Datenanalyse mit Data Mining

- von Jan-Christoph Meier

- Hamburg, 19.01.2012

Ablauf

- Motivation
- Speicherung der Daten für das Data Mining
- Data Mining Algorithmen
- Ausblick auf die Masterarbeit
- Konferenzen und Papers

Ablauf

- **Motivation** 
- Speicherung der Daten für das Data Mining
- Data Mining Algorithmen
- Ausblick auf die Masterarbeit
- Konferenzen und Papers

Was ist Data Mining?

- Extrahieren von Wissen aus **großen Mengen** von Daten
- Alternativ auch: “**Knowledge Discovery from Data**”

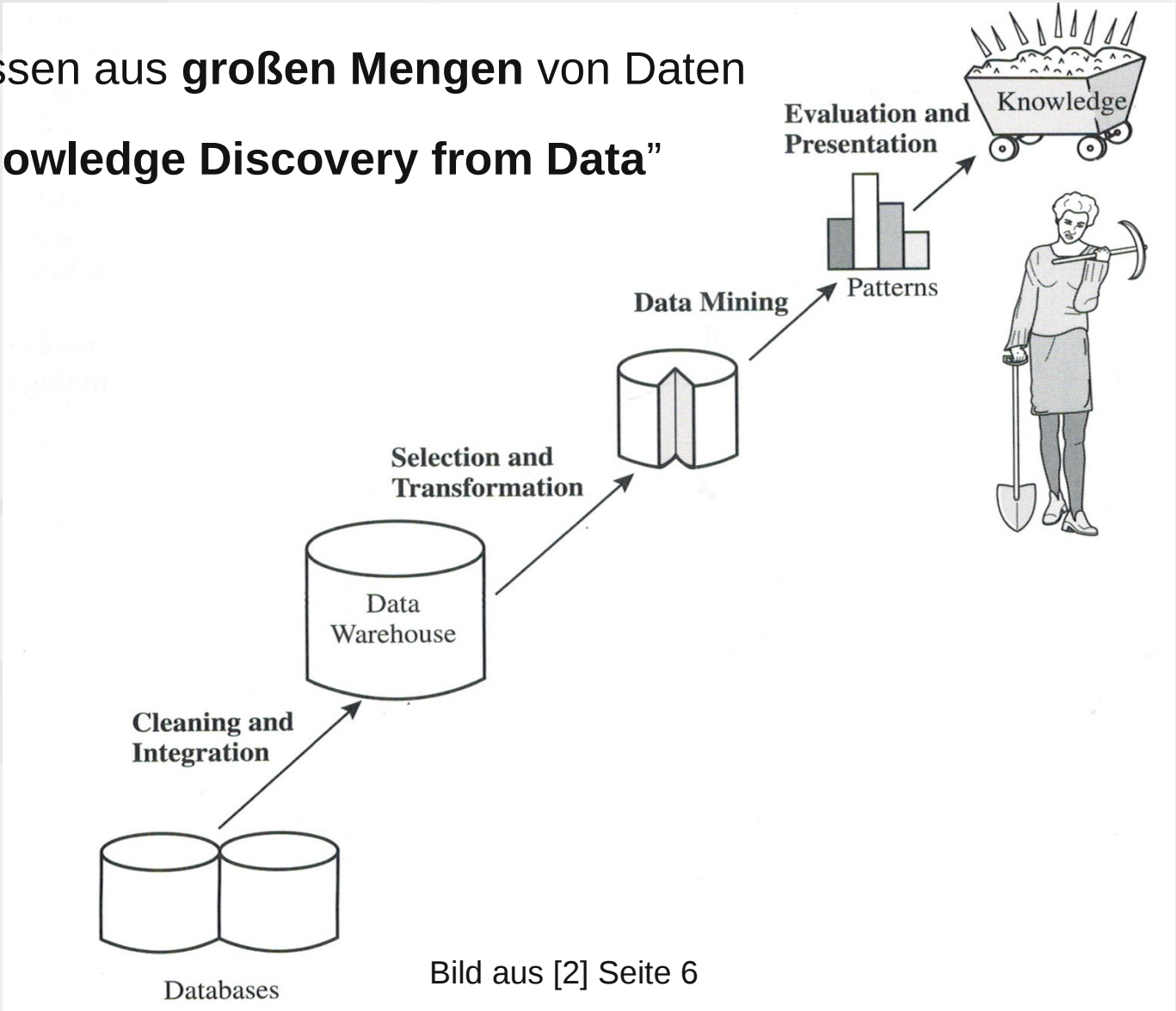


Bild aus [2] Seite 6

Was für Daten werden generiert?

- Es werden permanent **große** Datenmengen generiert
  - In **Webshops** (X hat Y gekauft)
  - In **medizinischen Datenbanken** (X hat das Medikament Y bekommen, X hat Symptom Z)
  - In **sozialen Netzwerken** (X ist mit Y befreundet)
  - In **Suchmaschinen** (X hat nach Y gesucht)

Welches Wissen kann extrahiert werden?

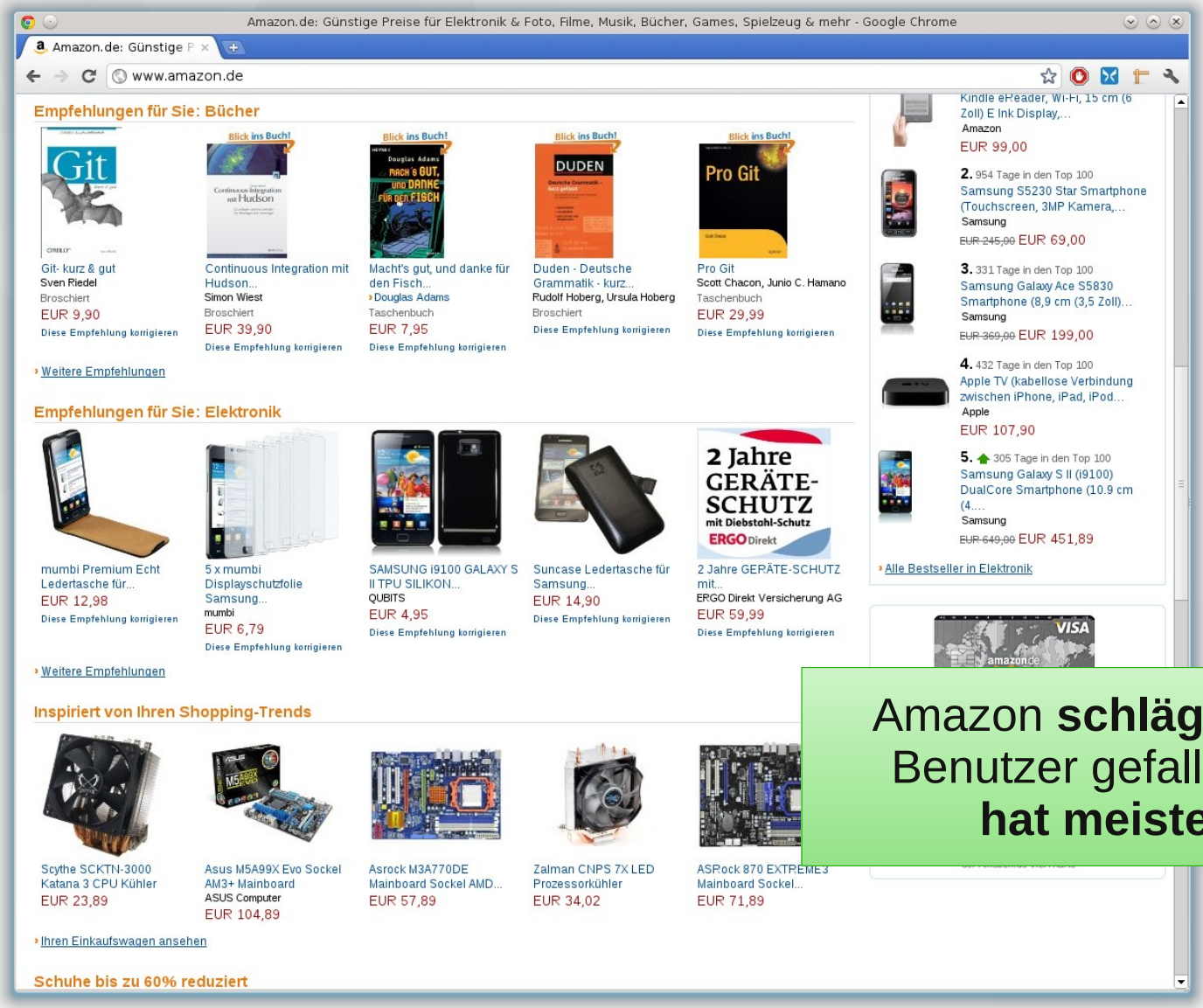
- Korrelationen in den Daten herstellen
  - In Webshops (X hat Y gekauft)

