



# Datenanalyse mit Data Mining

- ▼ von Jan-Christoph Meier
- ▼ Hamburg, 19.01.2012

# Ablauf

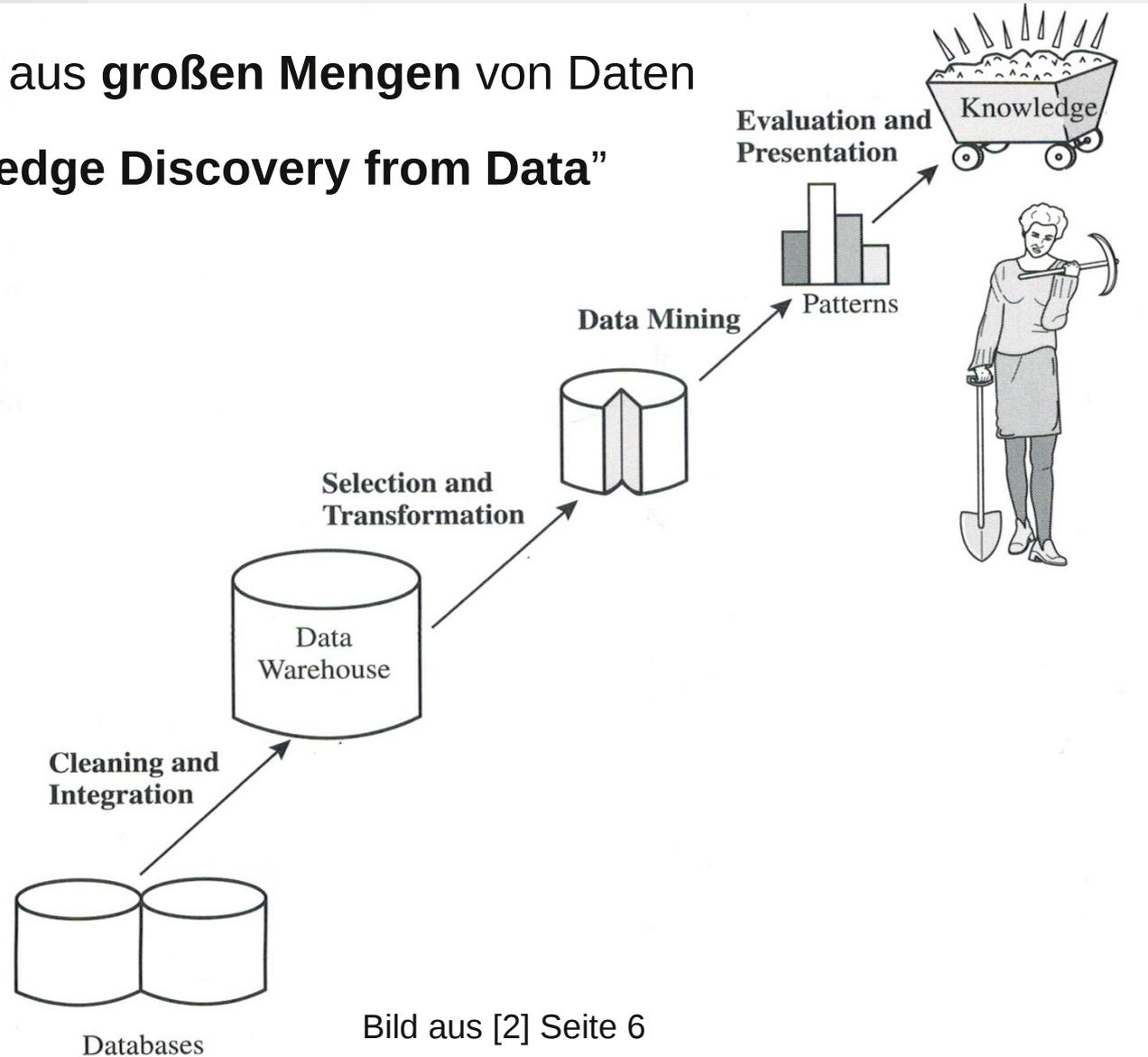
- ▼ Motivation
- ▼ Speicherung der Daten für das Data Mining
- ▼ Data Mining Algorithmen
- ▼ Ausblick auf die Masterarbeit
- ▼ Konferenzen und Papers

# 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”**



# 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 Symptom 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?

The screenshot shows the Amazon.de homepage with several recommendation sections:

- Empfehlungen für Sie: Bücher**
  - Git** - kurz & gut von Sven Riedel, EUR 9,90
  - Continuous Integration mit Hudson** von Simon Wiest, EUR 39,90
  - Macht's gut, und danke für den Fisch** von Douglas Adams, EUR 7,95
  - DUDEN - Deutsche Grammatik** von Rudolf Hoberg, Ursula Hoberg, EUR 7,95
  - Pro Git** von Scott Chacon, Junio C. Hamano, EUR 29,99
- Empfehlungen für Sie: Elektronik**
  - mumbi Premium Echt Ledertasche** für Samsung, EUR 12,98
  - 5 x mumbi Displayschutzfolie** für Samsung, EUR 6,79
  - SAMSUNG I9100 GALAXY S II TPU SILIKON...** QUBITS, EUR 4,95
  - Suncase Ledertasche für Samsung**, EUR 14,90
  - 2 Jahre GERÄTE-SCHUTZ mit Diebstahl-Schutz** von ERGO Direkt, EUR 59,99
- Inspiziert von Ihren Shopping-Trends**
  - Scythe SCKTN-3000 Katana 3 CPU Kühler**, EUR 23,89
  - Asus M5A99X Evo Socket AM3+ Mainboard**, EUR 104,89
  - Asrock M3A770DE Mainboard Socket AMD...**, EUR 57,89
  - Zalman CNPS 7X LED Prozessorkühler**, EUR 34,02
  - ASRock 870 EXTREME3 Mainboard Socket...**, EUR 71,89

