DATA MINING

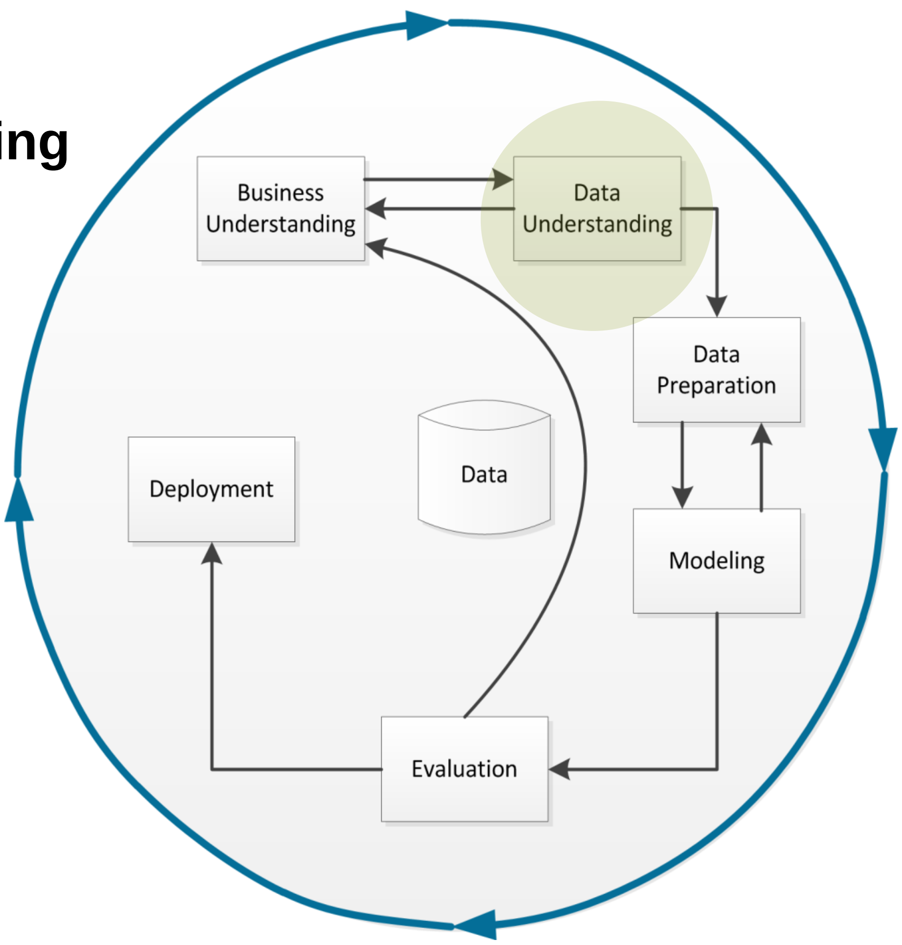
Data Understanding

JOHANNES KEPLER
UNIVERSITY LINZ

Data & Knowledge Engineering

CONTENTS

- Necessity for Data Understanding
- Levels of Measurement
- Basic Statistical Descriptions
- Data Visualization
- Digression:
  Explorative Data Analysis



NECESSITY FOR DATA UNDERSTANDING

Surface Impression of Data

Understanding of Problem Domain

GAIN SURFACE IMPRESSION OF DATA

- Know basic properties of data one operates with
  - Volume of data
  - Type of data (relational, transactional, …)
  - Level of measurement of attributes (nominal, ordinal, cardinal)

- Surface impression of data is important for
  - Judging applicability of data mining techniques, e.g., classification algorithms typically rely on non-cardinal data.
  - Estimating efforts for preprocessing, e.g., transformation of data in order to meet requirements of certain data mining techniques.

BETTER UNDERSTANDING OF THE PROBLEM

- What does the standard case look like?
  - Characterization of data
  - Central tendency and dispersion of data
  - Example: typically 40h/week (central tendency), 1 % works 45+h/week (dispersion)
- Are there any special cases?
  - Discrimination of data
  - Example: Someone working 55 hours represents an exceptional case in this example – but perhaps only in this example!
- Contrast central tendencies and dispersions !!
  - Different subsets of the data, e.g., profit in 2011 vs. profit in 2012
  - Different attributes, e.g., profit vs. benefits (in 2011 and 2012)

BETTER UNDERSTANDING OF THE PROBLEM

- Knowing characteristics, special cases, and dispersions
  - Asking the right questions (typically special cases are interesting)
  - Narrowing the problem domain (everything that is dispersed as expected is often of less interest)

- Example
  - Question: Why are daily sales of a company with five stores declining?
  - Quick look at the data may reveal that average daily sales of four stores are as usual, whereas the sales of the fifth store drastically declined.
  - Right question: Why are daily sales of the fifth store declining?

BETTER UNDERSTANDING OF THE PROBLEM

- Effects
  - Less and/or more appropriate results
    - Completeness
    - Soundness

  - Less demand for computational power and human resources

LEVELS OF MEASUREMENT

LEVELS OF MEASUREMENT

Level	Ranking	Distance	Example	
Nominal	–	–	hair color	(categorical)
Binary	–	–		(dichotomous, boolean)
symmetric	–	–	gender	both cases equally "interesting"
asymmetric	–	–	carcinogenic	causing cancer is more "interesting"
Ordinal	✓	–	drink size	small, medium, large
Cardinal	✓	✓	temperature	
interval-scaled	✓	difference	degree Centigrade	3 – 2 = 1 but 3 * 2 = ?
ratio-scaled	✓	multiplicity	degree Kelvin	3 – 2 = 1 and 3 * 2 = 6

OTHER CATEGORIZATIONS AND TERMINOLOGY

- Discrete vs. Continuous
  - Discrete = Nominal and Ordinal
  - Continuous = Cardinal
- Subgroups of nominal data
  - Binominal: nominal data with exactly two different values
  - Polynomial: nominal data with at least two different values
- Subgroups of cardinal (numeric) data
  - Integer
  - Real
  - Decimal
  - …

BASIC STATISTICAL DESCRIPTIONS

Central Tendency Measures

Dispersion Measures

Visualization

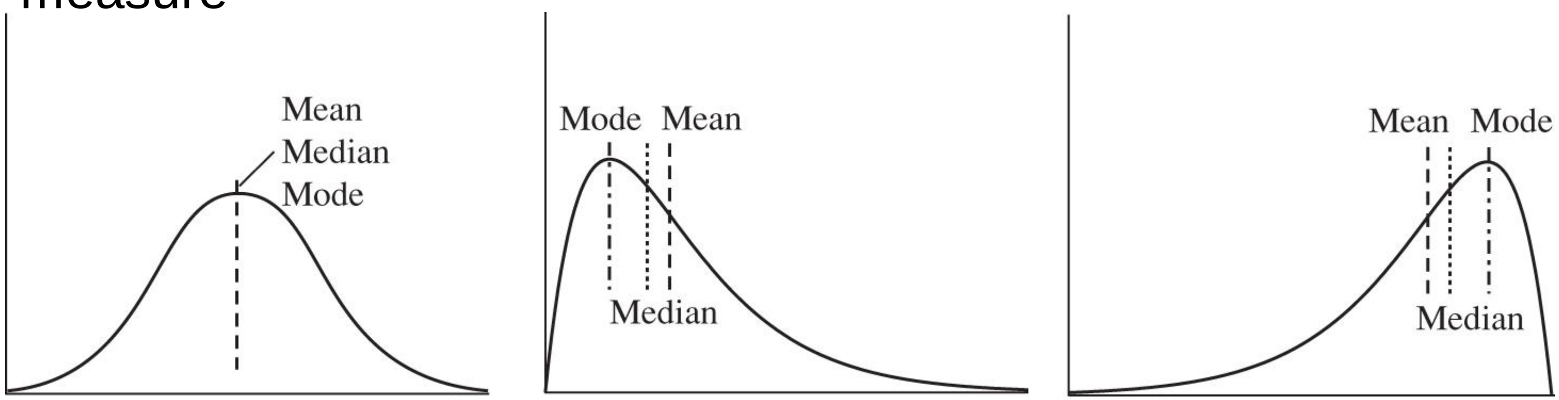
CENTRAL TENDENCY MEASURES

	Mean (average)	**Median** (value in the middle)	**Mode** (most frequent value)	**Example Data**
Nominal	–	–	*blue*	{*blue, red, blue*}
Ordinal	–	*med* Count = 5 Median = 3rd val	*small* and *large*	{*small, small, med, large, large*}
Cardinal	*4* Sum = 16 Count = 4 16/4 = 4	*3* Count = 4 Median = Mean of 2nd and 3rd val	*3*	{*1, 3, 3, 9*}

