

Skript zur Vorlesung  
**Knowledge Discovery in Databases**  
im Sommersemester 2015

# Kapitel 1: Einleitung

Vorlesung: PD Dr. Arthur Zimek  
Übungen: Dr. Tobias Emrich

Skript © 2015 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, 9.30-12.00 Uhr, M 010 (Hauptgebäude)
- Übungen (Tutor: Janina Bleicher):
  - Mittwoch, 14-16 Uhr, Raum U151 (Oettingenstr. 67) (ab 22.04.)
  - Donnerstag, 14-16 Uhr, Raum U151 (Oettingenstr. 67) (ab 23.04.)

- Anmeldung zur Vorlesung 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))

- Sprechstunden:

- PD Dr. Arthur Zimek: Dienstag, 15:00-16:00, Raum F109 (Oettingenstr. 67)
- Dr. Tobias Emrich: nach Vereinbarung, Raum F103 (Oettingenstr. 67)

- Klausur: Der Stoff der Klausur wird in der Vorlesung und in den Übungen besprochen.  
(Das Skript ist lediglich eine Lernhilfe.)

Digitalkameras



Kreditkarten



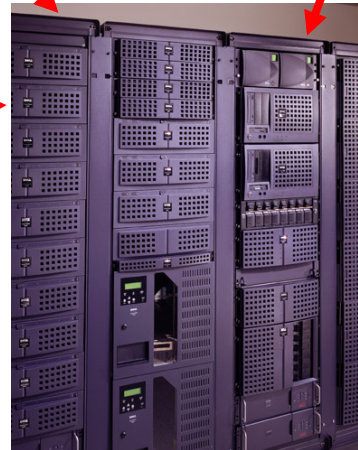
Scanner-Kassen



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

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

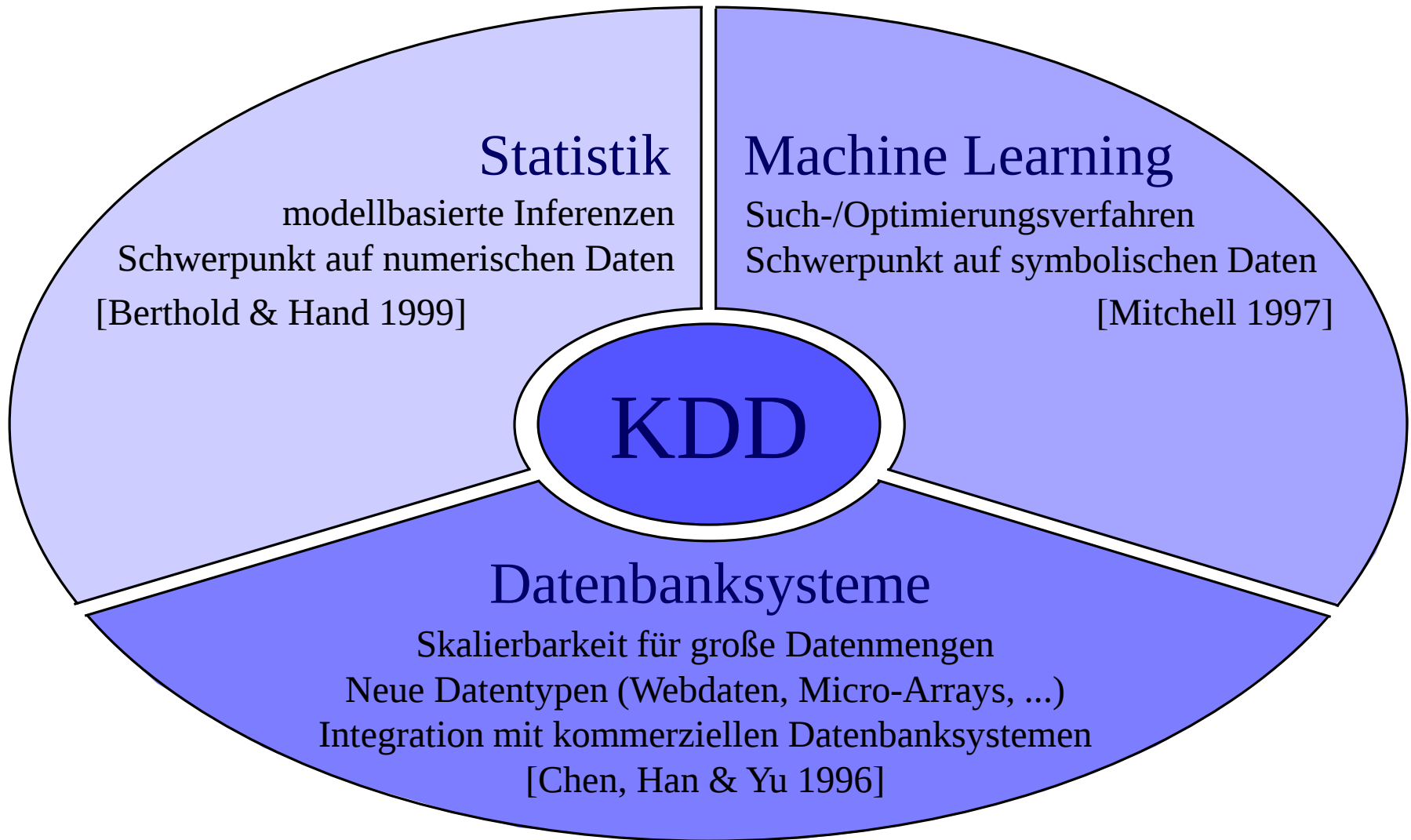
[Fayyad, Piatetsky-Shapiro & Smyth 1996]

