

## 5. Оценка качества рекомендательных систем

# Введение



В этом уроке мы поговорим о бизнес-метриках.

# План

---

1

Метрики  
машинного  
обучения

2

Специальные  
оценки  
качества

3

Многокритериа  
льные  
рекомендации

4

Онлайн и  
оффлайн оценки  
качества

5

Бизнес-  
метрики

6

Проблема  
«пузыря»

# Метрики машинного обучения

# Метрики качества



Все метрики качества можно условно разделить на три категории:

- **Prediction Accuracy** — оценивают точность предсказываемого рейтинга
- **Decision support** — оценивают релевантность рекомендаций
- **Rank Accuracy** — оценивают качество ранжирования выдаваемых рекомендаций

# Метрики качества



Все метрики качества можно условно разделить на три категории:

- **Prediction Accuracy** — RMSE, MAE
- **Decision support** — Precision@n, Recall@n, MAP
- **Rank Accuracy** — MRR, NDCG@n

# RMSE



$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{x}_i - x_i)^2}$$

# MAE

---

$$MAE = \frac{1}{m} \sum_{i=1}^m |\hat{x}_i - x_i|$$



# Метрики ранжирования

## MAP - Mean Average Precision

MAP помимо усреднения по  $n$ , дополнительно усредняет по всем пользователям или по всем запросам ( $u$ ).

Это делается из предположения из соображения, что все пользователи или запросы равноценны.

MAP является достаточно популярной, учитывает и порядок, и количество релевантных объектов.

$$MAP@n = \frac{1}{U} \sum_{u=1}^U AvgPrecision@n(u)$$

# Метрики ранжирования

## MRR - Mean Reciprocal Rank

$rank_u$  - означает  
положение первого релевантного объекта  
для пользователя  $u$ .

Чаще используется для поисковых  
запросов, где в место пользователя  
фигурируют запросы.

$$MRR = \frac{1}{U} \sum_{u=1}^U \frac{1}{rank_u}$$

# Метрики ранжирования

## nDCG@n - Normalized Discounted Cumulative Gain

DCG@n является популярной метрикой в информационном поиске. Она учитывает и порядок, и количество релевантных объектов.

nDCG@n – это нормированная метрика.  
nDCG@n = 1 означает, что объекты идеально отранжированны.

$$DCG@n = \sum_{k=1}^n \frac{rel(k)}{\log_2(k + 1)}$$

$$nDCG@n = \frac{DCG@n}{IDCG@n}$$

$$IDCG@n = \sum_{k=1}^n \frac{1}{\log_2(k + 1)}$$

- Нормировочная константа

Специальные оценки качества?

# Специальные метрики качества



Очевидной идеей кажется «взвешивать»  
каким-то образом метрики.

Например:

0.5 NDCG@10 +

0.25 NDCG@3 +

0.25 RMSE

# Специальные метрики качества

Другой вариант, когда нужно показывать только хорошие рекомендации:

**1-ый шаг.** Оптимизация RMSE.

Для каждого пользователя получаем число объектов с оценкой выше заданного порога (параметр алгоритма):  $(u, k)$

**2-ой шаг.** Оптимизация ранжирования.

Решаем задачу с метрикой качества  $\text{NDCG}@(\min(n, k))$ .

**3-ий шаг.** Оптимизация ранжирования.

Выводим в блок не больше  $\min(n, k)$  для каждого пользователя.

# Зачем нужны разные метрики?



Стоит выбирать метрику в зависимости от специфики задачи и типов данных.

Неправильно выбранная метрика может привести к неправильным решениям.

# Зачем нужны разные метрики?



Стоит выбирать метрику в зависимости от специфики задачи и типов данных.

Неправильно выбранная метрика может привести к неправильным решениям.

Спасибо, кэп! Но как ее выбрать?



# Многокритериальные рекомендации

# Многокритериальная задача оптимизации

$$\min_x \{f_1(x), f_2(x), \dots, f_k(x)\}, \quad x \in X$$

где  $f_i: R^n \rightarrow R$  это  $k$  ( $k \geq 2$ ) целевых функций.

Векторы решений  $x = (x_1, x_2, \dots, x_n)^T$  относятся к непустой области определения.