    Kunde, der ein Mainboard gekauft hat, hat auch eine CPU gekauft
  - in medizinischen Datenbanken (X hat das Medikament Y bekommen, X hat Symptom Z)

    Patienten, die Medikament X bekommen, haben Sympthom Z
  - In Suchmaschinen (X hat nach Y gesucht)

    Benutzer meint mit Schloss nicht das Vorhängeschloss, sondern das Schloss von Ludwig dem 14.

Wofür kann das Wissen verwendet werden?



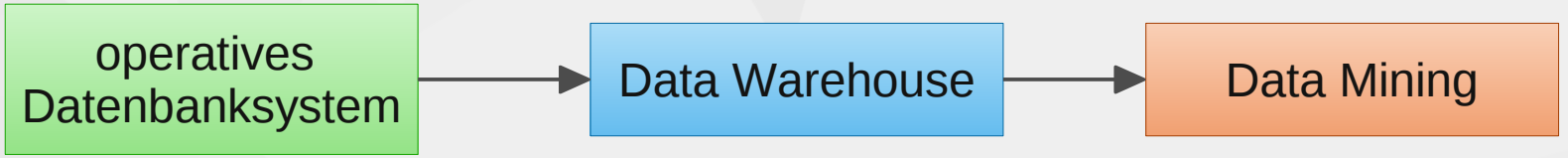
Amazon **schlägt vor,** was dem Benutzer gefallen könnte und **hat meistens Recht!**

Ablauf

- Motivation
- **Speicherung der Daten für das Data Mining** ←
- Data Mining Algorithmen
- Ausblick auf die Masterarbeit
- Konferenzen und Papers
- Fazit

Data Warehouse

- Daten werden für die Analyse in ein Data Warehouse übertragen



- Die Datenbank des Data Warehouse dient rein zu Analyse, es findet nur lesender Zugriff statt.
- Verschiedene Datenquellen, nicht zwingend Datenbanken

Data Warehouse (2)

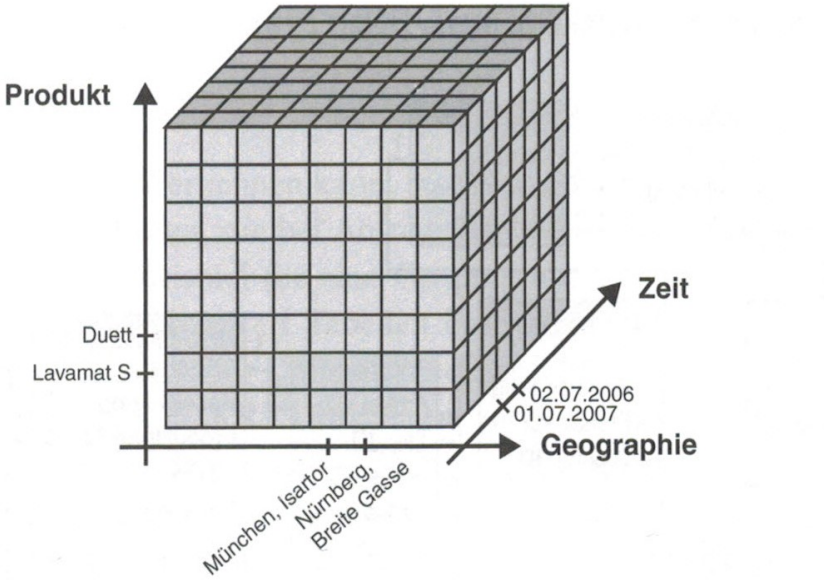
- Online Analytical Processing (OLAP) ist ein Oberbegriff für die Analyse multidimensionaler Daten
  - Wieviele Mainboards wurden im 1. Quartal 2011 in der K&M-Elektronik Filiale Berliner Tor verkauft?
  - Wieviele Patienten haben 2011 das Medikament M1 bekommen?

Beispiel:
Verkauf von verschiedenen Lebensmitteln in einem Supermarkt

	Measures		
Product	Unit Sales	Store Cost	Store Sales
−All Products	266.773	225.627,23	565.238,13
+Drink	24.597	19.477,23	48.836,21
−Food	191.940	163.270,72	409.035,59
+Baked Goods	7.870	6.564,09	16.455,43
+Baking Goods	20.245	15.370,61	38.670,41
+Breakfast Foods	3.317	2.756,80	6.941,46
+Canned Foods	19.026	15.894,53	39.774,34
+Canned Products	1.812	1.317,13	3.314,52
+Dairy	12.885	12.228,85	30.508,85

Data Warehouse (3)

