

Skript zur Vorlesung
Knowledge Discovery in Databases
im Wintersemester 2008/2009

Kapitel 1: Einleitung

Vorlesung+Übungen:
Dr. Matthias Schubert

Skript © 2003 Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Jörg Sander und Matthias Schubert

<http://www.dbs.ifi.lmu.de/Lehre/KDD>

- **Aktuelles**
 - Vorlesung: Dienstag, 13-16 Uhr (1.14, Oettingenstr. 67)
 - Übung: Freitag, 12-14 Uhr (0.37, Oettingenstr. 67)
Freitag, 14-16 Uhr (0.37, Oettingenstr. 67)
- Anmeldung für den Übungsbetrieb auf der Homepage unter <http://www.dbs.ifi.lmu.de/Lehre/KDD>
(Anmeldeschluß : **7 Dezember 2008**)
- Klausur: Der Stoff der Klausur wird in der Vorlesung und in den Übungen besprochen.
(Das Skript ist lediglich eine Lernhilfe.)

Digitalkameras



Kreditkarten



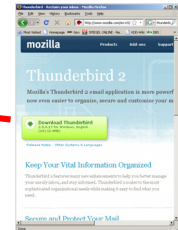
Scanner-Kassen



Astronomie



Telefongesellschaft



WWW

- Riesige Datenmengen werden in Datenbanken gesammelt
- Analysen können nicht mehr manuell durchgeführt werden

	Daten	Methode	Wissen
	Verbindungs- Rechnungserst.	Outlier Detection	Betrug
	Transaktionen Abrechnung	Klassifikation	Kreditwürdigkeit
	Transaktionen Lagerhaltung	Assoziationsregeln	Gemeinsam gekaufte Produkte
	Bilddaten Kataloge	Klassifikation	Klasse eines Sterns

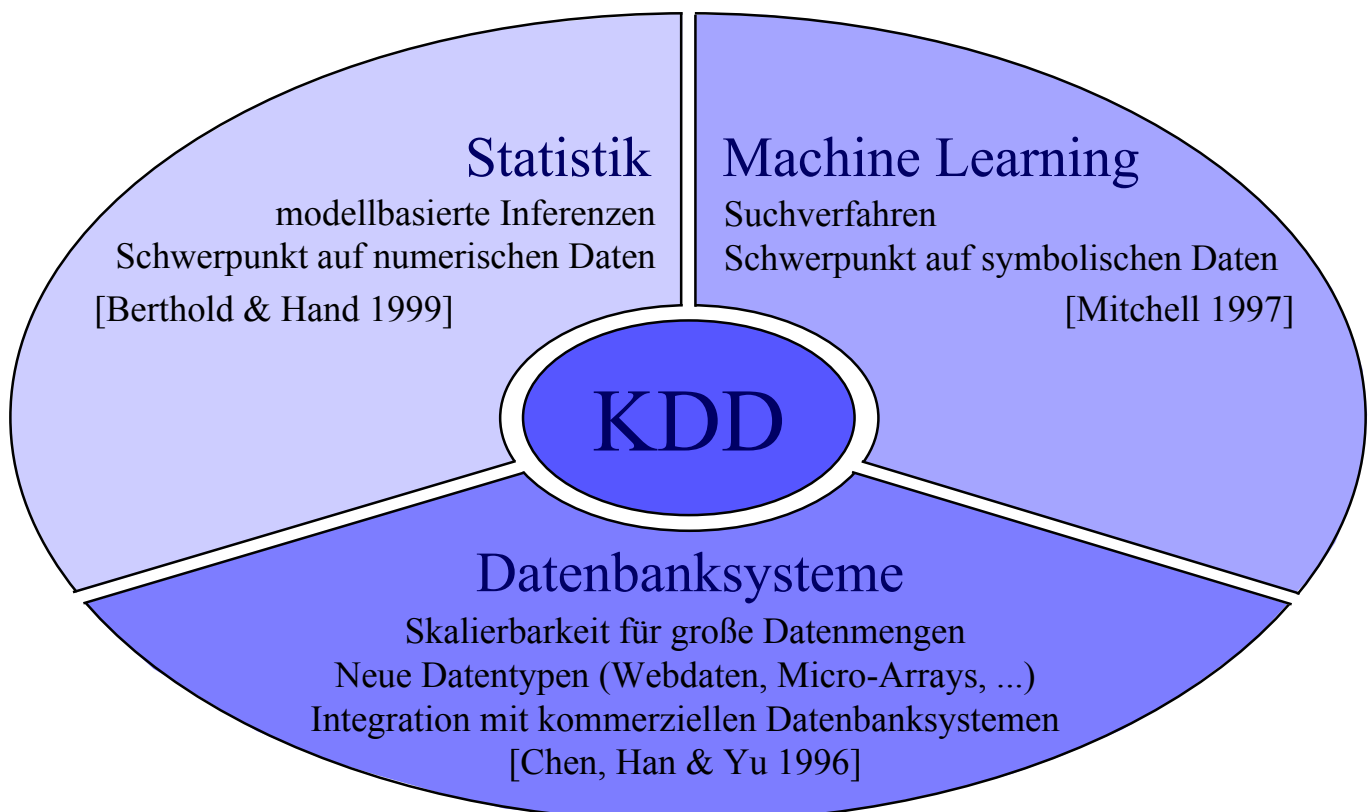
[Fayyad, Piatetsky-Shapiro & Smyth 1996]

