High dimensionality

Evgeny Maksakov

CS533C
Department of Computer Science
UBC

Today

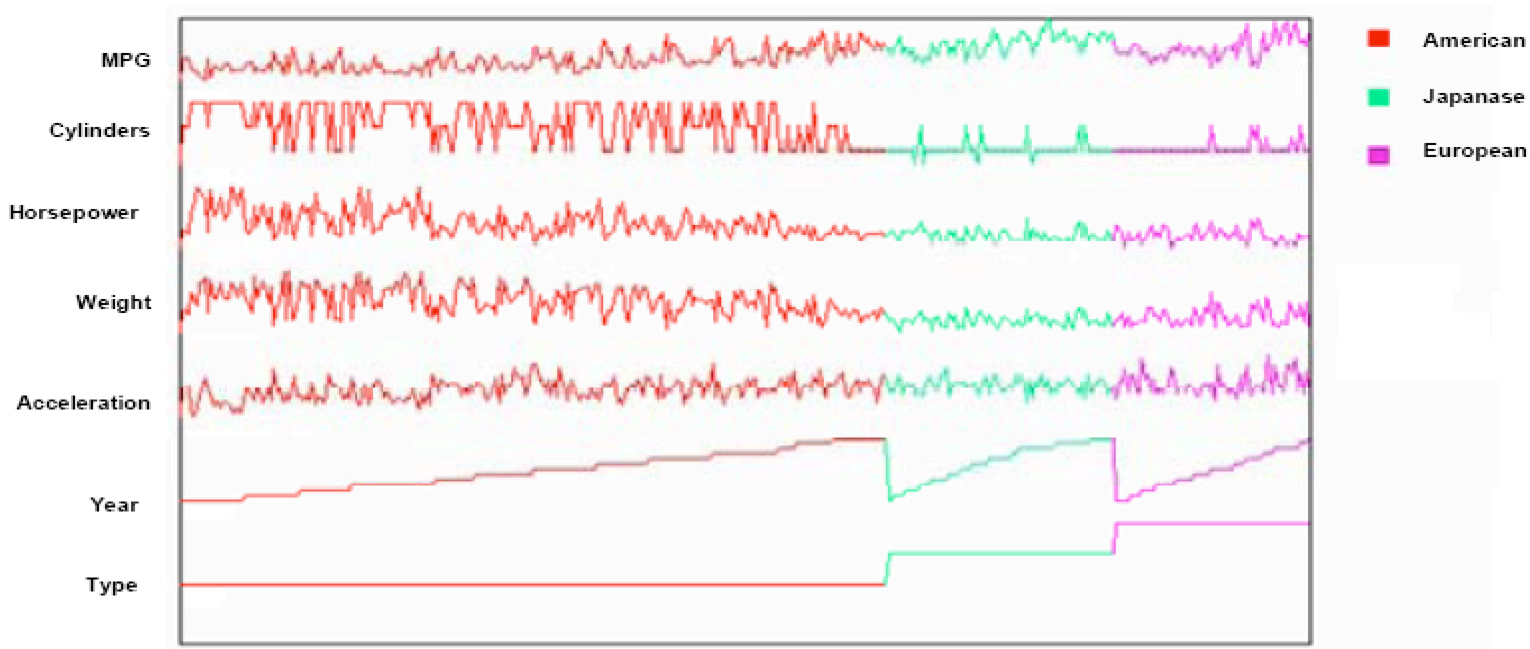
- Problem Overview
- Direct Visualization Approaches
  - Dimensional anchors
  - Scagnostic SPLOMs
- Nonlinear Dimensionality Reduction
  - Locally Linear Embedding and Isomaps
  - Charting manifold

Problems with visualizing high dimensional data

- Visual cluttering
- Clarity of representation
- Visualization is time consuming

Classical methods

Multiple Line Graphs



Pictures from Patrick Hoffman et al. (2000)

Multiple Line Graphs

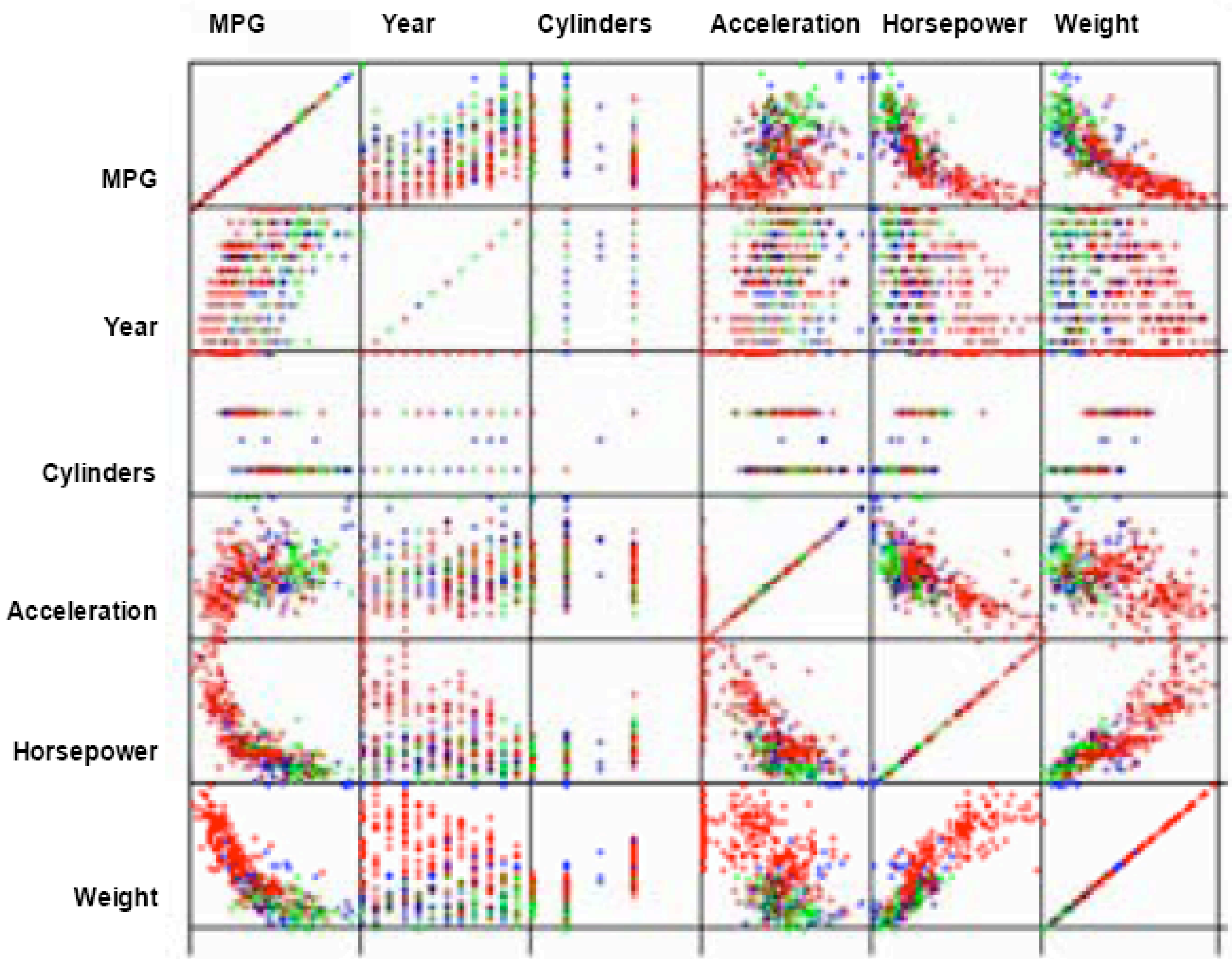
Advantages and disadvantages:

- Hard to distinguish dimensions if multiple line graphs overlaid

- Each dimension may have different scale that should be shown

- More than 3 dimensions can become confusing

Scatter Plot Matrices



Pictures from Patrick Hoffman et al. (2000)

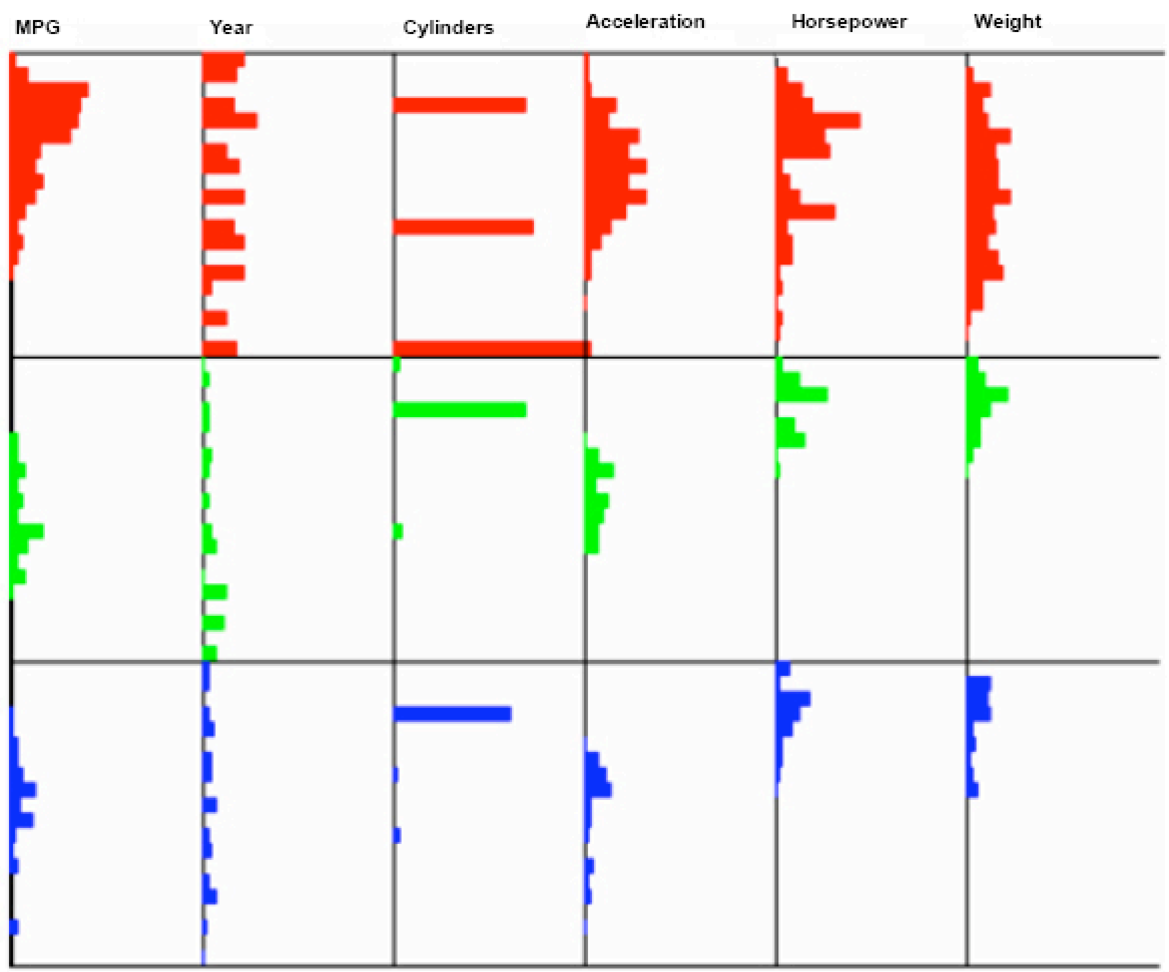
Scatter Plot Matrices

Advantages and disadvantages:

+ Useful for looking at all possible two-way interactions between dimensions

- Becomes inadequate for medium to high dimensionality

Bar Charts, Histograms



Pictures from Patrick Hoffman et al. (2000)

