Is *life2vec* a doom calculator?

Germans Savcisens

Copenhagen Data Beers

March 19th, 2024

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

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

Main contributions of the research:

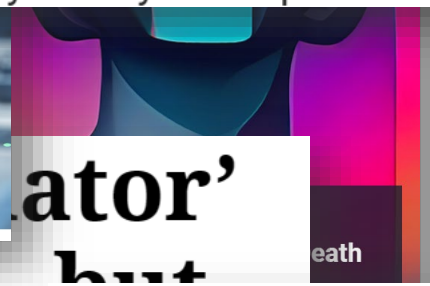
1. **Propose a framework** (*transformer-based*) to analyze large-scale socioeconomic and health data
2. Demonstrate the **power of dense representation**
3. **Adapt explainability methods** to understand predictions

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

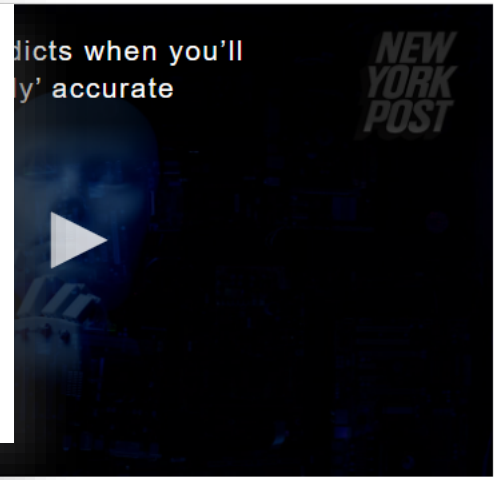
Published Dec. 20, 2023
Updated Dec. 20, 2023, 4:11 p.m. ET

N ator'
AI predicts death, but
do we really want
kr

Den
your

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.

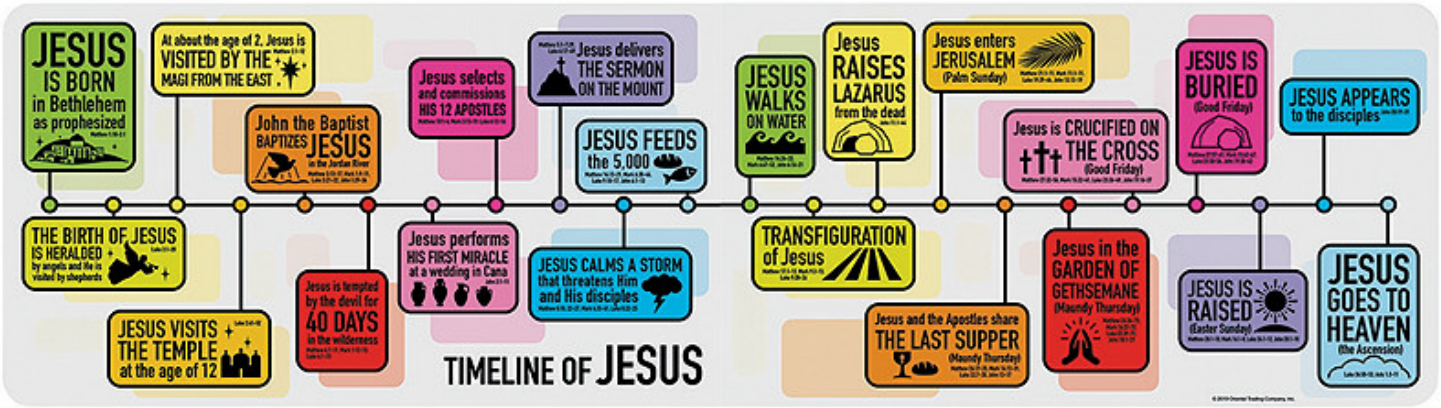
23 Dec 2023

Can it accurately predict the time of death?

Short answer: Not really!

So...what is the paper about?

Life Trajectories

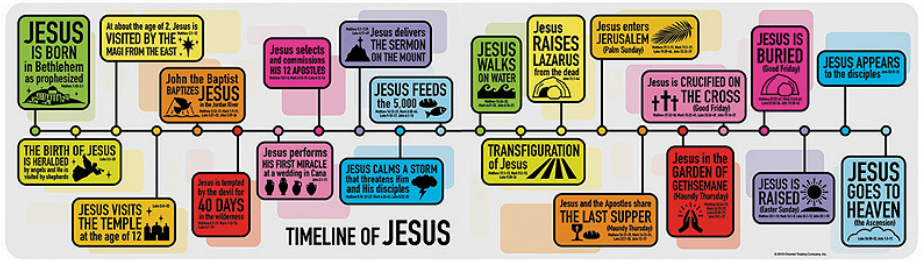


Issues associated with **longitudinal data**:

- 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

Image Source: Oriental Trading

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

Probability of readmission to a hospital?

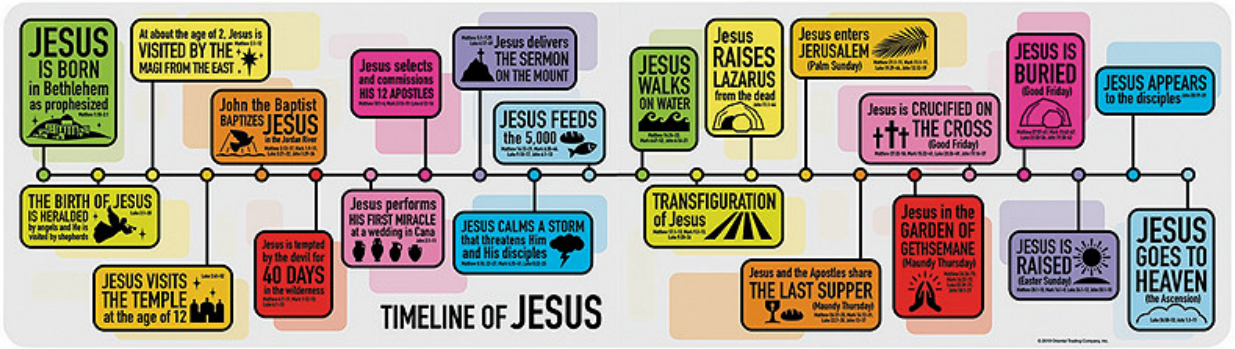
Model 2

Income level within the next year?

•••

Model N

* simplified



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)

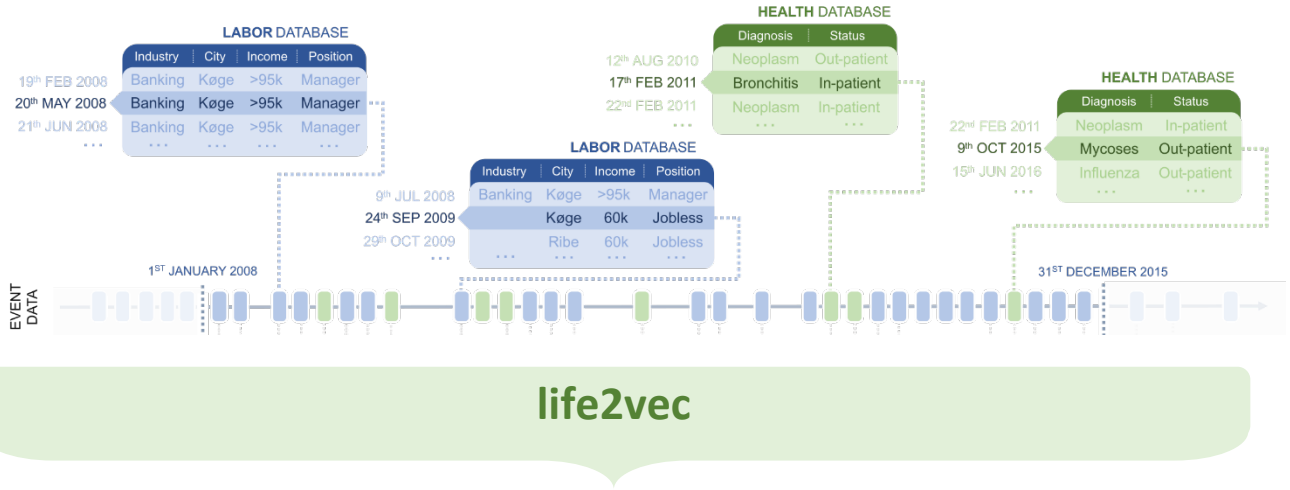
Give comprehensive insight into the data

Study sociological phenomena
(on a *global* scale)

Our Work

We are **not there yet,**

…but we have done the **first steps**



Main Components:

Text like encoding of data

Large Language Model

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...

