Data-Warehouse-Technologien

Prof. Dr.-Ing. Kai-Uwe Sattler[1]

Prof. Dr. Gunter Saake[2]
Dr. Veit Köppen[2]

[1]TU Ilmenau
FG Datenbanken & Informationssysteme

[2]Universität Magdeburg
Institut für Technische und Betriebliche Informationssysteme

Letzte Änderung: 18.10.2019

Teil X

Business Intelligence Anwendungen

Business Intelligence Anwendungen

Business Intelligence

Vielfältige Begrifflichkeit:

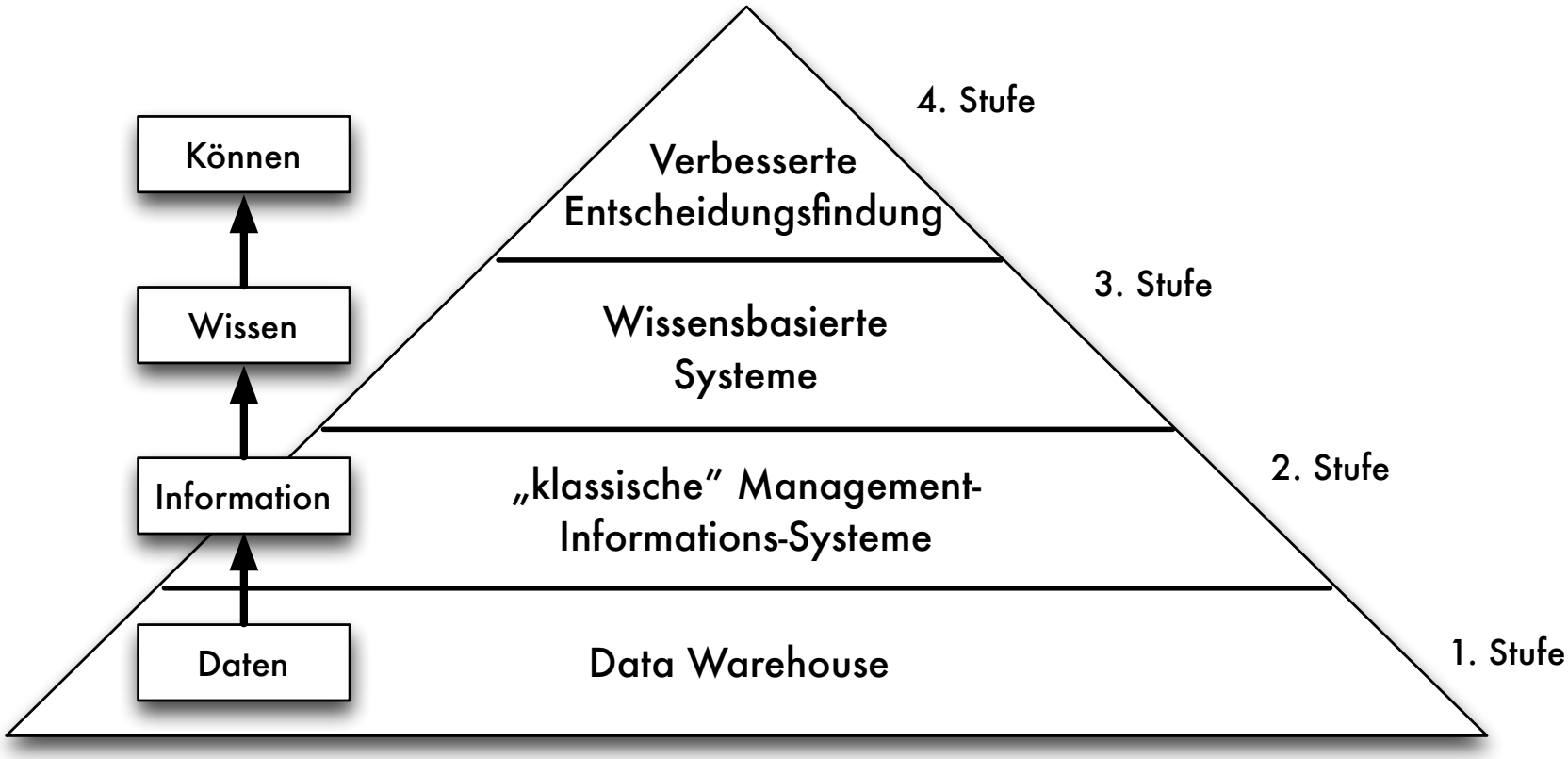
- 1989 Begriff Business Intelligence geprägt [Dresner 1989]
- ab den 60er Jahren (seit der Datenverarbeitung):
  - Management-Informations-Systeme
  - Management-Support-Systeme
  - Executive-Information-Systeme
- Unterscheidung:
  - Im engeren Sinne
  - Analyseorientiert
  - Im weiteren Sinne

Intelligence

Begrifflichkeit:

- Auffinden von Ordnungen,
- Regeln für Gemeinsamkeiten (Zusammentreffen),
- Regeln für Neben- und Nacheinanderauftreten von Ereignissen,
- Gezielte Sammlung und Weitergabe von Informationen,
- Informationslogik

Wissenspyramide



Business Intelligence

- Daten- und Informationsverarbeitung für die Unternehmensleitung
- Informationslogistik: Filterung von Informationen
- MIS: schnelle und flexible Auswertungen
- Frühwarnsystem im Unternehmen („Alerting“)
- BI = Data Warehousing
- Informations- und Wissensspeicherung
- Prozess von Erhebung $\rightarrow$ Diagnose $\rightarrow$ Therapie $\rightarrow$ Prognose $\rightarrow$ Kontrolle

[Mertens 2002]

Business Intelligence

Business Intelligence bezeichnet den analytischen Prozess, der – fragmentierte – Unternehmens- und Wettbewerbsdaten in handlungsgerichtetes Wissen über die Fähigkeiten, Positionen, Handlungen und Ziele der betrachteten internen oder externen Handlungsfelder (Akteure und Prozesse) transformiert.

