Master on Foundations of Data Science

Recommender Systems

Non-Personalized Recommender Systems

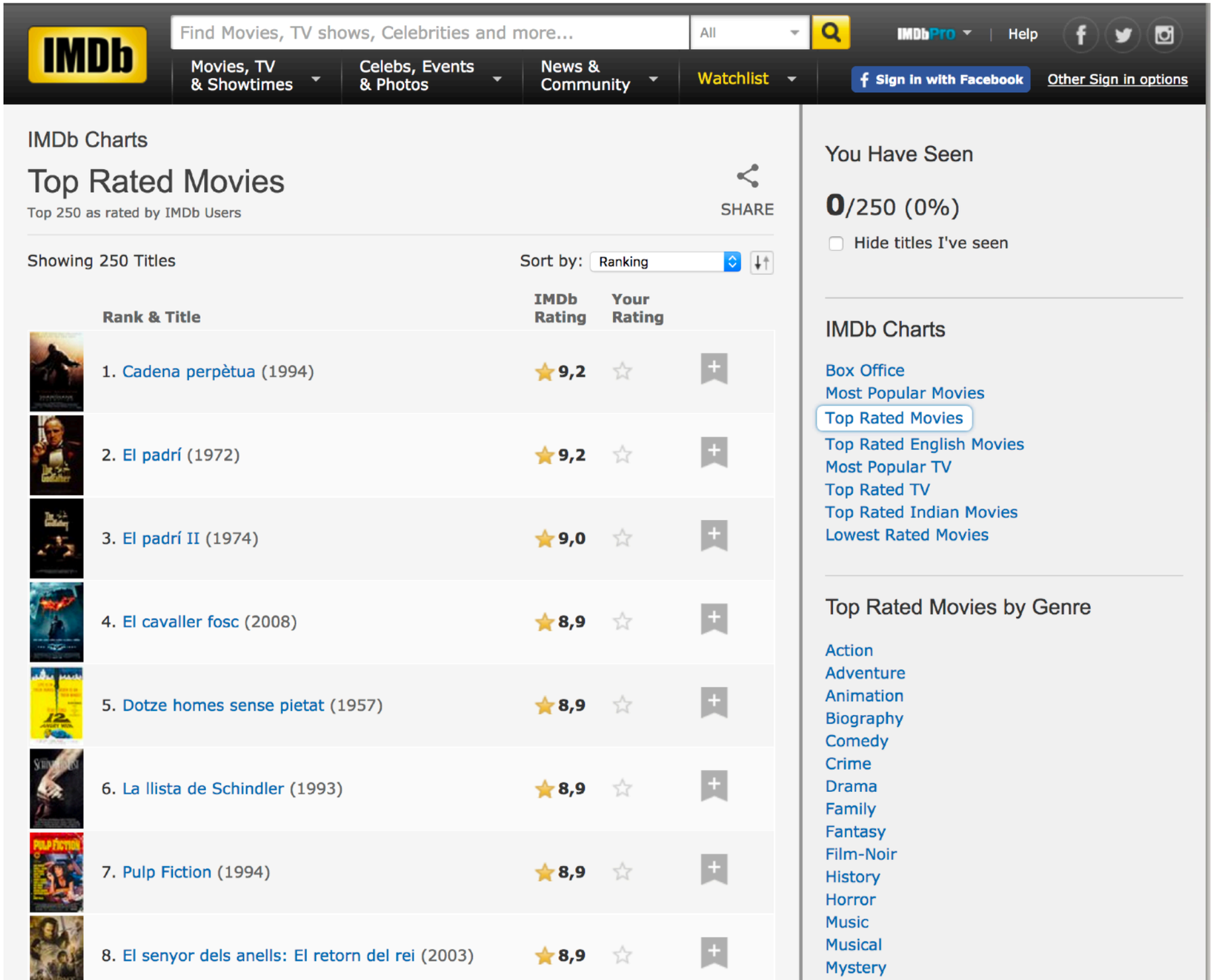
Santi Seguí | 2017-2018

What is a non-personalized recommender system

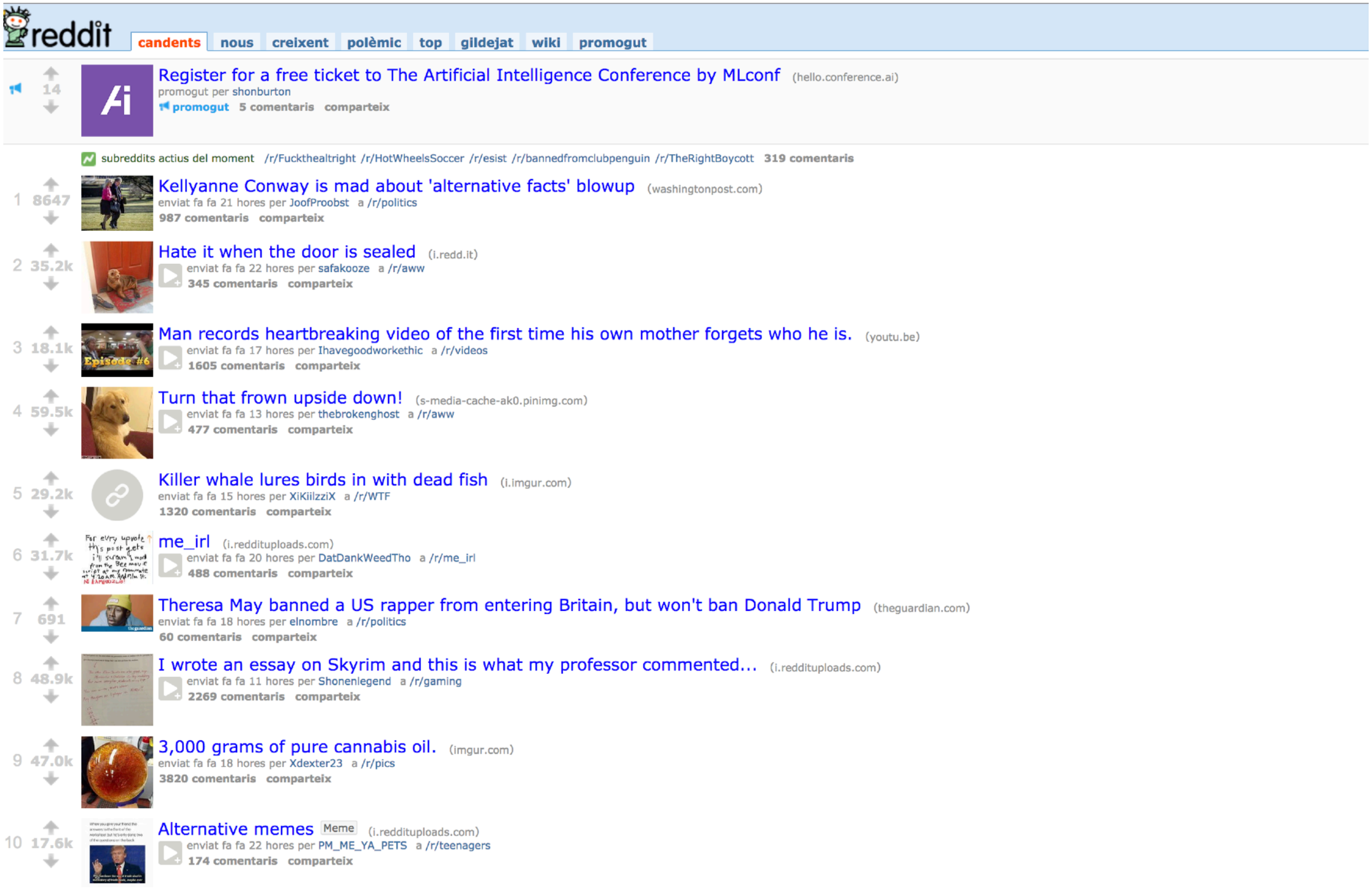
A non-personalized recommender system is one that **makes the same recommendations for everyone**.

