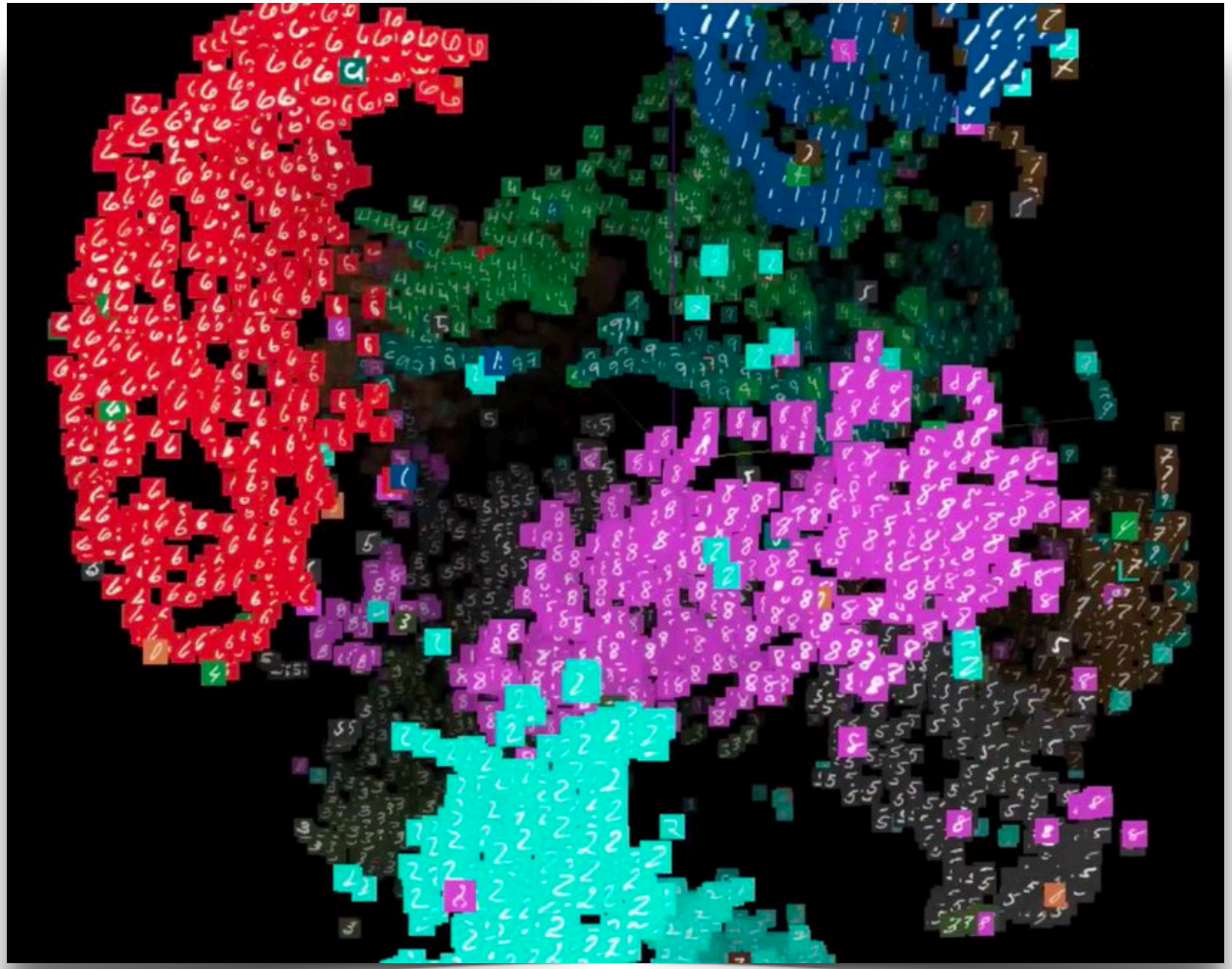
dimensionality reduction for dataviz



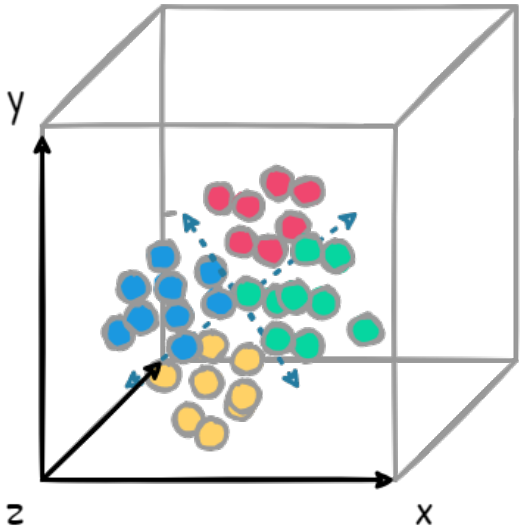
introduction

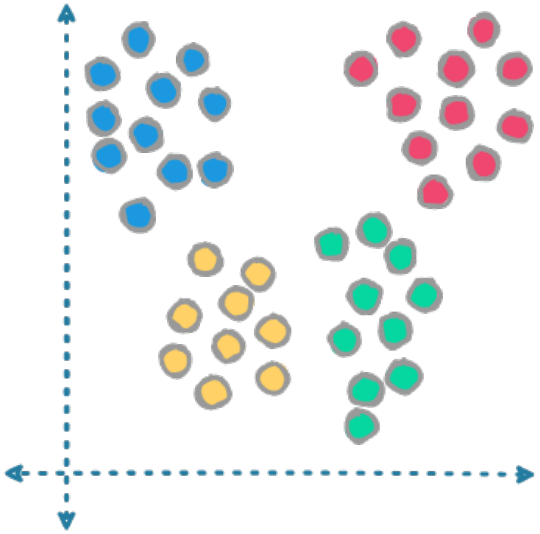
Dimensionality reduction

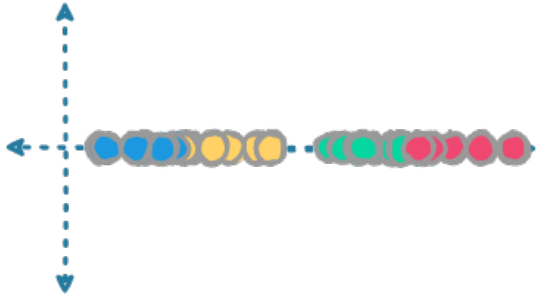
A definition

Dimensionality reduction is a method for representing a given dataset using a lower number of features (that is, dimensions) while still capturing the original data's meaningful properties.

Lih-Yuan Deng, Max Garzon, and Nirman Kumar,
Dimensionality Reduction in Data Science, Springer, 2022.



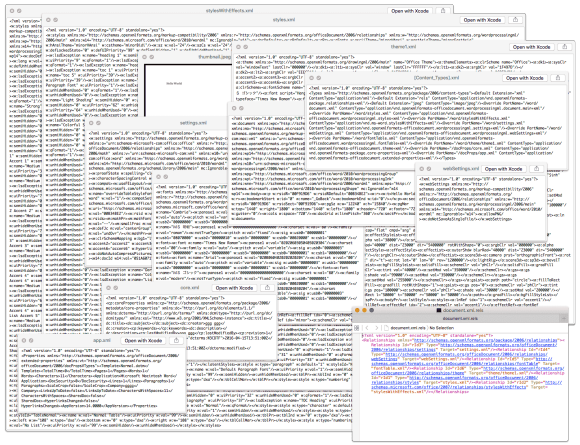




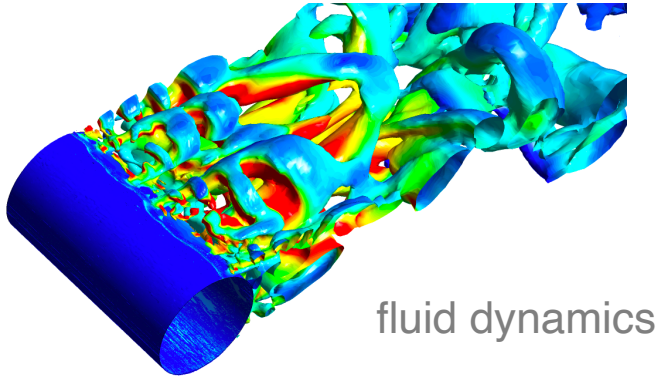
Dimensionality reduction

Why?

