DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Skript zur Vorlesung

Knowledge Discovery in Databases

im Sommersemester 2013

Kapitel 1: Einleitung

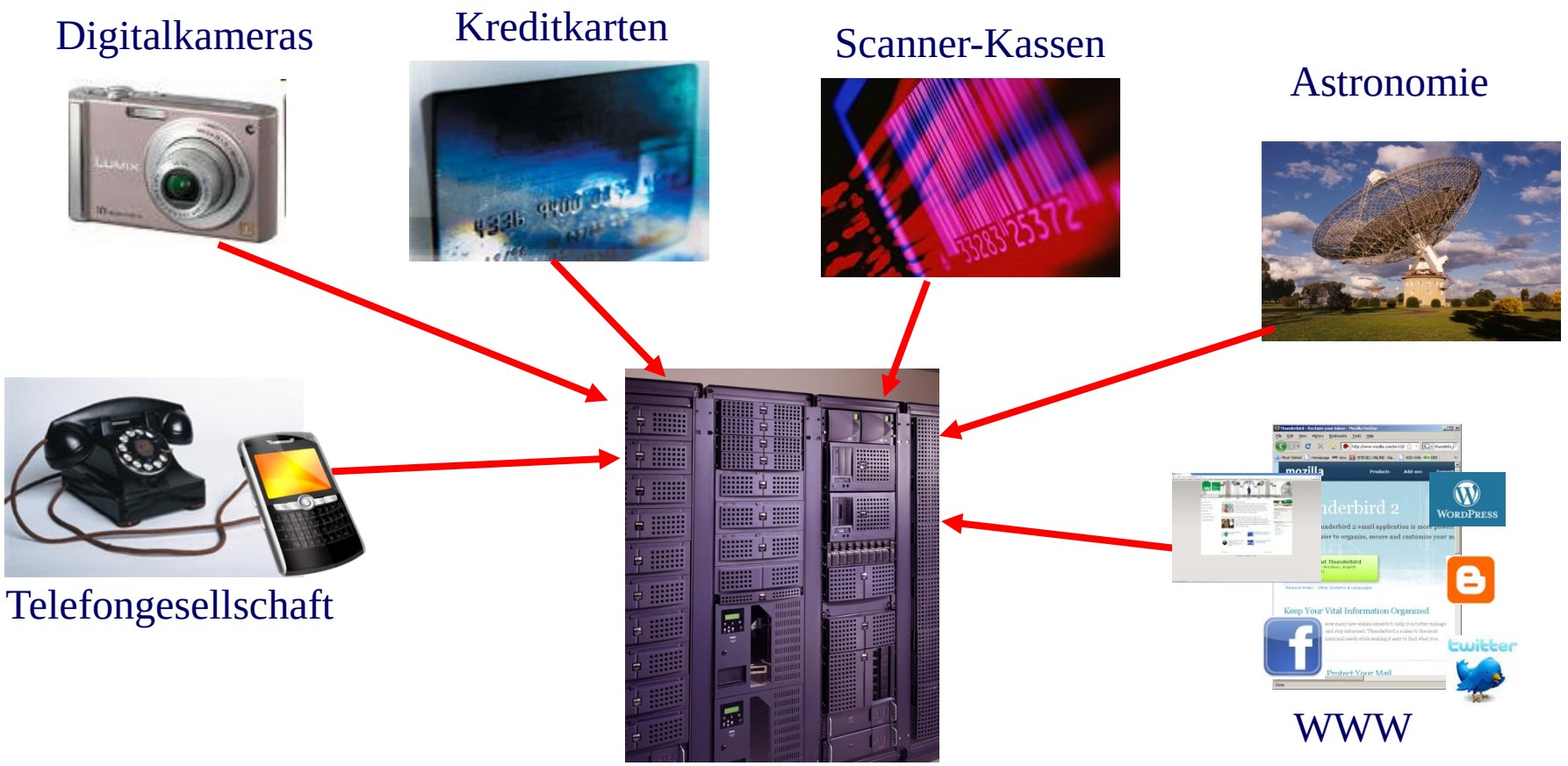
Vorlesung: Dr. Arthur Zimek
Übungen: Erich Schubert

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

Organisatorisches

- **Aktuelles**
  - Vorlesung: Dienstag, 9.00-12.00 Uhr, Raum B U101 (Oettingenstr. 67)
  - Übungen: Freitag, 12-14 Uhr, Raum U151 (Oettingenstr. 67)
    Freitag, 14-16 Uhr, Raum U151 (Oettingenstr. 67)
- Anmeldung für die Klausur auf der Homepage unter
  http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)
- Sprechstunden:
  - Dr. Arthur Zimek: Dienstag, 15:00-16:00, Raum F10x (Oettingenstr. 67)
  - Erich Schubert: Mittwoch, 14:00-15:00, Room F10x (Oettingenstr. 67)
- Klausur: Der Stoff der Klausur wird in der Vorlesung und in den Übungen besprochen.
  (Das Skript ist lediglich eine Lernhilfe.)

