

# Data Mining “light” mit JMP

Bertram Schäfer  
STATCON

- **Deutsche Sektion gegründet**
- [www.enbis.org](http://www.enbis.org)
  - Eintragen
  - Workshops geplant
  - News aus der Statistik können kostenfrei auf die website ([boris.kulig@statcon.de](mailto:boris.kulig@statcon.de))

## Data Mining

- Begriffe
- Verfahren
- Beispiele in JMP

## Was ist Data Mining?

- Wüsste ich auch gerne...
- Data Mining ist im allgemeinen recht ungenau definiert (Buzz Word)
- Beinhaltet:
  - Suche nach *vorteilhaften Mustern in Daten*
  - Nicht nur, aber insbesondere für große Datenmengen und Variablenanzahl
  - Anwendung von Data Mining Verfahren

## Data Mining - Definitionsversuche

- "the science of extracting useful information from large data sets or databases." (D. Hand, H. Mannila, P. Smyth, 2001)
- „Data-Mining ist ein integrierter Prozess, der durch Anwendung von Methoden auf einen Datenbestand Muster entdeckt“ (Bensberg, 2004)
- SEMMA  
sample, explore, modify, model, assess
- „score your production data on any machine, and deploy the scoring code in batch or real-time on the Web directly in relational databases“

## Typische Verfahren

- Clusteranalyse
- Tree-Verfahren
- Neuronale Netzwerke
- Logistische Regression
- Genetische Algorithmen
- Support Vector Machines

## Ziele:

- Mustererkennung
- Modelle für Vorhersagen

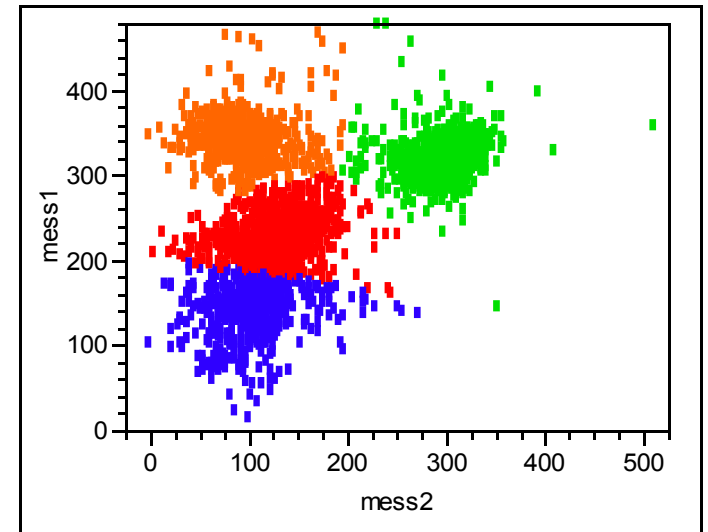
## Data Mining mit JMP

- Begriffe
- **Verfahren**
- Beispiele in JMP

## Cluster-Analyse

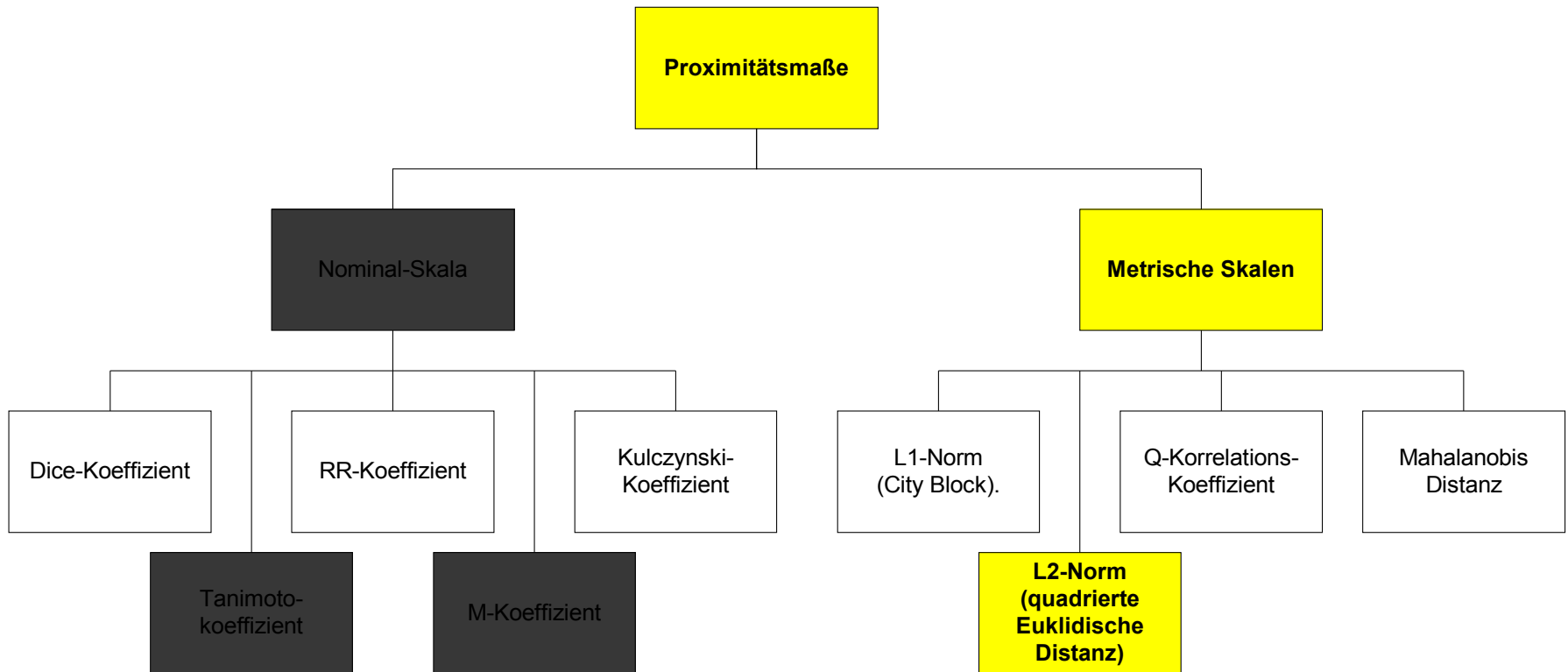
### Verfahren zur Bildung von Gruppen

- Daten nicht gleichmäßig im n-dimensionalen Raum verteilt
- Klumpungen, lokale Dichtezonen
- Gruppen untereinander möglichst unähnlich
- aller vorliegenden Merkmale werden genutzt
- Keine Zielgröße!

