# Introduction to the Social Web

*Recommendation and Mining*

**Sihem Amer-Yahia**

**CNRS/LIG**

**Nov 9th, 2016**

**Nov 16th, 2016**

# Social Content Sites

- **Web destinations that let users:**
  - Consume and produce content
    - Videos / photos / articles /…
    - tags / ratings / reviews /…
  - Engage in social activities with
    - friends / family / colleagues / acquaintances /…
    - people with similar interests / located in the same area /…

- **Questions today:**
  - Understand and explore user populations on those sites.
  - Extract useful content from user actions.

# Course Outline

- Nov 9th, 2016: Recommendation
- Nov 16th, 2016: Social data mining

# Last week's outline

- Recommender Systems
  - What are recommender systems and how do they work?
  - Example application: Hotlist Recommendation on Delicious
  - How are recommender systems evaluated?

- Recommendation challenges
  - Well-known challenges
  - Recommendation diversity
  - Group recommendation

# Social Data Mining Outline

- **Understand**: Mine and explore user groups
  - Target social content site: social rating sites and social tagging aites
- **Exploit**: Extract useful content from user actions
  - Taget social content site: social tagging site

# Collaborative rating systems

- **Places where users express their opinion on a content item in the form of a rating**
- **Of great interest to:**
  - Analysts who seek to explore users' opinion on items
  - End-users who seek to make choices, find similar/dissimilar users

# User data on collaborative rating systems

– a set of rating records:

<item attributes, user attributes, rating>

| 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*

# MovieLens and IMDb

| ID | | Title | Genre | Director | | Name | Gender | Location | | Rating |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Titanic | Drama | James Cameron | | Amy | Female | New York | | 8.5 |
| 2 | | Schindler's List | Drama | Steven Speilberg | | John | Male | New York | | 7.0 |

| | MovieLens (+IMDb) | BookCrossing |
|---|---|---|
| #Users | 6,040 | 38,511 |
| #Items | 3,900 | 260 |
| #Ratings | 1,000,209 (million) | 196,842 |
| Rating Scale | 1 to 5 | 1 to 10 |

# MovieLens
## *http://grouplens.org/datasets/movielens/*

### MovieLens 100k

100,000 ratings from 1000 users on 1700 movies.

- README.txt
- ml-100k.zip
- Index of unzipped files

### MovieLens 1M

1 million ratings from 6000 users on 4000 movies.

- README.txt
- ml-1m.zip

### MovieLens 10M

10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users.

- README.html
- ml-10m.zip

# Social data mining: definition

- **Define social data mining as group-based exploration**
  – because labeled user groups exhibit less sparsity and less noise than individual records
  – because labeled groups provide new insights
- **Group: set of rating records describable by a set of attribute values**

# Example user groups

- *Young people who rated Woody Allen movies*
- *Middle-aged females in California*
- *People who rated movies starring Scarlett Johannson*
- *Female engineers who rated Star Wars*
- *[25-35] year-old professionals who live in Grenoble and who rated movies starring Sean Penn*

# Social data mining problem

**Given a rating dataset, discover *good* groups**

# Challenges and outline

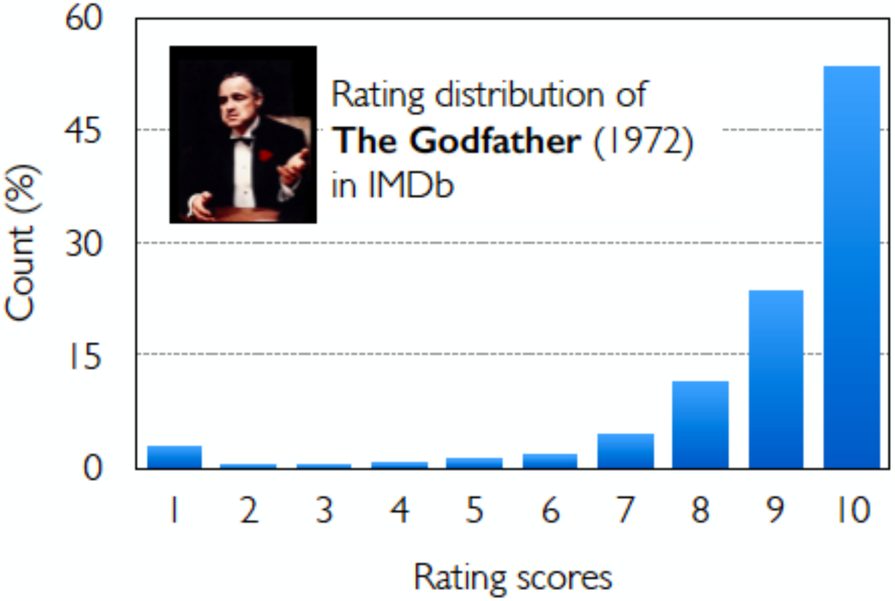
- **Challenges:**
  - How to express group quality?
  - How to find groups quickly?

- **Outline:**
  - One-shot exploration
  - Interactive exploration
  - Some perspectives

