

Vorlesung

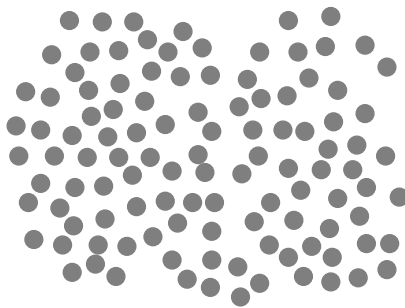
## Statistische Methoden der Datenanalyse

Prof. Dr. Dr. Wolfgang Rhode

### Data-Mining – Teil 1

#### Motivation

- Ziel ist es die Punkte in zwei Populationen zu trennen

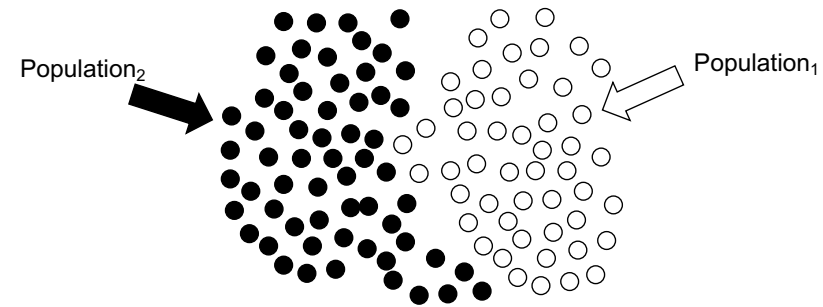


### Data-Mining – Teil 1

- Diskriminanzanalyse
- Grundbegriffe des Data-Minings
- Typischer Aufbau eines Data-Mining-Prozesses
- Datenauswahl
- Datenbereinigung
- Datenreduktion und –transformation
  - Hauptkomponentenanalyse
  - Feature Selection

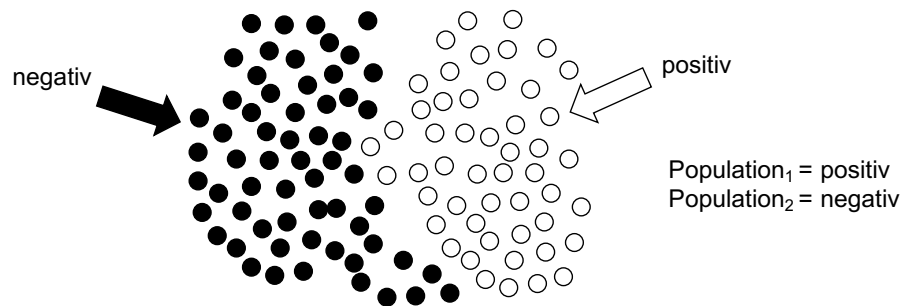
#### Motivation

- Ziel ist es die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt



## Motivation

- Ziel ist es die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt



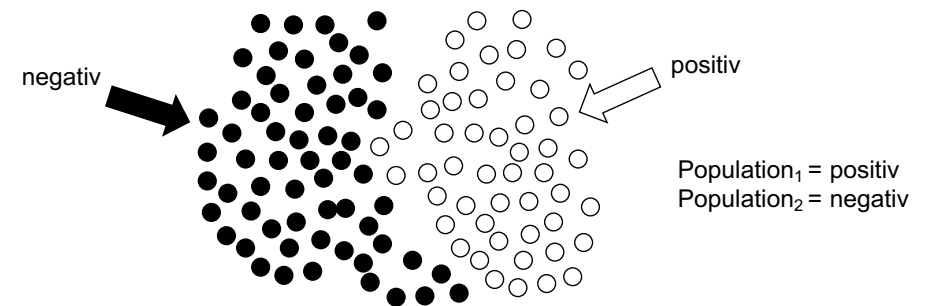
Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Motivation

- Ziel ist es die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt
- Idee: Suche im Monte-Carlo den *besten* eindimensionalen Schnitt



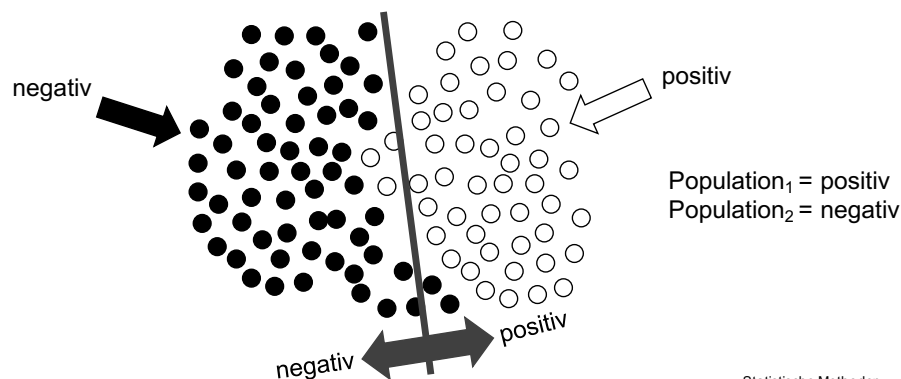
Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Motivation

- Ziel ist es die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt
- Idee: Suche im Monte-Carlo den *besten* eindimensionalen Schnitt

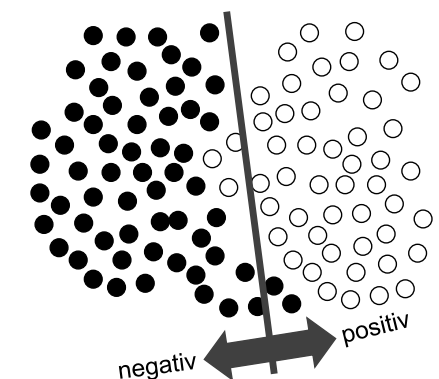


Prof. Dr. Dr. W. Rhode

Statistische Methoden  
der Datenanalyse

## Motivation

- Was ist der beste Schnitt?