Motivation



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

Von den Daten zum Wissen

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

Definition KDD

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

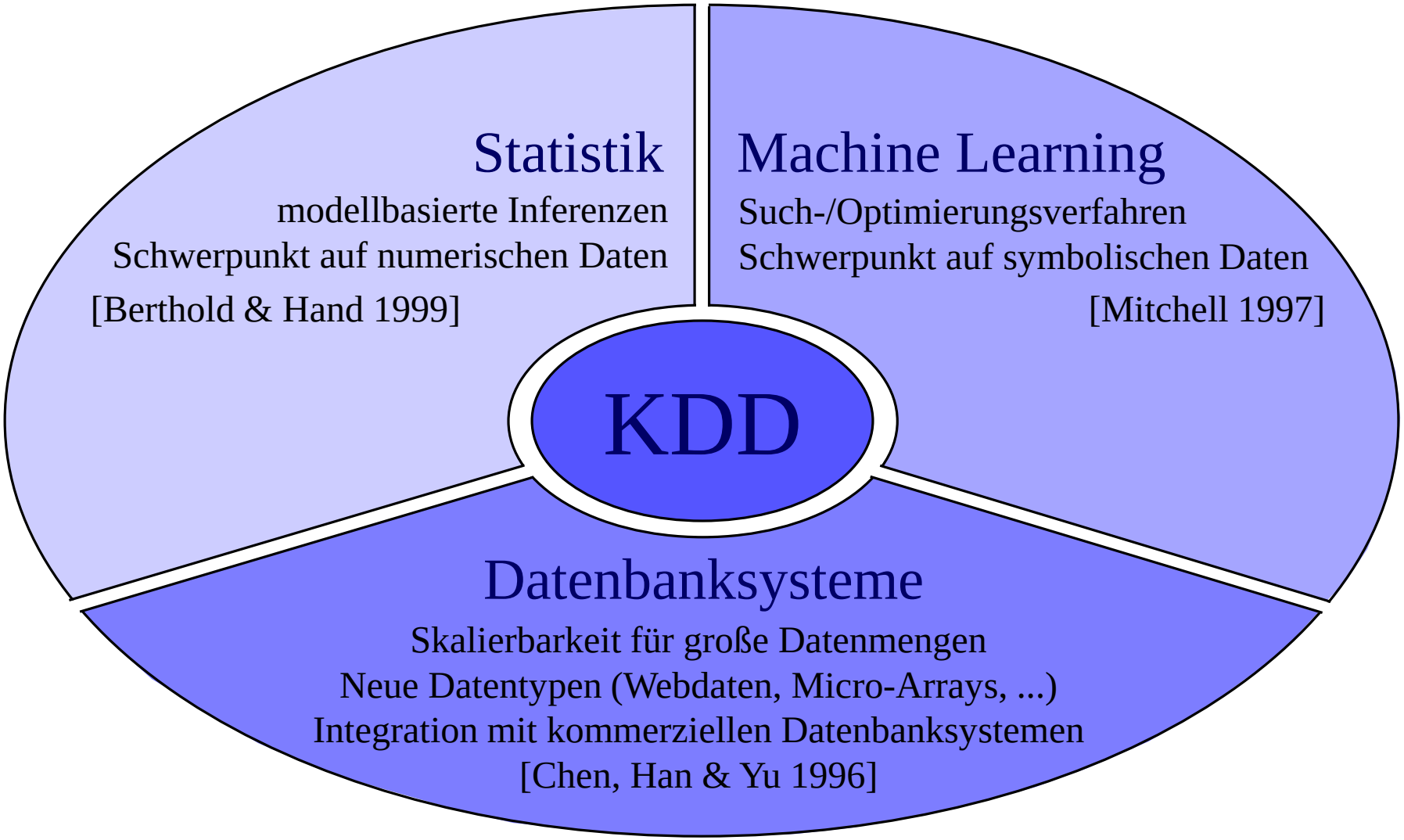
[Fayyad, Piatetsky-Shapiro & Smyth 1996]

- *data:* set of facts (z.B. Einträge in Datenbank)
- *pattern:* Ausdruck in einer Sprache um eine Teilmenge der Daten oder ein Model, das auf diese Teilmenge angewendet werden kann, zu beschreiben
- *process*: Vorgang in mehreren Schritten und Iterationen
- *nontrivial:* involviert komplexere Vorgänge wie Suche oder Folgerungen (Inferenz), nicht einfache Aggregationen wie Durchschnitt
- *valid:* anwendbar auf neue Daten mit einem bestimmten Grad an Zuverlässigkeit
- *novel:* für das System, besser noch auch für den Anwender
- *potentially useful:* vorteilhaft für den Anwender oder die Anwendung
- *ultimately understandable:* wenn nicht sofort, dann nach geeigneter Nachbereitung

understandability ⇔ simplicity?

