Arranging space - II

Visualization
Sebastian Ratzenböck

Last week

- Expressing quantitative data
- Use of **spatial channel** for **categorical** data
  - 1D → List arrangement
  - 2D → Matrix or Dimensional stacking
  - Graphical perception

Recap

Principle of expressiveness

The visual encoding should express all of, and only, the information in the dataset attributes.

Effectiveness principle

The most important attributes should be encoded with the most effective channels in order to be most noticeable.

Recap

Graphical perception

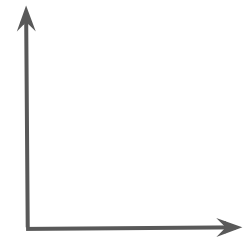
The visual decoding of information encoded in graphs.

Today

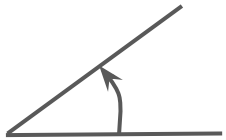
- Spatial layouts of axes
- Multivariate data visualization

Spatial layout

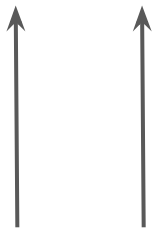
- **Rectilinear layout**: scatter plots, heatmap
- **Radial**: star plots, pie charts
- **Parallel**: parallel coordinates



Rectilinear



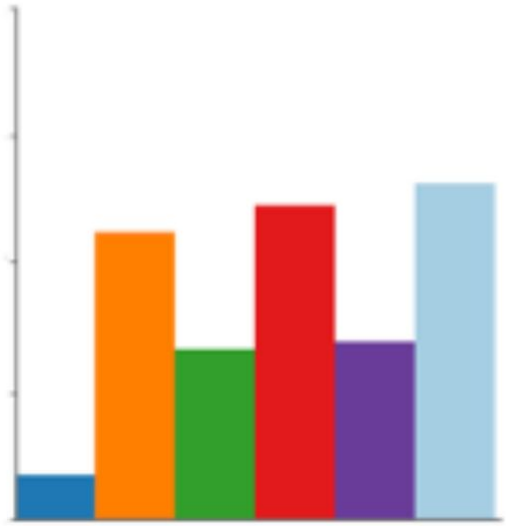
Radial



Parallel

Radial layouts

- use angular channel for dimensions
- rectilinear bar chart vs. radial star plot vs. radial multipodes plot



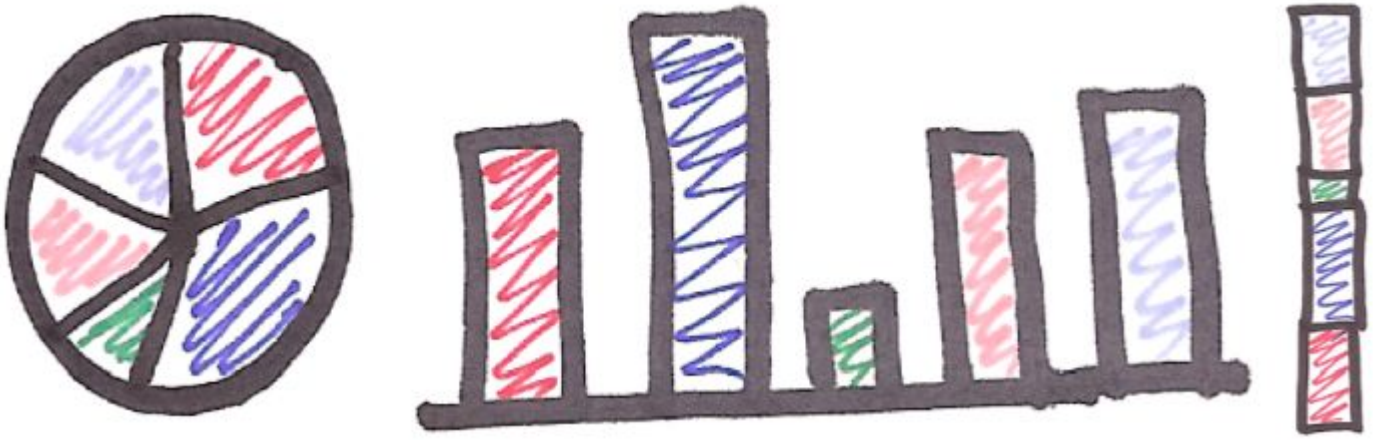
Booshehrian et al, EuroVis 2012

Radial layouts

- angular channel not as accurate as horizontal/vertical
- Used for periodic & percent data
- most common: pie chart
  - perceptually problematic
  - show relative contributions

Bar chart
M: Line
C: Length

Pie chart
M: Area
C: Angle



Radial layouts

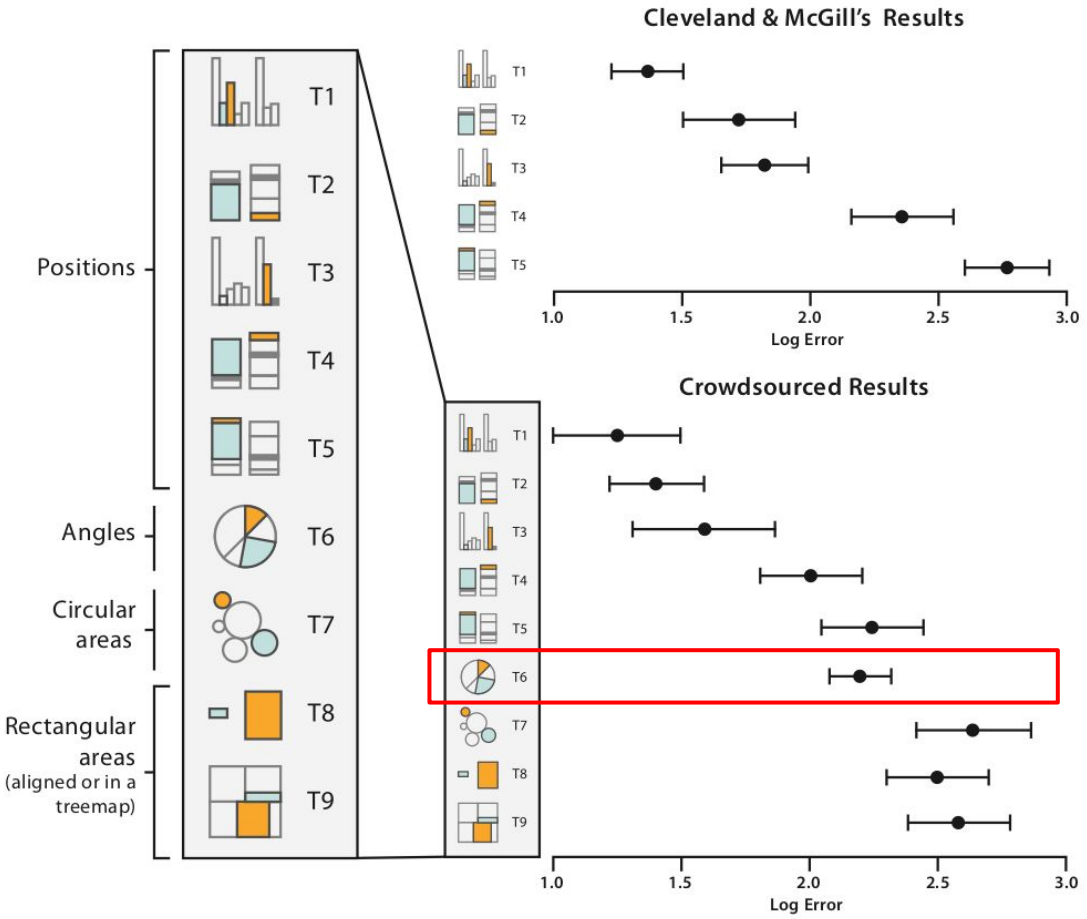


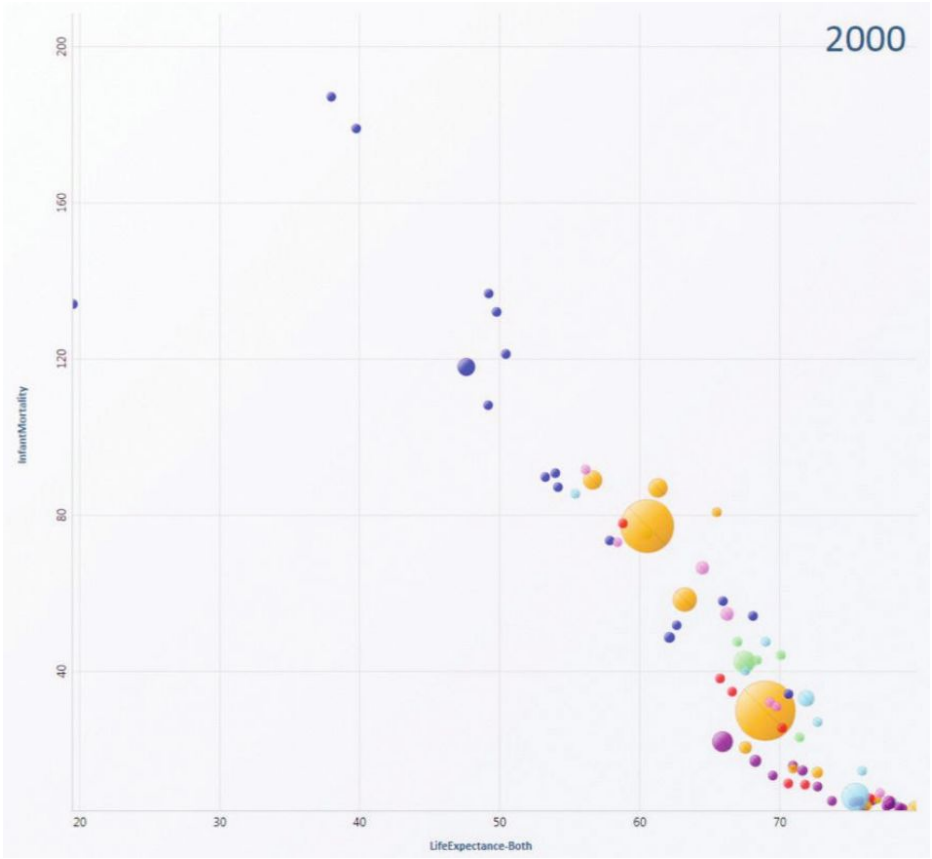
Figure 5.8. Error rates across visual channels, with recent crowdsourced results replicating and extending seminal work from Cleveland and McGill [Cleveland and McGill 84a]. After [Heer and Bostock 10, Figure 4].

Multivariate data visualization

Multivariate data visualization

Attributes

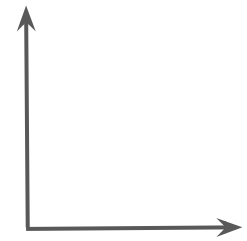
- Life expectancy
- Infant mortality
- Country size
- Continent



Want to **use effective spatial channel** for all attributes!

Spatial layout

- **Rectilinear layout**: scatter plots, heatmap
- **Radial**: star plots, pie charts
- **Parallel**: parallel coordinates



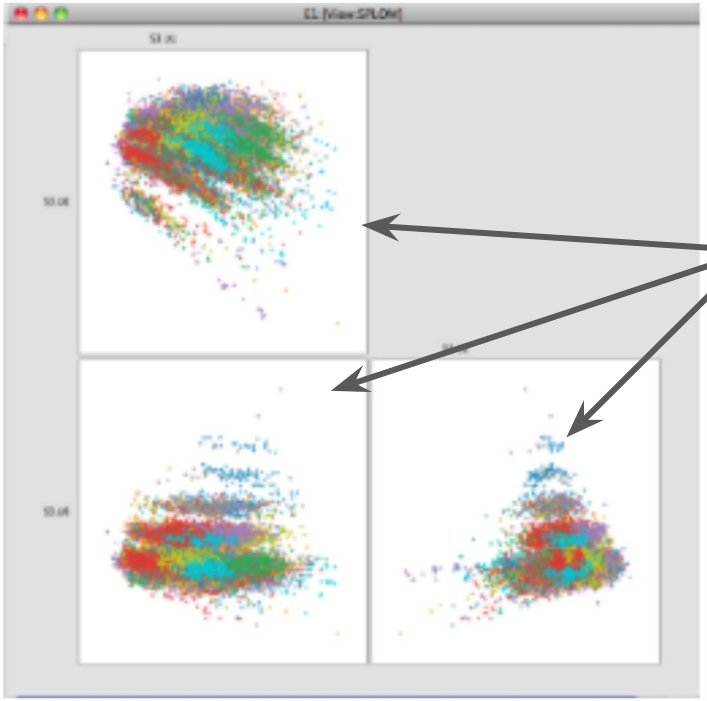
Rectilinear

Radial

Parallel

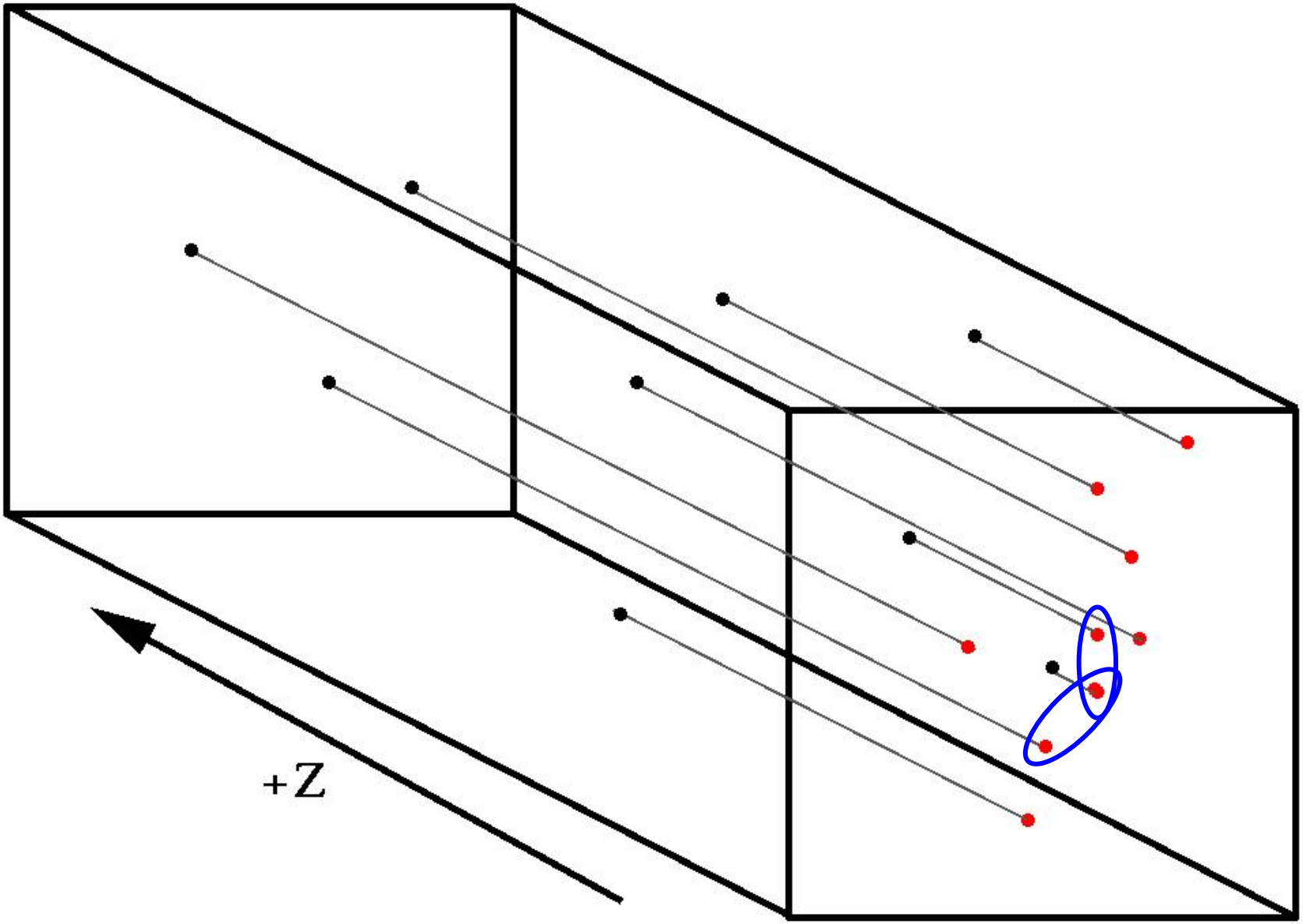
SPloMs

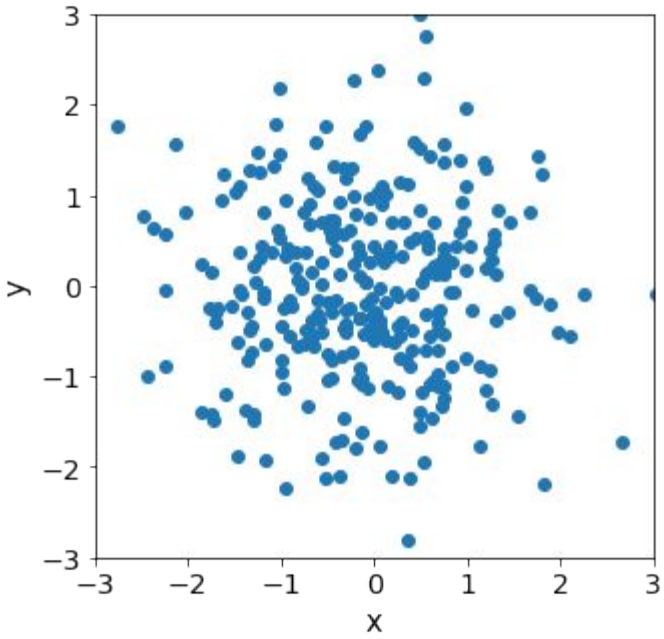
- one scatter plot: choose two dimensions
- SPloM: Scatter Plot Matrix
- No distortion but projection effects



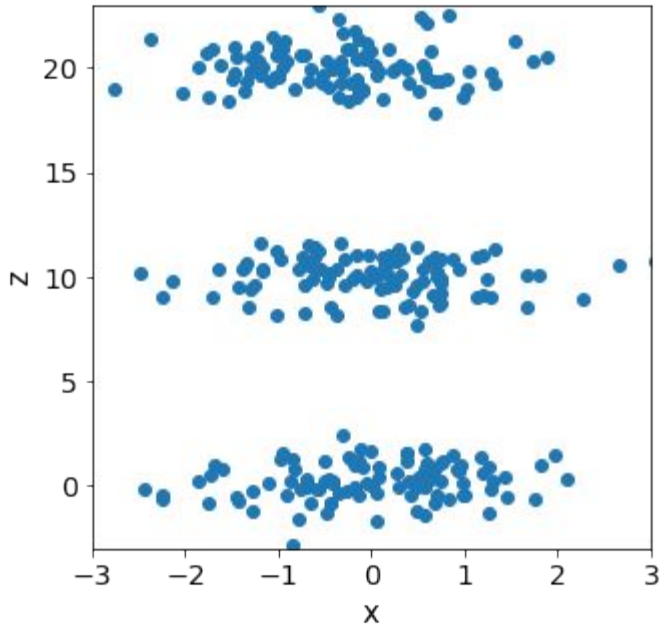
Show pairwise combinations of all features

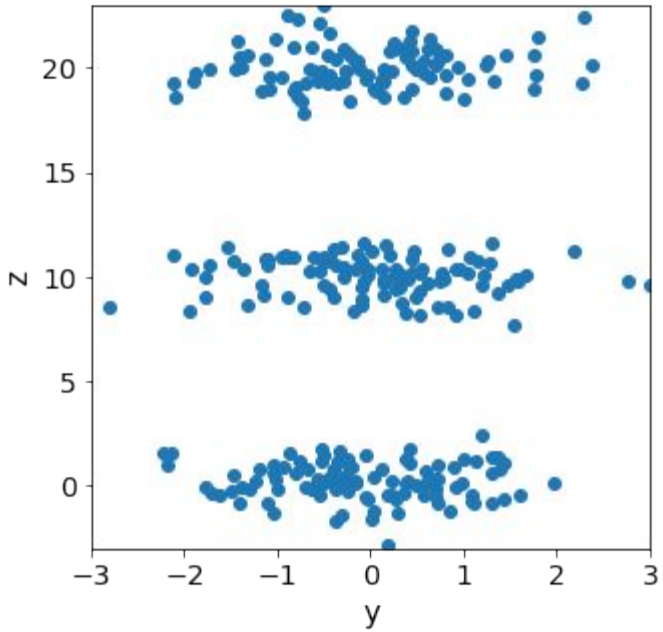
What is the “projection effect”?

