Рекомендательные системы

Андрей Зимовнов

Яндекс, ВШЭ

Search vs Discovery

"Говорят, что Интернет покидает эпоху поиска и входит в эпоху открытий. В чем разница? Поиск - это когда вы ищете что-то. Открытие - это когда что-то замечательное, о существовании которого вы не знали, находит вас."

CNN Money.

Постановка задачи

Рекомендации нужны, когда:

- пользователь не знает точно, что он хочет (I know it when I see it)
- число доступных объектов превышает возможности человека их обозреть
- критерии поиска не алгоритмизируются (вкус)

Примеры

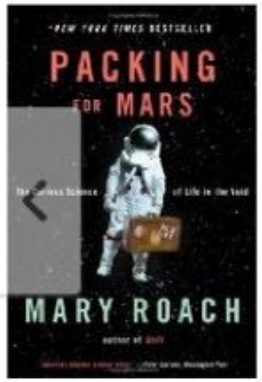
Торговля

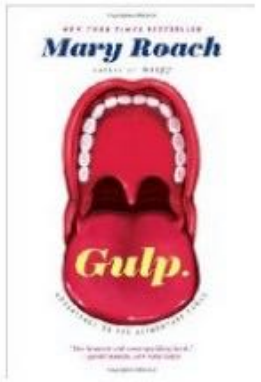
Amazon, Юлмарт, ozon.ru

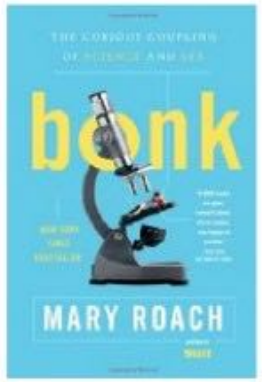
www.amazon.com

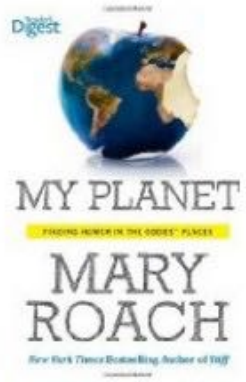
Доставка Корзина

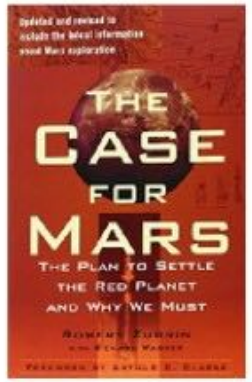
Related to Items You've Viewed See more





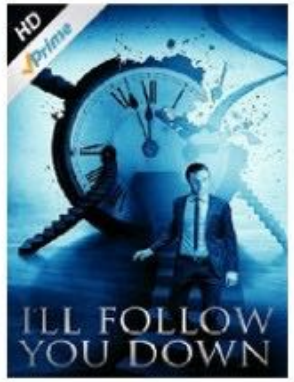


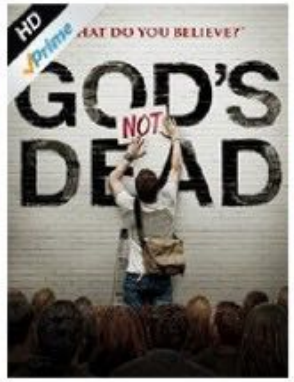


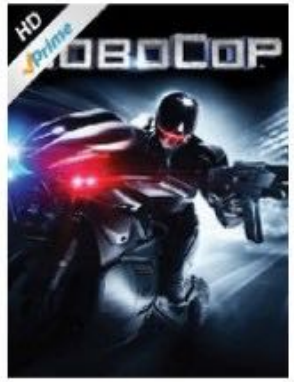


Ad feedback

Movies Included with Prime Membership at No Additional Cost See more





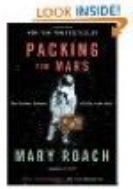


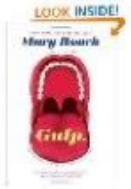
Mother's Day is May 10

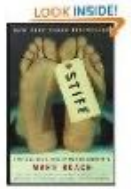
Add to Wish List

Have one to sell? Sell on Amazon

Frequently Bought Together



+


+


Price for all three: $30.62

Add all three to Cart

Add all three to Wish List

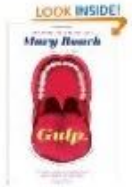
Show availability and shipping details

- ☑ **This item:** Packing for Mars: The Curious Science of Life in the Void by Mary Roach Hardcover $10.38
- ☑ Gulp: Adventures on the Alimentary Canal by Mary Roach Paperback $10.34
- ☑ Stiff: The Curious Lives of Human Cadavers by Mary Roach Paperback $9.90

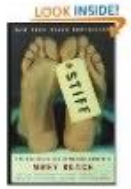
Ad feedback

Customers Who Bought This Item Also Bought

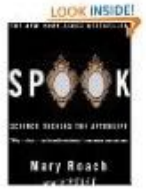
Page 1 of 19



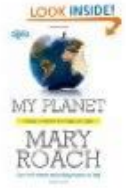
Gulp: Adventures on the Alimentary Canal
› Mary Roach
4.5 out of 5 stars 867
Paperback
$10.34 Prime



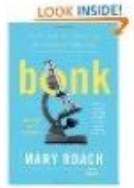
Stiff: The Curious Lives of Human Cadavers
› Mary Roach
4.5 out of 5 stars 1,136
#1 Best Seller in Forensic Medicine
Paperback
$9.90 Prime



Spook: Science Tackles the Afterlife
› Mary Roach
3.5 out of 5 stars 293
Paperback
$13.02 Prime



My Planet: Finding Humor in the Oddest Places
› Mary Roach
4.5 out of 5 stars 151
Paperback
$13.49 Prime



Bonk: The Curious Coupling of Science and...
› Mary Roach
4.5 out of 5 stars 432
Paperback
$13.55 Prime

Тип SIM-карты	Nano-SIM
Количество поддерживаемых SIM-карт	1
Цвет	Черный
Операционная система	Android
Связь 3G	Есть
4G (LTE)	Есть
GPS-модуль	Есть
Сенсорный экран	Ёмкостный
Тип поддерживаемых карт памяти	Нет

Подробные характеристики

Похожие товары

Смартфон Apple iPhone 6 Plus 64Gb Space Gray, MGAH2RU/A

57 490 руб.

Еще похожие товары

Лучшие аксессуары

Рекомендуем | Bluetooth и проводные гарнитуры | Авто и вело держатели | Автомобильные держатели | Аккумуляторы | Аккумуляторы и зарядные устройства | Аксессуары для смартфонов, телефонов | Зарядные устройства | Защитные стекла и пленки | Кабели синхронизации | Чехлы и сумки для смартфонов, телефонов

Вы выбрали 1 товар на сумму

54990 руб.

Купить

Отметьте интересующие

Музыка

Яндекс.Музыка, Apple Music

https://music.yandex.ru

Яндекс | Музыка | Трек, альбом, исполнитель | Найти | Рекомендации | Жанры | Подборки | Моя музыка | yndx.cheshire

Сегодня, 7 мая

3 часа 42 минуты музыки — послушать рекомендованные сегодня треки

Если вам по душе русский рок, обратите внимание на эти треки

Её глаза (из Шекспира) Би-2	4:52
Весна Сурганова и Оркестр	4:14
Лучшая песня о любви Високосный Год	2:52
Любовь и боль Сергей Шнуров	2:18
Останусь Город 312	3:44
Родина АнимациЯ	2:57
На берегу безымянной реки	4:14

Яндекс.Директ

Спасибо. Объявление скрыто.

Спасибо. Объявление скрыто.

Появились новые доклады, тезисы!

21 и 22 мая 2015 в Москве состоится IT-Фестиваль РИТ ++! Примите участие!

fpconf.ru Адрес и телефон

Руль Logitech Driving Force GT!

Руль Logitech Driving Force GT - 7577 руб.

Моя звезда

Вячеслав Бутусов & Deadушки

https://music.yandex.ru

Не остановиться
Вячеслав Бутусов
✓ 3:56

Похоже, вам нравится Ёлка, попробуйте послушать и эти треки

Знаки вопроса
Ёлка
3:13

Цепи-ленты
Ёлка
3:55

Город Обмана
Ёлка
3:41

Я в печали
Ёлка
3:54

Судя по вашим предпочтениям, Джанго вам понравится

Была не была
Джанго
3:53

До тебя
Джанго
3:22

Венгерка
Джанго
3:37

Спасибо. Объявление скрыто.


Моя звезда
Вячеслав Бутусов & Deadушки

Социальные сети

VK, Facebook, …

Search Facebook

Grigory | Home

People You May Know

Katrin Kras
Александр Баранов is a mutual friend.

Add Friend

Yulia Horoshaya (Yulia)

Add Friend

Nikita Pustovoytov
Alex Dainiak and 5 other mutual friends

Add Friend

Dina Rubinshtein
Valeriy Platonov and 3 other mutual friends

Add Friend

Di Guseva

Add Friend

Дионисий Невероятный
Александр Баранов and Svetlana Pavlova are mutual friends.

Add Friend

Туризм

booking.com, …

Берген — популярные отели


[Piano Hostel](#)
Последнее бронирование: 2 часа назад


[P-Hotels Bergen](#) ★★★
Последнее бронирование: 15 минут назад
Сейчас 1 пользователь просматривает этот отель.


[Scandic Byparken (formerly Rica Hotel Bergen)](#) ★★★★
Последнее бронирование: 47 минут назад
Сейчас 1 пользователь просматривает этот отель.


[Det Hanseatiske Hotel](#) ★★★★
Последнее бронирование: 1 час назад


[Best Western Hotel Hordaheimen](#) ★★★
Последнее бронирование: 8 часов назад
Сейчас 1 пользователь просматривает этот отель.

Часто задаваемые вопросы

[Типы номеров](#)
[Цены](#)
[Оплата](#)

Отель Grand Terminus находится в Бергене, в 10 минутах ходьбы от площади Торгалменнинген. К услугам гостей круглосуточная стойка регистрации и терраса в саду. Классические номера располагают спутниковым телевидением, бесплатным WiFi и мини-баром.

Каждое утро в отеле Grand Terminus подают завтрак «шведский стол». В виски-баре Terminus можно заказать различные напитки и закуски по барному меню. Кофе и чай предоставляются круглосуточно.

В отеле к вашим услугам оздоровительный центр с тренажерным залом и сауна. Гости отеля могут бесплатно арендовать велосипед для осмотра Бергена.

Гавань Брюгген, объект Всемирного наследия ЮНЕСКО, расположена в 10 минутах ходьбы от отеля. Фуникулер Флейбанен, на котором можно легко подняться на гору Флёйен, находится в 850 метрах.

Номеров в отеле: 131, Сеть отелей: Historical Hotels. На Booking.com с 3 февр. 2010.

Особенности объекта размещения

Последнее бронирование: 1 час назад

Бесплатный Wi-Fi

Ориентиры:

Зал Хакона и башня Розенкранц (100 m)

Музей проказы (150 m)

Зафиксируйте отличную цену для своей предстоящей поездки

Получите мгновенное подтверждение бронирования и возможность БЕСПЛАТНОЙ отмены для большинства номеров!

Наличие мест

Grand Terminus — Когда вы желаете здесь остановиться?

Дата заезда День ▾ Месяц ▾

Дата отъезда День ▾ Месяц ▾

Гарантия лучшей цены

Проверить наличие мест

www.booking.com Booking.com: Отель Grand Terminus - Берген, Норвегия

Посетители, просмотревшие этот вариант (Grand Terminus), также заинтересовались следующими:

Scandic Ørnen ★★★★

Открытый в мае 2014 года отель Scandic Ørnen расположен рядом с железнодорожным вокзалом Бергена. К услугам гостей современные номера с бесплатным WiFi и телевизором с плоским экраном.

Оценка по **911** отзывам

Очень хорошо 8,4/10

Забронировать

Radisson Blu Royal Hotel, Bergen ★★★★

Отель Radisson Blu Royal, Bergen расположен возле пристани Брюгген, которая внесена в список всемирного наследия ЮНЕСКО, а также в 300 метрах от рыбного рынка.

Последнее бронирование в данном отеле произведено сегодня в 12:47

Оценка по **1368** отзывам

Очень хорошо 8,3/10

Забронировать

Augustin Hotel ★★★

Отель Augustin, украшенный оригинальными произведениями норвежских художников и скульпторов, находится в центральной части города Берген в 5 минутах ходьбы от рыбного рынка.

Последнее бронирование в данном отеле произведено сегодня в 13:54

Оценка по **777** отзывам

Очень хорошо 8/10

Забронировать

Thon Hotel Bergen Brygge ★★★

Отель находится возле верфи Брюгген, внесенной в список Всемирного наследия ЮНЕСКО. Гостей приглашают остановиться в номерах с хорошим соотношением цены и качества в центре Бергена.

Последнее бронирование в данном отеле произведено сегодня в 12:32

Оценка по **1690** отзывам

Хорошо 7,2/10

Забронировать

Посмотреть другие отели в этом городе | Посмотреть другие отели класса люкс в этом город | Посмотреть другие романтические отели в городе Берген | Посмотреть другие шале в этом городе | Посмотреть другие семейные отели в городе Берген | Посмотреть другие бизнес-отели в городе Берген

Grand Terminus — самые последние отзывы

Посмотреть все отзывы гостей: Grand Terminus

5 августа 2014 г.

Nina

10 Великолепно

Видео

TED, YouTube, …

Similar topics TED Conference | Children | Creativity | Culture | Dance | Education | Parenting

This talk was presented at an official TED conference, and was featured by our editors on the home page.

The creative spark
Playlist (6 talks)

Re-imagining school
Playlist (13 talks)

11 must-see TED Talks
Playlist (11 talks)

Related talks

Gever Tulley
5 dangerous things you should let your kids do

Ken Robinson
Bring on the learning revolution!

