Perception and visualization. Preprocessing.

Lecture 2

732A98

Human perception

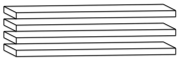
- How are visualizations perceived by different humans?
- How do we know that a given visualization is correctly interpreted?

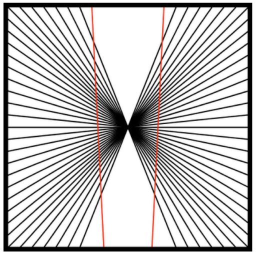
Perception:

- Recognizing
- Organizing (gathering, storing)
- Interpreting (binding to knowledge)

Illusions

- Human perceptual system is not perfect





Perception mechanism

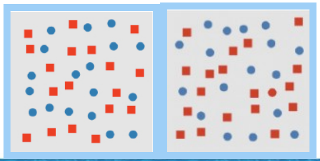
- Preattentive
  - Fast (250 ms)
  - Performed in parallel
- Attentive
  - Slow
  - Uses short term memory
  - Transforms simple visual features into structured objects
  - Compares to memory models (ex. door)

Preattentive processing

- Preattentive feature= shape
- How quickly do you see a red circle?

Preattentive processing

- **Important:** Combination (conjuction) of nonunique features can not be detected preattentively
  - Many red objects
  - Many circle objects
- How quickly can you find a unique object here?



Preattentive features

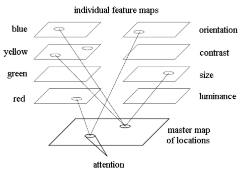
- Length
- Width
- Size
- Curvature (shape)
- Hue
- Intensity
- Flicker
- Direction of motion
- 3D depth
- Lighting direction

Preattentive visual tasks

- Presense or absense of object with a unique visual feature among distractors is detected preattentively
- Boundary between two groups of elements with the same visual feature is detected preattentively
- Movement of an object with a unique visual feature is tracked preattentively
- Amount of elements with a unique visual feature is estimated preattentively

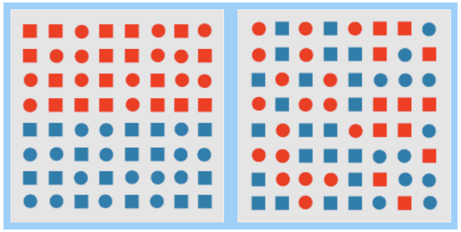
Treisman's theory of preattentive processing

- A figure is processed in parallel by checking individual feature maps
- A specific preattentive task is performed in each feature map
- Conjuction of features requires searial search between maps - takes time



Treisman's theory of preattentive processing

- How quickly can you identify a boundary?



Metrics

- What graphical features can be accurately perceived by humans?
- How many distinct entities can be visualized without confusion?
- How should we use color?
- How should we combine features in a complex phenomenon?

Channel capacity: how many different levels of a feature we can perceive

- 8 levels = 3 bits

Metrics

- Position on a line: 10-15 levels (3.25 bits)
- Size of squares: 4-5 levels (2.2 bits)
- Color: hue 10 levels, brightness: 5 levels (3.1 bits, 2.1 bits)
- Line length: 2.8 bits
- Line orientation: 3 bits
- Line curvature: 1.6-2.2 bits

Summary: 6-7 unique values max.

Metrics

Note: Combining metrics does not sum up the capacity!...

- Hue and saturation: 3.6 bits
- Size, brightness and hue: 4.1 bits
- Position in a square: 4.6 bits

Metrics

Relative judgement: comparing two values of a feature

Errors (in increasing order)

- Position along a common scale
- Length
- Angle
- Area
- Volume
- Color hue

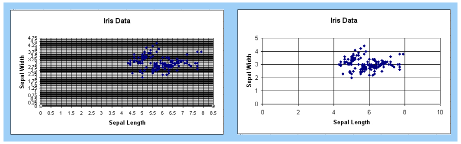
–> *Pie charts* are less effective than *Bar Charts*

Principles of good visualization

- Use intuitive mapping to aesthetics
  - Visualization type is adopted to user's background
  - Geographical coordinates –> X,Y, temperature–>color
  - Use correct mapping
    - Ordinal variables- X,Y, saturation, orientation
    - Nominal variables - shape, texture, hue
- Support view modifications
  - Scrolling, zooming -scales
  - Color map -Level of details
  - Mapping aesthetics

