

Skript zur Vorlesung
Knowledge Discovery in Databases
im Sommersemester 2011

Kapitel 1: Einleitung

Vorlesung+Übungen:
PD Dr. Matthias Schubert, Dr. Arthur Zimek

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

[http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_\(KDD_I\)](http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I))

- **Aktuelles**
 - Vorlesung: Dienstag, 9.00-12.00 Uhr (Raum E 216 Hauptgebäude)
 - Übung: Dienstag, 16-18 Uhr (Raum D Z007 Hauptgebäude)
Donnerstag, 16-18 Uhr (Raum C 113 Theresienstr. 41)
- Anmeldung für die Klausur auf der Homepage unter [http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_\(KDD_I\)](http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I))
- 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

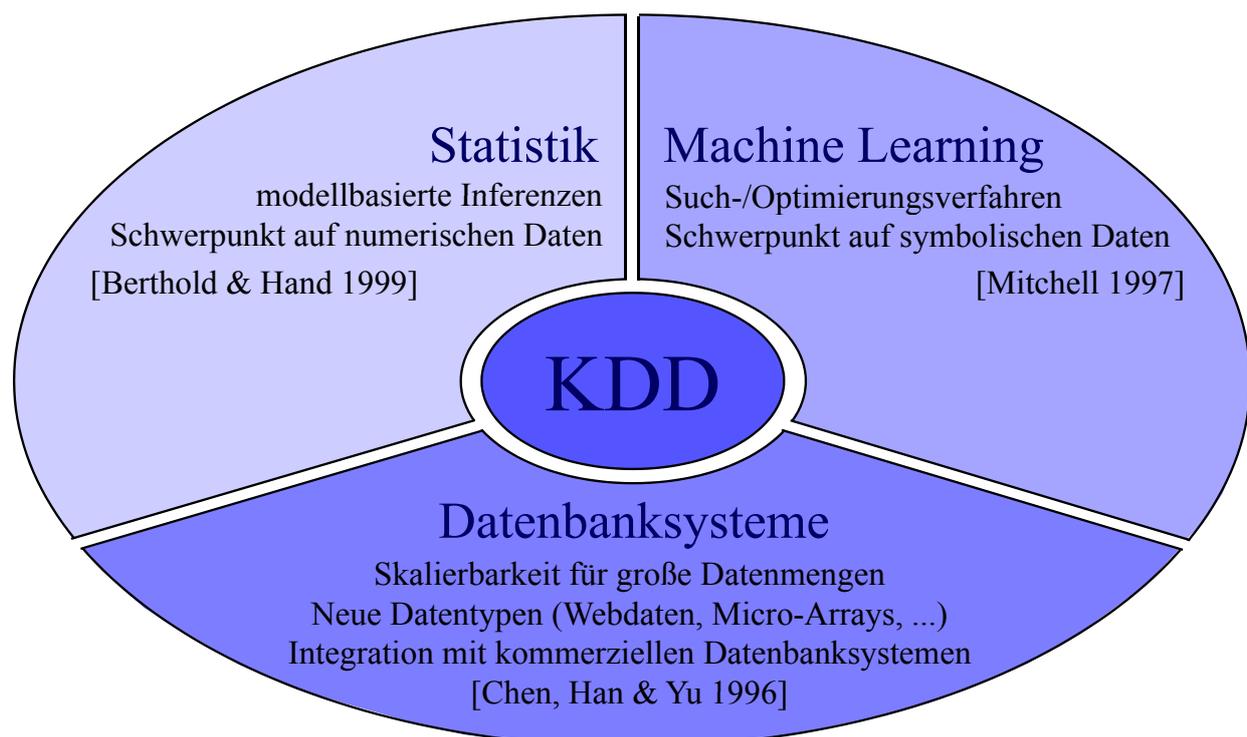
[Fayyad, Piatetsky-Shapiro & Smyth 1996]

