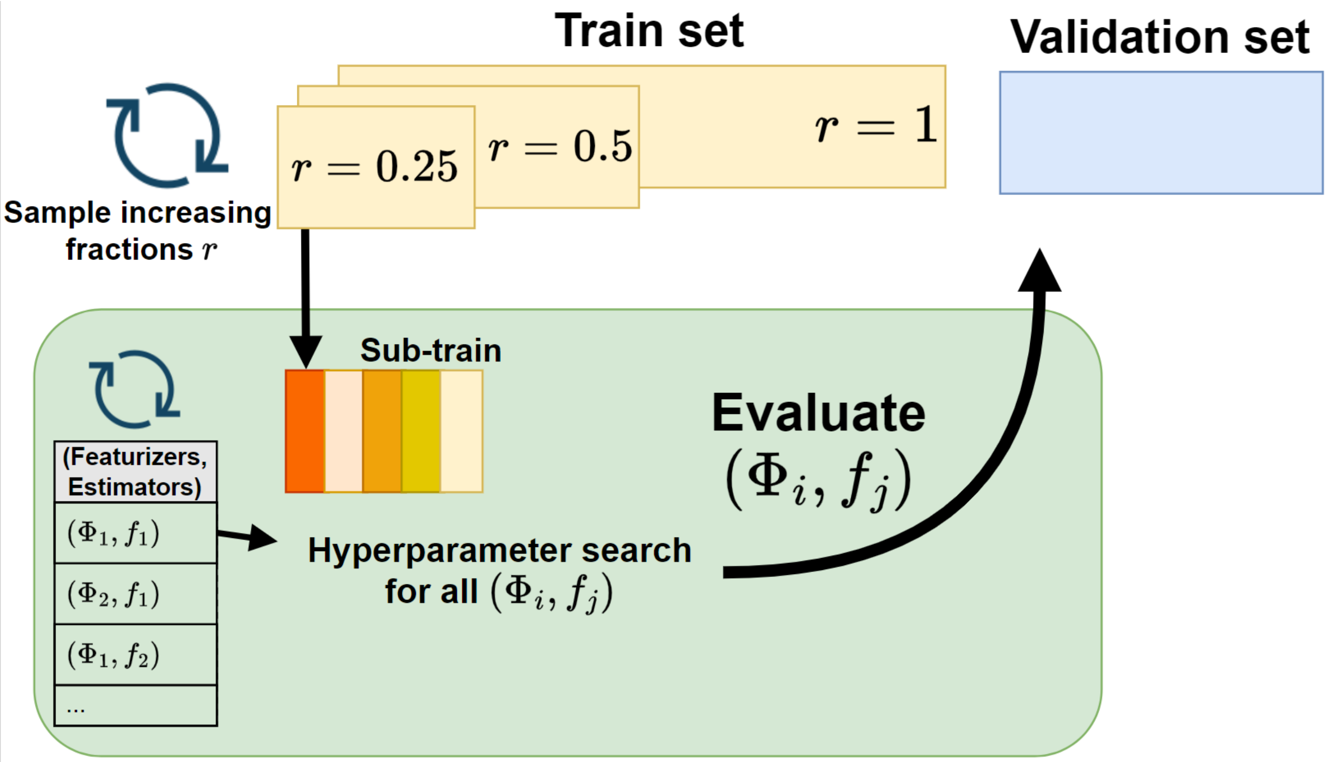
At least one event in **4734 codes occurring** at least 10 times:

- **drug exposure administrations:** 663, 027 events
- **procedure occurrences:** 222, 770 events
- **condition occurrences:** 203, 779 events

