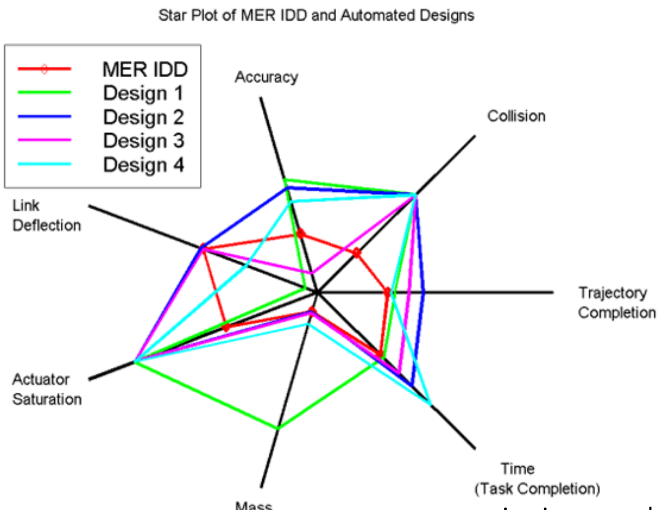
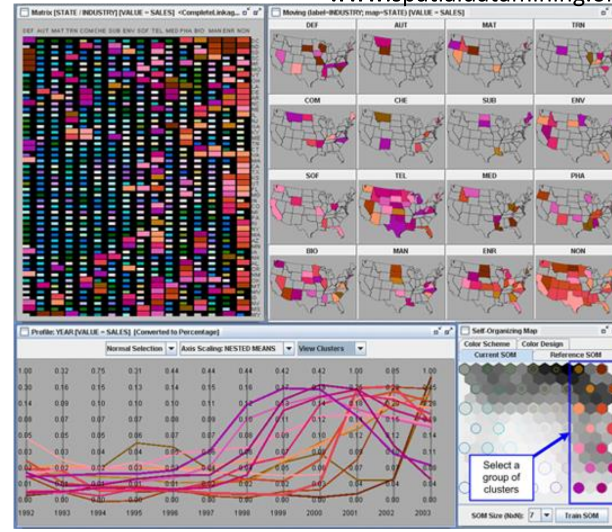
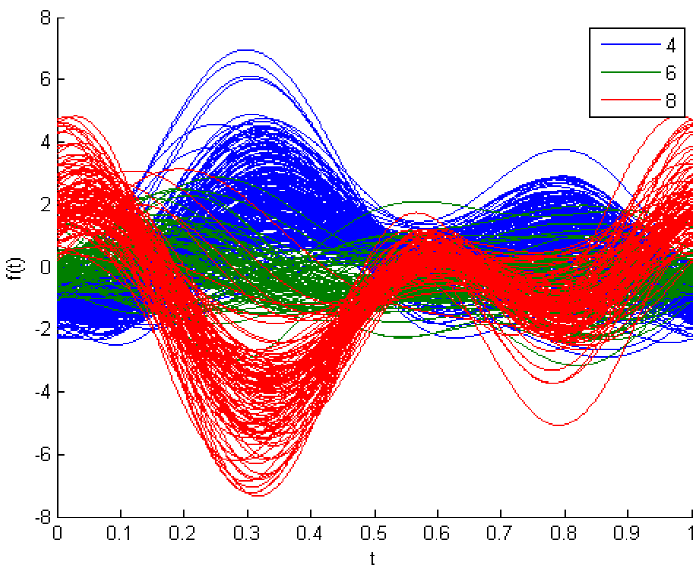


jessicasmaps.blogspot.com

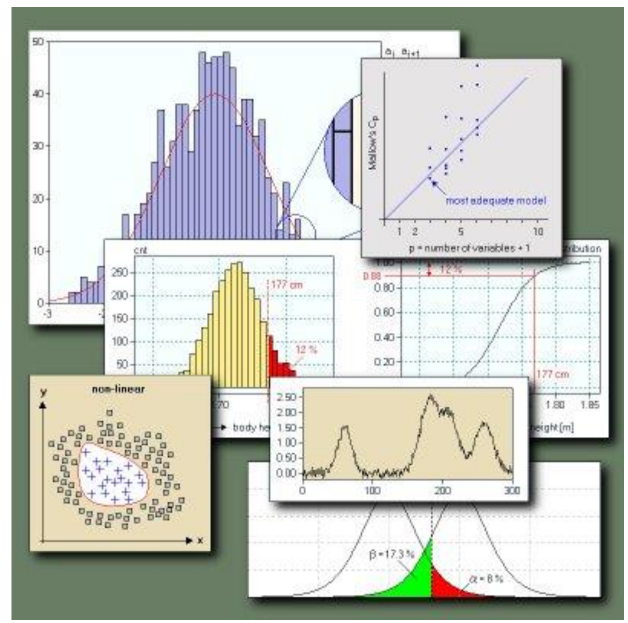
www.spatialdatamining.org



7. Visualization of multivariate data



www.mathworks.com



http://www.statistics4u.com/fundstat_eng/wrapnt3EE177_basic_knowledge.html

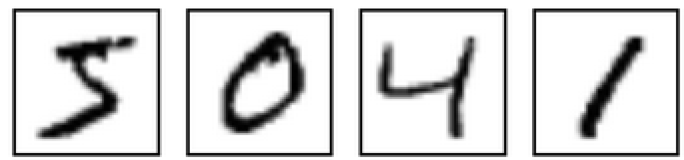
Credits: Manuela Waldner, TU Wien

High-Dimensional Data

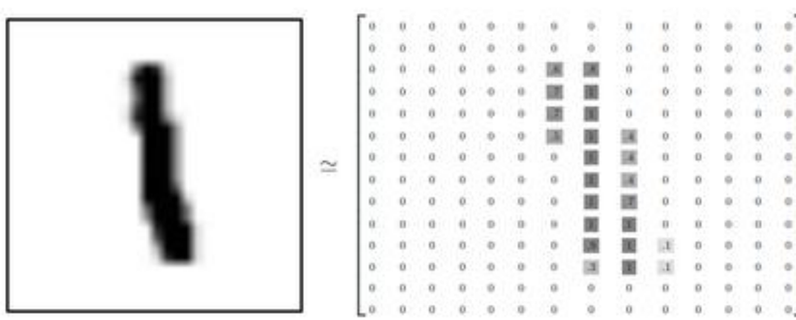
- Consist of multiple types of attributes
  - E.g., weight *w*, height *h*, shoe size *s* of randomly selected sample of people
  - The triples $(w_1, h_1, s_1), (w_2, h_2, s_2)$ then form a set of multivariate data
- Techniques for visualization of lists and tables of data that generally do not contain explicit spatial attributes

High-Dimensional Data

- Image features
  - Vocabulary of visual words
  - Classification
  - Example: MNIST



    - 10 000 hand-written digits
    - 28x28 pixels → 784-dimensional feature vector (intensity values) per image



https://www.tensorflow.org

High-Dimensional Data

- Gene expression data
  - Dimensions: genes
  - Samples: experimental conditions / species /...

http://cancerres.aacrjournals.org/content/64/23/8558

Curse of Dimensionality

- Efficiency of many algorithms depends on the number of dimensions
- With increasing number of dimensions, data becomes sparse
- Number of required training samples grows exponentially with the number of dimensions

Goals of Visualization

- Visual exploration of high-dimensional data sets
  - Detecting clusters
  - Finding regularities and irregularities
  - Identifying relevant data dimensions
- Visual inspection of classification results
- Understanding and quality assessment of algorithms

Exemplary Dataset

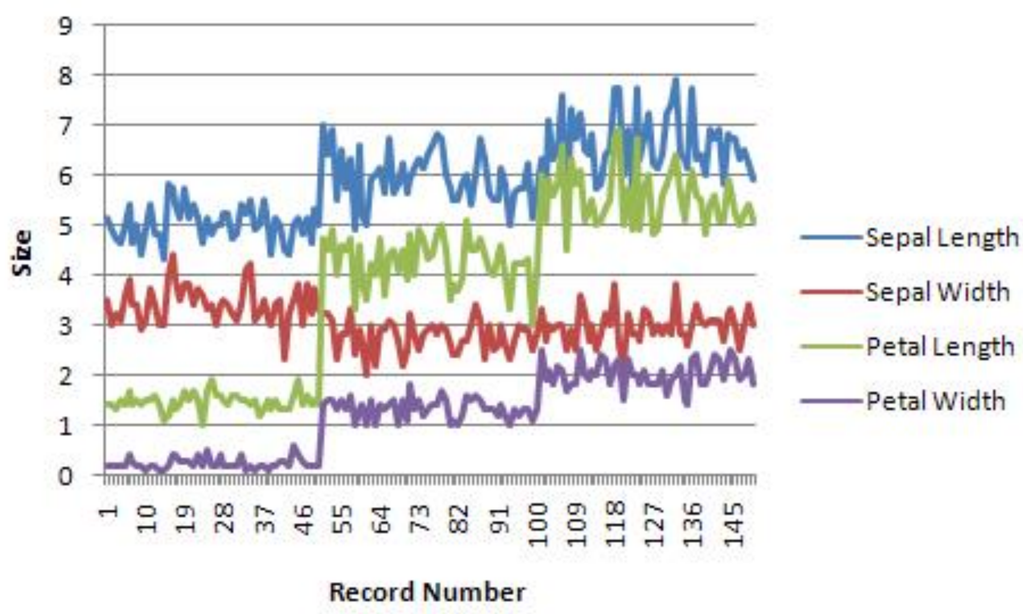
- Example: Iris dataset
  - 3 species
  - 50 samples per species
  - 4 features: length and width of sepals and petals



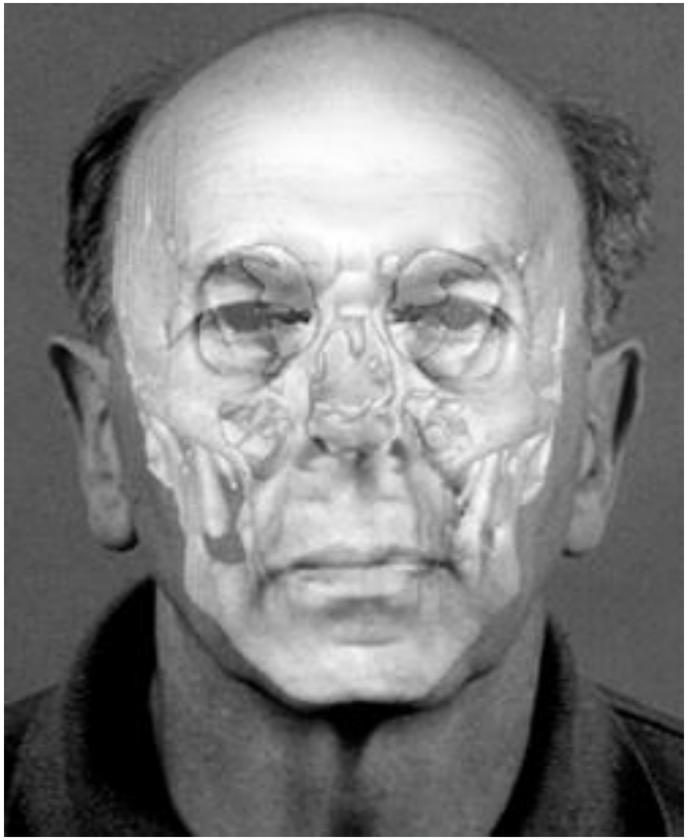
Wikipedia: Iris flower data set

Line-Based Representations

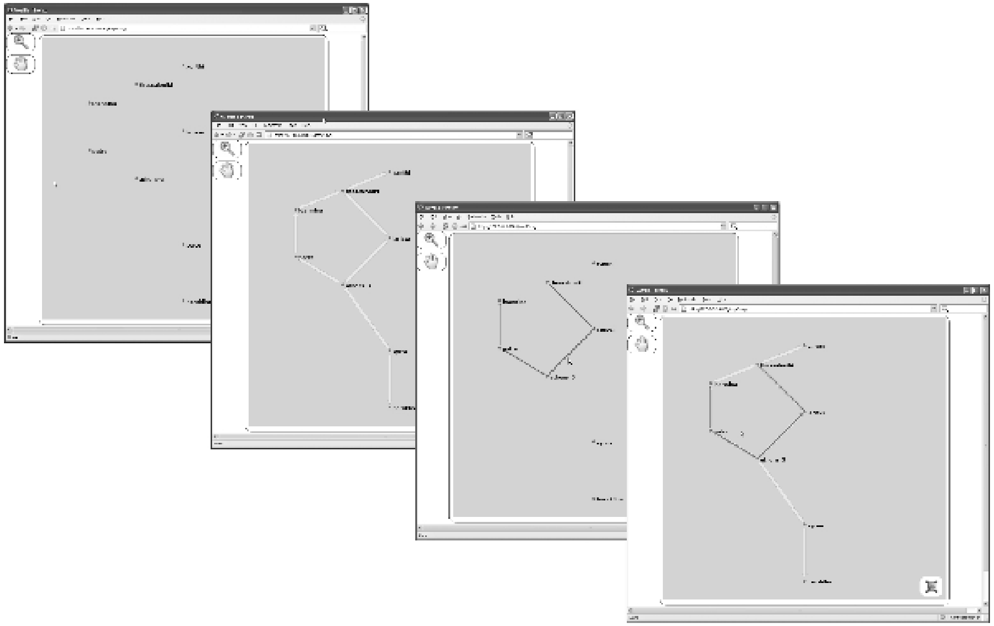
- Visualization technique for single variable, where vertical axis represents possible range of variable values and horizontal axis represents certain ordering of records in a given dataset
- Extension for multivariate data – superimposition, juxtaposition



