DTU

PHD THESIS DEFENCE
GERMANS SAVCISENS
13TH MARCH 2024

LIFE TRAJECTORIES AS SYMBOLIC LANGUAGE

Exploring Human Behaviour with Language Models

Special Thanks


Agata Wlaszczyk


Nikolaos Nakis


Sune Lehmann


Lars Kai Hansen


Anna Rogers


Tina Eliassi-Rad


Laust Hvas Mortensen


Ingo Zettler


Lau Lilleholt


Social Complexity Lab Members

Agenda

Introduction

Part I: Data

Part II: Representation Learning and NLP

Part III: Forming Labour and Health Language

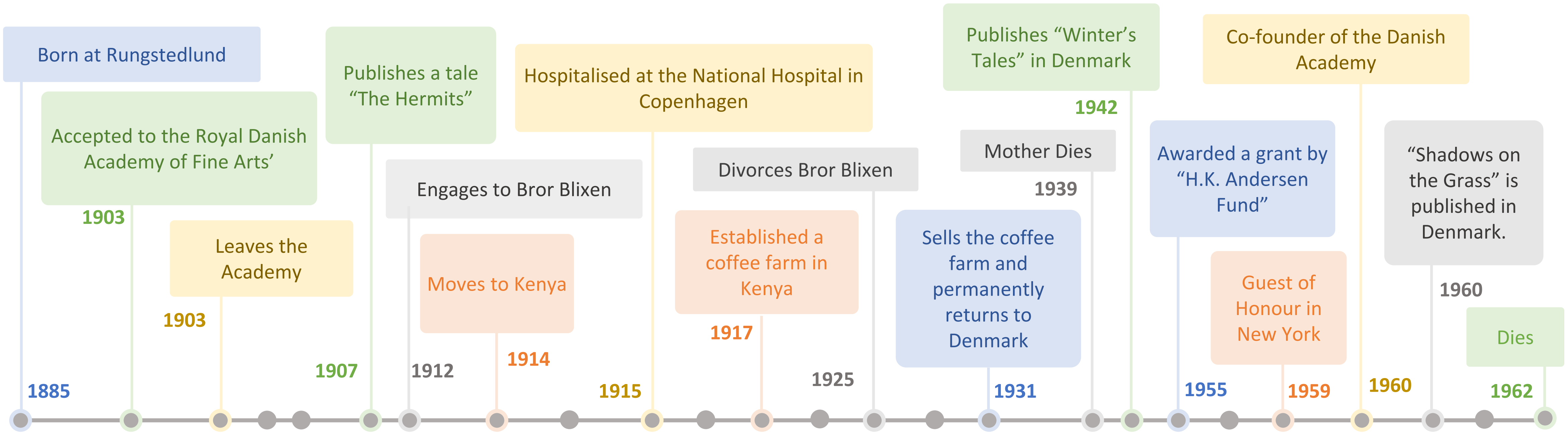
Part IV: Capturing the structure with the `life2vec`

Part V: `life2vec` as a foundation model

Conclusion

Life Trajectories

Life of Karen Blixen (Danish author)*



** simplified*

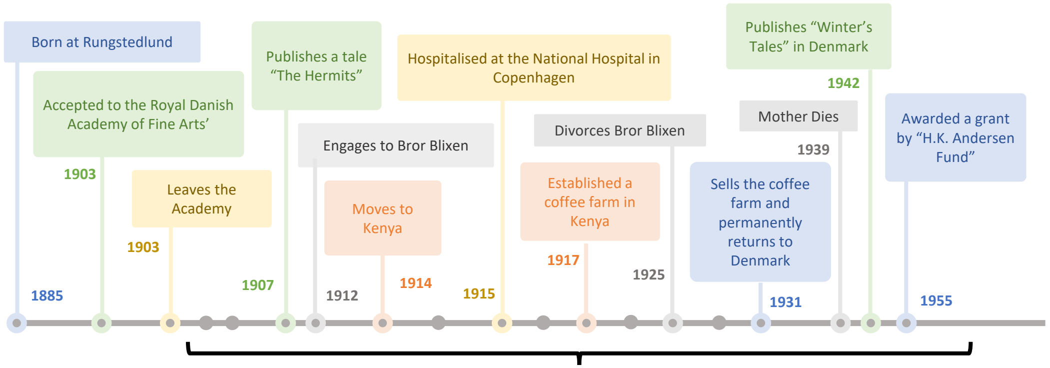
The Problem

Issues associated with **longitudinal data modelling**:

- Features have **mixed formats** (continuous and categorical).
- Various data sources
- Events have an “**uneven**” sampling rate.
- **Missing values**
- **The number of records** per person **varies** a lot

Classical models are not that good at handling it!

The Problem



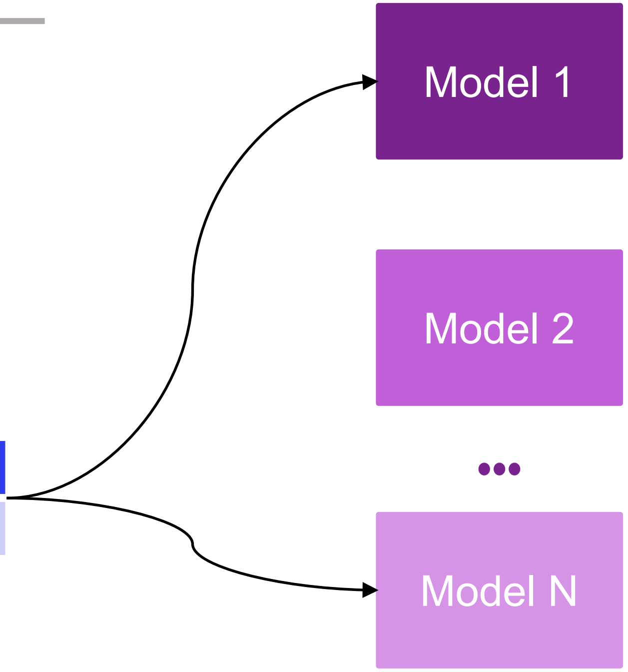
Simplifying data

How many times admitted to a hospital?

Career changes?

Traveling abroad?

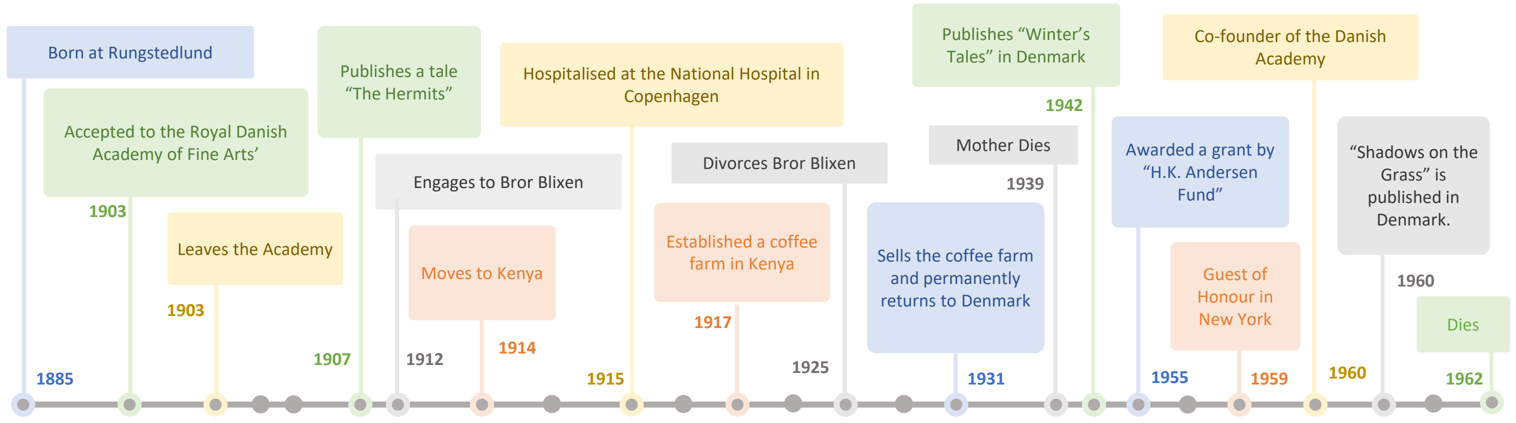
Travelled within a year	…	Married	Hospital Admission
1	…	1	2



Probability of readmission to a hospital?

Income level within the next year?

* simplified



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

General Purpose Model

Compressed **representation of life** progression

Predict the human behaviour

(on an *individual* level)

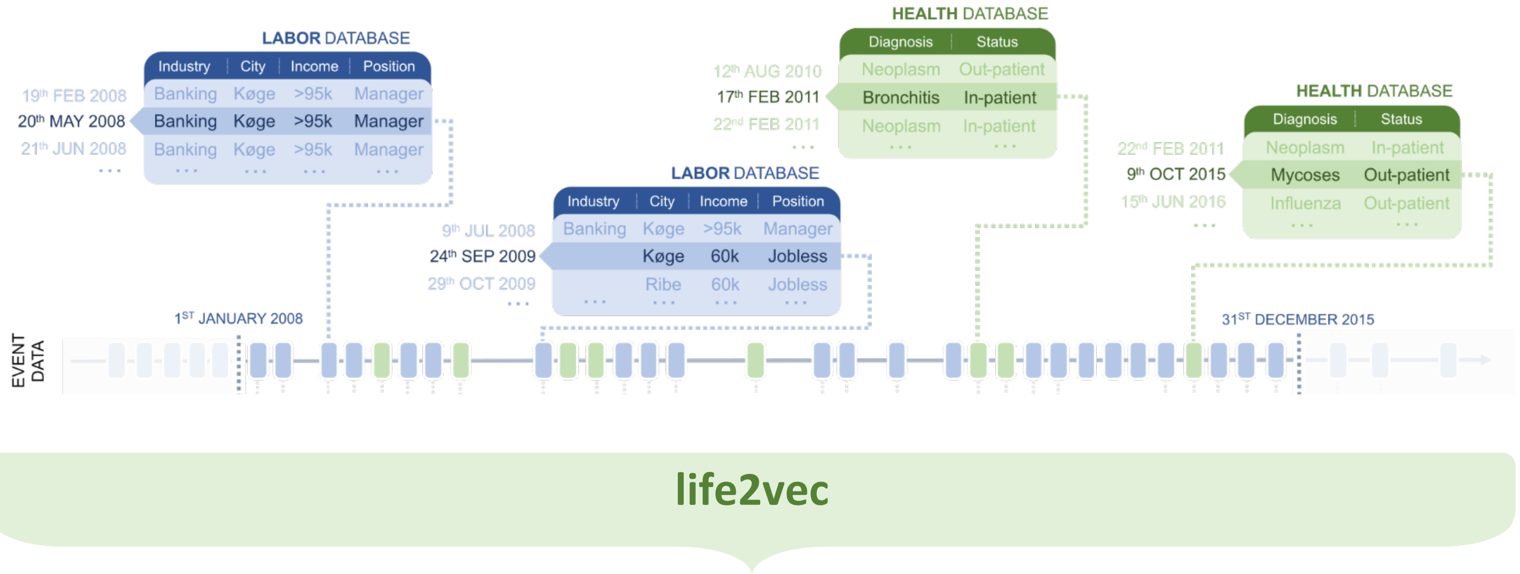
Give comprehensive insight into the data

Study sociological phenomena

(on a *global* scale)

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

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



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...

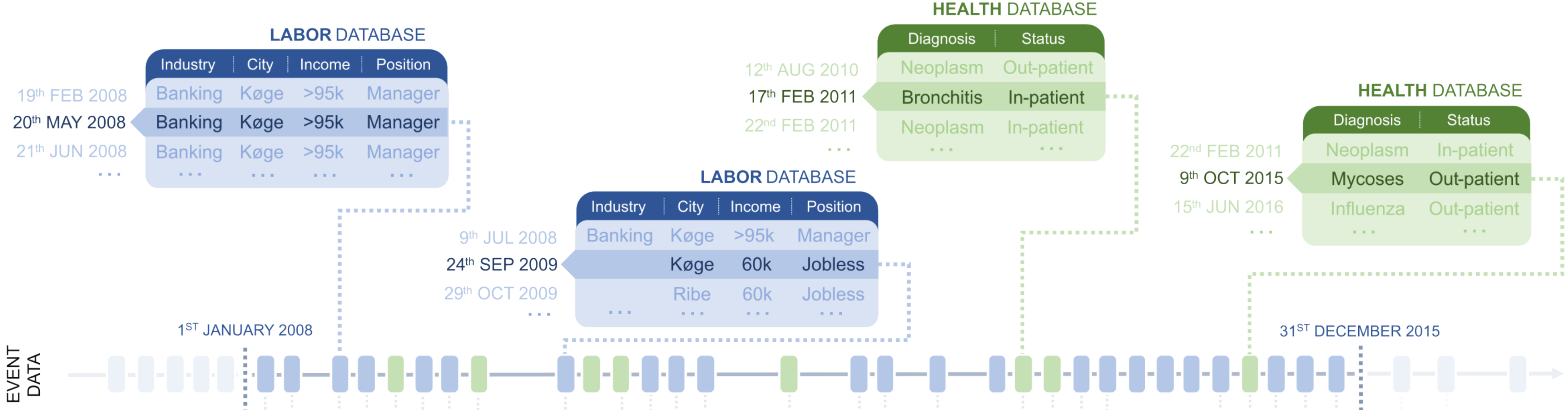
Labour and income
Employment, unemployment, earnings, income, wealth...

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



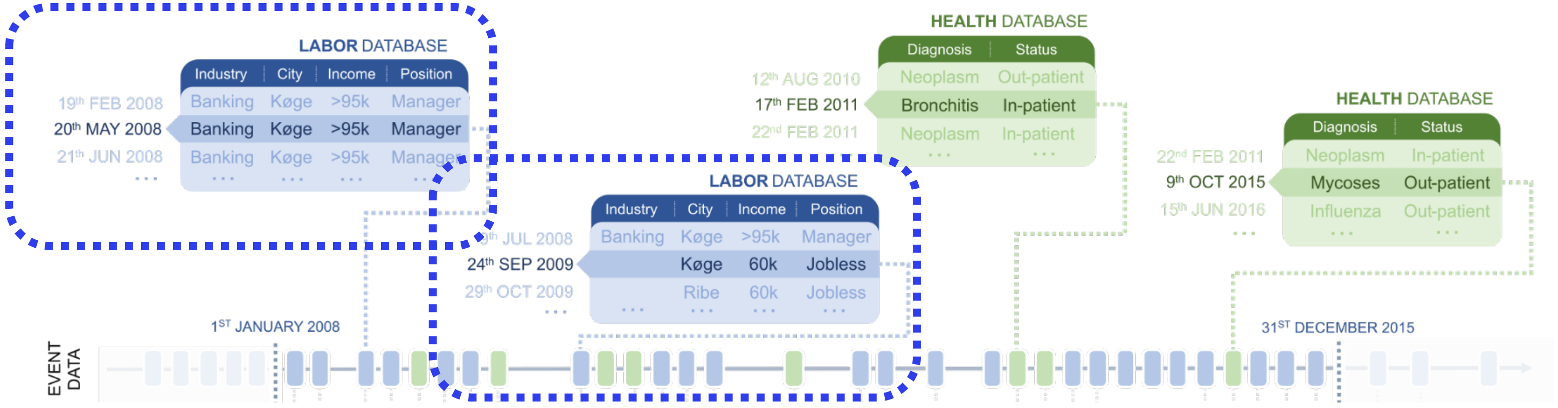
DANMARKS STATISTIK

Labour Data

Health Data

Detailed reconstruction of labor and health life trajectories

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

Labor Data: Hierarchies

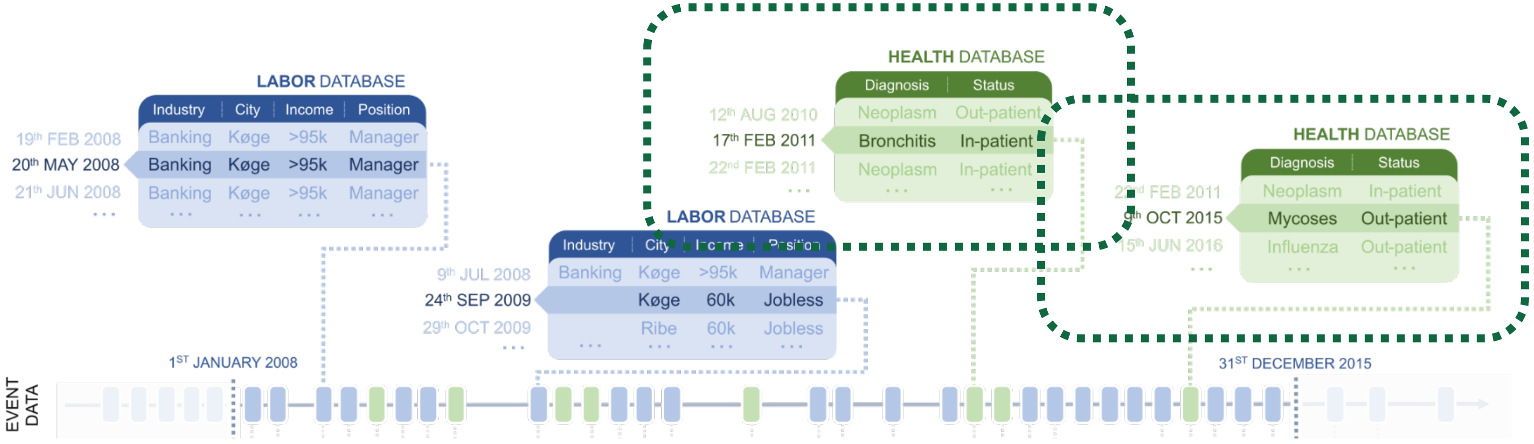
Example of codes describing the **Industry**

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



Records of visits to a health practitioner or hospital:

– Focus on 3 features

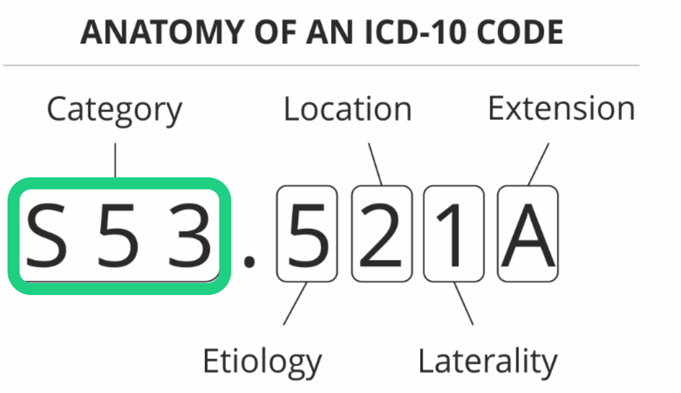
– Diagnoses encoded in the ICD10 System

Features we use:

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

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)



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

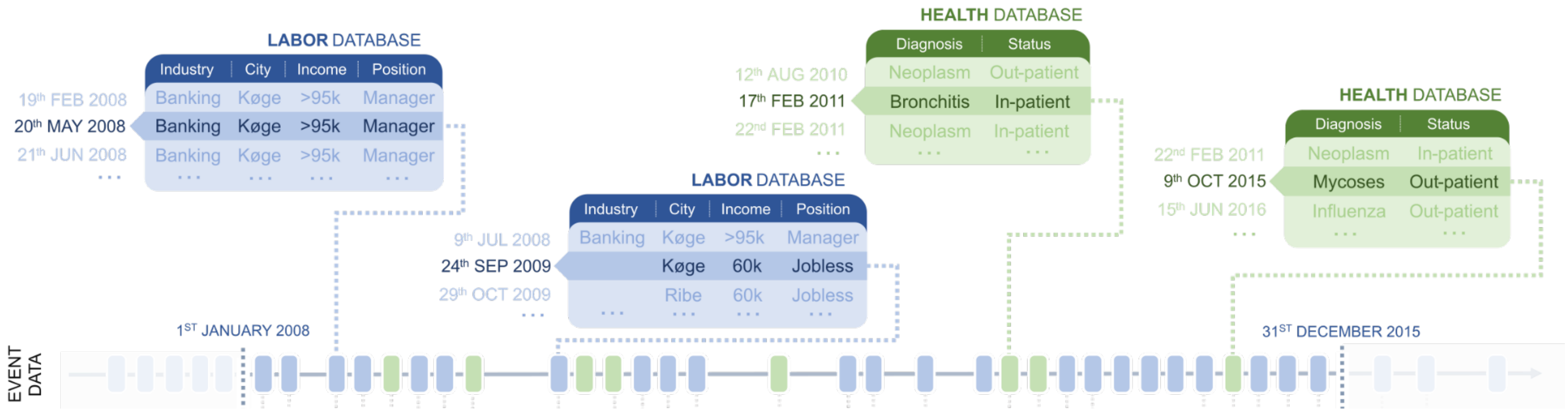
Examples of ICD10 codes:

- **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

Power of National Registry

The National Registry is a source of **fine-grained information** about **the progression of one life**.

Unique possibility to study life progression and life outcomes.



How do we analyze?

Part II

Representation Learning and NLP

Natural Language Processing

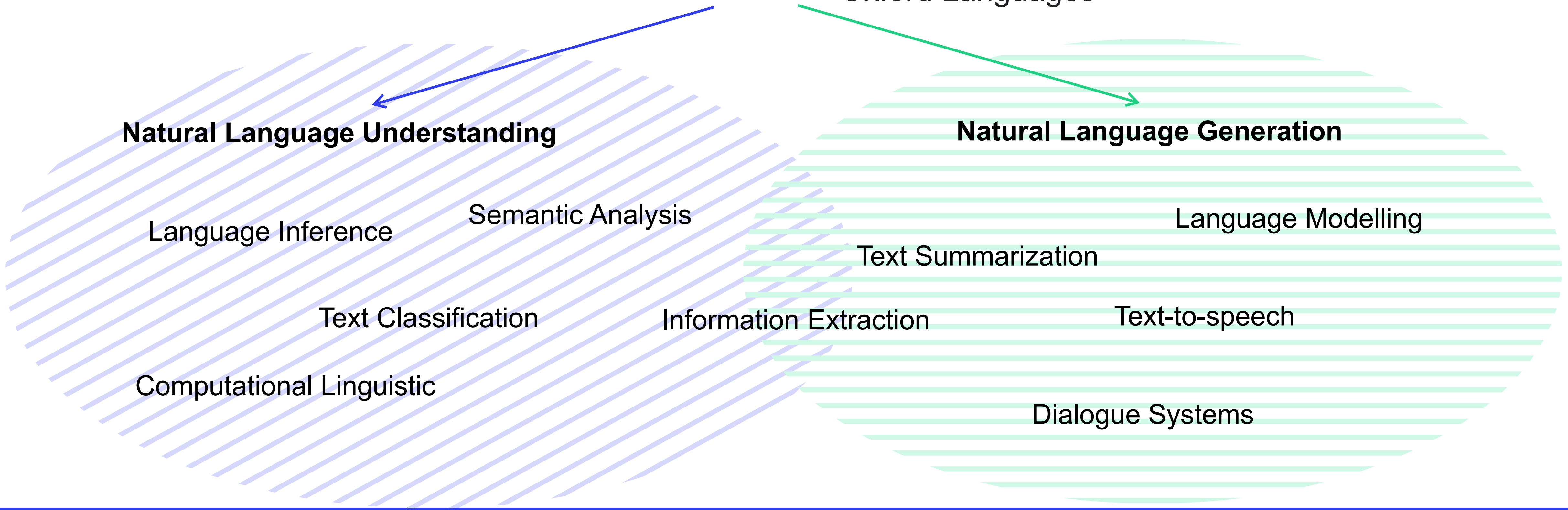
“[..] *the application of computational techniques to the **analysis** and **synthesis** of natural language and speech.”*

- Oxford Languages

Natural Language Processing

"[..] *the application of computational techniques to the* ***analysis*** *and* ***synthesis*** *of natural language and speech."*

- Oxford Languages



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.

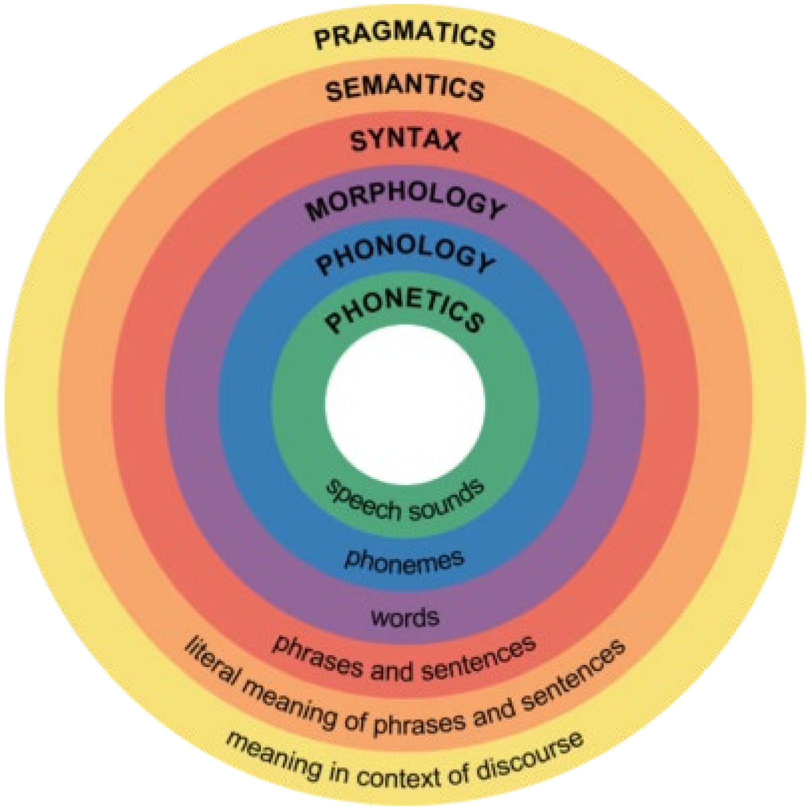
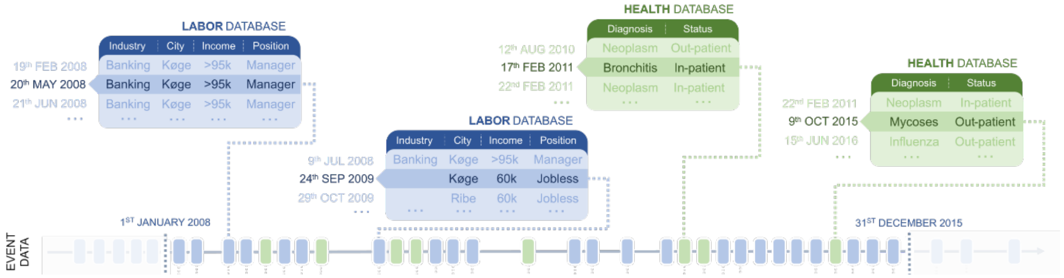


Image: Luchmee, D. (2019, July 25*). The Complex Skill of Language*. HappyNeuron. Retrieved March 5, 2024, from https://news.happyneuronpro.com/the-complex-skill-of-language/

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”

“Document” representation

$$\begin{Bmatrix} 0.1 \\ 0.2 \\ 0.0 \end{Bmatrix} + \begin{Bmatrix} 0.1 \\ 0.0 \\ 0.0 \end{Bmatrix} + \begin{Bmatrix} 0.0 \\ 0.2 \\ 0.5 \end{Bmatrix} \rightarrow \begin{Bmatrix} 0.2 \\ 0.4 \\ 0.5 \end{Bmatrix}$$

Aggregate the word representations

“Susanne likes kiwi”

$$\rightarrow \begin{Bmatrix} 0.1 \\ 0.2 \\ 0.6 \end{Bmatrix}$$

“Maria likes spaceships”

$$\rightarrow \begin{Bmatrix} 0.2 \\ -.1 \\ 0.0 \end{Bmatrix}$$

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

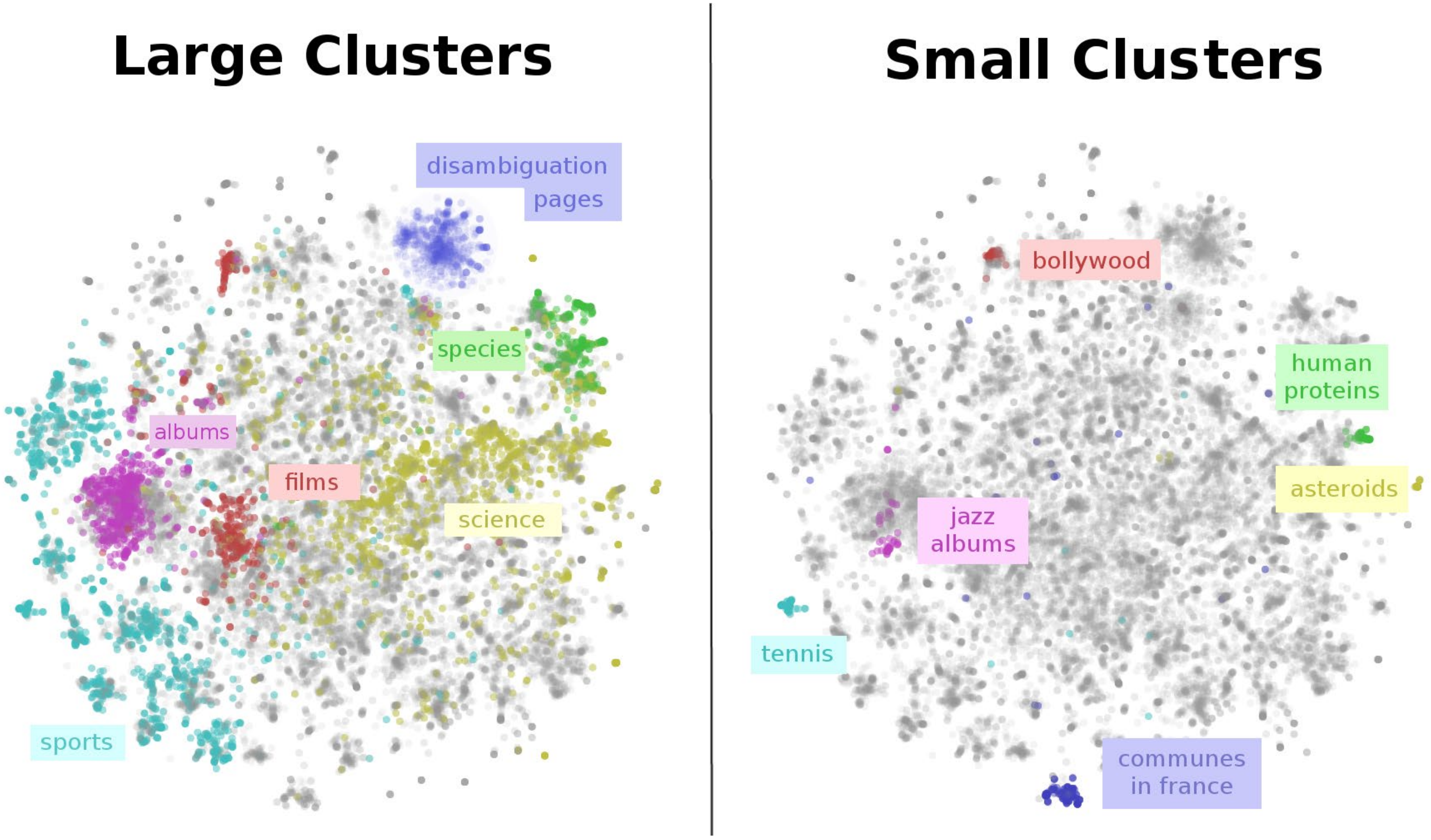
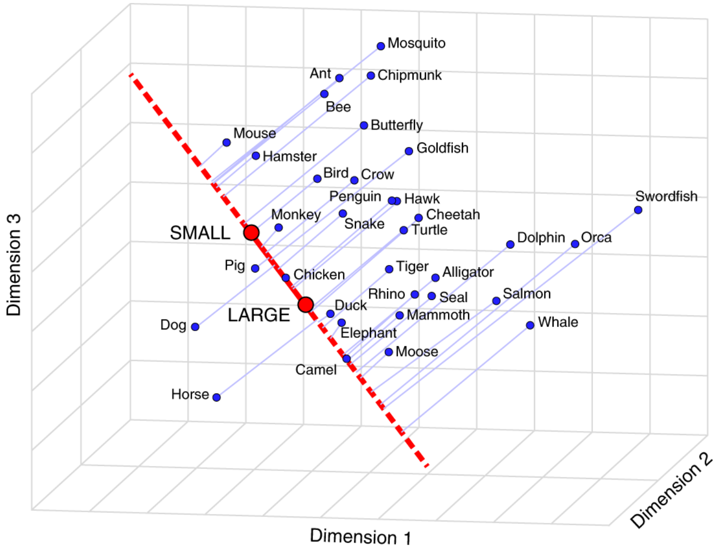
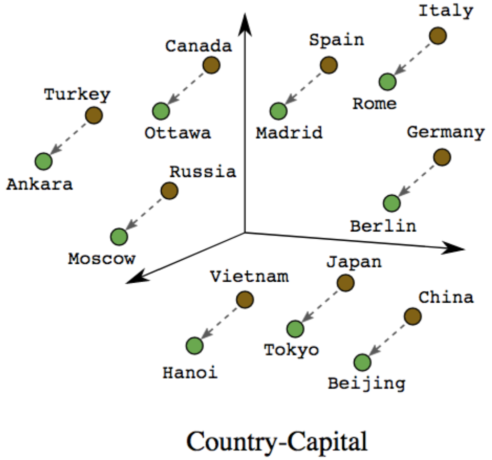