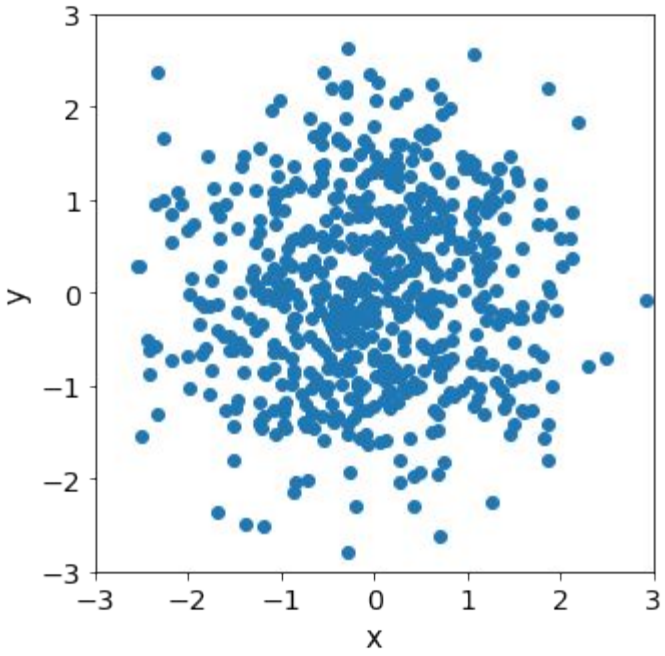

+Z


3
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3
x

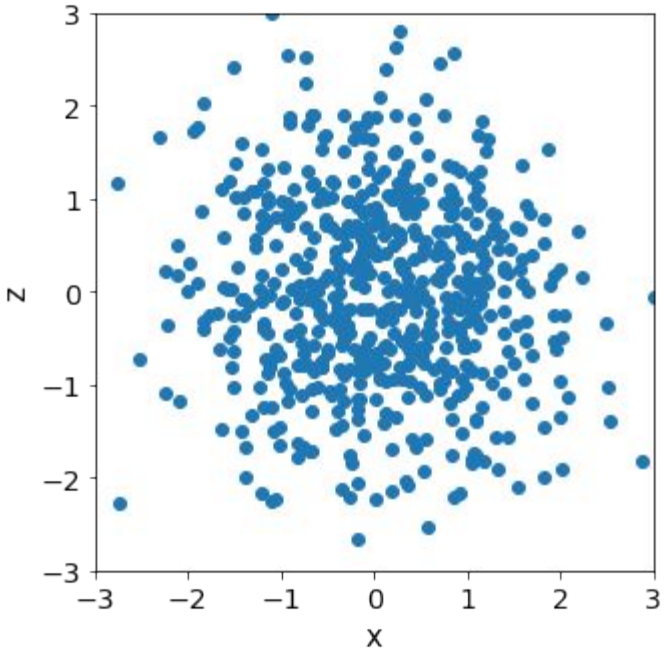
Data set 1

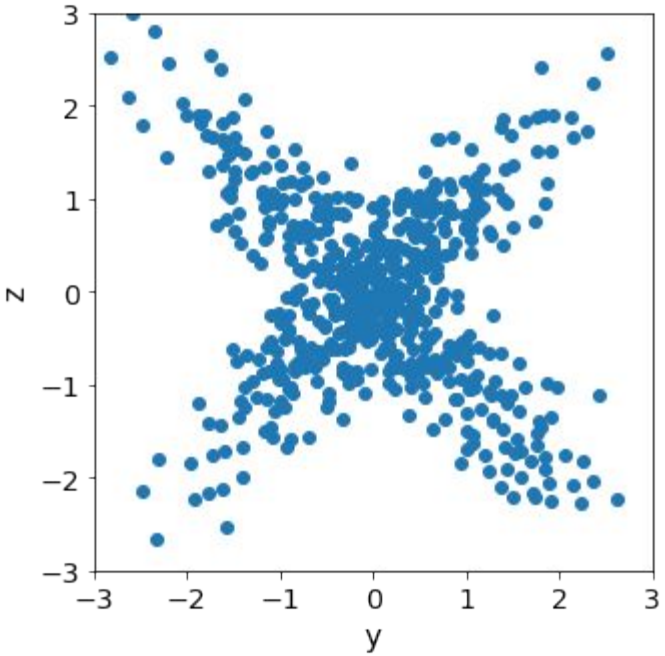

20
15
10
5
0
z
−3
−2
−1
0
1
2
3
x


20
15
10
5
0
z
−3
−2
−1
0
1
2
3
y



Data set 2

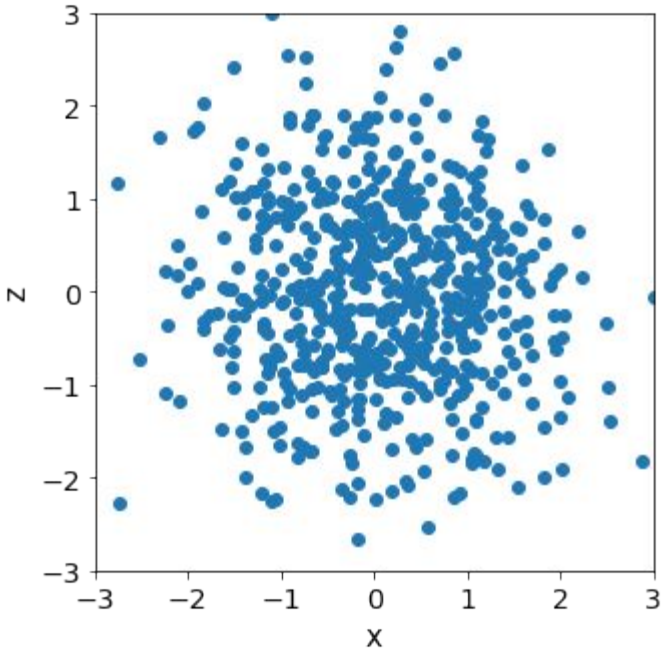


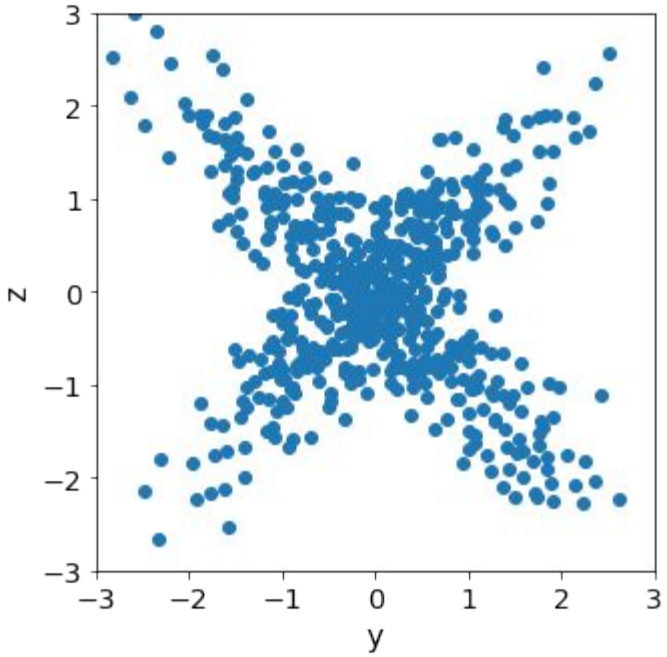


→ Need to have full view of data

BUT: large data set can limit number of combinations to show

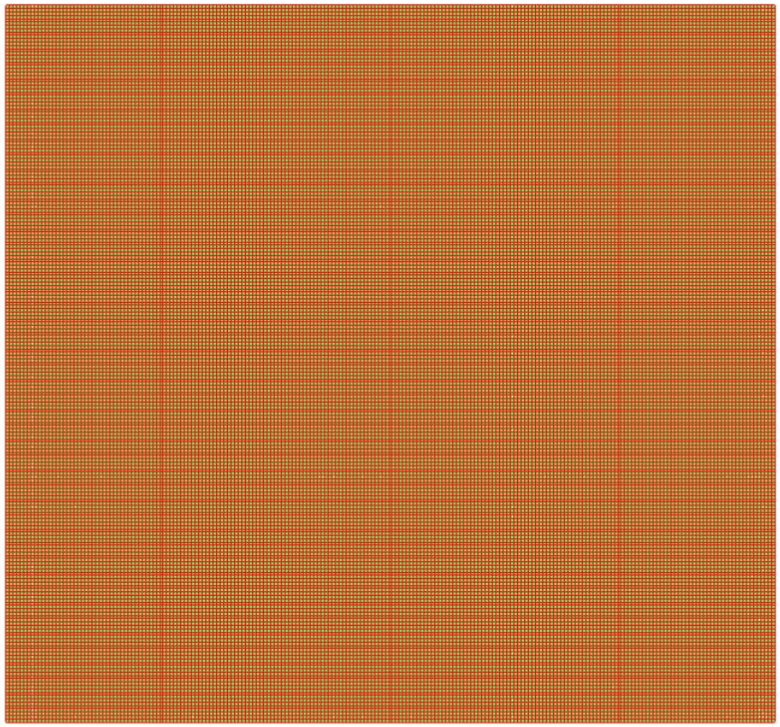
→ Filtering only systematically – not randomly





SPloM: dimension management

- Raw



(a)

Figure 4: Scatterplot Matrices. (a): OHSUMED dataset without DOSFA. Individual plots cannot be discerned without significant zooming. (b): after filtering.

[Yang et al. InfoVis 2003]

Feature selection: Problem definition

Input: Vector space with dimensions $D = \{d_1, ..., d_n\}$

Output: a minimal subspace over $D' \subseteq D$ which is optimal for a given task

Challenges

- There are d(d -1)/2 possible combinations
- Features might only be useful in combination with certain other features

SPloM: dimension management

- Raw

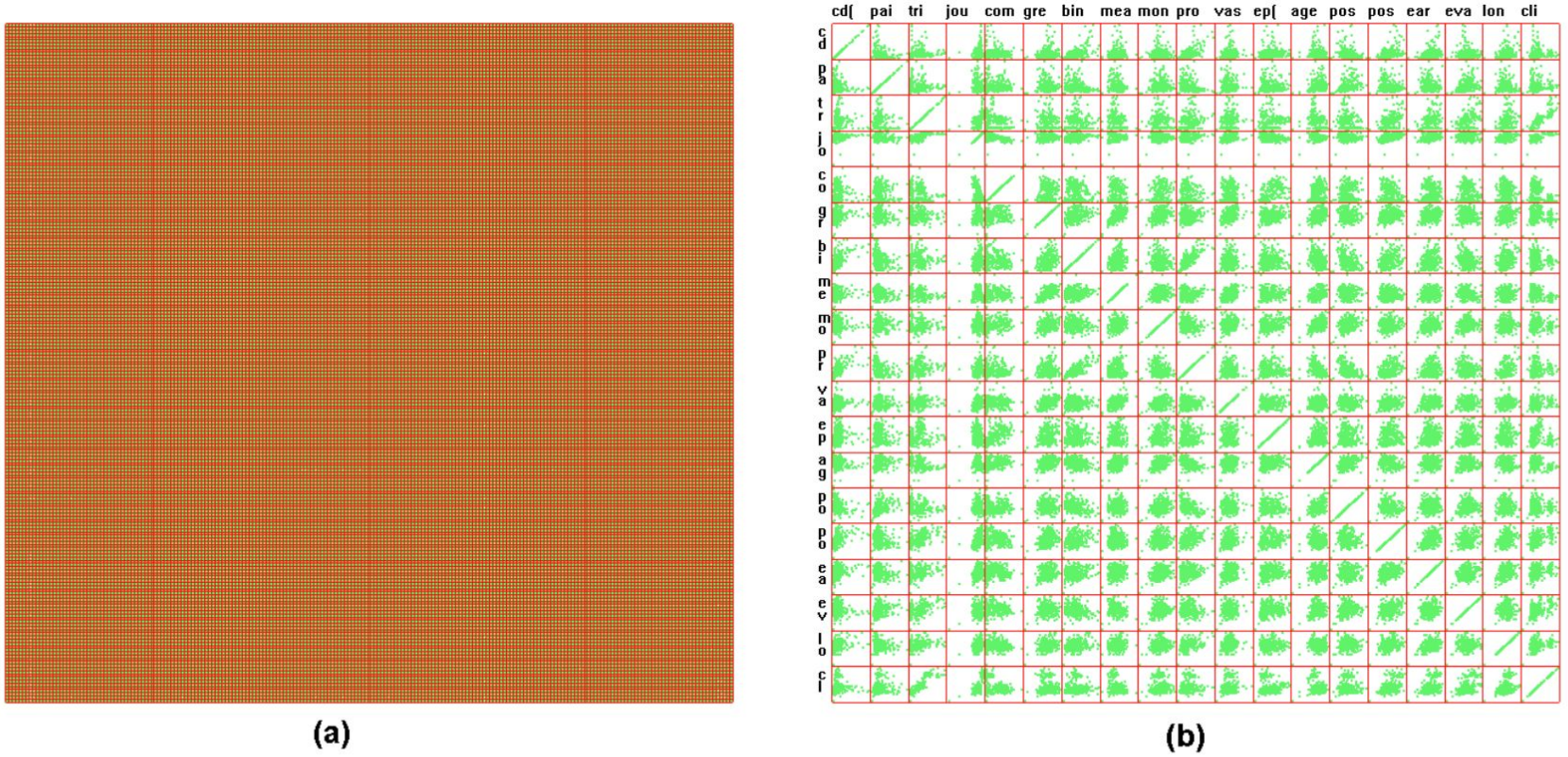


Figure 4: Scatterplot Matrices. (a): OHSUMED dataset without DOSFA. Individual plots cannot be discerned without significant zooming. (b): after filtering.

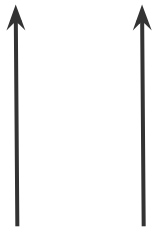
[Yang et al. InfoVis 2003]

Spatial layout

- **Rectilinear layout**: scatter plots, heatmap
- **Radial**: star plots, pie charts
- **Parallel**: parallel coordinates

Rectilinear

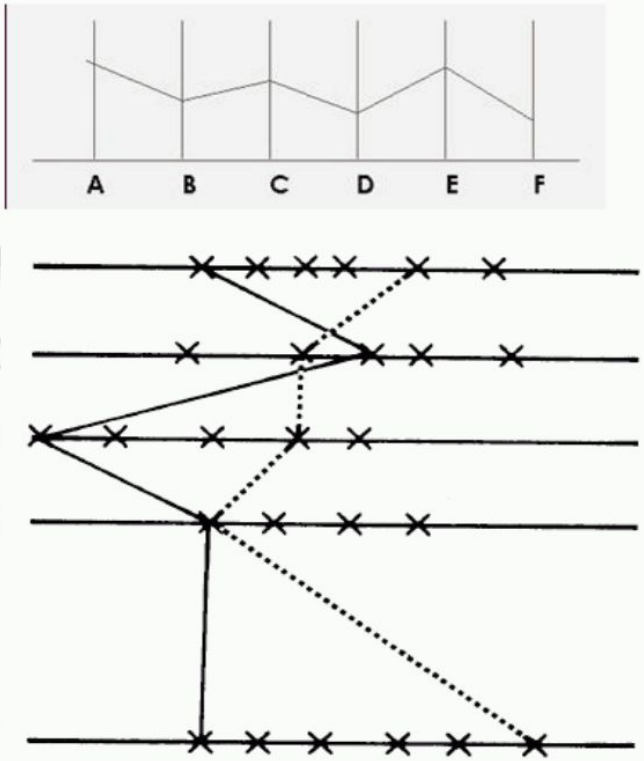
Radial



Parallel

Parallel coordinates

- only 2 orthogonal axes in the plane
- instead, use parallel axes!



[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, Vol. 85, No. 411. (Sep., 1990), pp. 664-675.]

Parallel correlation

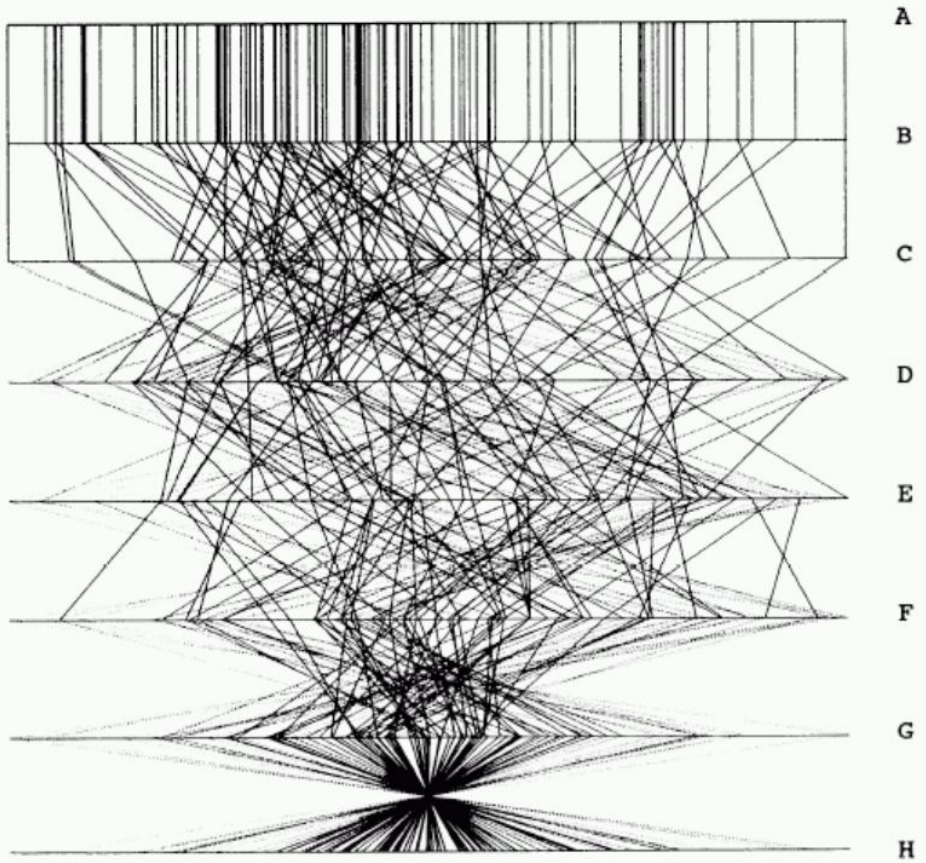
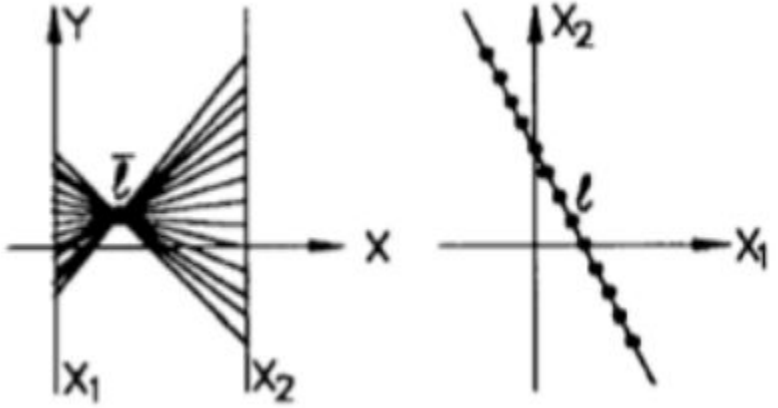


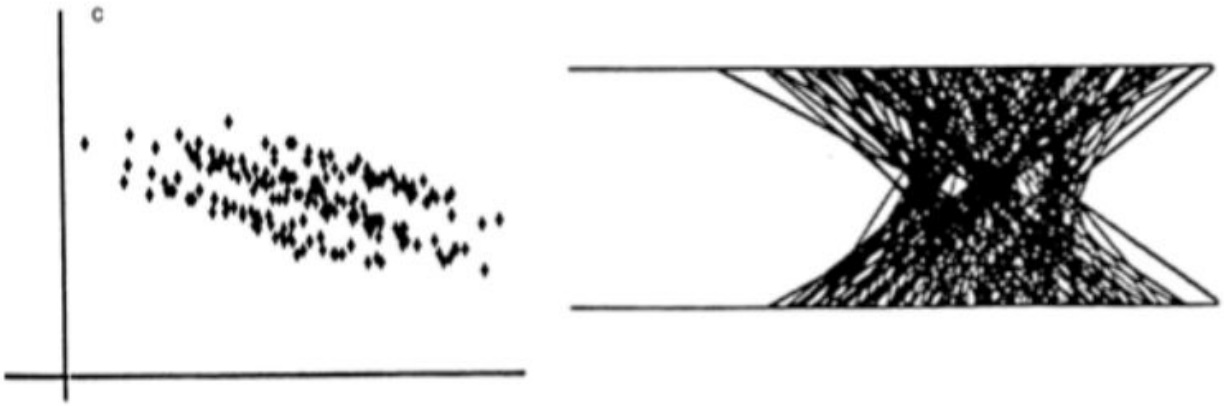
Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $\rho = 1, .8, .2, 0, -.2, -.8,$ and -1.

[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, Vol. 85, No. 411. (Sep., 1990), pp. 664-675.] 4

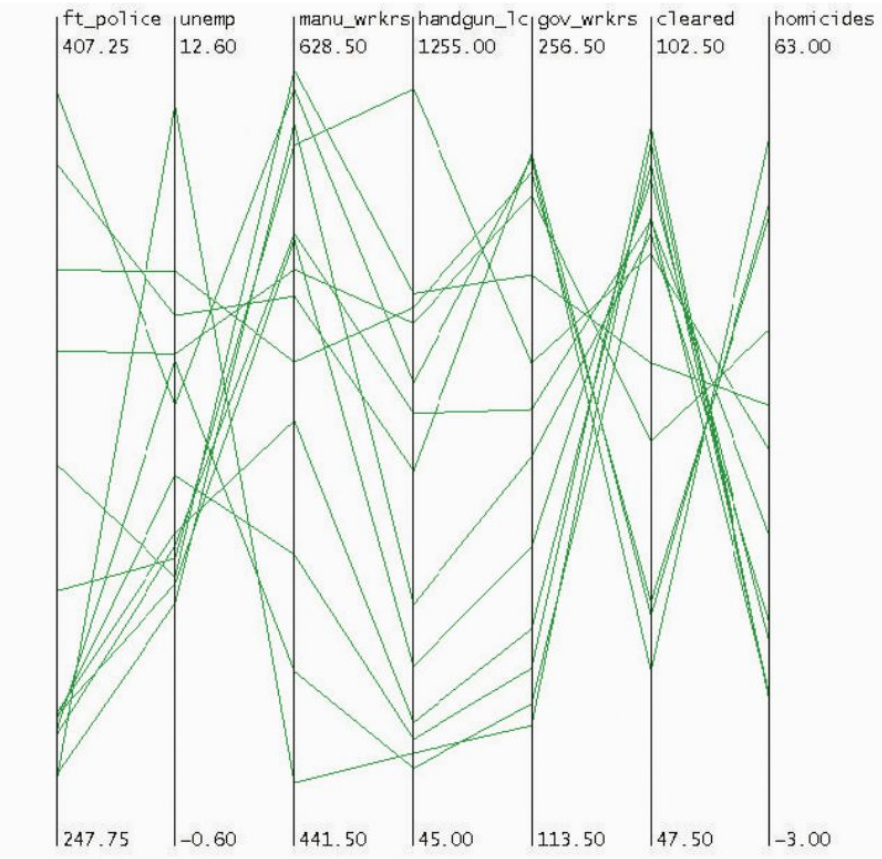
PC: Duality

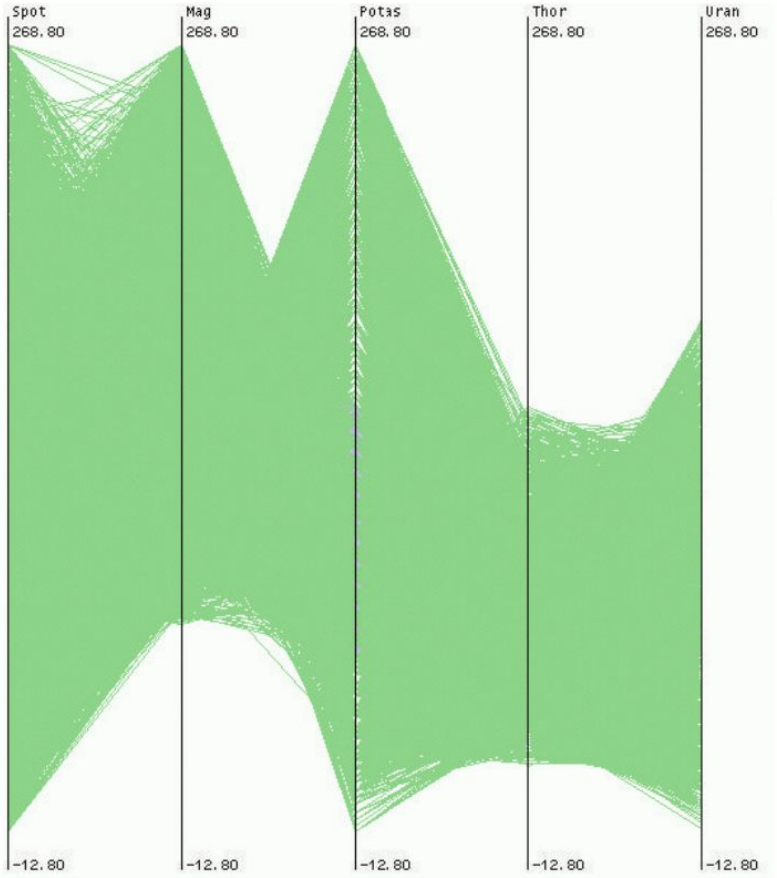
- Rotate-translate
- point-line



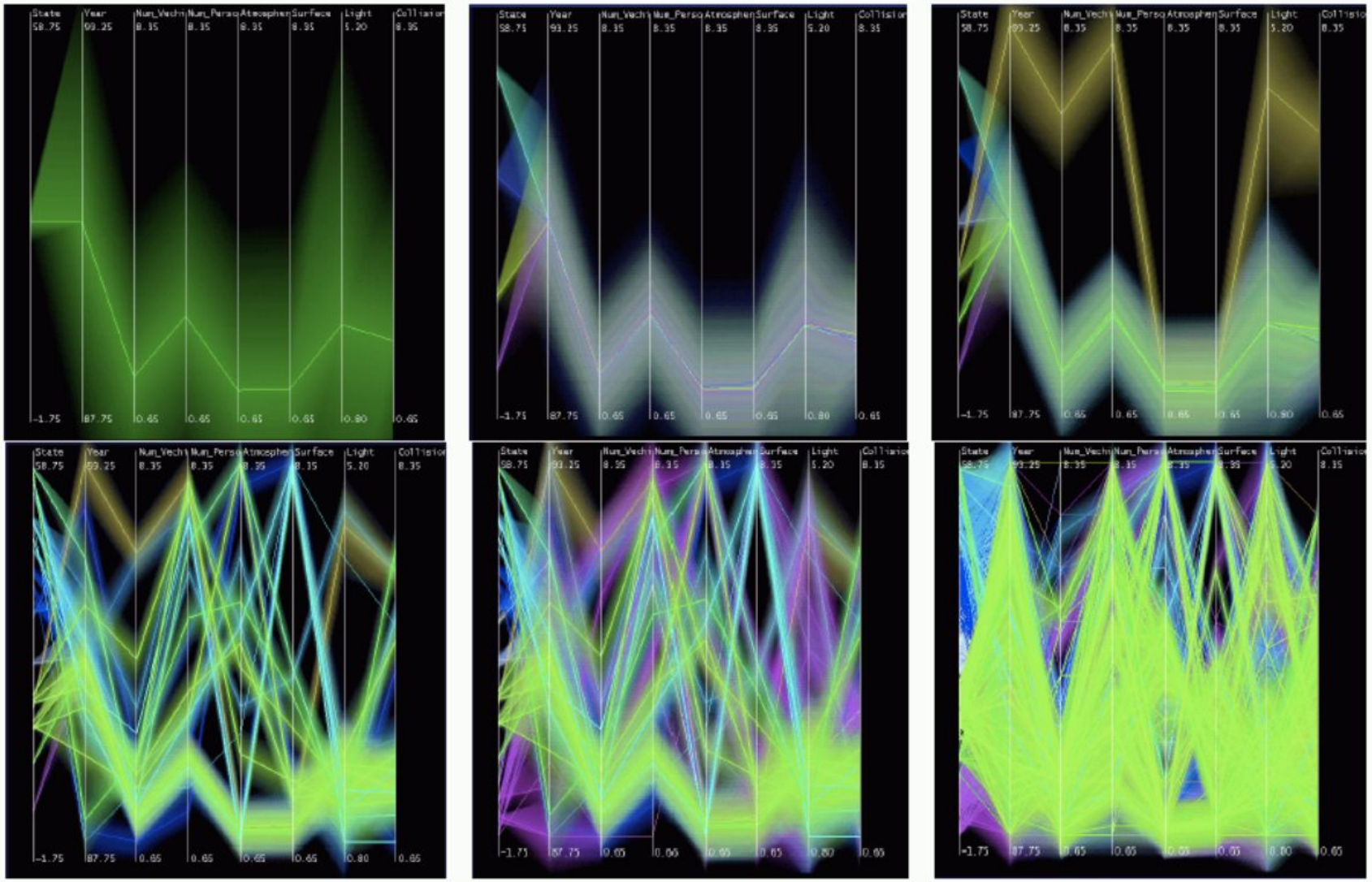


PCs and large data





Hierarchical PC

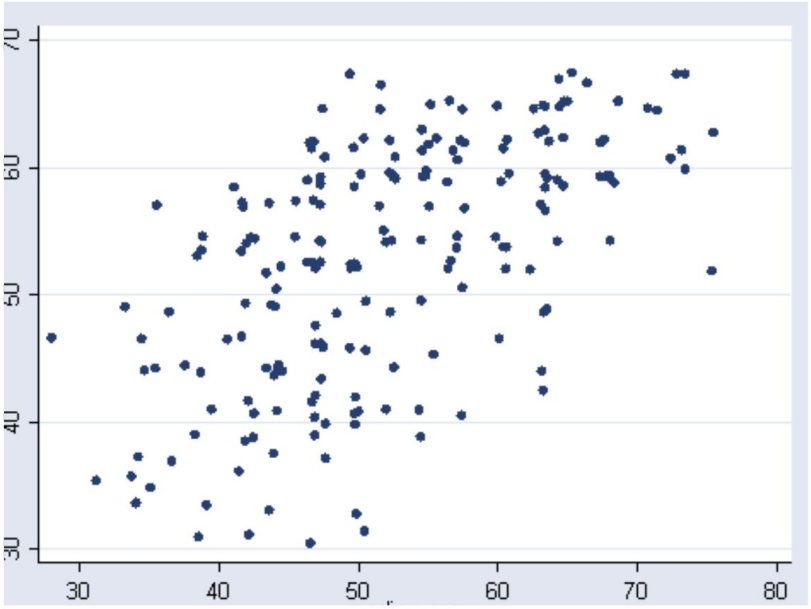


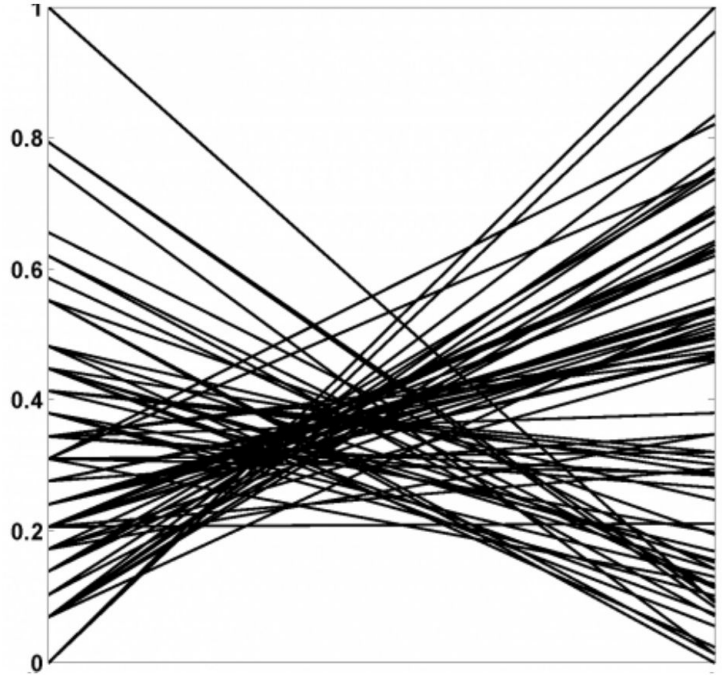
[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, IEEE Visualization '99.]

