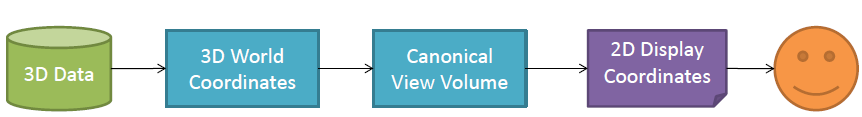
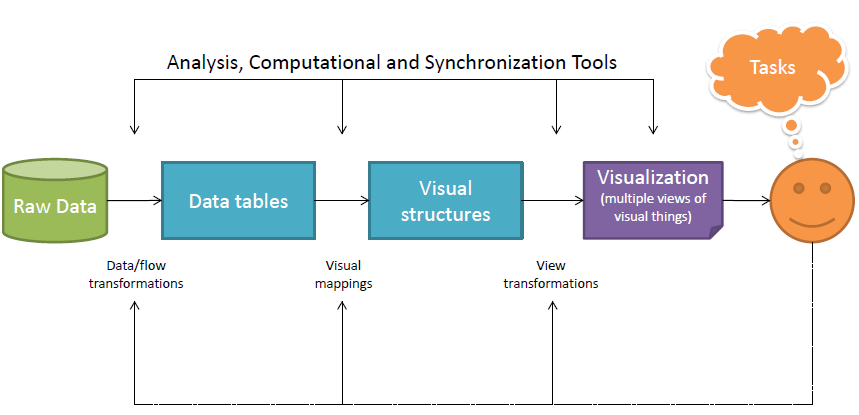
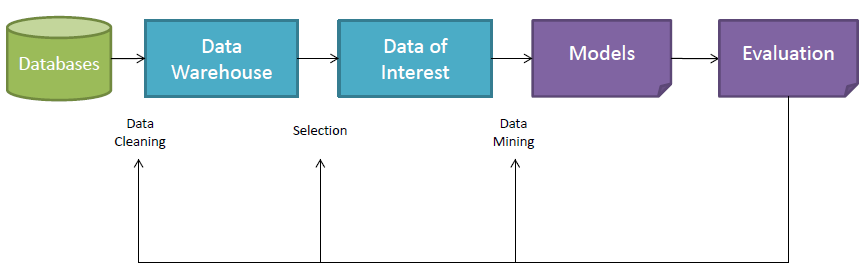
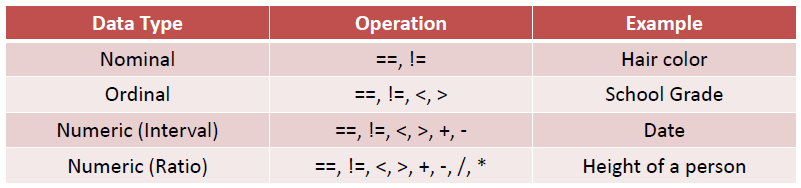
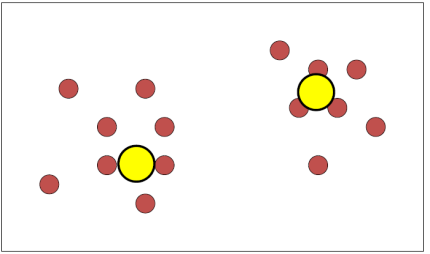
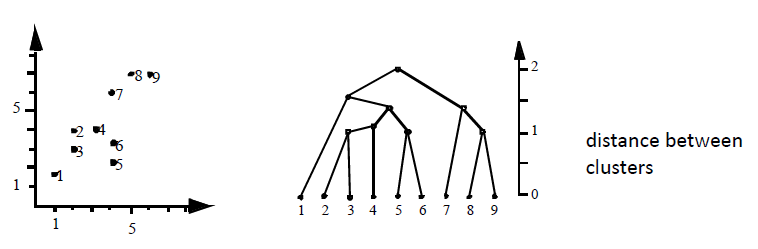
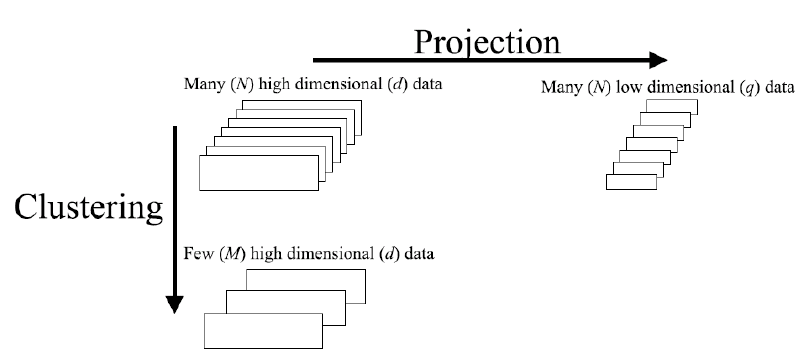
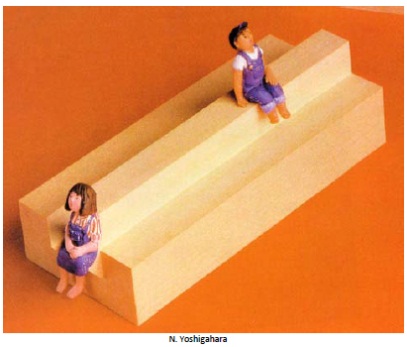
1. **Introduction**
   1. **The Visualization Process**

