

Scientific and Large Data Visualization
8-9 November 2018
High Dimensional Data

Massimiliano Corsini

Visual Computing Lab, ISTI - CNR - Italy

Overview

- **Graphs Extensions**
- **Glyphs**
 - Chernoff Faces
 - Multi-dimensional Icons
- **Parallel Coordinates**
- **Star Plots**
- **Dimensionality Reduction**
 - Principal Component Analysis (PCA)
 - Locally Linear Embedding (LLE)
 - IsoMap
 - t-SNE

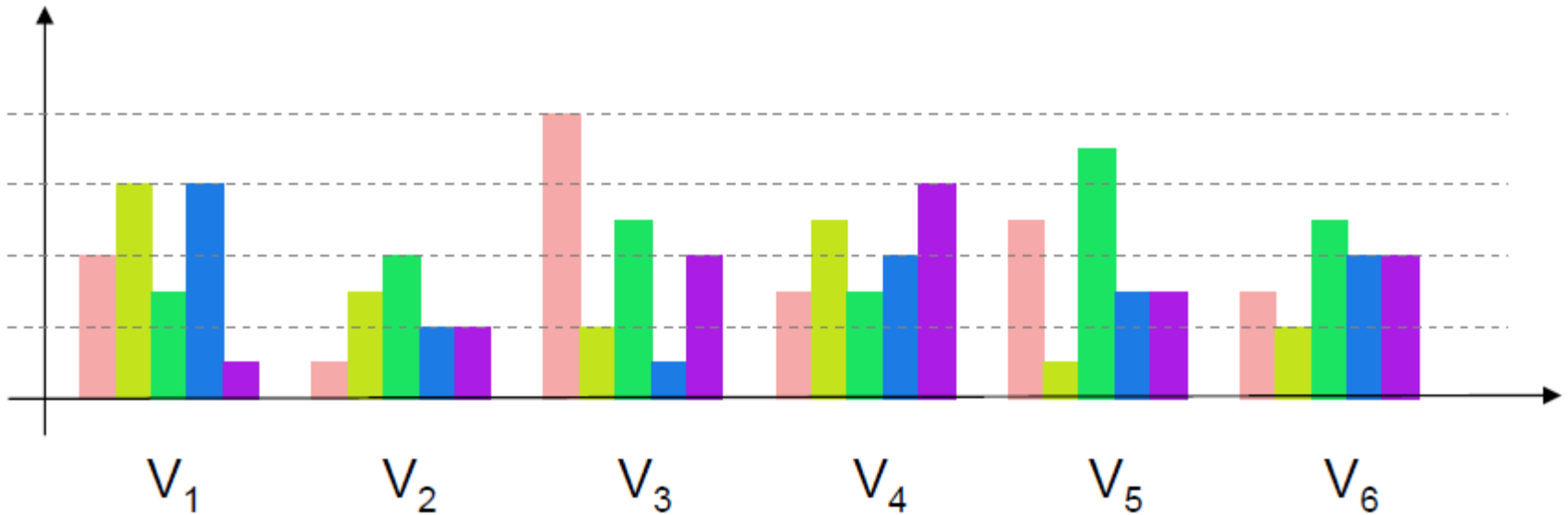
Multi-set Bar Charts

- Also called *grouped bar charts*.
- Like bar chart for more datasets.
- Allow accurate numerical comparisons.
- It can be used to show mini-histograms.



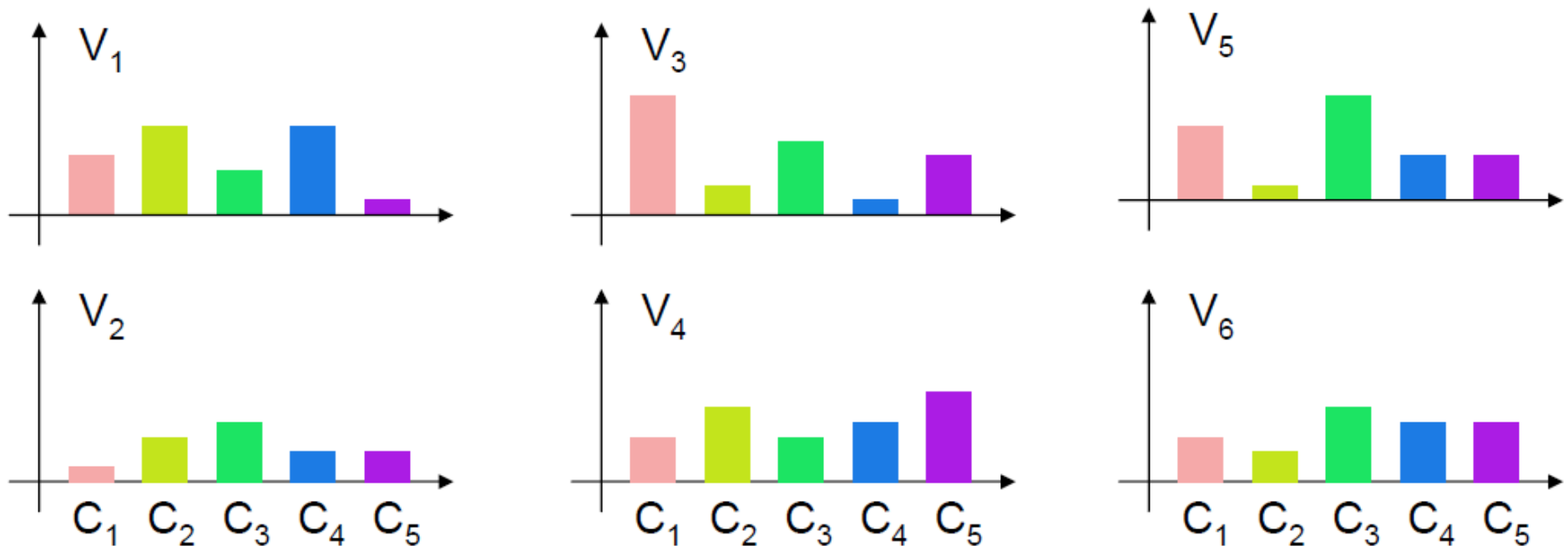
Multiple Bars

- Distributions of each variable among the different categories/data points.



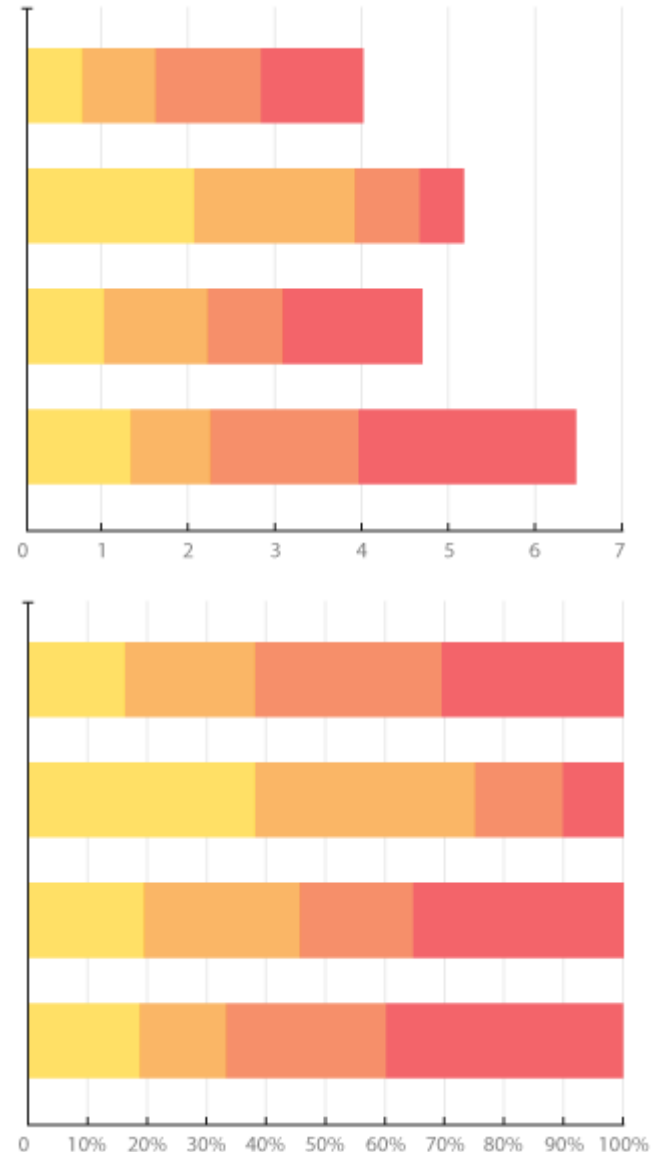
Multiple Views

- Distributions of each variable among the different categories/data points (*each variable has its own display*).



Stacked Bar Charts

- Each dataset is drawn on top of each other.
- Two types:
 - *Simple Stacked Bar Charts*
 - *Percentage Stacked Bar Charts*.
- More segments each bar more difficult to read.



Simple vs Percentage Bar Charts

- Simple Stacked Bar Charts
 - Useful if the visualization of the absolute values (and their sum) is meaningful.
- Percentage Bar Charts
 - Better to show the relative differences between quantities in the different groups.

Spineplots

- Generalization of stacked bar charts.
- Special case of *mosaic plot*.
- Permit to show both percentages and proportions between variables.

Spineplots – Example

- Car data:

Nationality	Repair records	# models
domestic	1	2
domestic	2	8
domestic	3	27
domestic	4	9
domestic	5	2
foreign	1	0
foreign	2	0
foreign	3	3
foreign	4	9
foreign	5	9

Spineplots – Example

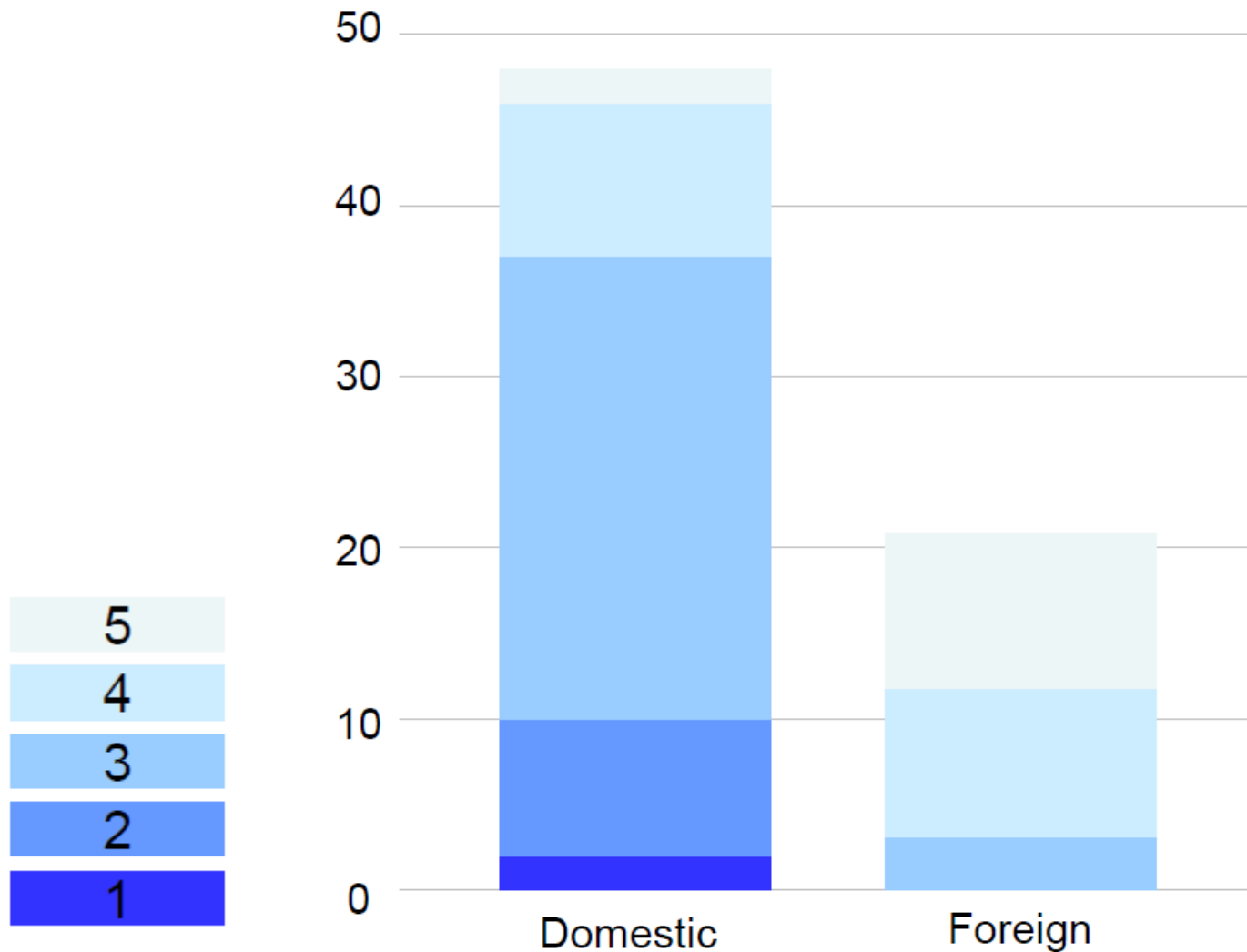


Figure from *Information Visualization Course*, Università Roma 3.

Spineplots – Example

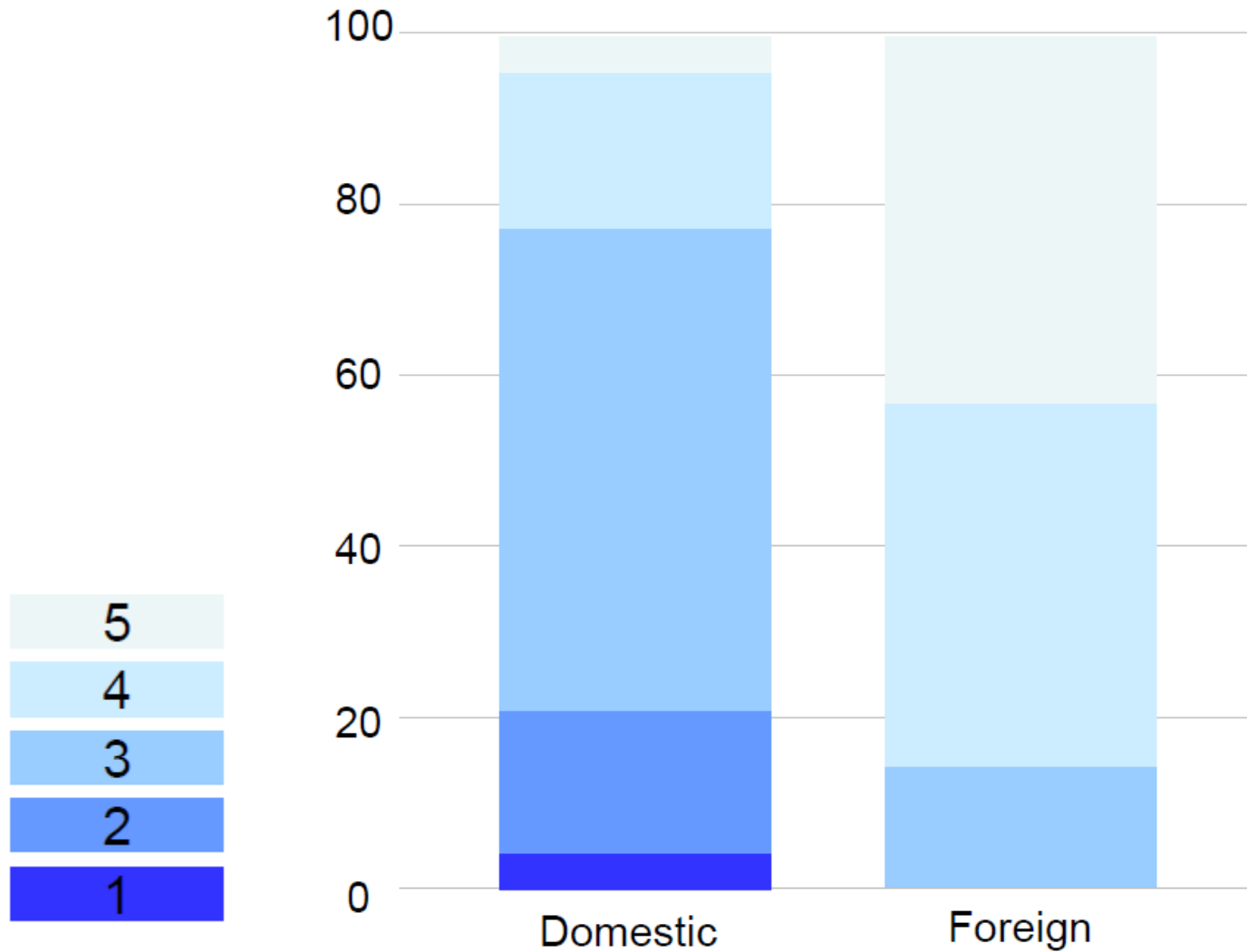
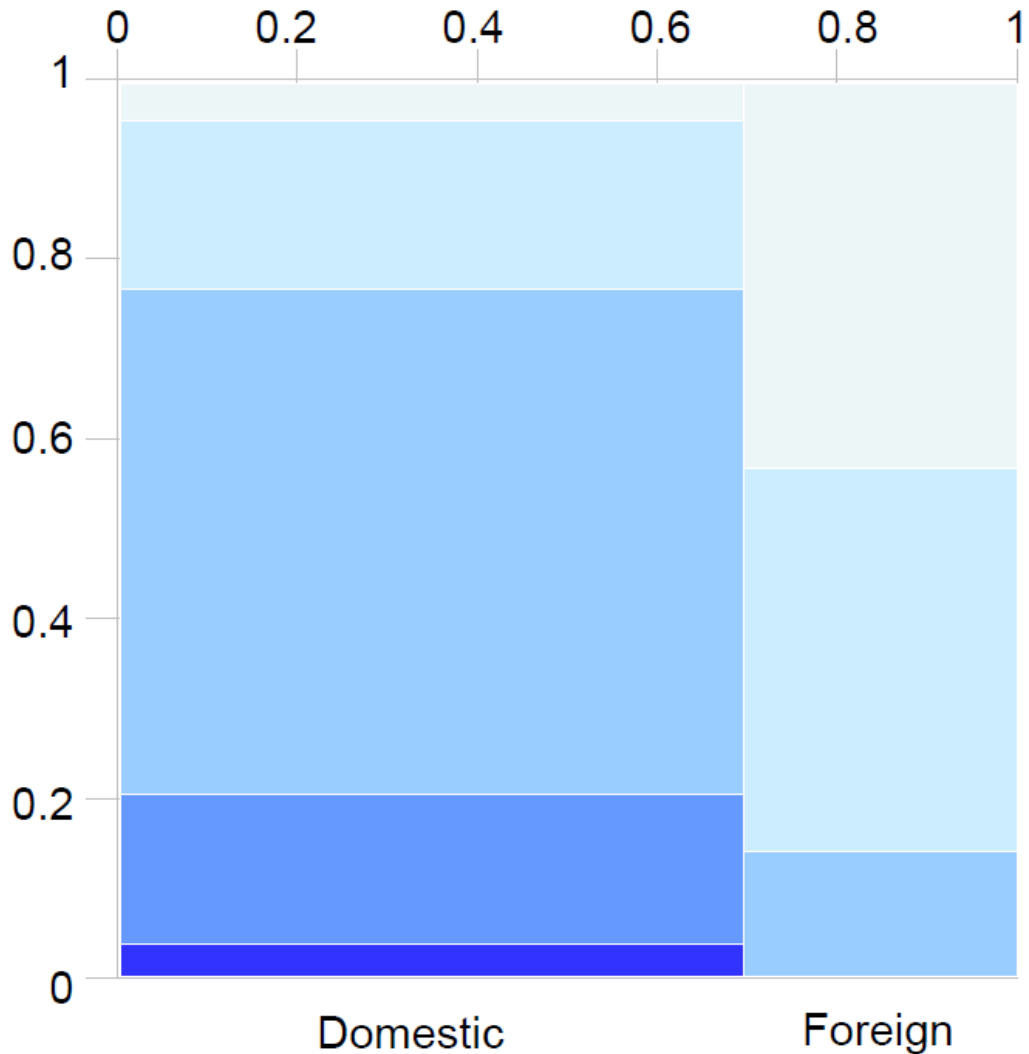
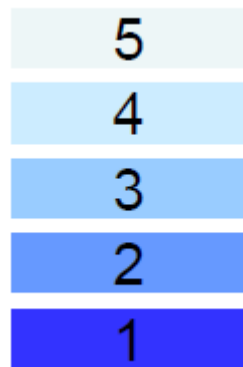


Figure from *Information Visualization Course*, Università Roma 3.