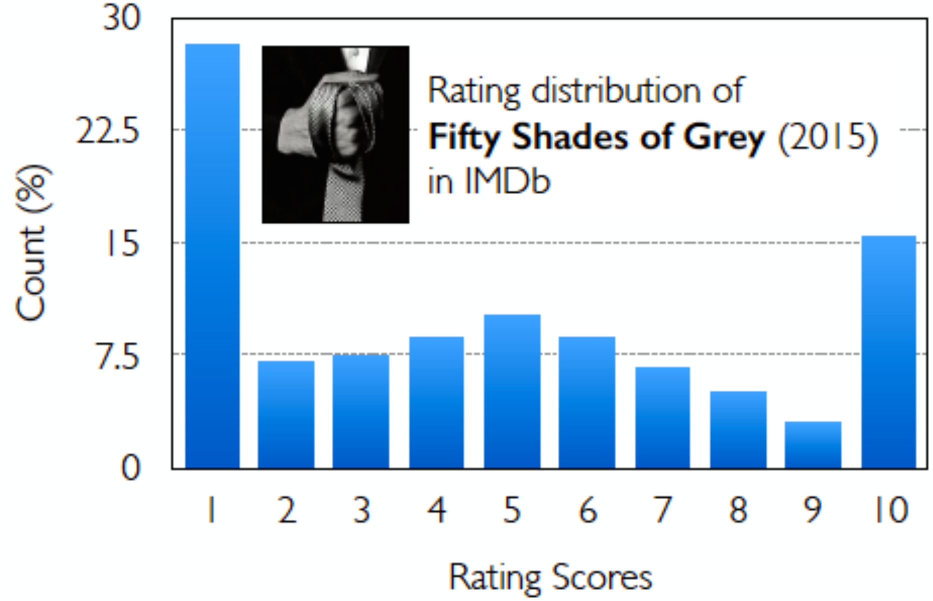
# *Pre-defined* user groups on IMDb

# What is a good group, locally speaking?

**Homogeneous Rating Distribution**




**Polarized Rating Distribution**



# What is a good group, globally speaking?

IMDb

*Ratings for romance genre movies.*

*I believe romantic movies are mostly watched by females.*

***Elena***
*social scientist*

*young females average rating:* ***3.7***

*variance:* ***2***

*females in DC average rating:* ***4.6***

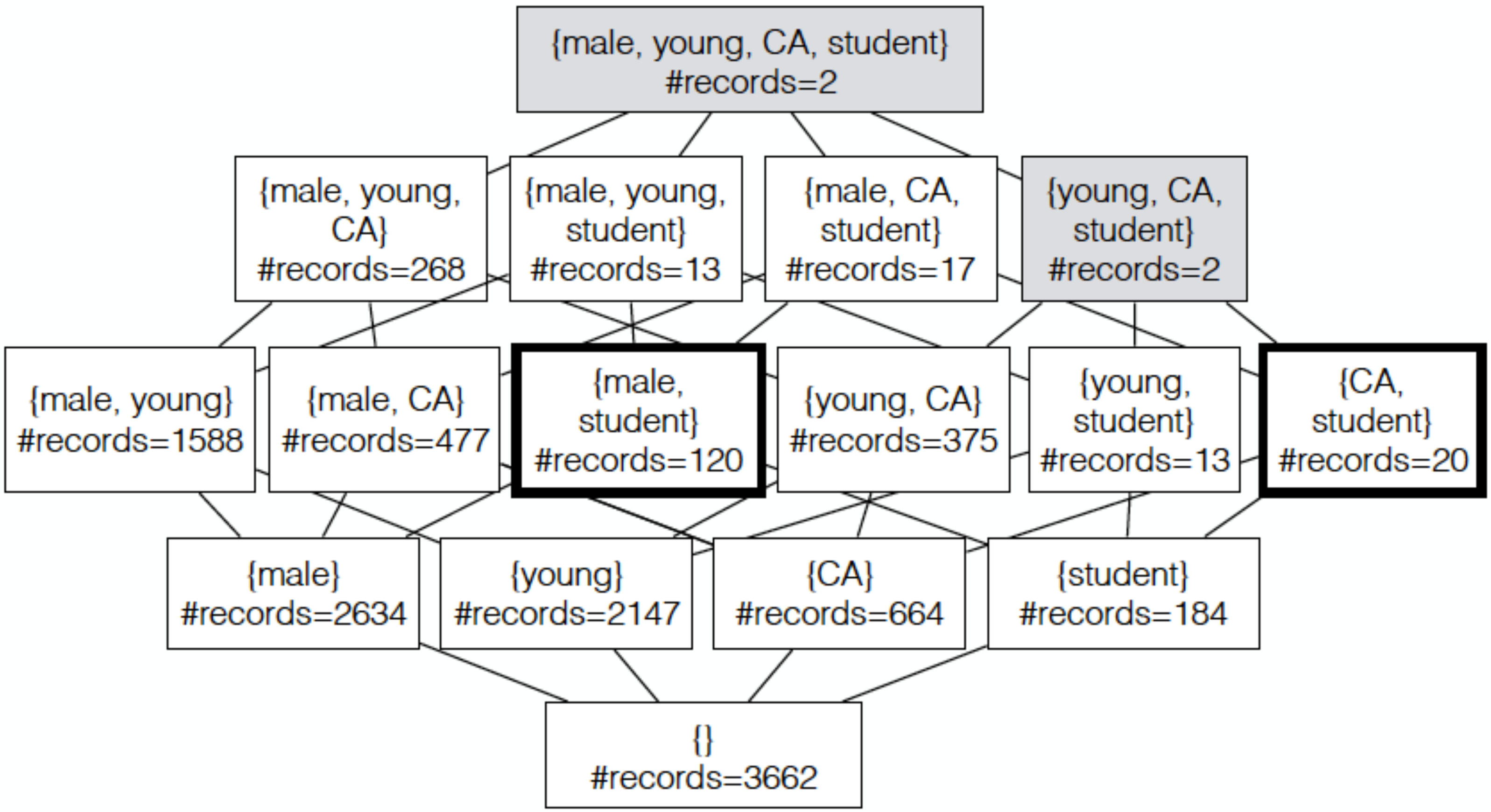
*variance:* ***1.5***

*male teenagers average rating:* ***3.1***

*variance:* ***3.4***

*— user groups that cover most ratings*

# Partial lattice for the movie Toy Story
# An exponential search space



# 1. An example of a one-shot formulation Meaningful Description Mining *[1]*

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

**description erro** $\boxed{\mathtt{error}(C, R_I)}$ **is minimized, subject to:**

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

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

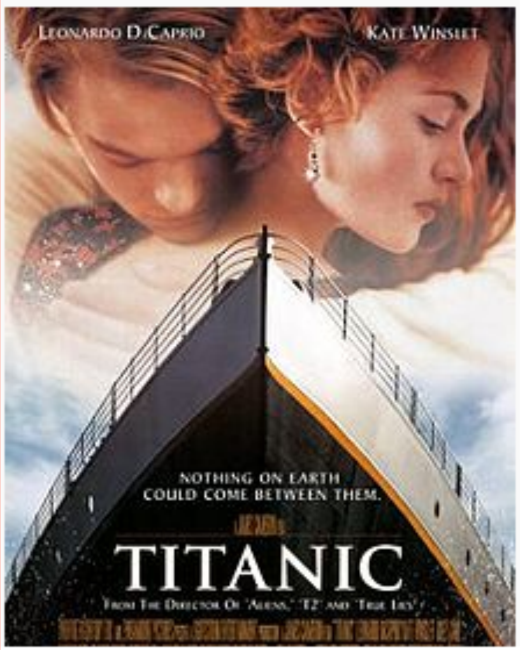
*[1] MRI: Meaningful Interpretations of Collaborative Ratings, S. Amer-Yahia, Mahashweta Das, Gautam Das and Cong Yu. In PVLDB 2011.*

# Meaningful Description Mining

## k=1

Identify groups of reviewers who consistently share similar ratings on items

*Titanic*




**Titanic** (1997)

PG-13 194 min - Adventure | Drama | History - 19 December 1997 (USA)

7.4 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**

# Meaningful Description Mining

## k=3

*Identify groups of reviewers who consistently share similar ratings on items*


**Black Swan** (2010)

R 108 min - Drama | Mystery | Thriller - 17 December 2010 (USA)

