Überblick

4.1 Einleitung
4.2 Clustering
~~4.3 Klassifikation~~

4.4 Outlier Detection

Outlier Detection

Was ist ein Outlier?

Definition nach Hawkins [Hawkins 1980]:

> "Ein Outlier ist eine *Beobachtung*, die sich von den anderen *Beobachtungen* so deutlich unterscheidet, daß man denken könnte, sie sei von einem anderen Mechanismus generiert worden."

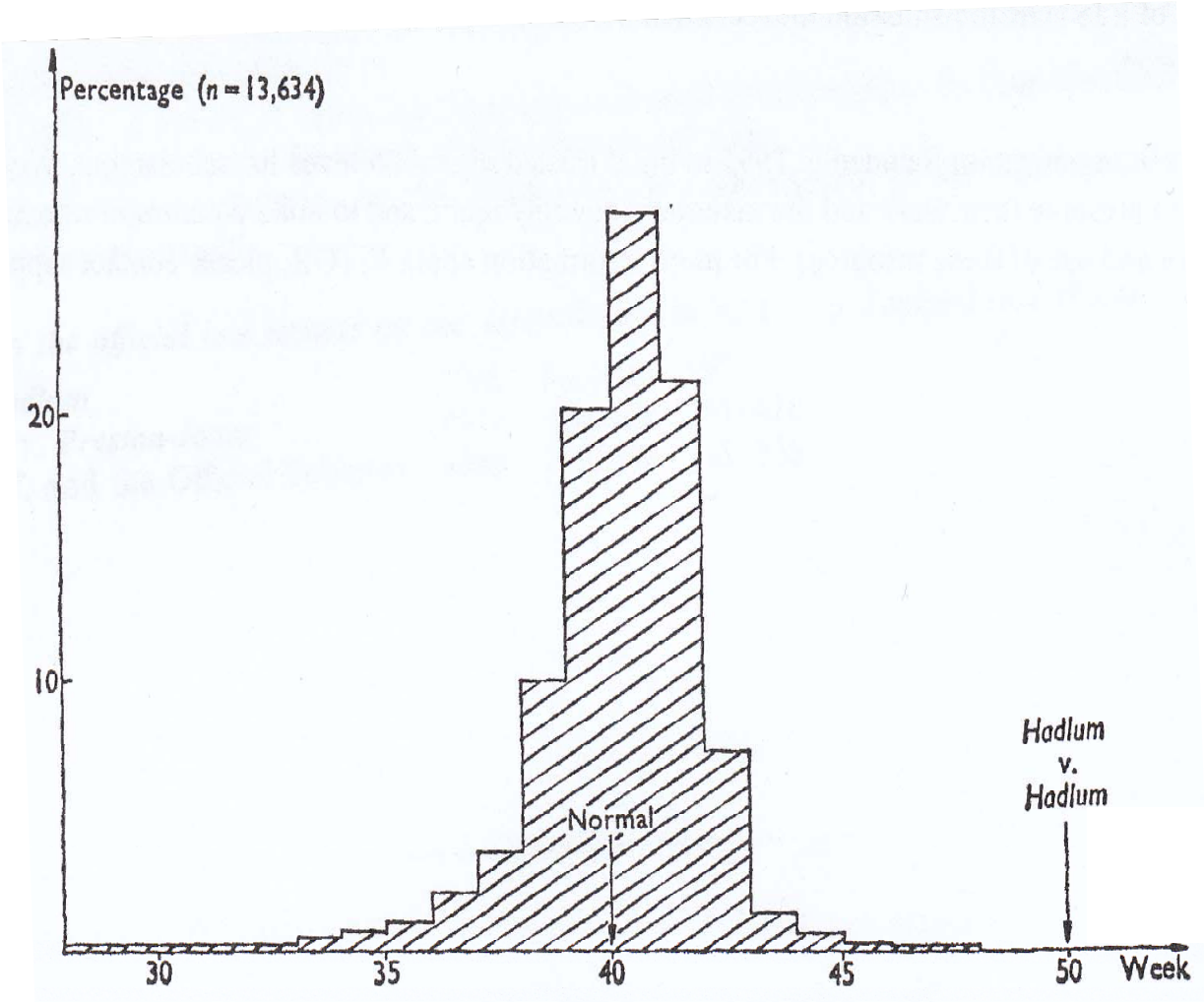
Was meint "Mechanismus"?

