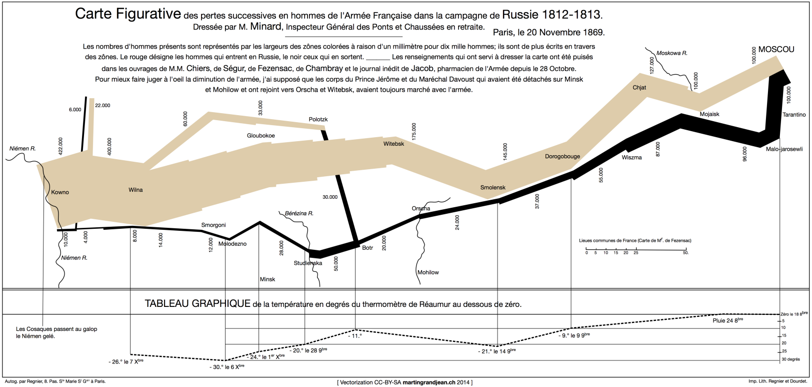
Multidimensional Data Visualization

Introduction

A Good Sketch is Better Than a Long Speech



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ______ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop
le Niémen gelé.

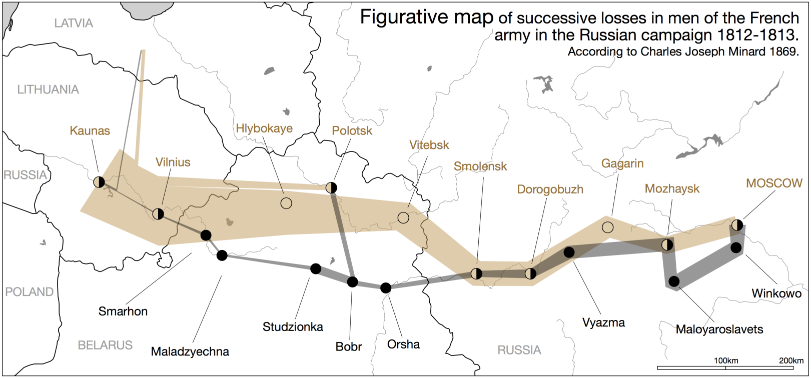
Autog. par Regnier, 8. Pas. S^te^ Marie S^t^ G^ain^ à Paris.

[Vectorization CC-BY-SA **martingrandjean.ch** 2014]

Imp. Lith. Regnier et Dourdet.

Charles Joseph Minard's 1869 diagram of Napoleon's March. The original map shows the road to Moscow in brown and the way back in black. The path is simplified into a single stream, as explained in the description of Minard, under the title. Only a few cities are displayed, the path is summarized in segments between these points. On the way back, a graph shows the temperatures at irregular intervals.

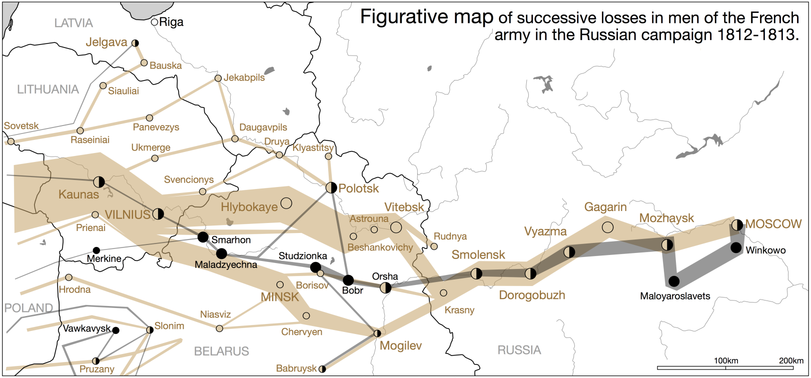
Minard’s Map: Geographic Display



[CC-BY-SA martingrandjean.ch 2014]

