Data Preparation for Data Mining

Lesson 7

Lesson 7 Overview

- Bringing it All Together
  - Data Survey
  - Preparing the Dataset
  - Using the Prepared Data
- Graphical Examination of Data
- Let’s Do it Ourselves!

What is Data Survey?

- Preparation -> Survey -> Mining
- Dataset enfolds information
  - Purpose of the Data Survey is to show and clarify shape, substance, structure, relationships and limits of this information
  - "Just what information is in here anyway?"
- Most surveying techniques can be automated

Surveying the data

- The goal
  - to find the problem areas in the data, so that the mining can be planned optimally.
- Main tools
  - Confidence analysis
  - Entropy analysis
  - Analysis of sparsity and variability
  - Cluster analysis
  - Distribution analysis

What is Information?

- Data contains information, but…
- What is information?
  - Communication of instructive knowledge (Oxford)
- Information comes in two pieces

  1) Informing communication (signal)

  2) Framework in which to interpret information
- Measuring information -> Shannon's Information theory (1948)

Measuring information: Signals

- Signals carry information of system state
- Simplest signal has binary nature (on/off)
  - Binary digit or Bit of information
- Bits are most common way to measure information
- Redundancy measures duplicate information in system states

Measuring information : System states

- Number of possible system states can be usually easily discovered
- Number of bits needed to present all the system states:

  $n = log_2(\# \text{ of system states})$
- If each state is equally likely, this is also information content of the system

Measuring information: Entropy

- Probability distribution of system states should be taken into account when measuring information content -> Entropy E
- $E=-\Sigma p_i log(p_i)$
- Entropy measures uncertainty in system
  - Maximum entropy when all of the system states equally probable
- **Total information content cannot be measured!**
  - Requires evaluating the significance of the signal

Using Entropy Measurements

- Entropy measurements are the foundation for evaluating and comparing information content
  - Maximum entropy
  - Whole data set entropy $E(x)$
  - Conditional entropy $E(y|x)$
  - Mutual information $E(x;y)=E(x)-E(x|y)=E(y)-E(y|x)$

Using Entropy Measurements

- Entropy can be used to
  - To evaluate the quality and problem areas of the input and output variables
  - To estimate independence of variables
  - To select the input variables with maximum predictive information about the outputs
  - To estimate the maximum possible quality of a model

Identifying Problems with a Data Survey

- While entropy measurements identifies *where* problems are, they say little about *what* they are
- Possible problems
  - The data set doesn't enfold sufficient information
  - Part of the data is not well defined (sparsity)
  - High variance or noise obscures relationships
  - The relationships are very complex (nonlinearity)

Is data sufficient?

- Data survey makes assumption that dataset is multivariably representative
- If not get better data…
- If yes entropy measures can be used to determine if information content of input data is sufficient to define output
- Entropy measures also suggest way to reduce variables

Detecting Sparsity

- State space density variation is positive necessity
- However there could be a problem if sparsity falls to such level that data doesn’t carry sufficient information to define relationships
- Low entropy, and antisymmetric forward and reverse entropy values could be indicator for this
- Also state space density maps are useful for pointing problem areas

High variance and Noise

- High entropy values indicate high variance or noise
- Model building (estimating manifold) could be problematic
- Also measuring skewness gives information about problems

Complex relationships

- Determining the shape of manifold and measuring its complexity are hard tasks and are not jobs for data survey but modeling
- However output of the model can be surveyed against dataset using information measures to indicate potential problem areas

What’s Next in the Data Survey?

Surveying the data

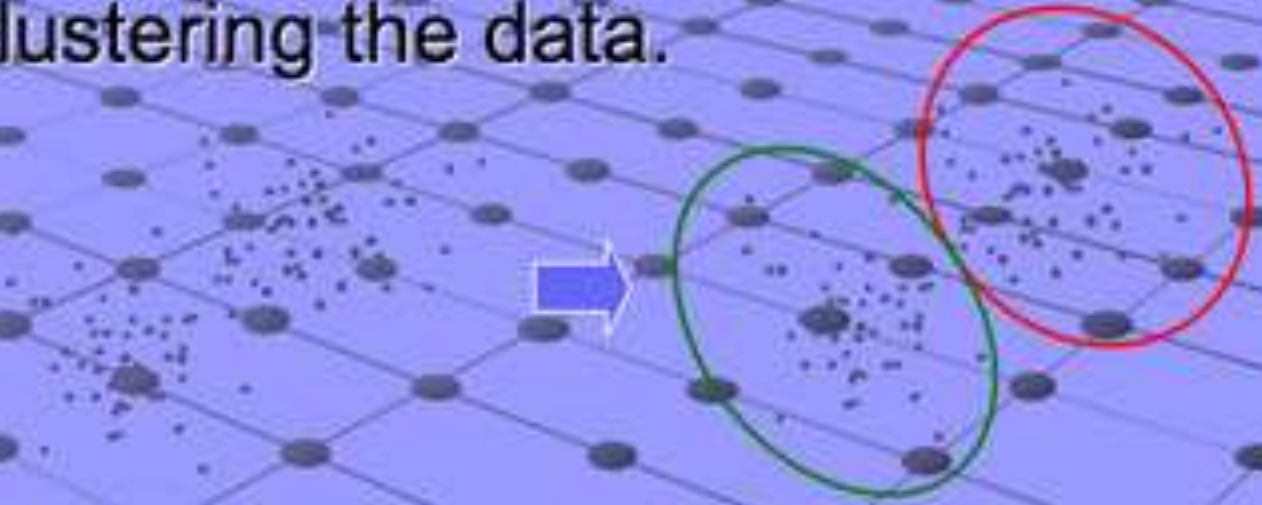
- The goal
  - to find the problem areas in the data, so that the mining can be planned optimally.
- Main tools
  - Confidence analysis
  - Entropy analysis
  - Analysis of sparsity and variability
  - Cluster analysis
  - Distribution analysis

Sampling Bias

- Sampling bias is one of the most common error sources in data analysis.
- Sampling bias is generated, when
  - data points that should be included are left out from the analysis (omission)
  - data points that should be excluded are taken in to the analysis process (commission).
- Analysis of the clusters and variable distributions reveal the possible problems.

Cluster Analysis

- States of the system can be studied by clustering the data.



- Clustering may help to detect possible problems in the data.

- Clusters represent the likely system states
    - Finding an explanation for the data clusters helps to understand the data.
- Clusters may also reveal a sampling bias
    - Clusters can be created by an omission or a commission error.