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

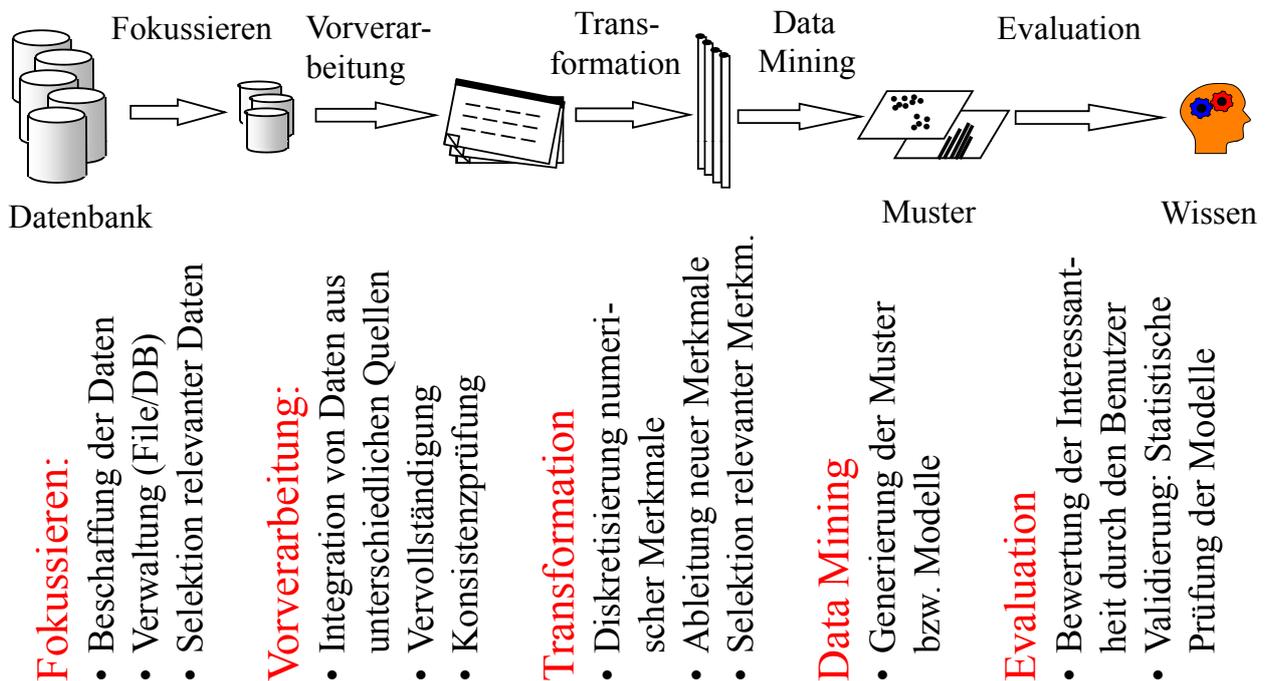
- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

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

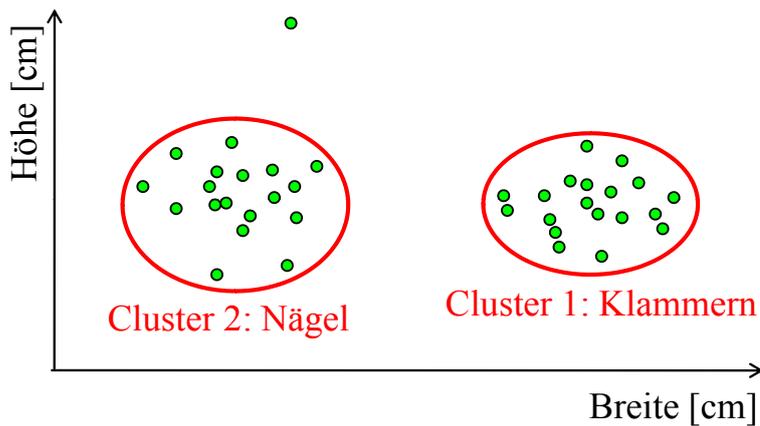


Die wichtigsten Data-Mining-Techniken:

Supervised: z.B. Klassifikation, Regression, Outlier Detection
Ein Ergebnis-Merkmal soll aufgrund von Vorwissen gelernt/geschätzt werden.
Das Vorwissen steht typischerweise als Trainingsdaten bereit.