- **Data Cube** – Abstraktionsmodell zur logischen Darstellung der Daten



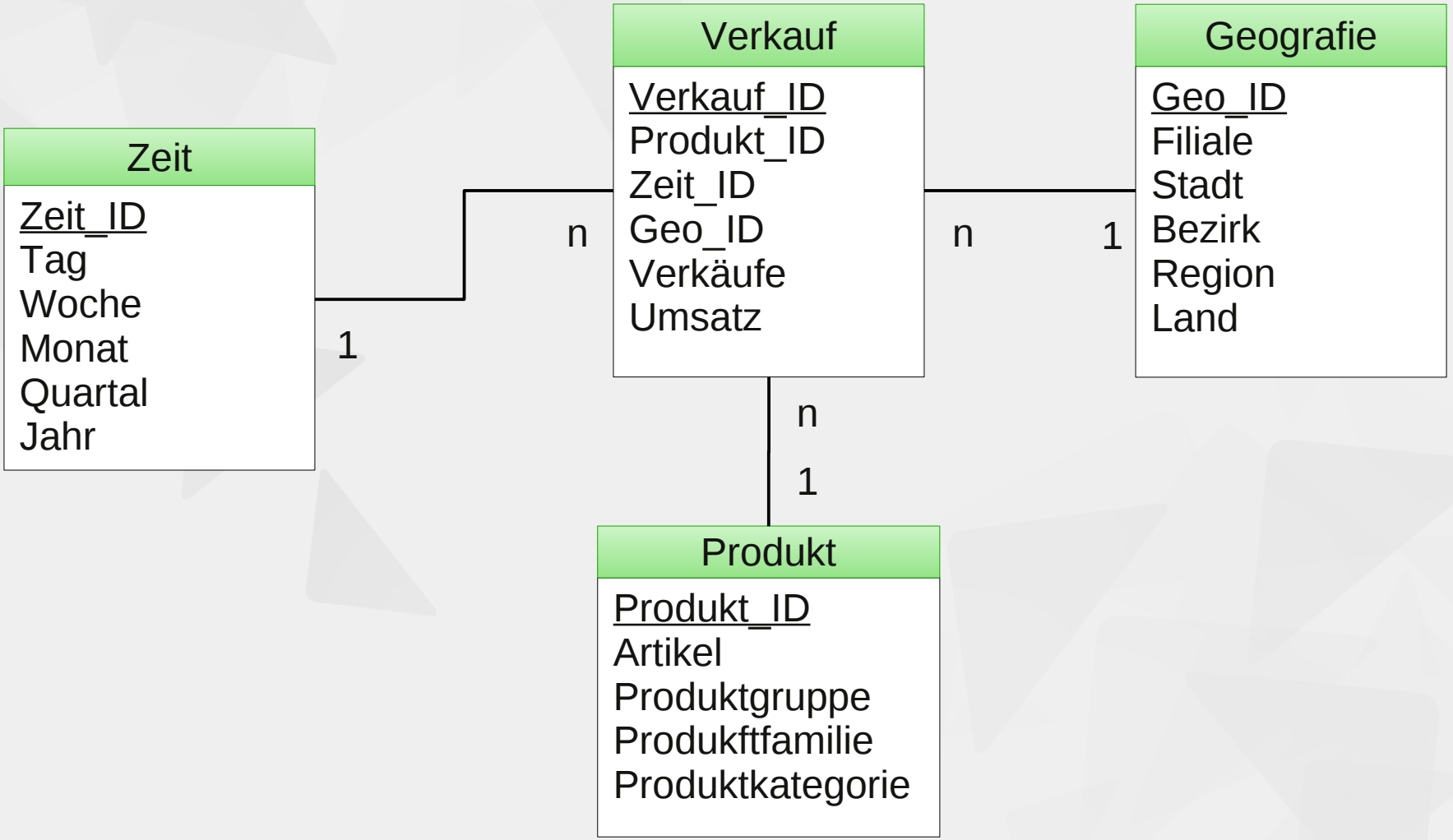
Artikel	Filiale	Tag	Verkäufe
Duett	Nürnberg, Breite Gasse	01.07.2007	7
Duett	München, Isartor	01.07.2007	3
Lavamat S	München, Isartor	01.07.2007	2

Bild aus [1] Seite 215

Beispiel:
Absatz von Waschmaschinen an verschiedenen Standorten.

Data Warehouse (3)

- Im Data Warehouse werden die Daten in einem **Star-Schema** gespeichert



Ablauf

- Motivation
- Speicherung der Daten für das Data Mining
- **Data Mining Algorithmen** 
- Ausblick auf die Masterarbeit
- Konferenzen und Papers
- Fazit

Support und Confidence

- **Berechnung der Korrelation zwischen verschiedenen Datensätzen.**
- Als Basis dienen *n* Transaktionen, die A und B enthalten können.

$$support\left(A\right)=P\left(A\right)$$

$$confidence\left(A\Rightarrow B\right)=P\left(B|A\right)=\frac{support\left(A\cup B\right)}{support\left(A\right)}$$

- **Support**: Anzahl der Transaktionen, die A enthalten.
- **Confidence**: Mit welcher Wahrscheinlichkeit sind A und B zusammen in einer Transaktion

Support und Confidence (2)

- Beispiel

Transaktionen	
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

$$support\left(I2\right)=\frac{7}{9}$$

$$confidence\left(I2\Rightarrow I1\right)=\frac{support\left(I2\cup I1\right)}{support\left(I2\right)}=\frac{\left(4/9\right)}{\left(7/9\right)}=\frac{4}{7}$$

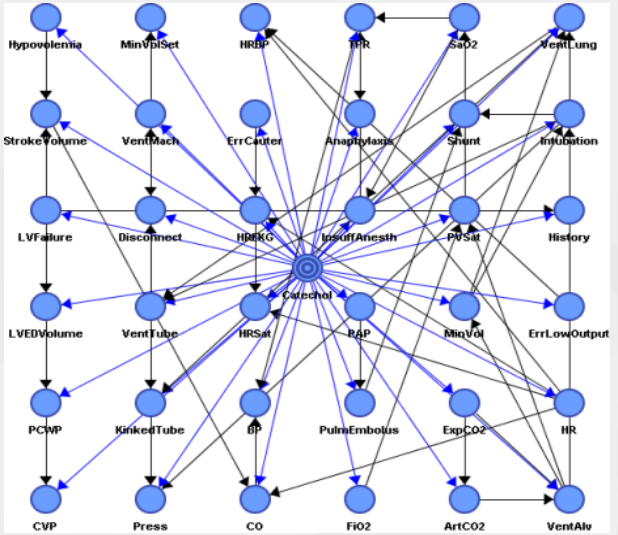
Finden von Korrelationen in Daten

- **Problem**

  Welche Elemente in sehr vielen Transaktionen haben eine hohe Korrelation?

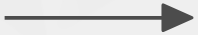
- **Lösung**

  Apriori Algorithmus, FP-Tree Algorithmus



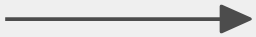
Apriori Algorithmus (1)

