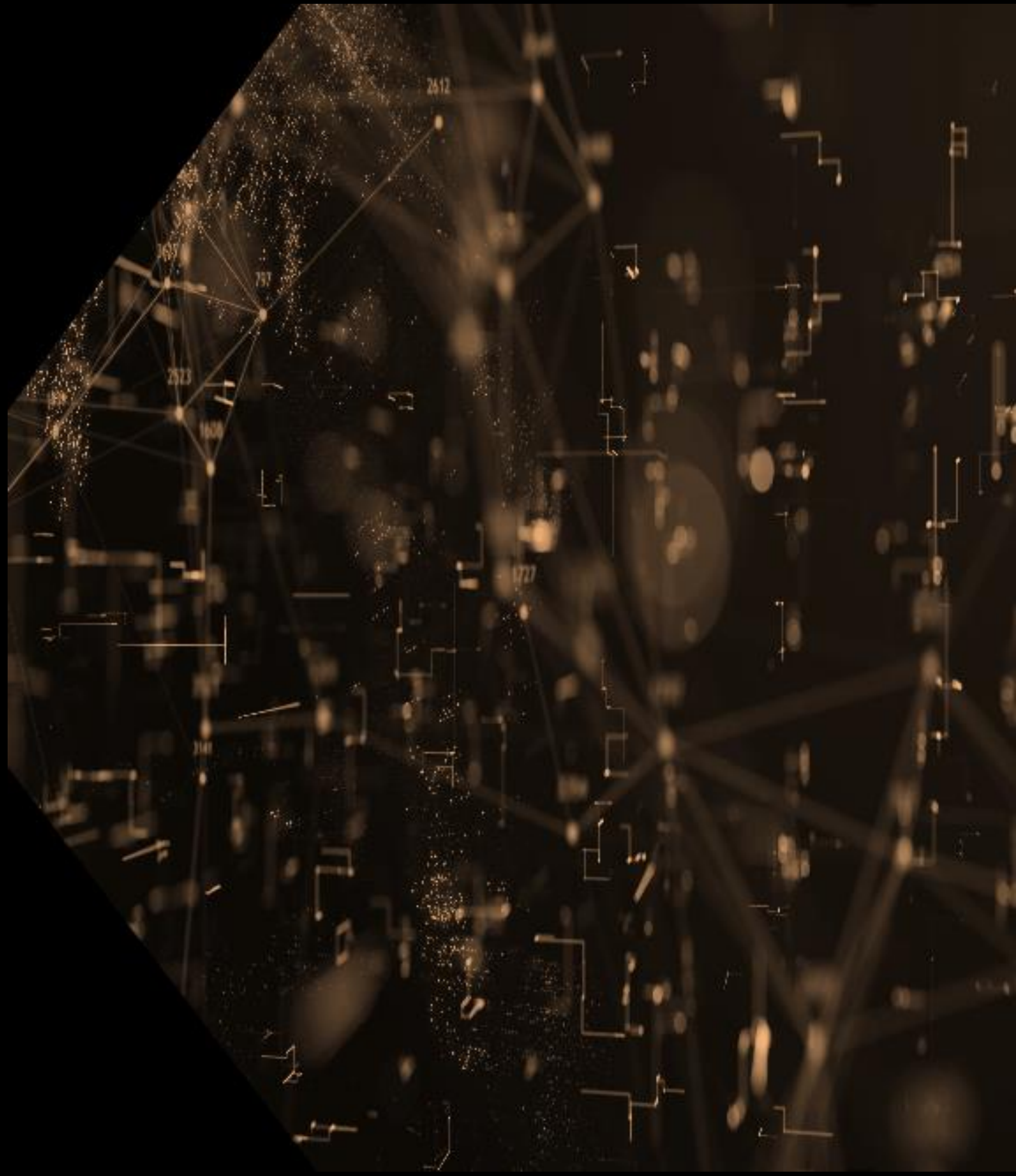


09 августа 2023

Машинное обучение и приложения





Петр Сокерин

Сколковский институт науки и технологий,
РЭУ им. Г.В. Плеханова

Опыт в анализе данных:

- Участник проектов в области машинного обучения в Skoltech
- Автор нескольких публикаций в области анализа данных
- Ex-senior data Scientist в финансовой лаборатории
- Основатель DS клуба в РЭУ им. Плеханова

О чем мы сегодня поговорим

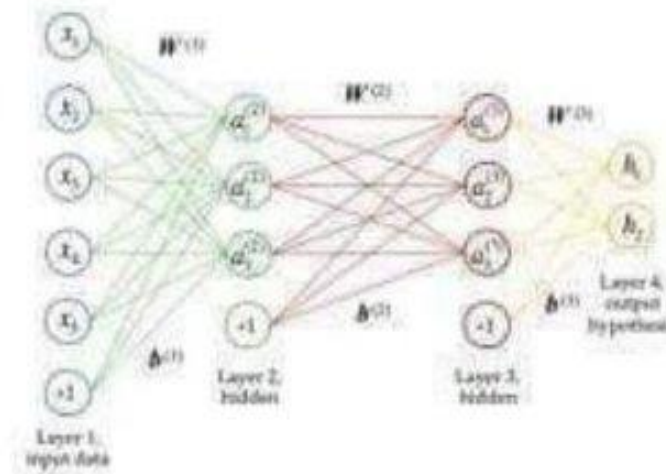
Длительность презентации – 1.5 часа

1. Повторение основных терминов машинного обучения
2. Зачем бизнесу нужно машинное обучение
3. Рекомендательные системы
4. Uplift-моделирование
5. Кейс с оценкой цены автомобиля

Попугай



Машинное обучение



Виды задач в машинном обучении



Обучение с учителем

- Регрессия
- Классификация
- Ранжирование



Что-то между

- Генерация
- Обучение представлений (word2vec)
- Обучение с подкреплением



Обучение без учителя

- Кластеризация
- Детекция аномалий
- Сокращение размерности

Виды задач в машинном обучении



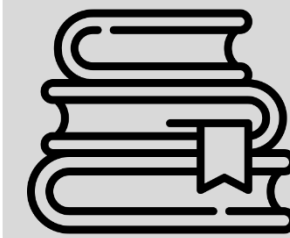
**Обучение
с учителем**

- Регрессия
- Классификация
- Ранжирование



**Что-то
между**

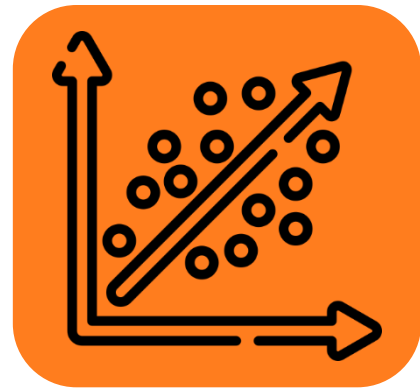
- Генерация
- Обучение представлений (word2vec)
- Обучение с подкреплением



**Обучение
без учителя**

- Кластеризация
- Детекция аномалий
- Сокращение размерности

Обучение с учителем



Регрессия - предсказываем число
Пример: предсказание цены квартиры



Классификация - предсказываем класс
Пример: распознавание лица



Ранжирование - сортировка объектов
Пример: запрос в поисковике



Рекомендации

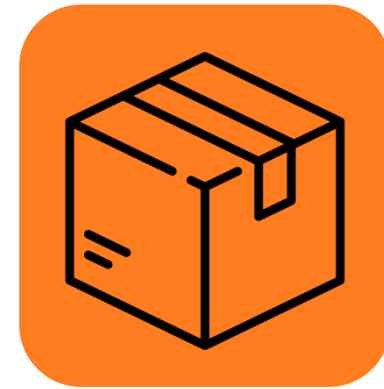
Задача построения рекомендаций



Пользователи

Клиенты, покупатели

- Активные агенты, совершают взаимодействие с предметами
- Зачастую реальные люди, например, покупатели



Предметы

Объекты, товары

- Пассивные объекты, с которыми пользователи взаимодействуют
- Товары, фильмы, аудиозаписи, тексты и т.д.

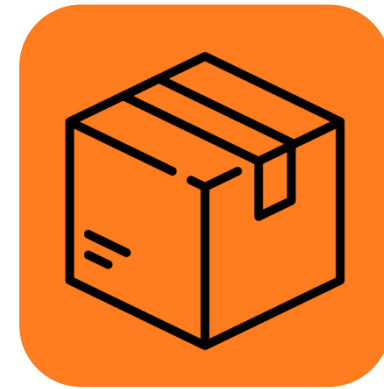
Задача построения рекомендаций



Пользователи

Клиенты, покупатели

- Активные агенты, совершают взаимодействие с предметами
- Зачастую реальные люди, например, покупатели



