Dimensionality Reduction

Raoul Grouls, 10 November 2023

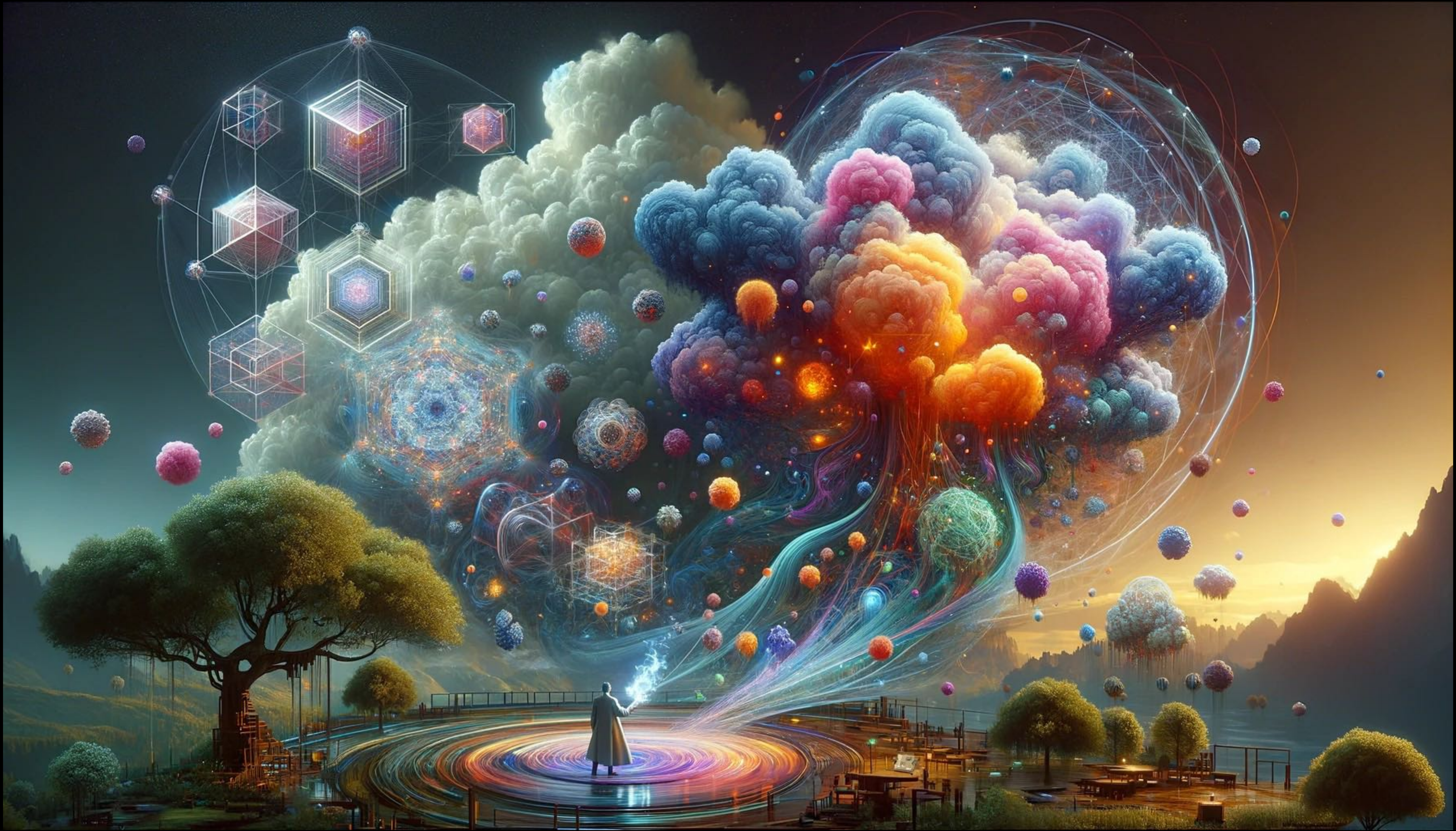
Motivation for embedding data in high dimensional vector spaces as a design pattern

Mapping to and fro

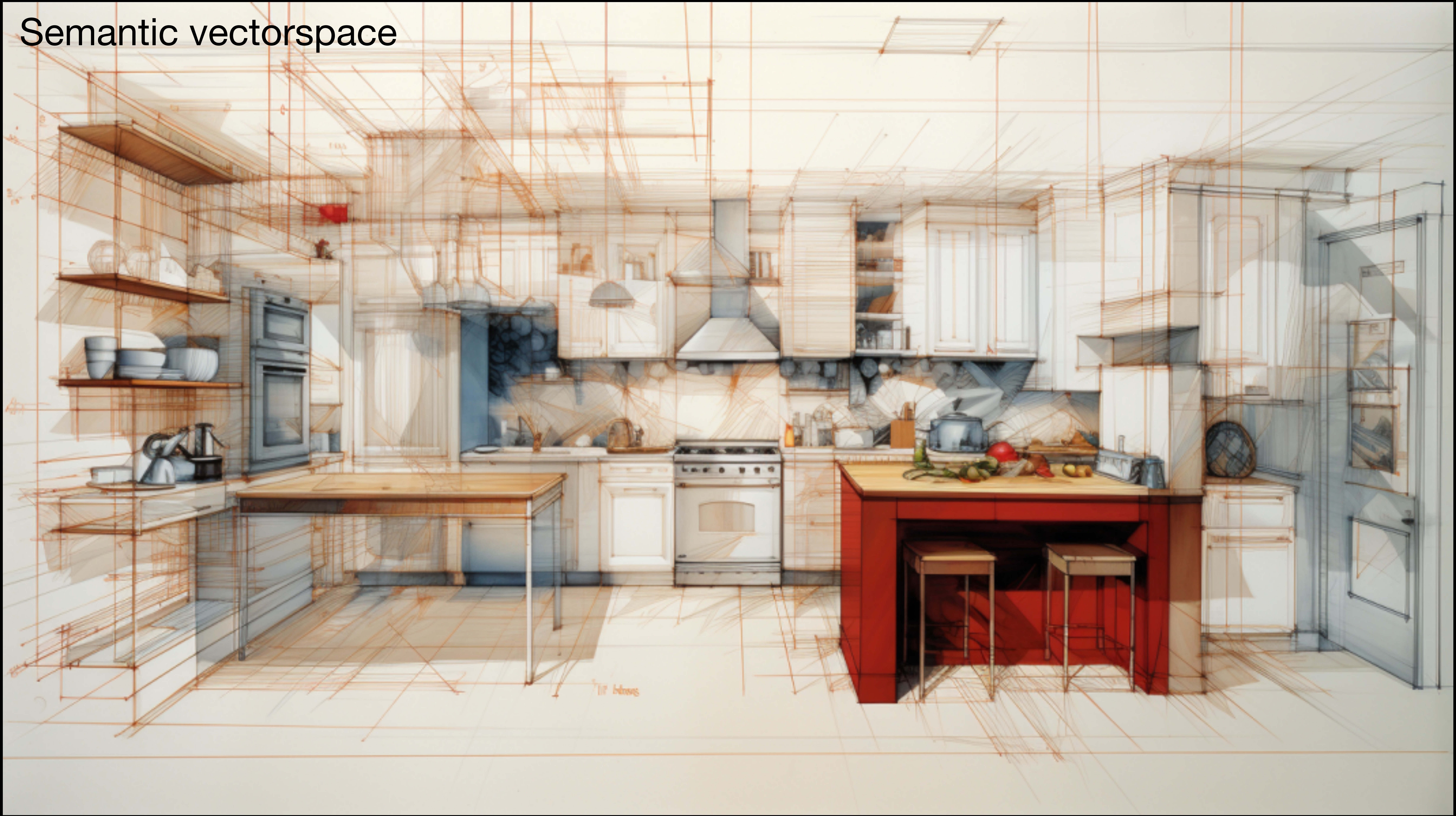
First, map data to a high dimensional space Z.

