Dimensionality Reduction

Raoul Grouls, 10 November 2023

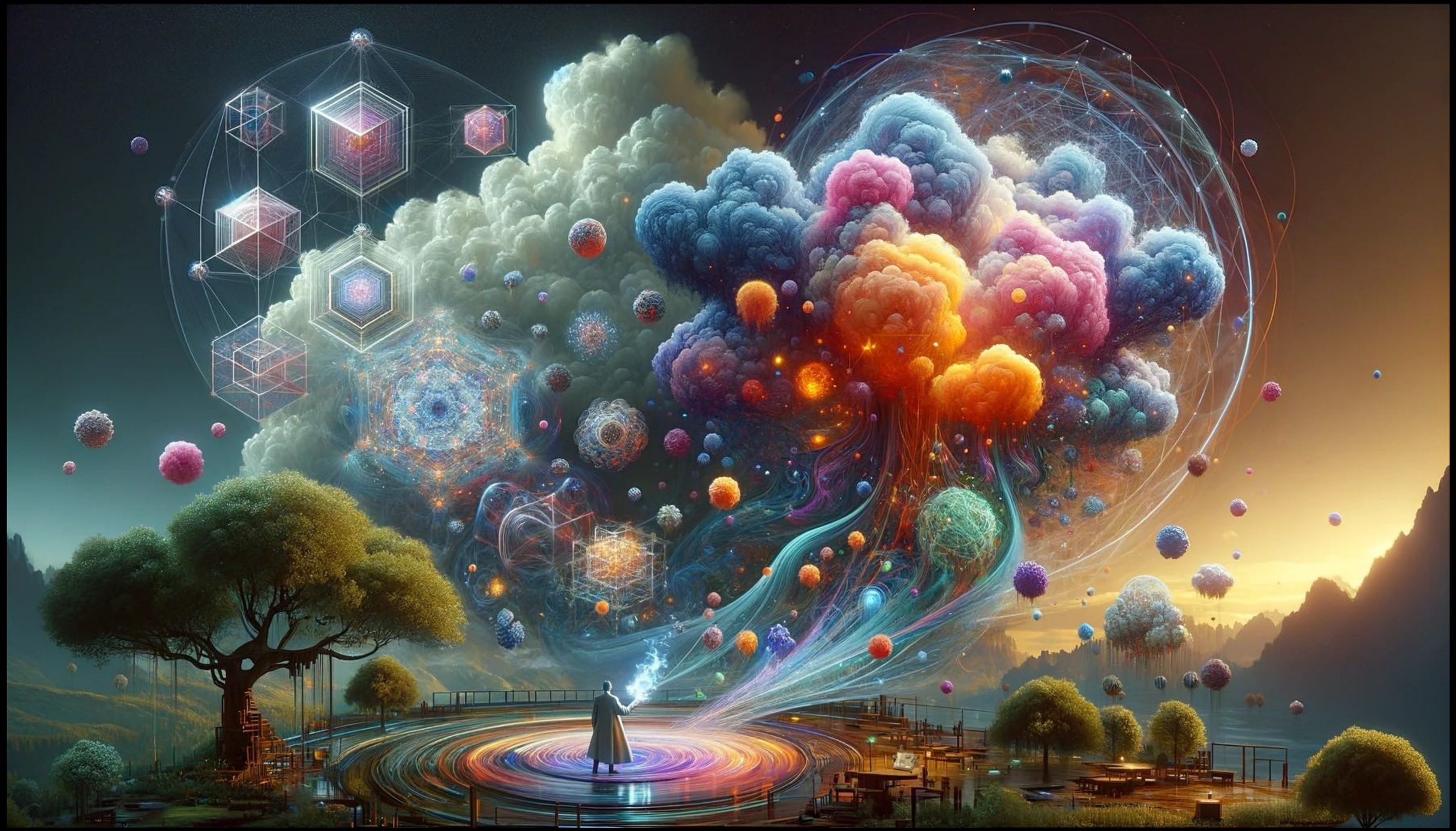
Motivation for embedding data in high dimensional vector spaces as a design pattern