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

When is it useful?



Case 1: News Recommendations



Case 2: Top Recommendations

See more choices

Lucky Bums Snow Sport Helmet with Fleece Liner
by Lucky Bums
$32.93 - $63.87 Prime
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
$64.95 - $120.00 Prime
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
$59.95 - $212.69 Prime
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

See more choices

SUNVP Ski Helmet Ultralight Integrally Warmest Windproof Snowboards Snow Sports Helmet Unisex Adult
by SUNVP
$36.99 ~~$110.00~~ Prime
Some sizes/colors are Prime eligible
FREE Shipping on eligible orders
Show only SUNVP items
12

Smith Optics SO-H15AS Men's Aspect Snow Helmet, Matte Charcoal-Large
by Smith
$99.95 Prime
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
$69.99 - $338.22 Prime
Some sizes/colors are Prime eligible
More Buying Choices
$159.94 new (20 offers)
Show only Giro items
118

Case 3: 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 Prime

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 Snow Goggles for Men &…
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 Ski / Snowboard Goggles for Men, Women & Youth…
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 Prime

Several cases but Two main approaches

1. Recommenders based on **aggregated opinions**

2. Recommenders based on **basic product associations**

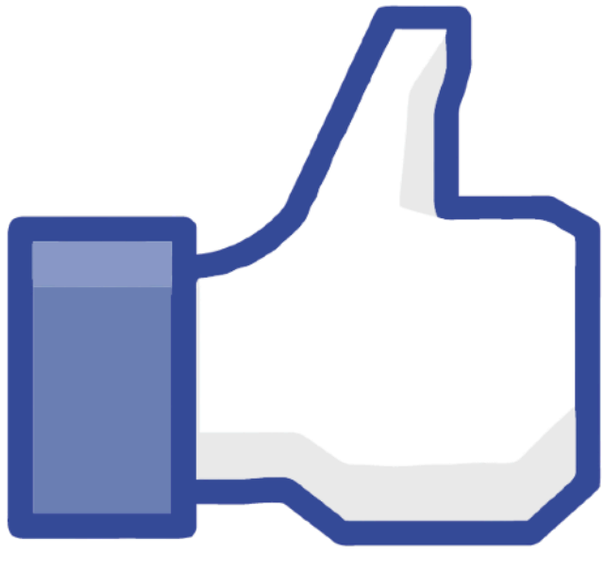
Recommenders based on **aggregated opinions**

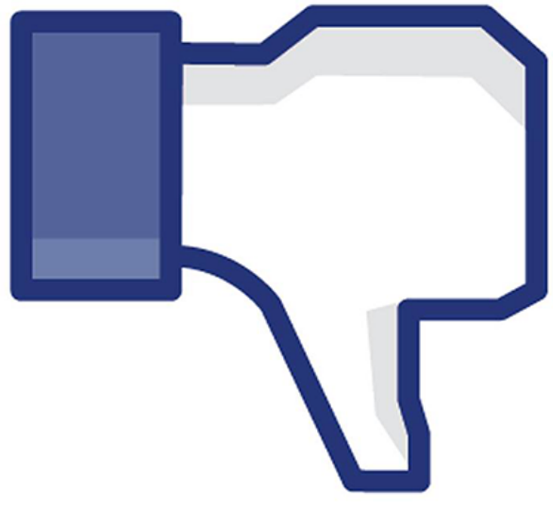
How to rank?

Data for Non-Personalized Recommenders

Widely used

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





Thumbs and likes

Vote up/down

Like (+1)

How to score/rank items?

- 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?

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

tripadvisor®

How the ranking works?

tripadvisor®

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.

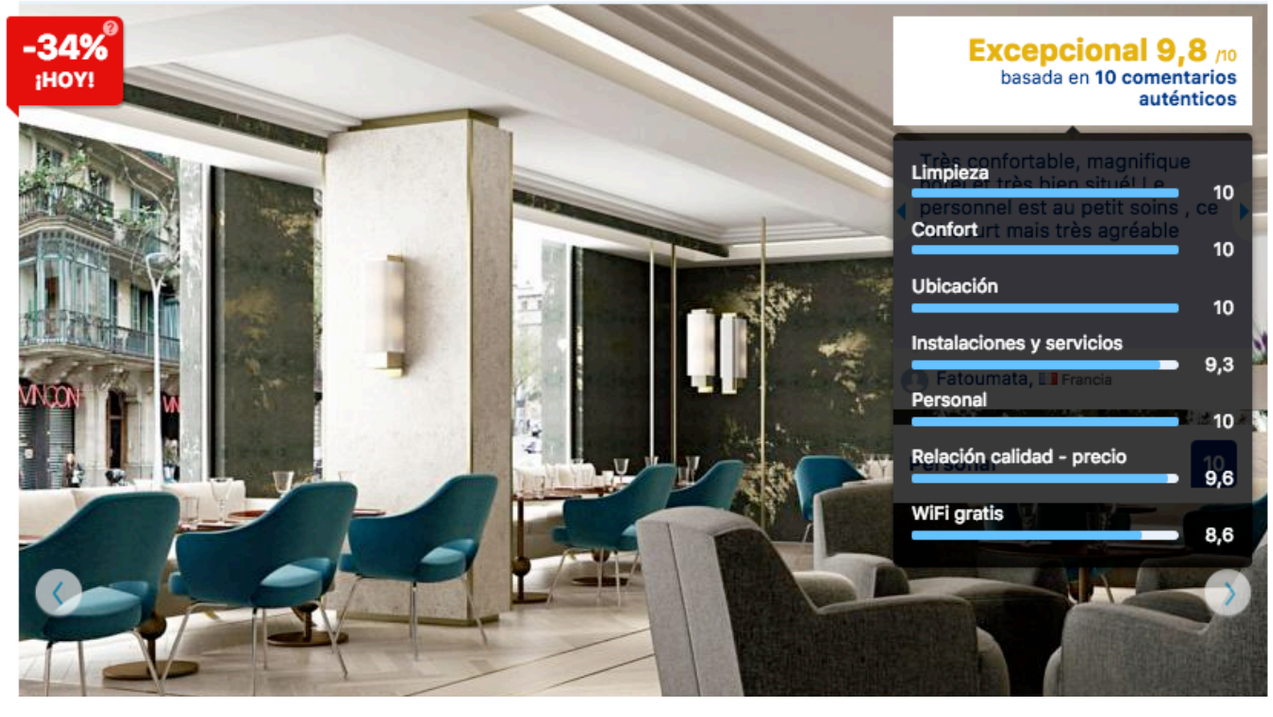
We're confident that our system provides a reasonably faithful representation of our users' opinions. This scheme is applied uniformly to all films listed in the database, without exception; and since it has proved to be very effective we do not plan to abandon it, though we periodically revisit and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

