VISUALIZATION OF MULTIVARIATE DATA

Dr. Yan Liu
Department of Biomedical, Industrial and Human Factors Engineering
Wright State University

Introduction

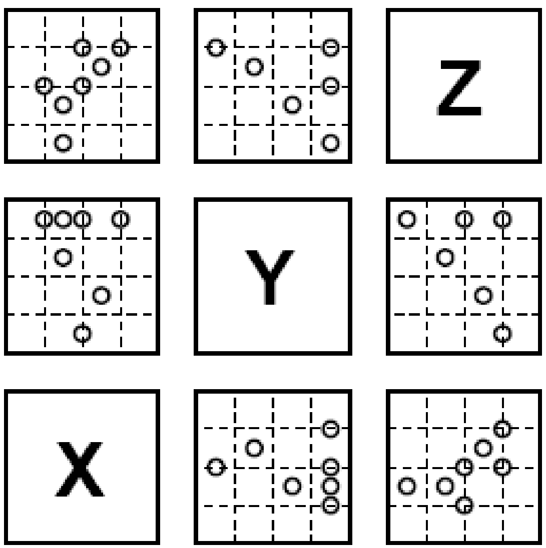
- Multivariate (Multidimensional) Visualization
  - Visualization of datasets that have more than three variables
  - "Curse of dimension" is a trouble issue in information visualization
    - Most familiar plots can accommodate up to three dimensions adequately
    - The effectiveness of retinal visual elements (e.g. color, shape, size) deteriorates when the number of variables increases
- Categories of Multivariate Visualization Techniques
  - Different approaches to categorizing multivariate visualization techniques
    - The goal of the visualization, the types of the variables, mappings of the variables, etc.
  - Categories used in Keim and Kriegel (1996)
    - Geometric projection techniques
    - Icon-based techniques
    - Pixel-oriented techniques
    - Hierarchical techniques
    - Hybrid techniques

Geometric Projection Techniques

- Basic Idea
  - Visualization of geometric transformations and projections of the data
- Examples
  - Scatterplot matrix
  - Hyperbox
  - Trellis display
  - Parallel coordinates

Scatterplot Matrix

- Organizes all the pairwise scatterplots in a matrix format
- Each display panel in the matrix is identified by its row and column coordinates
  - The panel at the ith row and jth column is a scatterplot of X_j versus X_i

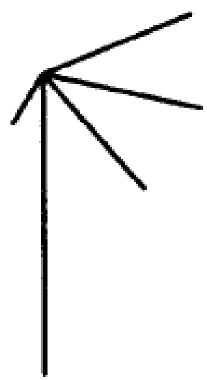


Scatterplot matrix with three variables X, Y, and Z

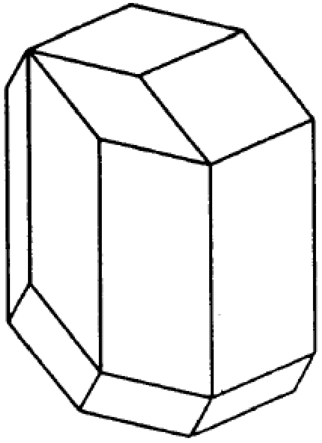
- The panel at the 3rd row (the top row) and 1st column is a scatterplot of Z versus X
- Panels that are symmetric with respect to the XYZ diagonal have the same variables as their coordinates, rotated 90°
  - The redundancy is designed to improve visual linking
  - Patterns can be detected in both horizontal and vertical directions
- Can only visualize the correlation between two variables, without using retinal visual elements or interaction techniques

Hyperbox (Alpern & Carter, 1991)

- Like the scatterplot matrix, it also involves pairwise 2D plots of variables
- A hyperbox is a 2D depiction of a *k*-D box
  - A very constrained picture, starting with *k* line segments radiating from a point which are contained within an angle less than 180°
  - The length of the line segments and the angles between them are arbitrary, although they should ideally follow the banking to 45° principle (a line segment with an orientation of 45° or -45° is the best to convey linear properties of the curve)



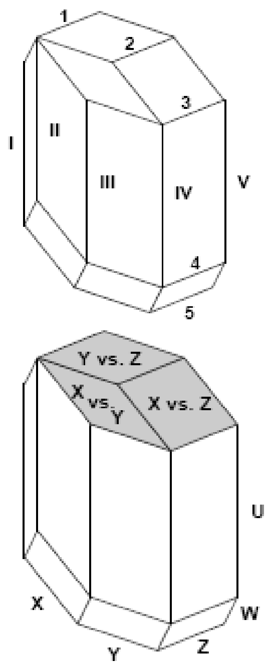
Five line segments of a hyperbox



A 5-D hyperbox

Hyperbox (Cont'd)

- Properties
  - Contains k^2 lines and $k(k-1)/2$ faces
    - e.g. there are $5^2=25$ lines and $5(5-1)/2=10$ faces in a 5-D hyperbox
  - For each line in a hyperbox, there are k-1 other lines with the same length and orientation; lines with the same length and orientation form a direction set



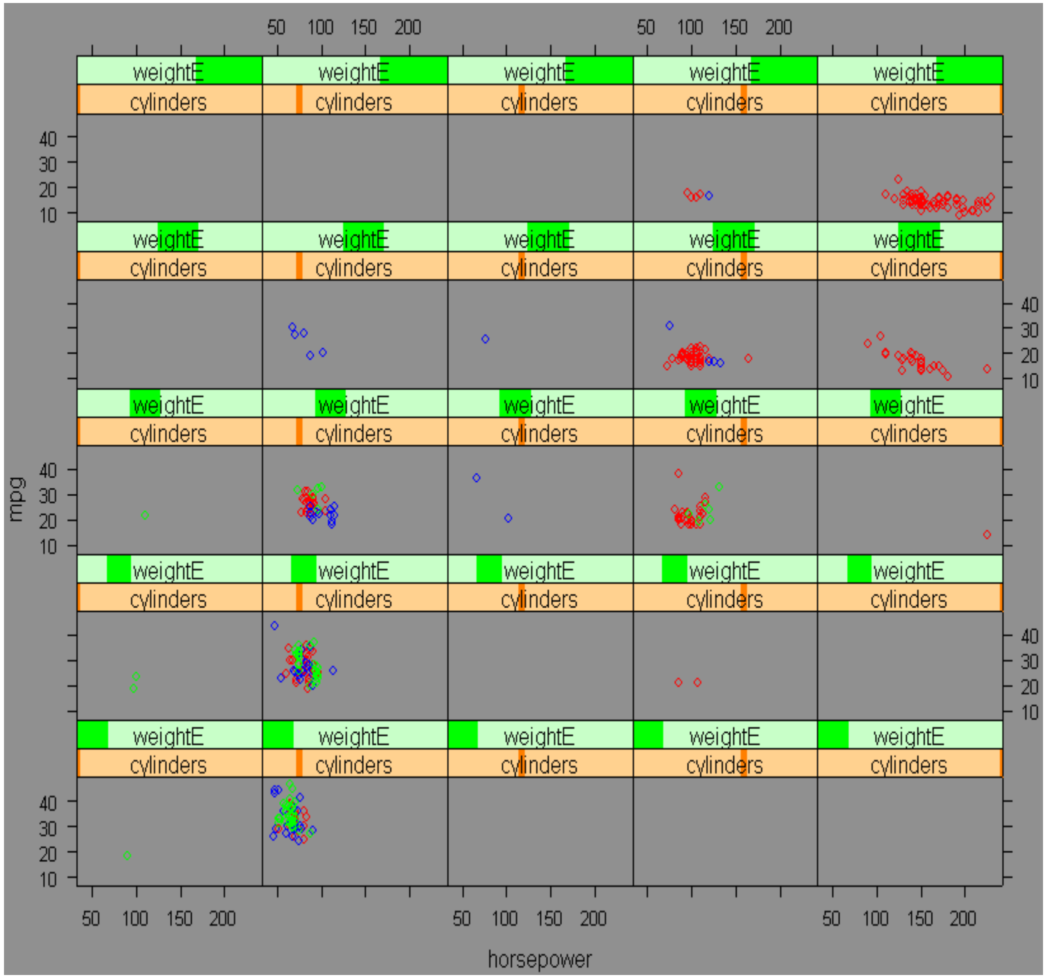
A 5-D hyperbox

• lines 1, 2, 3, 4, and 5 form a direction set
• lines I,II, III, IV, and V form a direction set

• Five variables X, Y, Z, W, and U are mapped to five direction sets
• Each face of the hyperbox can be used to display 2D plots (e.g. scatterplot, line chart)

Trellis Displays (Becker and Cleveland, 1996)