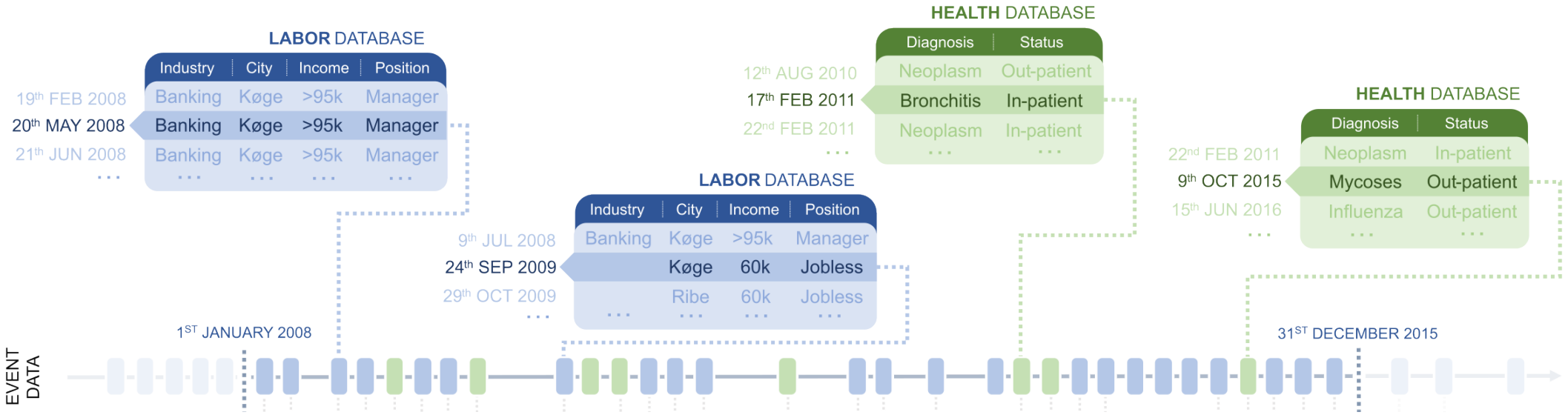
Culture and leisure
Film, media, museums, music, digital behaviour, sports...

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

**AI-Generated Image

Power of National Registry

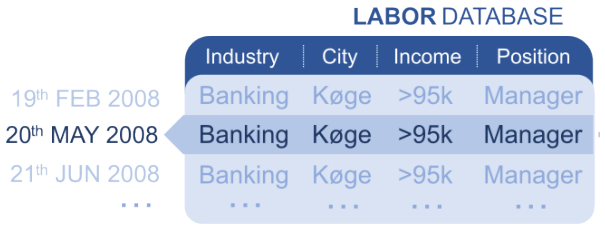
The National Registry is a source of **fine-grained information** about **the progression of one life**.
Unique possibility to study life progression and life outcomes.



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

still…how do we model it?

Forming a Language

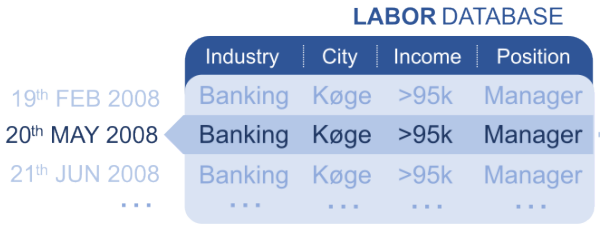


Convey the content in a spoken language

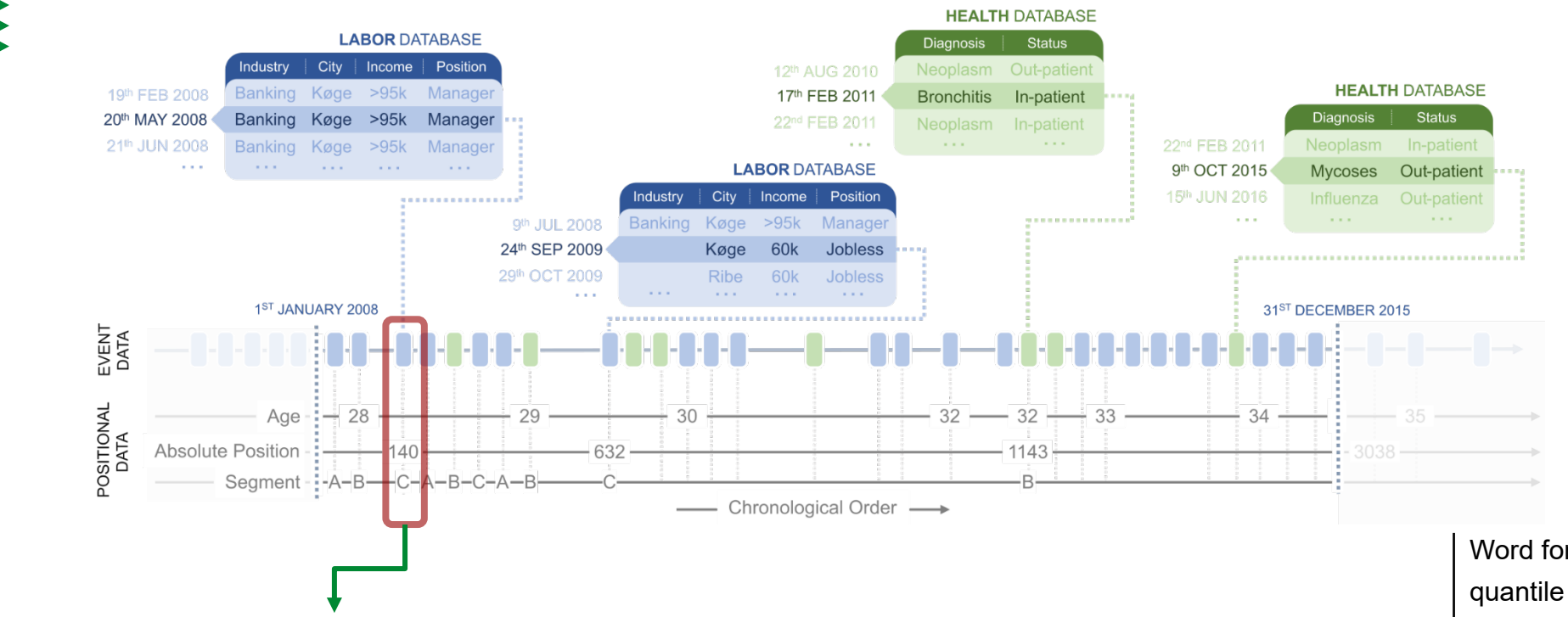
"In May 2008, Riley received >95k as a manager in Bank."

Language allows for super flexible and nuanced communication

Modelling Tabular Records as Language



Convey the content in an artificial symbolic 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

1. Describe the content of the record

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

Word for the quantile of the *income*

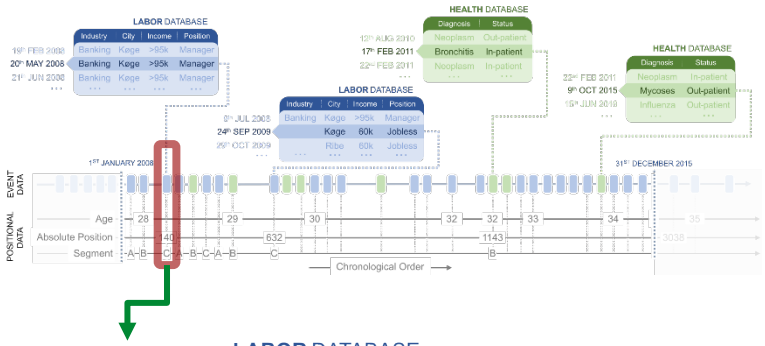
A "dot"

