Data Warehousing und Data Mining

Einführung in Data Mining

Ulf Leser

Wissensmanagement in der Bioinformatik

Wo sind wir?

- Einleitung & Motivation
- Architektur
- Modellierung von Daten im DWH
- Umsetzung des multidimensionalen Datenmodells
- Extraction, Transformation & Load (ETL)
- Physische und logische Optimierung
- Materialisierte Sichten
- Data Mining
  - Klassifikation
  - Warenkorbanalyse
  - Clustering

Inhalt dieser Vorlesung

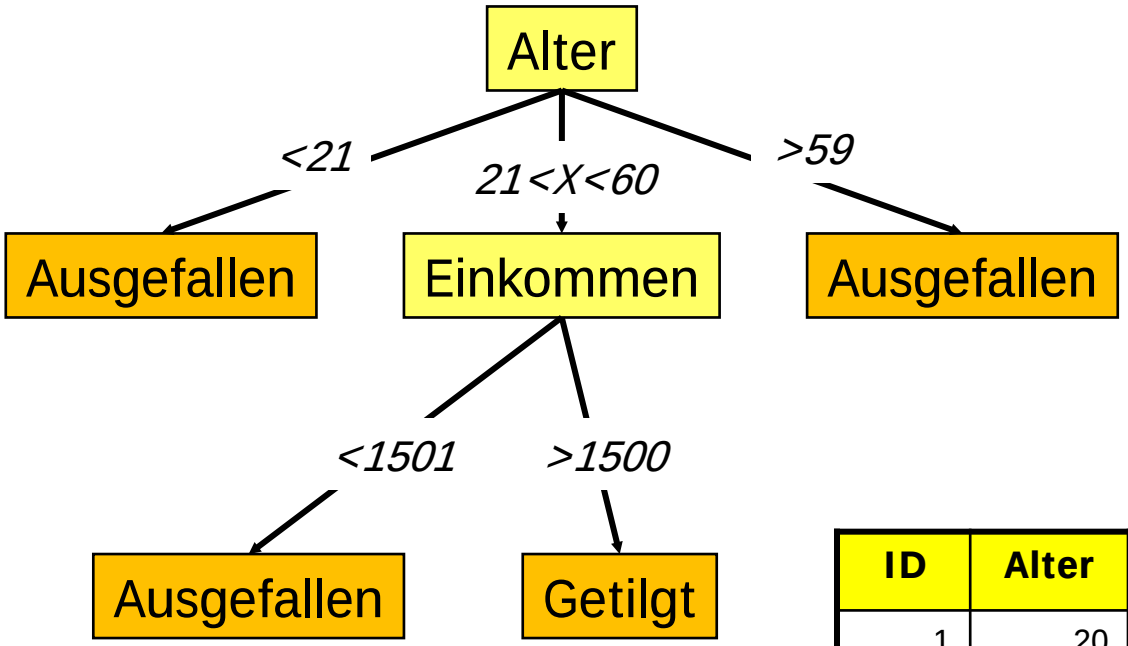
- Was ist Data Mining?
- Typische Problemstellungen & Anwendungen
- Datenaufbereitung und Exploration
- Data Mining Tools

Beispiel

ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen

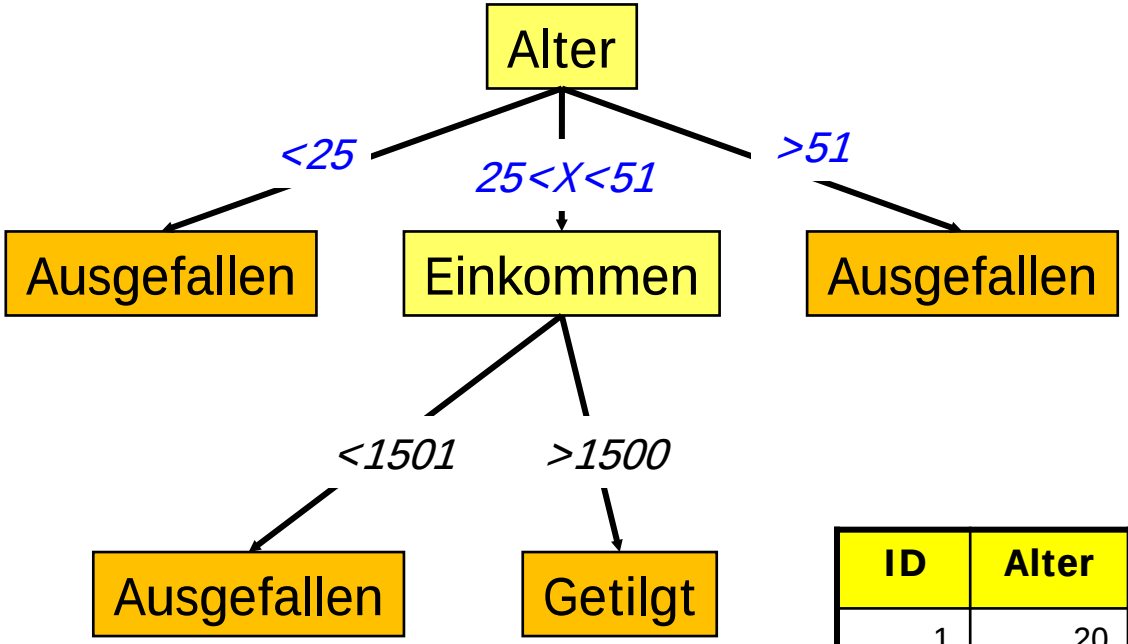
- Welches Risiko schätzen wir für eine Person von 45 Jahren mit 4000 Euro Einkommen?
- Vorhersage aufgrund bisheriger Erfahrungen

Intuitive Idee: Entscheidungsbäume



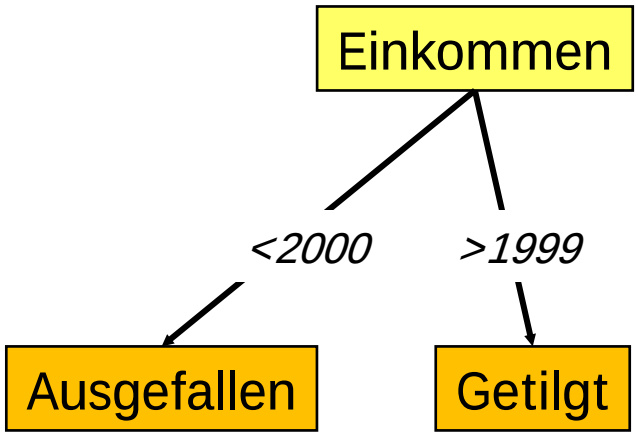
ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen

Oder ...



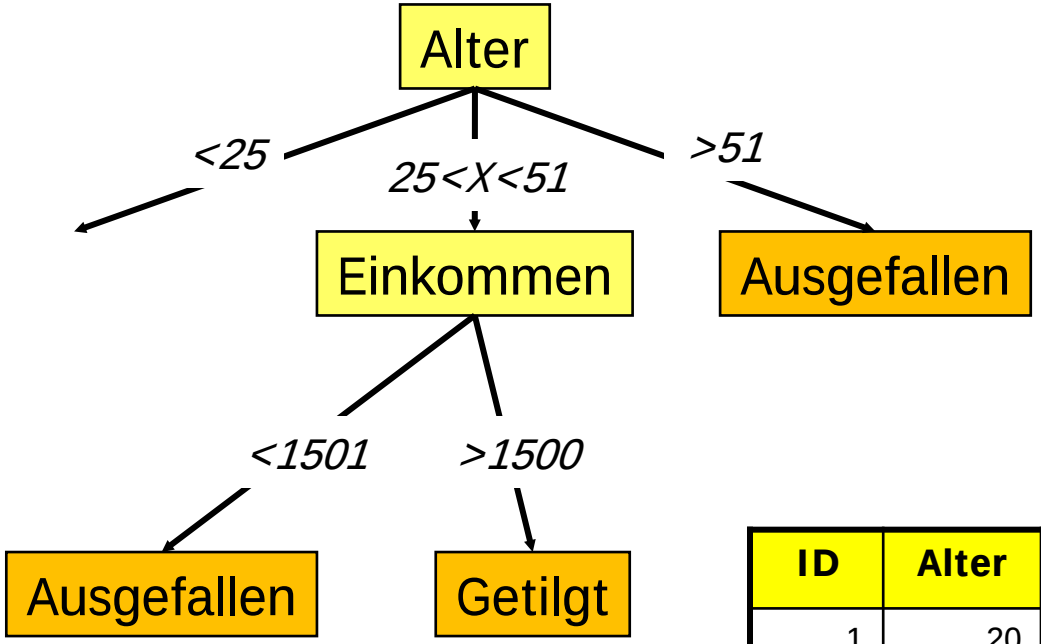
ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen

Oder ...