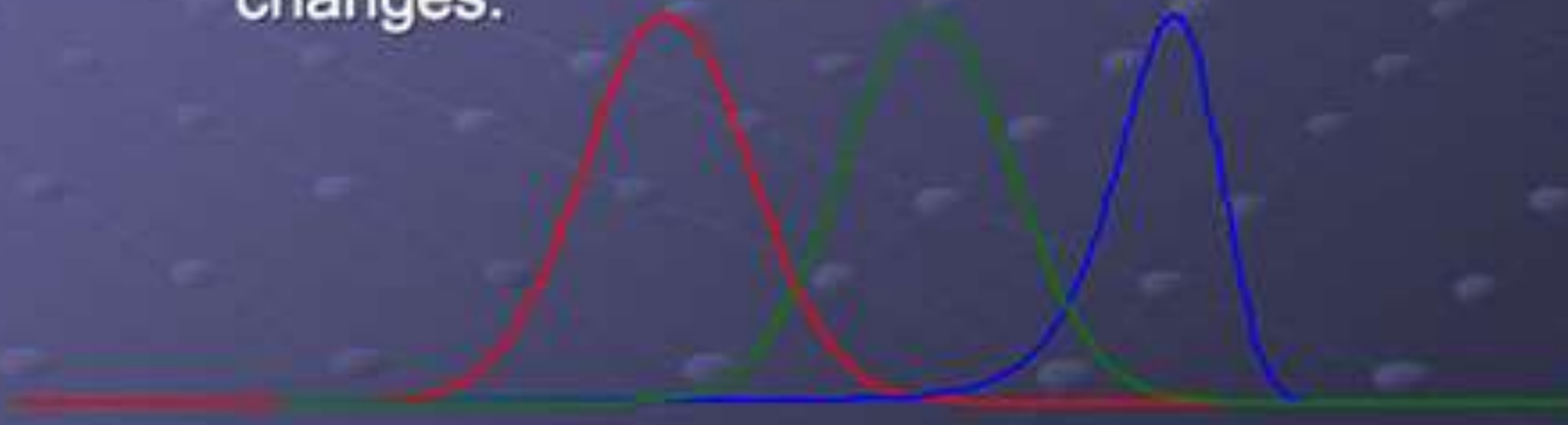
- In general, the input clusters should map to the output clusters
    - if knowing the input cluster doesn't help in predicting the output cluster, problems are to be expected.
- Knowing the possible strict dependencies between the input and output clusters allows the miner to focus on more problematic areas of the data.

Distribution Analysis

- In general, if the data is unbiased, the shape of the distribution of the output variables should remain the same across different input variable values.
  - Changing the input value changes the output value, but not the behavior of the system.

- An example
  - When trying to define the amount of potential restaurant customers among a concert hall audience by analyzing the dependence between the number of customers in the restaurant and the number of concert tickets sold, full house hours may bias the results as some of the potential customers can't be served.
  - This may be diagnosed as an omission (some potential customers are left out of the data) or as a commission (full house hours should be left out of the analysis). One explanation would be that a variable containing information of the vacant tables is missing.

- Sampling bias may be observed as a change in the distribution of dependent (output) variables
    - when the number of concert tickets sold is high, the skewness of the distribution of the number of customers in the restaurant changes.



Basic Data Survey Procedure

- Estimate how well the data represents and covers the true population
- Analyze the entropy of and between the variables
- Try to explain the clusters
  - Check the mapping between input and output clusters.
- Check sparsity and uncertainty
- Check variable distributions
  - Try to explain the possible changes in the distributions.

Additional Methods

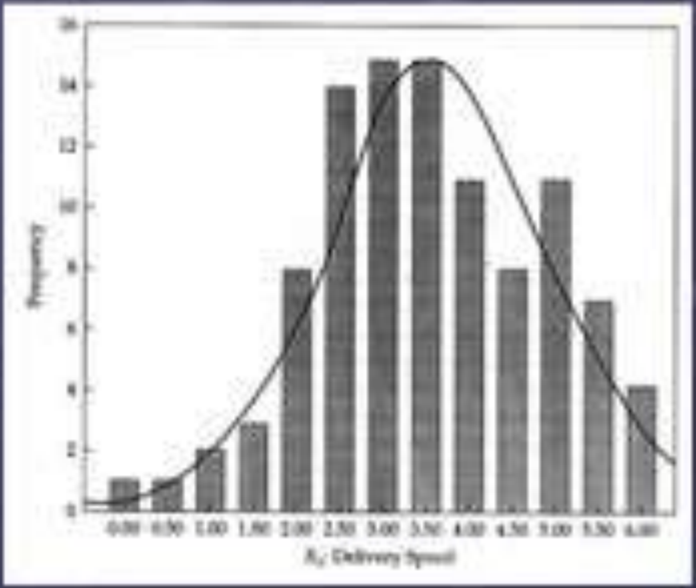
- Novelty detection
  - mainly used when exploiting the mining results
  - estimates the probability that a certain input is drawn from the same population as the training data
- Tensegrity structures
- Fractals (used as manifolds)
- Chaotic attractors

Graphical Examination of Data

- Examining one variable
- Examining the relationship between two variables
- 3D visualization
- Visualizing multidimensional data

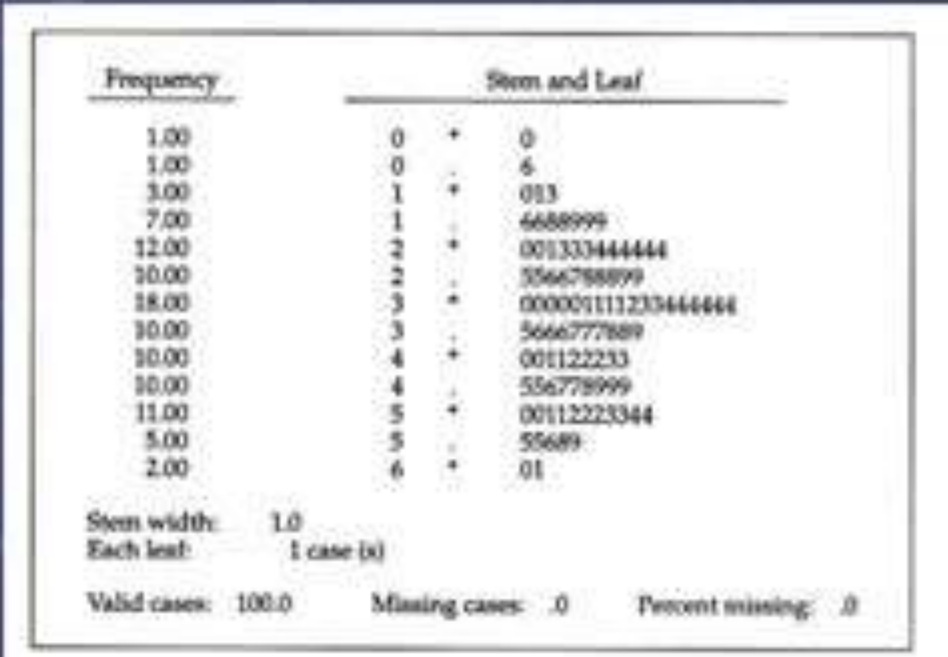
Examining one variable

- Histogram
  - Represents the frequency of occurences within data categories
    - one value (for discrete variable)
    - an interval (for continuous variable)