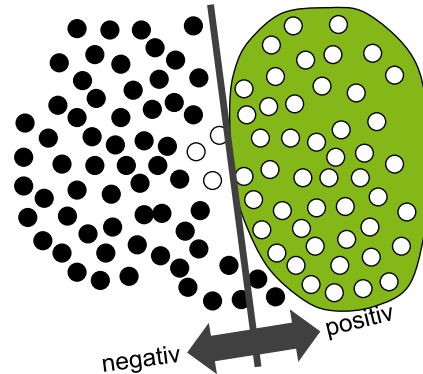
Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1



Statistische Methoden  
der Datenanalyse

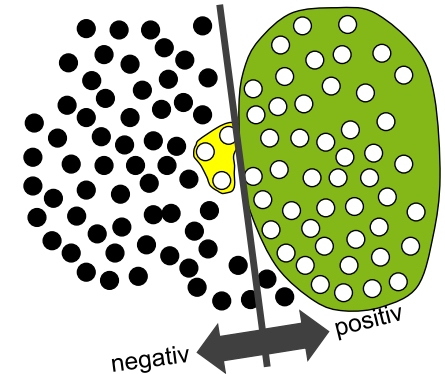
## Motivation

- Was ist der beste Schnitt?
  - true positiv (tp) 
    - "positiv" Elemente die nach der Trennung im "positiv" Bereich liegen






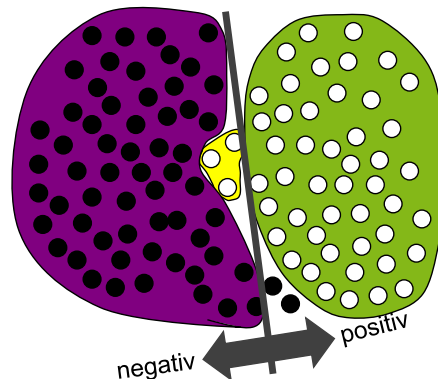
## Motivation

- Was ist der beste Schnitt?
  - true positiv (tp) 
    - "positiv" Elemente die nach der Trennung im "positiv" Bereich liegen
  - false negativ (fn) 
    - "positive" Elemente die nach der Trennung im "negativ" Bereich liegen







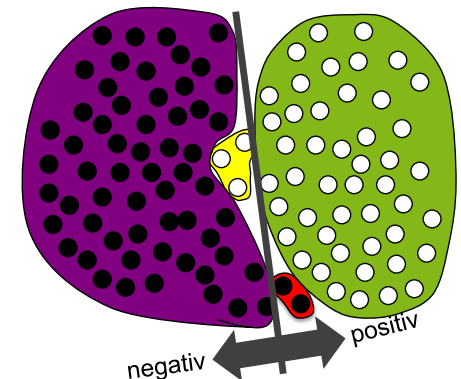
## Motivation

- Was ist der beste Schnitt?
  - true positiv (tp) 
    - "positiv" Elemente die nach der Trennung im "positiv" Bereich liegen
  - false negativ (fn) 
    - "positive" Elemente die nach der Trennung im "negativ" Bereich liegen
  - true negativ (tn) 
    - "negativ" Elemente die nach der Trennung im "negativ" Bereich liegen



## Motivation

- Was ist der beste Schnitt?
  - true positiv (tp) 
    - "positiv" Elemente die nach der Trennung im "positiv" Bereich liegen
  - false negativ (fn) 
    - "positive" Elemente die nach der Trennung im "negativ" Bereich liegen
  - true negativ (tn) 
    - "negativ" Elemente die nach der Trennung im "negativ" Bereich liegen
  - false positiv (fp) 
    - "negativ" Elemente die nach der Trennung im "positiv" Bereich liegen

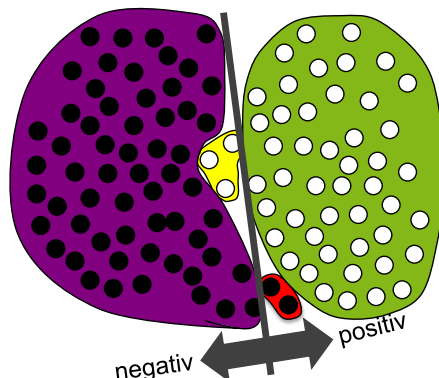


## Motivation

- Was ist der beste Schnitt?
  - Qualitätsmaße für zwei Populationen:
    - Reinheit (bzgl. Population<sub>1</sub>):

$$\text{Reinheit} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

(„precision“)



## Motivation

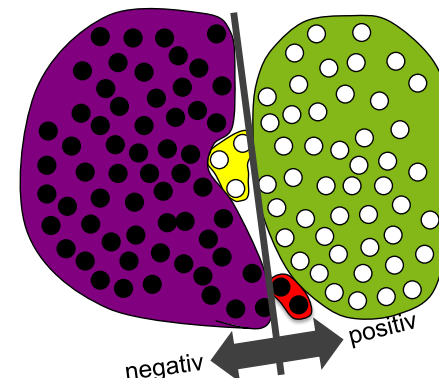
- Was ist der beste Schnitt?
  - Qualitätsmaße für zwei Populationen:
    - Reinheit (bzgl. Population<sub>1</sub>):

$$\text{Reinheit} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

(„precision“)

$$\text{Effizienz} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

(„recall“)



## Motivation

- Was ist der beste Schnitt?
  - Qualitätsmaße für zwei Populationen:
    - Reinheit (bzgl. Population<sub>1</sub>):

$$\text{Reinheit} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

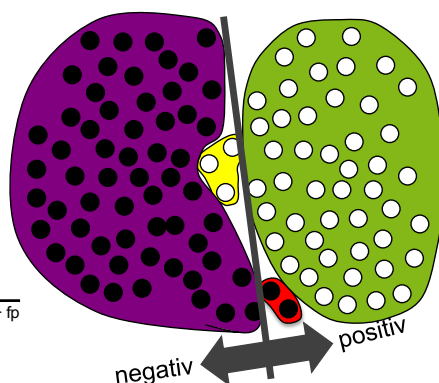
(„precision“)

$$\text{Effizienz} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

(„recall“)

$$\text{Genauigkeit} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}}$$

(„accuracy“)