Superimposition vs. juxtaposition



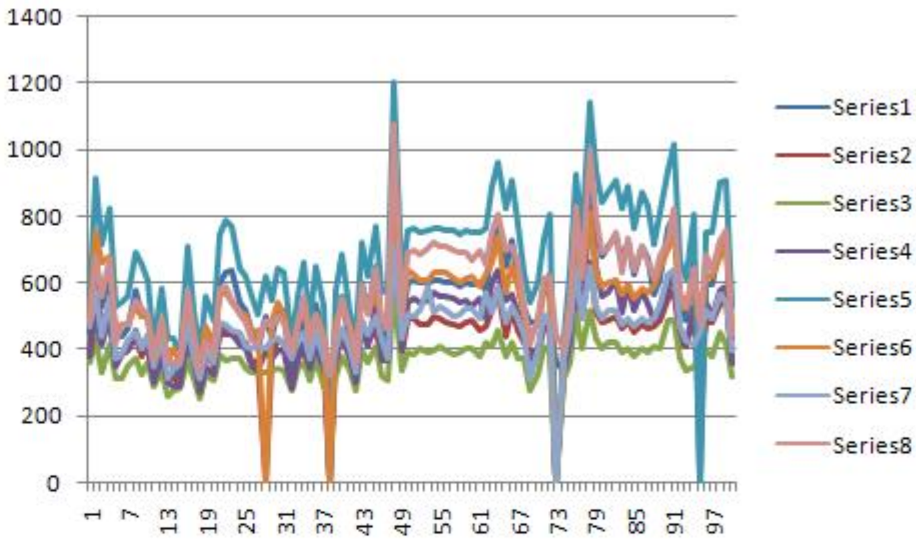
www.craniofacial-id.com

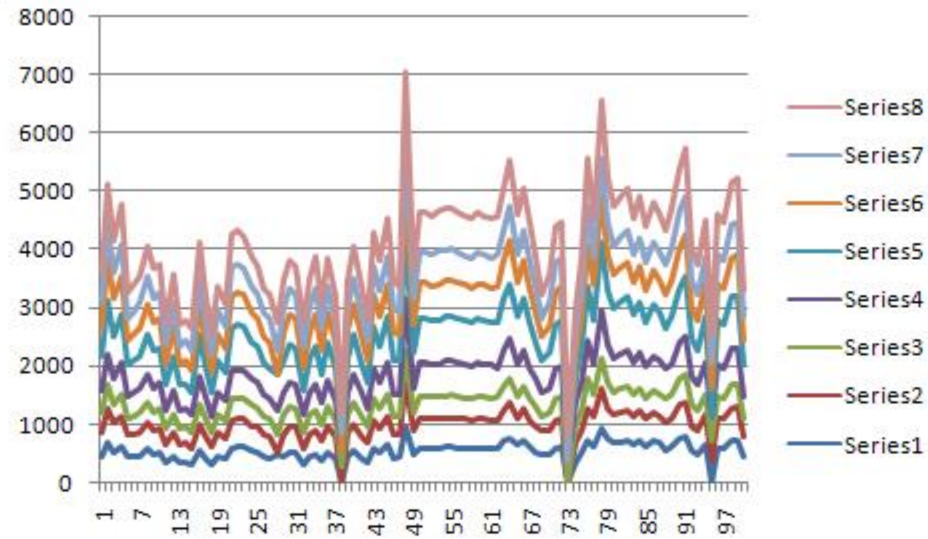


www.usenix.org

Line Charts

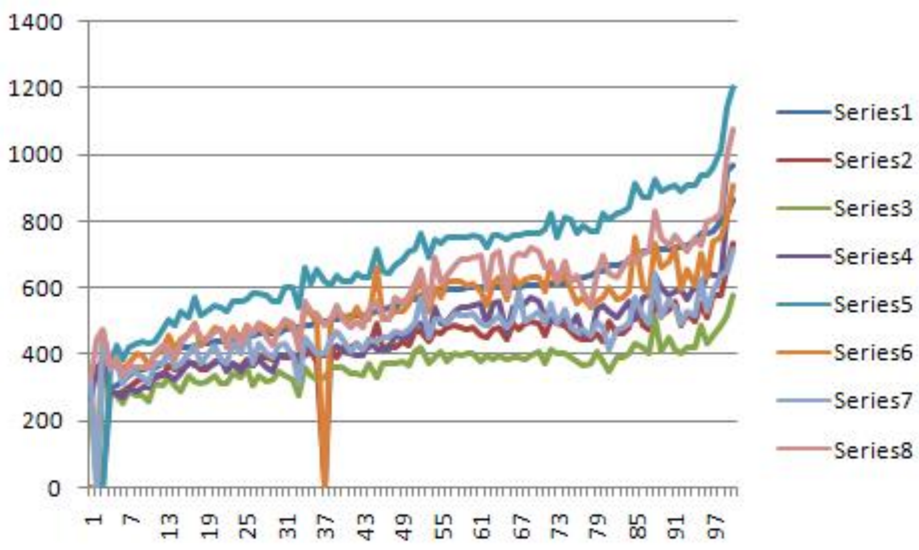
- Classic line chart for 8-dimensional dataset vs. stacked line chart (for each added dimension the chart of previous dimension serves as the base)

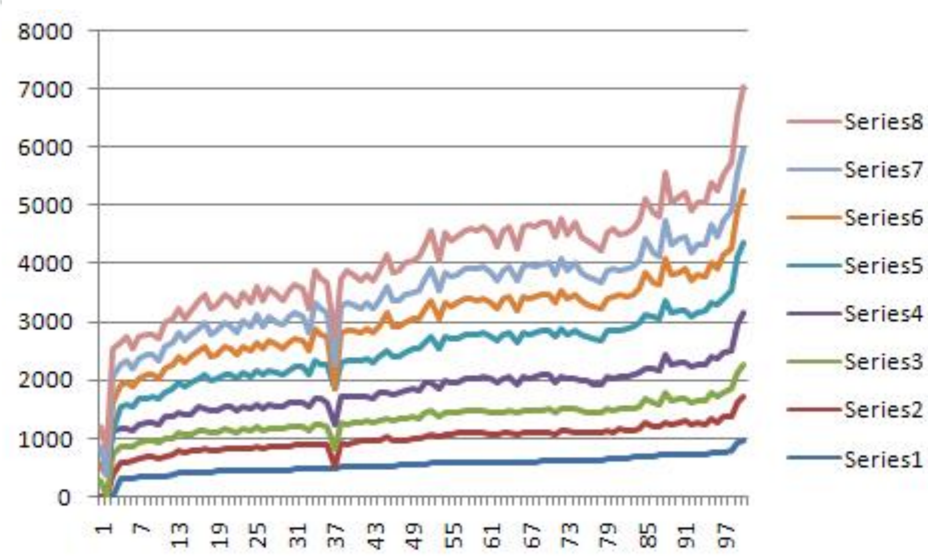




Line Charts

- Sorting of records by single dimension



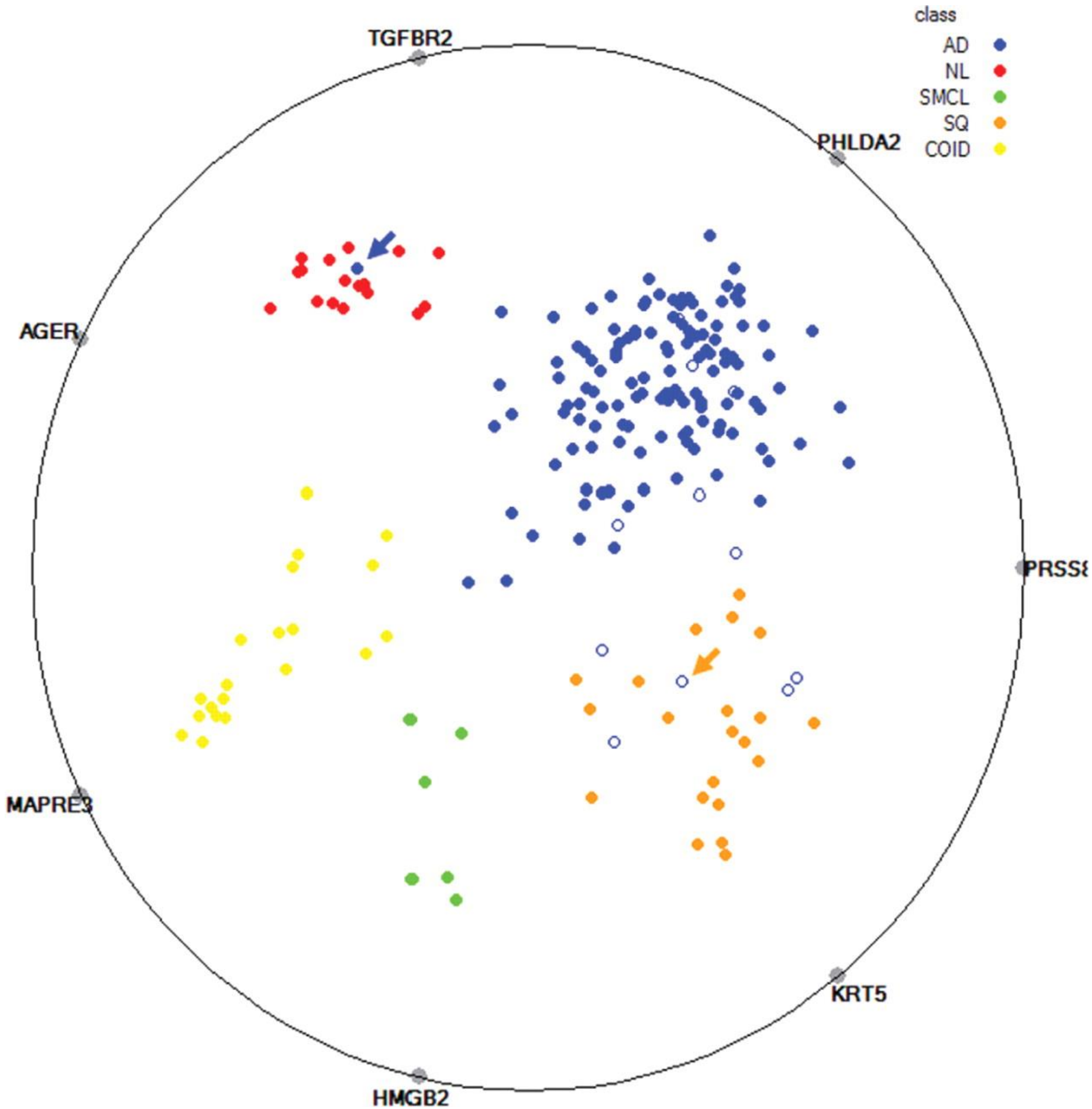


Line Charts

- If the dimensions have the same units, it is possible to use one of the previous techniques
- However, if the individual variables have different units, it is necessary to use different approach, e.g.:
  - Using multiple vertical axes
  - Vertical stacking of charts for individual dimensions

RadViz

- Based on Hooke's law of elasticity for finding equilibrium position of the point.
- For N-dimensional dataset, N so-called "anchor" points are placed on the circumference of a circle (for simplicity we consider a unit circle placed at the origin of the coordinate system) – these represent fixed ends of N strings assigned to each data point.



bioinformatics.oxfordjournals.org

RadViz

- For a given normalized vector of data record $D_i = (d_{i,0}, d_{i,1}, \ldots, d_{i,N-1})$ and a set of vectors *A*, where A_j is the j-th anchor point, we get the equilibrium equation:

$$\sum_{j=0}^{N-1} (A_j - p) d_j = 0$$

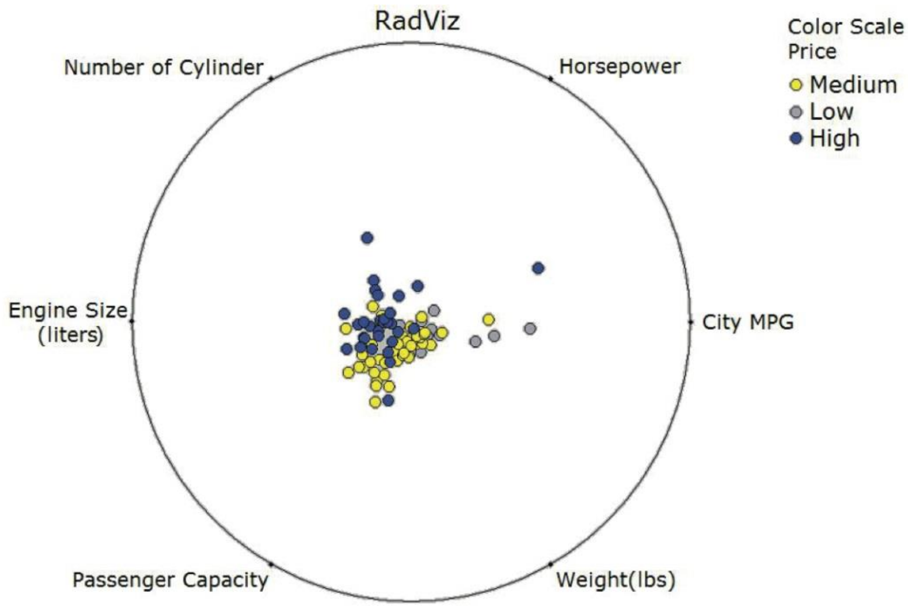
where p is the vector for the point in equilibrium position and can be found as:

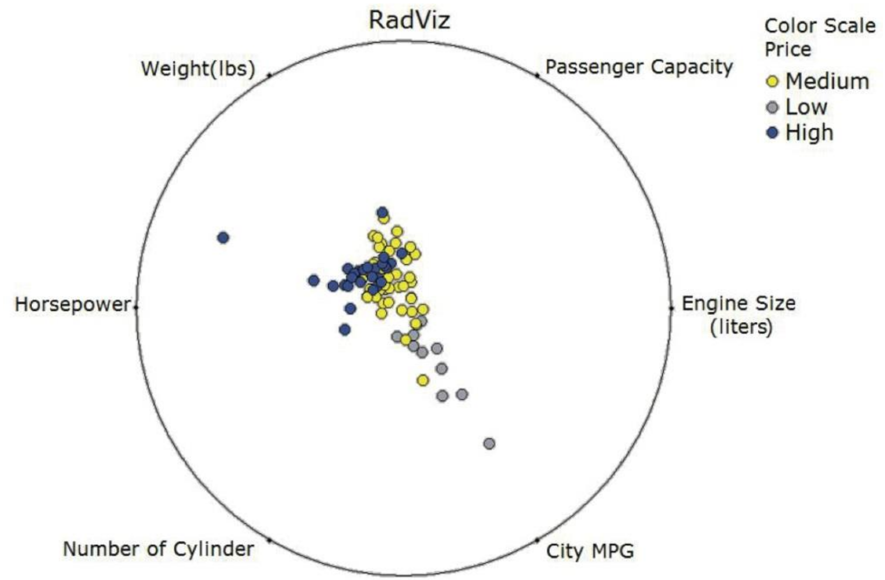
$$p = \frac{\sum_{j=0}^{N-1} (A_j d_j)}{\sum_{j=0}^{N-1} d_j}$$

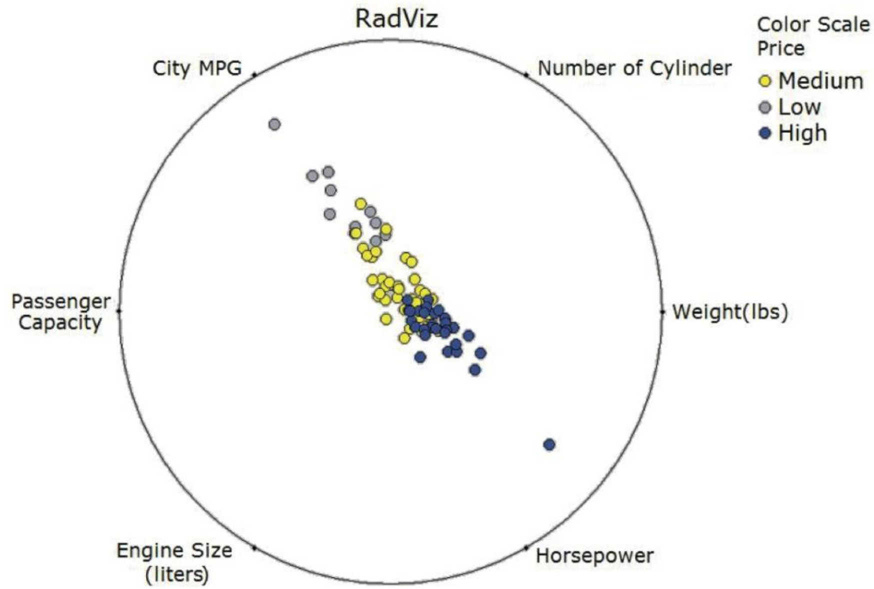
RadViz

- Different placement and order of anchor points leads to different results
- Points with different position in the N-dimensional space can be mapped to the same position in 2D space
- These problems concern all the techniques for projection and dimension reduction
- The simple solution for RadViz is enabling the user to interact (manipulate) with anchor points