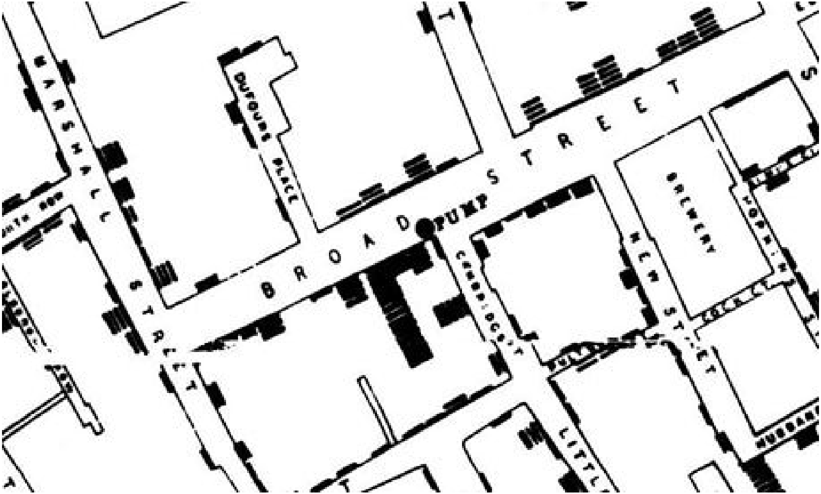
Using data from Minard, this map projects the path taken by Napoleon’s troops in the geographical reality. To make this map understandable, places and borders reflect the current situation (2014). Brown/Black dots indicate the cities crossed twice.

Minard’s Map: Historical Map



The reality is not as simple as the visualization of 1869 suggests: Napoleon’s army was divided into several corps which followed different paths and fortunes. This third map combines Minard’s codes and the most accurate informations we have about the actual route of the different corps of the “Great Army”. The small dots indicate the places that Minard didn’t mention.

Visualization Based Discovery



A map by John Snow showing the clusters of cholera cases in the London epidemic of 1854. Snow used a dot map to illustrate the cluster of cholera cases around the pump.

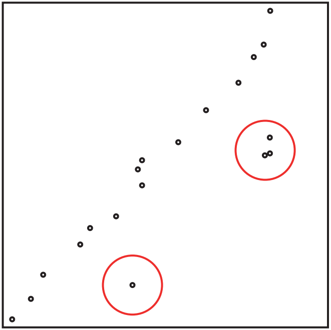
Multidimensional Data and Visualization

- It is often desirable to visualize a data set the items of which are described by more than three features. Therefore, we have multidimensional data and our goal is to make some visual insight into the data set analyzed.
- For human perception, the data must be represented in a low-dimensional space, usually of two or three dimensions. The goal of visualization methods is to represent the multidimensional data in a low-dimensional space so that certain properties (e.g. clusters, outliers) of the structure of the data set were preserved as faithfully as possible.
- Such a visualization of data is highly important in data mining because recent applications produce a large amount of data that require specific means for knowledge discovery.
- The dimensionality reduction or visualization methods are recent techniques to discover knowledge hidden in multidimensional data sets.

Example of Visualization

5.86	2.91	−4.19	−8.49	0.43	−1.13
0.31	−1.14	−2.25	−2.60	−1.58	2.17
11.58	2.97	−14.31	−14.18	3.46	−0.99
15.14	5.46	−20.15	−16.61	0.87	0.31
−1.25	0.39	0.40	2.50	0.16	−0.13
−14.42	−3.81	12.65	13.92	1.94	0.93
5.90	3.36	−10.09	−7.96	−0.85	0.89
−9.55	−0.93	9.71	11.53	2.54	−1.41
13.98	3.41	−20.10	−11.60	−0.59	−1.55
0.85	0.37	−2.40	−3.83	−1.15	0.86
−5.96	−2.06	7.90	9.44	1.06	−1.46
6.39	6.82	−12.52	−8.35	2.05	0.49
−3.92	−1.66	6.54	2.82	−1.70	0.65
3.99	−0.83	−3.87	−1.85	−1.05	1.08
−10.36	−2.47	12.88	10.64	0.76	−0.75
2.62	3.72	9.95	7.88	−0.91	−0.37
0.76	2.63	9.47	10.40	0.35	1.02
2.71	2.99	8.75	10.28	−0.59	2.34
13.84	7.71	−7.00	−6.33	−0.68	1.57

→



- The data visualization allows us to detect the presence of clusters, outliers, or regularities in the analyzed data.
- It is evident that some items of the data set form separate clusters – outliers and the remaining ones are scattered near to a line. These clusters – outliers and distribution around the line can be clearly observed visually on a plane and cannot be recognized directly from the table without a special analysis.

