Statistical Problems of Manifold Learning in Predictive Modeling

Evgeny Burnaev
Associate Professor, Skoltech
Head of ADASE group

Skoltech
Skolkovo Institute of Science and Technology

ADASE group

- 30 researchers
- DL for
  - 3D Computer Vision
  - Predictive Analytics

~ 100 **papers** in major venues, incl. NIPS, ICML, CVPR, etc.

Moscow government prize for Scientific Achievements, 2018

Ilya Segalovich prize for Scientific Achievements, 2020

with the **President of Russian Academy of Science**
State Kremlin palace, 2018

Industrial Expertise: since 2007

AIRBUS, LOUIS VUITTON, AIRBUS HELICOPTERS, IPSEN Innovation for patient care, ALTRAN, SAHARA Force India, PORSCHE, SBERBANK, HUAWEI, Alibaba Group 阿里巴巴集团, BOSCH

and many others...

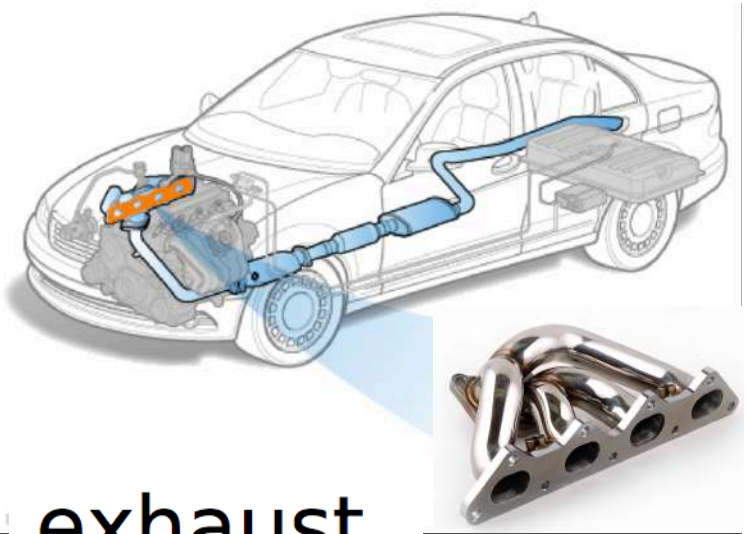
Overview

- Intro
- Dimension Reduction Problem Statements
- PCA, MDS and Sammon Mapping, Autoencoders
- ISOMAP and LLE
- TDA for Time Series Analysis
- References

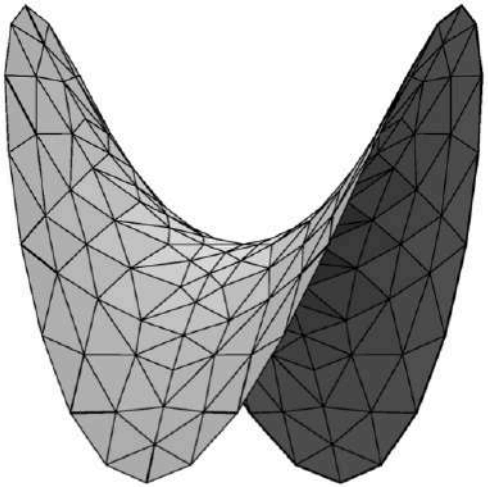
Overview

- **Intro**
- Dimension Reduction Problem Statements
- PCA, MDS and Sammon Mapping, Autoencoders
- ISOMAP and LLE
- TDA for Time Series Analysis
- References

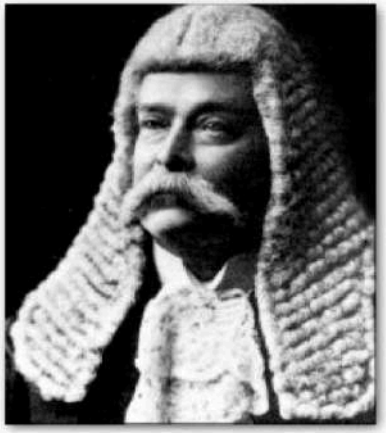
Types of manifold



- exhaust manifold

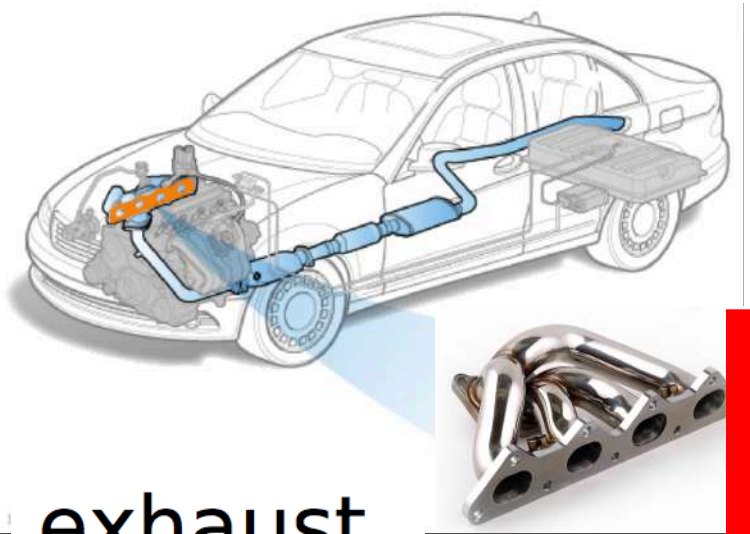


low-D surface embedded in high-D space

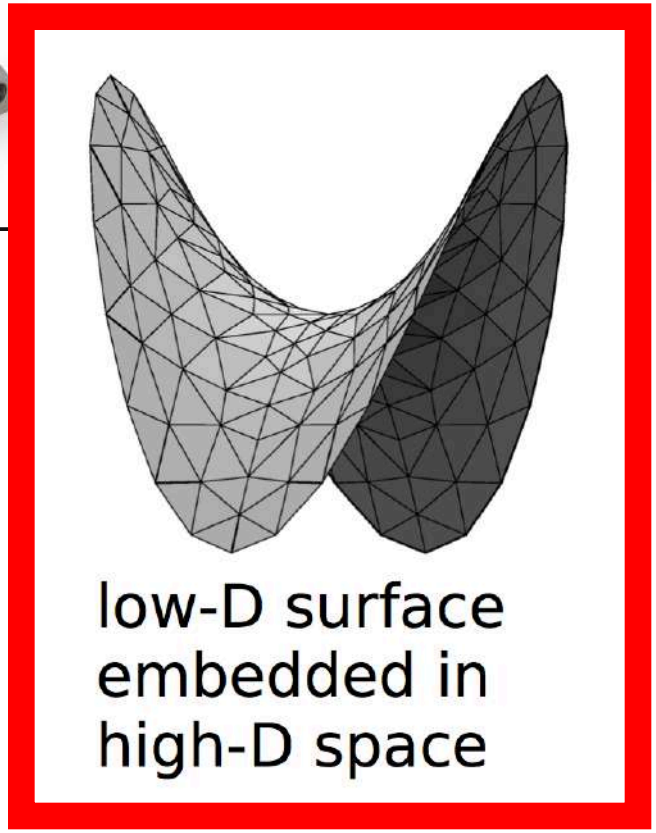


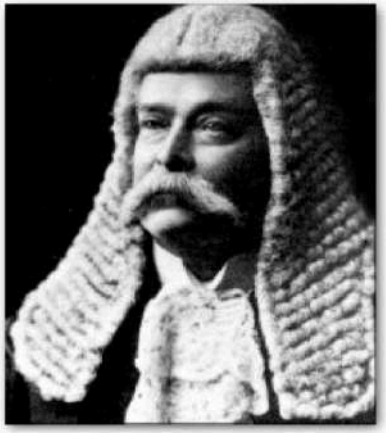
Sir Walter Synnot Manifold 1849-1928

Types of manifold



- exhaust manifold





Sir Walter
Synnot Manifold
1849-1928

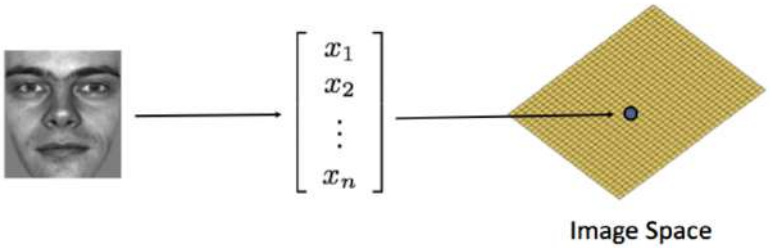
Manifold Learning

Manifold learning – Data Analysis technology based on **geometrical model** about high-dimensional data [1]

A. **The world is multidimensional**
B. **Multidimensional data are difficult to use**
C. **Real-world data have low-dimensional structure**
D. **The world is not flat (nonlinear)**

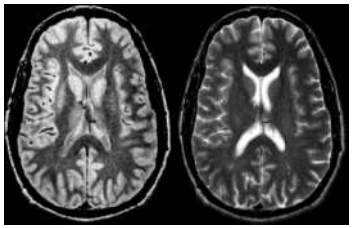
A. The world is multidimensional

Real-world data



1024×1024: $d \approx 10^6$

64×256: d = 16 384



fMRI: $d \approx 1.4 \times 10^6$/sec

B. Multidimensional data is difficult to analyze

1) **Regression:** (Ibragimov, Khasminskii (1979); Stone (1982); etc.)

If $\mathbf{F} = \{\psi : [0,1]^d \to \mathbb{R}^1,\ \psi \text{ is Lipschitz}\}$

then for any estimator $\hat{\psi}$ of any kind from n known measurements $\{(x_i, \psi(x_i))\}$:

$$\sup_{\psi \in \mathbf{F}} \mathbb{E}\left(\psi(x) - \hat{\psi}(x)\right)^2 \geq \text{Const} \times n^{-2/(2+d)}$$

The lower bound is nonasymptotic!

2) MSE in case of KDE ~ $O(n^{-4/(d+4)})$

Multidimensional data are difficult to analyze (cont.)

Data Analysis works for real multidimensional data – when and why?

Which properties of real-world data help? How?

- concentration of measures
- sparse representation
- latent structure models
- low-dimensional structure of data support

Multidimensional data are difficult to analyze (cont.)

Data Analysis works for real multidimensional data – when and why?

Which properties of real-world data help? How?

- concentration of measures
- sparse representation
- **latent structure models**
- **low-dimensional structure of data support**

C. Low-dimensional structure of real-world data

- Data from '**natural**' sources occupy usually a **small part** $\mathbf{X}$ in the 'observation space' $\mathbb{R}^d$
- $\mathbf{X}$ has **small 'intrinsic dimension'** $s < d$
- Data can be described by a small number s of parameters (features)

$d \approx 10^6$

$s = 84$

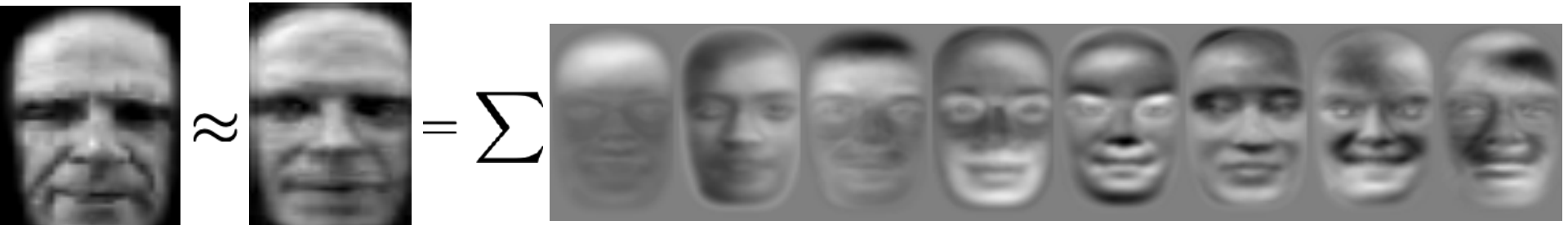
$s = 40$

Low-dimensional structure of real-world data

- **to find a low-dimensional structure of Data space**
  - ✓ to estimate an Intrinsic dimension s of $\mathbf{X} \subset \mathbb{R}^d$
  - ✓ to construct a s-dimensional features $z = h(x)$ describing $x \in \mathbf{X}$
- **to use extracted low-dimensional structure** to solve **specific Data analysis tasks**

Principal Component Analysis

Face-vector $x \in \mathbb{R}^{2061}$



$\approx \quad = \sum$

$(z_1 \times \mathbf{e}_1)+(z_2 \times \mathbf{e}_2)+(z_3 \times \mathbf{e}_3)+(z_4 \times \mathbf{e}_4)+(z_5 \times \mathbf{e}_5)+(z_6 \times \mathbf{e}_6)+(z_7 \times \mathbf{e}_7)+(z_8 \times \mathbf{e}_8)$

x_{PCA}= projection on L_{PCA} defined by features $z = (z_1, z_2, \ldots, z_s)$

$$x(face) \leftrightarrow z = (z_1, z_2, \ldots, z_s) \in \mathbf{R}^s$$

$$L_{PCA,s}(faces) = \{mean\ face + \sum_{i=1}^{s} EigenFace_i \times z_i\}$$

Principal Component Analysis (cont.)

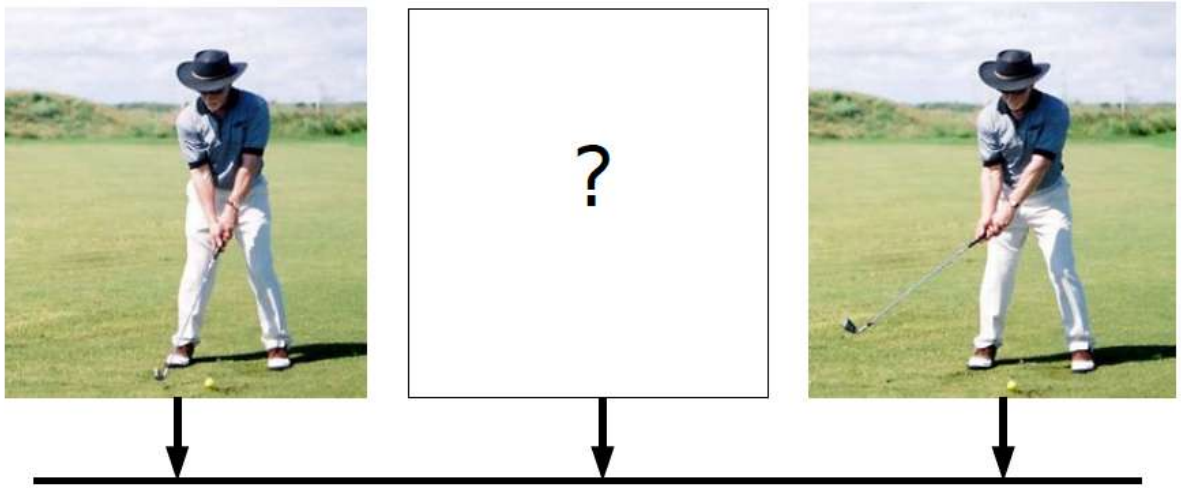
Original face described by 10^6-dimensional vector

Left to right: the same face described by ***s*** reduced features

s = 84 ***s*** = 40 ***s*** = 20 ***s*** = 3 ***s*** = 2 ***s*** = 1

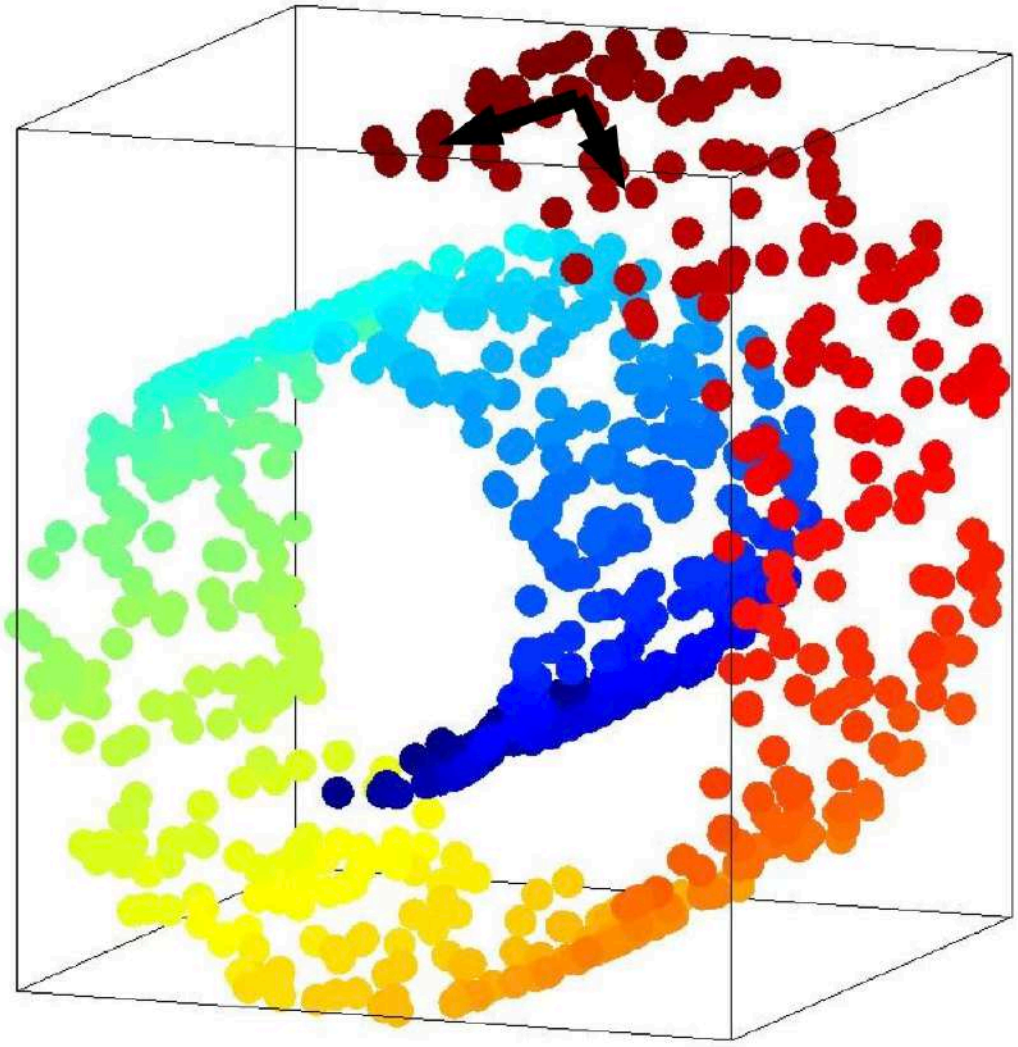
D. The world is not flat

Video stream

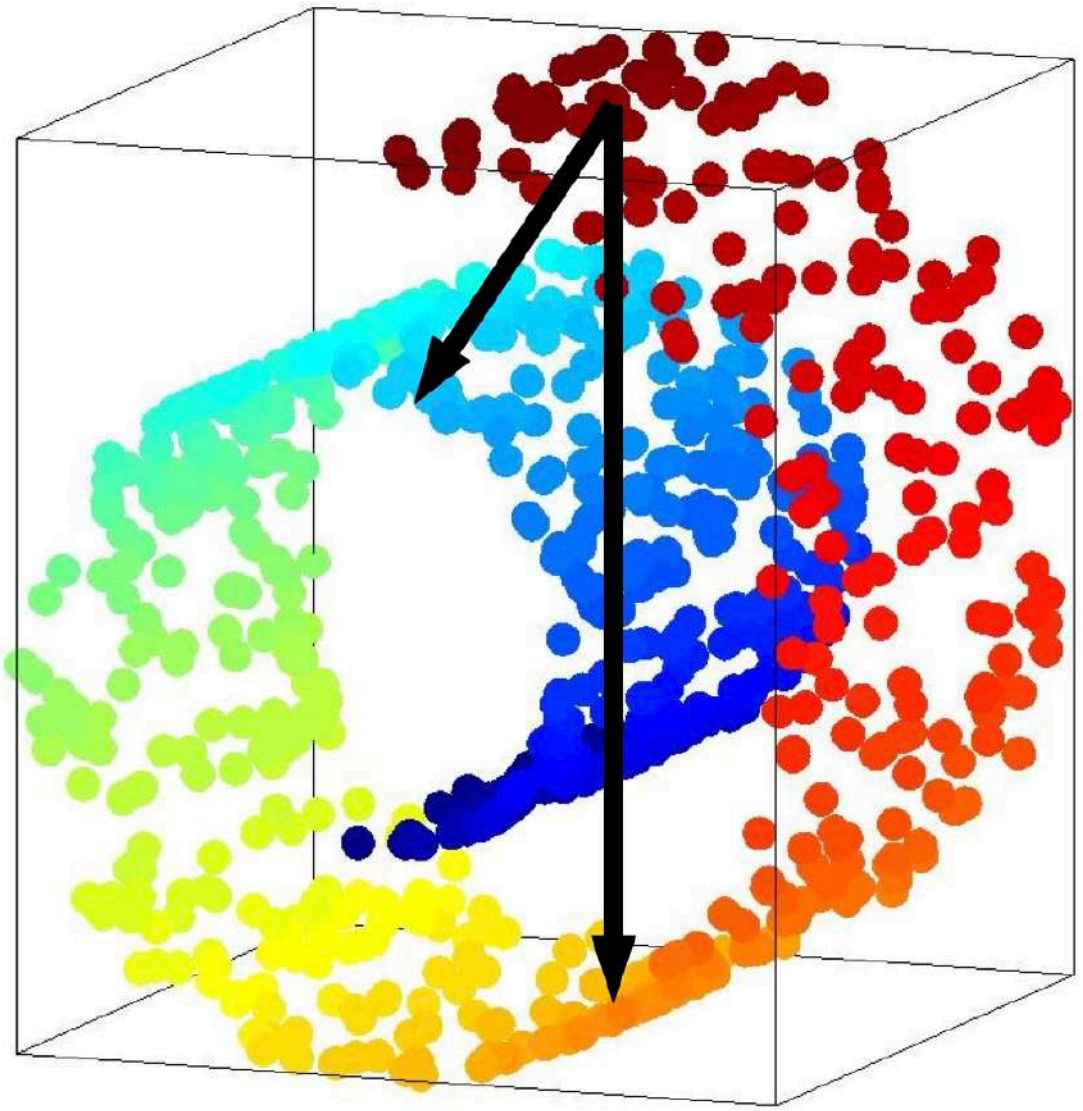


Frame rate conversion based on inter-frame interpolation

What is “Reasonable distance metrics”?

