Friedrich-Alexander-Universität
Technische Fakultät

3. Getting to Know Your Data

Knowledge Discovery in Databases

Dominik Probst, `dominik.probst@fau.de`
Chair of Computer Science 6 (Data Management), Friedrich-Alexander-University Erlangen-Nürnberg
Summer semester 2025

Outline

Data Objects and Attribute Types

Types of Data Sets

Records:

- Relational records.
- Data matrix, e.g. numerical matrix, crosstabs.
- Document data: text documents, typically represented as *term-frequency vectors*.
- *Transaction data*.

Graph and network:

- World wide web.
- Social information networks.
- Molecular structures.

Stud. ID.	**Forename**	**Surname**	**Study Programm**
22305910	Jane	Doe	Data Science
23061995	Dominik	Doe	Informatics
23312331	Friedrich	Doe	Data Science
20312131	Melanie	Doe	Informatics

TID	**Items**
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk

Types of Data Sets

Friedrich-Alexander-Universität
Technische Fakultät

Ordered data:

- Video data: sequences of images.
- Temporal data: time series.
- Sequential data: transaction sequences.
- Genetic sequence data.

Semester	WS22	SS23	WS23	SS24	WS24
Students	592	550	520	524	553

Spatial, image and multimedia:

- Spatial data: maps.
- Image data.
- Video data.

Important Characteristics of Structured Data

Dimensionality:
Curse of dimensionality (sparse high-dimensional data spaces).

Sparsity:
Only presence counts.

Resolution:
Patterns depend on the scale.

Distribution:
Centrality and dispersion.

Data Objects

Data Object

A data set consists of data objects. A single *data object* represents an entity.

Also known as: Samples, examples, instances, data points, objects, tuples.

Examples:

- Sales database: customers, store items, sales.
- Medical database: patients, treatments.
- University database: students, professors, courses.

Data objects are described by attributes:

- Database rows $\rightarrow$ data objects.
- Columns $\rightarrow$ attributes.

Attributes

Attribute

An *attribute* represents a specific characteristic such as customer unique identifier, customer name, and customer address.

Also known as: Variable, feature, field, dimension.

Further terminology:

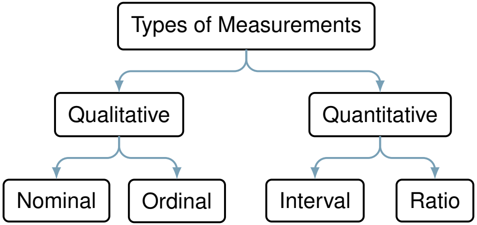
- *Observations*: Observed or measured values of an attribute.
- An *attribute vector* or *feature vector* describes a data object by its set of attributes.
- Data sets with one attribute are referred to as *univariate* (more than one: *multivariate*).

Two different views on attributes:

- The view of the *statistics* community.
- The view of the *informatics* community.

Types of Attributes: Statistics

- Most common classification based on Stevens[1]:



Qualitative Measurements:
Describes an attribute without providing a size or quantity.

Quantitative Measurements:
Describes an objective, measurable attribute. Often numerical.

[1] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946, ISSN: 00368075, 10959203. Accessed: Apr. 9, 2025. [Online]. Available: http://www.jstor.org/stable/1671815

Qualitative Measurements

Friedrich-Alexander-Universität
Technische Fakultät

Nominal Measurements:

- Categories, states, or "names of things".
- E.g. $\texttt{color} = \{\text{blue,red,green}\}$.
- Other examples: `marital_status`, `occupation`, `ID`, `ZIP code`.
- **Special case -** Binary measurements:
  - Only two states (0 and 1).
  - **Symmetric binaries**: both outcomes equally important, such as `sex`.
  - **Asymmetric binary**: outcomes not equally important.
    E.g. medical test (positive vs. negative).
    Convention: assign 1 to most important outcome (e.g. diabetes, HIV positive).

Ordinal Measurements

- Values have a meaningful order (ranking), but magnitude between successive values is not known.
- E.g. $\texttt{size} = \{\text{small, medium, large}\}$
- Other examples: `grades`, `army rankings`.

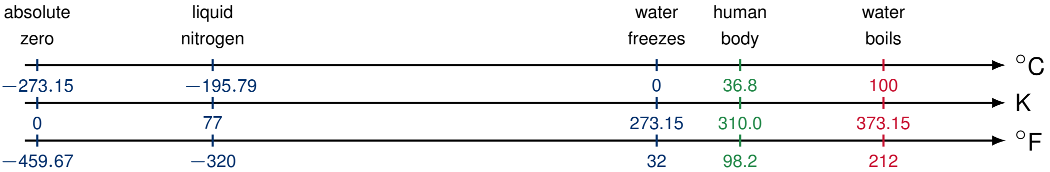
Quantitative Measurements

Interval-Scaled Measurements

- Measured on a scale of **equally sized** units.
- No true zero-point.
- Values have order.
- E.g. temperature in $^\circ C$ or $^\circ F$.

