INT 303 BIG DATA ANALYTICS

Lecture11: Recommendation System & Summary

Jia WANG

Jia.wang02@xjtlu.edu.cn

Xi'an Jiaotong-Liverpool University

西交利物浦大學

RECOMMENDATION

Search

Recommendations

Items

Products, web sites, blogs, news items, …

EXAMPLES:

amazon.com

PANDORA

StumbleUpon

del.icio.us

NETFLIX

movielens
helping you find the *right* movies

last·fm
the social music revolution

Google News

YouTube

XBOX LIVE

RECOMMENDATION MODEL

- X = SET OF **CUSTOMERS**
- S = SET OF **ITEMS**

- **Utility function** $u: X \times S \rightarrow R$
  - R = set of ratings
  - R is a totally ordered set
  - e.g., **0-5** stars, real number in **[0,1]**

UTILITY MATRIX

	AVATAR	LOTR	Matrix	Pirates
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

TYPES OF RECOMMENDATION

- **KEY PROBLEM:** UTILITY MATRIX ***U*** IS **SPARSE**
  - Most people have not rated most items
  - **Cold start:**
    - New items have no ratings
    - New users have no history
- **Three approaches to recommender systems:**
  - **1)** Content-based
  - **2)** Collaborative

} **Today!**

CONTENT-BASED RECOMMENDATION

Matrix Factorization

CONTENT-BASED RECOMMENDATION

- **MAIN IDEA:** RECOMMEND ITEMS TO CUSTOMER ***X*** SIMILAR TO PREVIOUS ITEMS RATED HIGHLY BY ***X***

Example:

- **Movie recommendations**
  - Recommend movies with same actor(s), director, genre, …
- **Websites, blogs, news**
  - Recommend other sites with “similar” content

MATRIX FACTORIZATION

MATRIX FACTORIZATION

- User vectors:

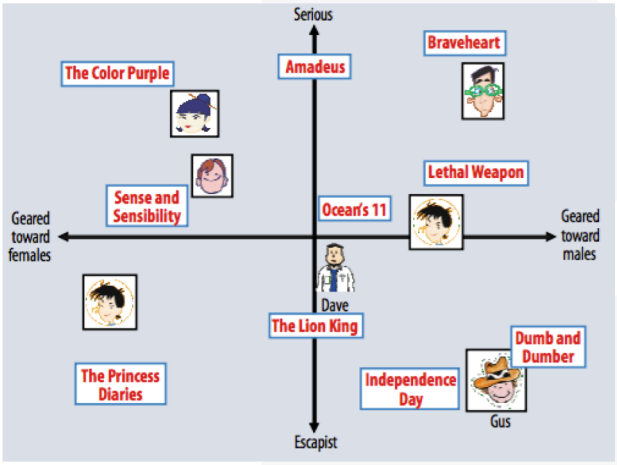
$$(W_{u*})^T \in \mathbb{R}^r$$

- Item vectors:

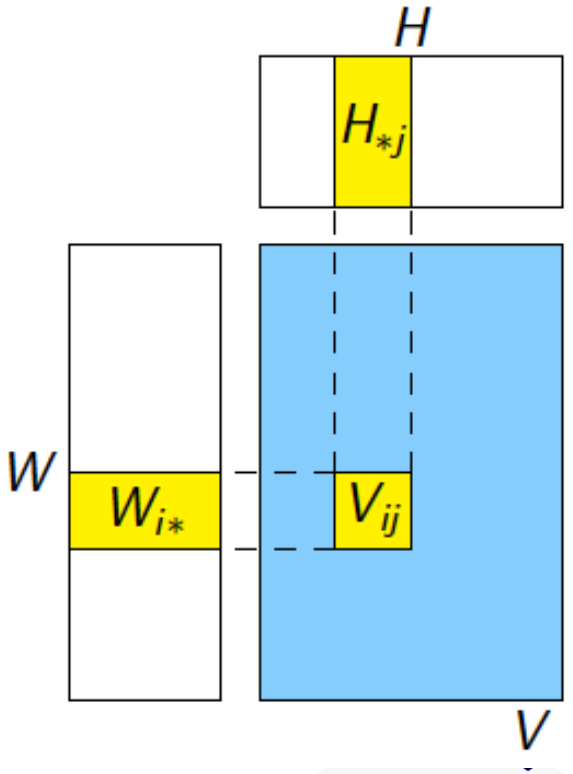
$$H_{*i} \in \mathbb{R}^r$$

- Rating prediction:

$$V_{ui} = W_{u*} H_{*i}$$
$$= [WH]_{ui}$$



Figures from Koren et al. (2009)