- Median is meaningless for nominal data
  - Arbitrarily sorted hair colors of swedish students:
    blonde, blonde, black, red, brunette -> Median = black
  - Yet, we all know that Swedes are typically blonde

CHOICE OF CENTRAL TENDENCY MEASURES

- Level of measurement is decisive
  - Median is for example meaningless for nominal data
- For cardinal and ordinal data, several measures for central tendency are meaningful in general
  - Which one describes the standard case best ??
- Skewedness is decisive for choosing the central tendency measure



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

CHOICE OF CENTRAL TENDENCY MEASURE

- Example: Prosperity indicator
  - Measure: income per citizen
  - What is the typical income of a citizen per year?

	Average income	Median income
Croatia	18.000 USD	15.000 USD
Equatorial Guinea	16.000 USD	1.000 USD

(fictitious data)

BASIC STATISTICAL DESCRIPTIONS

Central Tendency Measures

Dispersion Measures

Visualization

DISPERSION MEASURES

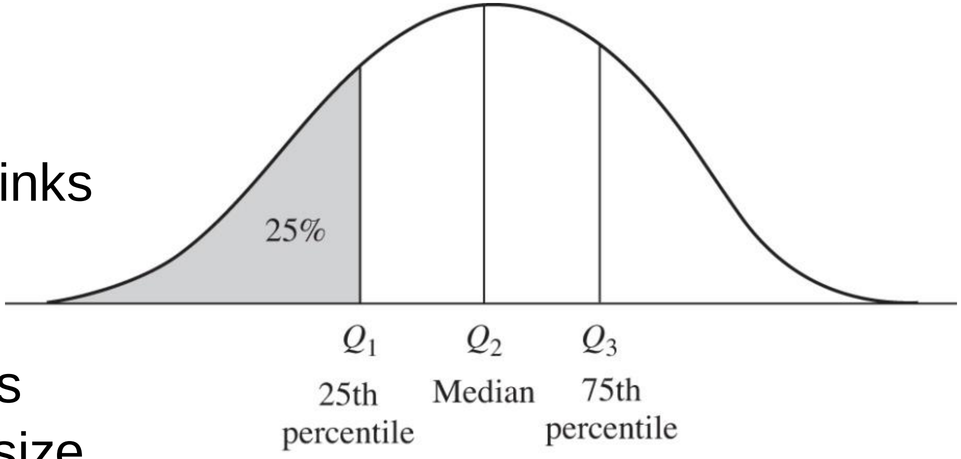
- How do data spread out around the center?
- Requires ability to relate data to each other
  - Nominal data are unordered; they cannot be related to each other
  - Example: how is brunette hair related to blonde hair?
  - Consequence: no dispersion measures for nominal data
- Dispersion measures for ordinal data
  - Range
  - Quantiles
- Additional dispersion measure for cardinal data
  - Standard deviation

ORDINAL DATA DISPERSION – RANGE

- Simply states the minimum and maximum value
  - For cardinal data the range is often regarded as the difference between the minimum and the maximum value
- Gives a rough impression of the dispersion
- Example
  - Available sizes of soft drinks: small, medium, large and extra large
  - Median size of sold soft drinks: large
  - Range of size of sold soft drinks: from medium to large
  - What does this tell us?

ORDINAL DATA DISPERSION – QUANTILES

- Quantiles extend the idea of the median
  - Median: 50 % of all vales are less than a certain value
  - q-Quantile: q % of all values are less than a certain value
- Frequently used types of q-Quantiles
  - Percentile: q = 100
  - Quartile: q = 4 (Q1, Q2, Q3, Q4)
    - Q2 = Median
- Example: Soft drink size
  - Q2 = medium: 50 % of drinks sold are at most of medium size (incl. small)
  - Q3 = large, 75 % of drinks sold are at most of large size



JKU

ORDINAL DATA DISPERSION – QUANTILES

- Five-Number Summary
  - Minimum, Q1, Q2, Q3, Maximum
- Example: Size of sold soft drinks
  - Min = Q1 = Q2 = medium
  - Q3 = large
  - Max = extra large
- Interpretation?

CARDINAL DATA DISPERSION – STANDARD DEV.

- For cardinal data, one can compute the distance (difference) between two values.
- In order to characterize how data spread around the center (average) one can "sum up" the distances between the values and the average value
- Variance: for technical reasons the distances are squared before summing them up

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x_i - \bar{x})^2 \qquad s^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_i - \bar{x})^2$$

- Standard deviation: For easier interpretation, the square root of the variance is used as indicator for the dispersion

CARDINAL DATA DISPERSION – STANDARD DEV.

- Example: Standard deviation in duration of phone calls

Call duration	**Difference to average**	**Squared difference**
2	2	4
2	2	4
3	1	1
4	0	0
4	0	0
5	-1	1
6	-2	4
6	-2	4
32		**18**

- Average: 32 / 8 = 4
- Variance: 18 / 8 = 2,25 "square minutes"
- Standard dev. = SQRT(2,25) = 1,5 minutes, i.e., most phone calls last between 2,5 (avg – sdev) and 5,5 (avg + sdev) minutes

CARDINAL DATA DISPERSION – STANDARD DEV.

- Question: How many phone calls indeed last between 2,5 and 5,5 minutes?
- Answer: Depends on the actual distribution of the data !?!?
- However, using Chebyshev's inequality, the following holds approximately true for any distribution (at least for n > 1000)
  - 50 % within +/- 1,4 σ
  - 75 % within +/- 2 σ
  - 89 % within +/- 3 σ
- If the distribution is known, even tighter bounds exist, e.g., if data follows the normal distribution:
  - 68 % within +/- 1 σ
  - 95 % within +/- 2 σ
  - 99 % within +/- 3 σ

CHOICE OF DISPERSION MEASURE

- Level of measurement is decisive !!
  - Standard deviation only applicable to cardinal data
- Standard deviation gives a good first impression of the dispersion
- 5-number summary is more precise but hard to comprehend when used for comparing dispersions of different variables