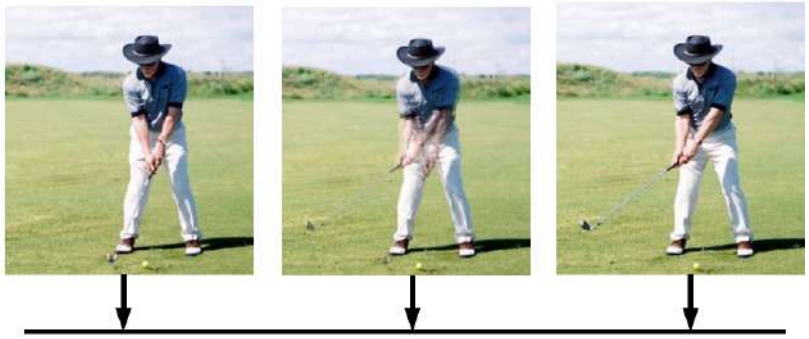


What is “Reasonable distance metrics”?



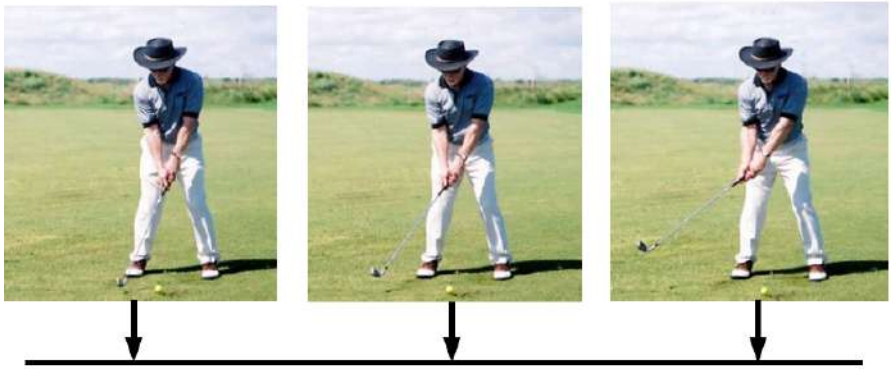
The world is not flat (cont)

‘Linear’ inter-frame interpolation

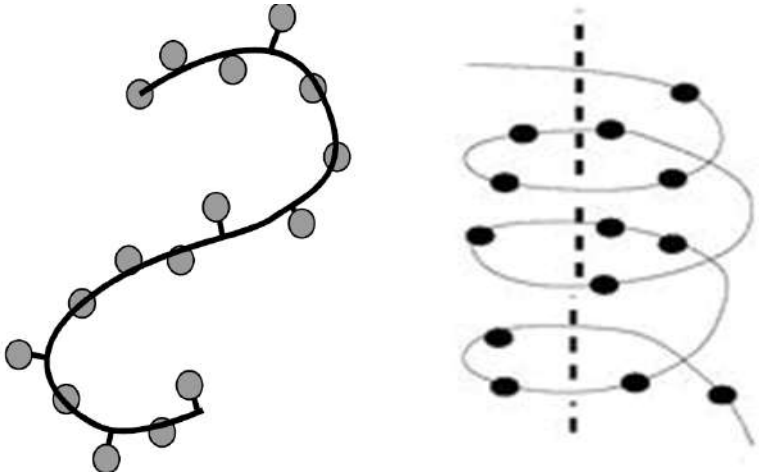


The world is not flat (cont.)

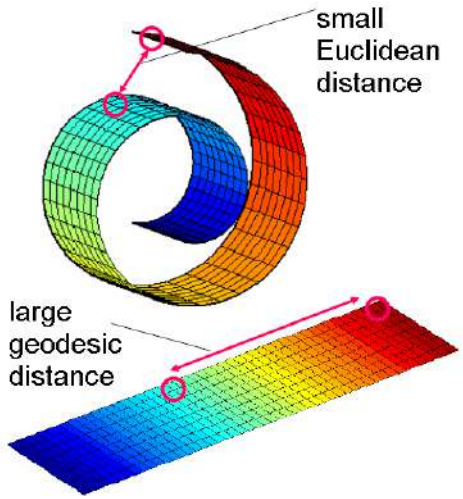
‘Nonlinear’ inter-frame interpolation



Data analysis for nonlinear data



- **Linear methods like PCA do not work**



- **Math: Multivariate Statistical analysis consider mainly ‘linear methods’**

Manifold learning

- **Electricity price curve modeling** and forecasting by **manifold learning** (2008)
- On the **manifold structure** of the space of **brain images** (2009)
- **Manifold Modeling** for **Brain Population Analysis** (2010)
- Simultaneous Learning of **Nonlinear Manifold** and **Dynamical Models** for **High-dimensional Time Series** (2007)
- A kernel entropy **manifold learning** approach for **financial data analysis** (2014)
- **Manifold learning** algorithms for localization in **wireless sensor networks** (2004)
- **Manifold Learning** for **Object Tracking with Multiple Motion Dynamics** (2010),
- **Manifold Learning** and Representations for **Image Analysis and Visualization** (2006)
- **Learning an Image Manifold** for **Retrieval** (2004)
- **Manifold Learning** for **medical image registration segmentation and classification** (2012),
- The **Manifold Tangent Classifier** (2011),
- **Inferring 3D body pose** from silhouettes using activity **manifold learning** (2004)
- Nonlinear **manifold learning** for **visual speech recognition** (2005)
- **Separating style and content** on a **nonlinear manifold** (2004)
- **Learning Nonlinear Manifolds** from **Time Series** (2006)
- A New **Manifold Learning** Technique for **Face Recognition** (2012)
- Regional **Manifold Learning** for **Deformable Registration of Brain MR Images** (2012)
- Regional **Manifold Learning** for **Disease Classification** (2014)
- Overview of **Manifold Learning** and Its **Application in Medical Data set** (2014)
- **Sparse pose manifolds** (2014)

Manifold model: mobile robot navigation [2]

64×256 pixels: d = 16384

Robot localization θ = **(2D Coordinates, Orientation)** $\in \mathbb{R}^3$

$x = \varphi(\theta) \in \mathbb{R}^d$ - captured image at Robot localization θ

Appearance space $\mathbf{M} = \{x = \varphi(\theta),\, \theta \in \Theta\}$
consisting of images which may be captured

under **all possible localizations** $\theta \in \Theta \subset \mathbb{R}^3$ is 3D-surface (**Appearance manifold**) in $\mathbb{R}^d$

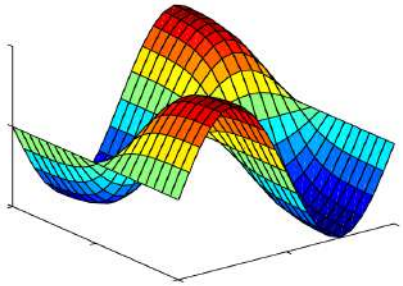
Manifold covered by single chart (surface in $\mathbb{R}^d$)

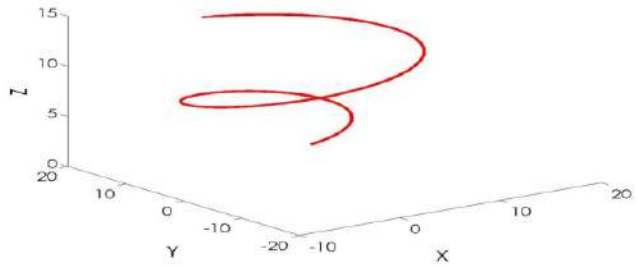
$$\mathbf{M} = \{x = g(z) \in \mathbb{R}^d : \; z \in \mathbf{Z} \subset \mathbb{R}^s\}$$

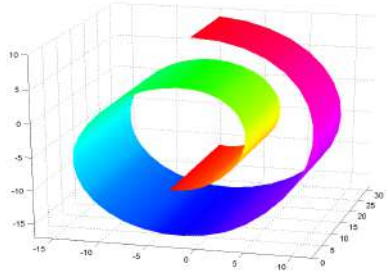
well-behaved **unknown** s -dimensional surface - **Data manifold**

covered by **single chart** g defined on **Coordinate space** $\mathbf{Z} \subset \mathbb{R}^s$

and embedded in ambient d-dimensional space, $s < d$

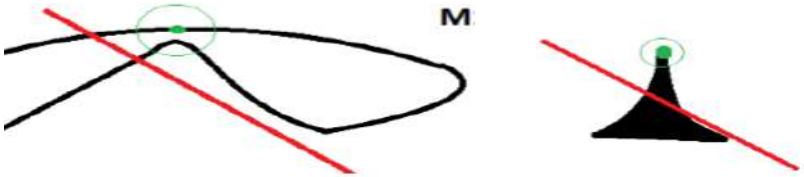






$h = g^{-1} : \mathbf{M} \to \mathbf{B}$ - inverse mapping – a parameterization

$z = h(x)$ on the Data manifold



Statistical analysis of manifold valued data

Let μ be some **unknown** probability measure on **unknown** s-dimensional manifold $\mathbf{M} = \mathrm{supp}(\mu)$ with **unknown** value of s

Based on given sample of independent observations

$$\mathcal{D}_n = \{x_1, x_2, \ldots, x_n\} \subset \mathbf{M}$$

solve various statistical problems such as:

- to estimate intrinsic dimension s
- to estimate low-dimensional parameterization h on the manifold $\mathbf{M}$
- to estimate the manifold $\mathbf{M}$
- to estimate tangent space $L(x)$ to the manifold $\mathbf{M}$ at point x
- to estimate a density $f(x) = \frac{d\mu}{dm}$, etc.

Manifold Learning in Data analysis

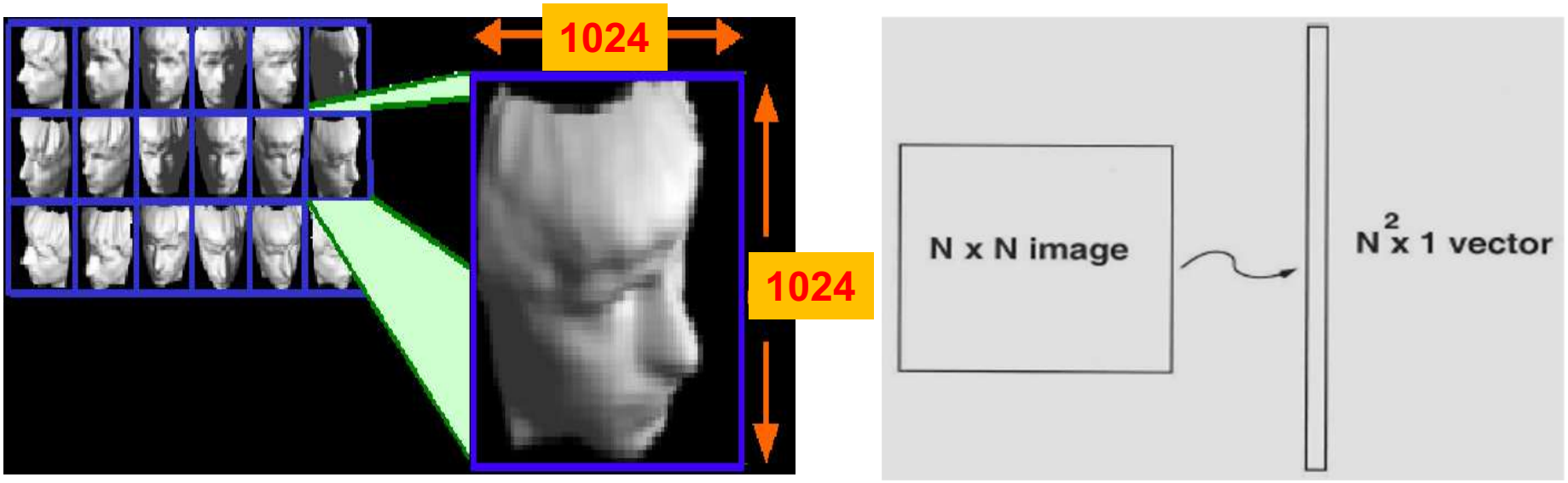
Data Analysis under Manifold Assumption

- Parameterization of the Data manifold
- Estimation of the Data manifold
- Denoising/Super-resolution/Impainting
- Classification
- Regression
- Anomaly Detection
- Other applications

Dimensionality Reduction Problem

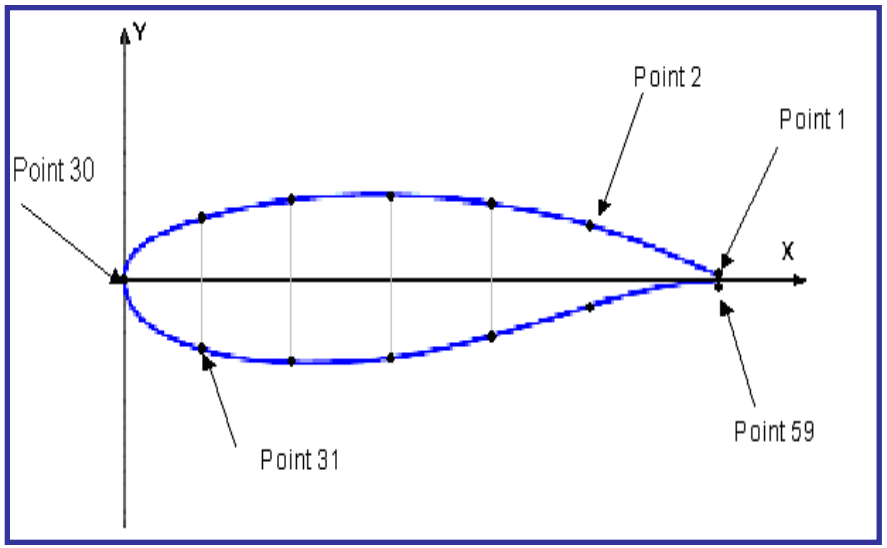
- object **O** is described by a **p**-dimensional vector **X(O)** $\in$ R^p. Components of **X(O)** are features of **O**

Example 1 (human face): grey-scale pixel-wise representation



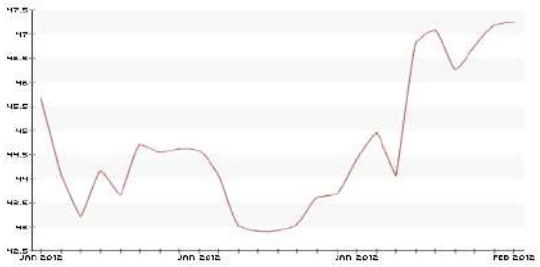
Face is represented by 1024×1024 pixels – dimension p = 10^{20} ~ 1 000 000

Example 2 (wing airfoil description).

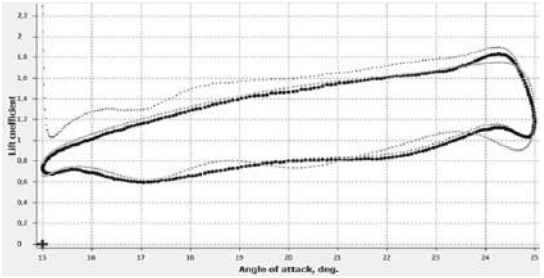


Airfoil **O** $\rightarrow$ **X(O)** = $(x_1, x_2, \dots, x_p)^T$ - ordinates of upper and lower contours, **p ~ 50 ÷ 200**

Example 3 (plots of dependences, multidimensional time-series)



Electricity price curve



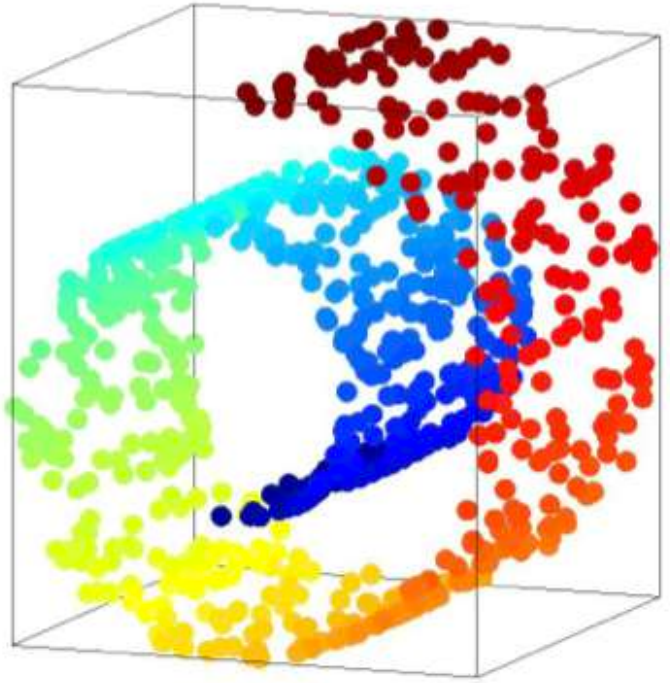
Lift force vs. AoA curve

Speech recognition

Curve f(x) is described by the vector $f = (f_1, f_2, \dots, f_p)^T \in R^p$, $f_j = f(t_j)$; time-series segment $(x_1, x_2, \dots, x_p)$ is considered as a **p-**dimensional vector

Example 4 (MNIST)

Example 5 Famous Toy Problem
"Swiss Roll"



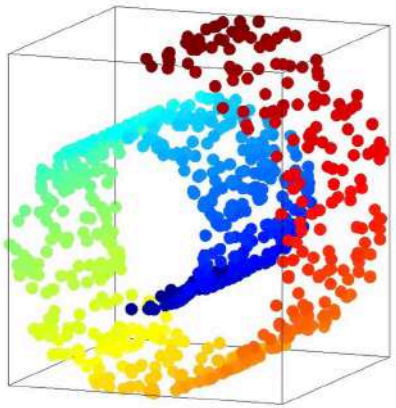
- High dimensionality **p** of **X(O)** is critical for efficient learning
- We can visualize only in 2D/3D

- Dimension Reduction = construct reduced dimension representation **$y(O) \in R^q$**, **$q << p$**, of **O** without "significant loss of information"

- DR for
  - ✓ Visualization
  - ✓ Data compression
  - ✓ "curse of dimensionality"
  - ✓ De-noising
  - ✓ Reasonable distance metrics

$X \rightarrow X'$ S.T.
$\dim(X') << \dim(X)$

uncovers the intrinsic dimensionality (invertible)



Overview

- Intro
- **Dimension Reduction Problem Statements**
- PCA, MDS and Sammon Mapping, Autoencoders
- ISOMAP and LLE
- TDA for Time Series Analysis
- References

Dimension Reduction as an Embedding Problem

Embedding Problem. Using a sample

$$\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset R^p$$

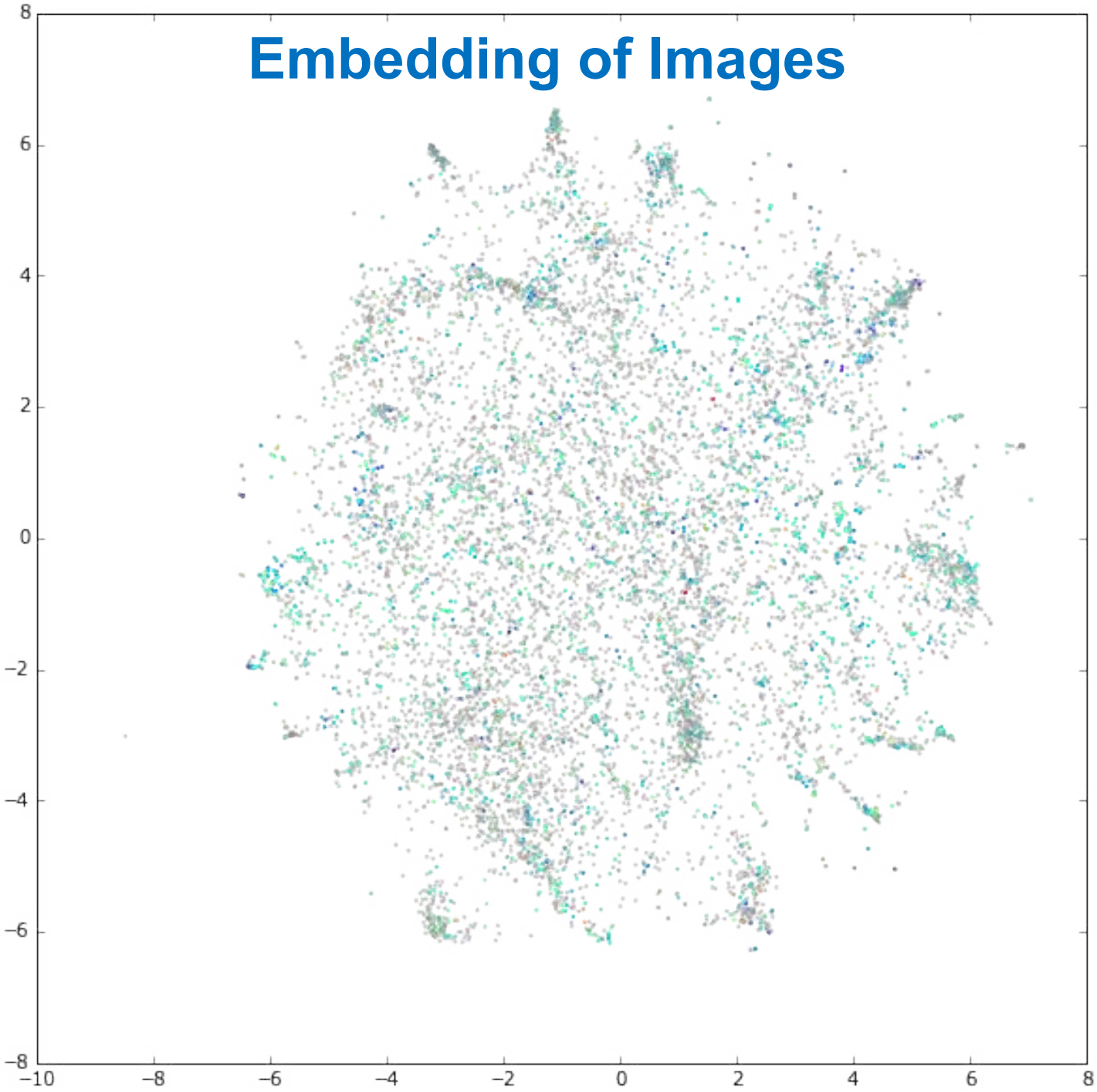
construct an embedding

$$h: \mathbf{X}_n \to \mathbf{Y}_n = h(\mathbf{X}_n) = \{y_1, y_2, \ldots, y_n\} \subset R^q$$