Knowledge Discovery in Databases (KDD) ist der Prozess der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das

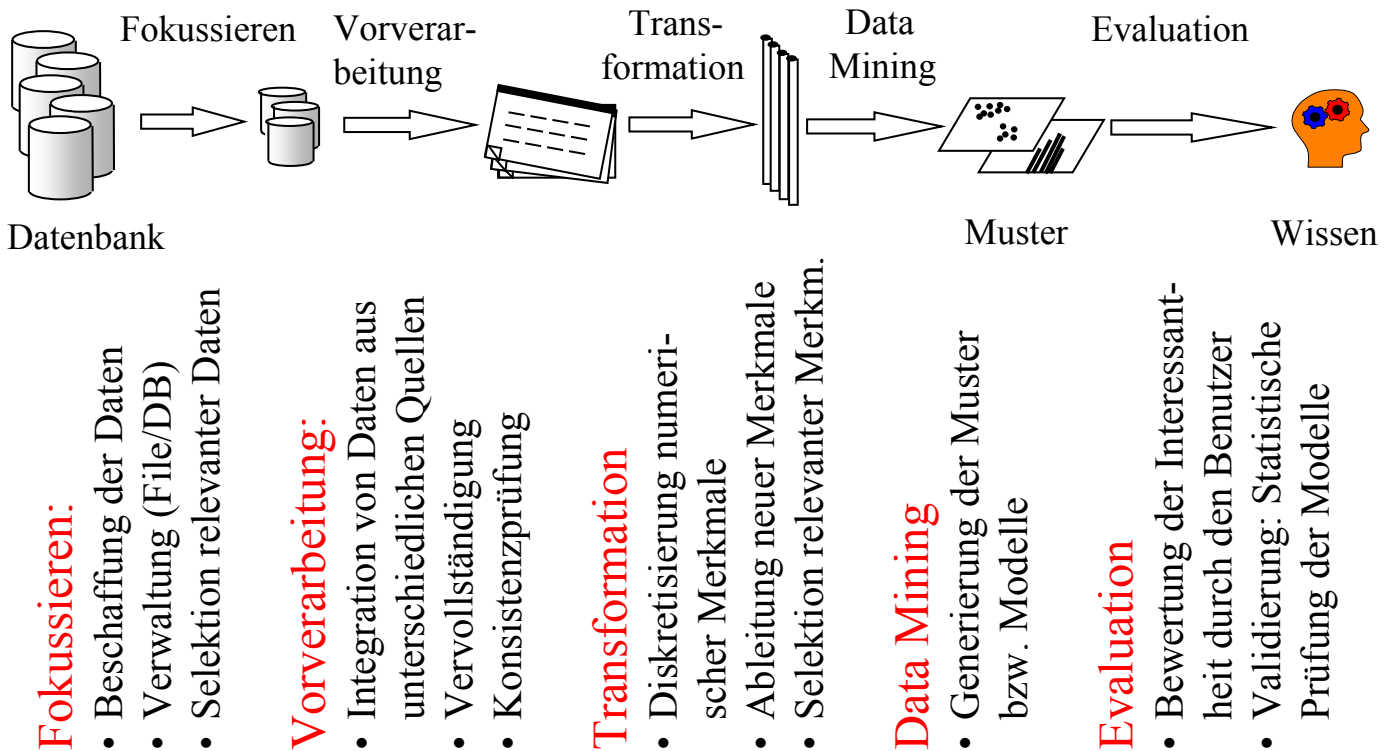
- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