- *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

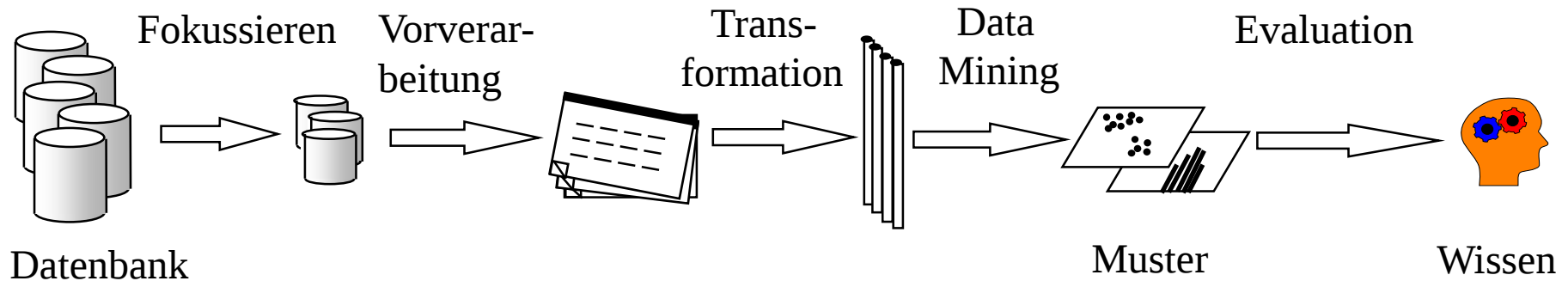
understandability  $\Leftrightarrow$  simplicity?

(validity, novelty, usefulness, simplicity)  $\Leftrightarrow$  „interestingness“

# Teilbereiche KDD



## 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

Die wichtigsten Data-Mining-Techniken:

Supervised: z.B. Klassifikation, Regression, Outlier Detection

Ein Ergebnis-Merkmal soll aufgrund von Vorwissen gelernt/geschätzt werden.

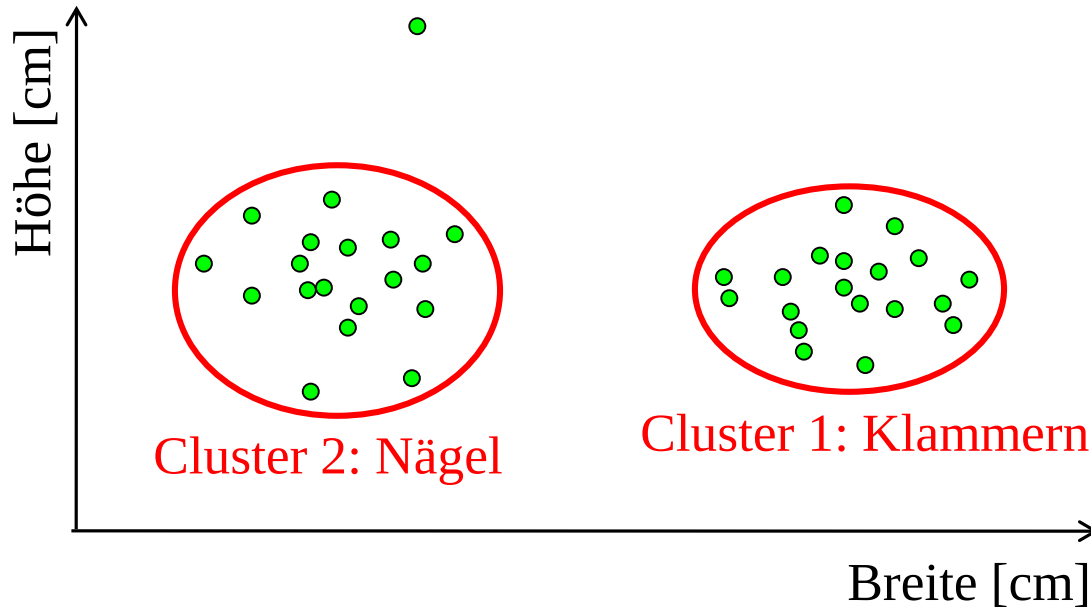
Das Vorwissen steht typischerweise als Trainingsdaten bereit.

