Multidimensional Data Visualization

Direct Visualization Methods

Strategies for Multidimensional Data Visualization

- Direct visualization methods, where each feature, characterizing a multidimensional object, is represented in a visual form;
- Projection, so-called dimensionality reduction, methods, allowing us to represent the multidimensional data on a low-dimensional space. Artificial neural networks may also be used for visualizing multidimensional data. They realize various nonlinear projections.
- Our target is not to enumerate all methods and describe them in detail, but to present the most typical approaches and representatives of each group.

Direct Visualization Methods

- There is no formal mathematical criterion to estimate the visualization quality in direct visualization methods.
- All the features that characterize multidimensional data are represented in a visual form acceptable to a human.
- These methods may be classified into geometric, iconographic, and hierarchical visualization techniques.

Direct Visualization Methods

1. Geometric methods:
   a) scatter plots,
   b) matrix of scatter plots,
   c) multiline graphs,
   d) permutation matrix,
   e) survey plots,
   f) Andrews curves,
   g) parallel coordinates,
   h) radial visualization (RadViz) and its modifications GridViz and PolyViz.
2. Iconographic displays:
   a) Chernoff faces,
   b) star glyphs,
   c) stick figure,
   d) color icon.
3. Hierarchical displays:
   a) dimensional stacking,
   b) trellis display,
   c) hierarchical parallel coordinates.

Projection Methods

- Methods that allow us to represent multidimensional data from $\mathbb{R}^n$ in a low-dimensional space $\mathbb{R}^d$, $d < n$, are called projection (dimensionality reduction) methods.
- If the dimensionality of the projection space is small enough ($d = 2$ or $d = 3$), these methods may be used to visualize the multidimensional data. In such a case, the projection space can be called a *display*, *embedding* or *image space*.
- The projection methods usually invoke formal mathematical criteria by which the projection distortion is minimized.

Projection Methods

1. Linear projection methods:
   a) principal component analysis,
   b) linear discriminant analysis,
   c) projection pursuit.
2. Nonlinear projection methods:
   a) multidimensional scaling,
   b) locally linear embedding,
   c) isometric feature mapping,
   d) principal curves.

Direct Visualization

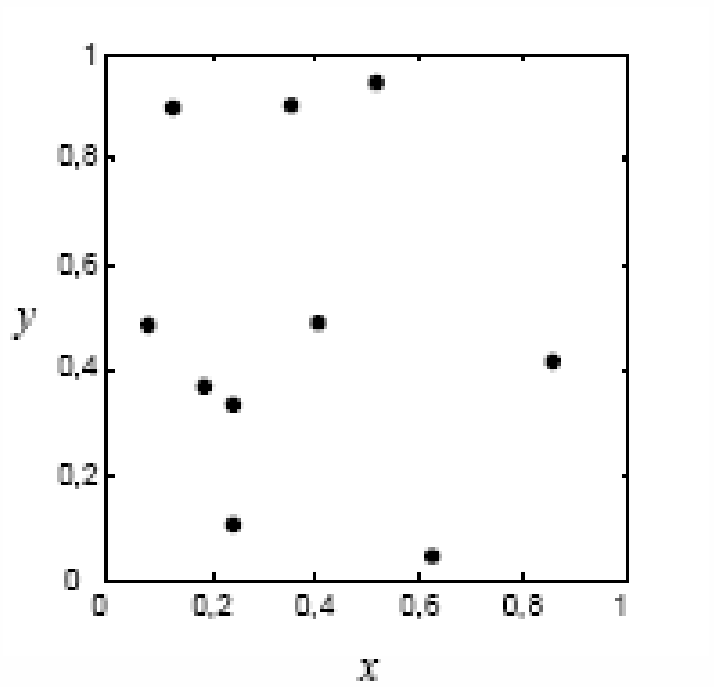
- The direct data visualization is a graphical presentation of the data set that provides a qualitative understanding of the information contents in a natural and direct way.
- The commonly used methods are scatter plot matrices, parallel coordinates, Andrews curves, Chernoff faces, stars, dimensional stacking, etc.
- The direct visualization methods do not have any defined formal mathematical criterion for estimating the visualization quality.
- Each of the features $x_1, x_2, \ldots, x_n$ characterizing the object $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i \in \{1, \ldots, m\}$, is represented in a visual form acceptable for a human being.

Geometric Methods

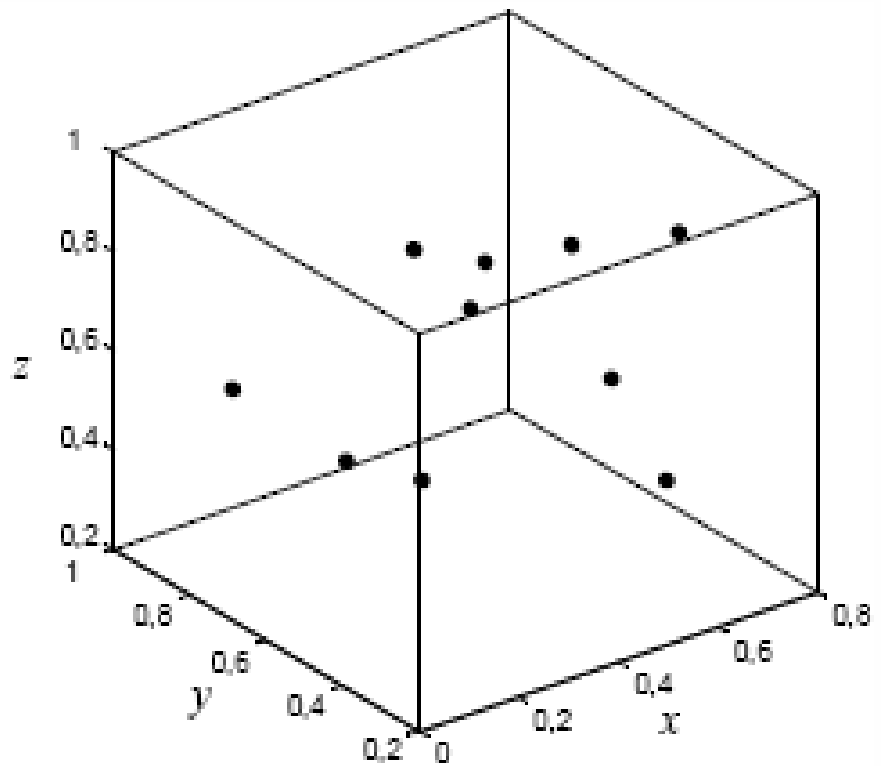
- Geometric visualization methods are the methods where multidimensional points are displayed using the axes of the selected geometric shape.

Scatter Plots

- *Scatter plots* are one of the most commonly used techniques for data representation on a plane $\mathbb{R}^2$ or space $\mathbb{R}^3$. Points are displayed in the classic (x, y) or (x, y, z) format.
- Usually, the two-dimensional ($n = 2$) or three-dimensional ($n = 3$) points are represented by this technique.



a)

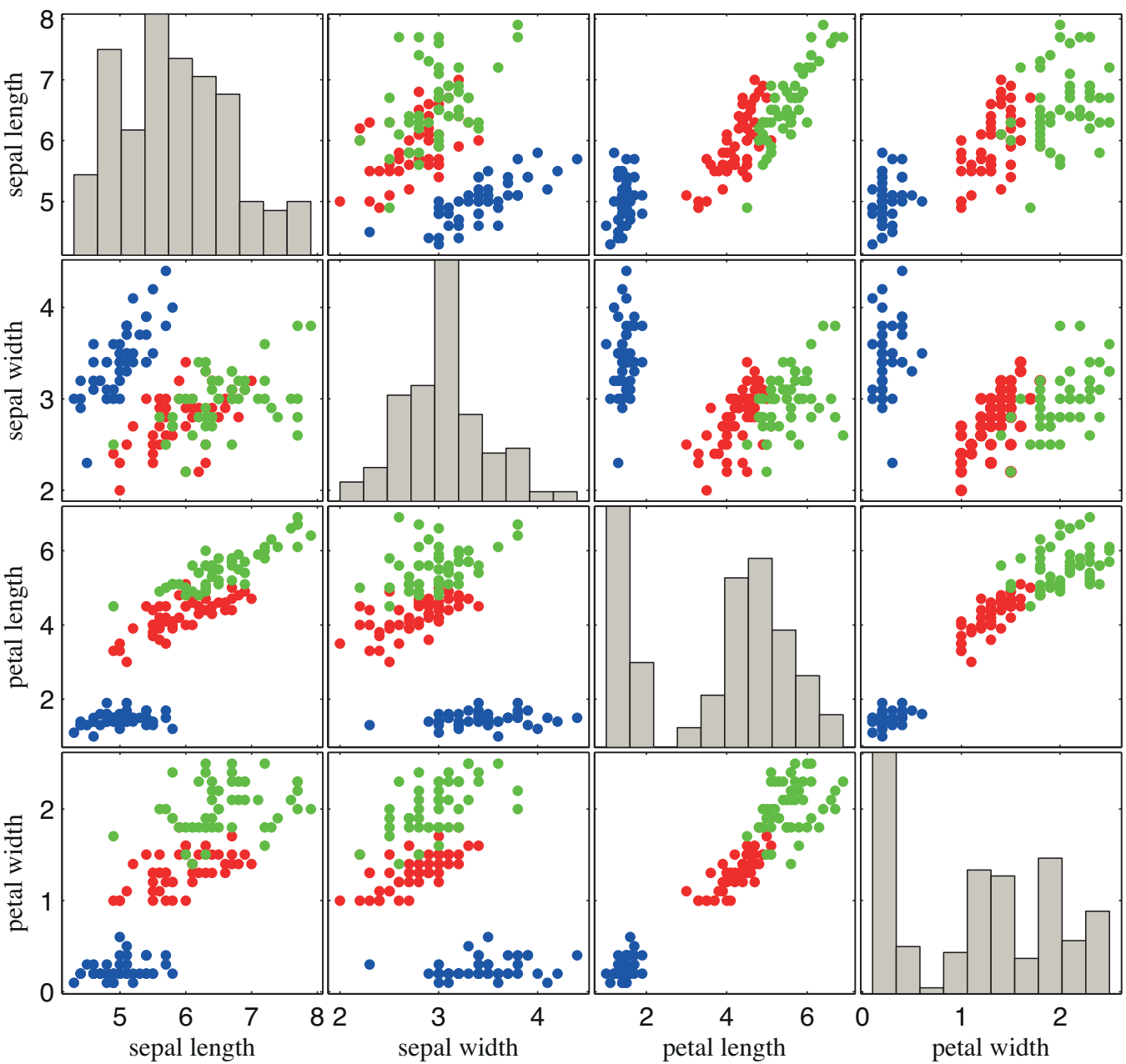


b)

Matrix of Scatter Plots

- Using a *matrix of scatter plots*, the scatter plots can be applied to visualize more higher dimensionality data.
- The matrix of scatter plots is an array of scatter plots displaying all possible pairwise combinations of features.
- If n-dimensional data are analyzed, the number of scatter plots is equal to $\frac{n(n-1)}{2}$.
- In the diagonal of the matrix of scatter plots, a graphical statistical characteristic of each feature can be presented, for example, a histogram of values.
- The matrix of scatter plots is useful for observing all possible pairwise interactions between features.
- The scatter plots can also be positioned in a non-array format (circular, hexagonal, etc.).
- The matrix of scatter plots of the Iris data is presented. We can see that Setosa flowers (blue) are significantly different from Versicolor (red) and Virginica (green).