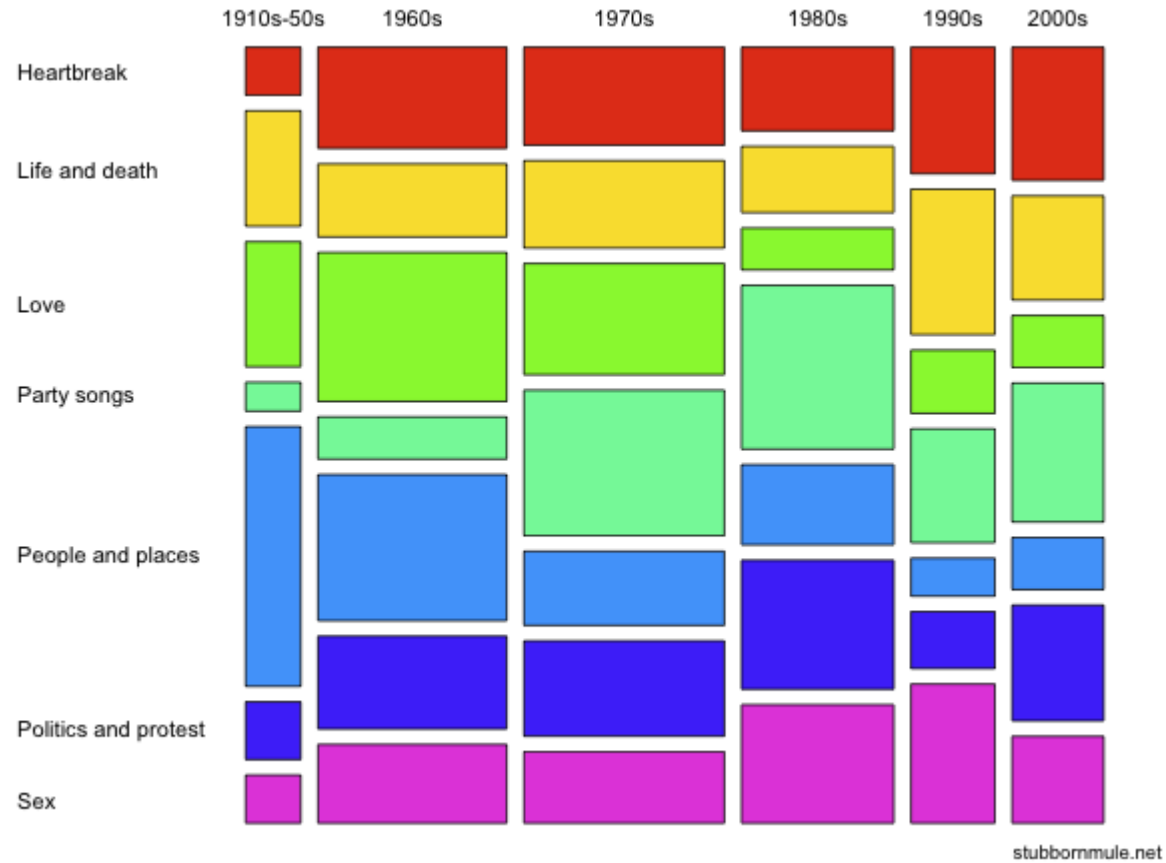
Spineplots – Example

- Convey both percentages and proportions.



Mosaic Plots

- They give an overview of the data by visualizing the relative proportions.
- Also known as Mekko charts.



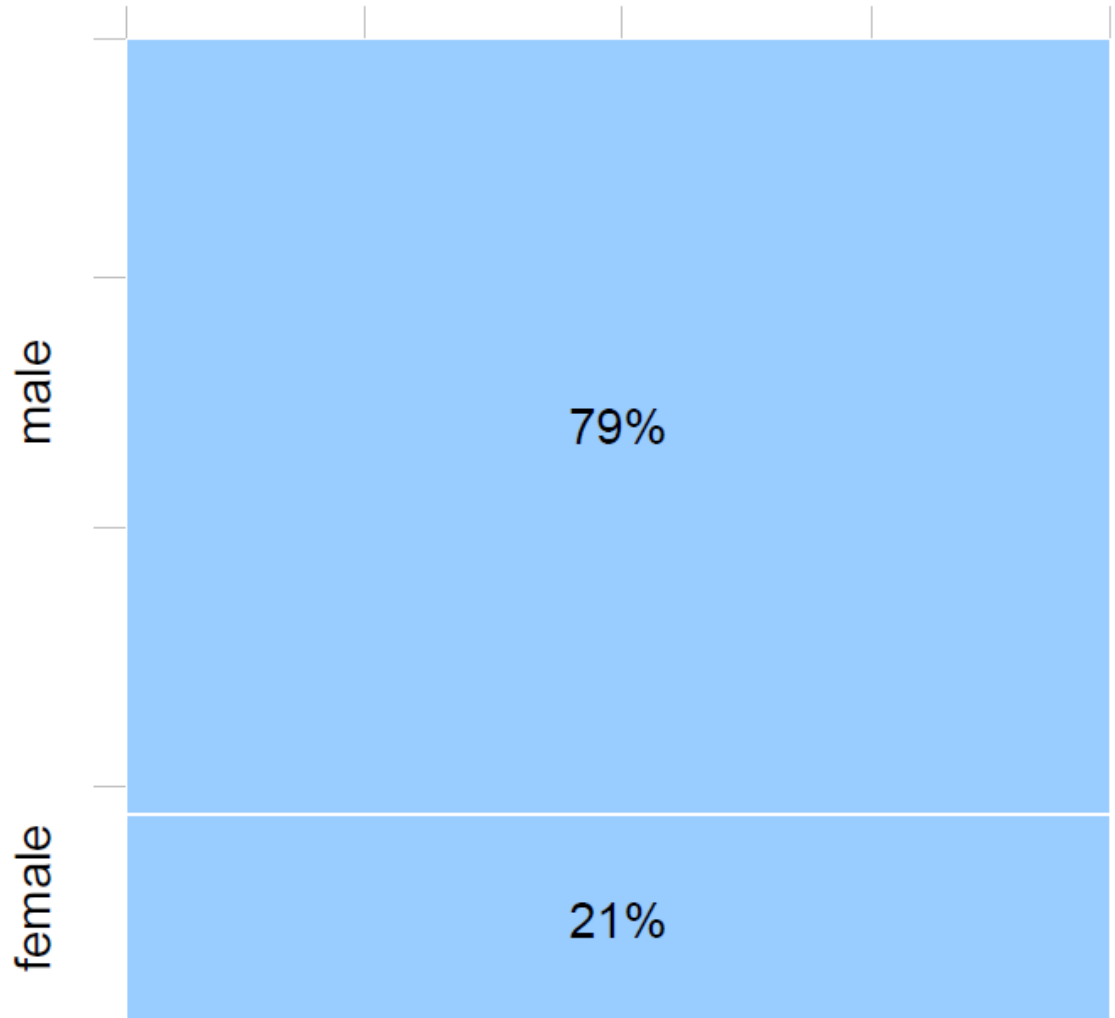
Mosaic Plots

- Titanic data (from Wikipedia):

Gender	Survived	1° class	2° class	3° class	Crew
Male	No	118	154	422	670
Male	Yes	62	25	88	192
Female	No	4	13	106	3
Female	Yes	141	93	90	20

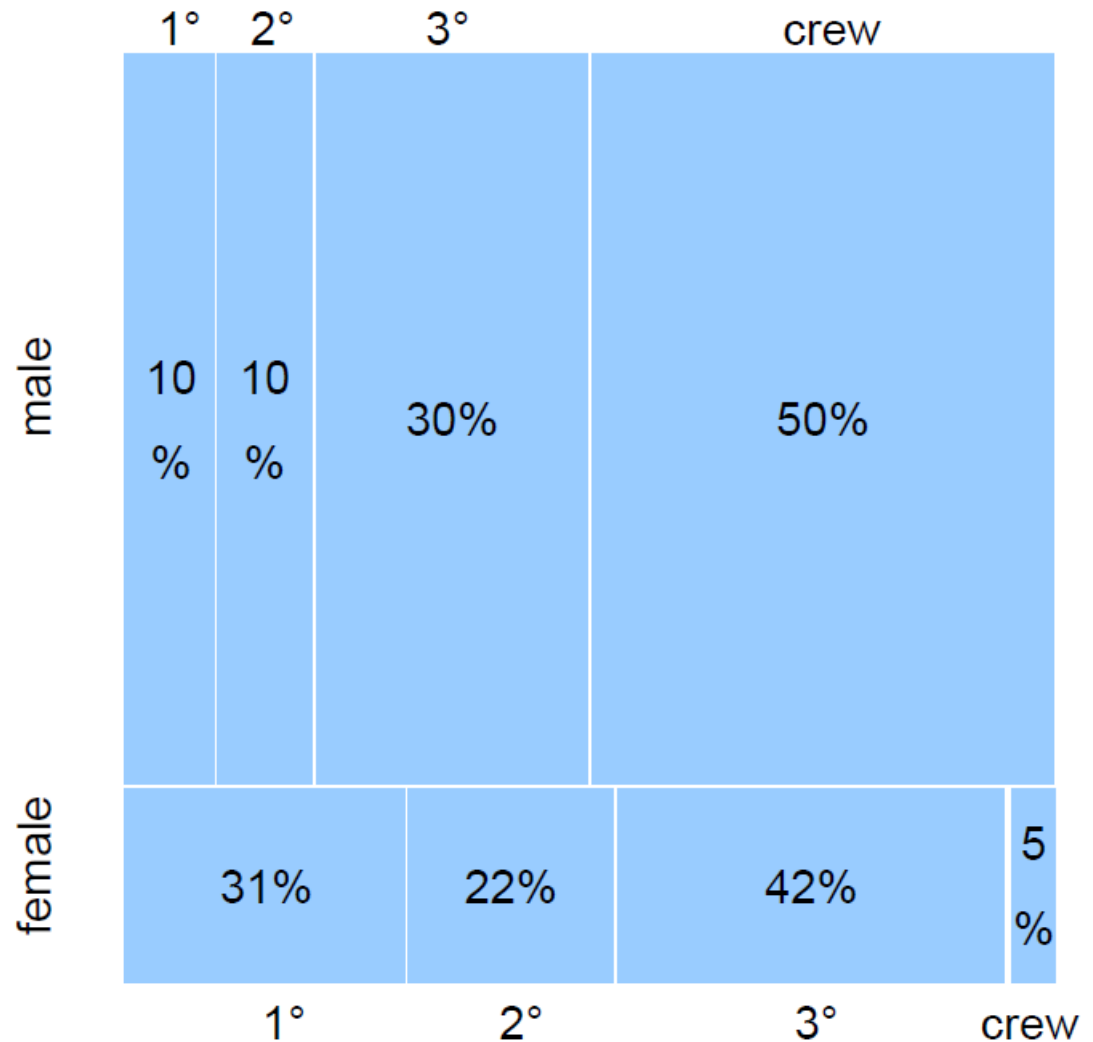
Mosaic Plots

- Variable *gender* → vertical axis.



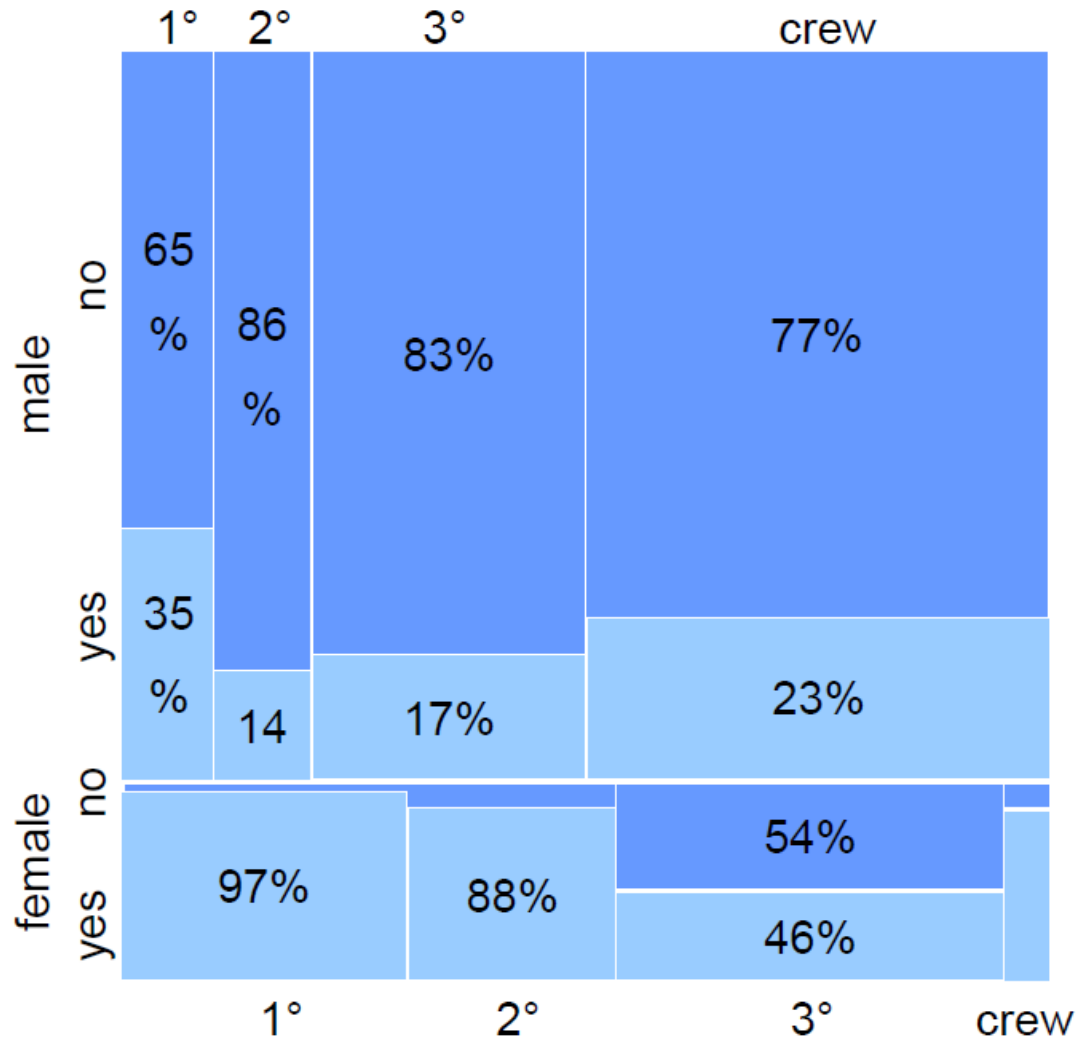
Mosaic Plots

- Add variable *Class* → Horizontal axis.



Mosaic Plots

- Add variable *survived* → vertical axis.



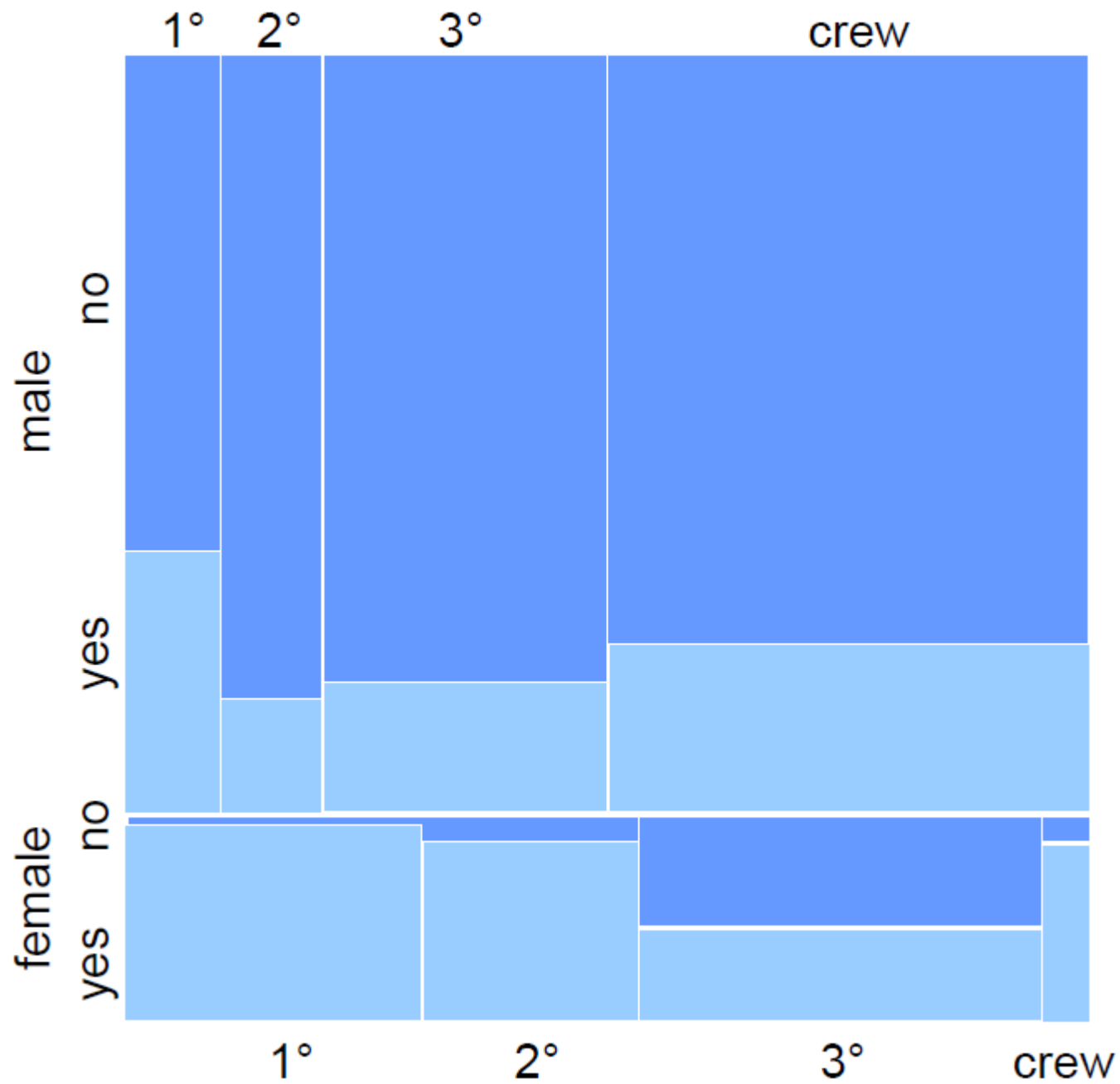


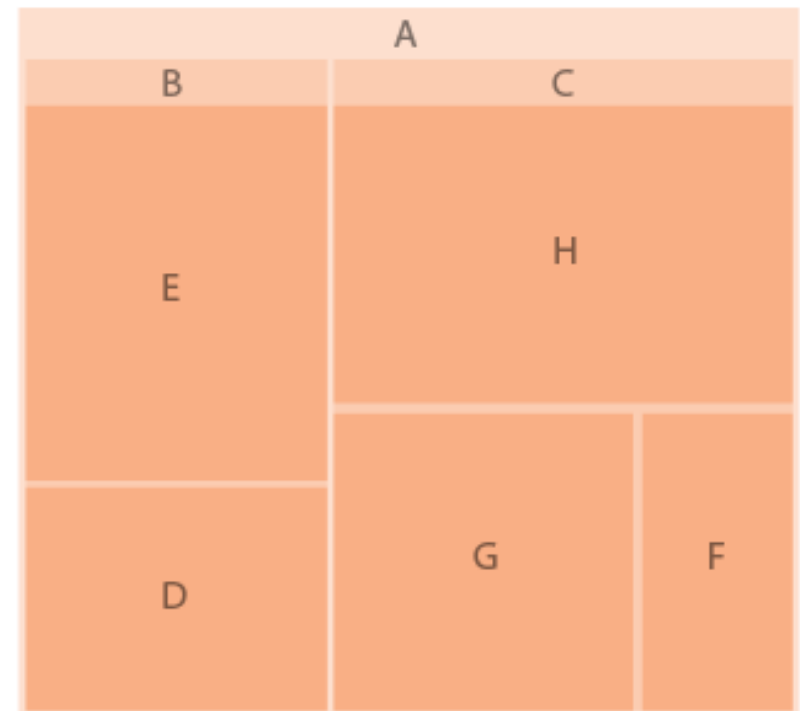
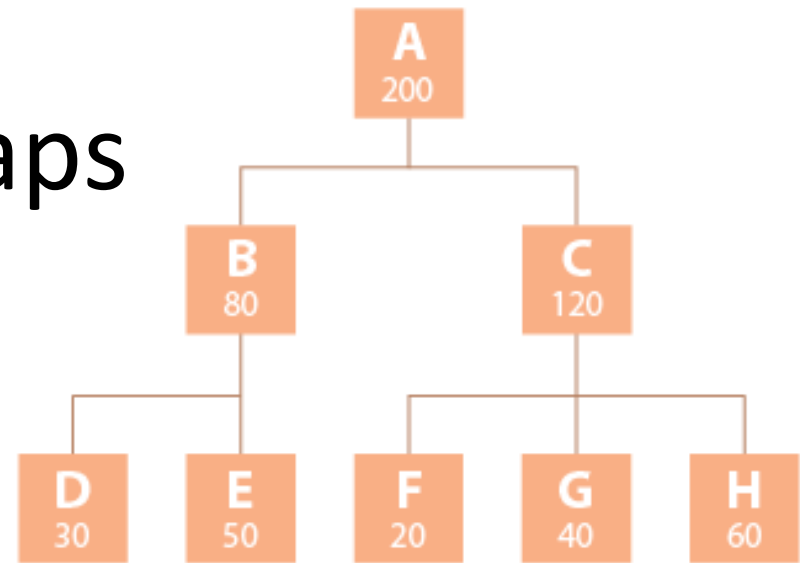
Figure from *Information Visualization Course*, Università Roma 3.