Ken Robinson
How to escape education's death valley

Bill T. Jones
The dancer, the singer, the cellist ... and a moment of creative magic

Alice Goffman
How we're priming some kids for college — and others for prison

Clint Smith
How to raise a black son in America

Learn more

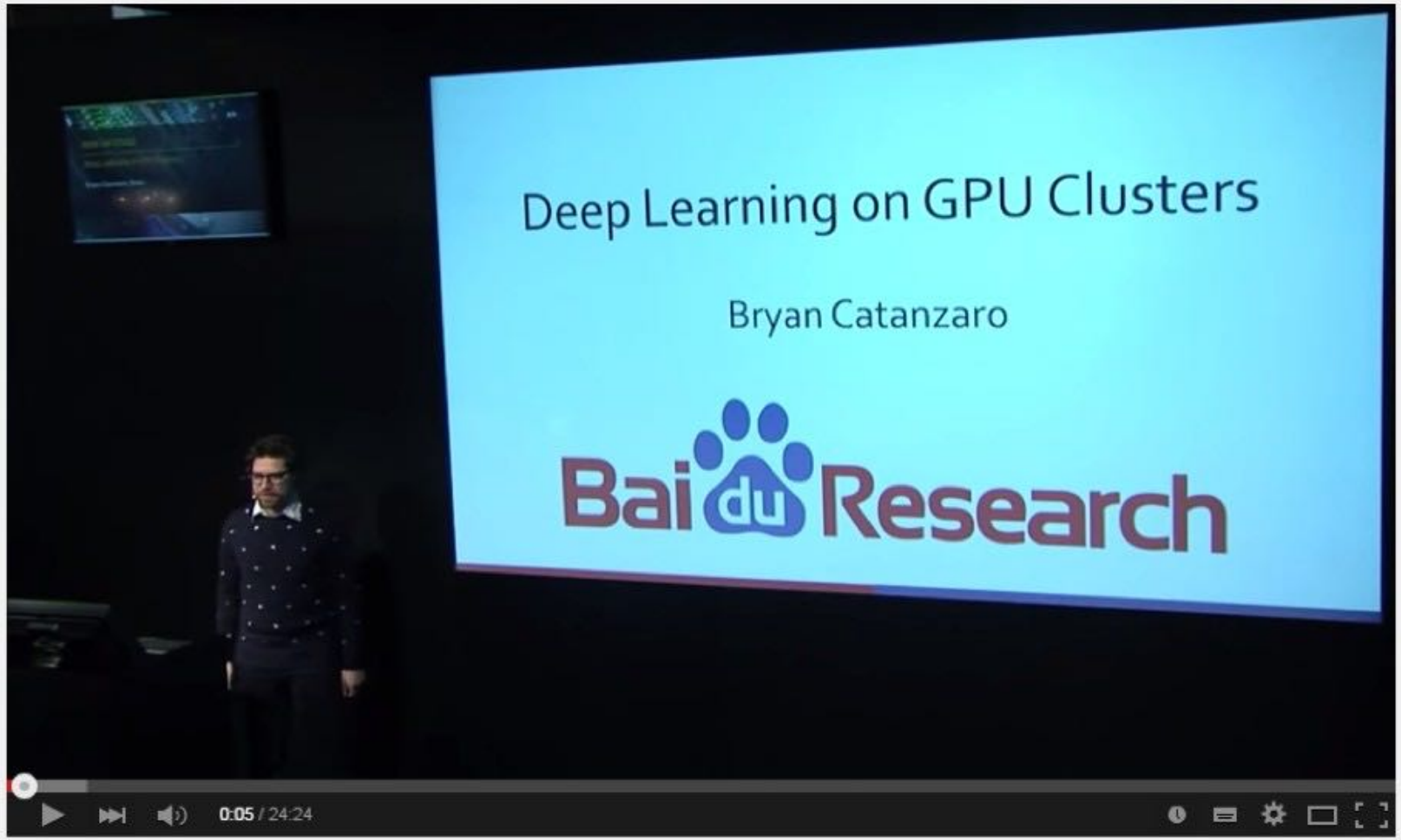
Out of Our Minds

Dance: Just as important as mathematics

www.youtube.com/watch?v=brui4N2orlI Копировать Поделиться

YouTube RU

Добавить видео



Deep Learning on GPU Clusters

RichReport

Следующее Автовоспроизведение

Taming Latency Variability and Scaling Deep Learning, by Jeff Dean, Google,
sfbayacm
6 281 просмотр

Large-Scale Deep Learning for Building Intelligent Computer Systems
UWTV
491 просмотр

Deep Learning through Examples Screencast with Audio 9/11/14
H2O.ai
2 667 просмотров

Deep Learning Lecture 9: Neural networks and modular design in Torch
Nando de Freitas
1 873 просмотра

What is Deep Learning AI and How You Should Use it Today?
AlchemyAPI
3 451 просмотр

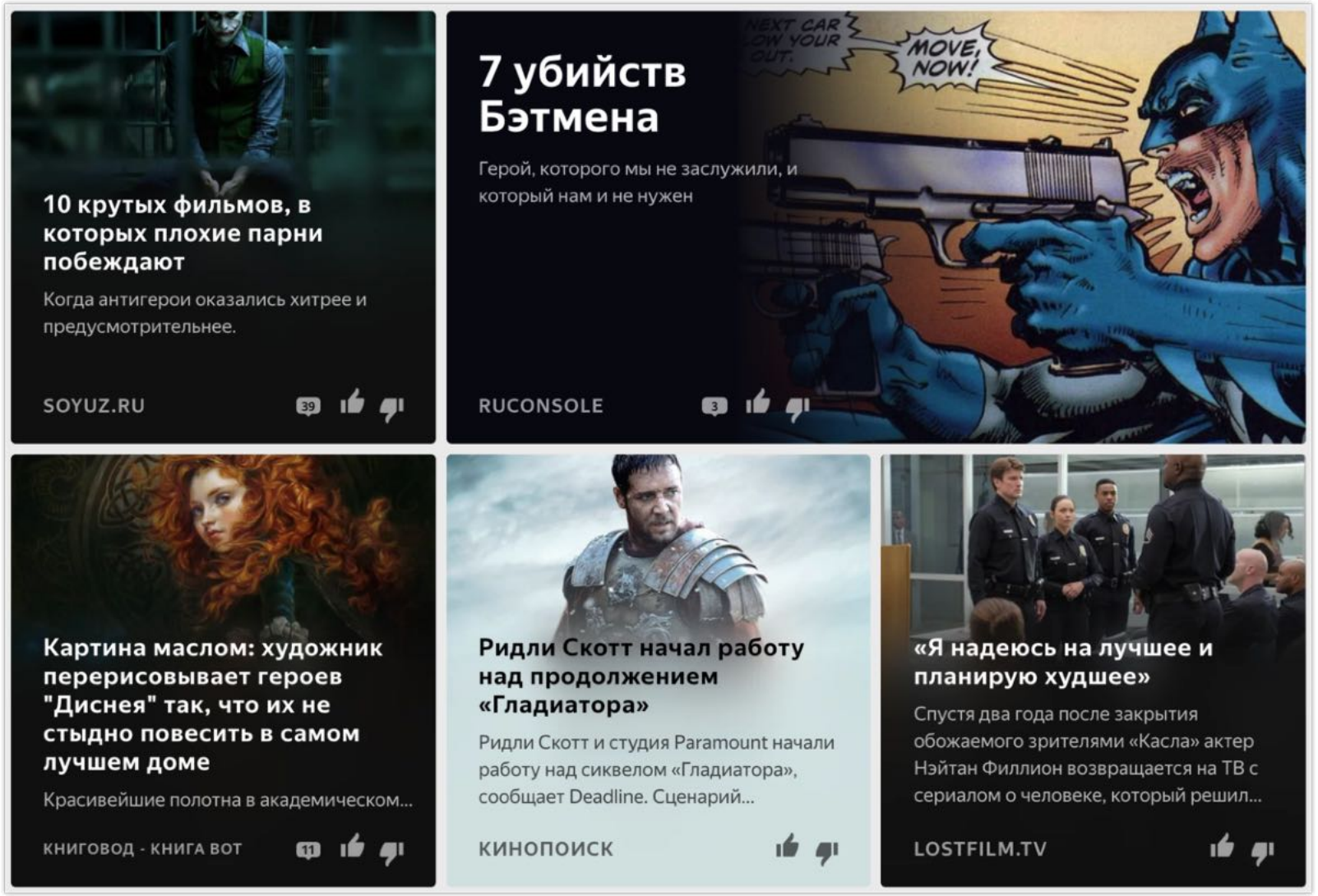
Andrew Ng: Deep Learning, Self-Taught Learning and Unsupervised Feature
黄鑫
125 534 просмотра

Deep Learning: The Theoretician's Nightmare or Paradise? (LeCun, NYU,
npresearch

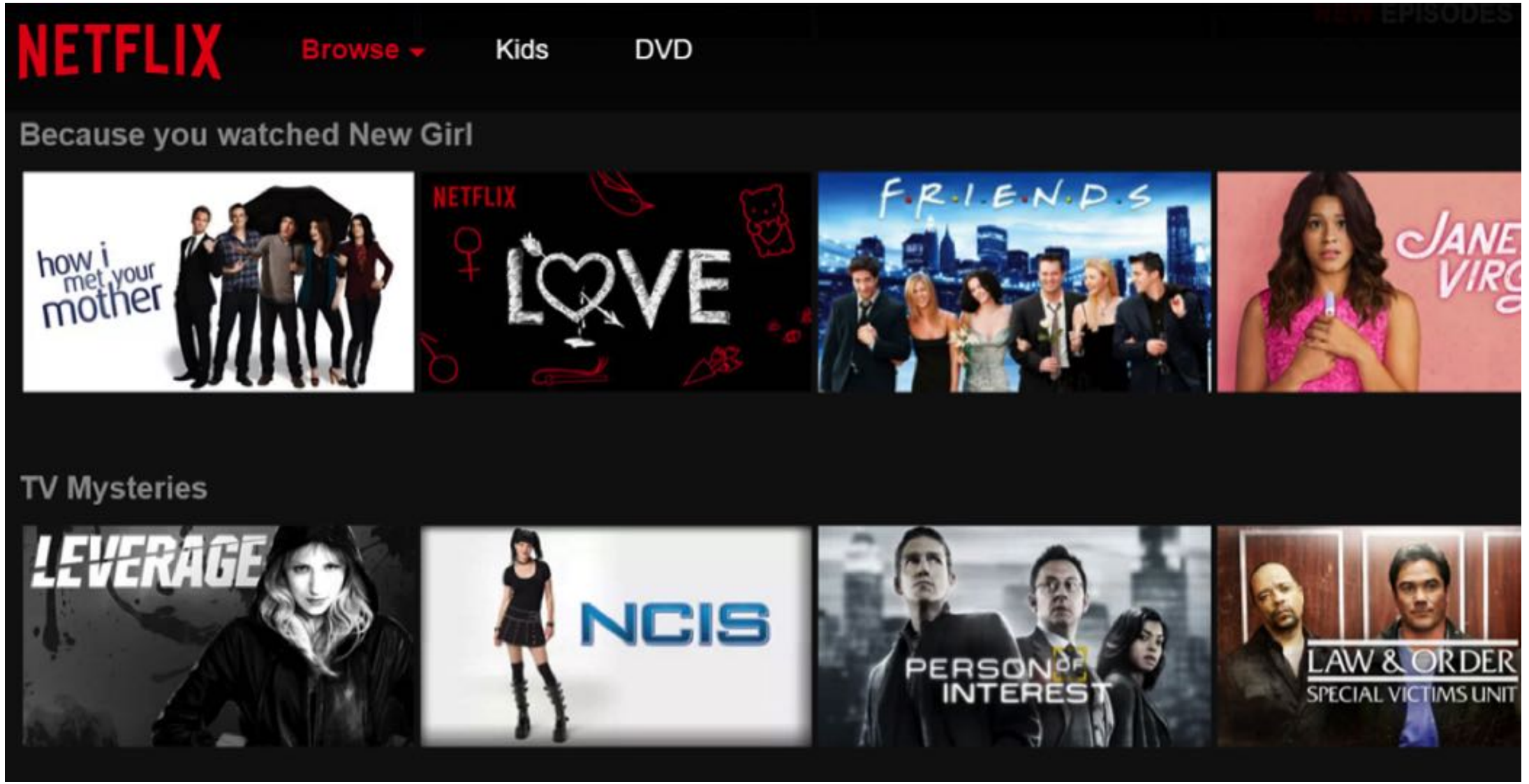
Рекомендательные сервисы

Яндекс.Дзен, Кинопоиск, Netflix

Яндекс.Дзен


10 крутых фильмов, в которых плохие парни побеждают
Когда антигерои оказались хитрее и предусмотрительнее.
SOYUZ.RU
39
7 убийств Бэтмена
Герой, которого мы не заслужили, и который нам и не нужен
RUCONSOLE
3
MOVE, NOW!
Картина маслом: художник перерисовывает героев "Диснея" так, что их не стыдно повесить в самом лучшем доме
Красивейшие полотна в академическом...
КНИГОВОД - КНИГА ВОТ
11
Ридли Скотт начал работу над продолжением «Гладиатора»
Ридли Скотт и студия Paramount начали работу над сиквелом «Гладиатора», сообщает Deadline. Сценарий...
КИНОПОИСК
«Я надеюсь на лучшее и планирую худшее»
Спустя два года после закрытия обожаемого зрителями «Касла» актер Нэйтан Филлион возвращается на ТВ с сериалом о человеке, который решил...
LOSTFILM.TV

Видеопрокат в Netflix



Netflix Prize: View Leaderboard - Chromium

www.netflixprize.com/leaderboard

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

$1,000,000 за улучшение алгоритма на 10%.

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09

Зачем это все

Какова цель? Какую задачу решаем?

Зачем это все

Какова цель? Какую задачу решаем?

- Увеличить продажи
- Повысить время, проведённое на сайте
- Предоставить клиенту удобный сервис
- Построить сообщество

Неперсональные рекомендации

Неперсональные

Для всех пользователей показываем
одни и те же рекомендации

Рейтинги kinopoisk.ru

Гарри Поттер и философский камень

Проблемы с рейтингами

- **Явные рейтинги**
  - разная шкала (субъективная)
  - разброс рейтингов
- **Неявные рейтинги**
  - покупки (понравилось или нет?)
  - время на сайте (а если отвлекся?)
  - клик (является ли «не клик» сигналом, а что после клика?)
- **Накрутки**

Средний рейтинг

Явный рейтинг

It's ok

I love it

I like it

I hate it!