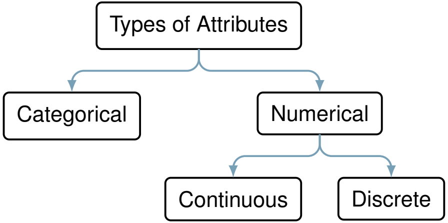
Ratio-Scaled Measurements

- Inherent **zero point**.
- We can speak of values as being an order of magnitude larger than the unit of measurement.
- E.g. temperature in Kelvin.



Types of Attributes: Informatics

- The view of the *informatics* community focuses on the **data type** of the attribute:



Categorical Attributes:
Corresponds to qualitative measurements.
Most often strings.

Numerical Attributes:
Corresponds to quantitative measurements.
E.g. integers, floats.

Numerical Attributes

Continuous Attributes

- Has real numbers as attribute values. E.g. temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Discrete Attributes

- Has finite or countably infinite elements. E.g. ZIP code, profession, or the set of words in a collection of documents.
- Sometimes represented as integer variables.

Note

Binary attributes are a special case of discrete attributes.

Friedrich-Alexander-Universität
Technische Fakultät

Basic Statistical Descriptors of Data

Basic Statistical Descriptions of Data

Motivation:

- To better understand the data: central tendency, variation and spread.

Data dispersion characteristics:

- Median, max, min, quantiles, outliers, variance etc.

Numerical dimensions correspond to sorted intervals.

- Data dispersion: analyzed with multiple granularities of precision.
- Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures.

- Folding measures into numerical dimensions.
- Boxplot or quantile analysis on the transformed cube.

Population vs. Sample

Population

- Collection of *all* data objects of interest. E.g. all people currently living in Germany or all enrolled students at FAU.
- Population size often denoted by N.
- Hard to define and to observe. E.g. a survey requiring all enrolled students is often not feasible as not all will participate.

Note

Unlikely that we have data of a whole population. Thus, we can assume we always have a sample.

Sample

- Randomly selected and representative subset of a population obtained by a sampling method[2].
- Sample size often denoted by n.
- Sample guards against (unconscious) investigator biases.
- Samples are faster and less costly to obtain.
- More control regarding (missing) values and outlier.

[2] Example include but not limited to simple random sampling, stratified sampling.

Central Tendency: Mean and Trimmed Mean

Arithmetic Mean

Simple and most commonly used measure of central tendency:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Sensitive to extreme values (outliers).

Trimmed Mean

Simple and robust variation of the arithmetic mean:

$$\bar{x} = \frac{x_{([n\alpha]+1)} + \dots x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

where $[n\alpha]$ denotes the greatest index less than or equal to $n\alpha$. In other words:

- Order data.
- Discard lowest and highest $100\alpha\%$.
- Compute arithmetic mean.

Note: 50% trimmed mean corresponds to median.

Central Tendency: Median (I)

The **median** $\tilde{x}$ is the middle value of ordered observations where

- A minimum of 50% of all values are lesser or equal $\tilde{x}$.
- A minimum of 50% of all values are greater or equal $\tilde{x}$.

$$\tilde{x} = \begin{cases} x_{\frac{N+1}{2}} & \text{if } N \bmod 2 \neq 0, \\ \frac{x_{\frac{N}{2}} + x_{\frac{N+1}{2}}}{2} & \text{if } N \bmod 2 = 0. \end{cases}$$

More robust compared to mean as extreme values does not have such drastic affects.

Central Tendency: Median (II)

Friedrich-Alexander-Universität
Technische Fakultät

Median for interval grouped data (An Example)

Class i	Age x_i (x_i^l - x_i^u)	Class Width Δ_i	Absolute Frequency n_i	Relative Frequency f_i	Cumulative rel. Frequency F_i
1	1 — 5	5	200	0.06262	0.06262
2	6 — 15	10	450	0.14089	0.20351
3	16 — 20	5	300	0.09393	0.29743
4	21 — 50	30	1500	0.46963	0.76706
5	51 — 80	30	700	0.21916	0.98622
6	81 — 110	30	44	0.01378	1.00000
$\sum$			3194	1.00000	

The median lies within the fourth group, i.e. in the age group from 21 to 50 years

Central Tendency: Median (III)

Median for interval grouped data

For grouped data we typically have no information about the underlying distribution. However, we can assume that the data is equally distributed. In this case we can approximate the median with the following steps:

- Order data (in our example by age, not by frequencies!).
- Compute relative frequencies, that is: frequency divided by sum of all frequencies.
- Compute cumulative relative frequencies.