Principles of good visualization

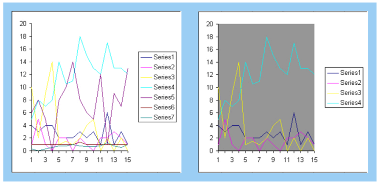
- Do not put too much information in the display (occlusion)
- Add keys, labels, legends, grids with care
- Use display efficiently (0%-100% scale vs actual domain)



Principles of good visualization

Color:

- Keep the number of colors low (5-6 distinct)
- Use redundant mappings (color+size)
- Include labeled color key
- Use resonant colors



Principles of good visualization

Aesthetics:

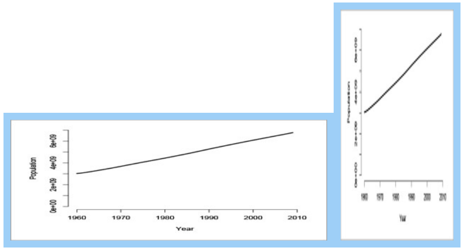
- Important findings should be visually emphasized
- Most important components in the center
- Do not put much information into one display

Other:

- The size of the plot should be normally Horizontal:Vertical=1.5:1
- Text in the graph is normally horizontal
- Caption and Source should be present and informative
- In bar charts, bars are normally sorted
- Axis labels present.

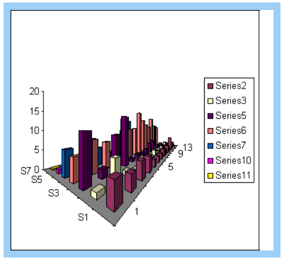
Misleading graphs

- Scaling and perspective problem



Misleading graphs

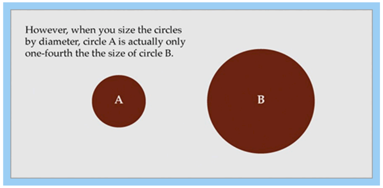
- Scaling and perspective problem



Misleading graphs

Abusing dimensionality/wrong mapping

- A scalar is mapped to a size of a cube
- Mapping is wrong: a scalar is mapped to radius, not area
  - R1=2R2, A1=4A2 !



Misleading graphs

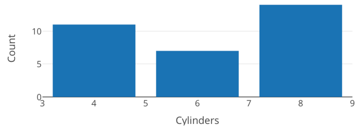
- Mixing data of different nature/scales
  - Ex: One time series plots with two series: Price and Amount both on Y axis
- Smoothed/filtered data interpreted as raw data
  - How good was the smoothing?
- Using of insufficient sampled data

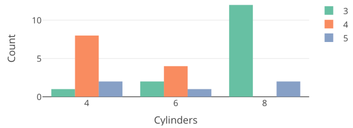
Basic plots

- Quantiative variable:
  1. Computing summaries (ex. frequencies)
  2. Visualizing as bar or pie charts
- What to analyse:
  - Largest and smallest bar or slice
  - For sorted bars, sudden shifts in level
  - Compare first within groups and then difference between groups

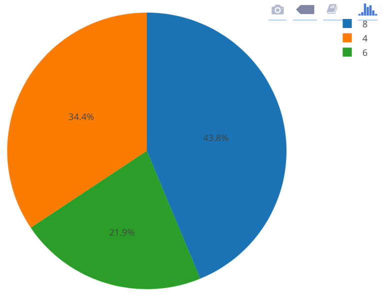
Basic plots

Example: Visualizing number of gears and number of cylinders in cars

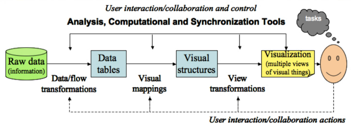




Basic plots

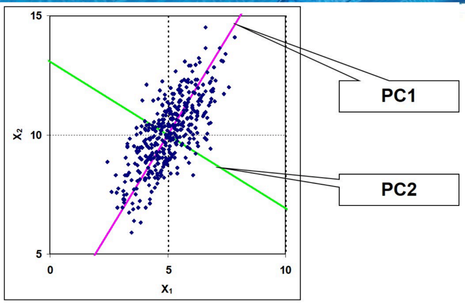


Visualization pipeline



- Dimension reduction
  - PCA
  - MDS
  - Correspondence analysis (nominal)
  - Other techniques (ex. ICA, Autoencoders), welcome to Machine Learning course..