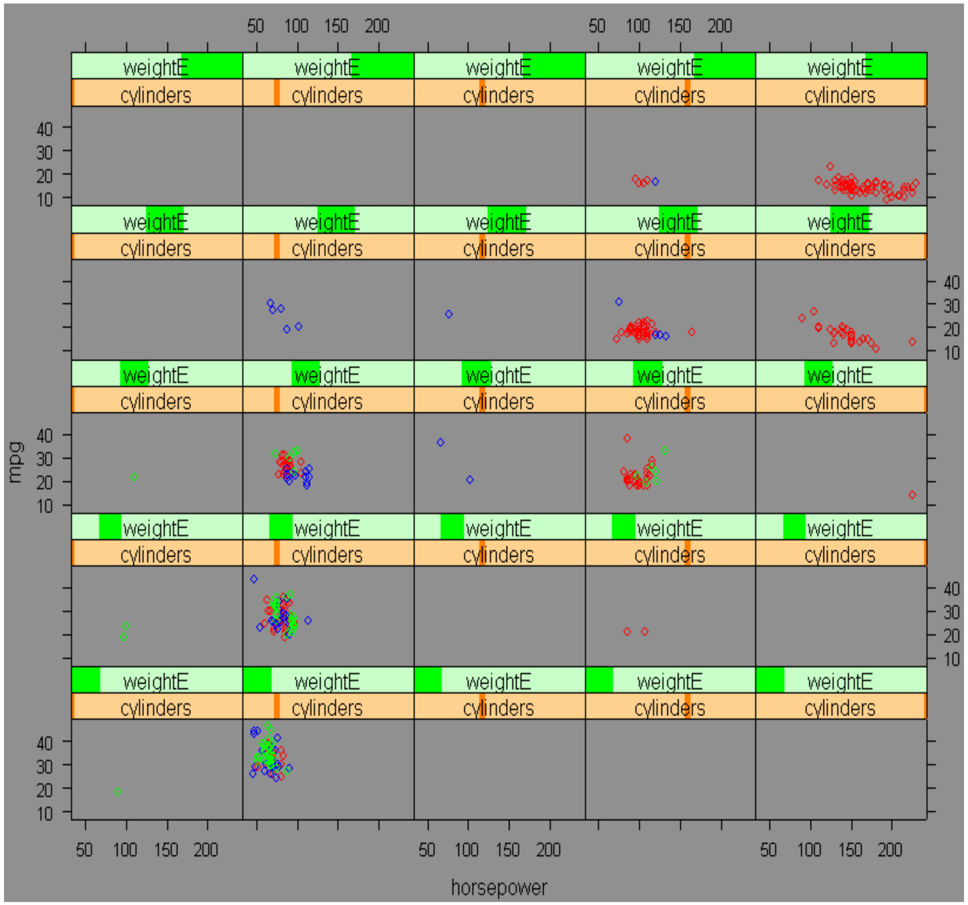
- Display any one of the large variety of 1D, 2D and 3D plot types in a trellis layout of panels, where each panel displays the selected plot type for a level or interval on additional discrete or continuous conditioning variables
- Panels are laid out into columns, rows and pages
- Mapping of Variables
  - Axis variable
    - Mapped to one of the coordinates in the panels
  - Conditioning variable
    - Mapped to a horizontal bar at the top of each panel, representing one of its levels (discrete variable) or intervals (continuous variable)
    - Continuous variables have to be divided into intervals
      - The intervals are usually overlapped a little to improve the effectiveness of visualizing interrelationships
  - Superposed variable
    - Mapped to color or symbol of points in the panels



Trellis Display of an Auto Dataset

● American ● European ● Japanese

- Five Variables
  - mpg (continuous)
  - cylinders (3/4/5/6/8)
  - horsepower (continuous)
  - weight (continuous)
  - origin (American/European/Japanese)
- Axis variables
  - horsepower and mpg
- Conditioning variables
  - weight and cylinders
- Superposed variable
  - origin



Trellis Display of an Auto Dataset

● American ● European ● Japanese

- Effective in demonstrating the relationships between axis variables, considering all the conditioning variables

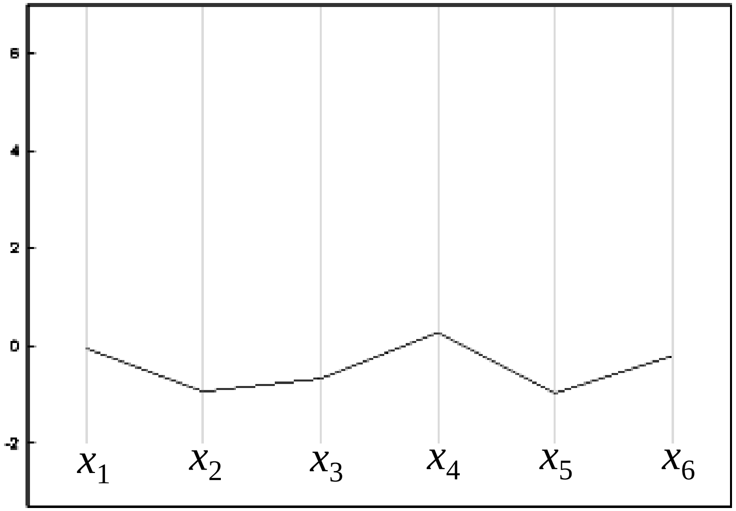
What patterns can you see?

- The generated visualization may be greatly affected by how the continuous conditioning variables are categorized

- Data overlapping occurs when many data records have the same or similar values or the number of data points is large relative to the size of a panel

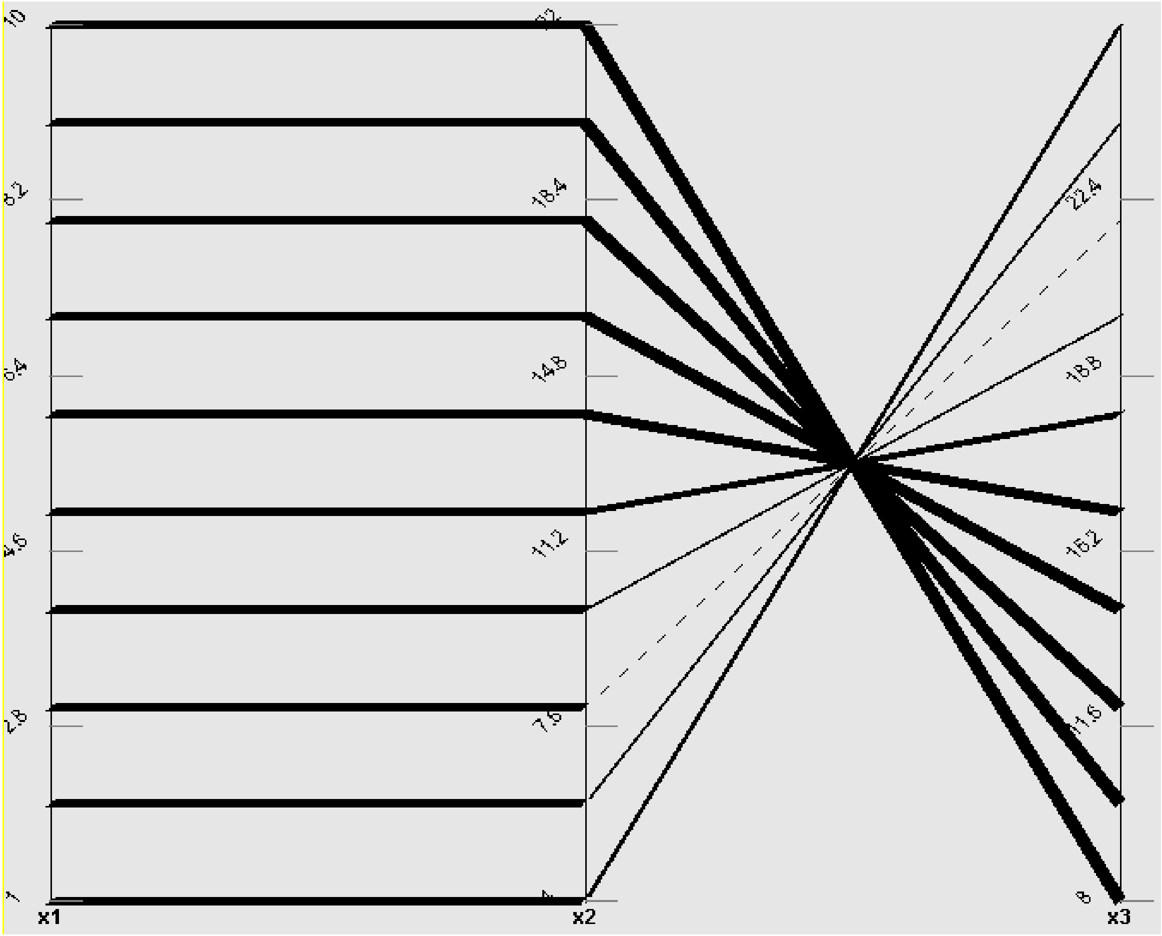
Parallel Coordinates (Inselberg, 1985)

