

TOWARD INTERACTIVE USER DATA ANALYTICS

Sihem Amer-Yahia

CNRS Research Director, LIG, France

Keynote at BICOD 2017



User data

- Data about users
 - Socio-demographics: *age, location, occupation, etc*
 - Behavior: *interests, ratings, etc*
- Generated by users, by sensors
 - Collaborative rating sites
 - Quantified-self
 - Research papers

User data analytics

- A set of methods and tools to extract value from user data
- Behavioral analytics

https://en.wikipedia.org/wiki/Behavioral_analytics

“a recent advancement in business analytics that reveals new insights into the behavior of consumers on eCommerce platforms, online games, web and mobile applications, and IoT.”

Behavioral analytics

examples https://en.wikipedia.org/wiki/Behavioral_analytics

- **Ecommerce and retail** – Product recommendations and predicting future sales trends
- **Online gaming** – Predicting usage trends, load, and user preferences in future releases
- **Application development** – Determining how users use an application to predict future usage and preferences
- **Cohort analysis** -- Breaking users down into similar groups to gain a more focused understanding of their behavior
- **Security** – Detecting compromised credentials and insider threats by locating anomalous behavior
- **Suggestions** – People who liked this also liked...

Behavioral analytics

examples https://en.wikipedia.org/wiki/Behavioral_analytics

- **Ecommerce and retail** – Product recommendations and predicting future sales trends
- **Online gaming** – Predicting usage trends, load, and user preferences in future releases
- **Application development** – Determining how users use an application to predict future usage and preferences
- **Cohort analysis** -- Breaking users down into similar groups to gain a more focused understanding of their behavior
- **Security** – Detecting compromised credentials and insider threats by locating anomalous behavior
- **Suggestions** – People who liked this also liked...

Labeled user groups

- See user data as *labeled groups*
 - Because labeled groups are more informative than individuals
 - Because user data is sparse
 - Because user data is noisy
- Examples
 - *Young people who rated Woody Allen movies*
 - *Middle-aged females in California*
 - *People who rated movies starring Scarlett Johansson*
 - *Female engineers who rated Star Wars*
 - *[25-35] year-old professionals who live in Grenoble and who rated movies starring Sean Penn*

Interactive analytics, why?

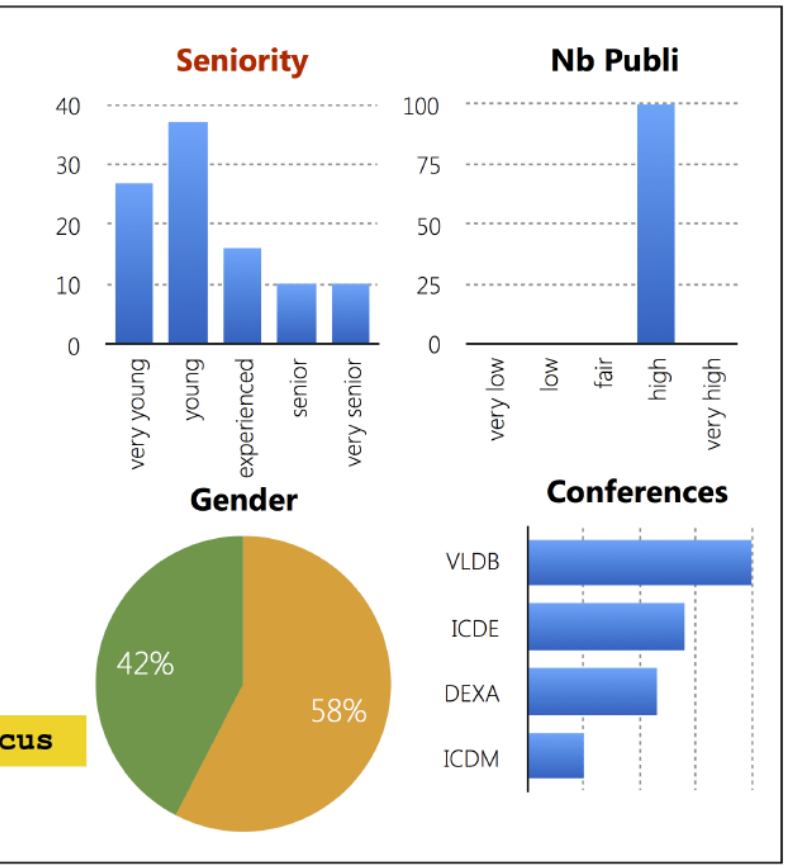
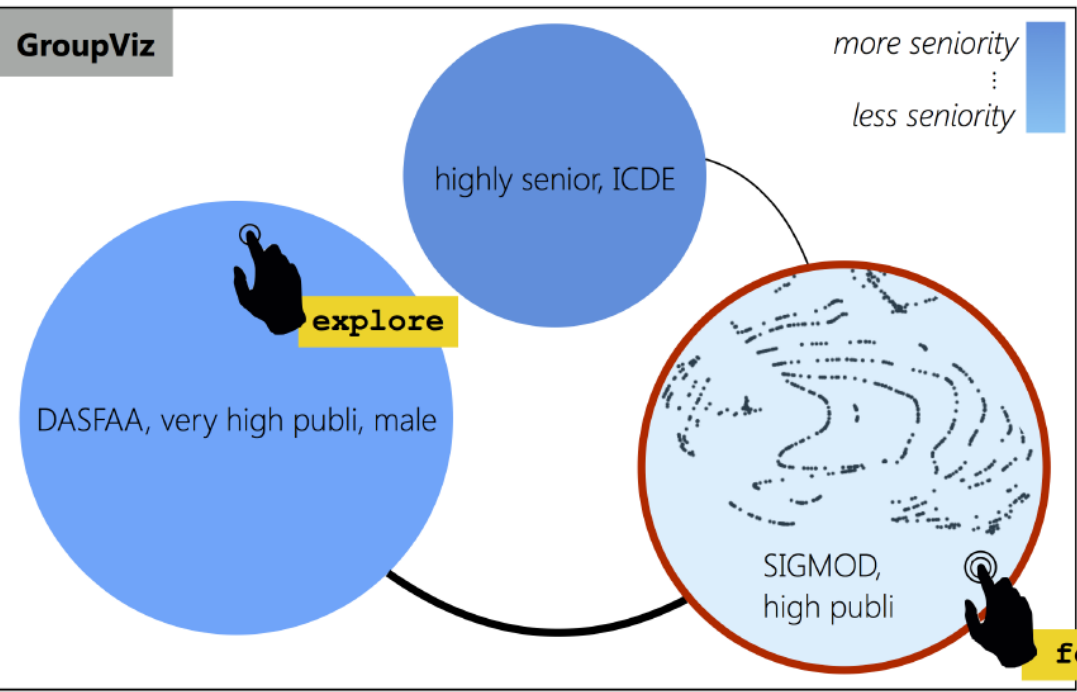
- Often, analysts need to see results to proceed further
- Often, they can express only partial needs

- Examples
 - A social scientist who seeks to verify a hypothesis
 - A computer scientist who wants to build a conference PC
 - A person looking for someone about whom she only remembers some details
 - A data scientist who wishes to query/mine the same dataset in different ways

Towards interactive user analytics

- Dedicated spaces
- Ability to switch between spaces
- A memory

Context high publi active VLDB male



Tracker , data integration → SIGMOD, male

Users Table Seniority **backtrack** _publi pub_rate

Users Table	Seniority	_publi	pub_rate
C. Bohm	young	fair	active
F. T. Liu	very young	fair	active
E. Ronchetti	experienced	very high	very active
A. Raffio	young	high	active

Save Area Alexander S. Szalay

David A. Shamma Djamel Benslimane

G male, extreme_active

C. Bohm

G very_young, female

save

Toward interactive user analytics

- Dedicated spaces
 - for group members, i.e. users
 - for groups
 - for analytics
- Ability to switch between spaces
 - from group to groups
 - from users to groups
 - from analytics to groups
- A memory
 - to gather individual group members
 - to remember exploration paths

