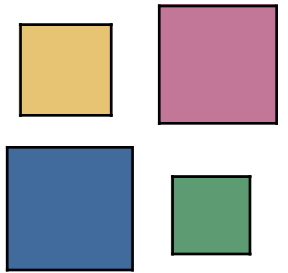


# Multivariate Statistische Verfahren

(Statistik III)



# Syllabus

- **Inhalt**

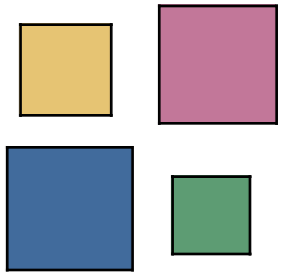
- Betrachtung von mehreren Variablen gleichzeitig.
- Reduktion von Dimensionen und Fällen.
- Klassifikation und Gruppierung von Daten.
- Graphik ist in allen Fällen sehr wichtig!

- **Vorlesung**

- Dienstag: 10:15-11:45
- Donnerstag: 12:30-14:00
- Raum: 1009

- **Übung**

- Montag: 10:15-11:45
- Raum: ???? oder Mac-Pool
- Keine Abgabe von Übungsblättern ...
- ... aber Präsenzübung – meist praktisch.



## Syllabus (cont.)

- **Software**

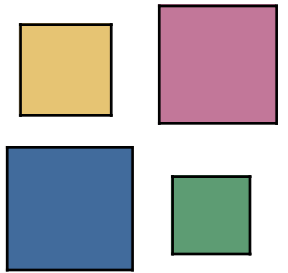
- R
- ggobi
- Mondrian

- **Ziele**

- Grundlegendes Verständnis von multivariaten Problemen
- Erfolgreiche Anwendung von statistischer Standardsoftware
- Grundwissen und Terminology für Lösung weiterer multivariater Probleme
- Anwendung der gelernten Methoden auf neue, unbekannte Daten

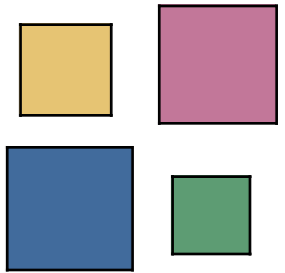
- **Leistungsnachweis**

- 20% Beteiligung in der Vorlesung
- 30% Übungen
- 50% mündliche Prüfung (20min.)



