



UNIVERSITAT DE  
BARCELONA



## Master on Foundations of Data Science



# Recommender Systems

Non-Personalized Recommender Systems

Santi Seguí | 2016-2017

# What is a non-recommender system

A non-personalized recommender system is one that **makes the same recommendations for everyone.**

The simplest example is a retailer that shows the ten (or some number) most popular products on their homepage.

Examples: TripAdvisor, IMDB

# Where is it useful?

IMDb

Find Movies, TV shows, Celebrities and more...

All

IMDbPro Help

f Sign in with Facebook Other Sign in options

## IMDb Charts

### Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating	Action
1. Cadena perpètua (1994)	★ 9,2	☆	[+]
2. El padrí (1972)	★ 9,2	☆	[+]
3. El padrí II (1974)	★ 9,0	☆	[+]
4. El cavaller fosc (2008)	★ 8,9	☆	[+]
5. Dotze homes sense pietat (1957)	★ 8,9	☆	[+]
6. La llista de Schindler (1993)	★ 8,9	☆	[+]
7. Pulp Fiction (1994)	★ 8,9	☆	[+]
8. El senyor dels anells: El retorn del rei (2003)	★ 8,9	☆	[+]

SHARE

You Have Seen

0/250 (0%)

Hide titles I've seen

---

### IMDb Charts

Box Office

Most Popular Movies

Top Rated Movies

Top Rated English Movies

Most Popular TV

Top Rated TV

Top Rated Indian Movies

Lowest Rated Movies

---

### Top Rated Movies by Genre

Action

Adventure

Animation

Biography

Comedy

Crime

Drama

Family

Fantasy

Film-Noir

History

Horror

Music

Musical

Mystery

# Case 1: Top Recommendations



[See more choices](#)

## Lucky Bums Snow Sport Helmet with Fleece Liner

by Lucky Bums

\$32<sup>93</sup> - \$63<sup>87</sup>

Some sizes/colors are Prime eligible

More Buying Choices

\$32.93 new (4 offers)

\$27.99 used (4 offers)

FREE Shipping on eligible orders

Show only Lucky Bums items

118



[See Color & Size Options](#)

## Smith Aspect Helmet Mens

by Smith Optics

\$64<sup>95</sup> - \$120<sup>00</sup>

Some sizes/colors are Prime eligible

More Buying Choices

\$95.00 new (25 offers)

FREE Shipping on eligible orders

Show only Smith Optics items

78



[See more choices](#)

## Giro Sestriere Snow Helmet

by Giro

\$59<sup>95</sup> - \$212<sup>69</sup>

Some sizes/colors are Prime eligible

More Buying Choices

\$69.95 new (3 offers)

\$64.95 used (1 offer)

FREE Shipping on eligible orders

Show only Giro items

41



Giro Seam Snow Helmet - Men's M...  
\$159<sup>95</sup>   
 118



Smith Optics Unisex Adult Vantage...  
\$270<sup>00</sup>   
 51



[See more choices](#)

## SUNVP Ski Helmet Ultralight Integrally Warmest Windproof Snowboards Snow Sports Helmet Unisex Adult

by SUNVP

\$36<sup>99</sup> \$110.09

Some sizes/colors are Prime eligible

FREE Shipping on eligible orders

Show only SUNVP items

12



[See more choices](#)

## Smith Optics SO-H15AS Men's Aspect Snow Helmet, Matte Charcoal-Large

by Smith

\$99<sup>95</sup>

More Buying Choices

\$99.95 new (2 offers)

FREE Shipping on eligible orders

Show only Smith items

4



[See more choices](#)

## Giro Seam Snow Helmet

by Giro

\$69<sup>99</sup> - \$338<sup>22</sup>

Some sizes/colors are Prime eligible

More Buying Choices

\$159.94 new (20 offers)

Show only Giro items

118



Giro Sestriere Snow Helmet (Black...  
\$69<sup>95</sup> \$70.00   
 41

# Case 2: Product Association



## Lucky Bums Snow Sport Helmet

by [Lucky Bums](#)

788 customer reviews | 69 answered questions

Price: \$30.59 - \$65.88

Size:

Select ▾

Color: Kryptek Typhon, Glossy



- ALL-AROUND PROTECTION FOR ALL SNOW SPORTS NEEDS - The Lucky Bums Snow Sports Helmet is everything you want in a helmet at an affordable price: it's comfortable, stylish, durable, and most of all, functional. It features two protective layers, a padded chin strap, and goggle loop for extreme downhills. Ready to go right out of the box, find your size and color today.
- ABS AND EPS CONSTRUCTION THAT MEETS EN1077 STANDARDS - This helmet has two reinforced layers. The external cap is made from strong ABS material, which is covered by a supporting EPS outer shell for dual protection. This helmet fully complies with EN1077 standards and is CE certified.
- NUMEROUS FEATURES FOR A FULL AND ENJOYABLE DAY ON THE SLOPES - There's more than meets the eyes with the Lucky Bums Snow Sports Helmet. Inside, the internal fabric lining and ear padding include a hypoallergenic and antibacterial treatment while the ESP inner shell features multiport with mesh screens to prevent snow buildup and allow for ample airflow.
- COMFORTABLE WITH ADJUSTABLE FIT AND MULTIPLE SIZING OPTIONS - In addition to being vented and hypoallergenic, this helmet is comfortable to wear all day long. Its padded chin strap and integrated goggle loop enhance comfort and function. There are 4 sizes to choose from and each helmet includes a micro adjustable strap for the perfect fit for just about anyone.
- DURABLE AND DEPENDABLE WITH MANUFACTURER'S LIMITED LIFETIME WARRANTY - We know how upsetting it is to have a great day on the mountain spoiled by seemly, uncomfortable, and undependable gear. That's why we created Lucky Bums. We believe in our quality and put our money where our mouth is. Each helmet comes with Manufacturer's Limited Lifetime Warranty, which protects against defects in materials or workmanship. Look good, feel good, be protected.

