RWTH Aachen University
Chair of Computer Science VI
Prof. Dr.-Ing. Hermann Ney

Seminar Data Mining WS 2003/2004

Preprocessing and Visualization

Jonathan Diehl

January 19, 2004

Preprocessing and Visualization

- Introduction
- Theoretical Fundamentals
- Visualization
- Preprocessing

Literature

- D. Hand, H. Manila, P. Smyth: Principles of Data Mining. MIT Press, Cambridge, MA, 2001, Chapters 2 and 3
- J. Han, M. Kamber: Data Mining: Concepts and Techniques. Academic Press, San Diego, CA, 2001, Chapter 3
- S. Roweis, L.Saul: Locally Linear Embedding Homepage. http://www.cs.toronto.edu/∼roweis/lle/

Overview

Data Preprocessing:

- data manipulation prior to mining
- improves quality or speed of actual mining

Data Exploration:

- combined human and computer analysis
- utilizes human's natural abilities

Data Visualization:

- methods to display data to the human
- data is plotted graphically

Preprocessing and Visualization

- Introduction and Motivation
- Theoretical Fundamentals
- Visualization
- Preprocessing

Data Representation

- **data consists of N samples (measures, etc.) and M attributes (variables, dimensions, etc.)**
- **for each sample n and each attribute m there exists one data value $x_{n,m}$**

Data Representation

Data Set:

$$\boldsymbol{X} := \Big(\boldsymbol{X}(1), \dots, \boldsymbol{X}(N)\Big)^{\mathrm{T}} \quad \textbf{with } \boldsymbol{X}(n) = (\boldsymbol{x}_{n,1}, \dots, \boldsymbol{x}_{n,M})$$

Data Matrix:

$$\boldsymbol{X} := \begin{pmatrix} \boldsymbol{x}_{1,1} & \dots & \boldsymbol{x}_{1,M} \\ \vdots & & \vdots \\ \boldsymbol{x}_{N,1} & \dots & \boldsymbol{x}_{N,M} \end{pmatrix}$$

with $x_{n,m}$ value of n-th sample and m-th attribute

Measures of Location

Sample Mean:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} X(n)$$ **with X set of N data values**

Median:

50% of values below median

Mode:

data point which occurs most often

Measures of Dispersion

Variance:

$$\hat{\sigma}^2 = \frac{1}{N} \cdot \sum_{n=1}^{N} \left(X(n) - \mu\right)^2$$ **with X set of N data values and mean μ**

Quartiles:

$25\%/75\%$ **of values below quartile**

Range:

difference between lowest and highest data value

Measures of Correlation

Covariance (of attributes A and B):

$$\mathrm{Cov}(A, B) = \frac{1}{N} \cdot \sum_{n=1}^{N} \big(A(n) - \mu_A\big) \cdot \big(B(n) - \mu_B\big)$$

Covariance Matrix:

$$V = \frac{1}{N} \cdot X^{\mathrm{T}} X$$ **with X zero-centered data matrix**

Correlation Coefficient:

$$\rho_{A,B} = \frac{\mathrm{Cov}(A, B)}{\sigma_A \cdot \sigma_B}$$

Modelling Data

Linear Regression (for two-dimensional data set):

$Y = a + b \cdot X$ **with** a, b **coefficients and** X, Y **attributes**

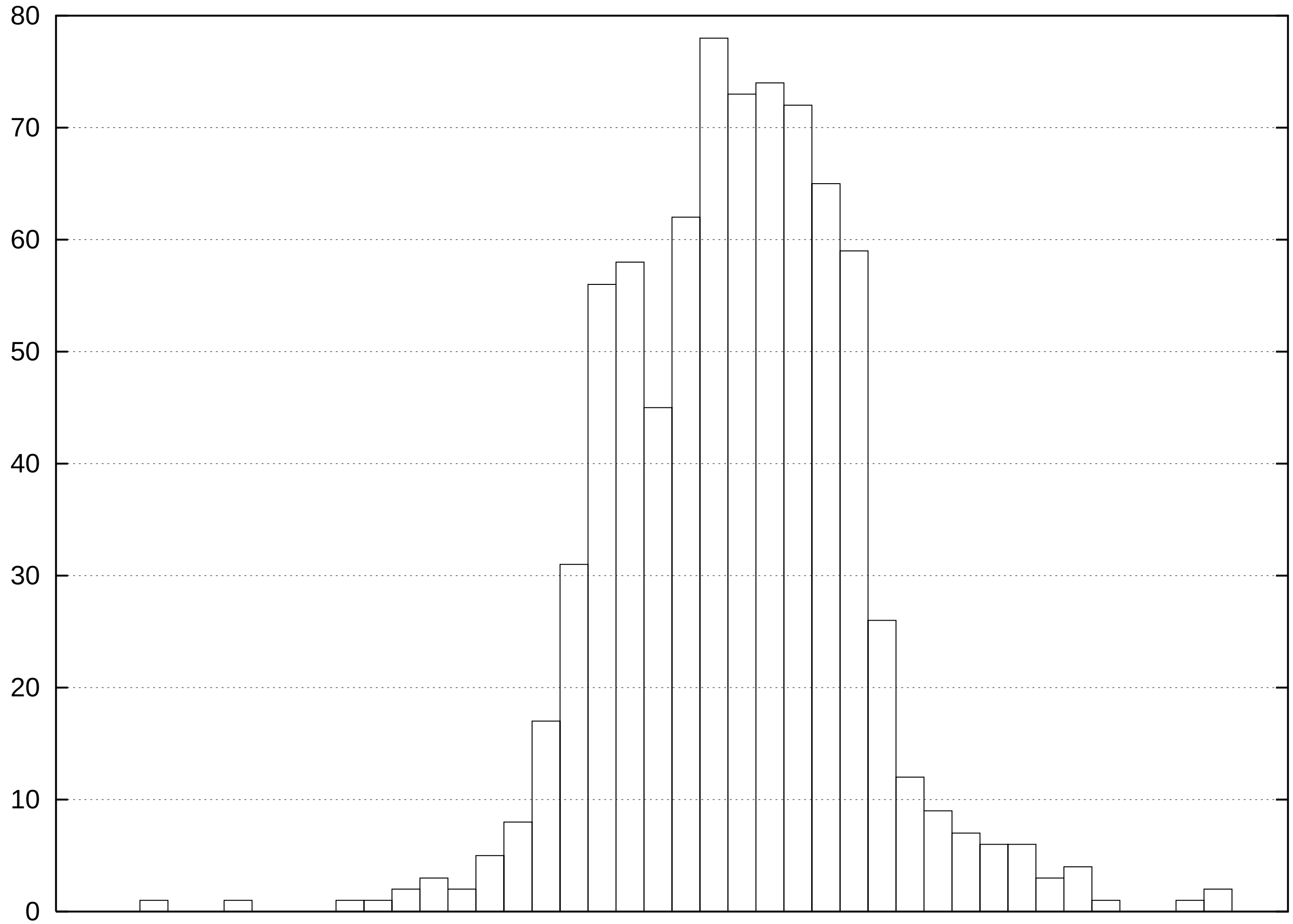
Multiple Linear Regression:

$$Y = a + \sum_{n=1}^{N} b_n \cdot X_n$$

Preprocessing and Visualization

- Introduction and Motivation
- Theoretical Fundamentals
- Visualization
- Preprocessing

Histogram



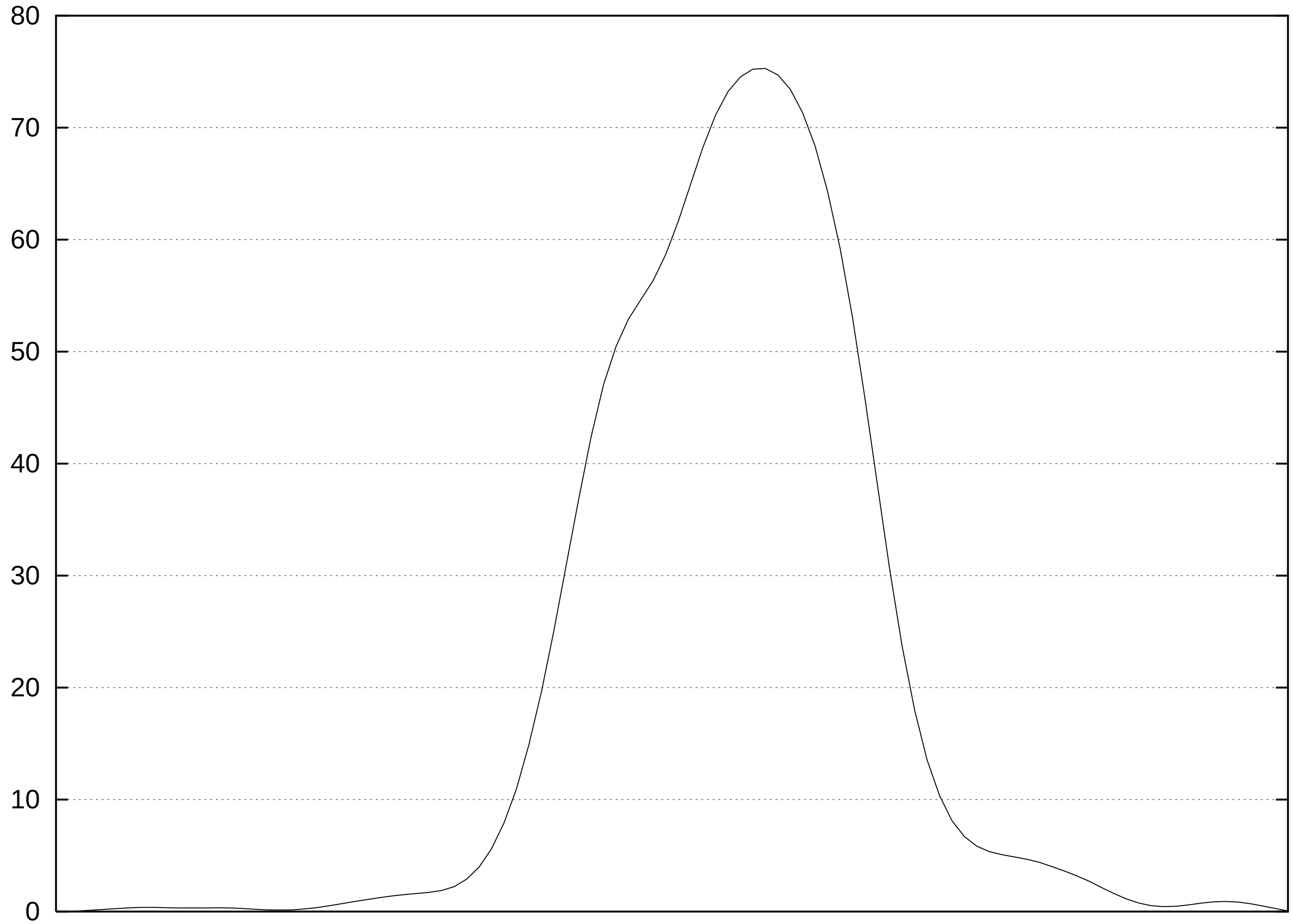
Kernel Method

Estimated density of kernel function K with bandwidth h:

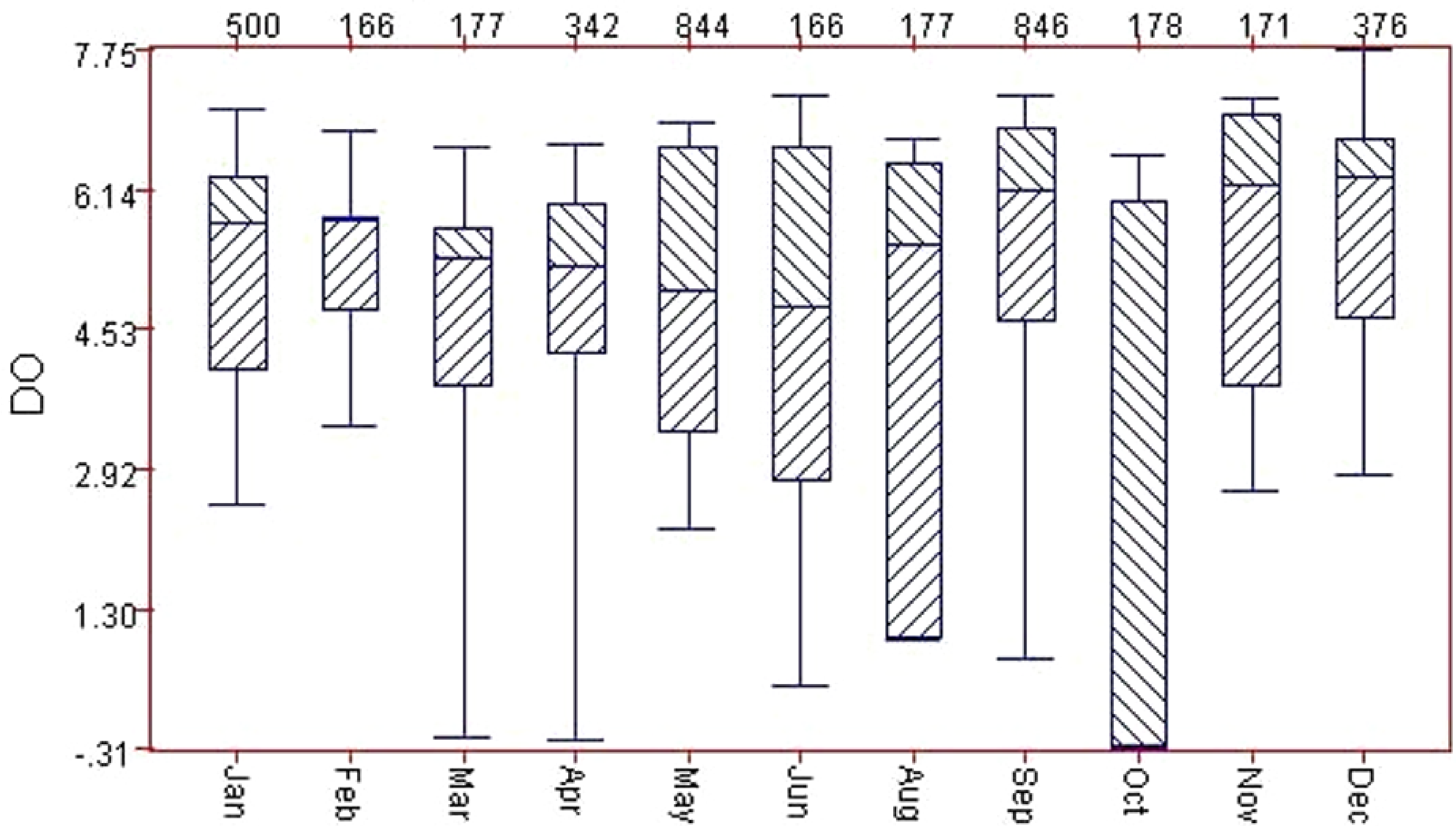
$$f(x) = \frac{1}{N} \sum_{n=1}^{N} K\left(\frac{x - X(n)}{h}\right)$$ for given data set X of N values