Bemerkungen:

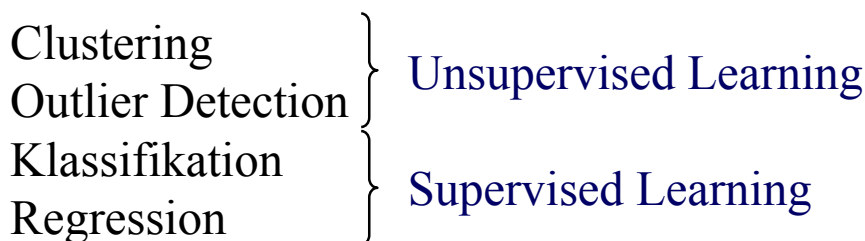
- *(semi-) automatisch*: im Unterschied zu manueller Analyse. Häufig ist trotzdem Interaktion mit dem Benutzer nötig.
- *gültig*: im statistischen Sinn.
- *bisher unbekannt*: bisher nicht explizit, kein „Allgemeinwissen“.
- *potentiell nützlich*: für eine gegebene Anwendung.



Prozessmodell nach Fayyad, Piatetsky-Shapiro & Smyth



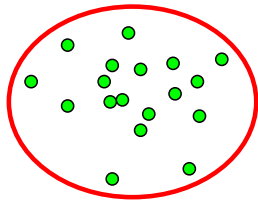
Wichtigste Data-Mining-Verfahren auf **Merkmals-Vektoren**:



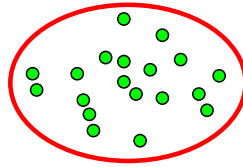
Supervised: Ein Ergebnis-Merkmal soll gelernt/geschätzt werden

Unsupervised: Die Datenmenge soll lediglich in Gruppen unterteilt werden

Darüber hinaus gibt es zahlreiche Verfahren, die nicht auf Merkmalsvektoren, sondern z.B. auf **Texten**, **Mengen**, **Graphen** arbeiten.



Cluster 1: Kammern



Cluster 2: Nägel

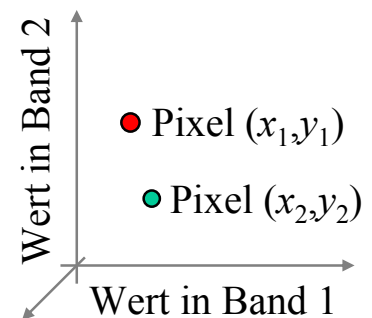
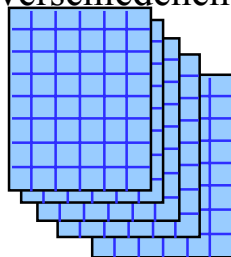
Clustering heißt: Zerlegung einer Menge von Objekten (bzw. Feature-Vektoren) so in Teilmengen (Cluster), dass

- die Ähnlichkeit der Objekte innerhalb eines Clusters maximiert
- die Ähnlichkeit der Objekte verschiedener Cluster minimiert wird

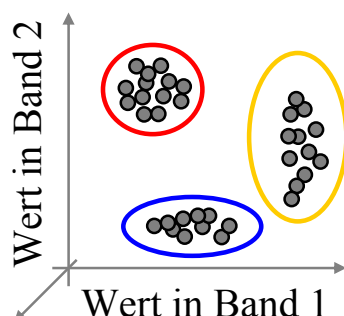
Idee: Die verschiedenen Cluster repräsentieren meist unterschiedliche Klassen von Objekten; bei unbek. Anzahl und Bedeutung der Klassen



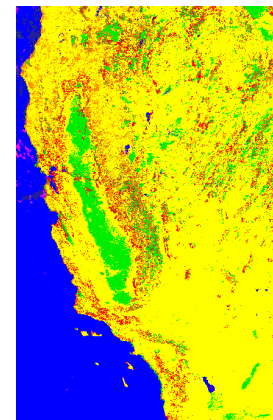
Aufnahme der Erdoberfläche
in 5 verschiedenen Spektren



Cluster-Analyse

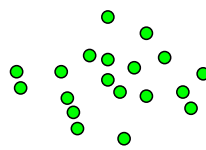
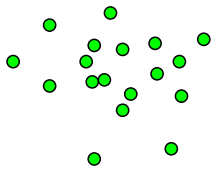


Rücktransformation
in xy -Koordinaten
Farbcodierung nach
Cluster-Zugehörigkeit





Datenfehler?
Betrug?