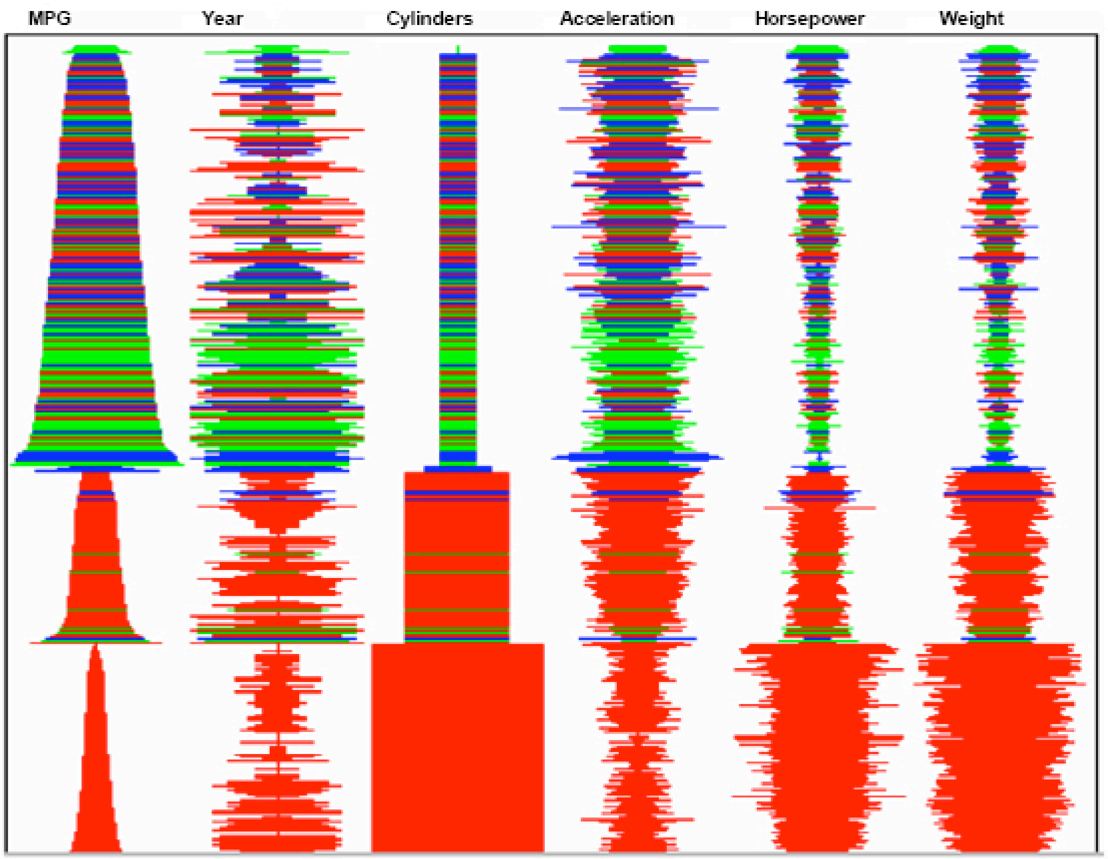
Bar Charts, Histograms

Advantages and disadvantages:

+ Good for small comparisons

- Contain little data

Survey Plots



Pictures from Patrick Hoffman et al. (2000)

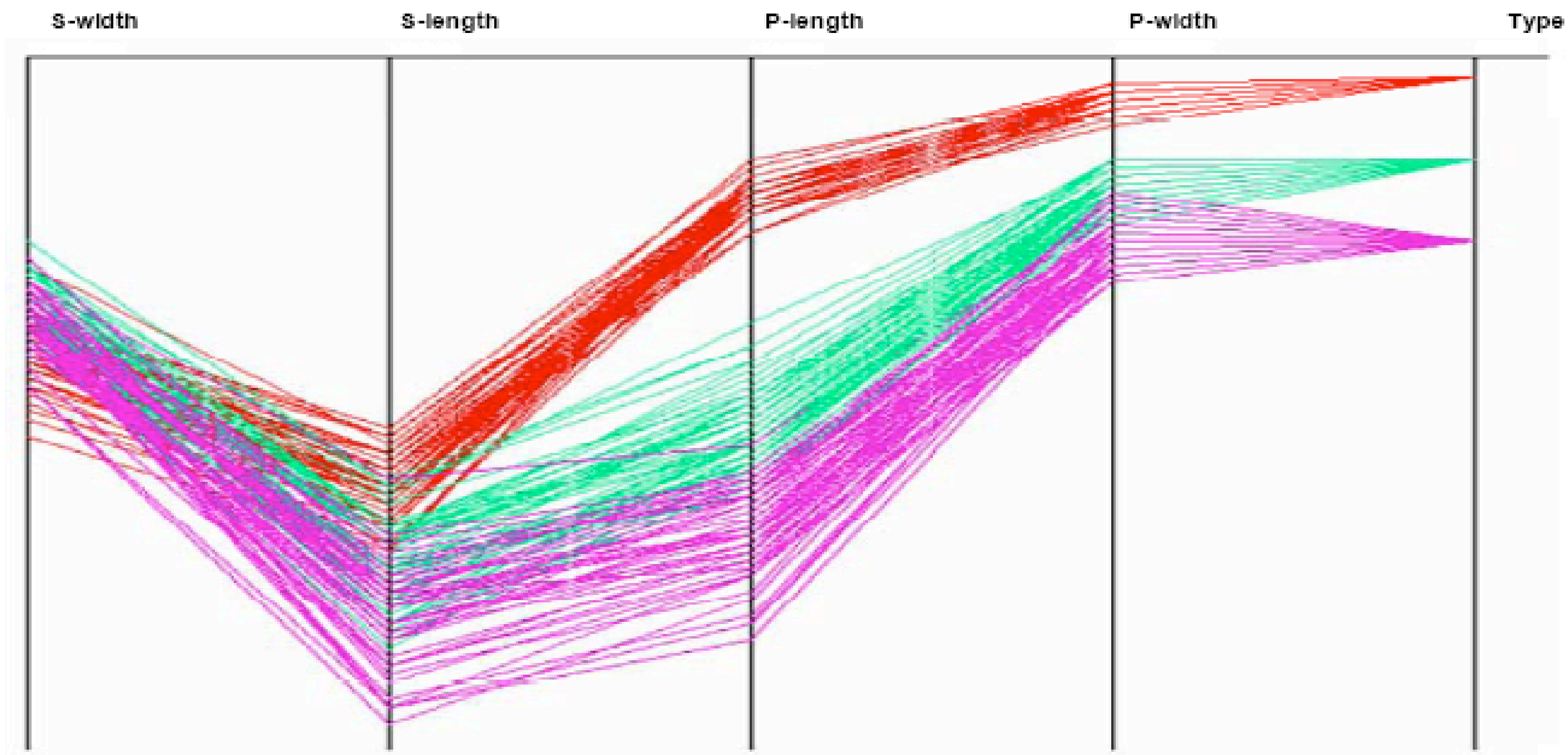
Survey Plots

Advantages and disadvantages:

+ allows to see correlations between any two variables when the data is sorted according to one particular dimension

- can be confusing

Parallel Coordinates



Pictures from Patrick Hoffman et al. (2000)

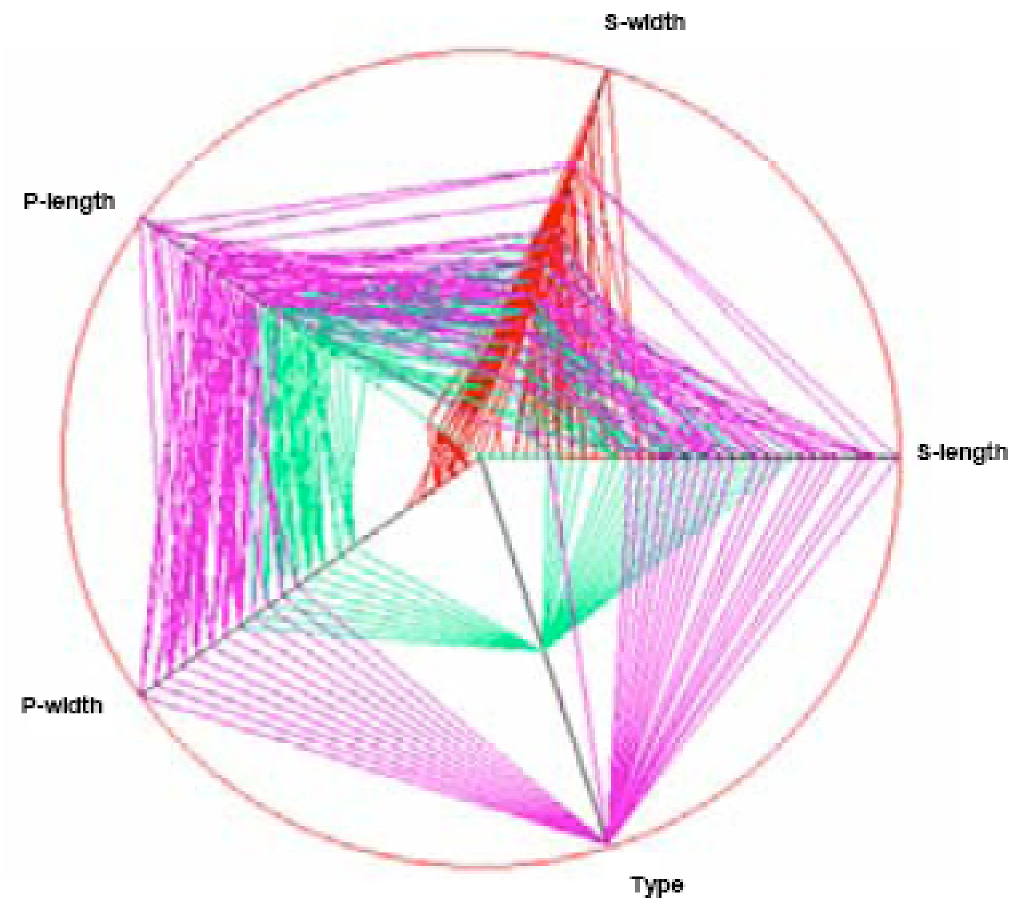
Parallel Coordinates

Advantages and disadvantages:

+ Many connected dimensions are seen in limited space
+ Can see trends in data

- Become inadequate for very high dimensionality
- Cluttering

Circular Parallel Coordinates



Pictures from Patrick Hoffman et al. (2000)

Circular Parallel Coordinates

Advantages and disadvantages:

+ Combines properties of glyphs and parallel coordinates making pattern recognition easier

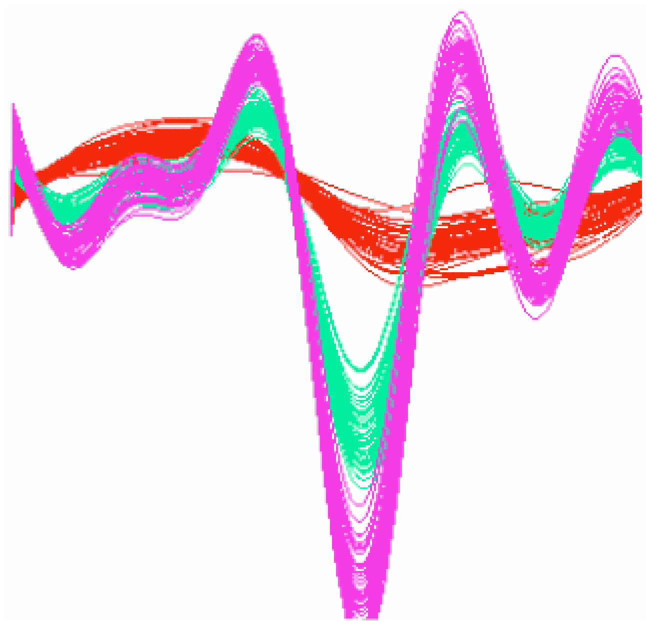
+ Compact

- Cluttering near center

- Harder to interpret relations between each pair of dimensions than parallel coordinates

Andrews' Curves

$$f(t) = \frac{x1}{\sqrt{2}} + x2 * \sin(t) + x3 * \cos(t) + x4 * \sin(2t) + x5 * \cos(2t) + \ldots$$



Pictures from Patrick Hoffman et al. (2000)