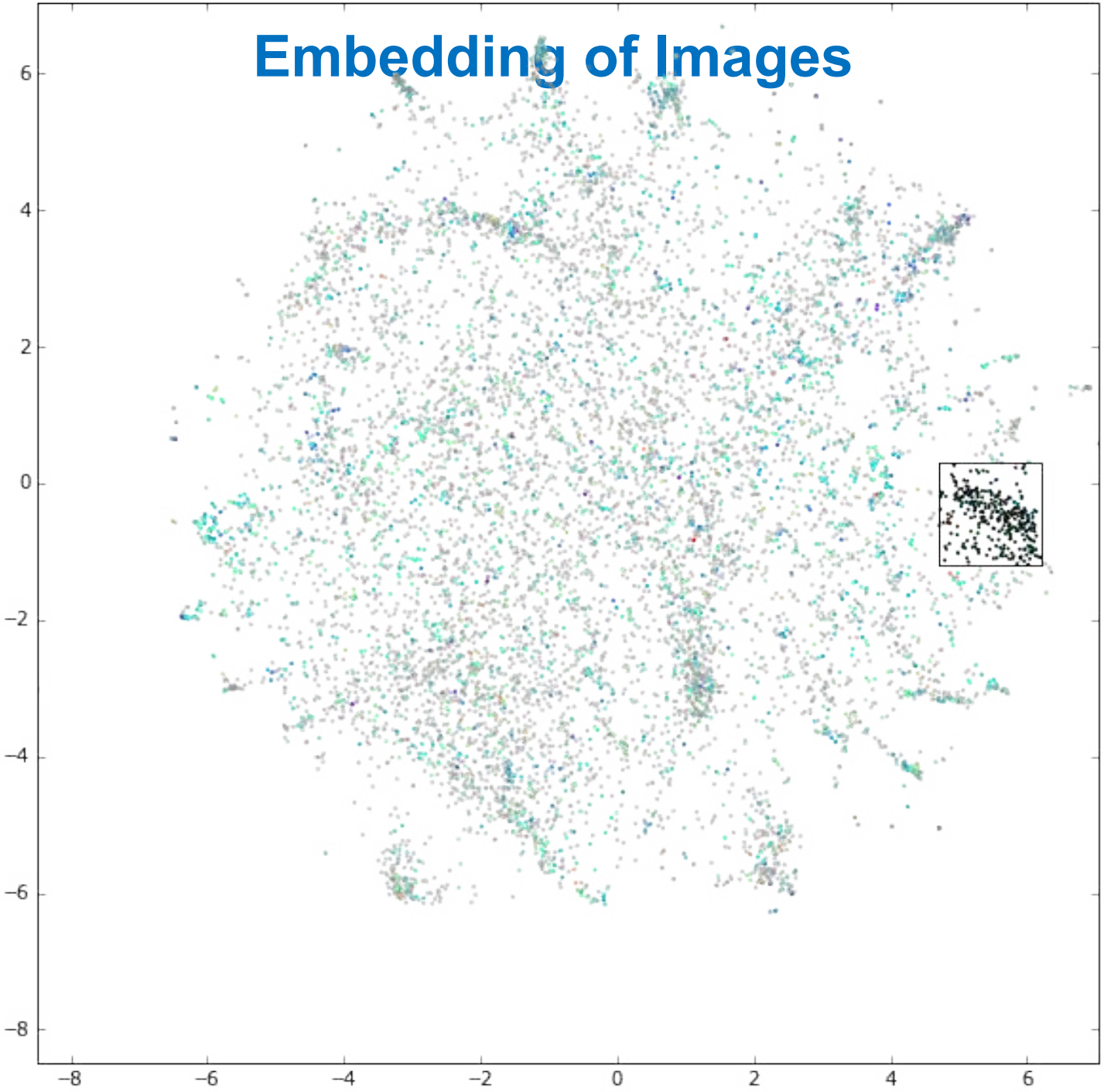
of sample points $\mathbf{X}_n$ в R^q, **q < p**, such that the set $\mathbf{Y}_n$ **«consistently represents»** the set $\mathbf{X}_n$

(e.g., the set $\mathbf{Y}_n$ should preserve some geometric structure of $\mathbf{X}_n$, etc.)

Embedding of Images



Embedding of Images






10 ДНЕЙ
ДО ЛЕТА













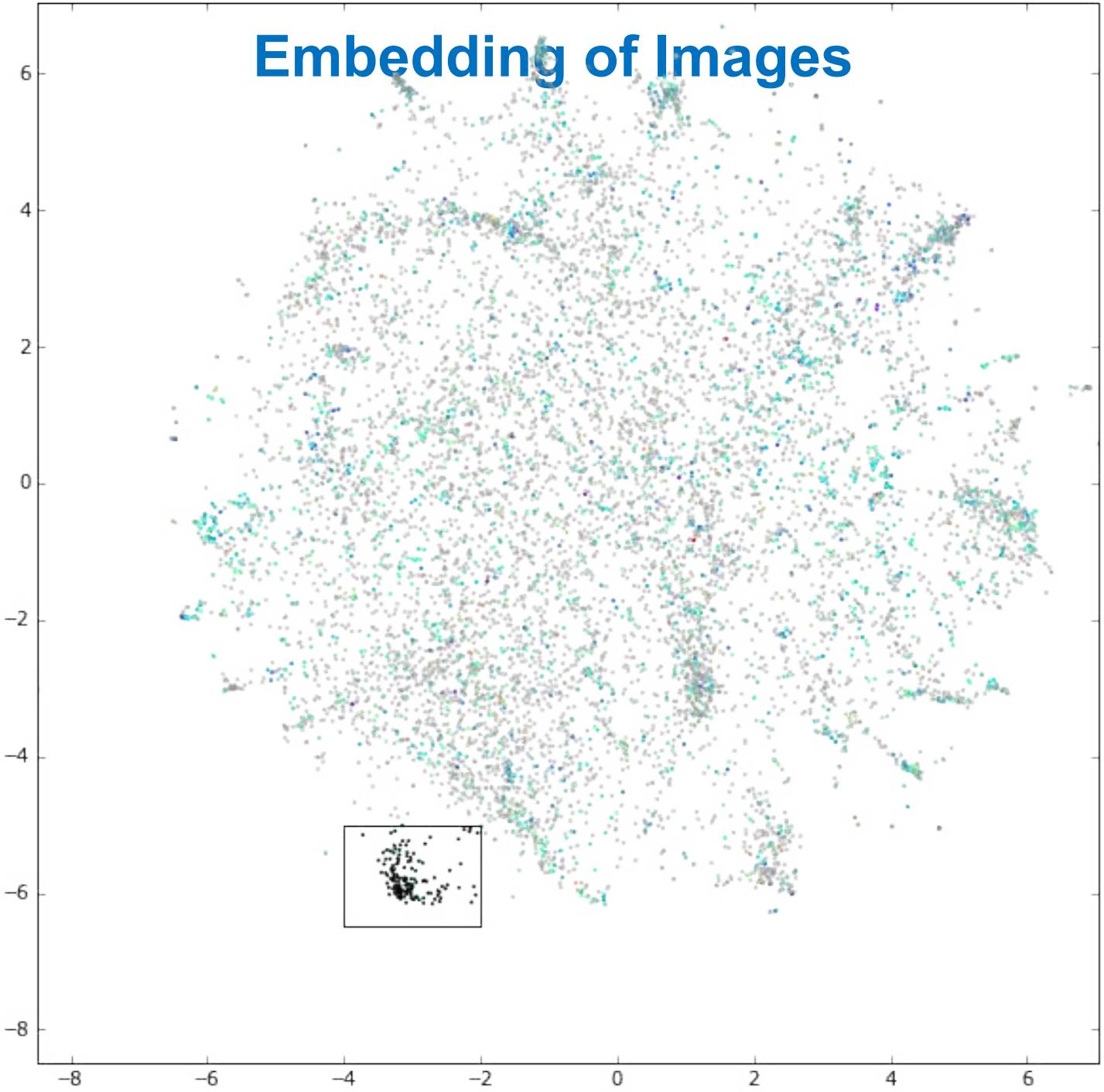


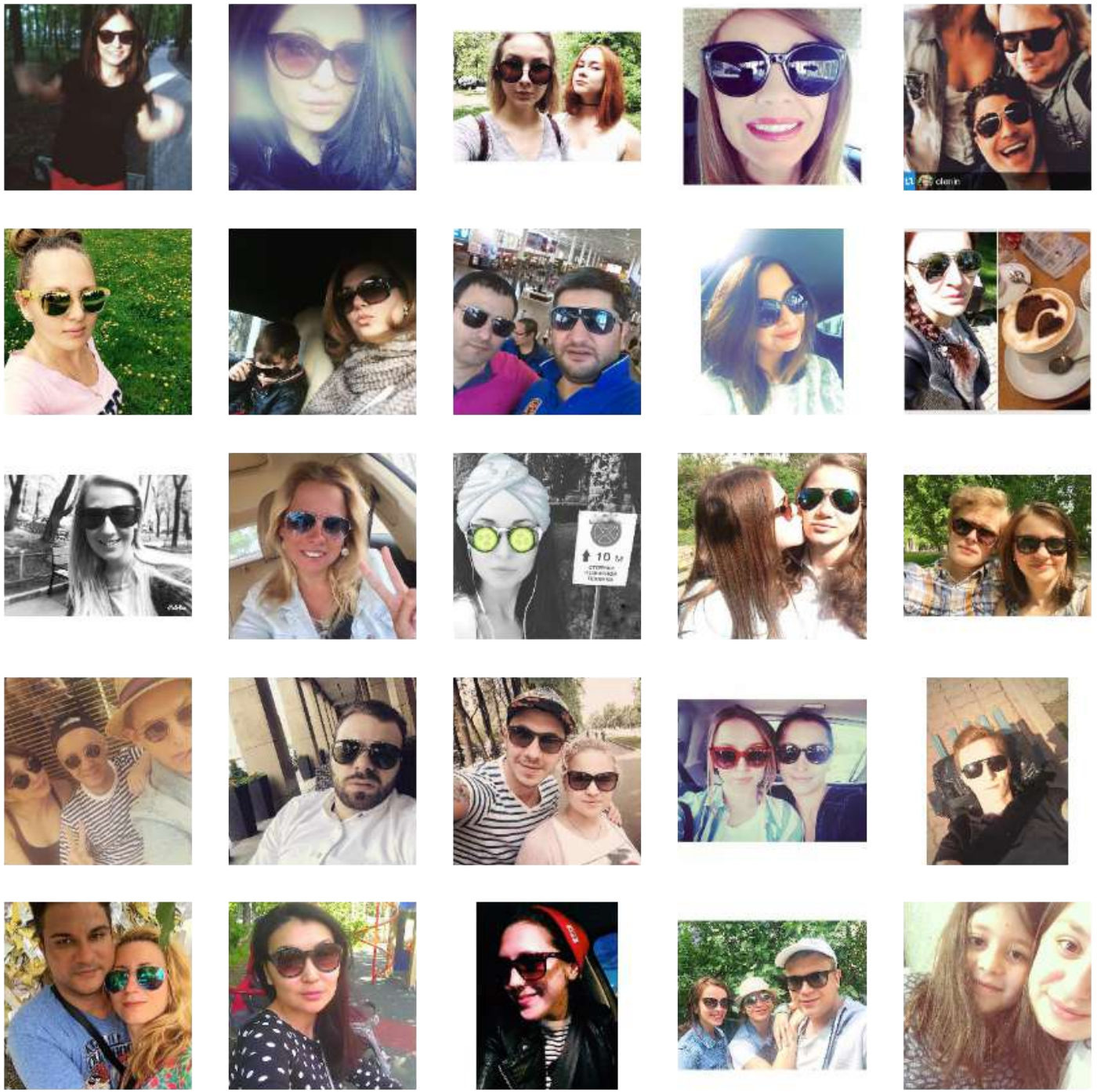




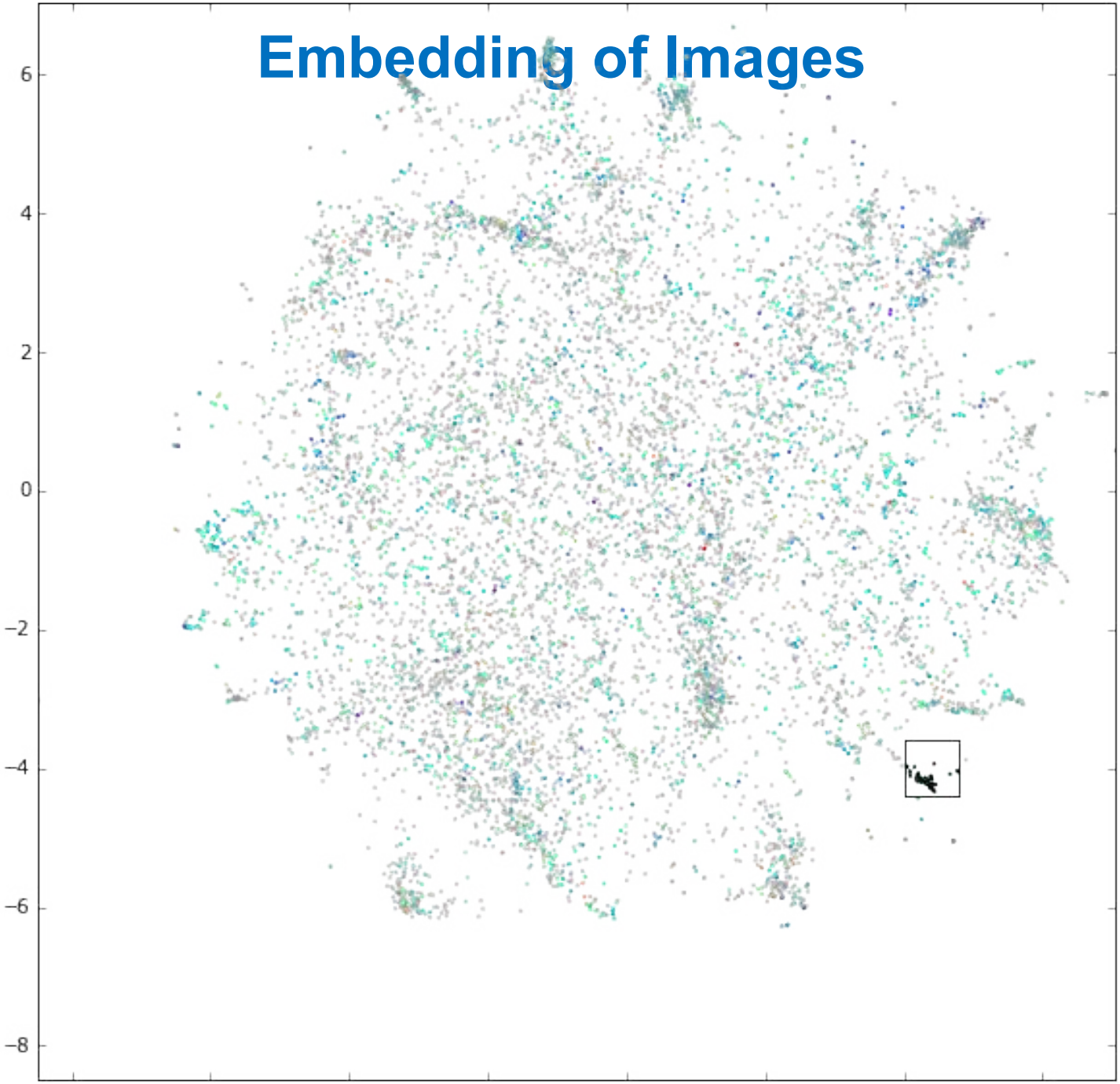


Embedding of Images





Embedding of Images




8 963 653 8888
8 963 653 8888
Беза-гель!!! С 01.06
по 07.06
Освободилось 1
место на курс!!!

Typical scheme:

• construct some cost function $L(y_1, y_2, \ldots, y_n)$. E.g. for Multi Dimensional Scaling, MDS

$$L(y_1, y_2, \ldots, y_n) = \sum_{i,j} \left(\rho(X_i, X_j) - \| y_i - y_j \| \right)^2$$

• optimize cost function

• visually analyze results (Swiss Roll, Spiral, …)

Popular methods to solve embedding problem:

- Pursuit Projection; Principal Component Analysis
- Locally Linear Embedding, LLE; Conformal Eigenmaps
- Laplacian Eigenmaps LE
- Hessian Eigenmaps, HE
- ISOmetric MAPing, ISOMAP; Landmark ISOMAP
- Kernel PCA, KPCA
- Local Tangent Space Alignment, LTSA

Extended Embedding Problem

Using a sample

$$\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset R^p$$

construct an embedding transformation

$$h: \mathbf{X} \subset R^p \rightarrow \mathbf{Y} = h(\mathbf{X}) \subset R^q,$$

both for sample points $\mathbf{X}_n$, and for new (out-of-sample) points

$$X_{new} \in \mathbf{X} / \mathbf{X}_n$$

without solving embedding problem for $\mathbf{X}_n \cup \{X\}$ anew

Example – face recognition

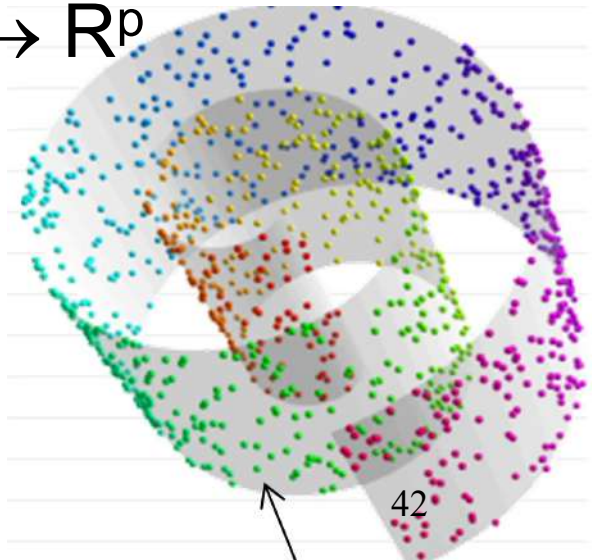
For extended embedding problem we need a Data Model **X:** description of **X**, and of a mechanism, which generates points $\mathbf{X}_n$ and new points X_{new} from **X**

Usually we assume that real data is well approximated by some manifold, i.e.

Manifold Data Model:

$$\mathbf{X} = \{X = f(b) \in R^p : b \in \mathbf{B} \subset R^q\} \subset R^p$$

- open set **B** (inner coordinate space)
- smooth bijective transformation $f: \mathbf{B} \to R^p$



Manifold Learning Problem!!!