- Each variable is represented by a vertical axis
- *k* variables are organized as *k* uniformly spaced vertical lines in a 2D space
- A data record with *k* variables is manifested as a connected set of *k* points, one on each axis
- Variables are usually normalized so that their maximum and minimum values correspond to the top and bottom points on their corresponding

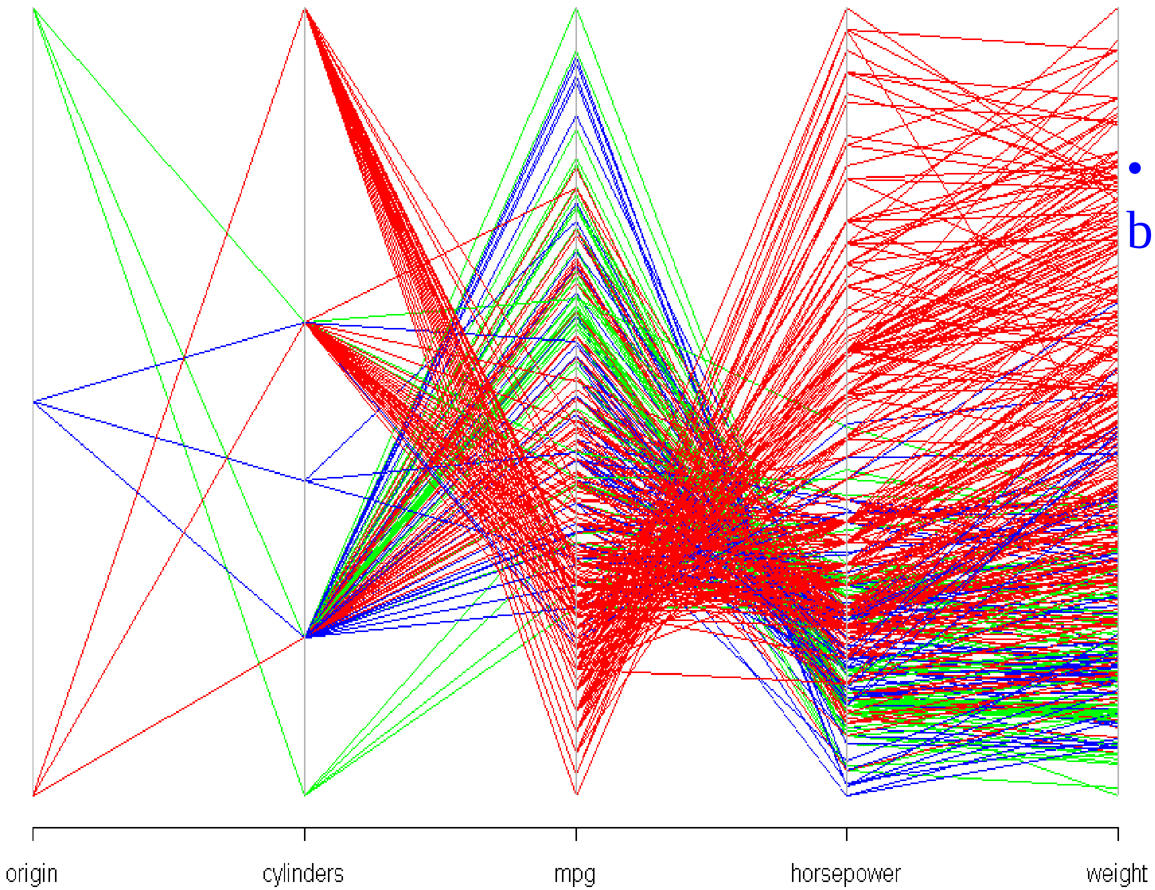


A parallel coordinate representation of a point with 6 variables

• The point represented in this figure is (0,-1,-0.75,0.25,-1, -0.25)



Perfect positive linear relationship between X1 and X2
Perfect negative linear relationship between X2 and X3



A parallel coordinate representation of the auto dataset

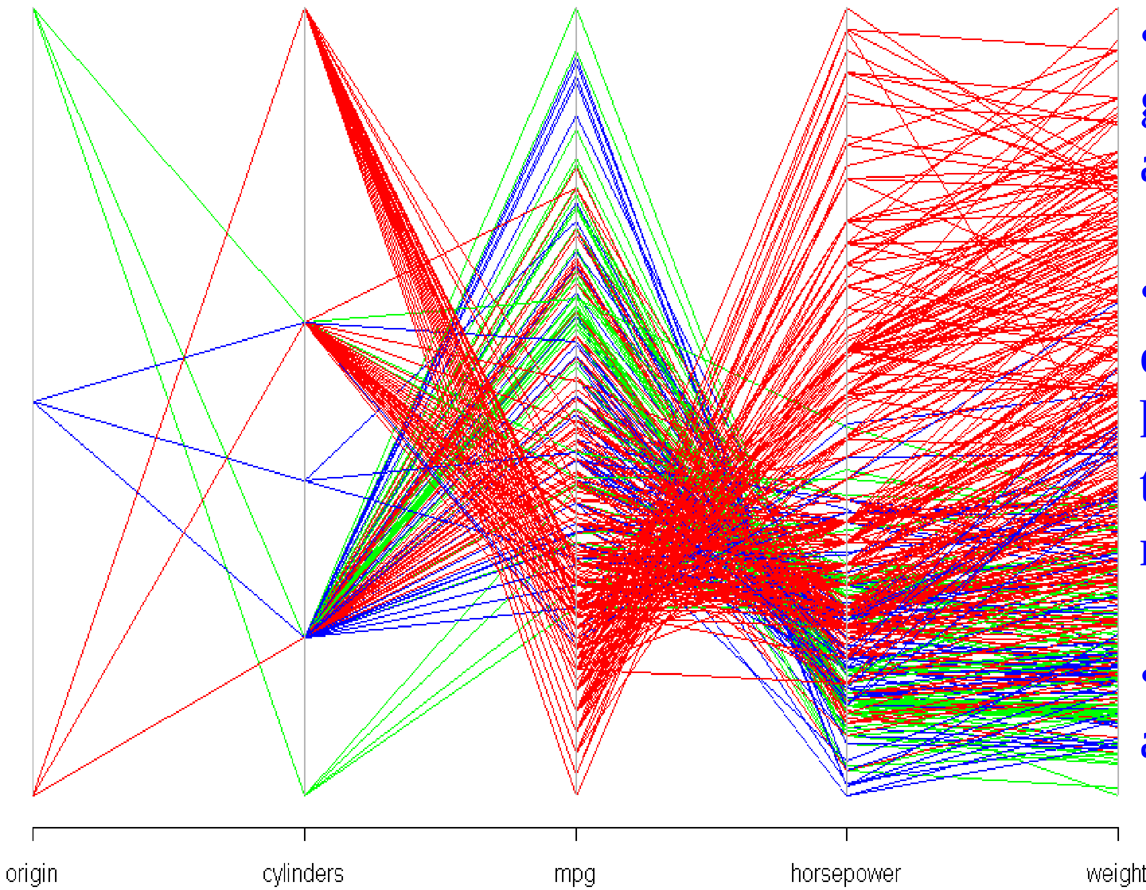
● American ● European ● Japanese

- Effective in revealing relationships between adjacent axis variables

  Relationship between mpg and horsepower, between horsepower and weight?

- Effective in showing the distributions of attributes

  Distribution of cylinders , mpg, horsepower, and weight in US cars?



