

Skript zur Vorlesung:

Einführung in die Informatik: Systeme und Anwendungen

Sommersemester 2014

Kapitel 4: Data Mining

Vorlesung: Prof. Dr. Christian Böhm

Übungen: Sebastian Goebel

Skript © 2010 Christian Böhm, Peer Kröger, Arthur Zimek

<http://www.dbs.ifi.lmu.de/cms/>

[Einführung in die Informatik Systeme und Anwendungen](http://www.dbs.ifi.lmu.de/cms/)



4.1 Einleitung

4.2 Clustering

4.3 Klassifikation

Motivation



Kreditkarten



Scanner-Kassen



Telefongesellschaft



Datenbanken



Astronomie

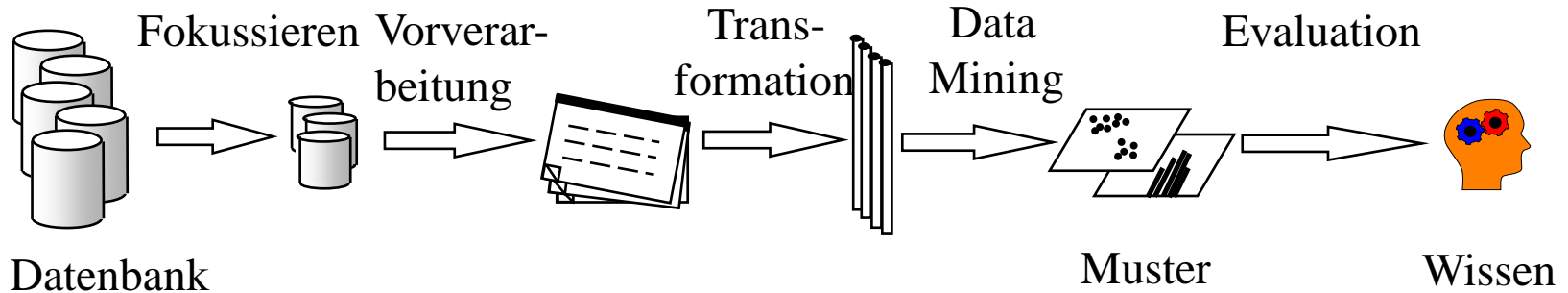
- Riesige Datenmengen werden in Datenbanken gesammelt
- Analysen können nicht mehr manuell durchgeführt werden

Definition KDD

- *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.

Der KDD-Prozess (Modell)

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

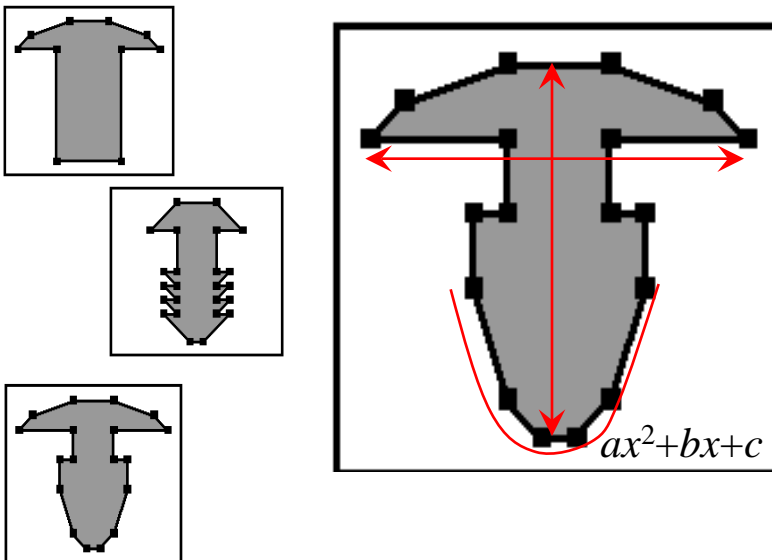
Evaluation

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

Objekt-Merkmale (Feature)

- Oft sind die betrachteten Objekte komplex
- Eine Aufgabe des KDD-Experten ist dann, geeignete Merkmale (*Features*) zu definieren bzw. auszuwählen, die für die Unterscheidung (Klassifikation, Ähnlichkeit) der Objekte relevant sind.

