Geometry of representations in self-supervised deep learning models

Alessio Ansuini and Alberto Cazzaniga

alessio.ansuini@areasciencepark.it

alberto.cazzaniga@areasciencepark.it

AREA Science Park,
Institute for Research and Innovation Technology

DTU Advanced ML summer school 2023

Copenhagen, 23/08/2023

AREA Science Park

Since more than 40 years:

- Research Infrastructures
- Scientific Park
- Technology transfer

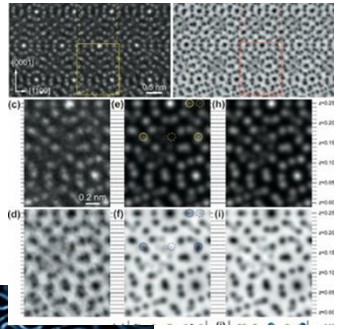
Last 3 years added:

FUNDAMENTAL AND APPLIED RESEARCH

Institute for Research and Innovative Technologies: one heart and three souls

LAGE

LAME



LADE

Laboratory of Data Engineering

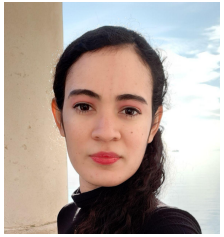
Last three years at LADE





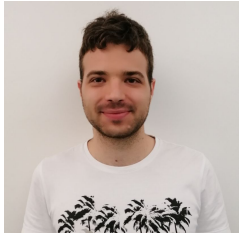









Plan of the day

Morning:
Intro and methods of analysis (1h 15’)

- Generalities about representations and the manifold hypothesis
- Intrinsic dimension
- Neighborhood structure

Pause 15’

Geometry of representations in large transformer models (1h 15’)

Afternoon:
Introduction to the exercises and Q/A (1h 30’)

Representations

Q: What do we mean by representations?

Representations

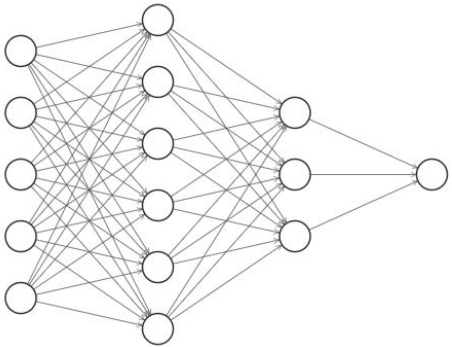
Q: What do we mean by representations?

A: Patterns of activity of artificial of biological units (neurons) caused by external or internal signals

Representations

Q: What do we mean by representations?

A: Patterns of activity of artificial of biological units (neurons) caused by external or internal signals

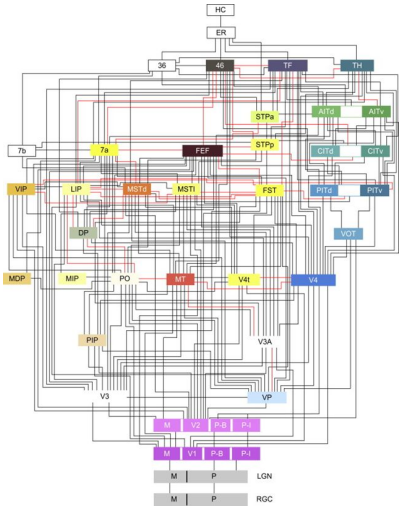


$\{\text{cat}, \text{dog}\}$

Representations

Q: What do we mean by representations?

A: Patterns of activity of artificial of biological units (neurons) caused by external or internal signals



press red button if "cat" [...]

Representations in deep networks

N.of pixels $\times$ channels = D

$$x_1, x_2, \ldots, \ldots, x_D \quad \in \mathbb{R}^D$$

$$h_1^{(1)}, h_2^{(1)}, \ldots, \ldots, h_{n_1}^{(1)} \quad \in \mathbb{R}^{n_1}$$

$$h_1^{(2)}, h_2^{(2)}, \ldots, \ldots, h_{n_2}^{(2)} \quad \in \mathbb{R}^{n_2}$$

$$\ldots \rightarrow h^{(L)} \quad \in \{0, 1\}$$

Representations in deep networks

N.of pixels $\times$ channels = D

$$\begin{matrix} x_1 \\ x_2 \\ \cdots \\ \cdots \\ \cdots \\ x_D \end{matrix} \quad \mathbb{R}^D$$

$$\begin{matrix} h_1^{(1)} \\ h_2^{(1)} \\ \cdots \\ \cdots \\ \cdots \\ h_{n_1}^{(1)} \end{matrix} \quad \mathbb{R}^{n_1}$$

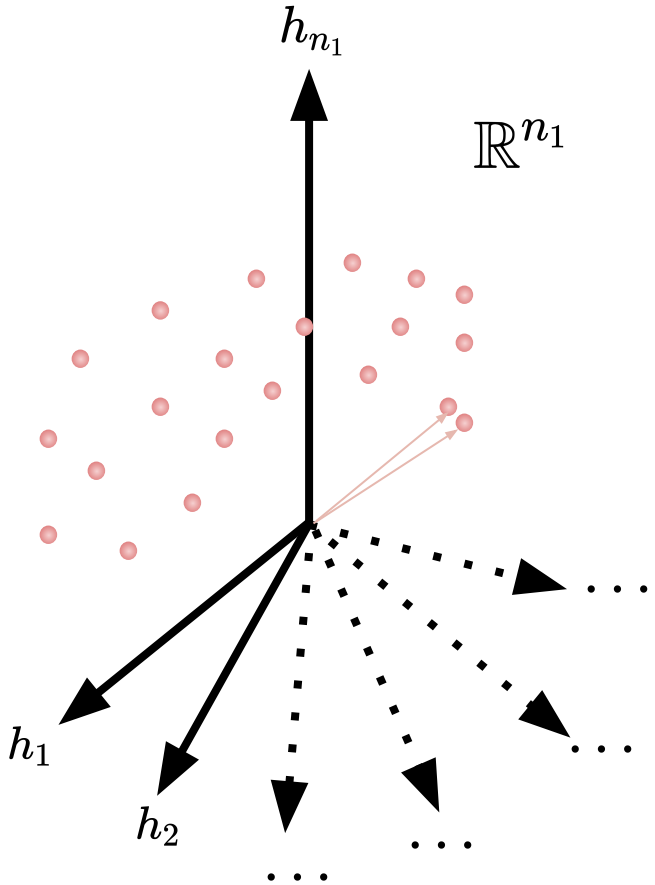
$$\begin{matrix} h_1^{(2)} \\ h_2^{(2)} \\ \cdots \\ \cdots \\ \cdots \\ h_{n_2}^{(2)} \end{matrix} \quad \mathbb{R}^{n_2}$$

$$\cdots$$

$$\begin{matrix} h_1^{(L)} \\ h_2^{(L)} \\ \cdots \\ \cdots \\ \cdots \\ h_D^{(L)} \end{matrix} \quad \mathbb{R}^D$$

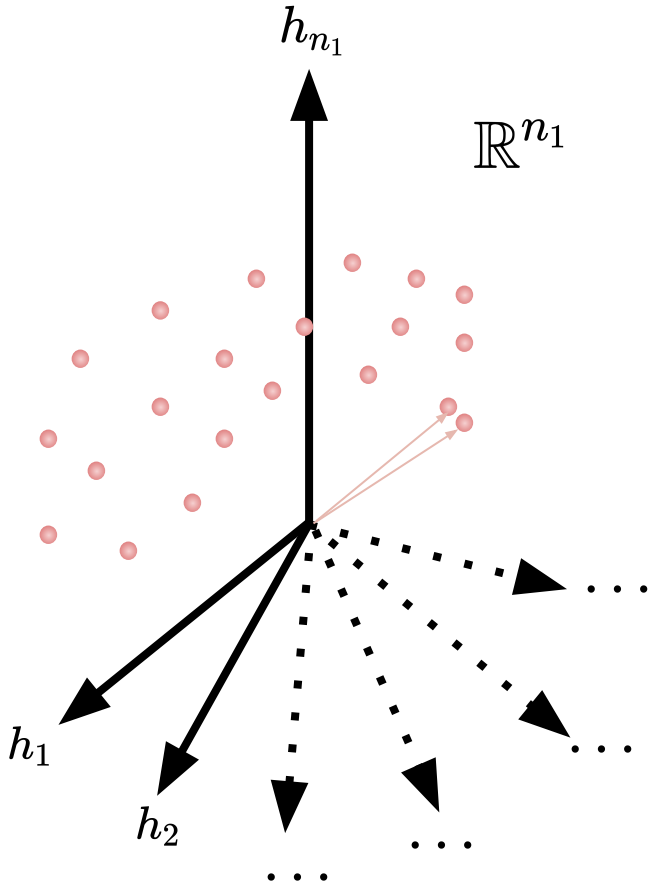
Representation matrix

$$A_{N\times n_1} = \begin{bmatrix} h_1(x^{(1)}) & h_2(x^{(1)}) & \dots & h_{n_1}(x^{(1)}) \\ h_1(x^{(2)}) & h_2(x^{(2)}) & \dots & h_{n_1}(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x^{(N)}) & h_2(x^{(N)}) & \dots & h_{n_1}(x^{(N)}) \end{bmatrix}$$



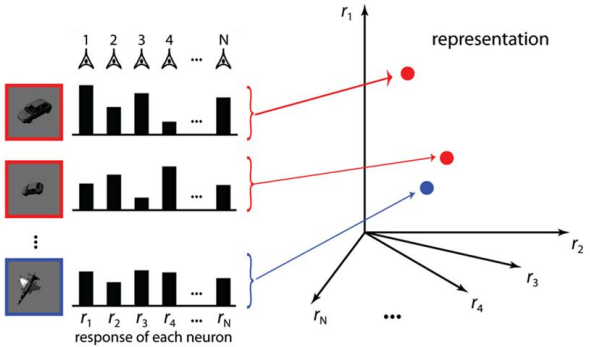
Representation matrix

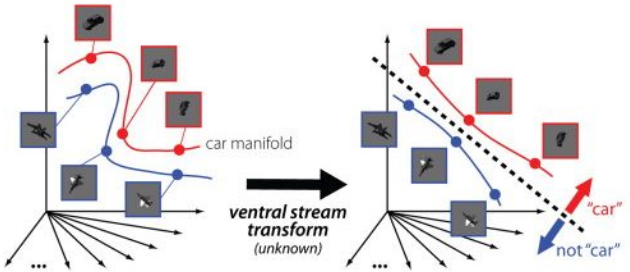
$$A_{N\times n_1} = \begin{bmatrix} h_1(x^{(1)}) & h_2(x^{(1)}) & \dots & h_{n_1}(x^{(1)}) \\ h_1(x^{(2)}) & h_2(x^{(2)}) & \dots & h_{n_1}(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x^{(N)}) & h_2(x^{(N)}) & \dots & h_{n_1}(x^{(N)}) \end{bmatrix}$$

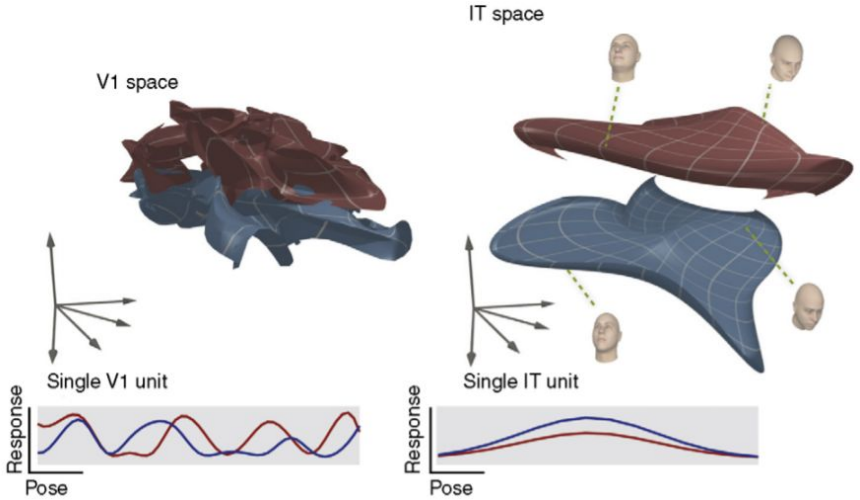


Similar vectors = Similar data

A natural question in this “view”

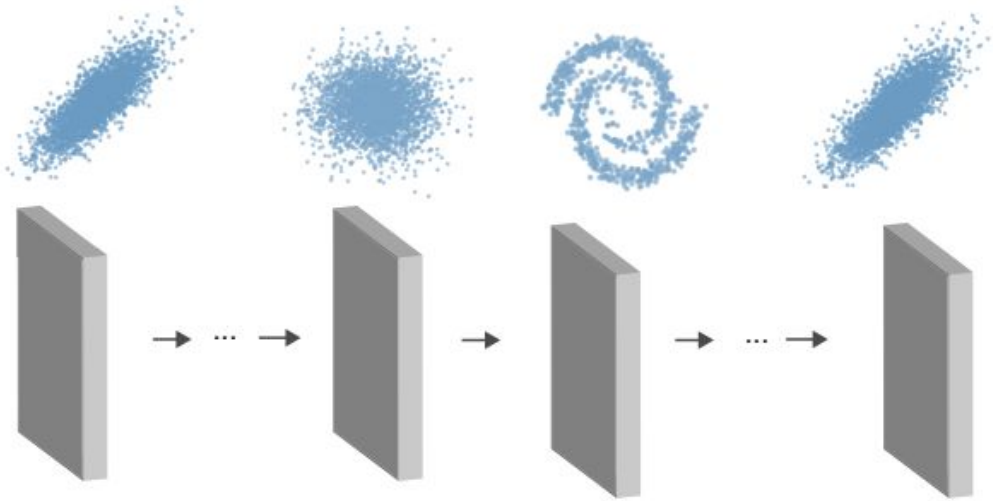




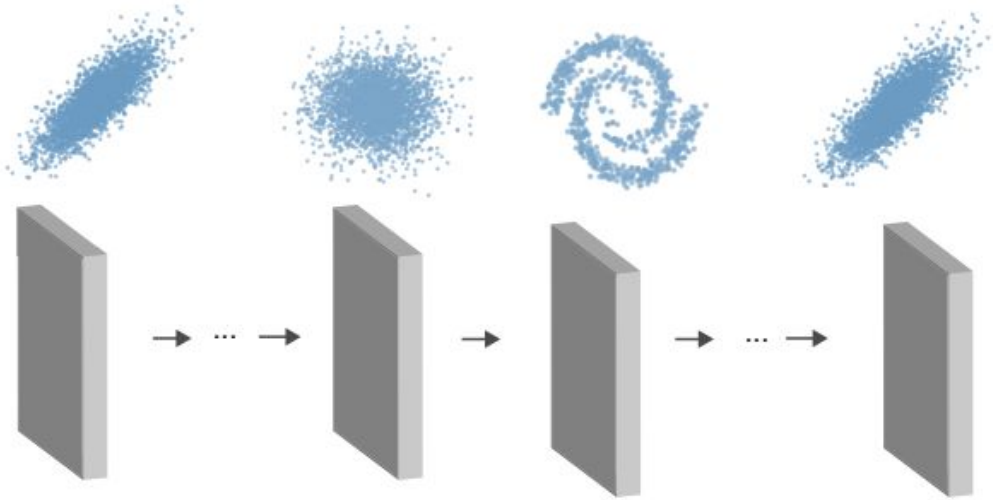


J. DiCarlo et al, “How does the brain solve visual object recognition?, 2012

A natural question in this “view”

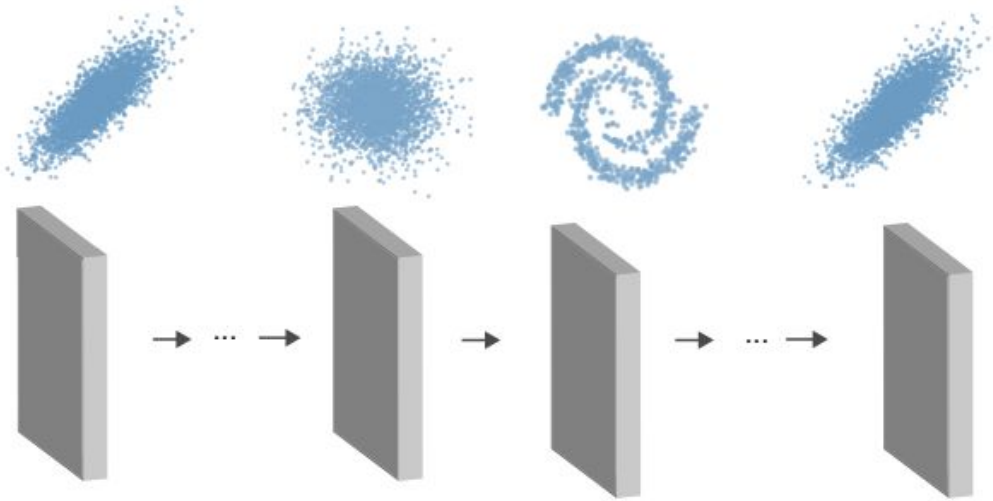


A natural question in this “view”



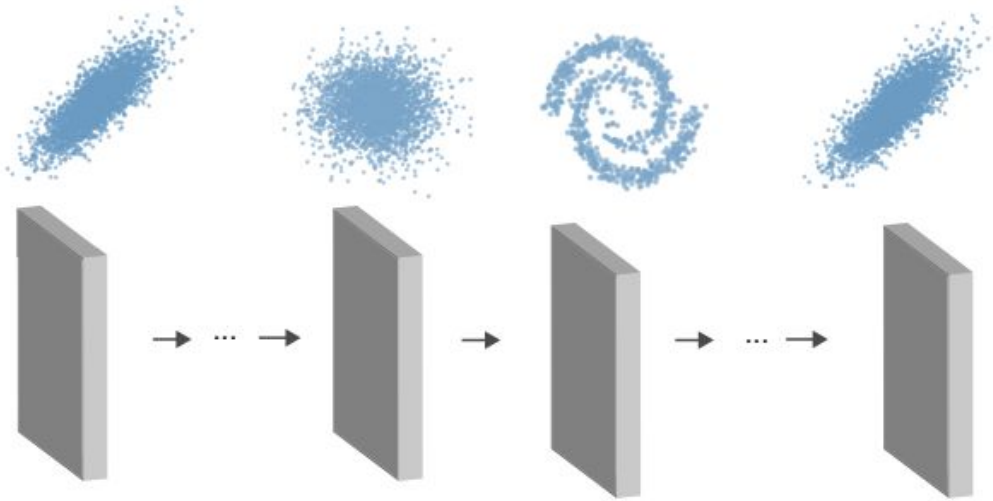
- How the geometry of representations change?

