# Similarity: Numerical Data

Mining Massive Datasets

Prof. Carlos Castillo

Topic 06

Universitat Pompeu Fabra Barcelona

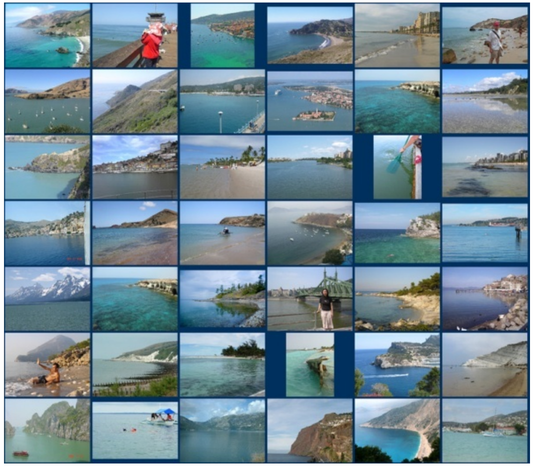
# Main Sources

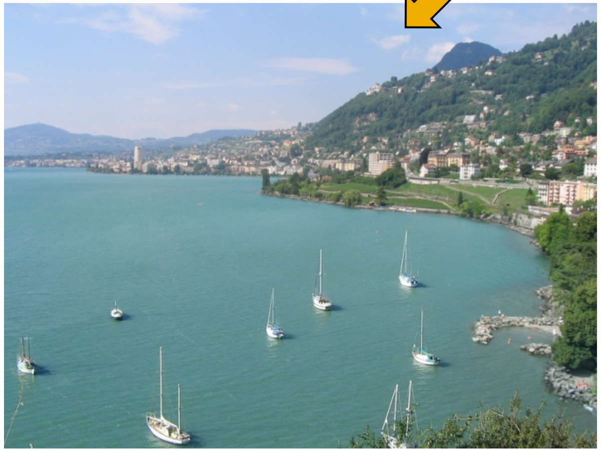
- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 3) + slides by Lijun Zhang
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Section 2.4)
- Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 3)

# Example: scene completion

Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 3)