On the right side, there is a list of best-selling electronics items:

- Kindle eReader, Wi-Fi, 15 cm (6 Zoll) E Ink Display, ... Amazon, EUR 99,00
- Samsung S5230 Star Smartphone (Touchscreen, 3MP Kamera, ...), EUR 245,00 EUR 69,00
- Samsung Galaxy Ace S5830 Smartphone (8,9 cm (3,5 Zoll) ...), EUR 369,00 EUR 199,00
- Apple TV (kabellose Verbindung zwischen iPhone, iPad, iPod ...), EUR 107,90
- Samsung Galaxy S II (9100) DualCore Smartphone (10.9 cm (4.1 Zoll) ...), EUR 649,00 EUR 451,89

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

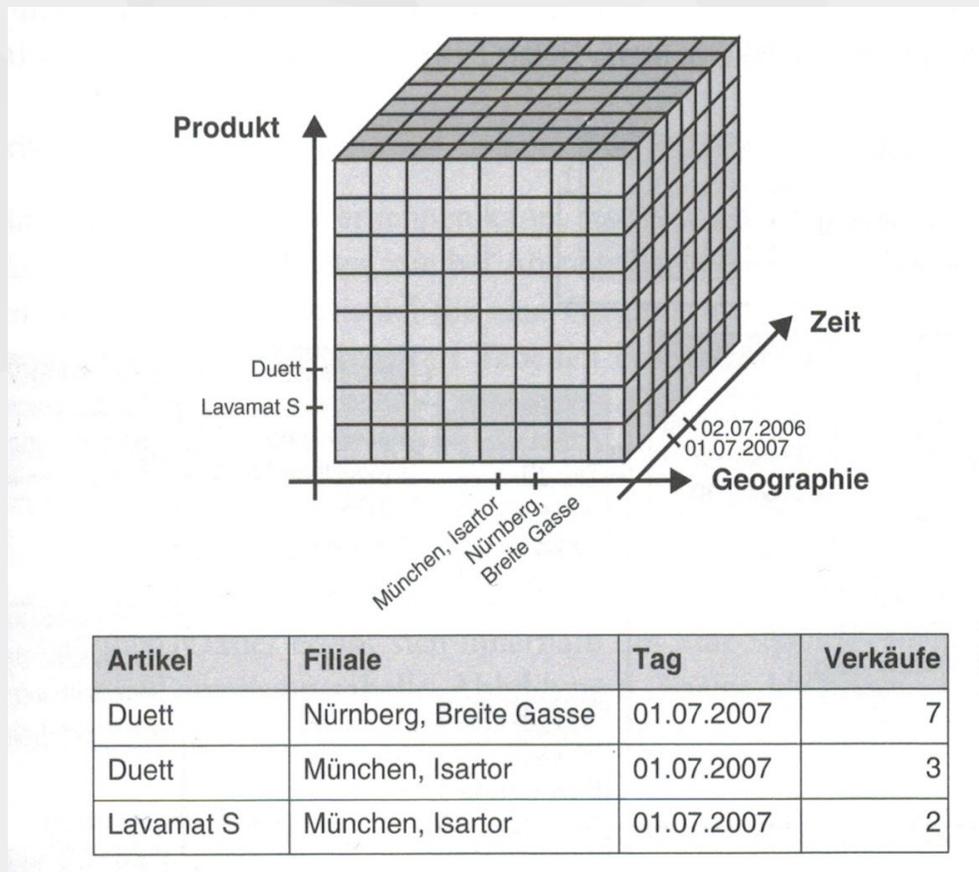
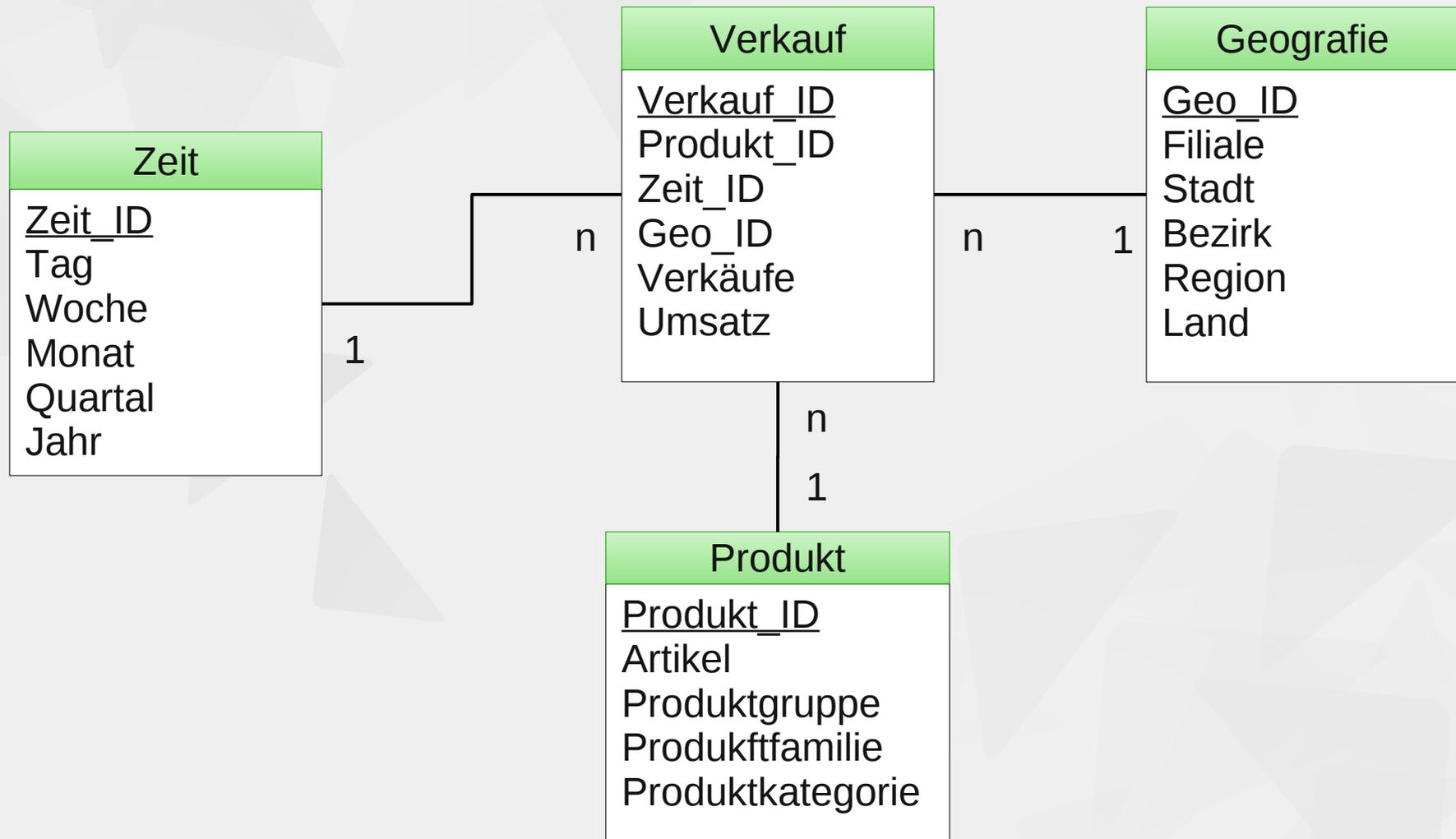