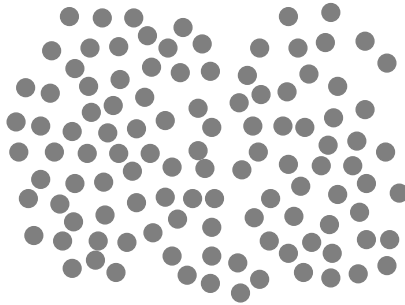
## Motivation

- Was ist der beste Schnitt?
  - Qualitätsmaße für zwei Populationen:

|  |                              | True condition  |   |   |  |
|--|------------------------------|---|---|---|--|
|  |                              | Condition positive  | Condition negative  | Prevalence<br>$= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$                                 |  |
| Predicted condition  | Predicted condition positive | True positive   | False positive<br>(Type I error)  | Positive predictive value (PPV), Precision<br>$= \frac{\sum \text{True positive}}{\sum \text{Test outcome positive}}$ | False discovery rate (FDR)<br>$= \frac{\sum \text{False positive}}{\sum \text{Test outcome positive}}$     |
|  | Predicted condition negative | False negative<br>(Type II error)   | True negative   | False omission rate (FOR)<br>$= \frac{\sum \text{False negative}}{\sum \text{Test outcome negative}}$                 | Negative predictive value (NPV)<br>$= \frac{\sum \text{True negative}}{\sum \text{Test outcome negative}}$ |
| Accuracy (ACC) =<br>$\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$ |                              | True positive rate (TPR), Sensitivity, Recall<br>$= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$ | False positive rate (FPR), Fall-out<br>$= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$        | Positive likelihood ratio (LR+)<br>$= \frac{\text{TPR}}{\text{FPR}}$  | Diagnostic odds ratio (DOR)<br>$= \frac{\text{LR+}}{\text{LR-}}$   |
|  |                              | False negative rate (FNR), Miss rate<br>$= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$         | True negative rate (TNR), Specificity (SPC)<br>$= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$ | Negative likelihood ratio (LR-)<br>$= \frac{\text{FNR}}{\text{TNR}}$  |  |

## Motivation

- Ziel ist es bei die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt
- Idee: Suche im Monte-Carlo den *besten* eindimensionalen Schnitt



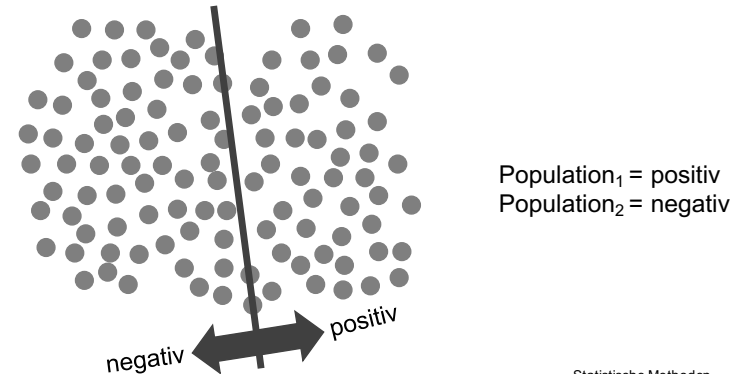
Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Motivation

- Ziel ist es bei die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt
- Idee: Suche im Monte-Carlo den *besten* (n-1)-dimensionalen Schnitt

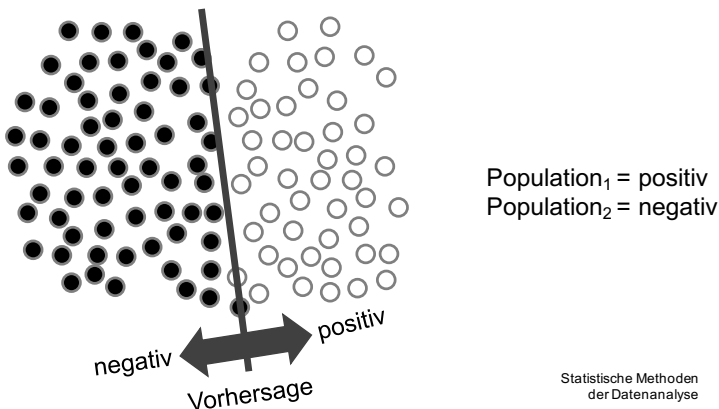


Prof. Dr. Dr. W. Rhode

Statistische Methoden  
der Datenanalyse

## Motivation

- Ziel ist es bei die Punkte in zwei Populationen zu trennen
- Im Monte Carlo ist die Zugehörigkeit der Elemente bekannt
- Idee: Suche im Monte-Carlo den *besten* (n-1)-dimensionalen Schnitt



Prof. Dr. Dr. W. Rhode

Statistische Methoden  
der Datenanalyse

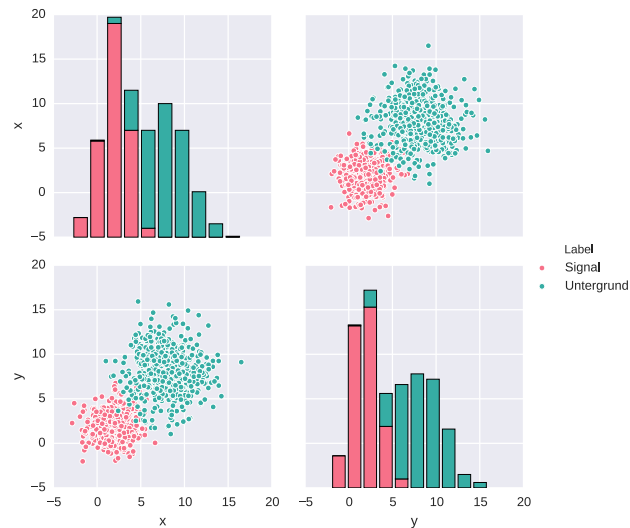
## Beispiel:

- Aufgabe: Trennen zweier Populationen
  - Grün: „Untergrund“
  - Rot: „Signal“
- Elemente beider Populationen über Wertepaare (x, y) beschrieben
  - Untergrund: Gaußverteilung mit dem Mittelwert (8, 8) und der Standardabweichungen (2.5, 2.5)
  - Signal: Gaußverteilung mit dem Mittelwert (2, 2) und der Standardabweichungen (1.5, 1.5)
- Suche den *besten* eindimensionalen Schnitt (trennende Hyperebene)

Data-Mining – Teil 1

Prof. Dr. Dr. W. Rhode

Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Beispiel:

- Aufgabe: Trennen zweier Populationen
  - Grün: „Untergrund“
  - Rot: „Signal“
- Elemente beider Populationen über Wertepaare (x, y) beschrieben
  - Untergrund: Gaußverteilung mit dem Mittelwert (8, 8) und der Standardabweichungen (2.5, 2.5)
  - Signal: Gaußverteilung mit dem Mittelwert (2, 2) und der Standardabweichungen (1.5, 1.5)
- Suche den *besten* eindimensionalen Schnitt (trennende Hyperebene)
  - Projektion auf den Normalenvektor der Hyperebene muss die Klassen *maximal* trennen

Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Lineare Fisher Diskriminanzanalyse

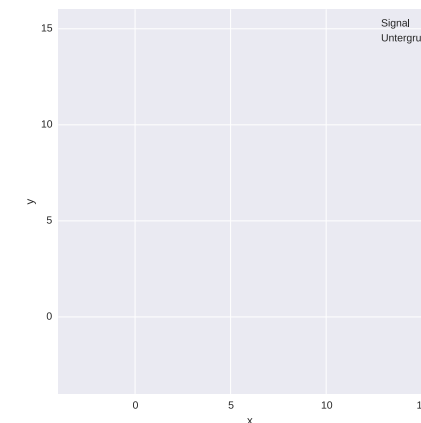
- Um eine gute Projektion  $\vec{\lambda}$  ( $\vec{x}' = \vec{\lambda}^T \vec{x}$ ) zu finden, muss ein Maß für die Trennbarkeit der Klassen definiert werden
  - Erste (naive) Idee:  
Abstand der Mittelwerte der Klassen auf der Projektionsachse

$$D_{\text{naiv}}(\vec{\lambda}) = |\vec{\mu}'_1 - \vec{\mu}'_2| = \left| \vec{\lambda}^T (\vec{\mu}_1 - \vec{\mu}_2) \right|$$

Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

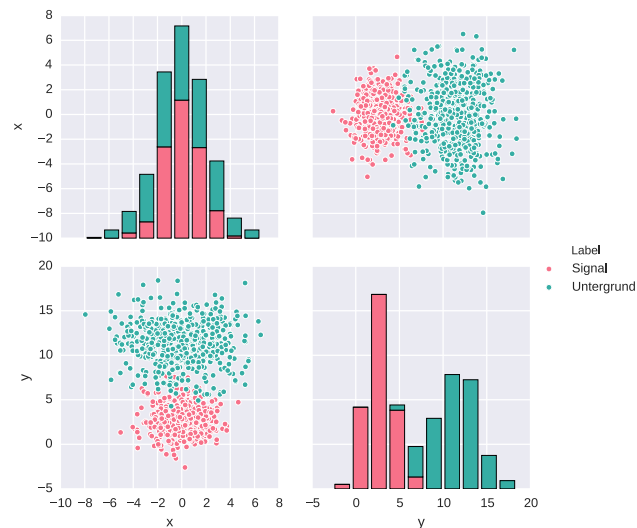
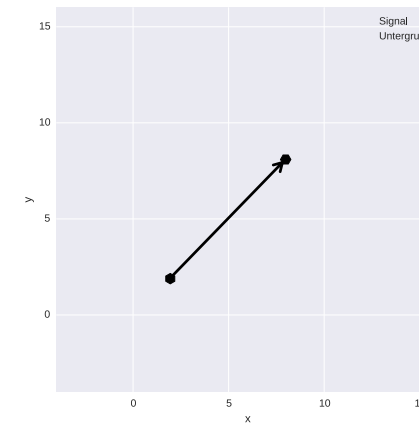
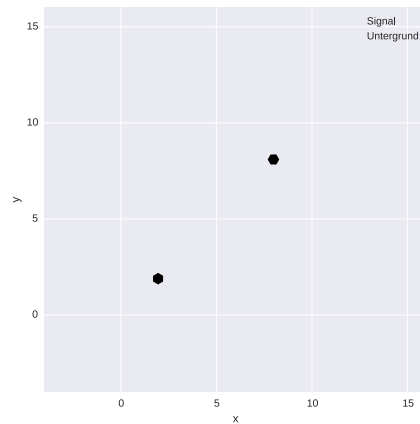
Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

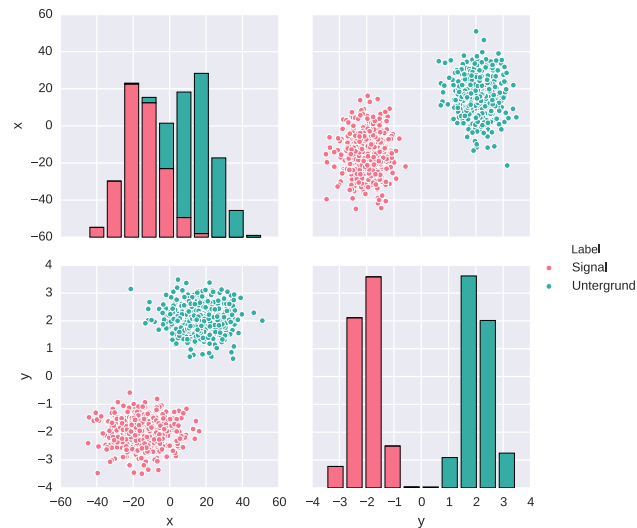


## Lineare Fisher Diskriminanzanalyse

- Um eine gute Projektion  $\vec{\lambda}$  ( $\vec{x}' = \vec{\lambda}^T \vec{x}$ ) zu finden, muss ein Maß für die Trennbarkeit der Klassen definiert werden
  - Erste (naive) Idee:  
Abstand der Mittelwerte der Klassen auf der Projektionsachse

$$D_{\text{naiv}}(\vec{\lambda}) = |\vec{\mu}'_1 - \vec{\mu}'_2| = \left| \vec{\lambda}^T (\vec{\mu}_1 - \vec{\mu}_2) \right|$$

