life2vec: Life trajectories in high-dimensional spaces

Presented by:

Germans Savcisens (NEU)

Collaborators:

Tina Eliassi-Rad (NEU)

Lars Kai Hansen (DTU)

Laust Mortensen (KU)

Lau Lilleholt (KU)

Ingo Zettler (KU)

Anna Rogers (ITU)

Sune Lehmann (DTU)

Other Contributors:

Søren Mørk Hartmann (DST)

Northeastern University

About Me

Germans Savcisens

PhD, Computational Social Science

Germans Savčišens

Герман Савчишен

Ghe-r-men Saf-chi-shen

- MSc in Human-Centered AI (DTU)
- PhD in **Computational Social Science** (DTU):
  - ML in Social Science
  - AI Fairness and Ethics
- **Postdoctoral Associate Researcher** at the Khoury College of Computer Sciences (Northeastern University):
  - LLMs for health interventions,
  - Beliefs and knowledge in LLMs,
  - Scientific co-authorship networks,
  - Uncertainty of embeddings (nodes and graphs)

Northeastern University

Article | Published: 18 December 2023

Using sequences of life-events to predict human lives

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler & Sune Lehmann

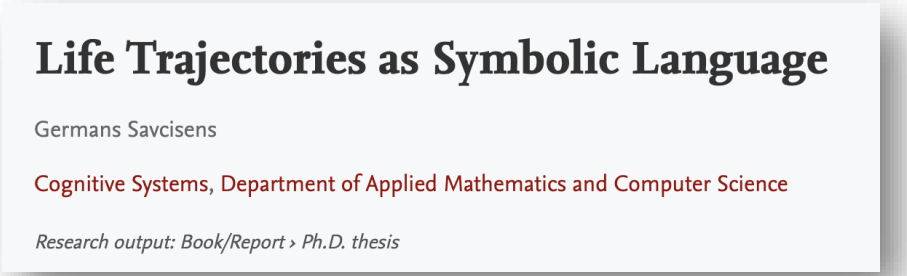
Nature Computational Science **4**, 43–56 (2024) | Cite this article

26k Accesses | **13** Citations | **2326** Altmetric | Metrics

Code Availability: SocialComplexityLab/life2vec (github.com)
carlomarxdk/life2vec-light (github.com)

Main contributions of the project:

1. **Propose a framework** (*transformer-based*) to analyze large-scale socioeconomic and health data
2. Demonstrate the **power of dense representation**
3. **Adapt explainability methods** to understand predictions



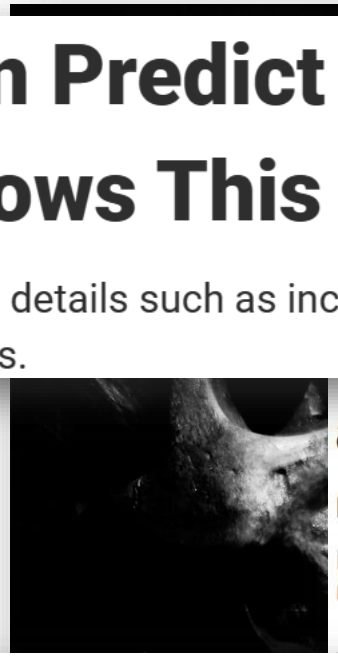
(Bonus) PhD Thesis: Link to DTU Orbit

This AI calculator can predict when you'll die with 'ext accuracy'

BY HAROLD LEMON TUBIANO
PUBLISHED DEC 22, 2023 2:54 PM

AI Tool That "Can Predict Almost Anything", Even Death, Follows This Procedure

The algorithm incorporates various details such as income, occupation, location, injuries, and pregnancy history for its predictions.



accurate

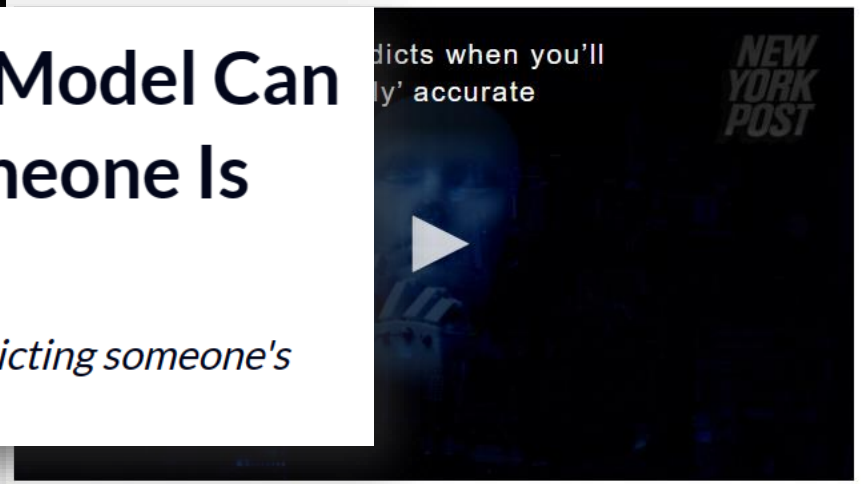
By Asia Grace

Published Dec. 20, 2023
Updated Dec. 20, 2023, 4:11 p.m. ET

AI predicts death, do we really want kn

Life2vec: This New AI Model Can Someone Is

of predicting someone's



Business Insider

AI can accurately predict death about 80% of the time, new study finds

A new research study using a large dataset of 6 million people in Denmark used machine learning to predict when someone is likely to die.

23 Dec 2023

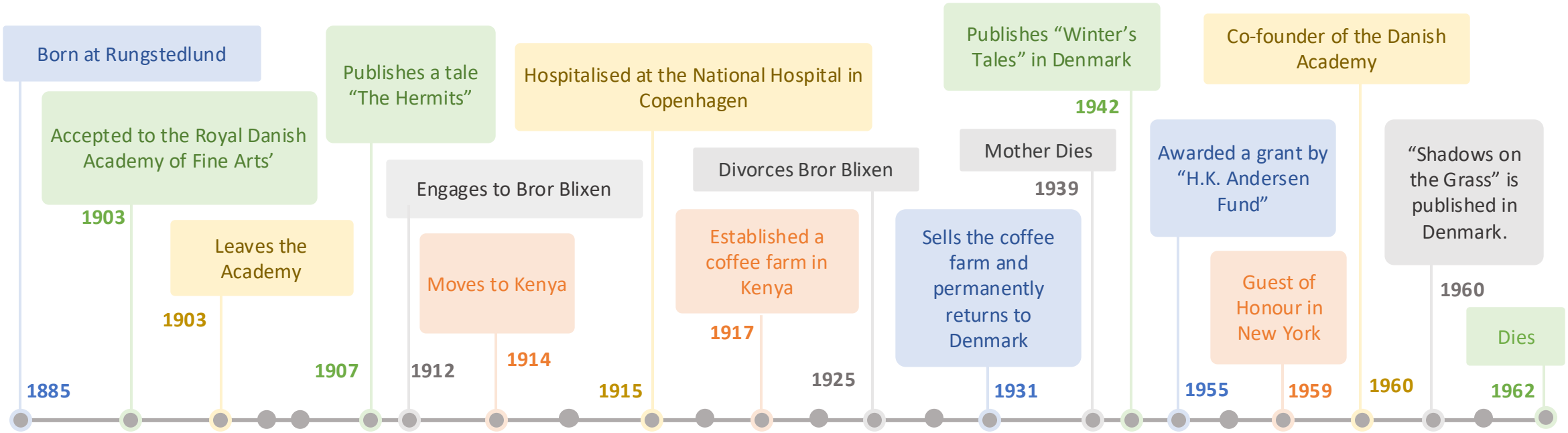
Agenda

Part 0	**Introduction**
Part I	Data
Part II	Representation Learning and NLP
Part III	Socio-economic and health *language*
Part IV	Capturing the structure *with* the `life2vec`
Part V	`life2vec` *as* a foundation model

Northeastern University

Life Trajectories

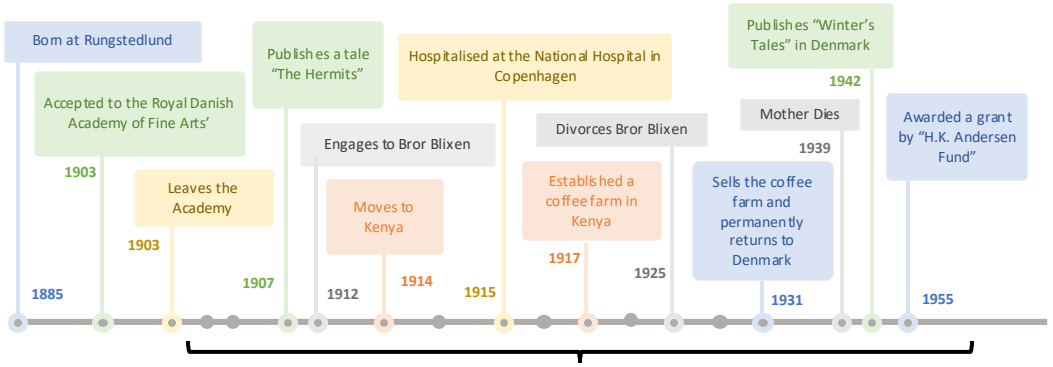
Life of Karen Blixen (Danish author)*



** simplified*

Northeastern University

The Problem



Simplifying data

How many times admitted to a hospital?

Career changes?

Traveling abroad?

Travelled within a year	…	Married	Hospital Admission
1	…	1	2

Model 1

Model 2

•••

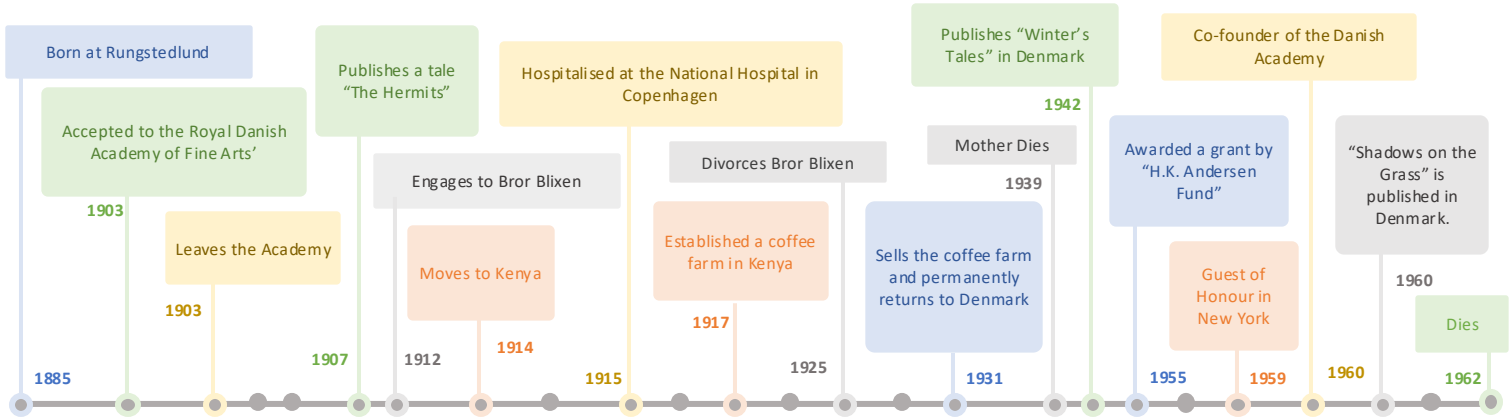
Model N

Probability of readmission to a hospital?

Income level within the next year?

* simplified

Northeastern University



We want a **single** model that takes **nuanced life trajectories**

General Purpose Model

12	-2	10	5	39	-30	20	33

Compressed **representation of life** progression

Predict the human behaviour
(on an *individual* level)

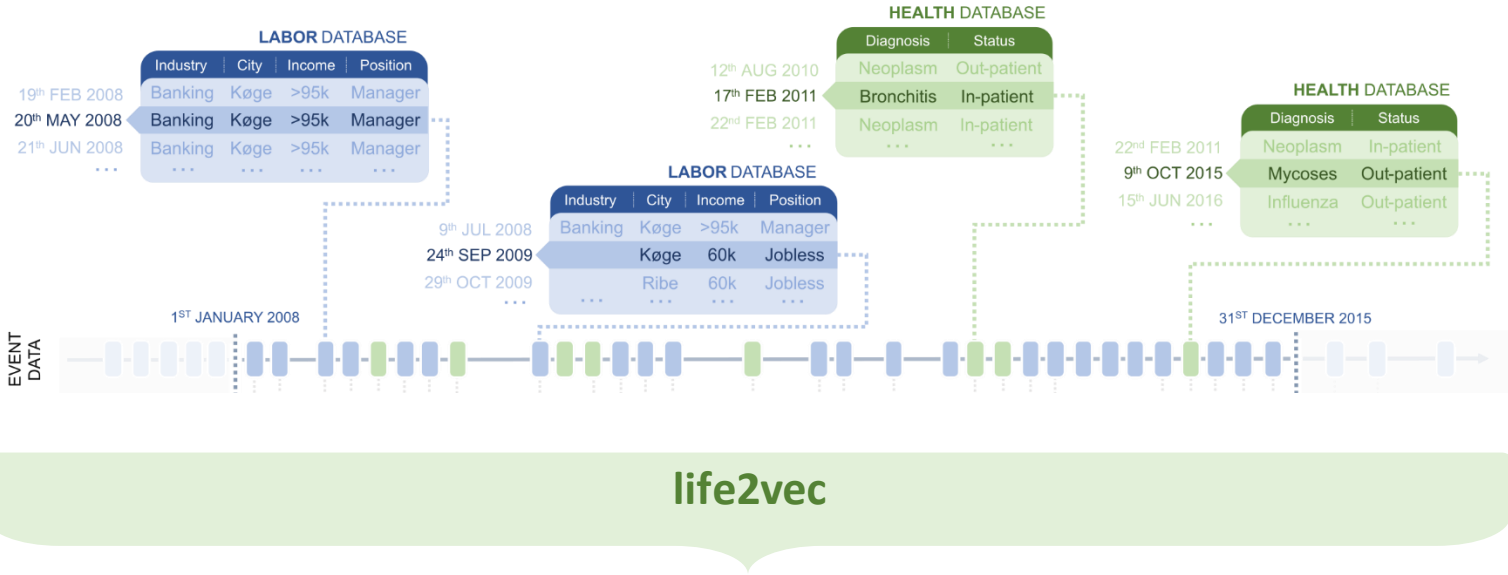
Give comprehensive insight into the data

Study sociological phenomena
(on a *global* scale)

Northeastern University

Our Work: *life2vec as a proof-of-concept*

Life Progression from the point of view of Labor and Health Records



Main Components:

← Text- like encoding of data

← Encoder Model

Novel way to understand
The structure of the data

Process complex-structure
Such as Life-Sequences

Explainable predictions

Part I

Life-Trajectories and Data

Danish National Registry

People
Names, population, health, elections, housing, church, gender equality...

Social conditions
Criminal offences, social benefits for senior citizens, cash benefits, placements...

Transport
Cars, goods transport, passenger transport, infrastructure, traffic accidents...

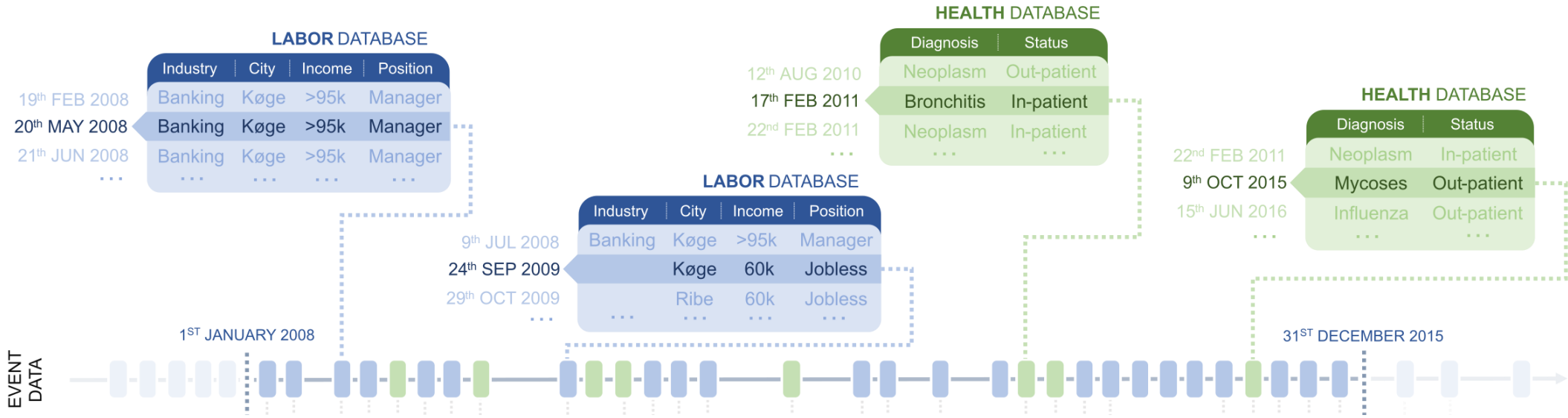
Labour and income
Employment, unemployment, earnings, income, wealth...

Education and research
Number of students, education programmes, innovation...

Culture and leisure
Film, media, museums, music, digital behaviour, sports...

Personal raw data is
tied to the **Social Security Number** (CPR)

**AI-Generated Image



DANMARKS STATISTIK

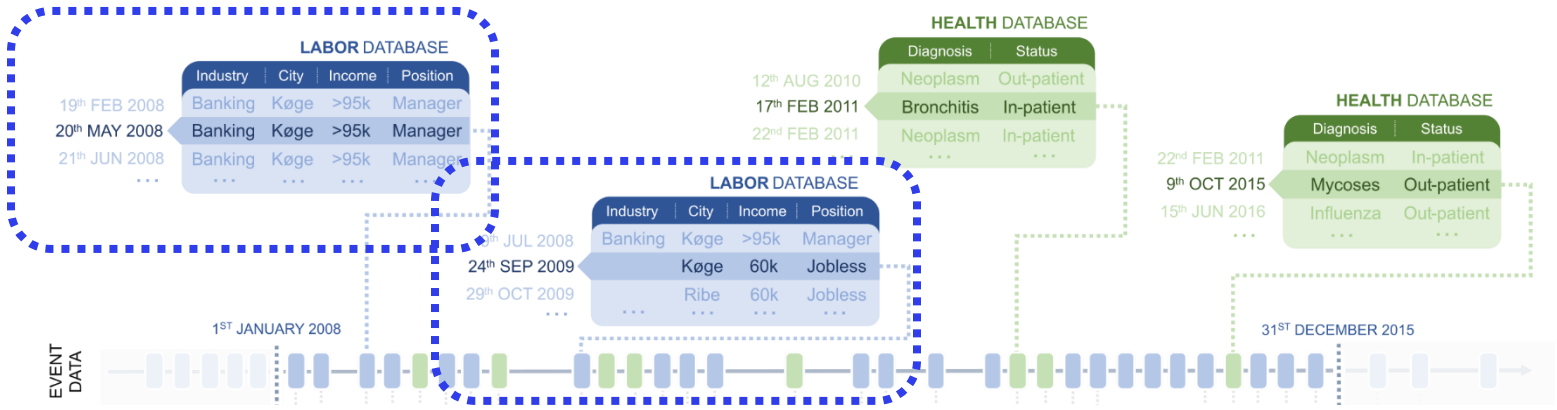
Labour Data

Health Data

Detailed reconstruction of labor and health life trajectories

Northeastern University

Labour Data



Records of any reported and taxable income:

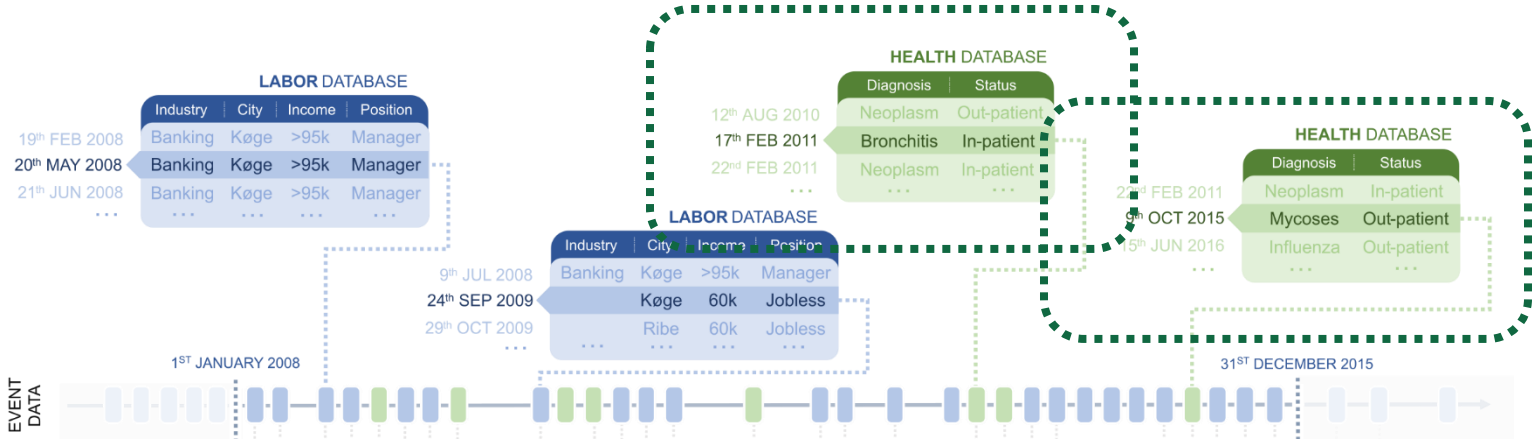
- Each record has around 70 features
- Hourly precision
- Timespan: 2008-2020
- Features have underlying structure

We focus on:

- **Income** (if applicable):
- **Residence**
  - Country of Origin / Citizenship
  - Address in Denmark
- **Socio-economic status**:
  - Age and sex
  - Employment status

Northeastern University

Health Data



Records of visits to a health practitioner or hospital:

- Focus on 3 features
- Diagnoses encoded in the ICD10 System

Features we use:

- **Diagnosis** (Initial, no follow-ups)
- **Patient type**: inpatient, outpatient, and emergency
- **Urgency**: Urgent, Non-urgent

Labor Data: Hierarchies

Example of codes describing the **Industry**

DB07 Code	Interpretation
C	Manufacturing
18	Printing and Reproduction of Recorded Media
18.1	Printing and Related Services
18.14	Bookbinding and Similar Services

Example of codes describing the **Occupation**

ISCO-08 Code	Interpretation
2	Professionals
26	Legal, Social and Cultural Professional
265	Creative and Performing Artists
2654	Dancers and Choreographers

Health Data: ICD-10

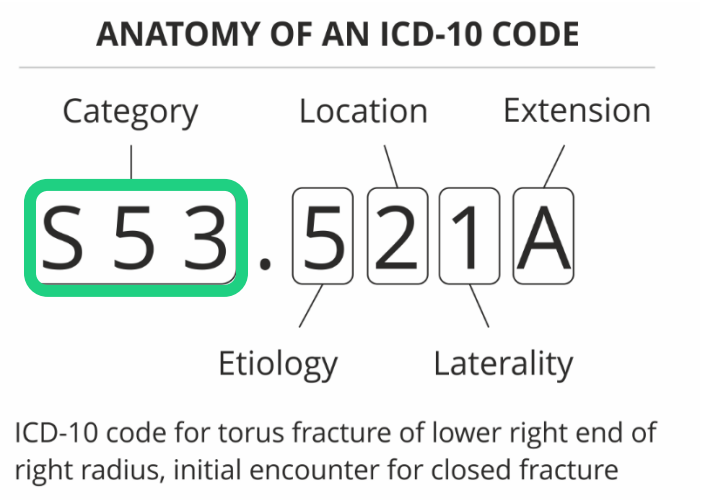
ICD-10 Code	Interpretation
S01	Open wound of head
S01.3	Open wound of ear
S01.35	Open bite of ear
S01.352	Open bite of left ear
S01.352D	Open bite of left ear (subsequent encounter)

Examples of ICD10 codes:

Y93.D: Activities involved arts and handcrafts

W61.62XD: Struck by duck, subsequent encounter

H47.51: Disorders of visual pathways in (due to) inflammatory disorder

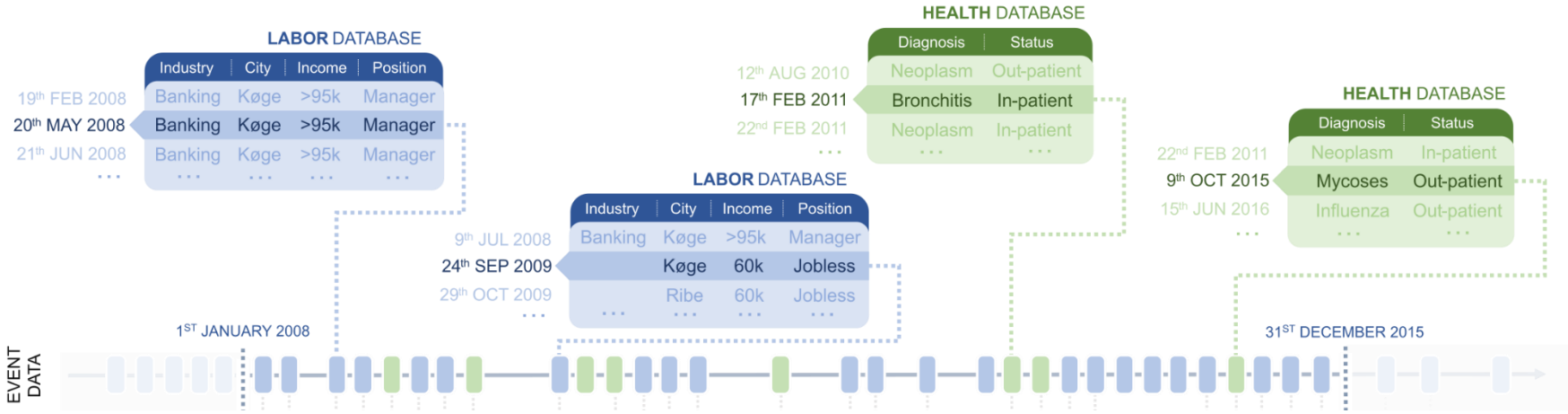


ICD-10 code for torus fracture of lower right end of right radius, initial encounter for closed fracture

Northeastern University

Power of National Registry

The National Registry is a source of **fine-grained information** about **the progression of one's life**.
Hence, unique possibility to study life progression and life outcomes.



How do we analyze?

Part II

Representation Learning and NLP

Northeastern University

Language and Machines

"*Everything was beautiful and nothing hurt*"[1]

**AI-Generated Image

1. Slaughterhouse-Five, Kurt Vonnegut (1969)

Language and Machines

"*Everything was beautiful and nothing hurt*"[1]

↓

Create a numerical representation of the text!