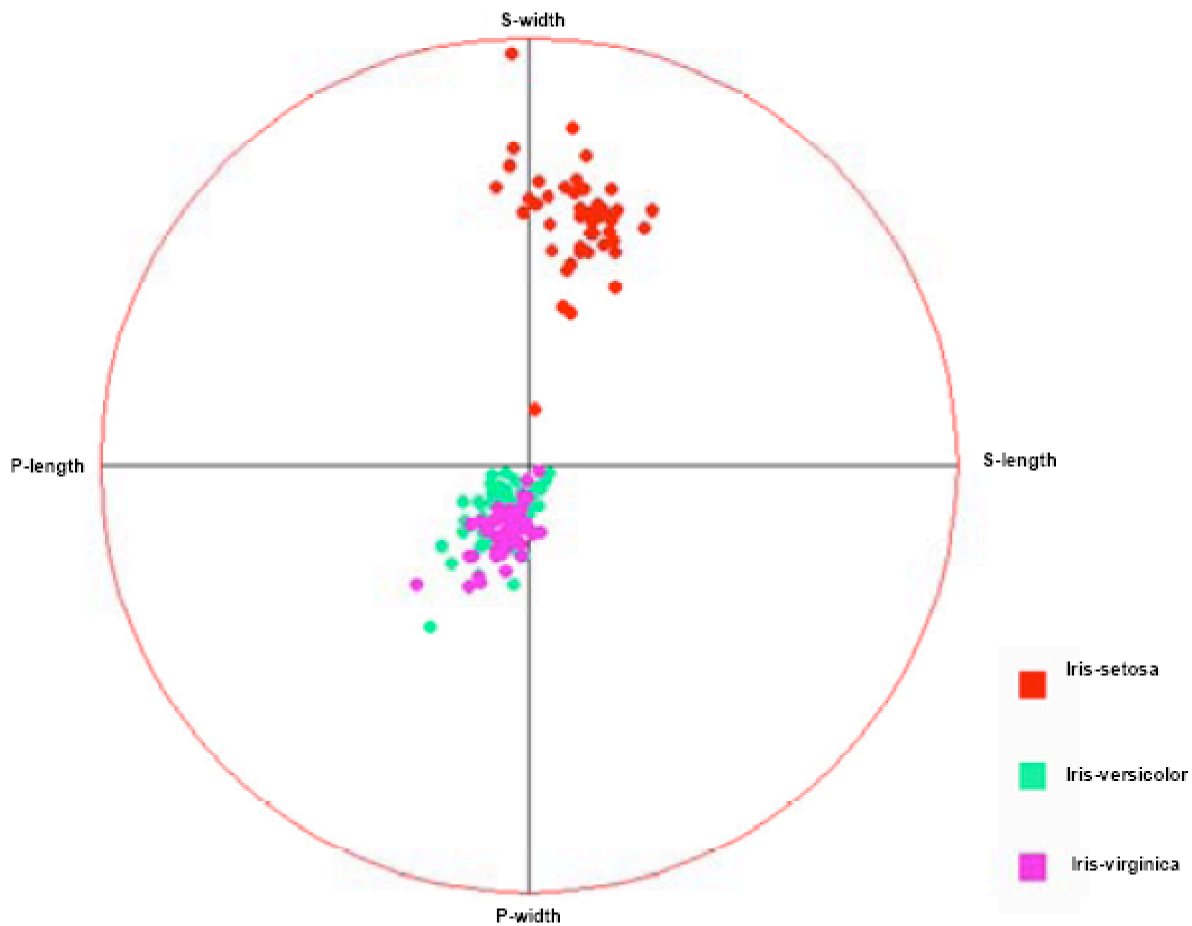
Andrews' Curves

Advantages and disadvantages:

+ Allows to draw virtually unlimited dimensions

- Hard to interpret

Radviz



Radviz employs spring model

Pictures from Patrick Hoffman et al. (2000)

Radviz

Advantages and disadvantages:

+ Good for data manipulation

+ Low cluttering

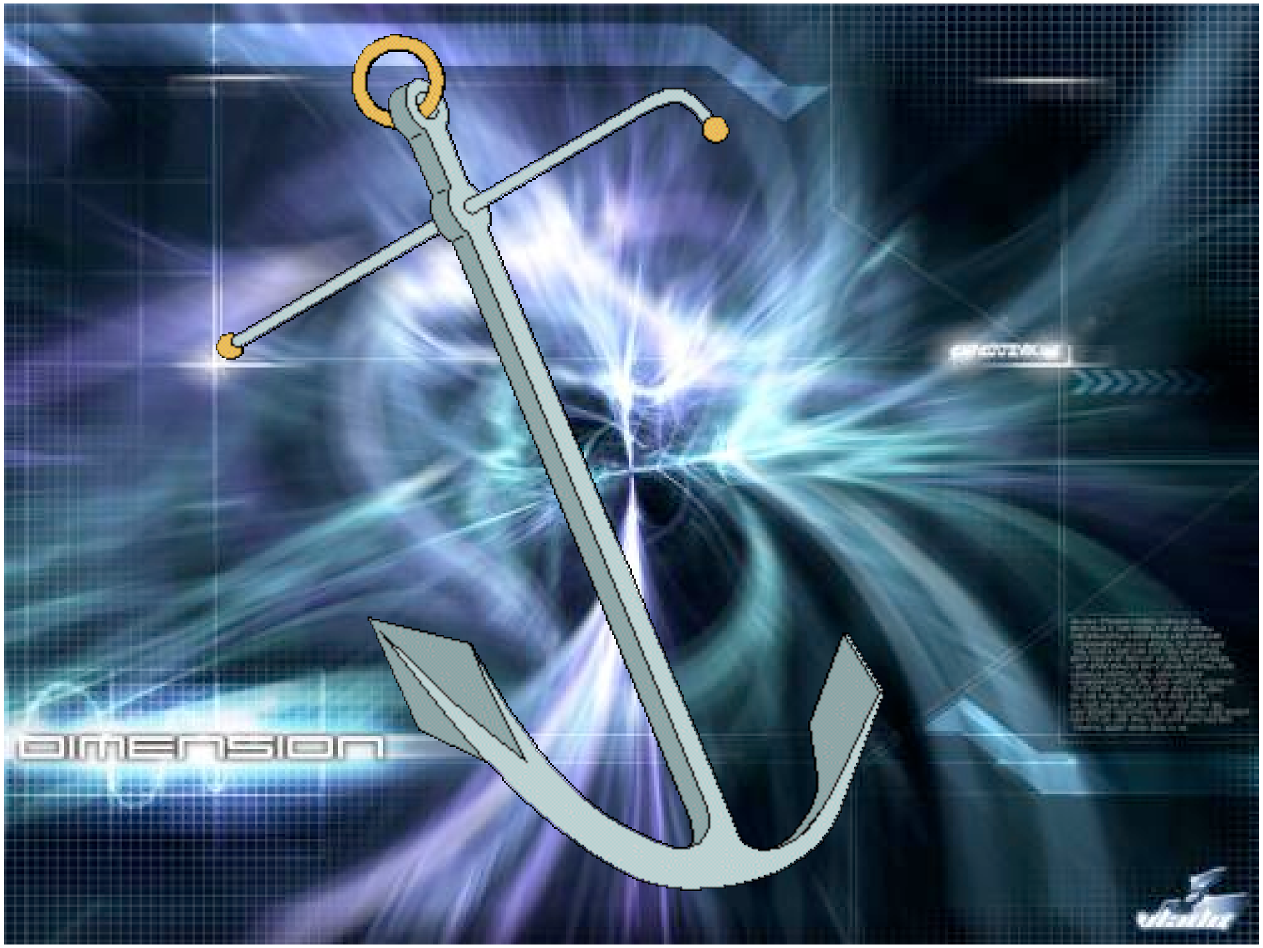
- Cannot show quantitative data

- High computational complexity

Dimensional Anchors

Attempt to Generalize Visualization Methods for High Dimensional Data

What is dimensional anchor?



Picture from members.fortunecity.com/agreeve/seacol.htm & http://kresby.grafika.cz/data/media/46/dimension.jpg_middle.jpg

What is dimensional anchor?

Nothing like that

DA is just an axis line… ϑ

Anchorpoints are coordinates… ϑ

Parameters of DA

Scatterplot features

- Size of the scatter plot points
- Length of the perpendicular lines extending from individual anchor points in a scatter plot
- Length of the lines connecting scatter plot points that are associated with the same data point

Parameters of DA

Survey plot feature

4. Width of the rectangle in a survey plot

Parallel coordinates features

5. Length of the parallel coordinate lines

6. Blocking factor for the parallel coordinate lines

Parameters of DA

Radviz features

7. Size of the radviz plot point

8. Length of “spring” lines extending from individual anchor points of radviz plot

9. Zoom factor for the “spring” constant K

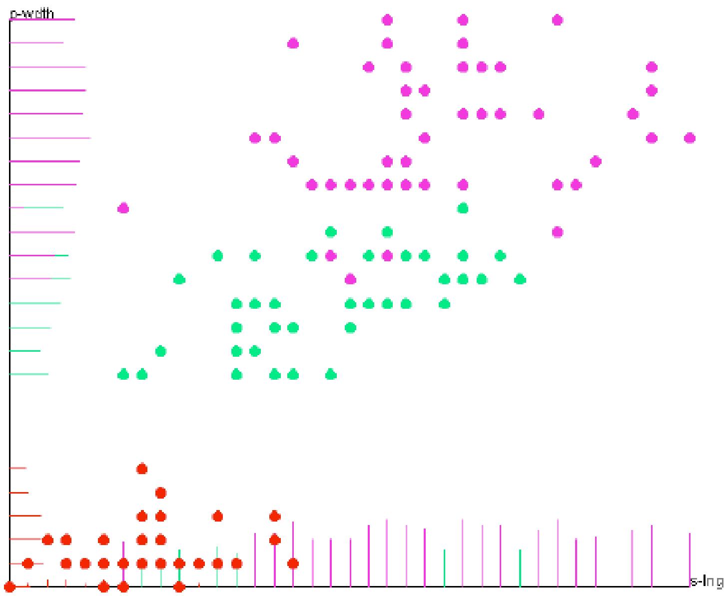
DA Visualization Vector

$$P\ (p_1,p_2,p_3,p_4,p_5,p_6,p_7,p_8,p_9)$$

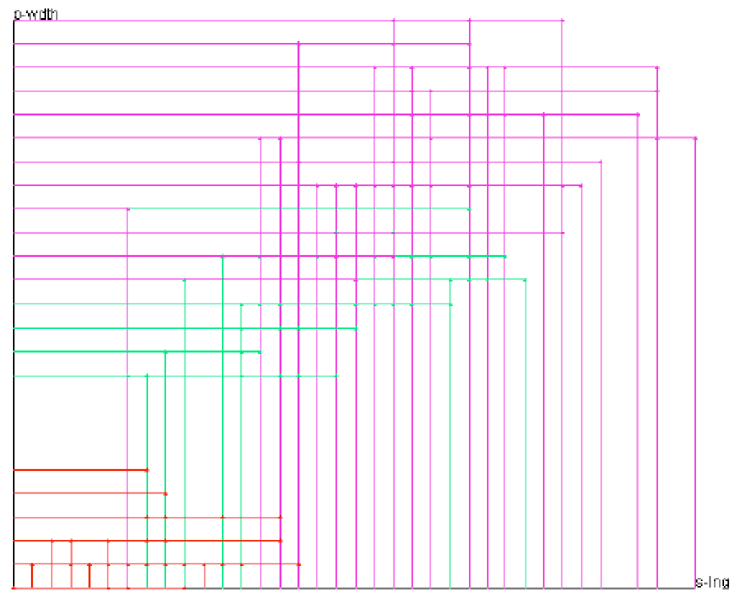
DA describes visualization for any combination of:

- Parallel coordinates
- Scatterplot matrices
- Radviz
- Survey plots (histograms)
- Circle segments

Scatterplots



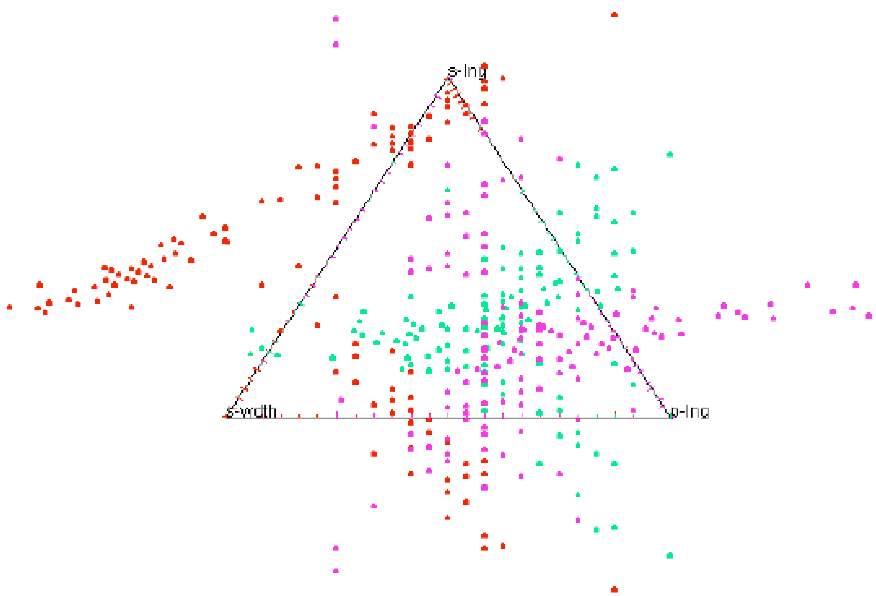
2 DAs, P = (0.8, 0.2, 0, 0, 0, 0, 0, 0, 0)