Determine median with these considerations:

- $F_i = 0.5$: We have no clear median. Therefore: Take this class ($F_i = 0.5$) and the next one ($F_i > 0.5$) to obtain the median.
- $F_i > 0.5$: Median lies within the class where the cumulative relative frequency exceeds 50% for the first time. Compute the approximate median with:

Central Tendency: Median (IV)

Median for interval grouped data

$F_i > 0.5$: Median lies within the class where the cumulative relative frequency exceeds 50% for the first time. Compute the approximate median with:

$$\tilde{x} \approx x_i^l + \left(\frac{\frac{1}{2}\sum_{i=1}^{N} n_i - \sum_{k=1}^{N-1} n_k}{n_i} \right) \Delta_i$$

where

- i denotes the class number in which the median lies,
- x_i^l is the lower boundary,
- $\sum_{i=1}^{N} n_i$ is the sum of all absolute frequencies,
- $\sum_{k=1}^{N-1} n_k$ is the cumulative sum of absolute frequencies below class i, and
- $\Delta_i = x_i^u - x_i^l$ the class width.

In our example: $\tilde{x} \approx 21 + \left(\frac{\frac{3194}{2} - 950}{1500} \right) * 30 \approx 33.94$, i. e. 33 years and 11 months

Central Tendency: Mode

Mode

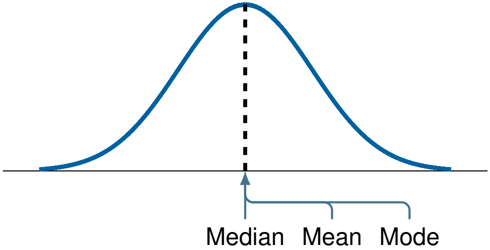
- Value that occurs most frequently within the data set.
- Can be unimodal, bimodal, multimodal.
- Also possible that no mode exists when each value is unique, i. e. occurs only once.
- Empirical formula for unimodal modes:

$$\bar{x} - \text{mode} \approx 3(\bar{x} - \tilde{x}).$$

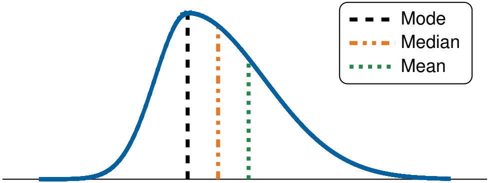
Example of Mode, Median and Mean (I)

Normal distribution

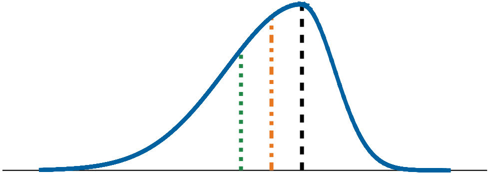
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



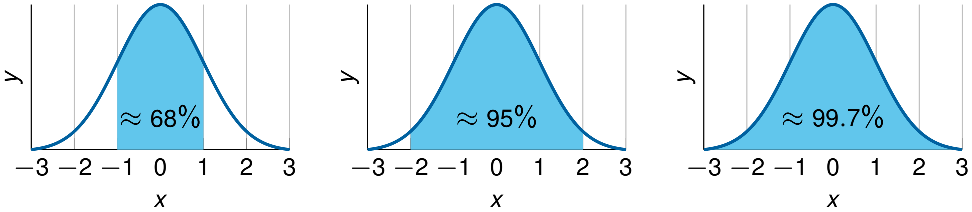
Positively Skewed Data Distribution



Negatively Skewed Data Distribution



- **The normal distribution:**
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68$\%$ of the measurements.
    - μ: mean,
    - σ: standard deviation.
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95$\%$ of the surface under the curve.
  - $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7$\%$ of the surface under the curve.



Measuring the Dispersion of Data (I)

Variance σ^2 and standard deviation σ:

- Empirical sample variance is the mean: $\overline{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Standard deviation is the square root $\sigma = \sqrt{\sigma^2}$.

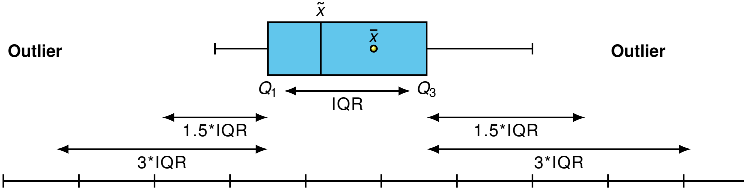
Measuring the Dispersion of Data (II)

- **Range** is the difference between the largest and smallest value.
- **Quartiles:** Also known as *quantiles*. Generalization of the median. Median is the 50$^{\text{th}}$ quartile. Other common quartiles include Q_1 (25$^{\text{th}}$ percentile) and Q_3 (75$^{\text{th}}$ percentile).