Задача многокритериальной оптимизации состоит в **поиске вектора целевых переменных**, удовлетворяющего наложенным ограничениям и оптимизирующего векторную функцию, элементы которой соответствуют целевым функциям. Эти функции образуют математическое описание критерия удовлетворительности и, как правило, взаимно конфликтуют. Отсюда, «оптимизировать» означает найти такое решение, при котором значения целевых функций были бы приемлемыми для постановщика задачи.

# Многокритериальная задача оптимизации

$$\min_x \{f_1(x), f_2(x), \dots, f_k(x)\}, \quad x \in X$$

где  $f_i: R^n \rightarrow R$  это  $k$  ( $k \geq 2$ ) целевых функций.

Векторы решений  $x = (x_1, x_2, \dots, x_n)^T$  относятся к непустой области определения.

Допустимое решение  $\tilde{x} \in X$  называется **эффективным по Парето** или **Парето-оптимальным**, если не существует другого решения  $x \in X$  такого, что  $f_p(x) \leq f_p(\tilde{x})$  для всех  $p = 1, \dots, k$ ,  $f_i(x) \leq f_i(\tilde{x})$  хотя бы для одного  $i = 1, \dots, k$

Множество всех эффективных решений называется **эффективным множеством** и обозначается  $X_E$ .

# Свертка критериев



Стандартный приём «борьбы» с многокритериальным выбором это переход к однокритериальной задаче с использованием метода свёртки критериев.

# Свертка критериев

**Свёртка критериев** означает построение интегрального показателя на основе частных критериев. Интегральный показатель  $I$  рассчитывается или как взвешенная сумма частных показателей или как их произведение, нормированное на соответствующие веса (важность критериев).

Решение многокритериальных задач на основе линейной свертки критериев состоит в назначении тем или иным способом неотрицательных (а чаще положительных) коэффициентов  $\mu_1, \mu_2, \dots, \mu_m$ , в сумме дающих единицу (хотя это не обязательно), и последующей минимизацию линейной комбинации критериев

$$\sum_{i=1}^m \mu_i f_i(x)$$

на множестве  $X$ .

Вопрос: как можно использовать  
многокритериальную оптимизацию в  
задачах рекомендаций?

Бизнес-метрики

**Нужны ли бизнесу все  
рассмотренные  
метрики?**



~~Нужны ли бизнесу все  
рассмотренные  
метрики?~~

**НЕТ!**

# Бизнес-метрики

---

## Что нужно бизнесу?

- Конверсия
- Кликабельность
- Увеличение времени на сайте или в приложении
- LTV – ценность за период
- Стоимость привлечения клиента
- (CAC – customer acquisition cost)
- Коэффициент удержание клиента
- Время возмещения CAC  
(количество месяцев)
- Прибыль



# Бизнес-метрик **ОЧЕНЬ** много



На практике в бизнесе используются десятки метрик.  
Мы рассмотрим наиболее частые, которые используются в рекомендательных системах.

<https://vc.ru/marketing/239889-cpiski-metrik-dlya-7-biznes-modeley-ot-saas-do-e-commerce>

# CTR (click-through rate)



**CTR (click-through rate — показатель кликабельности)** — метрика в интернет-маркетинге. CTR определяется как отношение числа кликов на баннер или рекламное объявление к числу показов, измеряется в процентах.

**Формула вычисления CTR:**

$$\text{CTR} = (\text{количество кликов} / \text{количество показов}) * 100$$

# Конверсионные метрики



- Добавление в корзину из рекомендательного блока
- Покупки из рекомендательного блока
- Добавление в корзину просмотренных в рекомендательном блоке
- Покупки просмотренных в рекомендательном блоке
- Другие конверсии

# «Денежные» метрики



- Выручка
- Прибыль
- Средний чек
- Средняя прибыль
- LTV – life-time value (ценность клиента за все время пользования сервисом)
- Число платящих пользователей

# Клиентские метрики



## Клиентские метрики

показывают клиентский опыт.

Например:

- удовлетворенность клиентов — CSAT
- лояльность клиентов — NPS
- усилия клиентов — CES

# Лояльность клиентов

**Net Promoter Score** - «общее количество промоутеров» или «индексом искренней лояльности», который помогает просчитать, как много у вашей компании людей, готовых рекомендовать вас своим знакомым и, следовательно, привести других клиентов.

Для того чтобы рассчитать NPS, потребителям задается вопрос: «Насколько вероятно, что вы порекомендуете нашу компанию?» — по шкале от 0 до 10 баллов.