face recognition

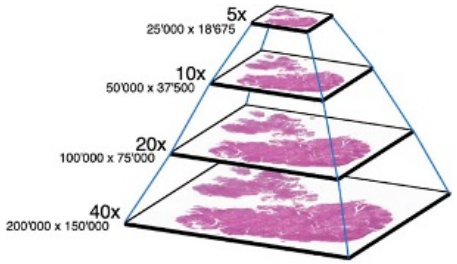


text processing



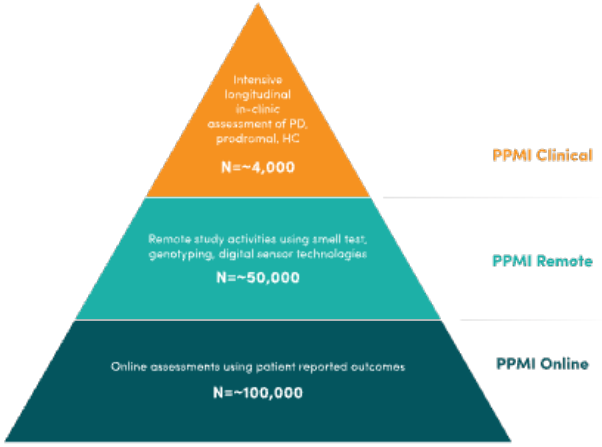
fluid dynamics

omics



digital pathology

Parkinson's Progression Markers Initiative



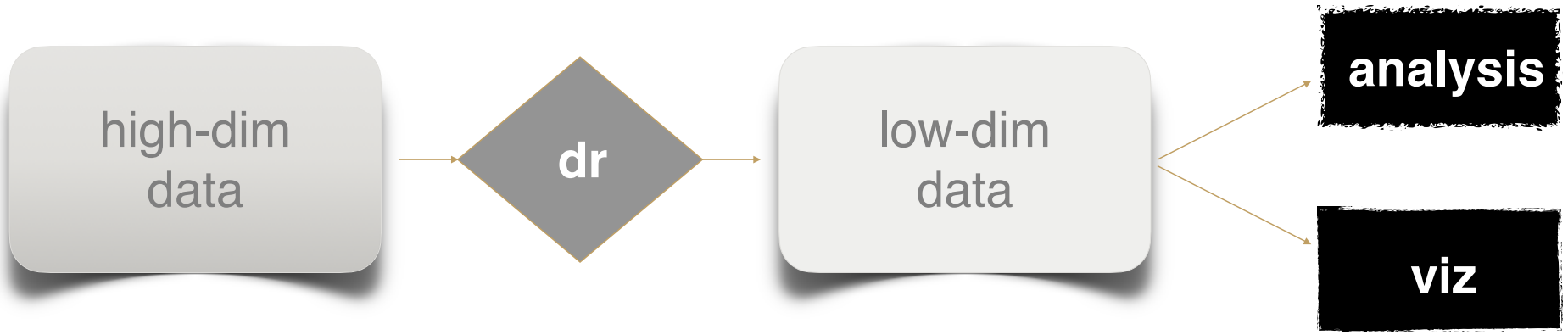
clinical data

Dimensionality reduction

Why?

Many data analysis pipeline cannot be used for high dimensional data for computational and storage issues.

Many models perform poorly for high-dimensional data.



… also for data visualisation!

olivetti face dataset:
10 people, 100 faces,
64x64 resolution

example 1

imaging

every image with *resolution N=m x n*
can be transformed into a vector of
3*N* elements

a dataset with k images is represented
by

$$K = \{x_1, \ldots, x_k\} \subset \mathbb{R}^N$$

a possibility is to reduce N to 2, so the two parameters represents the
pose and the face expression

example 2

Neuroscience

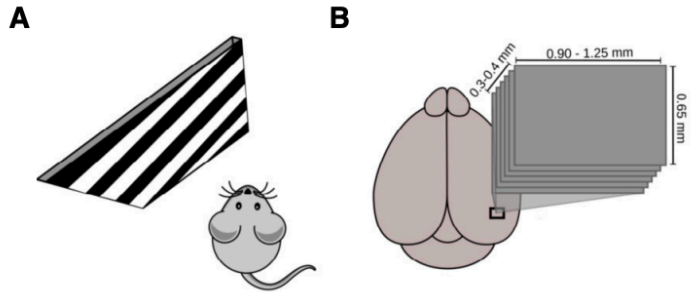
Cell

CellPress

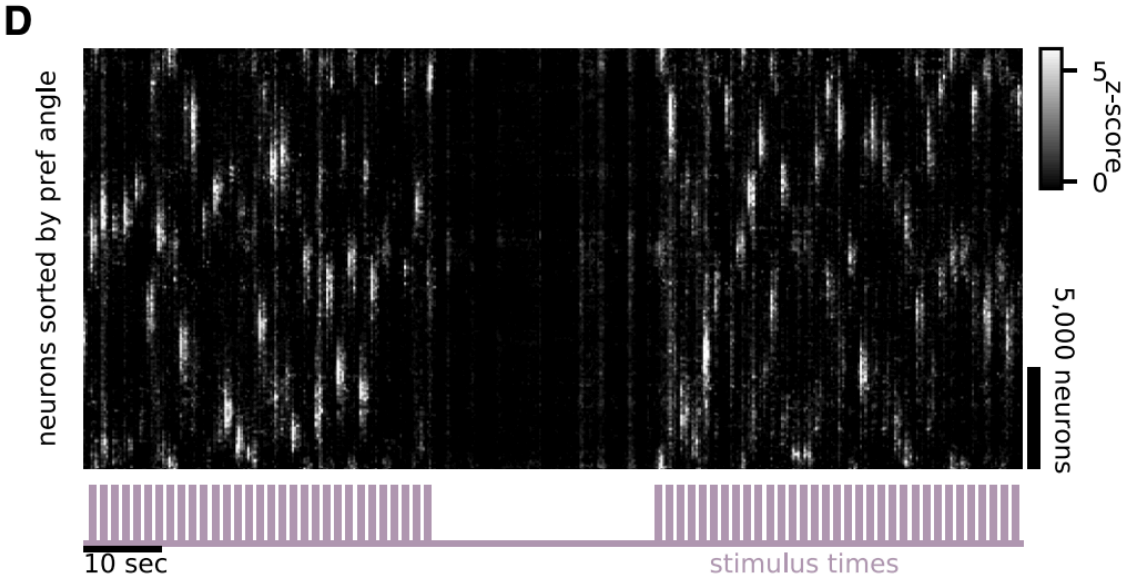
Article

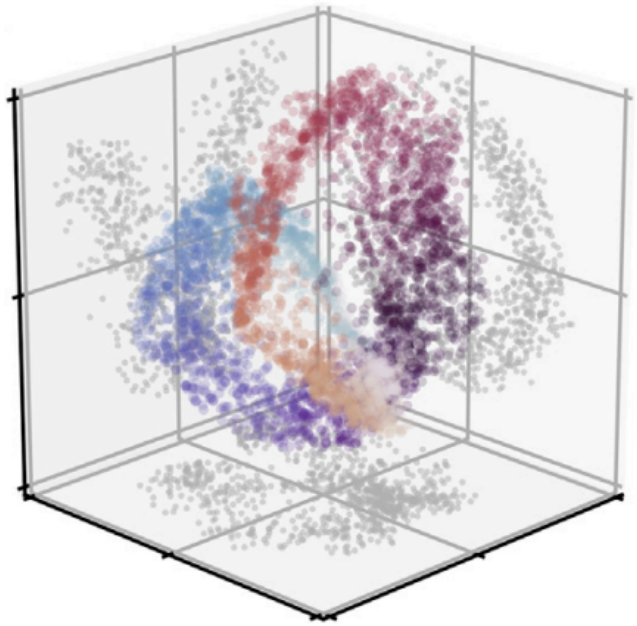
High-precision coding in visual cortex

Carsen Stringer,[1] Michalis Michaelos,[1] Dmitri Tsyboulski,[1] Sarah E. Lindo,[1] and Marius Pachitariu[1,2,*]
[1]HHMI Janelia Research Campus, Ashburn, VA 20147, USA
[2]Lead contact
*Correspondence: pachitarium@janelia.hhmi.org
https://doi.org/10.1016/j.cell.2021.03.042



- ~50,000 neurons
- 3 Hz

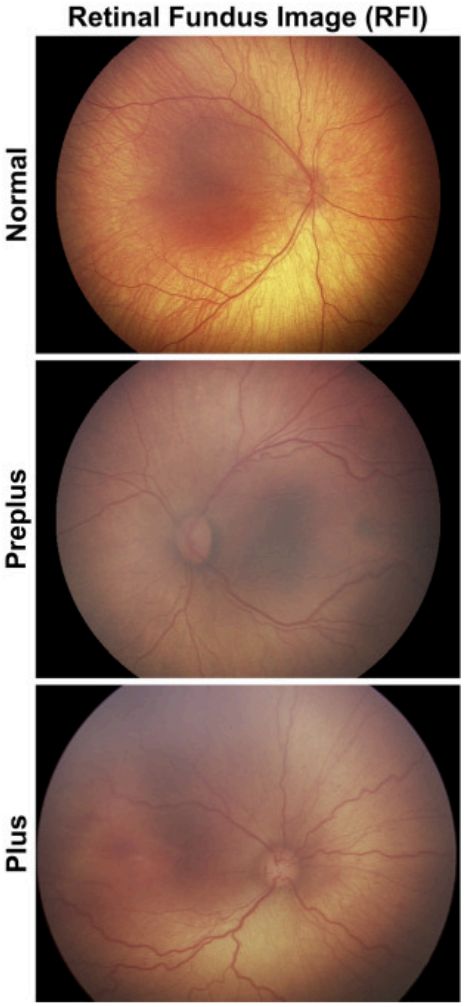




RESEARCH QUESTION

Is the grating direction encoded into the neural activity?