Предметы

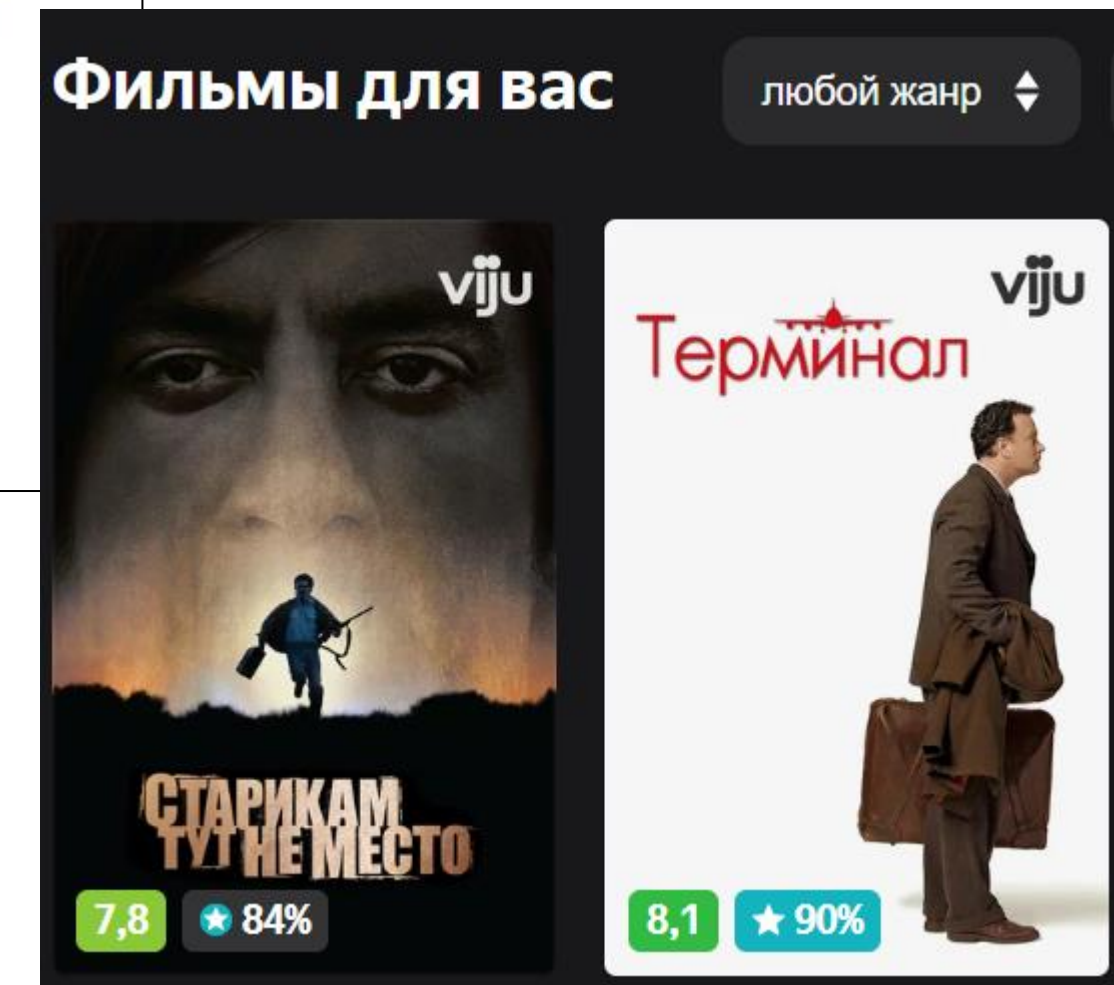
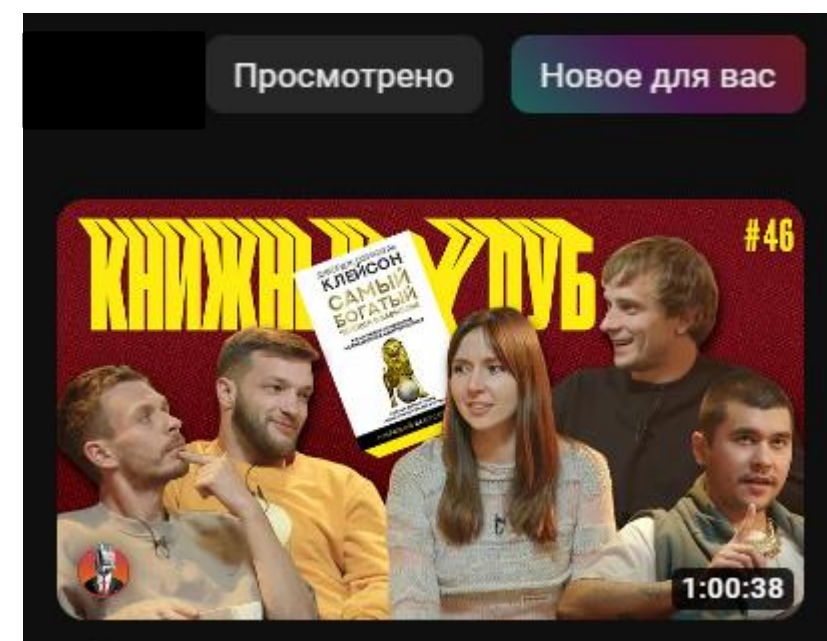
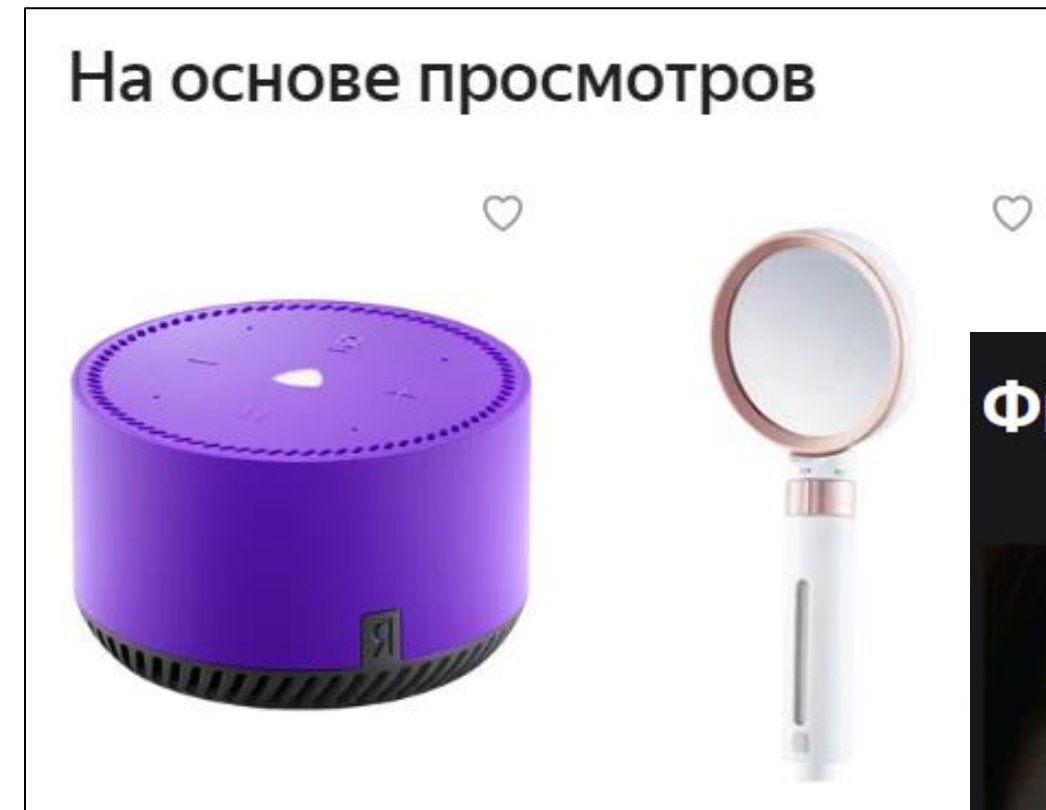
Объекты, товары

- Пассивные объекты, с которыми пользователи взаимодействуют
- Товары, фильмы, аудиозаписи, тексты и т.д.

Задача рекомендательной системы – подобрать топ K релевантных предметов для пользователя

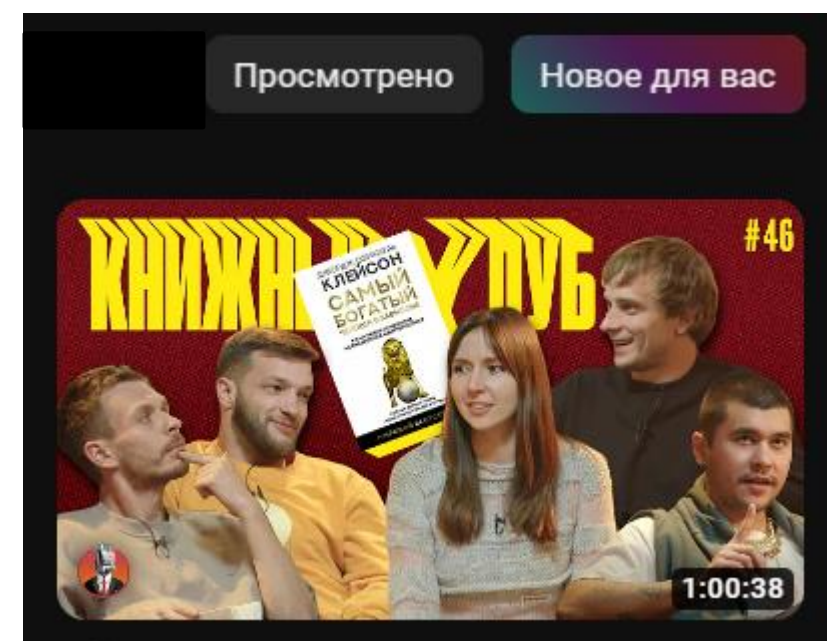
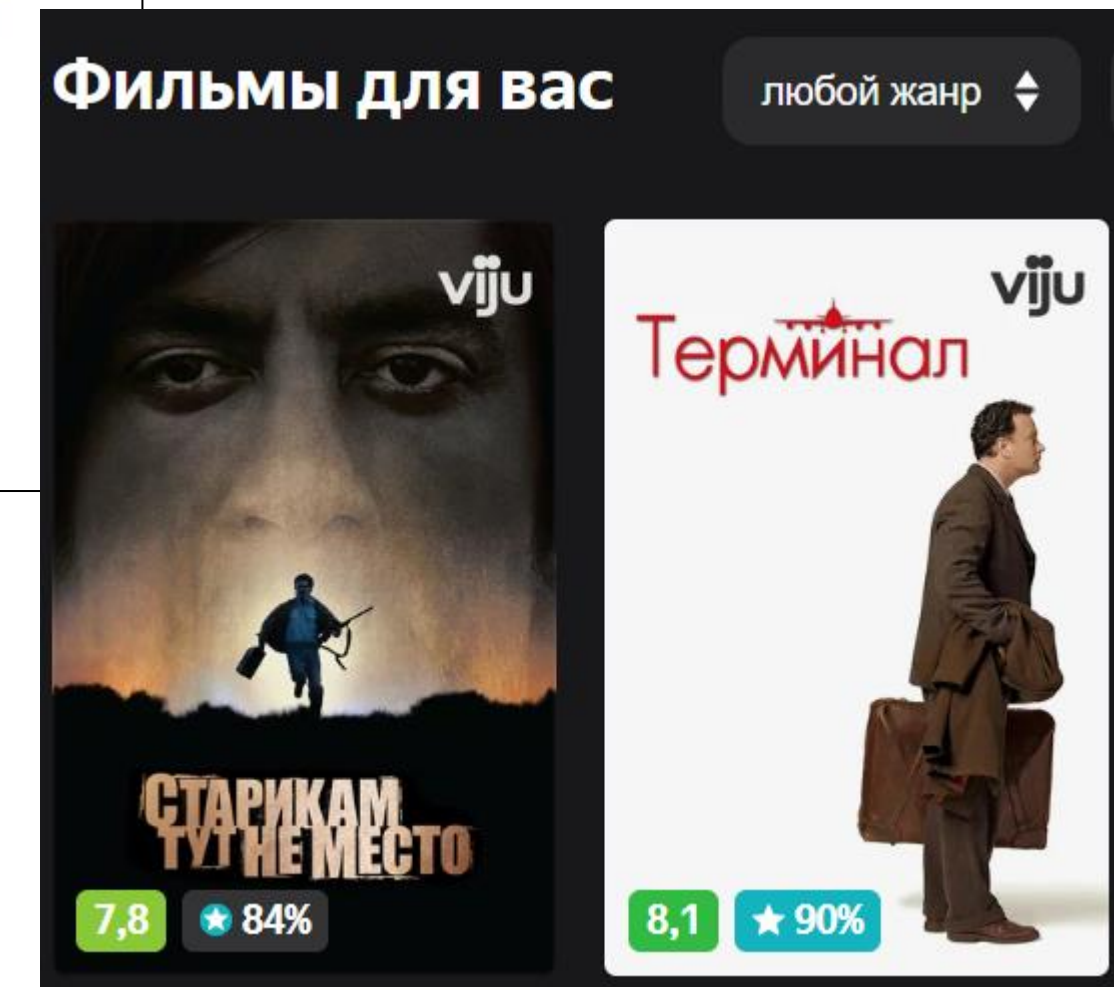
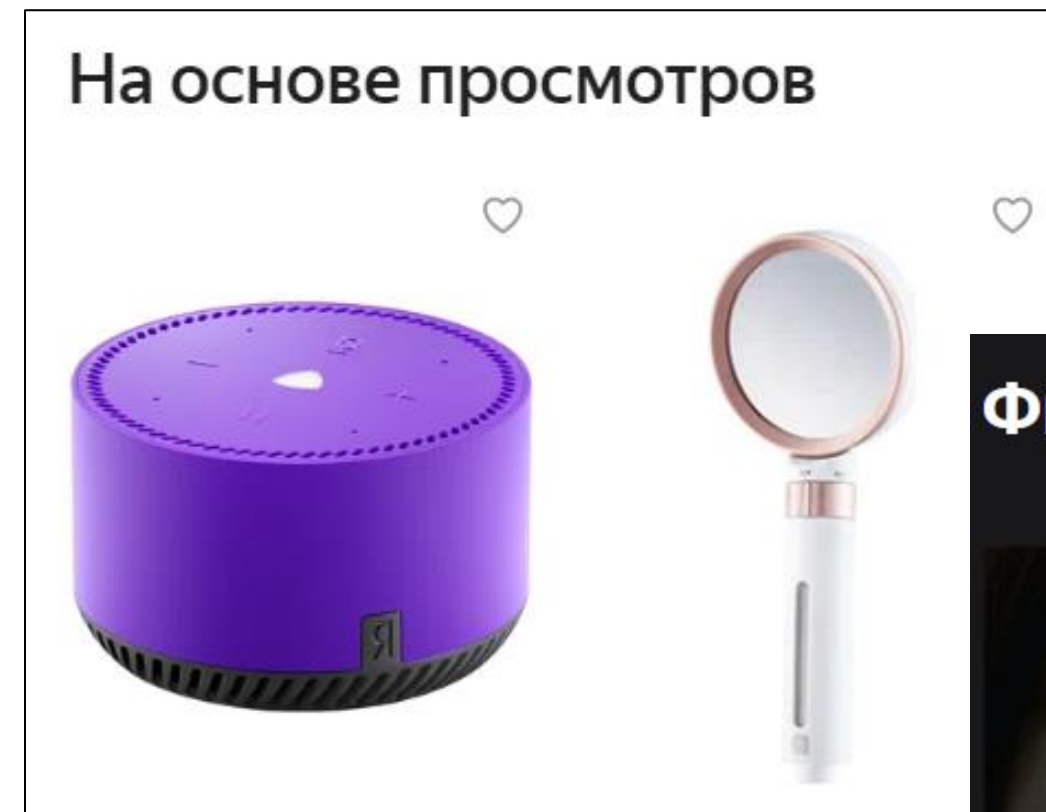
Где применяются рекомендательные системы