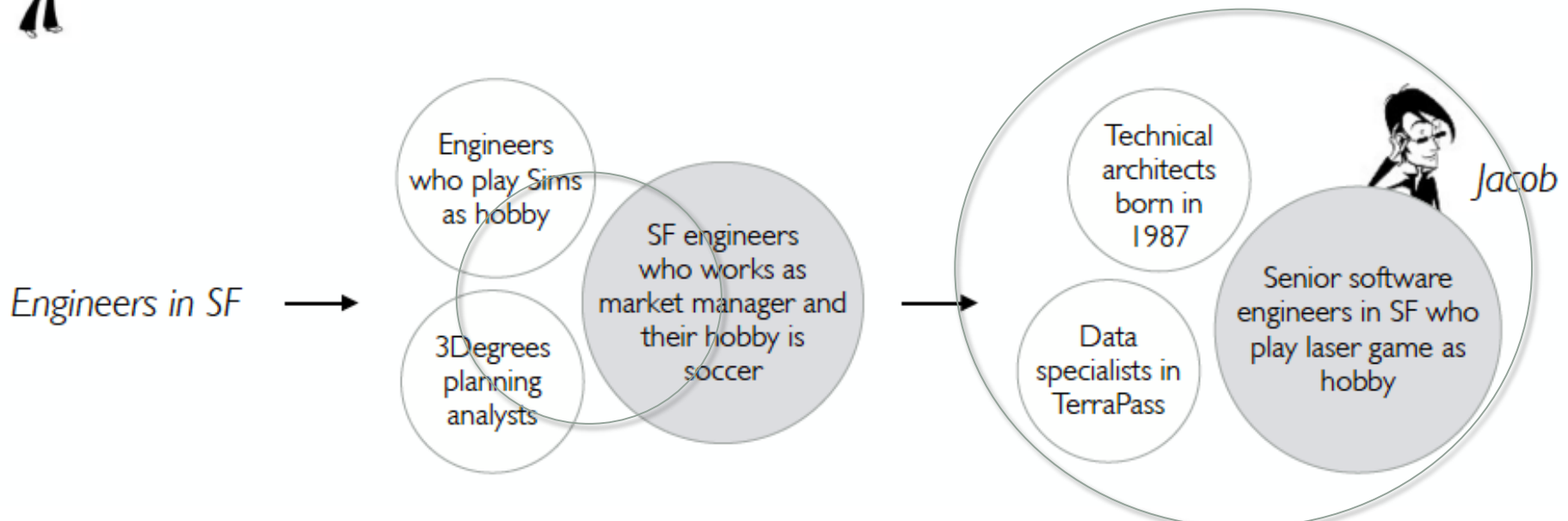
Group Exploration

from group to groups

Looking for a person [1] *explore/exploit*

Julia

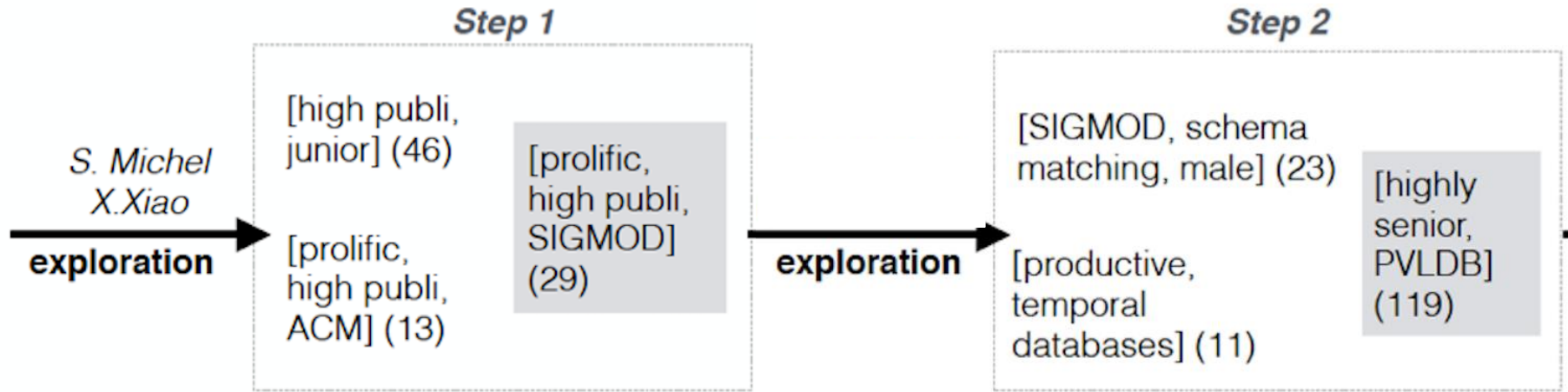
I met a guy in last night party in San Fransisco (SF) but lost his phone number and I don't remember his name! I only remember that he works as engineer.



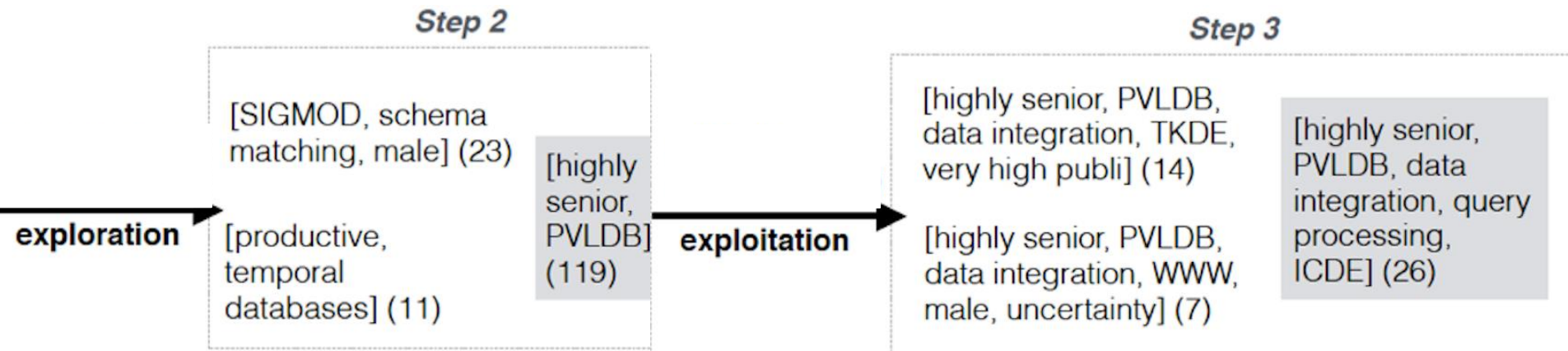
Group exploration primitives

- Given a group, the analyst switches between
 - *Tell me more about that group*
 - *Show me related groups*
- Given a group
 - *Exploit it*: finding k groups that maximize a combination of *Coverage* and *Diversity* is NP-Complete (shown using the Maximum Coverage Problem)
 - *Explore it*: find k groups that *Overlap* with it and that maximize *Diversity* (NP-Complete using the Maximum Edge Subgraph Problem)

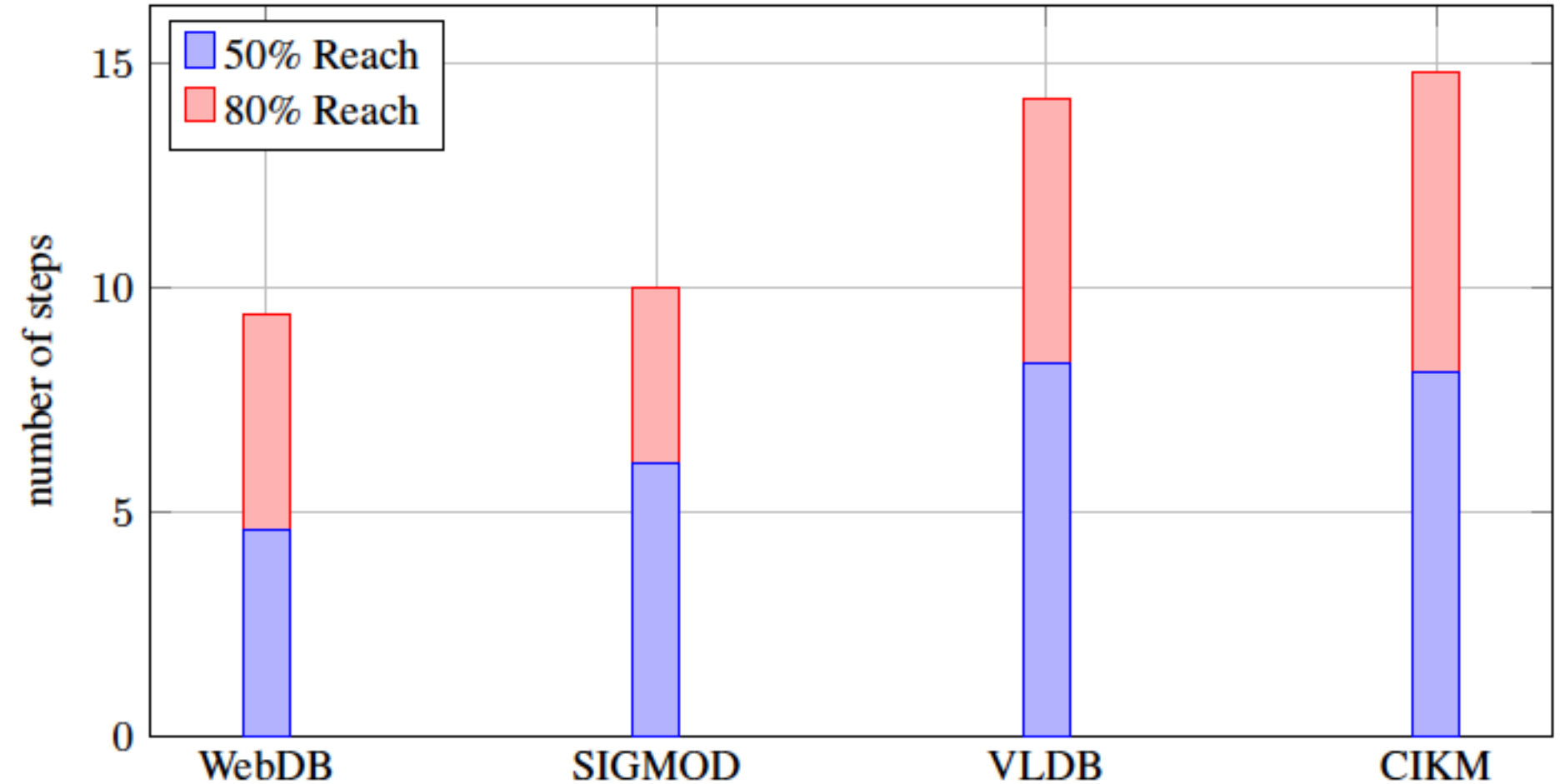
Building a PC (Martin T., WebDB PC chair) explore/exploit