Scatter Plot Matrix of the Iris Data



Setosa

Versicolor

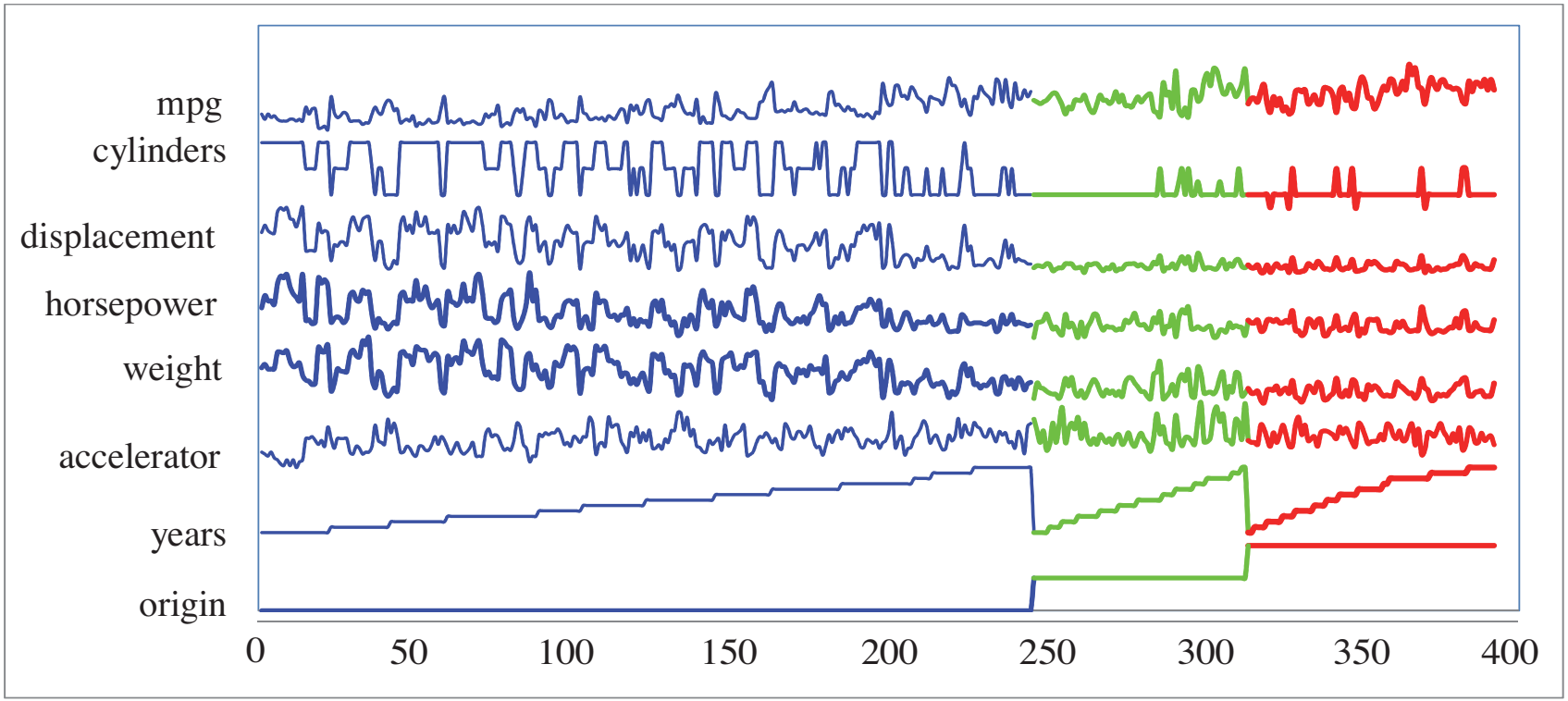
Virginica

Multiline Graphs

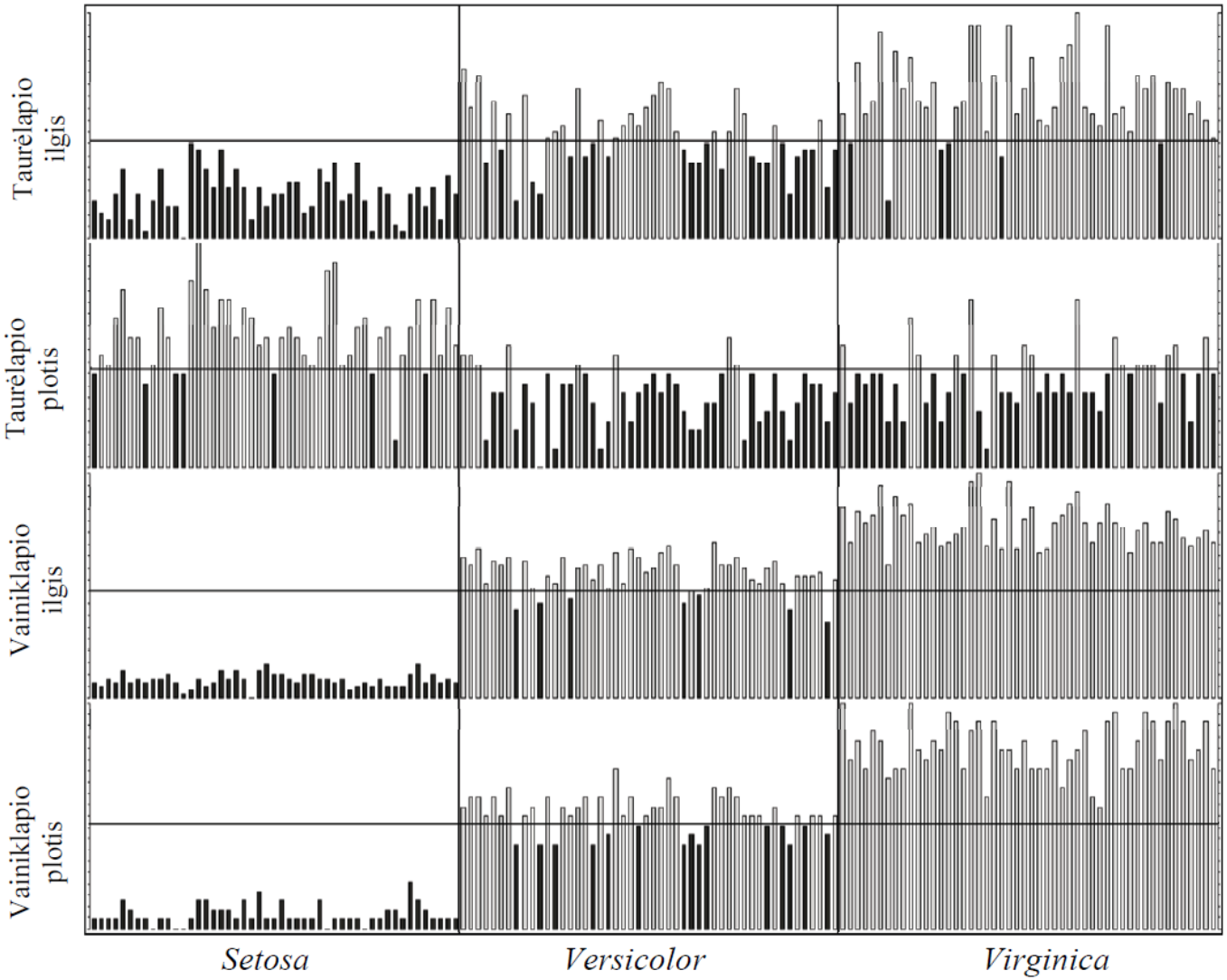
- In *Multiline graphs*, we draw n curves (line graphs) that represent the features depending on the order number of objects.
- An example for Auto MPG data is presented in the figure. The data set is the data on the car produced in the USA, Europe and Japan in 1970-1982 (398 cars). The cars are described by nine features:
  - MPG (miles per gallon) (x_1),
  - the number of cylinders (x_2),
  - displacement (x_3),
  - horsepower (x_4),
  - weight (x_5),
  - acceleration (x_6),
  - model year (x_7),
  - the origin (x_8),
  - the car name (x_9).

Multiline Graphs of the Auto MPG Data

- The data are aligned according to origin (USA, Japan, Europe).