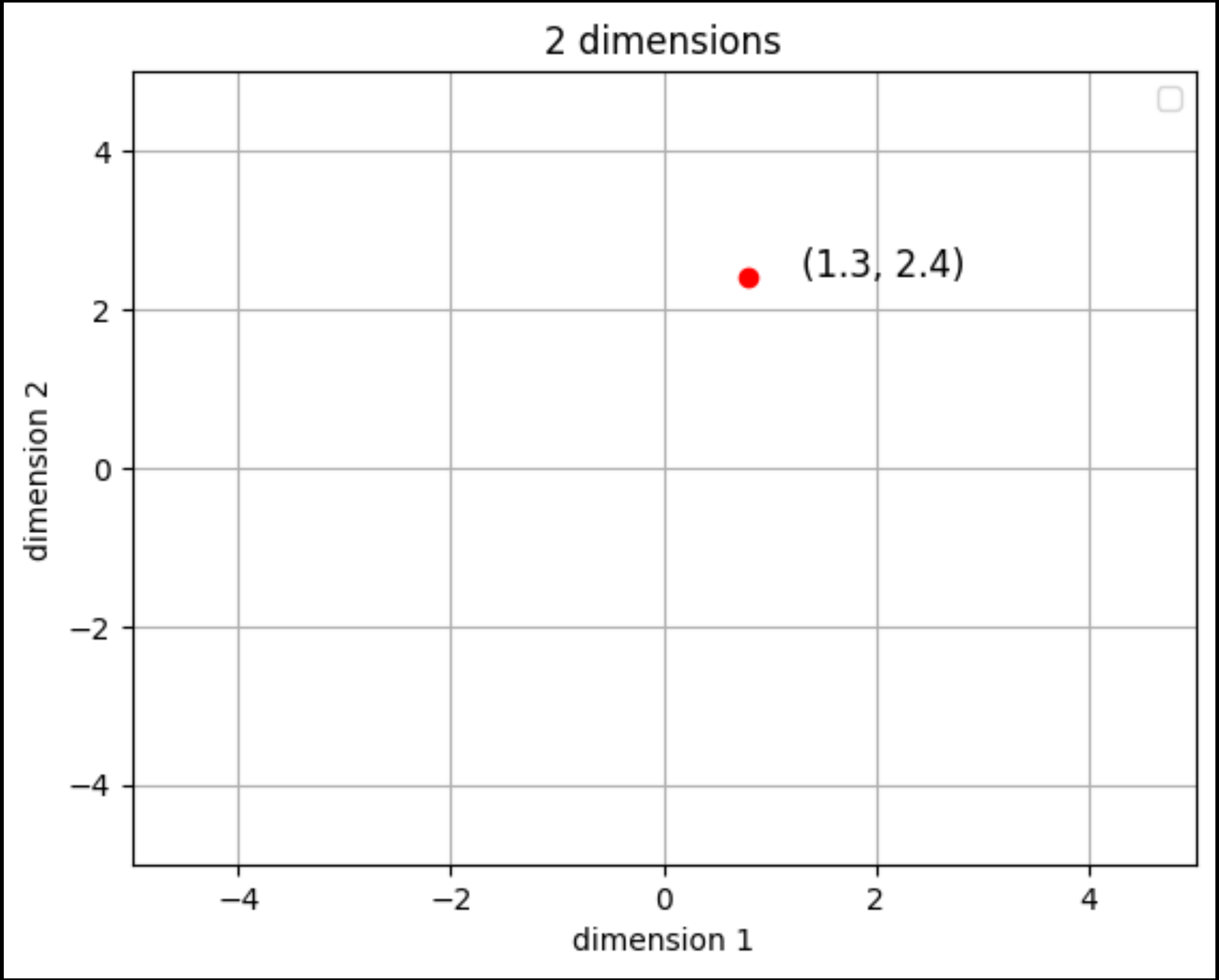
Do some transformations, and map it back to a low dimensional manifold.

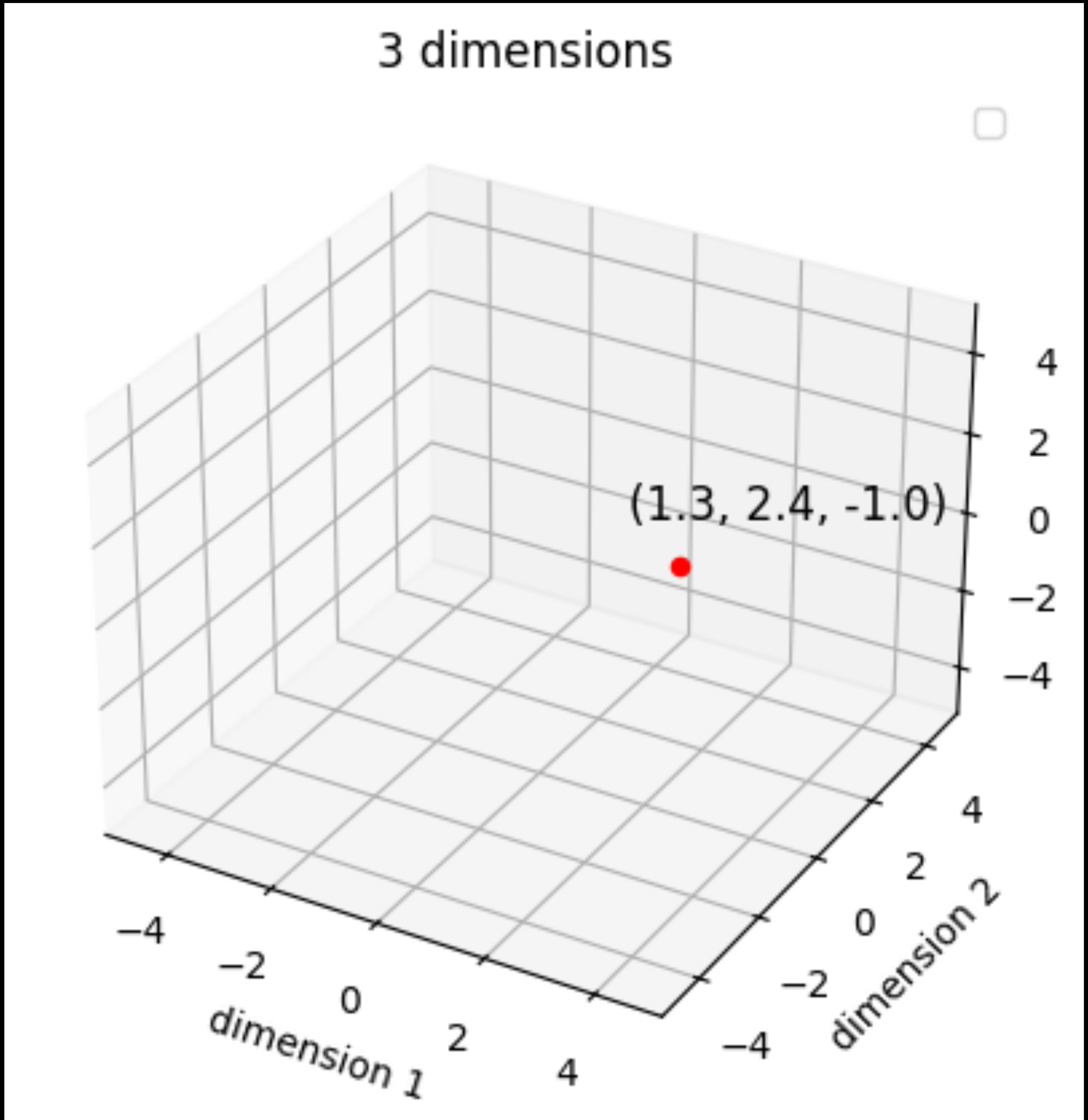
- $f: X \to Z$, with $Z \in \mathbb{R}^d$
- $g: Z \to M$, with $M \in \mathbb{R}^2$