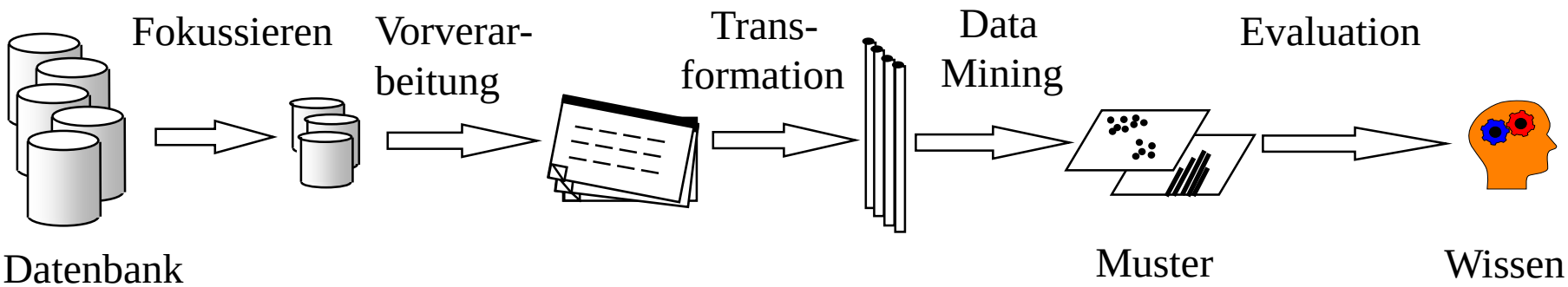
(validity, novelty, usefulness, simplicity) ⇔ „interestingness“

Teilbereiche KDD



Das KDD-Prozessmodell

Prozessmodell nach Fayyad, Piatetsky-Shapiro & Smyth



Fokussieren:
- Beschaffung der Daten
- Verwaltung (File/DB)
- Selektion relevanter Daten

Vorverarbeitung:
- Integration von Daten aus unterschiedlichen Quellen
- Vervollständigung
- Konsistenzprüfung

Transformation
- Diskretisierung numerischer Merkmale
- Ableitung neuer Merkmale
- Selektion relevanter Merkm.

Data Mining
- Generierung der Muster bzw. Modelle

Evaluation
- Bewertung der Interessantheit durch den Benutzer
- Validierung: Statistische Prüfung der Modelle

Data Mining Aufgaben

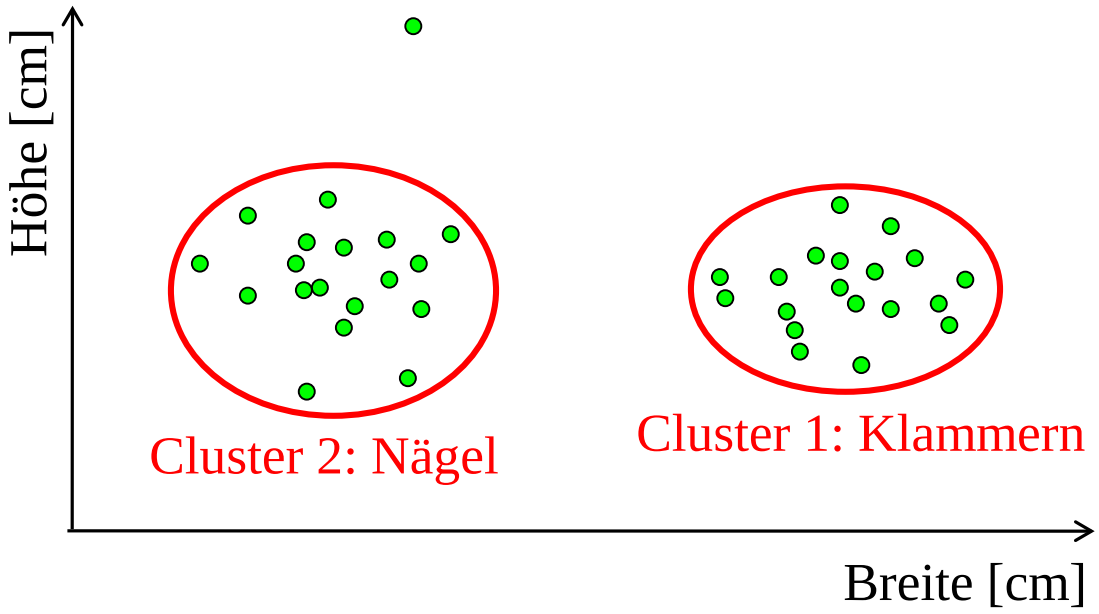
Die wichtigsten Data-Mining-Techniken:

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 (feature vectors). Darüber hinaus gibt es zahlreiche Verfahren, die nicht auf Merkmalsvektoren, sondern z.B. auf Texten, Mengen, Graphen arbeiten.

