Sistemas de Recomendação não personalizados

SISREC

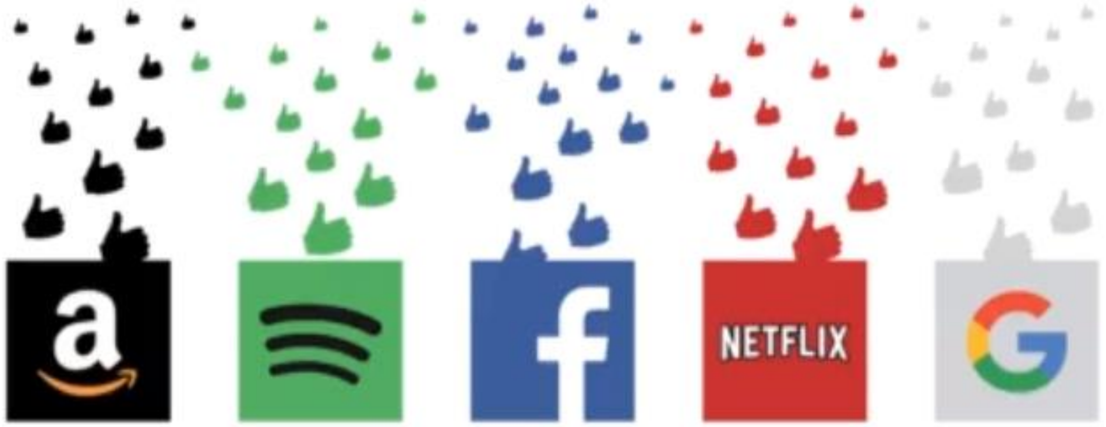
Mestrado em Engenharia Informatica

2023/2024

Joaquim Santos, Catarina Figueiredo, Dulce Mota, Constantino Martins

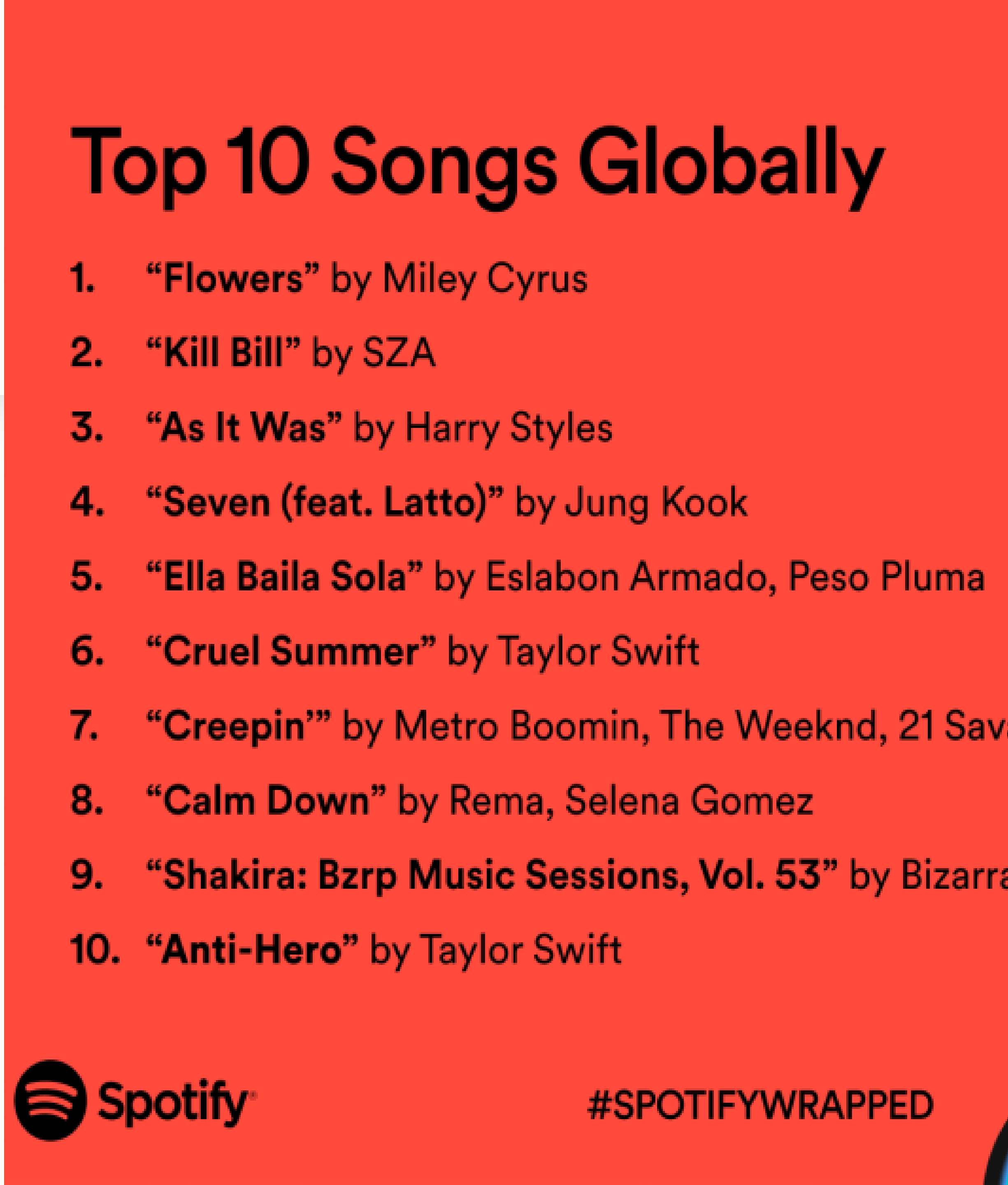
What are non-personalized SR?

- Even if we had never visited a specific website, it's expected that we receive recommendations. These recommendations are non-personalized.
- Use the population behavior of a whole in order to infer what you might like. *“other people like this, therefore you’ll probably like it too”*
- Items recommended to you are recommended



Popularity

- Determine items popularity based on multiple criteria.
  - Ex. Songs: Number of times users listened to the sound per month
  - Ex. Films: Film revenue on movie theatres
- Popular items will probably be appreciated by most people.
- So, if we present popular items, the probability of someone liking it will be higher.
- **Example**: Find the top 10 songs played on Spotify, present those 10 songs to new users.