where $\int K(t)dt = 1$

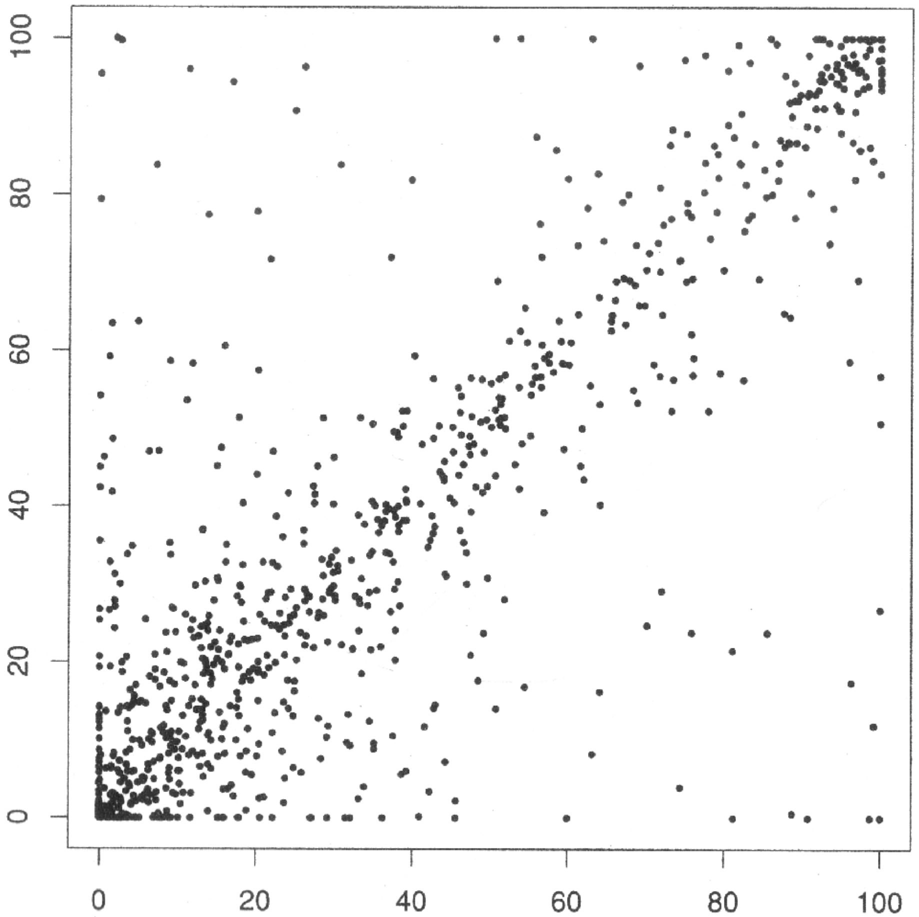
Kernel Method



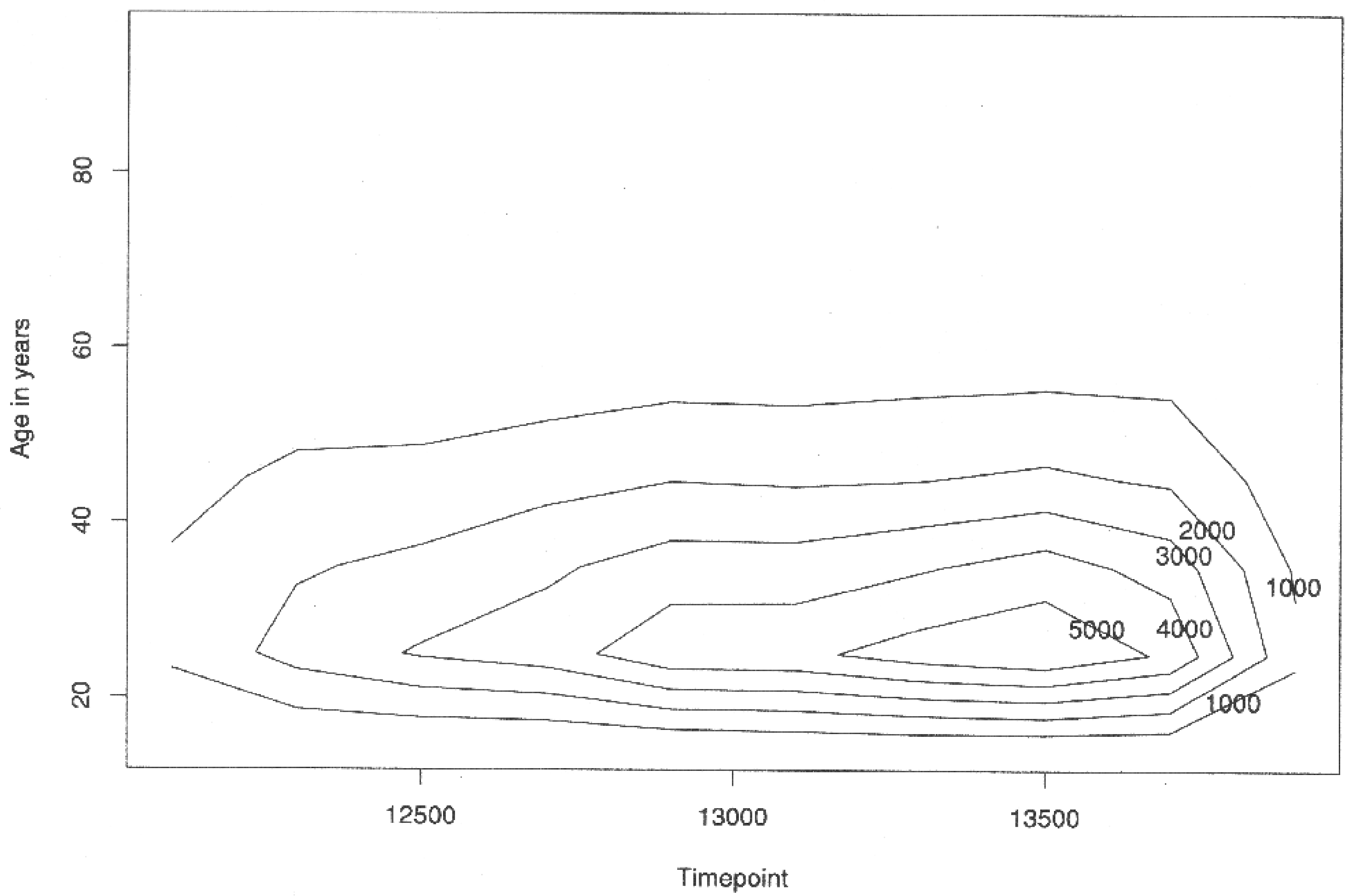
Boxplot


500
166
177
342
844
166
177
846
178
171
376
7.75
6.14
4.53
2.92
1.30
-.31
DO
Jan
Feb
Mar
Apr
May
Jun
Aug
Sep
Oct
Nov
Dec