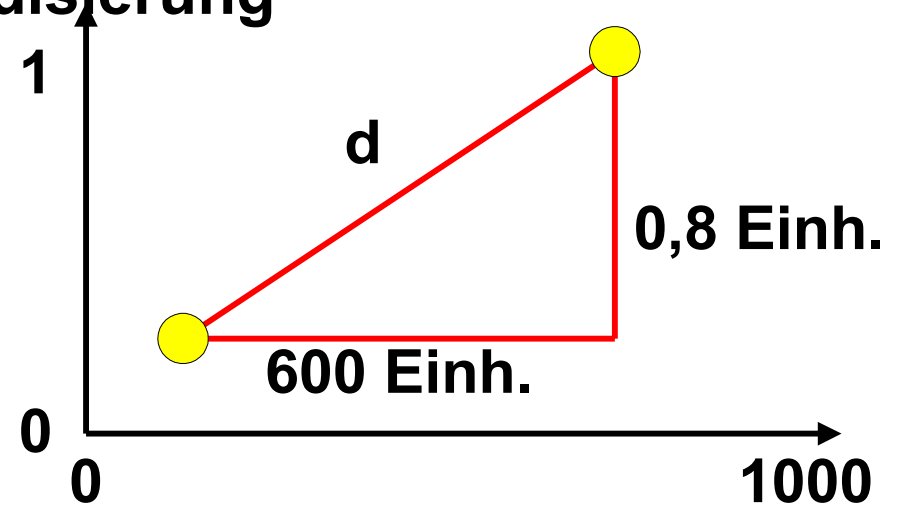


## Clusterverfahren – Proximitätsmaße

Wert, der den Unterschied bzw. die Ähnlichkeit von Fällen ausdrückt



## Clusterverfahren - Standardisierung



- **Euklidischen Distanz**

- ohne Standardisierung

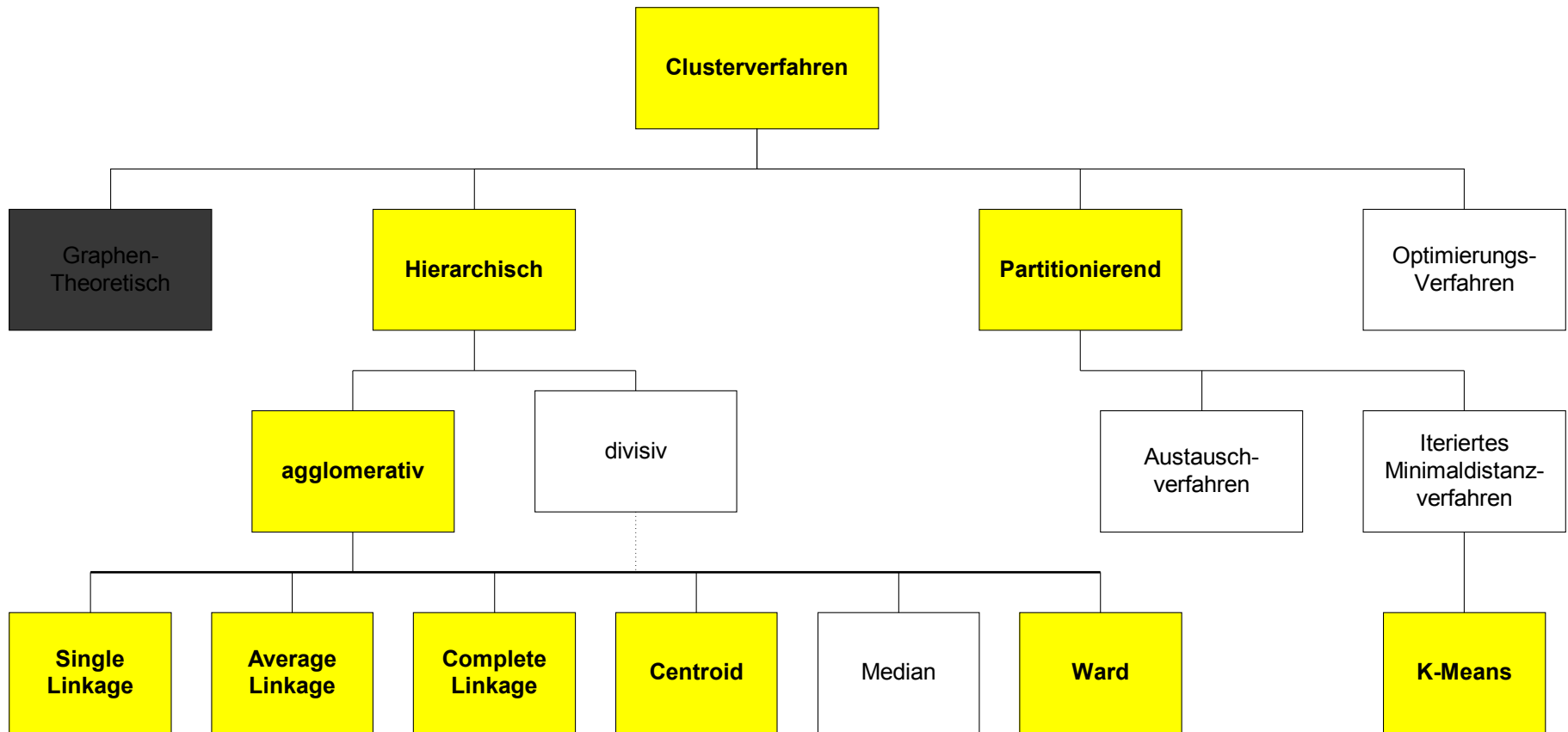
$$d_{(i,j)} = \sqrt{(600^2 + 0,8^2)} = \sqrt{(36000 + 0,64)} = \sqrt{600,0007}$$

- mit Standardisierung

$$d_{(i,j)} = \sqrt{(0,6^2 + 0,8^2)} = \sqrt{(0,36 + 0,64)} = \sqrt{1}$$

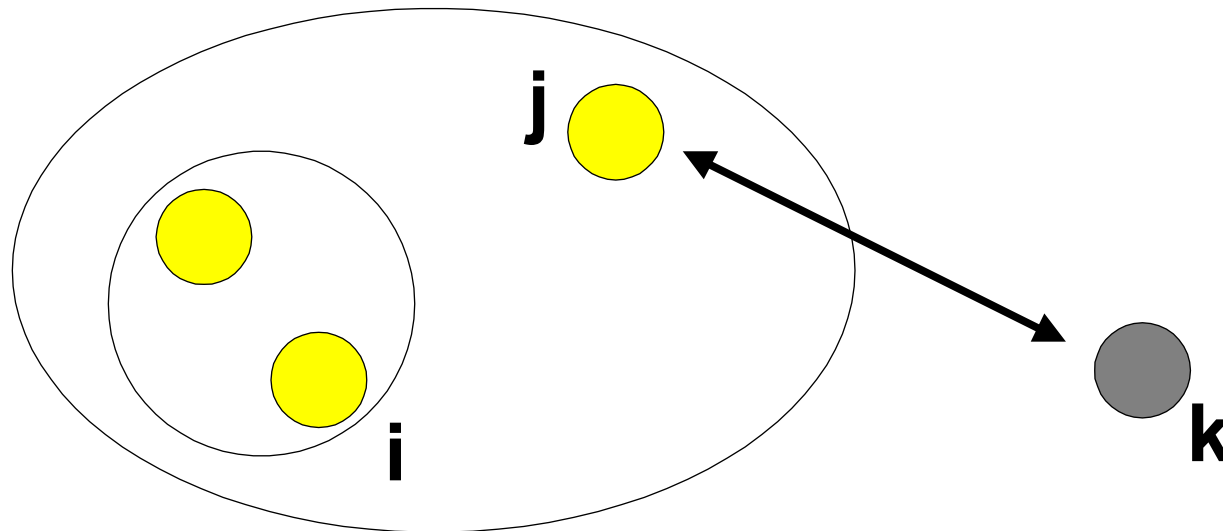
## Clusterverfahren – Fusionsalgorithmen

- Verfahren zum Zusammenfassen von Fällen



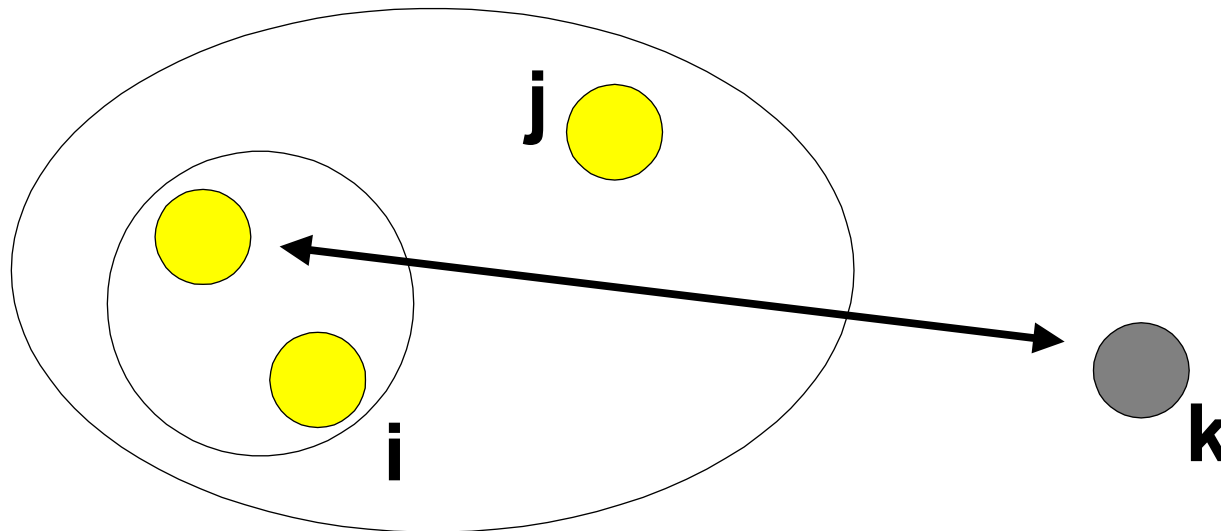
## Clusterverfahren – Fusionsalgorithmen

- **Single Linkage (Nearest-Neighbour-Verfahren / Nächstgelegener Nachbar)**



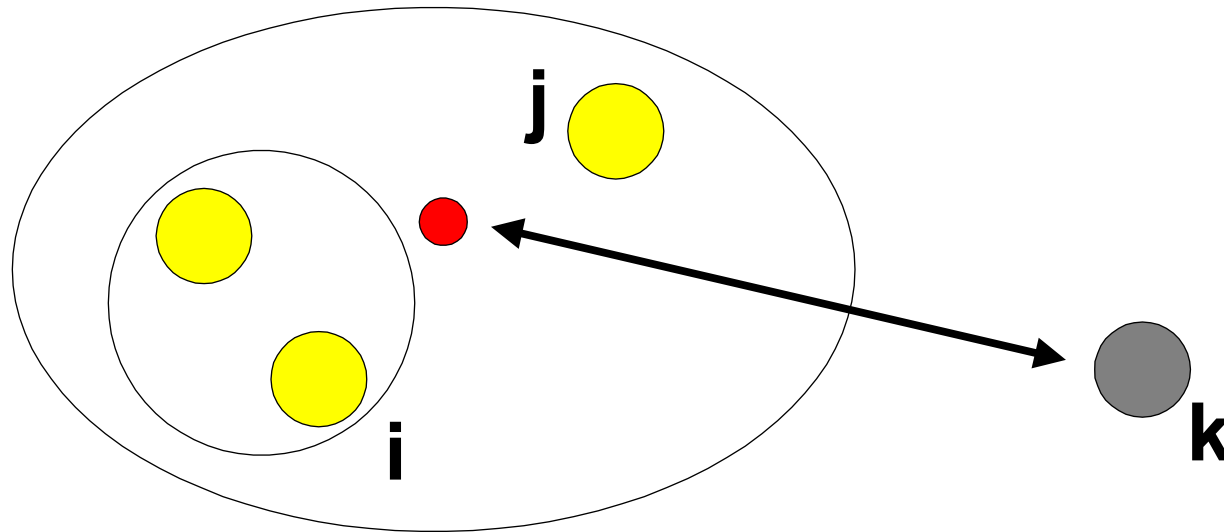
## Clusterverfahren – Fusionsalgorithmen

- **Complete Linkage (Furthest-Neighbour-Verfahren / Entferntester Nachbar)**



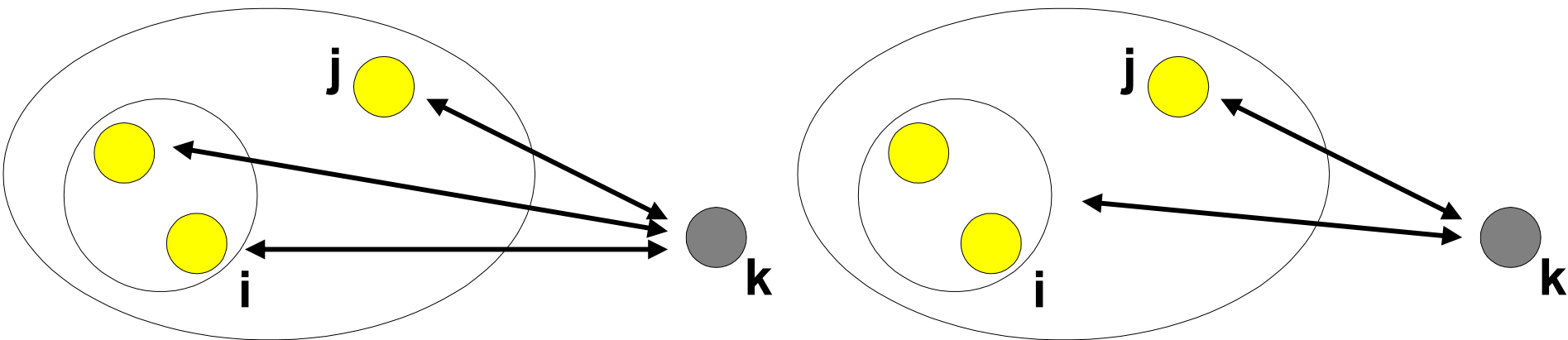
## Clusterverfahren – Fusionsalgorithmen