↓

a	…	and	…	beautiful	…	everything	…	hurt	…	no	nothing	…	was	…	zyzzyva
0	…	1	…	1	…	1	…	1	…	0	1	…	1		0

↓

Computers can work with numbers

1. Slaughterhouse-Five, Kurt Vonnegut (1969)

Language and Machines

a	…	and	…	beautiful	…	everything	…	hurt	…	no	nothing	…	was	…	zyzzyva
0	…	1	…	1	…	1	…	1	…	0	1	…	1		0

If we reconstruct the sentence

“Beautiful was nothing and everything hurt”

“Everything beautiful hurt and was nothing”

“Everything hurt nothing and was beautiful”

1. Slaughterhouse-Five, Kurt Vonnegut (1969)

Language and Machines

It is even more obvious issues if we look here.

Let's match people based on their description

"Viktor prefers apples"

- apples: 1
- prefers: 1
- likes: 0
- spaceships:0
- kiwi: 0

"Maria likes spaceships"

- apples: 0
- prefers: 0
- likes: 1
- spaceships:1
- kiwi: 0

"Susanne likes kiwi"

- apples: 0
- prefers: 0
- likes: 1
- spaceships: 0
- kiwi: 1

1. Slaughterhouse-Five, Kurt Vonnegut (1969)

Complexity of Language

Language is a super complex signal…

…and it inherits many issues associated with the longitudinal data.

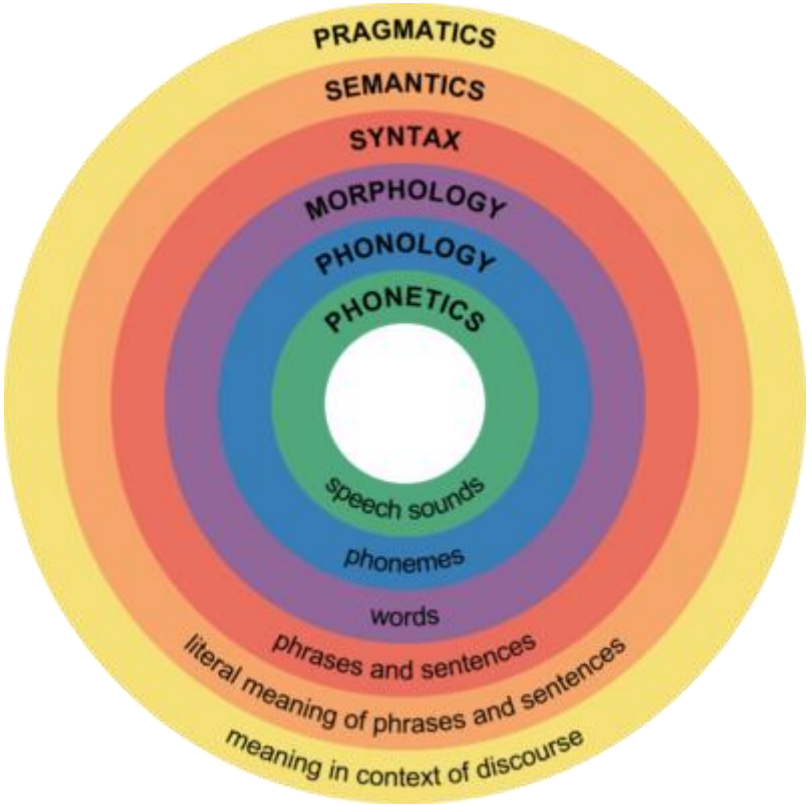
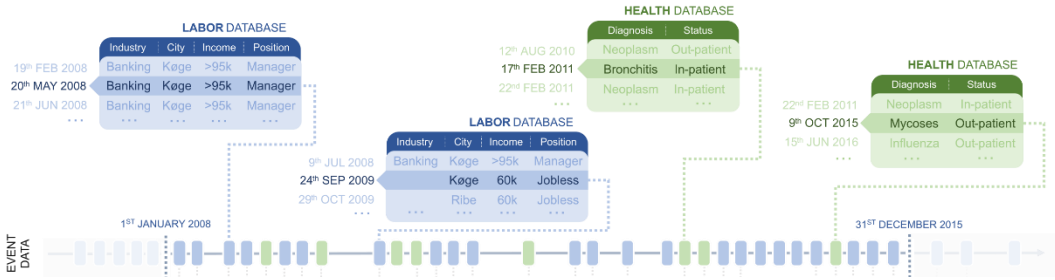


Image: Luchmee, D. (2019, July 25*). The Complex Skill of Language*. HappyNeuron. Retrieved March 5, 2024, from https://news.happyneuronpro.com/the-complex-skill-of-language/

Northeastern University

Language and Life Sequences

"*Everything was beautiful and nothing hurt*"



These two cases have similar issues!

The field of NLP has two great solutions!

Word Representations

Captures aspects of words

Large Language Models

Handles structured sequences

Representation of Places

	longitude*	latitude*
Great Pyramid	31.08	29.58
Petra	30.19	35.26
Machu Picchu	13.09	35.26
Colosseum	12.29	41.53

These values **capture spatial location,**
and allow us to **reason about the distances** (“*similarity*”).

** simplified*

Word Representations

Solution in NLP: Take a step back and assign **coordinates** to **words** (capture meaning)

	liveliness	vehicle-(ness)	artificiality
spaceship	0.0	1.0	1.0
apple	**0.3**	0.0	0.2
kiwi	**0.3**	0.0	0.3
dog	1.0	0.3	0.1

Representation of Documents

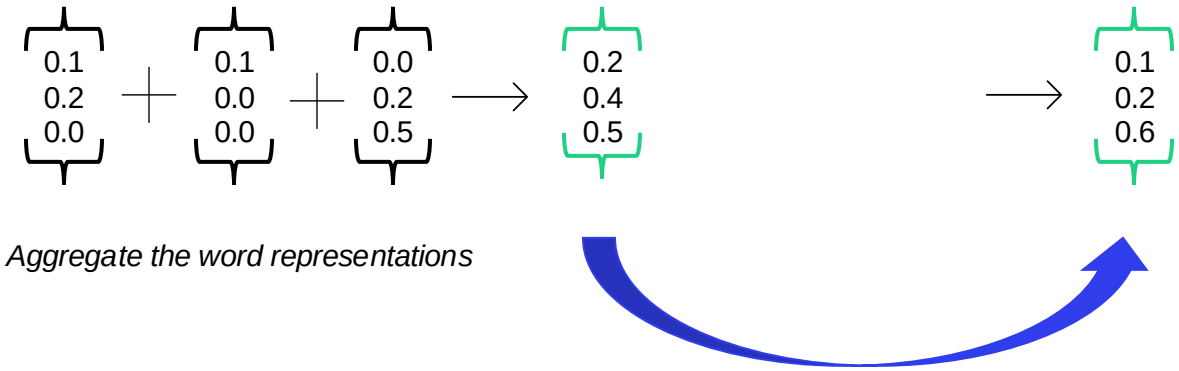
Using these nuanced word embeddings, we can create document embeddings

"Viktor prefers apples"

"Susanne likes kiwi"

"Maria likes spaceships"

"Document" representation



Aggregate the word representations

Learning Embeddings

We can employ different methods to create the word embeddings:

1. **Manually** assign values to each dimension (based on questionaries)
2. **Frequency-based**: Count-Vectors, TF-IDF, N-grams
3. **Prediction-based**: SkipGram, CBOW, GLoVE, by-products of training ML algorithms (e.g. RNNs)

Embedding Spaces and Structure

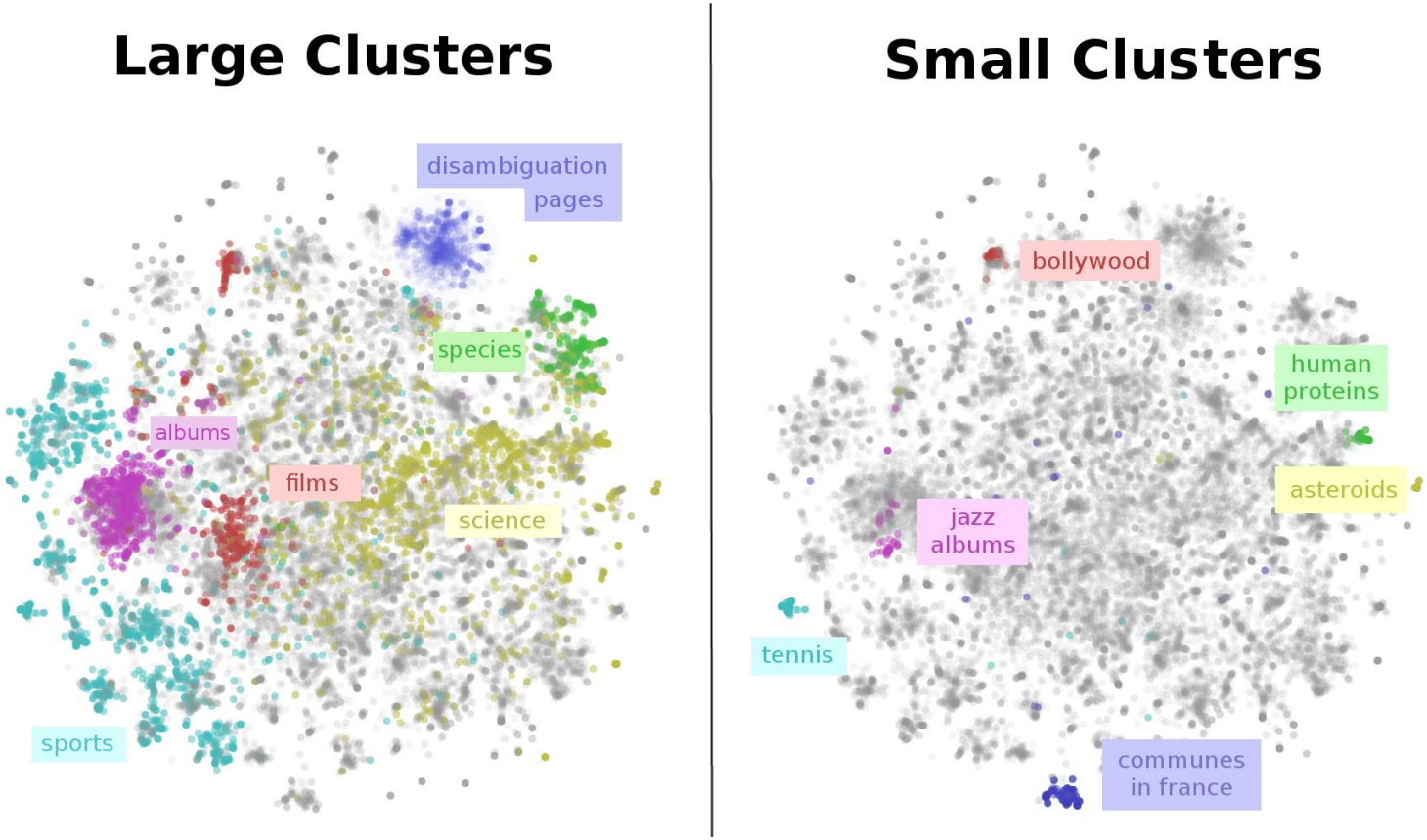


Fig 1: Two-dimensional projection of the word embeddings (word2vec)[1]

1. Olah, C. (2015, January 16). *Visualizing Representations: Deep Learning and Human Beings*. Colah's Blog. Retrieved March 3, 2024, from https://colah.github.io/posts/2015-01-Visualizing-Representations/

Northeastern University

Embedding Spaces and Structure

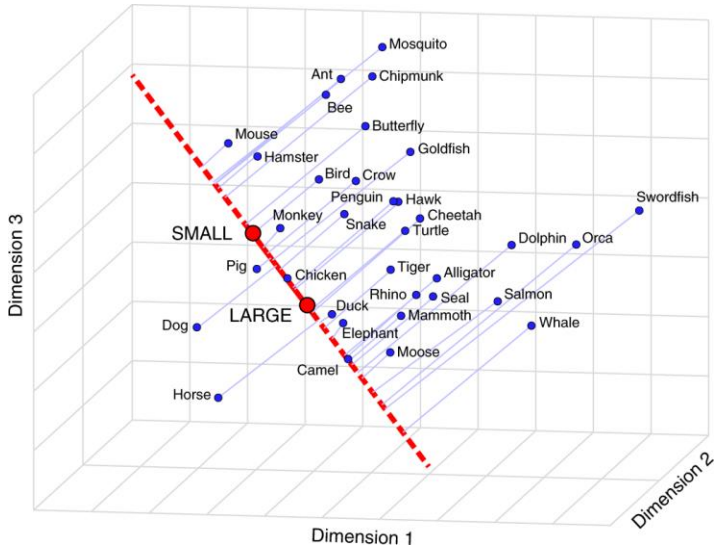


Fig.1: Schematic illustration of semantic projection[1]

In the embedding space (GloVe), "animal"-related words projected onto the "small-large" direction

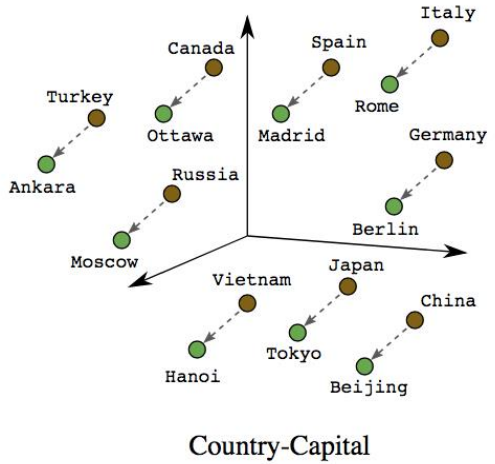


Fig.2: Embeddings can produce remarkable analogies[2]

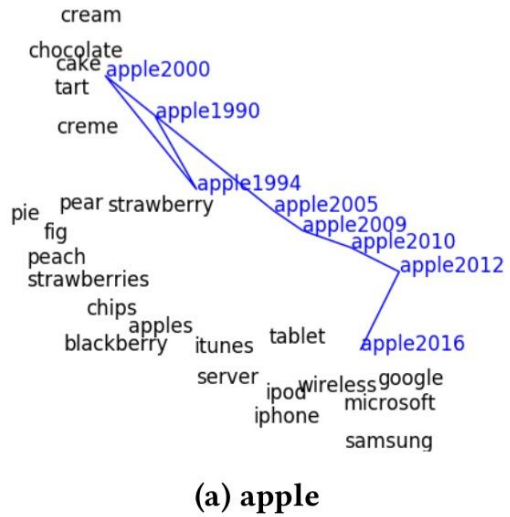


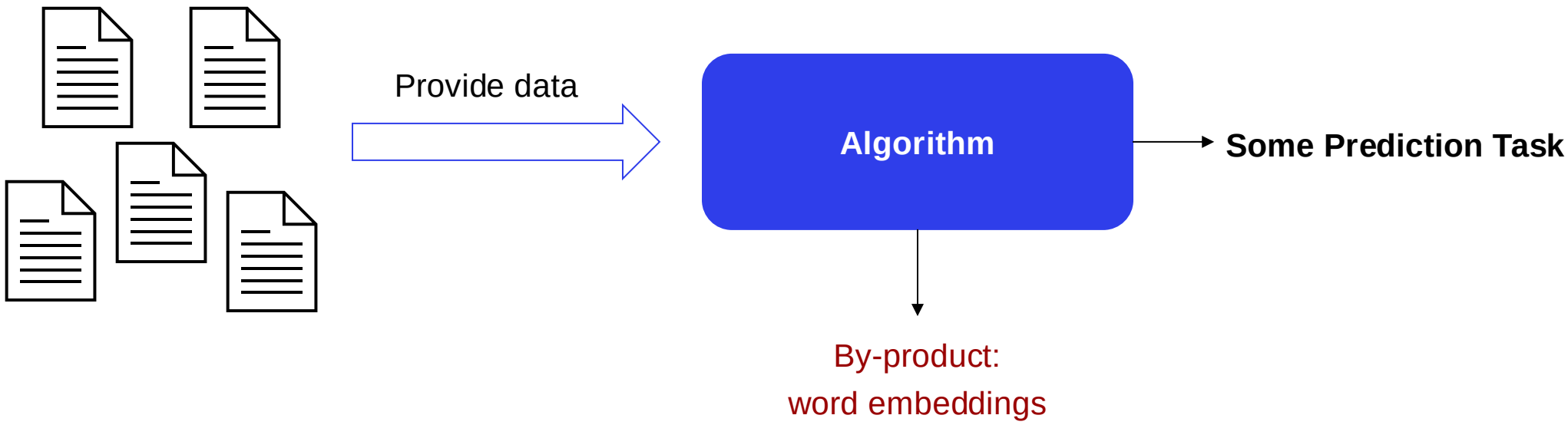
Fig.3: Trajectories of brand names[3]

Temporal evolution of terms with word2vec

1. Grand, G., Blank, I.A., Pereira, F. *et al.* Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat Hum Behav* **6**, 975–987 (2022). https://doi.org/10.1038/s41562-022-01316-8
2. *Embeddings: Translating to a Lower-Dimensional Space*. Google for Developers. Retrieved March 3, 2024, from https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space
3. Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018, February). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining* (pp. 673-681).

General Purpose Embeddings

- But how to make sure that we have a **meaningful space**?
- The nature of the task influences the representations

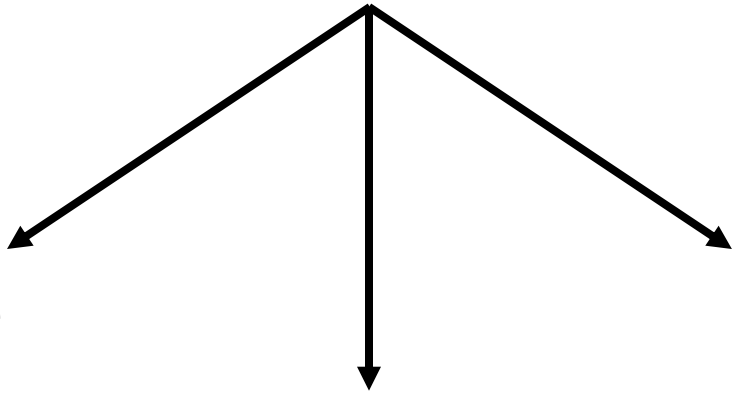


Transformer-based Models

Powerful Sequence Models already exist:

Large Language Models

Bidirectional Encoder Representations from Transformers (**BERT**)

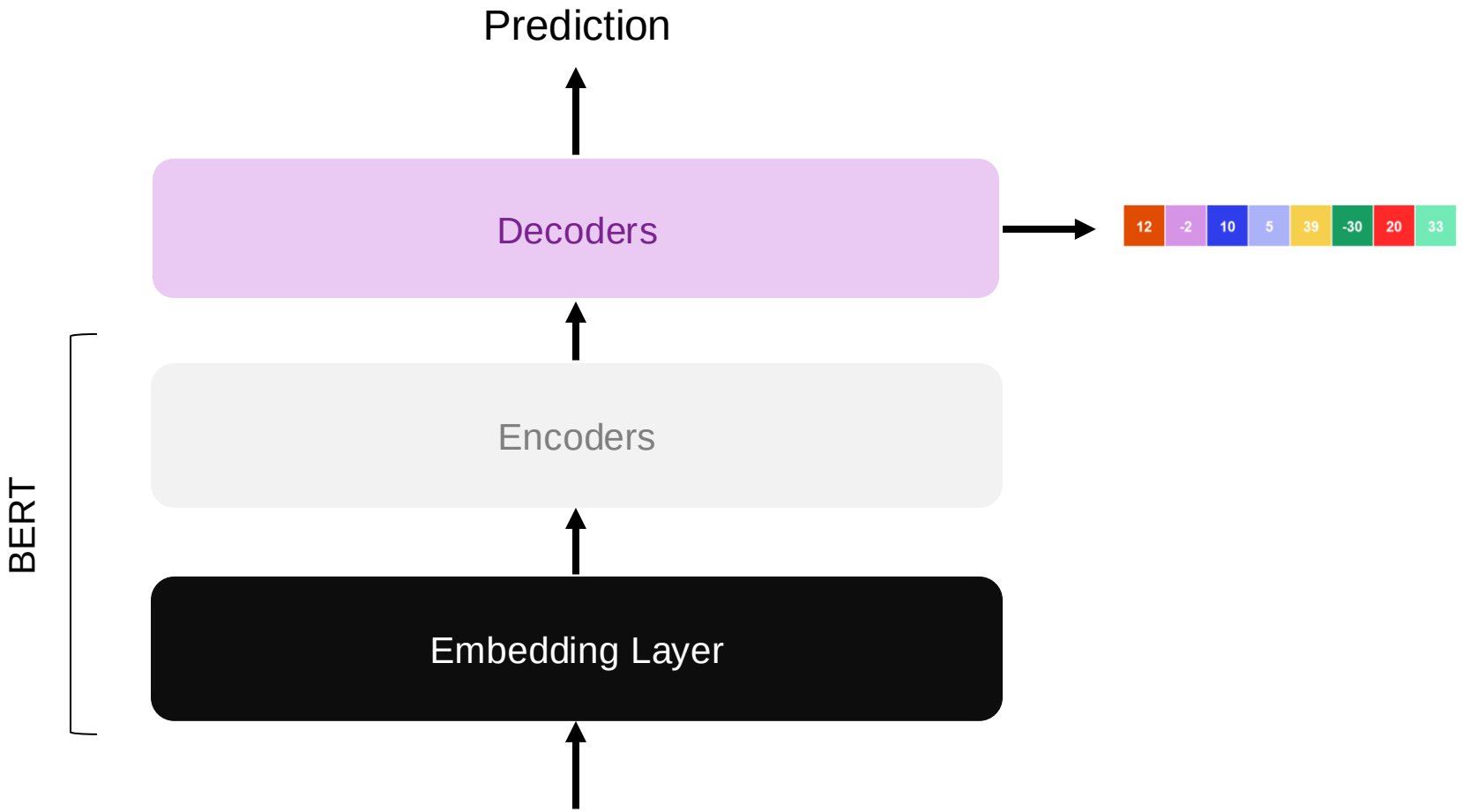


Create ***nuanced*** **word embeddings** and handle **complex sequences**

General-purpose model, adaptable to new tasks

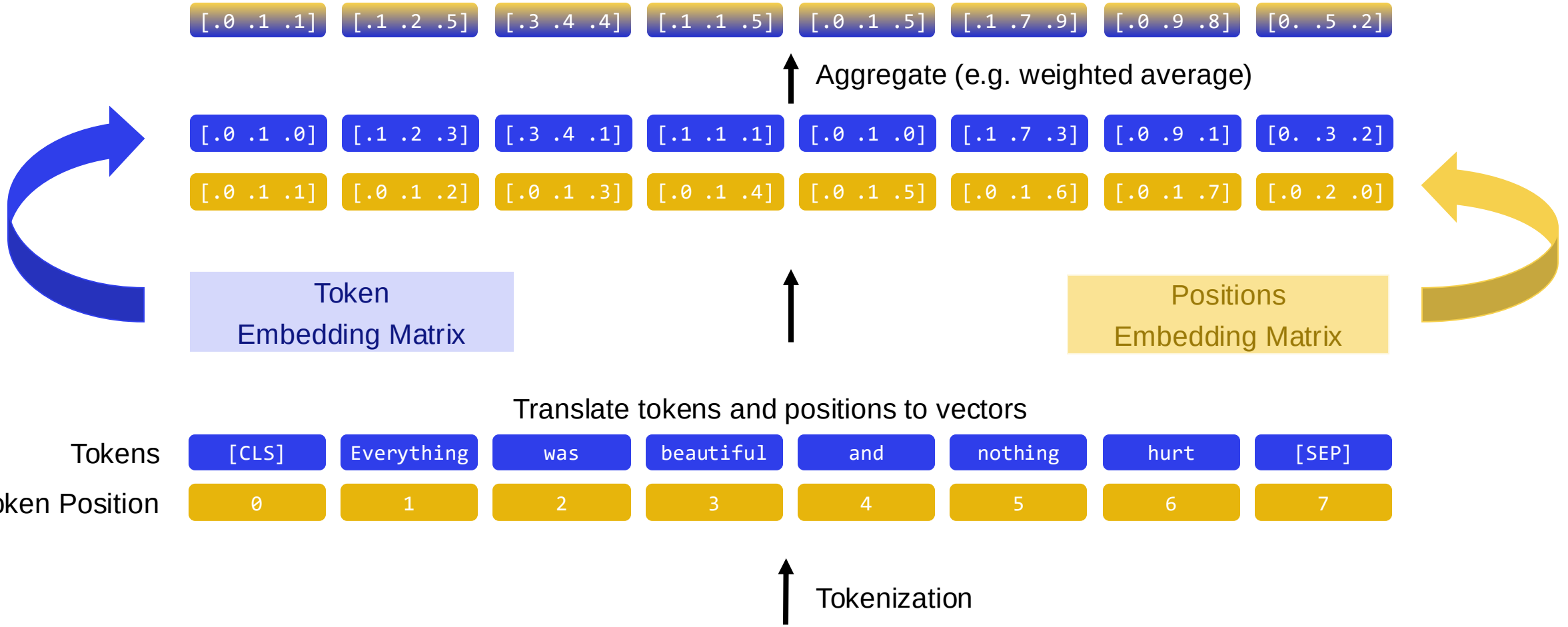
Great **predictive performance** on many NLP tasks

Transformer Architecture (BERT)



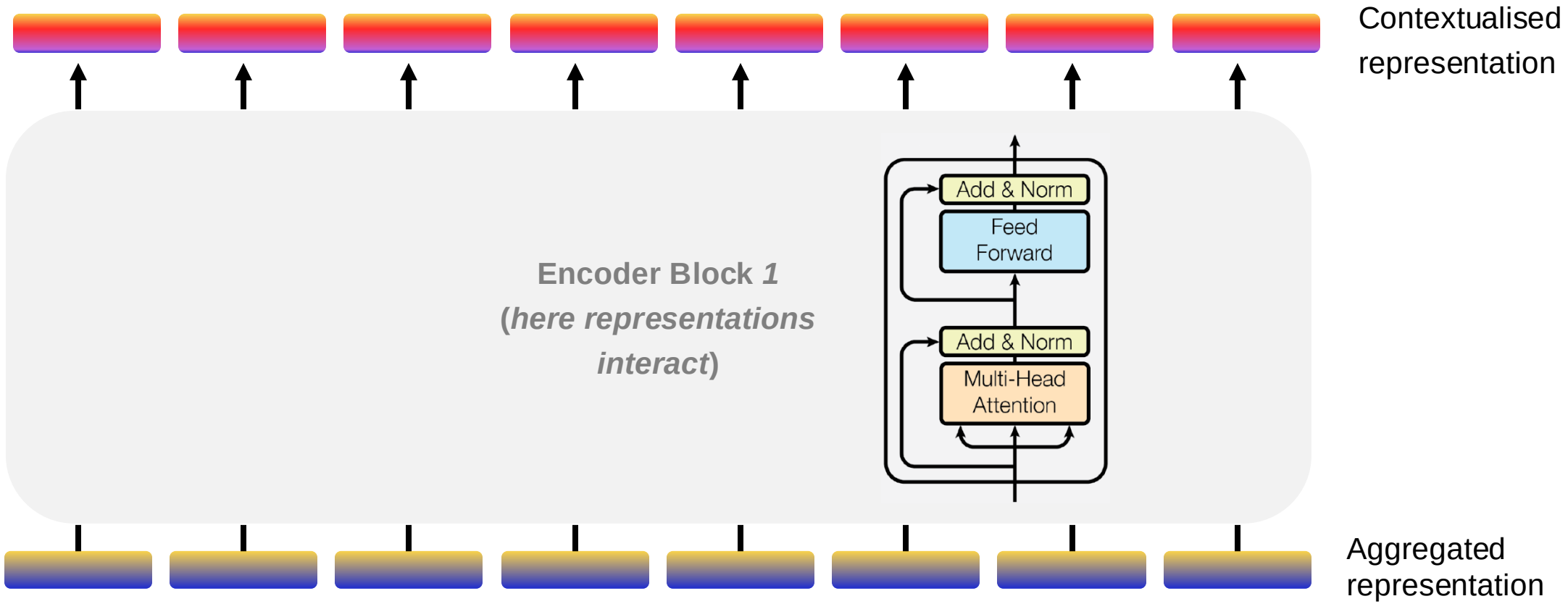
"Everything was beautiful and nothing hurt"

