RECOMMENDER SYSTEMS

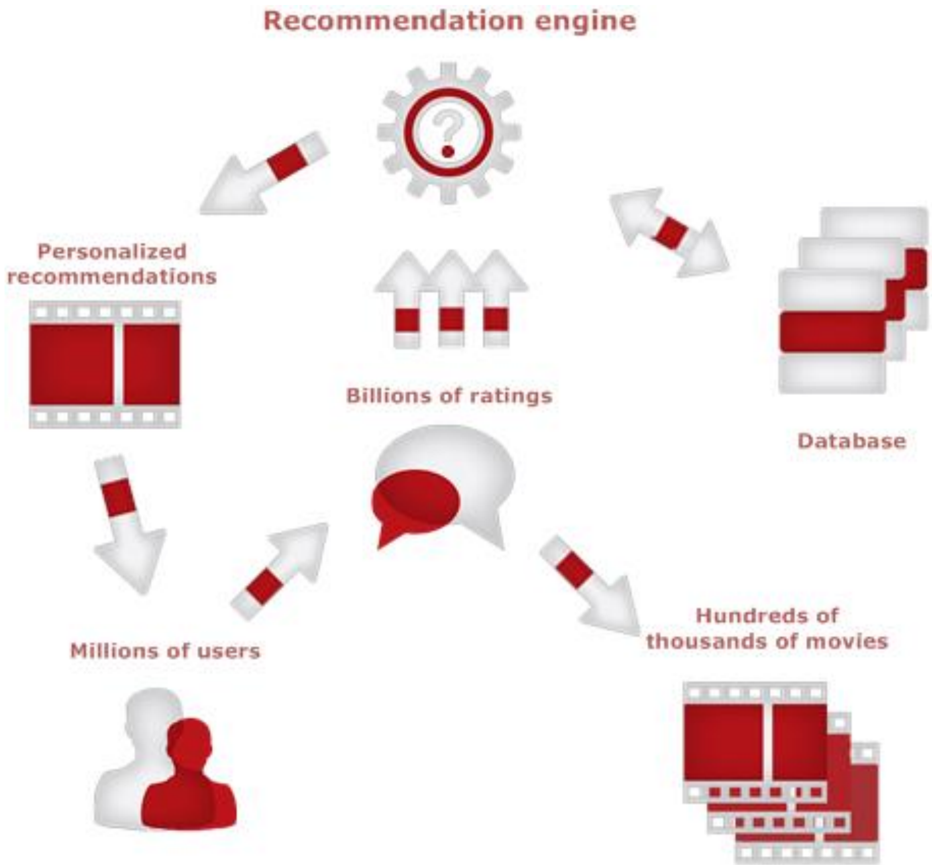
HOW ALGORITHMS PREDICT FUTURE CONSUMPTION BEHAVIOR

MSc. Christopher Krauss
Senior Project Manager R&D
Competence Center Future Applications and Media
christopher.krauss@fokus.fraunhofer.de

Fraunhofer FOKUS

making cities smart

BUT DID YOU KNOW HOW IT WORKS?



Fraunhofer FOKUS

http://www.ni-tekna.com/biz/Netflix_Recommendation_Engine.html *

AGENDA

- Introduction into Data Mining
- Recommendation Engines
- Filtering Approaches
- Demonstration: TV Predictor
- Machine Learning
- Demonstration: RTV & multithek
- Technical issues
- Legal and ethical issues

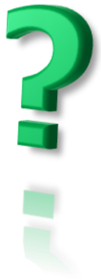
PREDICTIVE DATA MINING AND RECOMMENDATION ENGINES

MSc. Christopher Krauss
Competence Center Future Applications and Media
www.fokus.fraunhofer.de/go/fame