Mosaic Plots

- Advantages:
 - Maximal use of the available space
 - Good overview of the proportions between data
 - Good overview of the variable dependency
- Disadvantages:
 - Extension to many variables is difficult

Tree Maps

- An alternative way to visualize a tree data structure.
- Space-efficient (!)
- The way rectangles are divided and ordered into sub-rectangles is dependent on the tiling algorithm used.

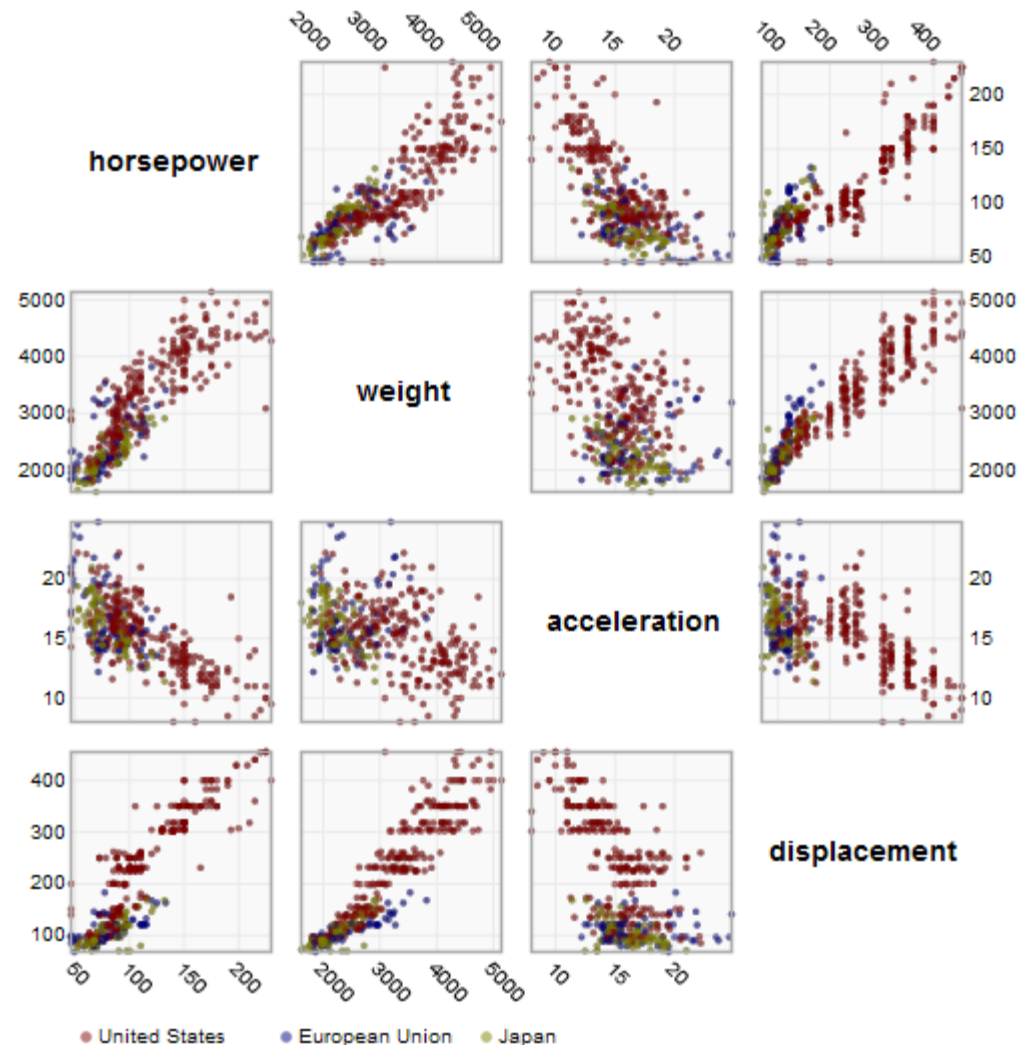


Tree Maps – Example

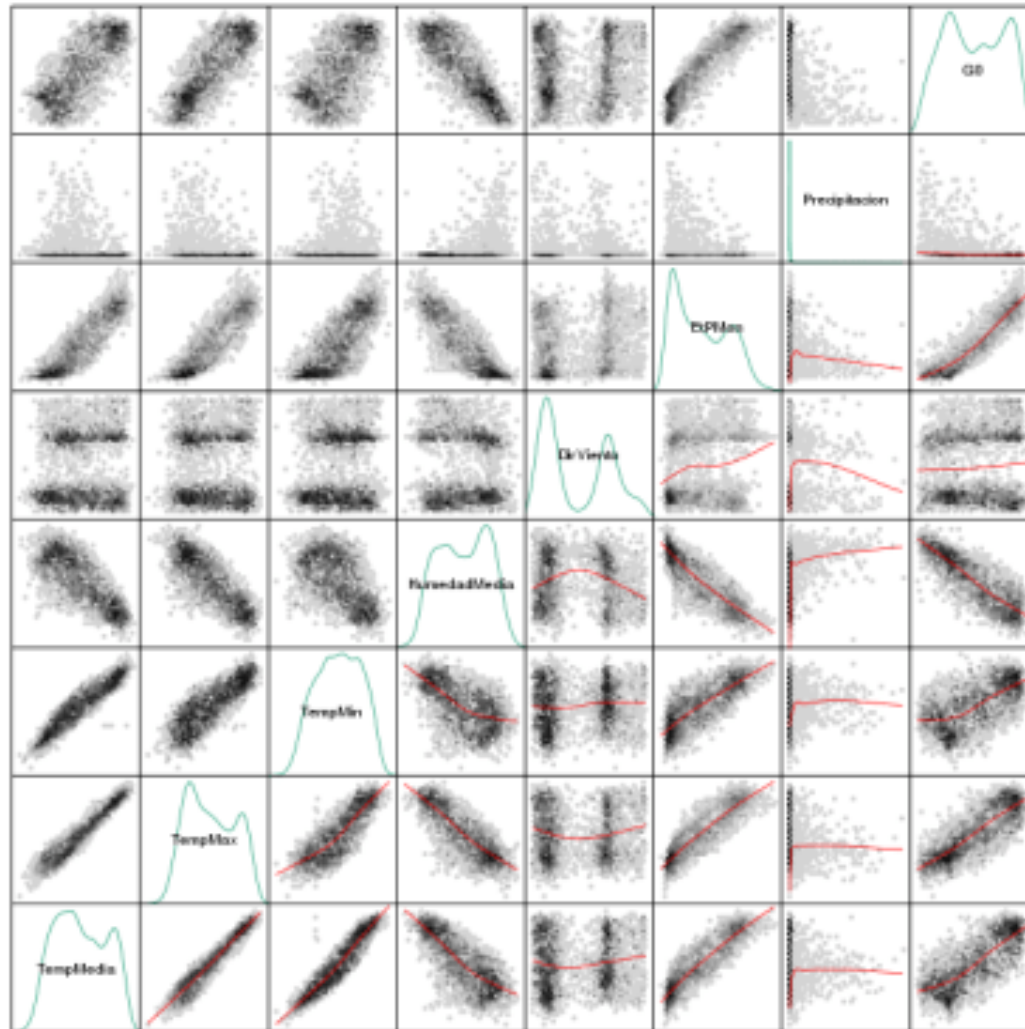


Scatter Plot Matrix (SPLOM)

- Each possible pair of variables is represented in a standard 2D scatter plot.
- Many works in literature to find efficient way to present them.



SPLOM

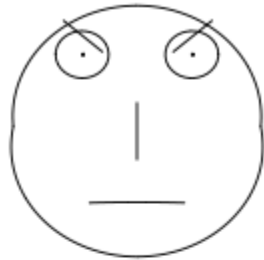


Scatter Plot Matrix

Chernoff Faces

- Introduced by Herman Chernoff in 1973.
- We are very good at recognize faces.
- Variables are mapped on facial features:
 - Width/curvature of mouth
 - Vertical size of the face
 - Size/slant/separation of eyes
 - Size of eyebrows
 - Vertical position of eyebrows

Chernoff Faces (lawyers judge)



AARONSON, L.H.



ALEXANDER, J.M.



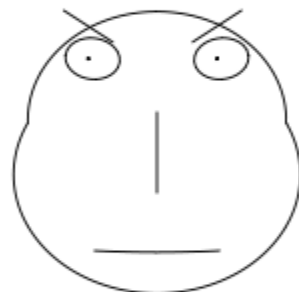
ARMENTANO, A.J.



BERDON, R.I.



BRACKEN, J.J.



BURNS, E.B.



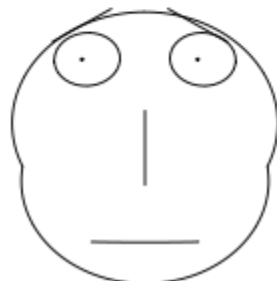
CALLAHAN, R.J.



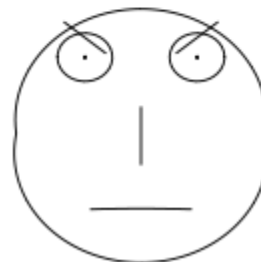
COHEN, S.S.



DALY, J.J.



DANNEHY, J.F.



DEAN, H.H.



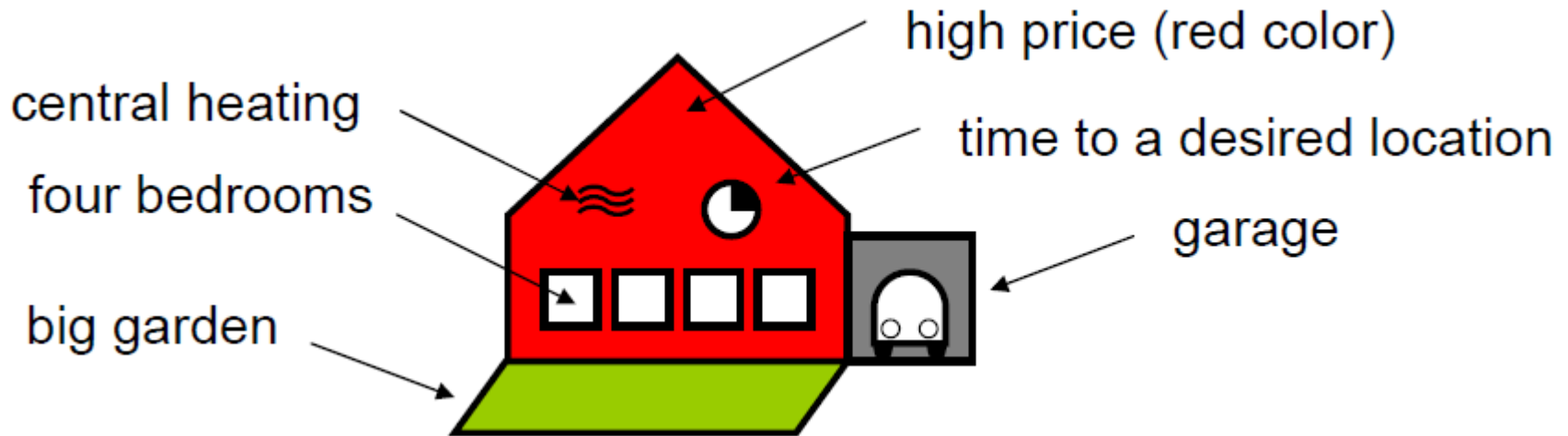
DEVITA, H.J.

Chernoff Faces

- Chernoff faces received some criticism
 - Difficult to compare.
 - A legend is necessary.
- Moreover, the mapping should be chosen carefully to work properly.

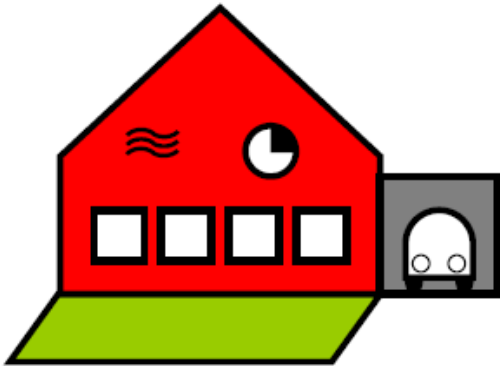
Multi-dimensional Icons

- Spence and Parr (1991) proposed to encode properties of an object in simple iconic representation and assemble them together.
- They applied this approach to check quickly dwell offers.

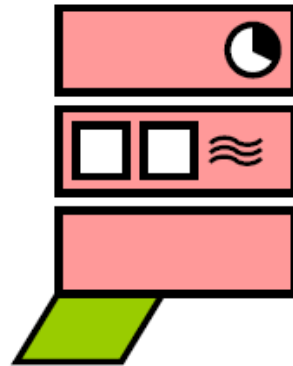


Robert Spence and Maureen Parr, *“Cognitive assessment of alternatives”*, *Interacting with Computers* 3, 1991.

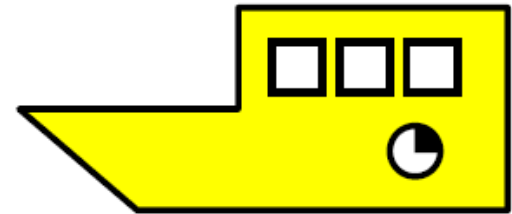
Spence and Parr (1991) (examples)



house
£400,000
garage
central heating
four bedrooms
good repair
large garden
Victoria 15 mins



flat
£300,000
no garage
central heating
two bedrooms
poor repair
small garden
Victoria 20 mins



houseboat
£200,000
no garage
no central heating
three bedrooms
good repair
no garden
Victoria 15 mins

Petals as a Glyph

- The idea of Moritz Stefaner to visualize a life index is to map several variables into *petals of different size*.
- Web site: www.oecdbetterlifeindex.org



Create Your Better Life Index

What is your recipe for a better life — a good education, clean air, nice home, money?
See how your country measures up on the topics important to you. [Help](#)

Start with all topics rated equally or set your own preferences [here](#).



How's life?

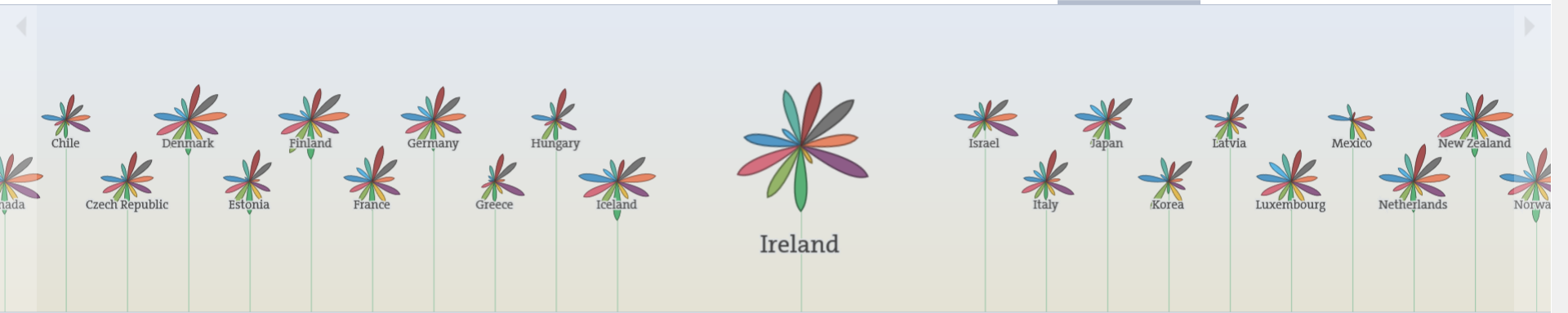
Create Your Better Life Index

Rate the topics according to their importance to you:

- Housing
- Income
- Jobs
- Community
- Education
- Environment
- Civic Engagement
- Health
- Life Satisfaction
- Safety
- Work-Life Balance

[Reset](#) [Help](#)

- [Gender differences](#)
- [Compare with others](#)
- [Share your index](#)



Ireland

→ Learn even more about Ireland at oecd.org

Did you know?

Population	4.6 mil.
Visitors per year	6.0 mil.
Renewable energy	6.1 %

How's Life?

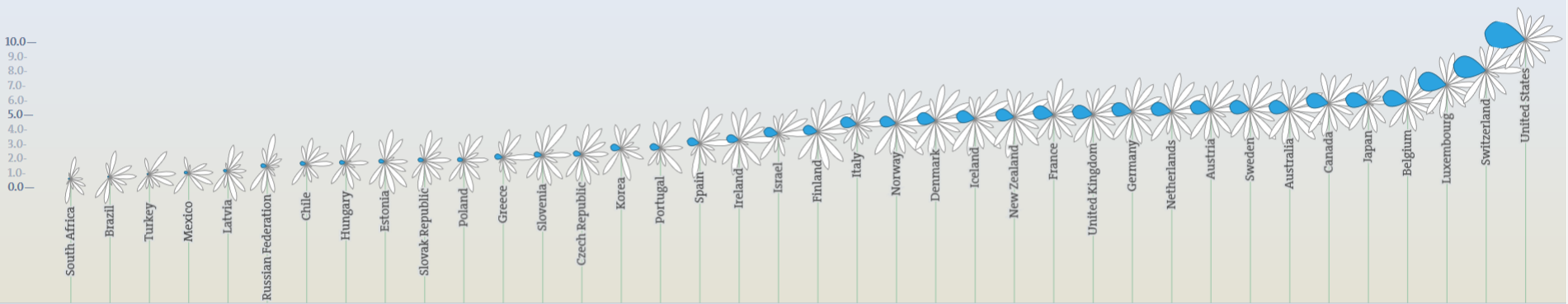
Ireland performs well in many measures of well-being relative to most other countries in the Better Life Index. Ireland ranks above the average in housing, personal security, health status, education and skills, social

Topics

Housing 6.9

OECD in Action





Income

Background

While money may not buy happiness, it is an important means to achieving higher living standards and thus greater well-being. Higher economic wealth may also improve access to quality education, health care and housing.