Embedding Layer



Northeastern University

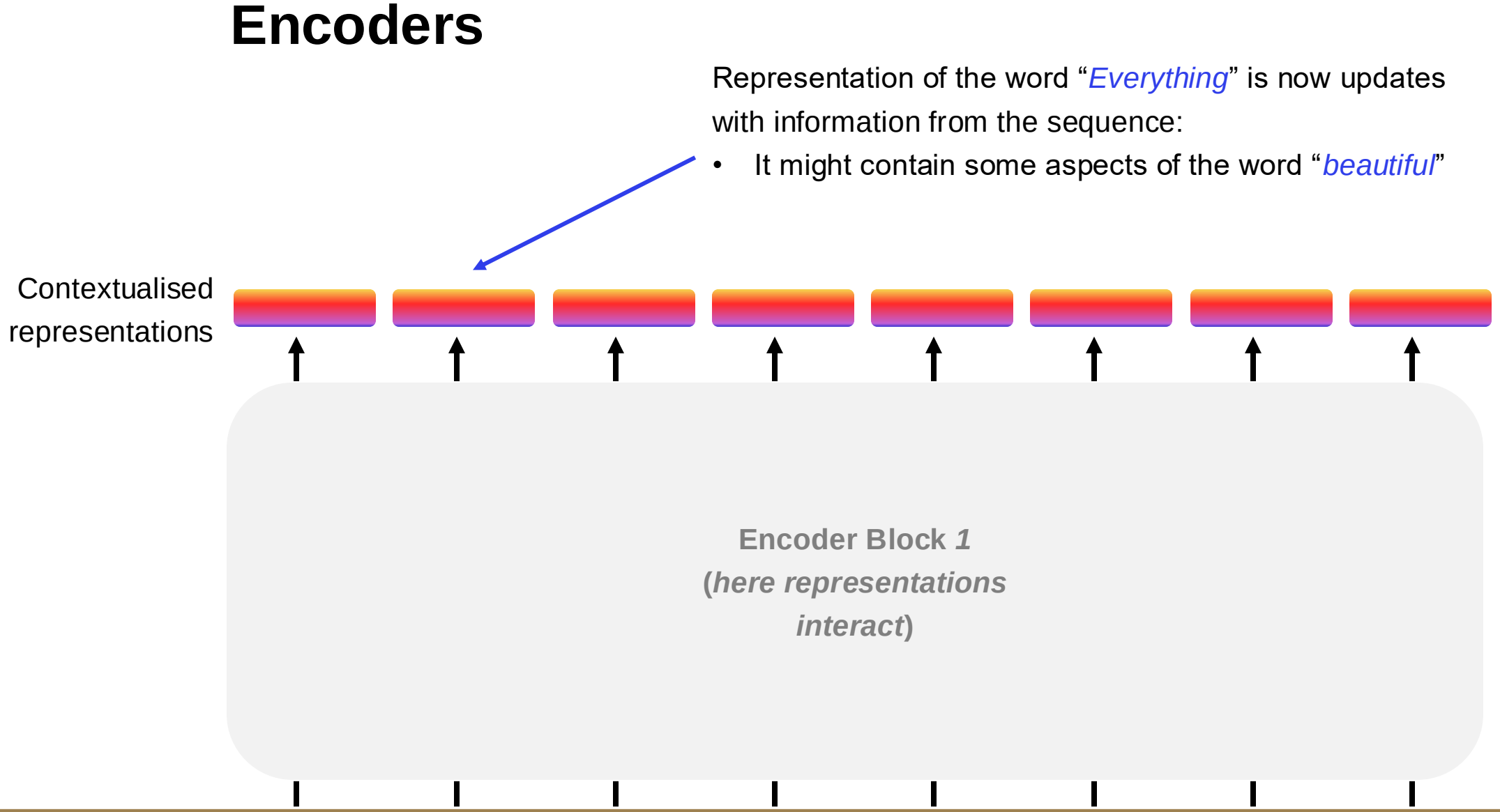
Encoders



Encoders

Representation of the word "*Everything*" is now updates with information from the sequence:

- It might contain some aspects of the word "*beautiful*"

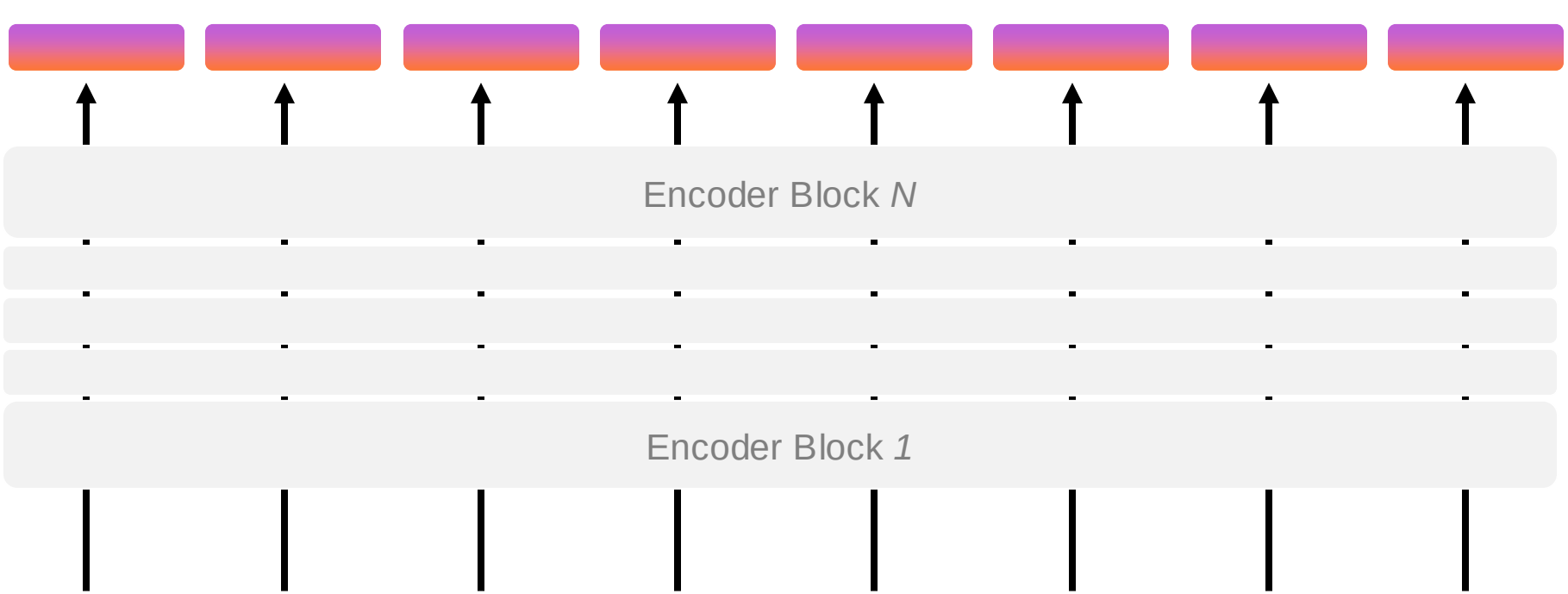


BERT Encoders

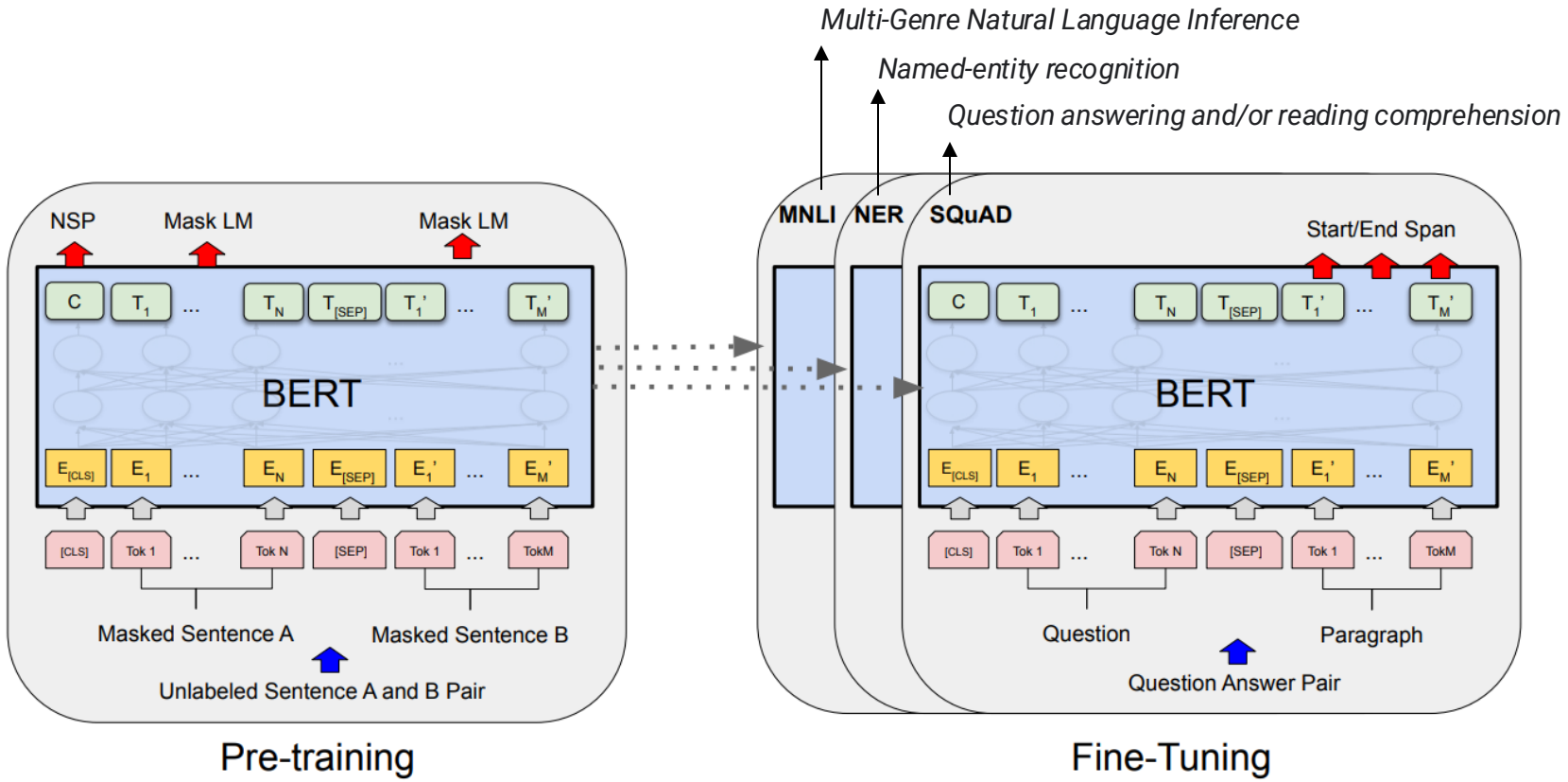
Contextualized token representations **contain rich and nuanced information** about the role of a token in a sequence.

What you can do with the output of decoders:

- Make predictions on the first token (CLS, more about that later)
- Using any ML model



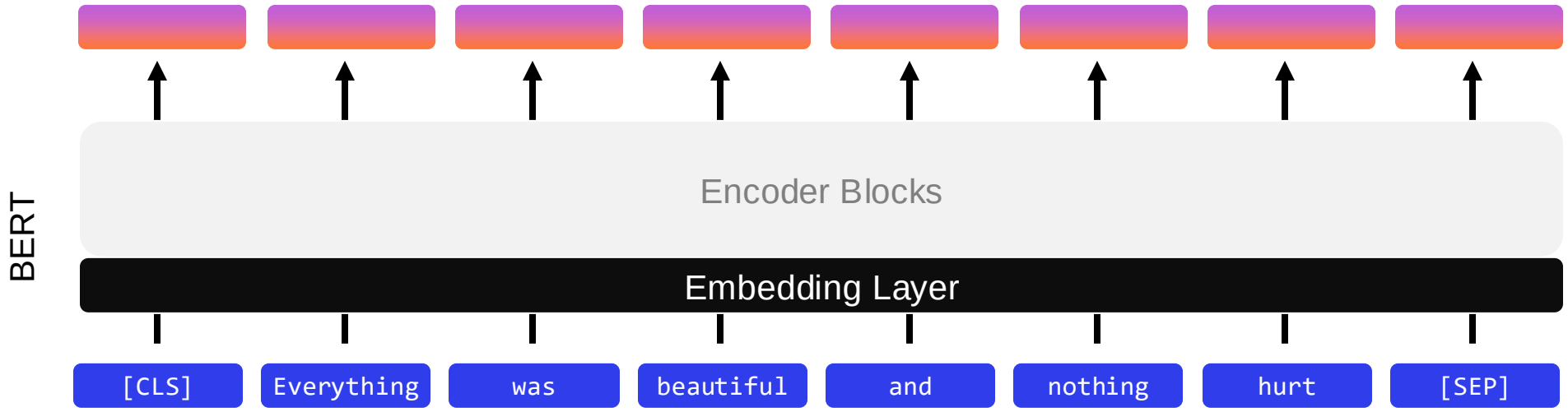
BERT: Training Stages



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

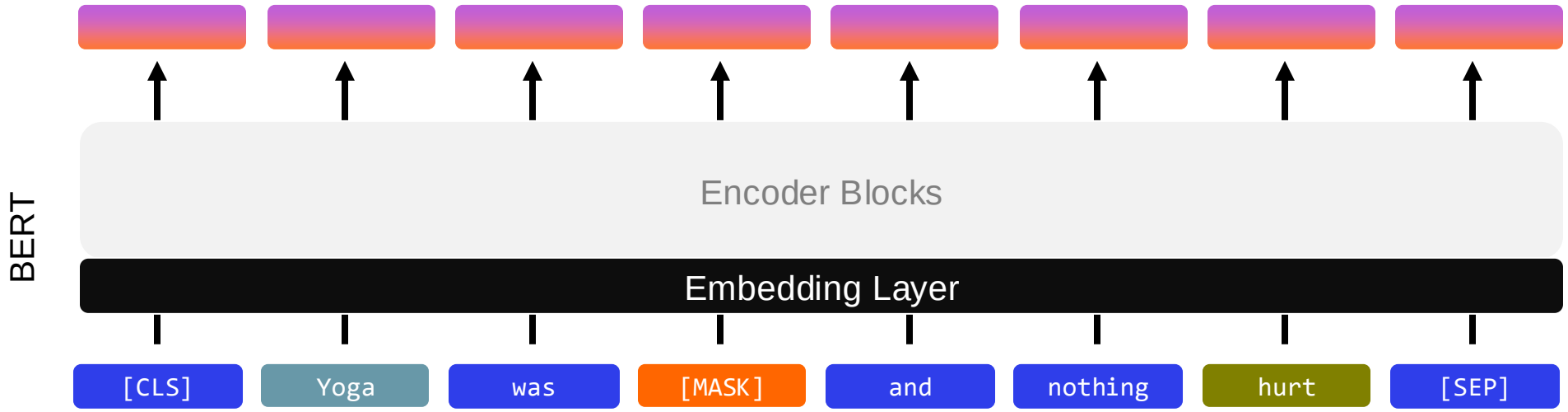
BERT: Pretraining

- Mask 15% of tokens (not including **[PAD],[SEP],[CLS]**):
  - 10% unchanged
  - 10% substituted with random tokens
  - 80% substituted with the **[MASK]** token



BERT: Pretraining

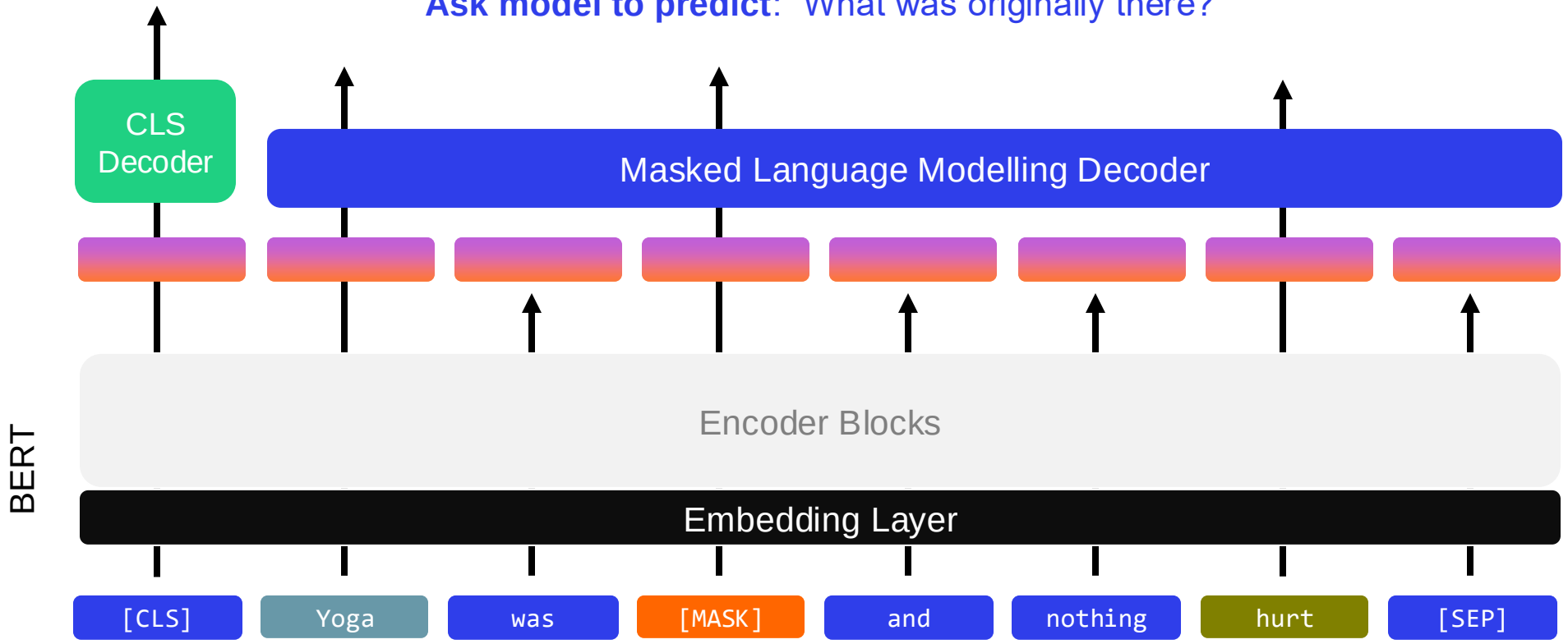
- Mask 15% of tokens (not including **[PAD],[SEP],[CLS]**):
  - 10% unchanged
  - 10% substituted with random tokens
  - 80% substituted with the **[MASK]** token



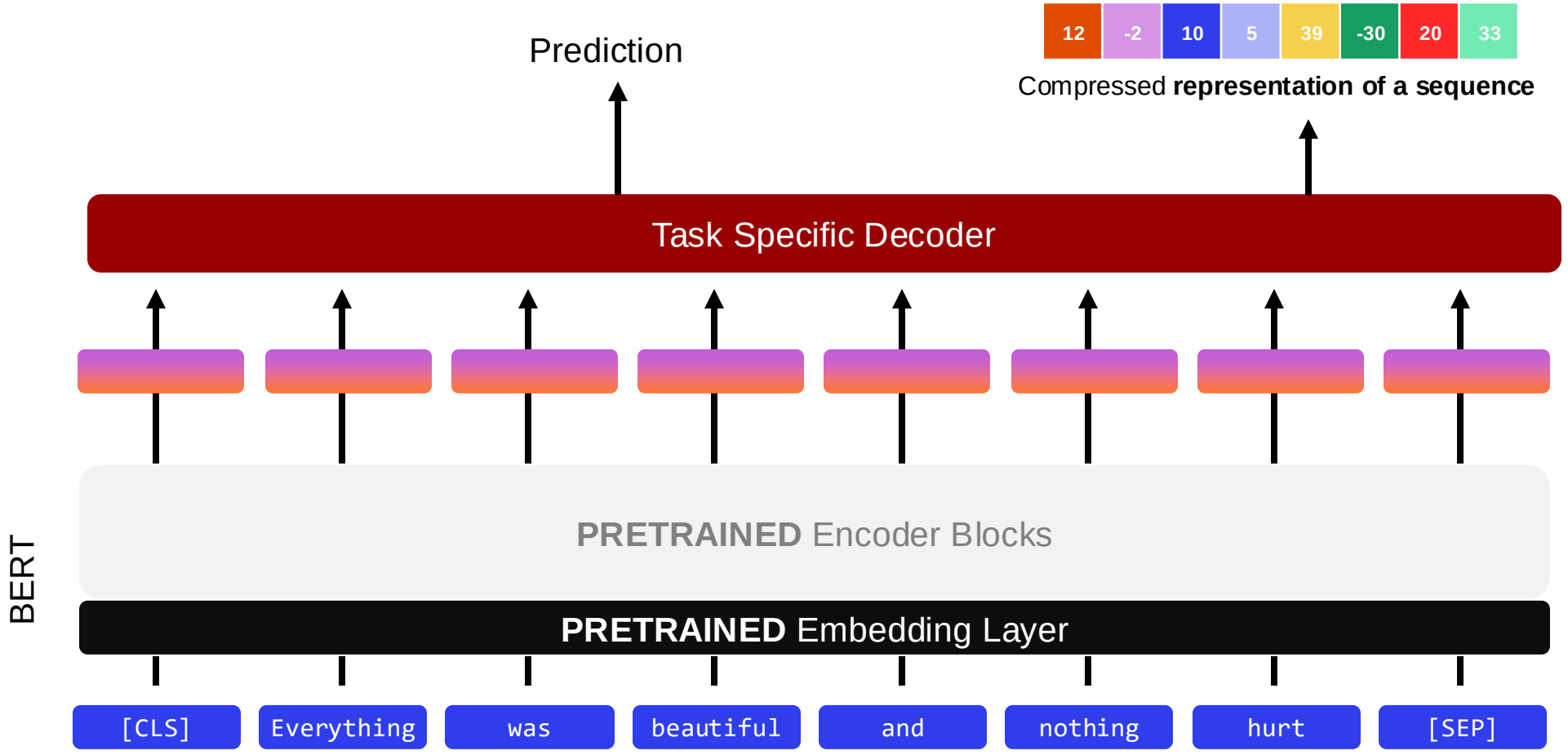
BERT: Pretraining

[CLS] usually has some task assigned

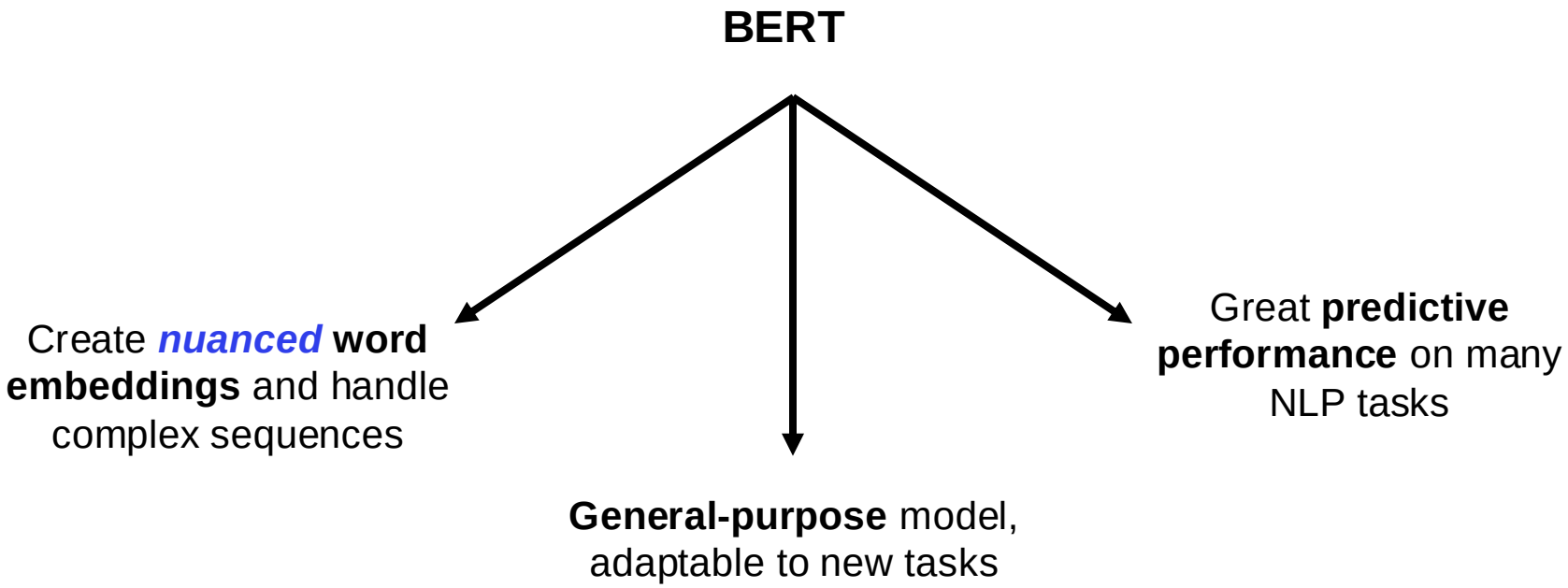
Ask model to predict: "What was originally there?"