Fig 1: Two-dimensional projection of the word embeddings (word2vec)[1]

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/

Embedding Spaces and Structure





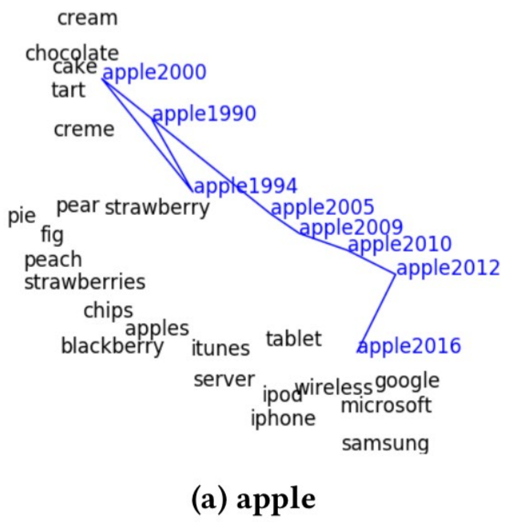


Fig.1: Schematic illustration of semantic projection[1]

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

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

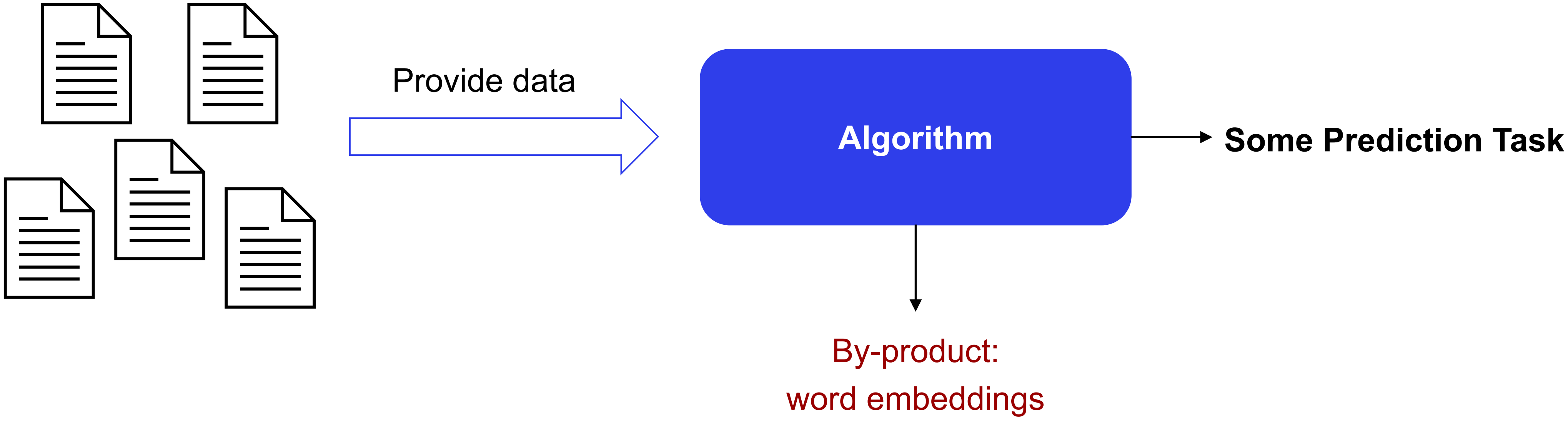
Fig.3: Trajectories of brand names[3]

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

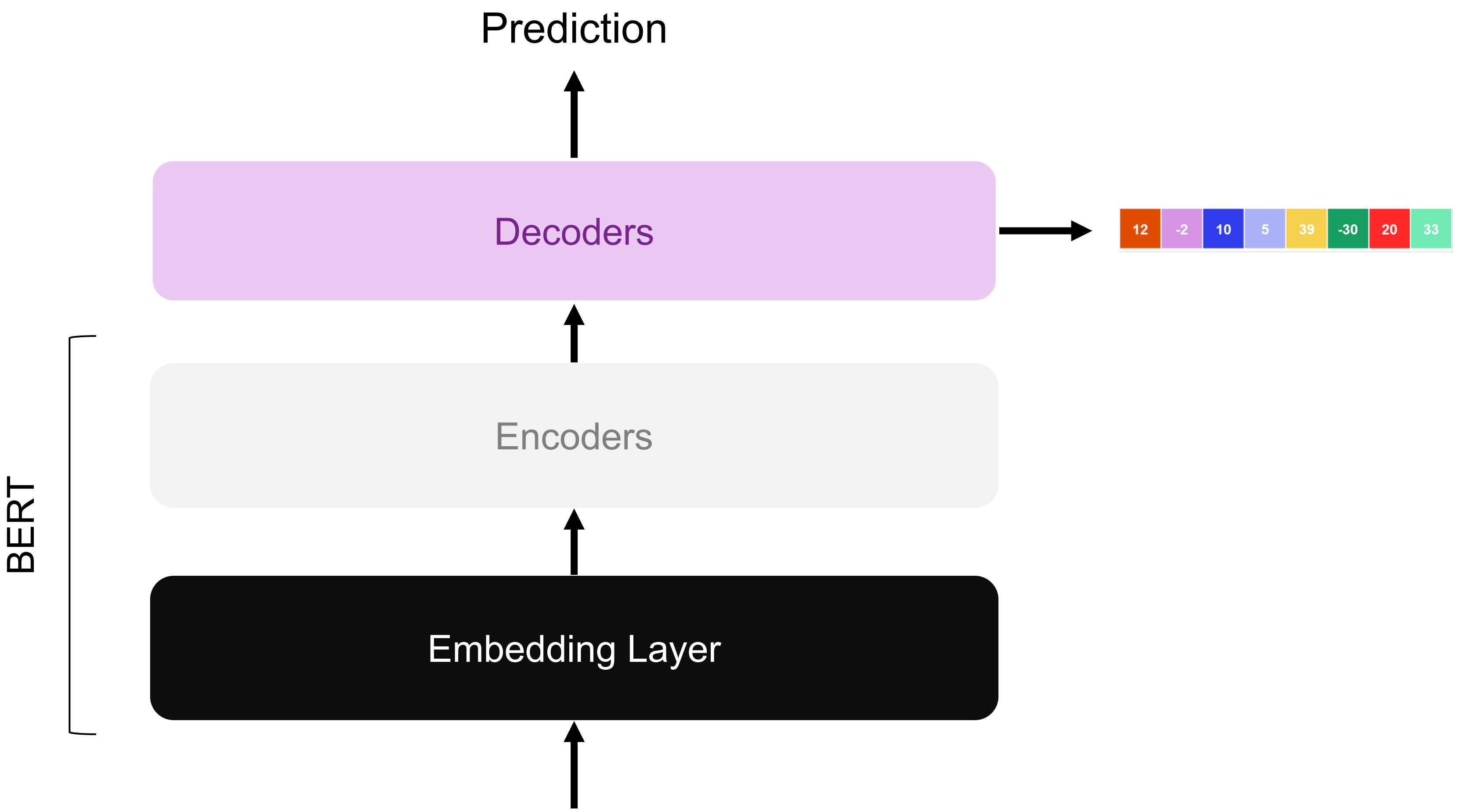
Bidirectional Encoder Representations from Transformers (**BERT**)

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

General-purpose model, adaptable to new tasks

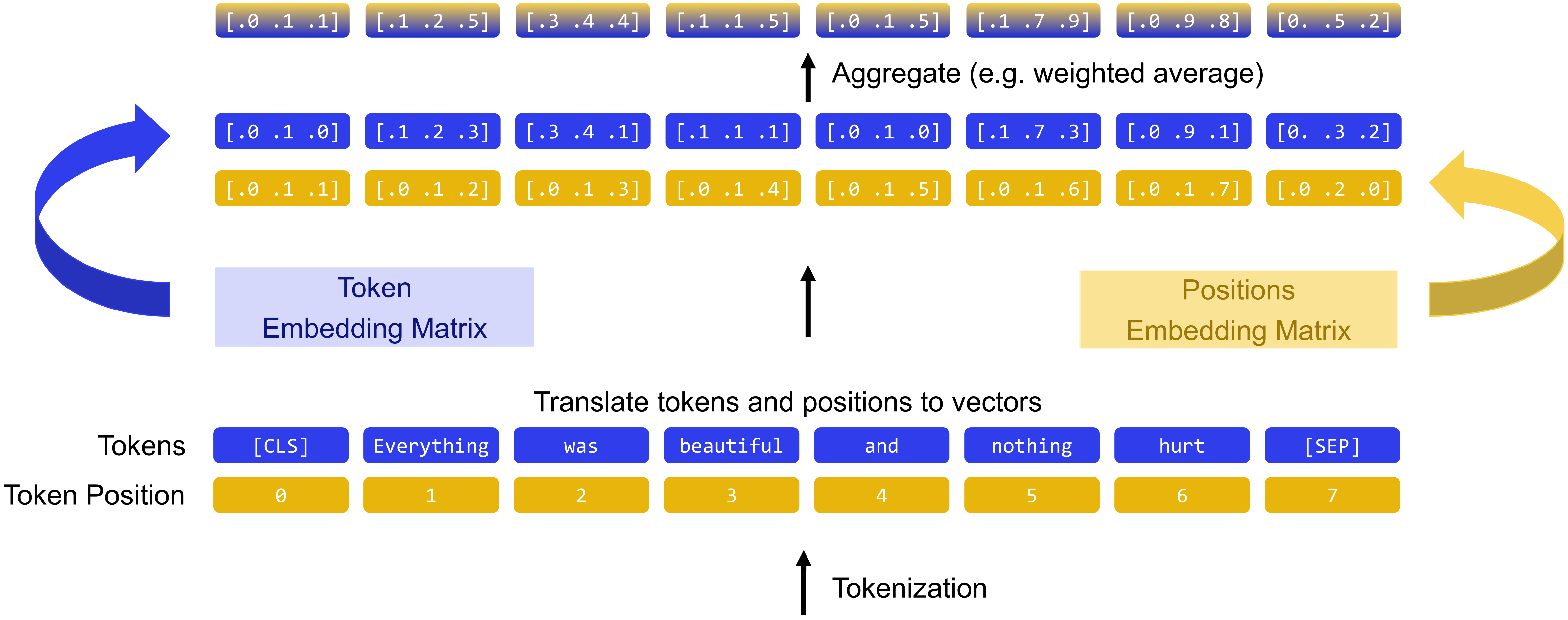
Great **predictive performance** on many NLP tasks

Transformer Architecture (BERT)



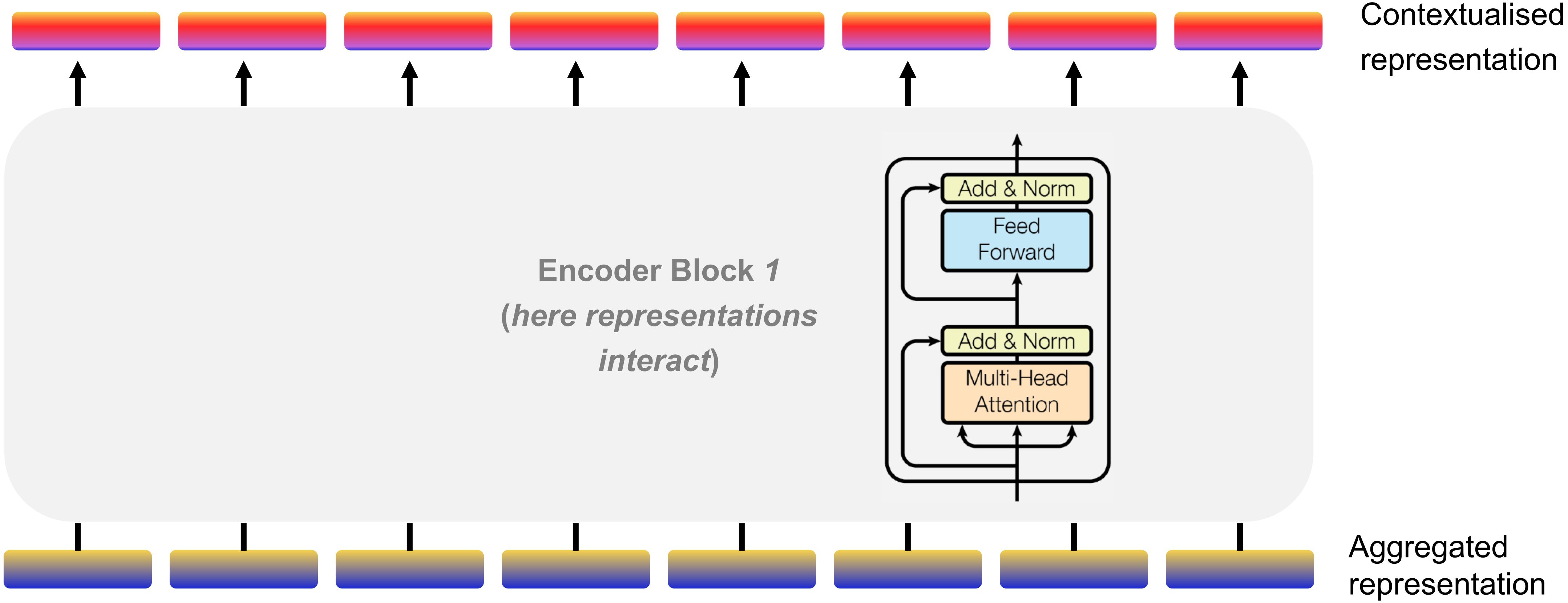
"Everything was beautiful and nothing hurt"

Embedding Layer



"Everything was beautiful and nothing hurt"

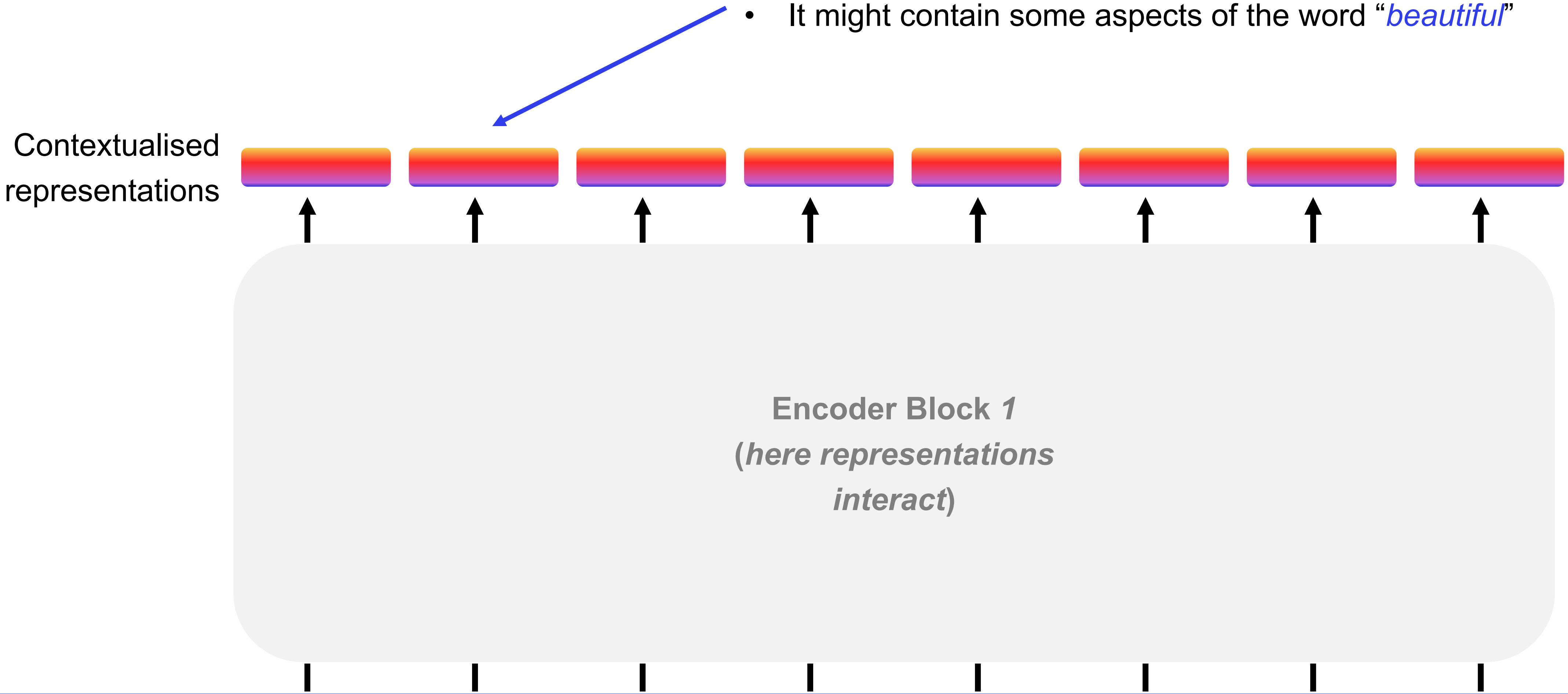
Encoders


Contextualised representation
Encoder Block *1*
(*here representations interact*)
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Aggregated representation

Encoders

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

- It might contain some aspects of the word “*beautiful*”

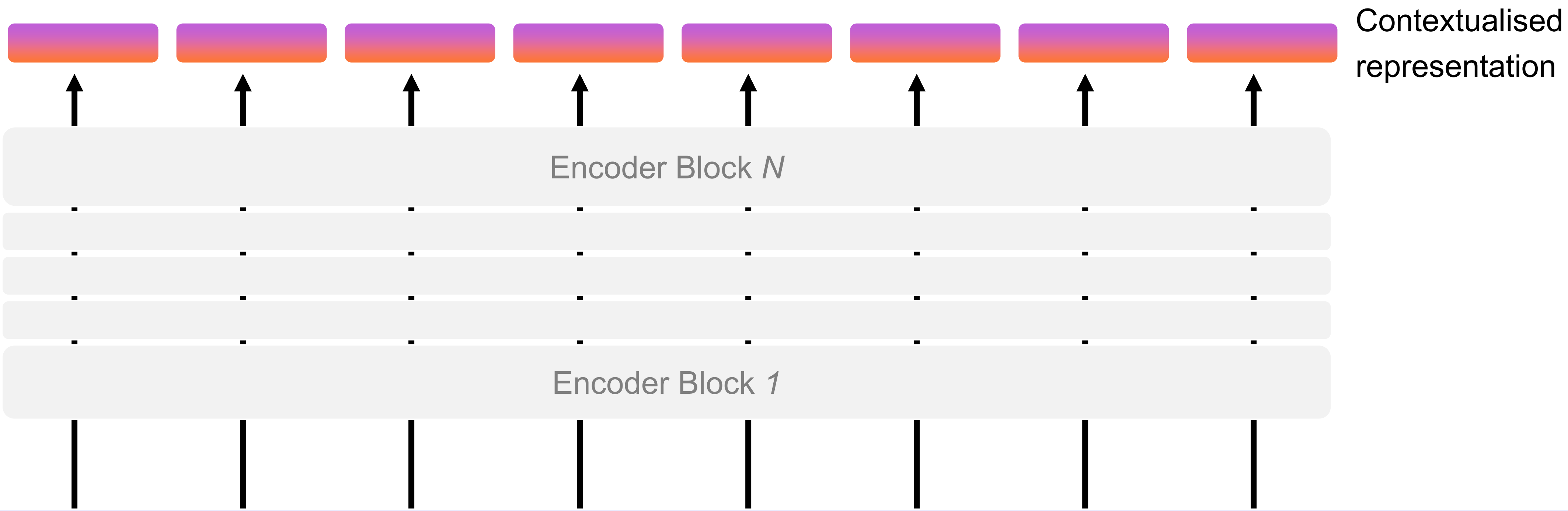


BERT Encoders

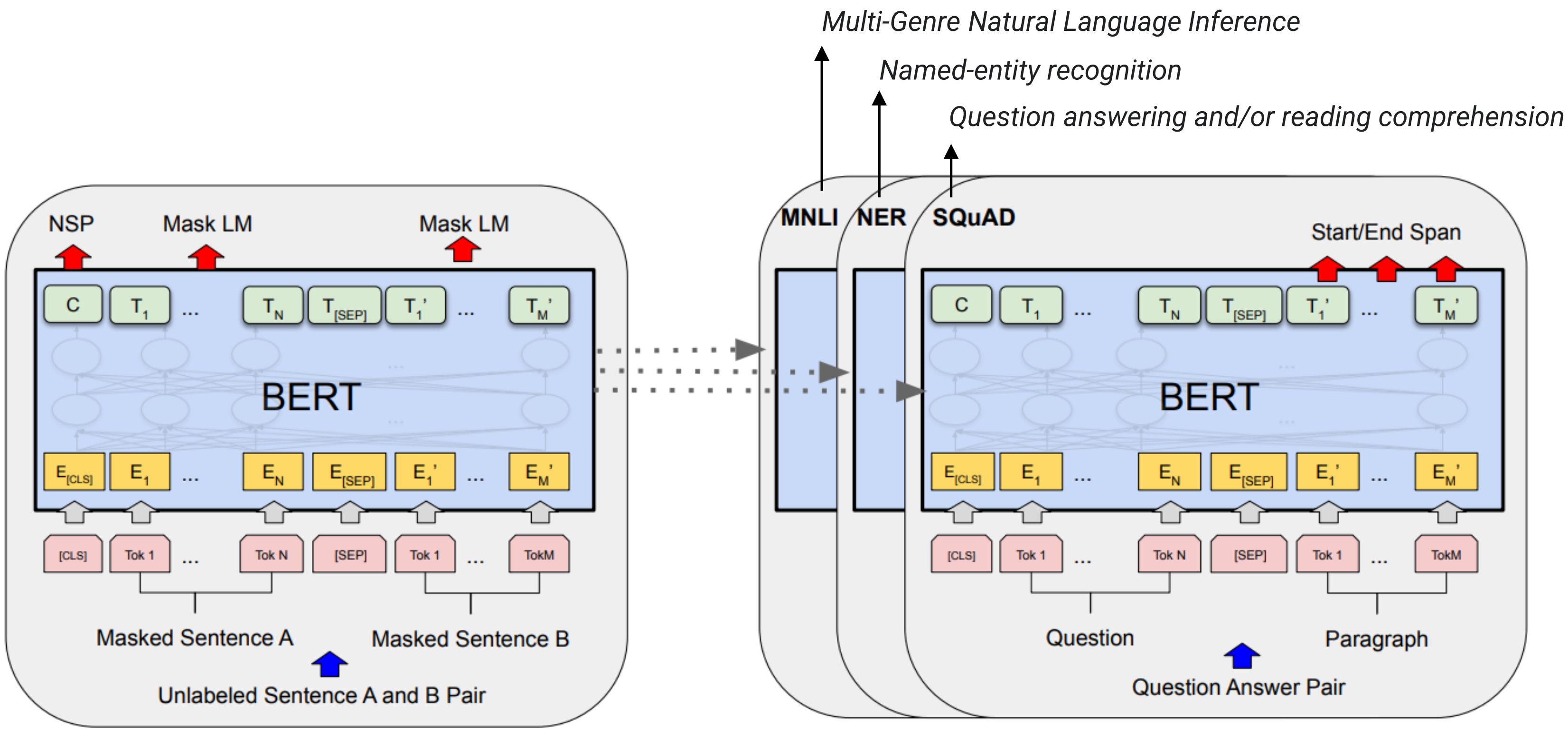
Contextualized token representations **contain rich and nuanced information** about the role of a token in a sequence.