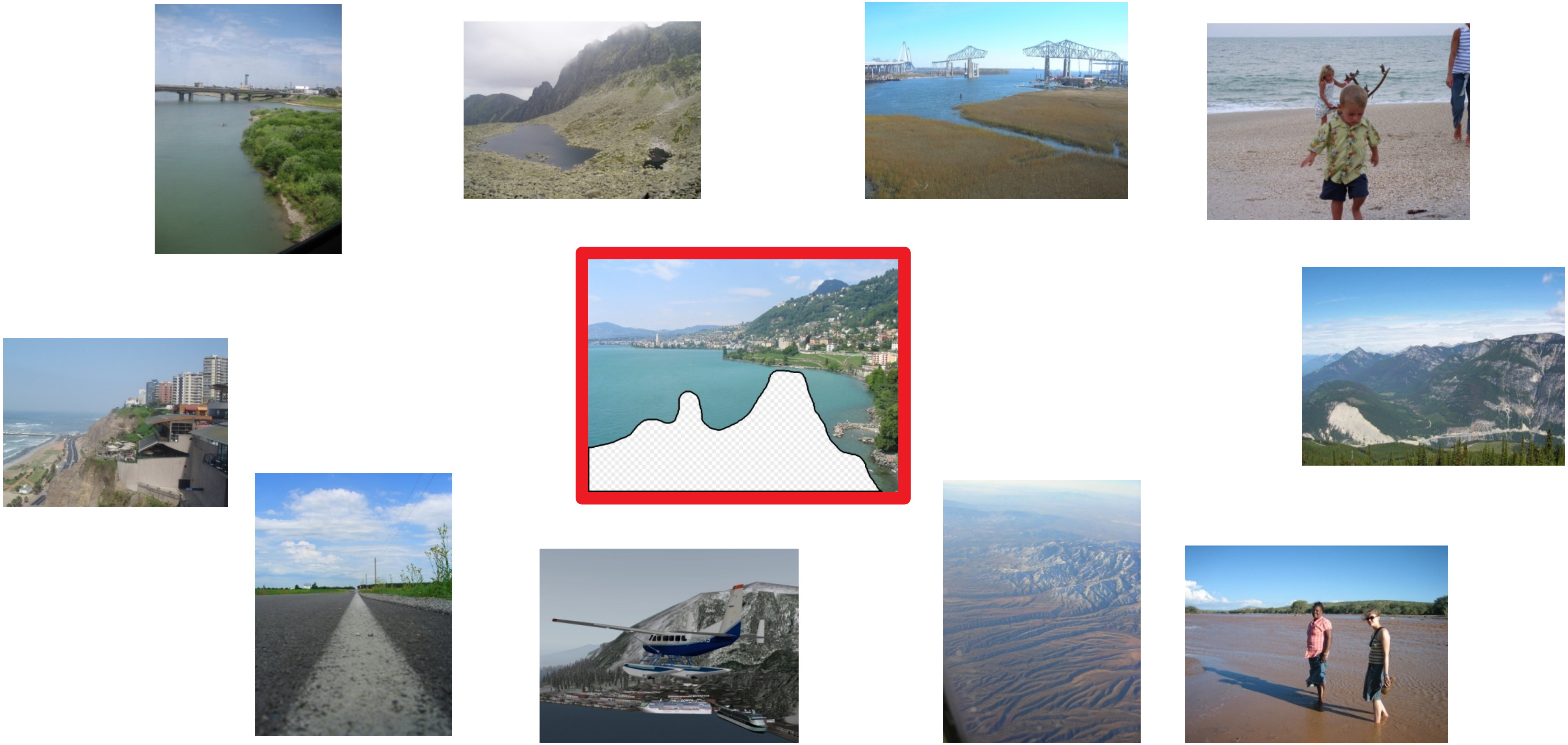
# Scene completion problem



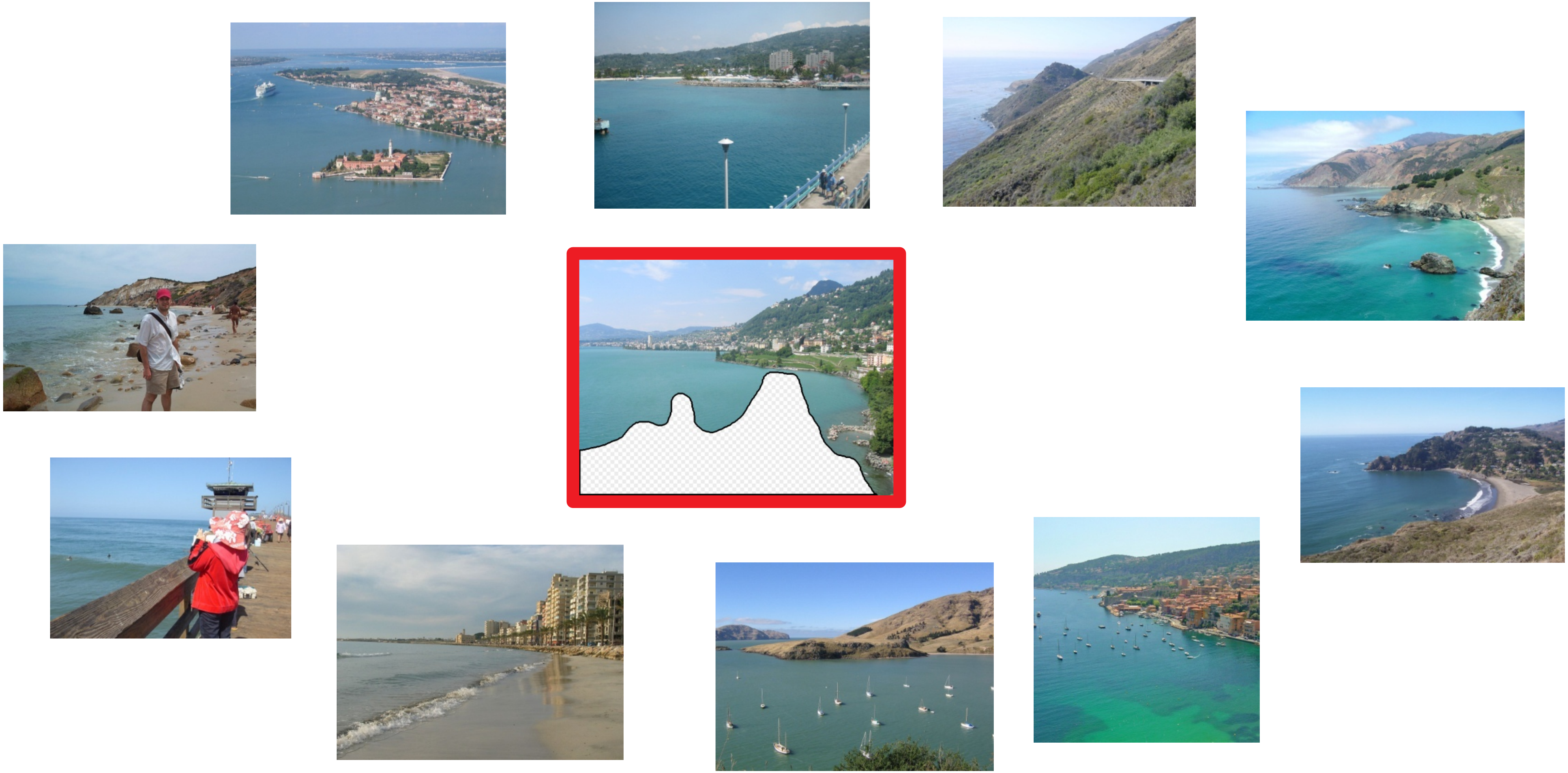


[Hays and Efros, SIGGRAPH 2007]

# 10 closest items in a collection of 20K images



# 10 closest items in a collection of 2M images



# Computing similarity

# Computing similarity is important

- Many problems can be expressed as finding “similar” sets:
  - Find near-neighbors in high-dimensional space
- Examples:
  - Pages with similar words
    - For duplicate detection or for classification by topic
  - Customers who purchased similar products
    - Products with similar customer sets
  - Images with similar features
  - Users who visited similar websites