  Quartile with order p with $(0 < p < 1)$ have following characteristics:
  - A minimum of $p * 100\%$ of values are lesser or equal to Q_p.
  - A minimum of $(1 - p) * 100\%$ of values are greater or equal Q_p.
- **Inter quartile range:** $\text{IQR} = Q_3 - Q_1$.
- **Five number summary:** minimum, Q_1, median, Q_3, maximum.
- **Outlier**: usually assigned to values higher/lower than $1.5 \cdot \text{IQR}$.

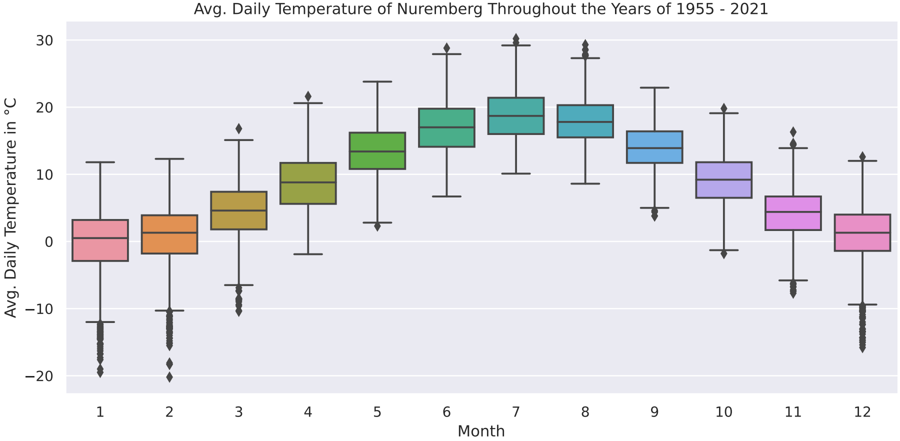
Visualization of choice of these measures: **boxplot**

Boxplot Analysis



- Data is represented with a box vertically or horizontally.
- The ends of the box are at the first and third quartiles, i.e. the width of the box is IQR. The median $\tilde{x}$ is marked by a line within the box.
- Whiskers: two lines outside the box, reaching the minimum and the maximum.
- Outliers: points beyond a specified outlier threshold, plotted individually.

Boxplot Analysis: Example



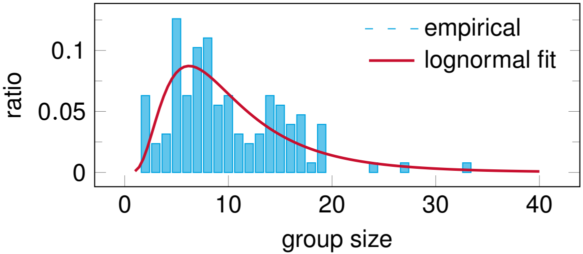
Data courtesy to German Meteorological Service (Deutscher Wetterdienst). `www.dwd.de`

Visualization of Basic Statistical Descriptions

- **Histogram**: x-axis are values, y-axis represent frequencies.
- **Quantile plot**: Each value x_i is paired with some q_i indicating that approximately $q_i \cdot 100\%$ of data are $\leq x_i$.
- **Quantile-quantile (q-q) plot**: Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- **Scatter plot**: Each pair of values is a pair of coordinates and plotted as points in the plane.

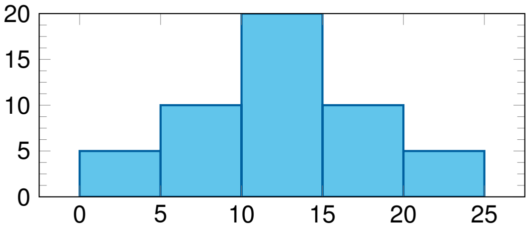
Histogram Analysis

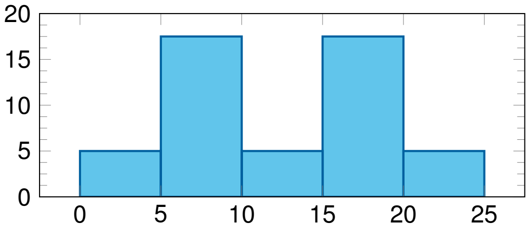
- **Histogram**: Visualization of tabulated frequencies, shown as bars.
- It shows what proportion of cases fall into each of several categories.
- Differs from a **bar chart** in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width.
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent.