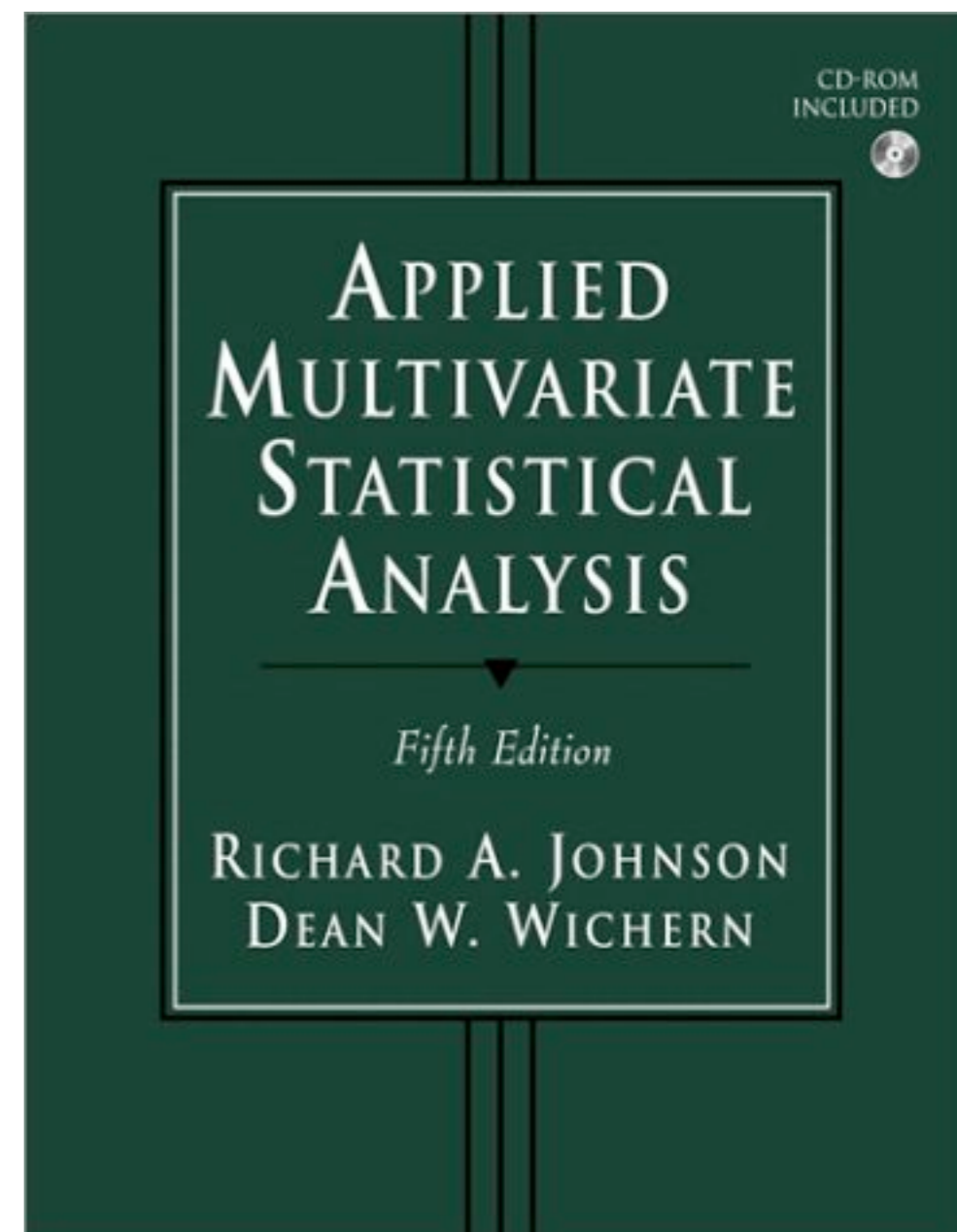
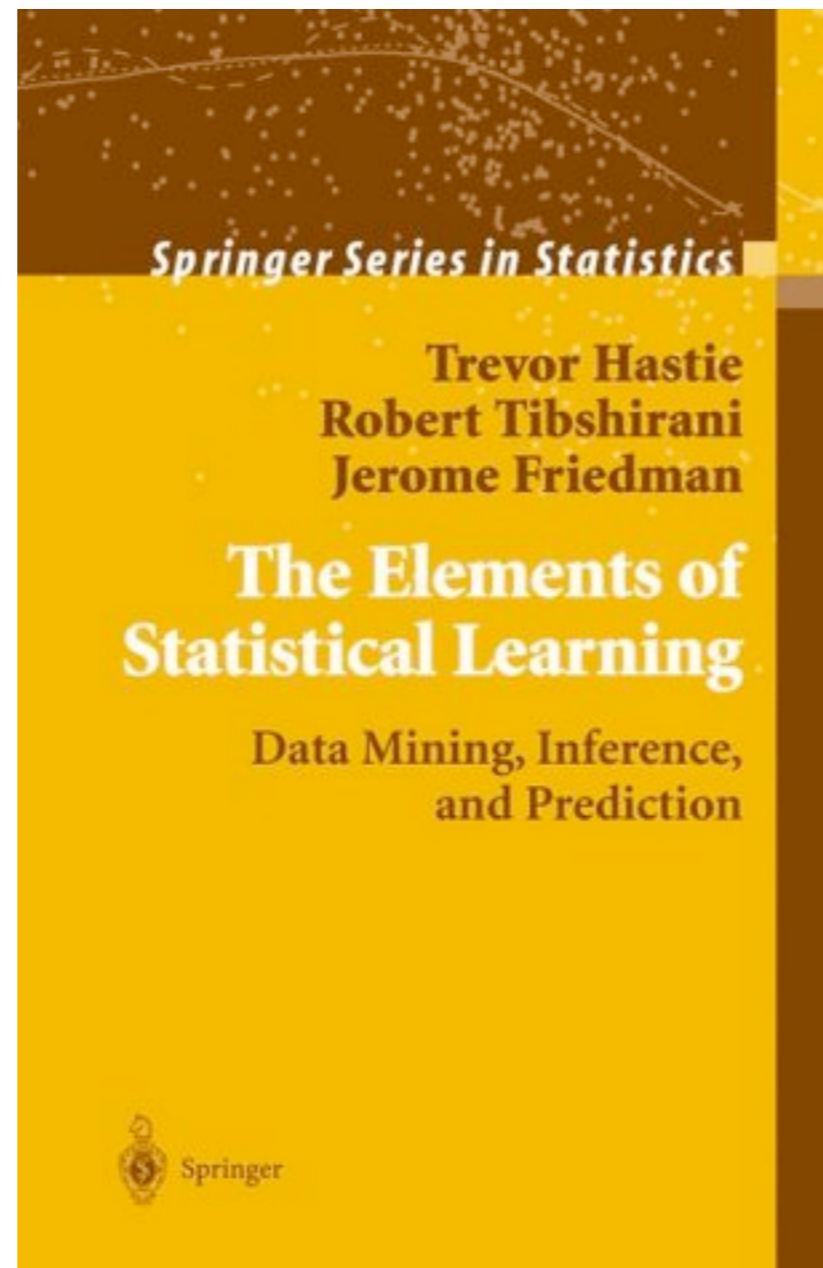
# Syllabus: Themen