8.3 Ratings: **8.3**/10 from 156,148 users Metascore: **79**/100

Reviews: 892 user | 523 critic | 42 from Metacritic.com

**Young female reviewers love this movie, average rating: 9.3**

**Reviewers from New York love this movie, average rating: 8.7**

**Young male student reviewers hate this movie, average rating: 6.1**

# Meaningful Description Mining

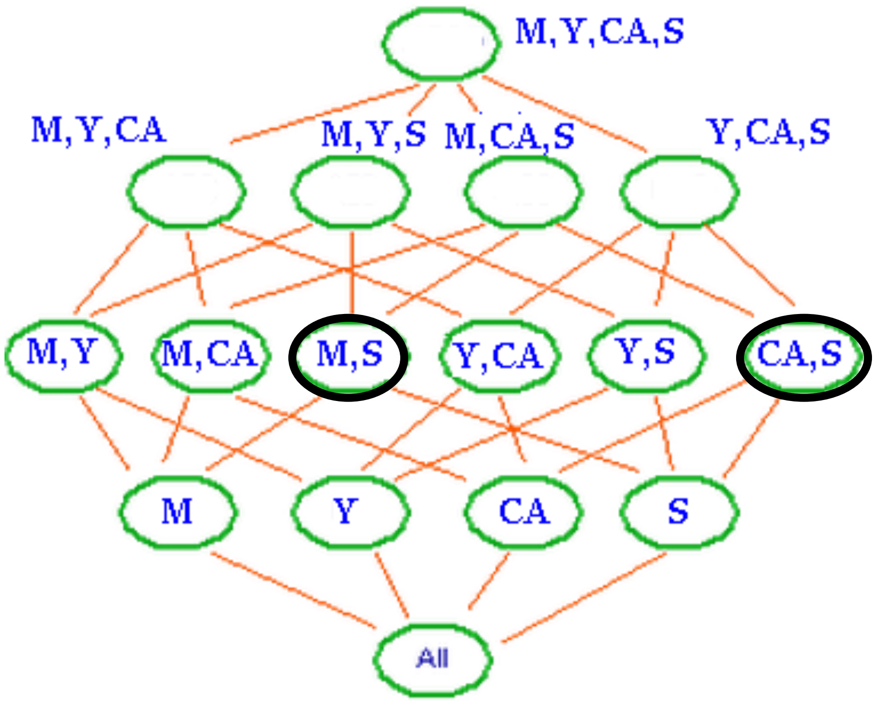
THEOREM 1. *The decision version of the problem of meaningful description mining (DEM) is NP-Complete even for boolean databases, where each attribute $ia_j$ in $\mathcal{I}_A$ and each attribute $ua_j$ in $\mathcal{U}_A$ takes either 0 or 1.*