RadViz

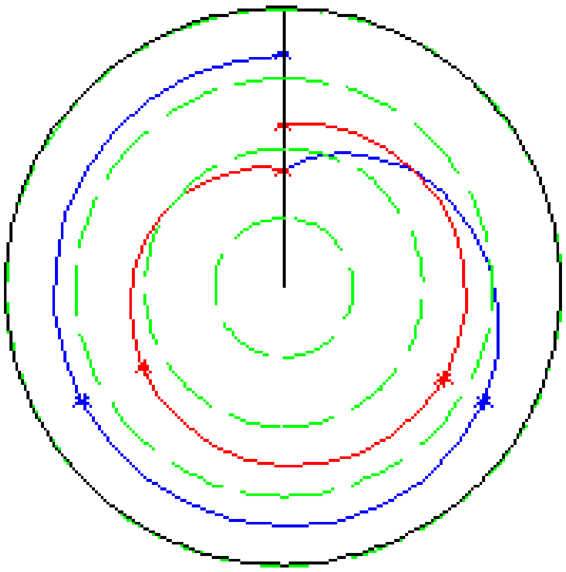

RadViz
Number of Cylinder
Horsepower
Engine Size (liters)
City MPG
Passenger Capacity
Weight(lbs)
Color Scale
Price
Medium
Low
High


RadViz
Weight(lbs)
Passenger Capacity
Horsepower
Engine Size (liters)
Number of Cylinder
City MPG
Color Scale
Price
Medium
Low
High


RadViz
City MPG
Number of Cylinder
Passenger Capacity
Weight(lbs)
Engine Size (liters)
Horsepower
Color Scale
Price
Medium
Low
High

Radial Axis Techniques

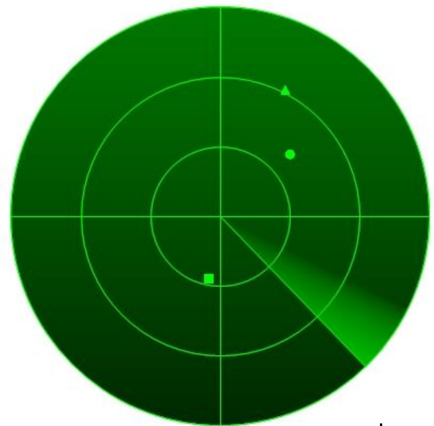
- For each technique with horizontal and/or vertical orientation of coordinate system there exists equivalent technique using radial orientation
- Radial line chart



publib.boulder.ibm.com

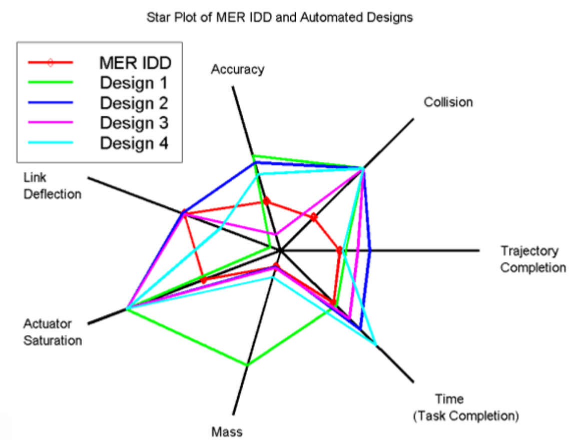
Radial Techniques

- Radar



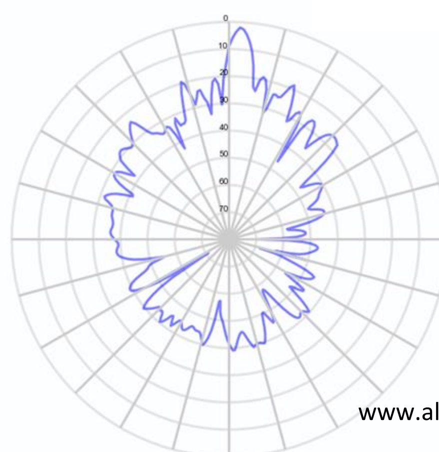
www.prlog.org

- Star chart



commons.wikimedia.org

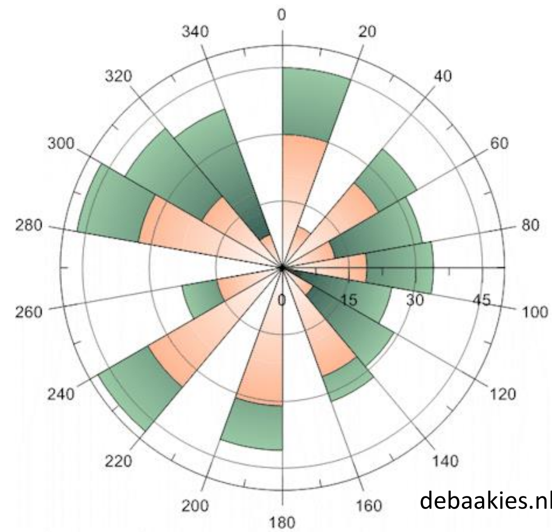
- Polar chart
  - Displaying polar coordinates



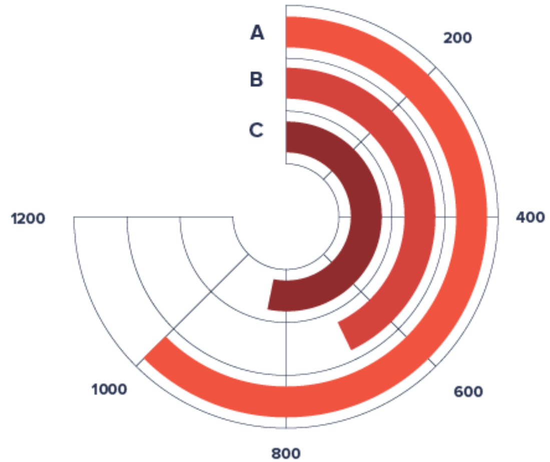
www.alteryx.com

Radial Techniques

- Radial column charts



- Radial bar charts

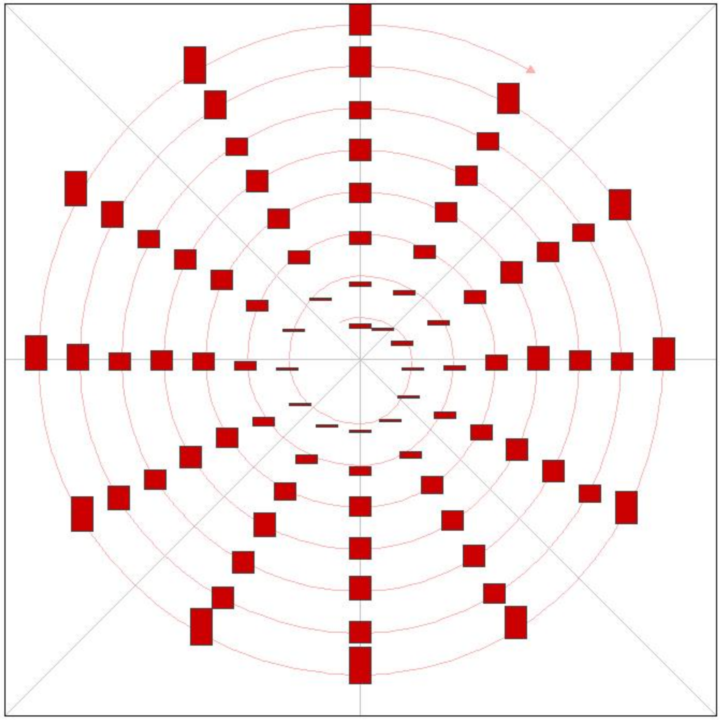


- Radial area charts

datavizproject.com

Types of Techniques for Radial Axes

- Concentric circles
- Continuous spiral – does not exhibit disconti-nuity at the end of each cycle
- Compared to traditional bar representation enables observation of patterns between elements at the same position in different cycles

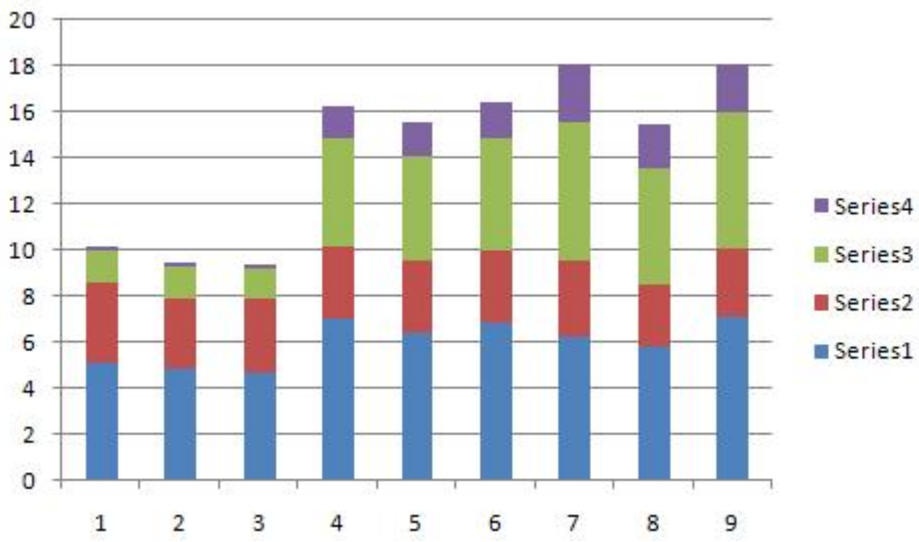


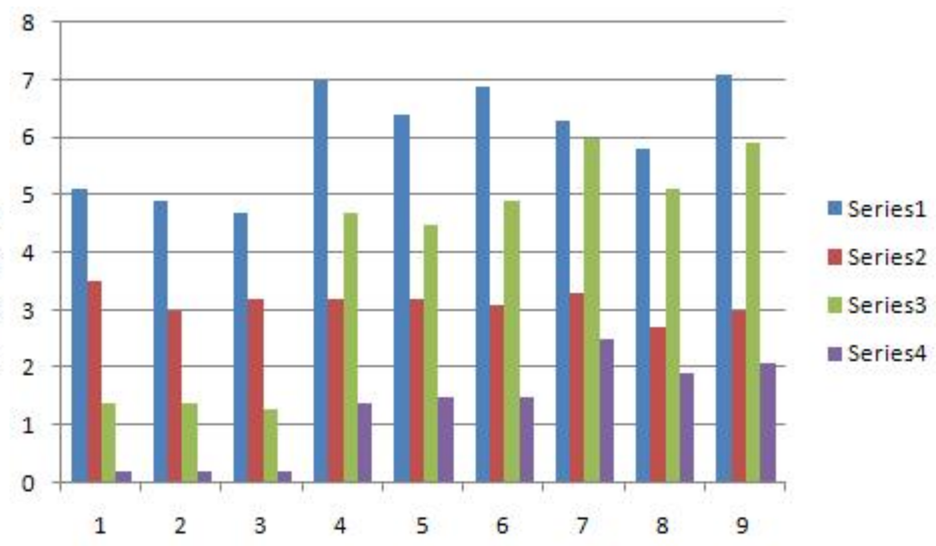
Techniques for Area Data

- Usage of filled polygons of given size, shape, color, …
- The aim of some of these techniques is not showing individual data records, but their clusters and distribution
- Originally designed for univariate data (single variable) – pie charts and bar charts. Subsequently extended for multiple dimensions.

Bar Charts/Histograms

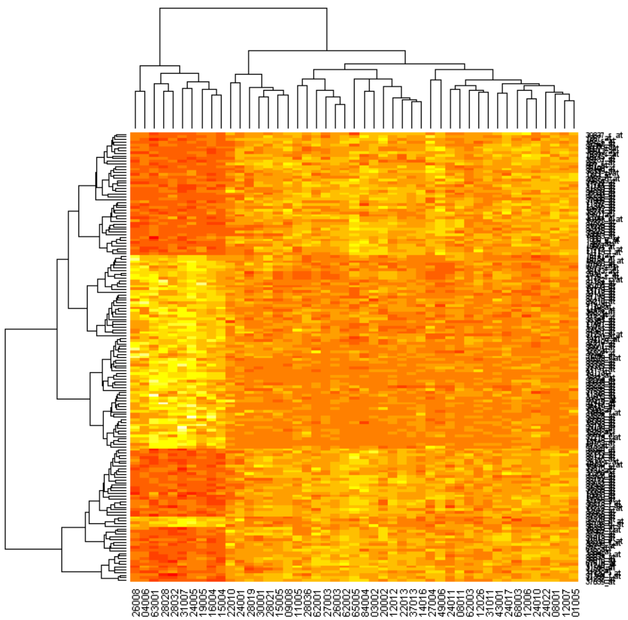
- Multivariate data – stacked bar chart



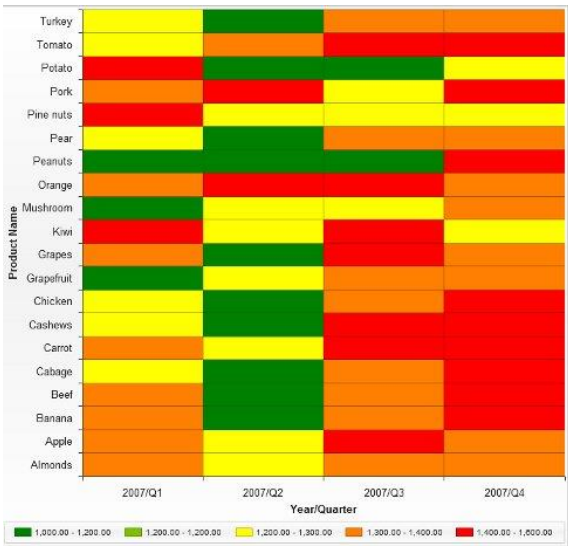


Tabular Visualizations

- Multivariate data often in tables
- **Heatmaps**
  - – displaying records using color instead of text
  - – each value is rendered as a colored rectangle



www2.warwick.ac.uk



akweebeta.com

Tabular Visualizations

- **Survey plot**
  - Instead of color, the size of the cell depicts the value
  - Centres of the cells are aligned to individual attributes
  - Measurement of area is more prone to errors than measurement of length

Dimensional Stacking

- Mapping of data from discrete N-dimensional space to 2D image in such way that the data occlusions are minimalized, while the majority of the spatial information is preserved

