3. Multidimensional Information Visualization II
Concepts for visualizing univariate to hypervariate data

Lecture „Informationsvisualisierung”

Prof. Dr. Andreas Butz, WS 2012/13

Concept and slides: Thorsten Büring,

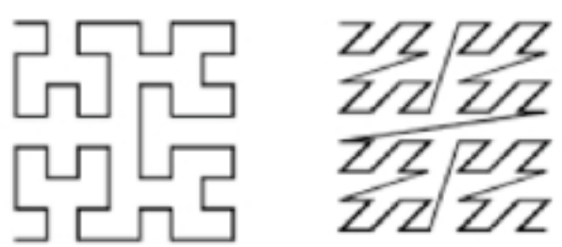
3rd, revised edition

Outline

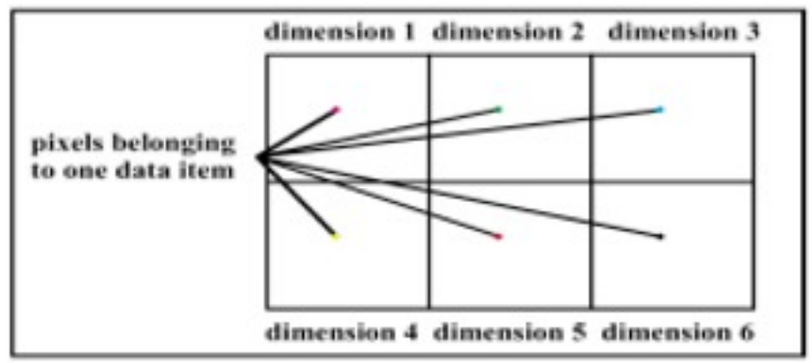
- Reference model and data terminology
- Visualizing data with < 4 variables
- Visualizing multivariable data
  - Geometric transformation
  - Glyphs
  - Pixel-based
  - Downscaling of dimensions
- Case studies: support for exploring multidimensional data
  - Rank-by-feature
  - Dust & magnet
- Clutter reduction techniques
- Exploring Design Decisions - demo

Pixel-Based Techniques

- Idea: each data value is represented by one colored pixel
- Value ranges are mapped to a fixed color sequence of full color (hue) scale but monotonically decreasing brightness
- Data values belonging to one attribute are displayed in a separate view – only one pixel per data value without need for a border
- But: users need to relate to different portions of the screen to perceive correlations
- Optimization Goal (OG) 1: arrangement of pixels in the subwindows should preserve the 1D ordering into 2D plane as best as possible
  - Simple left-right or top-down arrangement do usually not provide useful results on a pixel-level
  - Space-filling curves (Paneo-Hilbert and Morton) provide maximum of locality preservation, but are difficult to follow and thus to relate between the subwindows
- Recursive pattern
- Circle segments



Paneo-Hilbert Morton



Recursive Pattern

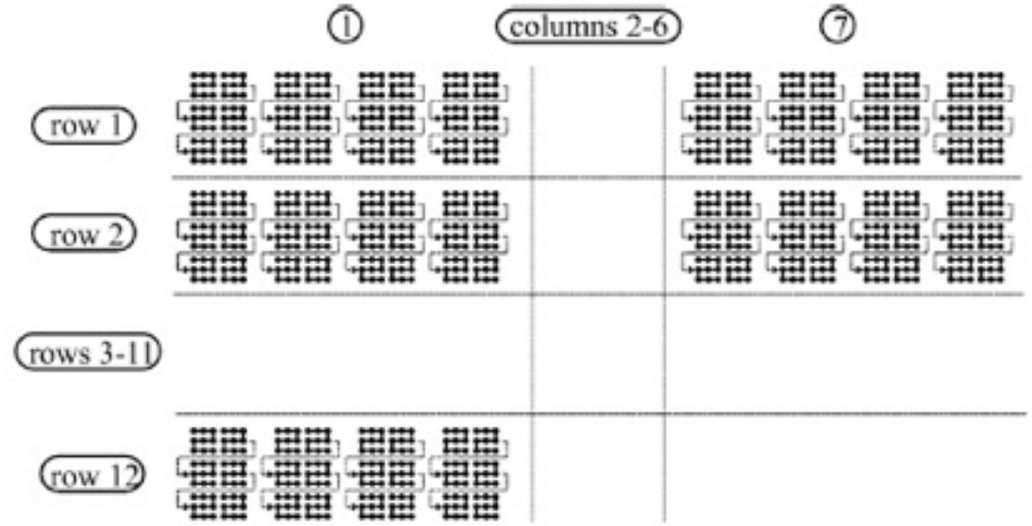


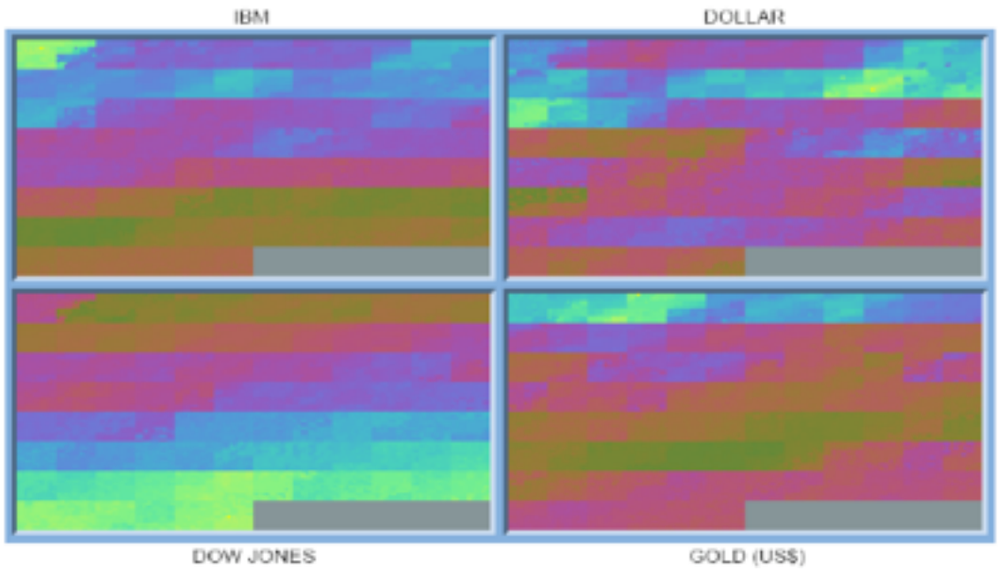
Fig. 9. Schematic representation of a highly structured arrangement $[(w_1, h_1) = (3, 3),\ (w_2, h_2) = (2, 3),\ (w_3, h_3) = (4, 1),\ (w_4, h_4) = (1, 12),\ (w_5, h_5) = (7, 1)]$. (Adapted from [39] ©IEEE.)

Keim 2000

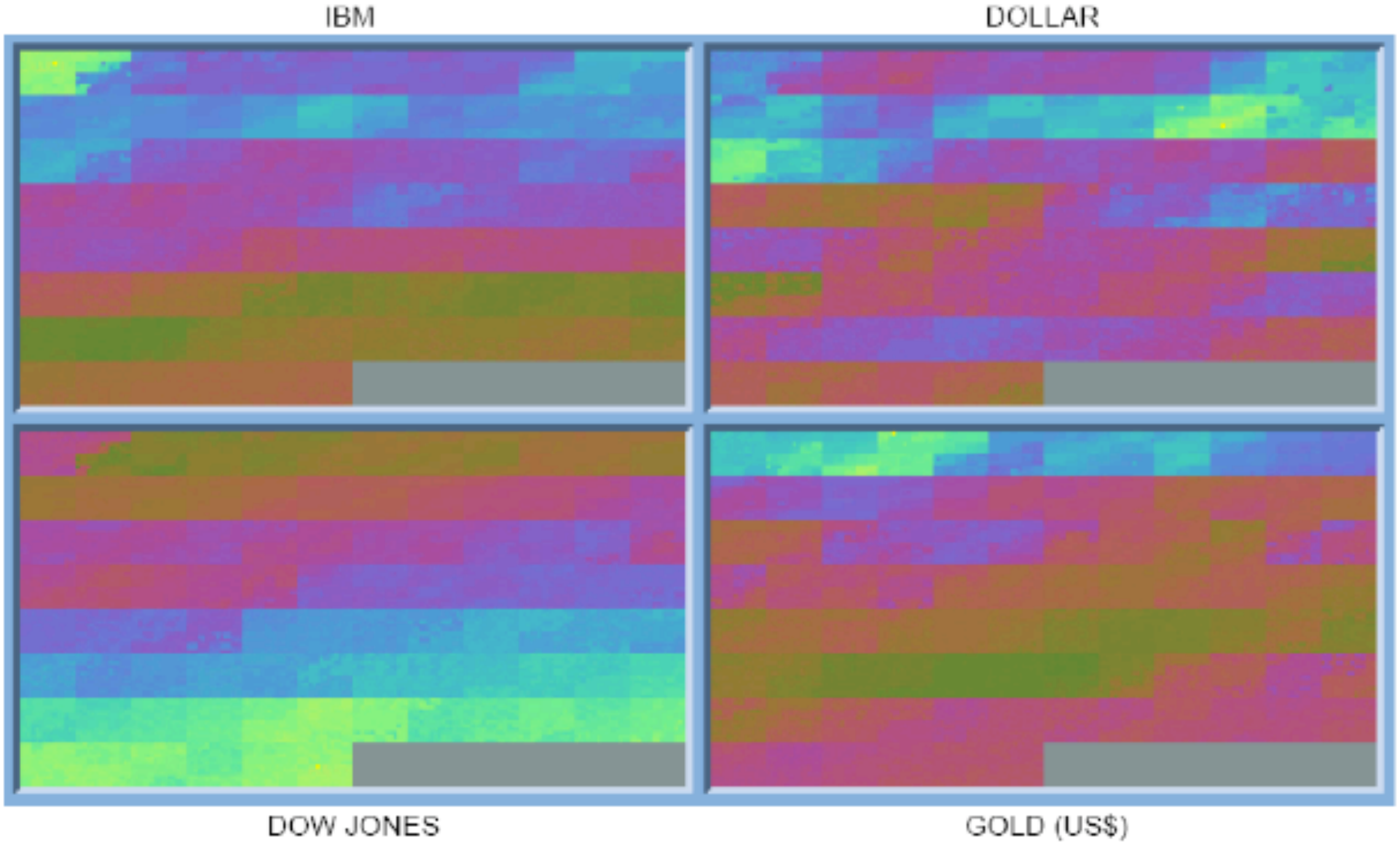
- Keim et al. 1995
- Naturally ordered data set
  - Prices of IBM stock, Dow Jones index, Gold, exchange rate US-Dollar
  - September 1987-February 1995
  - 9 daily measurements for each stock
- Recursive pattern visualization
  - Lower-level patterns used as building blocks for higher level patterns
  - LP1 one day, LP2 one week, LP3 one month, LP4 one year