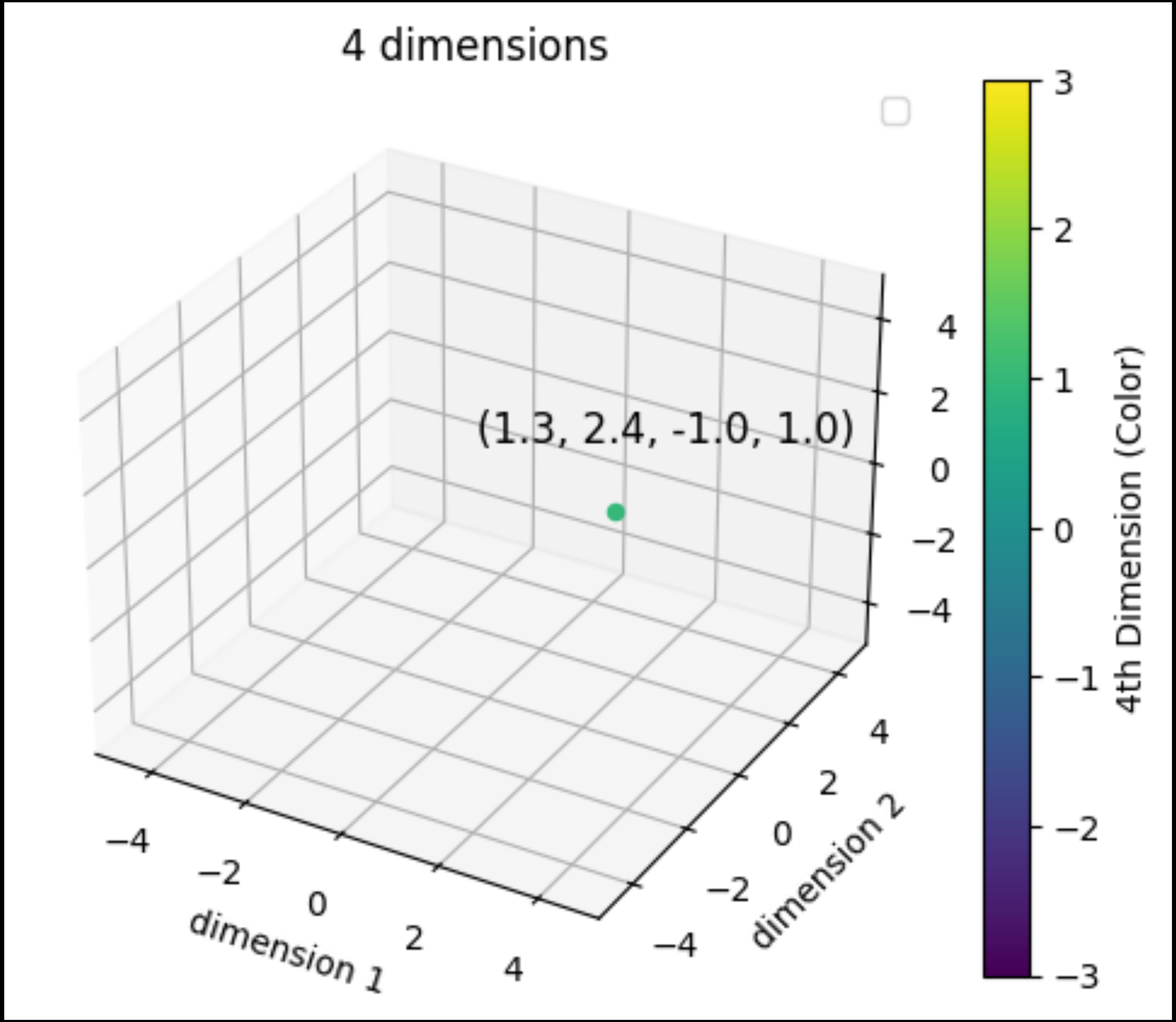
Semantic vectorspace




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



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


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}

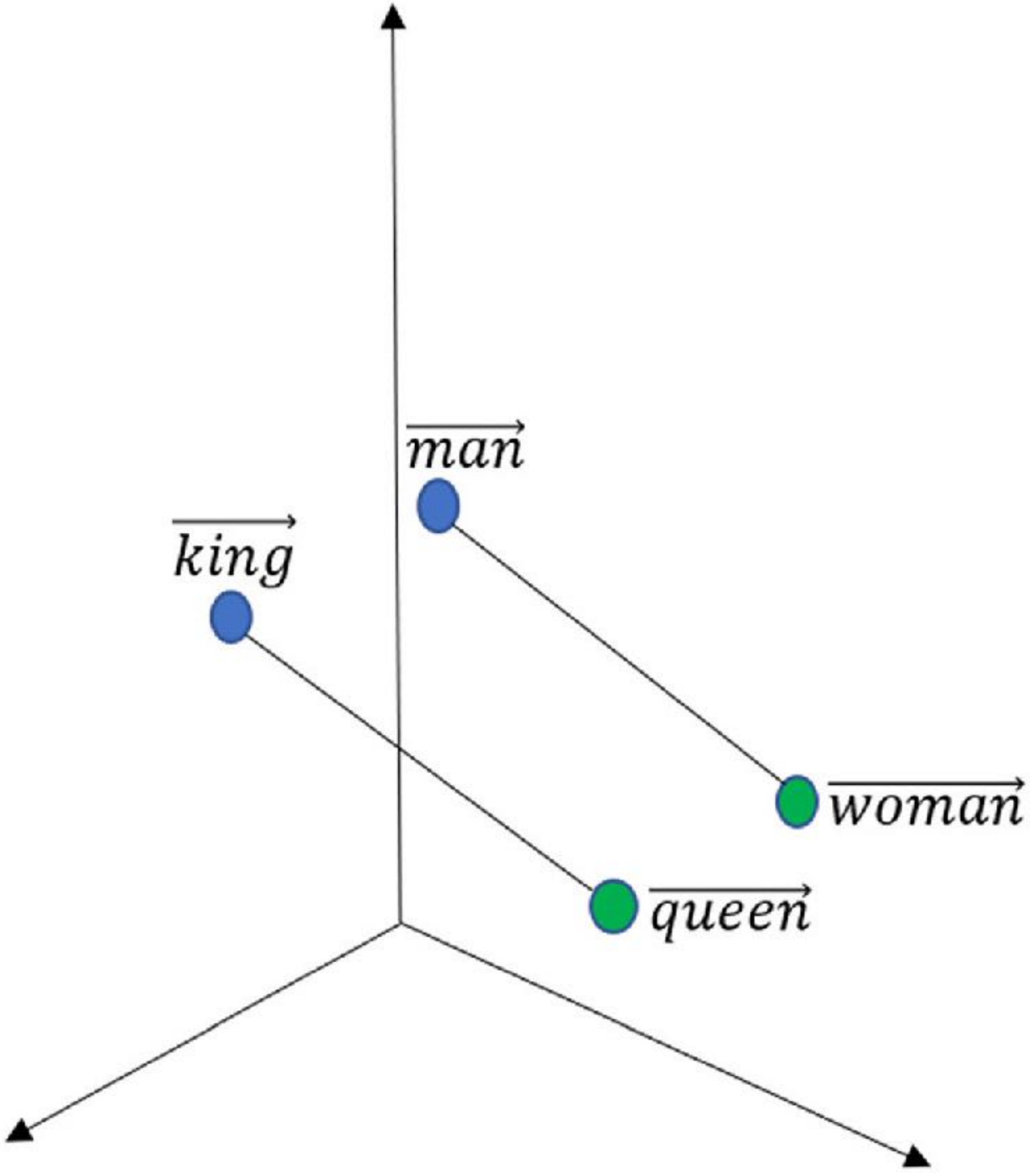
How big is 10^{68} (shuffle a stack of cards)

1. Every 10^9 years (3.2×10^{16} sec), 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 ($1.5 \times 10^{11} m$)