Top 10 Songs Globally

1. **“Flowers”** by Miley Cyrus
2. **“Kill Bill”** by SZA
3. **“As It Was”** by Harry Styles
4. **“Seven (feat. Latto)”** by Jung Kook
5. **“Ella Baila Sola”** by Eslabon Armado, Peso Pluma
6. **“Cruel Summer”** by Taylor Swift
7. **“Creepin’”** by Metro Boomin, The Weeknd, 21 Sav
8. **“Calm Down”** by Rema, Selena Gomez
9. **“Shakira: Bzrp Music Sessions, Vol. 53”** by Bizarra
10. **“Anti-Hero”** by Taylor Swift

Spotify

#SPOTIFYWRAPPED

Movie recommender

- **Step 1**: Find the best-selling/most-watched movies.
- **Step 2**: Filter out the ones the user has seen and present the remaining movies.
- What if in the Top 5 of best-selling movies there is a horror movie?
- Will this be a good suggestion for a child or for someone that dislikes horror movies? **No**
- What can we do? Only use those recommenders when we still don't have any information regarding the user.

Popularity-based Problems

- Using the behavior of the population may be too broad.
- Not everyone will like the recommendations provided by this type of recommendation system.

Why?

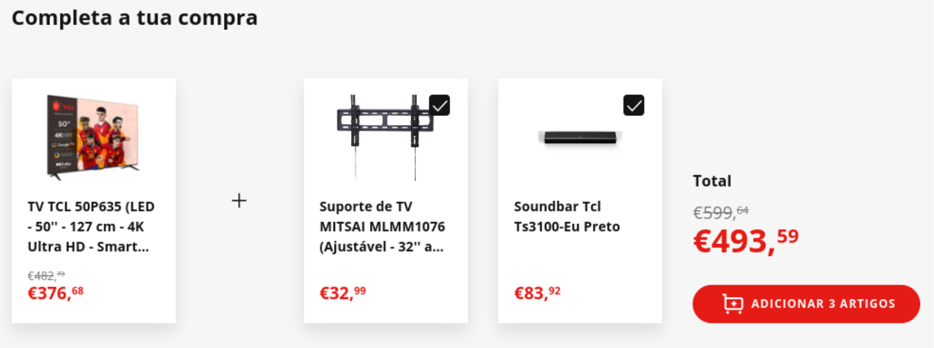
Because instead of using personal preferences, we are using the preferences of the whole population.

Simple solution yet not the best one for most use cases.

worten fnac amazon

Associations

- Being aware of products that are bought together.
- How can we find product associations?
    - Apriori algorithm
- If Buying B makes it more likely to buy A, then Lift > 1.
- If p(A|B) = 20% and p(A) = 10%
- Lift = 2



Completa a tua compra

TV TCL 50P635 (LED - 50'' - 127 cm - 4K Ultra HD - Smart...
~~€482,^{79}~~
€376,^{68}

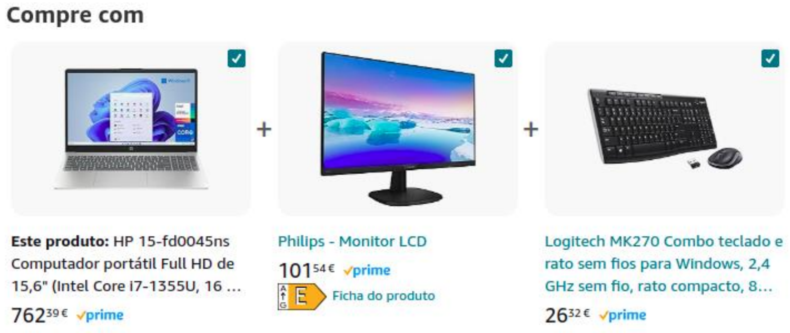
+

☑ Suporte de TV MITSAI MLMM1076 (Ajustável - 32'' a...
€32,^{99}

☑ Soundbar Tcl Ts3100-Eu Preto
€83,^{92}

Total
~~€599,^{64}~~
€493,^{59}

ADICIONAR 3 ARTIGOS



Compre com

☑ **Este produto:** HP 15-fd0045ns Computador portátil Full HD de 15,6" (Intel Core i7-1355U, 16 ...
762^{39} € prime

