Dimensionality Reduction

Raoul Grouls, 1 April 2025

Motivation for embedding data in high dimensional vector spaces as a design pattern

Supervised learning

$X \in \mathbb{R}^{c \times w \times h}$

$y \in \{0,1\}$

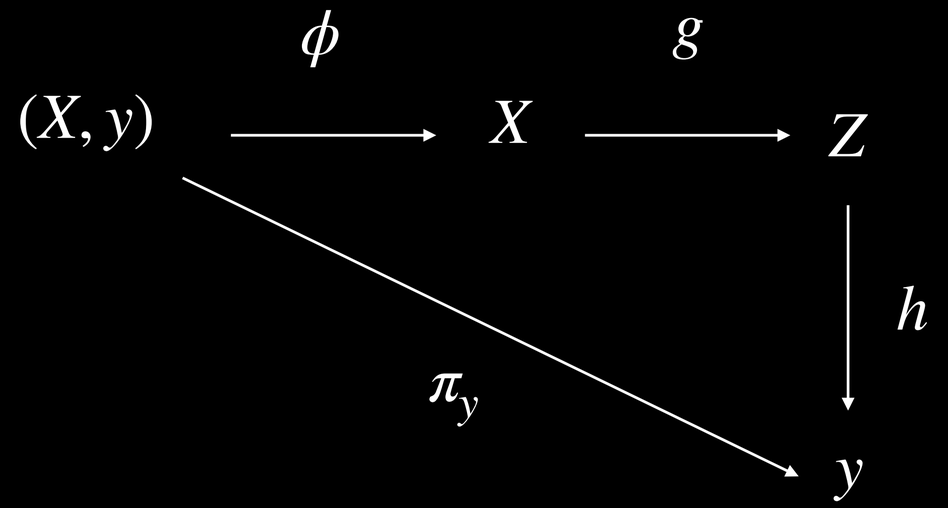
$Z \in \mathbb{R}^{d}$

$\phi \in \mathcal{T} = \{\text{mask}, \text{crop}, \text{flip}, \pi, \text{id}, \ldots\}$

$\phi : X \to X$

$g : X \to Z$

$h : Z \to y$



Autoencoder

$X \in \mathbb{R}^{c \times w \times h}$

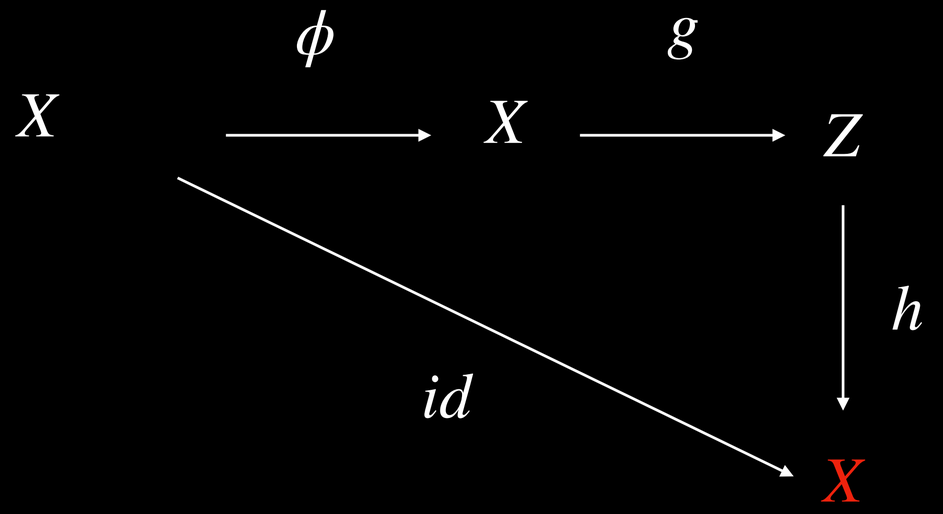
$Z \in \mathbb{R}^{d}$

$\phi \in \mathcal{T} = \{\text{mask}, \text{crop}, \text{flip}, \pi, \text{id}, \dots\}$