Beispiel: CAD-Zeichnungen:

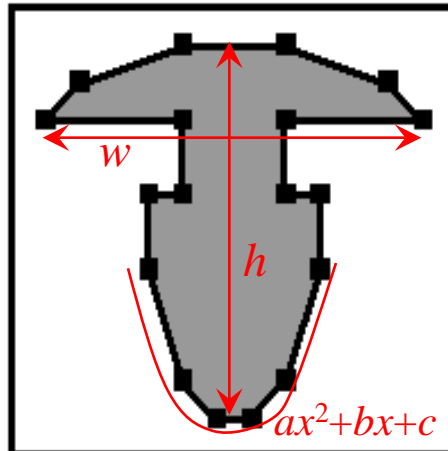


Mögliche Merkmale:

- Höhe h
- Breite w
- Krümmungs-Parameter (a, b, c)

Feature-Vektoren

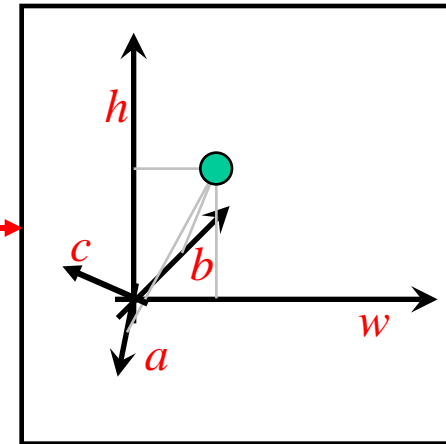
Objekt-Raum



(h, w, a, b, c)



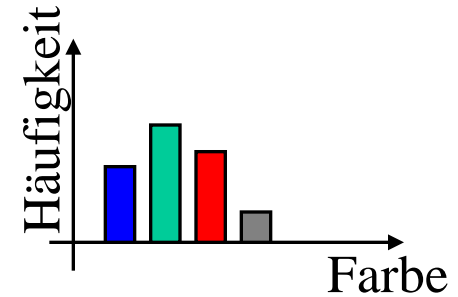
Merkmals-Raum



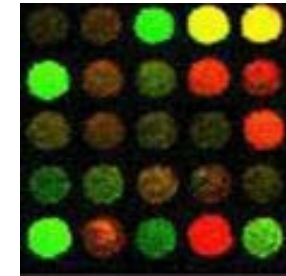
- Im Kontext von statistischen Betrachtungen werden die Merkmale häufig auch als *Variablen* bezeichnet
- Die ausgewählten Merkmale werden zu Merkmals-Vektoren (*Feature Vector*) zusammengefasst
- Der Merkmalsraum ist häufig hochdimensional (im Beispiel 5-dim.)

Feature-Vektoren (weitere Beispiele)

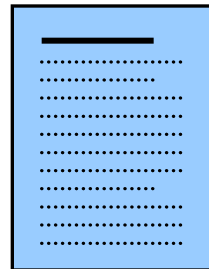
Bilddatenbanken:
Farbhistogramme



Gen-Datenbanken:
Expressionslevel



Text-Datenbanken:
Begriffs-Häufigkeiten



Data	25
Mining	15
Feature	12
Object	7
...	

Der Feature-Ansatz ermöglicht einheitliche Behandlung von Objekten verschiedenster Anwendungsklassen

Feature: verschiedene Kategorien

Nominal (kategorisch)

Charakteristik:

Nur feststellbar, ob der Wert gleich oder verschieden ist. Keine Richtung (besser, schlechter) und kein Abstand.

Merkmale mit nur zwei Werten nennt man *dichotom*.

Beispiele:

Geschlecht (dichotom)
Augenfarbe
Gesund/krank (dichotom)

Ordinal

Charakteristik:

Es existiert eine Ordnungsrelation (besser/schlechter) zwischen den Kategorien, aber kein einheitlicher Abstand.

Beispiele:

Schulnote (metrisch?)
Gütekategorie
Altersklasse

Metrisch

Charakteristik:

Sowohl Differenzen als auch Verhältnisse zwischen den Werten sind aussagekräftig. Die Werte können diskret oder stetig sein.