Multidimensional Data Visualization

- A natural idea arises to present multidimensional data, stored in such a table (matrix), in some visual form. It is a complicated problem followed by extensive researches, but its solution allows a human being to gain a deeper insight into the data, draw conclusions, and directly interact with the data.
- Such a possibility to present multidimensional data in a visual form is not one and only. A large number of methods have been developed for multidimensional data visualization.
- It is desirable to preserve certain properties of the structure of the data set as faithfully as possible when transferring from several dimensions to two.
- In this course we will review and discuss various methods for multidimensional data visualization.

Principal Notations

- At first, we determine the principal notions and terms used in this course.
- Here we confront with two principal terms: *object* and *feature*.
- The term *object* can cover various things: people, equipment, products of manufacturing, plants, natural phenomena, etc.
- An object is characterized by some *features*. For example, the patient is an object, he (she) can be described by a number of features, such as name, sex, age, and diagnostic test results like blood pressure, cholesterol level, etc.
- If the data set consists of a lot of objects, then the data set is called a large data set. If the number of features is large, then the data set is called a high-dimensional data set.

Multidimensional Data

- Objects are also called *items*, *instances*, *samples*, *observations*.
- Features are called *attributes*, *parameters*, *properties*, *variables dimensions*.
- Objects described by the same features $x_1, x_2, \ldots, x_n$ form a *data set*. Assume that any feature may take some numerical values.
- A combination of values of all features characterizes a particular object

  $$X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \; i \in \{1, \ldots, m\},$$

  where n is the number of features, m is the number of objects, i is the order number of the object.
- If the objects $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = 1, \ldots, m$ are described by more than one feature, the data characterizing the objects are called *multidimensional data*.
- If the number of features is n, then $X_1, X_2, \ldots, X_m$ are the n-dimensional data items.

Multidimensional Points

- Often $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ are interpreted as points in the multidimensional space $\mathbb{R}^n$, where n defines the dimensionality of the space.
- The coordinate values of point X_i are values of the features $x_{i1}, x_{i2}, \ldots, x_{in}$. In such a case, we have a matrix (table) X of numerical data:

$$X = \{X_1, X_2, \ldots, X_m\} = \left\{x_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\right\}, \tag{1}$$

  and the ith row of this matrix is a point $X_i \in \mathbb{R}^n$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i \in \{1, \ldots, m\}$ and x_{ij} is the jth coordinate of the ith point, m is the number of points in the data set. The data point X_i contains feature values of corresponding object.
- Sometimes it does not suffice to refer to X_i as a point, so the notion of a *vector* can be useful to enlarge the properties of points. The points $X_1, X_2, \ldots, X_m$ can be conceived as vectors bound to the origin $(0, 0, \ldots, 0)$.

Proximity of Data

- Note that there are cases where we do not have and cannot get a set of numerical values of the features characterizing a particular object. However, we can estimate proximities between two objects.
- Let us determine the notion of *proximity* between two objects X_i and X_j.
- A (dis)similarity is a proximity that indicates how two objects X_i and X_j are (dis)similar. The (dis)similarity is denoted by δ_{ij}.
- If δ_{ij} is a similarity, a high δ_{ij} value indicates that the objects X_i and X_j are very similar.
- For dissimilarities, a small δ_{ij} value indicates that the objects are very similar.
- When the proximities are known, the visualization of objects $X_1, X_2, \ldots, X_m$ may be carried out using the matrix of their proximities $\Delta = \{\delta_{ij}, i, j = 1, \ldots, m\}$. The advantage is that the dimensionality n can be unknown. This often happens, for example, in psychological tests.