Clustering

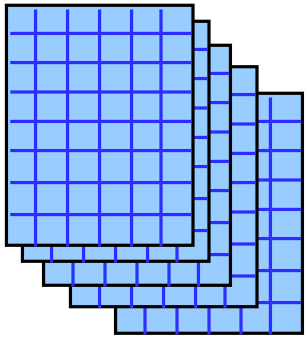


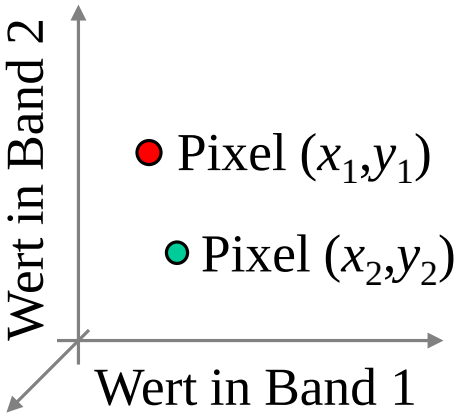
Clustering heißt: Zerlegung einer Menge von Objekten (bzw. Feature-Vektoren) in Teilmengen (Cluster) ähnlicher Objekte

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

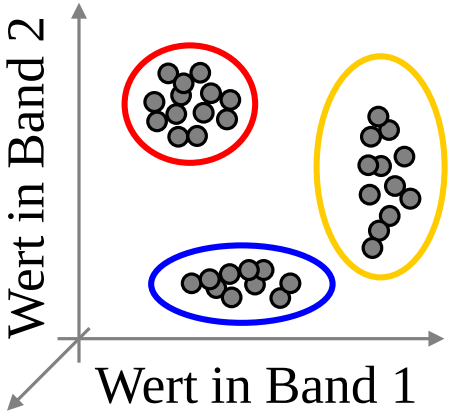
Anwendung: Thematische Karten

Aufnahme der Erdoberfläche
in 5 verschiedenen Spektren



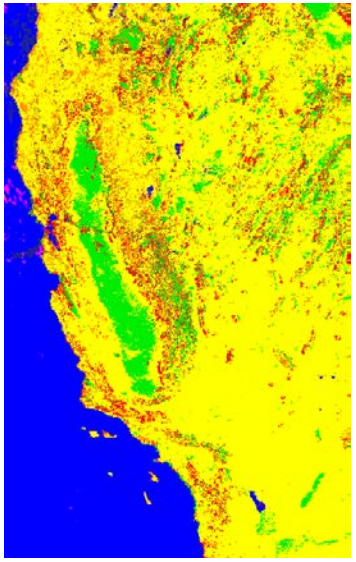


Cluster-Analyse

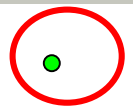


Rücktransformation
in *xy*-Koordinaten

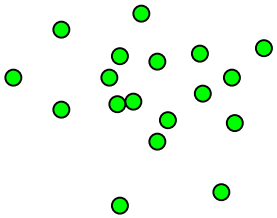
Farbcodierung nach
Cluster-Zugehörigkeit

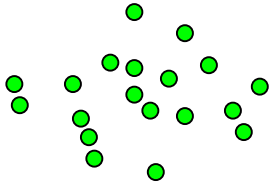


Outlier Detection



Datenfehler?
Betrug?





Outlier Detection bedeutet:
Ermittlung von untypischen Daten

Idee: Outlier könnten hindeuten auf

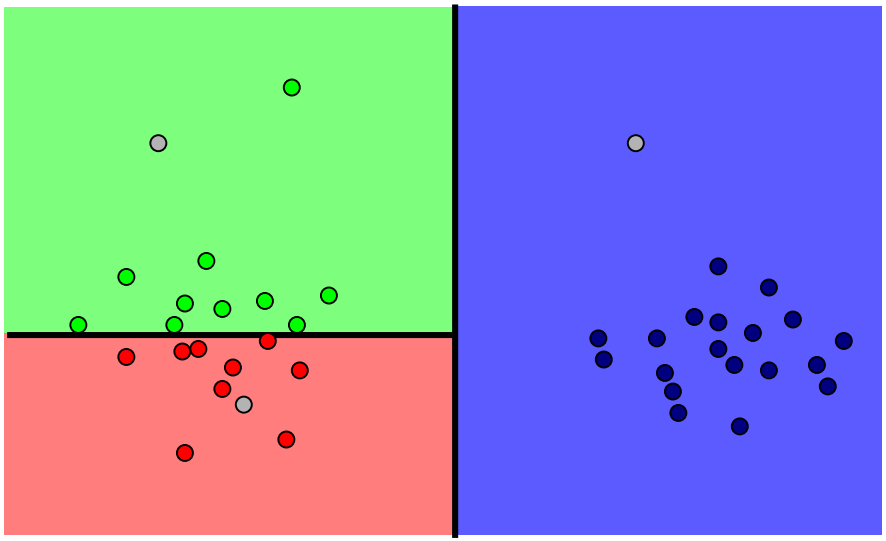
- Missbrauch etwa bei
  - Kreditkarten
  - Telekommunikation
- Datenfehler

Anwendung

- 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	Giovanne Elber	21	13	Sturm	Hohe Tore-pro-Spiel Quote

