New Perspectives in Social Data Management

Sihem Amer-Yahia

Research Director

CNRS @ LIG

Sihem.Amer-Yahia@imag.fr

TSUKUBA University

May 20th, 2014

Traditional data management stack

Application Layer

High-level specification

Logical Layer

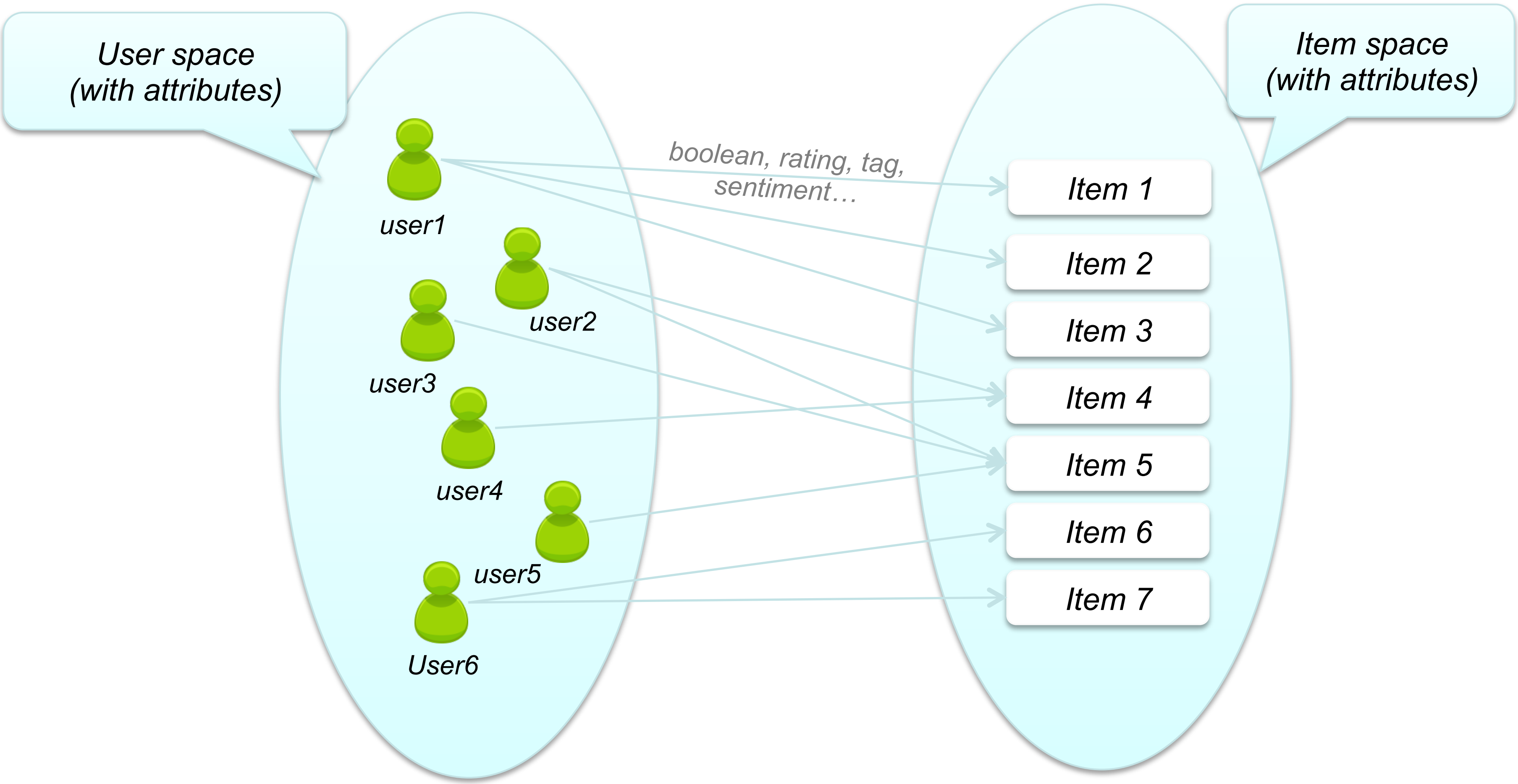
Logical and physical optimizations

Physical Layer

Index creation

relational tables++
native XML backend

Collaborative data model



Let’s examine a canonical social application

Extracting travel itineraries from Flickr

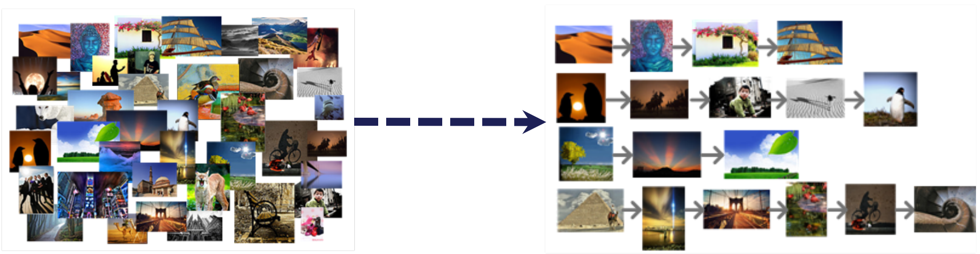
Goal: extract the itinerary of each traveler by mapping photos into Points Of Interest (POIs) and aggregate actions of many travelers into coherent queryable itineraries

***Automatic construction of travel itineraries using social breadcrumbs*: with Munmun De Choudhury (Arizona State University), Moran Feldman (Technion), Nadav Golbandi, Ronny Lempel (Yahoo! Research), Cong Yu (Google Research)**. *HyperText Conference 2010*

***Interactive Itinerary Planning*: with Senjuti Basu Roy (Univ. of Washington), Gautam Das (Univ. of Texas at Arlington), Cong Yu (Google Research)**. *ICDE 2011*

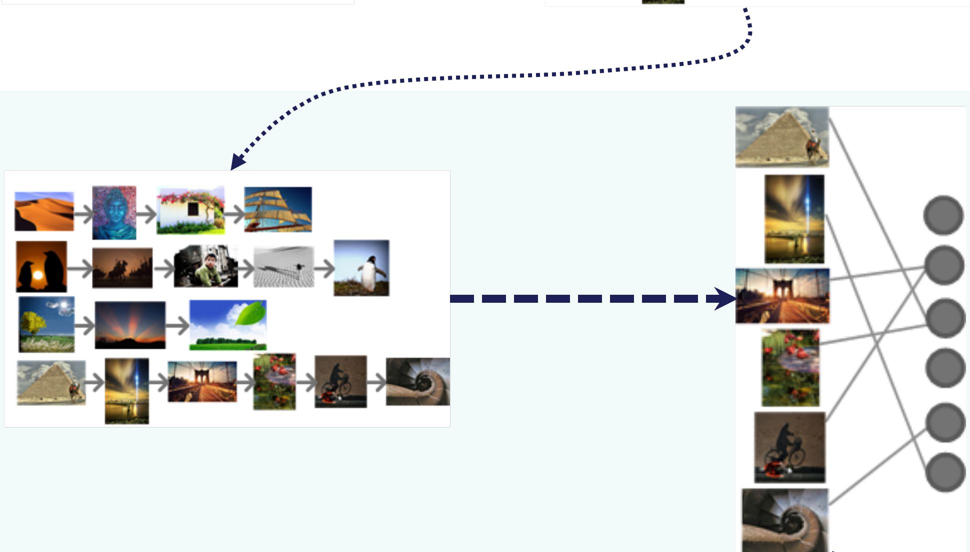
Deployed on Yahoo! Mobile

Photo Streams



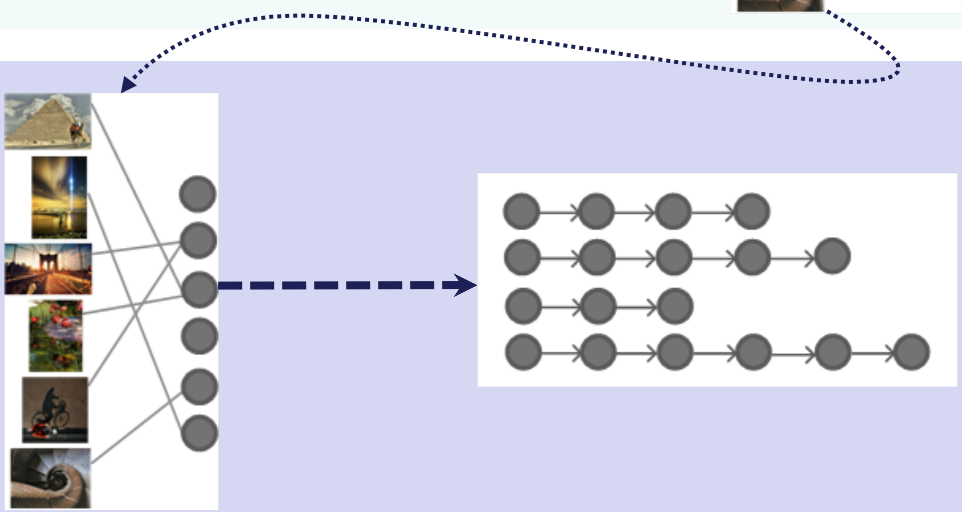
- *Identify photos of a given city*
- *Filter out residents of a city*
- *Validate photo timestamps*