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*



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

2. Extract positional information about the event

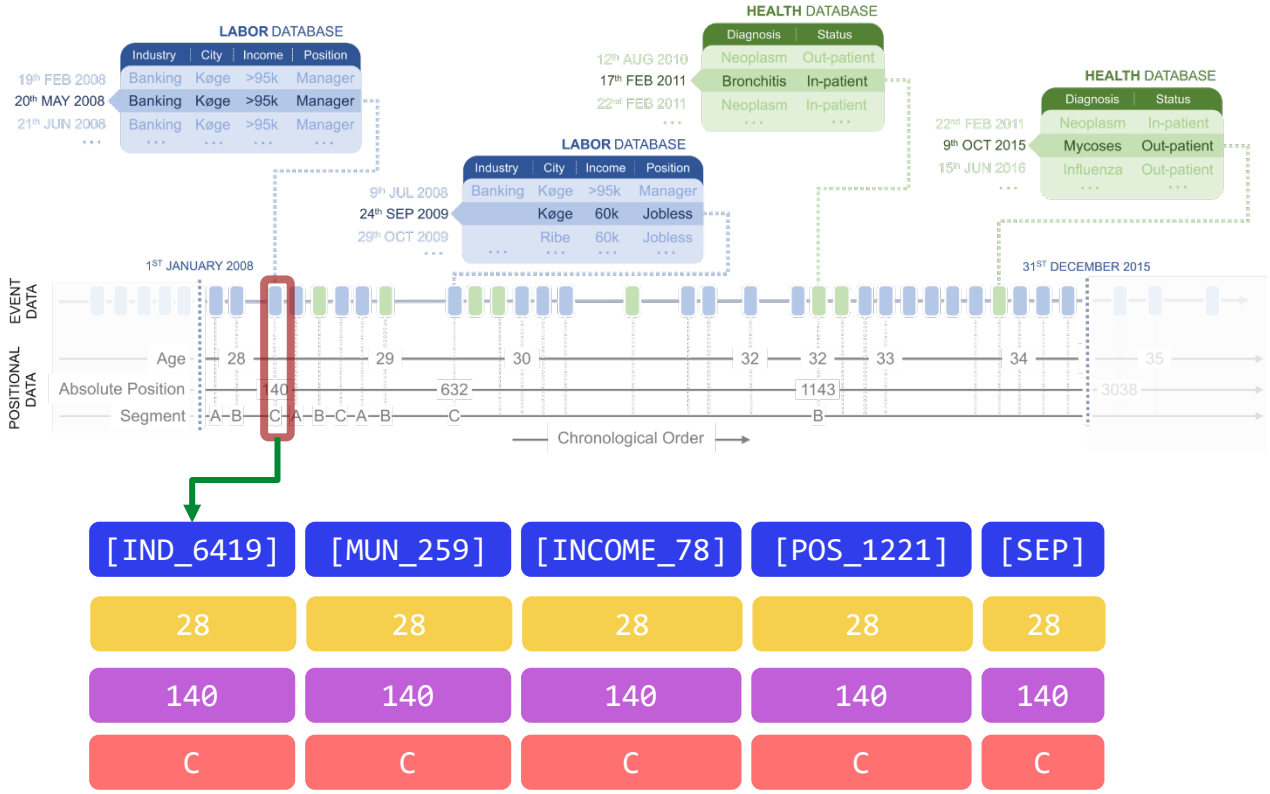
Age: 28

Global timestep: 140

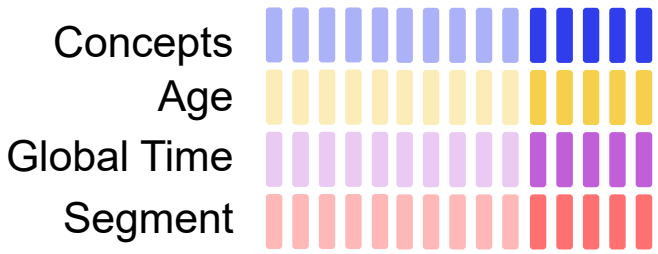
Segment: C

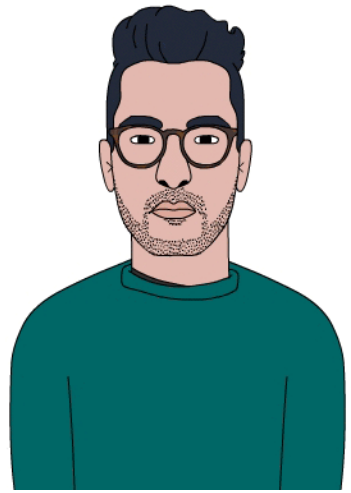
[IND_6419]	[MUN_259]	[INCOME_78]	[POS_1221]	[SEP]
28	28	28	28	28
140	140	140	140	140
C	C	C	C	C

** slightly simplified overview*

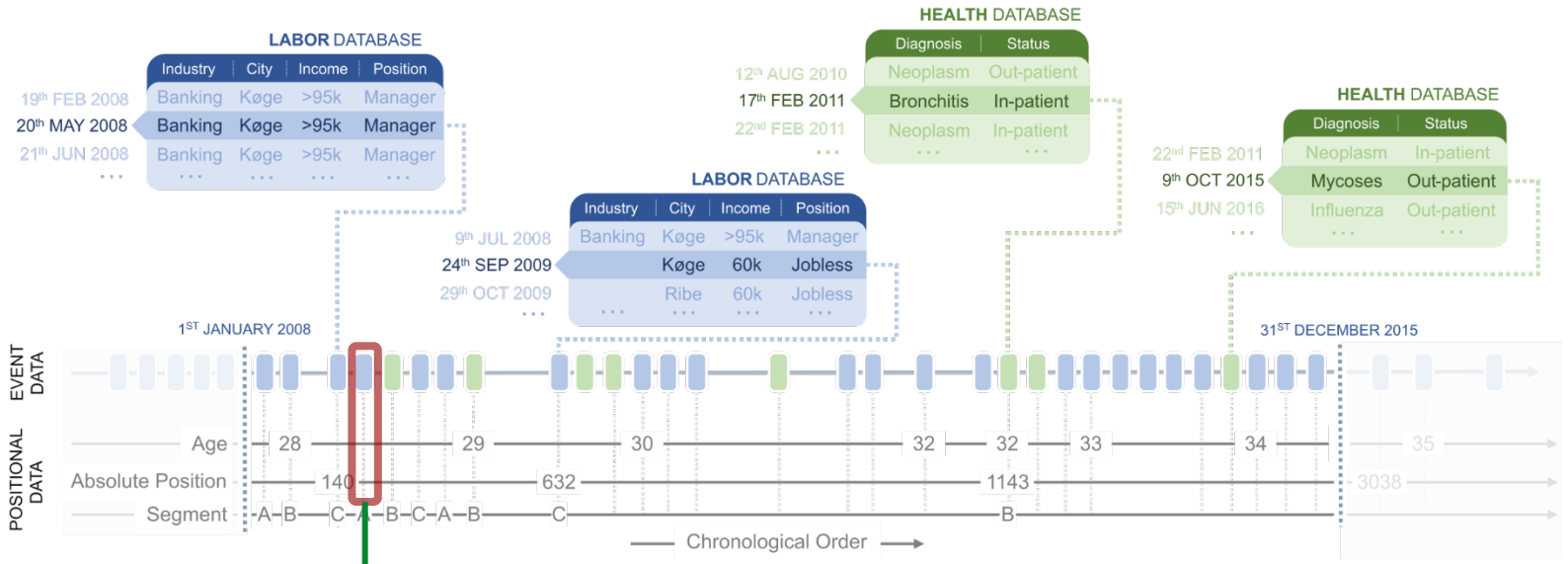


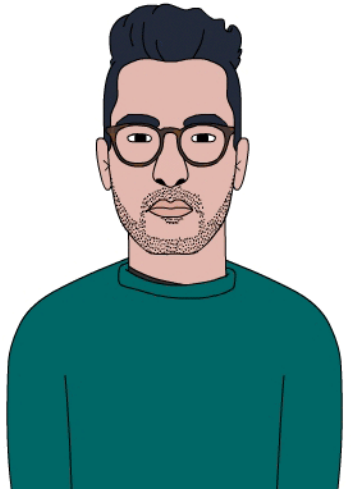
4. Insert data into the Life-Sequence (person document)