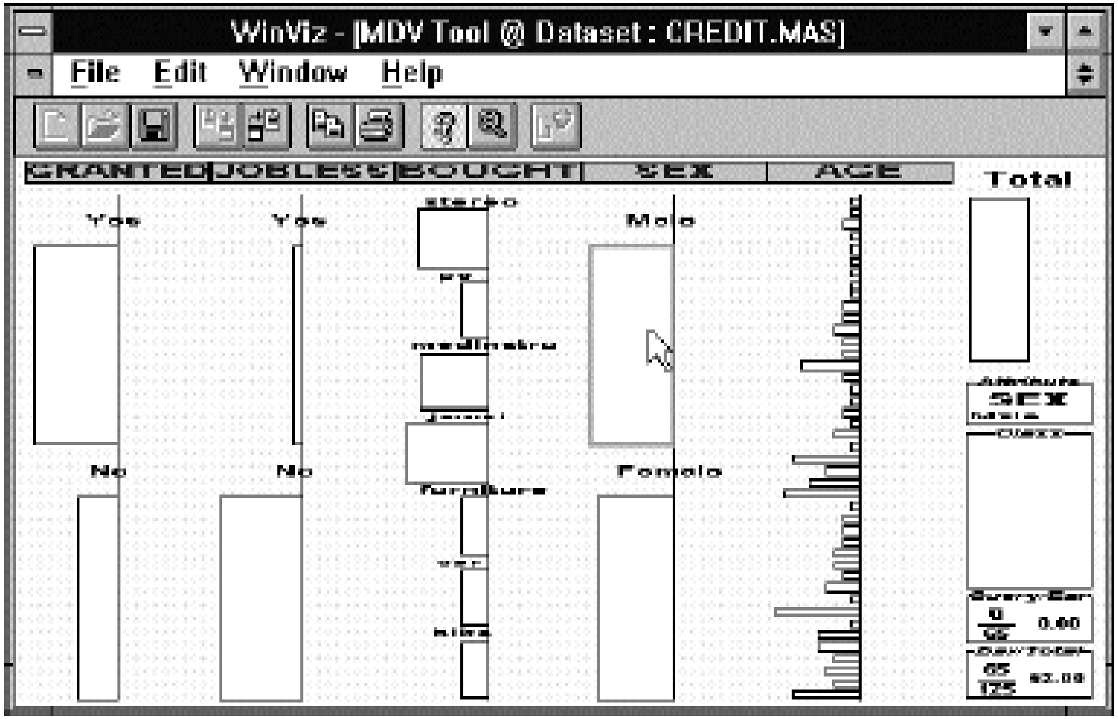
A parallel coordinate representation of the auto dataset

● American ● European ● Japanese

- Effectiveness of visualization is greatly impacted by the order of axes
- Overlapping of line segments occurs when many data records have the same or similar values or the number of data records is large relative to the display
- Interaction techniques are often applied to address the problems
  - changing the order of the axes, selecting a subset of data for visualization

Parallel Coordinates (Cont'd)

- Applications
  - visualize discrete variables, present classification rules, etc.



Parallel coordinate representation of a credit screening dataset (Lee et al., 1995)

- Variables
  - Application Granted (Yes/No)
  - Jobless (Yes/No)
  - Items Bought (Stereo/PC/Bike/Instrument/ Jewel/Furniture/Car)
  - Sex (Male/Female)
  - Age (categorized into intervals)

- Width of a bar indicates the No. of records in its corresponding category; height of the bar has no significance

Summary of Geometric Projection

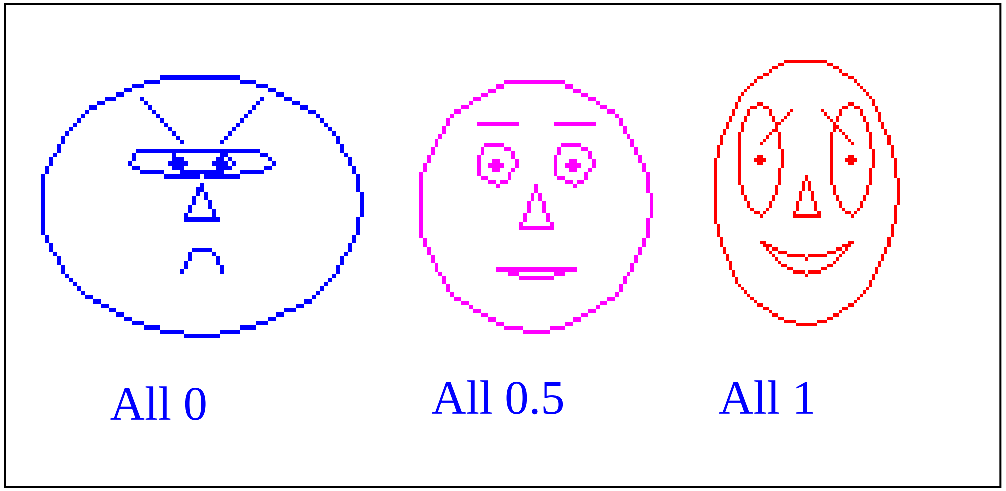
- Can handle large and very large datasets when coupled with appropriate interaction techniques, but visual cluttering and record overlap are severe for large datasets
- Can reasonably handle medium- and high- dimensional datasets
- All data variables are treated equally; however, the order in which axes are displayed can affect what can be perceived
- Effective for detecting outliers and correlation among different variables

Icon-Based Techniques

- Basic Idea
  - Visualization of data values as features of icons
- Examples
  - Chernoff faces
  - Stick figures
  - Star plots
  - Color icons

Chernoff Faces (Chernoff, 1973)

- Named after their inventor Herman Chernoff (1973)
- A simplified image of a human face is used as a display
- Data variables (attributes) are mapped to different facial features



Chernoff faces with 10 facial characteristic parameters:
1. head eccentricity, 2. eye eccentricity, 3. pupil size, 4. eyebrow slant, 5. nose size, 6. mouth shape, 7. eye spacing, 8. eye size, 9. mouth length, and 10. degree of mouth opening

Stick Figures (Pickett & Grinstein, 1988)

- Two most important variables are mapped to the two display dimensions
- Other variables are mapped to angles and/or length of limbs of the stick figures
- Stick figure icons with different variable mappings can be used to visualize the same dataset

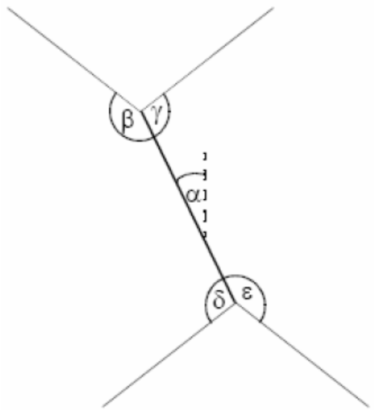
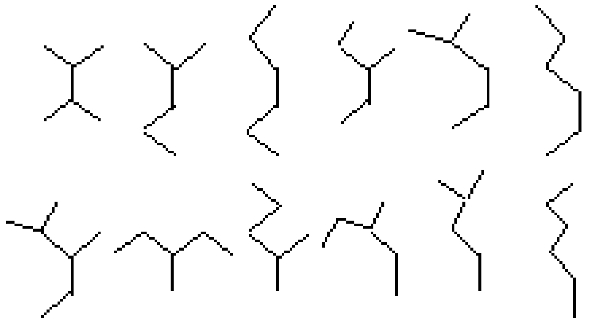


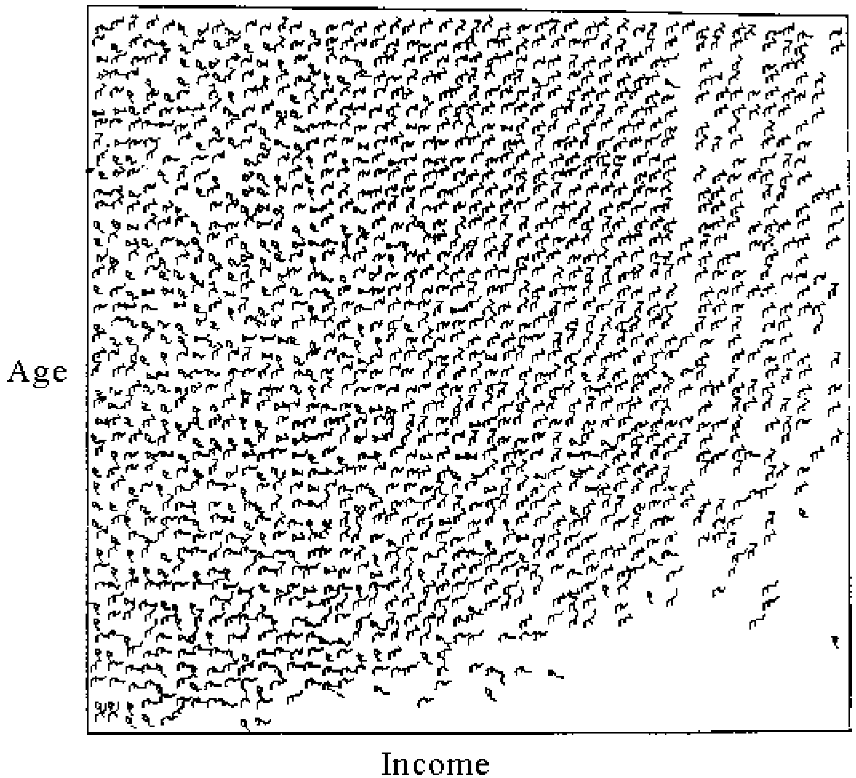
Illustration of a stick figure
(5 angles and 5 limbs)



A family of 12 stick figures that
have 10 features

Stick Figures (Cont'd)

