

# Data Mining

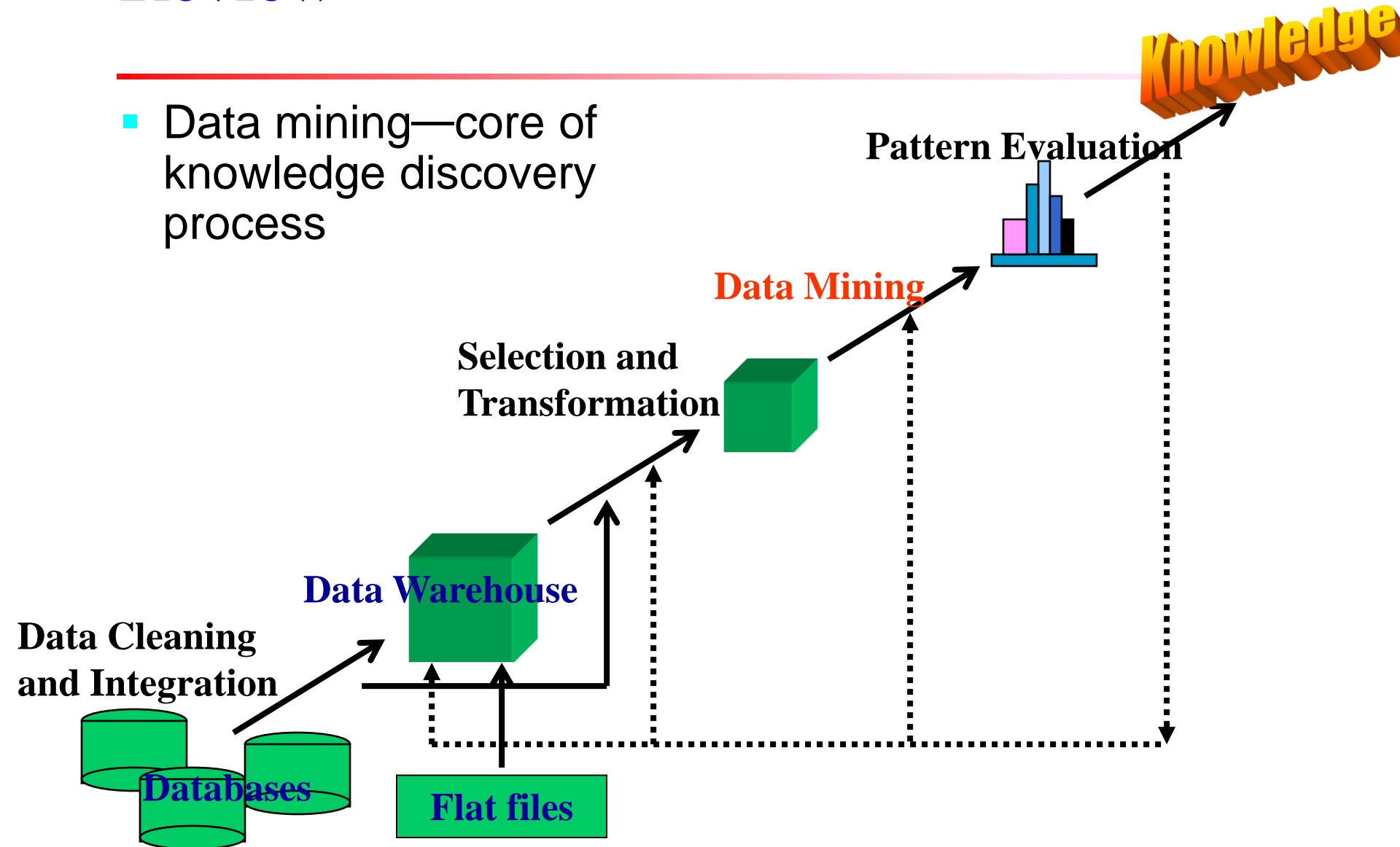
---

**Ying Liu, Prof., Ph.D**

*School of Computer Science and Technology  
University of Chinese Academy of Sciences  
Data Mining and High Performance Computing Lab*

# Review

- Data mining—core of knowledge discovery process



# Outline

---

- What is Recommender System?
- Recommendation Algorithms
- Evaluation of Recommender Systems

# Motivation

---

- Which digital camera should I buy?
- Where should I spend my holiday?
- Which movie should I see?
- Whom should I follow?
- Where should I find interesting news article?