A natural question in this “view”



- How the geometry of representations change?
- How change the type of information encoded?

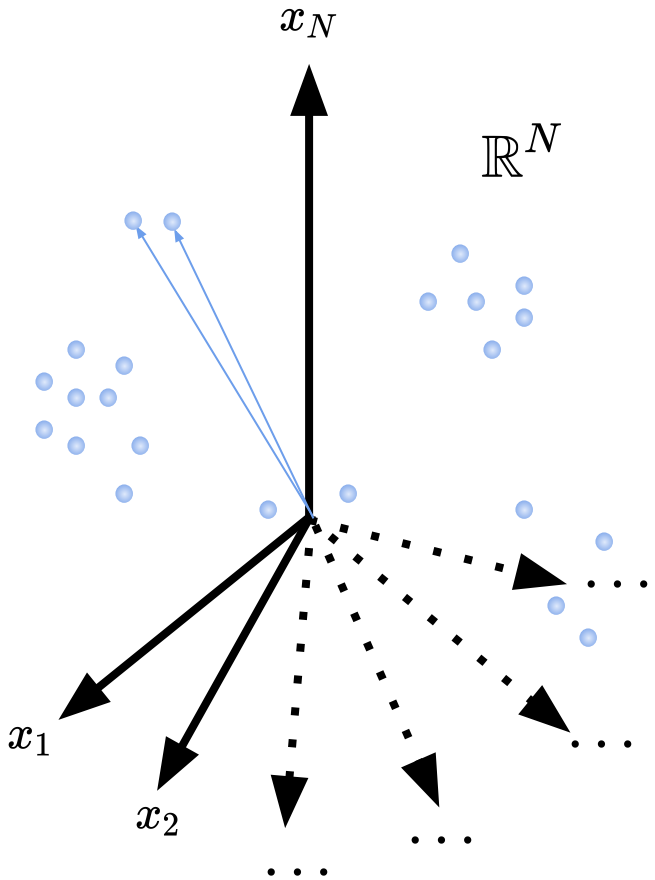
A natural question in this “view”



- How the geometry of representations change?
- How change the type of information encoded?
- How geometry and information are connected?

Representation matrix: a different "view"

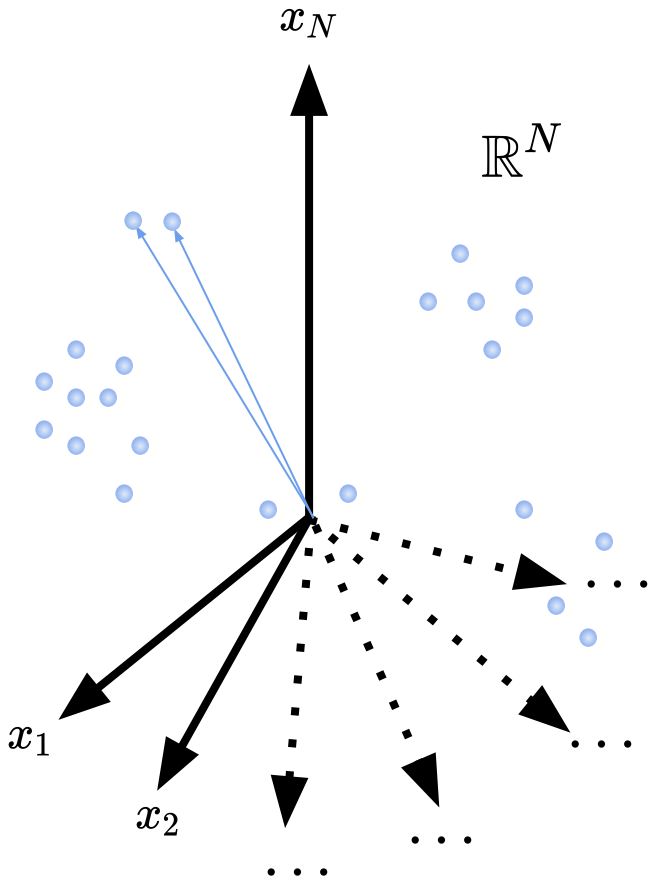
$$A_{N\times n_1} = \begin{bmatrix} h_1(x^{(1)}) & h_2(x^{(1)}) & \dots & h_{n_1}(x^{(1)}) \\ h_1(x^{(2)}) & h_2(x^{(2)}) & \dots & h_{n_1}(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x^{(N)}) & h_2(x^{(N)}) & \dots & h_{n_1}(x^{(N)}) \end{bmatrix}$$



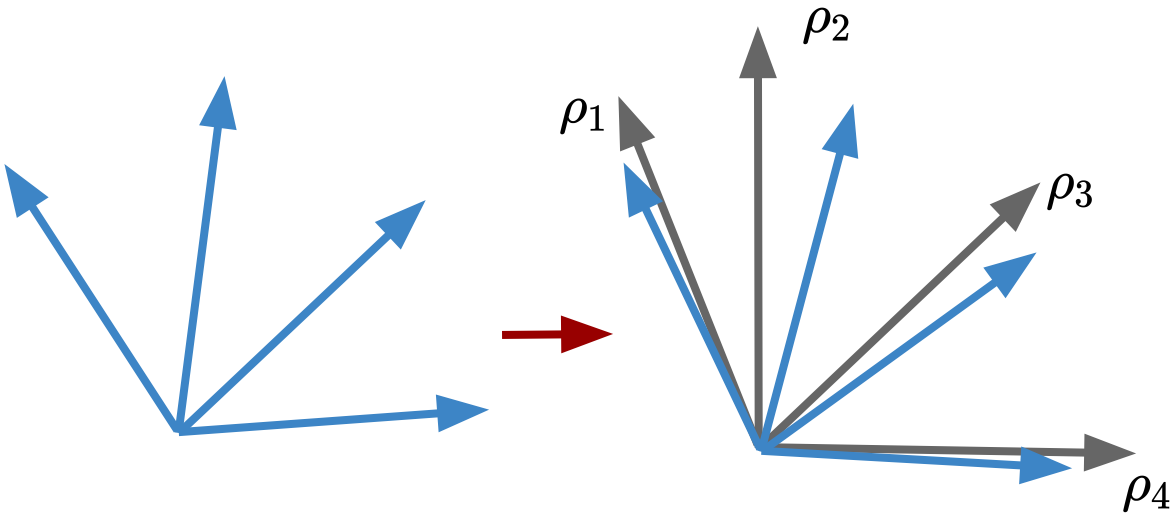
Representation matrix: a different “view”

$$A_{N\times n_1} = \begin{bmatrix} h_1(x^{(1)}) & h_2(x^{(1)}) & \dots & h_{n_1}(x^{(1)}) \\ h_1(x^{(2)}) & h_2(x^{(2)}) & \dots & h_{n_1}(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x^{(N)}) & h_2(x^{(N)}) & \dots & h_{n_1}(x^{(N)}) \end{bmatrix}$$

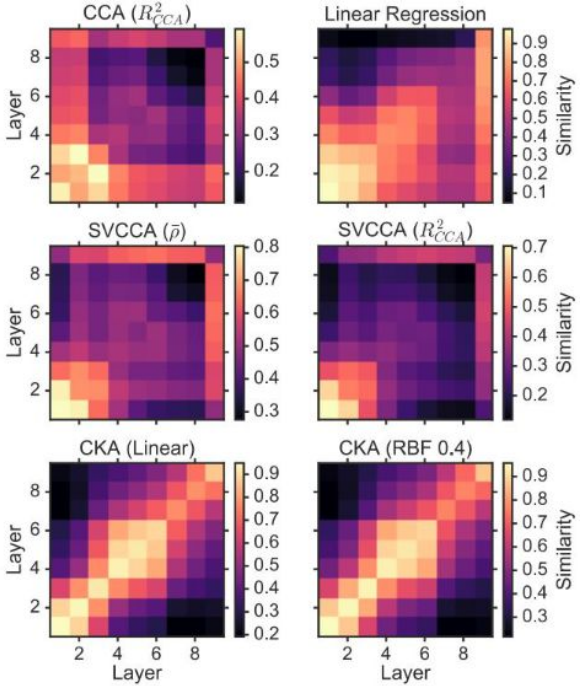
Similar vectors = Similar units



Natural questions in this “view”



$$\bar{\rho} = \frac{1}{\min(m_1, m_2)} \sum_i \rho_i$$



M. Raghu et al., SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability, 2017

Kornblith, Similarity of Neural Network Representations Revisited, 2019

Questions on representations?

Geometry (dimension)

Robustness

“Shape” of data

Information

Computational strategies

Lightweight models

Similarity

Diagnostics

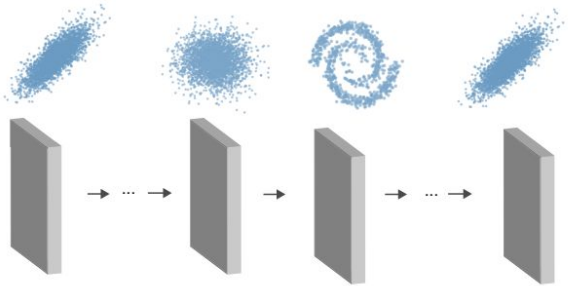
Biological plausibility

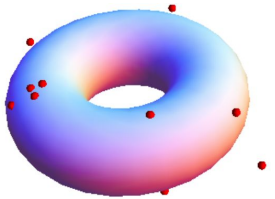
Transfer learning

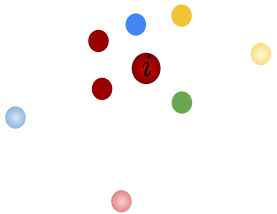
Alignment

Data “cartography”

What we will do with representations?







Geometry (dimension)

Transfer learning

“Shape” of data

Computational strategies

Information

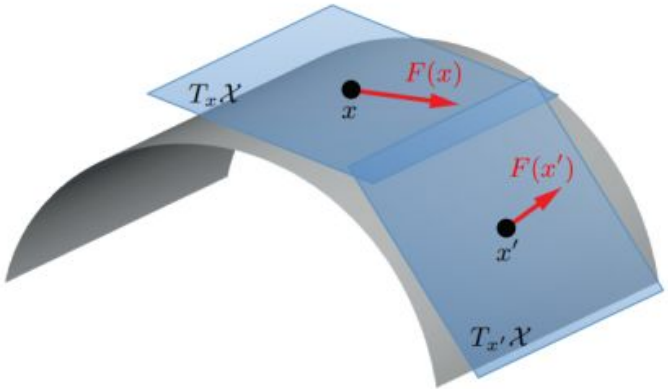
Similarity

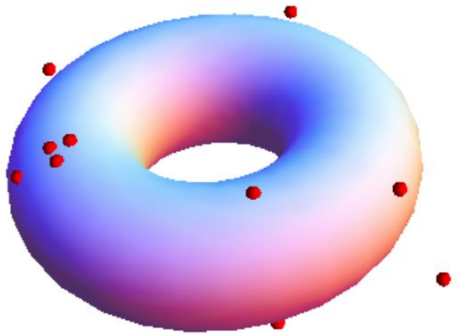
Transfer learning

Computational strategies

The Manifold Hypothesis

Most “naturally occurring” datasets lie on a low-dimensional manifold



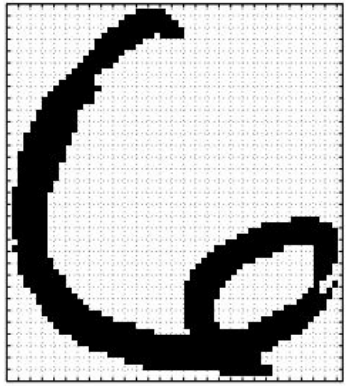


M. Bronstein et al., Geometric Deep Learning: Going beyond Euclidean data, 2017

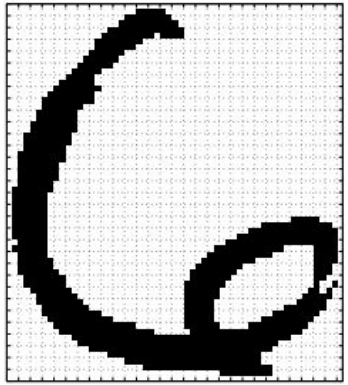
Fefferman et al., Testing the manifold hypothesis, 2016

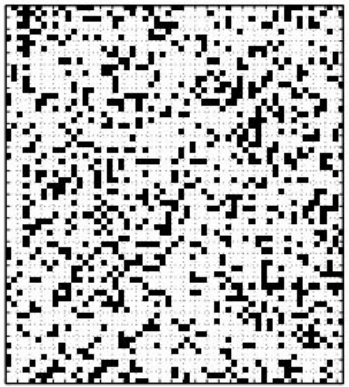
Experimental evidences of the Manifold Hypothesis

Experimental evidences of the Manifold Hypothesis

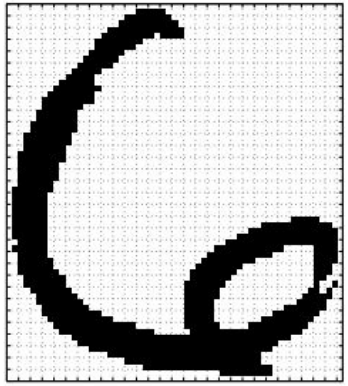


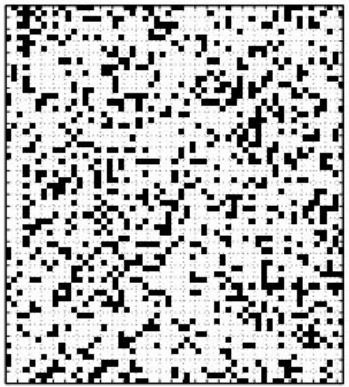
Experimental evidences of the Manifold Hypothesis

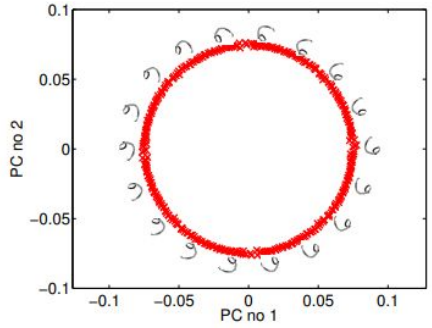




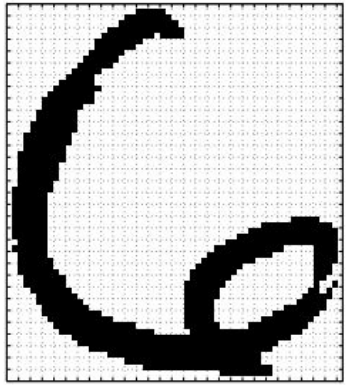
Experimental evidences of the Manifold Hypothesis