Fraunhofer FOKUS

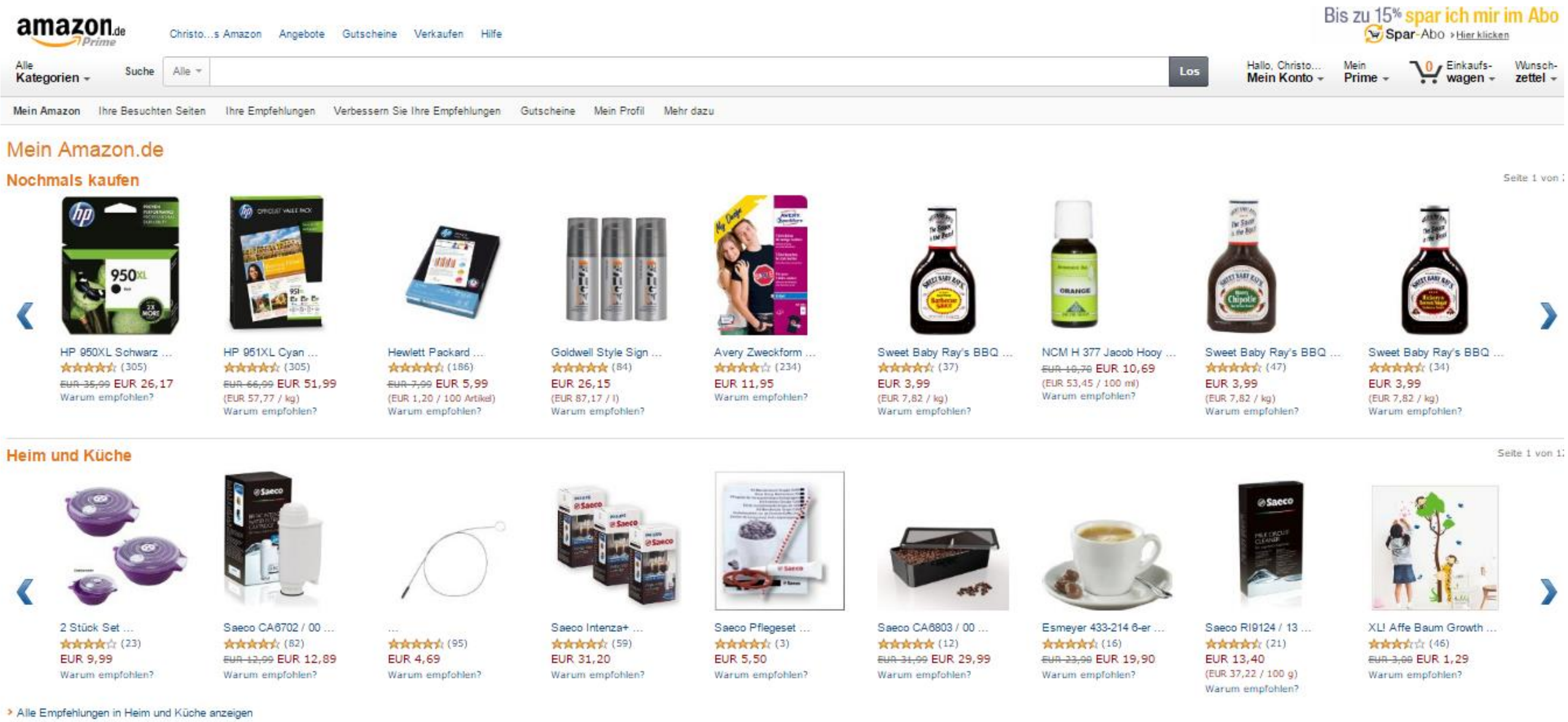
making cities smart

APPLICATION AREAS



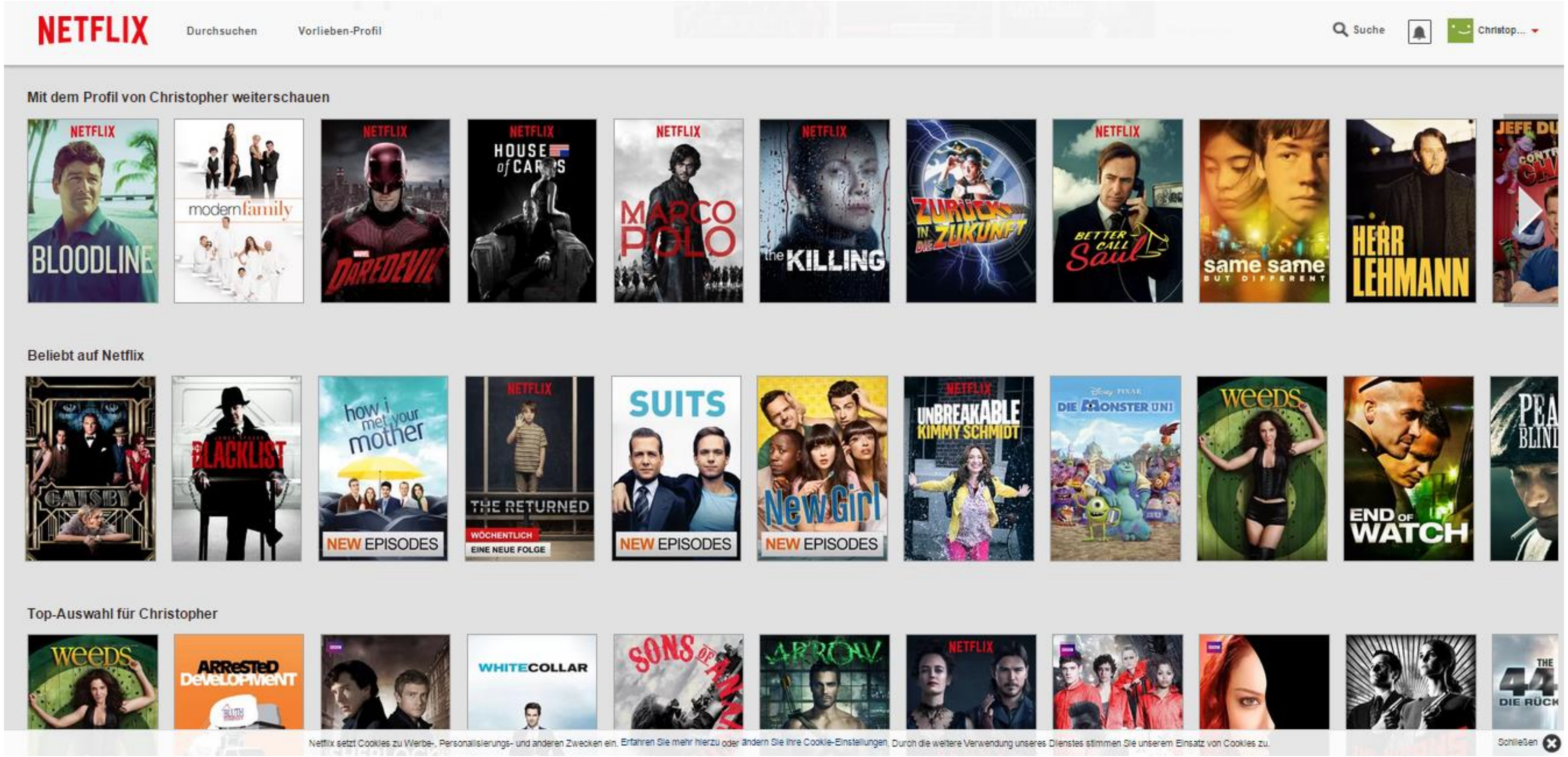
* https://www.google.com/analytics/

APPLICATION AREA - E-COMMERCE



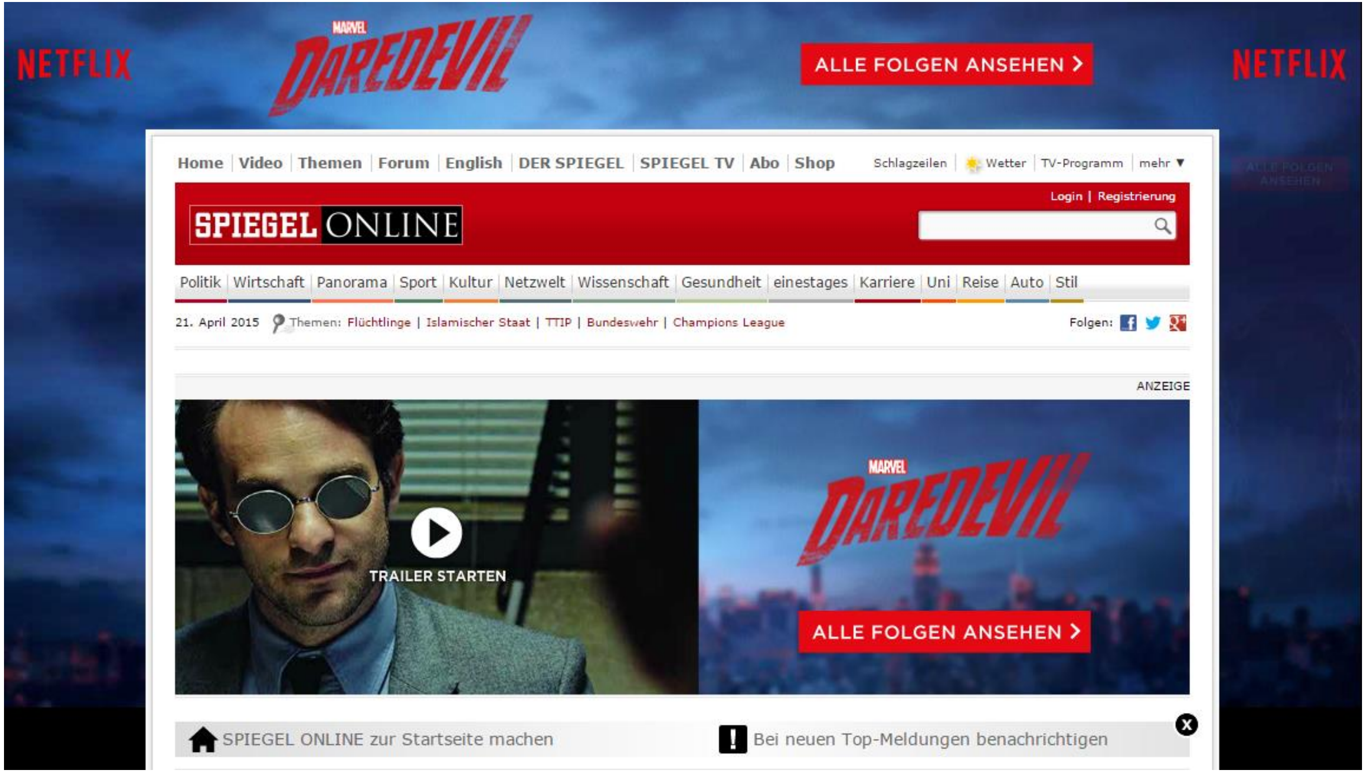
* http://www.amazon.de/

APPLICATION AREA - MEDIA ON DEMAND



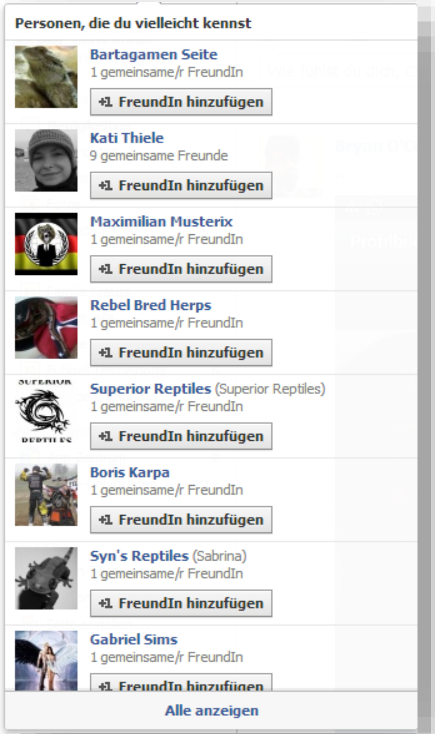
* http://www.netflix.com/

APPLICATION AREA - PERSONALIZED COMMERCIALS

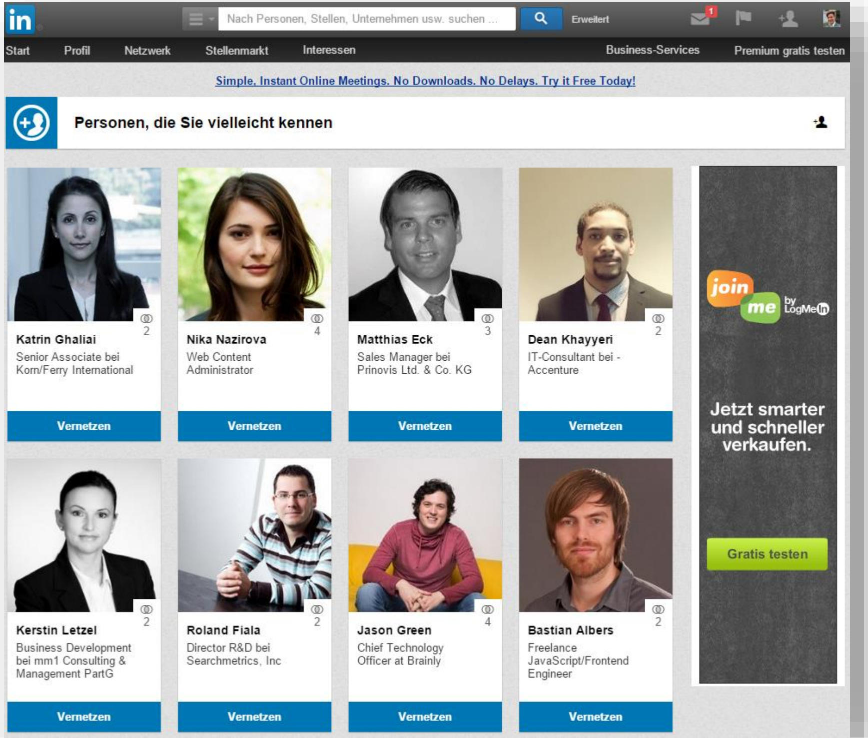


* http://www.spiegel.de/

APPLICATION AREA - SOCIAL-NETWORK ANALYSIS



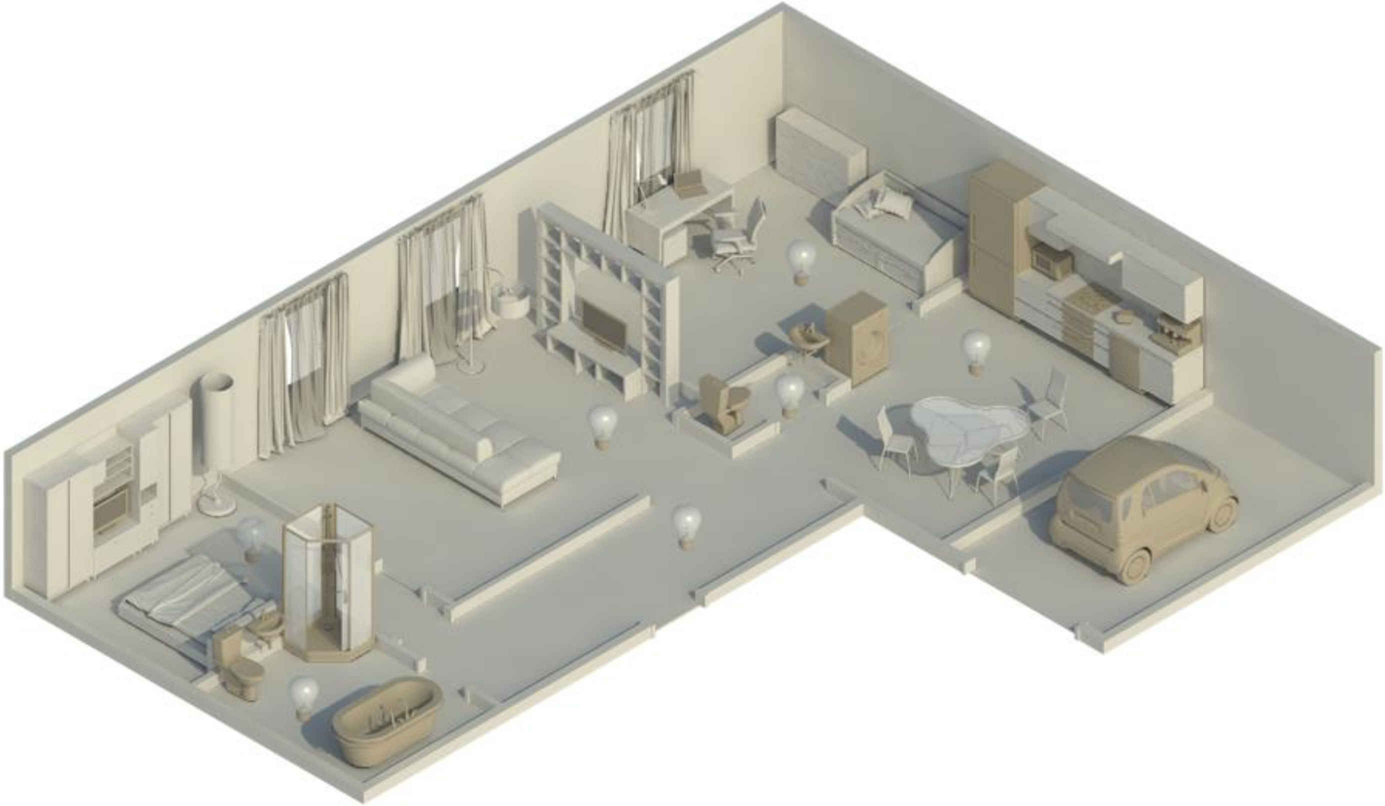
* https://www.facebook.com/



* https://www.linkedin.com/

Fraunhofer FOKUS

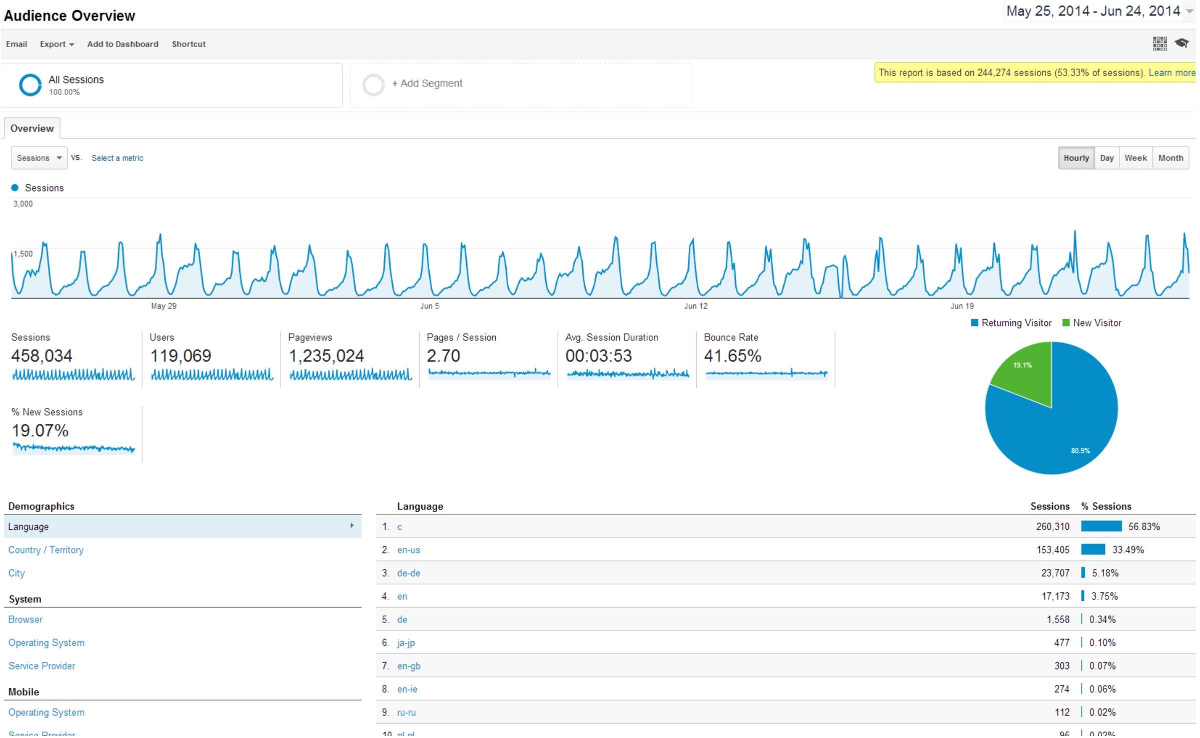
APPLICATION AREA – SMART ENVIRONMENTS



APPLICATION AREA – SPORT ANALYTICS

* http://predictz.com/

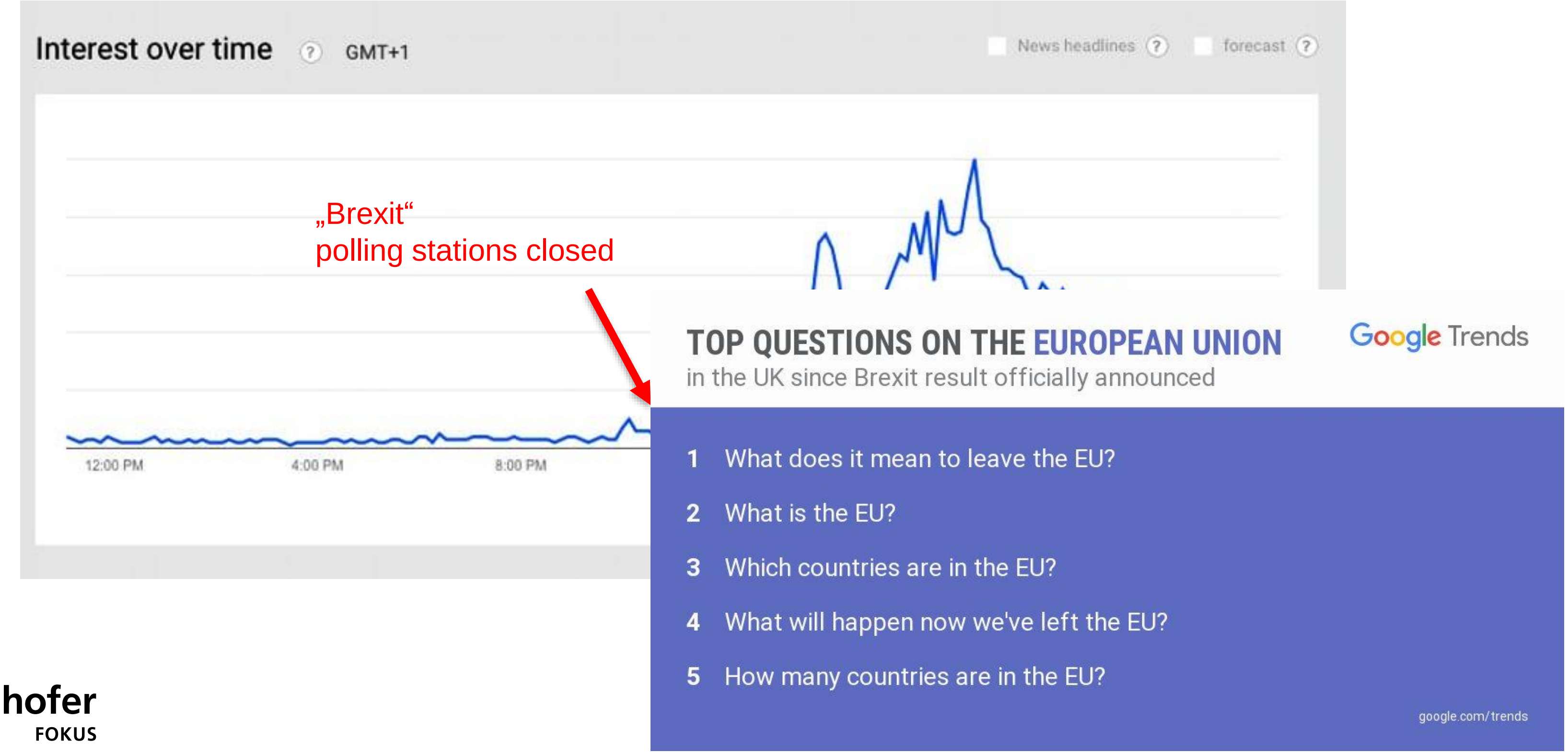
APPLICATION AREA – BUSINESS ANALYTICS



* https://www.google.com/analytics/

Fraunhofer FOKUS

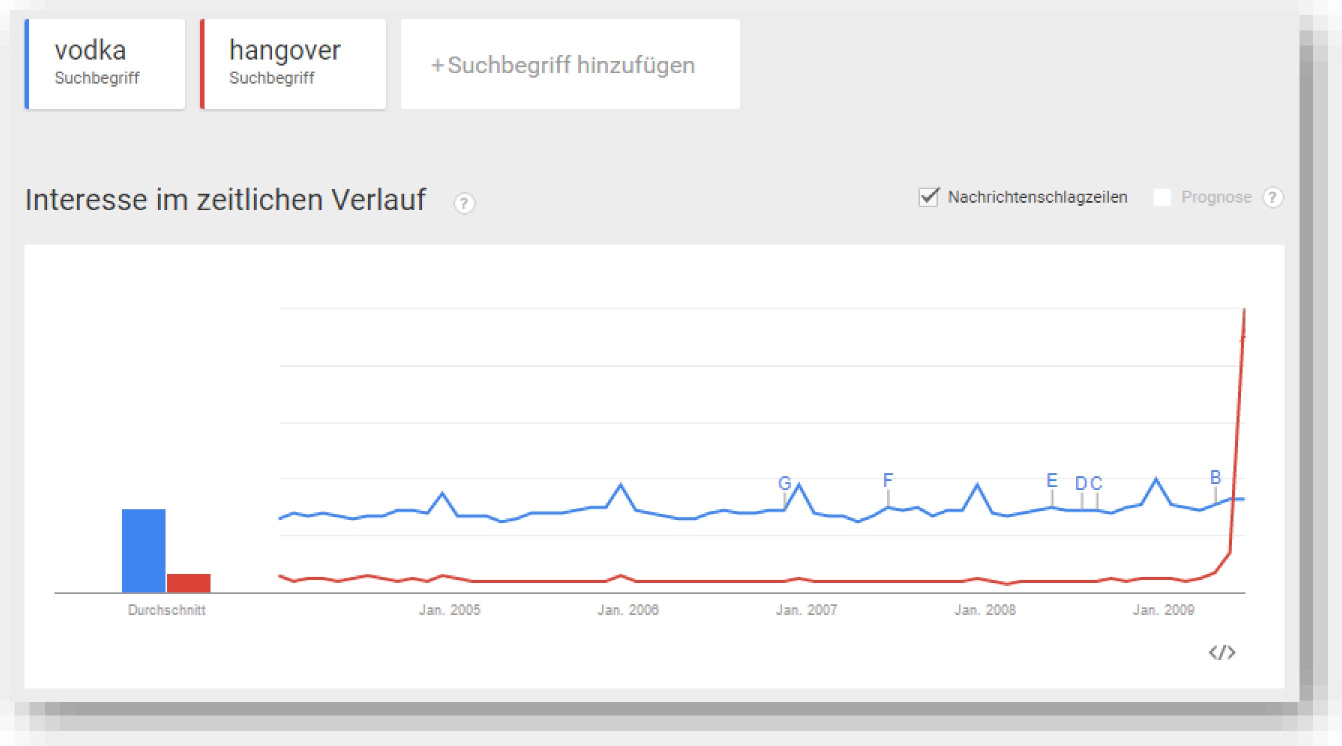
TREND DETECTION



TREND DETECTION/ PREDICTION

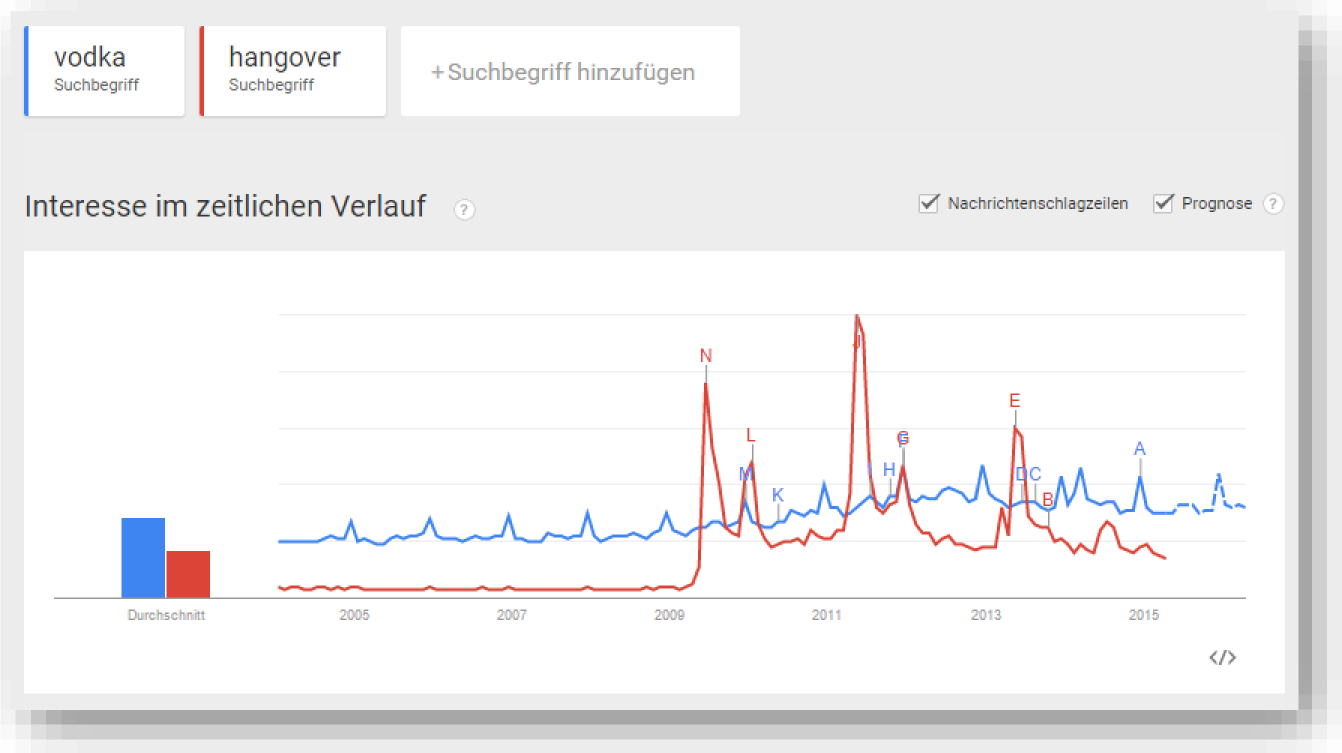
Fraunhofer FOKUS

TREND DETECTION/ PREDICTION

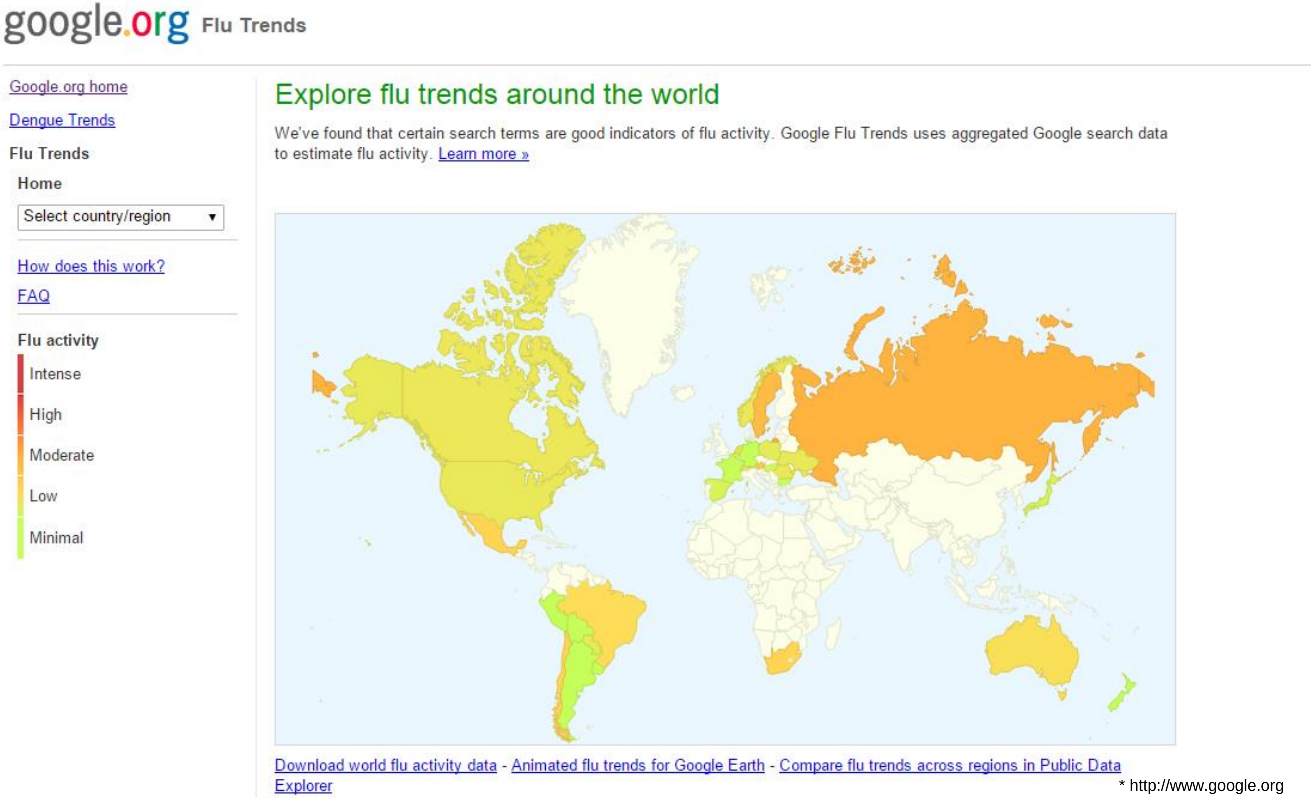


Fraunhofer FOKUS

TREND DETECTION/ PREDICTION



APPLICATION AREA – EMERGENCY PREDICTION

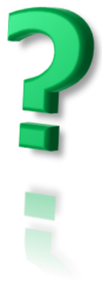


* http://www.google.org

APPLICATION AREA – PRE-CRIME

* http://www.computer-bild.de/

OVER…



691mio products on Amazon

100k movies on MovieLens

450mio pages on MySpace

1bn users on Facebook

300k tracks on Jamendo

100mio websites on Del.icio.us

333mio programs on Hulu

5bn videos on Youtube

67k movies on LoveFilm

30mio movies on IMDb

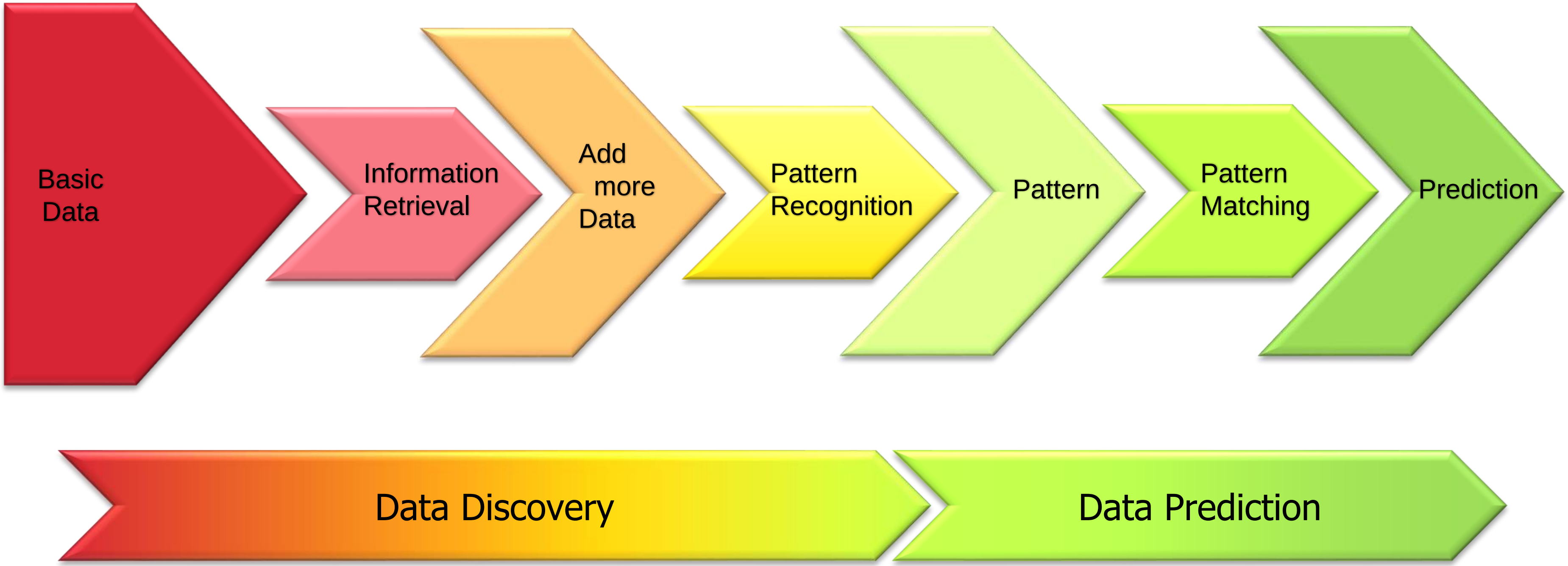
10mio products on eBay at the same time

68mio books on LibraryThing

4mio tracks on Last.fm

70k programs on Netflix

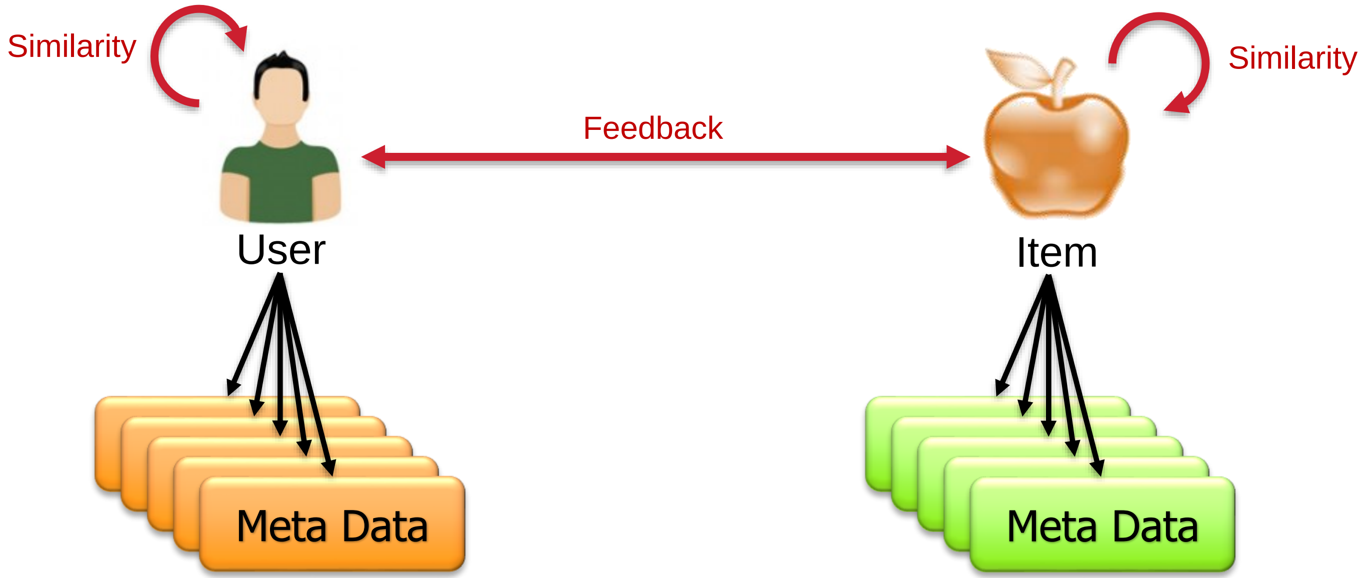
DATA MINING: KNOWLEDGE DISCOVERY IN DATABASES



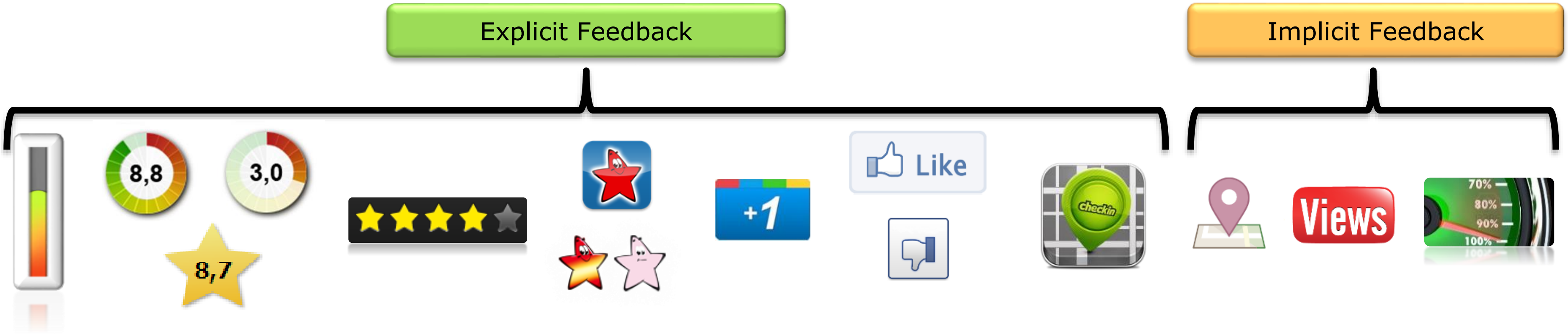
Fraunhofer FOKUS

Recommendation Engines

RECOMMENDATIONS

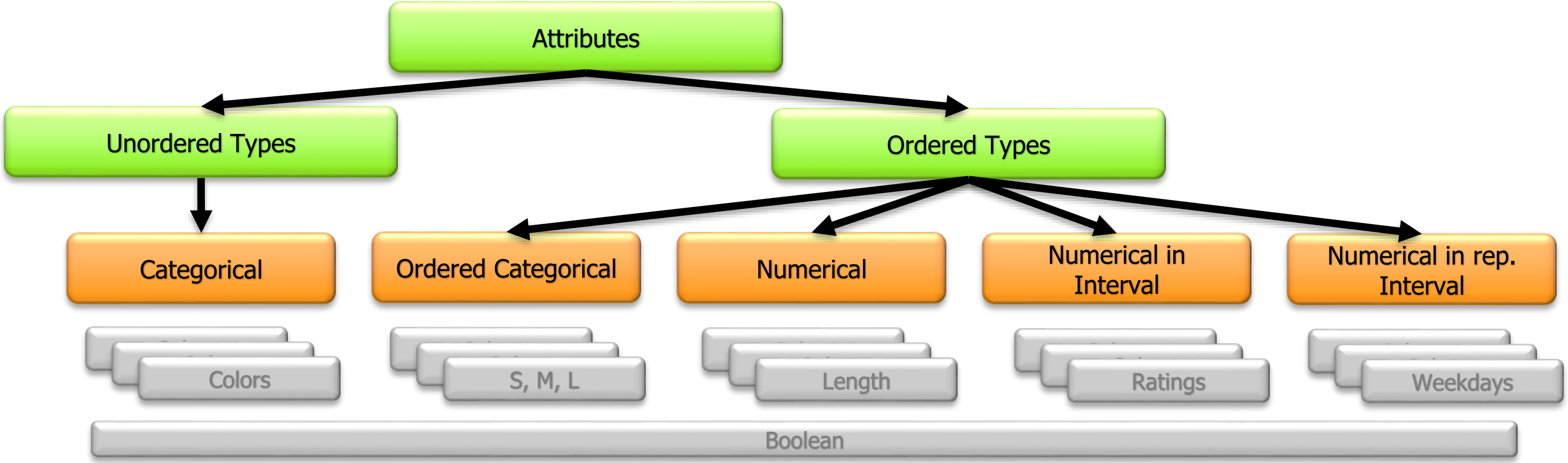


FEEDBACK TYPES



Fraunhofer
FOKUS

META DATA ATTRIBUTES



- Distance and Similarity

$$dist(p,q) = 1 - sim(p,q)$$

According to this formula the similarity value is in range $[0,1]$ and so the distance has the same limitations. But in contrast to the similarity a distance of 0 means an equality of an attribute and 1 stands for the highest dissimilarity.

p and q are the values taken by two elements on the same attribute. n is the number of all possible values of an ordered categorical. For Numericals $max(A)$ is the highest and $min(A)$ is the lowest possible value of an attribute.

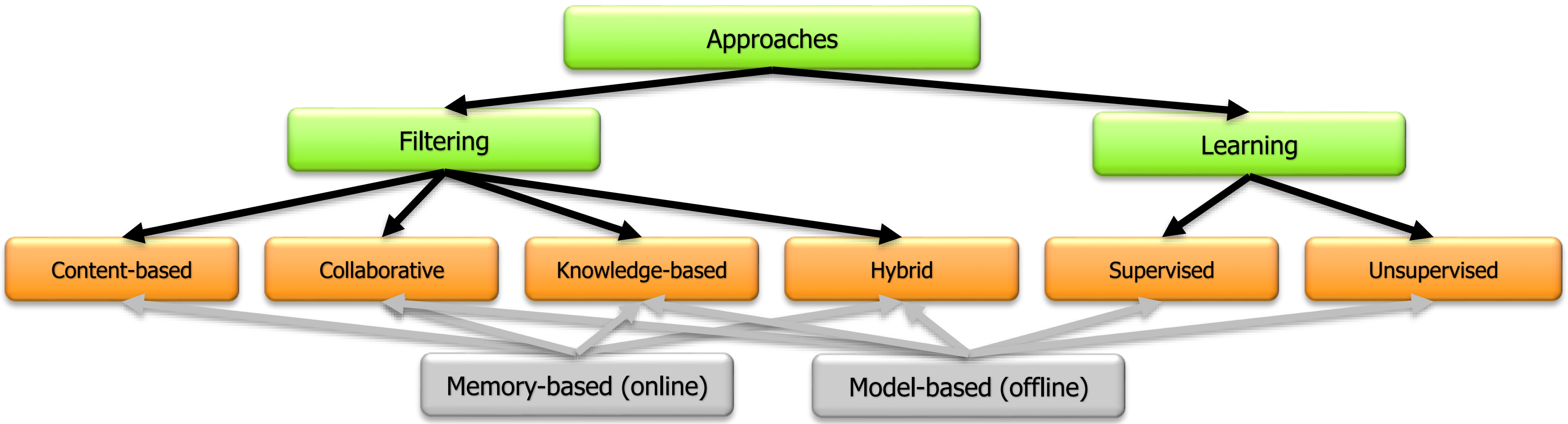
$$sim(p,q) = \begin{cases} 1, & \text{if p=q} \\ 0, & \text{if p} \neq \text{q} \end{cases}$$

$$sim(p,q) = 1 - \frac{|p-q|}{n-1}$$

$$sim(p,q) = 1 - \frac{|p-q|}{max(A) - min(A)}$$

Fraunhofer FOKUS

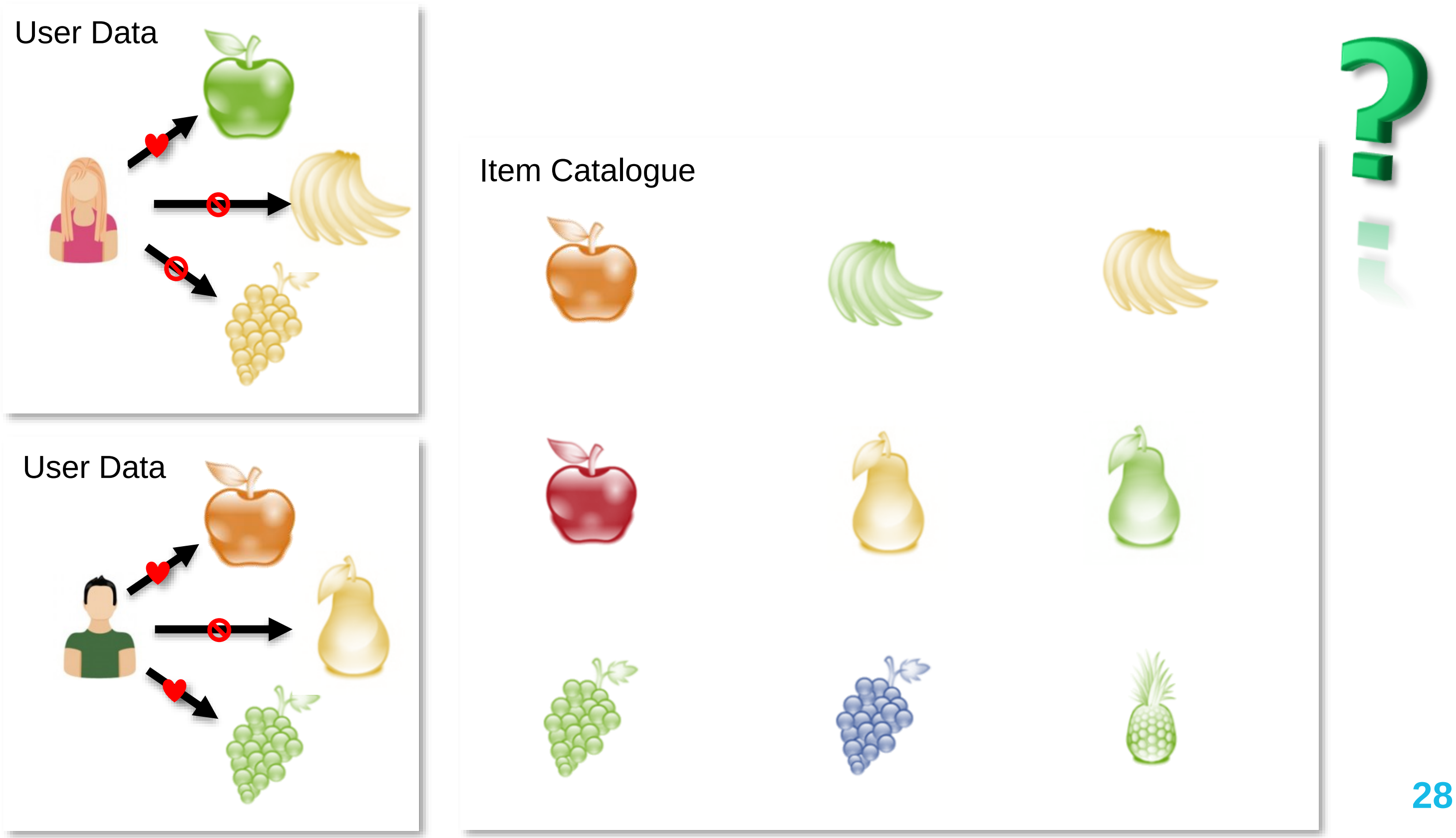
RECOMMENDATION APPROACHES



- **Memory-based (online) Filtering** predicts items depending on the given feedback data set. So the recommendation is calculated on the fly by using just the current database values and no retrospectively transformed data sets.