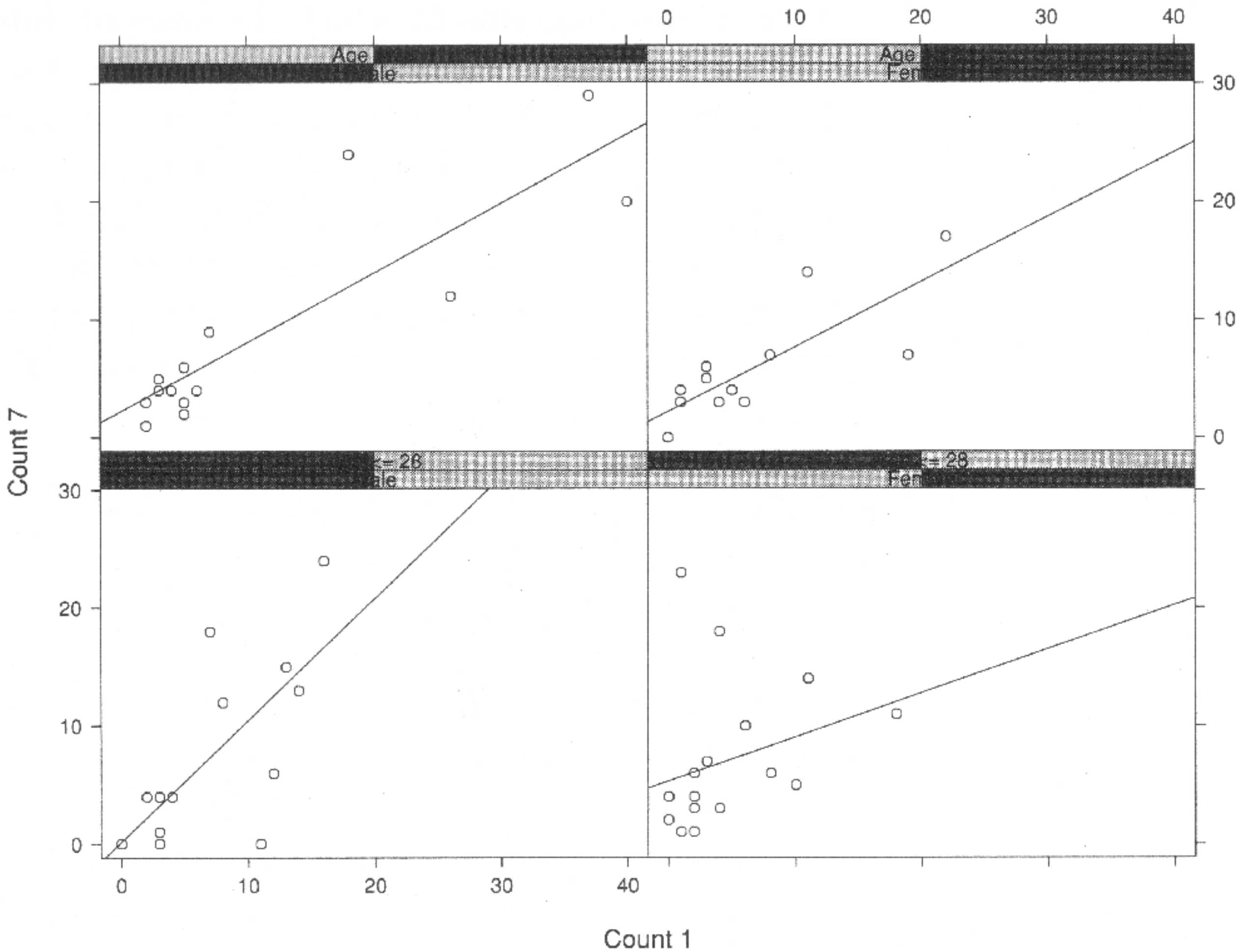
Scatterplot



Contour Plot



Trellis Plot



Preprocessing and Visualization

- Introduction and Motivation
- Theoretical Fundamentals
- Visualization
- Preprocessing

Data Cleaning

Problems:

- incomplete data (missing values)
- erroneous (noisy) data
- inconsistent data ($\rightarrow$ data integration)

Missing Values

Solutions:

- ignore samples → loose important information
- determine most likely value:
  1. make use of statistical measures (mean/median, class mean/median)
  2. construct model of attribute relations (regression) and calculate value
  3. construct decision tree and derive value

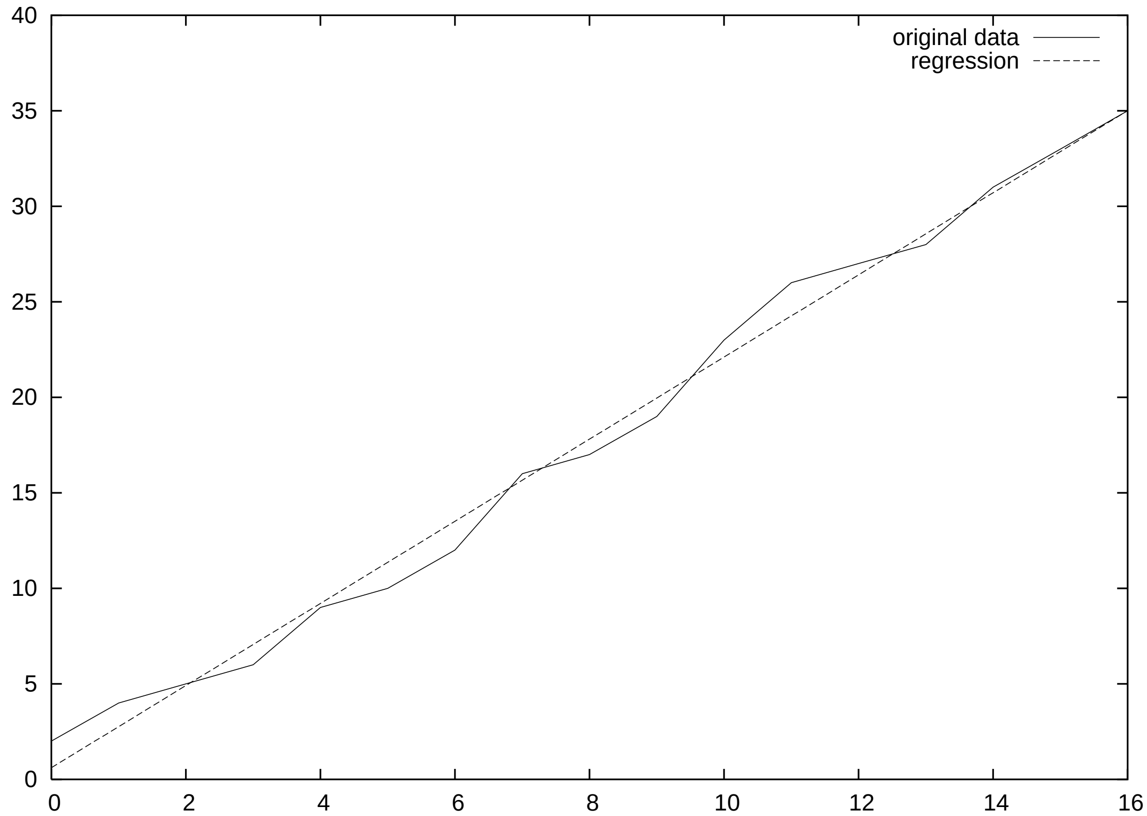
  ...