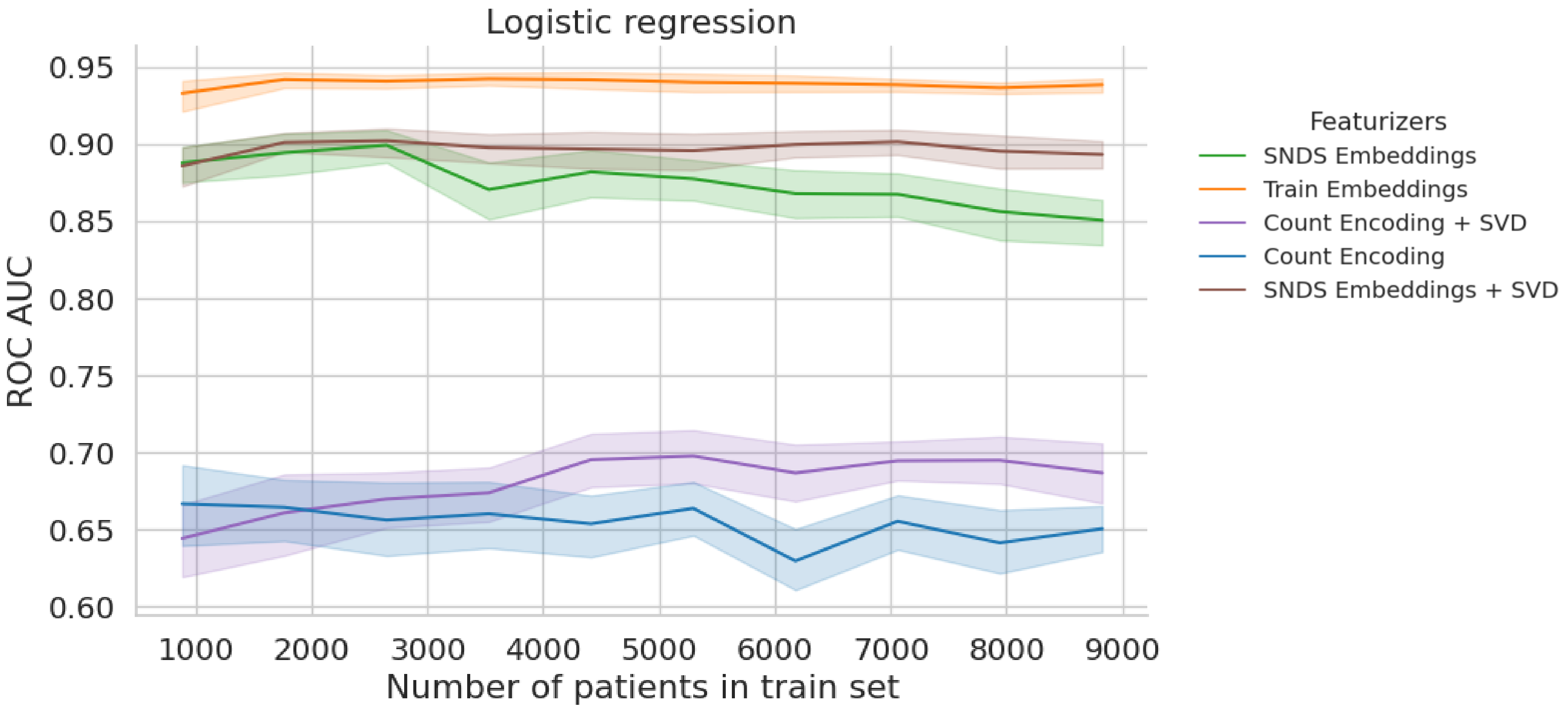
25, 063 patients: mean age = 54.4, female ratio = 54.1%, mean LOS>7 days: 20.80%