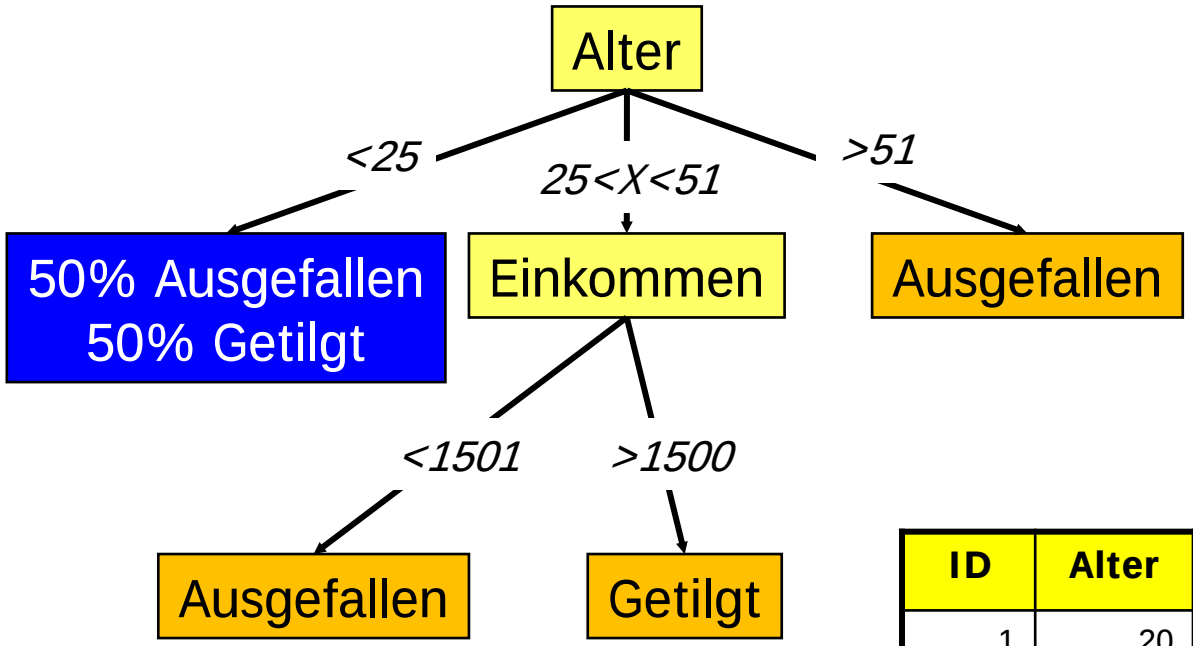
ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen

Was nun?



ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen
7	20	1500	Getilgt

Was nun?



ID	Alter	Einkommen	Risiko
1	20	1500	Ausgefallen
2	30	2000	Getilgt
3	35	1500	Ausgefallen
4	40	2800	Getilgt
5	50	3000	Getilgt
6	60	1900	Ausgefallen
7	20	1500	Getilgt

Lernen von Entscheidungsbäumen

- In welcher Reihenfolge sollen Attribute verwendet werden?
- Wie viele Splits pro Attribut?
- Wie sollen die Grenzen gewählt werden?
- Was tun bei widersprüchlichen Daten?
- Was tun bei 50.000.000 Datensätzen?
- ...

Traditionelle Analysemethode

- Manuell ausgeführte statistische Analyse
- Eher wenige Datensätze, eher wenig Attribute
- Formulieren von Hypothesen und deren Überprüfung
  - „Hypothesis-driven“
  - „Wie viele Kunden über 45 mit einem Einkommen unter 2000 Euro hatten einen Kreditausfall?“
  - Hypothesen werden vor der Datenanalyse formuliert
- DWH: Man überlegt sich mögliche Zusammenhänge und überprüft sie durch Formulieren der entsprechenden (SQL-)Anfrage

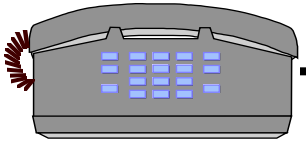
Data Mining

- „We are drowning in data and starving for knowledge“
  - „Was machen Kunden eigentlich auf meiner Webseite?“
- Riesige Datenberge
  - Business: Weblogs, Telefonate, Einkäufe, Börsendaten, ...
  - Forschung: Astronomie, Teilchenphysik, Bioinformatik, ...
  - Jeder: Nachrichten, Blogs, Webseiten, Fotos, ...
  - Millionen oder Milliarden von Datensätzen
  - Hochdimensionale Daten mit Hunderten von Attributen
- Formulierung von Hypothesen schwierig: Es gibt zu viele
- „Data-Driven“: Automatische Generierung und Prüfung von Hypothesen
  - Vorsicht: Irgendwas findet man immer

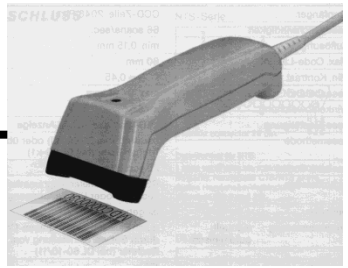
Beispiele

Welche Kunden erreiche ich mit welcher Werbung am Besten?

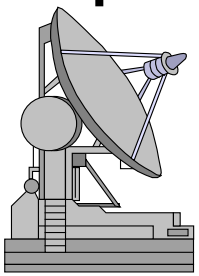
Welche Assoziationen bestehen zwischen den in einem Supermarkt gekauften Waren?



DWH



Bei welchen Telefonkunden besteht der Verdacht eines Betrugs?