- **Centroid**



## Clusterverfahren – Fusionsalgorithmen

- **Average Linkage / Weighted Average Linkage**



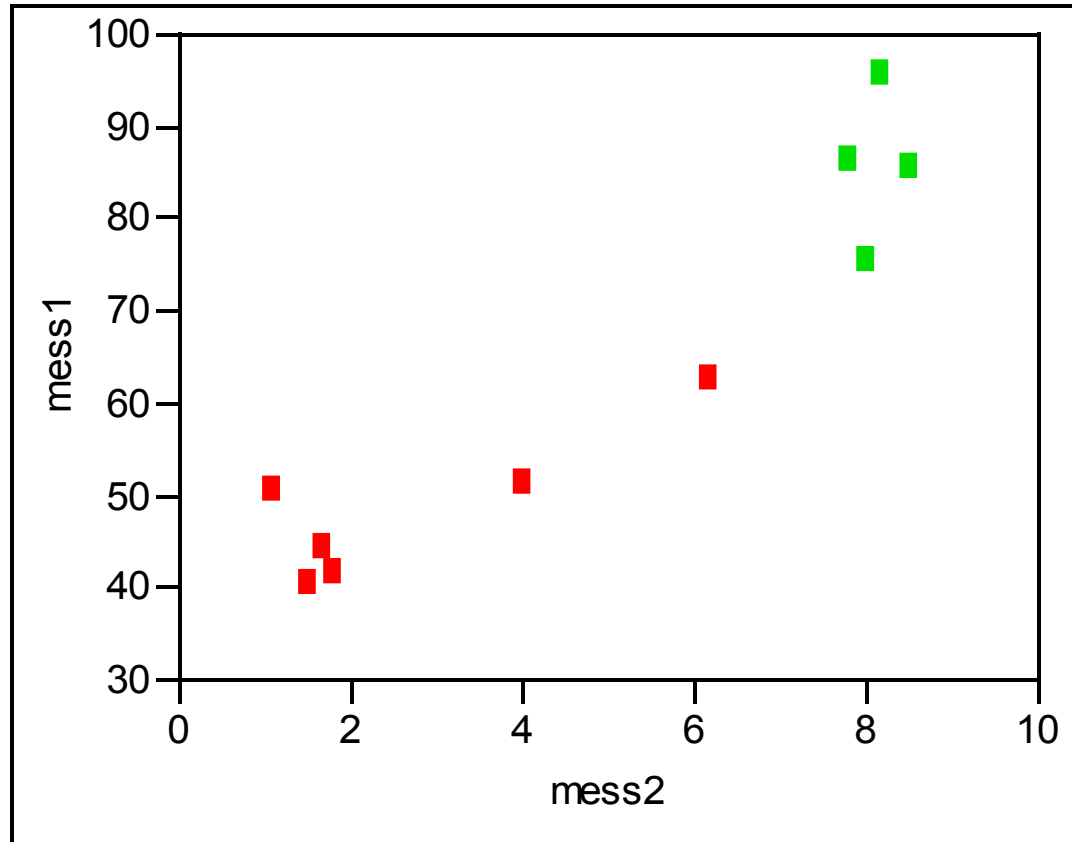
## Clusterverfahren – Fusionsalgorithmen

- **Ward**
  - Varianzkriterium
  - Geringste Erhöhung der Fehler-Quadratsumme

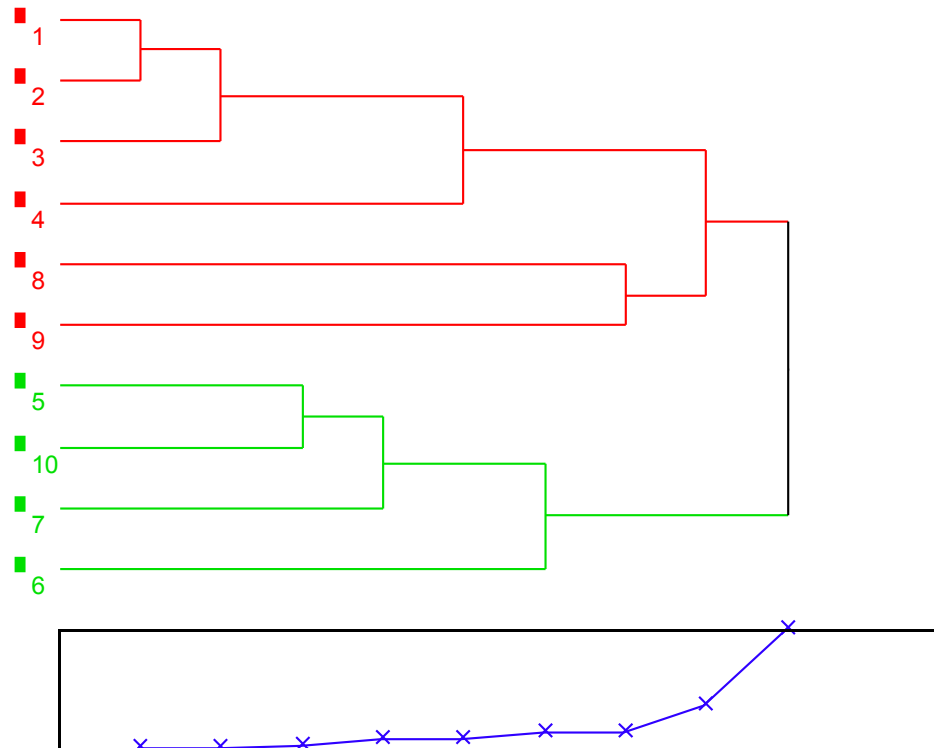
## K-Means Clustering

- Clusterbildung ist ein iterativer alternierender Prozess
- Zugehörigkeit eines Datensatzes zu einer Partition ist variabel
- Für große Datensätze
- Festlegung der Anzahl von Partitionen ist nötig
- Fusionierung ähnlich dem Centroid Verfahren