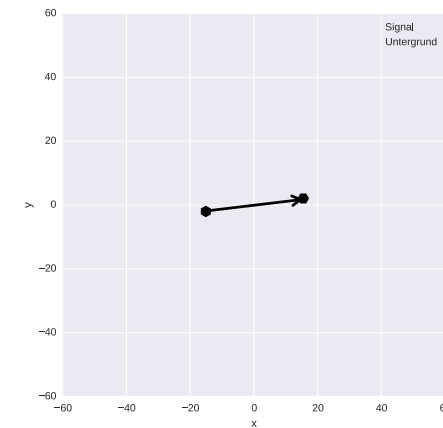
Problem: **Varianz innerhalb der Klassen wird nicht berücksichtigt!**



Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

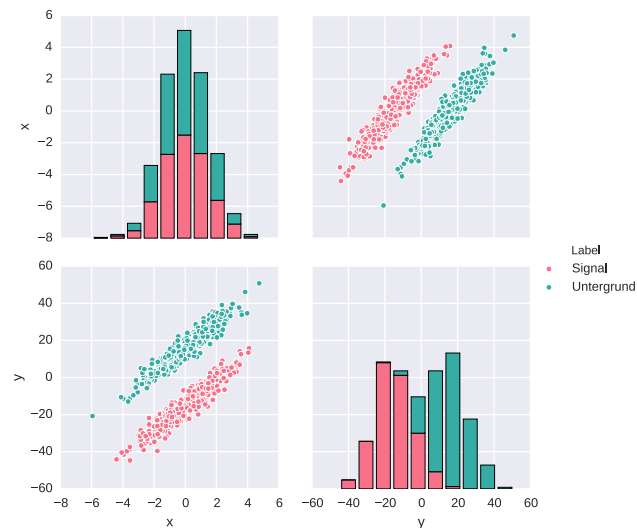
Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Lineare Fisher Diskriminanzanalyse

- Um eine gute Projektion  $\vec{\lambda}$  ( $\vec{x}' = \vec{\lambda}^T \vec{x}$ ) zu finden, muss ein Maß für die Trennbarkeit der Klassen definiert werden

- Erste (naive) Idee:  
Abstand der Mittelwerte der Klassen auf der Projektionsachse

$$D_{\text{naiv}}(\vec{\lambda}) = |\vec{\mu}'_1 - \vec{\mu}'_2| = \left| \vec{\lambda}^T (\vec{\mu}_1 - \vec{\mu}_2) \right|$$

Problem: **Varianz innerhalb der Klassen wird nicht berücksichtigt!**

- Idee nach Fisher:  
Quadrat des Abstandes der Mittelwerte der Klassen auf der Projektionsachse, normalisiert mit der Streuung der Klassen

$$D(\vec{\lambda}) = \frac{|\vec{\mu}'_1 - \vec{\mu}'_2|^2}{s_1'^2 + s_2'^2}$$

Prof. Dr. Dr. W. Rhode

## Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse



## Lineare Fisher Diskriminanzanalyse

- Optimale Trennung zweier Klassen mit je n Observablen durch eine (n-1)-dimensionale Hyperebene
- Gesucht wir die Projektion  $\vec{\lambda}$  die  $D(\vec{\lambda})$  maximiert
  - Berechnung der n-dimensionalen Mittelwertvektoren

## Lineare Fisher Diskriminanzanalyse

- Berechnung der n-dimensionalen Mittelwertvektoren
  - Allgemein

$$\vec{\mu}_j = \begin{pmatrix} \bar{x}_{j,1} \\ \vdots \\ \bar{x}_{j,n} \end{pmatrix} = \frac{1}{N_j} \begin{pmatrix} \sum x_{j,1,i} \\ \vdots \\ \sum x_{j,n,i} \end{pmatrix}$$

- Beispiel

$$\vec{\mu}_1 = \begin{pmatrix} \bar{x}_1 \\ \bar{y}_1 \end{pmatrix} = \frac{1}{N_1} \begin{pmatrix} \sum x_{1,i} \\ \sum y_{1,i} \end{pmatrix}$$

$$\vec{\mu}_2 = \begin{pmatrix} \bar{x}_2 \\ \bar{y}_2 \end{pmatrix} = \frac{1}{N_2} \begin{pmatrix} \sum x_{2,i} \\ \sum y_{2,i} \end{pmatrix}$$

## Lineare Fisher Diskriminanzanalyse

- Optimale Trennung zweier Klassen mit n Observablen durch eine (n-1)-dimensionale Hyperebene
- Gesucht wir die Projektion  $\vec{\lambda}$  die  $D(\vec{\lambda})$  maximiert
  - Berechnung der n-dimensionalen Mittelwertvektoren
  - Berechnung der Streumatrizen

## Lineare Fisher Diskriminanzanalyse

- Berechnung der Streumatrizen  $S_W$  und  $S_B$

- Streuung innerhalb der Klassen ("Within-class scatter matrix")

$$\text{Gesamtstreuung: } S_W = \sum_{j=1}^{N_{\text{Klassen}}} S_j$$

$$\text{Streuung der Klasse j: } S_j = \sum_i^{n_j} (\vec{x}_i - \vec{\mu}_j)(\vec{x}_i - \vec{\mu}_j)^T$$