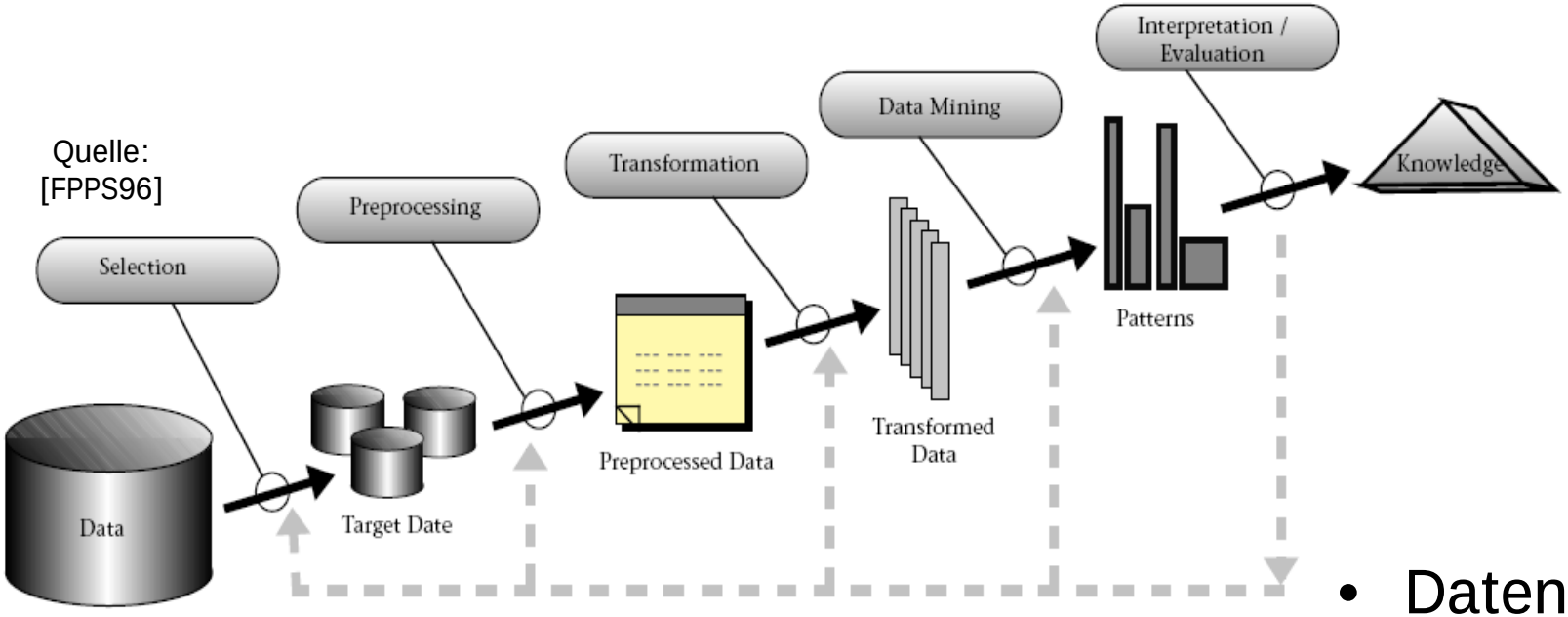
Zu welcher Klasse gehört dieser Stern?

Knowledge Discovery in Databases [FPSS96]

- "KDD is the non-trivial process of identifying valid, novel, useful and ultimately understandable patterns in data"
  - Valid: Muster sind im statistischen Sinne valide
    - Signifikant = wahrscheinlich nicht durch Zufall erzeugt
  - Novel: Bisher unbekannt
  - Useful: Man kann damit Wertschöpfung erzielen
  - Understandable: Benutzer verstehen die Muster (und die Daten)
- Viel Interpretationsspielraum

KDD als Prozess

Quelle:
[FPPS96]



- Datenauswahl
- Datenvorverarbeitung
- Datenreduktion
- Explorative Datenanalyse
- Data Mining
- Interpretation und Anwendung

Begriffe

- KDD, Data Mining, Machine Learning, Business Intelligence, Artificial Intelligence
- BI umfasst alle Techniken, die dem Business helfen und auf Datenanalyse betonen
  - Klassisch: OLAP, heute auch Data Mining und ML
- KDD heute praktisch synonym zu Data Mining
- Machine Learning: Bestimmte Formen des DM ohne den Datenmanagementaspekt
  - ML ist praktisch immer main memory bound und will pre-processing gerne ignorieren
- ML ist ein Teilgebiet des AI
  - Aber heute fast synonym

Inhalt dieser Vorlesung

- Was ist Data Mining?
- Typische Problemstellungen
  - Klassifikation
  - Clustering
  - Assoziationsregeln
- Datenaufbereitung und Exploration
- Data Mining Tools

Eingabe

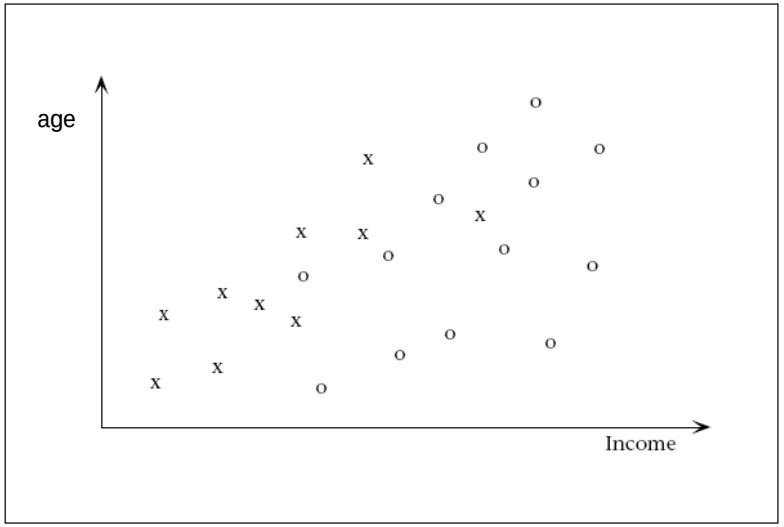
- Eine Menge $O=\{o_1, o_2,\ldots ,o_n\}$ von Objekten
- Jedes Objekt o_i wird beschrieben durch Werte für eine Menge von Attributen $A=\{a_1, a_2,\ldots ,a_m\}$
  - Heißen auch Dimensionen oder Feature
- Attributwerte können kategorial oder kontinuierlich sein
- Attributwerte können geordnet, halbgeordnet, ungeordnet sein

Drei klassische DM Aufgaben

- Klassifikation
  - Gegeben eine Menge von Objekten und eine Menge von Klassen
  - Welcher Klasse gehören die unklassifizierten Objekte an?
  - Beispiel: Fraud-Detection bei Kreditkarten
- Clustering
  - Gegeben eine Menge von Objekten
  - Gibt es Gruppen (Cluster) ähnlicher Objekte?
  - Beispiel: Segmentierung von Kunden
- Assoziationsregeln
  - Geg. Menge von jeweils gemeinsam durchgeführten Aktionen
  - Welche Aktionen kommen besonders häufig zusammen vor?
  - Beispiel: Welche Produkte werden häufig gemeinsam gekauft?

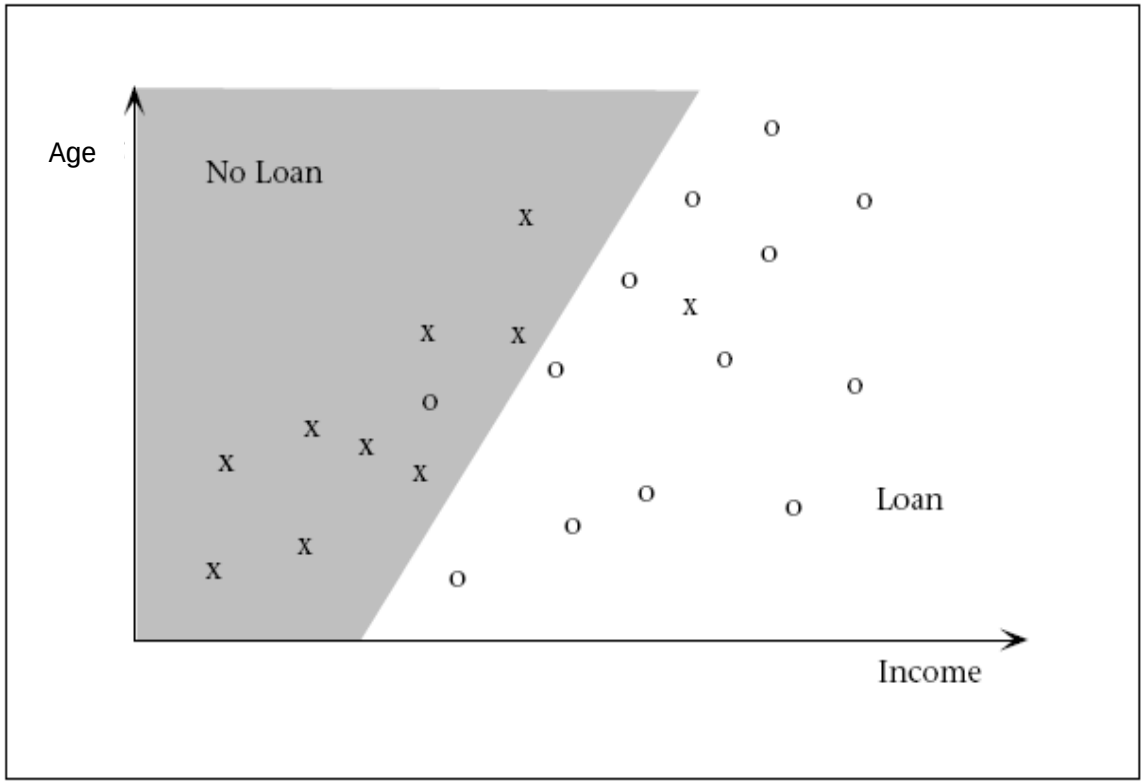
Klassifikation

- Attribute **`age, income`**
- Jeder Kunde als Punkt im zweidimensionalen Raum
- Zwei Klassen
  - Getilgt: „o“
  - Ausgefallen: „x“
- Für historische Objekte ist Klassenzugehörigkeit bekannt