Photo-POI Mapping



- *Extract Candidate POIs*
  - *Lonely Planet/Y! Travel to extract landmarks*
  - *Yahoo! Maps API to retrieve their geo-locations*
- *Tag & geo-based POI association*

Timed Paths



- *Photo Streams Segmentation*
  - *Split the stream whenever the time difference between two successive photos is "large"*
- *Distillation of Timed Visits*
  - *<POI, start time, end time>*
- *Construction of Timed Paths*
  - *A sequence of Timed Visits*

Problem definition

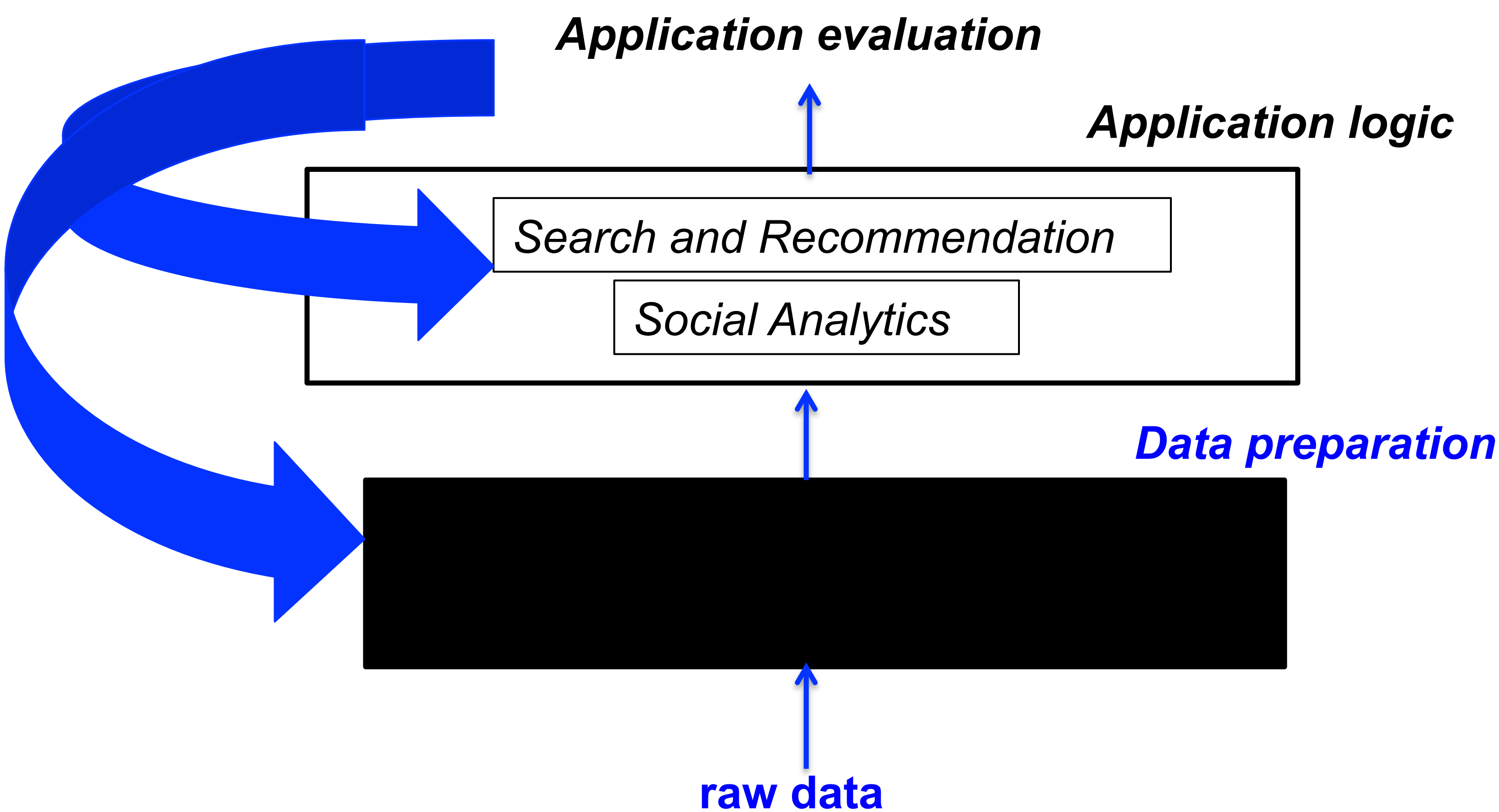
- **Definitions**
  - Each itinerary is a timed path
  - The set of timed paths implies a *weighted graph* G over POIs
  - An *itinerary* is a path in the graph G
  - The *value* of an itinerary is the sum of popularities of its POIs
  - The *time* of an itinerary is the sum of POI visit and transit times

- **Problem Instance ("Orienteering")**
  - Find an itinerary in G from a *source* POI to a *target* POI of budget (=time) at most B maximizing total value
  - The time budget B is typically whole days
  - *source* and *target POIs* provided by user (e.g. hotel)

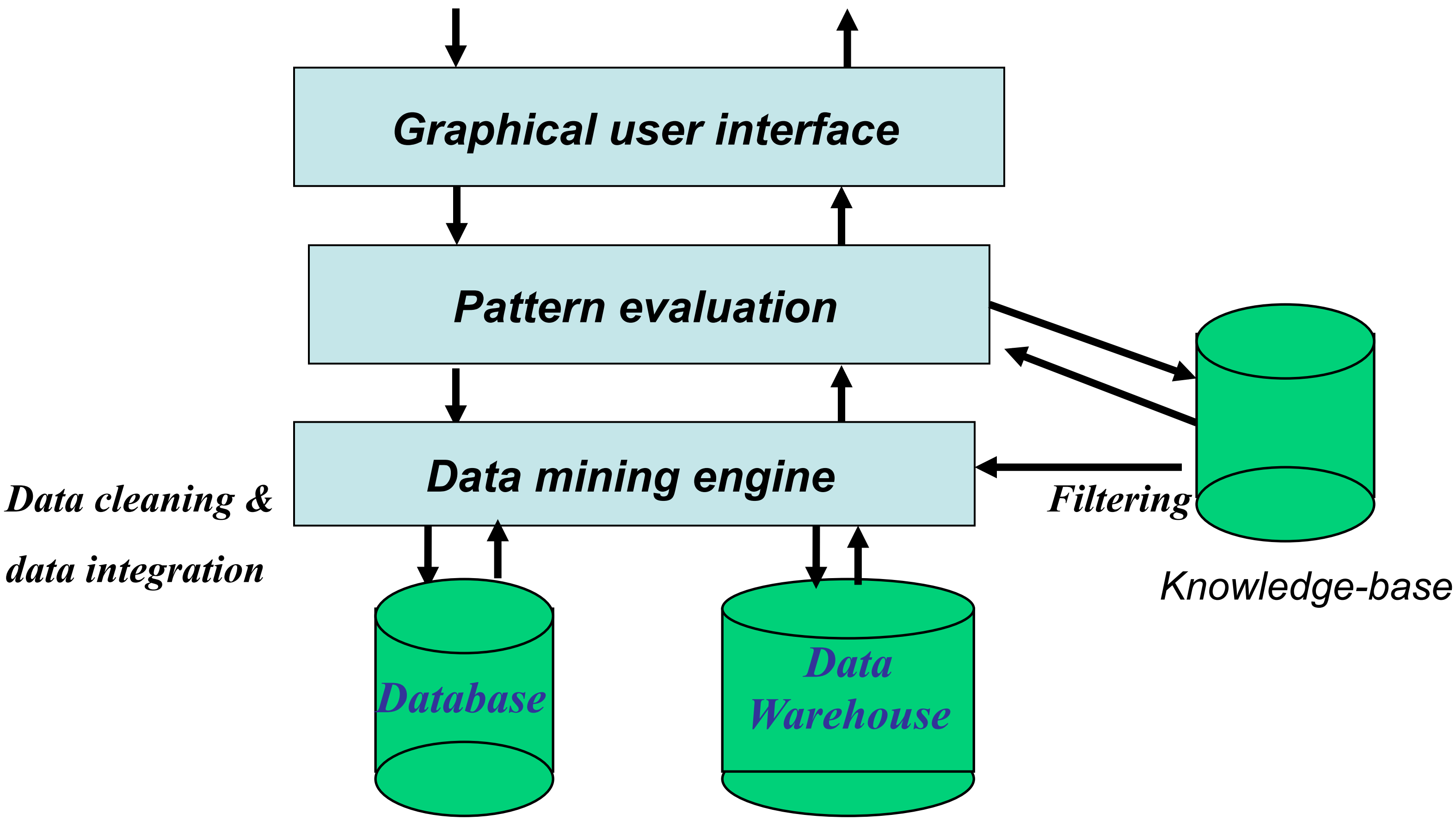
Example itinerary for NYC (single-day)