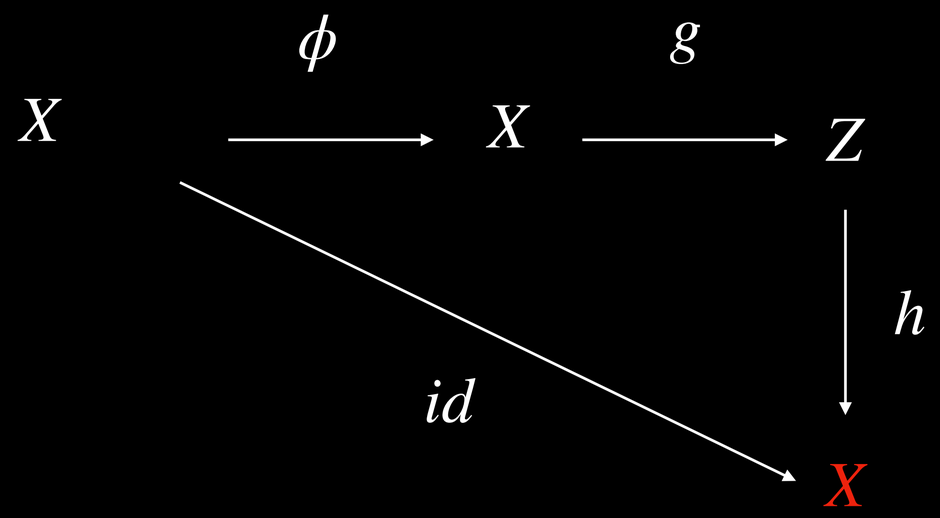
$\phi : X \to X$

$g : X \to Z$

$h : Z \to X$



Training

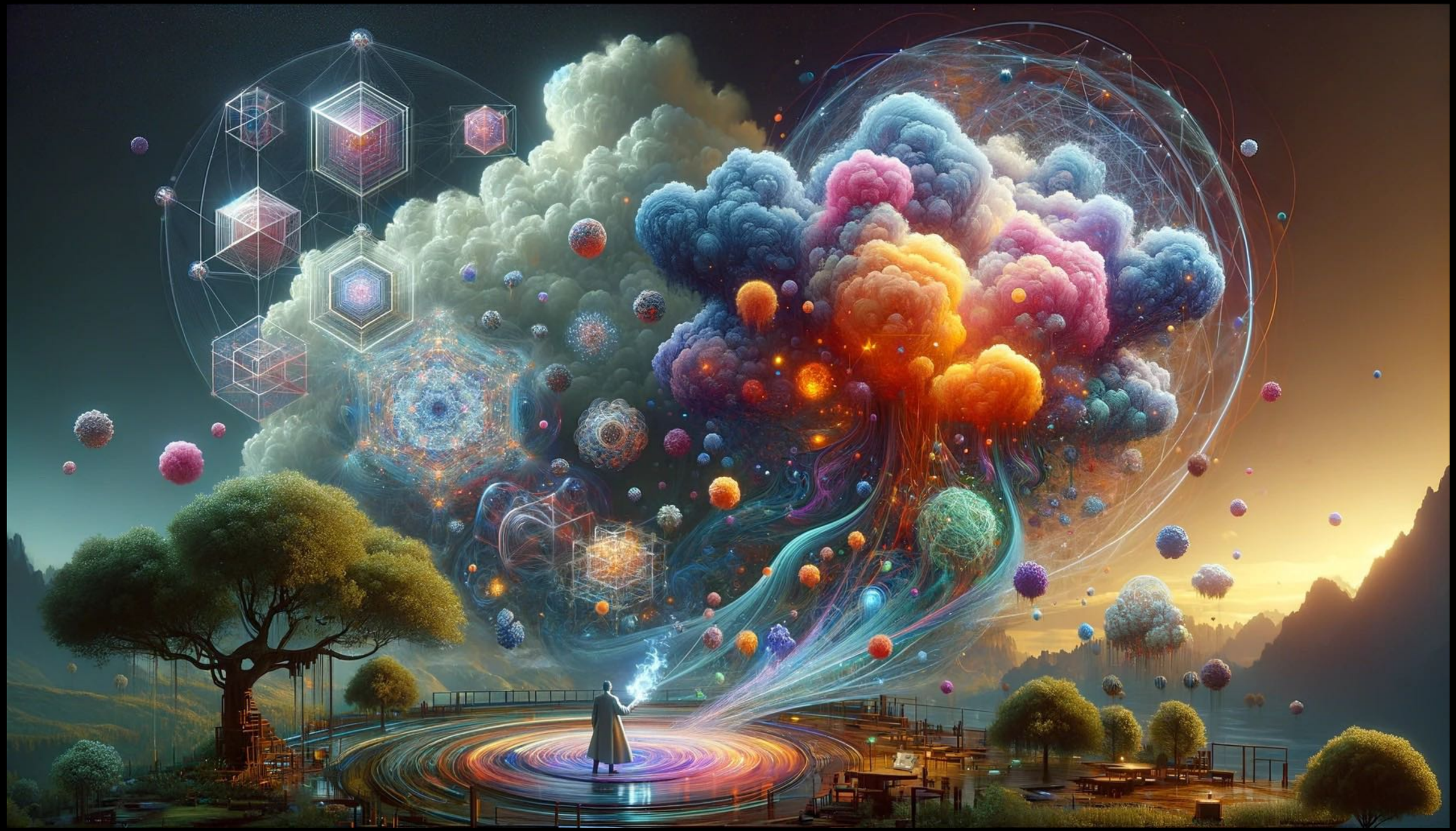


What you use

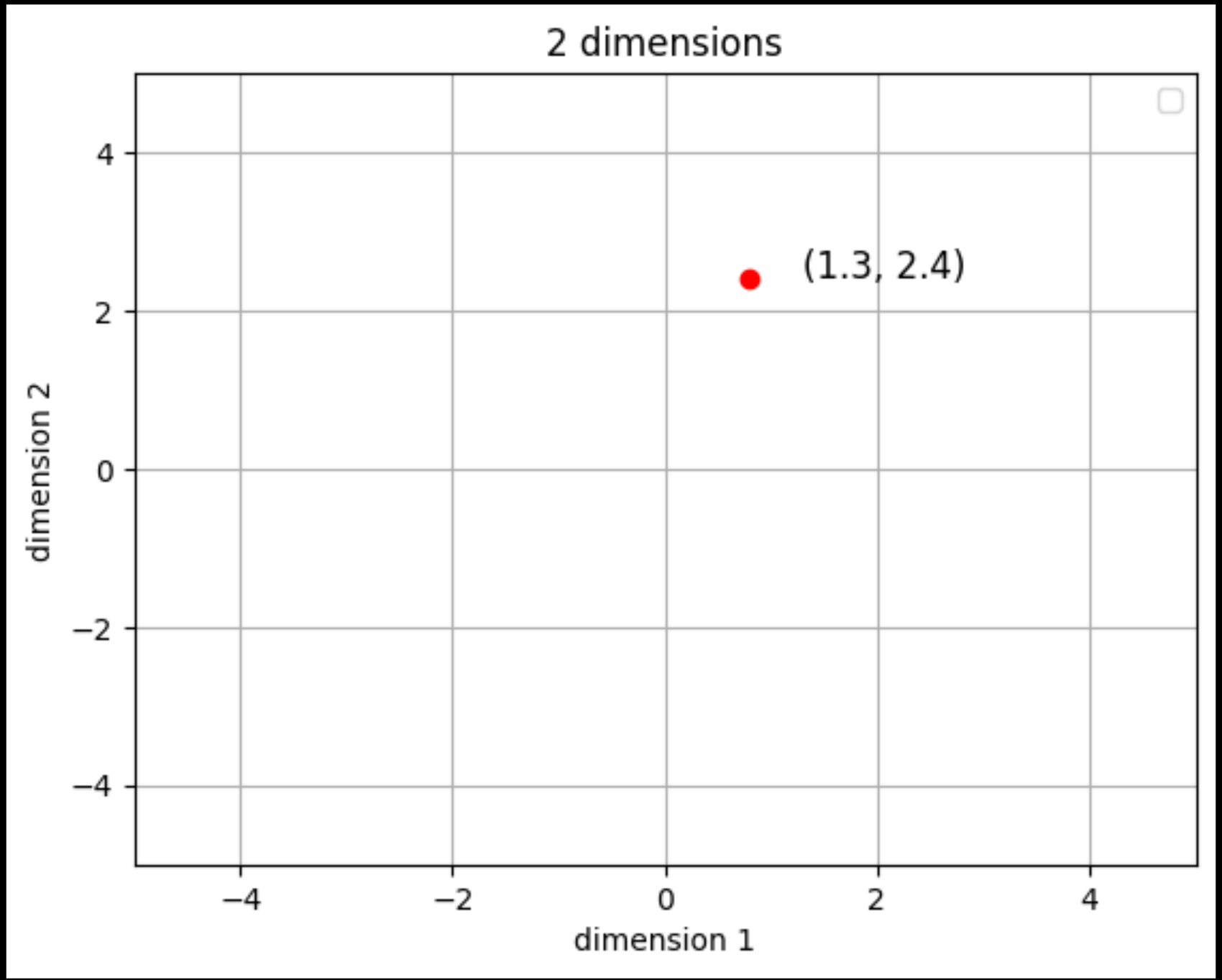
What can you do with $Z \in \mathbb{R}^d$?