- Онлайн маркетплейсы
- Онлайн кинотеатры
- Видеохостинги



Где применяются рекомендательные системы

- Онлайн маркетплейсы
- Онлайн кинотеатры
- Видеохостинги

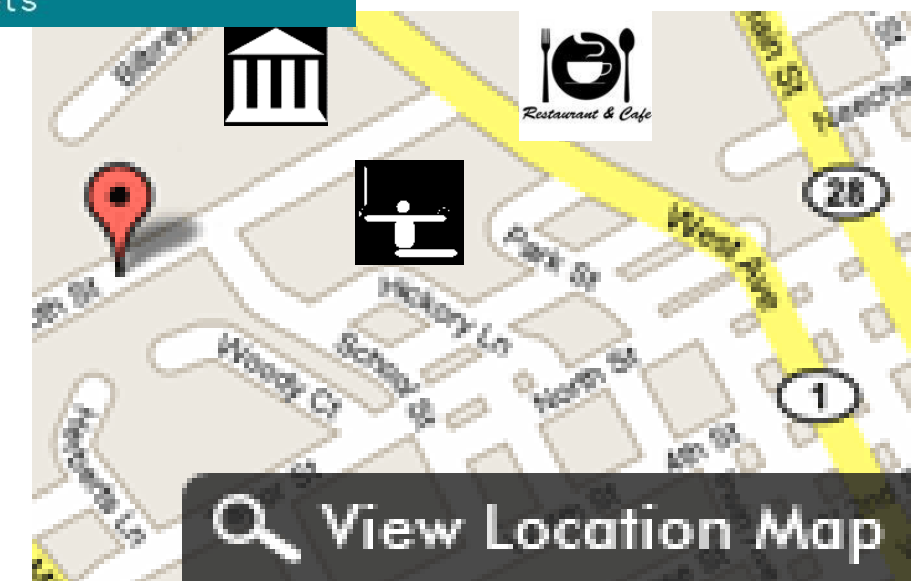
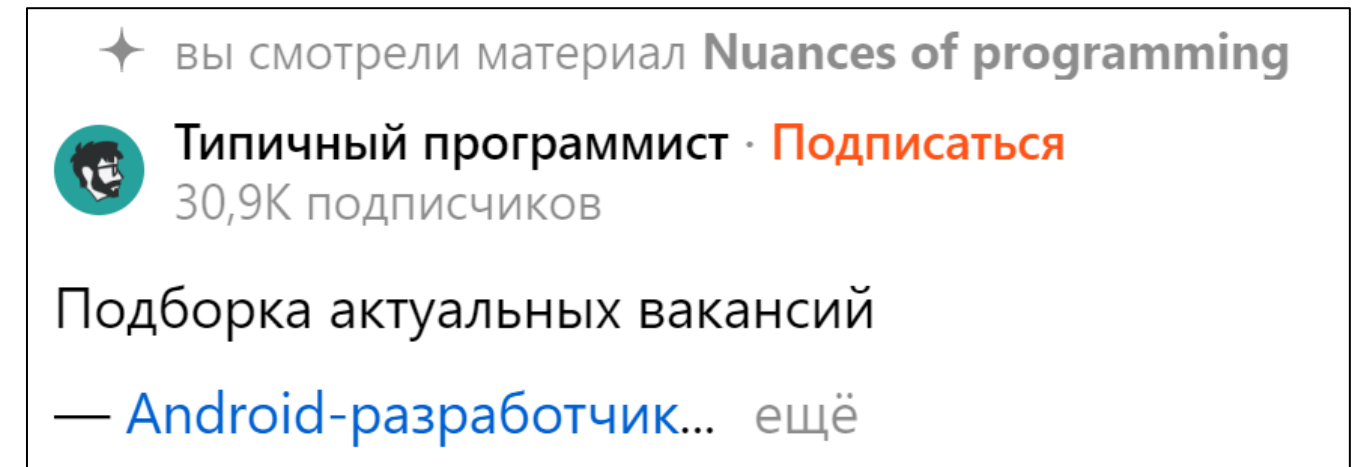


А где еще
применяются
рек. системы ?



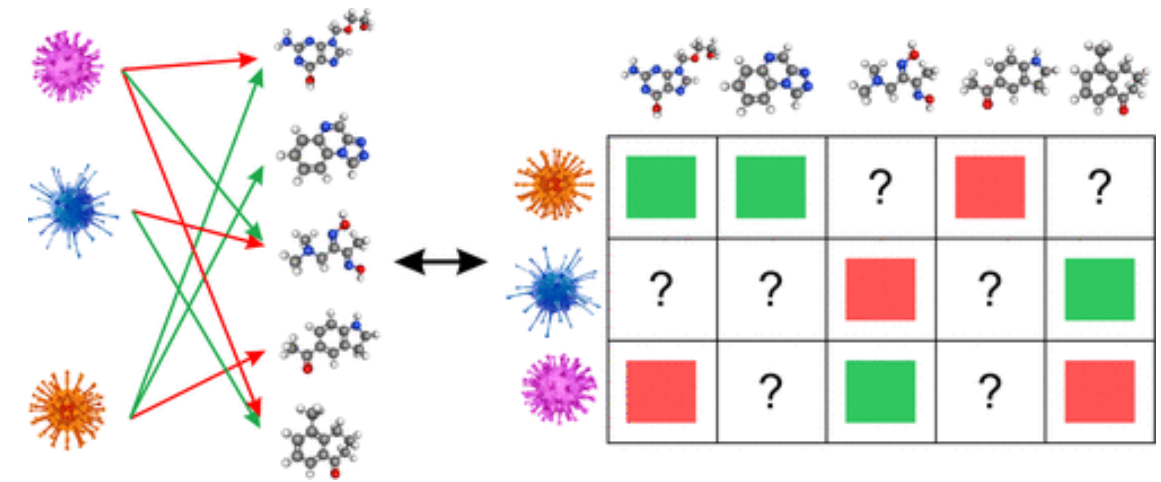
Где применяются рекомендательные системы

- Онлайн маркетплейсы
- Онлайн кинотеатры
- Видеохостинги
- Тестовые рекомендации
- Рекомендации на основе графов
- Рекомендации на основе геолокаций

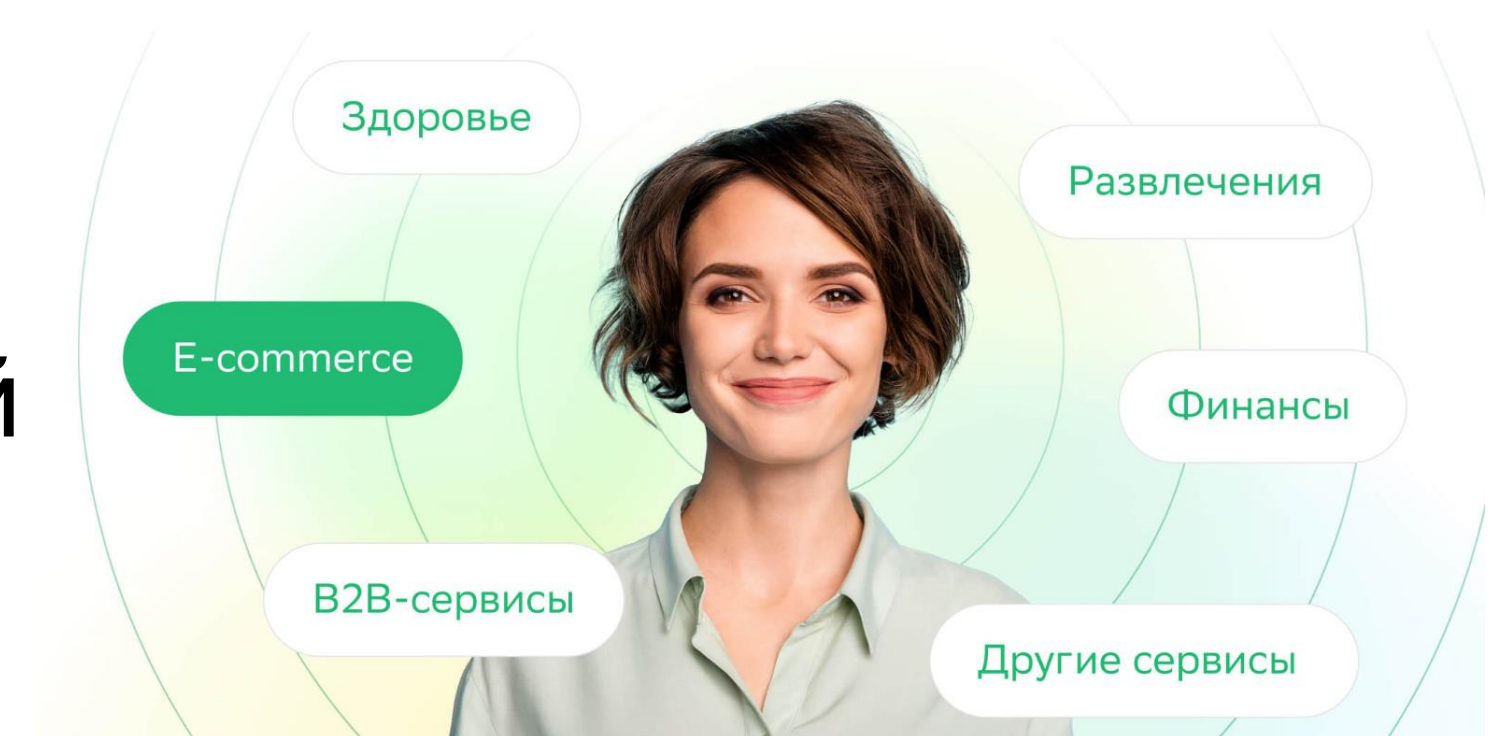


Где применяются рекомендательные системы

- Онлайн маркетплейсы
- Онлайн кинотеатры
- Видеохостинги
- Тестовые рекомендации
- Рекомендации на основе графов
- Рекомендации на основе геолокаций
- Создание лекарств
- Рекомендации внутри экосистем