* Analysis of: data type and the information the viewer hopes to extract
* Preprocess the data
* Define a mapping
* Provide interactive controls (if necessary)
* Visualization as port of a larger process:
  + Exploratory data analysis
  + Knowledge discovery
  + Visual analytics
* Goal: Building a model
* Process (data -> image/visualization/model) is called pipeline
  1. **Computer Graphics Pipeline**
  2. **The Visualization Pipeline**
  3. **Knowledge Discovery Pipeline**
  4. **The Role of the User**
* **Presentation:**
  + Starting point: presented facts are a priori
  + Process: choice of appropriate presentation techniques
  + Result: high-quality visualization of data to present facts
* **Confirmatory Analysis**
  + Starting point: hypothesis about the data
  + Process: goal-oriented examination of the hypothesis
  + Result: visualization of data to confirm or reject the hypothesis
* **Exploratory Analysis:**
  + Starting point: no hypothesis about the data
  + Process: interactive, usually undirected search for structures, trends
  + Result: visualization of data to lead to hypothesis about the data

1. **Data Foundations**
   1. **Types of Data**

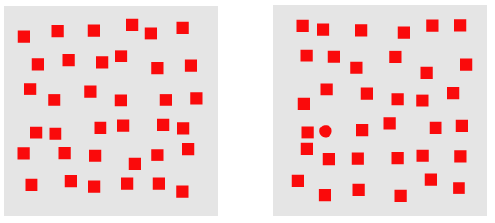
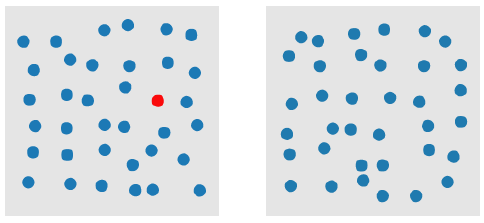
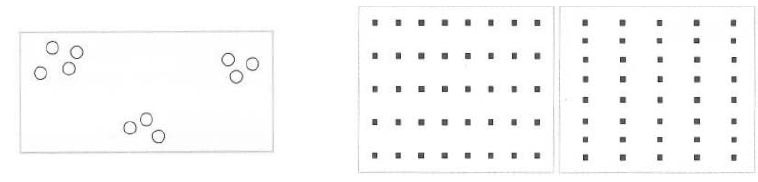
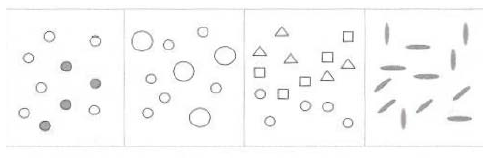
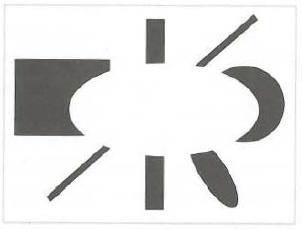
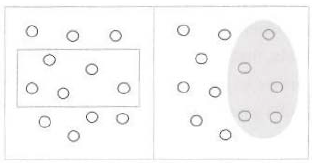
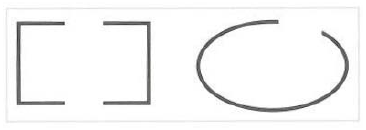
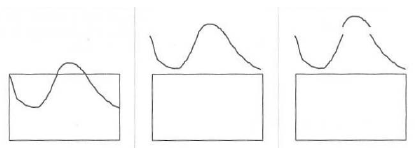
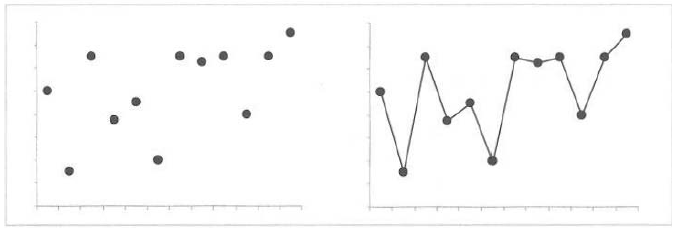