Now let’s compare!

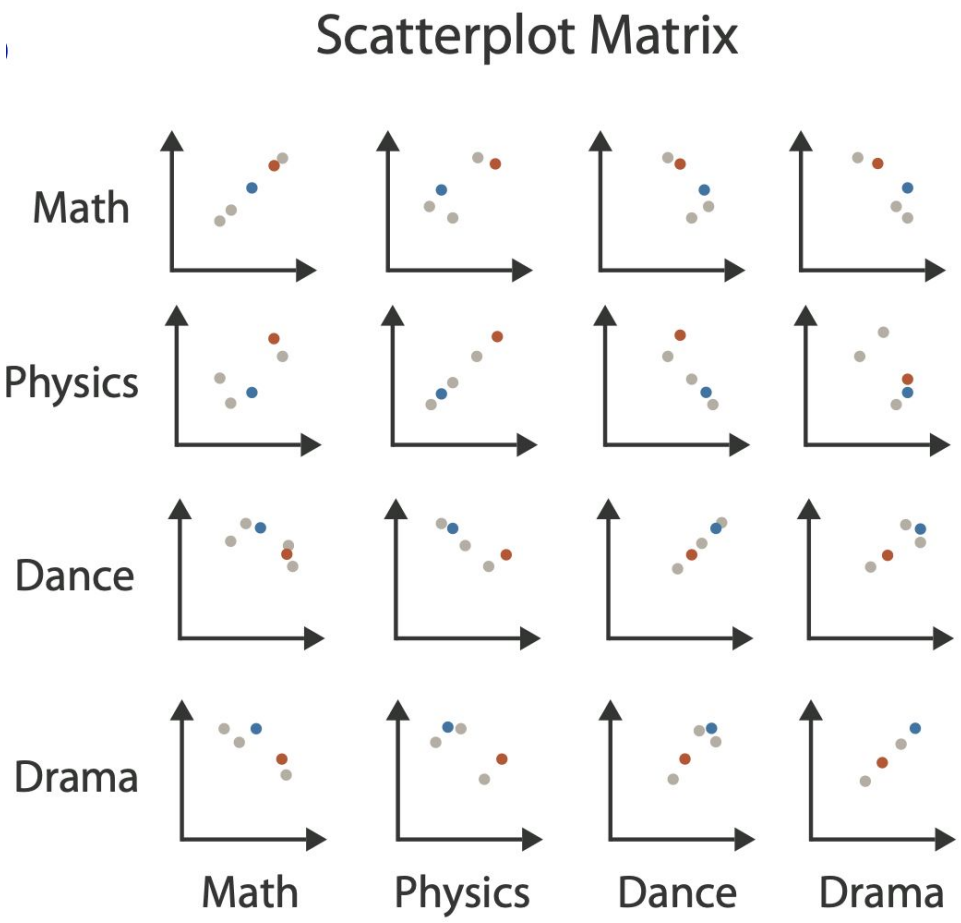
Scatterplots vs. Parallel coordinates

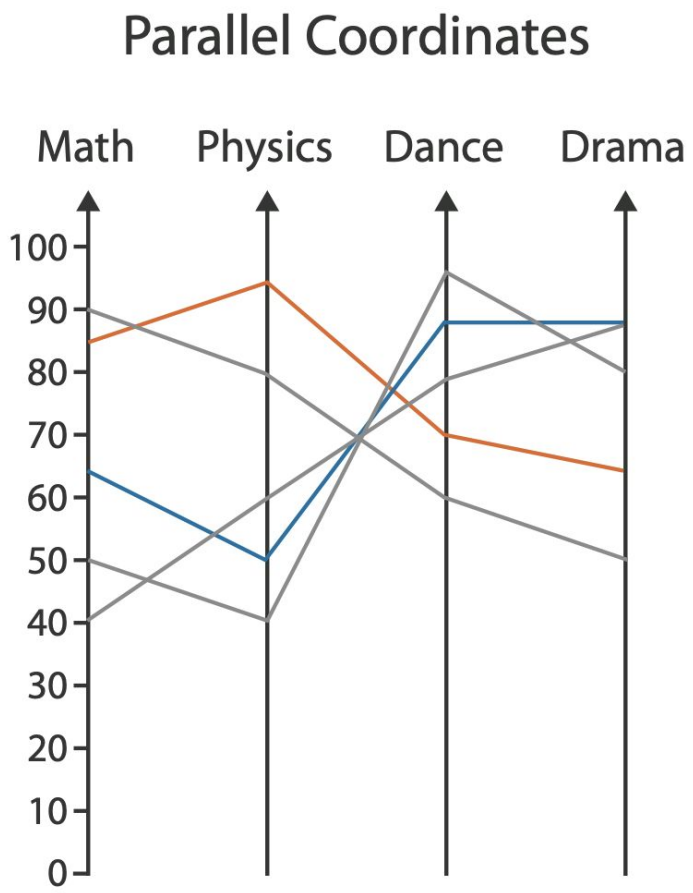
- Which do you prefer?





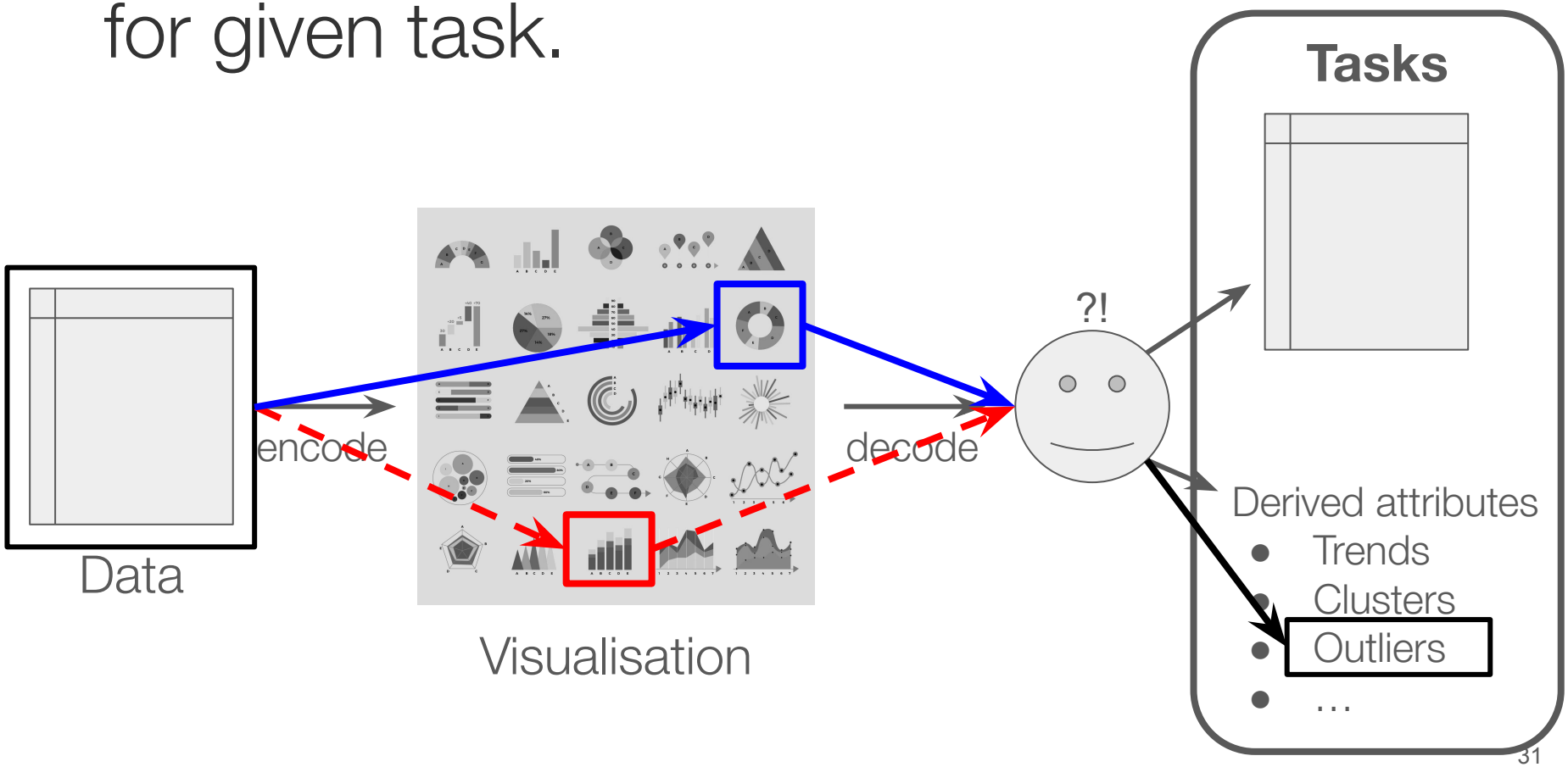
Scatterplots vs. Parallel coordinates





“Best” way to visualise data

- accurately decodes displayed information for given task.



Tasks

- **Cluster**
- Filter
- Compute derived value
- Find extremum
- Sort
- Determine range
- Characterize distribution
- Find anomalies
- **Retrieve value**
- Correlate

Cluster detection

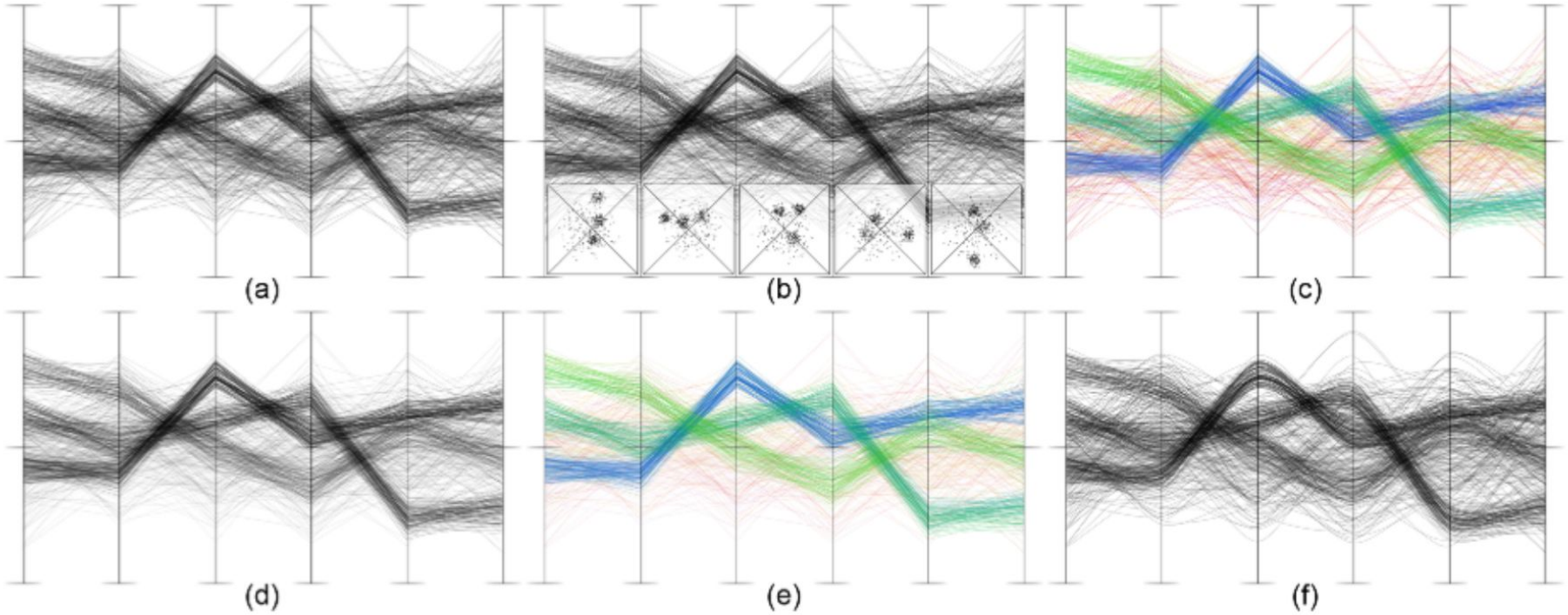


Figure 4: *The six non-animated PCP variations that were evaluated: (a) standard PCP, (b) scatter plots embedded into a PCP, (c) colored polylines, (d) blended polylines, (e) colored and blended polylines, and (f) curves instead of polylines.*

[Evaluation of Cluster Identification Performance for Different PCP Variants; Holten and van Wijk, EuroVis 2010]

Results

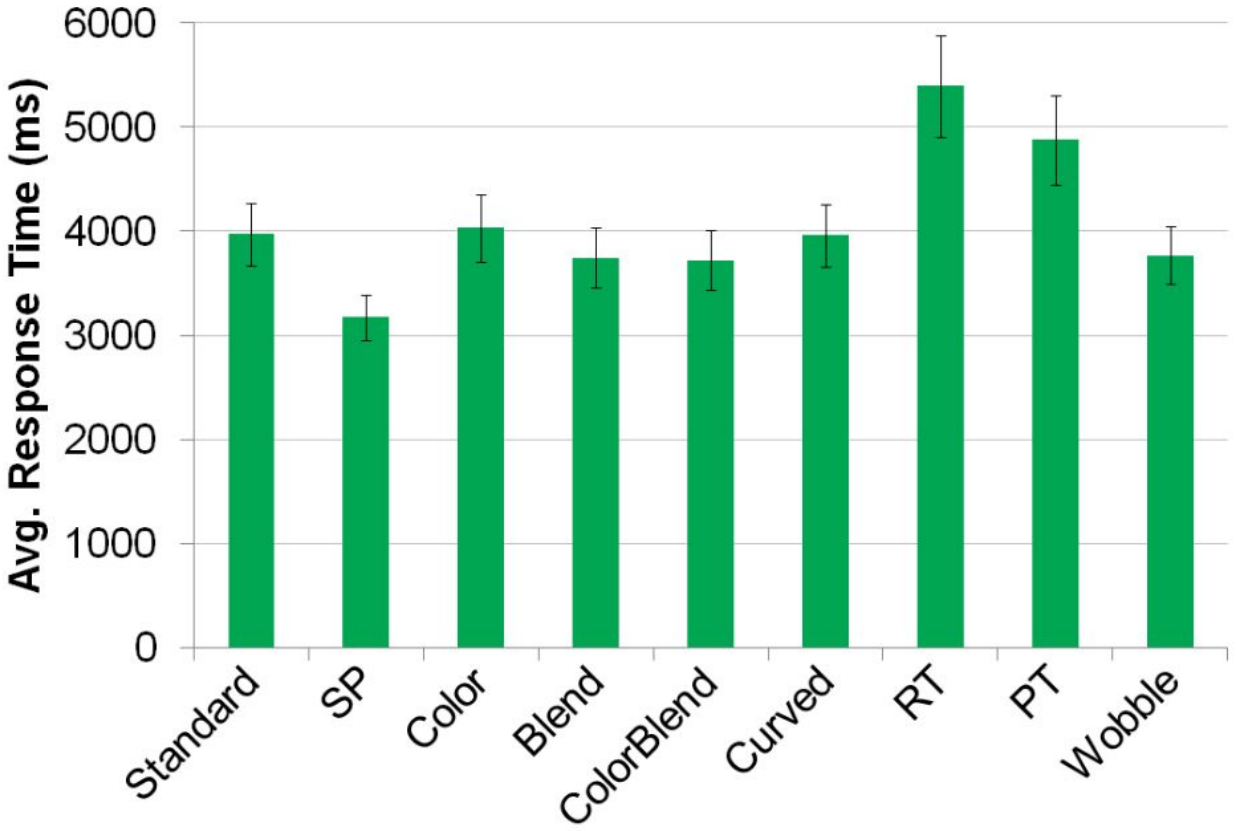


Figure 8: *Average response time for each of the PCP variations (shown with a 95% confidence interval).*

Results

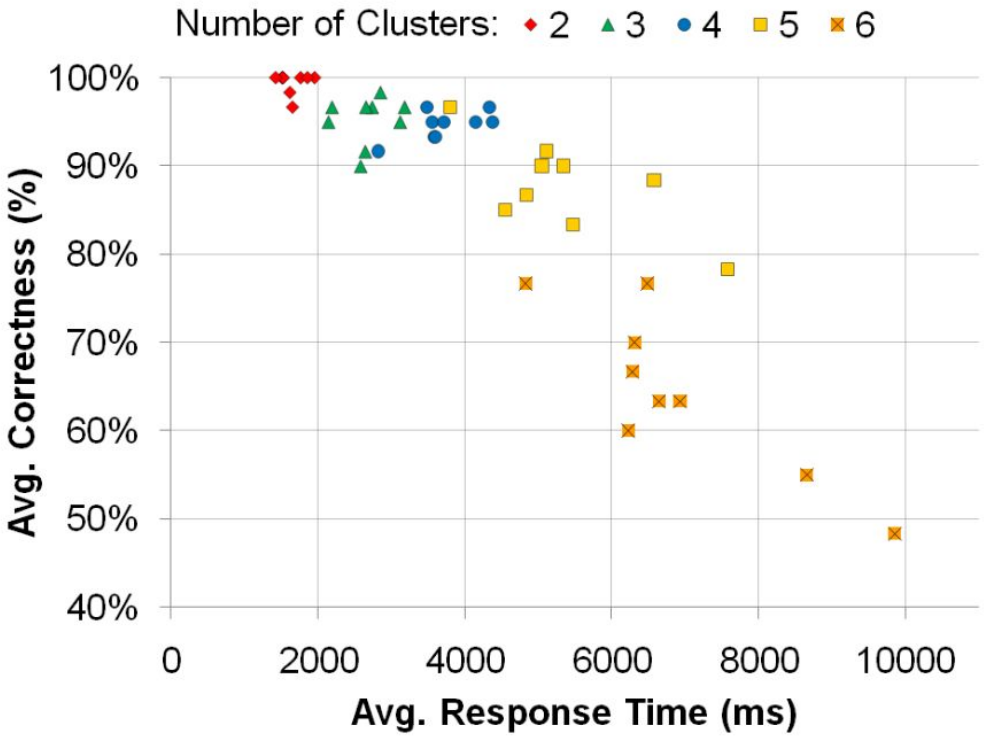
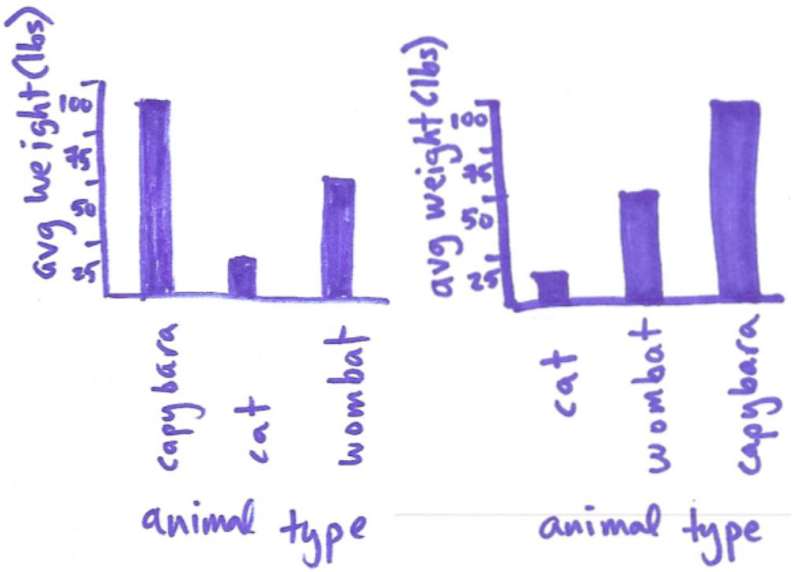


Figure 7: *There is a strong correlation ($r = -0.88$) between response time and correctness. Each point represents a unique combination of PCP variation and cluster count.*

Tasks

- Cluster
- Filter
- Compute derived value
- Find extremum
- Sort
- Determine range
- Characterize distribution
- Find anomalies
- **Retrieve value**
- Correlate

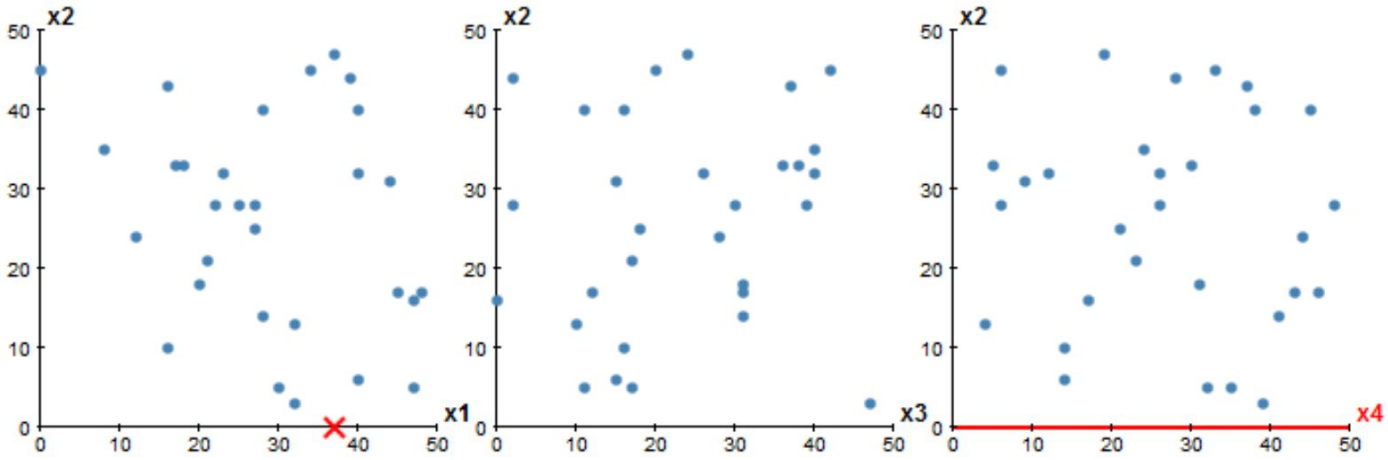


Retrieve value

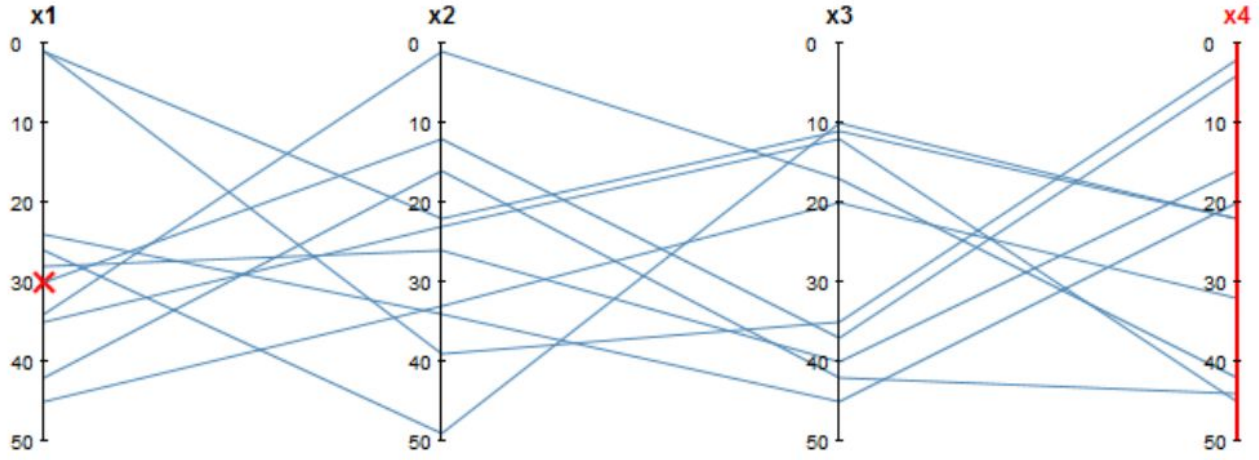
- Given the numerical value of one attribute of a data tuple, find the numerical value of another attribute of the same data tuple.