Booking.com

How the ranking works?

Booking.com

How the ranking works?



Booking.com

If it was known everyone would try to hack it!

Booking.com

The way we present properties **needs to be relevant**. We aim to **match the right product**, with the **right person**, at the **right time**. Therefore, ranking is a complex and **scientific approach to optimizing this for both partners and customers**.

Our goal is to give customers **suggestions of properties that will make them happy**, therefore driving the right customers and high-quality traffic to partners. **When partners and guests are well matched and proper expectations are set, this leads to higher conversion.**

What is included in ranking?

There are many different factors and elements that influence ranking. This includes data from partners and customers, **ranging from click and booking behavior to conversion**, **availability** and **pricing data**, as well as the **facilities of a property**. Ranking, therefore, isn't just dependent on one thing.

Basically, anything we know about our customers and partners is factored into how our search results are ordered on Booking.com.

Millions visit Booking.com every day and we want to make your property stand out as much as possible. Every customer and accommodation partner is unique, so our dedicated team is here to make sure that your property appeals to the right audience.

How do you influence visibility?

We have various tools to help you get the most out of your Booking.com page. But first, make sure you have all the basics set up, so you can grow organically.

We offer personalized advice based on 4 specific areas: pricing, availability, content, and extra features. All of these create a complete profile so we can match you with the right audience.

Through continued improvement of these topics, plus a good guest review score, you'll help your property get a higher position in search results – making it stand out to more bookers looking for your accommodation type.

A higher ranking position **will help lead to more bookings**, **fewer cancellations**, and more **satisfied guests**.

How do I get better visibility on Booking.com?

Updated 3 days ago

How you appear in search results is a very dynamic process that takes the millions of searches each day—with their own individual needs—and matches them to the best property based on their criteria. We know our customers. Whether they're family travelers, business travelers, budget travelers, or city lovers, we cater to all and will use our knowledge to match you well for a great experience. We have more info on how ranking works on Booking.com to explain it all.

Do you want to appear higher in search results on Booking.com? Here are a few tips:

High property page score:

First impressions matter, so making sure you give bookers all the info they expect can bring more bookings since they understand what exactly is offered.

Pricing:

Keep your rates competitive and attract more bookers. A recent TripAdvisor study showed that price is a key factor in 95% of bookings. Learn more about rates in your area with RateIntelligence.

Availability:

Don’t miss out on potential new bookings. By making sure you’re constantly bookable on Booking.com, you’ll help improve your ranking. Remember to always show rates for high-demand dates. Learn why availability is important with this video.

Photos:

Customers are always attracted to properties with great, high-resolution photos. You can check the size of your photos, as well as add and remove images, by checking the Property tab. Check out this photography guide.

Extra features:

How to crack the Booking.com algorithm

Comment Print Dec 18.2014

If we want to improve our positioning on Booking.com as rapidly as possible in order to increase sales, it is obvious what needs to be done.

Increase commissions!

This maximizes the profits of Booking.com and in turn its web algorithm will reward us with good search results.

NB: This is an analysis by Francesco Canzoniere, ex-CEO of Viajar and now founder and managing director of Travel Performance.

But do increases in commissions have consequences? Yes, of course.

Just as when hotels go headlong into a price war (downwards), the same happens when they enter into a commission war, they all end up with the same market share, but with lower net income.

At times It is possible that the hotel feels obliged to pay higher commissions but it is also clear in many cases that this solution is chosen for its “speed and simplicity”.

But is it sustainable to continue working with a strategy of “get some work and improvise”?

Hands on time!