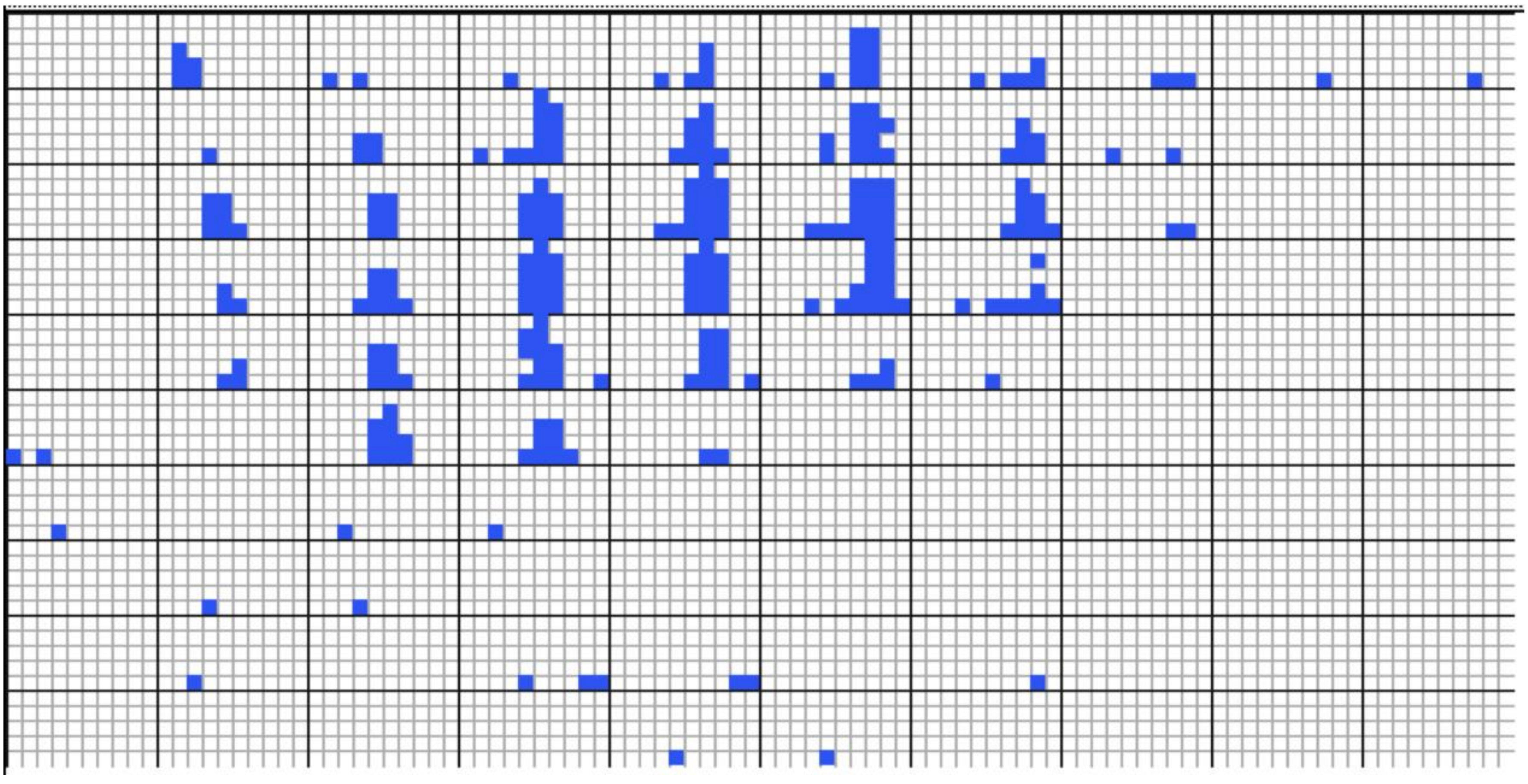
Dimensional Stacking

- Data of 2N+1 dimensions
- Select final cardinality for each dimension
- Select one dimension as dependant variable, the rest of the dimensions are independent
- Create ordered pairs of independent variables (N pairs) and assign unique value (speed) to each pair – from 1 to N
- Pair corresponding to speed 1 creates virtual image with size corresponding to the cardinality of its dimensions
- In each position of this virtual image, new virtual image corresponding to the dimensions of pair with the speed 2 is created
- The process is repeated, until all dimensions are not included

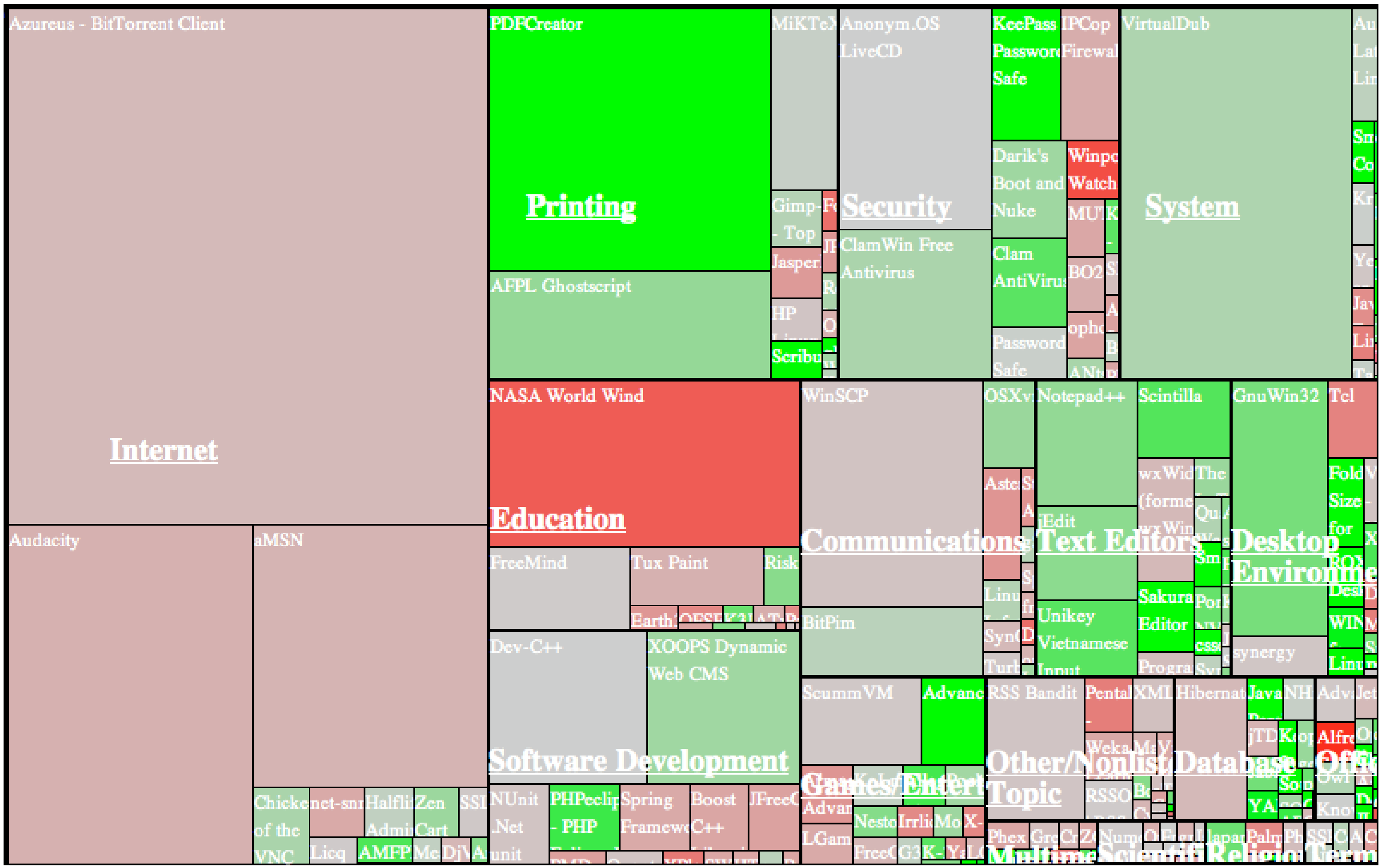
Dimensional Stacking

- Begins with discretisation of the range of each dimension. Orientation and order is then assigned to each dimension. Dimensions with two lowest orders are then used to split the virtual screen into sections - the cardinality of the dimensions indicates, how many sections are generated on horizontal and vertical axes. Each generated section is then used for recursive splitting of virtual screen in next two dimensions in the same way. This process is repeated until all the dimensions are not processed and the data are not placed to their corresponding positions on the screen.

Dimensional Stacking



Treemap



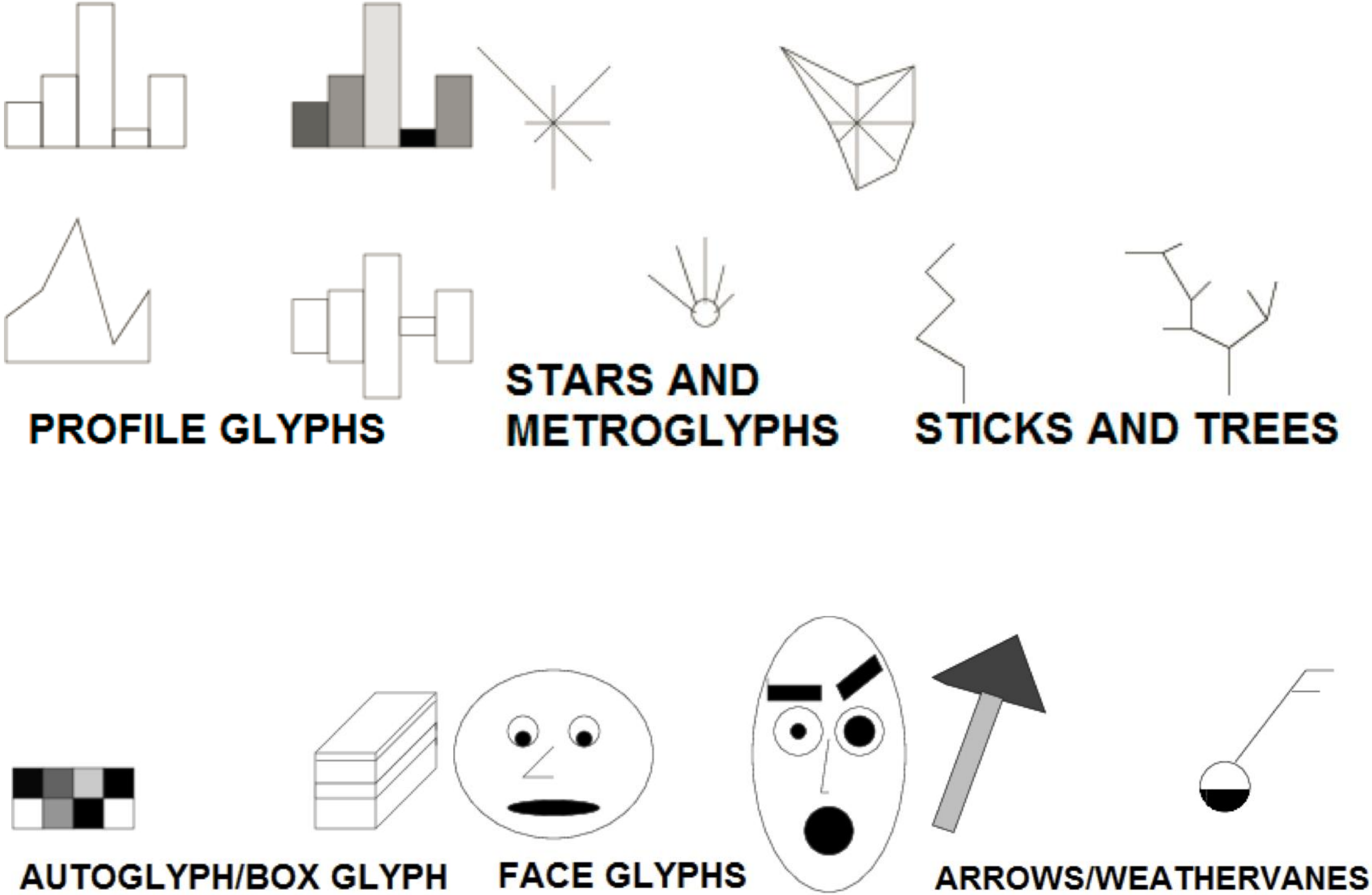
Azureus - BitTorrent Client
Internet
Audacity
aMSN
PDFCreator
Printing
AFPL Ghostscript
MiKTeX
Anonym.OS LiveCD
Security
ClamWin Free Antivirus
KeePass Password Safe
IPCop Firewall
Darik's Boot and Nuke
Clam AntiVirus
Password Safe
VirtualDub
System
NASA World Wind
Education
FreeMind
Tux Paint
Risk
WinSCP
Communications
BitPim
Notepad++
Scintilla
jEdit
Text Editors
Unikey Vietnamese Input
Sakura Editor
GnuWin32
Tcl
Desktop Environment
synergy
Dev-C++
XOOPS Dynamic Web CMS
Software Development
NUnit
Spring Framework
Boost C++
ScummVM
Games/Entertainment
RSS Bandit
Other/Nonlisted Topic
Hibernate
Database

Glyphs and Icons

- Visual representation of parts of data or information, where graphical entity and its attributes are driven by one or more attributes of input data
- Graphical attributes, to which the data values can be mapped:
  - position, size, shape, orientation, material, line style, dynamics

Glyphs and Icons

- Types of mapping:
  - 1:1 – each data attribute is mapped to unique graphical attribute
  - 1:N – set of redundant mappings (e.g., mapping data attribute simultaneously to size and color)
  - M:N – multiple or all data attributes mapped to a common type of graphical attribute


PROFILE GLYPHS
STARS AND METROGLYPHS
STICKS AND TREES
AUTOGLYPH/BOX GLYPH
FACE GLYPHS
ARROWS/WEATHERVANES

Glyphs and Icons

- We must be aware of inaccuracies and restrictions of these techniques:
  - Inaccuracy of perception – depends on the type of used graphical attributes
  - Distance between graphical attributes influences the accuracy of their comparison – the closer, the more precise comparison
  - Number of dimensions and data records which can be effectively displayed using glyphs is limited

Glyphs and Icons

- After selection of the type of glyph there are N! possible orderings of the dimensions, which can be used when mapping
- Several strategies for selection of suitable order exist:
  - Sorting of dimensions based on their correlation
  - Increasing influence of glyph with symmetrical shape
  - Sorting by the values of dimensions in a single record
  - Manual sorting based on knowledge of the domain

Placement of Glyphs

- Three basic types of strategies for placement of glyphs on the screen:

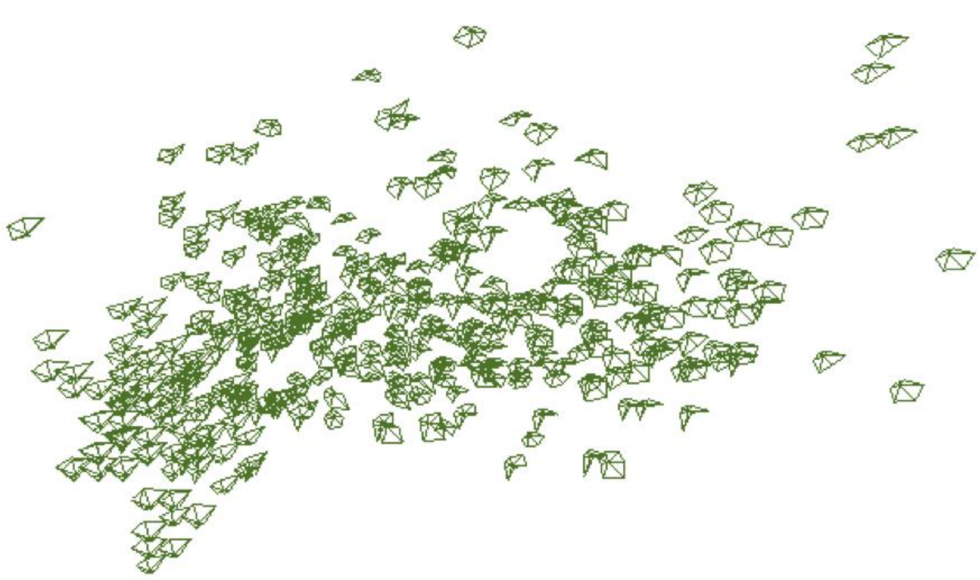
1. Uniform
2. Data-driven
3. Structure-driven