# Motivation

---

- There are many choices
  - There are no obvious advantages among them
  - We do not have enough resources to check all options (*information overload*)
  - We do not have enough knowledge and experience to choose
- Solution
- ***Recommendation: automatically come up with a short list of items that fits user's interests!***

# Examples

## Book recommendation in Amazon

The screenshot shows the Amazon product page for 'Networks: An Introduction' by Mark Newman. The page includes a 'Frequently Bought Together' section, a 'Customers Who Bought This Item Also Bought' section (highlighted with a red box), and an 'Editorial Reviews' section. The 'Also Bought' section displays several related books with their prices and ratings.

Book Title	Author	Price	Rating
Networks, Crowds, and Markets: Reasoning About a Highly Connected World	David Easley	\$69.40	4.5 stars (13)
Dynamic Processes on Complex Networks	Alain Barrat	\$64.18	4.5 stars (3)
Simply Complexity: A Clear Guide to Complexity Theory	Neil Johnson	\$9.81	4.5 stars (6)
Social Network Analysis: Methods and Applications	Stanley Wasserman	\$44.52	4.5 stars (10)
Networks of the Brain	Olaf Sporns	\$32.28	4.5 stars (4)

## Video clip recommendation in YouTube

The screenshot shows a YouTube video player for 'Ariz. Wildfire Near Flagstaff at 10,000 Acres'. The video has 510 views. A red box highlights the 'Suggestions' sidebar on the right, which lists related videos about wildfires in Arizona.

Video Title	Channel	Views
Schutz Fire - Flagstaff, AZ - June 20, 2010	Ball2Venus	7,211 views
Flagstaff Father's Day Fire #2 - Schutz Wildfire	GoalBusters	8,327 views
Winds Driving Fire in Ariz., Homes Threatened	AssociatedPress	1,491 views
Arizona wildfires rage on	NewsOnABC	141 views
Arizona wildfires third largest in state history	CBSNewsOnline	815 views
Arizona Governor Tours Growing Wildfire Near NM	AssociatedPress	9,118 views
Arizona wildfire barely contained	WMAZabc2news	55 views

## Product Recommendation in ebay

The screenshot shows an eBay product page for 'Dr. Seuss's Second Beginner Book Collection'. The page includes a 'Recommendations for you' section and a 'Popular on eBay' section. The 'Recommendations' section displays several related books with their prices and ratings.

Book Title	Price	Rating
Dr. Seuss's Second Beginner Book Collection	\$21.90	4.5 stars (13)
Dr. Seuss's Beginner Book Classic Four	\$1.99	4.5 stars (3)
Ten Apples Up on Top!	\$1.99	4.5 stars (6)
Dr. Seuss's ABC by Dr. Seuss 1951	\$3.50	4.5 stars (10)
The Big Green Book of Beginner Books	\$11.35	4.5 stars (4)
The Cat in the Hat by Dr. Seuss	\$0.99	4.5 stars (13)

## Restaurant Recommendation in Yelp

The screenshot shows a Yelp search results page for 'Tempé, AZ'. The page displays a list of restaurants with their ratings and reviews. A map on the right shows the location of the restaurants in the area.