Building a PC (Martin T., WebDB PC chair) explore/exploit



Number of steps for PC selection on average

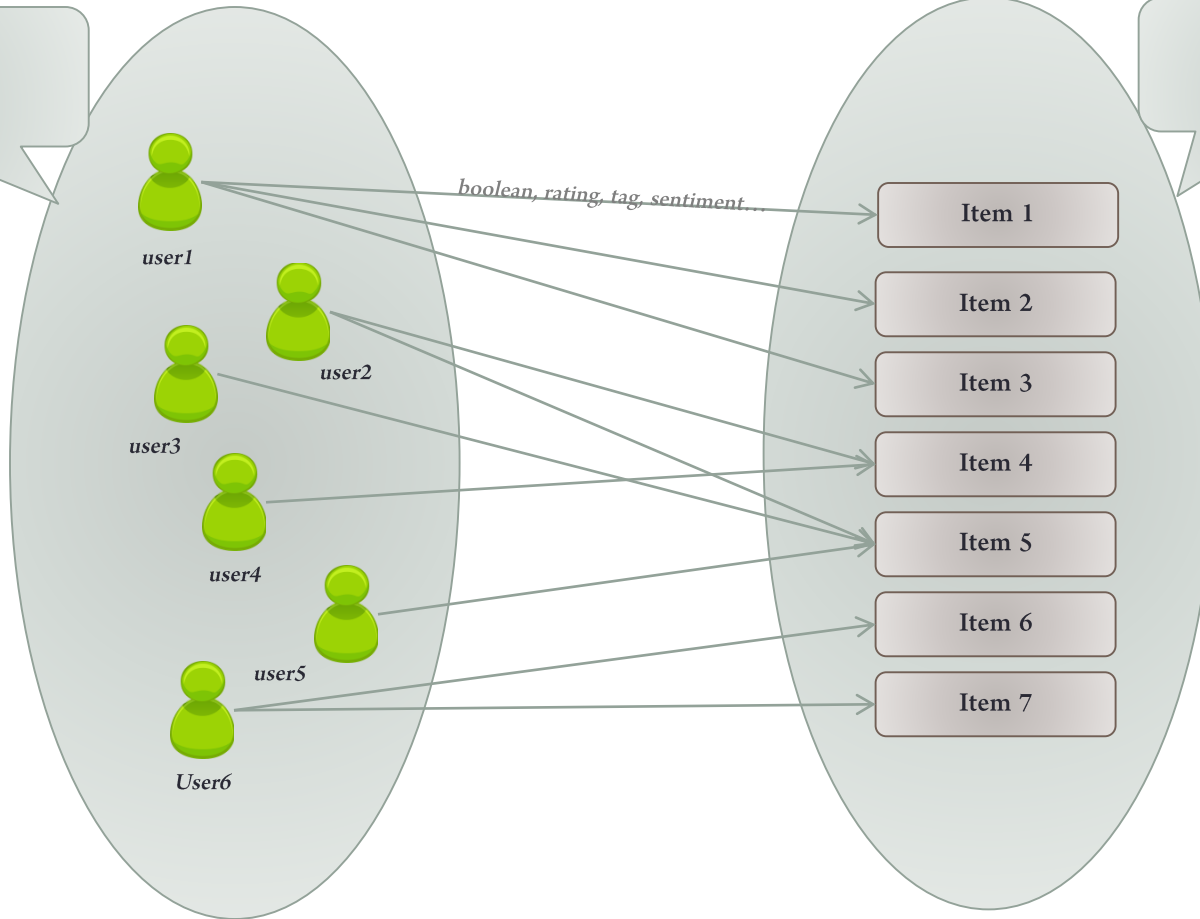


Group Discovery

from users to groups
from analytics to groups

User data model

User space
(with attributes)



Item space
(with attributes)

User data on collaborative rating systems

- a set of rating records: $\langle \text{item attributes}, \text{user attributes}, \text{rating} \rangle$

ID	Movie	Name	Gender	Age	Occup.	Rating
r ₁	Toy Story	John	M	young	teacher	4
r ₂	Toy Story	Jennifer	F	old	teacher	3
r ₃	Toy Story	Mary	F	old	teacher	2
r ₄	Titanic	Carine	F	old	other	4
r ₅	Toy Story	Sara	F	young	student	3
r ₆	Toy Story	Martin	M	young	student	5
r ₇	Titanic	Peter	M	young	student	1

Data from MovieLens (additional attributes from IMDb)

User group discovery problem

Given raw user data, discover *good* groups

Input: a set of user records, (and analytics)

Output: a set of groups

What is a good group, *globally* speaking?

IMDb



*rating records for
romance genre
movies*

I believe romantic
movies are mostly
liked by females.



Elena
social
scientist

young females
average: 3.7
variance: 2

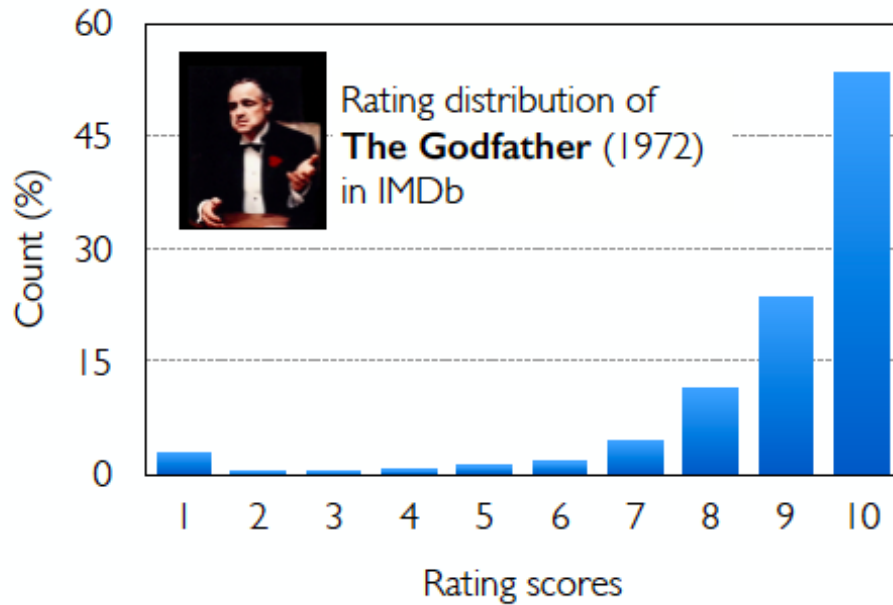
females in DC
average: 4.6
variance: 1.5