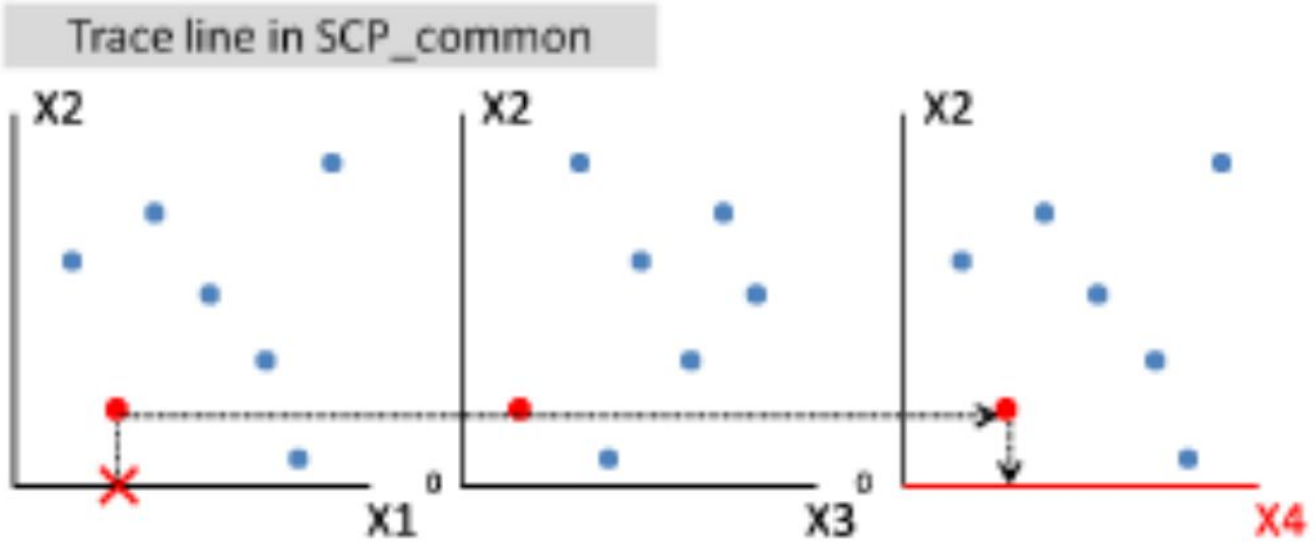
Multi-Variate Data Tuple (X_1, X_2, X_3, …. , X_n)

38 **?**

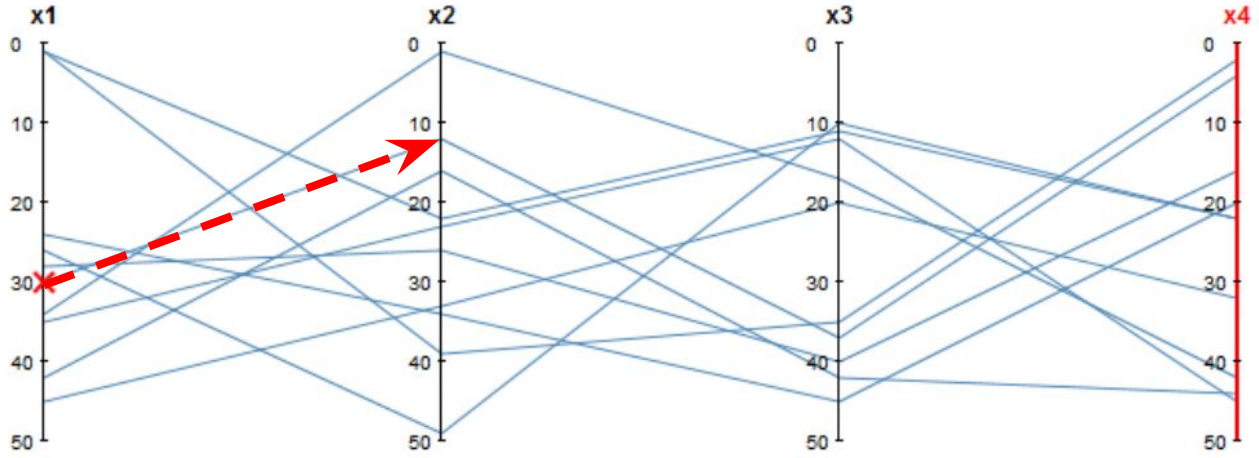


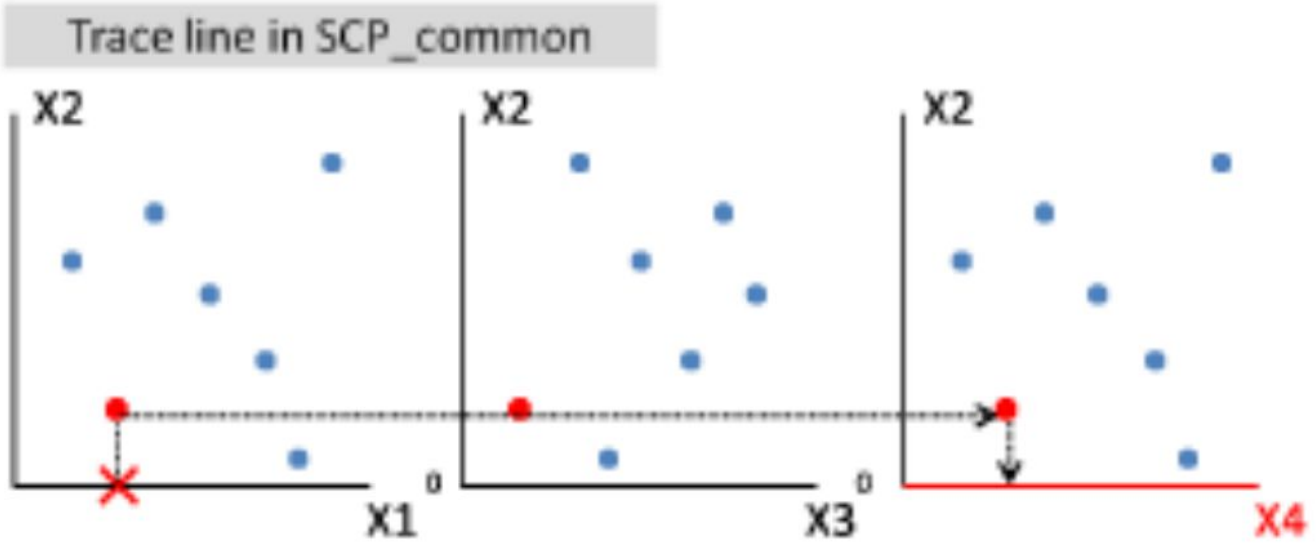
Retrieve value



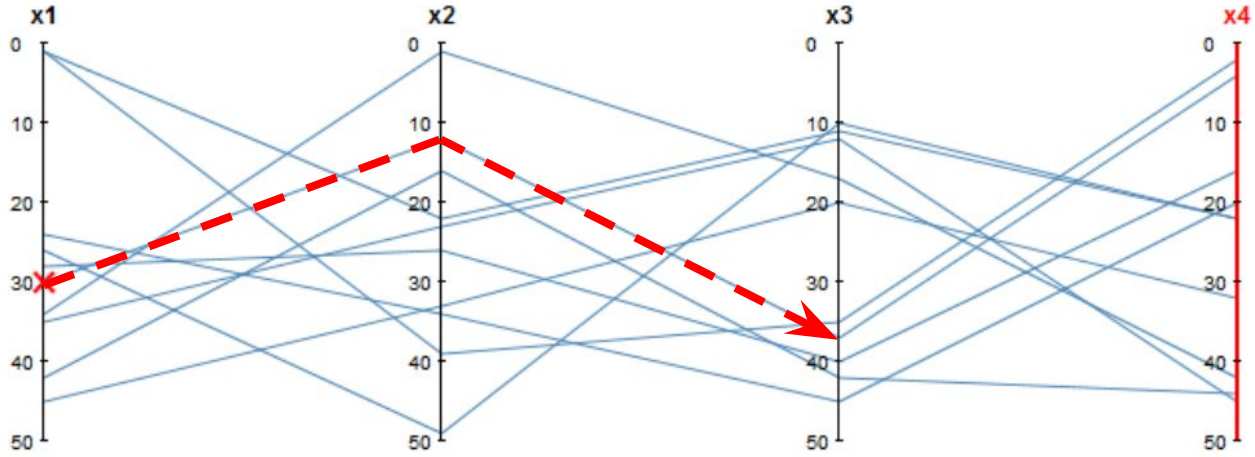


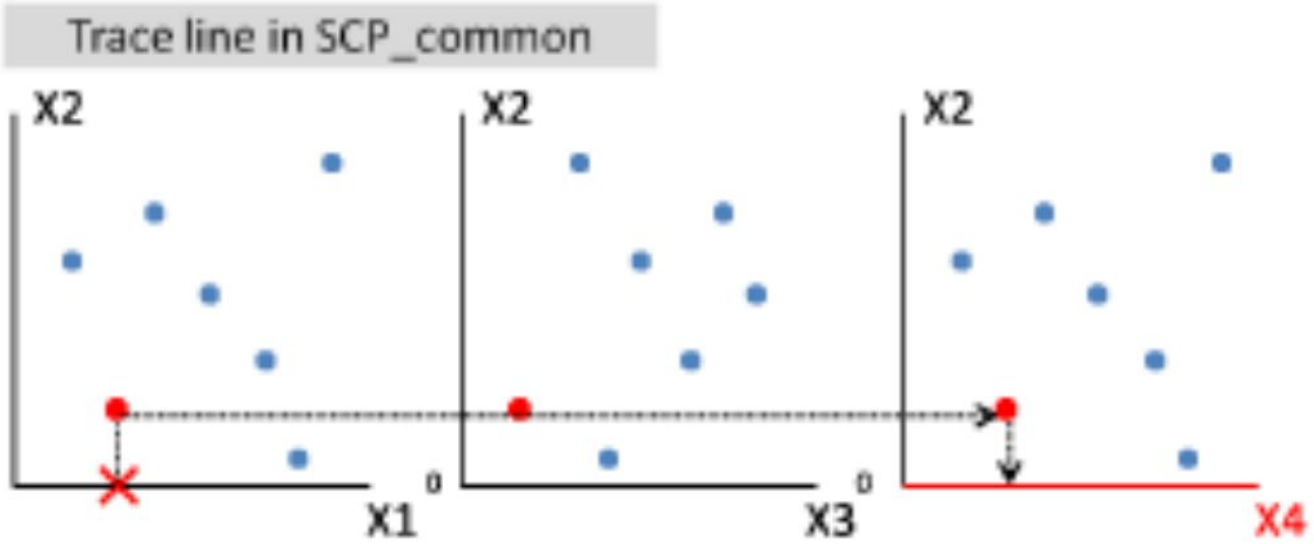
Retrieve value



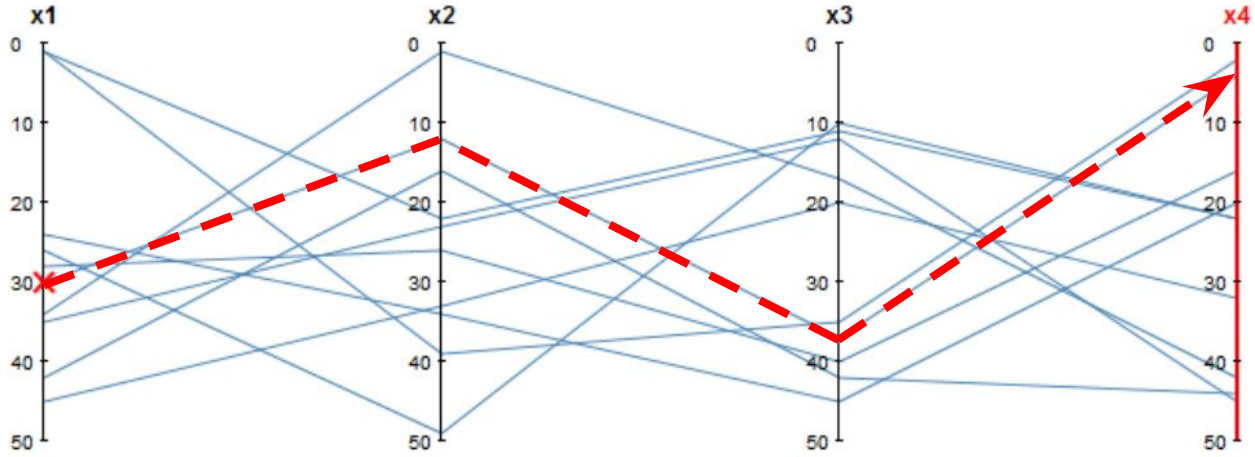


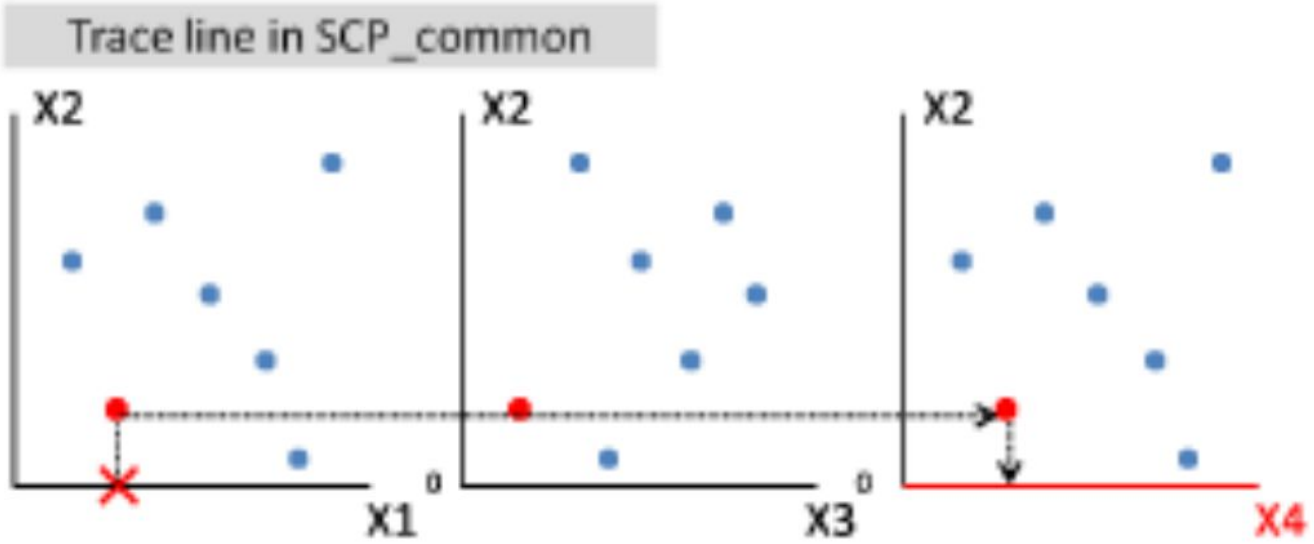
Retrieve value



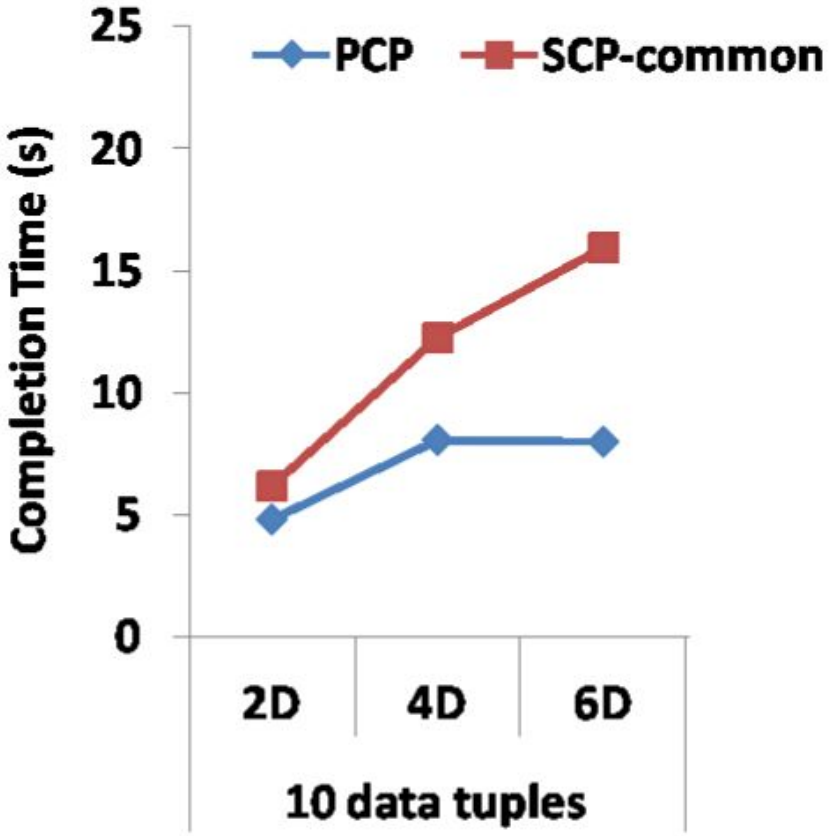


Retrieve value

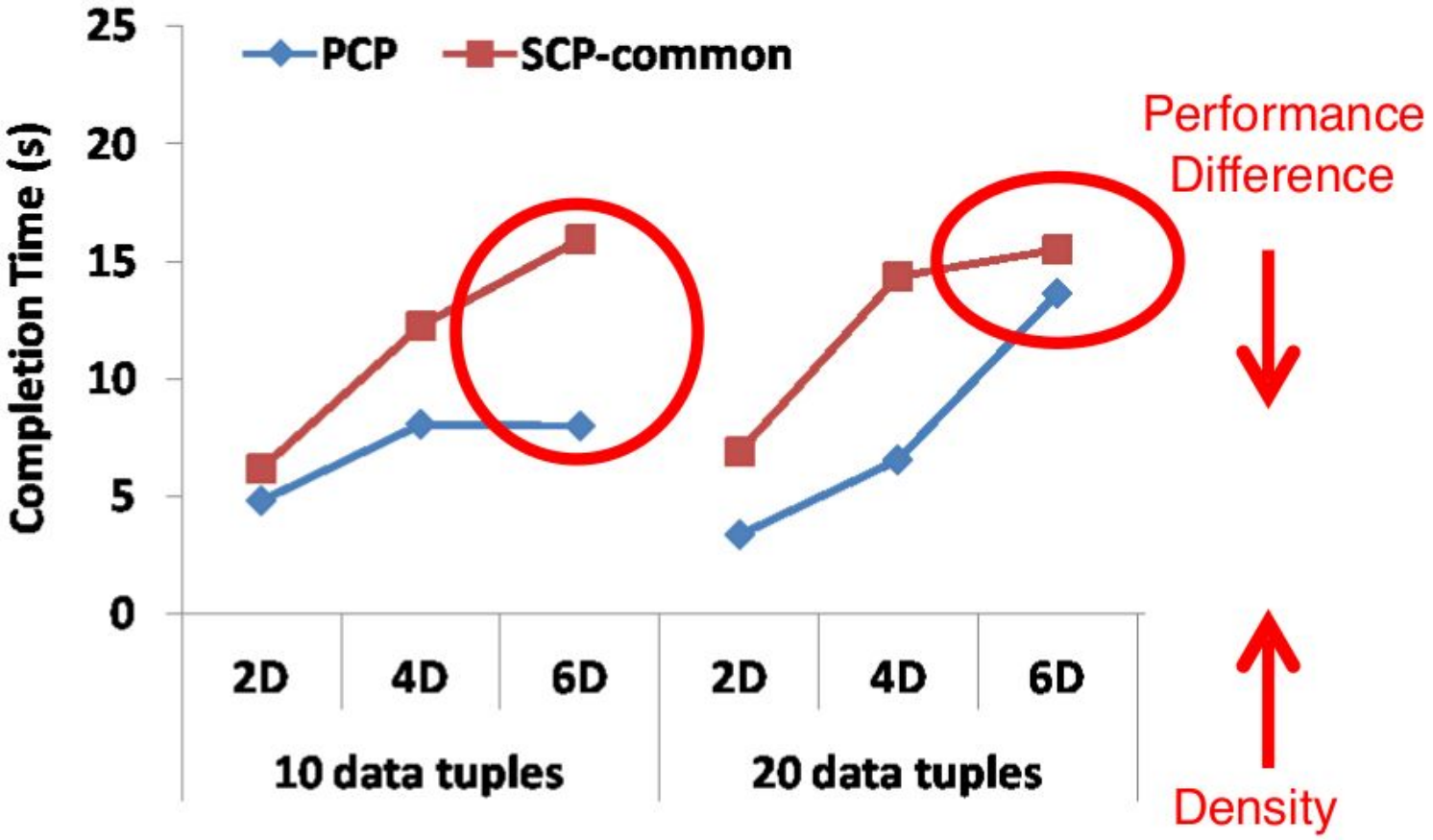




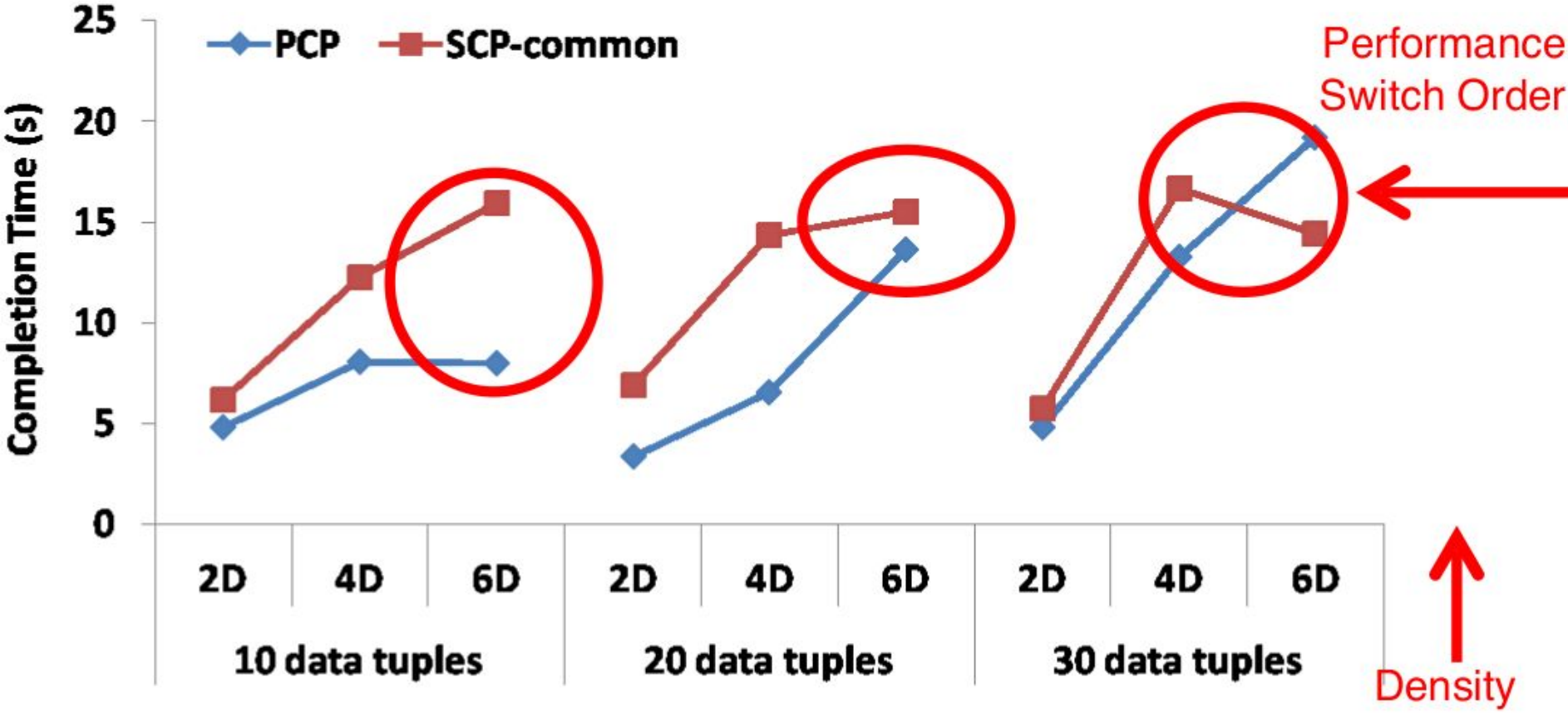
PC vs. SC: completion time



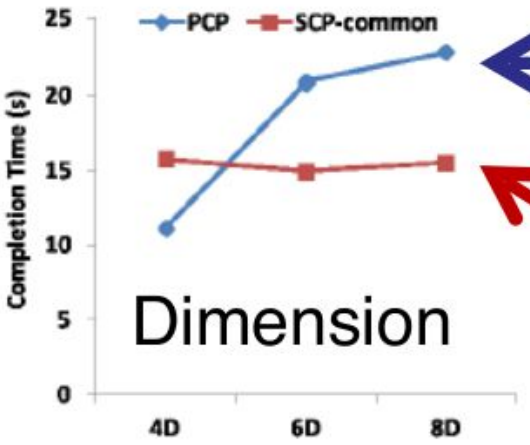
PC vs. SC: completion time



PC vs. SC: completion time

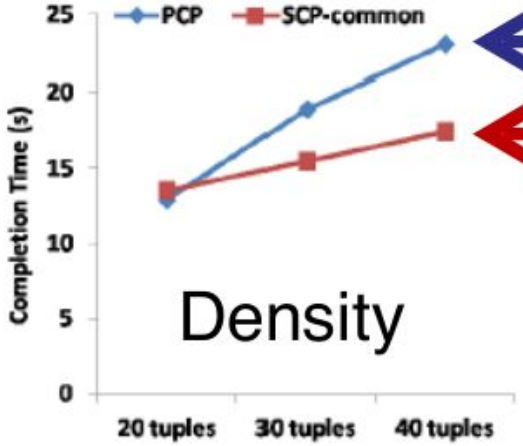


Take-away lessons



The value retrieval performance of PCP increases depending on dimensionality.