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.

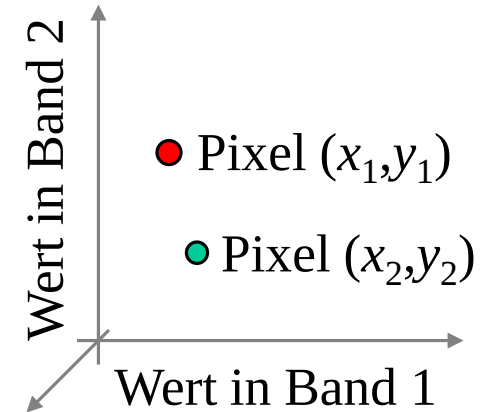
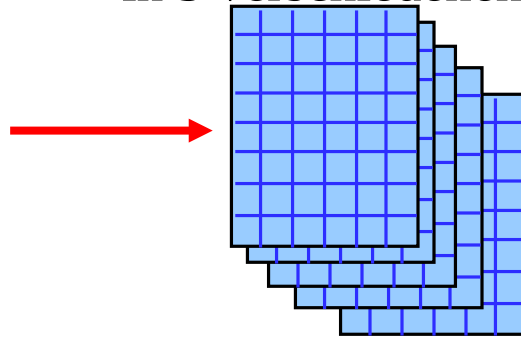


Clustering heißt: Zerlegung einer Menge von Objekten (bzw. Feature-Vektoren) in Teilmengen (Cluster) ähnlicher Objekte

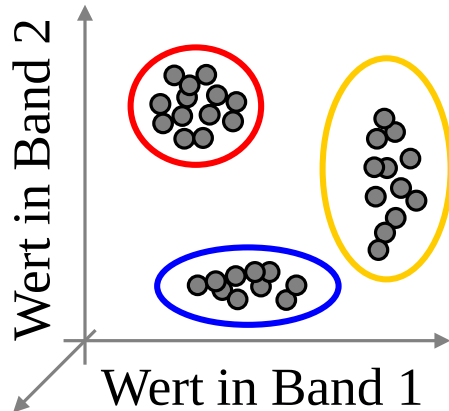
Idee: Die verschiedenen Cluster repräsentieren meist unterschiedliche Klassen von Objekten; bei unbek. Anzahl und Bedeutung der Klassen



Aufnahme der Erdoberfläche  
in 5 verschiedenen Spektren

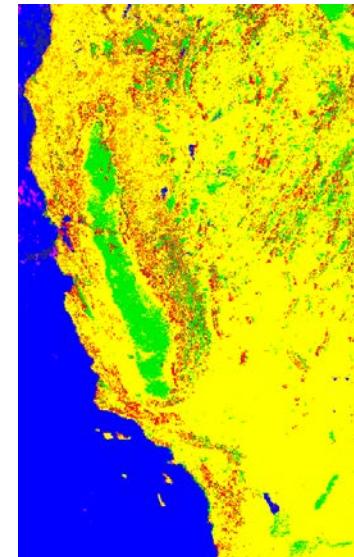


Cluster-Analyse

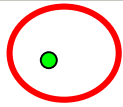


Rücktransformation  
in xy-Koordinaten

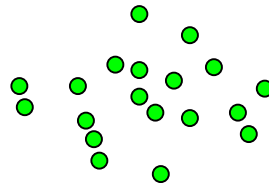
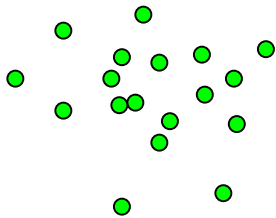
Farbcodierung nach  
Cluster-Zugehörigkeit



# Outlier Detection



Datenfehler?  
Betrug?