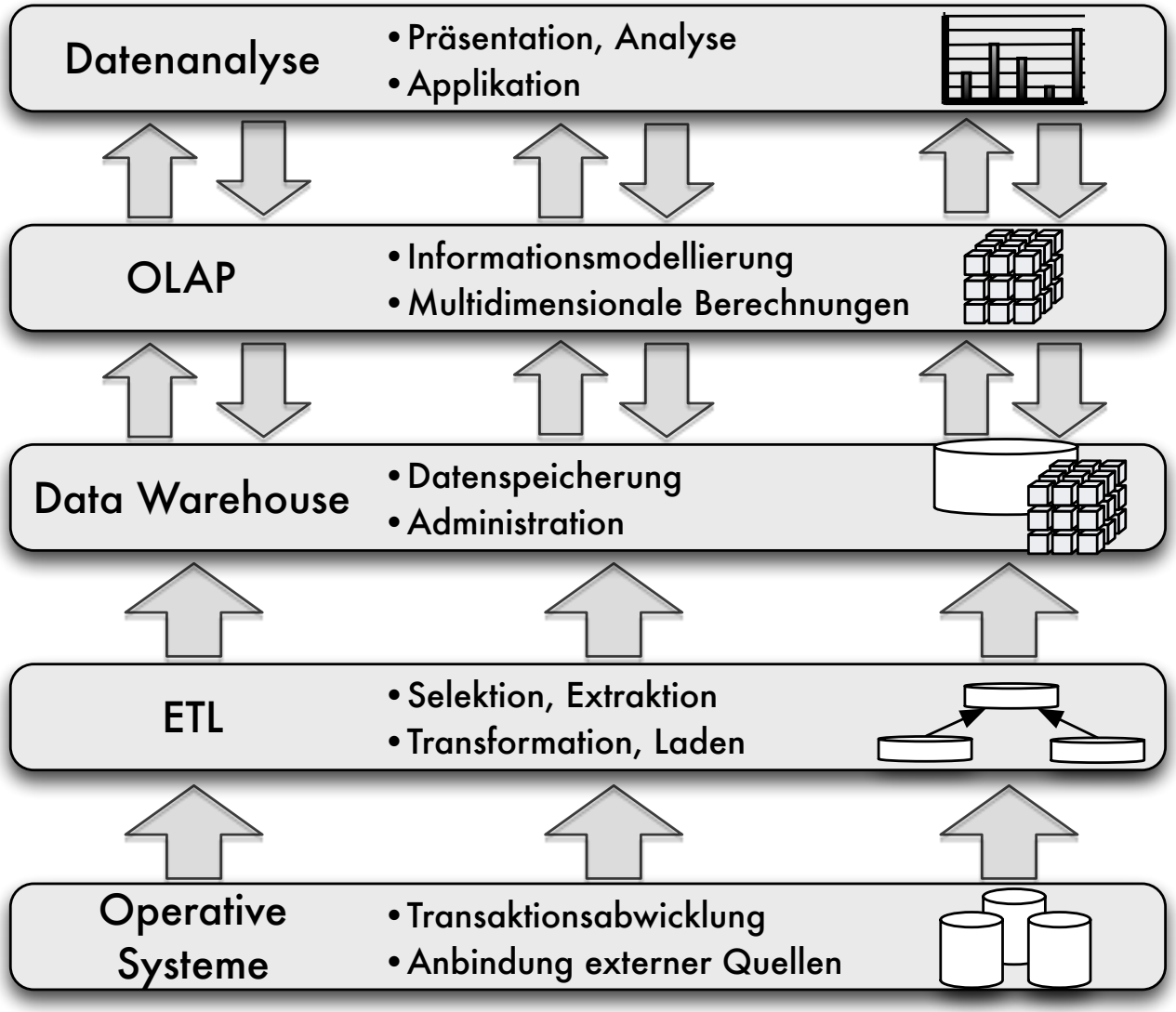
[Grothe & Gensch 2000]

- Analytischer Prozess: Planen, Entscheiden und Steuern
- Allgegenwärtige Datenintegration und -bereitstellung
- Handlungsgerichtetes Wissen: Kommunikation + Information + Wissensdarstellung

Business Intelligence Portfolio

	Unternehmens-, Markt, und Wettbewerbsanalyse	
Ausprägung der Datengrundlage:	quantitativ	überwiegend qualitativ
Entdeckungsprozess:	strukturiert	semi-strukturiert
	hypothesengestützt	hypothesenfrei
Bereitstellung (data delivery)	Data-Warehouse-Systeme Multidimensionale Modelle für: Planung, Budgetierung, Analyse, Reporting	Internet Agententechnologie (ex- und) implizites Wissen
Entdeckung (knowledge discovery)	OLAP Analysen, Balanced Scorecards ABC-Analyse, Abweichungsanalyse	Business-Simulatoren Früherkennungssysteme Data Mining, Text Mining Fallbasiertes Schließen
Kommunikation (knowledge sharing)	standardisiertes und ereignisgesteuertes Reporting Informationssysteme	Interessenprofile Issue Management traditionelles Wissens-management
	Pull und Push-Service Competitive Intelligence: Unternehmens-, Markt- und Wettbewerbsanalyse	

Business Intelligence Prozess



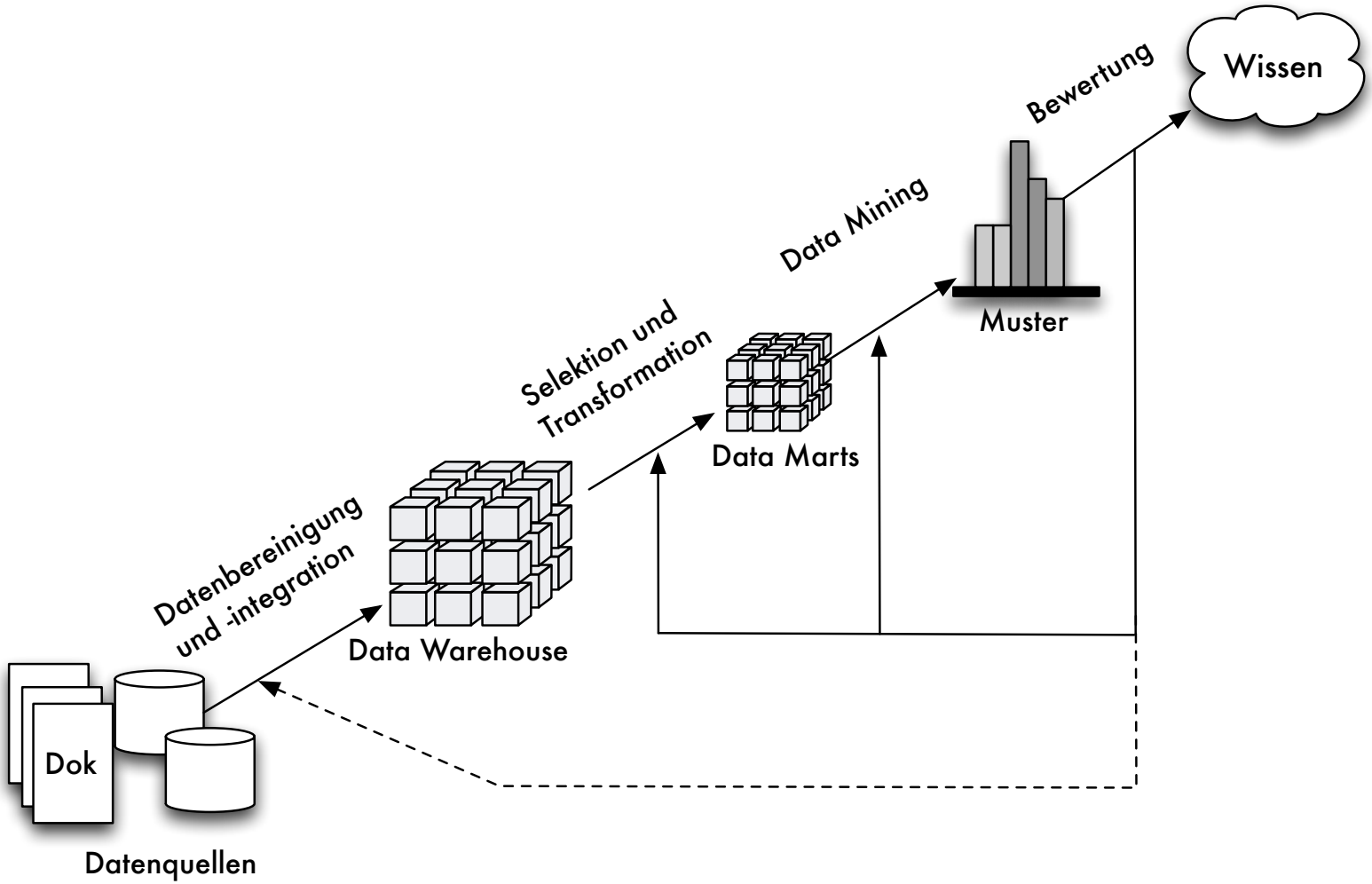
Data Warehouse und Business Intelligence

- Data Warehouse ist zentraler Informationsspeicher
- BI: Methoden zur Verbindung quantitativer, qualitativer, interner und externer Informationen
- Menge der DW-Daten muss geeignet gefiltert und aggregiert werden, um personalisierte Informationen / Wissen darzustellen
- Data Mart stellt Ausgangspunkt für domänenspezifische Analyse dar

Hohes Datenaufkommen:

- Datenbestände im OLAP-Bereich wachsen ständig
  $\hookrightarrow$ Überblick über Strukturen der Daten mittels explorativen Verfahren
- Data Mining und Mustererkennung

Knowledge Discovery Prozess



[Han & Kamber 2006]

Business Intelligence

Business Intelligence ist die entscheidungsorientierte Sammlung, Aufbereitung und Darstellung geschäftsrelevanter Informationen.