YES

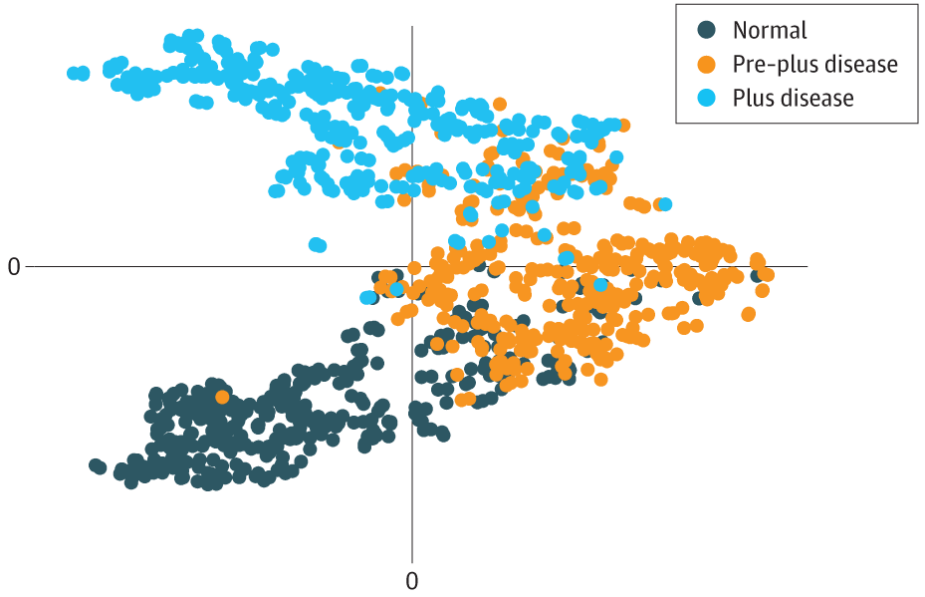


example 3

Retinopathy of Prematurity

Retinopathy of Prematurity (ROP) is a potentially blinding disease that affects prematurely born infants. ROP happens when abnormal blood vessels grow in the retina (the light-sensitive layer of tissue in the back of your eye). A screening method is through Retinal Fundus Images (RFI)

N = 5511 images



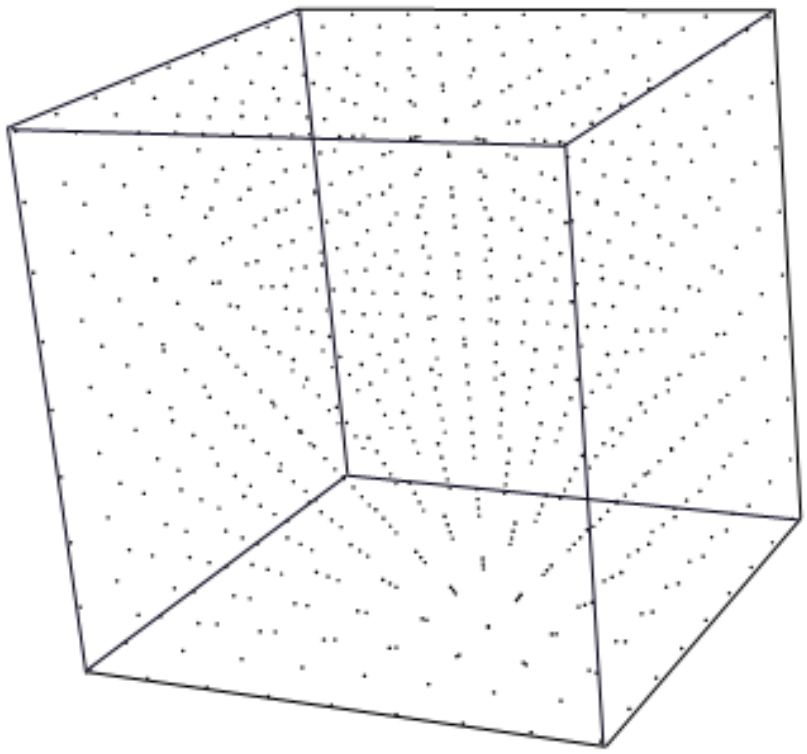
richard bellman

curse of dimensionality

adaptive control processes: a guided tour, 1961

in the absence of simplifying assumptions, the sample size required to estimate a function of several variables to a given degree of accuracy (i.e., to get a reasonably low variance estimate) grows exponentially with the increasing number of variables.

curse of dimensionality

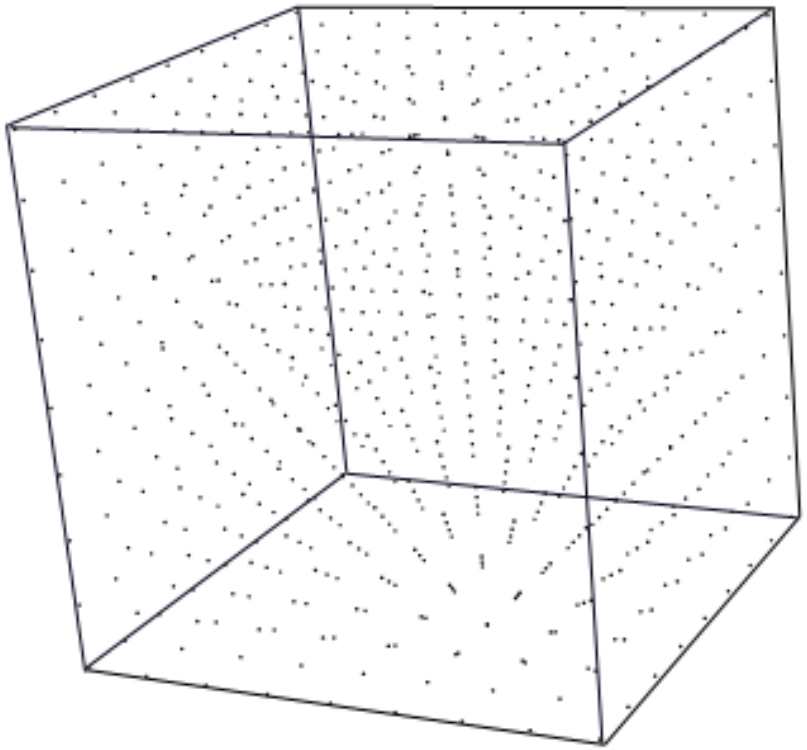


grid of spacing 1/10 the $[0,1)^D$ cube in $\mathbb{R}^D$ has 10^D points

curse of dimensionality

the empty space phenomenon [scott & thompson, 1983]