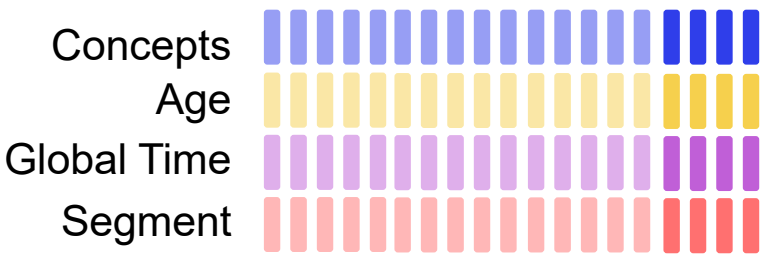




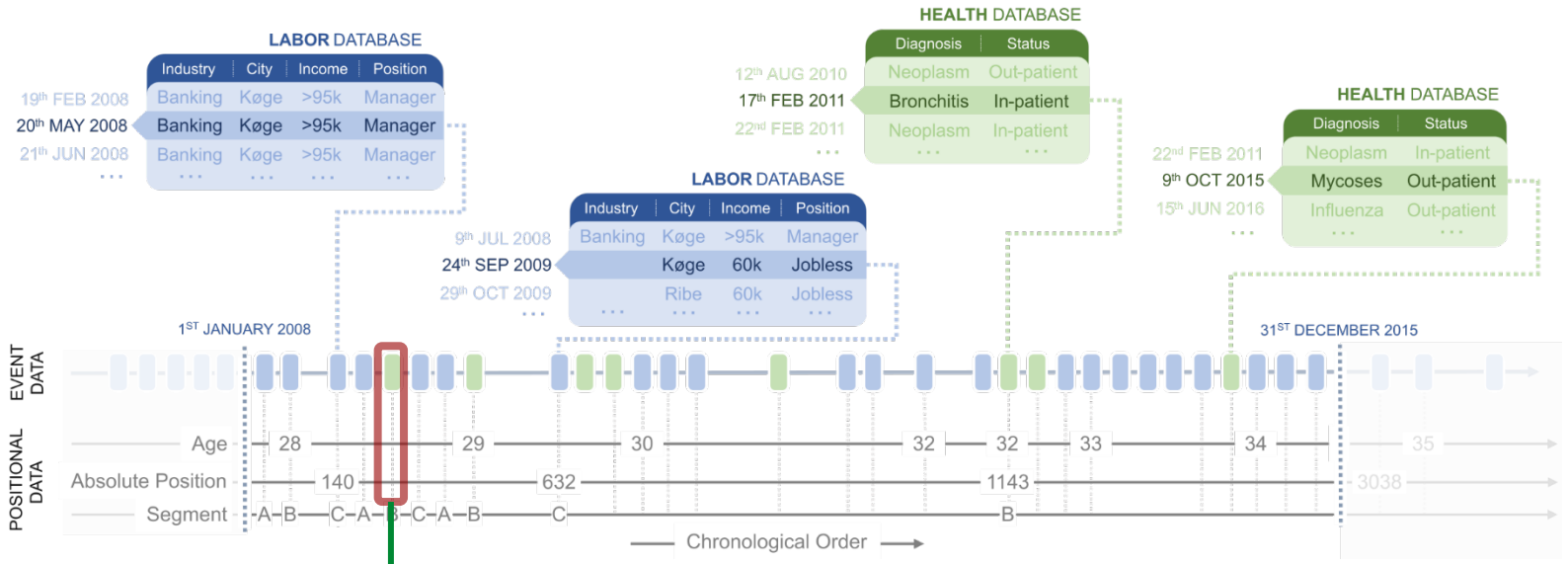
** slightly simplified overview*

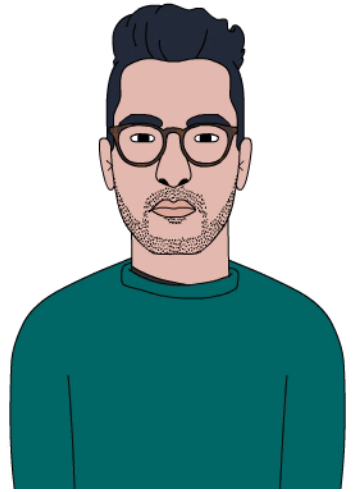


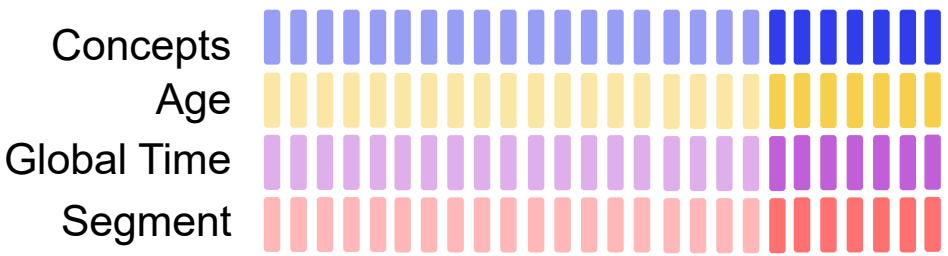




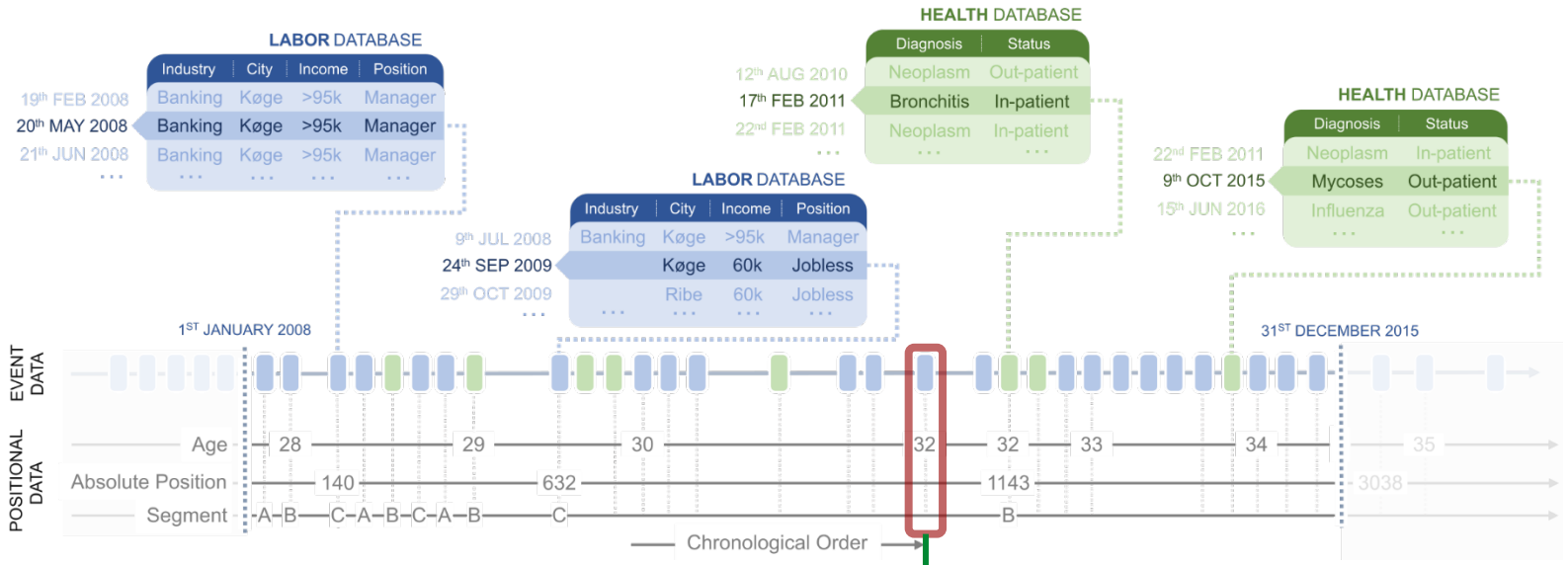
** slightly simplified overview*

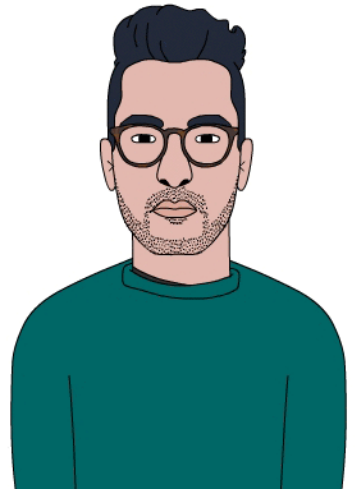


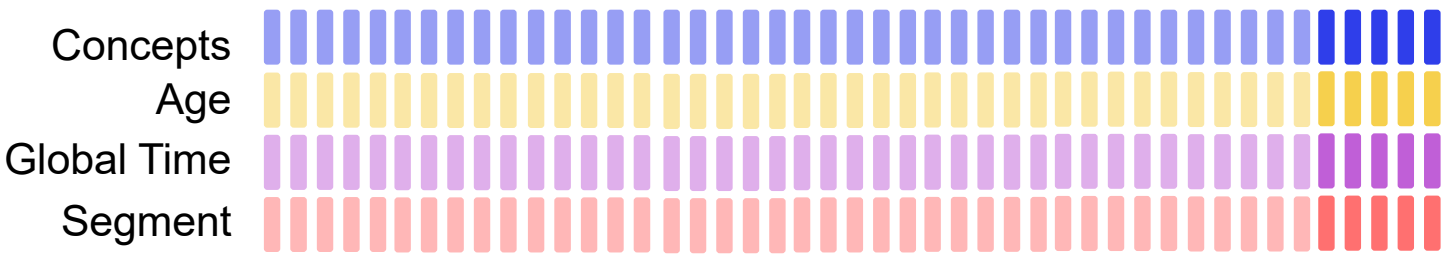




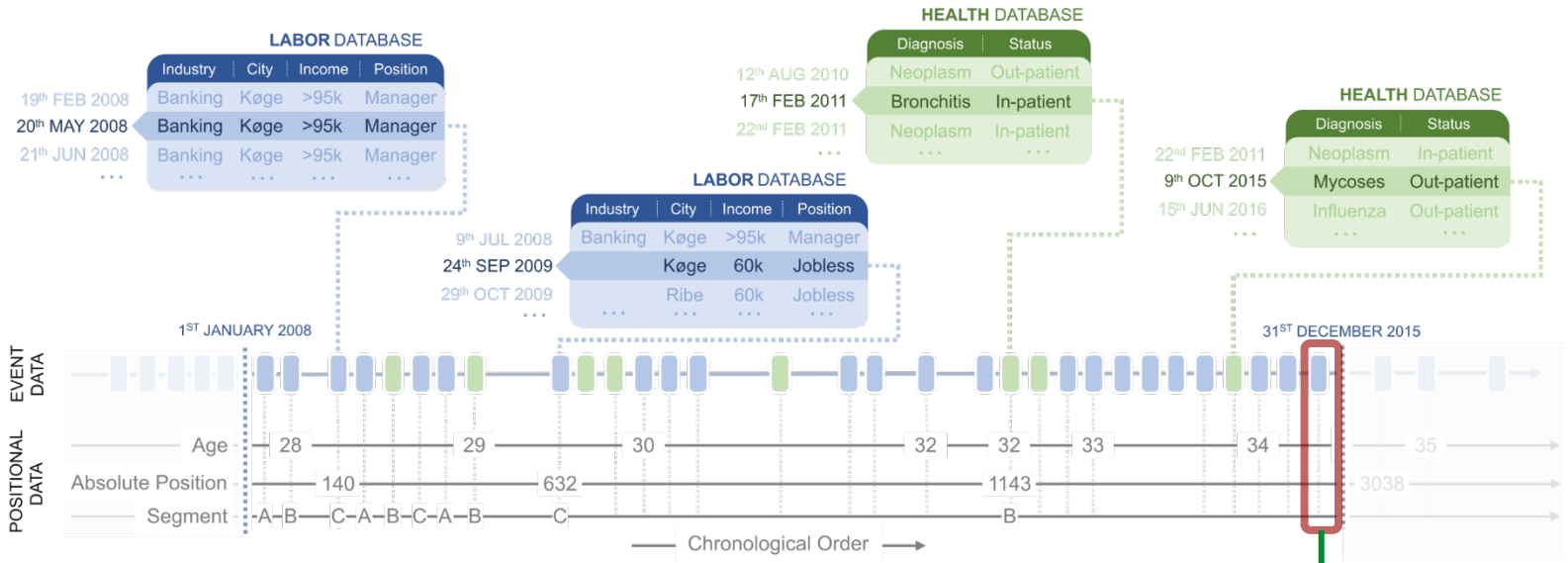
** slightly simplified overview*

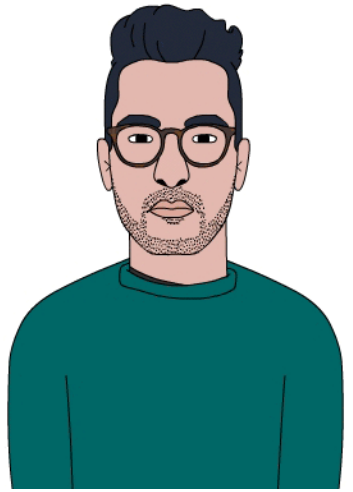


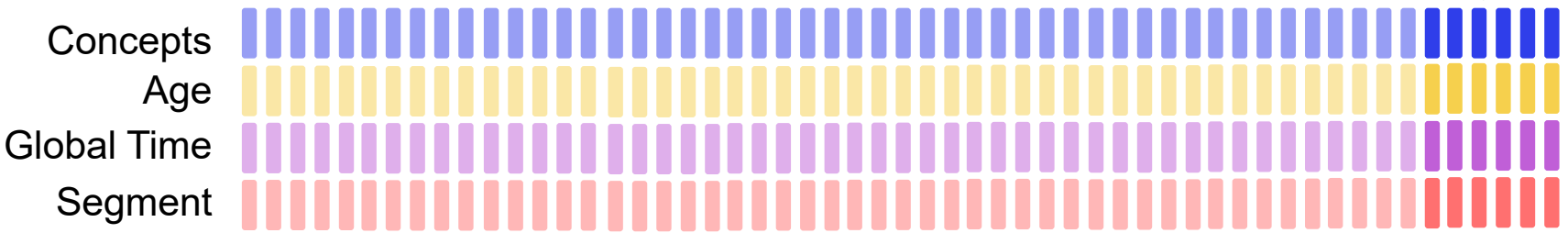




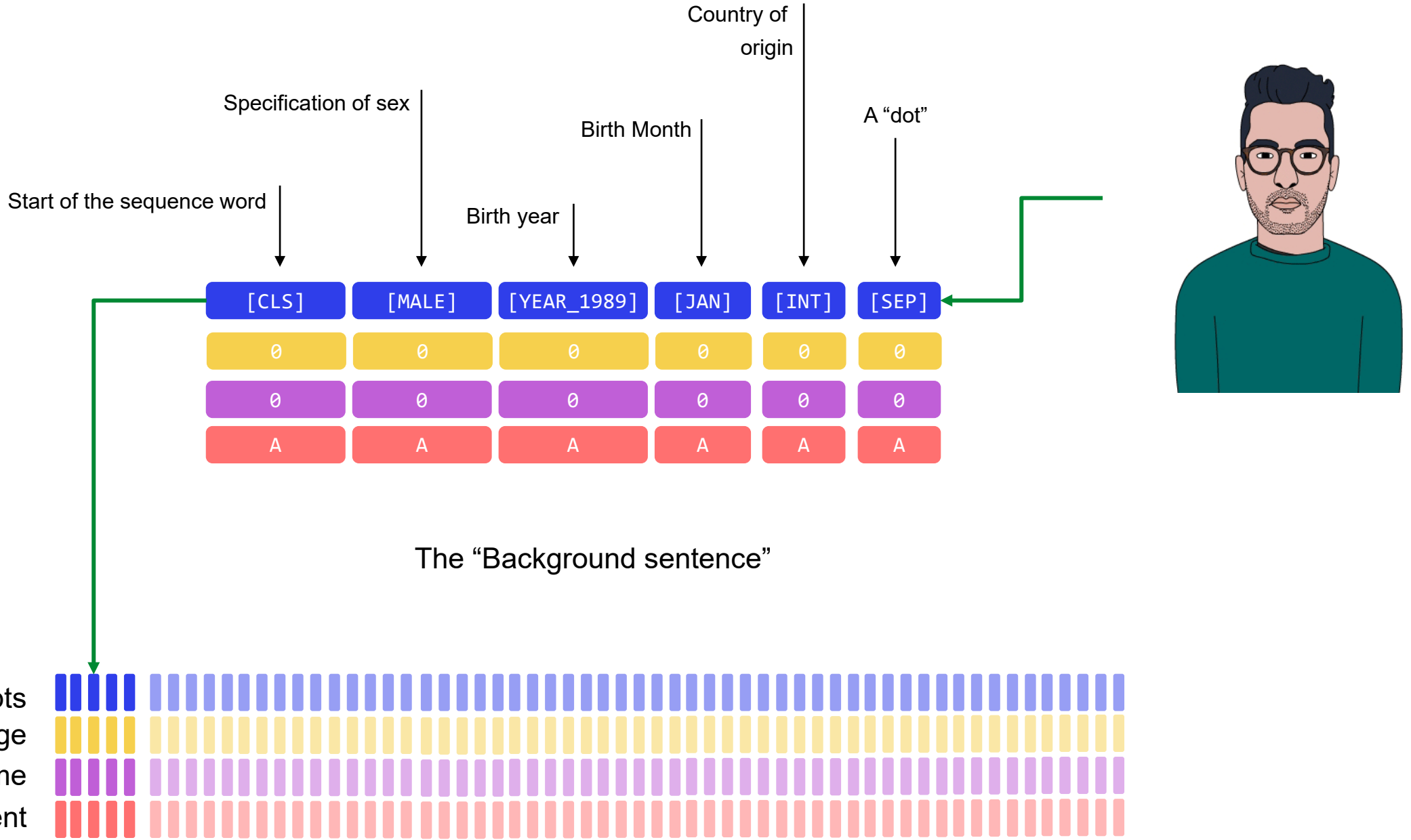
** slightly simplified overview*





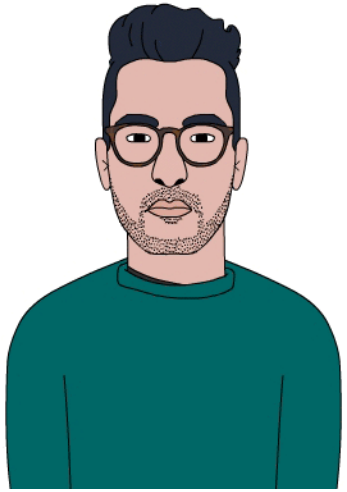