Proximity Measures

- A proximity matrix can be obtained from matrix X applying some proximity measure, too.
- Often the proximity is measured using the Euclidean distance, which belongs to the group of Minkowski distances. The Minkowski distance between two objects $X_k = (x_{k1}, x_{k2}, \ldots, x_{kn})$ and $X_l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ is defined by the formula:

$$d_q(X_k, X_l) = \left\{ \sum_{j=1}^{n} |x_{kj} - x_{lj}|^q \right\}^{\frac{1}{q}} .$$

- Some other proximity measures are also possible: Canberra distance, Bray-Curtis dissimilarity, correlation, etc.

Minkowski Distances

- The following Minkowski distances may be derived for different q:
- City-block or Manhattan distance, $q = 1$:

$$d_1(X_k, X_l) = \sum_{j=1}^{n} |x_{kj} - x_{lj}|.$$

- Euclidean distance, $q = 2$:

$$d_2(X_k, X_l) = \sqrt{\sum_{j=1}^{n} |x_{kj} - x_{lj}|^2}.$$

- Chebyshev distance, $q = \infty$:

$$d_\infty(X_k, X_l) = \max_j \left| x_{kl} - x_{lj} \right| .$$

Conditions of Distances

- Here the distances between two objects X_k and X_l satisfy the following conditions:
- $d(X_k, X_l)$ is a nonnegative real number;
- $d(X_k, X_k) = 0$;
- $d(X_k, X_l) = d(X_l, X_k)$, i.e. the distance from object X_k to object X_l is equal to the distance from object X_l to object X_k;
- $d(X_k, X_l) \leq d(X_k, X_j) + d(X_j, X_l)$, i.e. the distance between any two objects X_k and X_l cannot be larger than a sum of distances between objects X_k, X_j and X_l, X_j (triangle inequality).

Data Normalization

- Let the analyzed data set $X = \{X_1, X_2, \ldots, X_m\}$ contains n-dimensional vectors $X_i = (x_{i1}, x_{i2}, \ldots, x_{in}),\ i = 1, \ldots, m$, i.e. i-th row of the matrix X is a vector X_i.
- Often the values of features $x_1, x_2, \ldots, x_n$ of a multidimensional data set $X = \{X_1, X_2, \ldots, X_m\}$ are of different scales or measured in different units (e.g. kilograms, meters, degrees).
- Therefore the scales must be unified/normalized before analysis of the data set. There are several ways of normalization.

Data Normalization: Normalizing Residuals

1. Values of parameters are changed so that the mean was 0 and standard deviation was 1. For each feature x_j the average

$$\overline{x_j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$$

and standard deviation

$$\sigma_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_{ij} - \overline{x_j})^2$$

is estimated. Every value of feature x_{ij} is transformed using

$$x_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}.$$

Data Normalization: Feature Scaling

2. Feature scaling is used to bring all values into the range $[0,1]$. It is also called unity-based normalization. For each feature x_j minimal $x_{j\,\min}$ and maximal $x_{j\,\max}$ values are estimated and every value of feature x_{ij} is transformed using
$$x_{ij} = \frac{x_{ij} - x_{j\,\min}}{x_{j\,\max} - x_{j\,\min}}.$$

Methods of Multidimensional Data Visualization

- The methods of multidimensional data visualization can be divided into two groups.
- 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.
- Our aim 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) Andrews curves,
   e) parallel coordinates,
   f) radial visualization (RadViz) and its modifications GridViz and PolyViz.
2. Iconographic displays:
   a) Chernoff faces,
   b) star glyphs.
3. Hierarchical displays:
   a) dimensional stacking,
   b) trellis display,
   c) hierarchical parallel coordinates.

Projection (Dimensionality Reduction) 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*.
- These methods usually invoke formal mathematical criteria by which the projection distortion is minimized.

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.

Artificial Neural Networks

- Artificial neural networks may also be used for visualizing multidimensional data. They realize various nonlinear projections.