+

☑ Philips - Monitor LCD
101^{54} € prime
E Ficha do produto

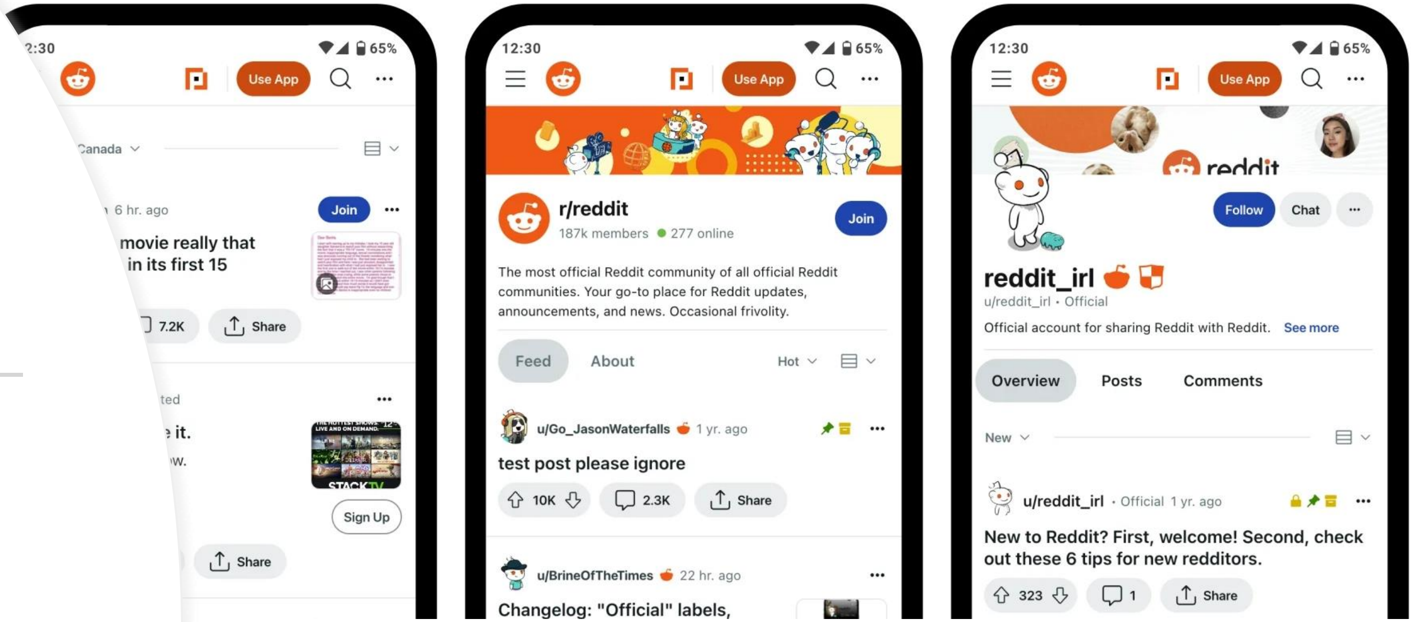
+

☑ Logitech MK270 Combo teclado e rato sem fios para Windows, 2,4 GHz sem fio, rato compacto, 8...
26^{32} € prime

$$Lift = \frac{p(A,B)}{p(A)p(B)} = \frac{p(A\,|\,B)}{p(A)} = \frac{p(B\,|\,A)}{p(B)}$$

News Feeds

- Normally up votes and down votes.
- Use popularity.
- Must consider recency.
- We should not present the most popular article from 3 years ago on the front page, recency should be considered to prevent this behavior.
- Higher the popularity, the higher the ranking, but the higher the age, the lower the ranking.



The most official Reddit community of all official Reddit communities. Your go-to place for Reddit updates, announcements, and news. Occasional frivolity.

test post please ignore

Changelog: "Official" labels,

Official account for sharing Reddit with Reddit.

New to Reddit? First, welcome! Second, check out these 6 tips for new redditors.

$$\frac{f(popularity)}{g(age)}$$

Challenges

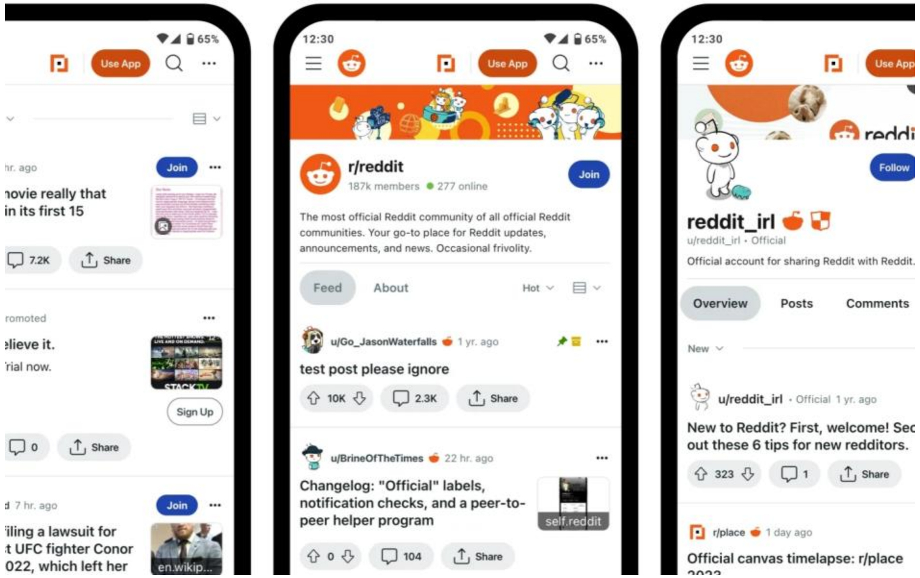
- How to deal with hate and conspiracy posts?
- We should avoid giving too much importance to those posts.

Hacker News Formula:

$$score = \frac{(ups - downs - 1)^{0.8}}{(age+2)^{gravity}} \times penalty$$

gravity = 1.8

penalty = multiplier to implement "business rules" (e.g. penalize self-posts, "controversial" posts, etc...)



novie really that
in its first 15

elieve it.
rial now.

iling a lawsuit for
t UFC fighter Conor
022, which left her

r/reddit
187k members • 277 online
The most official Reddit community of all official Reddit communities. Your go-to place for Reddit updates, announcements, and news. Occasional frivolity.

u/Go_JasonWaterfalls • 1 yr. ago
test post please ignore

u/BrineOfTheTimes • 22 hr. ago
Changelog: "Official" labels, notification checks, and a peer-to-peer helper program