Recursive Pattern

- 8 horizontal bars correspond to 8 years
- Subdivisions between the bars represent 12 months within each year
- Example analysis results
  - Gold price was very low in the sixth year
  - IBM price fell quickly after the first 1 ½ month
  - US-Dollar exchange rate was highest in the third year



Keim et al. 1995

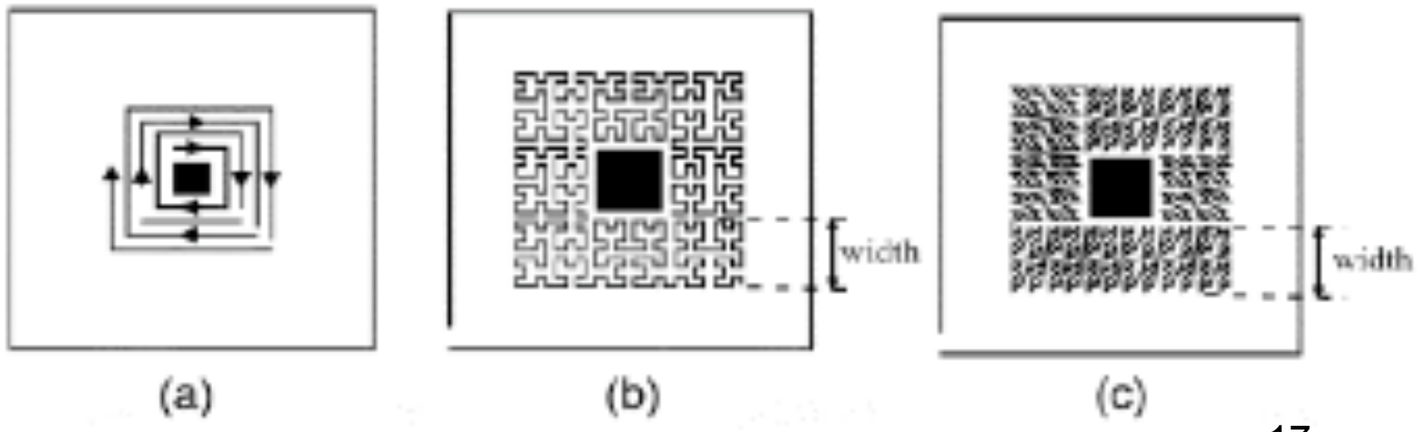


Keim et al. 1995

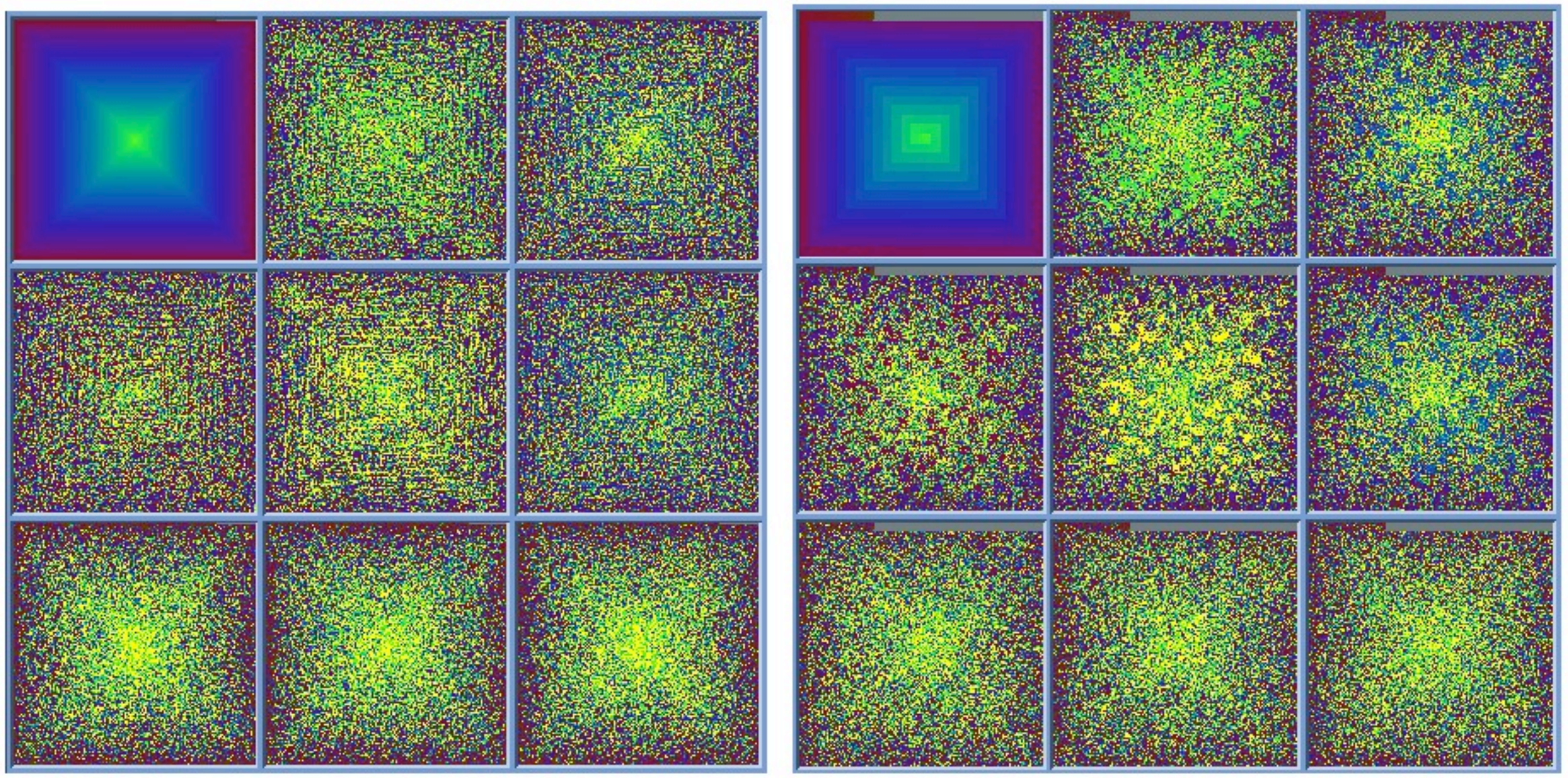
Query Dependent Arrangement

- Ordering of data objects based on relevance to a given query
- Most relevant data object is placed in the center of the screen
- OG 2: for the pixel arrangement in each subwindow the distance to the center should correspond to the ordering of the data objects
- Simple spiral arrangement fulfills OG 2, but local clustering properties (OG 1) are weak, i.e. low probability that two pixels close on the screen are also close in the 1D ordered sequence of the query result set
- Generalized spiral technique: enhance the clustering qualities of the spiral technique by using screen-filling curves locally