## Customers Who Bought This Item Also Bought

Page 1 of 10



Bolle Mojo Snow Goggles

1,560  
\$11.69 - \$62.77



Ski Goggles, 2-Pack Skate Glasses for Kids, Boys & Girls, Youth, Men & Women, with UV 400...

114  
#1 Best Seller in Snowboarding Equipment  
\$11.99



Lucky Bums Snow Sport Helmet with Fleece Liner  
 118  
\$32.93 - \$63.87



Lucky Bums Kid's Alpine Series Doodlebug Helmet  
 14  
\$34.99 - \$79.99



OutdoorMaster Ski Goggles PRO - Frameless, Interchangeable Lens  
 417  
\$24.99 - \$49.99



Bolle Carve Snow Goggles  
 492  
\$17.49 - \$102.94



Traverse Dirus 2-in-1 Convertible Ski & Snowboard/Bike & Skate Helmet  
 78  
#1 Best Seller in Adult Bike Helmets  
\$39.98 - \$40.00



OutdoorMaster OTG Ski Goggles - Over Glasses  
 295  
\$19.99 - \$25.99



Arctix Men's SnowSports Cargo Pants  
 1,537  
#1 Best Seller in Men's Snowboarding Clothing  
\$26.95 - \$84.74



SUNVP Ski Helmet  
Upscale Warmest Windproof Adult Integrally Snow Sports Snowboard...  
 12  
\$42.99

# Case 3: News Recommendations

redd<sup>dit</sup> cidents nous creixent polèmic top gildejat wiki promogut

1 14 Register for a free ticket to The Artificial Intelligence Conference by MLconf (hello.conference.ai)  
promogut per shonburton 1 promogut 5 comentaris comarteix

2 8647 Kellyanne Conway is mad about 'alternative facts' blowup (washingtonpost.com)  
enviat fa fa 21 hores per JoofProobst a /r/politics 987 comentaris comarteix

3 35.2k Hate it when the door is sealed (i.reddit.it)  
enviat fa fa 22 hores per safakooze a /r/aww 345 comentaris comarteix

4 18.1k Man records heartbreaking video of the first time his own mother forgets who he is. (youtu.be)  
enviat fa fa 17 hores per Ihavegoodworkethic a /r/videos 1605 comentaris comarteix

5 59.5k Turn that frown upside down! (s-media-cache-ak0.pinimg.com)  
enviat fa fa 13 hores per thebrokenghost a /r/aww 477 comentaris comarteix

6 29.2k Killer whale lures birds in with dead fish (i.imgur.com)  
enviat fa fa 15 hores per XiKilzziX a /r/WTF 1320 comentaris comarteix

7 31.7k me\_irl (i.reddituploads.com)  
enviat fa fa 20 hores per DatDankWeedTho a /r/me\_irl 488 comentaris comarteix

8 691 Theresa May banned a US rapper from entering Britain, but won't ban Donald Trump (theguardian.com)  
enviat fa fa 18 hores per elnombre a /r/politics 60 comentaris comarteix

9 48.9k I wrote an essay on Skyrim and this is what my professor commented... (i.reddituploads.com)  
enviat fa fa 11 hores per Shonenlegend a /r/gaming 2269 comentaris comarteix

10 47.0k 3,000 grams of pure cannabis oil. (imgur.com)  
enviat fa fa 18 hores per Xdexter23 a /r/pics 3820 comentaris comarteix

11 17.6k Alternative memes Meme (i.reddituploads.com)  
enviat fa fa 22 hores per PM\_ME\_YA\_PETS a /r/teenagers 174 comentaris comarteix