male teenagers
average: 3.1
variance: 3.4

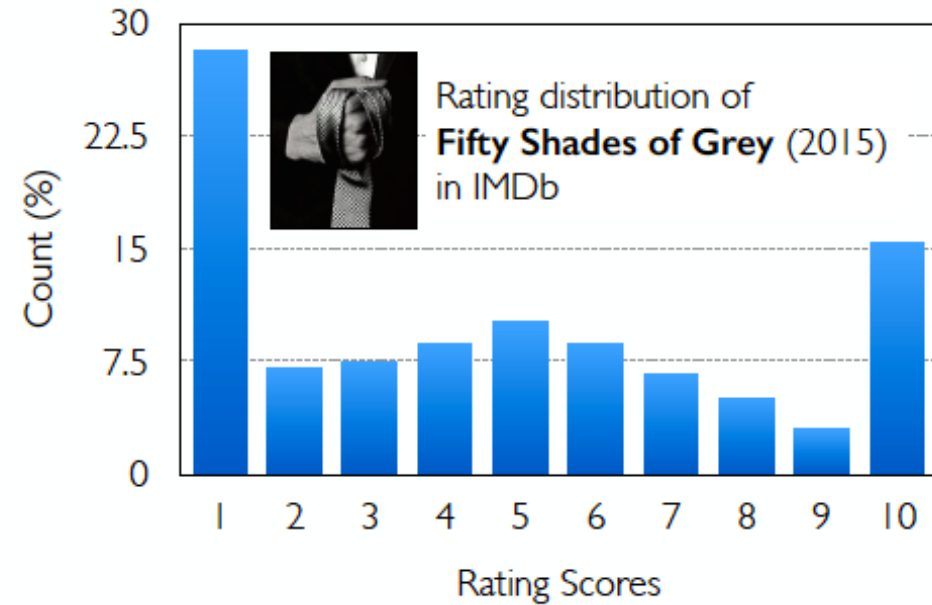
— user groups
that **cover** most
romance ratings

What is a good group, *locally* speaking?

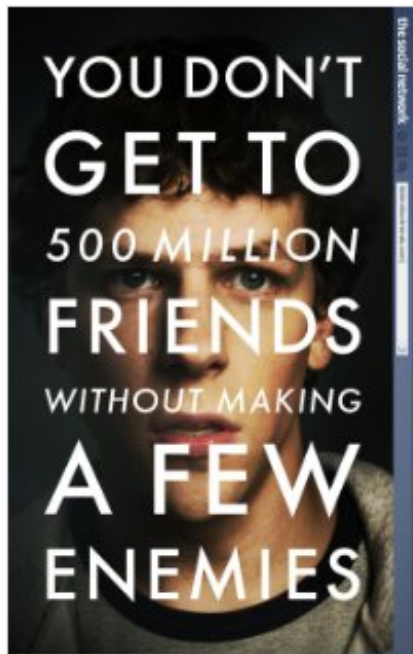
Homogeneous Rating Distribution



Polarized Rating Distribution



(pre-defined) user groups on IMDb



The Social Network (2010)

PG-13 120 min - [Biography](#) / [Drama](#) - 1 October 2010 (USA)



Ratings: **8.0/10** (circled)
Reviews: 522 users

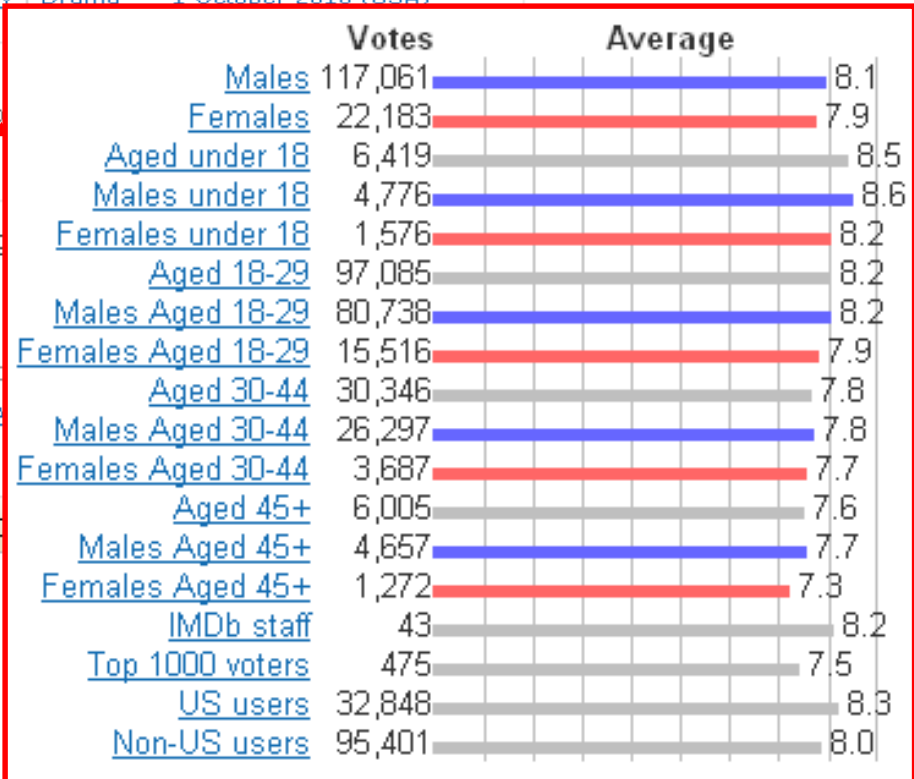
A chronicle of the founding of the social networking Web site.

Director: [David Fincher](#)

Writers: [Aaron Sorkin](#) (screenplay)

Stars: [Jesse Eisenberg](#), [Andrew Garfield](#), [Justin Timberlake](#)

[Watch Trailer](#) [Add to Watchlist](#)

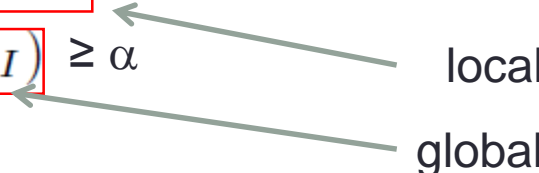


Group discovery: a possible formulation [2]

For an input item covering R_I ratings, return a set C of k groups, s.t.

description error $\text{error}(C, R_I)$ is minimized, subject to:

coverage $\text{coverage}(C, R_I) \geq \alpha$



local

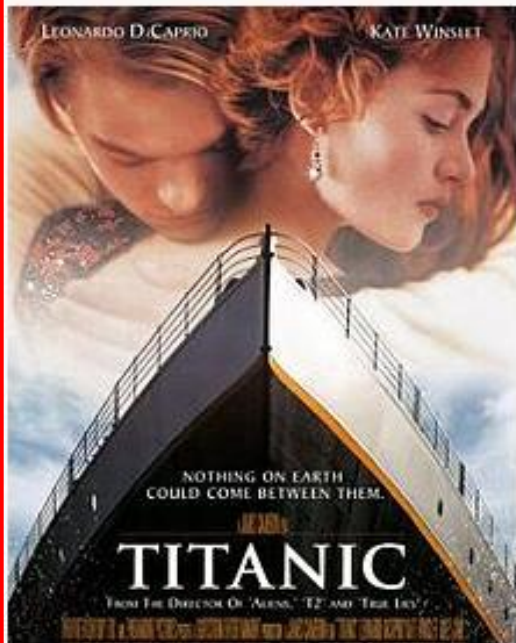
global

$$\begin{aligned}\text{error}(C, R_I) &= \sum_{r \in R_I} (E_r) \\ &= \sum_{r \in R_I} \text{avg}(|r.s - \text{avg}_{c \in C \wedge r \in c}(c)|)\end{aligned}$$

[2] MRI: Meaningful Interpretations of Collaborative Ratings.
S. Amer-Yahia, Mahashweta Das, Gautam Das and Cong Yu. PVLDB 2011.

Group discovery: $k=1$

Titanic



Titanic ([1997](#))

PG-13 194 min - [Adventure](#) | [Drama](#) | [History](#) - [19 December 1997 \(USA\)](#)



Ratings: **7.4/10** from 288,334 users Metascore: **74/100**
Reviews: 2,284 user | 174 critic | 34 from [Metacritic.com](#)

**Teen-aged female reviewers have rated this movie uniformly
Their average rating: 9.2**

Group discovery problem

THEOREM 1. The decision version of the problem of meaningful description mining (DEM) is NP-Complete even for boolean databases, where each attribute ia_j in \mathcal{I}_A and each attribute ua_j in \mathcal{U}_A takes either 0 or 1.