Beispiele:

Gewicht (stetig)
Verkaufszahl (diskret)
Alter (stetig oder diskret)

Ähnlichkeit von Objekten

- Spezifiziere Anfrage-Objekt $q \in DB$ und...
 - ... suche ähnliche Objekte – Range-Query (Radius ε)

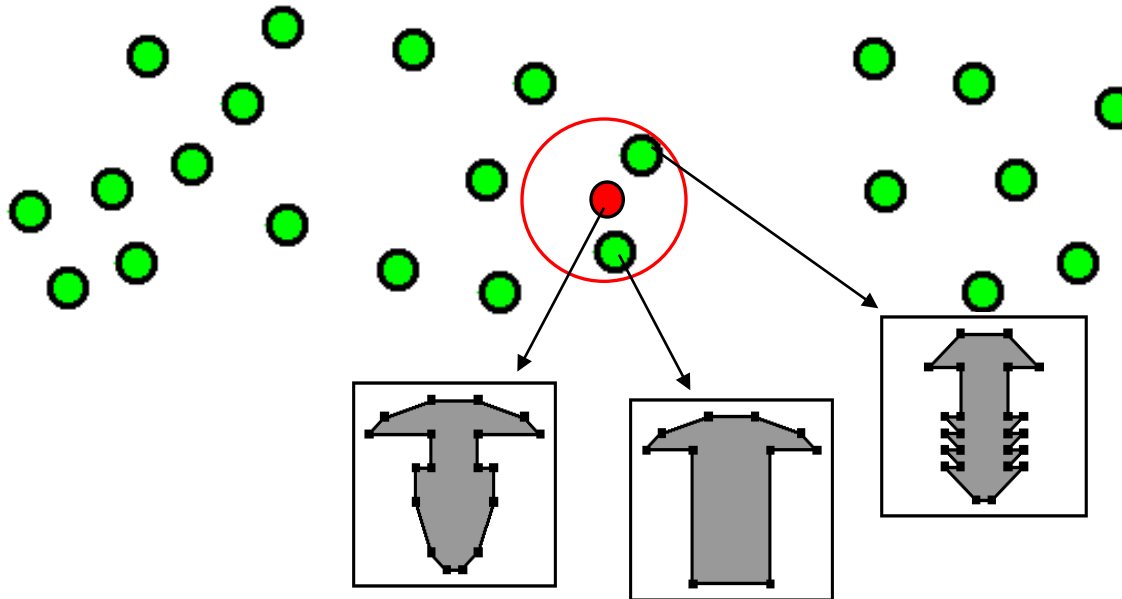
$$RQ(q, \varepsilon) = \{ o \in DB \mid \delta(q, o) \leq \varepsilon \}$$

alternative Schreibweise
für Mengendifferenz:
 $A \setminus B = A - B$

- ... suche die k ähnlichsten Objekte – Nearest Neighbor Query

$NN(q, k) \subseteq DB$ mit k Objekten, sodass

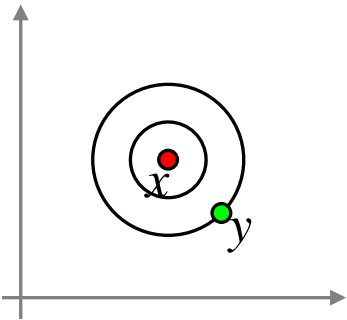
$$\forall o \in NN(q, k), p \in DB \setminus NN(q, k) : \delta(q, o) \leq \delta(q, p)$$



Ähnlichkeitsmaße im Feature-Raum

Euklidische Norm (L_2):

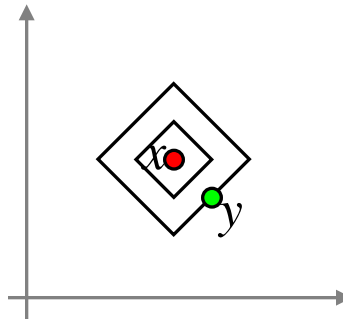
$$\delta_1(x,y) = ((x_1-y_1)^2+(x_2-y_2)^2+\dots)^{1/2}$$



Abstand in Euklidischen Raum
(natürliche Distanz)

Manhattan-Norm (L_1):

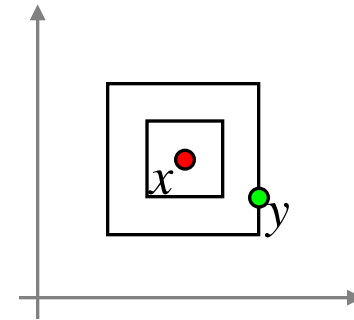
$$\delta_2(x,y) = |x_1-y_1|+|x_2-y_2|+\dots$$