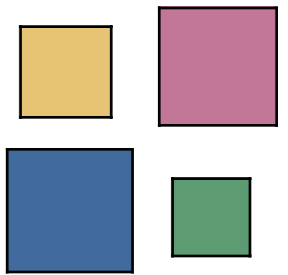
- **Graphik**
  - Grundlagen und Konzepte Interaktive Graphik
  - Multivariate Plots für stetige Daten: Parallele Koordinaten
  - Multivariate Plots für diskrete Daten: Mosaik Plots
- **Klassische Methoden**
  - Multivariate Normalverteilung
  - Inferenz
  - Hauptkomponenten Analyse (PCA)
  - Faktoren Analyse
  - Multidimensionale Skalierung (MDS)
  - Kanonische Korrelation (CC)
- **Lernmethoden**
  - Clusteranalyse
  - Lineare und quadratische Diskriminanz Analyse (LDA/QDA)
  - Klassifikations und Regressionsbäume (CART)
  - Feed Forward Neural Networks (FFN)
  - Support Vector Machines (SVM)



# Syllabus (cont.)

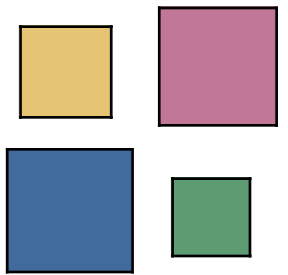
- Literatur





## Data Mining: Ursprung

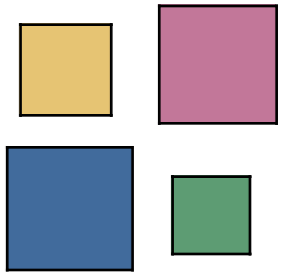
- John Tukey (1962)
  - Statistics - concerned with data analysis - should be defined in terms of a set of problems (as are most fields) rather than a set of tools, namely those problems that pertain to data.
- Brad Efron (19??)
  - “Statistics has been most successful in information science.”
  - “Those who ignore statistics are condemned to reinvent it.”
- Daryl Pregibon (1999)
  - KDD = statistics at scale & speed
- Jerry Friedman (1999)
  - “Every time the amount of data increases by a factor of ten, we should completely rethink how we analyze data.”
  - “We will also have to expand our curriculum to include current computer oriented data analysis methodology ...”
- Bill Cleveland (1999)
  - Data Science: Expanding the technical Areas of Statistics