- **Streudiagramm / Scatterplot**

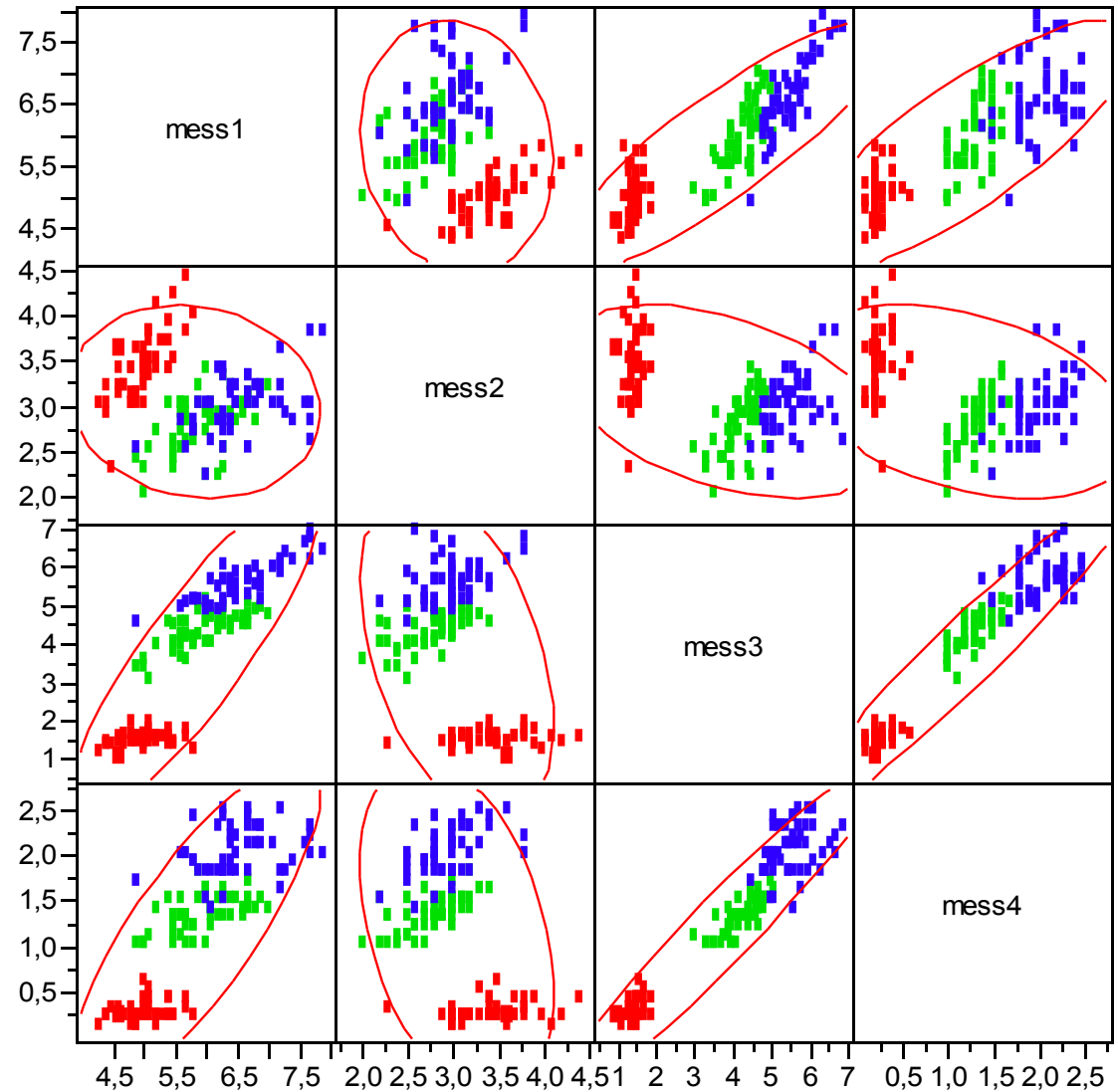


- **Dendrogramm**

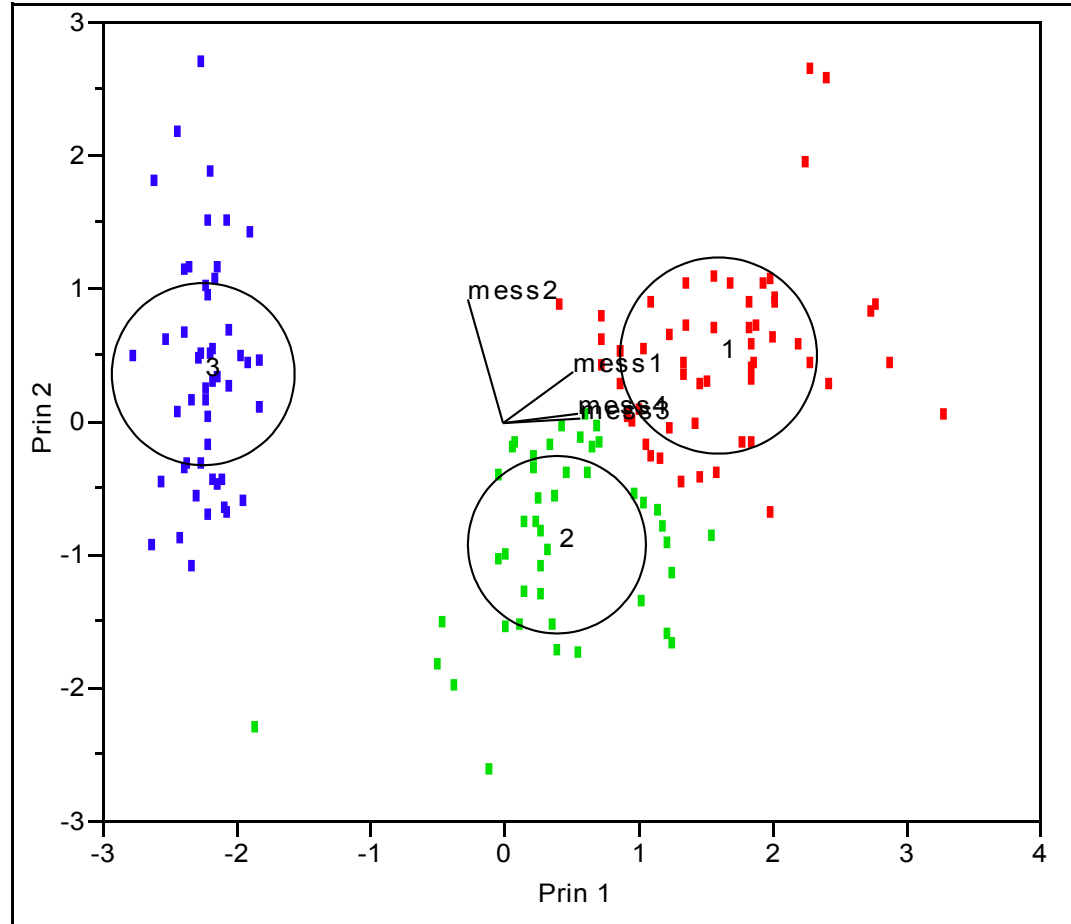


# Darstellung von Clustern

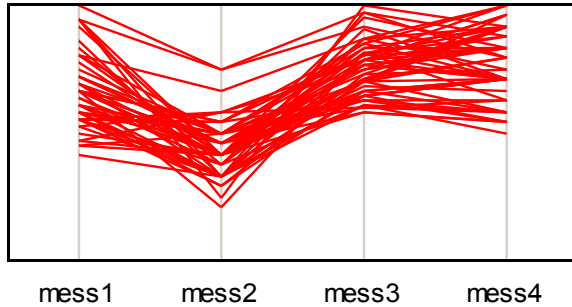
- **Scatterplot  
-Matrix**



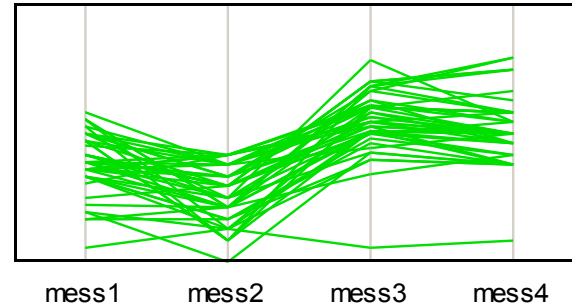
- **Biplot**



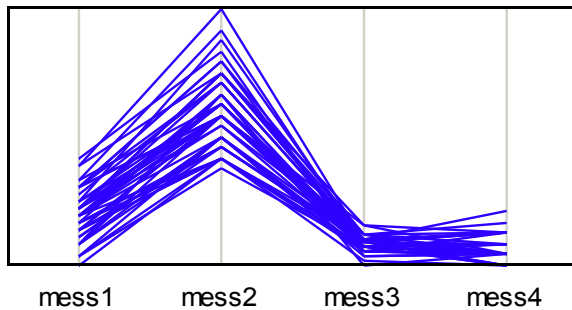
- **Parallel Coordinate Plots**



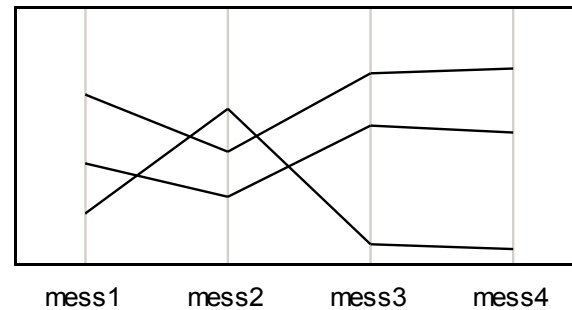
Cluster1



Cluster2



Cluster3



Cluster Means

## ***Partition - Entscheidungsbäume***

- Decision Trees / Classification Tree / Regression Trees
- *Partitioning* bezieht sich auf die Segmentierung in Untergruppen, die so homogen wie möglich

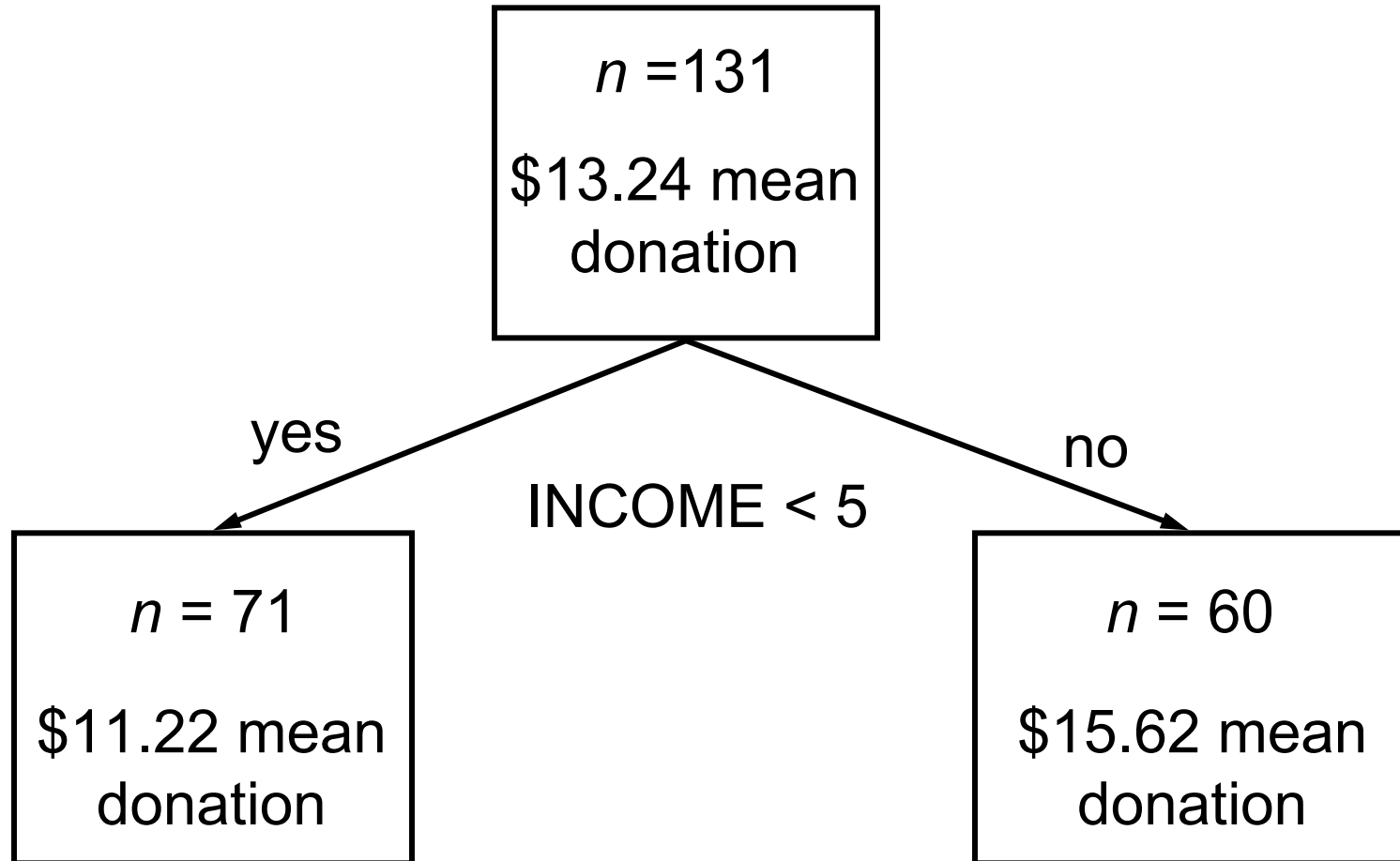
**in Bezug auf eine Zielgröße (Y)**

sind.