* 1. **Data Preprocessing**
* Metadata can help interpreting (format, unit, …)
* Statistical analysis can provide useful information (outliers, clusters, …)
  + 1. ***Missing Values and Data Cleansing***
* **Missing value**  is a variable not in the data set, but existing in the real world
* **Empty value**  is a variable in the data set without a value in the real world
* Ignore the tuple
* Fill in the missing value manually
* Use global constant or attribute mean to fill
* Use most probable value to fill (determined with regression, interpolation, …)
  + 1. ***Normalization***
* Very few outstanding values out of the data set
* Could have huge influence of e.g. a heat-map color mapping
* Linear Mapping:
* Square root mapping:
* Logarithmic mapping is similar
  + 1. ***Segmentation***
* **Given:** Data set with N d-dimensional data items
* **Task:**determine natural partitioning of the data set into a numbers of clusters (k) and noise
* **Manual Segmentation:**
  + Based upon Attribute values/ranges and topological properties
* **Automatic Segmentation = Clustering Algorithms**
  + K-means
  + Linkage-based methods
  + Kernel density estimation
    1. ***Sampling and Subsetting***
* **Motivation:** data set is much larger than possible to work on
* **Example:** voters of an election (use an representative sample)
* **Important:** subset must represent some well-defined characteristics of whole data set
* **Types:**
  + Non-probabilistic samples: sample on random-basis (volunteers, …)
  + Probabilistic samples: sample on random-basis, but so that every element has equal chance to being selected
    - Simple random sampling: is least biased method
    - Systematic random sampling: elements are numbered 1 to N in some order -> numbers randomly chosen
    - Stratified Random sampling: data set divided into non-overlapping subsets called *strata*, subsets are random
    - Cluster random sampling: sample consists of randomly chosen groups of neighboring elements (clusters)
    1. ***Approximation and Interpolation***
* **Approximation:** 
  + Problem: spatially random distributed weather stations
  + Temperature data approximation based on triangulation
  + Regression (linear, quadratic, …)
    - Linear: tries to discover straight line equation, that best fits the data point (y = a +bx), minimizes the least square error
* Interpolation:
  + Polynomial (Lagrange basis, Newton form)
  + Piecewise polynomial (cubic splines, …)
    - Passing single polynomial through many data points can lead to oscillations in the interpolant
    - Interpolant is cubic polynom
  + Orthogonal polynomials (Legendre, …)
  + Trigonometric functions
    1. ***Dimension Reduction:***
* **Major problems**:
  + Large number of features represent an object
  + Data difficult to visualize, especially if some features not characteristic
  + Irrelevant features my cause reduction of algorithm accuracy
* **Idea: Projection =** Identify most important features
  + Simplifies processing without quality loss
  + Directly visualizes two/three most important features
* **Goal:**
  + Discover hidden factors that explain the data
  + Reduce dimensionality of the data
* Similar to cluster centroids
* **PCA:**
  + There are n observations
  + Projections are called
  + Projection is linear:
  + Assume zero mean, we want to find the **W**  which:
    - Decorrelates the projected points **u**
    - Preserves most values of the variance in the data
    - Minimizes reconstruction error
    - Arbeitsannahme: “*Die Richtungen mit der größten Streuung (Varianz) beinhalten die meiste Information.*”
    1. ***Mapping Nominal Dimensions to Numbers***
* Find mapping which not introduces artificial relationships that not exists
* Low number of different values (color, shape, …)
* Use multi-dimensional scaling (MDS) to map different nominal values to positions
* Only one nominal attribute: label the graphical elements
  + 1. ***Aggregation and Summarization***
* **Count** the items in the data set
* **Sum** the items in a list
* **Average (avg)** of all items in a data set
* **Measurement and Error:**
  + Random + systematic error + the true value gives the observation result
  + Only random error (noise) does not affect average
  + Only systematic error (bias) affect the average
    1. ***Smoothing and Filtering***
* Smooth & filter data to reduce noise and to blur sharp discontinuities
* Convolution: values that are significantly different from their neighbors will be modified to be more similar
* Binning
* Smoothing Noisy Data
  + Noise: random error or variance in a measured attribute
  + Causes:
    - Faulty data collection instruments
    - Data entry problems
    - Data transmission problems
    - Technology limitation
    - Inconsistency in naming convention
  + Binning:
    - Sort data and partition into (equi-depth) bins
    - Smooth by bin means, bin median, bin boundaries, etc.
  + Regression:
    - Smooth by fitting a regression function
  + Clustering:
    - Detect and remove outliers
  + Combined computer and human inspection:
    - Detect suspicious values and check by human
    1. ***Raster to Vector Conversion***
* Why?
  + Compressing the contents for transmission
  + Comparing the contents of two or more images
  + Transforming and/or segmenting the data
* How?
  + Tresholding: Identify values to break data into regions – boundaries can be traced to generate edges and vertices
  + Region-growing: merge pixels into clusters if they are sufficiently similar
  + Boundary-detection: convolve the image with particular pattern matrix
  + Thinning: Reduce wide linear features, such as arteries, to a single pixel width
    1. ***Summary of data preprocessing***
* Preprocessing can improve the effectiveness of the visualization.
* Convey to the user that these processes have been applied to the data.
* This helps interpreting the results.
* Otherwise, misinterpretation or erroneous conclusions can be drawn from the data.

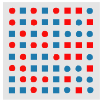
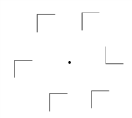
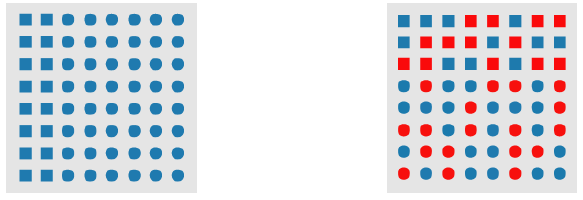
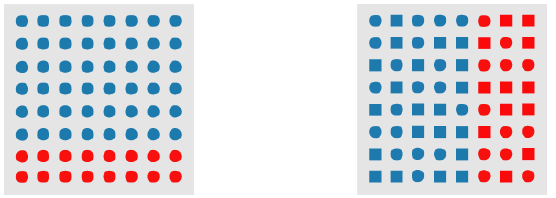
1. **Human Perception and Information Processing**
   1. **Definitions**