Restaurant Name	Category	Rating	Reviews
1. The Dhaba	Curry, Indian, Pakistani	4.5 stars	10 reviews
2. China Farm Chinese Buffet	Chinese, Buffet, Food Delivery Services	4.5 stars	10 reviews
3. Capriotti's Sandwich Shop	Sandwiches	4.5 stars	12 reviews

# Recommender Systems

---

- Idea: Use historical data such as the user's past preferences or similar users' past preferences to predict future likes
- Basic assumption
  - Users' preferences are likely to remain stable, and change smoothly over time
  - Users with similar tastes have similar ratings for an item
- By watching the past users' or groups' preferences, try to predict their future likes
  - Then we can recommend items of interest to them

# Recommender Systems

---

- Formally, a recommender system takes a set of users  $U$  and a set of items  $I$  and *learns a function  $f$*  such that:

$$f : U \times I \rightarrow \mathbb{R}$$



# Recommendation vs. Search

---

- One way to get answers is using search engines
- Search engines find results that match the query provided by the user
- The results are generally provided as a list ordered with respect to the relevance of the item to the given query
- Consider the query “best 2014 movie to watch”
  - The same results for an 8 year old and an adult

*Search engines' results are not customized!*

# Outline

---

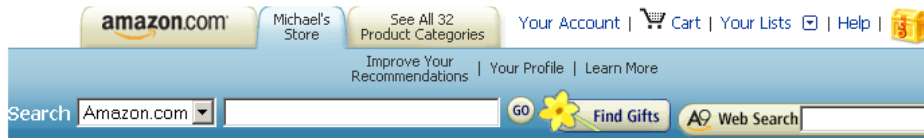
- What is Recommender System?
- Recommendation Algorithms
- Evaluation of Recommender Systems

# Content-based Methods

---

- Content-based methods are based on the fact that **a user's interest should match the description of the items** that she should be recommended
- The more similar the item's description to that of the user's interest, the more likely the user finds the item's recommendation interesting
- **Core idea:** Find the similarity between the user and all of the existing items

# Example



## Edit Favorites

Mark the categories that interest you the most.

☒ Books

Submit

### Your Books Favorites

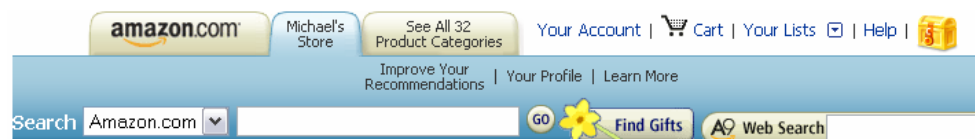
#### Categories

- ☒ Biographies & Memoirs
- ☒ Business & Investing
- ☒ Computers & Internet

☒ Nonfiction

#### Add to Your Favorites

- |  |   |
|--|---|
| <input type="checkbox"/> Arts & Photography      | <input type="checkbox"/> Outdoors & Nature        |
| <input type="checkbox"/> Children's Books        | <input type="checkbox"/> Parenting & Families     |
| <input type="checkbox"/> Comics & Graphic Novels | <input type="checkbox"/> Professional & Technical |
| <input type="checkbox"/> Cooking, Food & Wine    | <input type="checkbox"/> Reference                |
| <input type="checkbox"/> Entertainment           | <input type="checkbox"/> Religion & Spirituality  |



## Recommended For You > Books

### Recommendations by Category in Books

Your Favorites [Edit](#)

[Business & Investing](#)  
[Computers & Internet](#)  
[Biographies & Memoirs](#)  
[Nonfiction](#)

More Categories

[Arts & Photography](#)  
[Children's Books](#)  
[Comics & Graphic Novels](#)  
[Cooking, Food & Wine](#)  
[Entertainment](#)  
[Gay & Lesbian](#)  
[Health, Mind & Body](#)  
[History](#)  
[Home & Garden](#)

These recommendations are based on [items you own](#) and more.

view: **All** | [New Releases](#) | [Coming Soon](#)

[More results](#)

- 

**The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture**  
by John Battelle  
Average Customer Review: ★★★★★  
Publication Date: September 8, 2005