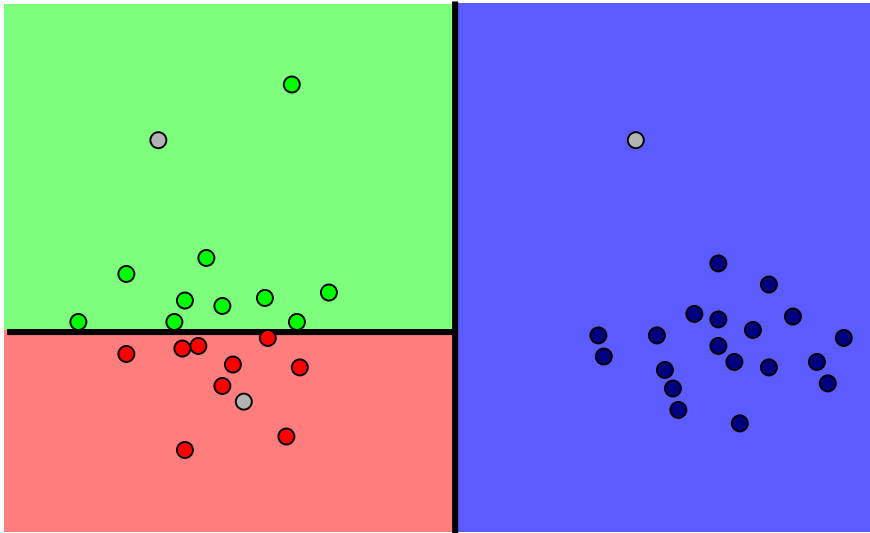
Outlier Detection bedeutet:  
Ermittlung von **untypischen** Daten

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           |



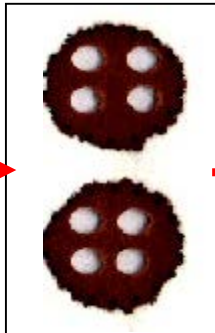
- Schrauben
 ● Nägel
 ● Klammern
 } Trainings-  
daten
- Neue Objekte

Aufgabe:

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

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

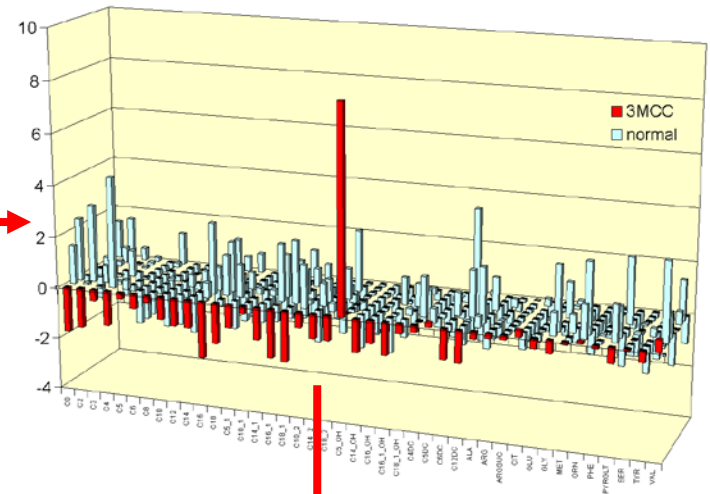
Blutprobe des  
Neugeborenen



Massenspektrometrie



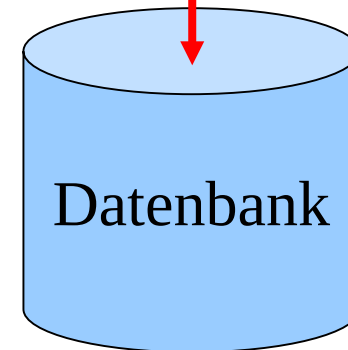
Metabolitenspektrum

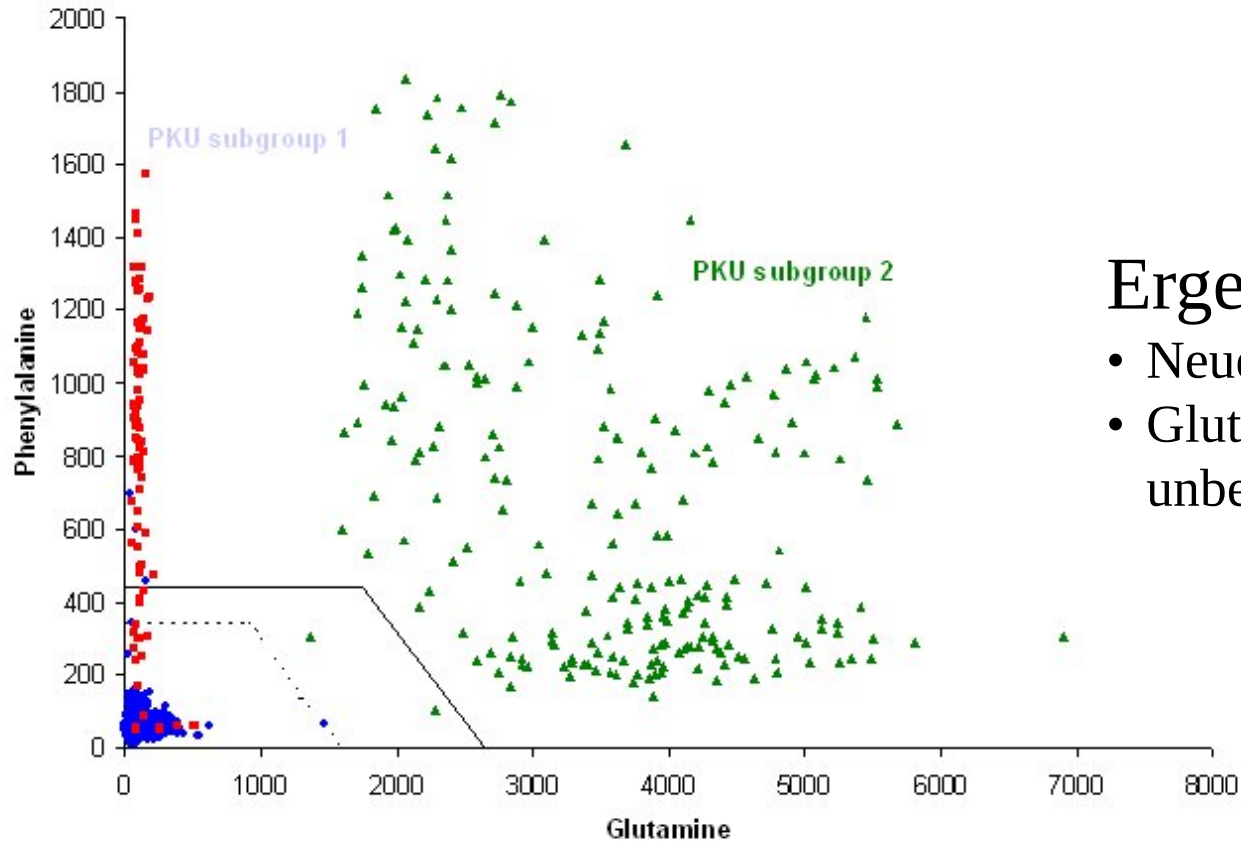


**14 analysierte Aminosäuren:**

alanine  
arginine  
argininosuccinate  
citrulline  
glutamate  
glycine  
methionine

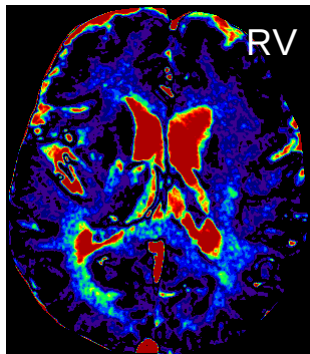
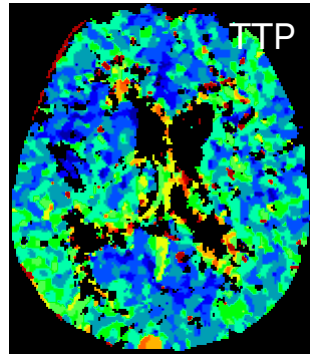
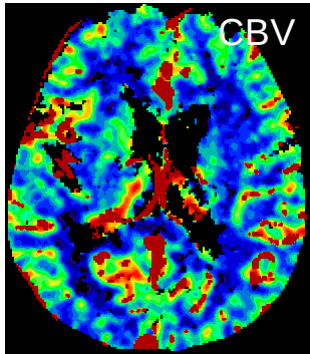
phenylalanine  
pyroglutamate  
serine  
tyrosine  
valine  
leucine+isoleucine  
ornitine