Examining one variable

- Stem and leaf diagram
  - Presents the same graphical information as histogram
  - provides also an enumeration of the actual data values

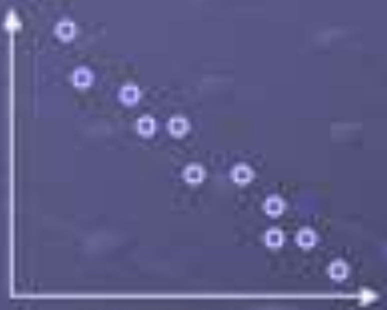


```
Frequency            Stem and Leaf

  1.00         0  *   0
  1.00         0  .   6
  3.00         1  *   013
  7.00         1  .   4688999
 12.00         2  *   001333444444
 10.00         2  .   5566788899
 18.00         3  *   000001111233444444
 10.00         3  .   5666777889
 10.00         4  *   001122233
 10.00         4  .   556778999
 11.00         5  *   00112223344
  5.00         5  .   55689
  2.00         6  *   01

Stem width:     1.0
Each leaf:        1 case (s)

Valid cases:  100.0      Missing cases:  .0      Percent missing:  .0
```

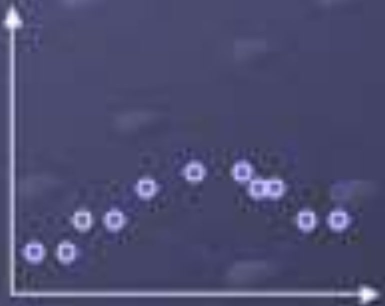
FIGURE 2.2 Stem and Leaf Plot of X_1 (Delivery Speed)

Examining the relationship between two variables

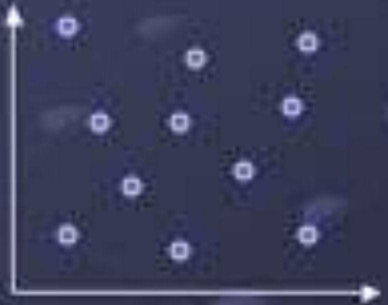
- Scatterplot
  - Relationship between two variables



Linear



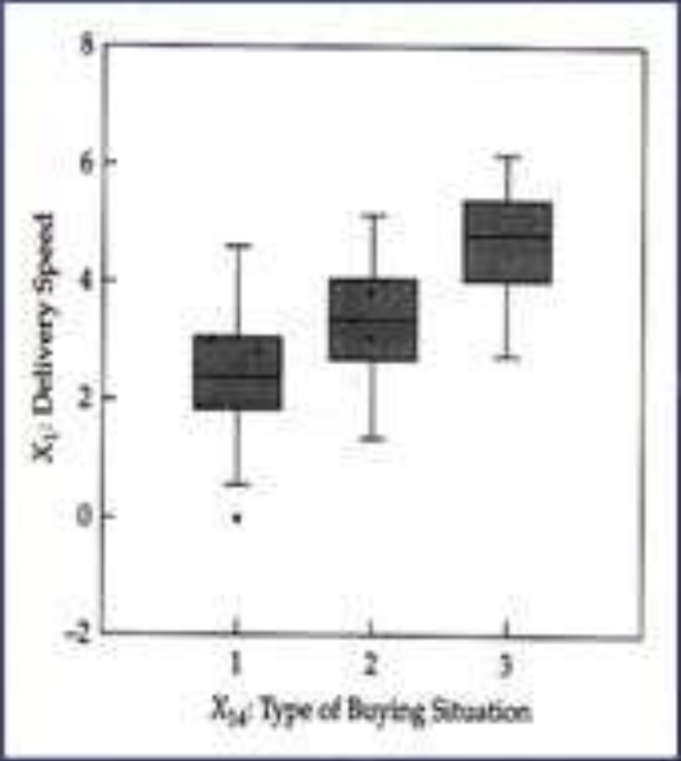
Non-linear



No correlation

Examining the relationship between two variables

- Boxplot
  - Representation of data distribution
  - Shows:
    - Middle 50% distribution
    - Median (skewness)
    - Whiskers
    - Outliers
    - Extreme values



3D visualization

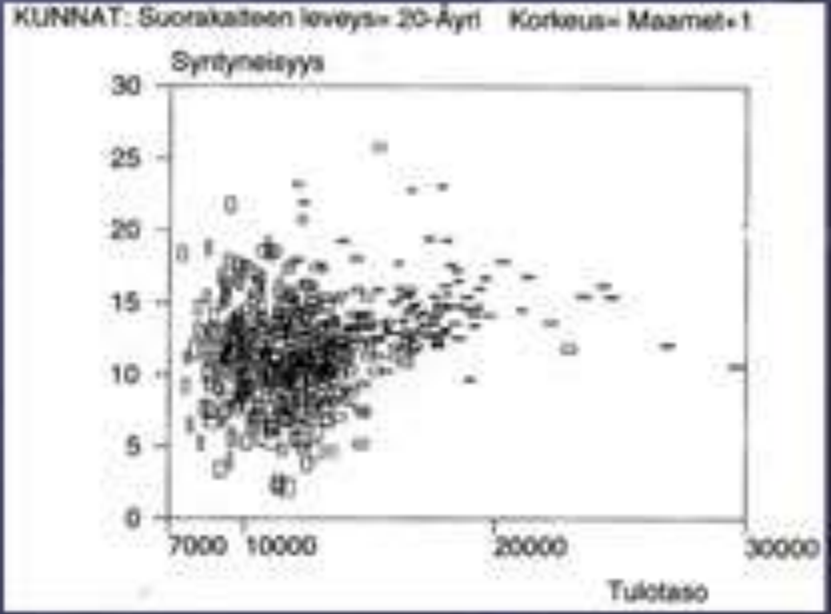
- Good if there are just 3 variables
- Mustonen: “Problems will arise when we should show lots of dimensions at the same time. Spinning 3D-images or stereo image pairs give us no help with them.”