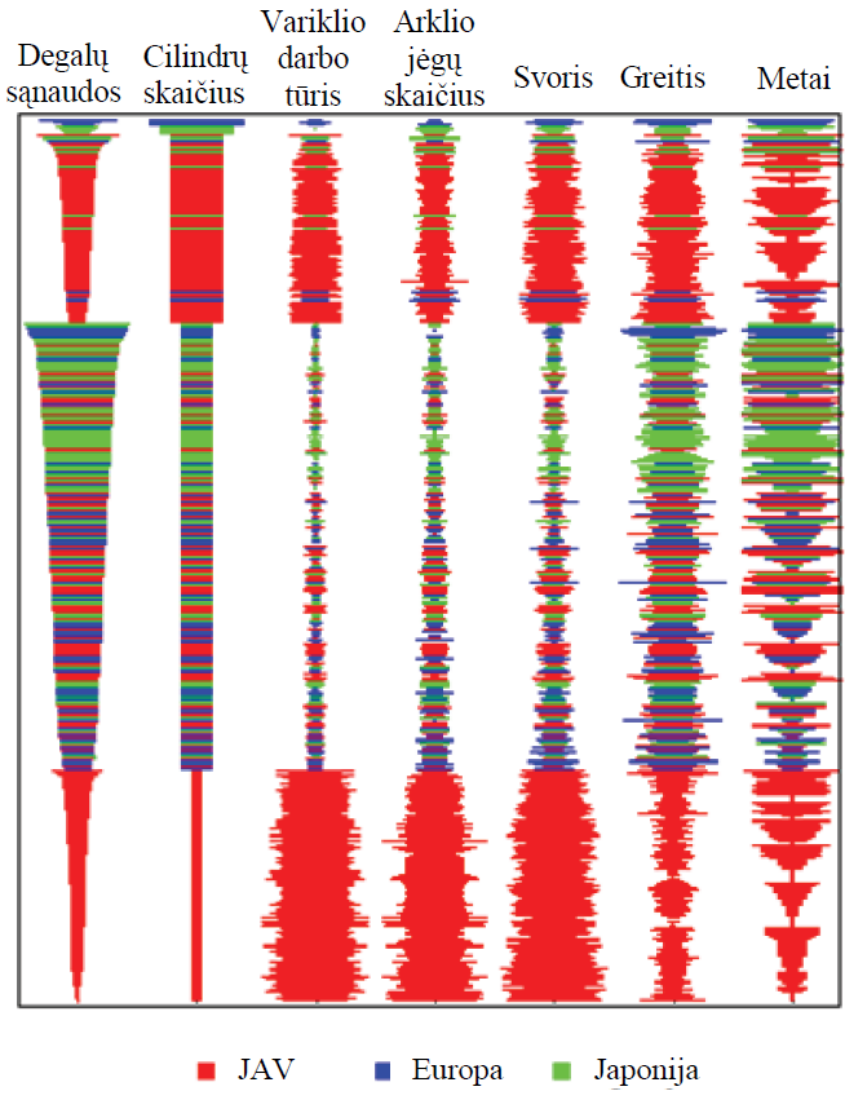


Permutation Matrix



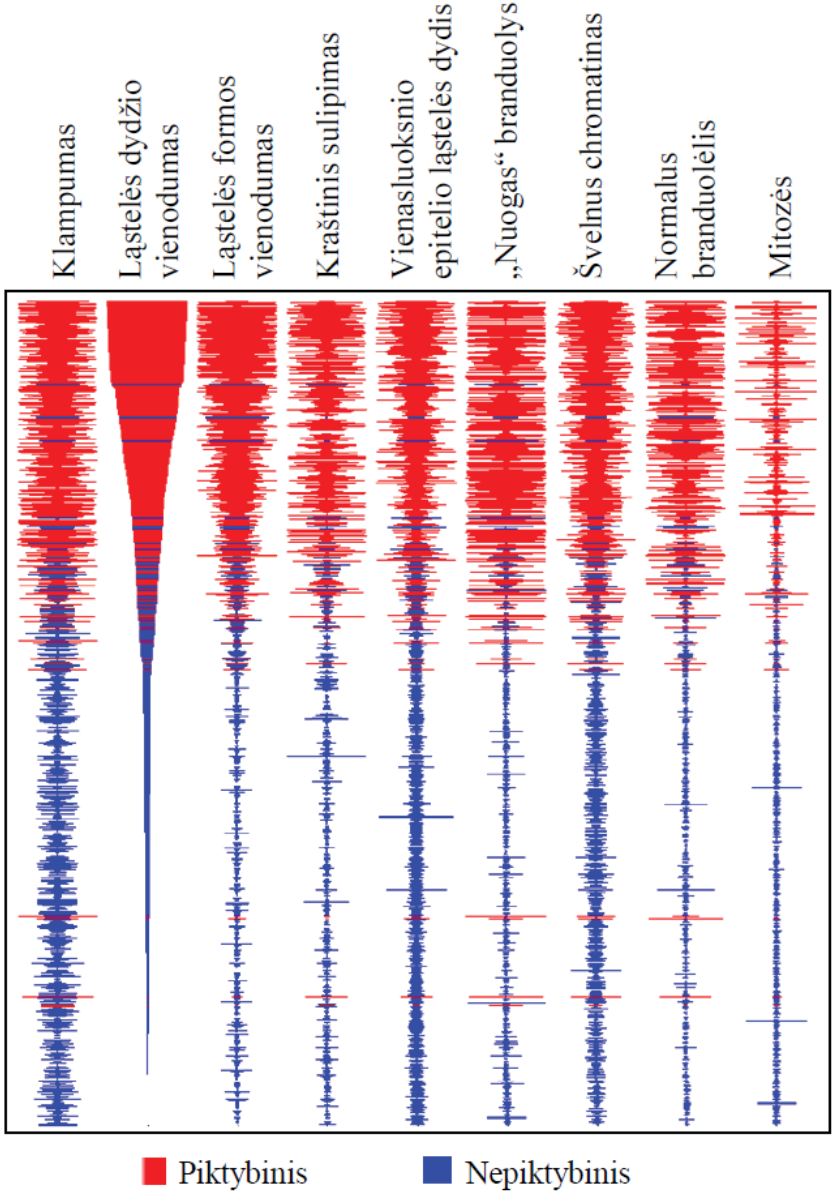
Survey Plot of Auto MPG Data

- Sorted according to the number of cylinders and MPG.



Survey Plot of Breast Cancer Data

- Sorted according to uniformity of cell size.



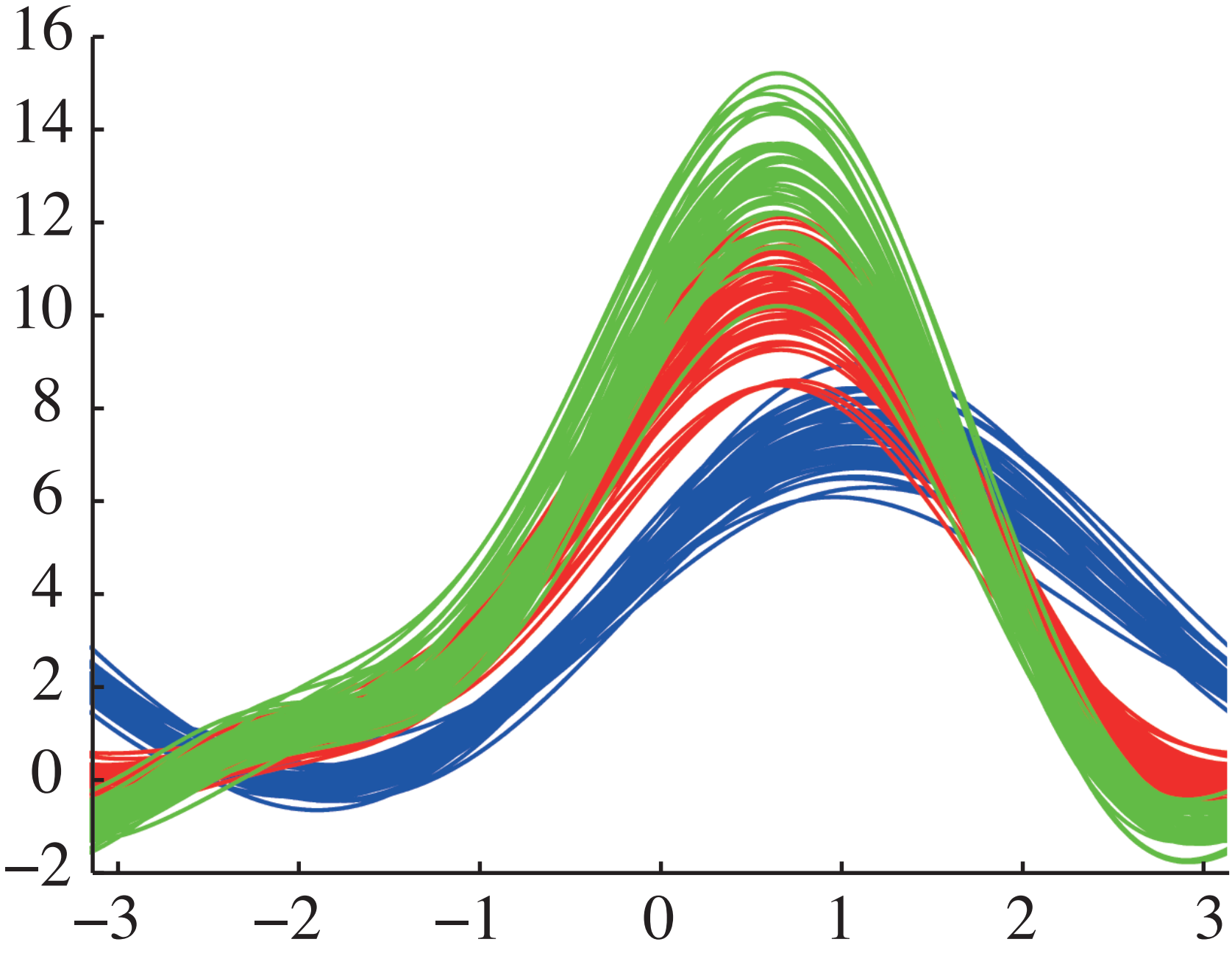
Andrews Curves

- *Andrews curves* plot each n-dimensional point X_i, $i \in \{1, \ldots, m\}$ as a curve (sum of sinusoids), using the function:

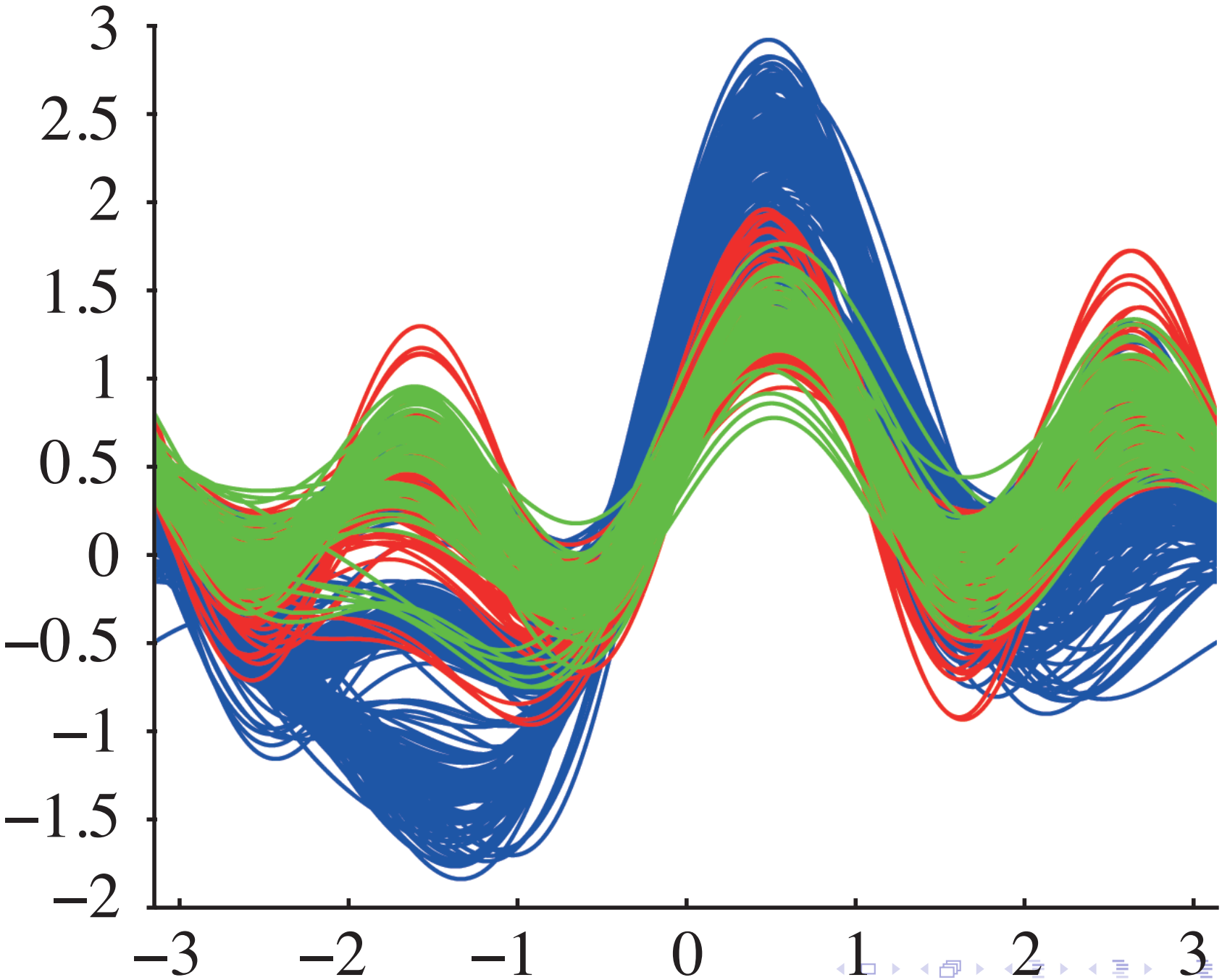
$$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \cdots, \quad -\pi <$$

  where $x_{i1}, x_{i2}, \ldots, x_{in}$ are the values of coordinates of the point X_i.
- Andrews curves of the Iris and Auto MPG data sets are presented. The curves are obtained by the *Matlab* system (`http://www.mathworks.com`), where different species of irises and classes of auto by the origin are painted in different colors.
- An advantage of this method is that it can be used for the analysis of data of a high dimensionality n. A shortcoming is that when visualizing a large data set, i.e. with large enough m, it is difficult to comprehend and interpret the results.

Andrews Curves of Iris Data

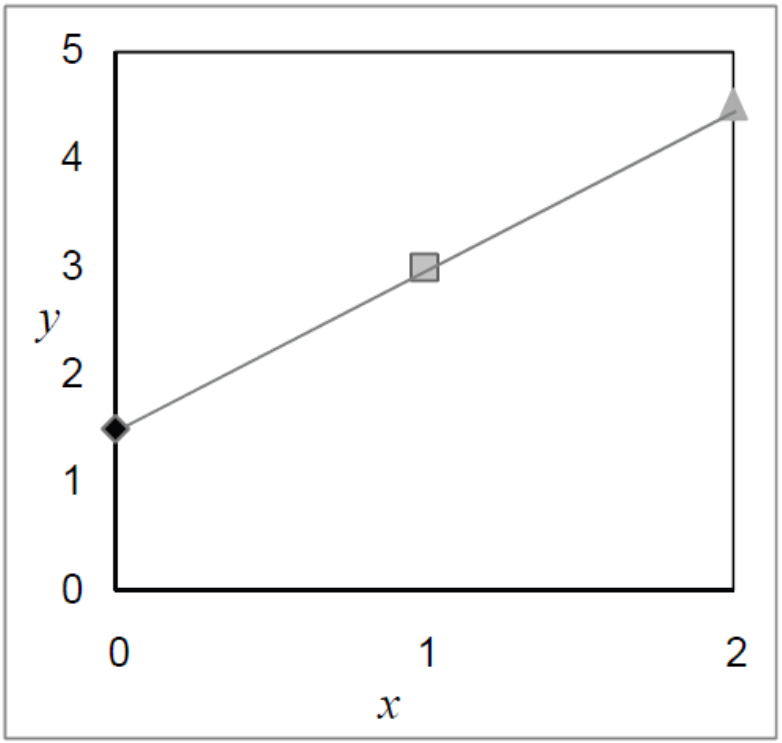


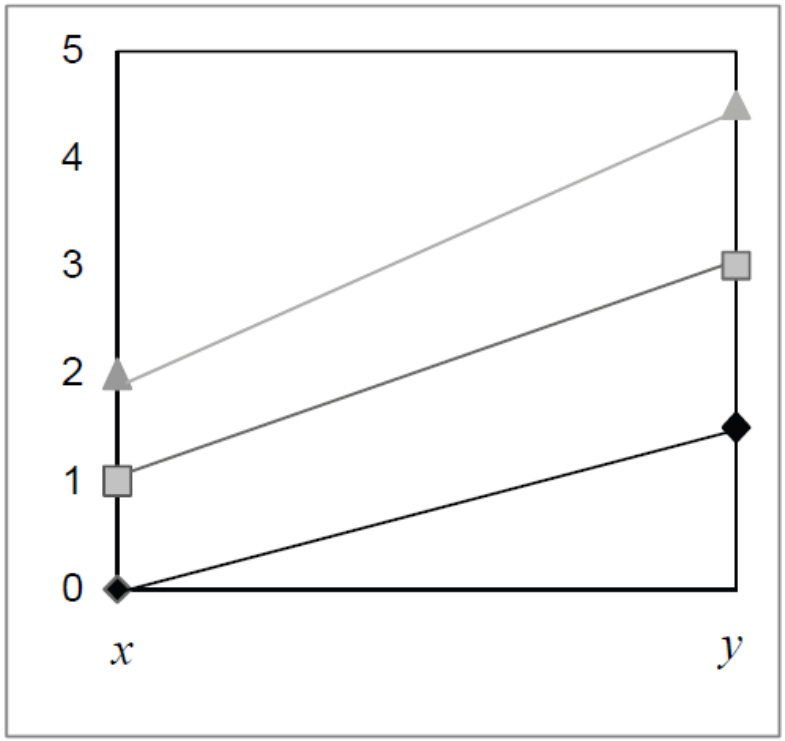
Andrews Curves of Auto MPG Data



Parallel Coordinates

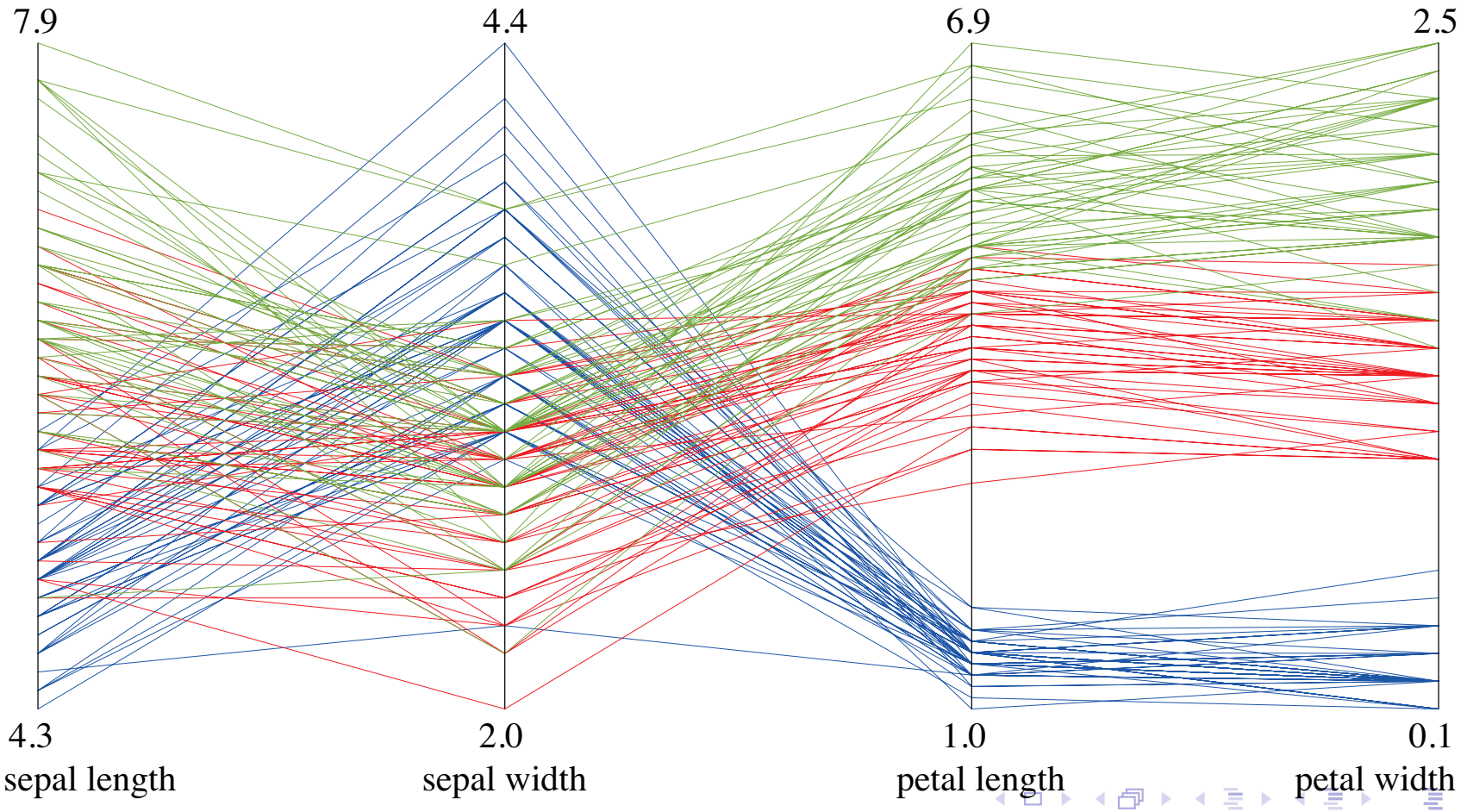
- *Parallel coordinates* as a way of visualizing multidimensional data are proposed by Inselberg in 1981.
- In this method, coordinate axes are shown as parallel lines that represent features.
- An n-dimensional point is represented as $n - 1$ line segments, connected to each of the parallel lines at the appropriate feature value.