Household net adjusted disposable income

Financial wealth makes up an important part of a household's economic resources, and can protect from economic hardship and vulnerability. For example, a low-income household having above-average wealth will be better off than a low-income household with no wealth at all. Across the OECD, the

Top Ranking



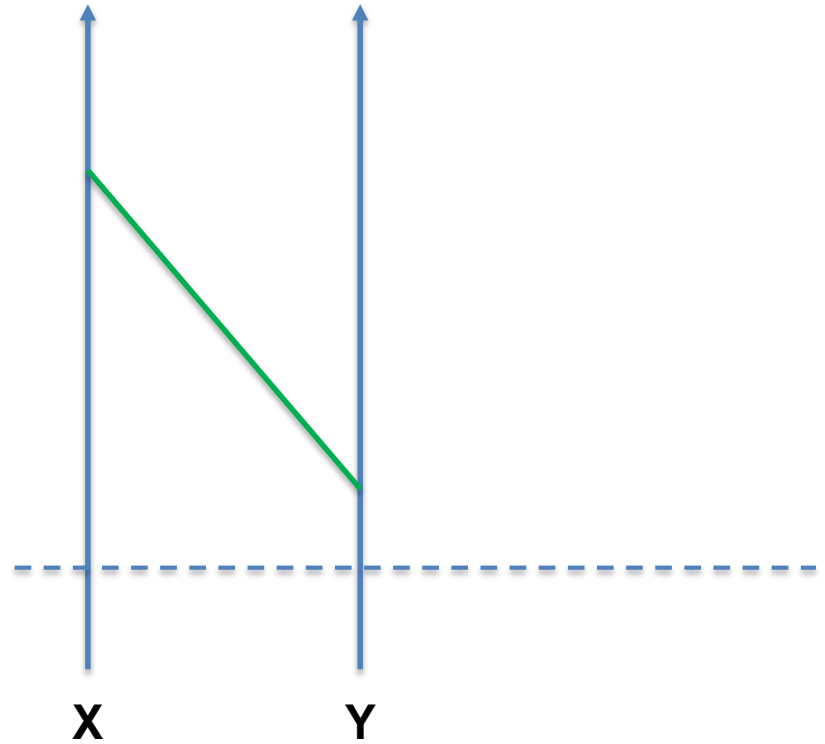
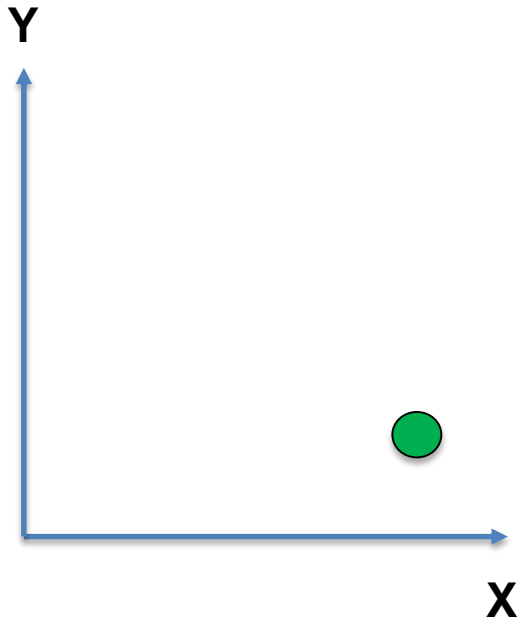
Indicators

Parallel Coordinates

- Originally attributed to Philbert Maurice d'Ocagne (1885).
- Extends classical Cartesian Coordinates System to visualize multivariate data.
- Re-discovered and popularized by Alfred Inselberg in 1970s.

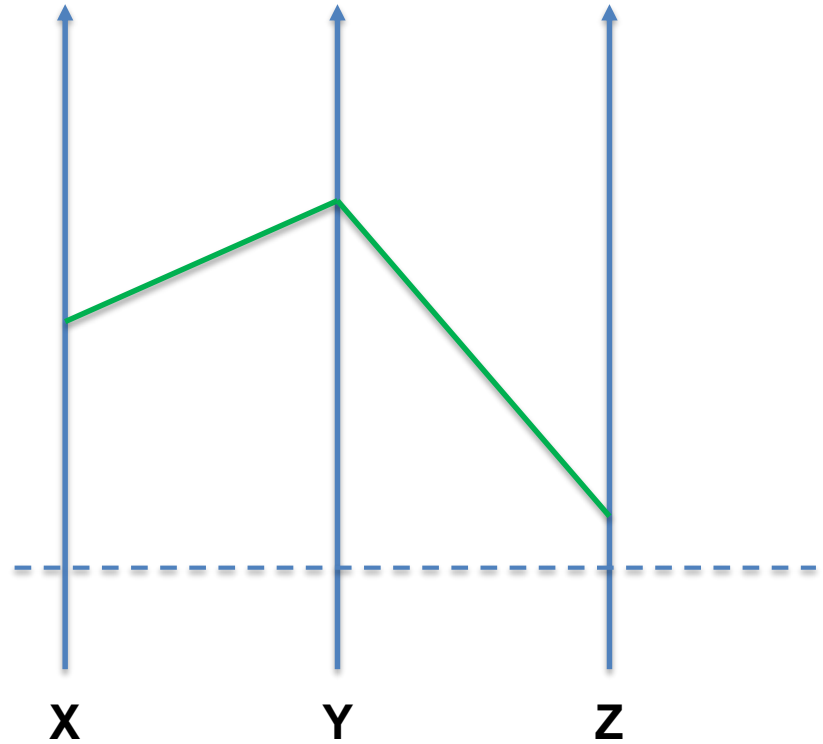
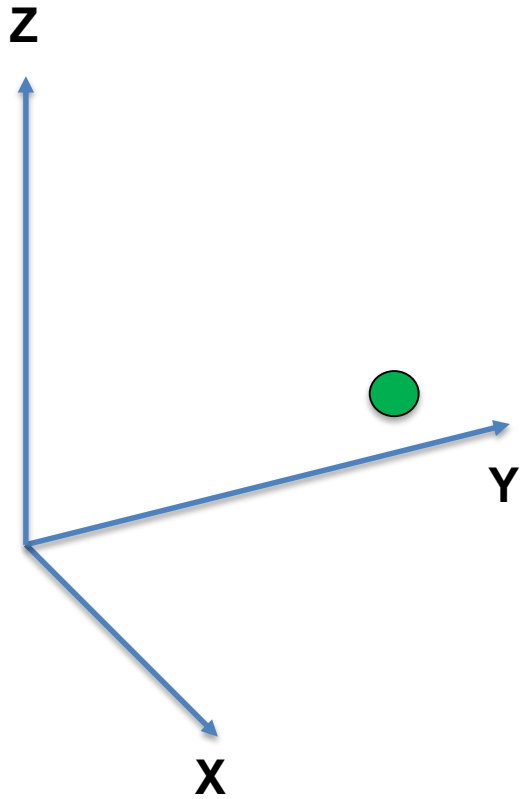
Parallel Coordinates

Two variables



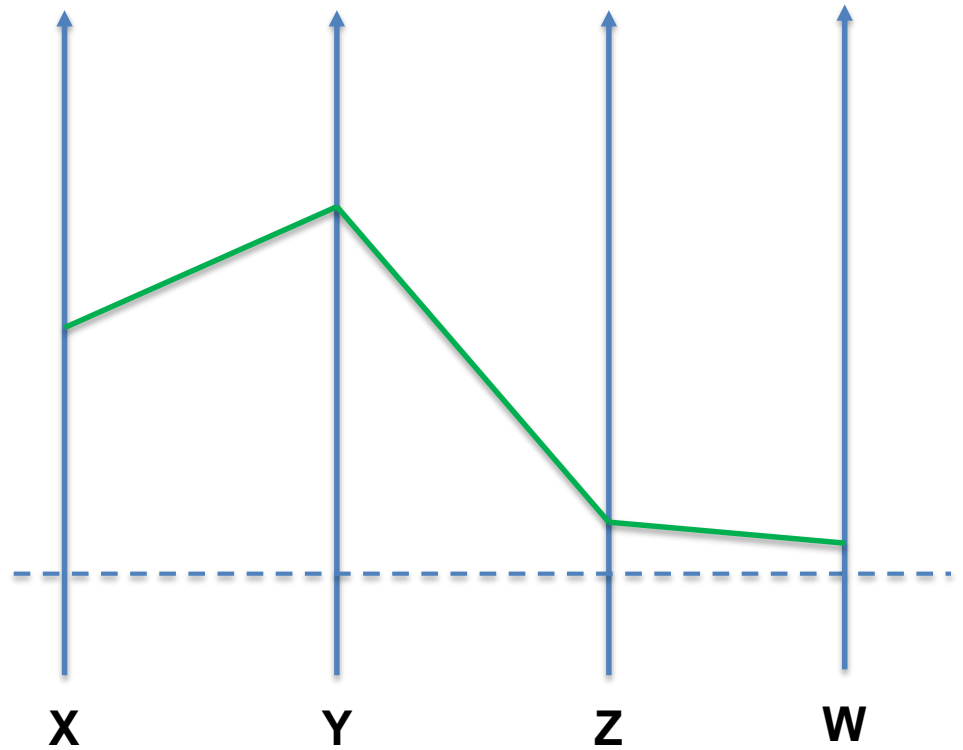
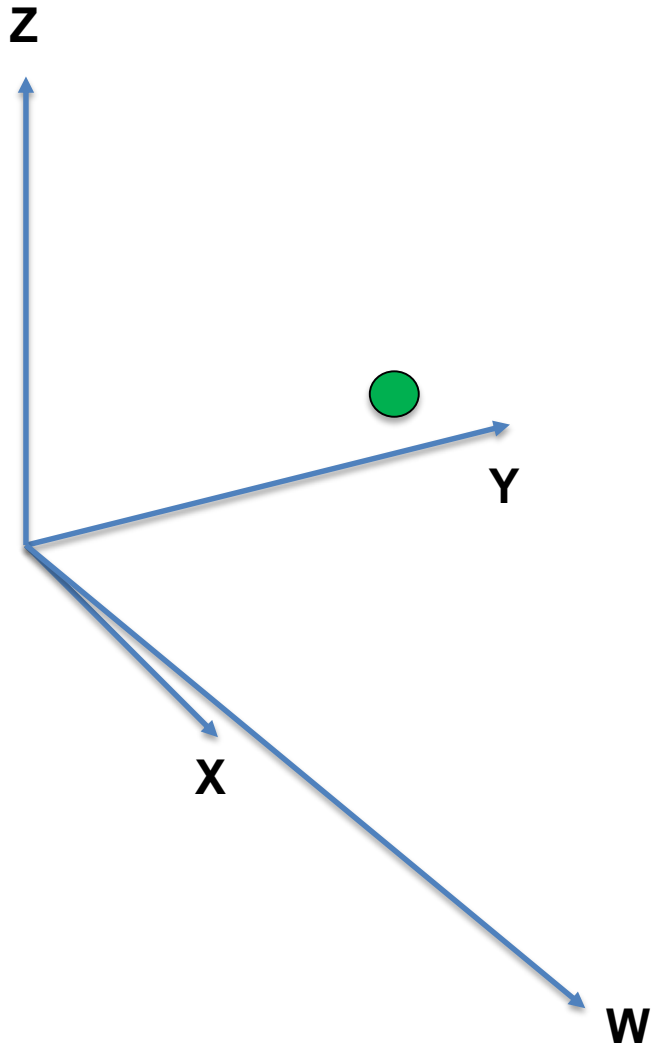
Parallel Coordinates

Three variables



Parallel Coordinates

Four Variables



Parallel Coordinates

N variables

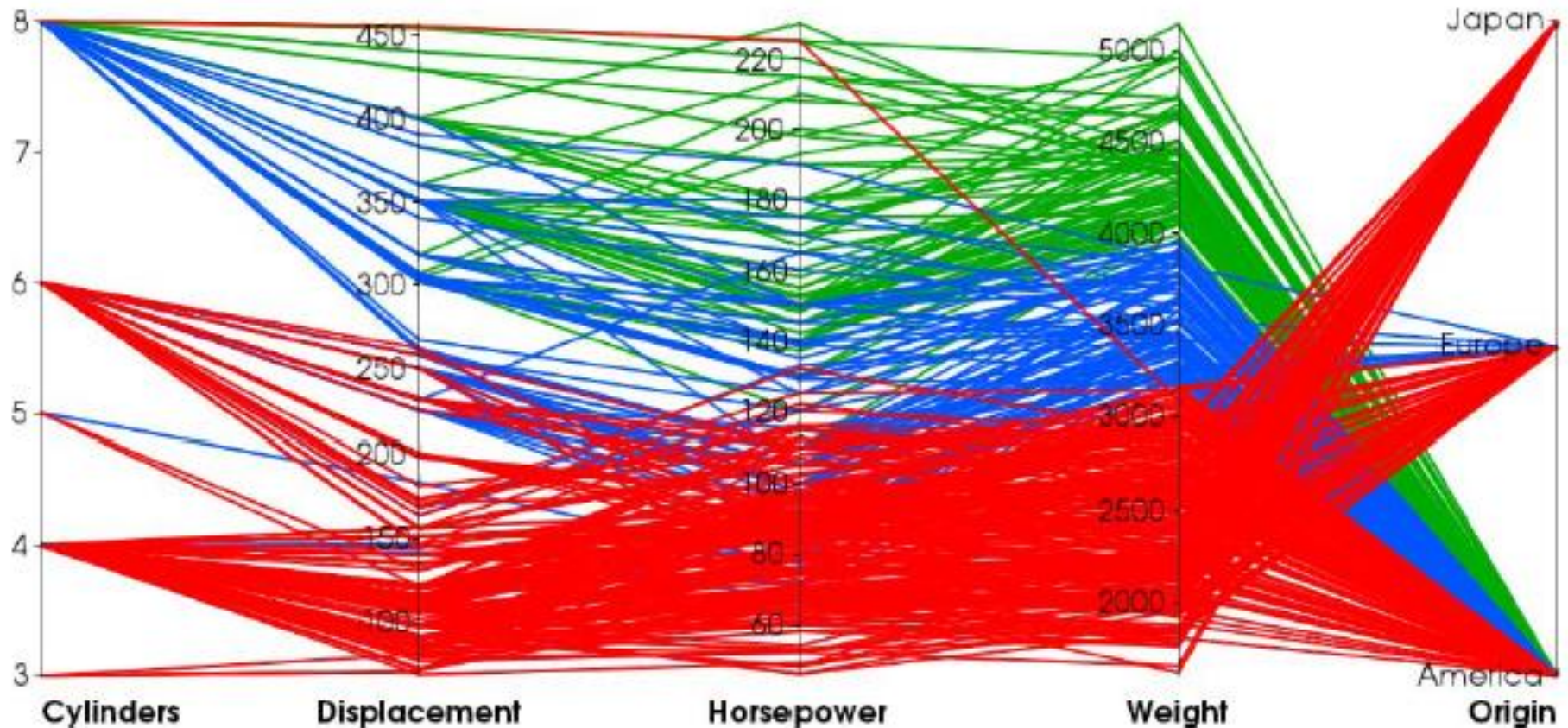



Figure from G. Palmas, M. Bachynsky, A. Oulasvirta, Hans Peter Seidel, T. Weinkauff, "An Edge-Bundling Layout for Interactive Parallel Coordinates", in PacificVis 2014.

Axis Order

- Order of the axes play a fundamental role in readability.
- But we have $n!$ combinations. 
- Try by yourself at this [website](#).
- Many strategies have been investigating for axes re-ordering.

Re-ordering and Edge-bundling

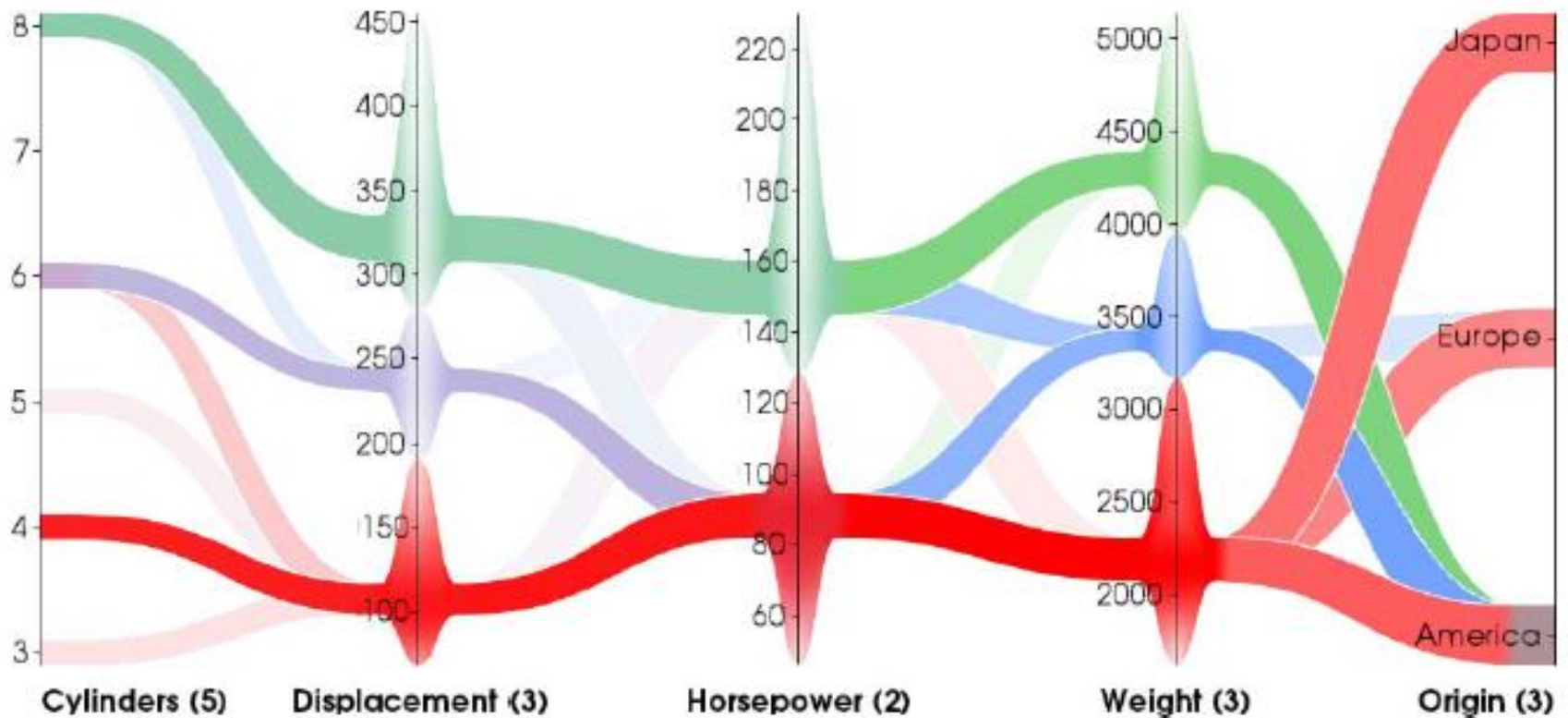


Figure from G. Palmas, M. Bachynsky, A. Oulasvirta, Hans Peter Seidel, T. Weinkauff, "An Edge-Bundling Layout for Interactive Parallel Coordinates", in PacificVis 2014.