Насколько вероятно, что вы нас порекомендуете знакомым?



0: никогда не порекомендую

10: с радостью порекомендую

Отправить



# Удовлетворенность клиентов

Эта метрика помогает  
определить степень  
удовлетворенности  
клиентов компании —  
**Customer SATisfaction.**

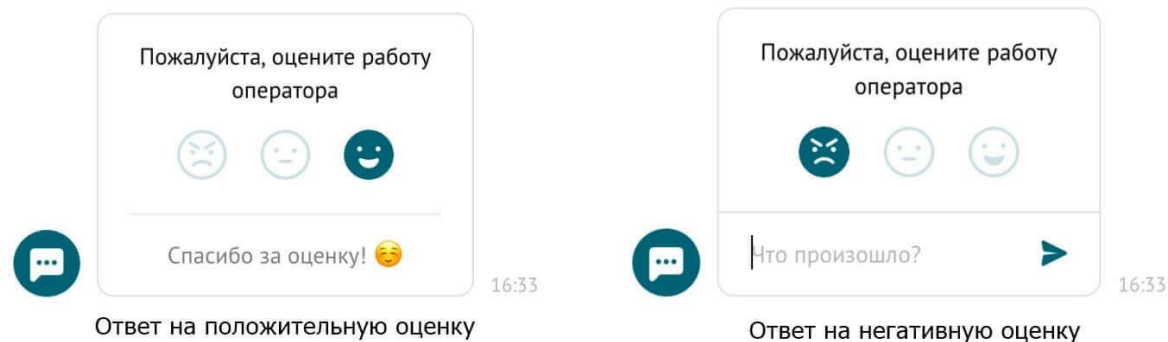
Каковы ваши впечатления  
от продукта?



Отправить

# Удовлетворенность клиентов

Эта метрика помогает определить степень удовлетворенности клиентов компании — **Customer SATisfaction**.



## Формула:

(количество довольных клиентов / количество опросов) x 100 = % удовлетворенных клиентов.

Чтобы вычислить CSAT, клиентов спрашивают, насколько они удовлетворены продуктом или услугой. А для ответа предлагают шкалу от 1 до 5 баллов.

# Усилия клиентов

**Customer effort score** — показатель усилий клиентов.

Метрика помогает обнаружить сложности для потребителя при взаимодействии с вашей компанией/продуктом/сервисом.

**Формула:**  
**ЛВ(%) — СВ(%),**

где *ЛВ* — легкое взаимодействие клиентов с компанией, а *СВ* — сложное, когда клиентам пришлось затратить много усилий для совершения действия.

Насколько вы согласны или не согласны с данным утверждением:

Компания организовала все так, что решить вопрос было легко

1 2 3 4 5 6 7  
Полностью не согласен Полностью согласен

Что мы можем улучшить в нашем сервисе для вас:

Спасибо!

Отправить

# Прокси-метрики

Как правило, при разработке рекомендательных систем используются **прокси-метрики**.

Задача.

Вариант 1. Тест на Выручку – 1 год ( $+2.5 \cdot 12 = 30$ ), но точно выберем.

Вариант 2. Тест на CTR – 2 месяца ( $+2.5 \cdot 2 + 5 \cdot 10 = 55$ ), но можем ошибиться.

$\text{corr}(\text{CTR}, \text{Выручку}) = 0.67$

**Какой вариант выбрать?**

Потенциальная прибыль 5 у.е.

Вероятность того что **CTR** прокрасился, а **Выручка** нет = **15%**

# «Другие» метрики

Также часто от рекомендательных систем других неформальных свойств.

Например:

- Разнообразие
- Не тривиальных рекомендаций
- Покрытия запроса

# Тестирование



Если мы возьмем рекомендательные системы в онлайн-бизнесе, то они, как правило, преследует **две цели**:

- проинформировать пользователя об интересном товаре
- побудить его совершить покупку (путем рассылки, составления персонального предложения и т.д.).

# Онлайн и оффлайн оценки качества

# Тестирование



Как и в любой модели, направленной на мотивацию пользователя к действию, **оценивать следует только инкрементальный прирост целевого действия.**

Т.е., например, при подсчете покупок по рекомендации нам нужно исключить те, которые пользователь и так сам бы совершил без нашей модели. Если этого не сделать, эффект от внедрения модели будет сильно завышен.