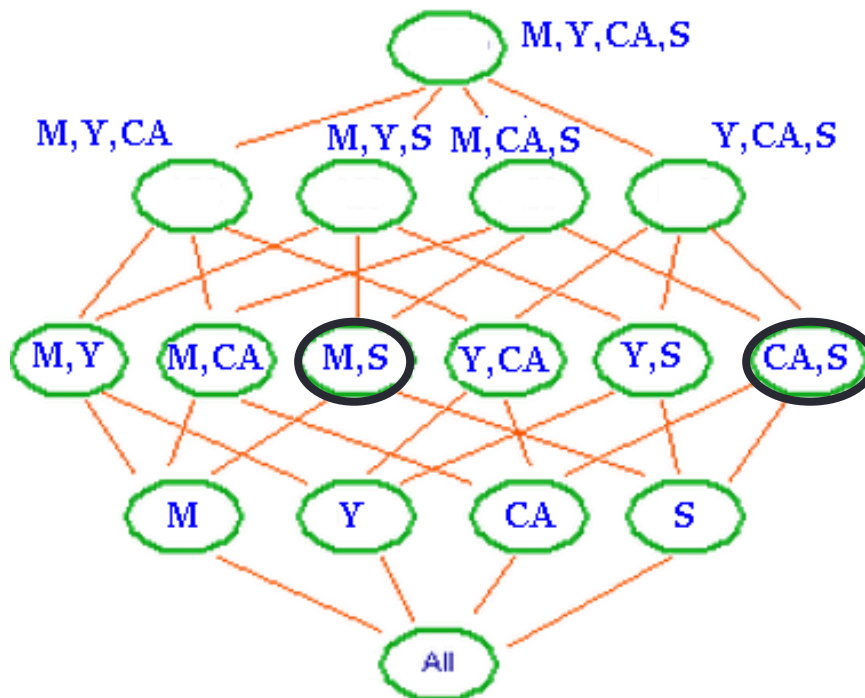
To verify NP-completeness, we reduce the Exact 3-Set Cover problem (EC3) to the decision version of our problem. EC3 is the problem of finding an exact cover for a finite set U , where each of the subsets available for use contain exactly 3 elements. The EC3 problem is proved to be NP-Complete by a reduction from the Three Dimensional Matching problem in computational complexity theory

Random Restart Hill Climbing Algorithm

$k = 2$

Satisfy Coverage

Minimize Error

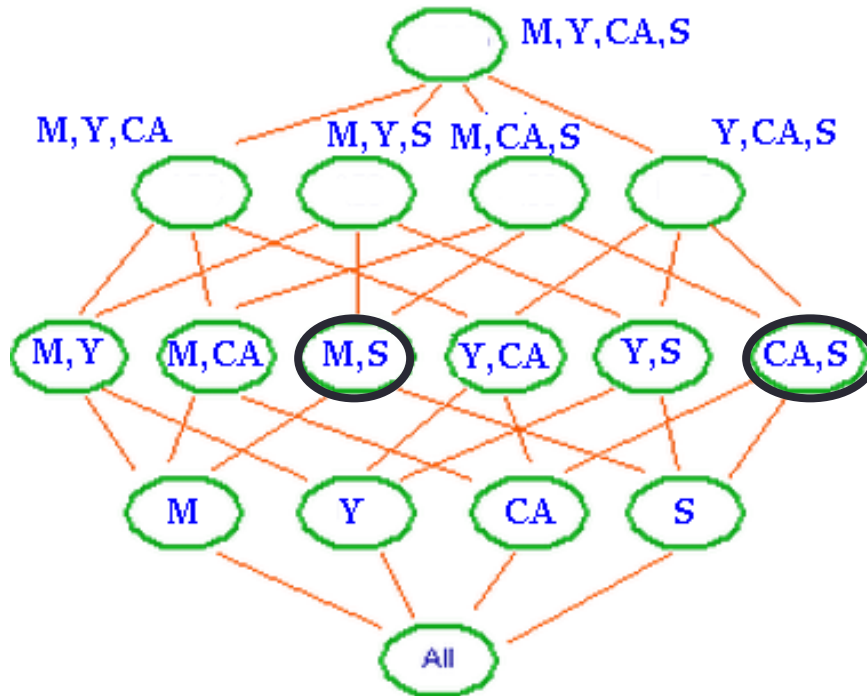


$C = \{\text{Male, Student}\}$
 $\{\text{California, Student}\}$

Random Restart Hill Climbing Algorithm

Satisfy Coverage

Minimize Error

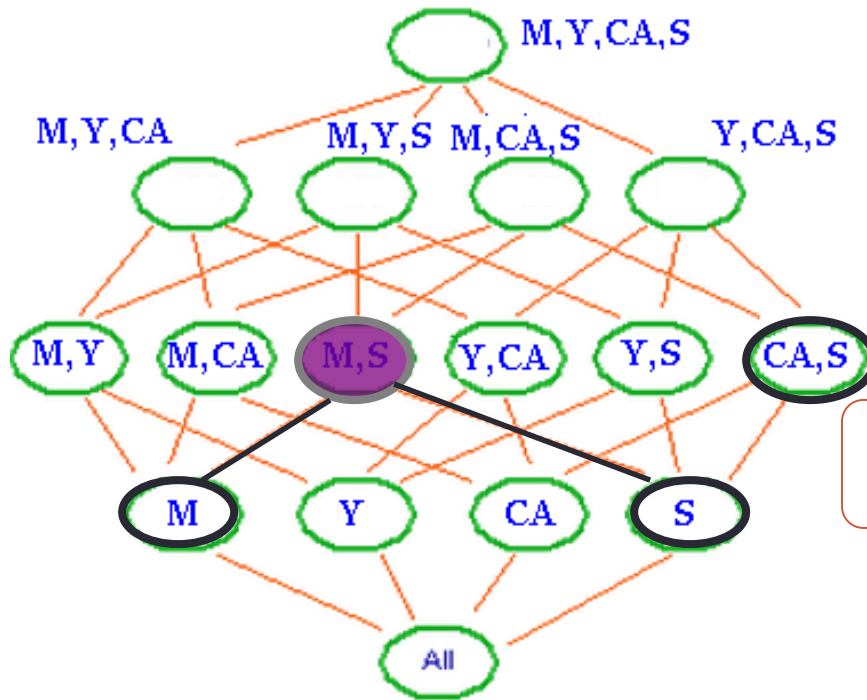


$C = \{ \text{Male, Student} \}$
 $\{ \text{California, Student} \}$

Say, C does not satisfy Coverage Constraint

Random Restart Hill Climbing Algorithm

Satisfy Coverage
Minimize Error



$C = \{\text{Male, Student}\}$
 $\{\text{California, Student}\}$

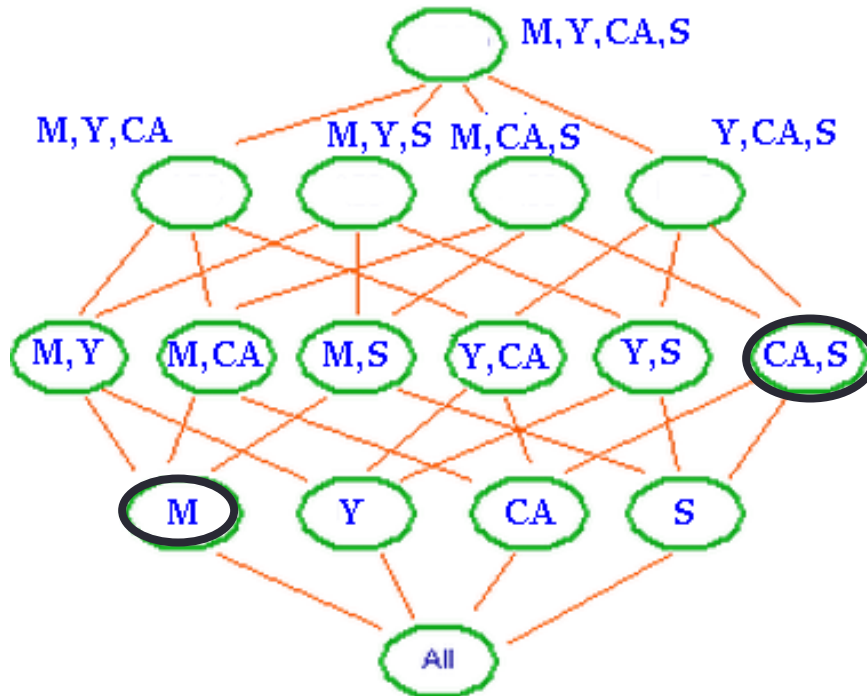
$C = \{\text{Male}\}$
 $\{\text{California, Student}\}$