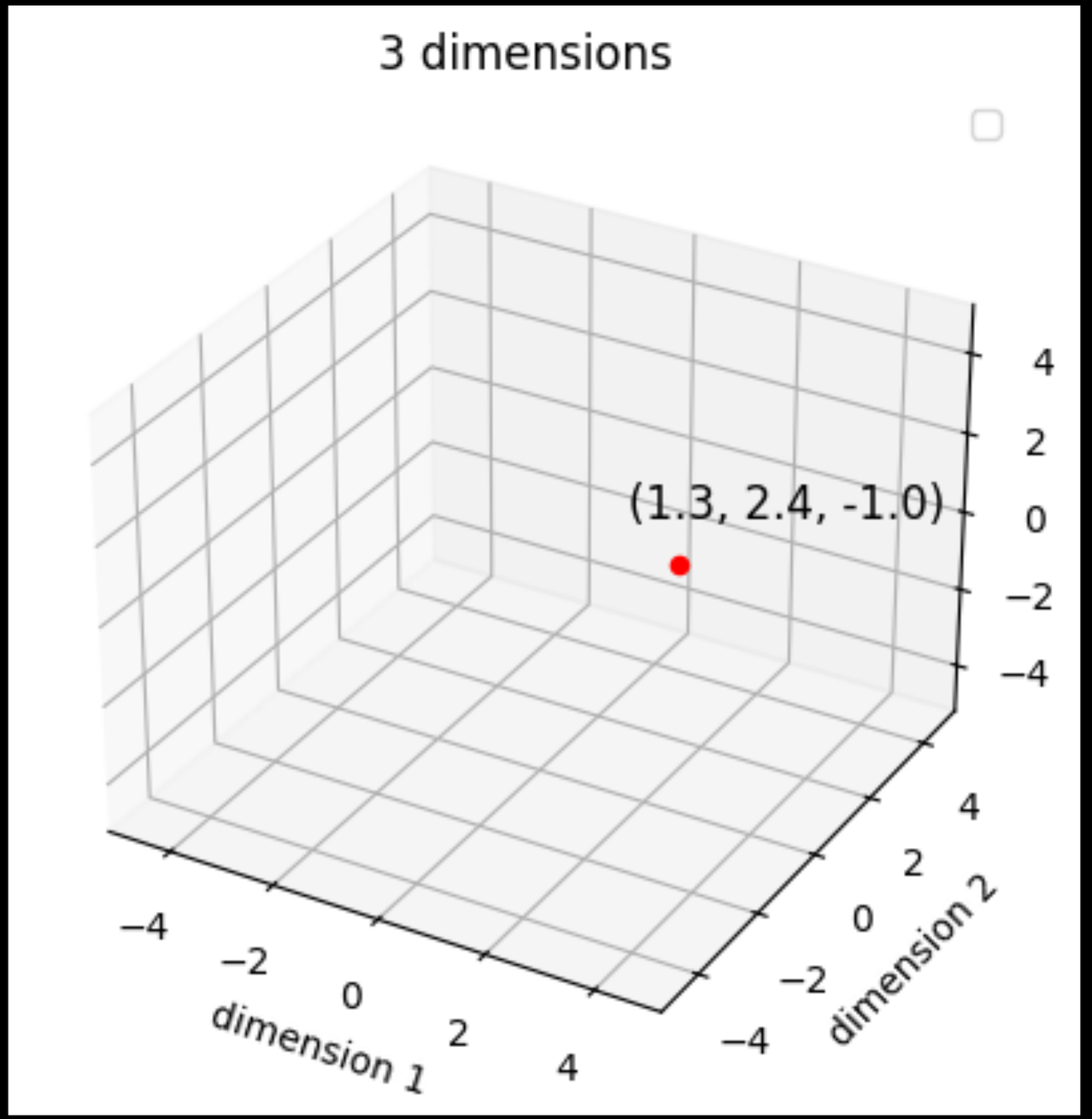
You can use Z as a starting point to

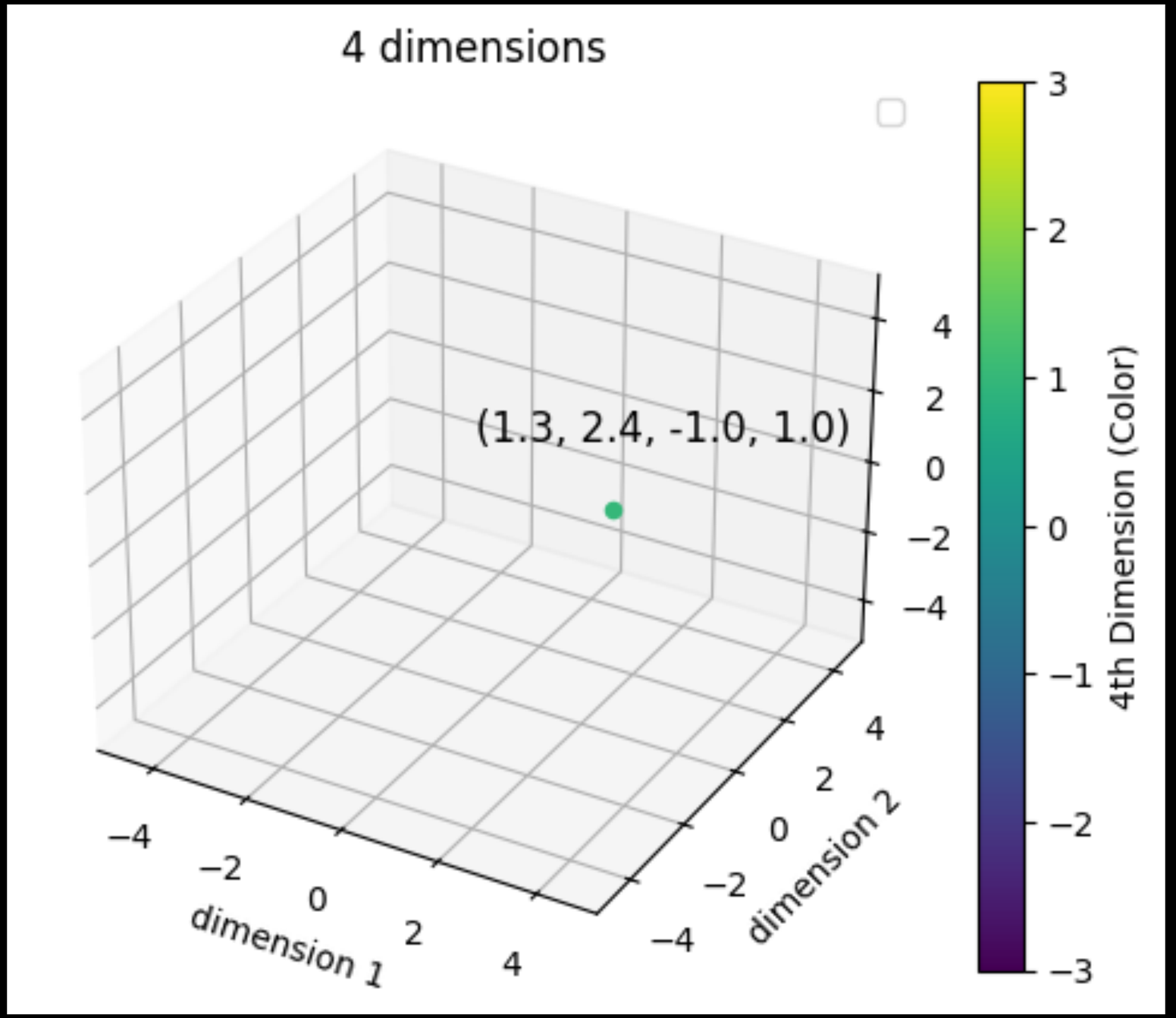
- Visualise your data (map to $\mathbb{R}^2$) to spot patterns
- Train your own function $h: Z \rightarrow y$
- Map a cluster of inputs to Z and perform clustering in z



Semantic vectorspace


2 dimensions
(1.3, 2.4)
dimension 1
dimension 2
4
2
0
−2
−4
−4
−2
0
2
4


3 dimensions
(1.3, 2.4, -1.0)
4
2
0
−2
−4
−4
−2
0
2
4
dimension 1
4
2
0
−2
−4
dimension 2


4 dimensions
(1.3, 2.4, -1.0, 1.0)
dimension 1
dimension 2
4th Dimension (Color)
−4
−2
0
2
4
3
1
−1
−3

Grote getallen

- cm^3 in een liter 10^3
- Stappen rond de Aarde 4×10^{10}
- 1.5×10^{11} m tot de zon
- Neuronen in een brein 10^{11}
- Cellen in het lichaam 10^{14}
- Mieren op aarde 10^{16}
- Seconden in een jaar 3.2×10^{16}
- Zandkorrels op aarde 10^{19}
- Druppels water in alle oceanen 10^{25}
- Atomen in het menselijk lichaam 10^{28}
- Bacterien 10^{30}
- Atomen in de Aarde 10^{50}
- Atomen in het zonnestelsel 10^{57}
- Manieren om een kaartendek te schudden 10^{68}
- Atomen in het zichtbare heelal 10^{80}

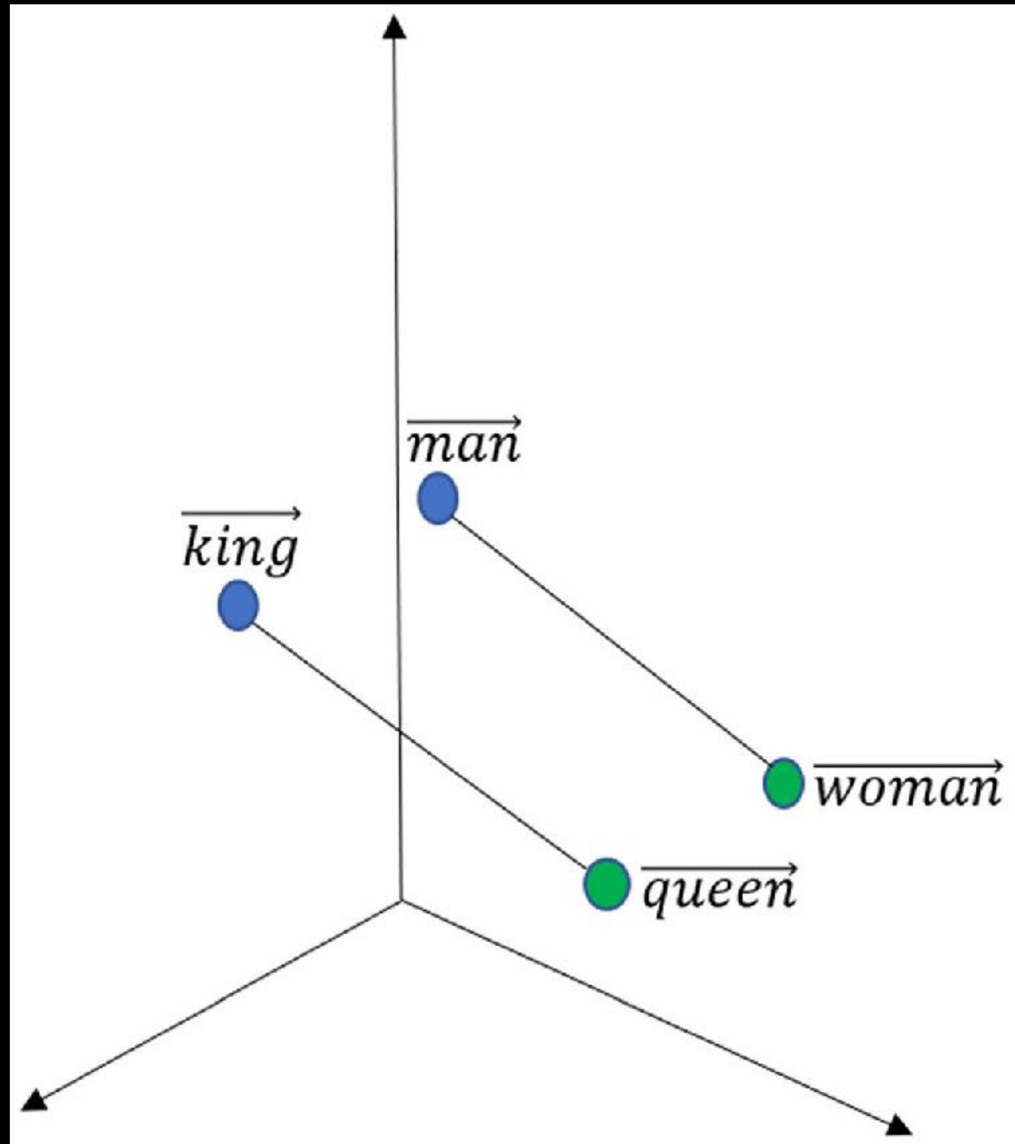
How big is 10^{68} (shuffle a stack of cards)

1. Count seconds for a year (3.2×10^{16} sec), then take one step forward (about 1 meter)
2. Once you've walked around the Earth's equator (which would take about 4×10^{10} steps), take a drop of water out of the Ocean.
3. When all the Oceans are empty (10^{25} drops), place one sheet of paper on the ground.
4. Repeat this until the stack of paper reaches the Sun ($1.5 \times 10^{11} m \times 10^{4} sheets/m$)
5. This gives us $16 + 10 + 25 + 11 + 4 = 66$, and $3.2 \times 4 \times 1.5 = 19.2$, so about 1.9×10^{67}, which means we need to do this whole process about 5 times.