# Тестирование



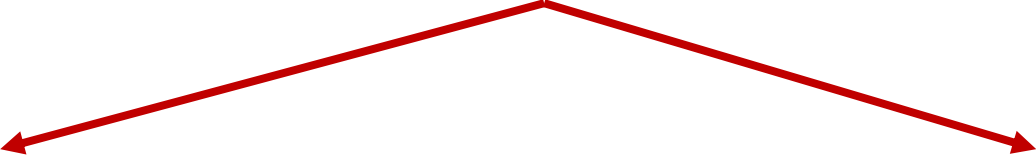
**Lift** — показатель того, во сколько раз точность модели превосходит некий baseline алгоритм. В нашем случае baseline алгоритмом может быть просто отсутствие рекомендаций или случайные рекомендации или «самые популярные» (см. лекцию №2).

Данная метрика хорошо отлавливает долю инкрементальных покупок и это позволяет эффективно сравнивать разные модели.

# Оценка качества системы



## Подходы к тестированию



offline тестирование  
модели на исторических  
данных с помощью  
ретро-тестов

тестирование готовой  
модели с помощью  
A/B тестирования

# Offline тестирование

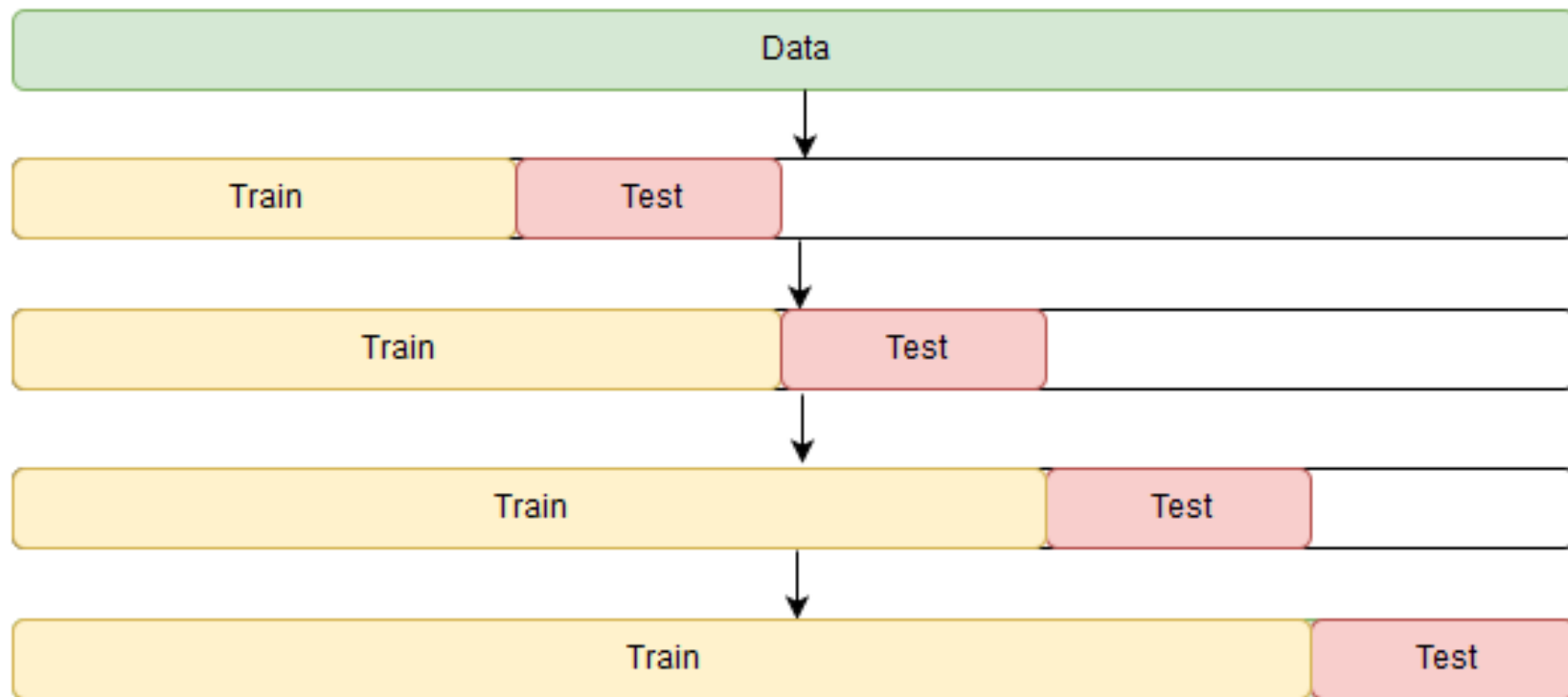


Стандартный подход — это кросс-валидация методами **leave-one-out** и **leave-p-out**.