[Schrödl 2006]

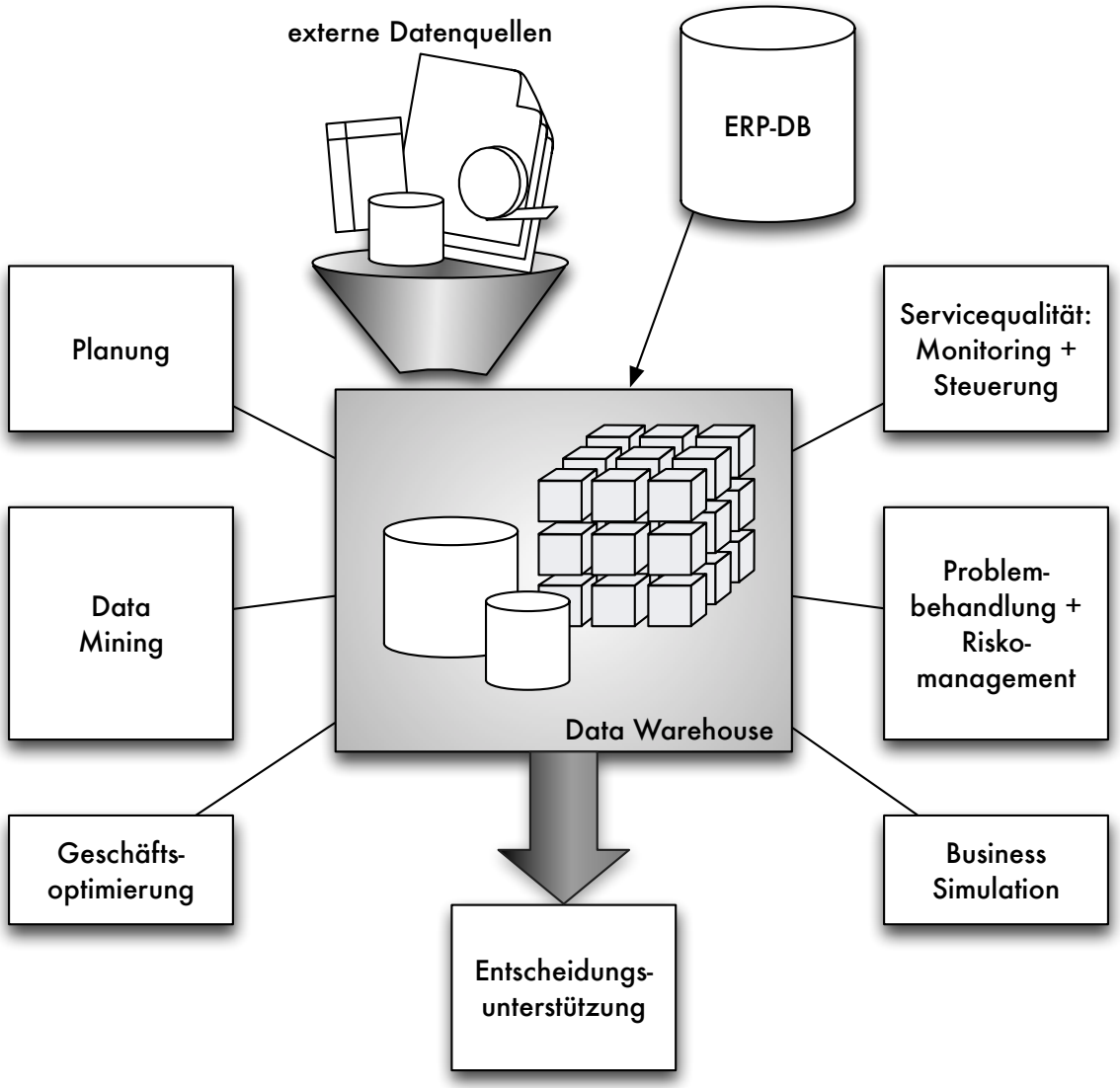
- Entscheidungsgrundlagen verbessern,
- Datensammlung: heterogene Quellen und Anforderungen (z.B. Sicherheit)
- Rohdaten zu Informationen transformieren (z.B. mathematisch, regelbasiert)
- Informationsdarstellung für Anwender
- Konzentration auf Geschäftsrelevanz (Optimierung Nutzen & Aufwand)

BI-Zyklus

1. Quantifizieren und Qualifizieren von Unternehmensinformationen
2. Analyse der gewonnen Daten
3. Ableiten von Erkenntnissen, welche die geschäftlichen Vorgänge unterstützen
4. Bewerten der Erkenntnisse in Bezug auf die Ziele
5. Umsetzen der relevanten Erkenntnisse in konkrete Maßnahmen

[Vitt et al. 2002]

Business Intelligence



Typische DW Anwendungsfälle

- Welche Kunden haben wir?
  ↪ Customer Relationship Management
- Wie entwickeln sich unsere Kosten?
  ↪ Supply Chain Management
- Wo existieren in unserem Produktsortiment weitere Potentiale?
  ↪ Warenkorbanalyse
- ...

Typische Data Mining Verfahren

- Assoziationsregeln – Was wurde gemeinsam in einem Warenkorb gekauft?
- Klassifikationsverfahren – Welchen Kundengruppen sollen wir Aktionen vorschlagen?
- Clustering – Welche Gemeinsamkeiten gibt es bei unseren Kunden / Lieferanten?
- . . .

Warenkorbanalyse

- Transaktionen an der Kasse (Transaktionsdatenbank):
  - T1: {Müller-Thurgau, Riesling, Dornfelder}
  - T2: {Riesling, Erfurter Bock, Ilmenauer Pils, Anhaltinisch Flüssig}
  - T3: {Müller-Thurgau, Riesling, Erfurter Bock }
- Warenkorbanalyse: Welche Waren werden häufig miteinander gekauft?
- Ziele:
  - Optimierung Laden-Layout
  - Cross-Marketing
  - Add-On Sales

Assoziationsregel

- Regeltyp:
  Rumpf $\rightarrow$ Kopf [support, confidence]
- Beispiel:
  - kauft(X, „Rotwein“) $\rightarrow$ kauft(X, „Erfurter Bock“) [0.5%, 60%]
  - 98% aller Kunden, die Müller-Thurgau und Riesling kaufen, bezahlen mit Kreditkarte.