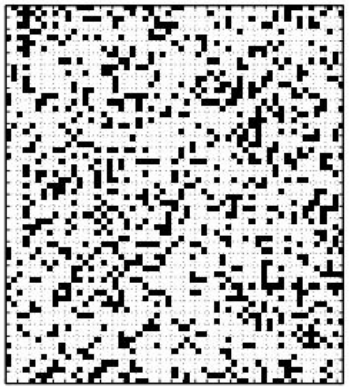


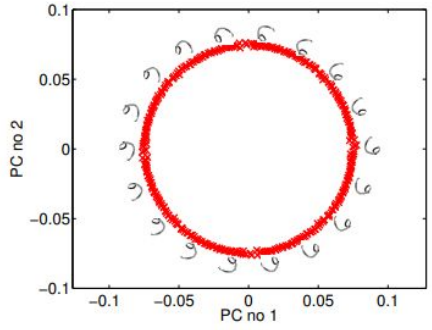


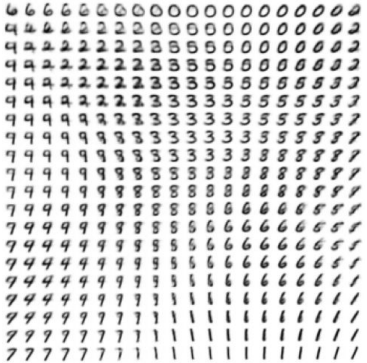


Experimental evidences of the Manifold Hypothesis



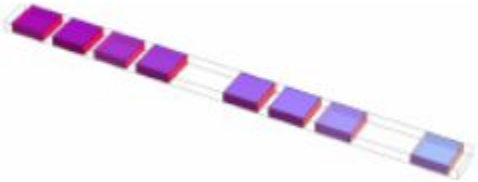




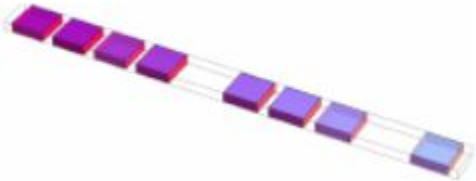


The curse of dimensionality

The curse of dimensionality



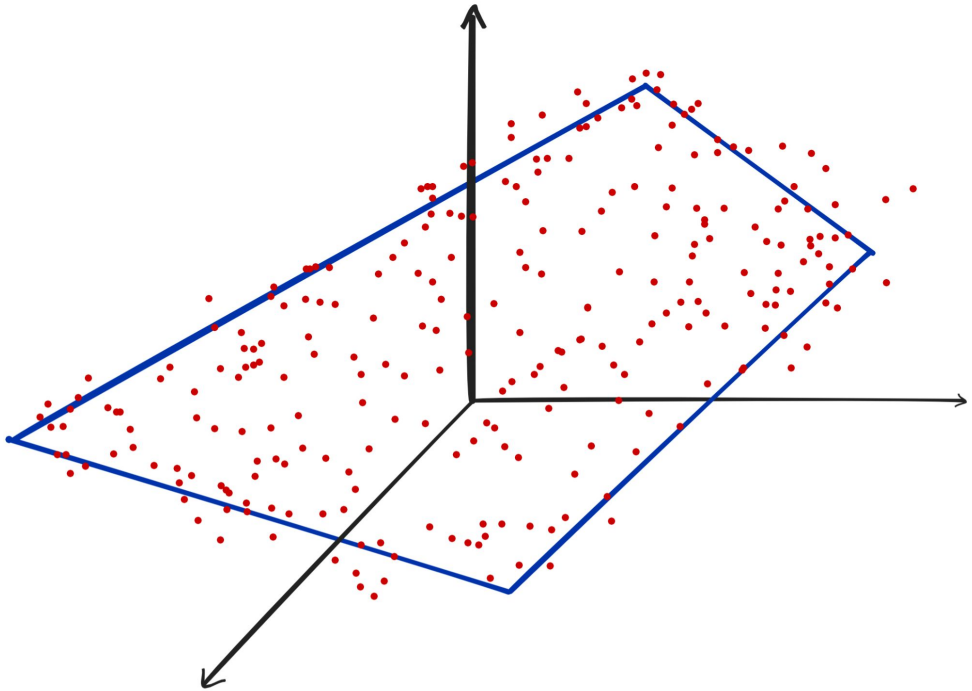
The curse of dimensionality

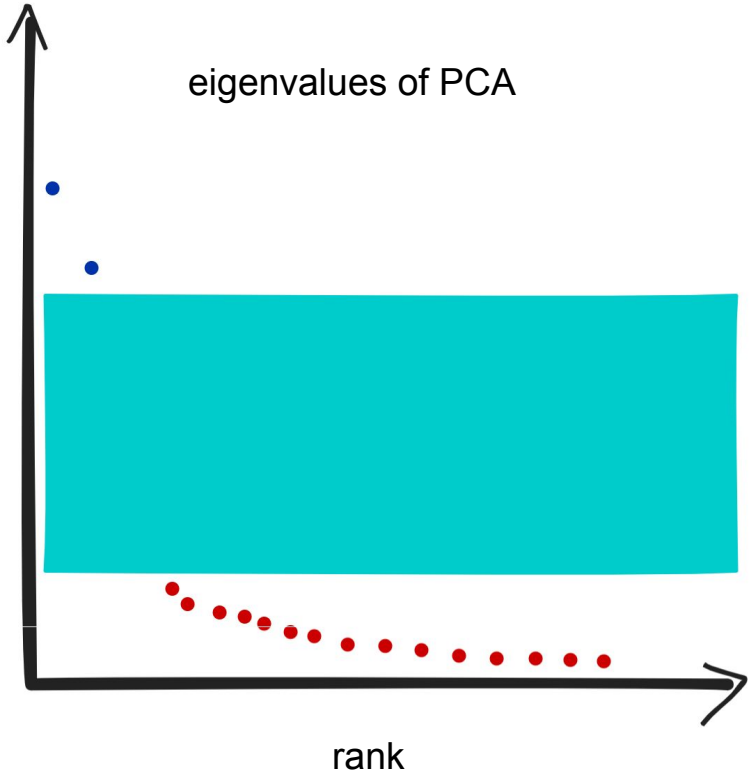


The curse of dimensionality

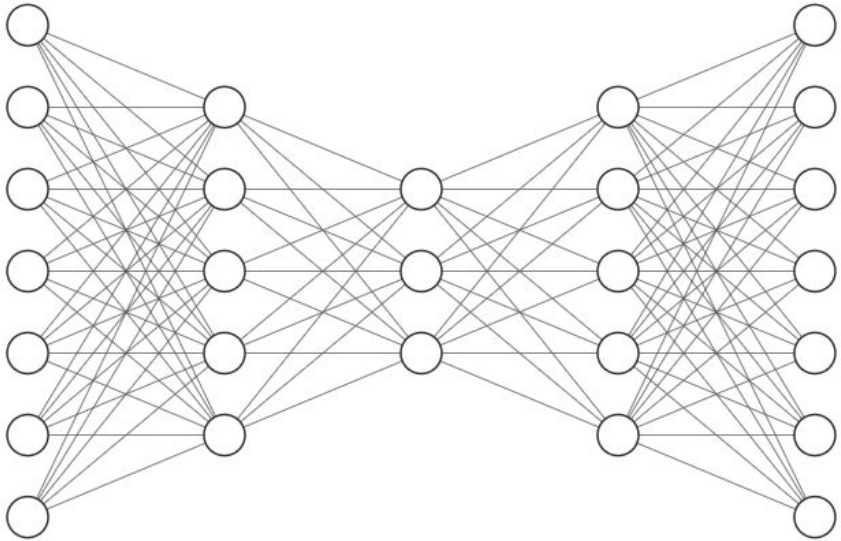
(Figure by Nicolas Chapados)

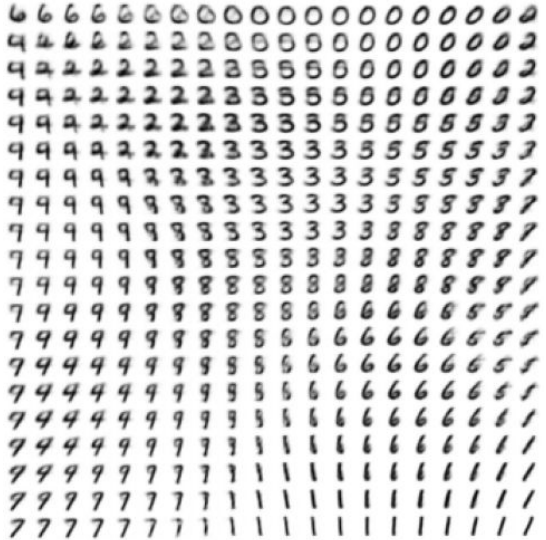
Intrinsic dimension with PCA



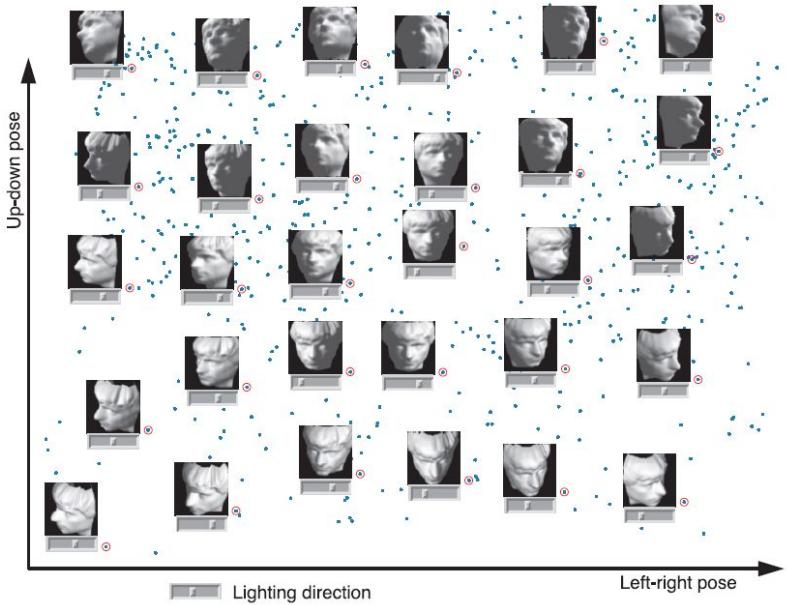


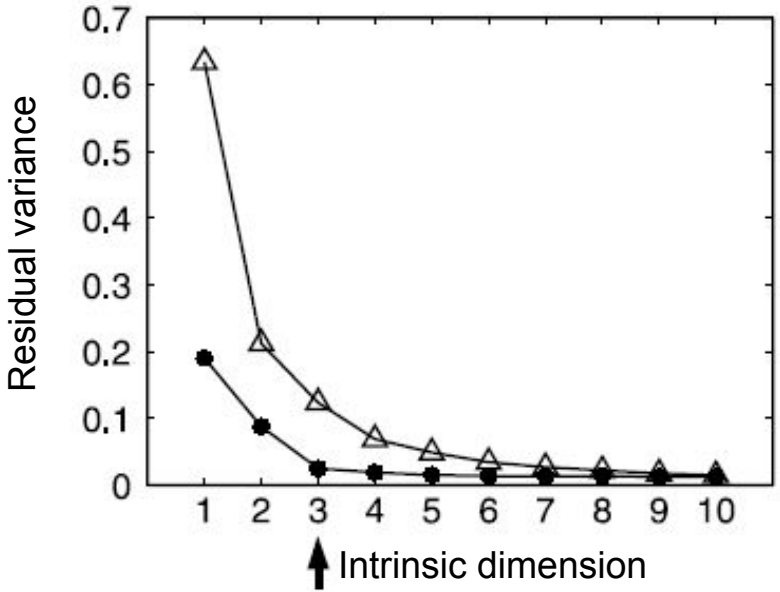
Autoencoders and non-linear PCA





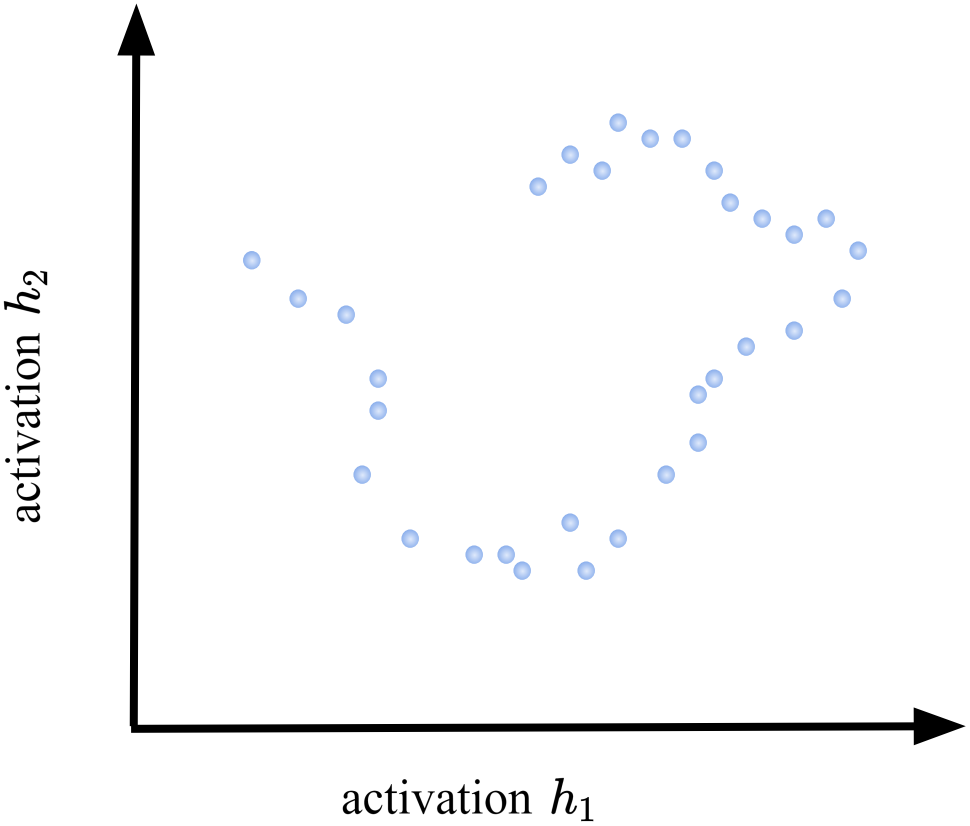
ISOMAP





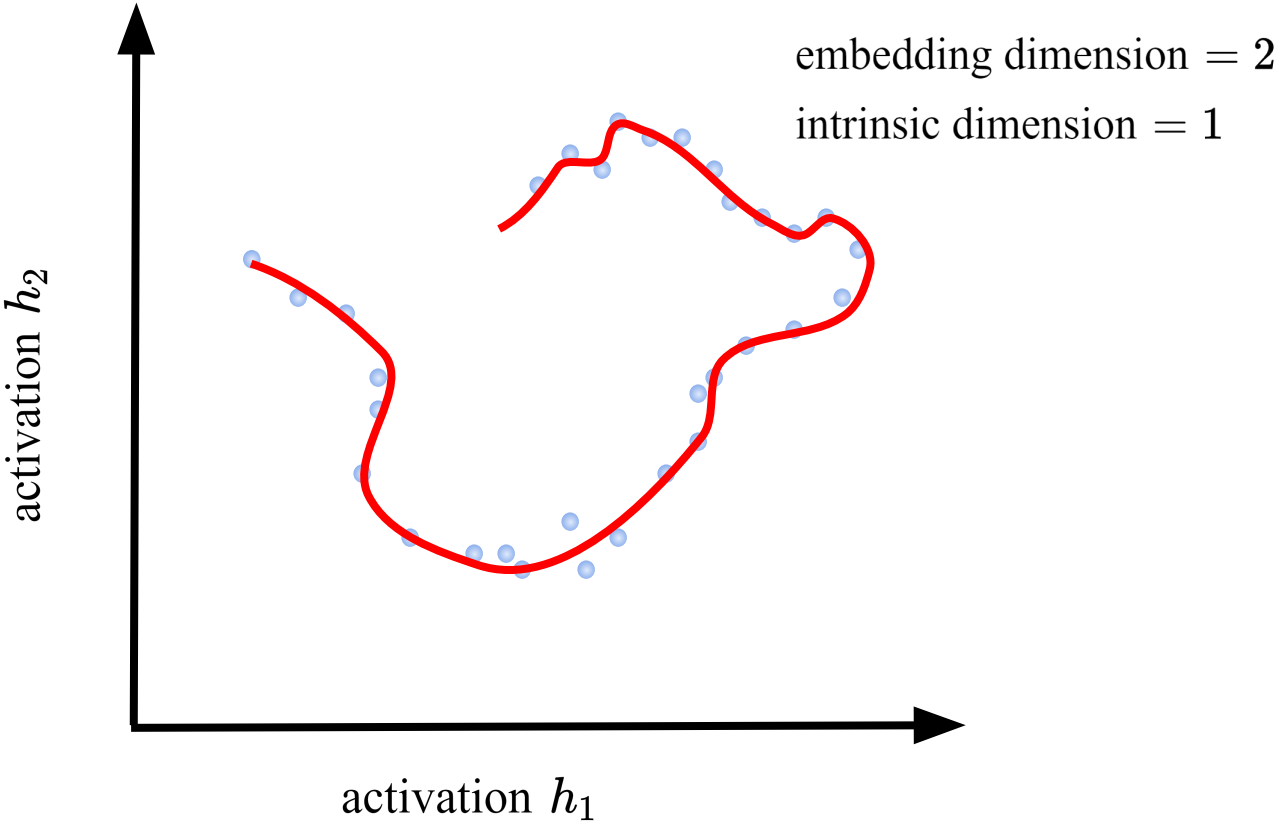
J. Tenenbaum, A global geometric framework for nonlinear dimensionality reduction, 2000

Intrinsic dimension with local information (Two-NN)