- **leave-one-out** — модель обучается на всех оцененных пользователем объектах, кроме одного, а тестируется на этом одном объекте. Так делается для всех  $n$  объектов, и среди полученных  $n$  оценок качества вычисляется среднее
- **leave-p-out** — то же самое, но на каждом шаге исключается  $p$  точек

Важно сделать правильно разделение по времени для фолдов!

# Offline тестирование



Важно сделать правильно разделение по времени для фолдов!

Проблема: как оценить бизнес-метрику если мы объект, который мы хотели показать не показывали на валидации?

# Offline тестирование



С помощью оффлайн тестирования мы можем построить модель только оценки известных показателей.

Например, мы хотим прогнозировать CTR (и ранжировать по нему товары). Но если наша ранжирующая модель показывает, что раньше надо было показывать товар X пользователю, а мы его не показывали – у нас нет никакой статистики.

С точки зрения задачи ранжирования нам не важен конкретный CTR, важны только ранги. Но, как правило, мы не можем валидировать ранги, т.к. доля товар по которым были клики очень мала.

Поэтому приходится валидироваться на значения CTR. То есть использовать pointwise подход.

Также очень полезно знать как  
скоррелирована ваша ML-метрика  
с бизнес-метрикой.

# Online тестирование



Онлайн-модели отличаются от оффлайн-моделей тем, что их главная цель – **как можно быстрее оценить качество алгоритма.**

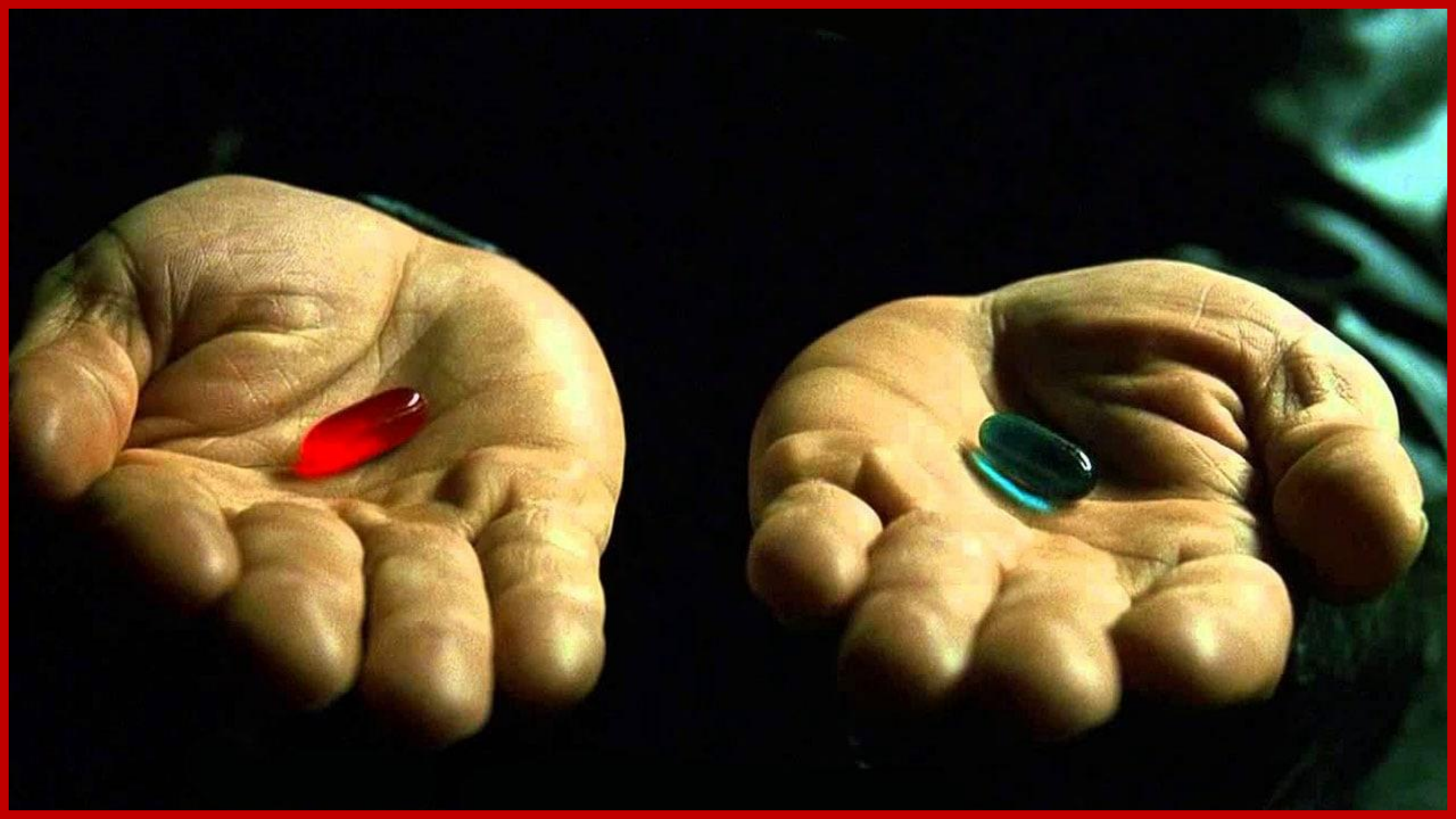
- Данных тут недостаточно, чтобы такие изменения можно было поймать методами коллаборативной фильтрации.
- Поэтому онлайн-методы обычно меньше персонализированы, индивидуальных данных просто не наберётся.