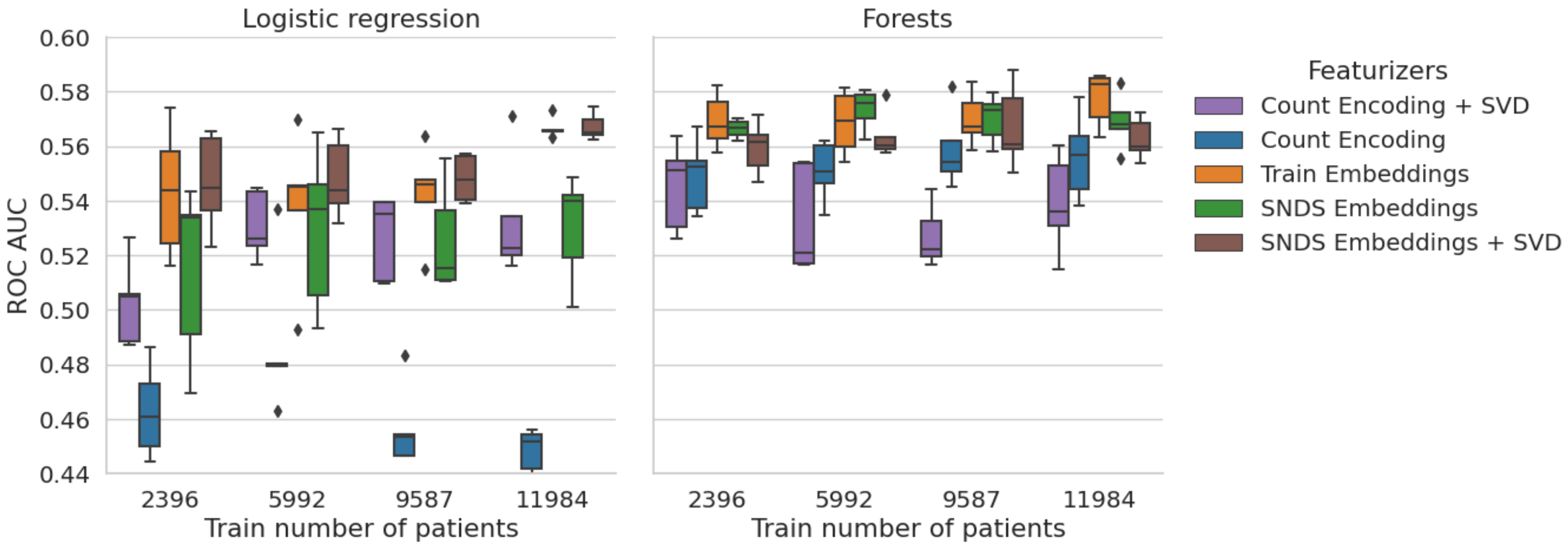
Selection procedure



Task: Length Of Stay interpolation : <=7 days vs. >7 days



Current results for rehospitalization or death @ 30 days



Overall, **performances are low** (too difficult task ?, badly defined ?)

Logistic regression: **in-domain embeddings are equivalent to SNDS embeddings**

Forests smooth these differences (leverages better the missing value mask ?)

Further work

- Study **transfer capabilities inside APHP**

- Study **transfer capibilities with international embeddings** such as cui2vec

- Perturbe the learning by dropping some codes

- An **intermediate task** to study predictive performance:

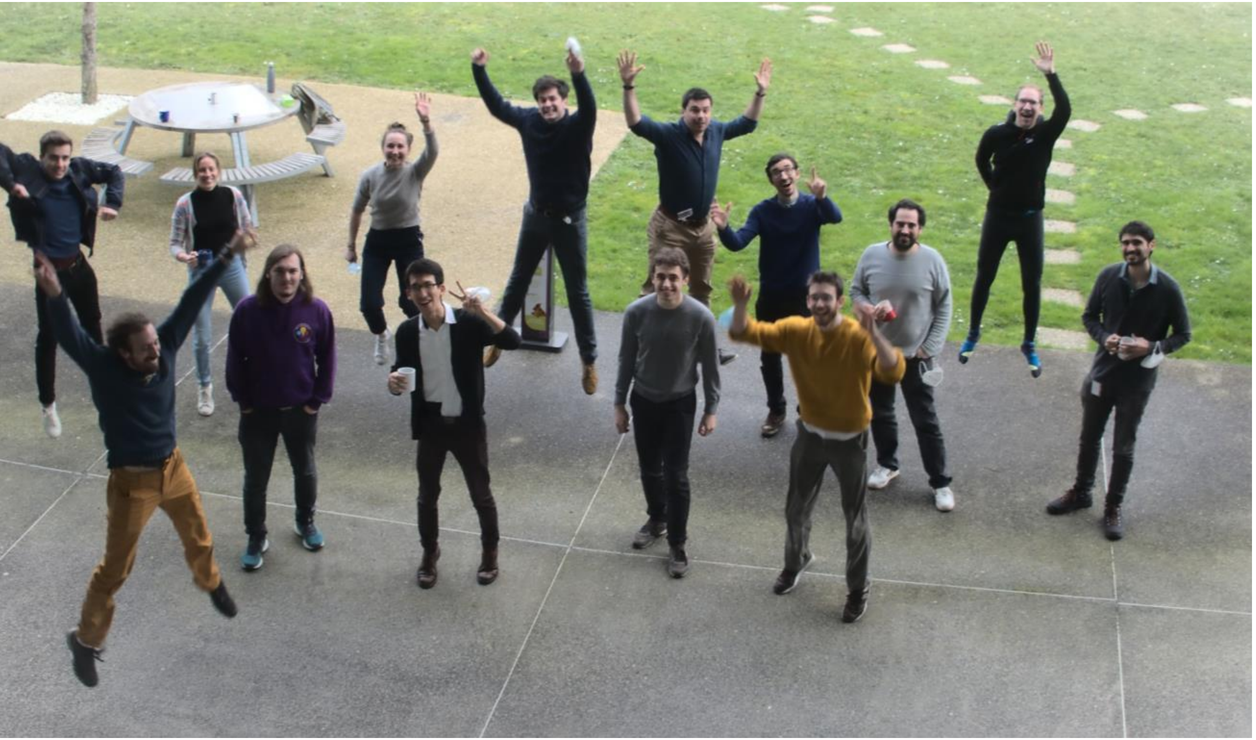
  mortality prediction ? Disease onset ? Computational phenotyping ?