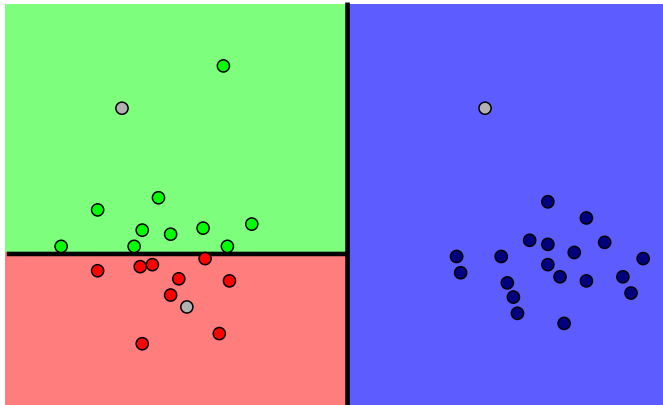
Outlier Detection bedeutet:
Ermittlung von **untypischen** Daten

Anwendungen:

- Entdeckung von Missbrauch etwa bei
 - Kreditkarten
 - Telekommunikation
- Datenfehler

- Analyse der SAT.1-Ran-Fußball-Datenbank (Saison 1998/99)
 - 375 Spieler
 - Primäre Attribute: Name, Einsätze, Tore, Spielposition (Torwart, Abwehr, Mittelfeld, Sturm),
 - Abgeleitetes Attribut: Tore pro Spiel
 - Outlier Analyse auf (Spielposition, Einsätze, Tore pro Spiel)
- Ergebnis: Top 5 Outliers

Rang	Name	Einsätze	Tore	Position	Erklärung
1	Michael Preetz	34	23	Sturm	Torschützenkönig
2	Michael Schjönberg	15	6	Abwehr	Abwehrspieler mit den meisten Toren
3	Hans-Jörg Butt	34	7	Torwart	Torwart mit den meisten Toren
4	Ulf Kirsten	31	19	Sturm	2. Torschützenkönig
5	Giovane Elber	21	13	Sturm	Hohe Tore-pro-Spiel Quote



- Schrauben
 - Nägel
 - Klammern
- } Trainingsdaten
- Neue Objekte

Aufgabe:

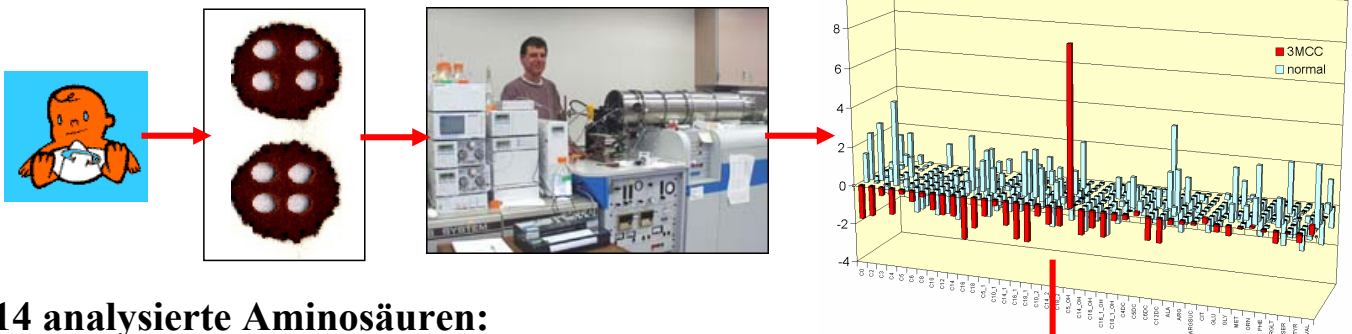
Lerne aus den bereits klassifizierten *Trainingsdaten* die *Regeln*, um neue Objekte nur aufgrund der Merkmale zu klassifizieren

Das Ergebnismerkmal (Klassenvariable) ist nominal (*kategorisch*)