BASIC STATISTICAL DESCRIPTIONS

Central Tendency Measures

Dispersion Measures

Visualization

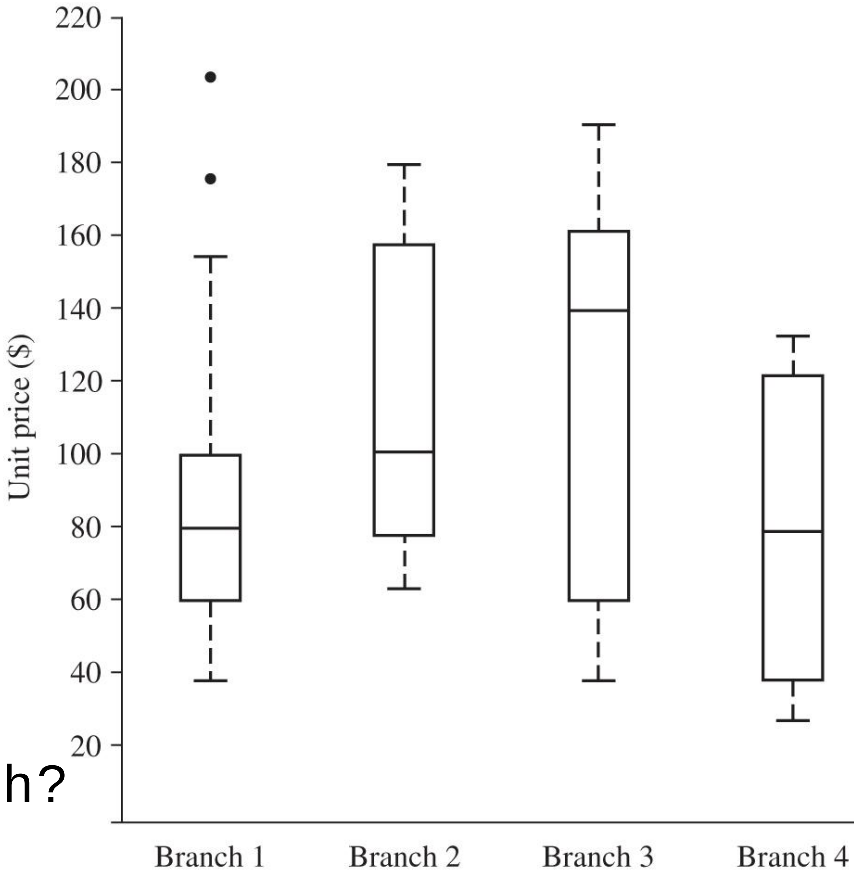
VISUALIZATION OF BASIC DESCRIPTIONS

- Visualization of uni-variate dispersions (single variable)
  - Box plot
  - Quantile Plot
  - Histogram

- Visualization of multi-variate dispersions (multiple variables, or multiple data sets of one variable)
  - Quantile-Quantile Plot
  - Scatter Plot (2D and 3D)

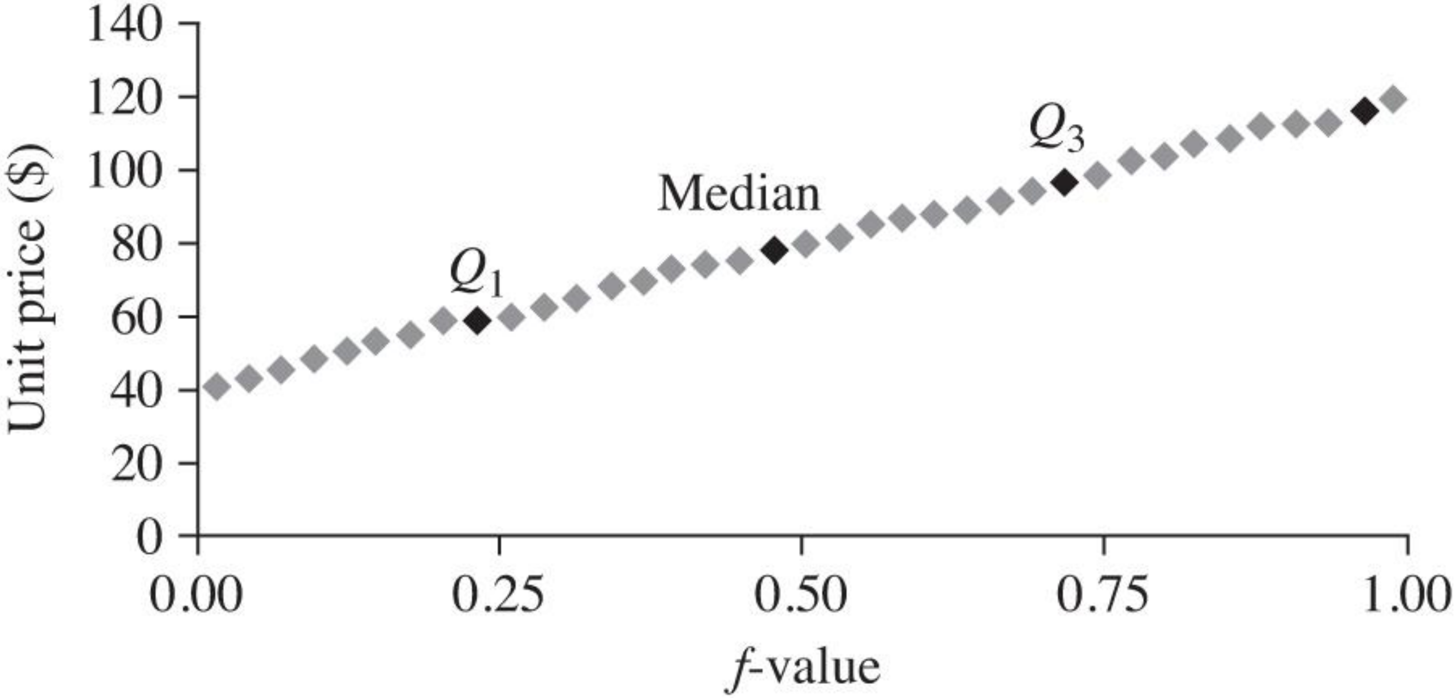
BOX PLOTS

- Five-Number summary
  - Min, Q1, Median, Q3, Max
- Box plot
  - Visualization of Five-Number summary
  - Whiskers indicate Extremes
  - Outliers
    - Whiskers then end at 1.5 * the range between Q1 and Q3, which is also called Interquartile Range
- Which is the cheaper branch?



QUANTILE PLOT

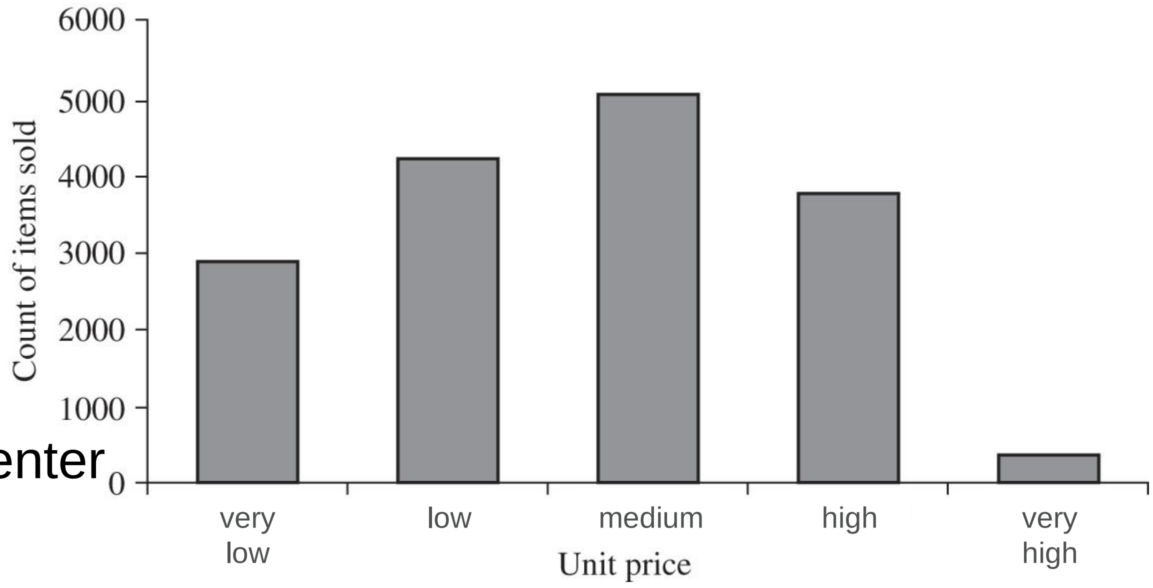
- Investigate univariate data distribution