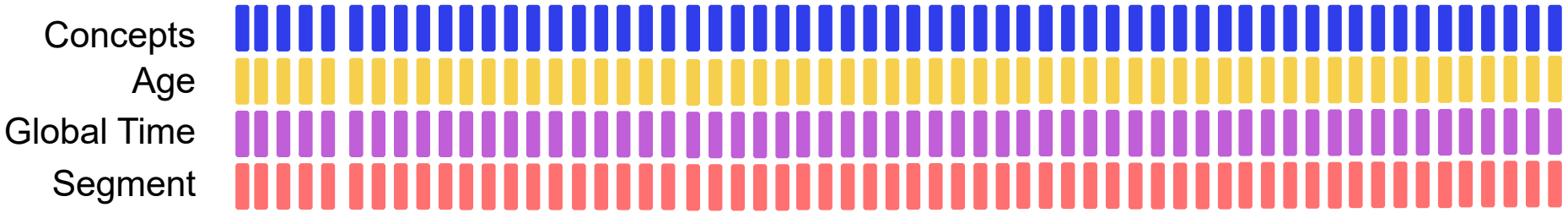
** slightly simplified overview*



** slightly simplified overview*

Individual Life-Sequence





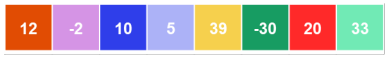
Input to the `life2vec` model

** slightly simplified overview*

Transformer-based Models

What do you think this quote means?

"*Everything Was Beautiful and Nothing Hurt*"*



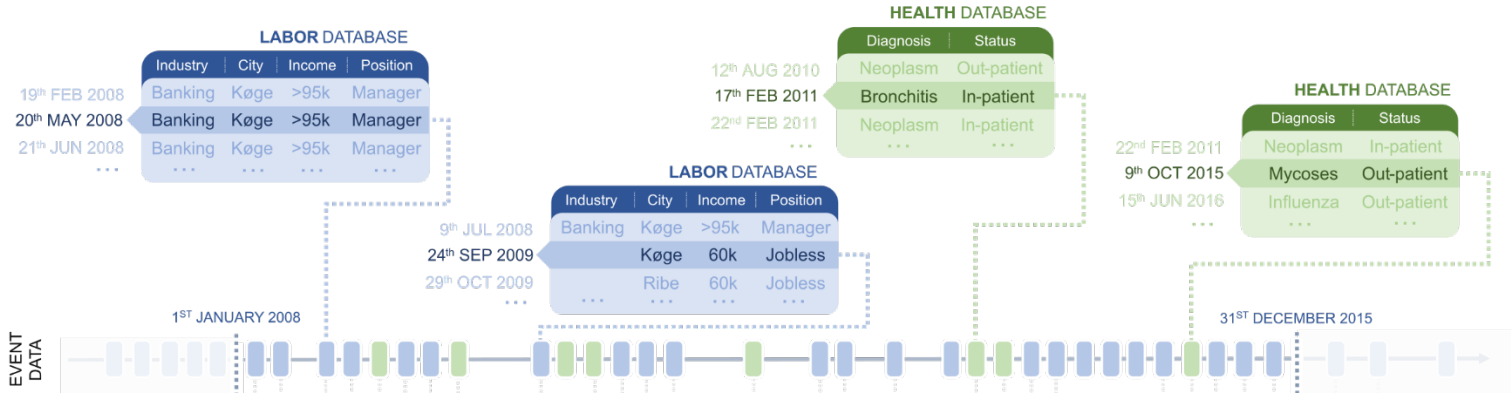
ChatGPT:
Text to text

[…]The line captures a complex array of sentiments, many of which are rooted in the themes of the book itself, such as the trauma of war, the nature of human experience, and the fluidity of time […]