Bild aus [1] Seite 215

**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.

$$\text{support}(A) = P(A)$$

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

- ▼ **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

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

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

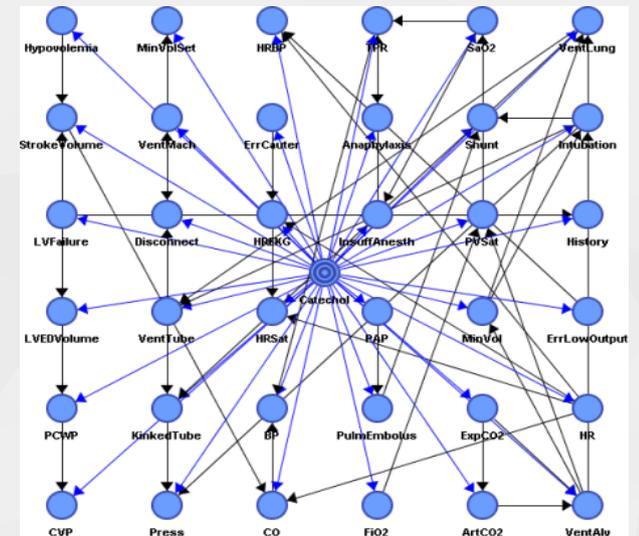
# Finden von Korrelationen in Daten

## ▼ 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  
 $\geq 3$



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

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

# Apriori Algorithmus (2)

Minimum Support Count = 3

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

Itemsets Bilden

Itemset
{I1, I2}
{I1, I3}
{I3, I2}

Support berechnen

Itemset	Support Count
{I1, I2}	4
{I1, I3}	4
{I3, I2}	4

Itemsets Bilden

Itemset
{I1, I2, I3}

Support berechnen

Itemset	Support Count
{I1, I2, I3}	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