Illustrative Rendering and Parallel Coordinates

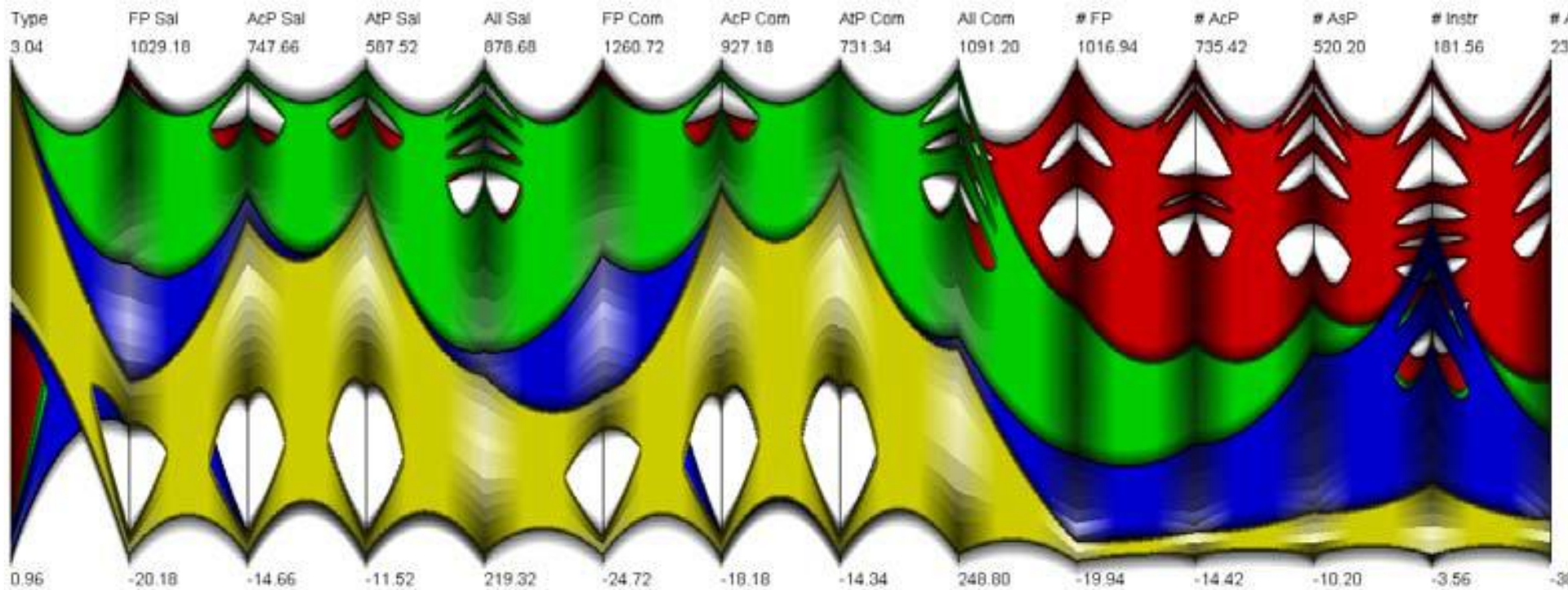
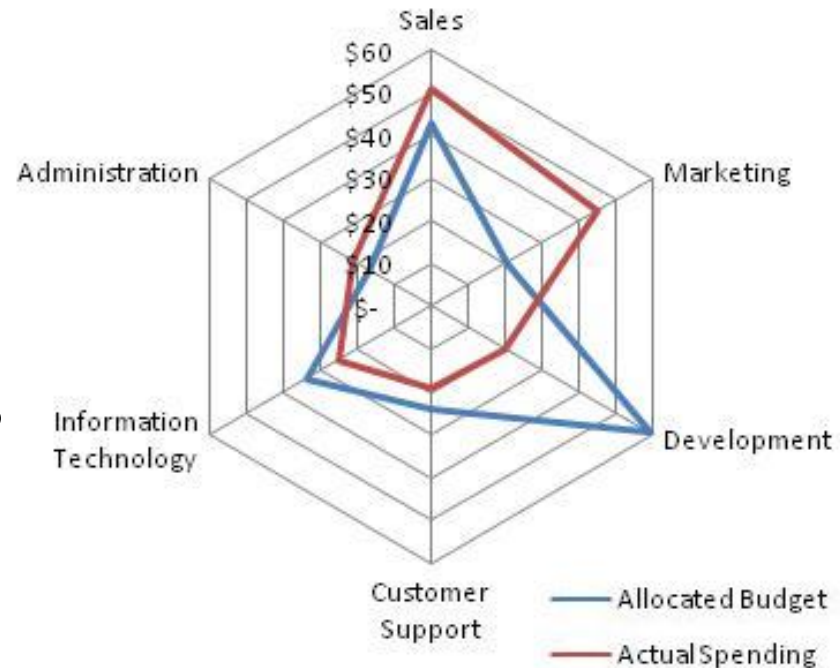


Figure from K. T. McDonnell and K. Mueller, *“Illustrative Parallel Coordinates”*, in *Computer Graphics Forum*, vol. 27, no. 3, pp. 1031-1038, 2008.

Star Plot

- Known with many names: radar chart, spider chart, web chart, etc.
- Analogous to parallel coordinates, but the axes are positioned in polar coordinates (equi-angular).
- Position of the first axis is uninformative.



Star Plot

- Easy to compare properties of a class of objects or a category.
- Not easy to understand trade-off between different variables.
- Not suitable for many variables or many data.

Star Plot

- Example: measure the quality of a photograph.



Original



Edited

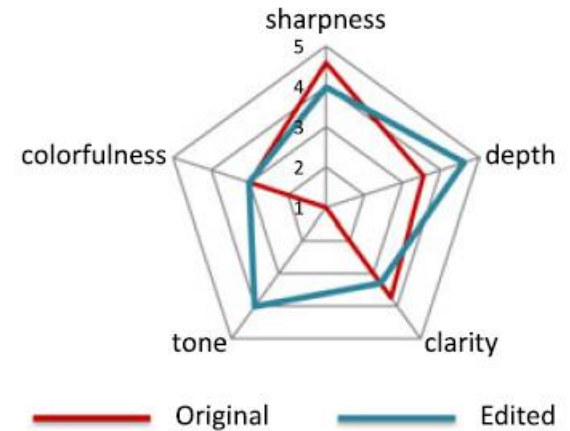


Figure from T. O. Aydın, A. Smolic and M. Gross, "*Automated Aesthetic Analysis of Photographic Images*", in IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 1, pp. 31-42.

Brief Recap

- Many solutions exists (a recent survey cites more than 250 papers).
- We have seen some of the most famous methods (e.g. SPLOM, parallel coordinates).
- A different approach is to reduce the dimensionality of the data → *dimensionality reduction*.

Dimensionality Reduction

- **Rationale:** N -dimensional data are projected to 2 or 3 dimensions for better visualization/understanding.
- Widely used strategy.
- In general, it is a mapping not a geometric transformation.
- Different mappings have different properties.

Principal Component Analysis (PCA)

- A classic multi-dimensional reduction technique is Principal Component Analysis (PCA).
- It is a linear non-parametric technique.
- The core idea to find a basis formed by the directions that maximize the variance of the data.

PCA as a change of basis

- The idea is to express the data in a new basis, that *best* express our dataset.

$$\mathbf{P}\mathbf{X} = \mathbf{Y}$$

- The new basis is a linear combination of the original basis.

PCA as a change of basis

$$\mathbf{P}\mathbf{X} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_m \end{bmatrix} [\mathbf{x}_1 \ \dots \ \mathbf{x}_n]$$

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{p}_1 \cdot \mathbf{x}_i \\ \vdots \\ \mathbf{p}_m \cdot \mathbf{x}_i \end{bmatrix}$$

Signal-to-noise Ratio (SNR)

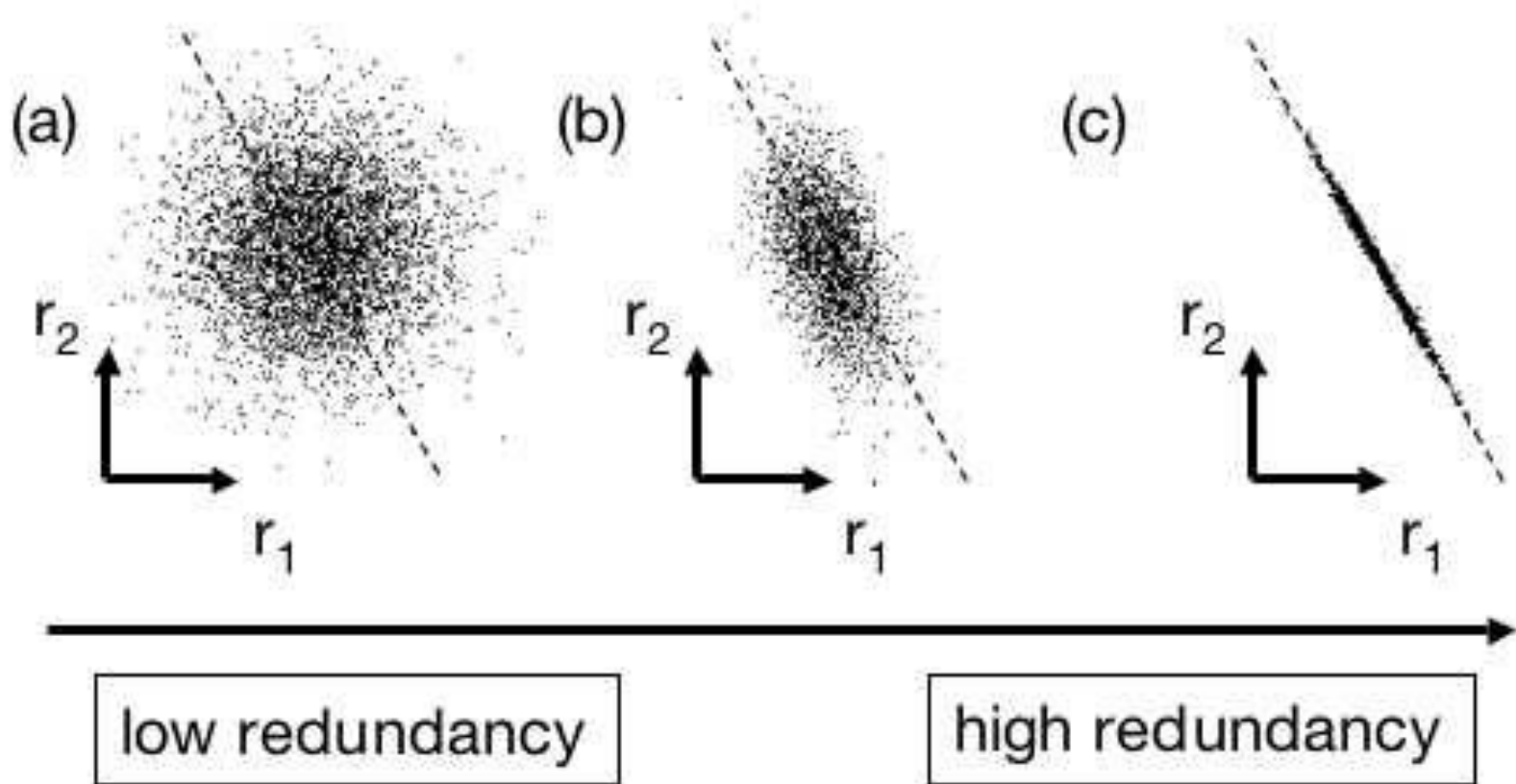
- Given a signal with noise:

$$SNR = \frac{P_{signal}}{P_{noise}}$$

- It can be expressed as:

$$SNR = \frac{\sigma_{signal}^2}{\sigma_{noise}^2}$$

Redundancy



Redundant variables convey no relevant information!

Figure From Jonathon Shlens, *"A Tutorial on Principal Component Analysis"*, arXiv preprint arXiv:1404.1100, 2015.

Covariance Matrix

$$\text{Cov}(\mathbf{X}) = \mathbf{C}_X = \frac{1}{n-1} \mathbf{X}\mathbf{X}^T$$

- Square symmetric matrix.
- The diagonal terms are the variance of a particular variable.
- The off-diagonal terms are the covariance between the different variables.

Goals

- How to select the best \mathbf{P} ?
 - Minimize redundancy
 - Maximize the variance
- Goal: to diagonalize the covariance matrix of \mathbf{Y}
 - High values of the diagonal terms means that the dynamics of the single variables has been maximized.
 - Low values of the off-diagonal terms means that the redundancy between variables is minimized.

Solving PCA

Remember that

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

$$\begin{aligned}\mathbf{C}_Y &= \frac{1}{n-1} \mathbf{Y}\mathbf{Y}^T \\ &= \frac{1}{n-1} (\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T \\ &= \frac{1}{n-1} \mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T \\ &= \frac{1}{n-1} \mathbf{P}(\mathbf{X}\mathbf{X}^T)\mathbf{P}^T \\ \mathbf{C}_Y &= \frac{1}{n-1} \mathbf{P}\mathbf{A}\mathbf{P}^T\end{aligned}$$

Solving PCA

- Theorem: a symmetric matrix \mathbf{A} can be diagonalized by a matrix formed by its eigenvectors as $\mathbf{A} = \mathbf{E}\mathbf{D}\mathbf{E}^T$.
- The column of \mathbf{E} are the eigenvectors of \mathbf{A} .

PCA Computation

- Organize the data as an $m \times n$ matrix.
- Subtract the corresponding mean to each row.
- Calculate the eigenvalues and eigenvectors of XX^T .
- Organize them to form the matrix P .

PCA for Dimensionality Reduction

- The idea is to find the k -th principal components ($k < m$).
- Project the data on these directions and use such data instead of the original ones.
- This data are the best approximation w.r.t the sum of the squared differences.

PCA as the Projection that Minimizes the Reconstruction Error

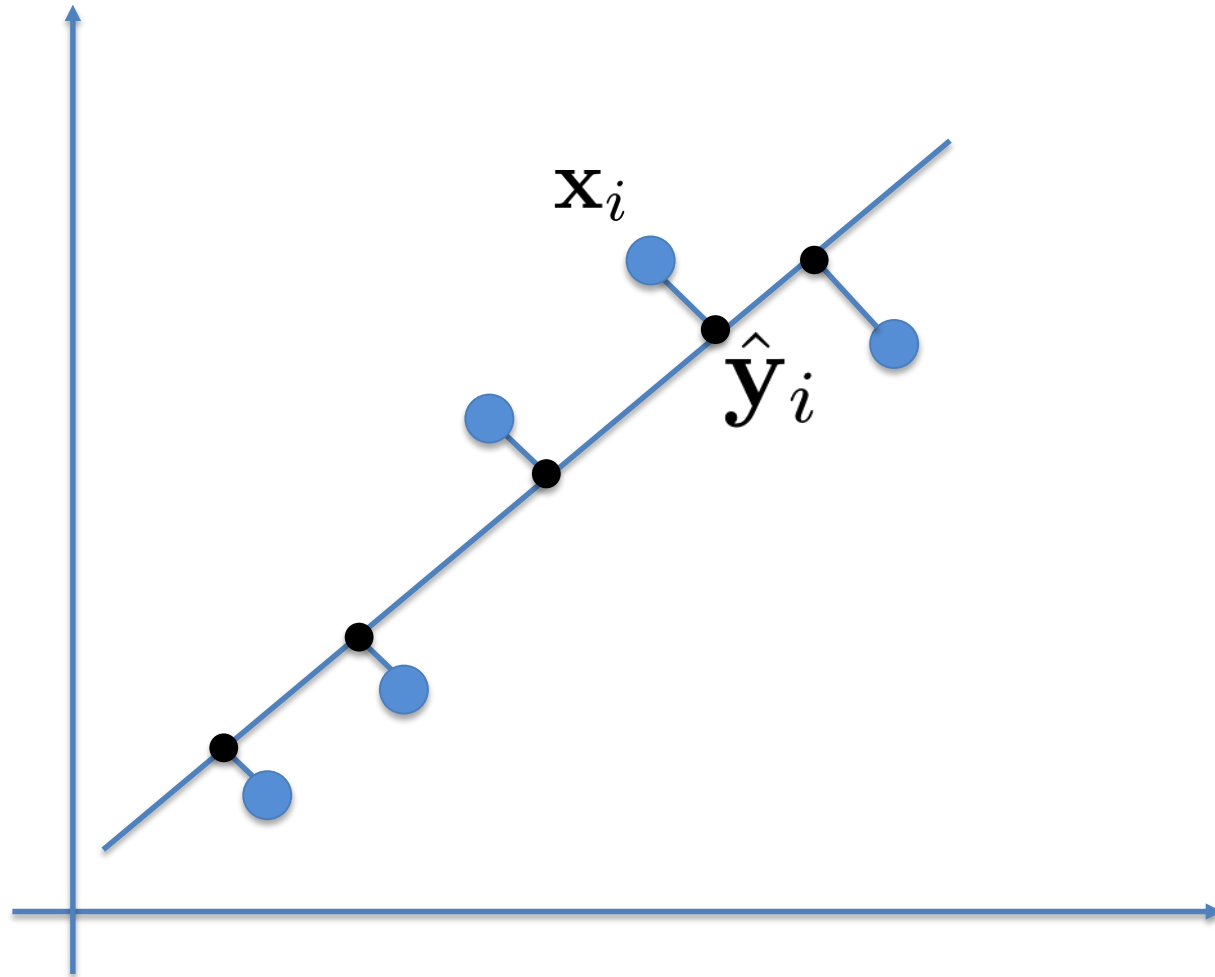
- If we use only the first $k < m$ components we obtain the best reconstruction in terms of squared error.

$$e = \sum_i (\hat{\mathbf{y}}_i - \mathbf{y}_i)^2$$

**Data point projected
on the first k components.**

**Data point projected
on all the components.**

PCA as the Projection that Minimizes the Reconstruction Error



Example

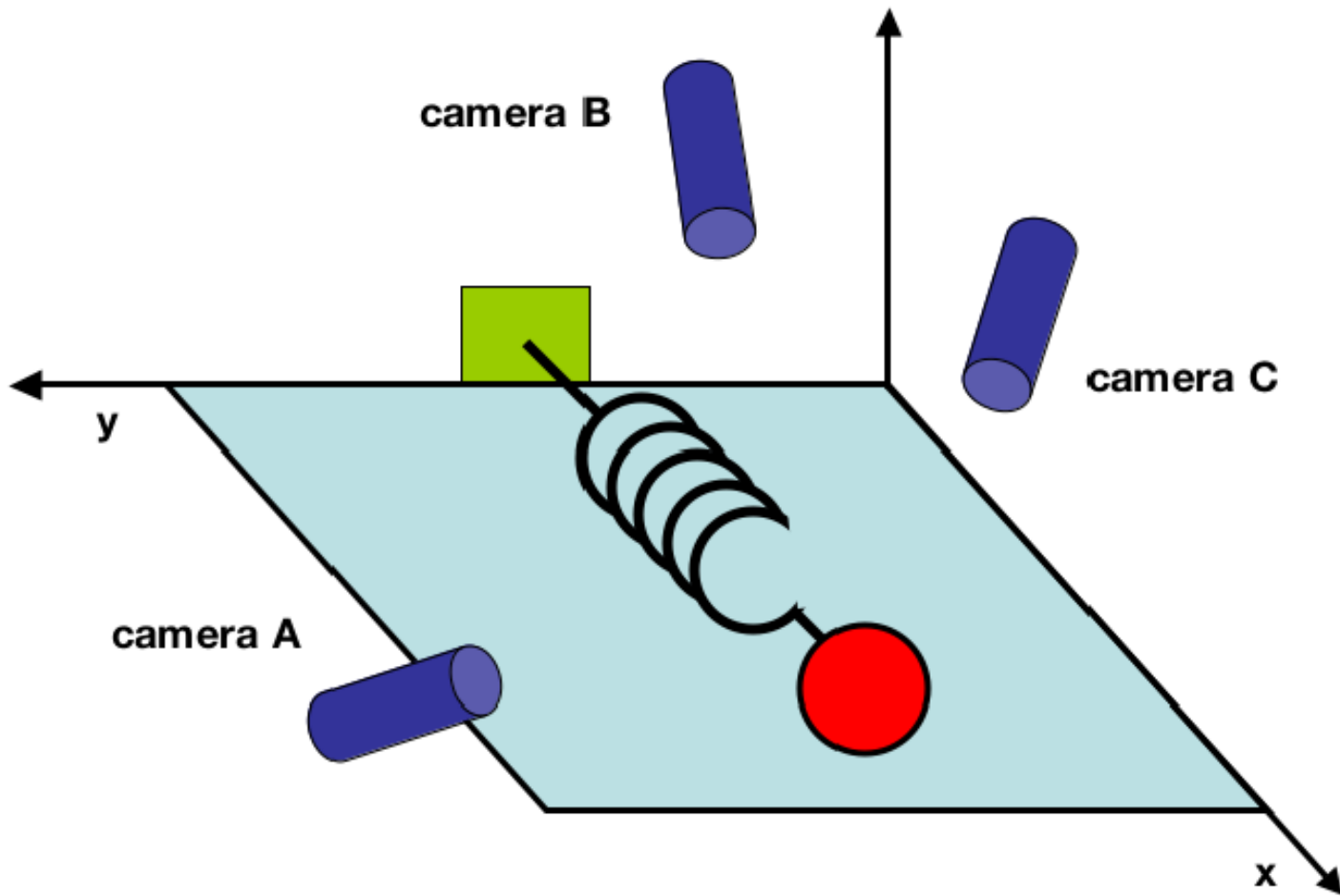


Figure From Jonathon Shlens, *"A Tutorial on Principal Component Analysis"*, arXiv preprint arXiv:1404.1100, 2015.