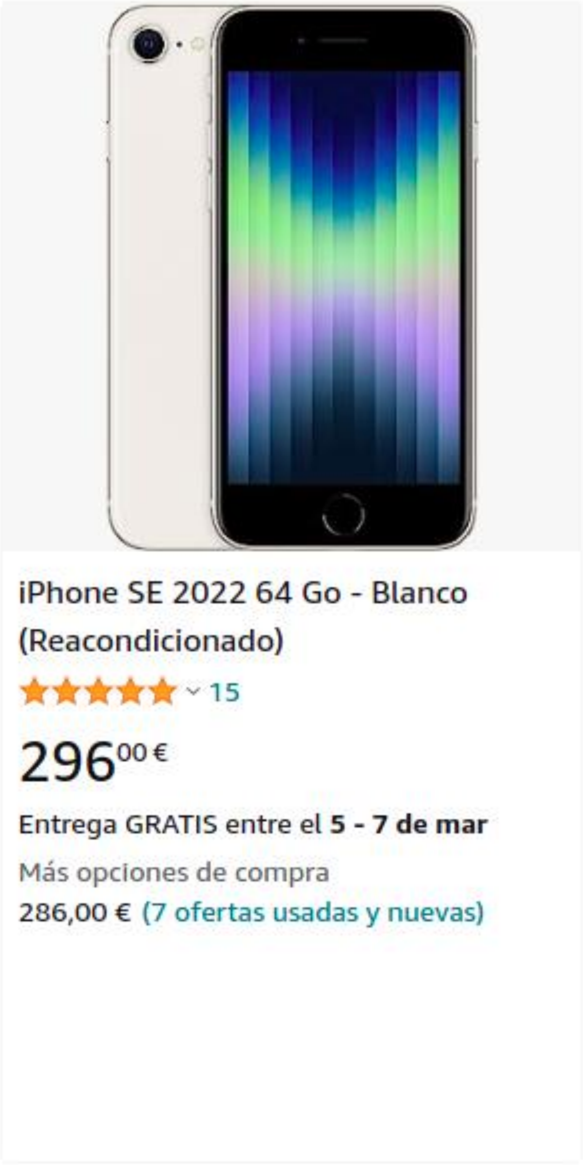
reddit_irl
u/reddit_irl • Official
Official account for sharing Reddit with Reddit.

u/reddit_irl • Official 1 yr. ago
New to Reddit? First, welcome! Sec
out these 6 tips for new redditors.

r/place • 1 day ago
Official canvas timelapse: r/place

Ratings

- 5 stars ratings systems: Amazon, Temu, etc.
- Sort the items by score.
- Simple approach: sort by average rating
- Is this a good approach?
- Which item is the best? Left, or right?
- The item with higher average rating has only 15 reviews while the other has a lot more but slightly lower average rating.



iPhone SE 2022 64 Go - Blanco (Reacondicionado)

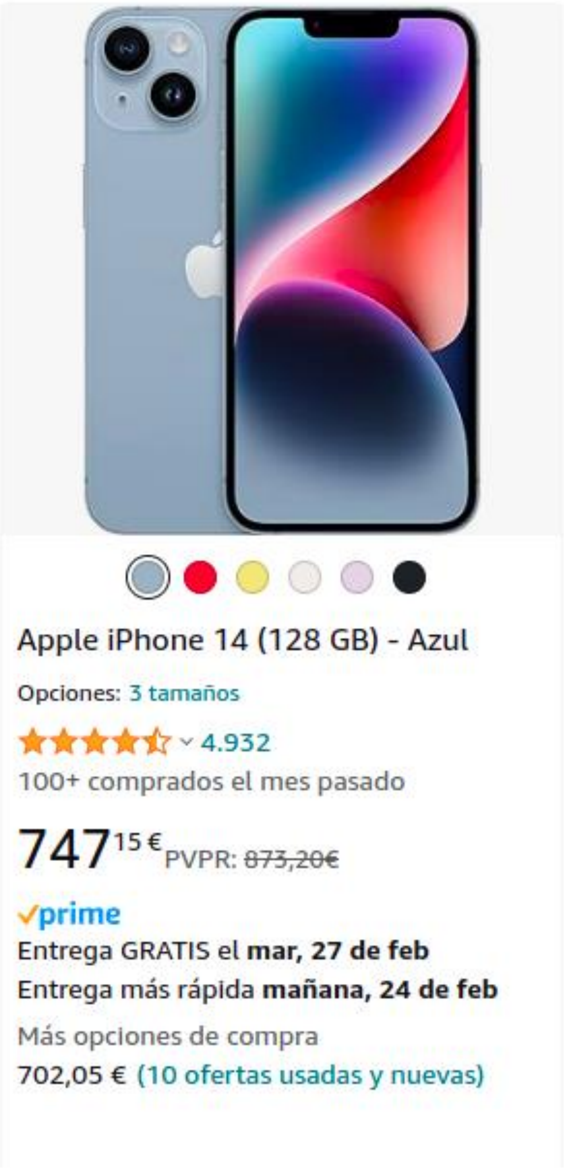
★★★★★ ⌄ 15

296^{00} €

Entrega GRATIS entre el **5 - 7 de mar**

Más opciones de compra

286,00 € (7 ofertas usadas y nuevas)