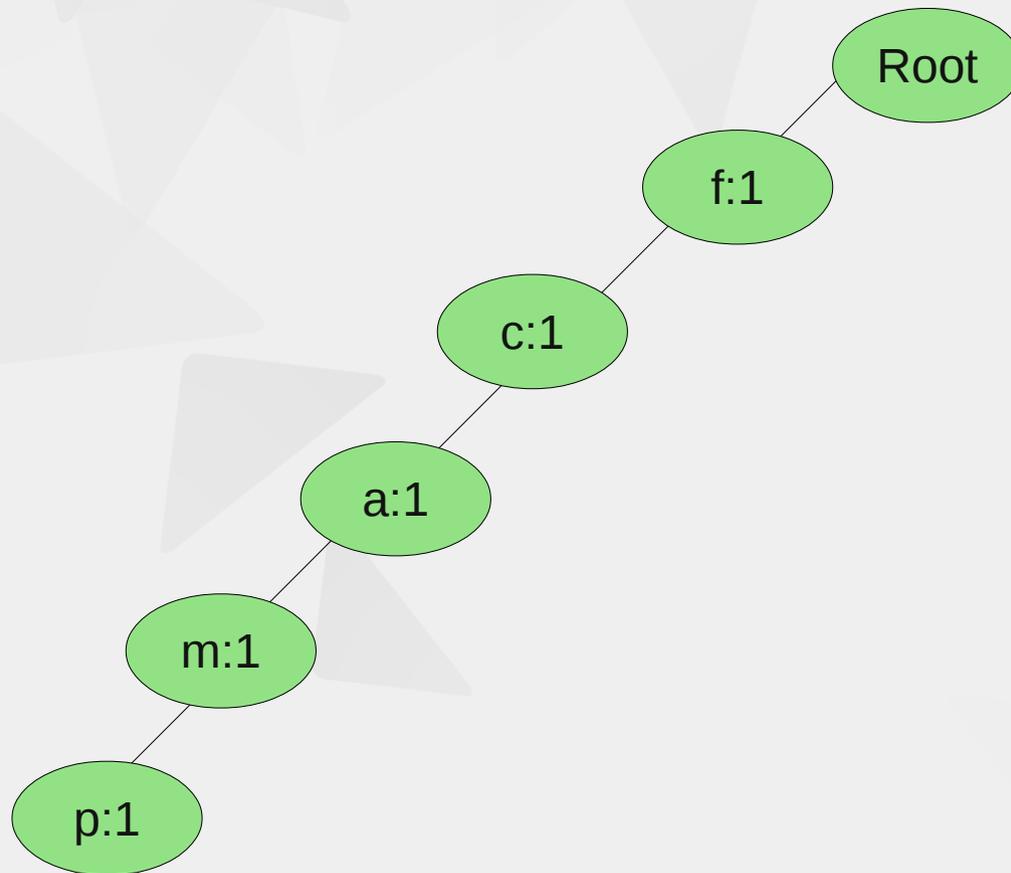
- ▼ FP = Frequent Pattern
- ▼ Verbesserung gegenüber Apriori zum Finden von Korrelationen in Daten.
- ▼ Eine Baumstruktur wird verwendet, um die Korrelationen zu bestimmen.

TID	Transaktionen	Elemente mit Support $\geq 3$
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