high-dim spaces are inherently sparse



most of the high-dimensional space remains empty

(hyper)-cubes & -spheres

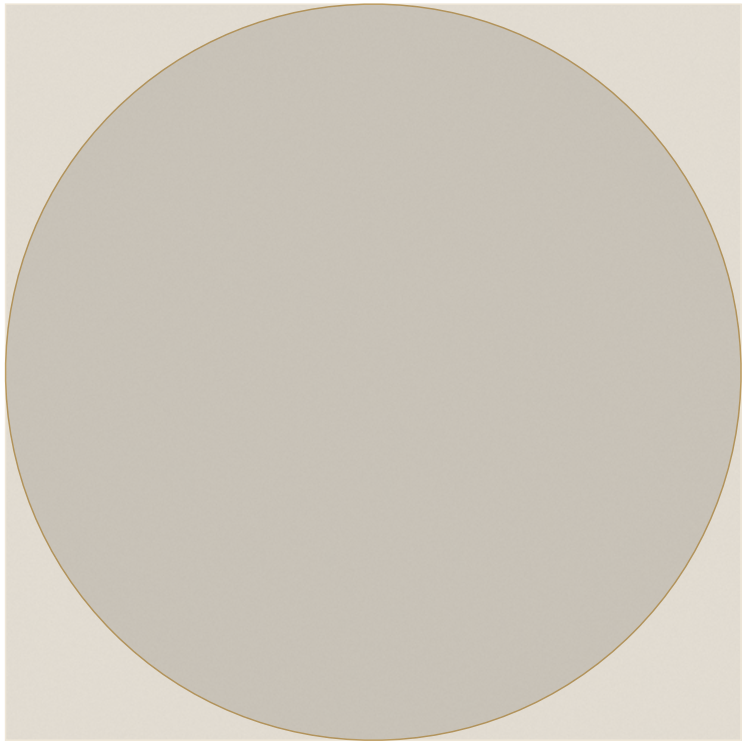
of diameter/side 2r

D=1

$$V_S^D = 2r \qquad V_C^D = 2r$$

(hyper)-cubes & -spheres

of diameter/side 2r



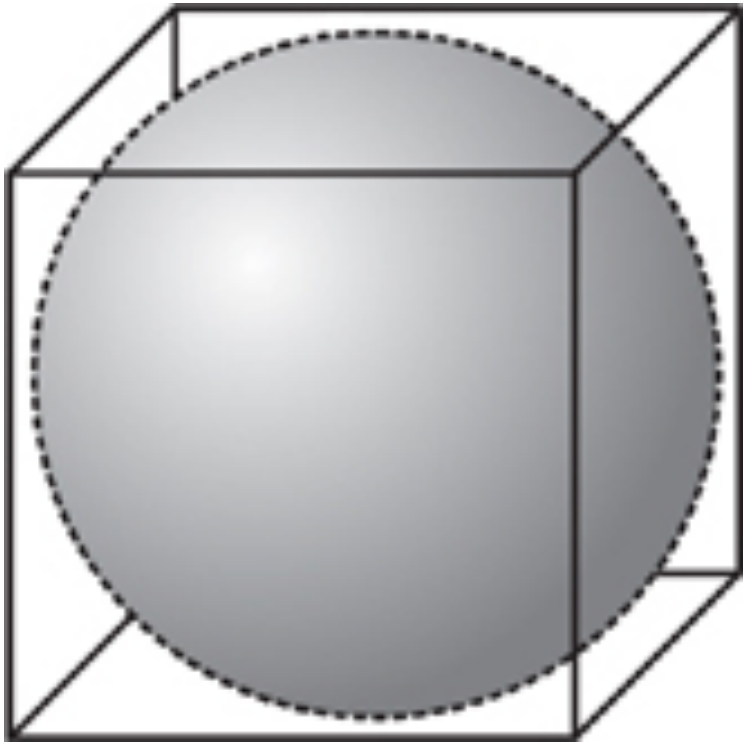
D=2

$$V_S^D = \pi r^2$$

$$V_C^D = 4r^2$$

(hyper)-cubes & -spheres

of diameter/side 2r



D=3

$$V_S^D = \frac{4}{3}\pi r^3 \qquad V_C^D = 8r^3$$

(hyper)-cubes & -spheres

of diameter/side 2r

$$V_S^D = \frac{\sqrt{\pi^D} r^D}{\Gamma\left(\frac{D}{2}+1\right)}$$

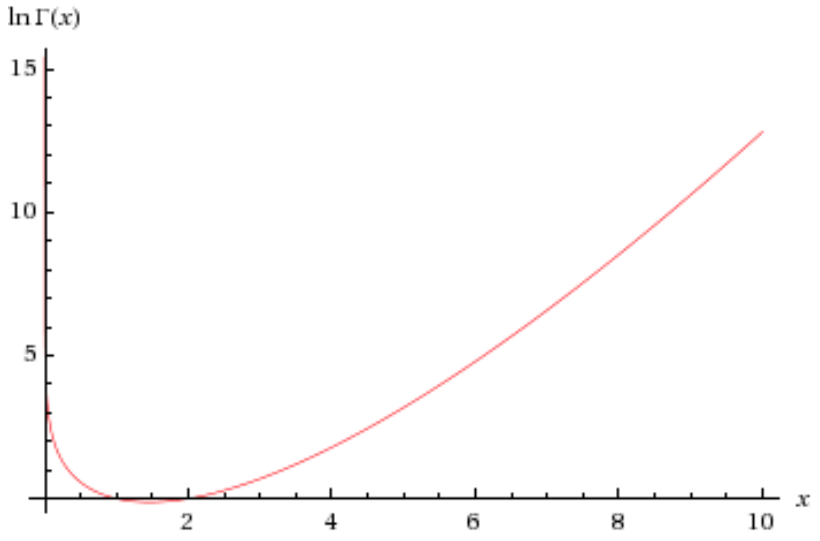
$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma\left(n+\frac{1}{2}\right) = \binom{(n-\frac{1}{2})}{n} n!\sqrt{\pi}$$

D

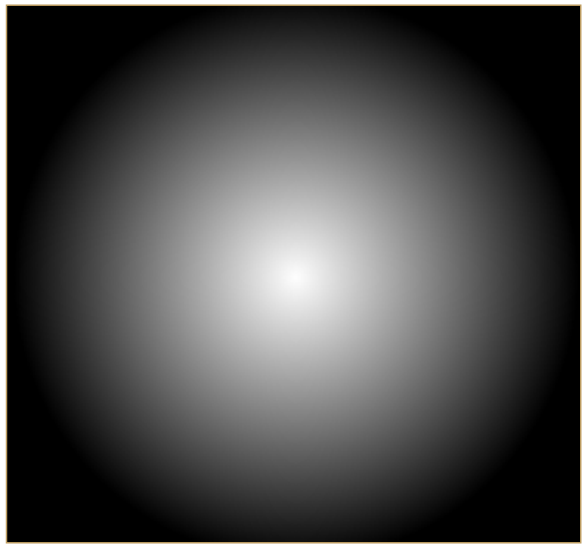
$$V_C^D = (2r)^D$$

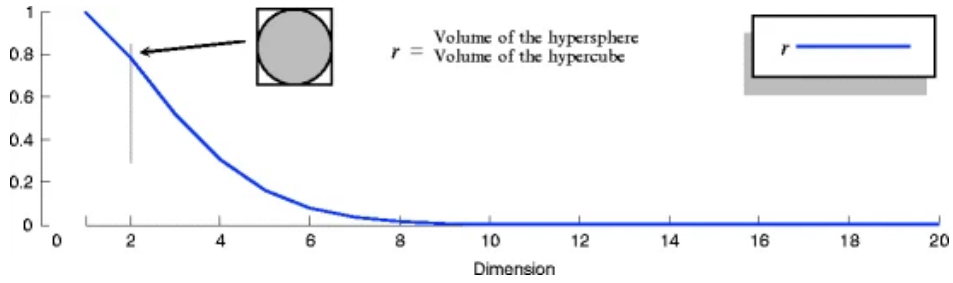


(hyper)-cubes & -spheres

of diameter/side 2r

$$\lim_{D\to\infty} \frac{V_S^D}{V_C^D} = \lim_{D\to\infty} \frac{\frac{\sqrt{\pi^D} r^D}{\Gamma\left(\frac{D}{2}+1\right)}}{(2r)^D} = 0$$

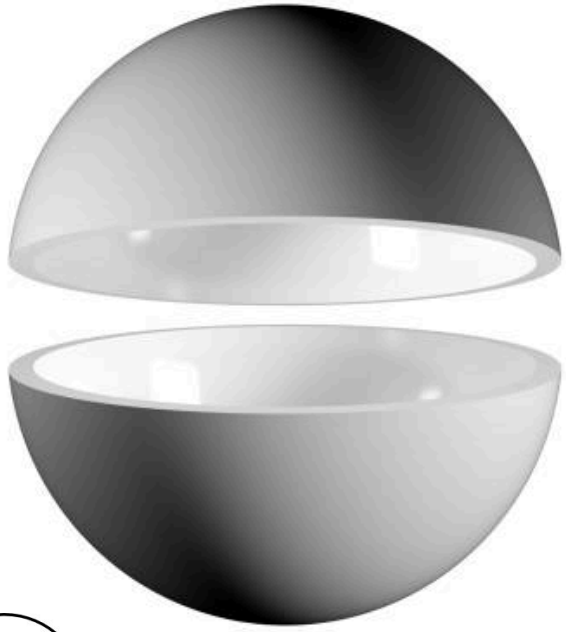




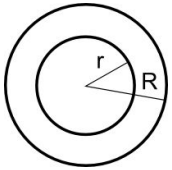
for increasing D, the volume of the cube concentrates more in its corners and less in the inscribed sphere.

(hyper)-spherical shells

$$V_{ss}^{D} = V_{s_R}^{D} - V_{s_r}^{D}$$



virtually all the content of a D dimensional sphere concentrates on its surface, which is only a (D – 1) dimensional manifold

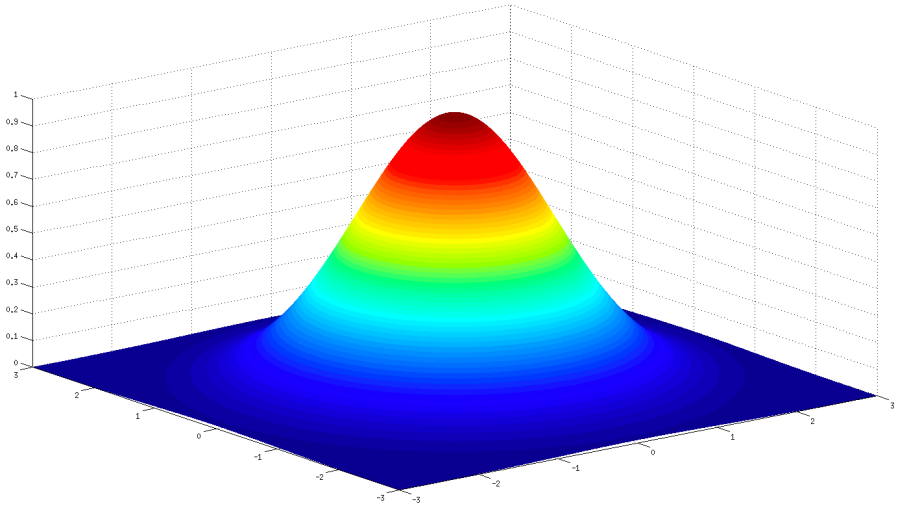


$$V_{s_r}^{D} = \frac{\sqrt{\pi^{D}}\, r^{D}}{\Gamma\left(\frac{D}{2}+1\right)}$$

$$\lim_{D\to\infty} \frac{V_{ss}^{D}}{V_{s_R}^{D}} = \lim_{D\to\infty} 1 - \frac{V_{s_r}^{D}}{V_{s_R}^{D}} = \lim_{D\to\infty} 1 - \left(\frac{r}{R}\right)^{D} = 1$$