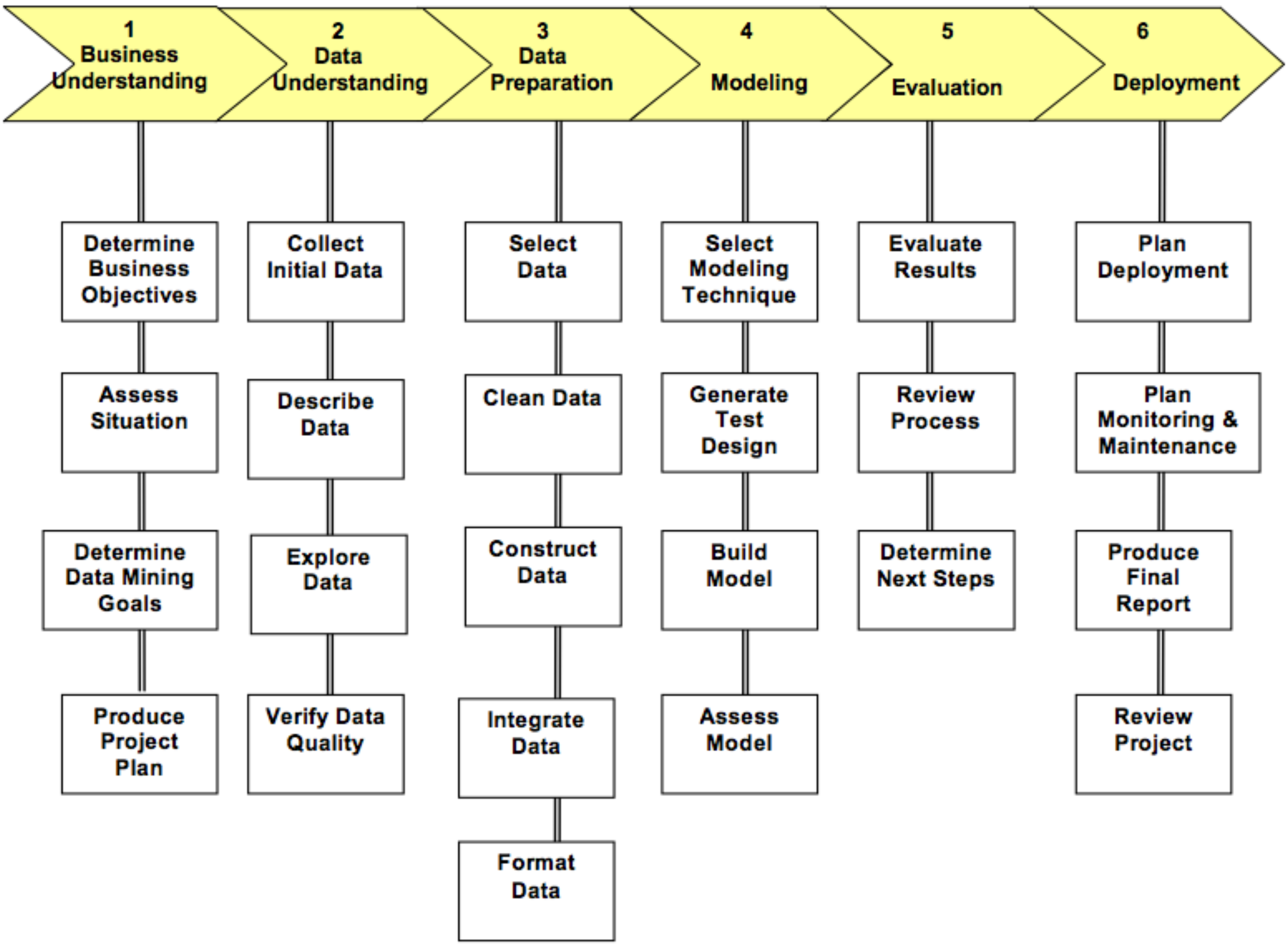
Time **09:00** : Start from **ground zero**
Time **09:00** : Spend 27 minutes at **ground zero**.
Time **09:27** : Transit to **empire state building** (estimated travel time: 52 minutes)
Time **10:19** : Spend 1 hour and 13 minutes at **empire state building**.
Time **11:32** : Transit to **new york public library** (estimated travel time: 15 minutes)
Time **11:47** : Spend 29 minutes at **new york public library**.
Time **12:16** : Transit to **radio city music hall** (estimated travel time: 24 minutes)
Time **12:43** : Spend 51 minutes at **radio city music hall**.
Time **13:34** : Transit to **central park** (estimated travel time: 23 minutes)
Time **13:57** : Spend 40 minutes at **central park**.
Time **14:37** : Transit to **rockefeller center** (estimated travel time: 33 minutes)
Time **15:10** : Spend 37 minutes at **rockefeller center**.
Time **15:47** : Transit to **grand central terminal** (estimated travel time: 22 minutes)
Time **16:09** : Spend 27 minutes at **grand central terminal**.
Time **16:36** : Transit to **chrysler building** (estimated travel time: 6 minutes)
Time **16:42** : Spend 31 minutes at **chrysler building**.
Time **17:13** : Transit to **brooklyn bridge** (estimated travel time: 32 minutes)
Time **17:45** : Spend 36 minutes at **brooklyn bridge**.
Time **18:21** : Transit to **statue of liberty** (estimated travel time: 21 minutes)
Time **18:42** : Spend 42 minutes at **statue of liberty**.
Time **19:24** : Transit to **little korea** (estimated travel time: 26 minutes)
Time **19:50** : Spend 31 minutes at **little korea**.
Time **20:21** : Transit to **ground zero** (estimated travel time: 38 minutes)