1



Задача рекомендаций

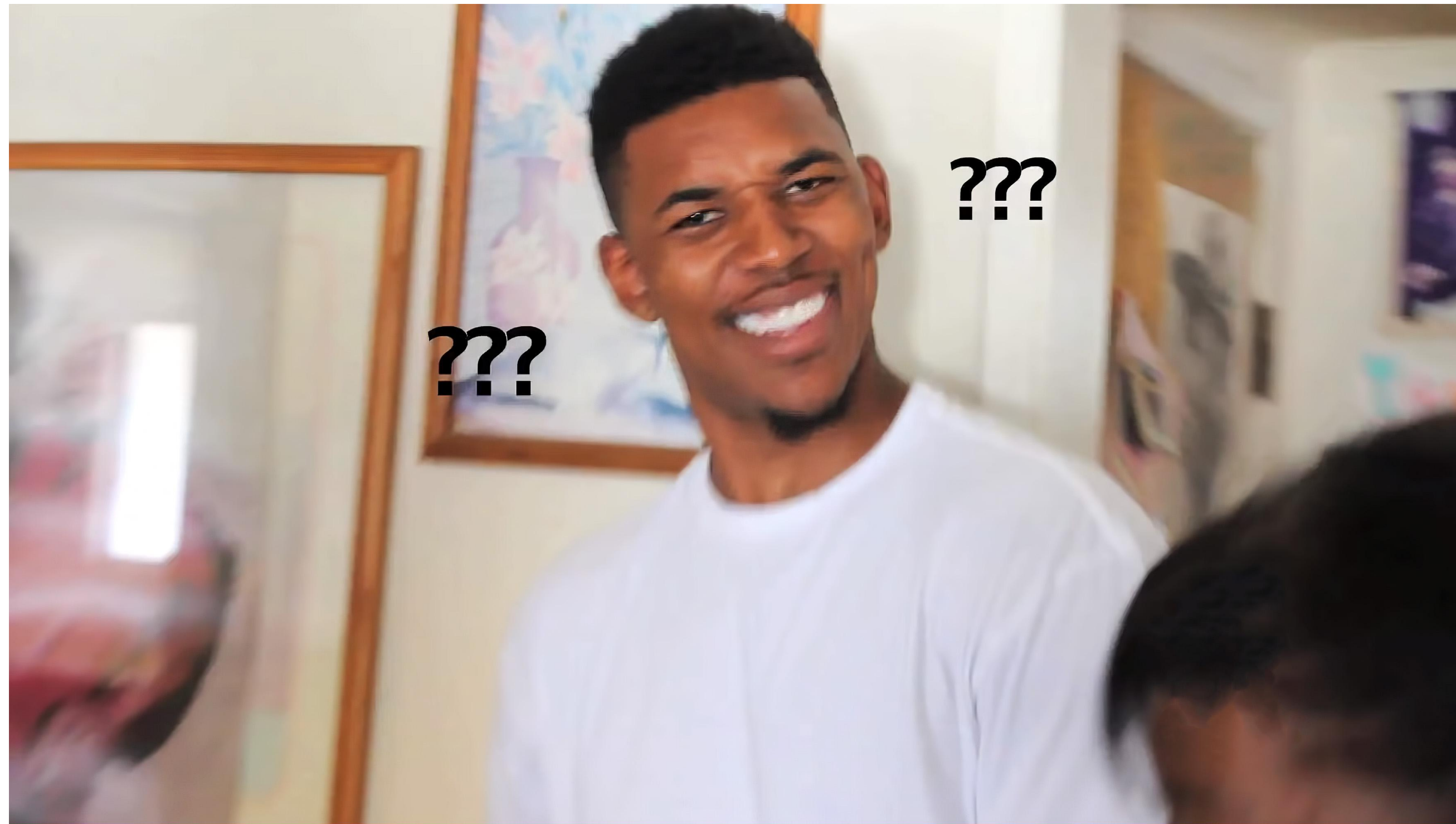
Задача рекомендаций – подобрать топ K предметов для пользователя по оценке релевантности.

Оценка релевантности – мера того, насколько товар подходит пользователю.

Подходы к решению:

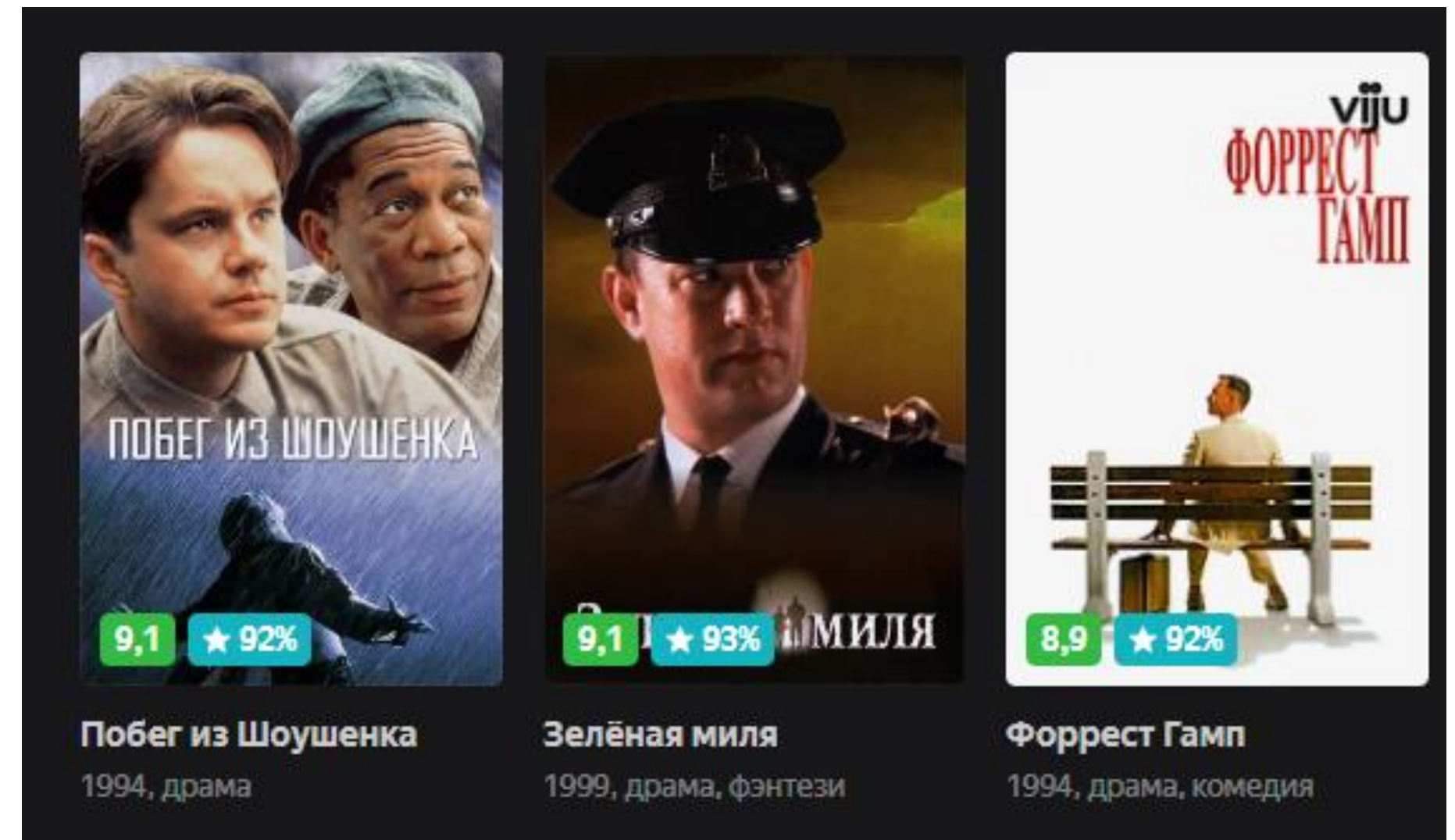
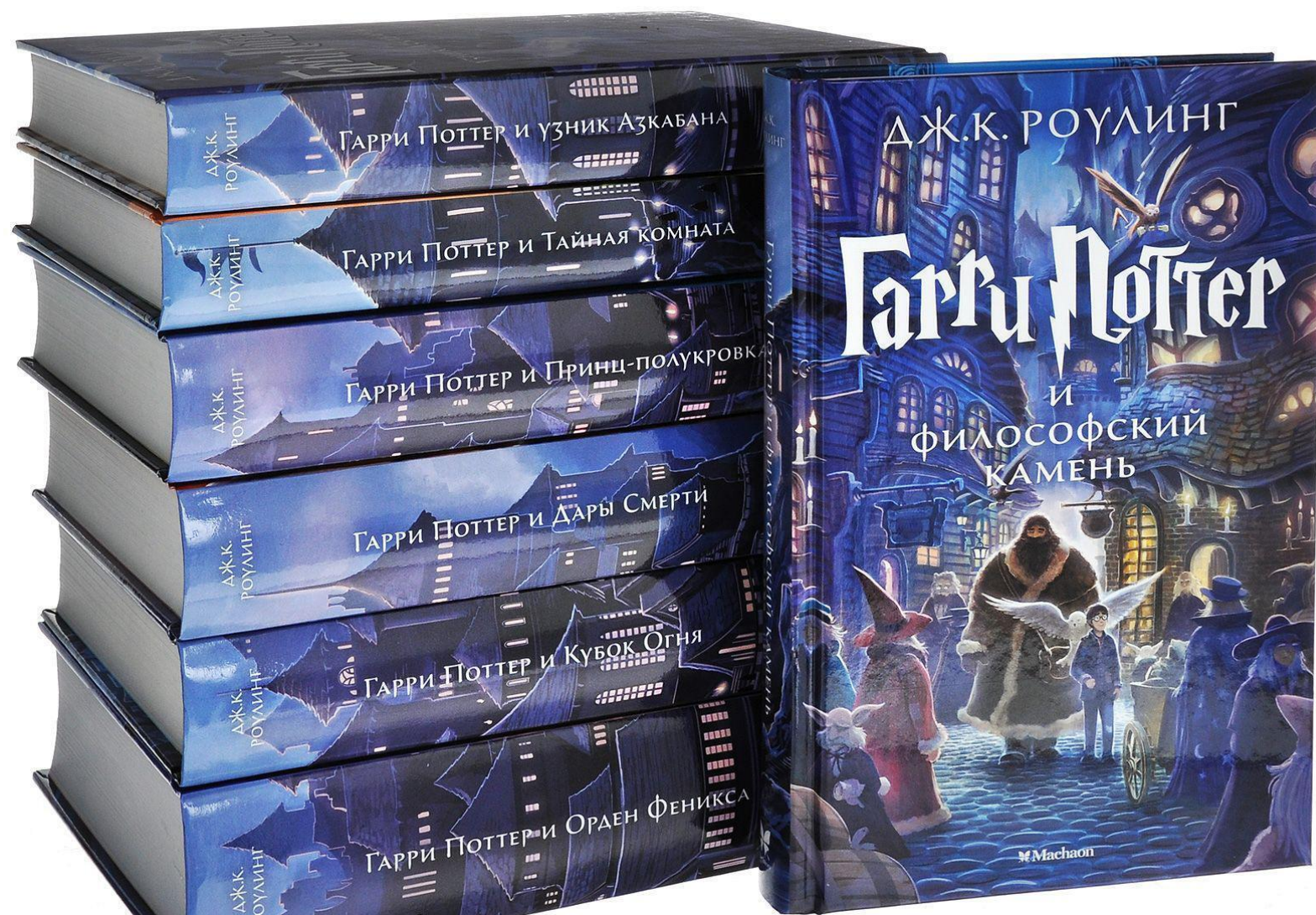
- **Задача регрессии:** предсказываем оценку релевантности
- **Задача бинарной классификации:** метка 1 – подходит, метка 0 – не подходит
- **Задача ранжирования:** топ 1 – самый релевантный предмет