* **Perception:** Process of organizing sensory data, deriving (dt.: ableiten) structure from the complex pattern of energy impinging on our sensory receptors.
* **Cognition:** Is the act or process of knowing including both awareness and judgment; also: a product if this act
  1. **Physiology**
* **Sensory of vision:** involves the gathering and recording of light from objects in the surrounding scene, and the forming of a two dimensional function on the photoreceptors.
  1. **Perceptual Processing**



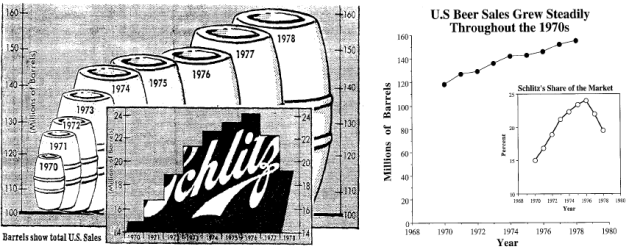
* **Classic Model** of the flow of sensory data for cognition:
  + 1. ***Preattentive Processing***
* Perception of visual features managed by the low-level visual machinery
* Extremely fast: < 200 msec (eyes take more than that time) -> proceed parallel
* Preattentive = before attention takes place? -> Attention plays a role!!
* **Gestalt laws**
  + Perceptual laws about how we **group visual objects** together to form visual entities
  + Law of **Proximity**
  + Law of **Similarity**
  + Law of **Enclosure**
  + Law of **Closure**
  + Law of **Continuity**
  + Law of **Connection**
    1. ***Theories of Preattentive Processing***

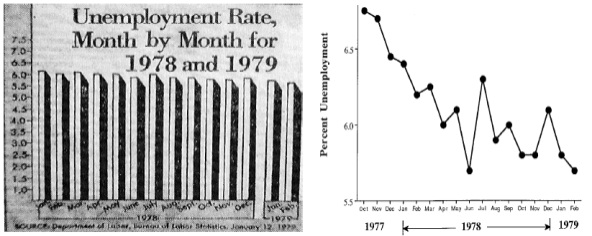
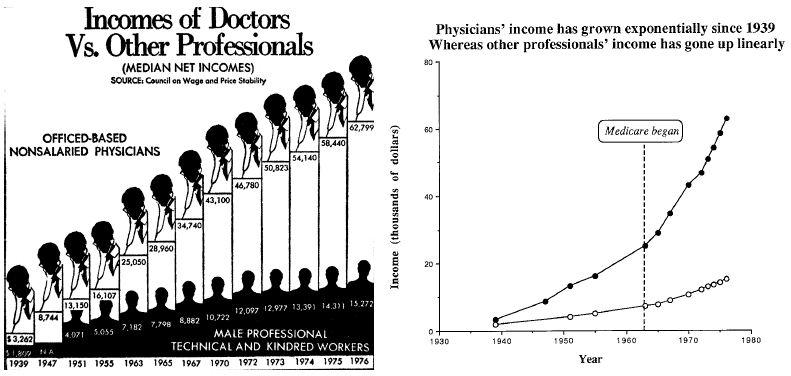
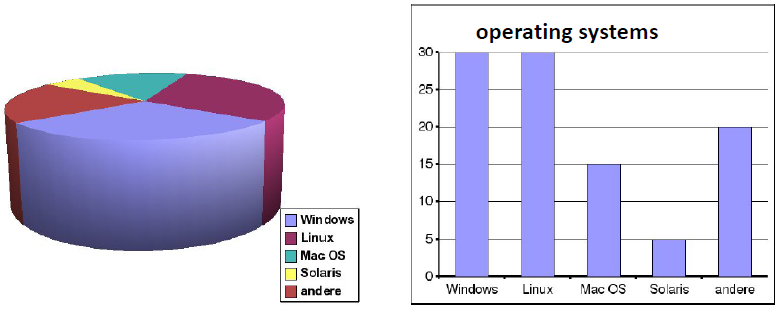


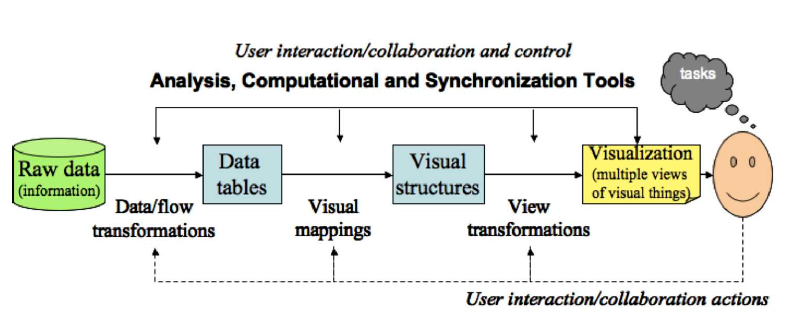
* Feature Integration Theory
  + Boundary defined by unique feature hue is preattentively classified as horizontal
  + Boundary defined by conjunction of features cannot be preattentively classified as vertical
* Similarity Theory
  + High nontarget-nontarget (N-N) similarity allows easy detection of target L
  + Low N-N similarity increases the difficulty of detecting the target L
    1. ***Feature Hierarchy***
* **Horizontal** hue boundary is preattentively identified when form is held constant
* **Vertical** hue boundary is preattentively identified when form varies randomly in the background
* **Vertical** form boundary is preattentively identified when hue is held constant
* Horizontal form boundary cannot be preattentively identified when hue varies randomly in the background
  1. **Perception in Visualization**
* Guidelines for **Color:**
  + Do not over- or underestimate the power of color
  + Always provide a color legend
  + Use color with extreme care and parsimony (Tufte: “above all do no harm”)
  + Learn to love grays and gray scales (grids!)
  + Do not represent unordered data with ordered colors
  + Keep an eye to skewed distribution
  + Do not use the (infamous) rainbow color scale
* **Texture:**
  + Often viewed as a single visual feature
  + Perceptual dimensions: regularity, directionality, contrast, size, etc.
  + Use to represent multiple data attributes
  1. **Metrics – Implications for visualizations**