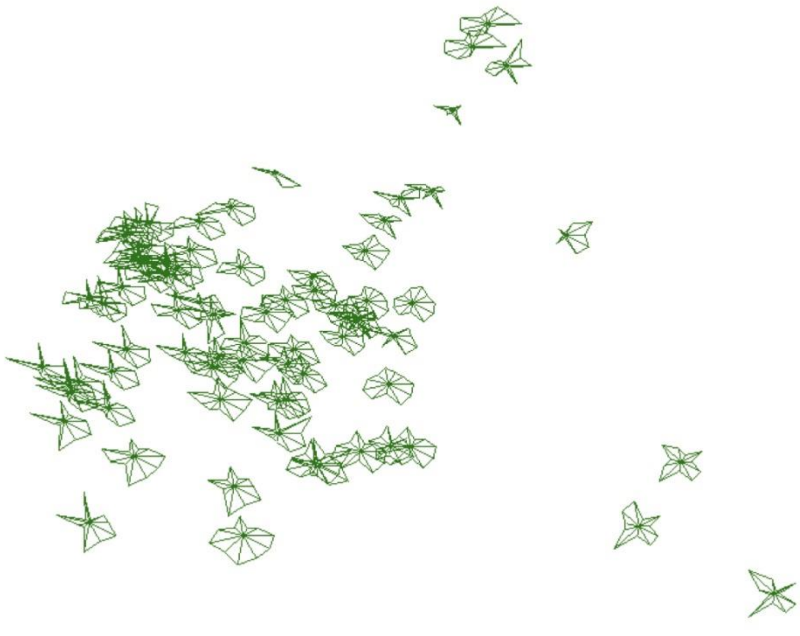
Uniform Placement

- Uniform placement on screen
- Elimination of overlaps, effective usage of screen space

Data-Driven Placement

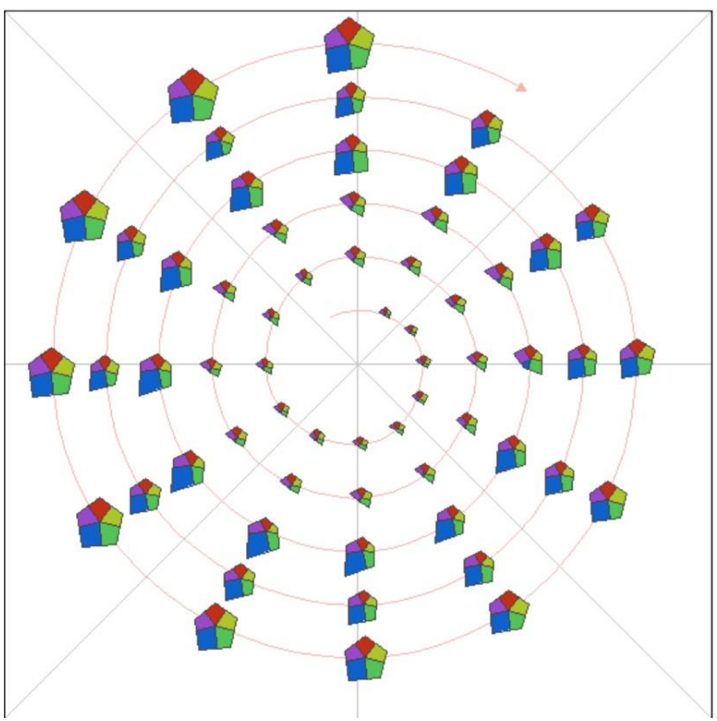
- Two approaches:
  - Select two dimensions to direct the placement (left)
  - Positions derived using PCA, MDS (right)

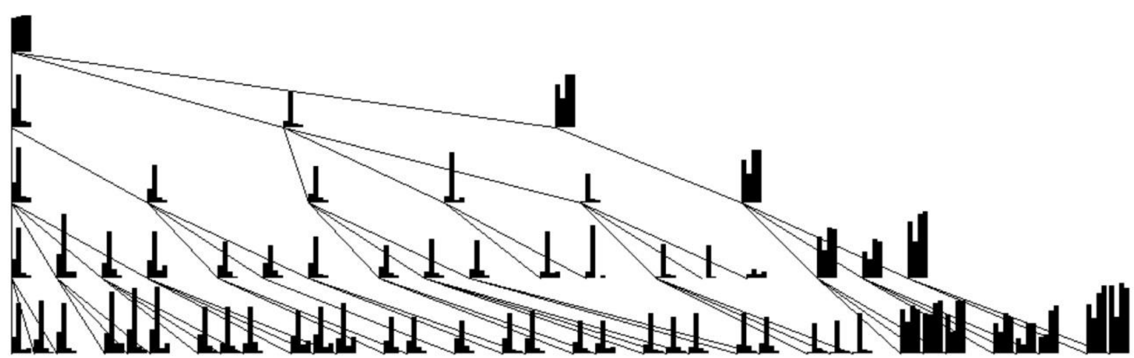




Structure-Driven Placement

- Using structure of the data – cyclic, hierarchical





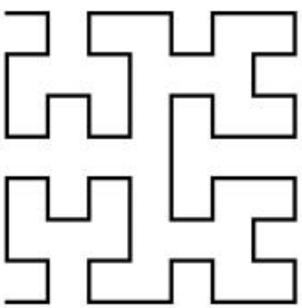
Dense Pixel Displays

- Hybrid method between point-based and regional (area-based) methods
- Maps each value to individual pixel and for each dimension creates filled polygon
- Displaying millions of values within one screen
- Number of data points determines the number of individual items in the image
- The technique relies on application of color

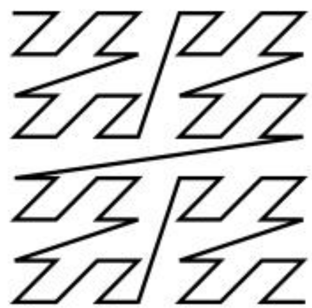
Dense Pixel Displays

- Simplest form:
  - Each dimension of dataset generates independent separated “sub-image” on the screen
  - Each dimension can be considered as an independent set of numbers, each set determines the color of the corresponding pixels
  - The placement of the items within the set (highlighting relationships between close points): alternating passes form right to left and from left to right; if the edge of the image is reached, move to the next line

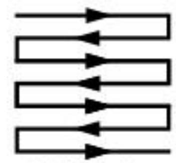
Dense Pixel Displays



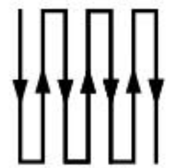
Peano-Hilbert



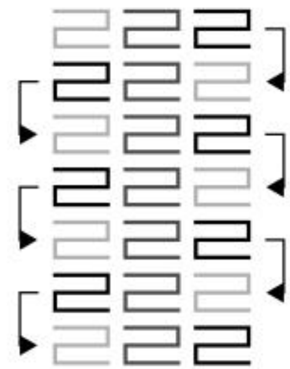
Morton



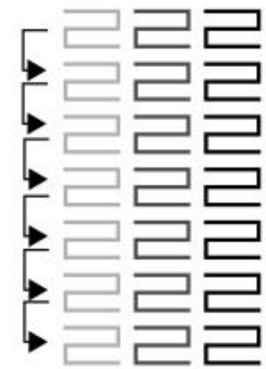
left-right



top-down



back-and-forth loop



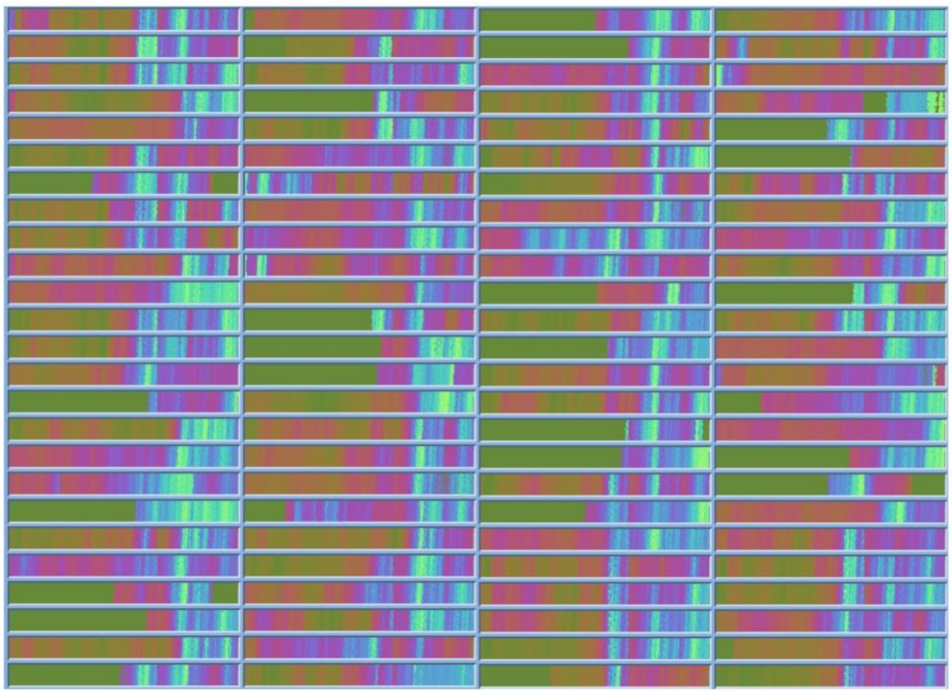
line-by-line loop

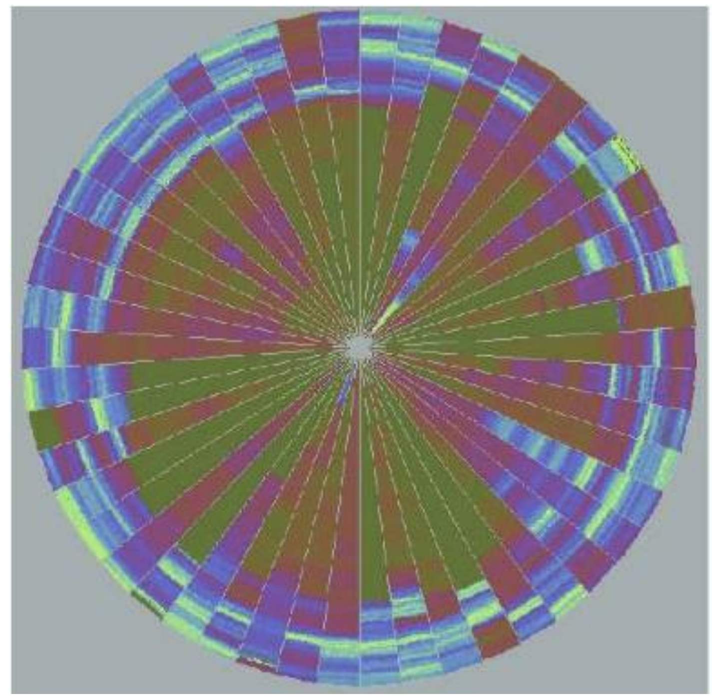
screen filling

recursive patterns

Recursive Patterns, Circular Segments

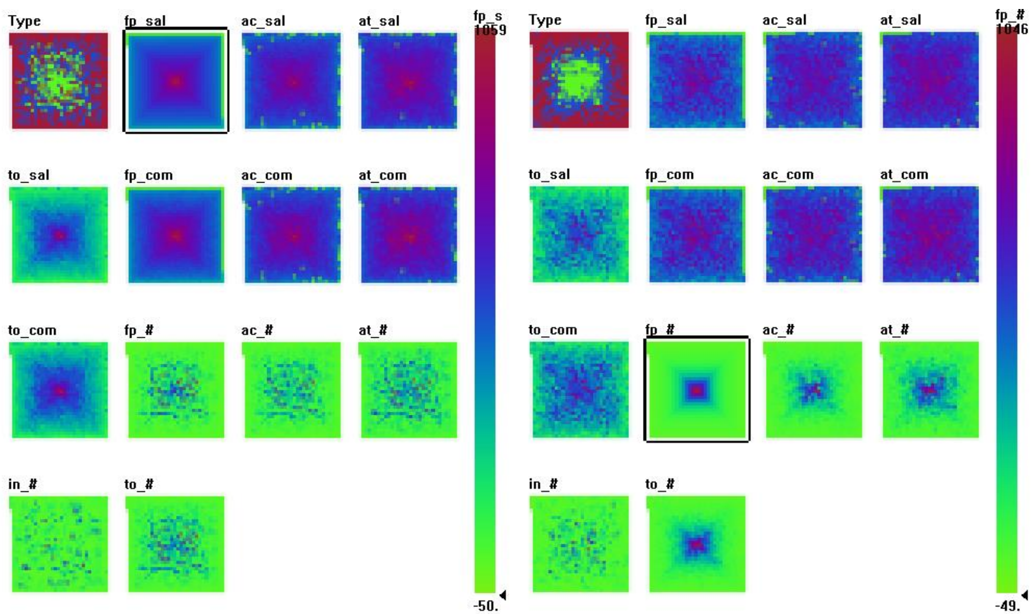
- Placement of sub-images using different approaches:





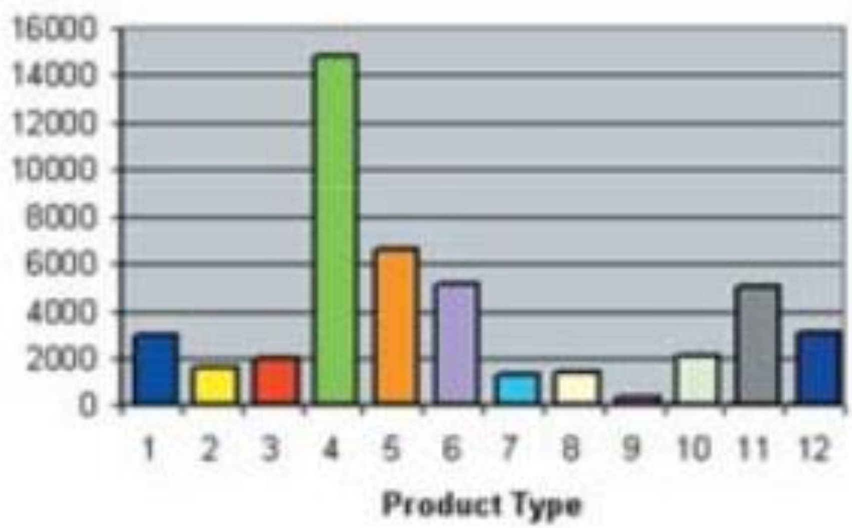
Dense Pixel Displays

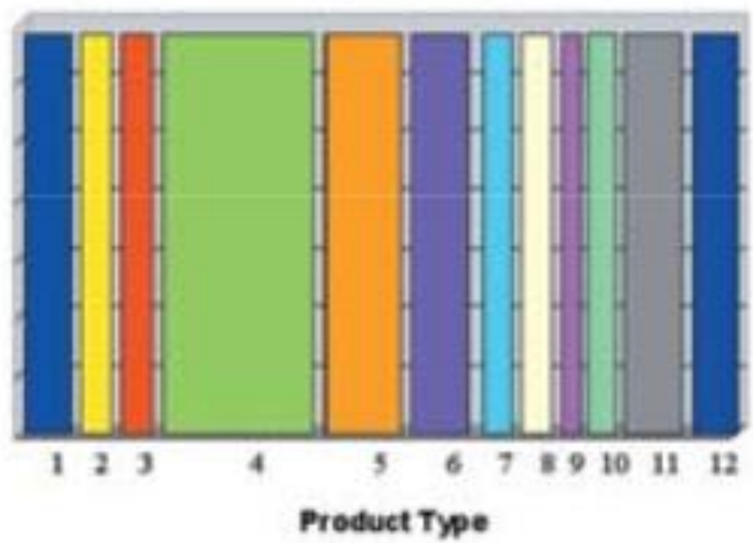
- Last important aspect is ordering of the data
- Time-series data have fixed ordering
- In other types of data the change of order can reveal interesting properties



Pixel Bar Charts

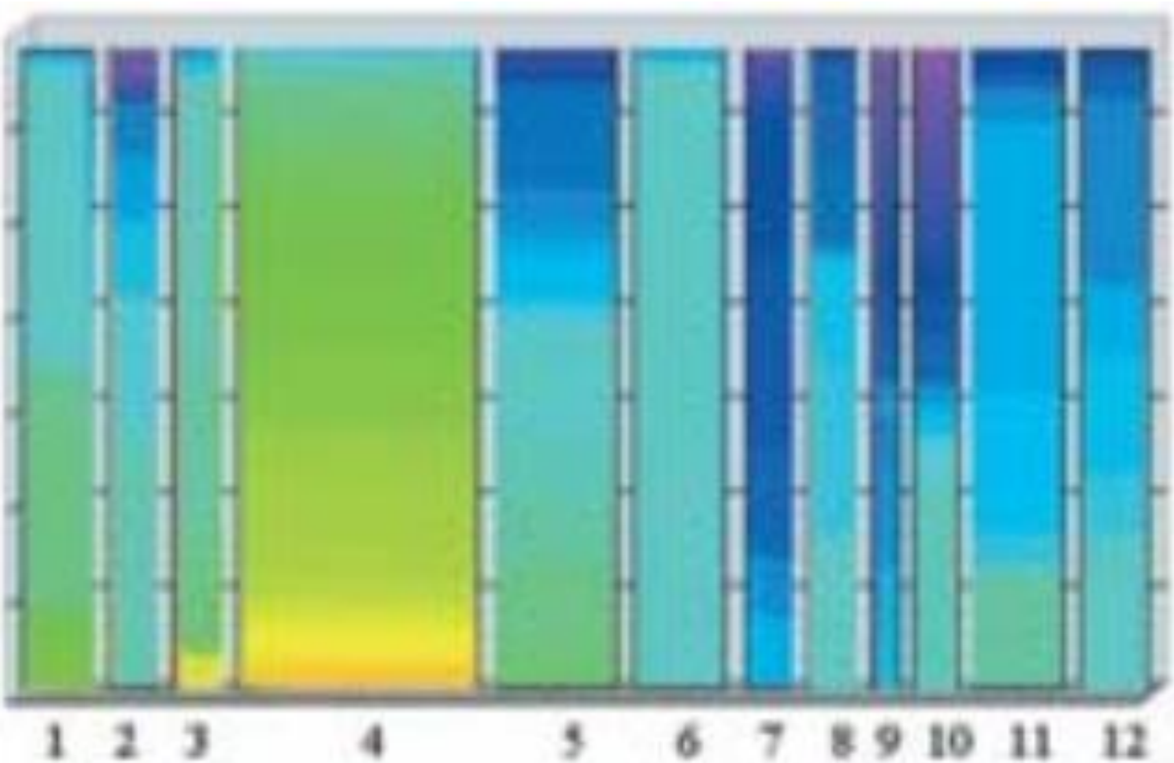
- Overloading of classical bar chart – including more information about individual items





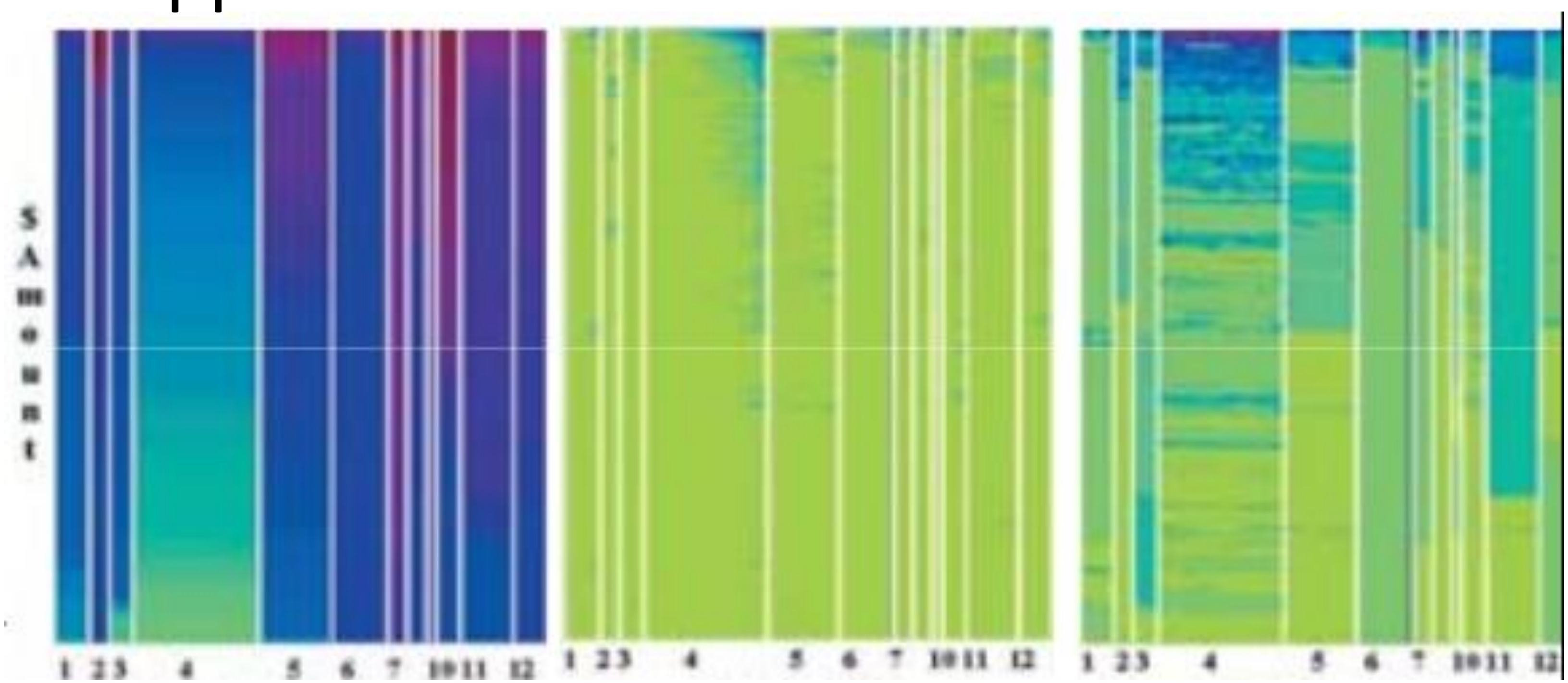
Pixel Bar Charts

- Each pixel of the bar represents a data point belonging to the group represented by this bar



Pixel Bar Charts

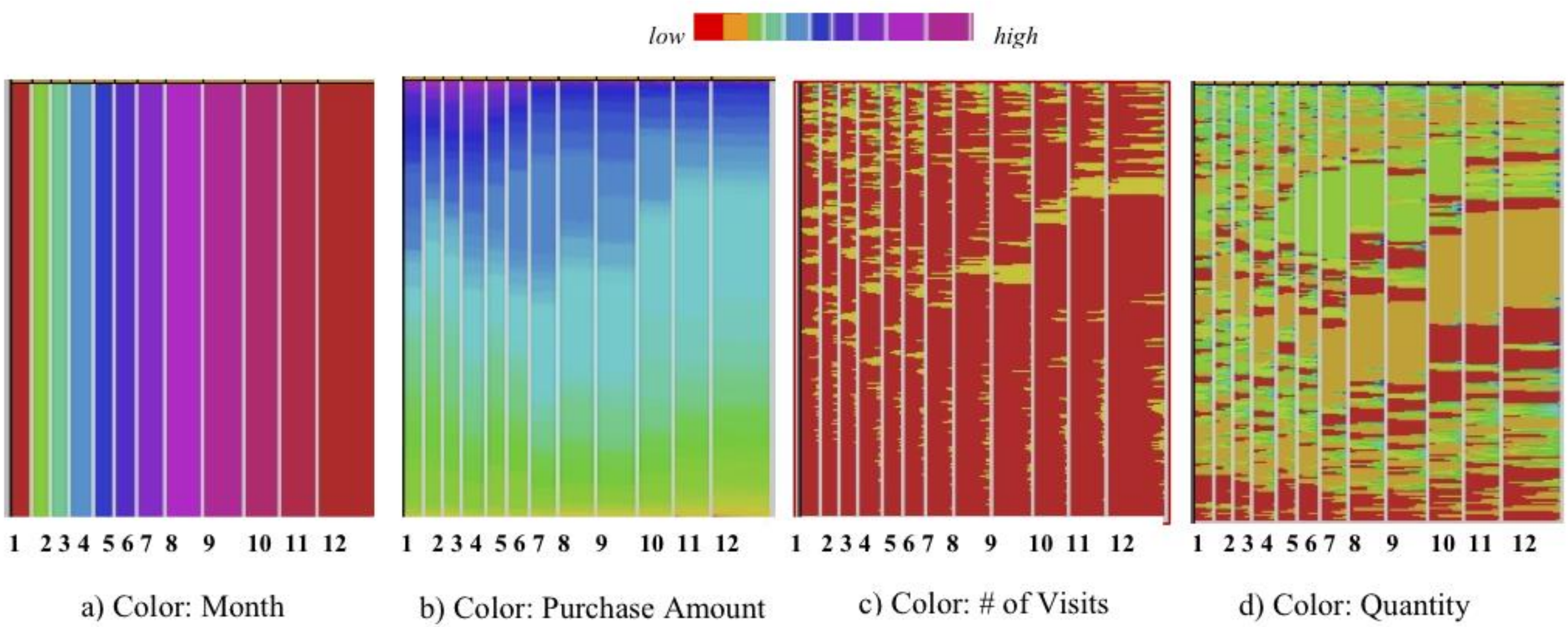
- Internet shopping – relationship between the type of product and the price. Color is mapped onto:



amount spent number of visits size of sales

Pixel Bar Charts

- Placement of dense pixels to bar chart



Pixel Bar Charts

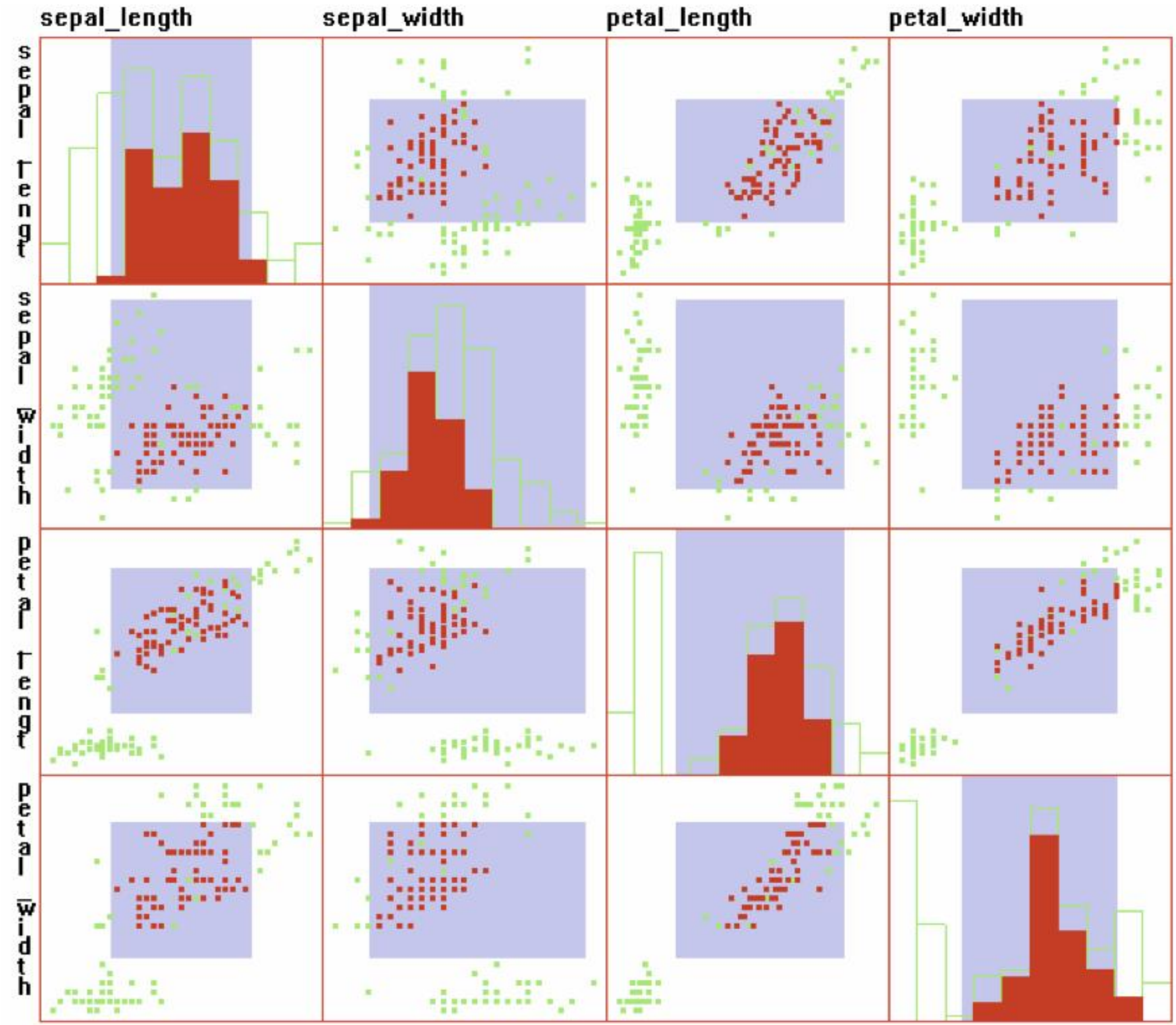
- We can derive, e.g.:
  - The largest amount of customers came in December, while in February, March, and May there was minimum of customers.
  - From February to May there were largest amounts of purchases.
  - Number of purchases in December is average.
  - From March to June the customers returned more frequently than in other months. December customers were mostly one-time customers.
  - Customers shopping the most are returning more often and buying more stuff.