● ● ● ● ● ●

Apple iPhone 14 (128 GB) - Azul

Opciones: 3 tamaños

★★★★☆ ⌄ 4.932

100+ comprados el mes pasado

747^{15} € PVPR: ~~873,20€~~

✓prime

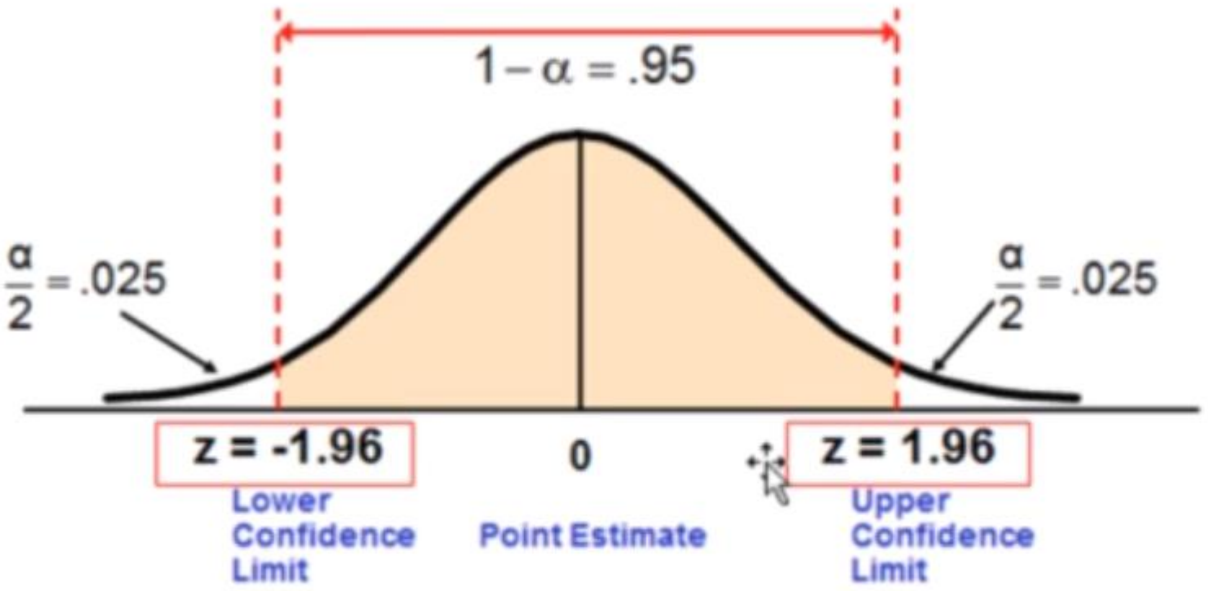
Entrega GRATIS el **mar, 27 de feb**

Entrega más rápida **mañana, 24 de feb**

Más opciones de compra

702,05 € (10 ofertas usadas y nuevas)

Average Rating

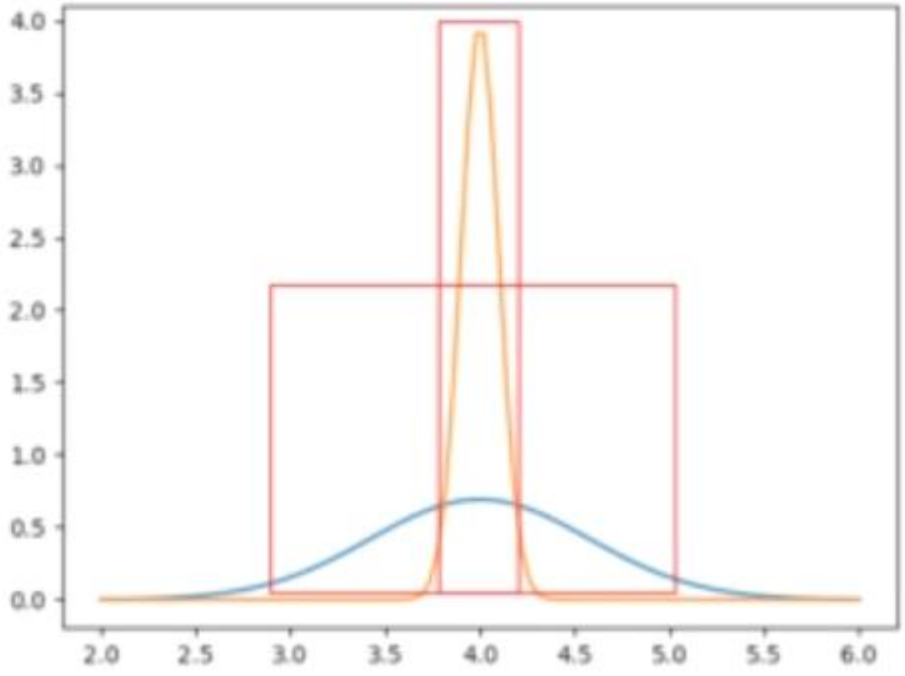


- **Problem**: How confidence are we?
- We should use confidence intervals.

$$\bar{X} = \frac{1}{N} \sum^{N} X_i$$

- More samples -> smaller variance = skinnier intervals
- As the number of ratings of a certain item grows, the more confident we are in the estimate of the average.

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

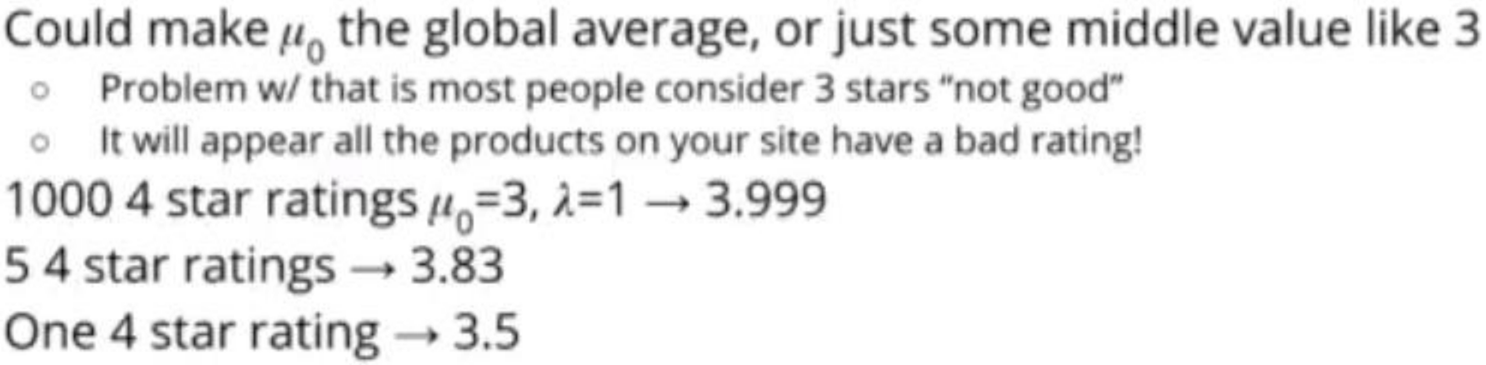


Problems with Average Rating

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$r = \frac{\sum_{i=1}^{N} X_i + \lambda\mu_0}{N+\lambda}$$

- What if N is very small or 0? (New item with no ratings)
- Add **Smoothing** by adding a small number to the numerator and denominator.



Could make μ_0 the global average, or just some middle value like 3

- Problem w/ that is most people consider 3 stars "not good"
- It will appear all the products on your site have a bad rating!