**Our Price: \$16.35**  
**Used & new** from \$10.95

[Add to cart](#)

[Add to Wish List](#)

☐ I Own It ☐ Not interested x|★★★★★ Rate it  
Recommended because you purchased [Amazonia](#) and more ([edit](#))

- 

**Writing Successful Science Proposals**  
by Andrew J. Friedland, Carol L. Folt  
Average Customer Review: ★★★★★  
Publication Date: June 10, 2000

[Add to cart](#)

# Content-based Methods

---

## ■ Steps

1. Describe the items to be recommended
2. Create a profile of the user that describes the types of items the user likes
3. Compare items with the user profile to determine what to recommend

# Content-based Algorithm

---

- 1. Represent both user profiles and item descriptions by vectorizing them using a set of  $k$  keywords
- 2. Vectorize (e.g., using TF-IDF) both users and items and compute their similarity

$$I_j = (i_{j,1}, i_{j,2}, \dots, i_{j,k})$$

$$U_i = (u_{i,1}, u_{i,2}, \dots, u_{i,k}).$$

$$\text{sim}(U_i, I_j) = \cos(U_i, I_j) = \frac{\sum_{l=1}^k u_{i,l} i_{j,l}}{\sqrt{\sum_{l=1}^k u_{i,l}^2} \sqrt{\sum_{l=1}^k i_{j,l}^2}}$$

- 3. Recommend the top most similar items to the user

# Collaborative Filtering

## ■ Assumption

### ■ User-based CF

- Users with similar previous ratings for items are likely to rate future items similarly

	I1	I2	I3	I4
U1	1	2	4	4
U2	1	2	4	?
U3	2	5	2	2
U4	5	2	3	3

### ■ Item-based CF

- Items that have received similar ratings previously from users are likely to receive similar ratings from future users (item-based CF)

	I1	I2	I3	I4
U1	1	2	4	4
U2	1	2	4	?
U3	2	5	2	2
U4	5	2	3	3

# Example

## Movies You've Rated

Based on your 745 movie ratings, this is the list of movies you've seen. As you discover movies on the website that you've seen, rate them and they will show up on this list. On this page, you may change the rating for any movie you've seen, and you may remove a movie from this list by clicking the 'Clear Rating' button.

Sort by > **Star Rating**

Jump to > **5 Stars**

	TITLE	MPAA	GENRE	STAR RATING
Add	<a href="#">12 Angry Men</a> (1957)	UR	Classics	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">The 39 Steps</a> (1935)	UR	Classics	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">An American in Paris</a> (1951)	UR	Classics	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">The Andromeda Strain</a> (1971)	G	Sci-Fi & Fantasy	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">Apollo 13</a> (1995)	PG	Drama	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">The Battle of Algiers</a> (1965) La Battaglia di Algeri	UR	Foreign	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">Being There</a> (1979)	PG	Drama	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">Big Deal on Madonna Street</a> (1958) I soliti ignoti	UR	Foreign	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">The Birds</a> (1963)	PG-13	Thrillers	5 stars <a href="#">Clear Rating</a>
Add	<a href="#">Blade Runner</a> (1982)	R	Sci-Fi & Fantasy	5 stars <a href="#">Clear Rating</a>

Value	Graphic representation	Textual representation
5	☆☆☆☆☆	Excellent
4	☆☆☆☆	Very good
3	☆☆☆	Good
2	☆☆	Fair
1	☆	Poor

Table 9.1: User-Item Matrix

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1



# Collaborative Filtering

---

## ■ Rating matrix

- **Explicit ratings:** entered by a user directly
  - i.e., “Please rate this on a scale of 1-5”



Rating: 5.2/10 (5 votes cast)



Rating: 5.2/10 (5 votes cast)



Rating: 8.8/10 (5 votes cast)

- **Implicit ratings:** inferred from other user behavior
  - Play lists or music listened to, for a music Rec system
  - The amount of time users spent on a webpage