- Finde Funktion, die neue Objekte einer Klassen zuordnet
  - Für neue Kunden also ihre Klasse vorhersagt

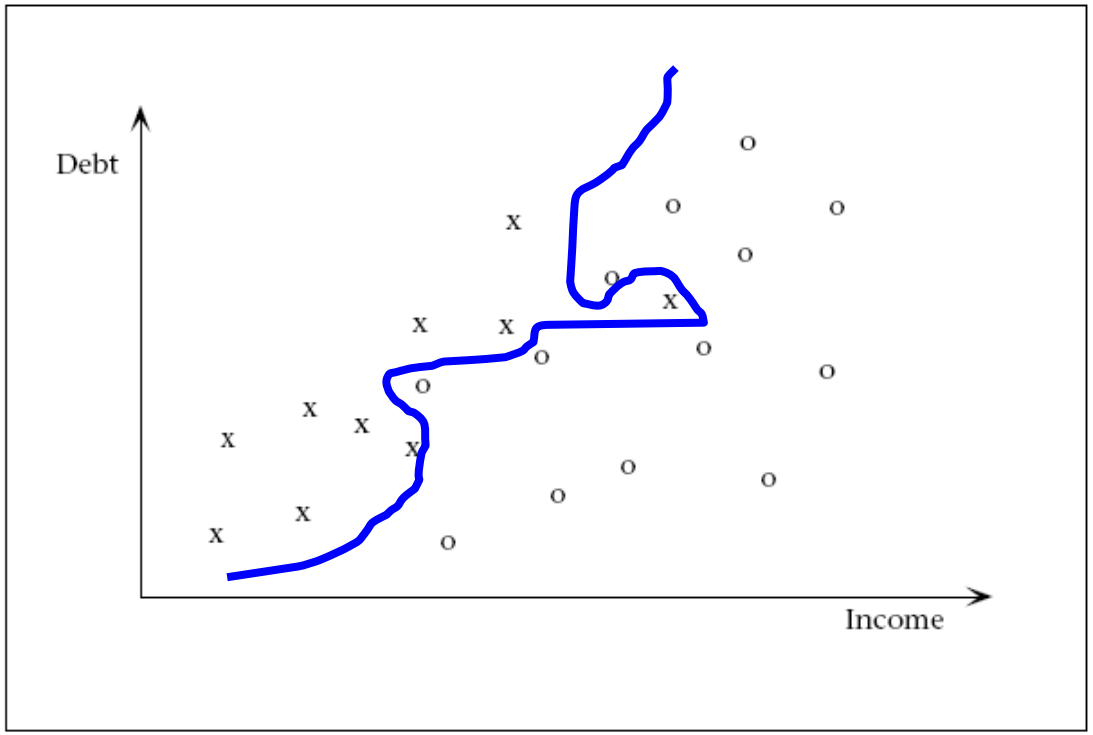
Lineare Trennung



Quelle:
[FPPS96]

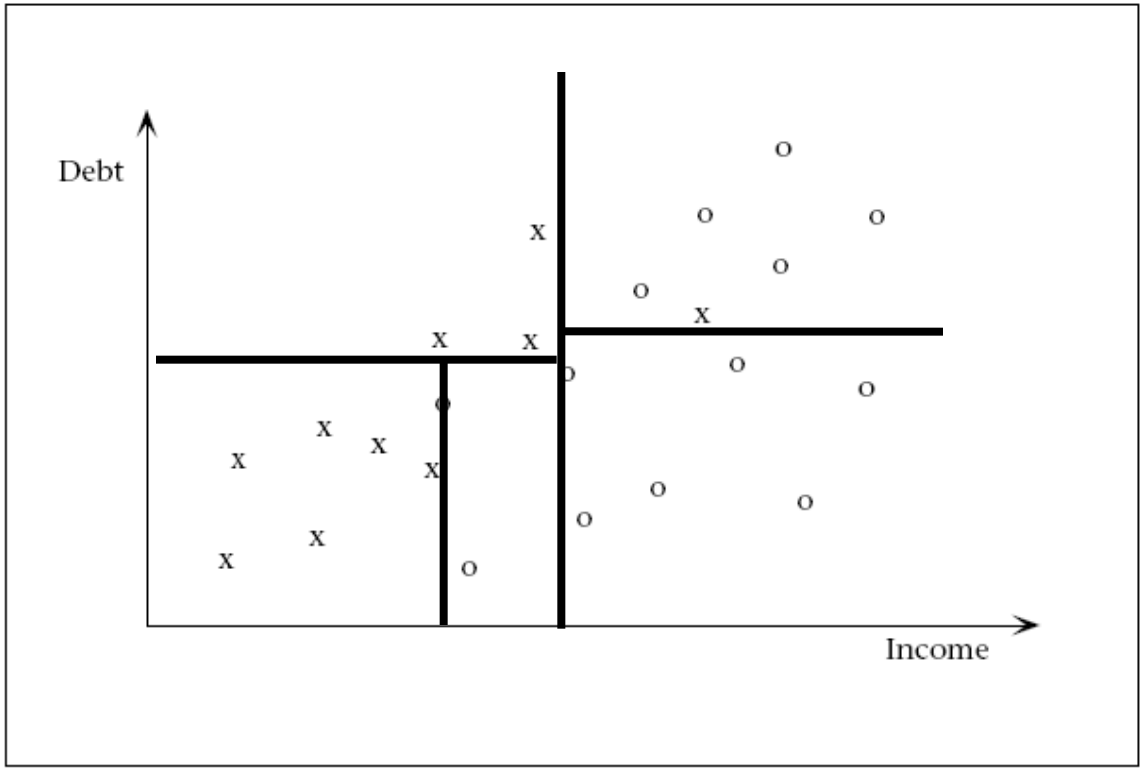
- Berechnung der Trennfunktion, die den Fehler minimiert
  - Komplexere Funktionen als lineare sind möglich
- Geht nur bei numerischen Attributen