- Mit dieser Matrix wird  $s_1'^2 + s_2'^2 = \vec{\lambda}^T S_W \vec{\lambda}$ , da:

$$\begin{aligned} s_j'^2 &= \sum (\vec{x}' - \vec{\mu}')^2 = \sum (\vec{\lambda}^T \vec{x} - \vec{\lambda}^T \vec{\mu})^2 = \sum (\vec{\lambda}^T (\vec{x} - \vec{\mu}))^2 \\ &= \sum (\vec{\lambda}^T (\vec{x} - \vec{\mu})) (\vec{\lambda}^T (\vec{x} - \vec{\mu}))^T = \sum \vec{\lambda}^T (\vec{x} - \vec{\mu}) (\vec{x} - \vec{\mu})^T \vec{\lambda} = \vec{\lambda}^T S_j \vec{\lambda} \end{aligned}$$

## Lineare Fisher Diskriminanzanalyse

### 2. Berechnung der Streumatrizen $S_W$ und $S_B$

- Streuung zwischen den Klassen ("Between-class scatter matrix")

$$S_B = (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T$$

- Mit dieser Matrix wird  $|\vec{\mu}'_1 - \vec{\mu}'_2|^2 = \vec{\lambda}^T S_B \vec{\lambda}$ , da:

$$\begin{aligned} |\vec{\mu}'_1 - \vec{\mu}'_2|^2 &= (\vec{\lambda}^T \vec{\mu}_1 - \vec{\lambda}^T \vec{\mu}_2)^2 \\ &= \vec{\lambda}^T (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T \vec{\lambda} \\ &= \vec{\lambda}^T S_B \vec{\lambda} \end{aligned}$$

## Lineare Fisher Diskriminanzanalyse

### 2. Berechnung der Streumatrizen $S_W$ und $S_B$

- Mit den Matrizen  $S_W$  und  $S_B$  gilt:

$$D(\vec{\lambda}) = \frac{|\vec{\mu}'_1 - \vec{\mu}'_2|^2}{s_1'^2 + s_2'^2} = \frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}$$

- Dieser Ausdruck soll nun maximiert werden

$$\vec{\lambda}^* = \arg \max \left[ \frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}} \right]$$

## Lineare Fisher Diskriminanzanalyse

- Optimale Trennung zweier Klassen mit n Observablen durch eine (n-1)-dimensionale Hyperebene
- Gesucht wir die Projektion  $\vec{\lambda}$  die  $D(\vec{\lambda})$  maximiert
  - Berechnung der n-dimensionalen Mittelwertvektoren
  - Berechnung der Streumatrizen
  - Projektion  $\vec{\lambda}^*$  berechnen

## Lineare Fisher Diskriminanzanalyse

### 3. Projektion $\vec{\lambda}^*$ berechnen

- Zu zeigen:  $\vec{\lambda}^* = \arg \max \left[ \frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}} \right] = S_W^{-1} (\mu_1 - \mu_2)$

Ableitung von  $D(\vec{\lambda})$  und mit 0 gleichsetzen:

$$\begin{aligned} \frac{d}{d\vec{\lambda}} [D(\vec{\lambda})] &= \frac{d}{d\vec{\lambda}} \left[ \frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}} \right] = 0 \\ \Leftrightarrow [\vec{\lambda}^T S_W \vec{\lambda}] \frac{d [\vec{\lambda}^T S_B \vec{\lambda}]}{d\vec{\lambda}} - [\vec{\lambda}^T S_B \vec{\lambda}] \frac{d [\vec{\lambda}^T S_W \vec{\lambda}]}{d\vec{\lambda}} &= 0 \\ \Leftrightarrow [\vec{\lambda}^T S_W \vec{\lambda}] 2S_B \vec{\lambda} - [\vec{\lambda}^T S_B \vec{\lambda}] 2S_W \vec{\lambda} &= 0 \end{aligned}$$

## Lineare Fisher Diskriminanzanalyse

### 3. Projektion $\vec{\lambda}^*$ berechnen (Fortsetzung)

$$\Leftrightarrow \left[ \vec{\lambda}^T S_W \vec{\lambda} \right] 2 S_B \vec{\lambda} - \left[ \vec{\lambda}^T S_B \vec{\lambda} \right] 2 S_W \vec{\lambda} = 0$$

- Durch  $\vec{\lambda}^T S_W \vec{\lambda}$  teilen:

$$\Leftrightarrow \frac{\left[ \vec{\lambda}^T S_W \vec{\lambda} \right]}{\left[ \vec{\lambda}^T S_W \vec{\lambda} \right]} S_B \vec{\lambda} - \frac{\left[ \vec{\lambda}^T S_B \vec{\lambda} \right]}{\left[ \vec{\lambda}^T S_W \vec{\lambda} \right]} S_W \vec{\lambda} = 0$$

(Skalar)

$$\Leftrightarrow S_B \vec{\lambda} - D S_W \vec{\lambda} = 0$$

$$\Leftrightarrow S_W^{-1} S_B \vec{\lambda} = D \vec{\lambda}$$

- Lösung des Eigenwert-Problems  $S_W^{-1} S_B \vec{\lambda} = D \vec{\lambda}$

## Lineare Fisher Diskriminanzanalyse

### 3. Projektion $\vec{\lambda}^*$ berechnen (Fortsetzung)

- Lösung des Eigenwert-Problems  $S_W^{-1} S_B \vec{\lambda} = D \vec{\lambda}$ :

Erinnerung:  $S_B = (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T$  d.h.  $S_B$  auf einen beliebigen Vektor  $\vec{v}$  angewendet liefert immer einen Vektor in Richtung  $(\vec{\mu}_1 - \vec{\mu}_2)$

$$S_B \vec{v} = (\vec{\mu}_1 - \vec{\mu}_2) \underbrace{(\vec{\mu}_1 - \vec{\mu}_2)^T \vec{v}}_k = k(\vec{\mu}_1 - \vec{\mu}_2)$$

$$\Rightarrow S_W^{-1} S_B \vec{\lambda} = k S_W^{-1} (\vec{\mu}_1 - \vec{\mu}_2) = D \vec{\lambda}$$

Offensichtlich ist eine mögliche Lösung:

$$\vec{\lambda}^* = \arg \max \frac{\left[ \vec{\lambda}^T S_B \vec{\lambda} \right]}{\left[ \vec{\lambda}^T S_W \vec{\lambda} \right]} = S_W^{-1} (\vec{\mu}_1 - \vec{\mu}_2)$$