5. This gives about 2×10^{63}, so we still need to repeat this about 5×10^4 times to get there….

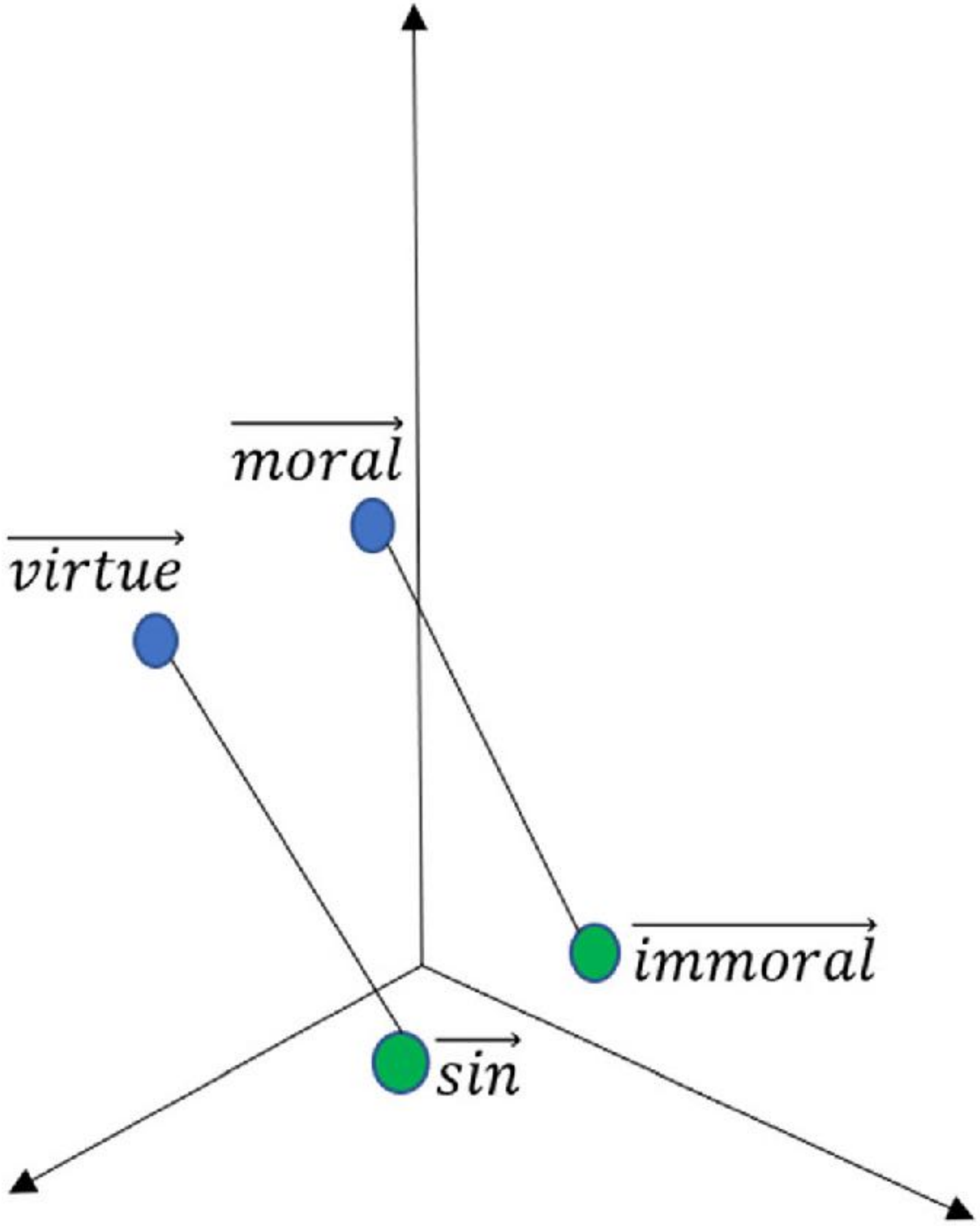
Semantic vectors

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

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



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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Ik	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

What is a vectorspace?