Mapping to and fro

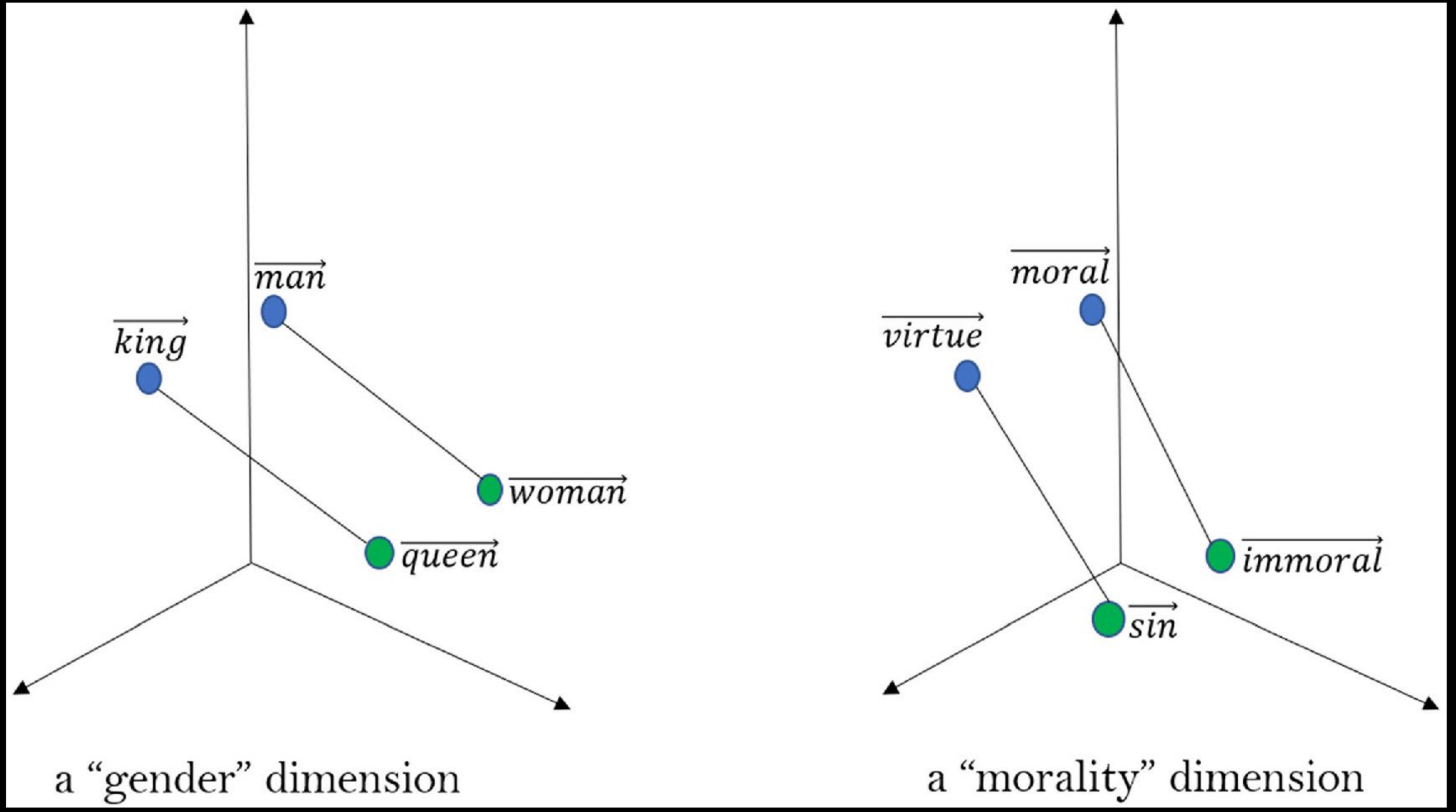
First, map data to a high dimensional space Z.

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

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



Semantic vectorspace


$\overrightarrow{man}$
$\overrightarrow{king}$
$\overrightarrow{woman}$
$\overrightarrow{queen}$
a “gender” dimension
$\overrightarrow{moral}$
$\overrightarrow{virtue}$
$\overrightarrow{immoral}$
$\overrightarrow{sin}$
a “morality” dimension

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)

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 **vectorspace** 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: $1 \cdot u = u$

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.