А как мы можем делать ранжирование



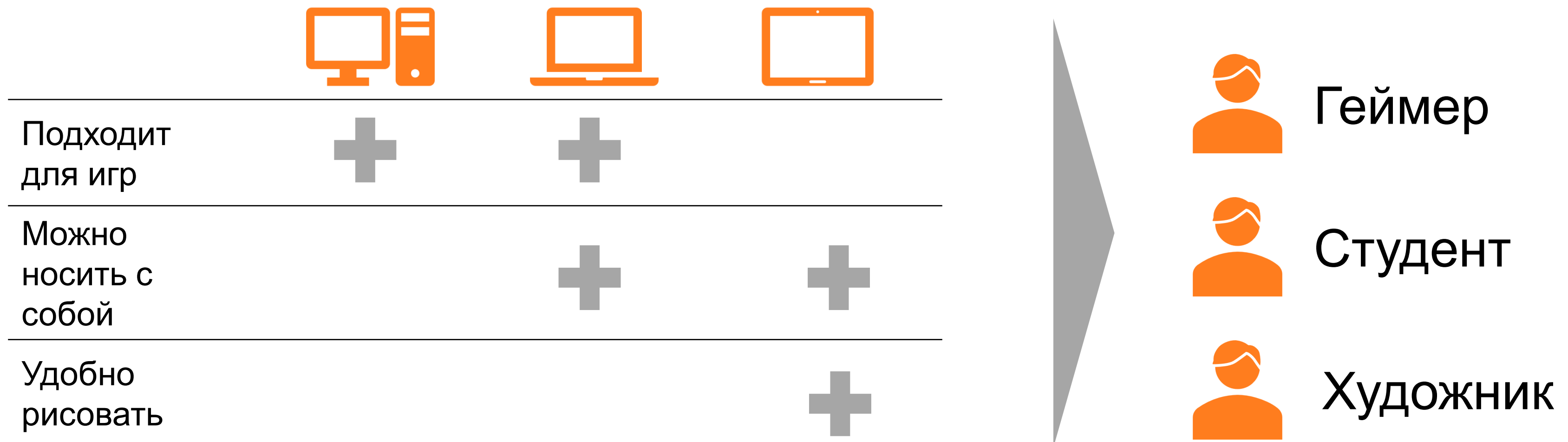
Подходы к построению рекомендаций

- Рекомендации на основе популярности



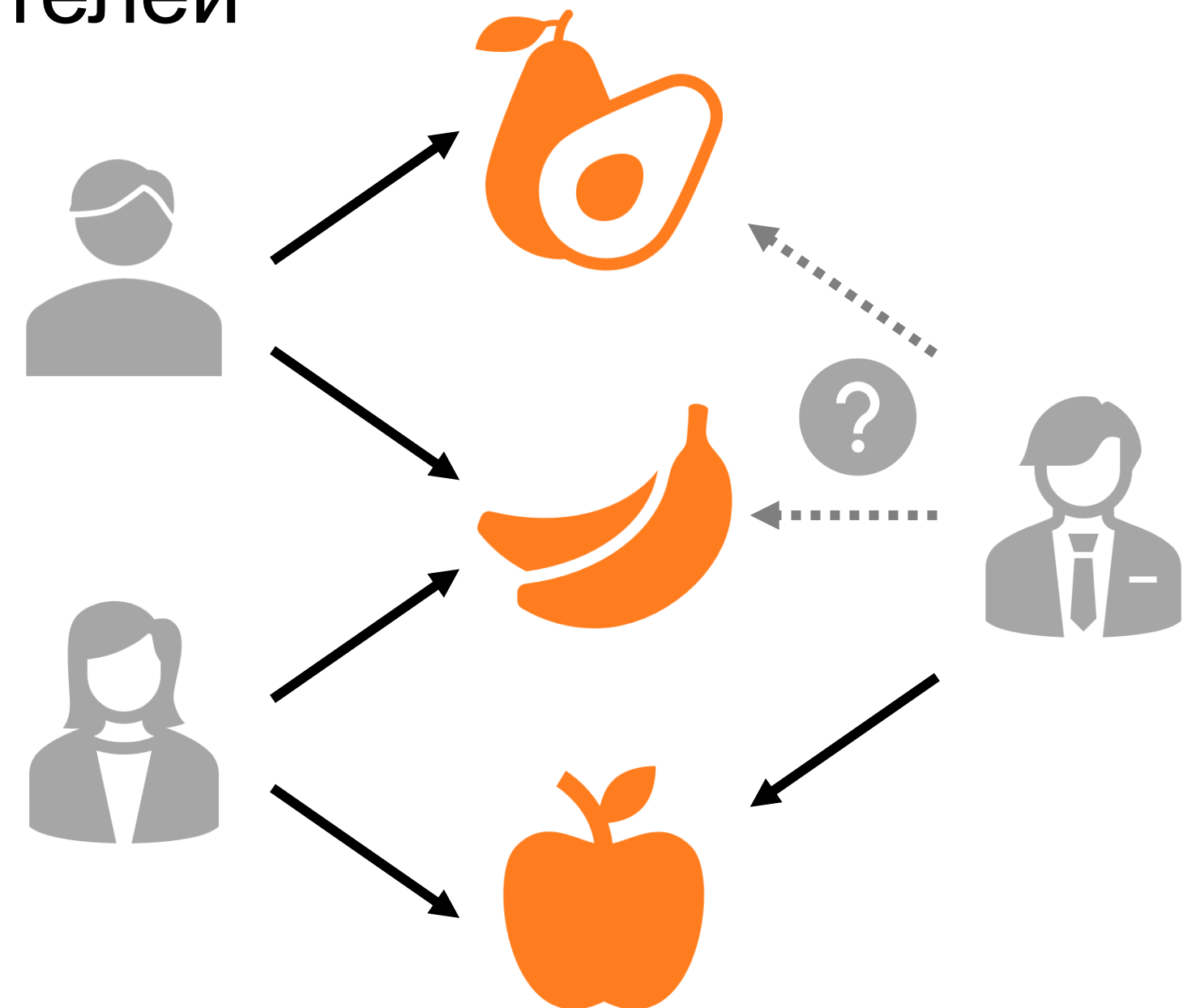
Подходы к построению рекомендаций

- Рекомендации на основе популярности
- Content Based подход: на основе характеристик товаров и пользователей



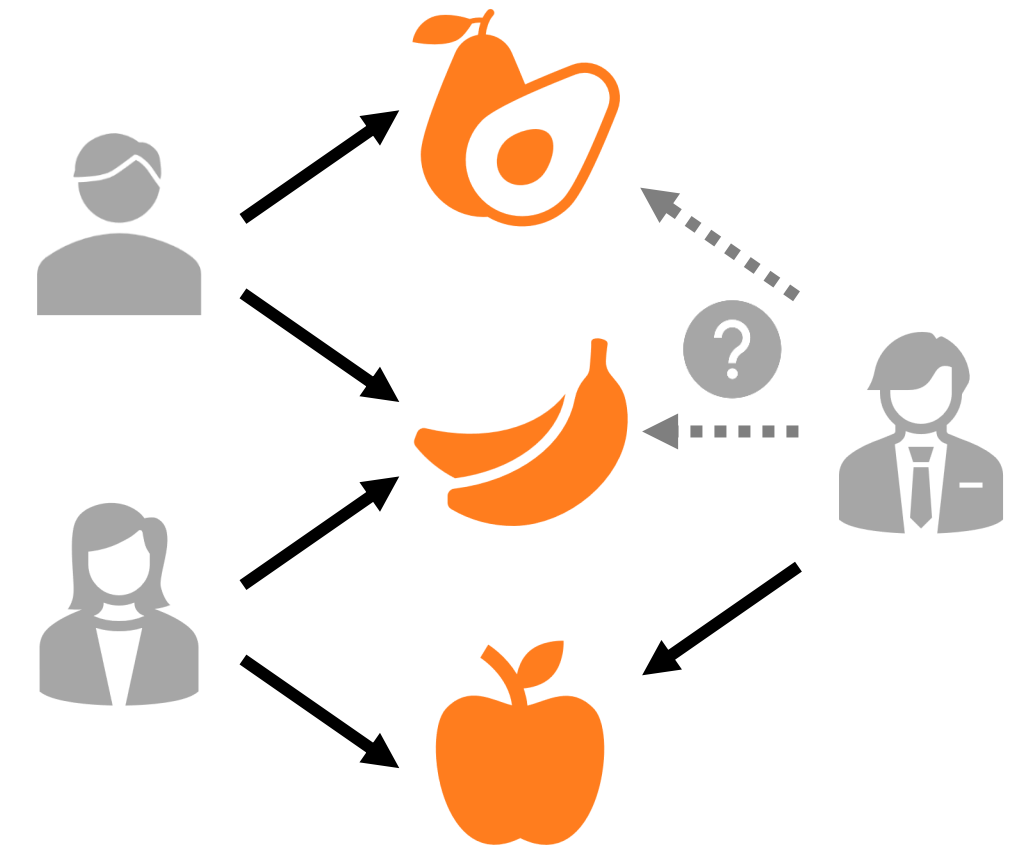
Подходы к построению рекомендаций




- Рекомендации на основе популярности
- Content Based подход: на основе характеристик товаров и пользователей
- Коллаборативная фильтрация: На основании взаимодействий



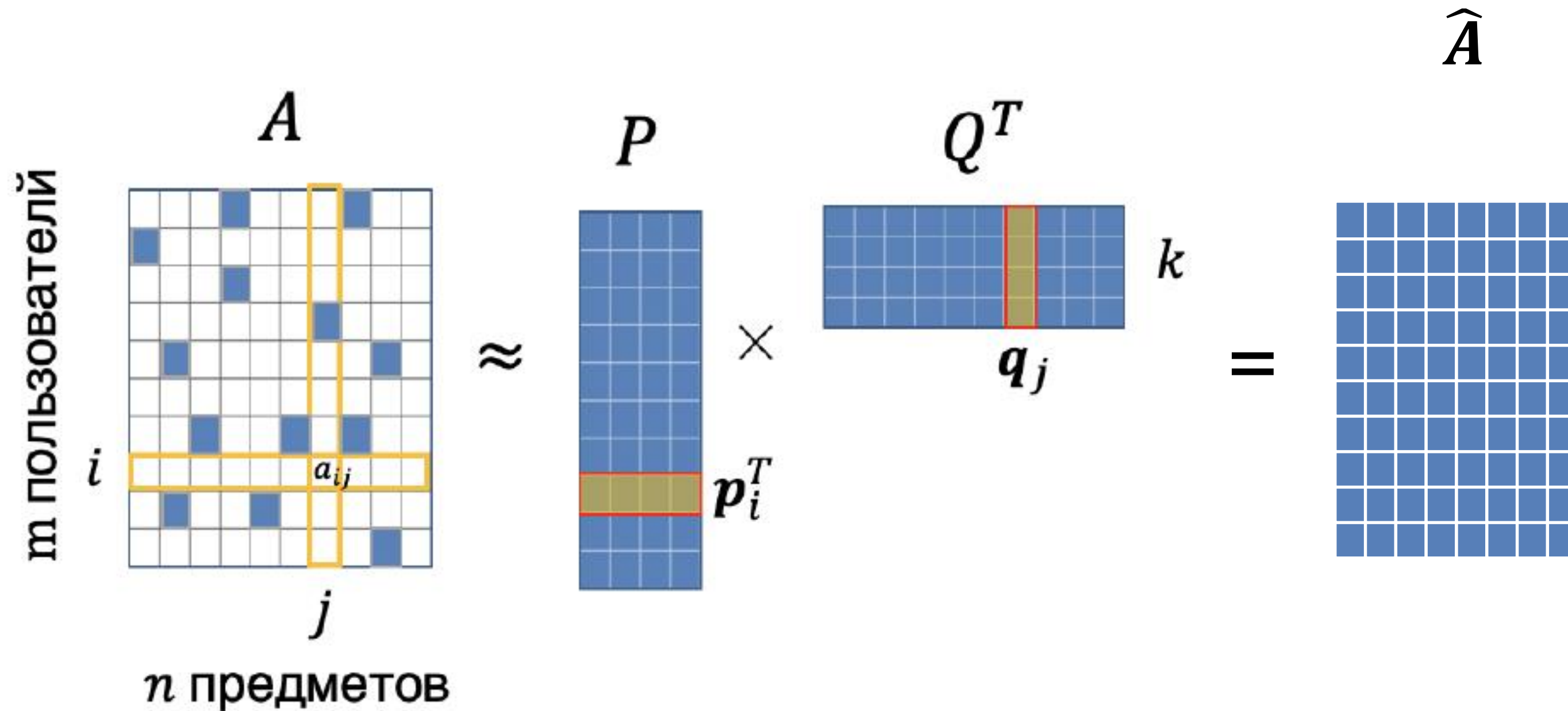
Подходы к построению рекомендаций

- Рекомендации на основе популярности
- Content Based подход: на основе характеристик товаров и пользователей
- Коллаборативная фильтрация: На основании взаимодействий
- **Смешанные варианты**