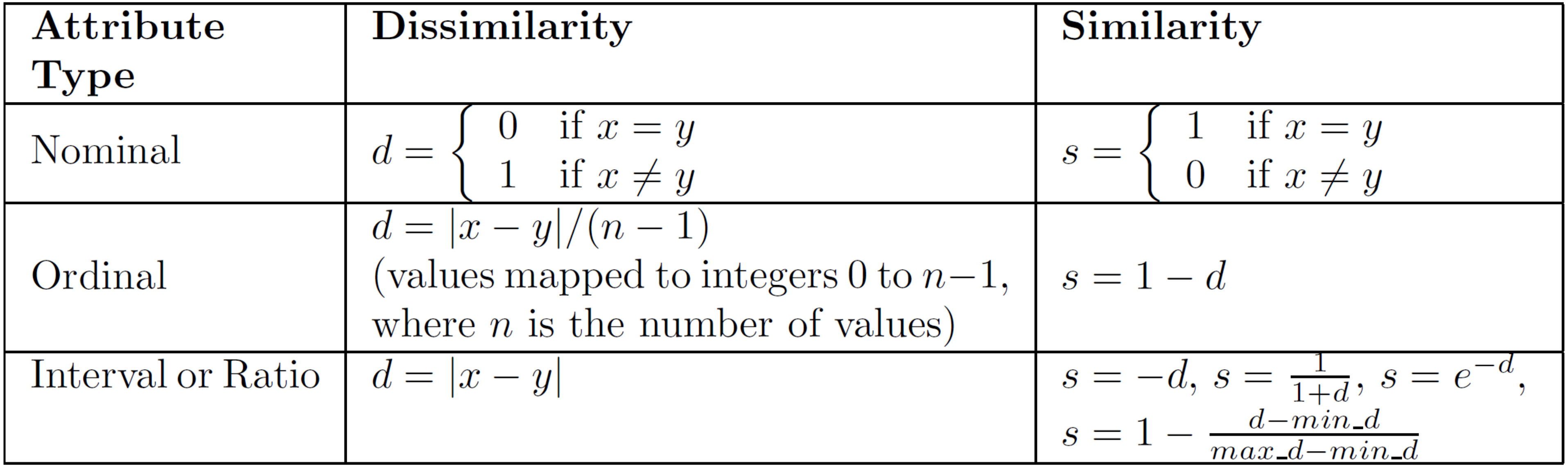
# Similarity computation task

- Given two objects u and v, determine the value of:

  similarity(u,v) and distance(u,v)

  (Often one is defined in terms of the other)
- Similar objects should have large similarity and small distance
- Dissimilar objects should have small similarity and large distance
- Closed-form functions (e.g., euclidean distance) or algorithm

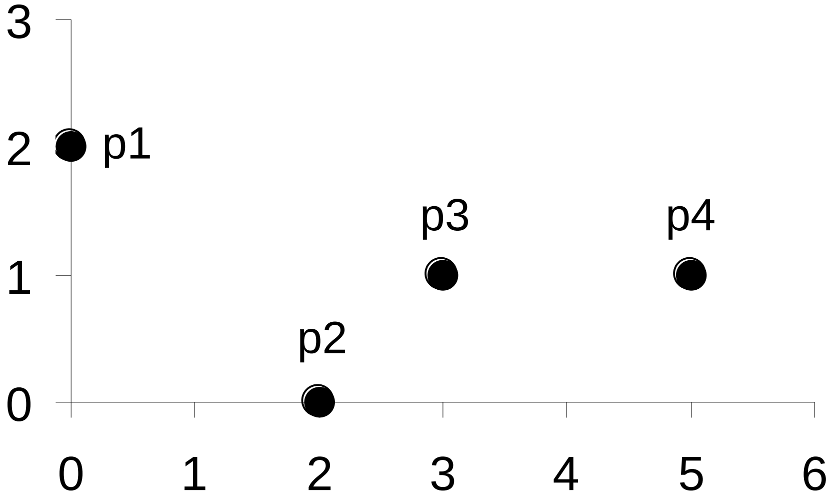
# Simple single-attribute similarity



| **Attribute Type** | **Dissimilarity** | **Similarity** |
|---|---|---|
| Nominal | $d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$ | $s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$ |
| Ordinal | $d = \lvert x - y \rvert / (n - 1)$<br>(values mapped to integers 0 to $n{-}1$, where $n$ is the number of values) | $s = 1 - d$ |
| Interval or Ratio | $d = \lvert x - y \rvert$ | $s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$,<br>$s = 1 - \frac{d - min\_d}{max\_d - min\_d}$ |

Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)

# Euclidean distance: $L_2$ norm



| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)

# THE CURSE OF DIMENSIONALITY

# $L_p$ norm, p ≥ 1

$$\mathrm{dist}(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- p=1 : Manhattan norm
  - Sum of absolute values
- p=2: Euclidean norm
  - Square root of sum of squares
  - Rotation-invariant
- p=∞ : Infinity norm
  - Largest absolute value

# Exercise: compute Lp distance

- Given vectors

  $u = (22, 1, 42, 10)$

  $v = (20, 0, 36, 8)$

- Compute:

  $L_1$ distance

  $L_2$ distance

  $L_\infty$ distance

Answer in
Nearpod Collaborate
Code to be given in class

# Generalized $L_p$ norm, $p \geq 1$

- Useful when some features are more important than others

$$\mathrm{dist}(x, y) = \left( \sum_{i=1}^{d} a_i \left| x_i - y_i \right|^p \right)^{\frac{1}{p}}$$

- E.g., in credit scoring,
  salary is more important than gender

- $a_i$ are domain-specific non-negative coefficients

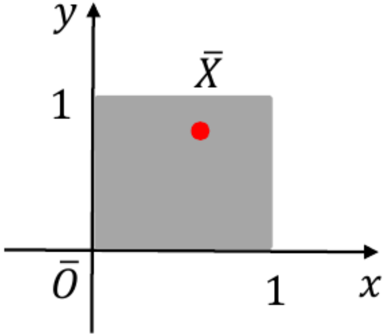
# The Curse of Dimensionality

- When the dimensionality is high, all points are at similar $L_p$ distances from each other

- Example: A unit cube of dimensionality $d$ in the nonnegative quadrant

  $\bar{X}$ is a random point in the cube

  Manhattan distance between $\bar{O}$ and $\bar{X}$



# THE CURSE OF DIMENSIONALITY

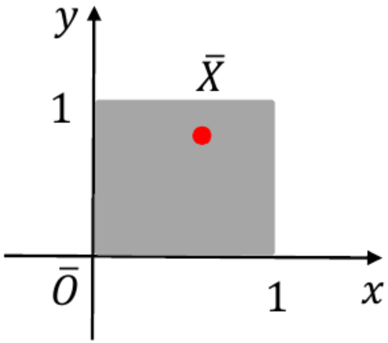
- Example (cont.):

