Advanced Data Visualization

CS 6965

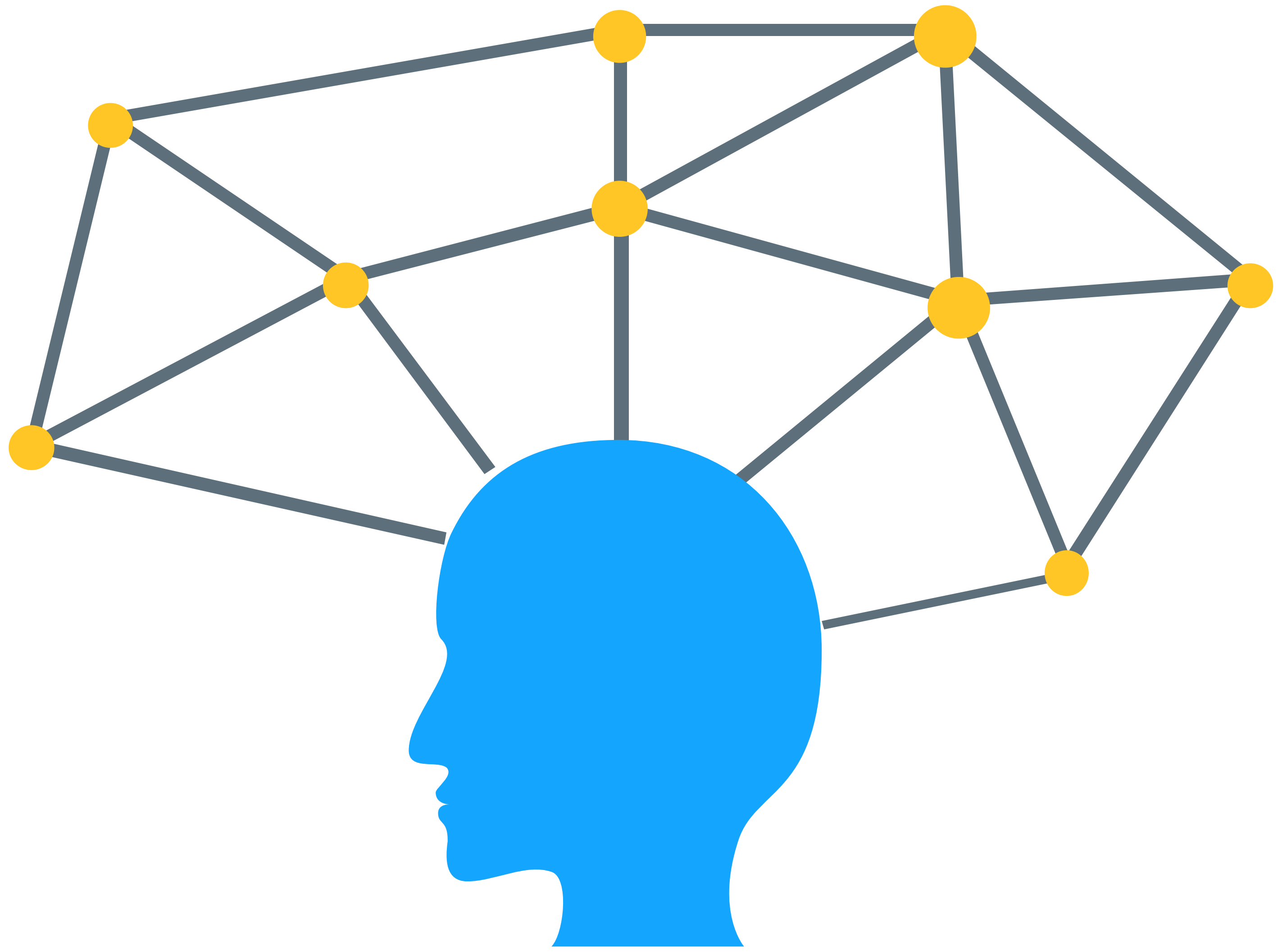
Fall 2019

Prof. Bei Wang Phillips

University of Utah

Lecture 05

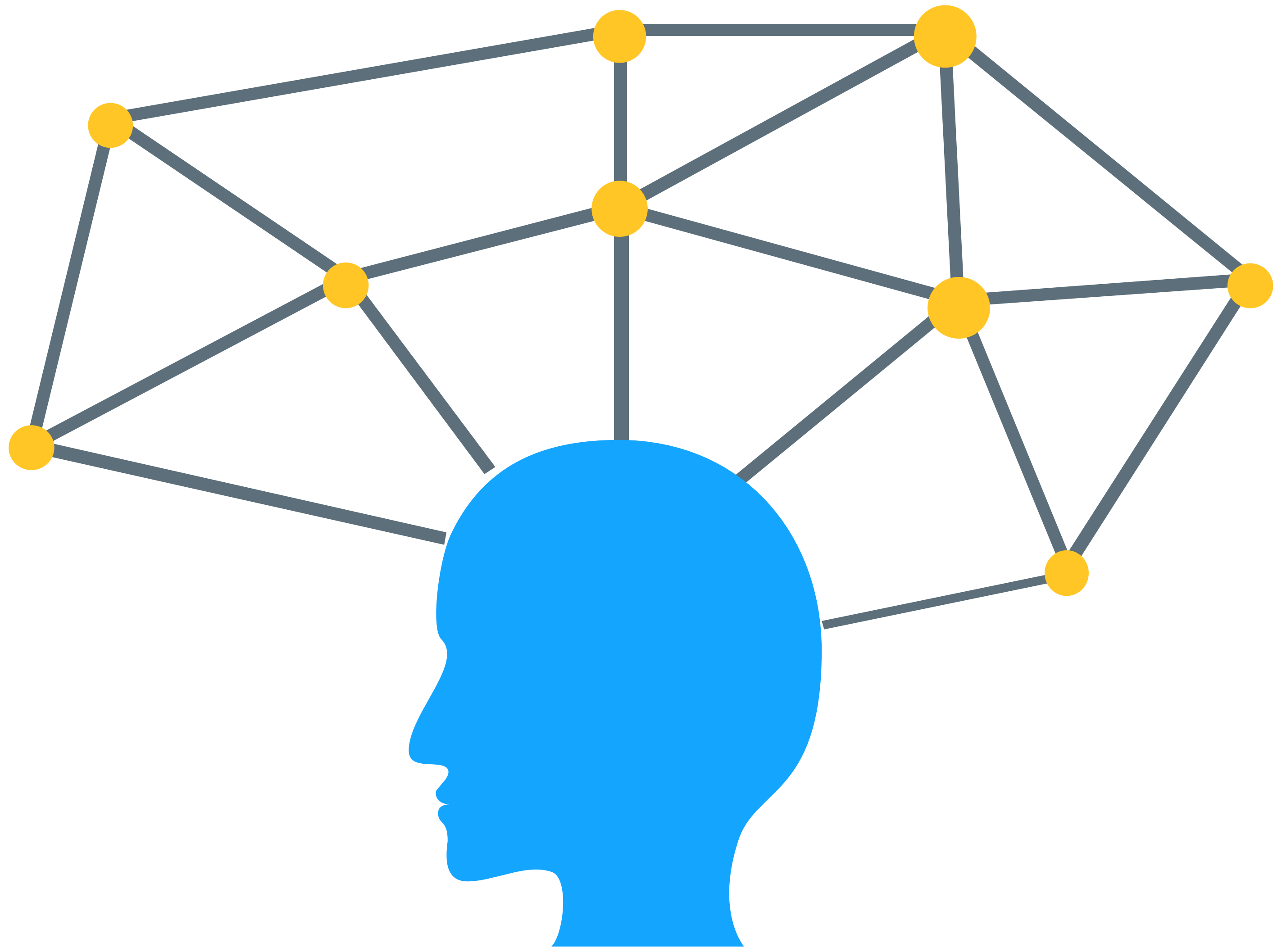
Brainstorming



Brainstorming

A spontaneous group discussion to produce ideas and ways of solving problems.

Brainstorming



Question: how to enhance an existing PCA-based visual analytic system?

+ Improve model interpretability
+ Improve data interpretability

Enhance PCA system

- Add DR techniques
- Outlier detection
- Distortion
- Dynamic projection, animation
- Capture user analytic history and user intensions
- Compare across multiple parameter settings
- Parameter tuning, parameter suggestion
- User study: guidance in exploratory analysis
- Add SVD information: visualize singular values and eigenvalues
- More…

More on DR, Clustering, and Vis

HD

A few more words on mapper algorithm

A tool for high-dimensional data analysis and visualization

Clustering algorithm

Let X be the original high-dimensional point cloud.

- Clustering algorithm applies to
  - The inverse image of the interval, which are points in the domain: a subset of X (the classic algorithm).
  - Alternatives, clustering can be applied to a transformed version of X, referred to as Y. For example, Y can be the result of DR of X.