Full Dimension Reduction problem

Using a sample

$$\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset R^p$$

construct embedding transformation:

$$h: \mathbf{X} \subset R^p \rightarrow \mathbf{Y} = h(\mathbf{X}) \subset R^q$$

and **Reconstruction transformation**

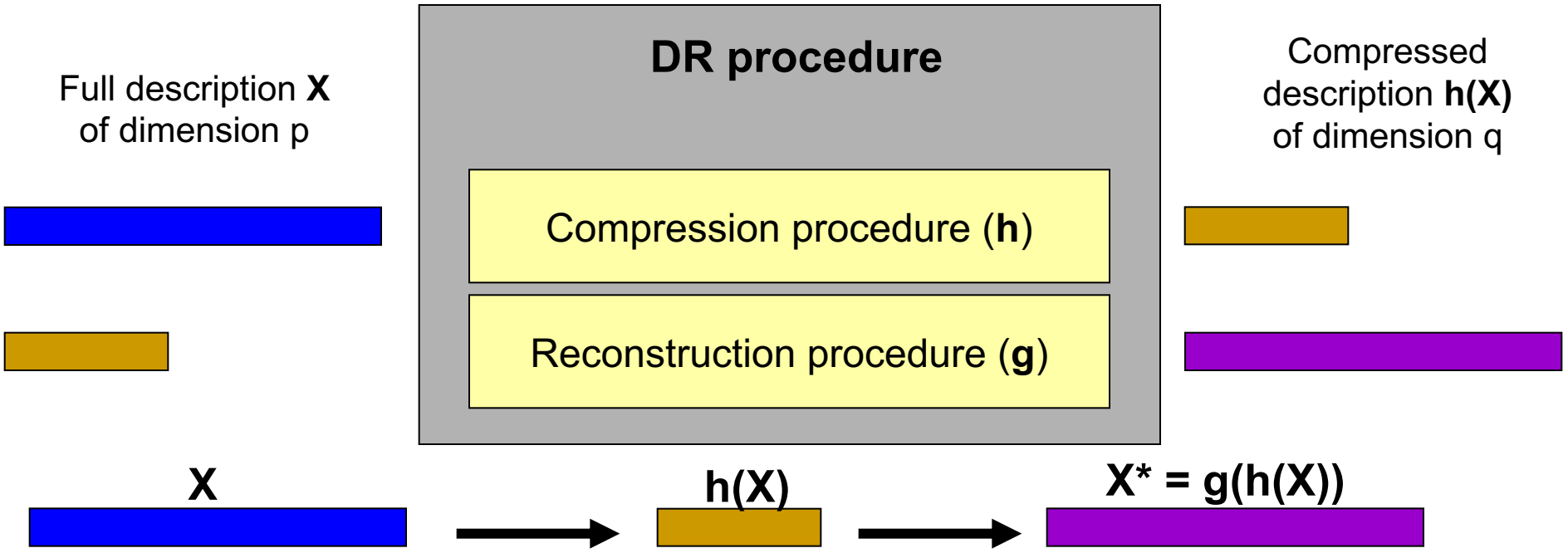
$$g: \mathbf{Y} \subset R^q \rightarrow \mathbf{X} \subset R^p$$

such that

$$g(h(X)) \approx X \quad \text{for } \mathbf{all}\ X \in \mathbf{X}$$

Example – dimension reduction of airfoils

Dimension Reduction



Requirements:

minimal dimension **q = Dim h(X)** of compressed description **h(X)**, providing required proximity between **X* = g(h(X))**) and **X**

or:

maximal proximity between **X** and **X* = g(h(X))** for a fixed dimension **q = Dim h(X)** of compressed description

Overview

- Intro
- Dimension Reduction Problem Statements
- **PCA, MDS and Sammon Mapping, Autoencoders**
- ISOMAP and LLE
- TDA for Time Series Analysis
- References

Principal Component Analysis, PCA
(K. Pearson, 1899)

Problem: find in R^p an affine subspace

$$L(q) = \left\{ x \in R^p : x = x_0 + \sum_{j=1}^{q} y_j \times e_j,\ y_1, y_2, \ \ldots, y_q \in R^1 \right\}$$

of dimension q < p, **which best approximates the set of points**

$$\mathbf{X}_n = \{X_i,\ i = 1, 2, \ldots, n\} \subset R^p.$$

in PCA: "the best" = minimize w.r.t. $\mathbf{x_0}$, $\mathbf{\{e_1, e_2, \ldots, e_q\} \subset R^p}$

$$\frac{1}{n} \sum_{j=1}^{n} \left\| X_j - P_{L(q)} X_j \right\|^2$$

$$\Pr_{L(q)}(X) = x_0 + \sum_{j=1}^{q} y_j(X) \times e_j, \qquad y_j(X) = (X - x_0, e_j)$$

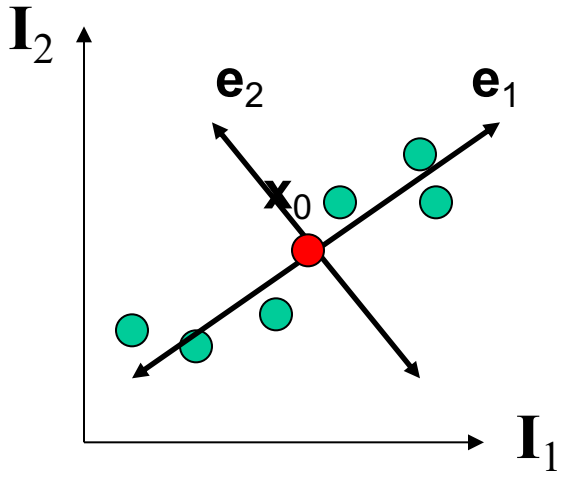
x_{mean} – empirical mean of $\{X_i, i = 1, 2, \dots, n\}$,

$\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p\}$ – eigenvalues of an empirical covariance (p×p)-matrix

$$\Sigma = \frac{1}{n}\sum_{j=1}^{n}(X_j - x_{mean}) \times (X_j - x_{mean})^T$$

providing orthonormal basis in R^p

p = 2
q = 1



$L(1) = \{X \in R^2: X = x_0 + y \times \mathbf{e}_1, y \in R^1\}$

Solution: $x_0 = x_{mean}$, $L(q) = x_0 \oplus \text{Span}(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_q)$,

where orthonormal eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_q\}$ of Σ

correspond to q highest eigenvalues of this matrix

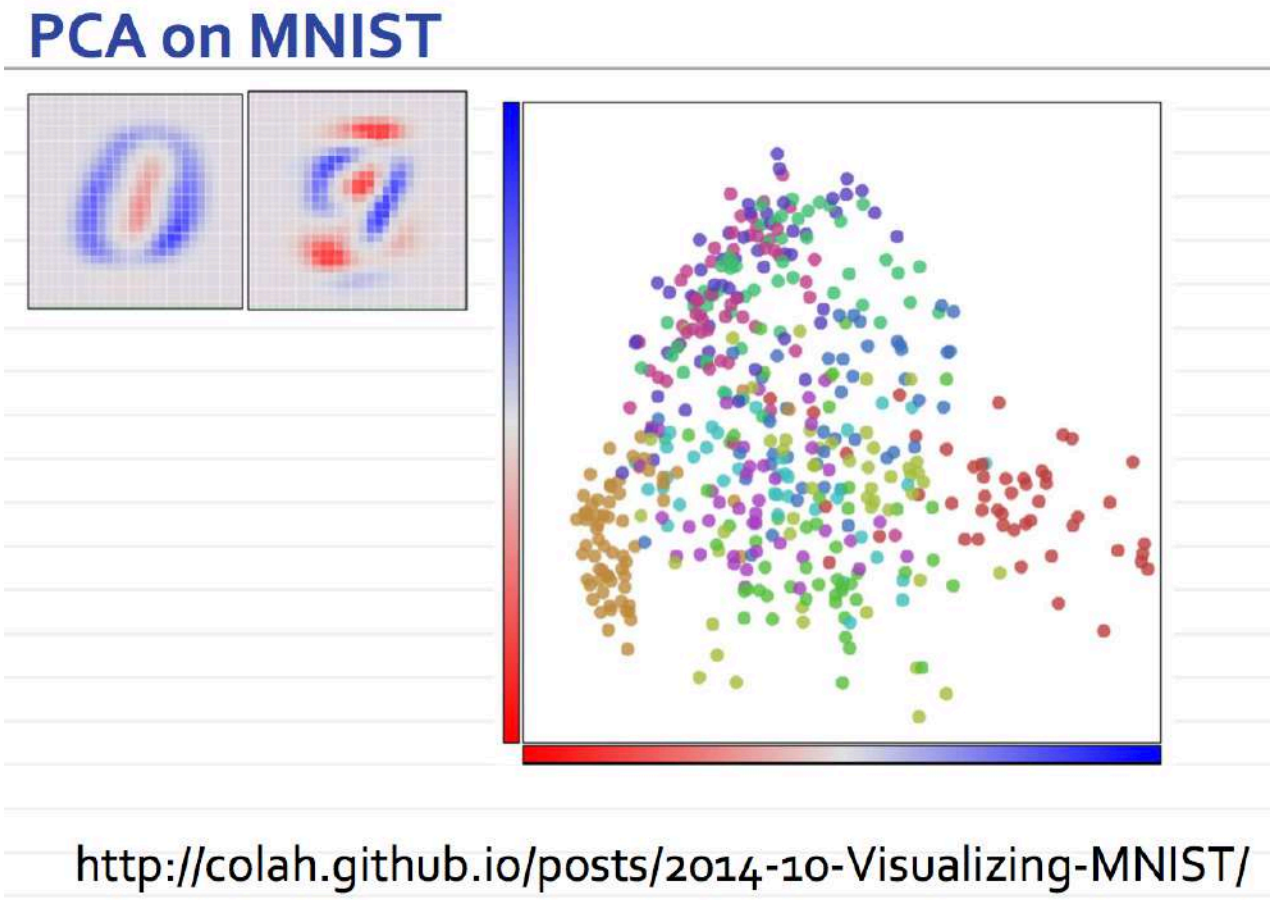
PCA solves E-Problem, EE-Problem and FULL DR problem:

embedding procedure:

$$X \in R^p \rightarrow h(X) = (y_1(X), y_2(X), \ldots, y_q(X))^T \in R^q$$

reconstruction procedure:

$$\mathbf{y} = (y_1, y_2, \ldots, y_q)^T \in R^q \rightarrow g(y) = x_0 + y_1 \times \mathbf{e}_1 + y_2 \times \mathbf{e}_2 + \ldots + y_q \times \mathbf{e}_q$$



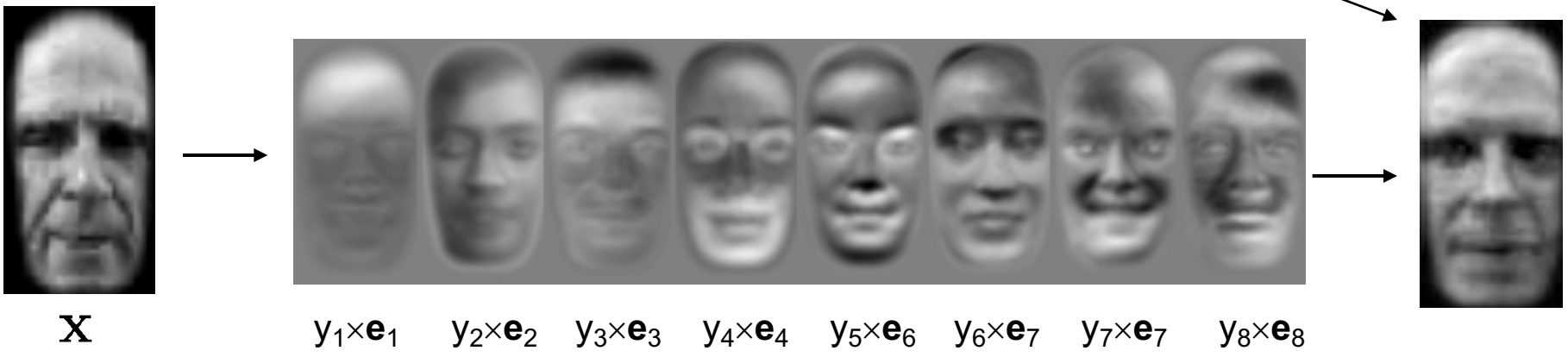
PCA for Face Recognition

- Eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p \in R^p$ in a space of faces: $p \sim 10^6$

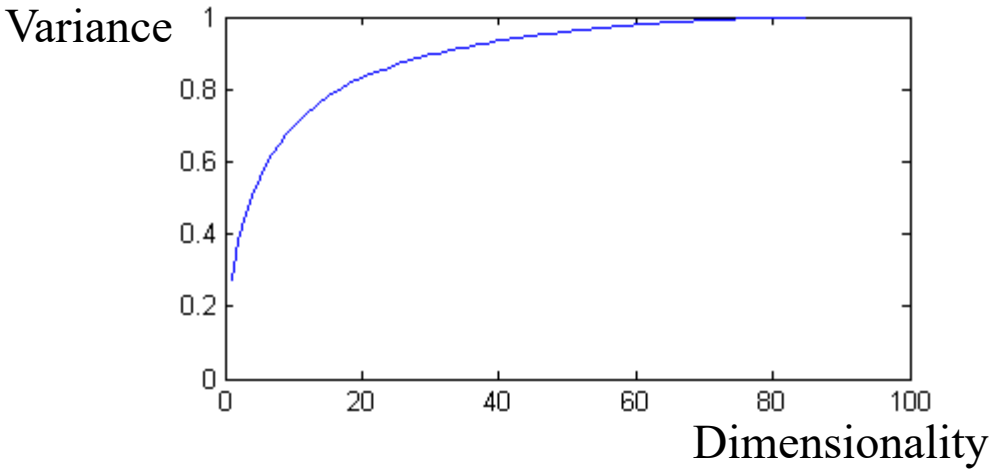
$$X \rightarrow (\underbrace{(X - x_0, \mathbf{e}_1)}_{y_1}, \underbrace{(X - x_0, \mathbf{e}_2)}_{y_2}, \ldots, \underbrace{(X - x_0, \mathbf{e}_p)}_{y_p})^T \in R^p$$

- We select first q, $q \sim 10^2$ eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_q \in R^q$

$$X^* \approx x_0 + y_1 \times \mathbf{e}_1 + y_2 \times \mathbf{e}_2 + \ldots + y_q \times \mathbf{e}_q$$



Left to right: original, reconstruction from 84, 40, 20, 3, 2, and 1 dimensions.



Pursuit Projection, PP: Similar to PCA, constructs **«the best» affine hyperplane** of smaller dimension

Multi Dimensional Scaling, MDS

Minimize quadratic form $\sum_{i,j}\left(\rho(O_i, O_j) - \|y_i - y_j\|\right)^2$

w.r.t. $y_1, y_2, \ldots, y_n \in R^q$, ρ is a some metrics in the object feature space

$y_1, y_2, \ldots, y_n \in R^q$ are defined **except for shift and rotation,** thus we use normalization, e.g.

$$\mathbf{Y}^T \times \mathbf{Y} = \mathbf{I_q} \quad \text{and} \quad \mathbf{Y}^T \times \mathbf{1} = \mathbf{0}$$

where $\mathbf{Y}^T = (y_1 : y_2 : \ldots : y_n)$ – (q×n)-matrix, $\mathbf{1} \in R^n$ – vectors of ones

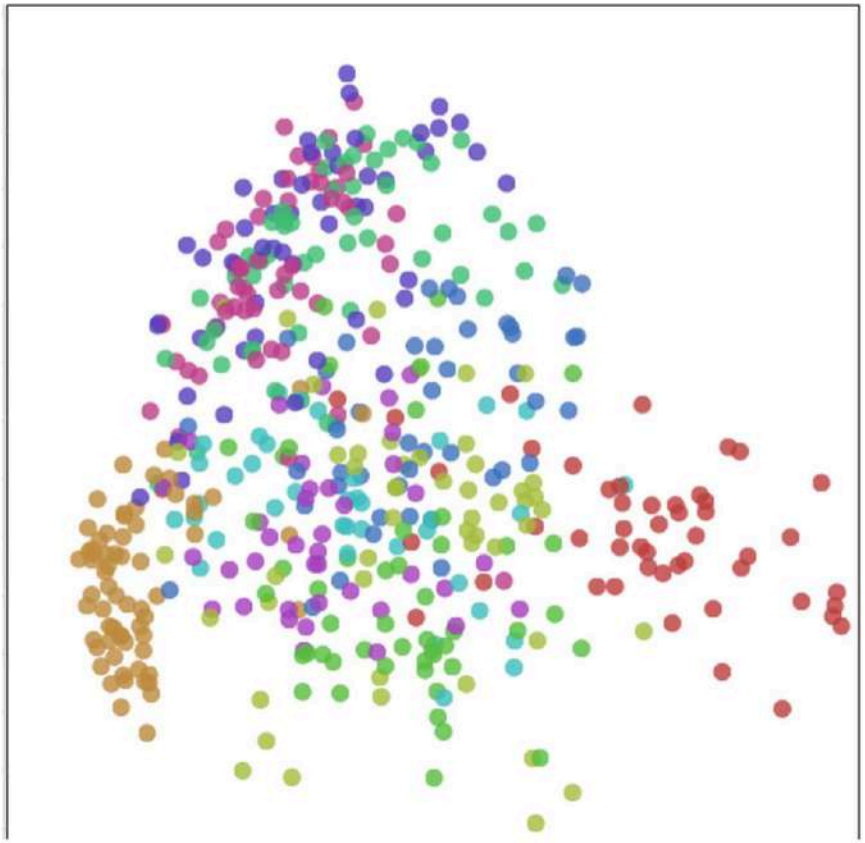
If $\rho(O_i, O_j) = \|X(O_i) - X(O_j)\|$, then MDS = PCA

Sammon mapping

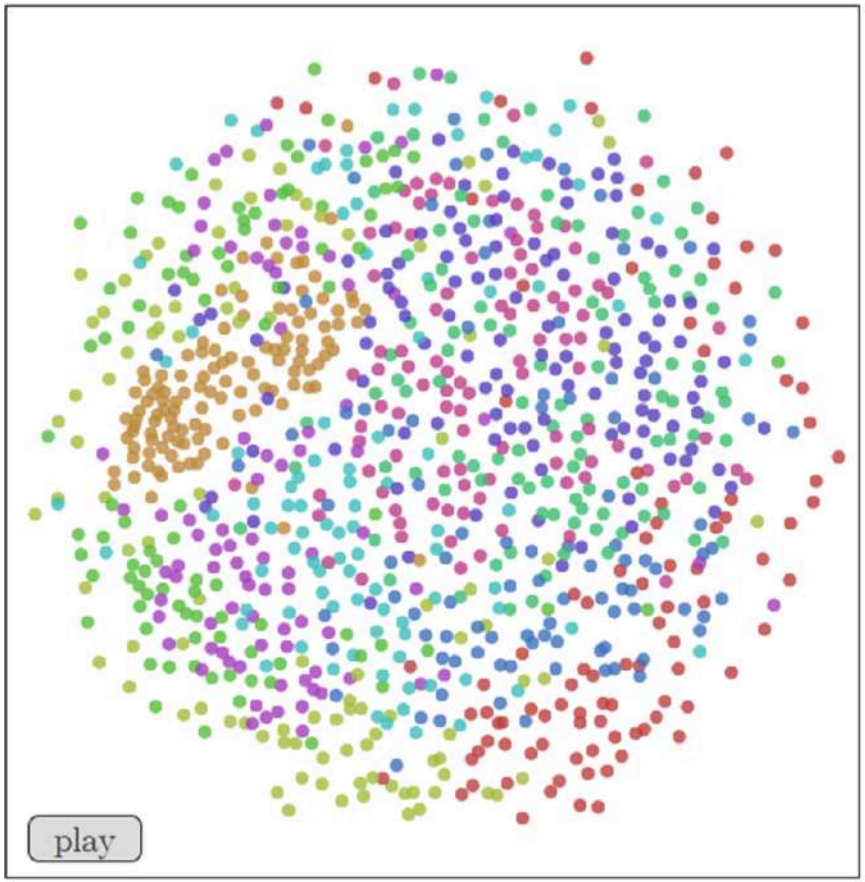
Common useful idea: local distances are often more important to preserve

$$\sum_{i,j} \frac{\left(\rho(O_i, O_j) - \left\| y_i - y_j \right\|\right)^2}{\left(\rho(O_i, O_j)\right)^2}$$

PCA vs. MDS

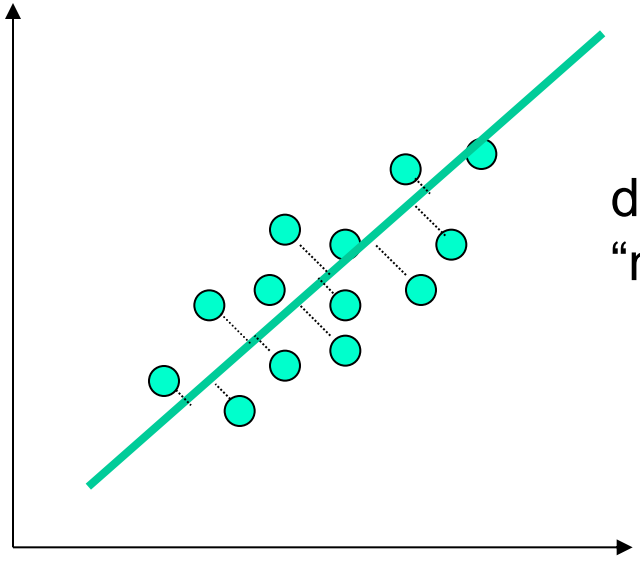


PCA



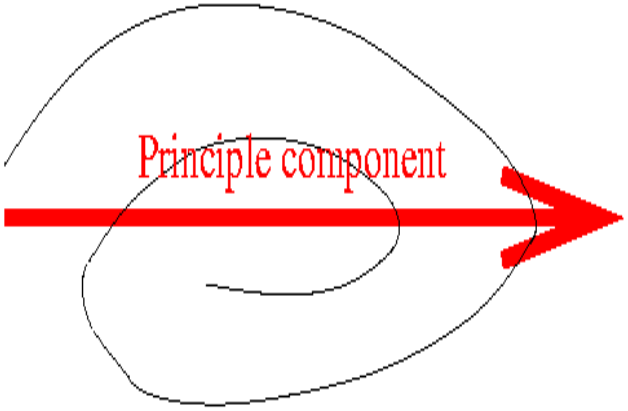
Sammon

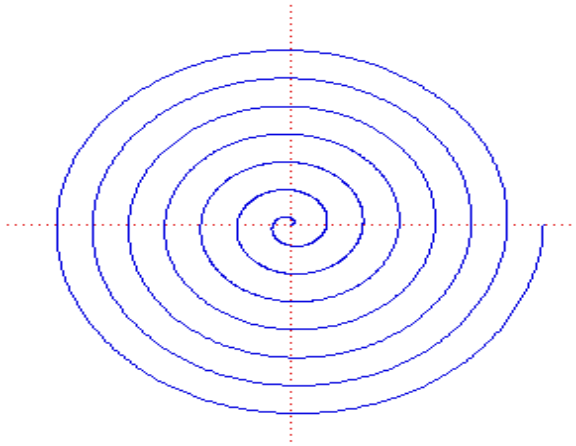
MDS, PCA, PP, MDS (for «Euclidean proximity») – linear methods,

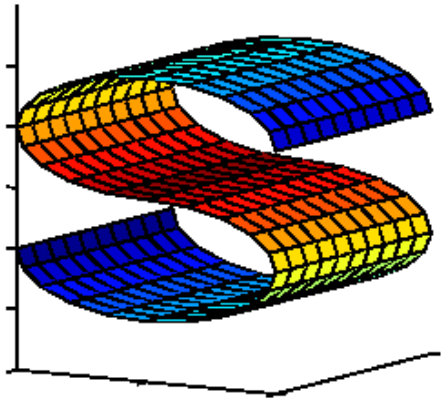


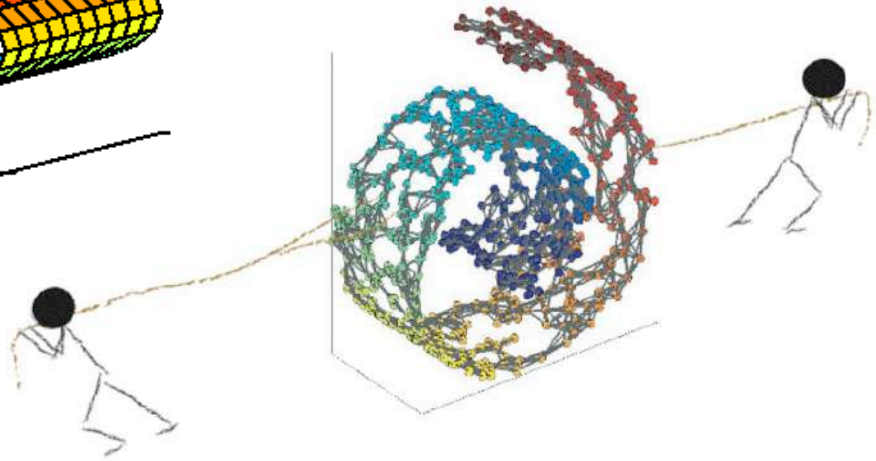
Spiral of Archimedes

do not work for
“nonlinear data”:

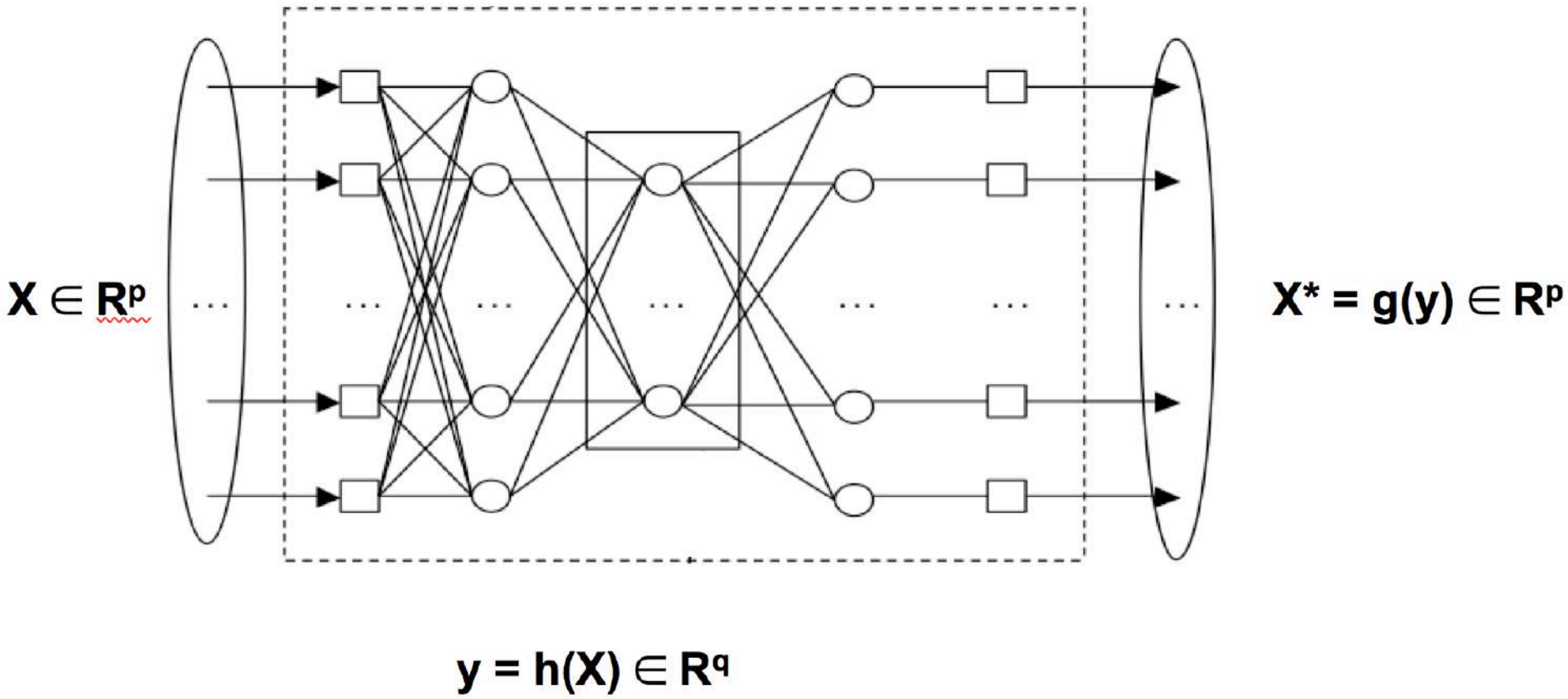




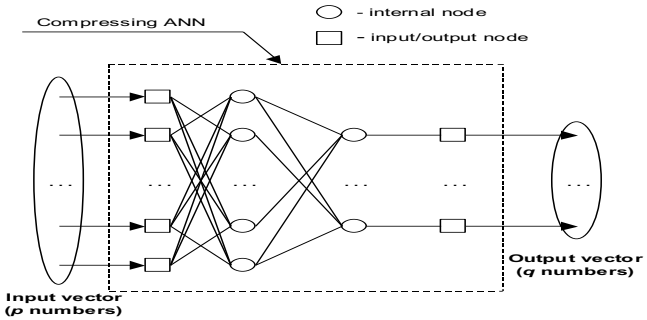


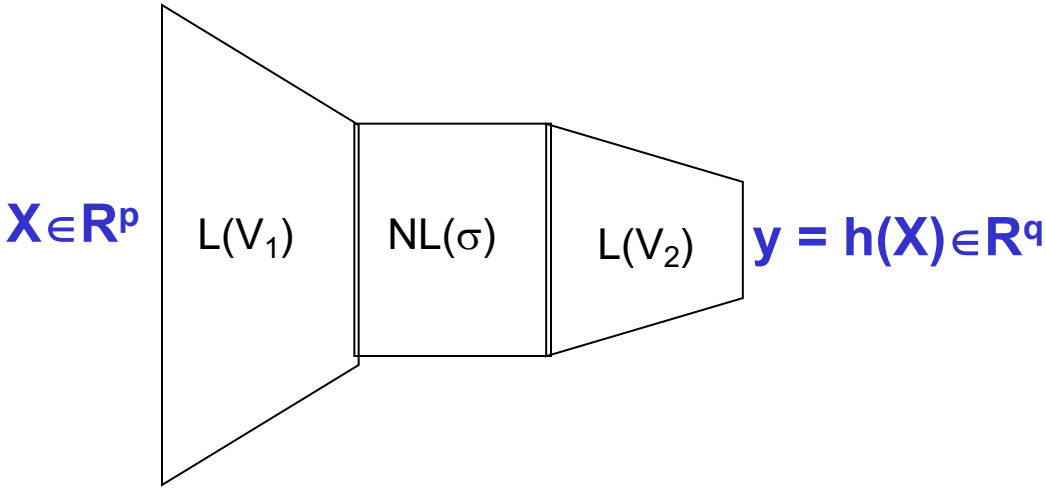


Replicative Neural Networks

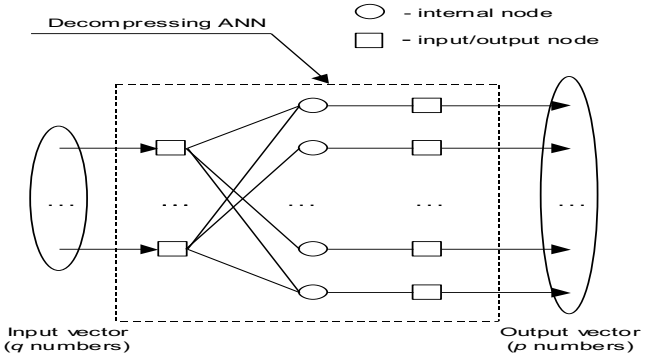


Compression transformation





Reconstruction transformation



$y \in R^q$ $\quad L(W_2) \quad NL(\sigma_R) \quad L(W_1)$ $\quad X^*=g(y) \in R^p$

Overview

- Intro
- Dimension Reduction Problem Statements
- PCA, MDS and Sammon Mapping, Autoencoders
- **ISOMAP and LLE**
- TDA for Time Series Analysis
- References

Nonlinear (local) DR methods

5) **Locally Linear Embedding, LLE**
6) **Laplacian Eigenmaps LE**
7) **Hessian Eigenmaps, HE**
8) **ISOmetric MAPing, ISOMAP**
9) Kernel PCA, KPCA - Spectral Embedding Algorithm, SEA
10) Riemannian Manifold Learning, RML
11) Local Tangent Space Alignment, LTSA, **…**

GRAPH: standard step for many “local” DR methods

Step 1. Construct neighborhoods (ε-Neighborhoods, k Nearest Neighbors)

$$U(X) = U(X|\rho, \varepsilon) = \{X_i \in \mathbf{X}_n: \rho(X, X_i) \le \varepsilon\}$$

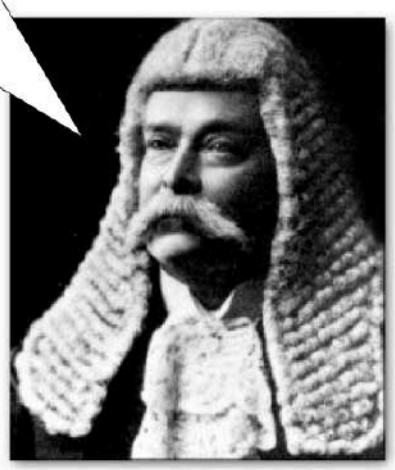
$\Rightarrow$ **Graph** $\Gamma(\mathbf{X}_n) = (\mathbf{X}_n, \mathbf{V})$: $(X_i, X_j) \in \mathbf{V} \Leftrightarrow X_j \in U(X_i)$ и $X_i \in U(X_j)$

Step 2. Construct weights $w_{ij} = w(X_i, X_j) = w(v)$, $v = (X_i, X_j) \in \mathbf{V}$

$w_{0,ij} = 1$ for all $(X_i, X_j) \in V$, $\quad w_{0,ij} = 0$ in other cases

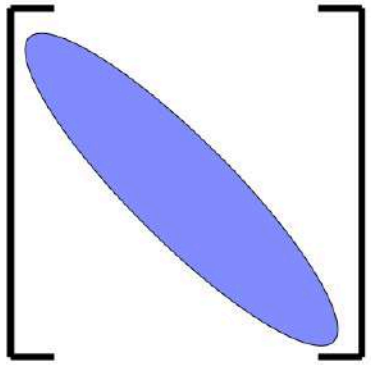
ISOMAP

Infer a distance matrix using distances along the manifold.

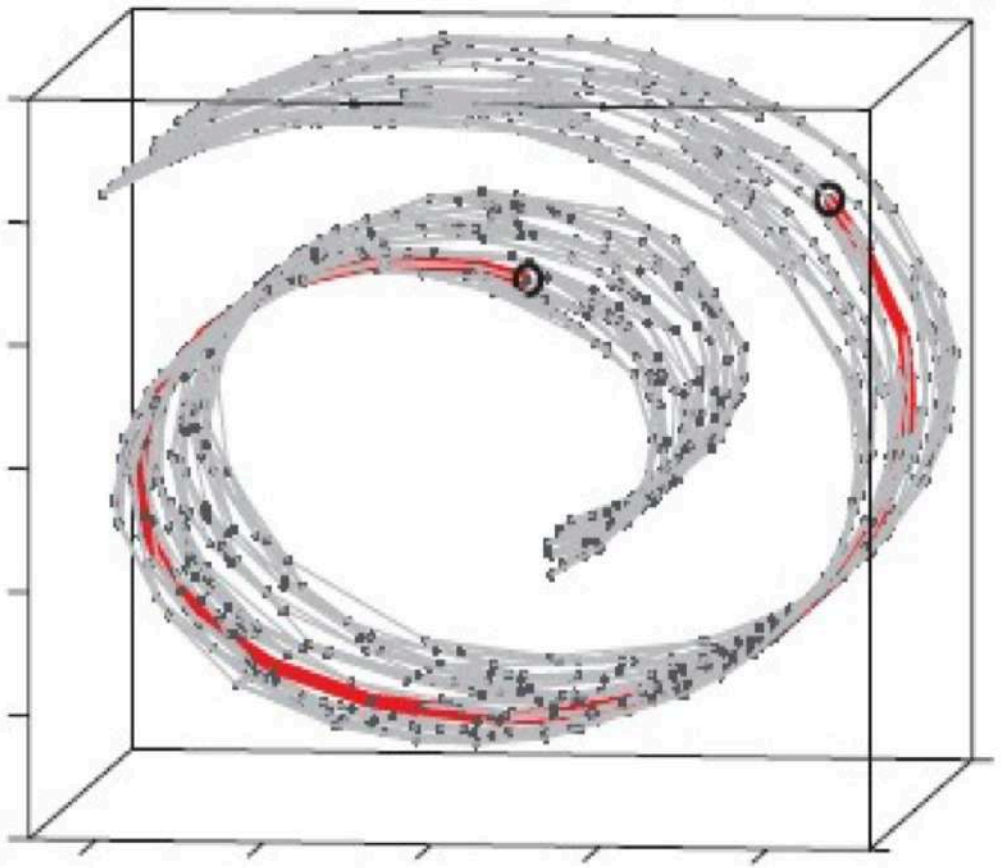


ISOMAP

1. Build a sparse graph with K-nearest neighbors

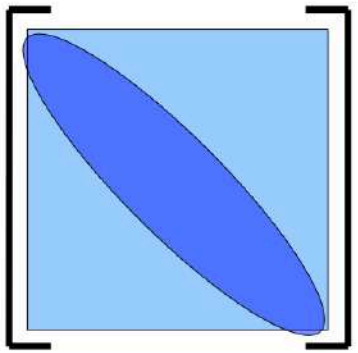
$D_g =$ 

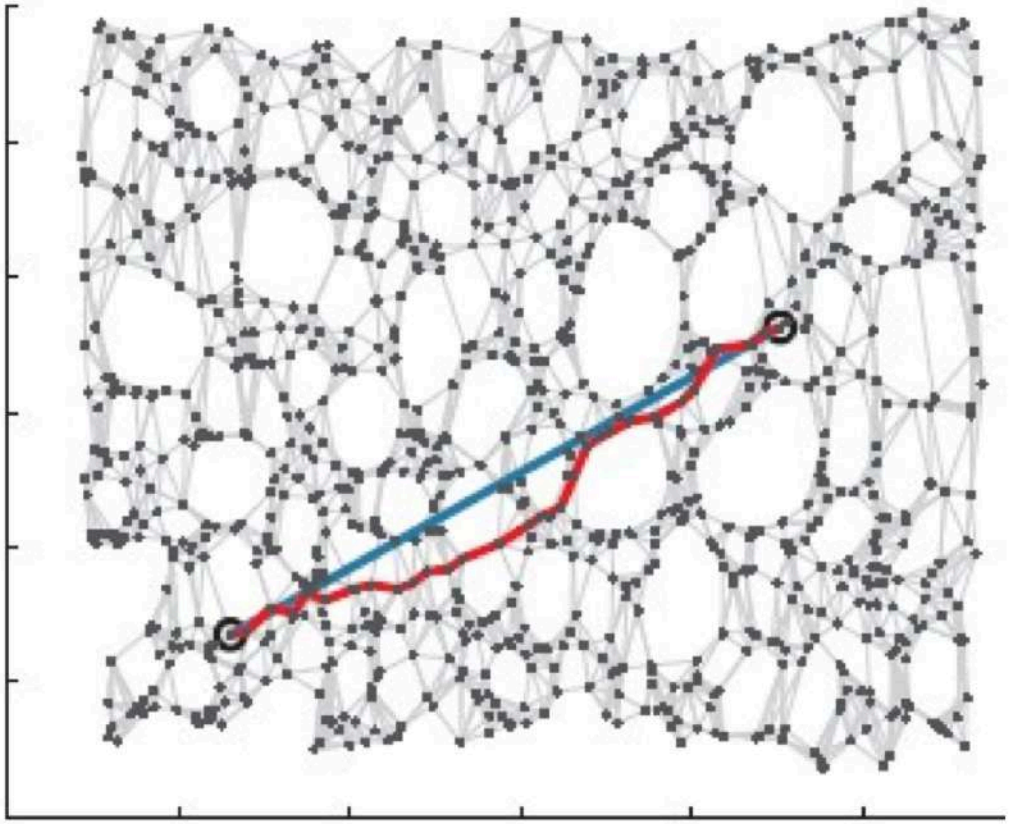
(distance matrix is sparse)



ISOMAP

2. Infer other interpoint distances by finding shortest paths on the graph (Dijkstra's algorithm).

$D_g =$ 



ISOMAP

Usual MDS

3. Build a low-D embedded space to best preserve the complete distance matrix.

Error function:

inner product distances in graph

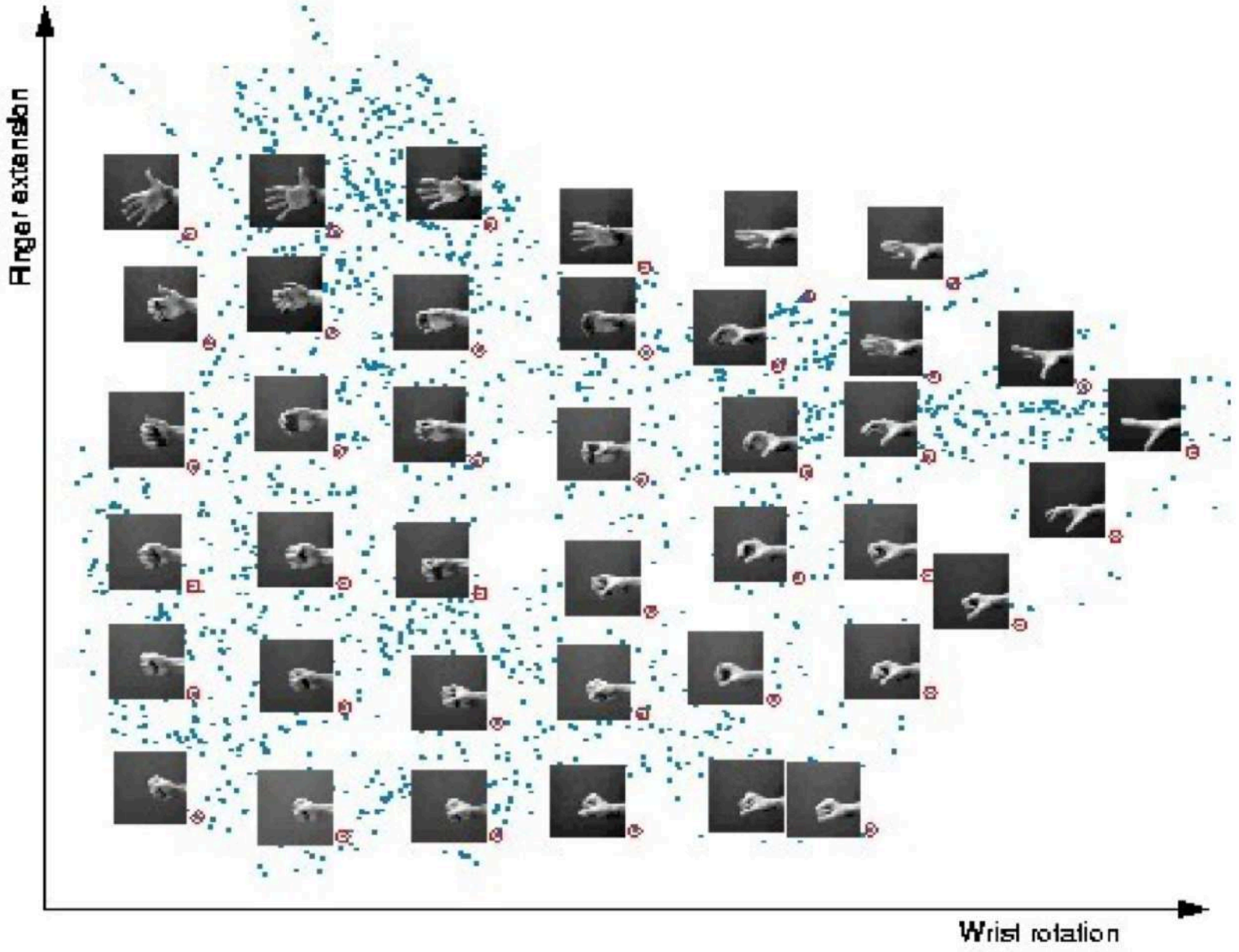
inner product distances in new coordinate system