- Allows comparison of different distributions based on quantiles

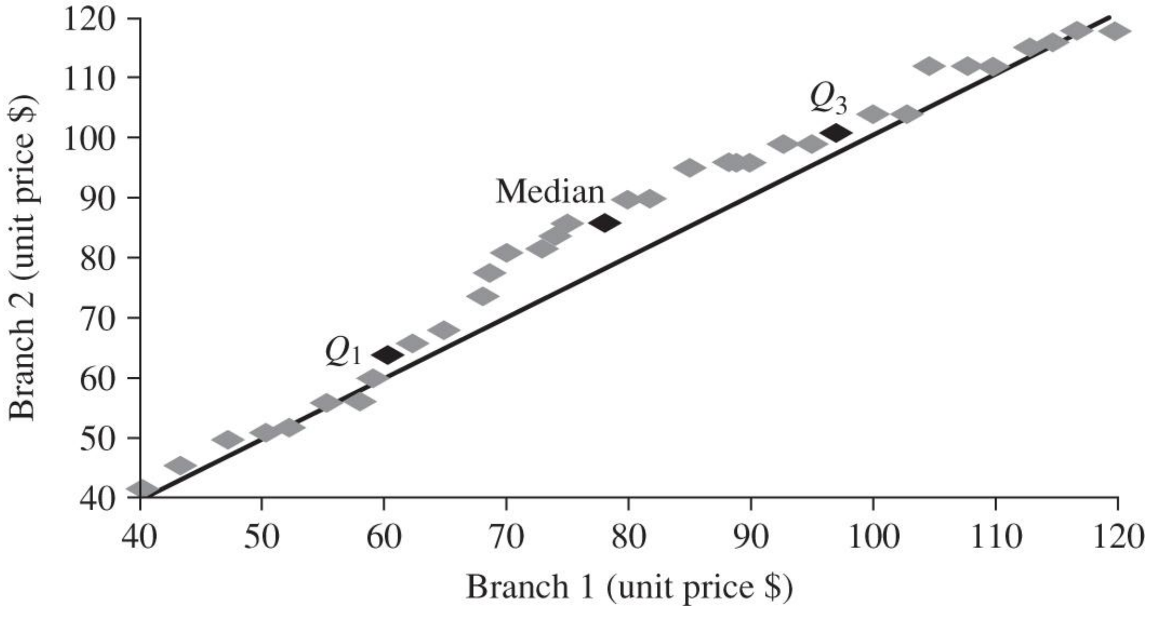
HISTOGRAM

- Display of tabulated frequencies, shown as bars
- Shows proportion of cases falling into each of several categories
  - Categories are non-overlapping intervals of some variable
- Bar chart vs. histogram
  - Bar chart: nominal data, i.e., categories are not ordered
  - Histogram shows categories in some order and thus relates data to the center



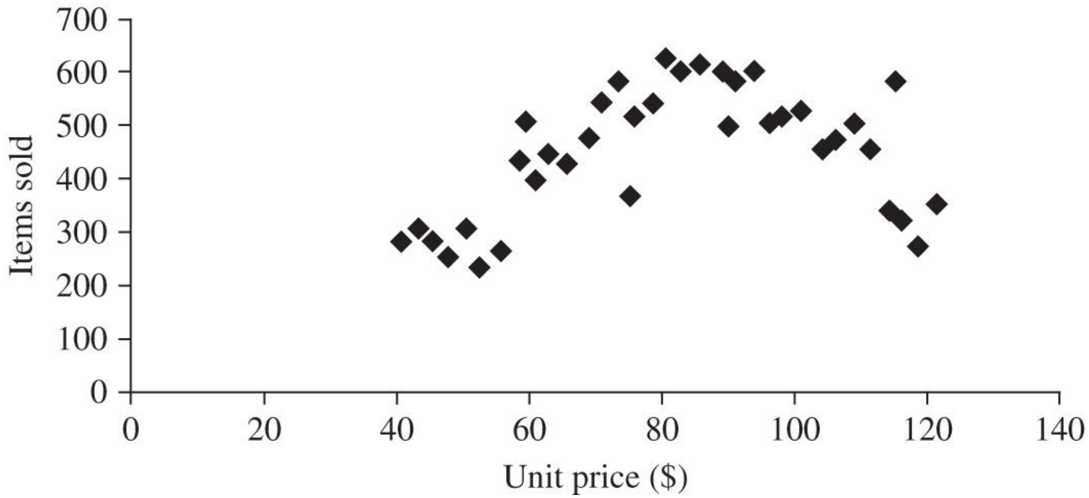
QUANTILE-QUANTILE PLOT
Q-Q PLOT

- Relates dispersion in two data sets w.r.t. the same variable
- Example: unit price (variable) in two branches (data sets)



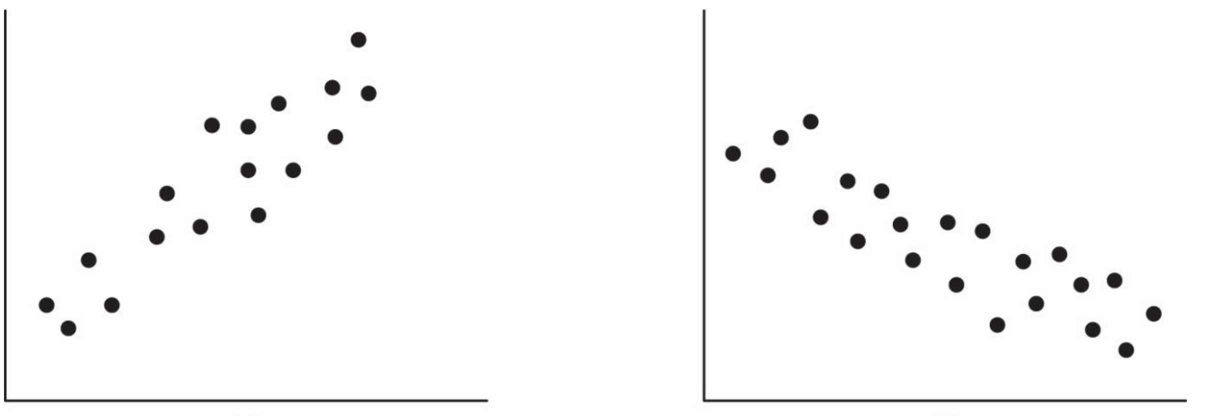
SCATTER PLOT

- Provides a first look at bi-variate data to see clusters, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



SCATTERPLOT AND CORRELATION

- correlation summarizes the *strength and the direction* of a *linear* (at least monotonic) relationship between two variables.



(a) (b)

(a) Positive correlation: higher x-value, higher y-value

(b) Negative correlation: higher x-value, lower y-value

BASIC STATISTICAL DESCRIPTIONS SUMMARY

- Central tendency:
  - Nominal data: mode
  - Ordinal data: median
  - Cardinal data: mean

- Dispersion
  - Nominal data: no measurement
  - Ordinal data: 5-number summary
  - Cardinal data: standard deviation and variance

- Visualization
  - Box plot, Quantile plot, Histogram, Q-Q plot, Scatter plot

DATA VISUALIZATION

Pixel-oriented techniques

Geometric projection techniques

Icon-based techniques

Hierarchical techniques

Complex data visualization techniques

WHY VISUALIZING DATA?