Mapper I/O and Parameters

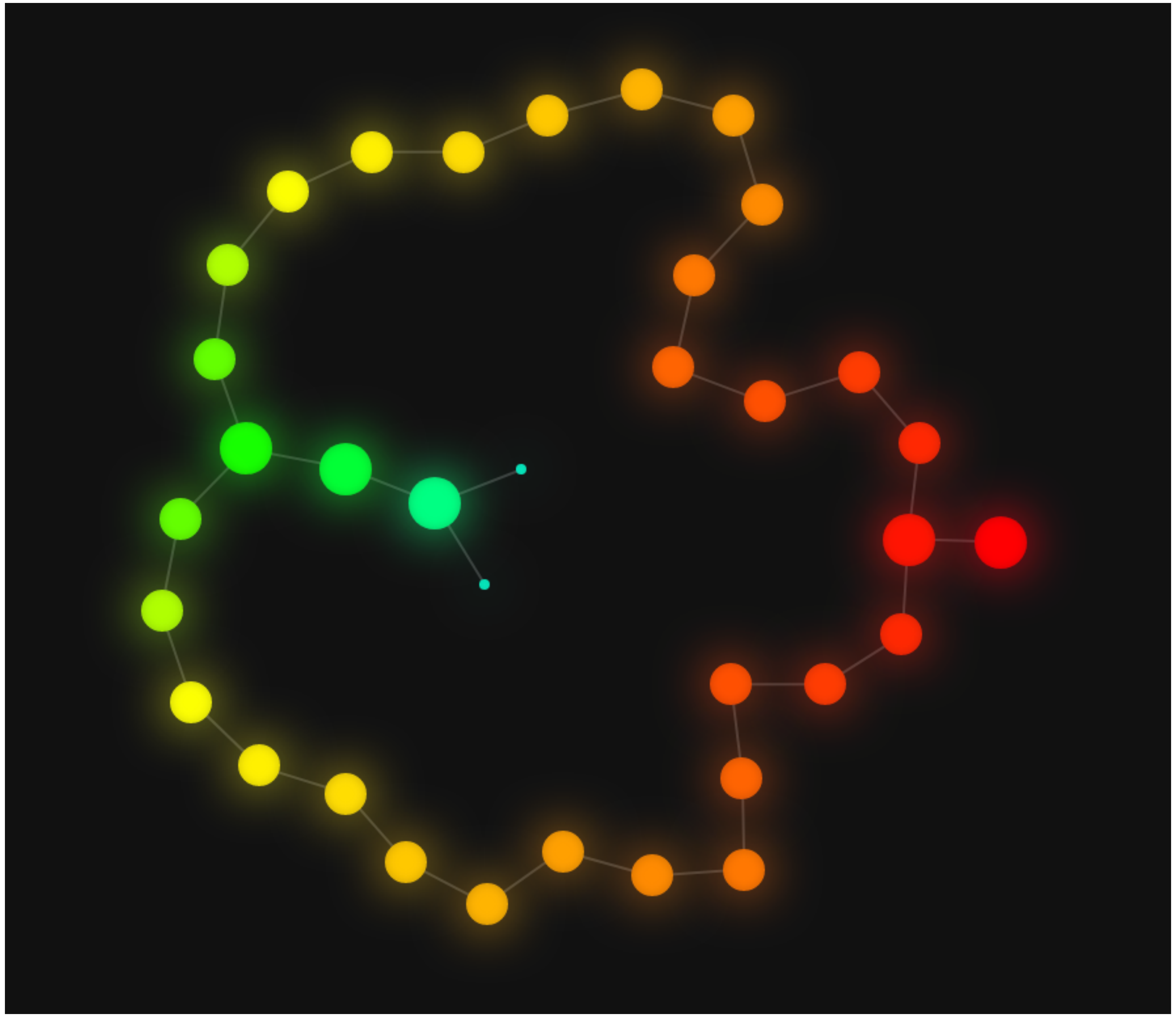
- Input:
  - Point cloud data X + distance metric on the point cloud
  - Filter functions f on X
- Output:
  - A graph or a simplicial complex representation
- Parameters:
  - Filter functions
  - Number of intervals
  - Amount of interval overlap
  - Color functions, etc.

[SinghMemoliCarlsson2007]