Manhattan distance between $\bar{O}$ and $\bar{X}$

$$Dist(\overline{O},\overline{X}) = \sum_{i=1}^{d}(Y_i - 0).$$

where $\bar{X} = [Y_1, \dots, Y_d]$

$Dist(\bar{O},\bar{X})$ is a random variable



✓ Since $\bar{X}$ is a random variable

✓ Mean is $\mu = d/2$

✓ Standard deviation $\sigma = \sqrt{d/12}$

# THE CURSE OF DIMENSIONALITY

Applying Chebyshev's inequality:

$$Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$Pr(|\operatorname{Dist}(\overline{O}, \overline{X}) - \mu| \geq 3\sigma) \leq \frac{1}{3^2}$$

$$Pr(\operatorname{Dist}(\overline{O}, \overline{X}) \in [\mu - 3\sigma, \mu + 3\sigma]) > 8/9$$

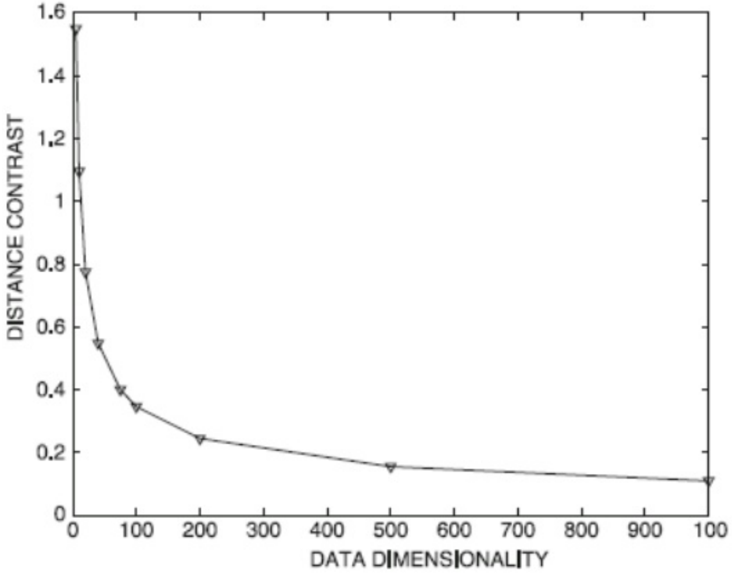
# THE CURSE OF DIMENSIONALITY

Applying Chebyshev's inequality:

$$Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$Pr(|\,\mathrm{Dist}(\overline{O},\overline{X})-\mu| \geq 3\sigma) \leq \frac{1}{3^2}$$

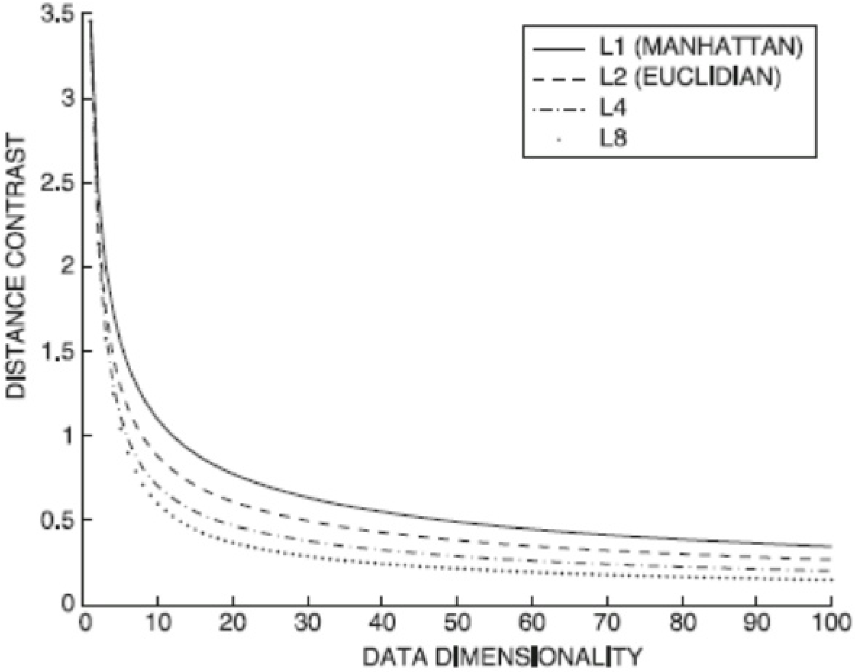
$$Pr(\mathrm{Dist}(\overline{O},\overline{X}) \in [\mu-3\sigma,\mu+3\sigma]) > 8/9$$



$$\mathrm{Contrast(d)} = \frac{D_{\max}-D_{\min}}{\mu} = \sqrt{12/d}$$

# Irrelevant features

- Many features are probably irrelevant for your purposes, specially in high-dimensional data
- $L_p$ norm suffers from irrelevant features
- Contrast worsens for large p



# Match-based similarity

Idea: to compute similarity(u,v) ignore dimensions in which they are “too far apart”

1) Discretize each dimension into $k_d$ equi-depth buckets
2) For two objects u, v, determine the dimensions in which they map to the same bucket
3) Compute $L_p$ norm on those dimensions only

# Match-based similarity (cont.)

$$PSelect(\overline{X},\overline{Y},k_d) = \left[ \sum_{i \in S(\overline{X},\overline{Y},k_d)} \left( 1 - \frac{|x_i - y_i|}{m_i - n_i} \right)^p \right]^{1/p}$$