$C = \{\text{Student}\}$
 $\{\text{California, Student}\}$

Random Restart Hill Climbing Algorithm

Satisfy Coverage

Minimize Error



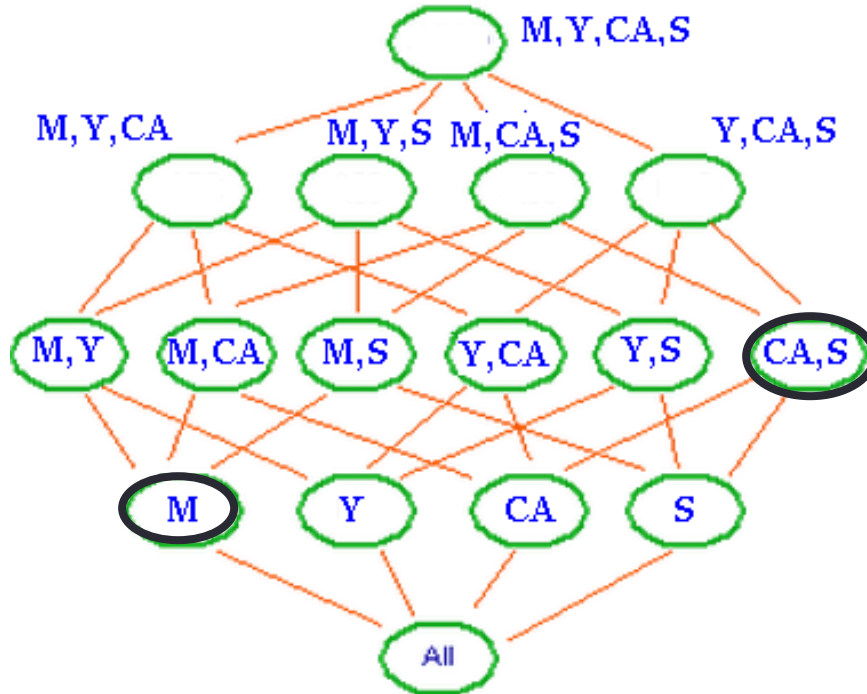
$C = \{\text{Male}\}$
 $\{\text{California, Student}\}$

Say, C satisfies
Coverage Constraint

Random Restart Hill Climbing Algorithm

Satisfy Coverage

Minimize Error

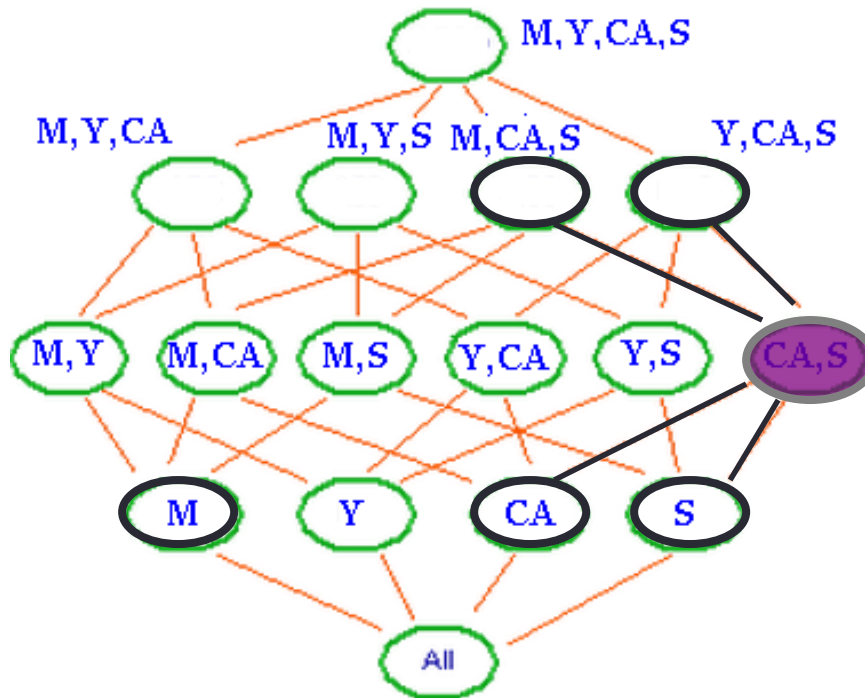


$C = \{Male\}$
 $\{California, Student\}$

Random Restart Hill Climbing Algorithm

Satisfy Coverage

Minimize Error

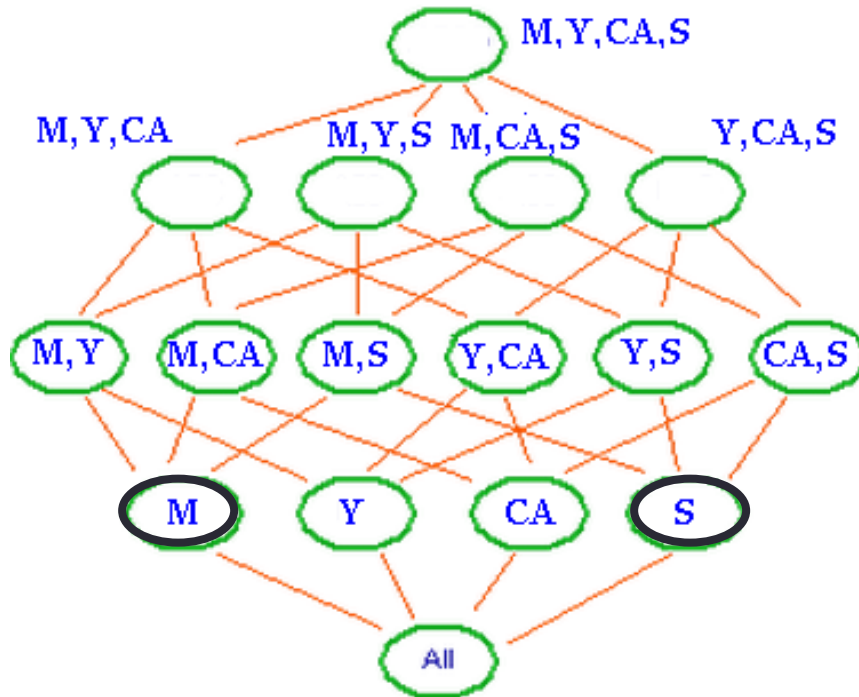


$C = \{Male\}$
 $\{California, Student\}$

Random Restart Hill Climbing Algorithm

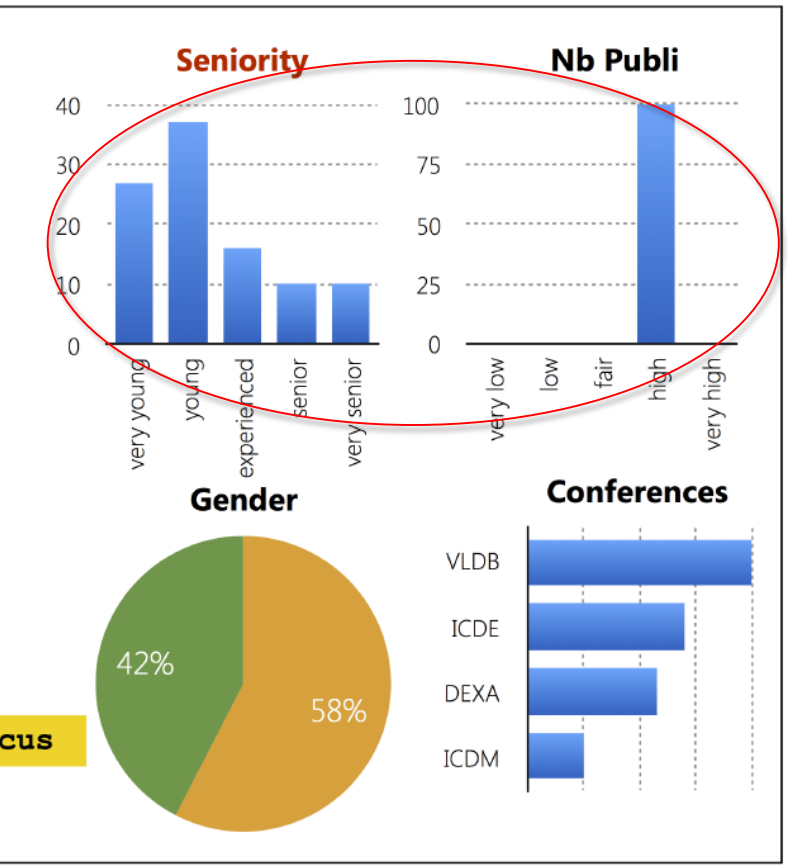
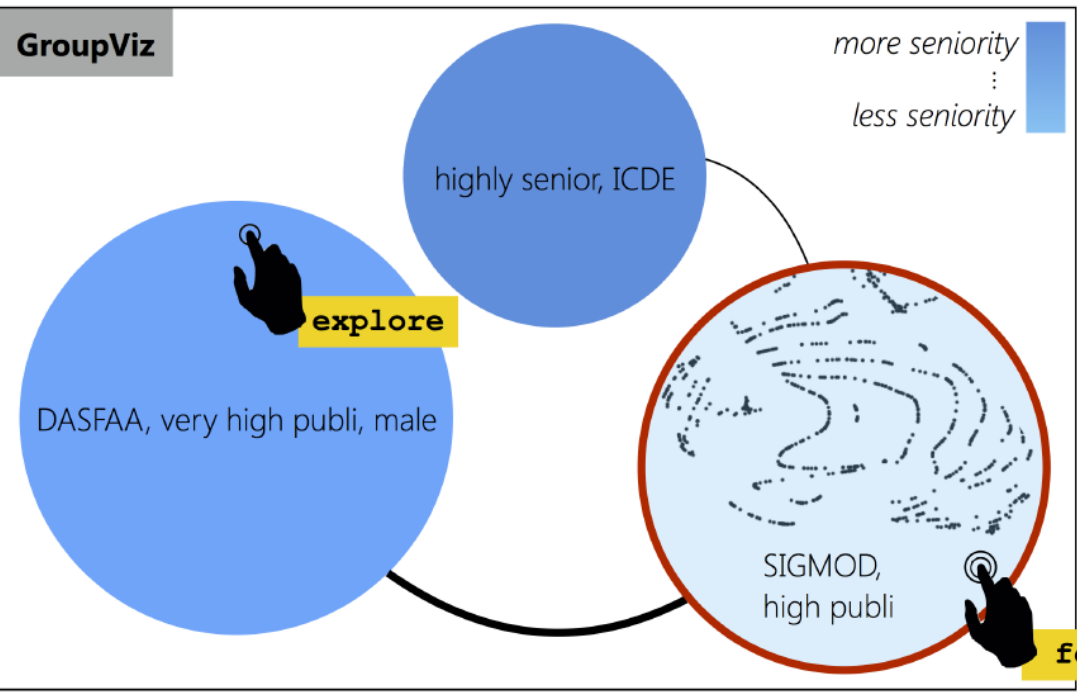
Satisfy Coverage