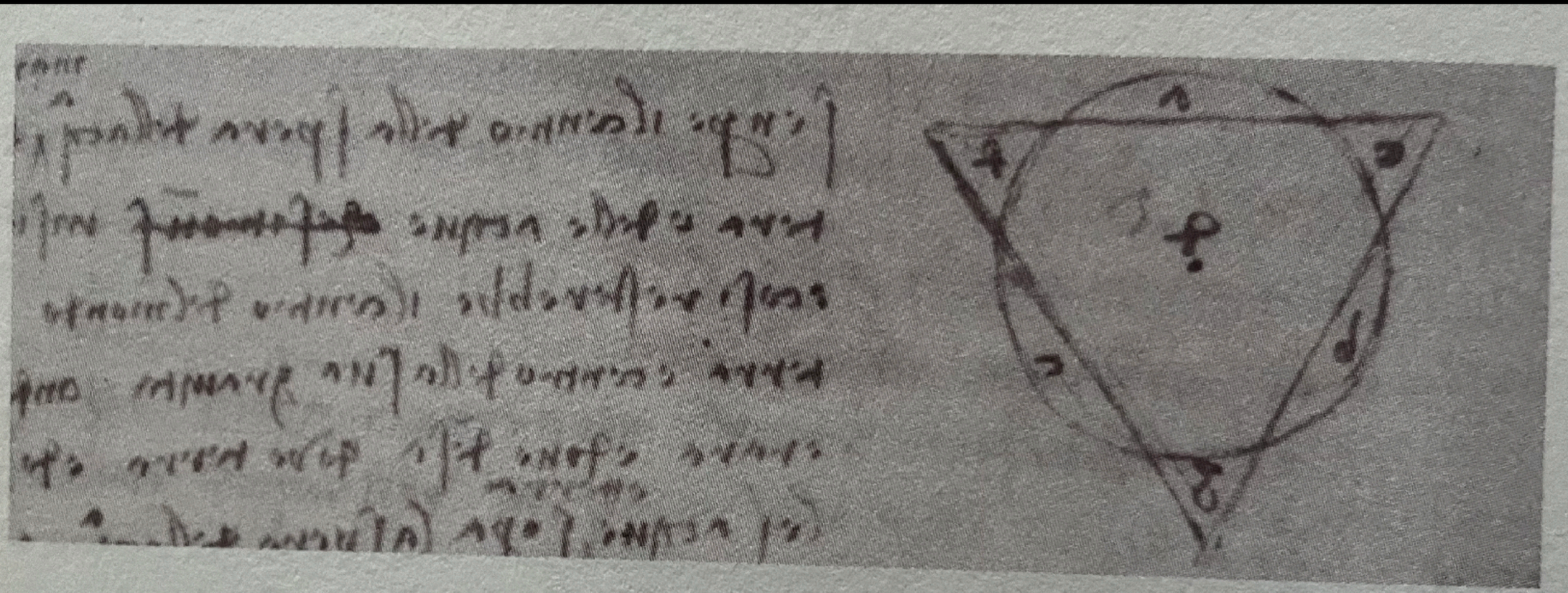
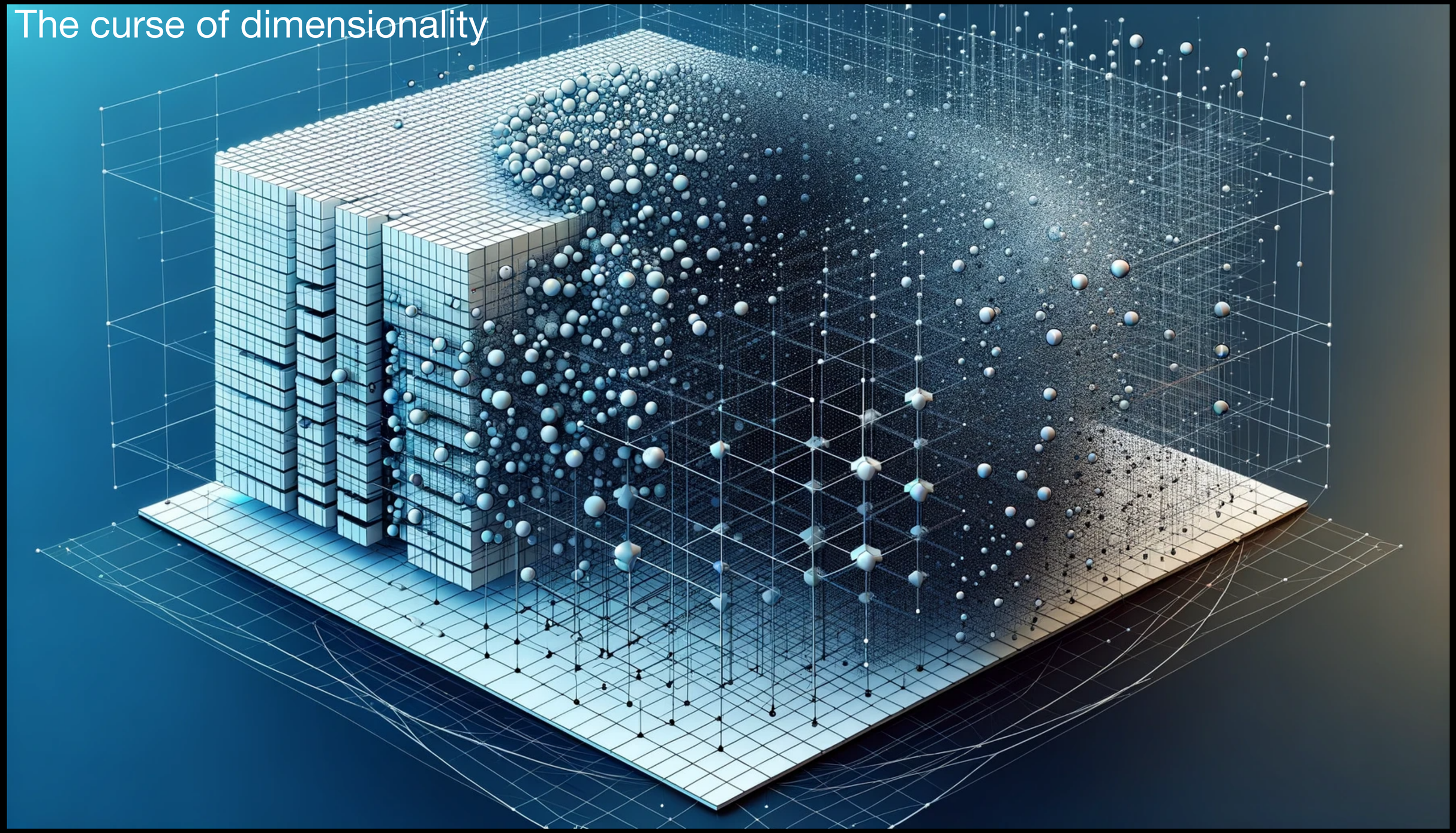