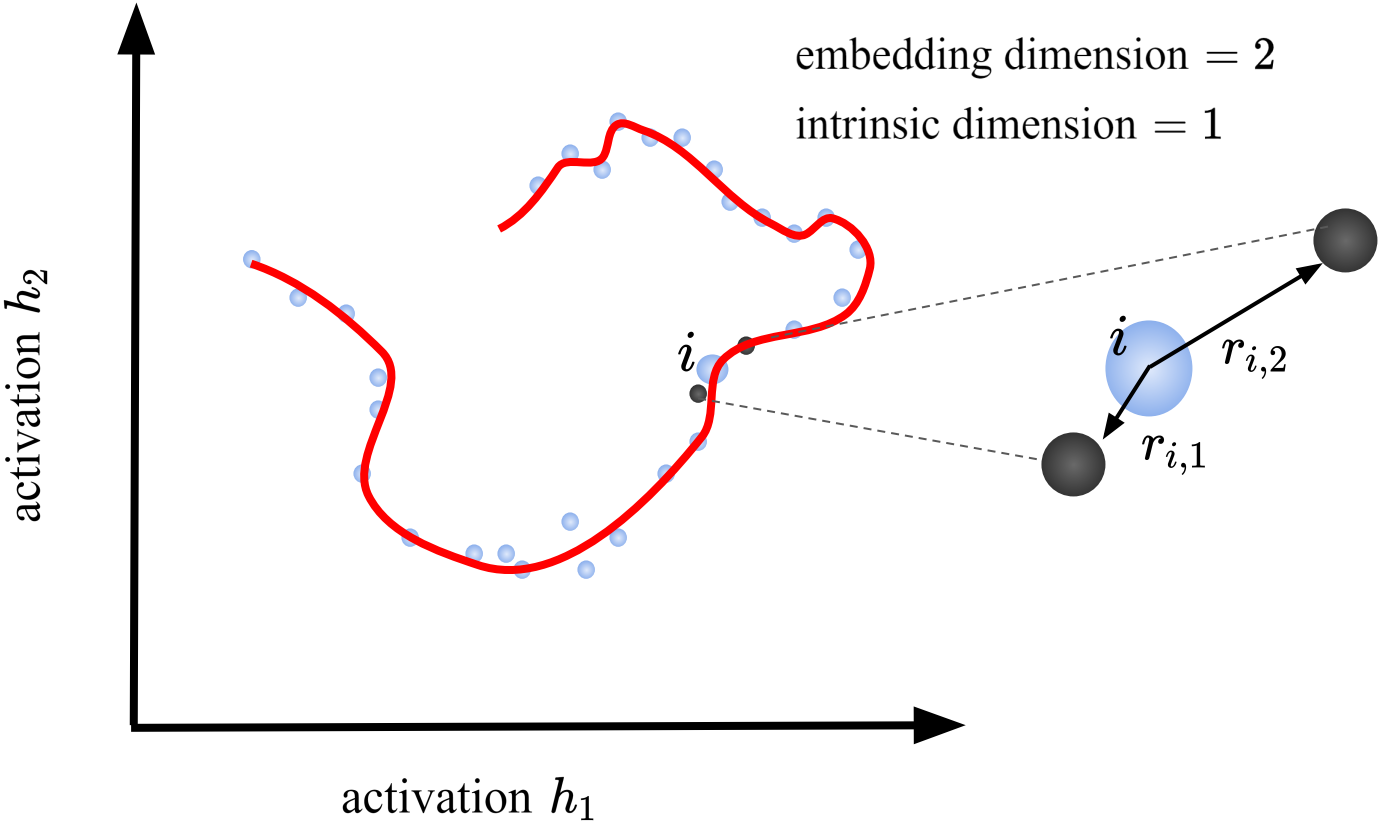


E. Facco et al., “Estimating Estimating the intrinsic dimension of datasets by a minimal neighborhood information”, 2017

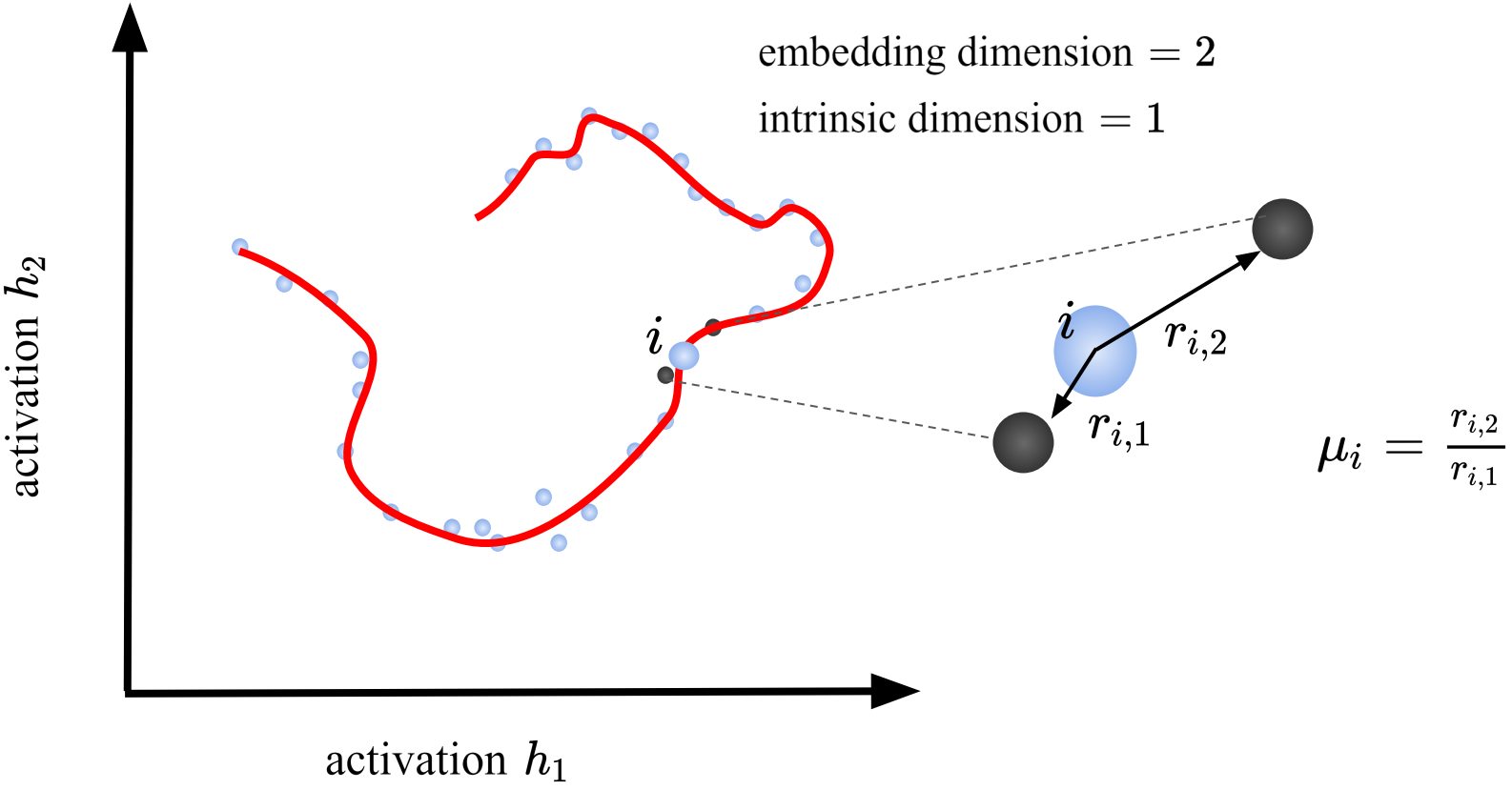
Intrinsic dimension with local information (Two-NN)



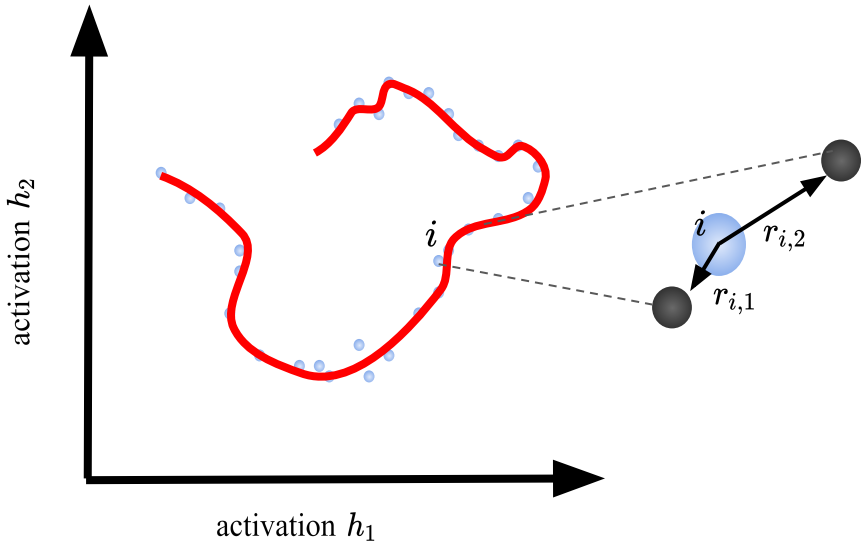
Intrinsic dimension with local information (Two-NN)



Intrinsic dimension with local information (Two-NN)

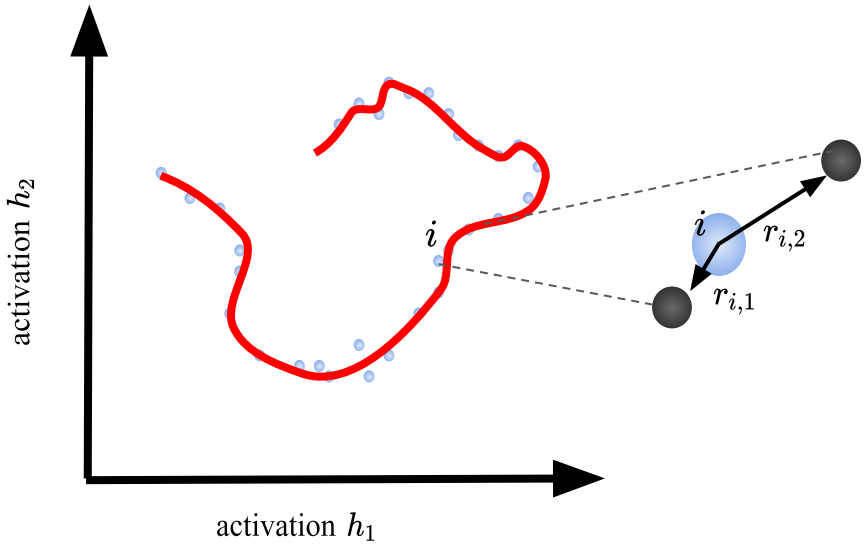


Two-NN algorithm

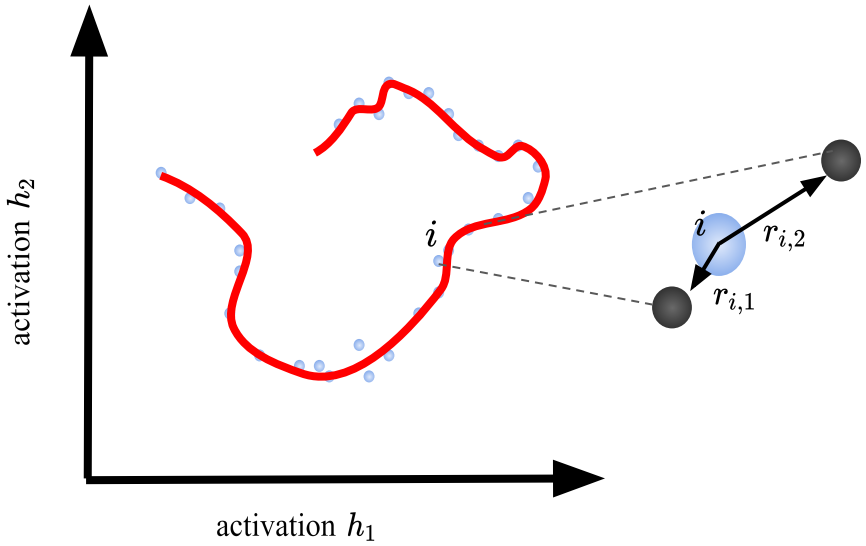


Two-NN algorithm

1) For each data point *i* compute the distance to its first and second neighbour



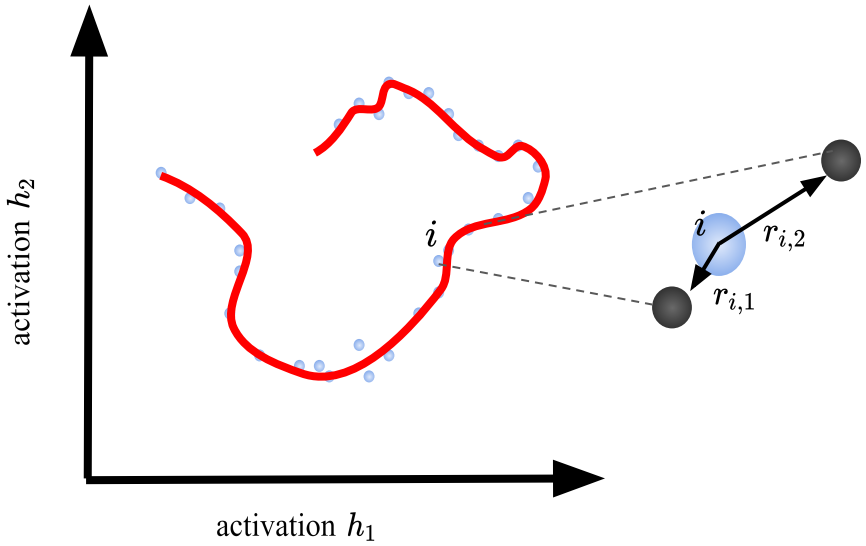
Two-NN algorithm



1) For each data point *i* compute the distance to its first and second neighbour

2) For each *i* compute $\mu_i = \frac{r_{i,2}}{r_{i,1}}$

Two-NN algorithm



1) For each data point *i* compute the distance to its first and second neighbour

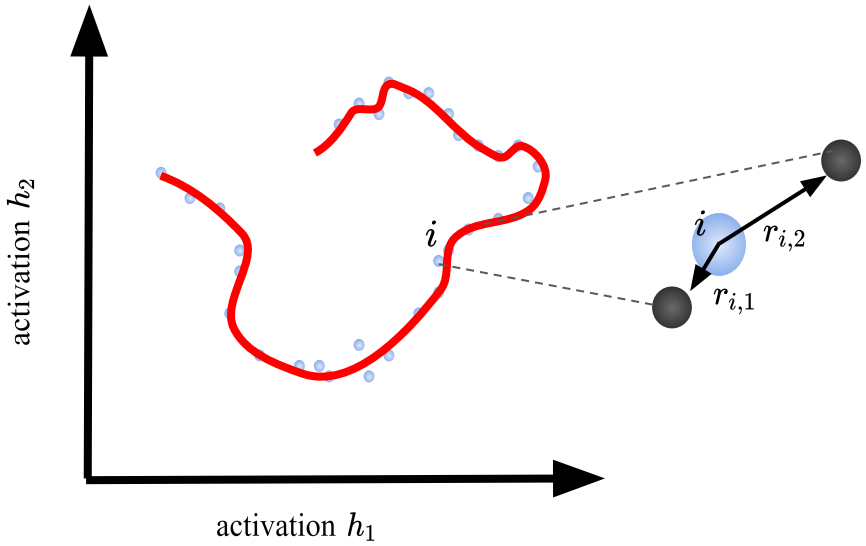
2) For each *i* compute $\mu_i = \frac{r_{i,2}}{r_{i,1}}$

3) The probability distribution of μ is:

$$P(\mu|d) = \frac{d}{\mu^{d+1}}$$

where d is the intrinsic dimension

Two-NN algorithm



1) For each data point *i* compute the distance to its first and second neighbour

2) For each *i* compute $\mu_i = \frac{r_{i,2}}{r_{i,1}}$

3) The probability distribution of μ is:

$$P(\mu|d) = \frac{d}{\mu^{d+1}}$$

where d is the intrinsic dimension

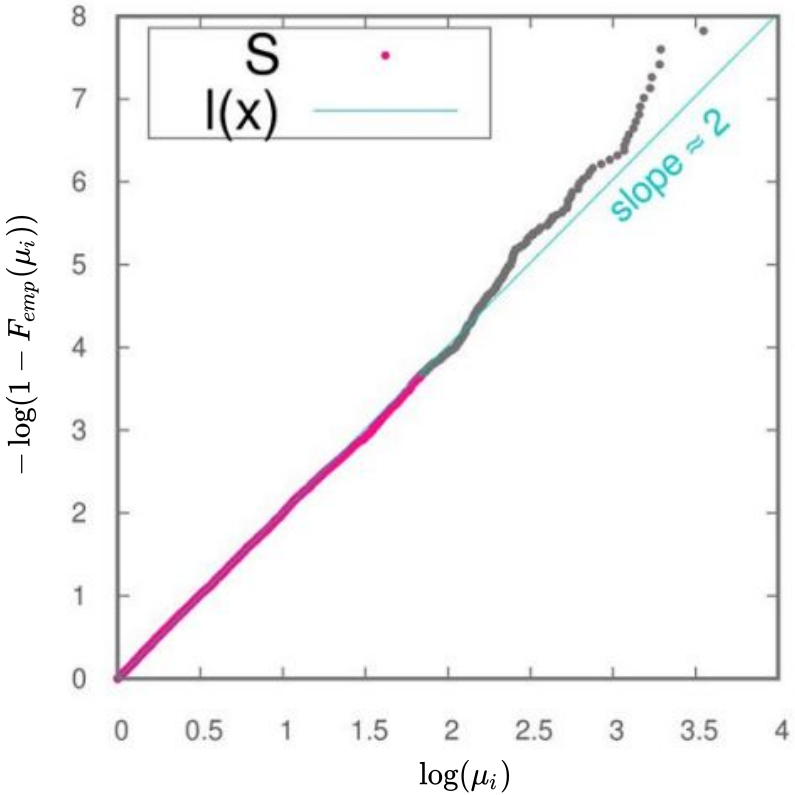
4) Infer d via maximum likelihood:

$$\log L(\{\mu_i\}|d) = \log \prod_i P(\mu_i|d)$$
$$\partial_d \log L = 0 \Rightarrow \hat{d} = \frac{N}{\sum_i \log \mu_i}$$

Two-NN algorithm

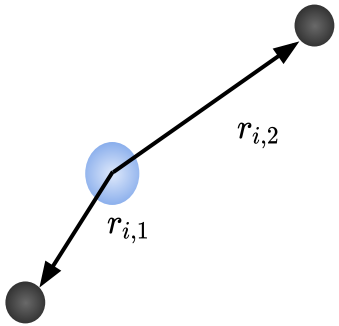
3) Compute the empirical cumulate of $F_{emp}(\mu)$ by sorting the values of μ in an ascending order through a permutation σ, then define $F_{emp}(\mu_{\sigma(i)}) = \frac{i}{N}$

4) Fit the points of the plane given by coordinates $\log(\mu_i), -\log(1 - F_{emp}(\mu_i)) \quad | \quad i = 1, 2, \ldots, N$ with a straight line passing through the origin