BERT: Finetuning



Transformer-based Models

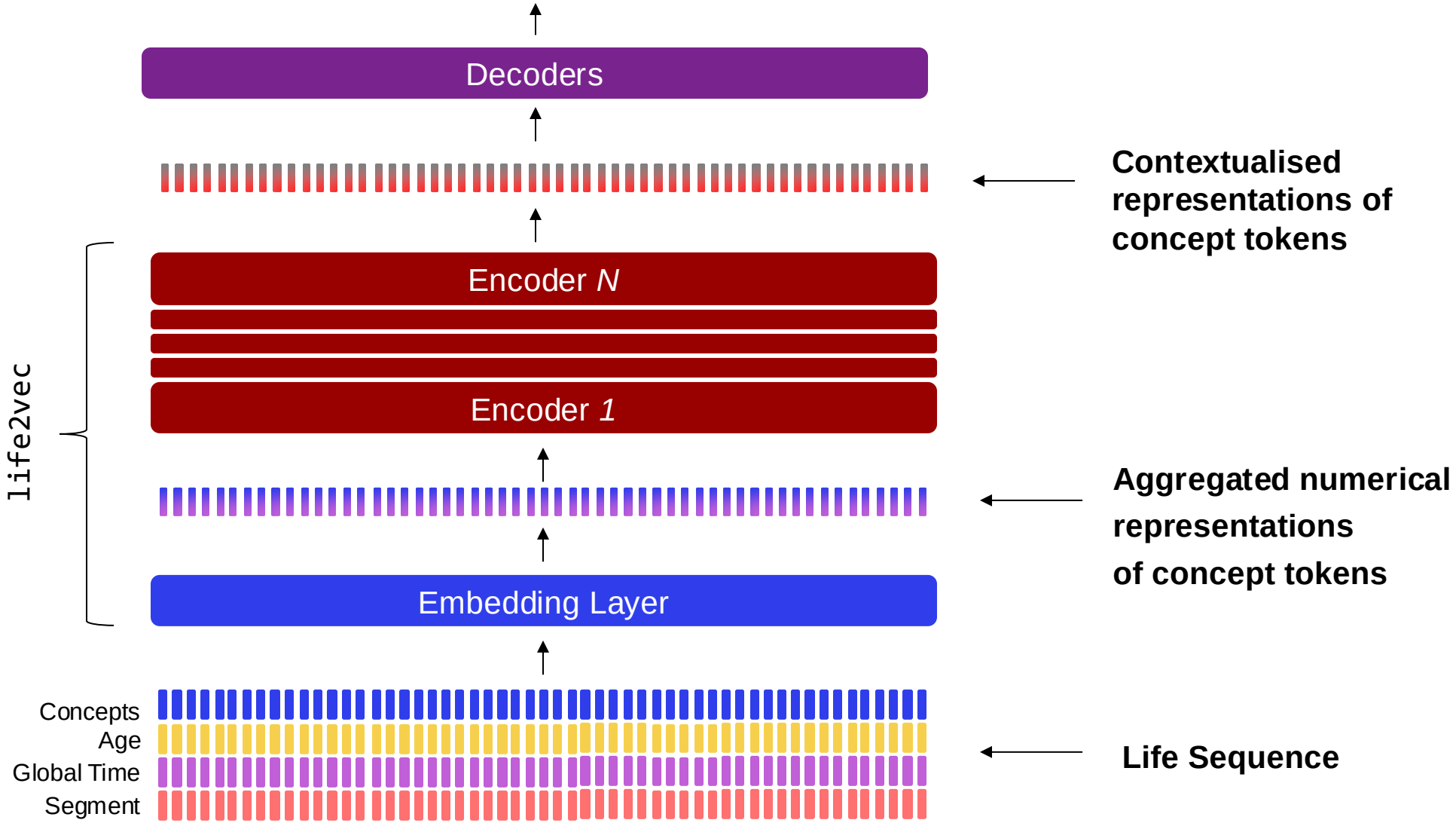


LIFE2VEC

Adapts BERT for life-sequences

Northeastern University

life2vec: Adaptation of BERT

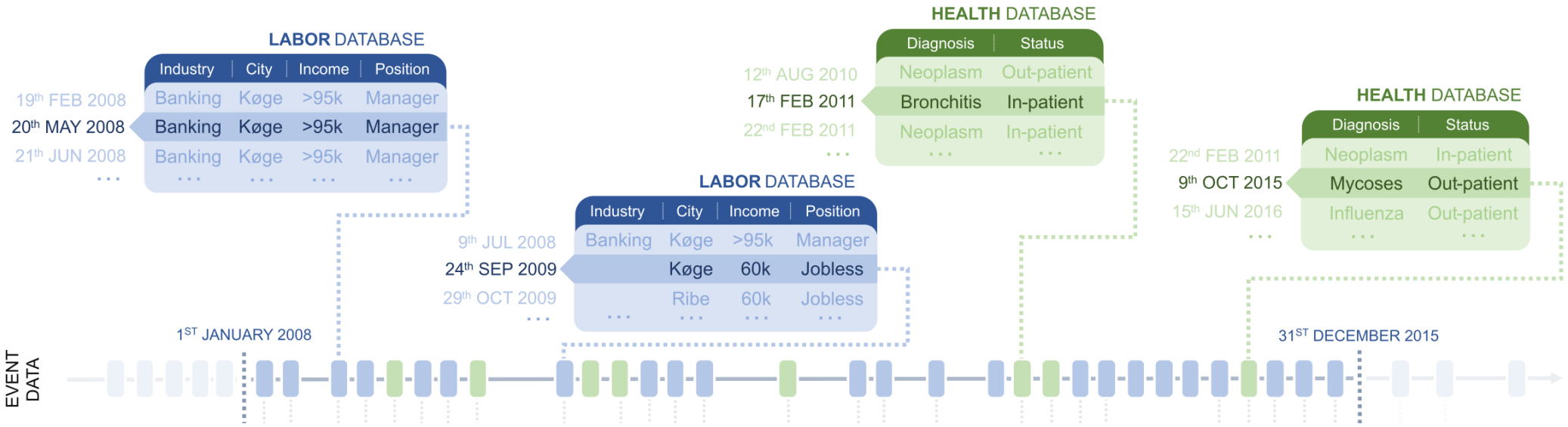


Part III

Socio-economic and health language

Northeastern University

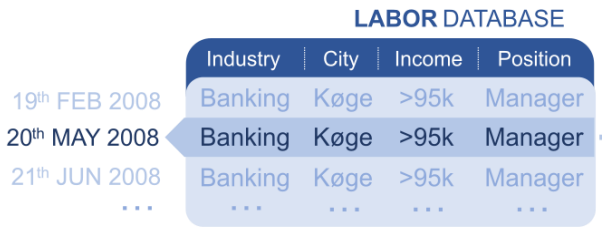
Unfolding the data



Tabular to Textual Representation?

** slightly simplified overview*

Forming a Language



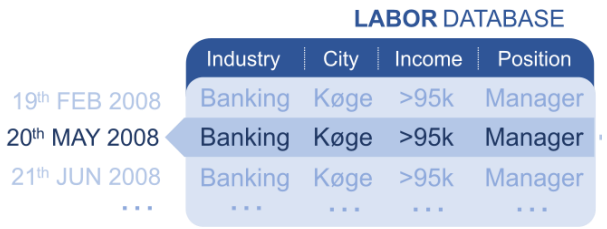
Convey the content in a spoken language

"In May 2008, Riley received >95k as a manager in Bank."

Language allows for super flexible and nuanced communication

Northeastern University

Forming a Language



Convey the content in a spoken language

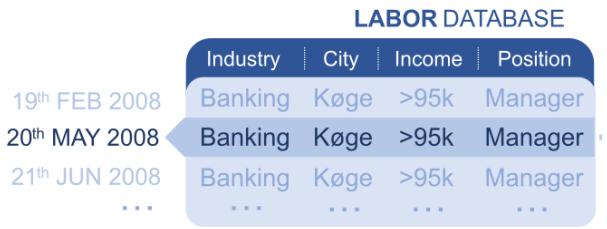
"*In May 2008, Riley received >95k as a manager in Bank.*"

Language allows for super flexible and nuanced communication

Not all of the structure in the English language is of interest to us

Northeastern University

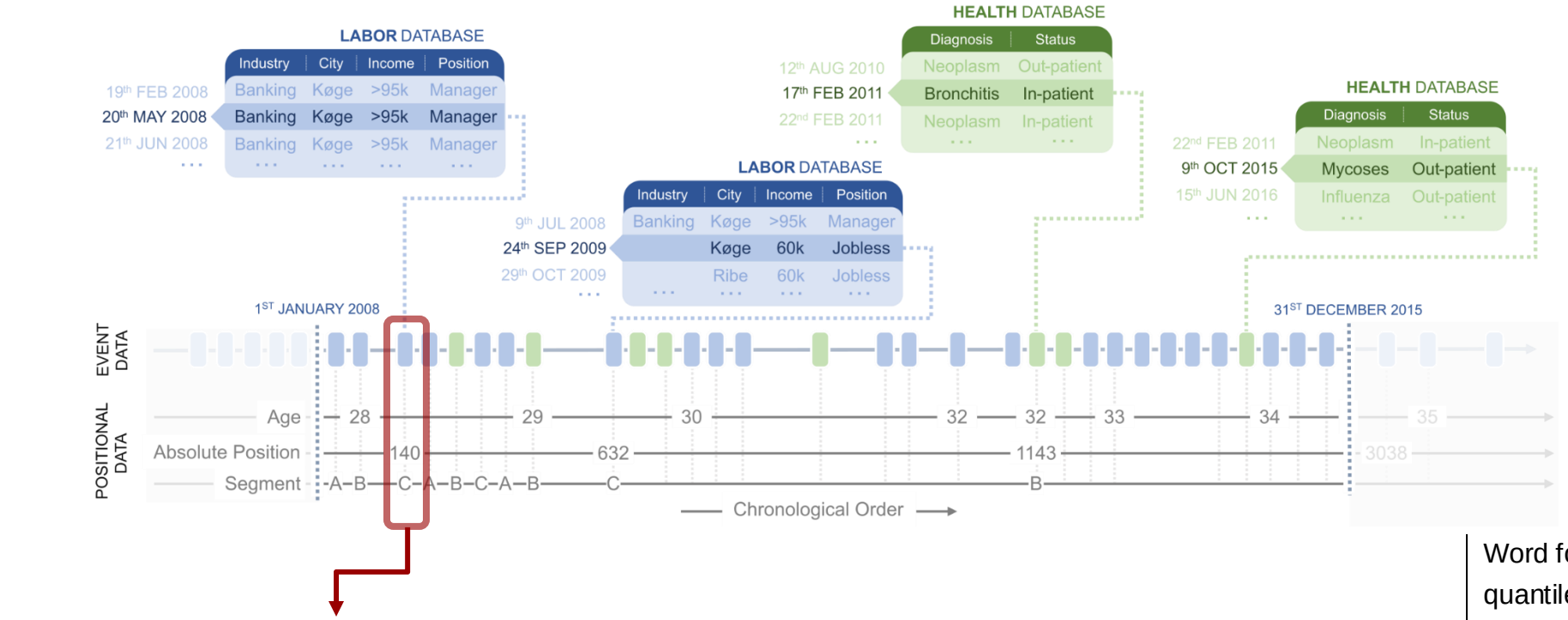
Forming a Language

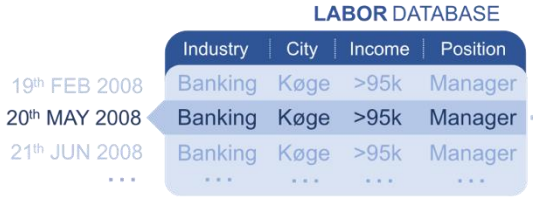


Convey the content in an artificial symbolic language

[IND_6419] [MUN_259] [INCOME_78] [POS_1221] [SEP]

Vocabulary consists of all the possible categories that any of the variable can take





1. Describe the content of the record

[IND_6419] [MUN_259] [INCOME_78] [POS_1221] [SEP]

Word for the quantile of the *income*

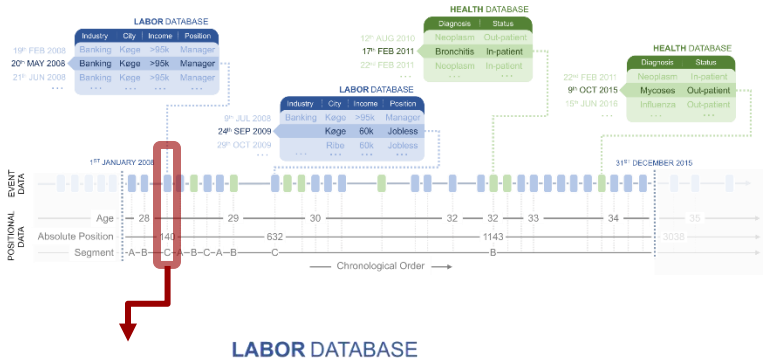
A “dot”

Word for the *banking* based on codes in “Danish Industries” system

Word for the municipality

Word for a *manager*, based on codes in ISCO-08

** slightly simplified overview*



1. Describe the content of the record

[IND_6419] [MUN_259] [INCOME_78] [POS_1221] [SEP]

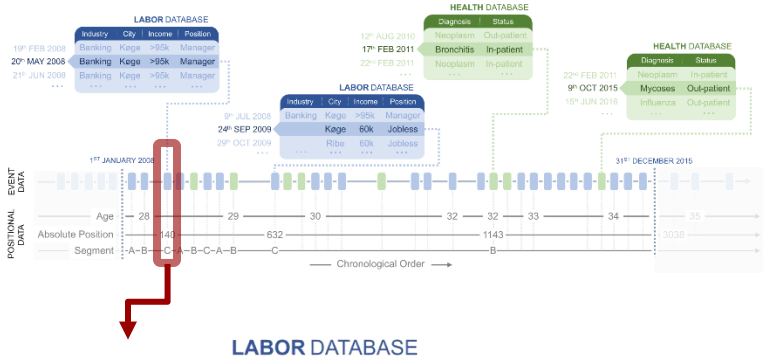
2. Extract positional information about the event

Age: 28 ← age at the time of the event

Global timestep: 140 ← number of days since 1st Jan 2008

Segment: C ← additional sentence identifier

** slightly simplified overview*



1. Describe the content of the record

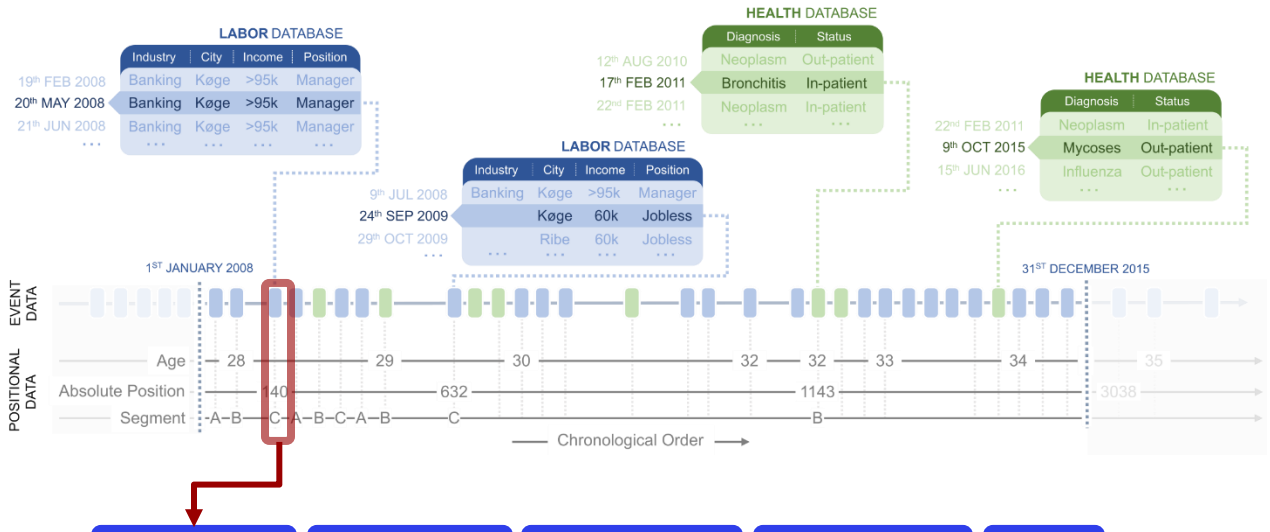
2. Extract positional information about the event

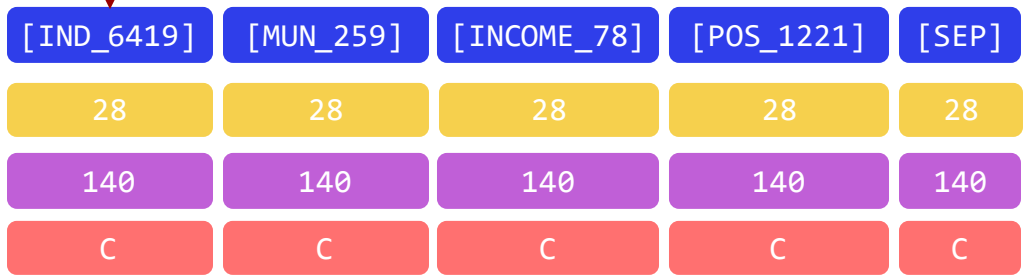
Age: 28
Global timestep: 140
Segment: C

3. Assign this information to tokens

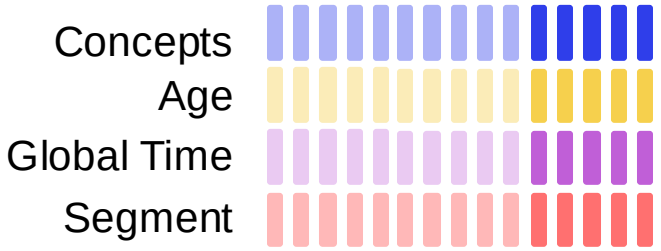
[IND_6419]	[MUN_259]	[INCOME_78]	[POS_1221]	[SEP]
28	28	28	28	28
140	140	140	140	140
C	C	C	C	C