Keim 2000

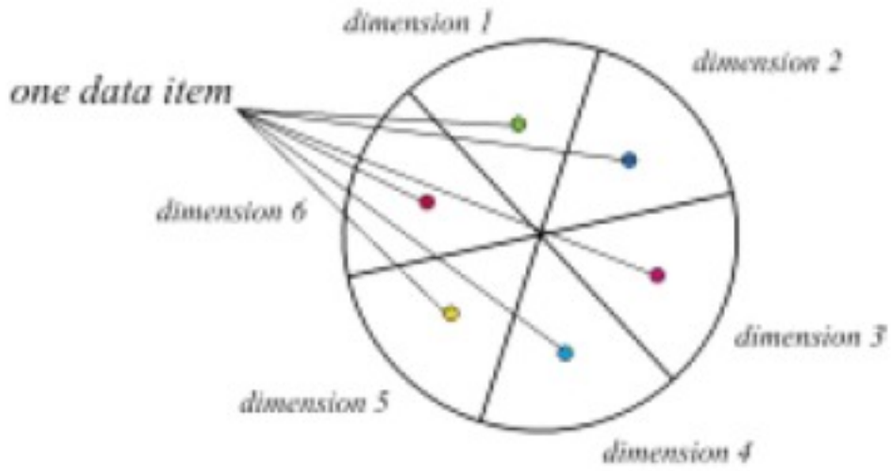


Spiral vs. Generalized Spiral

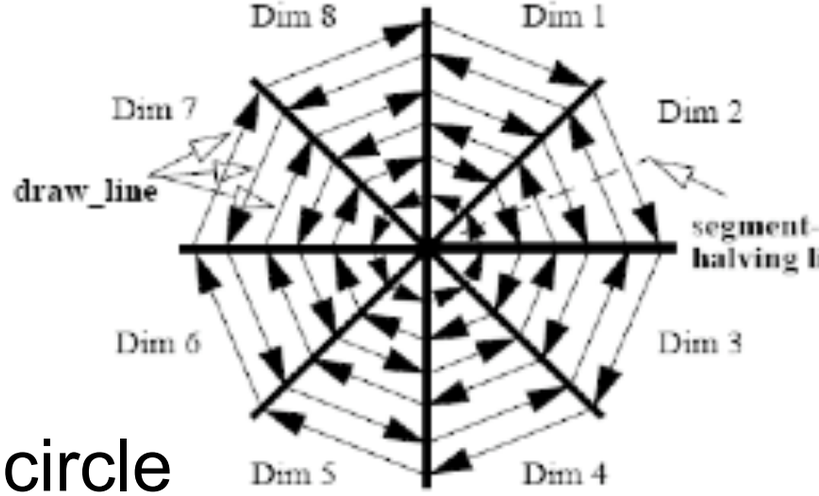


Keim1996

Circle Segment



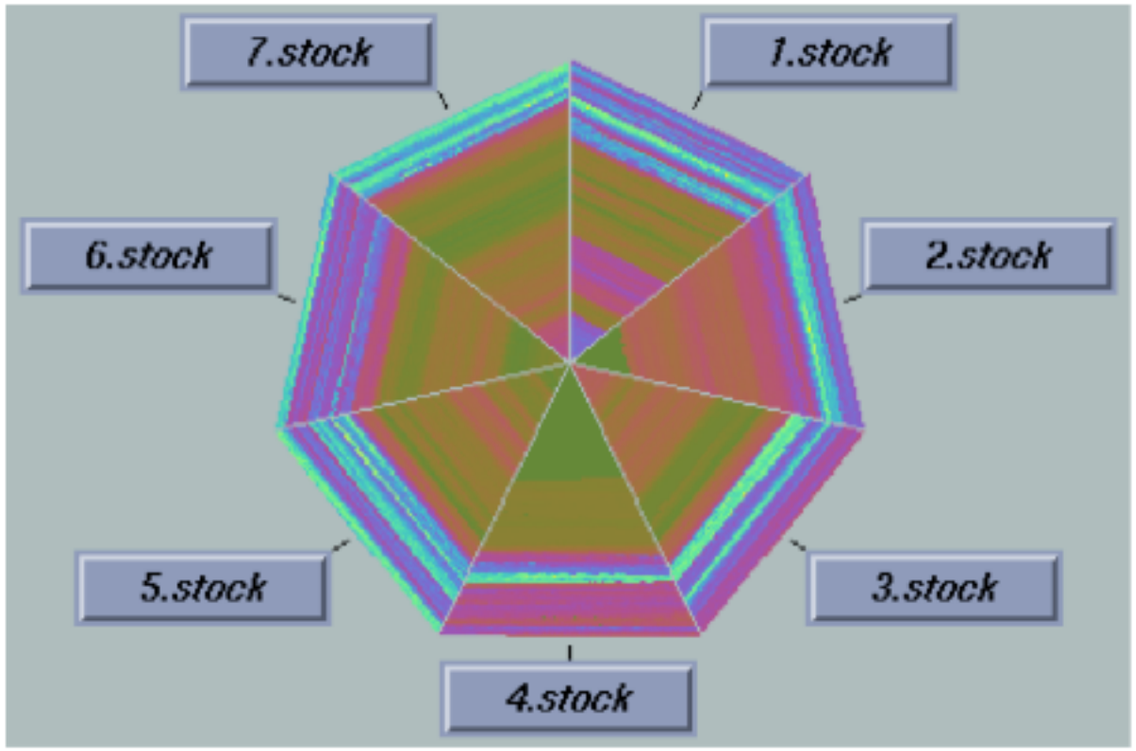
- Rethink the shape of subwindow
- Rectangular shape of subwindows makes efficient use of the screen
- For data sets with many dimensions, the pixels of one data object are rather far apart
- Makes it difficult to find patterns
- OP 3: minimize the average distance between the pixels (data values) belonging to one case

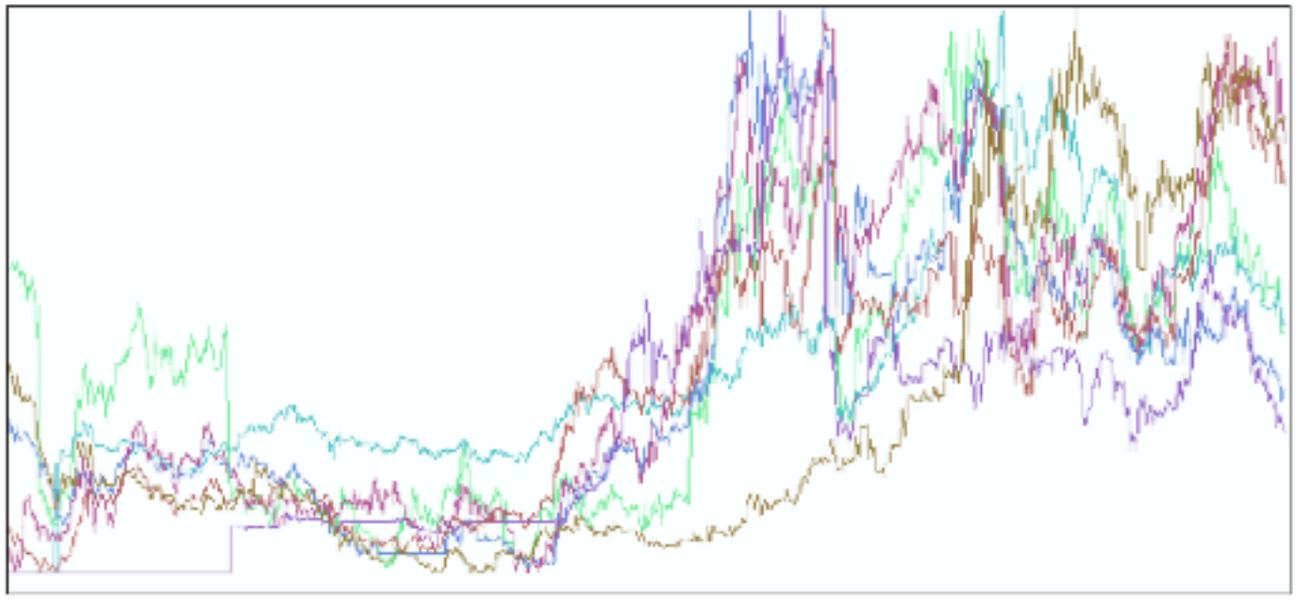


- Circle segment
- Each dimension corresponds to a segment of a circle
- Values of one dimension are drawn in a back and forth manner from the center of the circle to the outside

Circle Segment (CS) vs. Line Graphs

- 10 years of stock data for 7 stocks
- Line graph granularity is limited by the width of the screen
- CS: oldest data items in the middle of the circle, most recent ones are at the outside
- Easier to perceive patterns – no overlap of data





Keim et al. 1996b

Pixel-Based Techniques

- Advantages:
  - Large data sets can be visualized
  - Improved pattern detection due to non-overlap strategy
- Disadvantages
  - Labeling
  - Intuitiveness
  - Mapping color to quantitative data
  - ?
- Open questions
  - Drill-down to detail information?
  - ?