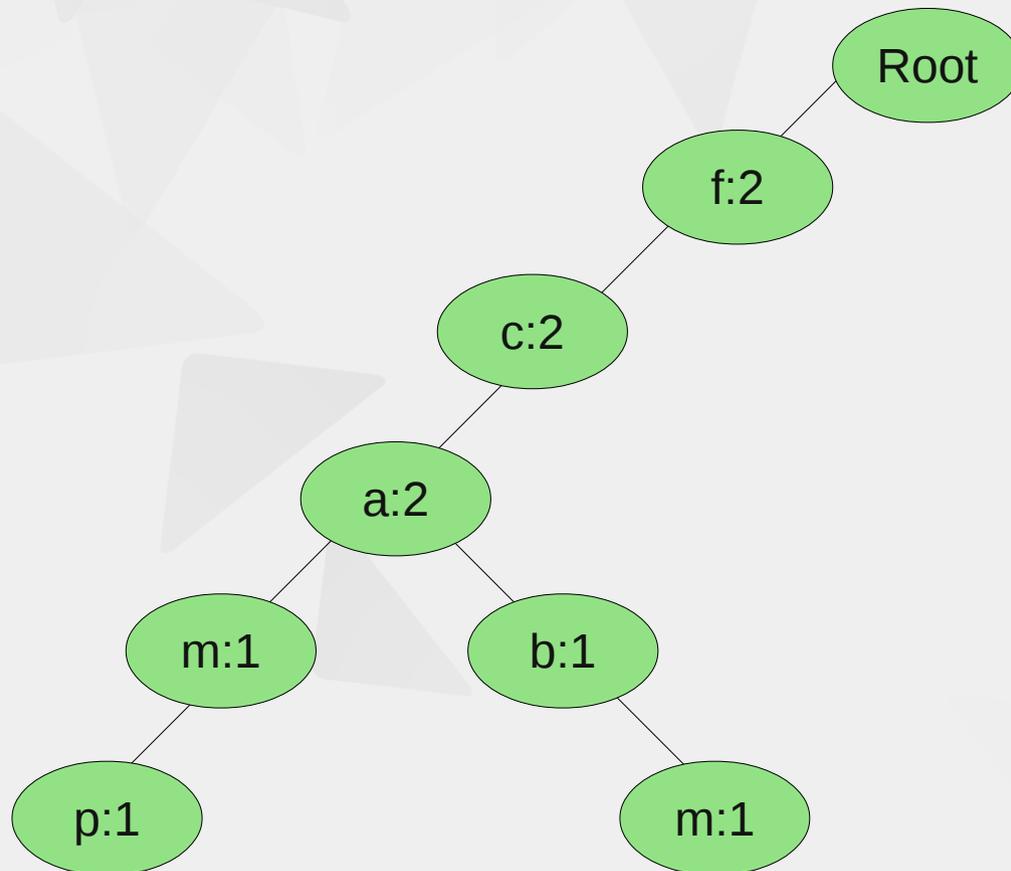
# FP-Tree – Konstruktion des Baums

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



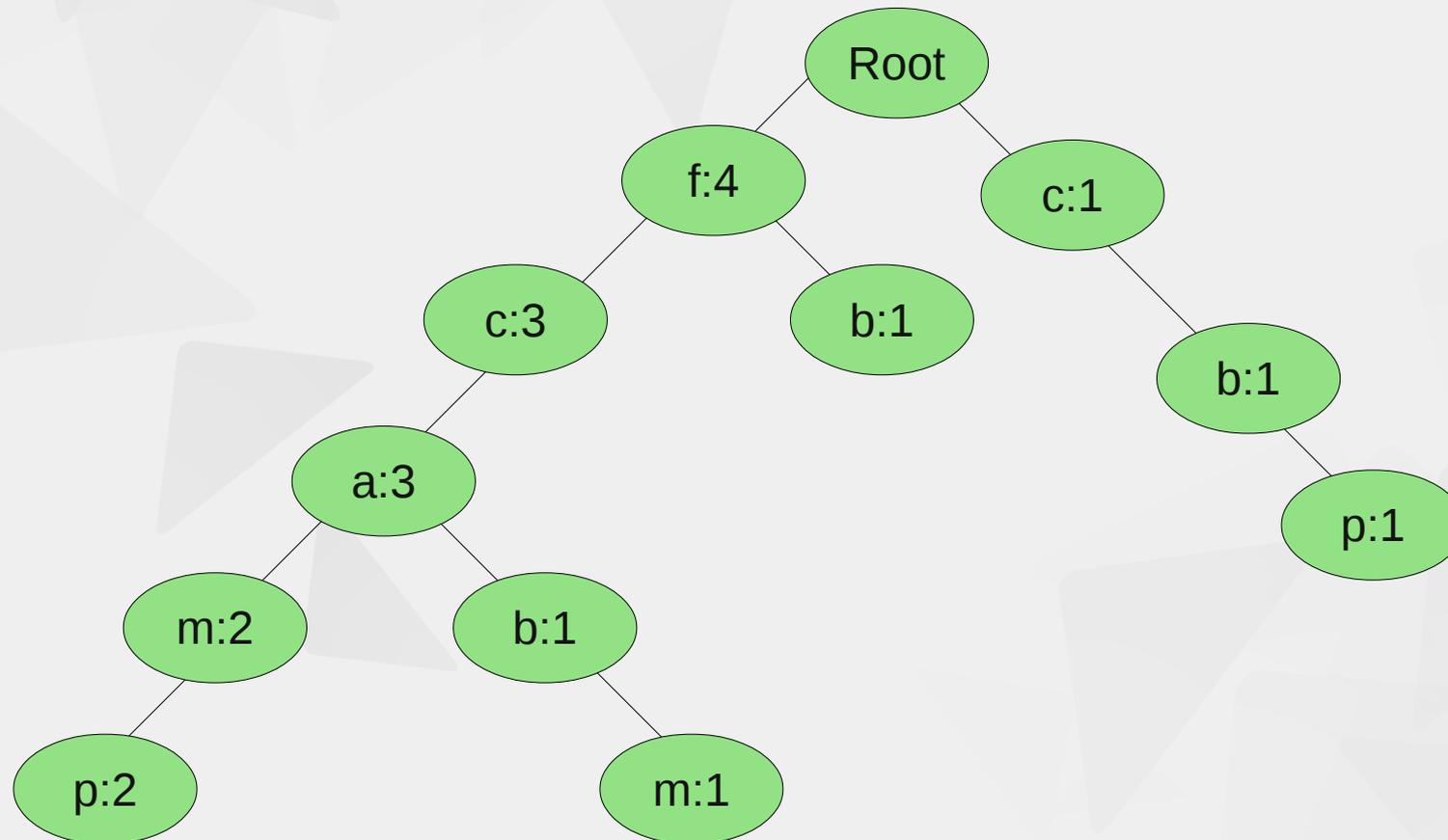
# FP-Tree – Konstruktion des Baums

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



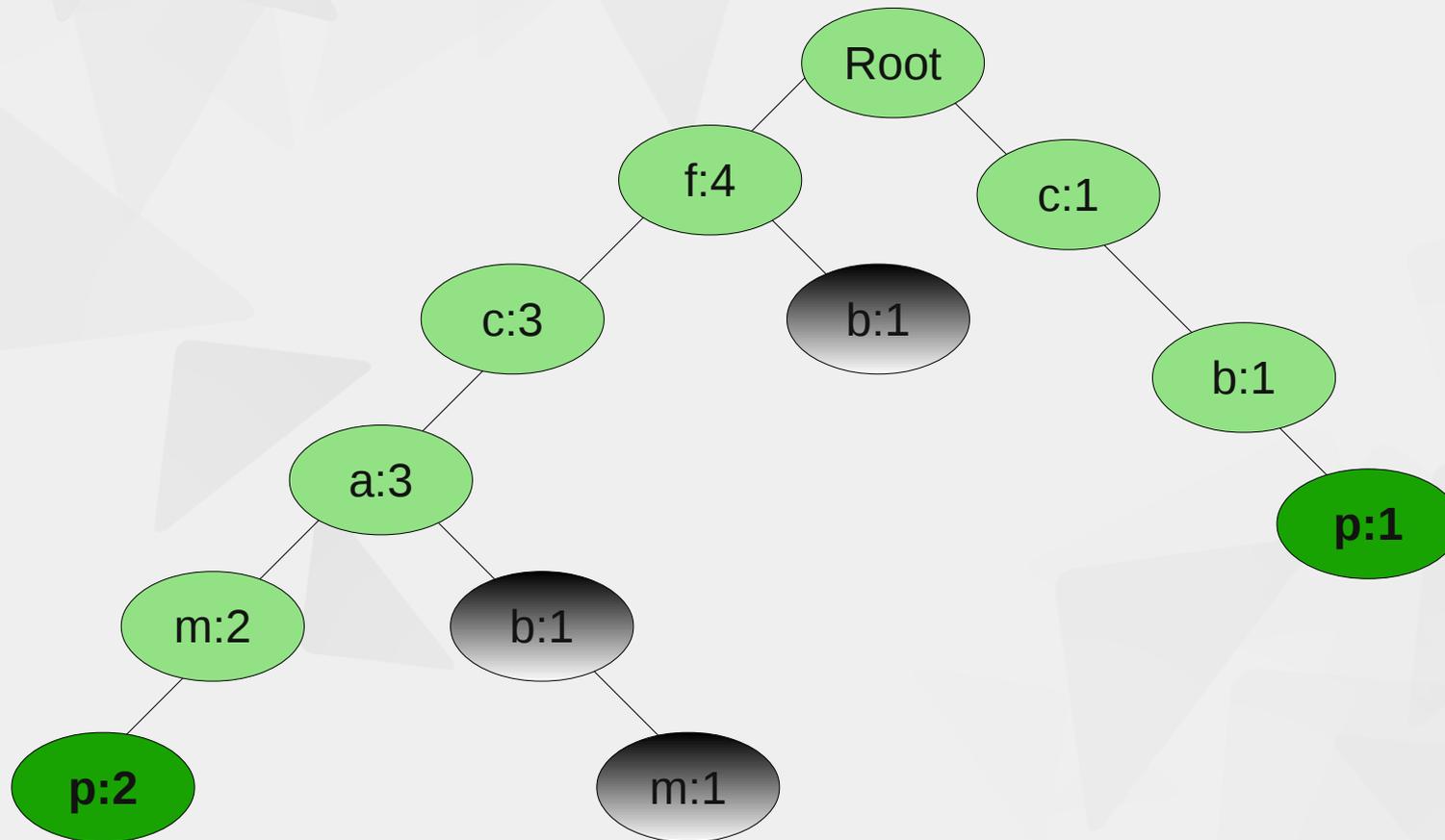