Transaktionen	
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3



Itemset	**Support Count**
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

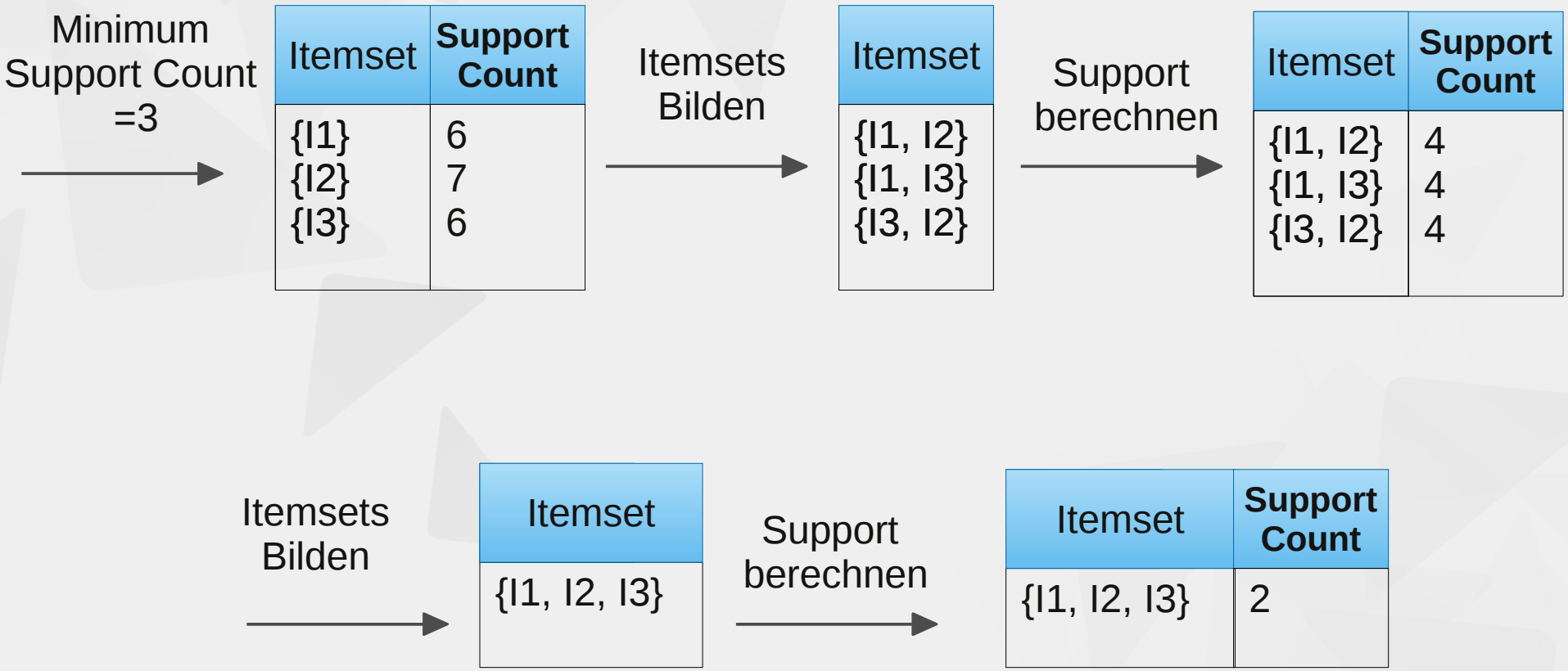
Minimum Support Count >=3



Itemset	**Support Count**
{I1}	6
{I2}	7
{I3}	6

Support Count: Gibt an, wie häufig das Element in allen Transaktionen vorkommt.

Apriori Algorithmus (2)



Apriori Algorithmus (3)

- **Probleme des Algorithmus**
  - Bei sehr vielen Elementen müssen sehr viele Kombinationen gebildet werden.
  - Bei sehr vielen Transaktionen muss häufig über diese iteriert werden, somit wird die Berechnung des Supports teuer.

FP-Tree Algorithmus

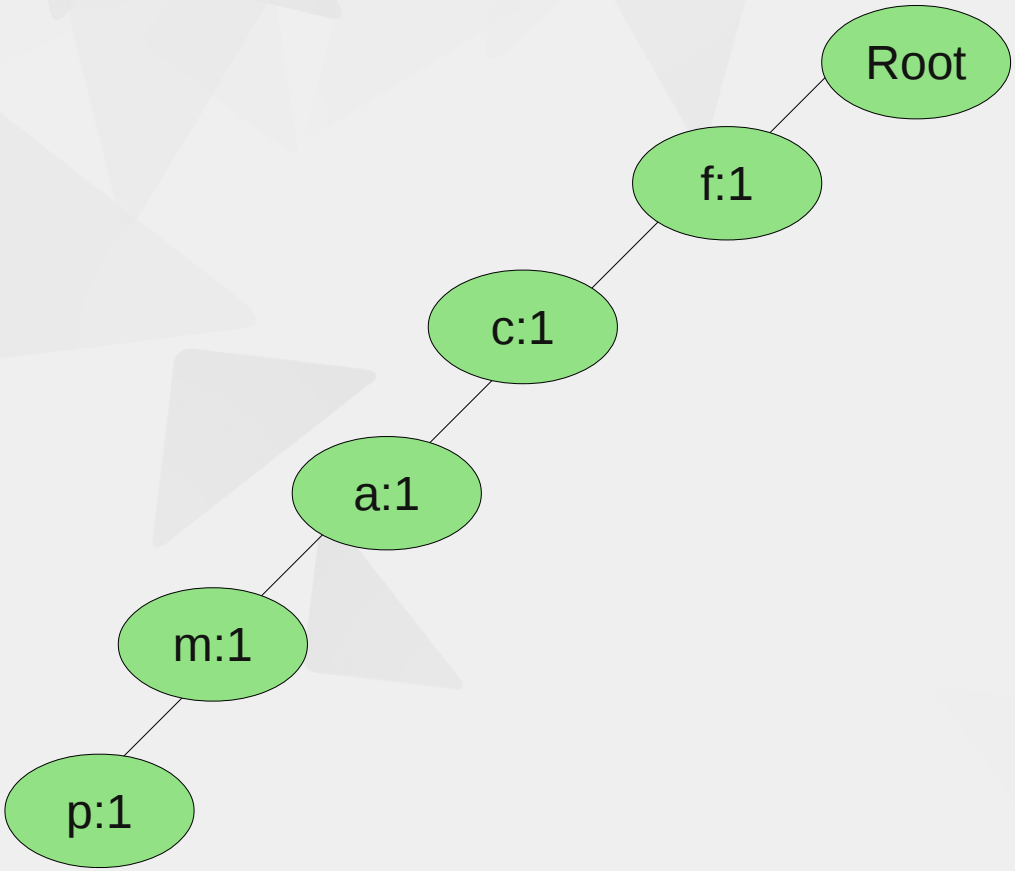
- FP = Frequent Pattern
- Verbesserung gegenüber Apriori zum Finden von Korrelationen in Daten.
- Eine Baumstruktur wird verwendet, um die Korrelationen zu bestimmen.

TID	Transaktionen	Elemente mit Support >= 3
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, l, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, k, s, p	c, b, p
5	a, f, c, e, l, p, m, n	f, c, a, m, p