Downscaling of Dimensions

- Projecting n dimensions down to a lower dimensionality while retaining as much of the original information as possible
- Principal components analysis, Factor analysis, Multidimensional scaling
  - Statistical approaches to reduce the number of dimensions by finding the data's main characteristics / patterns
  - Good tutorial: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- Self-organizing maps (SOM) aka Kohonen map
  - Reduce the dimensions of data by using self-organizing neural networks
  - Produces usually a 2D map which mirrors the similarity of cases (similar cases are grouped together)
  - Good tutorial: http://davis.wpi.edu/~matt/courses/soms/
- Problems: pruning of information; hard to interpret since display coordinates have no semantic meaning; SOM and MDS are iterative approaches (computationally hard, no unique result)


forest
soil
climat
chang
urban
model
spatial
temperatur
popul
transport
plan
city
data
precipit
surfac
gis
sediment
weather
state
hous
lake
measur
migrat
health
remot
beach
channel
flow
immigr
spatial
data
water
river
site analysi
home
flood
town
state
women
work
commun
manag
map
maps
geographi
polit
work
social
environment
develop
tourism
natur
coloni
natur
space
local
econom
region
student
geograph
american
landscap
polit
global
geographi
place
ident
nation
state
servic
product
industri
cultur
market

Skupin 2004

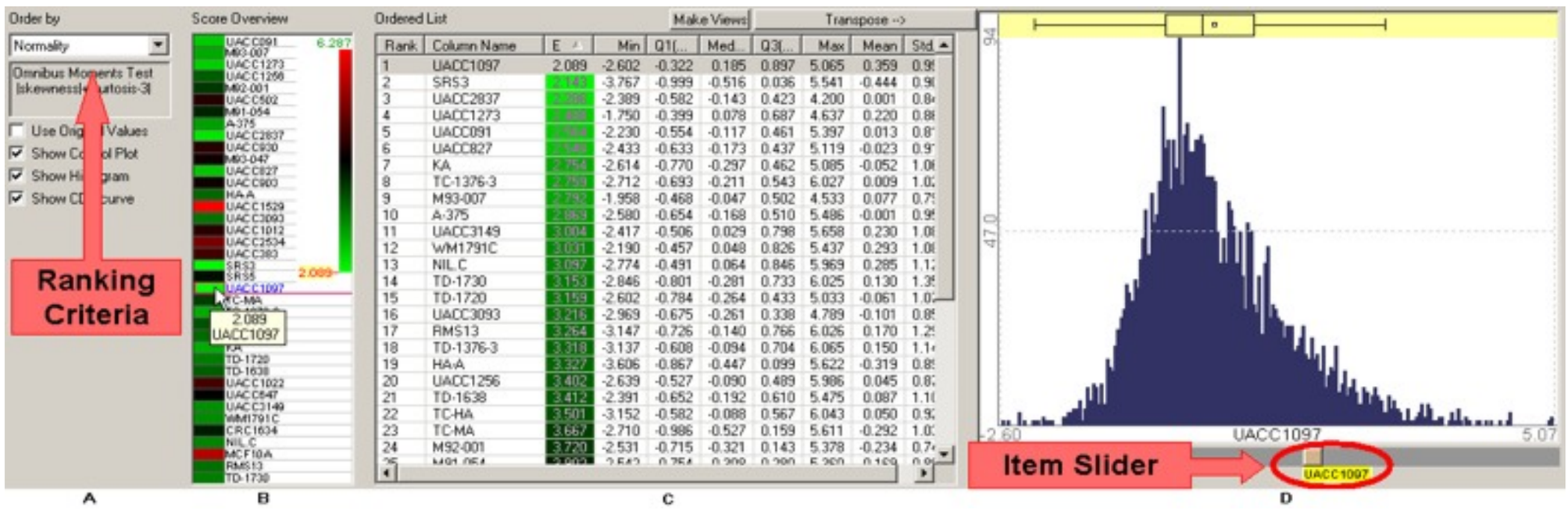
Rank-By-Feature

- Seo & Shneiderman 2004
- Part of the Hierarchical Clustering Explorer (HCE) (http://www.cs.umd.edu/hcil/multi-cluster/)
- Tabs: histogram and scatterplot ordering
- Implements systematic approach for data exploration
  - (1) study 1D, study 2D, then find features
  - (2) ranking guides insight, statistics confirm
- Tool provides low-dimensional projections as a histogram (1D) or scatterplot (2D)
- Users can select a feature detection criterion (e.g. test for normal distribution (1D), correlation coefficient (2D)) to rank projections
- The ranking facility is particularly helpful when the number of possible projections is too large to investigate: concentrate on the interesting ones

Rank-By-Feature

Rank-By-Feature

- Users start with 1D projections (histogram ordering)
- Four coordinated views
  - A: selection of ranking criterion
  - B: overview of scores for all dimensions (color coding: the brighter the color, the higher the score)
  - C: numerical / statistical detail for each dimension (e.g. score, mean, standard deviation)
  - D: display of histogram + boxplot (minimum, first quartile, median, third quartile, maximum)

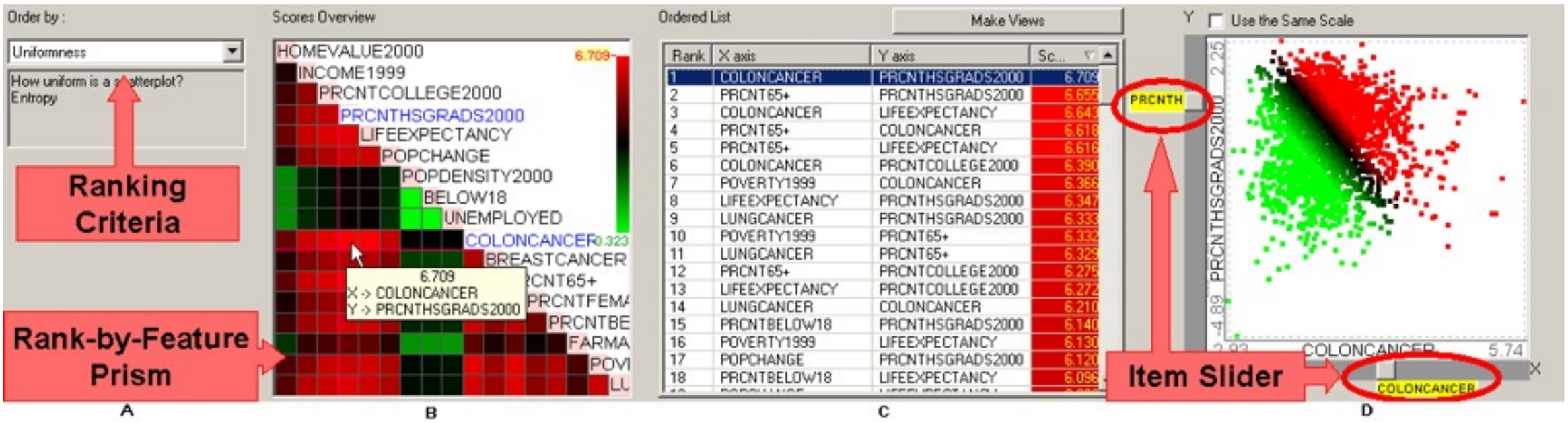


Rank-By-Feature

- Some basic statistical terms
  - Mean: Sum of all values divided by the number of values
  - Median: Middle value of a distribution of values when ranked in order of magnitude
  - Mode: Single most common value
  - Variance: average squared deviation between the mean and the values
  - Standard deviation: square root of the variance (translates the variance into the original units of measurement)
- Statistical tests supported by HCE for 1D ranking
  - Normality of the distribution: distribution of items forms a symmetric, bell-shaped curve
  - Uniformity of the distribution: all of the values of a random variable occur with equal probability (results in a flat histogram)
  - Number of potential outliers
  - Number of unique values

Rank-By-Feature

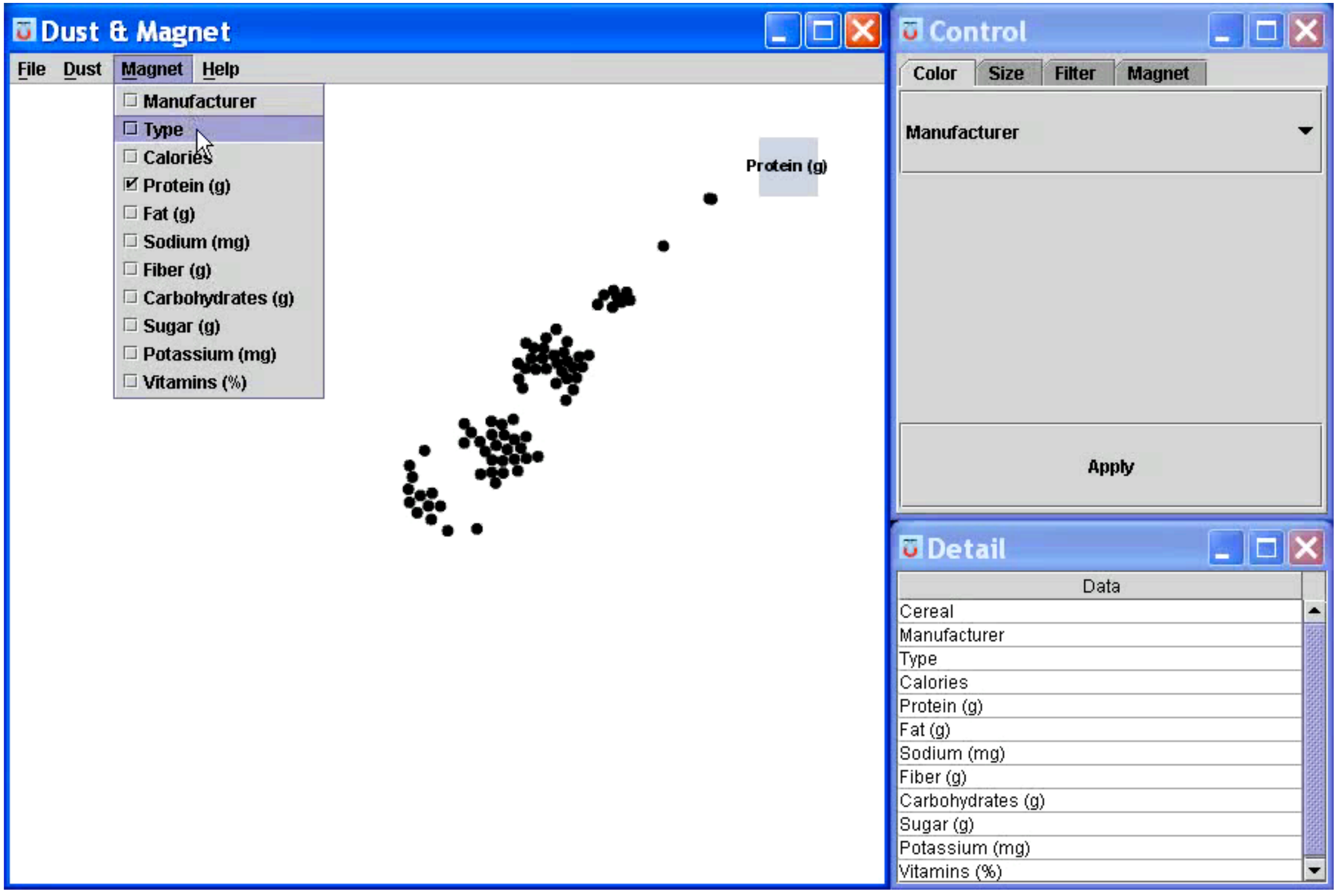
- Move on to 2D projections (scatterplot ordering)
- Identify pairwise relationships between dimensions
- B: prism provides overview of scores for dimension pairs; score is color coded
- D: scatterplot browser; multiple browsers are possible;
- Ranking creteria
  - Correlation coefficient: direction and strength of linear relationship
  - Least square error for simple linear / curvilinear regression: how well does the regression model fit
  - Number of items in a user-defined region of interest & uniformity of scatterplot