PCA – Example

$$m = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix}$$

Each measure has 6 dimensions (!)

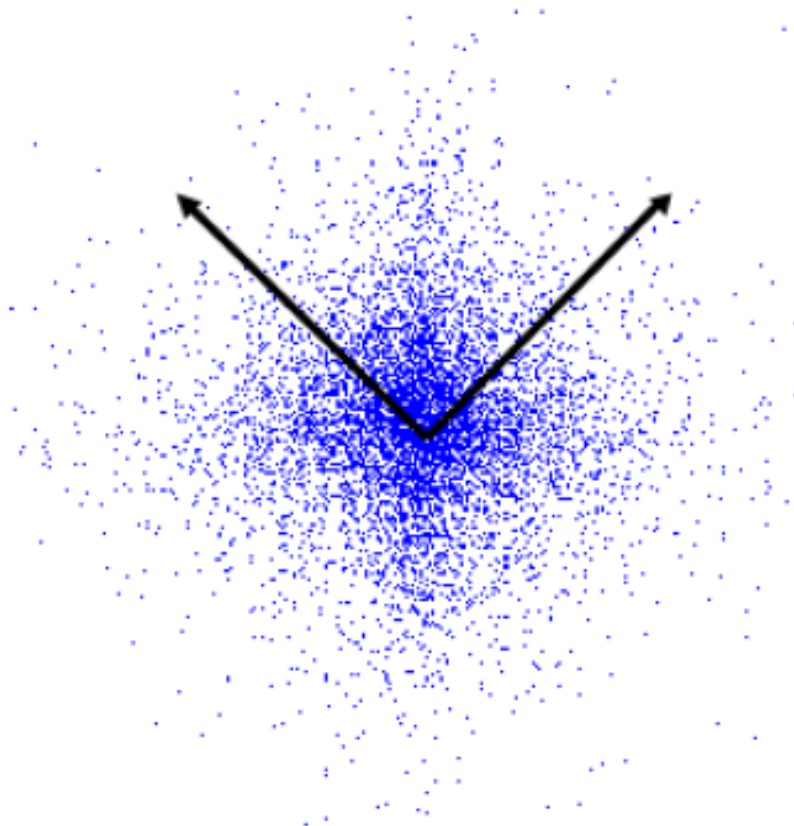
But the ball moves along the X-axis only..

Limits of PCA

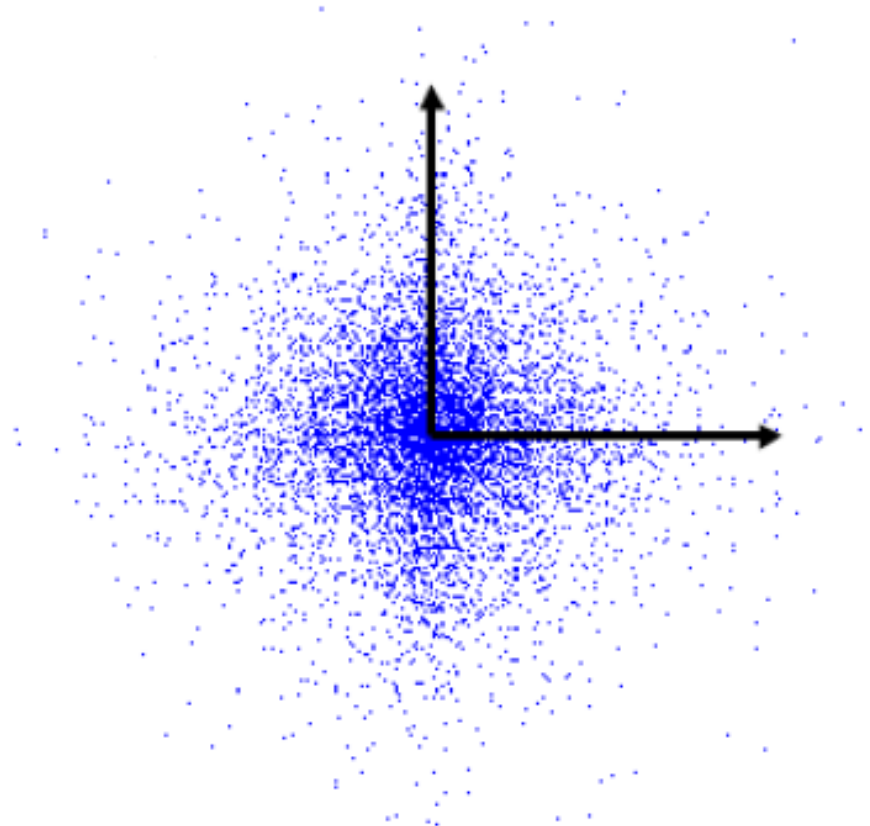
- It is non-parametric → this is a strength point but it can be also a weak point.
- It fails for non-Gaussian distributed data.
- It can be extended to account for non-linear transformation → *kernel PCA*.

Limits of PCA

PCA



ICA



ICA guarantees statistical independence $\rightarrow p(x, y) = p(x)p(y)$

Classic MDS

- Find the linear mapping $\mathbf{y}_i = \mathbf{M}\mathbf{x}_i$ which minimizes:

**Euclidean distance
in high dimensional space**

$$\phi(\mathbf{Y}) = \sum_{i,j} \overbrace{d_{ij}^2} - \underbrace{\|\mathbf{y}_i - \mathbf{y}_j\|^2}$$

**Euclidean distance
in low dimensional space**

PCA and MDS

- We want to minimize $\phi(\mathbf{Y})$, this corresponds to maximize:

$$\sum_{i,j} \|\mathbf{M}\mathbf{x}_i - \mathbf{M}\mathbf{x}_j\|^2$$

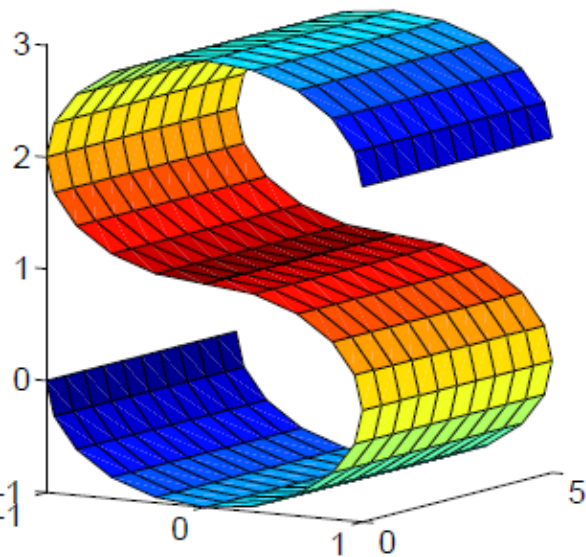
That is the variance of the low-dimensional points (same goal of the PCA).

PCA and MDS

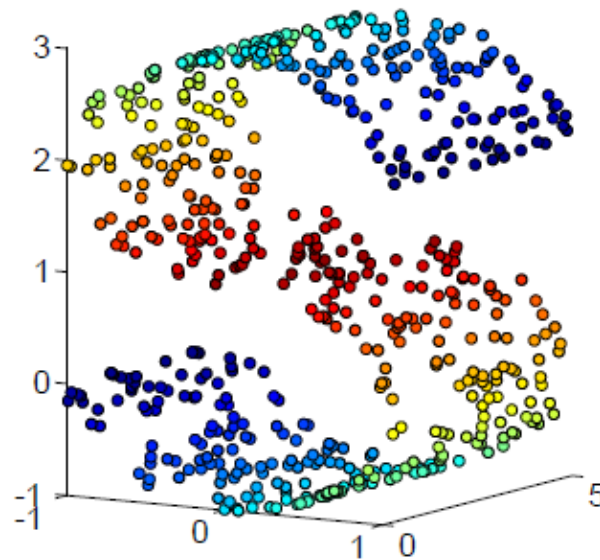
- The size of the covariance matrix is proportional to the dimension of the data.
- MDS scales with the number of data points instead of the dimensions of the data.
- Both PCA and MDS preserve better large pairwise distances.

Locally Linear Embedding (LLE)

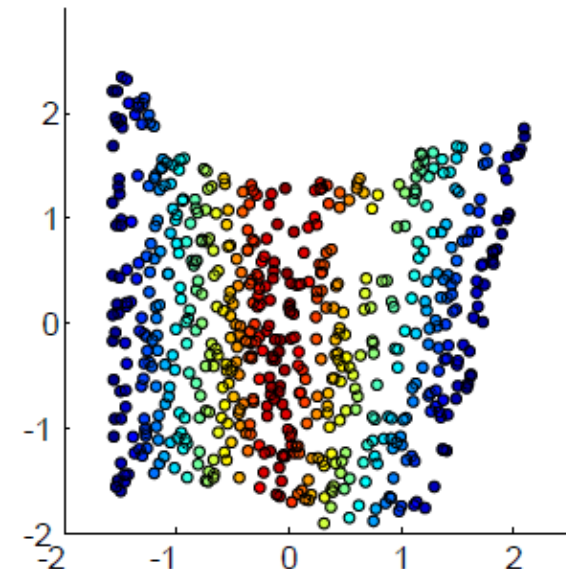
- LLE attempts to discover *nonlinear* structure in high dimension by exploiting local linear approximation.



Nonlinear Manifold M



Samples on M



Mapping Discovered

Locally Linear Embedding (LLE)

- *INTUITION* → assuming that there is sufficient data (well-sampled manifold) we expect each data point and its neighbors can be approximated by a local linear patch.
- The patch is represented by a weighted sum of the local data points.

Compute Local Patch

- Choose a set of data points close to a given one (ball-radius or K-nearest neighbours).
- Solve for W_{ij} :

$$\mathcal{E}(W) = \sum_i \left| \vec{X}_i - \sum_j W_{ij} \vec{X}_j \right|^2$$

LLE Mapping

- Find \vec{Y}_i which minimizes the embedding cost function:

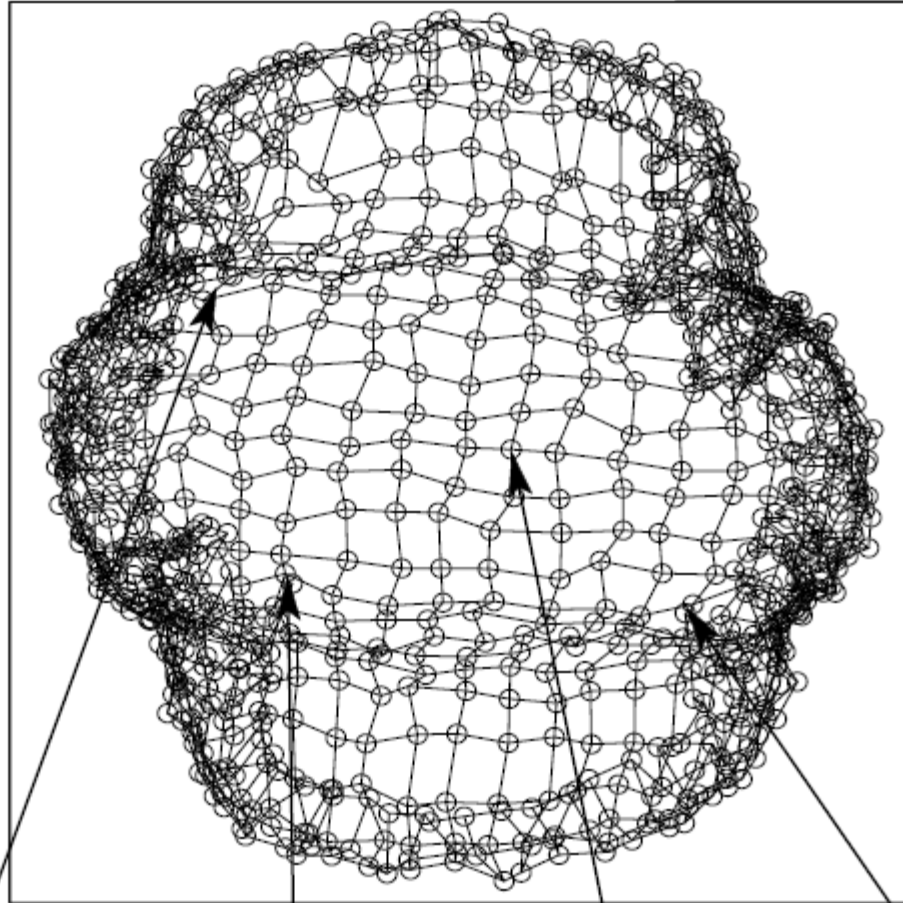
$$\Phi(Y) = \sum_i \left| \vec{Y}_i - \sum_j W_{ij} \vec{Y}_j \right|^2$$

Note that weights are fixed in this case!

LLE Algorithm

1. Compute the neighbors of each data point, \vec{X}_i .
2. Compute the weights W_{ij} that best reconstruct \vec{X}_i .
3. Compute the vectors \vec{Y}_i that minimizes the cost function.

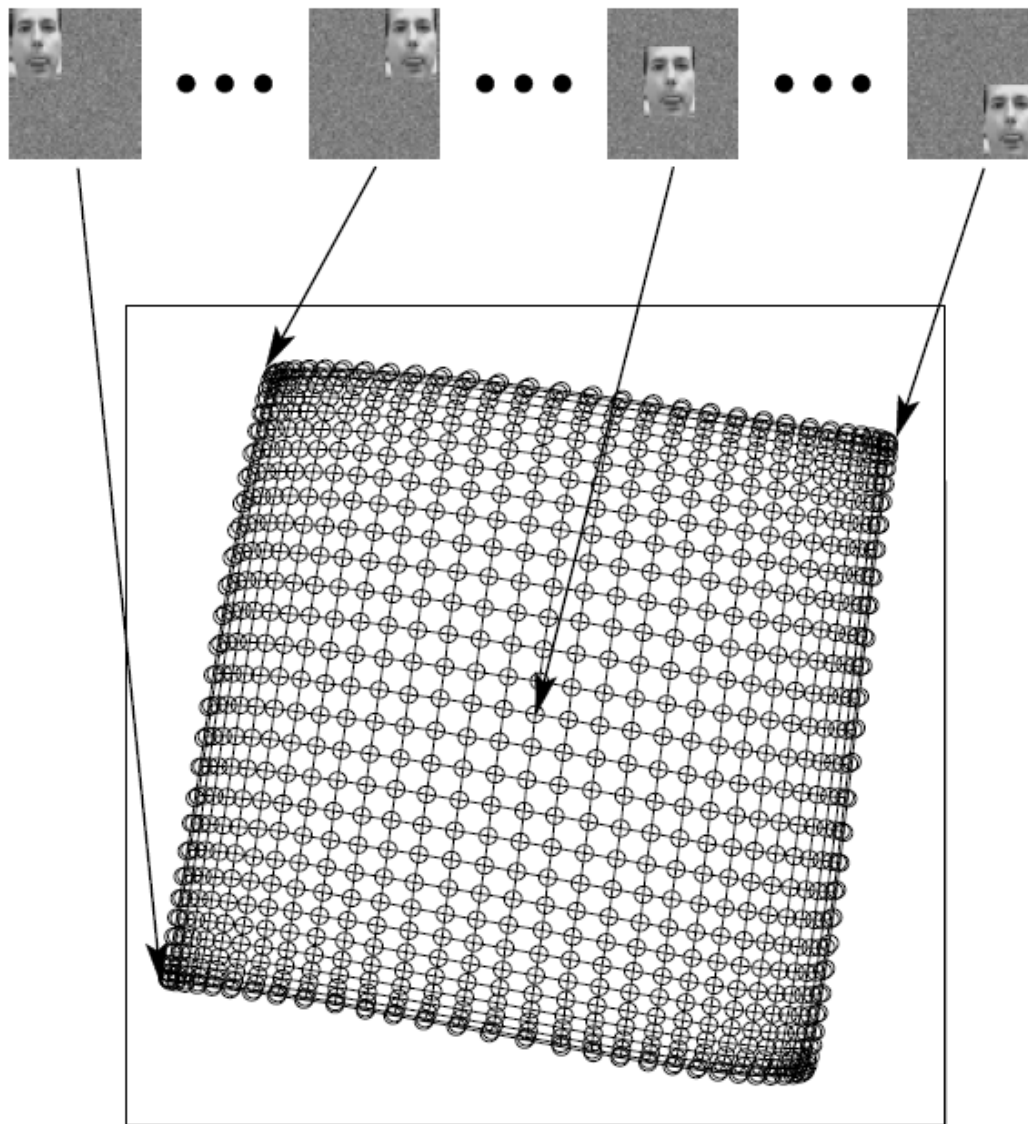
LLE – Example



PCA fails to preserve the neighborhood structure of the nearby images.

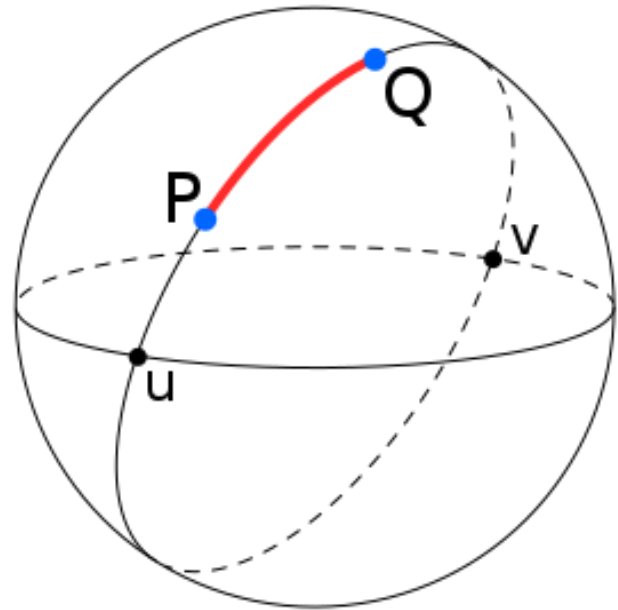


LLE – Example

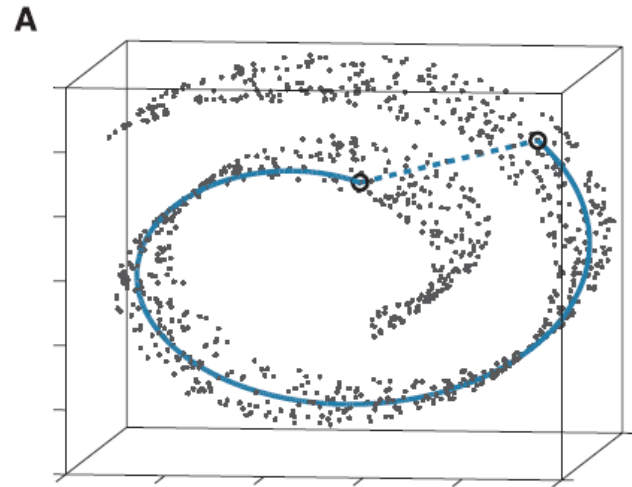


ISOMAP

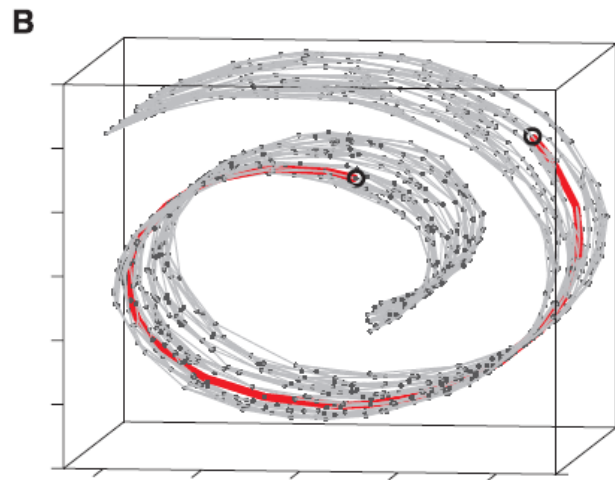
- The core idea is to preserve the geodesic distance between data points.
- Geodesic is the shortest path between two points on a curved space.