Unsupervised: z.B. Clustering, Outlier Detection, Assoziationsregeln
Die Datenmenge soll ohne weiteres Vorwissen in Gruppen unterteilt werden.
Die Gruppen haben je nach Aufgabe unterschiedliche Charakteristika.

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

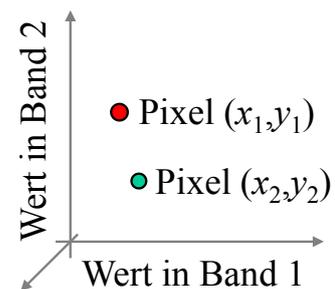
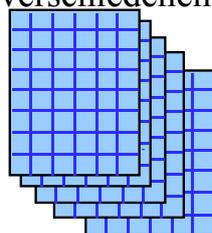


Clustering heißt: Zerlegung einer Menge von Objekten (bzw. Feature-Vektoren) in Teilmengen (Cluster) ähnlicher Objekte

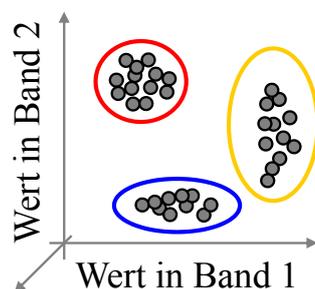
Idee: Die verschiedenen Cluster repräsentieren meist unterschiedliche Klassen von Objekten; bei unbek. Anzahl und Bedeutung der Klassen



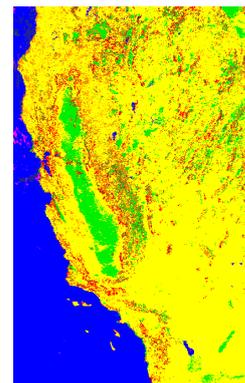
Aufnahme der Erdoberfläche in 5 verschiedenen Spektren



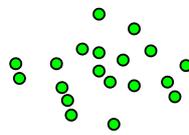
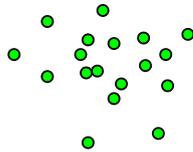
Cluster-Analyse



Rücktransformation in xy -Koordinaten
 Farbcodierung nach Cluster-Zugehörigkeit



 Datenfehler?
Betrug?



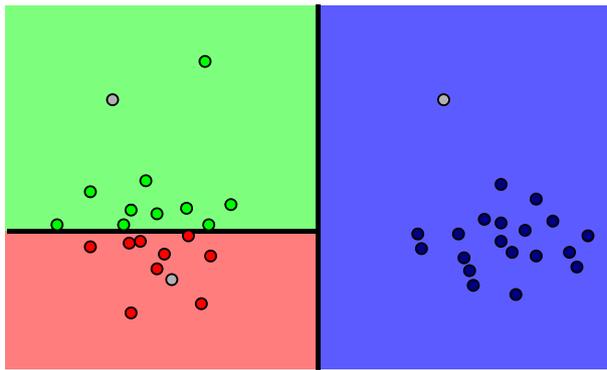
Outlier Detection bedeutet:
Ermittlung von **untypischen** Daten

Idee: Outlier könnten hindeuten auf

- 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



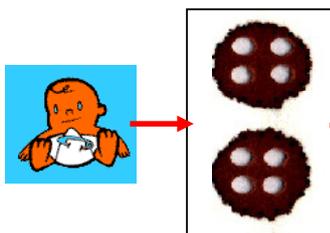
- Schrauben
 - Nägel
 - Klammern
- } Trainingsdaten
- Neue Objekte

Aufgabe:

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

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

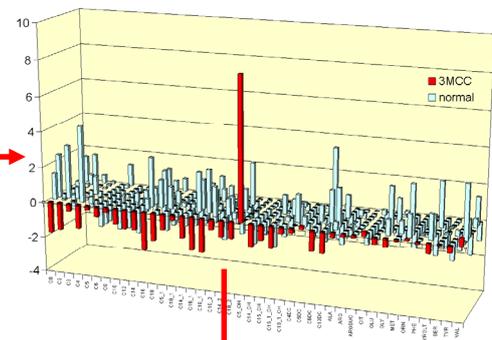
Blutprobe des Neugeborenen



Massenspektrometrie

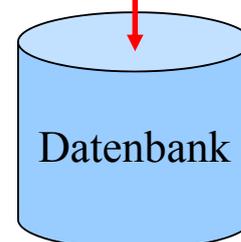


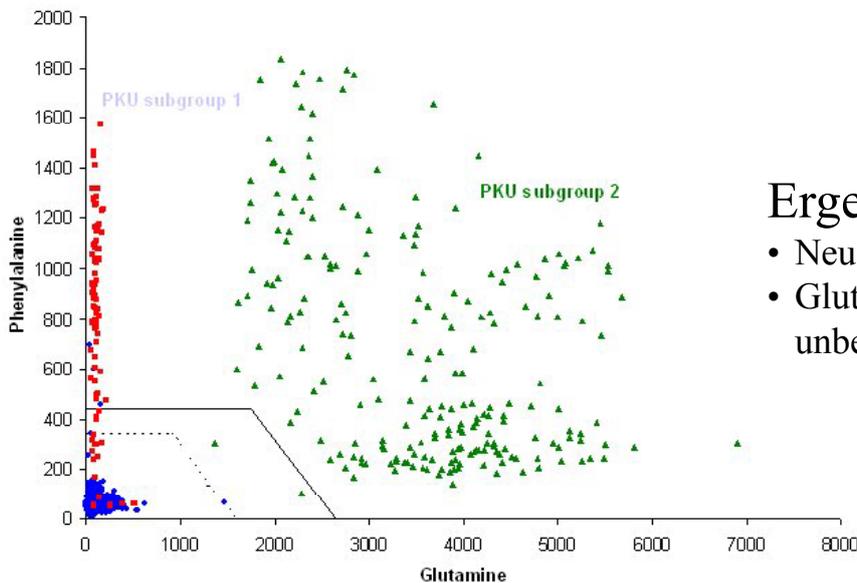
Metabolitenspektrum



14 analysierte Aminosäuren:

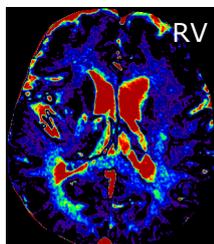
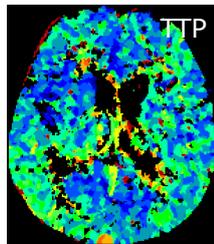
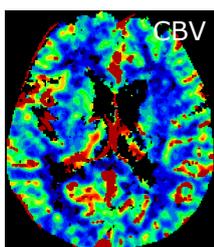
- | | |
|-------------------|--------------------|
| alanine | phenylalanine |
| arginine | pyroglutamate |
| argininosuccinate | serine |
| citrulline | tyrosine |
| glutamate | valine |
| glycine | leucine+isoleucine |
| methionine | ornitine |