Visualizing multidimensional data

- Scatterplot with varying dots
- Scatterplot matrix
- Multivariate profiles
- Star picture
- Andrews’ Fourier transformations
- Metroglyphs (Anderson)
- Chernoff’s faces

Scatterplot

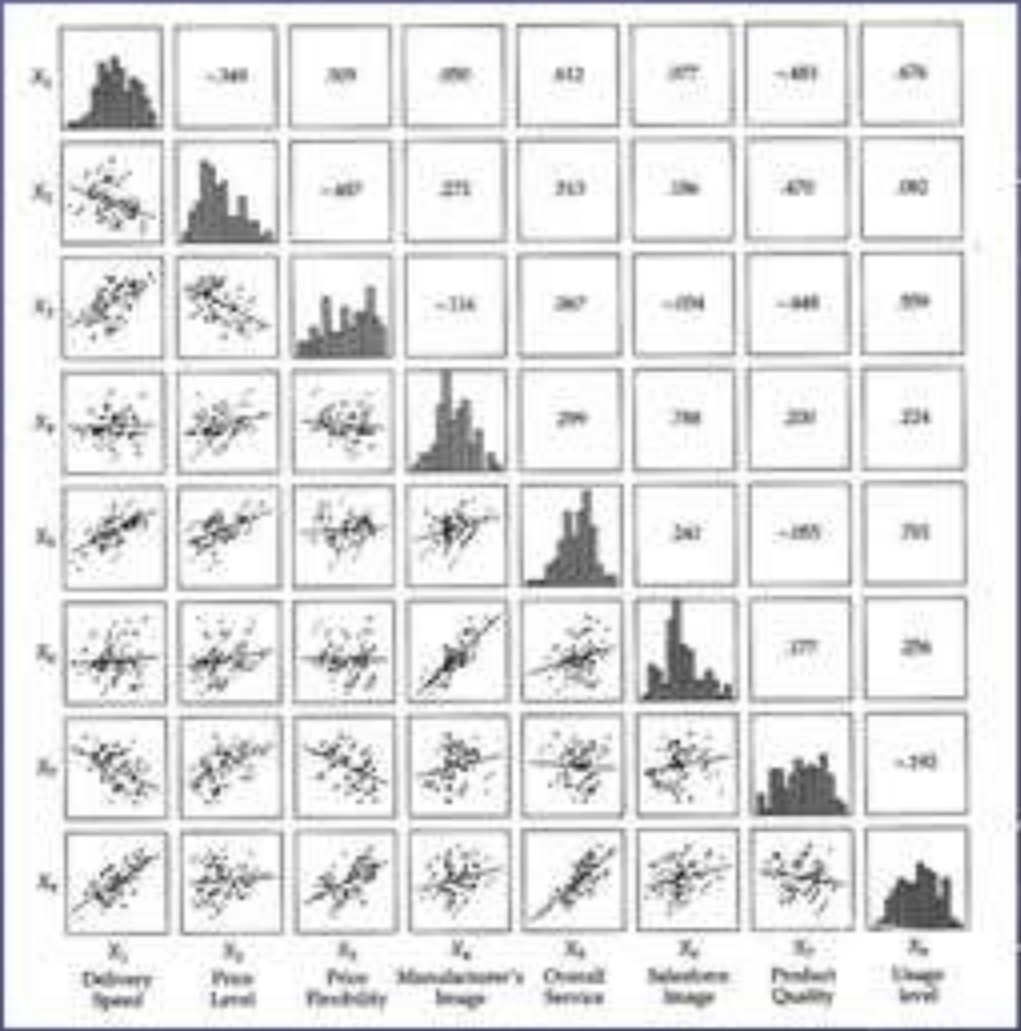
- Two variables for x- and y-axis
- Other variables can be represented by
  - dot size, square size
  - height of rectangle
  - width of rectangle
  - color



Scatterplot matrix

- Also named Draftsman's display
- Histograms on diagonal
- Scatterplot on lower portion
- Correlations on upper portion

Scatterplot matrix (cont…)



correlations

histograms

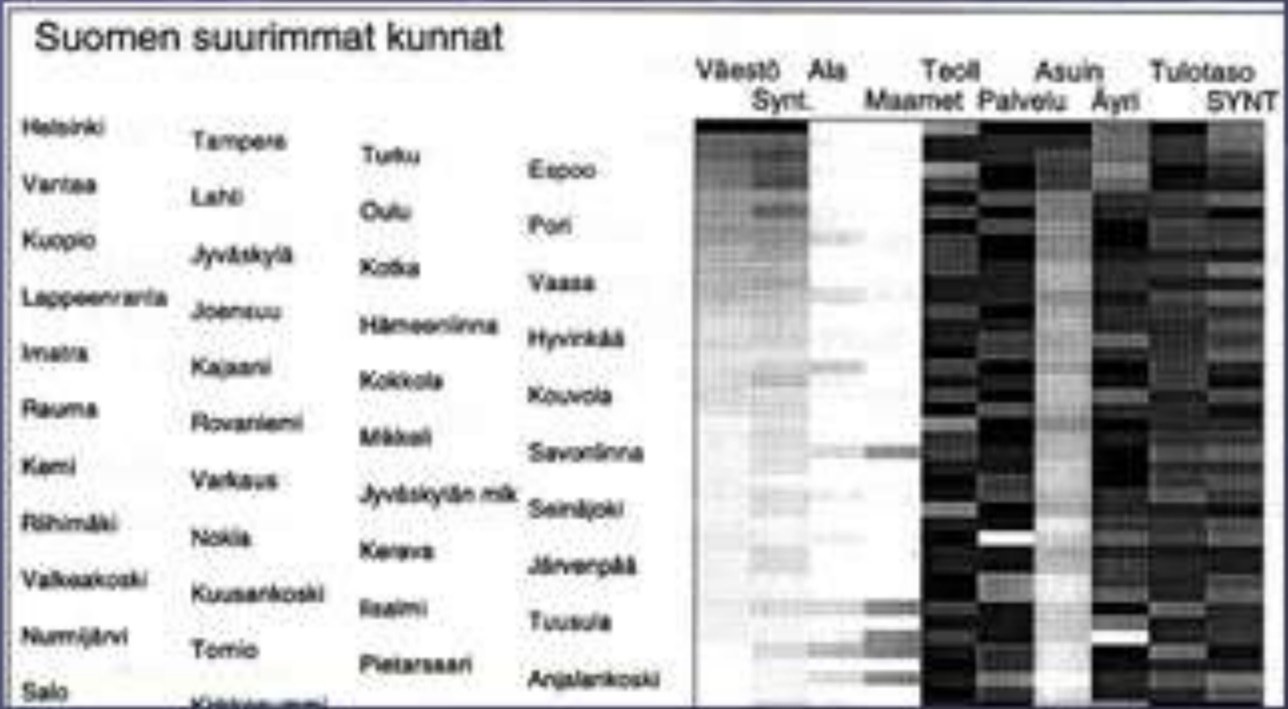
scatterplots

Scatterplot matrix (cont…)

- Shows relations between each variable pair
- Does not determine common distribution exactly
- A good mean to learn new material
- Helps when finding variable transformations