Figures from Gemulla et al. (2011)

MATRIX FACTORIZATION

- User vectors:

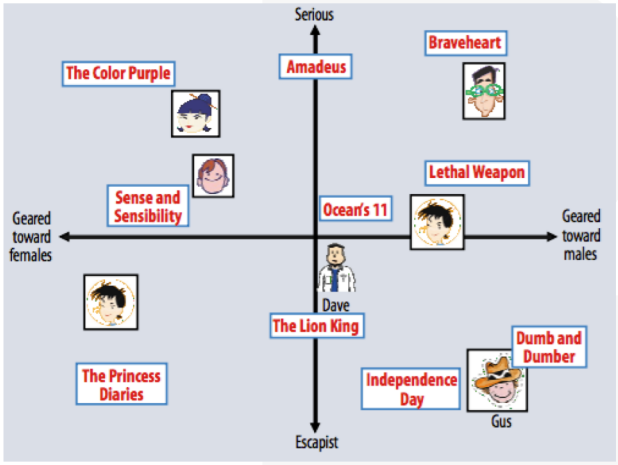
$$\mathbf{w}_u \in \mathbb{R}^r$$

- Item vectors:

$$\mathbf{h}_i \in \mathbb{R}^r$$

- Rating prediction:

$$v_{ui} = \mathbf{w}_u^T \mathbf{h}_i$$



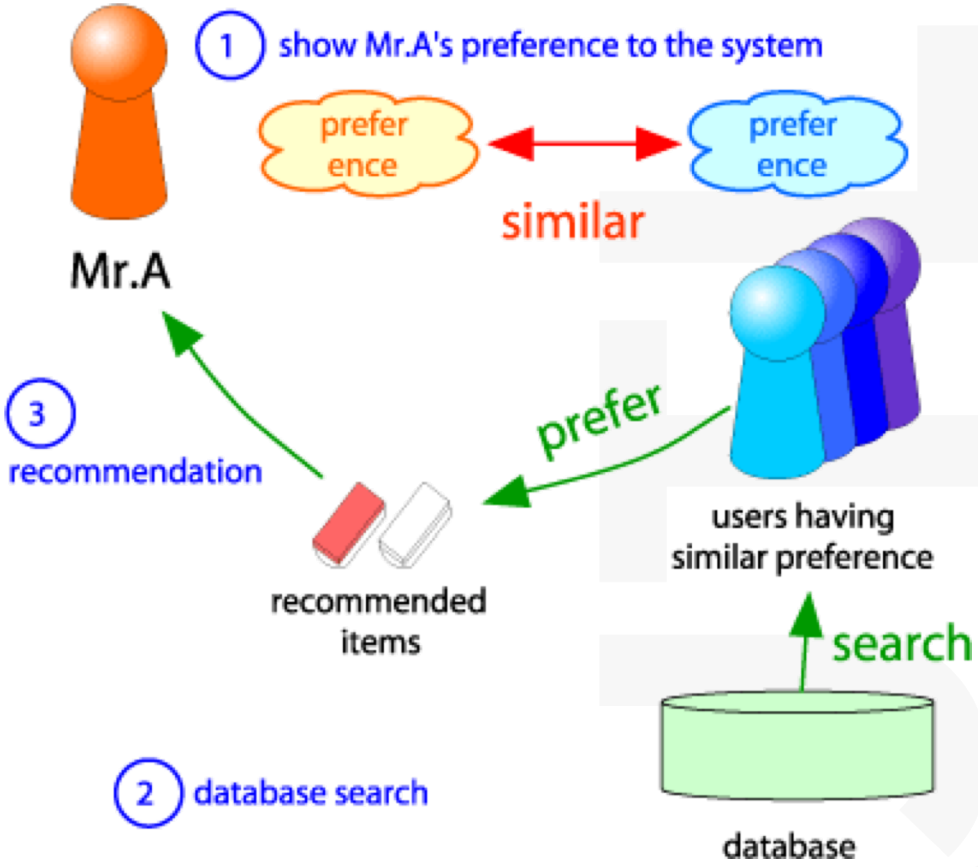
Figures from Koren et al. (2009)