# АБ-тестирование



**АВ-тестирование** – это проверка гипотезы, о том, что реализации случайной величины в одной версии (А) *отличается* от реализации случайной величины в другой версии (В).



# АБ-тестирование



**АВ-тестирование** — это эксперимент, целью которого является сравнение двух вариантов одного и того же объекта. Версии А и версии В.

# АБ-тестирование



**АВ-тестирование** – это эксперимент, целью которого является сравнение двух вариантов одного и того же объекта. Версии А и версии В.

Версия А лучше версии В – допустимый тест, но НЕ самый лучший.

# Как провести АБ-тестирование?

---

1. Определите цели
2. Определите метрику
3. Подготовка к эксперименту
4. Проведите эксперимент
5. Анализ результатов

# Гипотезы



Рассматривается пара гипотез:

$H_0$  - нул-гипотеза

$H_1$  - альтернативная гипотеза

Нул-гипотезой принято обозначать, такую гипотезу, что А и В не отличаются между собой.

Альтернативная гипотеза — это гипотеза о том, что А отличается от В.

# Матрица ошибок

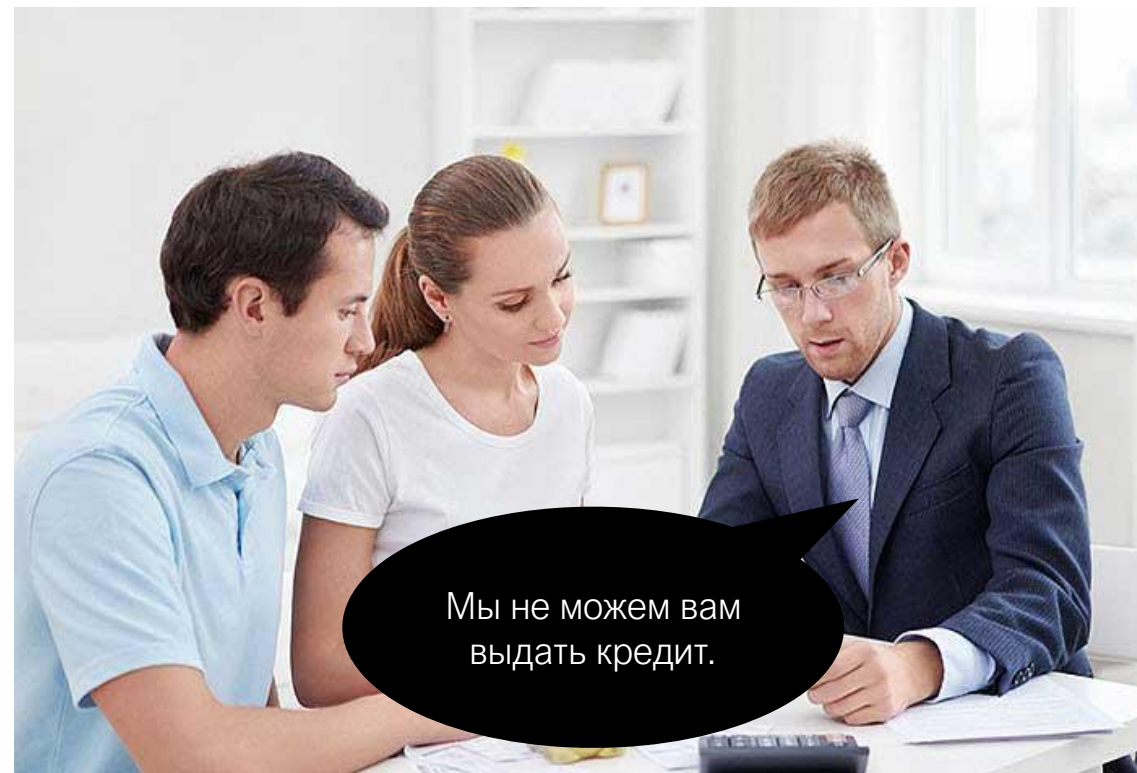
Результат  
применения  
критерия

Таблица бок	Верная гипотеза	
	$H_0$	$H_1$
$H_0$	$H_0$ Верно принята	$H_0$ Неверно принята (ошибка II рода)
$H_1$	$H_0$ Неверно отвергнута (ошибка I рода)	$H_0$ Верно отвергнута


## Ошибка I Рода



## Ошибка II Рода





# Ошибки I и II рода



Ошибки не симметричны!

Ошибка I рода – отвергнуть нуль-гипотеза, когда она верна.

Ошибка I рода, как правило, критичнее и она жестко ограничена сверху.

# Подготовка к эксперименту



1. Что мы измеряем
2. Минимальный размер эффекта