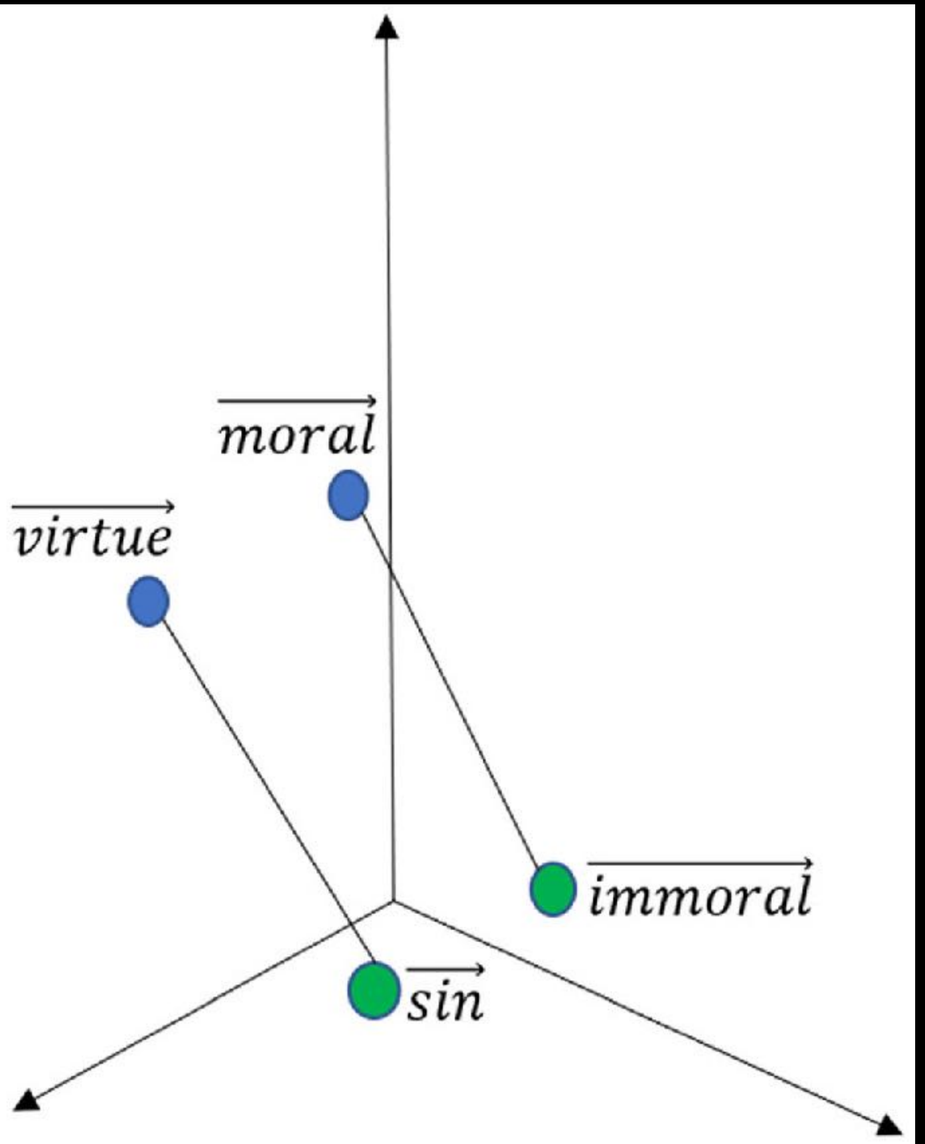
Semantic vectors

$$(x_1, x_2, x_3, \ldots, x_{766}, x_{767}, x_{768})$$

$$\mathbb{R}^{768}$$



a "gender" dimension



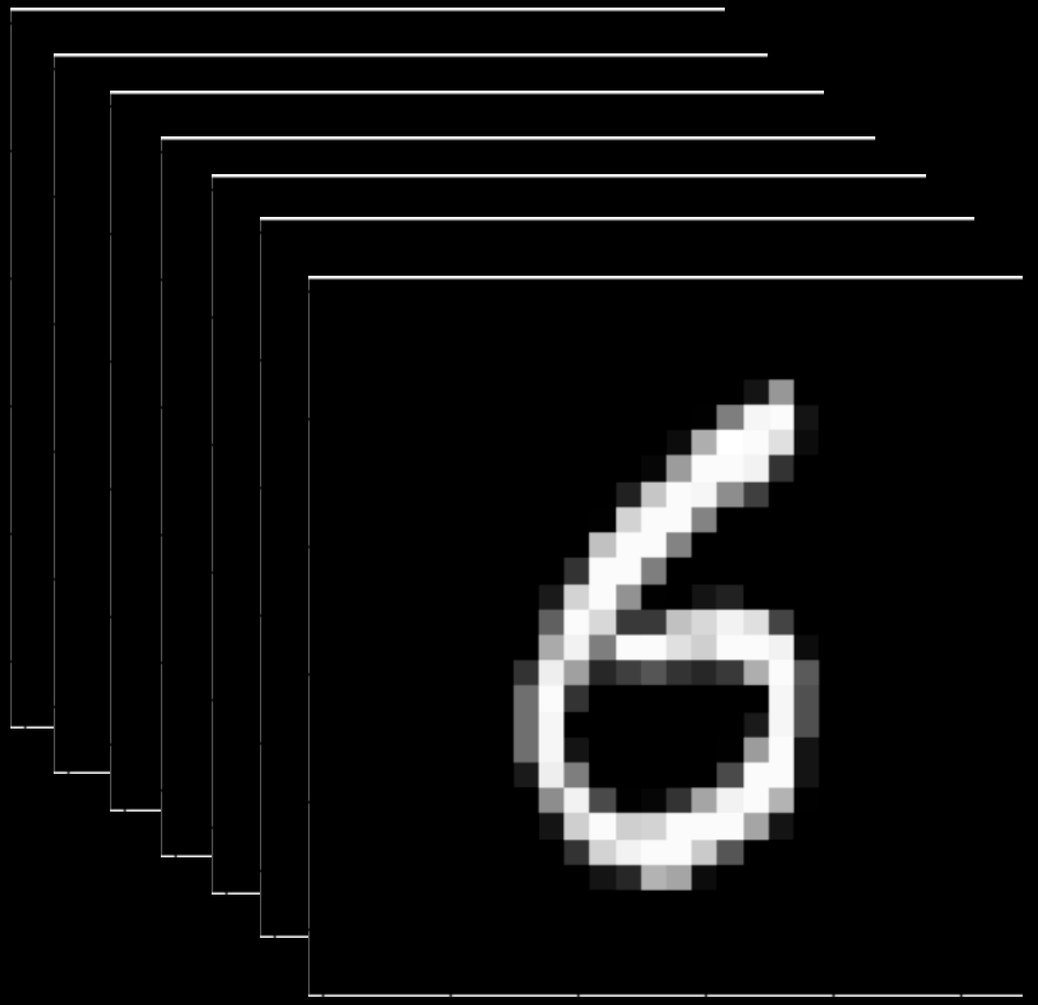
a "morality" dimension

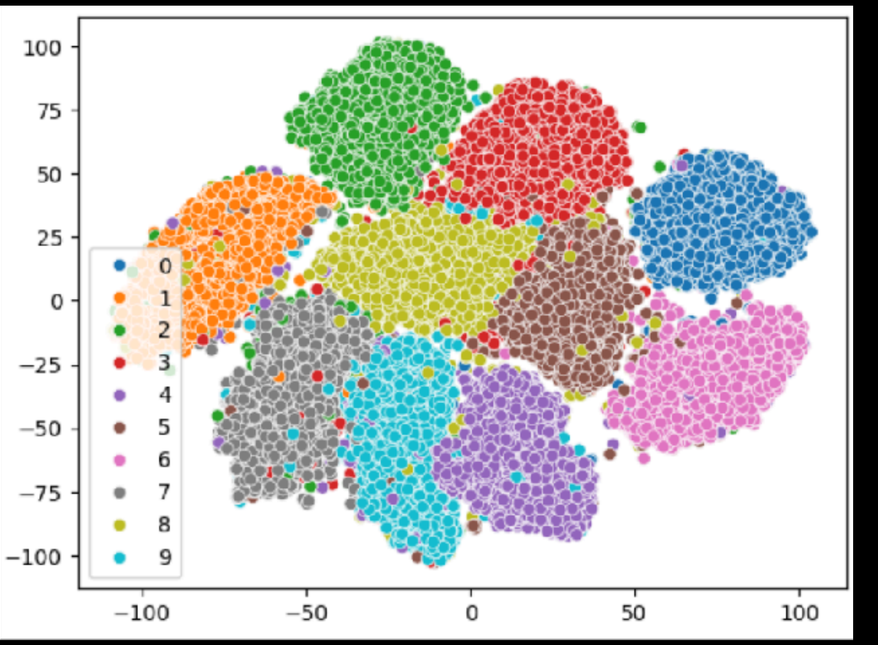
0.29	-0.50	-0.38	0.30	2.80	-1.67	2.36	-2.54
1.29	1.32	-2.84	1.25	0.28	2.22	-1.01	1.68
0.62	-3.00	0.30	-1.25	2.84	-1.76	1.93	-0.37
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
lk	kr	ijg	geld	van	de	ba	nk
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.29	0.92	0.42	0.06	0.95	-1.63	-1.01	-2.67
1.29	-2.31	0.39	0.51	2.07	-0.84	-2.30	2.99
0.62	2.70	-0.07	1.27	-2.06	1.37	1.31	1.42