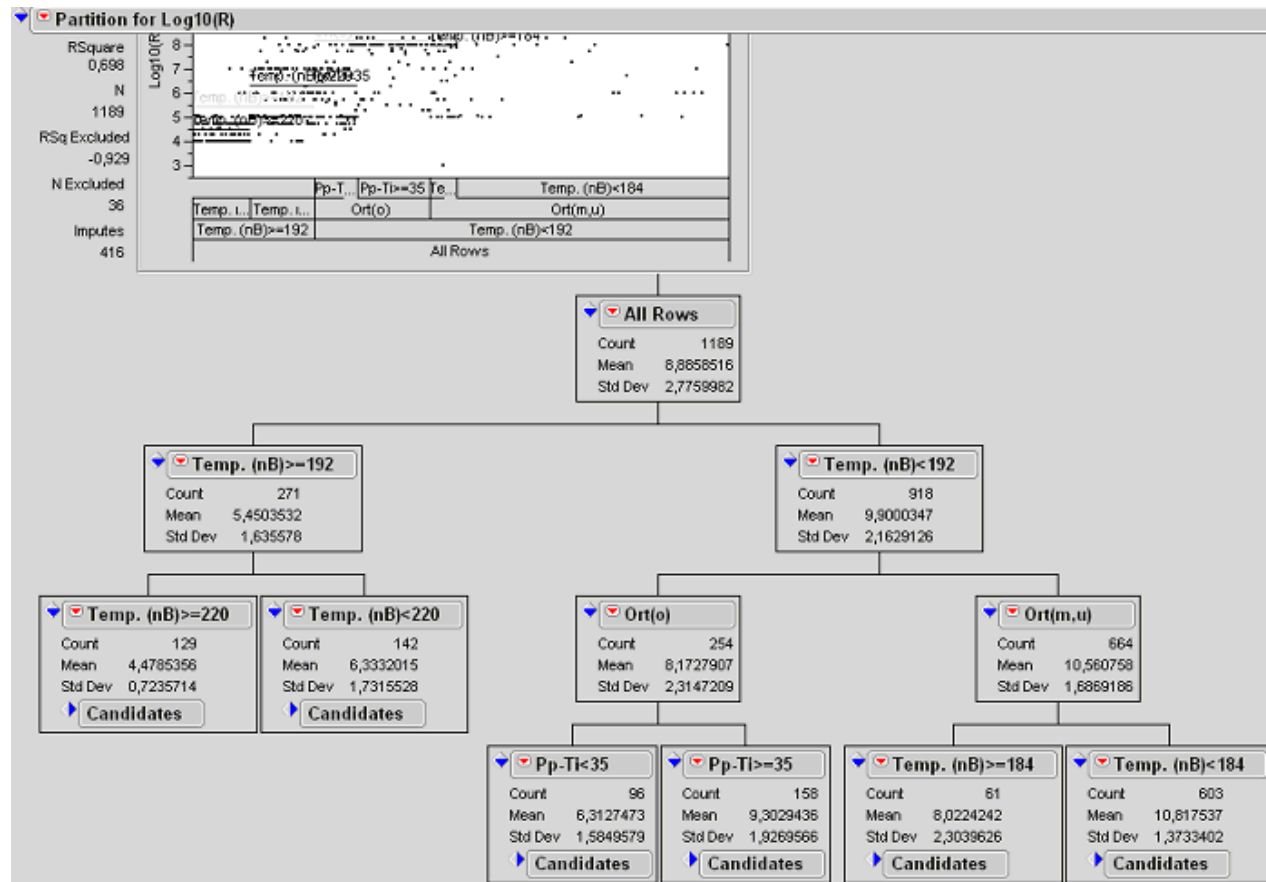
## ***Partition – Entscheidungsbäume***

- Rekursiv partitionierendes Verfahren
- Einflussgrößen und abhängige Variablen können jegliches Skalenniveau haben
- Ergebnis der Partitionierung ist ein Baum
- Partitionen werden durch Blätter symbolisiert
- Benötigt keine Modellannahme
- Läuft weitestgehend automatisch ab
- Eingriffe sind trotzdem möglich

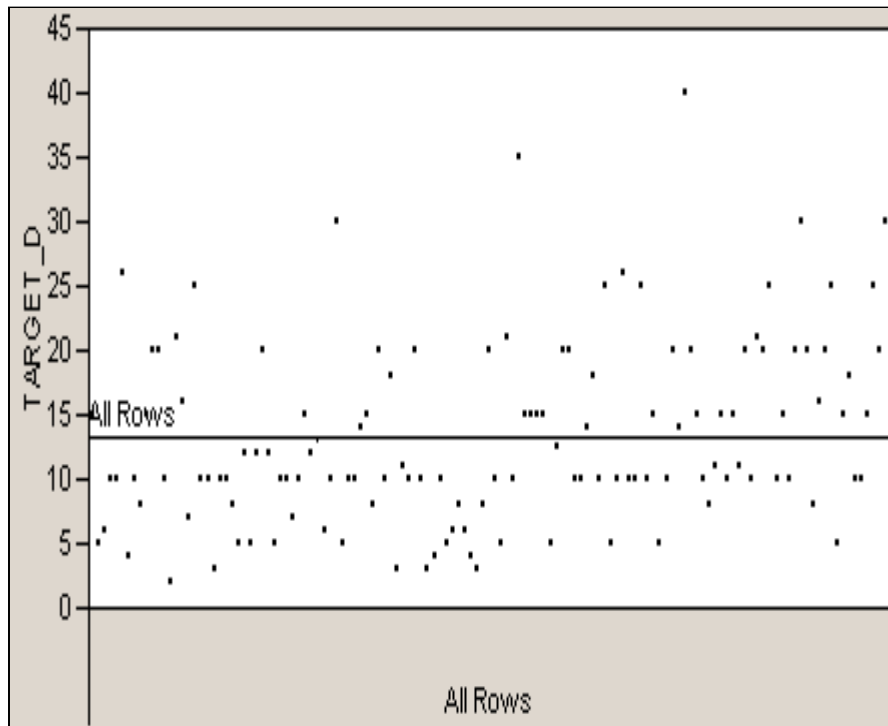
## Partition - „Teile und herrsche“



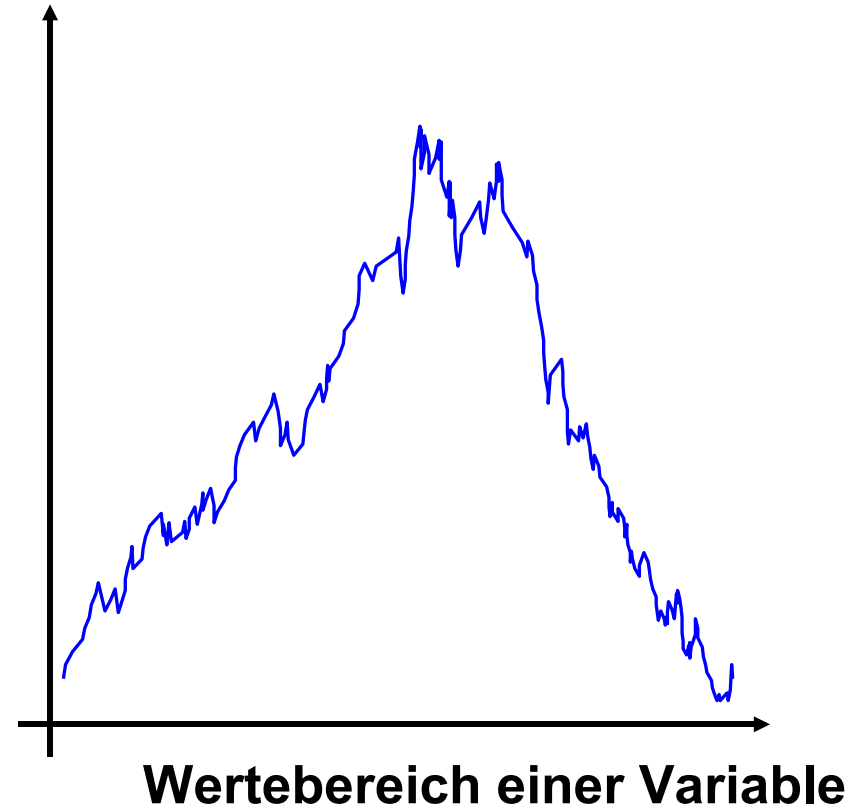
- **Beispiel eines Entscheidungsbaums in JMP**