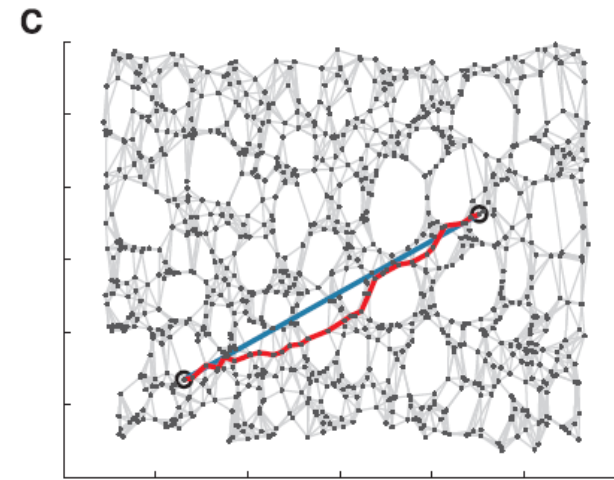
ISOMAP



**Euclidean distance
vs
Geodesic distance**



**Graph build
and
Geodesic distance
Approximation**

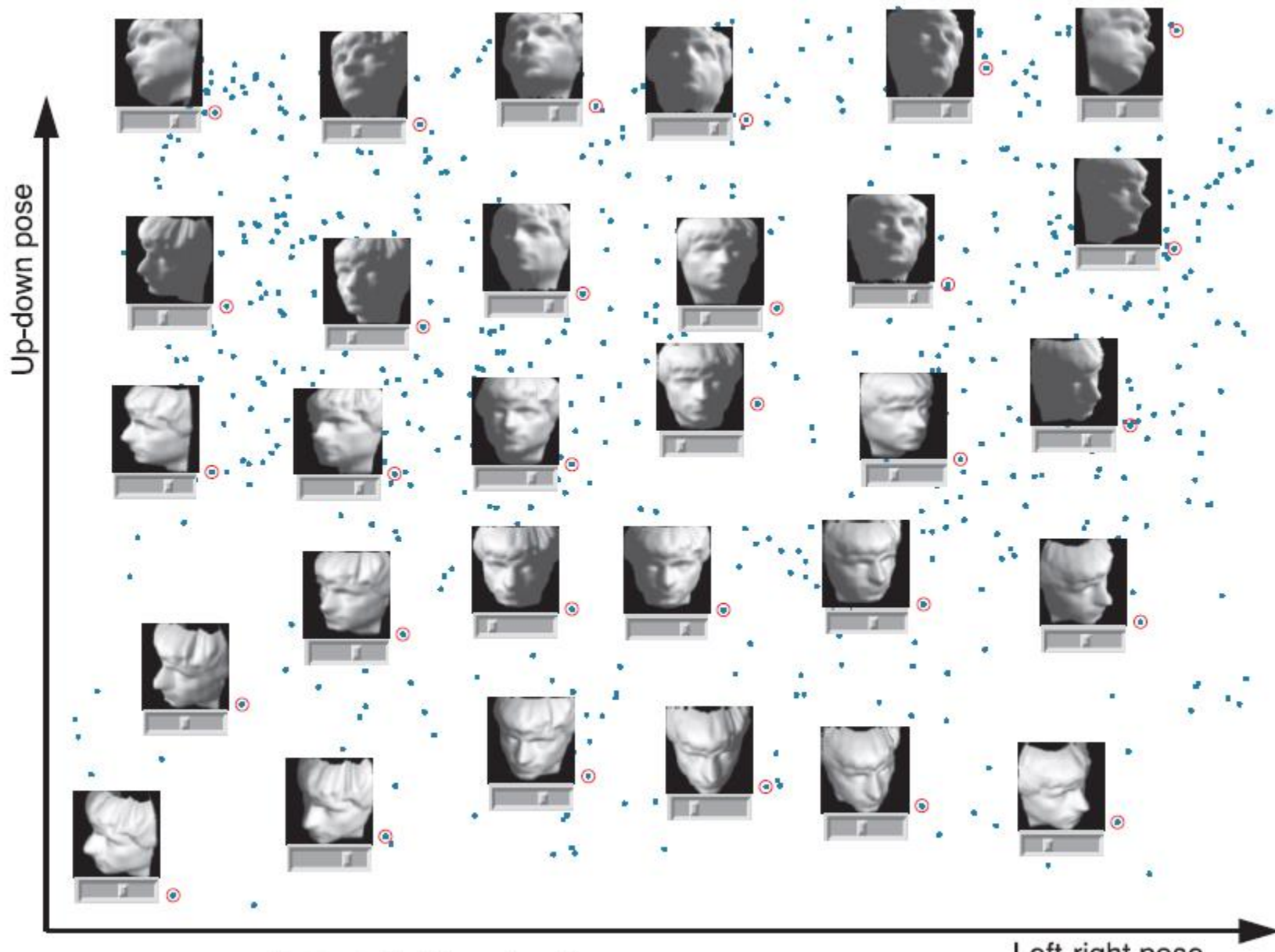


**Geodesic distance
vs
Approximated Geodesic**

ISOMAP

- Construct neighborhood graph
 - Define graph G over all data points by connecting points (i,j) if and only if the point i is a K nearest neighbor of point j
- Compute the shortest path
 - Using the Floyd's algorithm
- Construct the d -dimensional embedding

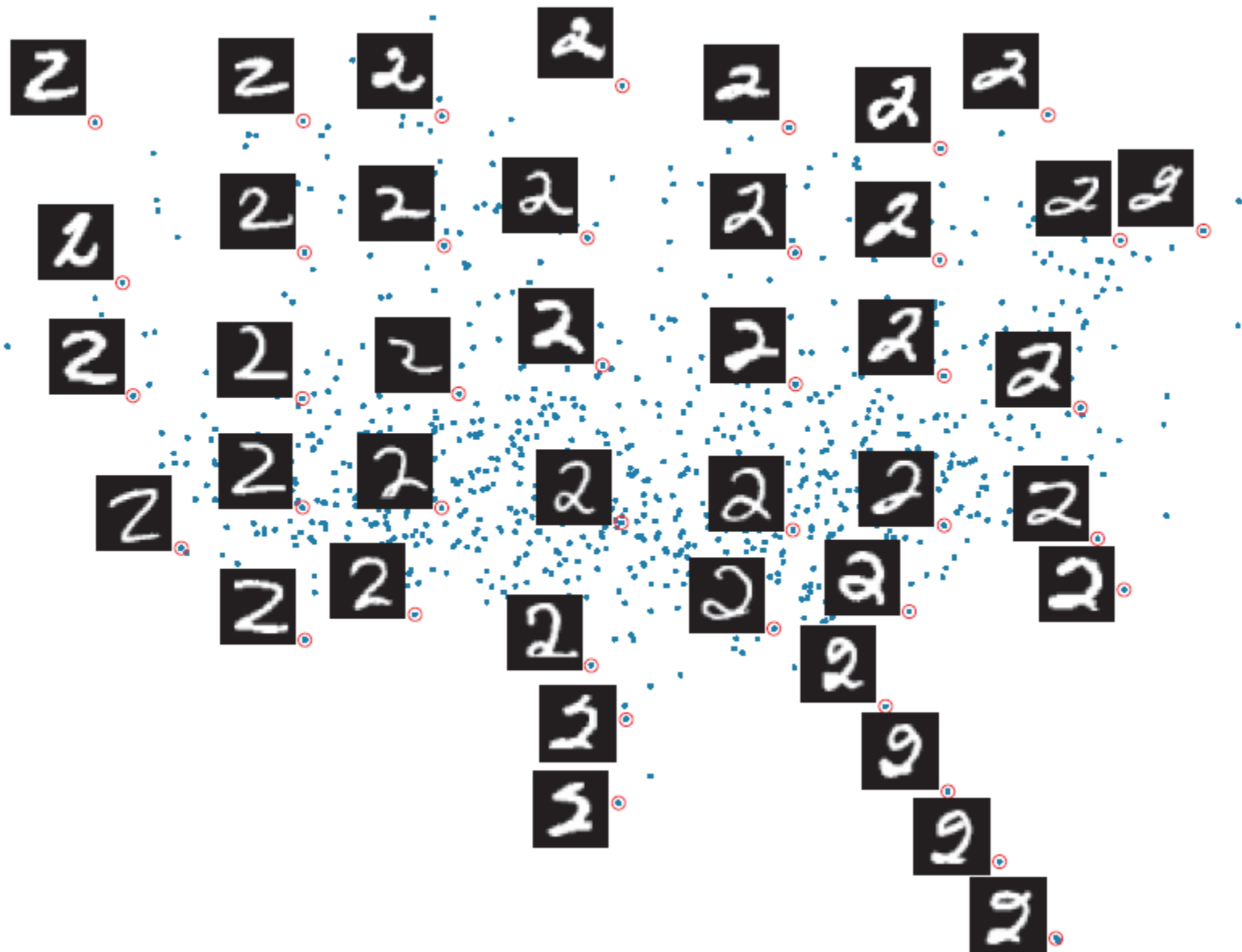
ISOMAP



ISOMAP

Bottom loop articulation →

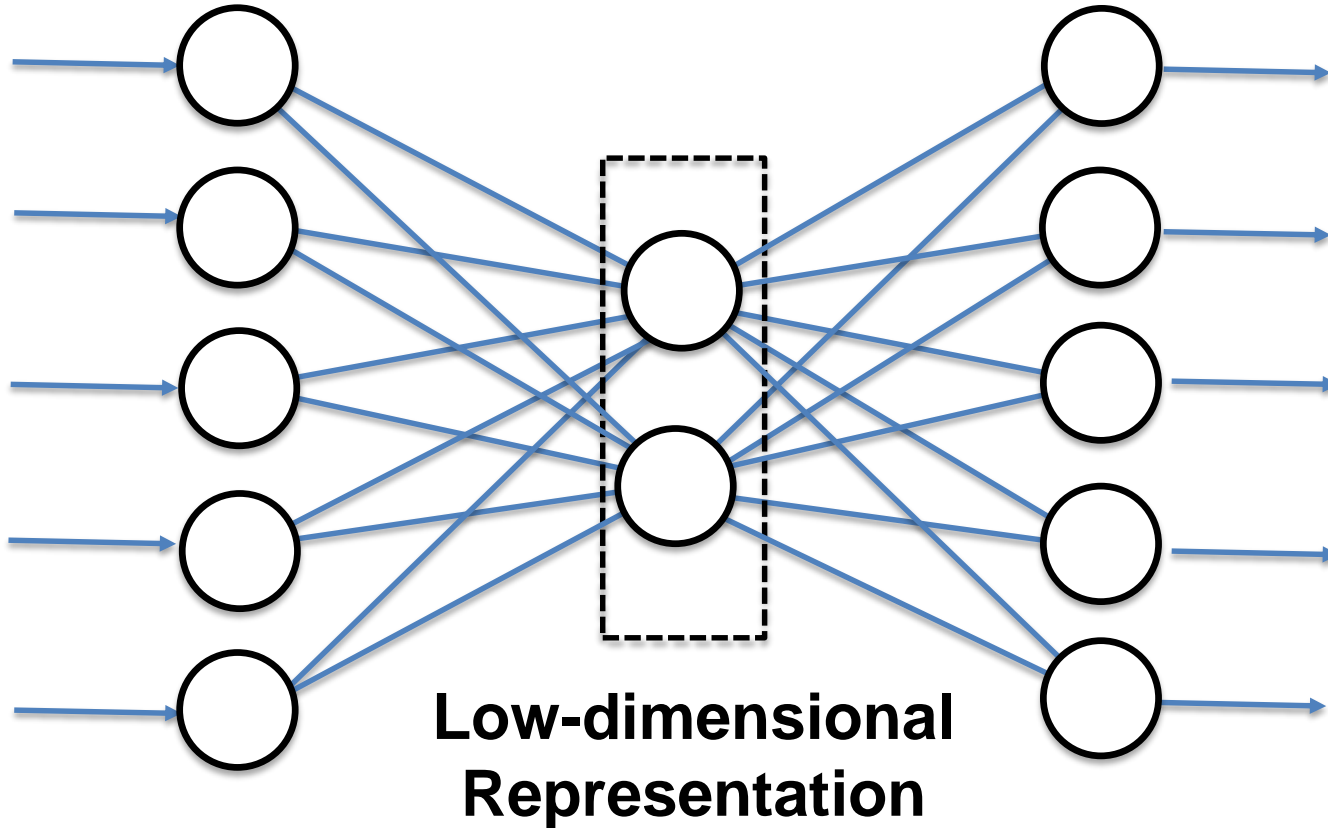
Top arch articulation ↓



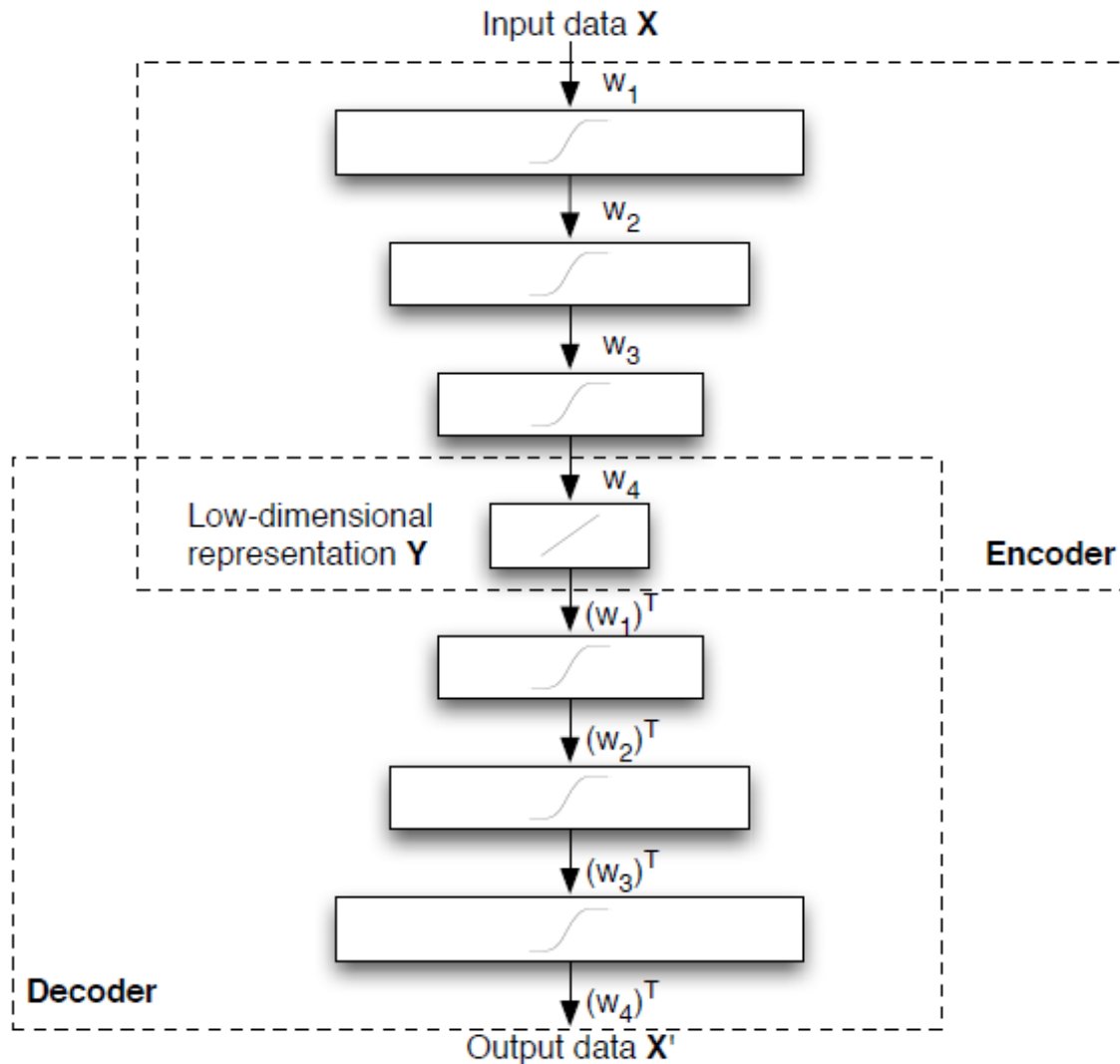
Autoencoders

- Machine learning is becoming ubiquitous in Computer Science.
- A special type of neural network is called *autoencoder*.
- An autoencoder can be used to perform dimensionality reduction.
- First, let me say something about neural network..

Autoencoder



Multi-layer Autoencoder



Summon Mapping

- Adaptation of MDS by weighting the contribution of each (i,j) pair:

$$\phi(\mathbf{Y}) = \frac{1}{\sum_{i,j} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \|\mathbf{y}_i - \mathbf{y}_j\|)^2}{d_{ij}}$$

- This allows to retain the local structure of the data better than classical scaling (the retain of high distances is not privileged).

t-SNE

- Most techniques for dimensionality reduction are not able to retain both the local and the global structure of the data in a single map.
- Simple tests on handwritten digits demonstrate this (Song et al. 2007).

Stochastic Neighbor Embedding (SNE)

- Similarities between high- and low-dimensional data points is modeled with conditional probabilities.
- Conditional probability that the point x_i would pick x_j as its neighbor:

$$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Stochastic Neighbor Embedding (SNE)

- We are interested only in pairwise distance

$$p_{i|i} = 0$$

- For the low-dimensional points an analogous conditional probability is used:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

Kullback-Leibler Divergence

- Coding theory: expected number of extra bits required to code samples from the distribution P if the current code is optimized for the distribution Q .
- Bayesian view: a measure of the information gained when one revises one's beliefs from the prior distribution Q to the posterior distribution P .
- It is also called *relative entropy*.

Kullback-Leibler Divergence

- Definition for discrete distributions:

$$D_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$$

- Definition for continuous distributions:

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Stochastic Neighbor Embedding (SNE)

- The goal is to minimize the mismatch between $p_{j|i}$ and $q_{j|i}$.
- Using the Kullback-Leibler divergence this goal can be achieved by minimizing the function:

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

Note that $KL(P || Q)$ is not symmetric !

Problems of SNE

- The cost function is difficult to optimize.
- SNE suffers, as other dimensionality reduction techniques, of the *crowding problem*.

t-SNE

- SNE is made symmetric:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- It employs a Student-t distribution instead of a Gaussian distribution to evaluate the similarity between points in low dimension.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

t-SNE Advantages

- The crowding problem is alleviated.
- Optimization is made simpler.