Scatterplot matrix as rasterplot

- Color level represents the value
  - e.g. values are mapped to gray levels 0-255

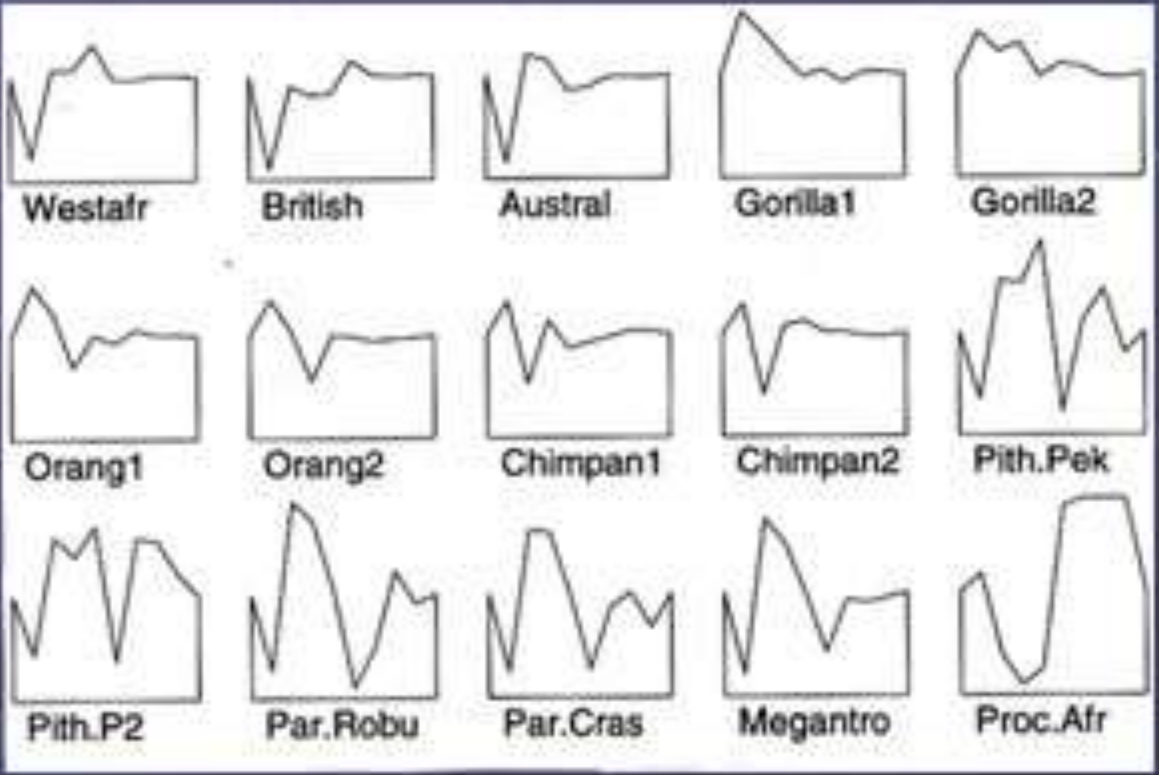


Multivariate profiles

- The objective of the multivariate profiles is to portray the data in a manner that enables each identification of differences and similarities.
- Line diagram
  - Variables on x-axis
  - Scaled (or mapped) values on y-axis

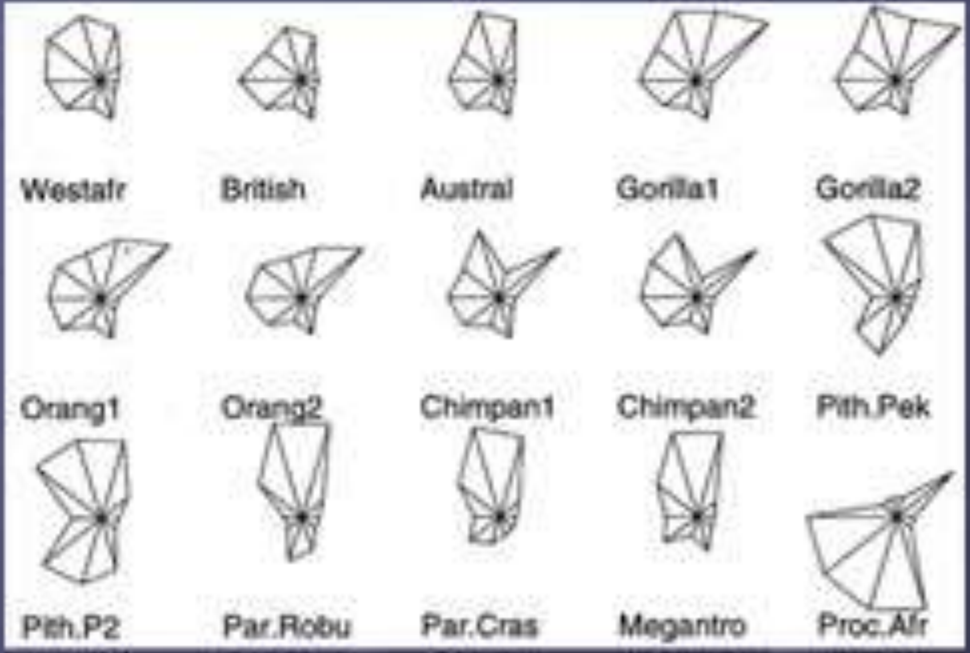
Multivariate profiles (cont…)

- A diagram for each measurement (or measurement group)



Star picture

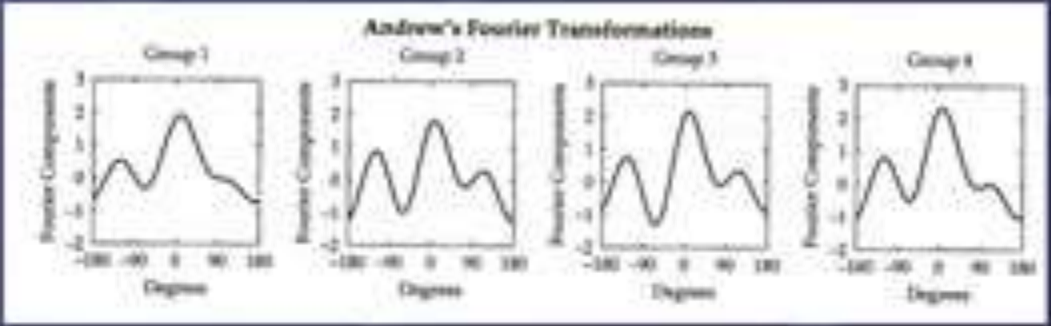
- Like multivariate profile, but drawn from a point instead of x-axis
- Vectors have constant angle



Andrews' Fourier transformations

- D.F. Andrews, 1972.
- Each measurement X = (X_1, X_2,..., X_p) is represented by the function below, where -π < t < π.