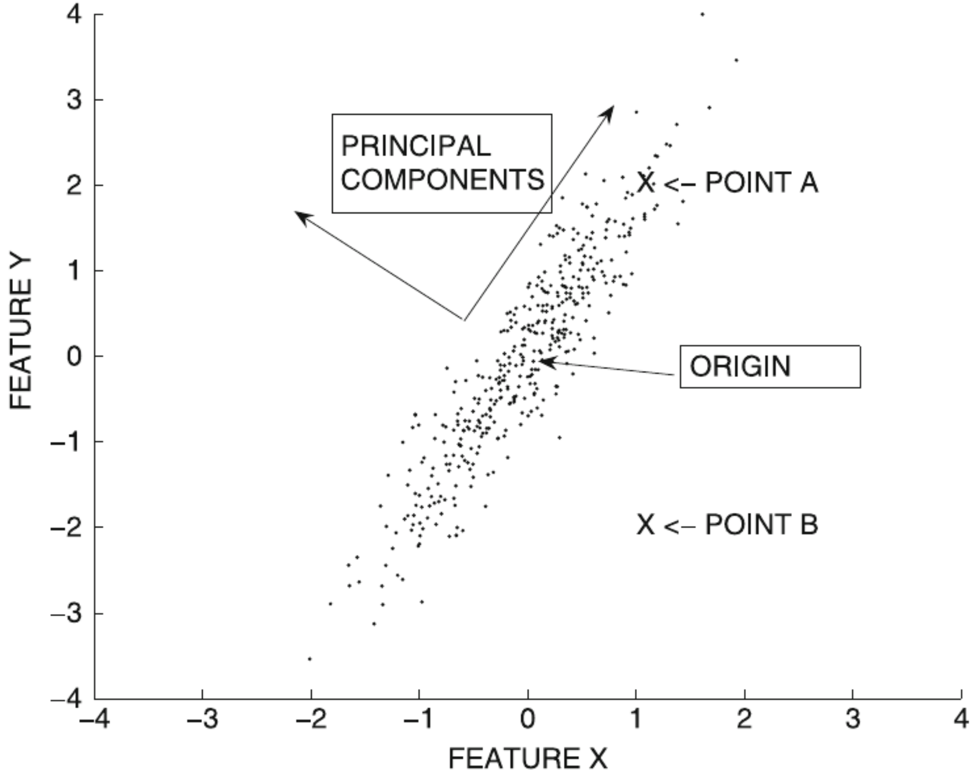
- $S(\overline{X}, \overline{Y}, k_d)$ is the set of features for which $\overline{X}$ and $\overline{Y}$ map to the same bucket
- $m_i$, $n_i$ are the max and min value of that bucket
- $k_d \propto d$ achieves a constant level of contrast in high dimensions for certain data distributions

# Distances and orientation

# Useful distances, in general, depend on data distributions

Points A and B are equidistant from the origin

However, point A should be considered closer to the origin than point B (think of a perfectly circular cloud of points)