* Slaughterhouse-Five, Kurt Vonnegut

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

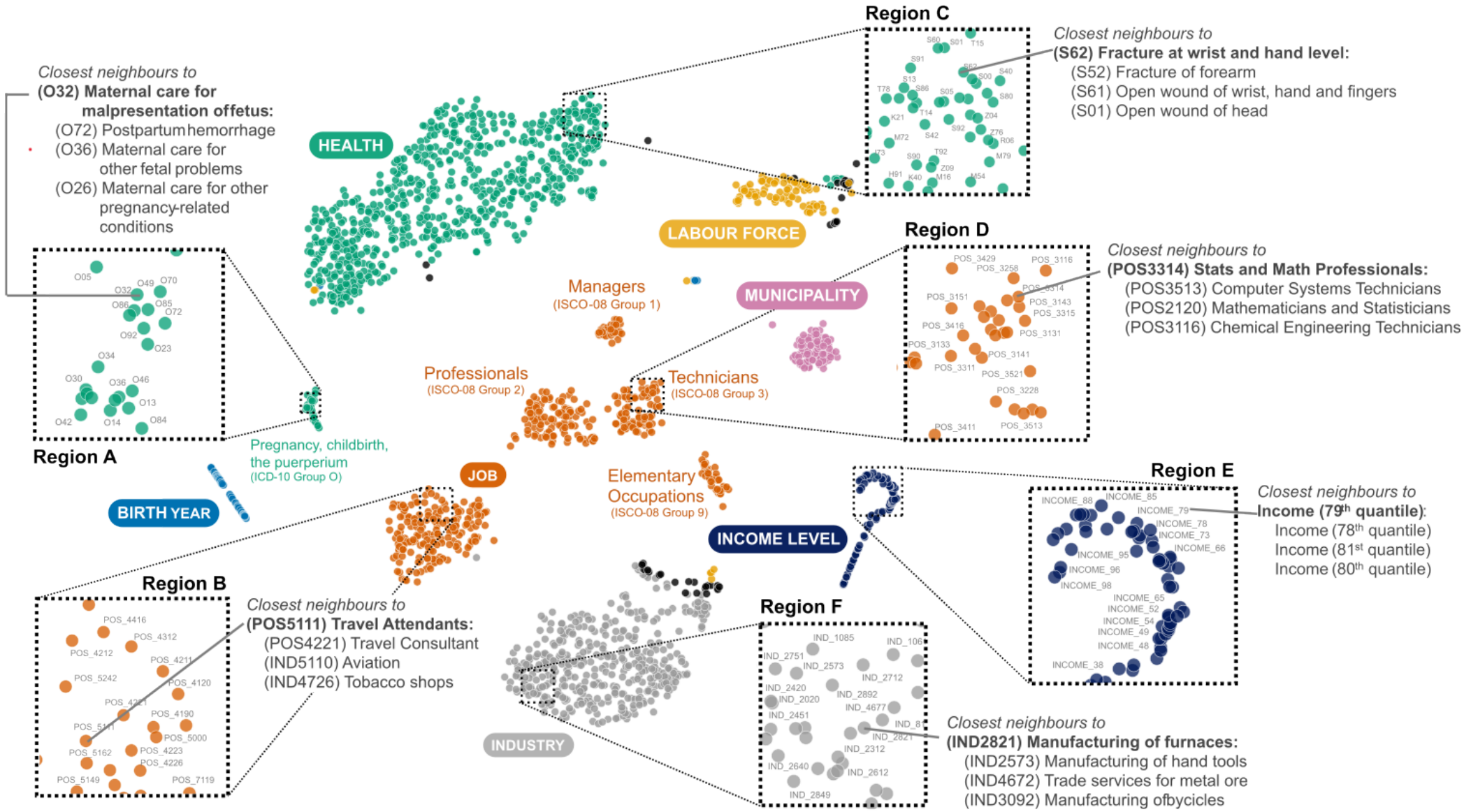
Life Progression from the point of view of Labor and Health Records



life2vec

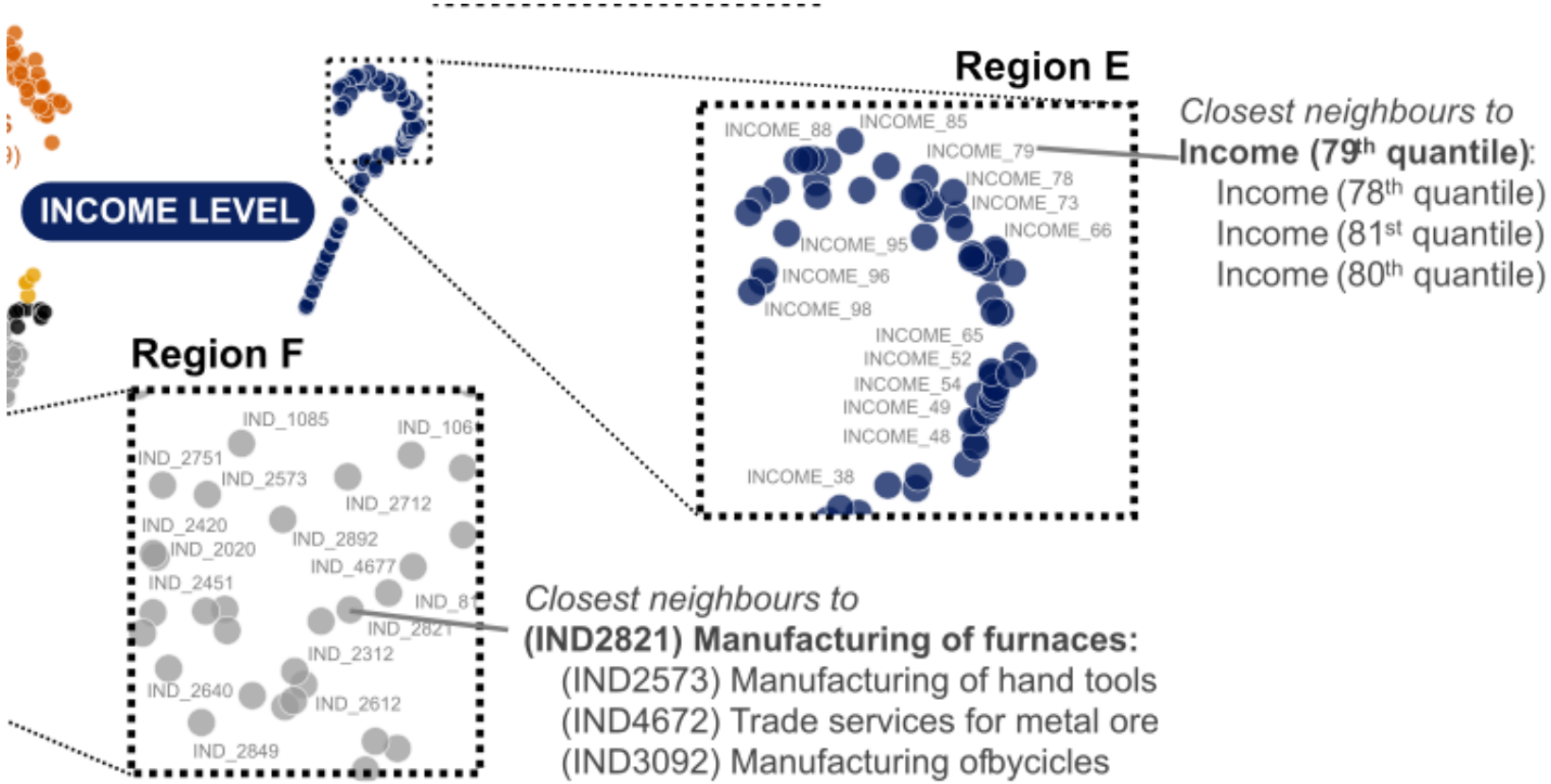
What did *life2vec* learn about the language?

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.

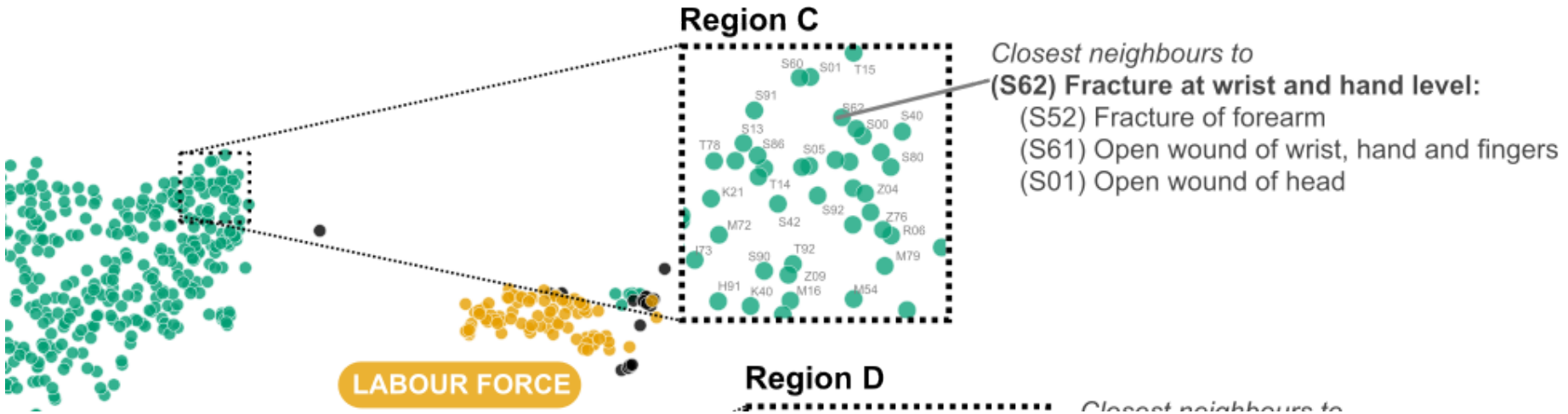
Space of concept tokens (with PaCMAP)