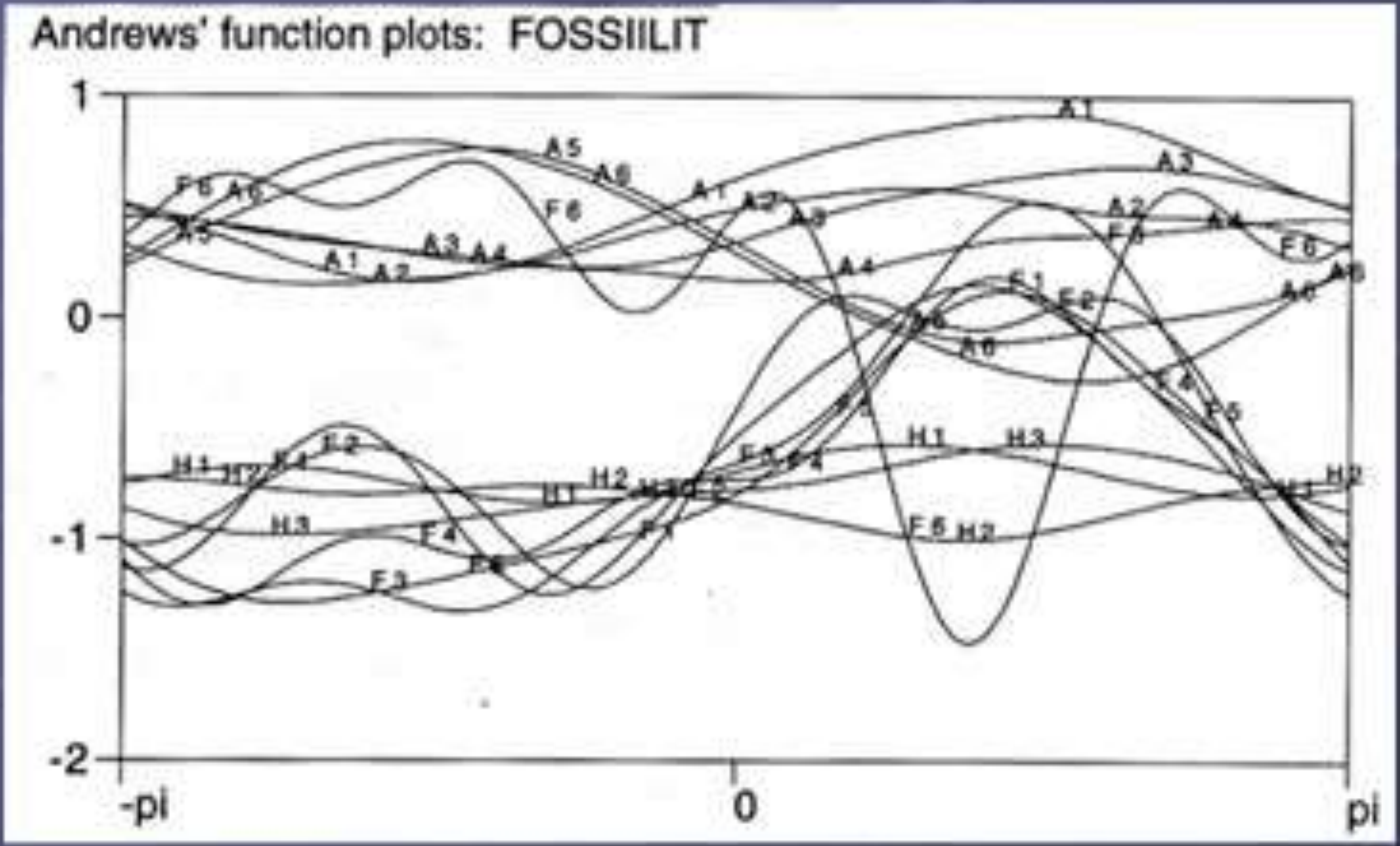
$$f_x(t) = \frac{X_1}{\sqrt{2}} + X_2 \sin(t) + X_3 \cos(t) + X_4 \sin(t) + X_5 \cos(t) + \ldots$$



Andrews' Fourier transformations (cont…)

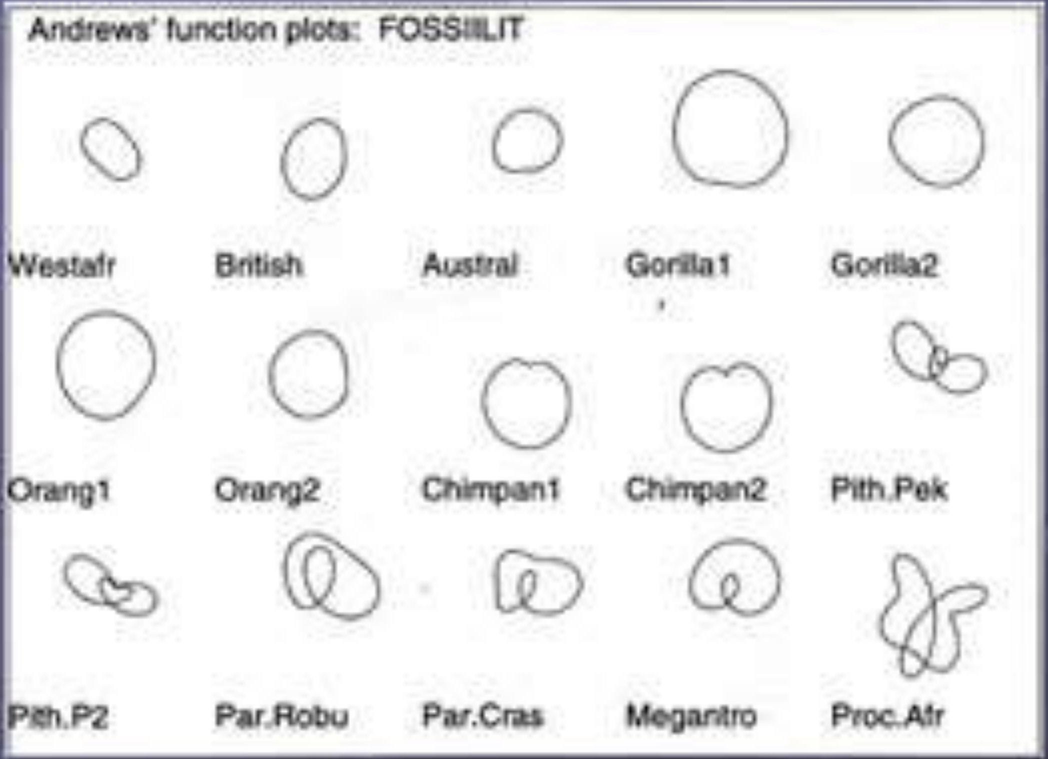
- If several measurements are put into the same diagram similar measurements are close to each other.
- The distance of curves is the Eucledean distance in p-dim space
- Variables should be ordered by importance