Social data management stack



Architecture of a typical Data Mining system

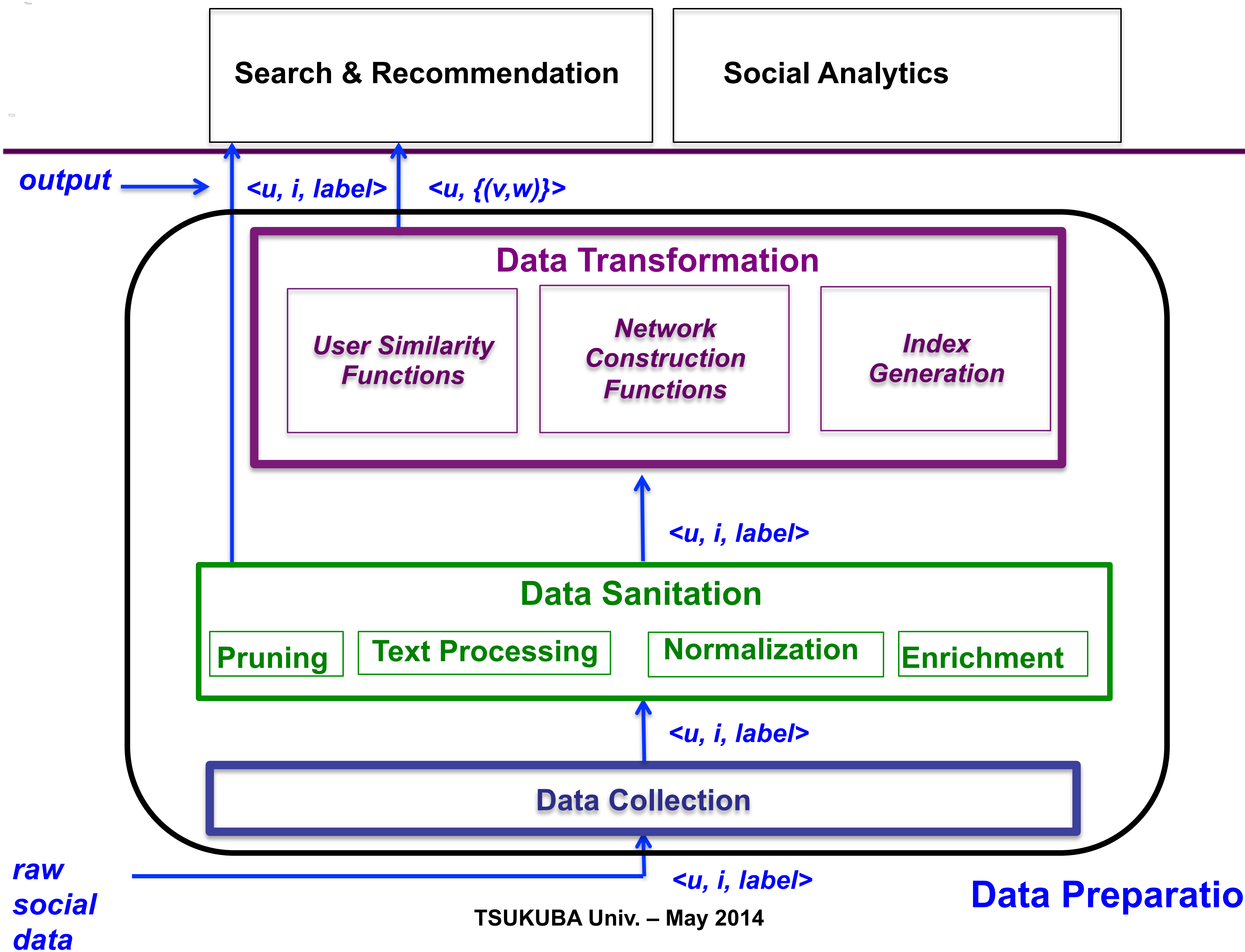



1 Business Understanding
2 Data Understanding
3 Data Preparation
4 Modeling
5 Evaluation
6 Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Format Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plan Monitoring & Maintenance
Produce Final Report
Review Project

SOCLE: A framework for social data preparation

with ***N. Ibrahim, C. Kamdem-Kengne, F. Uliana, M.C. Rousset***
submitted for publication

- Examined typical social applications: *URL recommendation in Delicious, group recommendation in MovieLens, social analytics on Twitter, itinerary extraction in Flickr*
  - **Data Collection**
    - mapping data into <*u,i,label*> triples
  - **Data Sanitation**
    - **Pruning**: cut long tails of user actions, remove photos taken by residents – *in delicious, removing URLs tagged with less than 5 tags reduces input data to 27% of input size*
    - **Text processing**: topic extraction
    - **Normalization**: of ratings– *in MovieLens, critics are more moderate than less-active reviewers*
    - **Enrichment**: POI-to-photo association, named entity extraction, twitter vocabulary expansion *(e.g., using Yahoo! Boss interface)*, sentiment analysis
  - **Data Transformation**
    - from <*u,i,label*> to <*u,i,label*> and <*u,{(v,w)}*> …



Data Preparation

SOCLE model

- Which data model? an extensible type system
- Which storage model?

SOCLE model and algebra

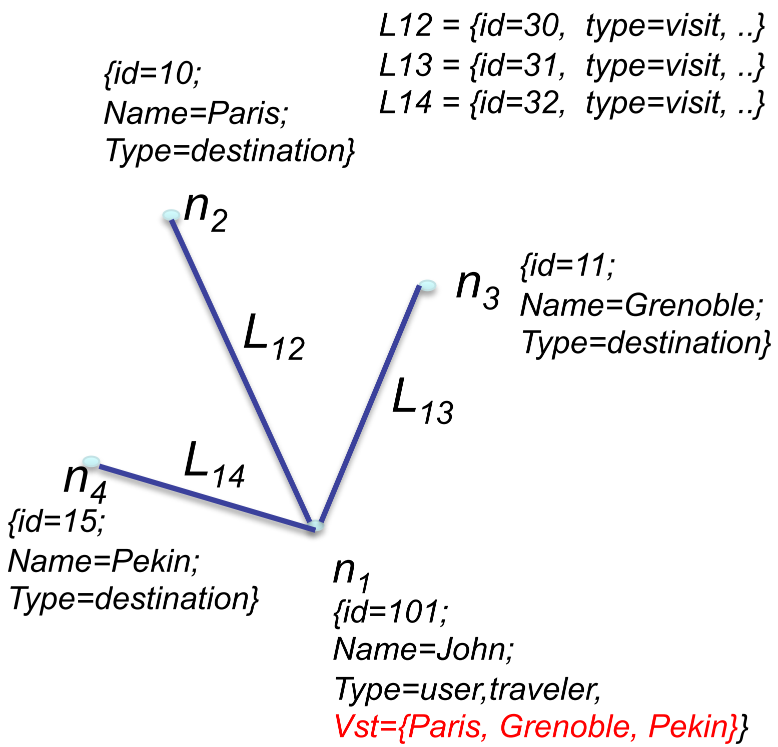
*with **L. Lakshmanan and C. Yu***

SocialScope: Enabling Information Discovery on Social Content Sites

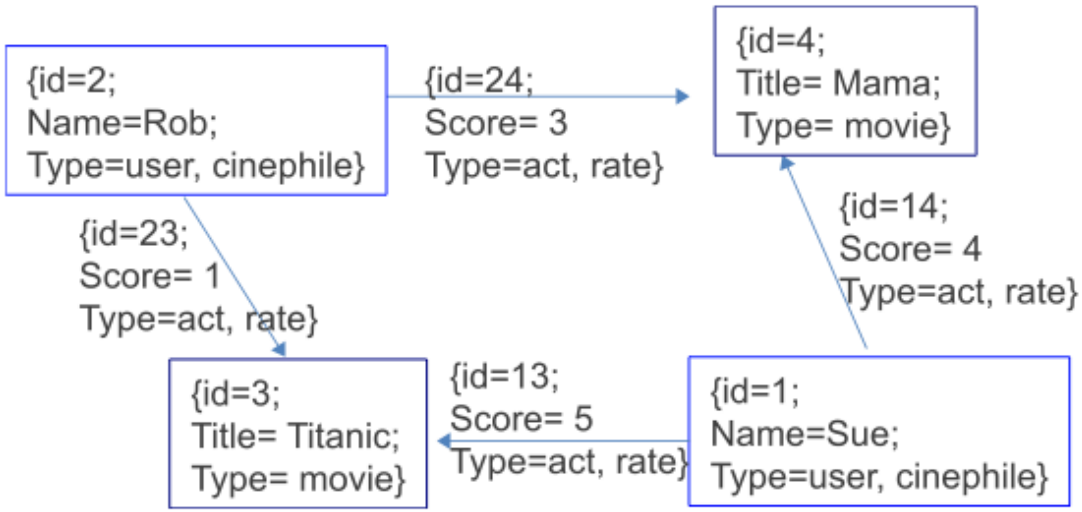
at CIDR 2009

Enrich a node with attributes -> new node type

- Algebra operator : $\gamma^{N}_{C,d,att,A}(G)$



Storage Model: native or relational++?



$T_{users}(idu, name)$

1	Sue
2	Rob

$T_{movies}(idm, title)$

3	Titanic
4	Mama

$T_{ratings}(idu, idm, rate)$

1	3	5
1	4	4
2	3	1
2	4	3

SOCLE algebra

- Examine how existing algebras/languages for querying social data can be used for data preparation
- Properties
  - Declarativity
  - Expressivity and closure
  - Provenance
  - Invertibility

What makes SDM different from DM?