- Visualizing data is necessary for
  - the data analyst to quickly get a first impression
  - the business analyst to communicate results (a picture is worth a thousand words)

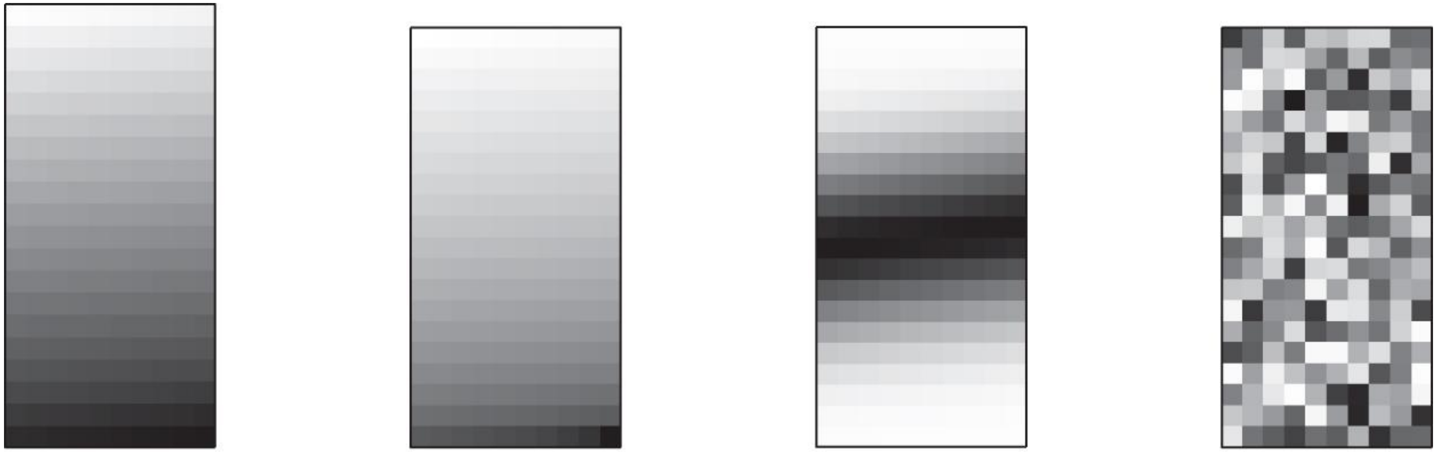
- Basic data descriptions and their visualizations focus on uni-variate or bi-variate data
  - They do not address high dimensional and complex data

WHY VISUALIZING DATA?

- Advanced visualization techniques include
  - ☐ Pixel-oriented visualization techniques
  - ☐ Geometric projection techniques
  - ☐ Icon-based visualization techniques
  - ☐ Hierarchical visualization techniques
  - ☐ Complex data visualization techniques

PIXEL-ORIENTED VISUALIZATION TECHNIQUES

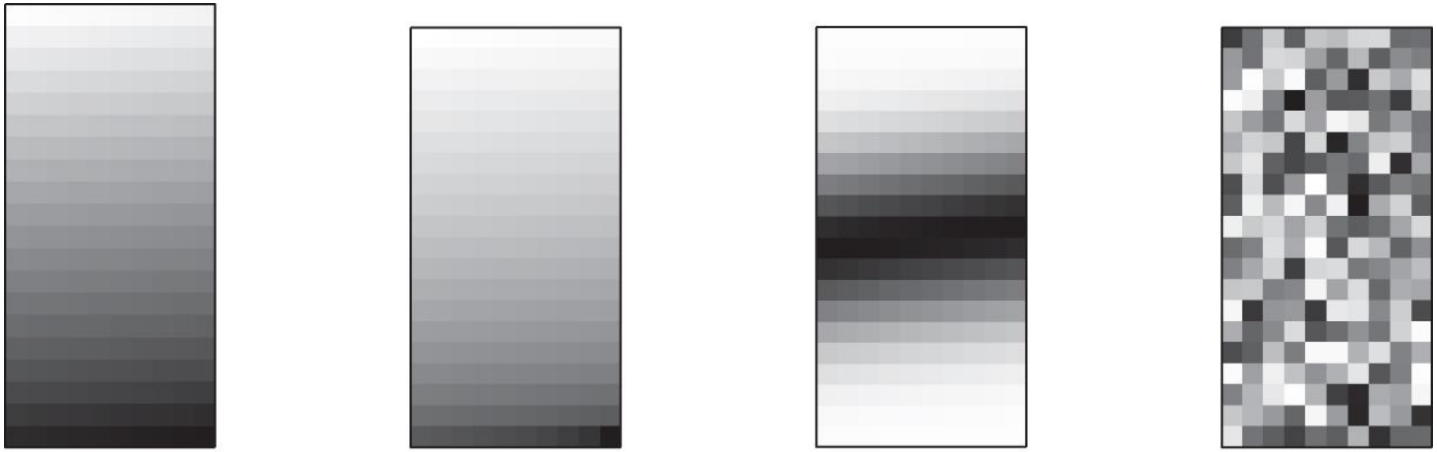
- Data are ordered globally
- For a data set of *m* dimensions, create *m* windows
- The *n* dimension values of a record are mapped to *n* pixels
- The colors of the pixels reflect the corresponding values



(a) *income* **(b)** *credit_limit* **(c)** *transaction_volume* **(d)** *age*

PIXEL-ORIENTED VISUALIZATION TECHNIQUES

- Problem of this particular technique:
  - Distance between pixels does not reflect global order
- Many other pixel-oriented techniques try to over come this limitation
  - Beyond the scope of this course



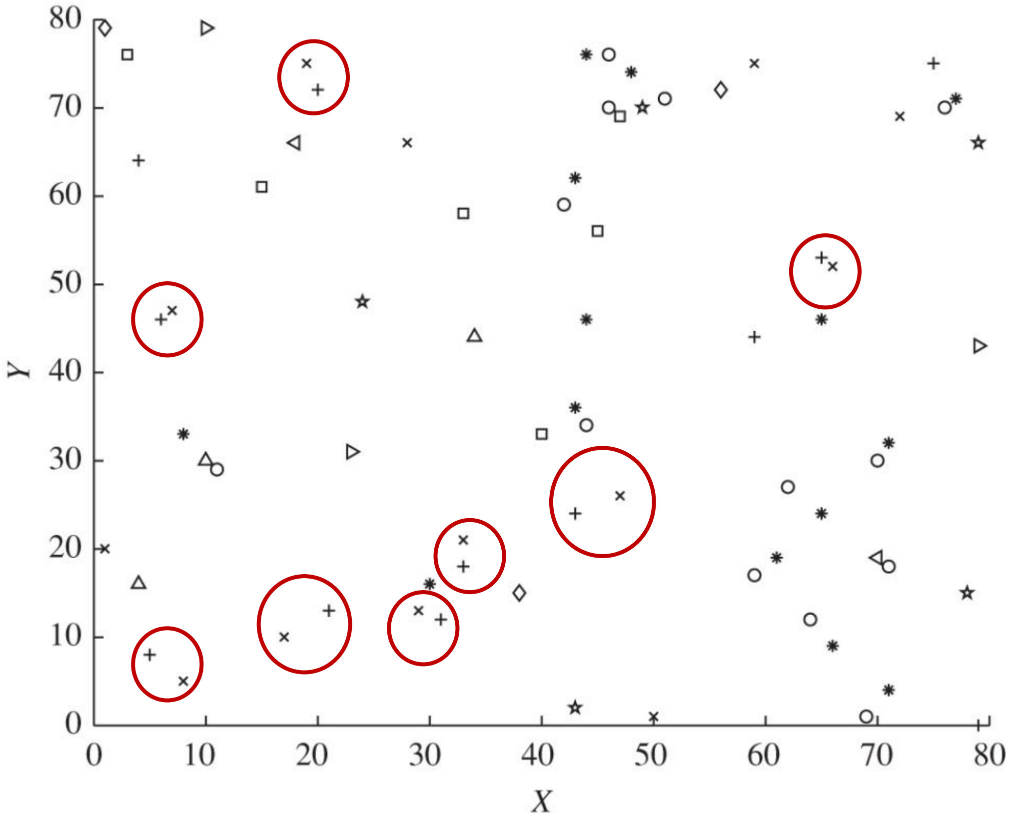
(a) *income* **(b)** *credit_limit* **(c)** *transaction_volume* **(d)** *age*