- Evaluation for **concept proximity :** eg. eds-scikit biology concepts as ground truth

Collaborations ?

- Better **inclusion of temporality** with **transformer-based models**
- Transfer **from APHP to SNDS ?**

Machine Learning for Health and Society



Research axes

Representation learning for heterogeneous databases

- Learning despite database normalization errors
- Tabular deep learning

Health and Social Sciences

- Electronic health records
- Epidemiological cohorts
- Educational data mining

Data-science with statistical learning

- Statistical learning with missing values
- Machine learning for causal inference

Turn-key machine-learning tools for socio-economic impact

Helping to maintain and grow tools such as scikit-learn, joblib…

HAS, mission Data

1. Données produites par la HAS
2. Données observationnelles
3. Connaissances textuelles
4. Organisation

Supplementary slides

Skip-gram with Negative Sampling

- Given (word, context) = (w, c) pairs, and random representation in the embedding space:

$$\vec{w} \in \mathbb{R}^{V_W \times d} \qquad \vec{c} \in \mathbb{R}^{V_c \times d}$$

- The probability of occurrence of a pair (w, c) is given by:

$$\mathbb{P}(D = 1 | w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1+e^{-\vec{w} \cdot \vec{c}}}$$

- We maximize for a pair:

$$log\, \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [log\, \sigma(-\vec{w} \cdot \vec{c})]$$

- On the whole corpus:

$$P_D(c) = \frac{\#(c)}{|D|}$$

$$l = \sum_{w \in V_W} \sum_{c \in V_C} \#(w, c) \left(log\, \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [log\, \sigma(-\vec{w} \cdot \vec{c})] \right)$$

SVD-PPMI as the solution of the SGNS objective

Given the **Pointwise Mutual Information matrix**: $$PMI(w,c) = log\frac{P(w,c)}{P(w)^{\alpha}P(c)^{\alpha}}$$

Rewrite **Cooccurence** as: $$PMI(w,c) = log\frac{\#(w,c)\cdot|D|^{2\alpha-1}}{\#(w)\cdot\#(c)}$$

Enforce **sparsity**: $$PPMI(w,c) = max(PMI(w,c),0)$$

Factorization: $$SPPMI(w,c) = U_d\Sigma_d V_d$$

Dense representations as singular components: $$\Phi(w) = \frac{1}{2}(U_d \cdot \sqrt{\Sigma_d} + (V_d \cdot \sqrt{\Sigma_d})$$

Thoerical arguments in favor of Glove model *(Pennington et al., 2014)*

$$\sum_{i,j} f(X_{ij})(w_i^T \tilde{w}_j - \log X_{ij})^2 ,$$

- Offline (like SVD-PPMI)
- Avoid high cost of softmax (computeation of normalization functions)
- No cross-entropy error (model poorly long tail distributions)