# **Several** cases but **Two main** approaches

- Aggregated opinion recommenders
- Basic product association recommenders

# Data for Non-Personalized Recommenders





**Excellent**



**Above Average**



**Average**



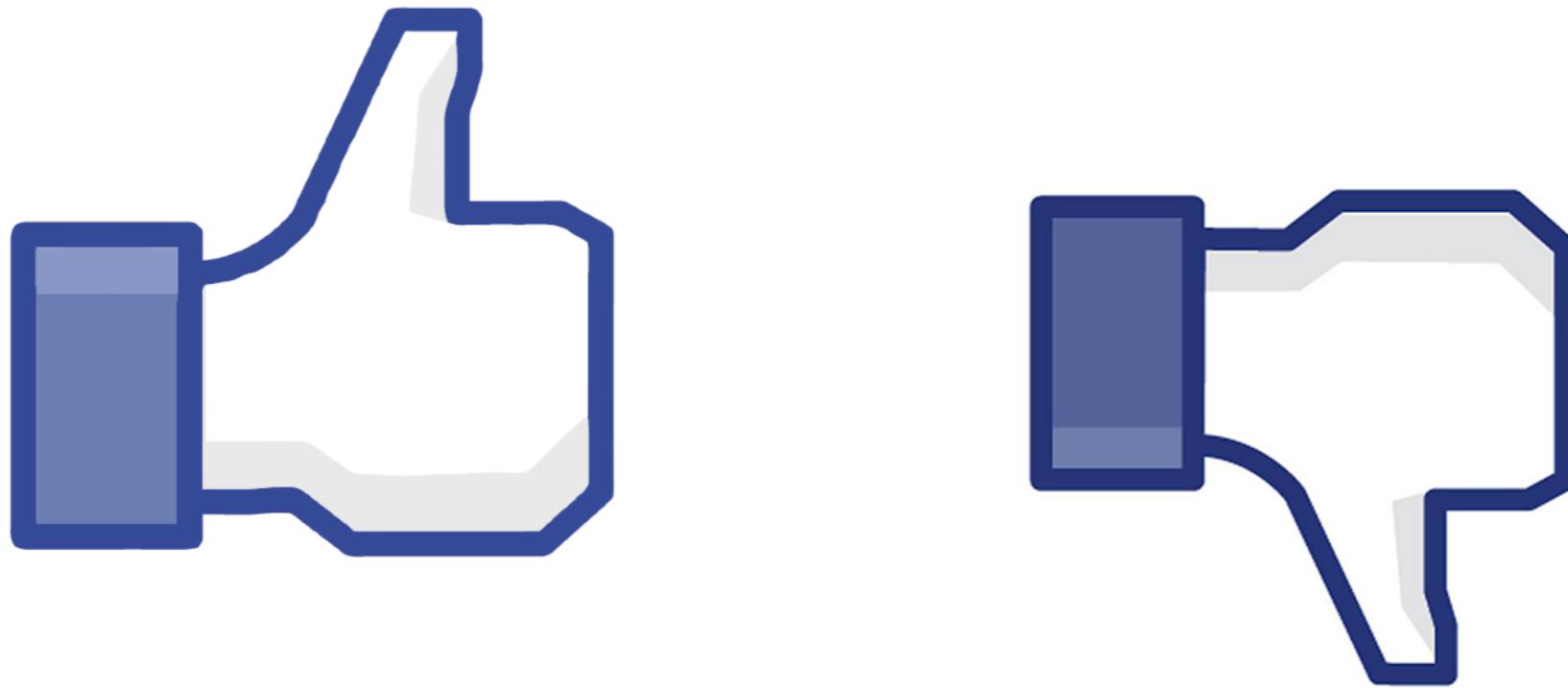
**Below Average**



**Poor**

Widely used

How many stars? 3 - 5 - 7 - 10 ?



# Thumbs and likes

Vote up/down

Like (+1)



# How to rank?

# How to score/rank?

- We first have to **understand the business case**. Several factors plays a role.
- Which information do we have about the items?
  - Bought / Seen / Rated / ...
  - From how many users do we have the info?
  - How old is that info?

# Ranking Considerations

- Confidence
  - How confident are we that this item is good?
- Risk tolerance
  - High-risk, high-reward
  - Conservative recommendations
- Domain and business considerations
  - Age
  - System goals



# How the ranking works?



# How the ranking works?

**quality**, **recency** and **quantity** of reviews.



# How the ranking works?



## The vote average for film or TV show "X" is clearly wrong! Why are you displaying another rating?

IMDb displays **weighted** vote averages rather than raw data averages. Various filters are applied to the raw data in order to eliminate and reduce attempts at "vote stuffing" by individuals more interested in changing the current rating of a movie or TV show than giving their true opinion of it.

Although the raw mean and median are shown under the detailed vote breakdown graph on the ratings pages, the user rating vote displayed on a film / show's page is a weighted average. In order to avoid leaving the scheme open to abuse, we do not disclose the exact methods used.

We can provide some more detail here to reassure you that our methods are both sound and fair. First of all, the same formula is applied universally across the database to all movies and shows without exception so there is no bias in when and where the scheme operates. Occasionally we receive mail from people who seem to assume that some favorite movie or show has been victimized by the weighted ratings whereas this is not the case. The objective of the scheme is to present a more representative rating which is immune from abuse by subsets of individuals who have combined together with the aim of influencing (either up or down) the ratings of specific movies or shows. This includes people involved in the production of a movie / TV show and their friends or fans trying to unduly raise the rating far above that of where the typical IMDb users would rate it.

The scheme combines a number of well-known and proven statistical methods, including a trimmed mean to reduce extreme influences and, most importantly a complex voter weighting system to make sure that the final rating is representative of the general voting population and not subject to over influence from individuals who are not regular participants in the poll. The scheme has been developed internally over the 25 years which the poll has been in operation and tuned on a regular basis to make sure it remains fair.

Most of the feedback we receive about our votes is based on the incorrect assumption that all votes cast by our users have the same impact on the final rating. This is not the case. Different votes may have different **weight** when they're used to calculate the final weighted rating. Most people think of the *arithmetic mean* when they hear the word *average*. When calculating an arithmetic mean, all votes are treated equally: the average is the sum of all the votes divided by the number of votes.

A weighted average, however, is defined as "an average that takes into account the proportional relevance of each component, rather than treating each component equally". There's nothing arcane or mysterious about it: it's a very simple, universally accepted statistical method, commonly used in a wide variety of fields (from financial analysis to student reports).