I don't like it

Средний рейтинг

$$P_i = \frac{\sum_{u=1}^{n} r_{ui}}{n}$$

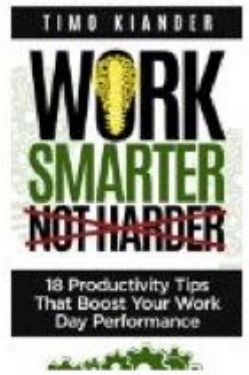
рейтинг пользователя u для товара i

item

количество пользователей оценивших i

Топ по среднему на Amazon

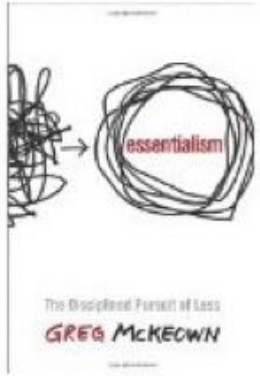
Все ОК?



Work Smarter Not Harder: 18 Productivity Tips That Boost Your Work Day Performance Mar 25, 2015
by Timo Kiander

Kindle Edition
$0.00
Auto-delivered wirelessly

★★★★★ ▾ 2



Essentialism: The Disciplined Pursuit of Less Apr 15, 2014
by Greg McKeown

Hardcover
$17.50 ~~$23.00~~ Prime
Get it by **Wednesday, May 6**

More Buying Choices
$10.61 used & new (62 offers)

Kindle Edition
$10.99
Whispersync for Voice-ready

★★★★½ ▾ 508

Trade-in eligible for an Amazon gift card
FREE Shipping on orders over $35

Excerpt
Page 5 : ... some new strategy in *time management*. It is about pausing ... See a random page in this book.

Проблема

Если мало рейтингов, то оценка неуверенная!

Регуляризация среднего

$$P_i = \frac{\sum_{u=1}^{n} r_{ui} + k\mu}{n + k}$$

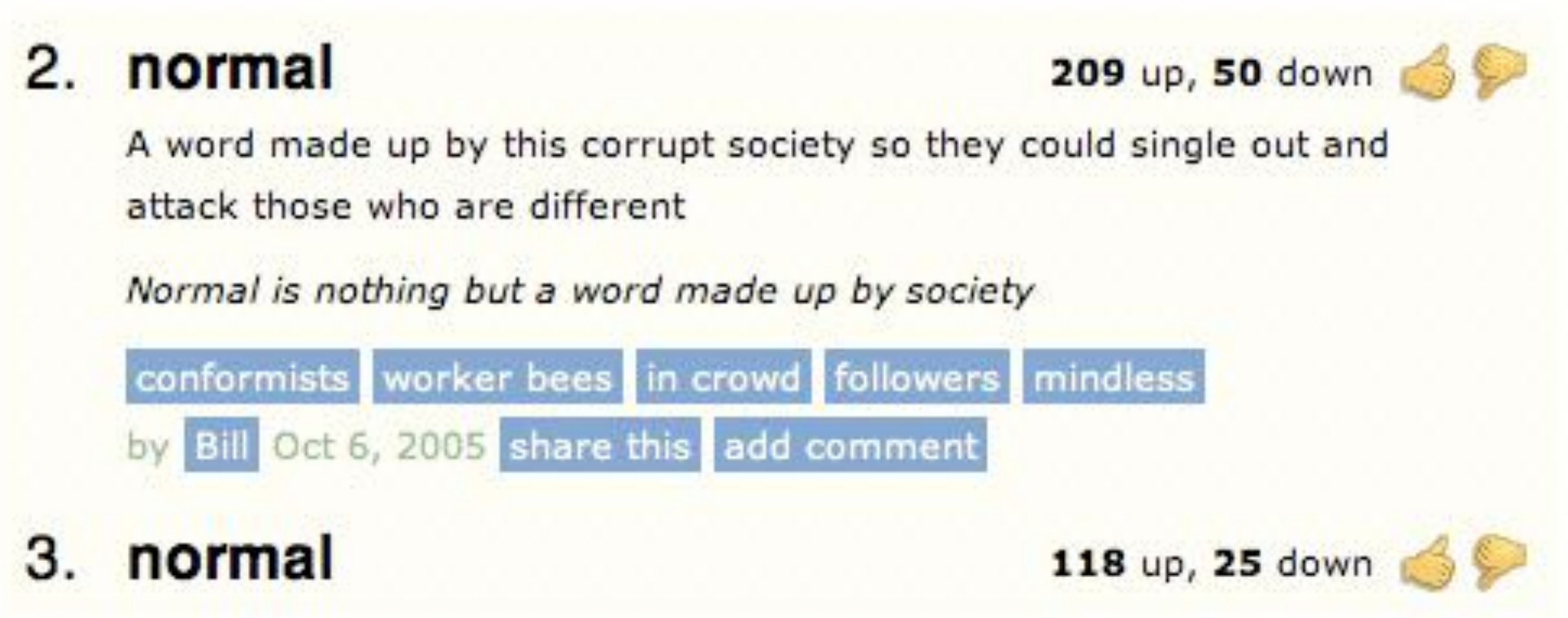
глобальное среднее

контролирует минимальное количество наблюдений

Лайки и дизлайки

Рекомендации на evanmiller.org

Все ОК?



Сортировка по *чистым* лайкам («like» - «dislike»)

Проблема

Разности («like» - «dislike») для разных товаров несравнимы!

Вероятность лайка

Каждый рейтинг принимает только два значения 1 и 0 (like, dislike)

Каждый рейтинг – случайная величина Бернулли с вероятностью p

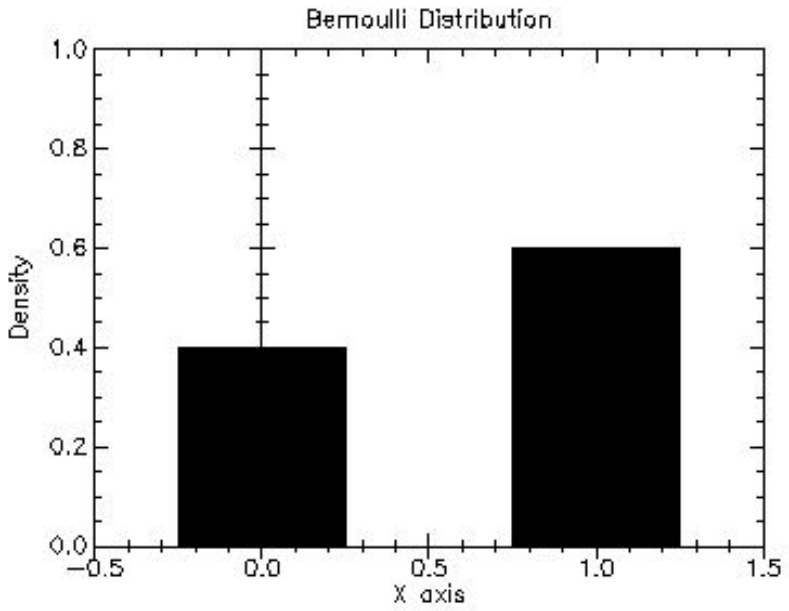
Распределение Бернулли

Вероятность лайка ($x = 1$): p (успех)

Вероятность дизлайка ($x = 0$): $1 - p$ (неуспех)

лайк дизлайк

Частоты:



дизлайк лайк

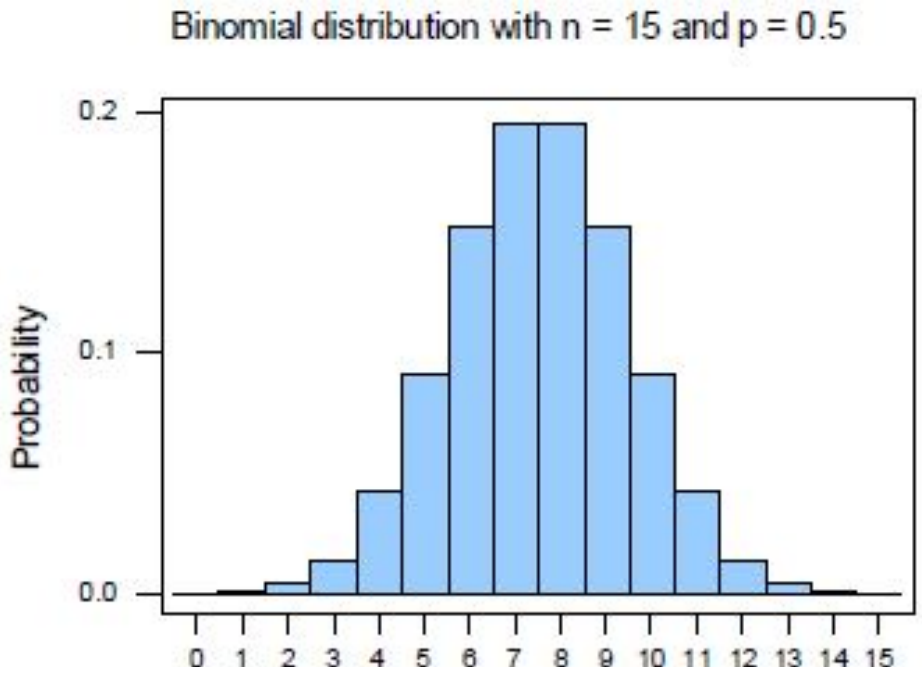
Сумма независимых величин

Проведем серию из n «подкидываний монеты»

Какова вероятность получить k лайков?