# FP-Tree – Konstruktion des Baums

- ▼ Ergebnis – Alle Transaktionen wurden eingefügt



# FP-Tree – Analyse des Baums

- ▼ In welchen Pfaden ist p enthalten?



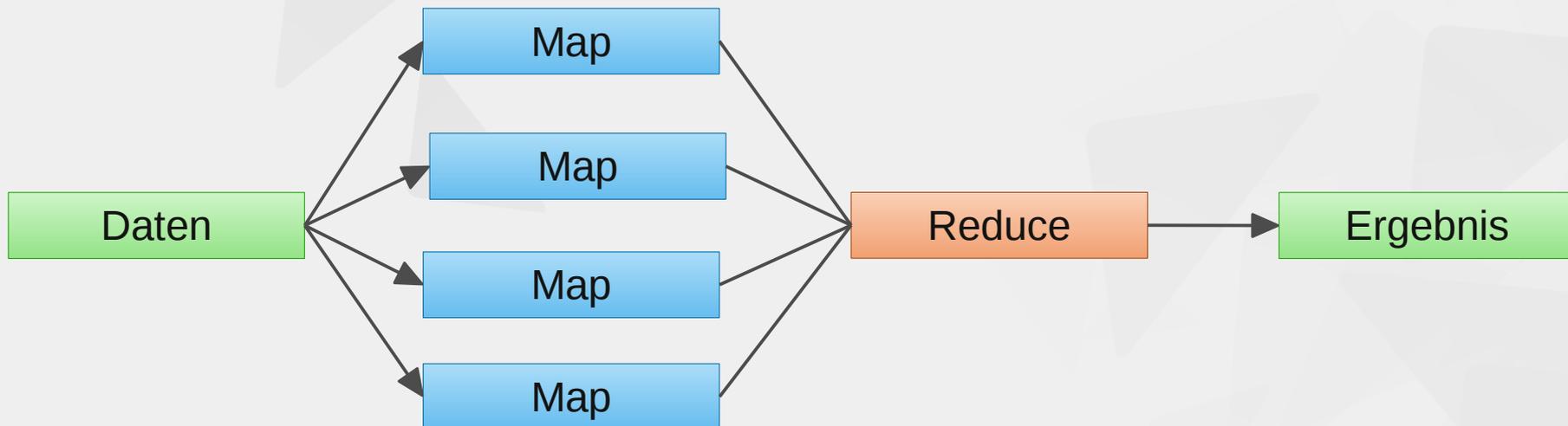
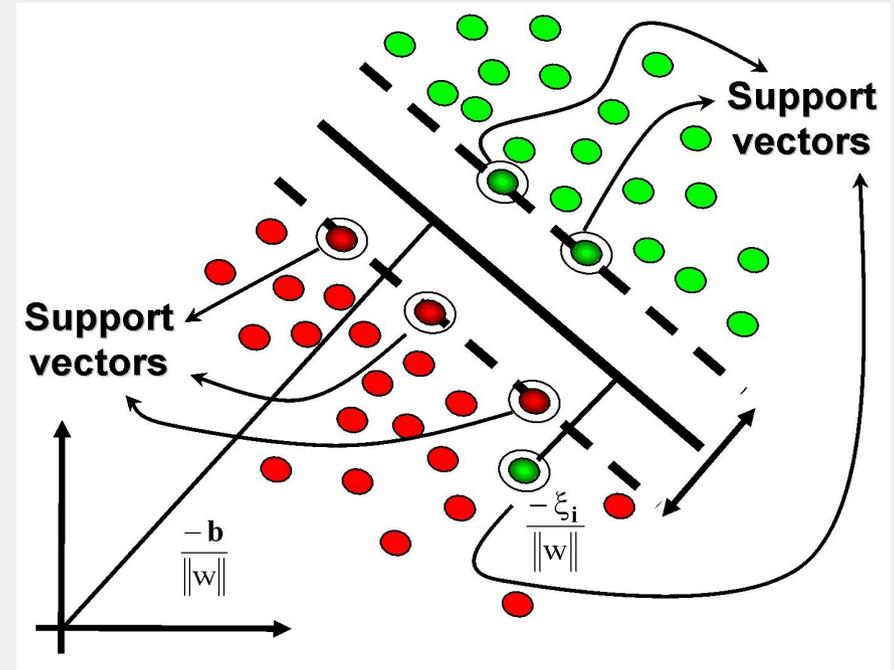
# FP-Tree Support berechnen