- SDM needs a different data management stack: *data preparation*
- In social computing, analysts do not always know what to look for
- In social computing, application output must be evaluated

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

Rating exploration

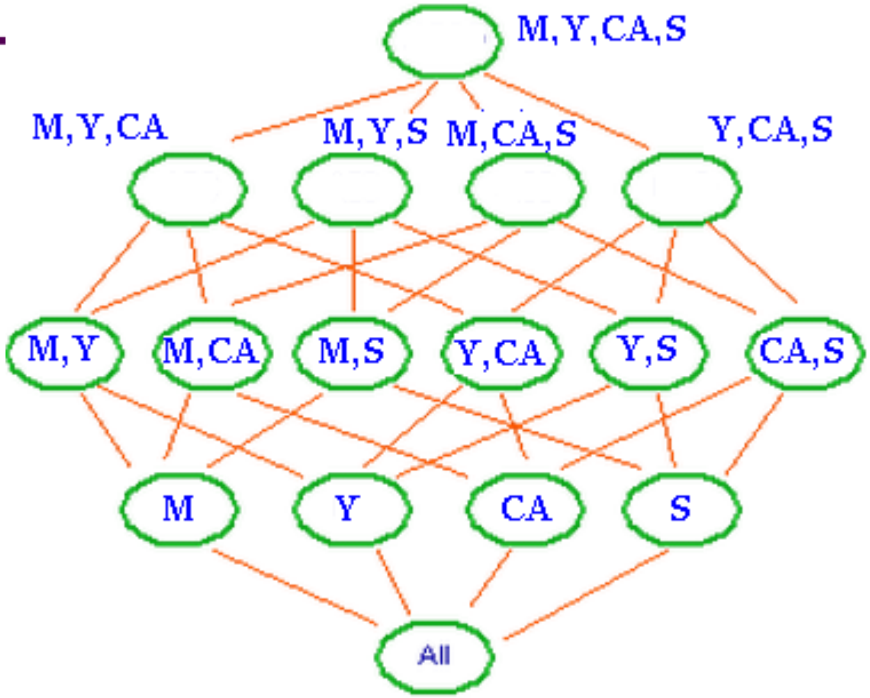
Collaborative rating model

- Rating tuple: <item attributes, user attributes, rating>

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

- Group: a set of ratings describable by a set of attribute values
  - Based on data cubes in OLAP (for mining multidimensional data)

Exploration space



Partial Rating Lattice for a Movie

(M:Male, Y:Young, CA:California, S:Student)

Each node/data cube/ cuboid in lattice is a group

Example group:
Gender: Male
Age: Young
Location: CA
Occupation: Student

Task
Quickly identify "good" groups in the lattice that help users understand ratings effectively


IMDb
The Internet Movie Database
Search All Go
Movies TV News Videos Community IMDb
YOU DON'T GET TO 500 MILLION FRIENDS WITHOUT MAKING A FEW ENEMIES
The Social Network (2010)
PG-13 120 min - Biography | Drama - 1 October 2010 (USA)
8.0
Ratings: 8.0/10 from 146,847 users Metascore: 95/100
Reviews: 522 user | 460 critic | 42 from Metacritic.com
A chronicle of the founding of
networking Web site.
Director: David Fincher
Writers: Aaron Sorkin (screen
Stars: Jesse Eisenberg, Andre
Timberlake
Watch Trailer
Add to
Votes Average
Males 117,061 8.1
Females 22,183 7.9
Aged under 18 6,419 8.5
Males under 18 4,776 8.6
Females under 18 1,576 8.2
Aged 18-29 97,085 8.2
Males Aged 18-29 80,738 8.2
Females Aged 18-29 15,516 7.9
Aged 30-44 30,346 7.8
Males Aged 30-44 26,297 7.8
Females Aged 30-44 3,687 7.7
Aged 45+ 6,005 7.6
Males Aged 45+ 4,657 7.7
Females Aged 45+ 1,272 7.3
IMDb staff 43 8.2
Top 1000 voters 475 7.5
US users 32,848 8.3
Non-US users 95,401 8.0

DEM: Meaningful Description Mining

- **For an input item covering *R_I* ratings, return set C of cuboids, s.t.:**
  - description error $\boxed{\mathtt{error}(C, R_I)}$ is minimized, subject to:
    - $|C| \leq k$;
    - coverage $\boxed{\mathtt{coverage}(C, R_I)} \geq \alpha$

Description Error: how well a cuboid average rating approximates the numerical score of each individual rating belonging to it

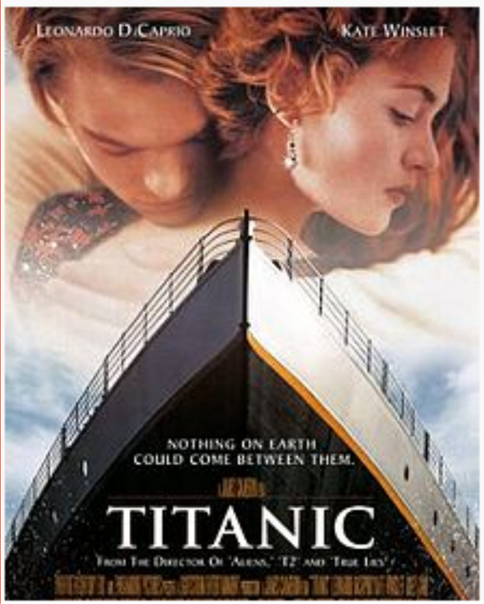
$$\boxed{\begin{aligned} \mathtt{error}(C, R_I) &= \Sigma_{r \in R_I}(E_r) \\ &= \Sigma_{r \in R_I} \mathtt{avg}(|r.s - \mathtt{avg}_{c \in C \wedge r \lessdot c}(c)|) \end{aligned}}$$

Coverage: percentage of ratings covered by the returned cuboids

DEM: Meaningful Description Mining

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

DEM: 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.

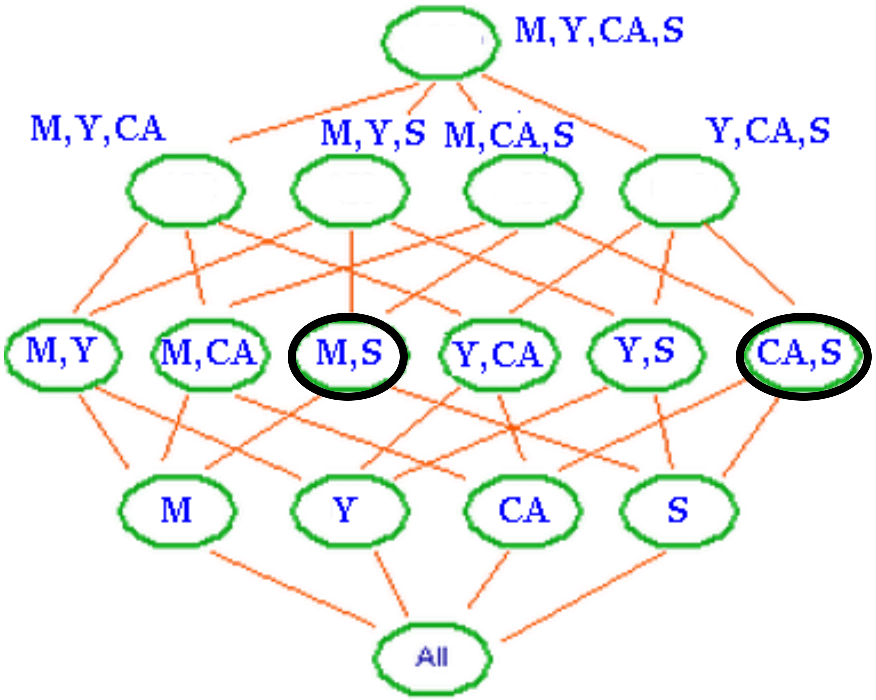
DEM Algorithms

- **Exact Algorithm (E-DEM)**
  - Brute-force enumerating all possible combinations of cuboids in lattice to return the exact (i.e., optimal) set as rating descriptions
- **Random Restart Hill Climbing Algorithm**
  - Often fails to satisfy Coverage constraint; Large number of restarts required
  - Need an algorithm that optimizes both Coverage and Description Error constraints simultaneously
- **Randomized Hill Exploration Algorithm (RHE-DEM)**