Overfitting



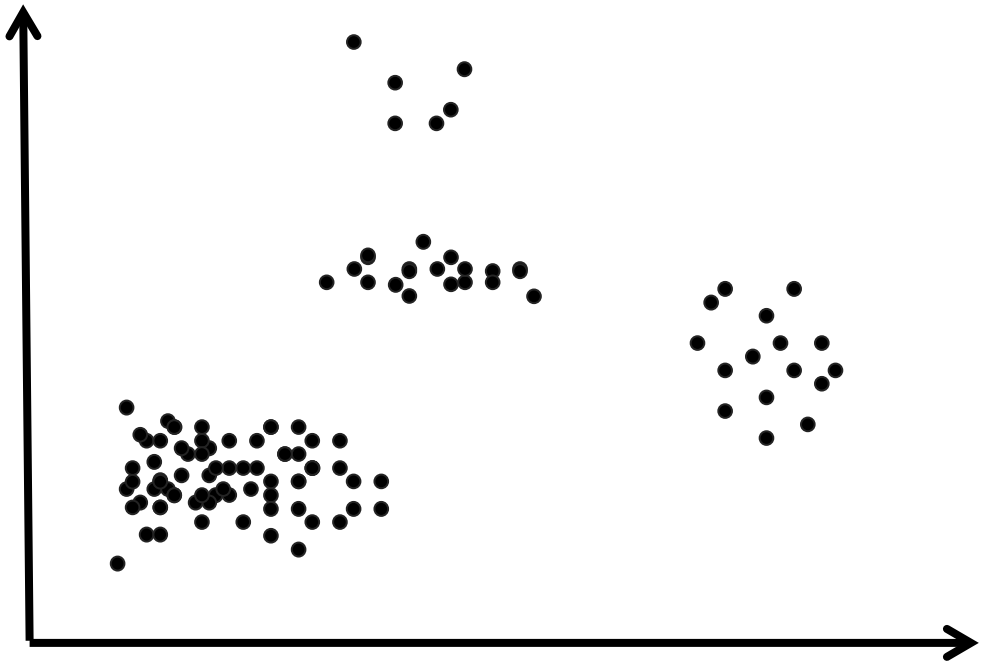
- Overfitting
  - Modell ist perfekt für Trainingsdaten
  - Aber sehr wahrscheinlich schlecht für andere Daten

Hierarchische Aufteilung



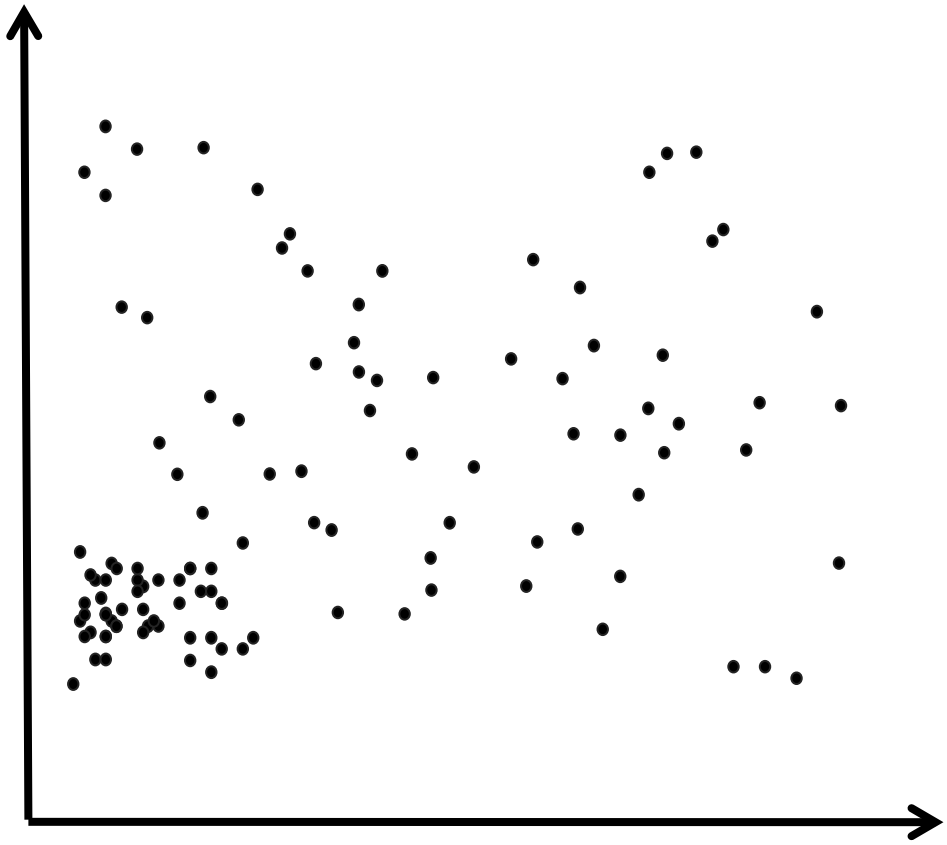
- Verwendung lokaler Trennfunktionen
- Siehe Entscheidungsbäume

Clustering



- Finde Gruppen zusammengehöriger Objekte
- Benötigt Abstandsmaß für Objekte definiert auf Attributen
  - Es soll gelten: zusammengehörend = „nahe“ bzgl. Maß

Clustern Daten?



Nicht immer einfach

Quelle:
[ES00]

- Problem schlechter definiert als Klassifikation
  - Wie groß sollen die Cluster ein?
  - Welche Form dürfen die Cluster haben?
  - Wie viele Cluster erwartet man?
  - Müssen alle Punkte geclustert werden?
  - Dürfen sich Cluster überlappen?
  - ...

Association Rule Mining

- Welche Items wurden häufiger als t Mal zusammen verkauft?

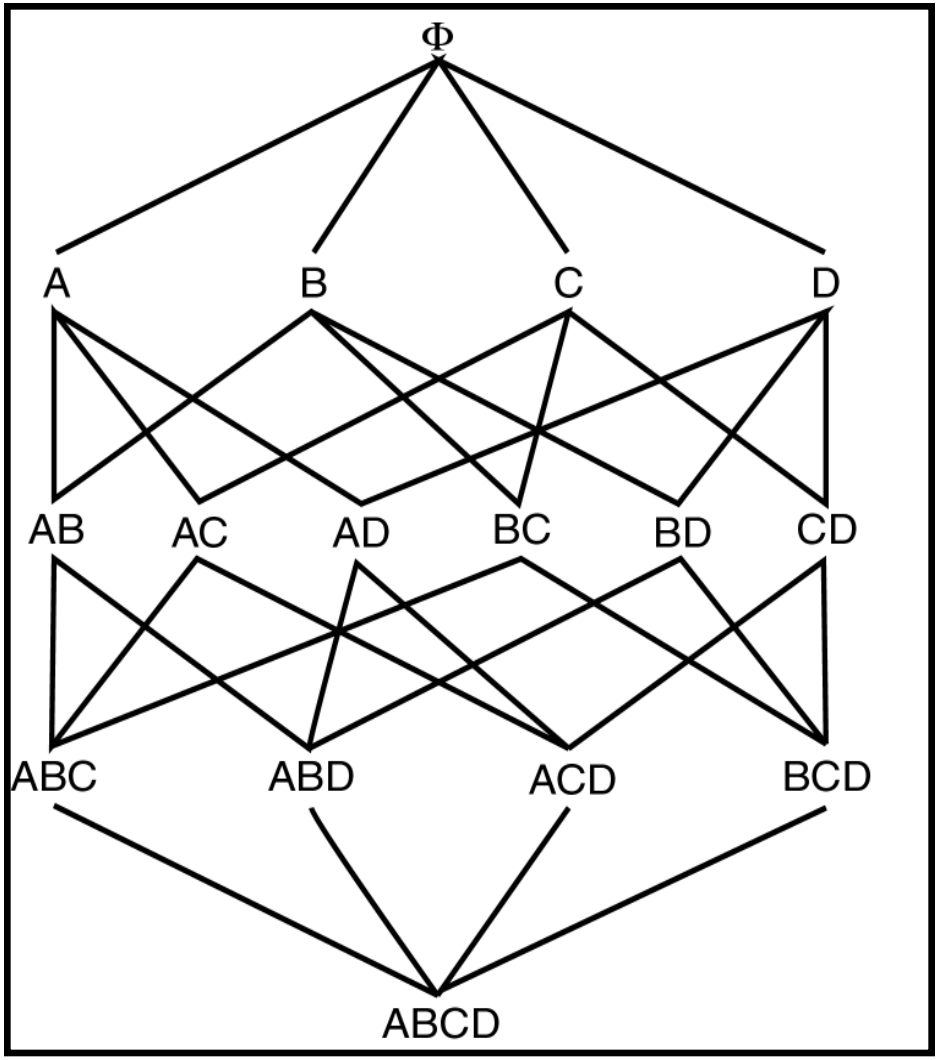
Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Quelle: [Dun02]

- Problem: Es gibt so viele mögliche Itemsets!
  - Wie viele?