GEOMETRIC PROJECTION TECHNIQUES

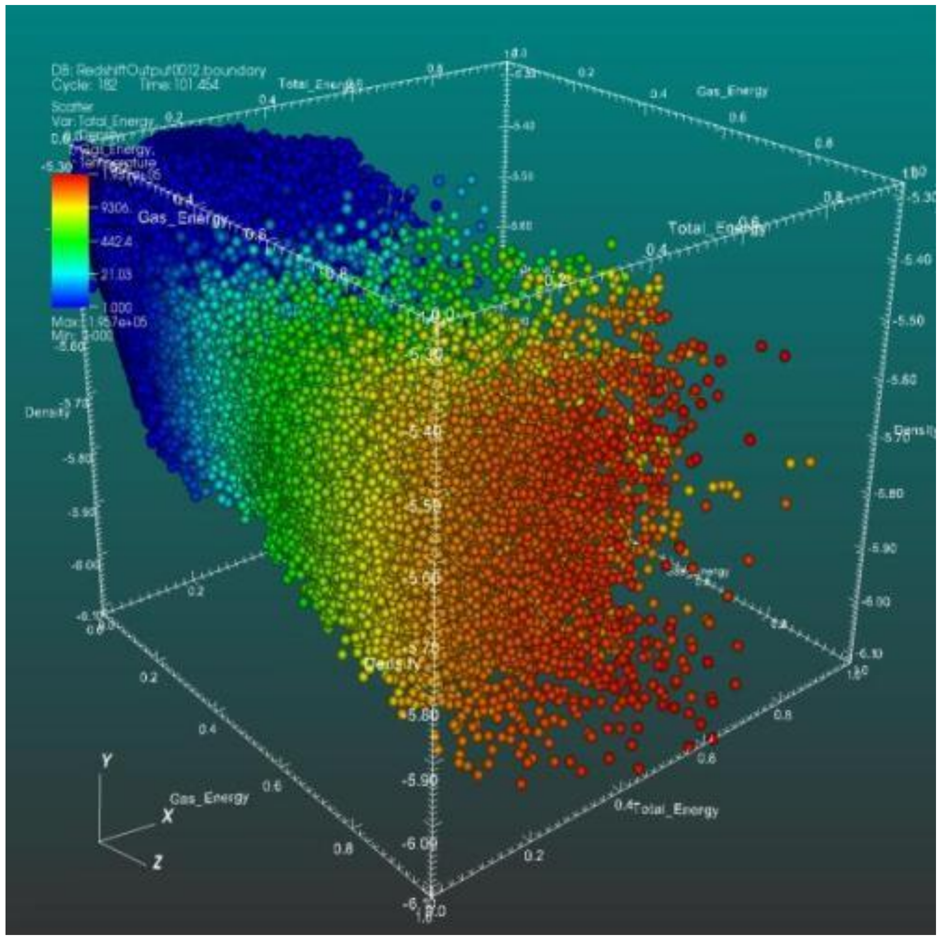
- Limitation of pixel-oriented visualization: does not show density in a multidimensional space
  - Geometric projection visualizations aim at overcoming this limitation
- Central challenge:
  - How to visualize high-dimensional space on a 2-D display?
- Prominent examples
  - Scatter plot (enhanced version)
  - Scatter plot matrix
  - Parallel coordinates
  - Principal Components Analysis (PCA)

SCATTER PLOT

- X = longitude
- Y = latitude
- Icons:
  - ○ = university
  - + = medical office
  - × = pharmacy
- Shows, e.g., that medical offices and pharmacies are frequently co-located
- Extension: 3D scatter plot