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



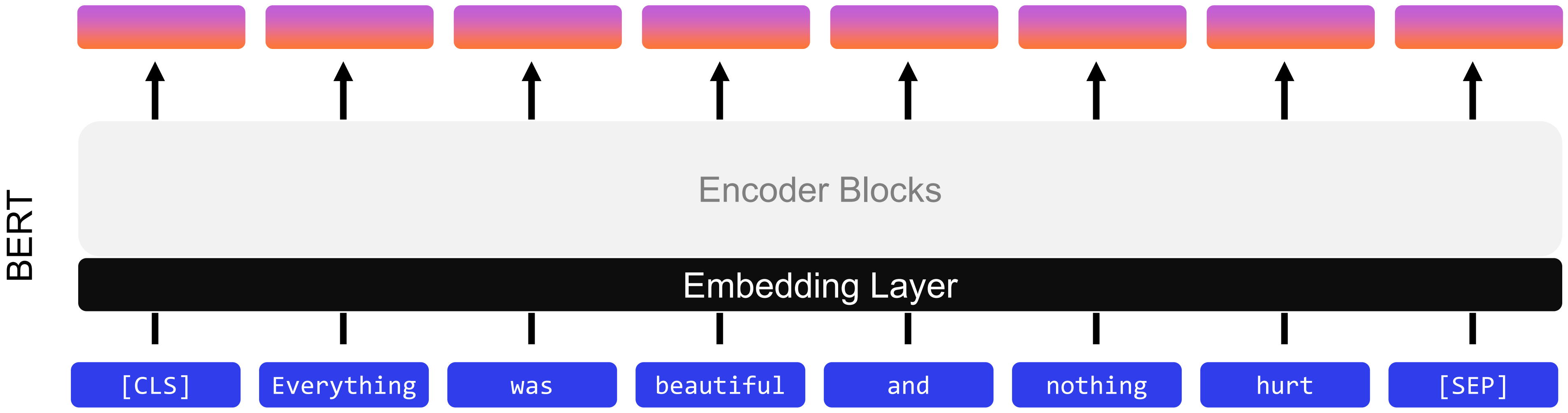
Pre-training

Fine-Tuning

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

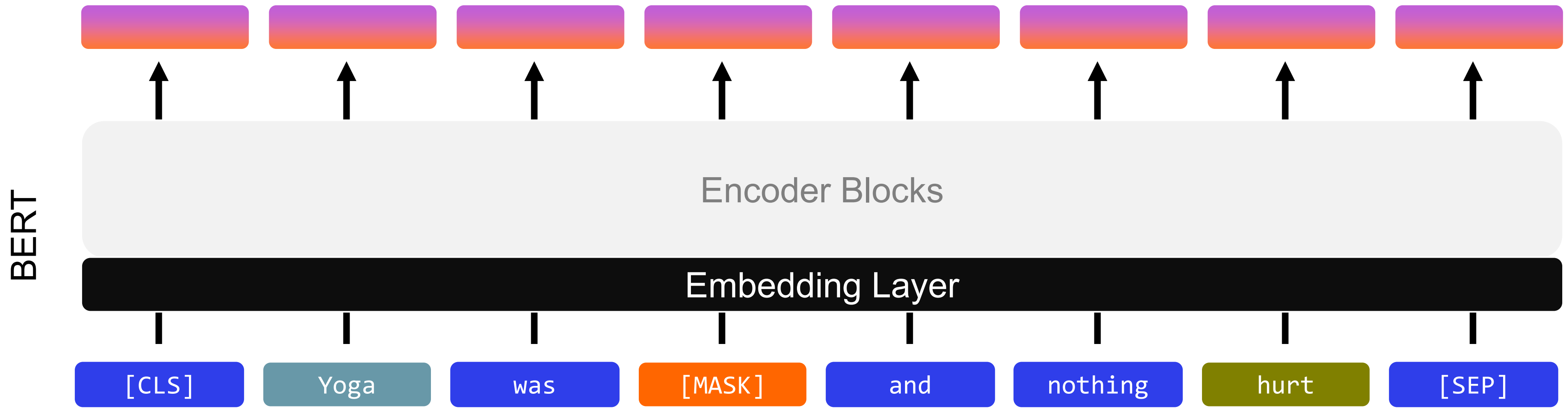
BERT: Pretraining

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

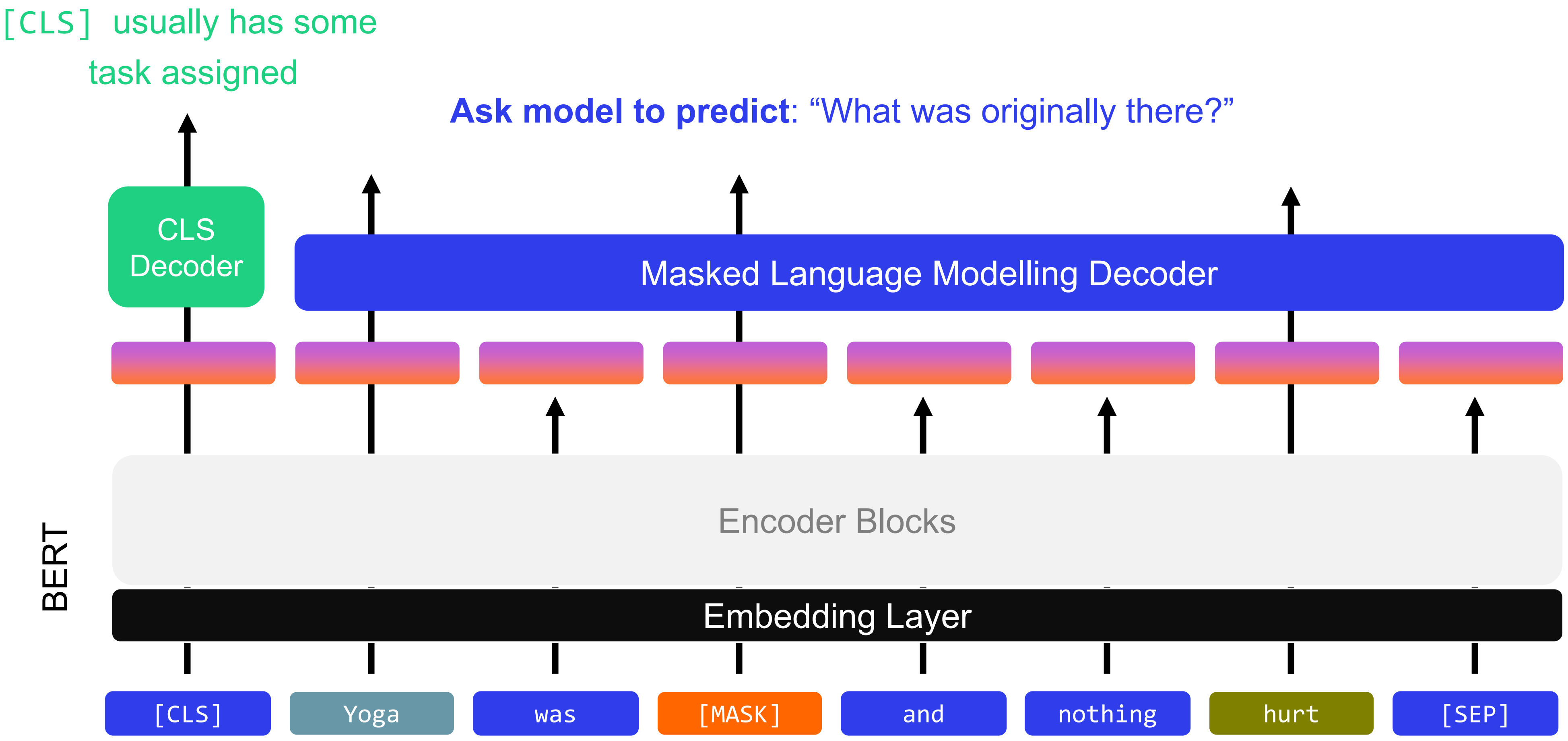


BERT: Pretraining

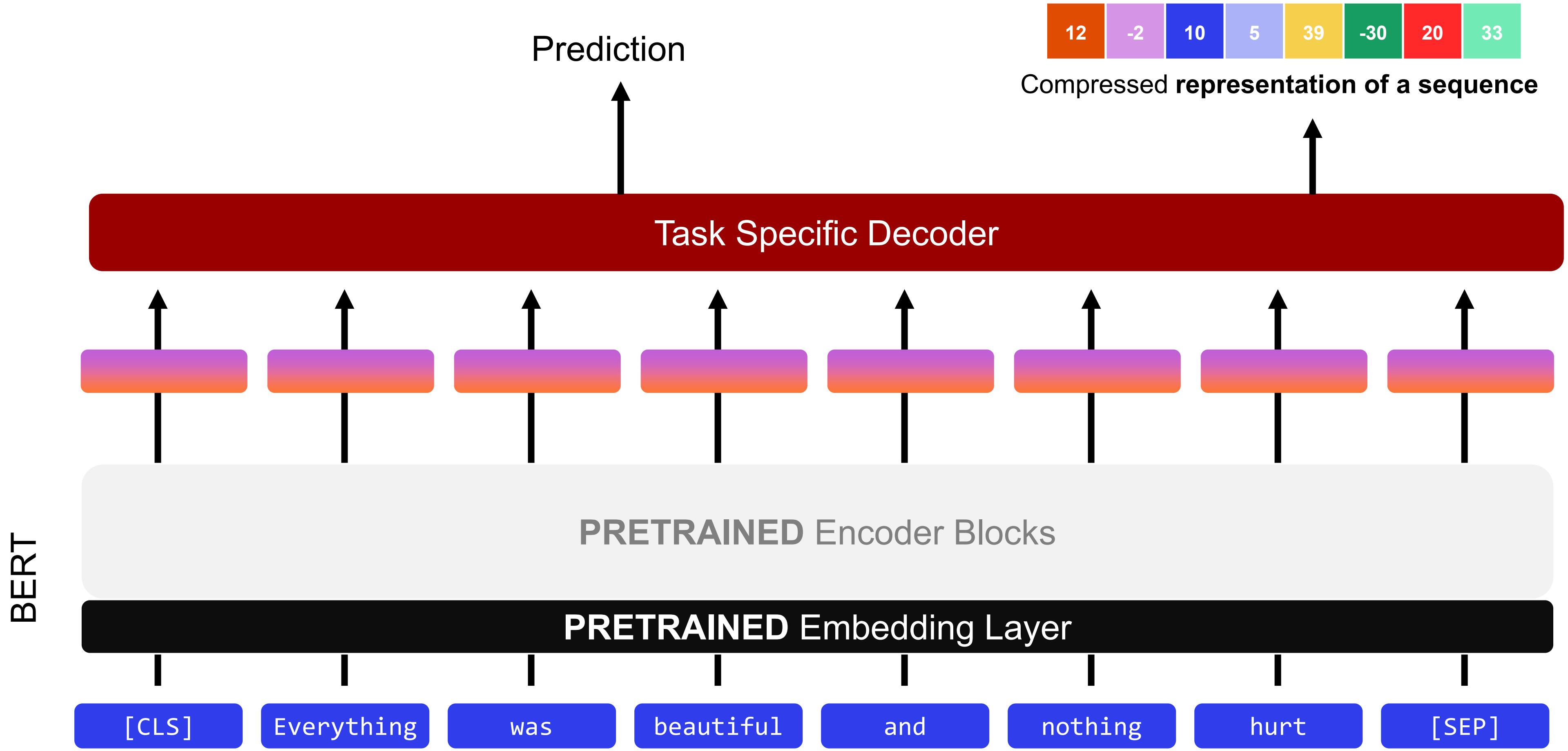
- Mask 15% of tokens (not including **`[PAD]`,`[SEP]`,`[CLS]`**):
  - 10% unchanged
  - 10% substituted with random tokens
  - 80% substituted with the **[MASK]** token



BERT: Pretraining



BERT: Finetuning



Transformer-based Models

BERT

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

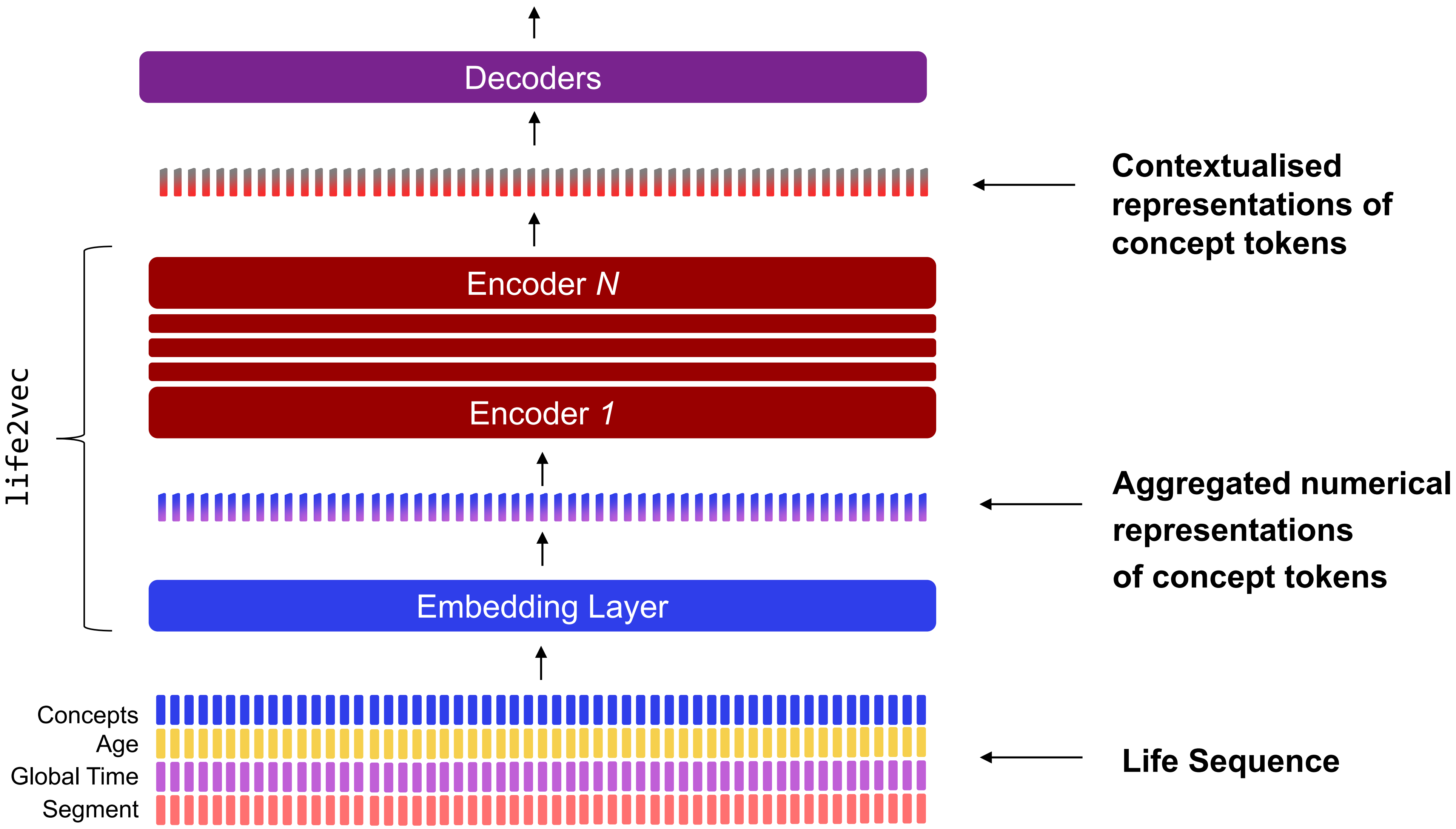
General-purpose model, adaptable to new tasks

Great **predictive performance** on many NLP tasks

LIFE2VEC

Adapts BERT for life-sequences

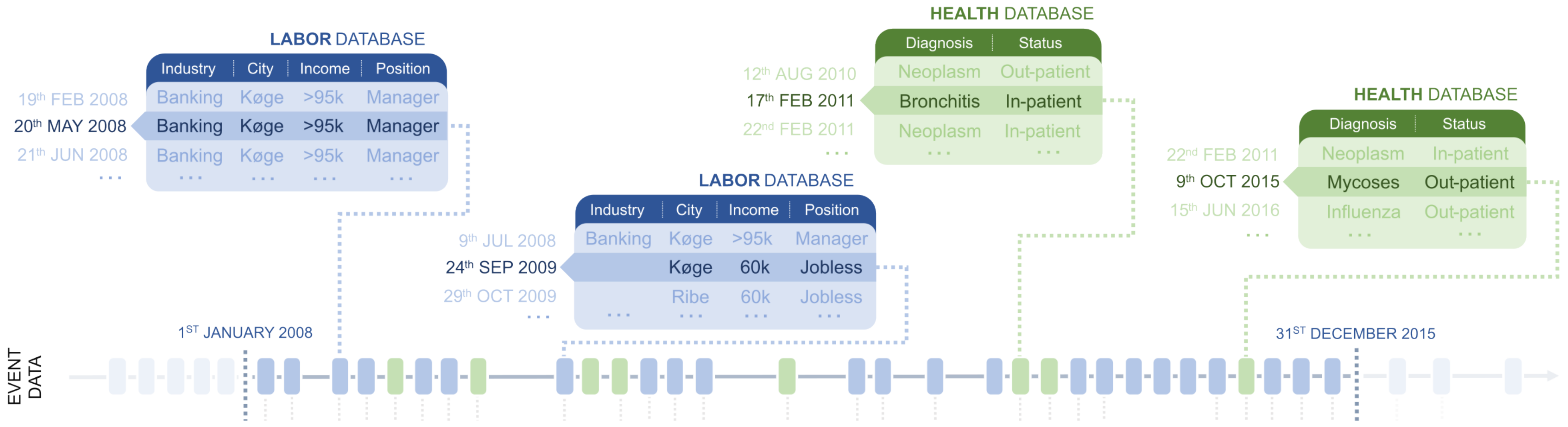
Life2vec: Adaptation of BERT



Part III

Creating Life-Sequences

Unfolding the data



Tabular to Textual Representation?

** slightly simplified overview*

Forming a Language

LABOR DATABASE

	Industry	City	Income	Position
19th FEB 2008	Banking	Køge	>95k	Manager
20th MAY 2008	Banking	Køge	>95k	Manager
21th JUN 2008	Banking	Køge	>95k	Manager
…	…	…	…	…

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

Forming a Language

LABOR DATABASE

	Industry	City	Income	Position
19th FEB 2008	Banking	Køge	>95k	Manager
20th MAY 2008	Banking	Køge	>95k	Manager
21th JUN 2008	Banking	Køge	>95k	Manager
...	...	...	...	...

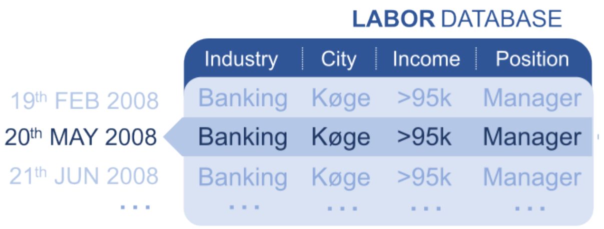
Convey the content in a spoken language

"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

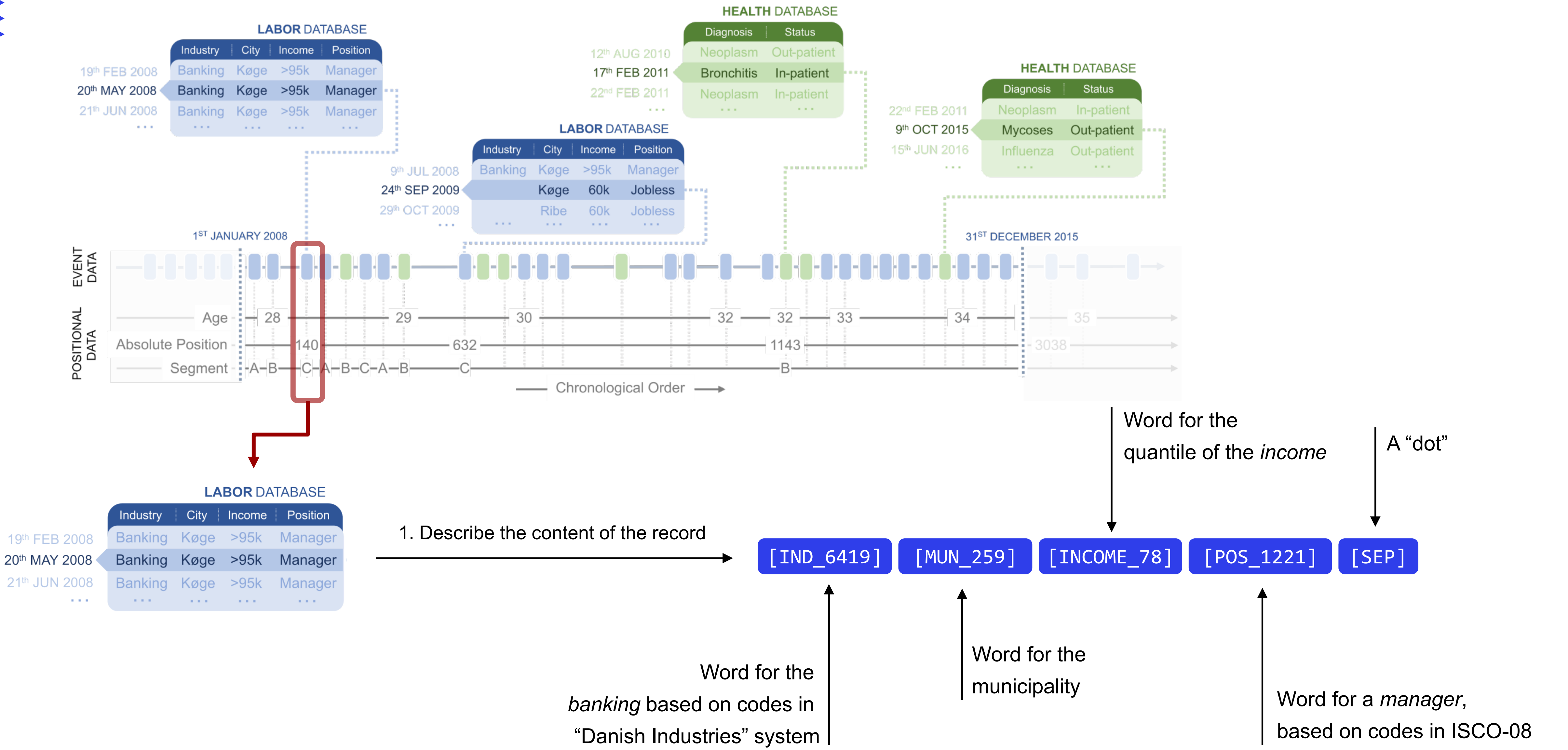
Forming a Language



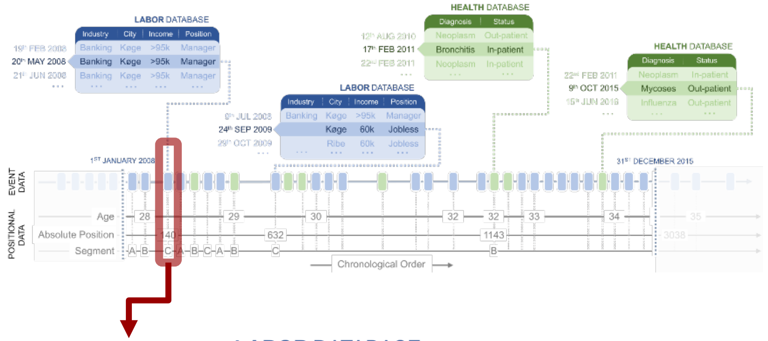
Convey the content in an
artificial symbolic language

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

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



* *slightly simplified overview*



LABOR DATABASE

	Industry	City	Income	Position
19th FEB 2008	Banking	Køge	>95k	Manager
20th MAY 2008	Banking	Køge	>95k	Manager
21th JUN 2008	Banking	Køge	>95k	Manager
...	...	...	...	...

1. Describe the content of the record

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

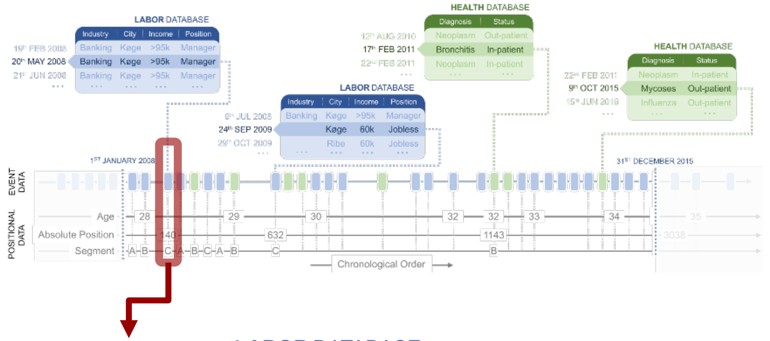
2. Extract positional information about the event

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*



LABOR DATABASE

	Industry	City	Income	Position
19th FEB 2008	Banking	Køge	>95k	Manager
20th MAY 2008	Banking	Køge	>95k	Manager
21th JUN 2008	Banking	Køge	>95k	Manager
...	...	...	...	...

1. Describe the content of the record

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

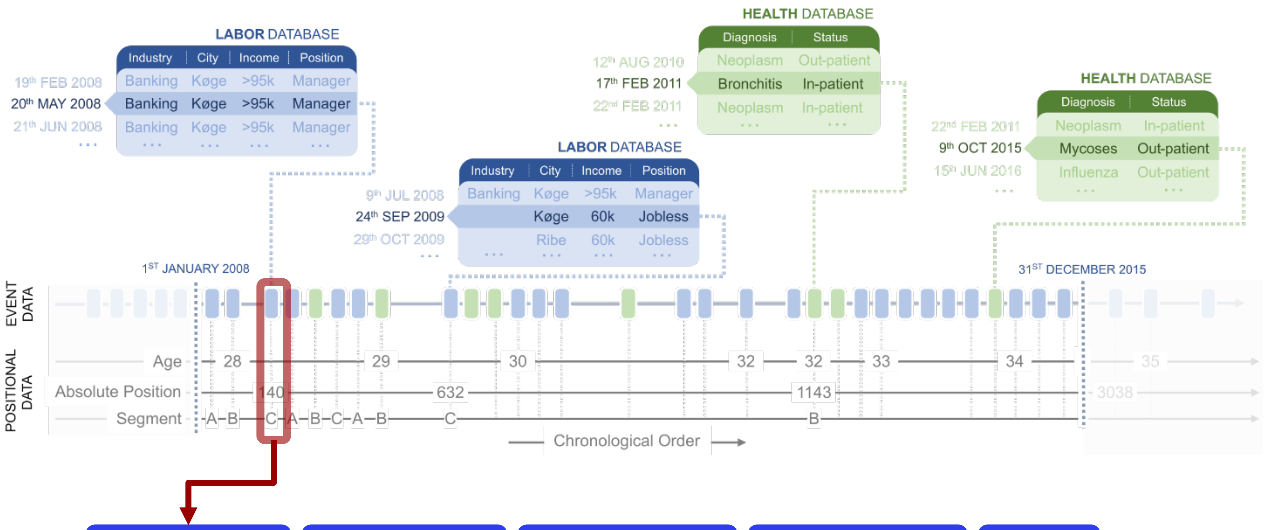
2. Extract positional information about the event

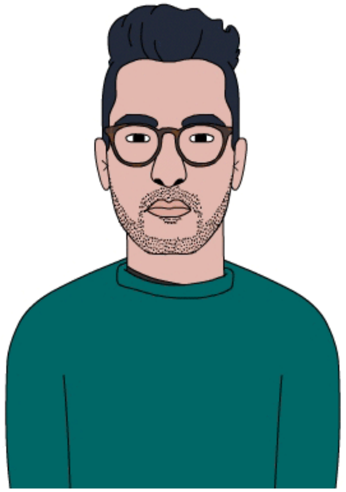
Age: 28
Global timestep: 140
Segment: C

3. Assign this information to tokens

28	28	28	28	28
140	140	140	140	140
C	C	C	C	C

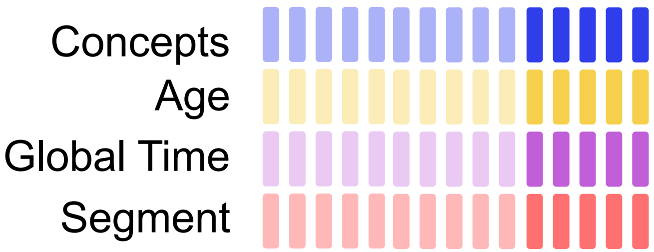
** slightly simplified overview*



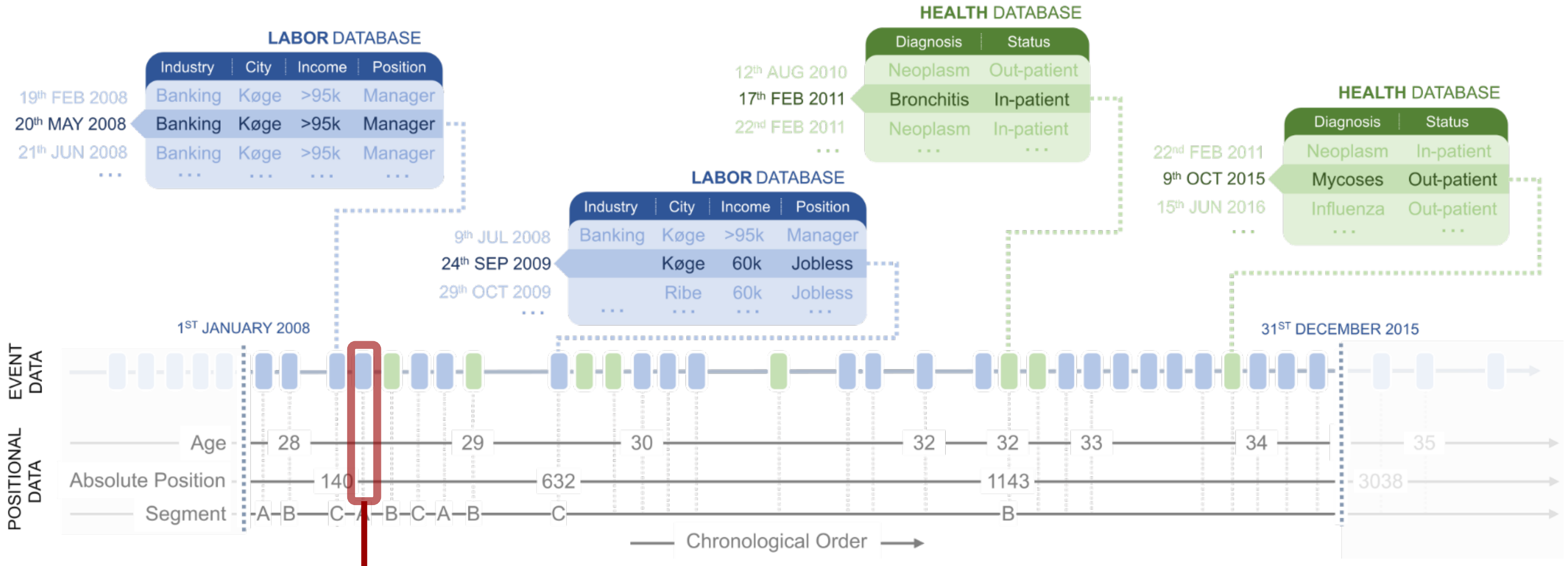


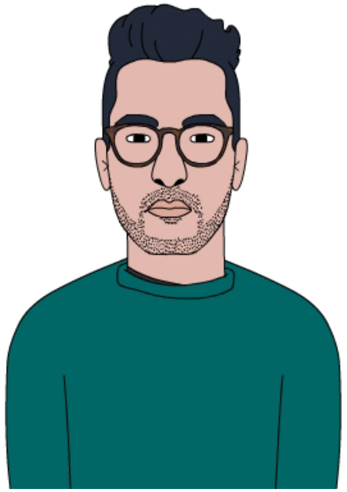
[IND_6419]	[MUN_259]	[INCOME_78]	[POS_1221]	[SEP]
28	28	28	28	28
140	140	140	140	140
C	C	C	C	C

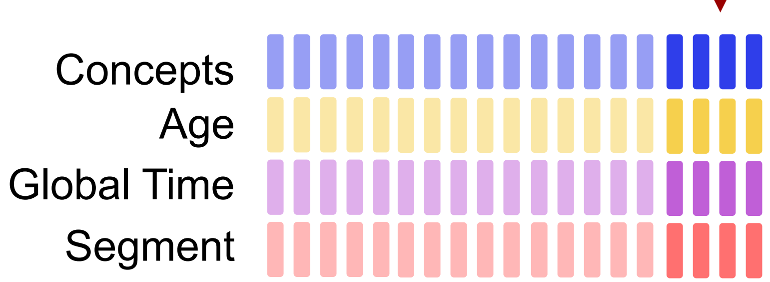
4. Insert data into the Life-Sequence (person document)