- **Model-based Filtering algorithms** mostly perform offline calculations, which is often an expensive operation. And only a few tasks are performed on demand.

Filtering Approaches

FILTERING

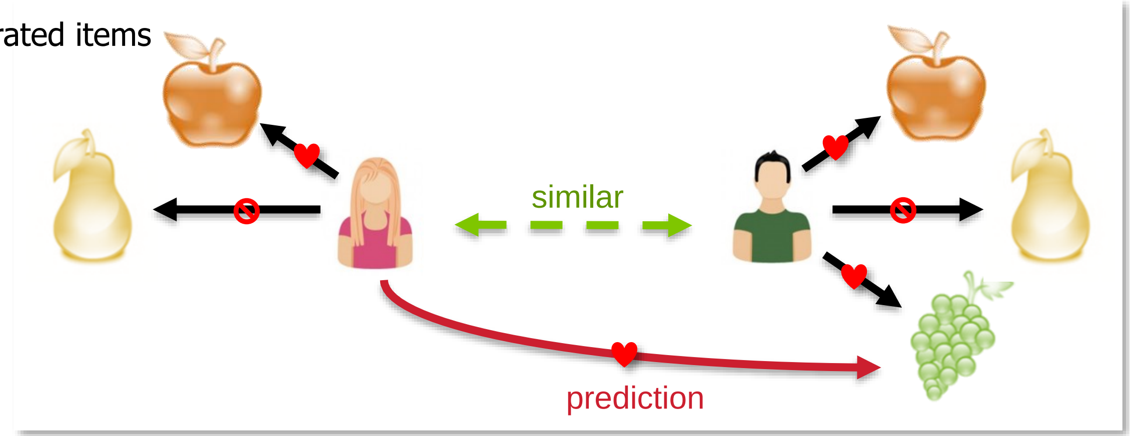


Fraunhofer
FOKUS

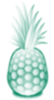
COLLABORATIVE FILTERING

- Item-based Filtering
  - Filters items by taking into account the ratings of all other users
- User-based Filtering (Neighbourhood-based Collaborative Filtering)
  - First searches for most similar users and retrieves their best rated items

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}
u_1	70	21	-	-	33.3	-	-	-	100	-	-	-
u_2	-	-	2	-	25	-	-	-	-	98	-	100
u_3	0	22	82	-	100	0	-	-	-	-	-	100
u_4	-	-	-	-	-	-	-	-	-	-	50	-
u_5	-	15	0	-	25	100	-	-	77	66.6	-	-
u_6	-	-	50	-	-	100	-	66	-	-	-	100



COLLABORATIVE FILTERING

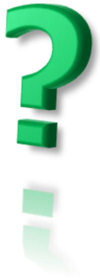
					
Alice	?	2		3	
Bob	5	4	3	5	4
Chris		4		4	4
Diana	4	2	5	3	
Elias	3		1		

COLLABORATIVE FILTERING

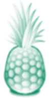
Alice		2	?	3	
Bob	5	4	3	5	4
Chris		4		4	4
Diana	4	2	5	3	
Elias	3		1		

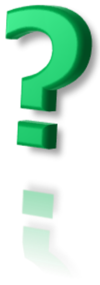
COLLABORATIVE FILTERING

					
Alice		2		3	
Bob	5	4	3	5	4
Chris		4		4	4
Diana	4	2	5	3	
Elias	3	?	1		



COLLABORATIVE FILTERING

					
Alice		2		3	
Bob	5	4	3	5	4
Chris		4	?	4	4
Diana	4	2	5	3	
Elias	3		1		



- Pearson Correlation Coefficient:

$$pSim(u_1, u_2) = \frac{\sum_{i \in I_{u_1 u_2}} (r_{i,u_1} - \overline{r}_{u_1})(r_{i,u_2} - \overline{r}_{u_2})}{\sqrt{\sum_{i \in I_{u_1 u_2}} (r_{i,u_1} - \overline{r}_{u_1})^2 (r_{i,u_2} - \overline{r}_{u_2})^2}}$$

r_{i,u_1} is the rating of user u_1 for the item i of item set $I_{u_1 u_2}$ (a set with existing ratings of both users for each item). $\overline{r}_{u_2}$ is the average rating of all the items of user u_2. (cf. [?, p. 738], [?, p. 153], [?, p. 619])

	1	3.5
	1	?
	4.5	1.5

$\triangleq$

$\neq$

	1	3.5
	1	?
	4.5	1.5

3.5

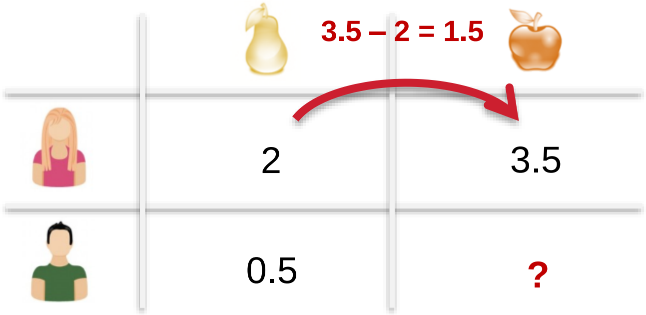
ITEM-BASED COLLABORATIVE FILTERING: RATING PREDICTION

Slope One

- Step 1: Deviation

$$dev(i_1, i_2) = \frac{\sum_{u \in U_{i_1 i_2}} (r_{u,i_1} - r_{u,i_2})}{|U_{i_1 i_2}|}$$

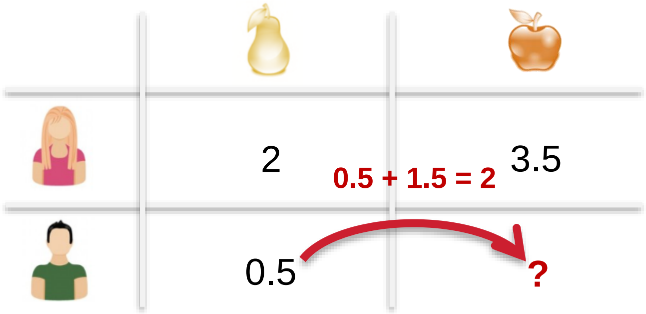
Where i_1 and i_2 are the items, r_{u,i_1} is the items rating user u gave. $U_{i_1 i_2}$ is the set of users who rated both items and $|U_{i_1 i_2}|$ is its cardinality. So the result is the ratings deviation of an item. If the average rating of item i_1 is higher than the average rating of item i_2 the value is positive, if it is lower the value is negative, or if the average ratings are equal the value is $dev(i_1, i_2) = 0$.



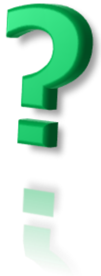
- Step 2: Rating Prediction

$$pre(u, j) = \frac{\sum_{i \in Ij} (r_{u,i} - dev(i,j))}{|Ij|}$$

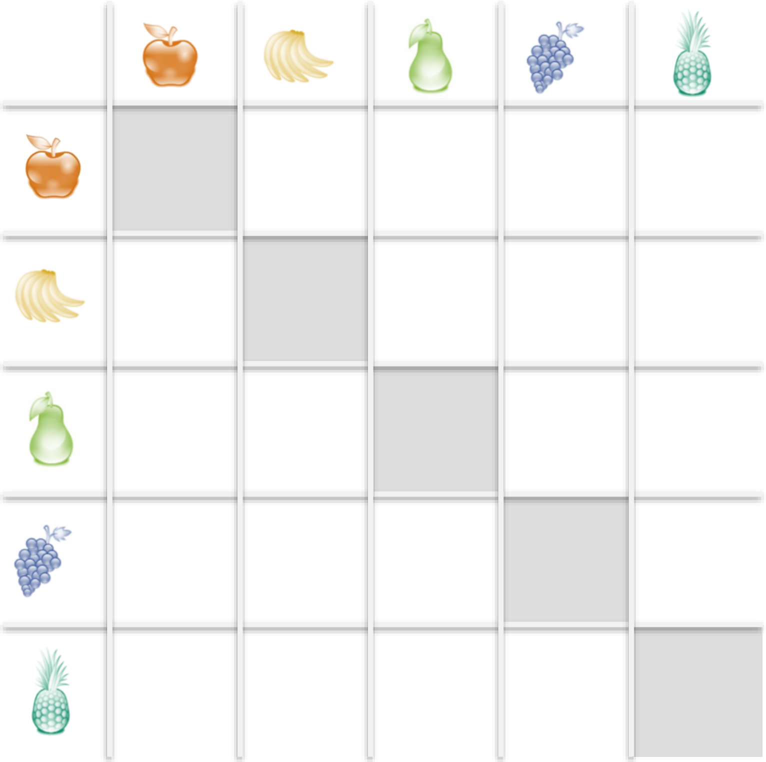
Ij is the set of all relevant items to be compared with item j and $|Ij|$ is its cardinality. The higher $|U_{i_1 i_2}|$, the better the prediction. $r_{u,i}$ is the rating of user u for item i. (cf. [LM05, p. 3]) The resulting value is the predicted rating of this user u for the current item j.