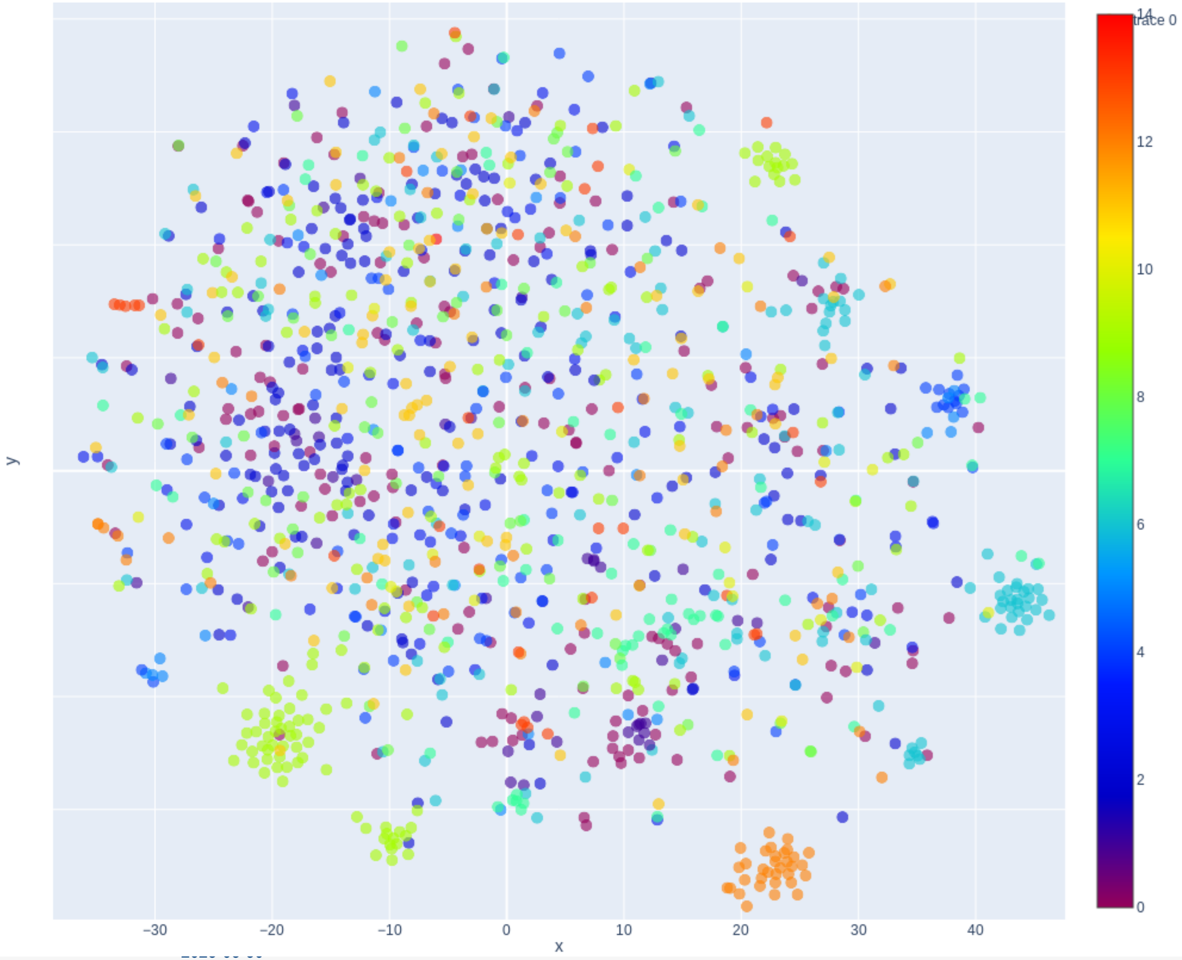
SNDS, details on data

- **Extraction**: Sample of three millions of patients followed 9 years
- **Sources**: DCIR (assurance maladie), PMSI (hospital billing codes) MCO, MCO_CE, SSR, SSR_CE, HAD
- **Events**: CIM10 (diagnostics), CCAM procedures (outpatient, inpatient, city care), city drugs, city biology
- **Granularity of codes**: ATC 7, CIM10 (4 characters), CCAM (7 characters), biology (4 characters) -> **15968 codes**
- **4416 codes** in common with APHP study cohort for rehospitalization@90 days