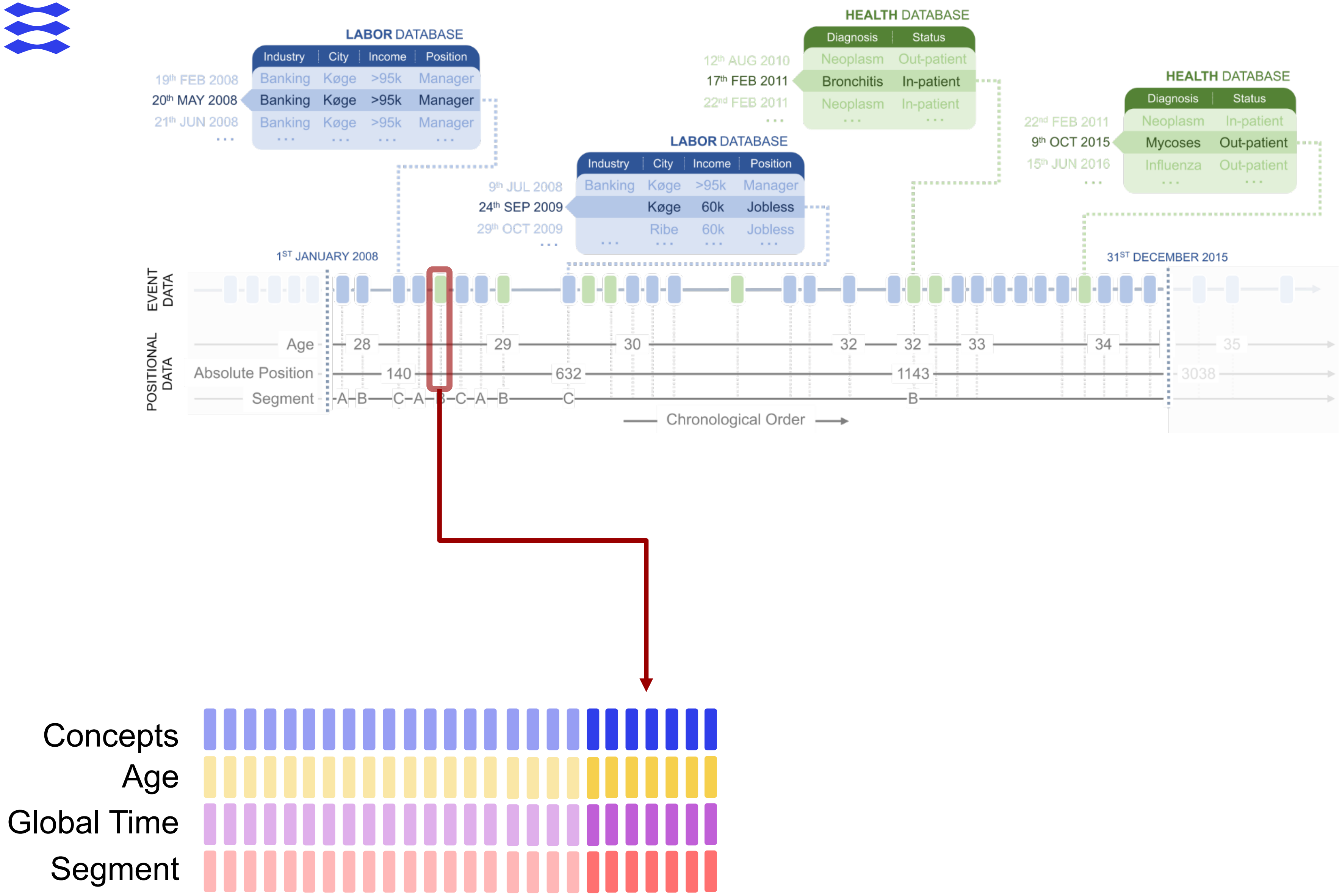
** slightly simplified overview*

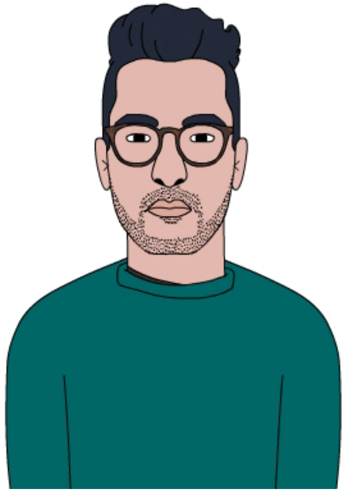




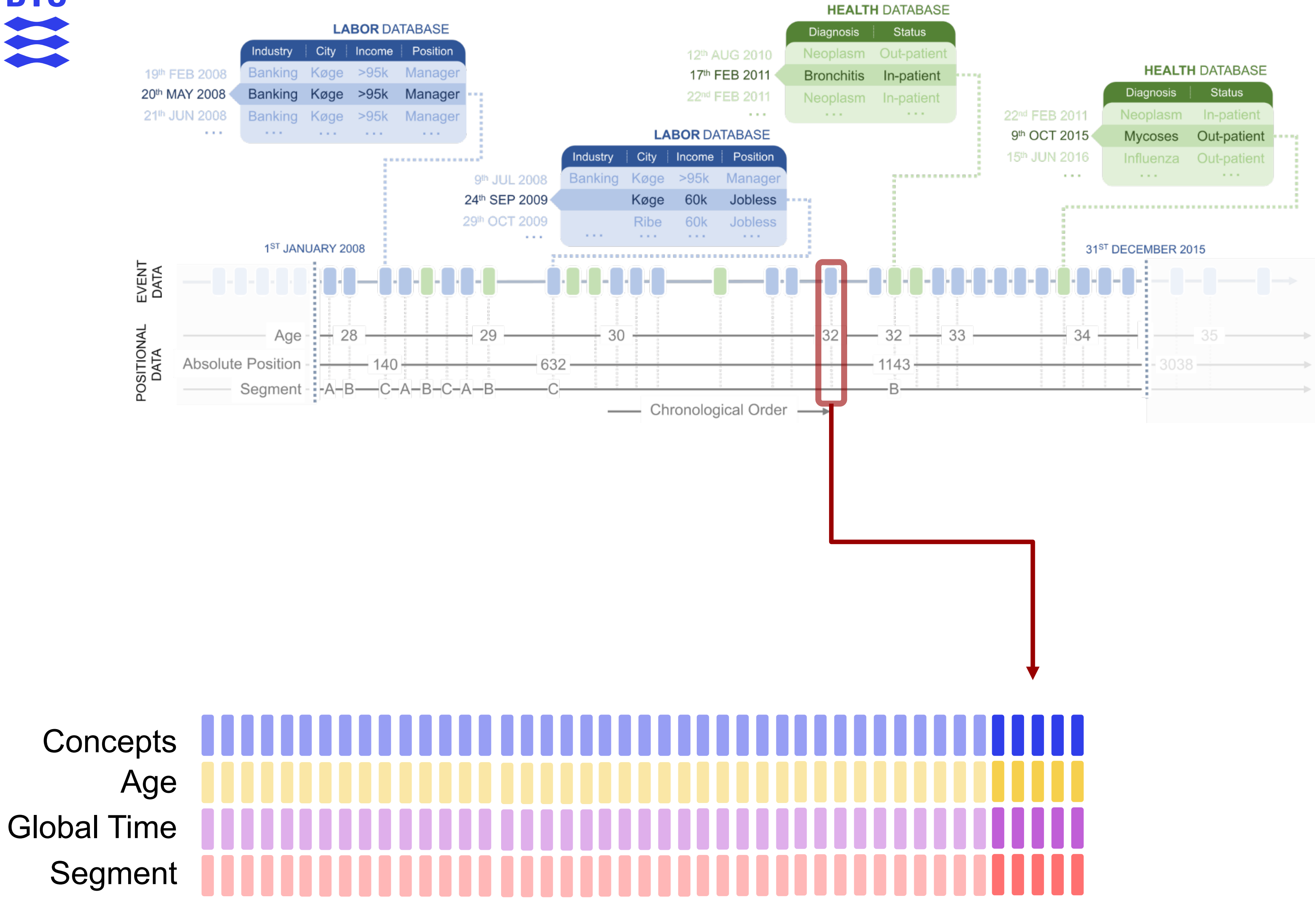


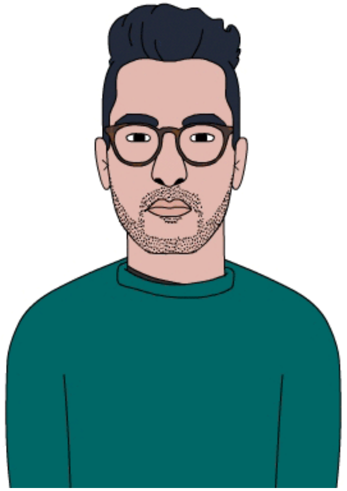
** slightly simplified overview*



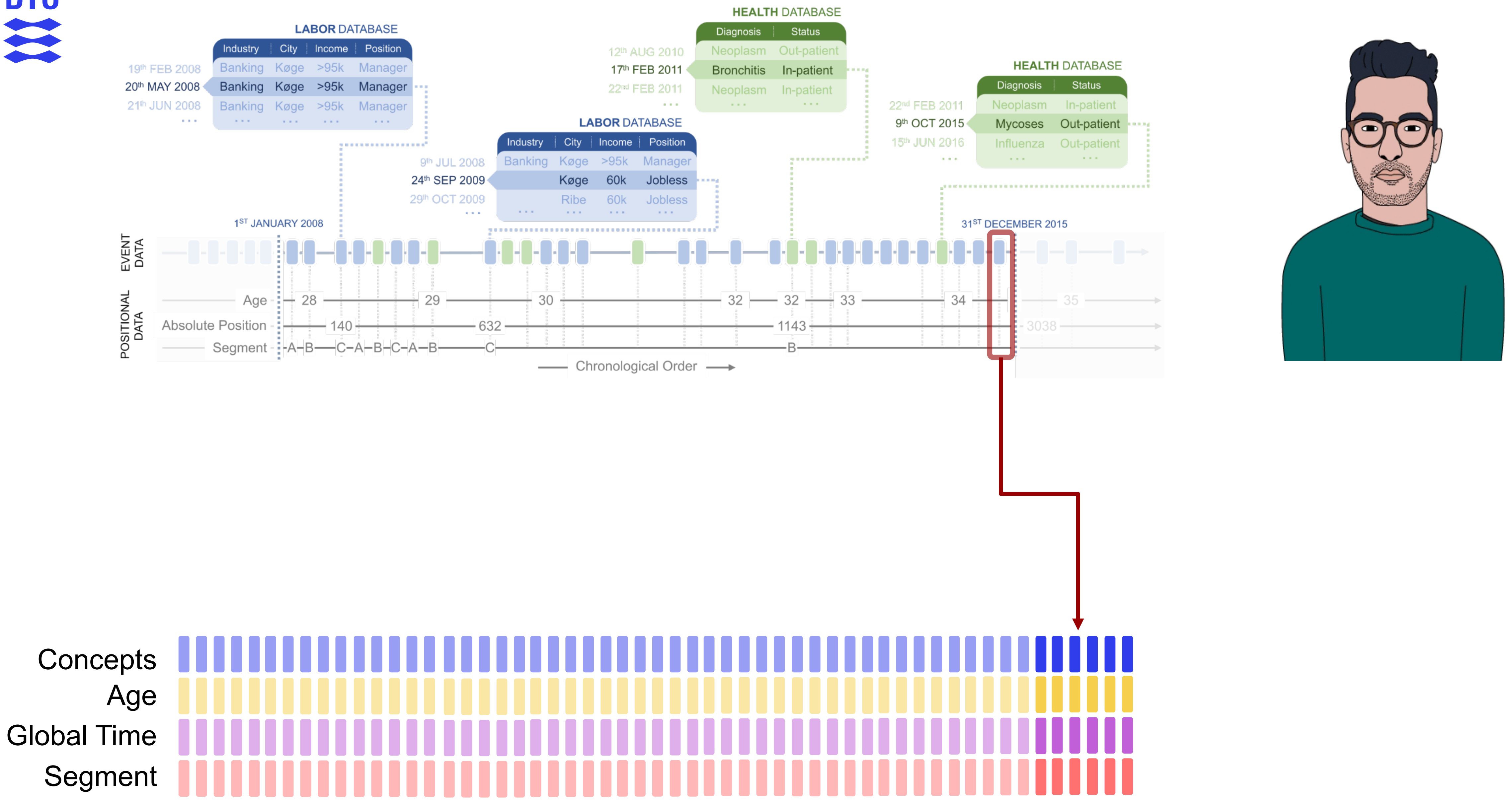


* *slightly simplified overview*

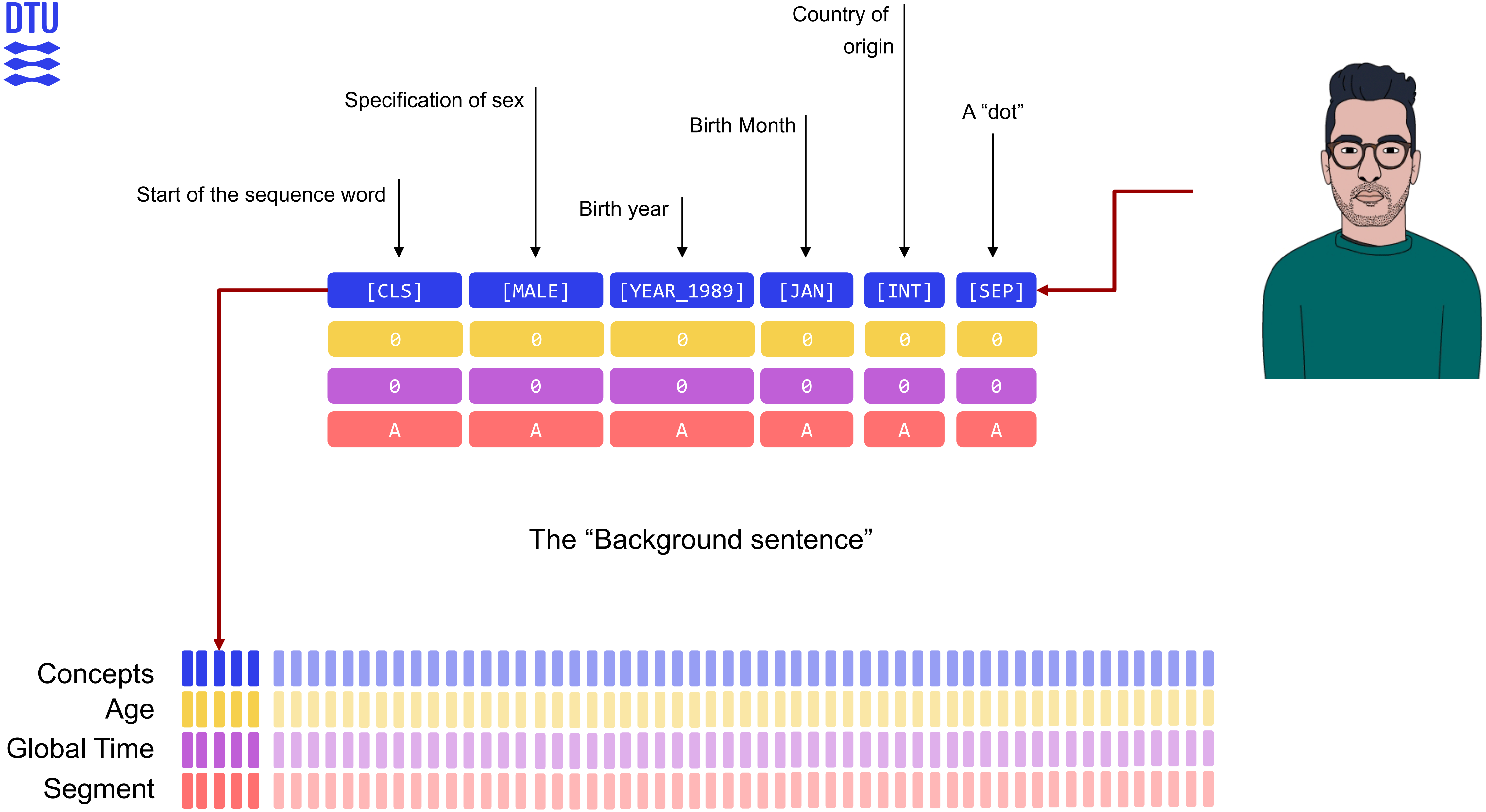




* *slightly simplified overview*

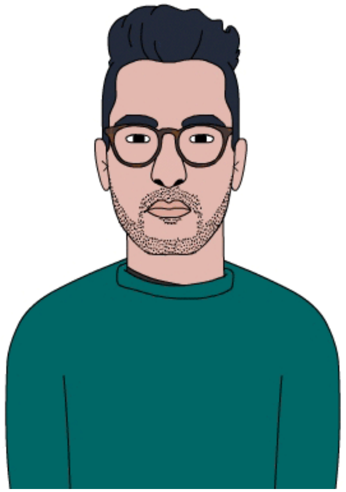


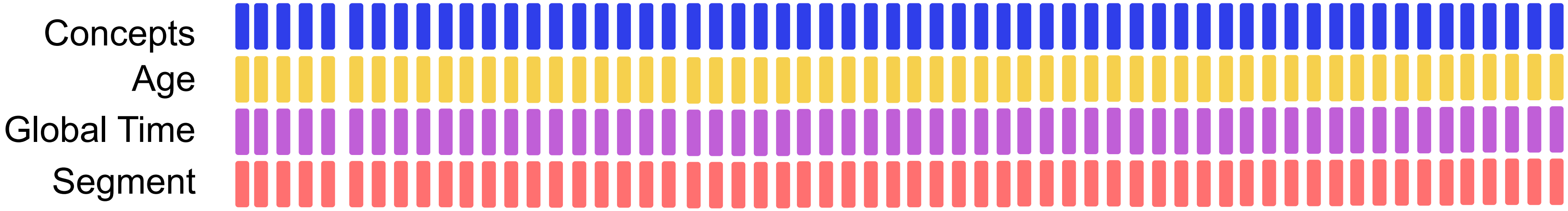
* *slightly simplified overview*



** slightly simplified overview*

Individual Life-Sequence





Input to the `life2vec` model

** slightly simplified overview*

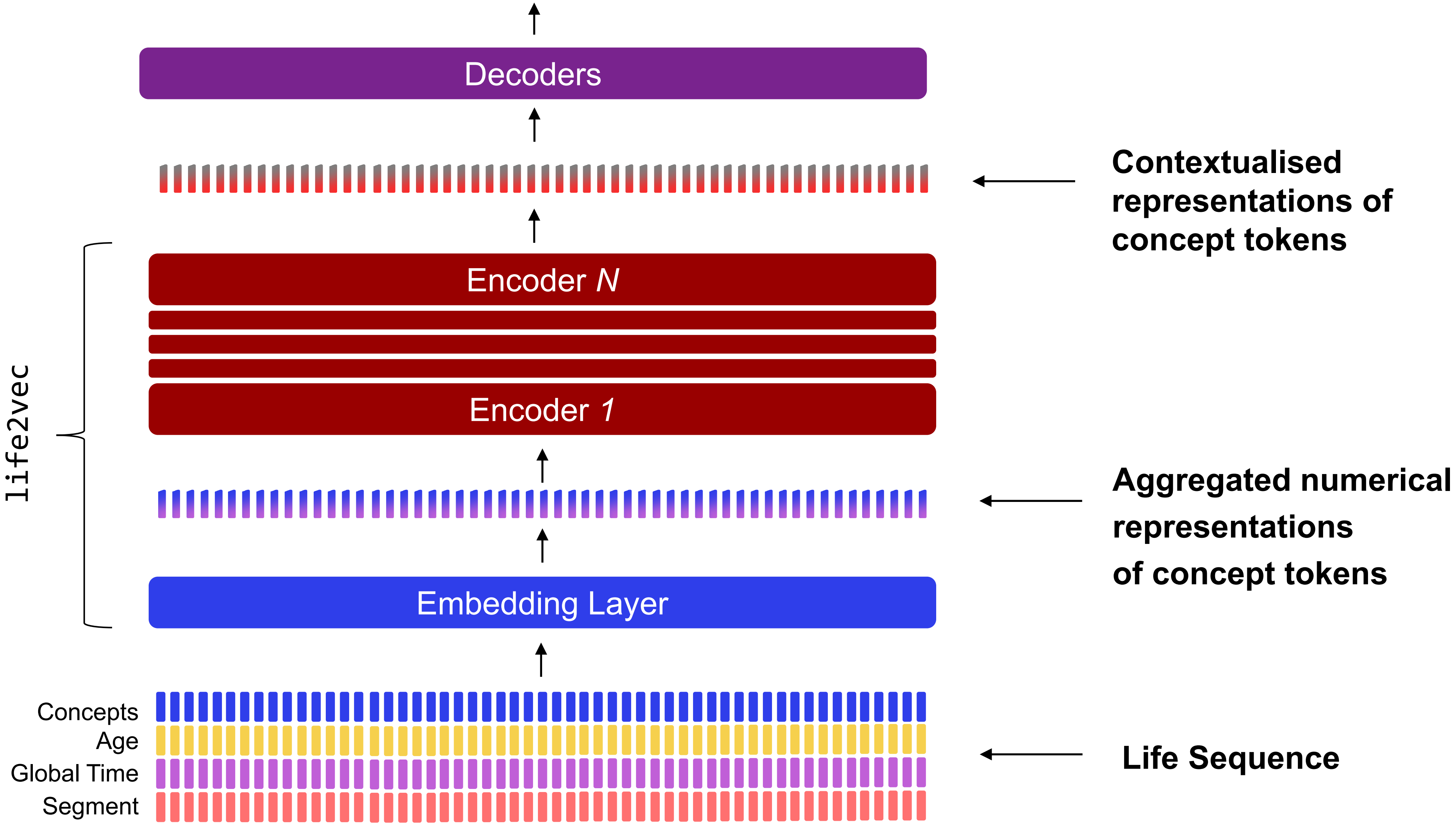
Vocabulary

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

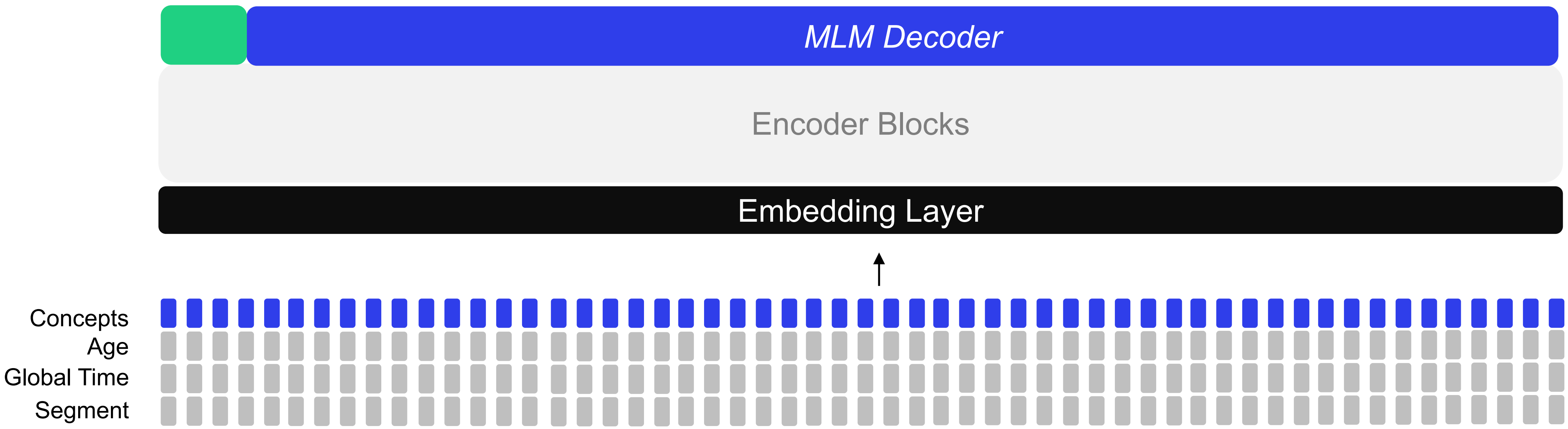
life2vec: capturing the structure

life2vec pipeline

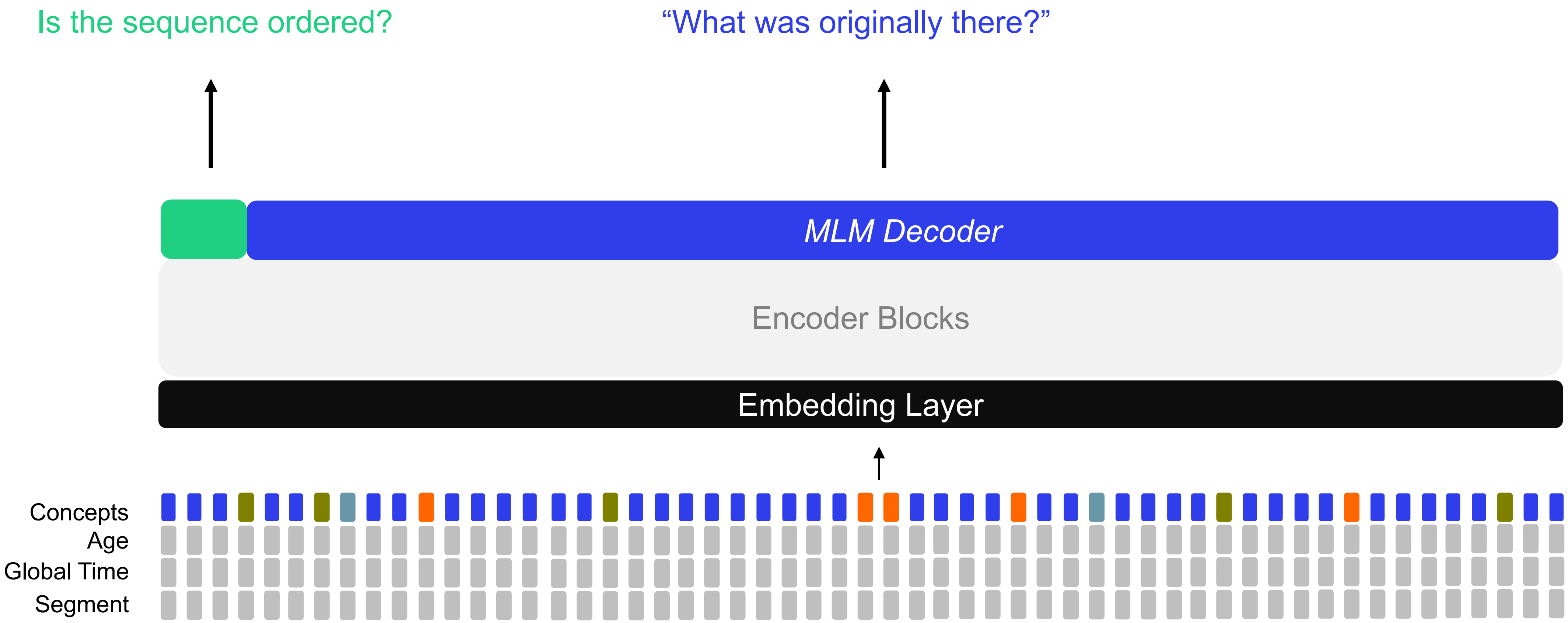


`life2vec`: pre-training

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

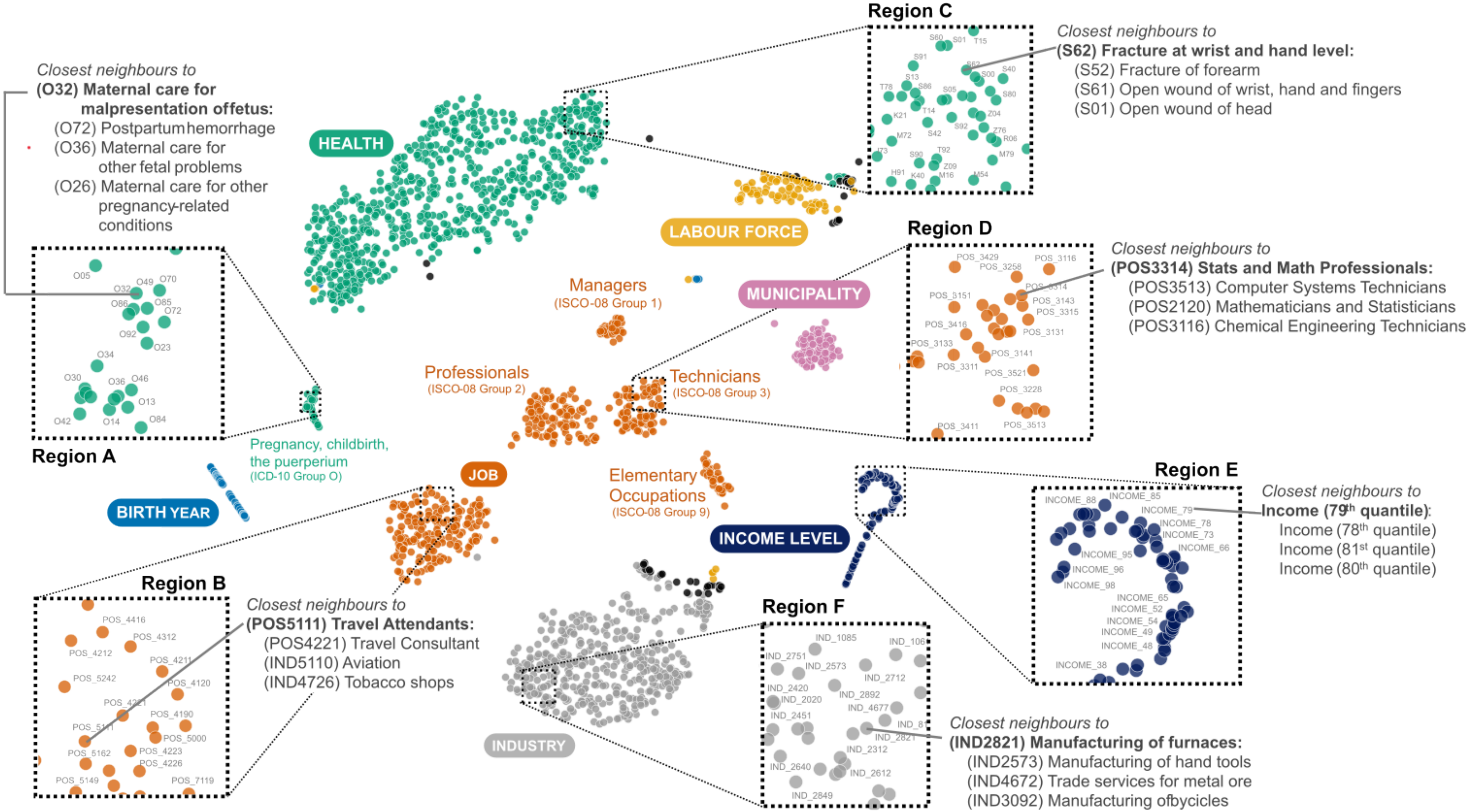


life2vec: pre-training



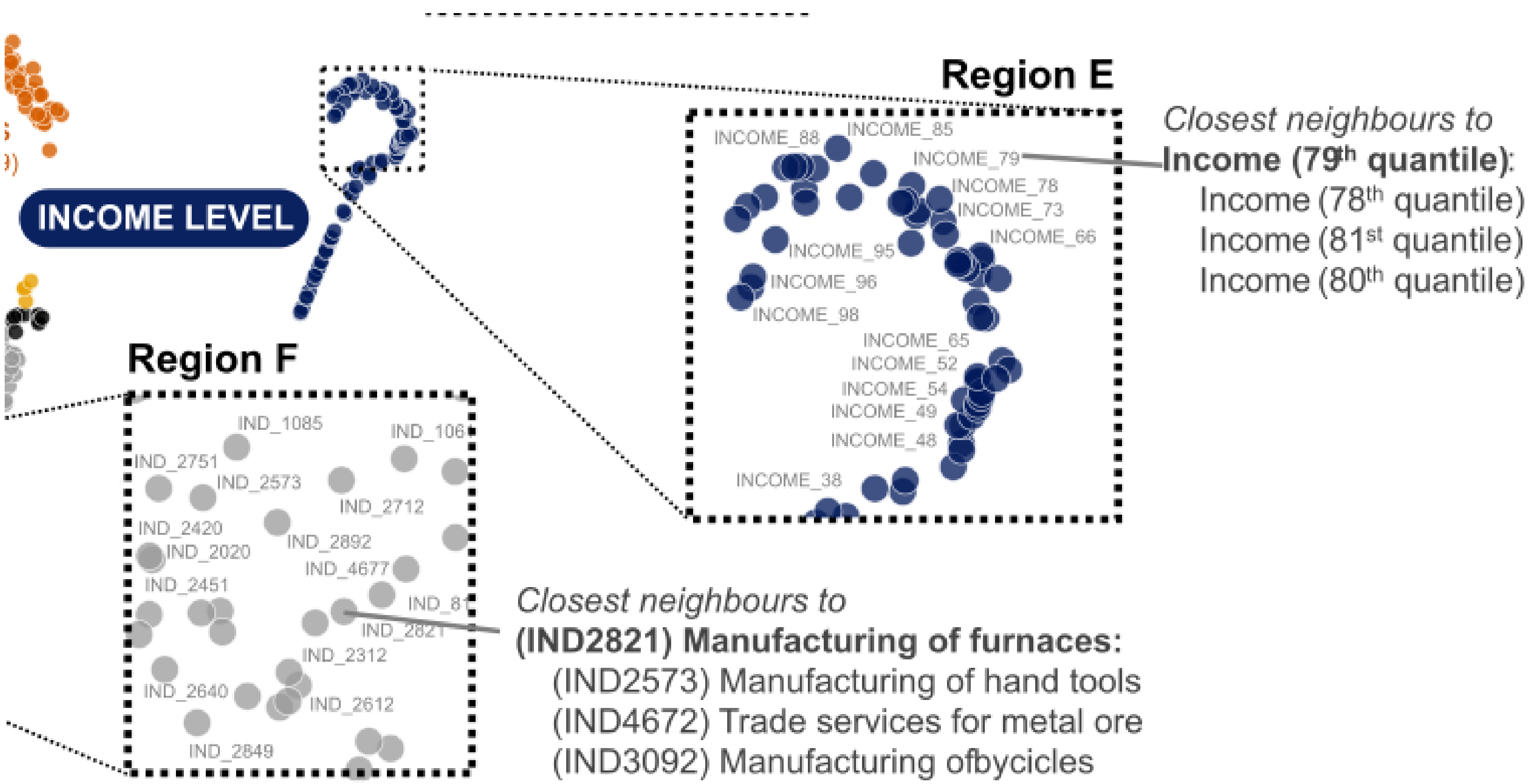
What did our model learn on pretraining?