Klassifikation



- Schrauben
- Nägel
- Klammern

} Trainings-daten

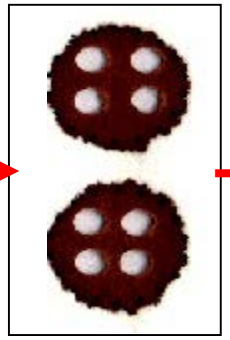
- Neue Objekte

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

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

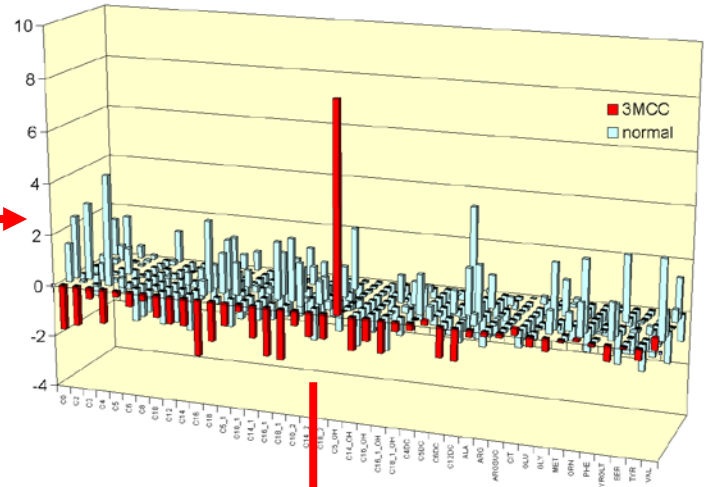
Anwendung: Neugeborenen-Screening

Blutprobe des Neugeborenen



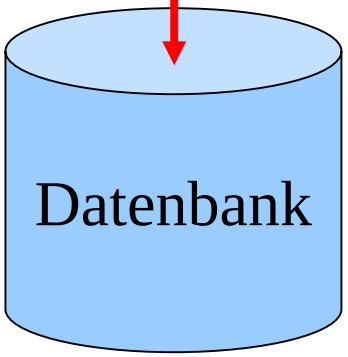
Massenspektrometrie

Metabolitenspektrum

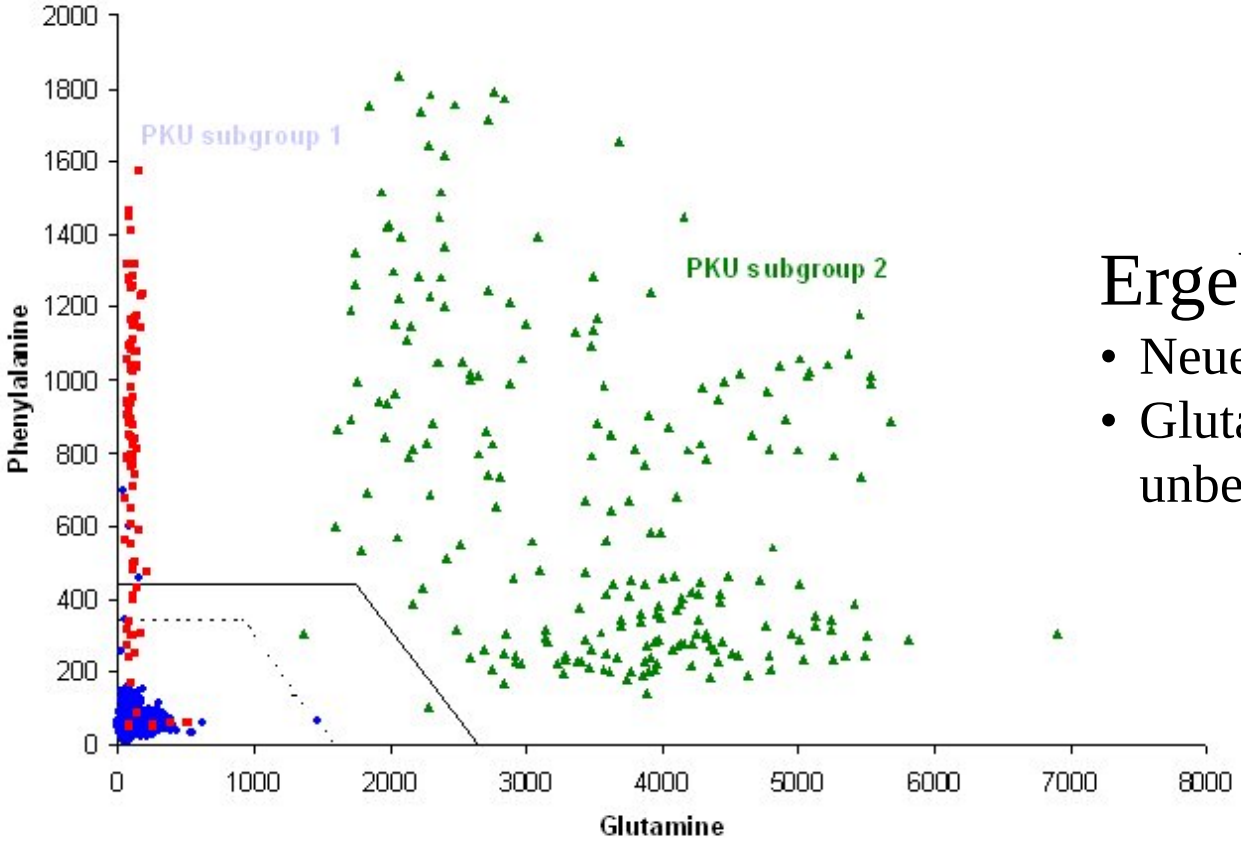


14 analysierte Aminosäuren:

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



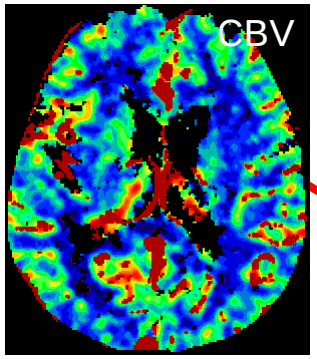
Anwendung: Neugeborenen-Screening

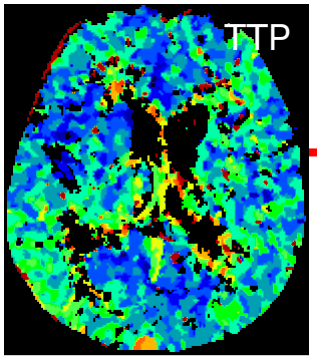


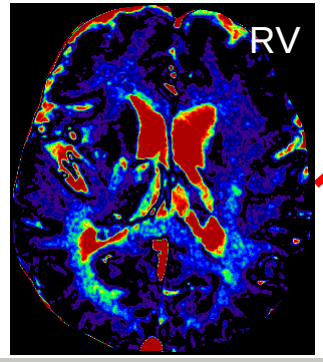
Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

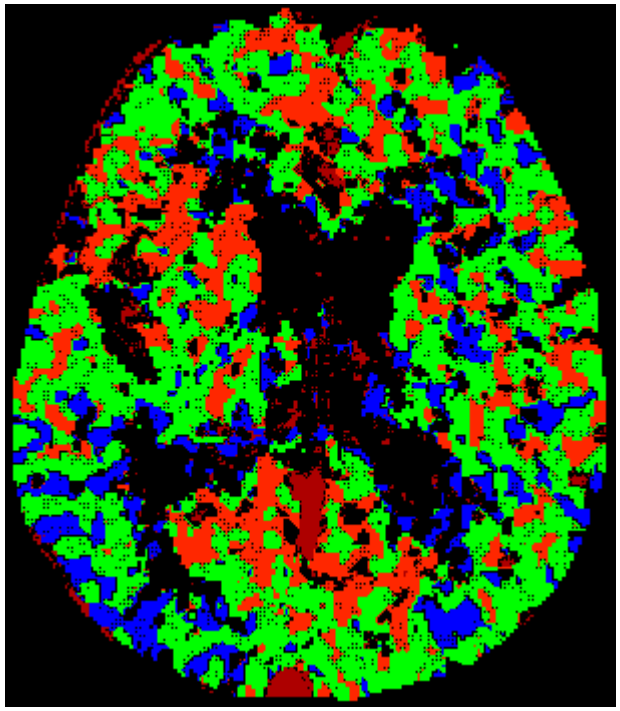
Anwendung: Gewebeklassifikation







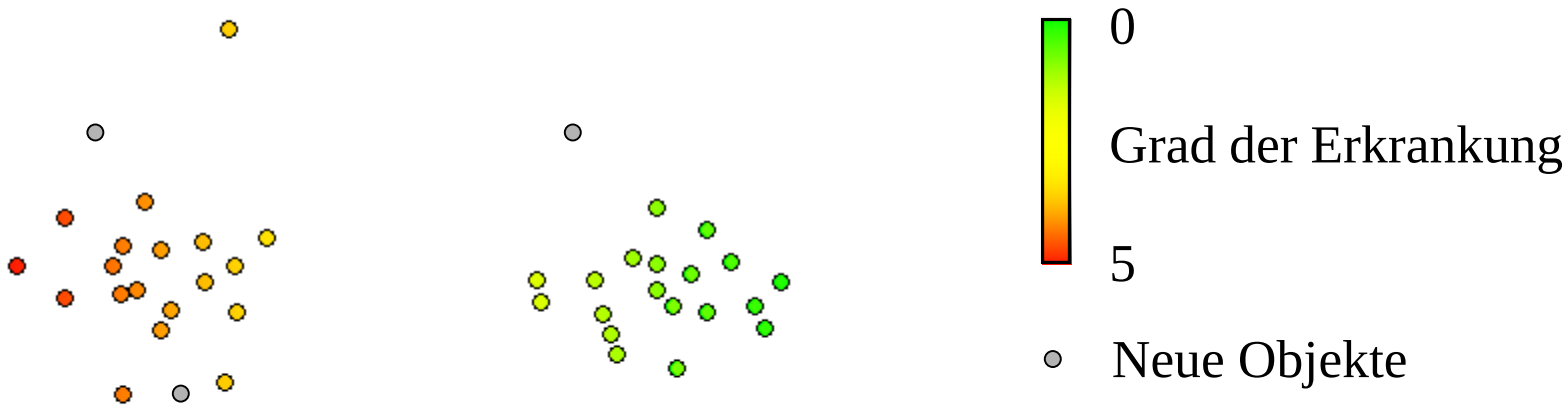
- Schwarz: Ventrikel + Hintergrund
- Blau: Gewebe 1
- Grün: Gewebe 2
- Rot: Gewebe 3
- Dunkelrot: Große Gefäße