Hypothesis underlying the Two-NN

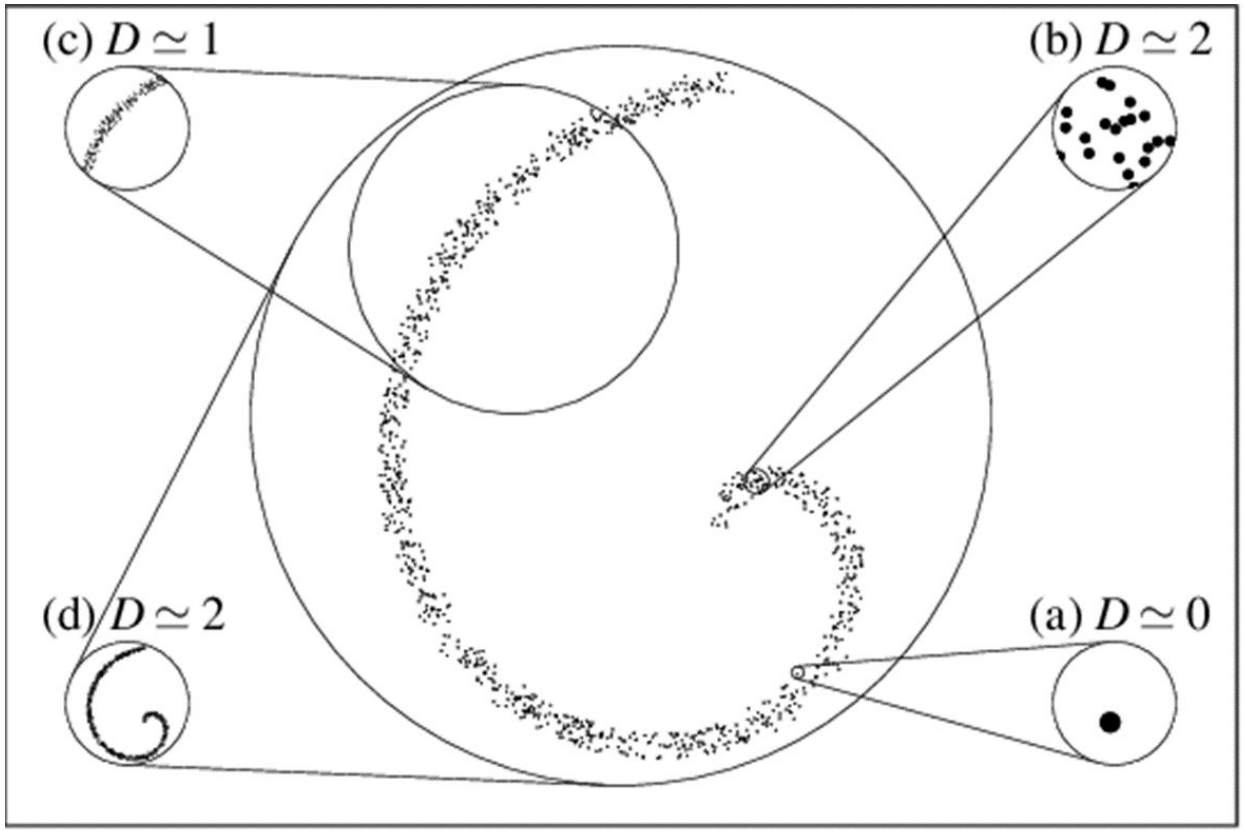
Theorem: $P(\mu|d) = \frac{d}{\mu^{d+1}}$



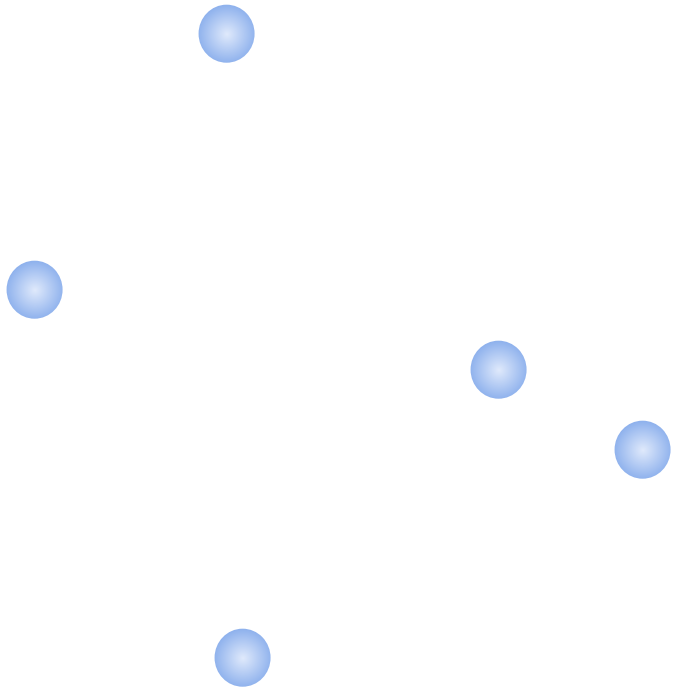
$\rho \approx const$

“on the lengthscale defned by the typical distance to the second neighbor”

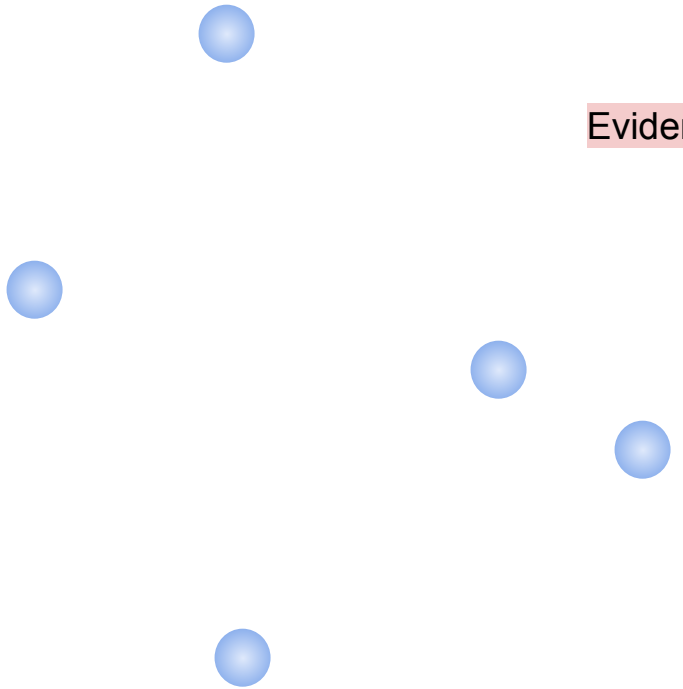
The importance of scale and persistence



A simple method to investigate scale: decimation

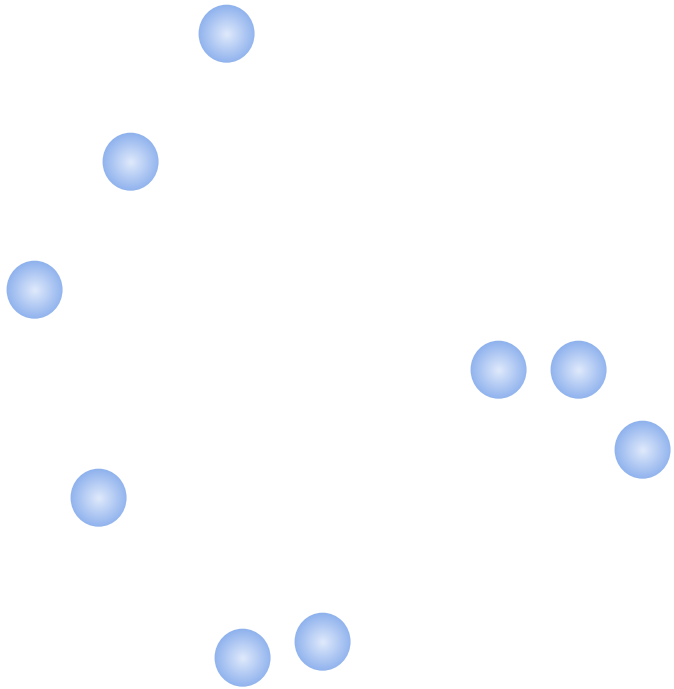


A simple method to investigate scale: decimation

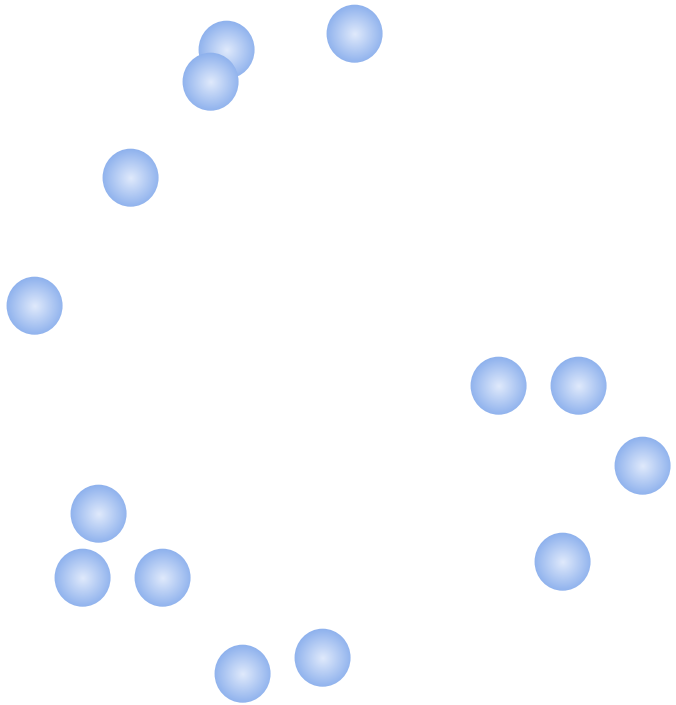


Evidence: 2-dimensional data

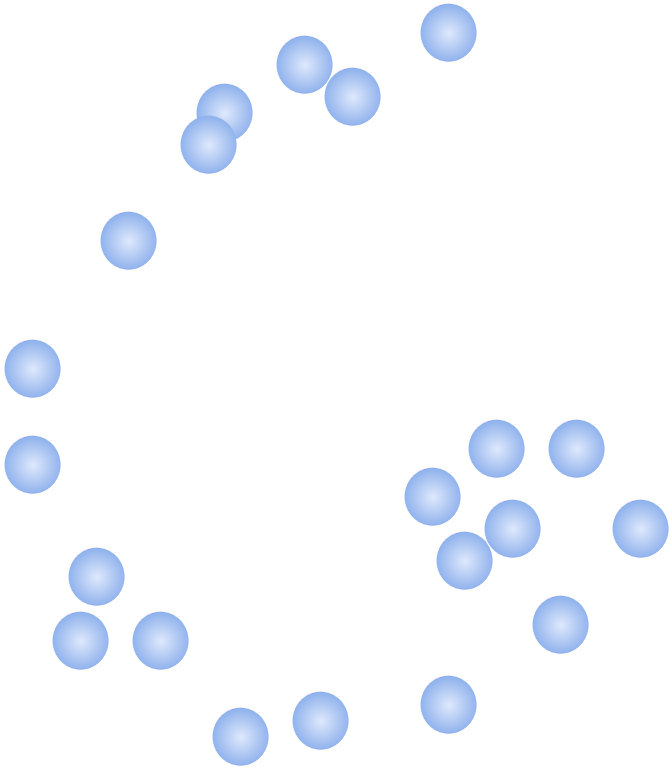
A simple method to investigate scale: decimation



A simple method to investigate scale: decimation



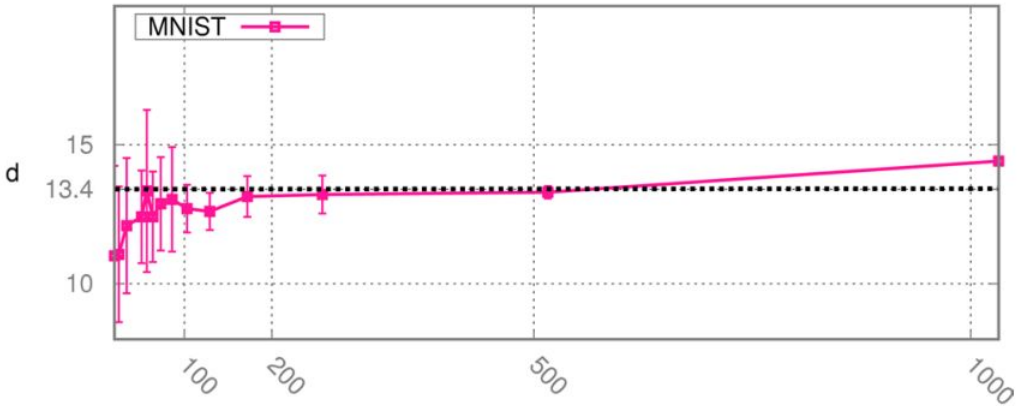
A simple method to investigate scale: decimation

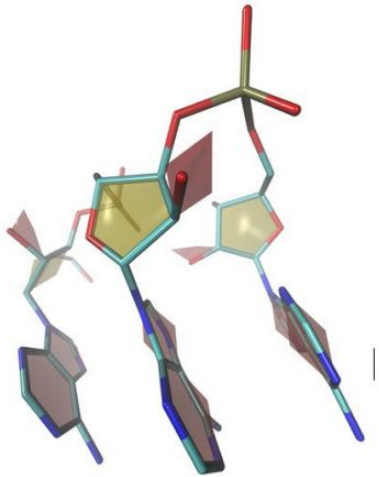


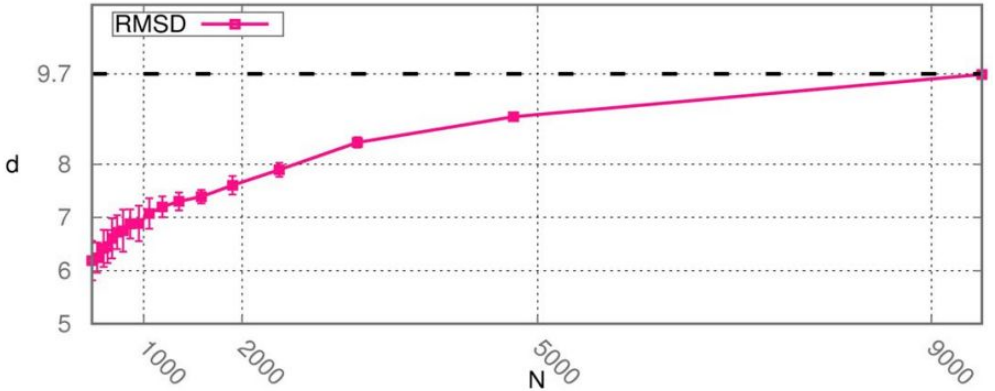
A simple method to investigate scale: decimation

Evidence: 1-dimensional data

A simple method to investigate scale: decimation







MNIST vs KMNIST

$$d_{\text{MNIST}} \approx 13 - 14$$

MNIST vs KMNIST

$d_{\text{MNIST}} \approx 13 - 14$

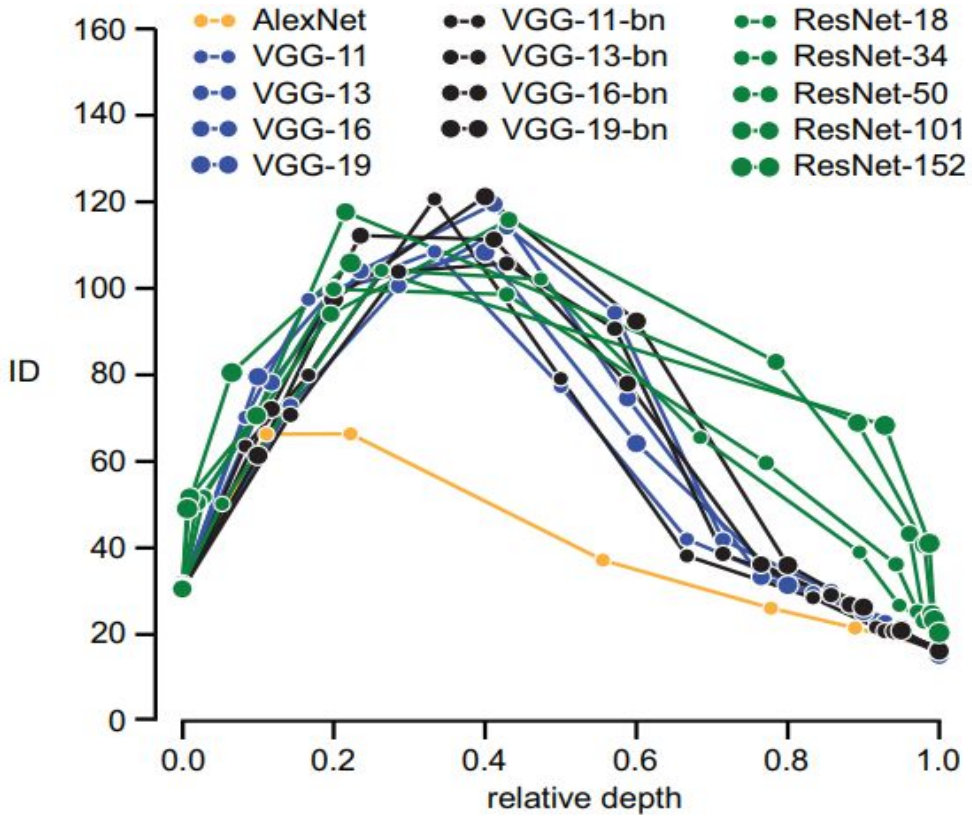
$d_{\text{KMNIST}} \approx \ ?$

MNIST vs KMNIST

$d_{\text{MNIST}} \approx 13 - 14$

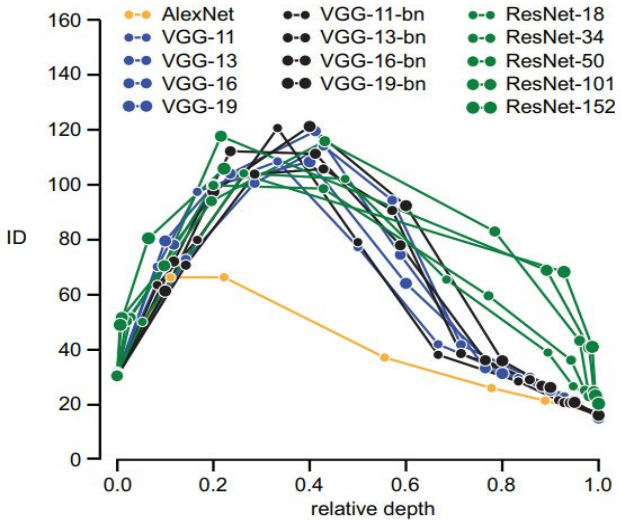
$d_{\text{KMNIST}} \approx 17 - 18$

Intrinsic dimension expansion and contraction

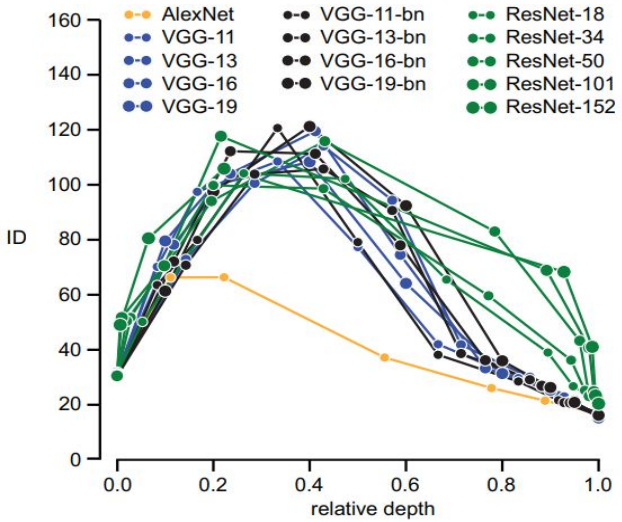


A. Ansuini et al., “Intrinsic dimension of data representations in deep neural networks”, 2019

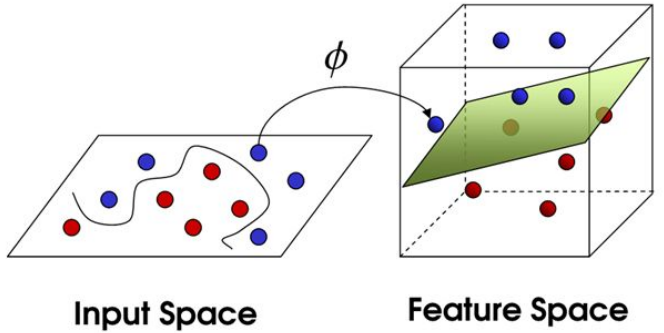
Why the dimensionality expands?