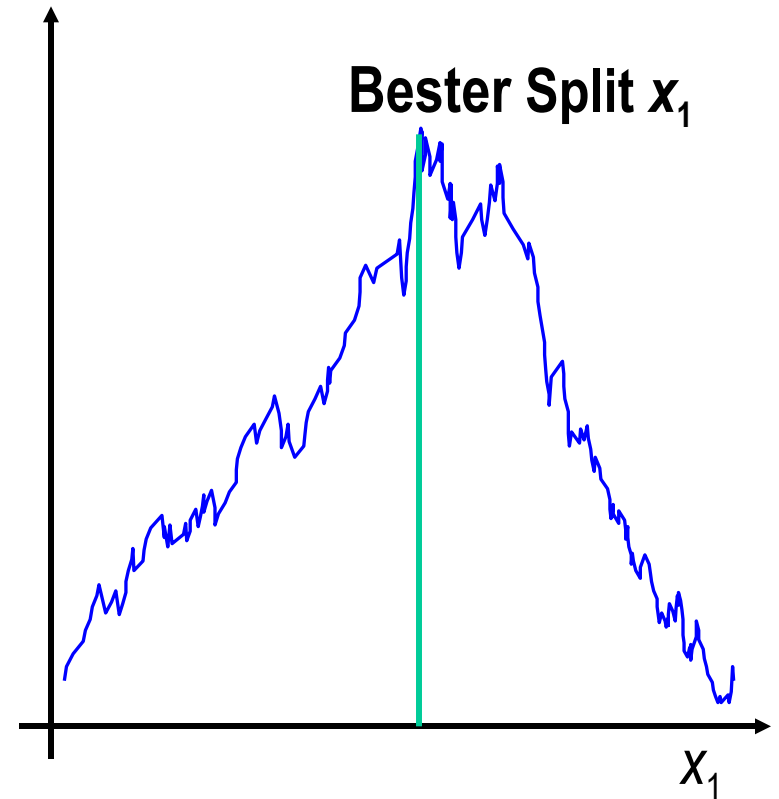
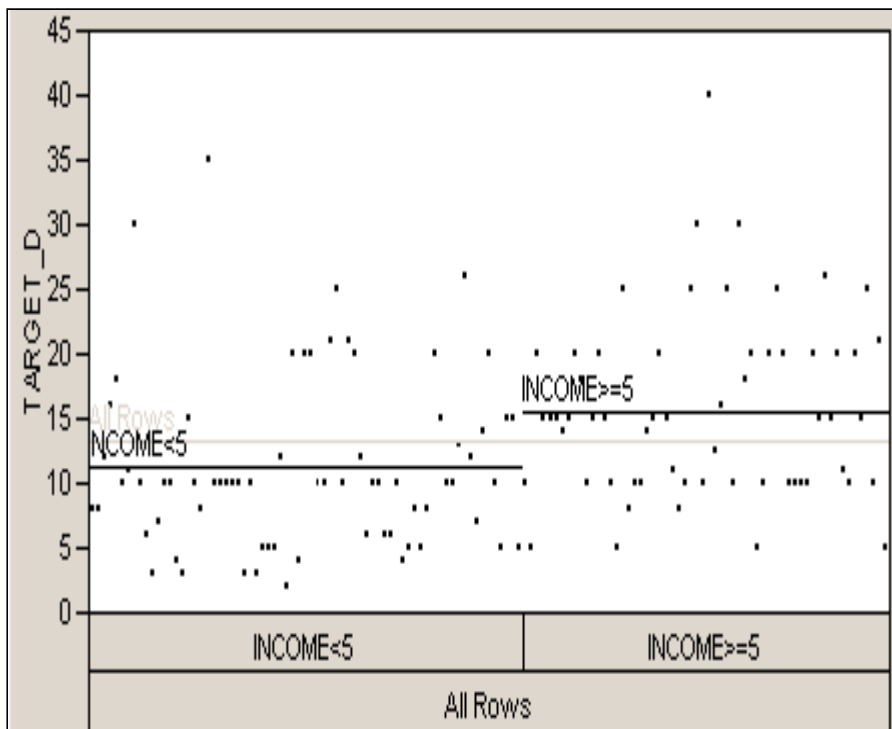
## Algorithmus 1: berechne Trennung der Mittelwerte



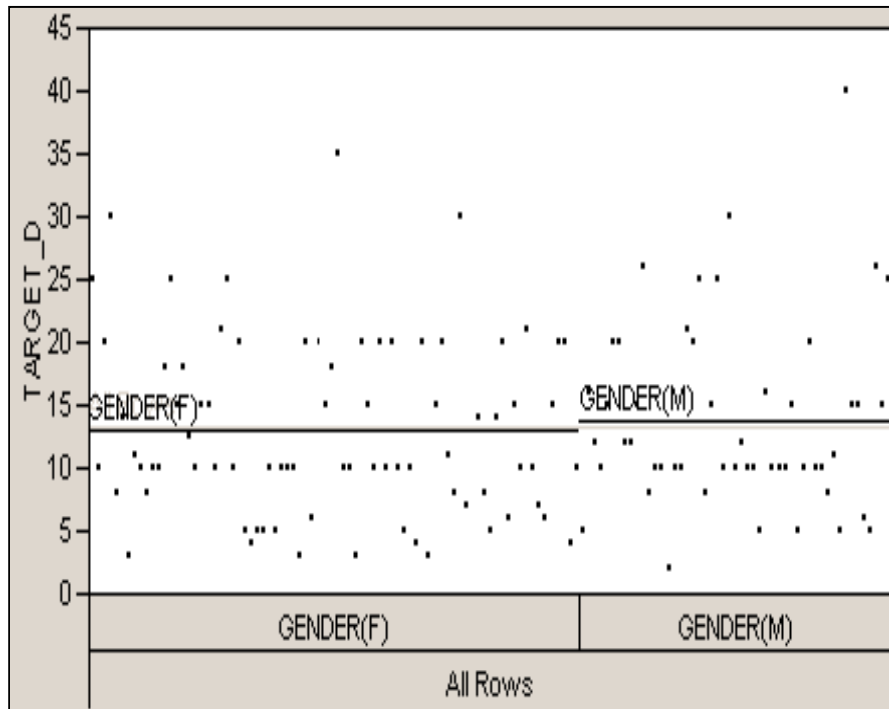
### Trennung der Mittelwerte



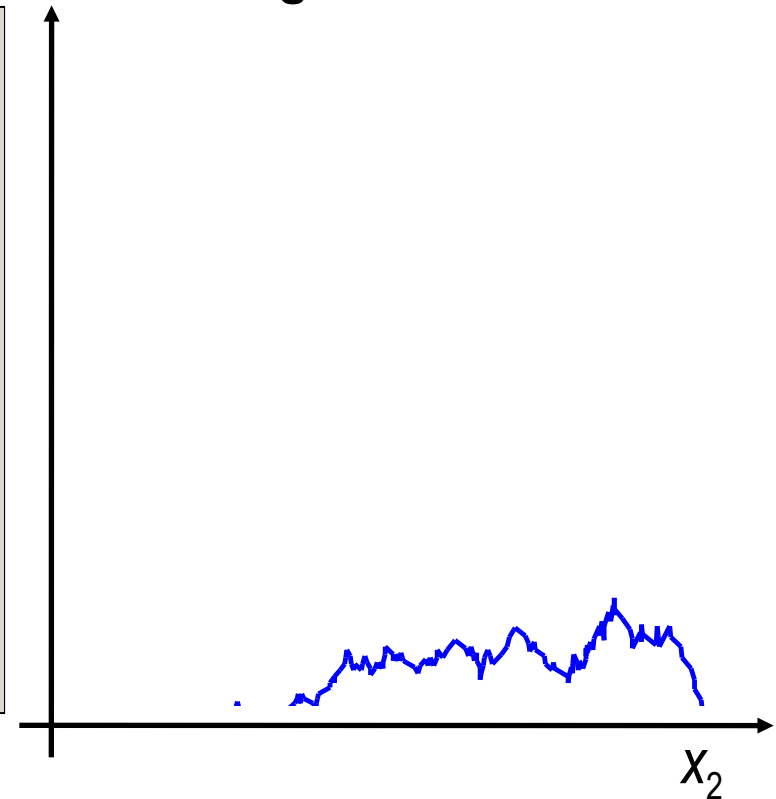
## Algorithmus 2: berechne besten Split



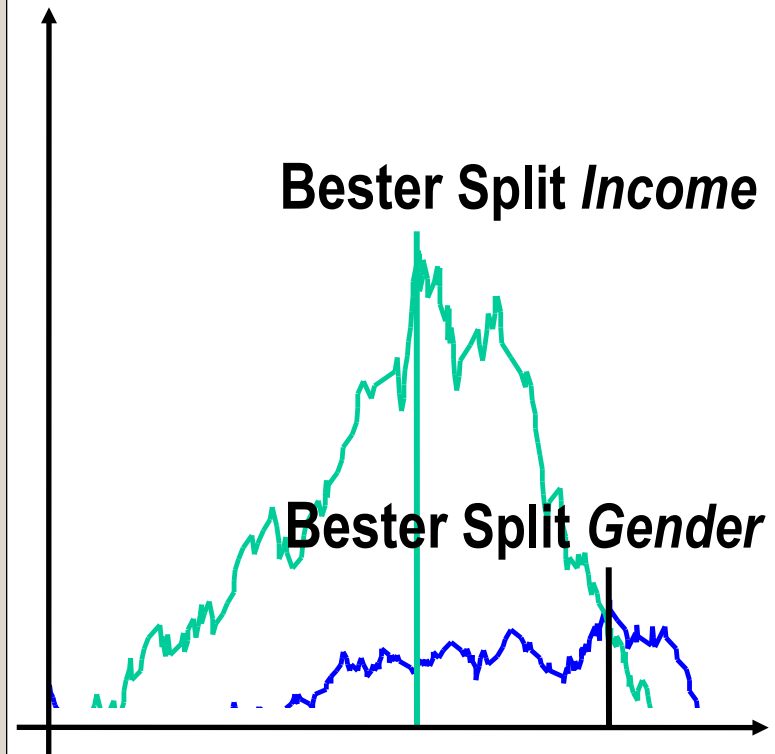
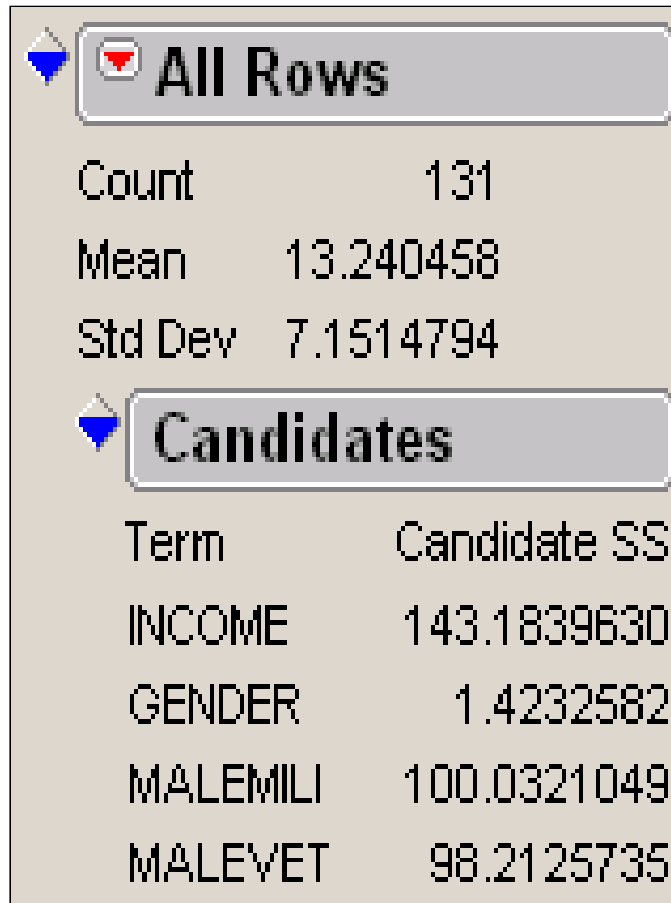
## Algorithmus 3: Wiederholung für andere Variablen