Iris Data Represented on the Parallel Coordinates

- The image is obtained using the system *Orange*.
- Different colors correspond to the different species.
- We see that the species are distinguished best by the petal length and width. It is difficult to separate the species by the sepal length and width.



Parallel Coordinates

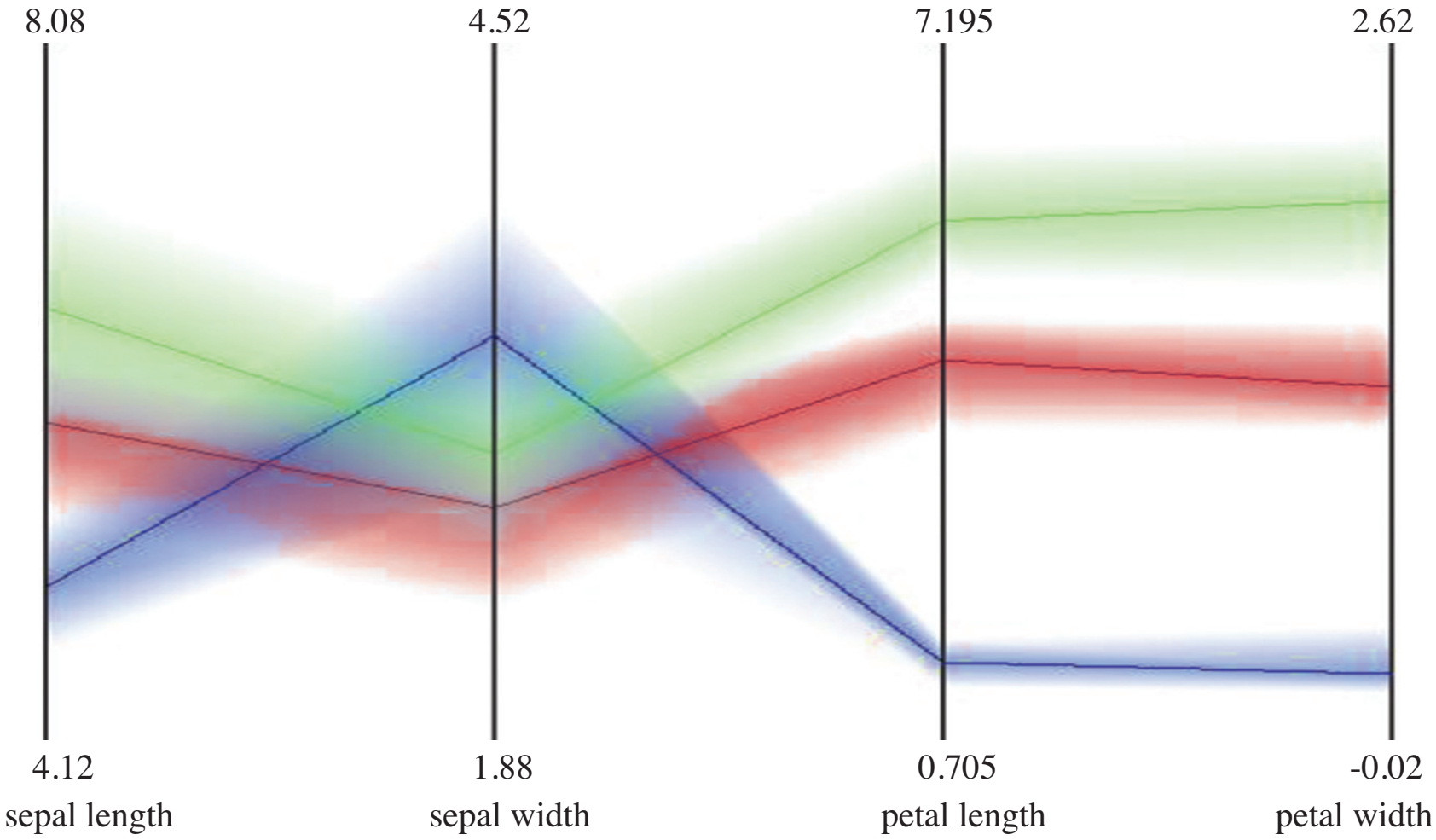
- The parallel coordinate method can be used for visualizing data of high dimensionality.
- However, then the coordinates must be spaced much nearer one to the other. When the coordinates are dense, it is difficult to perceive the data structure.
- When displaying a large data set, i.e. when the number m of objects is large, the interpretation of the results is very complicated, often it is almost impossible.

Hierarchical Parallel Coordinates

- *Hierarchical parallel coordinates* are one of variations of the parallel coordinates.
- When visualizing a large data set by the hierarchical parallel coordinates, the number of overlapping lines, obtained by the parallel coordinates, decreases.
- The data are represented on the hierarchical parallel coordinates as follows:
  - First, the data are grouped into some clusters by one of clustering methods;
  - Afterwards, the data are represented on the parallel coordinates, the centers of clusters are highlighted; the color intensity of the members of clusters depends on how far they are from the cluster center; different clusters are displayed by different colors.
- Hierarchical parallel coordinates allow a visual presentation of clustered data.

Iris Data Represented on the Hierarchical Parallel Coordinates

- The image is obtained using the system *Xmdv*.

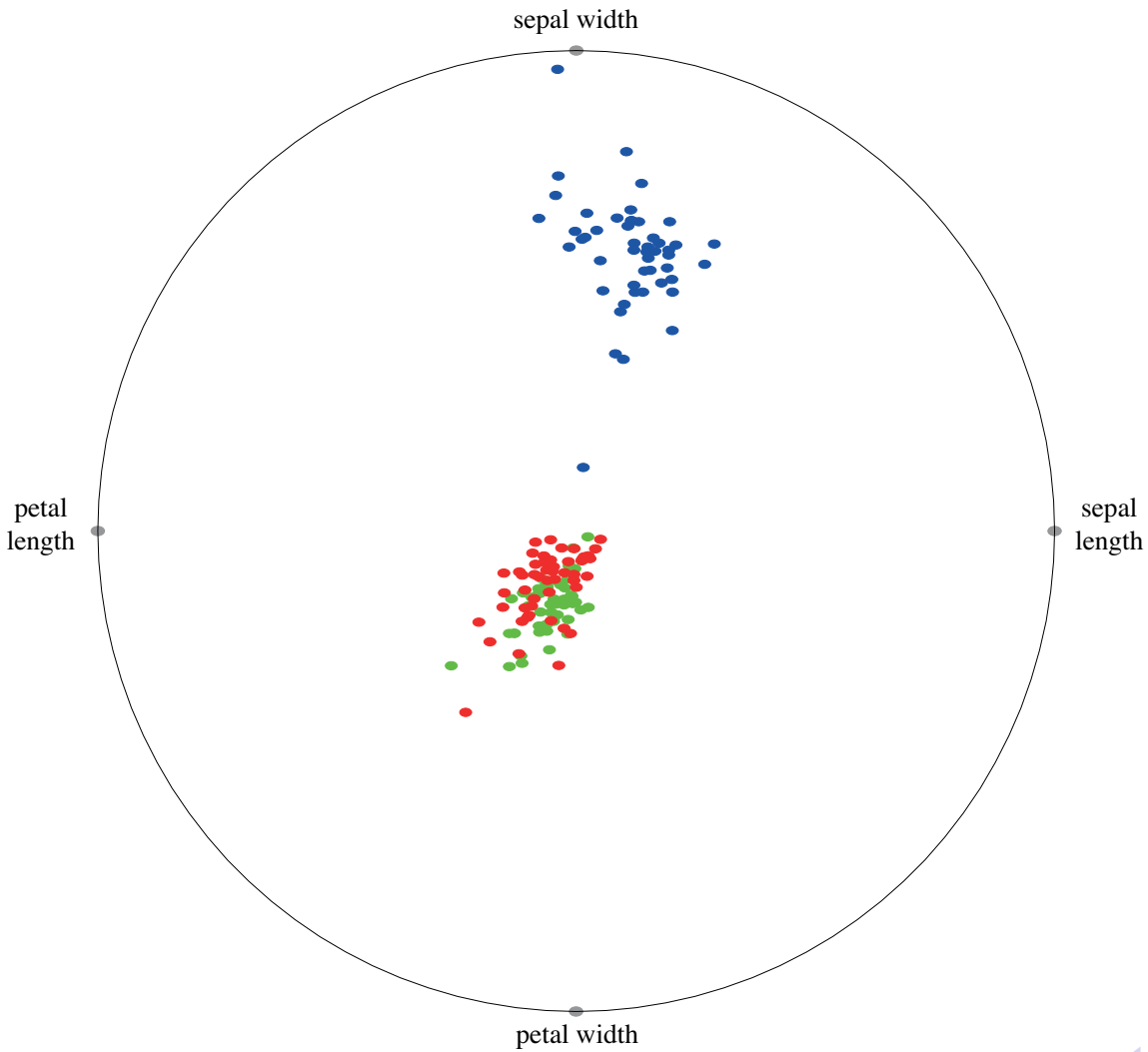


Radial Visualization – RadViz

- In *Radial visualization* (*RadViz*) method a circle is drawn and n so-called dimensional anchors representing features are fixed on this circle uniformly.
- The spring paradigm is applied to display multidimensional data. n springs are allocated to each n-dimensional object. One end of all the n springs is connected among them, other ends of the springs are connected to different dimensional anchors. The position of connection of n springs represents one object.
- The spring constants have the values of features of multidimensional objects. The values of features should be normalized in the range [0, 1] so that the minimal value of each feature were equal to 0, and the maximal one were equal to 1. Each object is displayed as a point in the position that produces a sum of spring forces equal to 0.