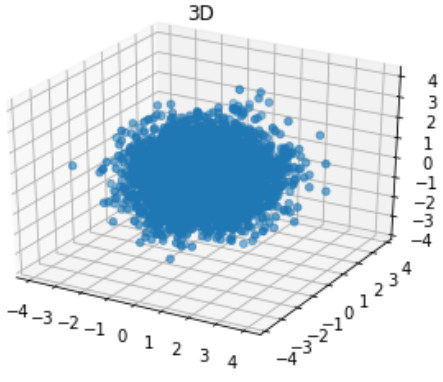
normal distributions

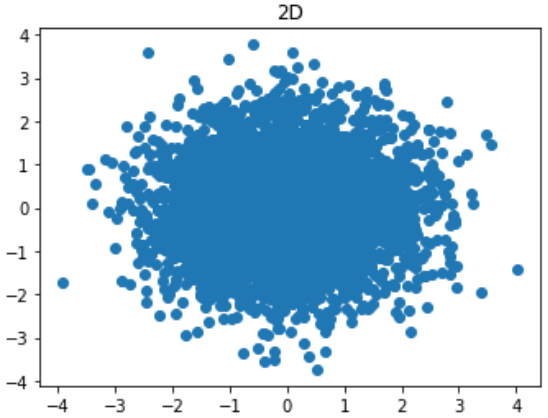
$N_D(0,1)$

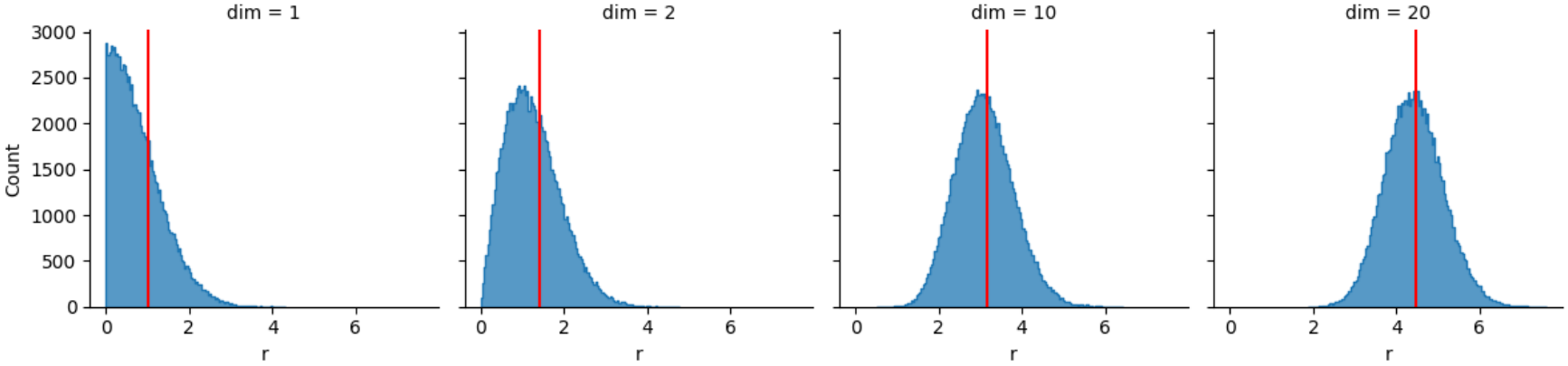


pdf $G(x) = K(r) = \frac{1}{\sqrt{(2\pi)^D}} e^{-\frac{r^2}{2}}, ||x|| = r$

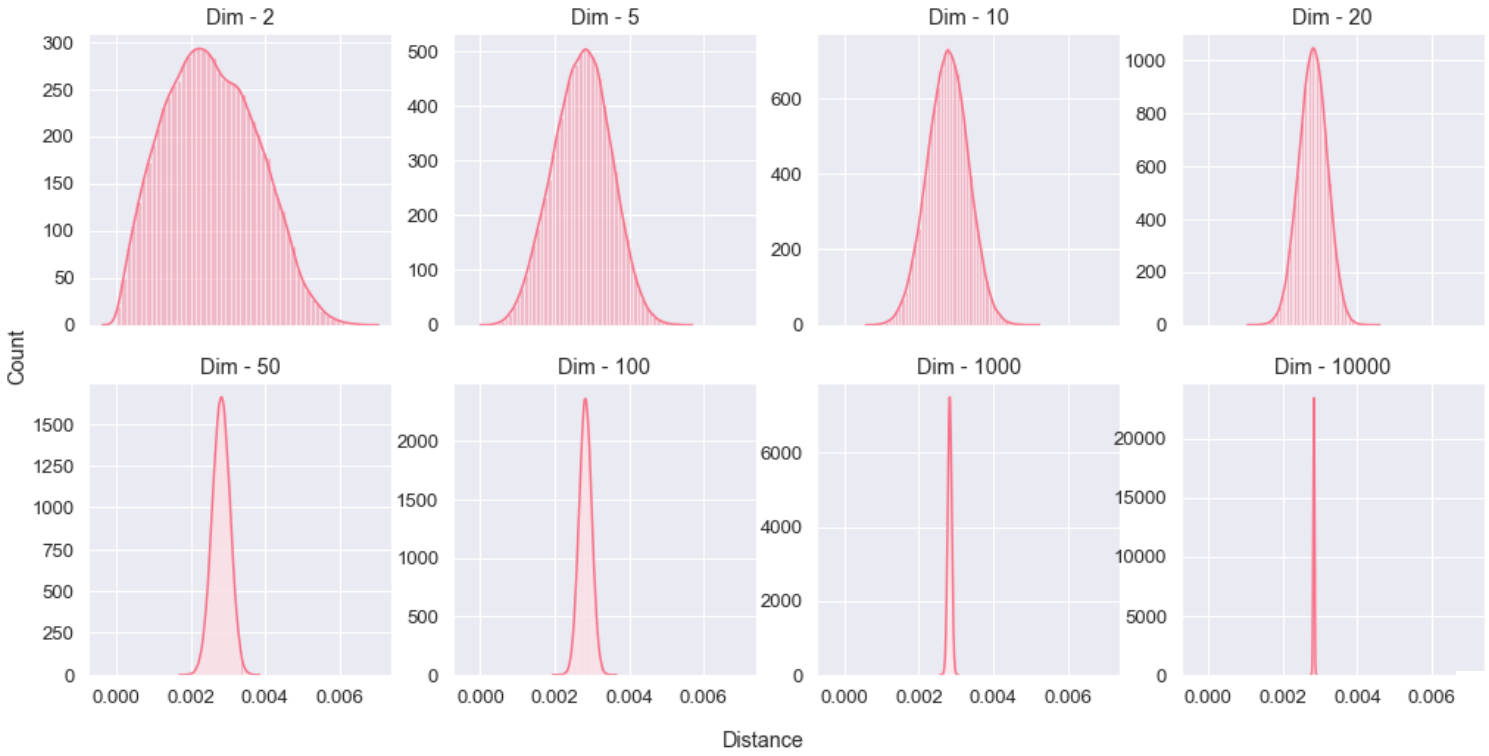
multivariate gaussian are soap bubbles as the dimension increases



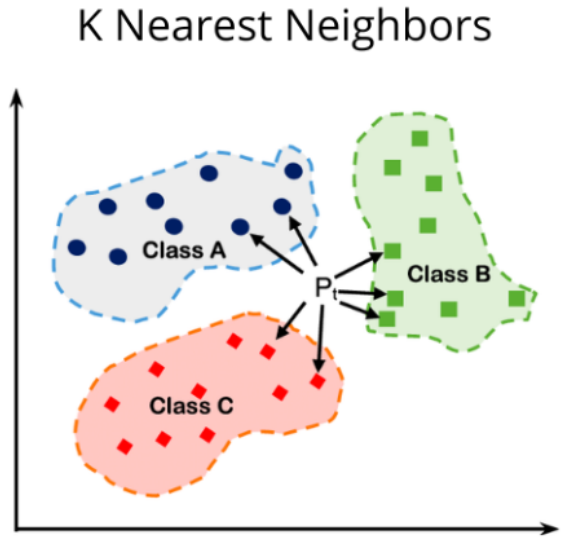




concentration of norms



euclidean distance between any two vectors is approximately constant.



intrinsic/extrinsic dim

the points of high-dimensional data usually reside on a much low-dimensional manifold

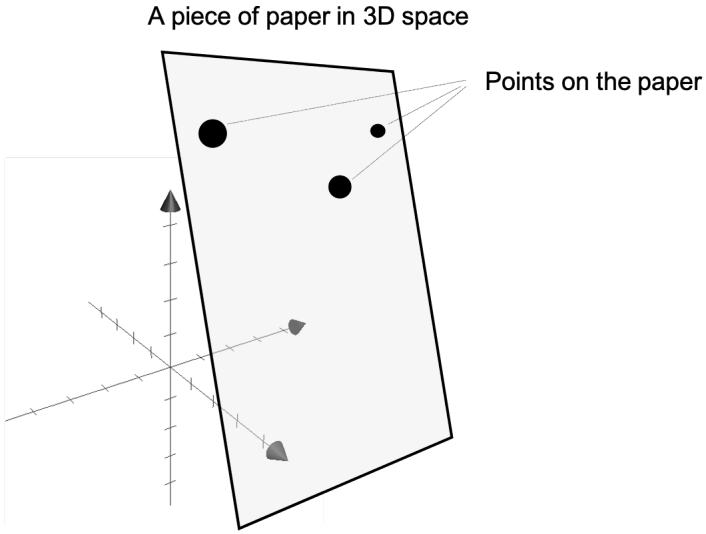
$$X = \{x_\alpha\}_{\alpha \in A}\ , x_\alpha \in M,\ \dim(M) = s,\ M \subset \mathbb{R}^D$$

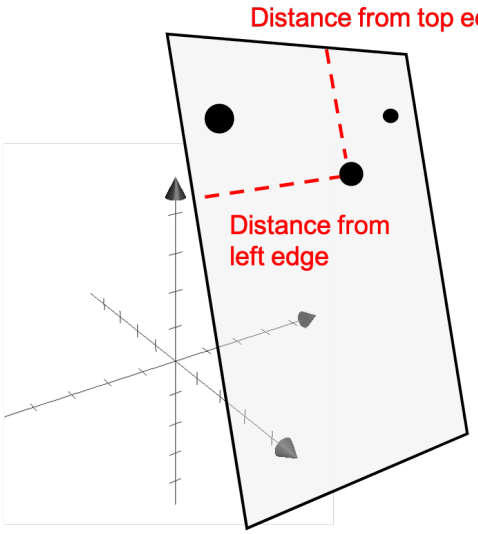
D extrinsic dimension of X, s intrinsic dimension of X