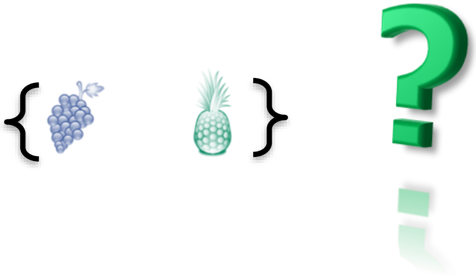


Item-based Collaborative Filtering: Transaction Prediction



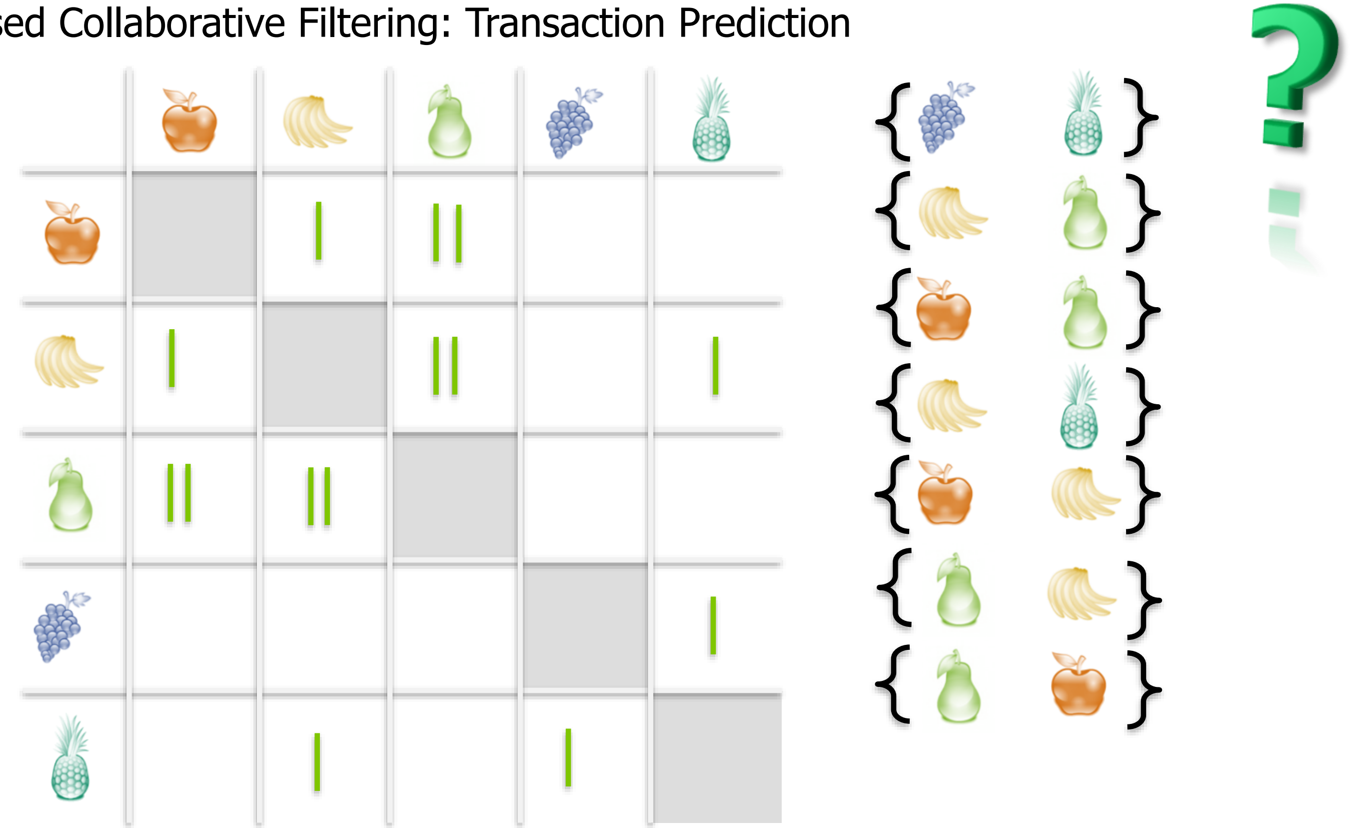
Item-based Collaborative Filtering: Transaction Prediction





Item-based Collaborative Filtering: Transaction Prediction

Item-based Collaborative Filtering: Transaction Prediction



Item-based Collaborative Filtering: Transaction Prediction

CONTENT-BASED FILTERING

1. An element is just compared to another element by respecting their content information on meta data
2. Euclidean Distance

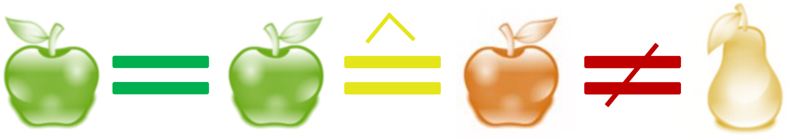
$$eDist(e_1, e_2) = \sqrt{\sum_{i=0}^{N} dist(a_{i,e_1}, a_{i,e_2})}$$

Each attribute value a_{i,e_1} at position i of element e_1 is compared by the distance value to the according attribute value a_{i,e_2} of element e_2. The Euclidean distance $eDist$ is the square root of the summations of these distances. (cf. [?, p. 305])

- Cosine-based Similarity

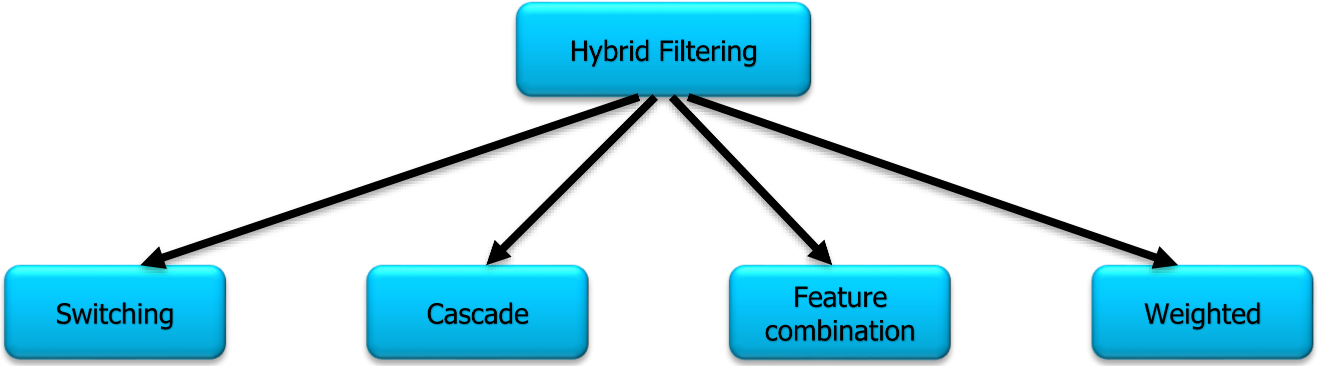
$$cSim(e_1, e_2) = cos(\vec{E}_1, \vec{E}_2) = \frac{\vec{E}_1 \cdot \vec{E}_2}{|\vec{E}_1| \times |\vec{E}_2|}$$

e_1 and e_2 are the elements to be compared like items or users. $\vec{E}_1$ and $\vec{E}_2$ are vectors representing all features of this element, so when n is the number of all attributes of an element, then $\vec{E}_1 = (a_{1,e_1}, a_{2,e_1}, a_{3,e_1}, ..., a_{n,e_1})$. (cf. [?, p. 619], [?, p. 929])



Fraunhofer FOKUS

HYBRID FILTERING



- Weighted Filtering

$$RV = \frac{\sum_{i=1}^{N} w_i * r_i}{\sum_{i=0}^{N} w_i}$$

The recommendation value RV results from the summation of each single recommendation value r multiplied with the according weight w of the N different algorithms and afterwards divided by the summation of all weights.

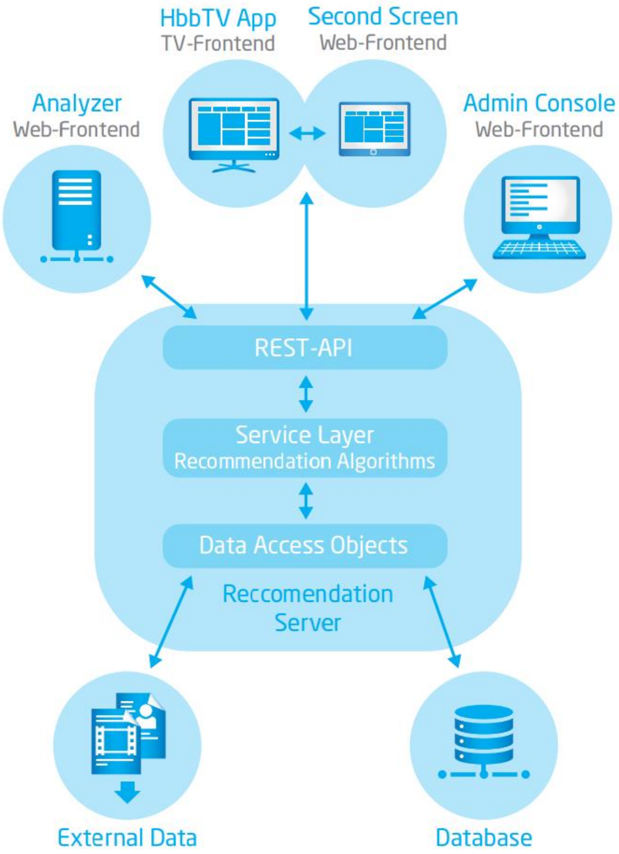
Fraunhofer FOKUS

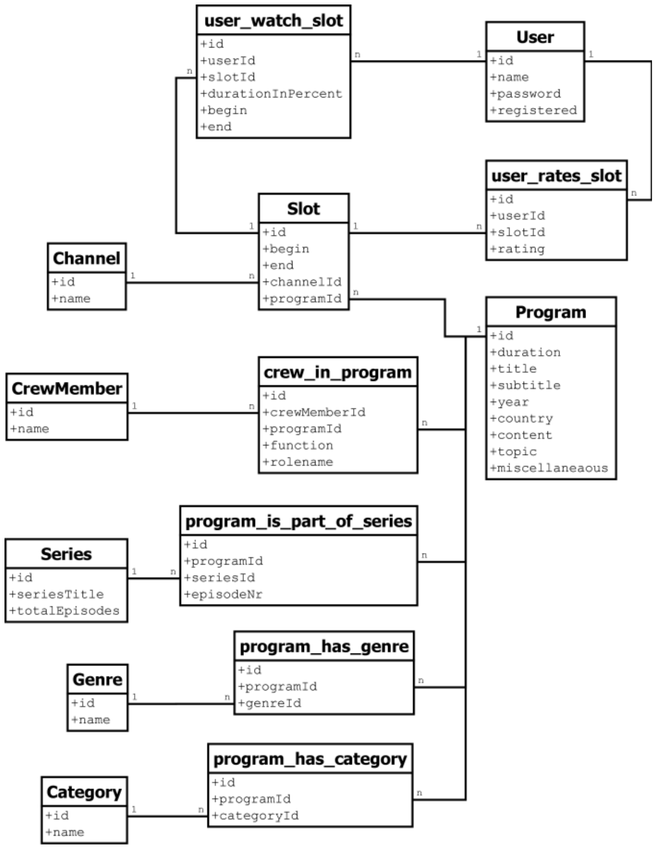
Demonstration: TV Predictor

MASTER THESIS: TV PREDICTOR

- HbbTV based personalized Recommender for TV Programs
- Approaches
  - Content-based Filtering
  - User Similarity
  - Rating Prediction (Slope One)
  - Clustering (K-Means)
  - Training of User Interests (SVM)
  - Association Rules (Apriori)
- Combines Item-Recommendations and Advertisements

MASTER THESIS: TV PREDICTOR – ARCHITECTURE & DATA MODEL

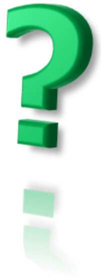




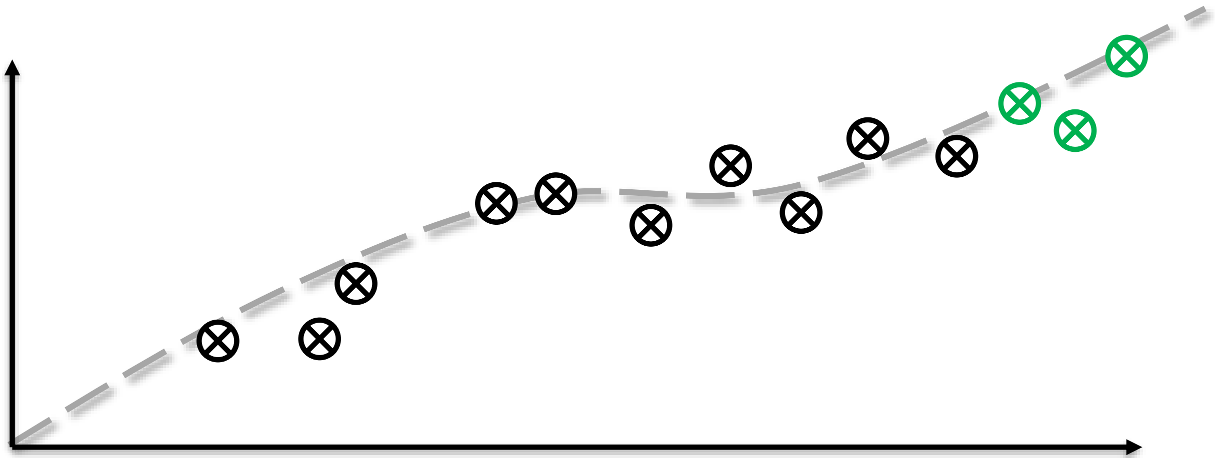
Fraunhofer
FOKUS

Machine Learning

REGRESSION VS. CLASSIFICATION

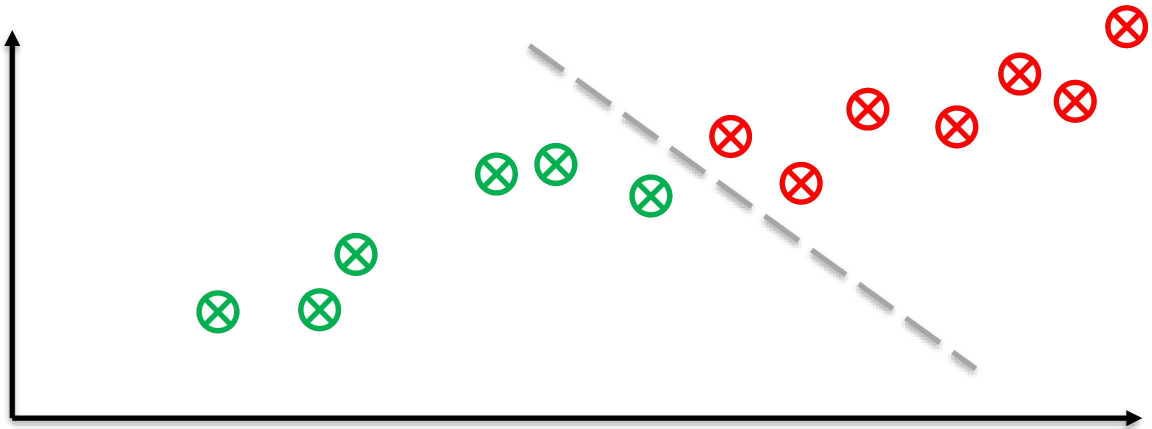


The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.

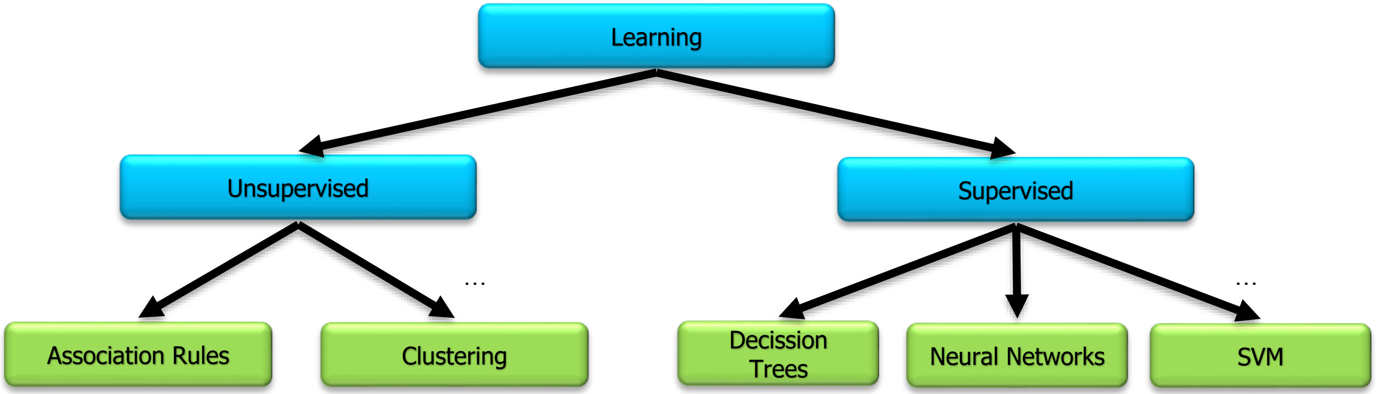


REGRESSION VS. **CLASSIFICATION**

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Popular classification techniques include decision trees and neural networks.



LEARNING ALGORITHMS



Unsupervised algorithms learn without confirmations and try to find patterns and structures of unlabeled data. This is used when it is not clear what aspect or aspects will divide the data. Instances of unsupervised algorithms are Association Rules and Clustering.

Supervised algorithms are using some kind of digital teacher, that confirms or corrects a prediction, for instance by using different training and evaluation data. (cf. [8, p. 437]) Instances of supervised algorithms are Decision Trees, Neural Networks and Support Vector Machines.

UNSUPERVISED - ASSOCIATION RULES

1. Transactions

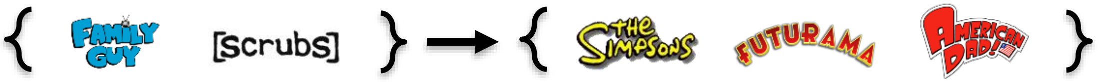
Lars watched:

Stefan watched:

Fabian watched:

2. Identify frequent item sets

3. Create Association Rules



UNSUPERVISED - ASSOCIATION RULES

$$X \implies Y$$

The item set X implies the item set Y. (cf. [Mer10, p. 4-7], [Han06, p. 1455]) Some thresholds are needed to adjust this algorithm. Especially the support value ($supp$ in percent or sometimes as absolute value) is necessary and defines the frequency of an item set.

$$supp(X \implies Y) = \frac{supp(X \cup Y)}{|T|} \geq minsupp$$

$|T|$ is the total number of transactions, so $supp(X \implies Y)$ is the fraction of all transactions that contain X and Y.

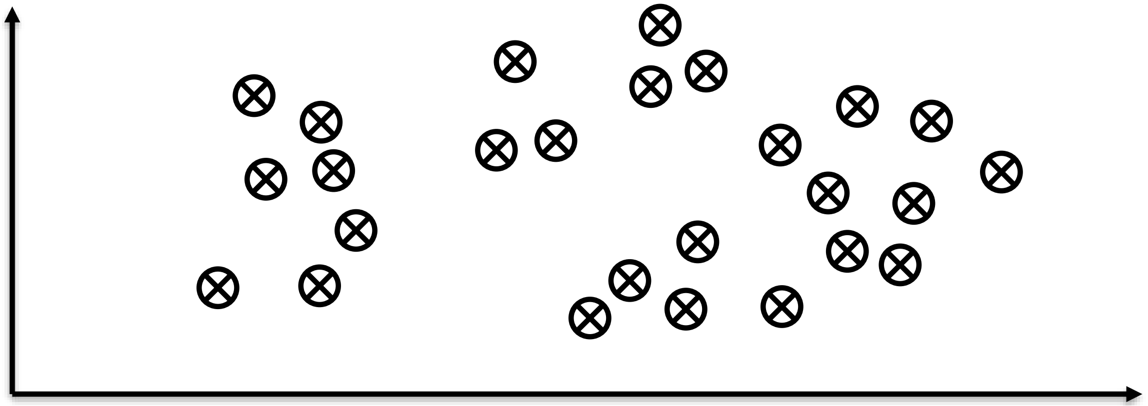
The confidence value expresses the conditional probability of Y knowing X and is defined as:

$$conf(X \implies Y) = \frac{supp(X \cup Y)}{supp(X)} \geq minconf$$

The minimum confidence value ($minconf$) and the minimum support value ($minsupp$) are the according thresholds, so the goal is to find rules with a support value equal to or greater than $minsupp$ and a confidence value equal to or greater than $minconf$.

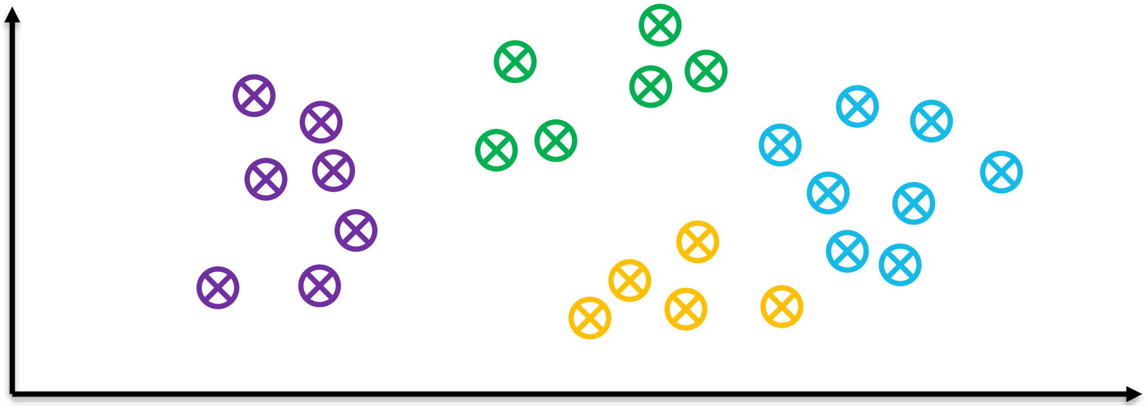
UNSUPERVISED - CLUSTERING

- Data Segmentation
- Devide a set into subsets



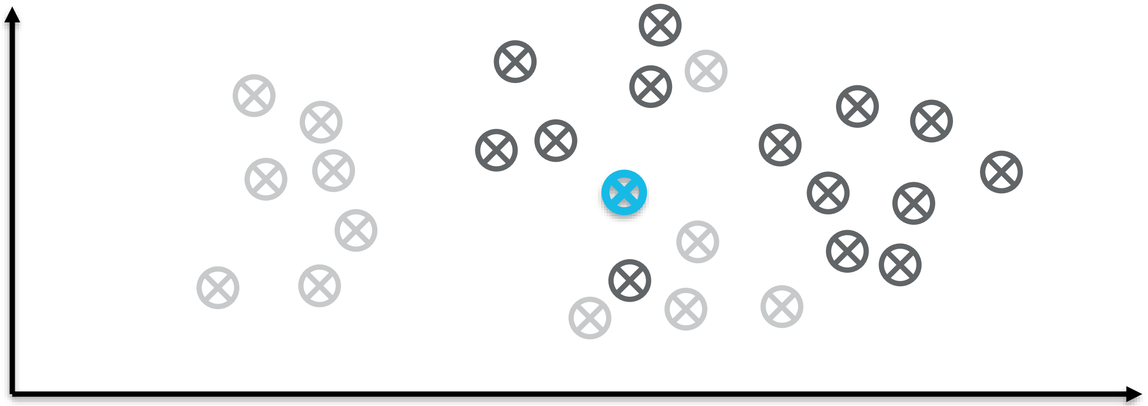
UNSUPERVISED - CLUSTERING

- Data Segmentation
- Devide a set into subsets



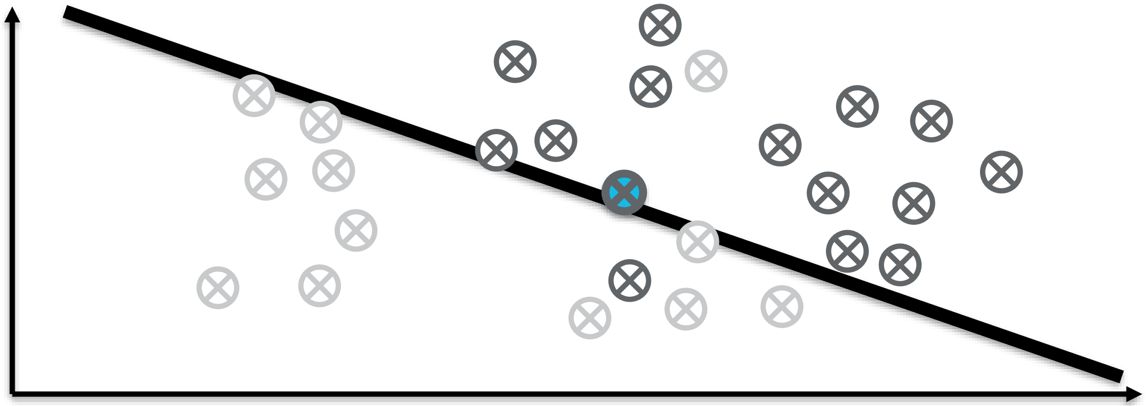
Supervised - Support Vector Machine

- Creates a hyperplane that divides specific space with according points into two spaces