For example, an automobile magazine reviewing a new car may give it high marks in several categories (appearance, comfort, fuel efficiency, price, number of cup holders). However a model with high ratings in all those categories may still get a low overall score if it gets a low vote in just one or two other areas (like speed or safety): even the cheapest, nicest-looking and most fuel-efficient car isn't worth buying if the fuel tank explodes when you hit the brakes or if its top speed is only 15 miles per hour. Clearly, a high (or low) vote in some categories has more **weight** than the same exact vote cast in another category, so the final rating takes these differences into account.

Our calculations follow a similar principle: some votes have more weight than others. We will **not** disclose exactly if/when/how certain votes are weighted differently because the idea is, among other things, to give a more objective rating and neutralize attempts to artificially inflate or deflate the average user rating on a film or show.

We're confident that our system provides a reasonably faithful representation of our users' opinions. This scheme is applied uniformly to all films listed in the database, without exception; and since it has proved to be very effective we do not plan to abandon it, though we periodically revisit and tweak it to make it even more accurate and tamper-proof.



# How the ranking works?



**-34% IHOY!**

**The One Barcelona GL** ★★★★   
Eixample, Barcelona – Cerca del metro  
Hay 8 personas mirando en este momento  
Reservado 157 veces esta semana  
 Piscina al aire libre  Spa y centro de bienestar  
 Tiene 1 restaurante  


**Habitación Doble - 1 o 2 camas**  
Puedes cancelar más tarde.  
Aprovecha y consigue un buen precio hoy.

**Muy solicitado. ¡Solo quedan 2 habitaciones en nuestra página!**

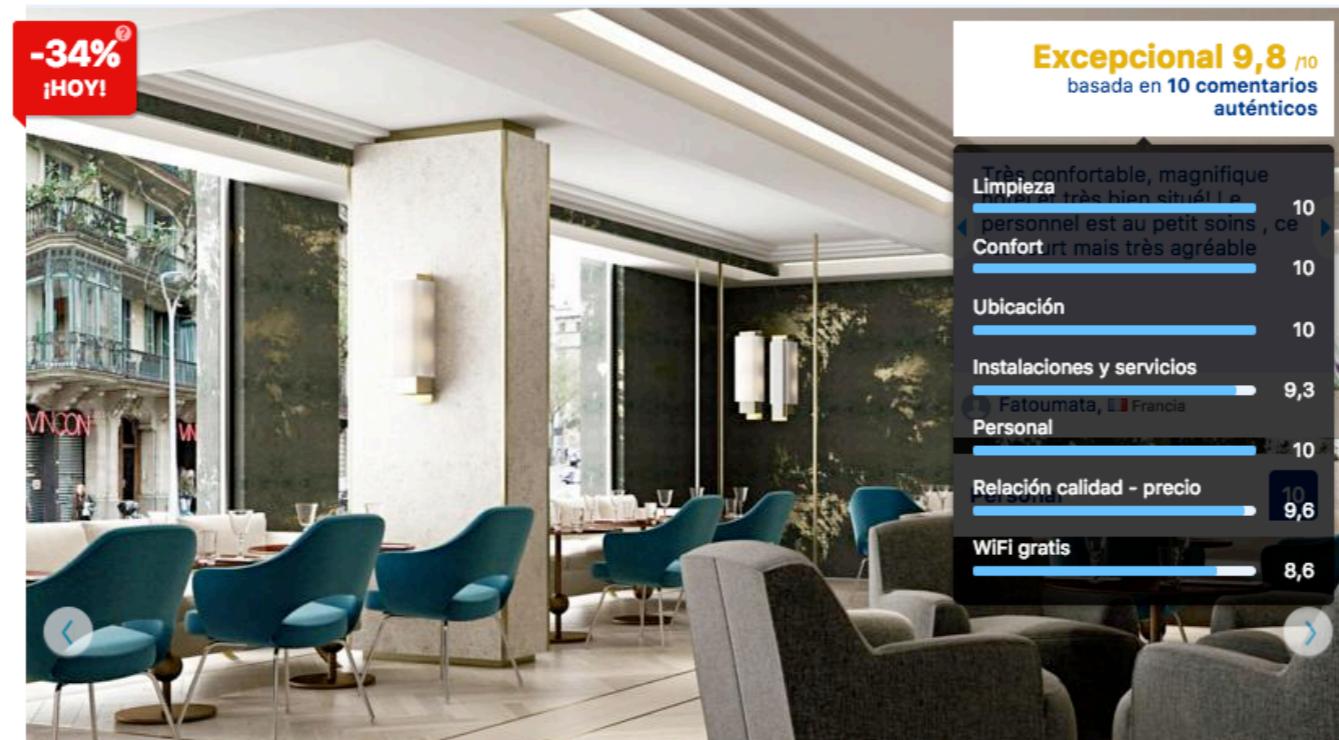
**Excepcional 9,8**  
**Ubicación 10,0**  
– según 10 comentarios

**Precio para 2 noches**  
~~€ 560~~ **€ 367**  
**Cancelación GRATIS.**  
Sin pago por adelantado.

**Ver las 8 habitaciones disponibles >**

# Booking.com

## How the ranking works?



# Hands on time!



# Damped means

- Problem: **There is low confidence with few ratings**
- Solution: **Assume that, without evidence, everything is average**
- Ratings are evidence of non-averageness
- $k$  controls strength of evidence required

$$damped\_mean = \frac{\sum_u r_{u,i} + k\mu}{n + k}$$

# Confidence Intervals

- From the reading: lower bound os statistical confidence interval
- Choice of bound affects risk/confidence
  - Lower bound is conservative: be sure it's good
  - Upper bound is risky: there's a chance of amazing