Motivation for dimensionality reduction

Manifold hypothesis

- although high-dimensional data $\mathbb{R}^{c \times w \times h}$ like images (or text and sound) might appear complex (eg 60.000 images of handwritten numbers)
- they actually lie on or near a much lower-dimensional manifold
- So what if we were able to uncover the much simpler representation? For example, could we map $\mathbb{R}^{28 \times 28} \rightarrow \mathbb{R}^{2}$




100
75
50
25
0
−25
−50
−75
−100
−100
−50
0
50
100
0
1
2
3
4
5
6
7
8
9

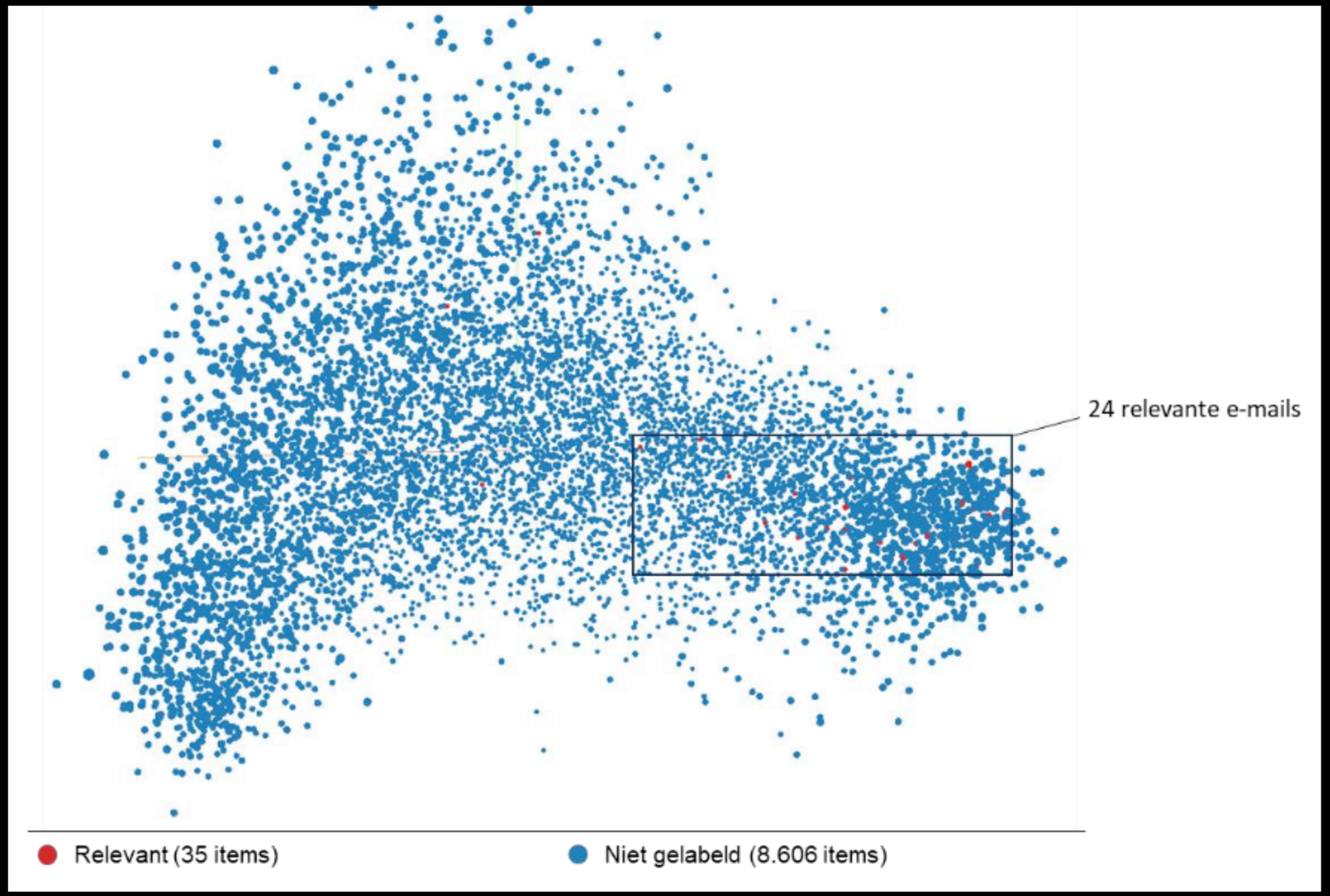
Sentiment classification of text

- tokenizer: $string \rightarrow \mathbb{N}$
- Vectorizer $g: \mathbb{N} \rightarrow \mathbb{R}^d$
- Result: $h: \mathbb{R}^d \rightarrow [0,1]$
- Model : $f = h(g(x))$, of $f = h \circ g$

Fraud detection

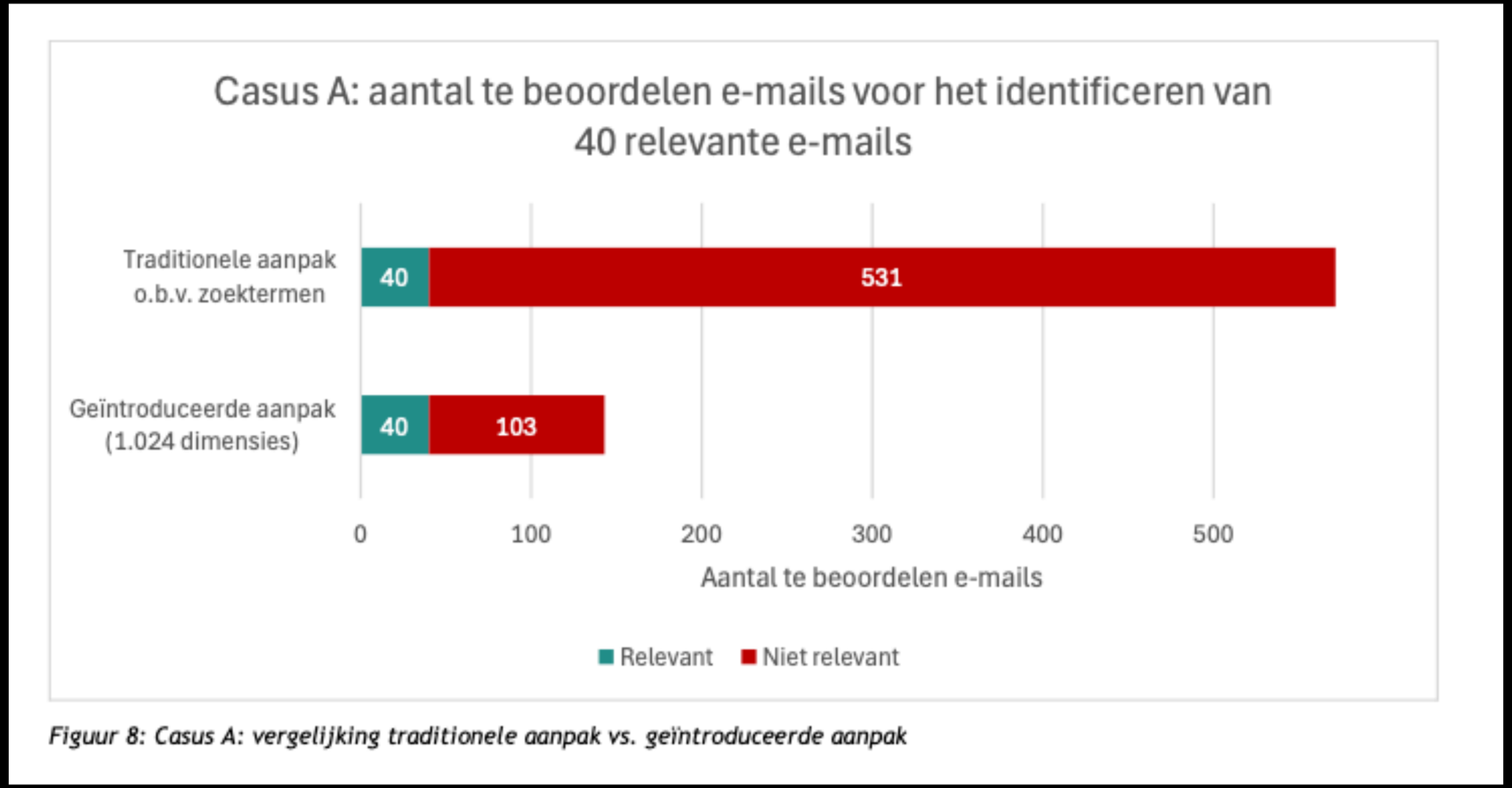
- tokenizer: $string \to \mathbb{N}$
- Vectorizer $g : \mathbb{N} \to \mathbb{R}^d$
- In $\mathbb{R}^d$ find a few relevant emails by hand
- Now search in $\mathbb{R}^d$ for emails close by