*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* ☐

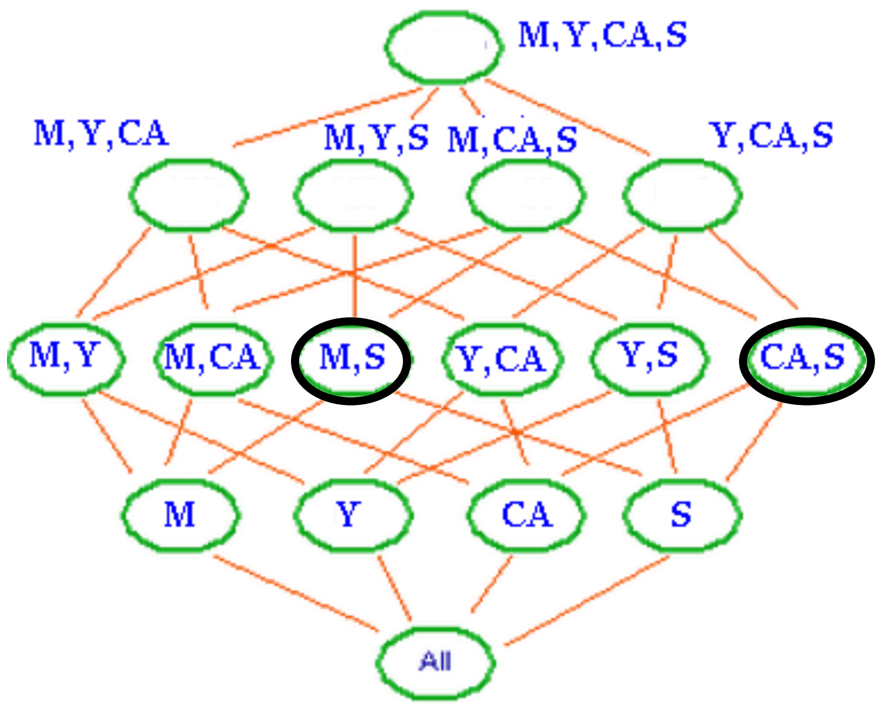



*C= {Male, Student}*
*{California, Student}*

# Random Restart Hill Climbing Algorithm

*Satisfy Coverage* ☐

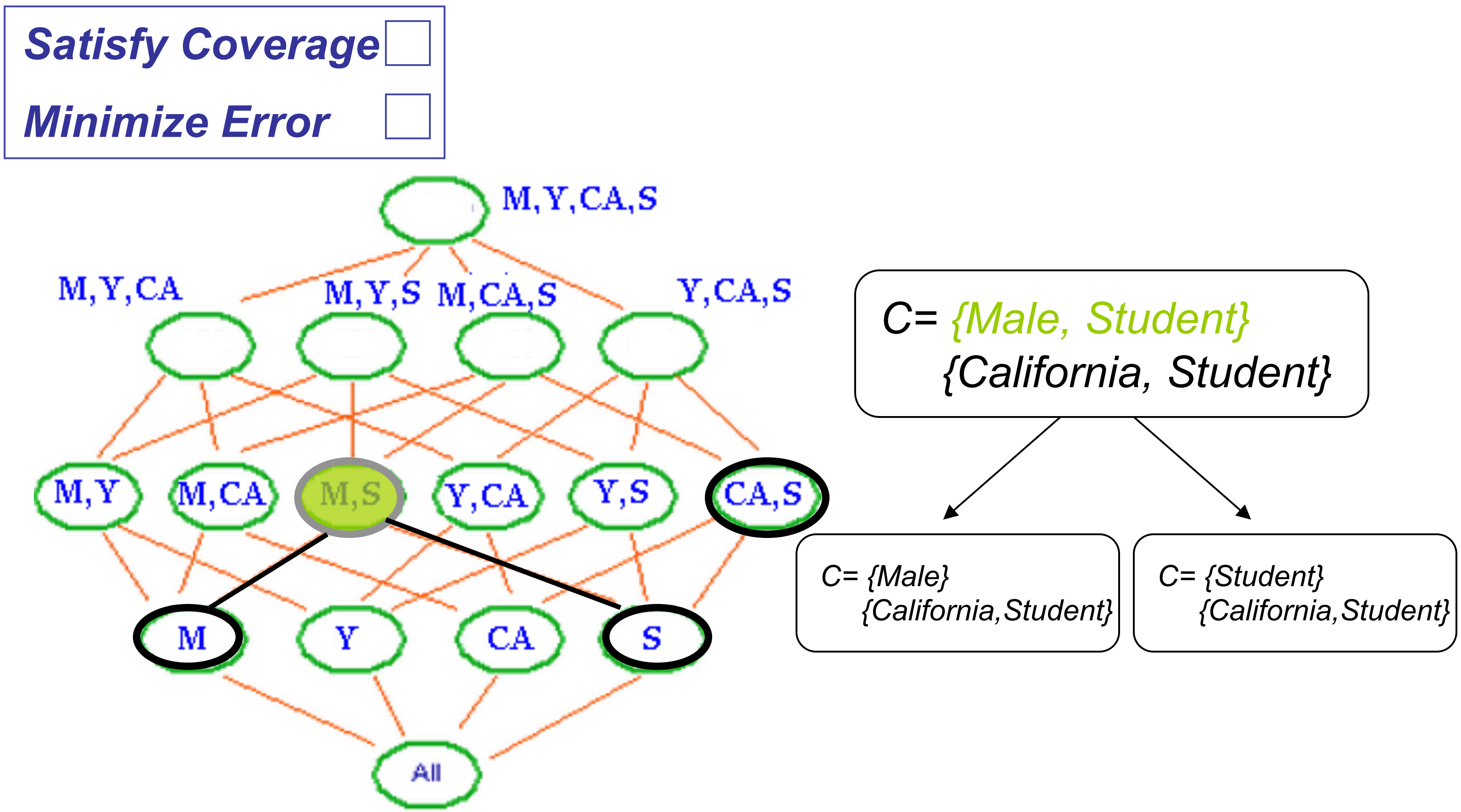
*Minimize Error* ☐




*C= {Male, Student}*
*{California, Student}*

*Say, C does not satisfy Coverage Constraint*

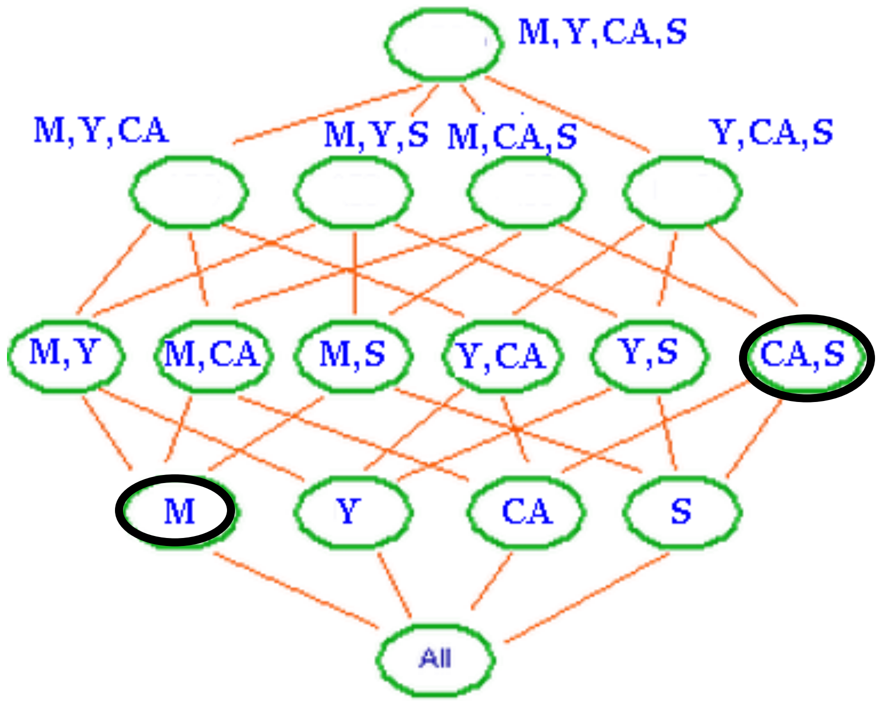
# Random Restart Hill Climbing Algorithm