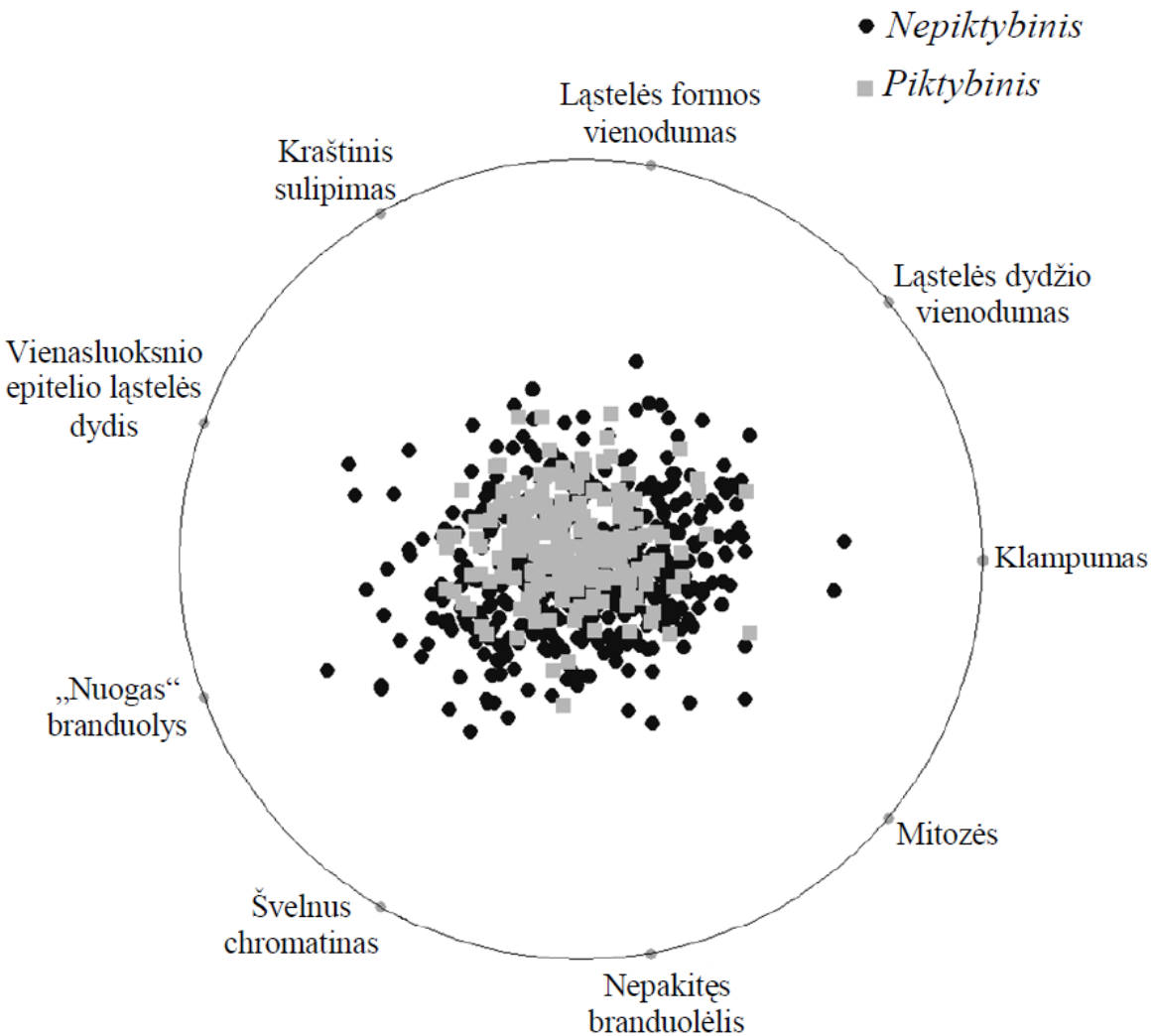
Iris data visualized by the RadViz method

- The petal length, petal width, sepal length, and sepal width are dimensional anchors. The image is obtained using the system *Orange*.



Breast Cancer Data Visualized by the RadViz Method

- Most of the malignant cases concentrate in the center, however, it is almost impossible to separate them from the benign cases.



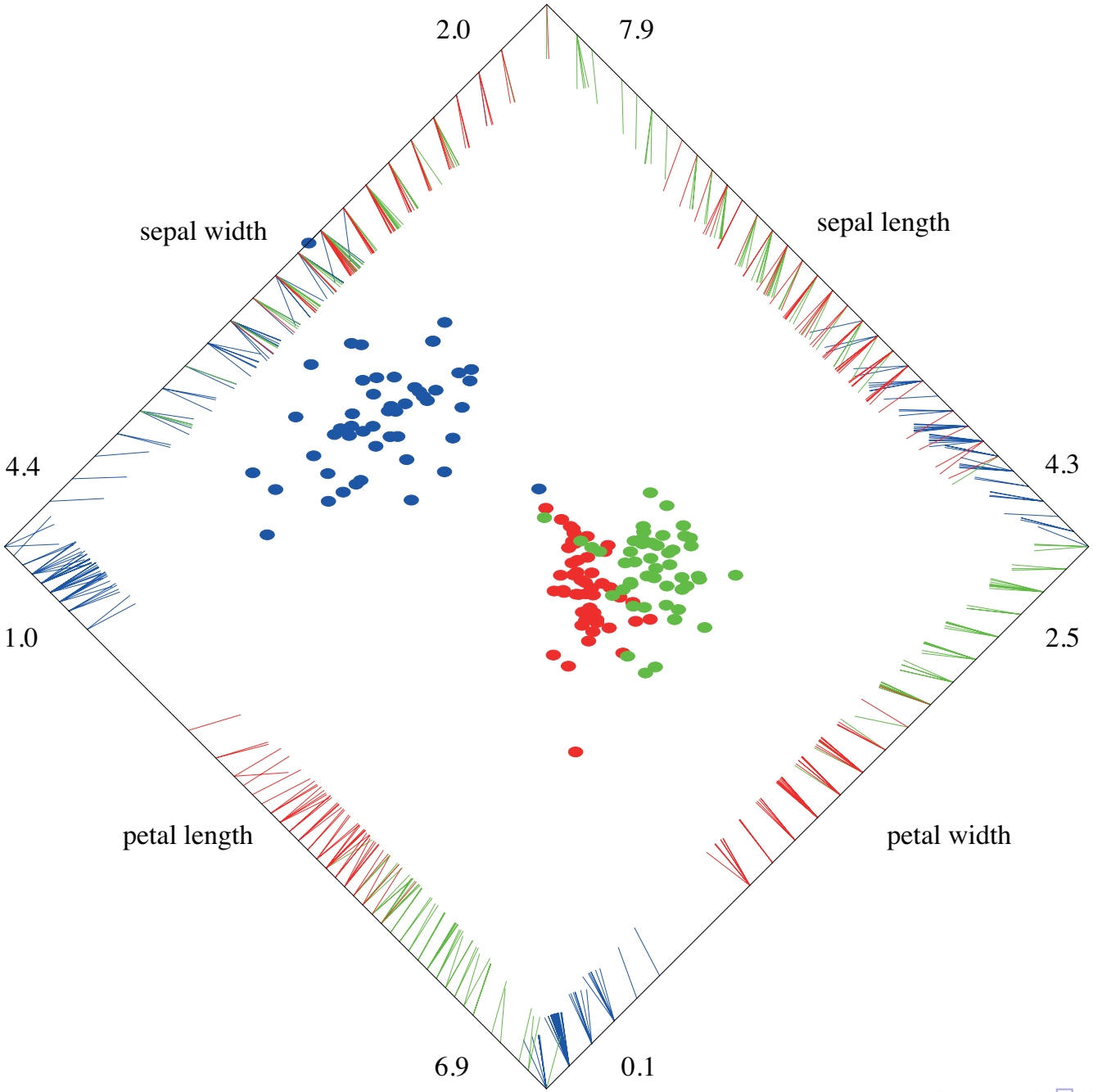
RadViz Modifications: GridViz

- Some modifications of the *RadViz* method are developed: *grid visualization* (*GridViz*) and *polygon visualization* (*PolyViz*).
- *GridViz* places the dimensional anchors (fixed spring end) on a rectangular grid but not on a circle. The spring paradigm is the same as in *RadViz*: the points are plotted where the sum of the spring forces is zero.
- In the *GridViz* case, feature labeling is difficult, but the displayed data dimensionality can be much higher.

RadViz modifications: PolyViz

- A shortcoming of *RadViz* is that *n*-dimensional objects with quite different values of features can appear at the same point.
- If the dimensional anchors are segments of lines, the overlapping of points is reduced. The segments are called anchor segments.
- The *Polygon visualization* (*PolyViz*) method was developed for this purpose.
- In the figure, springs start from the points, corresponding to the values of a particular feature on the anchor segment.