Fraud detection



M.M. van Zandwijk, 2024, Detecting potentially relevant e-mails for forensic irregularity investigations using Large Language Models (a FraudBERTa story), masterthesis HU

Fraud detection

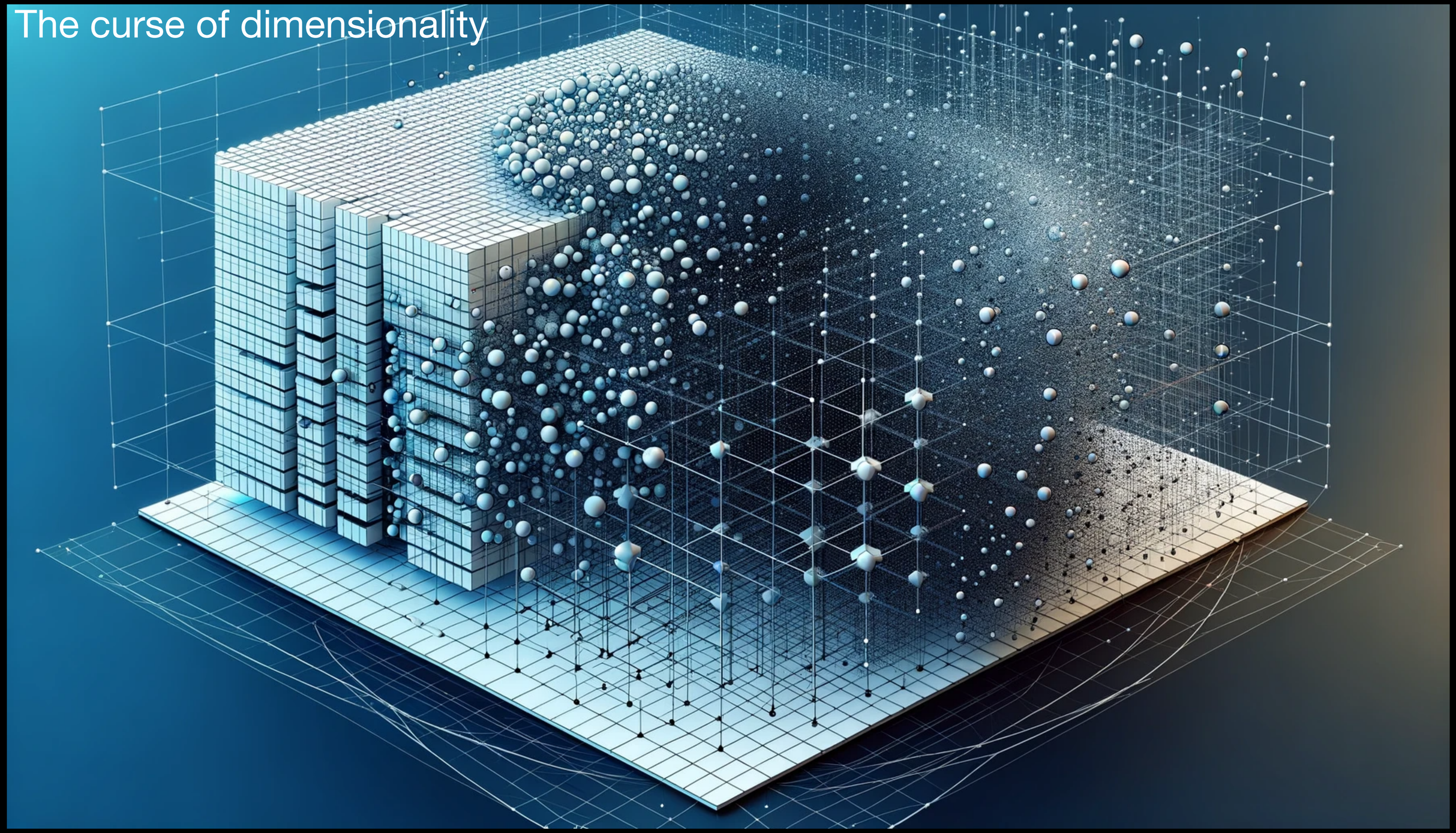


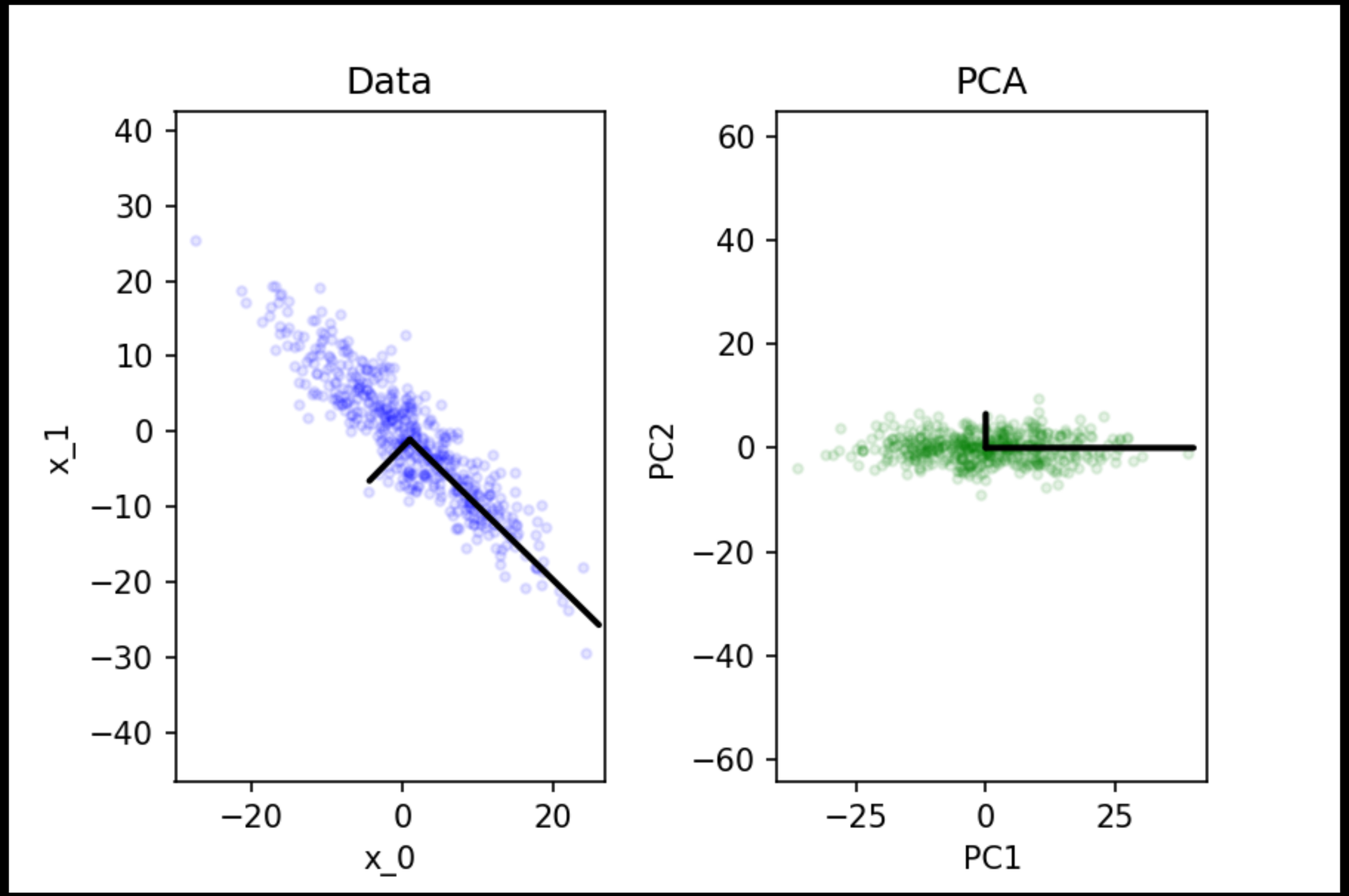
Figuur 8: Casus A: vergelijking traditionele aanpak vs. geïntroduceerde aanpak