** slightly simplified overview*

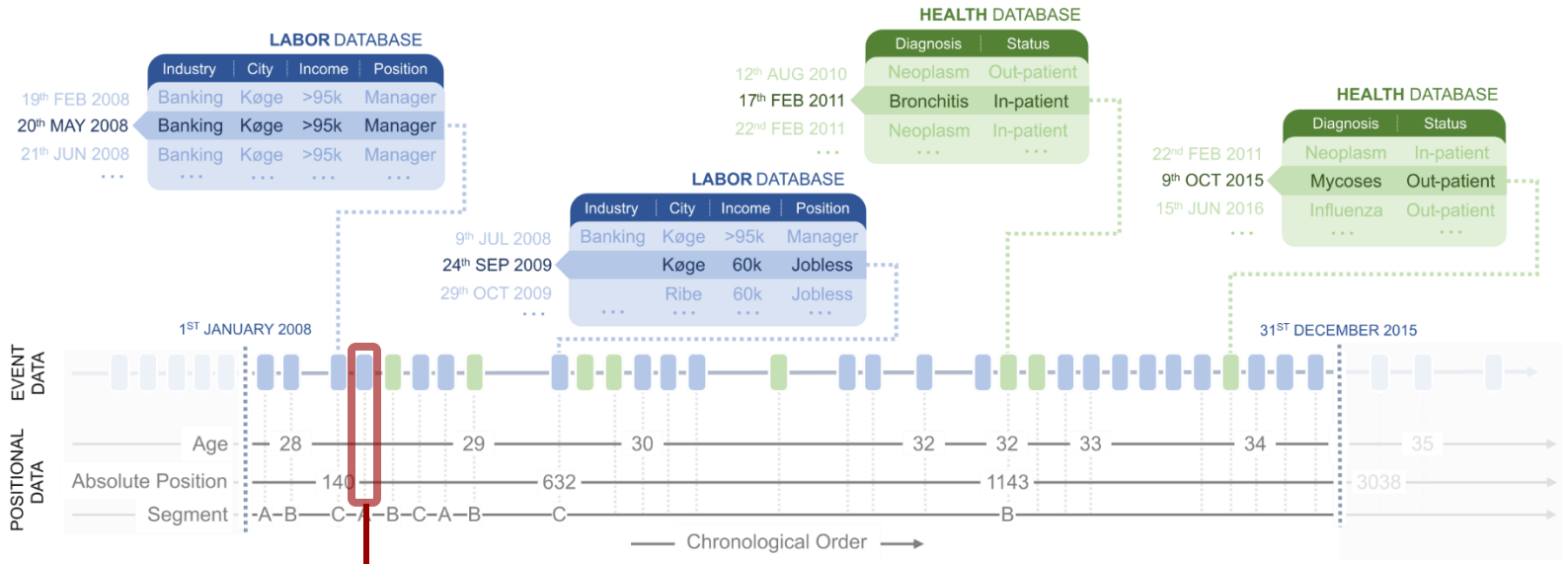


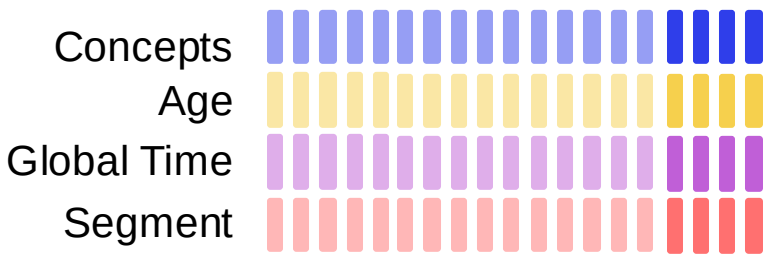


4. Insert data into the Life-Sequence (person document)

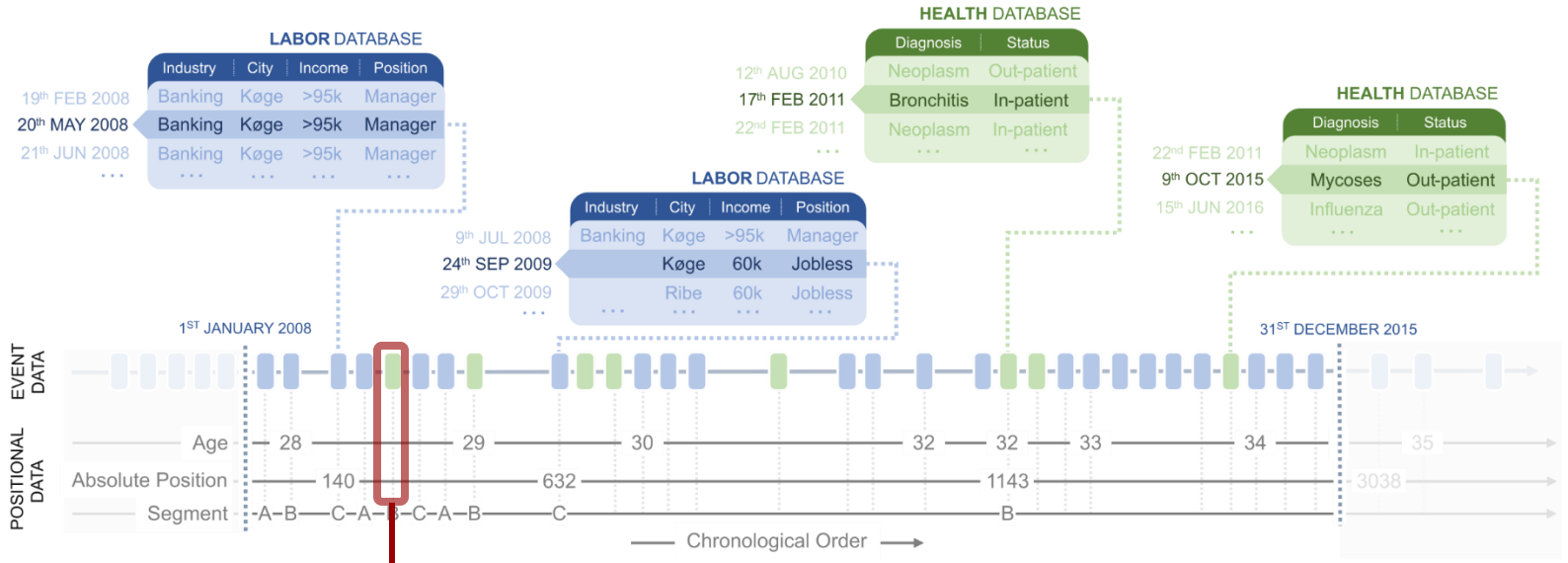


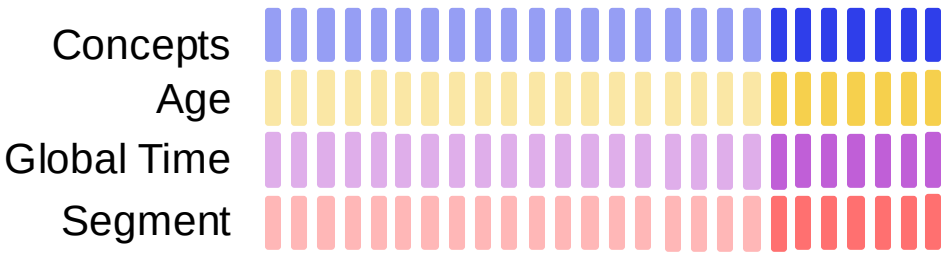
** slightly simplified overview*



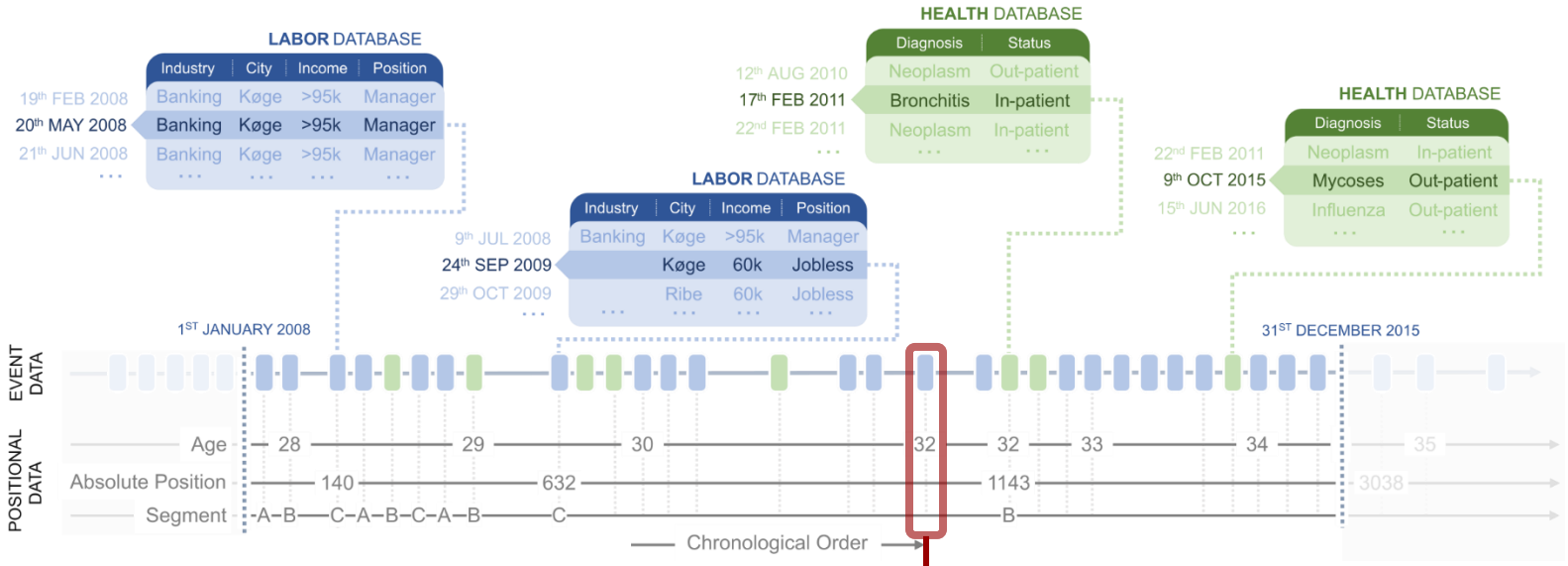


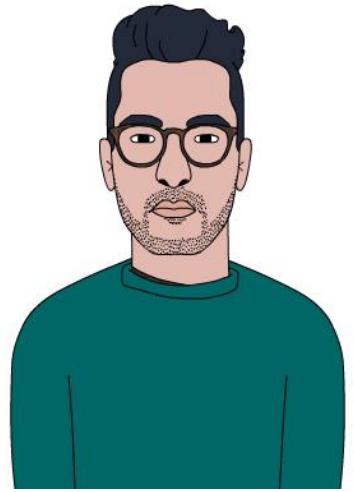
** slightly simplified overview*

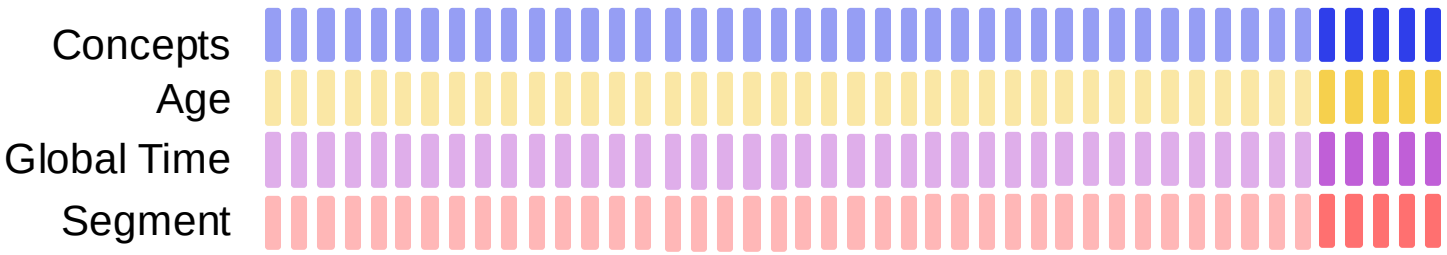




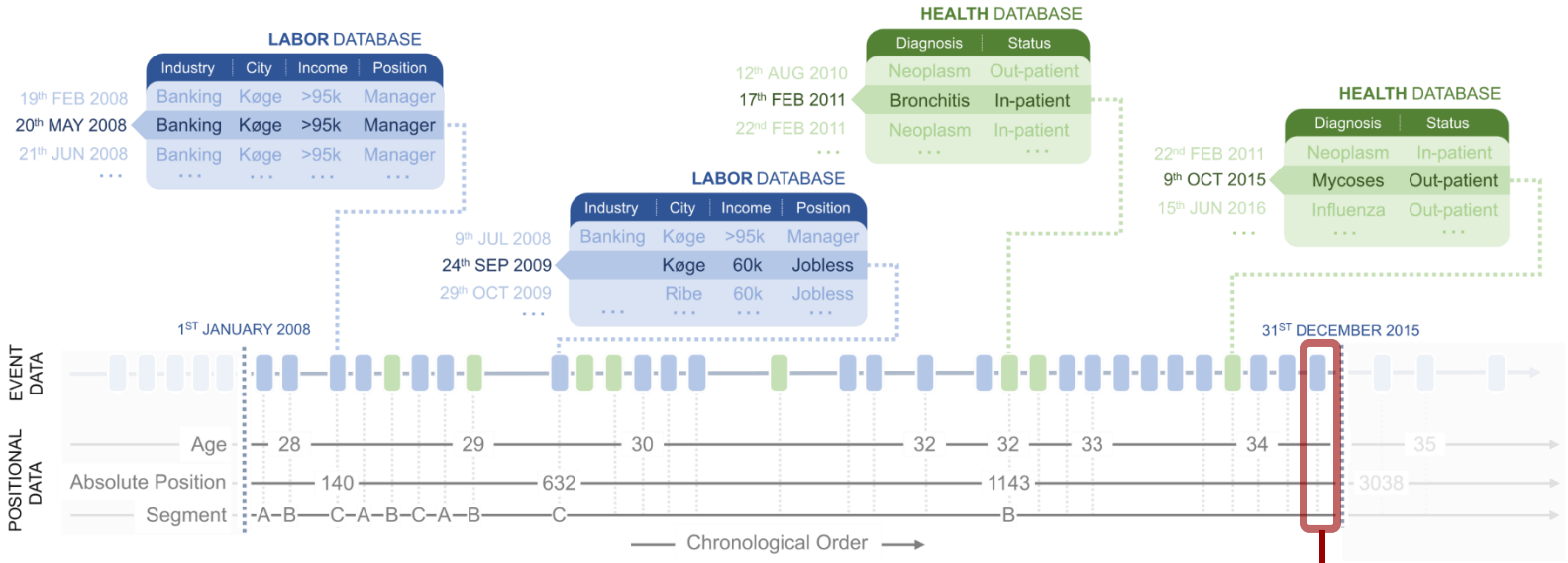
** slightly simplified overview*

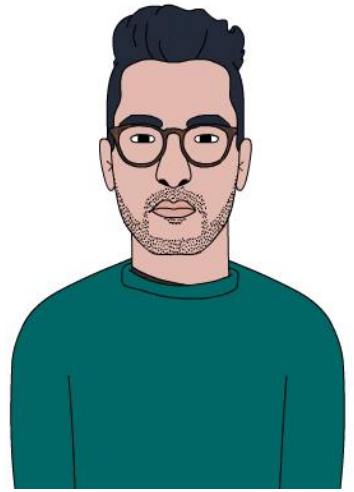


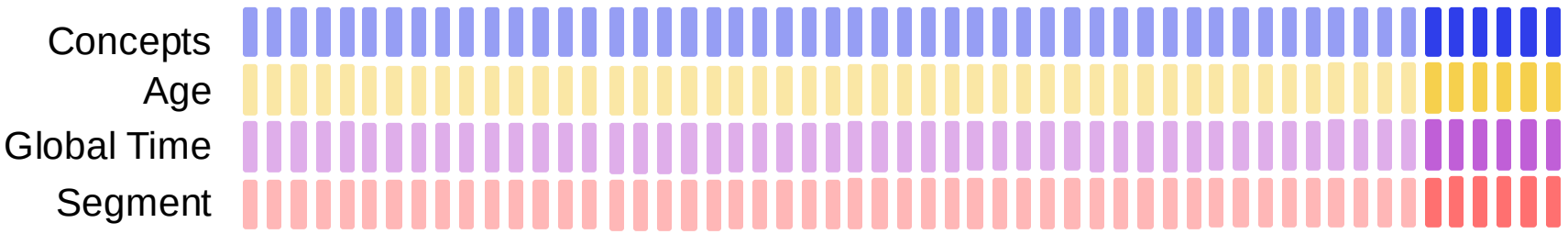




** slightly simplified overview*

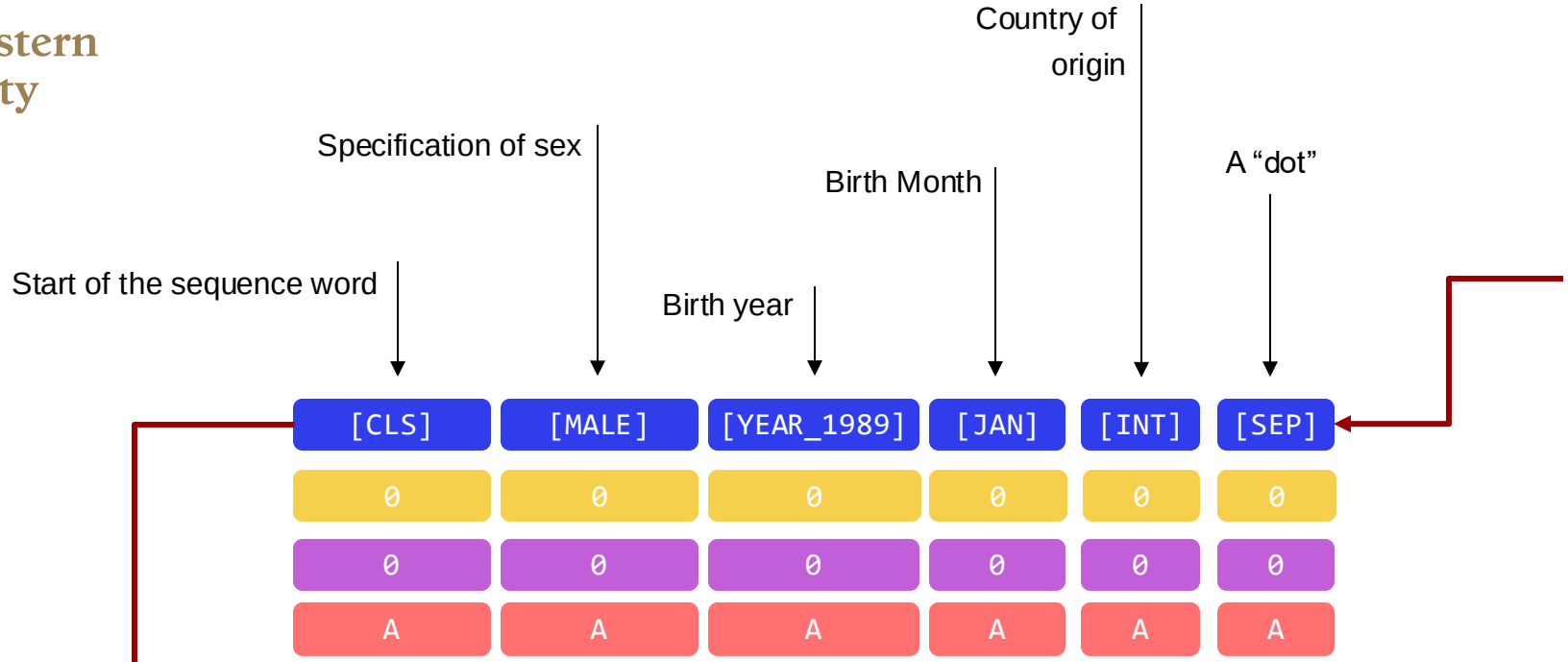


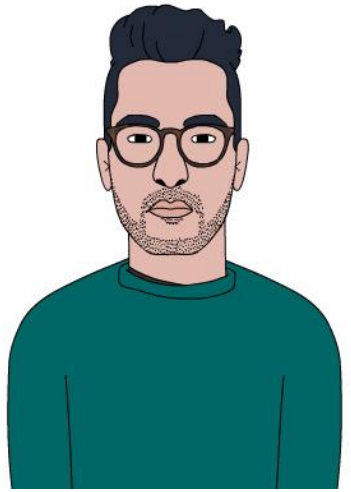




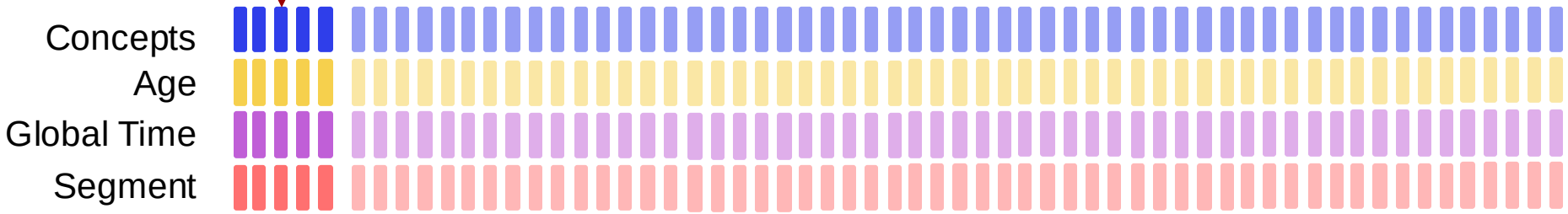
** slightly simplified overview*

Northeastern University





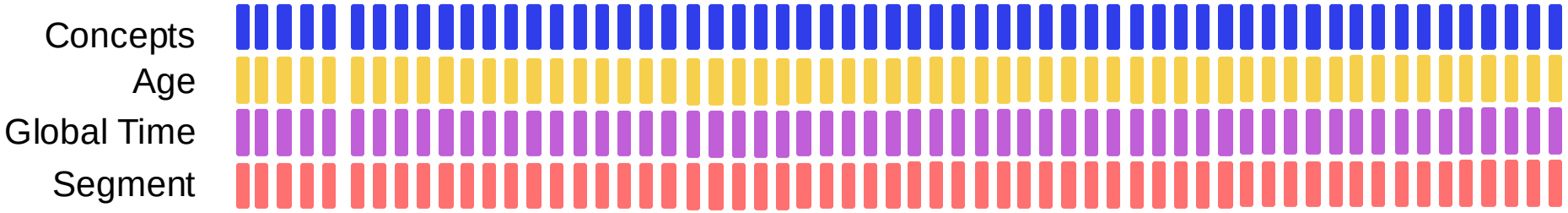
The “Background sentence”



** slightly simplified overview*

Northeastern University

Individual Life-Sequence



Input to the `life2vec` model

** slightly simplified overview*

Vocabulary

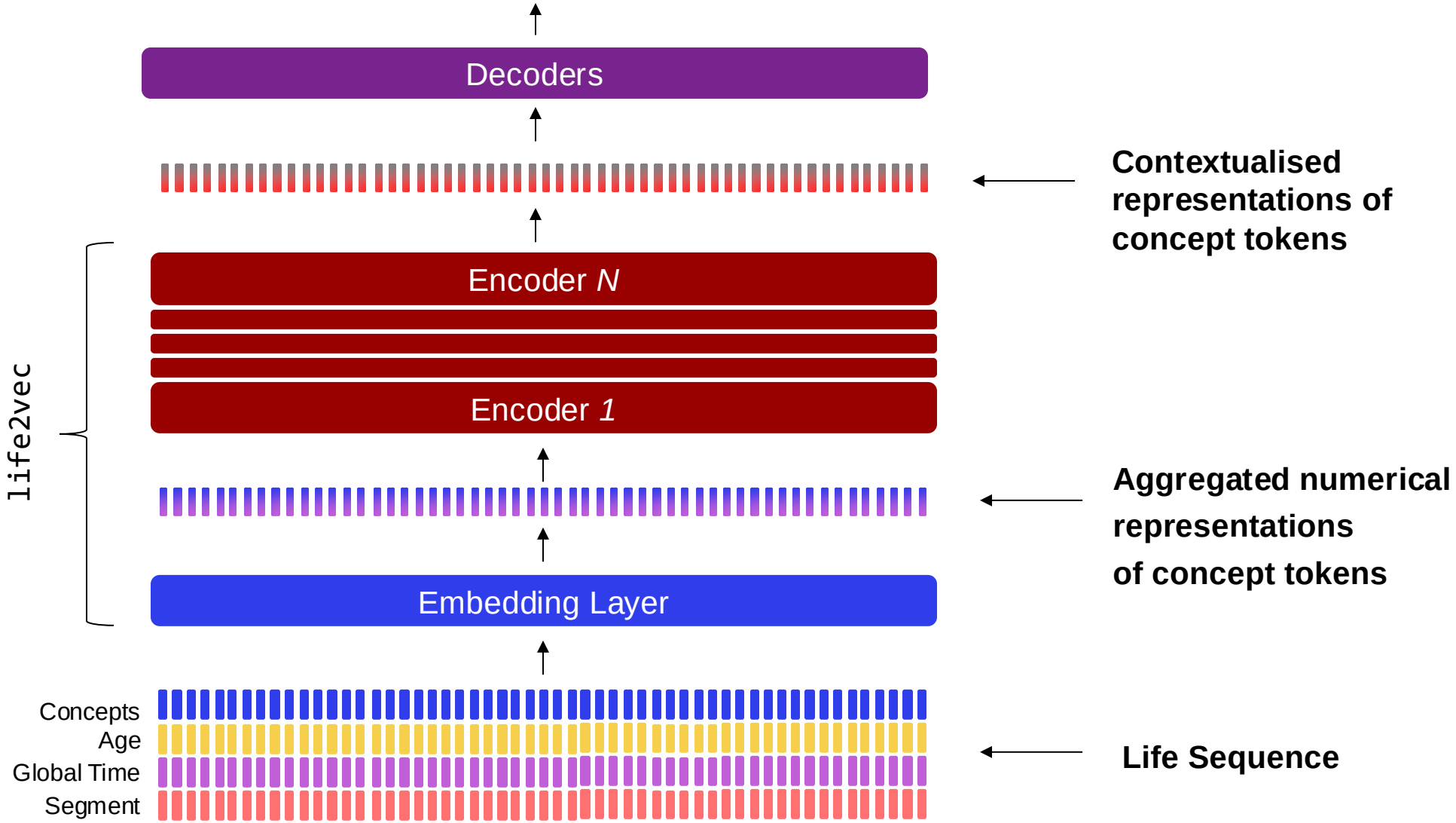
Type	Variables	# Categories	Encoding
Background Information	Sex	2 binary	Male, Female
	Birth Month	12	Jan-Feb
	Birth Year	45	1946-1991
	Country of Origin	2 binary	National or International
Labour Records	Municipality of Residence	97	Danish municipality codes
	Tax Bracket	6	DST definitions
	Income Level	100	Quantile-based
	Labour Force Status	35	DST definitions
	Labour Force Status (Modification)	58	DST definitions
	Labour-Force-Interval	10	Quantile based
	Industry Area (Company)	290	DB07
	Job type	359	ISCO-08
	Enterprise Type (Company)	15	ESA-2010
Health Records	Diagnosis	704	ICD-10
	Urgency	3	Urgent, Non-Urgent, Emergency
	Patient Type	2	In-, out- patient
Special	Special	10	`[PAD] ... [UNK]`

Part IV

life2vec: capturing the structure

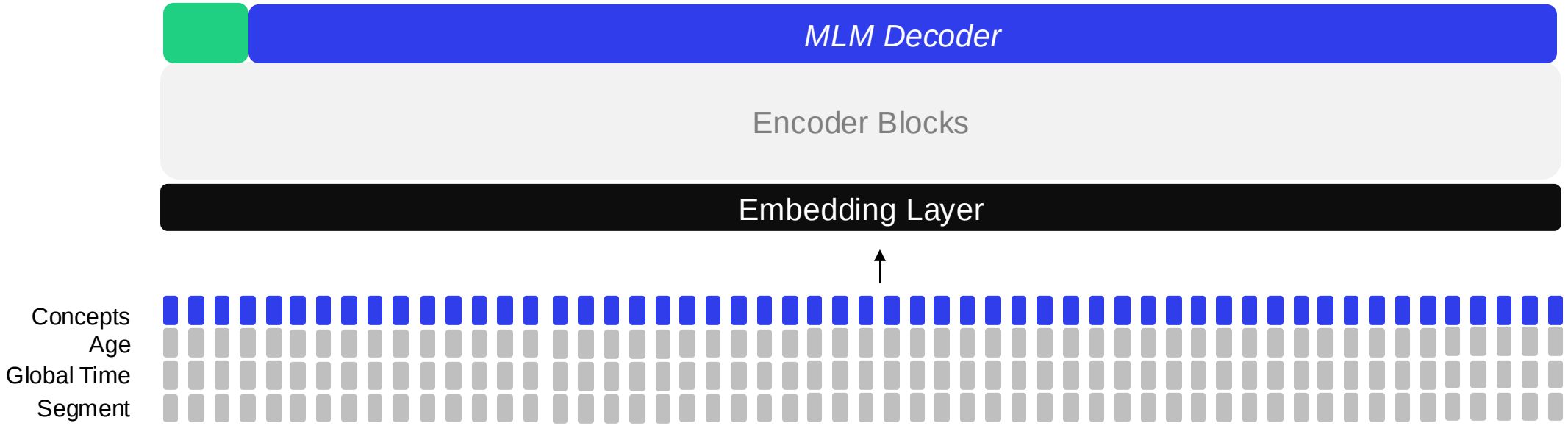
Northeastern University

life2vec pipeline

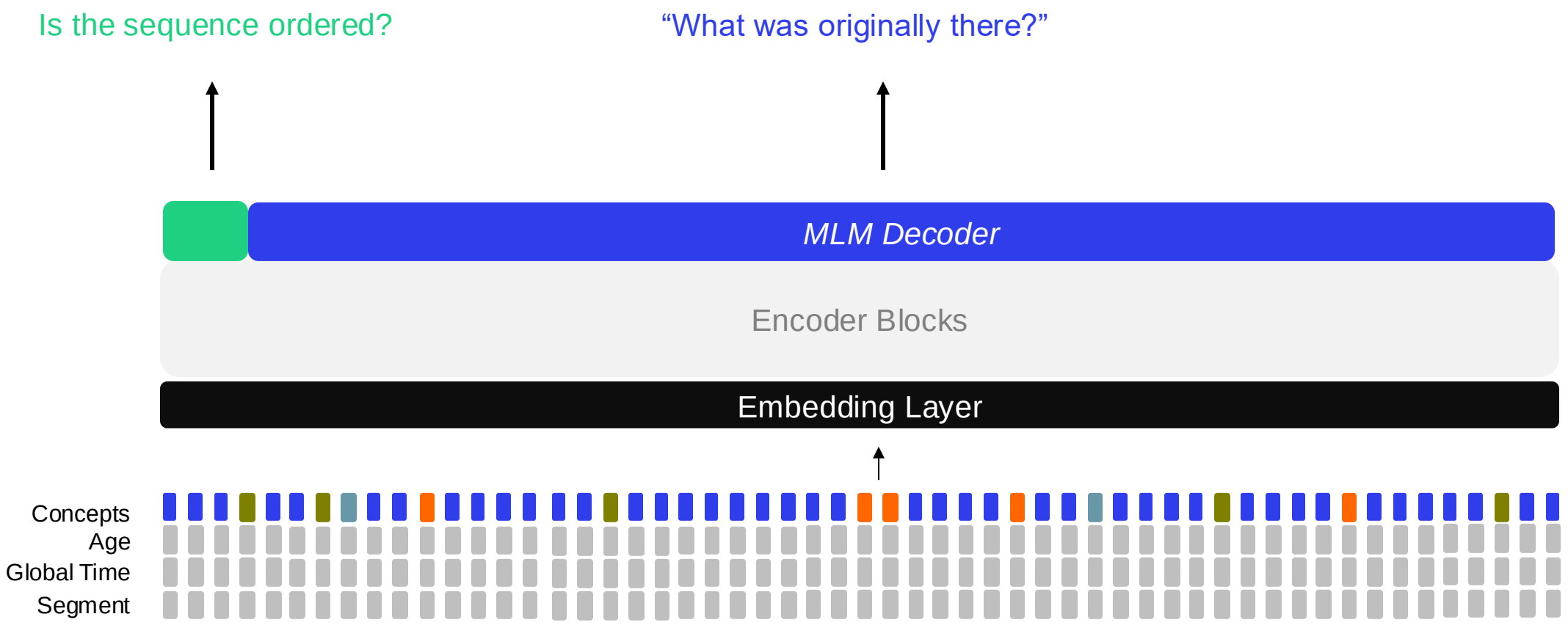


life2vec: pre-training

- **Mask 30%** of tokens (not including **[PAD],[SEP],[CLS]**):
  - 10% unchanged
  - 10% substituted with random tokens
  - 80% substituted with the **[MASK]** token



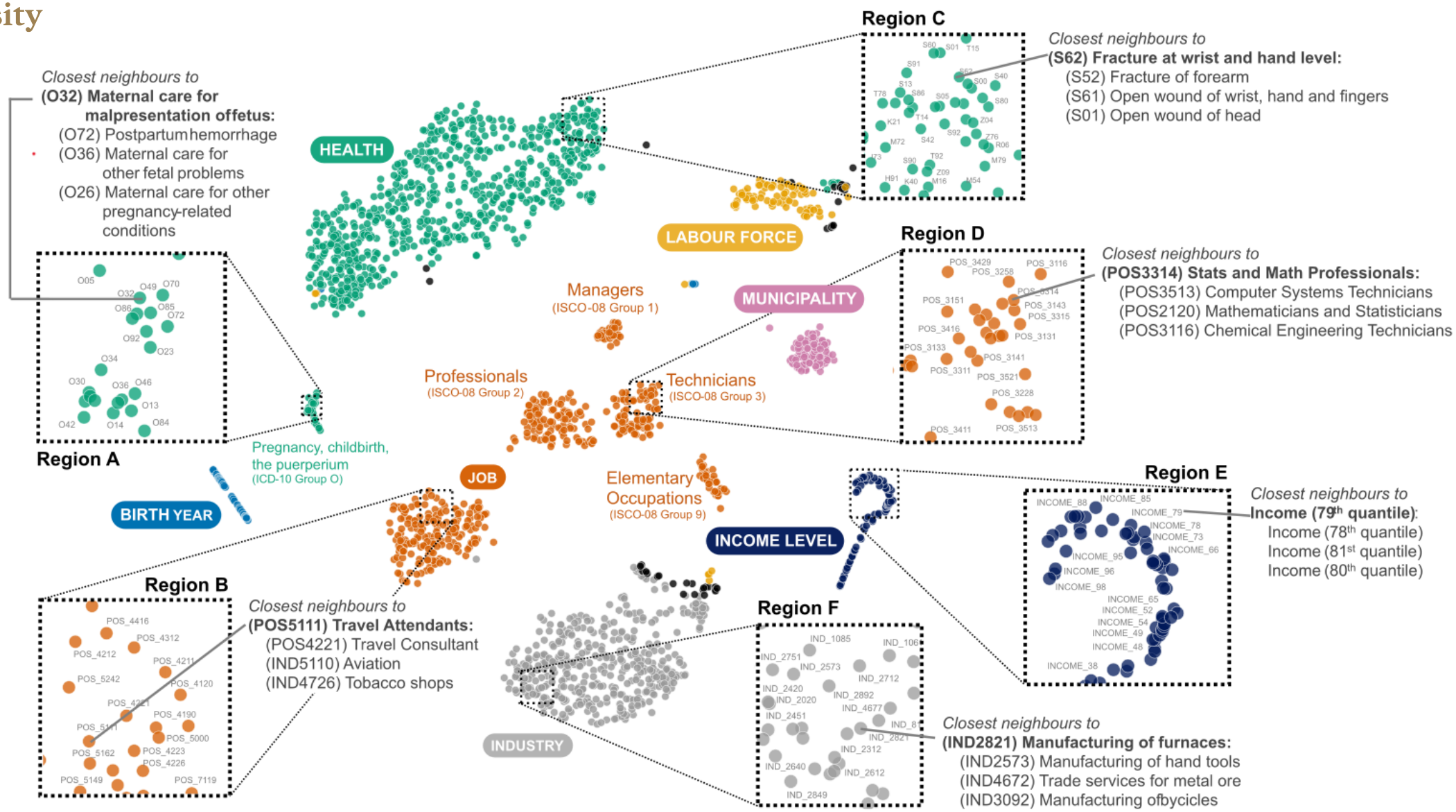
life2vec: pre-training



What did our model learn on pretraining?