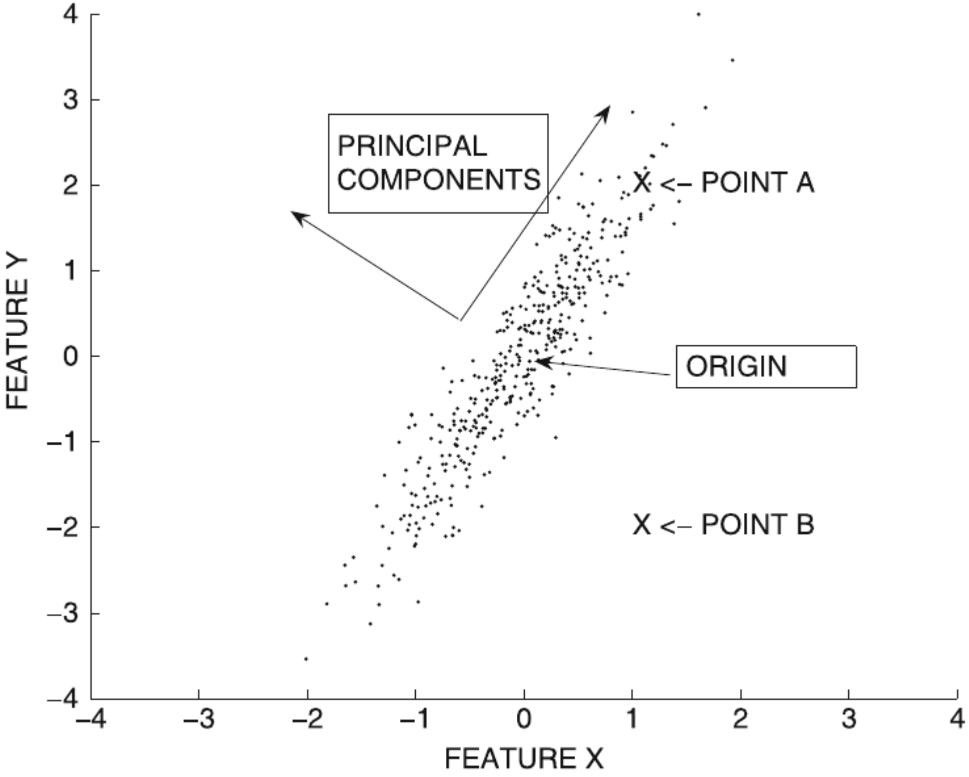


# Useful distances, in general, depend on data distributions (cont.)

The Mahalanobis distance, with Σ covariance matrix

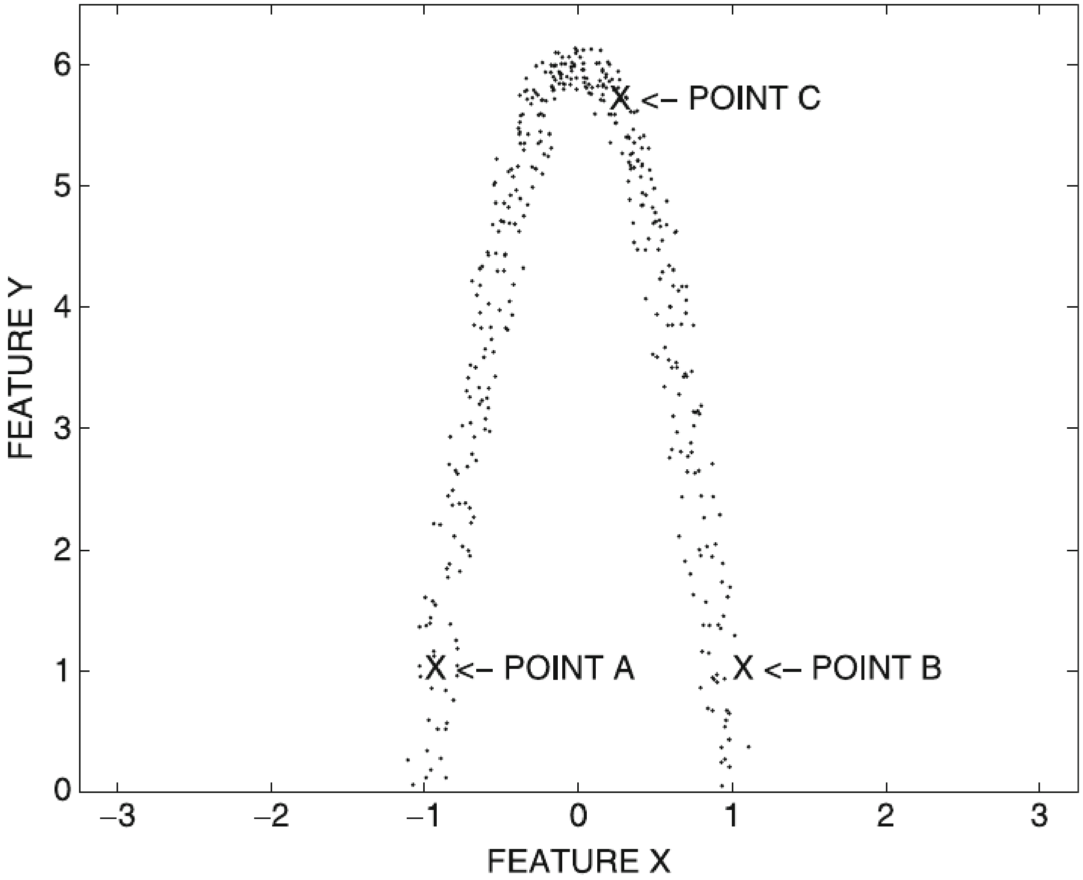
$$Maha(\overline{X}, \overline{Y}) = \sqrt{(\overline{X} - \overline{Y})\Sigma^{-1}(\overline{X} - \overline{Y})^T}.$$

is equivalent to applying PCA, dividing each coordinate by the standard deviation of that feature, and computing Euclidean distance



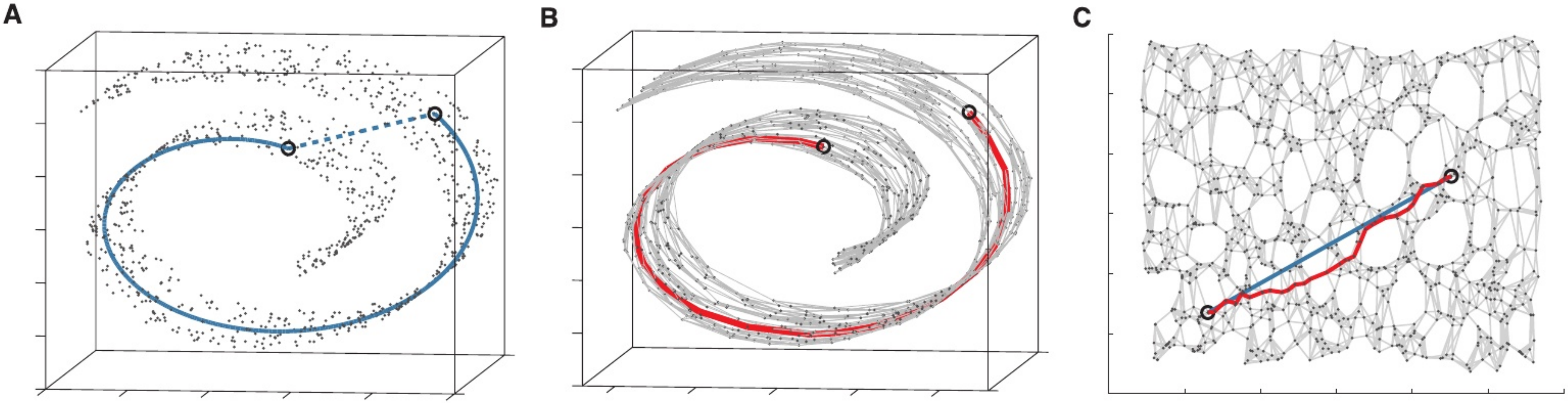
# Non-linear distributions

Which point would you consider as closer to A?