- Intuition aus der Statistik: "erzeugender Mechanismus" ist ein (statistischer) Prozess.
- Abnormale Daten (outlier) zeigen eine verdächtig geringe Wahrscheinlichkeit, aus diesem Prozess zu stammen.

Outlier Detection

Beispiel: Hadlum vs. Hadlum (1949) [Barnett 1978]

- Geburt eines Kindes von Mrs. Hadlum 349 Tage nachdem Mr. Hadlum zum Militärdienst abwesend war.
- Durchschnittliche Dauer einer menschlichen Schwangerschaft ist 280 Tage (40 Wochen)
- Ist eine Schwangerschaftsdauer von 349 Tagen ein Outlier?

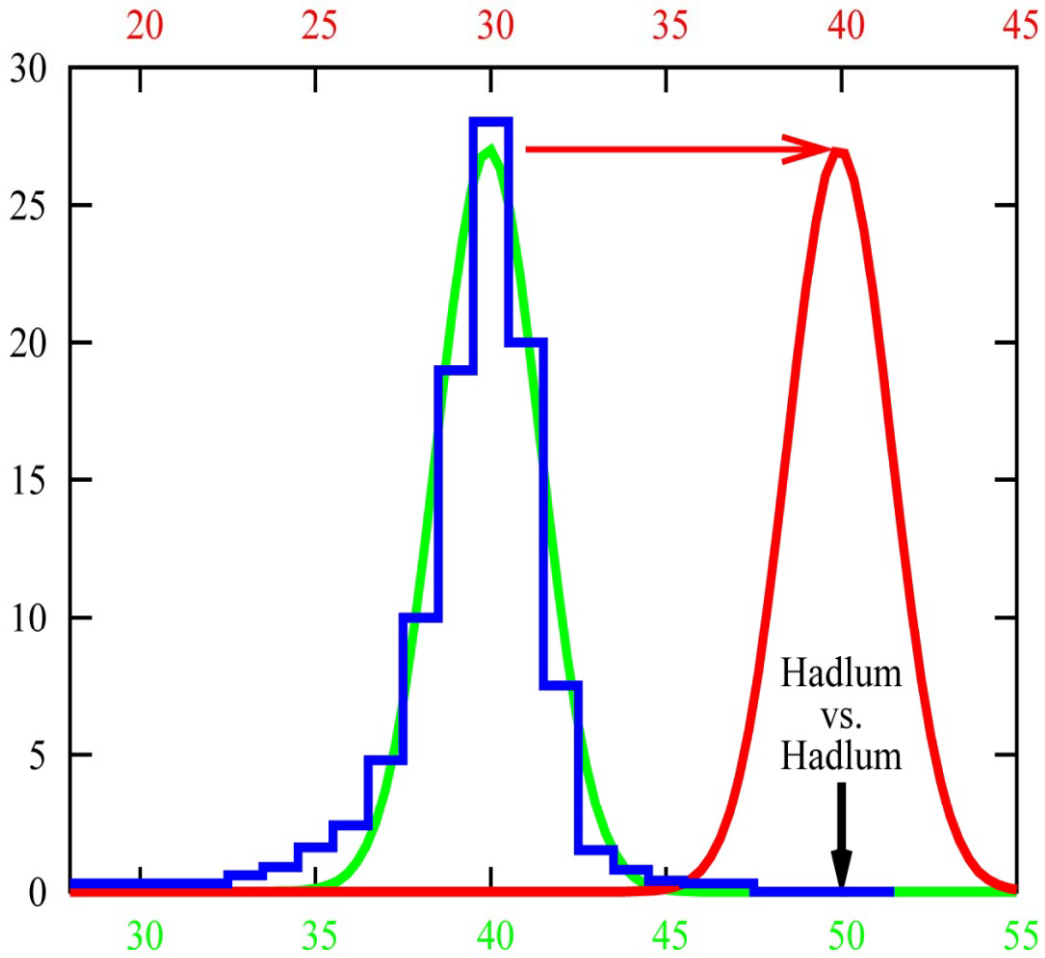


Distribution of human gestation periods.