- ▼ 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



Institute for Neuroimmunology and  
Clinical Multiple Sclerosis Research

# Pepseq

The screenshot shows the 'Sequence Manager' page of the Epitope Manager tool. It features a navigation menu with options like 'Sequence analyzer', 'Compare sequences', 'Sequence management', 'Motif finder', 'Statistical analysis', and 'About'. Below the menu, there are links for 'Remove all from comparison', 'Remove all from motif finder', and 'Export: Remove all from export' and 'Export to Excel'. The main content area is titled 'Viral' and contains a table with columns for 'Name', '# Amino acid', 'Pubmed url', and several action buttons. The table lists four entries for Human herpesvirus 4 (EBV).

	Name	# Amino acid	Pubmed url						
1	BDLF2 [Human herpesvirus 4].	420	<a href="#">Link</a>	updated at July 7, 2011, 10:40 a.m.	Delete	<a href="#">Select for comparison</a>	<a href="#">Analyze with motif finder</a>	<a href="#">Add to export</a>	<a href="#">Analyze</a>
2	BHLF1 [Human herpesvirus 4].	660	<a href="#">Link</a>	updated at March 16, 2011, 3:45 p.m.	Delete	<a href="#">Select for comparison</a>	<a href="#">Analyze with motif finder</a>	<a href="#">Add to export</a>	<a href="#">Analyze</a>
3	BKRF4 [Human herpesvirus 4].	217	<a href="#">Link</a>	updated at Sept. 14, 2010, 1:21 p.m.	Delete	<a href="#">Select for comparison</a>	<a href="#">Analyze with motif finder</a>	<a href="#">Add to export</a>	<a href="#">Analyze</a>
4	BLRF2 [Human herpesvirus 4].	162	<a href="#">Link</a>	updated at Sept. 14, 2010, 1:22 p.m.	Delete	<a href="#">Select for comparison</a>	<a href="#">Analyze with motif finder</a>	<a href="#">Add to export</a>	<a href="#">Analyze</a>

The screenshot shows the 'Sequence analyzer' page of the Epitope Manager tool. It includes a navigation menu and a form to enter a sequence. The 'Sequence loaded from' field shows the URL 'http://www.ncbi.nlm.nih.gov/protein/YP\_401695.1'. The 'Set category' is 'Viral' and the 'Set Subcategory' is 'EBV'. The 'Sequence' field contains a long amino acid sequence. Below the sequence, there is a list of HLA types.

Sequence loaded from: [http://www.ncbi.nlm.nih.gov/protein/YP\\_401695.1](http://www.ncbi.nlm.nih.gov/protein/YP_401695.1) Data loaded from cache, last update 10:40:40 07.07.2011

Set category: **Viral** Set Subcategory: **EBV** [Set category](#)

[Show datasheet](#)

Sequence:

```
MVDEQVAVEHGTVSHTISREEDGVVHEPVLASGERVEVFYKAPAPFPREGRASTFHDFTVPAAAAVPGPEPEPEPHFMP I HANGGGETKNTQDNQNTTTRTNKAERTAEMDD  
TMASGGQRGAPISADLLSLSLTGRMAANAPSWMKSEVCGERMRFKEDVTDGEATLAEPFCFMLS FVI IYCCYLAFALLAFGNPLFLPSPMPVGA KVLRGKGRDFGVPLSYGCP  
TNPFCVYTLIPAVVINNVITYPNNTDSHGHHGGFEAAALHVAALFESGCPNLQAVTNNRNFNVTRASGRVERLVDQGRVLAASAVVVHHHCYETIYVFDVGVGFEFTIP TPCFKD  
VLAFRPSLVTNCTAPLKTSSVKGPNWGAAGGMRKQCRVDRLTDRESFPAYLEEVMYVMVQ
```