Dust & Magnet

- Yi et al. 2005
- Data cases are represented as particles of iron dust
- Magnets represent the dimensions of the data set
- Users manipulate the magnets to move the dust
- Dust moves at different speed depending on its data values for the magnet dimensions

Dust & Magnet

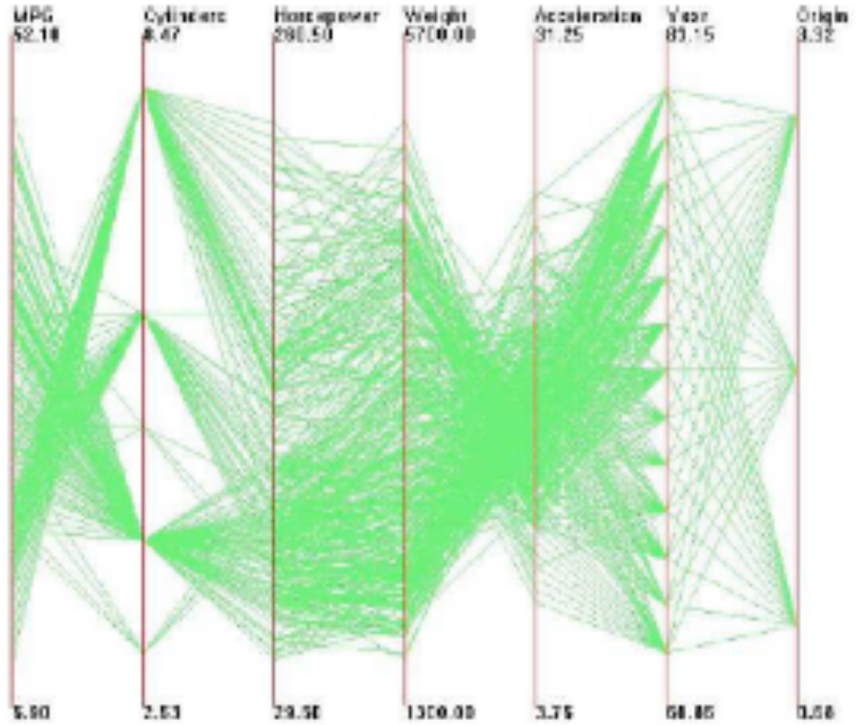

Dust & Magnet
File
Dust
Magnet
Help
Manufacturer
Type
Calories
Protein (g)
Fat (g)
Sodium (mg)
Fiber (g)
Carbohydrates (g)
Sugar (g)
Potassium (mg)
Vitamins (%)
Protein (g)
Control
Color
Size
Filter
Magnet
Manufacturer
Apply
Detail
Data
Cereal
Manufacturer
Type
Calories
Protein (g)
Fat (g)
Sodium (mg)
Fiber (g)
Carbohydrates (g)
Sugar (g)
Potassium (mg)
Vitamins (%)

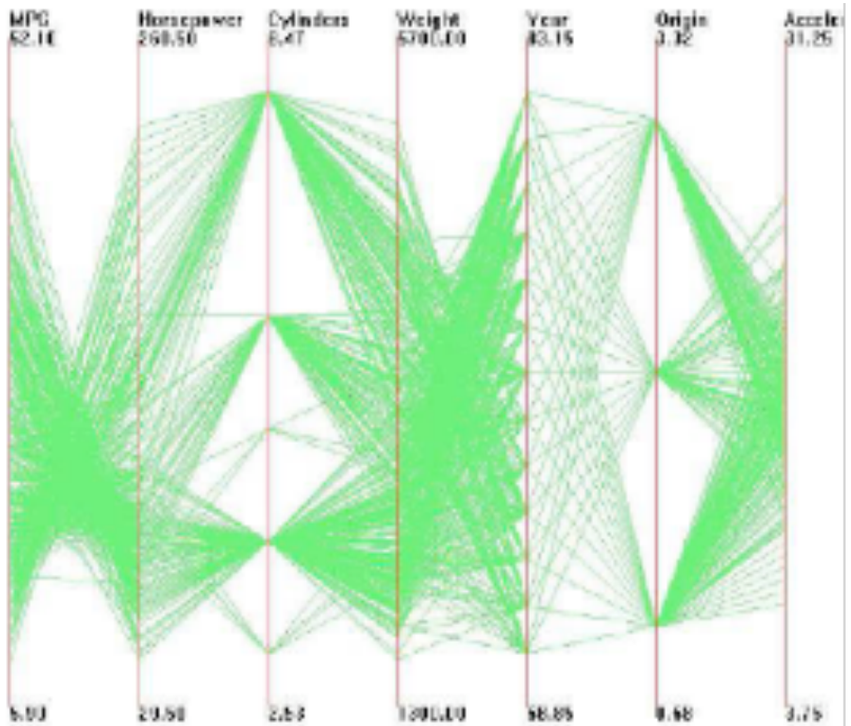
Clutter Reduction Techniques

- Clutter: crowded and disordered visual entities that obscure the structure in visual displays
- Avoiding clutter is one of the main challenges in Information Visualization
- Techniques to reduce clutter
  - Dimension reordering
  - Sampling
  - Constant-density visualization
  - Point displacement
  - Aggregation / Clustering (one mark represents more than one case, e.g. group day sales into months)
  - Filtering (Dynamic queries, zooming)

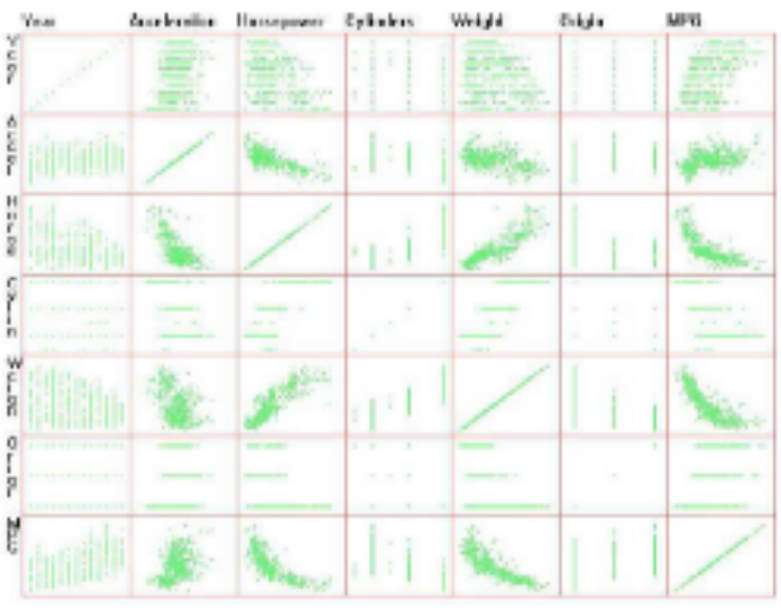
Dimension Reordering

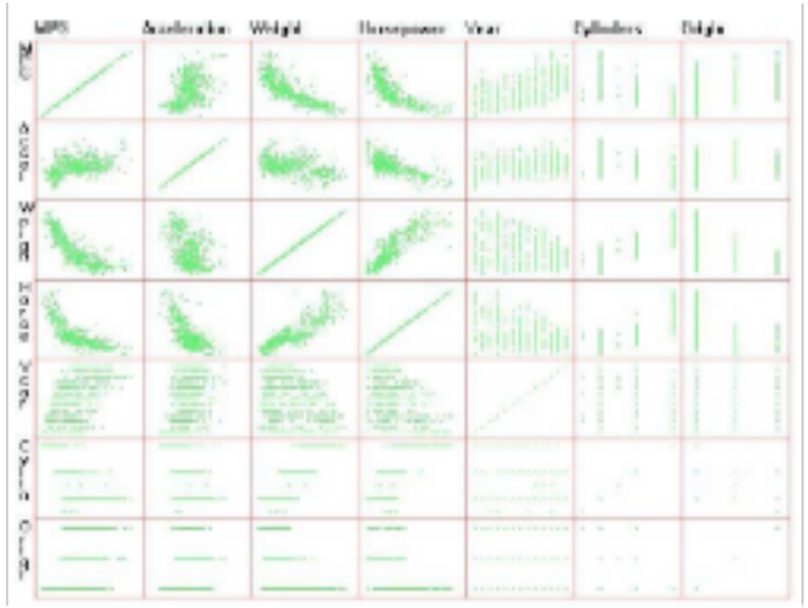
- Peng et al. 2004
- Automatically identify the views with the least amount of visual clutter
- Clutter definition and algorithms for different visualization techniques