https://github.com/ssegui/RecSysMaster/tree/master/notebooks

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

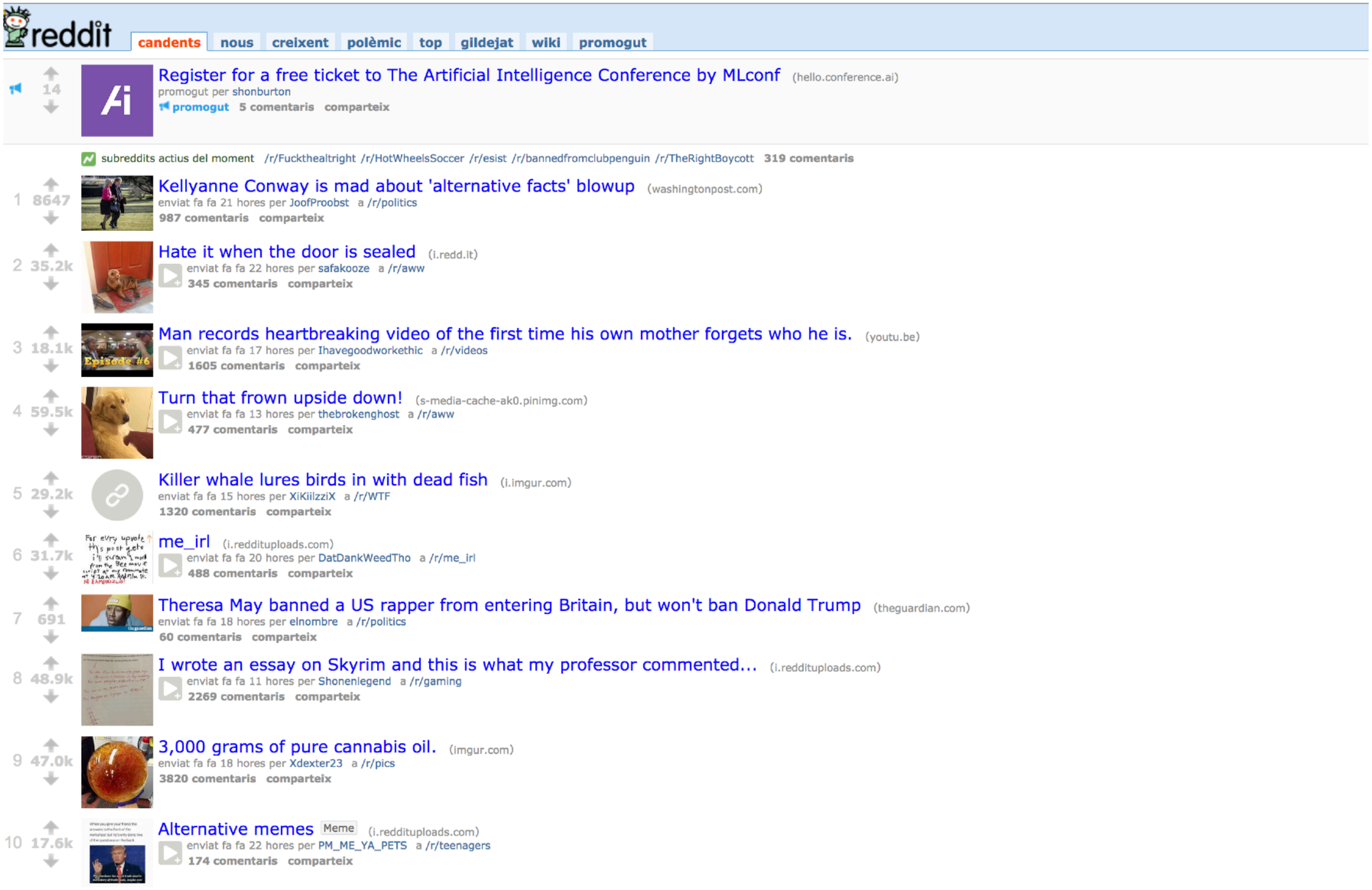
- From the reading: lower bound os statistical confidence interval
- Choice of bound affects risk/confidence
  - Lower bound is conservative: be sure it’s good
  - Upper bound is risky: there’s a change of amazing

Domain Considerations: Time

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

items have short lifetimes

Case 1: News Recommendations



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} it the age of the new. P is a penalty term related to the new. Where α and γ , are a decay polynomial parameter and the gravity parameter

- Net up-votes, polynomially decayed by time

How to score NEW stories

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

- 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

See more choices

Lucky Bums Snow Sport Helmet with Fleece Liner
by Lucky Bums

$32.93 - $63.87 Prime
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

5 star	62%
4 star	14%
3 star	8%
2 star	8%
1 star	8%

See all verified purchase reviews

500px Discover Blog About Business

Search for photos, location, or people

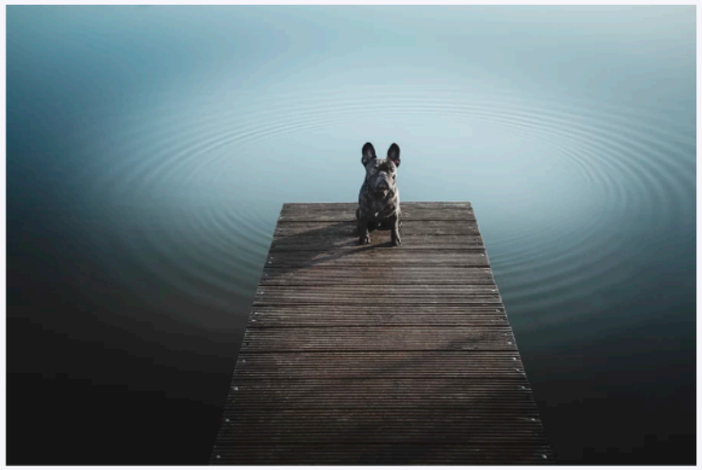
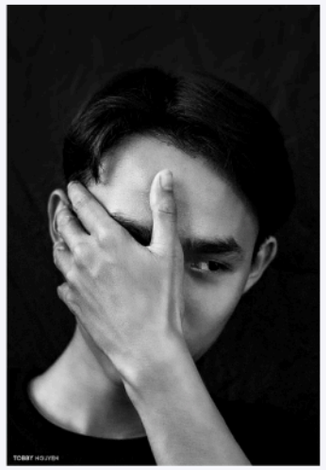
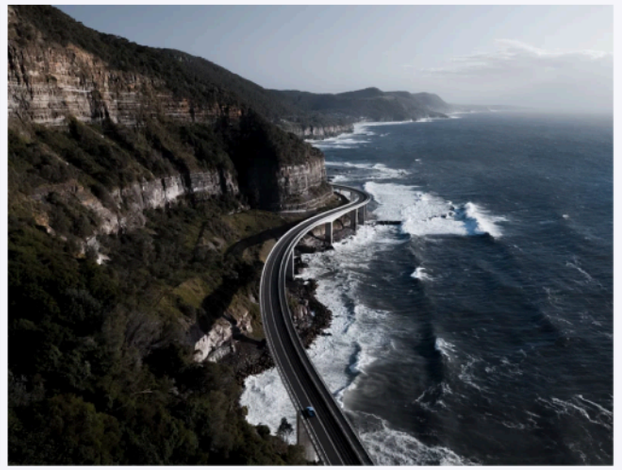
Log in Sign Up