Grundbegriffe

nach [Agrawal und Srikant (1994)]

- Items $I = \{i_1, i_2, \ldots, i_m\}$ – Grundgesamtheit an Literalen
- Itemset X: $X \subseteq I$
- Datenbank D – Menge von Transaktionen $X \subseteq I$
- $X \subseteq T$
- Lexikografische Sortierung in T und X
- Länge k eines Itemsets: Anzahl der Elemente
- k-Itemset: Itemset der Länge k

Grundbegriffe (2)

- Support der Menge X in D: Anteil der Transaktionen in D, die X enthalten:
  $supp(X) = \frac{|X|}{|D|}$
- Assoziationsregel: $A \rightarrow B$, mit $A \subseteq I$, $B \subseteq I$ und $A \cap B = \emptyset$
- Support s einer Assoziationsregel $A \rightarrow B$ in D: $s = supp(X \cup Y)$
- Konfidenz c einer Assoziationsregel $A \rightarrow B$ in D: Anteil der Transaktionen, die B enthalten, wenn sie in A enthalten sind – $c = conf(B|A) = \frac{supp(A \cup B)}{supp(A)}$

Problem: Bestimme alle Assoziationsregeln, die in D einen Support $\geq$ minsup und einen Konfidenz $\geq$ minconf besitzen.

Beispiel Assoziationsregeln

minsup = 20 %

TID	Items
1	Erfurter Bock, MT, Riesling
2	Erfurter Bock, MT, Dornfelder
3	Ilmenauer Pils, MT
4	Anhaltinisch Flüssig, Dornfelder, Riesling
5	Berliner Bräu, Dornfelder, Riesling
6	Kölnische Weiße, MT
7	Anhaltinisch Flüssig, Dornfelder

- $supp(\textit{MT}) \approx 57\%$
- $supp(\textit{Riesling}) = supp(\textit{Dornfelder}) \approx 43\%$
- $supp(\textit{Erfurter Bock}) = supp(\textit{Anhaltinisch Flüssig}) \approx 29\%$
- $supp(\textit{Ilmenauer Pils}) = supp(\textit{Berliner Bräu}) = supp(\textit{Köln. Weiße}) \approx 14\%.$
- potentielle Kandidaten: MT, Riesling, Dornfelder, Erfurter Bock, Anhaltinisch Flüssig

Beispiel Assoziationsregeln (2)

- mögliche Kombinationen aller Kandidaten:

Itemset	Support in %
(Erfurter Bock, MT)	≈ 29
(Erfurter Bock, Riesling)	≈ 14
(Erfurter Bock, Dornfelder)	≈ 14
(Erfurter Bock, Anhaltinisch Flüssig)	0
(MT, Riesling)	≈ 14
(MT, Dornfelder)	≈ 14
(MT, Anhaltinisch Flüssig)	0
(Riesling, Dornfelder)	≈ 29
(Riesling, Anhaltinisch Flüssig)	0
(Dornfelder, Anhaltinisch Flüssig)	≈ 29

Apriori Algorithmus

Input I, D, minsup
Output $\bigcup_k L_k$
C_k: zu zählende Kandidaten-Itemsets der Länge k
L_k: Menge aller häufig vorkommenden Itemsets
    der Länge k
initialisiere L_1:= 1-Itemsets aus I, k:= 2
WHILE $L_{k-1} \neq \emptyset$
    C_k := AprioriKandidatenGenerierung(L_{k-1});
    FOR EACH Transaktion $T \in D$
        CT := Subset(C_k, T)
        // alle Kandidaten aus C_k, die in T enthalten
        FOR jeden Kandidat $c \in CT$ c.count++
    $L_k := \{c \in C_k | (c.count/|D|) \geq minsup\}$
    k++

Effizienzsteigerung Apriori Algorithmus

- Zählen des Supports mittels Hashtabelle
  - [Park, Chen, Yu 1995]
  - Hashtabelle statt Hashbaum
  - k-Itemset, dessen Bucket einen Zähler kleiner den minimalen Support aufweist, kann nicht häufig auftreten
    effizienterer Zugriff auf Kandidaten, ungenauere Zählung
- Transaktionsreduktion
  - [Agrawal & Srikant 1994]
  - Transaktionen, die kein häufiges k-Itemset aufweisen, werden nicht benötigt, d.h. können entfernt werden
  - Datenbank-Scan effizienter, aber Schreibaufwand

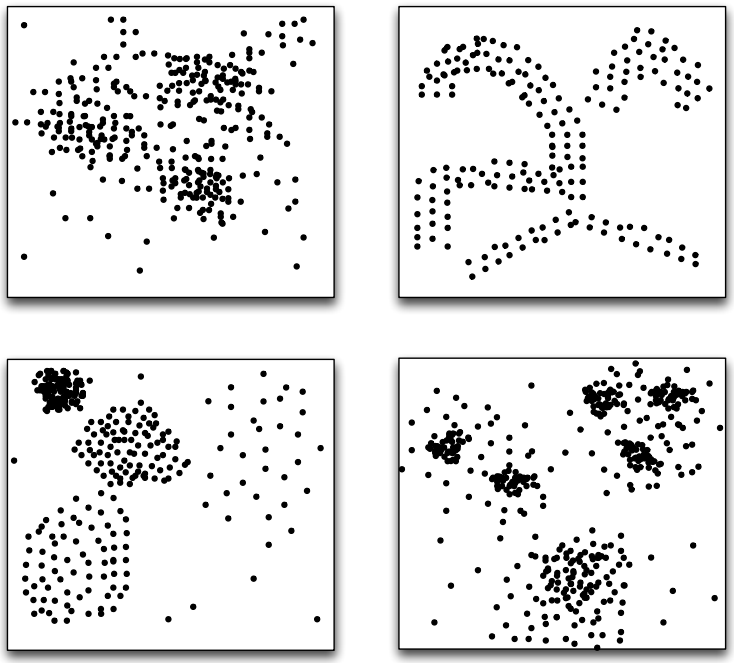
Effizienzsteigerung Apriori Algorithmus (2)

- Partitionierung
  - [Savasere, Omiecinski & Navathe 1995]
  - Itemset nur häufig, wenn es in einer Partition häufig ist
  - Ausnutzung des Hauptspeichers (Partition)
  - Partitionseffizient, aber Aufwand beim Zusammensetzen
- Sampling
  - [Toivonen 1996]
  - Anwendung Apriori auf Ausschnitt (Sample)
  - Zählen der gefundenen Regeln auf Gesamtdatenbank

Clusterverfahren

- Identifikation einer endlichen Menge von Gruppen in Daten → Suche nach Partitionierung
- Ähnlichkeit innerhalb Gruppe
- Möglichst Verschieden zwischen den Gruppen

Auftretende Muster (Größe, Form, Dichte):