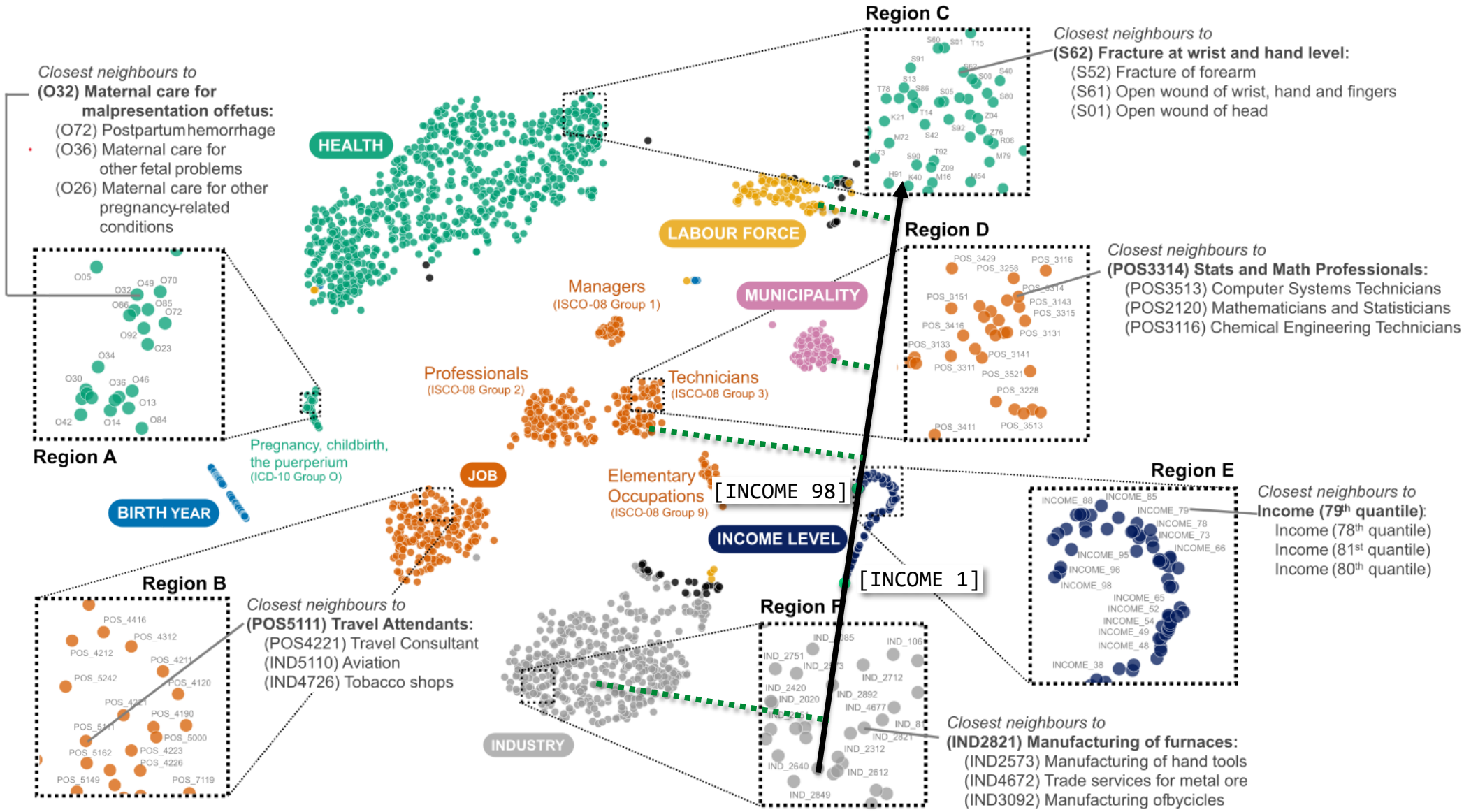
Closest neighbours to
Income (79th quantile):
Income (78th quantile)
Income (81st quantile)
Income (80th quantile)

Closest neighbours to
(IND2821) Manufacturing of furnaces:
(IND2573) Manufacturing of hand tools
(IND4672) Trade services for metal ore
(IND3092) Manufacturing ofbycicles

Space of concept tokens (with PaCMAP)



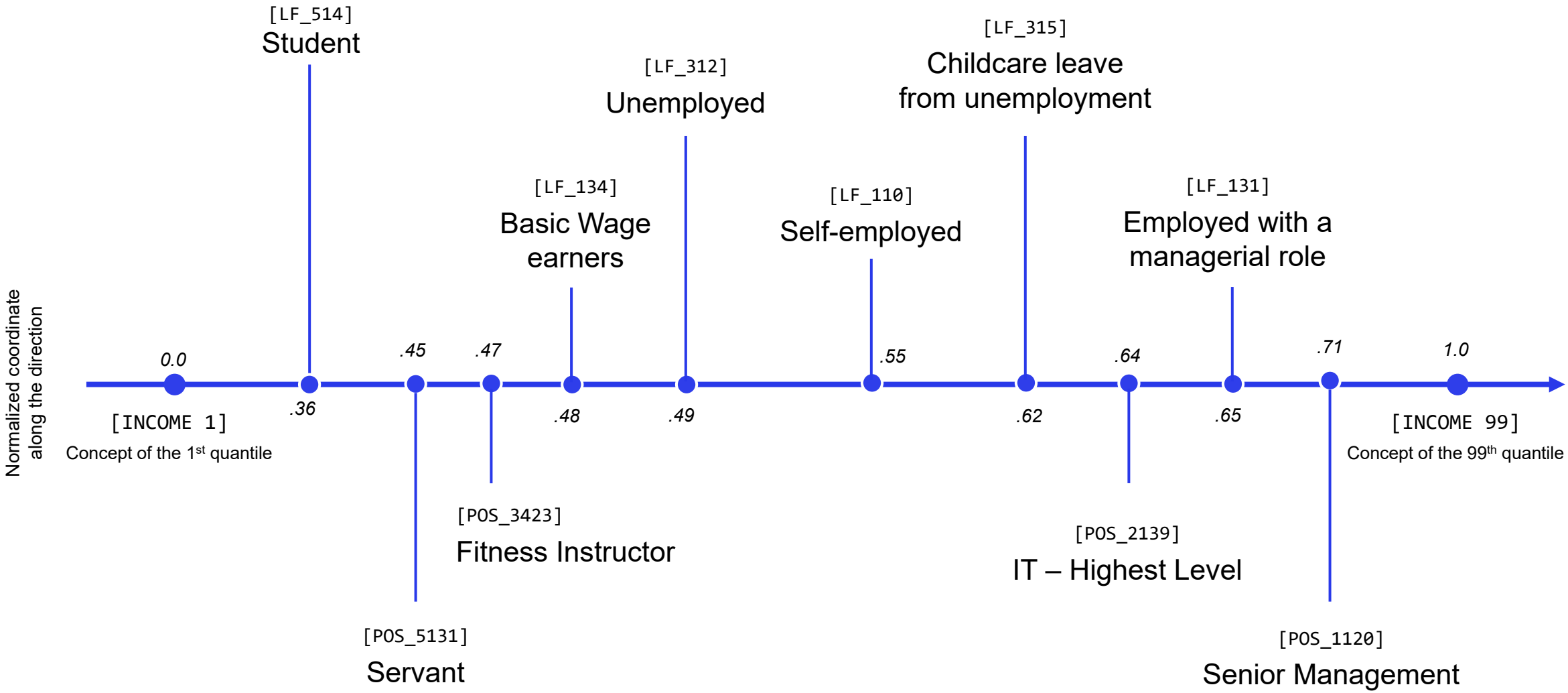
Space of Concept Tokens (with PaCMAP)



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.