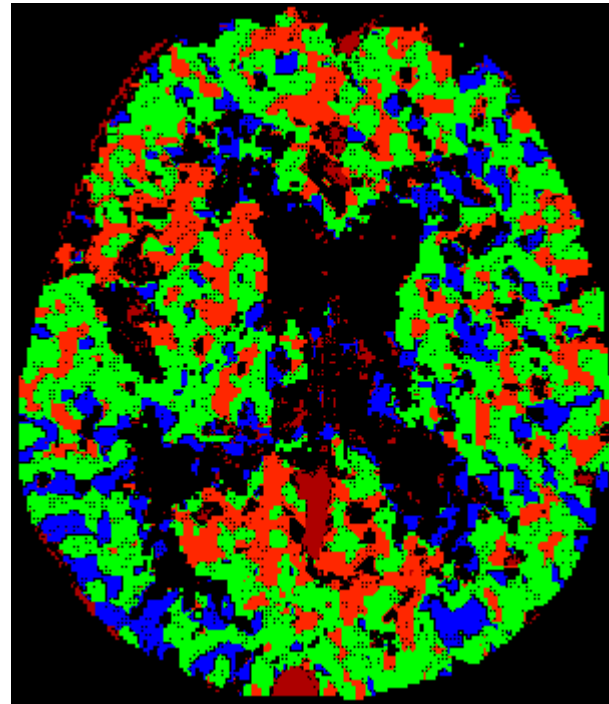


## Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

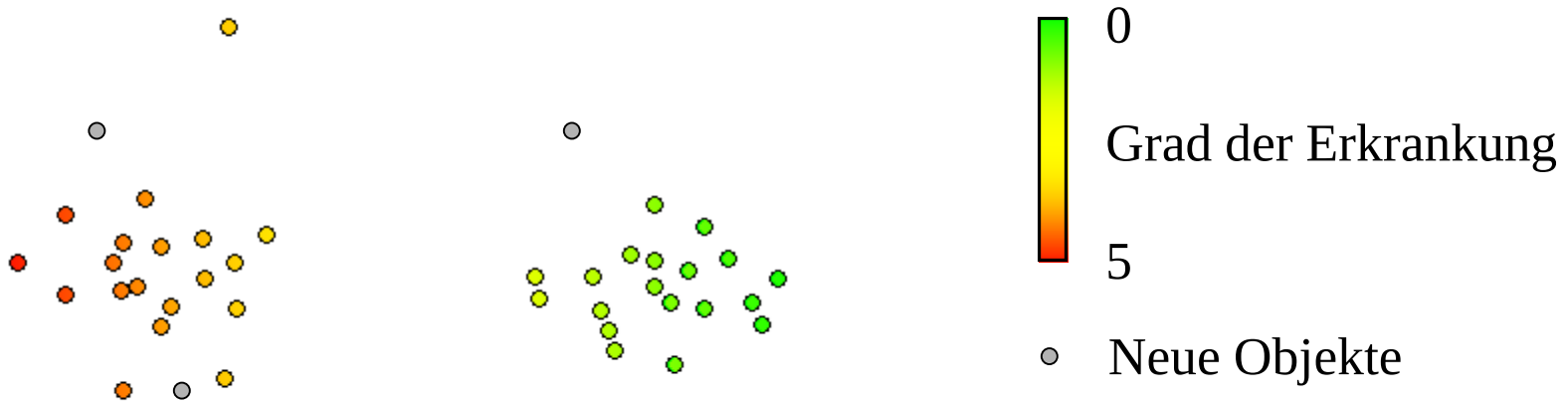


- Schwarz: Ventrikel + Hintergrund
- Blau: Gewebe 1
- Grün: Gewebe 2
- Rot: Gewebe 3
- Dunkelrot: Große Gefäße