Space of Concept Tokens (with PaCMAP)



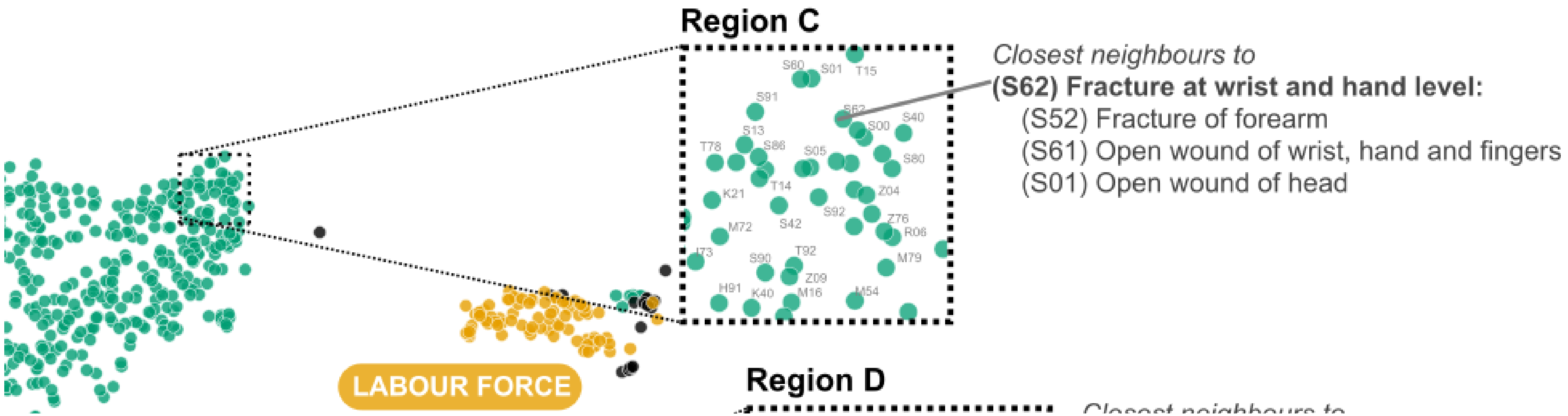
Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1-14.

Space of concept tokens (with PaCMAP)

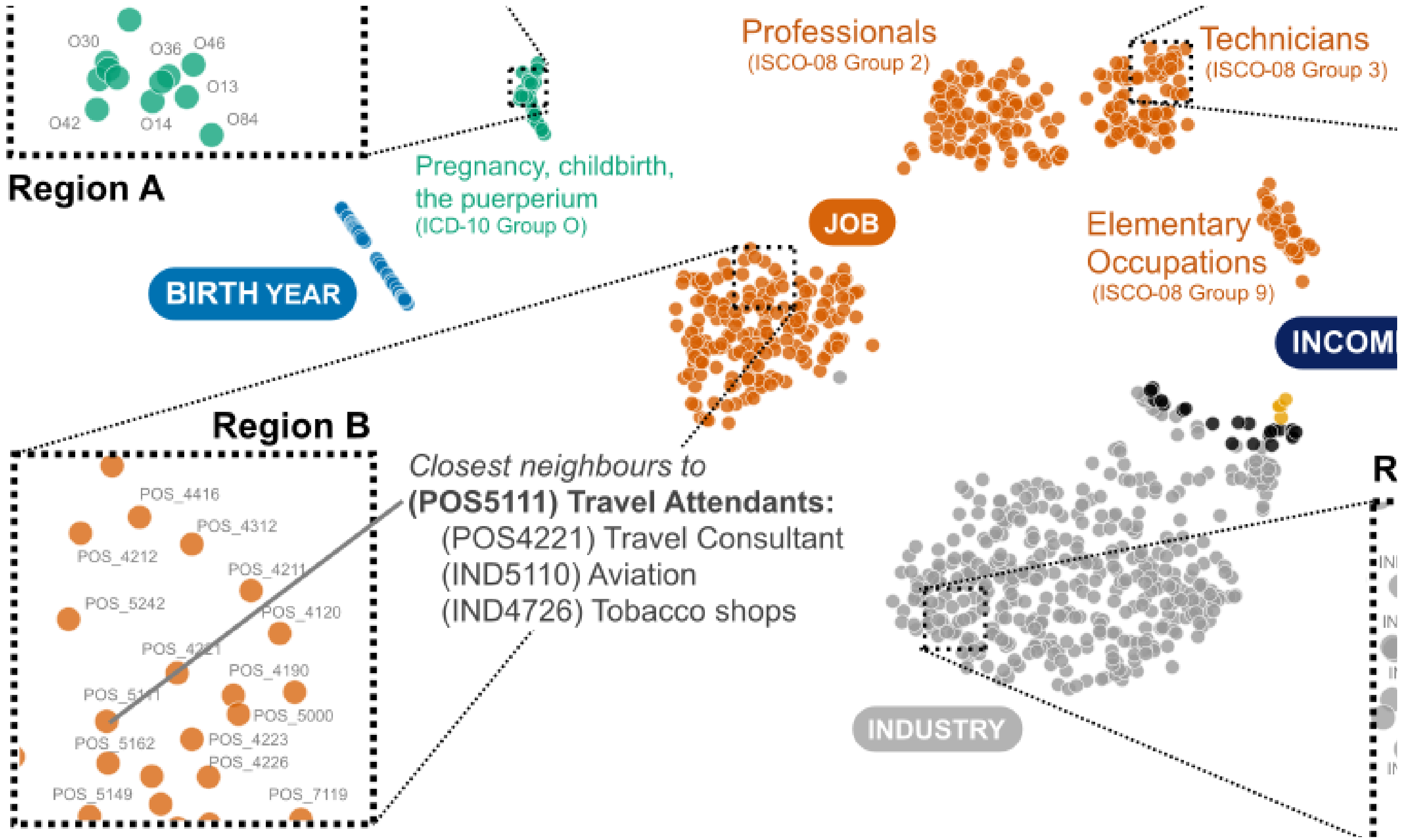


DTU

Space of concept tokens (with PaCMAP)

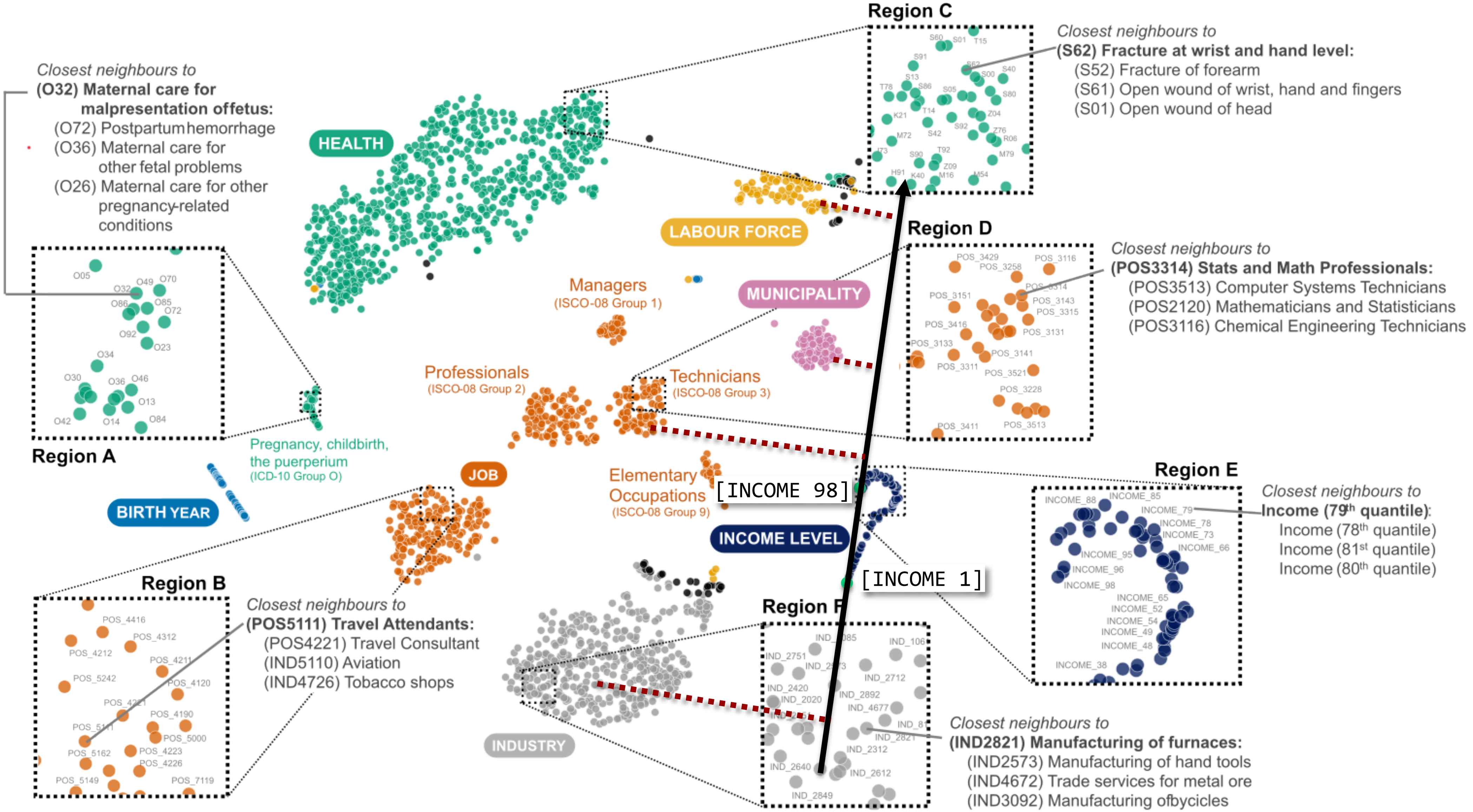


Space of concept tokens (with PaCMAP)



Visually structure corresponds to the structure of the variables

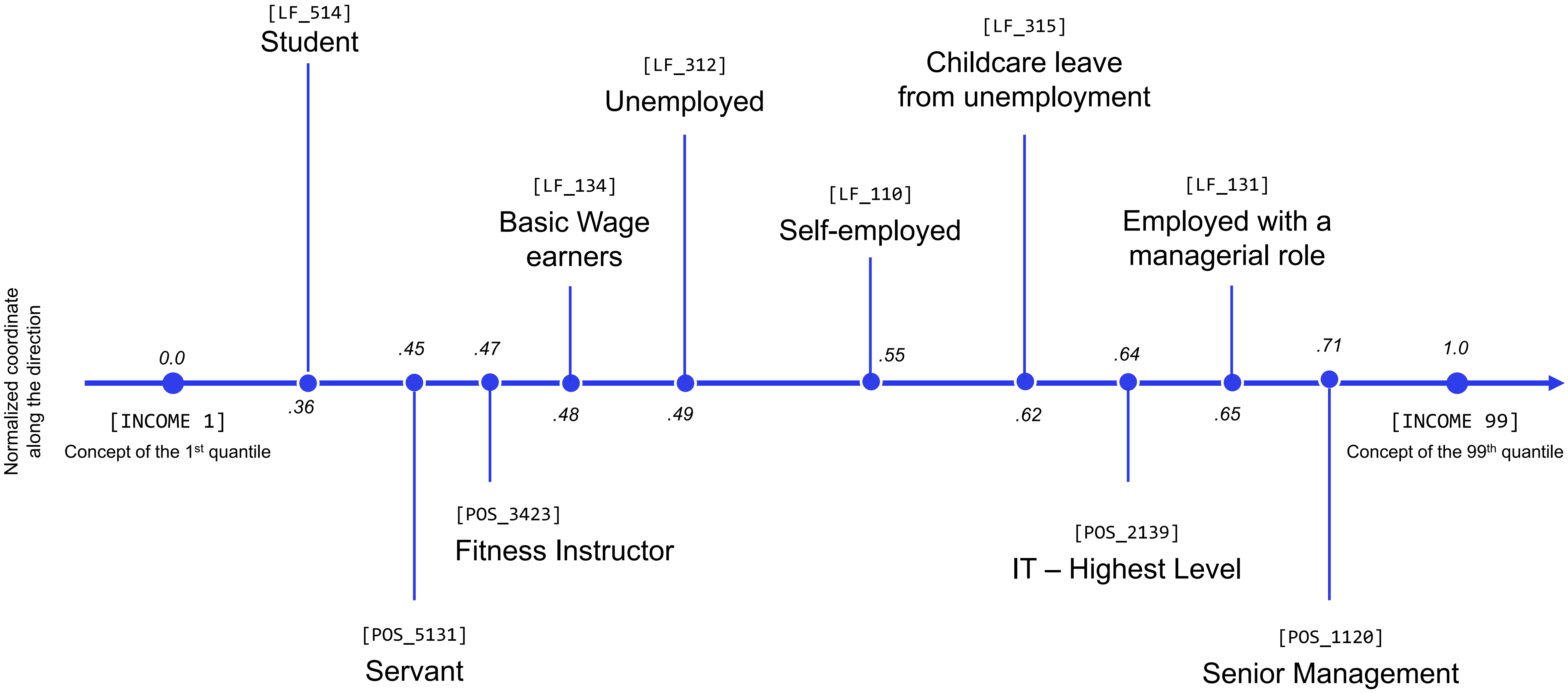
Space of Concept Tokens (with PaCMAP)



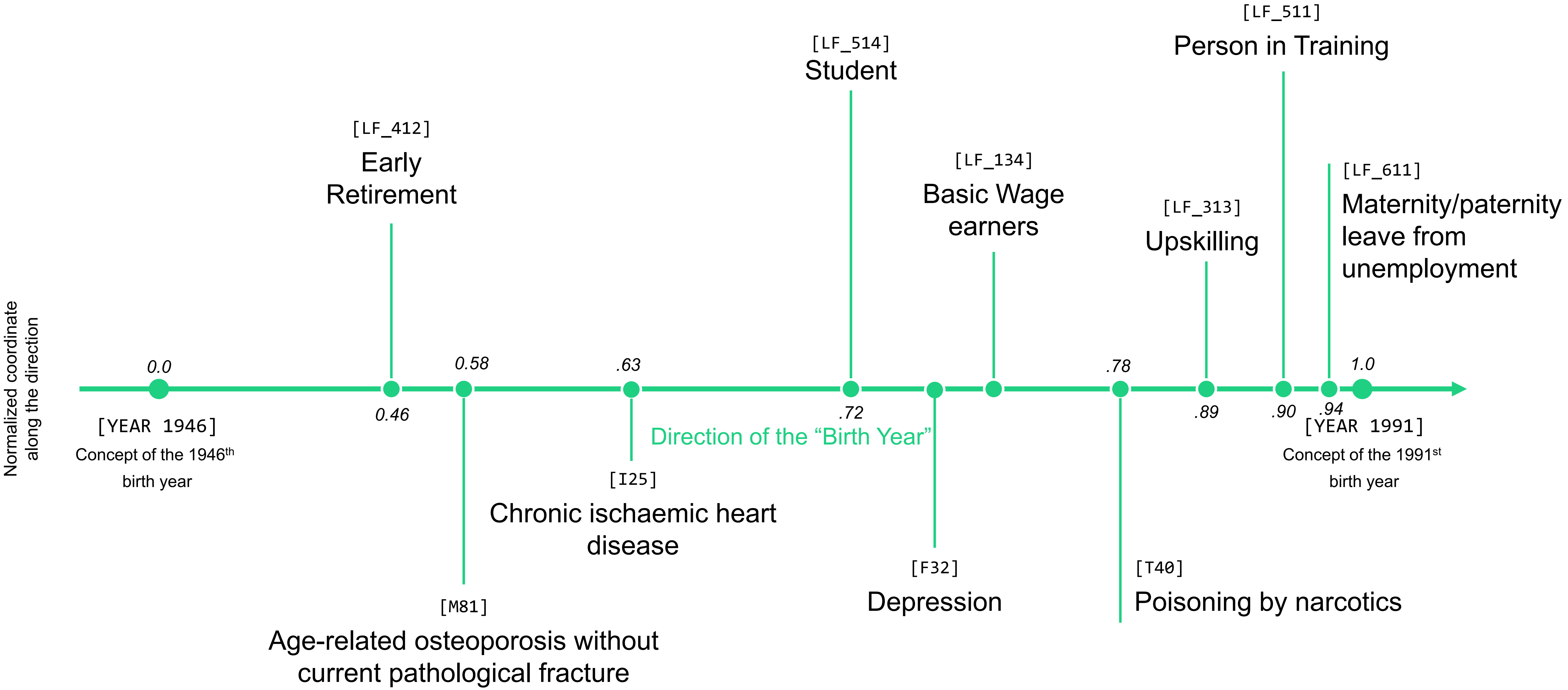
Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1-14.

LF – Labor Force Status
POS – Prof. Position

Projection to "*Income*" Direction



Projection to "*Year*" Direction



Projection to "*Occupation*" Direction

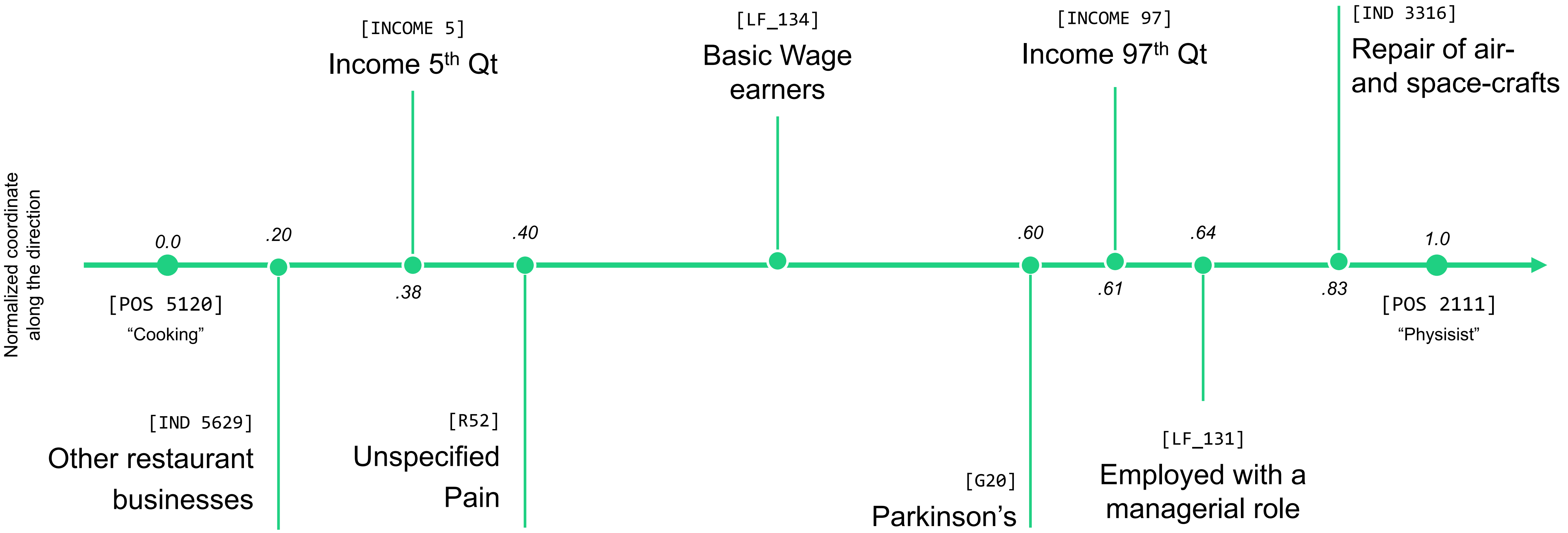
The opposite job of a chef and head cook is a physicist.

Chefs and Head Cooks use these skills the most	
1	Management of material resources
2	Management of financial resources
3	Management of personnel resources
4	Coordination
5	Negotiation
6	Monitoring
7	Time management
8	Persuasion
9	Social perceptiveness
10	Learning strategies

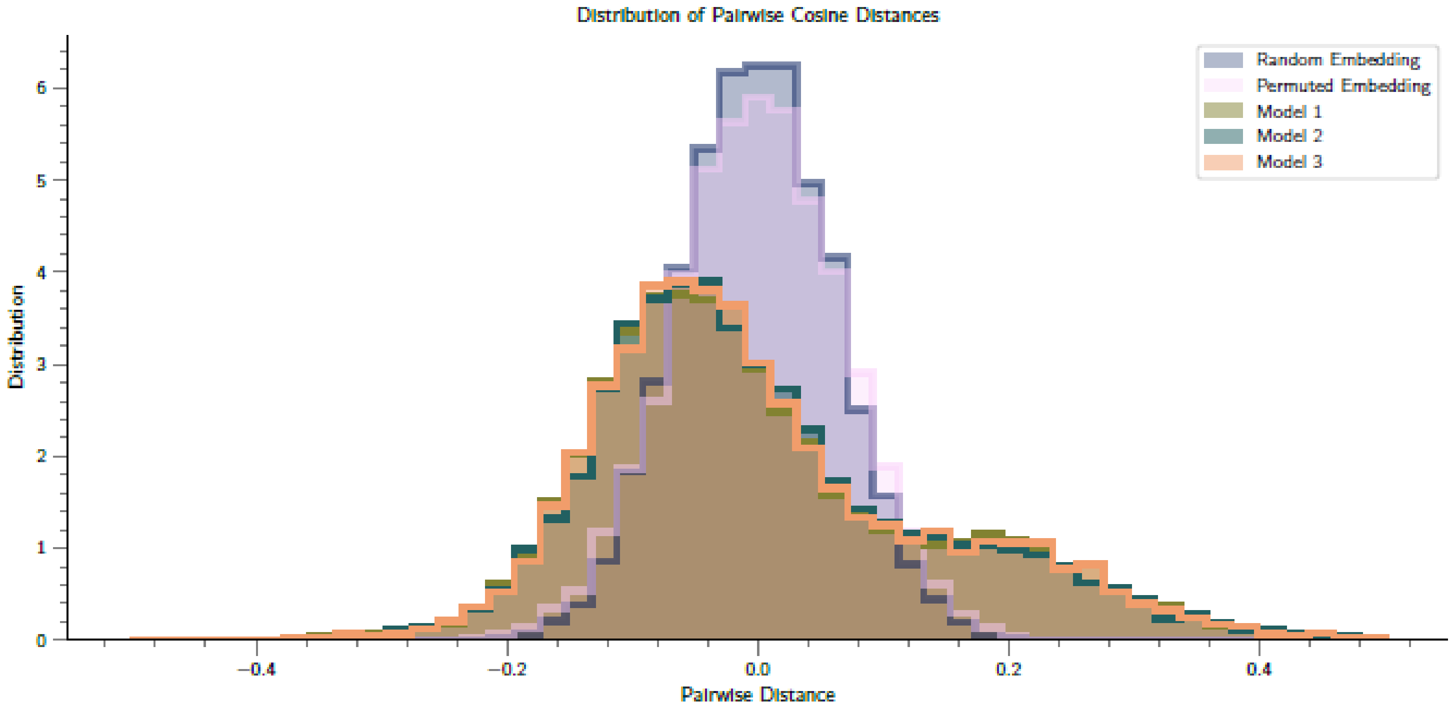
Physicists use these skills the most	
1	Physics
2	Mathematical reasoning
3	Number facility
4	Ability to organize groups in different ways
5	Information ordering
6	Mathematics
7	Oral comprehension
8	Mathematics
9	Originality
10	Speech clarity

(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

Projection to “*Occupation*” Direction



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

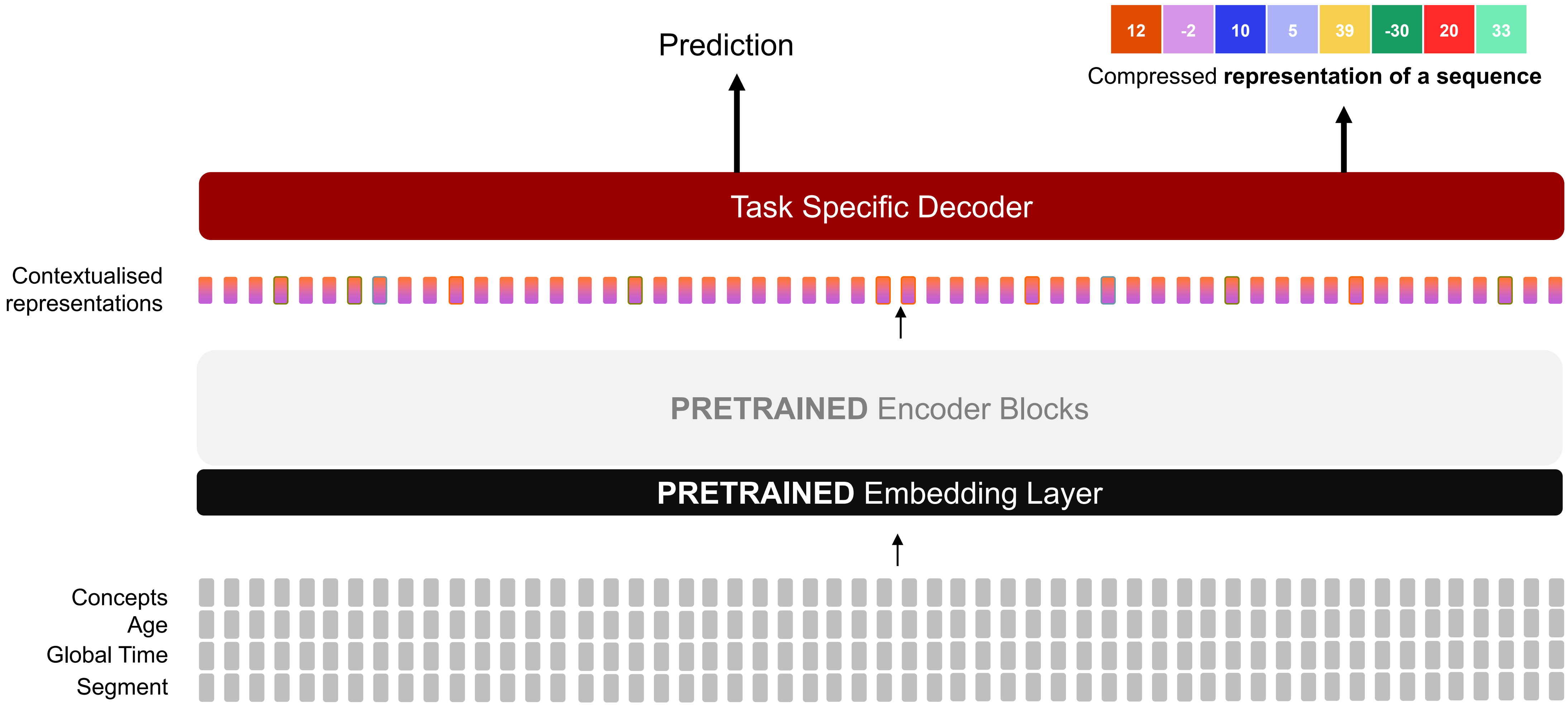
Foundation Models

"Train one model on a huge amount of data and ***adapt it to many applications****. We call such a model a foundation model."* [1]

"[…] *rather than developing a bespoke model for each specific use case (as was done traditionally), a single FM can instead be* ***reused across a broad range of downstream tasks*** *with minimal adaptation or retraining needed per task."*[2]

1. *Developing and understanding responsible foundation models.* Stanford CRFM. (n.d.). https://crfm.stanford.edu/
2. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, *6*(1), 135.

life2vec: finetuning



Life-Summaries

- We want **high predictive power** and **explainability**
- We condition life2vec on three tasks:
  - Early Mortality Prediction
  - Emigration Prediction
  - Self-reported personality assessment

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.