Quantitative results

ATC drug codes, r=30 days

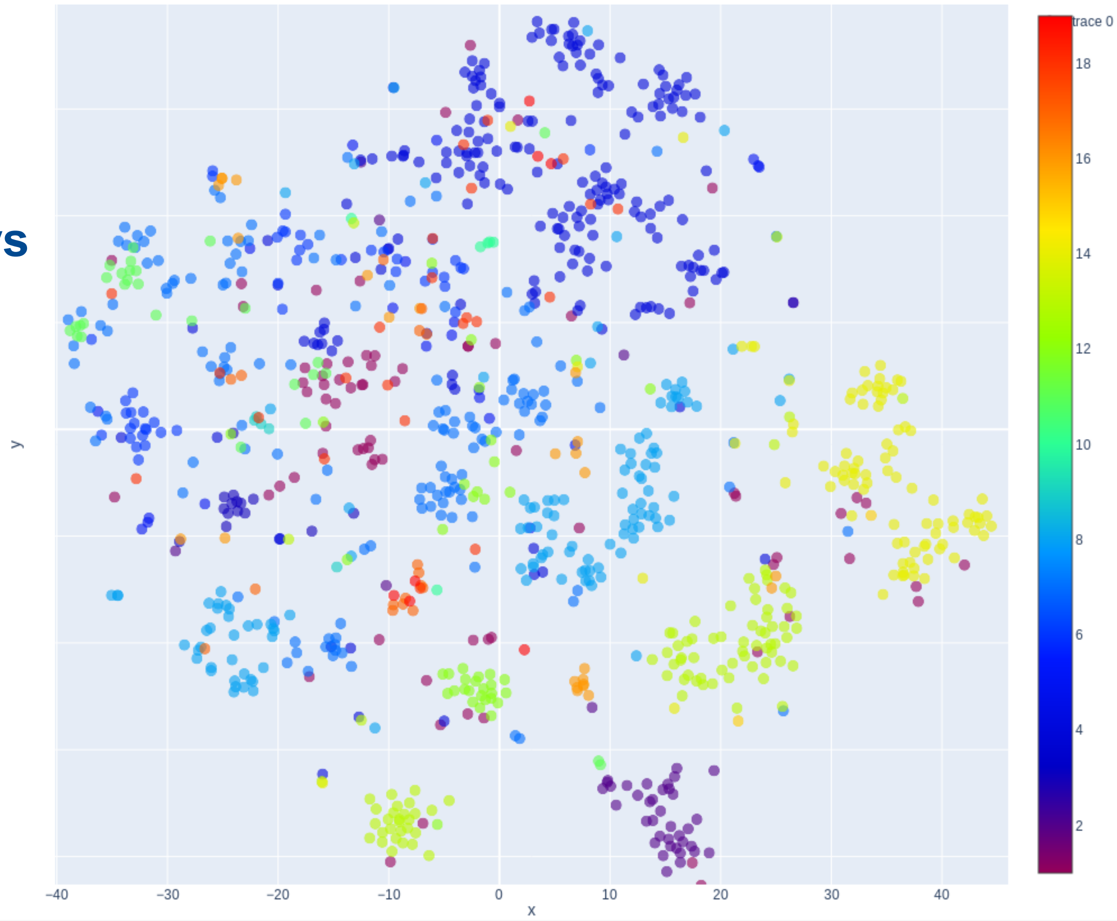
Colored by chapter



Qualitative results

CCAM billing procedures r=30 days

Colored by chapter



Benchmarks for intrisic evaluation

Metrics (Beam et al., 2019):

- **FMI:** quality of clustering
- **Medical Relatedness Measure:** How many close neighbors in the same hierarchical category ?
- **NDF-RT may treat / may prevent**
- **UMLS causative relationship**

granularity	radius	orientation	icd10	ccam	atc
L	30	centered	0.3818	*0.489*	0.2237
		future	*0.3878*	0.4821	0.1535
	90	centered	0.3677	0.4873	**0.3412**
F	30	centered	0.4533	0.5159	0.1891
		future	0.4465	0.4997	0.1368
	90	centered	**0.456**	**0.5306**	*0.3159*
		future	0.4528	0.5106	0.2376

Table 3: MRM computed on the embeddings grouped per terminology hierarchy. The best score is given in bold and the best score per granularity is given in italics.