Minimize Error



$C = \{\text{Male}\}$
 $\{\text{Student}\}$

Context high publi active VLDB male



Tracker , data integration → SIGMOD, male

Users Table Seniority _publi pub_rate

backtrack

Users Table	Seniority	_publi	pub_rate
C. Bohm	young	fair	active
F. T. Liu	very young	fair	active
E. Ronchetti	experienced	very high	very active
A. Raffio	young	high	active

Save Area Alexander S. Szalay

David A. Shamma Djamel Benslimane

G male, extreme_active

C. Bohm

G very_young, female

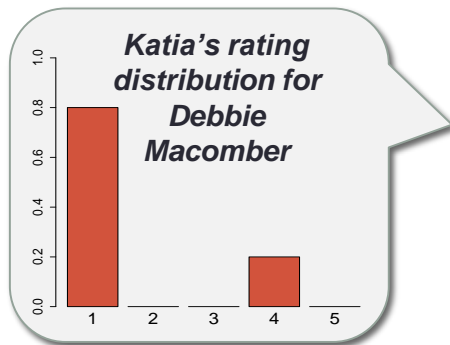
save

Looking for a book club [3]

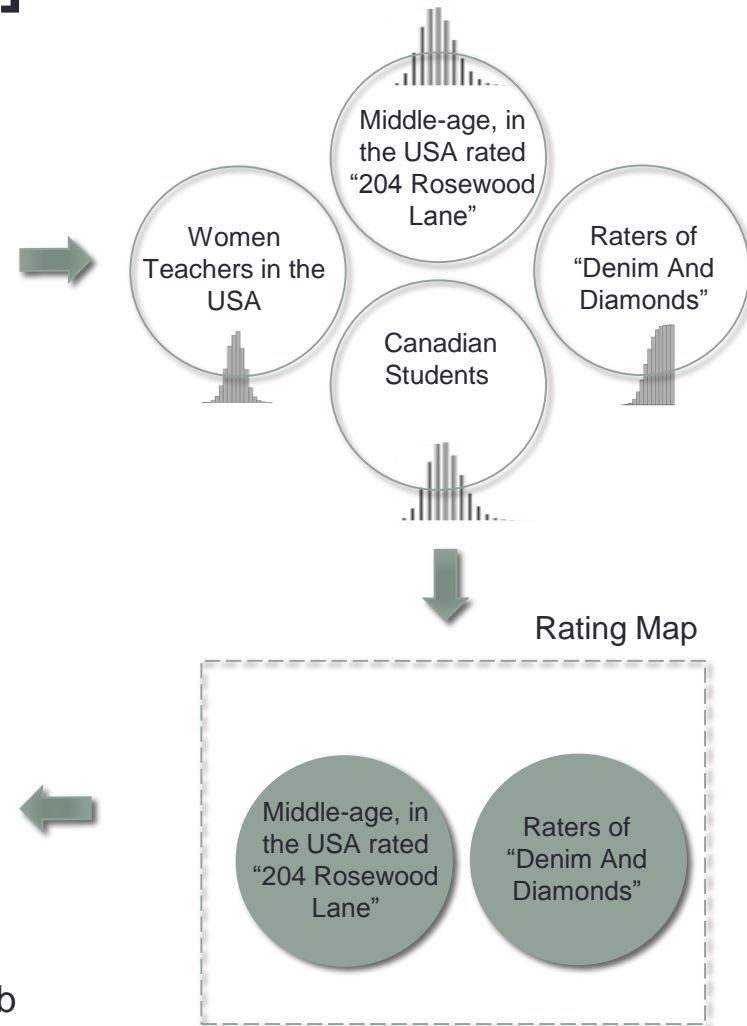
ID	Title	Author	Age	Gender	Occ.	Location	Rating
r_1	204 Rosewood Lane	Debbie Macomber	Young	M	Student	USA	1
r_2	204 Rosewood Lane	Debbie Macomber	Middle-age	F	Teacher	USA	5
r_3	Denim and Diamonds	Debbie Macomber	Middle-age	F	Teacher	USA	4



bookcrossing.com™



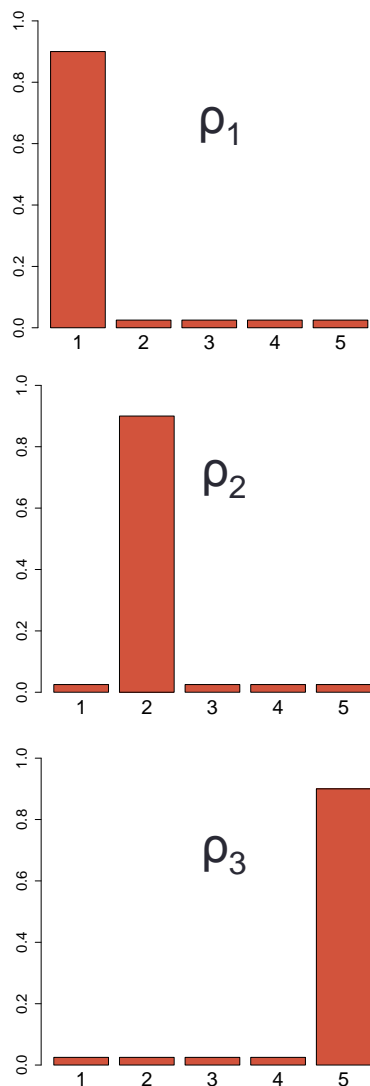
Katia wants to join a book club



[3] Exploring Rated Datasets with Rating Maps.