- If the data records are relatively dense with respect to the display, the resulting visualization presents texture patterns that vary according to the characteristics of the data and are therefore detectable by preattentive perception

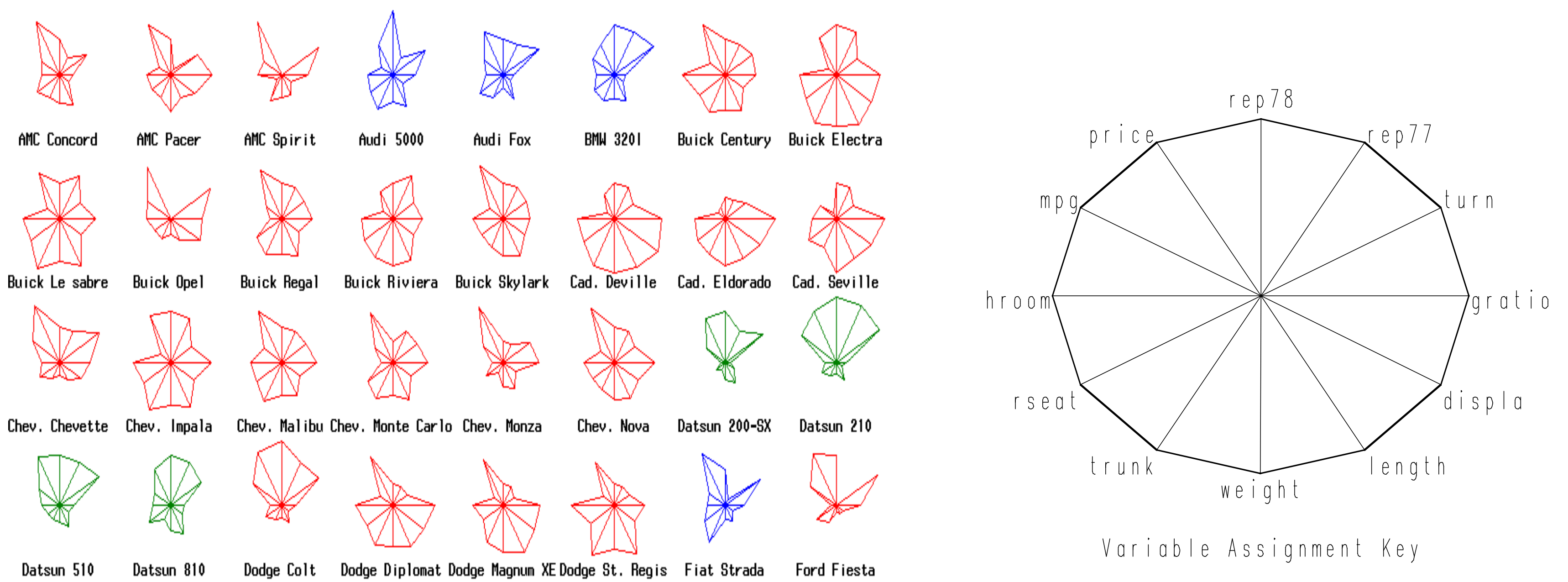


Stick figures of 1980 US census data

• Age and income are mapped to display dimensions
• Occupation, education levels, marital status, and gender are mapped to stick figure features
• A clear shift in texture over the screen, which indicates the functional dependencies of the other attributes on income and age

Star Plots (Chambers et al.,1983)

- Each data record is represented as a star-shaped figure with one ray for each variable
- The length of each ray is proportional to the value of its corresponding variable
- Each variable is usually normalized to between a very small number (close to 0) and 1
- The open ends of the rays are usually connected with lines



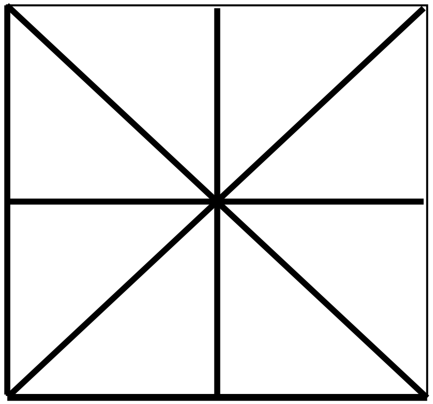
Star plots representation of an auto dataset with 12 variables

Star Plots (Cont'd)

- Issues
  - As the number of rays increases, it becomes more difficult to separate them
  - They should be separated at least 30° from each other to be distinguishable
  - The number of distinguishable arrays may be increased by adding retinal visual properties
    - e.g. hue, luminance, width, etc.

Color Icons (Levkowitz, 1991)

- An area on the display to which color, shape, size, orientation, boundaries, and area subdividers can be mapped by multivariate data



A square icon

- Linear mapping
  - Up to 6 variables can be mapped to the icon, shown as the thick lines
    - 2 of edges (one horizontal, one vertical)
    - 2 diagonals
    - 2 midlines
  - A color is assigned to each thick line according to the value of the corresponding variable
- Area mapping
  - Each subarea (totally 8 subareas) corresponds to one variable
  - A color is assigned to a subarea according to the value of its corresponding variable

Color Icons (Cont'd)

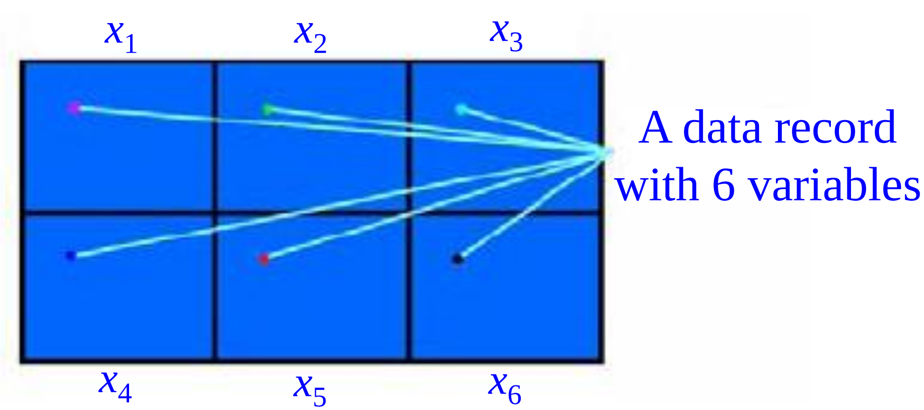
- The number of variables mapped to the color icon can be tripled by having each variable control one of the hue, saturation, and value (HSV) values
- More than one variable can be mapped to a linear feature by subdividing its length
  - Subdivision can be fixed globally (e.g. all linear features are subdivided in the middle)
  - Subdivision can be data-controlled, where the point of subdivision is controlled by the value of a variable
- Icons with different shapes can be used in place of the square icon
  - e.g. Triangular, hexagon

Summary of Icon-Based Techniques

- Can handle small to medium datasets with a few thousand data records, as icons tend to use a screen space of several pixels
- Can be applied to datasets of high dimensionality, but interpretation is not straightforward and requires training
- Variables are treated differently, as some visual features of the icons may attract more attention than others
  - The way data variables are mapped to icon features greatly determines the expressiveness of the resulting visualization and what can be perceived
- Defining a suitable mapping may be difficult and poses a bottleneck, particularly for higher dimensional data
- Data record overlapping can occur if some variables are mapped to the display positions

Pixel-Based Techniques

- Basic Idea (Keim, 2000)
  - Each variable is represented as a subwindow in the display which is filled with colored pixels
  - A data record with *k* variables is represented as *k* colored pixels, each in one subwindow associated with a variable
    - The color of a pixel demonstrates its corresponding value
  - The color mapping of the pixels, arrangement of pixels in the subwindows and shape of the subwindows depend on the data characteristics and visualization tasks