Population selection for prediction

Extraction of **200,000 random patients** from APHP EDS

Study period 2017-2022 (stable Information System)

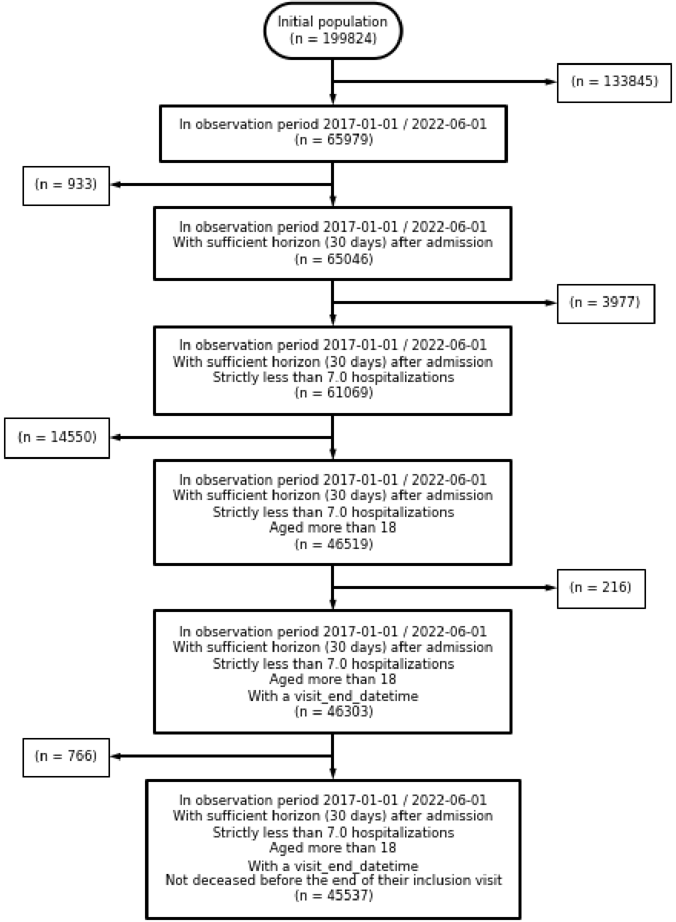
Sufficient horizon for followup (no right censure)

Exclusions: No children, not decesead during hospitalization

At least one event in **4923 codes occurring** at least 10 times:

- **drug exposure administrations:** 608, 577 events
- **procedure occurrences:** 252, 668 events
- **condition occurrences:** 219, 666 events

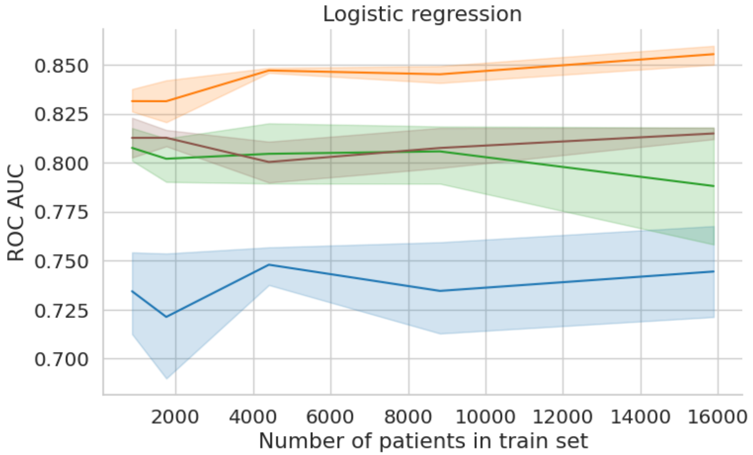
34, 063 patients: mean age = 54.4, female ratio = 55.8%, mean rehospitalization@30d=10.5%



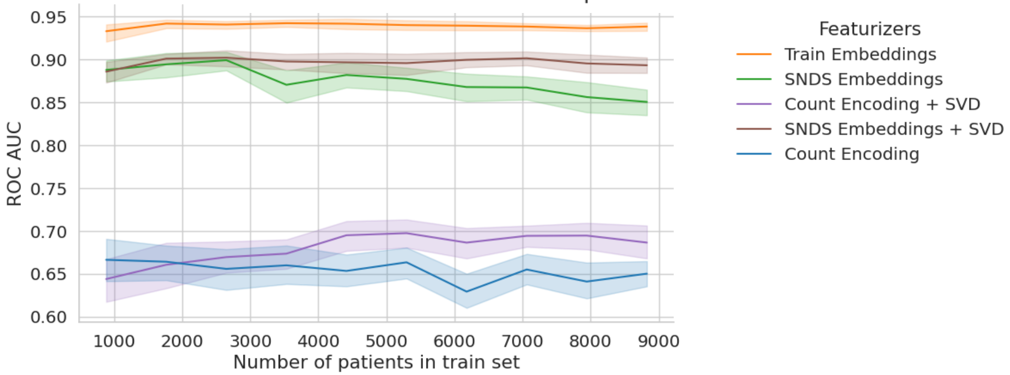
Task: Length Of Stay interpolation : <=7 days vs. >7 days

Adding a simple temporality decay seems very efficient.