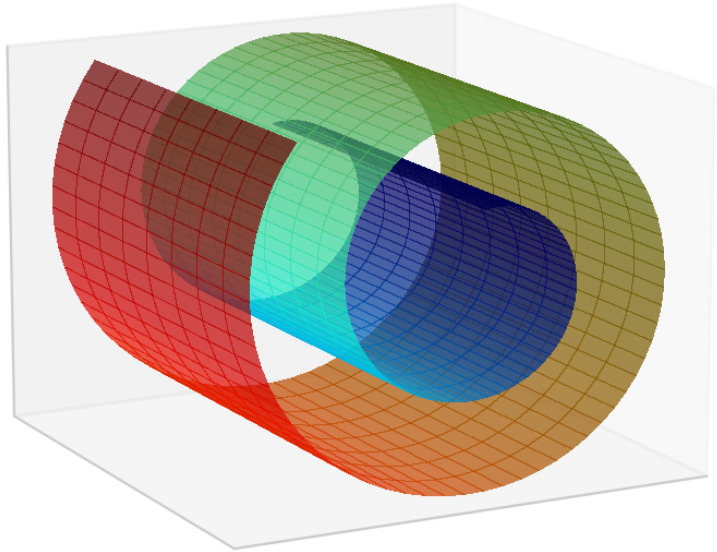
X samples set of random vectors **X** in dimension D

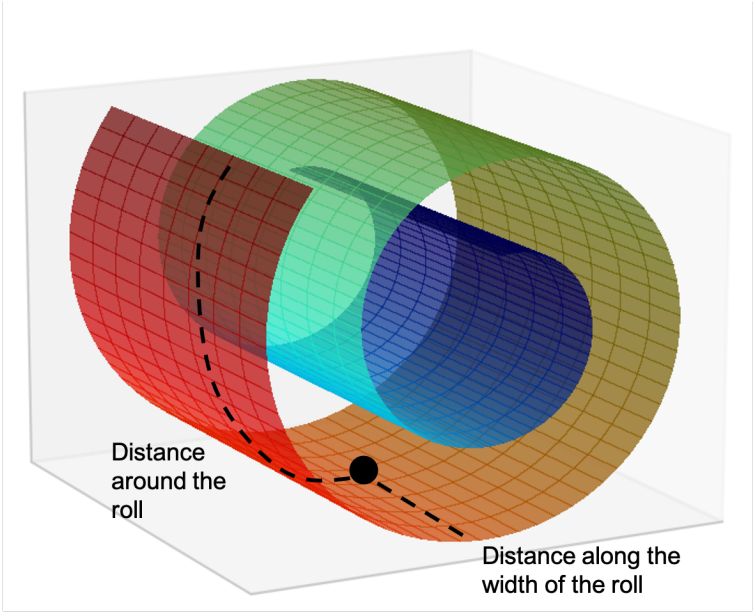
there exist a random vector **Y** in dim s and an invertible analytic function *f* s.t. *f*(**Y**)=**X**

intrinsic/extrinsic dim

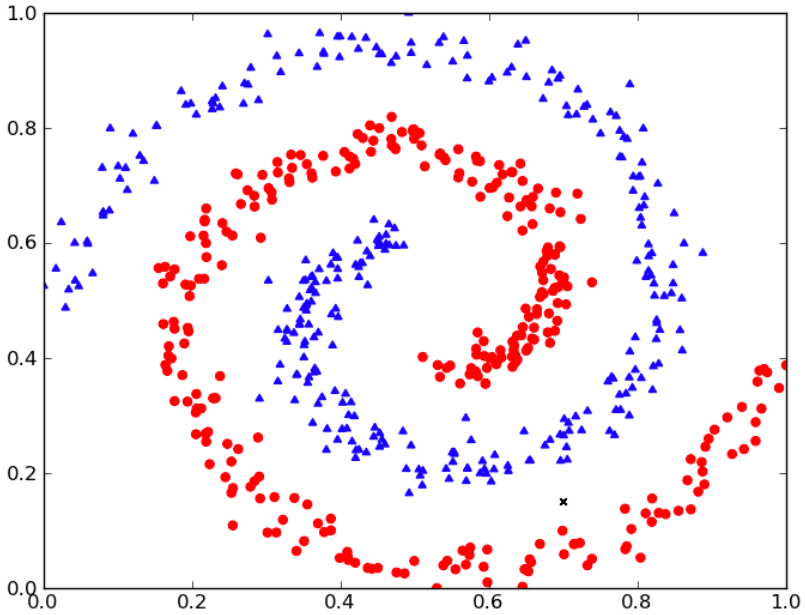


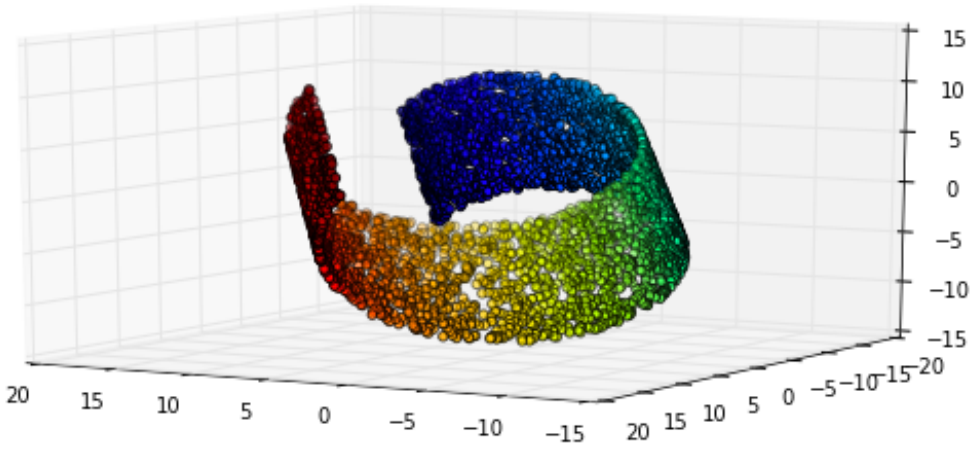






intrinsic/extrinsic dim





due to the low intrinsic dimension of data, we can reduce the (extrinsic) dimension without losing much information for many types of real-life high-dimensional data, avoiding many of the curses of dimensionality

dr is finding a parameterization of the manifold which the points of data reside on

Dimensionality reduction

How?

Removing irrelevant or redundant features, or simply noisy data.

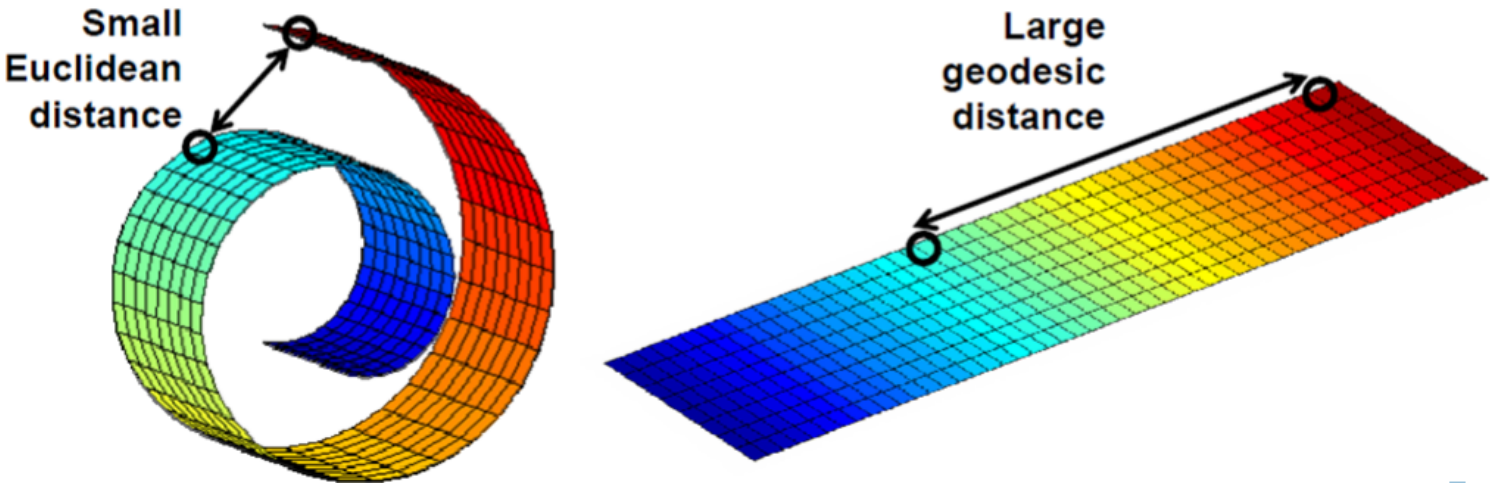
While dimensionality reduction methods differ in operation, they all transform high-dimensional spaces into low-dimensional spaces through variable extraction or combination.

Dimensionality reduction

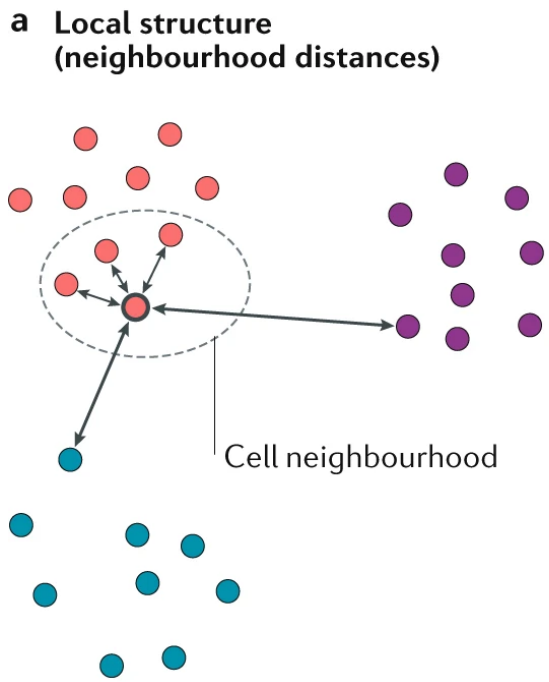
A definition

Dimensionality reduction is a method for representing a given dataset using a lower number of features (that is, dimensions) while still **capturing the original data's meaningful properties**.