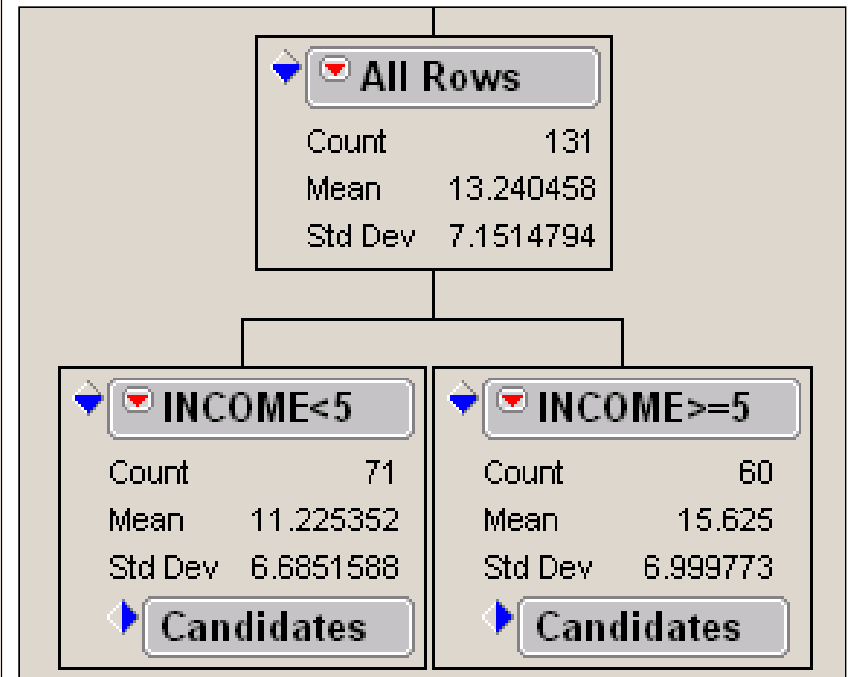
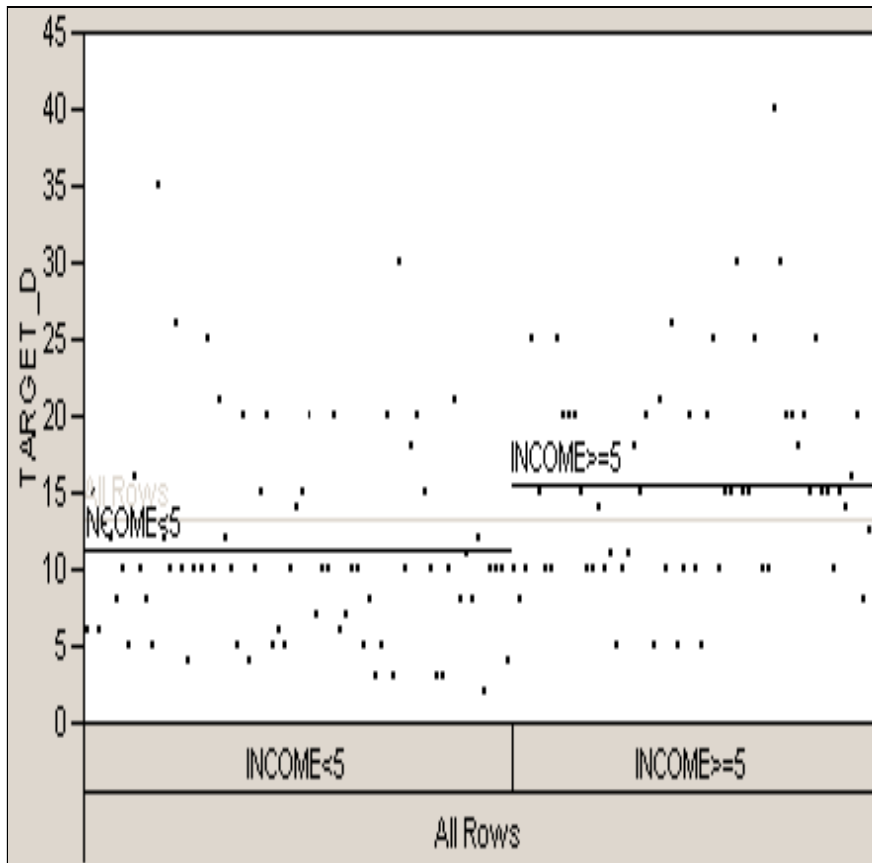
Trennung der Mittelwerte



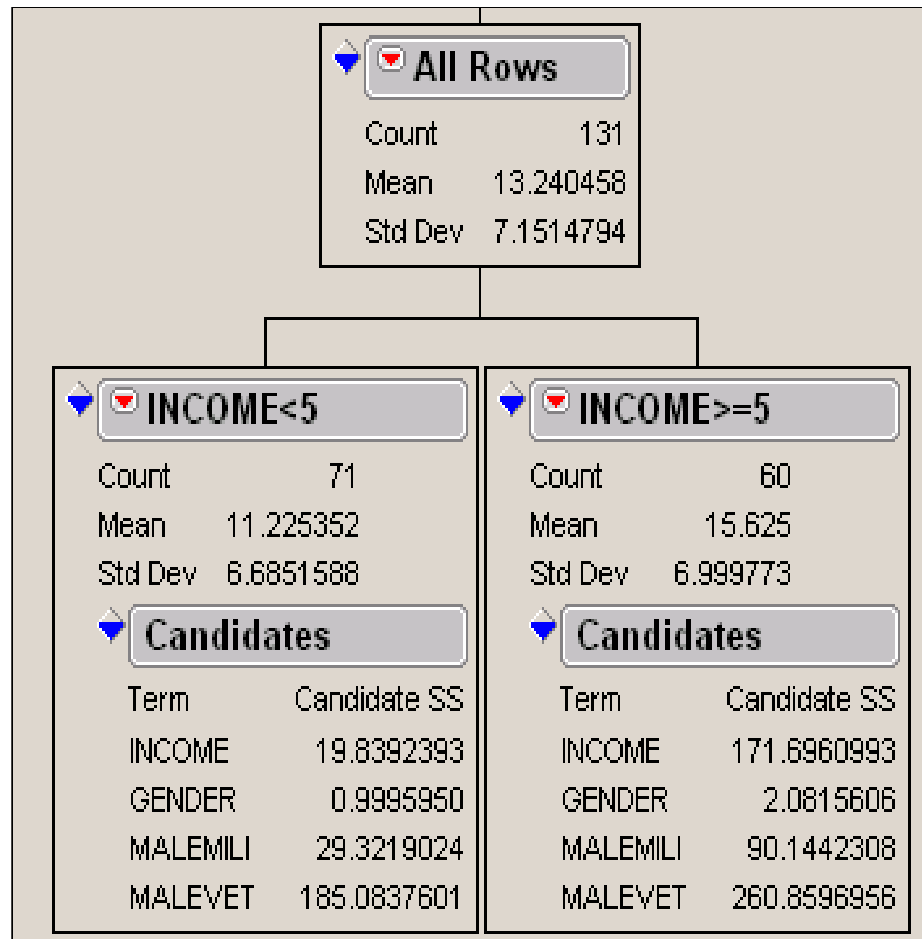
## Algorithmus 4: Vergleiche besten Split



## Algorithmus 5: Teile gemäß bestem Split



## Algorithmus 6: wiederhole innerhalb der Partitionen



## *Partition – Optionen*

- Split Here
- Split Specific
- Prune ...
- Minimum Size Split
- Split History
- Leaf Report
- Column Contribution
- ROC Curve
- Lift Curve

## Übersicht klassische Analysen

<b>Prediktor</b> (X, Regressor, Unabh. Variable) <b>Zielgröße</b> (Y, abh. Variable)	Kategorial	Stetig
Stetig	ANOVA	Regression
Kategorial	Kontingenz- tabelle	Logistische Regression