Wo temporality decay

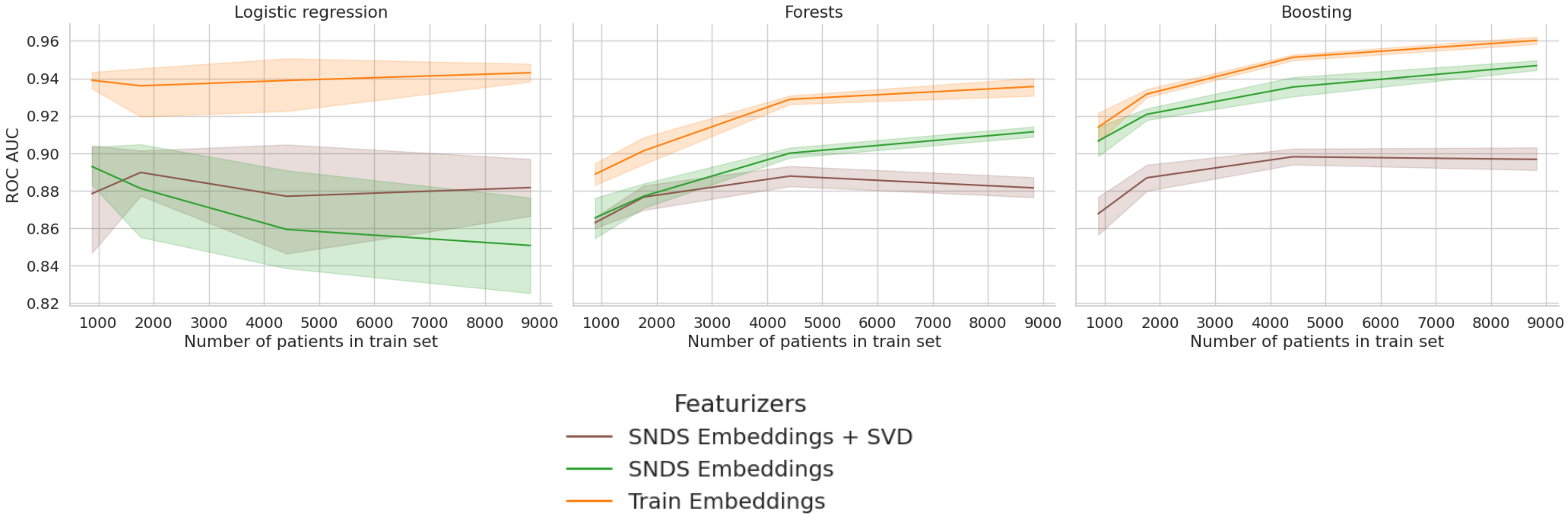


W temporality decay



Task: Length Of Stay interpolation : <=7 days vs. >7 days

There is sample gains at least for boosting and forests (no saturation of the task)



References

Event format:

- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 18.
- Bacry, E., Gaiffas, S., Leroy, F., Morel, M., Nguyen, D. P., Sebiat, Y., & Sun, D. (2020). SCALPEL3: a scalable open-source library for healthcare claims databases. *International Journal of Medical Informatics*, *141*, 104203.
- Chazard, E., Balaye, P., Balcaen, T., Genin, M., Cuggia, M., Bouzillé, G., & Lamer, A. (2022). Book Music Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare Structured Data. *Studies in Health Technology and Informatics*, *290*, 567-571.

G-estimation :

- Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, *37*(23), 3309-3324.
- Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

Medical concept embeddings

- Beam, A. L., Kompa, B., Schmaltz, A., Fried, I., Weber, G., Palmer, N., ... & Kohane, I. S. (2019). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing 2020* (pp. 295-306).
- Doutreligne, M., Leduc, A., Nguyen, D. P., & Vuagnat, A. (2020). Snds2vec, représentations continues pour les concepts médicaux du Système national des données de santé. *Revue d'Épidémiologie et de Santé Publique*, *68*, S35.

Word embeddings

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, *27*.
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

References

Embeddings for for predictive tasks

Ref. slides 8