Experiments

- Comparison with LLE, Isomap and Summon Mapping.
- Datasets:
 - MNIST dataset
 - Olivetti face dataset
 - COIL-20 dataset

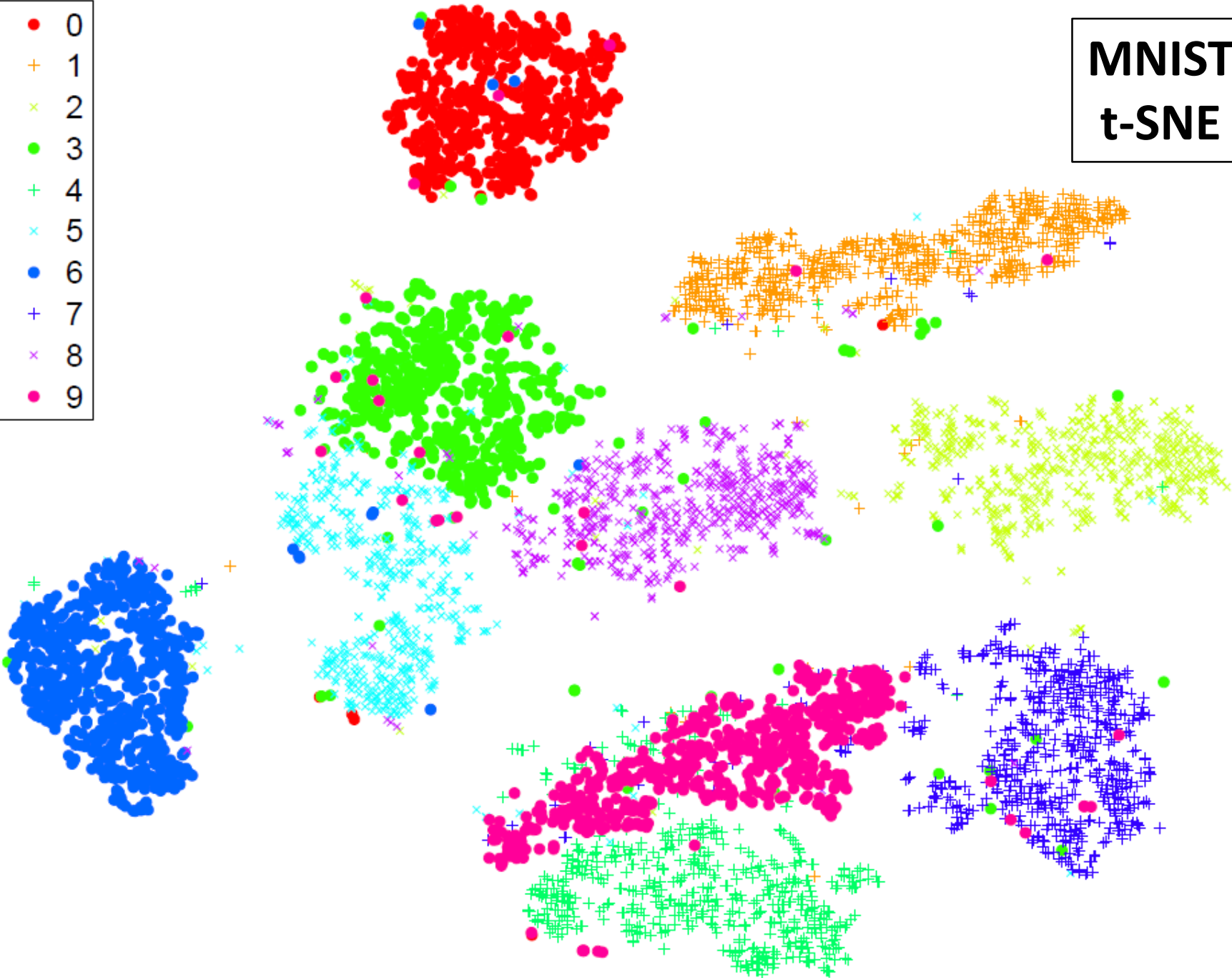
Comparison figures are from the paper L.J.P. van der Maaten and G.E. Hinton, *“Visualizing High-Dimensional Data Using t-SNE”*, Journal of Machine Learning Research, Vol. 9, pp. 2579-2605, 2008.

MNIST Dataset

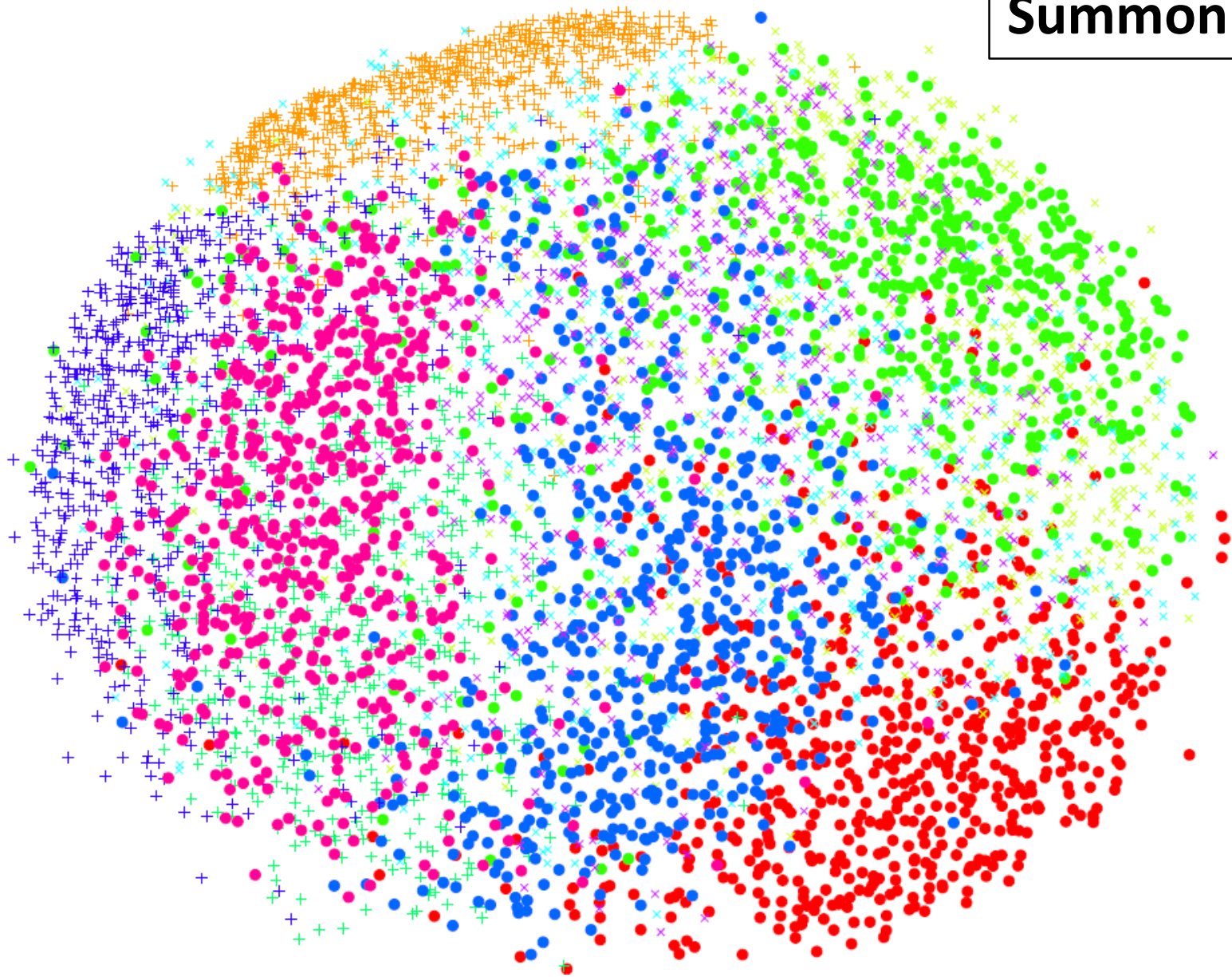
- 60,000 images of handwritten digits.
- Image resolution: 28 x 28 (784 dimensions).
- A subset of 6,000 images randomly selected has been used.

MNIST t-SNE

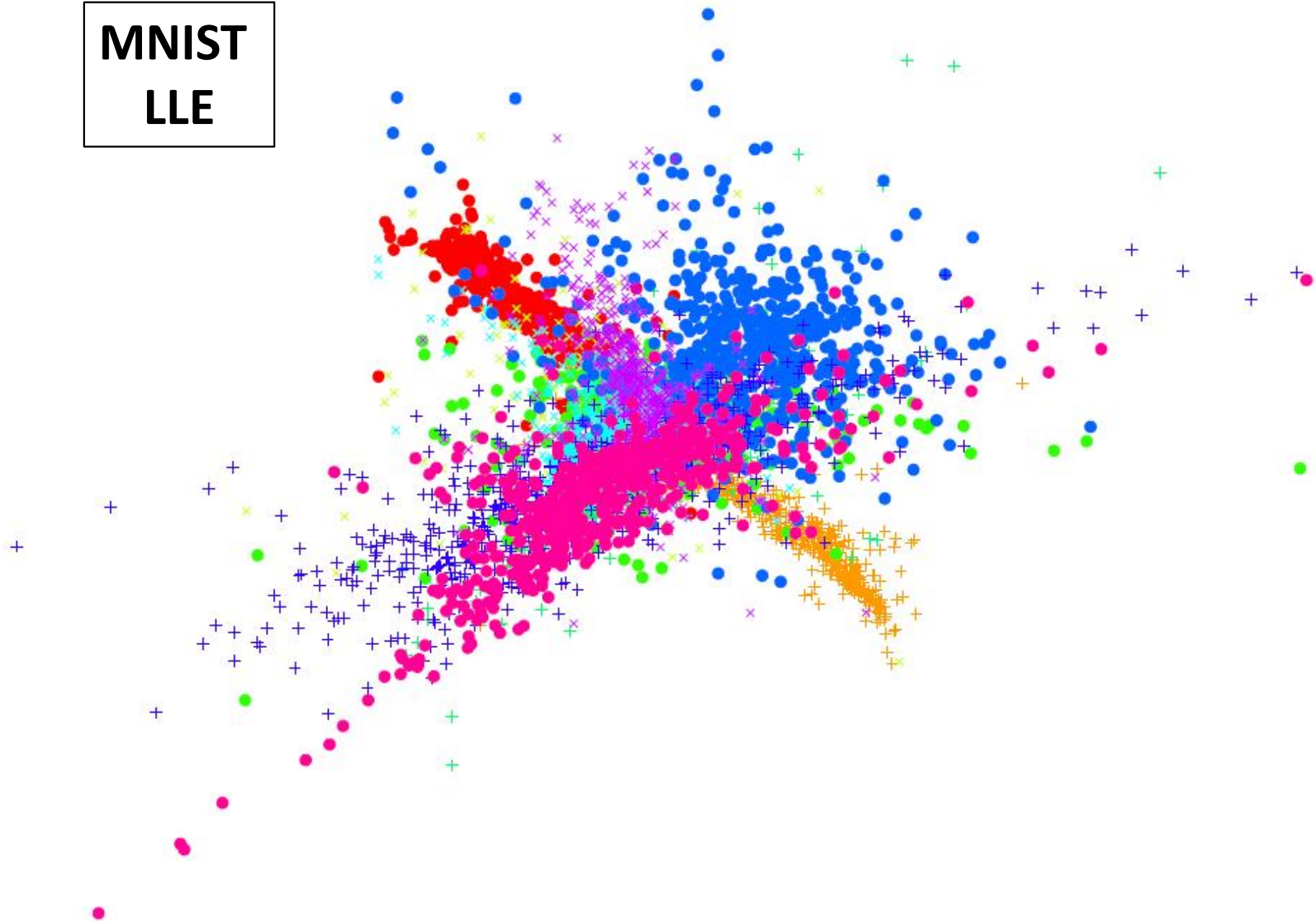
- 0
- + 1
- × 2
- 3
- + 4
- × 5
- 6
- + 7
- × 8
- 9



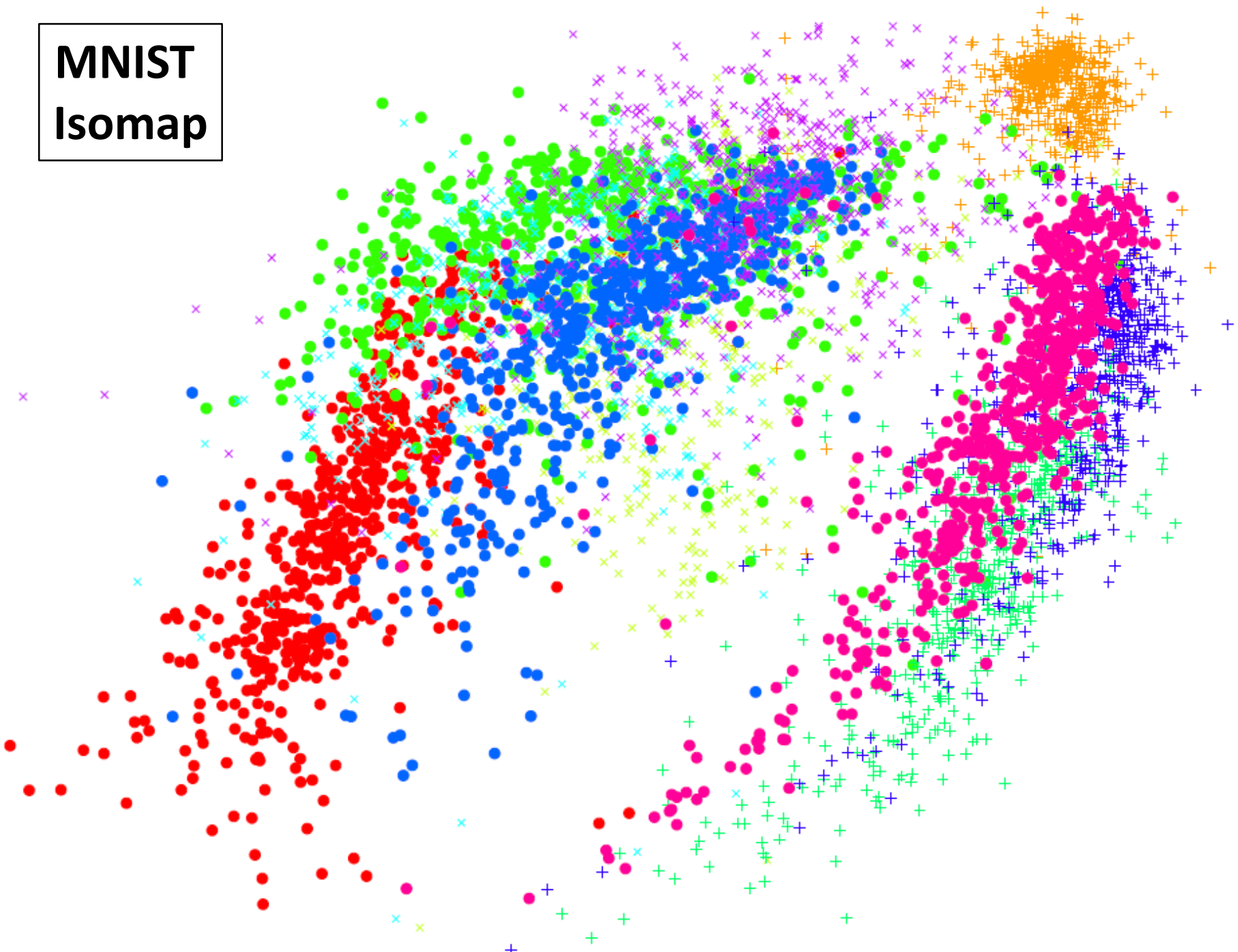
MNIST Summon Mapping



MNIST
LLE



**MNIST
Isomap**



COIL-20 Dataset

- Images of 20 objects viewed from 72 different viewpoints (1440 images).
- Image size: 32 x 32 (1024 dimensions).



COIL-20 Dataset

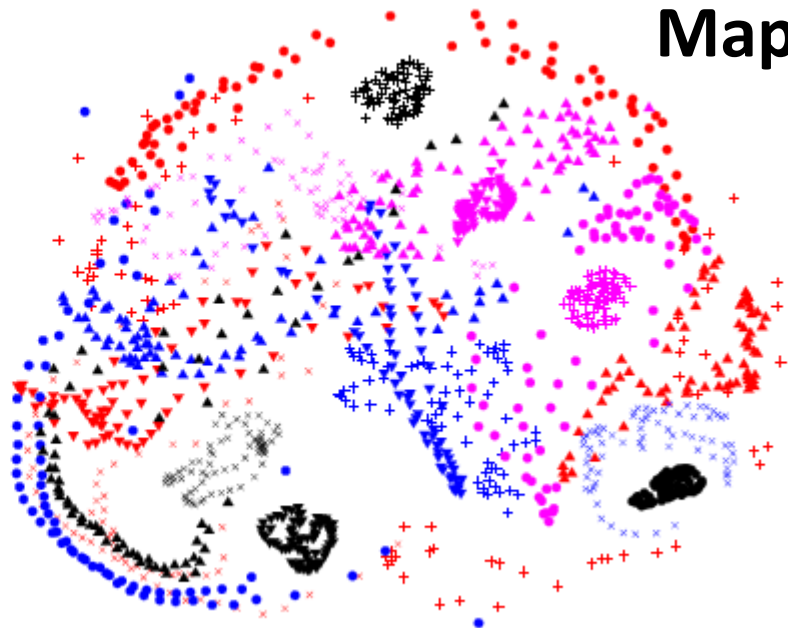


...

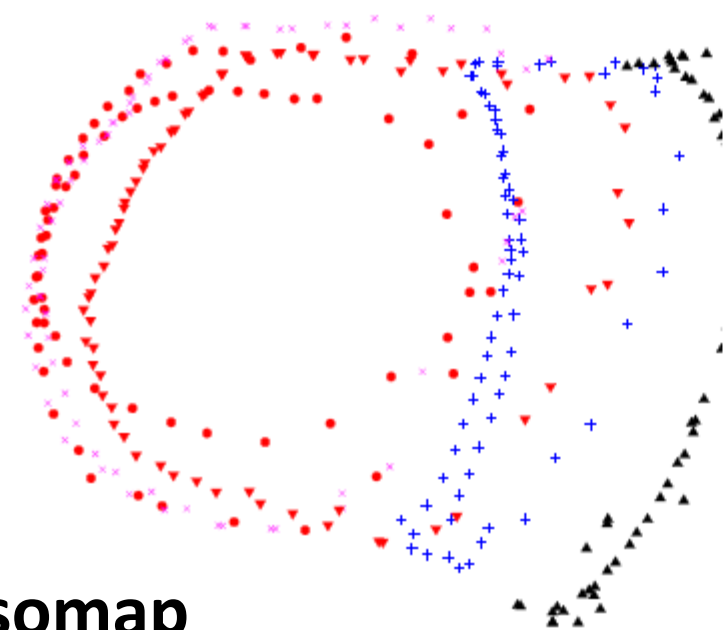
t-SNE



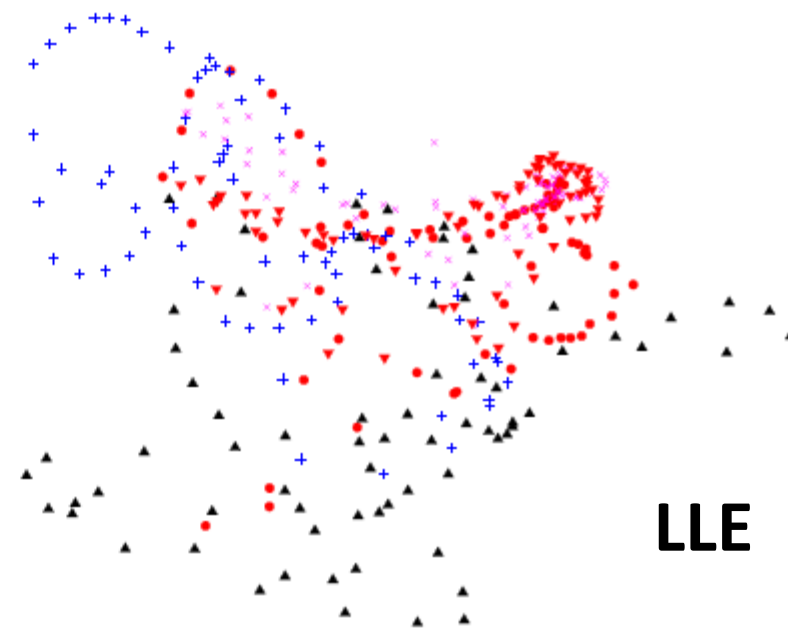
**Summon
Mapping**



Isomap



LLE



Questions ?