1000 4 star ratings μ_0=3, λ=1 → 3.999

5 4 star ratings → 3.83

One 4 star rating → 3.5

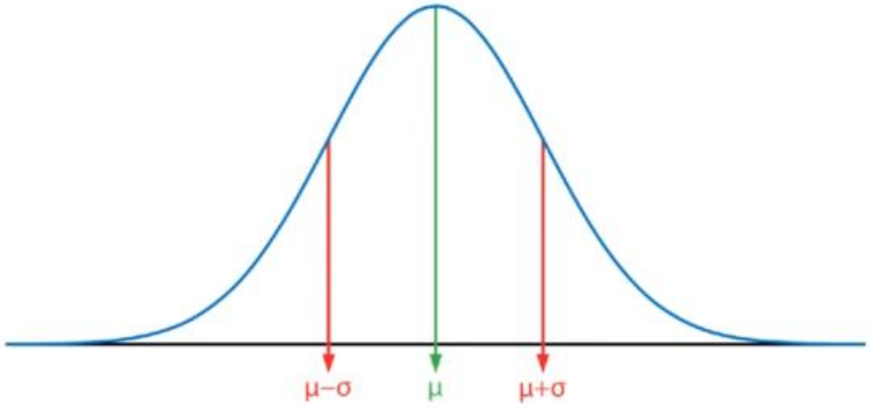
Explore – exploit Dilemma

- **Exploration** refers to trying out new options to gather more information about their potential rewards or outcomes. This involves taking risks and potentially sacrificing immediate gains for the sake of long-term learning and improvement.
- **Exploitation**, on the other hand, involves maximizing immediate rewards by choosing options that are already known to be effective based on existing knowledge or experience. This typically involves sticking with what's familiar and has shown positive results in the past.
- The dilemma arises because there's often a trade-off between exploration and exploitation. Too much exploration can lead to missing out on exploiting known good options, while too much exploitation can lead to stagnation and missing out on potentially better alternatives. Striking the right balance between exploration and exploitation is crucial for achieving optimal outcomes in various domains.

Explore – exploit Dilemma
Practical example

- We start watching a lot of videos on Youtube about how to cook. Youtube: exploits this fact and show tons of videos about cooking. After watching a ton of videos about cooking, we may be never suggested anything else. **Lots of exploiting**, but **no exploring.**
- After mastering cooking, we probably don't want to watch videos about cooking anymore. There should be a stronger exploration component. We might want to watch videos on machine learning which has nothing to do with cooking. Youtube could just as easily suggest something that we don't care about, like knitting. So, by exploring, you're not guaranteed to find something we like.
- It makes bad products appear better than they are, and it makes good products appear worse than they are. Until we have enough confidence for them to converge to their true values. Essentially what it leads to is a mix of probably good and probably bad products. Of course, the top recommendations will still be the very good products that have high confidence in.

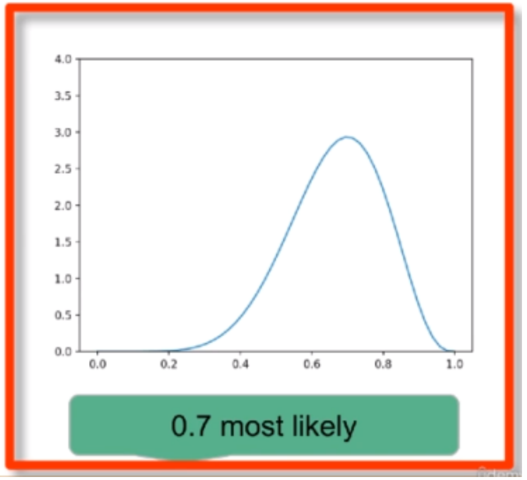
Bayesian Ranking

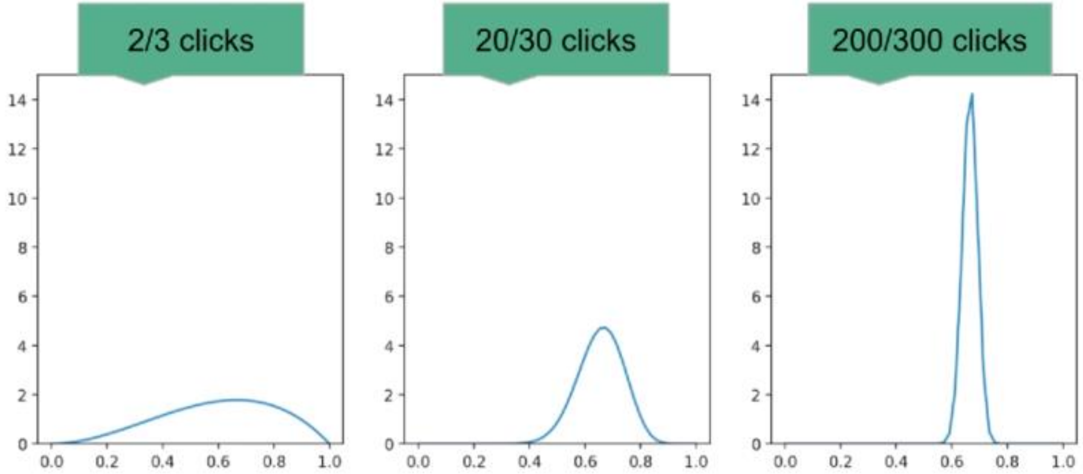


- Instead of fixed scores uses variable scores to rate and sort items.
- Pick random scores for every item.
- Random != "completely disordered"
- The random variable is still characterized by its distribution.
- The distribution tell us what values are more likely and what values are less likely.