Noisy Data: Regression

1. construct model of entire data set:
   - linear Regression
   - multiple Regression
     (regression over weighted linear combination of attributes)
2. calculate data points from regression equation

⇒ global smoothing, strong data reduction

Noisy Data: Regression



Data Integration

Problems:

- entity identification problem ($\rightarrow$ metadata)
- data redundancy ($\rightarrow$ correlation analysis)
- value conflicts ($\rightarrow$ data transformation)

Data Transformation

Problem:

- **data has to meet certain criteria before further processing**

e.g. attribute values must be weighted equally for many analysis methods

Normalization

Fit attribute A of data X into predefined range (e.g. $[0.0, 1.0]$)

- min-max normalization:

$$f(i) = \frac{A(i) - min_A}{max_A - min_A} \cdot (newmax_A - newmin_A) + newmin_A$$

- z-score normalization:

$$g(i) = \frac{A(i) - \mu_A}{\sigma_A}$$

Attribute Construction

Construct new (redundant) attributes summarizing stored information

- sums, products of samples/data classes

⇒ Fast online access to summarized data

Data Reduction

Problem:

- **huge databases with many attributes**

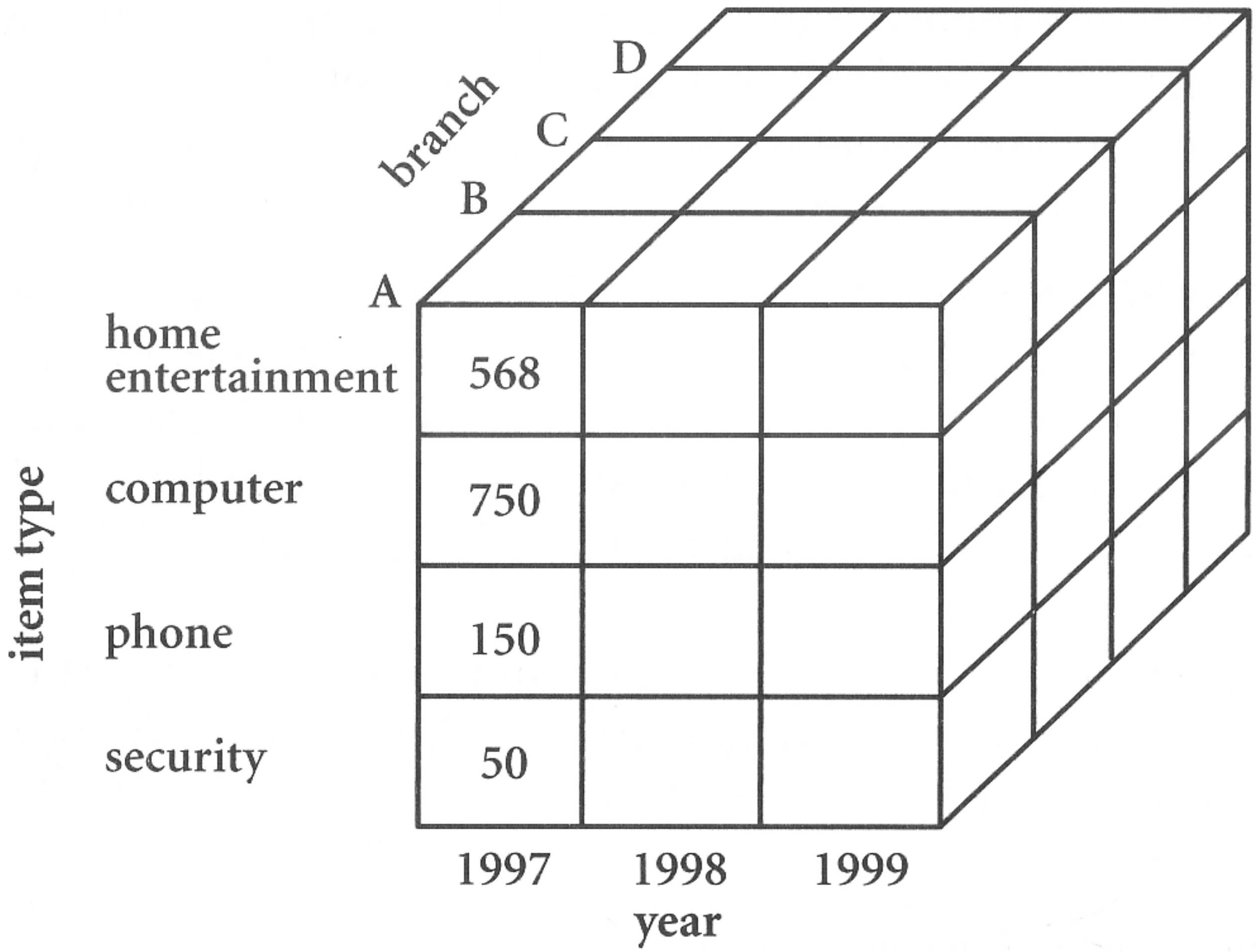
⇒ very slow mining process

Data Cube Aggregation

- multidimensional arrangement of aggregated information
- dimensions (axes) represent attributes
- multiple levels of abstraction possible (concept hierarchy)