# Random Restart Hill Climbing Algorithm

*Satisfy Coverage* √

*Minimize Error* ☐



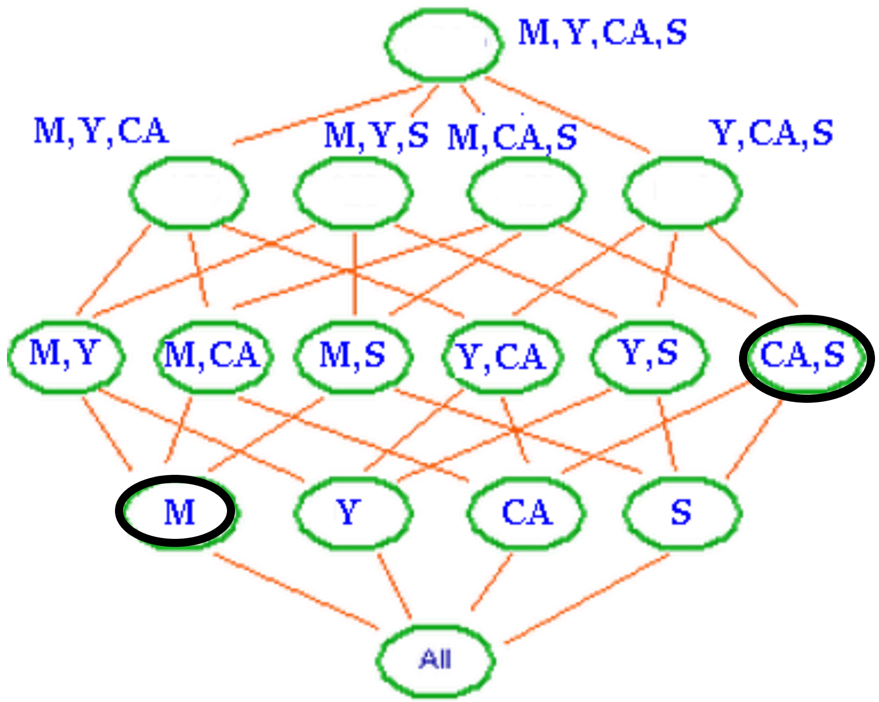

*C= {Male}*
*{California, Student}*

*Say, C satisfies Coverage Constraint*

# Random Restart Hill Climbing Algorithm

*Satisfy Coverage* √

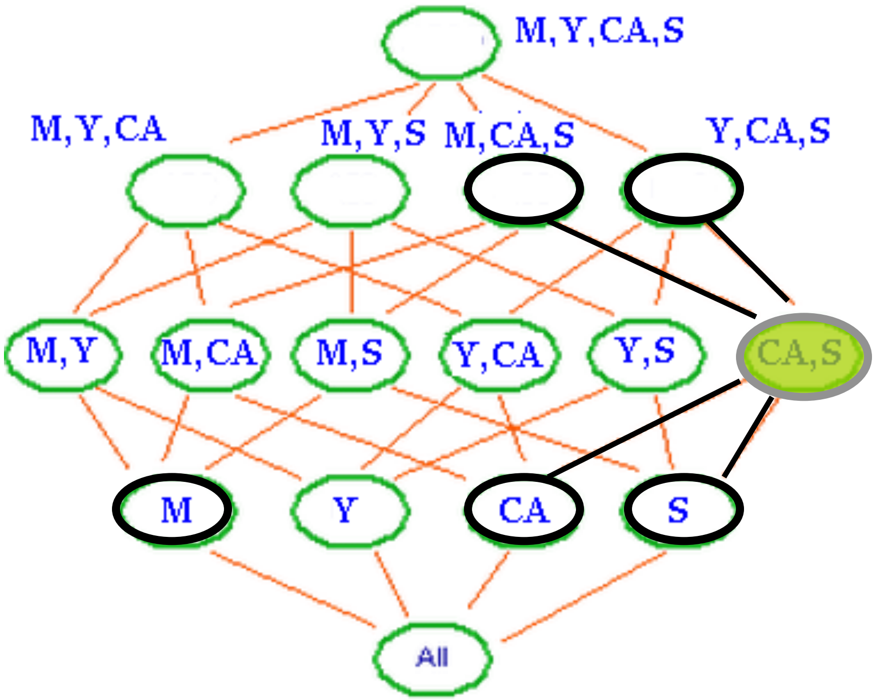
*Minimize Error*




*C= {Male}*
*{California, Student}*

# Random Restart Hill Climbing Algorithm

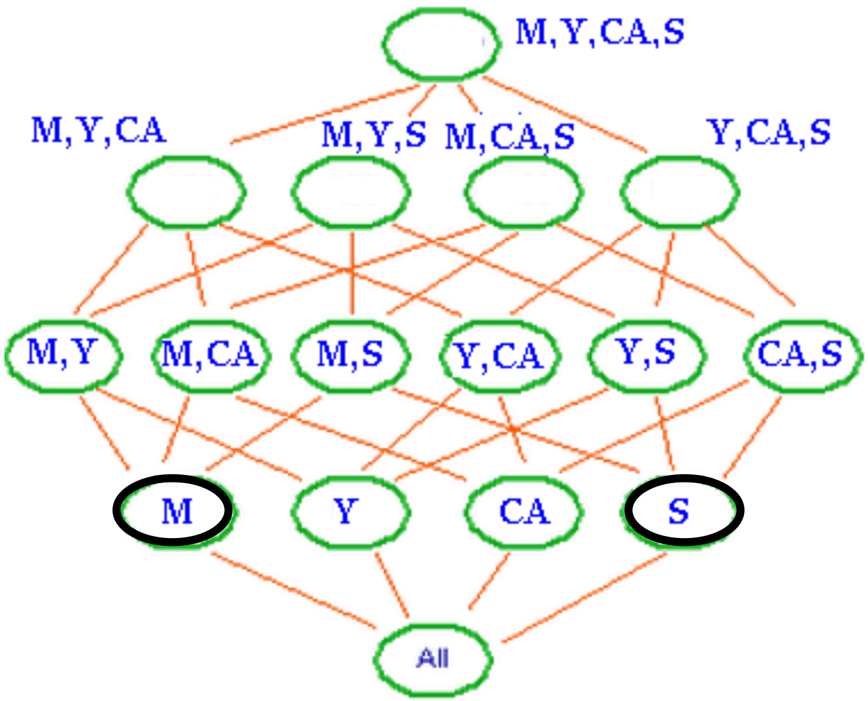
| | |
|---|---|
| *Satisfy Coverage* | √ |
| *Minimize Error* | |




*C= {Male}*
*{California, Student}*

# Random Restart Hill Climbing Algorithm

| | |
|---|---|
| *Satisfy Coverage* | √ |
| *Minimize Error* | √ |

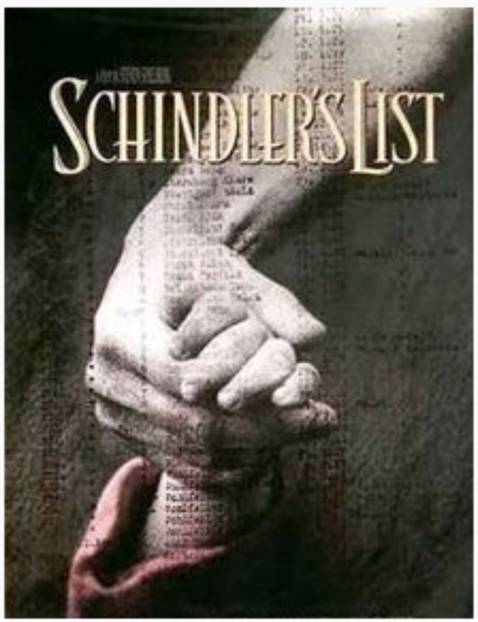



*C= {Male}*
*{Student}*

# 2. Another one-shot formulation
# Meaningful Difference Mining

Identify groups of reviewers who consistently disagree on item ratings

*Schindler's List*




**Schindler's List** (1993)

R 195 min - Biography | Drama | History - 15 December 1993 (USA)

8.9

Ratings: **8.9**/10 from 329,773 users Metascore: **93**/100

Reviews: 959 user | 95 critic | 23 from Metacritic.com

**Teen-aged female reviewers and male middle-aged reviewers have rated this movie inconsistently; their average rating: 7.5**

- **Middle-aged male reviewers love this movie, their average rating: 9.1**
- **Teen-aged female reviewers hate this movie, their average rating: 6.2**

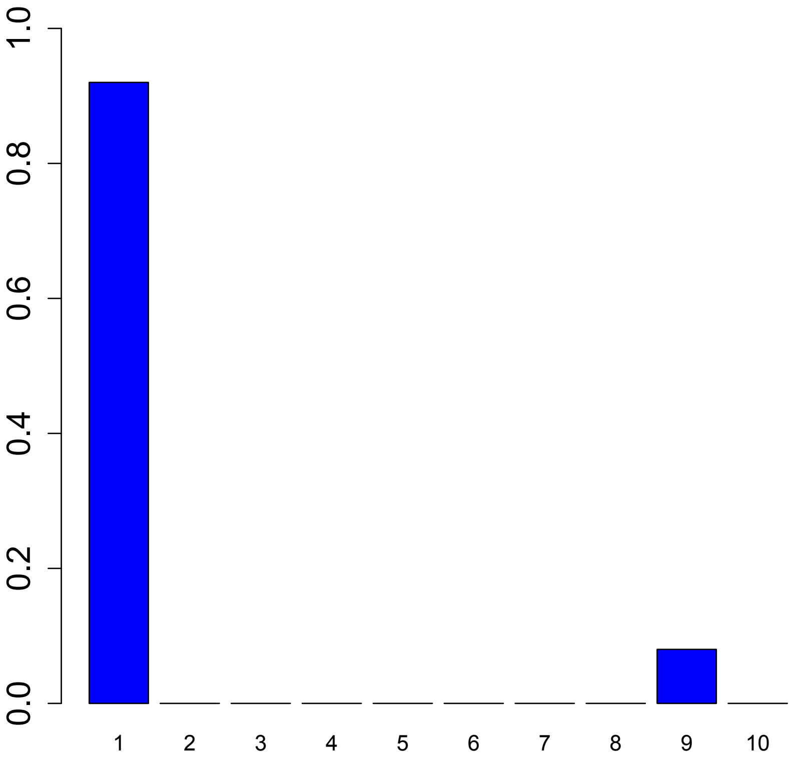
# 3. A third one-shot formulation People like/unlike me

Mary: 32 years, live in Bethlehem, USA
dislikes books by Debbie Macomber

- ✓ 25 middle-aged people, live in the **USA**, dislike “204 Rosewood Lane”
- ✗ 11 people, live in the **USA**, like “Changing Habits”

*We call this a rating map.*

# Mary's distribution for Debbie Macomber's books



# EMD as a rating comparison measure

***$\rho_1$ = [0.9, 0.025, 0.025, 0.025, 0.025]***
***$\rho_2$ = [0.025, 0.9, 0.025, 0.025, 0.025]***
***$\rho_3$ = [0.025, 0.025, 0.025, 0.025, 0.9]***

| Measure | $(\rho_1, \rho_2)$ | $(\rho_1, \rho_3)$ |
|---|---|---|
| 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 | 0.2500 | 0.2500 |
| Signal Noise Ratio | 2.0372 | 4.221 |
| Lukaszyk-Karmowski Metric | 1.1625 | 3.525 |
| EMD | 0.875 | 3.5 |