Northeastern University

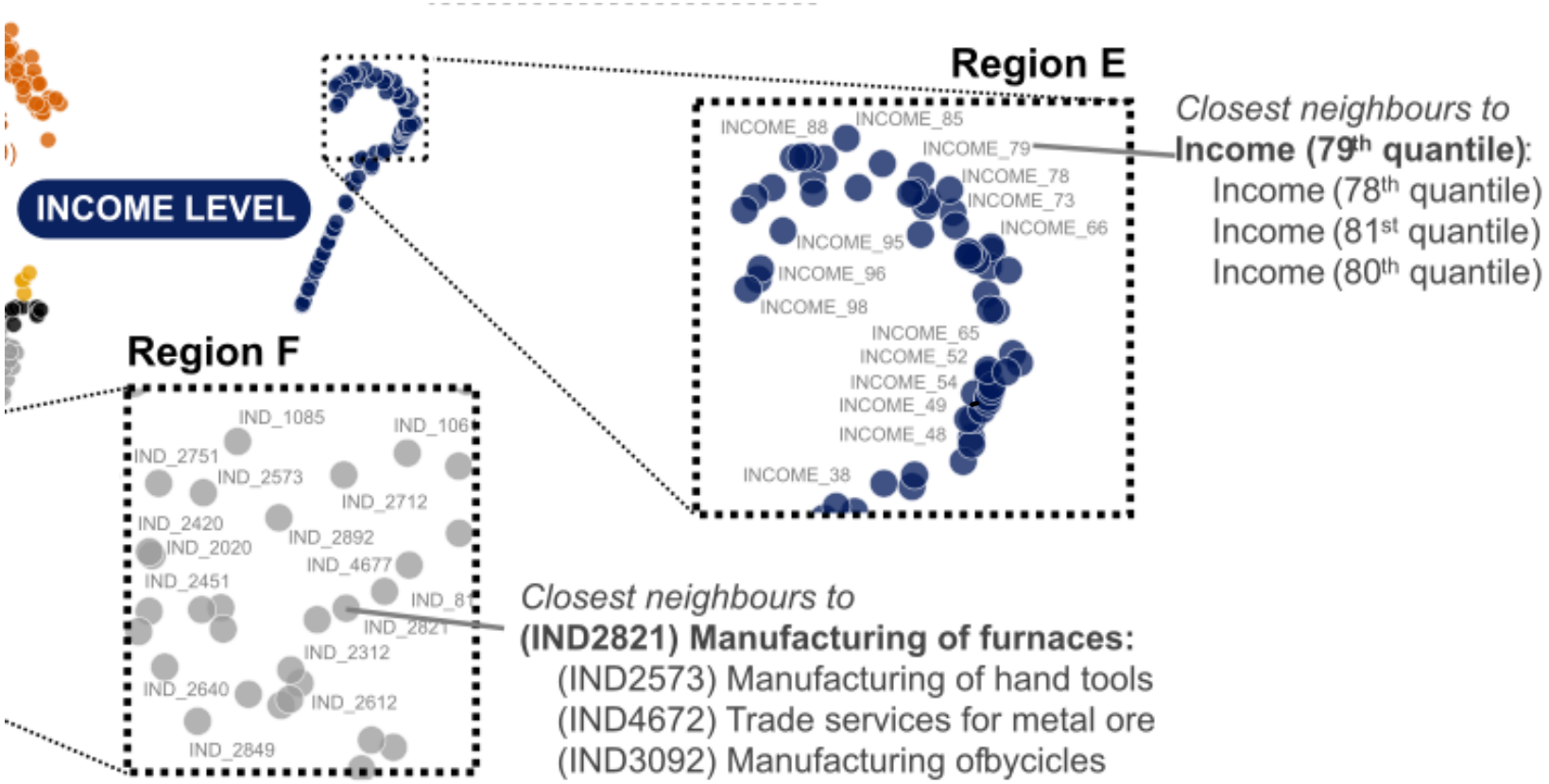
Space of Concept Tokens (with PaCMAP)



Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1-14.

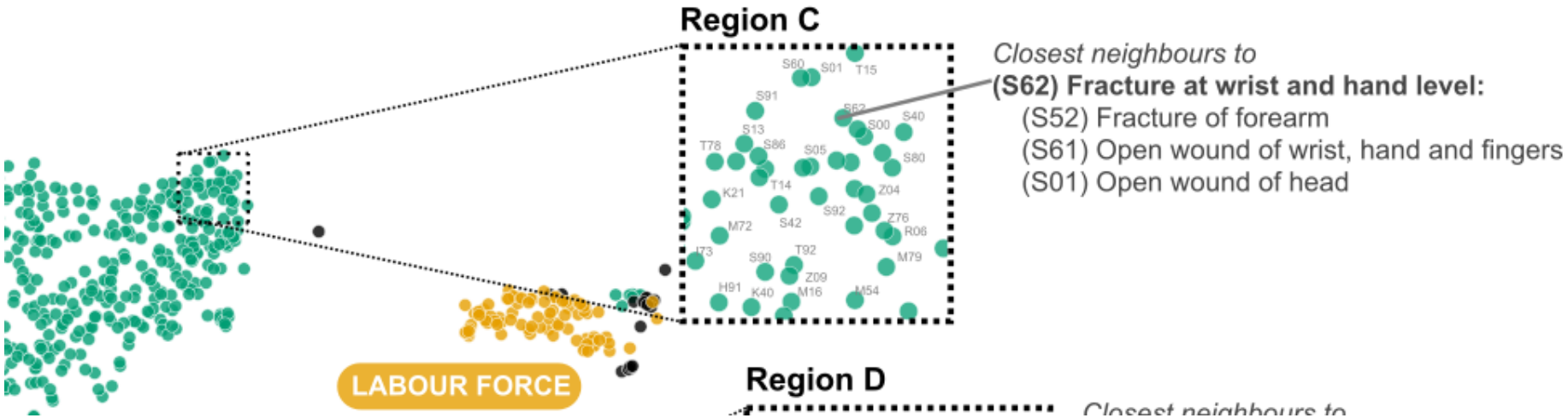
Northeastern University

Space of concept tokens (with PaCMAP)



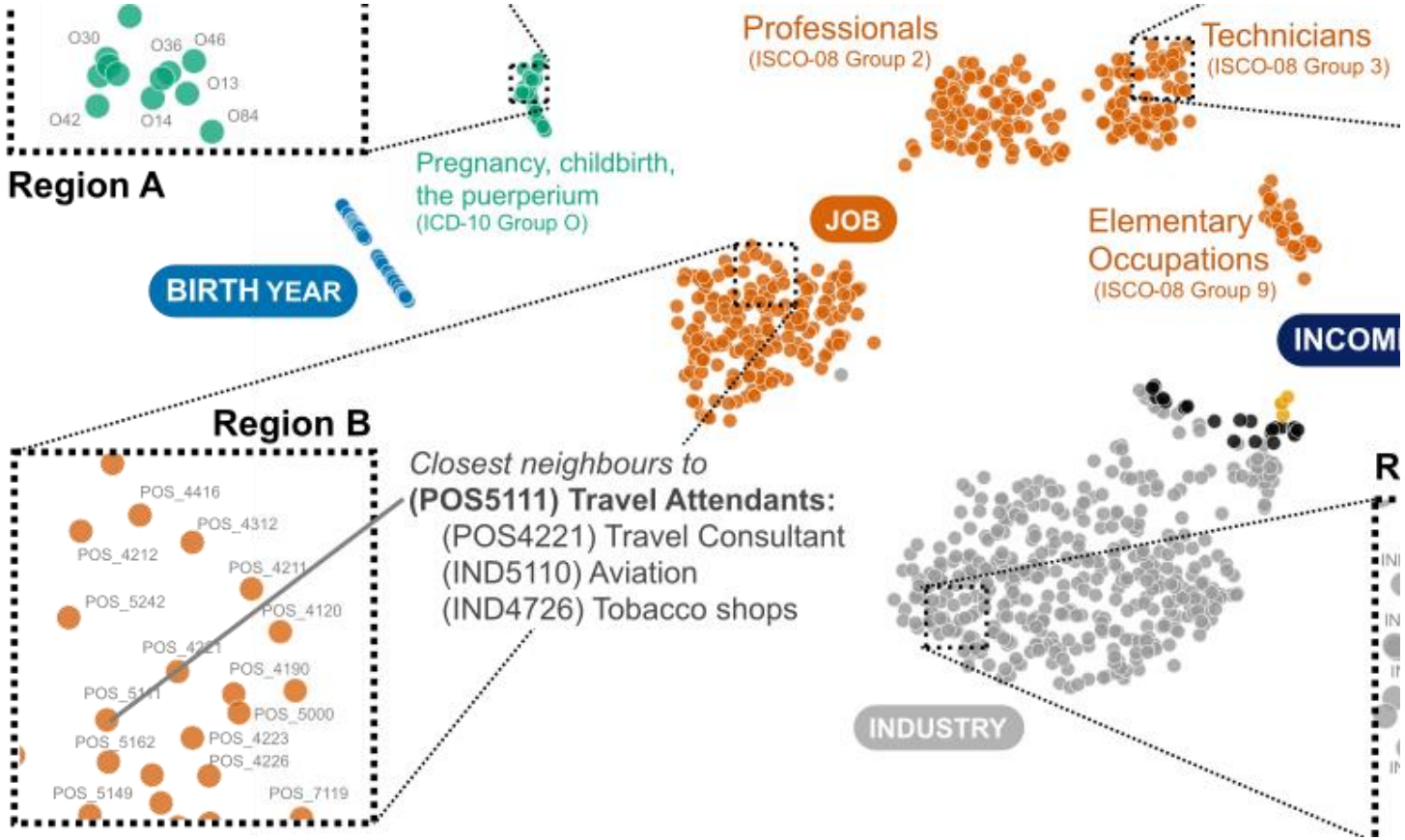
Northeastern University

Space of concept tokens (with PaCMAP)



Northeastern University

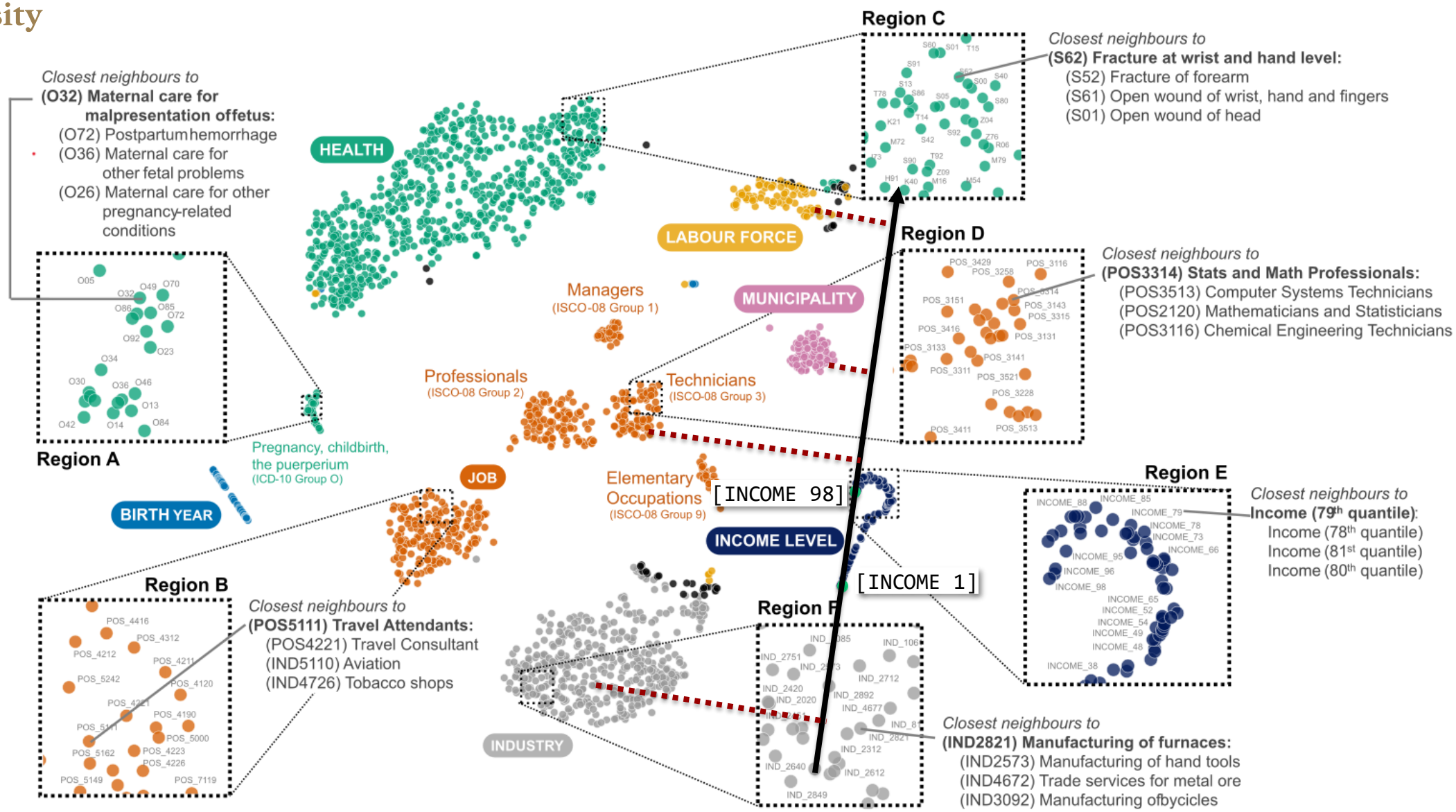
Space of concept tokens (with PaCMAP)



Visually structure corresponds to the structure of the variables

Northeastern University

Space of Concept Tokens (with PaCMAP)

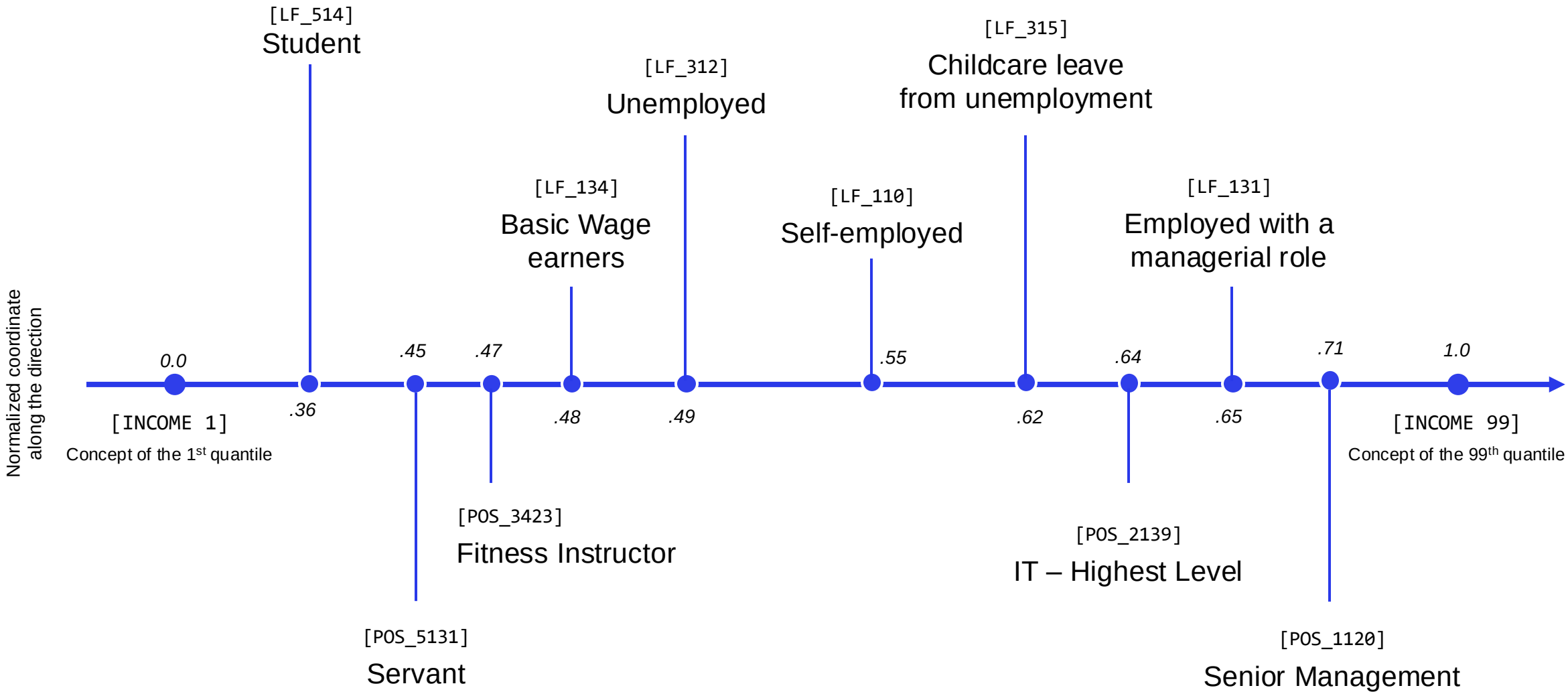


Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1-14.

Northeastern University

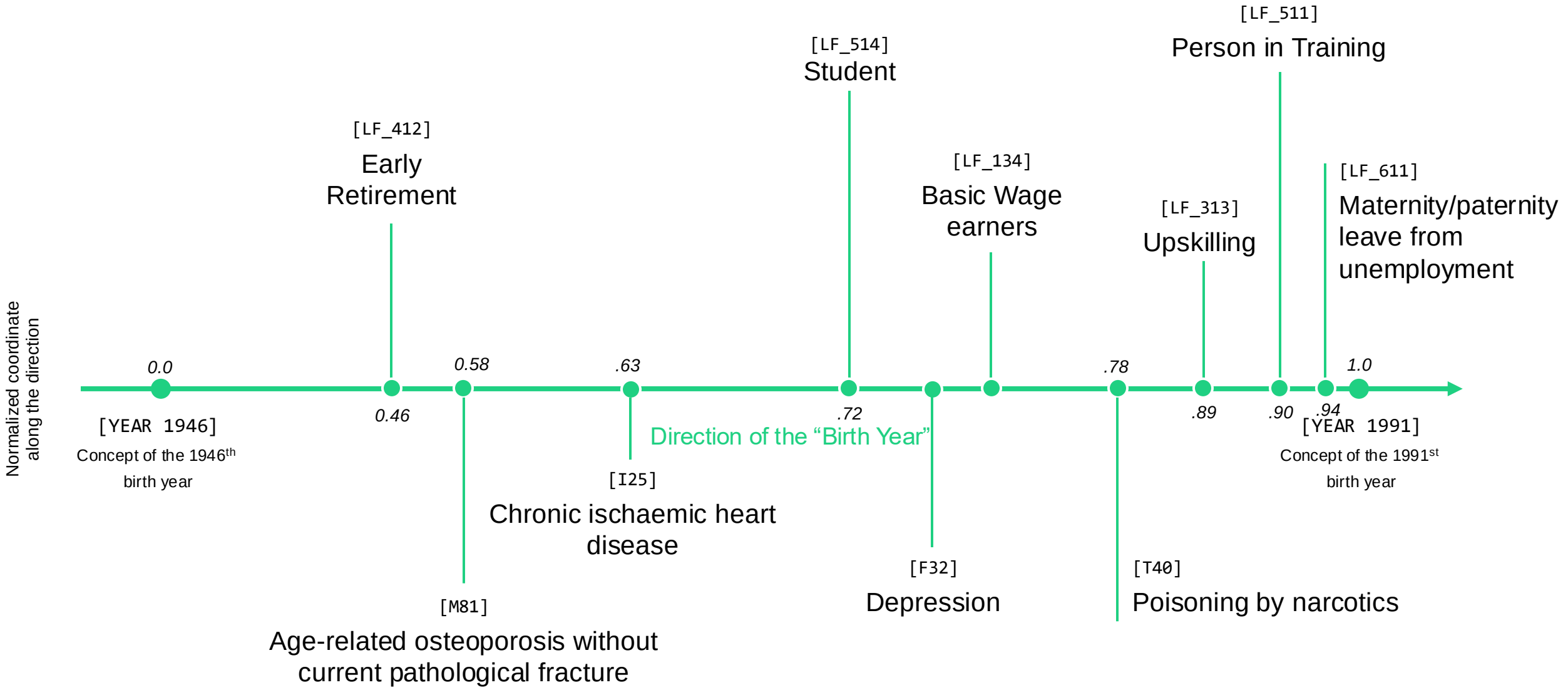
LF – Labor Force Status
POS – Prof. Position

Projection to "*Income*" Direction



Northeastern University

Projection to "*Year*" Direction



Northeastern University

Projection to "*Occupation*" Direction

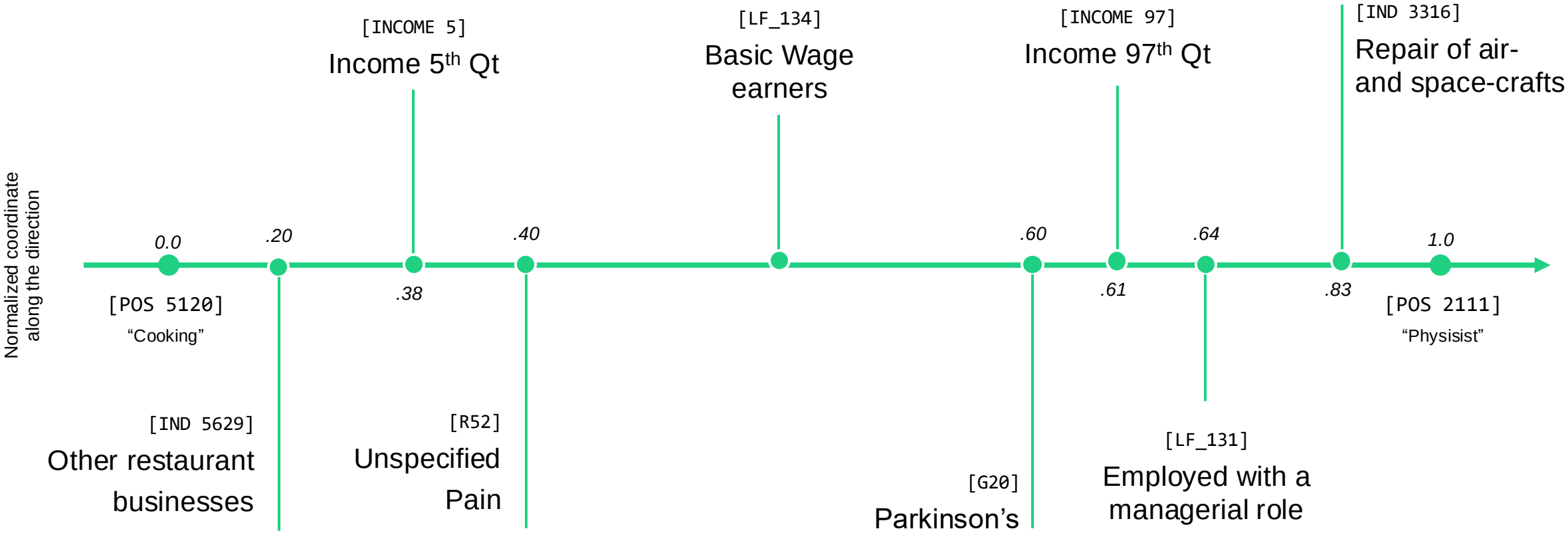
The opposite job of a chef and head cook is a physicist.

	Chefs and Head Cooks use these skills the most		Physicists use these skills the most
1	Management of material resources	1	Physics
2	Management of financial resources	2	Mathematical reasoning
3	Management of personnel resources	3	Number facility
4	Coordination	4	Ability to organize groups in different ways
5	Negotiation	5	Information ordering
6	Monitoring	6	Mathematics
7	Time management	7	Oral comprehension
8	Persuasion	8	Mathematics
9	Social perceptiveness	9	Originality
10	Learning strategies	10	Speech clarity

(n.d.). What Is Your Opposite Job? The New York Times. Retrieved March 11, 2024, from https://www.nytimes.com/interactive/2017/08/08/upshot/what-is-your-opposite-job.html

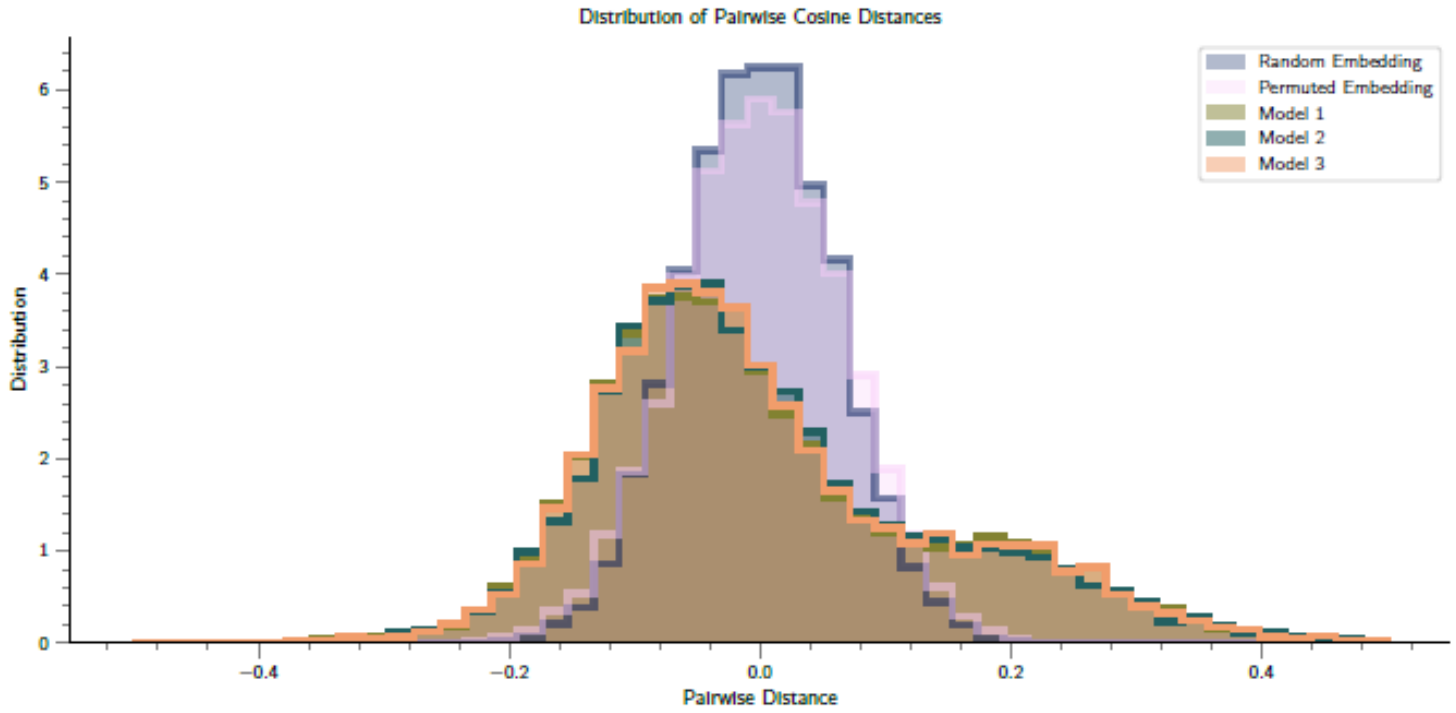
Northeastern University

Projection to "*Occupation*" Direction



Northeastern University

Concept Space Robustness: Permutation Test



Models trained on **separate datasets** and with **different initialization**

Model Comparison	Spearman's ρ
D^1 vs D^2	.668
D^1 vs D^3	.660
D^2 vs D^3	.661
D^1 vs D^R	-.001
D^1 vs D^{1p}	.000

rho > .6 (Strong *monotonic* correlation)[1]

1. Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768.