Iris Data Visualized by the PolyViz Method



Iconographic Displays

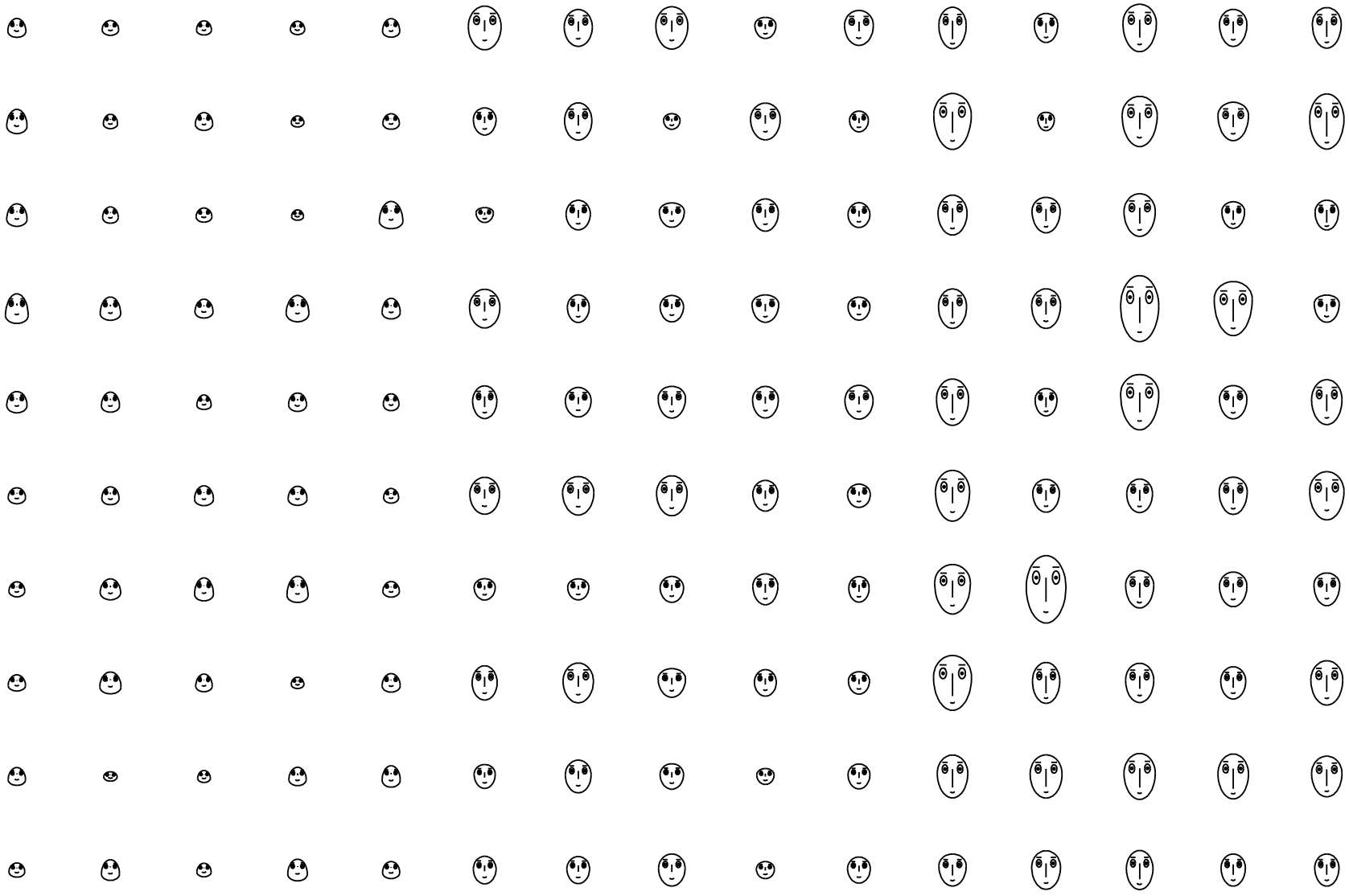
- The aim of visualization of multidimensional data is not only to map the data onto a two- or three-dimensional space, but also to help perceiving them.
- The second aim may be achieved visualizing multidimensional data by *iconographic display* methods. They are also called *glyph* methods.
- Each object that is defined by the n features is displayed by a glyph. Color, shape, and location of the glyph depend on the values of features.
- The most famous methods are *Chernoff faces* and the *star* method, however some methods of more complicated other glyphs may be used as well.

Chernoff Faces

- *Chernoff faces* are designed by Chernoff for visualization of multidimensional data.
- In Chernoff faces, data features are mapped to facial features, such as the angle of eyes, the width of a nose, etc.

Iris Data Visualized by Chernoff Faces

- Sepal length corresponds to the size of face, sepal width – shape of forehead, petal length – shape of jaw, and petal width – length of nose. *Matlab* is used to obtain the image.

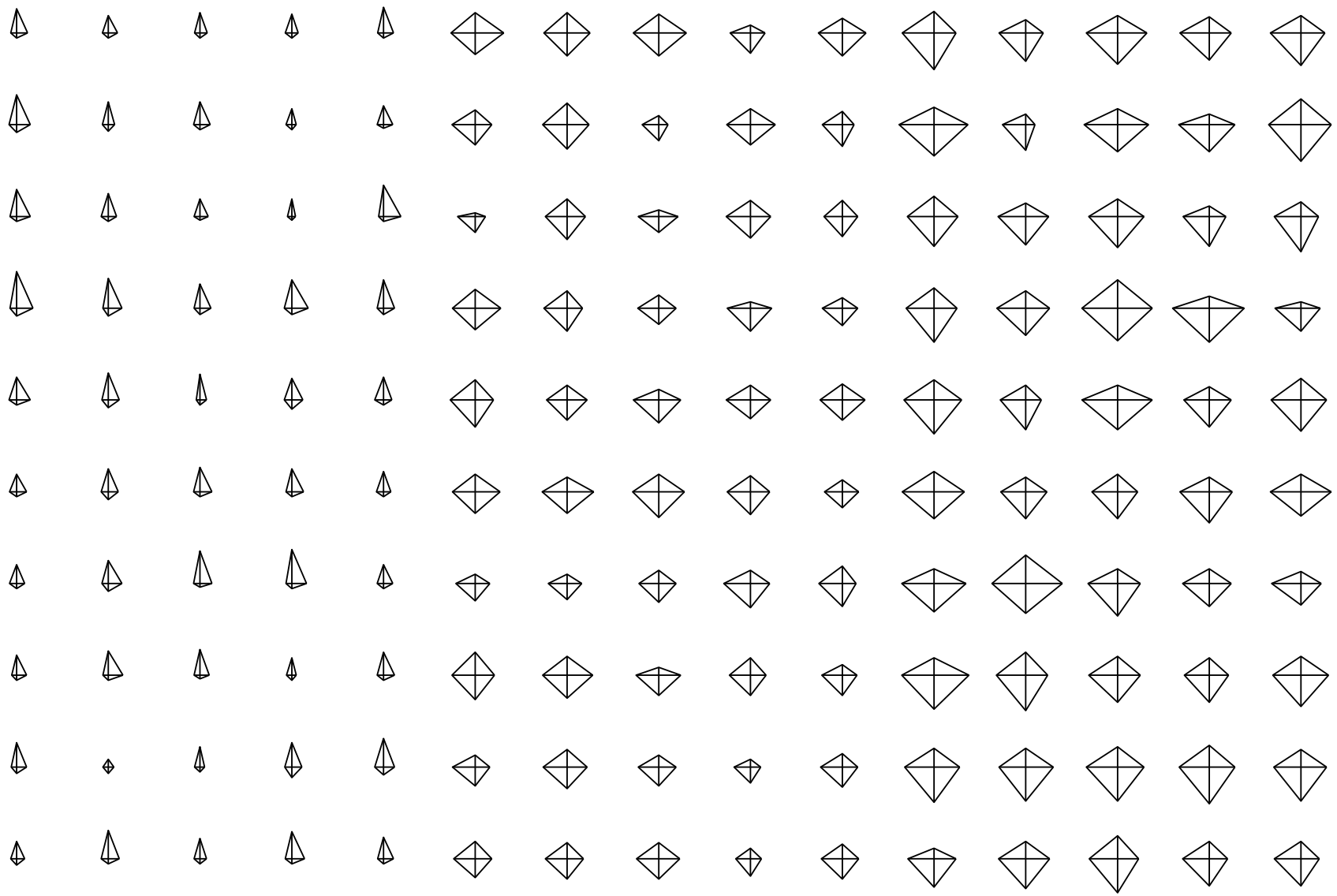


Star Glyphs

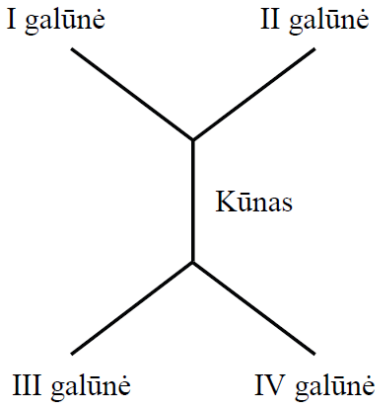
- Other glyphs commonly used for data visualization are *stars*.
- Each object is displayed by a stylized star.
- In the star plot, the features are represented as spokes of a wheel circle, but their lengths correspond to the values of features.
- The angles between the neighboring spokes are equal.
- The outer ends of the neighboring spokes are connected by line segments.

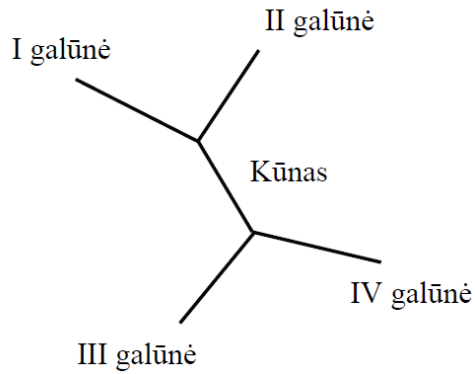
Iris Data Set Visualized by Star Glyphs

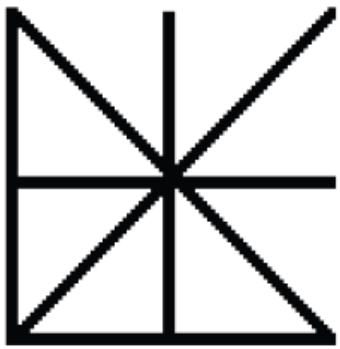
- The stars, corresponding to Setosa irises, are smaller than the other. The larger stars correspond to Virginica irises.