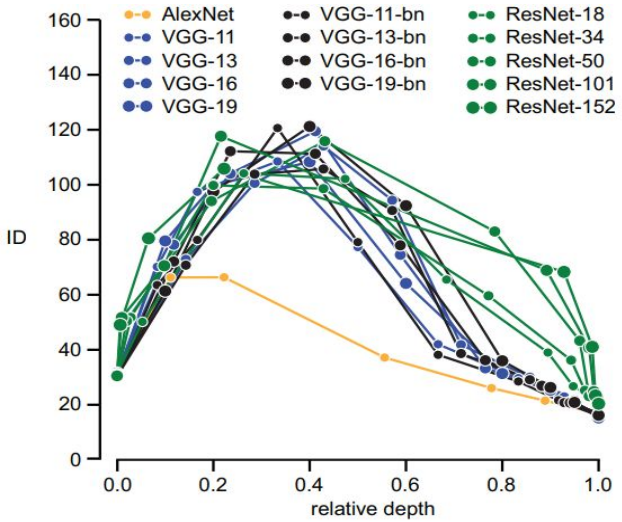
Why the dimensionality expands?



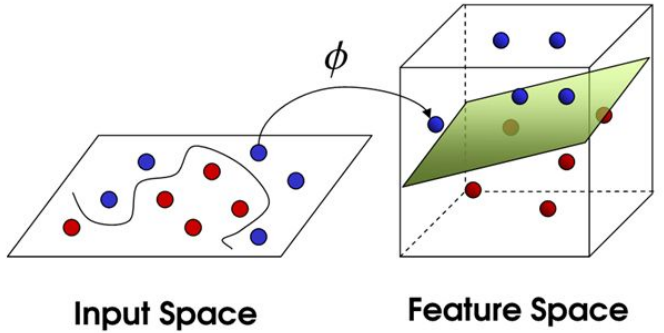
The kernel trick



Why the dimensionality expands?

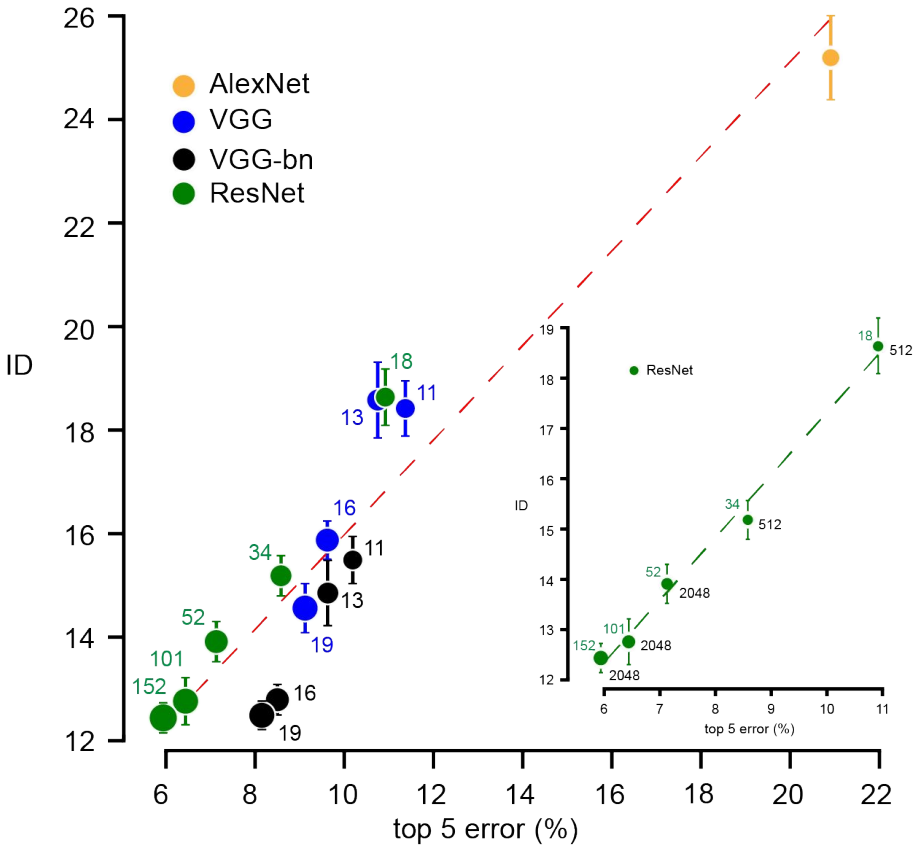


The kernel trick

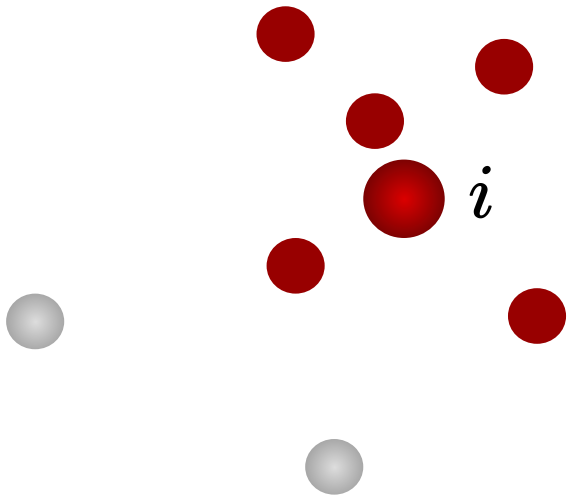


Cover's Theorem (1965): In sufficiently high dimension any dataset is *linearly separable*

Dimensionality contraction and performance



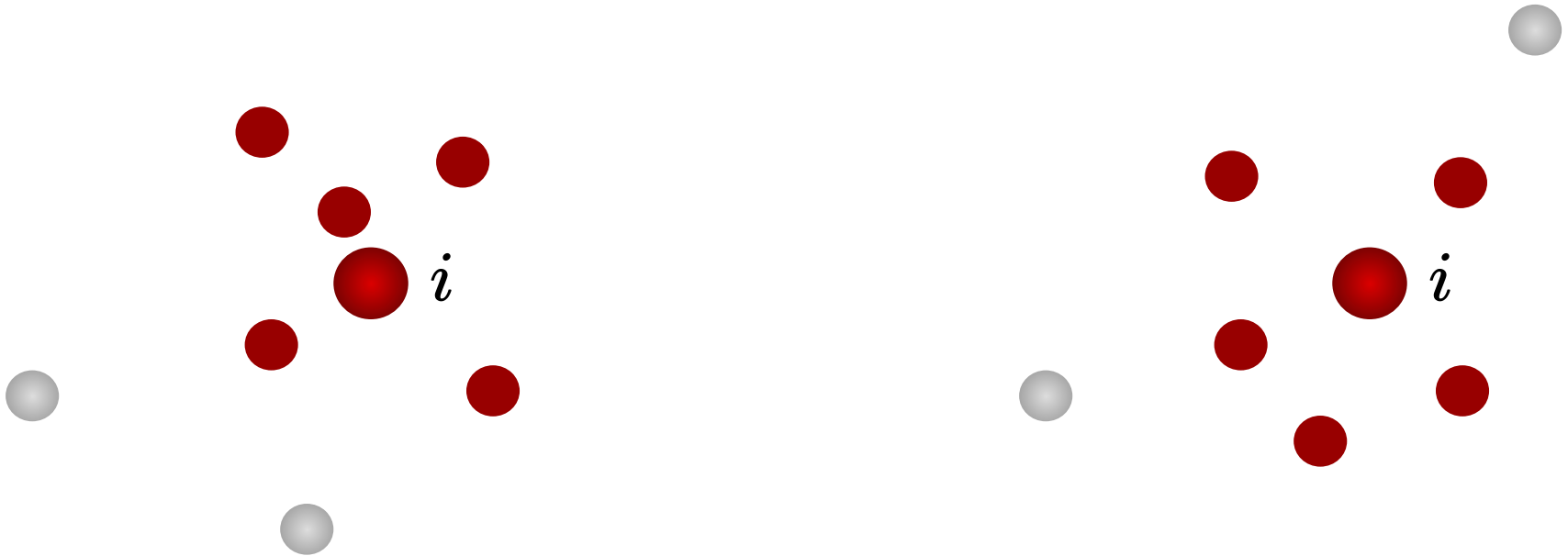
Local geometry: the neighbours “structure”



$$K = 5$$

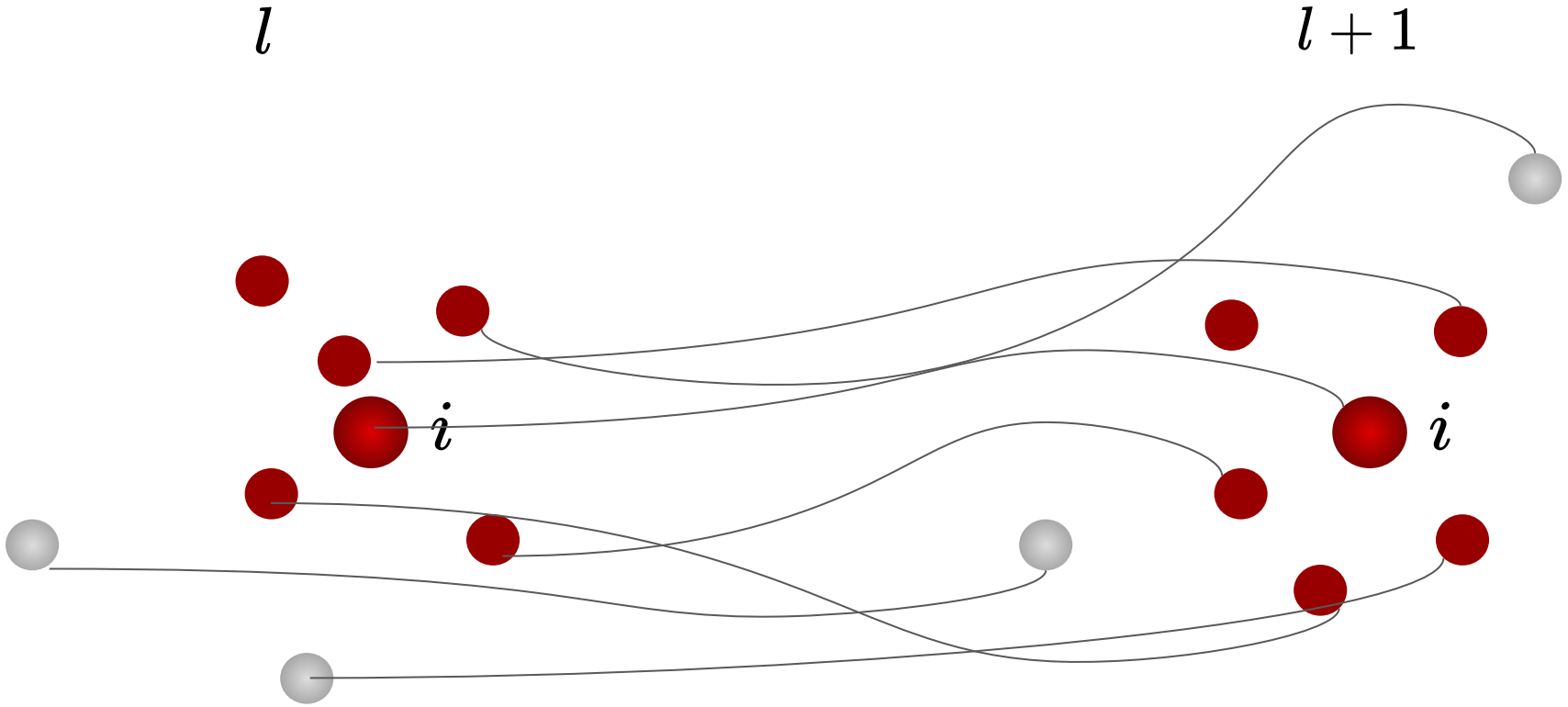
Local geometry: the neighbours “structure”