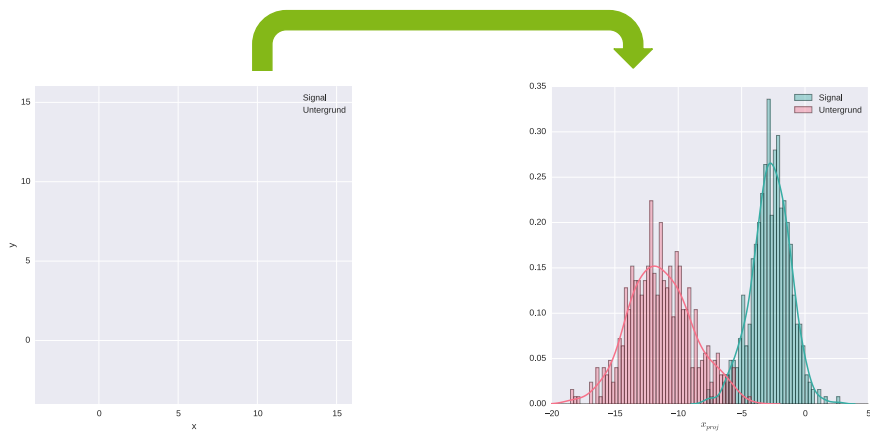
## Lineare Fisher Diskriminanzanalyse

- Optimale Trennung zweier Klassen mit n Observablen durch eine (n-1)-dimensionale Hyperebene
- Gesucht wir die Projektionsachse  $\vec{\lambda}$  die  $D(\vec{\lambda})$  maximiert
  - Berechnung der n-dimensionalen Mittelwertvektoren
  - Berechnung der Streumatrizen
  - Projektion  $\vec{\lambda}^*$  berechnen
  - Schnitt auf der Projektionsachse festlegen

## Lineare Fisher Diskriminanzanalyse

- Schnitt auf der Projektionsachse festlegen
  - Jeder n-dimensionale Punkt wird in eine Dimension projiziert
  - Gesucht ist ein Schnitt in auf der Projektionsachse, anhand dessen zwischen beiden Populationen entschieden wird

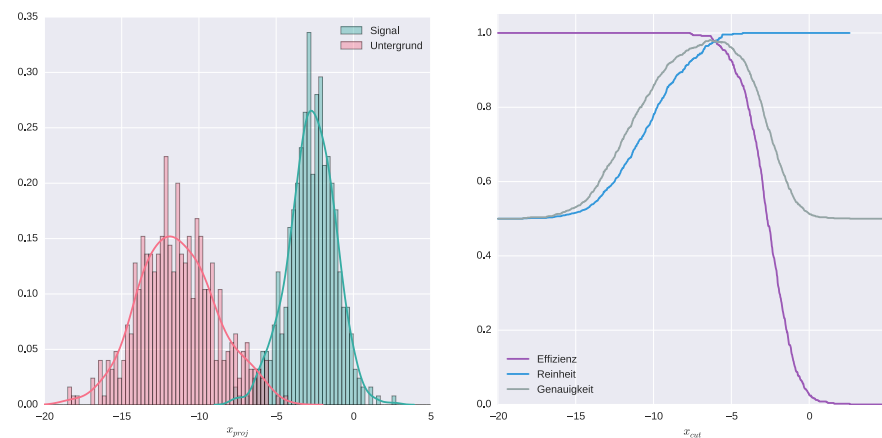
### Projektion der Fisher Diskriminanzanalyse



Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

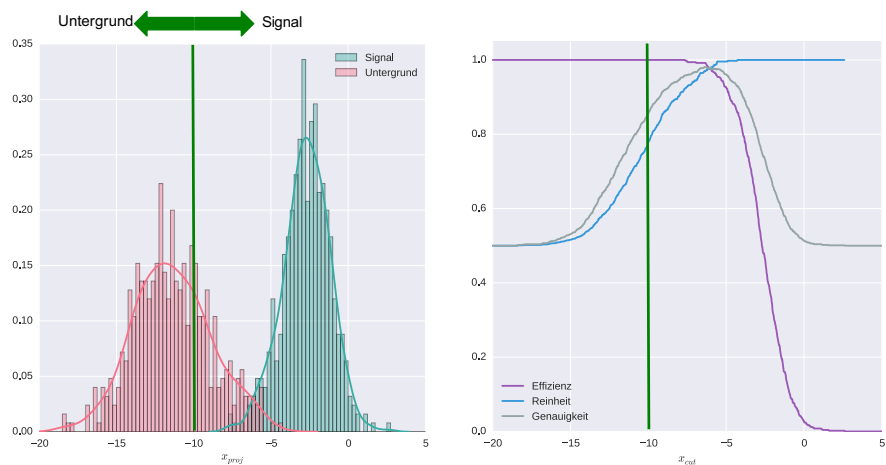
Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

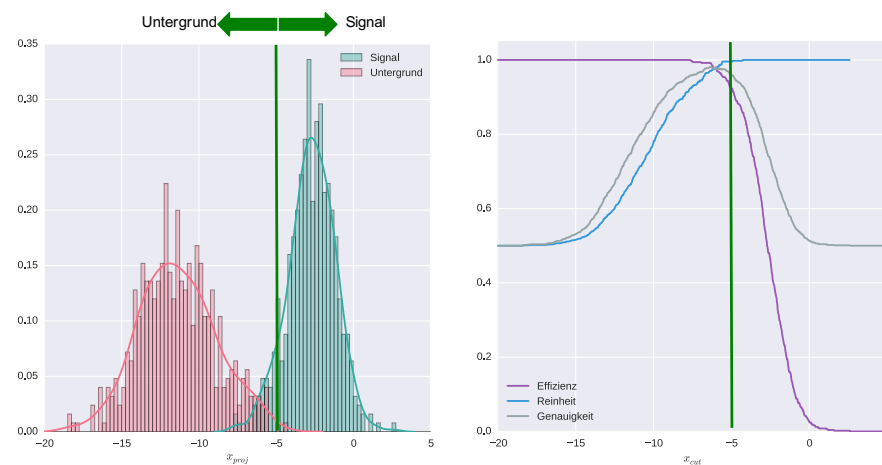
Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse