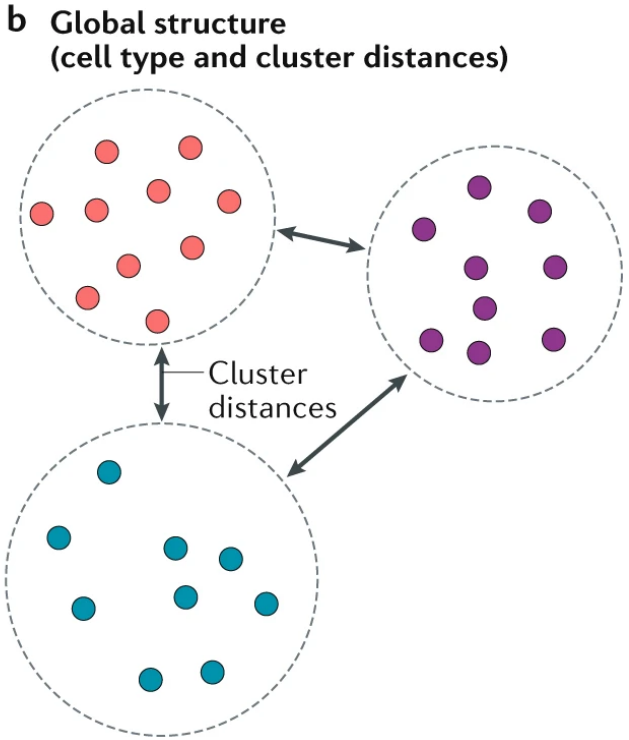
Lih-Yuan Deng, Max Garzon, and Nirman Kumar,
Dimensionality Reduction in Data Science, Springer, 2022.



Local and global structure



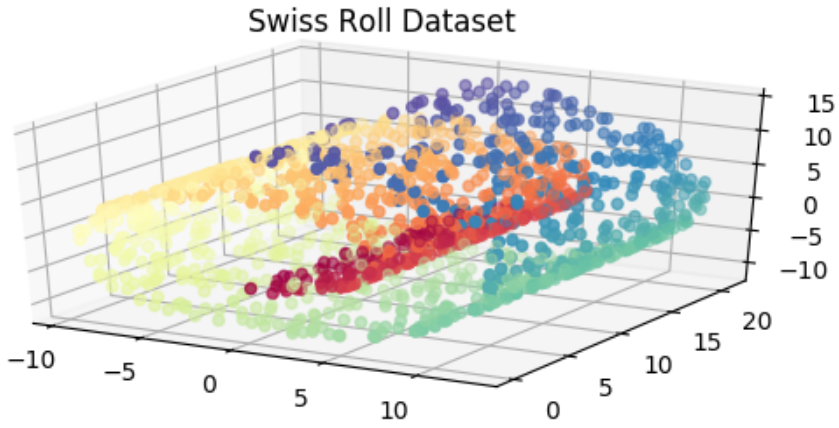
(a) Preserving the local structure of a dataset ensures that neighbouring points remain together in the visualization, rather than preserving the original high dimensional space.

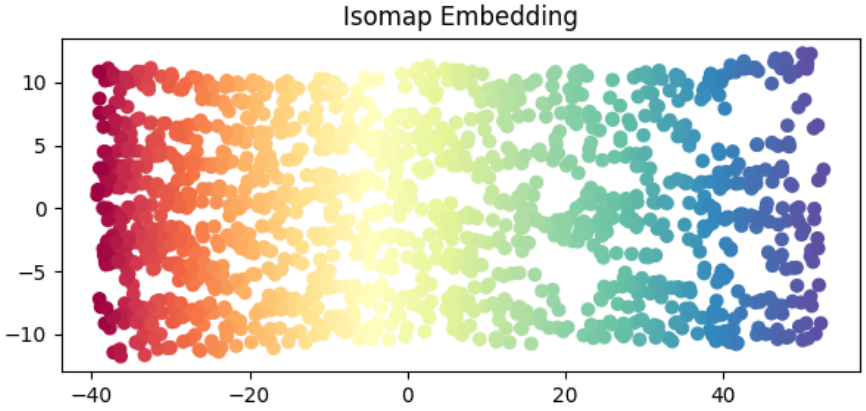


(b) Preserving the global structure of a dataset ensures that large-scale distances are maintained.

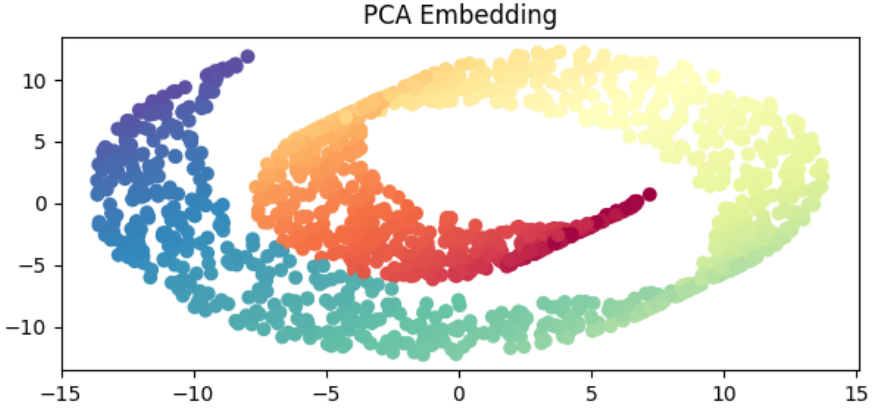
https://www.nature.com/articles/s41581-020-0262-0