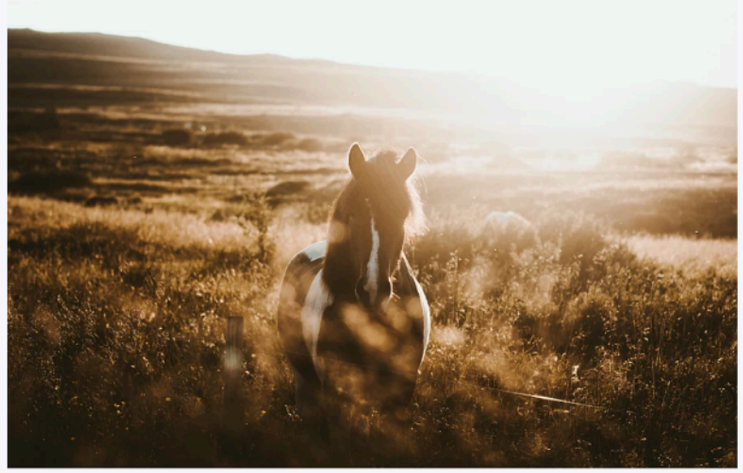
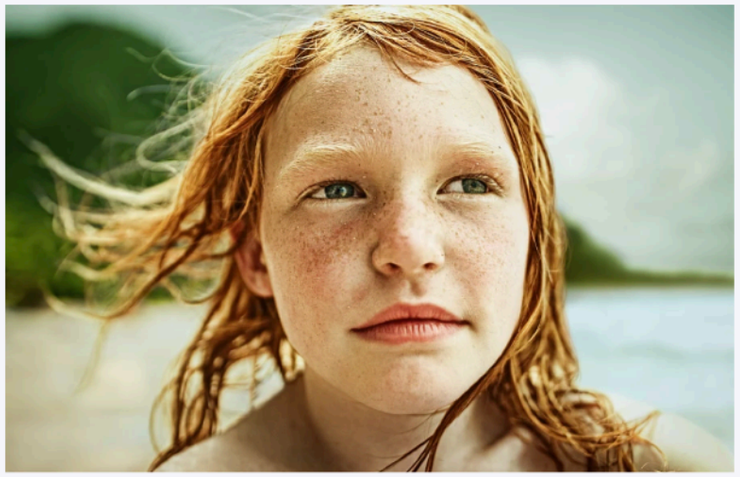
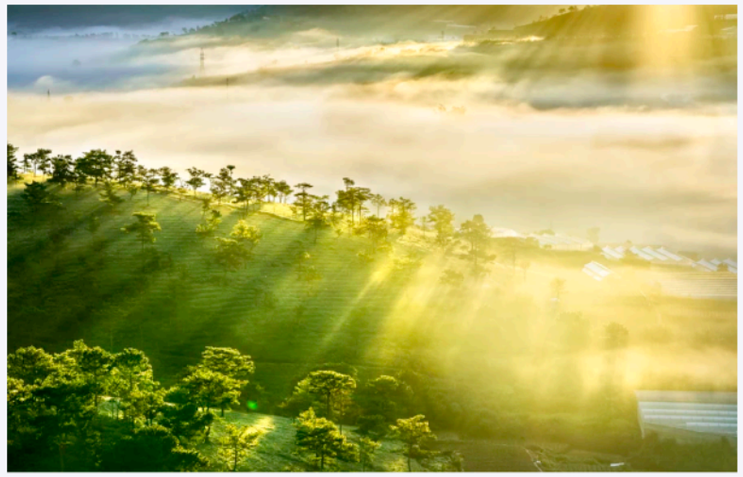
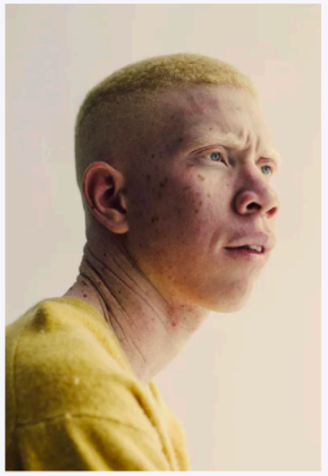
Hand-picked by our editors

Check out photos selected by our 500px Editors.

FOR YOU EDITORS' CHOICE POPULAR UPCOMING FRESH PLACES

All categories



UNIVERSITAT DE BARCELONA

DATA SCIENCE @ UNIVERSITAT DE BARCELONA

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
  - % >=4 stars
  - Full distribution
    - Complicated

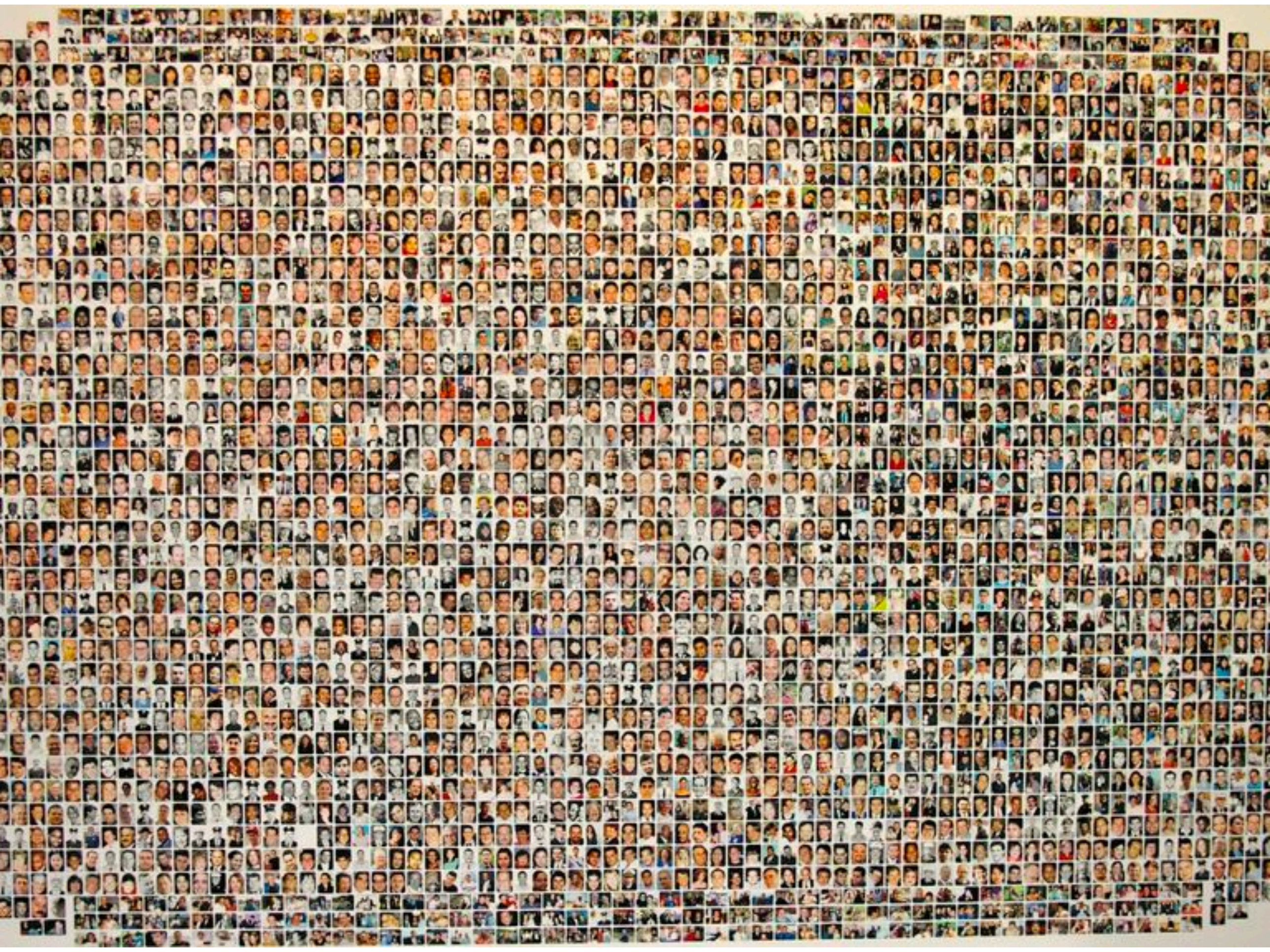
Popularity and controversy are important