What does it tell us?

- **Life2vec as proof of concept**
  - **Algorithms understand the textual representation of life-sequences**
  - **Transformers can capture structure in such a language**

Study the dynamic within the data source

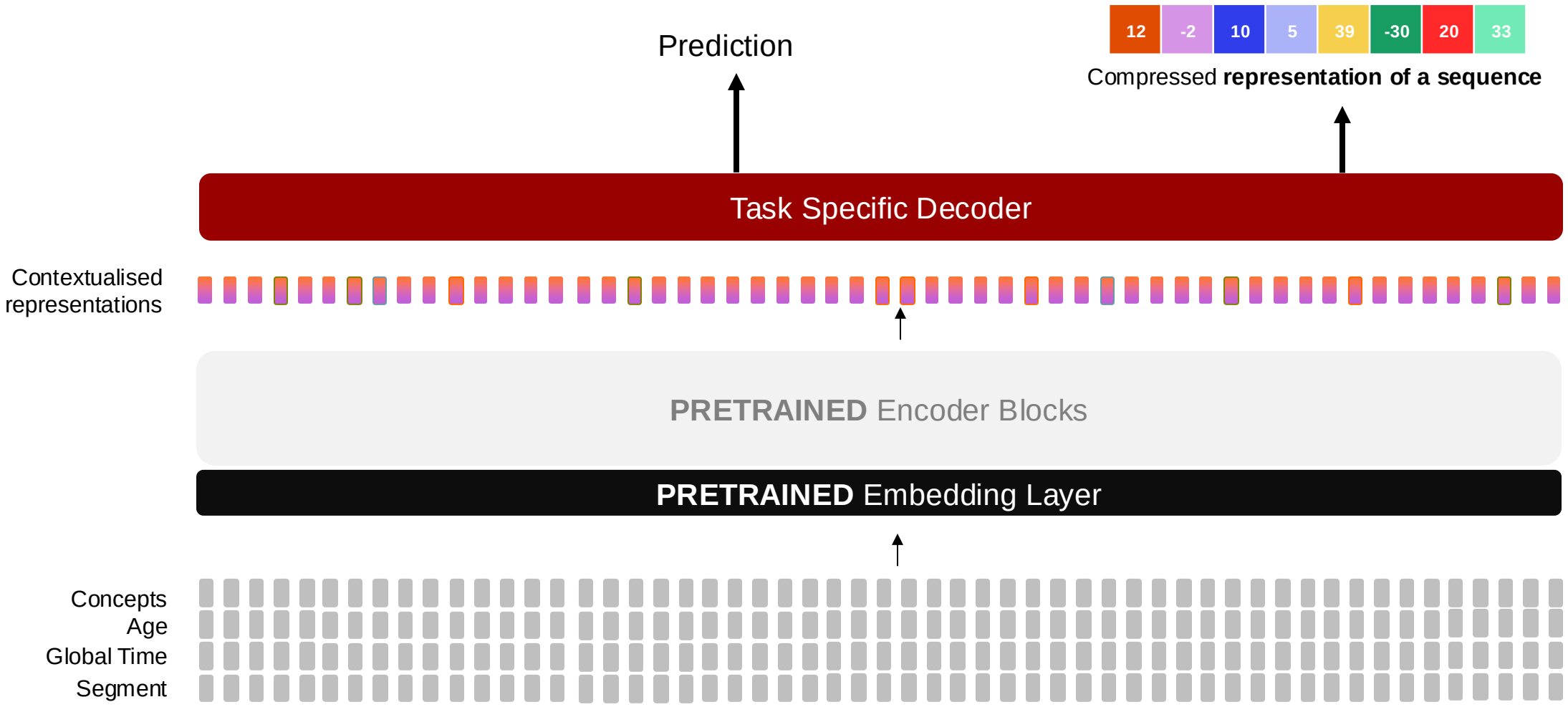
- Health and labor modelled in one space
- Can use embedding space to analyse relationships between categories

Part V

life2vec as a foundation model

Northeastern University

life2vec: finetuning

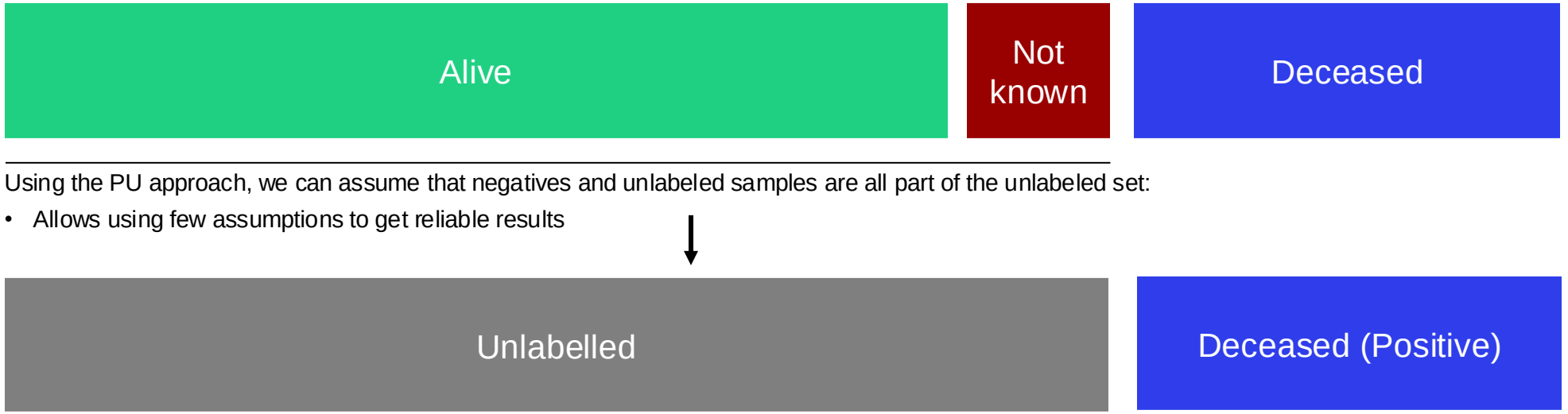


Wang, C., Pu, J., Xu, Z. and Zhang, J., 2021, July. Asymmetric Loss for Positive-Unlabeled Learning. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

Early Mortality Prediction

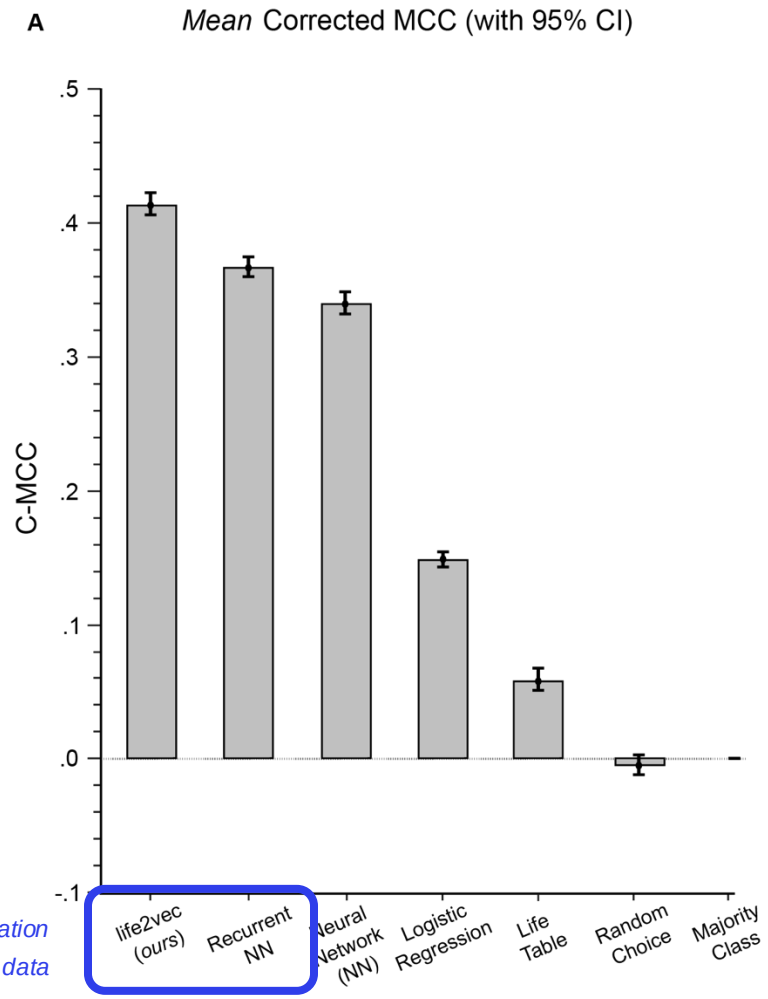
- **Task: "Is a person going to be deceased within the next 4 years after 31st December 2015?"**
  - Split people into ones who are marked as dead, and all others
  - Some people do not have "a label".
    - This is a Positive Unlabelled (PU)-Learning Problem

Why PU Learning? (Mortality Example)



Northeastern University

Early Mortality Prediction



Use textual representation of the data

True Labels

Predicted Labels	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$$\widehat{\text{mcc}} = \frac{\text{tp} \times \text{tn} - \text{fp} \times \text{fn}}{\sqrt{(\text{tp}+\text{fp})(\text{tp}+\text{fn})(\text{tn}+\text{fp})(\text{tn}+\text{fn})}} = \frac{\hat{\pi}(1-\hat{\pi})(\hat{\gamma}\cdot(1-\hat{\eta}) - \hat{\eta}\cdot(1-\hat{\gamma}))}{\sqrt{\theta\hat{\pi}(1-\hat{\pi})(1-\theta)}}$$

$$\widehat{\text{mcc}}_{cr} = \sqrt{\frac{\hat{\pi}_{cr}\,(1-\hat{\pi}_{cr})}{\theta(1-\theta)}}\,(\hat{\gamma}_{cr} - \hat{\eta}_{cr})$$

$$\text{Recall} = \hat{\gamma} = \frac{\text{tp}}{\text{tp}+\text{fn}}$$

$$\text{FPR} = \hat{\eta} = \frac{\text{fp}}{\text{tn}+\text{fp}}$$

$$\text{Positive Class Prior} = \hat{\pi} = \frac{\text{tp}+\text{fn}}{\text{tp}+\text{fn}+\text{tn}+\text{fp}}$$

$$\text{Positive Predictions} = \theta = \frac{\text{tp}+\text{fp}}{\text{tp}+\text{fn}+\text{tn}+\text{fp}}$$

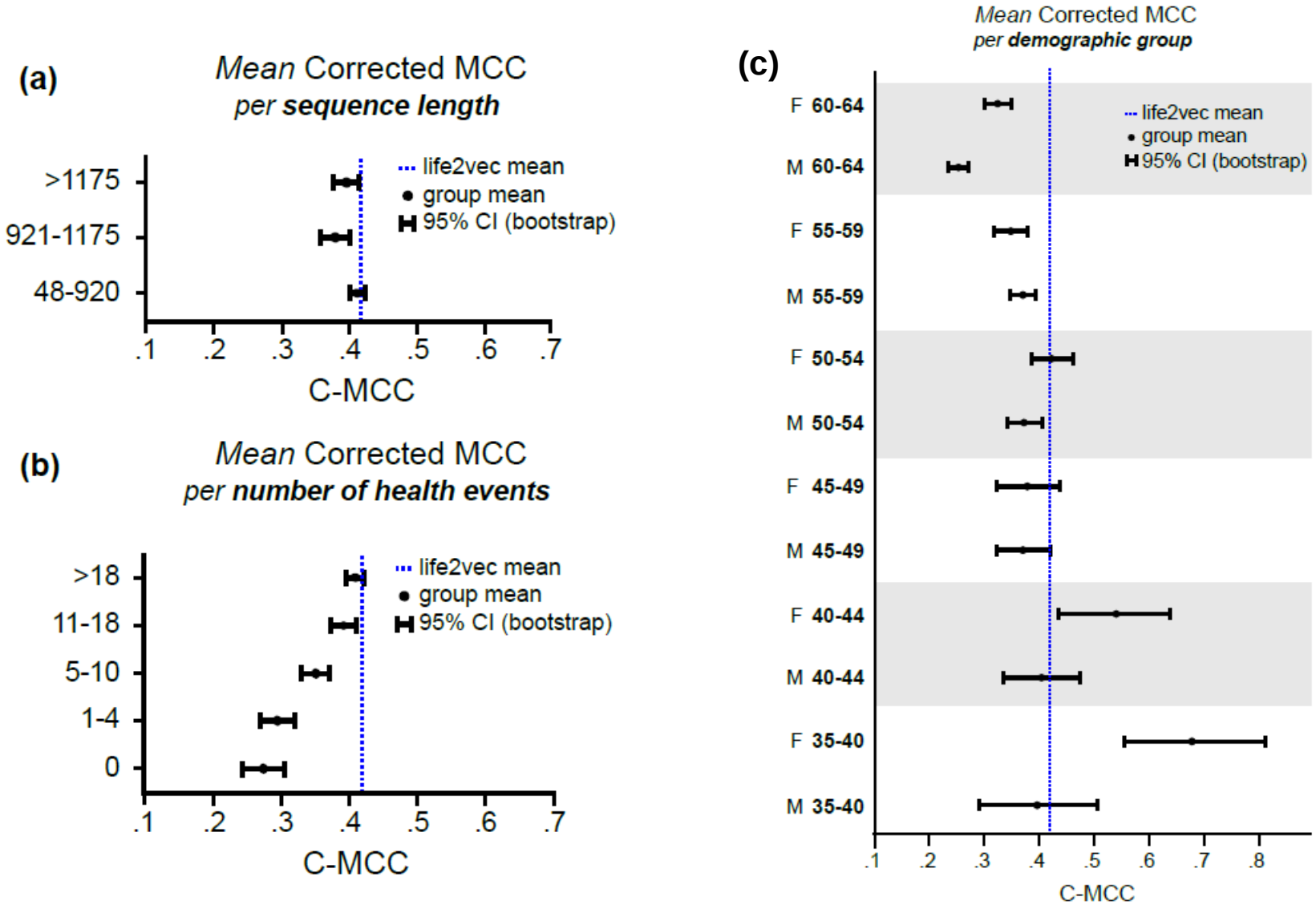
$$\hat{\gamma}_{cr} = (1-\hat{\alpha})^{-1}((1-\hat{\alpha})\cdot\hat{\gamma})$$

$$\hat{\eta}_{cr} = (1-\hat{\alpha})^{-1}(\hat{\eta} - \hat{\alpha}\cdot\hat{\gamma})$$

$$\hat{\pi}_{cr} = \hat{\pi} + (1-\hat{\pi})\cdot\hat{\alpha}$$

Northeastern University

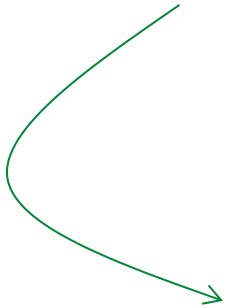
Early Mortality Prediction: Auditing



Early Mortality Prediction: Data Use

Retrain the model on different variations of the dataset

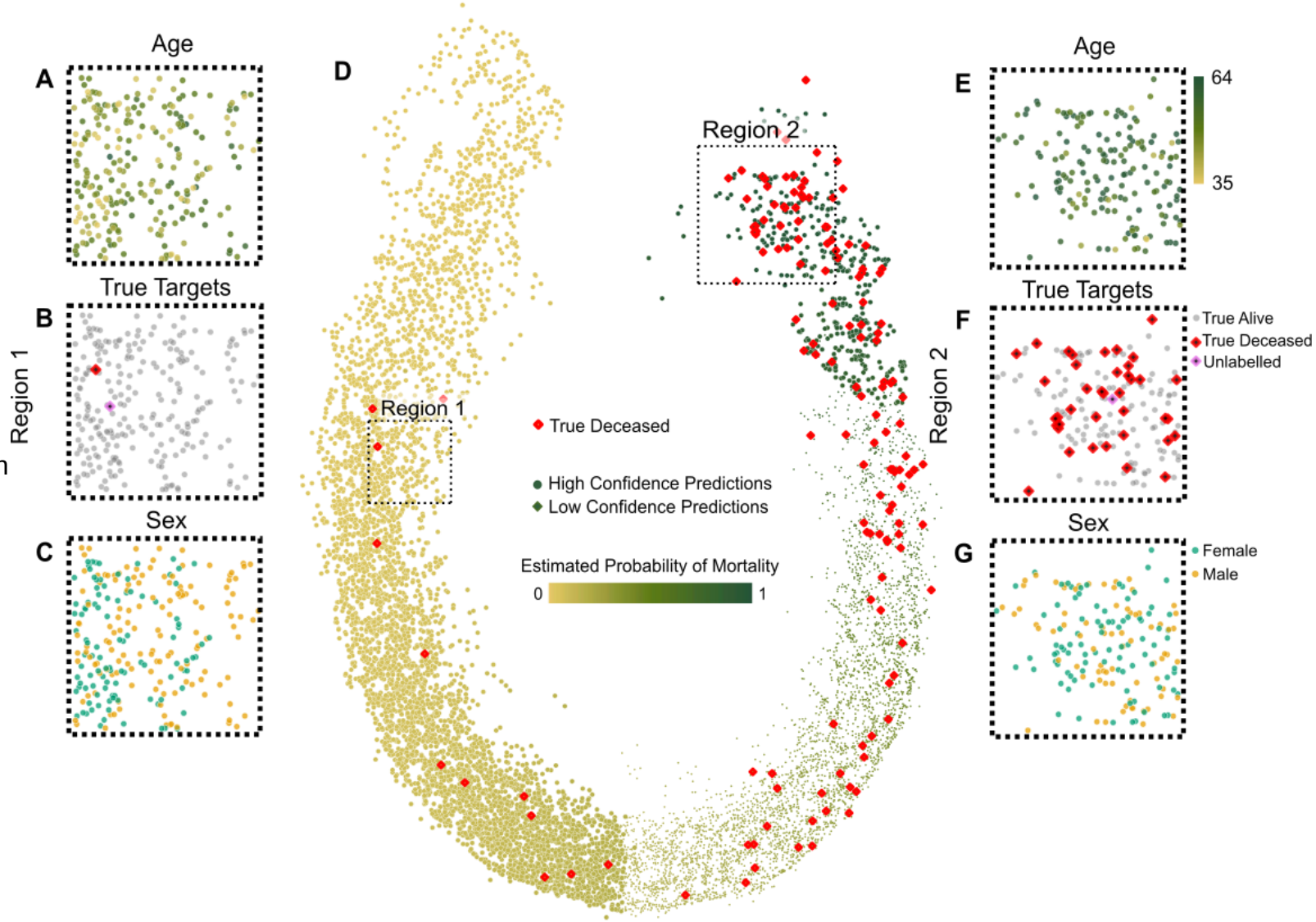
Data	C-MCC, 95%-CI	AUL	Vocab Size
Full Labor & Health	0.413 [0.410, 0.422]	0.845	2043
Partial Labor & Health	0.375 [0.367, 0.384]	0.837	1034
Only Full Labor	0.319 [0.312, 0.327]	0.809	1290
Only Partial Labor	0.278 [0.271, 0.285]	0.782	281



Partial Labor: no industry, sector, position and labour force

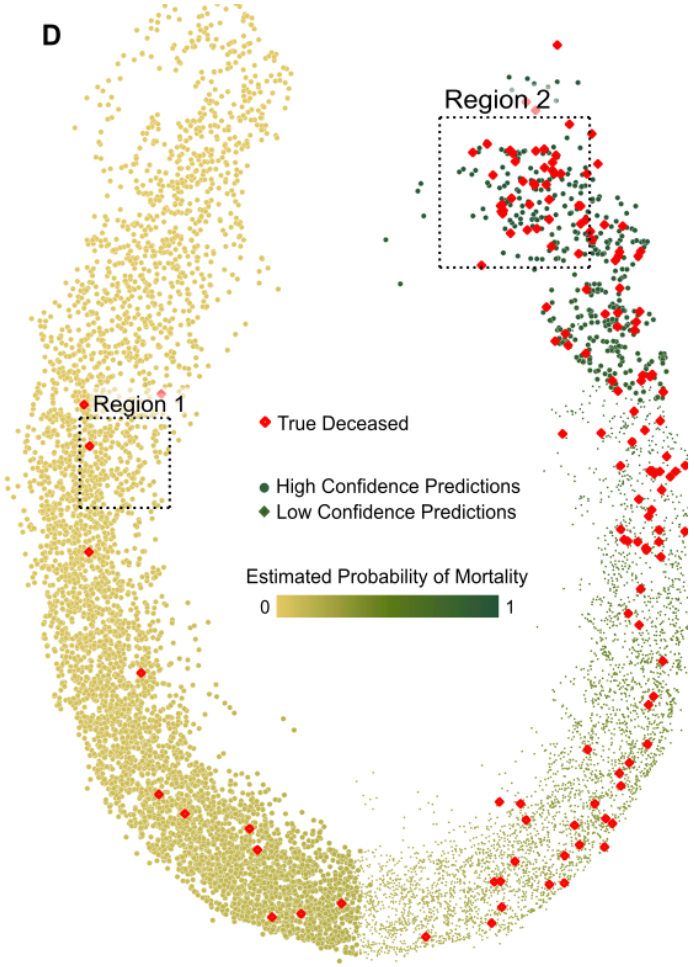
We can look at the low dimensional space of life-summaries.

2D Projection



Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Explainability with TCAV (Mortality Prediction):



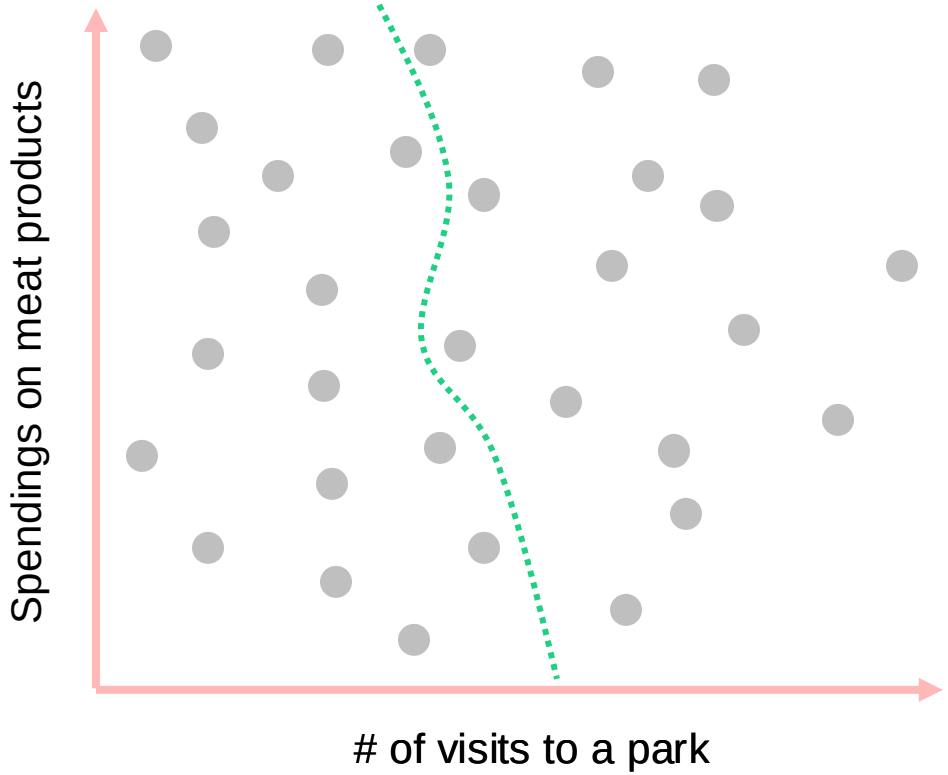
In the Concept space, we can find *somewhat* explainable directions!