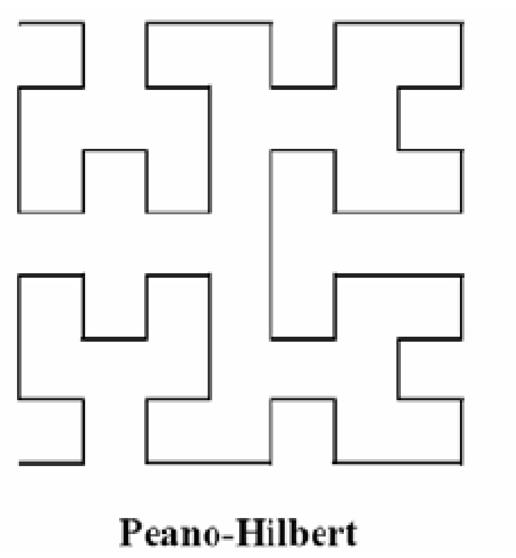


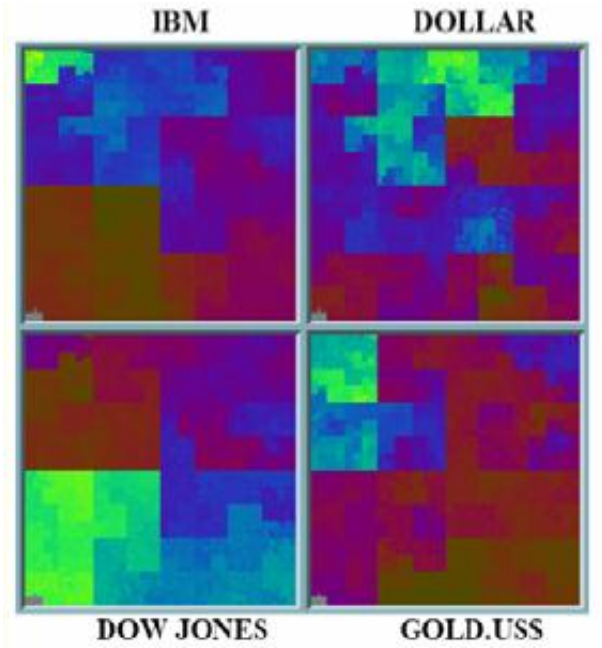
Pixel-Based Techniques (Cont’d)

- Types
  - Query-independent techniques: visualize the entire dataset
    - Space-filling curves
    - Recursive pattern technique
  - Query-dependent techniques: visualize a subset of data that are relevant to the context of a specific user query
    - Spiral technique
    - Circle segment
- Color Mapping
  - A HSI (hue, saturation, intensity) color model is used
  - A color map with colors ranging from yellow over green, blue, and red to almost black

Space Filling Curves

Assume there are n data records $\{R_1, R_2, \ldots, R_n\}$, each consisting of k data values $\{R_i^1, R_i^2, \ldots, R_i^k\}$ (i=1,2,…,k). In one of the subwindows, say m, we want to present all the values $\{R_1^m, R_2^m, \ldots, R_n^m\}$ in the window in such a way that the arrangement of pixels will best preserve the distance of between data values. This is an optimization problem, and the so called space-filling curves have been used to solve this problem. It is well-known that the Peano-Hilbert curve can provide the best optimization.

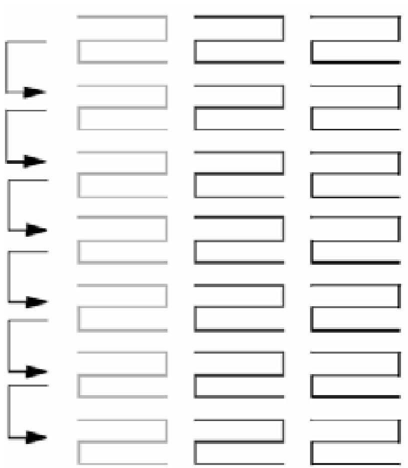




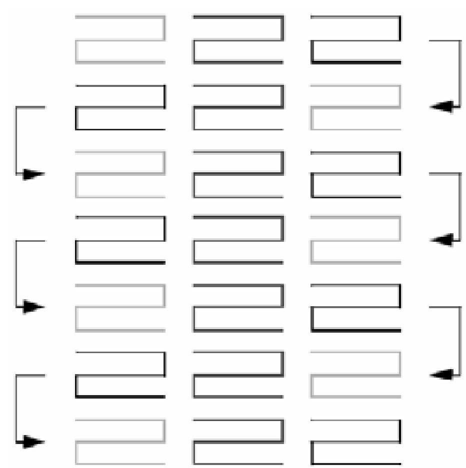
The pixel-based visualization of a financial dataset using Peano-Hilbert arrangement

Recursive Pattern Technique

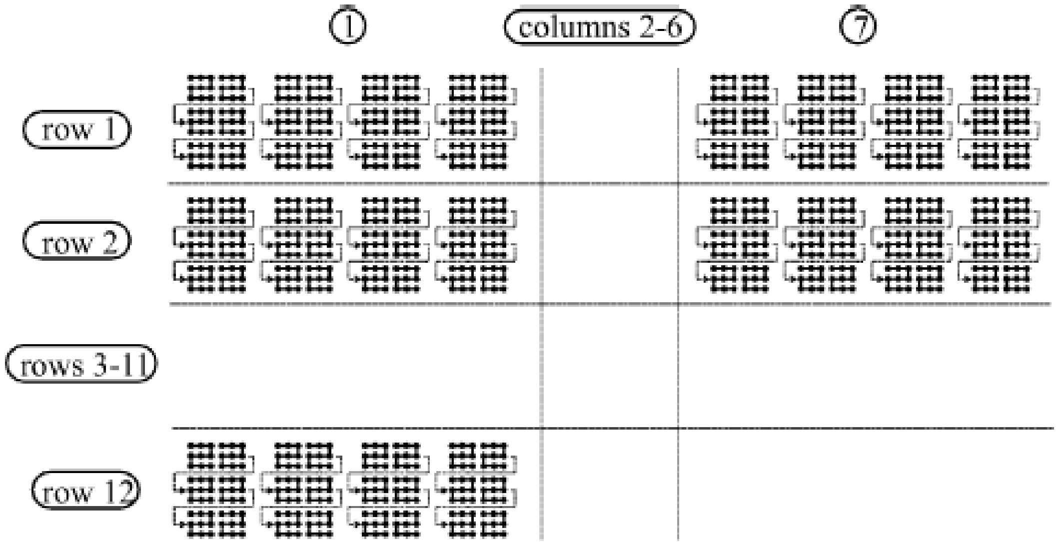
- Based on a general recursive scheme which allows lower-level patterns to be used as building blocks for higher-level patterns
  - e.g. For a time-series dataset which measures some parameters several times a day over a period of several months, it would be natural to group all data records belonging to the same day in the first-level pattern, those belonging to the same week in the second-level pattern, and those belonging to the same month in the third-level pattern



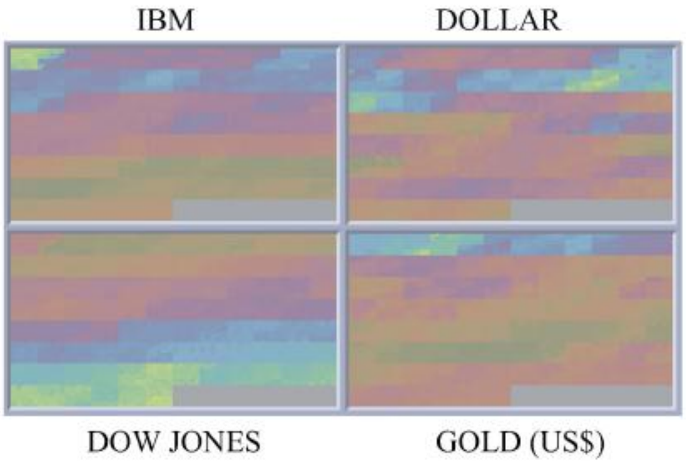
Line-by-line loop



Back-and-forth loop



Schematic representation of a 5-level recursive pattern arrangement



5-level recursive pixel-based visualization of a financial dataset

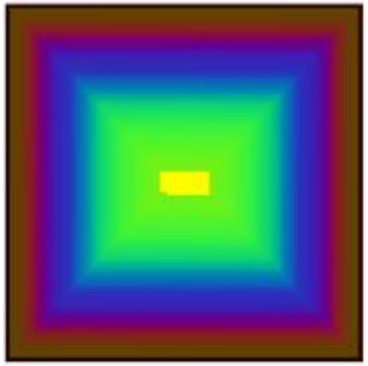
- First level: 3x3 pixels
- Second level: 3x2 level-1 groups
- Third level: 1x4 level-2 groups
- Fourth level: 12x1 level-3 groups
- Fifth level: 1x7 level-4 groups

Query-Dependent Techniques

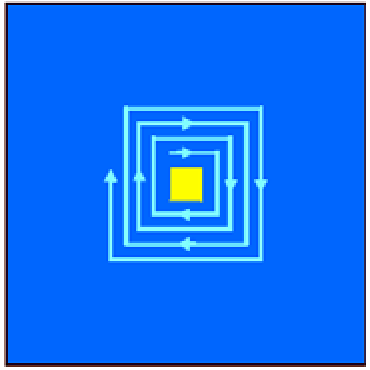
- Overview
  - k variables $(x_1, x_2, \ldots, x_k)$
  - Data records $(R_1, R_2, \ldots, R_n)$
    (i=1,2,…,n),each consisting of k data values $(R_i^1, R_i^2, \ldots, R_i^k)$
  - Query $(q_1, q_2, \ldots, q_k)$
    - e.g. q_1: x_1=5, q_2: x_2=3, …., q_k : x_k=7
  - Distance
    - For each data record, R_i, (i=1,2,…,n), its distance from the query is $(d_i^1, d_i^2, \ldots, d_i^k)$,where $d_i^j = R_i^j - q_j$, (j=1,2,..,k)
  - Overall distance
    - For each data record, R_i, (i=1,2,…,n), its overall distance is the weighted average of its individual distances
  - Sort the data records according to their overall distance, and only the $m/(n-k)$ quantile (m is the # of pixels in the display) of the most relevant data records are presented to the user