The performance of SCP-common seems independent of dimensionality.



Increasing density affects the performance of PCP more than it affects SCP-common.

Visualizing *very* high dimensional data

- Reducing dimensionality
- Topological data analysis

DR: Problem definition

Input: Vector space with dimensions $D = \{d_1, ..., d_n\}$

Idea: Instead of removing features, determines how all the dimensions may actually express a smaller subset.
→ Moreover, redundant features are summarized

Output: Transformed vector space with dimensions $D' \subseteq D$ where the transformation is task dependent

Examples of DR algorithms

- Linear
  - Principal component analysis (PCA)
  - Singular value decomposition (SVD)
  - Independent component analysis (ICA)
- Non linear
  - Multidimensional scaling (MDS)
  - Local linear embedding (LLE)
  - t-SNE

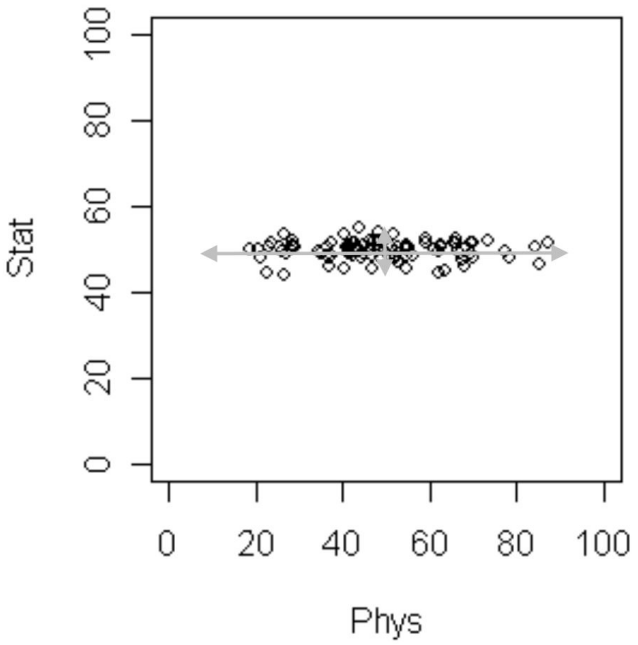
Principal component analysis

Idea: Variation/variance along an axis implies information content (zero variance → zero information)

Principal component analysis

Example: Grades of students in Physics and Statistics.

- If we want to compare the students, which grade is more discriminative? Statistics or Physics?



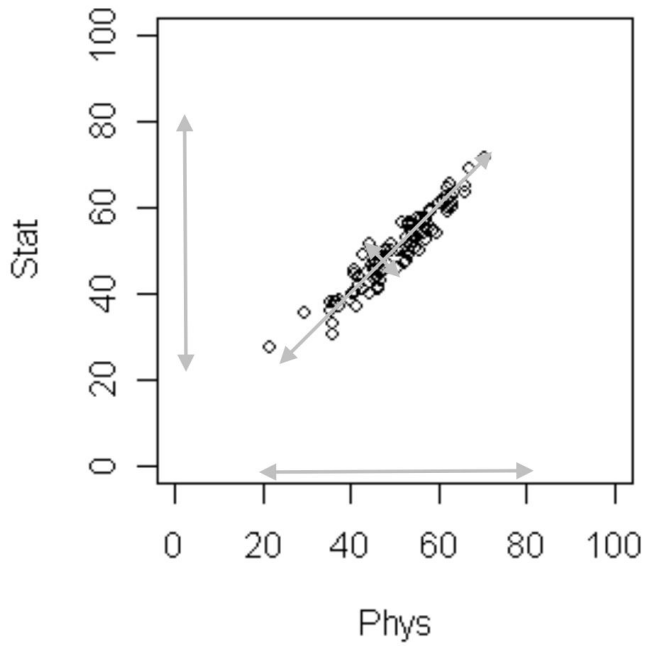
Feature Phys contains more information because the variation along that axis is larger.

Source: http://astrostatistics.psu.edu/su09/lecturenotes/pca.html

Principal component analysis

Example: Grades of students in Physics and Statistics.

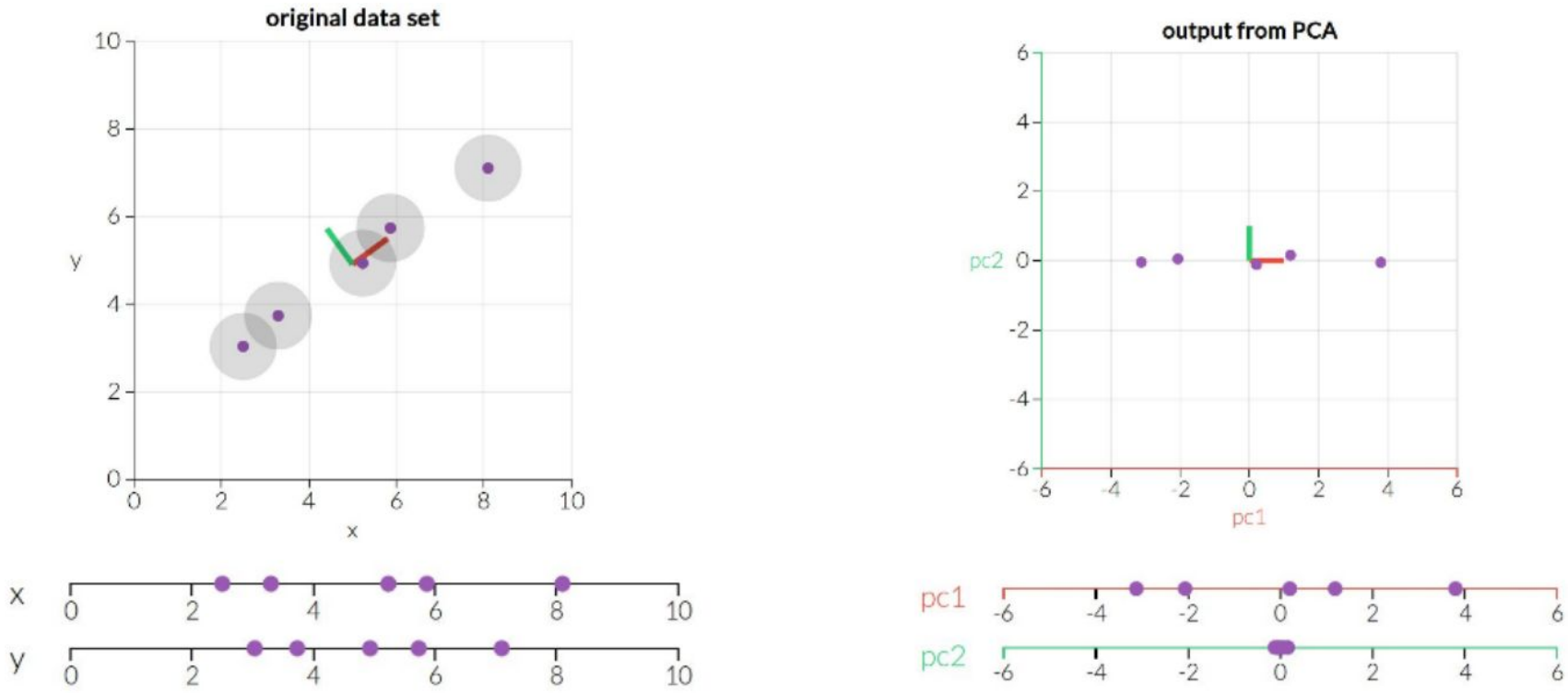
- If we want to compare the students, which grade is more discriminative? Statistics or Physics?



The most relevant feature is now a linear combination of Phys & Stat

Source: http://astrostatistics.psu.edu/su09/lecturenotes/pca.html

Principal component analysis



Source: https://miro.medium.com/max/2388/1*V3JWBvxB92Uo116Bpxa3Tw.png

Principal component analysis

Idea: Variation/variance along an axis implies information content (zero variance → zero information)

PCA determines

- Eigenvectors of covariance matrix
- Eigenvalues = variance along principle components

→ Stable & fast

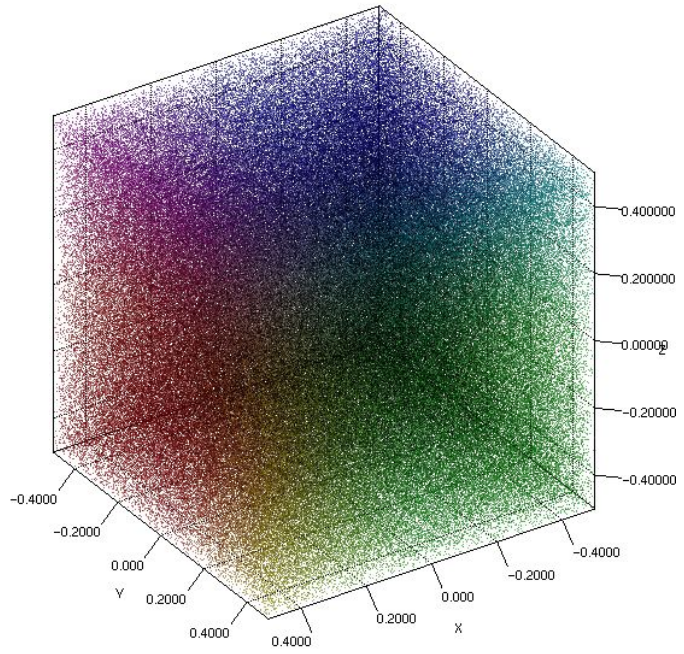
  - converges to same point, every time
  - no additional parameters

Intrinsic dimensionality

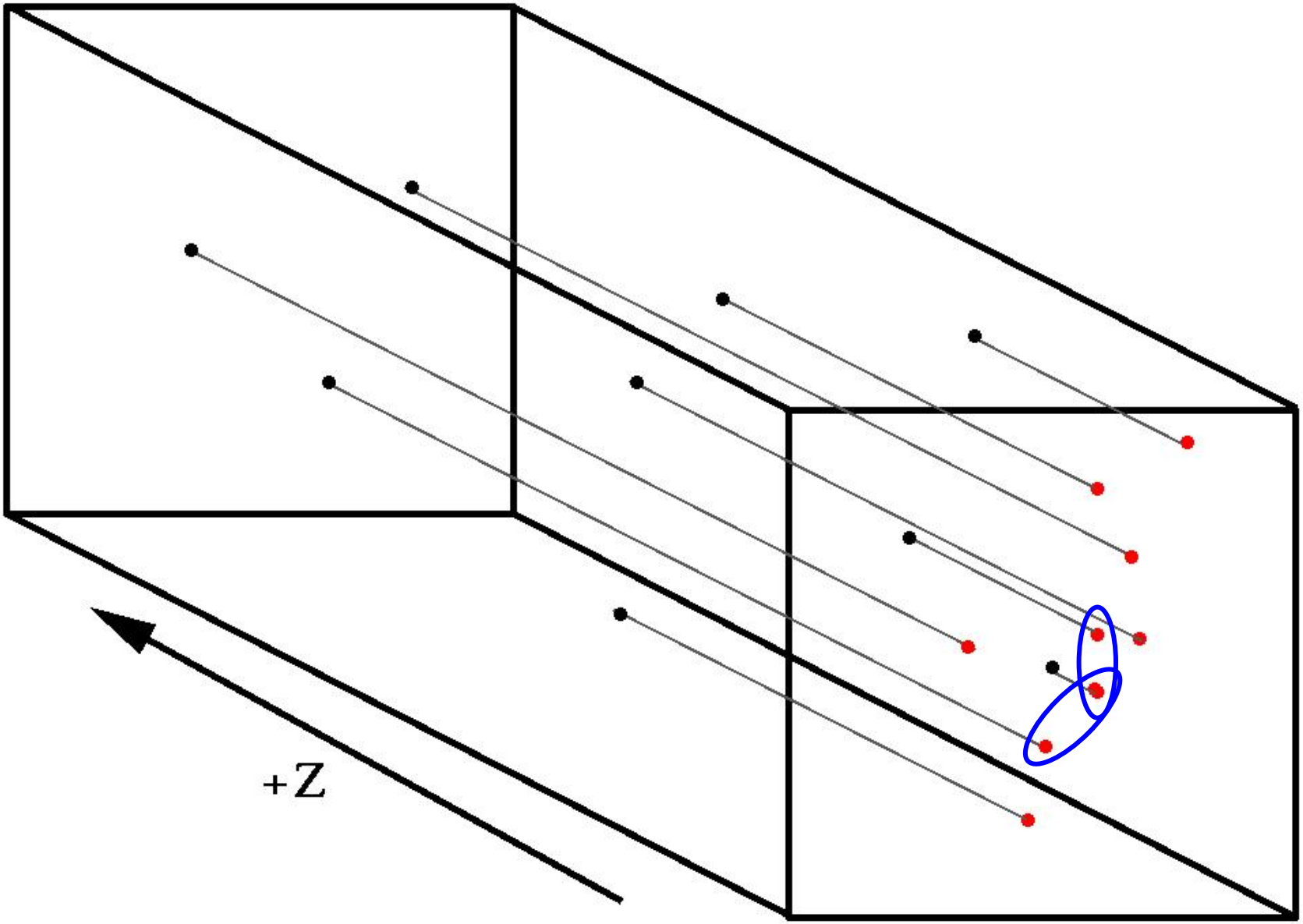
Attention: We generally do not know the dimensionality of the intrinsic manifold

→ Can lose valuable information via DR

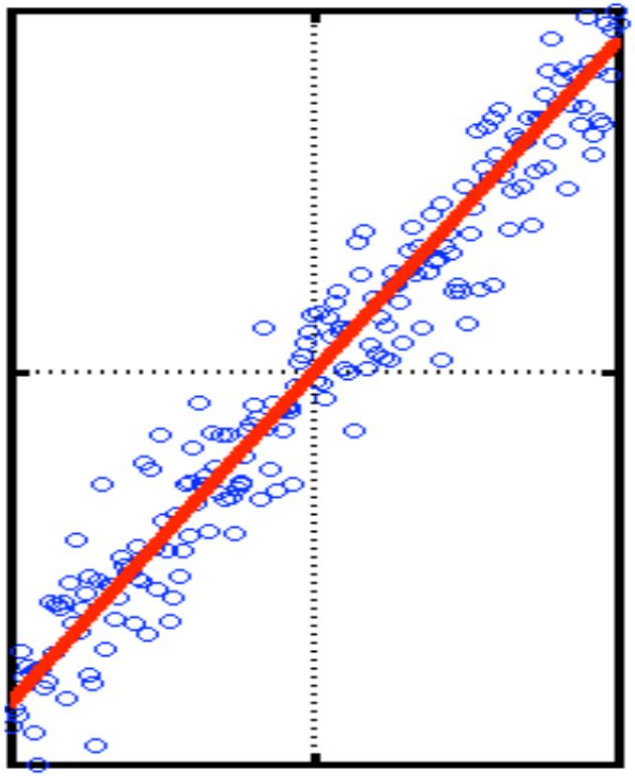
- Scree plot to determine ID
- Methods to estimate it

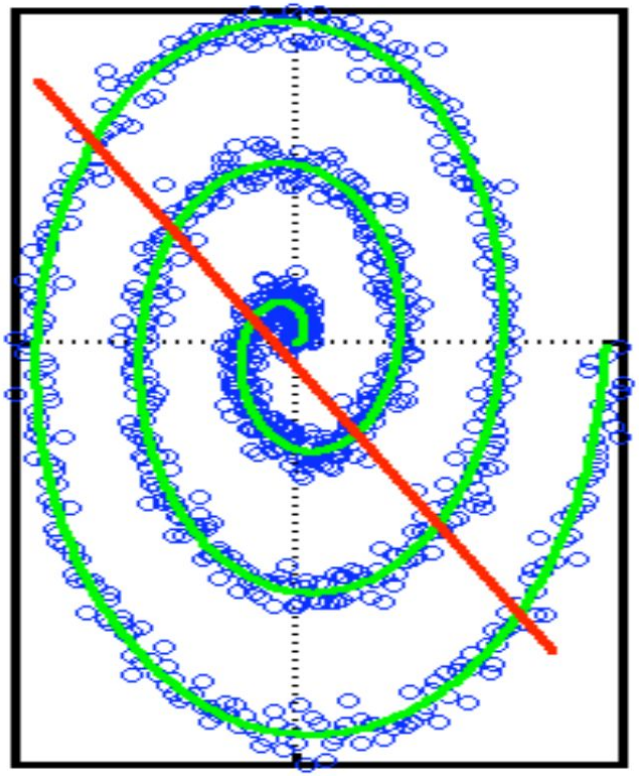


http://www.jzy3d.org/js/slider/images/ScatterDemo.png


+Z

Non-linearities

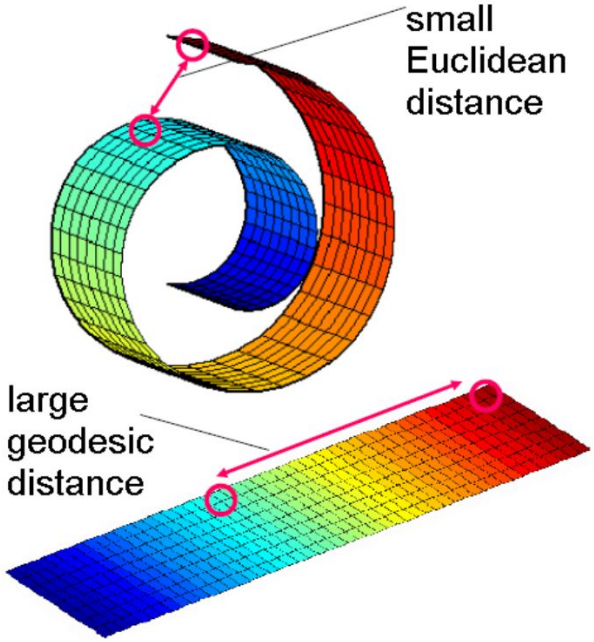




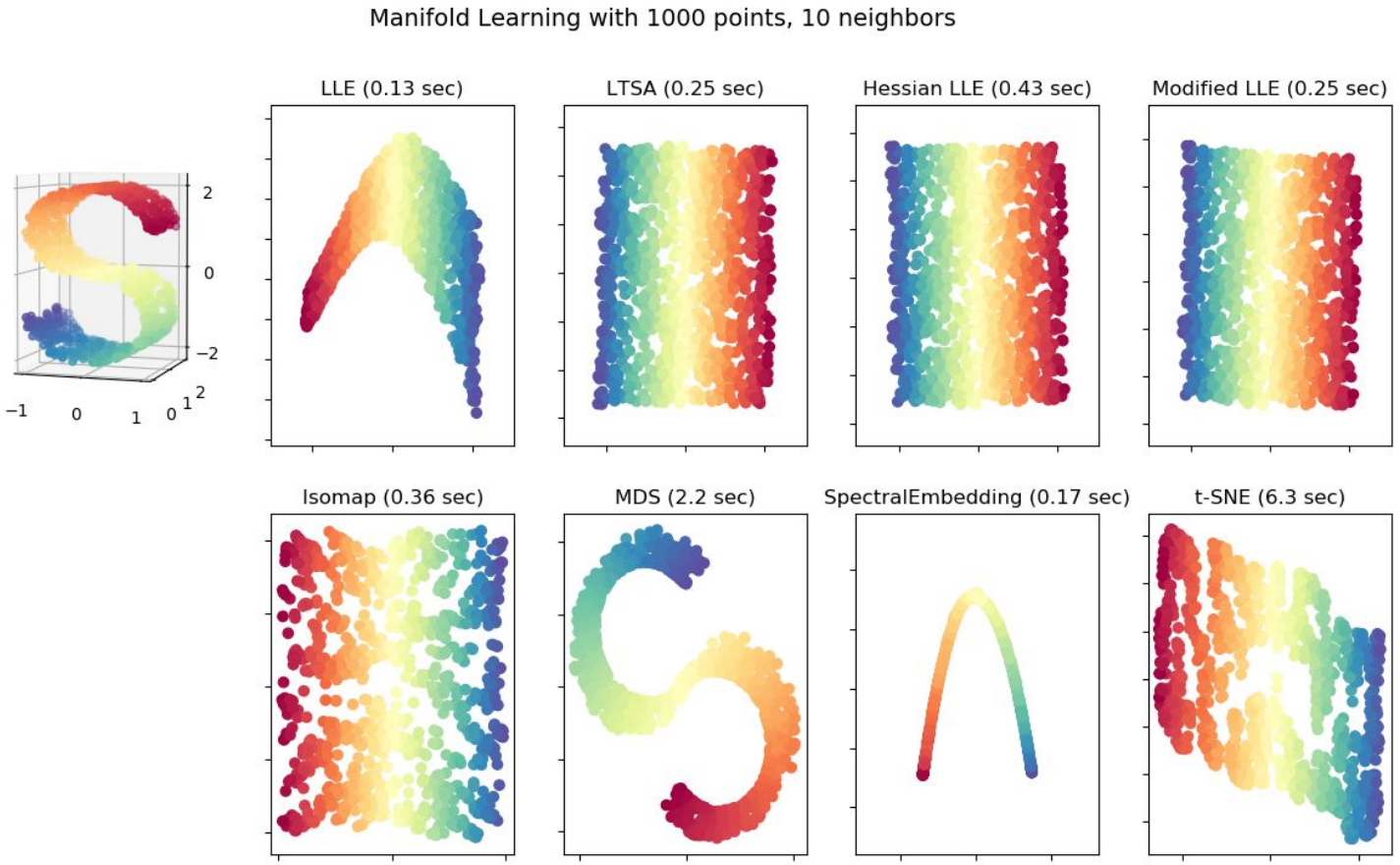
Manifold learning

If variables depend on each other their joint distribution does not span the whole space

→ data lies on (or around) lower dimensional manifold



Non-linear DR: Manifold learning



Source: Scikit Learn (scikit-learn.org/stable/modules/manifold.html)

Visualizing *very* high dimensional data

- Reducing dimensionality
- Topological data analysis

Topological data analysis

- Projection loss of dimensionality reduction (DR) methods
- Inconsistent Results
  - ↳ Different DR algorithms produce dissimilar projections because they encode different assumptions.