Построим распределение для $x_1 + \cdots + x_n$

лайк дизлайк

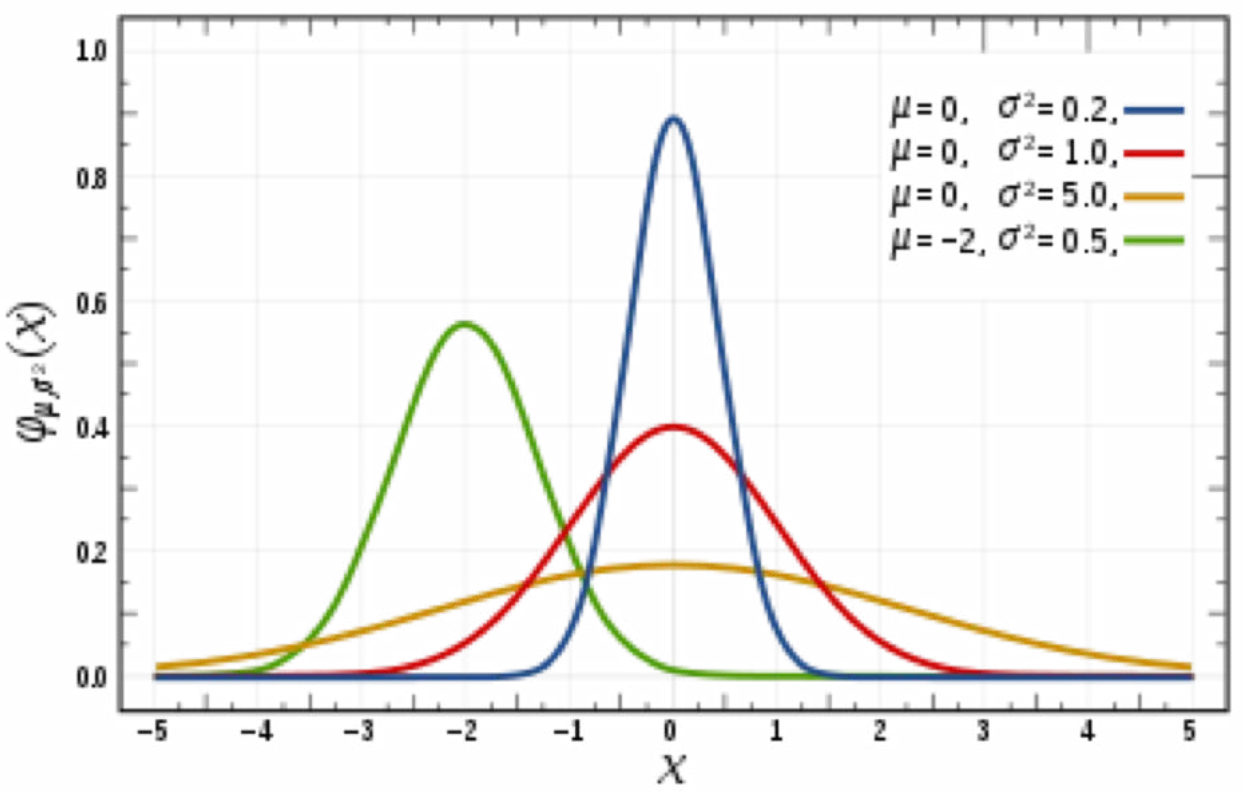


Биномиальное распределение

https://en.wikipedia.org/wiki/Binomial_distribution

Нормальное распределение

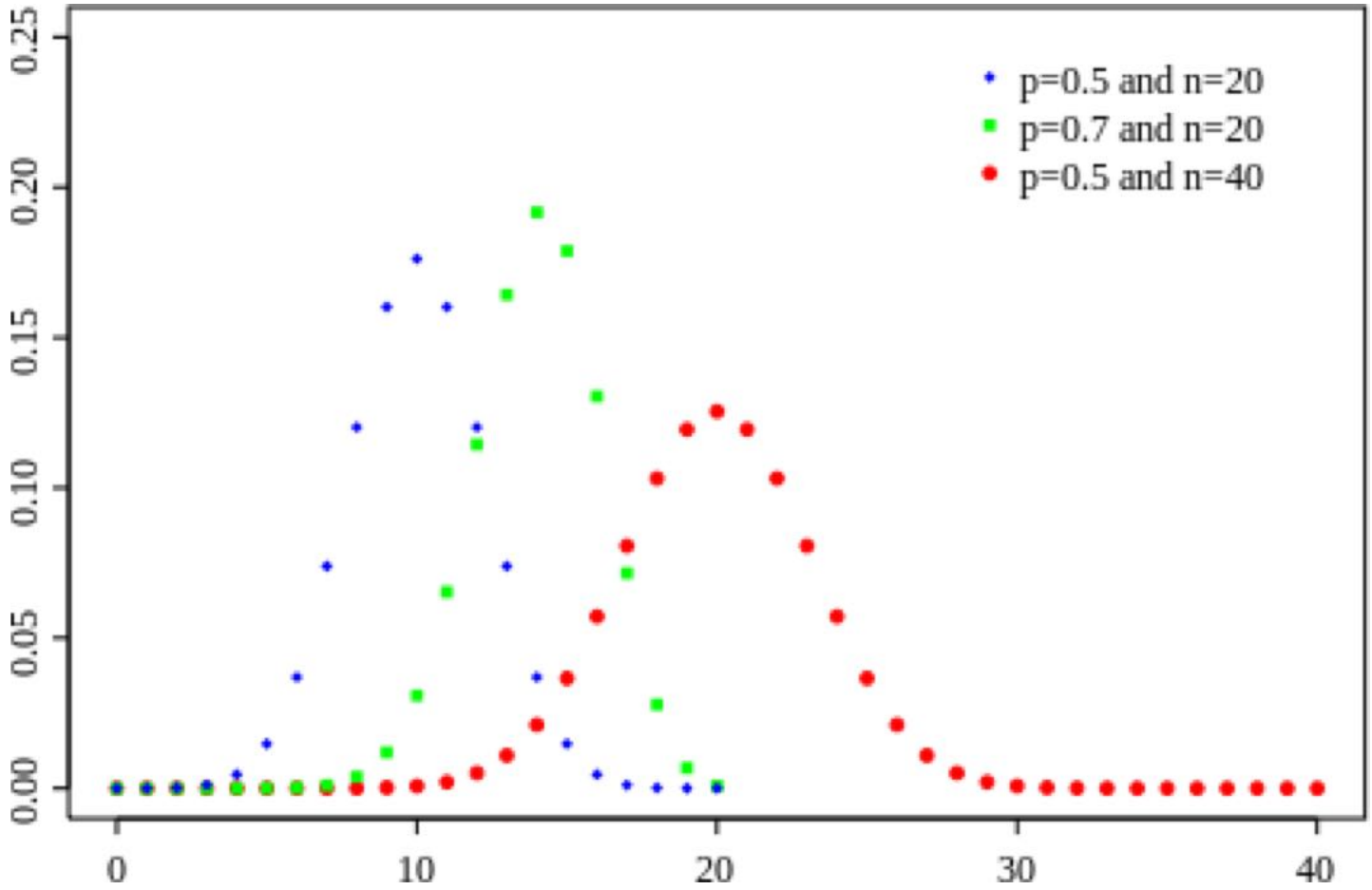
Величина объясняется суммой большого количества независимых компонент → ее распределение близко к нормальному



Приблизим нормальным

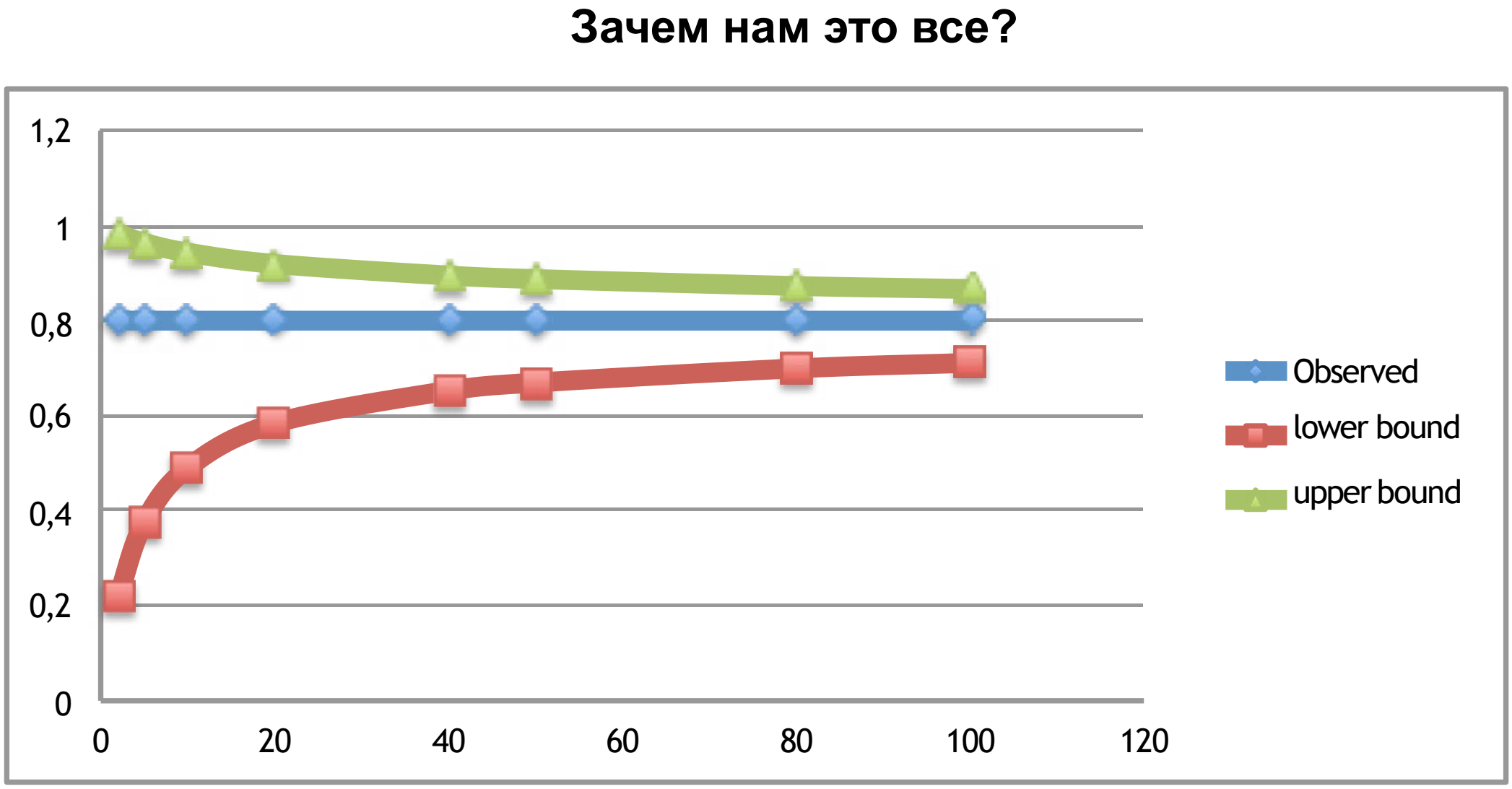
$$\mu = P$$

$$\sigma = \sqrt{\frac{1}{n} P(1-P)}$$



https://corplingstats.wordpress.com/2014/02/05/binomial-normal-wilson/#more-1162

Доверительный интервал



Учитываем уверенность

Ранжируем по нижней границе доверительного интервала!

Неперсональное ранжирование

Плюсы:

- Легко сделать
- Хорошо работает для новых пользователей

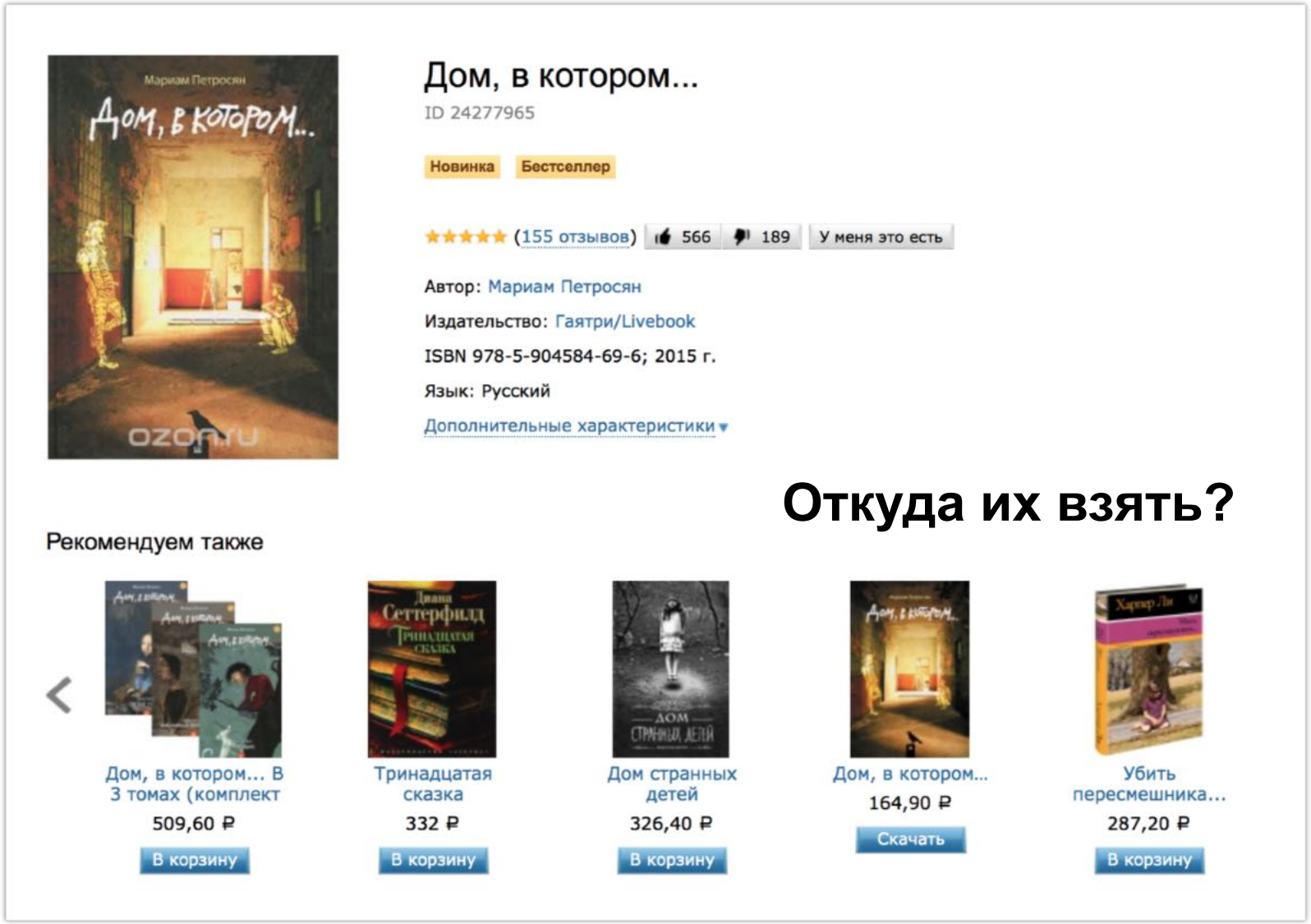
Минусы:

- Нет персонализации
- Смещенные оценки (люди больше жалуются, чем хвалят)

Неперсональные рекомендации в ozon.ru

Рекомендации к товару

Не зависит от пользователя



Наивный подход

Можно посчитать, как часто два товара покупают вместе.

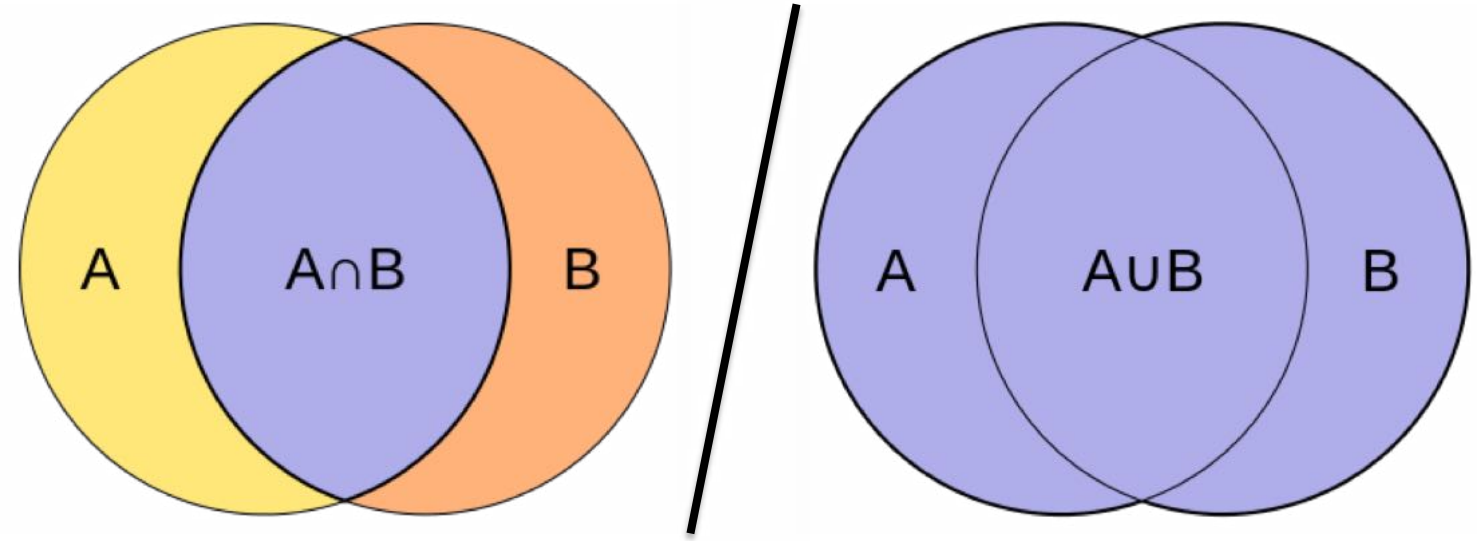
Наивный подход

Можно посчитать, как часто два товара покупают вместе.

Окажется, что туалетную бумагу покупают ко всему ☺

Мера Жаккара для множеств

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Что в нашем случае множество?

В нашем случае

Множество для товара – это все пользователи, которые его купили.

Тогда мы будем измерять похожесть двух товаров с точки зрения купивших их пользователей.

Чем чаще покупают вместе редкие товары, тем лучше.

В нашем случае

Рассмотрим матрицу Item-User, где в ячейке записана 1, если пользователь u покупал товар i.