Grundprinzip: „Large Itemset property“

- Jede Subgruppe eines häufigen Itemsets muss häufig sein
- ... oder ...
- Häufige große Itemsets müssen aus häufigen kleinen Itemsets bestehen



Weitere KDD Themen

- Oracle Advanced Analytics Whitepaper, 2014
  - Predict customer behavior *(Classification)*
  - Predict or estimate a continuous value *(Regression)*
  - Find profiles of targeted people or items *(Decision Trees)*
  - Identify most important factor *(Attribute Importance)*
  - Segment a population *(Clustering)*
  - Find fraudulent or “rare events” *(Anomaly Detection)*
  - Determine co-occurring items in a “baskets” *(Associations)*
- Recommendation engines

Weitere KDD Themen

- *Classification*
- *Regression*
- *Decision Trees*
- *Attribute Importance*
- *Clustering*
- *Anomaly Detection*
- *Associations*

Predictive

Descriptive

KDD auf anderen Datentypen

- Text-Mining: Clustering und Klassifikation von Texten
  - Patentanalyse; Sentimentanalyse; Marktbeobachtung; gezieltes Verschicken von Post; ...
- Web-Mining
  - Welche Webseiten werden häufig in einer bestimmten Reihenfolge besucht? Wann werden interaktive Elemente benutzt? Wie kommen Kunden mit meiner Webseite klar? ...
- Spatial Mining
  - Wo soll der nächste Supermarkt hin? Hat der Wohnort Einfluss auf Kreditwürdigkeit? Sind Cluster räumlich homogen? ...
- Graph-Mining
  - Struktur sozialer Netzwerke, Web als Graph, biologische Netzwerke, ...
- ...

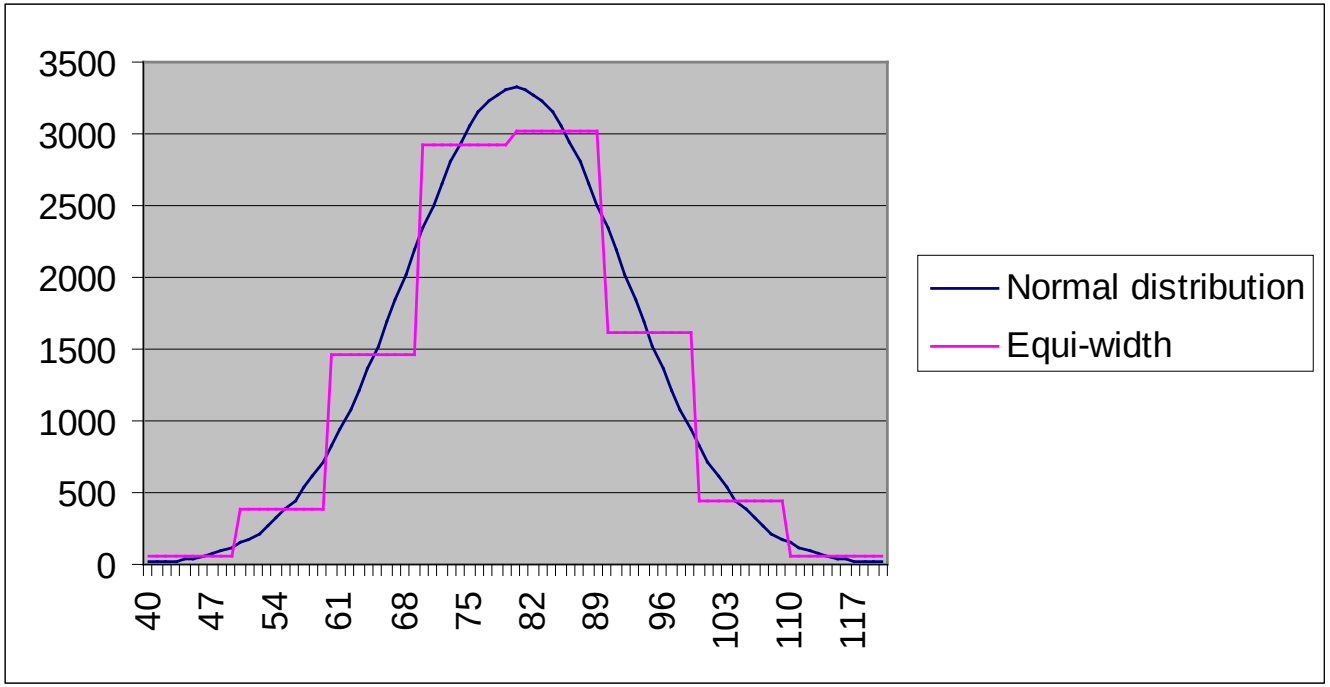
Inhalt dieser Vorlesung

- Was ist Data Mining?
- Typische Problemstellungen
- Datenaufbereitung und Exploration
- Data Mining Tools

Datenaufbereitung

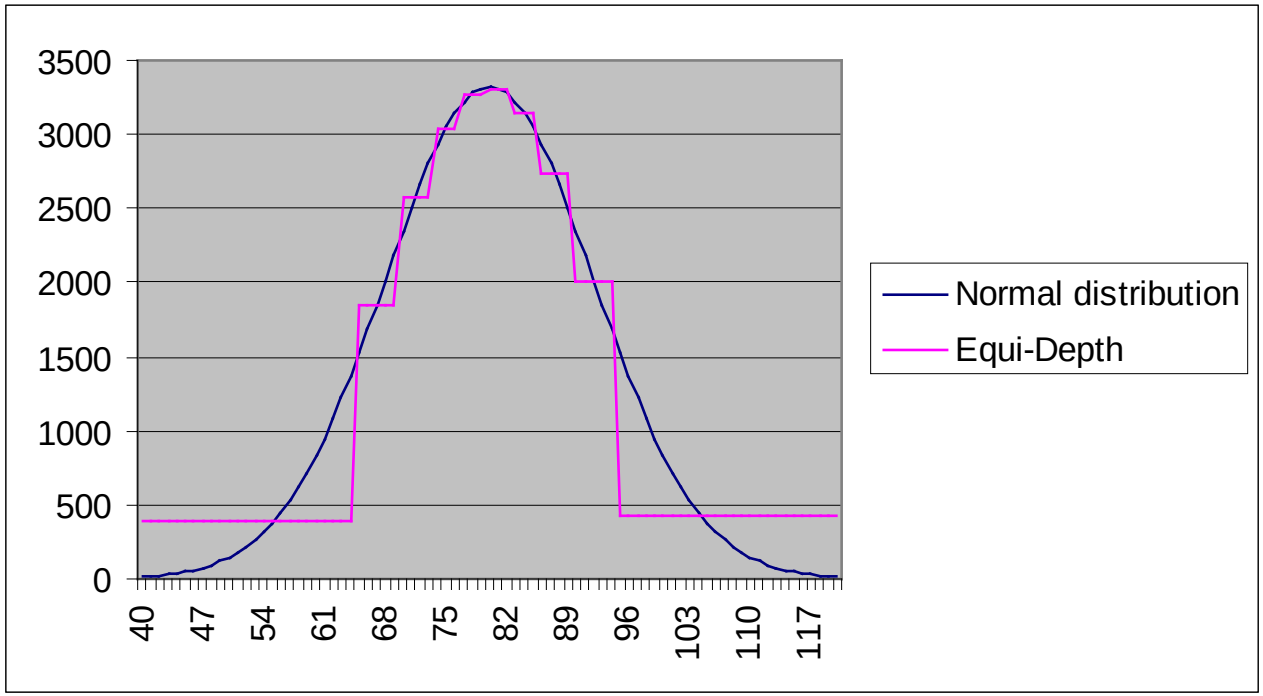
- Preprocessing: Herstellung einer homogenen, vollständigen und bereinigten Datenbasis
  - Alles aus ETL: Transformation, Plausibilität, Umrechnung, ...
  - Viele DM Verfahren reagieren empfindlich auf Ausreißer, fehlende Werte, Datenfehler etc.
  - Ersetzung von fehlenden Werten durch Schätzen, Extrapolation
  - Diskretisierung von Werten (Binning)
    - Z.B. Einteilung des Einkommens von Kunden in 5 Bereiche
    - Glättet Ausreißer, reduziert die Zahl verschiedener Werte