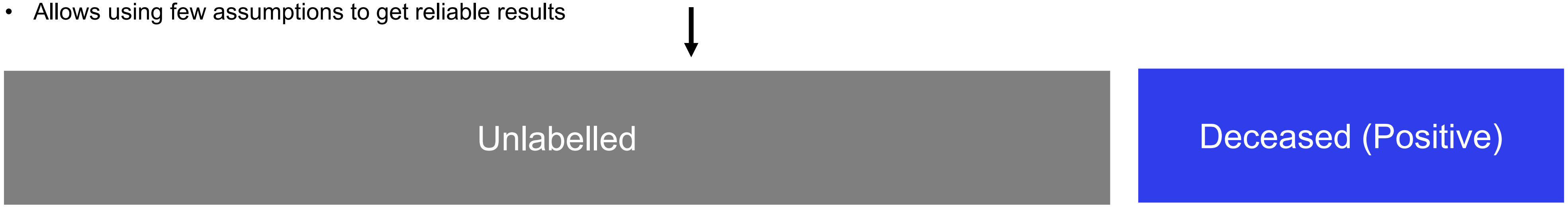
Early Mortality Prediction

– **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

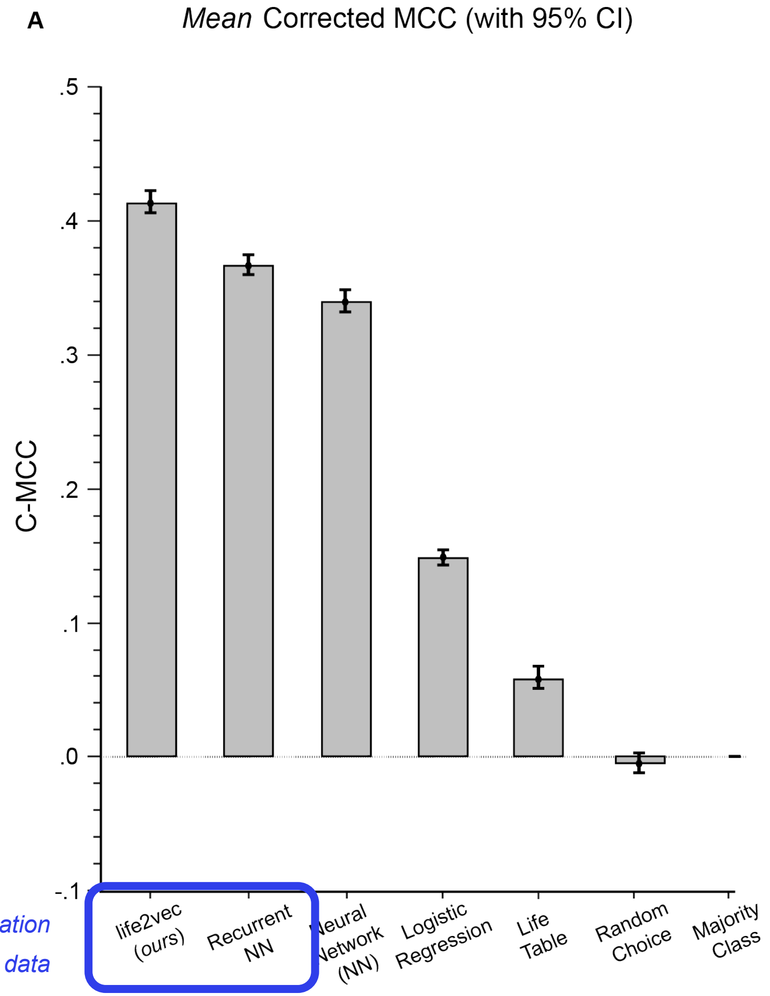
Why PU Learning? (Mortality Example)

Using the PU approach, we can assume that negatives and unlabeled samples are all part of the unlabeled set:

- Allows using few assumptions to get reliable results



Early Mortality Prediction



Use textual representation of the data

True Labels / **Predicted Labels**

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

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

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

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

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

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

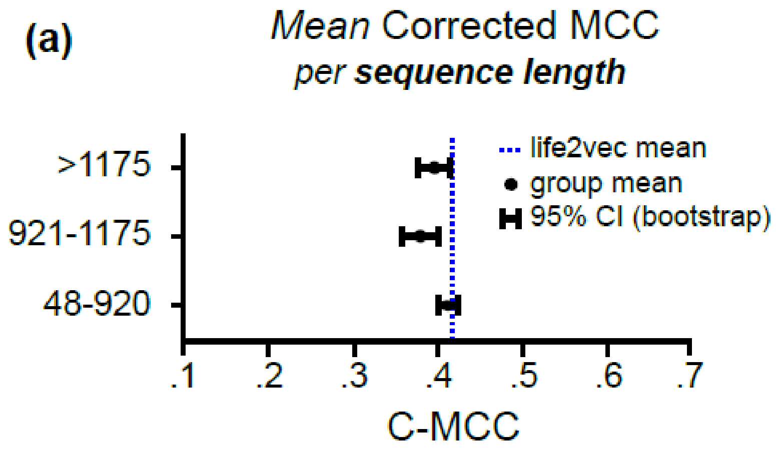
$$\text{Positive Predictions} = \theta = \frac{\text{tp}+\text{fp}}{\text{tp}+\text{fn}+\text{tn}+\text{fp}}$$

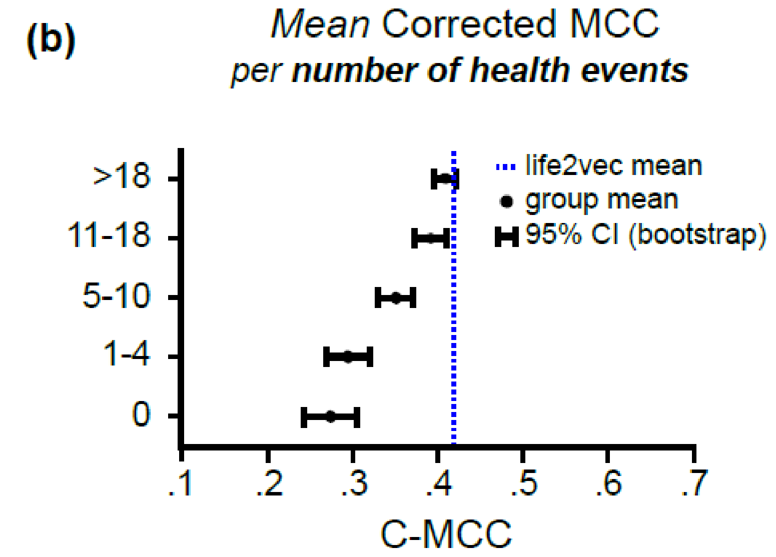
$$\hat{\gamma}_{cr} = (1-\hat{\alpha})^{-1}((1-\hat{\alpha})\cdot\hat{\gamma})$$

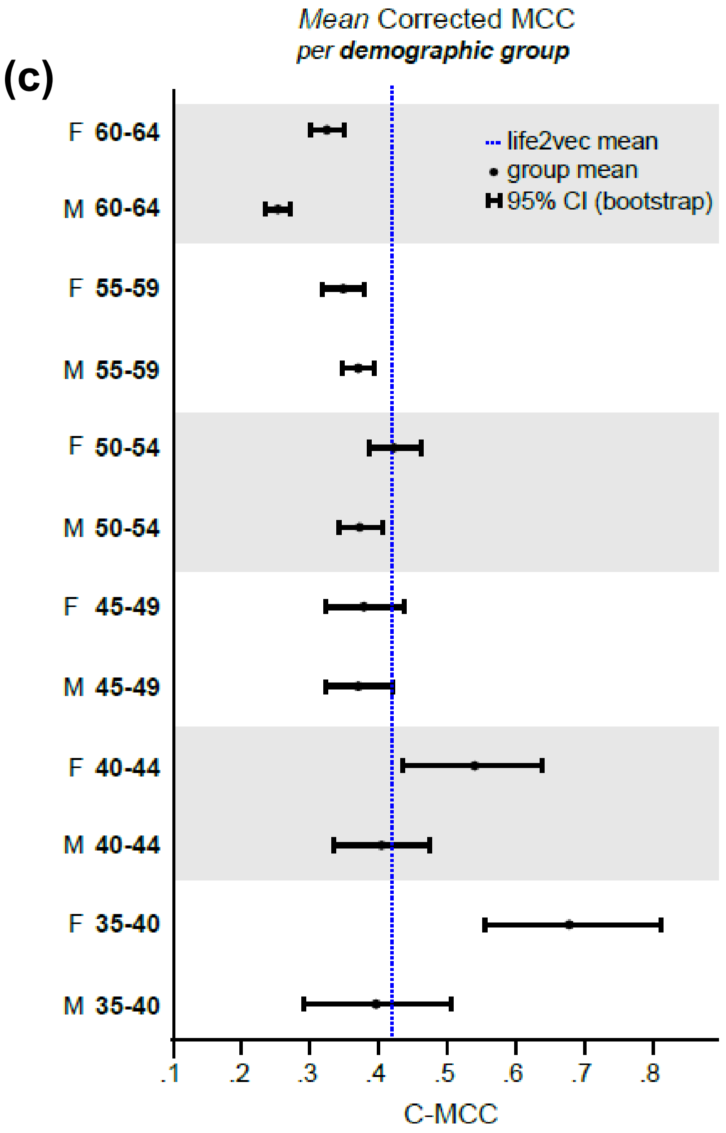
$$\hat{\eta}_{cr} = (1-\hat{\alpha})^{-1}(\hat{\eta} - \hat{\alpha}\cdot\hat{\gamma})$$

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

Early Mortality Prediction: Auditing







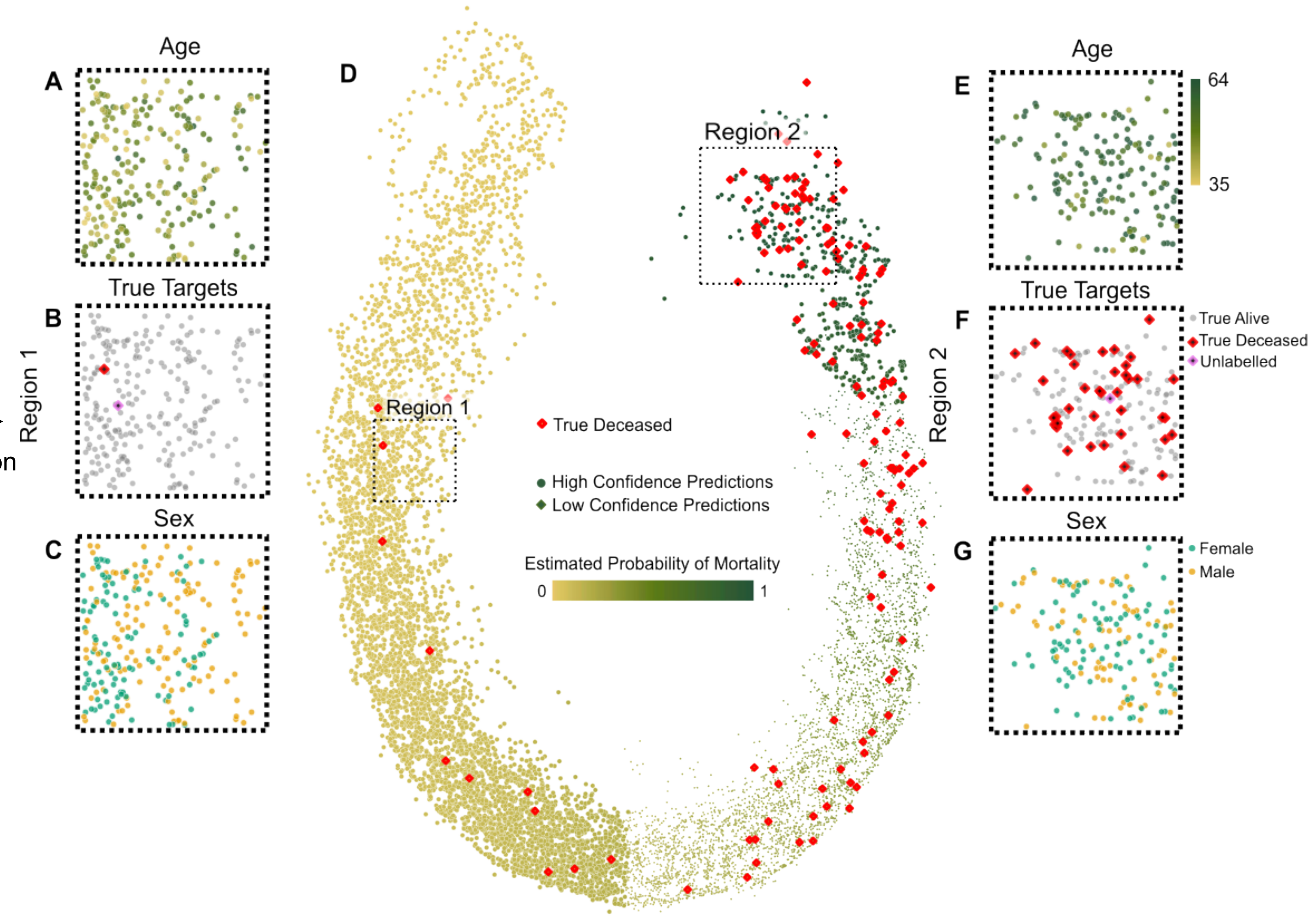
Early Mortality Prediction: Data Use

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

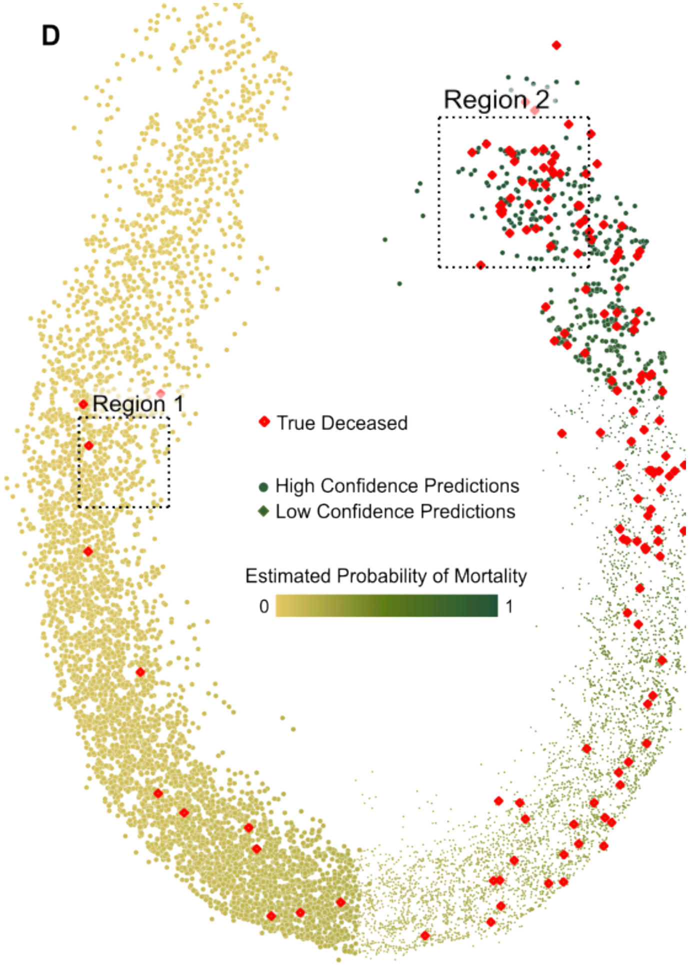
Partial Labor: no industry, sector, position and labour force

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



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):



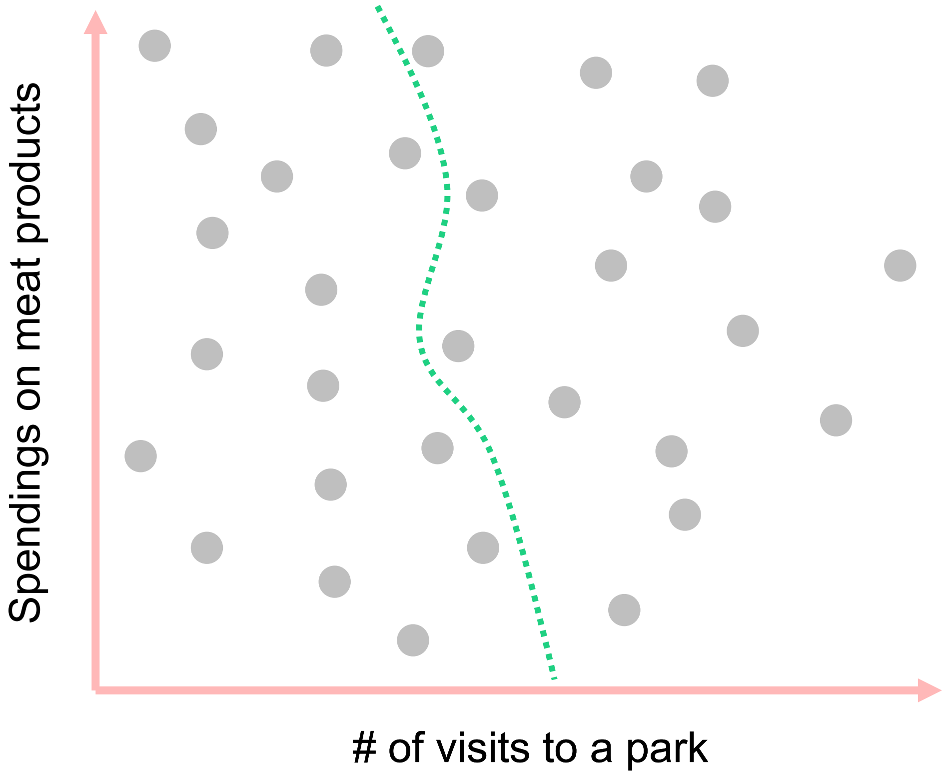
In the Concept space, we can find *somewhat* explainable directions!

- **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

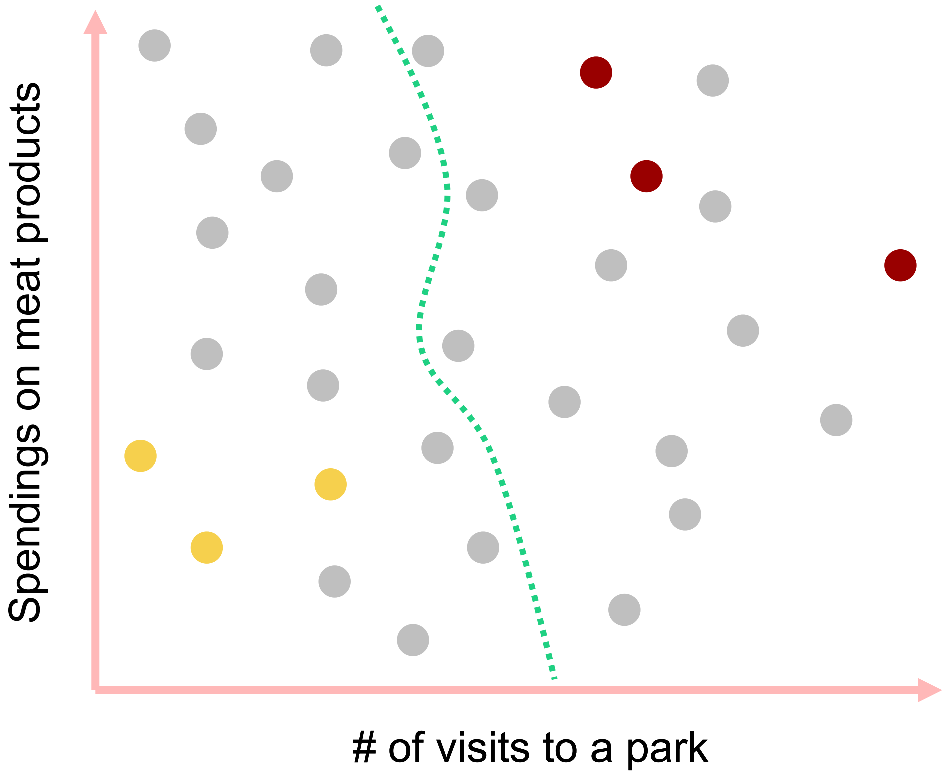
Overview of the TCAV method



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

decision boundary of the algorithm

Overview of the TCAV method



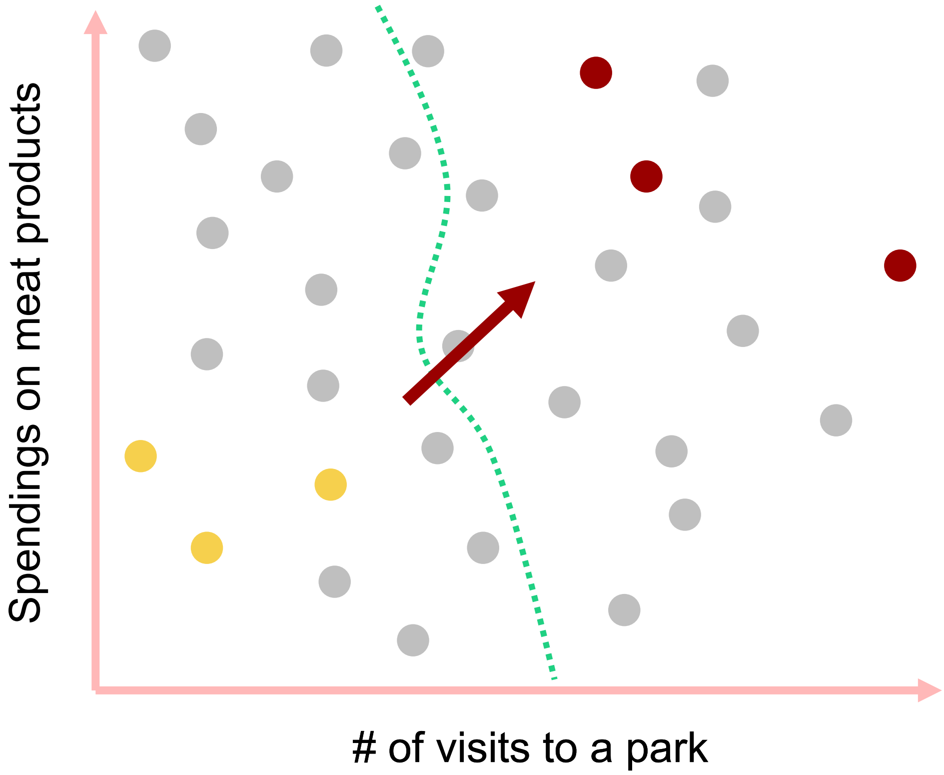
Let's imagine you have extra information

decision boundary of the algorithm

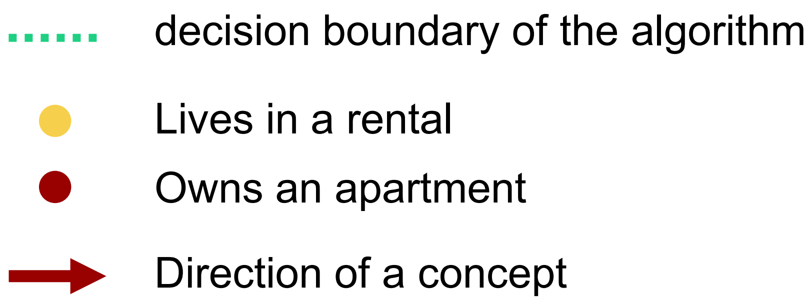
Lives in a rental

Owns an apartment

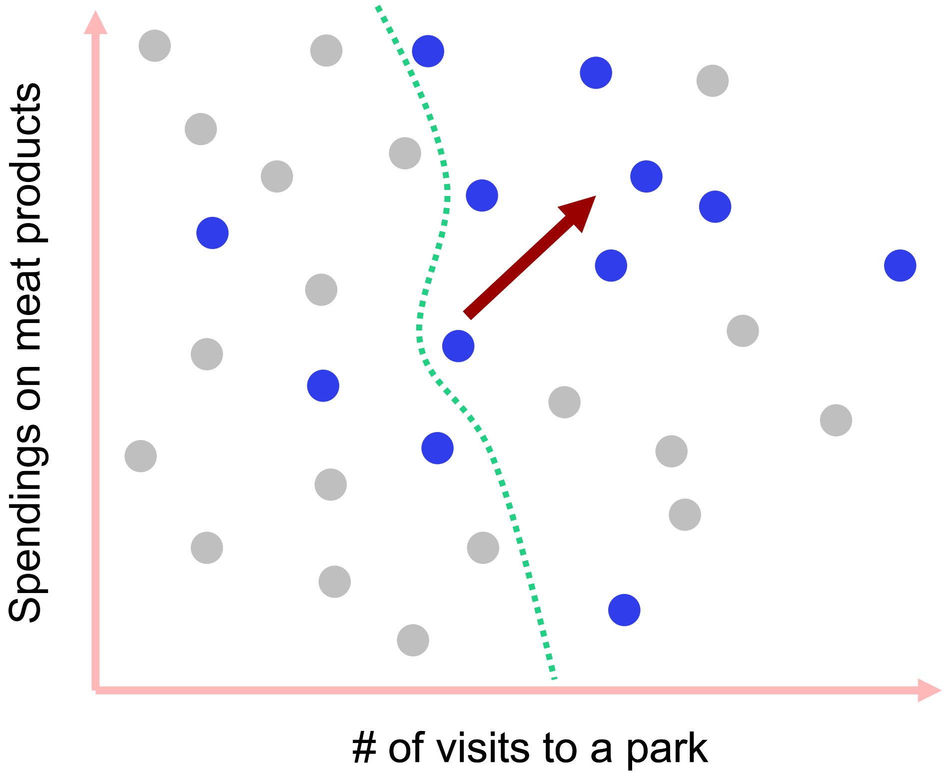
Overview of the TCAV method



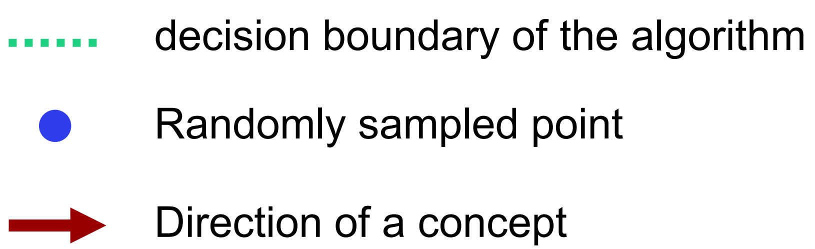
Let's imagine you have extra information



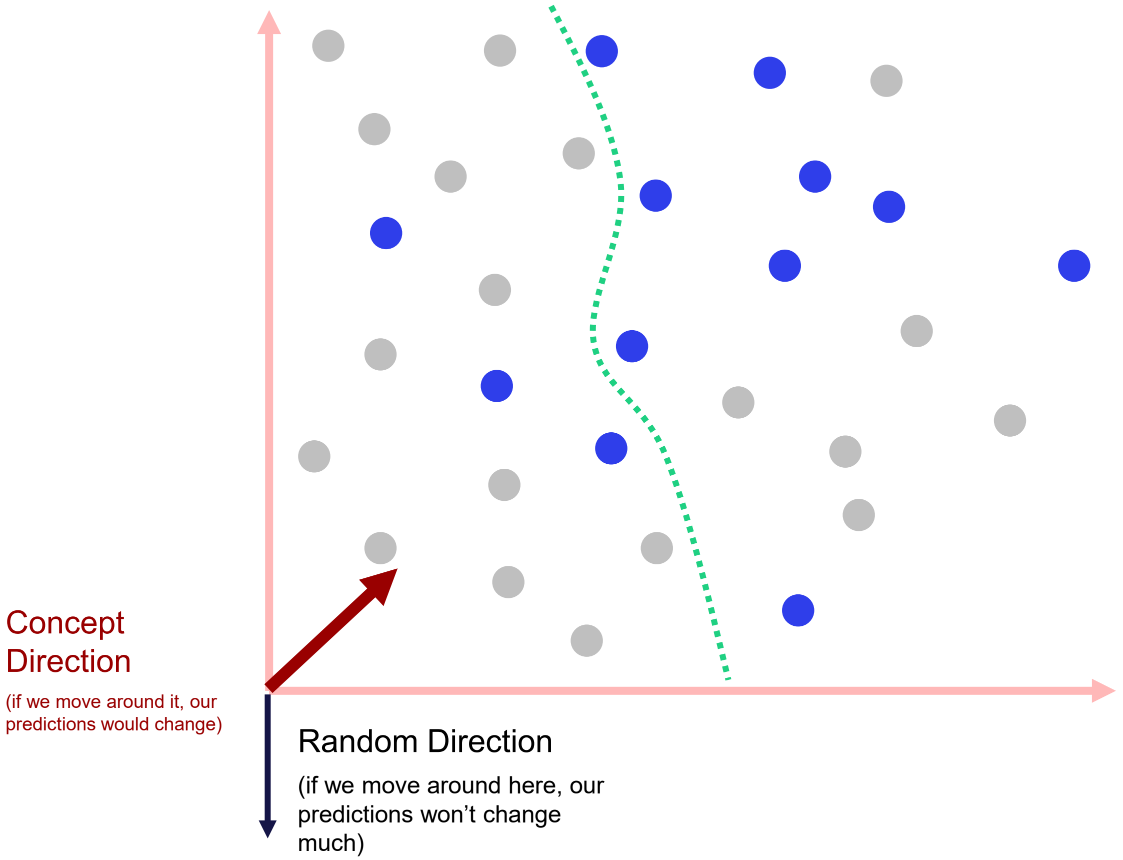
Overview of the TCAV method



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)

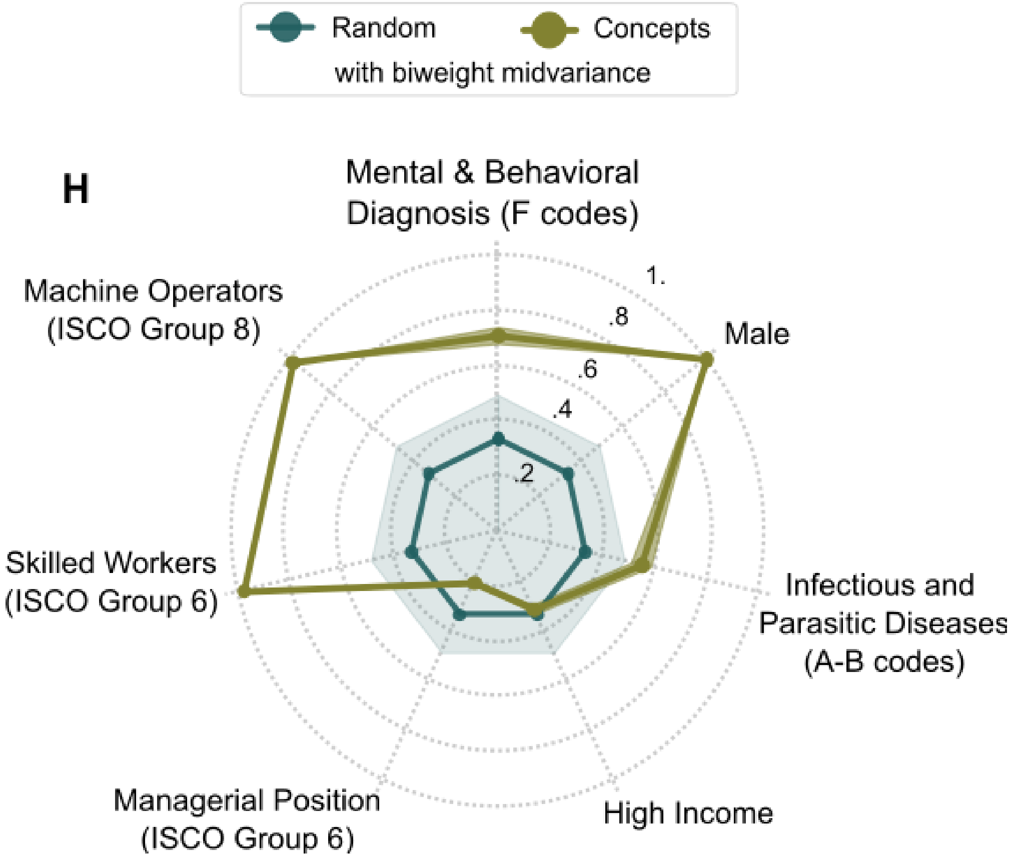


Overview of the TCAV method



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).