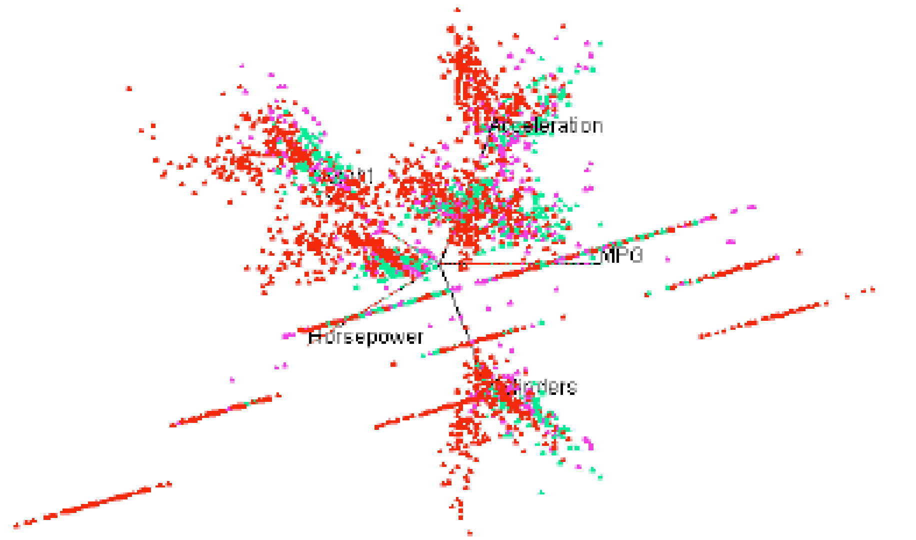
2 DAs, P = (0.1, 1.0, 0, 0, 0, 0, 0, 0, 0)

Picture from Patrick Hoffman et al. (1999)

Scatterplots with other layouts



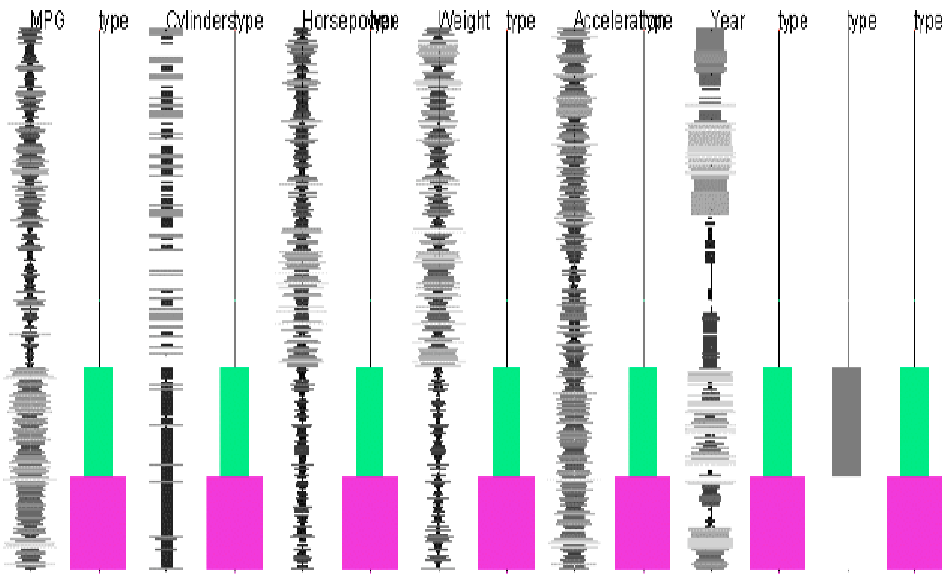
3 DAs, P = (0.6, 0, 0, 0, 0, 0, 0, 0, 0)



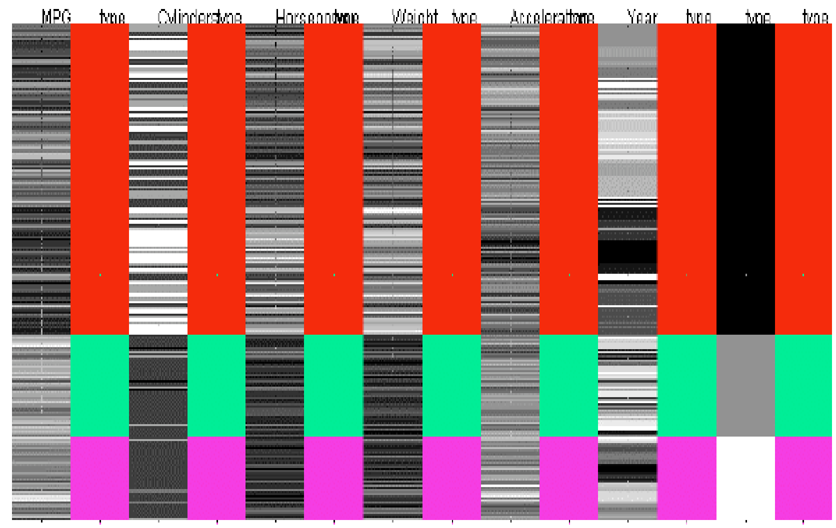
5 DAs, P = (0.5, 0, 0, 0, 0, 0, 0, 0, 0)

Picture from Patrick Hoffman et al. (1999)

Survey Plots



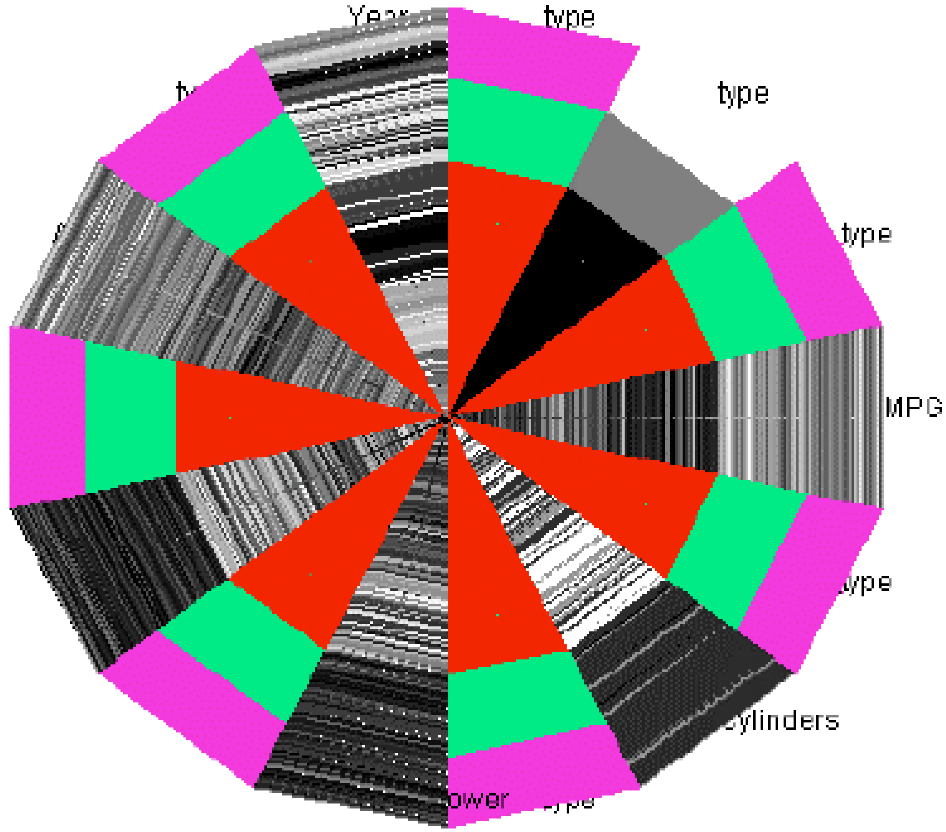
P = (0, 0, 0, 0.4, 0, 0, 0, 0, 0)



P = (0, 0, 0, 1.0, 0, 0, 0, 0, 0)

Picture from Patrick Hoffman et al. (1999)

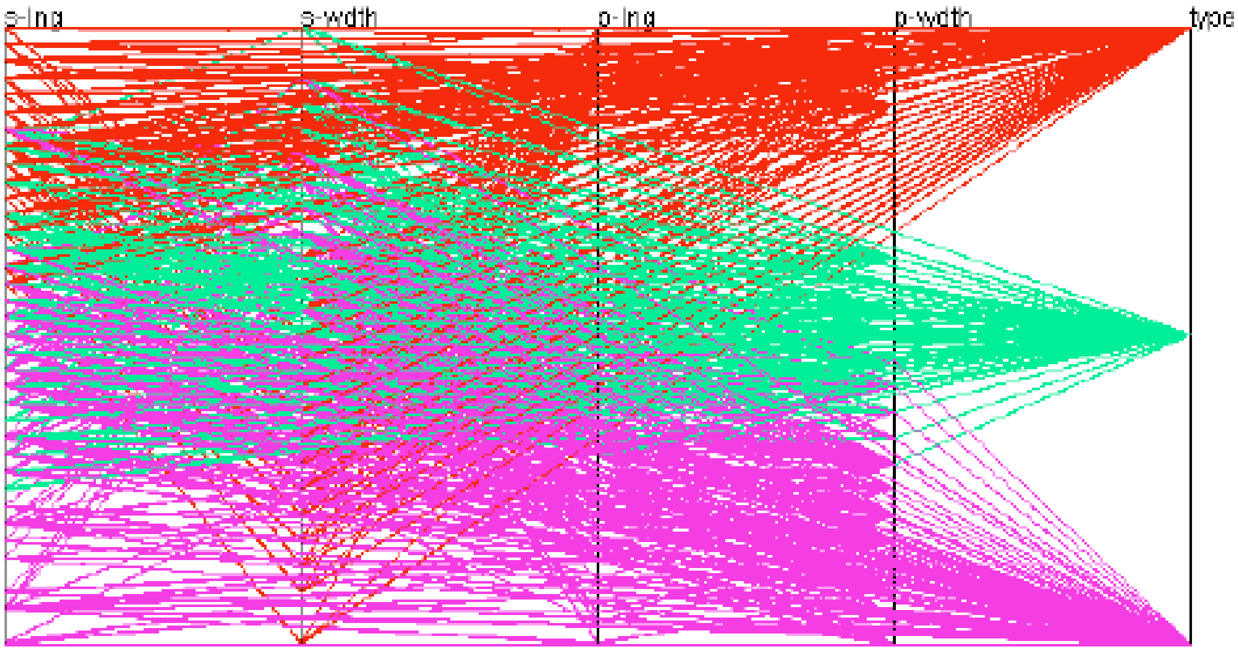
Circular Segments



P = (0, 0, 0, 1.0, 0, 0, 0, 0, 0)

Picture from Patrick Hoffman et al. (1999)

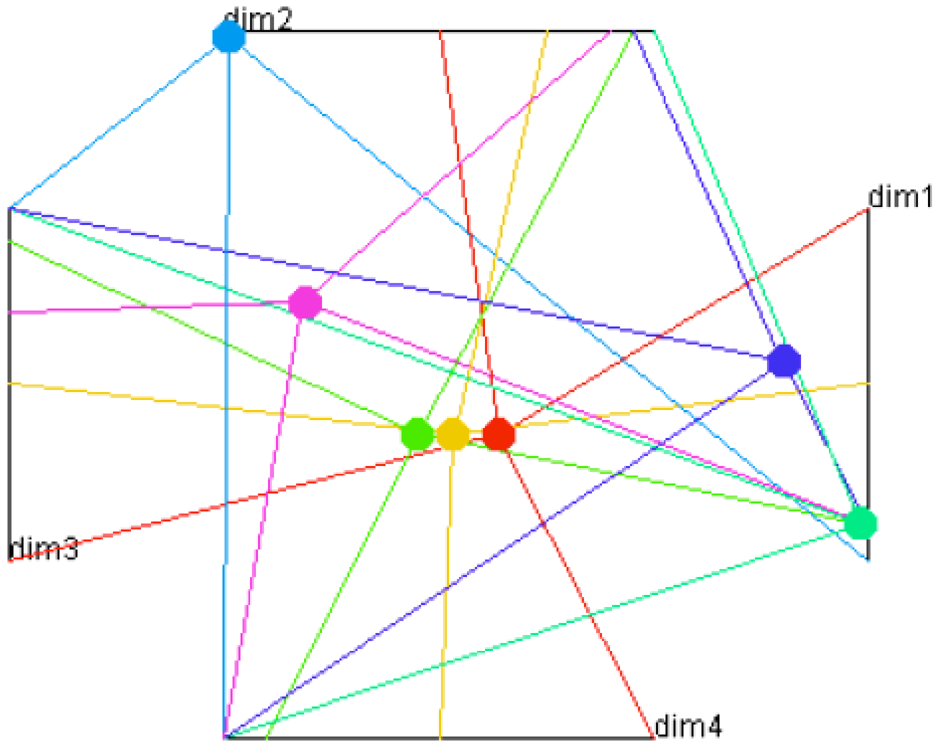
Parallel Coordinates



P = (0, 0, 0, 0, 1.0, 1.0, 0, 0, 0)