COLLABORATIVE FILTERING

User-user Filtering
Item-Item Filtering

COLLABORATIVE FILTERING

- CONSIDER USER ***X***
- Find set ***N*** of other users whose ratings are "**similar**" to ***x***'s ratings
- Estimate ***x***'s ratings based on ratings of users in ***N***



FIND "SIMILAR" USERS

- LET r_x BE THE VECTOR OF USER x'S RATINGS
- **JACCARD SIMILARITY MEASURE**
  - Jaccard distance = $1 - \frac{|v_1 \cap v_2|}{|v_1 \cup v_2|}$
  - **Problem:** Ignores the value of the rating
- **Cosine similarity measure**
  - sim(x, y) = arccos(r_x , r_y) = $\frac{r_x \cdot r_y}{||r_x|| \cdot ||r_y||}$
  - **Problem:** Treats missing ratings as "negative"

SIMILARITY METRIC

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- **Intuitively we want: sim(*A*, *B*) > sim(*A*, *C*)**
- **Jaccard similarity:** 1/5 < 2/4
- **Cosine similarity:** 0.386 > 0.322
  - Considers missing ratings as “negative”

QUESTION

- Calculate the following distance measures between the two users with the different ratings, r1 = [0, 1, 1, 0, 0, 0,. 1], and r2 = [1, 0, 1, 0, 1, 0, 0].

- (a) What is the Jaccard distance between two users?
  (b) What is the Cosine distance between two user? (You can use arccos(x) to present the answer).