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

Одним из признаков рекомендательной системы может быть мера Жаккара между строчками матрицы (товарами).

Пример расчета

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

$$J(i_1, i_2) = \frac{0}{4} = 0$$

Пример расчета

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

$$J(i_1, i_2) = \frac{0}{4} = 0$$

$$J(i_1, i_3) = \frac{1}{4} = 0.25$$

Алгоритмическая сложность

В реальной задаче:

- Миллионы пользователей (N)
- Миллионы товаров (M)

Как посчитать меру Жаккара для всех товаров?

Алгоритмическая сложность

В реальной задаче:

- Миллионы пользователей (N)
- Миллионы товаров (M)

Как посчитать меру Жаккара для всех товаров?

- Наивный подход: O(M * M * N)

Оптимальный алгоритм

Вклад в числитель только от совместных покупок каждого пользователя!

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

Инвертированный индекс

(1, 3) → +1 в пересечение

Так мы посчитаем мощность пересечения!

Как быть с мощностью объединения?

Оптимальный алгоритм

Вклад в числитель только от совместных покупок каждого пользователя!

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

Инвертированный индекс

(1, 3) → +1 в пересечение

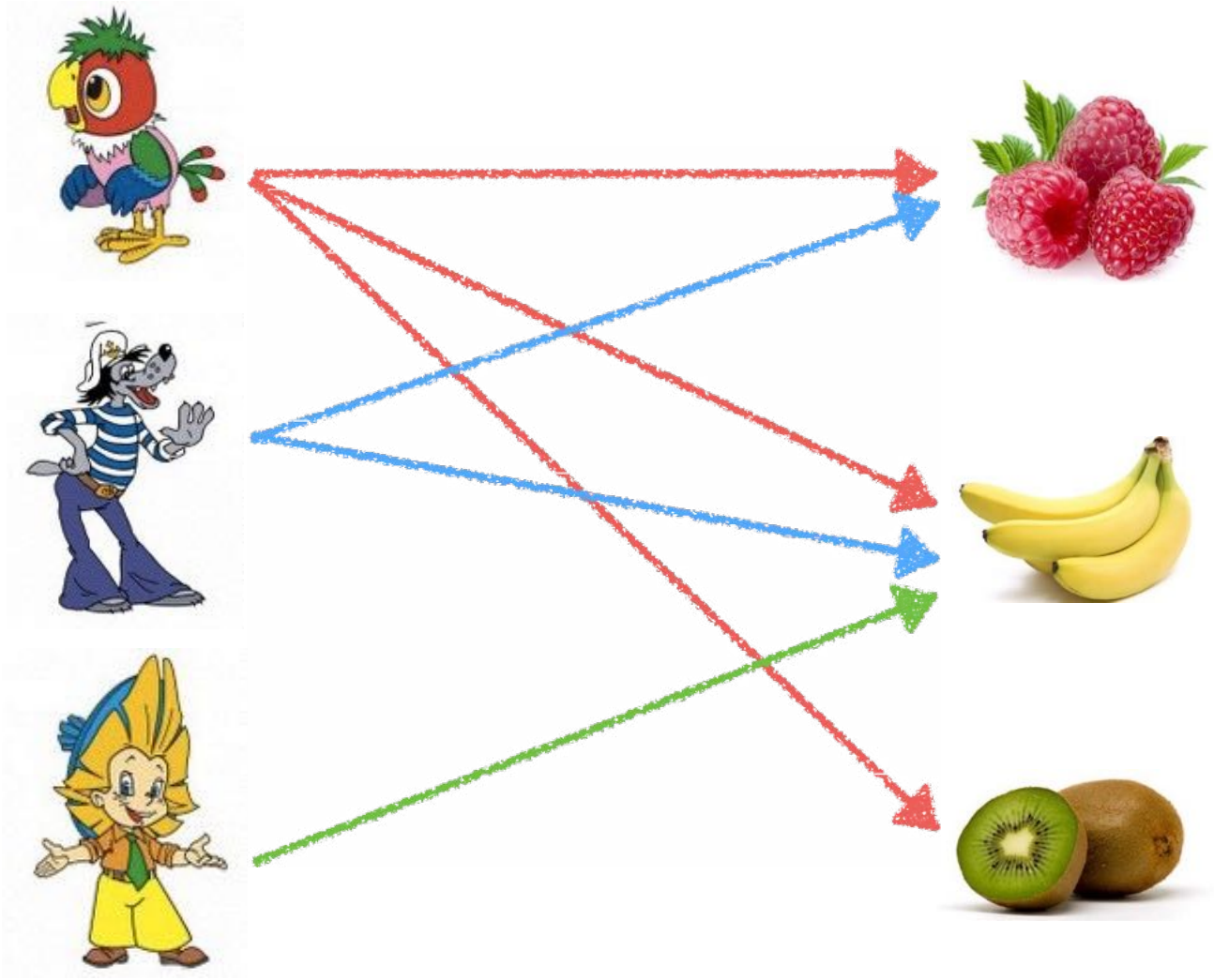
Так мы посчитаем мощность пересечения!

Как быть с мощностью объединения?

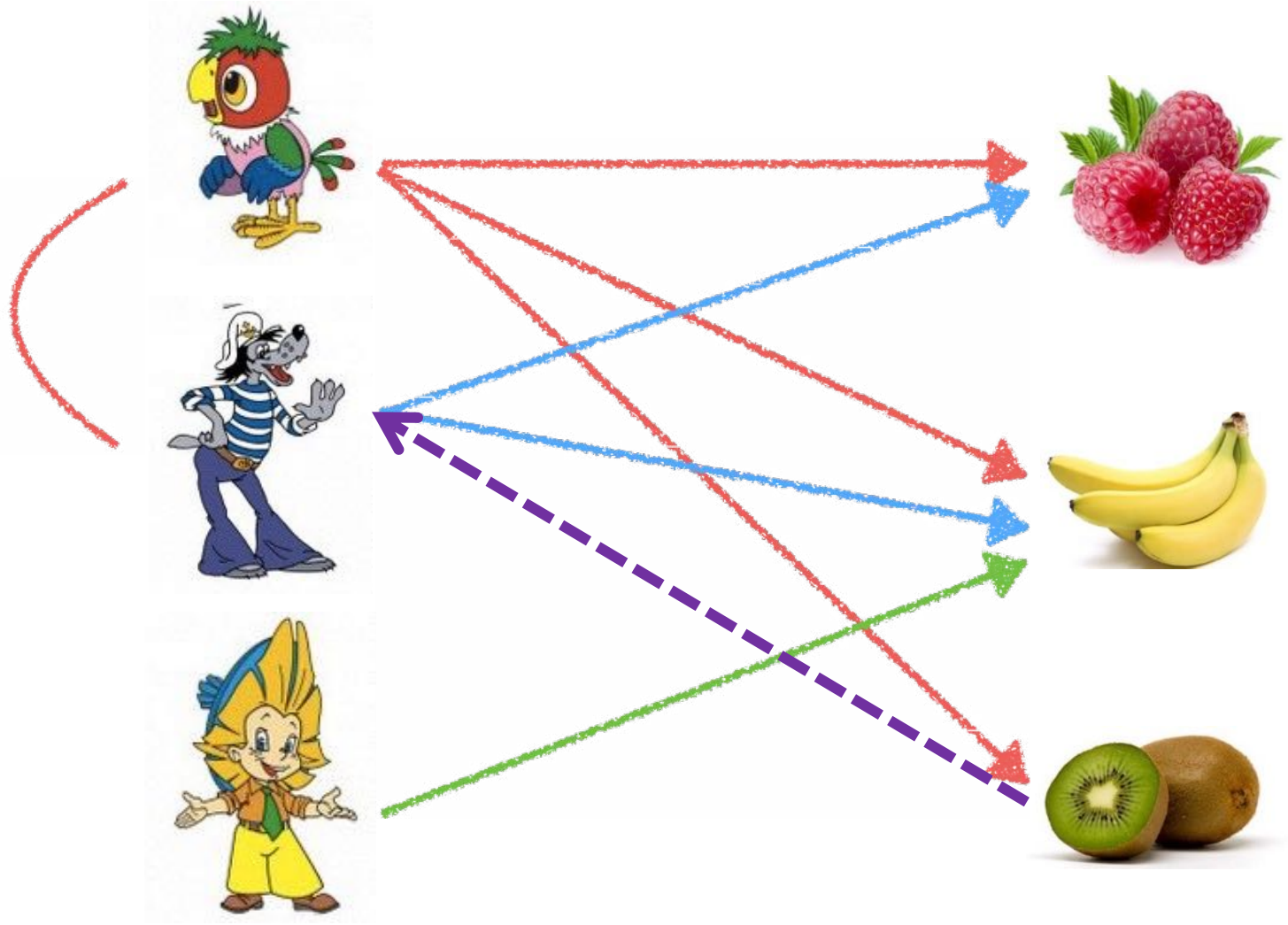
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Коллаборативная фильтрация

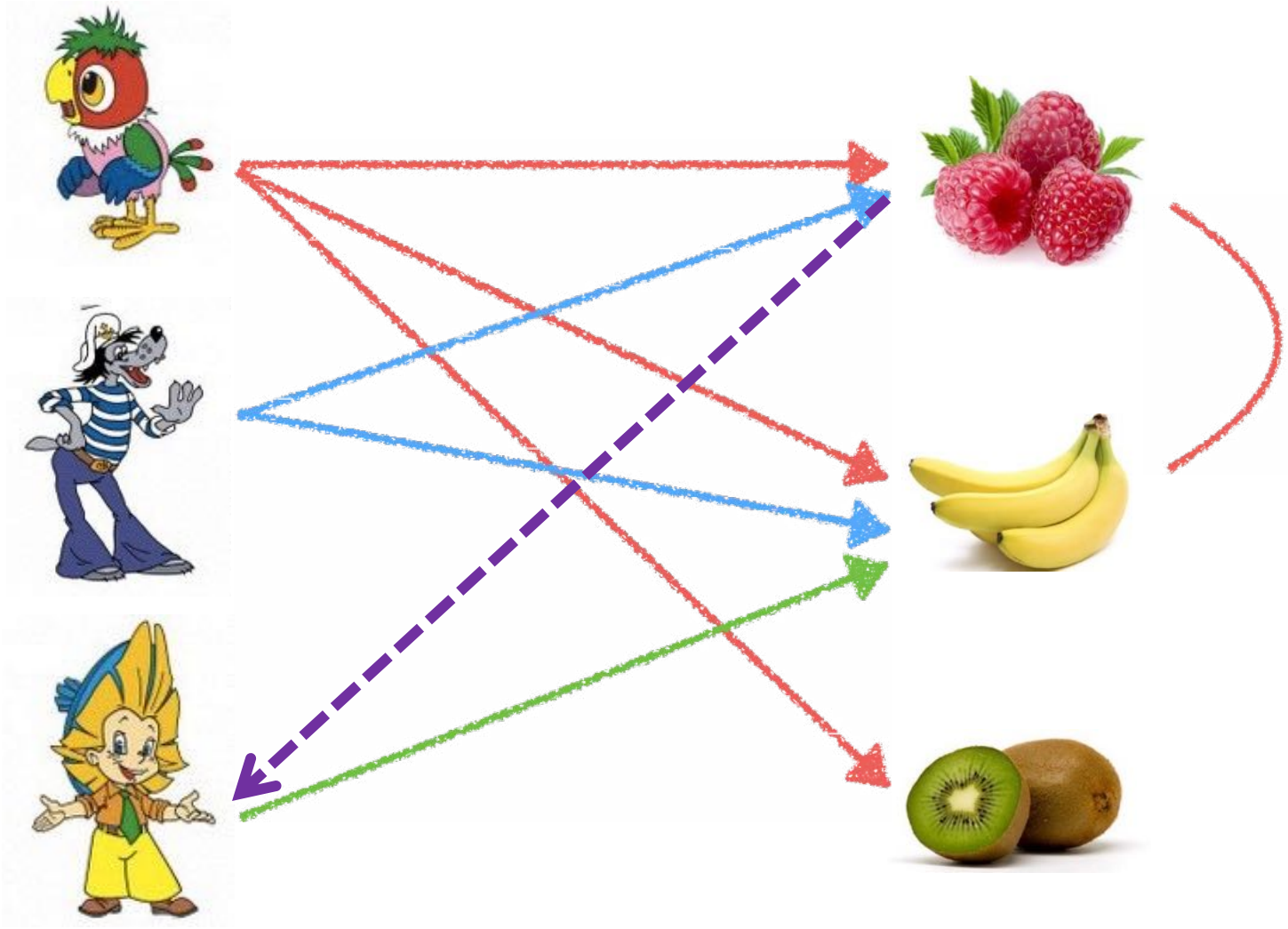
Как использовать эти данные?