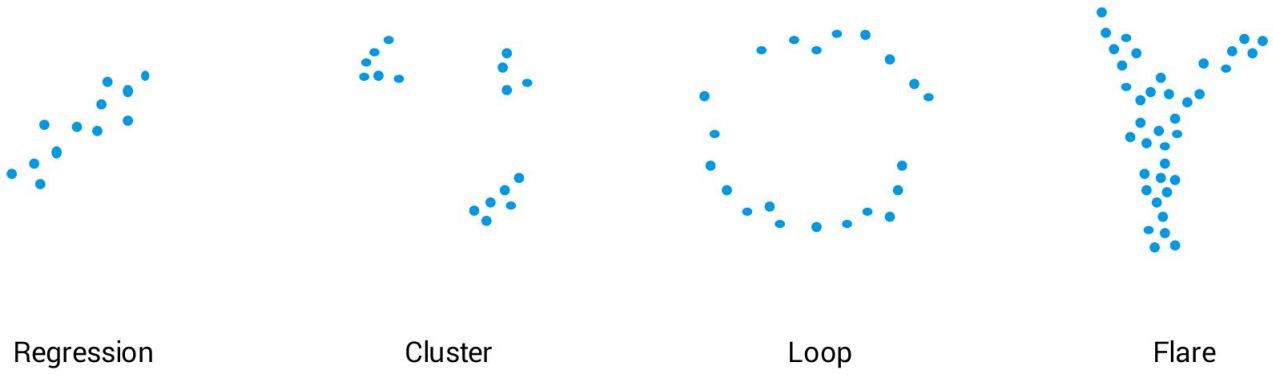
Goal of TDA

- Do not attempt to obtain fully accurate representation data set
  - ↳ Rather low-dimensional image that is easy to understand & highlights areas of interest

→ Data has **underlying shape** which carries meaning

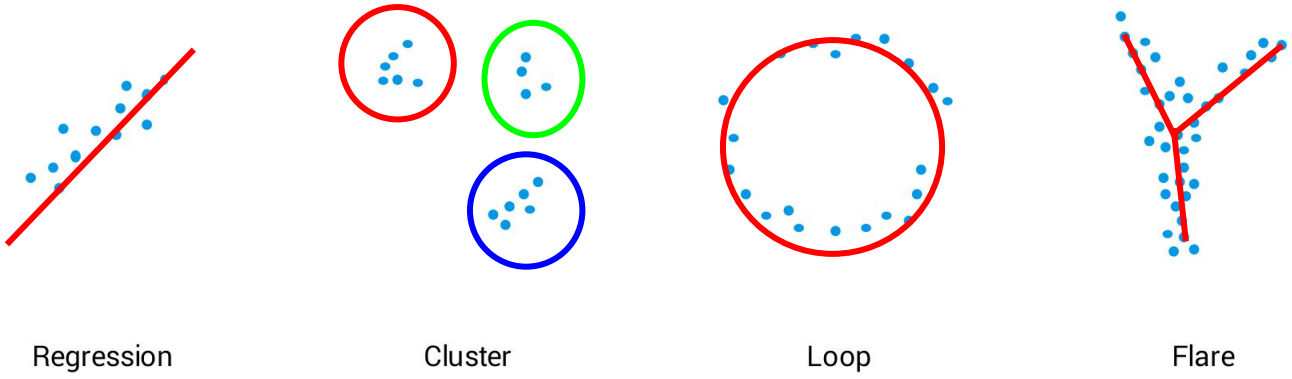
TDA philosophy

- All data has an underlying shape & that shape has meaning
- Data does not restrict itself to certain shapes.



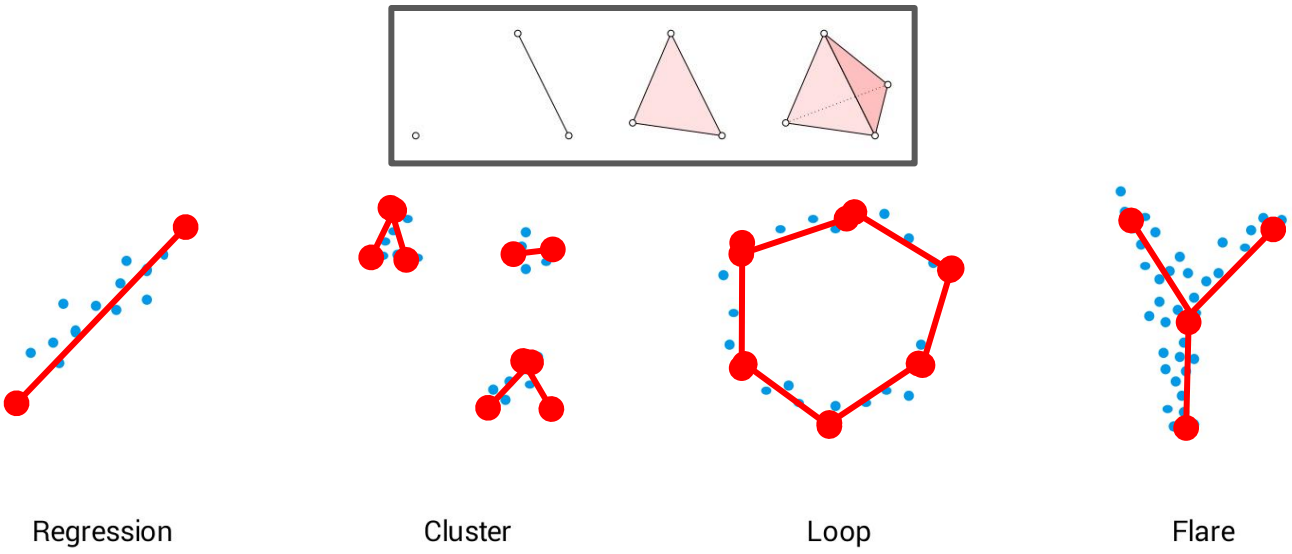
TDA philosophy

- All data has an underlying shape & that shape has meaning
- Data does not restrict itself to certain shapes.



TDA philosophy

- All data has an underlying shape & that shape has meaning
- Data does not restrict itself to certain shapes.



Mapper algorithm

Singh, G.; Memoli, F. & Carlsson, G. (2007), 'Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition' , The Eurographics Association

Mapper - Topological Construction

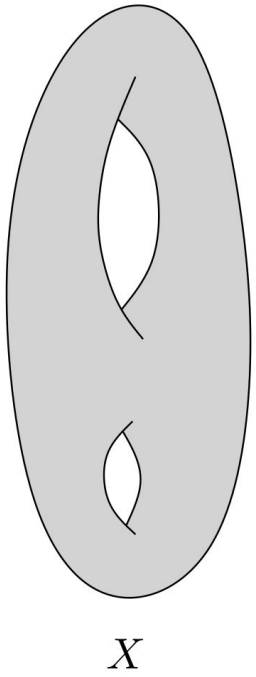
Representing a topological space *X* as graph needs 2 steps

1. Filter functions & finite covering
   i. Continuous filter function $f : X \to \mathbb{R}$
   ii. Cover the image of the filter function by open intervals

$$im(f) \subseteq \bigcup_{I \in \mathcal{I}} I$$

Singh, G.; Memoli, F. & Carlsson, G. (2007), 'Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition' , The Eurographics Association

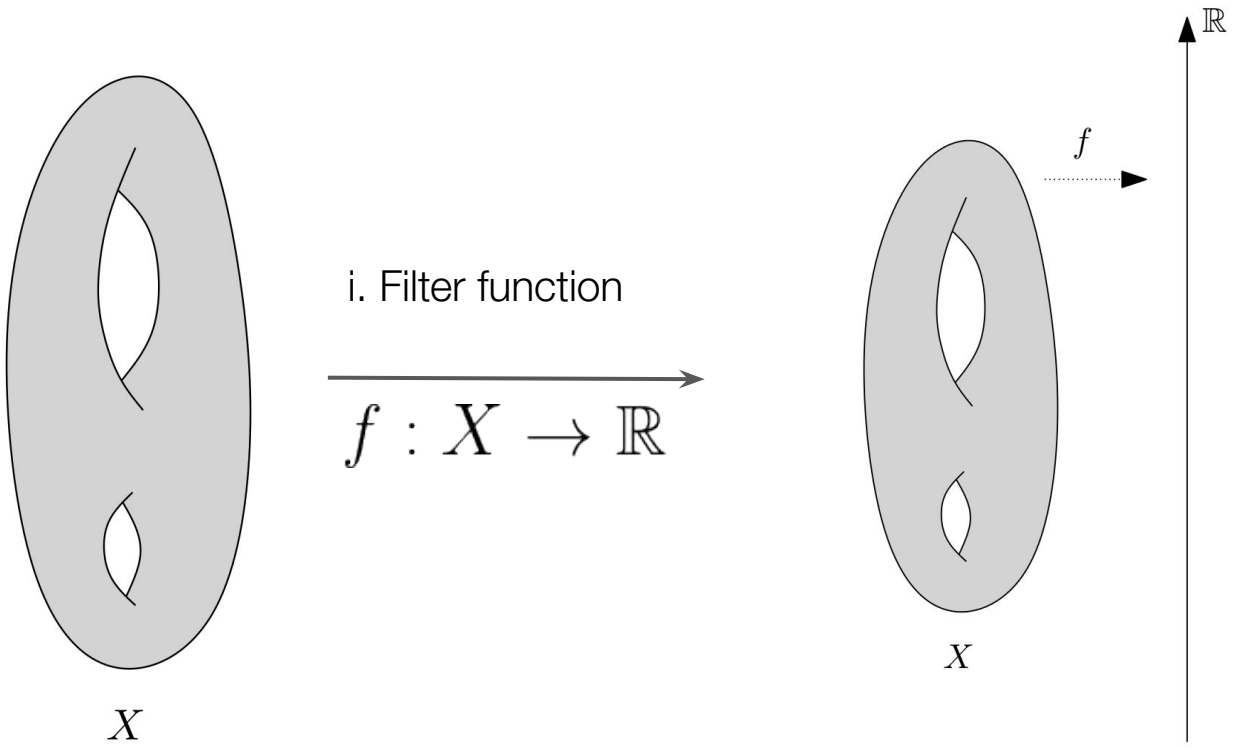
Filter functions & finite covering



X

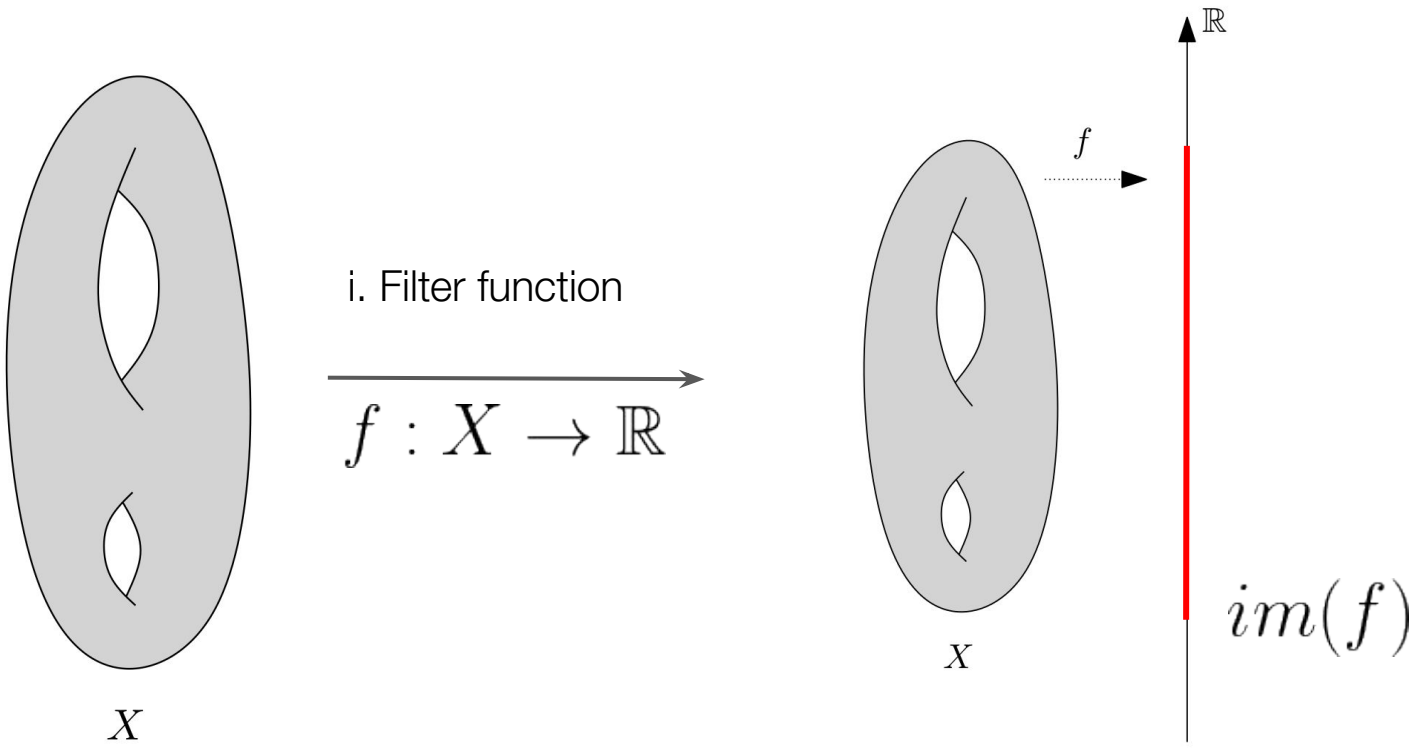
Topological space

Filter functions & finite covering



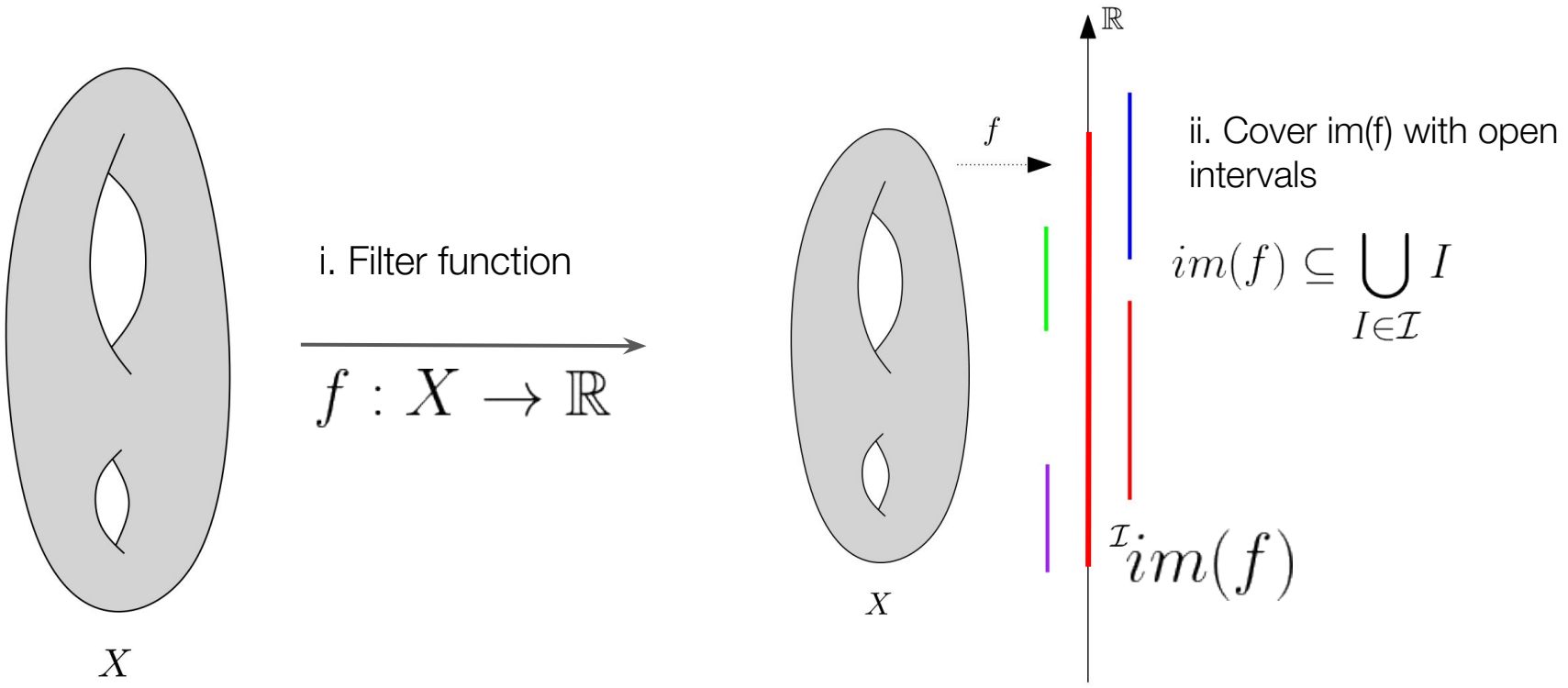
Topological space

Filter functions & finite covering



Topological space

Filter functions & finite covering



Topological space

Implementation - Topological Construction

2. Pullback cover, connected cover & Mapper

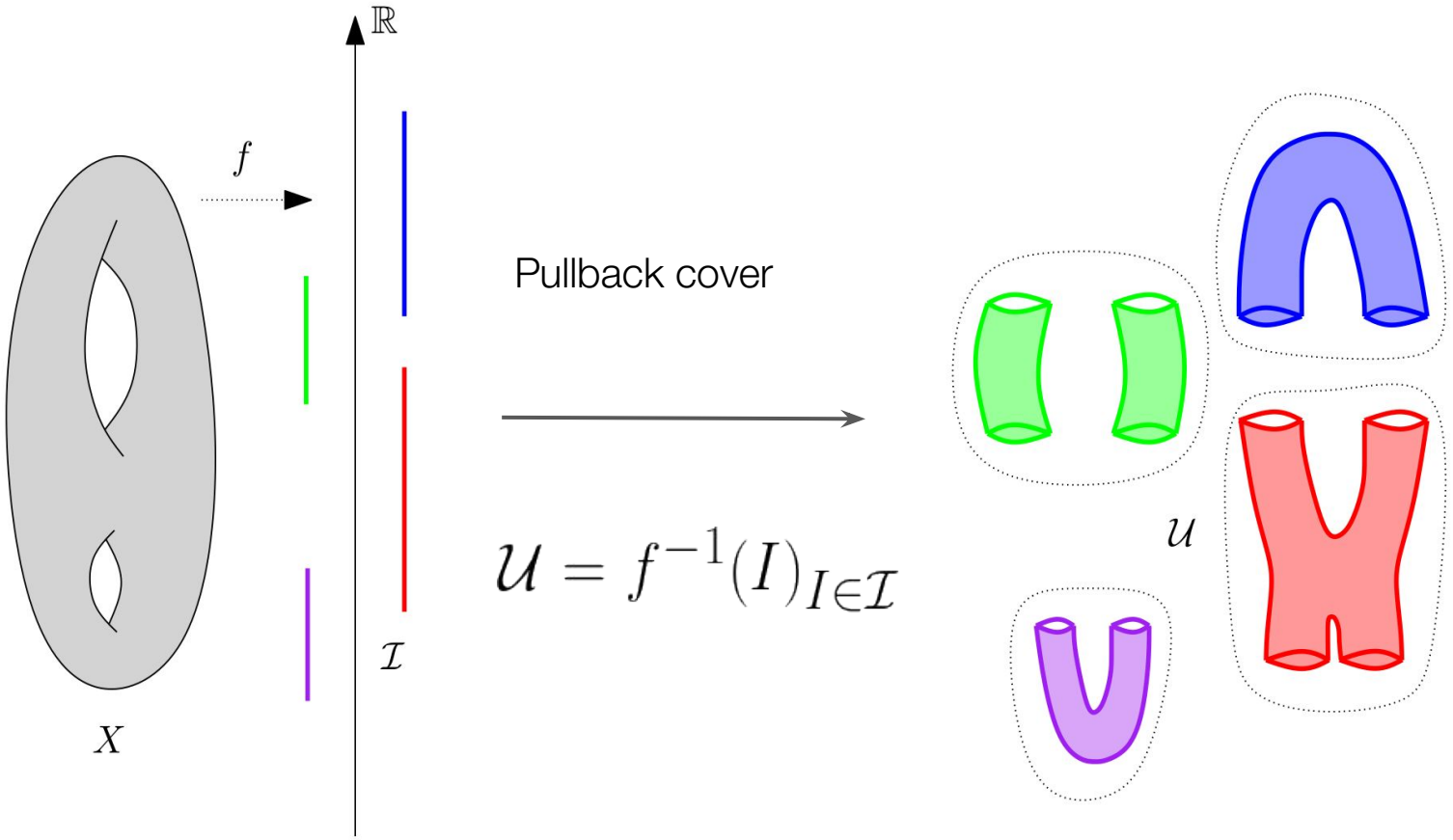
   i. Get pullback cover $\mathcal{U}$ of *X* for all intervals

$$\mathcal{U} = f^{-1}(I)_{I \in \mathcal{I}}$$

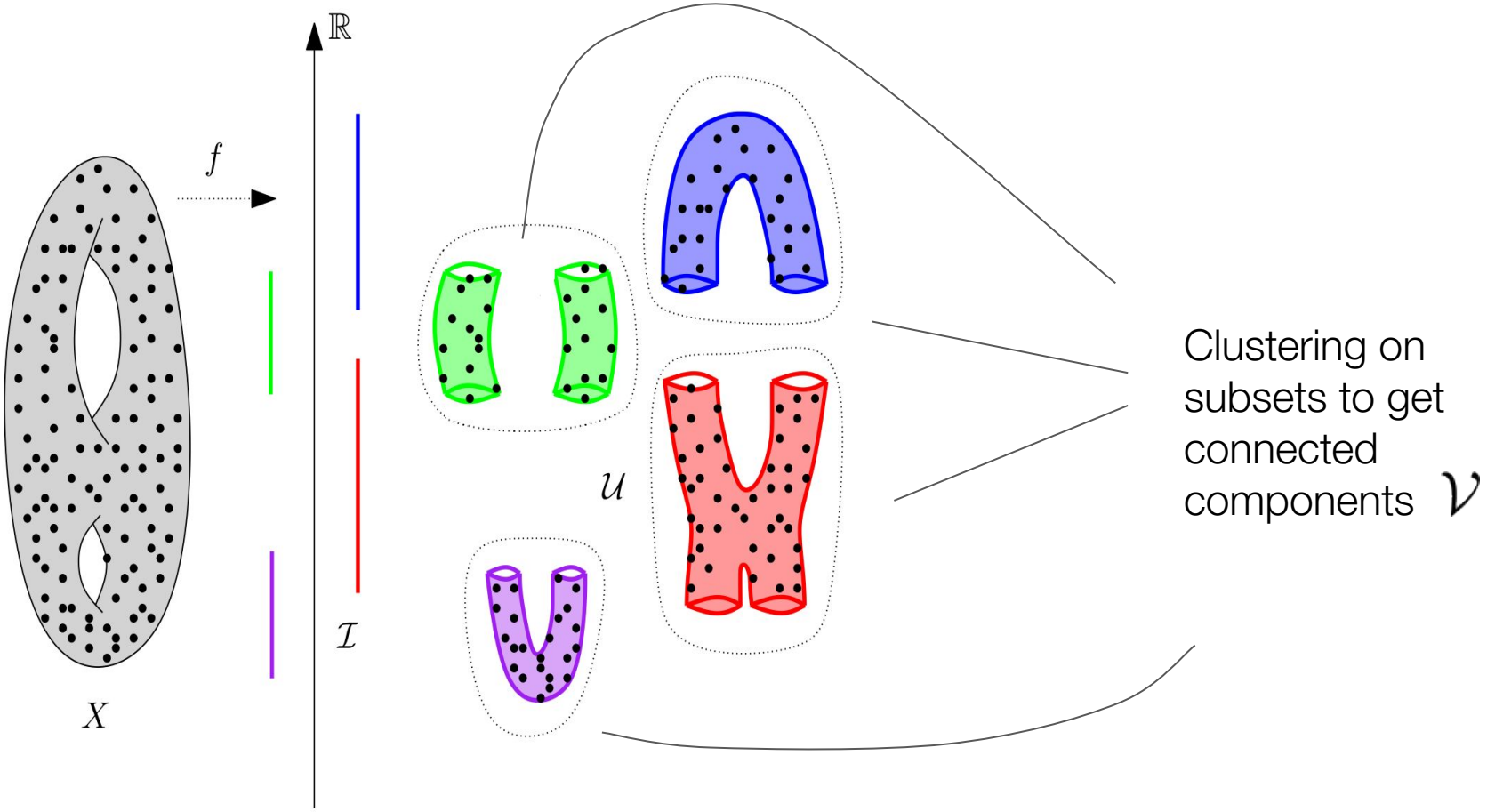
   ii. Subdivide $\mathcal{U}$ into its connected components in *X* → connected cover $\mathcal{V}$