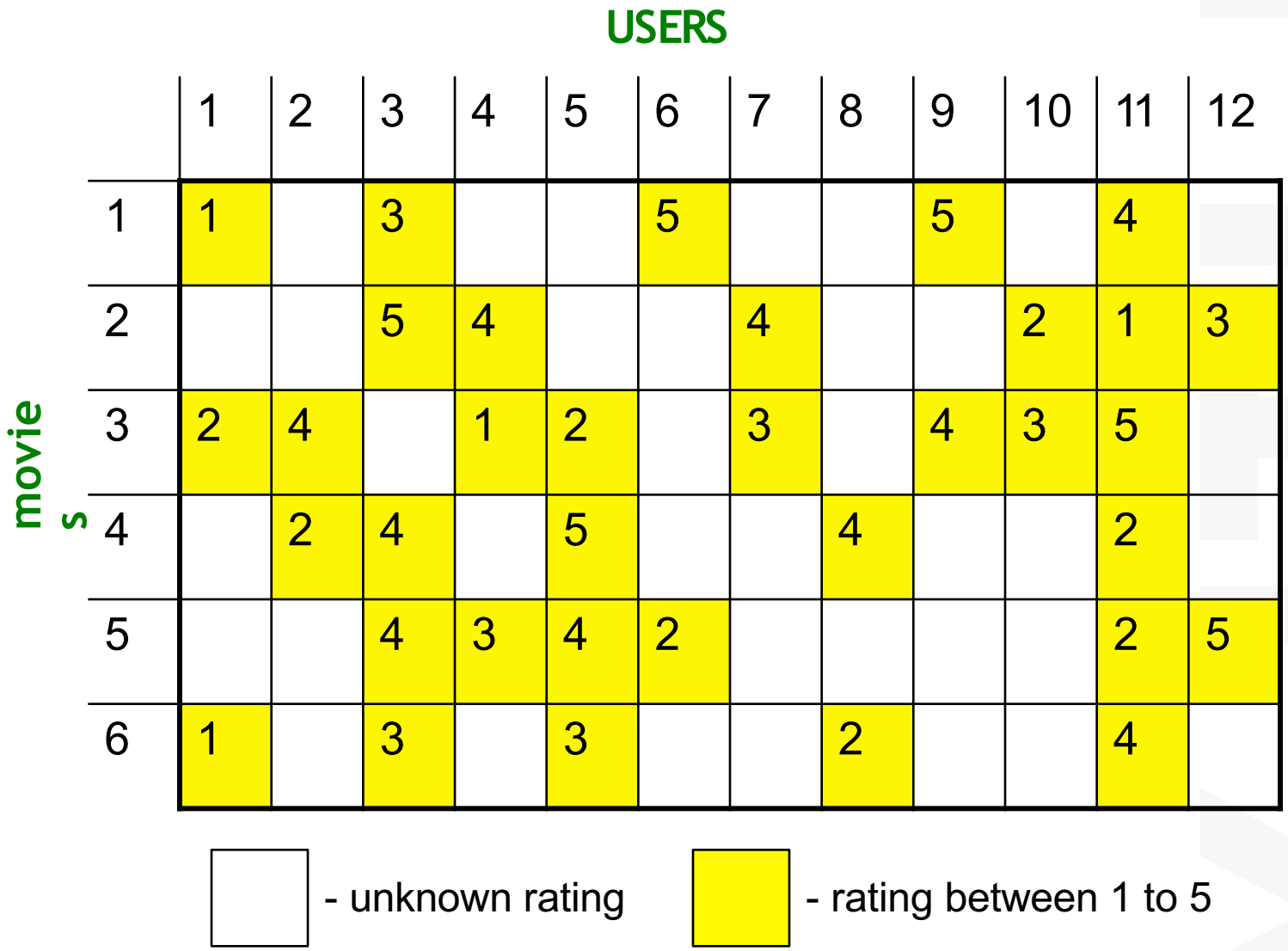
ITEM-ITEM COLLABORATIVE FILTERING

- **SO FAR: USER-USER COLLABORATIVE FILTERING**

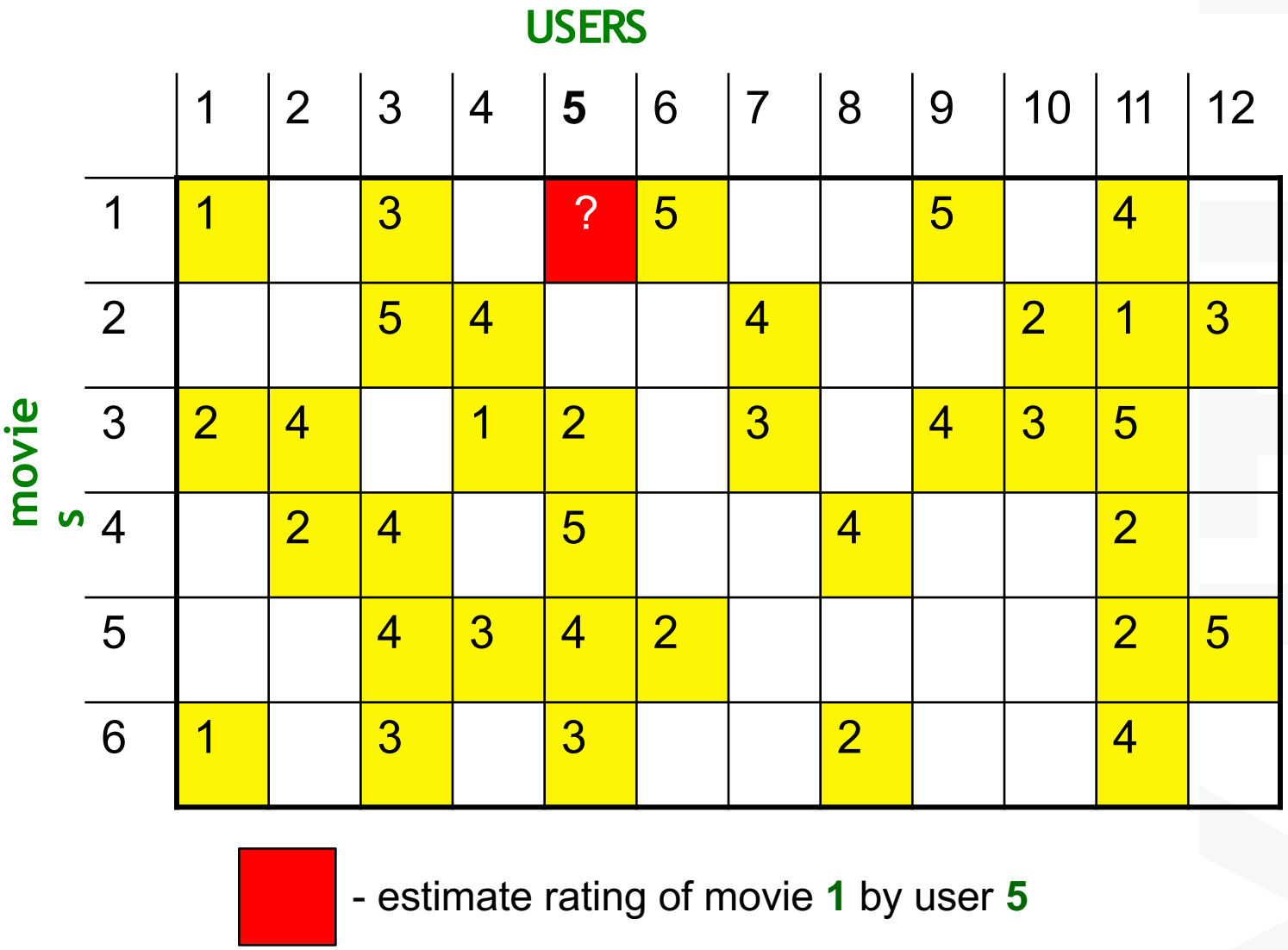
- **Another view: Item-item**
  - For item ***i***, find other similar items
  - Estimate rating for item ***i*** based on ratings for similar items
  - Can use same similarity metrics and prediction functions as in user-user model