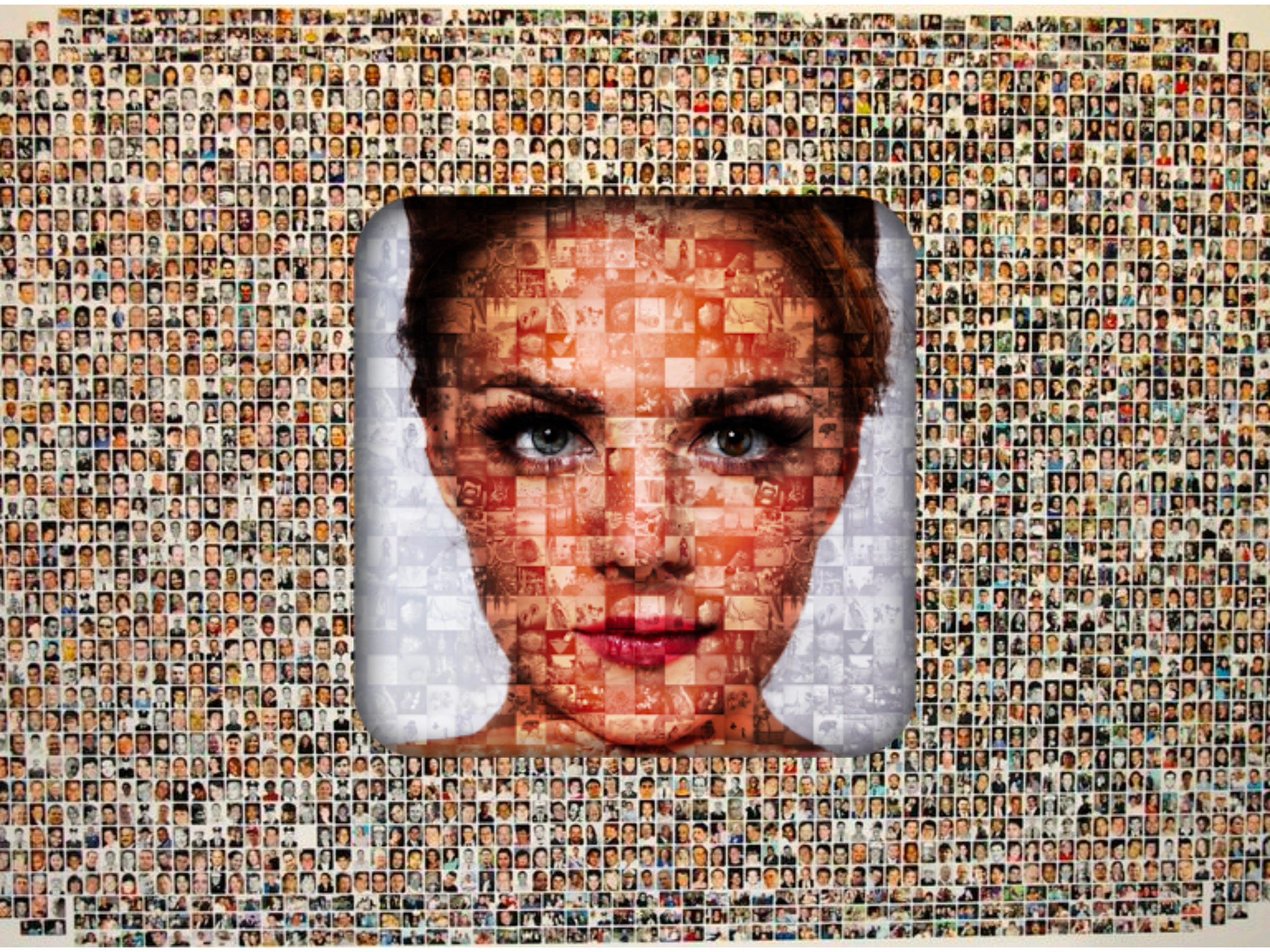
Predict with sophisticated score?

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





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
  - % >=4 stars
  - Full distribution

People who buy **X** also buy **Y**..

Product association recommenders

Market Basket Problem

Example

1	['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups']
2	['tropical fruit', 'yogurt', 'coffee']
3	['whole milk']
4	['pip fruit', 'yogurt', 'cream cheese ', 'meat spreads']
5	['whole milk', 'butter', 'yogurt', 'rice', 'abrasive cleaner']
6	['rolls/buns']
7	['other vegetables', 'UHT-milk', 'rolls/buns', 'bottled beer', 'liquor (appetizer)']
8	….

People who buy X also Y..

- How can we do 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?

Terms

- “IF” part = **Antecedent**
- “THEN” part = **Consequent**

- Item set” = the items (e.g., products) comprising the antecedent or consequent

- **Antecedent** and **consequent** are *disjoint* (i.e., have no items in common)

How to score the association rule?

- Percentage of **X**-buyers who also bought **Y**

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

Hands on time!

Hands on time!

How to score

- Percentage of X-buyers who also bought Y

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

Hands on time!

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

Solving <<***Bananas***>> Problems

How to score

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

X and **Y**

X

!X and **Y**

!X

- 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)