# People like me/unlike me problem *[2]*

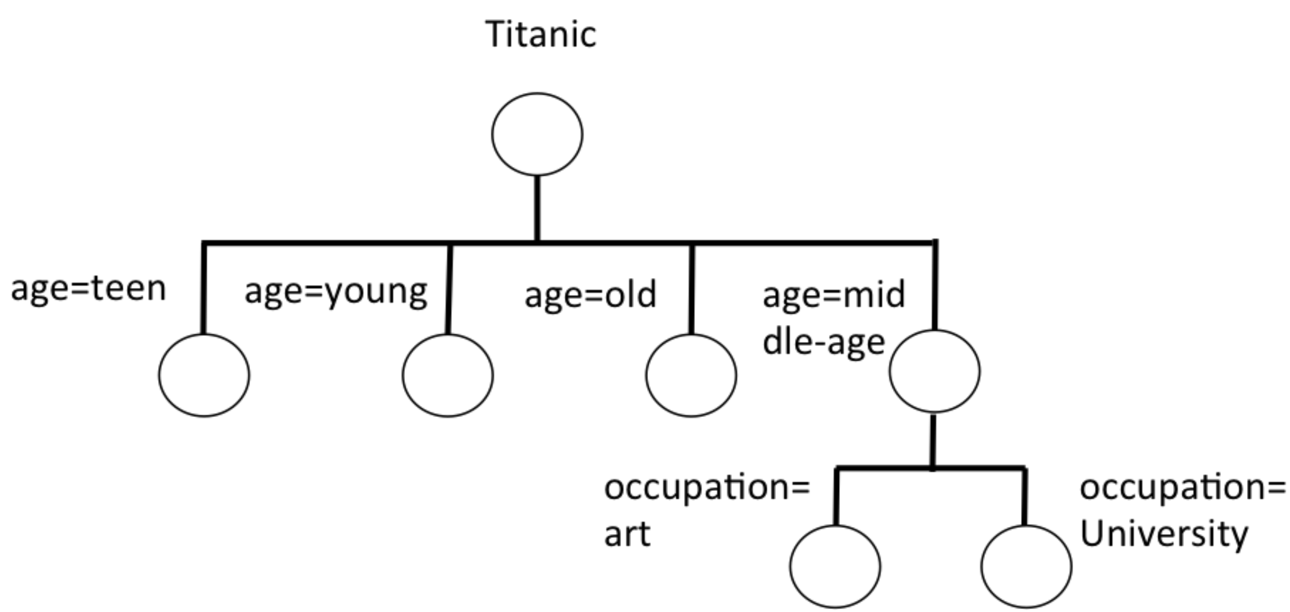
**Given a set of input distributions $\{\rho_1,\ldots,\rho_k\}$, find the *largest distinct user groups* whose distribution is the closest to one of $\{\rho_1,\ldots,\rho_k\}$ *(using an EMD threshold $\theta$)***

- *Groups with a shorter description are preferred*
- *Large groups are preferred*
- *Groups with different descriptions are preferred*

***[2] Exploring Rated Datasets with Rating Maps Sihem Amer-Yahia, Sofia Kleisarchaki, Naresh Kumar Kolloju, Laks V.S. Laskhmanan, Ruben H. Zamar (under review)***

# Partition Decision Tree (PDT)

- The set of rating records can be organized in a PDT
- Each node is a user group with a description



# Brief sketch of algorithms

- **DTAlg**
  - minimizes description length by finding a minimum height partition decision tree
  - classic decision trees driven by gain functions like entropy and gini-index.

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

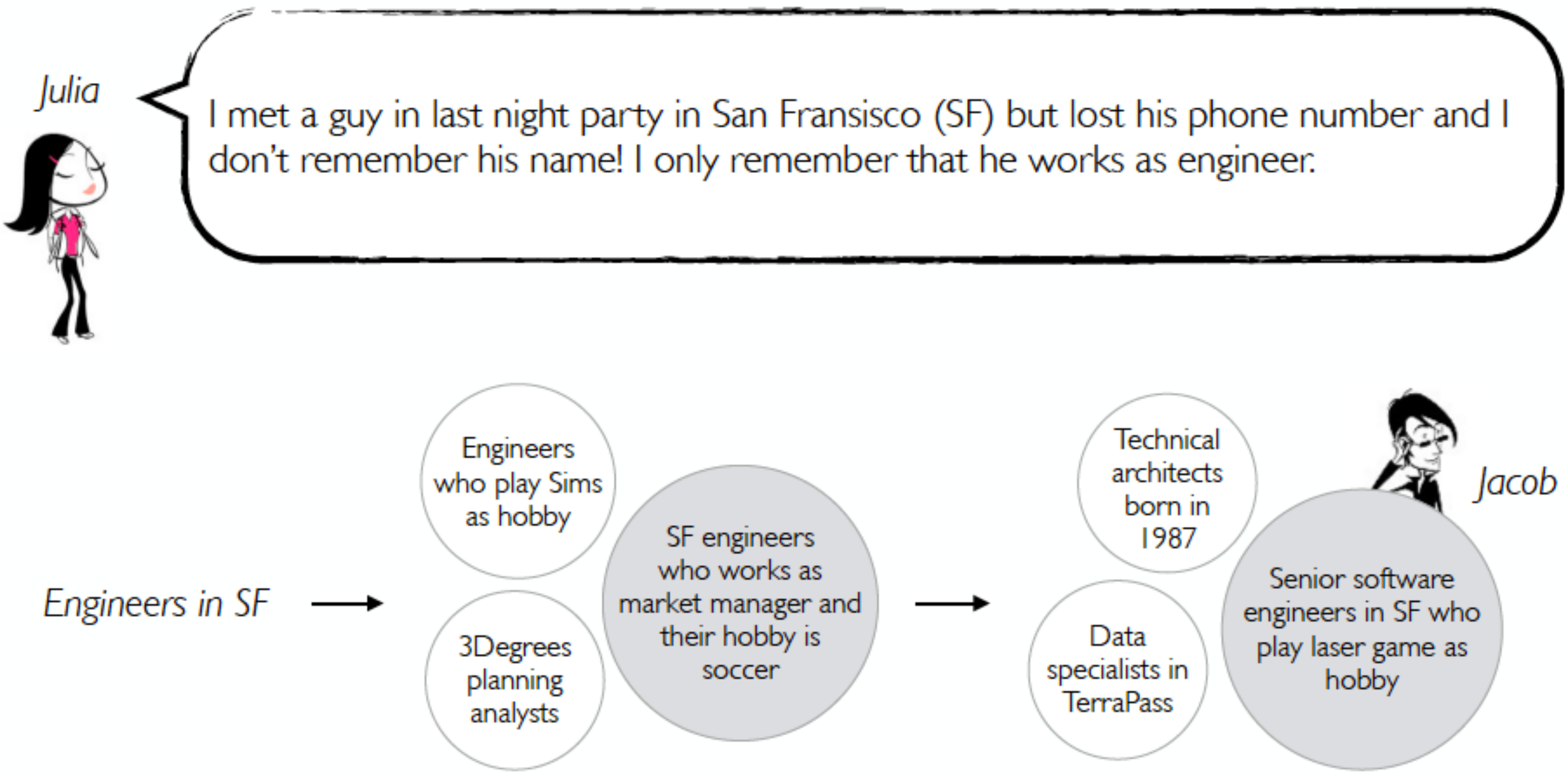
- **Random Forests**
  - splitting input dataset hurts coverage
  - runs multiple iterations of DTAlg with different splitting attributes and combines trees with RF-Cluster, RF-Desc, RF-Random, RF-Size, and RF-EMD

# Summary so far

1. **One-shot social exploration**
   – formulated as finding user groups
   – whose ratings are uniform/polarized
   – whose ratings are close to some input distribution
   – hard problems that necessitate appropriate heuristics