Die Unähnlichkeiten
der einzelnen Merkmale
werden direkt addiert

Maximums-Norm (L_∞):

$$\delta_\infty(x,y) = \max\{|x_1-y_1|, |x_2-y_2|, \dots\}$$



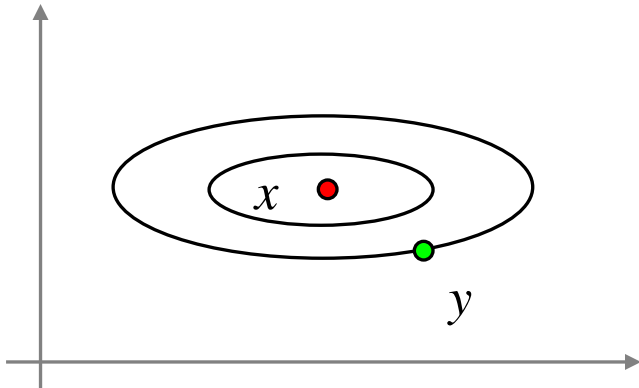
Die Unähnlichkeit des
am wenigsten ähnlichen
Merkmals zählt

Verallgemeinerung L_p -Abstandsmaß:

$$\delta_p(x,y) = (|x_1-y_1|^p + |x_2-y_2|^p + \dots)^{1/p}$$

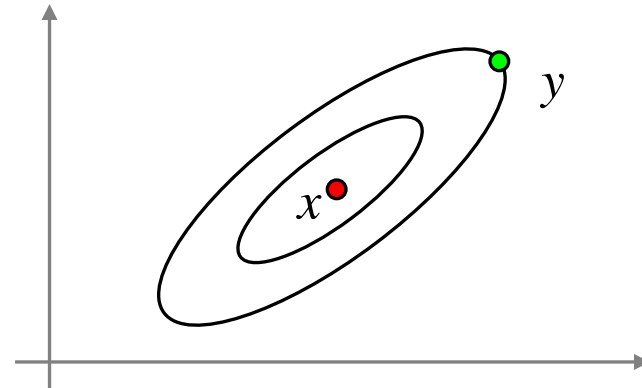
Gewichtete Ähnlichkeitsmaße

- Viele Varianten gewichten verschiedene Merkmale unterschiedlich stark.



Gewichtete Euklidische Distanz

$$\delta_{p,w}(x, y) = \sqrt[p]{\sum_{i=1}^d w_i \cdot |x_i - y_i|^p}$$




Mahalanobis Distanz

$$\delta_{\Sigma}(x, y) = \sqrt{(x - y)^T \cdot \Sigma^{-1} \cdot (x - y)}$$

Σ = Kovarianz-Matrix

Kategorien von Data Mining

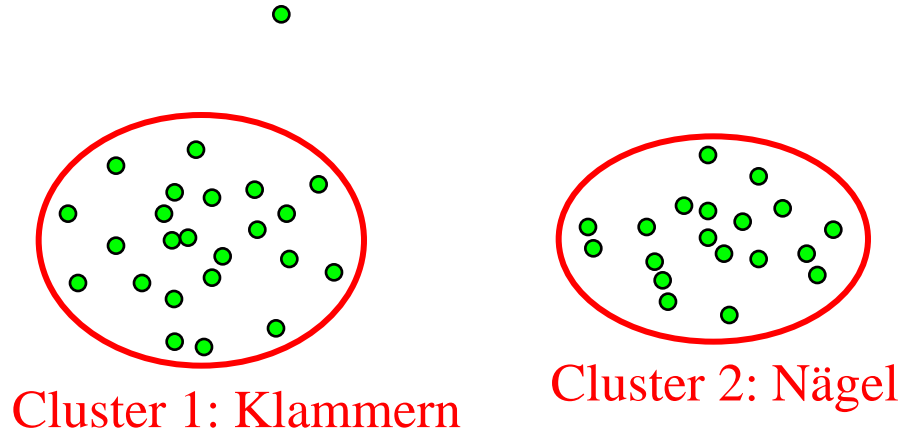
- Wichtigste Data-Mining-Verfahren auf **Merkmals-Vektoren**:
 - Clustering
 - Outlier Detection
 - Klassifikation
 - Regression



normalerweise unsupervised

normalerweise supervised
- Supervised: In Trainingsphase wird eine Funktion gelernt, die in der Testphase angewandt wird.
- Unsupervised: Es gibt keine Trainingsphase. Die Methode findet Muster, die einem bestimmten Modell entsprechen.
- Darüber hinaus gibt es zahlreiche Verfahren, die nicht auf Merkmalsvektoren, sondern direkt auf **Texten, Mengen, Graphen** usw. arbeiten.