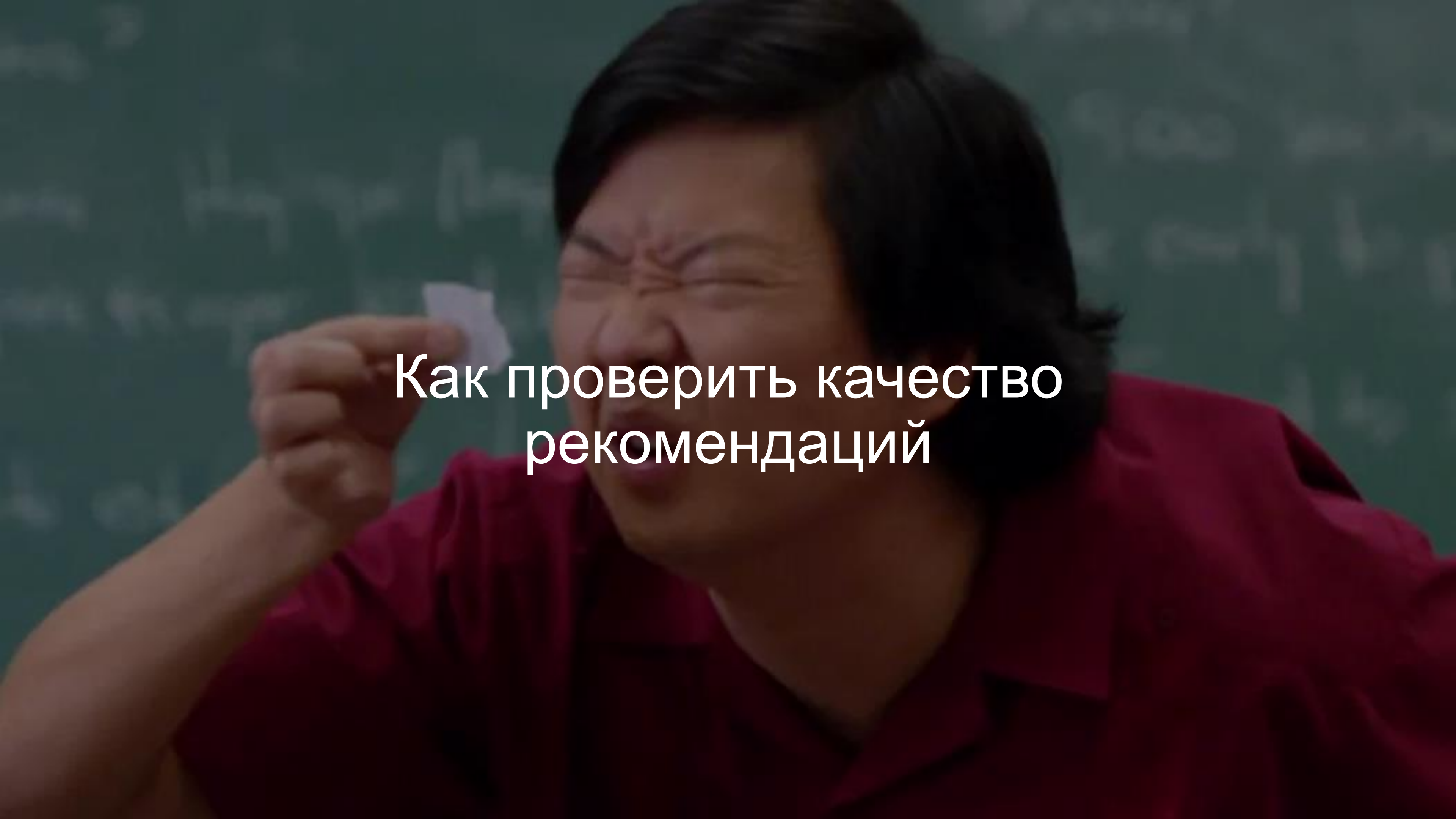
			
Подходит для игр	+	+	
Можно носить с собой		+	+
Удобно рисовать			+

Коллаборативная фильтрация



A – Исходная таблица с рейтингами

P, Q – таблицы представлений клиентов и объектов, которые мы подбираем

A person with dark hair, wearing a red shirt, is shown in a close-up shot. They have a frustrated or angry expression, with their eyes squinted and their mouth slightly open. They are holding a small, white, rectangular piece of paper in their right hand, which they appear to be tearing or crumpling. The background is a textured, light blue-green wall. The overall lighting is somewhat dim, and the image has a slightly grainy quality.

Как проверить качество
рекомендаций

Метрика качества рекомендаций

$$\text{HitRate (HR)} = \frac{n_{\text{relevant}}}{N},$$

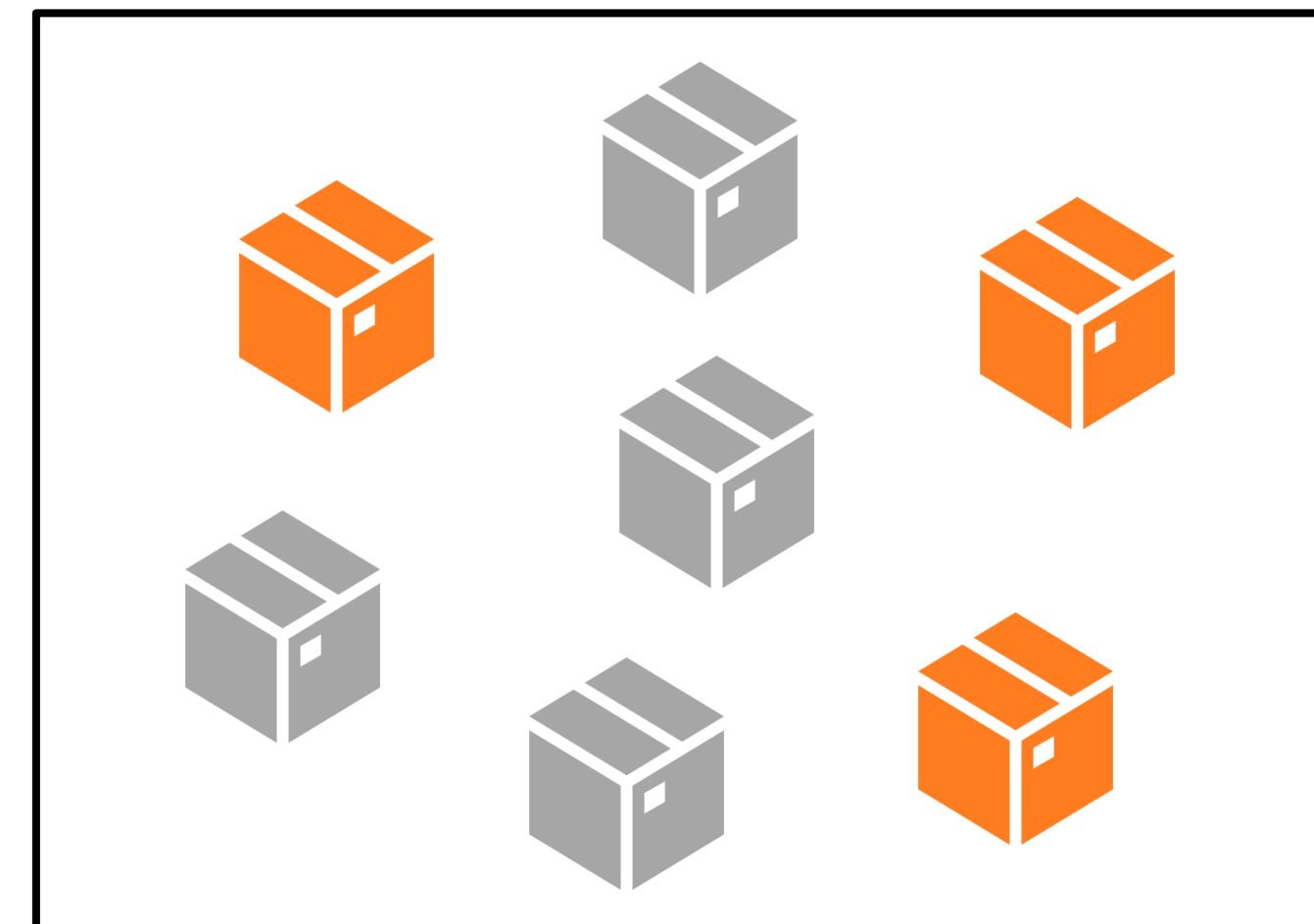
N – общее количество рекомендованных предметов

n_{relevant} – количество релевантных рекомендованных предметов

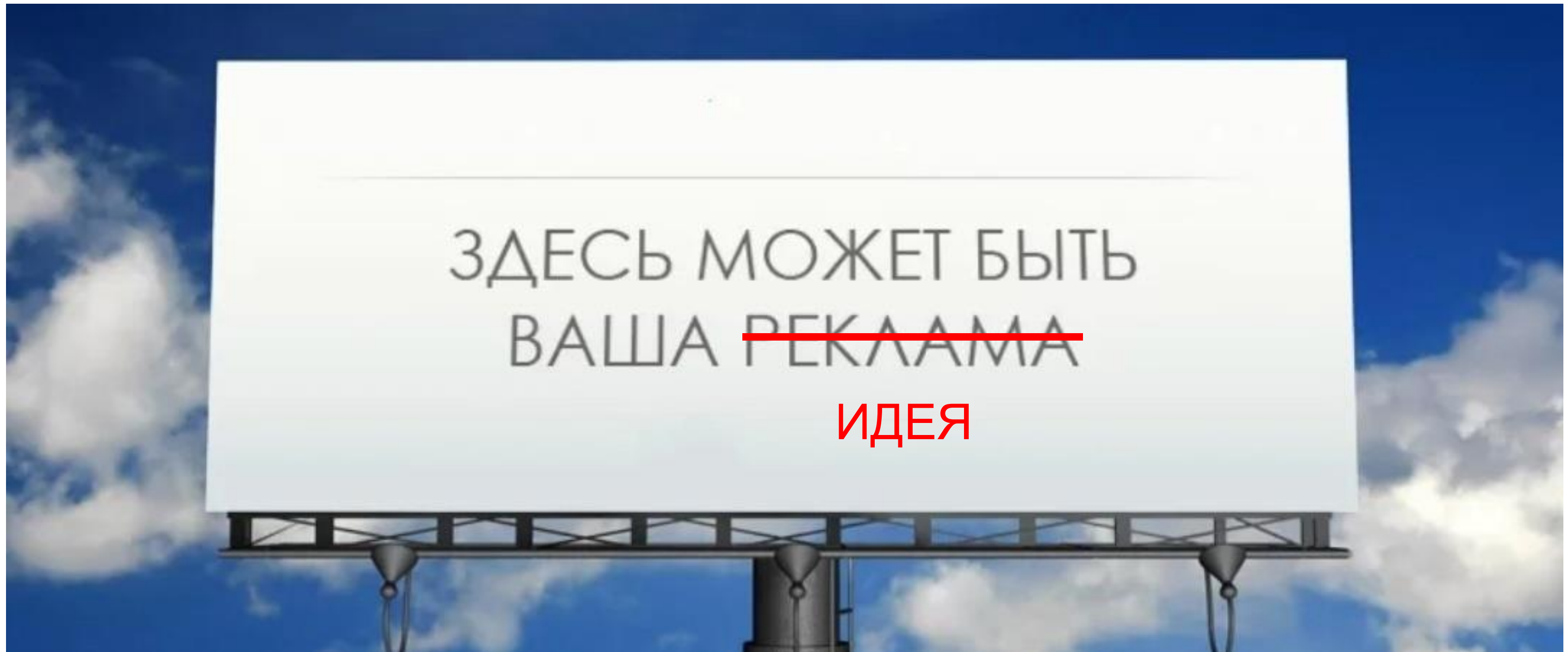


Релевантные
рекомендации

Все рекомендации



С какими проблемами можем столкнуться

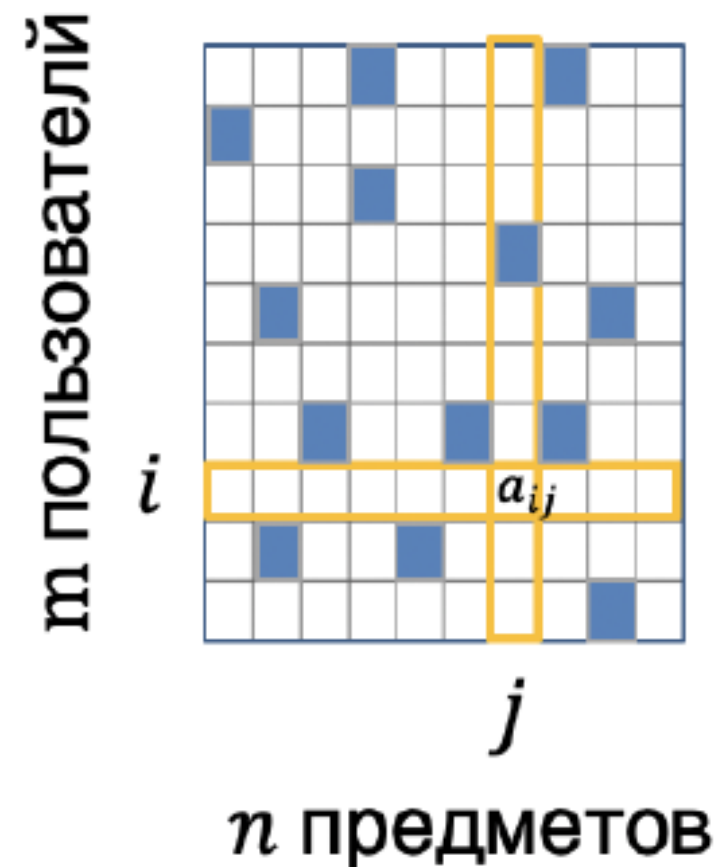


С какими проблемами можем столкнуться

- Проблема новых пользователей
- Проблема длинных хвостов:
 - проблема с памятью
 - проблема со скоростью работы
- Проблема оценки качества рекомендаций