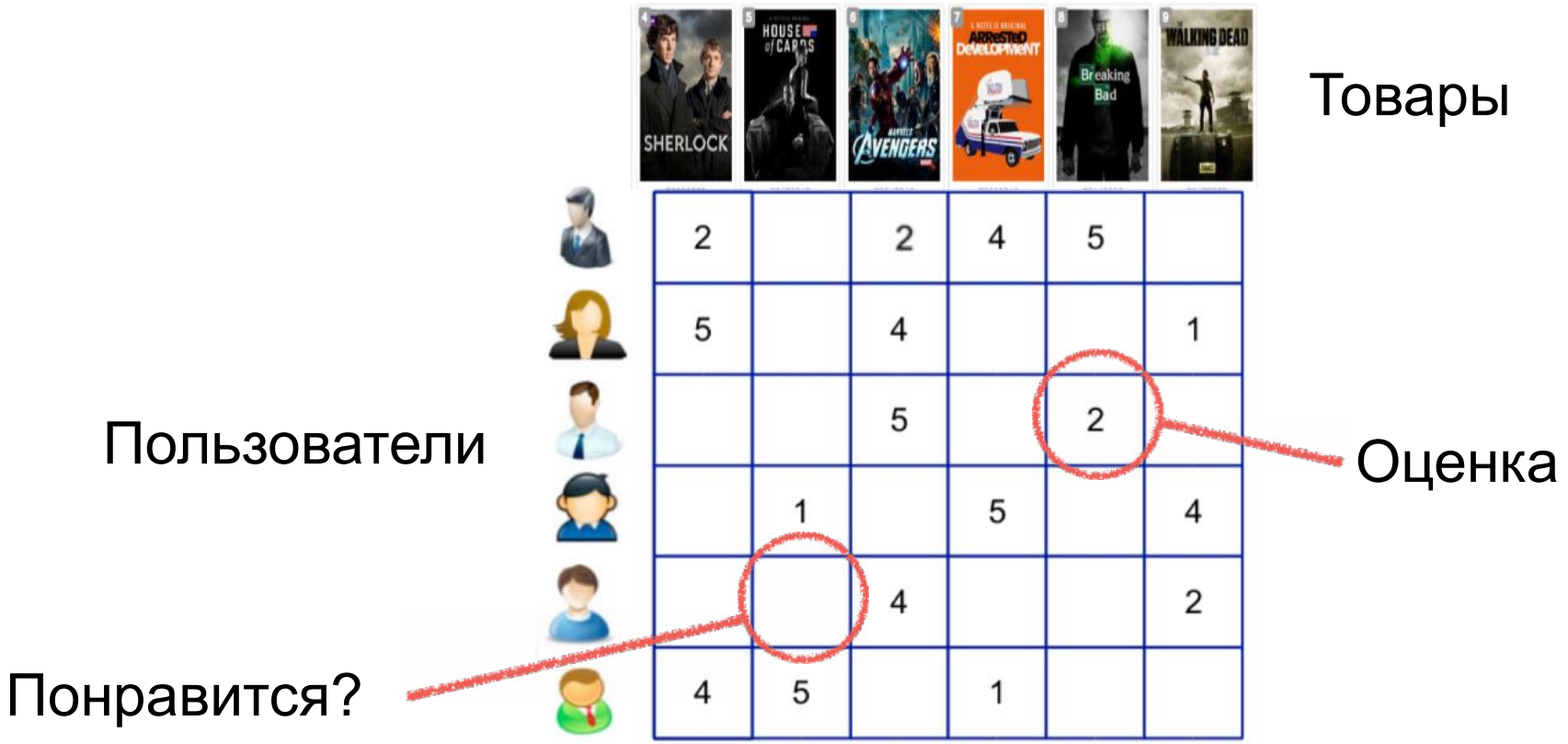
User-based Collaborative Filtering



Item-based CF

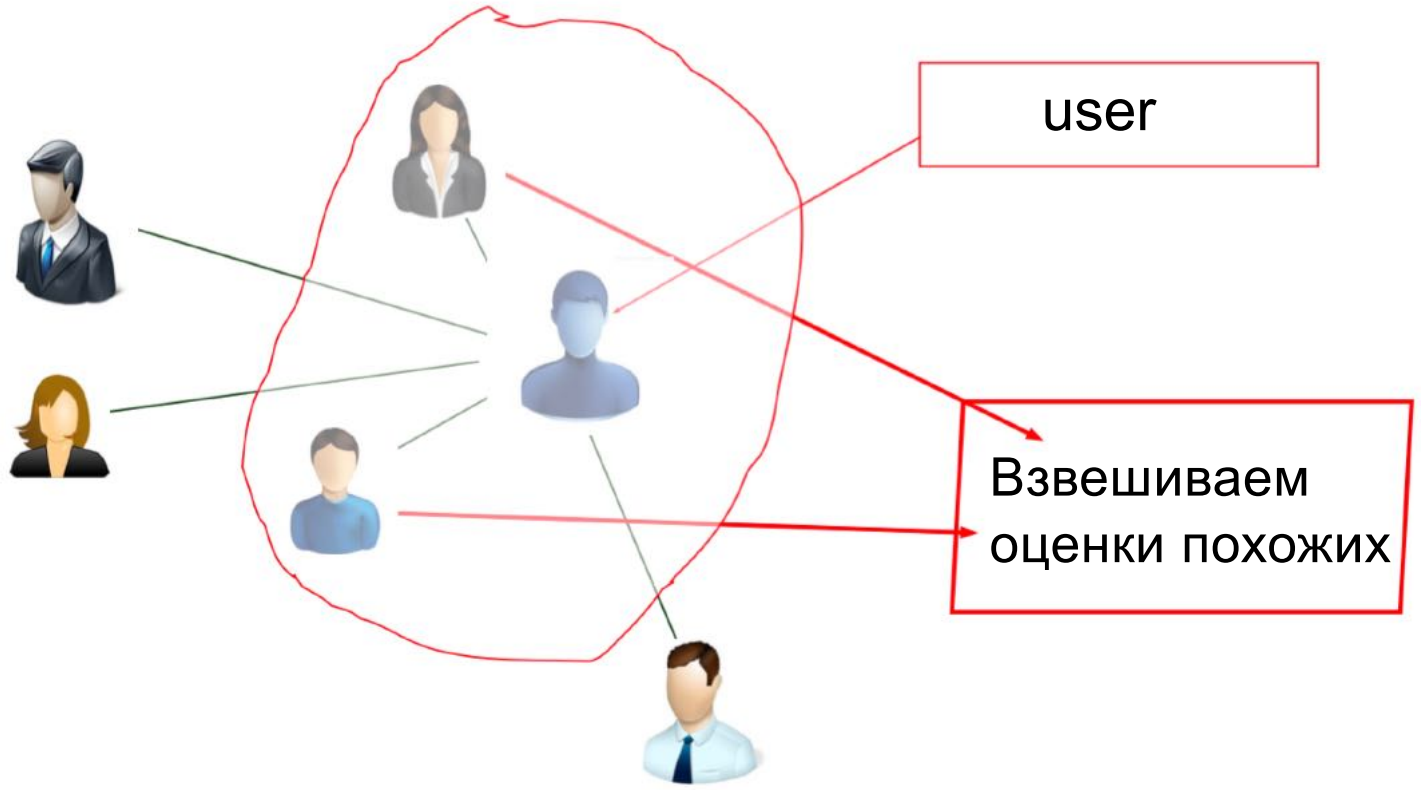


Матрица оценок



User-based CF

Идея: Найдем похожих на **user** пользователей и порекомендуем ему понравившиеся им товары.



Что такое похожесть юзеров?

	2		2	4	5	
	5		4			1
			5		2	
		1		5		4
			4			2
	4	5		1		

Идеи?

Корреляция оценок!

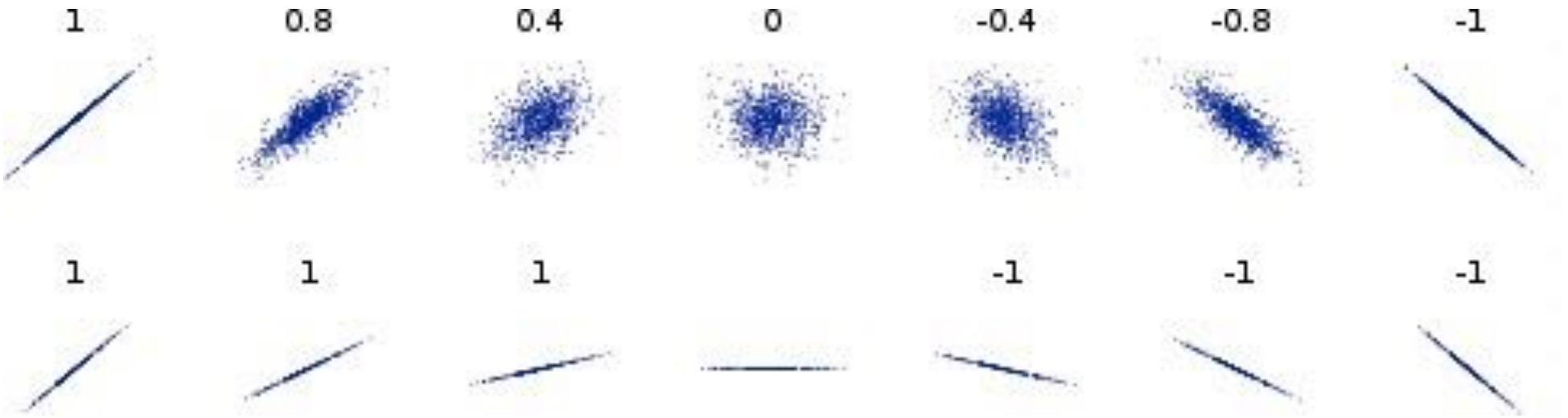
средний рейтинг юзера (по всем оценкам)

корреляция Пирсона

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)(r_{ui} - \overline{r}_u)}{\sqrt{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)^2} \sqrt{\sum_{i \in I_a \cap I_u} (r_{ui} - \overline{r}_u)^2}}$$