Clustering



Ein Grundmodell des Clustering ist:

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

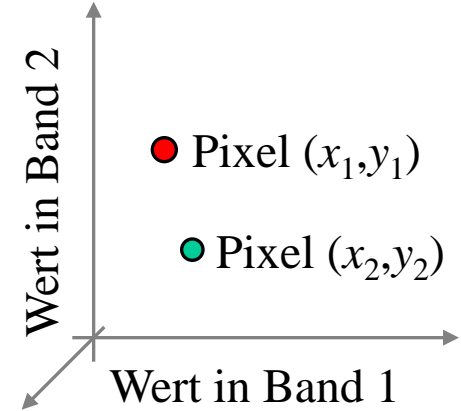
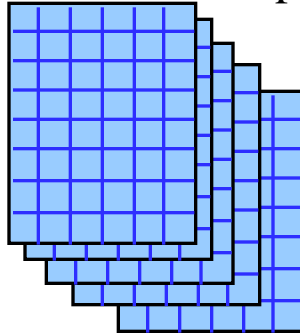
- die Ähnlichkeit der Objekte innerhalb eines Clusters maximiert
- die Ähnlichkeit der Objekte verschiedener Cluster minimiert wird

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

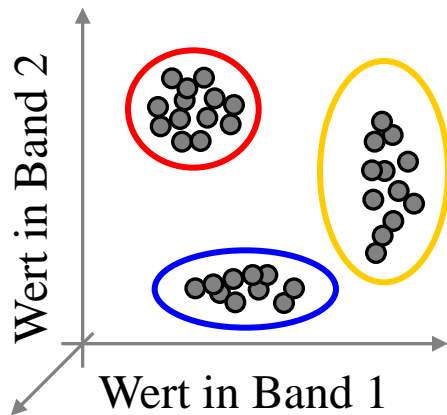
Anwendung: Thematische Karten



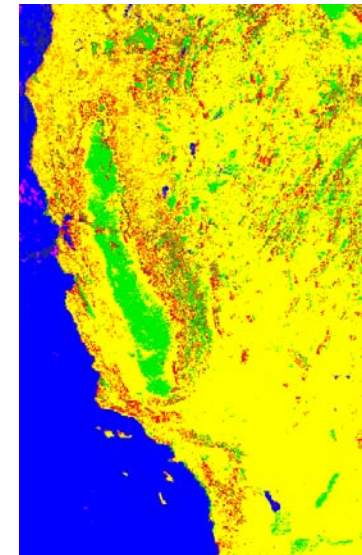
Aufnahme der Erdoberfläche in 5 verschiedenen Spektren



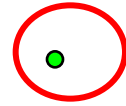
Cluster-Analyse



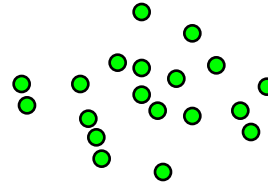
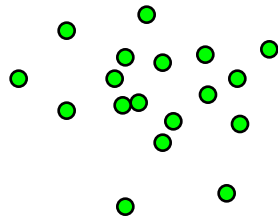
Rücktransformation in xy-Koordinaten
Farbcodierung nach Cluster-Zugehörigkeit



Outlier Detection



Datenfehler?
Betrug?



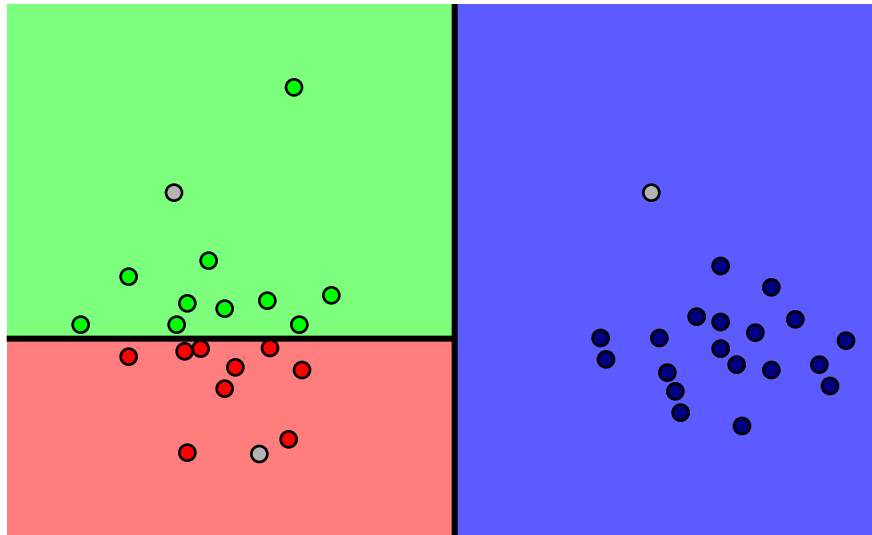
Outlier Detection bedeutet:

Ermittlung von **untypischen** Daten

Anwendungen:

- Entdeckung von Missbrauch etwa bei
 - Kreditkarten
 - Telekommunikation
- Datenbereinigung (Messfehler)

Klassifikation



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

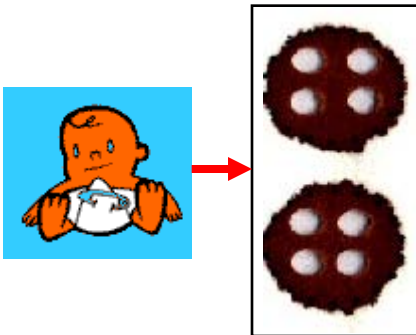
Aufgabe:

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

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

Anwendung: Neugeborenen-Screening

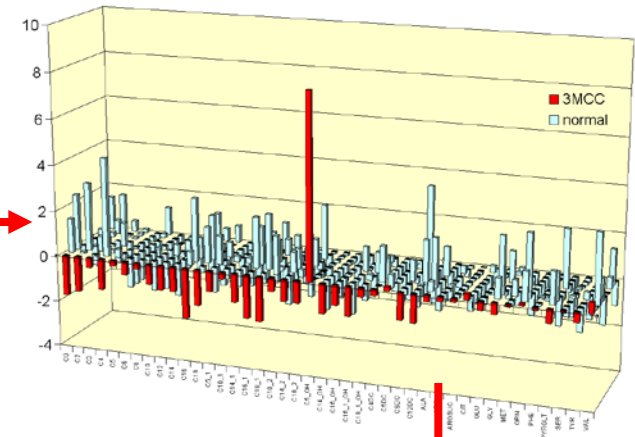
Blutprobe des
Neugeborenen



Massenspektrometrie