(Blackboard collaborate poll)

# ISOMAP (general idea)



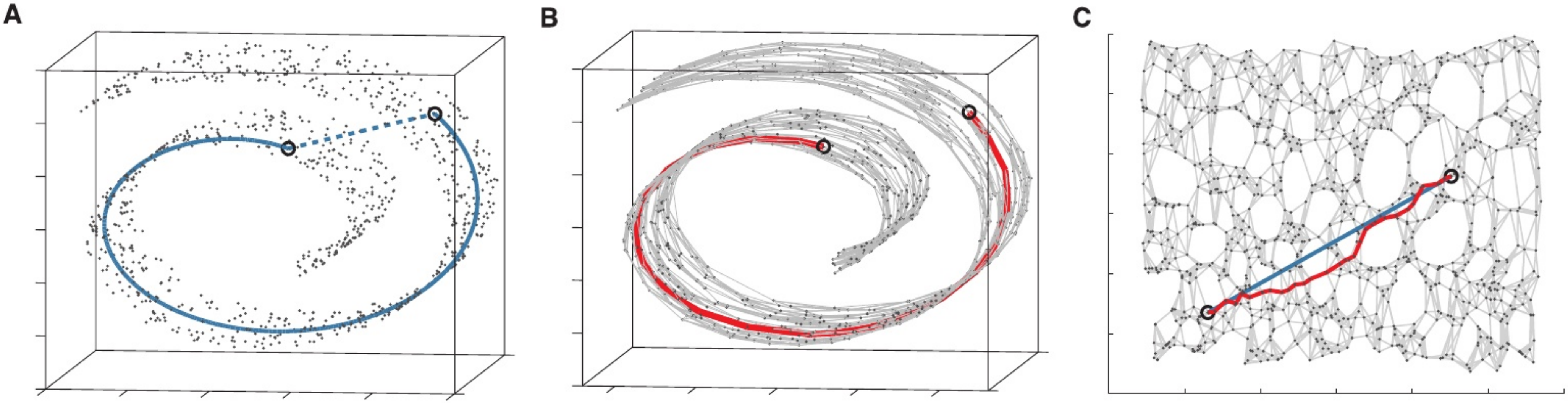
Original data

Nearest neighbors graph

Graph projection

Tenenbaum, Joshua B., Vin De Silva, and John C. Langford. "A global geometric framework for nonlinear dimensionality reduction." Science 2000.

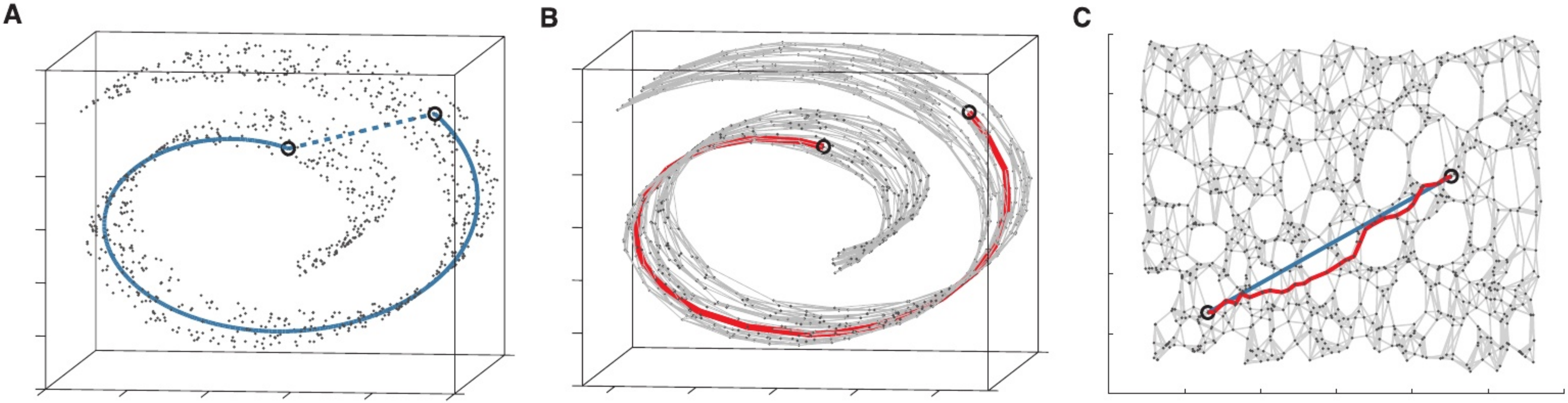
# ISOMAP (1/3)



The first step is to connect each point to its k nearest neighbors (here k=7)

Tenenbaum, Joshua B., Vin De Silva, and John C. Langford. "A global geometric framework for nonlinear dimensionality reduction." Science 2000.

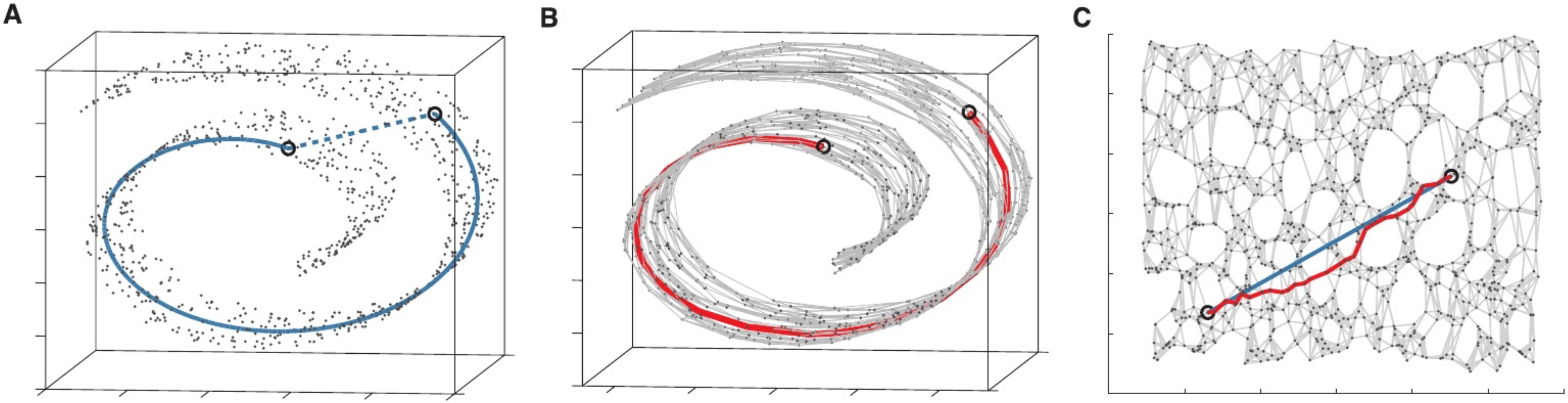
# ISOMAP (2/3)



Now, shortest path or *geodesic* distances can be computed on the graph (red color)

Tenenbaum, Joshua B., Vin De Silva, and John C. Langford. "A global geometric framework for nonlinear dimensionality reduction." Science 2000.