1. Self-organizing map.
2. Neural gas.
3. Curvilinear component analysis.
4. Multidimensional scaling using artificial neural networks:
   a) supervised learning strategy,
   b) unsupervised learning strategy,
   c) combinations of self-organizing map and neural gas with multidimensional scaling.
5. Auto-associative neural network.
6. NeuroScales.

Multidimensional Data Sets

- We have presented one of the possible classifications of methods for multidimensional data visualization.
- The best way to investigate the visualization methods is to use the test data sets with the known structure. The performance of the methods, discussed in this course, is illustrated on the real-life and artificial data sets.

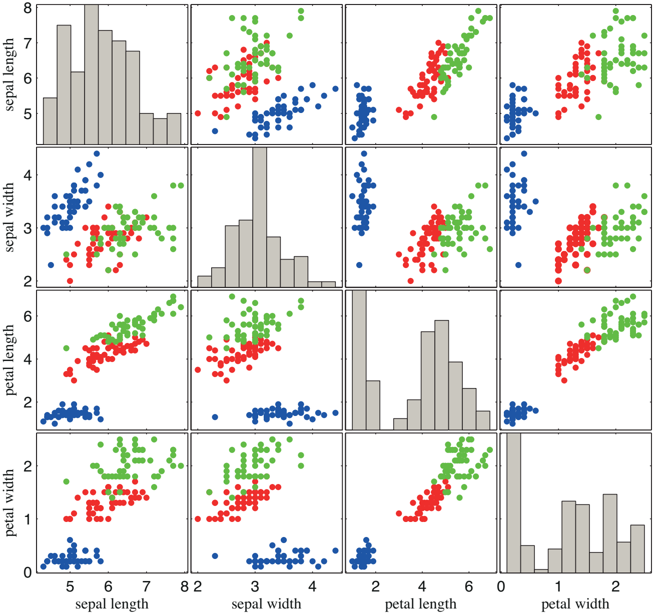
Examples of Multidimensional Data

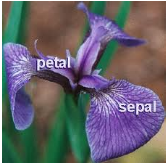
- Some data sets are used to illustrate the methods for visualizing multidimensional data and experimental investigations. The data sets are described by n-dimensional points $X_1, X_2, \dots, X_m$, where $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = 1, \dots, m$, or the dissimilarity matrix Δ of size $m \times m$. The coordinates of points are defined by the values of features $x_1, x_2, \dots, x_n$ of corresponding objects.
- The elements δ_{ij} of dissimilarity matrix describe the proximity of the ith and jth objects.
- Some multidimensional data examples from the database „UCI Repository of Machine Learning Databases“ (http://archive.ics.uci.edu/ml/).
  - Fisher Iris data set.
  - Auto MPG data set.
  - Breast Cancer data set.

Fisher Iris Data Set

- The *Iris* data set consists of 150 flowers of three species: *Setosa*, *Virginica* and *Versicolor*. Each species is represented by 50 flowers ($m = 150$, $n = 4$).
- Four features of each flower were measured:
    - sepal length (x_1),
    - sepal width (x_2),
    - petal length (x_3),
    - petal width (x_4).

Scatter Plot Matrix of Iris Data





Setosa

Versicolor

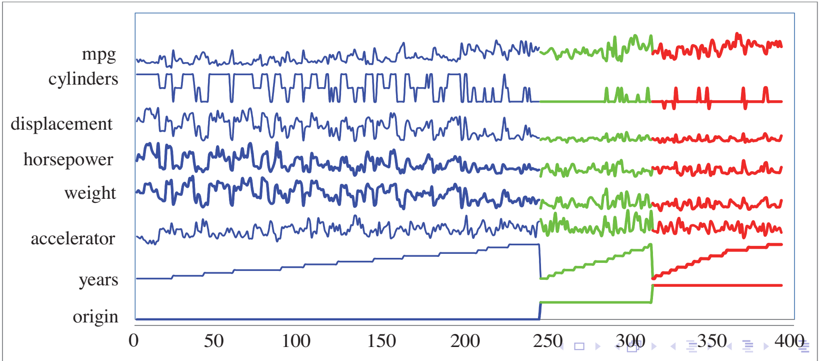
Virginica

Auto MPG Data Set

- The *Auto MPG* 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).
- The last two features are not numerical, therefore, they are not used in the visualization process. Therefore, the seven-dimensional ($n = 7$) points are used for visualization.