	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**

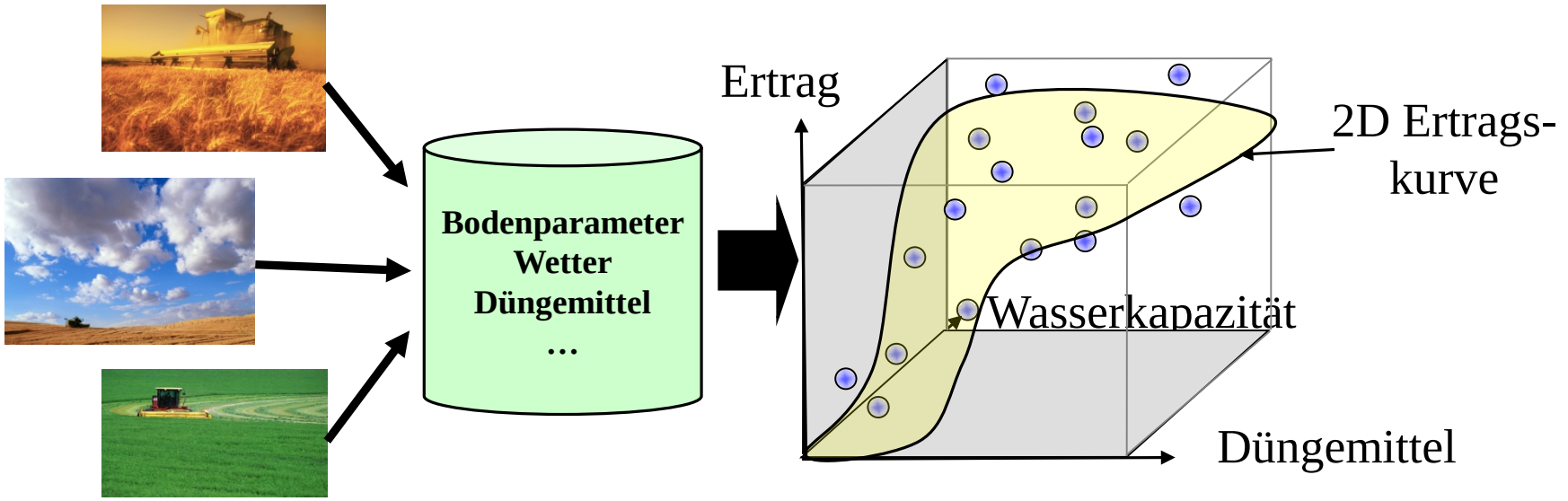
Ergebnis: Klassifikation cerebralen Gewebes anhand funktioneller Parameter mittels dynamic CT möglich.

Regression



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

Anwendung: Precision Farming

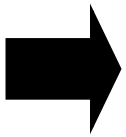


- Erstellen einer Ertragskurve, die von mehreren Parametern wie Bodenbeschaffenheit, Wetter und Düngemittelausbringung abhängt.
- Erst eine geeignete Anpassung der Düngemittelausbringung kann eine ertragsoptimale Nutzung in Abhängigkeit von Umweltfaktoren bewirken.
- Das Thema ist auch wegen der Umweltbelastung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

Assoziationsregeln

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

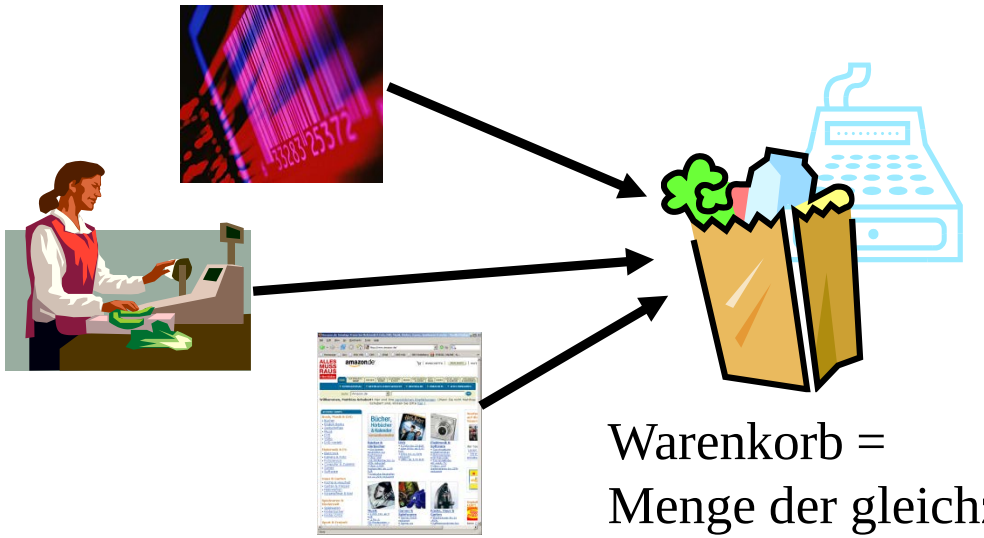
In 5 von 7 (ca. 71 %) der Fälle kommt **b,c,d** zusammen vor.