Supervised - Support Vector Machine

- Creates a hyperplane that divides specific space with according points into two spaces



Supervised – SVM – Offline Training of behavioural Patterns

Support Vector Machine

- Creates hyperplanes dividing specific space with according points into two spaces

- Used 109 features for each slot (program)

- Standardized values to range of [0;1]

using only valid intervals

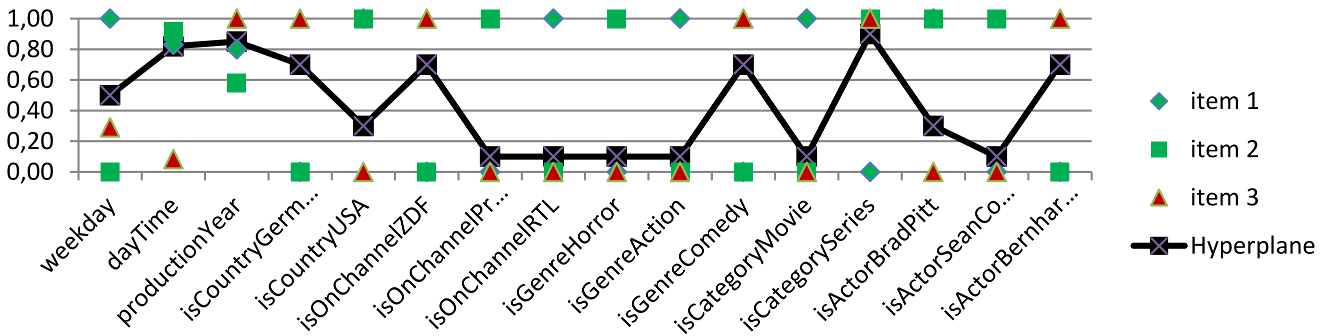
attributes	item 1	item 2	item 3
Liked (rated above average)	yes	yes	no
weekday	Sunday	Monday	Wednesday
dayTime	20:15	22:00	02:00
productionYear	2010	1985	2012
isCountryGermany	no	no	yes
isCountryUSA	yes	yes	no
isOnChannelZDF	no	no	yes
isOnChannelProSieben	no	yes	no
isOnChannelRTL	yes	no	no
isGenreHorror	no	yes	no
isGenreAction	yes	no	no
isGenreComedy	no	no	yes
isCategoryMovie	yes	no	no
isCategorySeries	no	yes	yes
isActorBradPitt	yes	yes	no
isActorSeanConnery	no	yes	no
isActorBernhardHoecker	no	no	yes

Fraunhofer FOKUS

Supervised – SVM – Offline Training of behavioural Patterns

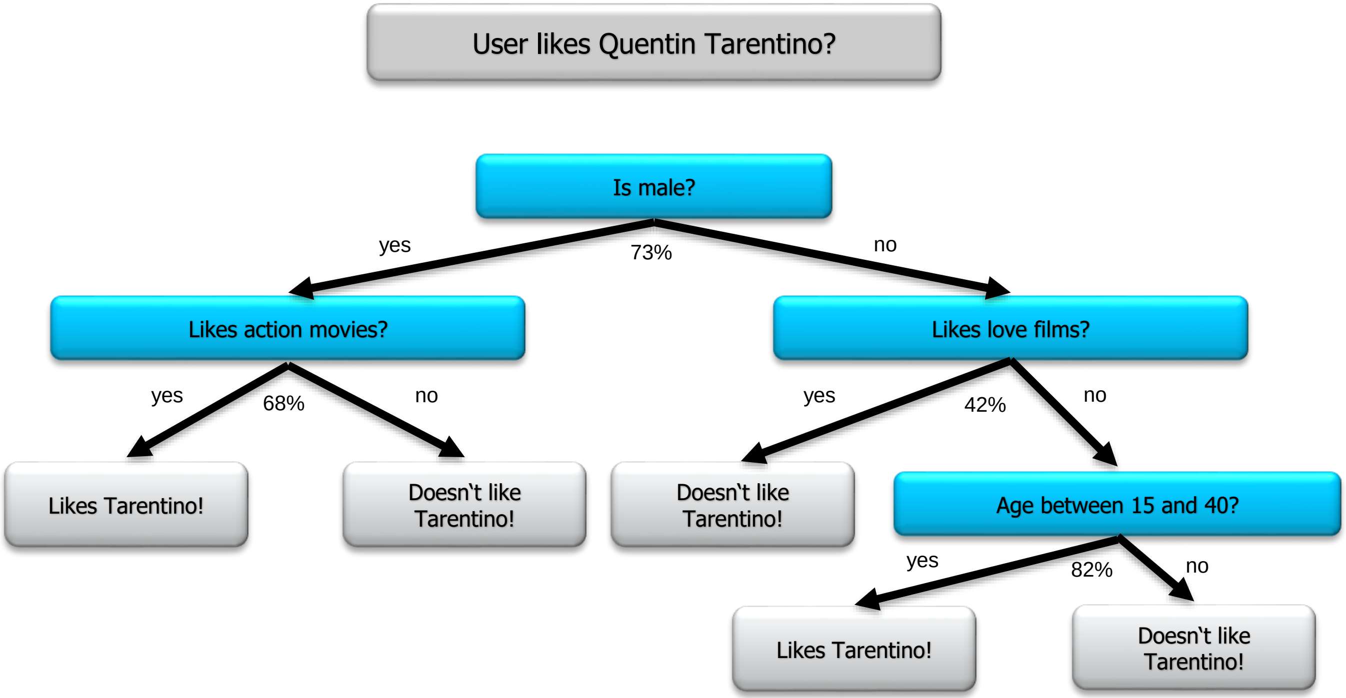
attributes	item 1	item 2	item 3
Liked (rated above average)	yes	yes	no
weekday	Sunday	Monday	Wednesday
dayTime	20:15	22:00	02:00
productionYear	2010	1985	2012
isCountryGermany	no	no	yes
isCountryUSA	yes	yes	no
isOnChannelZDF	no	no	yes
isOnChannelProSieben	no	yes	no
isOnChannelRTL	yes	no	no
isGenreHorror	no	yes	no
isGenreAction	yes	no	no
isGenreComedy	no	no	yes
isCategoryMovie	yes	no	no
isCategorySeries	no	yes	yes
isActorBradPitt	yes	yes	no
isActorSeanConnery	no	yes	no
isActorBernhardHoecker	no	no	yes

attributes	item 1	item 2	item 3	Hyperplane
weekday	1,00	0,00	0,29	0,5
dayTime	0,83	0,92	0,08	0,82
productionYear	0,80	0,58	1,00	0,85
isCountryGermany	0,00	0,00	1,00	0,7
isCountryUSA	1,00	1,00	0,00	0,3
isOnChannelZDF	0,00	0,00	1,00	0,7
isOnChannelProSieben	0,00	1,00	0,00	0,1
isOnChannelRTL	1,00	0,00	0,00	0,1
isGenreHorror	0,00	1,00	0,00	0,1
isGenreAction	1,00	0,00	0,00	0,1
isGenreComedy	0,00	0,00	1,00	0,7
isCategoryMovie	1,00	0,00	0,00	0,1
isCategorySeries	0,00	1,00	1,00	0,9
isActorBradPitt	1,00	1,00	0,00	0,3
isActorSeanConnery	0,00	1,00	0,00	0,1
isActorBernhardHoecker	0,00	0,00	1,00	0,7

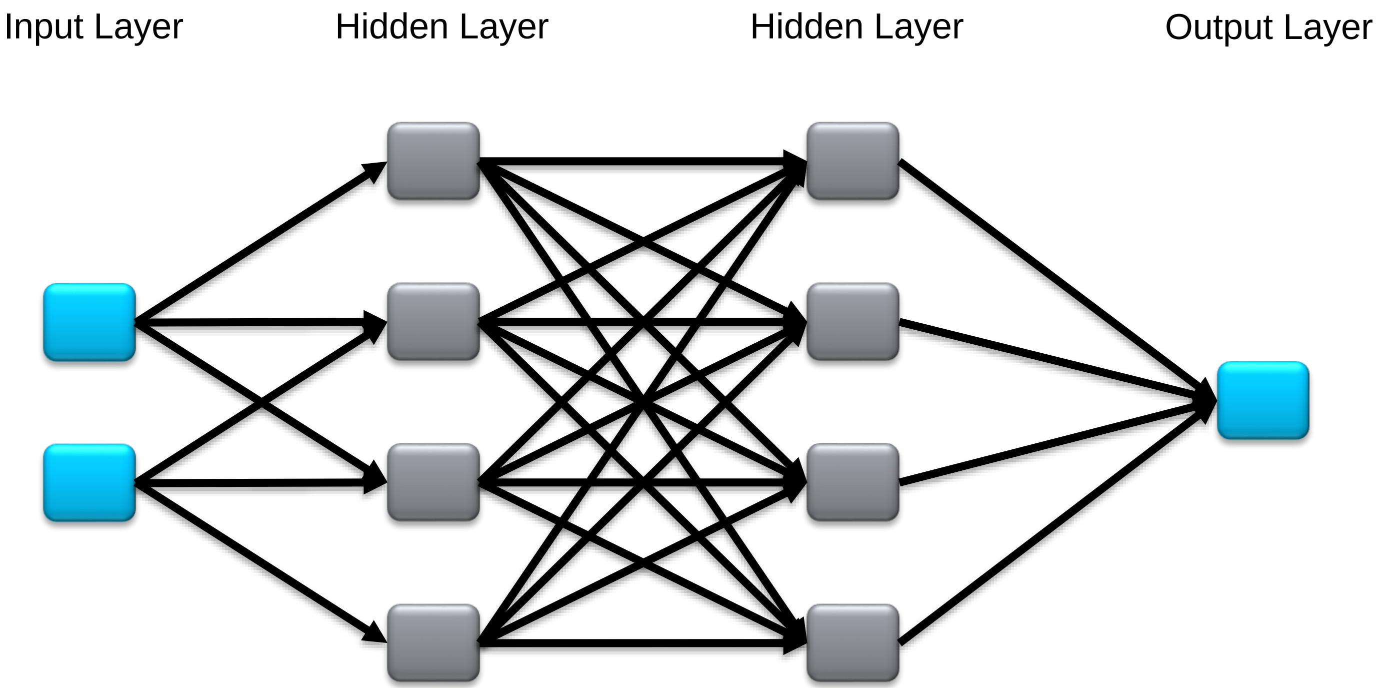


Fraunhofer FOKUS

SUPERVISED - DECISSION TREES



SUPERVISED – DEEP LEARNING: NEURAL NETWORKS



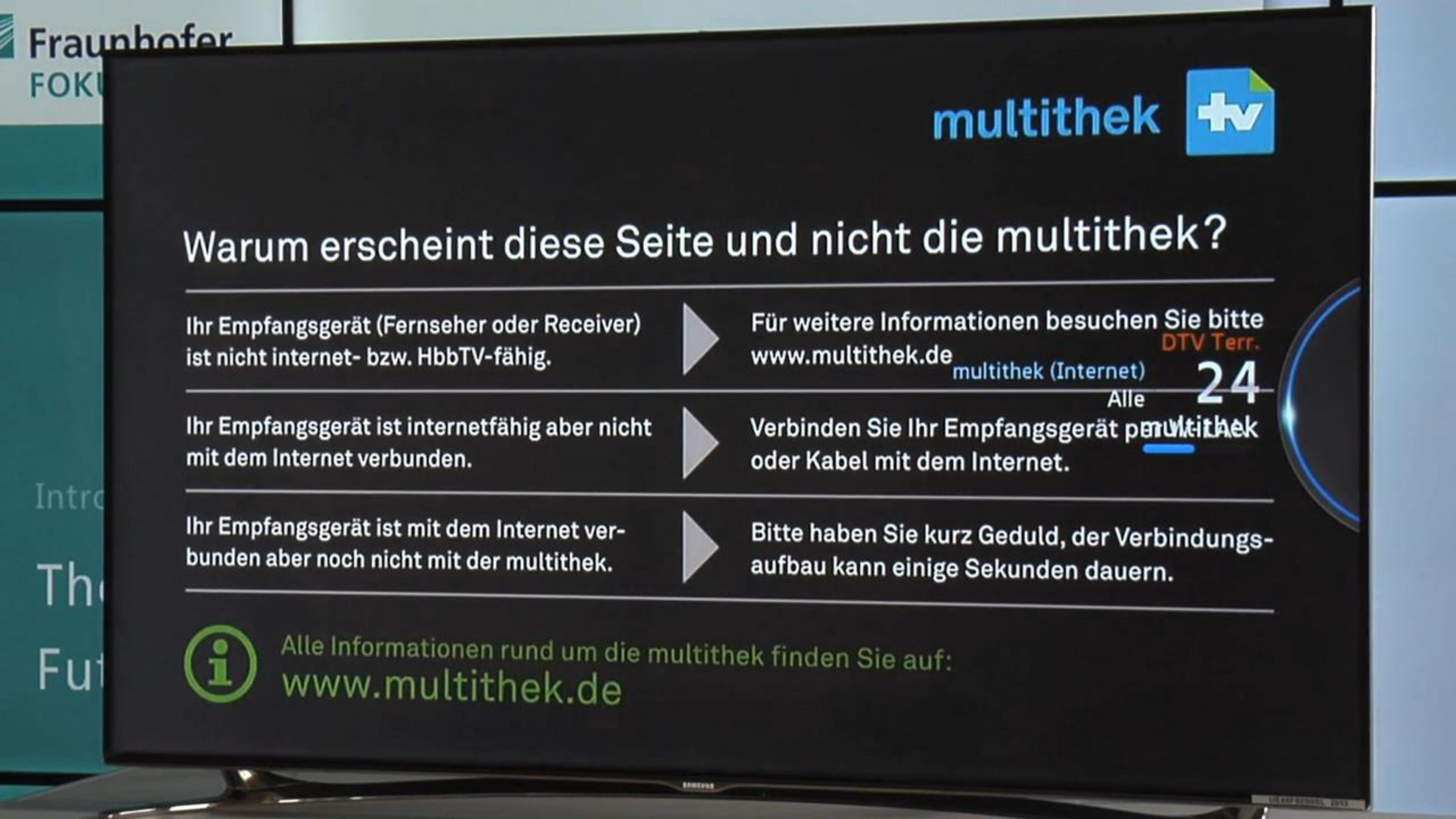
Demonstration: RTV & multithek

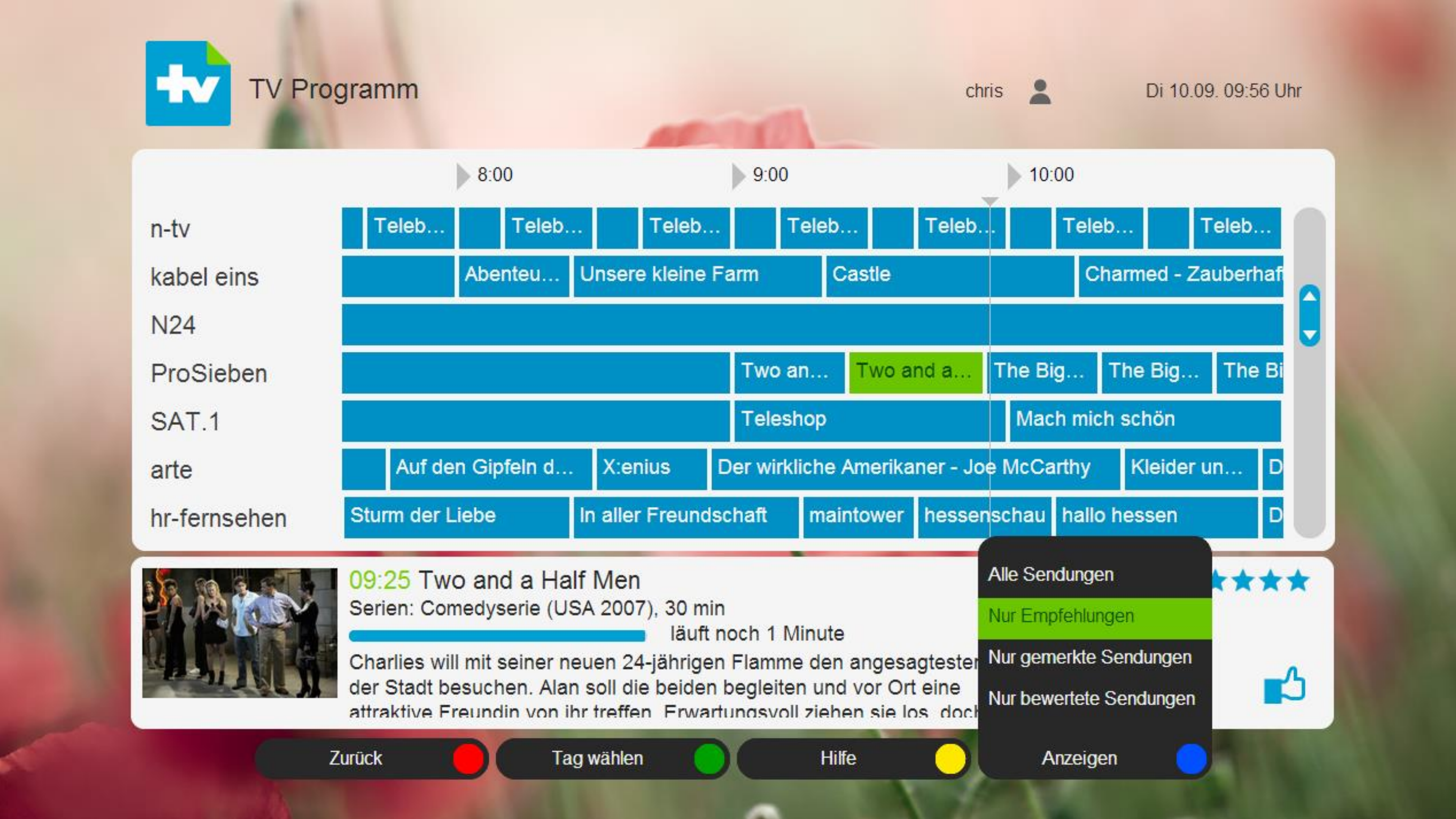
RTV MOVISTO

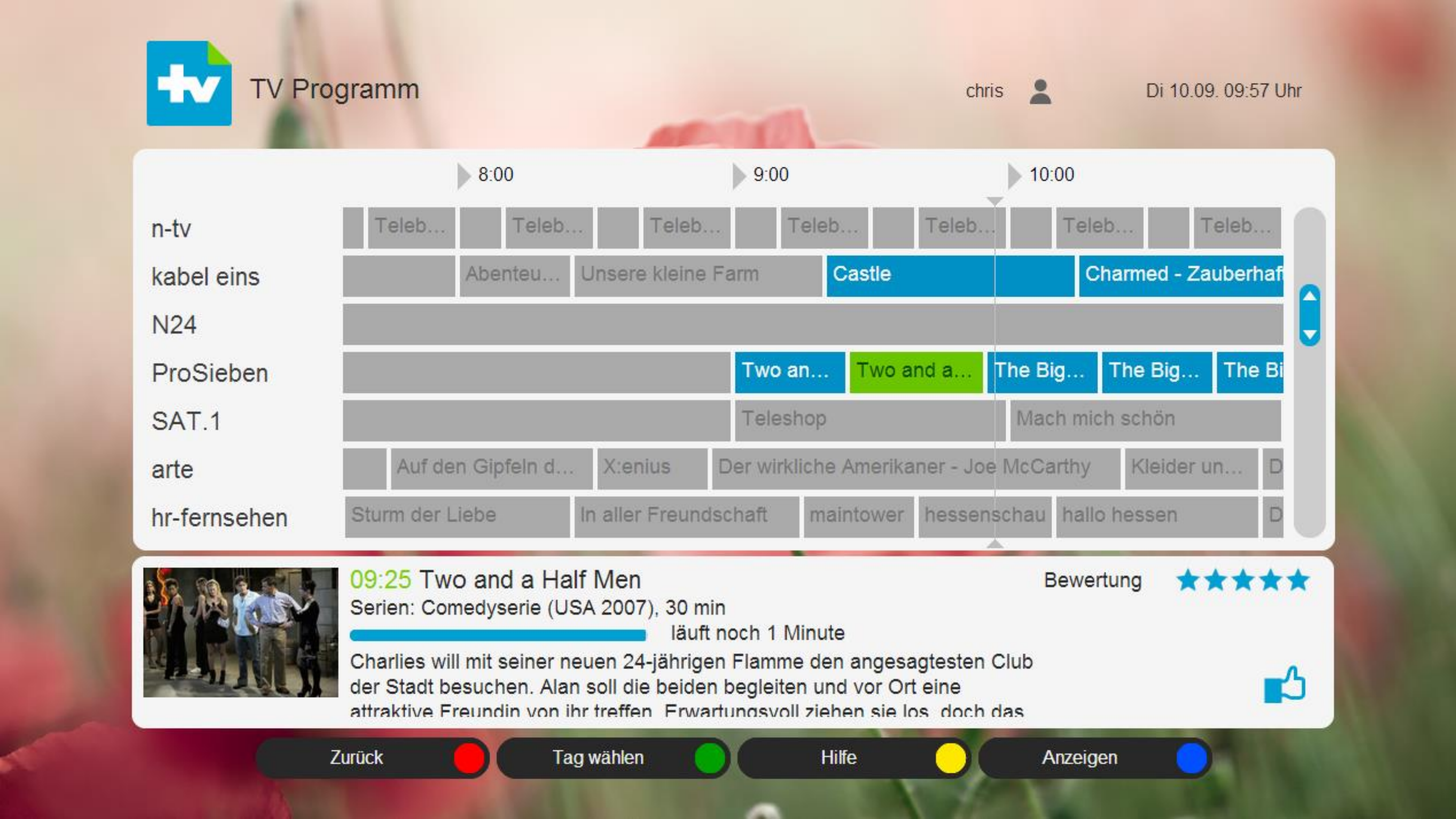
20:49 Samstag, 28. Juli 2012
Frank Seidel
Jetzt im TV
In den nächsten Minuten
Heute Abend
Tages Highlights
Personalisiertes Programm
Napoleon
Das Vierte
20:15 - 22:30
2002, F, GB, D
7,6
Underdogs
ARTE
20:45 - 22:15
2007, D
6,3
VoD
Peace Virus – Die Bedrohung
maxdome
VoD
2002, USA
6,1
Das einfache Mädchen
3sat
20:15 - 21:45
1957, D
5,4
NEU
8,9
Sendungs-Info
Ähnliche Sendungen