# Collaborative Filtering Algorithm

---

## ■ Steps

1. Weigh all users/items with respect to their similarity with the current user/item
2. Select a subset of the users/items (neighbors) as recommenders
3. Predict the rating of the user for specific items using neighbors' ratings for the same (or similar) items
4. Recommend items with the highest predicted rank

# Collaborative Filtering Algorithm

---

- Measure Similarity between Users (or Items)

$$\text{sim}(U_i, U_j) = \cos(U_i, U_j) = \frac{U_i \cdot U_j}{\|U_i\| \|U_j\|} = \frac{\sum_k r_{ik} r_{jk}}{\sqrt{\sum_k r_{ik}^2} \sqrt{\sum_k r_{jk}^2}}$$

- Pearson Correlation Coefficient

$$\text{sim}(U_i, U_j) = \frac{\sum_k (r_{ik} - \bar{r}_i)(r_{jk} - \bar{r}_j)}{\sqrt{\sum_k (r_{ik} - \bar{r}_i)^2} \sqrt{\sum_k (r_{jk} - \bar{r}_j)^2}}$$

# Collaborative Filtering Algorithm

---

Updating the ratings:

The diagram shows the formula for updating a user's rating for an item,  $r_{u,i}$ . The formula is: 
$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} \text{sim}(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} \text{sim}(u, v)},$$
 Annotations with arrows point to various parts of the formula: 

- An arrow points from "User u's mean rating" to  $\bar{r}_u$ .
- An arrow points from "User v's mean rating" to  $\bar{r}_v$ .
- An arrow points from "Observed rating of user v for item i" to  $r_{v,i}$ .
- An arrow points from "Predicted rating of user u for item i" to  $r_{u,i}$ .

# Example

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Predict Jane's rating  
for Aladdin

## 1- Calculate average ratings

$$\bar{r}_{John} = \frac{3 + 3 + 0 + 3}{4} = 2.25$$

$$\bar{r}_{Joe} = \frac{5 + 4 + 0 + 2}{4} = 2.75$$

$$\bar{r}_{Jill} = \frac{1 + 2 + 4 + 2}{4} = 2.25$$

$$\bar{r}_{Jane} = \frac{3 + 1 + 0}{3} = 1.33$$

$$\bar{r}_{Jorge} = \frac{2 + 2 + 0 + 1}{4} = 1.25$$

## 2- Calculate user-user similarity

$$sim(Jane, John) = \frac{3 \times 3 + 1 \times 3 + 0 \times 3}{\sqrt{10} \sqrt{27}} = 0.73$$

$$sim(Jane, Joe) = \frac{3 \times 5 + 1 \times 0 + 0 \times 2}{\sqrt{10} \sqrt{29}} = 0.88$$

$$sim(Jane, Jill) = \frac{3 \times 1 + 1 \times 4 + 0 \times 2}{\sqrt{10} \sqrt{21}} = 0.48$$

$$sim(Jane, Jorge) = \frac{3 \times 2 + 1 \times 0 + 0 \times 1}{\sqrt{10} \sqrt{5}} = 0.84$$

# Example

---

3- Calculate Jane's rating for Aladdin,  
Assume that neighborhood size = 2

$$\begin{aligned}r_{Jane, Aladdin} &= \bar{r}_{Jane} + \frac{sim(Jane, Joe)(r_{Joe, Aladdin} - \bar{r}_{Joe})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\&\quad + \frac{sim(Jane, Jorge)(r_{Jorge, Aladdin} - \bar{r}_{Jorge})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\&= 1.33 + \frac{0.88(4 - 2.75) + 0.84(2 - 1.25)}{0.88 + 0.84} = 2.33\end{aligned}$$

# Outline

---

- What is Recommender System?
- Recommendation Algorithms
- Evaluation of Recommender Systems