# Domain Considerations: Time



old stories are not interesting, even if they have many upvotes!



items have short lifetimes

# Case 3: News Recommendations

redd<sup>dit</sup> cidents nous creixent polèmic top gildejat wiki promogut

1 14 Register for a free ticket to The Artificial Intelligence Conference by MLconf (hello.conference.ai)  
promogut per shonburton 1 promogut 5 comentaris comarteix

2 8647 Kellyanne Conway is mad about 'alternative facts' blowup (washingtonpost.com)  
enviat fa fa 21 hores per JoofProobst a /r/politics 987 comentaris comarteix

3 35.2k Hate it when the door is sealed (i.reddit.it)  
enviat fa fa 22 hores per safakooze a /r/aww 345 comentaris comarteix

4 18.1k Man records heartbreaking video of the first time his own mother forgets who he is. (youtu.be)  
enviat fa fa 17 hores per Ihavegoodworkethic a /r/videos 1605 comentaris comarteix

5 59.5k Turn that frown upside down! (s-media-cache-ak0.pinimg.com)  
enviat fa fa 13 hores per thebrokenghost a /r/aww 477 comentaris comarteix

6 29.2k Killer whale lures birds in with dead fish (i.imgur.com)  
enviat fa fa 15 hores per XiKilzziX a /r/WTF 1320 comentaris comarteix

7 31.7k me\_irl (i.reddituploads.com)  
enviat fa fa 20 hores per DatDankWeedTho a /r/me\_irl 488 comentaris comarteix

8 691 Theresa May banned a US rapper from entering Britain, but won't ban Donald Trump (theguardian.com)  
enviat fa fa 18 hores per elnombre a /r/politics 60 comentaris comarteix

9 48.9k I wrote an essay on Skyrim and this is what my professor commented... (i.reddituploads.com)  
enviat fa fa 11 hores per Shonenlegend a /r/gaming 2269 comentaris comarteix

10 47.0k 3,000 grams of pure cannabis oil. (imgur.com)  
enviat fa fa 18 hores per Xdexter23 a /r/pics 3820 comentaris comarteix

11 17.6k Alternative memes Meme (i.reddituploads.com)  
enviat fa fa 22 hores per PM\_ME\_YA\_PETS a /r/teenagers 174 comentaris comarteix

# How to score **NEW** stories

$$\frac{(U - D - 1)^\alpha}{(t_{now} - t_{post})^\gamma} \times P$$

- where U is up-votes, D is down-votes and  $t_{now} - t_{post}$  is the age of the new. P is a penalty term related to the new. Where  $\alpha$  and  $\gamma$ , are a decay polynomial parameter and the gravity parameter
- Net up-votes, polynomially decayed by time

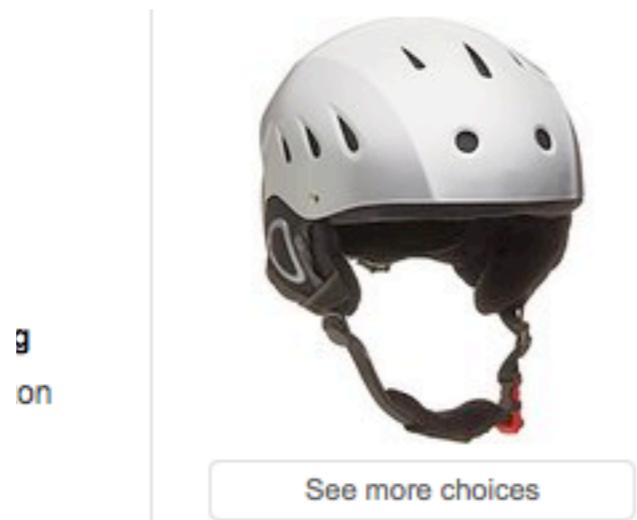
# How to score **NEW** stories

$$\log_{10} \max(1, |U - D|) + \frac{\text{sign}(U - D)t_{post}}{45000}$$

- Log term applied to votes
  - decrease marginal value of later votes
- A 1000-vote item is better than a 100-vote item. But is a 2000-vote item that much better yet?

# Difficulties with Rating





9  
on

### Lucky Bums Snow Sport Helmet with Fleece Liner

by Lucky Bums

\$32<sup>93</sup> - \$63<sup>87</sup> 

Some sizes/colors are Prime eligible

More Buying Choices

\$32.93 new (4 offers)

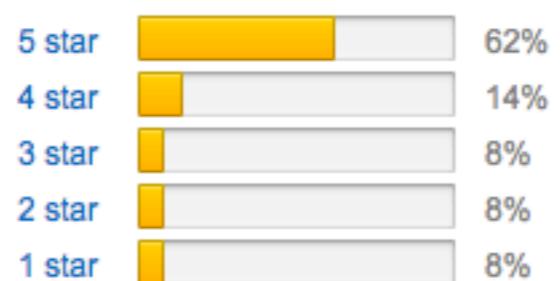
\$27.99 used (4 offers)