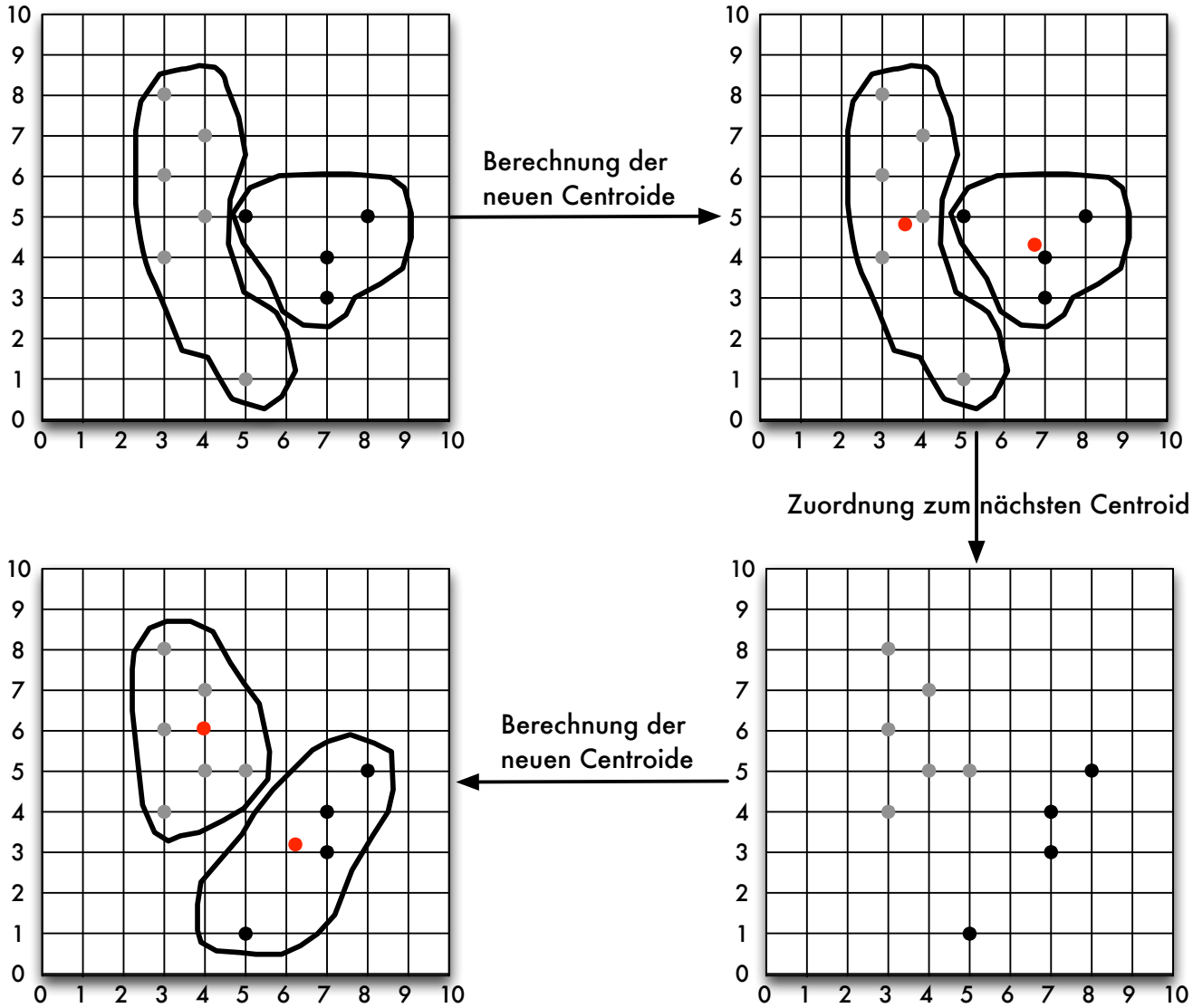
Distanzfunktionen

- Ähnlichkeitsmaß $sim(objekt_1, objekt_2)$
- Distanzfunktion $dist(objekt_1, objekt_2)$ $O \times O \rightarrow R_+$
  - kleine Distanz $\rightarrow$ ähnlich, große Distanz $\rightarrow$ unähnlich
  - $dist(objekt_1, objekt_2) = 0$, genau dann wenn $objekt_1 = objekt_2$
  - Symmetrie: $dist(objekt_1, objekt_2) = dist(objekt_2, objekt_1)$
  - Bei Metriken:
    $dist(objekt_1, objekt_3) \leq dist(objekt_1, objekt_2) + dist(objekt_2, objekt_3)$

Partitionierendes Clustering

ClusteringDurchVarianzMinimierung
Input: Tupelmenge D, Klassenanzahl k
Output: Cluster C
Erzeuge eine Anfangs-Zerlegung von D in k Klassen
Berechne Menge $C^* = \{C_1, \ldots, C_k\}$ der
    Centroide für die k Klassen
$C := \{\}$
repeat
    $C := C^*$
    Partioniere: Bilde k Klassen durch Zuordnung
    jedes Punktes zum nächstliegenden Centroid aus C
    Berechne Centroide: Berechne die Menge
    $C^* = \{C_1^*, \ldots, C_k^*\}$ der Centroide
    für die neu bestimmten Klassen
until C = C^*

Clusterverfahren: Illustration



Vor- und Nachteile

Vorteile:

- linearer Aufwand pro Iteration, wenige Iterationen
- einfache Implementierung
- k-means [MacQueen 1967]: populärster Clusteralgorithmus

Nachteile:

- Rauschen- und Ausreißeranfällig
- konvexe Form der Cluster
- Bestimmung Anzahl der Cluster
- Initialaufteilung wichtig für Laufzeit und Ergebnis

Klassifikation: Beispiel

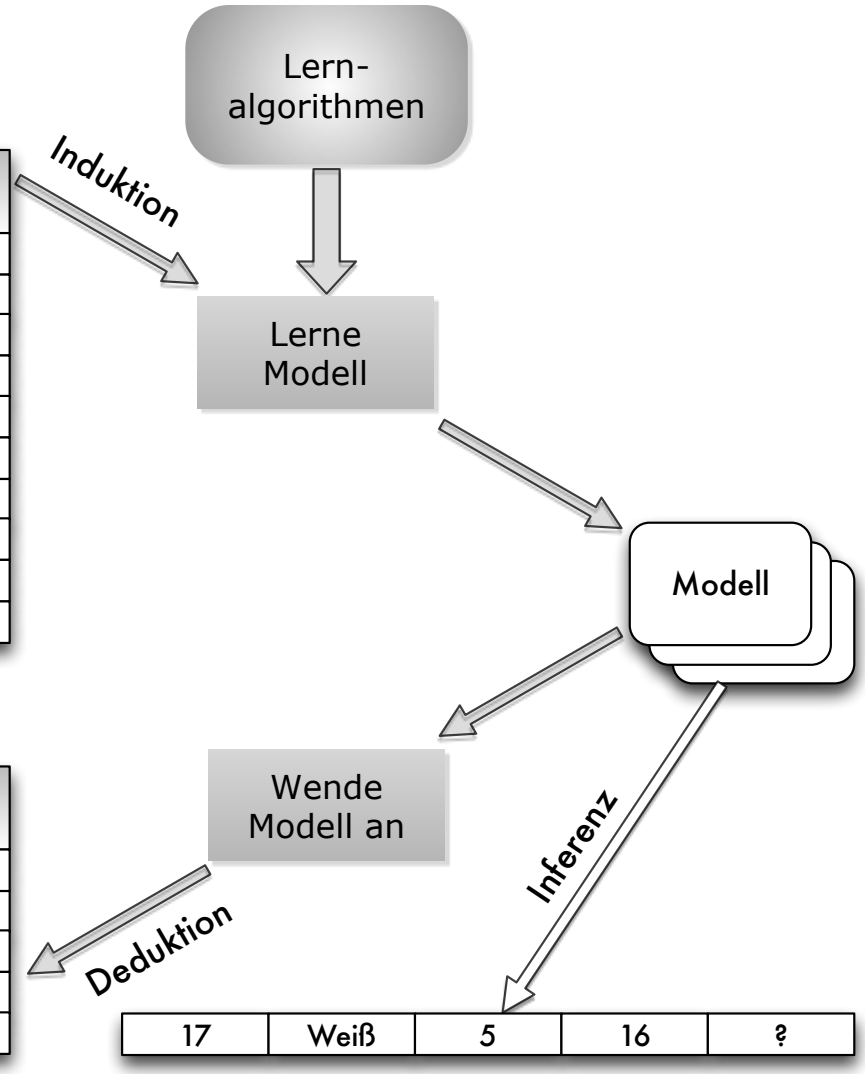
Schmeckt uns der Wein?