S. Amer-Yahia, S. Kleisarchaki, N. Kolloju, L. V.S. Lakshmanan, R. H. Zamar. WWW 2017.

Choosing a distribution comparison measure



Measure	(ρ_1, ρ_2)	(ρ_2, ρ_3)
Cosine	0.058	0.058
KL-Divergence	3.13	3.13
JS-Divergence	0.53	0.53
Euclidean Distance	1.24	1.24
Hellinger Distance	0.791	0.791
Total Variation Distance	0.875	0.875
Renyi Entropy Distance (0.5 order)	1.962	1.962
Battacharya Distance	0.981	0.981
Distance Correlation	0.25	0.25
Lukaszyk-Karmowksi Metric	1.1625	3.525
Signal Noise Ratio	2.0372	4.221
Earth Mover's Distance (EMD)	0.875	3.5

Choosing high quality groups

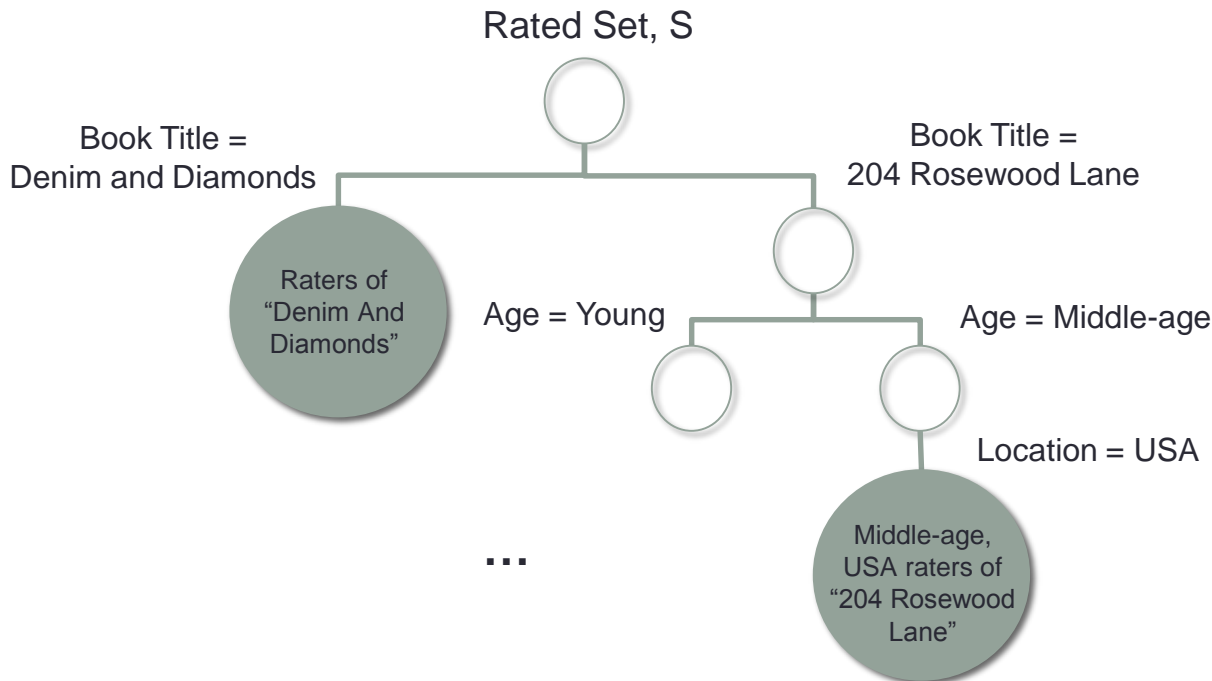
Optimization Problem: *Given a rated dataset S and distributions of interest $\{\rho_1, \dots, \rho_p\}$, find the largest and least overlapping groups whose rating distribution is close to one of $\{\rho_1, \dots, \rho_p\}$ (using an EMD threshold)*

- *Also a hard problem:* reduction from the classic Minimum Height Decision Tree problem

Brief sketch of algorithm

- Find a minimum height PDT with the following gain function:

$$\text{Gain}(\text{Attr}_i) = \frac{n}{\sum_{j=1}^n \min_{\rho \in \{\rho_1, \dots, \rho_k\}} \text{EMD}(c_j^i, \rho)}$$



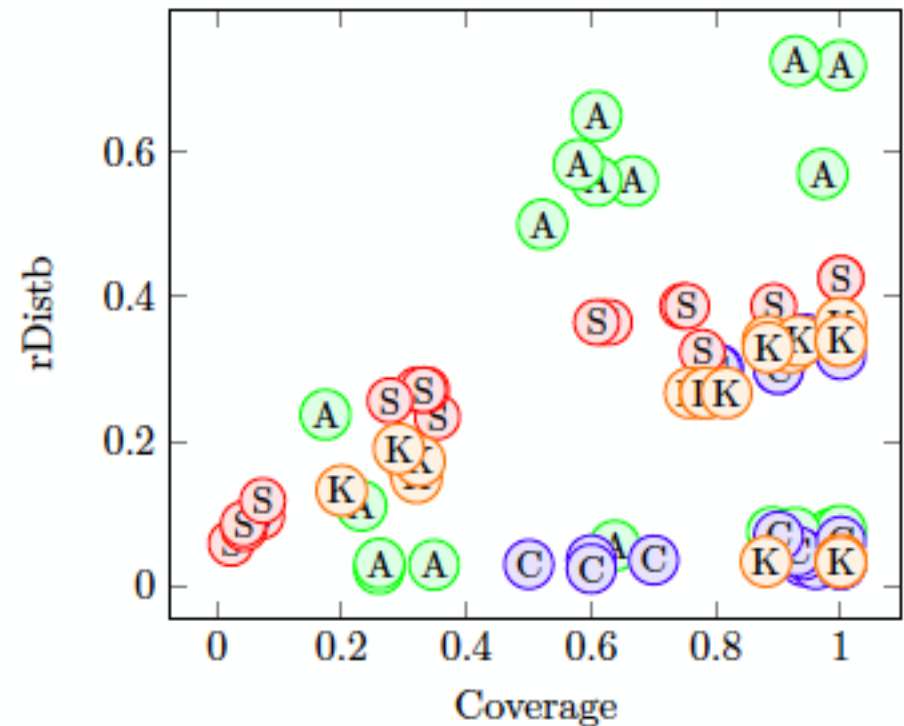
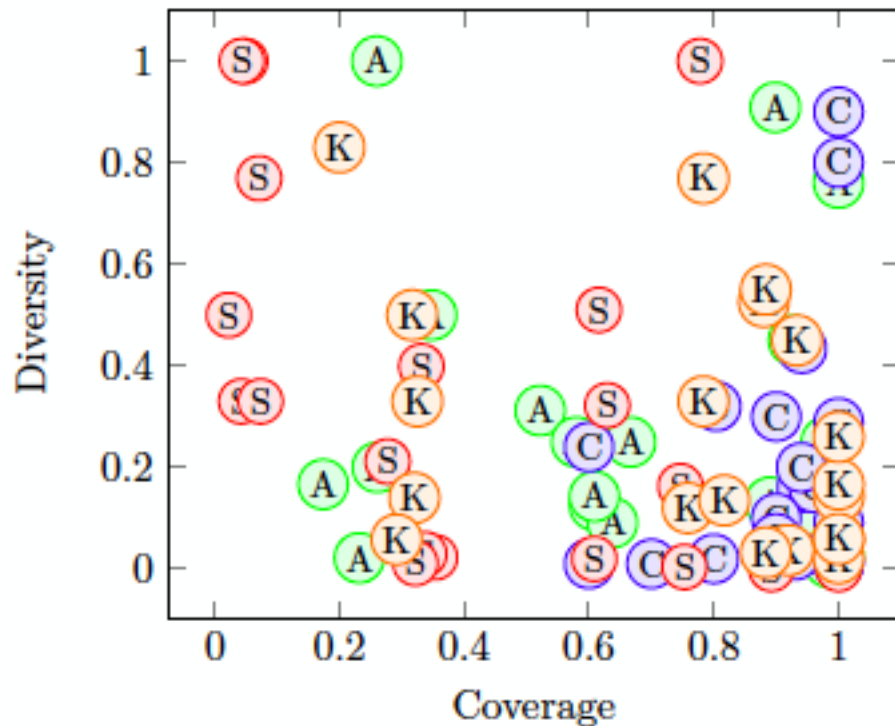
Gain(age)=0.7
Gain(gender)=1.2
Gain(book title)=1.86

Takeaways

- Several formalizations are possible both for group discovery and group exploration
 - Coverage: preserve train of thought of analyst
 - Diversity: provide exploration options
- Switching between spaces ensures seamless transitions

Research opportunity

Finding groups is *multi-objective* [4]



[4] Multi-Objective Group Discovery on the Social Web. B. Omidvar-Tehrani, S. Amer-Yahia, P. F. Dutot, and D. Trystram. ECML/PKDD 2016

Additional research opportunities

- Memory and feedback [5]
- An algebra for user data analytics [6]
- Groups over time [7]

[5] One click mining: Interactive local pattern discovery through implicit preference and performance learning. M. Boley, B. Kang, P. Tokmakov, M. Mampaey, S. Wrobel. IDEAS (ACM SIGKDD Workshop), 2013

[6] zenvisage and vega

[7] Querying Temporal Drifts at Multiple Granularities. Sofia Kleisarchaki, Sihem Amer-Yahia, Ahlame Douzal Chouakria, Vassilis Christophides. CIKM 2015