Multiline Graphs of Auto MPG data

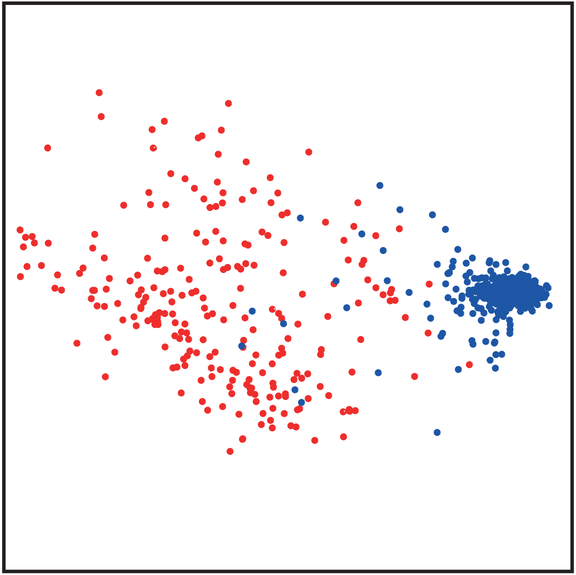
- MPG (miles per gallon) (x_1),
- the number of cylinders (x_2),
- displacement (x_3),
- horsepower (x_4),
- weight (x_5),
- acceleration (x_6),
- model year (x_7),
- the origin (x_8).



Breast Cancer Data Set

- The *Breast Cancer* data set was obtained from the University of Wisconsin Hospitals, USA. 699 observations of the breast cancer are collected. Each instance has one of the two possible classes: benign or malignant. There are nine features:
  - clump thickness (x_1),
  - uniformity of cell size (x_2),
  - uniformity of cell shape (x_3),
  - marginal adhesion (x_4),
  - single epithelial cell size (x_5),
  - bare nuclei (x_6),
  - bland chromatin (x_7),
  - normal nucleoli (x_8),
  - mitoses (x_9).
- There are some missing values of features, so the objects with missing values are eliminated from the data set for visualization. The visualized data set consists of 683 points ($m = 683$, $n = 9$).

Breast Cancer Data Visualized Using Principal Components

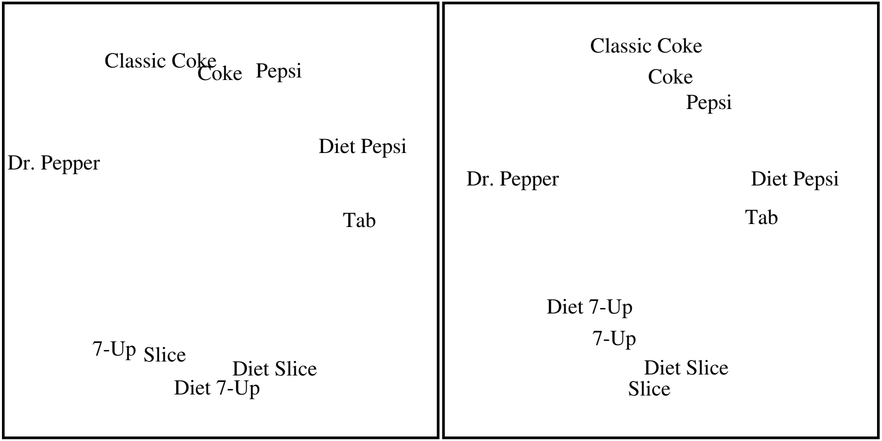


Soft Drinks Data Set

- The *Cola* data set is based on experimental testing of several soft drinks: Pepsi, Coke, Classic Coke, Diet Pepsi, Diet Slice, Diet 7-Up, Dr Pepper, Slice, 7-Up, Tab.
- 38 students have tested ten ($m = 10$) different brands of soft drinks.
- Each pair was judged on its dissimilarity in a nine-point scale (1 – very similar, 9 – completely different).
- The matrix of accumulated dissimilarities Δ_{cola}.