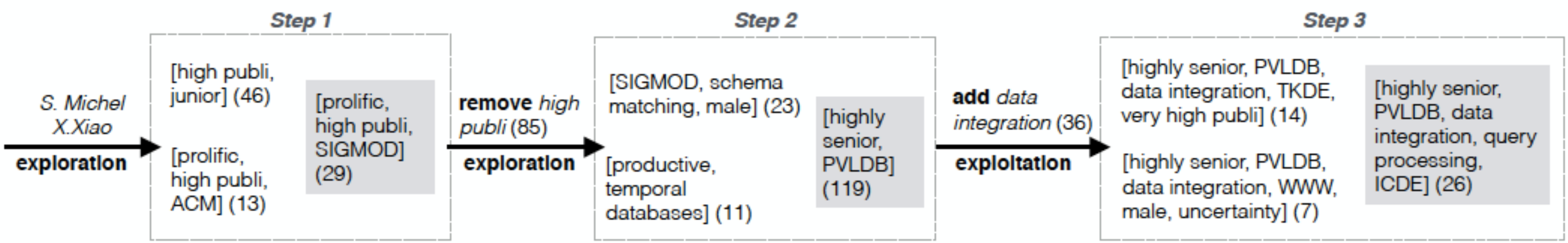
2. **Interactive social exploration**

# Interactive social exploration [3]




[3] *Interactive User Group Analysis. B. Omidvar-Tehrani, S. Amer-Yahia, A. Termier. CIKM 2015.*

# Interactive social exploration [3]




*[3] Interactive User Group Analysis. B. Omidvar-Tehrani, S. Amer-Yahia, A. Termier. CIKM 2015.*

# Some open questions

- **Immediate:**
  - personalized exploration [4]
- **A benchmark for evaluating interactive exploration**
  - Generic: number of steps
  - Task-driven: offline and online qualitative studies

*[4] 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.*

# Social data exploration instances

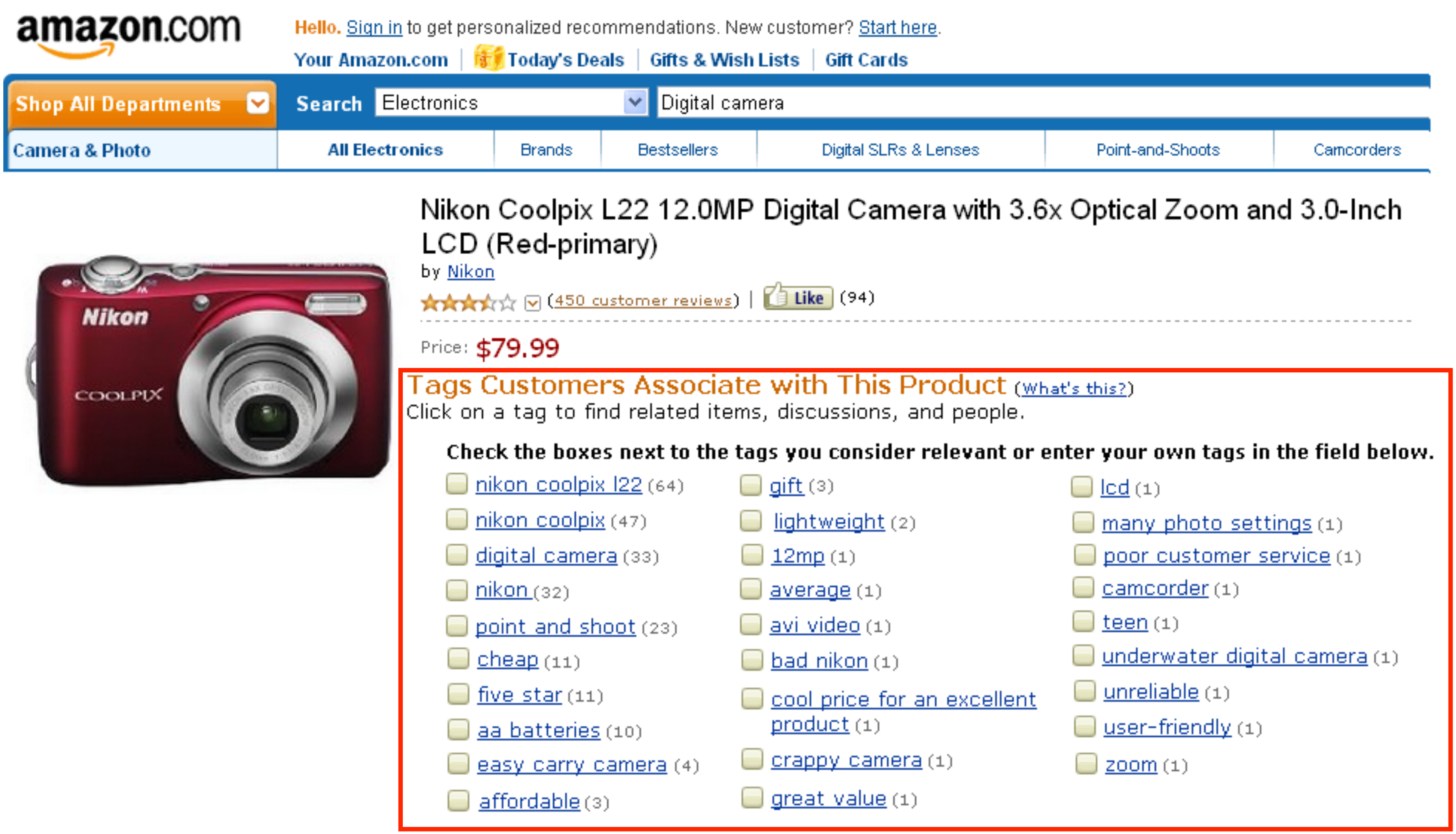
- **Since analysts do not know what to look for, let's examine some social data exploration instances**
  - Rating exploration

    **MRI: Meaningful Interpretations of collaborative Ratings**
    *with **M. Das, S. Thirumuruganathan, G. Das (UT Arlington), C. Yu (Google)***
    *at VLDB 2011*
  - Tag exploration