# ISOMAP (3/3)



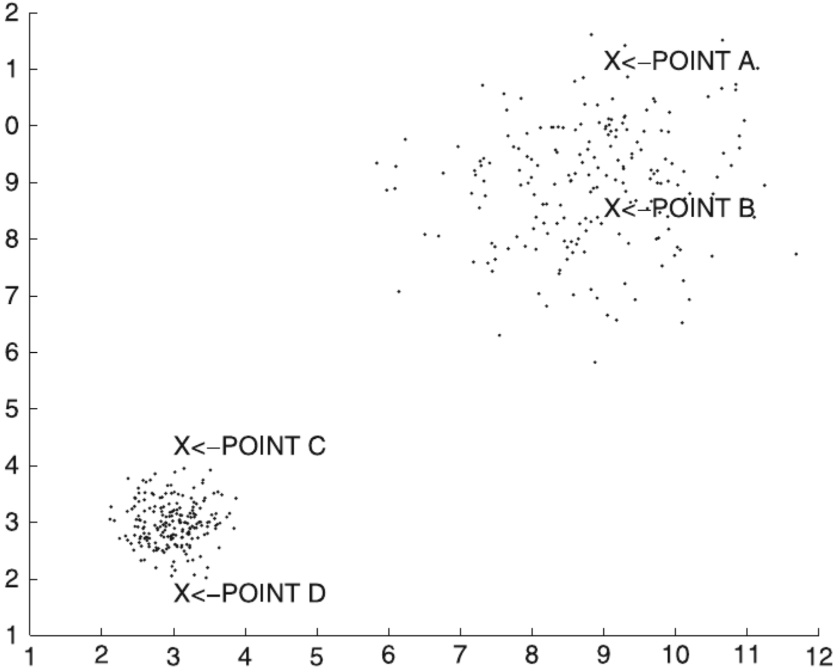
It is, however, more effective to project the graph and compute Euclidean distances in the projected graph (blue color)

Tenenbaum, Joshua B., Vin De Silva, and John C. Langford. "A global geometric framework for nonlinear dimensionality reduction." Science 2000.

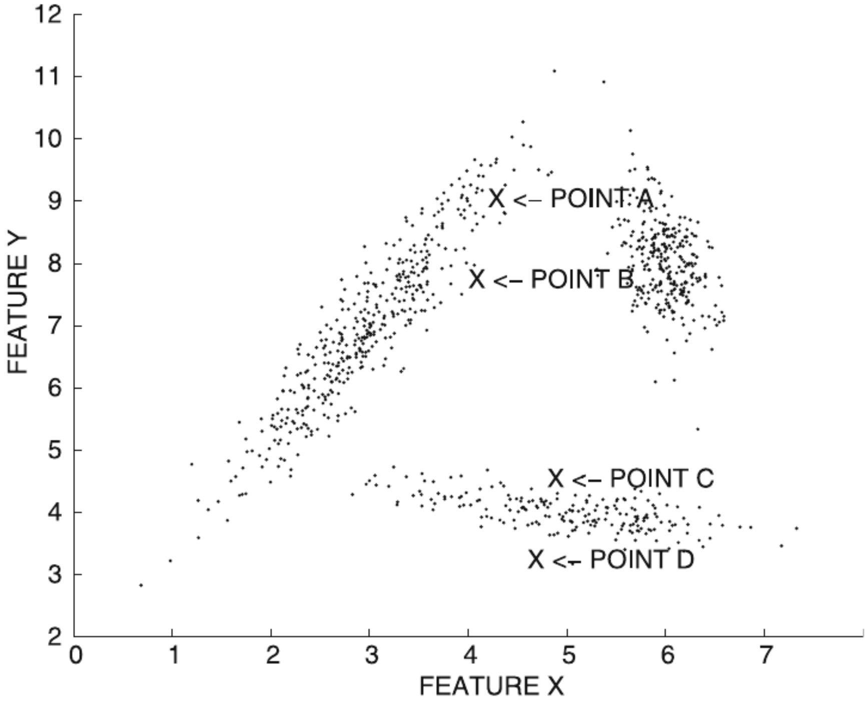
# Local variations

Which distance should be larger? A-B or C-D?

Which distance should be larger? A-B or C-D?



(a) local density variation



(b) local orientation variation

(Blackboard collaborate poll)

# Solution for local variations

- Partition the data into a set of local regions
  - (Nontrivial, which distance to use?)
- For any pair of objects, determine the most relevant region for the pair
- If they belong to the same region
  - Compute the pairwise distances using the local statistics of that region
  - E.g., local Mahalanobis distance
- If they belong to different regions
  - Global statistics or averaged statistics

# Summary

# Things to remember

- Distance/similarity is a key component of many data mining algorithms
- Sensitive to dimensionality, global/local nature of data distribution

# Exercises for TT06-TT07

- **Data Mining, The Textbook (2015) by Charu Aggarwal**
  - **Exercises 3.9 on similarity measures**
- Introduction to Data Mining 2nd edition (2019) by Tan et al.
  - Exercises 2.6 → 14-28
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al.
  - Exercises 3.5.7 on distance measures
- Data Mining Concepts and Techniques, 3rd ed. (2011) by Han et al.
  - Exercises 2.6 → 2.5-2.8