l $l+1$



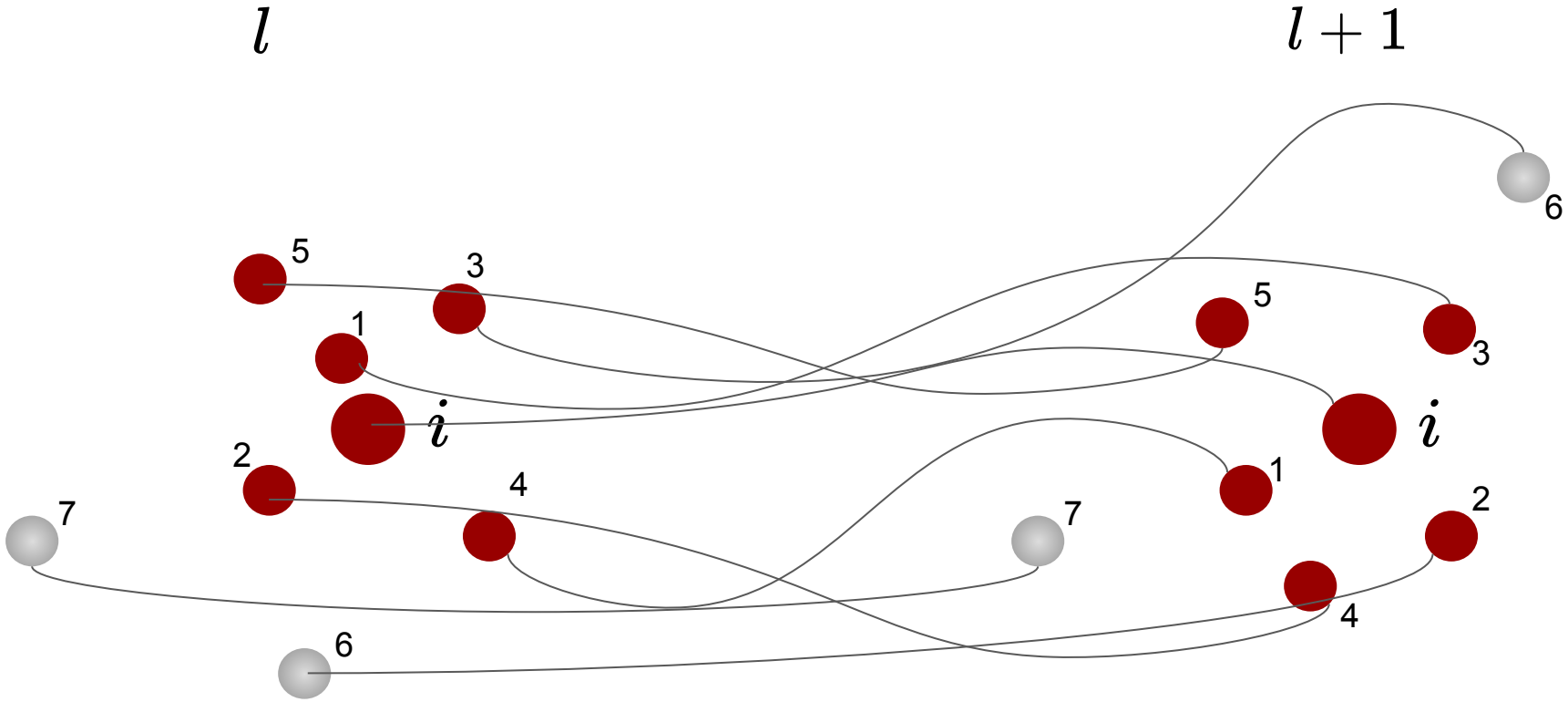
$K = 5$

Local geometry



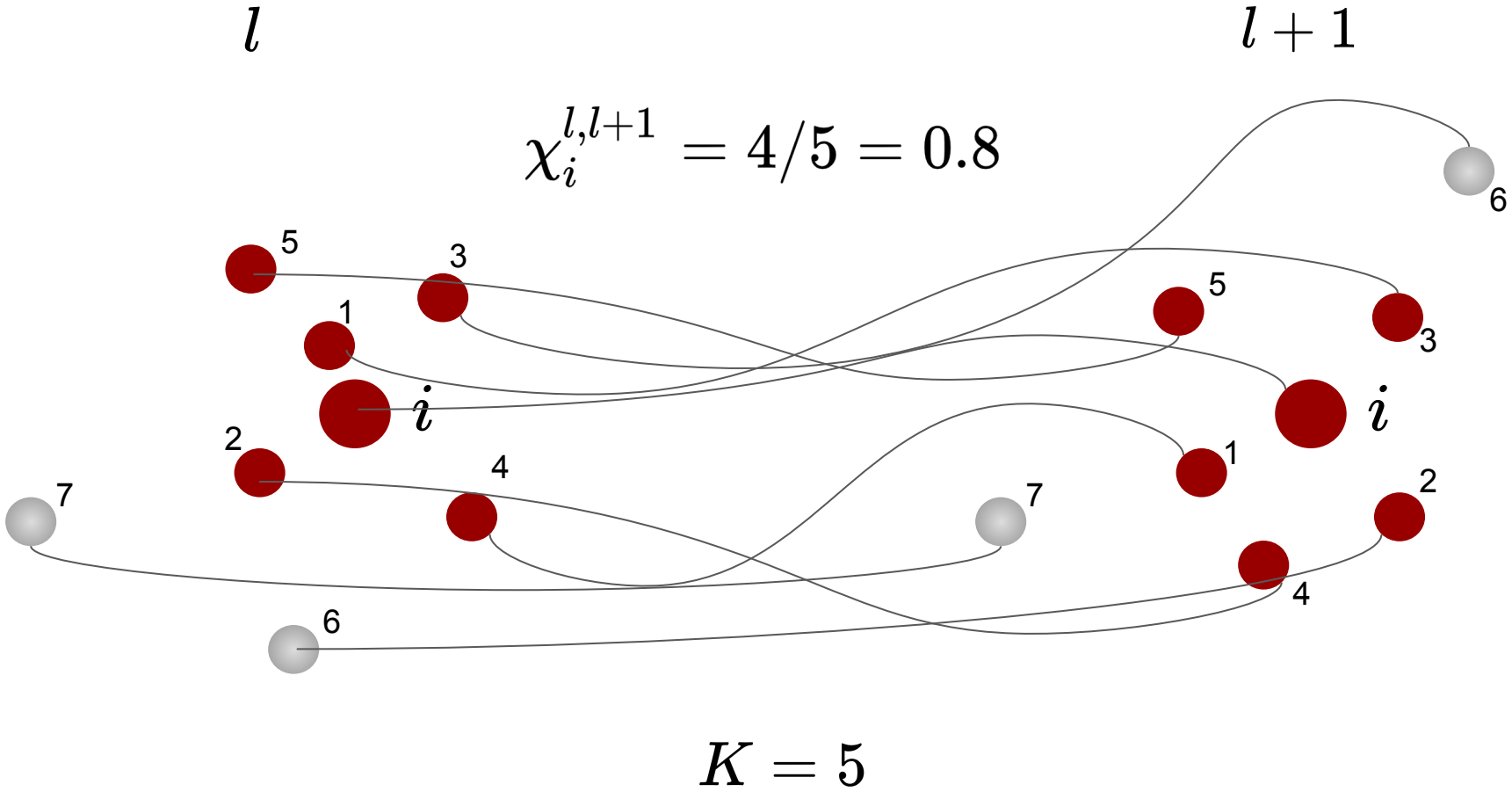
$K = 5$

Local geometry



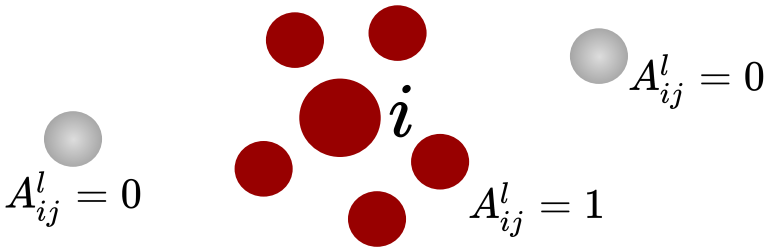
$K = 5$

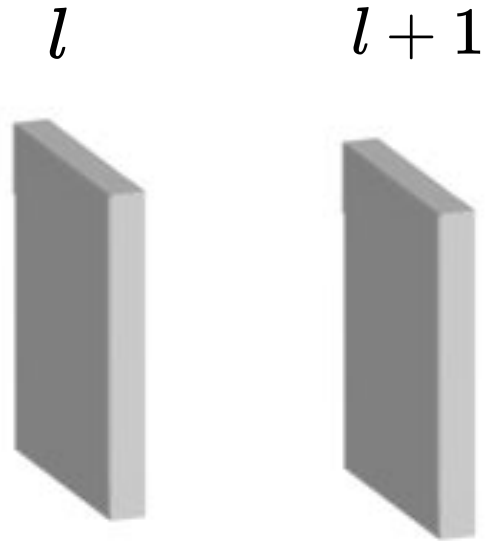
Local geometry



NO as a mesasure of representation similarity

$$\mathcal{N}^l_{k=5}(i)$$

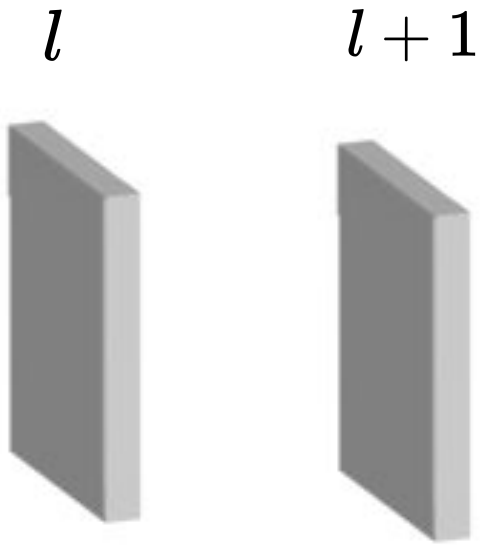




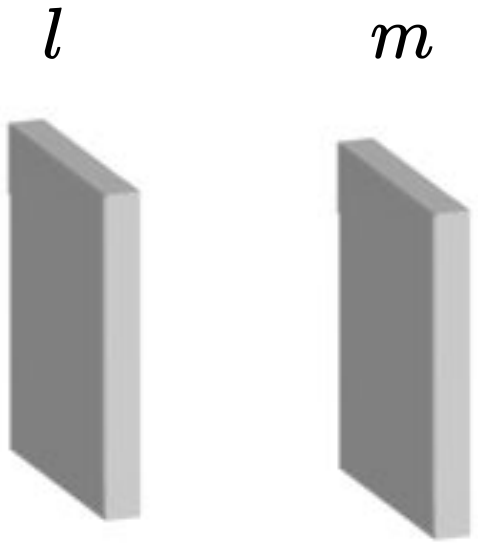
$$A^l_{ij}=1 \;\text{ iff }\; j\in\mathcal{N}^l_k(i)$$

$$\chi^{l,l+1}_k=\frac{1}{N}\sum_i\frac{1}{k}\sum_j A^l_{ij}A^{l+1}_{ij}$$

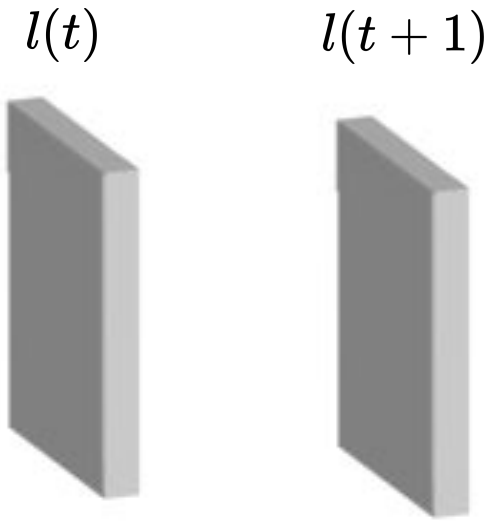
NO as a mesasure of representation similarity



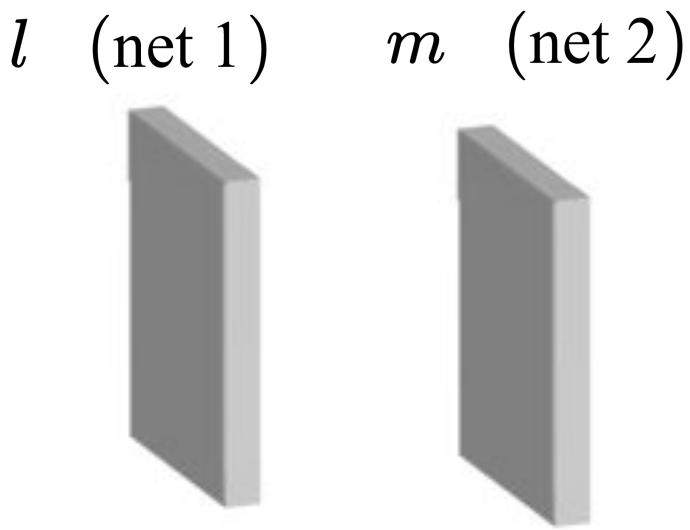
NO as a mesasure of representation similarity



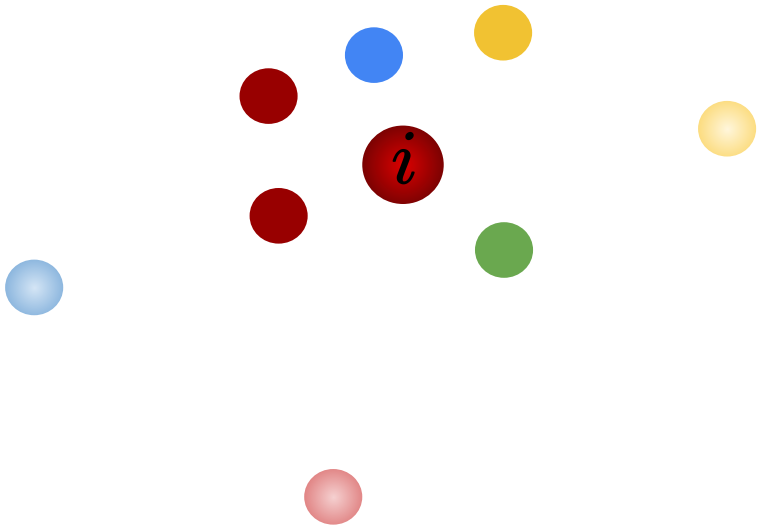
NO as a mesasure of representation similarity



NO as a mesasure of representation similarity



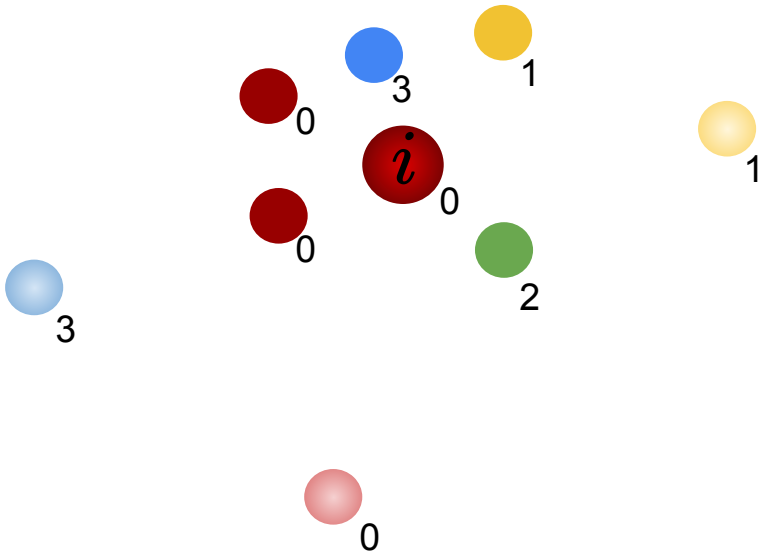
NO as a mesasure of information content



NO as a mesasure of information content

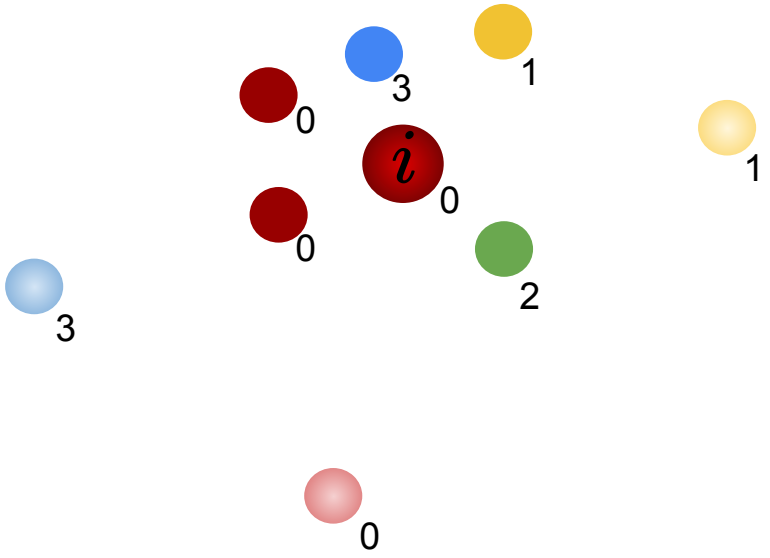
Labelled dataset

(x_1, y_1)
(x_2, y_2)
$\dots$
$(x_i, 0)$
$\dots$
(x_N, y_N)



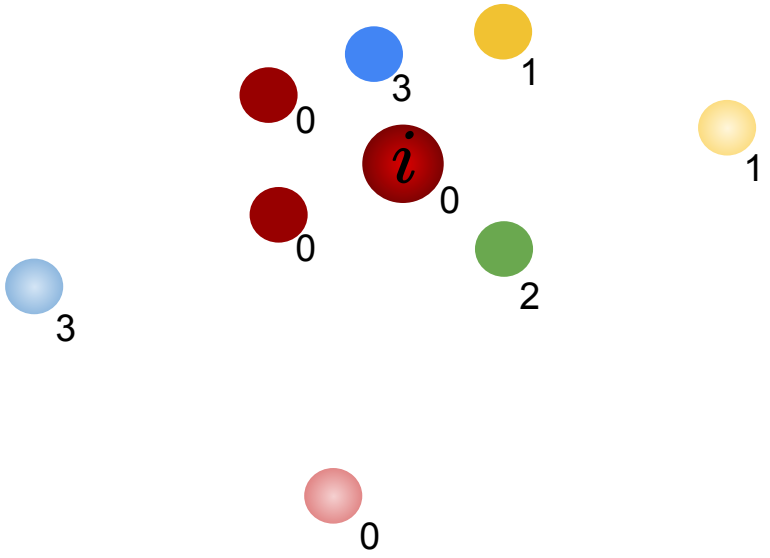
How many neighbors are in the same class (on average)?