## Data Mining: Definition

- Viele Definitionen ... die meisten enthalten die "Die Entdeckung von bisher unbekanntem Eigenschaften"
- Hier ist meine:

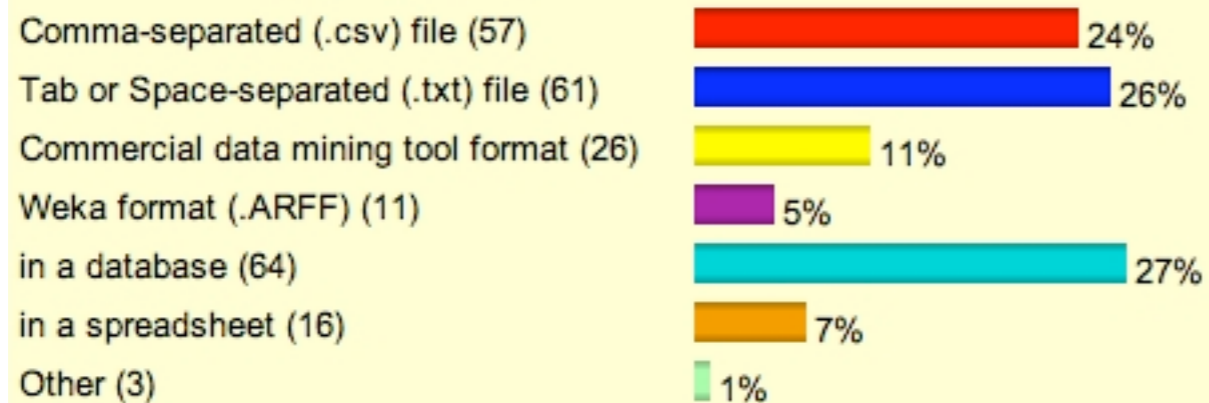
*Data Mining ist weder eine Methode noch eine Disziplin, sondern beschreibt die Schnittmenge von Methoden der Statistik und Informatik, in der computerintensive Methoden auf relativ große und komplexe Daten angewendet werden, um deren Eigenschaften mit statistischen Mitteln zu beschreiben.*



# Data Mining: Umfragen

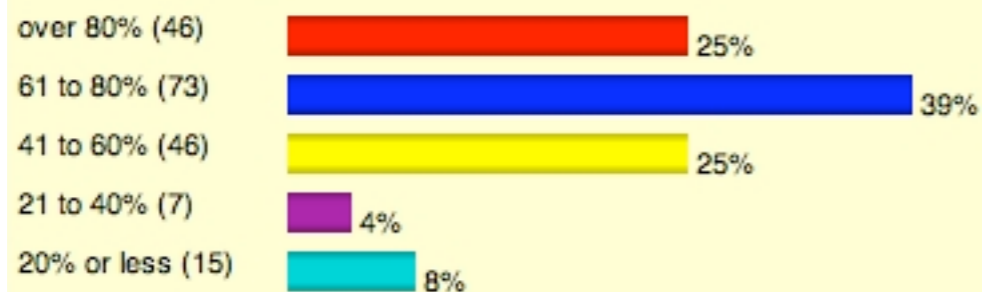
- Datenquellen

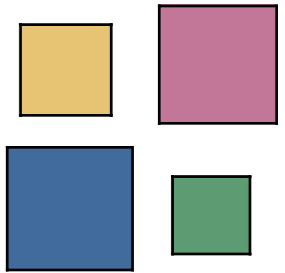
What dataset format you use the most when data mining? [238 votes total]