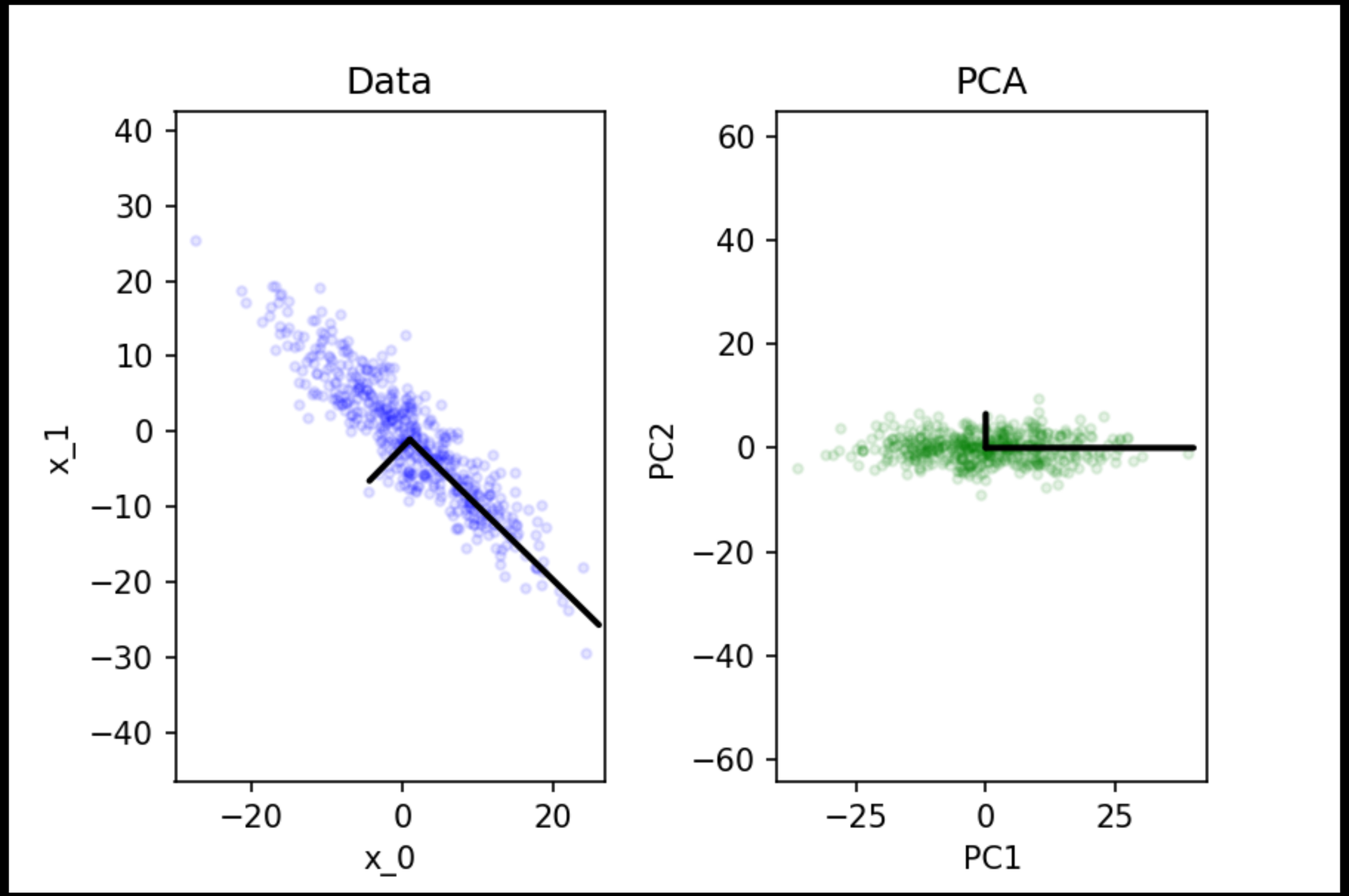