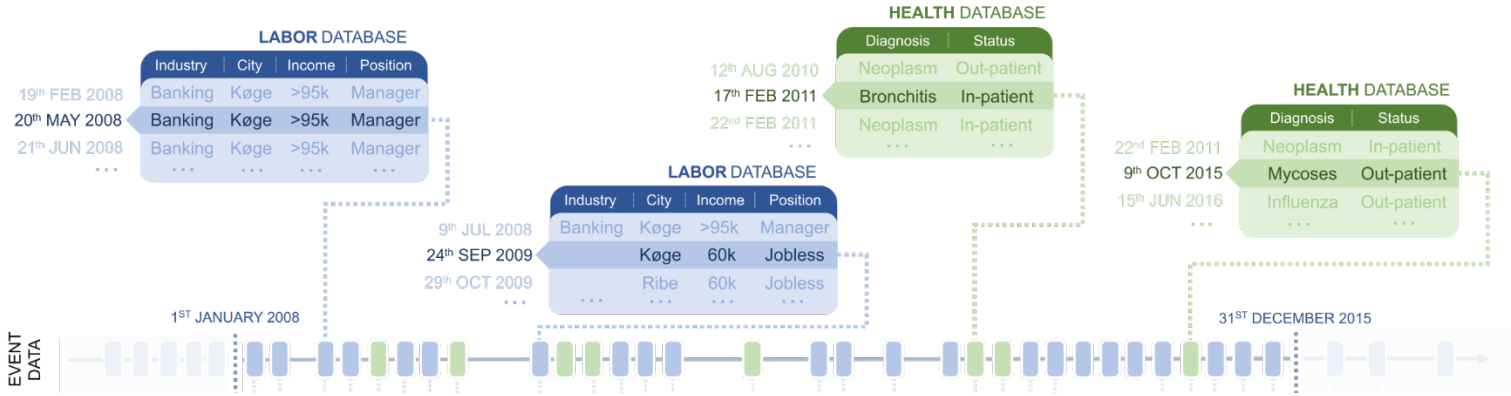
LF – Labor Force Status
POS – Prof. Position

Projection to "*Income*" Direction

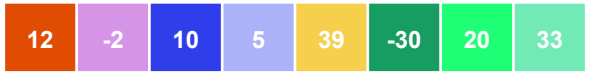


life2vec and mortality prediction

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



life2vec

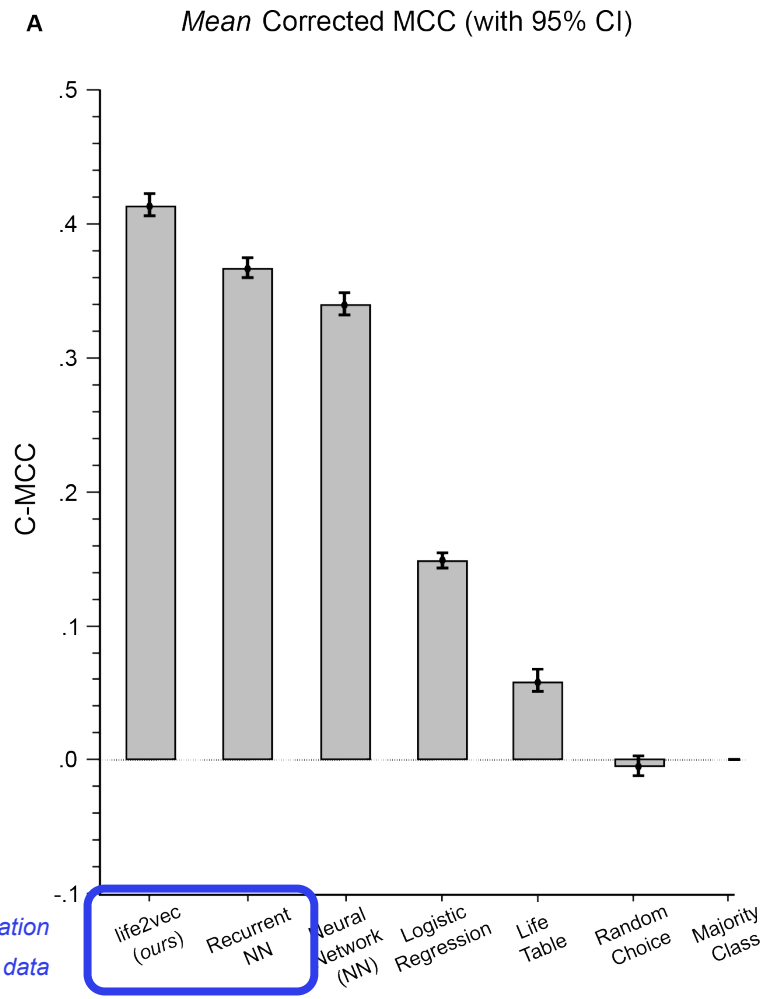


Compressed **representation of a sequence**

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

Early Mortality Prediction



Use textual representation of the data

True Labels

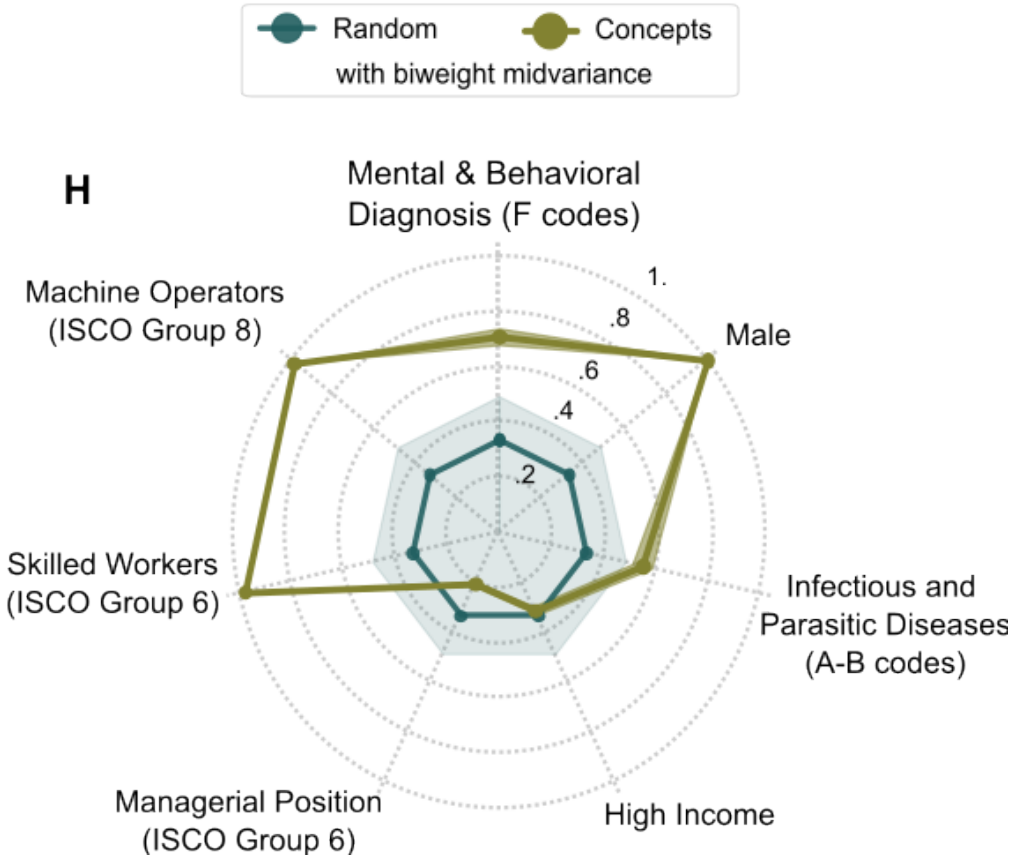
Predicted Labels

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

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

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

life2vec as interpretability tool

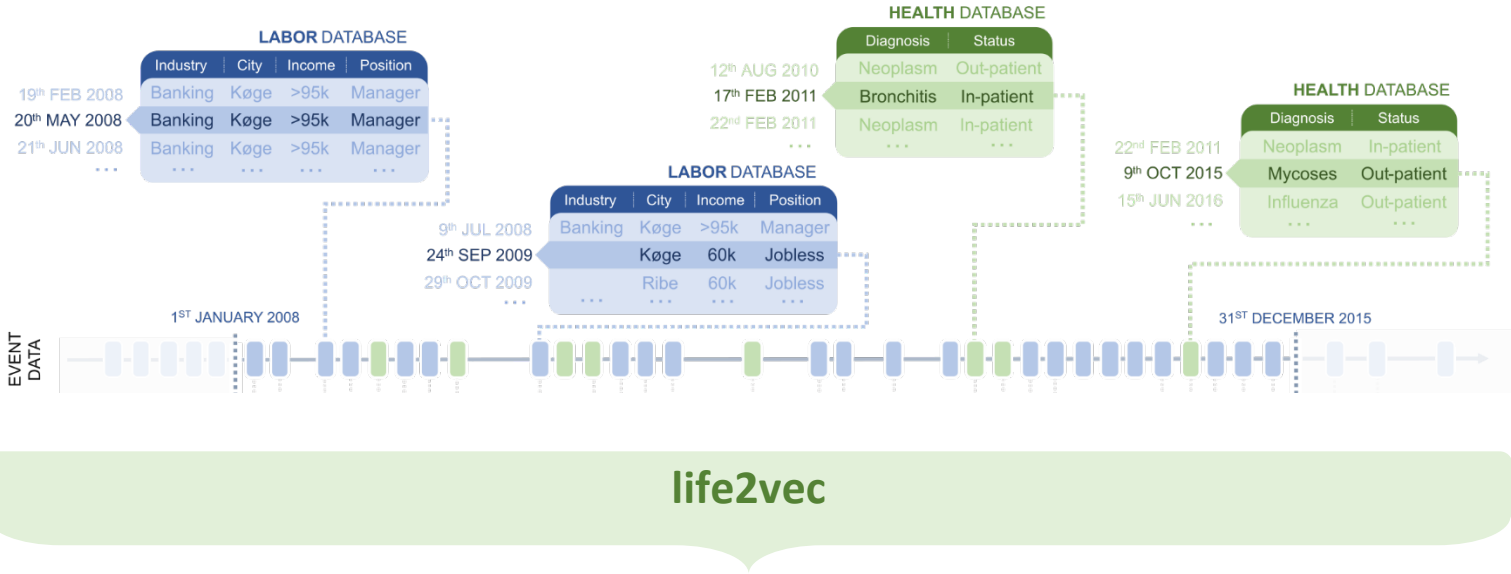


TCAV Score per "Direction"

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

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

Thank you for attention