- Datenaufbereitung

What % of time in your data mining project(s) is spent on data cleaning and preparation [187 votes total]

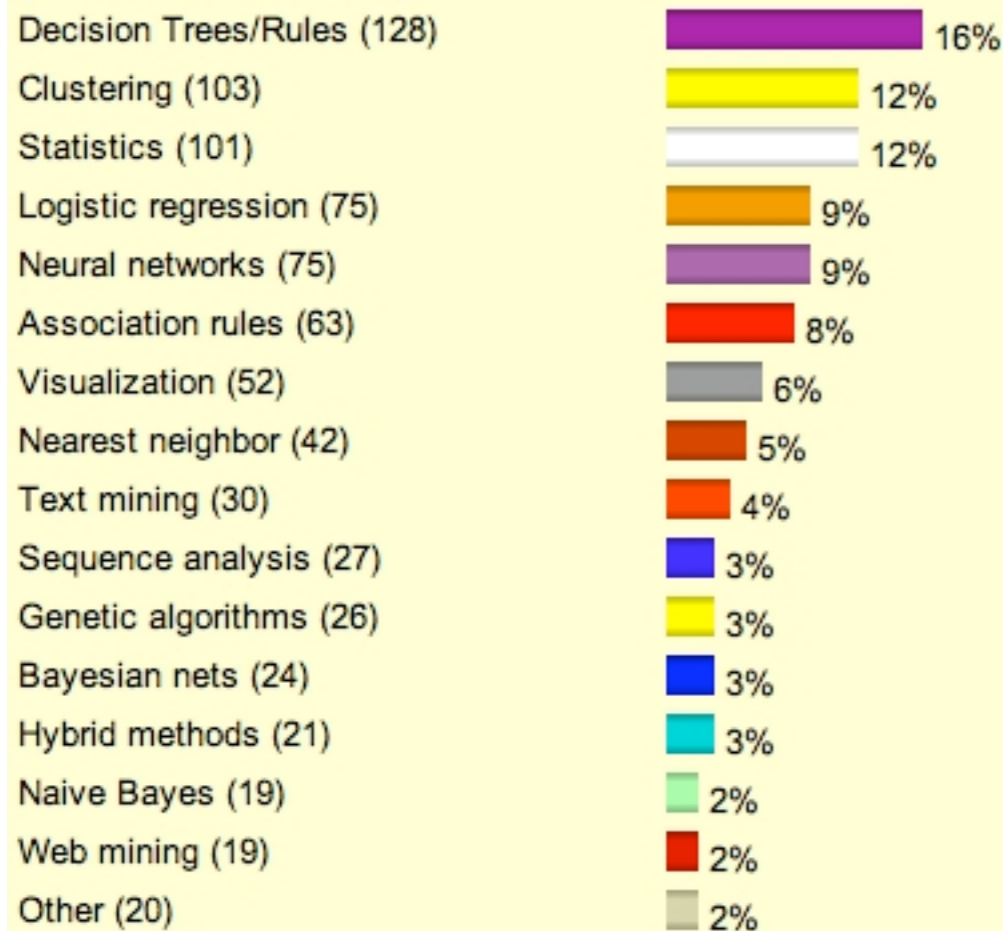




# Umfrage (cont.)

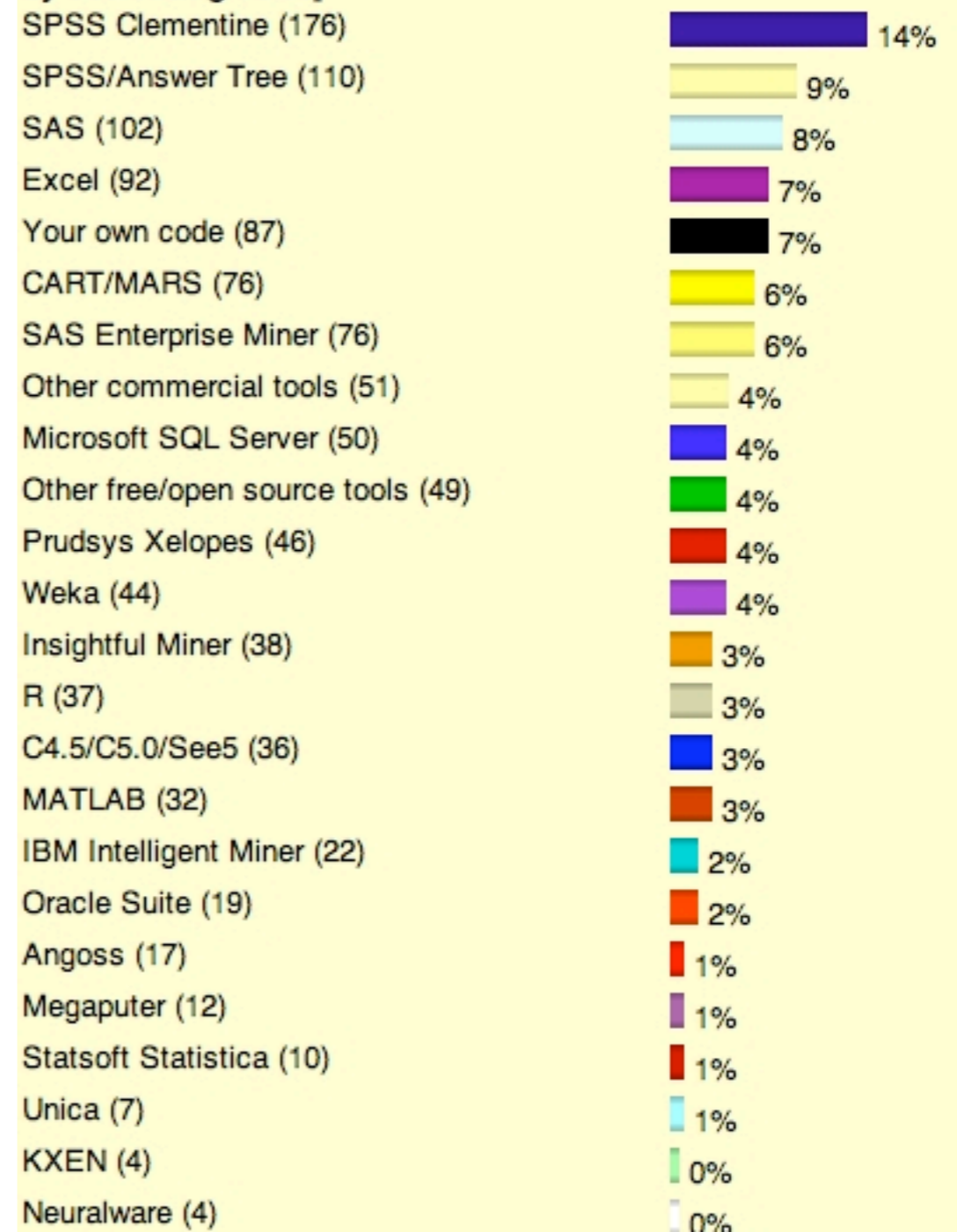
## • Techniken

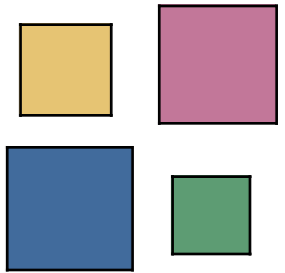
Which data mining techniques do you use regularly? (Choose several)  
[212 votes, 825 choices]



## • Tools

Data mining tools you regularly use: [628 responders, 1252 votes; sorted by decreasing votes]





## Beispiel: Australische Krabben

- **Variablen**

- Species (blue, orange)
- Sex (male, female)
- Index
- FL
- RW
- CL
- CW
- BD

- 200 Fälle,  
je 100 pro Gruppe und Geschlecht