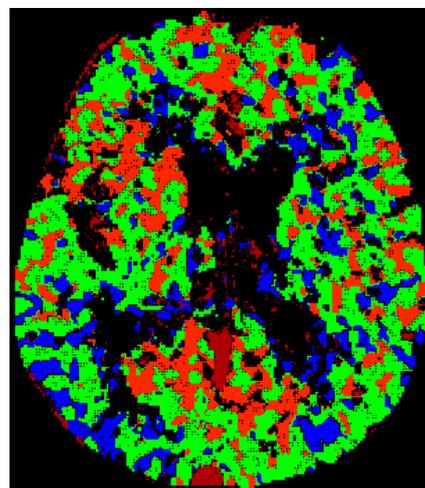


Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

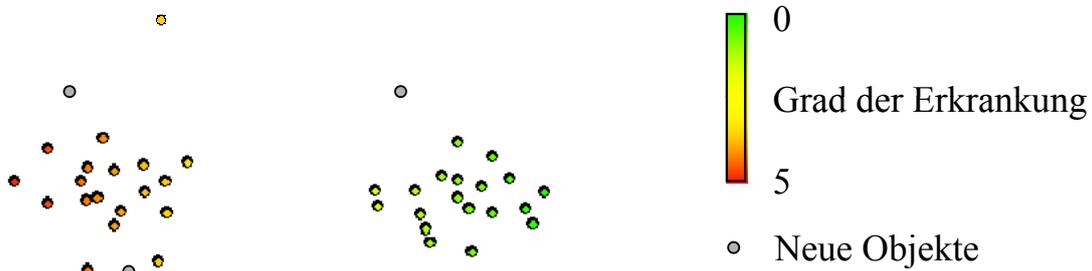


- Schwarz: Ventrikel + Hintergrund
- Blau: Gewebe 1
- Grün: Gewebe 2
- Rot: Gewebe 3
- Dunkelrot: Große Gefäße



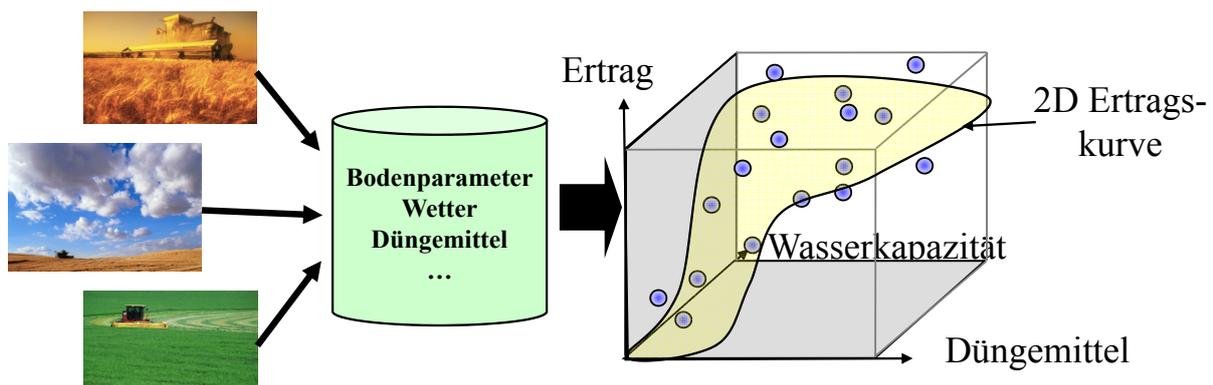
	Blau	Grün	Rot
TTP (s)	20.5	18.5	16.5
CBV (ml/100g)	3.0	3.1	3.6
CBF (ml/100g/min)	18	21	28
RV	30	23	21

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

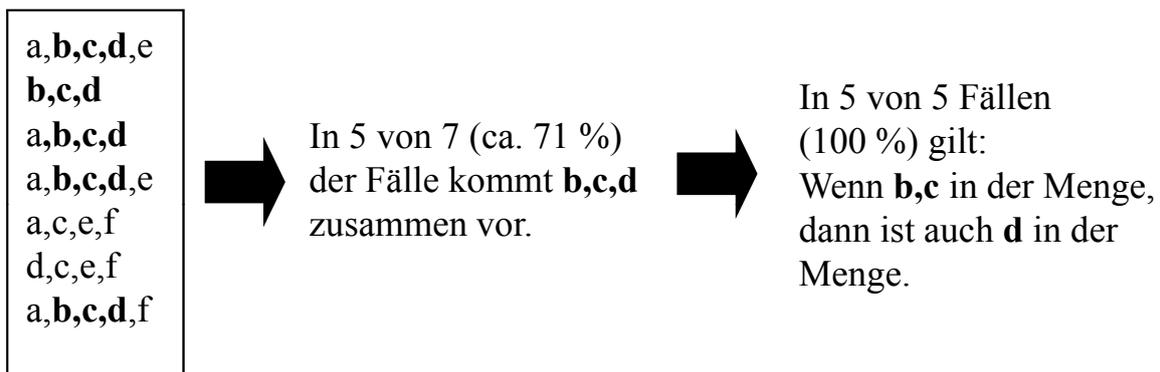


Aufgabe:

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



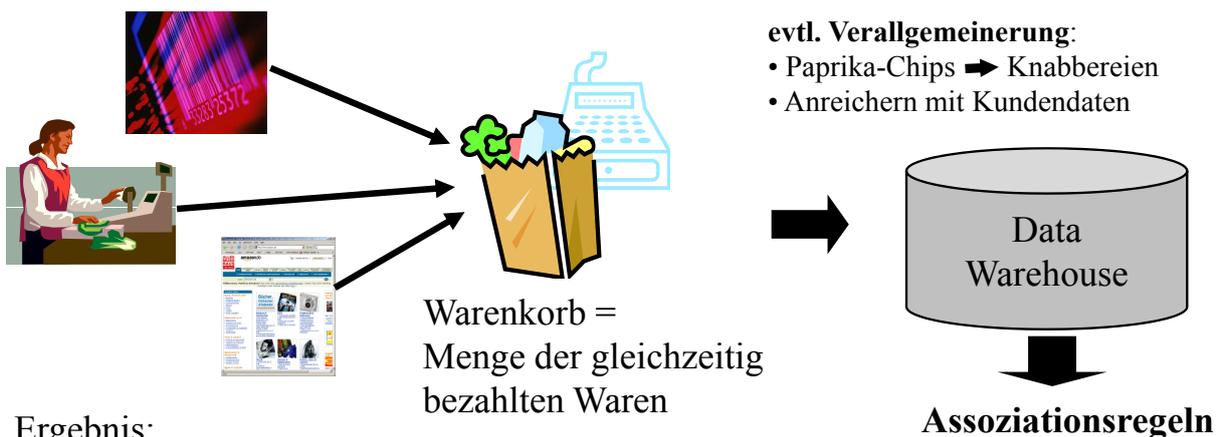
- Erstellen einer Ertragskurve, die von mehreren Parametern wie Bodenbeschaffenheit, Wetter und Düngemittelausbringung abhängt.
- Erst eine geeignete Anpassung der Düngemittelausbringung kann eine ertragsoptimale Nutzung in Abhängigkeit von Umweltfaktoren bewirken.
- Das Thema ist auch wegen der Umweltbelastung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



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 \Rightarrow 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. DB-Techniken zur Leistungssteigerung |
| 4. Regression | |
| 5. Clustering | 9. Ausblick: KDD 2 und Maschinelles Lernen und Data Mining |

Lehrbuch zur Vorlesung (deutsch):

Ester M., Sander J.

Knowledge Discovery in Databases: Techniken und Anwendungen

Springer Verlag, September 2000

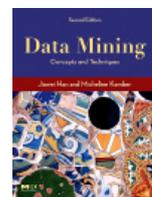


Weitere Bücher (englisch):

Han J., Kamber M.

Data Mining: Concepts and Techniques

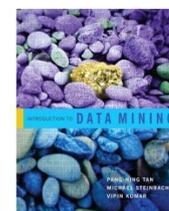
Morgan Kaufmann Publishers, March 2006



Tan P.-N., Steinbach M., Kumar V.

Introduction to Data Mining

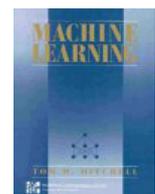
Addison-Wesley, 2006



Mitchell T. M.

Machine Learning

McGraw-Hill, 1997



Witten I. H., Frank E.

Data Mining: Practical Machine Learning Tools and Techniques

2. Auflage. Morgan Kaufmann Publishers, 2005