RHE-DEM Algorithm

Satisfy Coverage ☐

Minimize Error ☐



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

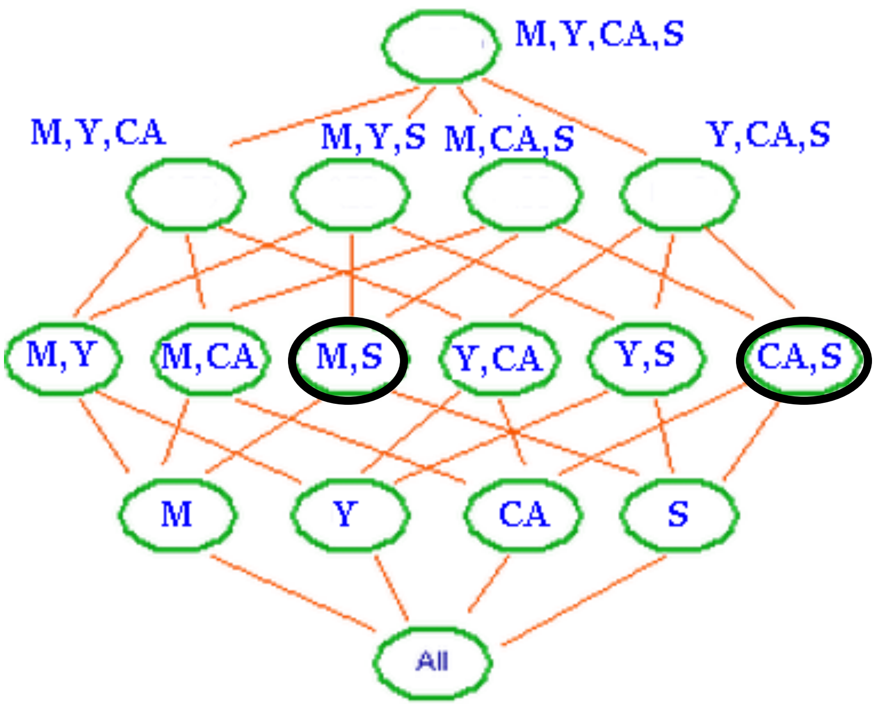
Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage

Minimize Error



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

Say, C does not satisfy Coverage Constraint

Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage	☐
Minimize Error	☐

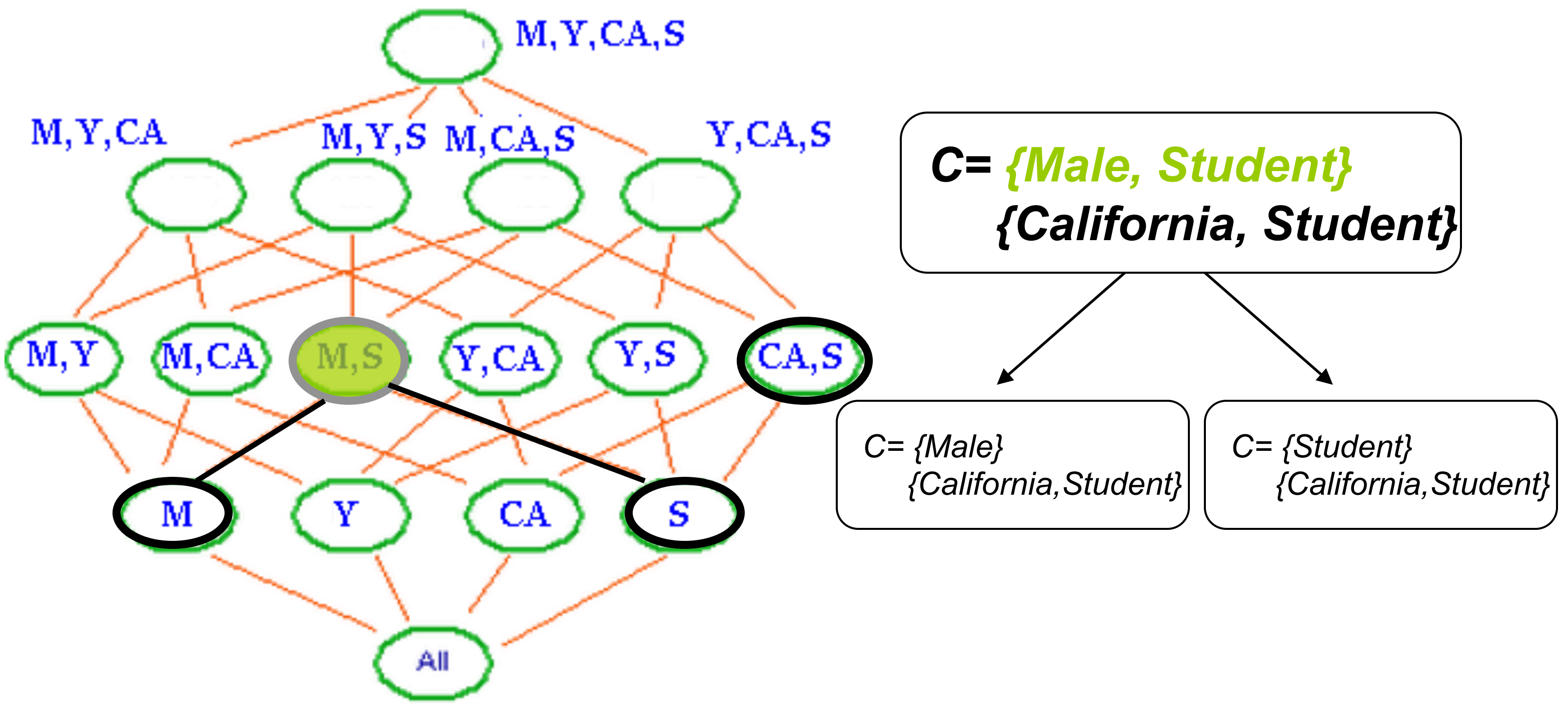


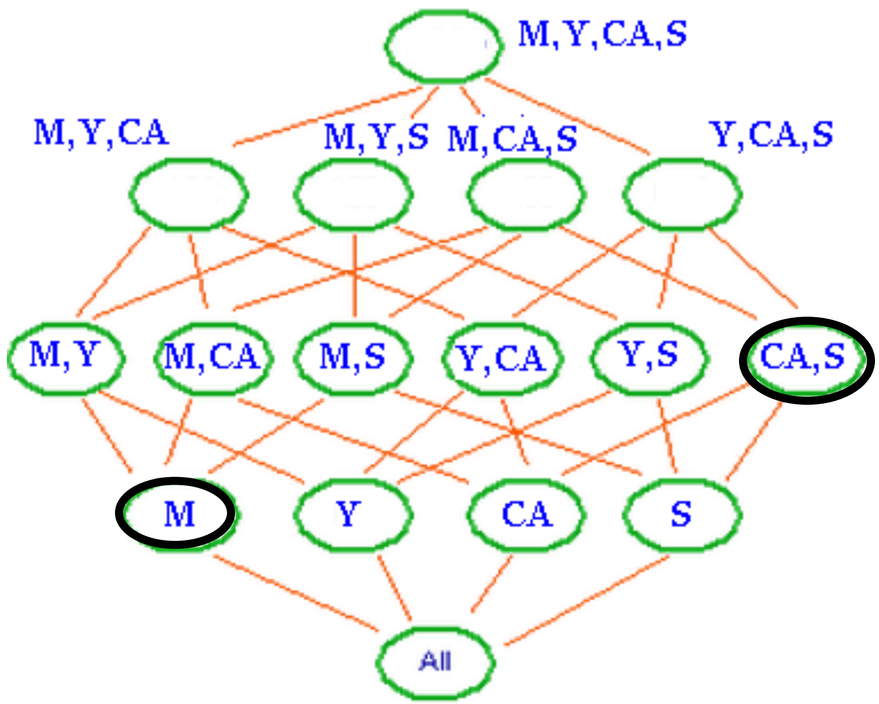
Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error