KepperMapper

A Demo

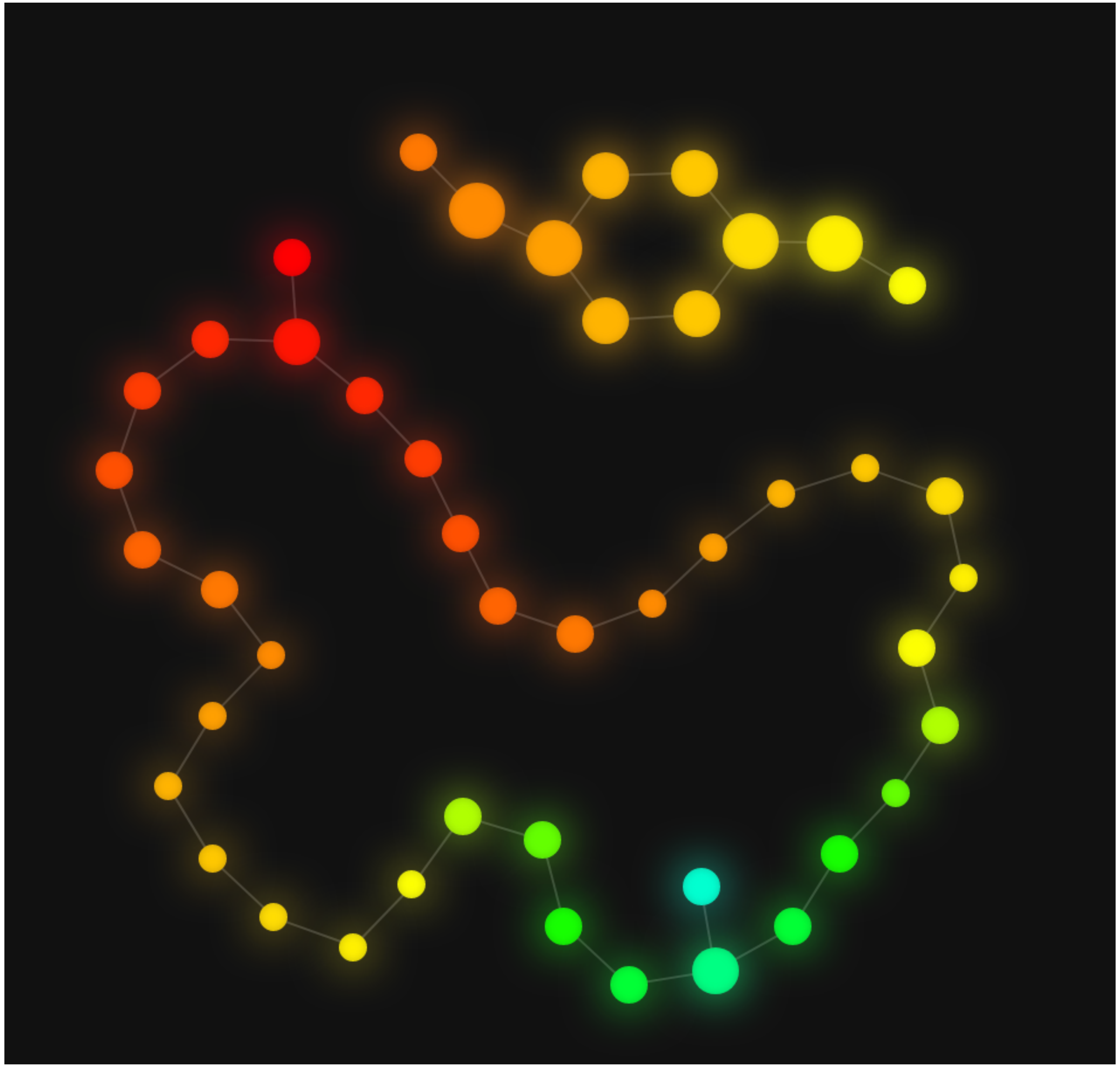
One Circle



Lens: x-values

circle-demo.py

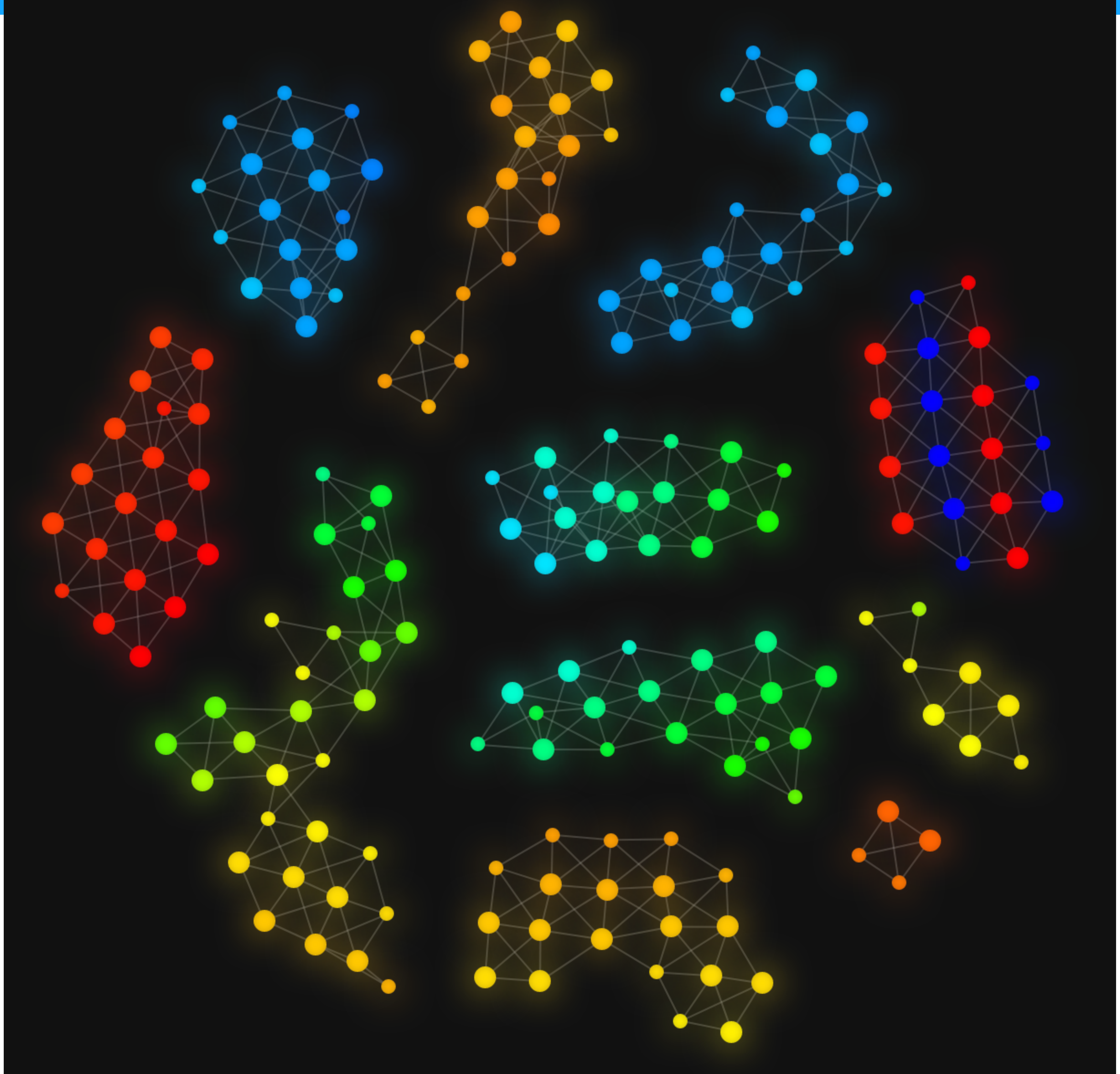
Two Circles



Lens: x-values

Color function: labels

double_circle_demo.py



Digits

digits-demo.py

Applying clustering to projected data

More discussions on DR

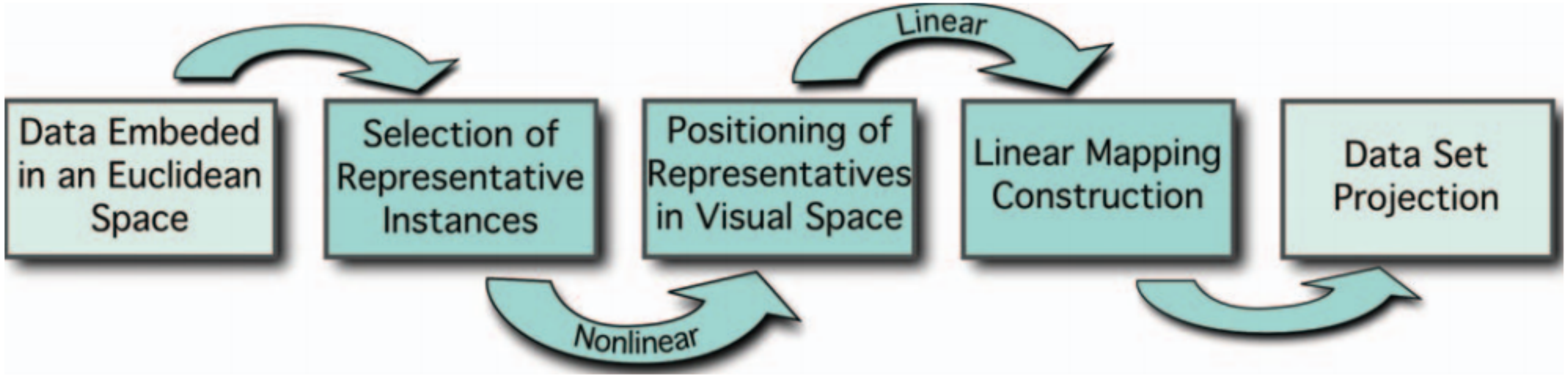
- Control point based projections
- Distance metric
- DR precision measure

Dealing with large data: control points

Control point based DR

- Improve efficiency of traditional linear/nonlinear DR
- 2 phase approach
  - Project a set of control points (anchor points)
  - Project the rest of the points based on the location of control points and preservation of local features
- Scalable system
- Allow users manipulate and modify the outcome of the DR

Part Linear Multi-dim. Proj. (PLMP)



[PaulovichSilvaNonato2010]

PLMP

$$\Phi = \underset{\hat{\Phi} \in \mathscr{L}_{m,p}}{\operatorname{argmin}} \left\{ \frac{1}{D} \sum_{ij} \left(d(h_i, h_j) - d(\hat{\Phi}(h_i), \hat{\Phi}(h_j)) \right)^2 \right\}$$

- Preserving distances between data instances as much as possible
- Approximate the above linear transformation using anchor points
- Sample selection: random vs clustering (cluster centers of k-means)
- If sampling rate increases, random and clustering produces similar results

[PaulovichSilvaNonato2010]

PLMP: steering projection

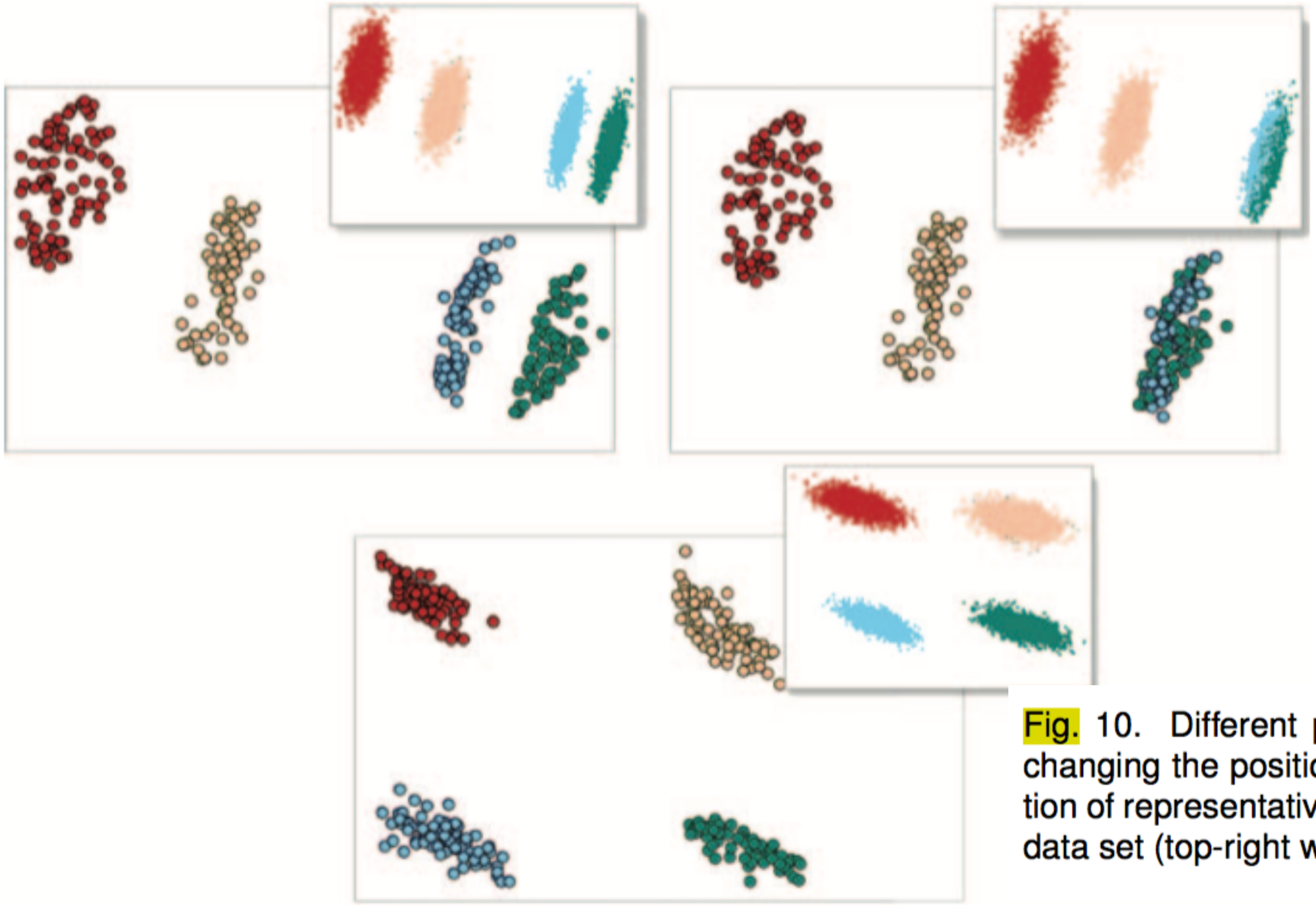


Fig. 10. Different projections of the Mammals data set produced by changing the position of representatives. Each picture shows the position of representatives (main frame) and the final projection of the whole data set (top-right window).