Spiral Technique

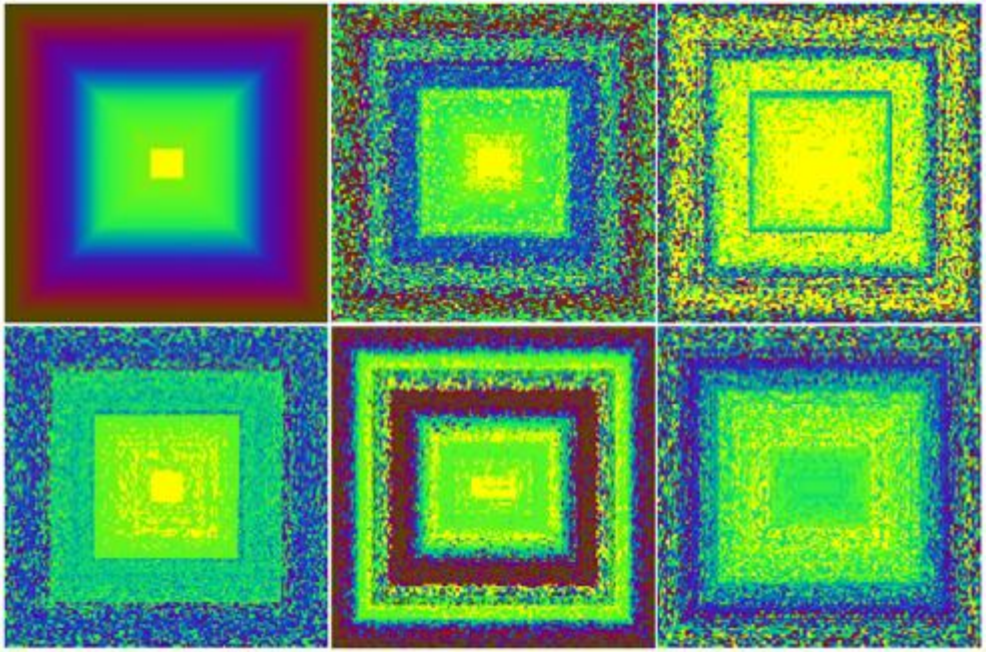
- Each variable is represented by a square window
- An additional window is used to represent the overall distances of all the presented data records
- The data records that have the smallest overall distances are placed at the center of the window, and the data records are arranged in a rectangular spiral-shape to the outside of the window



Window that shows the overall distance



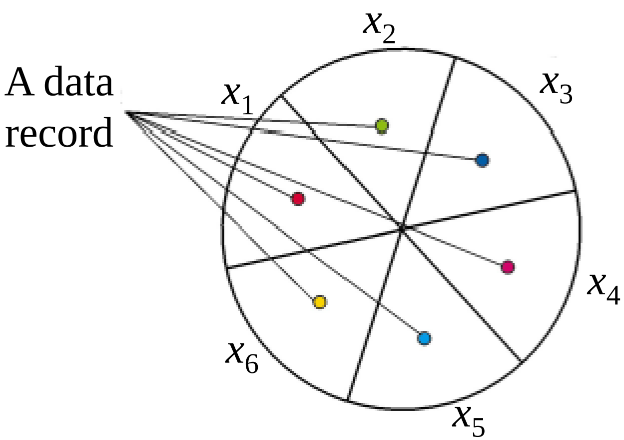
Spiral arrangement of pixels



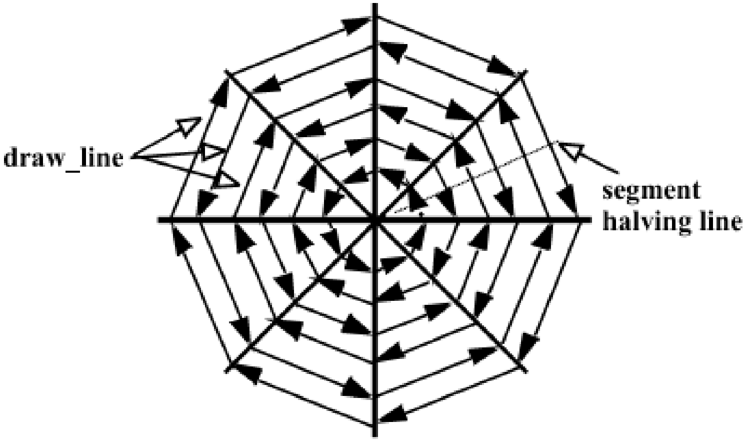
Spiral pixel-based visualization of a dataset with five variables

Circle Segments

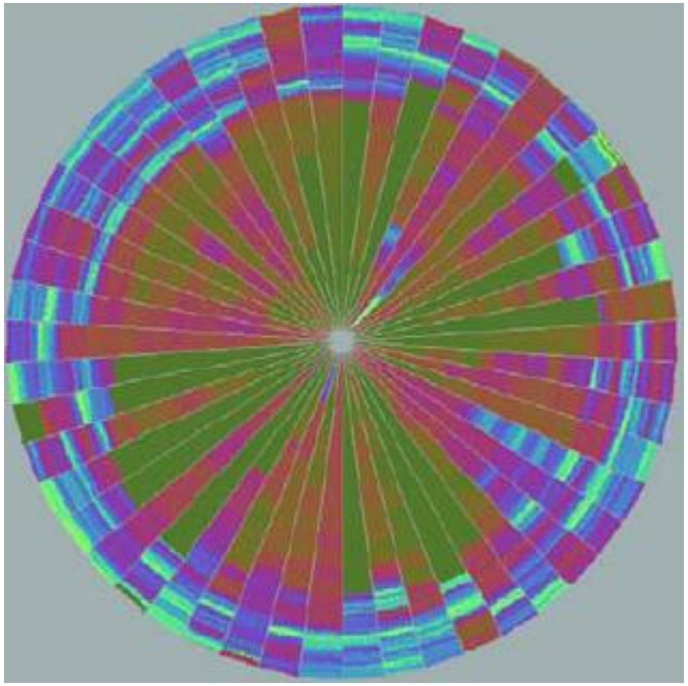
- Display the variables as segments of a circle
- If the dataset consists of k variables, the circle is partitioned into k segments, each representing one variable
  - The data records within each segment are arranged in a back-and-forth manner along the so called "draw_line" which is orthogonal to the line that halves the two border lines of the segment. The "draw_line" starts from the center of circle and moves to the outside of the circle



Circle segment representation of a dataset with 6 variables



Circle segment pixel arrangement for a dataset with 8 variables



Circle segment representation of
a dataset with 50 variables

Summary of Pixel-Based Techniques

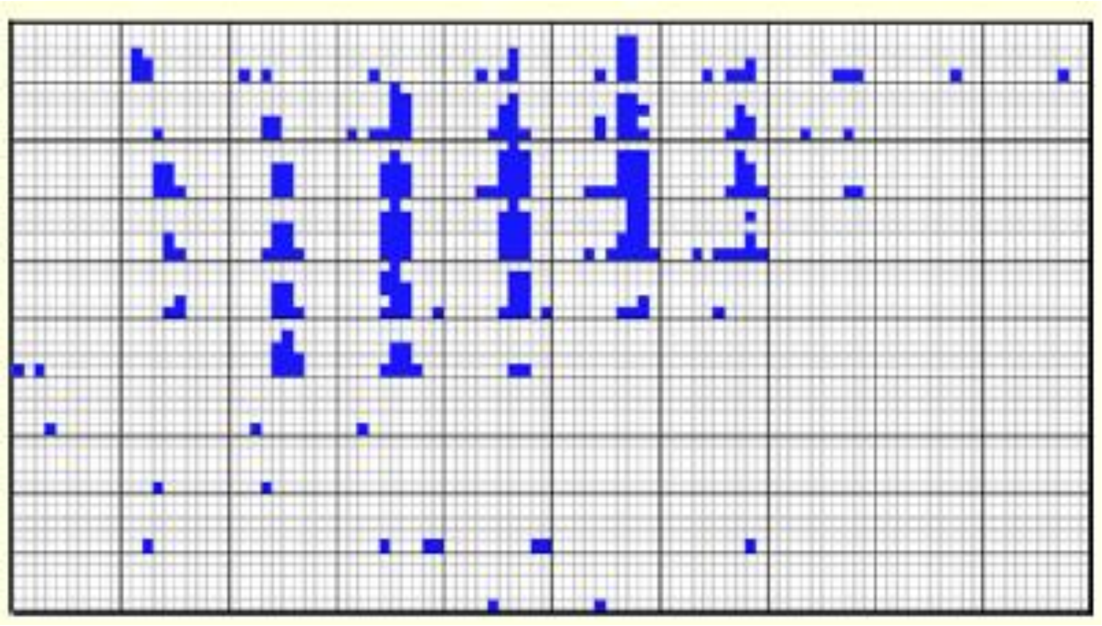
- Can handle large and very large datasets on high-resolution displays
- Can reasonably handle medium- and high- dimensional datasets
- As each data record is uniquely mapped to a pixel, data record overlapping and visual cluttering do not occur
- Limited in revealing quantitative relationships between variables because color is not effective in visualizing quantitative values,