Histograms Often Tell More than Boxplots

The two histograms shown below may have the same boxplot representation, thus the same values for min, Q_1, median, Q_3 and for the max. But they have rather different underlying distributions.

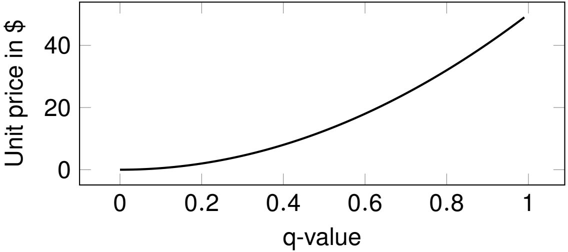




Quantile Plot

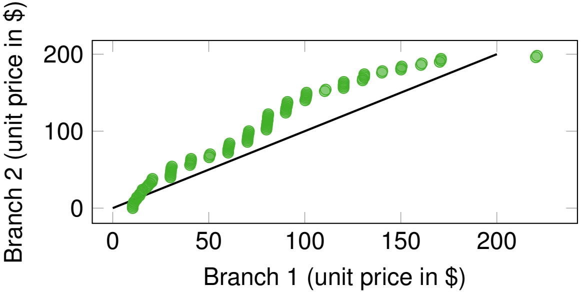
Displays all of the data.
A quantile plot allows the user to assess both the overall behaviour and unusual occurrences.

Plots quantile information.
For some data point x_i, sorted in increasing order, q_i indicates that approximately $q_i \cdot 100\%$ of the data are below or equal to the value of x_i.



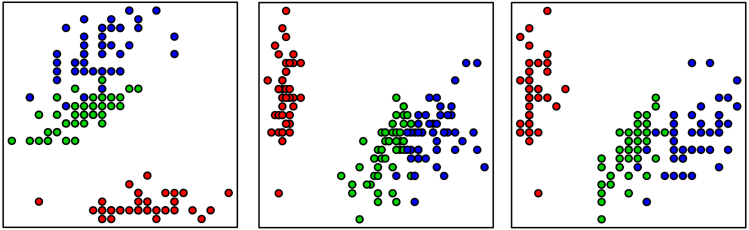
Quantile-quantile ($q - q$) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- View: Do these two distributions differ?
  Example shows unit price of items sold at Branch 1 vs. branch 2 for each quantile. Unit prices of items sold at branch 1 tend to be lower than those at branch 2.



Scatter Plots

Provides a first look at **bivariate data** to see clusters of points, outliers or similar.
Each pair of values is treated as a pair of coordinates and plotted as points in the plane.



Data Profiling

"Data profiling refers to the activity of collecting data about data, i.e., metadata."[3]

Derives metadata such as:

- Data types and value patterns such as most frequent values.
- Completeness and uniqueness of columns.
- Number of null values and distinct values in a column.
- Keys and foreign keys.
- Occasionally functional dependencies and association rules.
- Discovery of inclusion dependencies and conditional functional dependencies.

[3] Z. Abedjan et al., *Data Profiling*. Morgan & Claypool Publishers LLC, Nov. 2018, vol. 10, pp. 1–154. DOI: 10.2200/s00878ed1v01y201810dtm052. [Online]. Available: http://dx.doi.org/10.2200/s00878ed1v01y201810dtm052

Data Visualization

Data Visualization

Why visualize data?

- **Gain insight** into an information space by mapping data into graphical primitives.
- **Provide qualitative overview** of large data sets.
- **Search** for patterns, trends, structure, irregularities, relationships among data.
- **Help find interesting regions and suitable parameters** for further quantitative analysis.
- **Provide a visual proof** of computer representations derived.

Categorization of Visualization Methods

Visualization methods can be categorized into five groups:

1. Pixel Oriented Visualization
2. Geometric Visualization
3. Icon Based Visualization
4. Hierarchical Visualization
5. Complex Data and Relations Visualization

Data Visualization

Pixel Oriented Visualization

Pixel Oriented Visualization

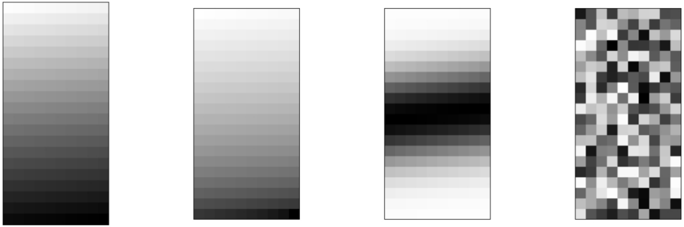
Very simple visualization techniques based on pixels.