L2 norm

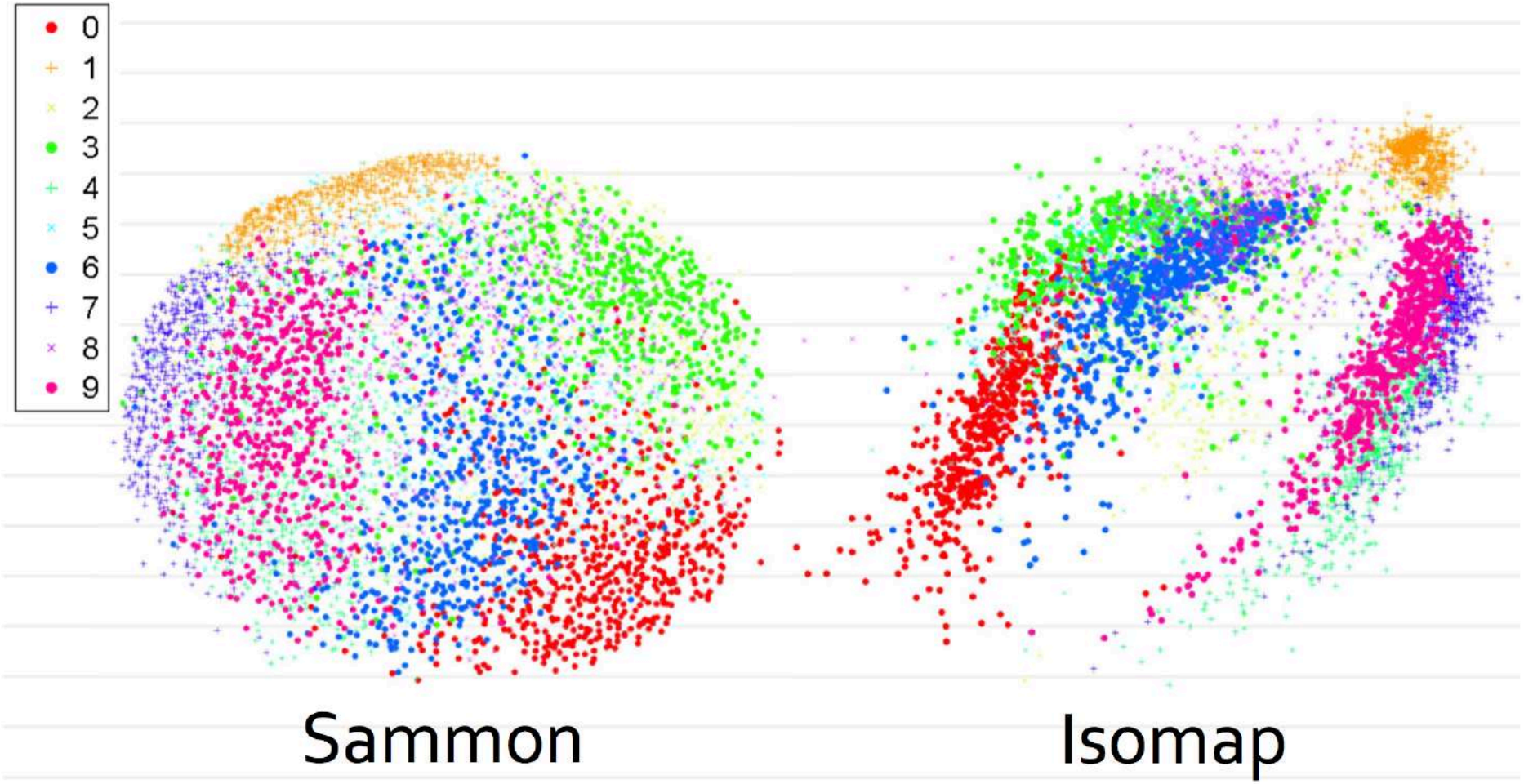
$$E = \|\tau(D_G) - \tau(D_Y)\|_{L^2}$$

Solution – set points Y to top eigenvectors of D_g

Isomap results: hands



MNIST: Sammon vs. ISOMAP



Isomap: pro and con

- preserves global structure

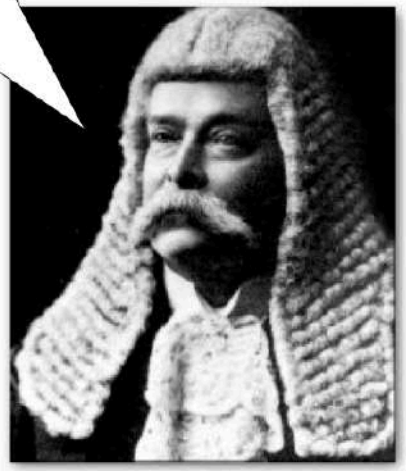
- few free parameters

- sensitive to noise, noise edges

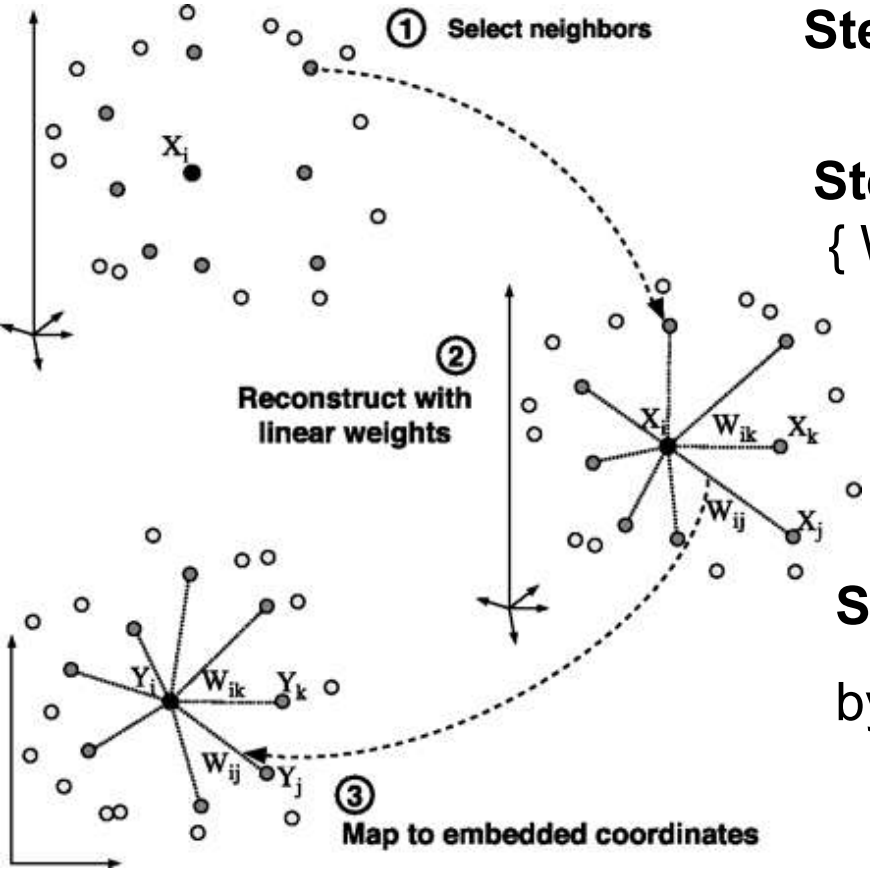
- computationally expensive (dense matrix eigen-reduction)

Locally Linear Embedding

Find a mapping to preserve local linear relationships between neighbors



Locally Linear Embedding, LLE (L.K. Saul, et al., 2000)



Step 1. K nearest neighbors

Step 2. Get "Baricentric" coordinates $\{ W_{ji}\}$ by minimizing

$$J_1(\mathbf{W}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{K} W_{ji} \mathbf{x}_{j(i)} \right\|^2$$

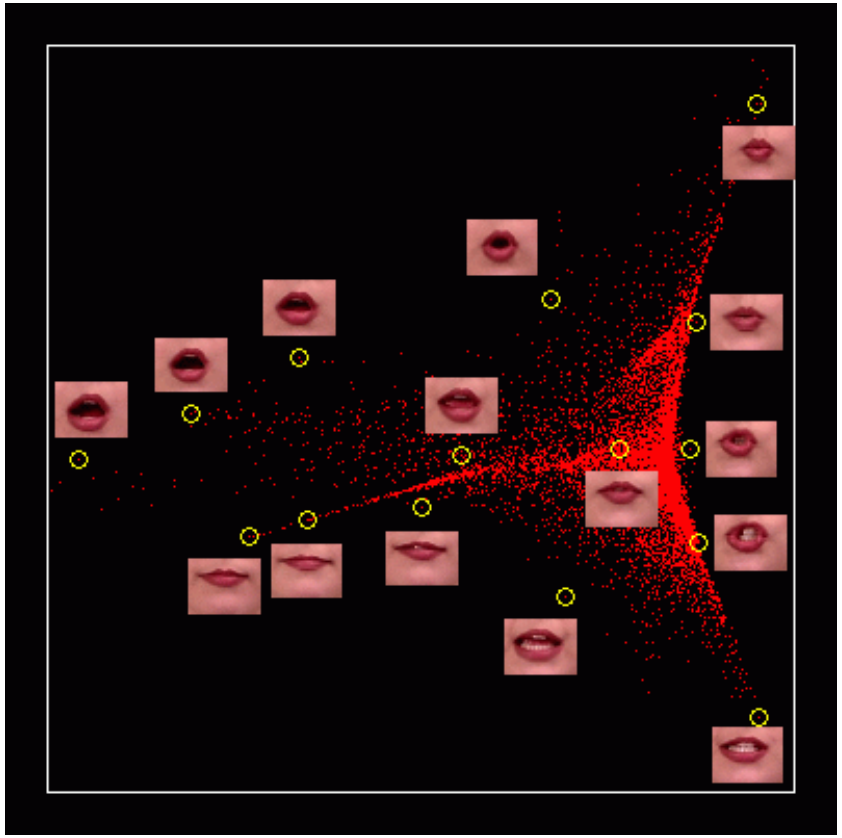
Step 3. Get $\{y_1, y_2, \ldots, y_n\} \in R^q$

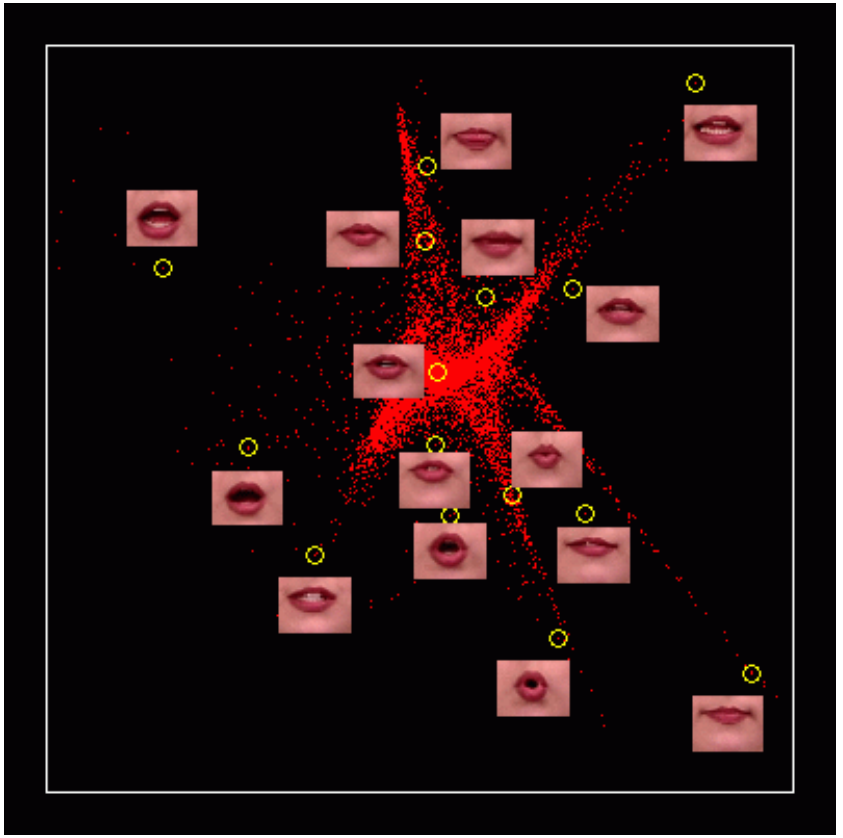
by minimizing

$$J_2(\mathbf{Y}) = \sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j=1}^{N} W_{ji} \mathbf{y}_{j(i)} \right\|^2$$

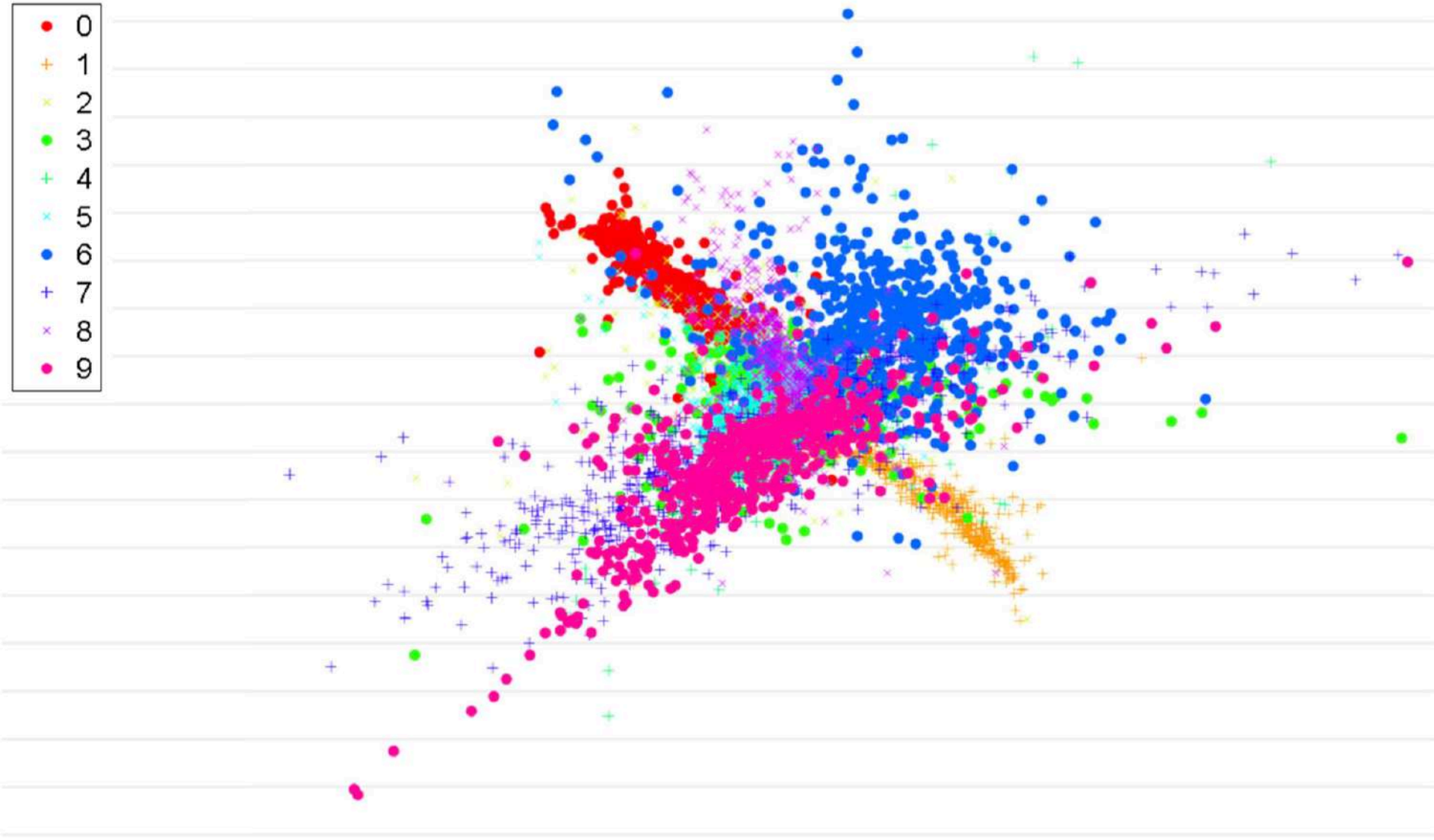
with standard normalizing conditions

LLE: Lips pictures





LLE: MNIST



LLE: pro and con

- no local minima, one free parameter

- incremental & fast

- simple linear algebra operations

- can distort global structure

Overview

- Intro
- Dimension Reduction Problem Statements
- PCA, MDS and Sammon Mapping, Autoencoders
- ISOMAP and LLE
- **TDA for Time Series Analysis**
- References

Demand Forecasting

Lots of customers

Limited resources

- Identifying **valuable customers** and **predicting demand**
- **Unified process** of marketing and demand forecasting

Use of the customer segmentation for demand forecasting

- **Millions** of time series to predict
- **Novelty** of the cloud computing, thus **historic data is limited**

This work proposes three different methods:

RFM, TS RFM and TDA RFM

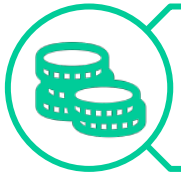
What is RFM?

Recency:

"**When** was the **last time** that a user was active?"

Frequency:

"**How frequently** has a user been active?"



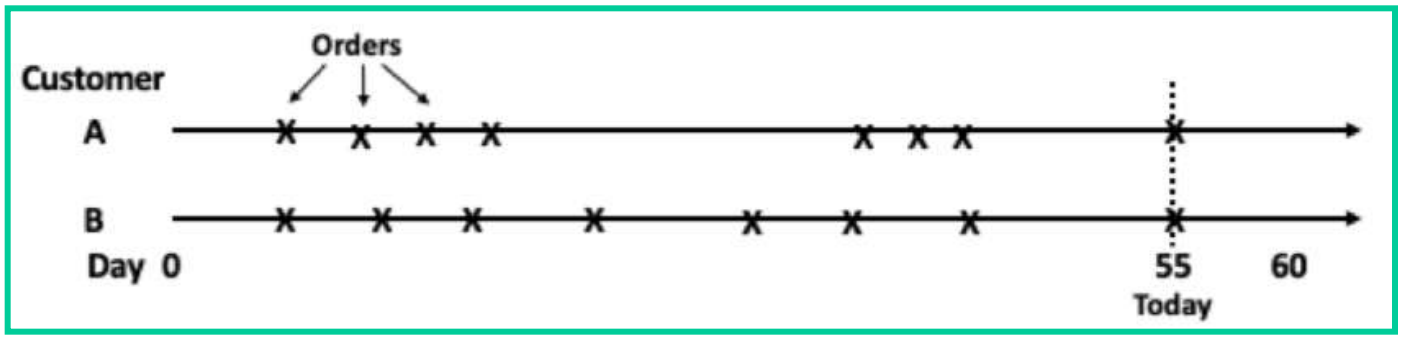
Monetary:

"**How much revenue** has this user generated?"

RFM score:

3 numbers

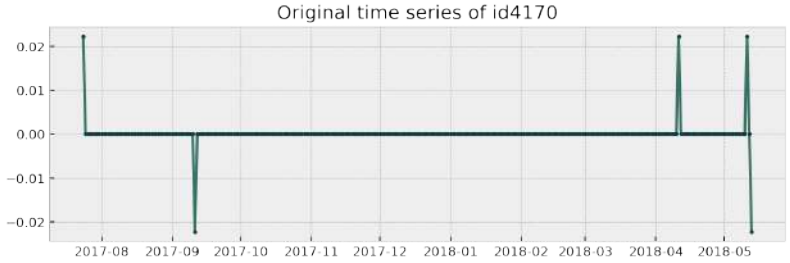
Why can it fail?

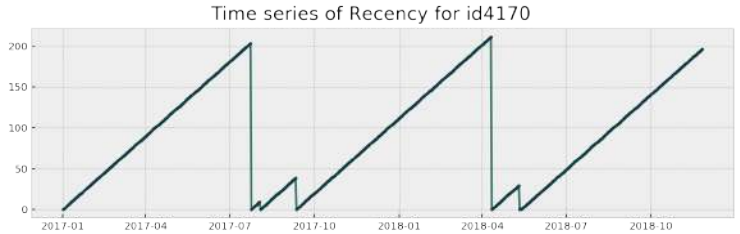


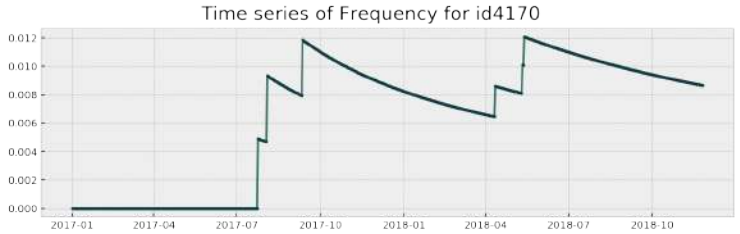
Two customers share the **same** RFM score

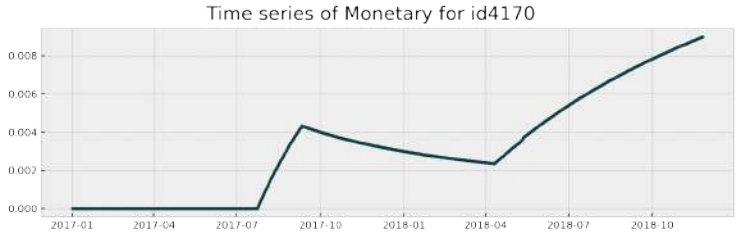
But **A** is likely to be **different** from **B**

How to improve it?