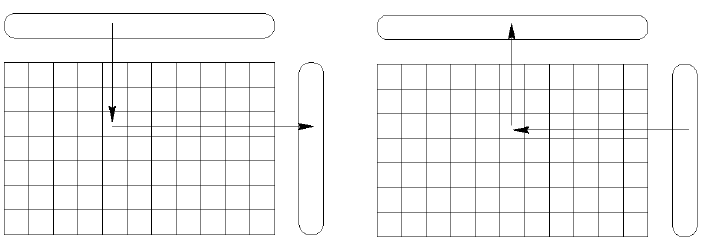
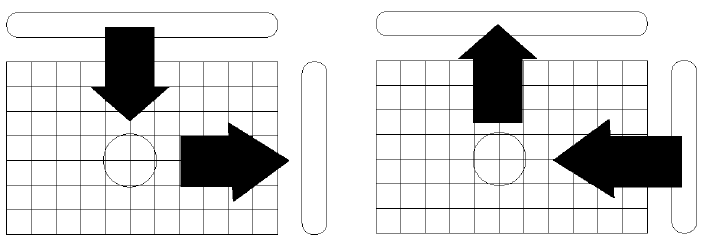
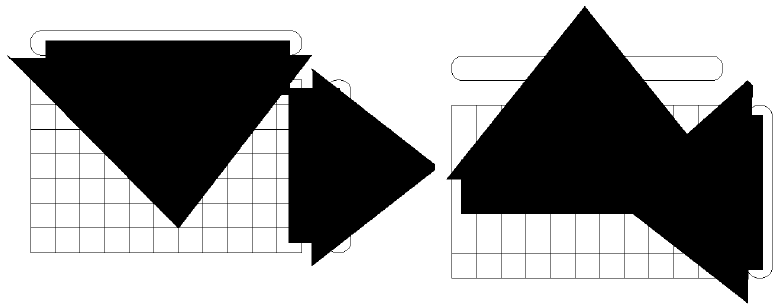
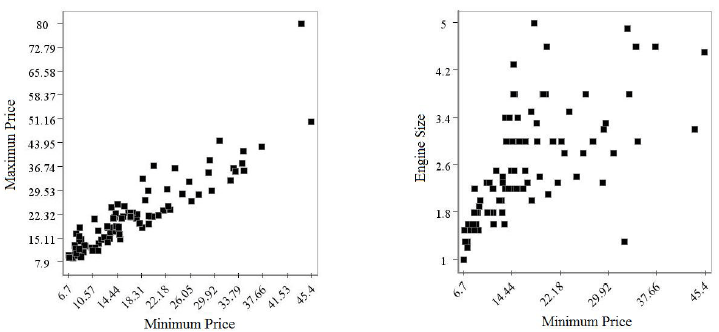
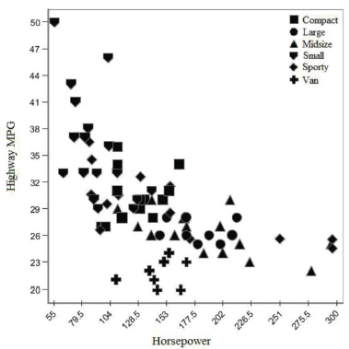
To enhance our absolute judgment

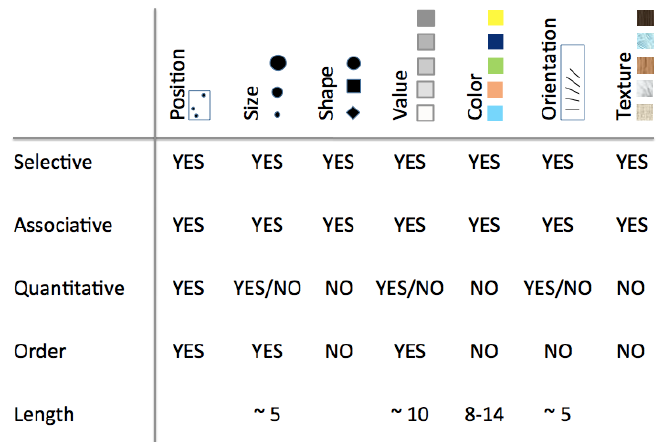
* Reduce graphical representation with one attribute to **4-7 values**
* Or repose problem in **multiple dimension**
* Or reduce problem to **sequence** of small problems
* Or **focus** first on relative judgment, then refine with absolute judgment