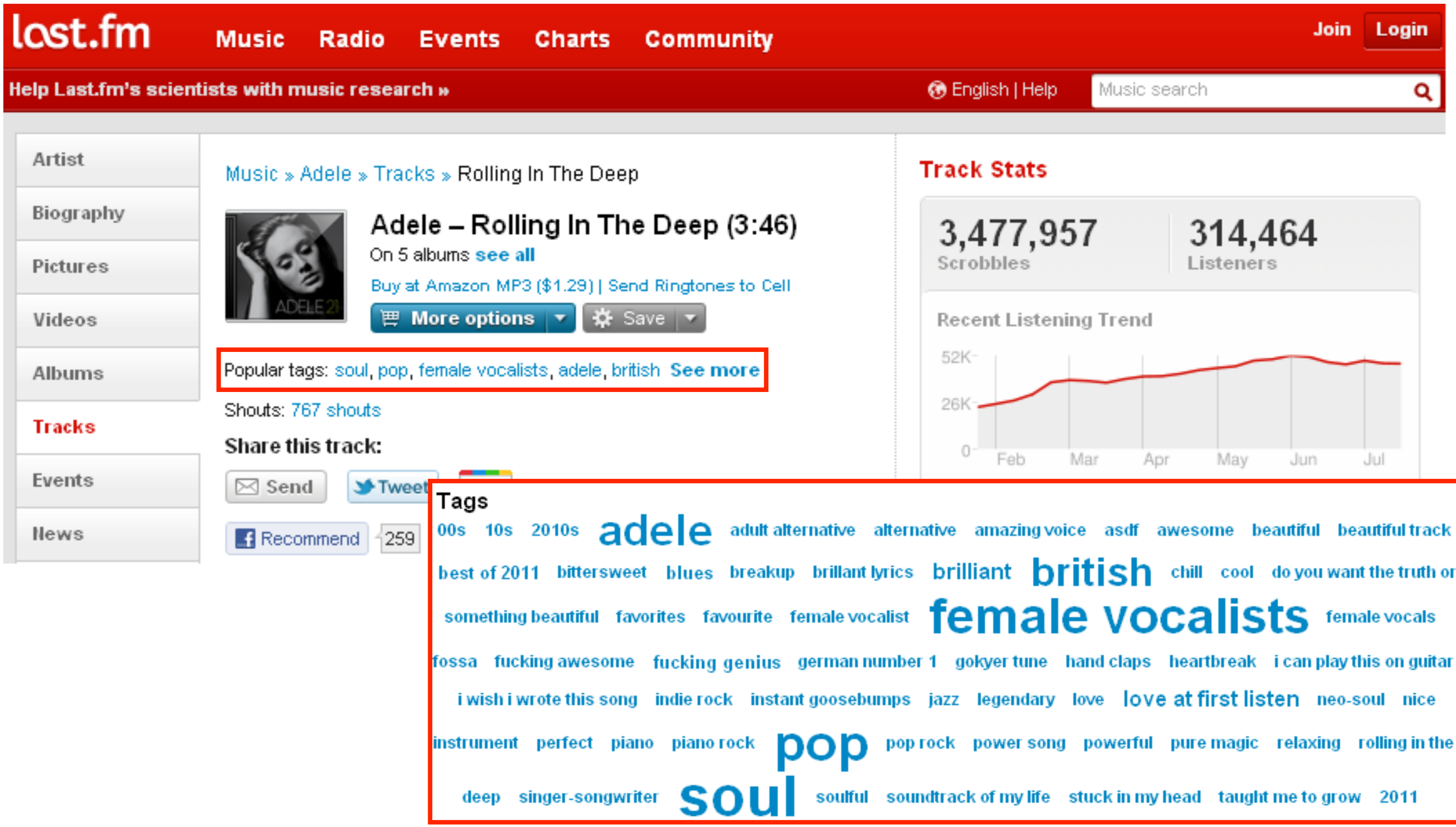
    **Who tags what? An analysis framework**
    *with **M. Das, S. Thirumuruganathan, G. Das (UT Arlington), C. Yu (Google)***
    *at VLDB 2012*
  - Temporal exploration

    **Efficient sentiment correlation for Large-scale Demographics**
    with ***M. Tsytsarau and T. Palpanas (Univ. of Trento)***
    *at SIGMOD 2013*

# Collaborative tagging system (Amazon)



# Collaborative tagging system (LastFM)



# MovieLens instances (with tags) *[3]*

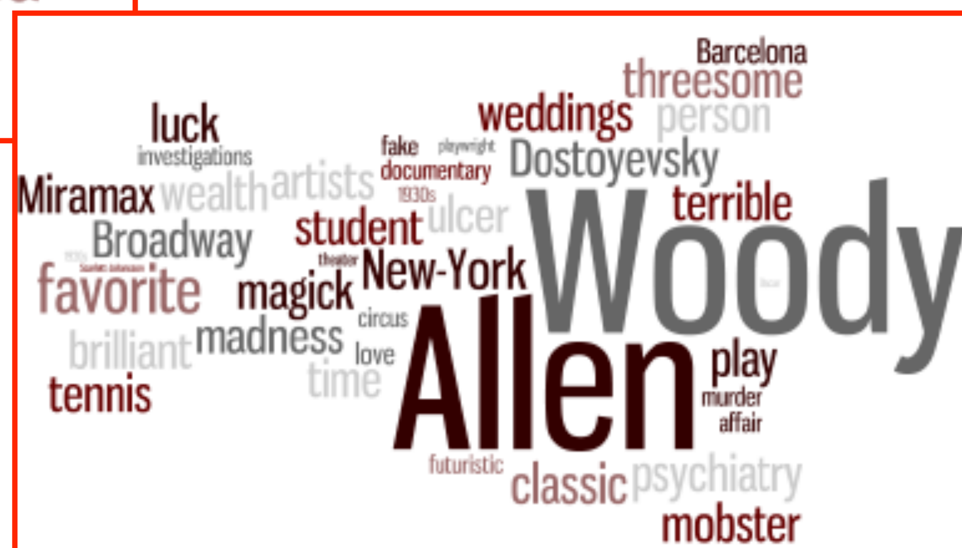
| ID | | Title | Genre | Director | | Name | Gender | Location | | Tags |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Titanic | Drama | James Cameron | | Amy | Female | New York | | love, Oscar |
| 2 | | Schindler's List | Drama | Steven Speilberg | | John | Male | New York | | history, Oscar |

*[3] An expressive framework and efficient algorithms for the analysis of collaborative tagging. Mahashweta Das, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das, Cong Yu. VLDB J. 23(2): 201-226 (2014)*

# Exploring collaborative tagging in MovieLens


*Tag Signature for all Users*

Woody Allen




*Tag Signature for all CA Users*

# Exploring collaborative tagging in MovieLens

**Identify similar groups of reviewers who share similar tagging behavior for different items**

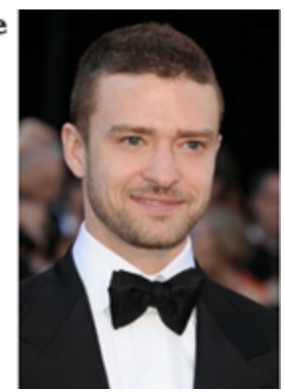
***Male Young***

**Jennifer Aniston**

***comedy, drama, friendship***

**Justin Timberlake**

***drama, friendship***



# Exploring collaborative tagging in MovieLens

**Identify diverse groups of reviewers who have different tagging behavior for the same items**

***Male Teen***

*gun, special effects*

***Female Teen***

*violence, gory*


**Genre: Action**

Most Popular Action Feature Films

1. Prometheus (2012) — Add to Watchlist
7.9/10
A team of explorers discover a clue to the origins of mankind on Earth, leading them on a journey to the darkest corners of the universe. There, they must fight a terrifying battle to save the future of the human race.
Dir: Ridley Scott With: Noomi Rapace, Logan Marshall-Green, Michael Fassbender
Action | Horror | Sci-Fi — 124 mins. R

2. The Avengers (2012) — Add to Watchlist
8.6/10
Nick Fury of S.H.I.E.L.D. brings together a team of super humans to form The Avengers to help save the Earth from Loki and his army.
Dir: Joss Whedon With: Robert Downey Jr., Chris Evans, Scarlett Johansson
Action | Adventure | Sci-Fi — 143 mins. PG-13

3. Snow White and the Huntsman (2012) — Add to Watchlist
6.7/10
In a twist to the fairy tale, the Huntsman ordered to take Snow White into the woods to be killed winds up becoming her protector and mentor in a quest to vanquish the Evil Queen.
Dir: Rupert Sanders With: Kristen Stewart, Chris Hemsworth, Charlize Theron
Action | Adventure | Drama | Fantasy — 127 mins. PG-13