Binning: Equi-Width Histograms



- Zahl der Bins festlegen und Raum äquidistant aufteilen
- Bins enthalten unterschiedlich viele Objekte
- Bei Ausreissern (z.B. ein falscher, viel zu großer Wert) sind viele Bins leer, weil der „Raum“ falsch abgeschätzt wird
- Berechnung durch einen Scan

Equi-Depth



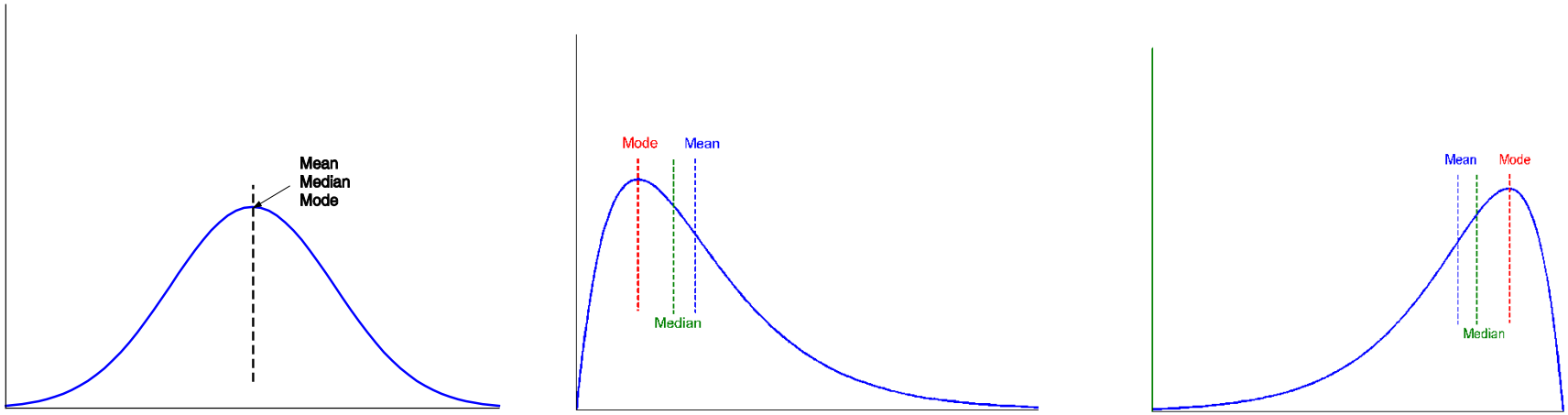
- Zahl der Bins festlegen und Raum so aufteilen, dass alle Bins gleich viele Tupel enthalten
- Führt zu gleichgroßen Bins mit unterschiedlicher Breite
- Unempfindlich gegenüber Ausreissern
- Berechnung durch Sortieren + Scan

Explorative (deskriptive) Datenanalyse

- Ziel: „Gefühl“ für die Daten bekommen
  - Welche Werte sind wie häufig?
  - Unterliegen die Werte einer bestimmten Verteilung?
  - Sind zwei (oder mehr) Attributwerte stark korreliert?
- Bei 2.000.000.000 Tupeln nicht einfach
- Vorbereitung zur Auswahl des Data Mining Verfahrens
- Hier: Nur ganz einfache statistische Kennwerte
  - Und deren Berechnung im DWH

Univariate Beschreibung

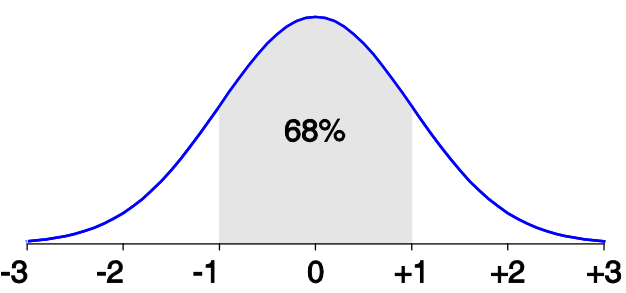
- Beschreibung der Verteilung der Werte eines Attributs
- Suche nach einer möglichst kompakten Beschreibung
- Alle Werte erfassen: Verteilungsfunktion
- Mit einem Wert charakterisieren: Mittelwert, Median, Modus

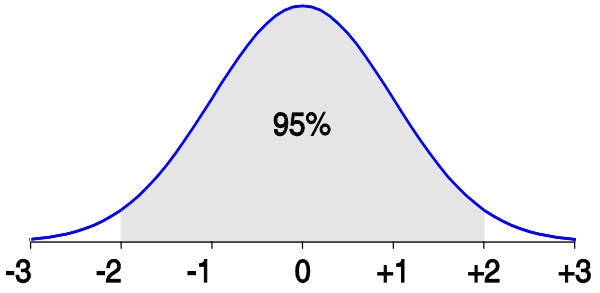


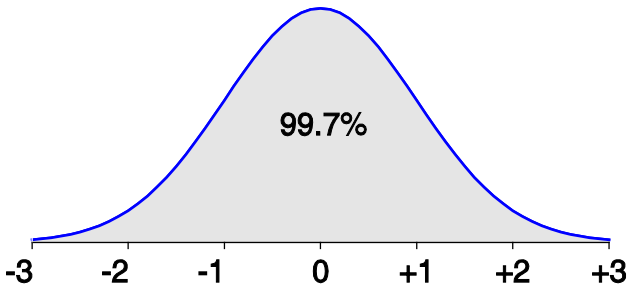
Quelle: [HK05]

Normalverteilte Daten

- Sehr viele Daten sind normalverteilt
- Zwei Werte: Mittelwert und Varianz
  - [μ–σ, μ+σ]: Ca. 68% der Datenpunkte
  - [μ–2σ, μ+2σ]: Ca. 95% der Datenpunkte
  - [μ–3σ, μ+3σ]: >99% der Datenpunkte
- Testen z.B. mit Shapiro-Wilk-Test

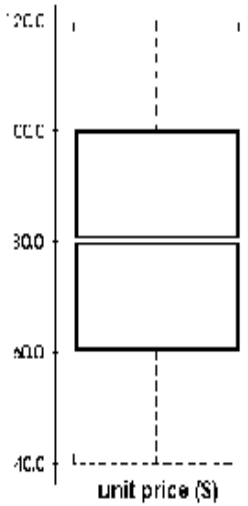




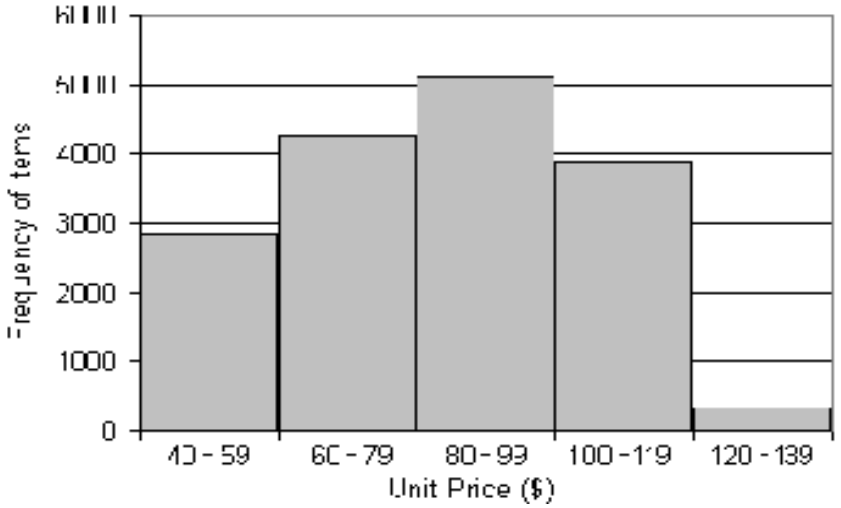


Visualisierung von Verteilungen

- Boxplots
  - Min und max
  - Erstes und drittes Quartil
    - Oder STDDEV bei normalverteilten Daten
  - Mittelwert und (meist) Median



- Histogramme



SQL

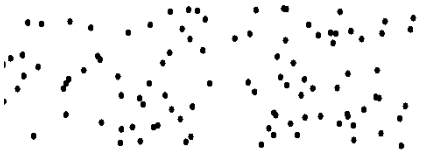
- Standard SQL: **`avg, stddev, median, quartile`**
- Wie findet man den **`mode`** eines Attributs **`t.a`**?