Let V be a set, let F be a field equipped with addition and multiplication

We define binary operations

- “+” on V, denoted $V \times V \to V$,
- “.” on $F \times V$ denoted $F \times V \to V$

A **<u>vectorspace</u>** satisfies for
$\forall c, d \in F, \forall u, v, w \in V$ the following:

Closure under addition: $u + v \in V$

Closure under multiplication: $c \cdot v \in V$

What is a vectorspace?

A **vectorspace** satisfies for
$\forall c, d \in F, \forall u, v, w \in V$ the following:

Addition (+):

1. Commutative: $u + v = v + u$
2. Associative: $(u + v) + w = u + (v + w)$
3. Identity: $u + 0 = 0 + u = u$
4. Inverse: There exists an element (-1) such that: $u + (-1)u = 0$

Multiplication (.):

1. Compatibility: $(cd)u = c(du)$
2. Distributivity: $c(u + v) = cu + cv$
3. Distributivity: $(c + d)u = cu + du$
4. Identity: $1 \cdot u = u$

What is a metric?

For $\forall x, y, z$:

1. Non-negativity: $d(x, y) \leq 0$
2. Identity of indiscernibles: e$d(x, y) = 0$ if and only if $x = y$.
3. Symmetry: $d(x, y) = d(y, x)$
4. Triangle inequality: d(x, y) + d(y, z) ≥ d(x, z)

Motivation for dimensionality reduction

Manifold hypothesis

- although high-dimensional data (like images, text, and sound) might appear complex and unwieldy,
- they actually lie on or near a much lower-dimensional manifold.

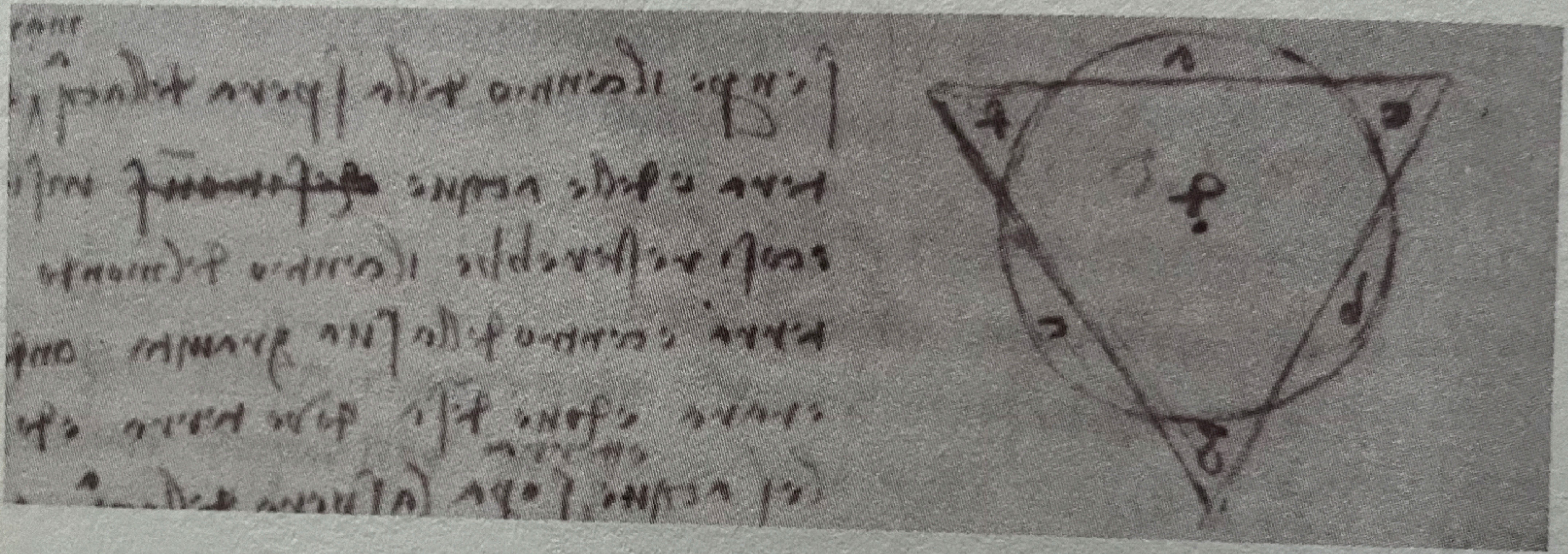
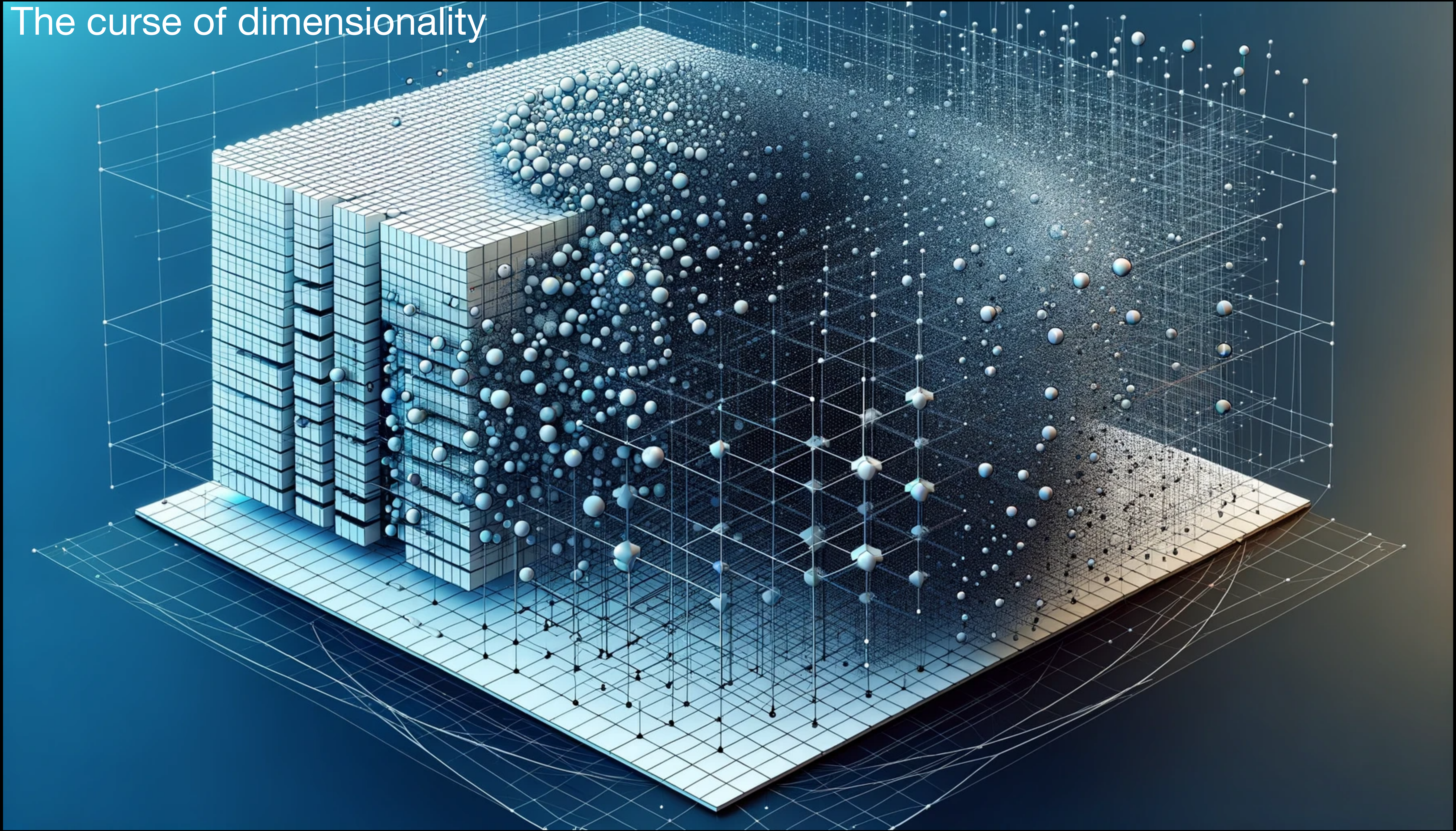
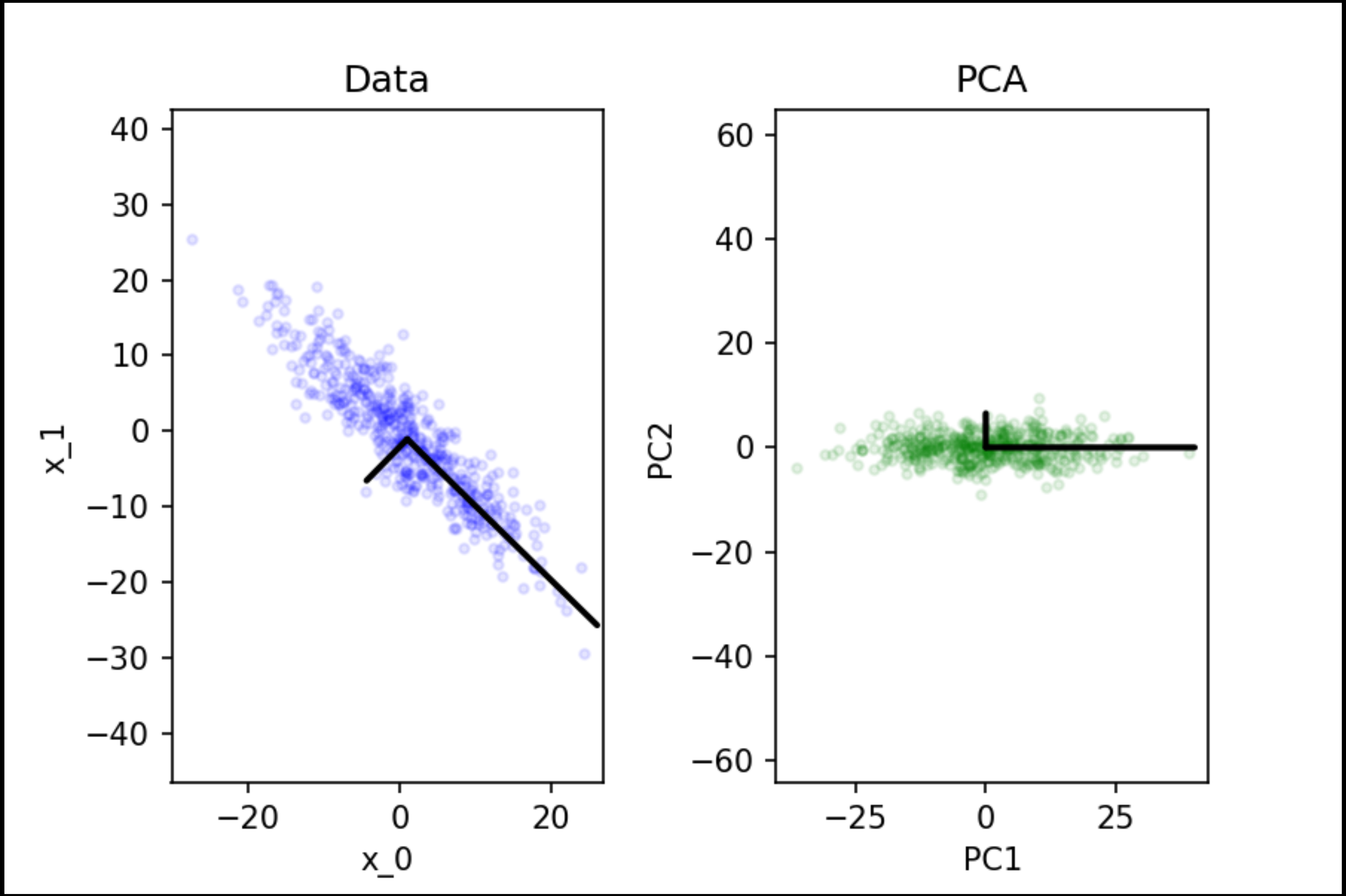