Blutprobe des
Neugeborenen

Massenspektrometrie

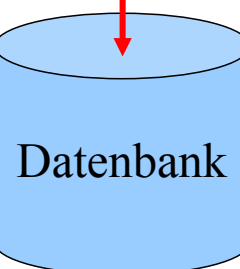
Metabolitenspektrum

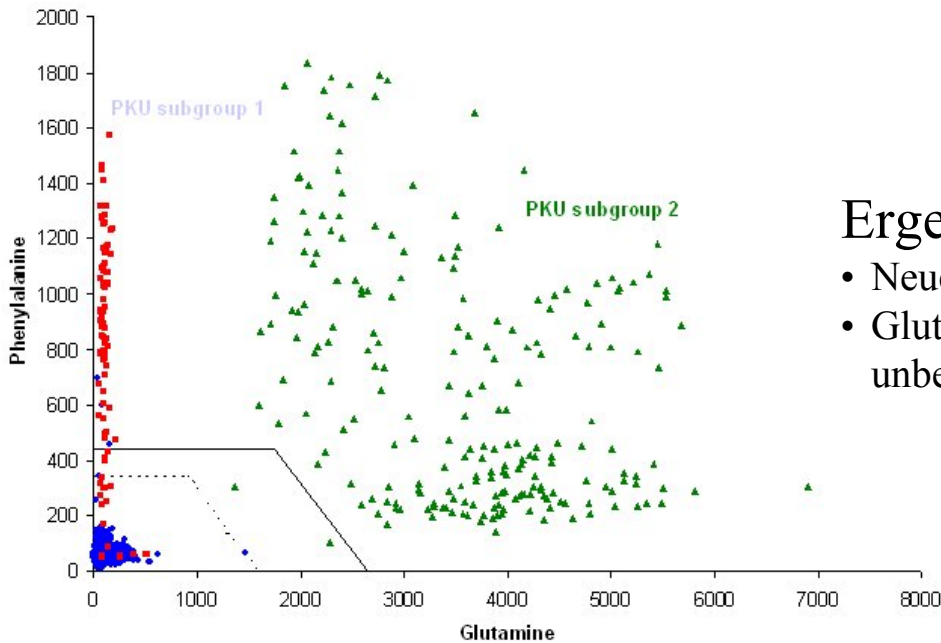


14 analysierte Aminosäuren:

alanine
arginine
argininosuccinate
citrulline
glutamate
glycine
methionine

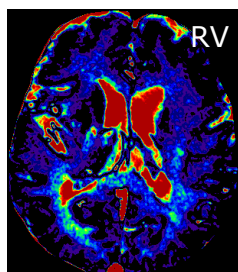
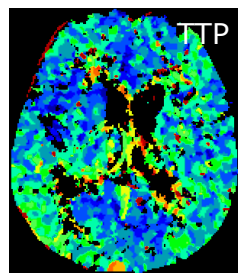
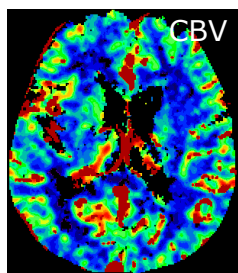
phenylalanine
pyroglutamate
serine
tyrosine
valine
leucine+isoleucine
ornitine