Проблема холодного старта

Коллаборативная фильтрация плохо справляется с **НОВЫМИ ПОЛЬЗОВАТЕЛЯМИ И ПРЕДМЕТАМИ**



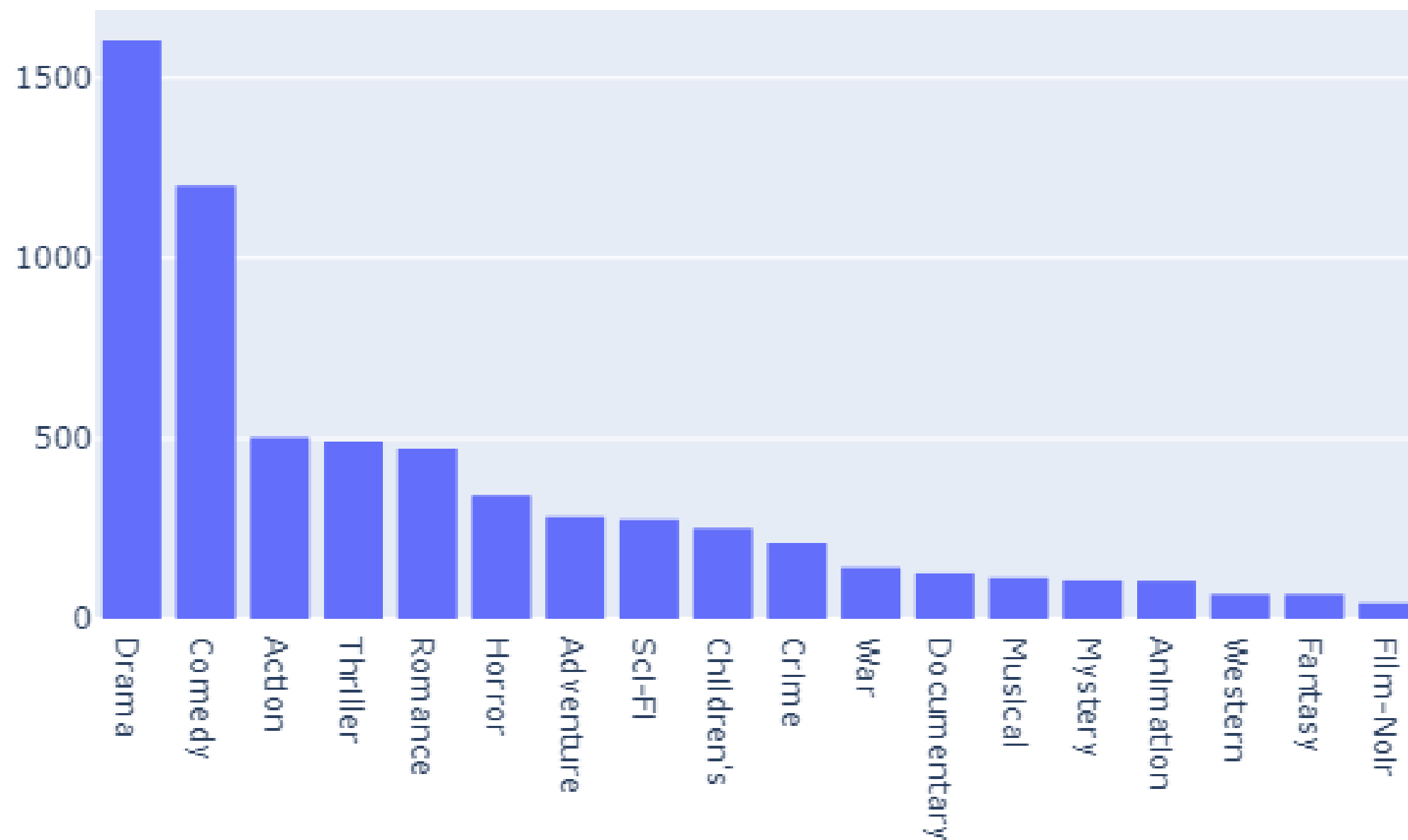
Решения:

- Использовать content-based подход: составлять анкеты для новых пользователей и рекомендовать по вкусам
- Увеличивать оценку для новых предметов, чтобы они чаще попадали в рекомендации
- Использовать смешанные модели, например, Light FM

Проблема длинных хвостов

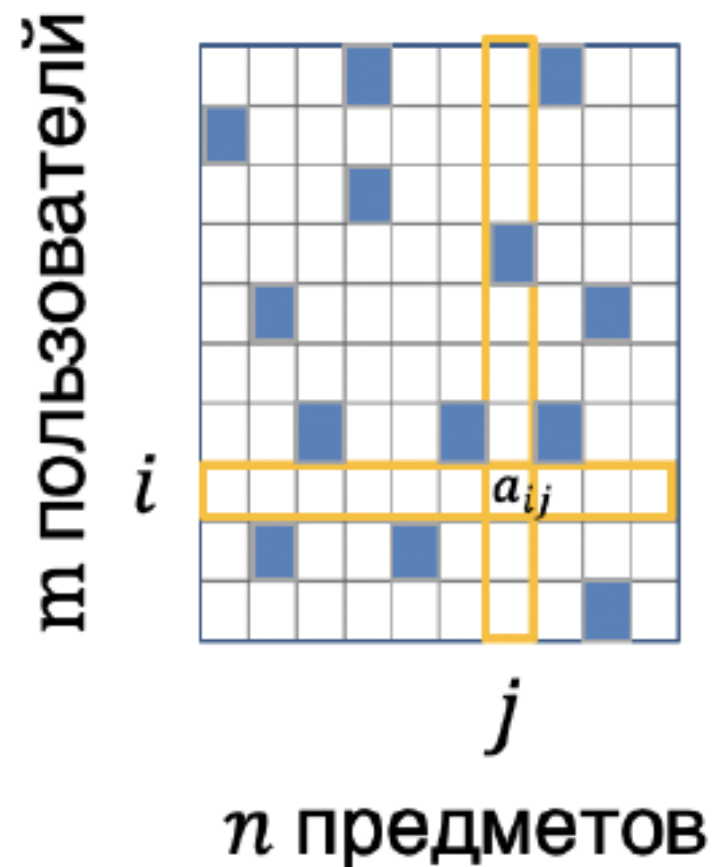
Обычно встречается большое количество нишевых товаров, которые тоже нужно рекомендовать

Гистограмма распределения фильмов по жанрам в выборке MovieLens1M



Проблема длинных хвостов

Таблицы предметы-пользователи получаются слишком разреженные (менее 5% данных заполнены)



Сложности:

- Проблемы с памятью для хранения данных.
- В реальных задачах часто необходимо строить **распределенную систему**.
- Проблемы в скорости подбора рекомендаций.

Проблема в скорости работы

При каждом вызове сети необходимо сравнивать тысячи/миллионы предметов.

Решения:

Заранее рассчитывать рекомендации, когда это возможно *возможно не во всех задачах*

Применять простые, но быстрые модели *потеря в качестве модели*

Применять двухэтапную модель рекомендаций

Двухэтапная модель факторизации

Этап 1: Простая быстрая

Пример модели:
SVD

Требование к модели:
скорость работы

Задача:
отобрать из всех
предметов ограниченное
число лучших (50 -
1000)

Этап 2: Медленная точная модель

Пример модели:
CatBoost

Требование к модели:
Качество работы

Задача:
переранжировать лучшие
предметы, отобранные
простой моделью, и выдать
финальные рекомендации

Оценка эффективности рекомендаций

Хорошая модель рекомендаций с точки зрения метрик рекомендаций (NR) может быть плоха для бизнеса с точки зрения прибыли

Проблема: метрики бизнеса сложно оценить на краткосрочном горизонте

Идея: использовать прокси-метрики, тесно связанные с бизнес метриками:

- Время на платформе
- Количество кликов
- Количество товаров в корзине

Оценка эффективности рекомендаций

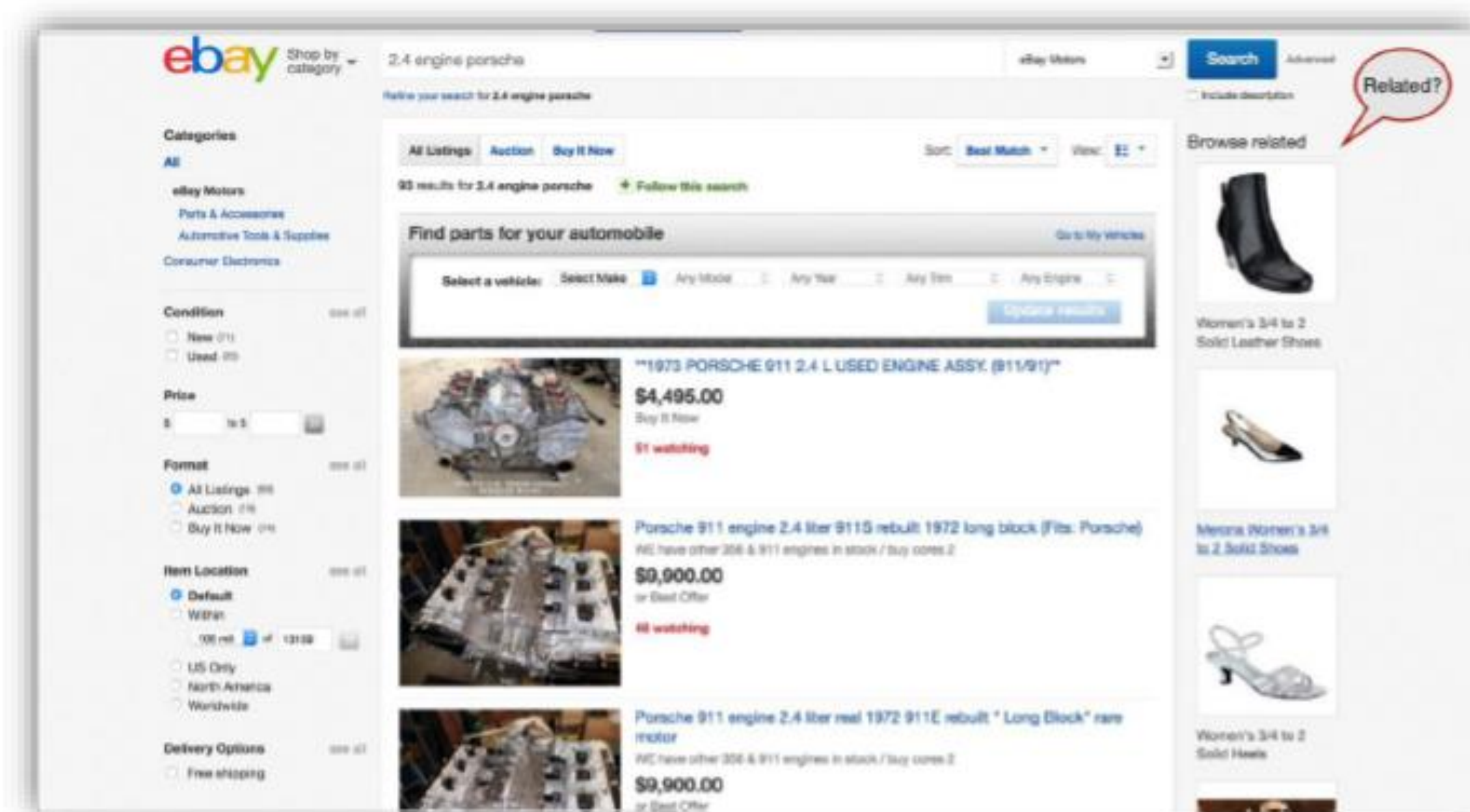
Помимо точности рекомендаций бизнесу важно большое количество других метрик.

Хорошие рекомендации должны:

- Быть разнообразными.
- Регулярно показывать новые предметы, не рекомендованные до этого.
- Быть неожиданными, уметь удивить пользователя.
- Покрывать потребности пользователя.