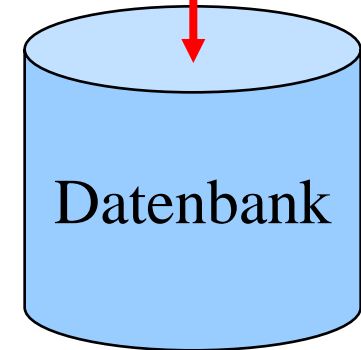
Metabolitenspektrum



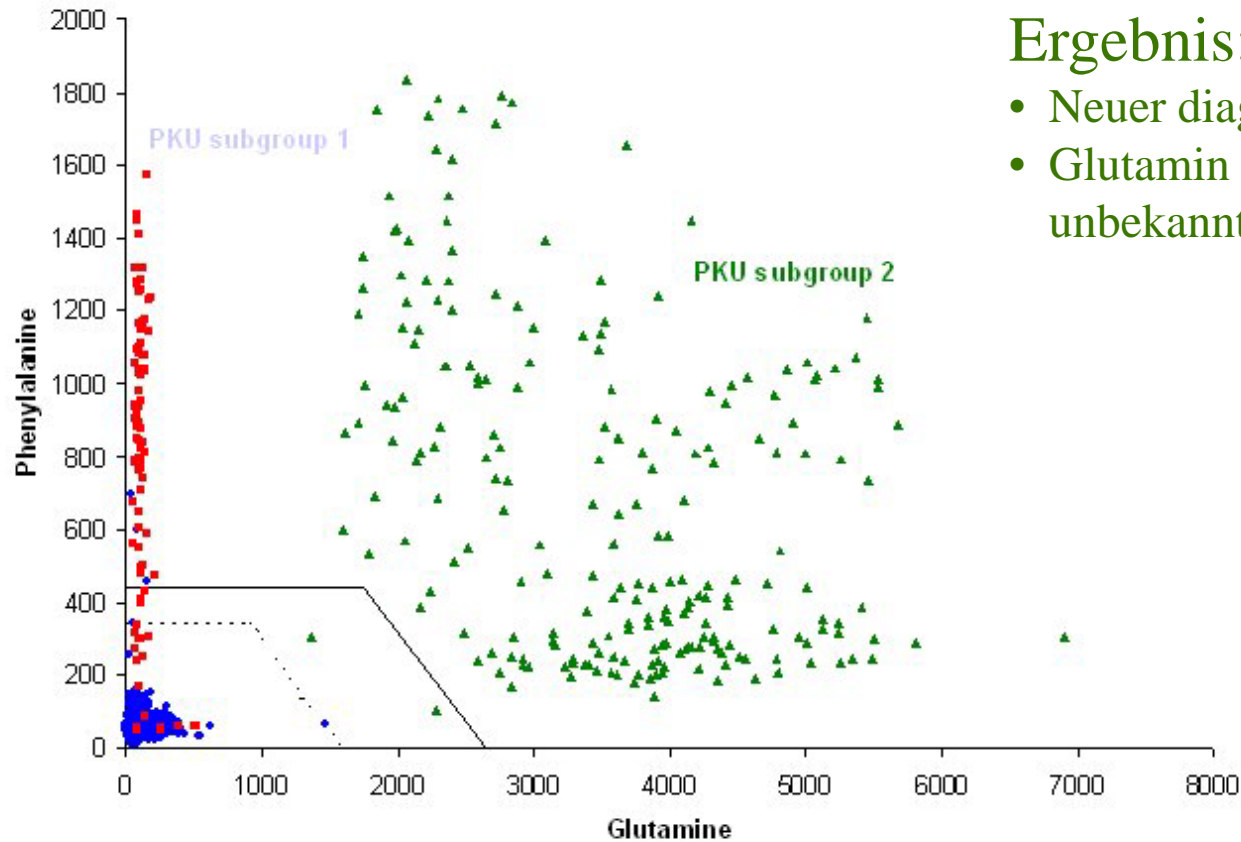
14 analysierte Aminosäuren:

alanine
 arginine
 argininosuccinate
 citrulline
 glutamate
 glycine
 methionine

phenylalanine
 pyroglutamate
 serine
 tyrosine
 valine
 leuzine+isoleuzine
 ornitine



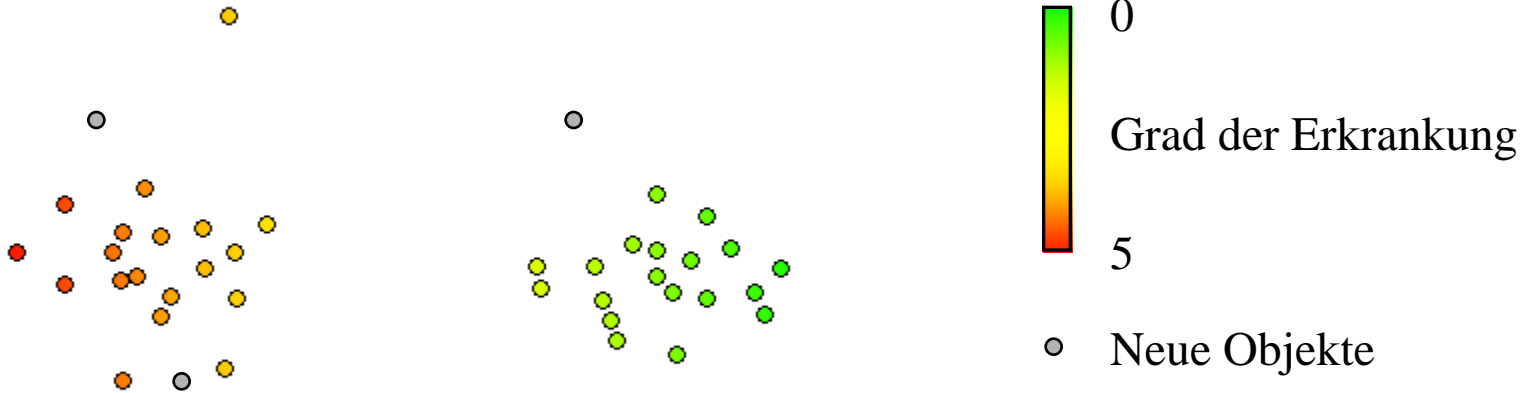
Anwendung: Neugeborenen-Screening



Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

Regression



Aufgabe:

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