   iii. Mapper is the "*nerve*" of $\mathcal{V}$

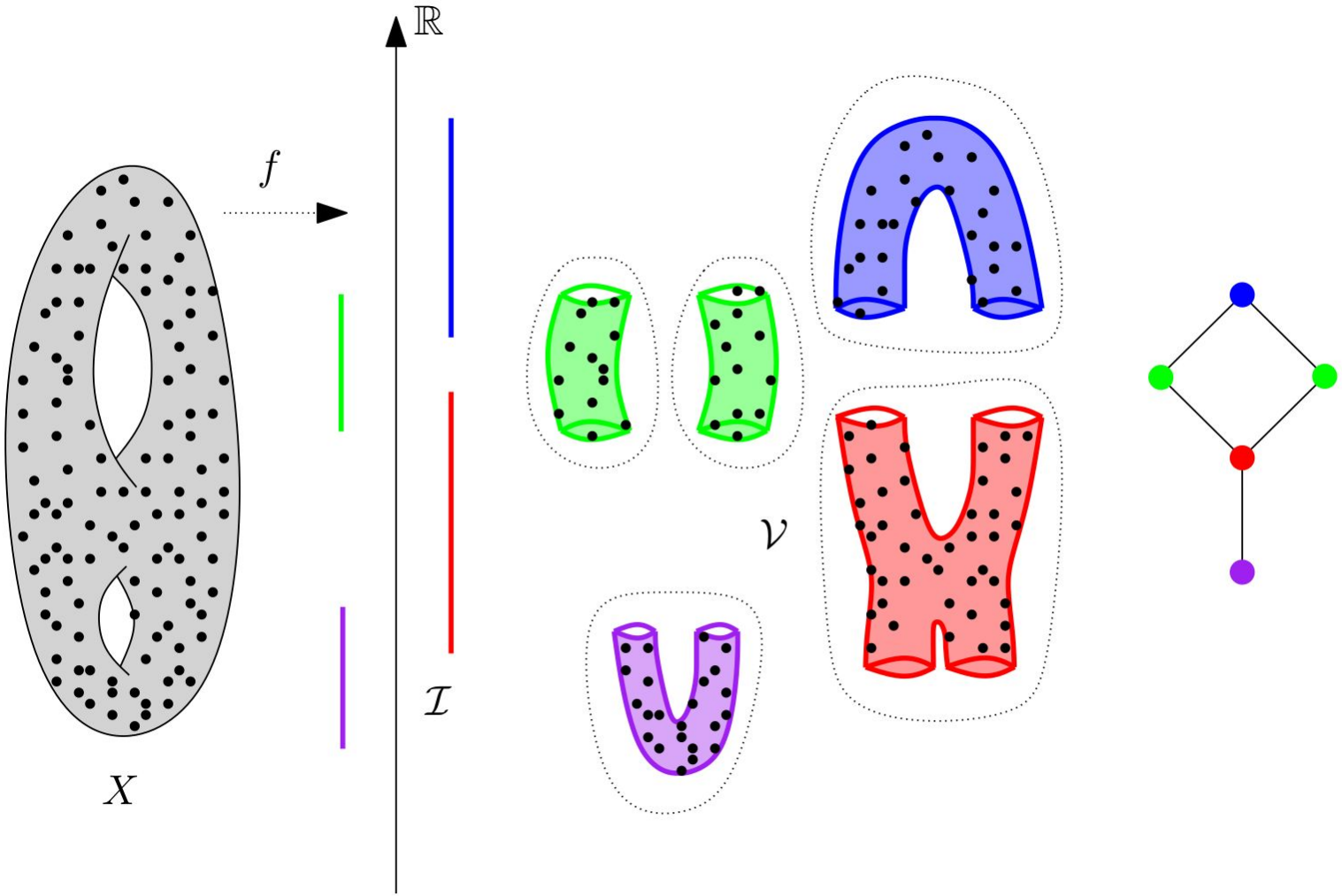
Pullback cover



Implementation

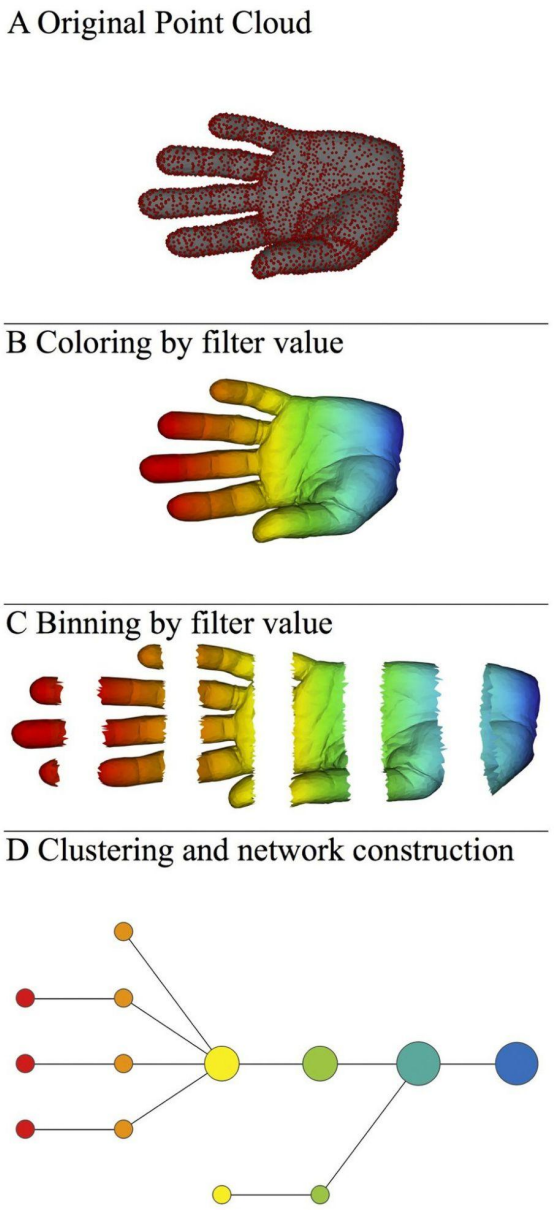


Implementation

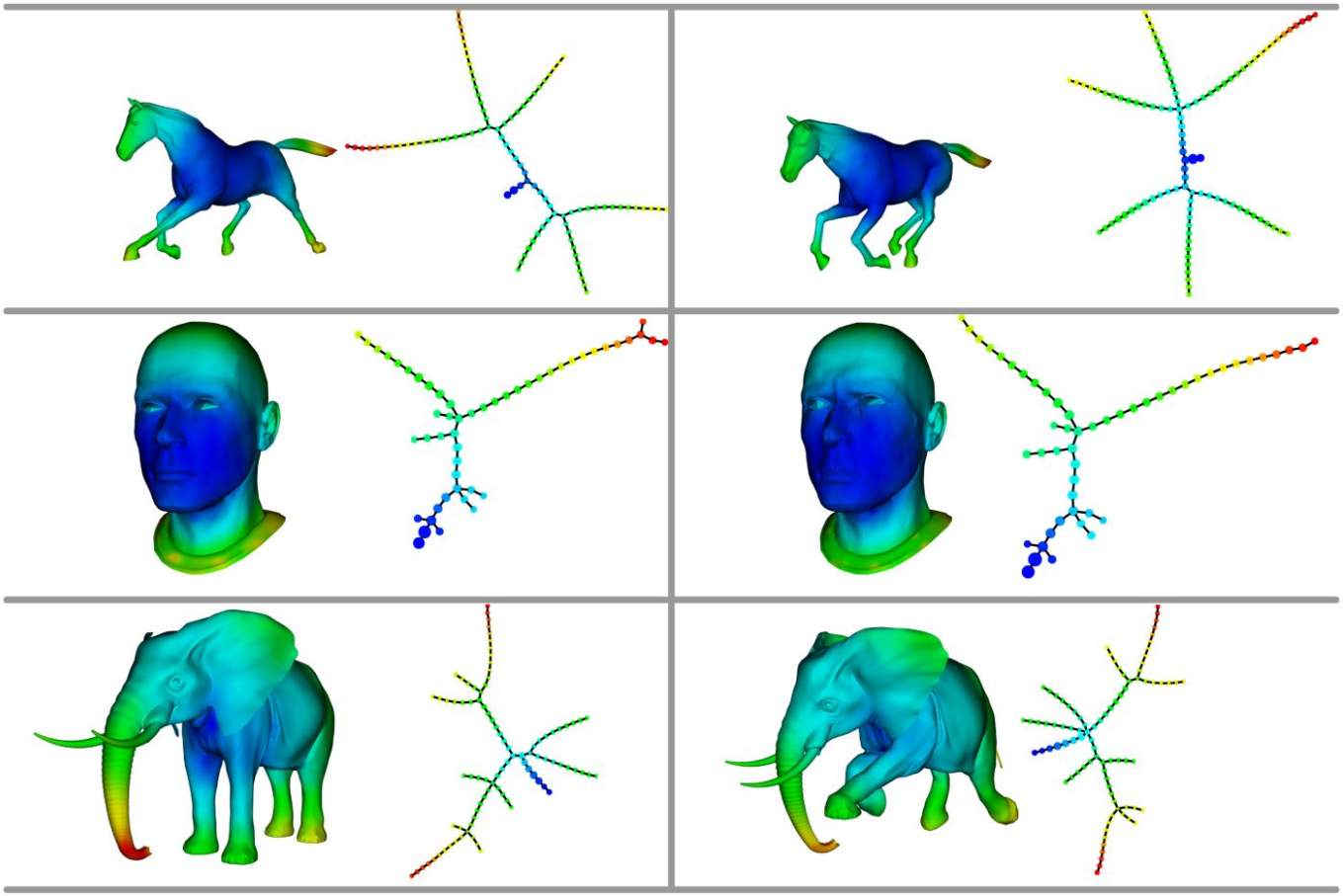


Filter functions

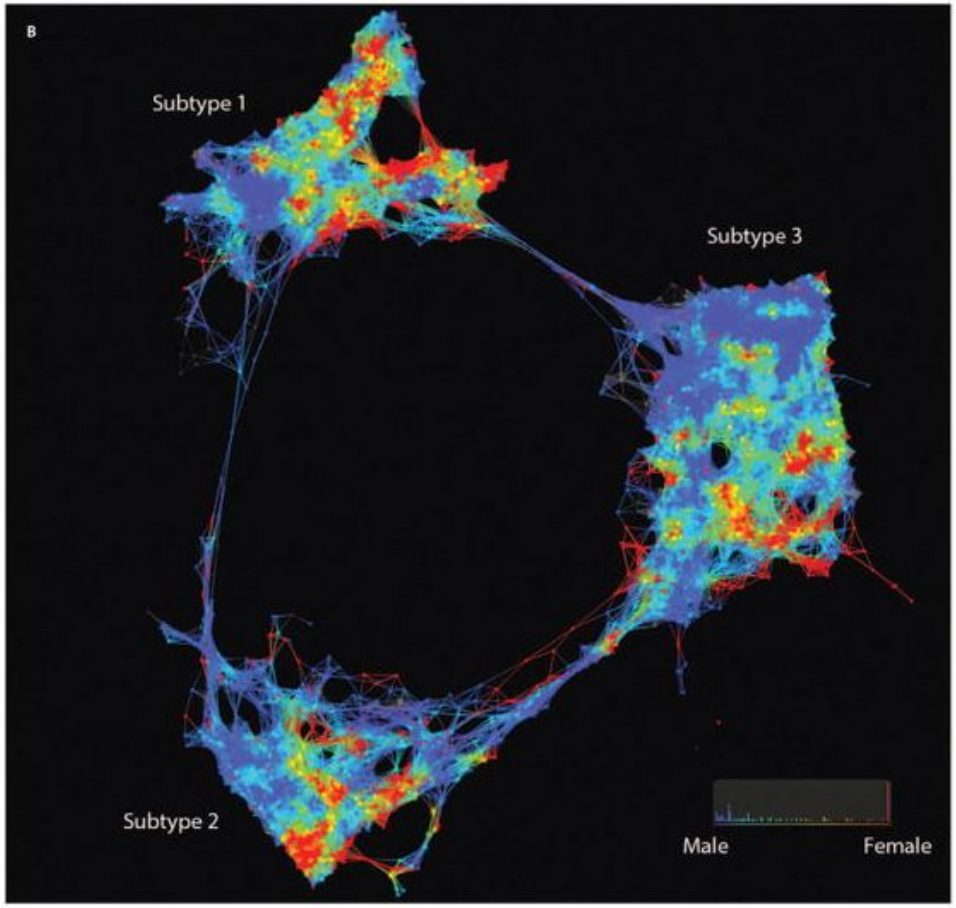
- Density value
- Distance to a point p
- Coordinate values:
  $f(x) = x_i : i \in \{1,...,d\}$



Poses



Identification of type 2 diabetes subgroups through topological analysis of patient similarity



Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine, 7(311), 311ra174.

Topological data analysis

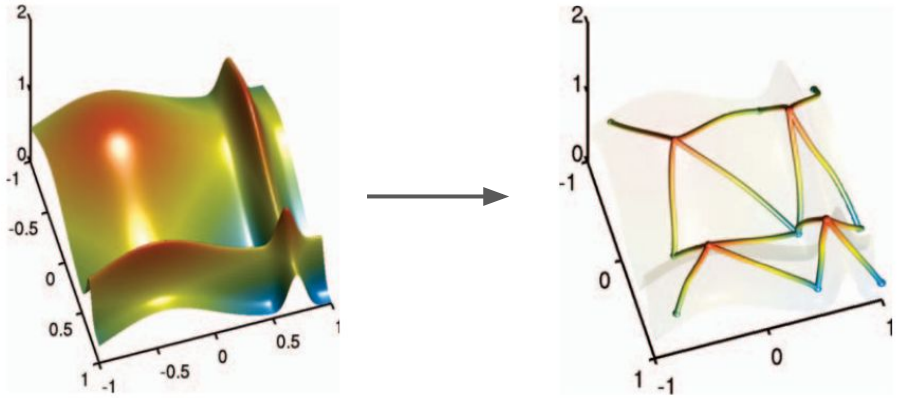
- Shape of data

- Extremal values of density function

Extremal values of density

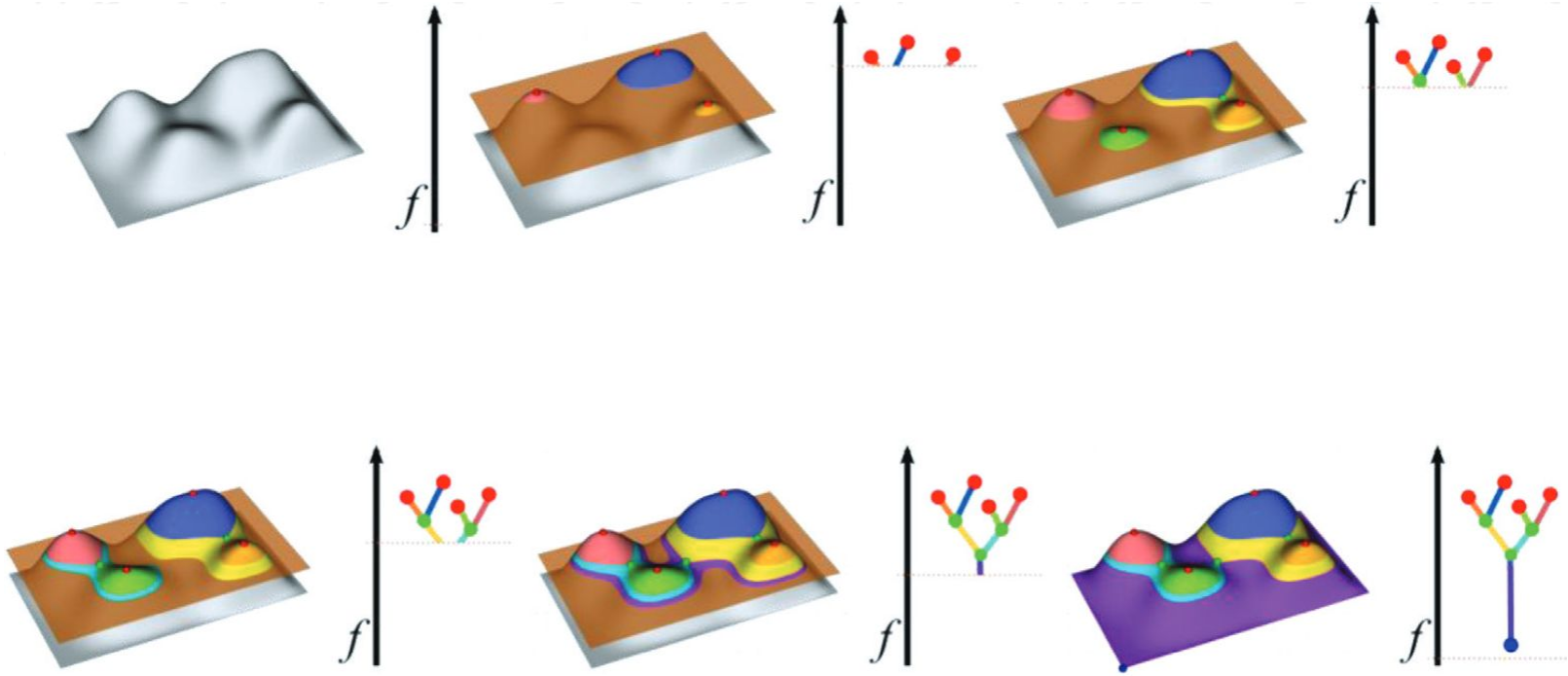
Interested in the Topology of scalar field f(**x**) = *y*

1. Understand extreme output values *y* (number & location)
2. Their connection in space
3. Which combinations of inputs **x** are responsible for which output *y*

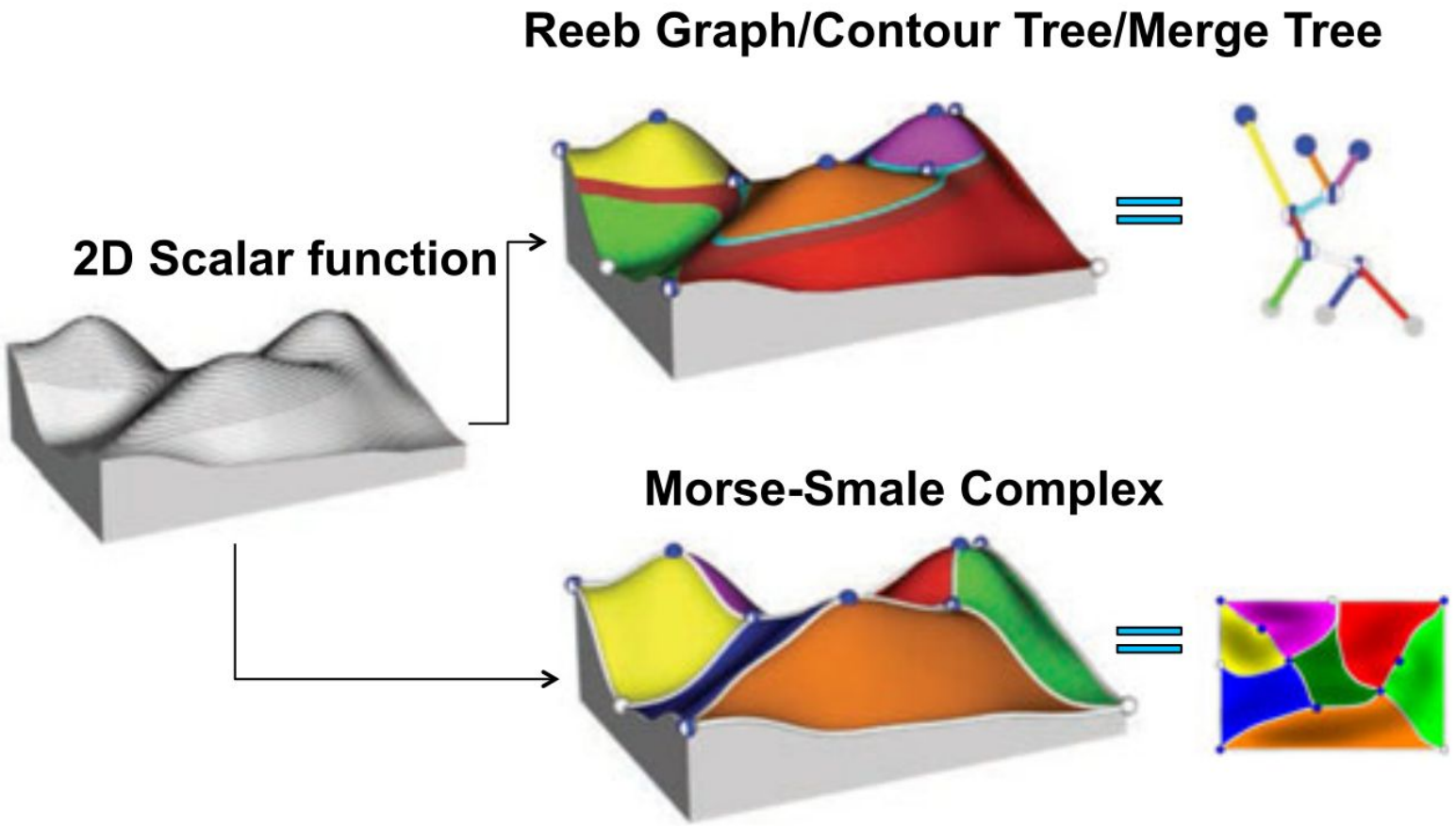


Cluster tree: density peaks in data

- Compact tree visualisation for arbitrary high dimensions
- Interesting for cluster analysis



Hierarchy of data



S. Liu, D. Maljovec, B. Wang, P. Bremer and V. Pascucci, "Visualizing High-Dimensional Data: Advances in the Past Decade," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 3, pp. 1249-1268, 1 March 2017, doi: 10.1109/TVCG.2016.2640960.