Support Count: f:4 c:4 a:3 b:3 m:3 p:3

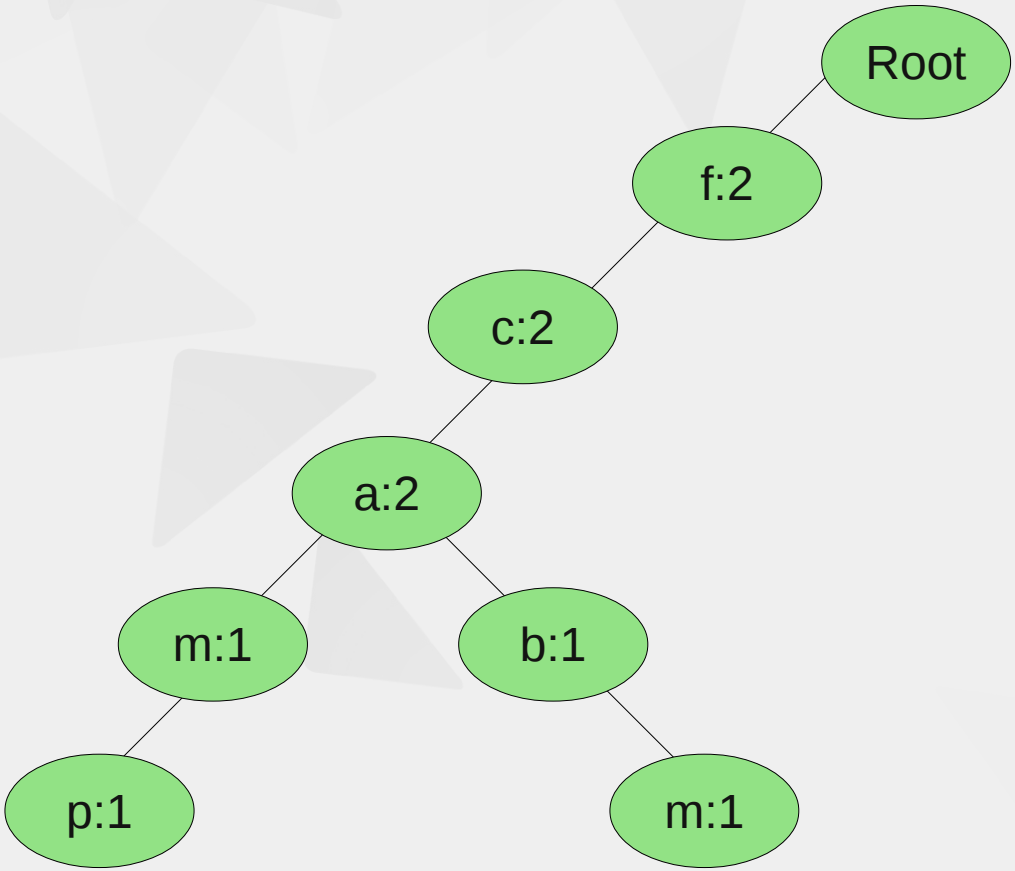
FP-Tree – Konstruktion des Baums

- Einfügen der 1. Transaktion: f, c, a, m, p



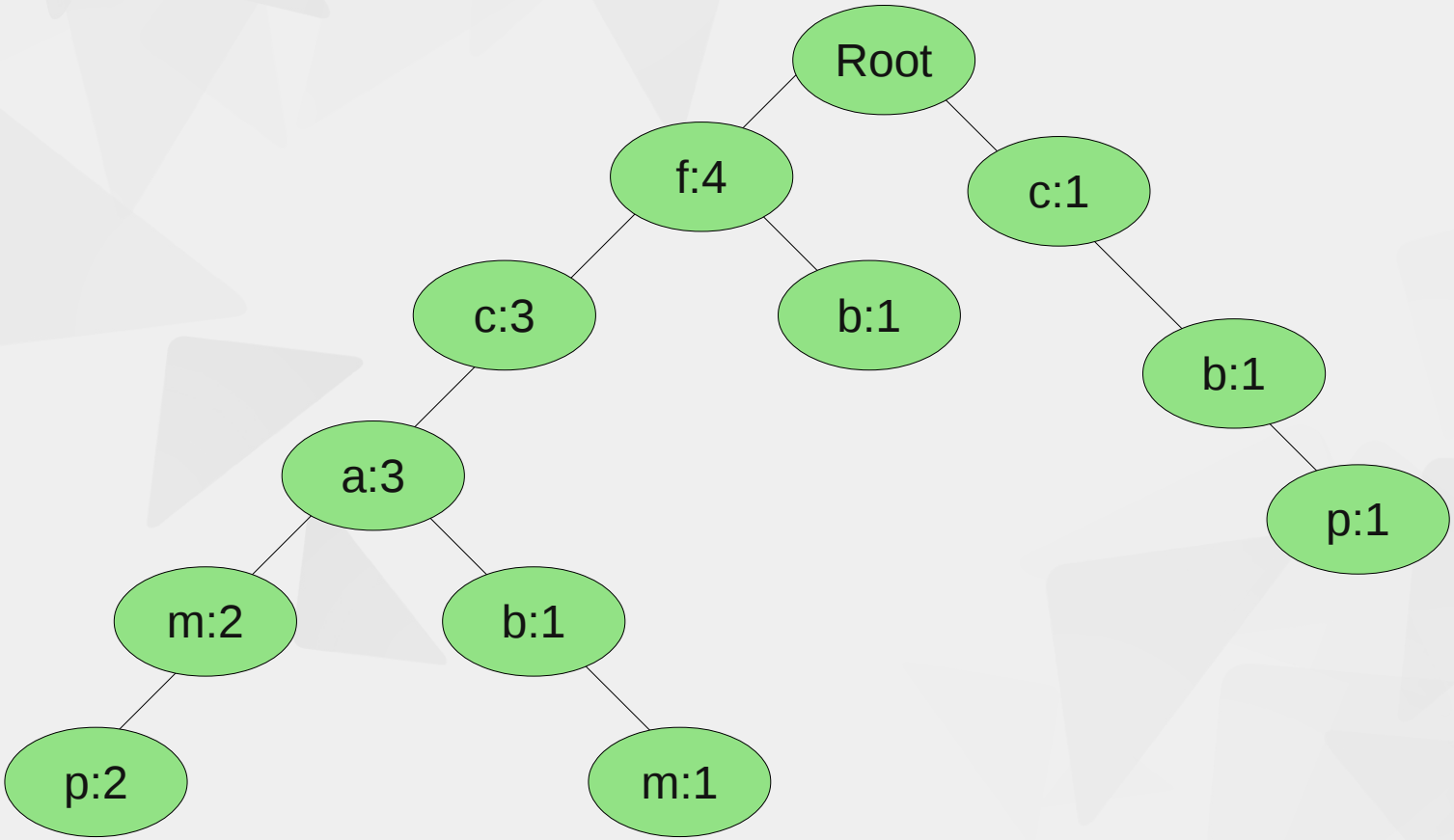
FP-Tree – Konstruktion des Baums

- Einfügen der 2. Transaktion: f, c, a, b, m



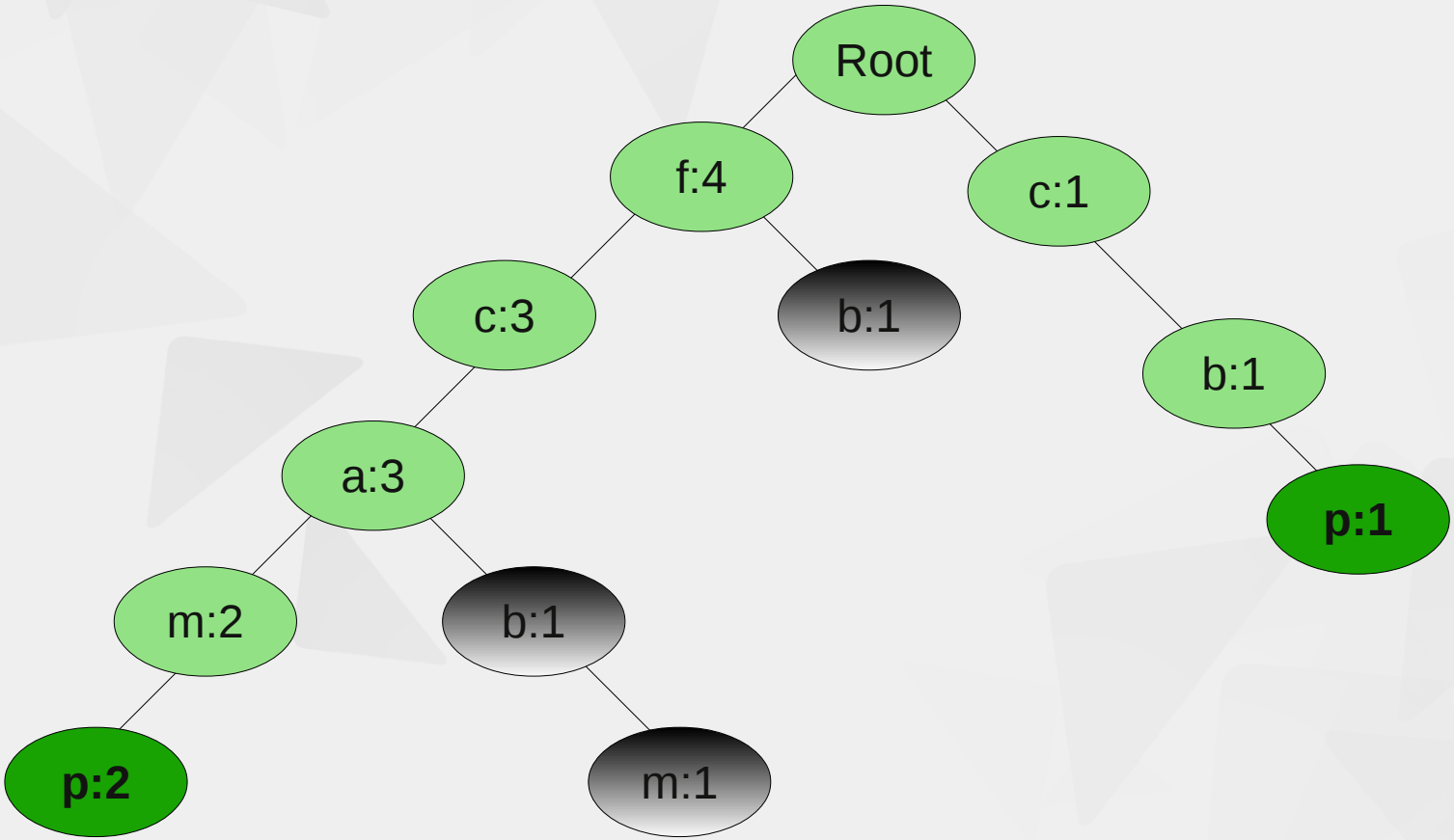
FP-Tree – Konstruktion des Baums

- Ergebnis – Alle Transaktionen wurden eingefügt



FP-Tree – Analyse des Baums

- In welchen Pfaden ist p enthalten?



FP-Tree Support berechnen

- Wir erhalten folgende Pfade für p:

  f:2, c:2, a:2, m:2 und **c:1, b:1**

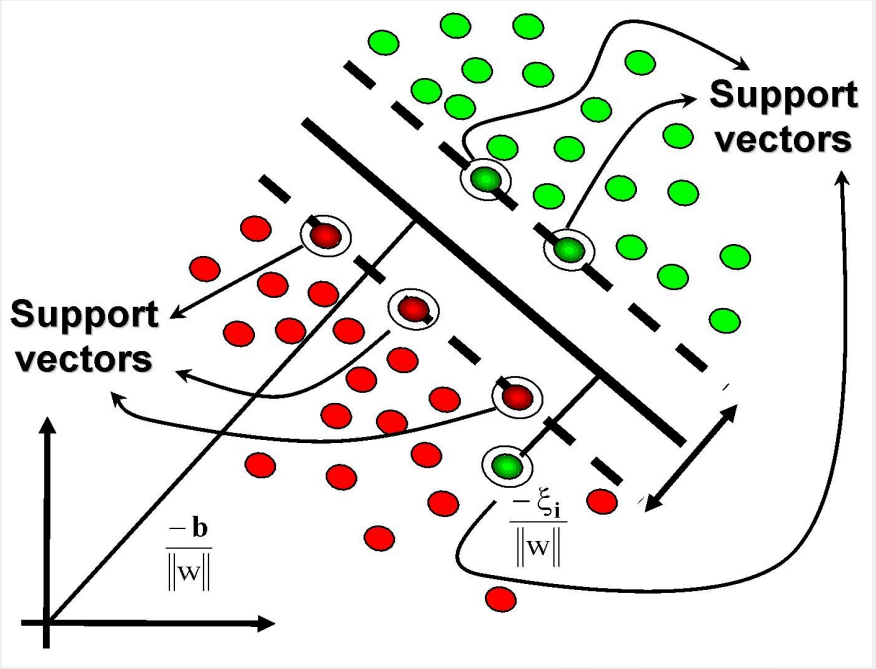
- Die Pfade werden als “Conditional pattern base” (CPB) bezeichnet

- Der minimale Support wird mit 3 festgelegt, somit erhalten wir

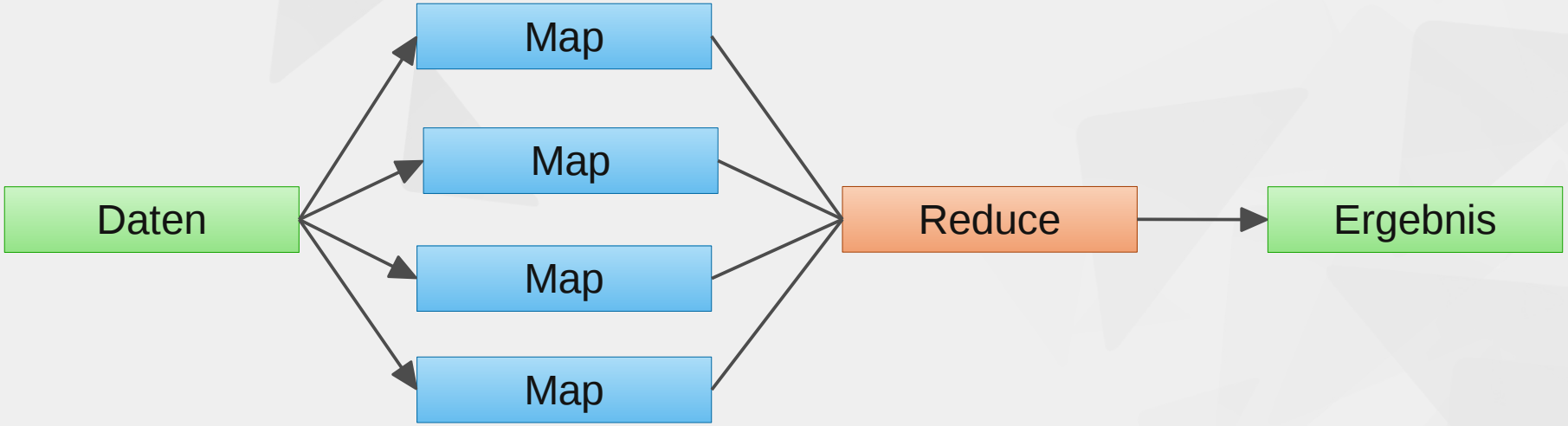
  support(cp) = 3

Weitere Ansätze

- Support Vector Machines



- MapReduce



Hadoop

- **Framework für die Entwicklung von MapReduce Algorithmen**
- Apache Mahout: Erweiterung für Hadoop mit Implementierungen von Data Mining Algorithmen und Machine Learning
- Hadoop wird unter anderem von Facebook und Yahoo eingesetzt.