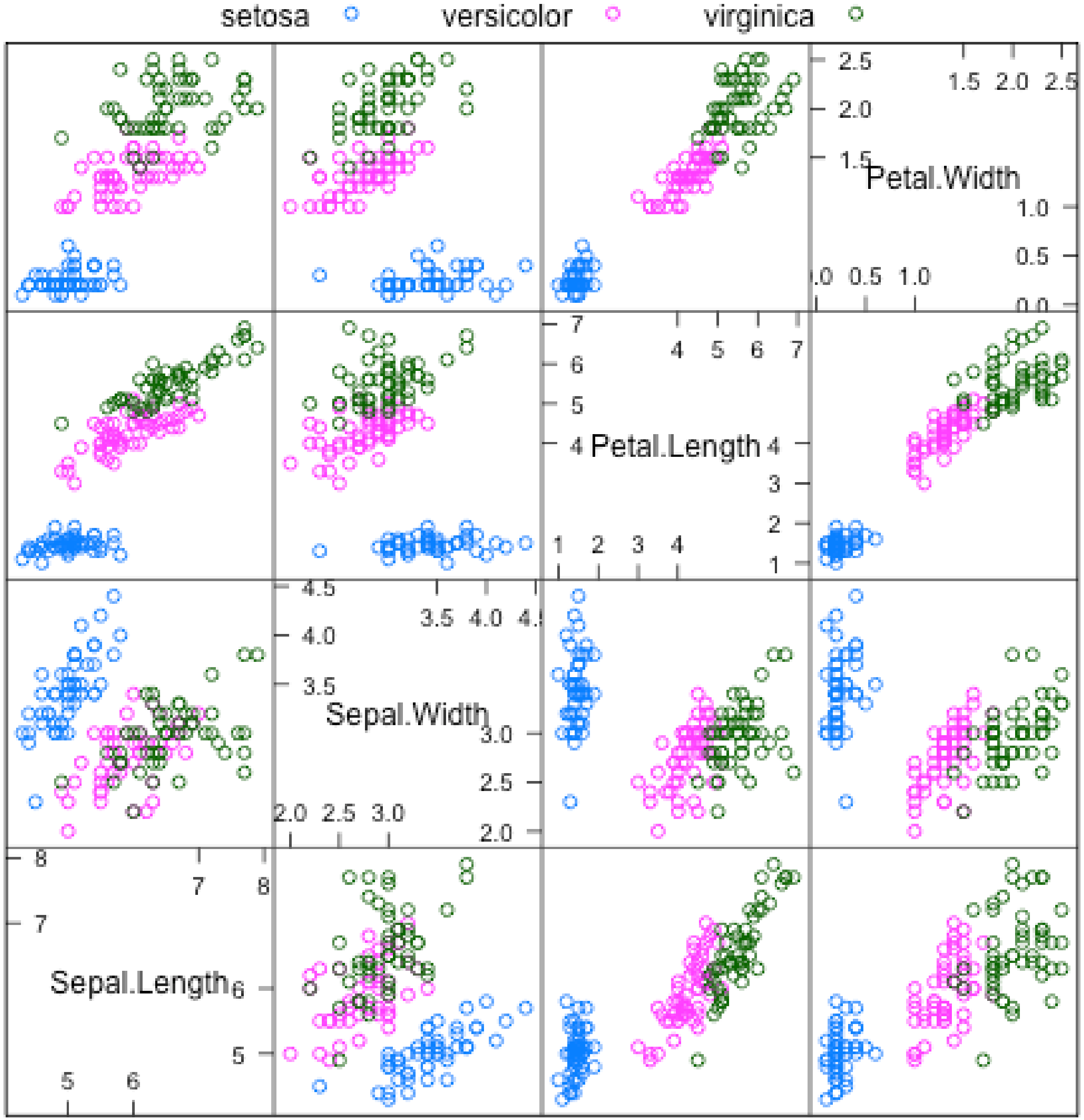
Scatterplots

- One of the first and most used visualization techniques used for data analysis
- Data analysis consists of:
  1. Multiple displays – displaying multiple plots together at once (superimposition, juxtaposition)
  2. Search for a subset of input data dimensions
  3. Dimension reduction (PCA, multidimensional scaling)
  4. Dimension embedding – mapping dimensions onto additional graphical atributes (color, size, shape)

Multiple Displays – Scatterplot matrix

- Grid containing scatterplots
- N^2 cells, where N is the number of dimensions
- Each dimension pair is displayed twice – just rotated by 90°
- Usually symmetric about the main diagonal
- Main diagonal displays
  - Description of corresponding dimension or
  - Histogram of the given dimension


sepal_length
sepal_width
petal_length
petal_width
sepal length
sepal width
petal length
petal width

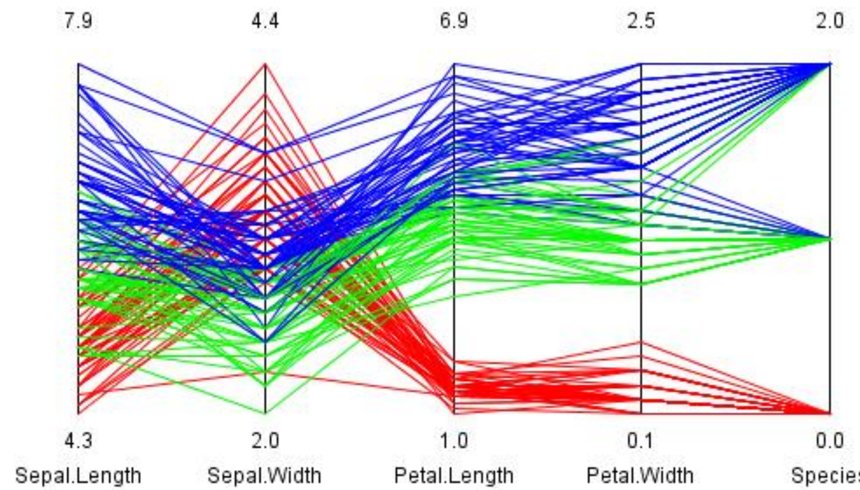


Scatter Plot Matrix

https://machinelearningmastery.com

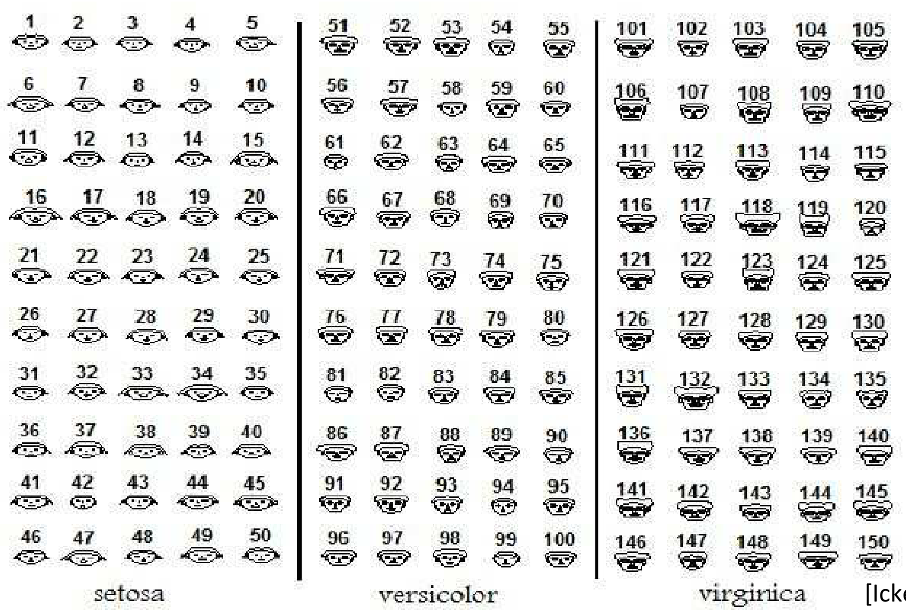
Other representations

- Parallel coordinates



https://www.wanderinformatiker.at

- Chernoff faces

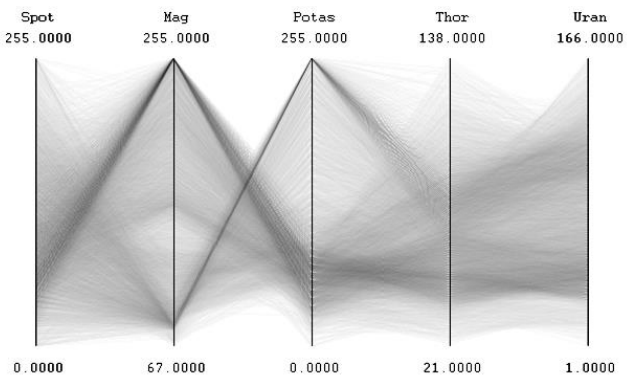


[Icke & Sklar, 2009]

Solving the scalability problem

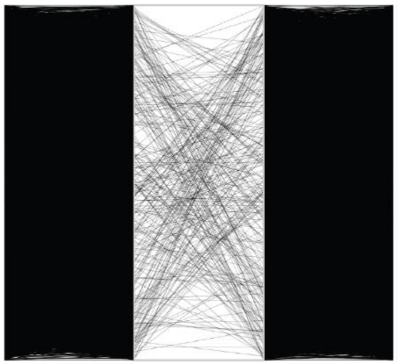
- Solutions for a lot of samples/items/records:

Density-based



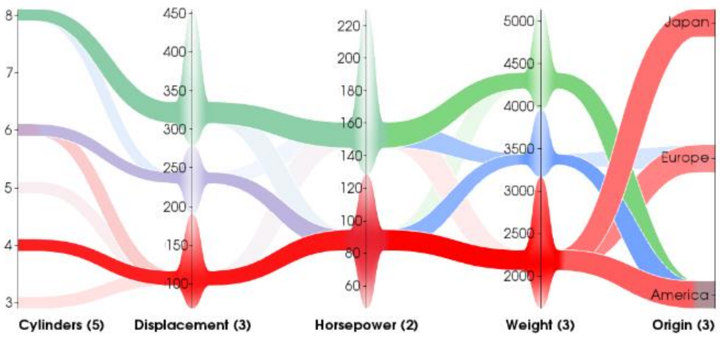
[Florek & Novotny, SCCG 2006]

Random sampling



[Ellis & Dix, CHI-EA 2005]

Edge bundling



[Palmas et al., 2014]

- But what about a lot of dimensions (axes)?

Approaches

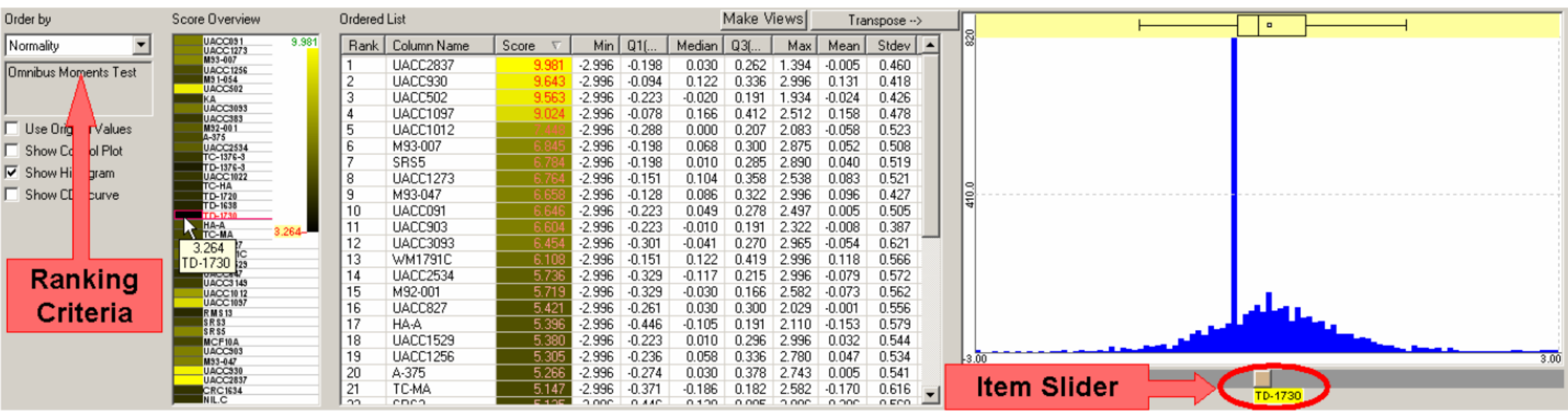
- **Feature selection (dimension subsetting)**
  - Selecting a subset of existing features without a transformation
  - Using multi-dimensional data visualization techniques
- **Feature extraction (dimension reduction)**
  - Transforming existing features into lower dimensional space
  - Using 1D / 2D / 3D /nD visualization technique
- **Hybrid approach**
  - Selecting a subset of existing features
  - Transforming feature subset into lower dimensional space

Feature Selection

- Selecting a subset of existing features without a transformation
- Dimensions (or dimension pairs) are ranked based on quality metric:
  - Number of outliers
  - Correlation between pair of dimensions
  - Image-based
  - …
- Quality metrics can be combined
- Using multi-dimensional data visualization techniques

Rank-by-Feature Framework

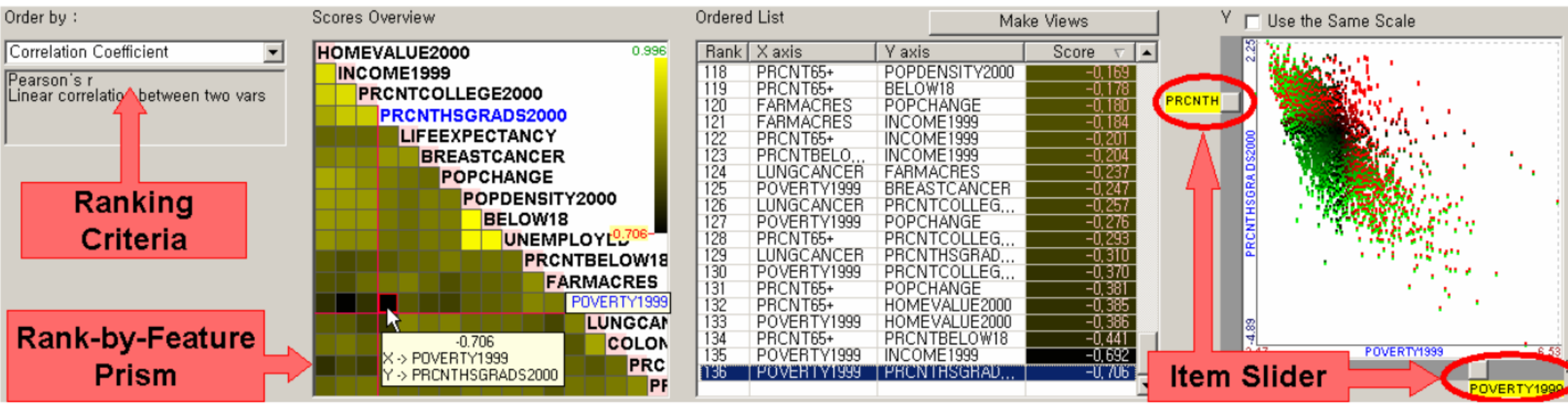
- Exploratory analysis of multidimensional data
- Based on ranking criteria
  - 1D ranking criteria: Normality or uniformity (entropy) of distribution, number of potential outliers, number of unique values