TID	Weinart	Restsüße g/l	Alkohol-gehalt	Class
1	Weiß	18	10	Yes
2	Rot	20	9	Yes
3	Rose	22	9	No
4	Rose	15	8	No
5	Rot	30	5	Yes
6	Weiß	18	10	Yes
7	Rot	15	15	No
8	Weiß	45	5	Yes
9	Weiß	18	14	Yes
10	Rot	8	10	No

Trainings Set

TID	Weinart	Restsüße g/l	Alkohol-gehalt	Class
11	Rot	23	10	?
12	Rose	15	12	?
13	Weiß	22	10	?
14	Weiß	30	6	?
15	Rot	12	14	?

Test Set

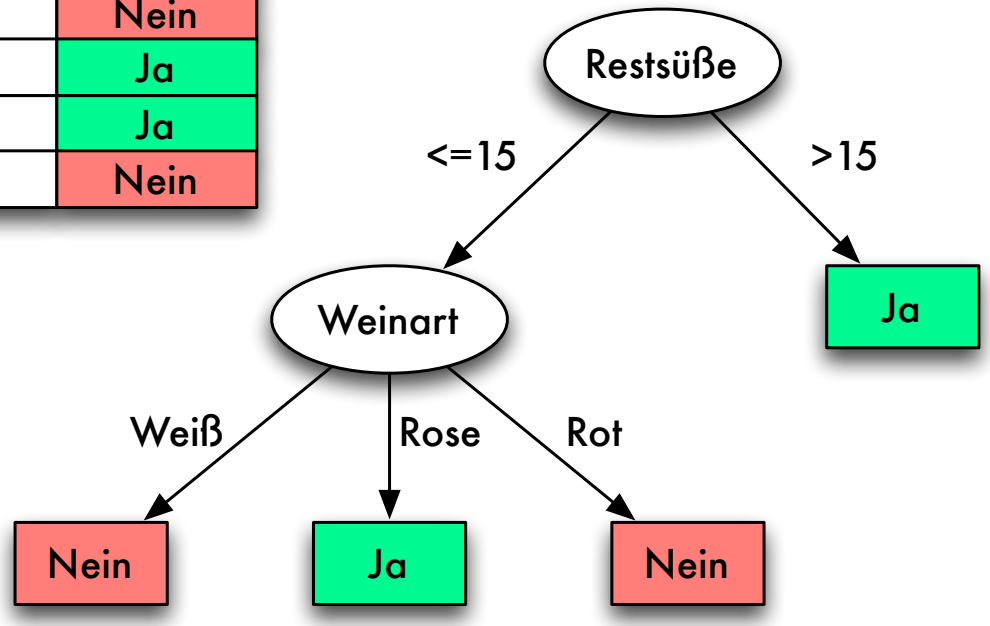


Klassifikation

- Gegeben sind Menge von Objekten mit Attributen $o = (x_1, \ldots, x_d)$ und Zugehörigkeit zu Klassenmenge C
- Gesucht: Klassifikator K für neue Objekte $\rightarrow K : Objekte_{neu} \rightarrow C$
- Klassenzugehörigkeit a-priori bekannt $\rightarrow$ Abgrenzung zu Clusterverfahren
- Ähnlich zu Prognose (z.B. lineare Regression)

Klassifikationsergebnis

TID	Weinart	Restsüße g/l	Alkohol-gehalt	Ja/Nein
1	Rot	23	12	Ja
2	Weiß	15	10	Nein
3	Rose	14	10	Ja
4	Weiß	30	6	Ja
5	Rot	12	14	Nein



Klassifikationsgüte

		Vorhersage	
		Klasse zugehörig	Klasse nicht zugehörig
wahre Werte	Klasse zugehörig	True Positive	False Negative
	Klasse nicht zugehörig	False Positive	True Negative

- Accuracy: $\frac{TP+TN}{TP+FN+FP+TN}$
- Precision: $p = \frac{TP}{TP+FP}$
- Recall: $r = \frac{TP}{TP+FN}$
- F-Measure: $F = \frac{2 \cdot TP}{2 \cdot TP+FN+FP}$

Klassifikationsmethoden

- Entscheidungsbaum
- Regelbasiert
- Lineare Diskriminanz nach Fisher
- Kategorielle Regression, Log-Lineare Modelle
- Neuronale Netzwerke
- Naive Bayes und Bayesian Belief Networks
- Support-Vektor-Maschinen

Entscheidungsbaum

- Vorgehen: Splitting und Partitionieren
- Explizites Wissen wird gefunden
- Leicht verständlich
- Gut visualisierbar

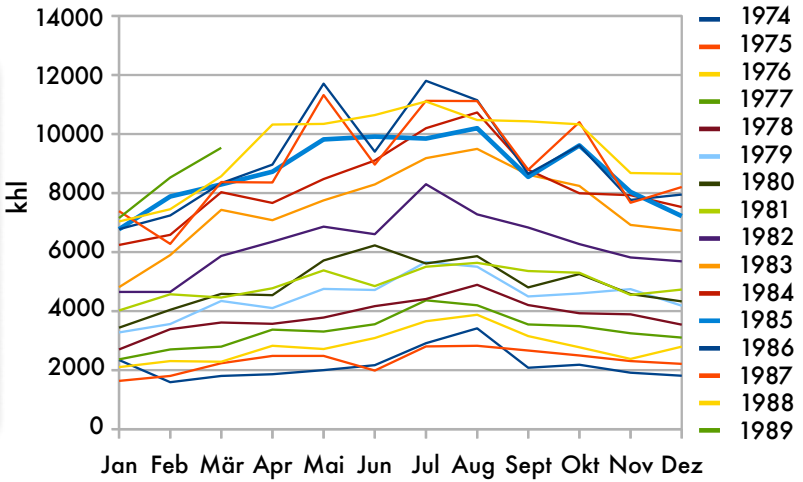
Algorithmus für den Entscheidungsbaum

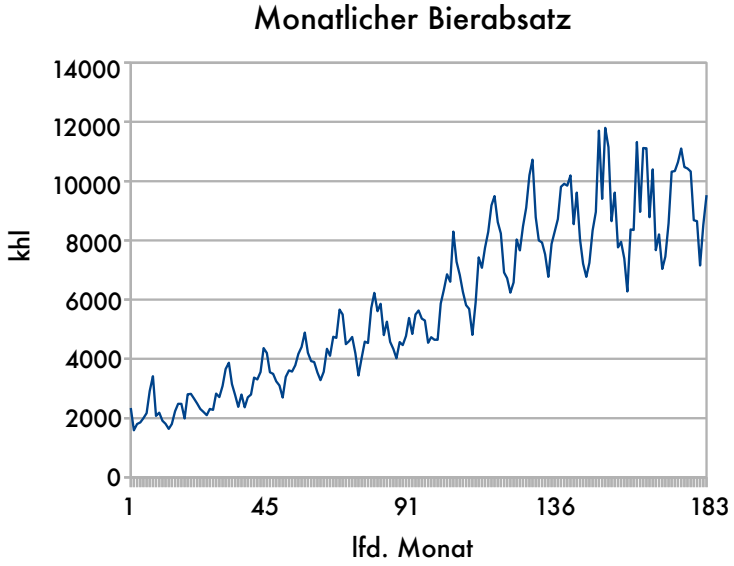
```
Input: Trainingsdatensätze
Initialisierung: alle Datensätze gehören
    zum Wurzelknoten
WHILE Splitattribut vorhanden OR Datensätze
eines Knoten in unterschiedlichen Klassen
    Wähle Splitattribut (Splittingstrategie)
    Partitioniere Datensätze des Knoten
    anhand Attribut
    Rekursion für alle Partitionen
```