Ablauf

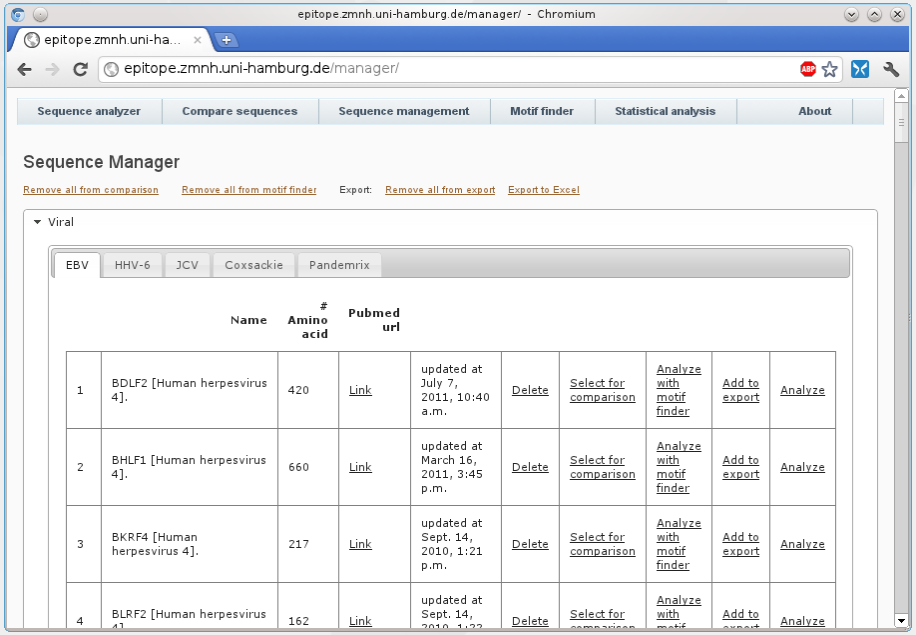
- Motivation
- Speicherung der Daten für das Data Mining
- Data Mining Algorithmen
- **Ausblick auf die Masterarbeit** 
- Konferenzen und Papers
- Fazit

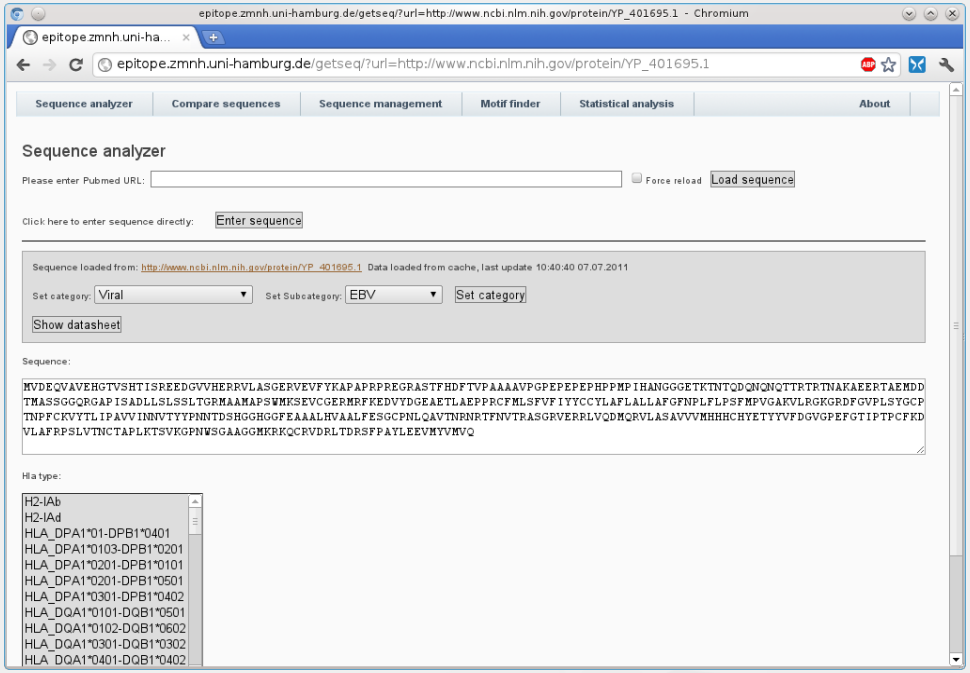
Data Mining in Pepseq

- Die in der Bachelorarbeit entwickelte Analyse-Software um Data Mining Algorithmen erweitern

Zentrum für molekulare
Neurobiologie Hamburg

Pepseq



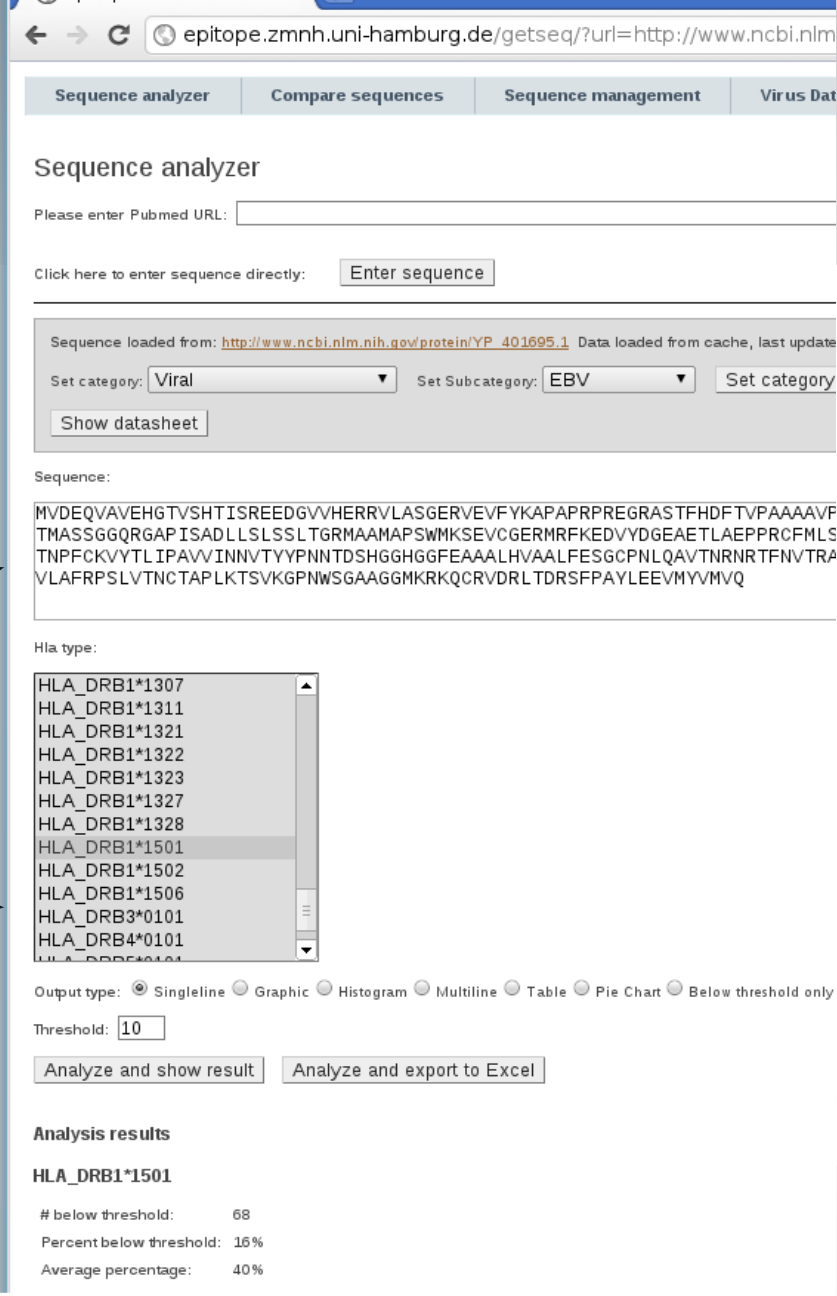


In Silico Vorhersage, ob ein Protein eine Immunrekation auslöst

Ziel:
Rückschlüsse ziehen, ob Proteine im eigenen Körper Ähnlichkeiten zu Proteinen von Viren haben