⇒ Efficient organization of summarized data

Data Cube Aggregation



Attribute Subset Selection

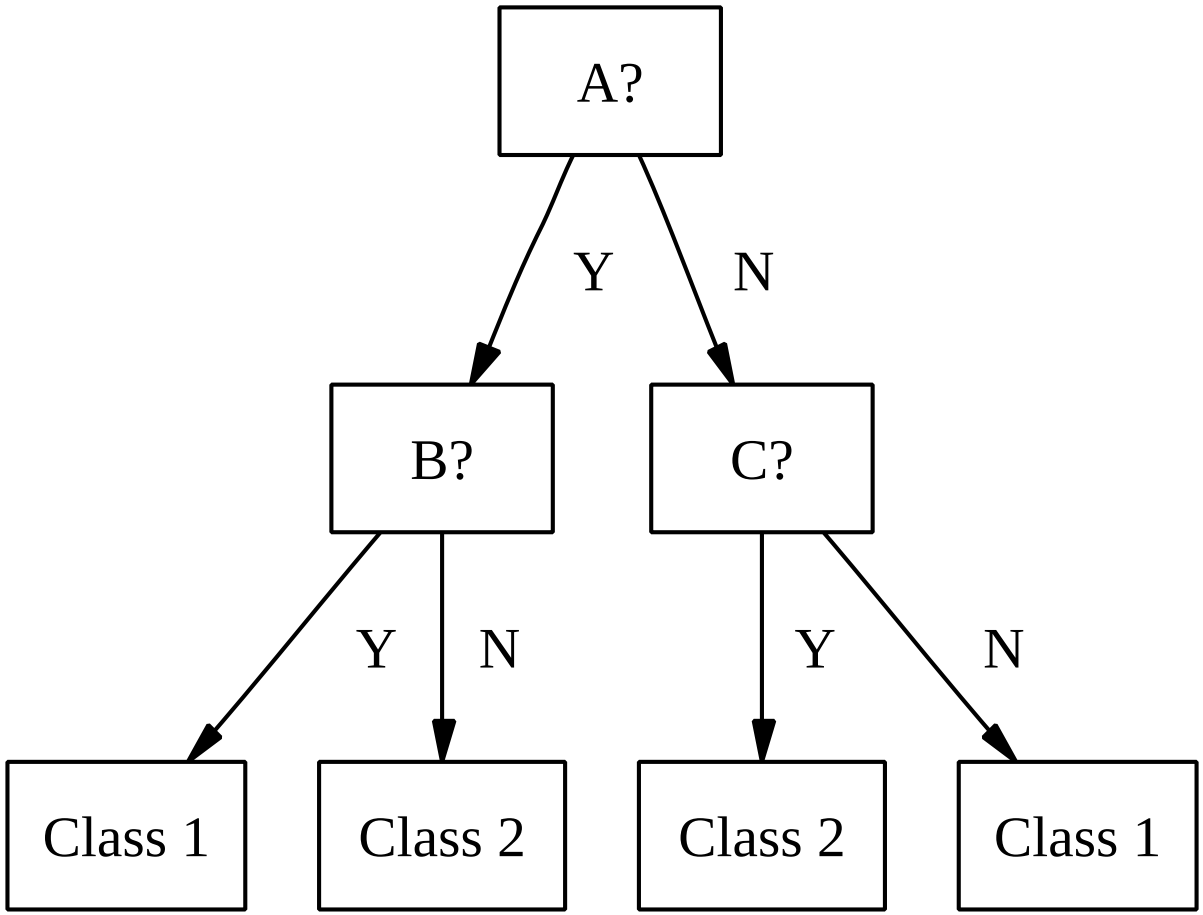
Determine significance of attributes:

- correlation coefficient
- information gain

Make selection:

- stepwise selection/elimination
- decision tree induction

Attribute Subset Selection



Principal Component Analysis

Aim: Maximize variability

- project data onto principal components (linear combination of attributes)
- normalize principal components
- zero-centered data set (subtract mean from all values)

Principal Component Analysis

Variability of projection vector a ($S := X^{\mathrm{T}}X$ scatter matrix):

$$
\begin{aligned}
s_a^2 &= \left(Xa\right)^{\mathrm{T}} \cdot \left(Xa\right) \\
&= a^{\mathrm{T}}X^{\mathrm{T}}Xa \\
&= a^{\mathrm{T}}Sa
\end{aligned}
$$

Impose normalization constraint and maximize:

$$
\begin{aligned}
u &= a^{\mathrm{T}}Sa - \lambda(a^{\mathrm{T}}a - 1) \quad \rightarrow max \\
\frac{\delta u}{\delta a} &= \underbrace{2Sa - 2\lambda a = 0}_{\text{eigenvalue form}}
\end{aligned}
$$

⇒ eigenvector a with largest eigenvalue is first principal component

Principal Component Analysis

- data is projected onto the first K eigenvectors
- small values of K sufficient because late principal components are insignificant (eigenvalue approaches zero)
- for $K = 2$ principal component analysis can be used to project data into a plane for visualization

Multidimensional Scaling

Aim: Preserve Distances

- **fit multidimensional data into plane**
- **minimize squared distance error**
- **zero-centered data set (subtract mean from all values)**

Multidimensional Scaling

Given $B = XX^{\mathrm{T}}$ solve (euclidian distance)

$$
\begin{aligned}
d_{ij}^2 &= ||x_i - x_j||^2 \\
&= x_i^T X_i - 2x_i^T x_j + x_j^T x_j \\
&= b_{ii} + b_{jj} - 2b_{ij}
\end{aligned}
$$