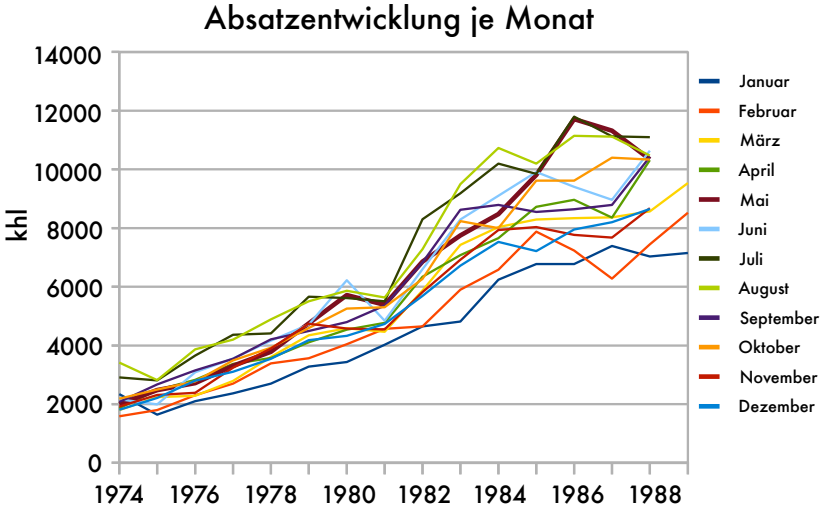
Prognose: Beispiel

Monatlicher Tankbierabsatz einer Brauerei (khl)

1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
2339	1638	2101	2363	2697	3279	3438	4021	4646	4811	6236	6770	6771	7386	7034	7150
1588	1798	2307	2700	3388	3561	4044	4570	4646	5896	6582	7881	7237	6279	7449	8525
1800	2235	2281	2794	3609	4343	4584	4461	5868	7426	8029	8290	8335	8370	8569	9530
1858	2481	2827	3371	3570	4103	4536	4771	6346	7076	7661	8720	8966	8356	10320	
2001	2479	2713	3303	3783	4749	5711	5383	6857	7749	8471	9813	11709	11318	10340	
2169	1988	3083	3555	4163	4711	6225	4843	6602	8293	9103	9913	9402	8964	10641	
2911	2804	3657	4364	4405	5661	5609	5504	8295	9183	10198	9847	11799	11119	11100	
3414	2820	3872	4198	4890	5503	5860	5633	7278	9496	10725	10196	11147	11113	10474	
2077	2666	3149	3547	4206	4494	4800	5360	6829	8620	8785	8546	8645	8783	10427	
2184	2494	2773	3491	3923	4595	5256	5297	6269	8237	7994	9613	9615	10397	10329	
1913	2308	2382	3246	3893	4740	4576	4546	5814	6919	7929	8038	7765	7672	8677	
1809	2212	2798	3102	3543	4179	4330	4733	5686	6721	7527	7217	7948	8202	8651	







Zeitreihenmodelle

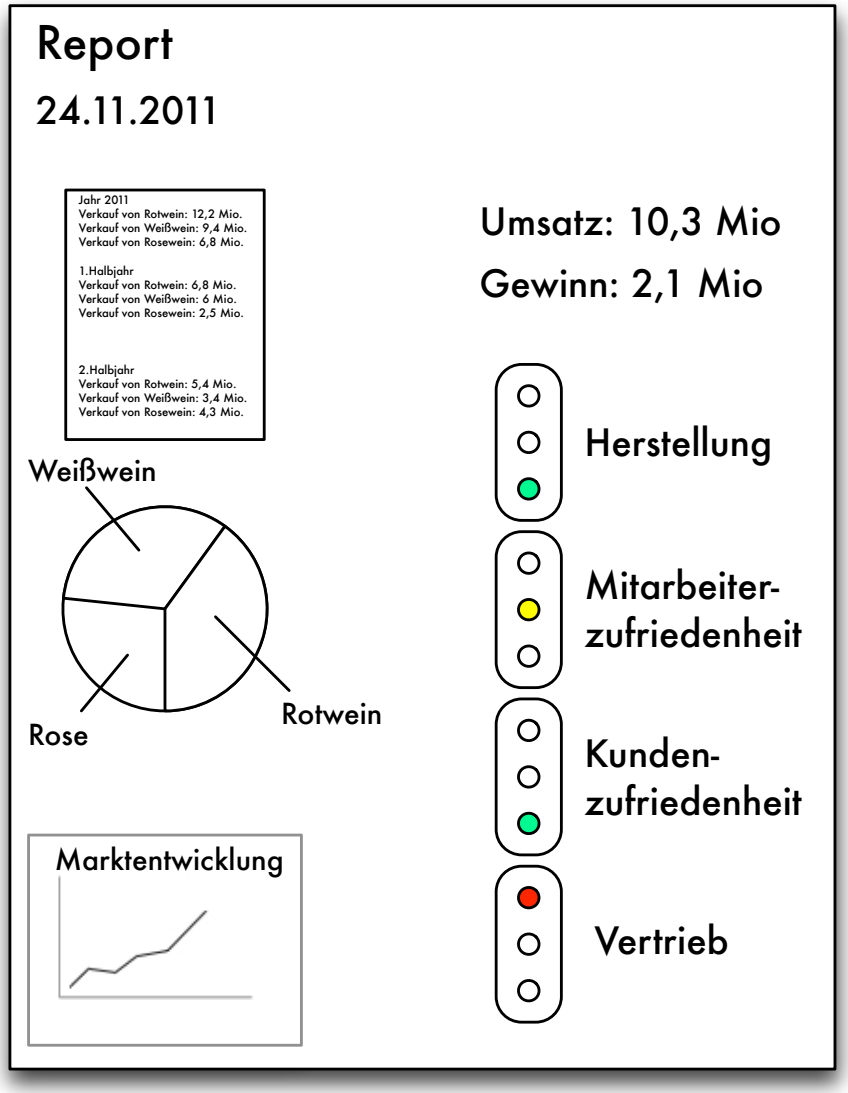
- Additiv: $X_t = G_t + S_t + e_t$
- Multiplikativ: $X_t = G_t \cdot S_t \cdot e_t$
- Gemischt: $X_t = G_t \cdot S_t + e_t$

Komponenten:

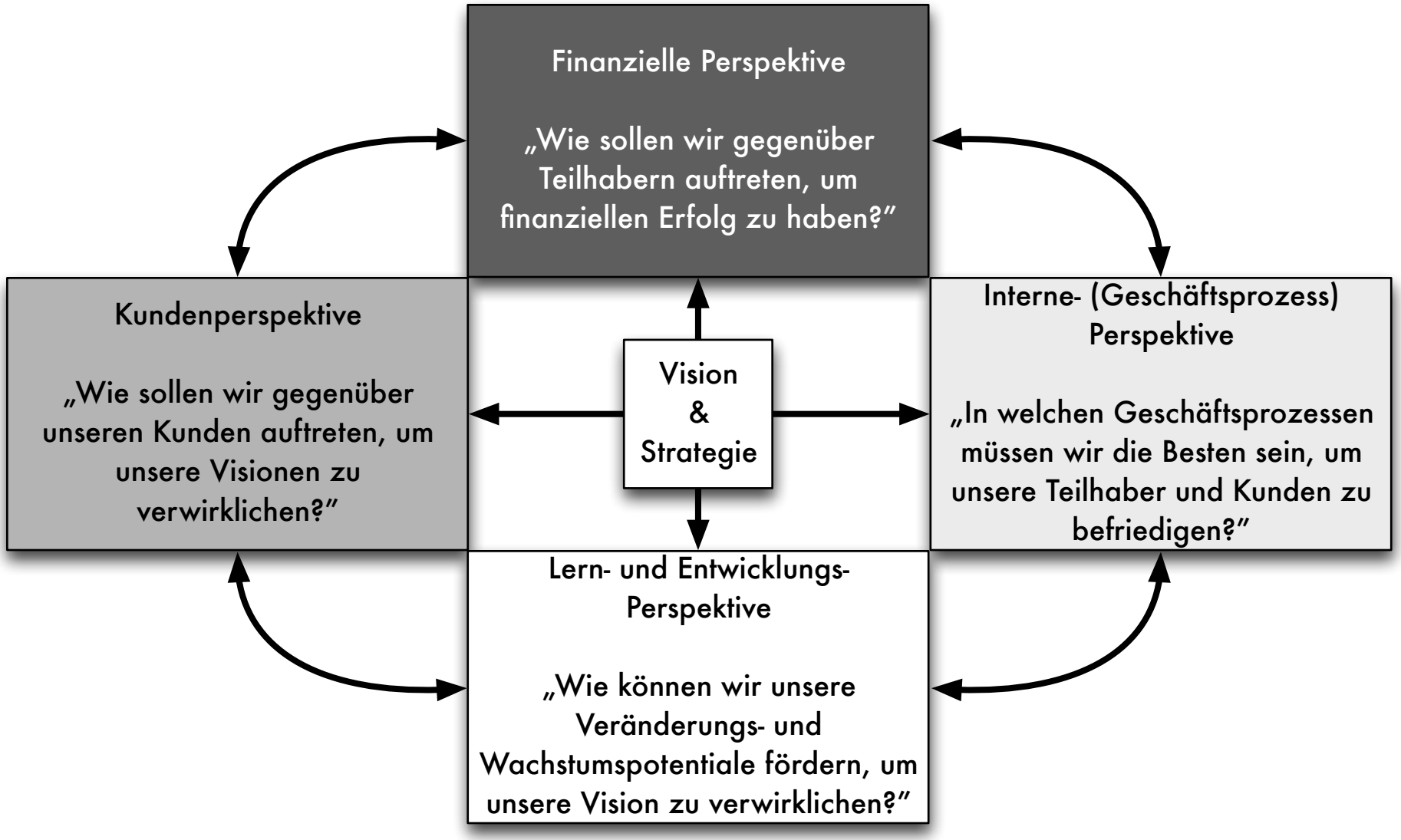
- Konstant: $dX_t/dS_t = 1$
- Niveauabhängig: $dX_t/dS_t = G_t$

- X_t: Ausprägung zum Zeitpunkt t
- G_t: Trend, Wachstum
- S_t: Saison, Konjunktur, Zyklen
- e_t: Fehlerterm

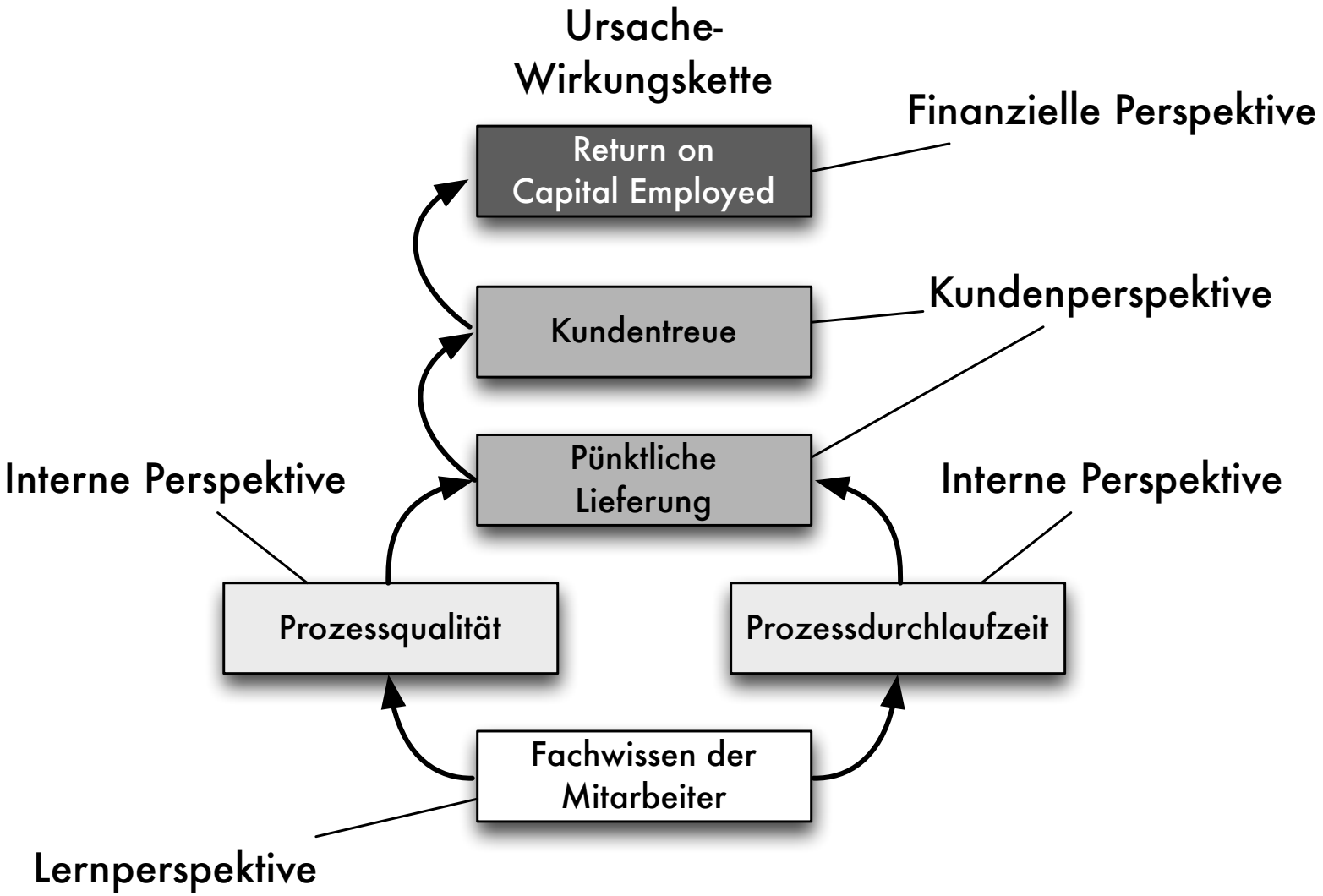
Reporting


Report
24.11.2011
Jahr 2011
Verkauf von Rotwein: 12,2 Mio.
Verkauf von Weißwein: 9,4 Mio.
Verkauf von Rosewein: 6,8 Mio.
1.Halbjahr
Verkauf von Rotwein: 6,8 Mio.
Verkauf von Weißwein: 6 Mio.
Verkauf von Rosewein: 2,5 Mio.
2.Halbjahr
Verkauf von Rotwein: 5,4 Mio.
Verkauf von Weißwein: 3,4 Mio.
Verkauf von Rosewein: 4,3 Mio.
Weißwein
Rose
Rotwein
Marktentwicklung
Umsatz: 10,3 Mio
Gewinn: 2,1 Mio
Herstellung
Mitarbeiter-
zufriedenheit
Kunden-
zufriedenheit
Vertrieb

Balanced Scorecard



Wirkungszusammenhänge



Entscheidungsunterstützung