C= {Male}
{California, Student}

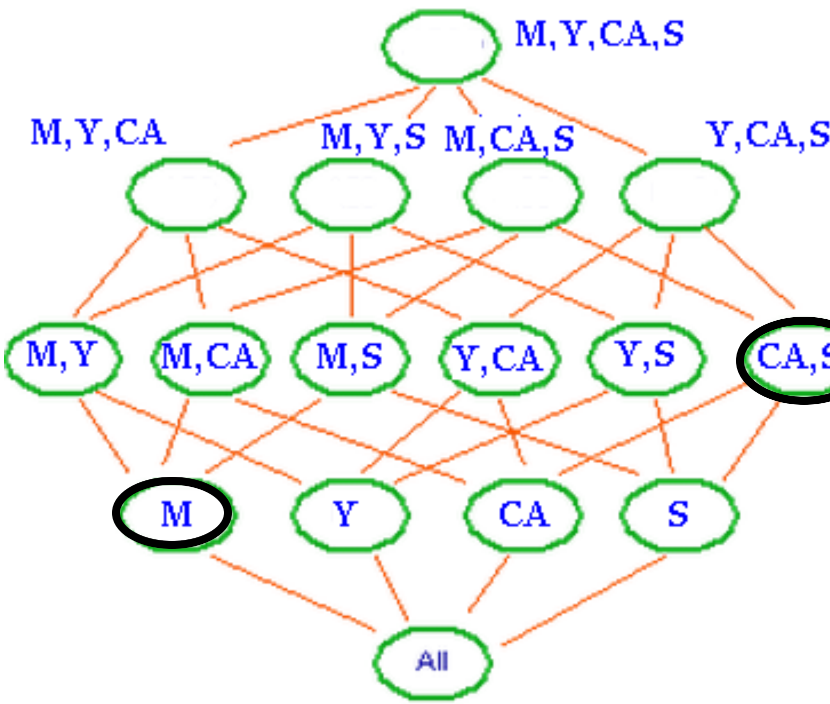
Say, C satisfies Coverage Constraint

Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage	√
Minimize Error	



C= {Male}
{California, Student}

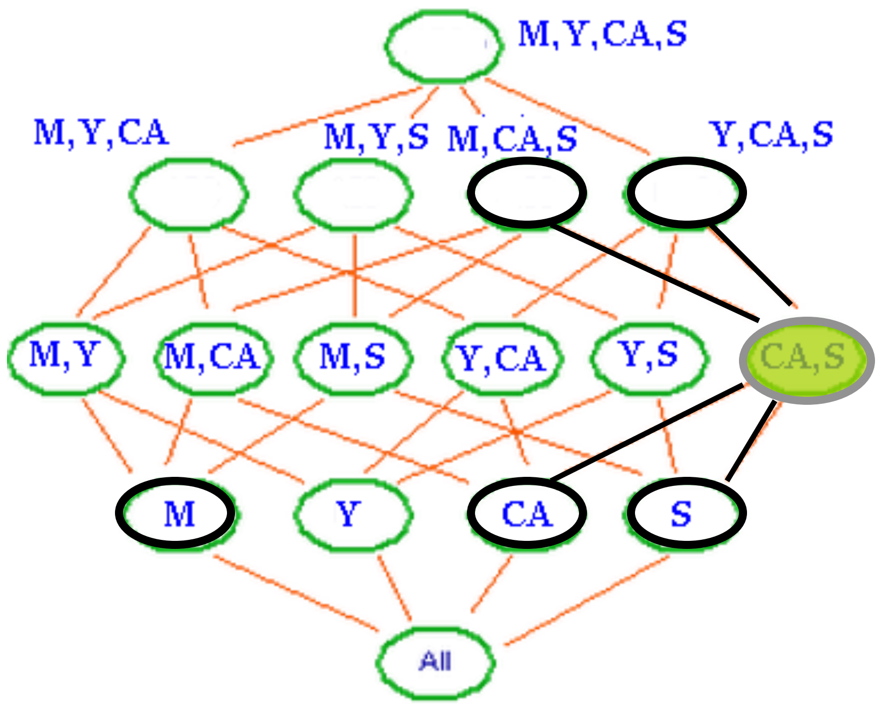
Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error ☐



C= {Male}

{California, Student}

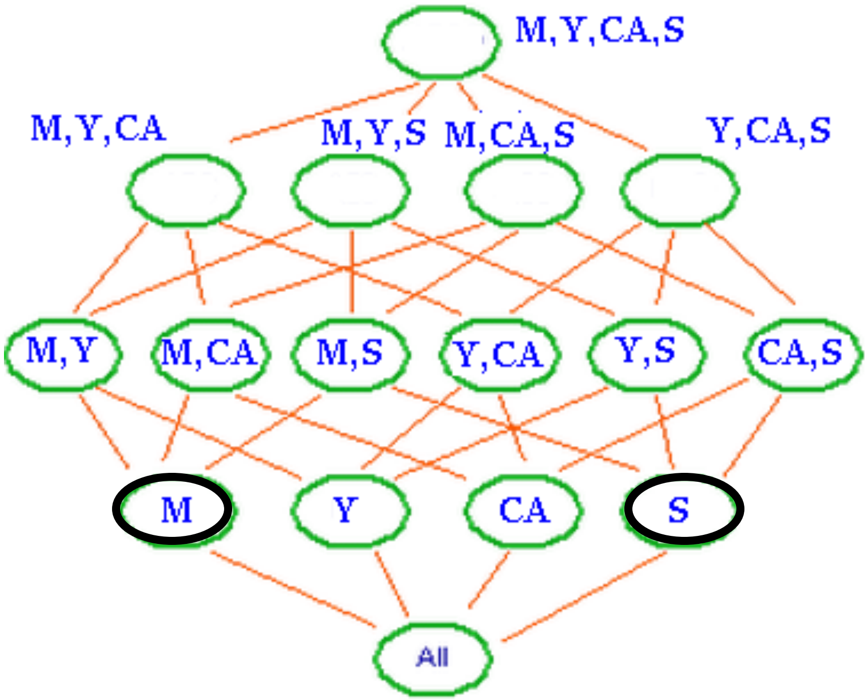
Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error √



C= {Male}
{Student}

Figure: Partial Rating Lattice for a Movie; k=2, α=80%

(M:Male, Y:Young, CA:California, S:Student)

What makes SDM different from DM?

- SDM needs a different data management stack: *data preparation*
- In social computing, analysts do not always know what to look for
- In social computing, application output must be evaluated

City	#POIs	#Timed Paths	Sample POIs
Barcelona	74	6,087	Museu Picasso, Plaza Reial
London	163	19,052	Buckingham Palace, Churchill Museum, Tower Bridge
New York City	100	3,991	Brooklyn Bridge, Ellis Island
Paris	114	10,651	Tour Eiffel, Musee du Louvre
San Francisco	80	12,308	Aquarium of the Bay, Golden Gate Bridge, Lombard Street

City	Ground Truth Sources
Barcelona	www.barcelona-tourist-guide.com
London	www.theoriginaltour.com
New York City	www.newyorksightseeing.com
Paris	www.carsrouges.com
San Francisco	www.allsanfranciscotours.com

Comparative evaluation

Evaluation Questions:

I. Overall, which one of the above two proposed itineraries you would rate higher?
- Itinerary 1 is significantly more useful than Itinerary 2.
- Itinerary 1 is somewhat more useful than Itinerary 2.
- Both are similar.
- Itinerary 2 is somewhat more useful than Itinerary 1.
- Itinerary 2 is significantly more useful than Itinerary 1.

Global comparison

II. How would you rate the set of points of interest included in the two itineraries?
- Itinerary 1 has significantly more appropriate points of interest than Itinerary 2.
- Itinerary 1 has somewhat more appropriate points of interest than Itinerary 2.
- Both are comparatively similar.
- Itinerary 2 has somewhat more appropriate points of interest than Itinerary 1.
- Itinerary 2 has significantly more appropriate points of interest than Itinerary 1.