RTV.DE

Samstag, 05. Januar 2013
13:15 Uhr
Ihr Jahreshoroskop 2013
Guten Tag chriskrauss! | Abmelden
Hilfe
mein rtv
meine Merkliste (0)
Partner von TVinfo HD
Sendungssuche
Home
Jetzt im TV
20:15 im TV
Spielfilme
Serien
Krimis
Sport
Dokumentation
Dritte
Pay-TV
Movisto® Beta
PROGRAMM
RATGEBER IM TV
STARS & AKTUELLES
MITMACHEN
GEWINNSPIELE
Movisto
Empfehlungsliste
Bewertungshistorie
Jetzt im TV
Demnächst
ab 20:15 Uhr
ab 22:15 Uhr
Tages-Highlights
Neujahrskonzert der Wien...
3sat
20:15 - 22:45 Uhr
9.4
Prosit Neujahr! Diesen traditionellen Gruß an die Welt entbieten die Wiener Philharmoniker unter der Leitung von Franz Welser-Möst, der für das Neujahrskonzert bereits zum zweiten Mal
Ugly Americans
VIVA
21:15 - 21:50 Uhr
8.2
ZDF-History
Phoenix
21:45 - 22:30 Uhr
7.9
Secretary - Wo...
Einsfestival
22:00 - 23:45 Uhr
7.6
Open Range - ...
sixx
08:15 - 10:35 Uhr
7.4
Königreich der...
VOX
20:15 - 22:45 Uhr
6.7
Das Dorf der V...
kabel1 classics
20:15 - 21:50 Uhr
6.4
Werden Sie Fan von rtv.de
rtv.de auf Facebook
Gefällt mir 4.377
4,377 Personen gefällt rtv.de.
Winie
Monika
Christina
Ste
Mimi
Jonas
Mark
EnergieMes
WebtvOne
Diana
Soziales Plug-in von Facebook
Die rtv-App fürs Handy

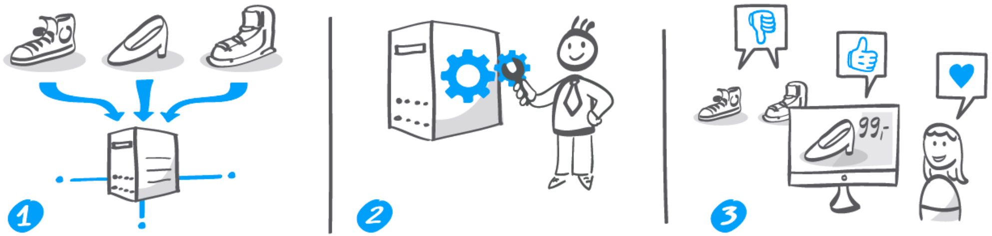

multithek +tv
Warum erscheint diese Seite und nicht die multithek?
Ihr Empfangsgerät (Fernseher oder Receiver) ist nicht internet- bzw. HbbTV-fähig.
Für weitere Informationen besuchen Sie bitte www.multithek.de
Ihr Empfangsgerät ist internetfähig aber nicht mit dem Internet verbunden.
Verbinden Sie Ihr Empfangsgerät per WLAN oder Kabel mit dem Internet.
Ihr Empfangsgerät ist mit dem Internet verbunden aber noch nicht mit der multithek.
Bitte haben Sie kurz Geduld, der Verbindungsaufbau kann einige Sekunden dauern.
Alle Informationen rund um die multithek finden Sie auf:
www.multithek.de
DTV Terr.
multithek (Internet)
24
Alle
multithek


TV Programm
chris
Di 10.09. 09:56 Uhr
8:00
9:00
10:00
n-tv
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
kabel eins
Abenteu…
Unsere kleine Farm
Castle
Charmed - Zauberhaf
N24
ProSieben
Two an…
Two and a…
The Big…
The Big…
The Bi
SAT.1
Teleshop
Mach mich schön
arte
Auf den Gipfeln d…
X:enius
Der wirkliche Amerikaner - Joe McCarthy
Kleider un…
D
hr-fernsehen
Sturm der Liebe
In aller Freundschaft
maintower
hessenschau
hallo hessen
D
09:25 Two and a Half Men
Serien: Comedyserie (USA 2007), 30 min
läuft noch 1 Minute
Charlies will mit seiner neuen 24-jährigen Flamme den angesagtester
der Stadt besuchen. Alan soll die beiden begleiten und vor Ort eine
attraktive Freundin von ihr treffen. Erwartungsvoll ziehen sie los, doch
Alle Sendungen
Nur Empfehlungen
Nur gemerkte Sendungen
Nur bewertete Sendungen
Anzeigen
Zurück
Tag wählen
Hilfe


TV Programm
chris
Di 10.09. 09:57 Uhr
8:00
9:00
10:00
n-tv
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
Teleb…
kabel eins
Abenteu…
Unsere kleine Farm
Castle
Charmed - Zauberhaf
N24
ProSieben
Two an…
Two and a…
The Big…
The Big…
The Bi
SAT.1
Teleshop
Mach mich schön
arte
Auf den Gipfeln d…
X:enius
Der wirkliche Amerikaner - Joe McCarthy
Kleider un…
D
hr-fernsehen
Sturm der Liebe
In aller Freundschaft
maintower
hessenschau
hallo hessen
D
09:25 Two and a Half Men
Bewertung
Serien: Comedyserie (USA 2007), 30 min
läuft noch 1 Minute
Charlies will mit seiner neuen 24-jährigen Flamme den angesagtesten Club
der Stadt besuchen. Alan soll die beiden begleiten und vor Ort eine
attraktive Freundin von ihr treffen. Erwartungsvoll ziehen sie los, doch das
Zurück
Tag wählen
Hilfe
Anzeigen

What is Context?
„Processing Semantics“

CONTEXT-AWARE RECOMMENDATIONS

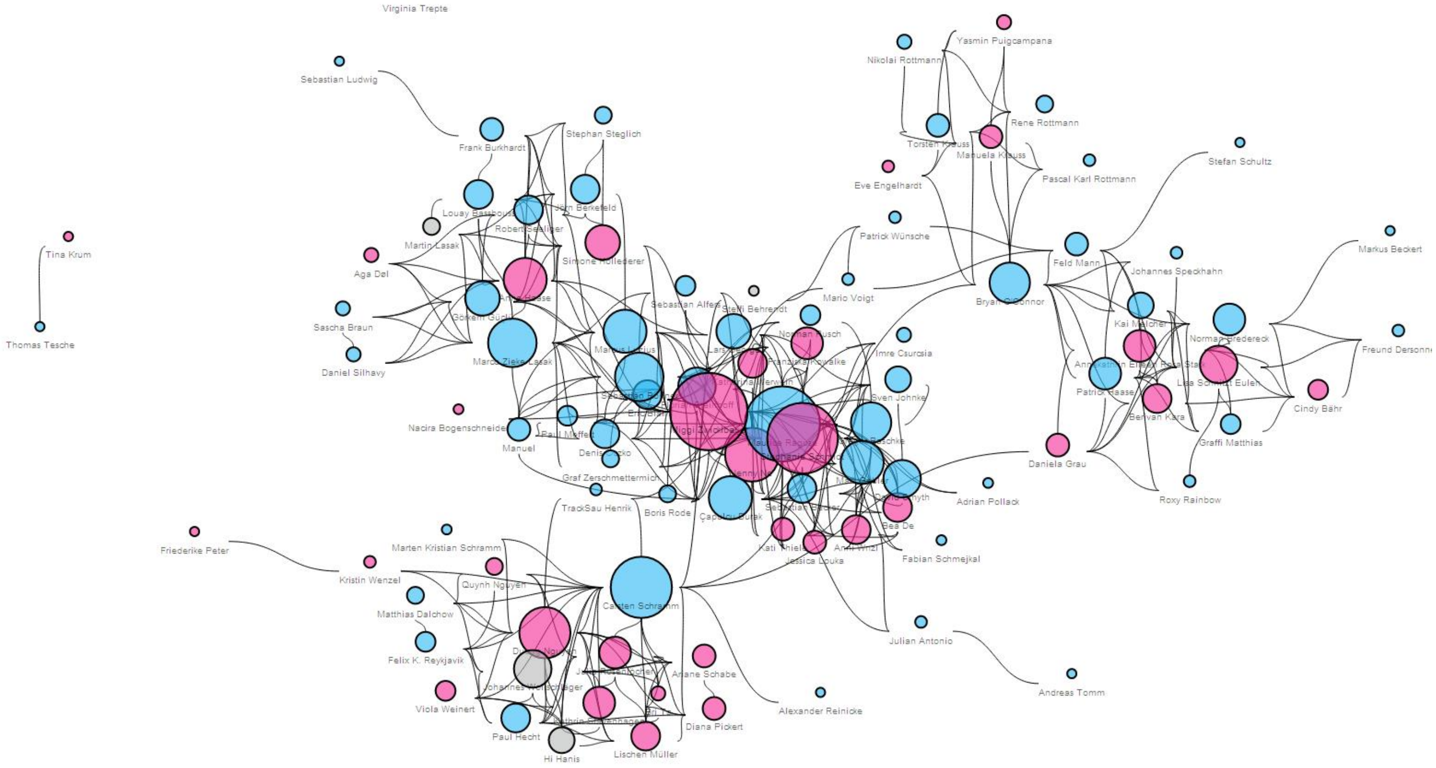


Processing Context-Data

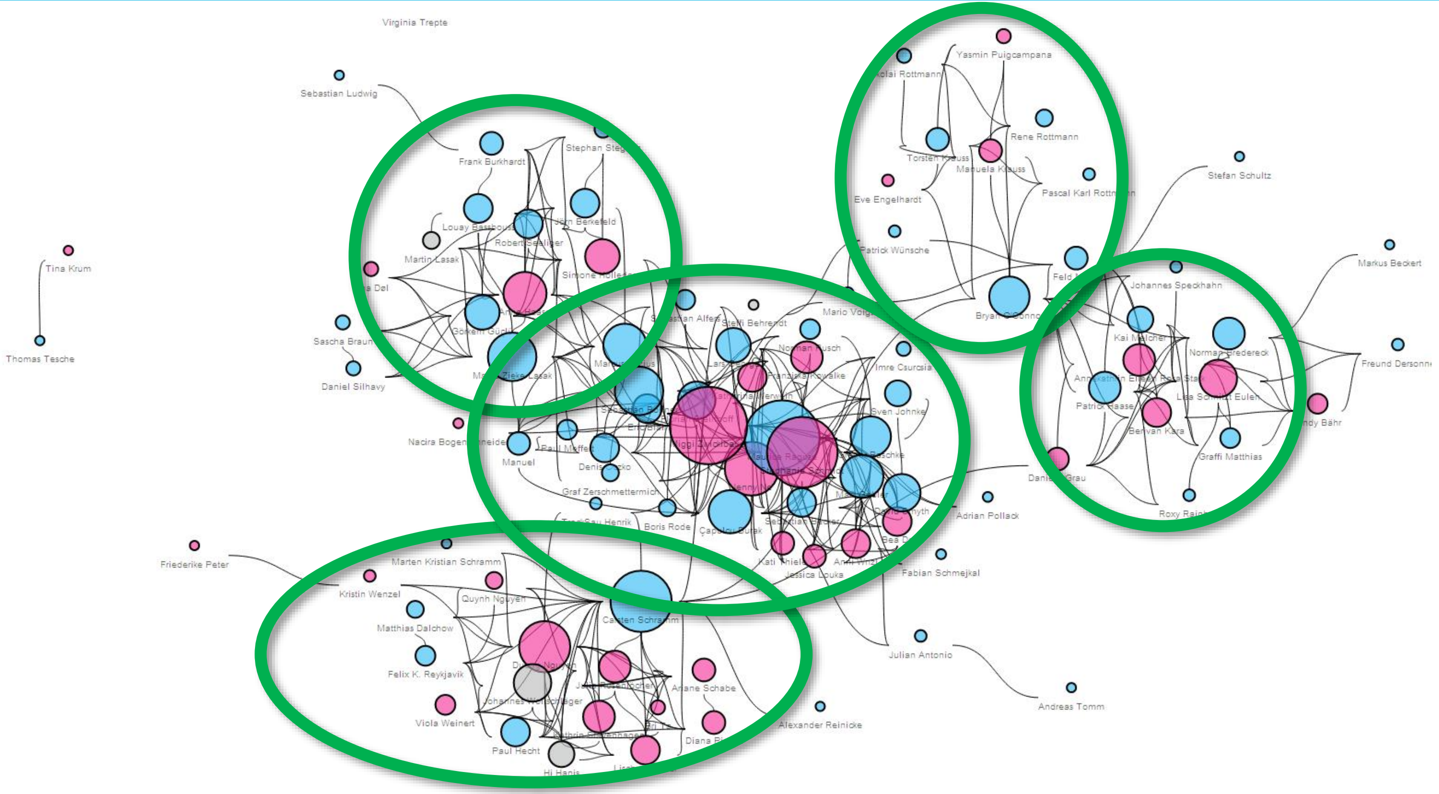
- User(s)
- Locations
- Time → Momentum
- Activities
- Social Interactions
- Moods
- External Influences
- Temporal Trends