Picture from Patrick Hoffman et al. (1999)

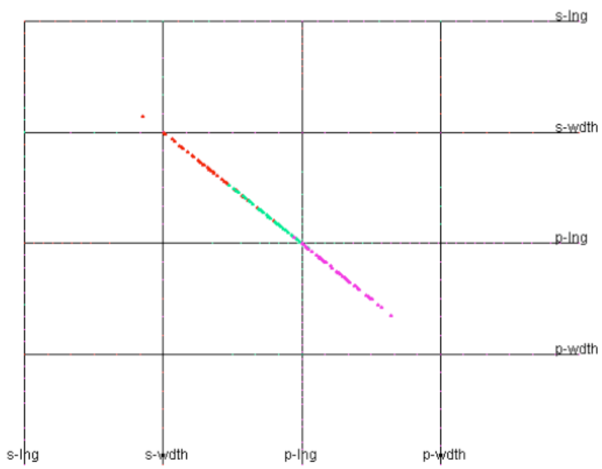
Radviz like visualization



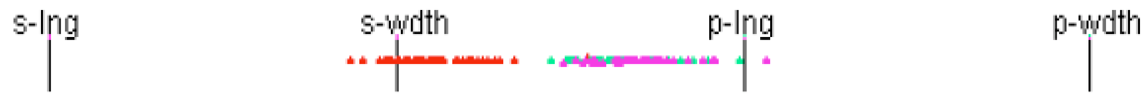
P = (0, 0, 0, 0, 0, 0, 0.5, 1.0, 0.5)

Picture from Patrick Hoffman et al. (1999)

Playing with parameters



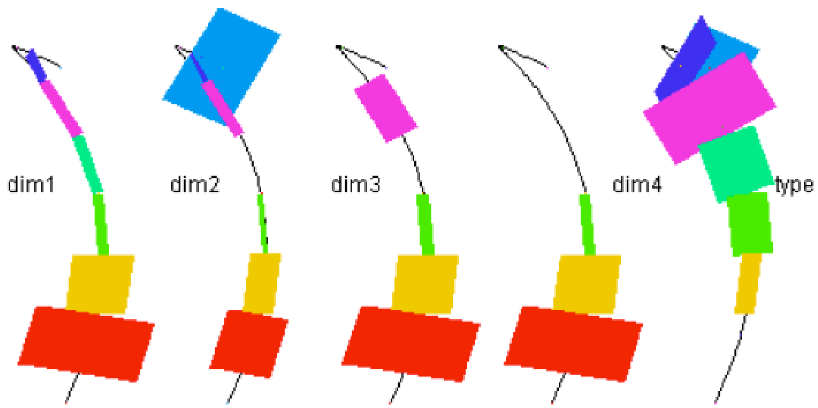
Crisscross layout with
P = (0, 0, 0, 0, 0, 0, 0.4, 0, 0.5)

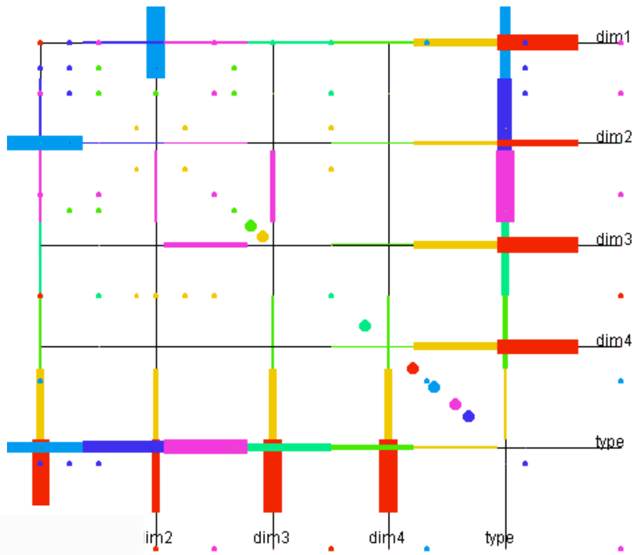


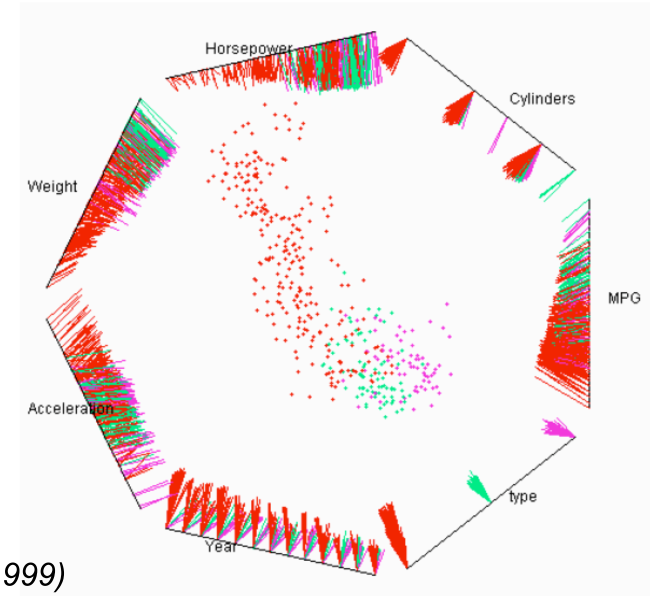
Parallel coordinates with
P = (0, 0, 0, 0, 0, 0, 0.4, 0, 0.5)

Pictures from Patrick Hoffman et al. (1999)

More?







Pictures from Patrick Hoffman et al. (1999)

Scatterplot Diagnostics

or

Scagnostics

Tukey's Idea of Scagnostics

- Take measures from scatterplot matrix
- Construct scatterplot matrix (SPLOM) of these measures
- Look for data trends in this SPLOM

Scagnostic SPLOM

Is like:

- Visualization of a set of pointers

Also:

- Set of pointers to pointers also can be constructed

Goal:

- To be able to locate unusual clusters of measures that characterize unusual clusters of raw scatterplots

Problems with constructing Scagnostic SPLOM

1) Some of Tukeys' measures presume underlying continuous empirical or theoretical probability function. It can be a problem for other types of data.

2) The computational complexity of some of the Tukey measures is $O(n\_)$.

Solution*

1. Use measures from the graph-theory.
   - Do not presume a connected plane of support
   - Can be metric over discrete spaces
2. Base the measures on subsets of the Delaunay triangulation
   - Gives O(nlog(n)) in the number of points
3. Use adaptive hexagon binning before computing to further reduce the dependence on *n*.
4. Remove outlying points from spanning tree

* *Leland Wilkinson et al. (2005)*

Properties of geometric graph for measures

- Undirected (edges consist of unordered pairs)
- Simple (no edge pairs a vertex with itself)
- Planar (has embedding in R2 with no crossed edges)
- Straight (embedded eges are straight line segments)
- Finite (*V* and *E* are finite sets)

Graphs that fit these demands:

- Convex Hull
- Alpha Hull
- Minimal Spanning Tree

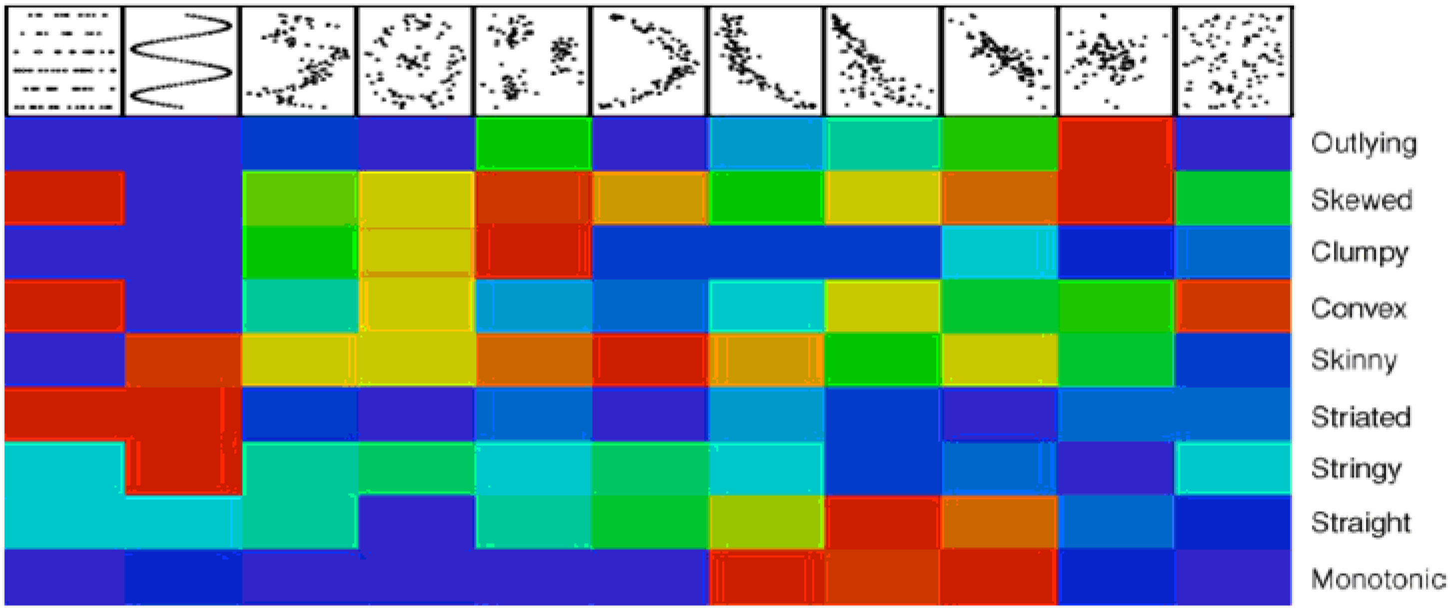
Measures:

- Length of en edge
- Length of a graph
- Look for a closed path (boundary of a polygon*)*
- Perimeter of a polygon
- Area of a polygon
- Diameter of a graph

Five interesting aspects of scattered points:

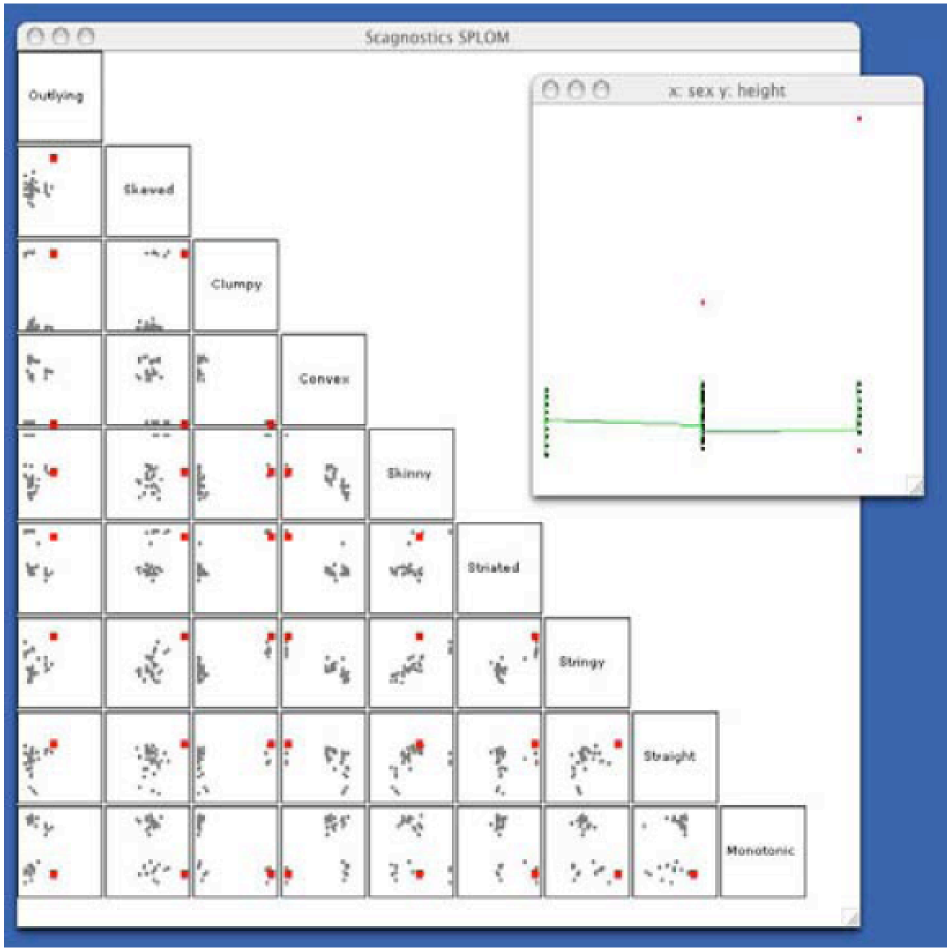
- ***Outliers***
  - Outlying
- ***Shape***
  - Convex
  - Skinny
  - Stringy
  - Straight
- ***Trend***
  - Monotonic
- ***Density***
  - Skewed
  - Clumpy
- ***Coherence***
  - Striated