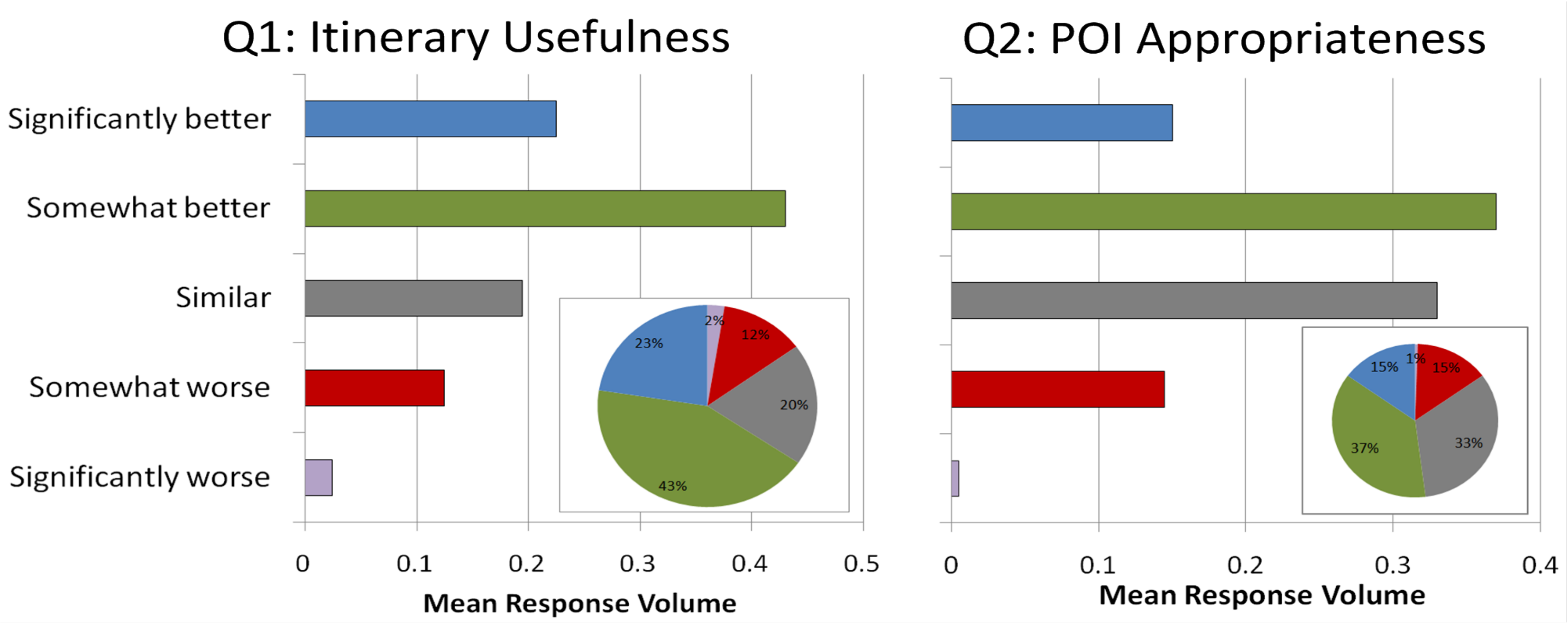
POI quality

III. How would you rate the transit times at the points of interest in the two itineraries (from a tourist perspective)?
- Itinerary 1 has significantly more accurate transit times than Itinerary 2.
- Itinerary 1 has somewhat more accurate transit times than Itinerary 2.
- Both are comparatively similar.
- Itinerary 2 has somewhat more accurate transit times than Itinerary 1.
- Itinerary 2 has significantly more accurate transit times than Itinerary 1.

Transit times

IV. Any additional comments?

Results for side-by-side comparison



Challenge 1:
Filtering expert AMT workers

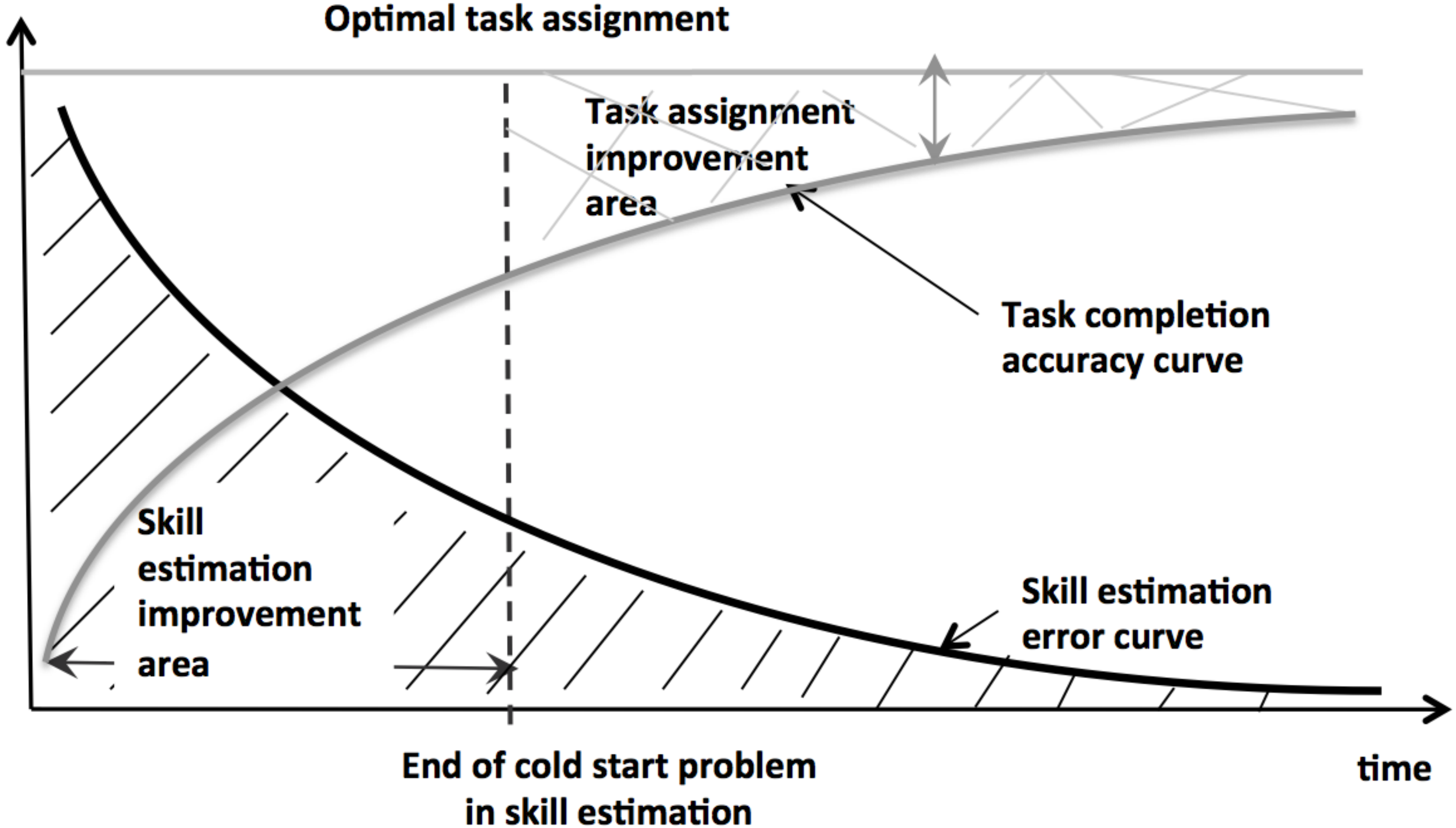
Multi-answer questions on “less-known” POIs

QUALIFICATION EVALUATION

Please choose the most suitable name of the point of interest based on your experience. This would judge your fitness to take the travel itinerary evaluation task in the next section.

○ Empire State Building	○ Flatiron Building	○ Herald Square
○ Rockefeller Center	○ Saint Patrick's Cathedral	○ Washington Sq Park
○ Chrysler Building	○ Trinity Church	○ Lincoln Center

Challenge 2:
How to better exploit the crowd?



Crowds, not drones: modeling human factors in crowdsourcing
with S. B. Roy (U. of Washington), G. Das, S. Thirumuruganathan (UT Arlington), I. Lykourentzou (Tudor Institute and INRIA) *at DBCrowd 2013*

Summary

- **There are three kinds of users in SDM**
  - *End user* who generates content of varying quality and demands high quality content
  - *Analyst (data scientist and application developer)* who needs a better understanding of the underlying data and users
  - *Worker* who helps relate to end user and evaluate content utility
- **Data preparation tools and efficient social exploration would help analysts**
  - *new opportunities for algebraic optimizations*
  - *a collection of optimization problems with data-centric or analyst-centric goals*
  - *often a reduction of hard problems with heuristics/approximation algorithms*
  - *but also appropriate indexing*
- **Application validation could benefit from worker profiling and crowd indexing**