## Binäre logistische Regression

**Gegeben:** zweistufige Variable  $Y$  (Stufen 0 und 1)

**Ziel:** Modellierung der Wahrscheinlichkeit für das Eintreten von Stufe 1 (bzw. 0)

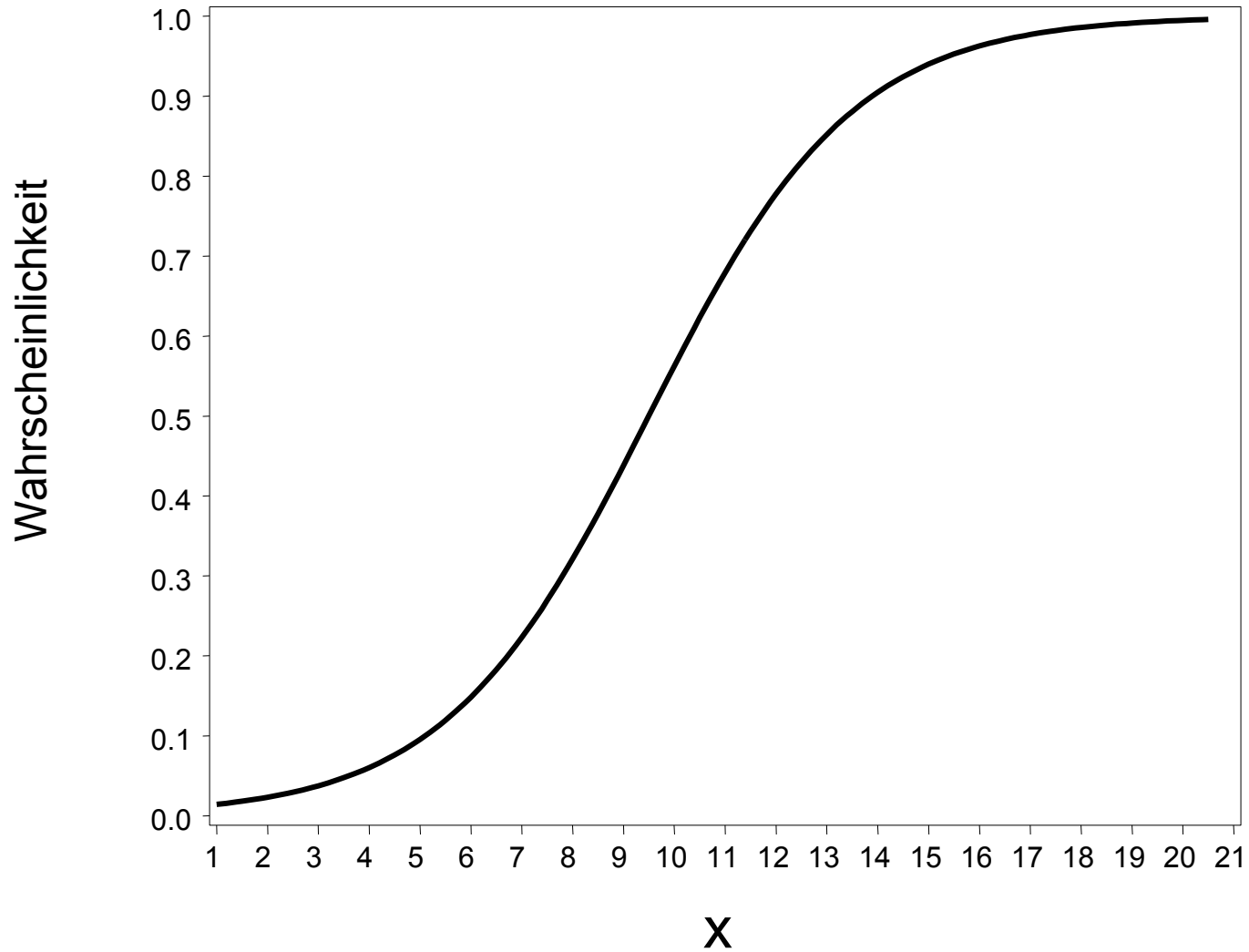
Naiver Ansatz:

$$\bullet \quad p_i = \beta_0 + \beta_1 X_1$$

- **Probleme**

- Wahrscheinlichkeiten sind beschränkt
- Beziehung zwischen Wahrscheinlichkeit und Prädiktoren im allgemeinen nicht linear

# Graph der logistischen Regression

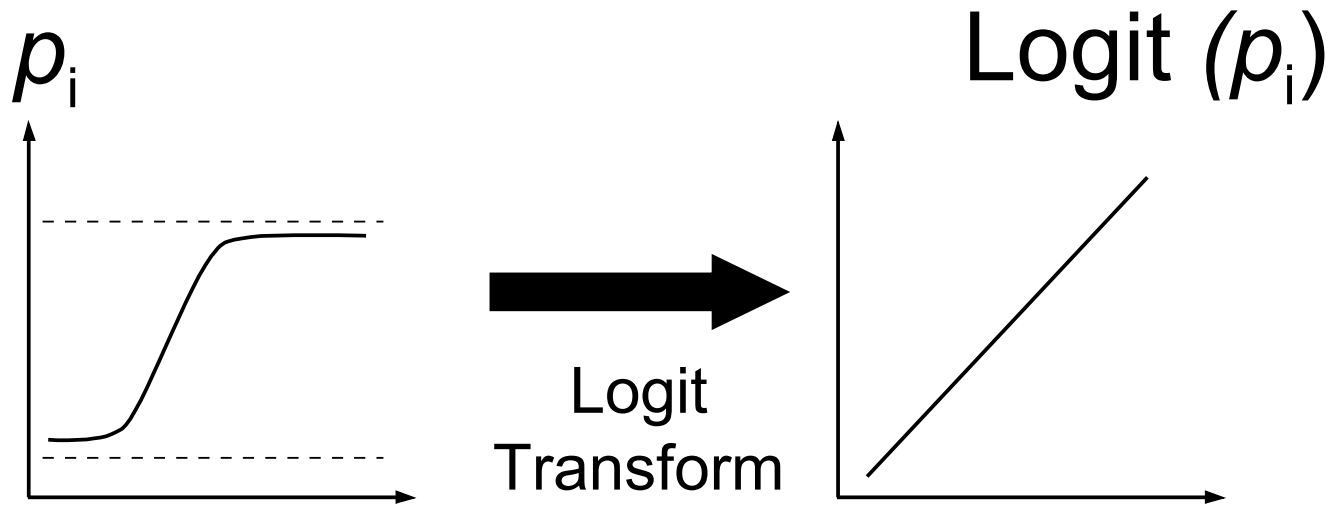


## Logistisches Regressionsmodell

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})}}$$

## Logit – Transformation: Linearisierung

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

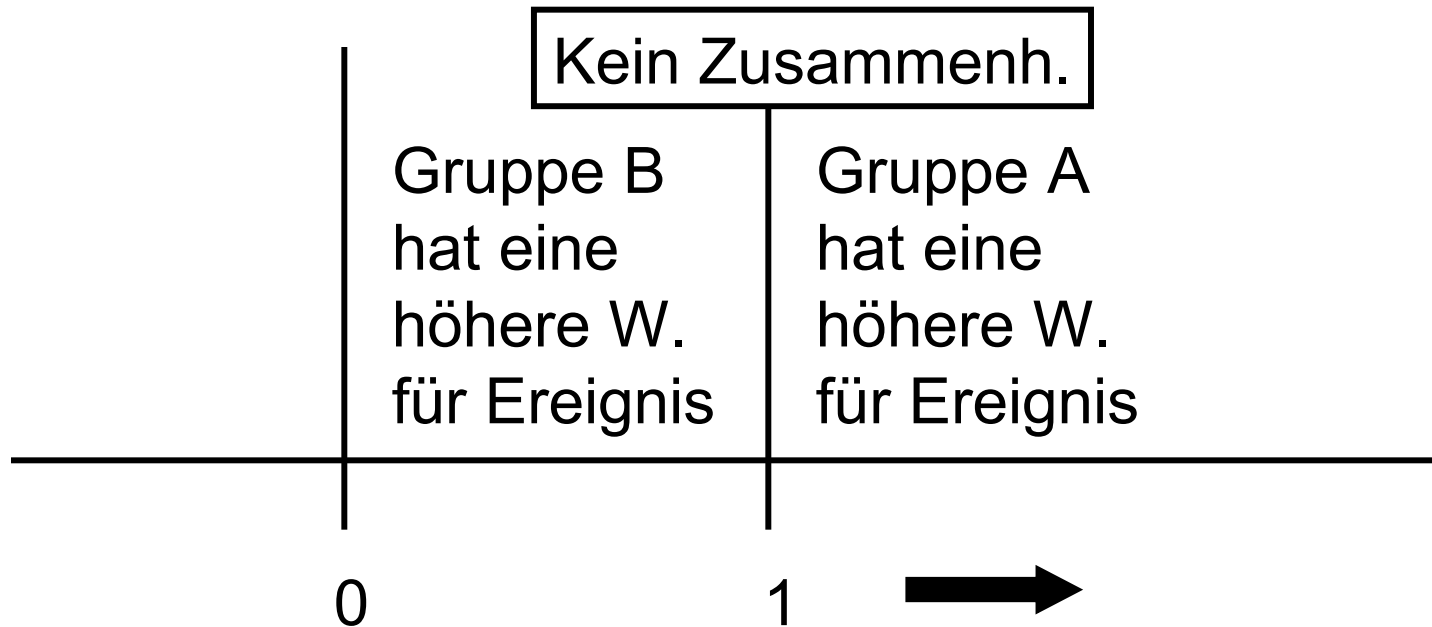


## Vorteile der Logit-Transformation

- Lineare Modelle haben schöne Eigenschaften
- Eingebettet in allgemeines Modell (GLM)
- Loglikelihoodfunktion und Maximum-Likelihood-Schätzer
- Konfidenzintervalle

## Odds Ratio - Chancenverhältnis

$$\text{Odds}_A = \frac{p_A}{1 - p_A} \quad \text{Odds}_B = \frac{p_B}{1 - p_B} \quad \text{Odds Ratio} = \frac{\text{odds}_A}{\text{odds}_B}$$



## Interpretation der Koeffizienten

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

**Koeffizient = Differenz der logit-Werte, daher:**

$$\text{Odds Ratio} = \frac{\text{odds}_{x_{1+1}}}{\text{odds}_{x_1}} = e^{\beta_1}$$

**Veränderung des Chancenverhältnisses bei Erhöhung des Prediktors um eine Einheit**

## Kriterien für Modellvergleich:

- **$R^2$  - Anteil erklärter Streuung**
- **AUC in ROC**
  - ROC: Receiver operating Characteristic
  - AUC: Area under the Curve (Trennschärfe)
- **Fehlklassifikationsrate**

## Data Mining mit JMP

- Begriffe
- Verfahren
- **Beispiele in JMP**