$$r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$$

ITEM-ITEM CF (|N|=2)



ITEM-ITEM CF (|N|=2)



ITEM-ITEM CF (|N|=2)

movies	1	2	3	4	**5**	6	7	8	9	10	11	12	sim(1,m)
1	1		3		?	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
3	2	4		1	2		3		4	3	5		0.41
4		2	4		5			4			2		-0.10
5			4	3	4	2					2	5	-0.31
6	1		3		3			2			4		0.59

Neighbor selection:
Identify movies similar to movie **1**, **rated by user 5**

1. Compute cosine similarities between rows

ITEM-ITEM CF (|N|=2)

USERS

movies

	1	2	3	4	5	6	7	8	9	10	11	12	sim(1,m)
1	1		3		?	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
3	2	4		1	2		3		4	3	5		**0.41**
4		2	4		5			4			2		-0.10
5			4	3	4	2					2	5	-0.31
6	1		3		3			2			4		**0.59**

Compute similarity weights:

$s_{1,3}=0.41$, $s_{1,6}=0.59$

ITEM-ITEM CF (|N|=2)

USERS

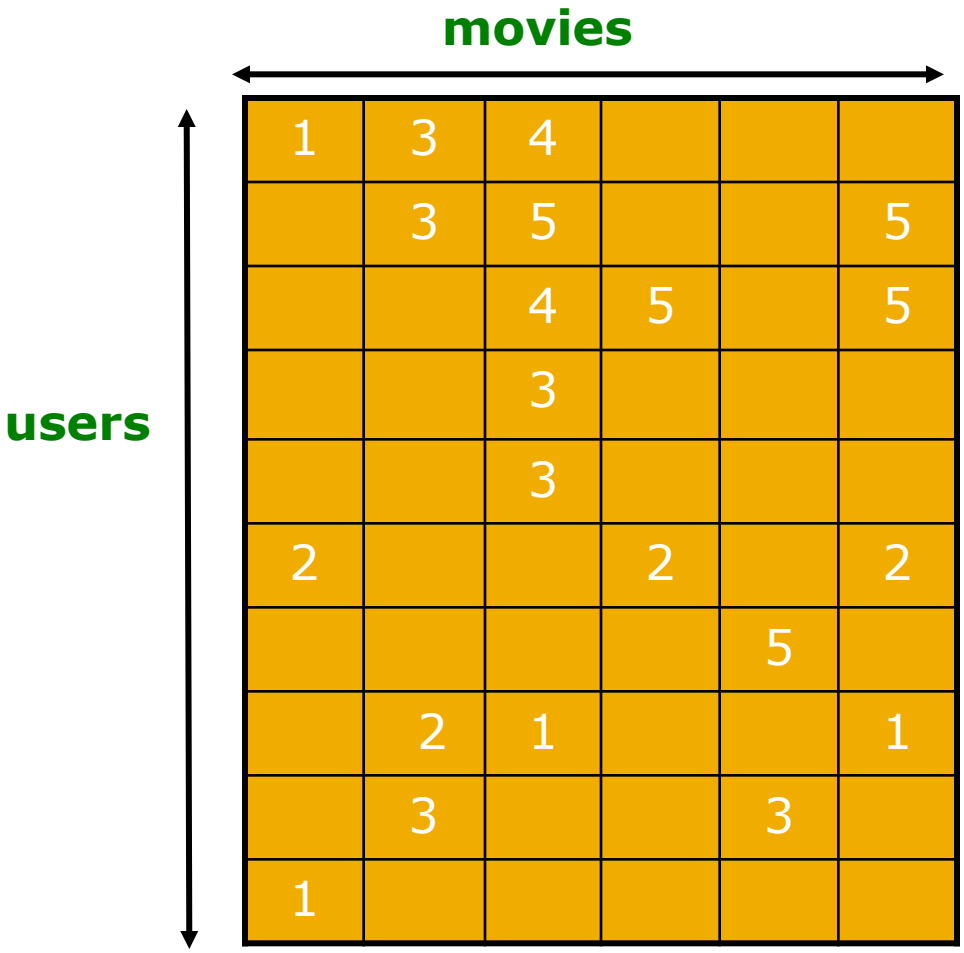
movies

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		**2.6**	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

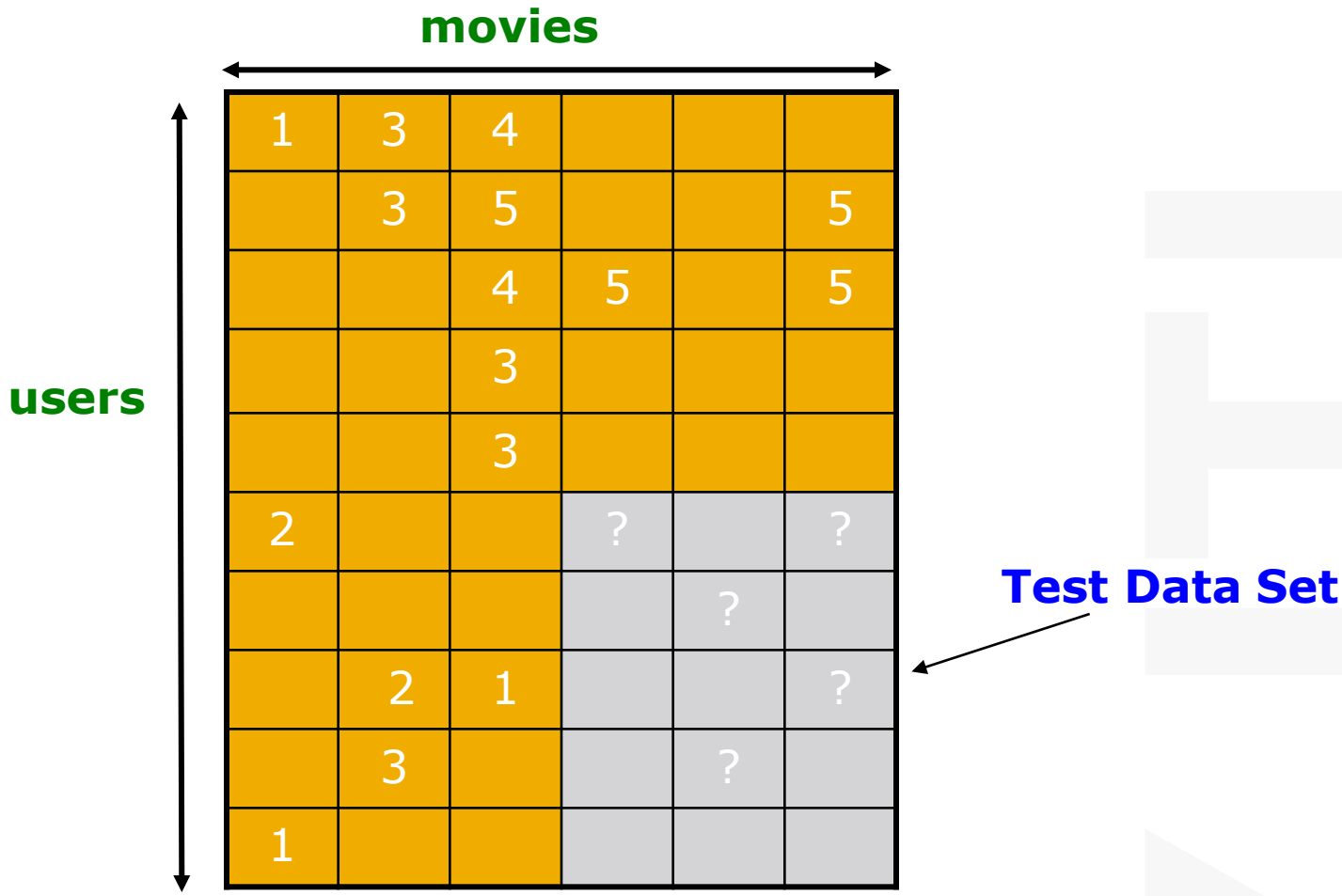
Predict by taking weighted average:

$r_{1.5} = (0.41*2 + 0.59*3) / (0.41+0.59) = 2.6$

EVALUATION



EVALUATION



QUESTION

- Consider a dataset containing information about movies: genre, director and release decade. We also have information about which users have seen each movie. The rating for a user on a movie is either 0 or 1.

Movie	Release decade	Genre	Director	Total number of ratings
A	1970s	Humor	D_1	40
B	2010s	Humor	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Humor	D_3	1

QUESTION

- Consider user U1=[2000s, D2, Humor]. We have some existing recommender system R that recommended the movie B to user U1.

(a) Given the above dataset, which one(s) do you think R could be?
  - User-user collaborative filtering.
  - Item-item collaborative filtering
  - Content-based recommender system.

(b) If some user U2 wants to watch a movie, under what conditions can our recommender system R recommend U2 a movie?

QUESTION

(c) If R recommends a movie, how to do it? If R cannot recommend a movie, please explain why it cannot be recommended.

(d) State any additional information R might want from U2 for predicting a movie for this user, if required.

REVIEW

WHAT?

The material of the course will integrate the five key facets of an investigation using data:

1. data collection; data wrangling, cleaning, and sampling to get a suitable data set
2. data management; accessing data quickly and reliably
3. exploratory data analysis; generating hypotheses and building intuition
4. prediction or statistical learning
5. communication; summarizing results through visualization, stories, and interpretable summaries.

WHAT WE HAVE LEARNED

Module 1:

<u>The basics.</u>

- What is big data analytics?
- What is data and big data?
- Big data Gramma

WHAT WE HAVE LEARNED

Module 2:

Big data collection and visualization.

- Data collection and Data scraping
- Data Virtualization

WHAT WE HAVE LEARNED

Module 3:

Systems and software . It introduces the popular platforms available for big data processing.

- Large-scale computing
- Distributed file system
- MapReduce: Distributed computing programming model
- Spark: Extends MapReduce

WHAT WE HAVE LEARNED

Module 4:

The data processing methods and algorithms.

- How to tell a good story
- Representing Data and Engineering Features
- Dimensionality Reduction
- Big Data Analysis Models
- Bagging
- Boosting

WHAT WE HAVE LEARNED

Module 5:

<u>The big data application.</u>

- Recommendation System

GRADES

Sequence	Method	Learning outcomes assessed	Duration	Timing	% of final mark	Resit
#1	Project 1: Data Scraping	All	See notice	S2	15%	NO
#2	Project 2: Big Data Competition	All	See notice	S2	15%	NO
#3	Written Exam	All	See notice	S2	70%	NO

GRADES

REVIEW TIPS

1. Review the slides.
2. The investigated knowledge points are all given on the LM.
3. The exam questions are similar to the questions in the slides.

THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大学