FREE Shipping on eligible orders

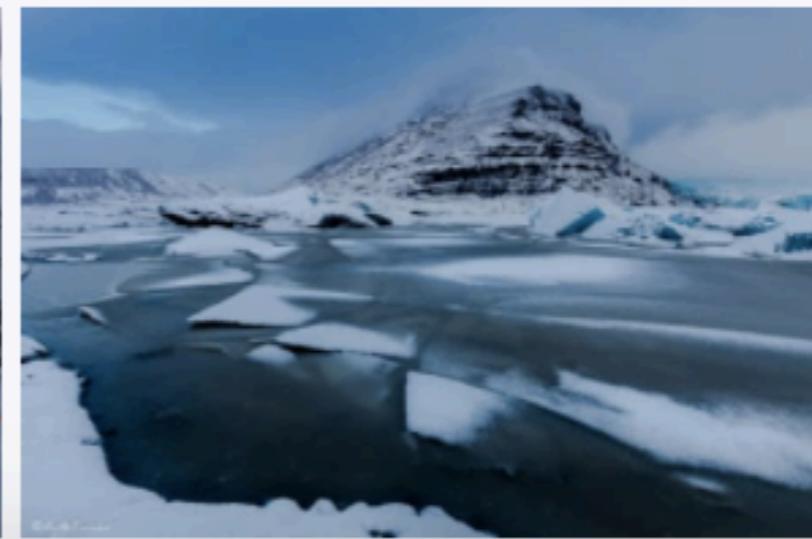
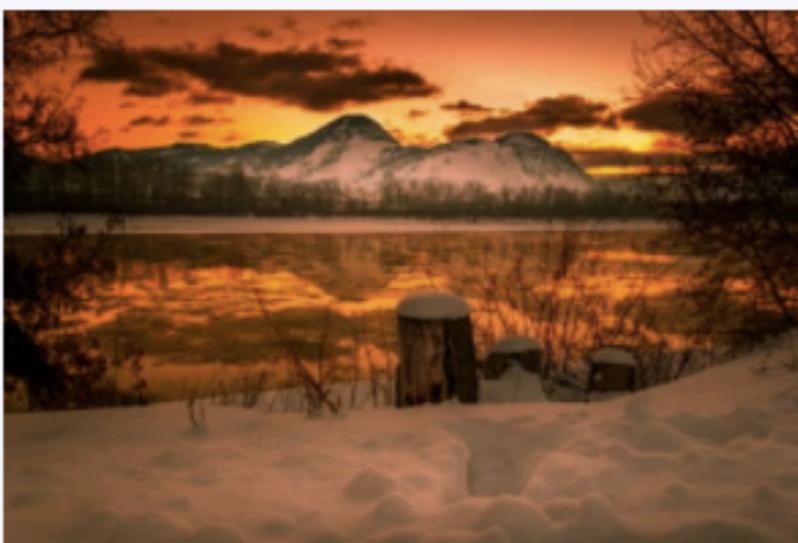
Show only Lucky Bums items

 118

4.2 out of 5 stars



[See all verified purchase reviews ▾](#)



# Scales and Normalization

- The most simple approaches
  - Average rating / Up-vote proportion
    - Does not show popularity
    - Of people who vote, do they like it?
  - # Up-vote / # of likes
    - Shows popularity
    - No controversy
  - %  $\geq 4$  stars
  - Full distribution
    - Complicated

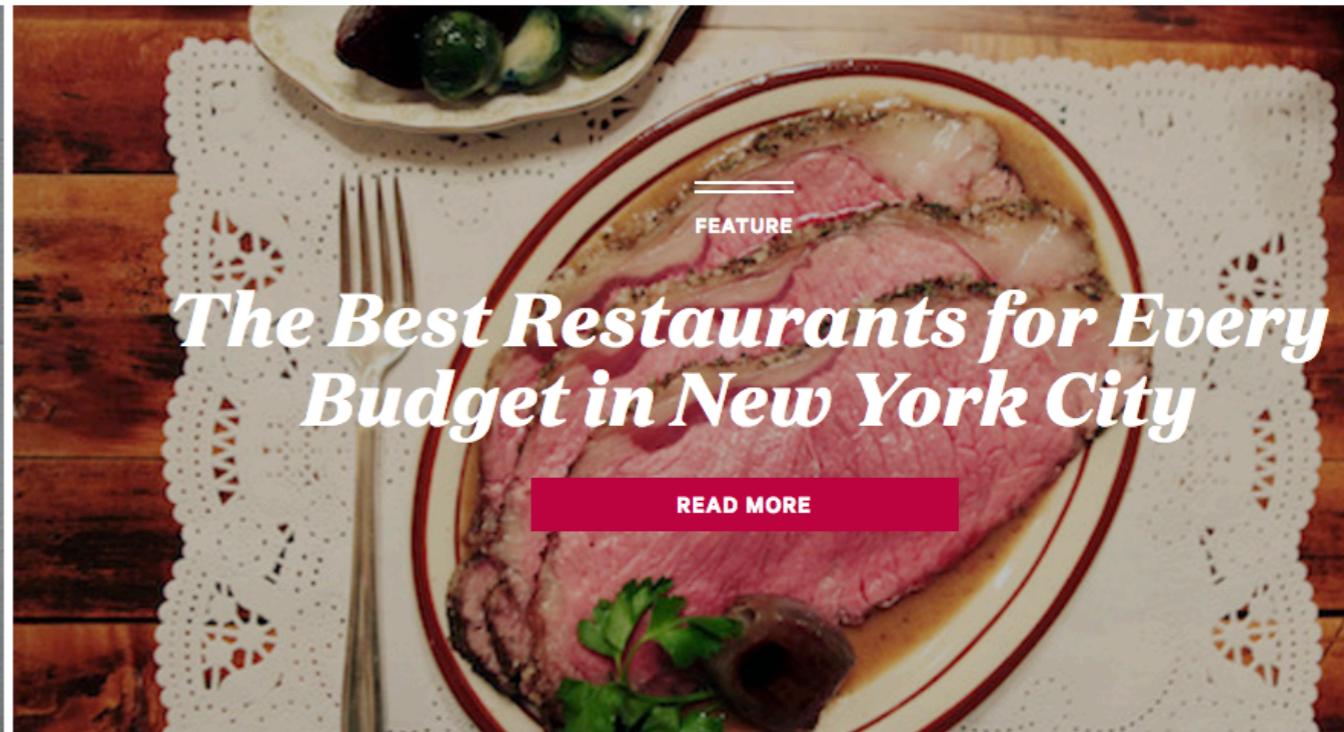
Popularity and  
controversy are  
important