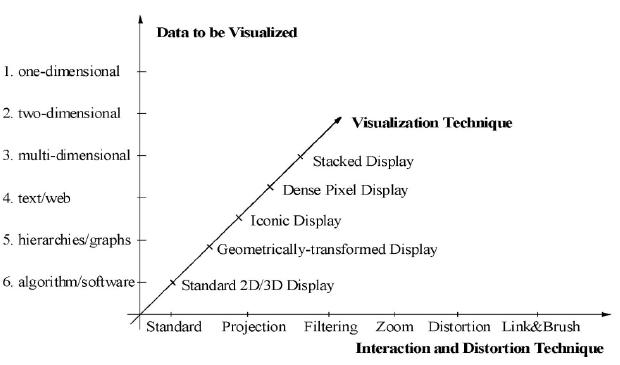
1. **Visualization Foundations**
   1. **Bad Visualizations**

* Use the effect of cubing and get a lie factor of over 131%
* Display Data out of context (hiding effect by careful choice of scale and origin)
* Change scales in mid-axis to make exponential growth linear
* Make clever use of 3D-effects: difficult to compare sizes of objects
  1. **The Visualization process in detail**



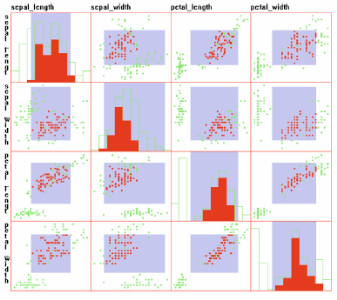
* **Expressiveness:** Visualization presents all the information and **only** the information.
* **Important:** Expressing additional information is potentially dangerous because it may not be correct.
* **Effectiveness:** Visualization is effective when it can be interpreted accurately or quickly and when it can be rendered in a cost-effective manner.
  + 1. ***Levels of information***
* **Level 1:** *Elementary level of information*
  + Relation between spate objects
  + Question: How many objects of this category exist?
  + Exists even in bad graphics
  + Human memory cannot store the multiplicity of elementary information
  + **Important:** number of elementary information must be reduced, similar elements must be discovered and combined to groups and classes
* **Level 2:** *Middle level of information*
  + Relation between groups/classes
  + Question: Which factors are crucial (entscheidend) ?
  + Analyzes the relationships within a group
* **Level 3:** *Upper level of information (overall information)*
  + Relationships between sets of objects
  + Questions: Which different sets do arise by the totality of all factors?
  + **Important:** The upper level of information is required for decisions!
  1. **Semiology of graphical symbols**
     1. ***Symbols and Visualizations***
* Without external (cognitive) identification a graphic is unusable
* External identification must be directly readable und understandable
* Meaningful images must have easily interpretable x-, y- and z-dimensions
* Graphics elements of image must be clear
* Similarity in data structures ↔ visual similarity of corresponding symbols
* Order between data items ↔ visual order between corresponding symbols
  + 1. ***Features of graphics***
* Aim of graphic is to discover groups of orders in x, and groups of orders in y, that are formed on z-values
* (x, y, z)-construction enables in all cases the discovery of these groups
* Within (x, y, z)-construction, permutations and classifications solve the problem of upper level of information
* Every graphic with more than three factors that differs from the (x, y, z)-construction destroys the unity of graphic and the upper level information
* Pictures must be read and understood by human
  + 1. ***The Eight Visual Variables***
* **(1) Position:** 
  + Unlike left, there does not appear to be a strong relationship right between the two variables
* **(2) Mark:**
  + Using shapes to distinguish between different types (e.g. of cars) to compare common characteristics (horsepower, MPG)
* **(3) Size (Length, Area and Volume)**
* **(4) Brightness**
* **(5) Color**
* **(6) Orientation** (to adjust mark orientation)
* **(7) Texture**
* **(8) Motion**
  + can be associated with any of th other visual variables
  + common use: varying speed at which a change is occuring, direction
* **Effects of visual variables**