Outlier Detection

Beispiel: Hadlum vs. Hadlum (1949) [Barnett 1978]

- Blau: statistische Beobachtungsbasis (13634 erhobene Schwangerschaften)
- Grün: angenommener zugrundeliegender Gauss-Prozess
  - sehr geringe Wahrscheinlichkeit, dass die Geburt aus diesem Prozess stammt
- Rot: Annahme von Mr. Hadlum (ein anderer Gauss-Prozess, in dem die Schwangerschaft später beginnt, ist für die Geburt verantwortlich)
  - unter dieser Annahme hat die Schwangerschaftsdauer einen Durchschnittswert und höchst-mögliche Wahrscheinlichkeit



Outlier Detection

- Anwendungsgebiete:
  - Betrugsentdeckung
    - Kaufverhalten mit einer Kreditkarte ändert sich, wenn die Karte gestohlen wurde
    - Ungewöhnliche Kauf-Muster können Kreditkarten-Mißbrauch anzeigen
  - Medizin
    - Ungewöhnliche Symptome oder Test-Ergebnisse können mögliche gesundheitliche Probleme eines Patienten anzeigen
    - Ob ein bestimmtes Testergebnis ungewöhnlich ist, kann von anderen Eigenschaften des Patienten abhängen (z.B. Geschlecht, Alter, Gewicht, …)
  - Öffentliches Gesundheitswesen
    - Auftauchen einer bestimmten Krankheit (z.B. Tetanus) verstreut über verschiedene Krankenhäuser einer Stadt zeigt Probleme mit dem zugehörigen Impfprogramm an
    - Ob das Auftreten der Krankheit unnormal ist hängt von verschiedenen Aspekten ab, z.B. Häufigkeit, räumliche Korrelation etc.

Outlier Detection

- Anwendungsgebiete:
  - Sport Statistiken
    - In vielen Sportarten werden diverse Parameter aufgezeichnet, um die Leistung eines Spielers zu bewerten
    - Außergewöhnliche (in positivem wie negativem Sinne) Spieler können durch ungewöhnliche Werte bestimmt werden
  - Erkennen von Messfehlern
    - Daten aus Sensoren (z.B. in einem wissenschaftlichen Experiment) enthalten häufig Meßfehler
    - Ungewöhnliche Werte können ein Hinweis auf Meßfehler sein
    - Solche Meßfehler aus den Daten zu entfernen, kann wichtig sein für erfolgreiche Datenanalyse und Data Mining
- Achtung:

  „One person‘s noise could be another person‘s signal.“

Outlier Detection

Statistischer Ansatz (vgl. Hadlum vs. Hadlum)

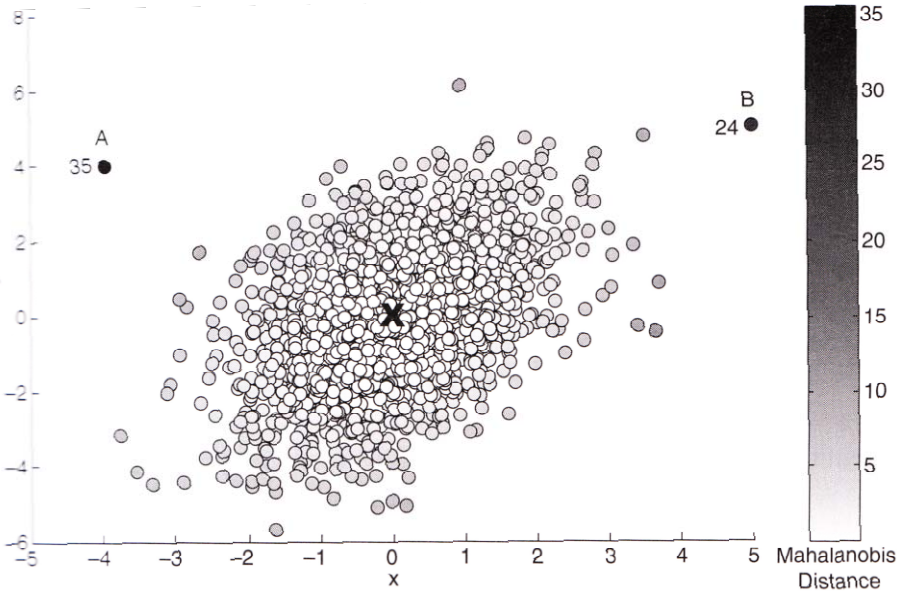
- Wahrscheinlichkeits-Dichte einer ***multivariaten*** Normalverteilung