Fraunhofer FOKUS

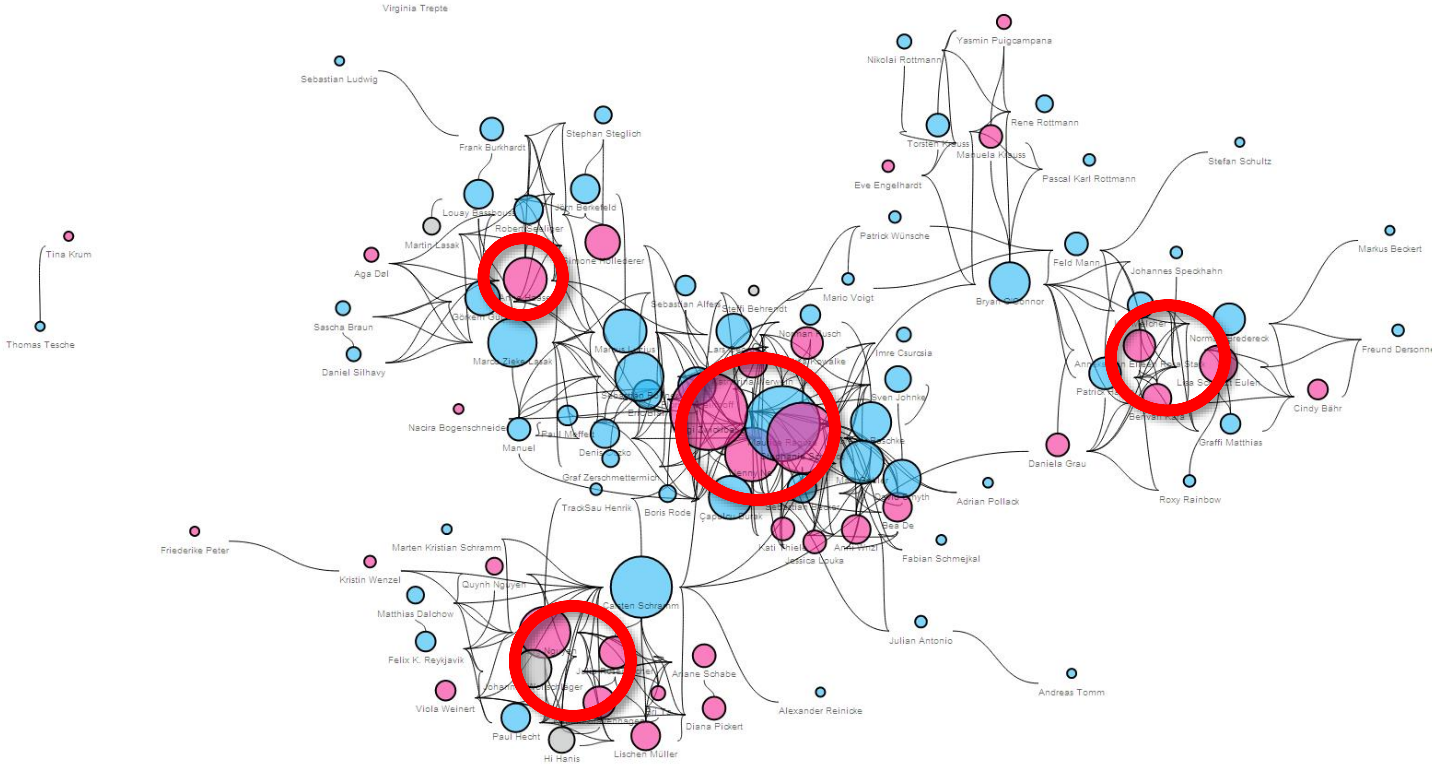
SOCIAL NETWORK ANALYSIS



COMMUNITY DETECTION

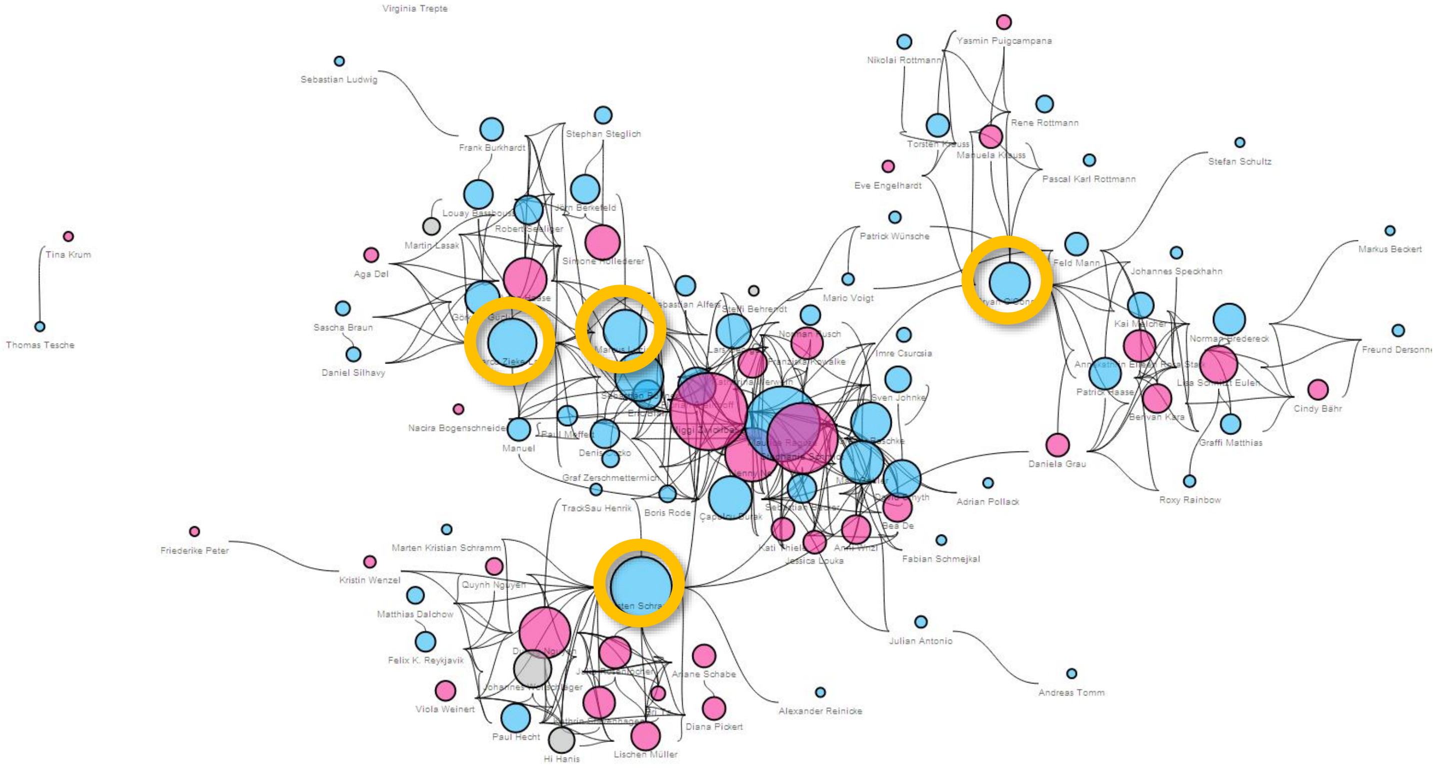


CENTRALITY



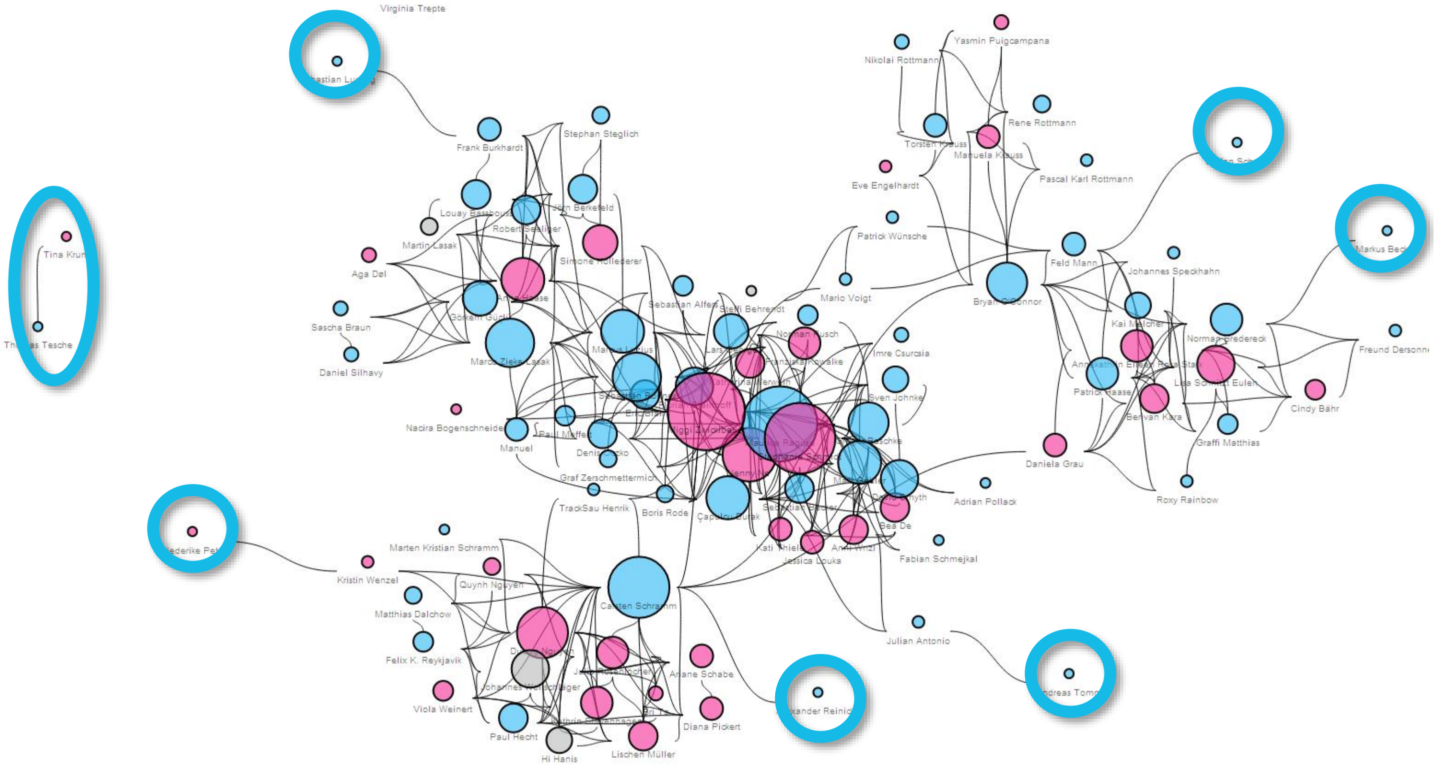
Virginia Trepte
Sebastian Ludwig
Frank Burkhardt
Stephan Steglich
Louay Bassbouss
Martin Lasak
Aga Dal
Tina Krum
Thomas Tesche
Sascha Braun
Daniel Silhavy
Nacira Bogenschneider
Manuel
Graf Zerschmettermich
TrackSau Henrik
Boris Rode
Sebastian Alfers
Steffi Behrendt
Mario Voigt
Patrick Wünsche
Eve Engelhardt
Nikolai Rottmann
Yasmin Puigcampana
Rene Rottmann
Pascal Karl Rottmann
Stefan Schultz
Bryan O'Connor
Feld Mann
Johannes Speckhahn
Markus Beckert
Freund Dersonn
Cindy Bähr
Graffi Matthias
Roxy Rainbow
Daniela Grau
Adrian Pollack
Imre Csurcsia
Sven Johnke
Fabian Schmejkal
Julian Antonio
Andreas Tomm
Alexander Reinicke
Jessica Louka
Bea De
Friederike Peter
Kristin Wenzel
Marten Kristian Schramm
Quynh Nguyen
Matthias Dalchow
Felix K. Reykjavik
Viola Weinert
Paul Hecht
Hi Hanis
Lischen Müller
Diana Pickert
Ariane Schabe

BETWEENNESS/ CLOSENESS



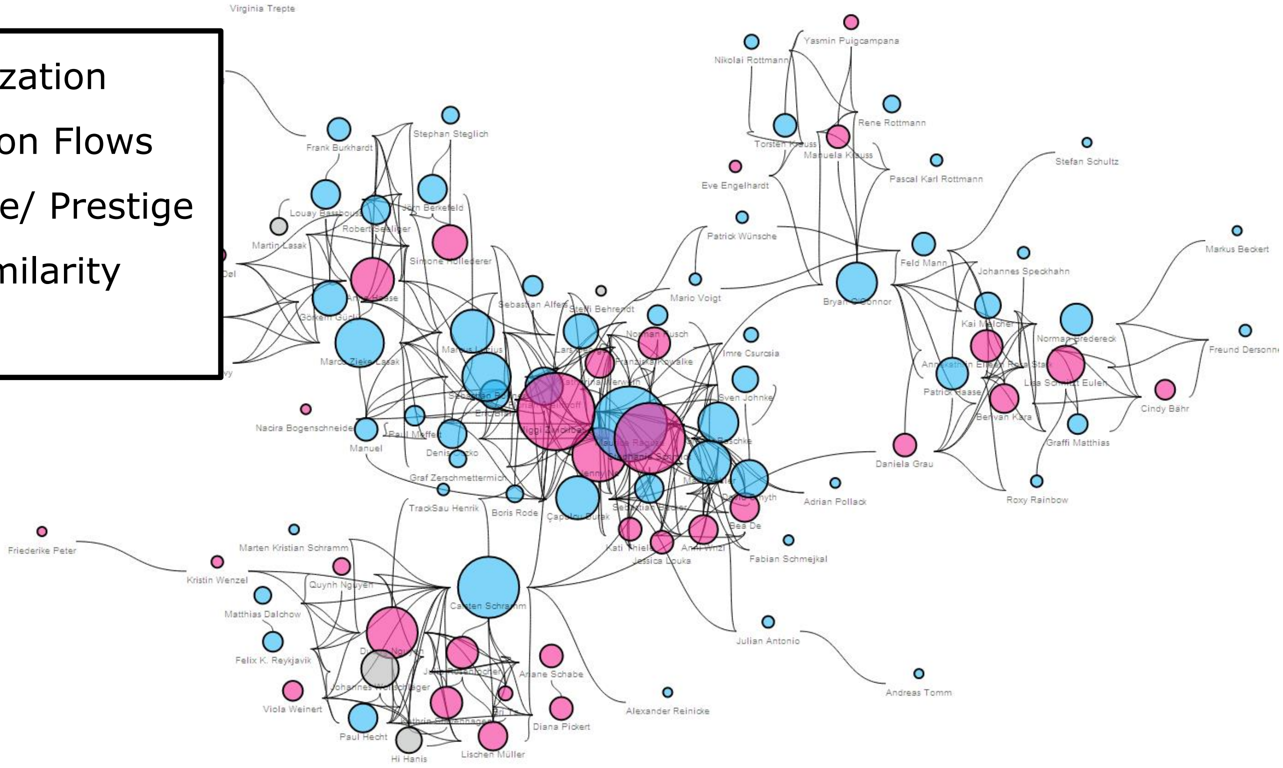
Fraunhofer
FOKUS

OUTLIER DETECTION



MORE INFORMATION

- Social Organization
- Communication Flows
- User Influence/ Prestige
- Structural Similarity
- Cohesion



Semantic Analysis in Social Networks

>> Robert wrote:

The magic ring of invisibility was found by the Hobbit Bilbo Baggins. His nephew Frodo tries to destroy the ring in Mordor to keep Sauron, the Dark Lord, from fulfilling his design to win domination over Middle-earth.

While "The Hobbit" isn't a bad movie, it isn't good either.

Semantic Analysis in Social Networks

>> Robert wrote:

The magic ***ring*** *of invisibility was found by the* ***Hobbit*** *Bilbo Baggins. His nephew Frodo tries to destroy the ring in* ***Mordor*** *to keep* ***Sauron****, the Dark Lord, from fulfilling his design to win domination over Middle-earth.*

While "The Hobbit" isn't a bad movie, it isn't good either.

Semantic Analysis in Social Networks

>> Robert wrote:

The magic ***ring*** *of invisibility was found by the* ***Hobbit*** *Bilbo Baggins. His nephew Frodo tries to destroy the ring in* ***Mordor*** *to keep* ***Sauron****, the Dark Lord, from fulfilling his design to win domination over Middle-earth.*

While "The Hobbit" isn't a bad movie, it isn't good either.

Sentiment Analysis

Semantic Keyword Detection

Semantic Analysis in Social Networks

>> Robert wrote:

*The magic **ring** of invisibility was found by the **Hobbit** Bilbo Baggins. His nephew Frodo tries to destroy the ring in **Mordor** to keep **Sauron**, the Dark Lord, from fulfilling his design to win domination over Middle-earth.*

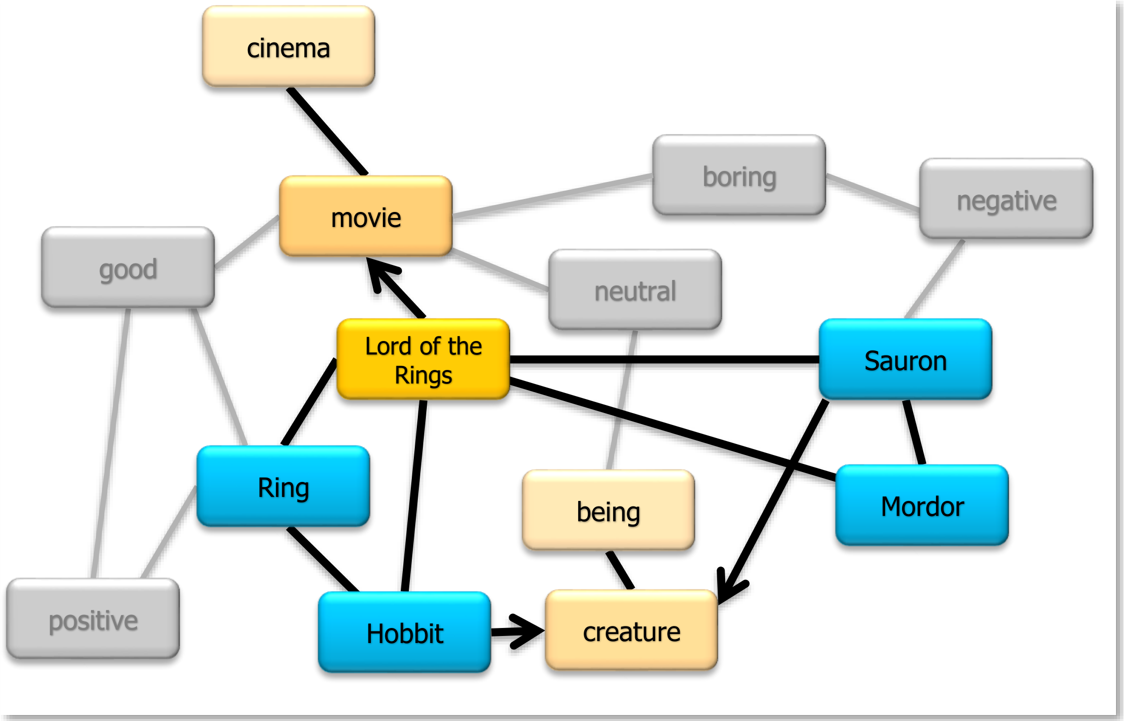
While "The Hobbit" isn't a bad movie, it isn't good either.

Sentiment Analysis

Semantic Keyword Detection

Roberts Profile in a Recommender System

Semantic Analysis (Automatic keyword extraction in semantic graphs)

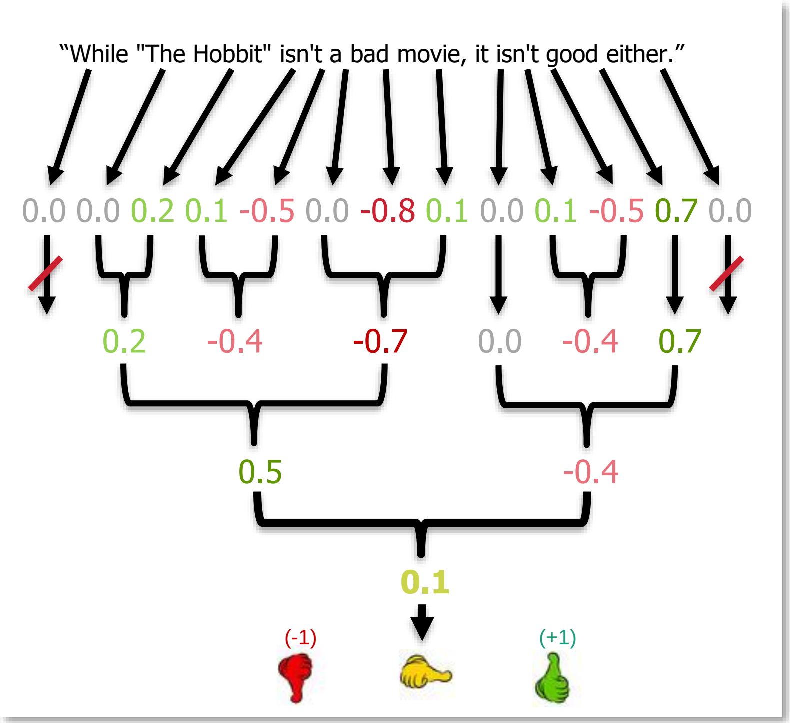


- **Synonymy** is the equality in meaning of two words.
- **Antonymy** stands for the opposite in meaning of two words.
- **Polysemy** characterizes a word with multiple different semantics, but a unique word origin.
- **Homonymy** characterizes the equality in pronunciation and spelling, but with a different semantic of two words. The origins of both words are diverse.
- **Hyperonymy** and **Hyponymy** are to describe a hierarchic relationship between a superordinate (hyperonym) and a subtopic (hyponym).
- **Holonymy** and **Meronymy** are to describe a hierarchic relationship between the whole thing (holonym) and a subset (meronym).
- **Associations** are unspecific relations of words sharing a common context. The definition of further meta data allow to concretize the type of relation.

Fraunhofer FOKUS

Semantic Analysis (Sentiment Detection e.g. in social comments)

- **Removal of stop words** to increase the systems performance
- **Stemming**: to reduce a word to its word stem, the so called "lemma"
- **Part of Speech Tagging**: Allocate a word to its part of speech (POS)
- **Identify Compounds** and Statistical Phrases to treat them as units
- **Compound Splitting** for differing multiple semantics of a concatenated word
- **Chunking** and **Shallow Parsing** separate words, sentences or whole text corpora into sub sets representing only cohesive topics and semantics
- **Word Sense Disambiguation** aims at identifying the correct sense of homonyms
- **Sentiment Analysis** is an approach to allocate the author's opinion of his writing to a numerical value



Technical Issues in Recommendation Systems

ACCURACY OF ALGORITHMS

- Mean Absolute Error

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

- Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} |p_i - q_i|^2}{N}}$$

n-fold cross validation to summarize results:

Original data set

1.	Prediction data set	Training data set	Training data set
2.	Training data set	Prediction data set	Training data set
3.	Training data set	Training data set	Prediction data set

NETFLIX PRIZE

“To qualify for the $1,000,000 Grand Prize, the accuracy of your submitted predictions on the qualifying set must be at least 10% better than the accuracy Cinematch can achieve on the same training data set at the start of the Contest.”*

“On September 21, 2009 we awarded the $1M Grand Prize to team ‘BellKor’s Pragmatic Chaos’.”*

*http://www.netflixprize.com/

RECOMMENDATION SYSTEMS IN DISTRIBUTION

NETFLIX

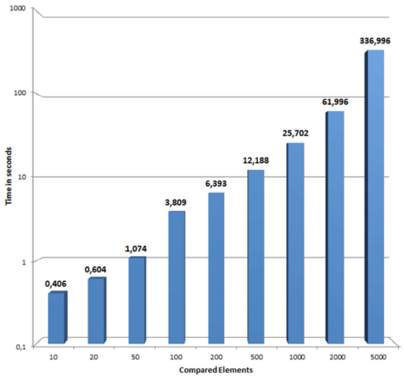
- Approx. 75% of view generated through recommendations
- 40 employees tag TV-Shows and movies by hand (80.000 subgenres)
- Usage of a variety of algorithms
- Several usage parameters being stored and analysed:
  - What has been watched
  - What was searched for
  - Ratings
  - Time, Date, day of the week
  - Devices
  - User interaction like scrolling and browsing

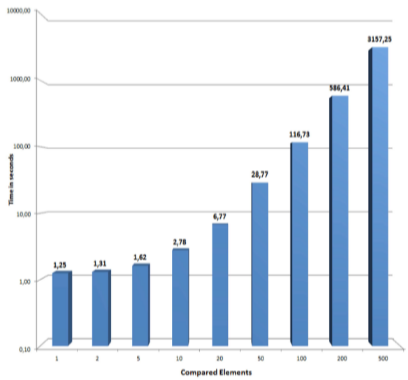
source: Carlos Gomez-Uribe, VP of product innovation and personalization algorithms
and Xavier Amatriain, engineering director (Interview on http://www.wired.com/2013/08/qq_netflix-algorithm/)

LIMITING BEHAVIOR/ PERFORMANCE/ COMPLEXITY/ SCALABILITY

1. Big O-Analysis (O-Notation)

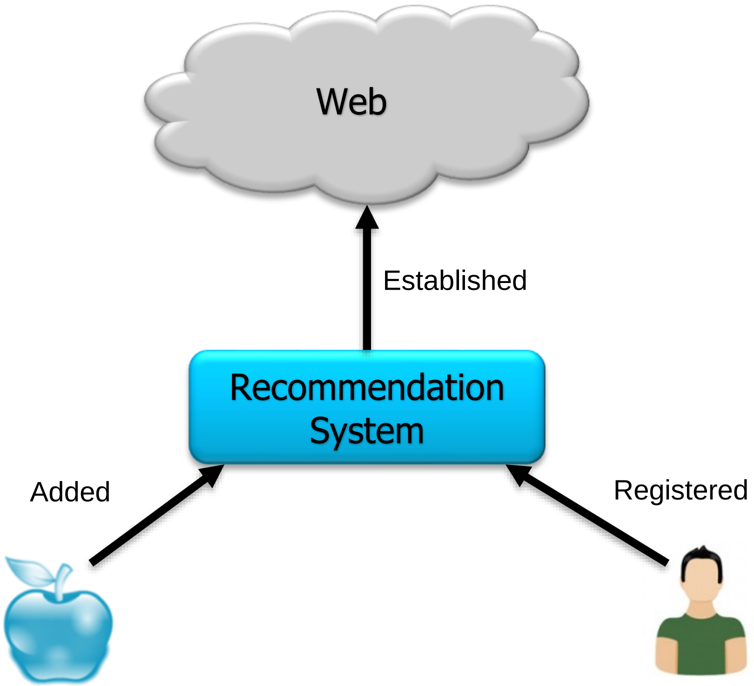
- Problem size-independent characteristics: $O(1)$
- Logarithmical growth: $O(log(n))$
- Proportional growth: $O(n)$
- Polynomial growth: $O(n^c)$
- Exponential growth: $O(c^n)$





Fraunhofer FOKUS

COLD START PROBLEM – NEW ITEM/ NEW USER PROBLEM



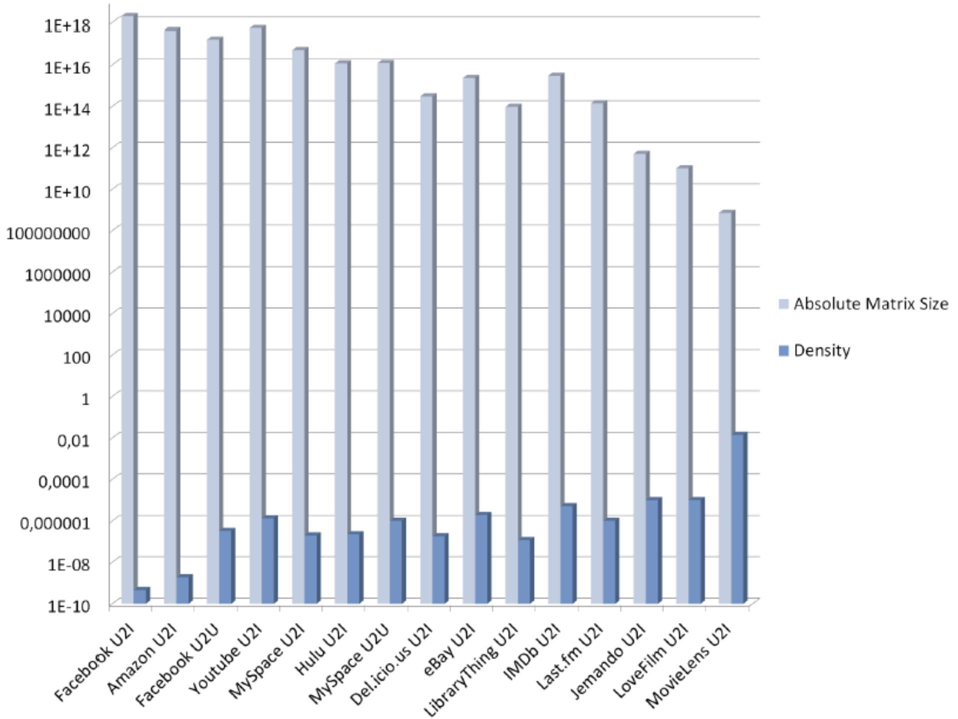
SPARSITY PROBLEM

- Density

$$dens = \frac{|F|}{|R| * |C|}$$

While $|F|$ is the amount of all feedbacks, $|R|$ and $|C|$ are the amount of the rows (e.g. Users) and the amount of columns (e.g. Items) of the according matrix. (cf. [ZL10, p.168]) So the divisor is the amount of all possible feedbacks.