	1	2	3	4	5	6	7	8	9	10
1. Pepsi	0	127	169	204	309	320	286	317	321	238
2. Coke	127	0	143	235	318	322	256	318	318	231
3. Classic Coke	169	143	0	243	326	327	258	318	318	242
4. Diet Pepsi	204	235	243	0	285	288	259	312	317	194
5. Diet Slice	309	318	326	285	0	155	312	131	170	285
6. Diet 7-Up	320	322	327	288	155	0	306	164	136	281
7. Dr Pepper	286	256	258	259	312	306	0	300	295	256
8. Slice	317	318	318	312	131	164	300	0	132	291
9. 7-Up	321	318	318	317	170	136	295	132	0	297
10. Tab	238	231	242	194	285	281	256	291	297	0

Soft Drinks Data Visualized Using Multidimensional Scaling

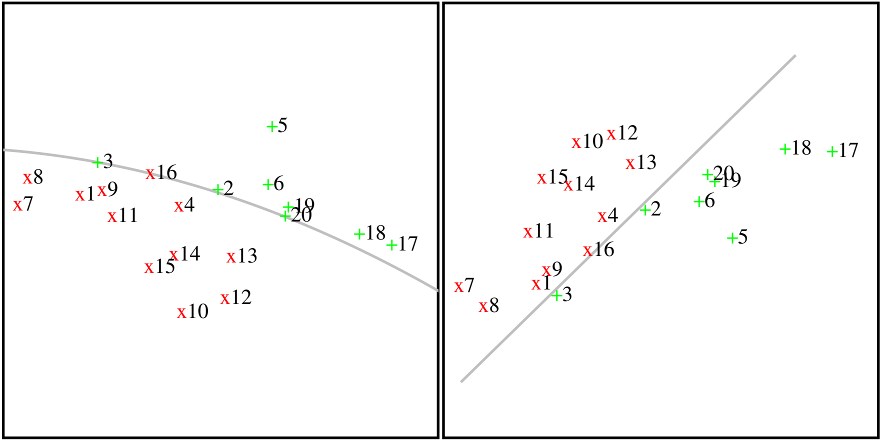


Pharmacological Data

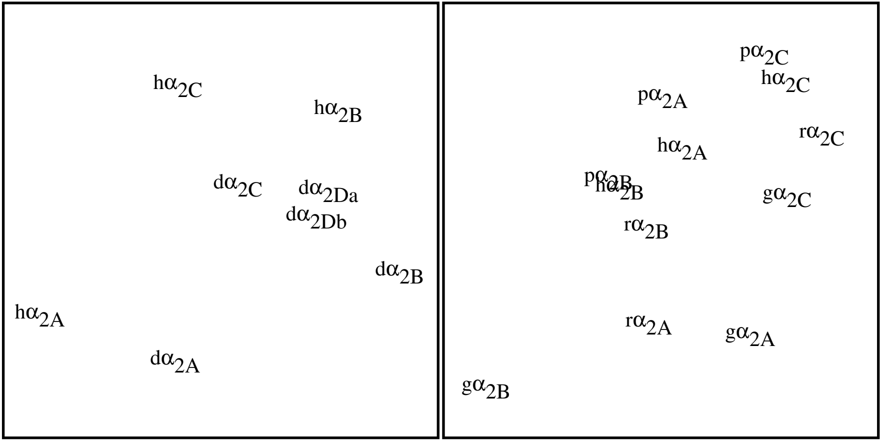
- Pharmacological data, e.g. the values of inhibition, are constants which under the given experimental conditions link the affinity of a given ligand to a given receptor protein.