[PaulovichSilvaNonato2010]

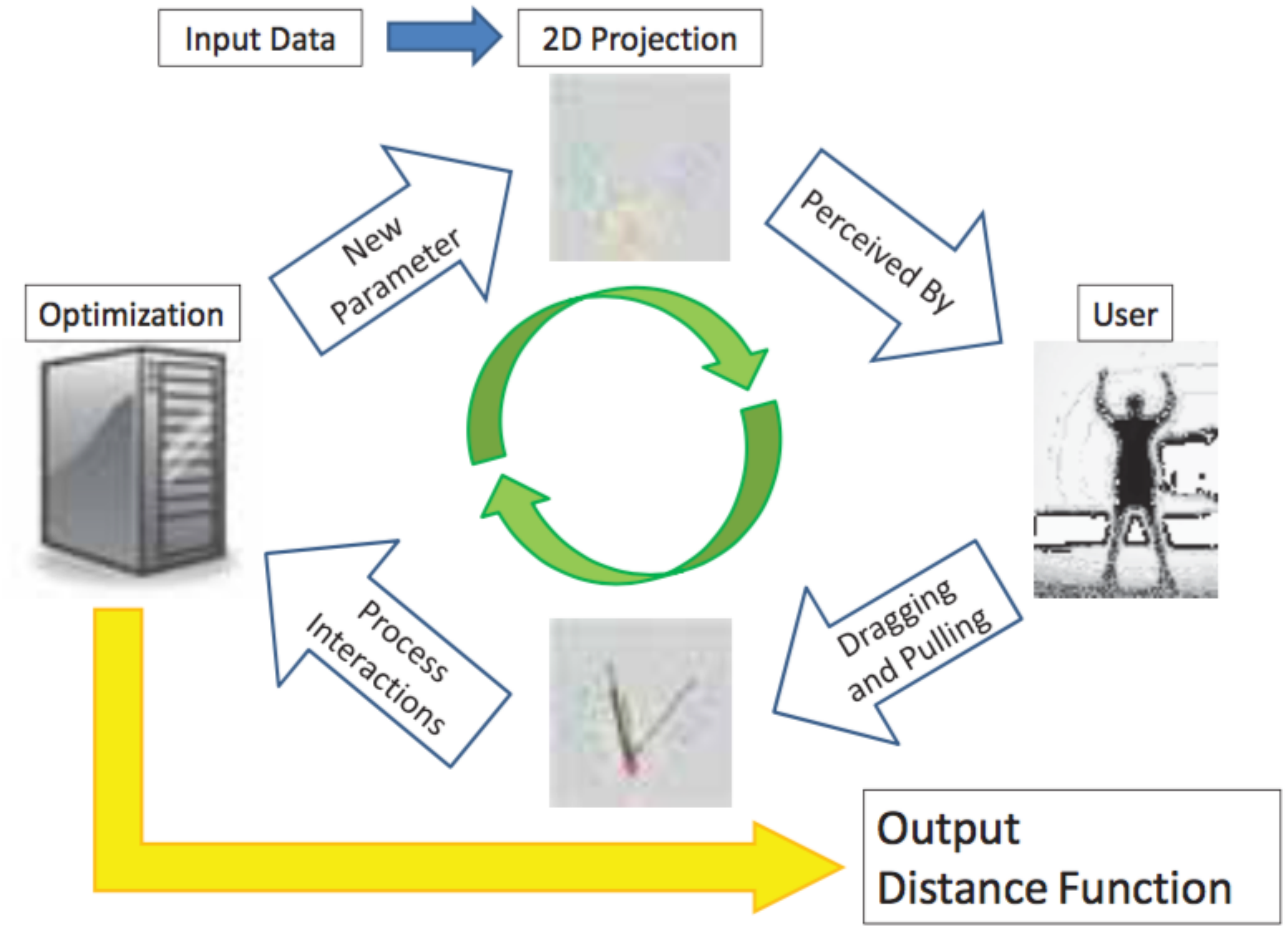
Distance metric

Learning distance interactively

- A suitable distance metric is essential for DR
- How to learn a distance function from data
- Distance function learning
  - A new distance function is calculated based on point layout manipulation by an expert user

[BrownLiuBrodley2012]

Learning distance interactively



1. 2D scatter plot visualization of the data
2. Find inconsistencies in data based on prior knowledge; drag/drop and selection to manipulate the data
3. Calculate a new distance function based on feedbacks from Step 2

[BrownLiuBrodley2012]

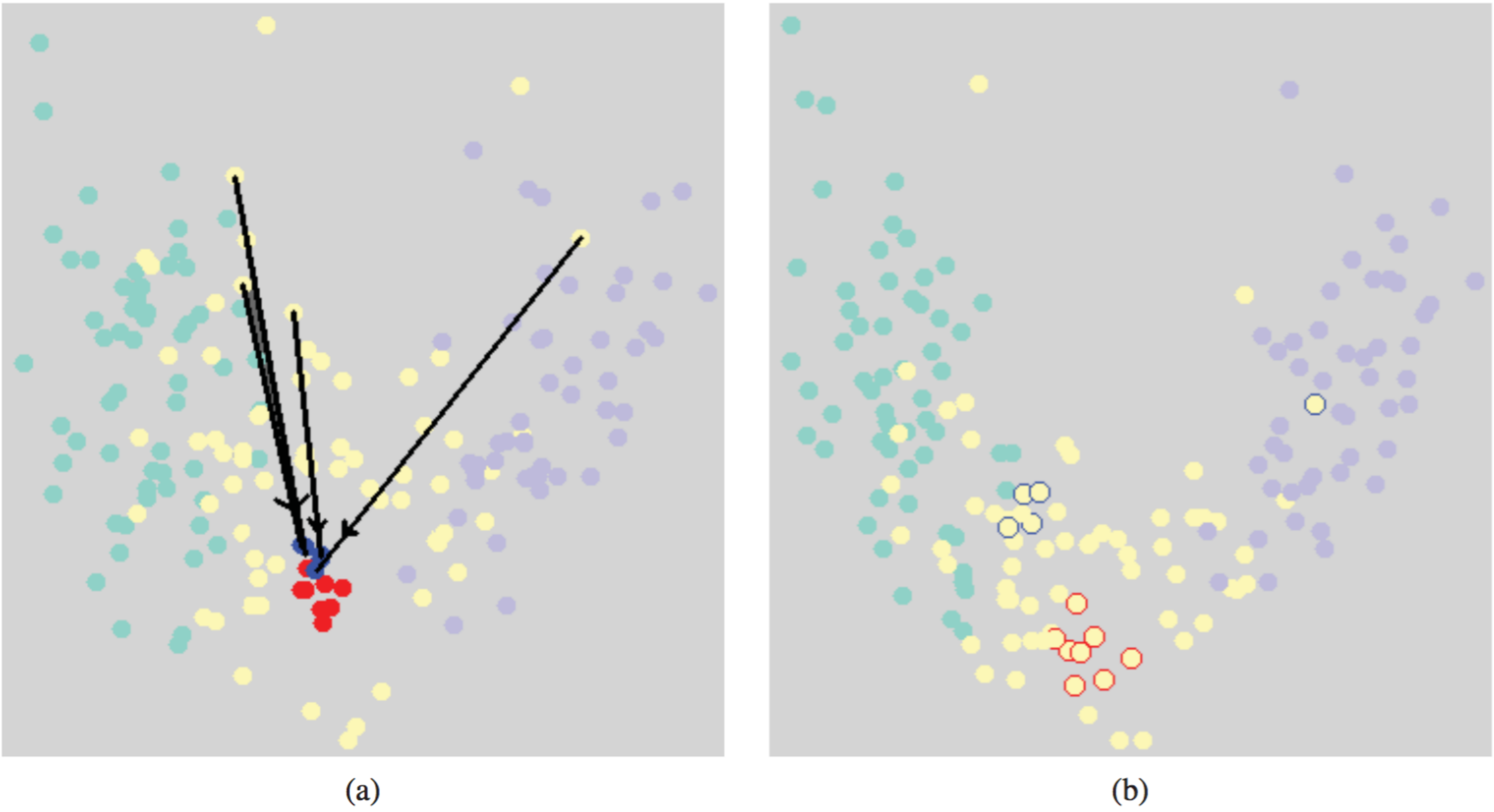


Figure 3: These images show an example of how a user manipulates the visualization. A handful of points have been marked in blue and dragged closer to another set of points, marked in red. After the update (on the right), the points in those groups are closer together, and the clustering with respect to different colors is more compact. The same red and blue points marked on the left are indicated in their new positions on the right with red and blue halos.

[BrownLiuBrodley2012]

http://www.cs.tufts.edu/~remco/publications/2012/videos/brown2012disfunction.mp4 [BrownLiuBrodley2012]

DR + Precision Measure

DR Quality Measures

- DR-dependent distortion measures
- DR-independent distortion measures

[LeeVerleysen2009] [LiuWangBremer2014]

DR-dependent distortion measures

- DR: optimizing a cost function f
- f incorporates a natural quality measure that assesses how much structure, in terms of relations among data points in high dimensions, stays consistent with the one inferred by the low-dimensional embedding
- Alternatively, how much cost is needed in transforming one to another.
- Global distortion measure: overall quality of DR, $\mathcal{E}$
- Local distortion measure: point-wise derivation of the global measure $\varepsilon : X \to \mathbb{R}$