# Predict with sophisticated score?

- Theoretically it is a fine thing to do.
- Be careful with transparency
  - If you say "average rating" and shows damped mean, users can be confused

# Difficulties with Rating

- Are ratings reliable and accurate?
- Do the users change his/her preferences?
- What is a rating mean?

Find a great place in New York City

## BEST OF NEW YORK CITY



OVERALL



PIZZA



STEAKHOUSE



BURGER



FRENCH



SEAFOOD



CHINESE



ITALIAN



MEXICAN



It was really famous and then it become to getting worse

# Why?

Your opinion diverges with the community averages

# Scales and Normalization

- The most simple approaches
  - Average rating / Up-vote proportion
  - # Up-vote / # of likes
  - %  $\geq$ 4 stars
  - Full distribution

People who buy X also Y..

# People who buy X also Y..

- How can we do this? How is our dataset (or a typical one) ?
  - User profiles?
  - Transaction data (people who bought them at the same time)?
  - User profile but time-constrained?

# How to score

- Percentage of X-buyers who also bought Y

$$\frac{\text{X and Y}}{\text{X}}$$

**Hands on time!**

# Hands on time!



# How to score

- Percentage of X-buyers who also bought Y

$$\frac{\text{X and Y}}{\text{X}}$$

**Hands on time!**

- What happens? Is it a good measure?
  - What happens with very popular items?

# How to score

- Let's adjust by looking at whether X makes more likely than not  $X(!X)$

$$\begin{array}{c} X \text{ and } Y \\ \hline X \\ \hline !X \text{ and } Y \\ \hline !X \end{array}$$

- It focuses on how Y increases buying X
- Value higher than one says that is more probable to be bought X when Y is bought than not.

# How to score

- Let's adjust by looking at whether  $X$  makes more likely than not  $X(!X)$

$$\begin{array}{c} X \text{ and } Y \\ \hline X \\ \hline !X \text{ and } Y \\ \hline !X \end{array}$$

- It focuses on the ratio of  $X$  to  $!X$
  - Value high probability of  $X$  and  $Y$  over low probability of  $X$  and  $!Y$
- Can it be used to recommend movies?**