3. Ошибка I (и II рода)

# Подготовка к эксперименту

---

1. Что мы измеряем
2. Минимальный размер эффекта
3. Ошибка I (и II рода)

На основе этих данных вычисляется: **Размер выборки**

# Как поставить АБ-тест

---

1. Что мы измеряем
2. Минимальный размер эффекта
3. Ошибка I (и II рода)

Пример:

1. CTR
2. Увеличится на 5%
3. Ошибка I рода: 0.05
4. 100 000 человек

# Проверка расчетов



Калькуляторы:

- [Driveback](#) на русском;
- [Hungrysites](#) на русском;
- [Mindbox](#) на русском;
- [ABTestGuide](#) на английском, но с визуализацией графиками;
- <https://yandex.ru/adv/statvalue> на русском.

Проблема пузыря

# Проблема взаимодействия с клиентом

Рекомендательная система считывает действия пользователя (разные «прокси»: лайки, переход на страницу и т.д.) и дальше начинает рекомендовать то, что и так нравится пользователю, при этом не предлагая каких-то других альтернатив, подходящих под его предпочтения.

Говоря более неформальным языком: рекомендательные системы максимизируют **краткосрочную** мотивацию пользователя.

# Bubble problem



Проблема «пузыря». Негативная сторона персонализированного поиска (и рекомендаций) состоит в том, что сервисы определяют, какую информацию пользователь **хотел бы увидеть**, основываясь на информации о его месторасположении, прошлых нажатиях и перемещениях мыши, предпочтениях и истории поиска. В результате сервисы показывают только информацию, которая согласуется с прошлыми точками зрения данного пользователя. Вся иная информация, как правило, пользователю не выводится.