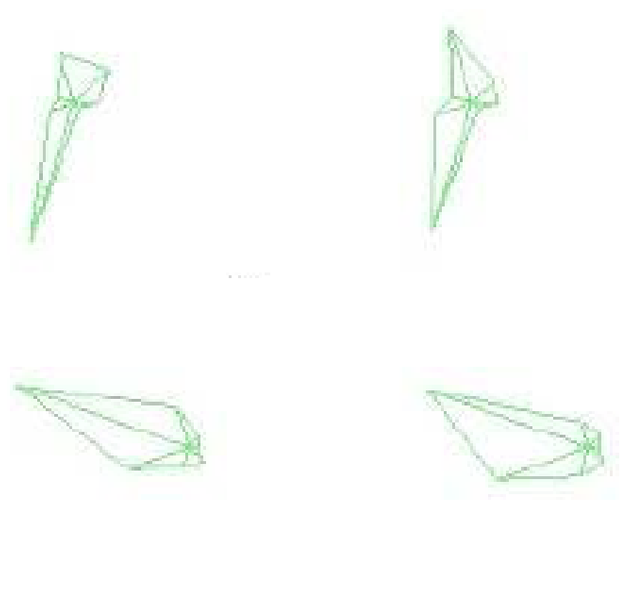


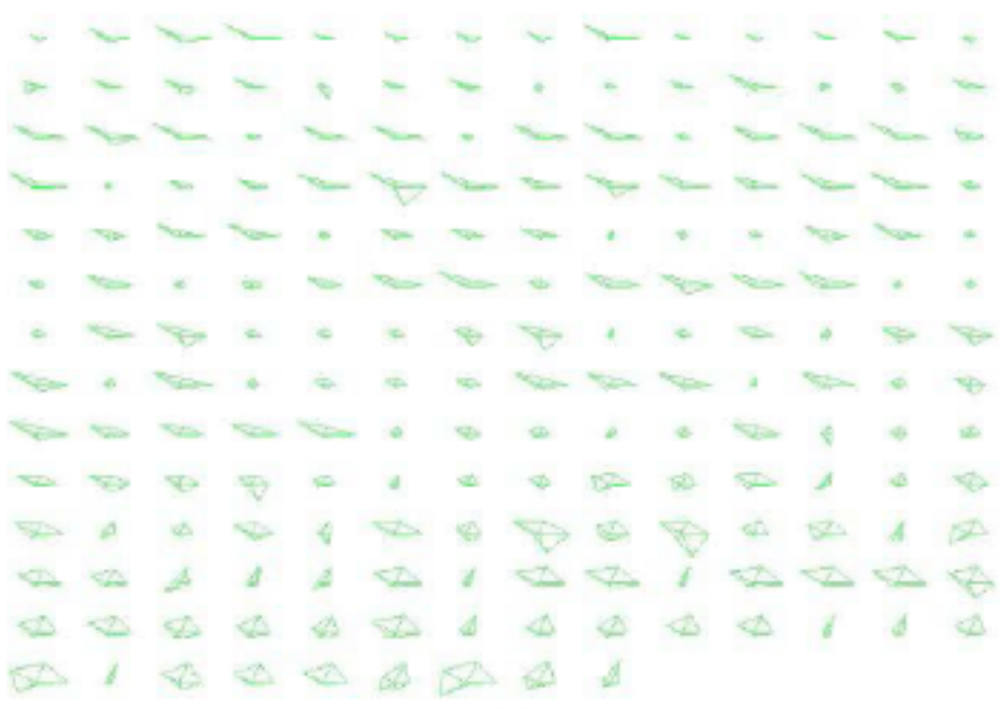


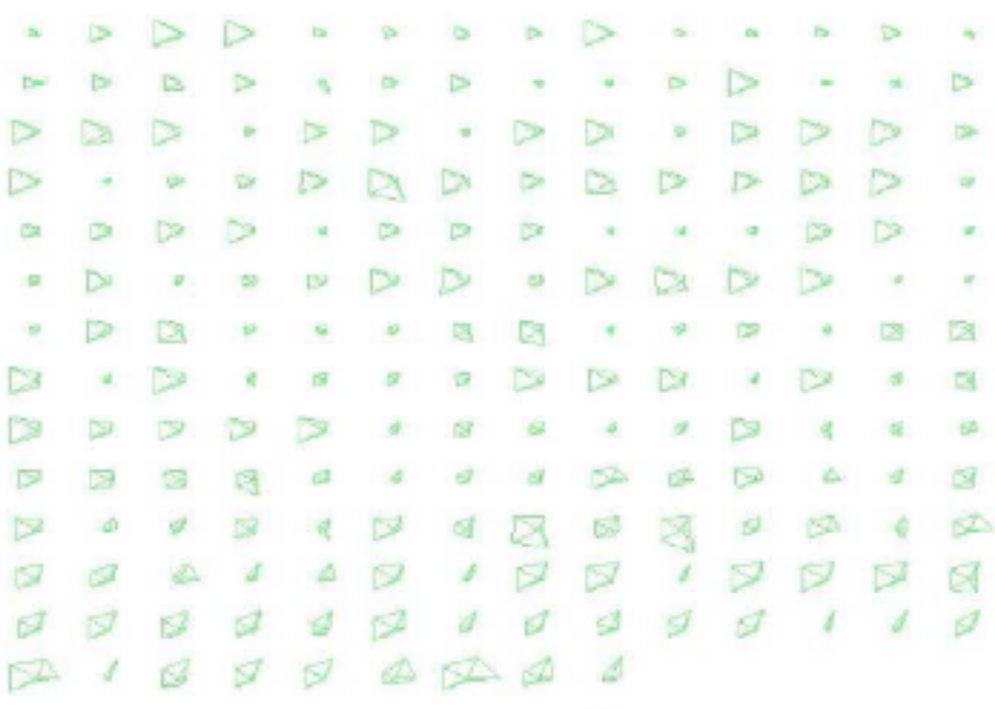
Dimension Reordering







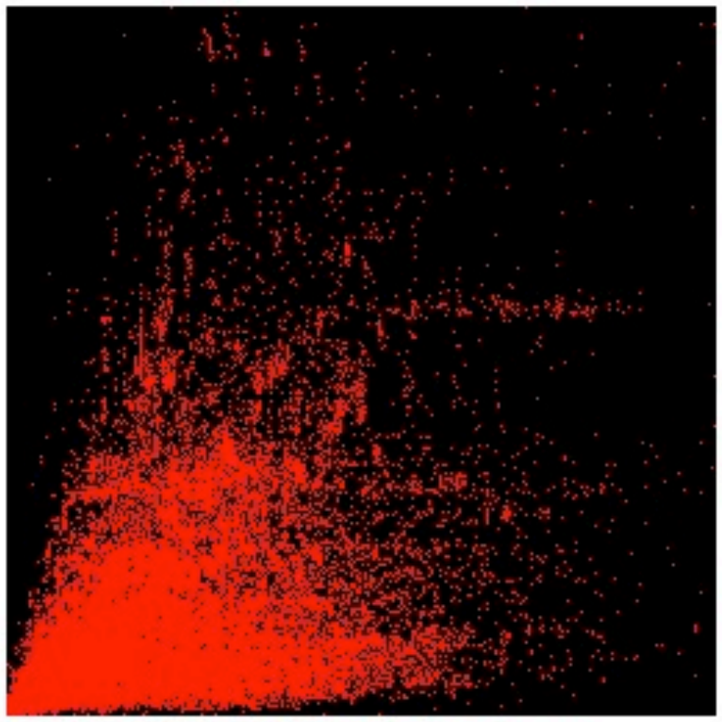




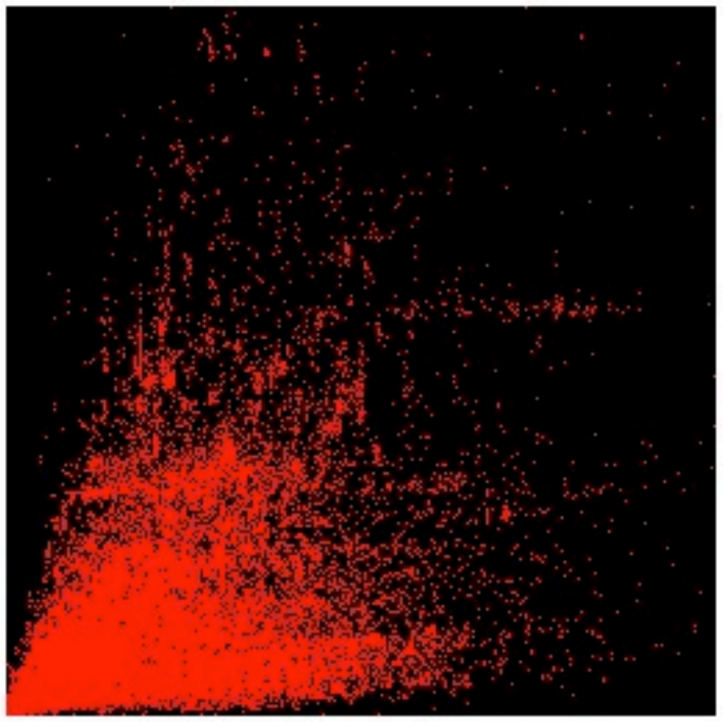
Sampling

- Reduce the density of visual representation by displaying a random subset of the data
- Random sampling preserves the distribution of data
- Overall trends (e.g. correlation) can still be detected at a reduced density
- Uniform sampling (Ellis & Dix 2004)
  - Applying the same (manually or automatically defined) sampling factor to the entire data space
  - Problem: areas with low density may become empty
- Non-uniform sampling (Bertini & Santucci 2004)
  - Preserving relative density
  - Model to compute where, how, and how much to sample to preserve image characteristics