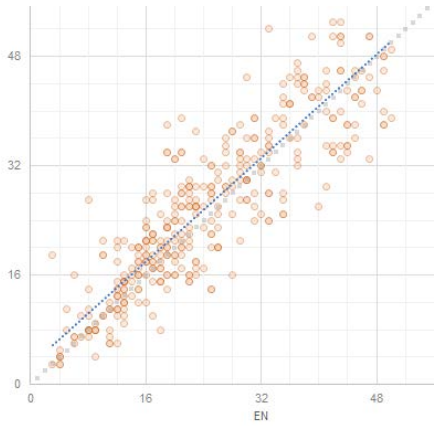
```
SELECT a, cnt
FROM (SELECT  a, count(a) cnt
      FROM    t
      GROUP BY a
      ORDER BY count(a) DESC)
WHERE ROWNUM=1;
```

```
SELECT a, count(a) cnt
FROM   t
GROUP BY a
ORDER BY count(a) DESC)
FETCH FIRST ROW ONLY;
```

Multivariate Beschreibung

- Betrachtung der gemeinsamen Verteilungen zweier oder mehr Attribute
- Einfachsten Fall: Statistische Unabhängigkeit
  - P(a|b)=p(a), p(b|a)=p(b), $p(a \wedge b) = p(a)*p(b)$
  - Dann reichen univariate Beschreibungen
  - Visuell erkennbar z.B. im Scatter-Plot





Kontingenztabellen

- Sehr oft sind Attribute aber nicht unabhängig
  - Trotzdem nimmt man das oft an um Dinge einfach zu halten
- Kontingenztabelle für kategoriale Attribute

	Mittelfristig Arbeitslos	Langfristig Arbeitslos	Summen
Ohne Ausbildung	19	18	37
Mit abgeschlossener Ausbildung	43	20	63
Summe	62	38	100

- Was erwartet man für unabhängige Attribute?

Kontingenztabellen

- Sehr oft sind Attribute aber nicht unabhängig
  - Trotzdem nimmt man das oft an um Dinge einfach zu halten
- Kontingenztabelle für kategoriale Attribute

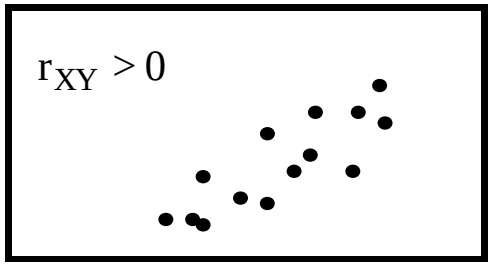
	Mittelfristig Arbeitslos	Langfristig Arbeitslos	Summen
Ohne Ausbildung	19 / 22	18 / 14	37
Mit abgeschlossener Ausbildung	43 / 39	20 / 24	63
Summe	62	38	100

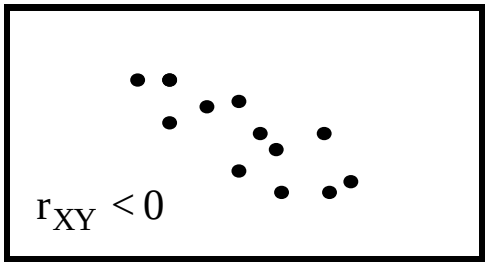
- Tests auf Unabhängigkeit, z.B. Chi-quadrat

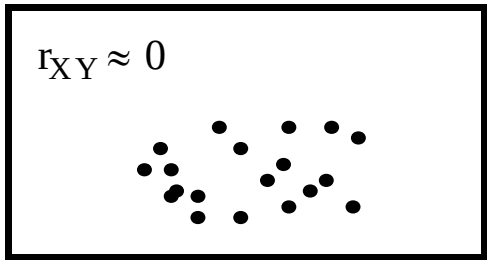
Korrelationskoeffizient

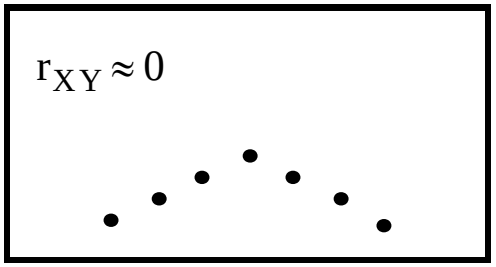
- Misst die lineare Korrelation zweier Attribute X und Y

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$









SQL

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Berechnung Kontingenztabelle für Attribute t.a und t.b?

```
SELECT   a,b,count(*)
FROM     t
GROUP BY cube(a,b);
```

- Berechnung des Korrelationskoeffizienten für t.a und t.b?

```
SELECT up/sqrt(down)
FROM  (SELECT sum((a-ma)*(b-mb)) up
       FROM t, (SELECT avg(a) ma, avg(b) mb
                FROM t) tm),
      (SELECT sum(sqr(a-ma))*sum(sqr(b-mb)) down
       FROM t, (SELECT avg(a) ma, avg(b) mb
                FROM t) tm);
```

Inhalt dieser Vorlesung

- Was ist Data Mining?
- Typische Problemstellungen
- Datenaufbereitung und Exploration
- Data Mining Tools

Data Mining Software

- Viele Open Source Machine Learning Bibliotheken
  - Meistens nicht auf Datenbanken ausgelegt – Files
  - Weka, SciKitLearn, RapidMiner, ...
- Spezielle Verfahren haben oft spezielle Tools
  - SVMLight, TensorFlow, Keras, NLTK, Mahout, ...
- Kommerzielle Tools
  - SPSS, EXCEL, MatLab, KNIME, ...
- Erweiterungen von Datenbankherstellern
  - Oracle Data Mining, SQL Server Analysis Services, DB2 Intelligent Miner

Literatur

- Han, J. and Kamber, M. (2006). "Data Mining. Concepts and Techniques", Morgan Kaufmann.
- Alpar, P. and Niedereichholz, J., Eds. (2000). "Data Mining im praktischen Einsatz". Braunschweig/Wiesbaden, Vieweg Verlagsgesellschaft.
- Dunham, A. M. H. (2002). "Data Mining". New Jersey, Pearson Education Inc.
- Ester, M. and Sander, J. (2000). "Knowledge Discovery in Databases". Berlin, Springer.
- Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases." AI Magazine 17(3): 37-54.
- Ganti, V., Gehrke, J. and Ramakrishnan, R. (1999). "Mining Very Large Databases." IEEE Computer: 38-45.

Selbsttest

- Nennen Sie einige deskriptive und einige prediktive Data Mining Verfahren
- KDD is a analytics process to find patterns in data that are ...(a) (b) ...
- Wenden Sie ein Equi-Depth Binning auf folgenden Daten an für 5 bins
- Wie wird eine Normalverteilung charakterisiert?
- Vermuten Sie bei den folgenden Beispieldaten, ob Sie einer Normalverteilung unterliegen oder nicht