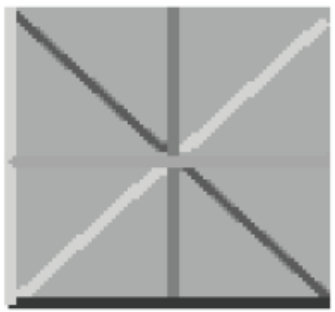


Stick Figure and Color Icon









Hierarchical Displays

- *Hierarchical displays* create a structure of an image such that some features are embedded in displays of other features.
- Visualization of some features is displayed in the structure depending on the values of other features.
- Here we introduce two such techniques: *dimensional stacking* and *trellis display*.

Dimensional Stacking

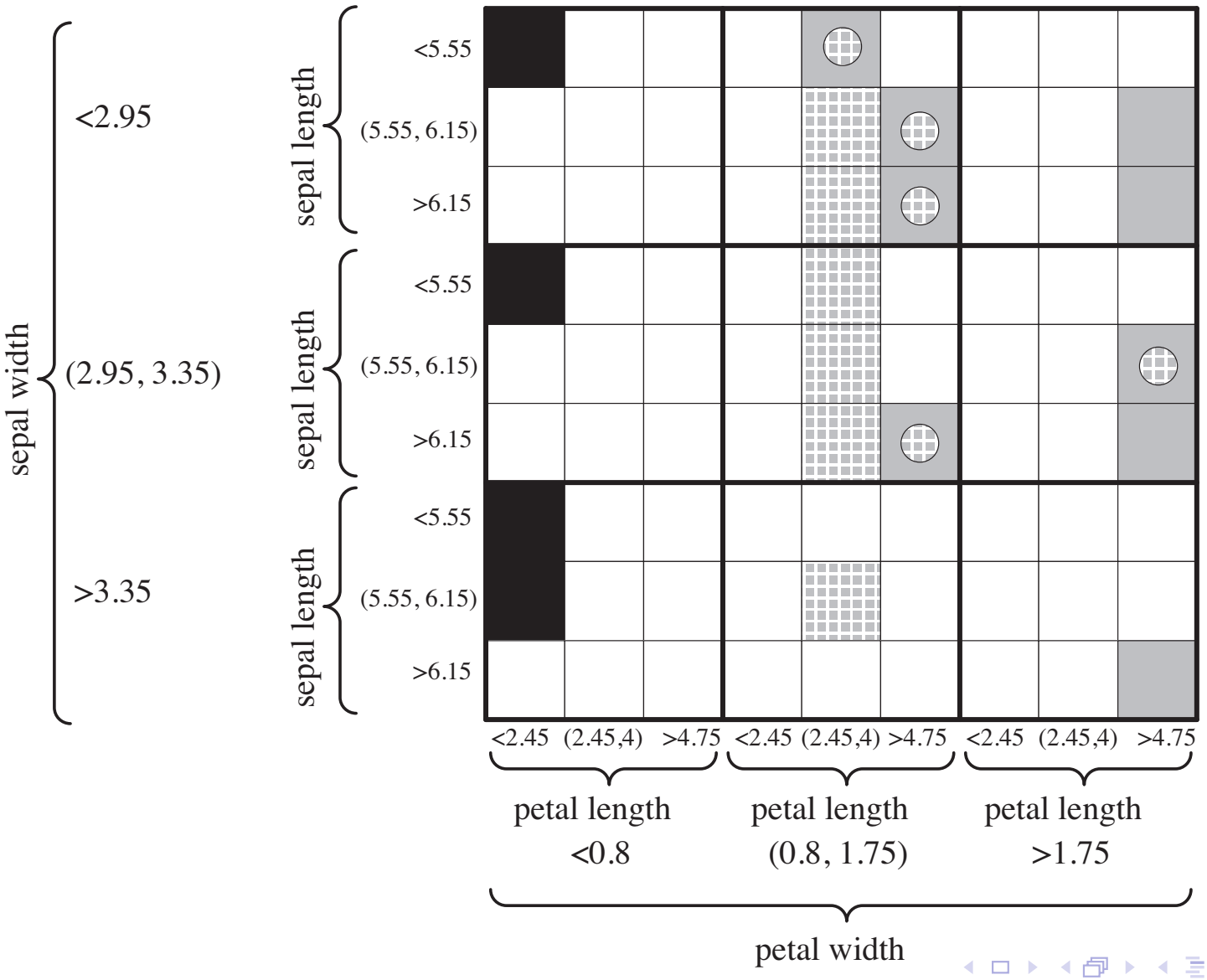
- A predecessor of *dimensional stacking* was the general logic diagrams. Only the Boolean data values 0 and 1 are displayed. M. Ward extends this method later on.
- The dimensional stacking method can be used for exploring clusters and outliers. However, when the dimensionality of data exceeds eight, the display of data and comprehension of the results are difficult.
- The dimensional stacking technique is implemented in the package *Xmdv*.

Scheme of Dimensional Stacking

- The ranges of values of a feature, characterizing the objects, are divided into subranges; a recommendation is that the number of such subranges is not more than five;
- two selected features, called the outer features, are represented by a grid, the numbers of rows and columns are equal to the numbers of subranges;
- when displaying other two features, called the inner features, a new grid is created at each cell of the outer grid; the grids, displaying the inner features, are embedded into all the cells of the outer grid; the recursive embedding continues until all features are displayed;
- the cell of the last embedding is colored, if there are objects the feature values of which are in subranges corresponding to this cell;
- if the classes of objects are known, the color of the cell is selected according to the class of the objects; moreover, the classes are overlapping, colors of the cell may overlap, too.

Iris data visualized by dimensional stacking

- Setosa irises (black cells) are displayed separately from the other two species. The other two species overlap.

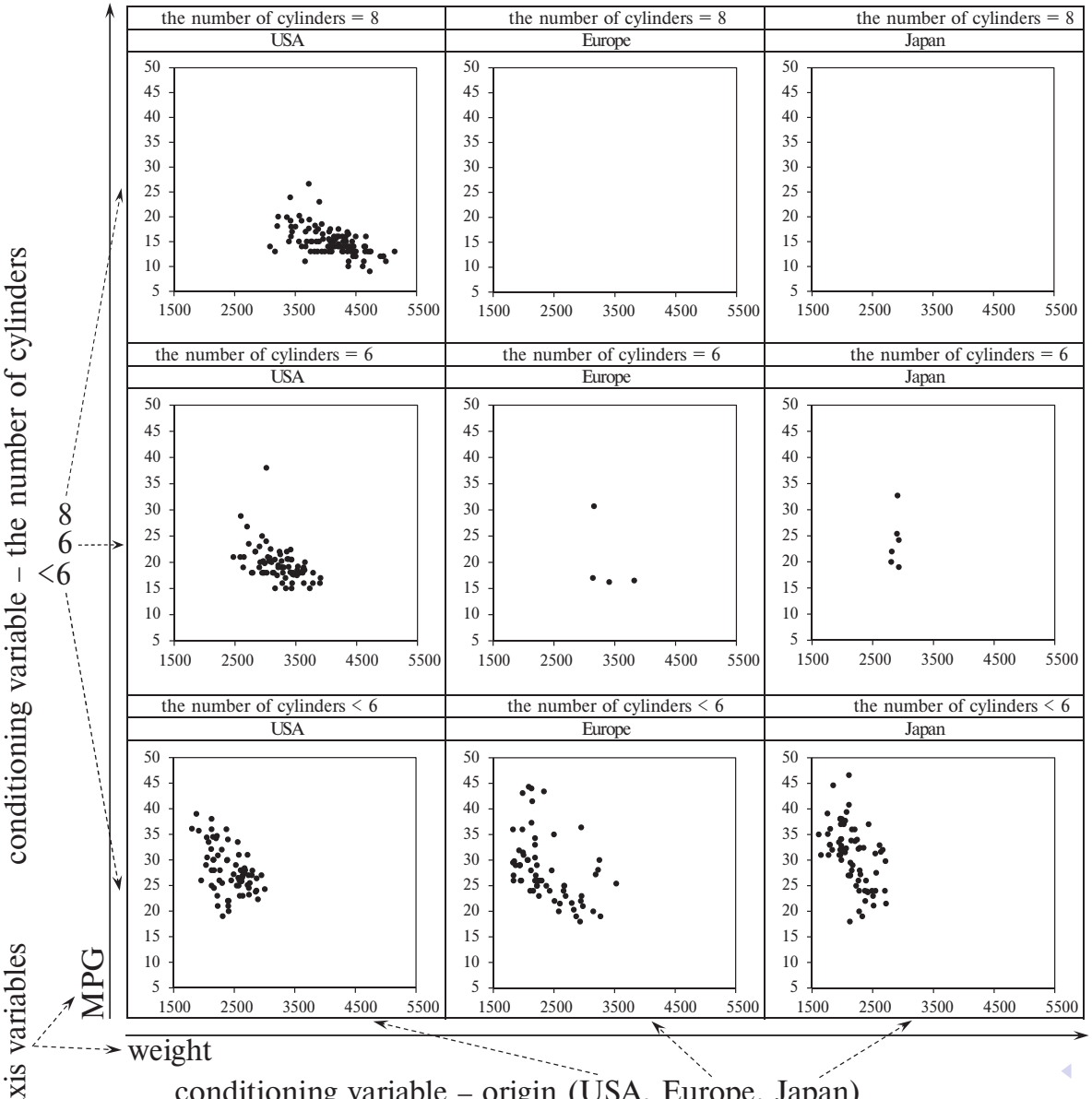


Trellis Display

- The *Trellis display* method is similar to the dimensional stacking. The name of the method is derived from the Latin word ‘tri-liceum’ which means a frame of lattice-work used for climbing plants.
- At first two features are selected. They are called *axis variables*. The ranges of the values of these features are divided into non-overlapping subranges.
- Other features are called *conditioning variables*.
- The panel plots are drawn for each pair of subranges. The panel plots can be scatter, bar, surface plots, etc.

Auto MPG Data Visualized by Trellis Display

- The axis variables are weight and miles per gallon (MPG).
- Origin and number of cylinders are conditioning variables.



Fractal Foam for Iris Data