X and **Y**

X

!X and **Y**

!X

- It focuses on how Y increases buying X

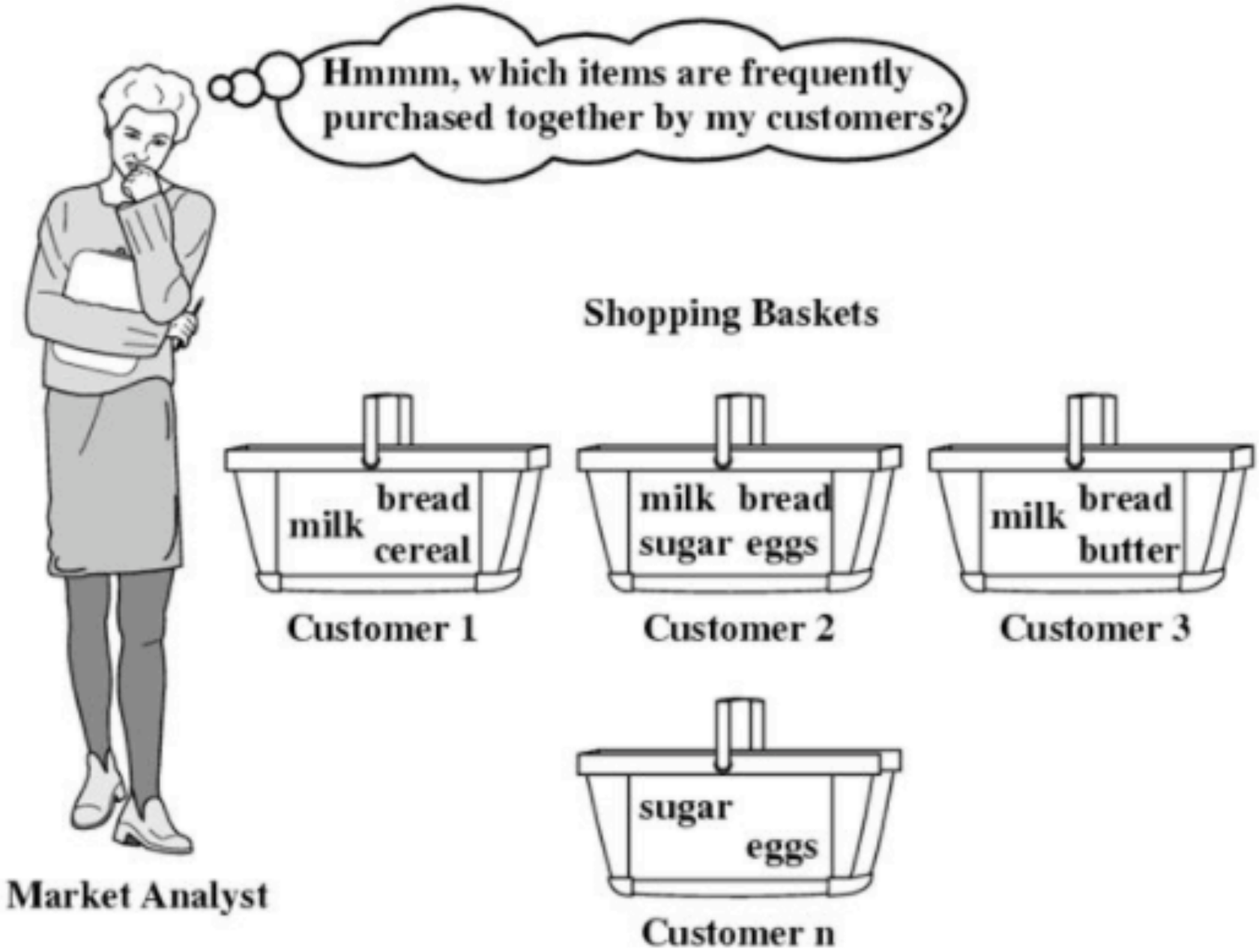
- Value higher than one says that is more probable to be bought X when Y is bought thant not.

Another Association Rule

- Association rule mining

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

Let’s see a much more complicated problem!


Hmmm, which items are frequently purchased together by my customers?
Shopping Baskets
milk bread cereal
Customer 1
milk bread sugar eggs
Customer 2
milk bread butter
Customer 3
Market Analyst
sugar eggs
Customer n

Association Rules

Terminologies:

items : 1, 2, 3,….

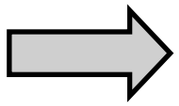
itemSet: {1},{1,2},{1,4,2,5},…

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}

⋮

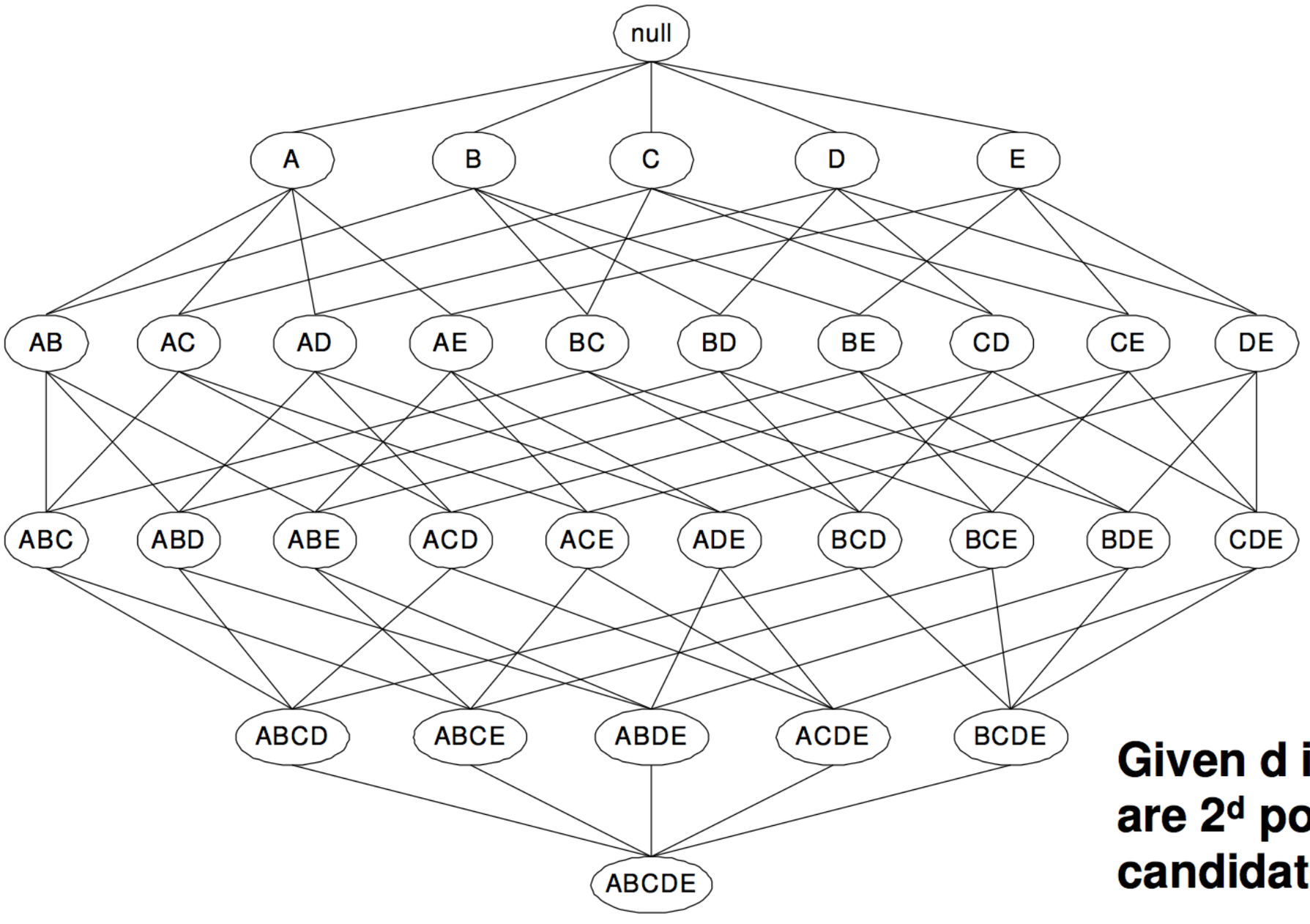
association rule metrics:

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

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

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

Association Rules



Given d items, there are 2^d possible candidate itemsets

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 the supermarket has 100 products

Some interesting numbers

Cardinality of Itemsets	Number of Itemsets
1	10,000
2	49,500,000
3	
4	
5	
6	
7	
8	

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

Frequent Item Sets

- Ideally, we want to create all possible combinations of items

- **Problem**: computation grows exponentially as # items increases… Brute force is not feasible

- **Solution**: Consider only “frequent items sets”

  - Criterion for frequent: **support**

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

Sugar -> **Milk**

Support = 13.5% and *confidence* = 70%

Apriori 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

Association Rules

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

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

support	itemsets
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}
0.5	{1}, {1,3}, {2,3}, {3,5}, {2,3,5}
0.75	{2}, {3}, {5}, {2,5}

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

support	confidence	rules
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

Apriori Method

Apriori(T, ϵ)

$L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}$

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$

for transactions $t \in T$

$C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$

for candidates $c \in C_t$

$count[c] \leftarrow count[c] + 1$

$L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\}$

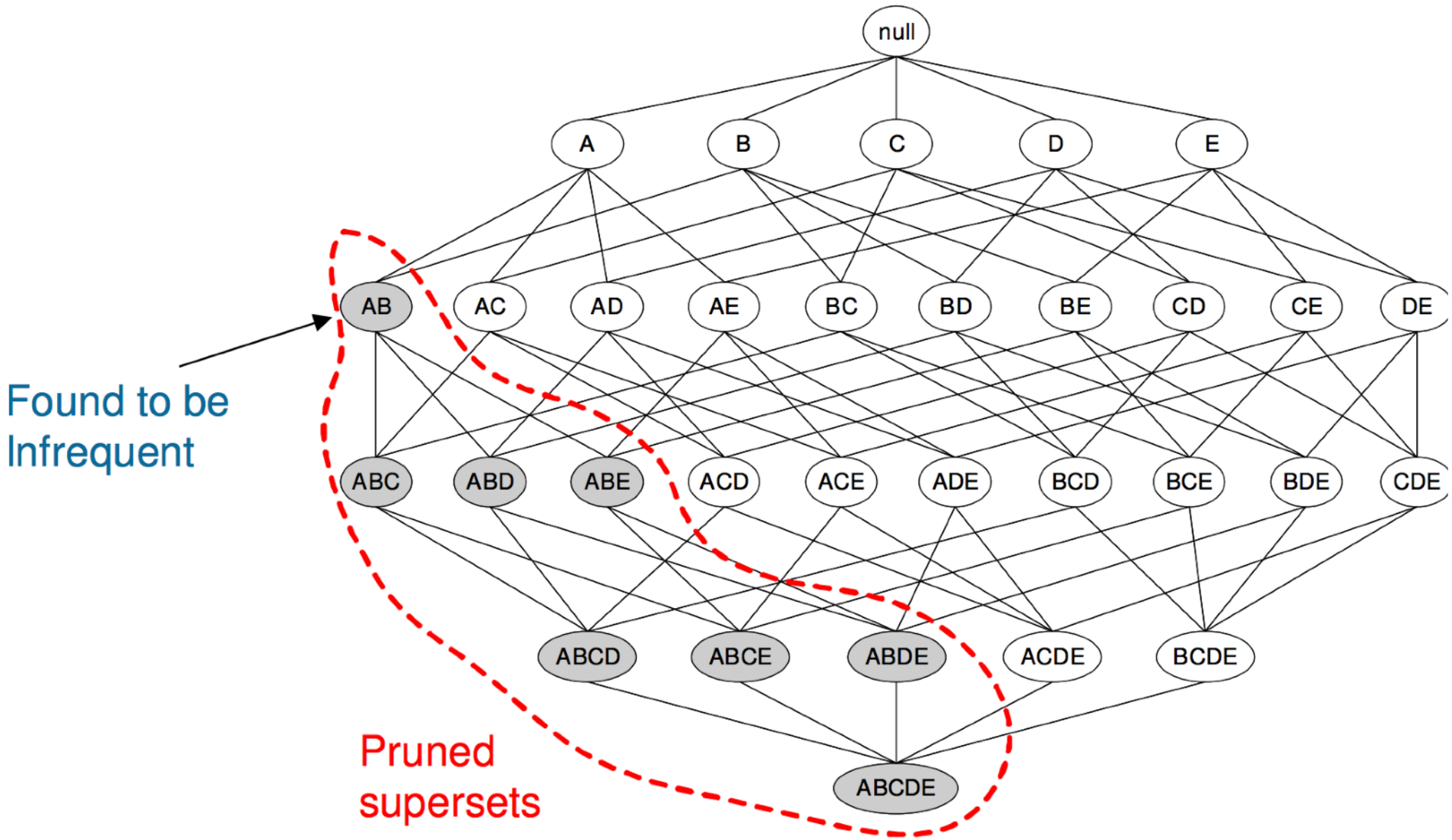
$k \leftarrow k + 1$

return $\bigcup_k L_k$

Fast algorithms for mining association rules
R Agrawal, R Srikant
Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499

21710 1994

Apriori Method



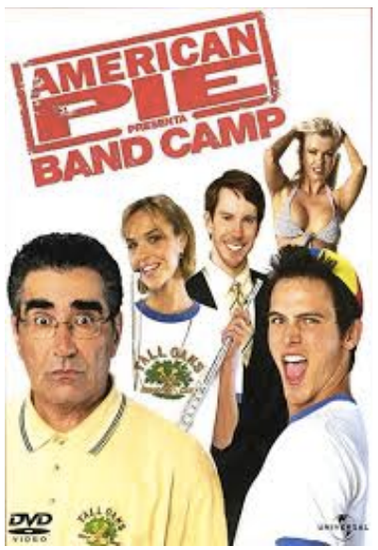
Caution:
The role of Chance

- Random data can generate apparently interesting association rules

- The more rules you produce, the greater this danger

- Rules based on large number of records are less subject to this danger

Association rules for movie recommendations?



Movie Recommendations based on Association Rules?

Hands on time!

Apriori on MovieLens?

market basket will have positive and negative samples.
>4 Like
<=4 Not like

More

When preferences are provided?

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

Quiz:
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