Bayesian Ranking – Click-Through Rate (CTR)

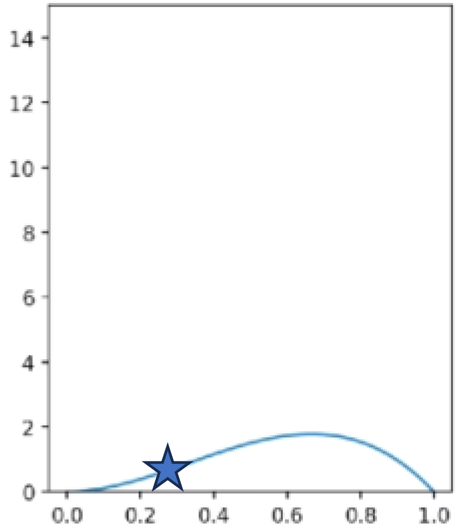
- Rank product based on which are more likely to be clicked -> Sort items by CTR.
- How to calculate CTR? Click=1 and View = 0 CTR = n_clicks/(n_clicks + n_views)
- Treat CTR not as a fixed number but as a random number (i.e. one that has a distribution)



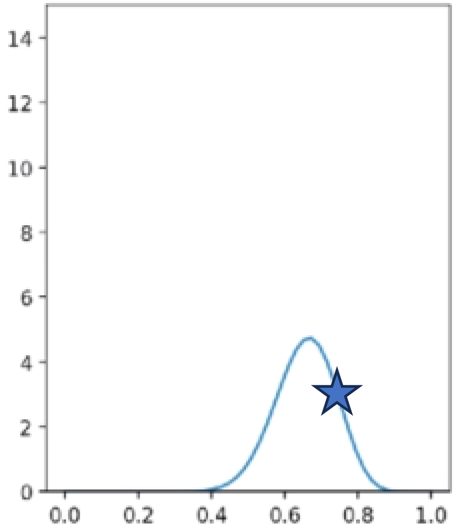


Bayesian Ranking – How do we rank the items?

- If we have fixed scores, items with higher scores on top.
- But if we have two distributions? Which item should appear on top? <u>Solution</u>: Sampling random numbers (Thompson Sampling)



Item A



Item B

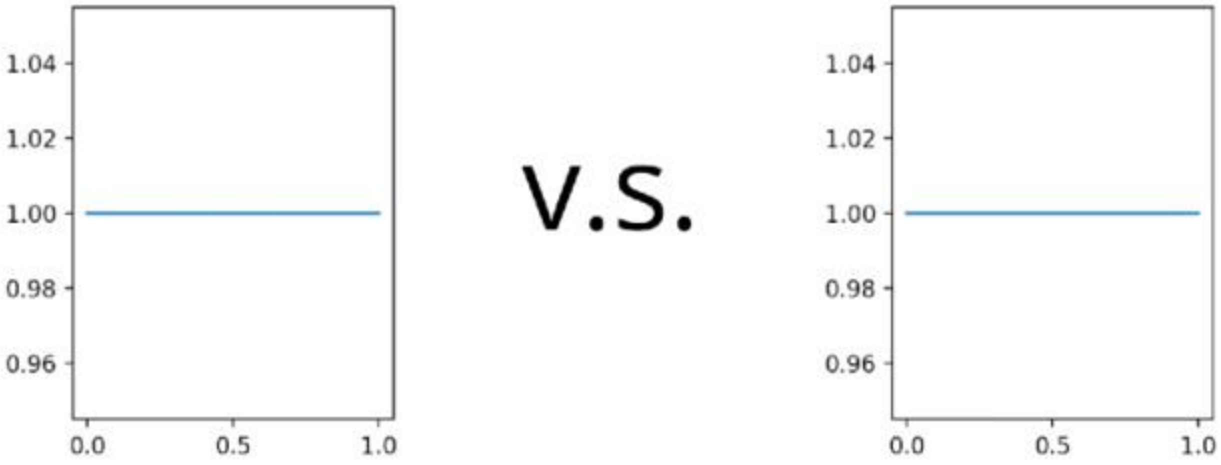
Ranked Items

1º Item B

2º Item A

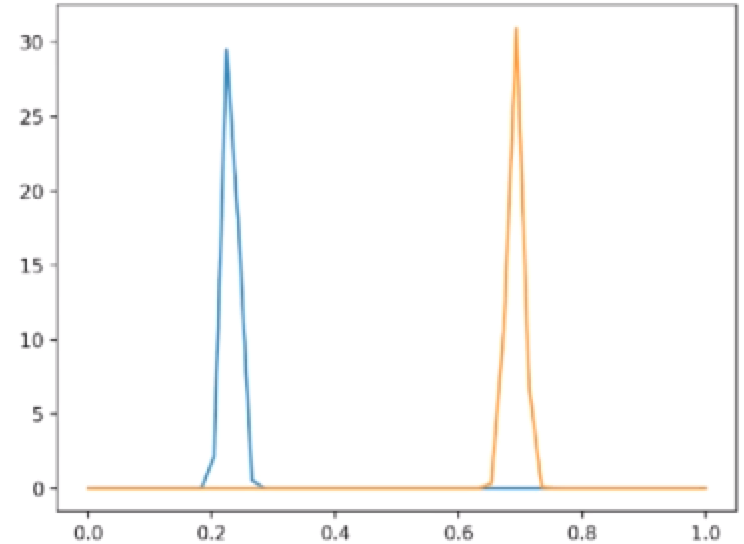
Bayesian Ranking – Extreme case #1 – Both Uniform

- Both samples are equally likely to be any number between 0 and 1.
- We need to show the items to users (explore and collect more data)



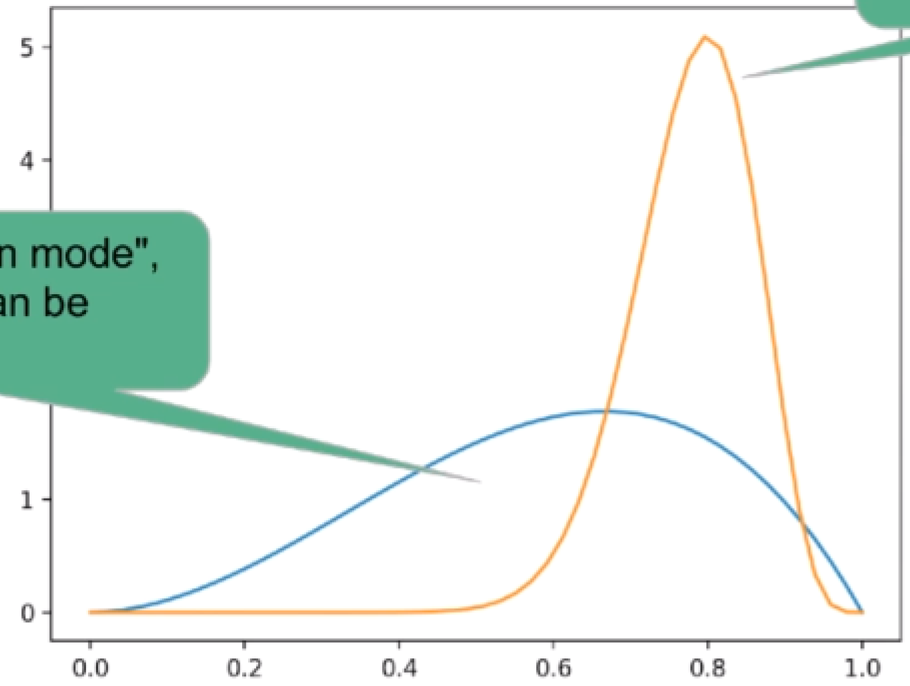
Bayesian Ranking – Extreme case #2 – Both Sharp Peaks