FIG. 2-7. Geometric model of the Earth.
Codex Leicester, folio 35v (detail).

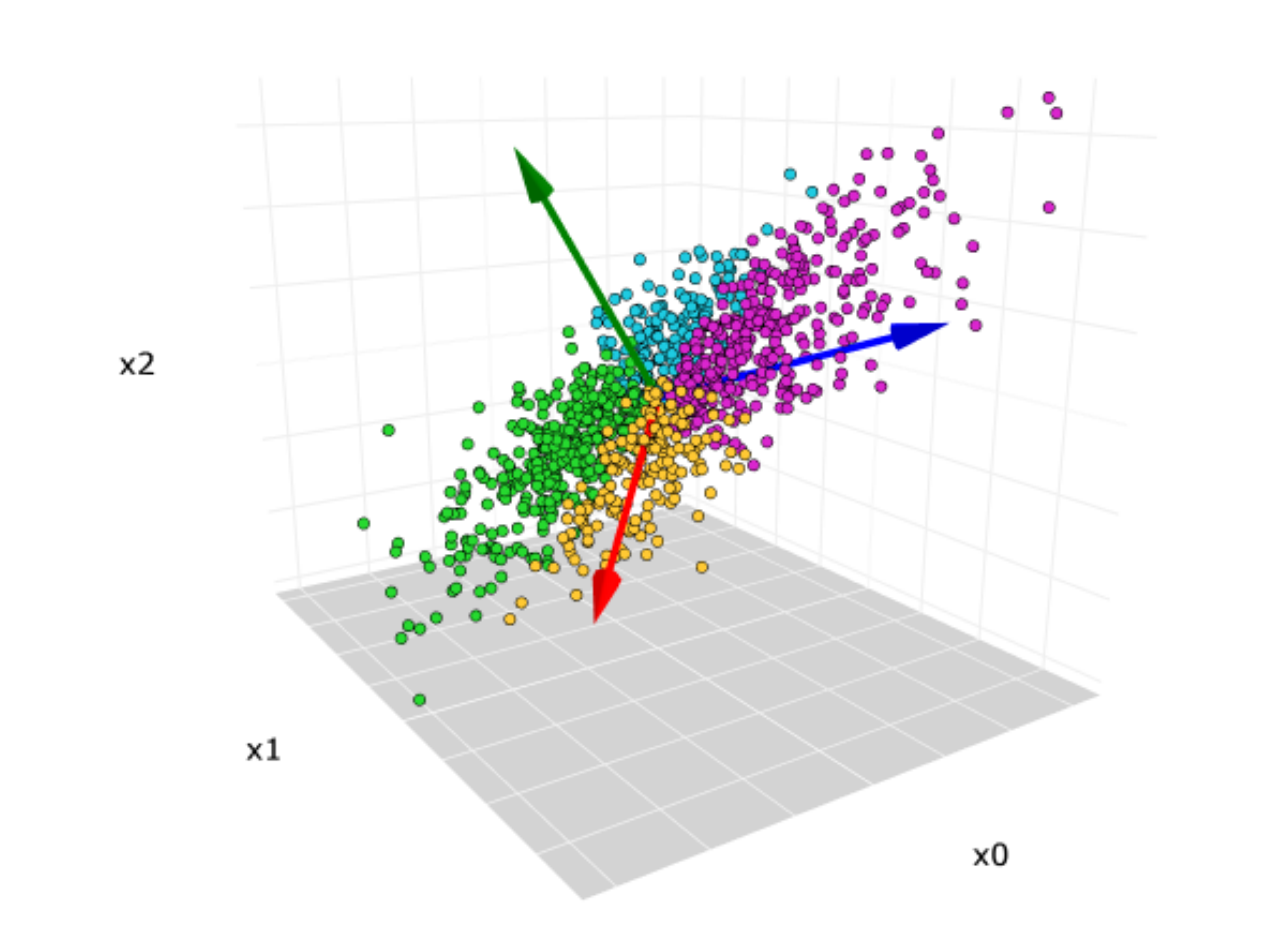
PCA

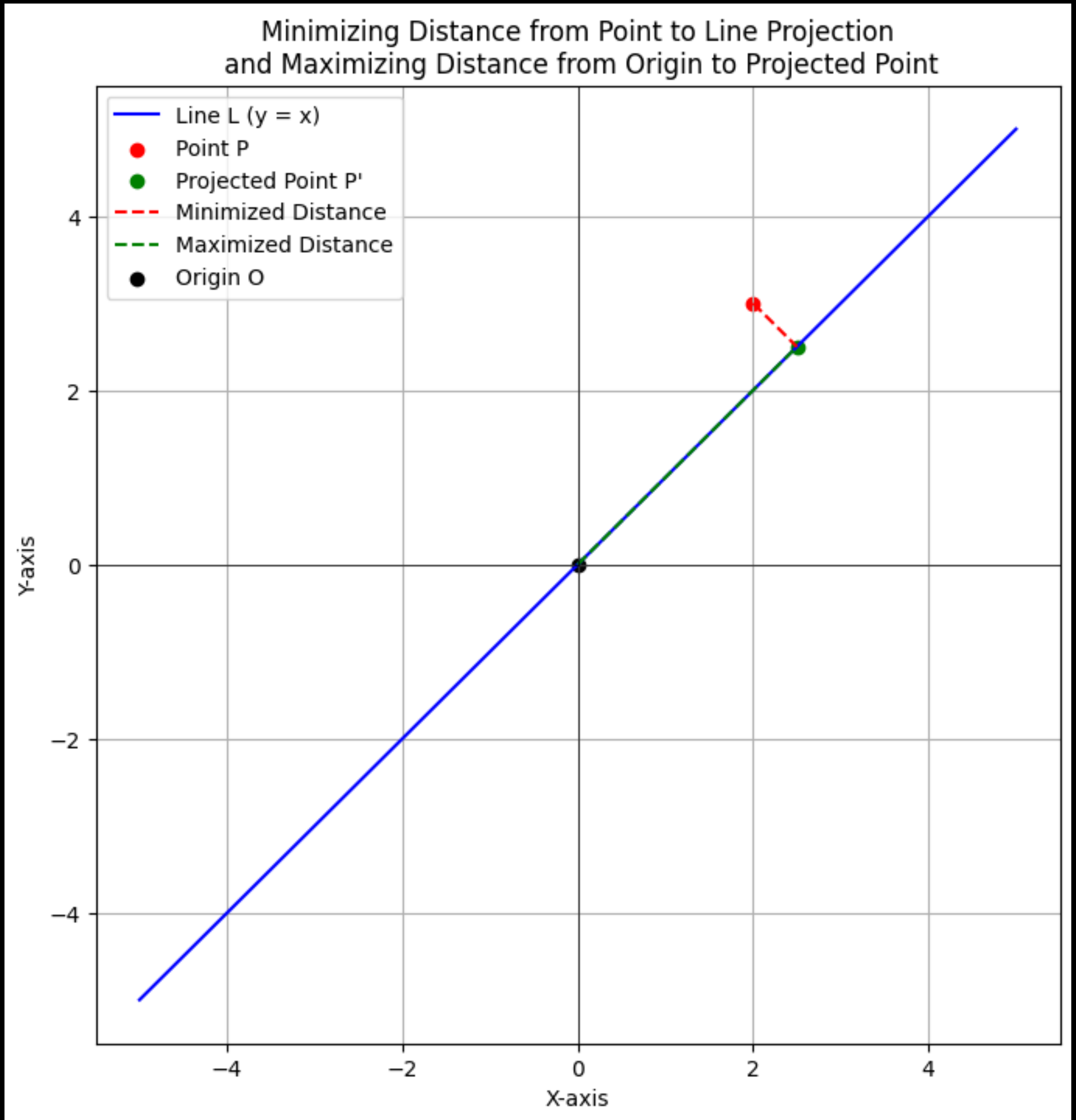
The curse of dimensionality





Data
x_1
x_0
PCA
PC2
PC1


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
4
2
0
−2
−4
Y-axis
−4
−2
0
2
4
X-axis

- Eigenvalue for PC1 = $\dfrac{\text{SS(distances for PC1)}}{n-1}$
- 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