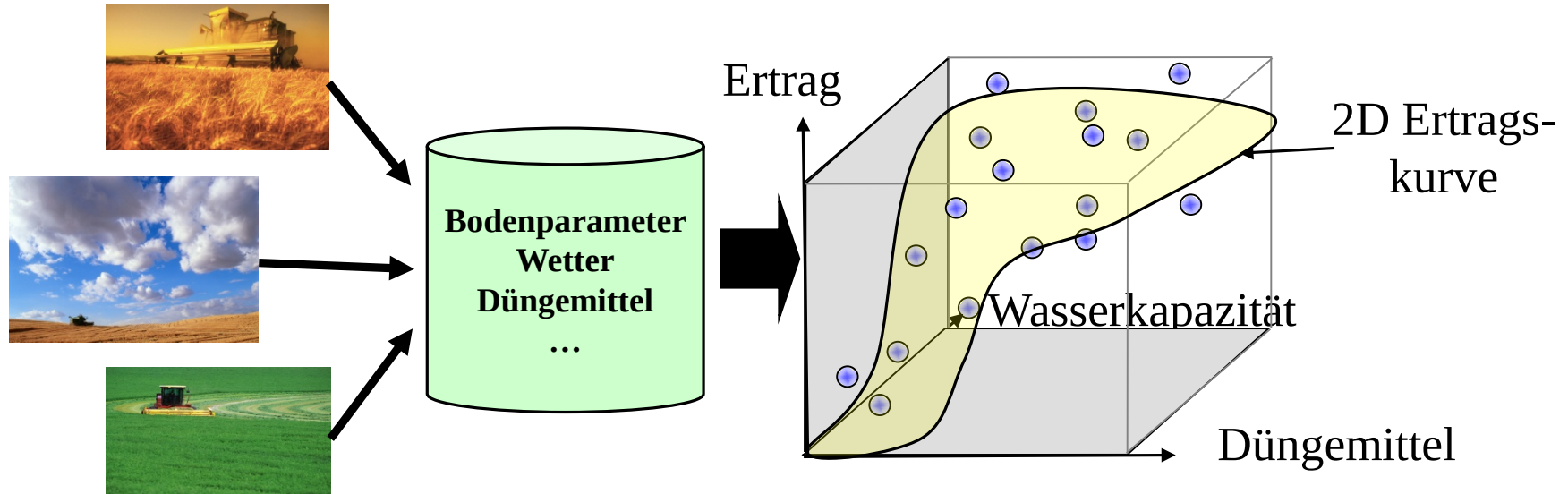
|                      | Blau | Grün | Rot  |
|----------------------|------|------|------|
| TTP<br>(s)           | 20.5 | 18.5 | 16.5 |
| CBV<br>(ml/100g)     | 3.0  | 3.1  | 3.6  |
| CBF<br>(ml/100g/min) | 18   | 21   | 28   |
| RV                   | 30   | 23   | 21   |

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

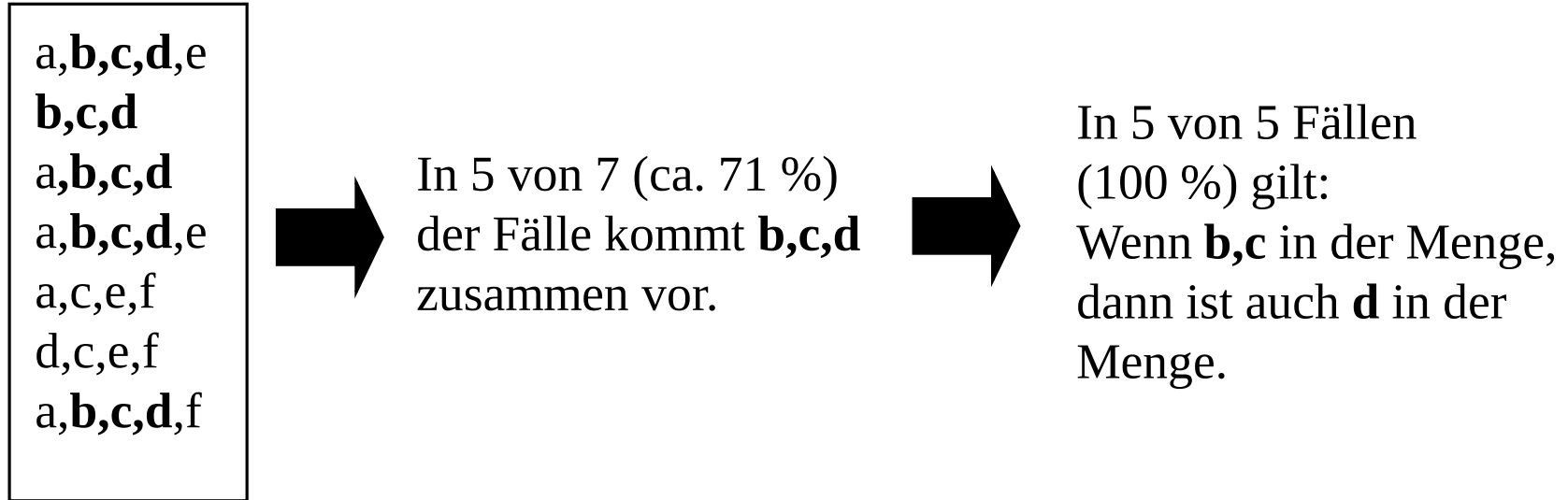


Aufgabe:

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



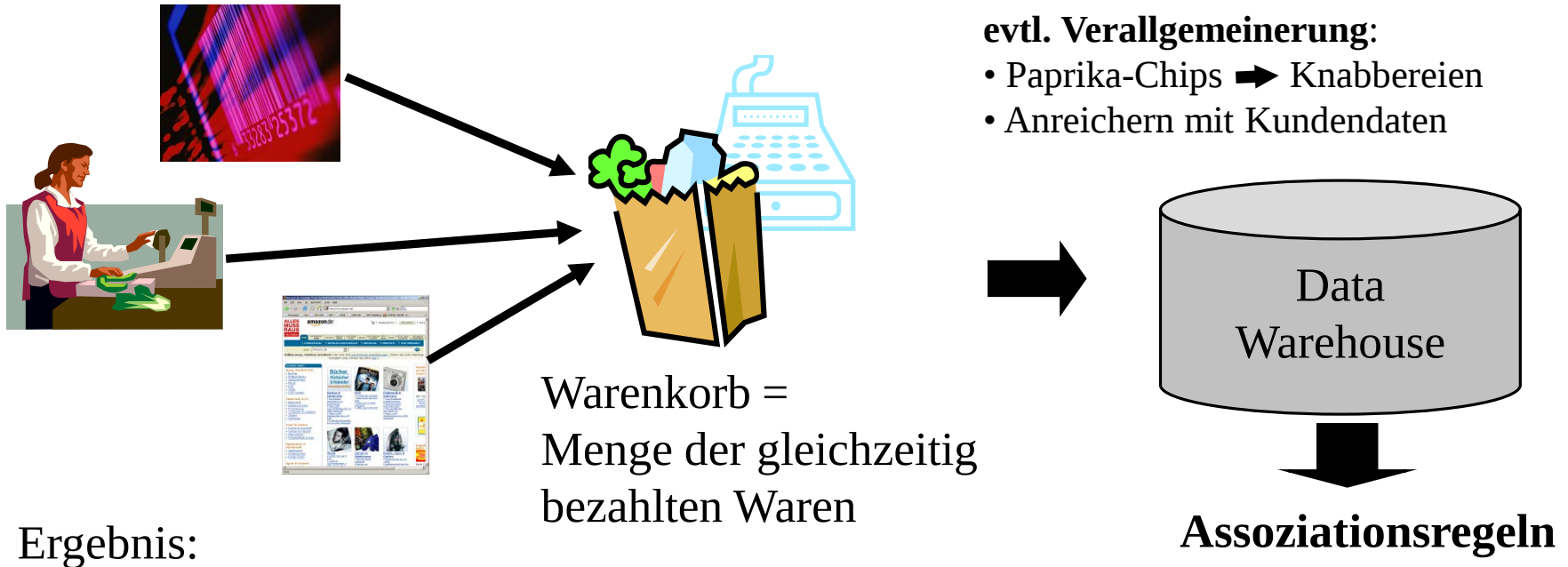
- Erstellen einer Ertragskurve, die von mehreren Parametern wie Bodenbeschaffenheit, Wetter und Düngemittelausbringung abhängt.
- Erst eine geeignete Anpassung der Düngemittelausbringung kann eine ertragsoptimale Nutzung in Abhängigkeit von Umweltfaktoren bewirken.
- Das Thema ist auch wegen der Umweltbelastung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



- 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.

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

- 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:



SciPy + NumPy



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

- Übungen:
  - Übungsblätter vorbereiten/lösen
  - Lösungen werden in den Übungen besprochen
- **Erste Übung nächste Woche!**
- Individuelle Studienprojekte (6 ECTS Praktikum)
  - z.B. Implementierung (in ELKI) und Evaluierung von Methoden aus der Literatur
- 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)

## 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