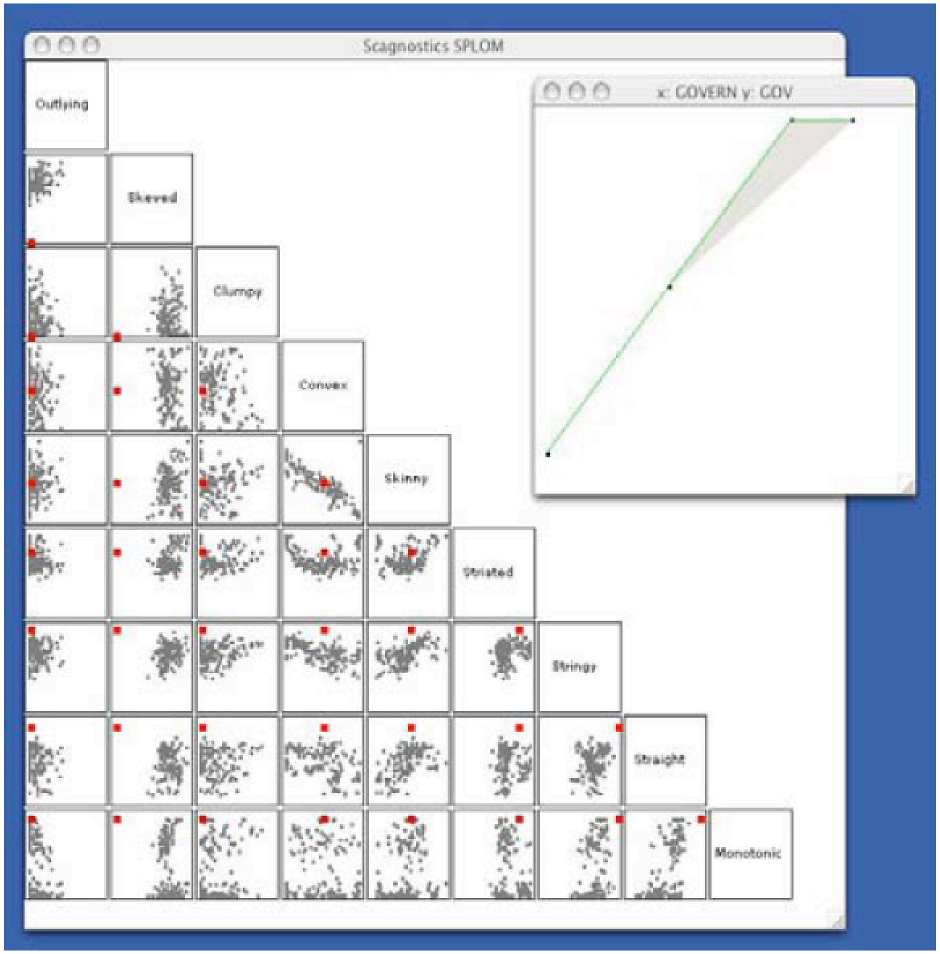
Classifying scatterplots



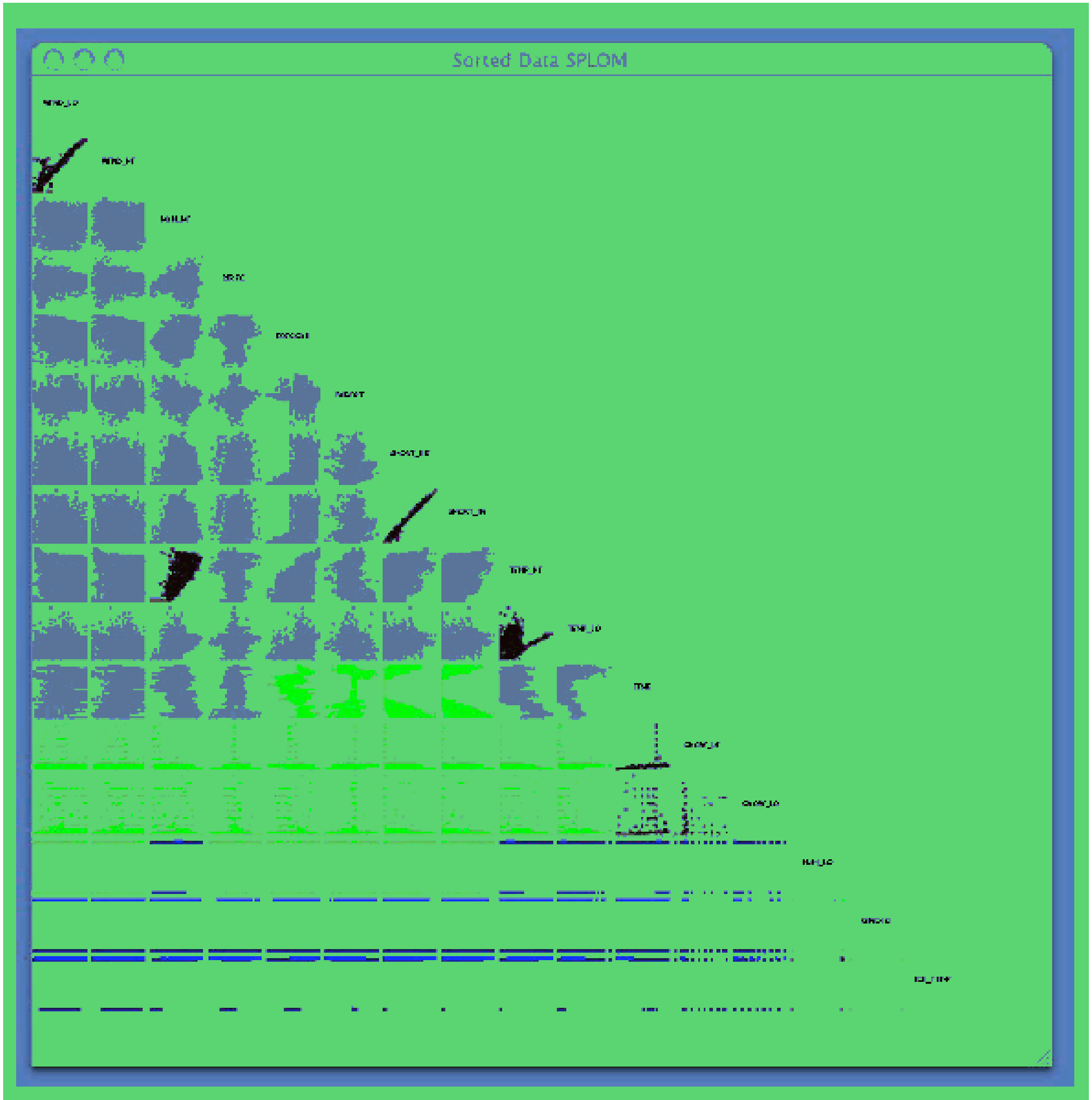
Picture from L. Wilkinson et al. (2005)

Looking for anomalies





Picture from L. Wilkinson et al. (2005)

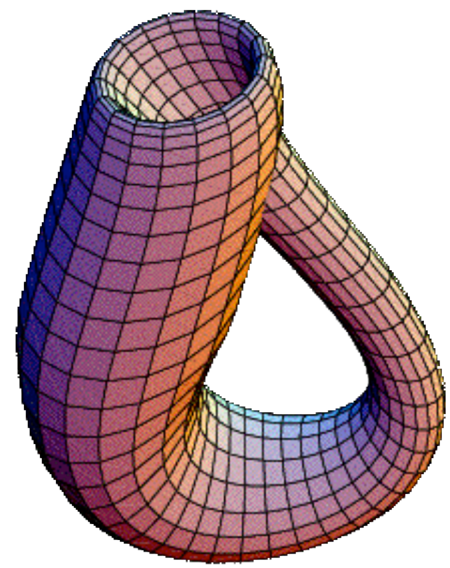


Picture from L. Wilkinson et al. (2005)

Nonlinear Dimensionality Reduction (NLDR)

Assumptions:

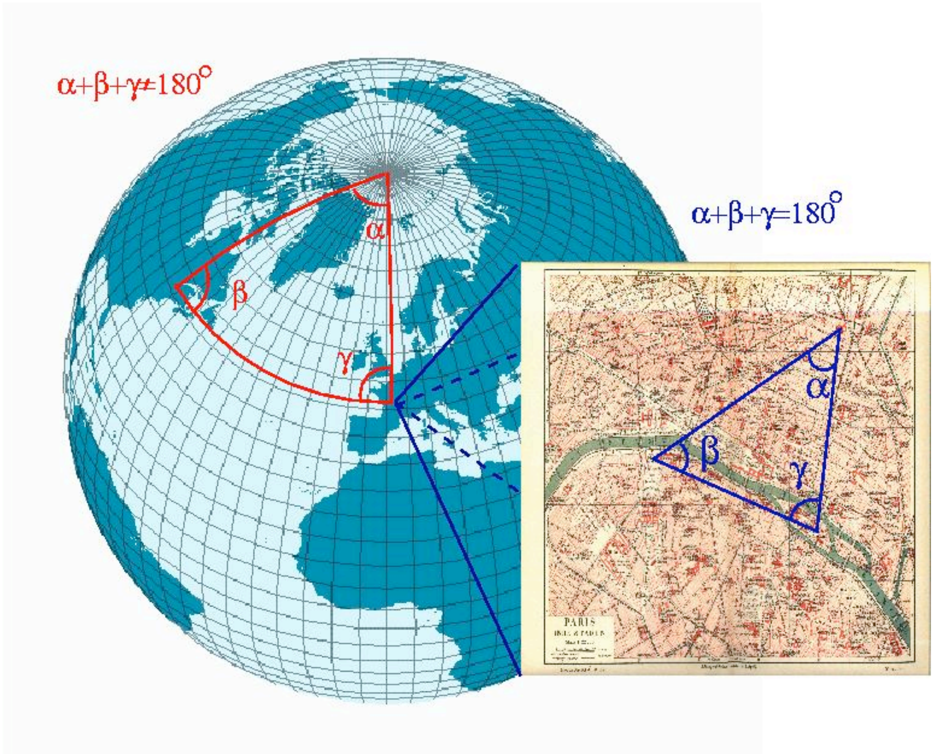
- data of interest lies on embedded nonlinear manifold within higher dimensional space
- manifold is low dimensional $\Rightarrow$ can be visualized in low dimensional space.



Picture from: http://en.wikipedia.org/wiki/Image:KleinBottle-01.png

Manifold

Topological space that is “locally Euclidean”.