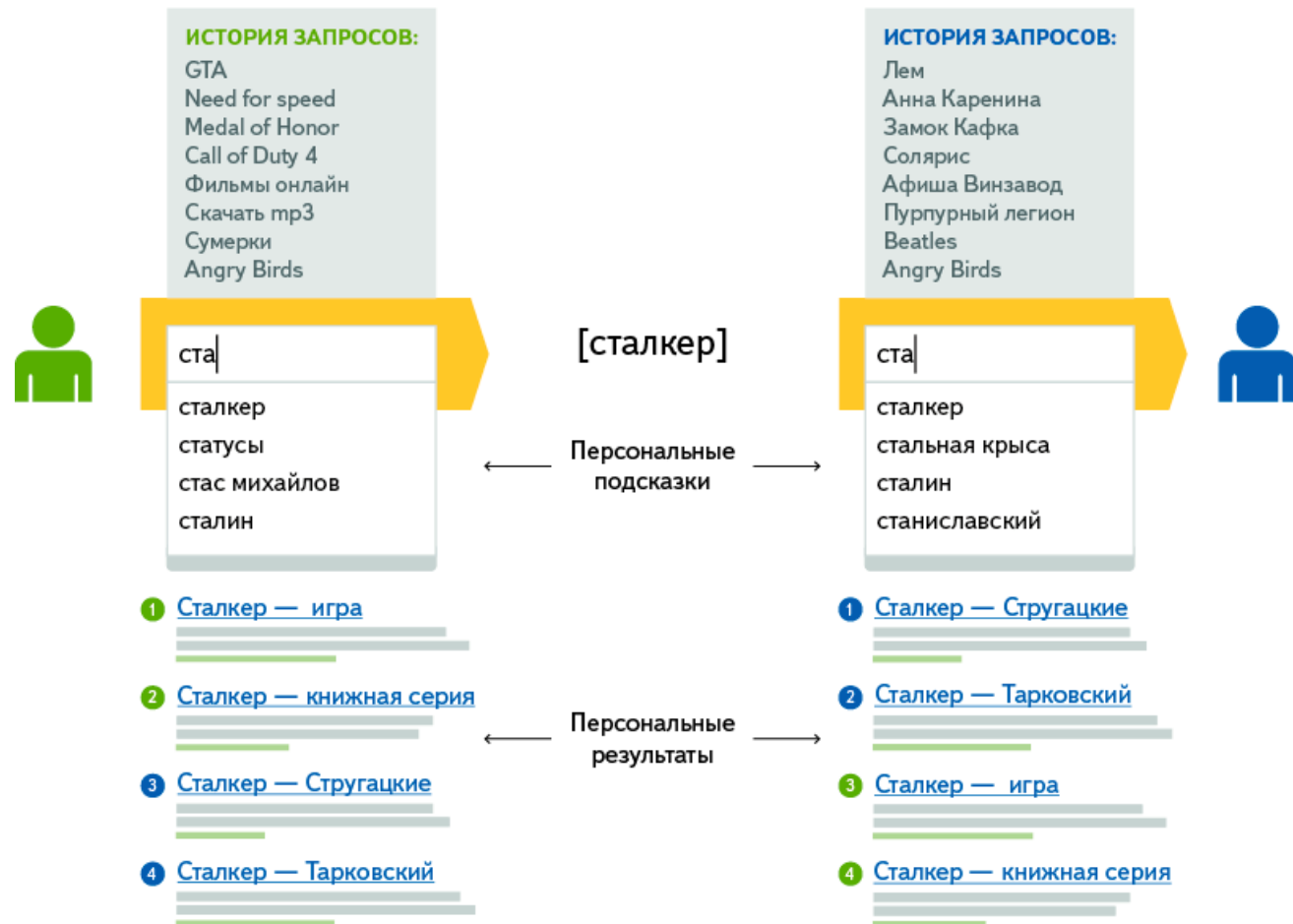
*Eli Pariser: The Filter Bubble (2011)*

<https://youtu.be/B8ofWfx525s>



# Bubble problem

Примером этого являются Google и другие поисковые системы с персонализированными результатами поиска, а также Facebook с персонализированной лентой новостей, которая с каждым действием пользователя наполняется всё более и более персональными результатами.



# Bubble problem



С данной проблемой сталкиваются множество различных сервисов:

- Социальные сети
- Поисковые
- Контентные
- Музыкальные и видео
- Интернет-магазины (в меньшей степени)

Как решать такую проблему?

# Заключение



- Нужно знать бизнес-метрики
  - Собирать разные метрики
  - Не только целевую
- Нужно использовать прокси-метрики
- Правильно проводить АБ
- Уметь строить зависимость между бизнес-метрикой и метрикой машинного обучения
- Работать с побочными эффектами от рекомендациями.

Вопросы

# Семинар: сравнение метрик ранжирования и прокси-метрик