Local and global structure



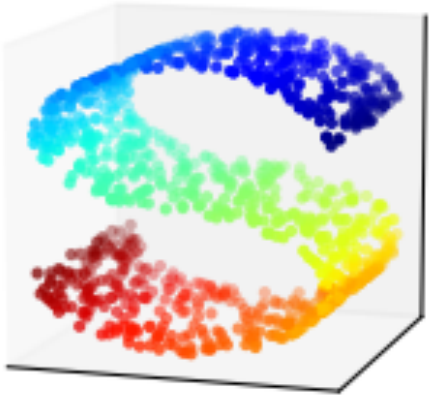


Local structure preserved

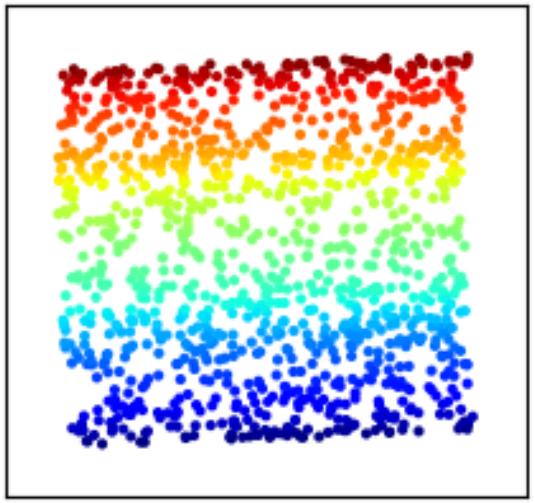


Global structure preserved

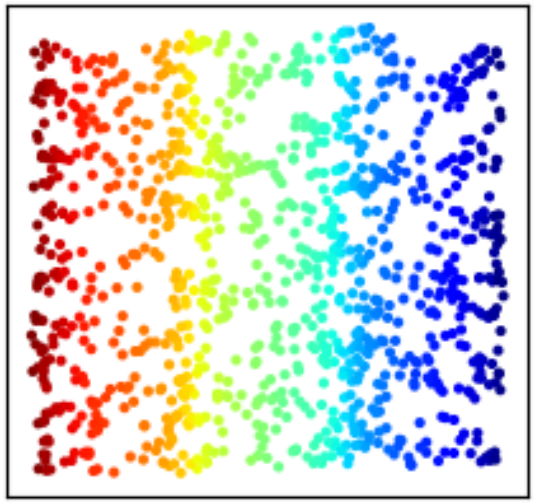
Local and global structure



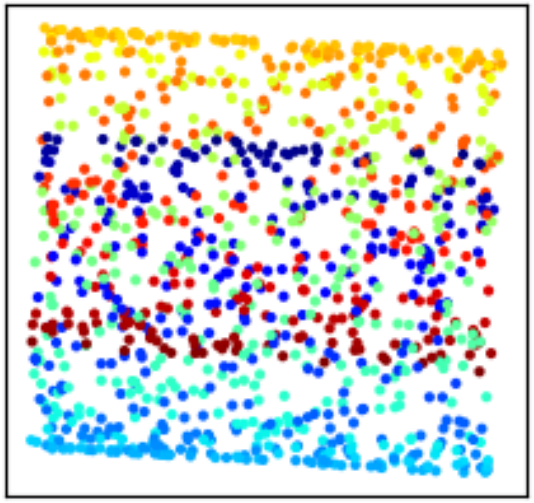
LLE projection



IsoMap projection



PCA projection



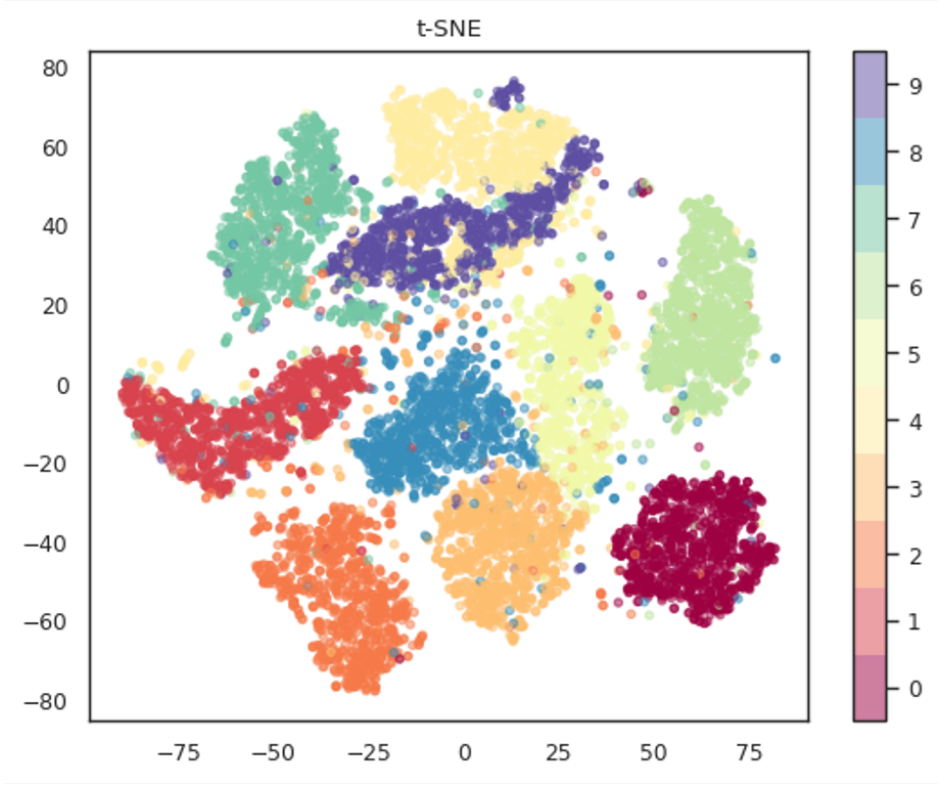
Local structure preserved

Global structure preserved

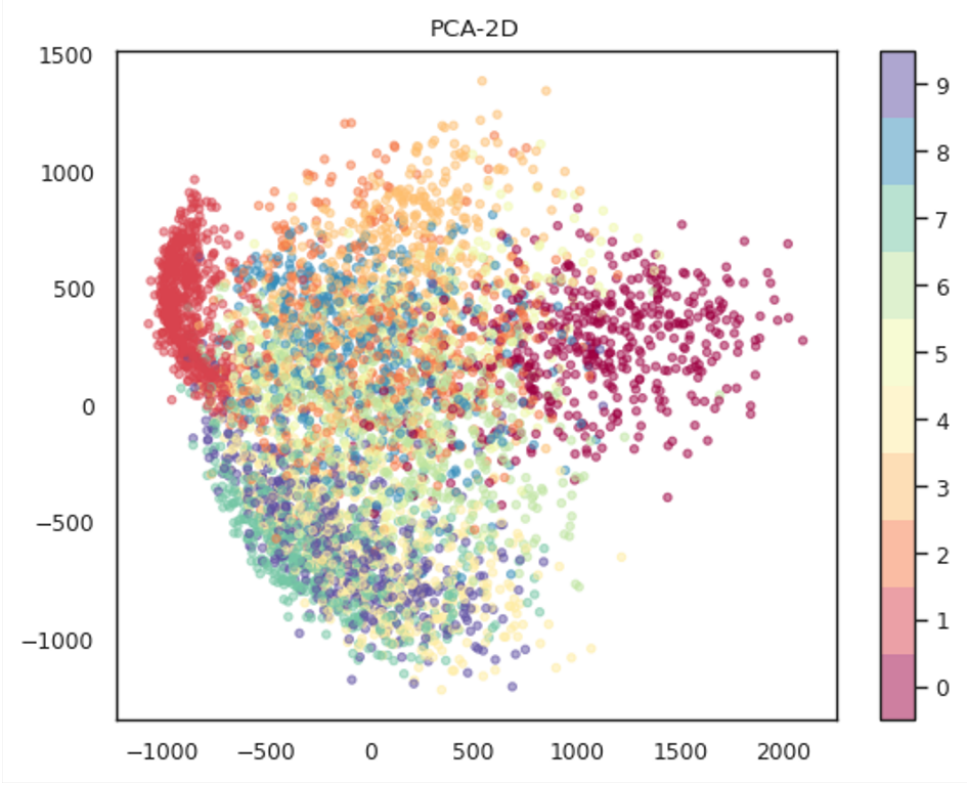
MNIST dataset

- Modified National Institute of Standards and Technology database
- Handwritten digits
- 1998
- Grayscale images
- 28 x 28 pixels (flattened to N=784)
- 60,000 training images
- 10,000 testing images

Local and global structure

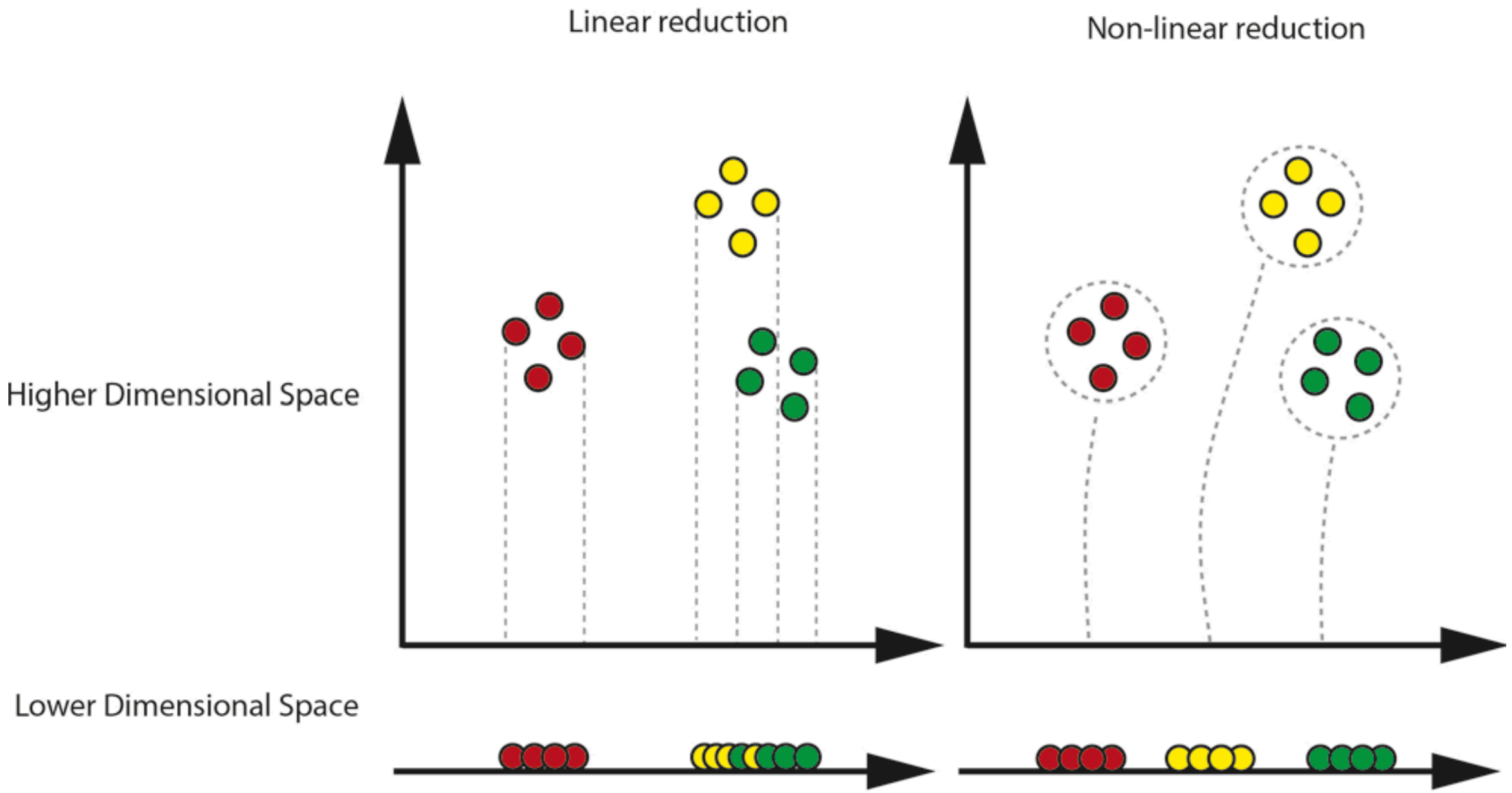


Local structure preserved



Global structure preserved

Linear and non linear



Dimensionality reduction

Methods

Principal Component Analysis	PCA	Linear	Global
Multidimensional scaling	MDS	Linear	Global
Isometric mapping	Isomap	Non linear	Local
Locally Linear Embedding	LLE	Non linear	Local
t-distributed Stochastic Neighbor Embedding	t-SNE	Non linear	Local
Uniform Manifold Approximation and Projection	UMAP	NNon linear	Local

Dimensionality reduction

Caveats

- Local structure preservation
- Global structure preservation
- Sensitivity to parameters (and user choices)
- Global efficiency

No “silver bullet”