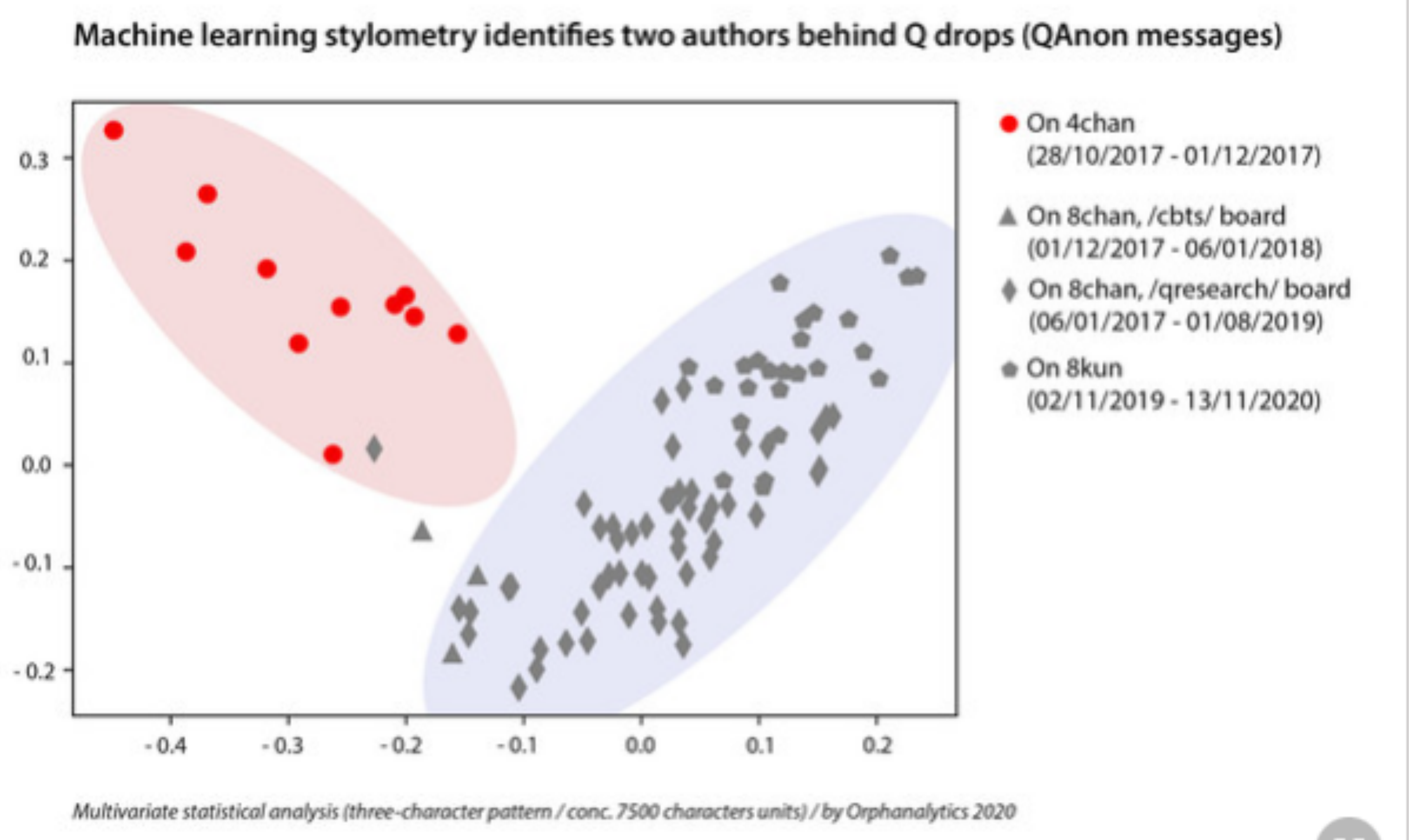
An algorithm-based stylometric approach provides new evidence to identify the authors of QAnon conspiracy theories

---

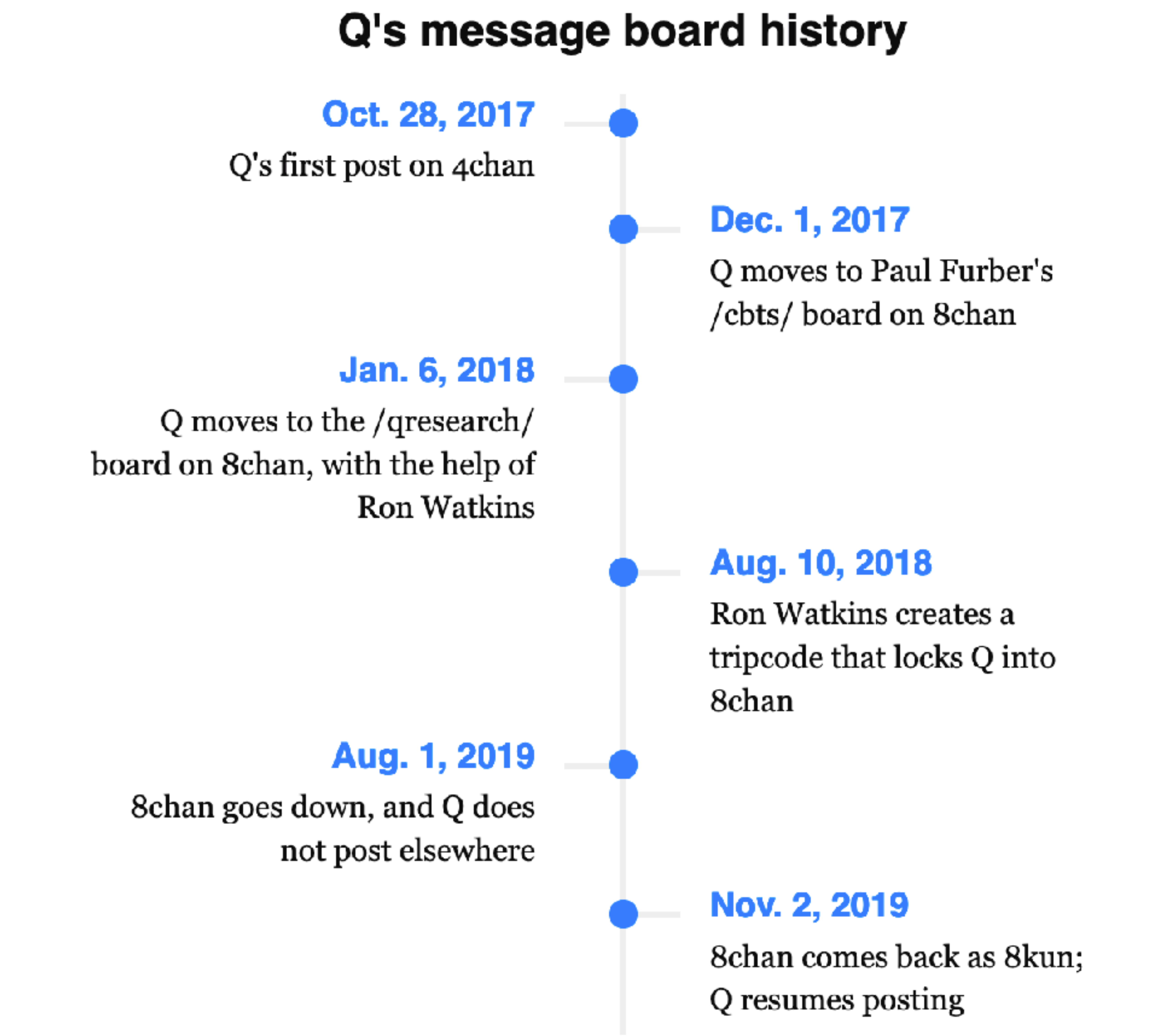
NEWS PROVIDED BY

OrphAnalytics →

15 Dec, 2020, 08:38 ET



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



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

```python
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)
```

```python
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
```