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

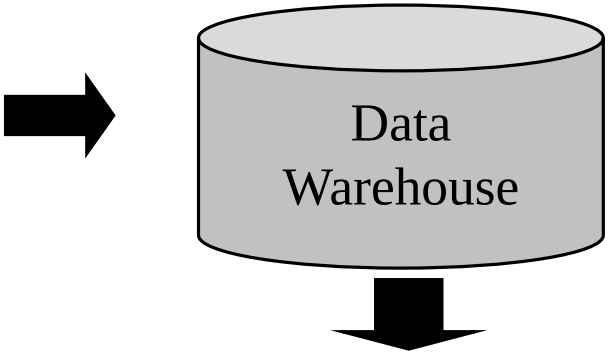
Aufgabe:
Finde alle Regeln in einer Datenbank von diskreten Mengen der folgenden Art:
Wenn *a*, *b*, *c* in der Menge M enthalten sind, dann ist auch *t* mit einer Wahrscheinlichkeit vom >X % in der Menge enthalten.

Anwendung: Warenkorbanalyse



Warenkorb =
Menge der gleichzeitig bezahlten Waren

evtl. Verallgemeinerung:
- Paprika-Chips ➡ Knabbereien
- Anreichern mit Kundendaten



Assoziationsregeln

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 Warenkörben:
  Kunden die „Krieg der Sterne“ I-VI gekauft haben, sind vielleicht auch an „Herr der Ringe“ I-III interessiert.

Überblick über die Vorlesung
(Momentaner Stand der Planung)

1. Einleitung
2. Merkmalsräume
3. Clustering
4. Outlier Detection
5. Klassifikation
6. Regression
7. Evaluation
8. Assoziationsregeln

Resourcen

- Kommerzielle und freie / open source Tools:
  - http://www.kdnuggets.com/software/suites.html
- Kommerzielle Tools werden von vielen großen Firmen angeboten: IBM, Microsoft, Oracle
- Freie/open source Tools:

Weka

R

SciPy + NumPy

ELKI

Orange

Rapid Miner (free, commercial versions)

Was haben Sie gelernt?

- Definition KDD
- KDD Prozess
- Data Mining-Schritt
- Supervised vs. Unsupervised Learning
- Wichtige Data Mining Aufgaben:
  - Clustering
  - Classification
  - Regression
  - Association rules mining
  - Outlier detection

Hausaufgabe/Übungen

- Übungen: Übungsblätter vorbereiten (darüber nachdenken, versuchen zu lösen), werden in den Übungen am Freitag besprochen
- Keine Übungen diese Woche!
- In Vorbereitung: individuelle Studienprojekte (Implementierung und Evaluierung von Methoden aus der Literatur) – Bonus-Punkte für Klausur
- Hausaufgabe: Denken Sie über Probleme aus dem täglichen Leben nach, auf die Sie KDD Methoden anwenden könnten
  - Warum?
  - Welche Art von Mustern wäre interessant?
- Lektürevorschlag:

  Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3): 37-54 (1996)

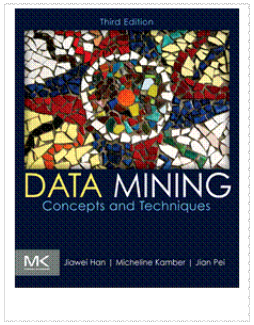
Literatur

Lehrbuch zur Vorlesung (deutsch):

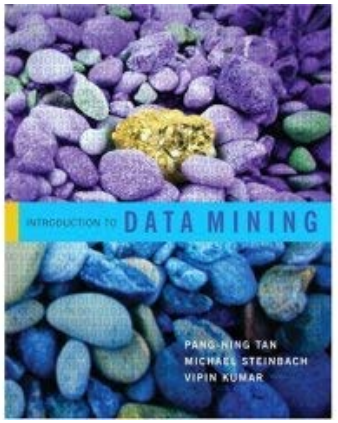
Ester M., Sander J.
Knowledge Discovery in Databases: Techniken und Anwendungen
Springer Verlag, September 2000

Weitere Bücher (englisch):

Han J., Kamber M., Pei J.
Data Mining: Concepts and Techniques
3. Auflage, Morgan Kaufmann, 2011



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., Hall M. A.
Data Mining: Practical Machine Learning Tools and Techniques
3. Auflage, Morgan Kaufmann, 2011