NO as a mesasure of information content



$$\chi_i^{l,gt} = 2/5 = 0.4$$

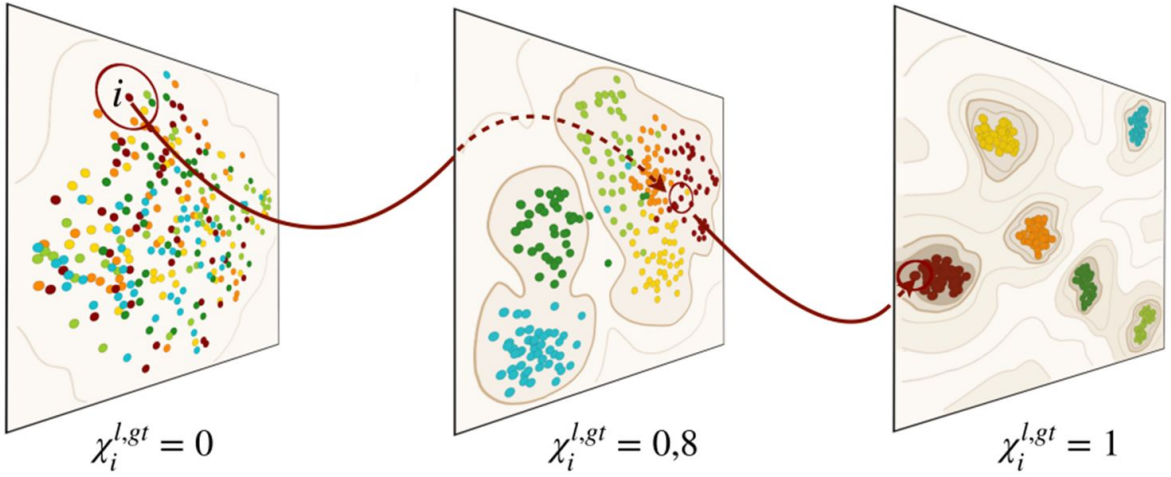
NO as a mesasure of information content

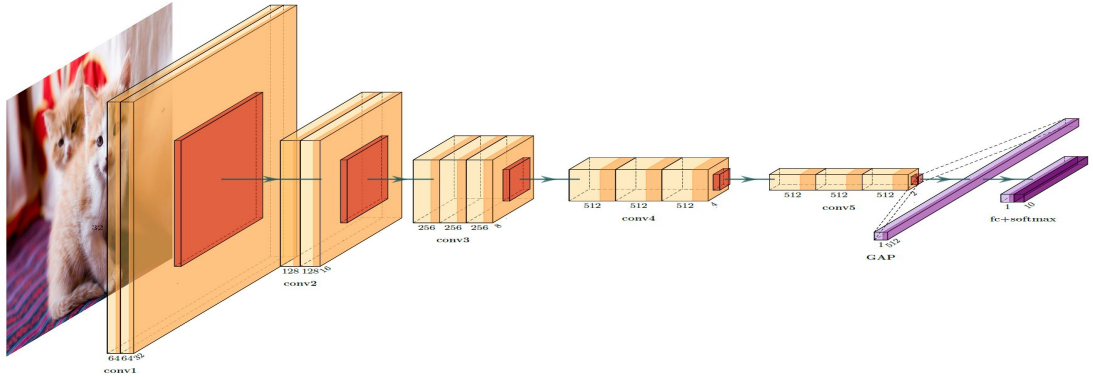


$$A_{ij}^{gt} = 1 \quad \text{iff gt(i) = gt(j), otherwise = 0}$$

$$\chi^{l,gt} = \frac{1}{N} \sum_i \frac{1}{k} \sum_j A_{ij}^{l} A_{ij}^{gt}$$

Neighbours rearrangements



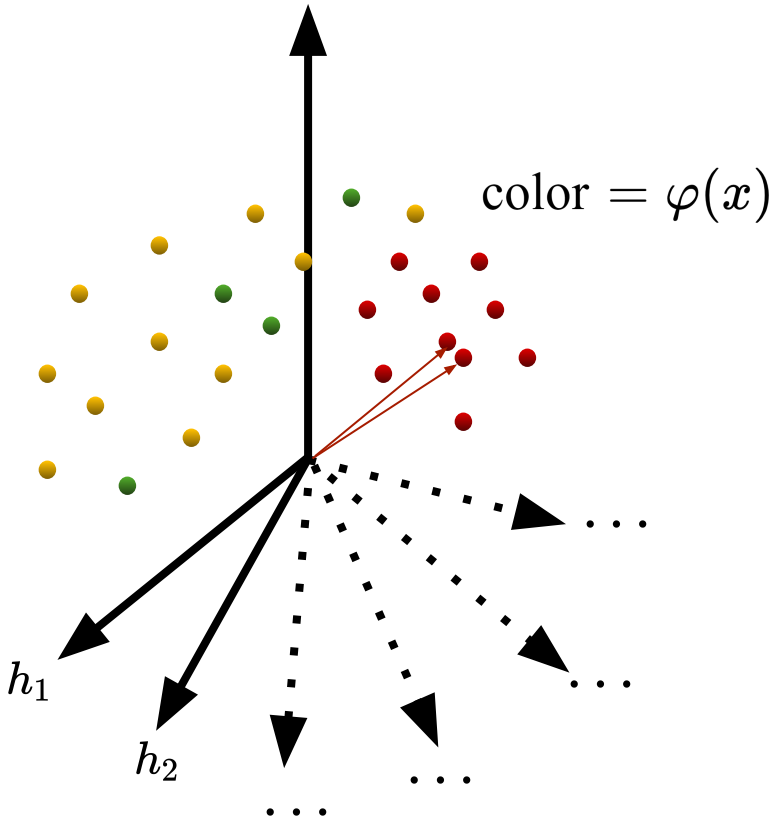


NO with general labelling

The class label is only an example

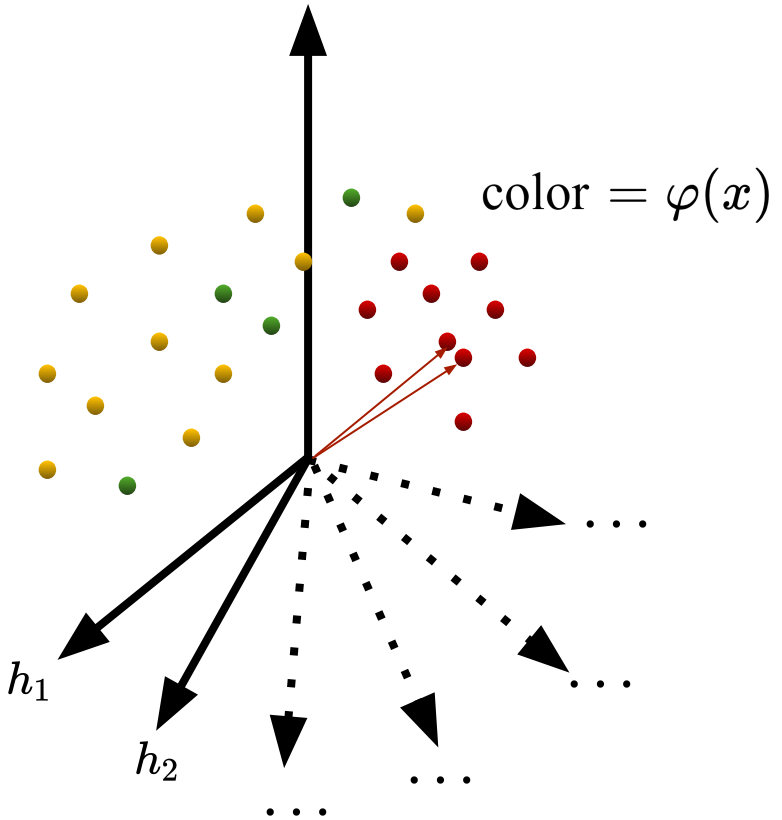
NO with general labelling

The class label is only an example



NO with general labelling

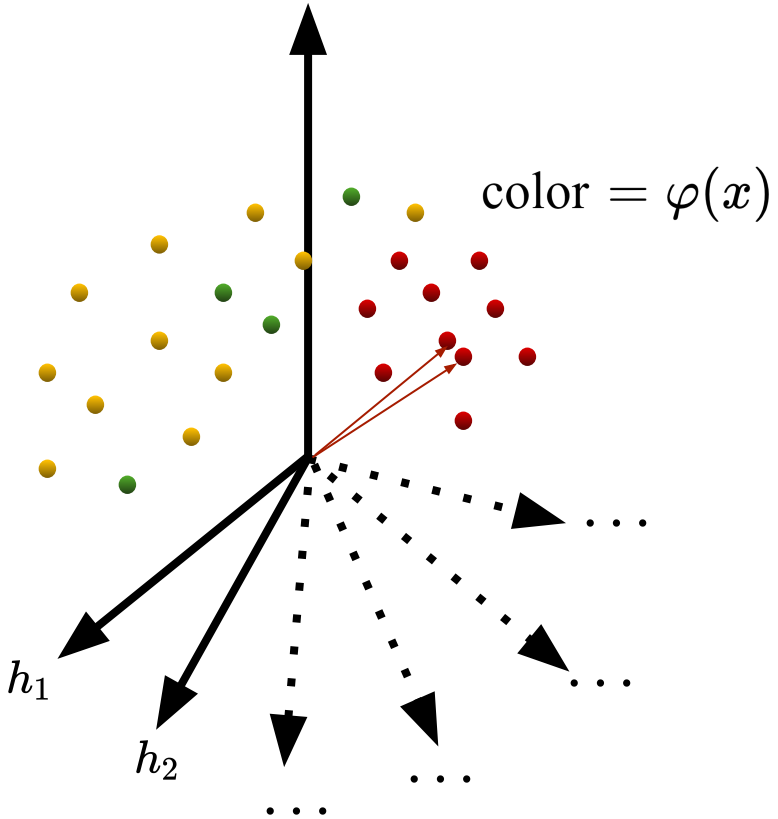
The class label is only an example



Examples of φ :

NO with general labelling

The class label is only an example

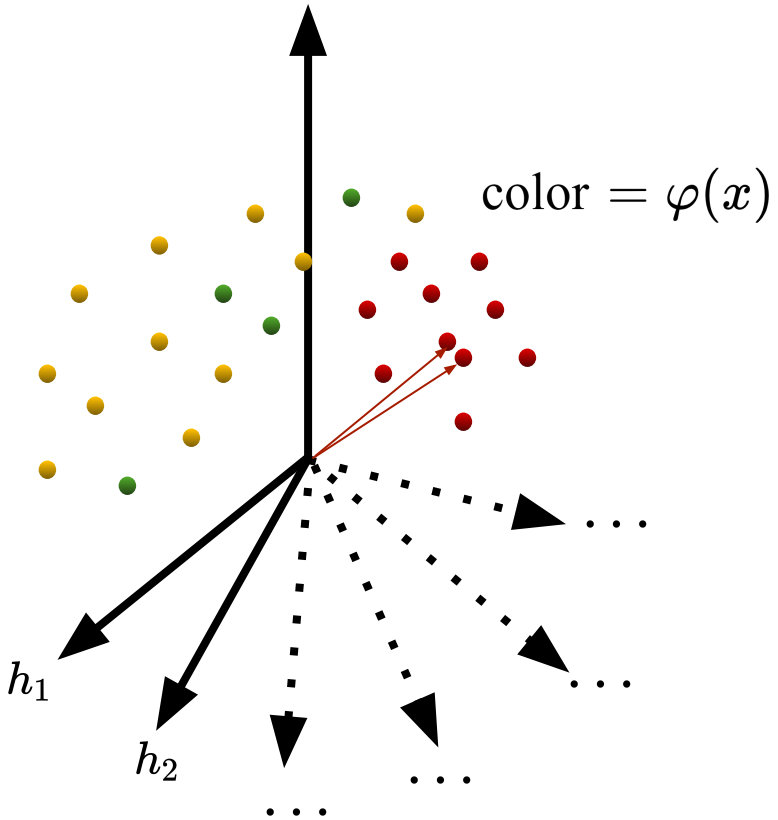


Examples of φ :

- luminance of an image

NO with general labelling

The class label is only an example

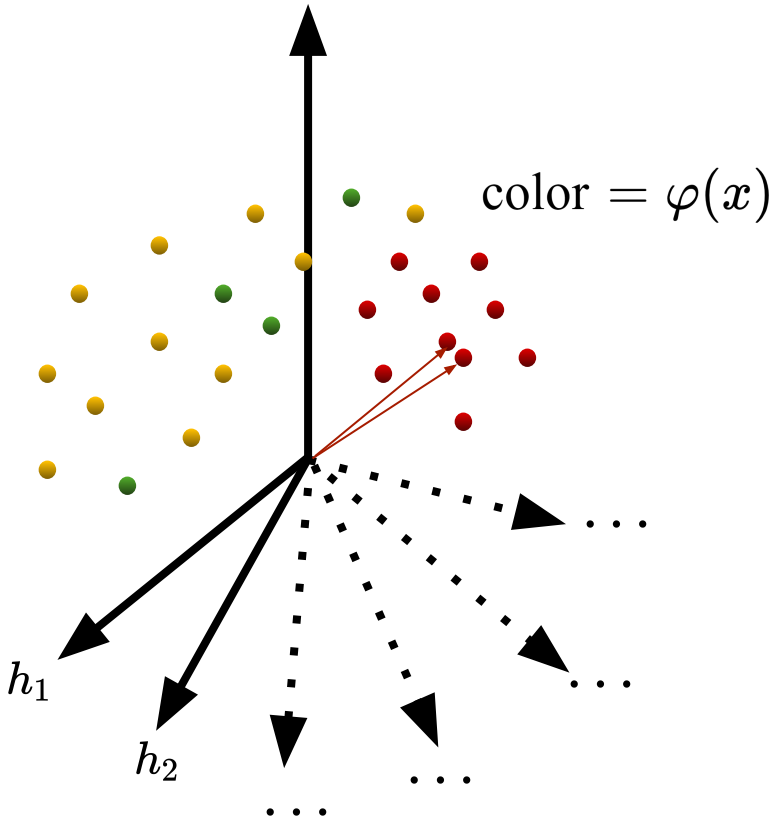


Examples of φ :

- luminance of an image
- sentiment of a sentence

NO with general labelling

The class label is only an example

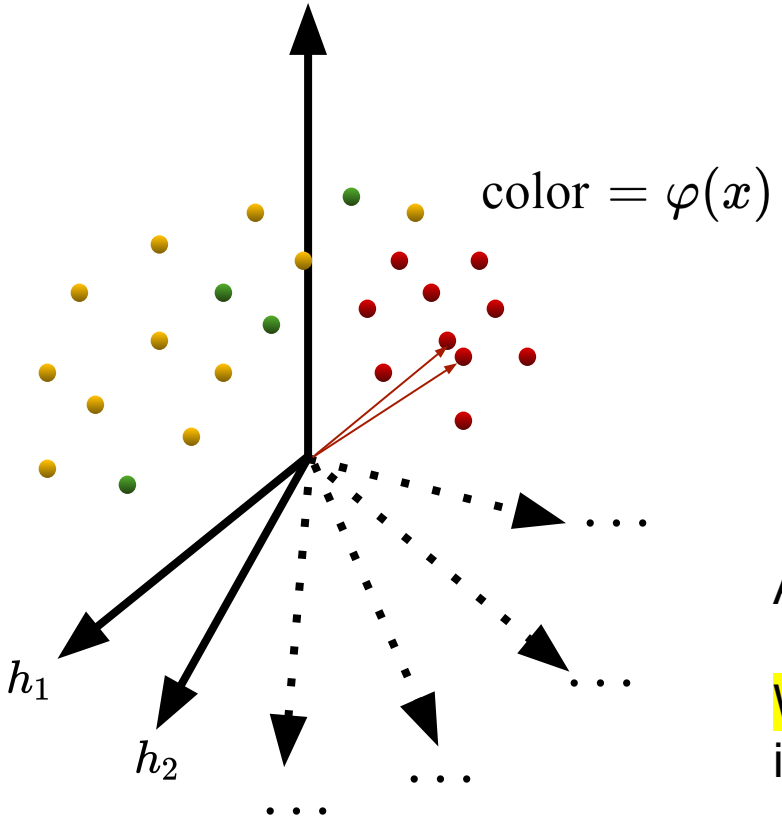


Examples of φ :

- luminance of an image
- sentiment of a sentence
- “family” of a protein, fold
- …

NO with general labelling

The class label is only an example



Examples of φ :

- luminance of an image
- sentiment of a sentence
- “family” of a protein, fold
- …

Address the question:

Where (and how much) the feature φ is encoded in the hidden representations?

NO with general labelling

Neighborood Overlap: possible extensions

(l)	1	2	3	4	5	6	7
(l+1)	2	1	3	4	5	6	7

Neighborood Overlap: possible extensions

(l)	1	2	3	4	5	6	7
(l+1)	2	1	3	4	5	6	7

(l)	1	2	3	4	5	6	7
(l+1)	1	2	3	4	5	7	6

Neighborood Overlap: possible extensions

(l)	1	2	3	4	5	6	7
(l+1)	2	1	3	4	5	6	7

(l)	1	2	3	4	5	6	7
(l+1)	1	2	3	4	5	7	6

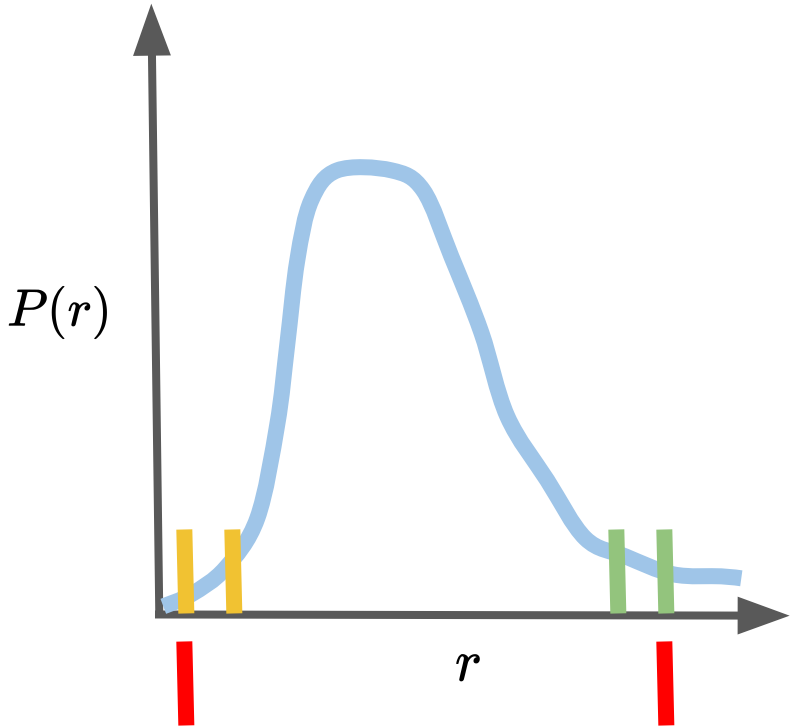
(l)	1	2	3	4	5	6	7
(l+1)	7	2	3	4	5	6	1

Neighborood Overlap: possible extensions

(l)	1	2	3	4	5	6	7
(l+1)	2	1	3	4	5	6	7

(l)	1	2	3	4	5	6	7
(l+1)	1	2	3	4	5	7	6

(l)	1	2	3	4	5	6	7
(l+1)	7	2	3	4	5	6	1



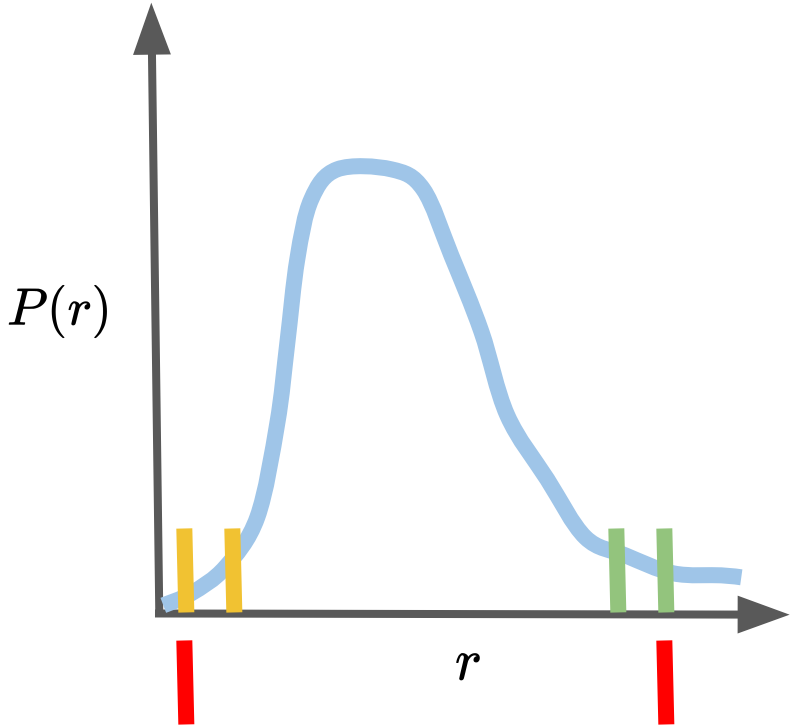
Neighborood Overlap: possible extensions

These three situations are indistinguishable for NO (at the moment)

(l)	1	2	3	4	5	6	7
(l+1)	2	1	3	4	5	6	7

(l)	1	2	3	4	5	6	7
(l+1)	1	2	3	4	5	7	6

(l)	1	2	3	4	5	6	7
(l+1)	7	2	3	4	5	6	1



Next topics: results of

- Intrinsic dimension
- Neighborhood overlap

on representations in self-supervised models (images, proteins, language)

Thanks for the attention!