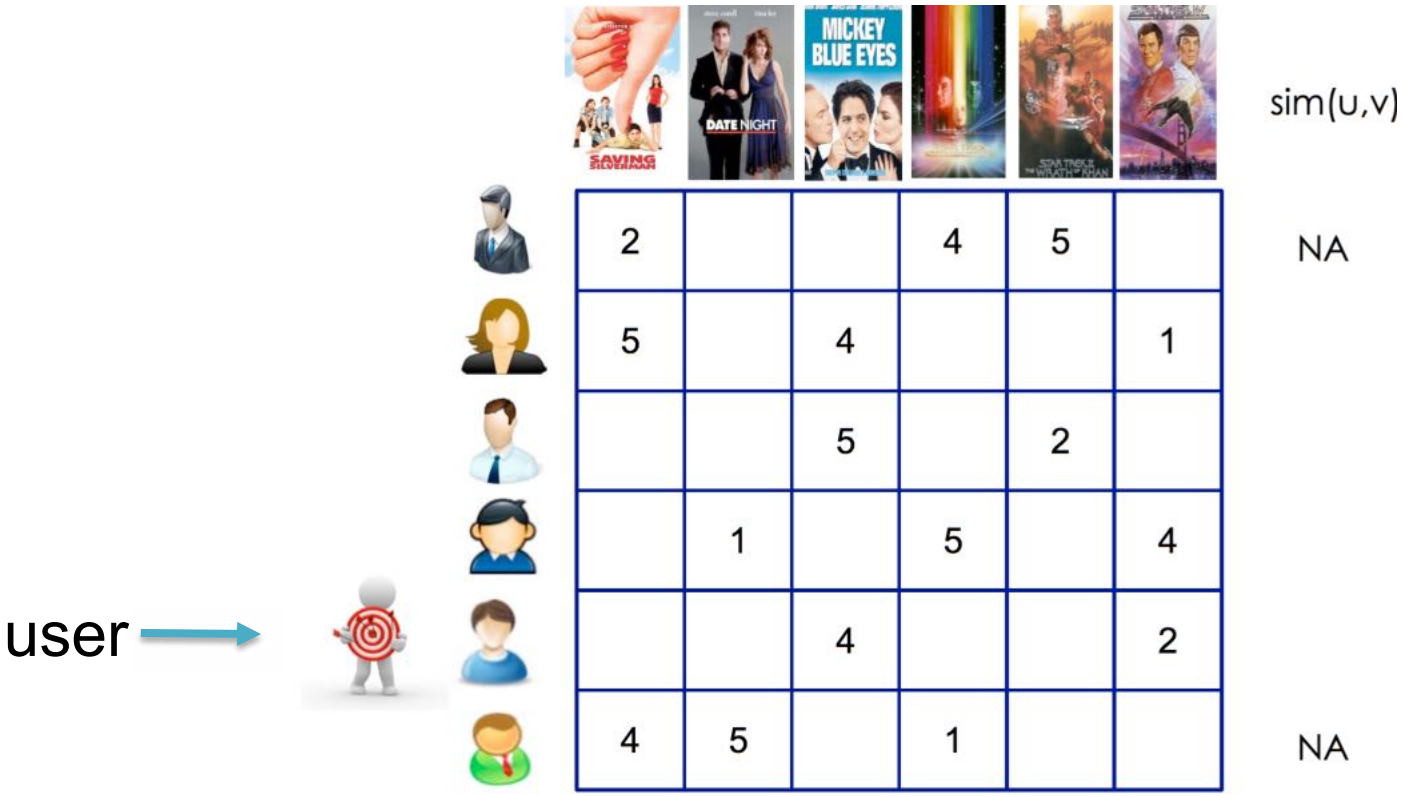
общие рейтинги

Корреляция Пирсона



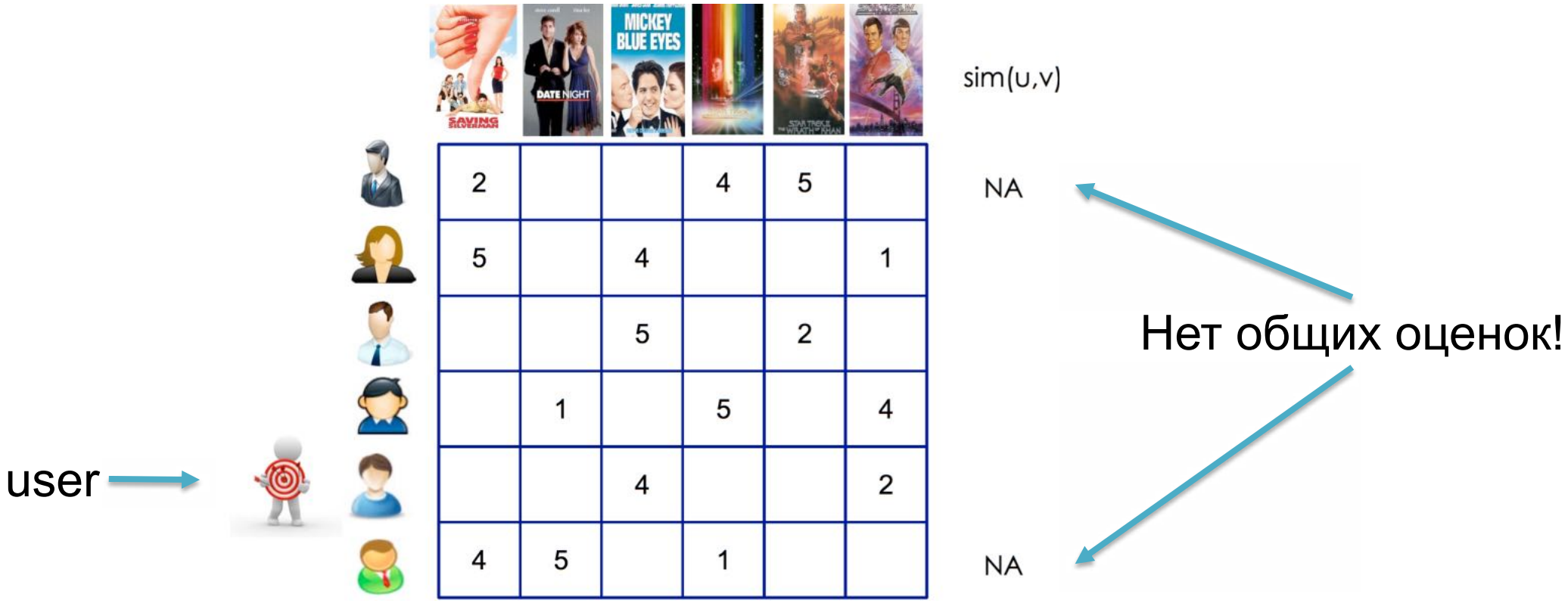
Изменяется от -1 до 1

Пример

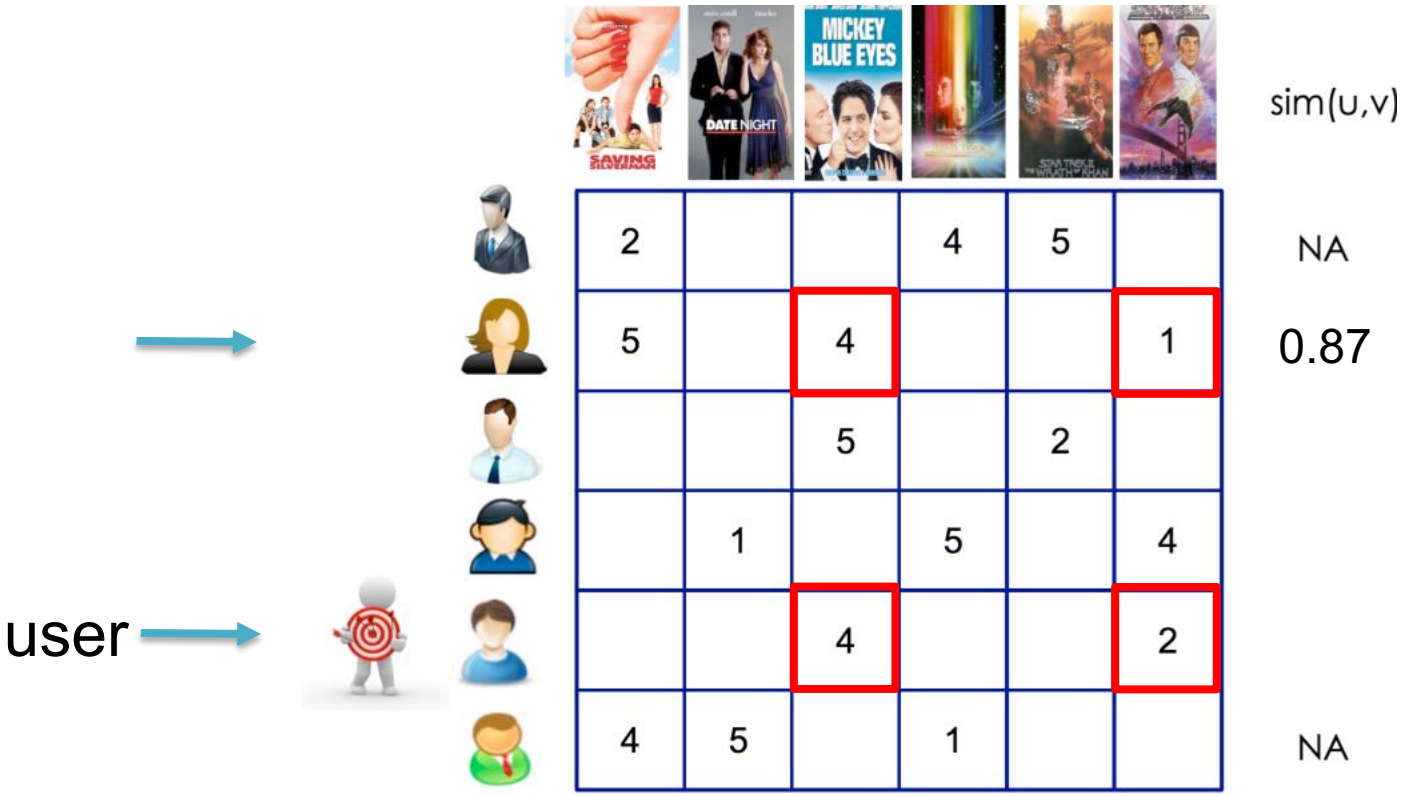


Почему?

Пример

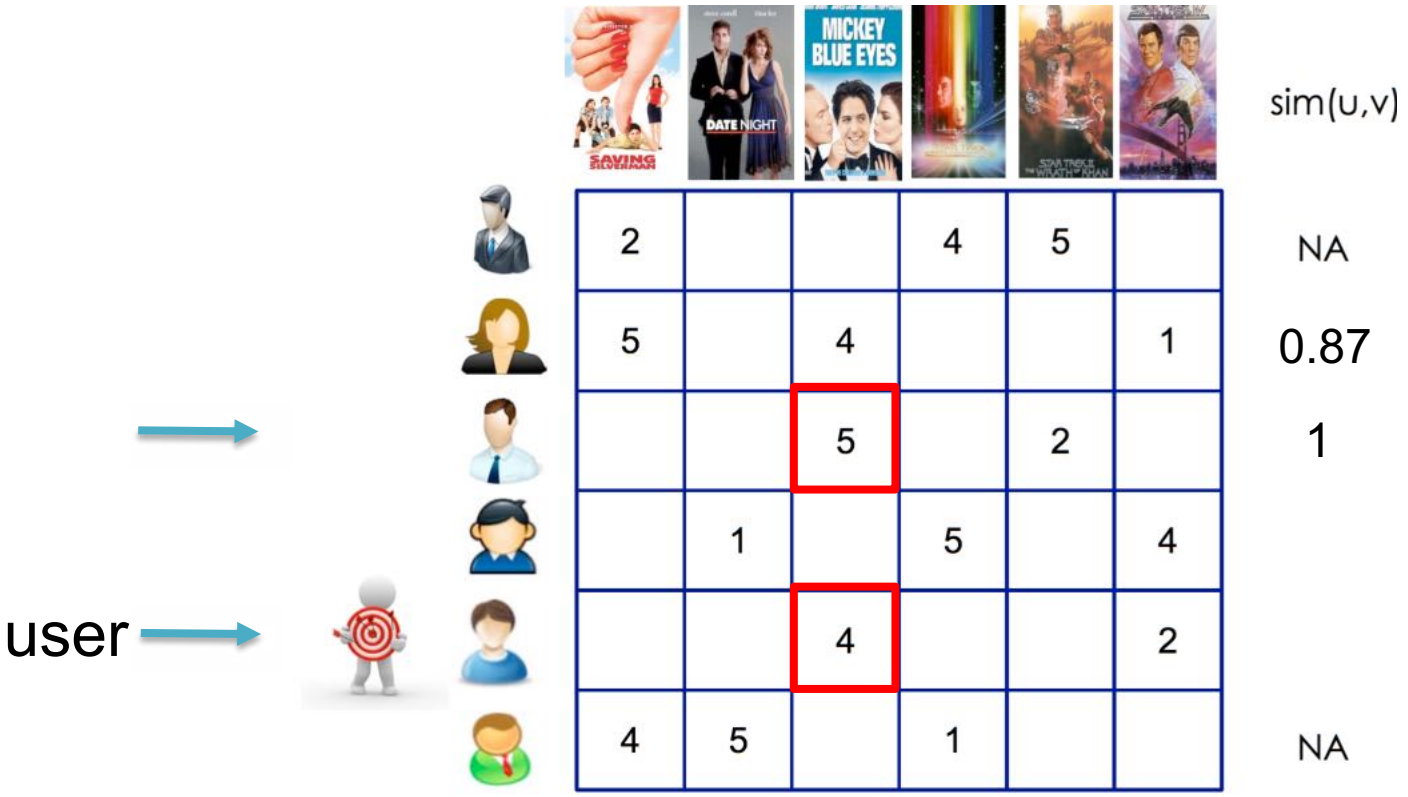

MICKEY BLUE EYES
DATE NIGHT
sim(u,v)
2 4 5
NA
5 4 1
5 2
1 5 4
4 2
4 5 1
NA
user
Нет общих оценок!

Пример



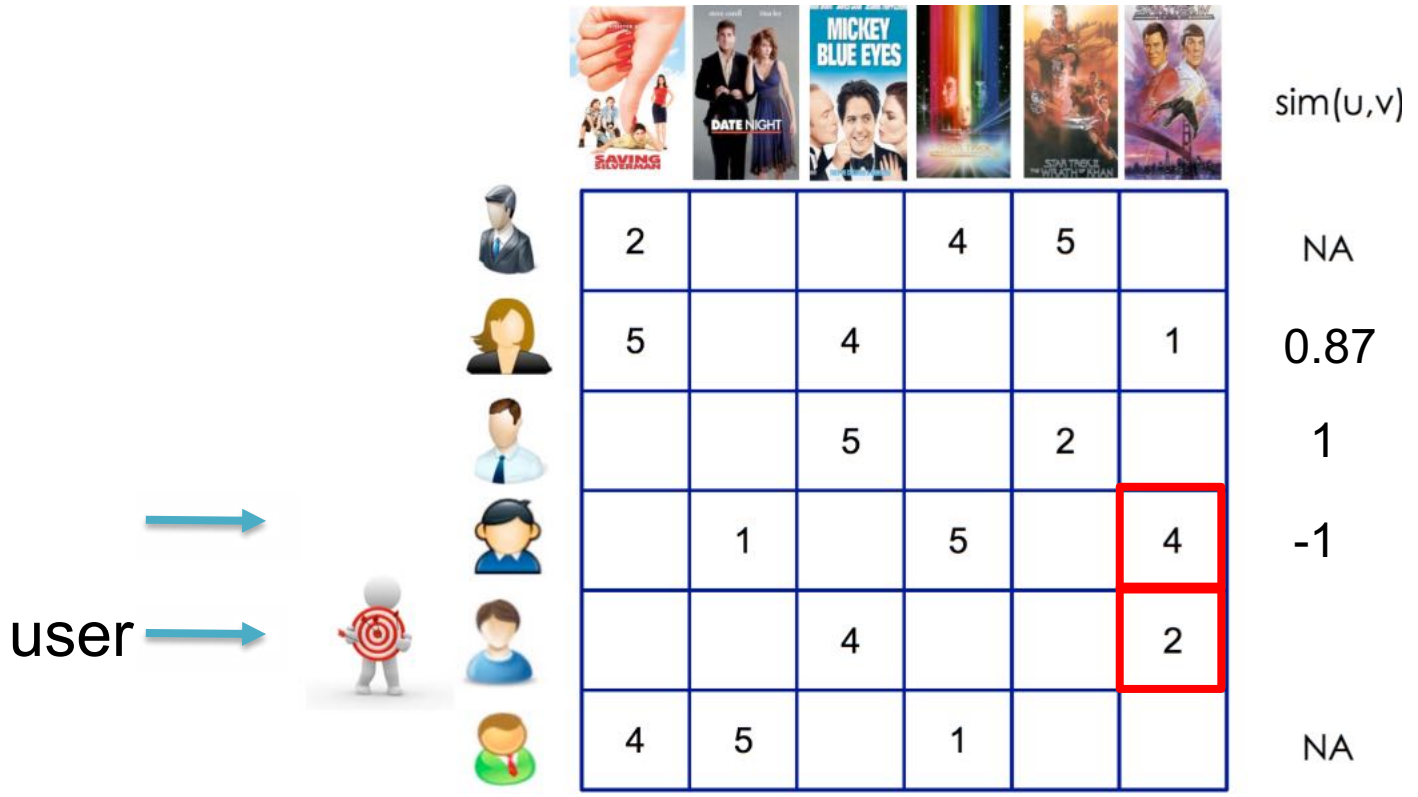
Не 1, потому что максимальная оценка у юзера 5

Пример



Если бы вычитали среднее по общим оценкам, получили бы деление на ноль!

Пример



И не нашли бы это…

Юзер с одинаковыми оценками

Если <u>все оценки</u> юзера одинаковые, то будет деление на ноль!

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)(r_{ui} - \overline{r}_u)}{\sqrt{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)^2} \sqrt{\sum_{i \in I_a \cap I_u} (r_{ui} - \overline{r}_u)^2}}$$

0

Нужно пропускать таких пользователей!

Мало оценок в пересечении

В случае одной общей оценки:

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)(r_{ui} - \overline{r}_u)}{\sqrt{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)^2} \sqrt{\sum_{i \in I_a \cap I_u} (r_{ui} - \overline{r}_u)^2}} = \frac{(r_{ai} - \overline{r}_a)(r_{ui} - \overline{r}_u)}{\sqrt{(r_{ai} - \overline{r}_a)^2} \sqrt{(r_{ui} - \overline{r}_u)^2}} = \pm 1$$

Произведение знаков

Проблема: большие неуверенные значения!
Что делать?