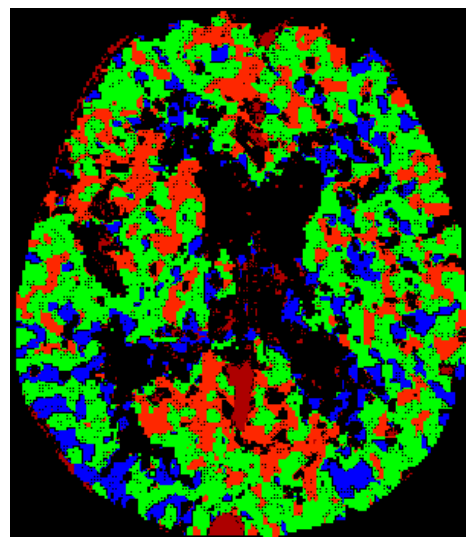


Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

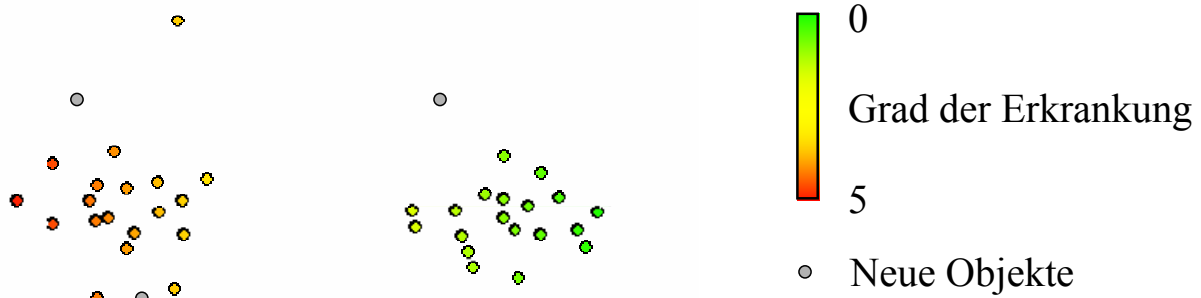


- Schwarz: Ventrikel + Hintergrund
- Blau: Gewebe 1
- Grün: Gewebe 2
- Rot: Gewebe 3
- Dunkelrot: Große Gefäße



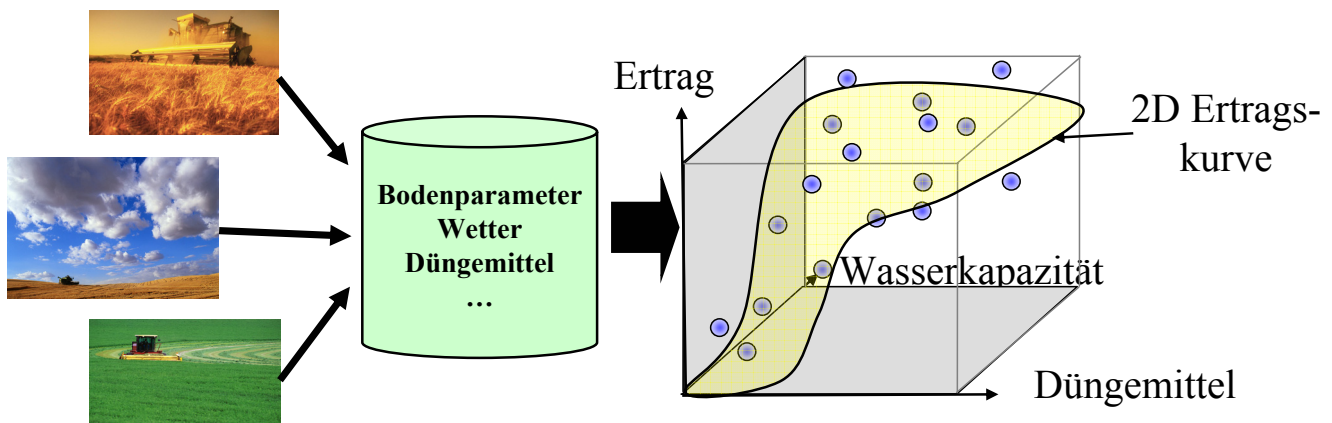
	Blau	Grün	Rot
TTP (s)	20.5	18.5	16.5
CBV (ml/100g)	3.0	3.1	3.6
CBF (ml/100g/min)	18	21	28
RV	30	23	21

Ergebnis: Klassifikation cerebralen Gewebes anhand funktioneller Parameter mittels dynamic CT möglich.



Aufgabe:

Ähnlich zur Klassifikation, aber das Ergebnis-Merkmal, das gelernt bzw. geschätzt werden soll, ist *metrisch*

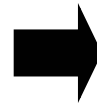


- Erstellen einer Ertragskurve, die von mehreren Parametern wie Bodenbeschaffenheit, Wetter und Düngemittelausbringung abhängt.
- Erst eine geeignete Anpassung der Düngemittelausbringung kann eine ertragsoptimale Nutzung in Abhängigkeit von Umweltfaktoren bewirken.
- Das Thema ist auch wegen der Umweltbelastung durch Überdüngung wichtig.

a,b,c,d,e
b,c,d
a,b,c,d
a,b,c,d,e
a,c,e,f
d,c,e,f
a,b,c,d,f



In 5 von 7 (ca. 71 %) der Fälle kommt **b,c,d** zusammen vor.