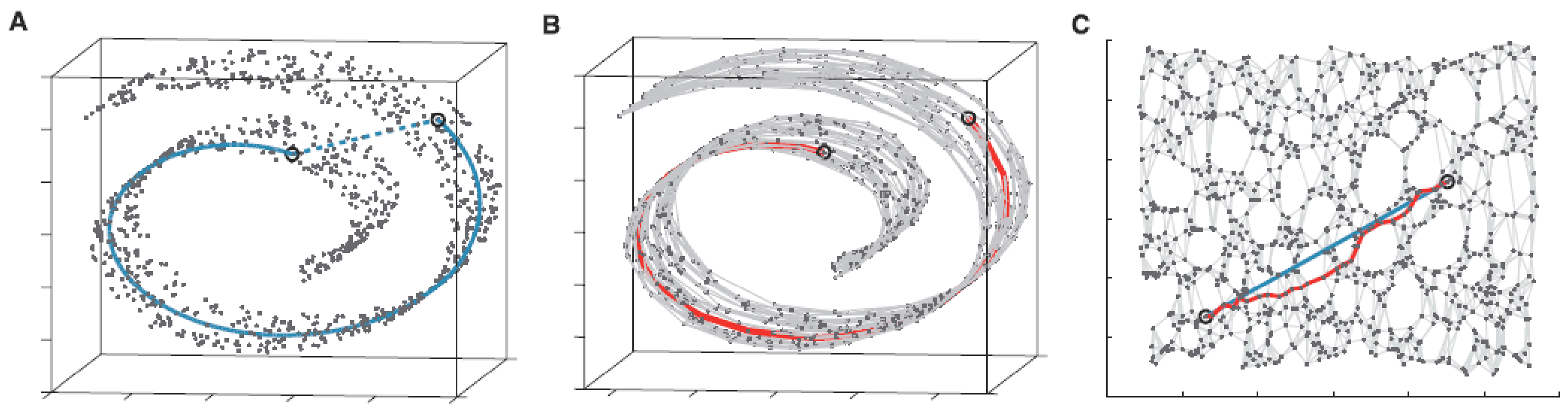


Picture from: http://en.wikipedia.org/wiki/Image:Triangle_on_globe.jpg

Methods

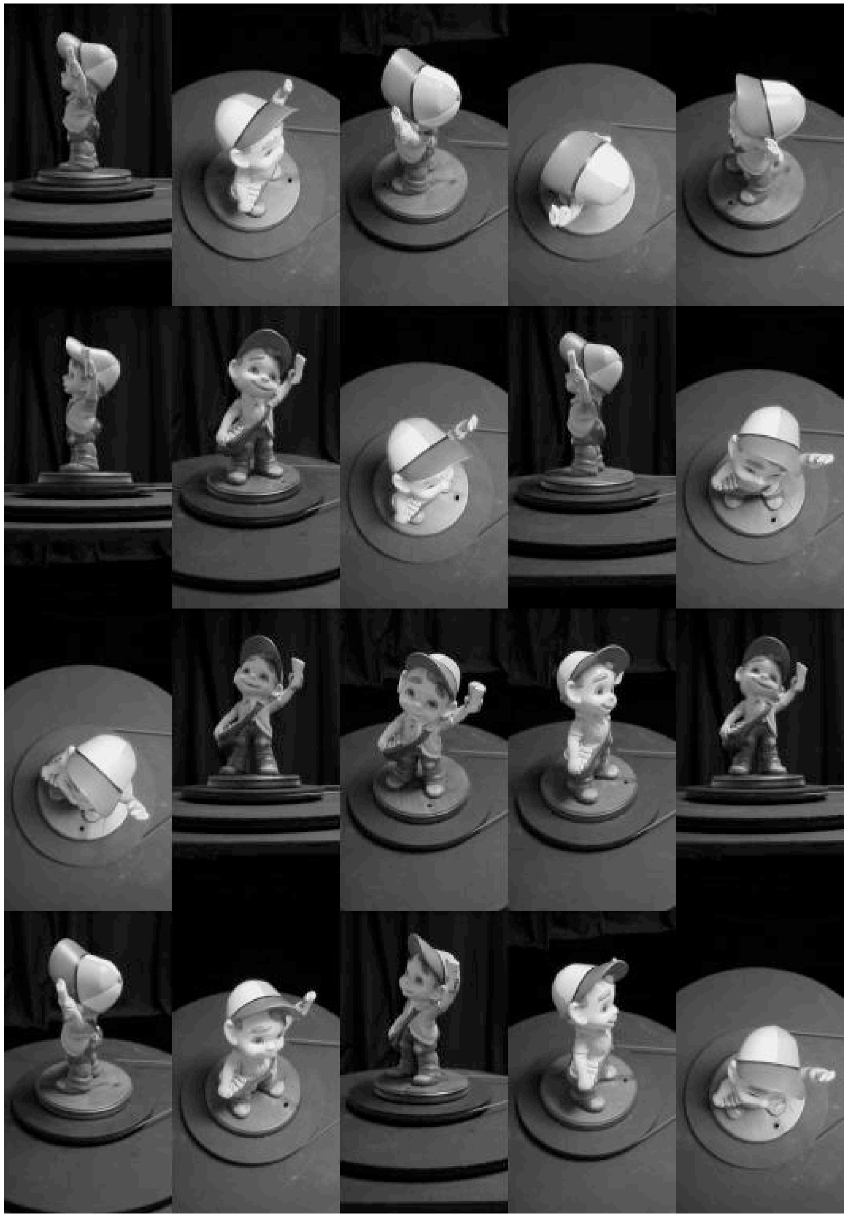
- Locally Linear Embedding
- ISOMAPS

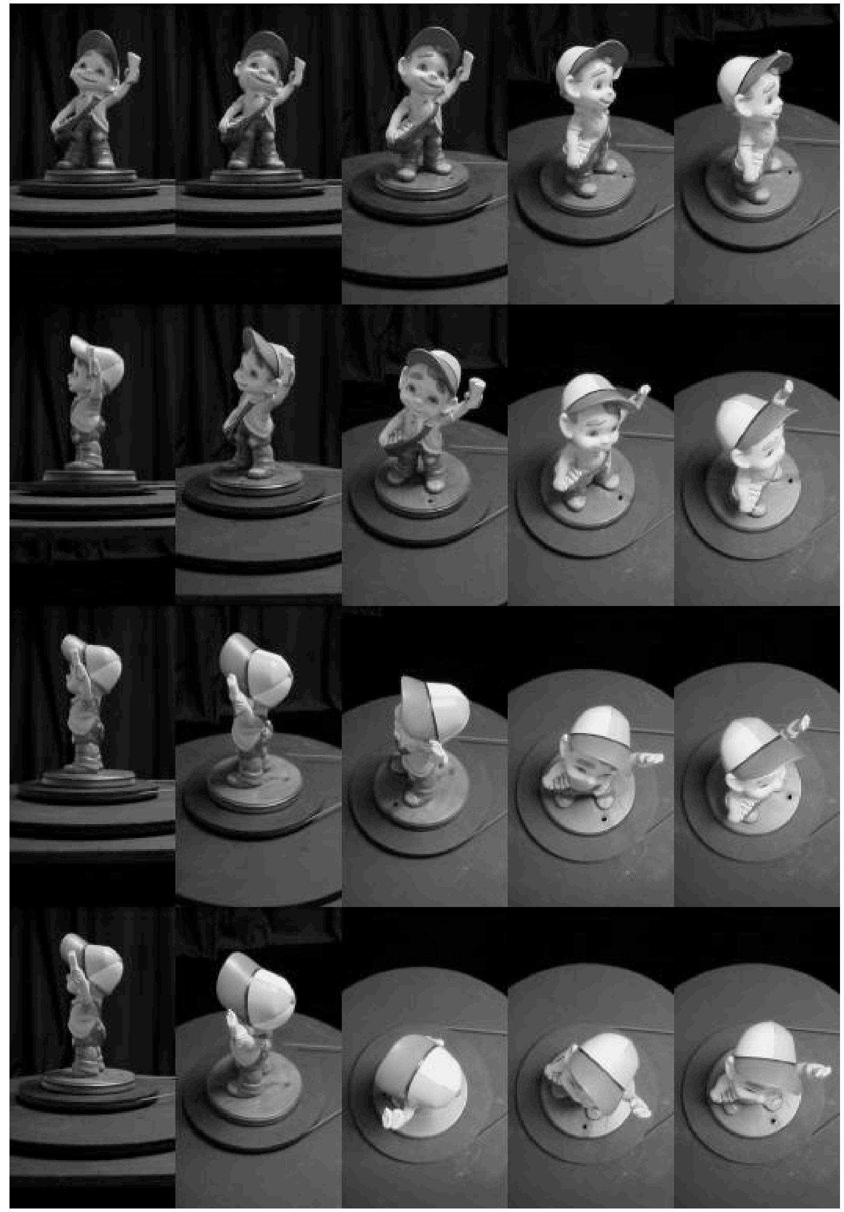
Isomaps Algorithm



1. Construct neighborhood graph
2. Compute shortest paths
3. Construct *d*-dimensional embedding (like in MDS)

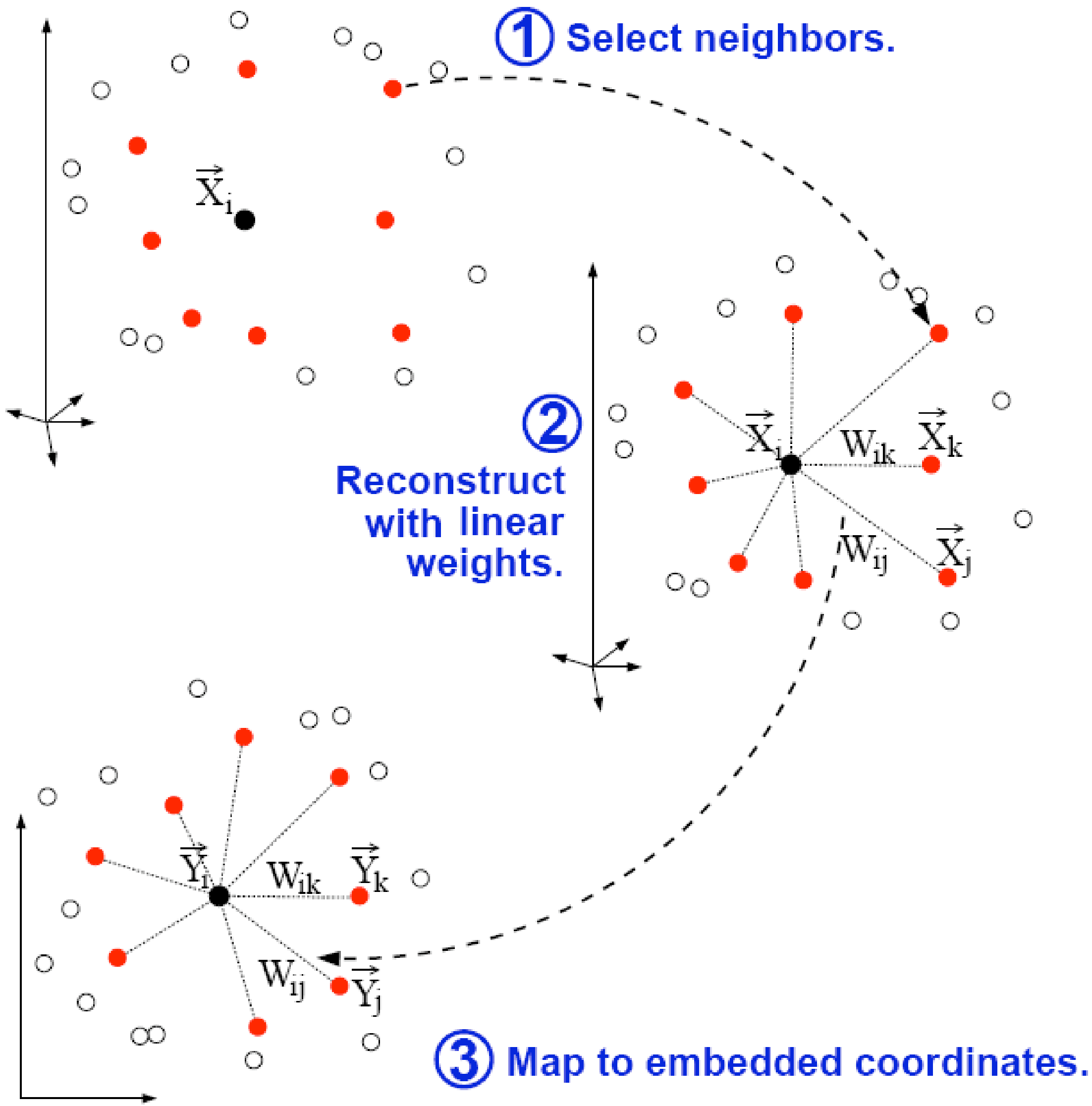
Picture from: Joshua B. Tenenbaum et al. (2000)