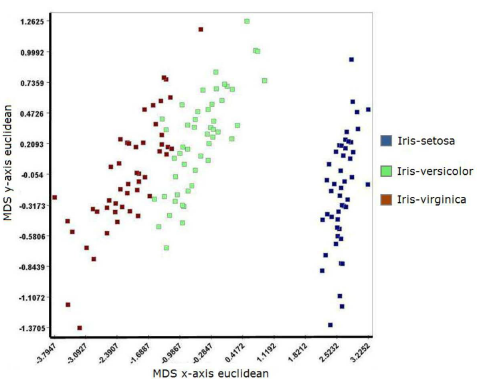


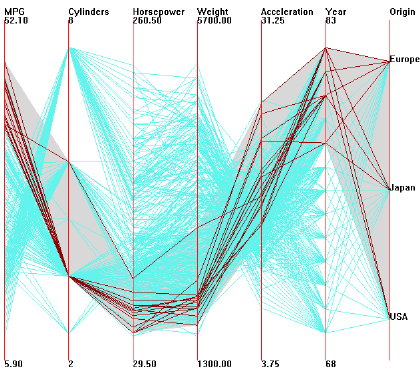
* 1. **Taxonomies (Klassifikationen)**
* Provides structure and understanding relationships in the large number of visualization techniques
* Reveal gaps (zeigt Lücken)
* Help understanding & to design systems
  + 1. ***Taxonomy of visualization goals (Keller & Keller)***
* **Data types:**
  + Scalar (or scalar field)
  + Nominal
  + Direction (or direction field)
  + Shape
  + Position
  + Spatially extended region or object (SERO)
* **Tasks:**
  + *Identify* – establish characteristics by which an object is recognizable
  + *Locate* – ascertain the position (absolute or relative)
  + *Distinguish* – recognize as distinct or different (identification is not needed)
  + *Categorize* – place into division or classes
  + *Cluster* – group similar objects
  + *Rank* – assign an order or position relative to other object
  + *Compare* – notice similarities and differences
  + *Associate* – link or join in a relationship that may or may not be of the same type
  + *Correlate* – establish a direct connection, such as casual or reciprocal
    1. ***Data Type by task taxonomy (Shneiderman)***
* **Data types:**
  + One-dimensional linear
  + Two-dimensional map
  + Three-dimensional world
  + Temporal
  + Multidimensional
  + Tree
  + Network
* **Tasks:**
  + Overview
  + Zoom
  + Filter
  + Details-on-demand
  + Relate
  + History
  + Extract
    1. ***Keim’s Taxonomy (2002)***
* **Classification Criteria:**
  + Data type to be visualized
    - Dimensionality (1D, 2D, Multidimensional)
    - Complex data types (Text/Web, graphs, etc.)
  + Visualization techniques
    - Support exploration of large data sets
    - Standard 2D/3D display, iconic display, etc.
  + Interaction and distortion techniques
    - User interacts with the data
    - Standard, projection, filtering, etc.
* **Data to be visualized:**
  + One-dimensional data:
    - Termporal data (i.e. news data, stock prices), text documents, …
  + Two-dimensional data:
    - Geographical maps, charts, floorplans, newspaper, layouts, …
  + Multidimensional data:
    - Relational tablets, …
  + Text and hypertext
    - News articles, web documents, …
  + Hirarchies/graphs:
    - Telephone calls, web documents, …
  + Algorithm/software:
    - Debugging operations, …

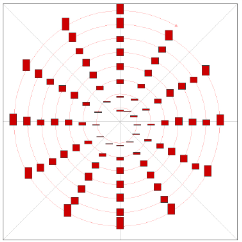
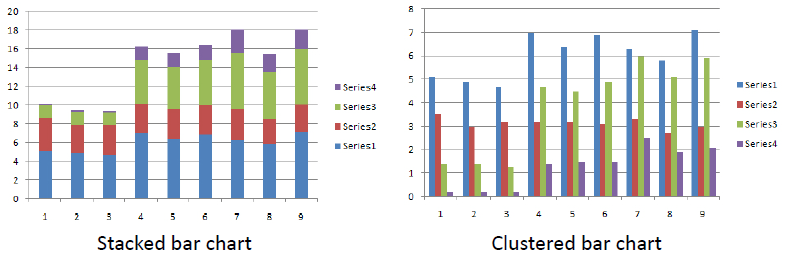
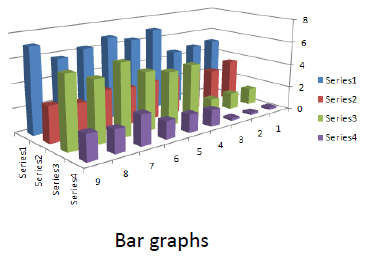
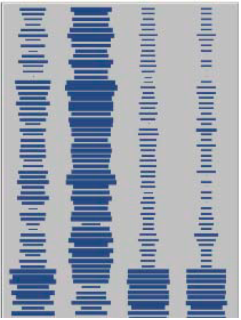
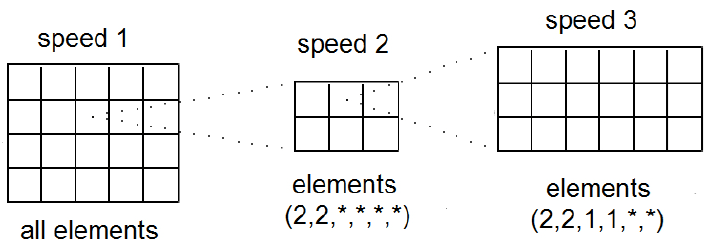
1. **Visualization Techniques or Multivariate Data**
   1. **Point-Based Techniques**

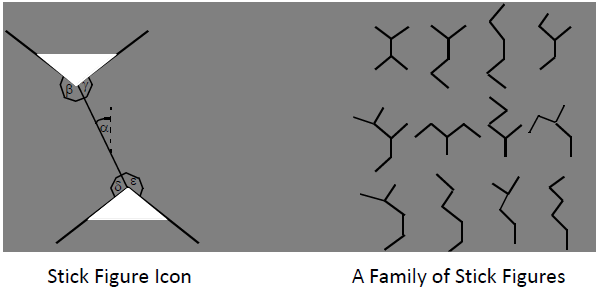
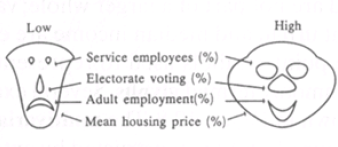
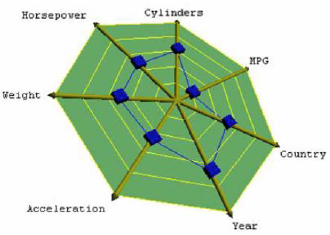
* Project records from n-dimensional data space to an arbitrary k-dimensional display space
* Each record is represented by a visual mark
* Can be structured by various projection techniques