General Methods:

- Heat map.
- *Circle segment diagram.*

Heat Map

- For a data set of *m* dimensions create *m* windows on the screen, one for each dimension.
- The values in dimension *m* of a record are mapped to *m* pixels at the corresponding positions.
- The color/intensity of the pixels reflect the corresponding values.



a) Income. b) Credit limit. c) Transaction volume. d) Age.

Friedrich-Alexander-Universität
Technische Fakultät

Data Visualization

Geometric Visualization

Geometric Visualization

Visualization of geometric transformations and projections of data.

General Methods:

- Scatter plot and scatter-plot matrices.
- Polar plot.
- *Parallel coordinates.*

Additional Methods:

- Projection pursuit[4]. Finds a linear projection (one- or two-dimensional) that are “highly revealing”.
- Prosection views[5].
- Hyperslice[6].

---

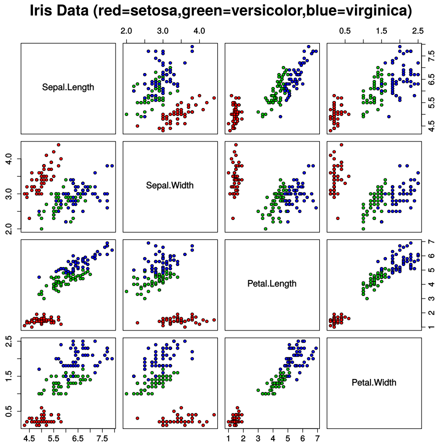
[4] J. Friedman and J. Tukey, “A projection pursuit algorithm for exploratory data analysis,” *IEEE Transactions on Computers*, vol. C-23, no. 9, pp. 881–890, Sep. 1974, ISSN: 0018-9340. DOI: 10.1109/t-c.1974.224051. [Online]. Available: http://dx.doi.org/10.1109/t-c.1974.224051

[5] G. W. Furnas and A. Buja, “Prosection views: Dimensional inference through sections and projections,” *Journal of Computational and Graphical Statistics*, vol. 3, no. 4, p. 323, Dec. 1994, ISSN: 1061-8600. DOI: 10.2307/1390897. [Online]. Available: http://dx.doi.org/10.2307/1390897

[6] J. J. van Wijk and R. van Liere, “Hyperslice - visualization of scalar functions of many variables,” in *4th IEEE Visualization Conference, IEEE Vis 1993, San Jose, CA, USA, October 25-29, 1993, Proceedings*, G. M. Nielson and R. D. Bergeron, Eds., IEEE Computer Society, 1993, pp. 119–125, ISBN: 0-8186-3940-7. DOI: 10.1109/VISUAL.1993.398859. [Online]. Available: https://doi.org/10.1109/VISUAL.1993.398859

Scatter Plot Matrices

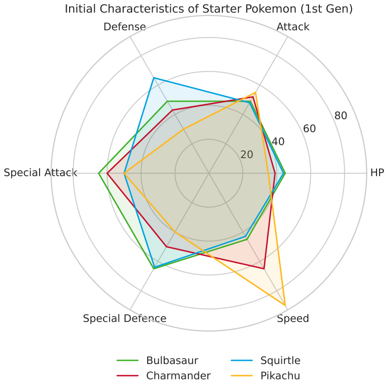
- Compare the values of more than two dimensions at once.
- Each column is drawn against each other column (twice).
- Makes use of scatter plots for each comparison.



Polar Plot

Friedrich-Alexander-Universität
Technische Fakultät

- Shows connections among multiple dimensions for each data record.
- Saves space.
- **Downside:** Can get crowded with too much data records.



Data Visualization

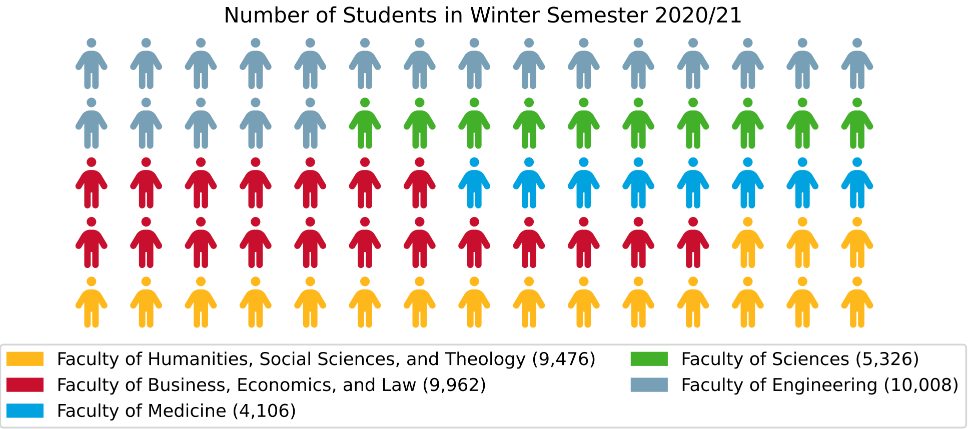
Icon Based Visualization

Icon Based Visualization

Visualization of the data values as features of icons:

- **General methods:**
  - Stick figures.
  - *Chernoff faces.*
- **General techniques:**
  - Shape coding: *Use shape to represent certain information encoding.*
  - Color icons: *Use color icons to encode more information.*
  - Tile bars: *Use small icons to represent the relevant feature vectors in document retrieval.*

Stick Figures



Data Visualization

Hierarchical Visualization

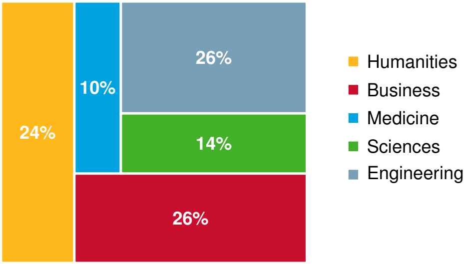
Hierarchical Visualization

Visualization of the data using a hierarchical partitioning into subspaces:

- **General Methods:**
  - Tree maps.
  - *Cone trees.*

Tree Maps

- Uses a hierarchical partitioning of the screen space depending on the importance of the data.
- In this example, there is only one level of hierarchy, but it can be extended to multiple levels (e.g. departments below the faculty level).



Friedrich-Alexander-Universität
Technische Fakultät

Data Visualization

Complex Data and Relations Visualization

Complex Data and Relations Visualization

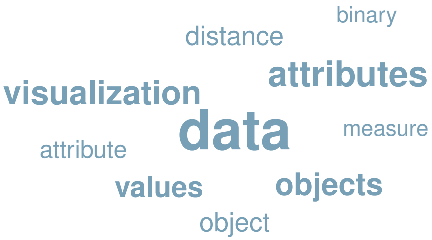
Friedrich-Alexander-Universität
Technische Fakultät

Visualizing non-numerical data:

- **General Methods:**
  - Word cloud.
  - *Networks.*

Word Cloud

- The importance of a word is represented by its size or color.
- In this example, the size of the word is proportional to its frequency in a sample text.



Measuring Data Similarity and Dissimilarity

Motivation

Similarity and Dissimilarity is inherently important for applications like

- **Nearest-Neighbor Classification** (c.f. lecture 7).
  Assign class label to similar objects.
  Example: spam classification or patient diagnosis.
- **Clustering** (c.f. lecture 8).
  Cluster customers based on similar properties such as income, age, area of residence).
  Useful for *marketing* campaigns.
- **Outlier Analysis** (c.f. lecture 9).
  Cluster data objects to identify outliers.

Cluster

A *cluster* subsumes data objects that are similar to each other yet dissimilar to data objects of other clusters.

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are.
- Value is higher when objects are more alike.
- Often chosen within the range of $[0, 1]$.

Dissimilarity

- E.g. distance.
- Numerical measure of how different two data objects are.
- Lower when objects are more alike.
- Minimum dissimilarity is often 0.
- Upper limit varies.

Proximity

Proximity can refer to similarity or dissimilarity.

Data Matrices and Dissimilarity Matrices

Data Matrix

- Stores data objects.
- Also called *object-by-attribute structure*.
- Each data object is described by m attributes.
- Having n data objects we have a $n \times m$ data matrix.

$$\left\{\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{array}\right\}$$

Dissimilarity Matrix

- Stores dissimilarity values of data object pairs.
- Also called *object-by-object structure*.

$$\left\{\begin{array}{ccccc} 0 & 0 & 0 & \cdots & 0 \\ d(x_1, x_2) & 0 & 0 & \cdots & 0 \\ d(x_1, x_3) & d(x_2, x_3) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(x_1, x_m) & d(x_2, x_m) & d(x_3, x_m) & \cdots & 0 \end{array}\right\}$$

with $d(i, j)$ as the measured dissimilarity between two objects i and j.
Similarity expressed with $\text{sim}(i, j) = 1 - d(i, j)$.

Proximity Measures for Nominal Attributes

Friedrich-Alexander-Universität
Technische Fakultät

Recall: Nominal attributes can take two or more states and values do not follow an order.