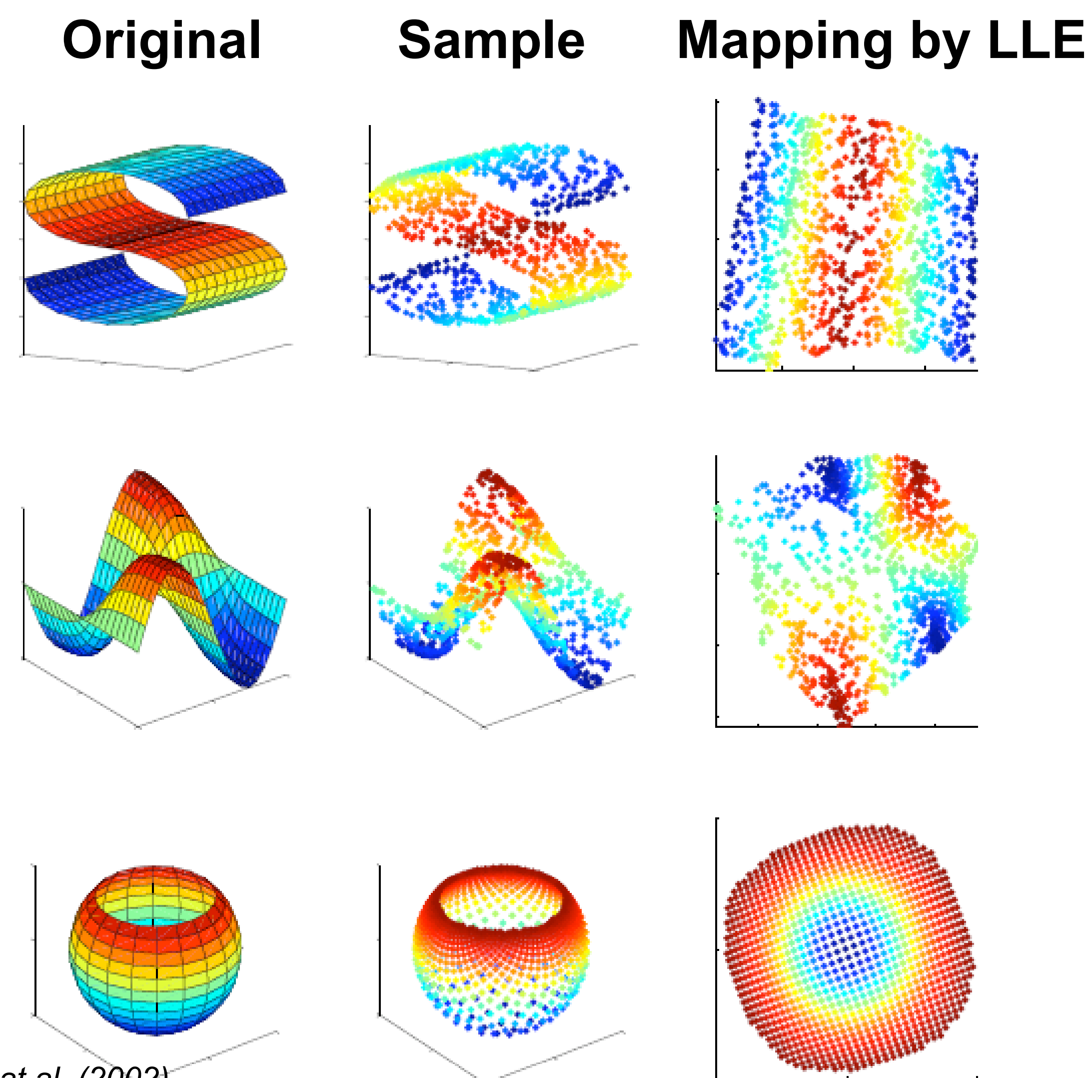
Pictures taken from http://www.cs.wustl.edu/~pless/isomapImages.html

Locally Linear Embedding (LLE) Algorithm



Picture from Lawrence K. Saul at al. (2002)

Application of LLE



Picture from Lawrence K. Saul at al. (2002)

Limitations of LLE

- Algorithm can only recover embeddings whose dimensionality, *d*, is strictly less than the number of neighbors, K. Margin between *d* and K is recommended.

- Algorithm is based on assumption that data point and its nearest neighbors can be modeled as locally linear; for curved manifolds, too large K will violate this assumption.

- In case of originally low dimensionality of data algorithm degenerates.

Proposed improvements*

- Analyze pairwise distances between data points instead of assuming that data is multidimensional vector
- Reconstruct convex
- Estimate the intrinsic dimensionality
- Enforce the intrinsic dimensionality if it is known a priori or highly suspected

** Lawrence K. Saul at al (2002)*

Strengths and weaknesses:

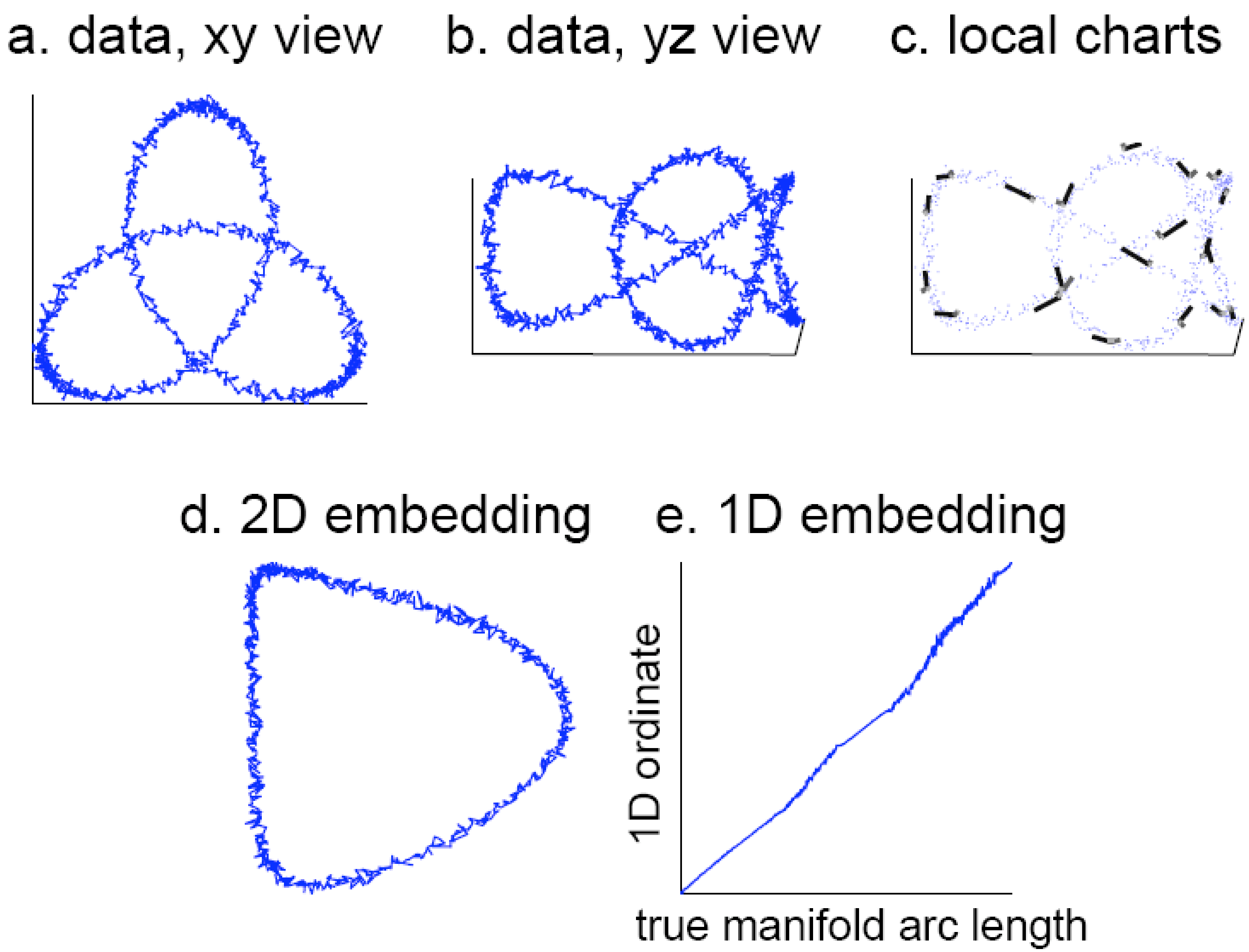
- ISOMAP handles holes well
- ISOMAP can fail if data hull is non-convex
- Vice versa for LLE
- Both offer embeddings without mappings.

Charting manifold

Algorithm Idea

1) Find a set of data covering locally linear neighborhoods (“charts”) such that adjoining neighborhoods span maximally similar subspaces

2) Compute a minimal-distortion merger (“connection”) of all charts

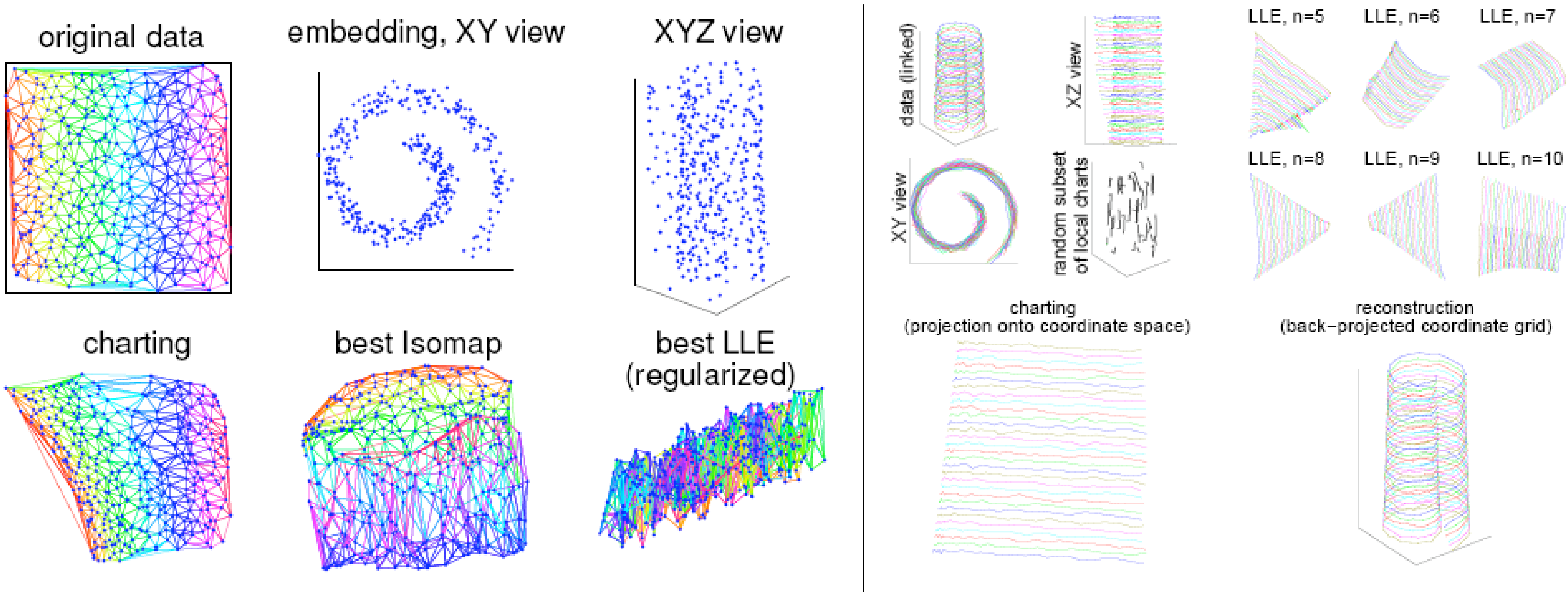


Picture from Matthew Brand (2003)

Video test

Picture from Matthew Brand (2003)

Where ISOMAPs and LLE fail, Charting Prevail



Picture from Matthew Brand (2003)

Questions?

Literature

Covered papers:

1. Graph-Theoretic Scagnostics L. Wilkinson, R. Grossman, A. Anand. Proc. InfoVis 2005.
2. Dimensional Anchors: a Graphic Primitive for Multidimensional Multivariate Information Visualizations, Patrick Hoffman et al., Proc. Workshop on New Paradigms in Information Visualization and Manipulation, Nov. 1999, pp. 9-16.
3. Charting a manifold Matthew Brand, NIPS 2003.
4. Think Globally, Fit Locally: Unsupervised Learning of Nonlinear Manifolds. Lawrence K. Saul & Sam T. Roweis. University of Pennsylvania Technical Report MS-CIS-02-18, 2002

Other papers:

- A Global Geometric Framework for Nonlinear Dimensionality Reduction, Joshua B. Tenenbaum, Vin de Silva, John C. Langford, SCIENCE VOL 290 2319-2323 (2000)