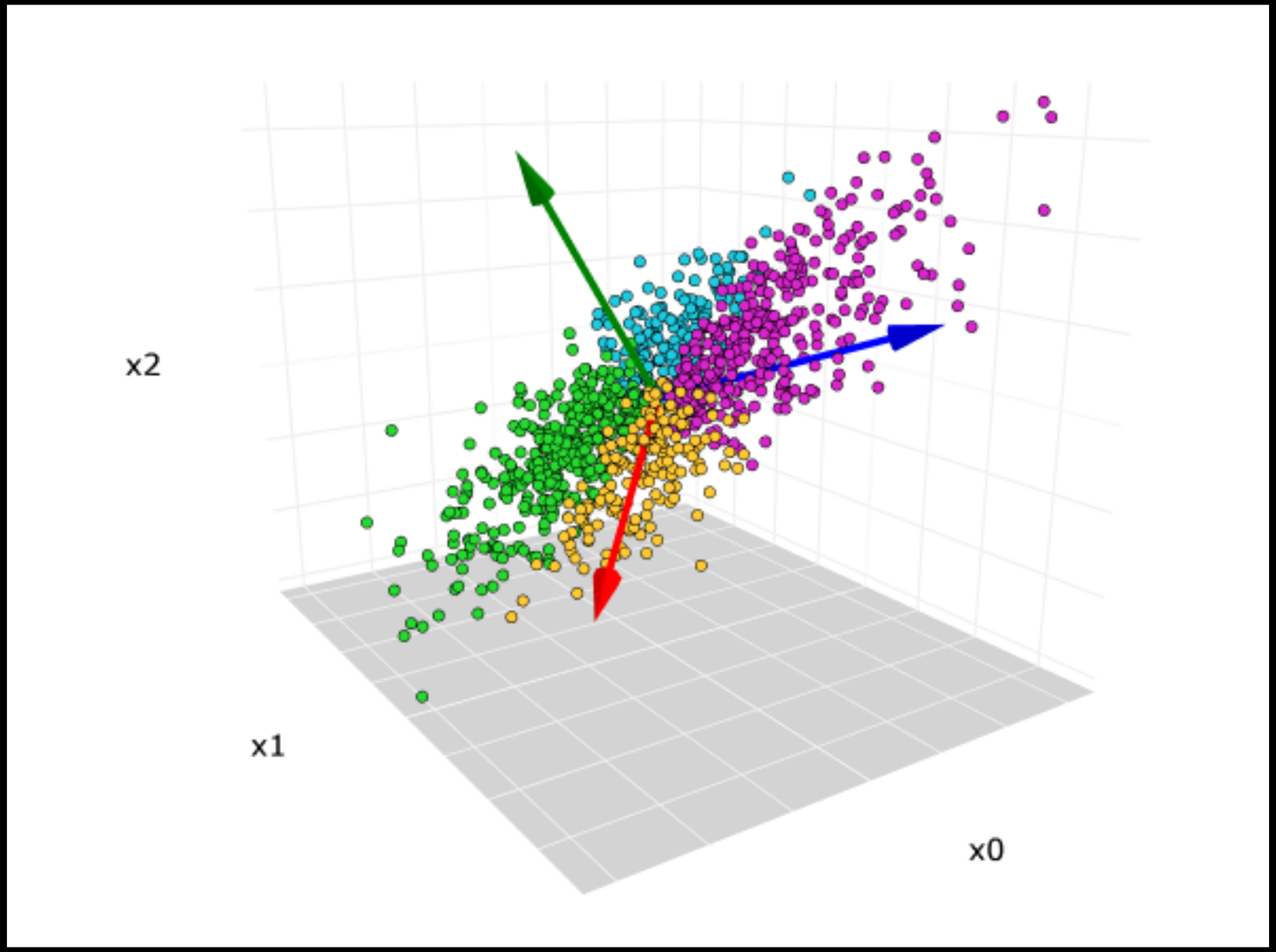
M.M. van Zandwijk, 2024, Detecting potentially relevant e-mails for forensic irregularity investigations using Large Language Models (a FraudBERTa story), masterthesis HU

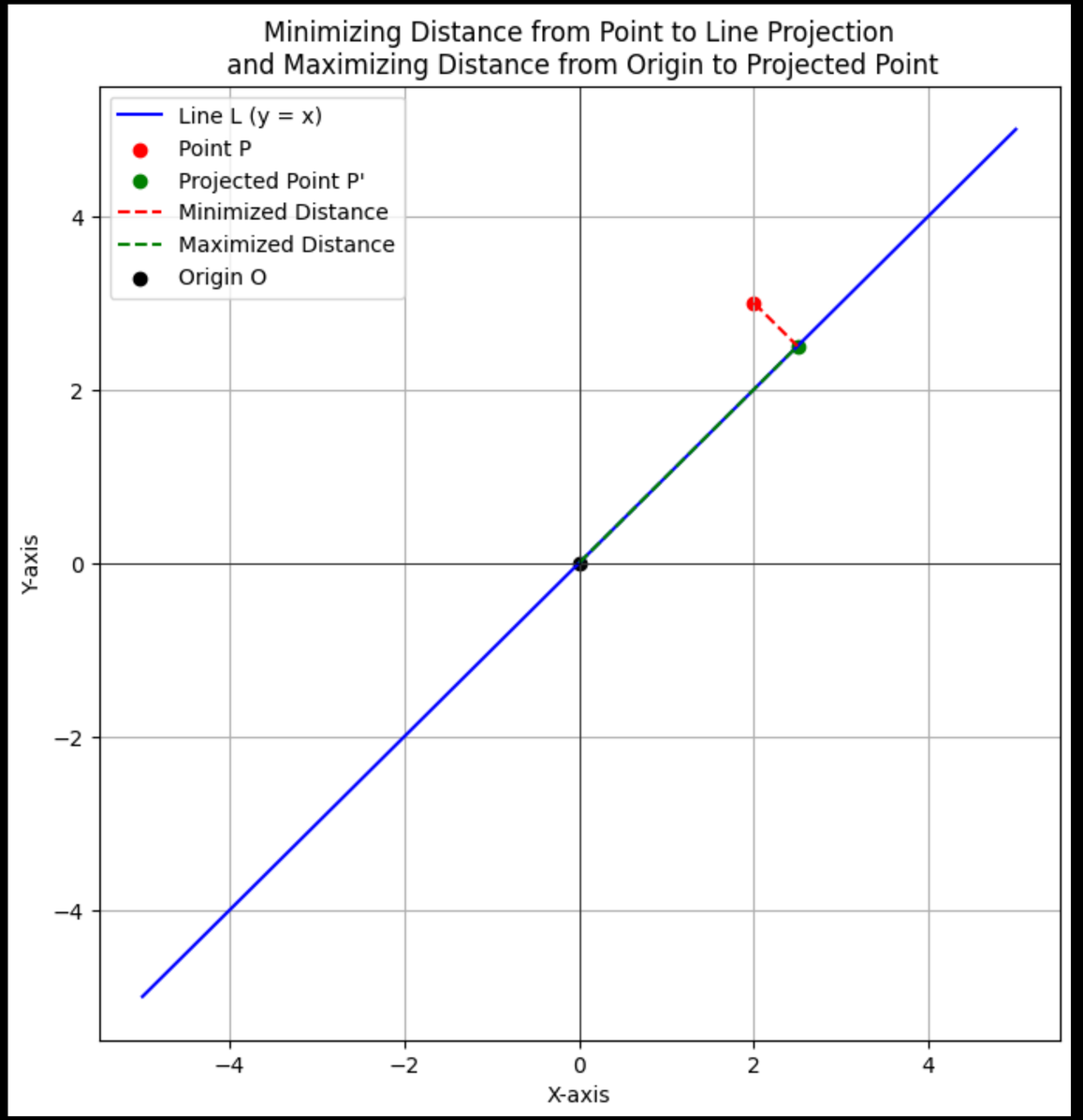
PCA

The curse of dimensionality




Data
PCA
x_1
x_0
PC2
PC1
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
60
40
20
0
−20
−40
−60
−25
0
25


x2
x1
x0


Minimizing Distance from Point to Line Projection
and Maximizing Distance from Origin to Projected Point
Line L (y = x)
Point P
Projected Point P'
Minimized Distance
Maximized Distance
Origin O
Y-axis
X-axis
4
2
0
−2
−4
−4
−2
0
2
4

- Eigenvalue for PC1 $= \frac{\text{SS(distances for PC1)}}{n-1}$
- If the sum of the squared distances of points projected on a vector are larger, that means points are closer to the vector
- What does it mean if an eigenvalue is lower or higher for an eigenvector?