Hierarchical Techniques

- Basic Idea
  - Subdivide the *k*-D data space and present subspaces in a hierarchical fashion
- Examples
  - Dimensional stacking
  - Mosaic Plot
  - Worlds-within-worlds (see lecture 1)
  - Treemap (see lecture 1)
  - Cone Trees (Later)

Dimensional Stacking (Leblanc et al., 1990)

- Partition the *k*-D data space in 2-D subspaces which are stacked into each other
- Adequate especially for data with ordinal attributes of low cardinality (the number of possible values)
- Procedures
  - Choose the most important pair of variables x_i and x_j, and define a 2D grid of x_i versus x_j
  - Recursive subdivision of each grid cell using the next important pair of parameters
  - Color coding the final grid cells
    - Using the value of a dependent variable, if applicable
    - Using the frequency of data in each grid cell

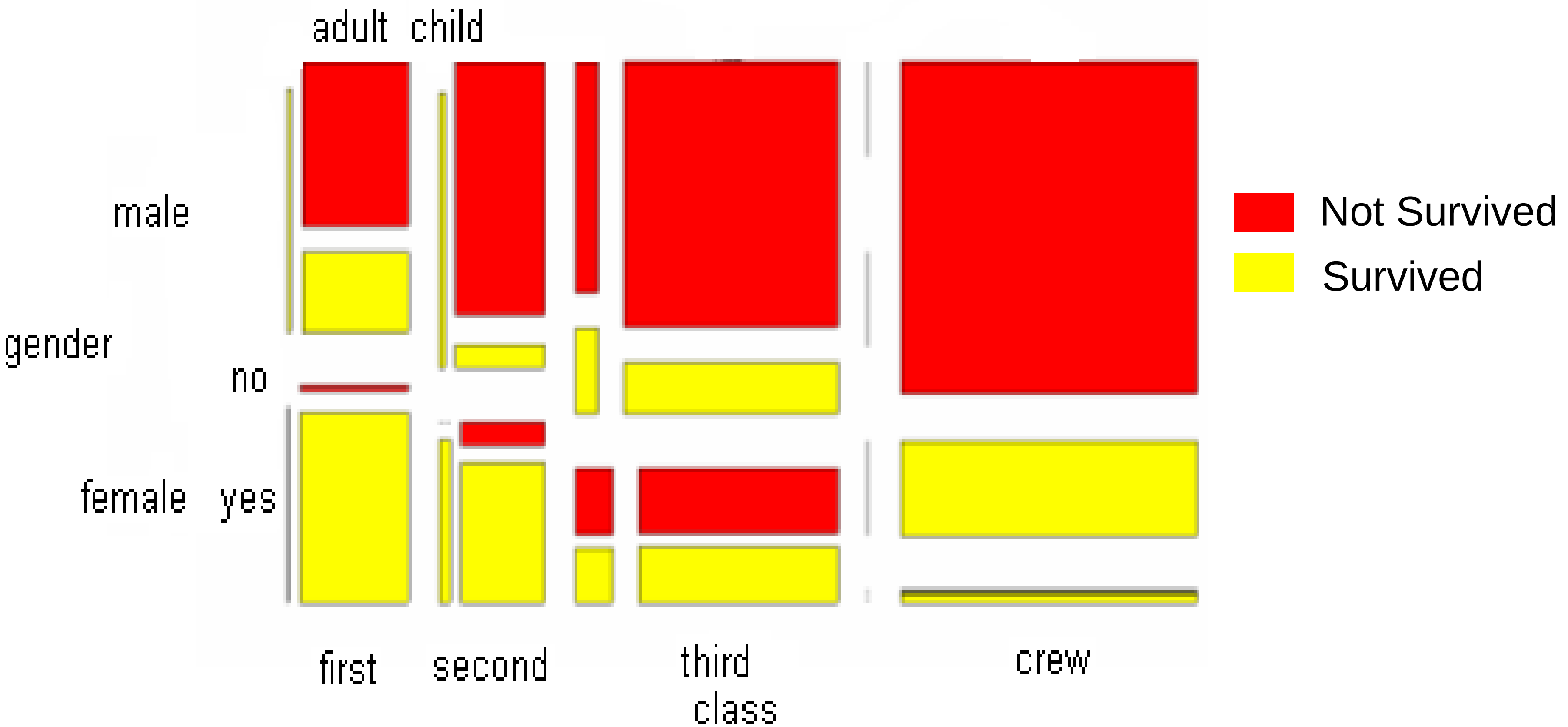


Dimensional stacking of an oil mining dataset

• Variables *longitude* and *latitude* are mapped to the horizontal and vertical axes of the outer grid
• Variables *ore grade* and *depth* are mapped to the horizontal and vertical axes of the inner grid

Mosaic Plot (Friendly, 1994)

- A well-recognized visualization method for categorical variables
- Shows frequencies in an *m*-way contingency table by nested rectangles
  - The area of a rectangle is proportional to its frequency (data counts)
- Procedures
  - First, divide a square in proportion to the marginal totals of variable X_1 along the horizontal axis
  - Next, the rectangle for each category of X_1 is subdivided in proportion to the conditional frequencies of variable X_2 along the vertical axis
  - Then, the rectangle for each combination of categories of X_1 and X_2 is subdivided in proportion to the conditional frequencies of X_2 along the horizontal axis
  - Repeat subdivisions until all variables of interest have been included in the plot



Mosaic Display of the Titanic Survival Dataset

Summary of Hierarchical Techniques

- Can handle small- to medium- sized datasets
- More suitable for handling datasets of low- to medium- dimensionality
- Variables are treated differently, with different mappings producing different views of data
- Interpretation of resulting plots requires training

Hybrid Techniques

- Integrate multiple visualization techniques, either in one or multiple windows, to enhance the expressiveness of visualization
  - Linking and brushing are powerful tools to integrate visualization windows (more in the next lecture)

References

- Alpern, B., & Carter, L. (1991). Hyperbox. *Proc. Visualization '91*, San Diego, CA, 133-139.
- Becker, R. A., & Cleveland, W. S., Shyu M.-J. (1996). The Visual Design and Control of Trellis Display, *Journal of Computational and Graphical Statistics*, 5(2), 123-155.
- Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983), *Graphical Methods for Data Analysis*, Wadsworth.
- de Oliveira, M., & Levkowitz, H. (2003). *IEEE Transactions on Visual and Computer Graphics*, 9(3), 378-394.
- Friendly, M. (2001). *Visualizing Categorical Data*. NC : SAS Institute.
- Inselberg, A. (1985). The Plane with Parallel Coordinates, Special Issue on Computational Geometry. *The Visual Computer*, 1, 69-97.
- Keim, D.A., & Kriegel, H-P. (1996) Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 923-936.
- Lee, H-Y, Ong, H-L, Toh, E-W, & Chan, S-K (1995). Exploiting visualization in knowledge discovery. *Proc. 19th International Computer Software and Applications Conference*, Washington D.C., 26-31.
- LeBlanc, J., Ward, M. O., & Wittels, N. (1990). Exploring n-dimensional databases. *Proc. Visualization '90*, San Francisco, CA, 230-239.

References

- Levkowitz, H. (1991). Color icons: merging color and texture perception for integrated visualization of multiple parameters. *Proc. Visualization '91*, San Diego, CA, 164-170.
- Pickett R. M., & Grinstein G. G. (1988). Iconographic Displays for Visualizing Multidimensional Data. *Proc. IEEE Conf. on Systems, Man and Cybernetics*, Piscataway, NJ, 514-519.
- Wong, P.C., & Bergeron, R. (1997). 30 Years of Multidimensional Multivariate Visualization. In G.M. Nielson, H. Hagan, and H. Muller (Eds), *Scientific Visualization - Overviews, Methodologies and Techniques* (pp.3-33) CA: IEEE Computer Society Press.