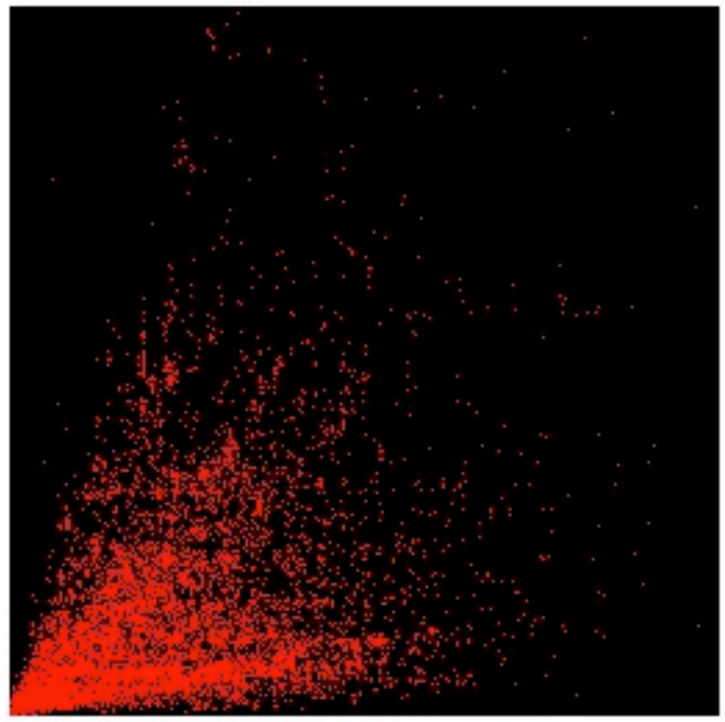
Bertini & Santucci 2004



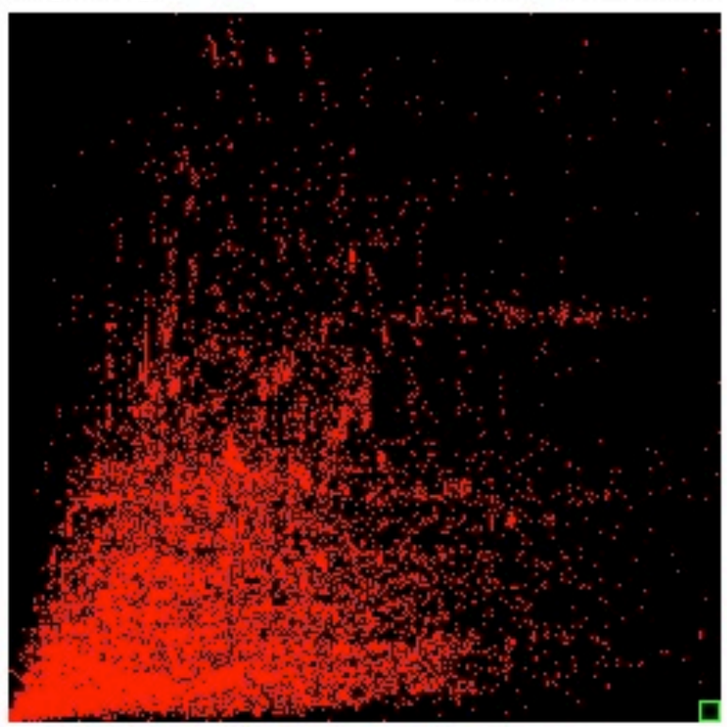
(a) - no sampling



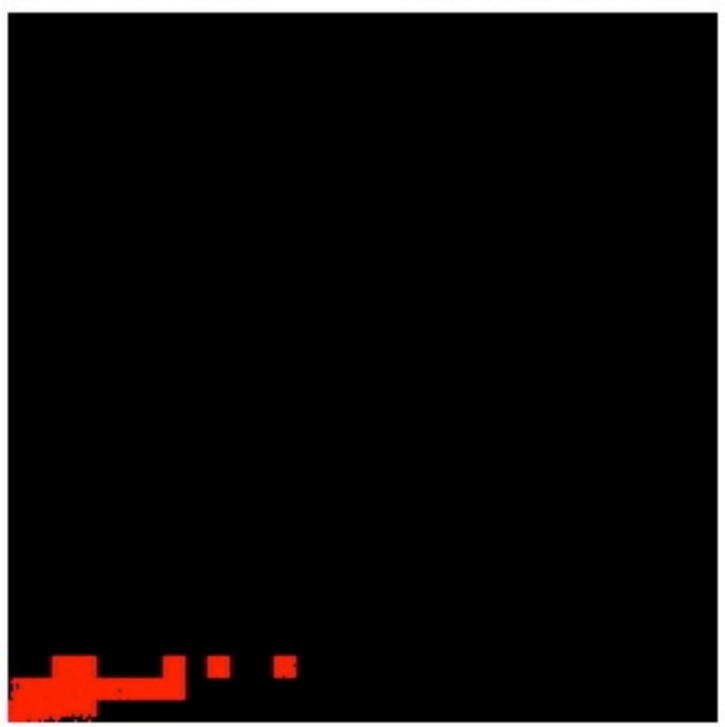
(b) uniform sampling 80%



(c) best uniform sampling 20%



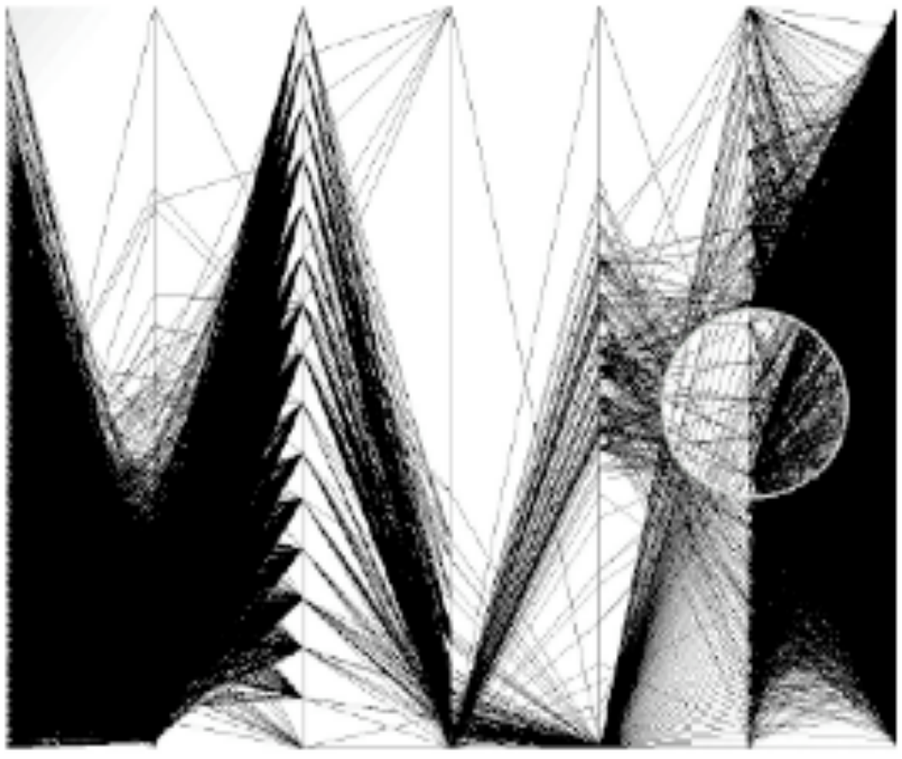
(d) non-uniform sampling

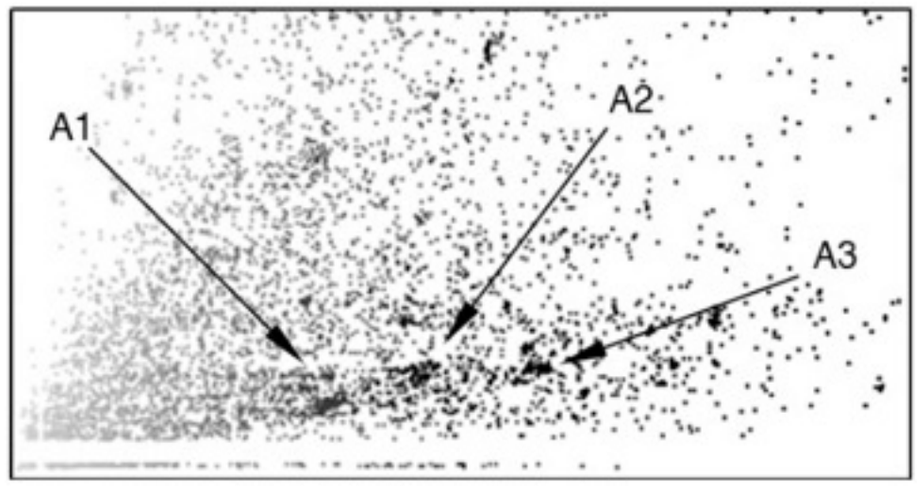


(e) most dense areas

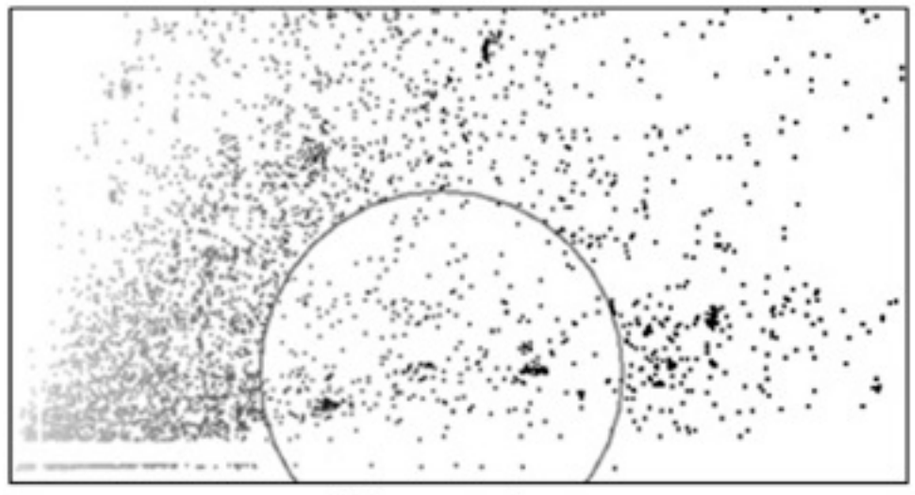
Sampling Lens

- Ellis et al. 2005
- Magic Lens approach to apply random sampling to user-defined regions while maintaining the original data density in the context





(a) without lens



(b) lens over clusters

Constant Information Density

- Woodruff et al. 1998
- Data objects in areas with high information density are represented by small-sized low-detail glyphs
- Objects in sparse-density areas are represented by larger, more detailed glyphs
- The number of information objects is kept constant as the user scales and moves the view

Constant Information Density

Figure 1a: DataSplash visualization.

Figure 1b: $VIDA_0$ visualization.

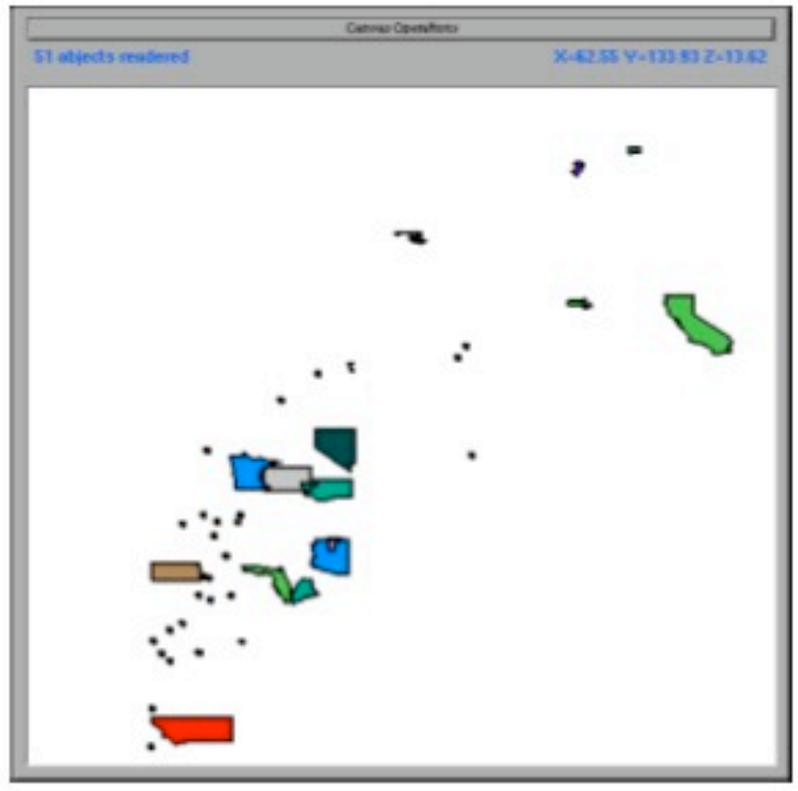


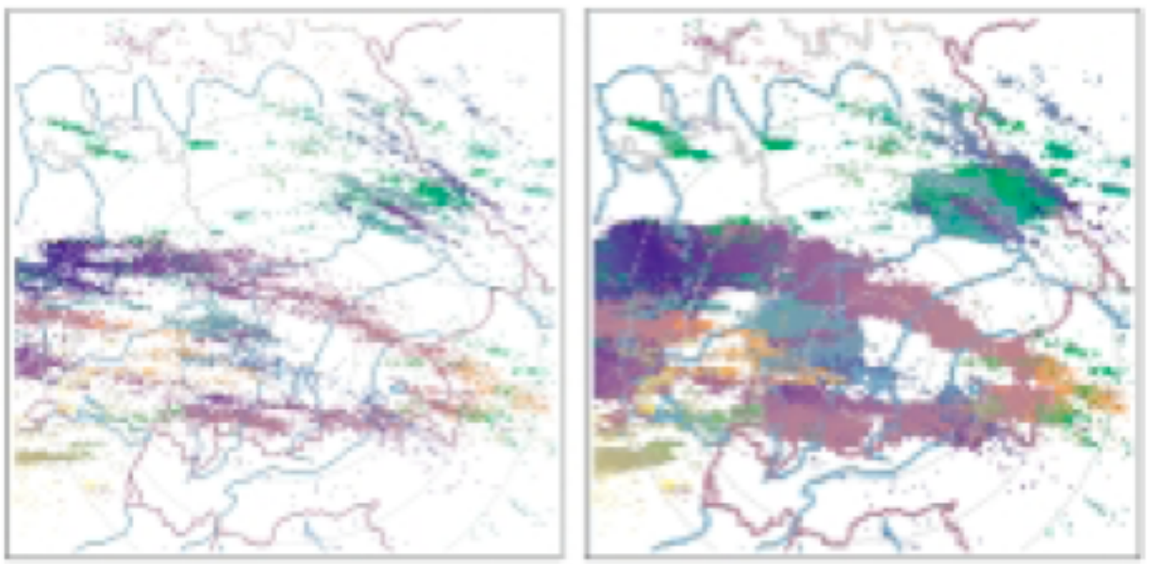
Figure 1c: VIDA visualization.

Figure 1: A visualization of states in the United States, displayed according to housing cost (*x* axis) and income (*y* axis). In Figure 1a, the visualization is obviously cluttered in many regions. In Figure 1b, objects are shown with a less detailed representation, as recommended by $VIDA_0$. Note