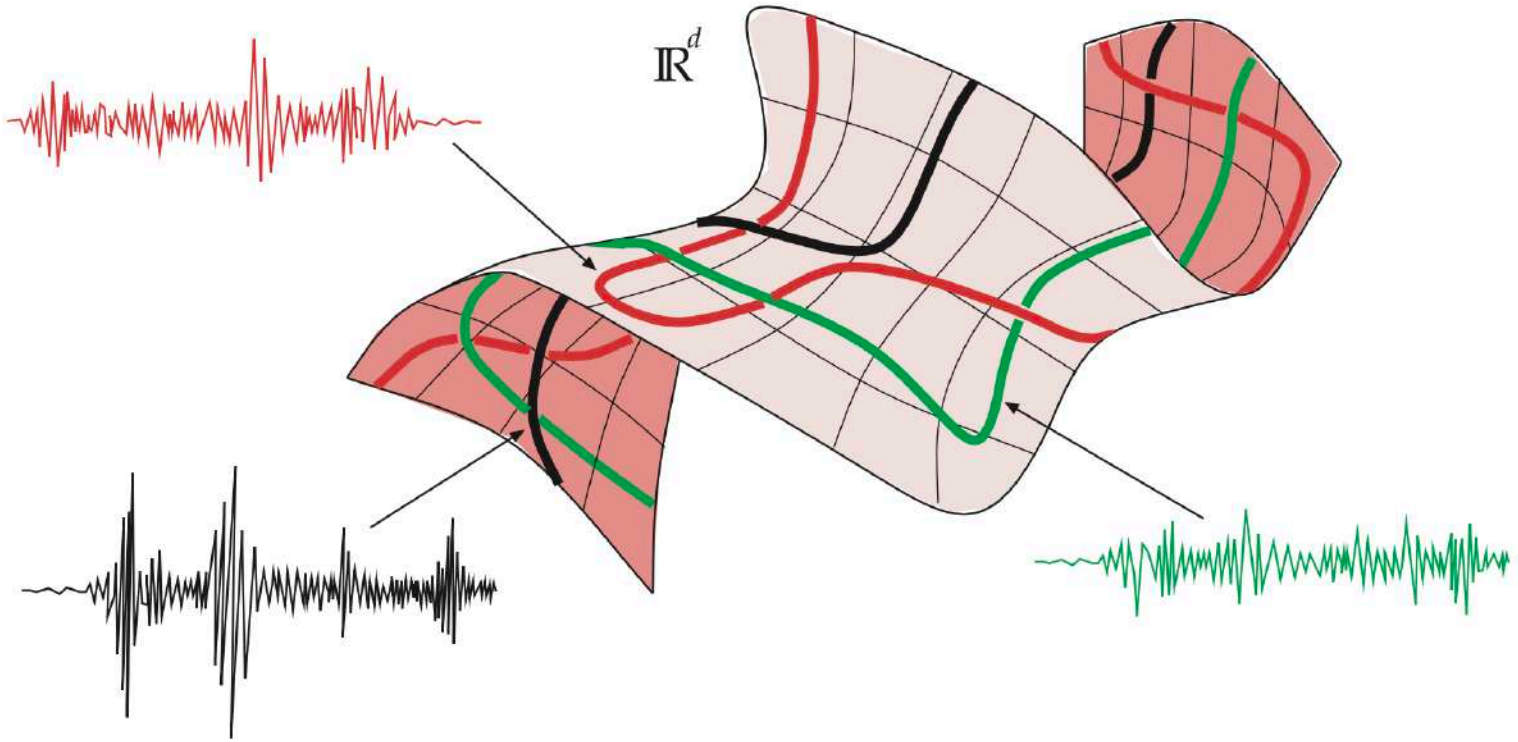


Point clouds representation

Sliding windows of size w are applied to each time series $X_j = \{x_1, x_2, \ldots, x_n\}$ in order to create point clouds Z_j of points in $\mathbb{R}^w$

$$Z_j = \begin{pmatrix} x_1, x_2, \ldots, x_w \\ \vdots \\ x_{n-w}, x_{n-w+1}, \ldots, x_n \end{pmatrix}$$

Trajectories from three different time-series lie along a smooth manifold

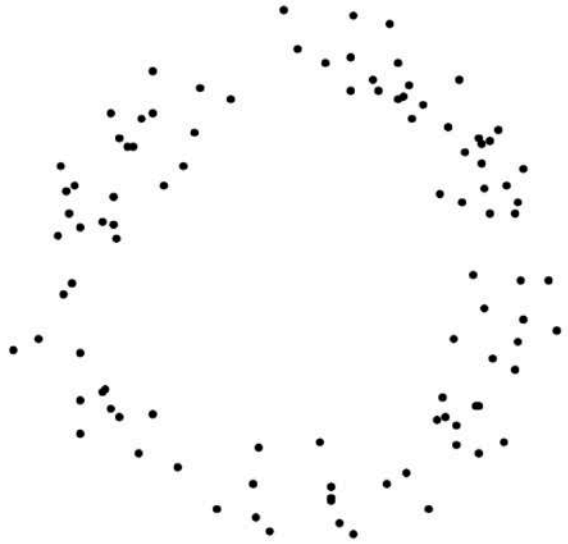


Picture credit to [Taylor et al. Geophys. J. Int. (2011) 185, 435–452]

Persistence homologies

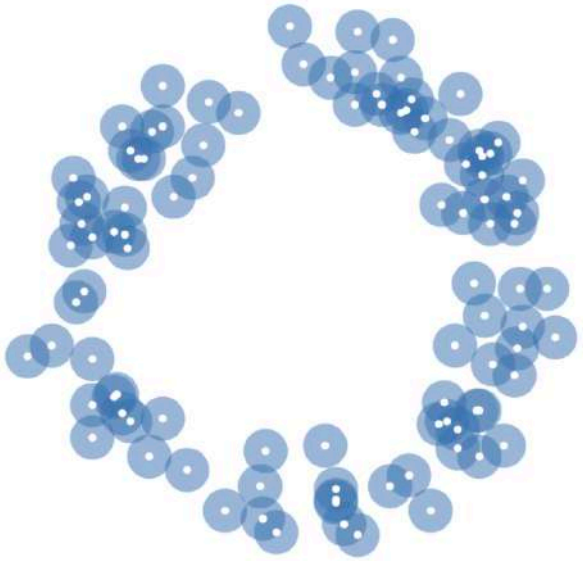
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



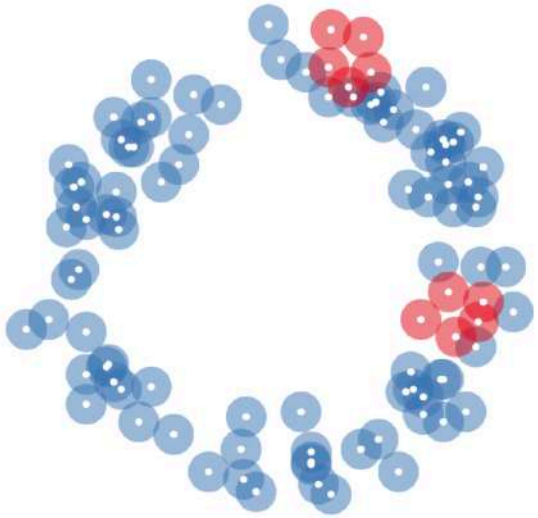
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



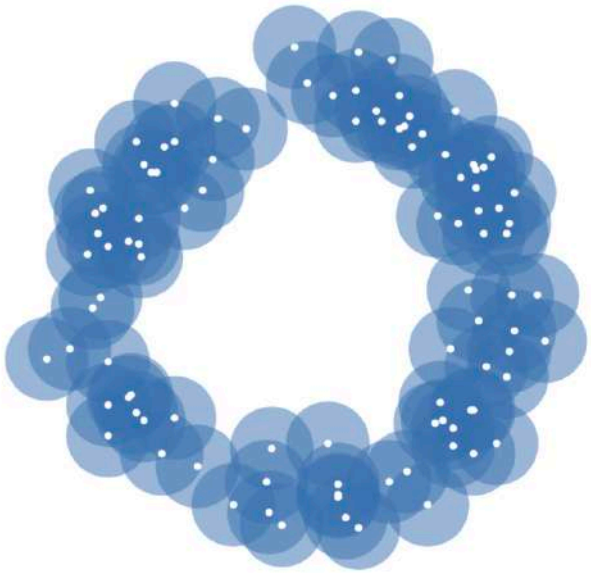
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



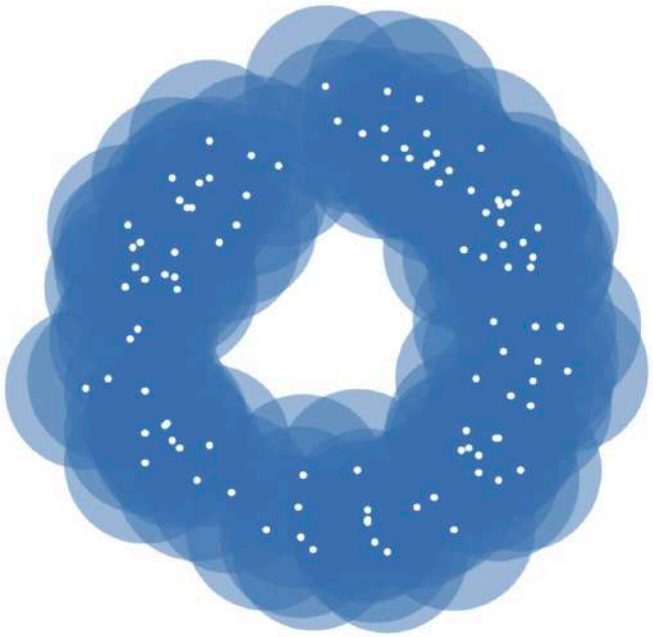
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



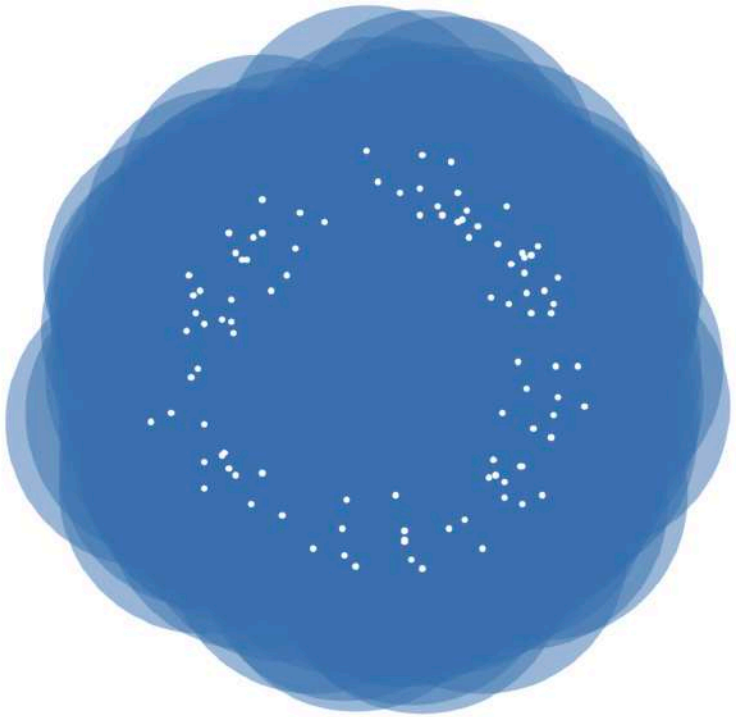
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



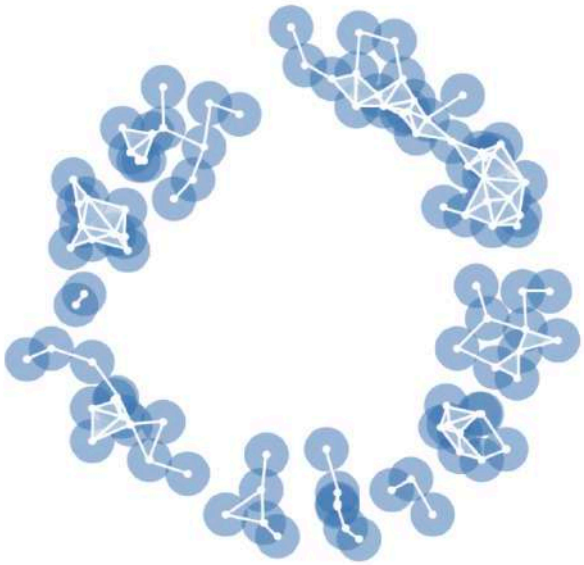
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



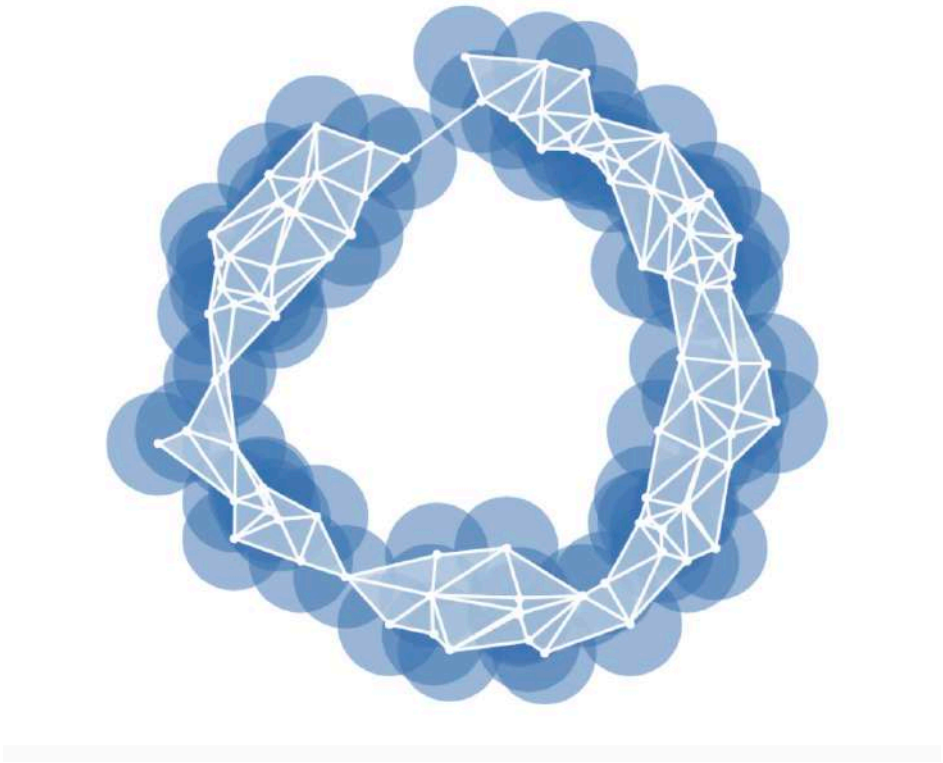
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies

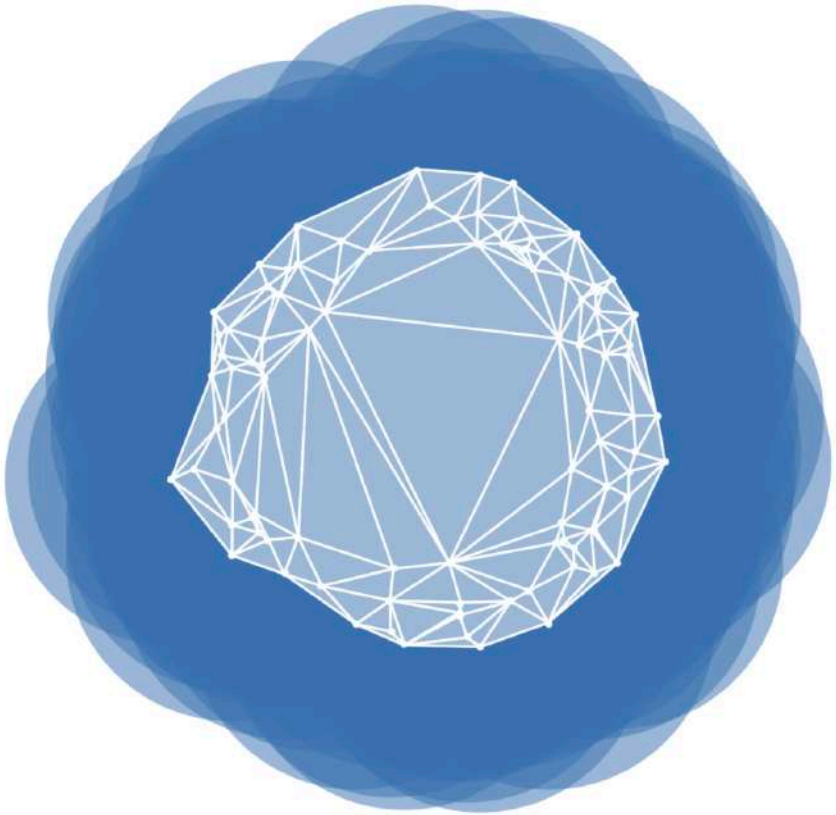


Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies

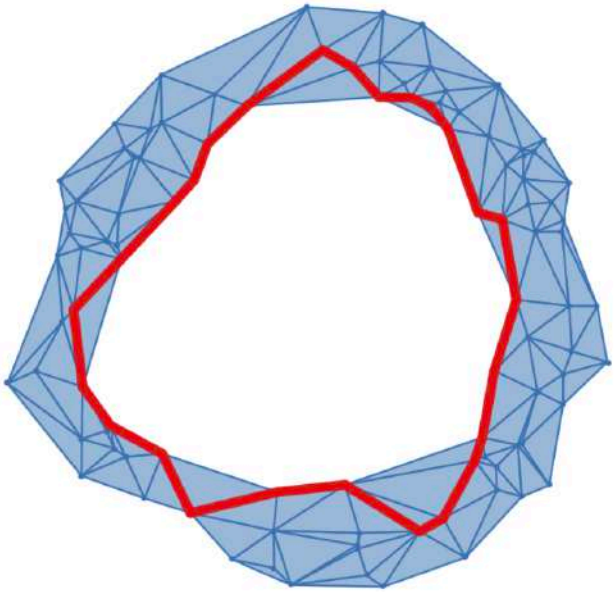
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



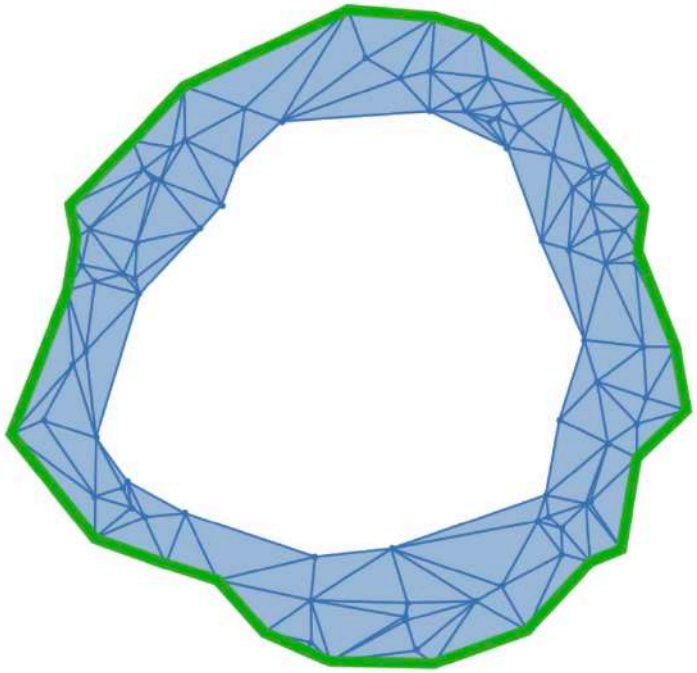
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



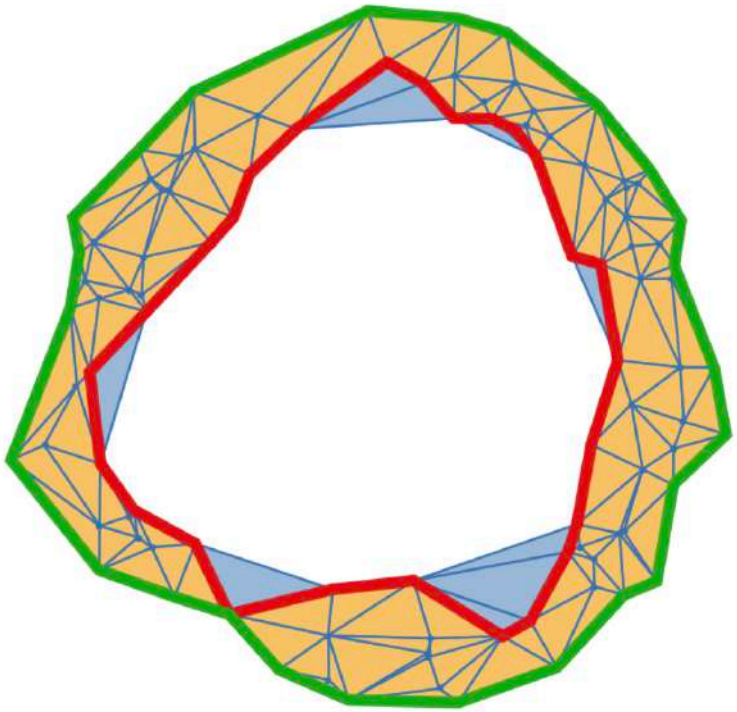
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



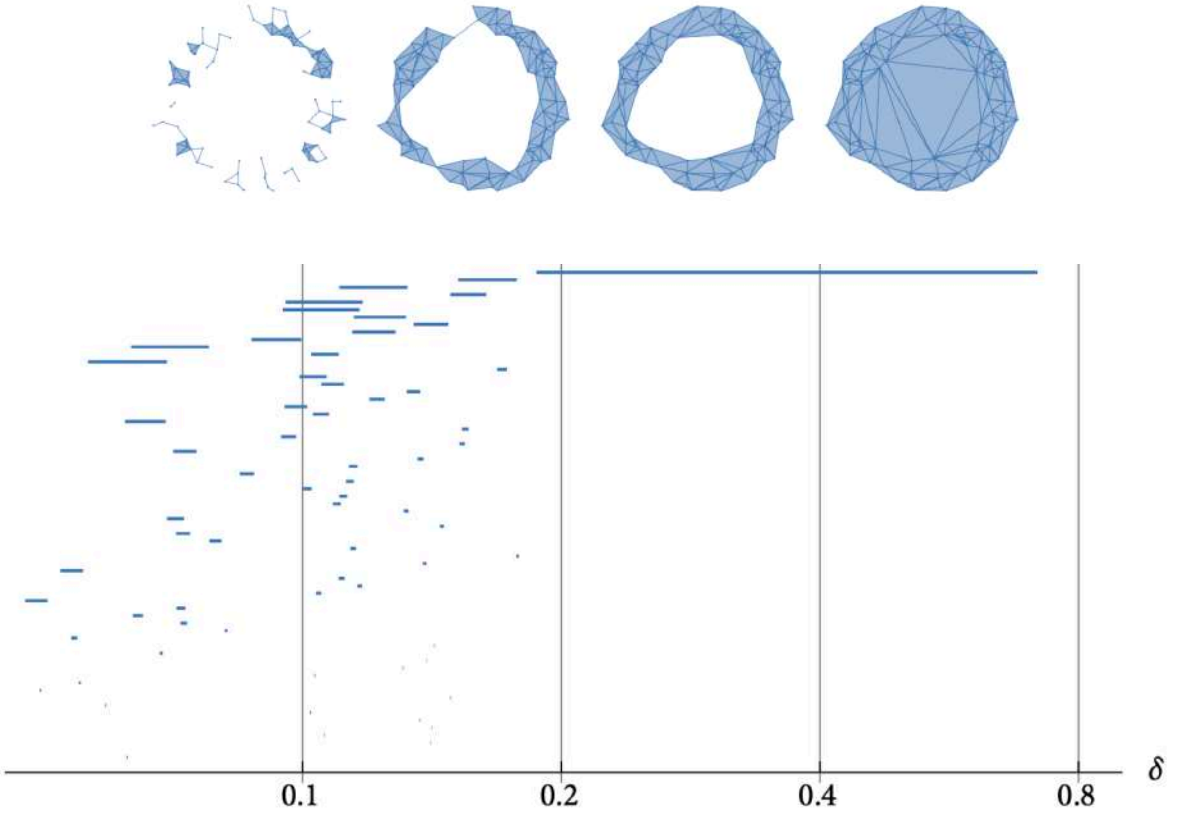
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