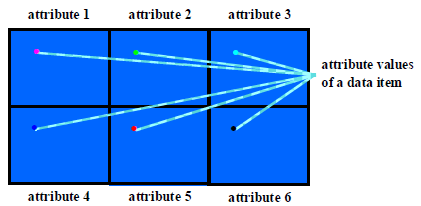


* + 1. ***Scatterplots and Scatterplot Matrices***
* Scatterplot matrix with diagonal plot showing a histogram of each dimension. (red points and histogram regions indicate selected data)
  + 1. ***Force-Based Methods***
* Maintain the N-dimensional features and chracteristics of the data through the projection process
* Difficult when number of dimension increases
* Unintentional artficats in this visualization but not in the data
* Multidimensional scaling (MDS)
  1. **Line-Based Techniques**
* Points are linked toghether with straight or curved lines
* Lines reinforce the relationships among data values
* Convey perceivable features of the data via slope, curvature, crossing, …
  + 1. ***Parallel Coordinates***

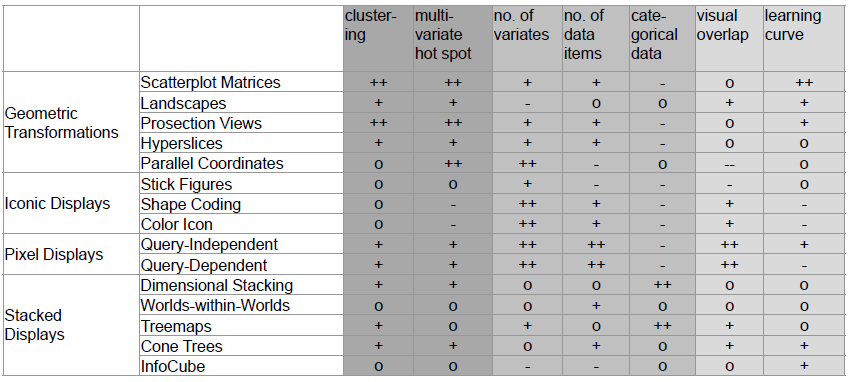


* N equidistant axes which are parallel to one of the screen axes and correspond to the attributes
* Axes are scaled to the [min, max]-range of the corresponding attribute
* Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute
  + 1. ***Radial Axis Techniques***
* **Circular line graphs:** the plotted lines are offset from a circular base
* **Polar graphs:** point plots using polar coordinates
* **Circular bar charts:** like (1) but plotting bars on the base line
* **Circular area graphs:** like (1), but area under line filled with a color ort texture
* **Circular bar graphs:** bars that are circular arcs + common center point and base line
  1. **Region-Based Techniques**
* Filled polygons are used to convey values (size, shape, color, …)
* Mostly not showing the real data, but summaries or distributions of the values
* One of the most common: bar chart
  + 1. ***Bar Charts/Histograms***
    2. ***Tabular Displays***
* Multivariate data is often stored in tables
* Visualization modeled on this tabular structure
* Color or size/length used to encode data value
* Example shows iris data set
  + 1. ***Dimensional Stacking***
* Mapping data from discrete N-dimensional space to two-dimensional image
* Start with data of dimension 2N+1
* Select finite cardinality for each dimension
* Choose one of the dimensions to be the dependent variable
* The rest will be considered independent
  1. **Combination of Techniques**
     1. ***Glyphs and Icons***
* A glyph is a visual representation of a piece of data or information where a graphical entity and its attributes are controlled by one or more data attributes.
* Stick Figures, Chernoff-Faces & Star Glyphs:



* + 1. ***Dense Pixel Displays***
* Basic Idea:
  + Each attribute value is represented by one colored pixel (the vlaue ranges of the attributes are ampped to a fixed color map)
  + Attribute values for each attribute are presented in separate subwindows
  1. **Comparison of the Techniques**
     1. ***Comparison***

Comparison based on the suitability for certain:

* **task characteristics:** clustering, multi variate hot spots, …
* **data characteristics:** no. of variates, no. of data items, categorical data, …
* **visualization characteristics:** visual overlap, learning curve, …
  + 1. ***Hybrid Approaches***
* **Basic Idea:**
  + Integrated use of multiple techniques in one or multiple windows to enhance the expressiveness of the visualizations.
  + Linking diverse visualizations techniques may provide additional information.
  + Virtually all visualizations techniques are combined with dynamics and interactivity
* **Guidelines for Using Multiple Views:**
  + *Rule of Diversity:*Use multiple views when there is a diversity of attributes, models, user profiles, level of abstraction or genres.
  + *Rule of Complementary:* Use multiple views when different views bring out correlations and/or disparities.
  + *Rule of Decompostion:* Partition complex data into multiple views to create manageable chunks and to provide insight into the interaction among different dimensions.
  + *Rule of Parsimony:* Use multiple views minimally.

Reasoning:

* + - Single view: stable context
    - Multiple views: additional complexity for user
  + *Rule of Space/Time Resource Optimization:* Balance the spacial and temporal costs of presenting multiple views with the spacial and temproal benefits of using the views.
  + *Rule of Self-Evidence:* Use perceptual cues to make relationships among multiple views more apparent to the user.
  + *Rule of Consistency:* Make the interfaces for multiple views consistent and make the states of multiple views consistent.
  + *Rule of Attention Management:* Use perceptual technqiues to focus und the user’s attention on the right view at the right time.