that within Figures 1a and 1b, all objects are displayed with a single graphical representation. In Figure 1c, VIDA displays dots for objects that appear in dense regions; objects in less dense regions are displayed as polygonal outlines.

Point Displacement

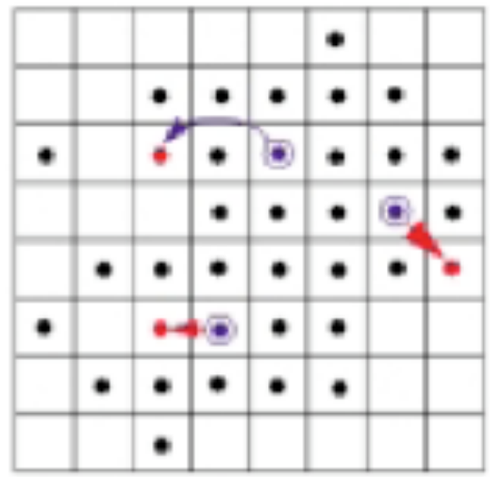
- Items that would overlap other items are moved to adjacent free positions
- Position of the items and their distance should be preserved as much as possible
- Three algorithms for displacing pixels (e.g. adding abstract data to geographical maps) (Keim & Hermann 1998)
  - Nearest-Neighbor
  - Curve-based
  - Gridfit
- Algorithms share the same procedure
  - All data points which have a unique position are placed on the display
  - New positions are determined for the remaining points

Keim 2000: data with and without overlap



Point Displacement

- Nearest-Neighbor algorithm
  - For all points which have not been set in the first step, place them on the nearest unoccupied position
  - Fast to compute, but limited effectiveness for very dense displays (pixels may be placed very far from their original positions)



- Curve-based algorithm
  - For all points which have not been set in the first step
    - Compute the nearest unoccupied position on a given screen-filling curve
    - Shift all points between the occupied pixel and the unoccupied position along the screen-filling curve
    - Place the point on the newly available unoccupied pixel