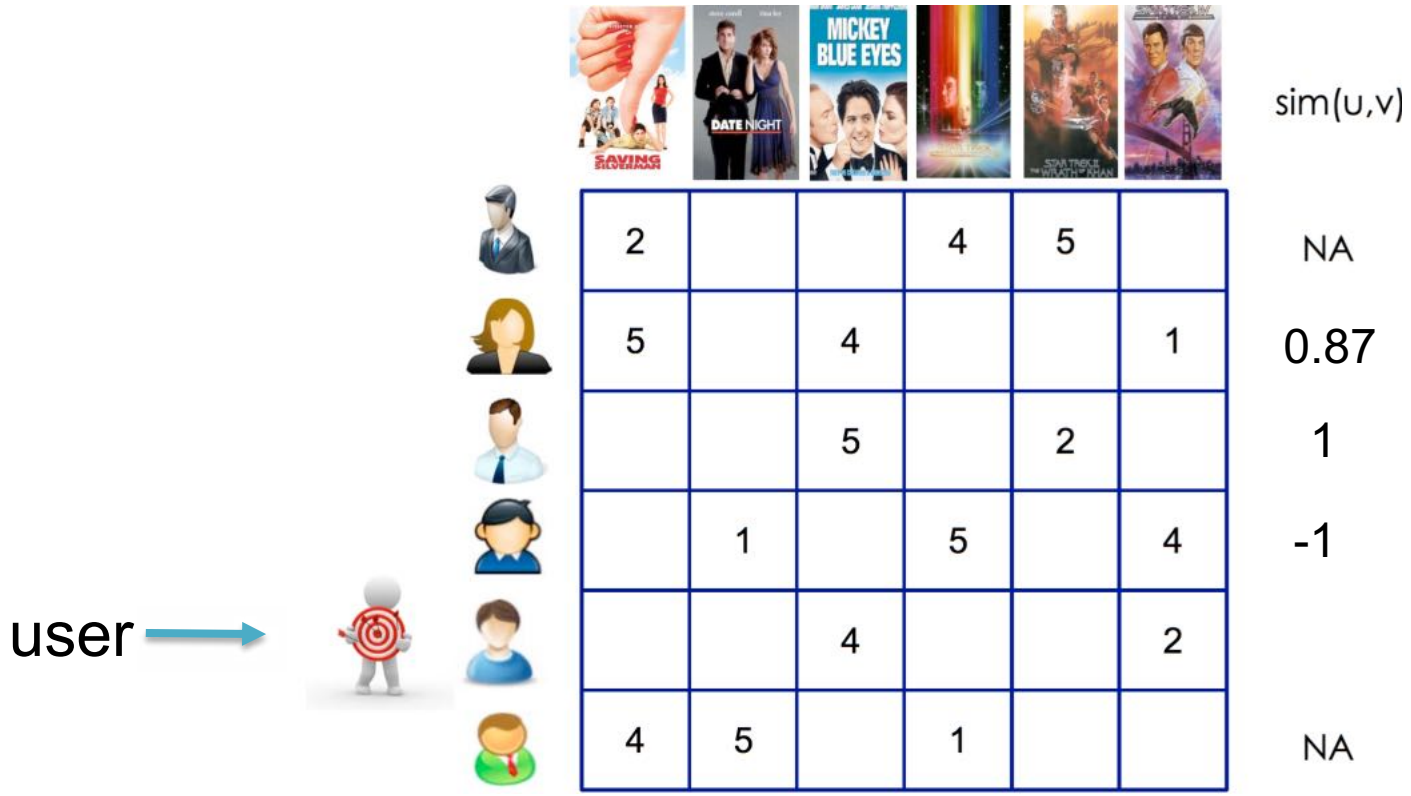
Мало оценок в пересечении

Решение: поправочный коэффициент!

$$s(a,u) = \min\left(\frac{|I_a \cap I_u|}{50}, 1\right) \frac{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)(r_{ui} - \overline{r}_u)}{\sqrt{\sum_{i \in I_a \cap I_u} (r_{ai} - \overline{r}_a)^2} \sqrt{\sum_{i \in I_a \cap I_u} (r_{ui} - \overline{r}_u)^2}}$$

50 – порог на количество общих рейтингов

Что дальше?



Идеи?

Среднее по соседям с весами

мера близости

рейтинг пользователя a для товара i

$$\hat{r}_{ai} = \frac{\sum_{u \in N(a)} s(a,u) r_{ui}}{\sum_{u \in N(a)} |s(a,u)|}$$

множество соседей

Проблема: юзеры ставят оценки в разной шкале!
Кто-то от 1 до 3, кто-то от 3 до 5! Что делать?

Учтем средний рейтинг!

мера близости

средний рейтинг

рейтинг пользователя a для товара i

$$\hat{r}_{ai} = \bar{r}_a + \frac{\sum_{u \in N(a)} s(a,u)(r_{ui} - \bar{r}_u)}{\sum_{u \in N(a)} |s(a,u)|}$$

множество соседей

Проблема: Кто-то от 1 до 5, кто-то от 2 до 4! Что делать?

И поделим на отклонение!

среднеквадратичное отклонение

$$p_{ai} = \bar{r}_a + \sigma_a \frac{\sum_{u \in N(a)} s(a,u)(r_{ui} - \bar{r}_u)/\sigma_u}{\sum_{u \in N(a)} |s(a,u)|}$$

$$\sigma_a = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (r_{ai} - \bar{r}_a)^2}$$

Может быть отрицательным!

$$\hat{r}_{ai} = \frac{\sum_{u \in N(a)} s(a,u) r_{ui}}{\sum |s(a,u)|}$$

- Можно выкинуть отрицательные *s(a,u)*, теряем информацию
- Или использовать формулу с поправками (прошлый слайд)
- Или заменять отрицательные прогнозы на ближайшие неотрицательные

Сколько соседей брать?

- Всех
- По порогу похожести
- Брать k ближайших, можно начать с k=30

Пример предсказаний

						sim(u,v)
	2			4	5		NA
	5		4			1	0.87
			5		2		1
		1		5		4	-1
	3.51*	3.81*	4	2.42*	2.48*	2	
	4	5		1			NA

Проблемы user-based CF

- Рейтингов у юзера мало → в пересечении еще меньше → неуверенные похожести

- При появлении новой оценки похожести могут сильно измениться
  → не получится посчитать заранее

Прикинем на примере

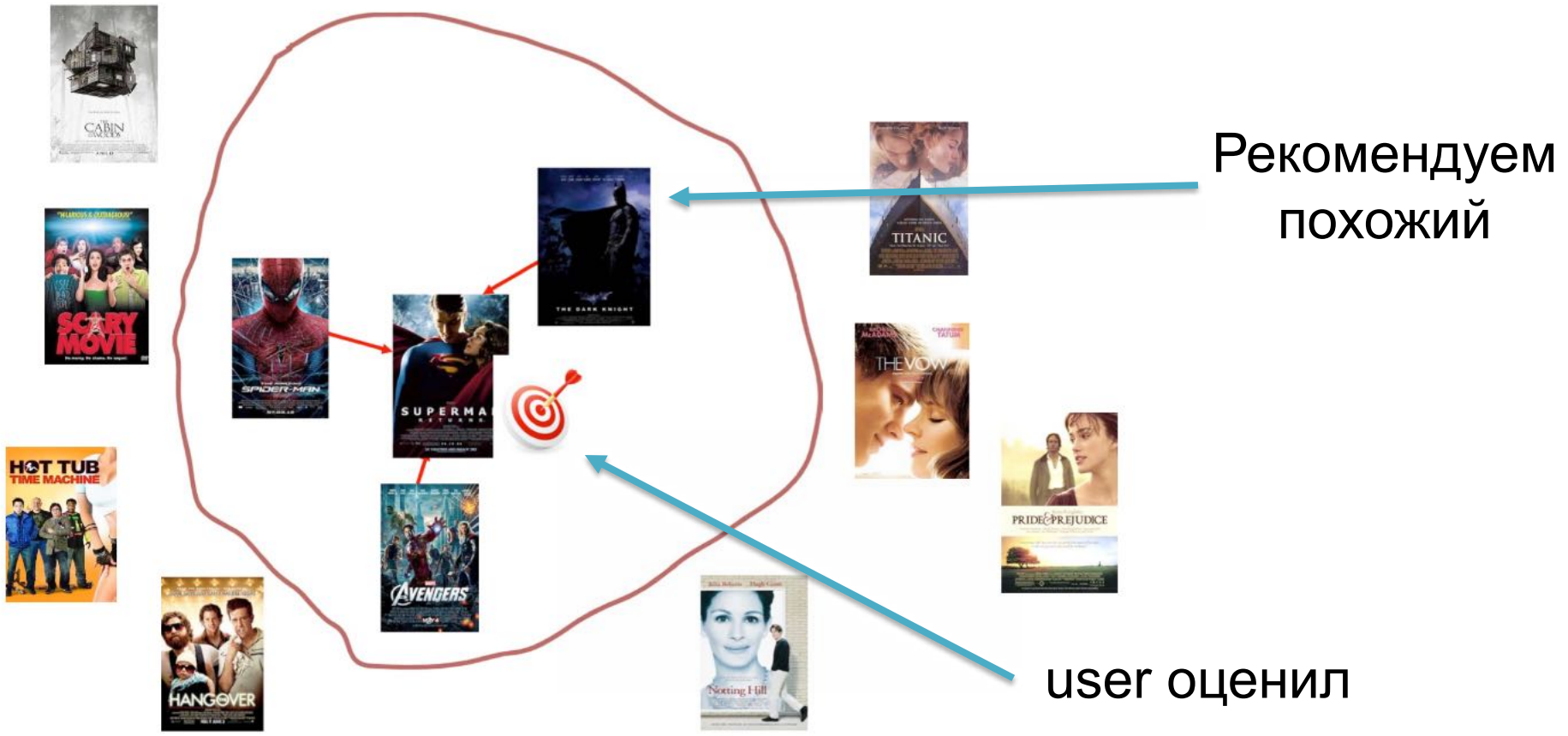
- 10000 рейтингов
- 1000 пользователей
- 100 товаров
- рейтинги распределены равномерно

Прикинем на примере

- 2 случайных пользователя в ожидании имеют 1 общий рейтинг
- 2 случайных товара в ожидании имеют 10 общих рейтингов

Item-based CF

Идея: К оцененным пользователем товарам найдем наиболее похожие на них и порекомендуем.



Похожесть товаров

2		2	4	5	
5		4			1
		5		2	
	1		5		4
		4			2
4	5		1		

Можно действовать также, как для юзеров

Косинусная мера

$$s(i,j) = \frac{\sum_u r_{ui} r_{uj}}{\sqrt{\sum_u r_{ui}^2} \sqrt{\sum_u r_{uj}^2}}$$

Важен только угол между векторами!

Числитель считается только по общим юзерам!

Знаменатель считается по всем юзерам!

То есть заменили пропуски на нули!

Что может пойти не так?

Проблемка

- Рейтинги [1,1] и [5,5] считает максимально близкими!
- Нужно пропускать такие случаи

Adjusted cosine similarity

Работает лучше (поправка на разный диапазон у юзеров):

$$s(i,j) = \frac{\sum_{u=1}^{n}(r_{ui} - r_u)(r_{uj} - r_u)}{\sqrt{\sum_{u=1}^{n}(r_{ui} - r_u)^2 \sum_{u=1}^{n}\left(r_{uj} - r_u\right)^2}}$$

Суммирование только по общим юзерам!

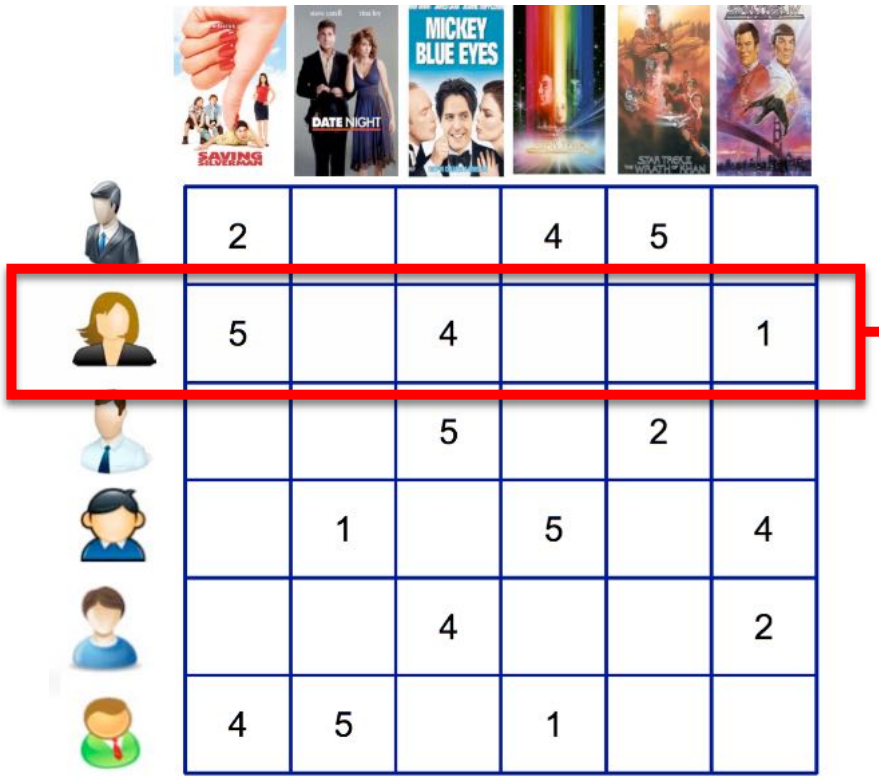
Плюсы item-based CF

- Для популярных товаров можно получить надежную оценку похожести.
- Можно обновлять похожести товаров реже, например раз в день.

https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

Item-item похожести в офлайне

Будем обновлять их раз в день, считаем при помощи инвертированного индекса:



Вклады:

(1, 3) –> (5 – 3.3) * (4 – 3.3)
(1, 6) –> (5 – 3.3) * (1 – 3.3)
(3, 6) –> (4 – 3.3) * (1 – 3.3)

В онлайне

- Быстро реагируем на новые оценки пользователя
- Похожести item-item храним в памяти (топ ~1000 для каждого товара)

Неявный фидбек

Для неявного фидбека, например, покупок, можно использовать меру Жаккара как похожесть!

Резюме CF

Плюсы:

- Неплохие рекомендации при большом количестве явных оценок.

Минусы:

- Плохо работает при сильной разреженности матрицы оценок
- Два пользователя должны оценивать одинаковые товары, оценка *сильно похожих* товаров не учитывается в их близости.
- Проблема холодного старта: не знаем, что делать с новым товаром или пользователем.

В следующей серии…

SVD для рекомендаций

Нейросетевые обобщения SVD

Спасибо за внимание!