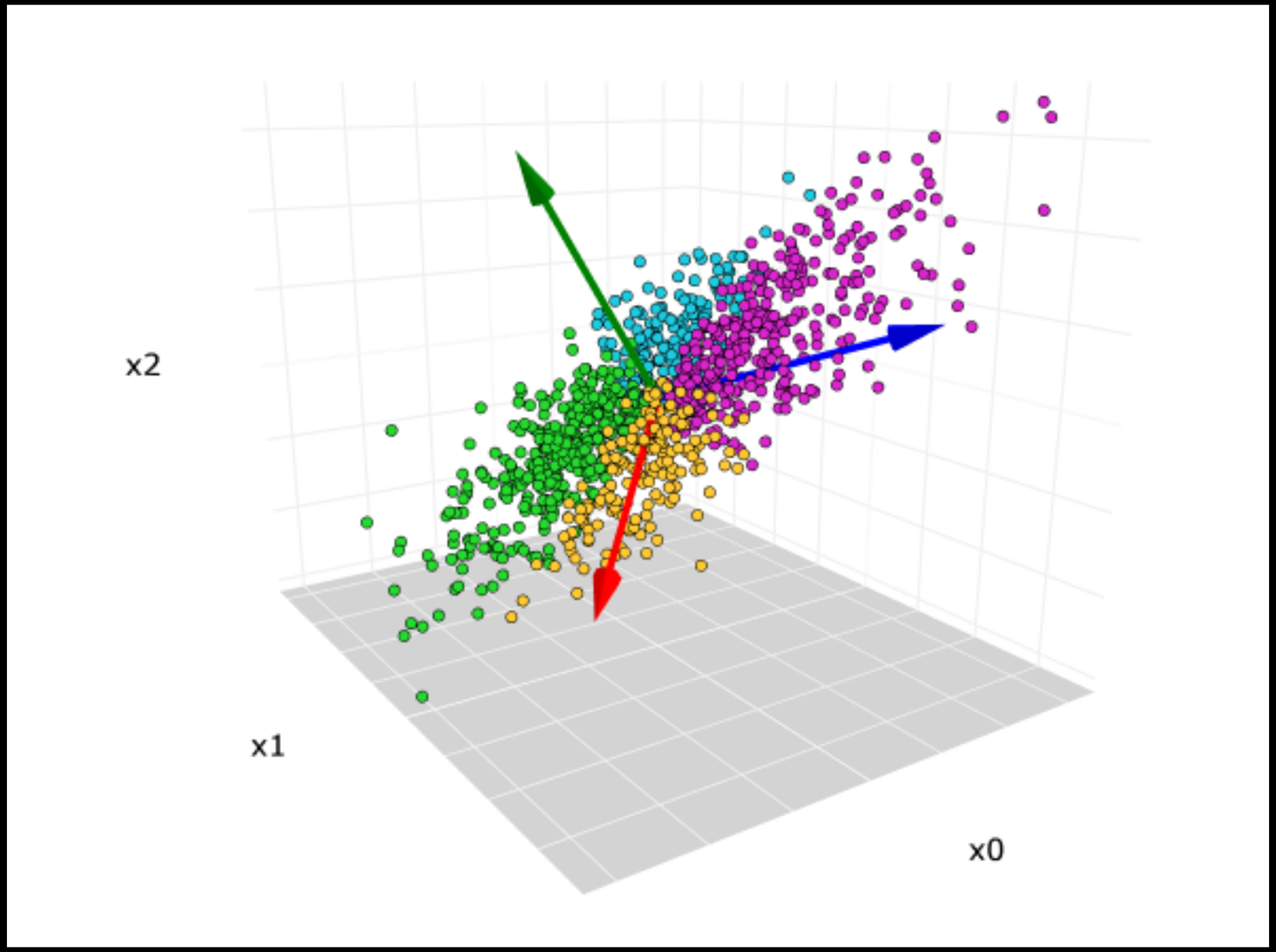
FIG. 2-7. Geometric model of the Earth.
Codex Leicester, folio 35v (detail).