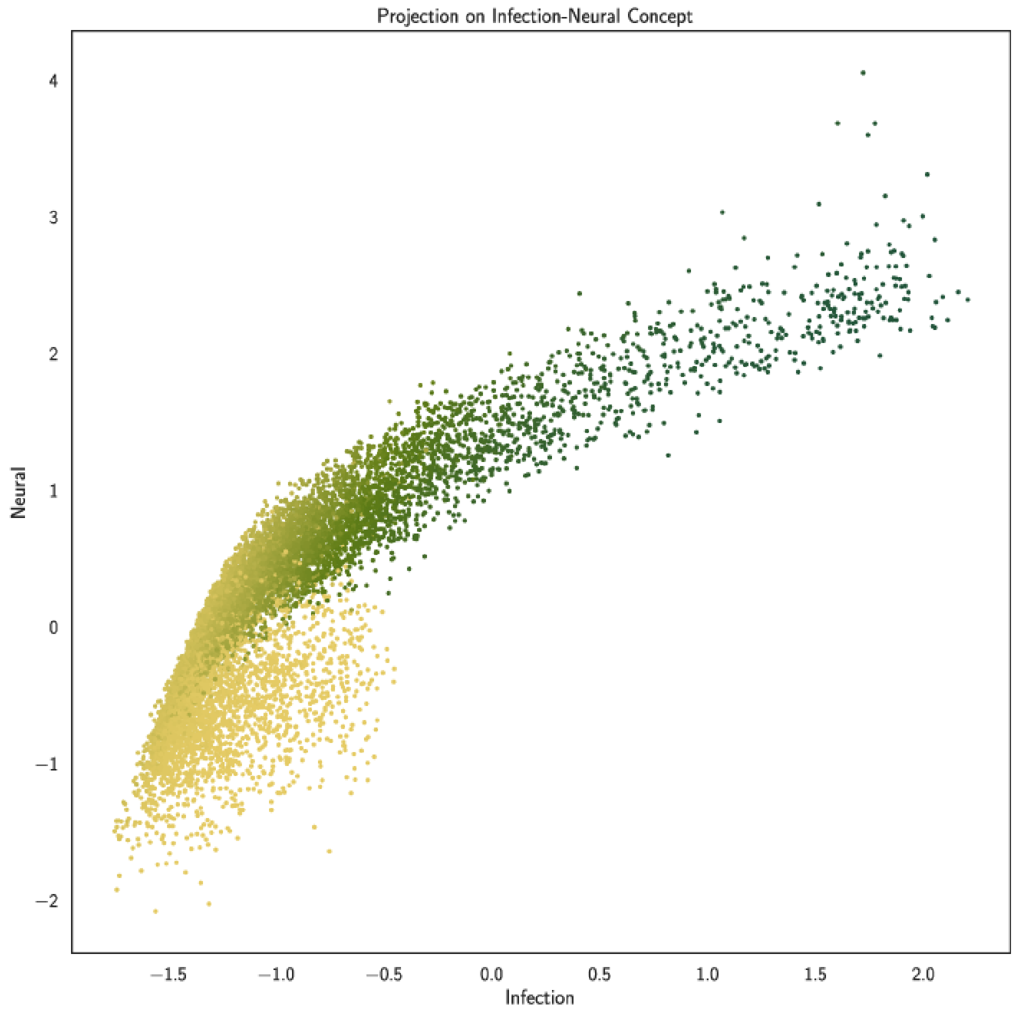
Explainability with TCAV (Mortality Prediction):

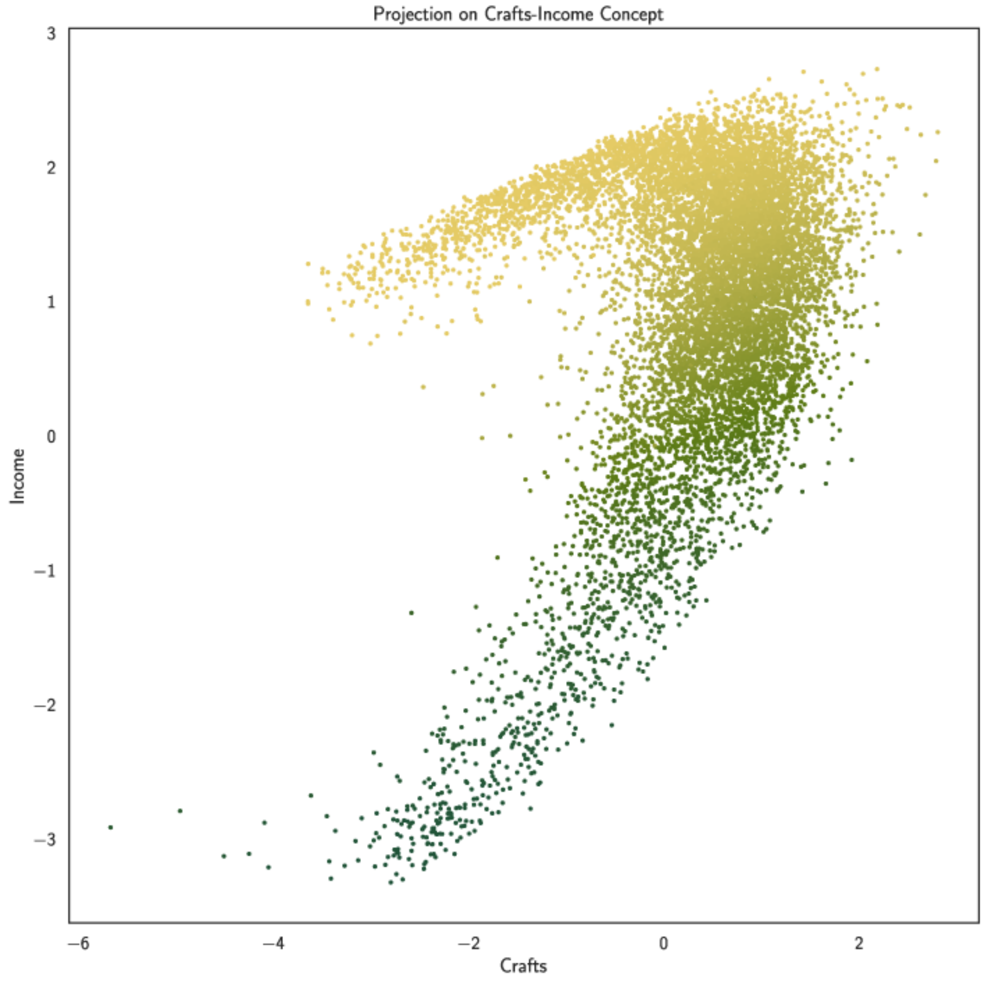


TCAV Score per "Direction"

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

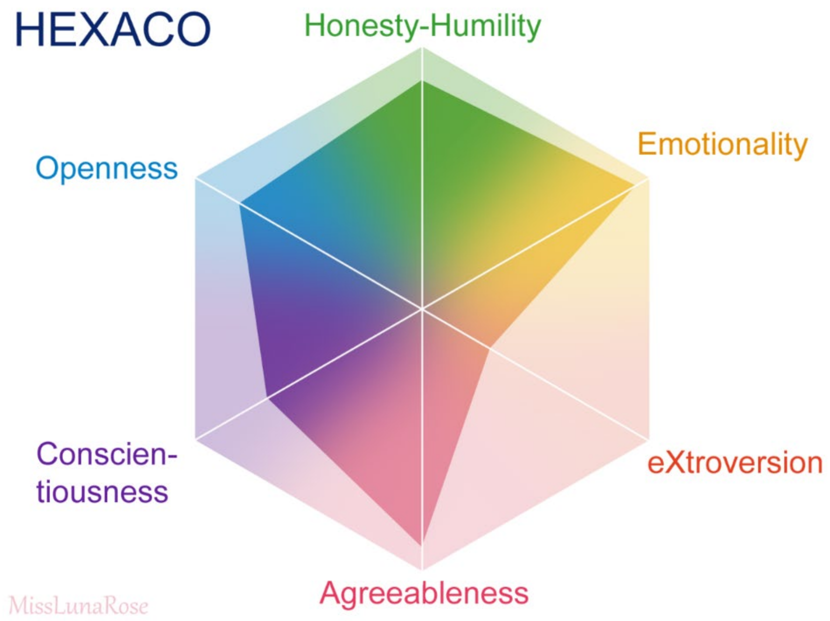
Projection to TCAV Directions





`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]

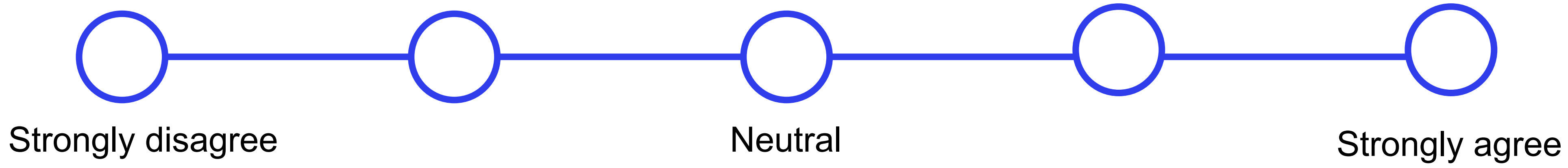
Inventory Descriptions: [The HEXACO Personality Inventory - Revised]

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



Personality Data

Quadratic Kappa Score

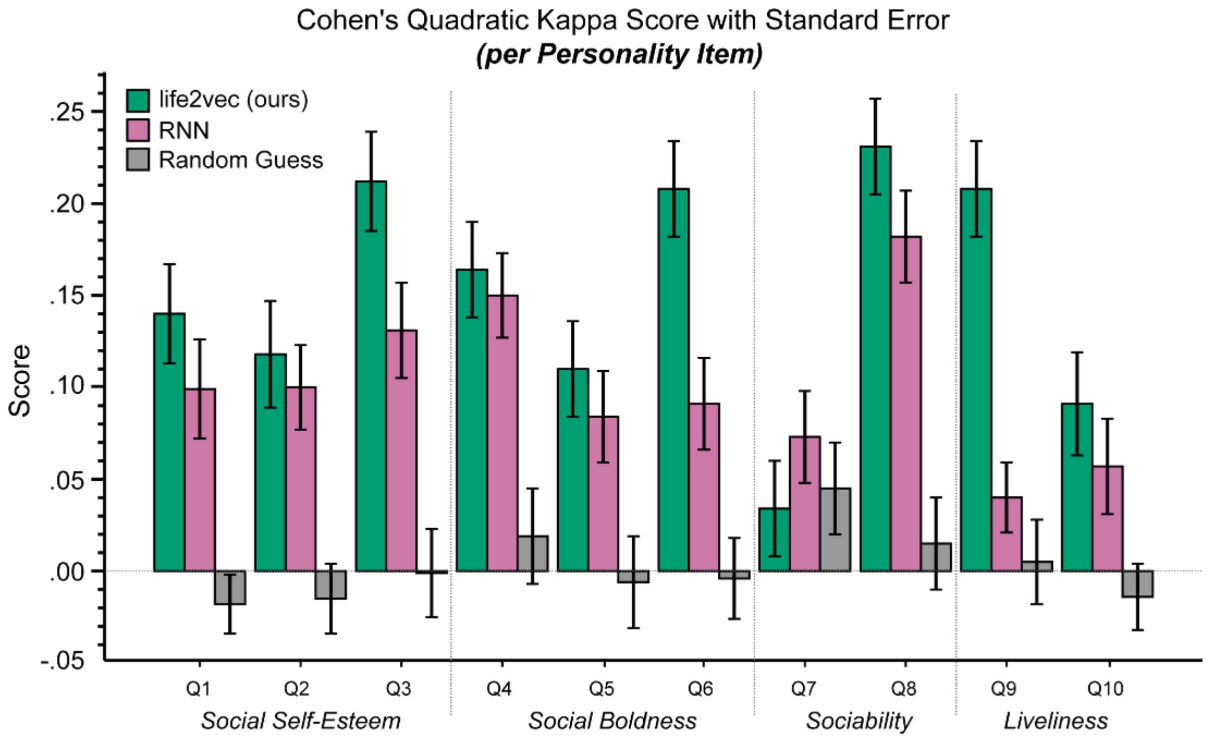
		Predicted				
		1	2	3	4	5
True	1	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}
	2	c_{21}	c_{22}	c_{23}	c_{24}	c_{25}
	3	c_{31}	c_{32}	c_{33}	c_{34}	c_{35}
	4	c_{41}	c_{42}	c_{43}	c_{44}	c_{45}
	5	c_{51}	c_{52}	c_{53}	c_{54}	c_{55}

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

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

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

Accounts for the distance
from predicted to target classes

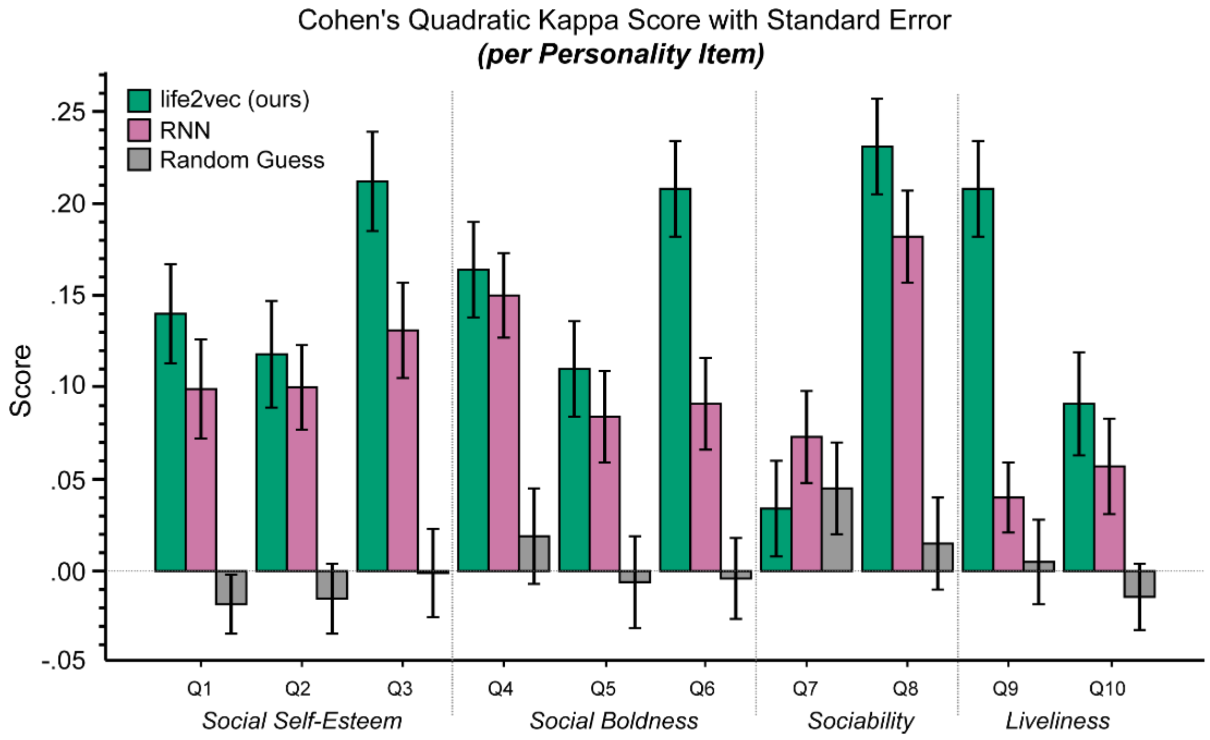


Personality Data

Questions:

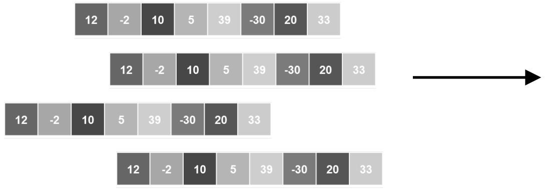
6. Most people are more upbeat and dynamic than I generally am (liveliness)

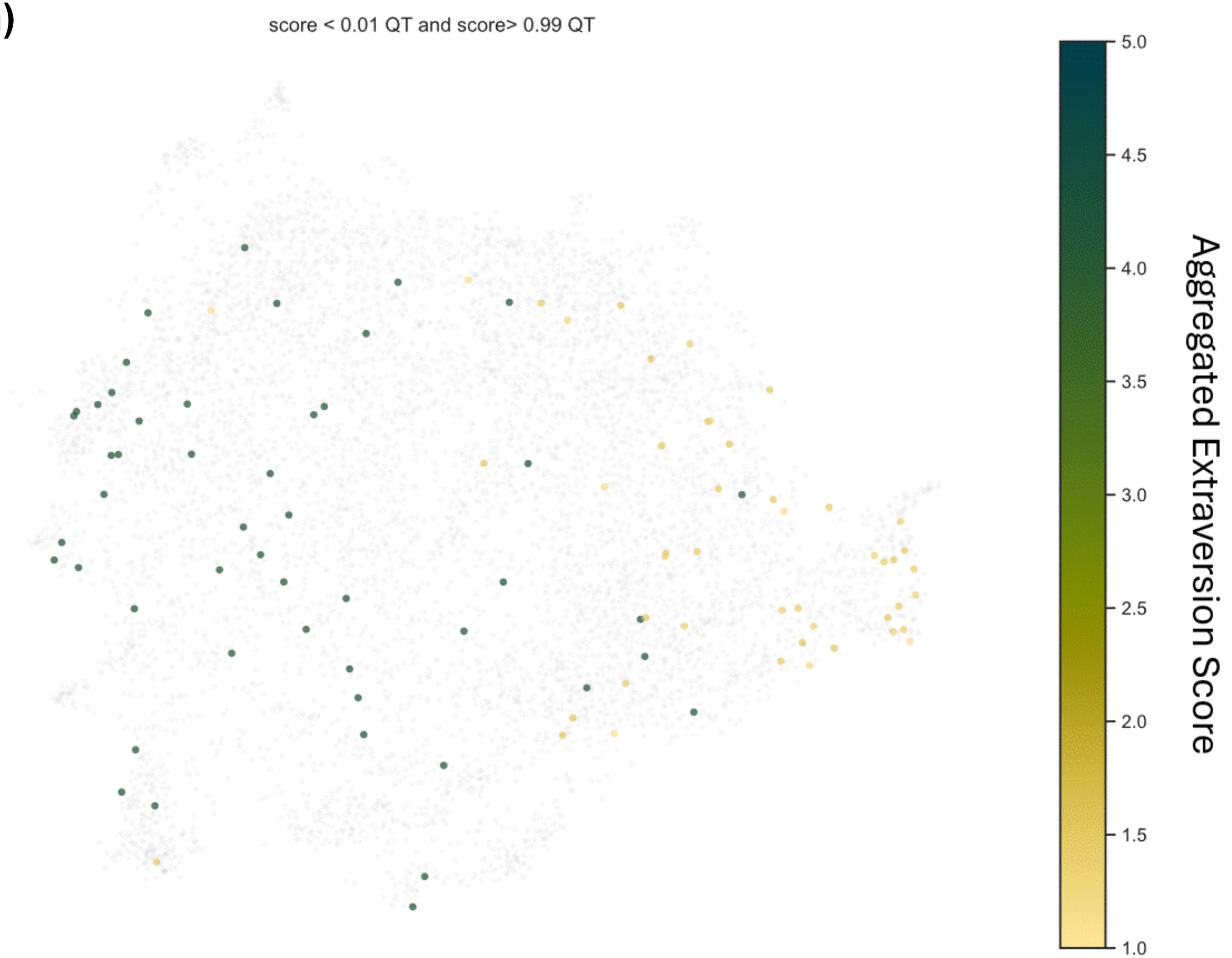
7. The first thing that I always do in a new place is to make friends (social I)



Personality Summaries
(PaCMAP projection)

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





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!

DTU

Exploring Embeddings

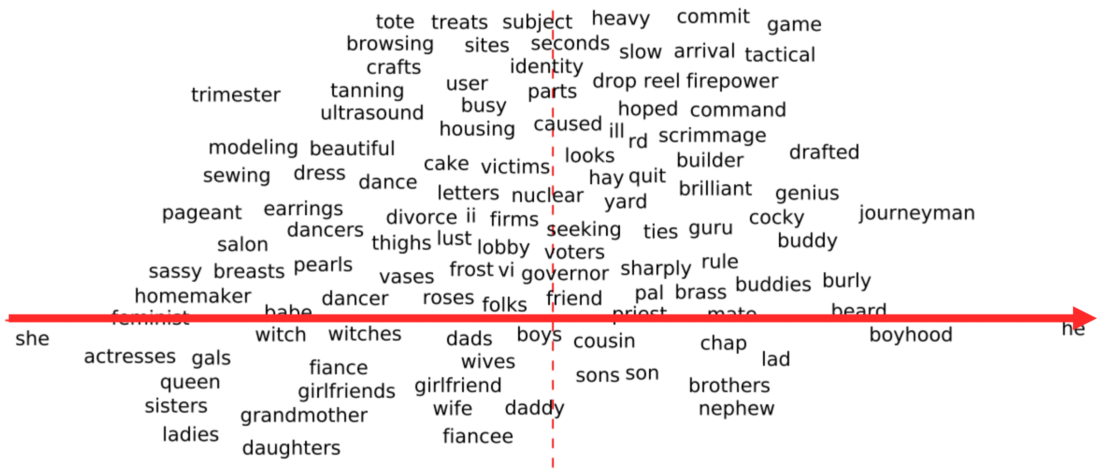
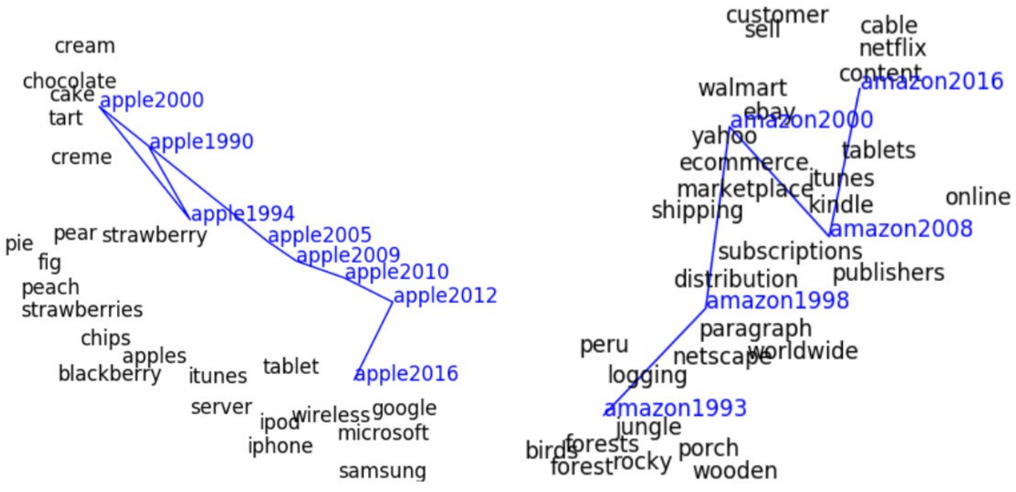


Fig.1: Words projected along the direction of "he"- "she" [1]

Study of gender bias in word2vec



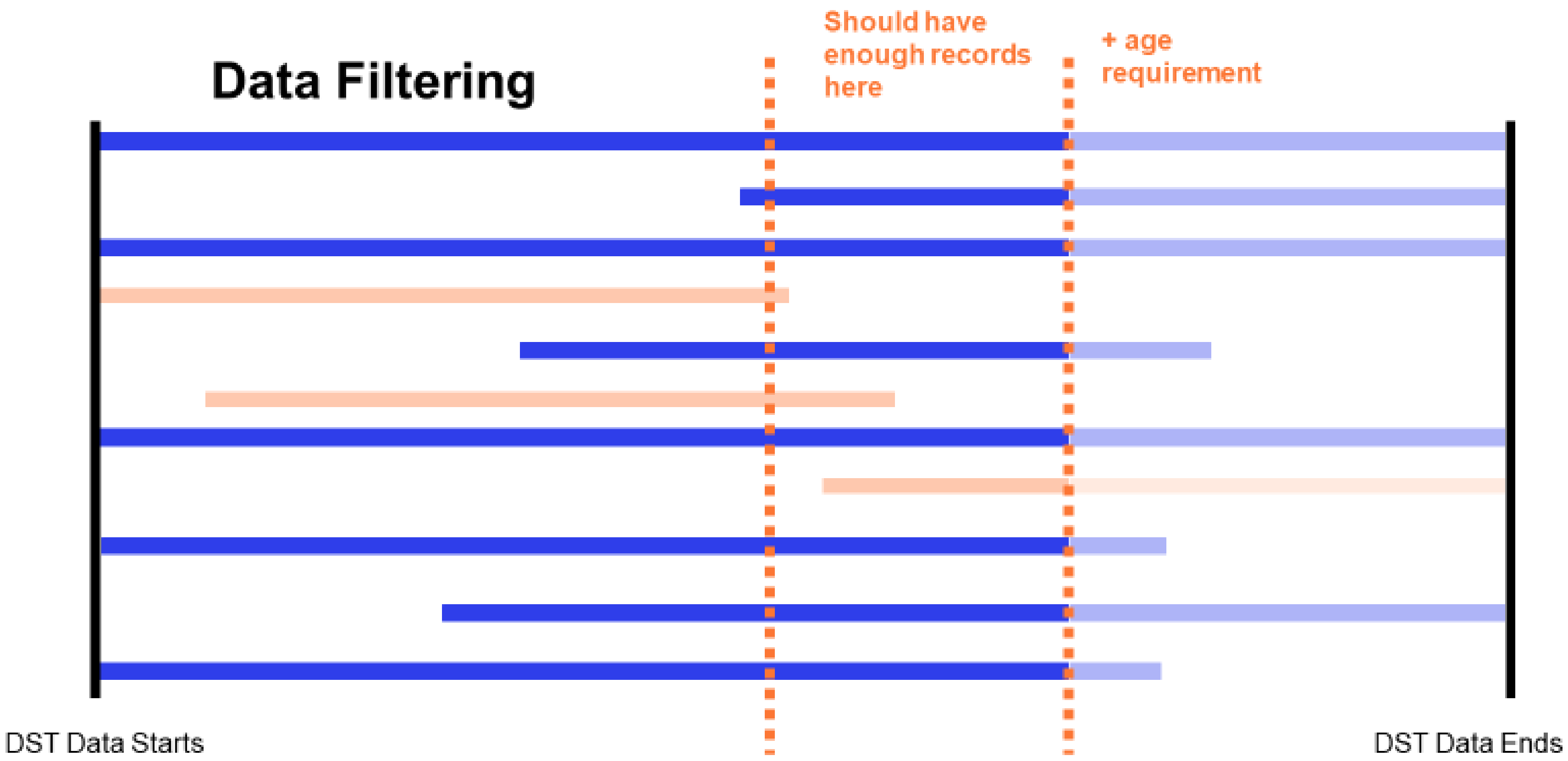
(a) apple (b) amazon

Fig.2: Trajectories of brand names[2]

Temporal evolution of terms with word2vec

1. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.
2. 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).

Data



Data Filtering
Should have enough records here
+ age requirement
DST Data Starts
DST Data Ends

Metric

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Balanced accuracy} = \frac{TPR + TNR}{2}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

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

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

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

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

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

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

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

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

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

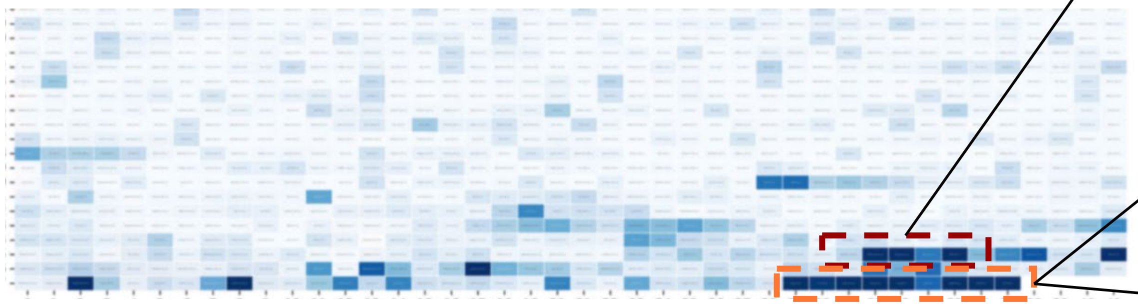
$$\widehat{\mathrm{mcc}}_{cr} = \sqrt{\frac{\hat{\pi}_{cr}(1 - \hat{\pi}_{cr})}{\theta(1 - \theta)}} (\hat{\gamma}_{cr} - \hat{\eta}_{cr})$$

True Labels (columns) / **Predicted Labels** (rows)

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Local Interpretability

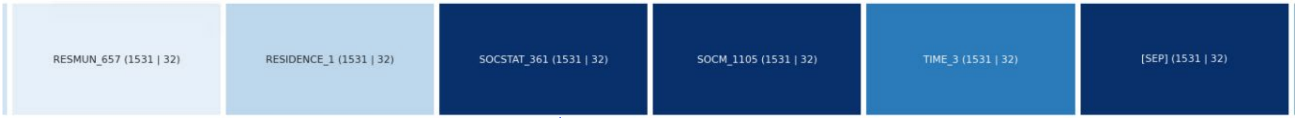
- **Interpretation of the scores**: how large is the change in the output likelihood if we slightly change the embedding of the token
- Only local explanation (e.g. per sequence), vague interpretation



Sequence of an individual in a textual format

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

If we zoom-in:



Maternity Leave

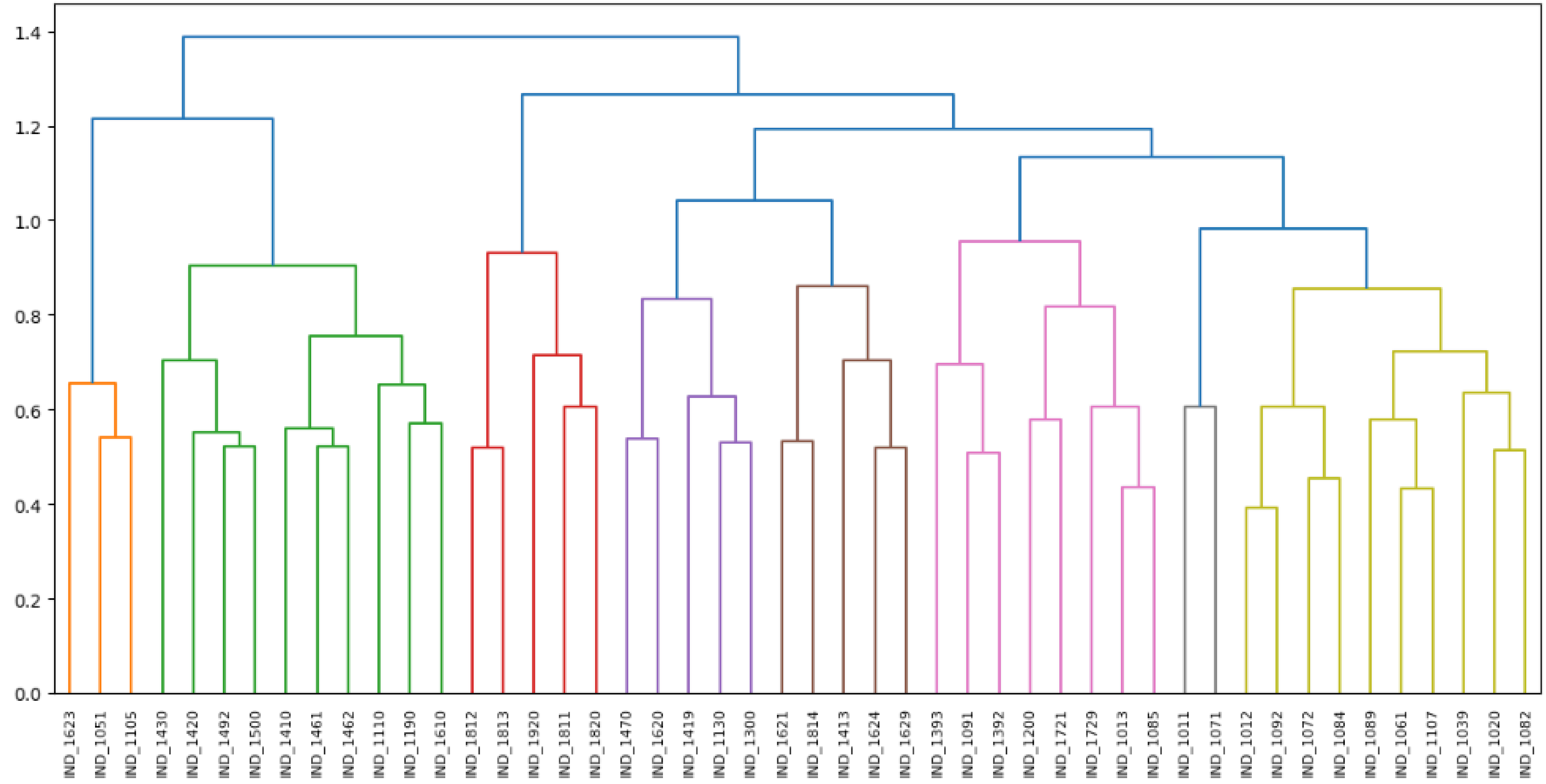
Flexjob, receives salary

Malignant neoplasm of brain (admitted to hospital)