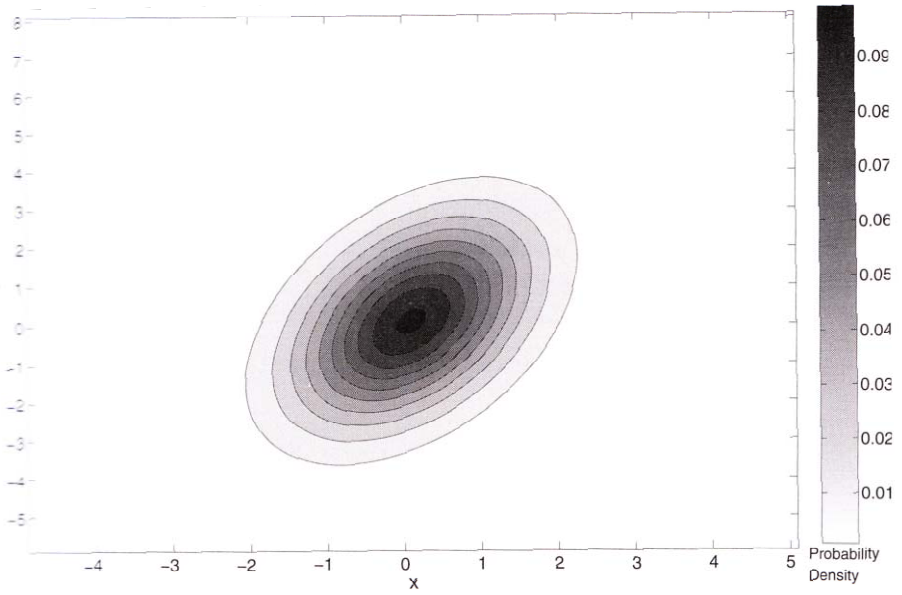
$$N(x) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)}{2}}$$

  - μ Mittelwert aller Punkte (normalerweise sind die Daten so normalisiert dass $\mu=0$)
  - Σ ist die Kovarianz-Matrix für die auf μ zentrierten Daten
  - $MDist(x,\mu) = (x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)$ ist die Mahalanobis Distanz von Punkt *x* zu μ
  - MDist folgt einer χ^2-Verteilung mit *d* Freiheitsgraden (*d* = Dimensionalität der Daten)
  - Alle Punkte *x*, mit $MDist(x,\mu) > \chi^2(0{,}975)$ $[\approx 3 \cdot \sigma]$ sind Outlier.

Outlier Detection

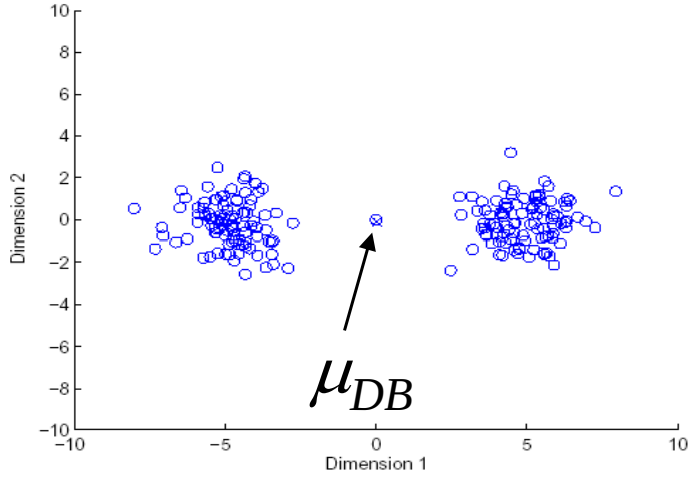
- Visualisierung (2D)





Outlier Detection

- Diskussion
  - Robustheit
    - Mittelwert/Standardabweichung, Kovarianzmatrix teilweise sehr anfällig bzgl. Ausreißern (aber diese Werte werden auf den gesamten Daten berechnet!!!)
    - The *MDist* is used to determine outliers although the *MDist* values are influenced by these outliers
  - Art der Datenverteilung muss bekannt sein (z.B. Gaussverteilung)
  - Mehrere erzeugende Prozesse? (=> Gaussian Mixture Model)