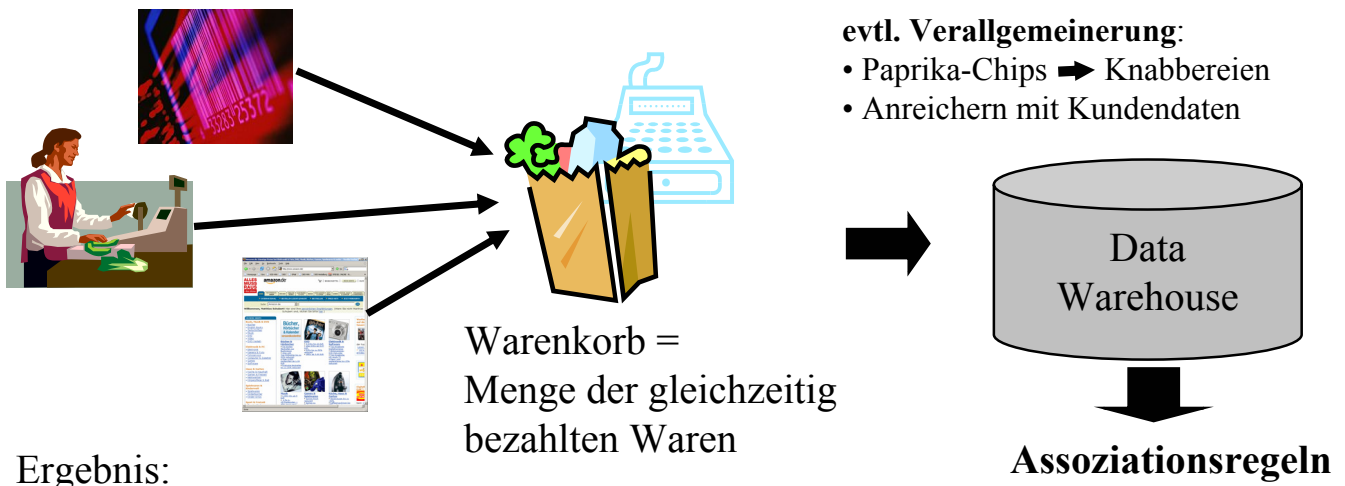


In 5 von 5 Fällen (100 %) gilt:
Wenn **b,c** in der Menge, dann ist auch **d** in der Menge.

Aufgabe:

Finde alle Regeln in einer Datenbank von diskreten Mengen der folgenden Art:

Wenn a, b, c in der Menge M enthalten sind, dann ist auch t mit einer Wahrscheinlichkeit vom $X\%$ in der Menge enthalten.



Ergebnis:

- Häufig zusammen gekaufte Artikel können im Supermarkt besser zueinander positioniert werden: Windeln werden häufig mit Bierkästen zusammen gekauft
=> Positioniere Bier auf dem Weg von Windeln zur Kasse
- Generiere Empfehlungen für Kunden mit ähnlichen Warenkörbe:
Kunden die „Krieg der Sterne“ I-VI gekauft haben, sind vielleicht auch an „Herr der Ringe“ I-III interessiert.

- | | |
|------------------|---|
| 1 Einleitung | 6. Outlier Detection |
| 2 Merkmalsräume | 7. Assoziationsregeln |
| 3 Klassifikation | 8. Data Warehousing und Generalisierung |
| 4 Regression | 9. High-Performance Data Mining |
| 5 Clustering | 10. Ausblick: KDD 2 und |
- Maschinelles Lernen und Data Mining

Lehrbuch zur Vorlesung (deutsch):

Ester M., Sander J.

Knowledge Discovery in Databases: Techniken und Anwendungen

ISBN: 3540673288, Springer Verlag, September 2000, € 39,95

Weitere Bücher (englisch):

Berthold M., Hand D. J. (eds.)

Intelligent Data Analysis: An Introduction

ISBN: 3540430601, Springer Verlag, Heidelberg, 1999, € 63,24

Han J., Kamber M.

Data Mining: Concepts and Techniques

ISBN: 1558609016, Morgan Kaufmann Publishers, March 2006, € 54,95

Mitchell T. M.

Machine Learning

ISBN: 0071154671, McGraw-Hill, 1997, € 61,30

Witten I. H., Frank E.

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations

ISBN: 1558605525, Morgan Kaufmann Publishers, 2000, € 50,93