then minimize

$$
\sum_{i=1}^{N} \sum_{j=1}^{N} (\delta_{ij} - d_{ij})^2
$$

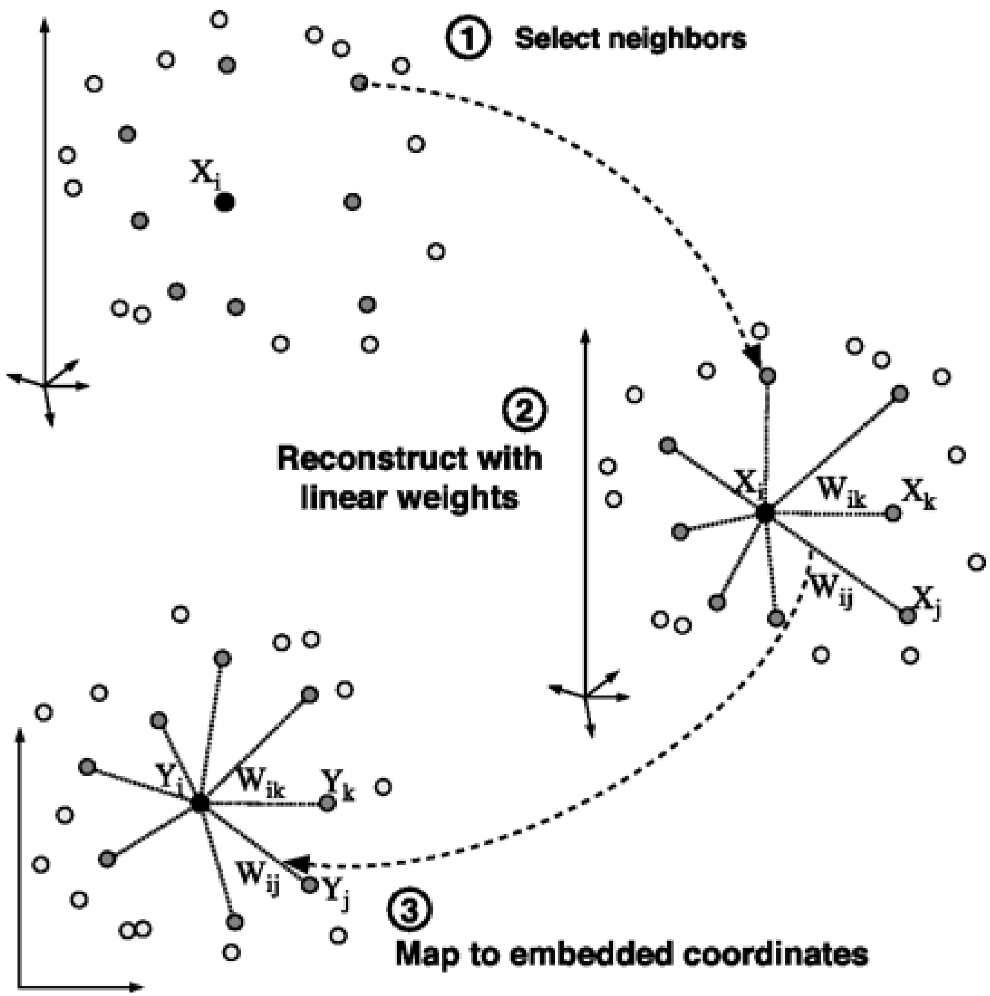
with δ_{ij} multidimensional distance between data point i and j

Locally Linear Embedding

Aim: Preserve Neighborhoods

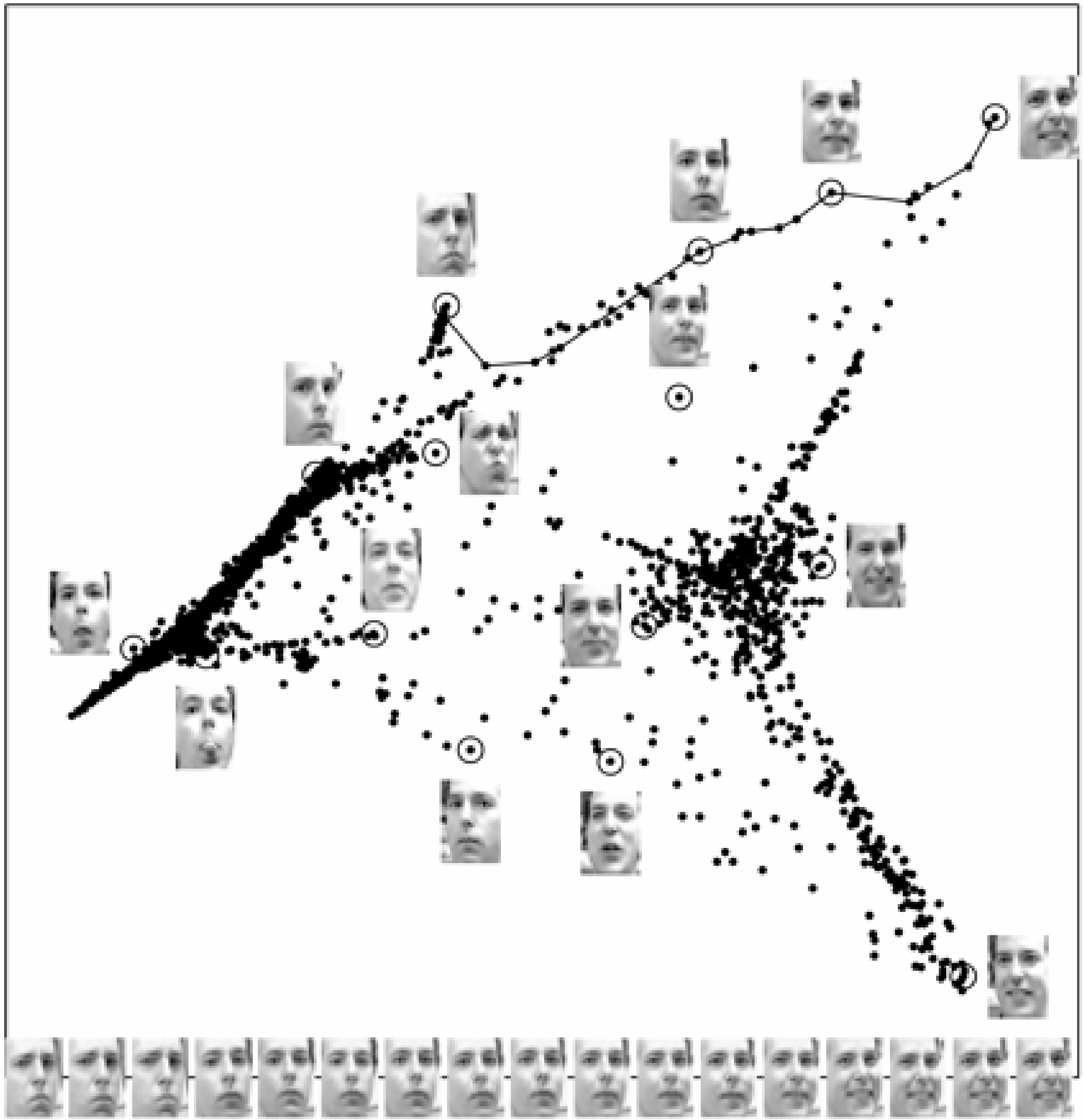
- reduce data to lower dimensional embedding
- find locally linear patches of the data
- reconstruct data points from its neighbors
- invariant to rotations, rescalings and translations

Locally Linear Embedding



Locally Linear Embedding

1. Find neighbors
2. Reconstruct data point as linear combination of neighbors
   $\Rightarrow$ Weight matrix
3. Calculate low-dimensional representation of weight matrix using eigenvector analysis



Summary and Discussion