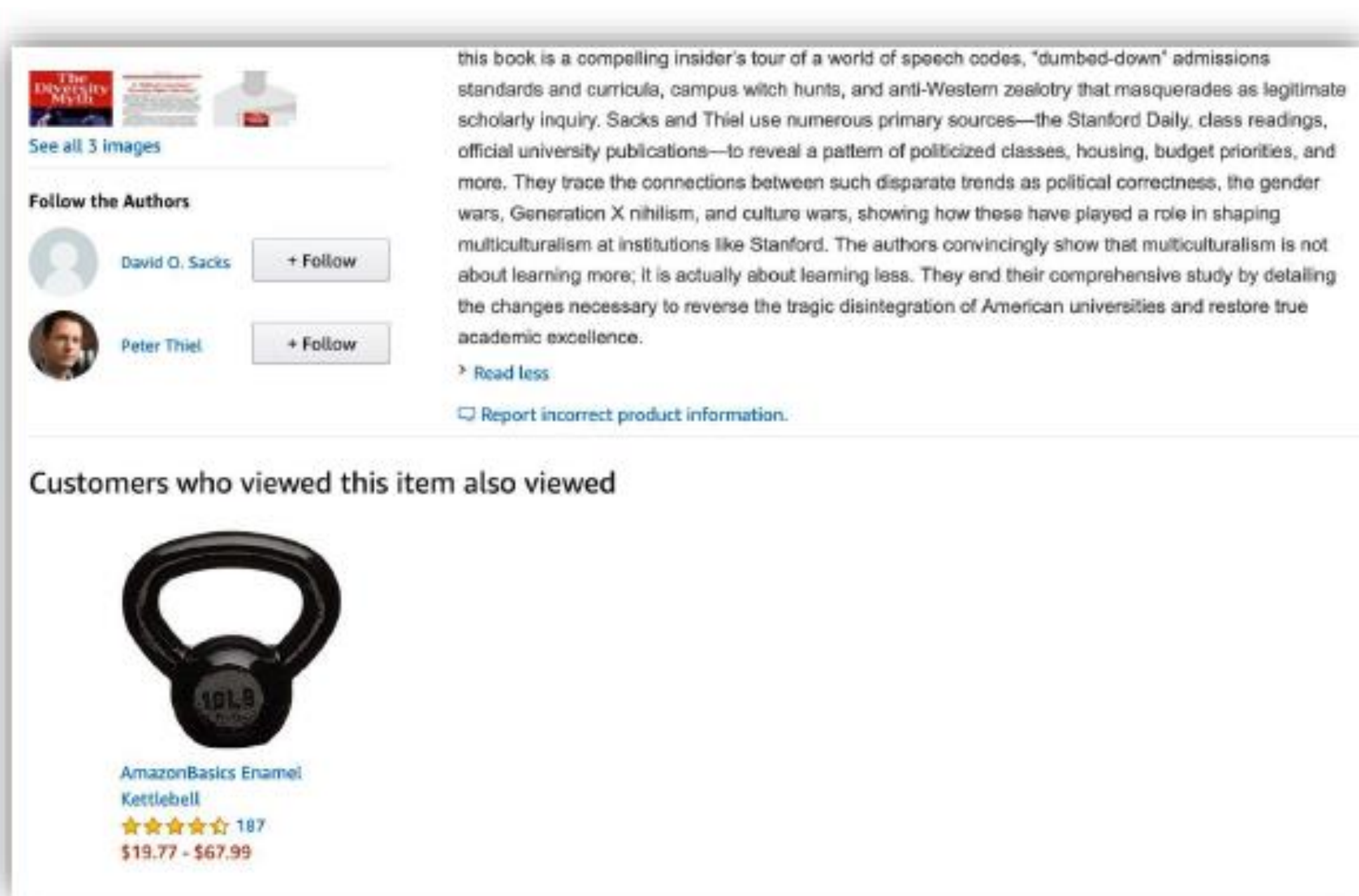
Примеры неудачных рекомендаций

Ищете запчасти для автомобилей? Вам также нужны женские туфли



Примеры неудачных рекомендаций

Если вам нравится книга Diversity Myth Питера Тиля, то вам может понравиться и гиря




The screenshot shows an Amazon product page for the book "Diversity Myth" by David O. Sacks and Peter Thiel. The page includes a book cover, author information, a description, and a recommendation for a kettlebell.

Follow the Authors

- David O. Sacks [+ Follow](#)
- Peter Thiel [+ Follow](#)

Customers who viewed this item also viewed


AmazonBasics Enamel Kettlebell
★★★★☆ 187
\$19.77 - \$67.99

this book is a compelling insider's tour of a world of speech codes, "dumbed-down" admissions standards and curricula, campus witch hunts, and anti-Western zealotry that masquerades as legitimate scholarly inquiry. Sacks and Thiel use numerous primary sources—the Stanford Daily, class readings, official university publications—to reveal a pattern of politicized classes, housing, budget priorities, and more. They trace the connections between such disparate trends as political correctness, the gender wars, Generation X nihilism, and culture wars, showing how these have played a role in shaping multiculturalism at institutions like Stanford. The authors convincingly show that multiculturalism is not about learning more; it is actually about learning less. They end their comprehensive study by detailing the changes necessary to reverse the tragic disintegration of American universities and restore true academic excellence.

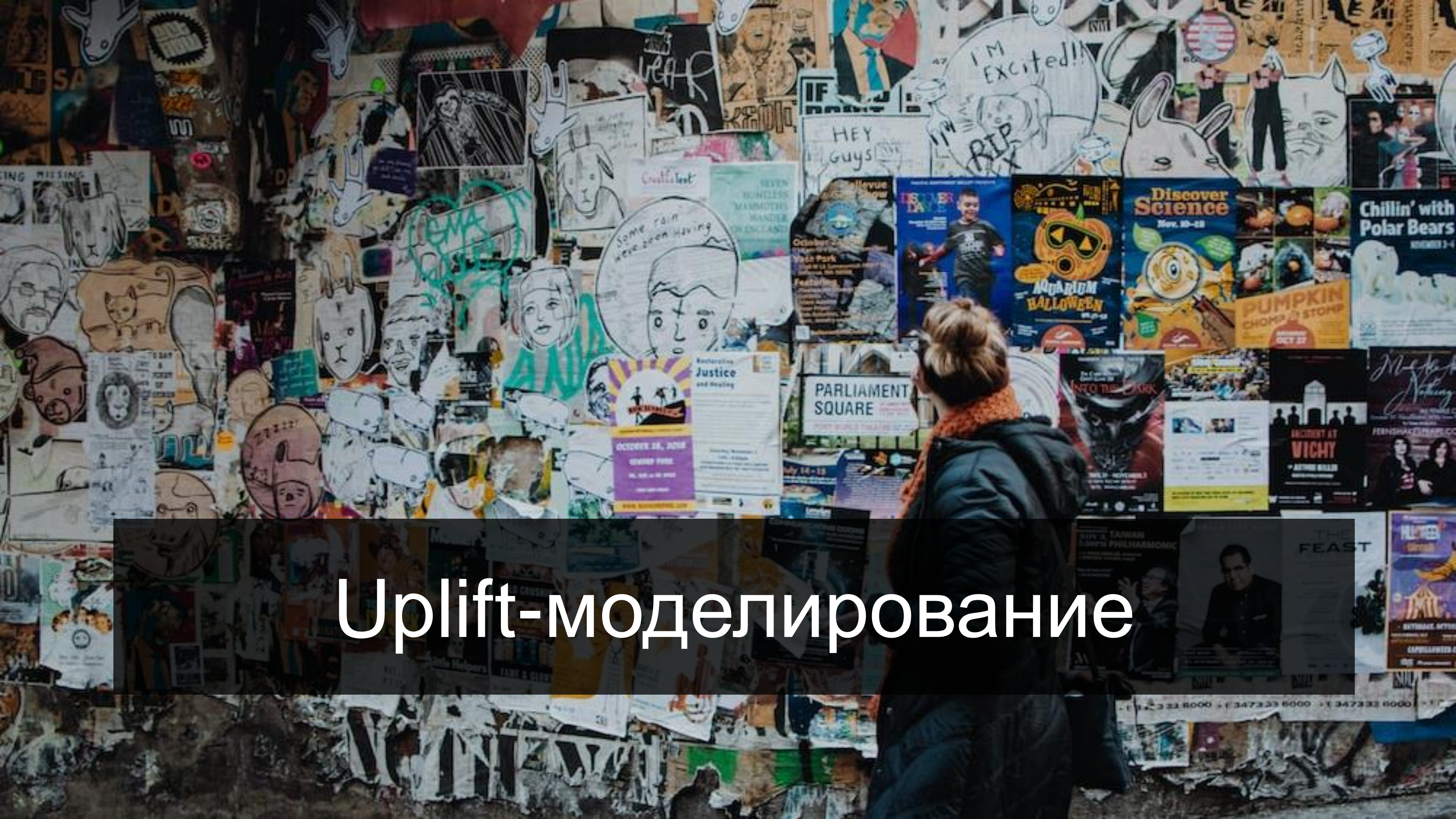
[Read less](#)

[Report incorrect product information.](#)

Inner loop problem



Решение: следить за разнообразием рекомендаций и искусственно повышать релевантность более редких предметов



Uplift-моделирование

Бизнес проблема

Пример

Мы продаем телефоны в интернет-магазине

Есть ограниченный бюджет на рекламу

Задача: Кому именно показывать рекламу?

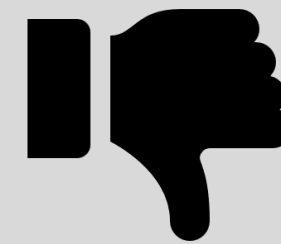
Типы клиентов

Взаимодействие: да
Таргет: нет



Спящие
котики

Взаимодействие: нет
Таргет: да



Потерянные

Взаимодействие: да
Таргет: нет

Взаимодействие: нет
Таргет: нет

Взаимодействие: да
Таргет: да



Лояльные

Взаимодействие: нет
Таргет: да



Убеждаемые

Взаимодействие: нет
Таргет: нет

Взаимодействие: да
Таргет: да

Задача

Найти пользователей, которые:

- **Только увидев рекламу купят телефон**
- **Без рекламы не купят телефон**

Эта задача – не задача определения того, кто купит

То есть – это не задача классификации

Бизнес проблема

Где может применяться данная задача?



Бизнес проблема

Где может применяться



Маркетинг



Медицина



Предвыборные компании

Основные термины

Взаимодействие

Treatment, воздействие

Мы пытаемся влиять на клиента

Пример:

- Звонок
- Показ рекламы
- Выдача таблетки

Целевое действие

Target, таргет

Клиент совершает нужное нам действие после влияния

Пример:

- Покупка
- Голос на выборах
- Выздоровление

Постановка задачи

P_1 = Вероятность:
было воздействие и купит



P_0 = Вероятность:
не было воздействие и купит



$$\text{Uplift (casual inference)} = P_1 - P_0$$

Как можно решить такую задачу





Кейс с автомобилем

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?
2. Какую задачу машинного обучения мы решаем?

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?
2. Какую задачу машинного обучения мы решаем?
3. Как мы поймем, что задача решена успешно?

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?
2. Какую задачу машинного обучения мы решаем?
3. Как мы поймем, что задача решена успешно?
4. Какие данные для обучения модели нам нужны и сколько?

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?
2. Какую задачу машинного обучения мы решаем?
3. Как мы поймем, что задача решена успешно?
4. Какие данные для обучения модели нам нужны и сколько?
5. Как мы поймем, что модель работает правильно?

Задача про справедливую цену автомобиля

Вы работаете на сайте по продаже БУ автомобилей. Ваша задача - добавить отображение **справедливой цены** автомобиля.

1. Какую бизнес-проблему мы решаем?
2. Какую задачу машинного обучения мы решаем?
3. Как мы поймем, что задача решена успешно?
4. Какие данные для обучения модели нам нужны и сколько?
5. Как мы поймем, что модель работает правильно?
6. Какие проблемы могут возникнуть при внедрении. Как их можно решить?

Задача про справедливую цену автомобиля

Вы р
- доб

1. К
2. К
3. К
4. К
5. К
6. К

р



stonks

но