- **Here, we do not – we need to find them!**

TCAV allows to find these directions

- Interpretation of the **directions of the person-summary** space
- **Sensitivity of the model** towards these directions
- Global Interpretability

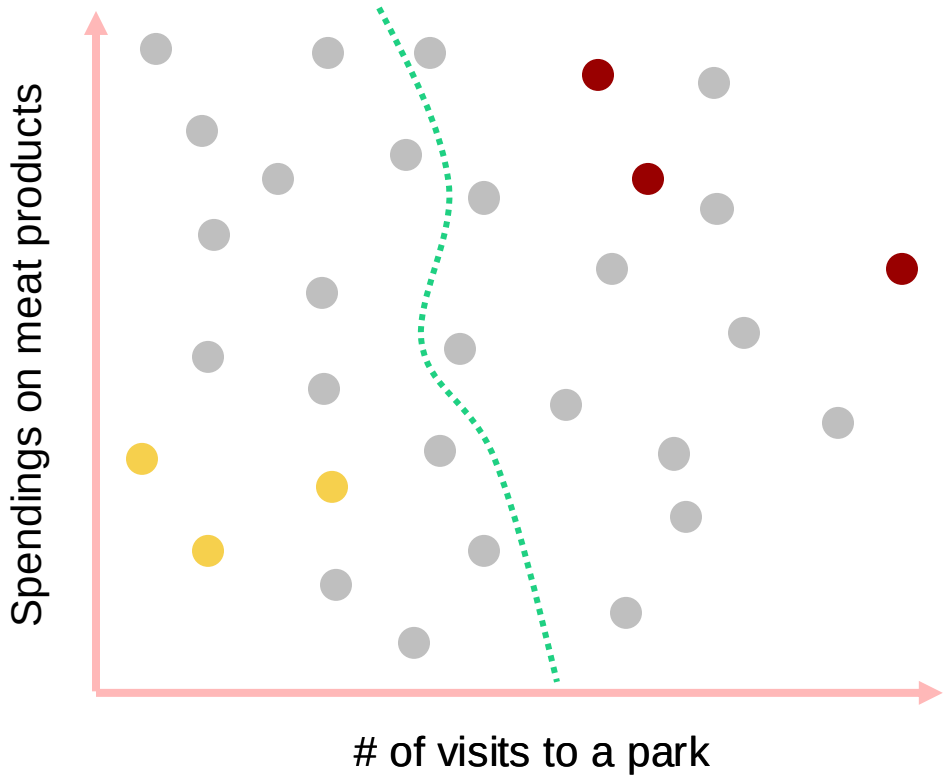
Overview of the TCAV method



Let's imagine an **algorithm that predicts whether a person has a dog**

decision boundary of the algorithm

Overview of the TCAV method



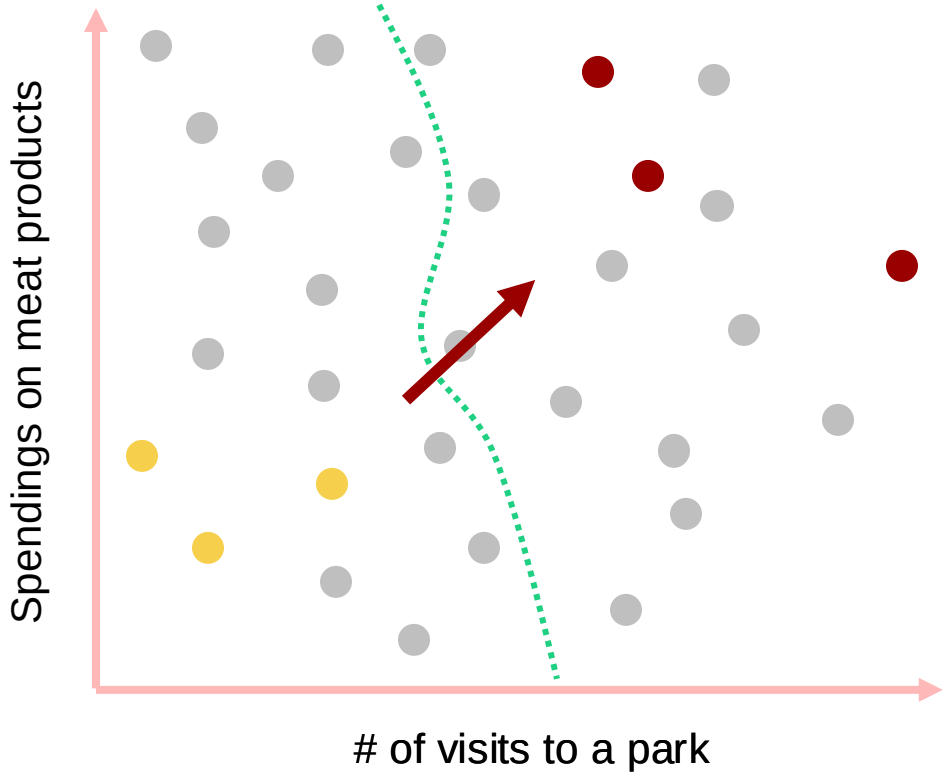
Let's imagine you have extra information

decision boundary of the algorithm

Lives in a rental

Owns an apartment

Overview of the TCAV method



Let's imagine you have extra information

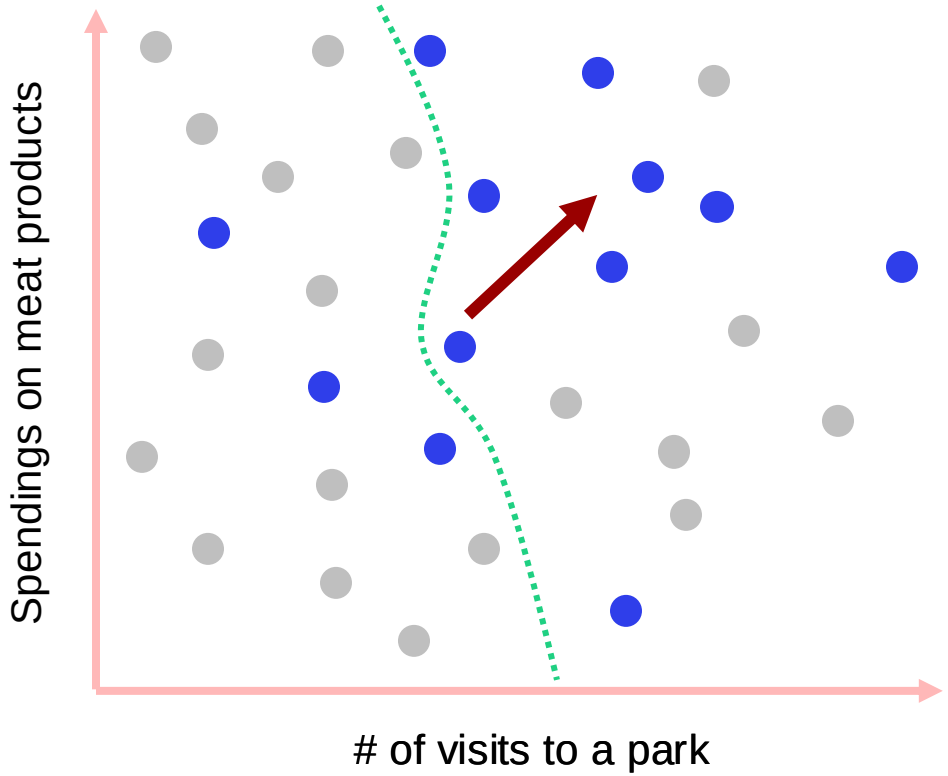
decision boundary of the algorithm

Lives in a rental

Owns an apartment

Direction of a concept

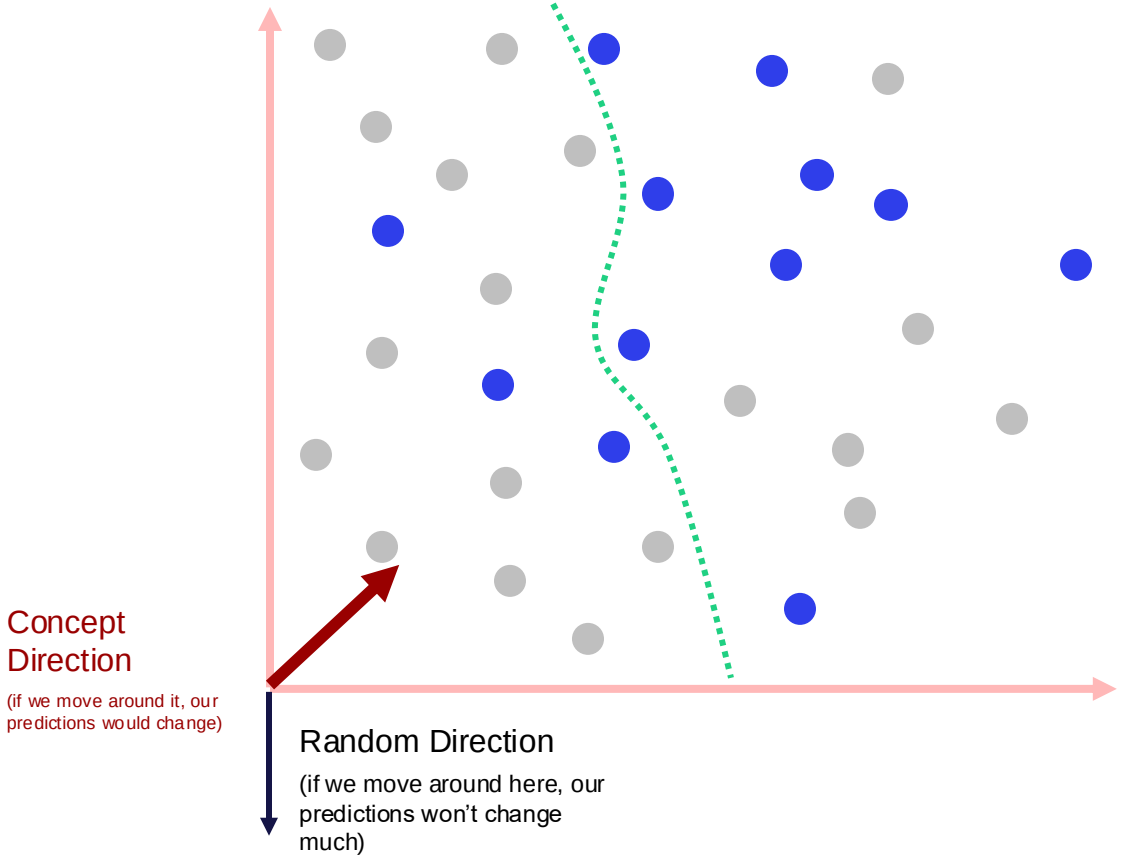
Overview of the TCAV method



Interpretation: If we move in a certain direction (the one that is associated with a concept), how strongly would it influence the output of our model (on average)

- decision boundary of the algorithm
- Randomly sampled point
- Direction of a concept

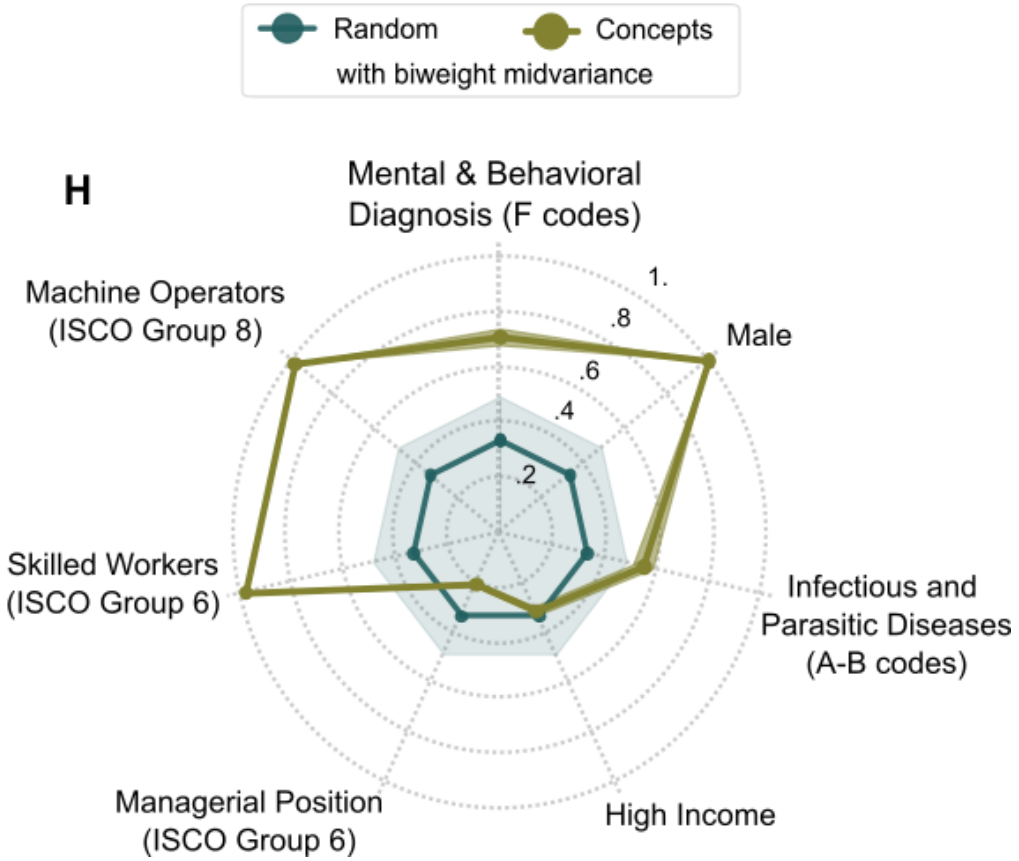
Overview of the TCAV method



Interpretation: If we move in a certain direction (the one that is associated with a concept), how strong would it influence the output of our model (on average).

Northeastern University

Explainability with TCAV (Mortality Prediction):



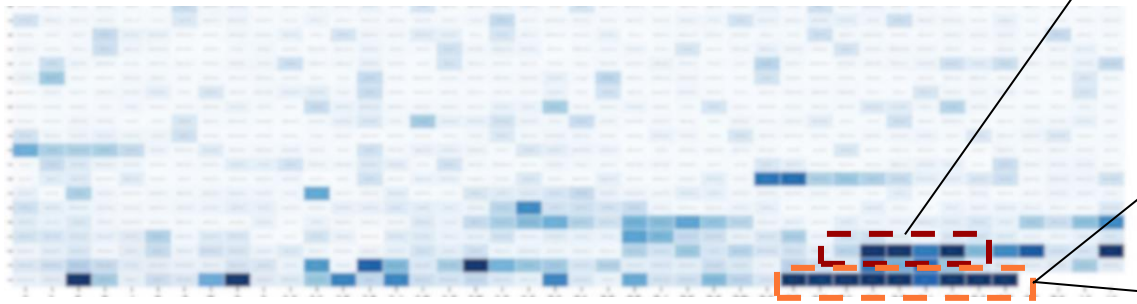
TCAV Score per "Direction"

- Interpretation of the **directions of the person-summary** space
- **Sensitivity of the model** towards these directions
- Global Interpretability

Northeastern University

Local Interpretability

- **Interpretation of the scores**: how large is the change in the output likelihood if we slightly change the embedding of the token
- Only local explanation (e.g. per sequence), vague interpretation



Sequence of an individual in a textual format

Read: left-to-right, top-to-bottom

If we zoom-in:

Maternity Leave

Flexjob, receives salary

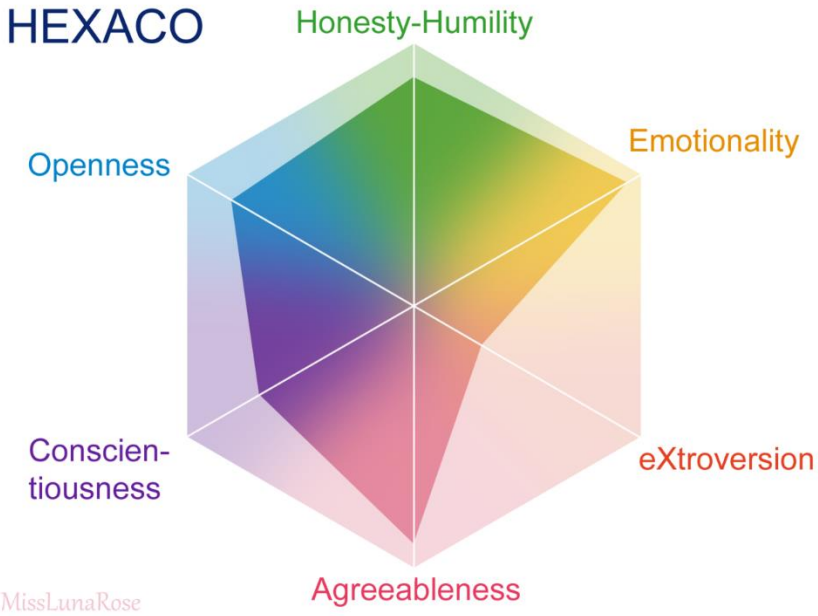
Malignant neoplasm of brain (admitted to hospital)

Pyothorax with fistula (admitted to hospital)

Northeastern University

`life2vec` and *Personality Traits*

- We focus on Extroversion Facets:
  - **Sociability** (tendency to enjoy social interactions)
  - **Liveliness** (one's typical enthusiasm and energy)
  - **Self-esteem** (tendency to have positive self-regard)
  - **Boldness** (comfort within a variety of social situations)



Example:

1. In social situations, I'm usually the one who makes the first move

Image source: Wikipedia

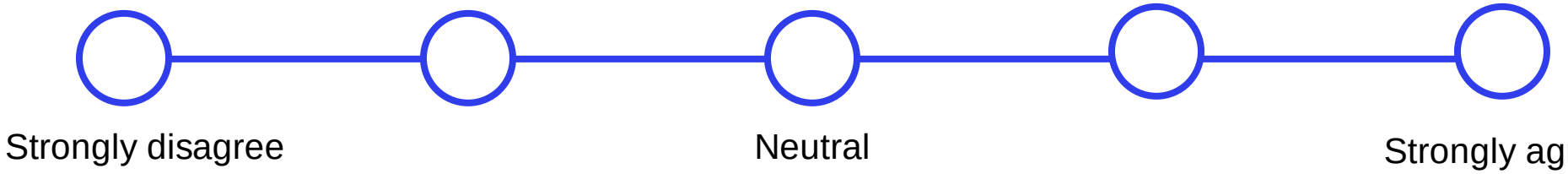
Inventory Descriptions: The HEXACO Personality Inventory - Revised

Extraversion Nuance Prediction

- **Task: "What kind of replies does the person give to the 10 questions evaluating their Extraversion? "**
  - Multiclass prediction
  - Ordinal Classification task (i.e. labels have ordered)
  - Highly Imbalanced Data
  - *We do not have much data*

Statement:

In social situations, I'm usually the one who makes the first move



Northeastern University

Personality Data

Quadratic Kappa Score

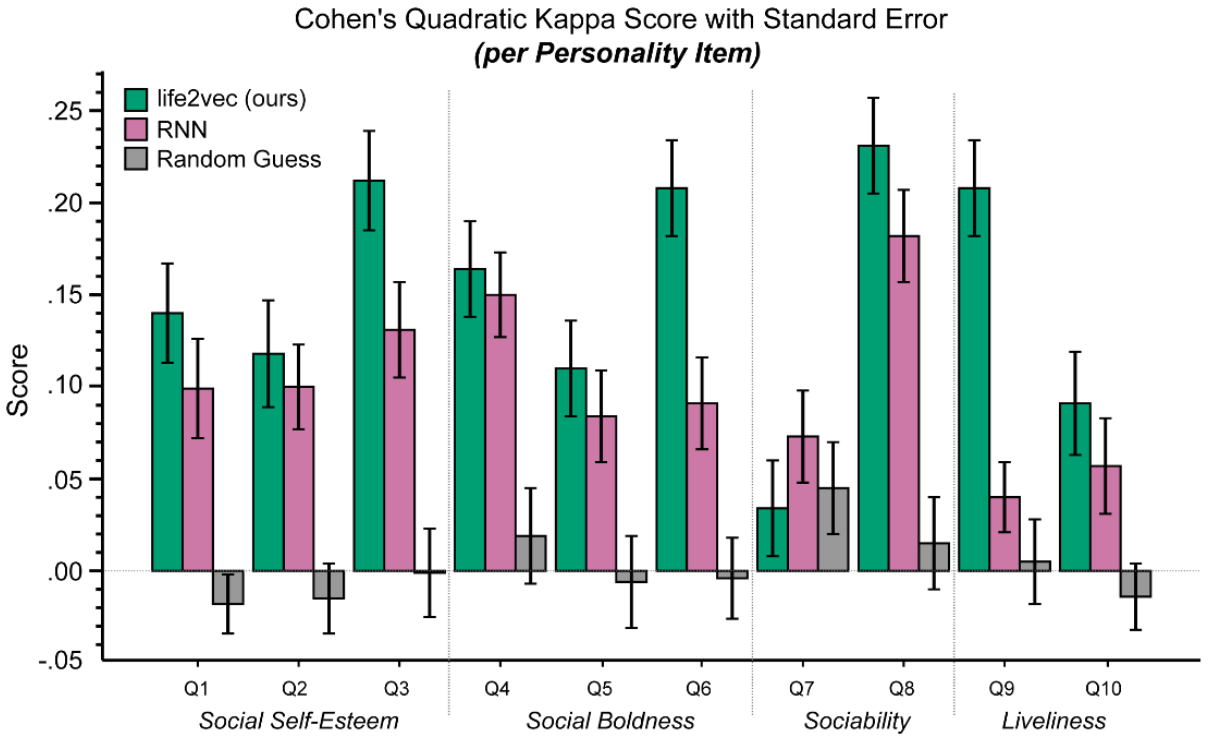
$$\kappa^2 = 1 - \frac{\sum_{i,j} w_{ij} \times c_{ij}}{\sum_{i,j} w_{ij} \times e_{ij}}$$

True \ Predicted	1	2	3	4	5
1	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}
2	c_{21}	c_{22}	c_{23}	c_{24}	c_{25}
3	c_{31}	c_{32}	c_{33}	c_{34}	c_{35}
4	c_{41}	c_{42}	c_{43}	c_{44}	c_{45}
5	c_{51}	c_{52}	c_{53}	c_{54}	c_{55}

$$e_{ij} = \frac{\sum_k c_{ik} \times \sum_k c_{ki}}{N}$$

$$w_{ij} = \left(\frac{i-j}{K-1}\right)^2$$

Accounts for the distance
from predicted to target classes



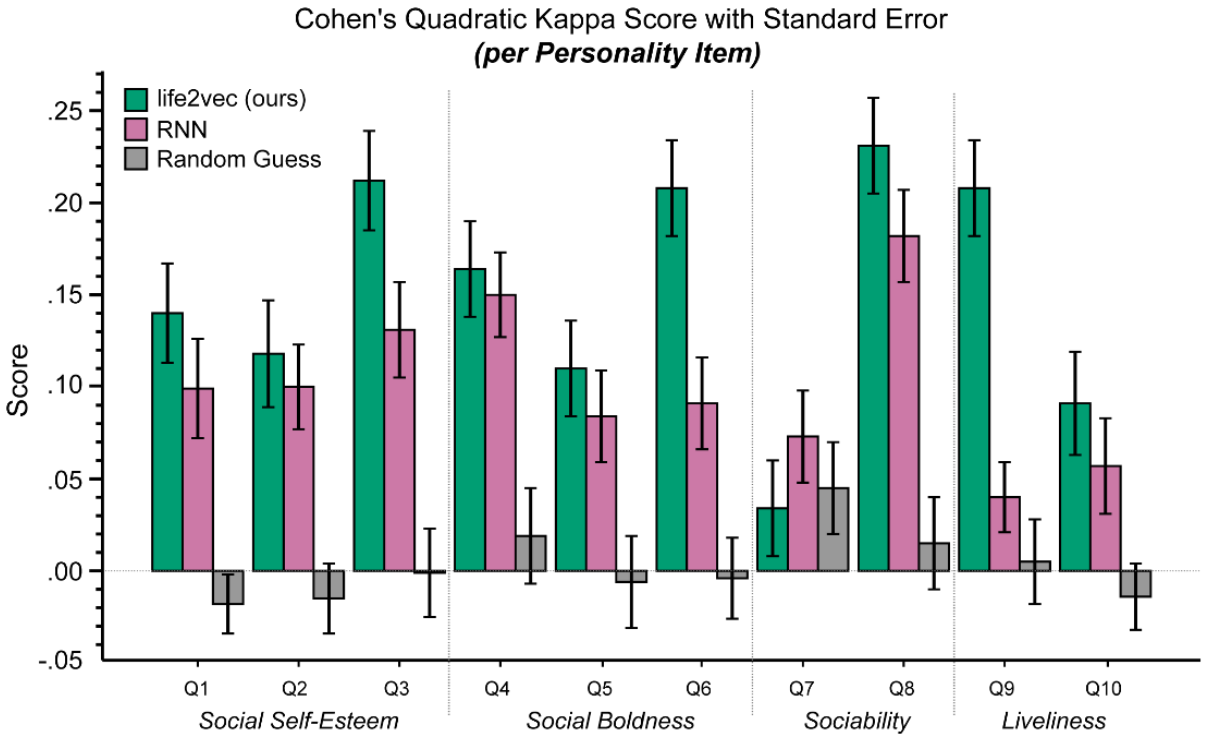
Northeastern University

Personality Data

Questions:

6. Most people are more upbeat and dynamic than I generally am (liveliness)

7. The first thing that I always do in a new place is to make friends (social I)



What does it tell us?

Performance:

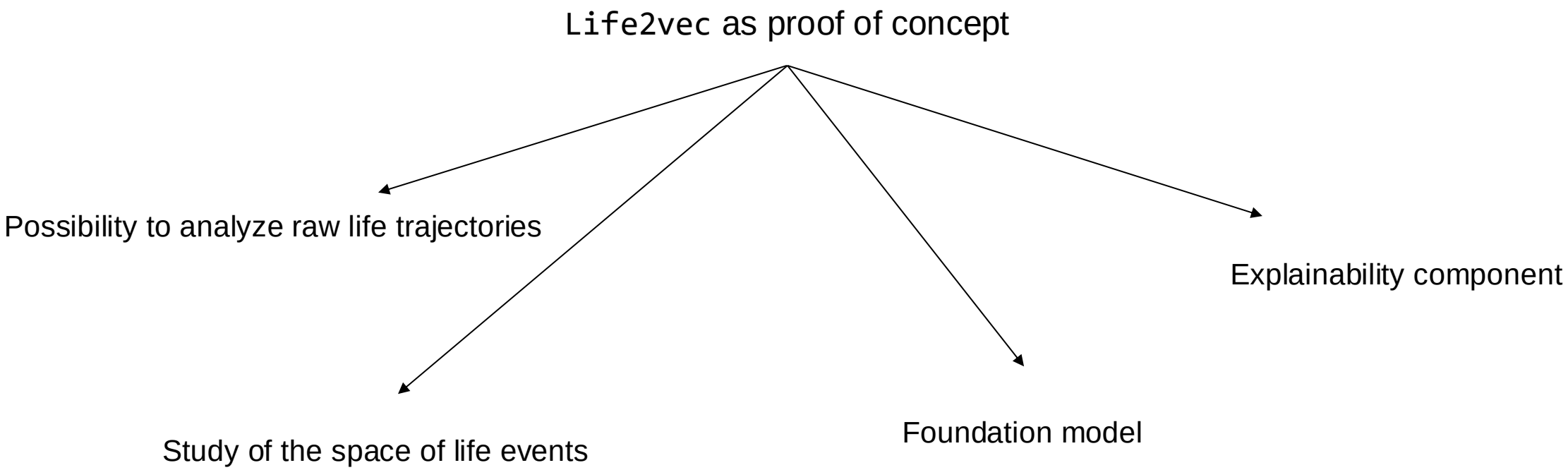
- You can use pretrained life2vec for downstream tasks
- Provides somewhat interpretable predictions
- Interpretations align with the literature

Person-summaries:

- Meaningful space
- Can be used to study various phenomena

Conclusion

Conclusion


Life2vec as proof of concept
Possibility to analyze raw life trajectories
Explainability component
Study of the space of life events
Foundation model

Thank you for attention!