Hierarchy of data

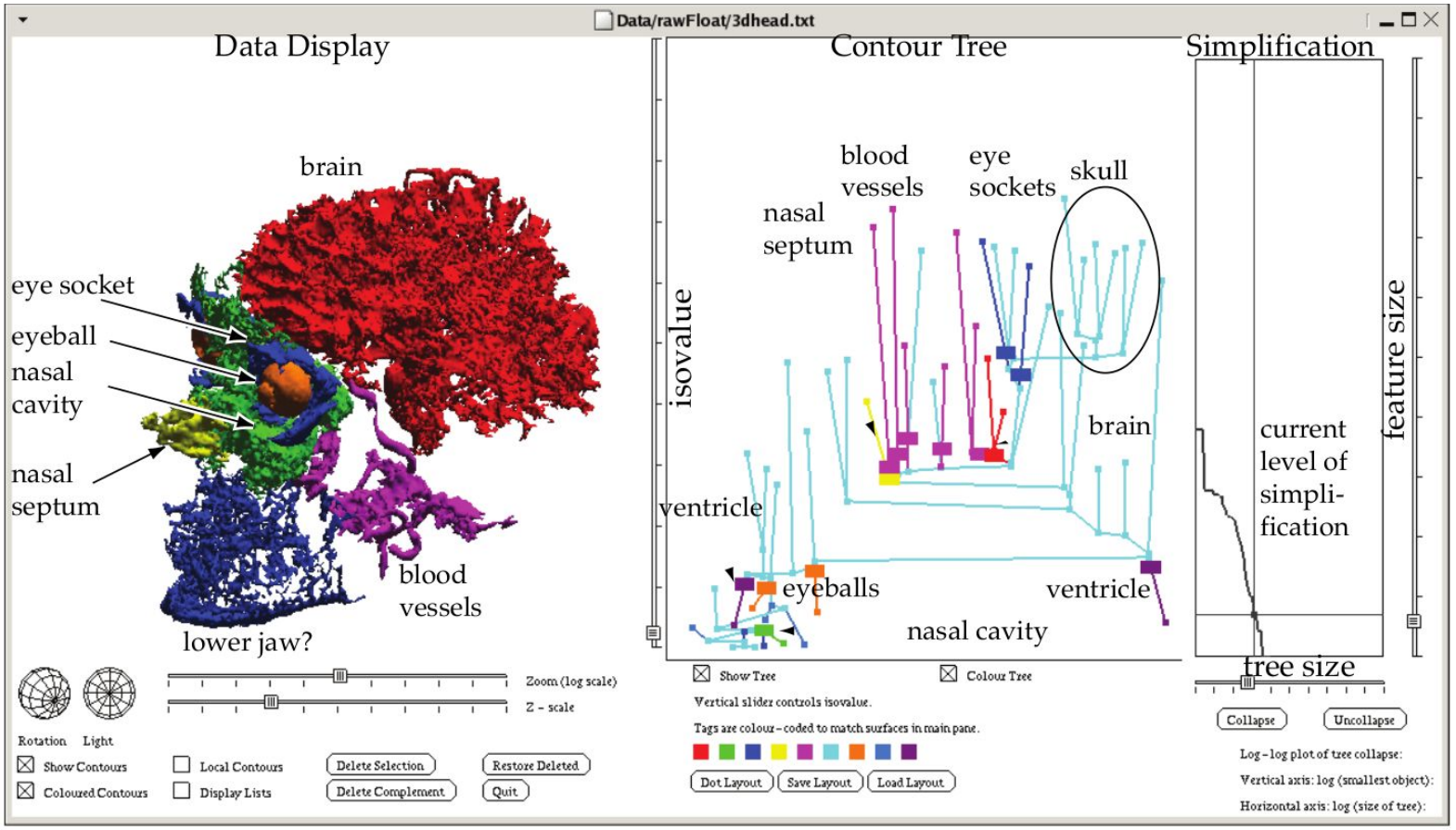
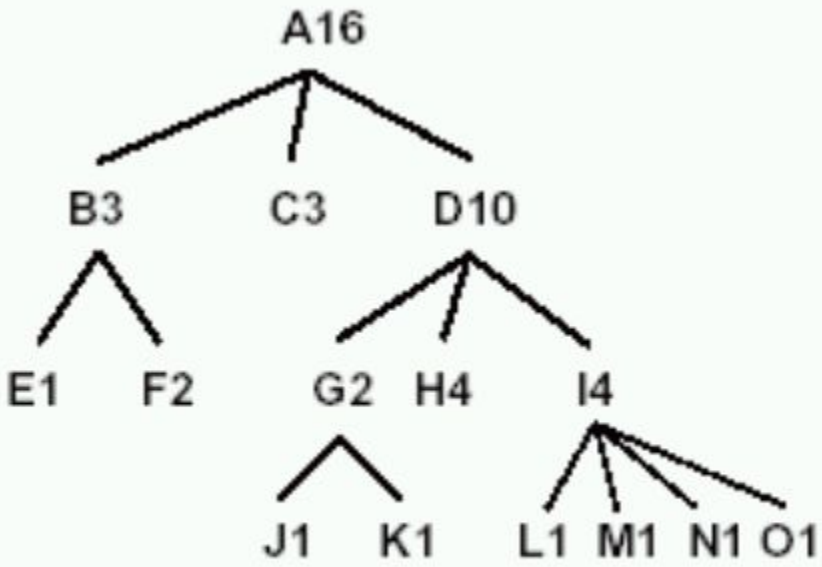


Figure 8.6. The flexible isosurfaces idiom uses the simplified contour tree of under 100 edges to help users identify meaningful structure. From [Carr et al. 04, Figure 1].

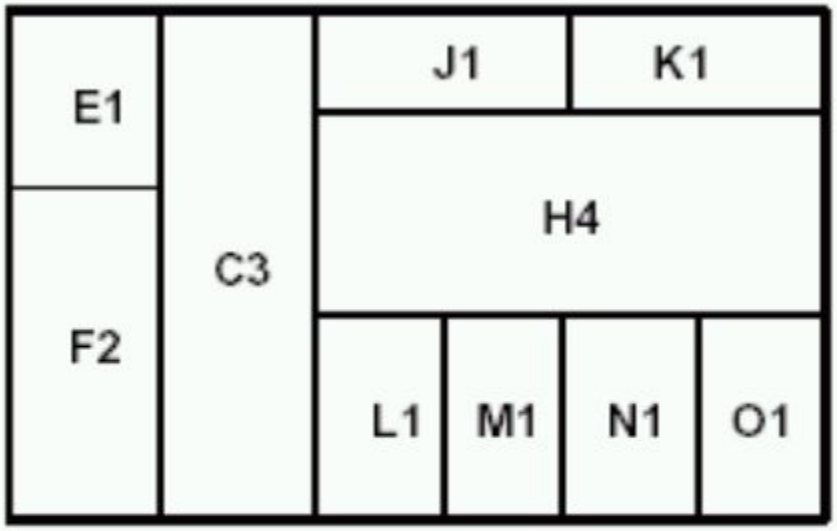
Hierarchical key-value structure:

Treemaps

Tree-map



Node and link diagram



Treemap

Tree-map

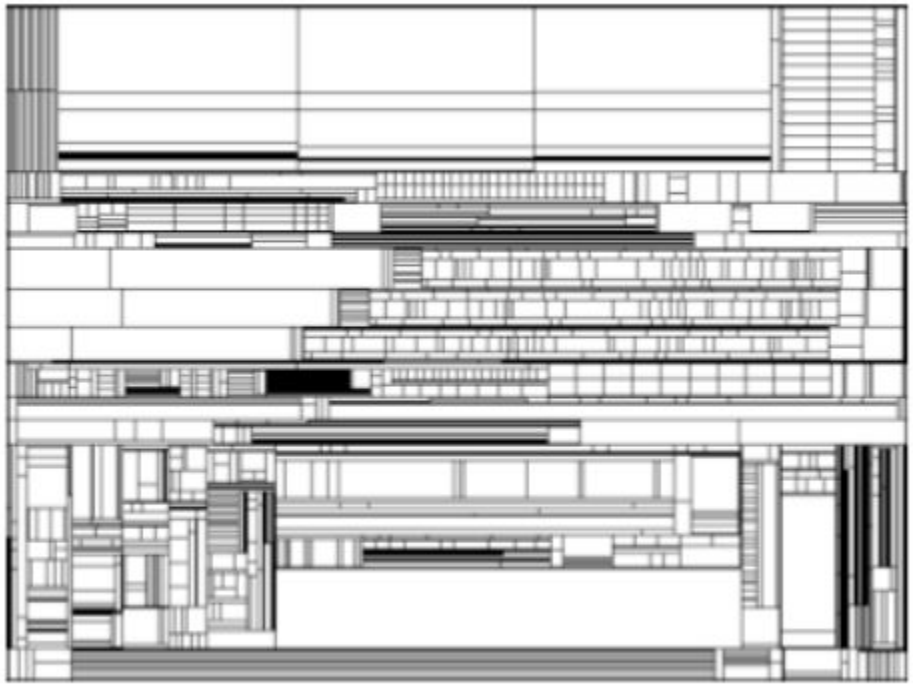


Figure 2. Treemap of file system

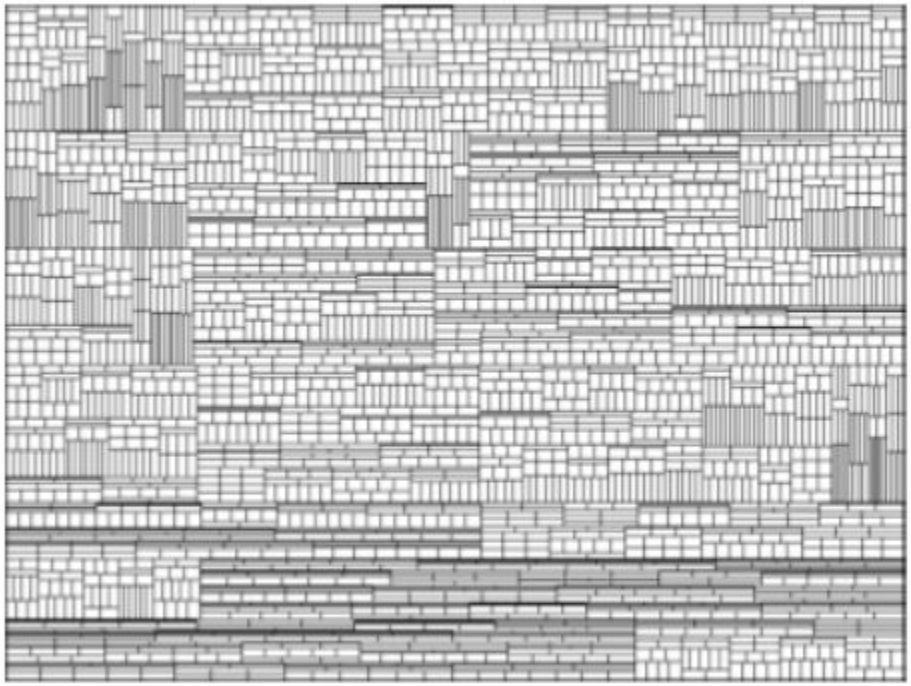


Figure 3. Treemap of organization

Cushion Tree-maps

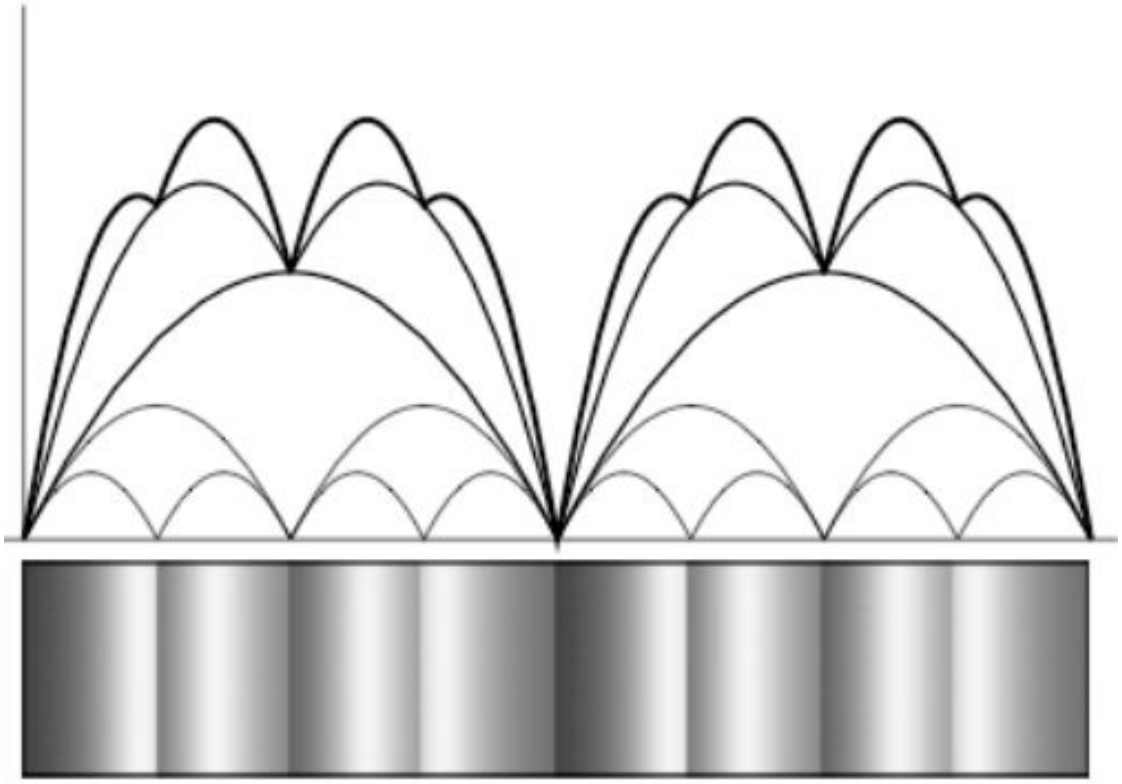
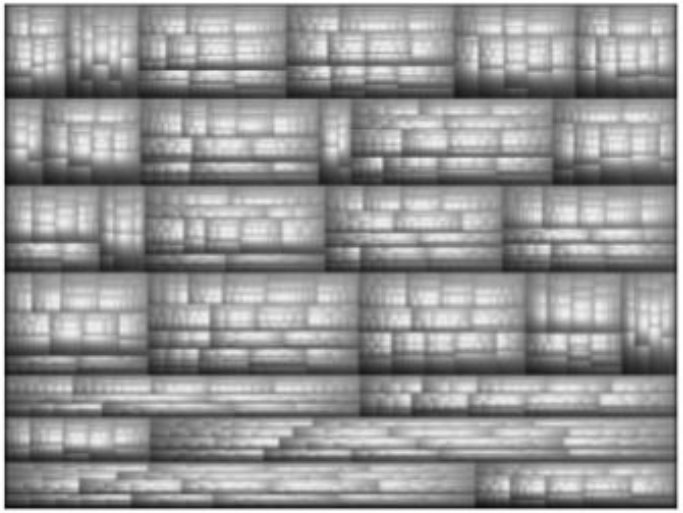


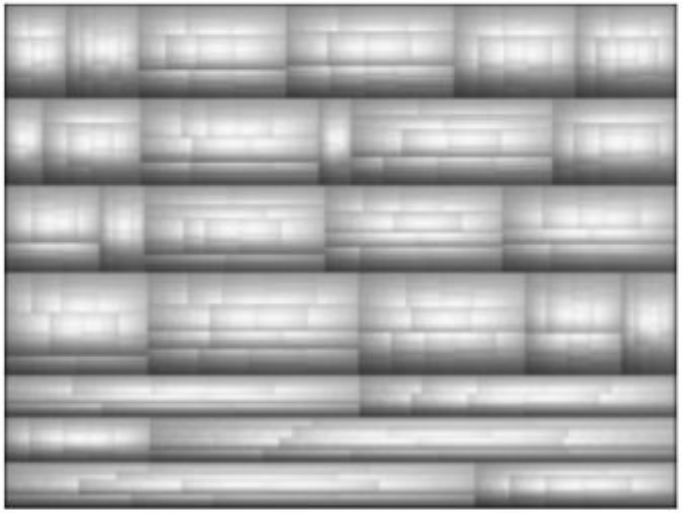
Figure 4. Binary subdivision of interval

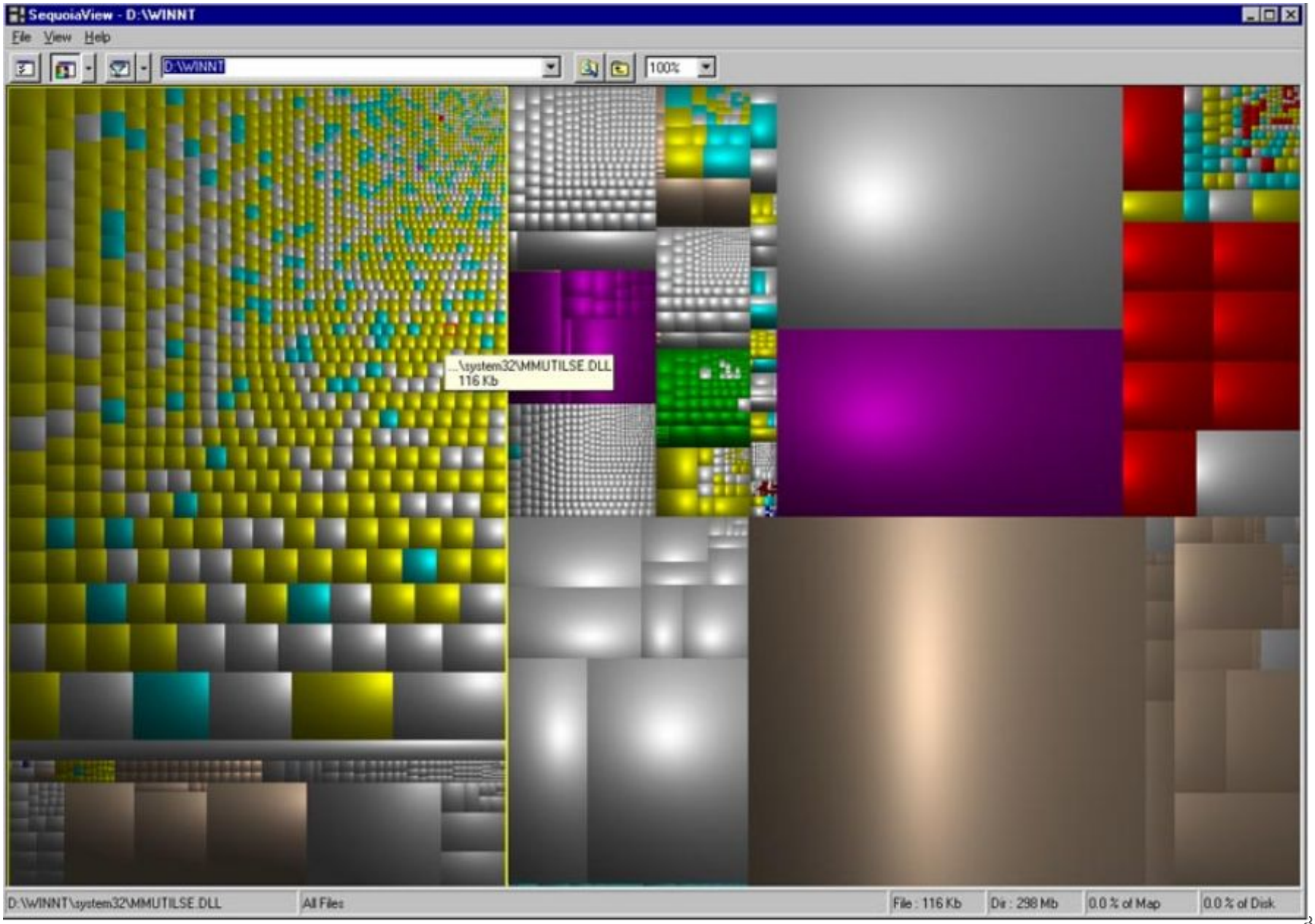
Cushion Tree-maps

$h = 0.5, f = 1$

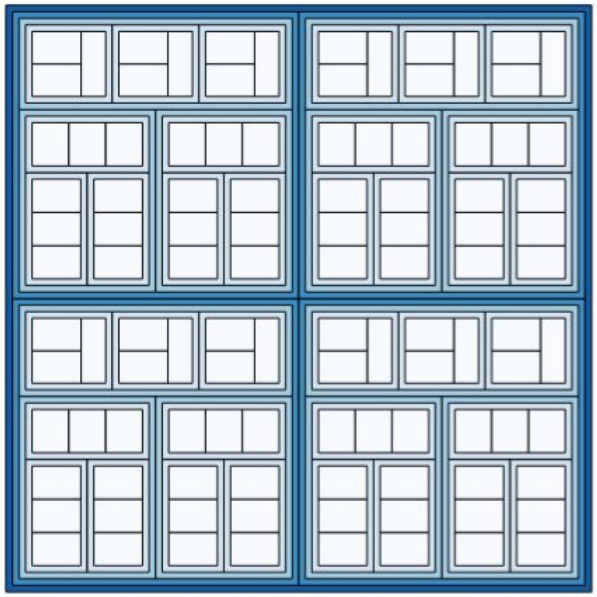


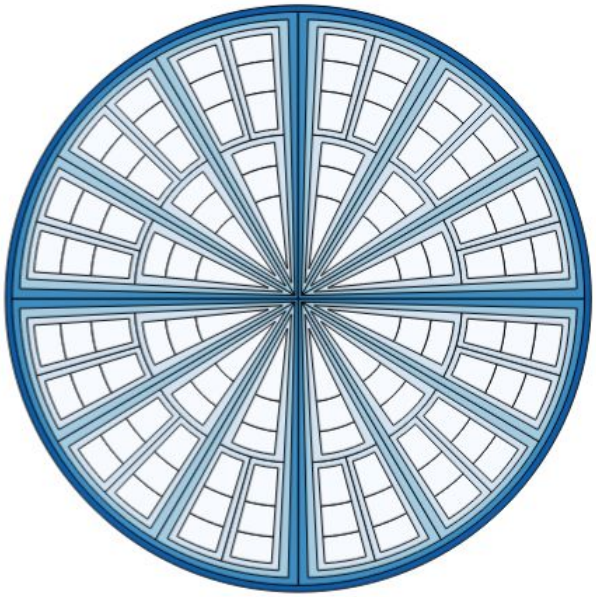
$h = 0.5, f = 0.75$

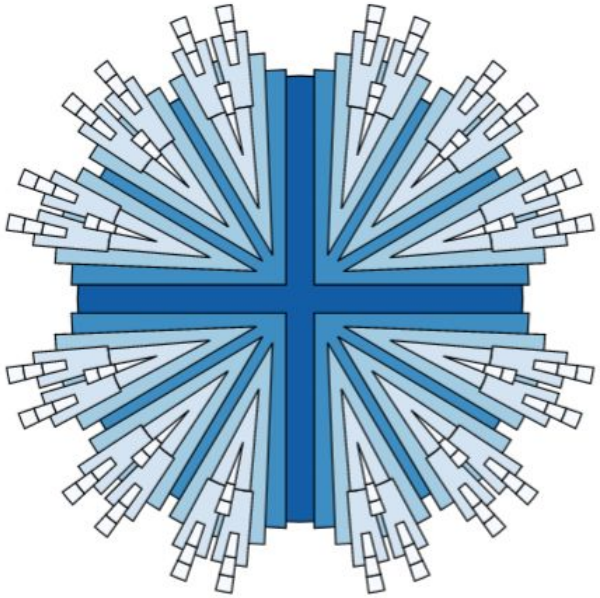




Rectilinear & radial layouts

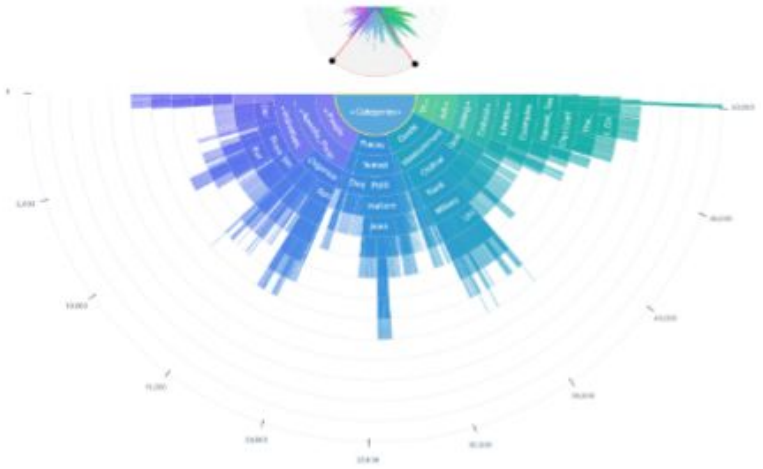




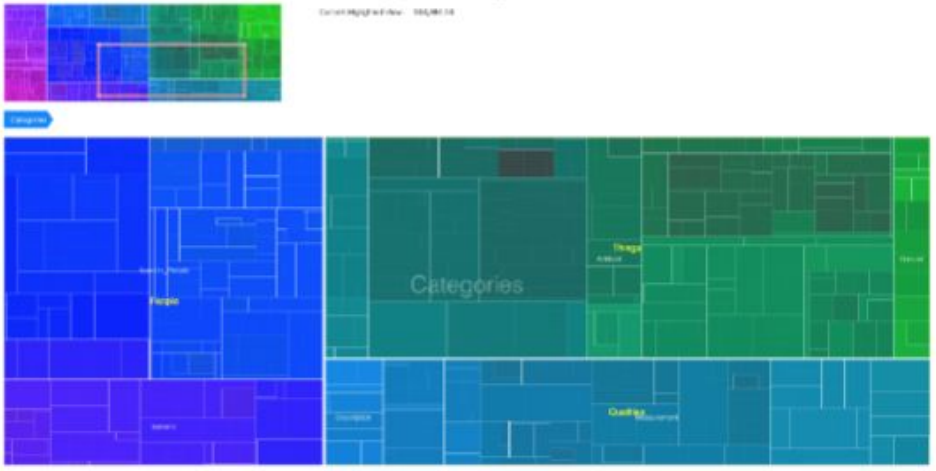


Rectilinear & radial layouts

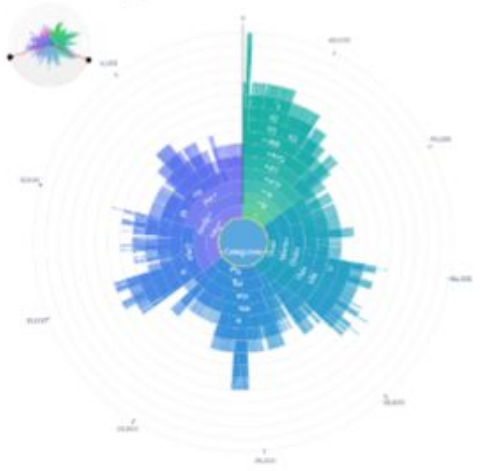
(a) Icicle plot



(b) Sundown chart



(c) Treemap



(d) Sunburst chart