[Seo and Shneiderman, 2004]

Rank-by-Feature Framework

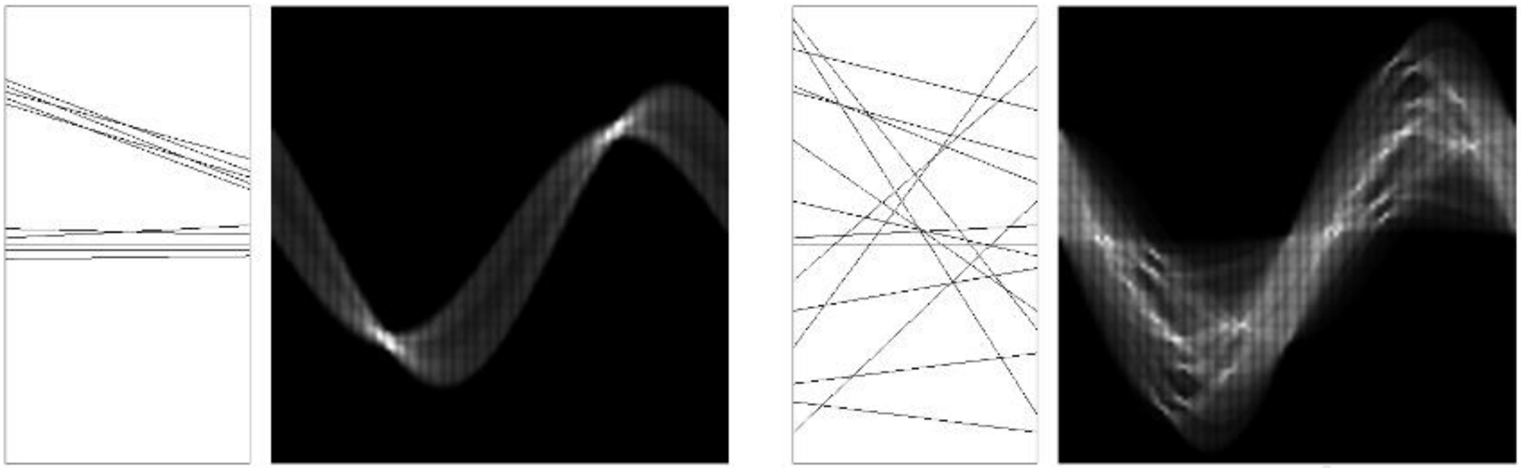
- 2D ranking criteria:
  - Correlation coefficient, least squares error for linear regression / curvilinear regression, number of items in region of interest, uniformity of scatterplots



[Seo and Shneiderman, 2004]

Selection and Ordering of Parallel Coordinate Axes

- Every dimension pair is converted to Hough space

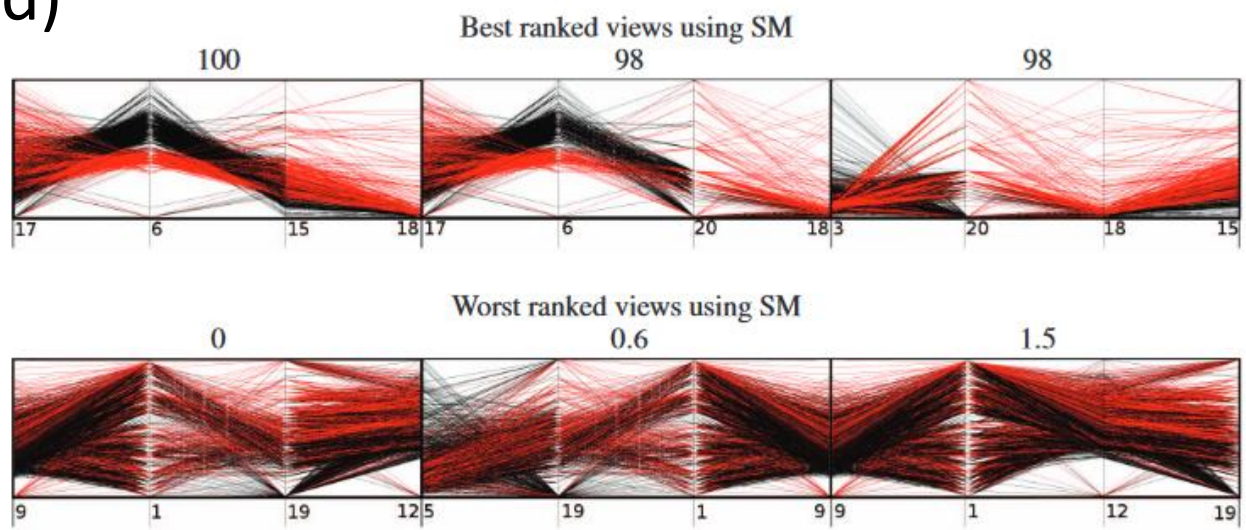


[Tatu et al., 2009]

- Quality metric: good dimension pairs have fewer, well-defined clusters in Hough space

Selection and Ordering of Parallel Coordinate Axes

- Example: dataset of cars
  - 7404 cars
  - 24 attributes
  - Classes separated into benzine (black) and diesel (red)



[Tatu et al., 2009]

Feature Extraction

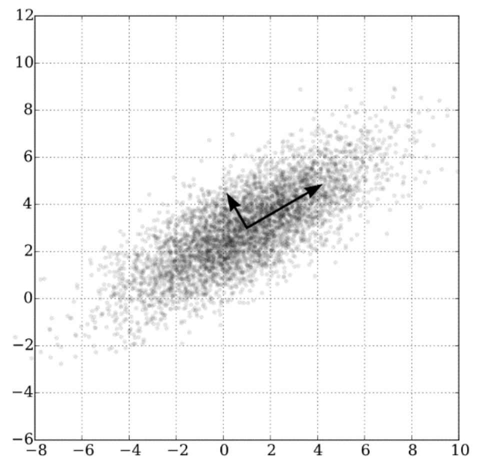
- Transforming existing features into lower dimensional space
- Dimensionality reduction
  - Linear
  - Non-linear
- Using 1D / 2D (/3D)/nD visualization technique

Dimensionality Reduction

- **Linear projection**
  - Linear transformation projecting data from high-dimensional space to low-dimensional space
  - Techniques:
    - Principal component analysis (PCA)
    - (metric) multi-dimensional scaling (MDS)
    - …

Principal Component Analysis

- Projecting data onto lower dimensions (= principal components)
- First principal component: as much variability of the data as possible
- Principal components are orthogonal



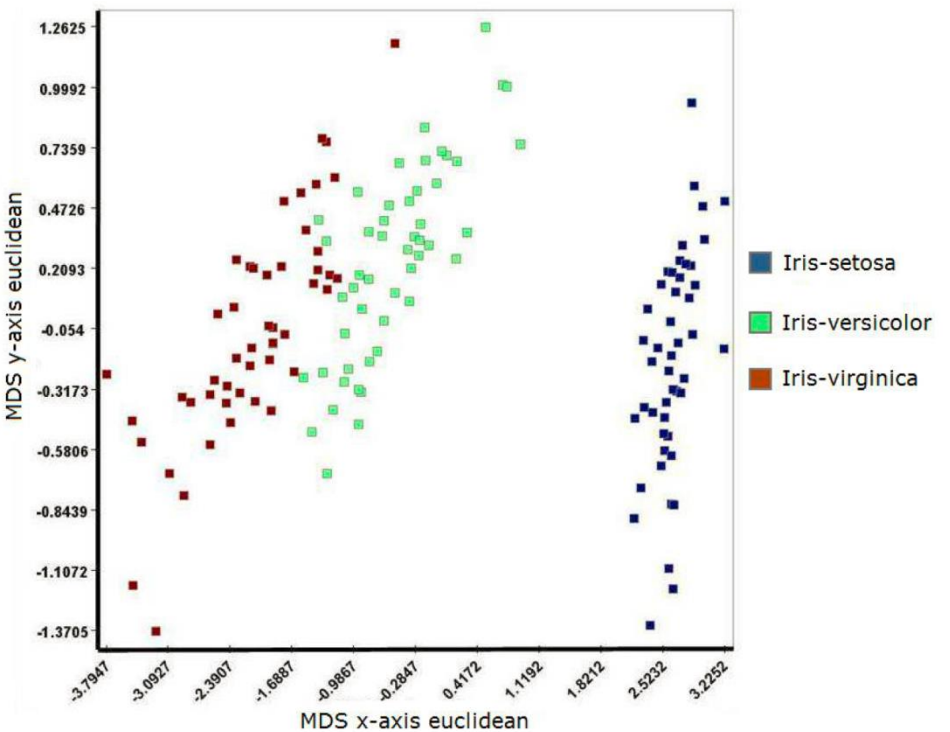
Wikipedia: principal component analysis

Force-Based Methods

- Projection of points from large dimensions into 2D or 3D space
- Aims to preserve the properties of N-dimensional data while projecting to different dimension
- Projection can introduce unwanted artifacts to appear in the resulting visualization

Multidimensional Scaling (MDS)

- Numerous variants of the algorithm exist. The main differences are in:
  - Method for similarity and stress computation
  - Definition of start and end conditions
  - Strategy for updating the position of points

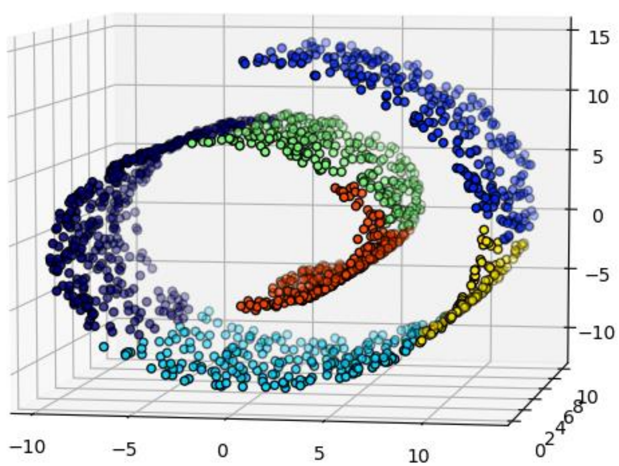


Problems

- Results are not unique – small changes in start conditions can lead to different results
- Coordinate system after the projection may not be easily understandable to the user – with respect to the dimensions of the original data
  - The most significant are the relative positions of individual points rather than their absolute positions, which may differ from algorithm to algorithm

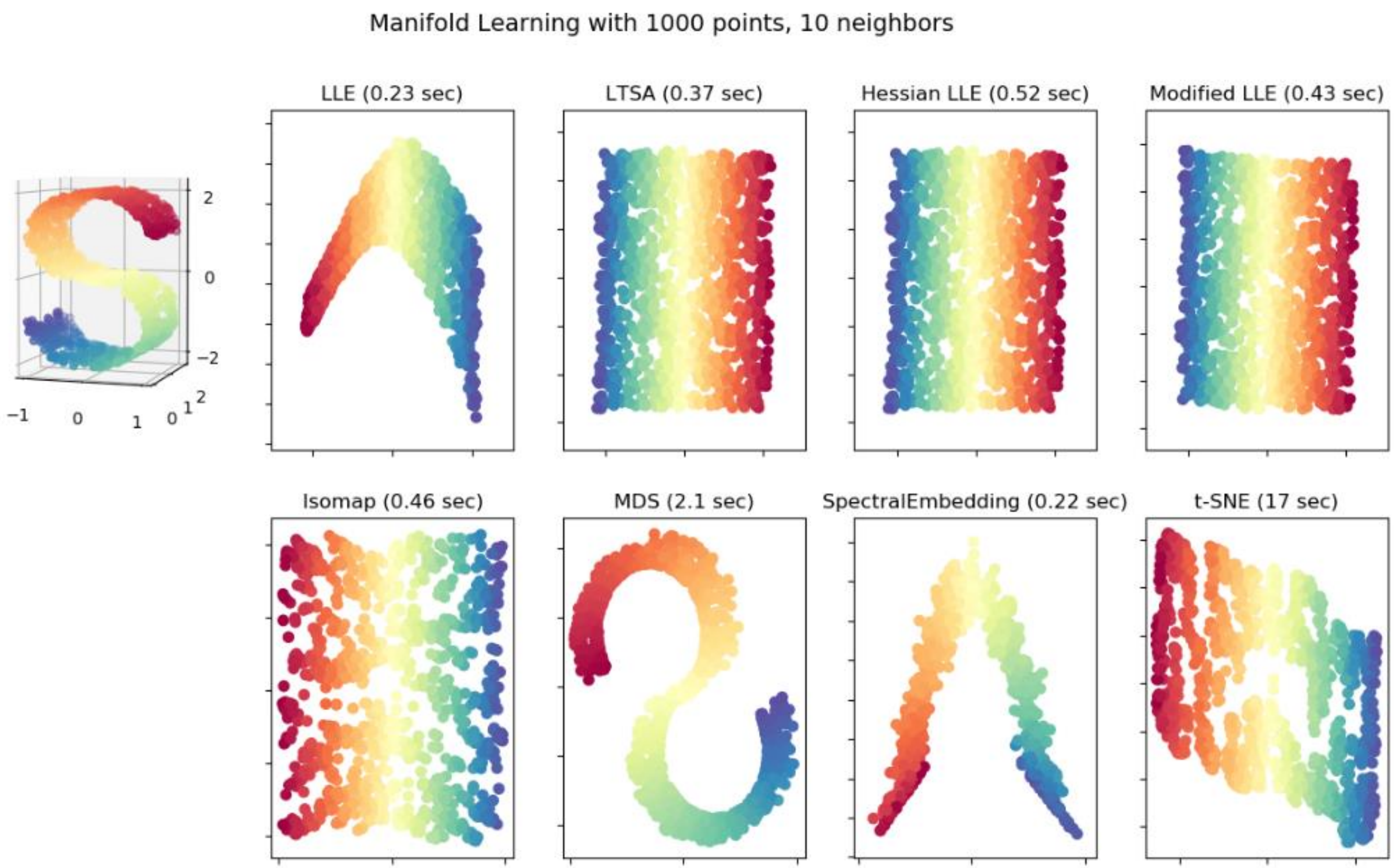
Non-Linear Dimensionality Reduction

- Low-dimensional surface embedded non-linearly in high-dimensional space
- Preserves the neighborhood information
  - Locally linear
  - Pairwise distances



„swiss roll“ http://scikit-learn.org

Non-Linear Dimensionality Reduction



scikit-learn.org

Comparison of Manifold Learning methods

Hybrid Approaches

- Dimensionality reduction often unwanted because domain knowledge is required to understand which dimension combinations make sense
- Combination of feature selection and feature extraction
- Feature selection:
  - User selection based on visual analysis
  - Quality metrics
- Feature extraction is performed on selected dimensions
- Using multi-dimensional data visualization techniques