Prof. Dr. Dr. W. Rhode

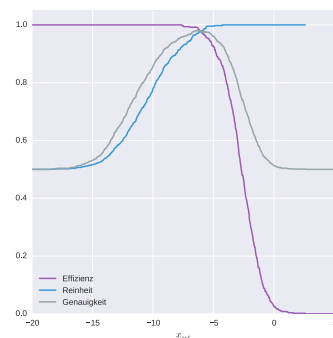
Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Lineare Fisher Diskriminanzanalyse

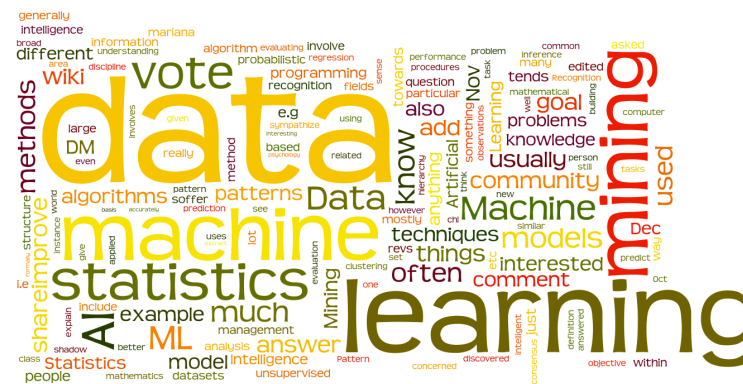
#### 4. Schnitt auf der Projektionsachse festlegen

- Jeder n-dimensionale Punkt wird auf eine Dimension projiziert
  - Gesucht ist ein Schnitt in auf der Projektionsachse, anhand dessen zwischen beiden Populationen entschieden wird
- 
- Allgemein kann kein bester Schnitt angegeben werden
  - Muss für jedes konkrete Problem motiviert werden
    - Trade-Off zwischen Effizienz und Reinheit



Prof. Dr. Dr. W. Rhode

# Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

„The capacity of digital data storage worldwide has doubled every nine months for at least a decade, at twice the rate predicted by Moore’s Law for the growth of computing power during the same period.“  
(Fayyad et al., 2002)

Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Data-Mining

- Ursprünglich ein Schritt von sogenannten „Knowledge Discovery in Databases“-Prozessen; mittlerweile meist gleichbedeutend zu KDD
- Data-Mining meint häufig die Anwendung von Algorithmen des maschinellen Lernens
  - „[Machine learning is a] field of study that gives computers the ability to learn without being explicitly programmed.“ (Arthur Smith, 1959)
  - „A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.“ (Tom M. Mitchell, 1997)
- Dieser Teil der Vorlesung soll einen **Einblick** in das große Feld des Data Minings geben


Prof. Dr. Dr. W. Rhode

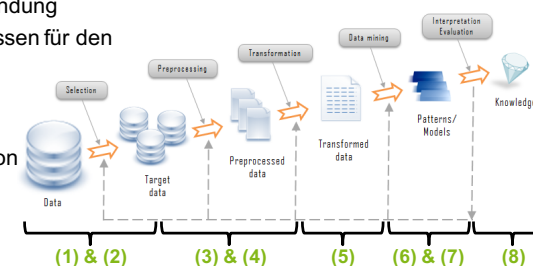
# Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Data Mining

- Die Teilschritte eines Data Mining-Prozesses sind (Fayyad et al.):
  - 1. Problemformulierung
  - 2. Datenaufbereitung
  - 3. Modellierung
  - 4. Modellbewertung
  - 5. Modellimplementierung

- (1) Definition der Ziele der Wissensfindung
  - (2) Bereitstellung von Hintergrundwissen für den jeweiligen Fachbereich
  - (3) Datenauswahl
  - (4) Datenbereinigung
  - (5) Datenreduktion und -transformation
  - (6) Auswahl eines Modells
  - (7) Data-Mining
  - (8) Interpretation
- 
- Das Diagramm zeigt den Prozess des Data Mining. Ein Stapel von blauen Kreisen ist mit 'Data' beschriftet. Ein orangefarbener Pfeil zeigt von der Mitte des Stapels nach oben zu einem grauen Kasten mit der Aufschrift 'Selection'. Ein weiterer orangefarbener Pfeil zeigt von der Spitze des 'Data'-Stapels nach oben zu einem weiteren orangefarbenen Pfeil, der nach rechts zeigt. Ein schwarzer Klammerbogen unter dem 'Data'-Stapel und dem orangefarbenen Pfeil, der nach rechts zeigt, ist mit '(1) & (2)' beschriftet.



- **KDD-/Data-Mining-Prozesse sind iterativ und interaktiv**
- In der Praxis sind manche Schritte nicht voneinander zu trennen und die Reihenfolge kann leicht unterschiedlich sein

Prof. Dr. Dr. W. Rhode

Data-Mining – Teil 1

Statistische Methoden  
der Datenanalyse

## Kleines Data Mining Wörterbuch

- **Feature:** Attribut, Observable, Messgröße, Merkmal
- **Label:** Zielgröße → **labeled/unlabeled:** Wert der Zielgröße bekannt/unbekannt
- **Klassen:** Werte einer diskreten Zielgröße (häufig wird Label und Klasse äquivalent genutzt)
- **Überwachtes Lernen:** „Supervised learning is the machine learning task of inferring a function from labeled training data.“ (Foundations of Machine Learning, 2012)
- **Unüberwachtes Lernen:** Erkennen von Strukturen unabhängig von Zielgrößen, Optimierungskriterien, Feedback Signalen oder sonstiger Informationen, die über die tatsächlichen Daten hinausgehen
- Warnung: Selten sind die Bedeutungen der Begriffe allgemeingültig definiert