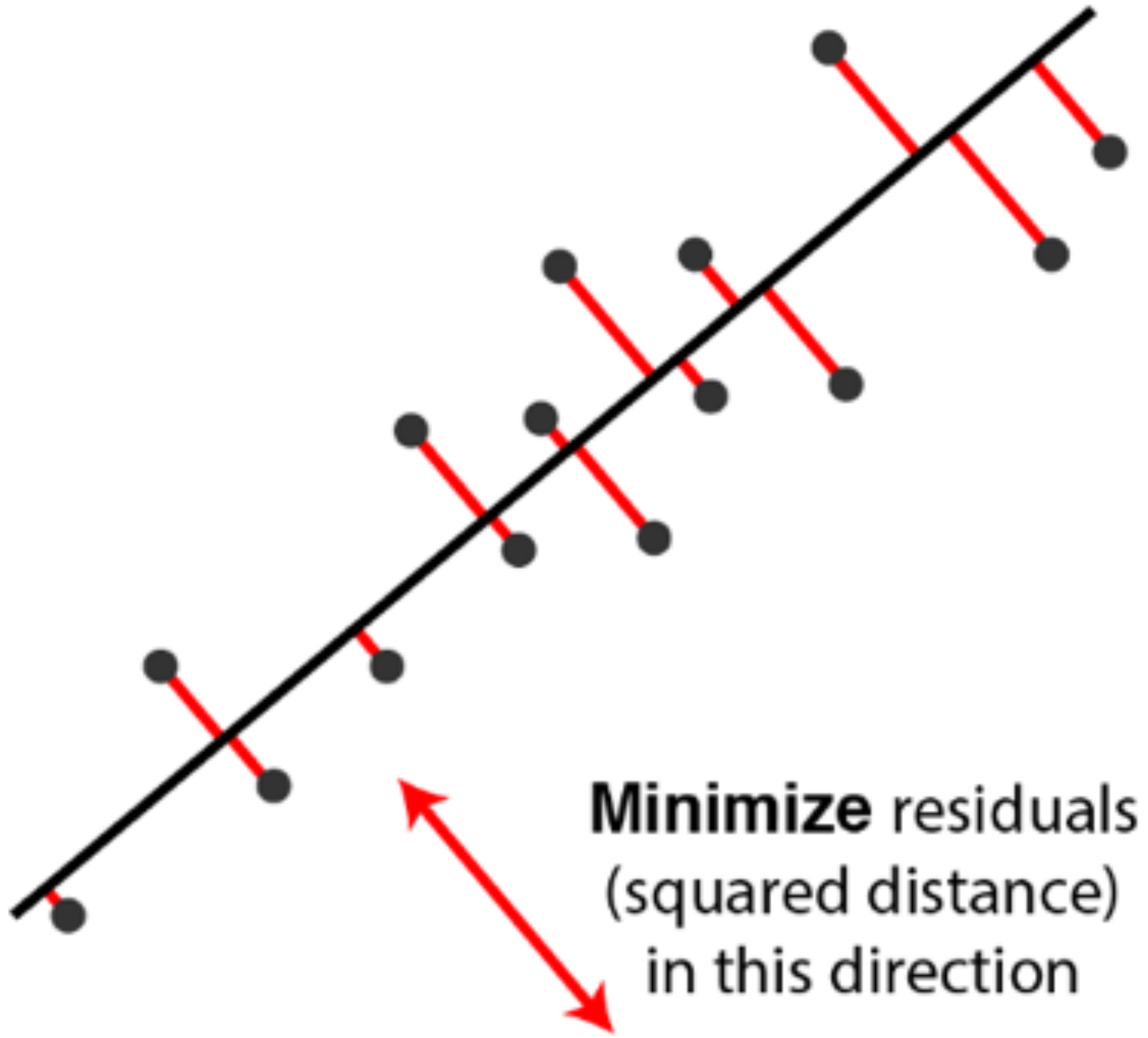
We further enforce $\mathcal{E} = \sum_i \varepsilon(x_i)$.

[LeeVerleysen2009] [LiuWangBremer2014]

PCA distortion measure

$$\mathcal{E} = \sum_i ||x_i - \mu(x_i)||^2$$

$$\varepsilon(x_i) = ||x_i - \mu(x_i)||^2.$$



[LiuWangBremer2014]

Locally linear embedding (LLE)

$$\mathcal{E} = \sum_i ||y_i - \sum_j W_{ij} y_j||^2$$

$$\varepsilon(y_i) = ||y_i - \sum_j W_{ij} y_j||^2.$$

- LLE represents each point as a weighted linear combination of its neighbors and tries to preserve this linear relationship in the reduced dimension.
- Wij: weight matrix that stores linear relationships

[LiuWangBremer2014]

DR-ind. distortion measures

- Kernel density estimate (KDE) distortion
- Stress
- Robust distance distortion
- Co-ranking distortion

[LeeVerleysen2009] [LiuWangBremer2014]

KDE distortion

Global KDE distortion $\mathcal{K} = \sum_i |KDE_X(x_i) - KDE_Y(y_i)|$

Local KDE distortion $k(x_i) = |KDE_X(x_i) - KDE_Y(y_i)|$

KDE $KDE_P(x) = \frac{1}{|P|} \sum_{p \in P} K(p, x)$

KDE w. Gaussian kernel $K(p, x) = \exp(-||p - x||^2 / 2\sigma^2)$

[LeeVerleysen2009] [LiuWangBremer2014]

Stress

Global stress $$\mathcal{S} = \frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}$$

Local stress $$s(x_i) = \frac{1}{2} \cdot \frac{\sum_{j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}$$

[LeeVerleysen2009] [LiuWangBremer2014]

Co-ranking distortion

Rank of xj w.r.t. xi $\rho_{ij} = |\{k \mid d_{ik} \leq d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$

Rank of yj w.r.t. yi $\gamma_{ij} = |\{k \mid \hat{d}_{ik} \leq \hat{d}_{ij} \text{ or } (\hat{d}_{ik} = \hat{d}_{ij} \text{ and } 1 \leq k < j \leq N)\}|$

Rank Error $R_{ij} = r_{ij} - \rho_{ij}$

Co-rank matrix: a histogram of all rank errors:

$$\mathbf{C}_{kl} = \left|\{(i,j) \middle| \rho_{ij} = k \quad \text{and} \quad r_{ij} = l\}\right|$$

dij: distance between xi and xj

[LeeVerleysen2009] [LiuWangBremer2014]

Co-ranking: global & local distortion

$$Q = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{l=1}^{K} C_{kl}$$

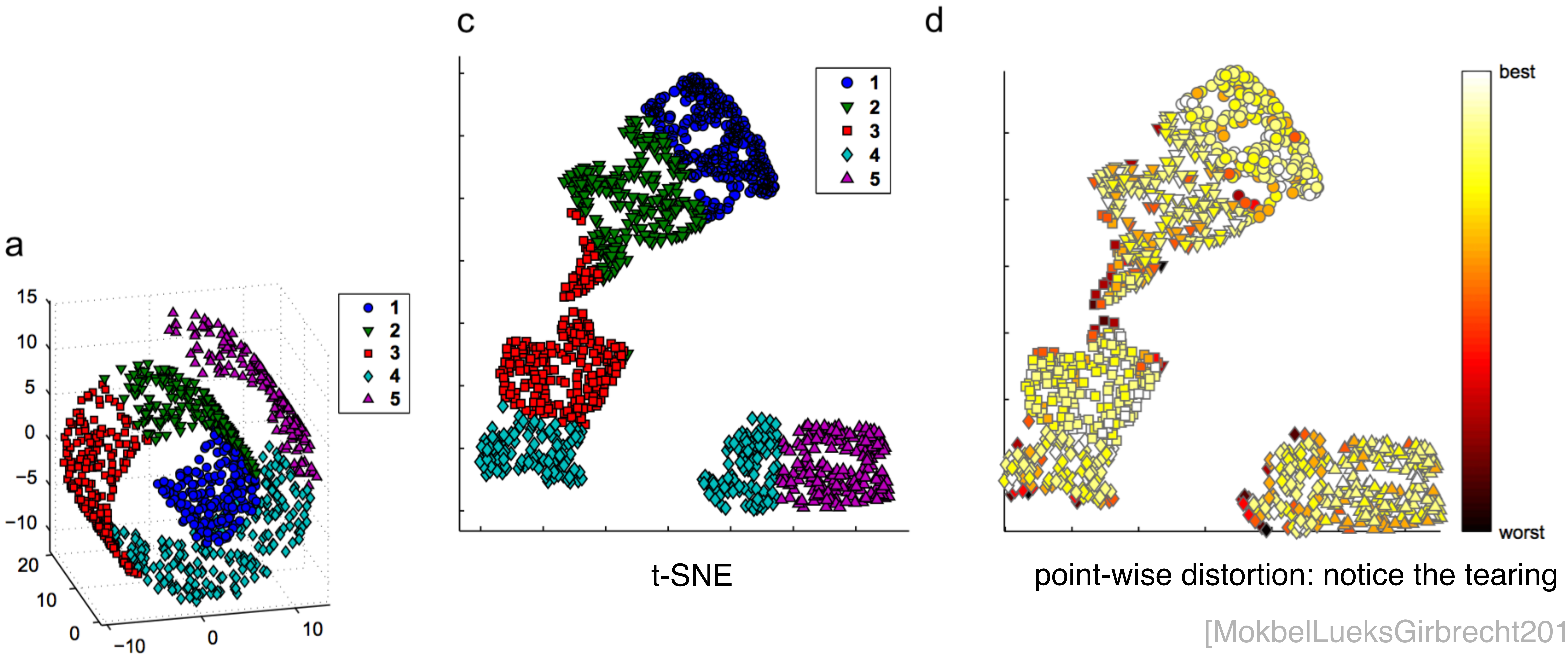
K: number of neighbors under consideration

$$C^i_{kl} = |\{j \mid \rho_{ij} = k \quad and \quad r_{ij} = l\}|$$

Point-wise contribution $$Q_i = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{K} C^i_{kl}$$ A larger Qi correspond to less distortion

[LeeVerleysen2009] [LiuWangBremer2014]

Visualizing the quality of DR



t-SNE

point-wise distortion: notice the tearing

[MokbelLueksGirbrecht2013]

Distortion-guided, structure-driven

interactive exploration of high-dim data [LiuWangBremer2014]