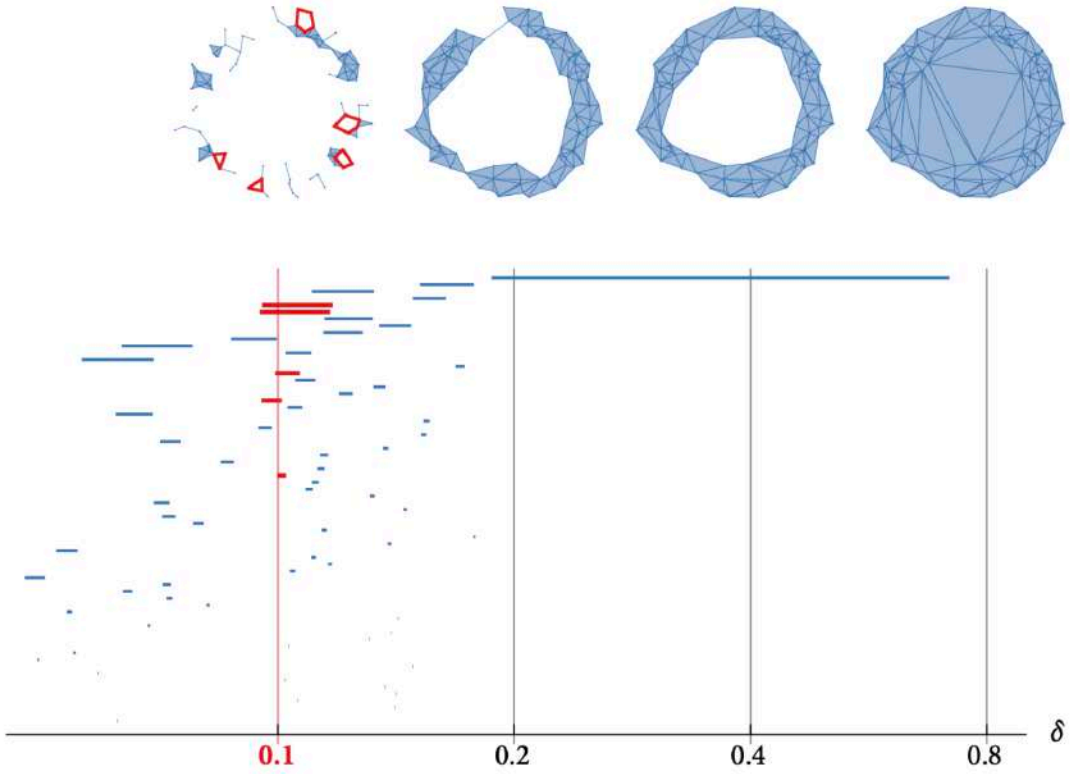
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



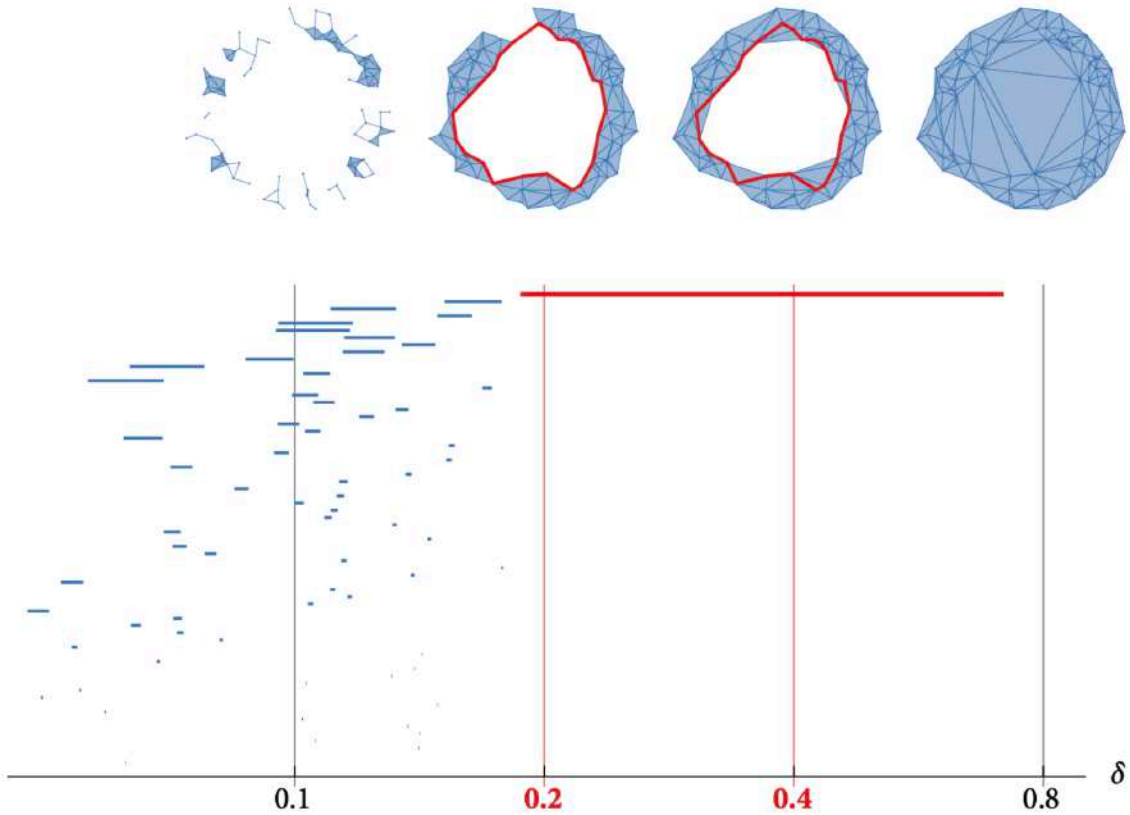
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence homologies



Picture credit to [Ulrich Bauer, MLSS 2019]

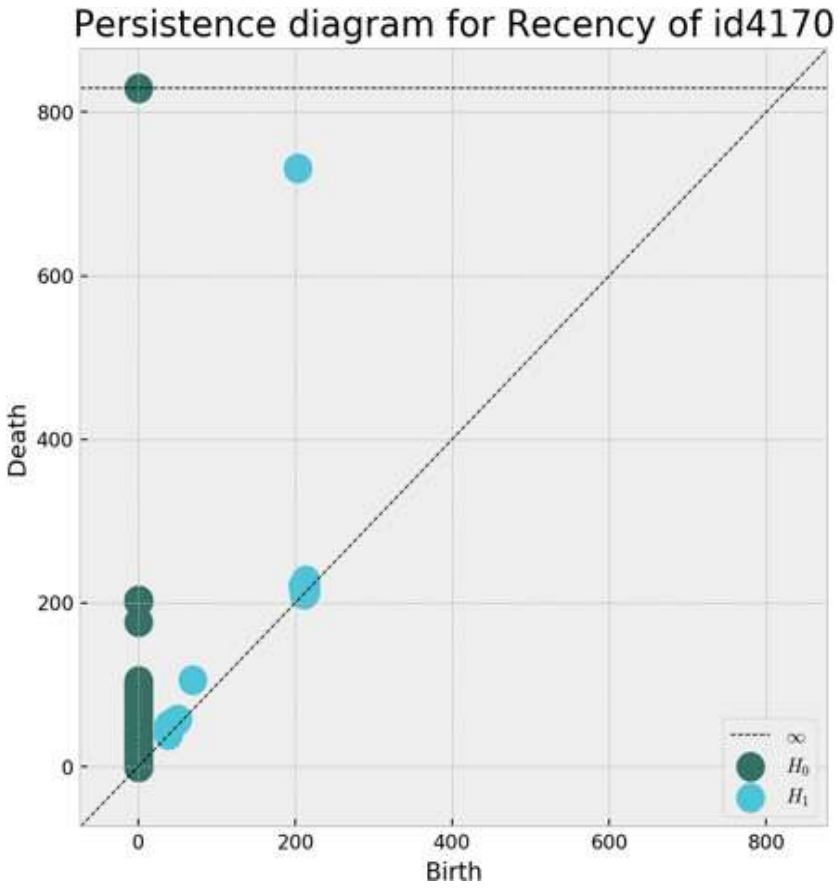
Persistence homologies



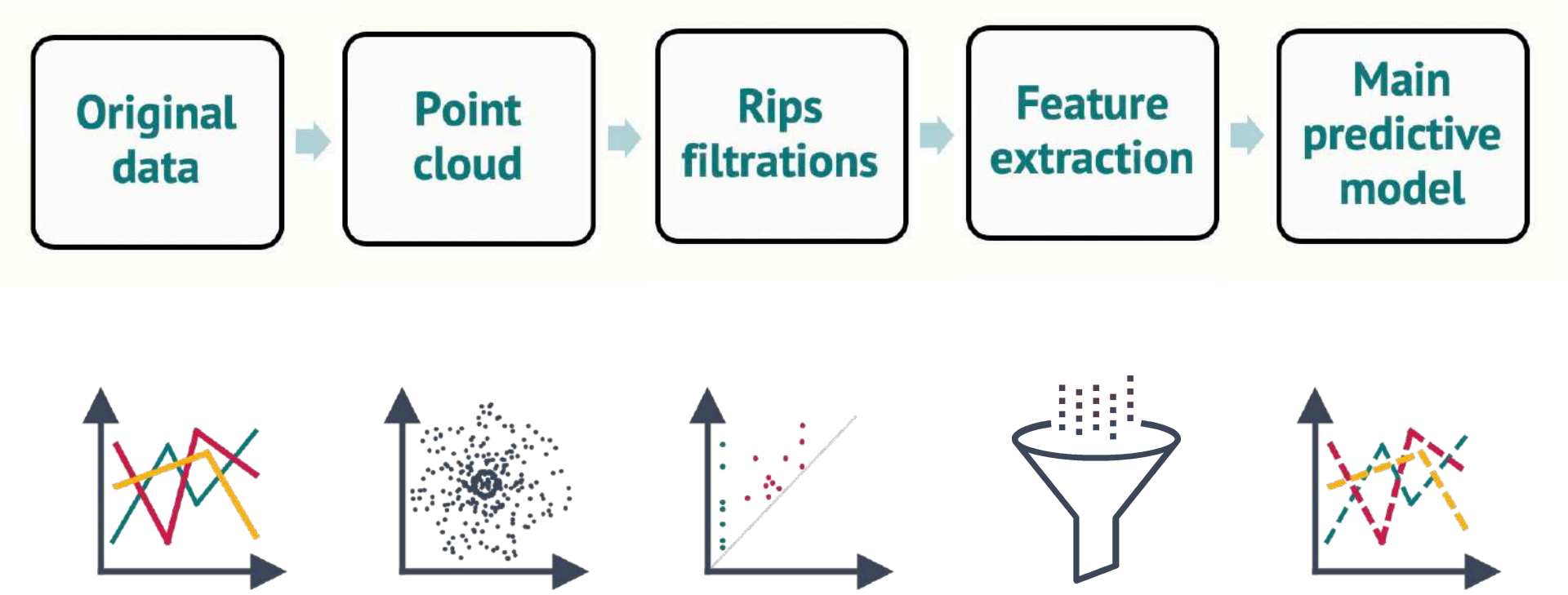
Picture credit to [Ulrich Bauer, MLSS 2019]

Persistence diagram

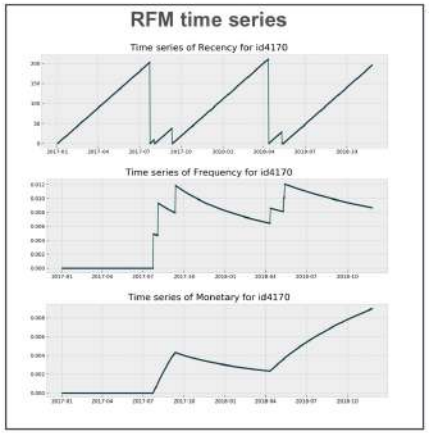
- Representation of **topological structure**
- Each point corresponds to a **homology**
- Each homology has **birth and death radius**
- Allows to extract **numerical features**

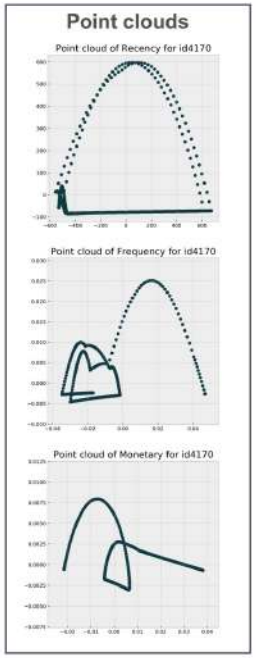


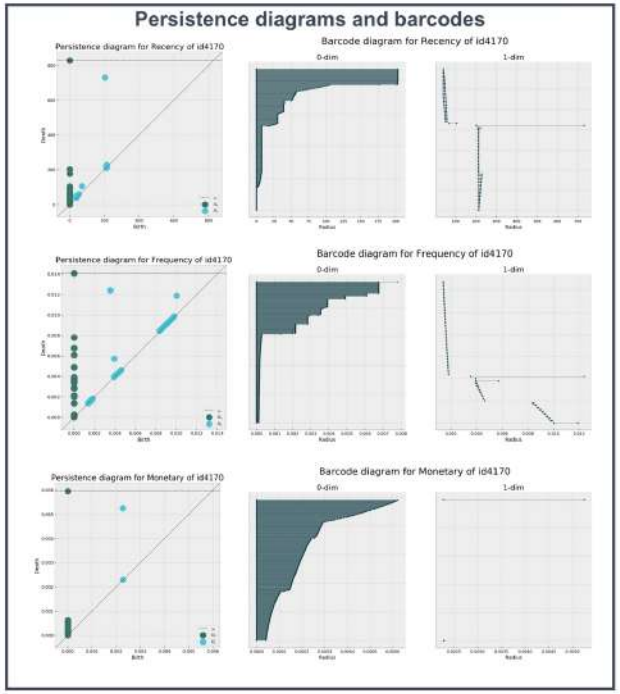
TDA pipeline

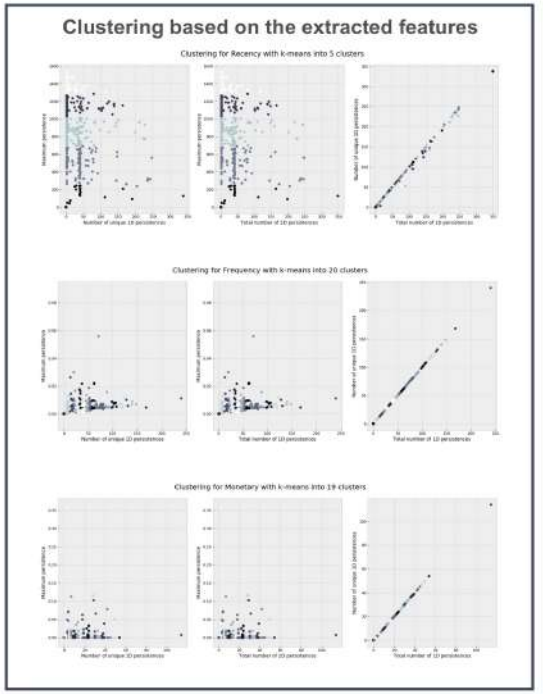


TDA RFM pipeline



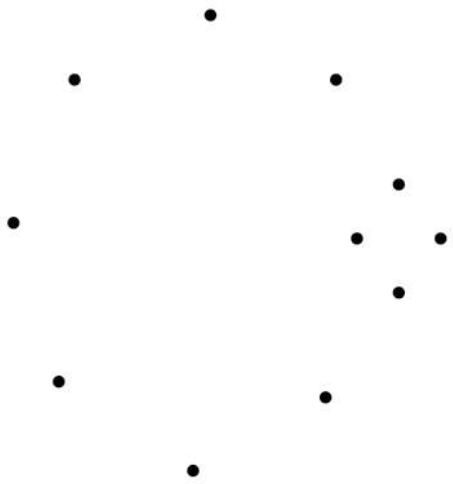






Why TDA?

- TDA has strong **mathematical foundation**
- Allows to **transform** time series into **features**, to use in common statistical tools
- **Unsupervised** learning with **no initial assumptions** on data
- TDA is **robust to noise** in data, which is especially useful for **financial data** and **client activities**



TS RFM pipeline

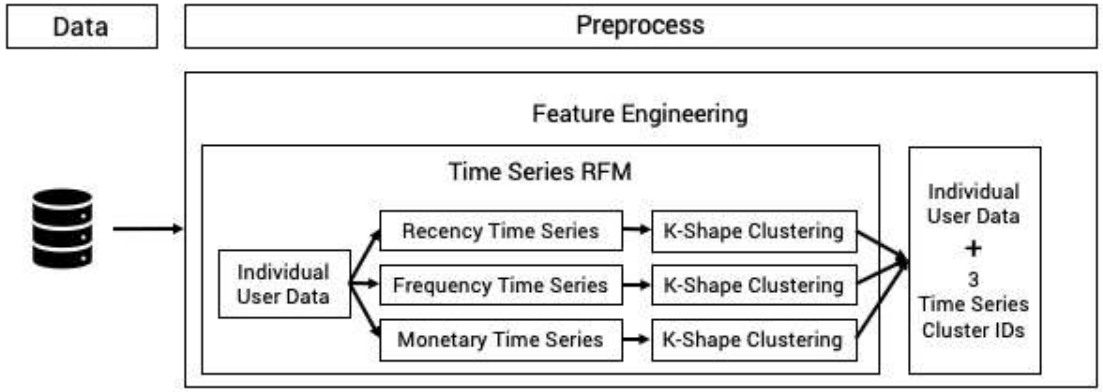
Obtain RFM time series

⇩

Apply k-shape clustering for each set of time series (RFM)

⇩

Use this clusters in further pipeline as features



Summary

RFM
(marketing person)

- Obtains RFM score as **3 numbers** for **each time series**
- Uses them in gradient boosting

TS RFM
(engineer)

- Creates **the time series of RFM scores**
- Uses **k-shape** clustering for each type of time series
- Uses this features in gradient boosting

TDA RFM

- Creates **the time series of RFM scores**
- **TDA pipeline** and feature extraction
- **k-means clustering** based on the features
- Gradient boosting on clusters

Datasets

CDnow — online store (public)

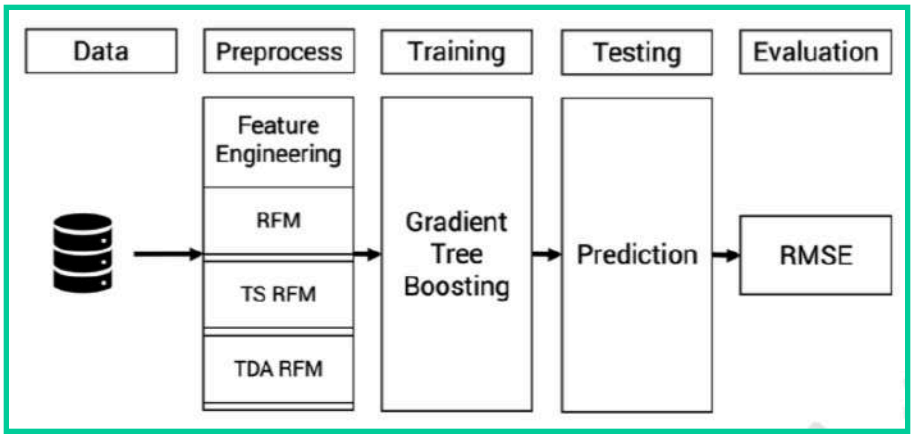
Bimbo — 2000 retail points of bakery (public)

Cloud — B2B cloud computing company (private)

Hospitality — B2B digital company in the goods procurement sector (private)

Experiments

- **Gradient boosting** to predict the value of the next activity
- $\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Predicted - Actual)^2}$



Results

- Pure **RFM** is **inferior** to all othe models
- Modified into **time series RFM** and **TDA RFM** are **superior** to others and **harshly decrease** the error
- Analysis of the **difference** between TDA and TS RFM is required

Dataset	Model	Error
CDNow	No RFM	13
	RFM	12,56
	TS RFM	**3**
	TDA RFM	18,87
Bimbo	No RFM	97
	RFM	172
	TS RFM	291
	TDA RFM	**18,97**
Cloud	No RFM	3,56
	RFM	3,98
	TS RFM	0,05
	TDA RFM	**0,03**
Hospitality	No RFM	219
	RFM	240
	TS RFM	**155**
	TDA RFM	275

Further steps

- **When** TDA is **better**
- **Too small** time series — how to **handle**
- Computational **intensity**
- **Other** statistics — which ones to use

- More applications
- Some other insights

Overview

- Intro
- Dimension Reduction Problem Statements
- PCA, MDS and Sammon Mapping, Autoencoders
- ISOMAP and LLE
- TDA for Time Series Analysis
- **References**

References

1. Yu. Ma, Yun. Fu. Manifold learning and applications, CRC Press 2011
2. Kuleshov A., Bernstein A., Burnaev E., Yanovich Yu. Machine Learning in Appearance-based Robot Self-localization, ICMLA, 2017
3. ShahRukh Athar, Evgeny Burnaev, Victor Lempitsky. Latent Convolutional Models. ICLR, 2019
4. Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014
5. A. Kuzina, E. Egorov, E. Burnaev. Bayesian generative models for knowledge transfer in MRI semantic segmentation problems. Journal: Frontiers in Neuroscience, section Brain Imaging Methods, 2019
6. Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitry Vetrov, Max Welling, The Deep Weight Prior, ICLR, 2019
7. A. Kuleshov, A. Bernstein and E. Burnaev. Kernel Regression on Manifold Valued Data, DSAA, 2018
8. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation Neural Computation, 2002, 15, 1373-1396
9. Carreira-Perpiñán, M. A. A review of dimension reduction techniques Dept. Computer Science, Univ. Sheffield, 1997
10. Gorban, A. N.; Kgl, B.; Wunsch, D. C. & Zinovyev, A. Principal Manifolds for Data Visualization and Dimension Reduction Springer, 2007
11. Zhang, Z. & Zha, H. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment SIAM J. Scientific Computing, 2004, 26, 313-338
12. John A. Lee, Michel Verleysen. Nonlinear Dimension Reduction. Springer, 2007
13. Frédéric Chazal, Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. https://arxiv.org/abs/1710.04019
14. Ulrich Bauer. Topological Data Analysis short course. https://smiles.skoltech.ru/event-recap https://github.com/mlss-skoltech https://www.youtube.com/watch?v=EeduNPexICk