HLA type:

```
H2-IAb  
H2-IAd  
HLA_DPA1*01-DPB1*0401  
HLA_DPA1*0103-DPB1*0201  
HLA_DPA1*0201-DPB1*0101  
HLA_DPA1*0201-DPB1*0501  
HLA_DPA1*0301-DPB1*0402  
HLA_DQA1*0101-DQB1*0501  
HLA_DQA1*0102-DQB1*0602  
HLA_DQA1*0301-DQB1*0302  
HLA_DQA1*0401-DQB1*0402
```

**In Silico Vorhersage, ob ein Protein eine Immunreaktion 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**

The screenshot shows the 'Sequence analyzer' interface. At the top, there are navigation tabs: 'Sequence analyzer', 'Compare sequences', 'Sequence management', and 'Virus Data'. The main content area is titled 'Sequence analyzer' and includes a 'Please enter Pubmed URL:' field, a 'Click here to enter sequence directly:' link with an 'Enter sequence' button, and a 'Sequence loaded from:' field showing the URL 'http://www.ncbi.nlm.nih.gov/protein/YP\_401695.1'. Below this, there are dropdown menus for 'Set category: Viral' and 'Set Subcategory: EBV', along with a 'Set category' button and a 'Show datasheet' button. The 'Sequence:' field contains a long protein sequence: MVDEQVAVEHGTVSHTISREEDGVVHERRVLAGSERVEVFYKAPAPRPREGRASTFHDF TVPAAAAVPTMSSGGQRGAPISADLLSLSL TGRMAAMAPSWMKSEVCGERMRFKEDVYDGEAETLAEPPRCFMLS TNPFCKVYTLIPAVVINNVYYPNNTDSHGHHGGFEAAALHVAALFESGCPNLQAVTNRNRTFNVTRAVLAFRPSLVTNCTAPLKTSVKGPWNSGAAGGMKRKQCRVDRLTDRSFPAYLEEVYVMVQ. Below the sequence, there is a 'Hla type:' section with a scrollable list of HLA types: HLA\_DRB1\*1307, HLA\_DRB1\*1311, HLA\_DRB1\*1321, HLA\_DRB1\*1322, HLA\_DRB1\*1323, HLA\_DRB1\*1327, HLA\_DRB1\*1328, HLA\_DRB1\*1501, HLA\_DRB1\*1502, HLA\_DRB1\*1506, HLA\_DRB3\*0101, HLA\_DRB4\*0101, and HLA\_DRB4\*0102. Below the list, there are radio buttons for 'Output type:' (Singleline, Graphic, Histogram, Multiline, Table, Pie Chart, Below threshold only) and a 'Threshold:' field set to 10. At the bottom, there are two buttons: 'Analyze and show result' and 'Analyze and export to Excel'. The 'Analysis results' section shows 'HLA\_DRB1\*1501' with the following statistics: '# below threshold: 68', 'Percent below threshold: 16%', and 'Average percentage: 40%'.

# 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 Bindungswahrscheinlichkeit

	Matching Canonical							
SGER	V	E	V	F	Y	K	A	APRPRE
AP	I	S	A	D	L	L	S	SSLTGRMA
RCFMLSF	V	F	I	Y	Y	C		
FMLSF	V	F	I	Y	Y	C	C	LAFLALLAFC
	L	L	A	F	G	F	N	FLPSFM
PLF	L	P	S	F	M	P	V	AKVLR
R	L	V	Q	D	M	Q	R	ASAVVVMF
	V	V	V	M	H	H	H	HYETYYVF
TPCFKD	V	L	A	F	R	P	S	VTNCTAP

Aminosäure

- ▼ 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

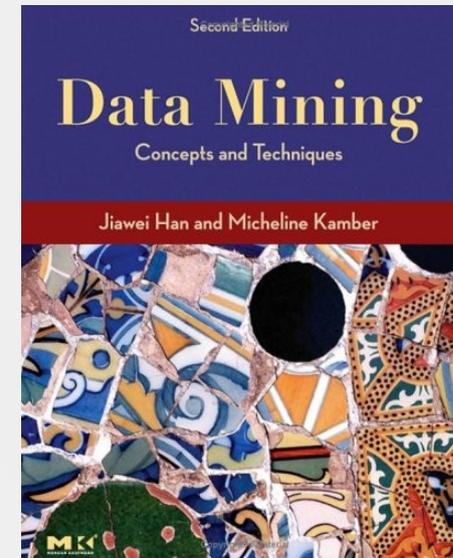
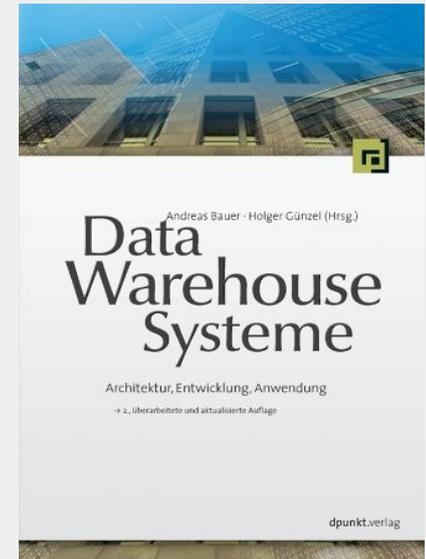
- ▼ **Support-Vector Networks**

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!