# Another Association Rule

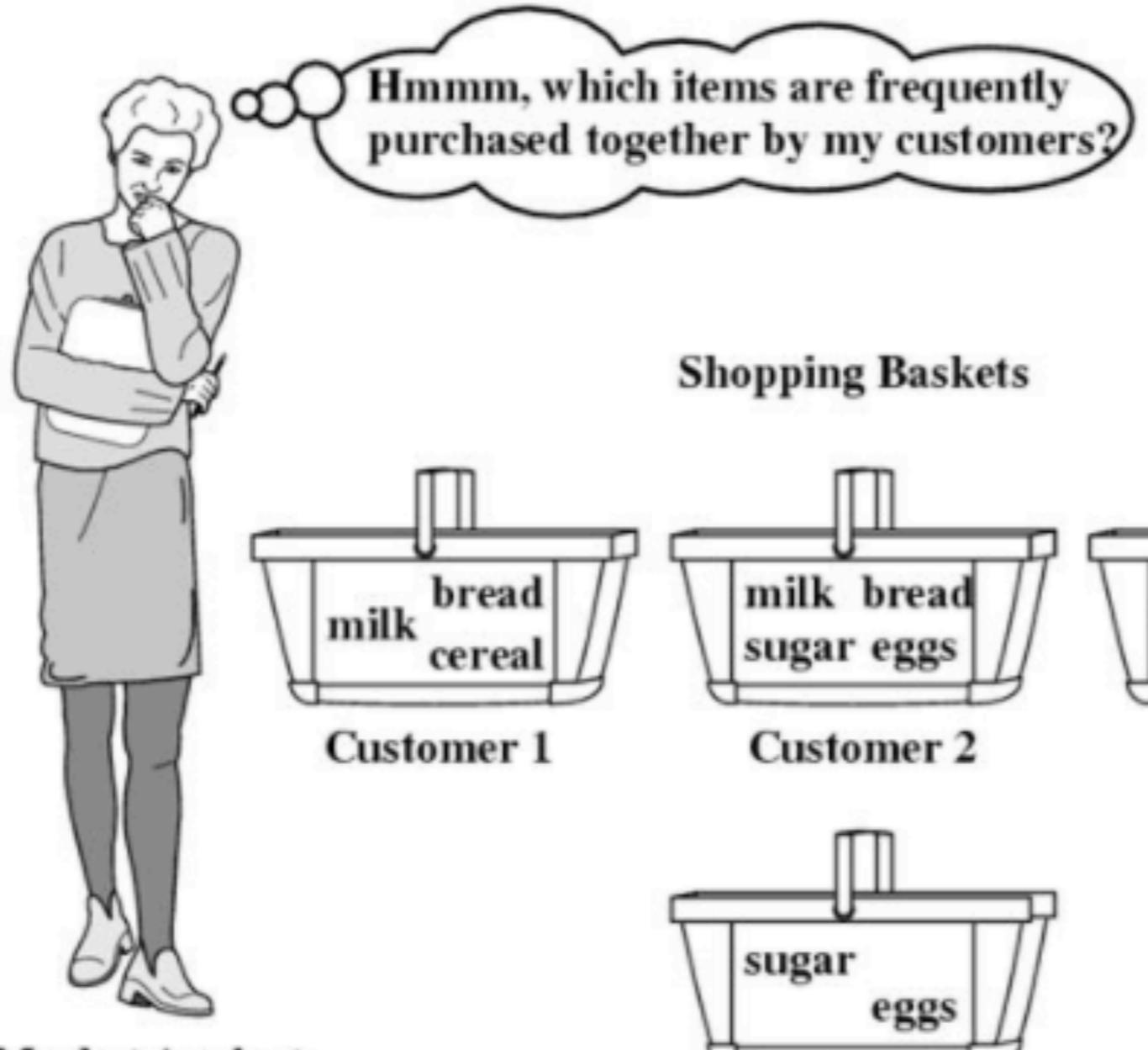
- Association rule mining

$$\frac{P(X \text{ and } Y)}{P(X) P(Y)}$$

Let's see a much more complicated problem!

# Market Basket





Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold

**Sugar -> Milk**

*Support = 13.5% and confidence = 70%*

Confidence

“When a customer who buys sugar, in **70%** of the cases, he or she will also buy milk!”

We find this happens in **13.5%** of all purchases”

support

# Some interesting numbers

Cardinality of Itemsets	Number of Itemsets
1	100
2	4,950
3	161,700
4	3,921,225
5	75,287,529
6	1,192,052,400
7	16,007,560,800
8	186,087,894,300

If a supermarket has at least 10,000 different items, there are almost 50,000,000 possibles 2-itemsets, imagine how many with 3-itemsets and 4-itemsets

# Association Rules

## Terminologies:

items : 1, 2, 3,....

itemSet: {1},{1,2},{1,4,2,5},...

1-itemset: {1},{2},{3}

2-itemset: {1,2},{2,3},{1,3}

DB of "Basket Data"

TID	items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5



association rules

{1} => {3}

{2,3} => {5}

{2,5} => {3}

⋮

## association rule metrics:

$$\text{confidence} \equiv \frac{\text{transactions containing } IS_a \cup IS_c}{\text{transactions containing } IS_a}$$

$$\text{support} \equiv \frac{\text{transactions containing } IS_a \cup IS_c}{\text{all transactions}}$$

### Fast algorithms for mining association rules

R Agrawal, R Srikant

Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499

21710 1994

# Association Rules

- **Step I:** Find all itemsets with *minimum support (minsup)*

<u>TID</u>	<u>items</u>	<u>support</u>	<u>itemsets</u>
100	1 3 4	0.25	{4}, {1,2}, {1,4}, {1,5}, {3,4}, {1,3,4}, {1,2,3}, {1,2,5}, {1,3,5}, {1,2,3,5}
200	2 3 5		
300	1 2 3 5	0.5	{1}, {1,3}, {2,3}, {3,5}, {2,3,5}
400	2 5	0.75	{2}, {3}, {5}, {2,5}

- **Step II:** Generate rules from *minsup'ed itemsets*

<u>support</u>	<u>confidence</u>	<u>rules</u>
0.5	66%	{3}=>{1}, {3}=>{2}, {2}=>{3}, {3}=>{5}, {5}=>{3}, {5}=>{2,3}, {3}=>{2,5}, {2}=>{3,5}, {5,2}=>{3}, {5,3}=>{2}
0.5	100%	{1}=>{3}, {5,3}=>{2}, {2,3}=>{5}
0.75	100%	{5}=>{2}, {2}=>{5}

Fast algorithms for mining association rules

R Agrawal, R Srikant

Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499

21710 1994

# A priori Method

**Apriori principle:** Any subset of a frequent itemset must be frequent

**Step 1:** Find the frequent itemsset: the set of items that have minimum support.

- A subset of a frequent itemset must also be a frequent itemset  
i.e. if  $\{1,2\}$  is a frequent itemset, both  $\{1\}$  and  $\{2\}$  should be a frequent itemset
- Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)

**Step 2:** Use the frequent itemsets to generate association rules

Fast algorithms for mining association rules

R Agrawal, R Srikant

Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499

21710

1994

# More

# When preferences are provided?

- Consumption - during or immediately after the consumption
- Memory - some time after experience
- Expectation - the item has not been experienced

# Which one are true?

- To increase the number of products a user buys
- To help a user find the most popular products
- To show a user complementary products
- To help users to find related products

# Conclusions

- Sparsity, inconsistency, temporal concerns, make data messy
- Simply scoring does not much necessarily match the domain score
  - good ways to deal with this: decay, time, penalties, damping