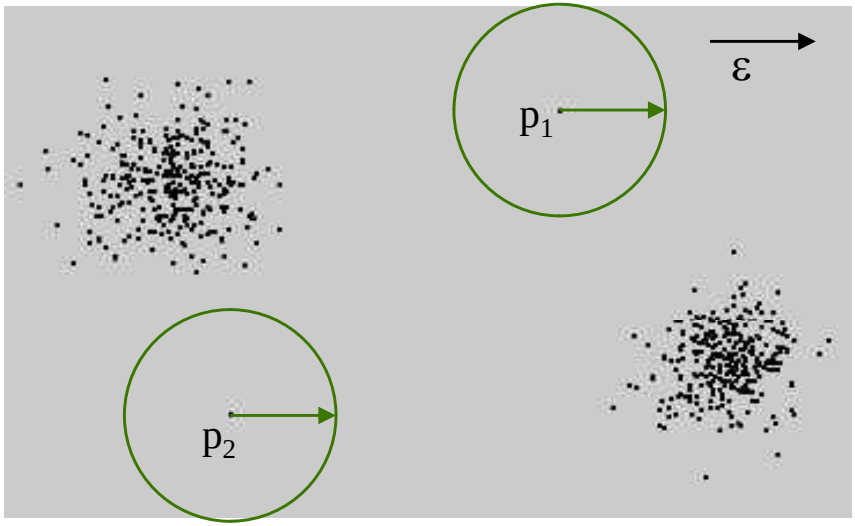
Outlier Detection

Distanzbasierte Ausreißer

- DB(ε,π)-Outlier
  - Gegeben: Radius ε und Prozentwert π
  - Punkt *p* ist ein Outlier wenn höchstens π Prozent aller Punkte eine Distanz zu *p* von weniger als ε haben

$$OutlierSet(\varepsilon,\pi) = \{p \mid \frac{Card(\{q \in DB \mid dist(p,q) < \varepsilon\})}{Card(DB)} \leq \pi\}$$

range-query mit Radius ε



Outlier Detection

- *k*NN-Distanzen
  - Die Distanz zum *k*-nächsten Nachbarn (*k*NN) eines Punktes ist dessen Ausreißer-Grad
  - Je größer die Distanz zum *k*NN, desto stärker ist der Punkt ein Ausreißer
  - Wahl von *k* kann kritisch sein (genause wie ε, π vorher)