Principal Component Analysis (PCA)



Distance between objects

- Meaning of "two objects are close"?
- Measure of proximity (ex: quantiative vars, Euclidian distance)
- Similarity measure s_{rs} (=1 if same object, <1 otherwise)
  - Ex: correlation
- Dissimilarity measure δ_{rs} (=0 if same object, >0 otherwise)
  - Ex: euclidian distance
- Problem of cosntructing the measures of proximity:
  - What if the variable is qualitative?
  - What if the object is a text document?

Multidimensional scaling (MDS)

Given n objects with known matrix of similarities or dissimilarities. Each object i is characterized by p-dimensional vector X_i

The aim:

- Present these objects in lower dimensions ($p' = 2$ or 3) such that the distance between the new points d_{rs} would reflect the matrix of similarities (or dissimilarities δ_{rs})
- See neighbour observations
- See clusters and outliers
- Have a "map" of your data

MDS

Two types of MDS:

- Metric MDS
- Non-metric MDS

Metric MDS
(algorithm is not discussed here)

Seaching for points $\chi_1, \ldots, \chi_n$, such that distances between $||\delta_{rs}||$ and $||d_{rs}||$ are minimized

Non-metric MDS

Given n objects $X_1, \ldots, X_n$ with known matrix of similarities $||\delta_{rs}||$ of dissimilarities.

For some configuration $\chi_1, \ldots, \chi_n$ (in lower dimension) with matrix $||d_{rs}||$, define **stress** $S(\chi_1, \ldots, \chi_n)$ by

1. Computing d'_{rs} as a a monotonic regression of $||d_{rs}||$ on $||\delta_{rs}||$
2. Computing $S = \sqrt{\frac{\sum_{r,s} (d_{rs} - d'_{rs})^2)}{\sum_{r,s} d_{rs}^2}}$

How to find optimal configuration?

- Use numeric optimization to minimize $S(\chi_1, \ldots, \chi_n)$

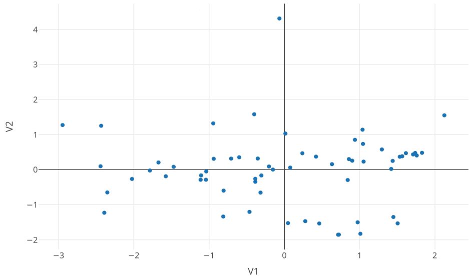
MDS- examples

Music data

- Artist (abba. Beatles. Wiwaldi, Mozart, Beethoven, Enya)
- Type (rock, classical, new wave)
- lvar, lave, lmax, lfener, lfreq - parameters of the music signal

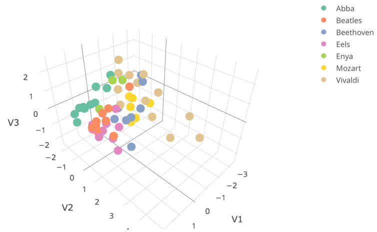
Metric MDS

- Mapping into two dimensions and using scatterplot



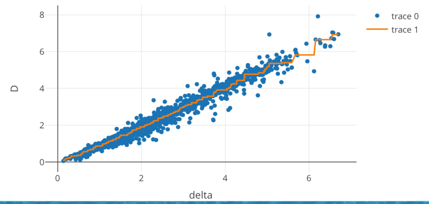
Non-metric MDS

- Mapping into three dimensions, coloring by Artist and using 3D-scatter:



Shephard plot

- Plot of d_{rs} vs δ_{rs}
- Displays also δ'_{rs} for non-metric MDS
- Shows the quality of MDS fit-> Best if scatter reminds a monotonic curve



Read at home

- Book, chapters 3.1, 3.3, 3.5, 13
- Cox, AA, and Cox, T.F.: "Multidimensional scaling." Handbook of data visualization. Springer, Berlin, Heidelberg, 2008. 315-347.
- Plotly book, ch 2.3