- Values can be the same (distance of 0) or different (distance of 1).
- More options for sets of nominal attributes (variables).
  1. **Simple Matching Coefficient** (SMC).

$$\text{SMC} = \frac{\#\text{of matching attributes}}{\#\text{number of attributes}}$$

  2. **One-Hot Encoding.**
     Convert nominal attributes to binary attributes,
     i.e. create one binary attribute for every unique nominal value.

Proximity Measure for Ordinal Attributes

Perform **Integer Encoding**, where every unique value is assigned an integer value.

Proximity Measure for Binary Attributes

- Contingency table for binary data that counts matches.

$$
\underbrace{\text{data object } i \left\{ \begin{array}{c|ccc}
 & 1 & 0 & \sum \\
\hline
1 & q & r & q+r \\
0 & s & t & s+t \\
\sum & q+s & r+t & q+r+s+t
\end{array} \right.}_{\text{data object } j}
$$

- Distance measure for *symmetrical* binary variables: $d(x, y) = \frac{r+s}{q+r+s+t}$
- Distance measure for *asymmetrical* binary variables: $d(x, y) = \frac{r+s}{q+r+s}$
- Similarity for asymmetrical binary values is called *Jaccard* coefficient and calculated as follows:

$$\mathrm{sim}(i, j) = \mathrm{jaccard}(i, j) = 1 - d(i, j) = \frac{q}{q+r+s}$$

Dissimilarity Between Binary Variables

Name	Sex	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Bob	M	Y	N	P	N	N	N
Alice	F	Y	N	P	N	P	N
Charlie	M	Y	P	N	N	N	N

Note

Attribute "Sex" is a symmetrical attribute (all values are of equal importance), whereas all remaining attributes are asymmetrical binary attributes.

Let values Y and P be equal to 1 and the value of N be 0, then we can compute:

$$d(\text{Bob}, \text{Alice}) = \frac{0+1}{2+0+1} \approx 0.33 \qquad d(\text{Charlie}, \text{Alice}) = \frac{1+2}{1+1+2} = 0.75$$

$$d(\text{Bob}, \text{Charlie}) = \frac{1+1}{1+1+1} \approx 0.67$$

Standardizing Numerical Data

Friedrich-Alexander-Universität
Technische Fakultät

- z**-Score:**

$$z = \frac{x - \mu}{\sigma}.$$

- x is the score to be standardized; μ is the population mean; σ is the standard deviation.
- The distance between the raw score and the population mean in units of the standard deviation.
- Negative when the raw score is below the mean, positive else.

Example: Data Matrix and Dissimilarity Matrix

- Data matrix:

Point	Attribute 1	Attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

- Dissimilarity matrix (with Euclidean distance):

	x_1	x_2	x_3	x_4
x_1	0			
x_2	$3,61$	0		
x_3	$2,24$	$5,1$	0	
x_4	$4,24$	1	$5,39$	0

Distance on Numerical Data

Friedrich-Alexander-Universität
Technische Fakultät

$$d(x, y) = \sqrt[h]{\sum_{i=1}^{n} |x_i - y_i|^h}$$

where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two n-dimensional data objects.
In fact, this distance induces a norm over real vector space, called L_n-norm.

Properties of a L_n-norm

- $d(x, y) \geq 0$, positive definiteness.
- $d(x, y) = d(y, x)$, symmetry.
- $d(x, y) \leq d(x, z) + d(z, y)$, triangle inequality.

A distance satisfying these properties is called **metric**.

Special Cases of Minkowski Distance

Friedrich-Alexander-Universität
Technische Fakultät

$h = 1$: **Manhattan Distance**

- Also known as city block, or L_1-norm.
- E.g. Hamming distance: number of bits that differ in two binary vectors.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

$h = 2$: **Euclidean**

- Also known as L_2-norm.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

$h \to \infty$: **Supremum**

- Also known as $L_{\max}$-norm, or L_∞-norm.
- Maximum difference between any attribute of two vectors.

$$d(x, y) = \lim_{h \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^h \right)^{\frac{1}{h}} = \max_i |x_i - y_i|.$$

Summary

Summary

- **Data attribute types:**
  Nominal, binary, ordinal, interval-scaled or ratio-scaled.
- **Many types of data sets:**
  E.g. numerical, text, graph, web, image.
- **Gain insight into the data by:**
  - Basic statistical data description: *Central tendency, dispersion and graphical display.*
  - Data visualization: *Map data onto graphical primitives.*
  - Measure data similarity.
- **Above steps are the beginning of data preprocessing.**
- **Many methods have been developed but still an active area of research.**

Any questions about this chapter?

Ask them now or ask them later in our forum:

https://www.studon.fau.de/studon/goto.php?target=lcode_OLYeD79h