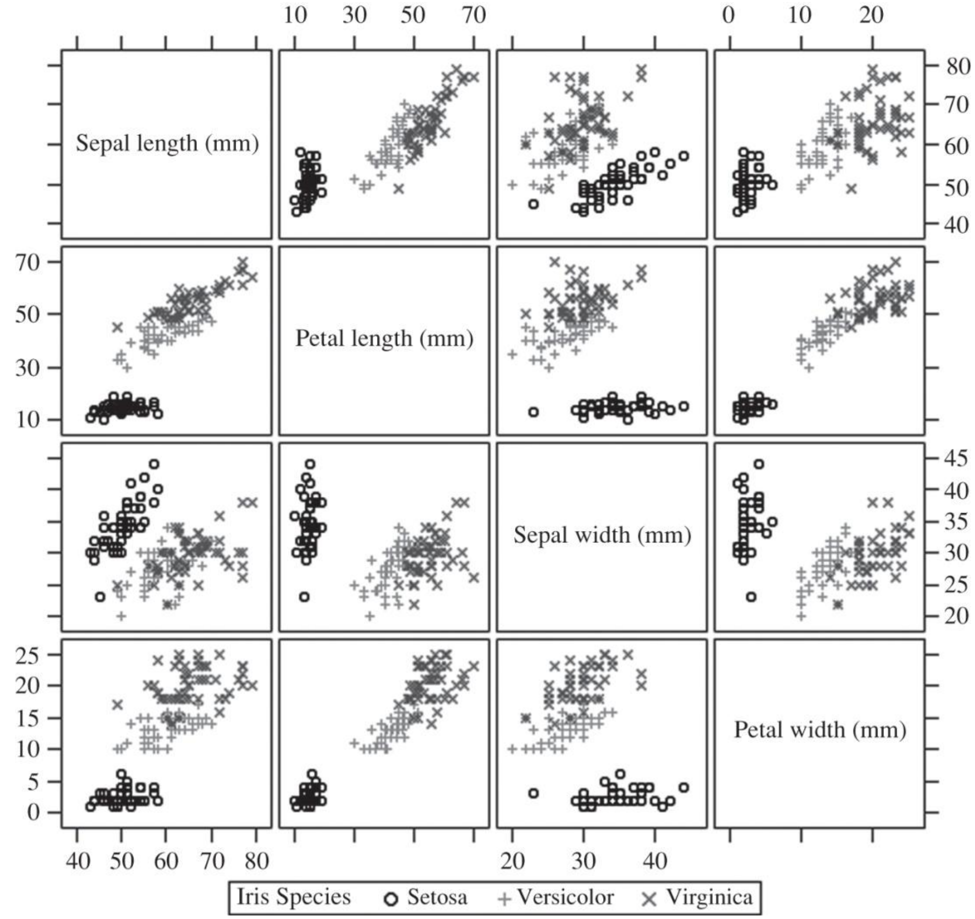


3D SCATTER PLOT



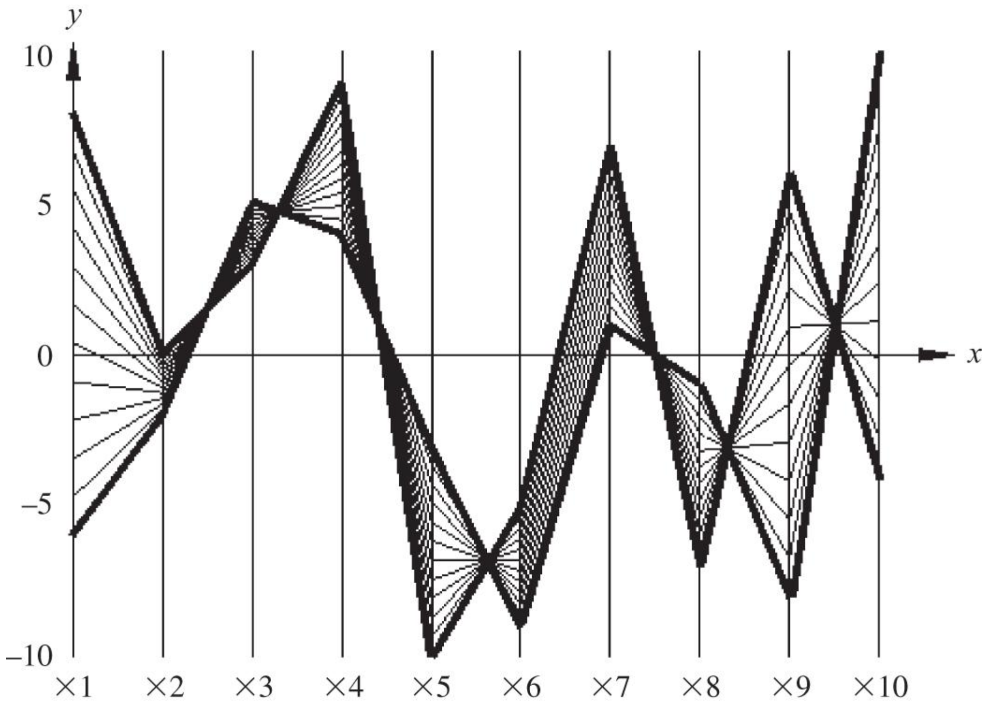
SCATTER PLOT MATRIX

- Relates dimensions to each other
- Reasonable only for few dimensions



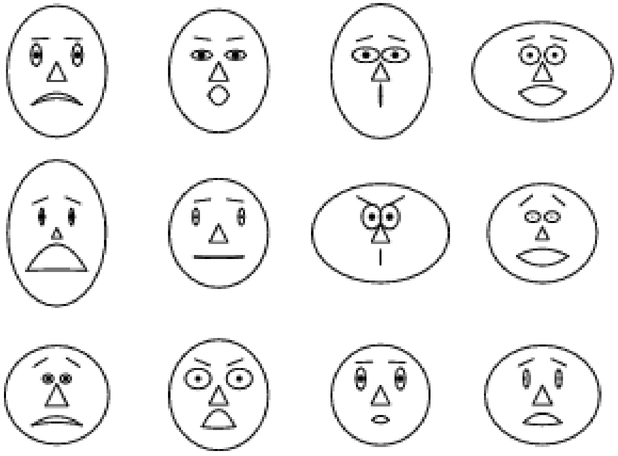
PARALLEL COORDINATES

- Equidistant axes
- The axes are scaled to the range of the attributes
- Every data item corresponds to a polygonal line intersecting the axes



ICON-BASED VISUALIZATION TECHNIQUES

- Visualization of the data values as features of icons
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
- Prominent example
  - Chernoff Faces
  - Basic idea: People are specialized on interpreting facial expressions



JKU

HIERARCHICAL VISUALIZATION TECHNIQUES

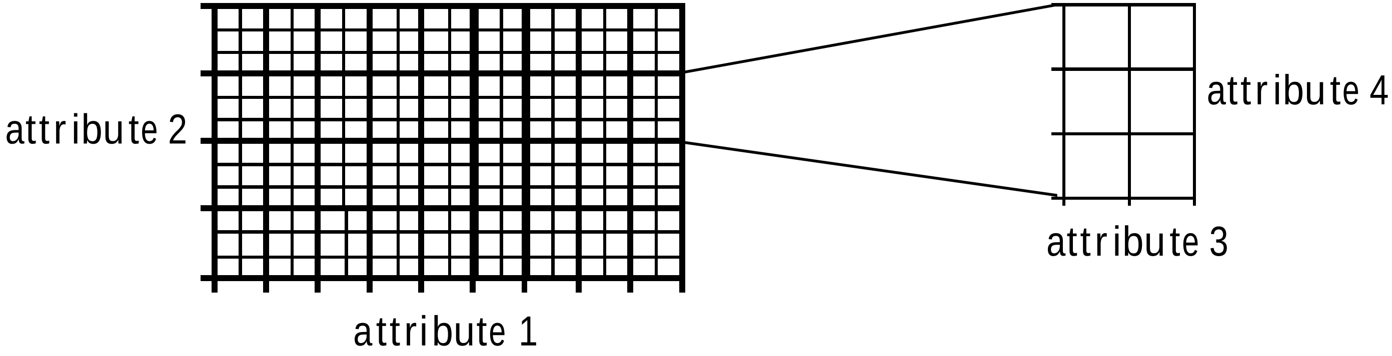
- Limitation of geometric projection visualization: get confusing for high-dimensional data

- Hierarchical visualization techniques address this limitation by partitioning dimensions into subspaces and visualizing them in a hierarchical manner.

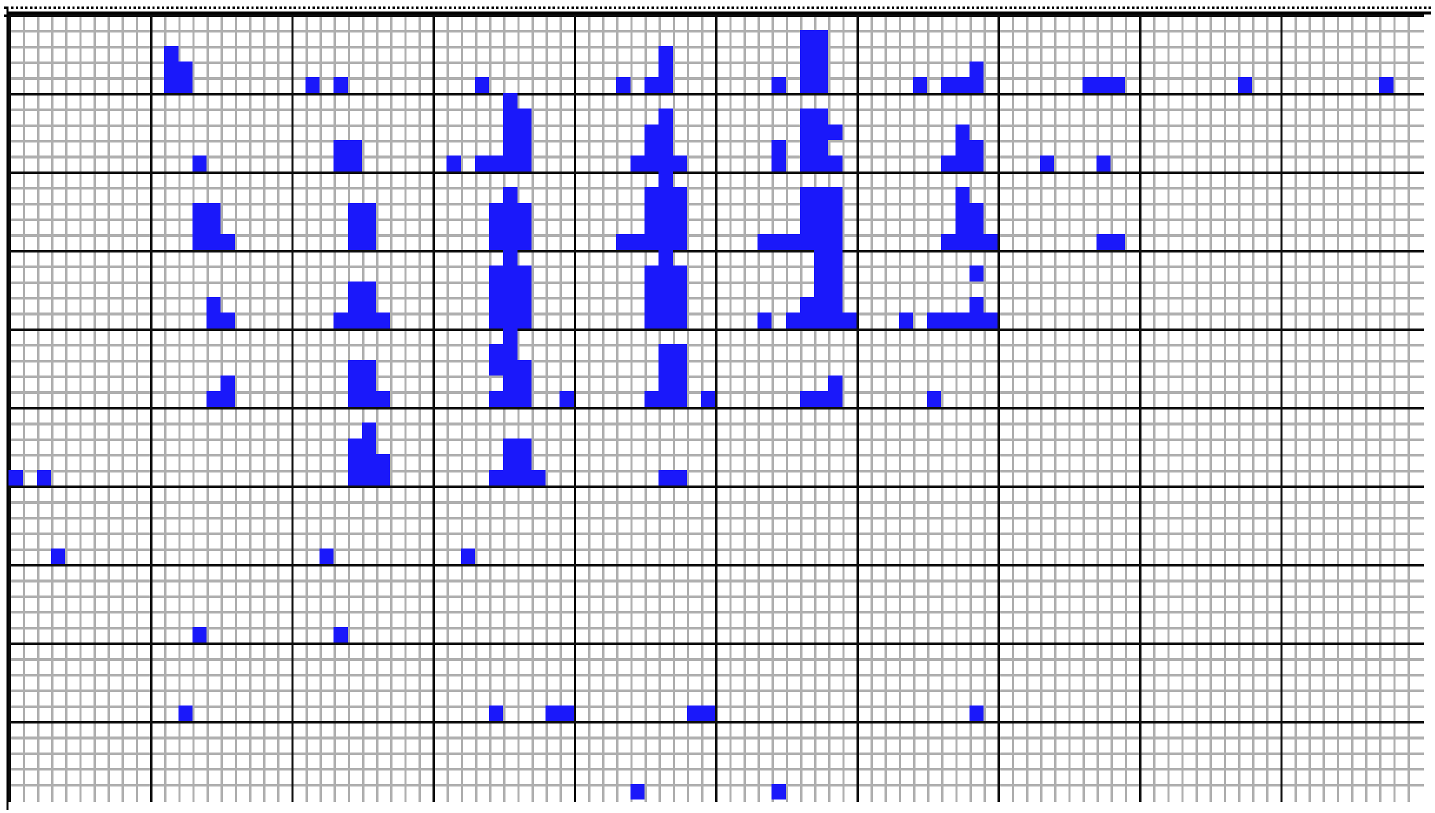
- Prominent examples:
  - Dimensional stacking
  - Worlds-within-Worlds
  - Tree Map

DIMENSIONAL STACKING

- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other
- Partitioning of the attribute value ranges into classes
- Important attributes should be used on the outer levels
- Adequate for data with low cardinality
- Difficult to display more than nine dimensions.