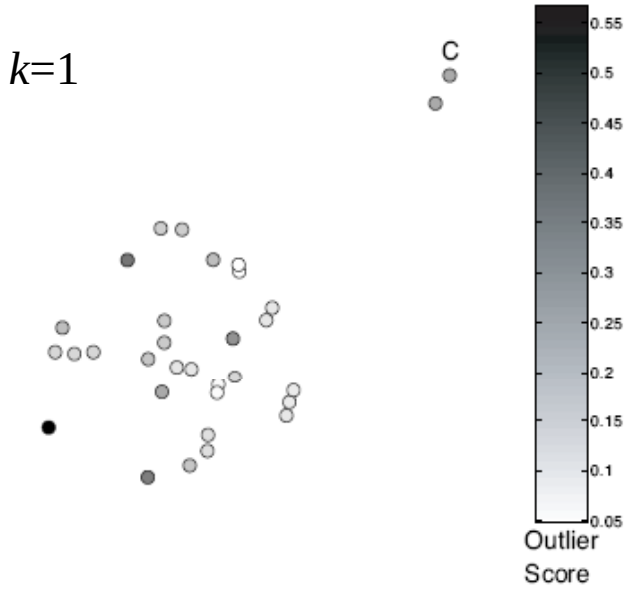


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

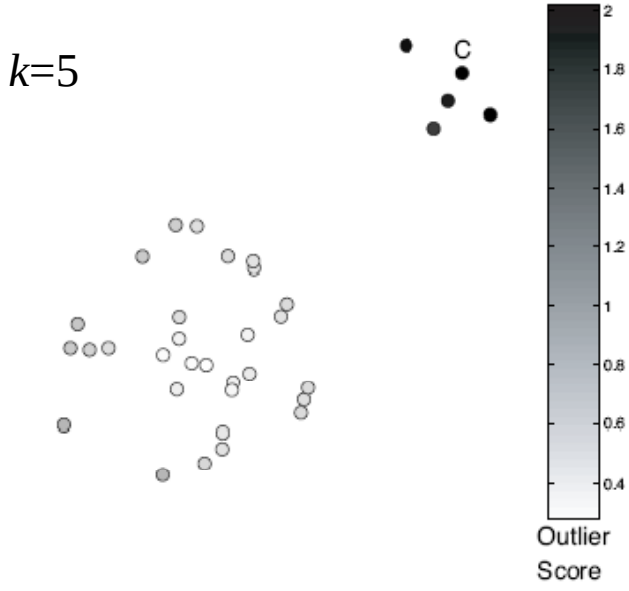


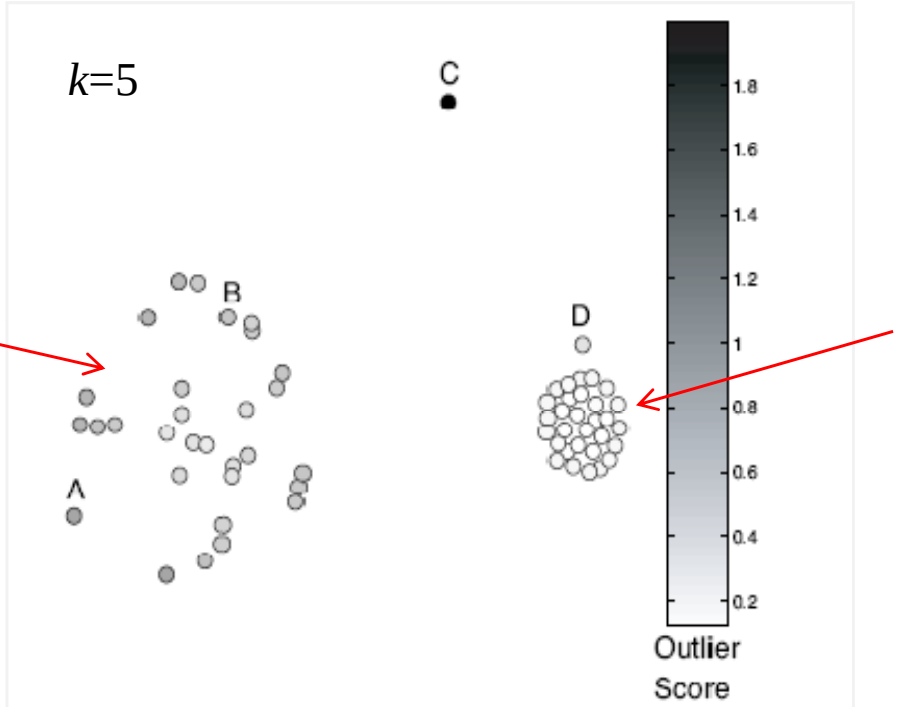
Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

Bilder aus:
[Tan, Steinbach, Kumar 2006]

Outlier Detection

- Problem: unterschiedlich dichte Cluster

sparse cluster



dense cluster

Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density. [Tan, Steinbach, Kumar 2006]

Ausblick

Data Mining und andere Wissenschaften

- Data Mining lebt von der Anwendung und muss für viele Anwendungsszenarien und Probleme zugeschnitten werden.

  Daher kann Data Mining nicht den (Anwendungs-)Experten ersetzen sondern benötigt den Dialog zwischen Experten und Informatiker

- Data Mining kann im Anwendungsgebiet (z.B. einer anderen Wissenschaft – Geographie, BWL, Kunst, Sprachwissenschaft, Physik, Biologie,…) zu neuen Erkenntnissen führen.

- Umgekehrt bietet ein konkretes Anwendungsszenario oft interessante Herausforderungen für die Forschung im Bereich Data Mining.

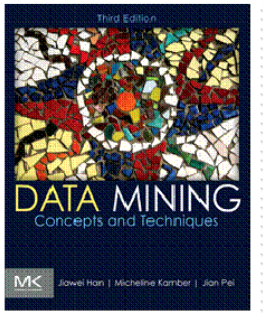
Literatur

Kurzes Lehrbuch (deutsch, aber leider schon etwas älter):

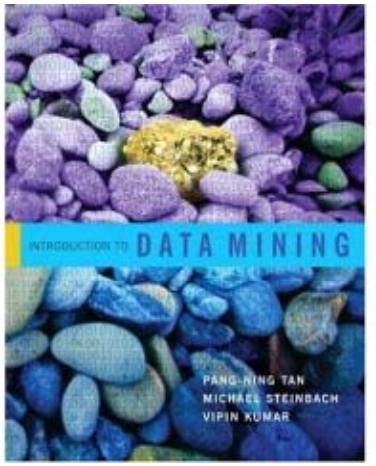
Ester M., Sander J.
Knowledge Discovery in Databases: Techniken und Anwendungen
Springer Verlag, September 2000

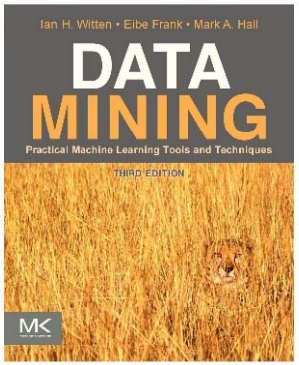
Weitere Bücher (englisch):

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