Distortion-Guided Structure-Driven
Interactive Data Exploration

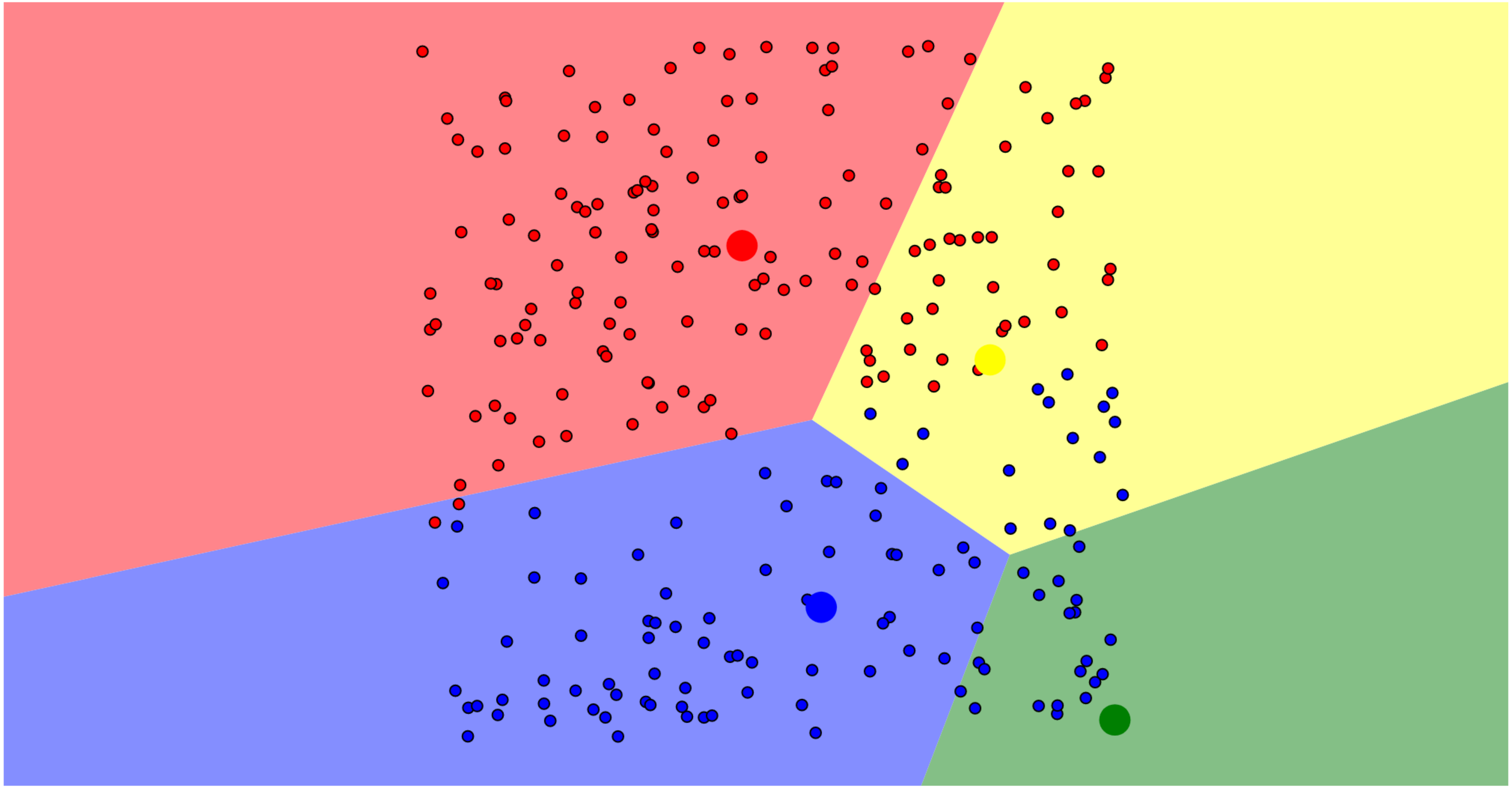
Shusen Liu, Bei Wang, Peer-Timo Bremer, Valerio Pascucci

Voice-over by Dan Maljovec

Clustering & Vis

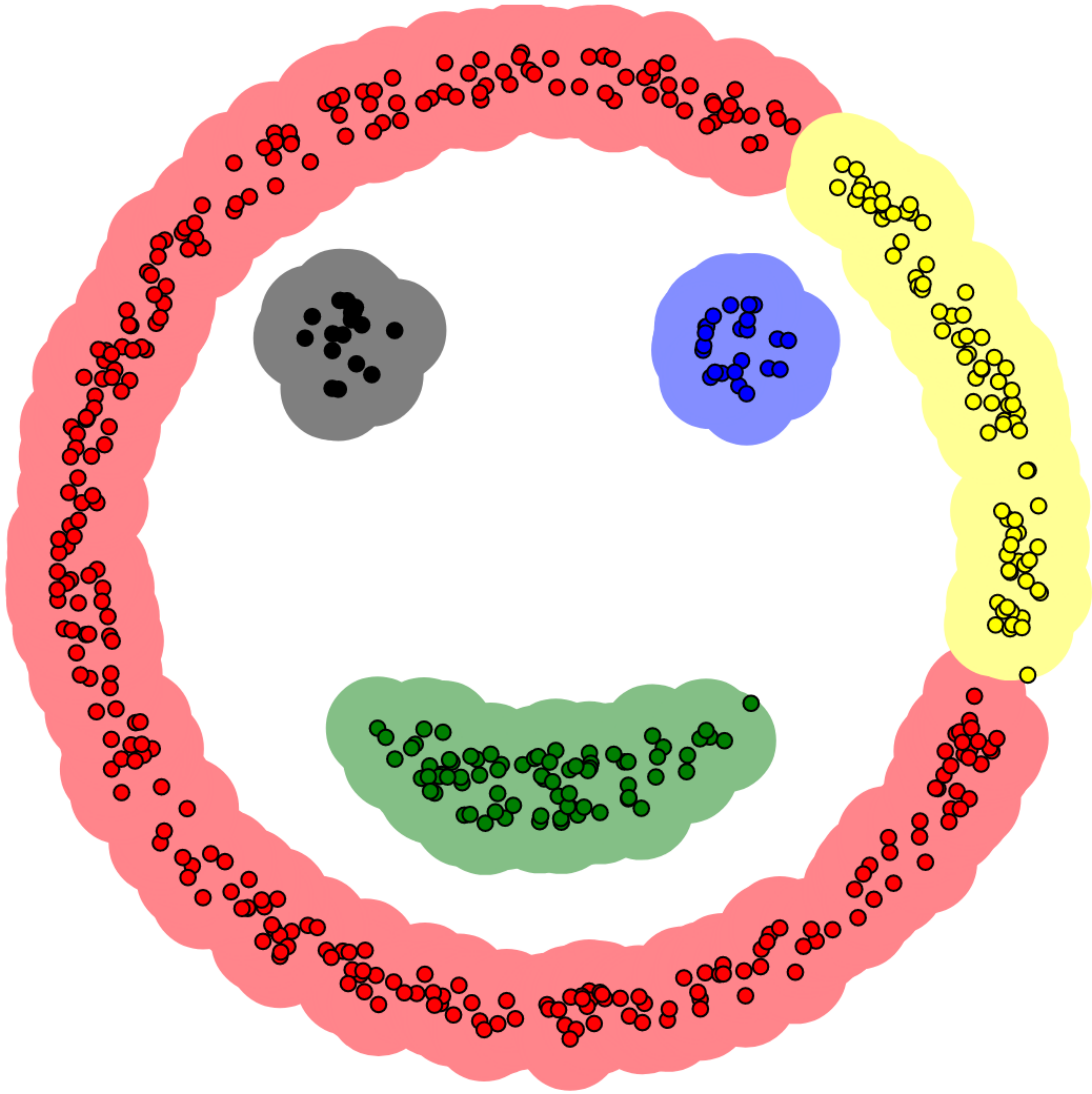
Visualizing Clustering Process

- Visualizing the algorithmic process for clustering (especially iterative ones)



https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Visualizing DBSCAN



https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Subspace clustering (SC)
& Vis

Subspace clustering vs DR

- Clustering: widely used data-driven analysis methods
- DR: compute one single embedding that best describes the structure of data
- Subspace clustering
  - Identify multiple embeddings, each capturing a different aspect of the data
  - Clustering either the dimensions or the data points

Subspace clustering & Vis

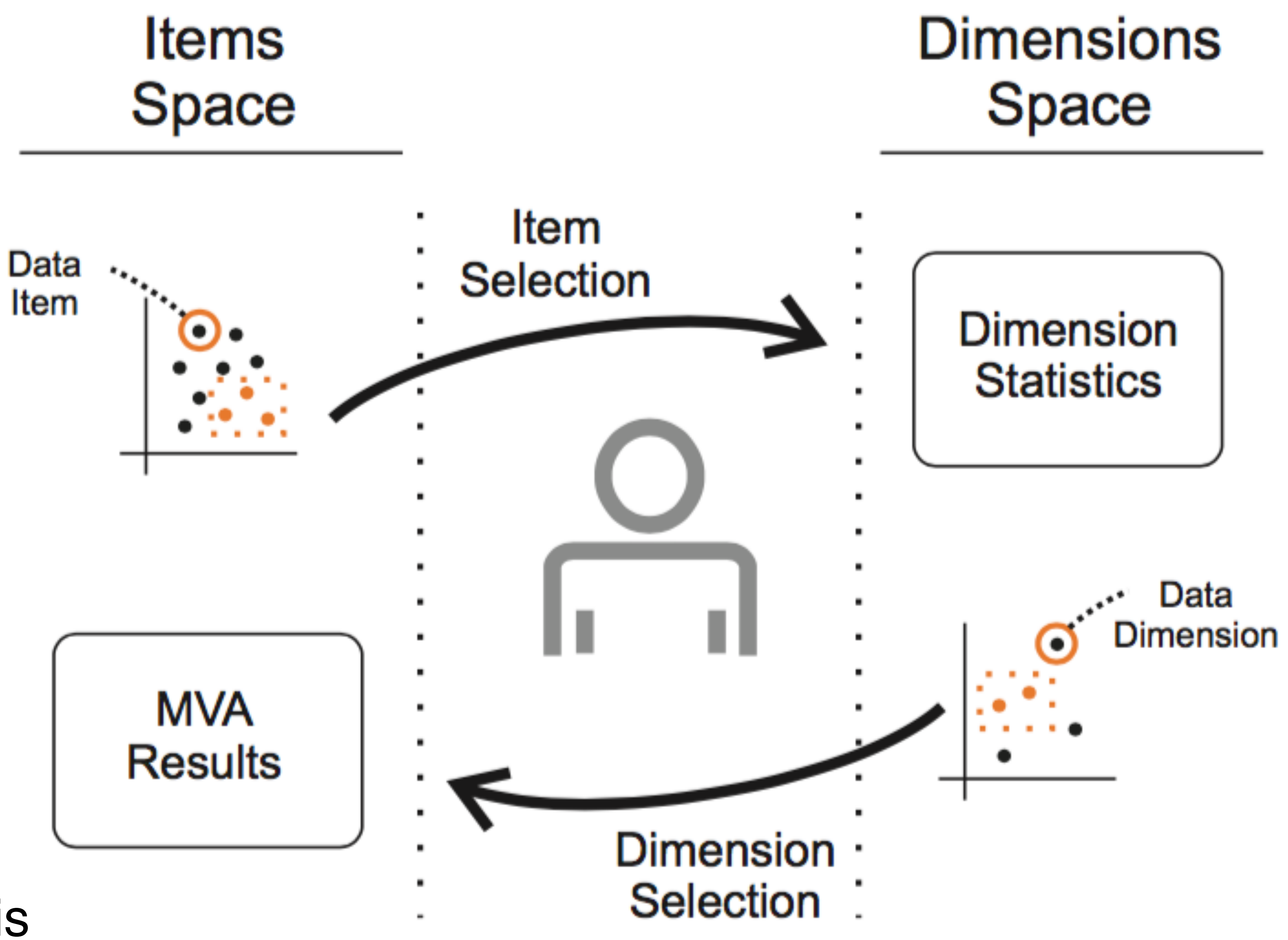
- Explore dimension space (this lecture)
- Explore subsets of dimensions (next lecture)
- Non-Axis-Aligned subspaces (next lecture)