t-SNE

t-SNE

- A linear recombination might not be the best way to visualise complex, non-linear data structures
- tSNE is optimized for visualisation (mapping to $\mathbb{R}^2$ or $\mathbb{R}^3$)

t-SNE

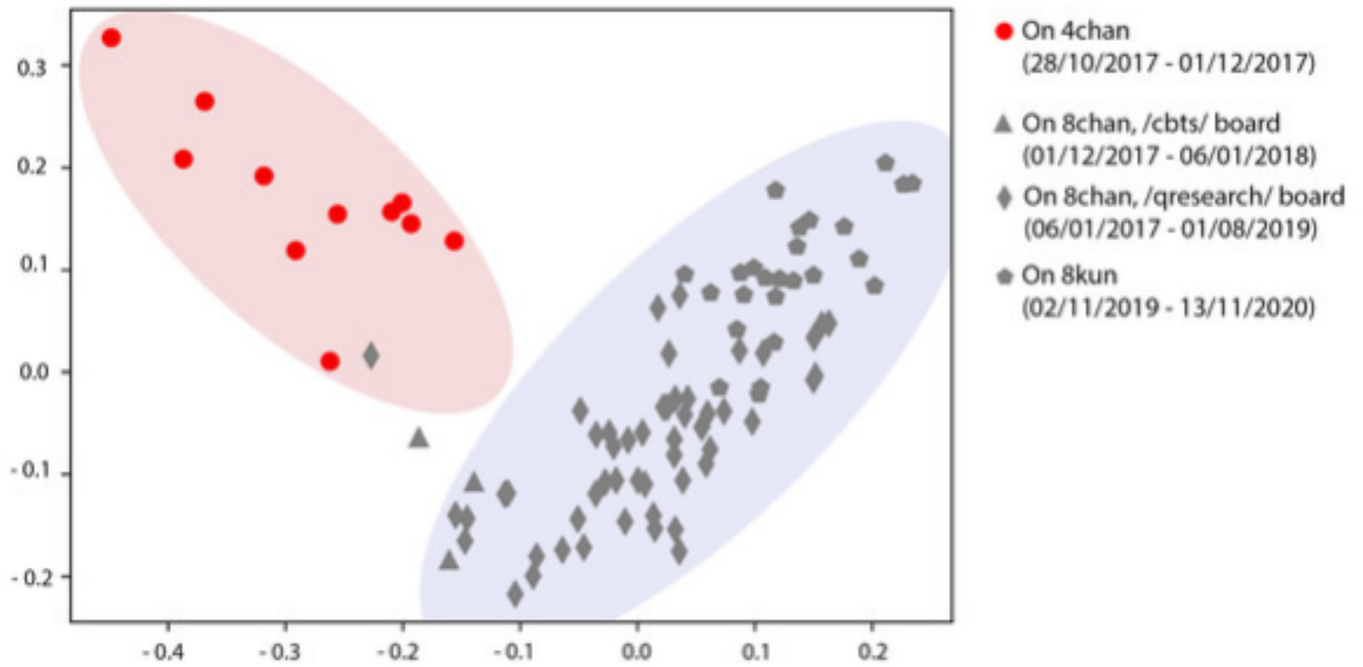
In a nutshell

- A high dimensional dataset $\mathcal{X} = \{x_1, \ldots, x_n \,|\, x \in \mathbb{R}^n\}$
- A low-dimensional mapping $\mathcal{Y} = \{y_1, \ldots, y_n \,|\, y \in \mathbb{R}^d\}$ with $d < n$
- The conditional probability $p_{j|i}$ that x_i would pick x_j as a neighbor
- The conditional probability $q_{j|i}$ that y_i would pick y_j as a neighbor
- A way to minimize the mismatch between P and Q

QAnon Is Two Different People, Shows Machine Learning Analysis from OrphAnalytics

An algorithm-based stylometric approach provides new evidence to identify the authors of QAnon conspiracy theories

NEWS PROVIDED BY
OrphAnalytics →
15 Dec, 2020, 08:38 ET



Two authors are behind QAnon messages, shows machine learning analysis from Swiss company Orphanalytics.

Q's message board history

Oct. 28, 2017
Q's first post on 4chan

Dec. 1, 2017
Q moves to Paul Furber's /cbts/ board on 8chan

Jan. 6, 2018
Q moves to the /qresearch/ board on 8chan, with the help of Ron Watkins

Aug. 10, 2018
Ron Watkins creates a tripcode that locks Q into 8chan

Aug. 1, 2019
8chan goes down, and Q does not post elsewhere

Nov. 2, 2019
8chan comes back as 8kun; Q resumes posting

Source: 4chan; 8chan; 8kun; qresear.ch; qagg.news
Chart: Sawyer Click/Business Insider

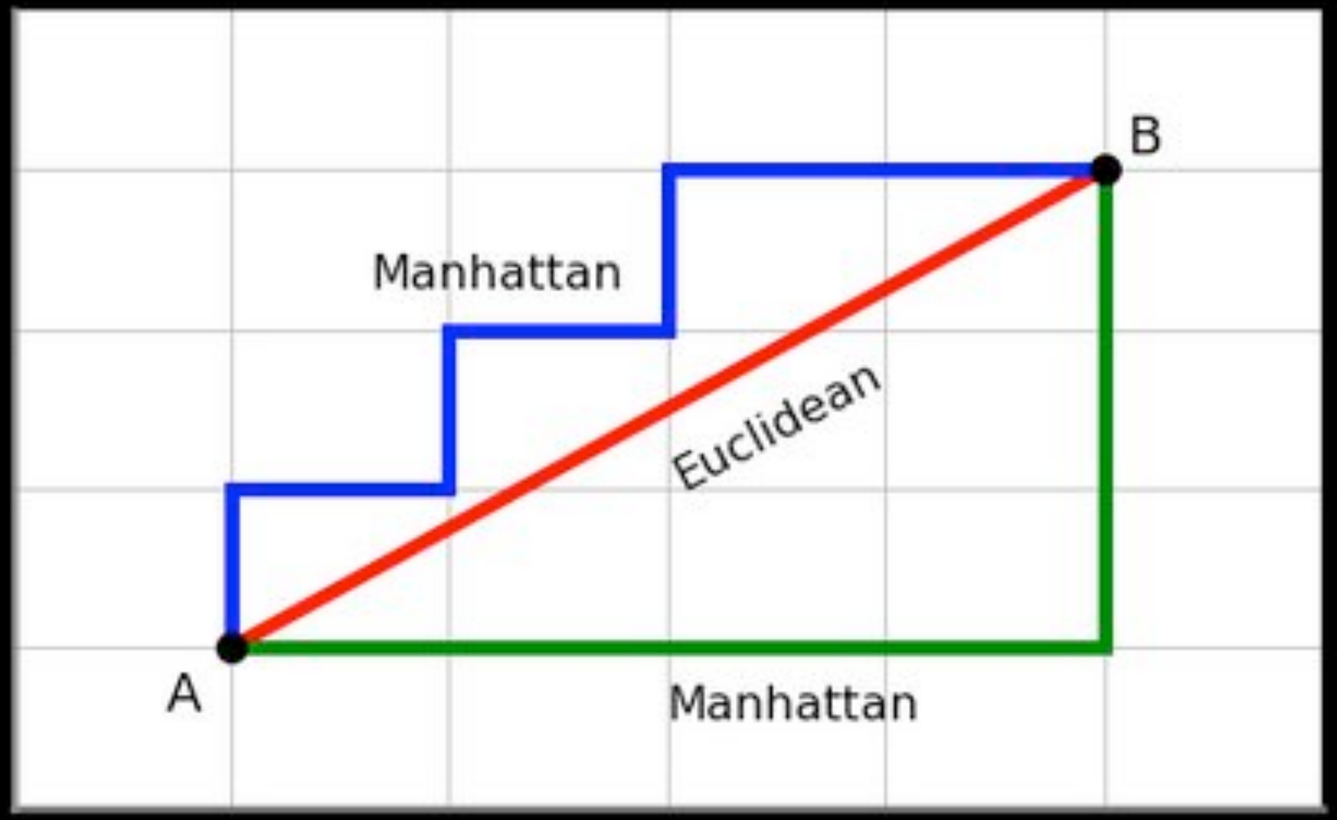
Trigram Vectorizer

- Chunks: "yellow banana", "papagena papaya banana”

Trigram Vectorizer

- Chunks: "yellow banana", "papagena papaya banana”
- Trigrams: ' ba' ' pa' 'a b' 'a p' 'age' 'ana' 'apa' 'aya' 'ban' 'ell' 'ena' ‘gen' 'llo' 'low' 'na ' 'nan' 'ow ' 'pag' 'pap' 'pay' 'w b' 'ya ' ‘yel’
- Countvector: [1, 0, 0, 0, 0, 2, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1]
- $f\colon string \rightarrow \mathbb{N}^{23}$

Manhattan distance



Manhattan distance

- $d\colon X \times X \to \mathbb{R}$
- De Manahttan distance tussen de trigram vectors van “yellow banana” en “papagena papaya banana” is 21

Authorship model

- n strings, each m characters long $\mathbb{T}^{n \times m}$
- Vectorizer: $f\colon \mathbb{T} \to \mathbb{R}^{n \times 23074}$
- Distance: $d\colon \mathbb{R}^{n \times 23074} \to \mathbb{R}^{n \times n}$
- Dim reduction: $PCA\colon \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times 2}$
- Assumption: for the same author, chunks will be closer in $\mathbb{R}^2$

```
def __call__(
    self, text: list[str], k: int, labels: list, batch: bool, method: str = "PCA"
) -> None:
    if batch:
        text = self.batch_seq(text, k)
    distance = self.fit(text)
    X = self.reduce_dims(distance, method)
    self.plot(X, labels)
```

```
def fit(self, parts: list[str]) -> np.ndarray:
    X = self.vectorizer.fit_transform(parts)
    X = np.asarray(X.todense())
    distance = manhattan_distances(X, X)
    return distance
```