System	Approach	Items	Users	I/U	Density
Amazon	Clustering, Item-to-Item CF	691mio Products	615mio	≈ 1.12	$U2I \approx 10^{-9}$
Del.icio.us	CF	100mio Websites	3 mio	≈ 33.33	$U2I \approx 10^{-7}$
eBay	Item-based CF	10mio Products	233mio	≈ 0.04	$U2U \approx 10^{-6}$
Facebook	User-based CF, Top-N Items	5bn Contents	400mio	≈ 12.5	$U2I \approx 10^{-9}$ $U2U \approx 10^{-6}$
Hulu	Item-based CF	333mio Programs	35mio	≈ 9.51	$U2I \approx 10^{-7}$
IMDb	Item-based CF	30mio Movies	100mio	≈ 0.3	$U2I \approx 10^{-5}$
Jamendo	CF	337,235 Tracks	1.5mio	≈ 0.22	$U2I \approx 10^{-5}$ $U2It \approx 10^{-4}$
Last.fm	Audioscrobbler, Feature-based F, User-to-Item CF	3.5mio Tracks	40mio	≈ 0.09	$U2I \approx 10^{-6}$
LibraryThing	Feature-based F, User-to-Item CF	68.3mio Books	1.4mio	≈ 48.79	$U2I \approx 10^{-7}$
LoveFilm	Item-based CF	67,000 Movies	1,5mio	≈ 0.04	$U2I \approx 10^{-5}$
MovieLens	Training data	10,000 Movies	72,000	≈ 0.14	$U2It \approx 10^{-2}$
MySpace	CF with MapReduce	463mio Pages	110mio	≈ 4.02	$U2U \approx 10^{-6}$ $U2I \approx 10^{-7}$
Netflix	Prize for best RE	70,000 Programs	6.3mio	≈ 0.01	$U2It \approx 10^{-2}$
Youtube	Feature-based F, CF	5.3bn Clips	99mio	≈ 53.54	$U2I \approx 10^{-6}$

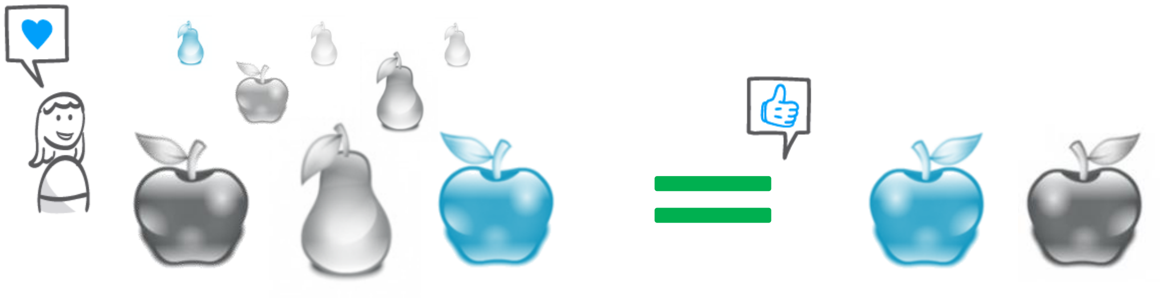


Fraunhofer FOKUS

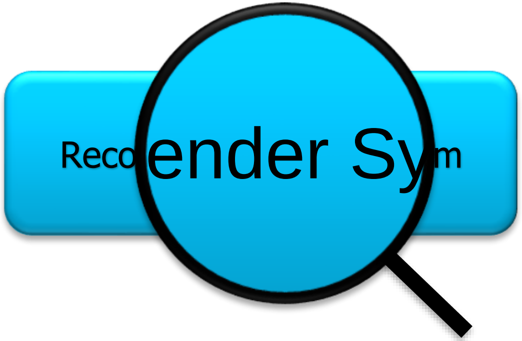
WHITE AND BLACK SHEEPS

Gray Sheep Problem:

PORTFOLIO EFFECT



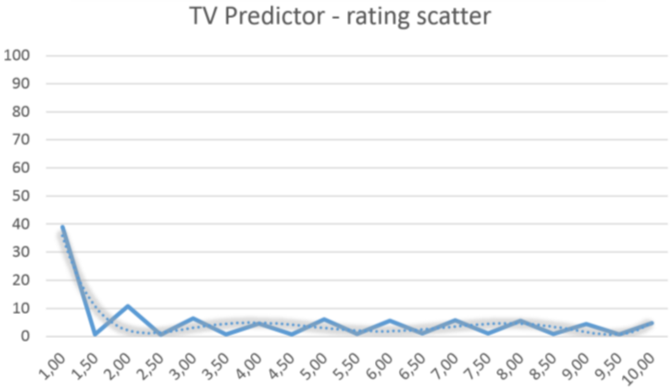
Overspecialization

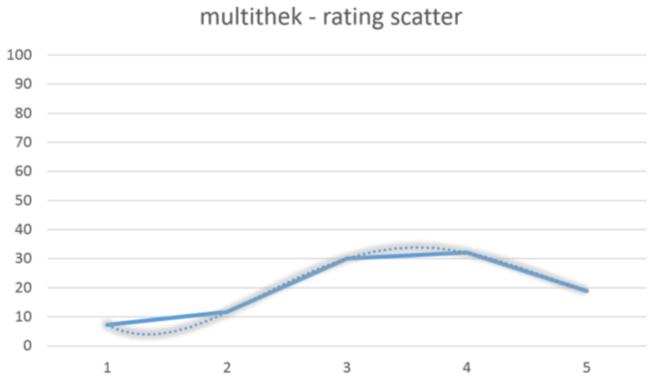


LONELY BEACH PARADOXON

USABILITY

vs.





Legal and ethical issues

LEGISLATION

Data Protection Act

- The inquiry, processing and use of personal data is only allowed if the person concerned **explicitly agrees** (§ § 12 TMG; 4 Abs. 1, 4a BDSG), for instance by filling in a so called opt-in field.
- The basic principle is to **reduce the amount of data** and handle them economically. (§ 3a BDSG)
- Personal data should only be inquired and processed, insofar as they are **necessary for the exercise of functions** (§ 13 BDSG). Useless data should be deleted or locked (§ § 13 Abs. 4 Nr. 2 TMG, 35 Abs. 2 Nr. 3 BDSG).
- The service provider has to **secure the data** with applicable precautionary measures from spying (§ 13 Abs. 4 Nr. 3 TMG).
- Personal data should only be used for issues of the **same telemedia service** (§ 13 Abs. 4 Nr. 4 TMG).
- The person concerned has the **right to get information** as to his personal data stored, the addressee and the intention of usage of their personal data (§ 13 Abs. 7 TMG; § 34 BDSG).

LEGISLATION

Self Privacy Protection

- Users may **change, limit or delete** their data, whenever they want (cf. [Fec11, p. 169])
- The service provider must not identify the user using their usage data, so the **user’s identity must be separated from the user’s behaviour** (cf. [Fec11, p. 366]).

FACEBOOK

Experimental evidence of

massive-scale emotional contagion

through social networks

„[…] Facebook conducted a massive psychological experiment

on 689,003 users,

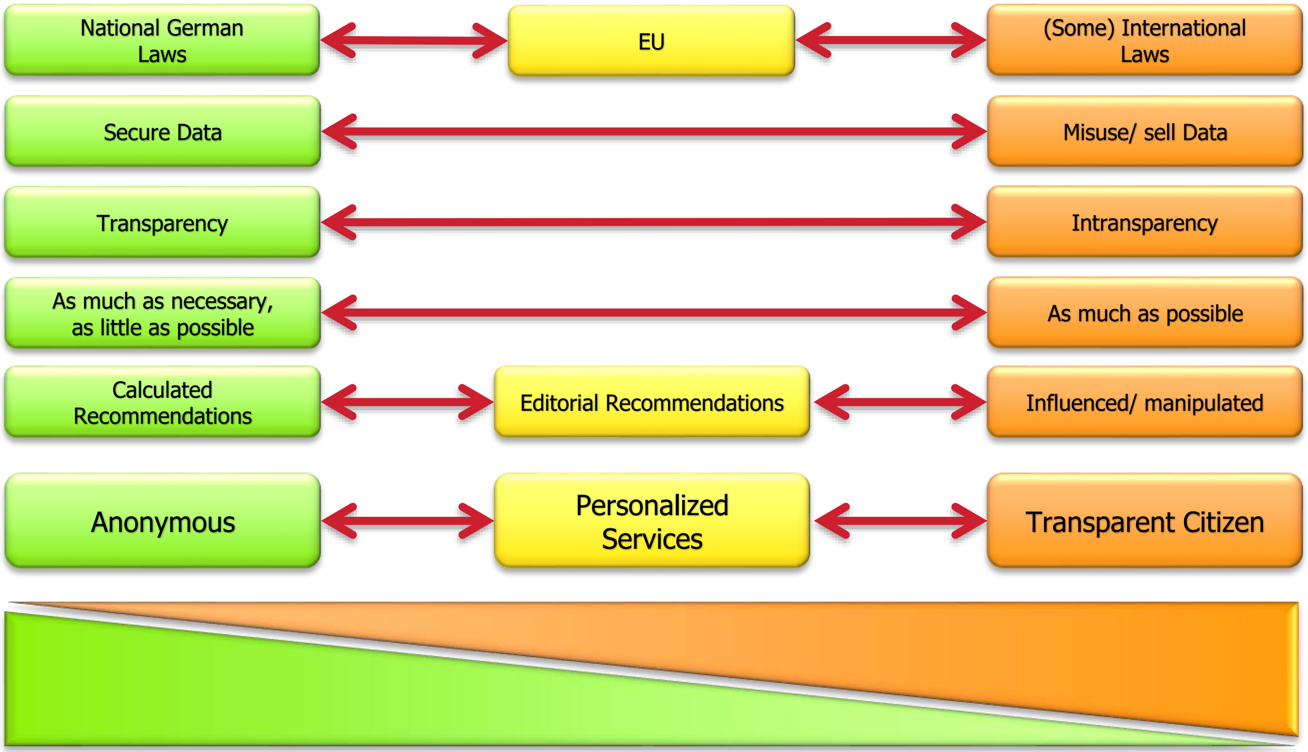
manipulating their news feeds

to assess the effects on their emotions. […]“ *

* Forbes.com (28.06.2014)

Fraunhofer FOKUS

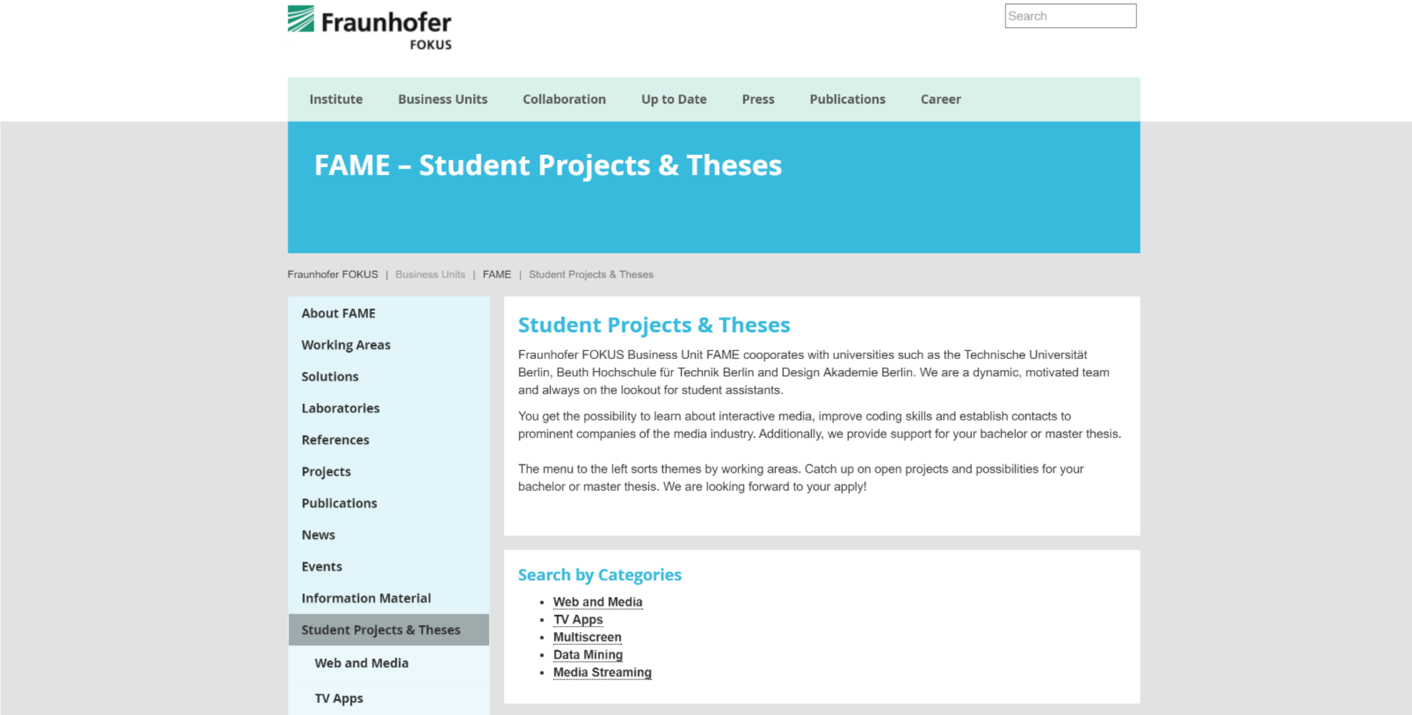
ETHICAL ISSUES



ETHIC: 7 GOLDEN RULES

1. **Transparency**: Explain how the system works
2. **Adjustment**: Allow users to tell the system it is wrong
3. **Trust**: Increase users' confidence in the system
4. **Effectiveness**: Help users make good decisions
5. **Persuasiveness**: Convince users to try or buy
6. **Efficiency**: Help users make decisions faster
7. **Satisfaction**: Increase the ease of usability or enjoyment

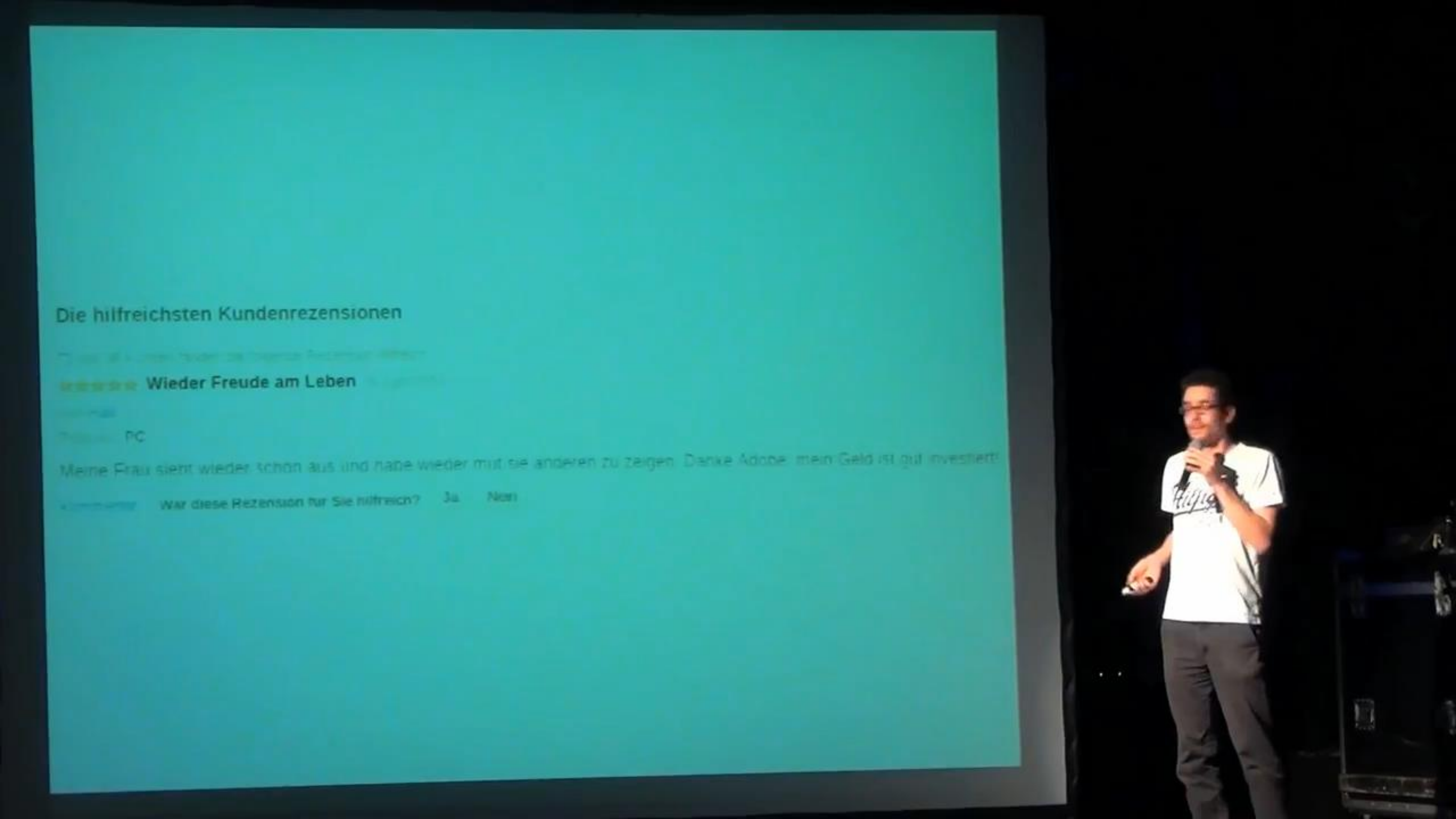
STUDENT JOBS AND THESIS TOPICS



https://www.fokus.fraunhofer.de/en/fame/student/datamining

MÖCHTEN SIE VIELLEICHT POMMES ZU DEN POMMES

https://www.youtube.com/watch?v=6-TpRmQtVEI


Die hilfreichsten Kundenrezensionen
Wieder Freude am Leben
PC
Meine Frau sieht wieder schön aus und habe wieder mut sie anderen zu zeigen. Danke Adobe, mein Geld ist gut investiert!
War diese Rezension für Sie hilfreich?
Ja
Nein

References

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, VOL. 17, NO. 6:734–749, 2005.

[Bai07] Duane A. Bailey. *Java Structures: Data Structures in Java for the Principled Programmer*. Williams College, 2007.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[CY10] Yan-ni Chen and Min Yu. A hybrid collaborative filtering algorithm based on user-item. *IEEE Computer Society*, pages 618–621, 2010.

[DFT10] Antonina Dattolo, Felice Ferrara, and Carlo Tasso. The role of tags for recommendation: a survey. *IEEE*, pages 548–555, 2010.

[Emb11] Ossama H. Embarak. A method for solving the cold start problem in recommendation systems. *IEEE - International Conference on Innovations in Information Technology*, pages 238–243, 2011.

[Han06] Jianchao Han. Learning fuzzy association rules and associative classification rules. *IEEE - International Conference on Fuzzy Systems*, pages 1454–1459, 2006.

[HMV07] Mark F. Hornick, Erik Marcadé, and Sunil Venkayala. *Java data mining - strategy, standard, and practice - A Practical Guide for Architecture, Design and Implementation*. Morgan Kaufmann Publishers, 2007.

[HTF01] Travor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science+Business Media, 2001.

[KA07] Jerome Kunegis and Sahin Albayrak. Adapting ratings in memory-based collaborative filtering using linear regression. *IEEE*, pages 49–54, 2007.

[KKH04] Hyungil Kim, Juntae Kim, and Jonathan Herlocker. Feature-based prediction of unknown preferences for nearest-neighbor collaborative filtering. *Proceddings of the Fourth IEEE International Conference on Data Mining*, IV, 2004.

[KPPK11] Eunhui Kim, Shinjee Pyo, Eunkyung Park, and Munchurl Kim. An automatic recommendation scheme of tv program contents for (ip)tv personalization. *IEEE TRANSACTIONS ON BROADCASTING*, VOL. 57, NO. 3:674–684, 2011.

[LJGL10] Guoling Liu, He Jiang, Runian Geng, and Huijie Li. Application of multidimensional association rules in personal financial services. *IEEE - International Conference On Computer Design And Appliations*, pages 500–503, 2010.

[LM05] Daniel Lemire and Anna Maclachlan. Slope one predictors for online rating-based collaborative filtering. *Proceedings of SIAM Data Mining (SDM'05)*, 2005.

[LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations - item-to-item collaborative filtering. *IEEE Internet Computing*, pages 76–80, 2003.

[LWQL09] Zhaobin Liu, Hui Wang, Wenyu Qu, and Weijiang Liu. Sparse matrix prediction filling in collaborative filtering. *IEEE International Conference on Scalable Computing and Communications; The Eighth International Conference on Embedded Computing;*, pages 304–307, 2009.

[Mar09] Gary M. Maranell. *Scaling: A Sourcebook for Behavioral Scientists*. Aldine Transaction, 2009.

[Mer10] Agathe Merceron. Data mining - association rules (slides). In *Data Mining*, 2010.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *GroupLens Research Group/Army HPC Research Center*, 2001.

[Ste46] S. S. Stevens. *On the Theory of Scales of Measurement*. American Association for the Advancement of Science, 1946.

[SW09] Choonsung Shin and Woontack Woo. Socially aware tv program recommender for multiple viewers. *IEEE Contributed Paper*, pages 927–932, 2009.

[TM07] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. *Data Engineering Workshop, 2007 IEEE*, 23rd International Conference:801 – 810, 2007.

[WY09] Pu Wang and Hong Wu Ye. A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering. *International Conference on Industrial and Information Systems*, pages 152–154, 2009.

[XA06] Jin An Xu and Kenji Araki. A svm-based personal recommendation system for tv programs. *IEEE*, pages 401–404, 2006.

Q&A

THANK YOU FOR YOUR ATTENTION

Fraunhofer FOKUS

FRAUNHOFER INSTITUTE FOKUS

Dr. Stefan Arbanowski
Director of Competence Center
Future Applications and Media
Tel. +49 (30) 34 63 – 71 97
stefan.arbanowski@fokus.fraunhofer.de

Fraunhofer Institute for Open Communication Systems FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin, Germany

Tel: +49 (30) 34 63 – 7000
Fax: +49 (30) 34 63 – 8000
www.fokus.fraunhofer.de

Christopher Krauss
Senior Project Manager R&D
Future Applications and Media
Tel. +49 (30) 34 63 – 72 36
christopher.krauss@fokus.fraunhofer.de

Fraunhofer Institute for Open Communication Systems FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin, Germany

Tel: +49 (30) 34 63 – 7000
Fax: +49 (30) 34 63 – 8000
www.fokus.fraunhofer.de

Fraunhofer
FOKUS

making cities smart