SC: Dimension Space Exploration

Dimension space exploration

- Guided by the user
- Interactively group relevant dimensions into subsets

Dual Analysis Model



MVA: Multivariate analysis

https://fmfatore.files.wordpress.com/2012/09/pres.pdf

[TurkayFilzmoserHauser2011]

Representative factor generation

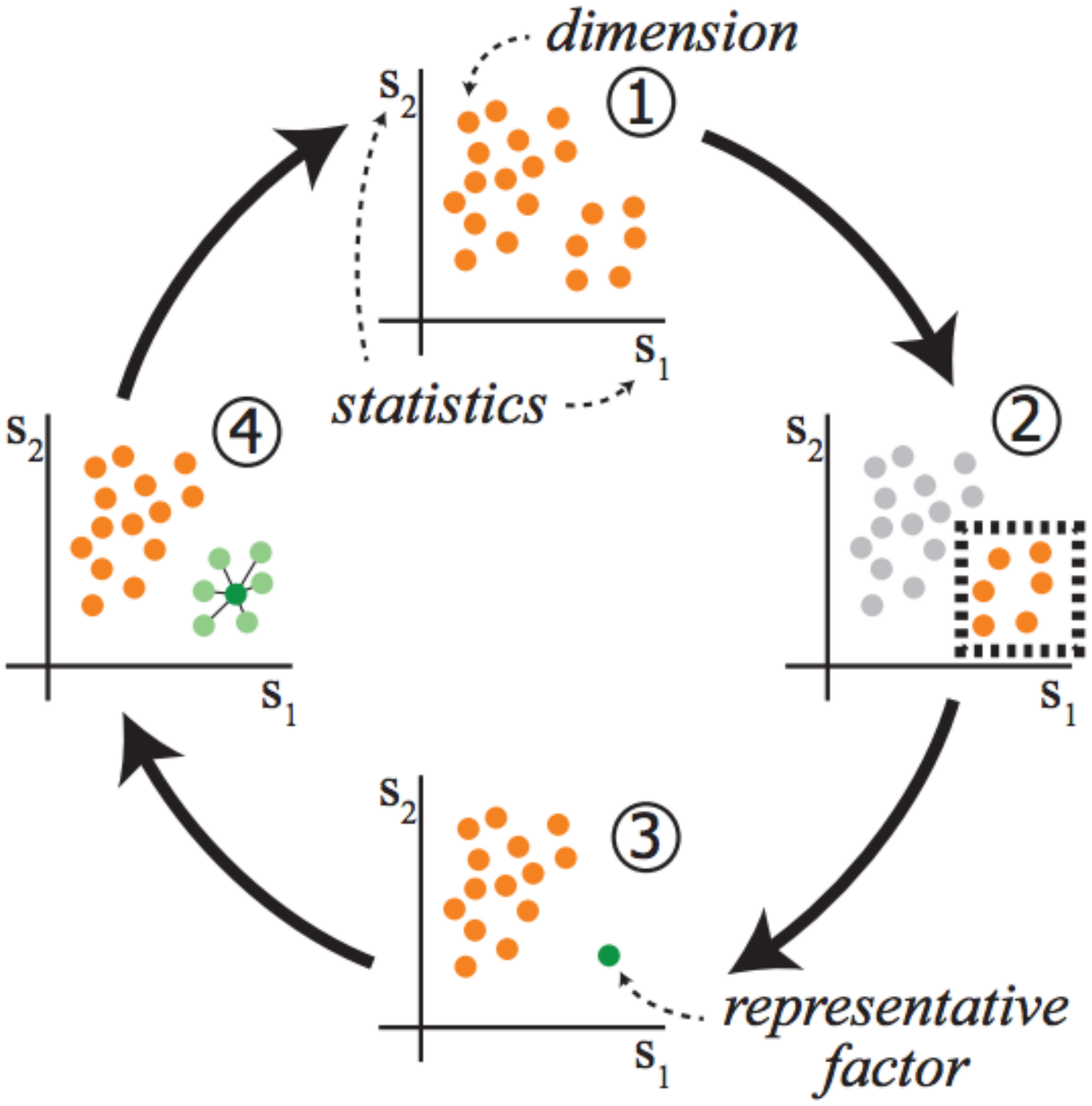


Fig. 1. An illustration of our representative factor generation method. Two statistics s_1 and s_2 are computed for all the dimensions and dimensions are plotted against these two values (1). This view reveals a group that shares similar values of s_1 and s_2 (2) and this group is selected to be represented by a factor. We generate a representative factor for this group and compute the s_1 and s_2 values for the factor (3). We observe the relation of the factor to the represented dimensions and the other dimensions (4). The analysis continues iteratively to refine and compare other structures in the data.

Grouping a collection of dimensions as a factor

[TurkayLundervoldLundervold2012]