Pyothorax with fistula (admitted to hospital)

Concept Space Robustness: Pairwise Distances

Concept Space Robustness: Tree Structures



Concept Space Robustness: Other Methods

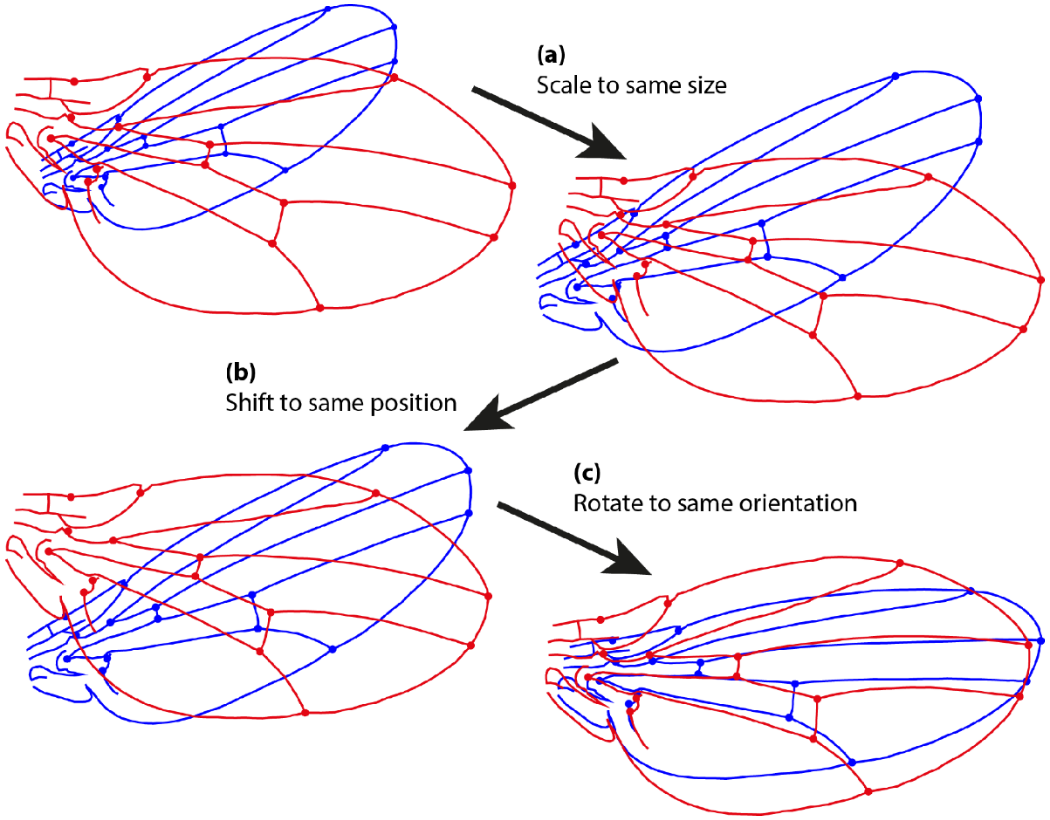


Fig 1: Pipeline behind Procrustes Analysis [1]

```
array([[0.        , 0.65814077, 0.64156468, 0.63280614, 0.65323986],
       [0.        , 0.        , 0.6460441 , 0.64274334, 0.68746551],
       [0.        , 0.        , 0.        , 0.63947709, 0.6323634 ],
       [0.        , 0.        , 0.        , 0.        , 0.65236674],
       [0.        , 0.        , 0.        , 0.        , 0.        ]])
```

```
procrustes(e_add[-1], permuted)[-1]
```

```
0.9422304059675406
```

Fig 1: Procsrustes Analysis on the Concept Spaces (SSE)

Procrustes analysis. (2023, July 1). In Wikipedia. https://en.wikipedia.org/wiki/Procrustes_analysis

Performer: Self-Attention for Long Sequences

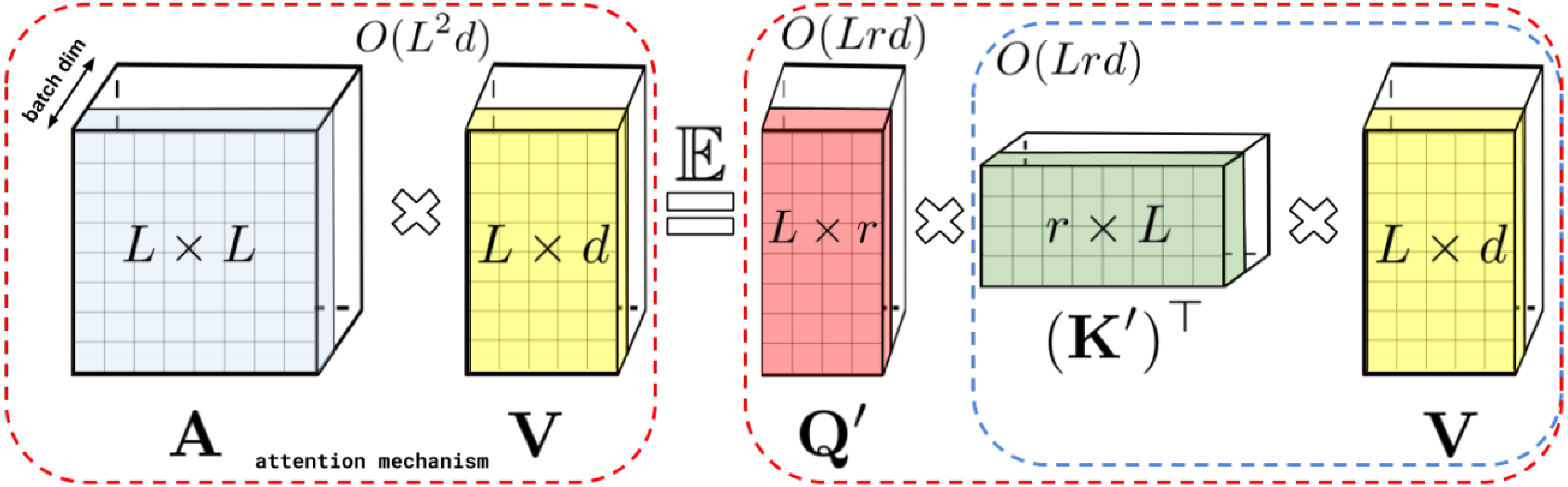


Figure 1: Approximation of the regular attention mechanism **AV** (before $\mathbf{D}^{-1}$-renormalization) via (random) feature maps. Dashed-blocks indicate order of computation with corresponding time complexities attached.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L. and Belanger, D., 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Time Encoding: Time2vec

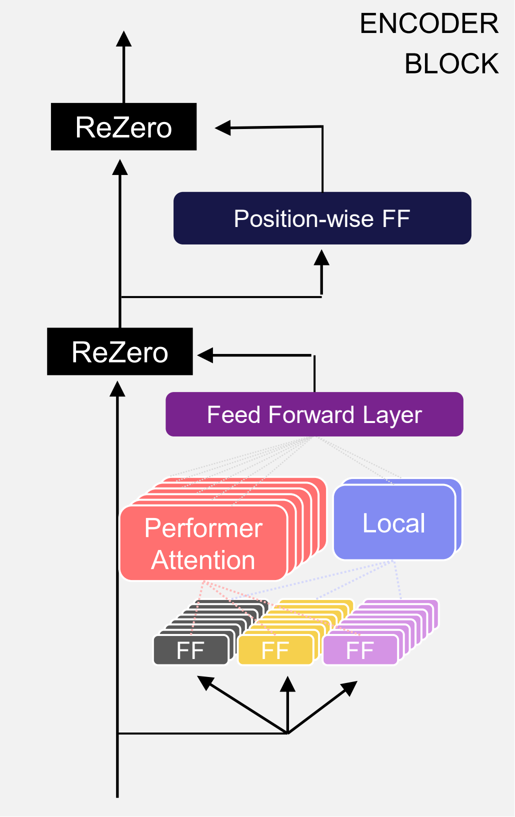
- We transform tokens/concepts into embeddings, but **what happens with AGE and ABSPOS?** We use time2vec embeddings:
- Two learnable parameters: ω and φ
- F is COS function (for age) and COS function (for abspos)
- "i" specifies the dimension of an embedding (k – number of dimensions)

Linear Component

$$\mathbf{t2v}(\tau)[i] = \begin{cases} \omega_i \tau + \varphi_i, & \text{if } i = 0. \\ \mathcal{F}(\omega_i \tau + \varphi_i), & \text{if } 1 \leq i \leq k. \end{cases}$$

Periodic component

Model Architecture



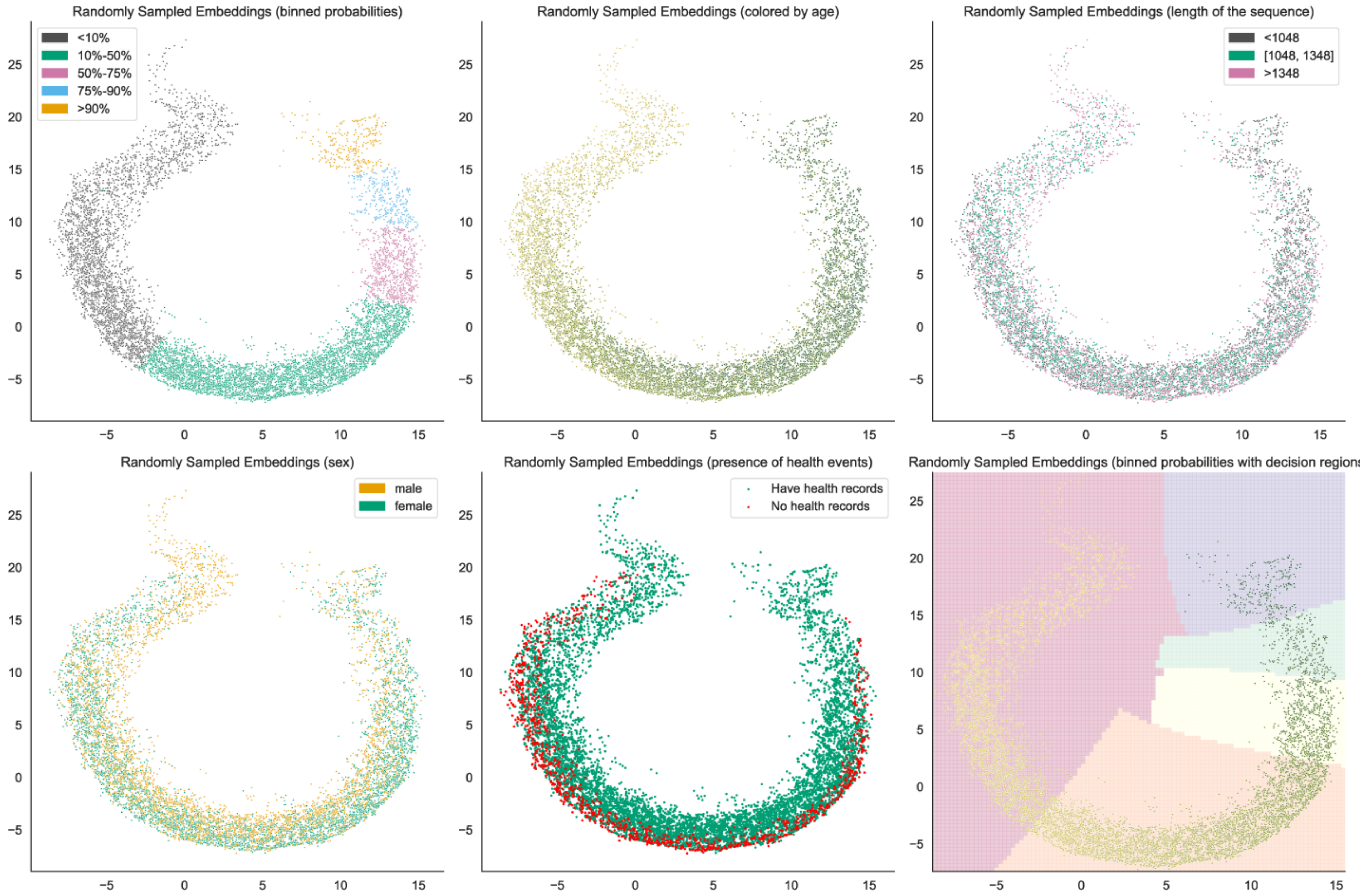
Details:

- Swish activations [2]
- Joint I-O Embedding [3]
- **Performer** Attention Unit [4]
- **ReZero** - Residual Connection [5]

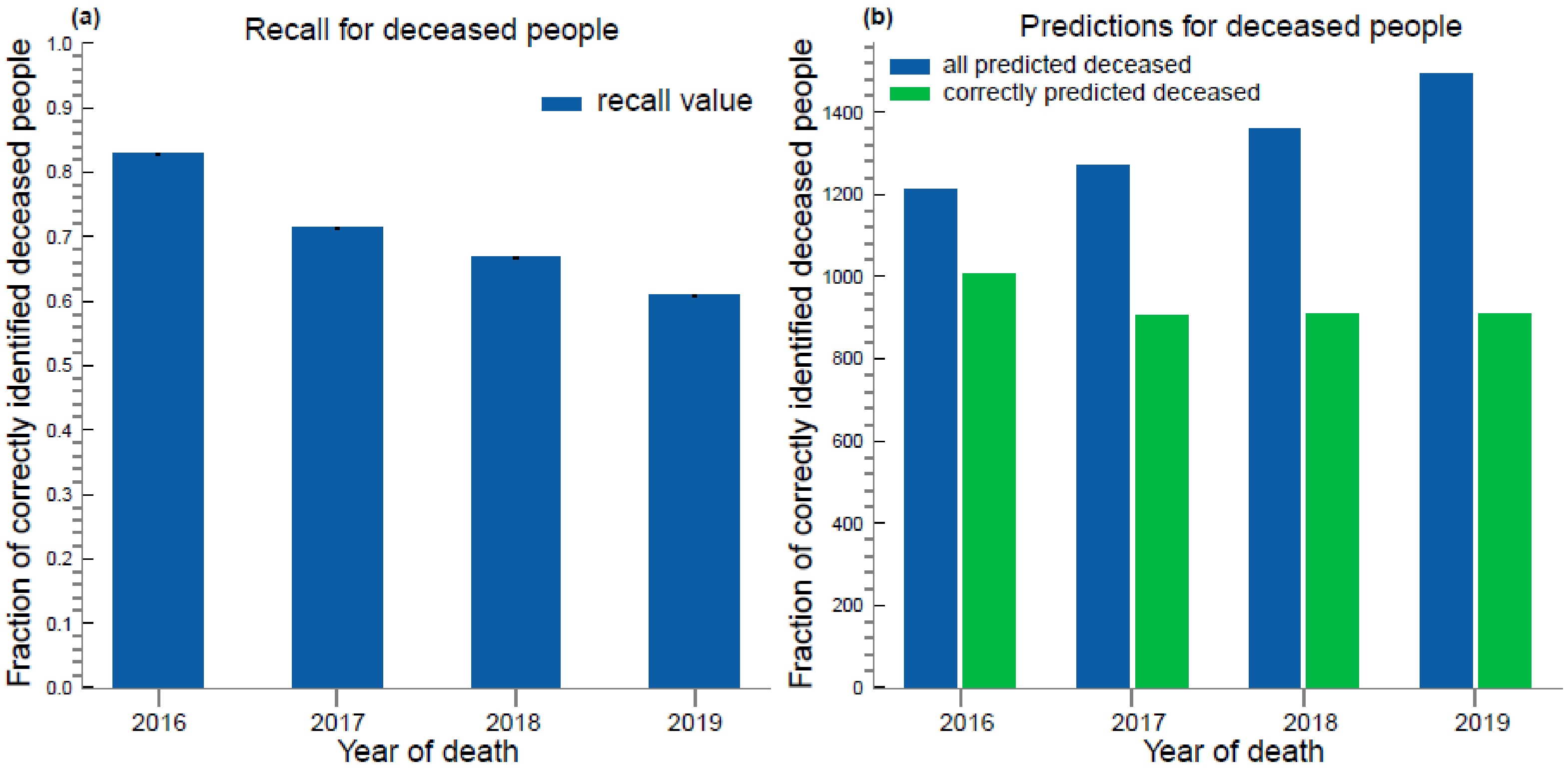
References:

1. Eger, S., Youssef, P. and Gurevych, I., 2019. Is it time to swish? Comparing deep learning activation functions across NLP tasks. arXiv preprint arXiv:1901.02671.
2. Nikolaos Pappas, Lesly Miculicich Werlen, and James Henderson. Beyond weight tying: Learning joint input-output embeddings for neural machine translation. arXivpreprint arXiv:1808.10681, 2018
3. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L. and Belanger, D., 2020. Rethinking attention with performers. arXiv preprint arXiv:2009.14794.
4. Parisotto, E., Song, F., Rae, J., Pascanu, R., Gulcehre, C., Jayakumar, S., Jaderberg, M., Kaufman, R.L., Clark, A., Noury, S. and Botvinick, M., 2020, November. Stabilizing transformers for reinforcement learning. In International Conference on Machine Learning (pp. 7487-7498). PMLR.
5. Bachlechner, T., Majumder, B.P., Mao, H., Cottrell, G. and McAuley, J., 2021, December. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence* (pp. 1352-1361). PMLR.

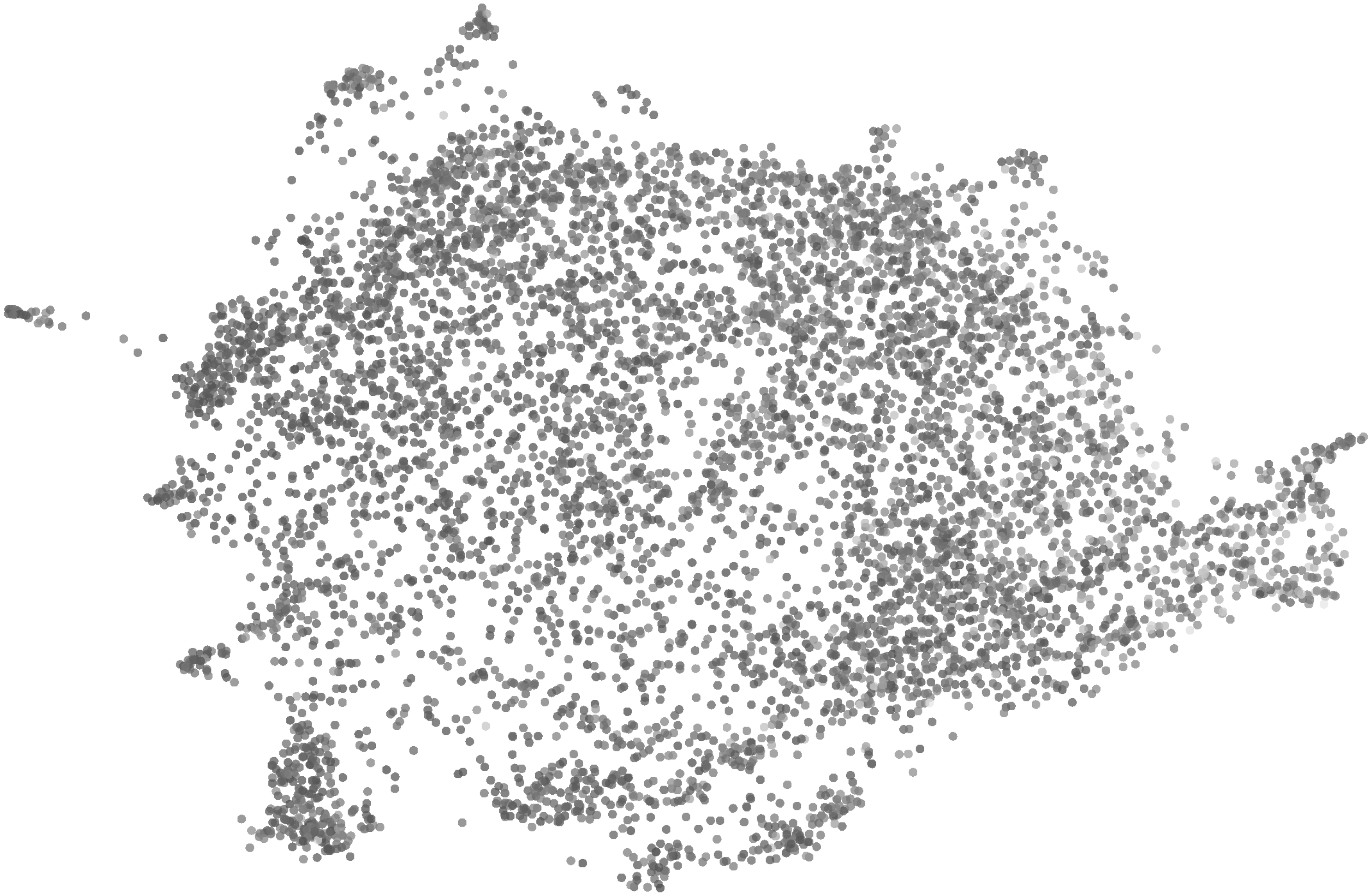
Embeding = Token + a * Age + b * Abs + c * Segment


Randomly Sampled Embeddings (binned probabilities)
<10%
10%-50%
50%-75%
75%-90%
>90%
Randomly Sampled Embeddings (colored by age)
Randomly Sampled Embeddings (length of the sequence)
<1048
[1048, 1348]
>1348
Randomly Sampled Embeddings (sex)
male
female
Randomly Sampled Embeddings (presence of health events)
Have health records
No health records
Randomly Sampled Embeddings (binned probabilities with decision regions

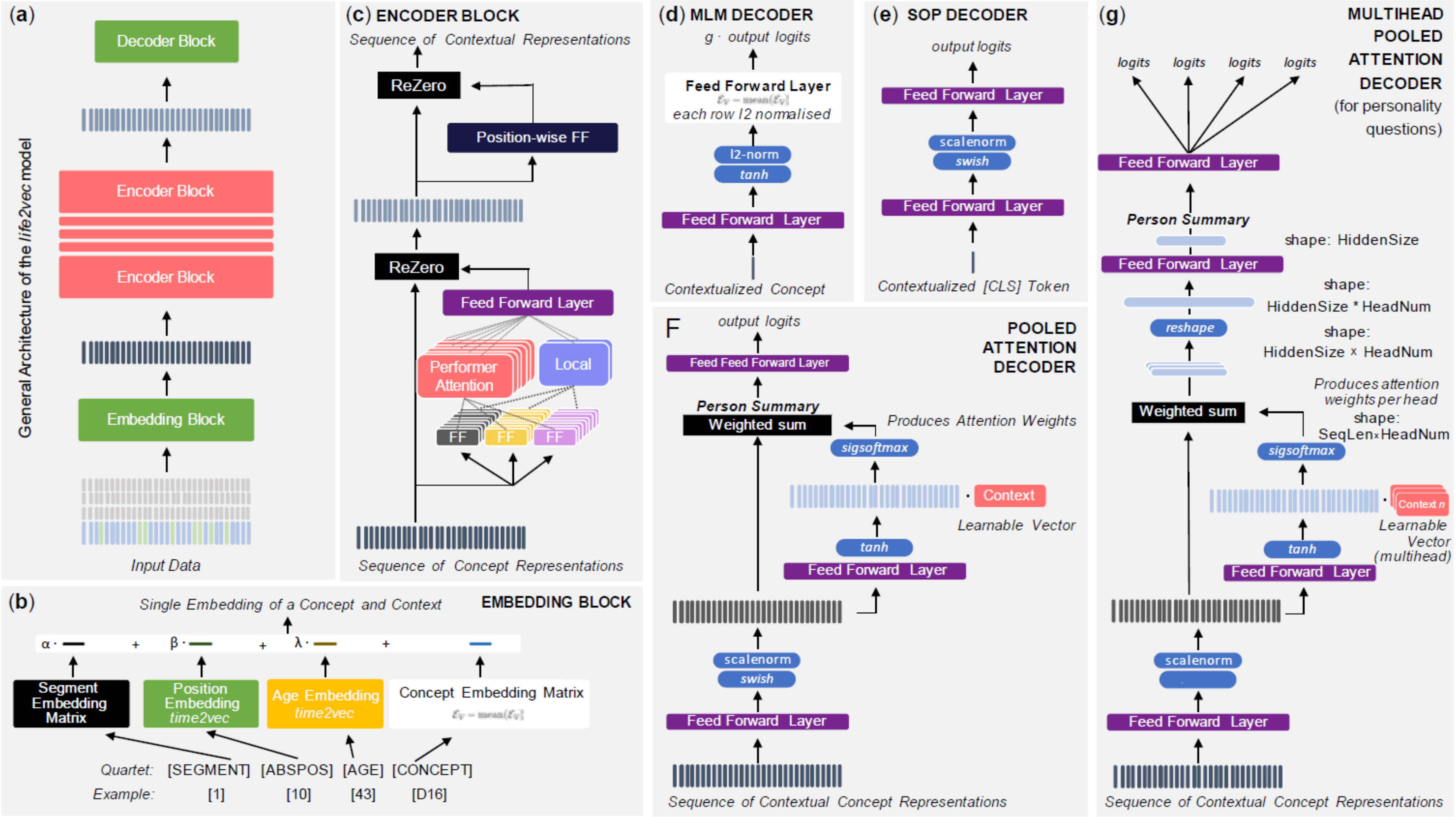
Early Mortality Prediction: Time-to-event



Person-Summary Space (based on Extraversion task)



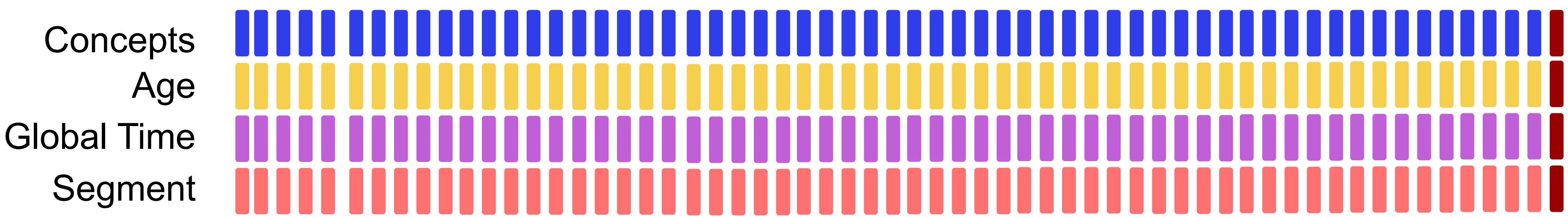
Architecture Overview



Placeholder Tokens `[PLCH]`

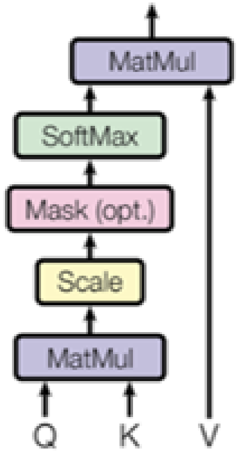
You would put it here as **a substitute for a question or a prompt**

`[PLCH1]` – might signal that the model needs to predict the response to Q1



Input to the `life2vec` model

Self-Attention I



How does it calculate the contextual-representation? With **Self-Attention**!

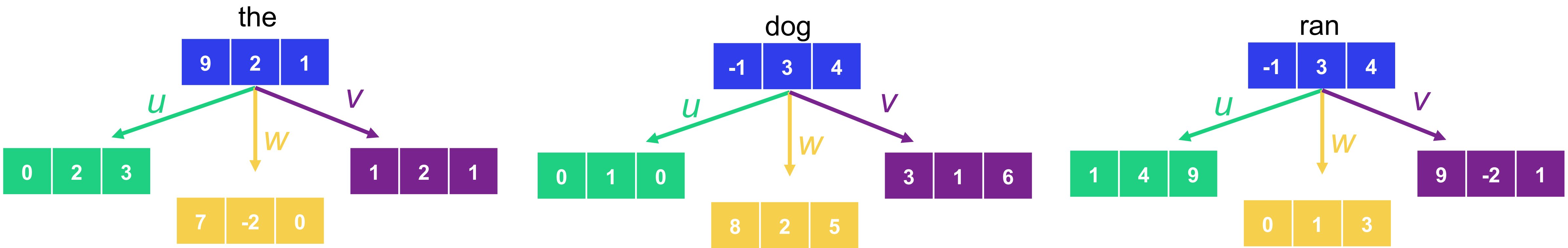
Let's assume we have sentence: **The dog run.**

Workflow (Simplified):

1. **Lookup embeddings** for each word (or take the ones from a previous block)

the			dog			ran		
9	2	1	-1	3	4	-1	3	4

2. **Transform** each embedding into Key, Query and Value (those are just names for transformed versions of embeddings)



u, w, v – Feed forward layer

Self-Attention II

3. Calculate **attention scores** for each word (dot product):

	query				key					score	softmax normalised
the	7	-2	0	•	0	2	3	*the*	=	-4	.05
the	7	-2	0	•	0	1	0	*dog*	=	-2	.25
the	7	-2	0	•	1	4	9	*ran*	=	-1	.70

4. Calculate **contextualized embedding:**

*This is a **contextualized** embedding for “**the**”*

$$\begin{bmatrix} 7 & 1 & 2 \end{bmatrix} = .05 \cdot \underset{\textit{the}}{\overset{\text{value}}{\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}}} + .25 \cdot \underset{\textit{dog}}{\overset{\text{value}}{\begin{bmatrix} 3 & 1 & 6 \end{bmatrix}}} + .70 \cdot \underset{\textit{ran}}{\overset{\text{value}}{\begin{bmatrix} 9 & -2 & 1 \end{bmatrix}}}$$

5. Do for each word
6. Pass embeddings to a next block and repeat (now with the contextualized)