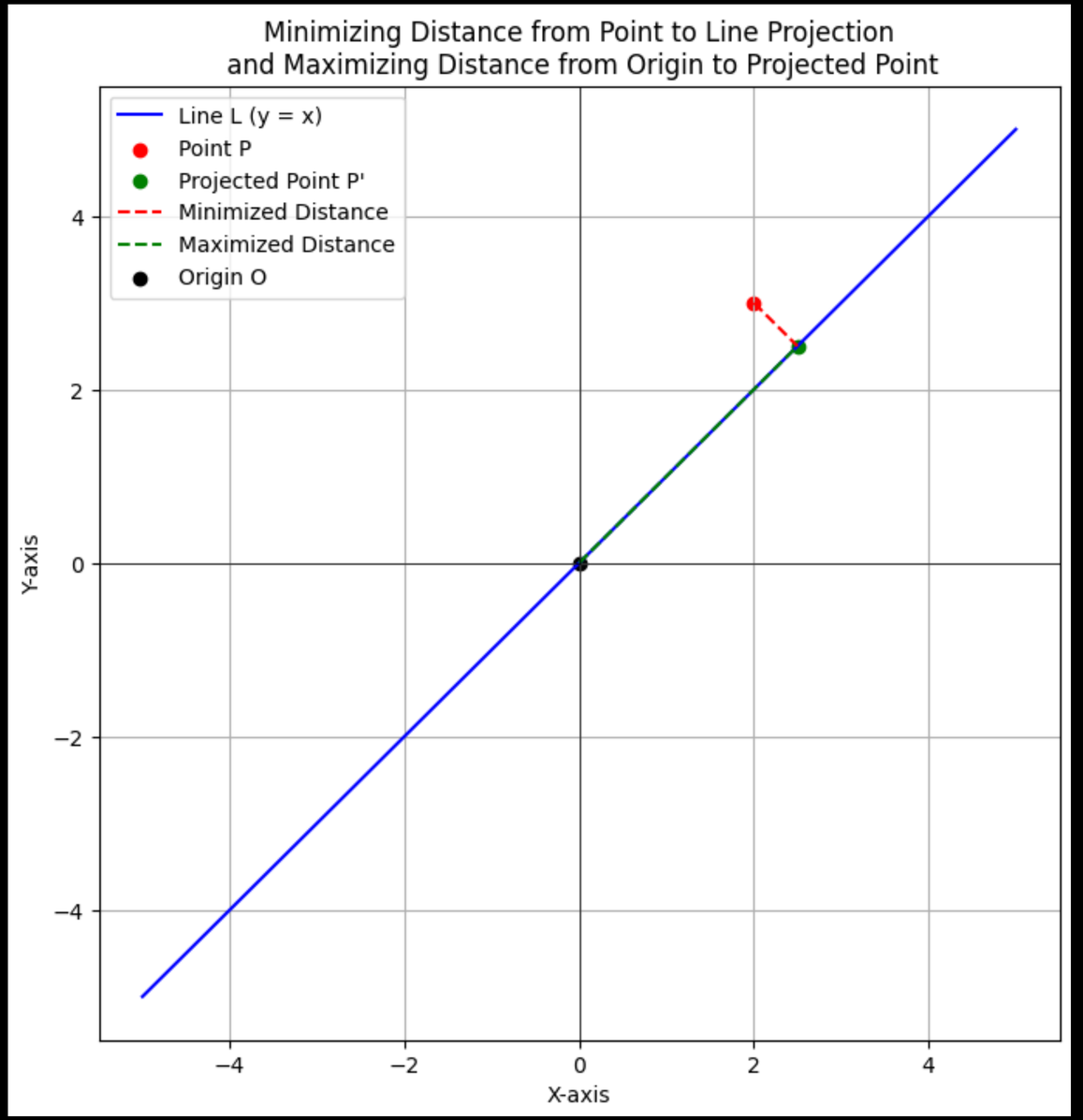
PCA

The curse of dimensionality




Data
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