Andrews' Fourier transformations (cont…)



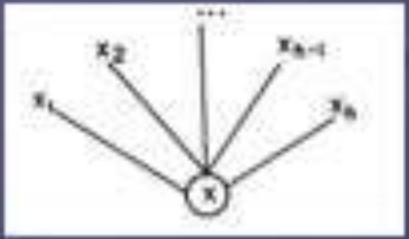
Andrews' Fourier transformations (cont…)

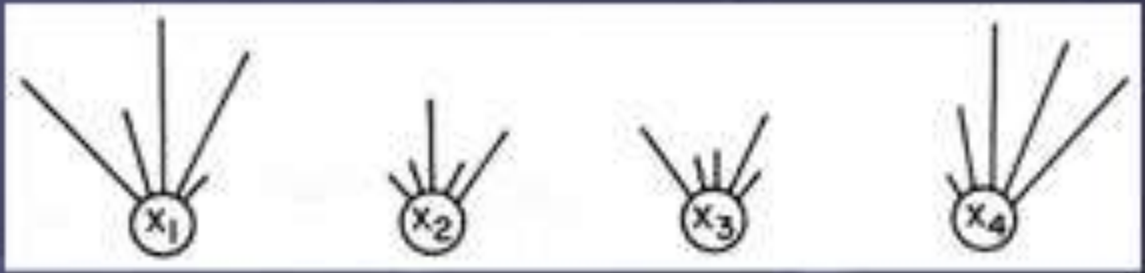
- Can be drawn also using polar coordinates



Metroglyphs (Andersson)

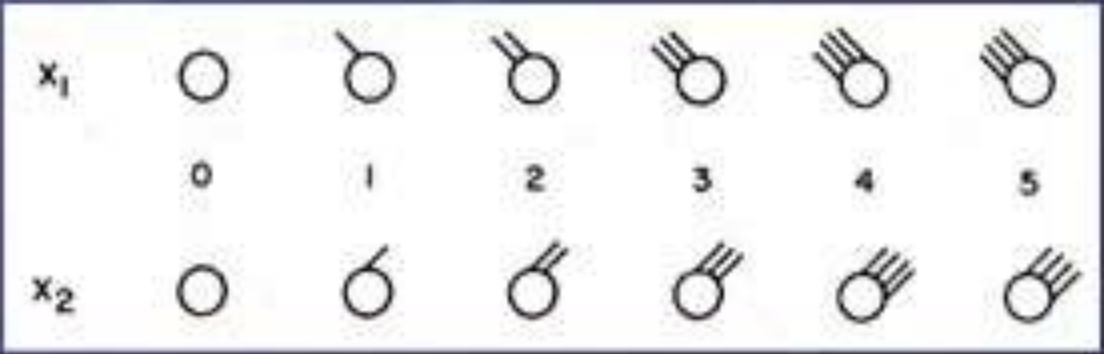
- Each data vector (X) is symbolically represented by a metroglyph
- Consists of a circle and set of h rays to the h variables of X.
- The lenght of the ray represents the value of variable





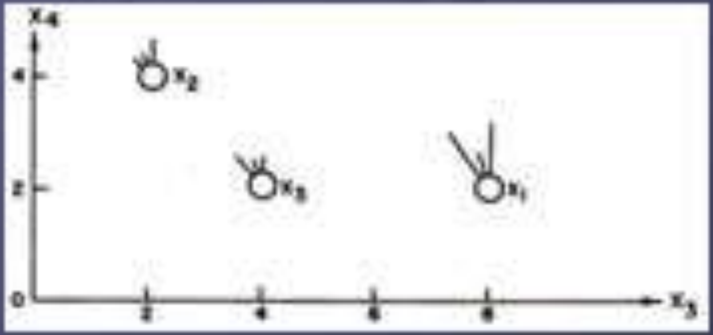
Metroglyphs (cont...)

- Normally rays should be placed at easily visualized and remembered positions
- Can be slant in the same direction
  - the better way if there is a large number of metrogyphs



Metroglyphs (cont...)

- Theoretically no limit to the number of vectors
- In practice, human eye works most efficiently with no more than 3-7 rays
- Metroglyphs can be put into scatter diagram => removes 2 vectors

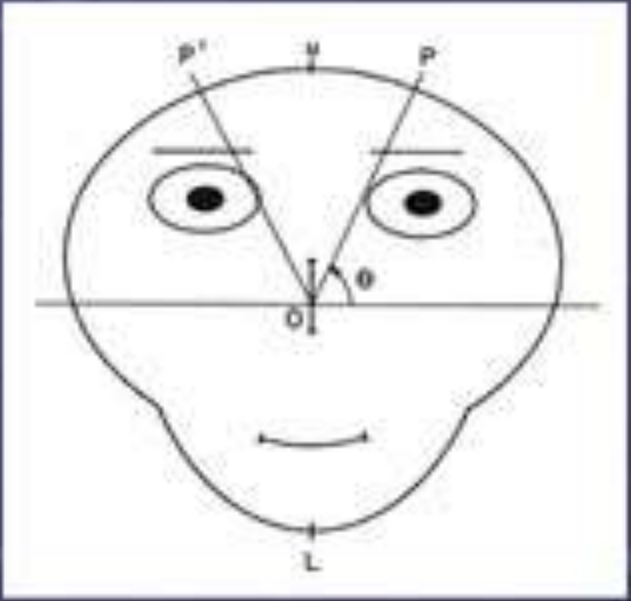


Chernoff's faces

- H. Chernoff, 1973
- Based on the idea that people can detect and remember faces very well
- Variables determine the face features with linear transformation

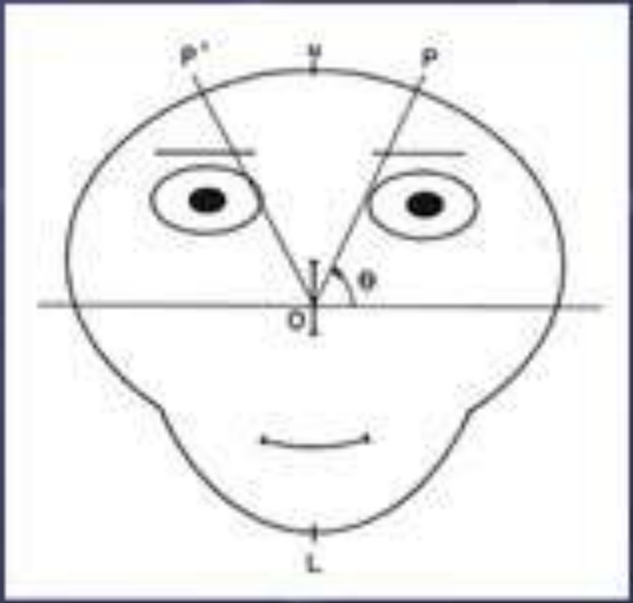
Chernoff’s faces (cont…)