Dimension projection matrix/tree

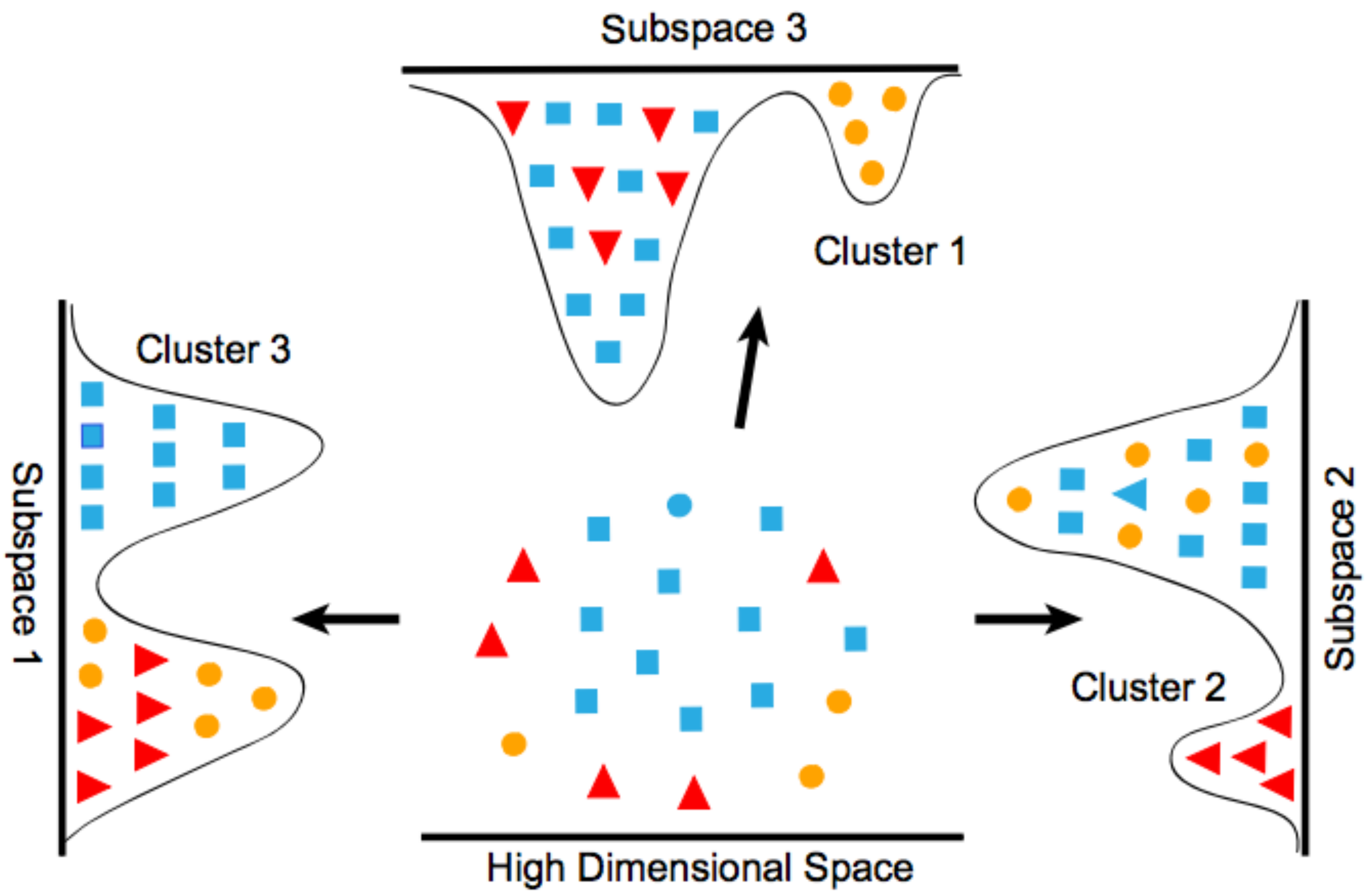
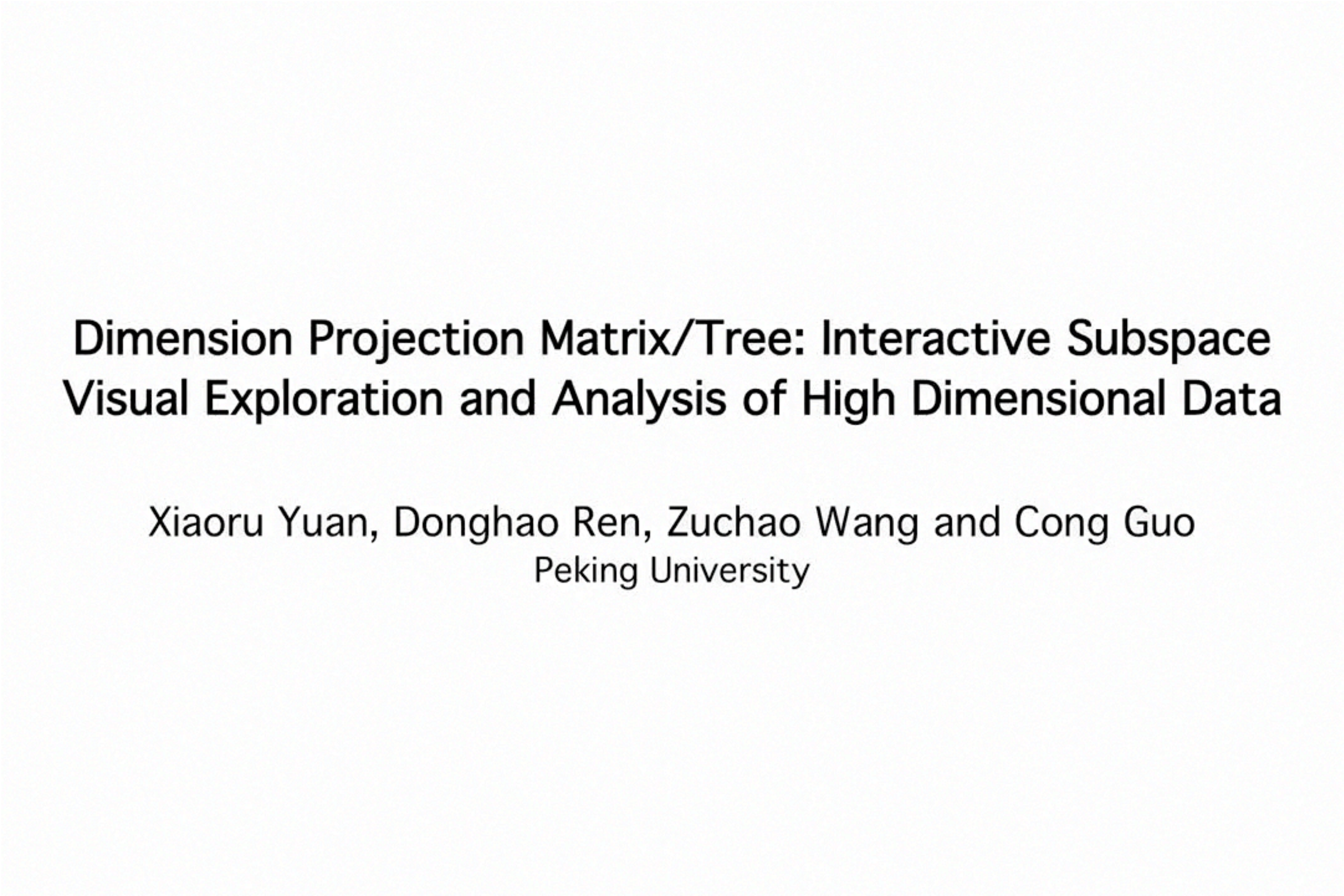


Fig. 2. Illustration of clustering in subspaces. Separation of clusters in appropriate selection of dimension subspaces can be much easier than that in the original high dimensional space.

[YuanRenWang2013]

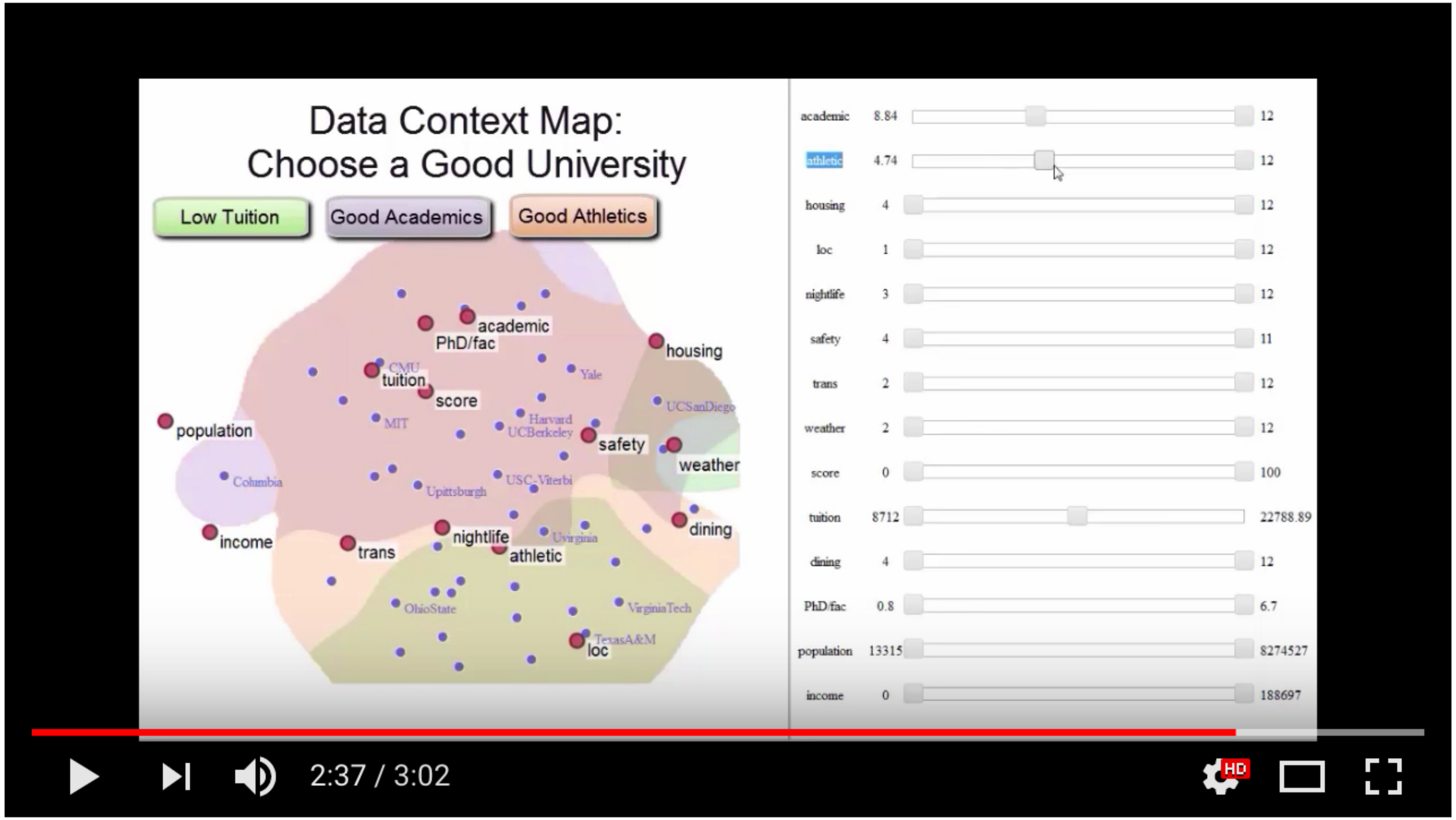
Dimension projection matrix/tree



Video: http://vis.pku.edu.cn/wiki/doku.php?id=publication:start [YuanRenWang2013]

Data Context Map

Observing the data points in the context of the attributes



Combining two similarity matrices typically used in isolation – the matrix encoding the similarity of the attributes and the matrix encoding the similarity of the data points.

https://www.youtube.com/watch?v=nnjkHA8xvbI&feature=youtu.be

http://www3.cs.stonybrook.edu/~mueller/research/pages/dataContextMap/ [ChengMuller2016]

Thanks!

Any questions?

You can find me at: beiwang@sci.utah.edu

CREDITS

Special thanks to all people who made and share these awesome resources for free:

- Presentation template designed by Slidesmash
- Photographs by unsplash.com and pexels.com
- Vector Icons by Matthew Skiles

Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:

http://www.1001fonts.com/oswald-font.html

https://www.fontsquirrel.com/fonts/open-sans

Colors used