DIMENSIONAL STACKING

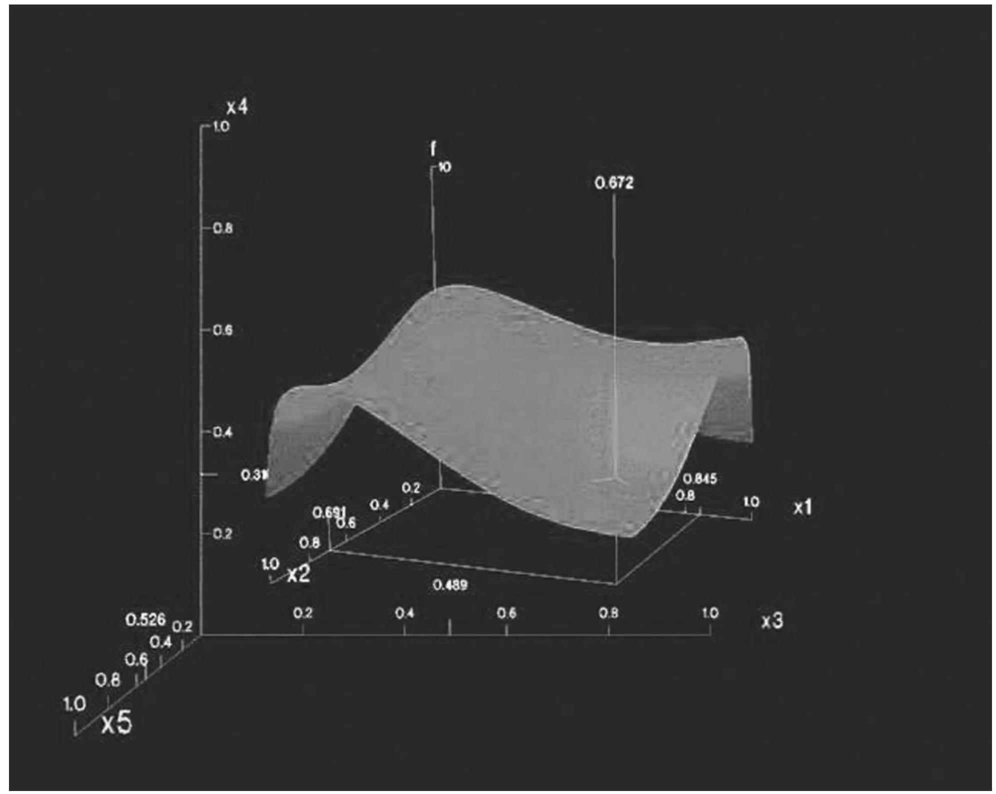


- Oil mining data: longitude and latitude (outer axis) vs. ore grade and depth (inner axis)

JKU

WORLDS-WITHIN-WORLDS

- Goal: visualize effect on one dimension if other dimensions are fixed to certain values
- Targeted dimension and two most important parameters are assigned to innermost world
- Outer worlds are created in dependence of
  the number of dimensions

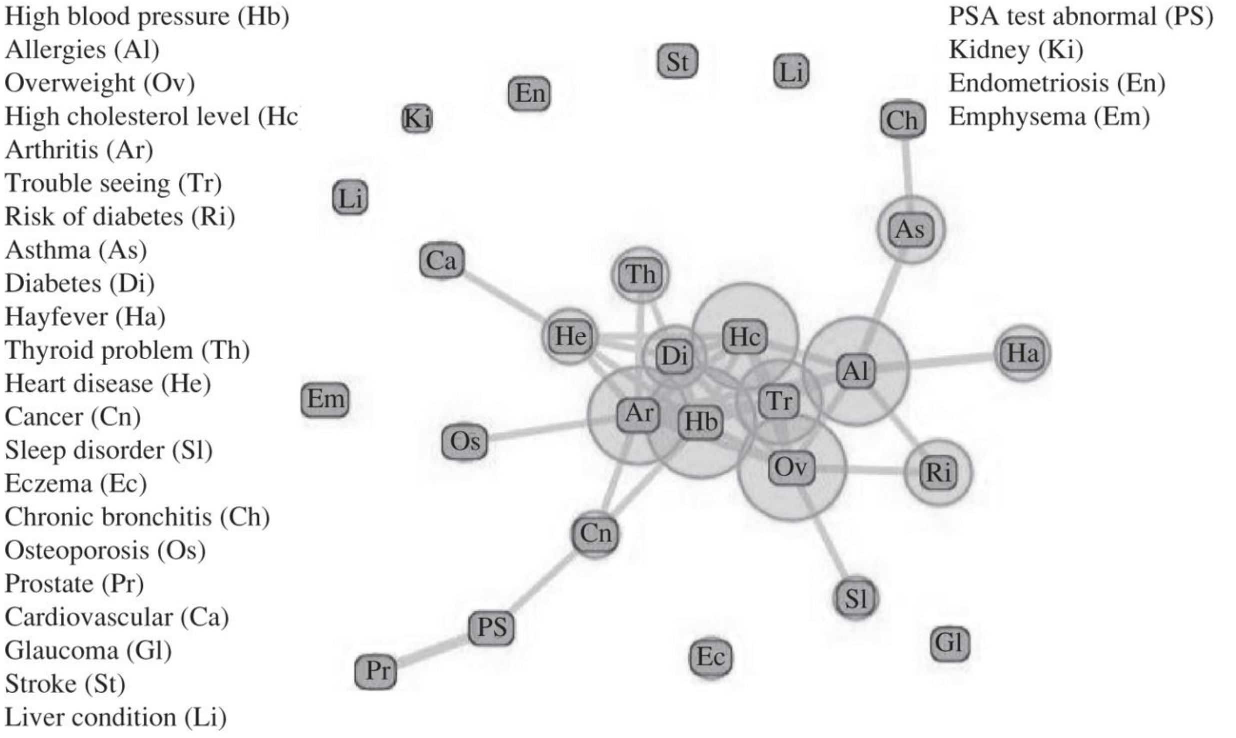


TREE MAP

- Visualizes hierarchical data as a set of nested rectangles
- Example:
  - Google news visualized as treemap using newsmap library
  - Main categories are the large rectangles with unique colors
  - Sub-categories are nested

COMPLEX DATA VISUALIZATION TECHNIQUES

- Earlier techniques focused on visualization of numeric data
- Now focus on non-numeric data like social network data
- Prominent example: Tag Clouds



EXPLORATIVE DATA ANALYSIS (EDA)

DIGRESSION: EXPLORATIVE DATA ANALYSIS

- Purpose: Gain interesting insights (verify hypothesis) by visualizing data instead of applying data mining algorithms

- Example: Are premium customers typically over 40 and wealthy?
  - Data mining: Apply a classification algorithm
  - EDA: Inspect 3D scatter plot of customer age, income and group

DIGRESSION: EXPLORATIVE DATA ANALYSIS

- Pros
  - Little technical knowledge required and thus also non expert users can explore data visually
  - Humans have expert knowledge about the problem domain neglected by data mining algorithms when searching for patterns

- Cons
  - Ability of finding patterns depends on ability to visualize possible complex data

SUMMARY

- Data understanding forms the basis data mining (CRISP-DM process)
- Statistics recap
  - Levels of measurement
  - Central tendencies
  - Dispersions
  - Visualization of statistical measures
- Data visualization
- Explorative Data Analysis

JKU