- Originally 18 features
  - Radius to corner of face OP
  - Angle of OP to horizontal
  - Vertical size of face OU
  - Eccentricity of upper face
  - Eccentricity of lower face
  - Length of nose
  - Vertical position of mouth
  - Curvature of mouth 1/R
  - Width of mouth

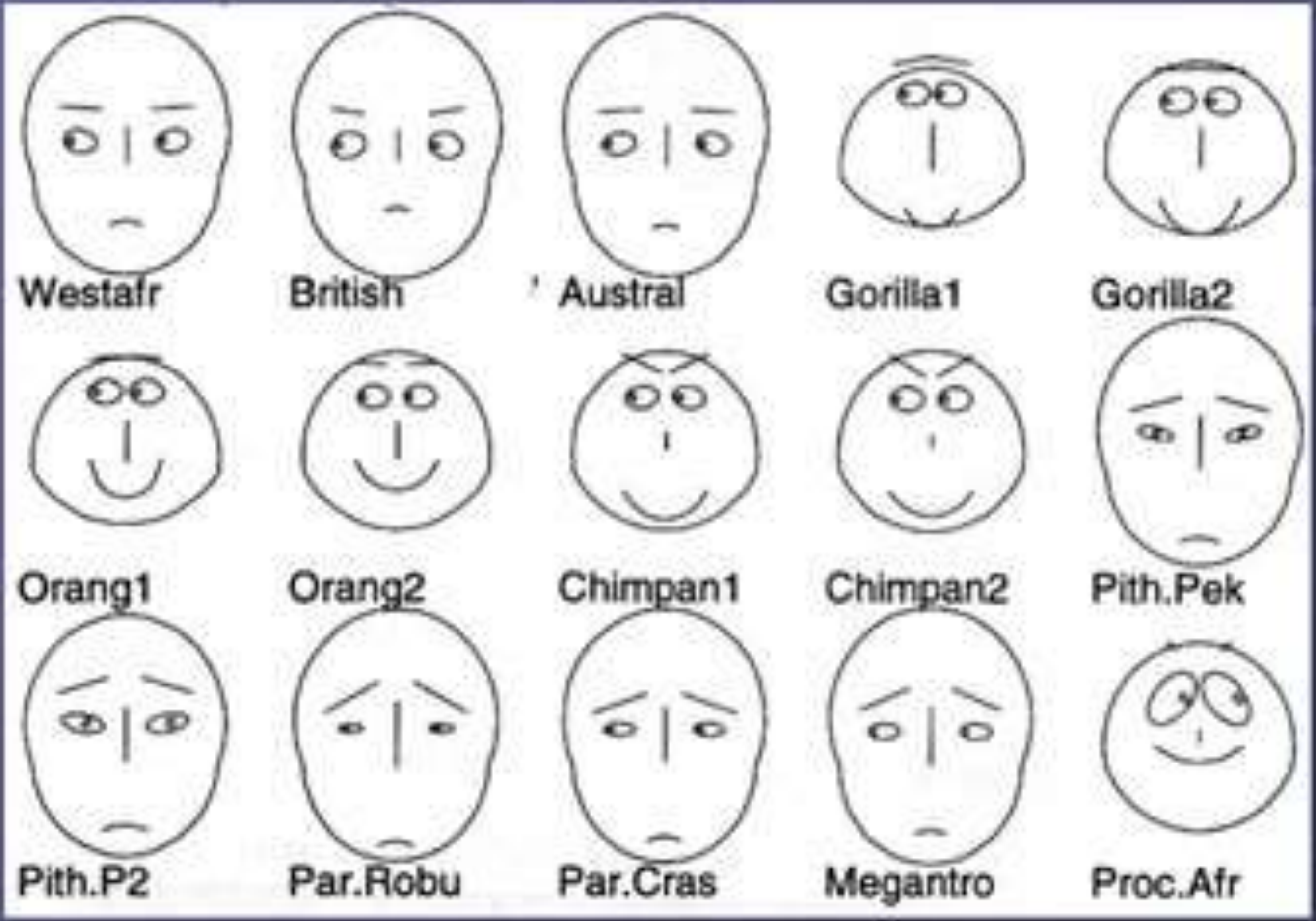


Chernoff's faces (cont…)

- Face features (cont…)
  - Vertical position of eyes
  - Separation of eyes
  - Slant of eyes
  - Eccentricity of eyes
  - Size of eyes
  - Position of pupils
  - Vertical position of eyebrows
  - Slant of eyebrows
  - Size of eyebrows



Chernoff's faces (cont…)



Conclusion

- Graphical Examination eases the understanding of variable relationships
- "Even badly designed image is easier to understand than data matrix."
- "A picture is worth of a thousand words"

Summary

- Data preparation looked at here has dealt with data in the form collected in mainly corporate databases.
- Clearly this is the focus of modern data mining.
- Overview of what we learned

Introduction to Data Preparation for Data Mining

- Motivation and Importance
- Data Preparation as a Part of the Data Mining Process
- Exploring the Problem Space
- Exploring the Solution Space
- The Nature of the World and How It Impacts Data Preparation

Next

- Data Preparation as a Process
  - Inputs, Outputs, Models and Decisions
  - Modelling Tools and Data Preparation
  - Stages of Data Preparation
  - Overview of Basic Data Preparation Techniques

Followed by

- Basic Data Preparation
  - Introductory example
  - Obtaining the Data
  - Data Characterization
  - Data Assembly
  - Example

Next

- Sampling
- Variability
- Confidence
- Numeric vs. Nominal Attributes
  - Dealing with Numeric Variables

The Thick of Things

- Handling non-numeric variables
- Normalizing variables
- Redistributing variables
- Replacing missing values
- Replacing empty values

Last

- Other Methods for Handling Alphas
  - Dimensionality Reduction
  - Multidimensional Scaling
- Let’s Do It Ourselves Considerations
  - Preparing the Dataset vs. Data

Finally

- Bringing it All Together
  - Data Survey
  - Preparing the Dataset
  - Using the Prepared Data
- Graphical Examination of Data

The Future of Data Preparation

- Development of automated data preparation tools
- Specific areas of interest:
  - Series variables
  - Text mining
  - www data
  - Multimedia data

Let's Do It Ourselves!

Assignment IV

(Final Assignment)