- Peak is very sharp, the sample will be very close to the peak.
- Very likely that the item with the highest-value peak wins.
- But if we have two distributions? Solution: Sampling random numbers (Thompson Sampling)
- We're collected so much data that we're very confident of the CTR



Bayesian Ranking – Mixed Case

- If we have fixed scores, items with higher scores



More likely to have a score (sample) near the peak

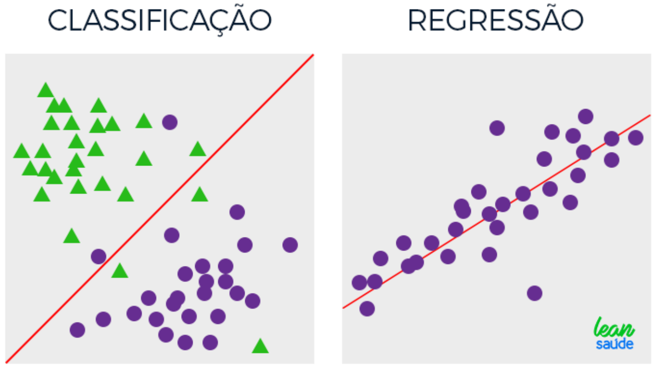
Still in "exploration mode", score (sample) can be high or low

Bayesian Ranking

- Bayesian method automatically balances need to explore and exploit.
- 2 fat distributions: explore both (totally random ranking)
- 2 skinny distributions: exploit both (nearly deterministic ranking)
- Mixed: explore and exploit co-exist

Supervised Machine learning

- We have some inputs (X) and corresponding targets(Y)
- Y might represent:
  - Did the user buy the product?
  - Click on the ad?
  - Click on the article?
  - Make an account?
  - What did the user rate this item?



- If the model predictions are accurate, then we can use it to recommend items the user is more likely to buy/click or rate highly.

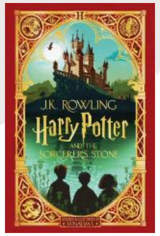
Supervised Machine learning

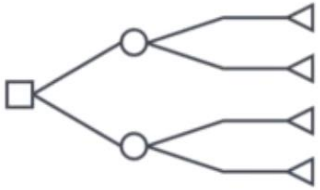
Input features

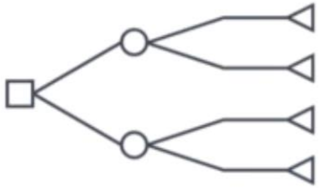
- Common features include demographics:
  - Age
  - Gender
  - Religion
  - Occupation
  - Education level
  - Etc.
- Data collected by the site
  - Purchase history, pages they viewed. Etc.

Supervised Machine learning

Problems







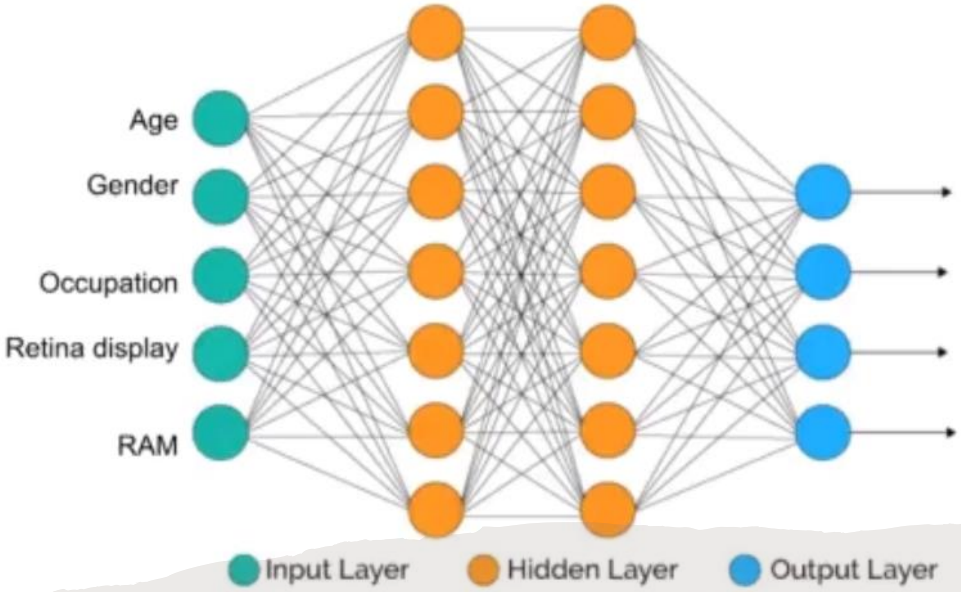
Those models take in consideration the user but now the item itself.

<u>**Solution:**</u>

Separate models for each product, **NOT** scalable.

Better Solution:

Add item attributes to the input features. So given user ***x*** demographics and item ***y*** attributes we will predict is the user will buy the product or not or the rating the use x will give to item y.



Supervised Machine learning
Getting Data

- Not easy.
- We can buy it, Not cheap!
- Privacy – ad and tracking blockers.
- Product data - dependent on vendor entering data correctly
- If free-form, lots of string parsing needed