# Evaluation is Challenging

---

- Different algorithms may be better or worse on different datasets (applications)
  - Many algorithms are designed specifically for datasets
  - Differences exist for rating density, rating scale, and other properties of datasets
- The goals to perform evaluation may differ
  - Early evaluation work focused specifically on the "accuracy" in "predicting"
  - Other properties also have important effect on user satisfaction and performance



# Evaluation is Challenging

---

- It is challenge in deciding what combination of measures should be used in comparative evaluation

# Predictive Accuracy Metrics

---

- Mean Absolute Error (*MAE*)  
measures the average absolute deviation between a predicted rating ( $\hat{r}$ ) and the user's true rating ( $r$ )

$$MAE = \frac{\sum_{i,j} |\hat{r}_{ij} - r_{ij}|}{n}$$

- $NMAE = MAE / (r_{\max} - r_{\min})$

- Root Mean Square Error (*RMSE*) is similar to *MAE*, but places more emphasis on larger deviation

$$RMSE = \sqrt{\frac{1}{n} \sum_{i,j} (\hat{r}_{ij} - r_{ij})^2}$$

# Example

---

Consider the following table with both the predicted ratings and true ratings of five items

<i>Item</i>	<i>Predicted Rating</i>	<i>True Rating</i>
1	1	3
2	2	5
3	3	3
4	4	2
5	4	1

$$MAE = \frac{|1 - 3| + |2 - 5| + |3 - 3| + |4 - 2| + |4 - 1|}{5} = 2$$

$$NMAE = \frac{MAE}{5 - 1} = 0.5$$

$$\begin{aligned} RMSE &= \sqrt{\frac{(1 - 3)^2 + (2 - 5)^2 + (3 - 3)^2 + (4 - 2)^2 + (4 - 1)^2}{5}} \\ &= 2.28 \end{aligned}$$

# Relevance: Precision and Recall

---

- **Precision:** a measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved

$$P = \frac{N_{rs}}{N_s}$$

- **Recall:** a measure of completeness, determines the fraction of relevant items retrieved out of all relevant items

$$R = \frac{N_{rs}}{N_r}$$

	Selected	Not Selected	Total
Relevant	$N_{rs}$	$N_{rn}$	$N_r$
Irrelevant	$N_{is}$	$N_{in}$	$N_i$
Total	$N_s$	$N_n$	$N$

# Example

---

	<i>Selected</i>	<i>Not Selected</i>	<i>Total</i>
<i>Relevant</i>	9	15	24
<i>Irrelevant</i>	3	13	16
<i>Total</i>	12	28	40

$$P = \frac{9}{12} = 0.75$$

$$R = \frac{9}{24} = 0.375$$

$$F = \frac{2 \times 0.75 \times 0.375}{0.75 + 0.375} = 0.5$$

# Evaluating Ranking

---

## ■ Spearman's Rank Correlation

$$\rho = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n^3 - n}$$

## ■ Kendall's $\tau$

- It checks the concordance of the items of the recommended ranking list against the ground truth ranking list
- If the two orders are consistent, it is concordant
- For top 4 items in ranking list, there are  $4 \cdot 3 / 2 = 6$  pairs

$$\tau = \frac{c - d}{\binom{n}{2}}$$

where  $c$  is the number of concordants and  $d$  of discordants

# Example

Consider a set of four items  $I = \{i_1, i_2, i_3, i_4\}$  for which the predicted and true rankings are as follows

	<i>Predicted Rank</i>	<i>True Rank</i>
$i_1$	1	1
$i_2$	2	4
$i_3$	3	2
$i_4$	4	3

Pair of items and their status  
{concordant/discordant} are

$(i_1, i_2)$  : concordant

$(i_1, i_3)$  : concordant

$(i_1, i_4)$  : concordant

$(i_2, i_3)$  : discordant

$(i_2, i_4)$  : discordant

$(i_3, i_4)$  : concordant

$$\tau = \frac{4 - 2}{6} = 0.33$$