Aufgaben der Software

- Analyse von **Proteinsequenzen** auf die Bindung an bestimmte **MHC-II-Moleküle**
- Berechnung einer **Bindungswahrscheinlichkeit**



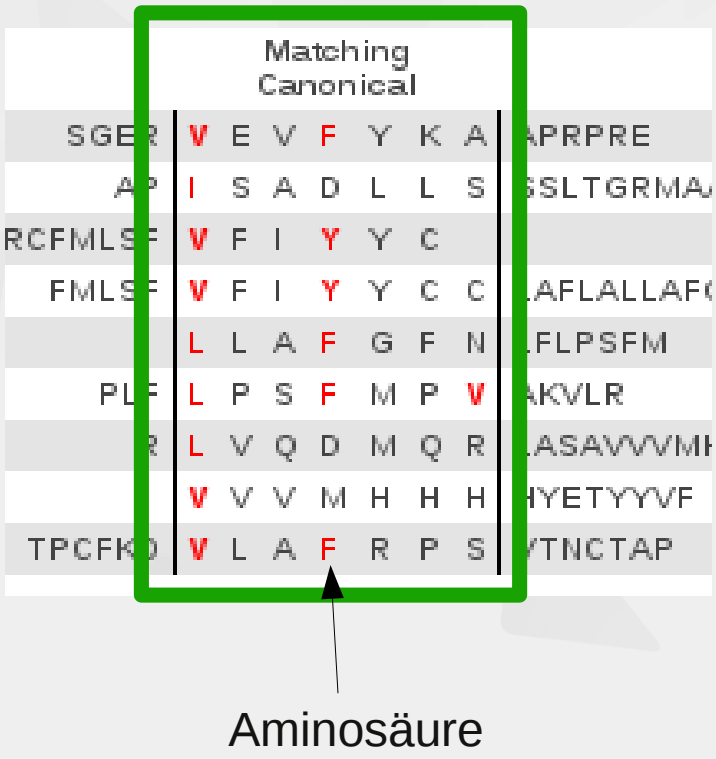
Bei einer hohen Bindungswahrscheinlichkeit kommt es zu einer Immunreaktion

Analyse

- Proteinsequenzen einer Virendatenbank mit Insgesamt **1200 Viren** sollen analysiert werden
- ca. **500.000 Datensätze,** die dabei entstehen und weiterverarbeitet werden müssen.
- Analyse einer größeren Virendatenbank mit **100.000 Viren** in Planung.

Data Mining der Analyseergebnisse

Proteinsequenzen mit einer hohen Bindungs-wahrscheinlichkeit



- Welche Aminosäuren treten häufig gemeinsam auf?
- Gibt es Ähnlichkeiten zwischen Proteinen im menschlichen Körper und Proteinen von Viren?
- **Erster Ansatz:**

  Implementierung eines FP-Trees zum Finden von Korrelationen
- **Später:**

  Charakterisierung der Aminosäuren anhand von Ladung, Größe, Hydrophobie

  Klassifizierung mit Support Vector Machines

Ablauf

- Motivation
- Speicherung der Daten für das Data Mining
- Data Mining Algorithmen
- Ausblick auf die Masterarbeit
- **Konferenzen und Papers** ←

Konferenzen

- 8. bis 12. Februar 2012
  Fifth ACM International Conference on Web Search and Data Mining
  http://www.wsdm2012.org/
  Seattle, USA

- 12. bis 16. August 2012
  18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
  http://www.kdd.org/kdd2012/
  Beijing, China

- 10. bis 13. Dezember 2012
  IEEE International Conference on Data Mining
  Brüssel, Belgien

... und viele mehr

Papers

- **Mining Frequent Patterns without Candidate Generation**
  Jiawei Han, Jian Pei, Yiwen Yin
  ACM 2000 1-58113-218-2/00/05

- **Fast Algorithms for Mining Association Rules**
  Rakesh Agrawal, Ramakrishnan Srikant
  Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994

- **Top 10 algorithms in data mining**
  Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg
  Springer 2007

Papers (2)

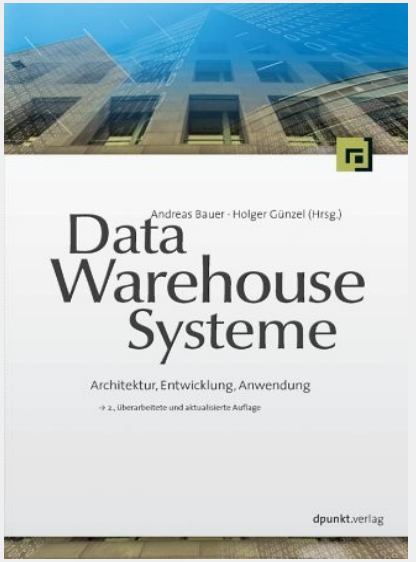
- **MapReduce: Simplified Data Processing on Large Clusters**
  Jeffrey Dean, Sanjay Ghemawat
  OSDI'04, Sixth Symposium on Operating System Design and Implementation
  Dezember 2004

- **A training algorithm for optimal margin classifiers**
  B. E. Boser, I. Guyon, and V. Vapnik
  Proceedings of the Fifth Annual Workshop on Computational Learning Theory
  ACM Press, 1992

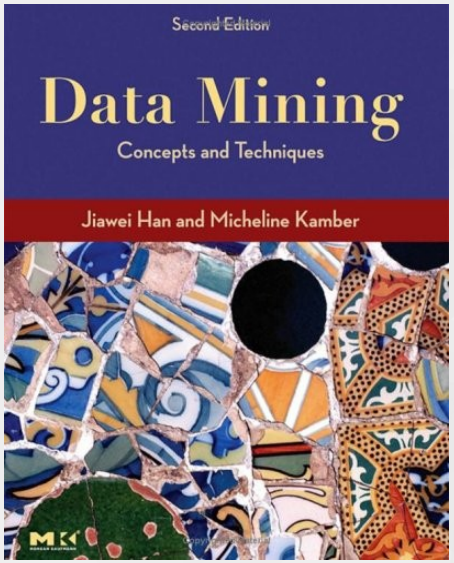
- **Support-Vector Networks**
  Corinna Cortes, Vladimir Vapnik
  Machine Learning, 20, 273-297, 1995

Literatur

- [1] Data Warehouse Systeme – Architektur, Entwicklung, Anwendung von Andreas Bauer, Holger Günzel, dpunkt Verlag



- [2] Data Mining – Concepts and Techniques von Jiawei Han, Michele Kamber – Morgan Kaufmann Verlag



Vielen Dank für die Aufmerksamkeit!