	Proteins							
Ligands	hα_{2A}	zα_{2A}	hα_{2B}	zα_{2B}	hα_{2C}	zα_{2C}	zα_{2Da}	zα_{2Db}
1. Atipamezole	1.6	13	1.5	5.0	4.3	2.1	5.1	6.9
2. Clonidine	10	89	44	250	110	55	120	150
3. Dexmedetomidine	1.3	2.2	4.7	7.6	6.5	12	4.1	3.7
4. Idazoxan	17	85	24	40	17	17	52	94
5. Oxymetazoline	2.1	5.1	1100	1200	130	1300	1100	440
6. UK14,304	32	40	320	1200	190	700	260	280
7. L657.743	0.8	6.9	0.7	1.2	0.09	1.0	1.6	1.3
8. Rauwolscine	1.9	1.0	1.1	1.4	0.2	0.5	2.3	2.3
9. Yohimbine	5.9	5.2	7.5	9.3	4.6	3.4	6.4	4.0
10. Chlorpromazine	990	110	43	1.1	330	83	18	19
11. Clozapine	32	3.3	12	9.3	2.1	3.2	12	24
12. ARC239	2100	1800	9.6	36	66	280	55	44
13. Prazosin	1030	330	66	300	31	100	68	64
14. Spiperone	540	45	12	51	11	63	15	18
15. Spiroxatrine	320	150	2.4	93	3.1	35	11	11
16. WB-4101	5.4	11	60	51	1.9	19	31	16
17. 2-Amino-1-phenylethanol	1300	5400	4200	9400	8100	5100	3700	4000
18. Dopamine	2000	790	6300	4400	1200	3900	1300	1700
19. (–)-Adrenaline	150	140	710	910	130	1080	500	470
20. (–)-Noradrenaline	110	260	680	647	250	580	380	510

Ligands Visualized Using Multidimensional Scaling



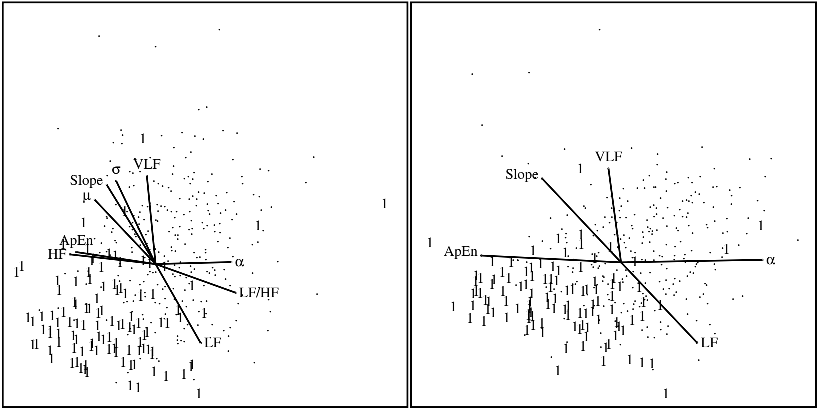
Proteins Visualized Using Multidimensional Scaling



Analysis of Heart Rate Oscillations with Respect to Characterization of Sleep Stages

- To diagnose sleep-related disorders and diseases, it is important to determine the sleep structure of a patient.
- The results of recent investigations show the dependence among characteristics of heart rate and sleep stages.
  - mean (x_1),
  - standard deviation (x_2),
  - very low frequency band 0.01–0.05 Hz (x_3),
  - low frequency band 0.05–0.15 Hz (x_4),
  - high frequency band 0.15–0.5 Hz (x_5),
  - ratio of normalized power in low and high frequency bands ($x_6 = x_4/x_5$),
  - approximate entropy that defines the complexity of behaviour of the time series (x_7),
  - fractal scaling exponent of a detrended fluctuation analysis (DFA) (x_8),
  - slope of a curve of a progressive detrended fluctuation analysis (x_9).

Heart Rate Data Visualized Using Multidimensional Scaling


VLF
σ
Slope
μ
ApEn
HF
α
LF/HF
LF
VLF
Slope
ApEn
α
LF