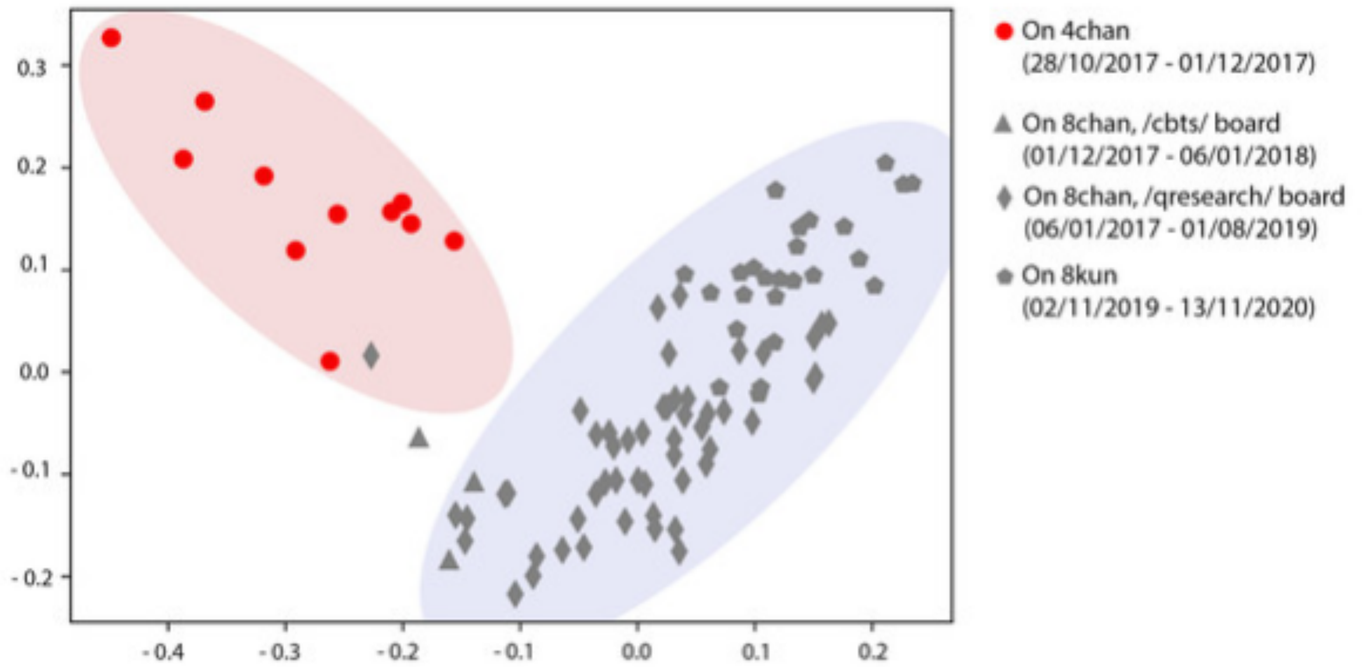
In a nutshell

- 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

```
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
```