

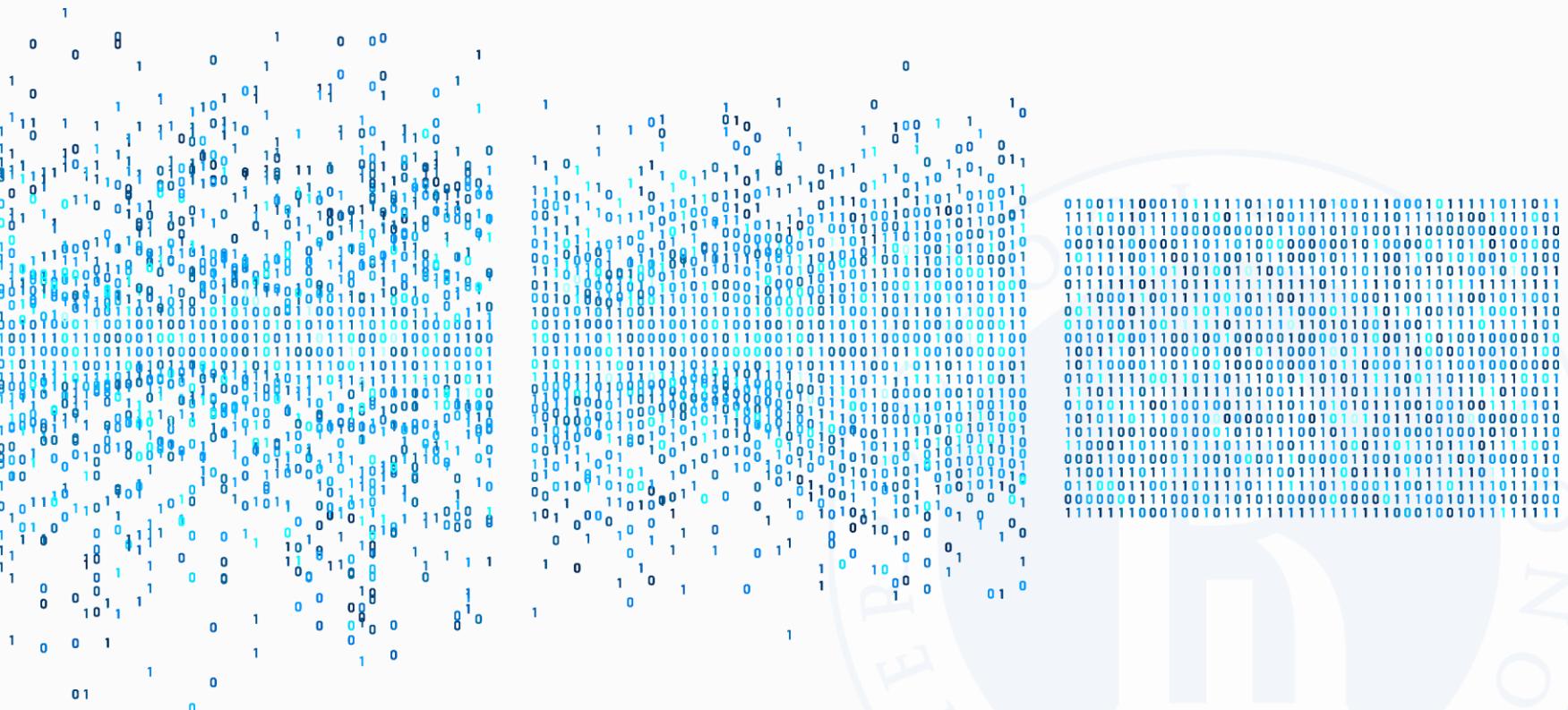
Аналитика



Аналитика



Создали свой сервис — что с ним делать дальше?



Метрики



Надо измерять какие-то показатели. Но какие и зачем?



Продукт

→ Как понять, что мы сделали что-то хорошее?



Что изучим?

- Как сделать метрики для своего сервиса
- Зачем проводить А/В-тесты
- Сможем использовать метрики для улучшения продукта
- Как работают с метриками разные команды:
разработчики, менеджеры и аналитики

Оффайн и онлайн-метрики



План



Что такое метрика



Типы метрик



Примеры метрик



Использование метрик



Что такое метрика

→ Метрика — инструмент измерения качества сервиса



Типы метрик

→ Online

- Считываются по логам пользователей
- Клики, показы, переходы и любая другая пользовательская активность

Типы метрик



Online

- Считываются по логам пользователей
- Клики, показы, переходы и любая другая пользовательская активность



Offline

- Считываются по самому сервису без пользовательской активности
- Замеряем качество продукта для всех пользователей сразу

Типы метрик

→ Online

- Считываются по логам пользователей
- Клики, показы, переходы и любая другая пользовательская активность

→ Offline

- Считываются по самому сервису без пользовательской активности
- Замеряем качество продукта для всех пользователей сразу

→ **Онлайн-метрики** считаются по логам пользователей,
оффлайн — без логов и активности пользователей

Пример онлайн-метрики



Эффективность подсказки при вводе запроса в поиск



Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками

Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками
- Сколько позиций внутри блока

Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками
- Сколько позиций внутри блока
- Клики по позициям

Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками
- Сколько позиций внутри блока
- Клики по позициям
- Клики вне блока для его закрытия

Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками
- Сколько позиций внутри блока
- Клики по позициям
- Клики вне блока для его закрытия
- Как быстро находят ответ **с** нашей подсказкой
- Как быстро находят ответ **без** подсказки

Пример онлайн-метрики



Что можем замерять?

- Количество показов блоков с подсказками
- Сколько позиций внутри блока
- Клики по позициям
- Клики вне блока для его закрытия
- Как быстро находят ответ **с** нашей подсказкой
- Как быстро находят ответ **без** подсказки
- Есть ли переформулировка запроса после подсказки

Пример онлайн-метрики



Что можем замерять?

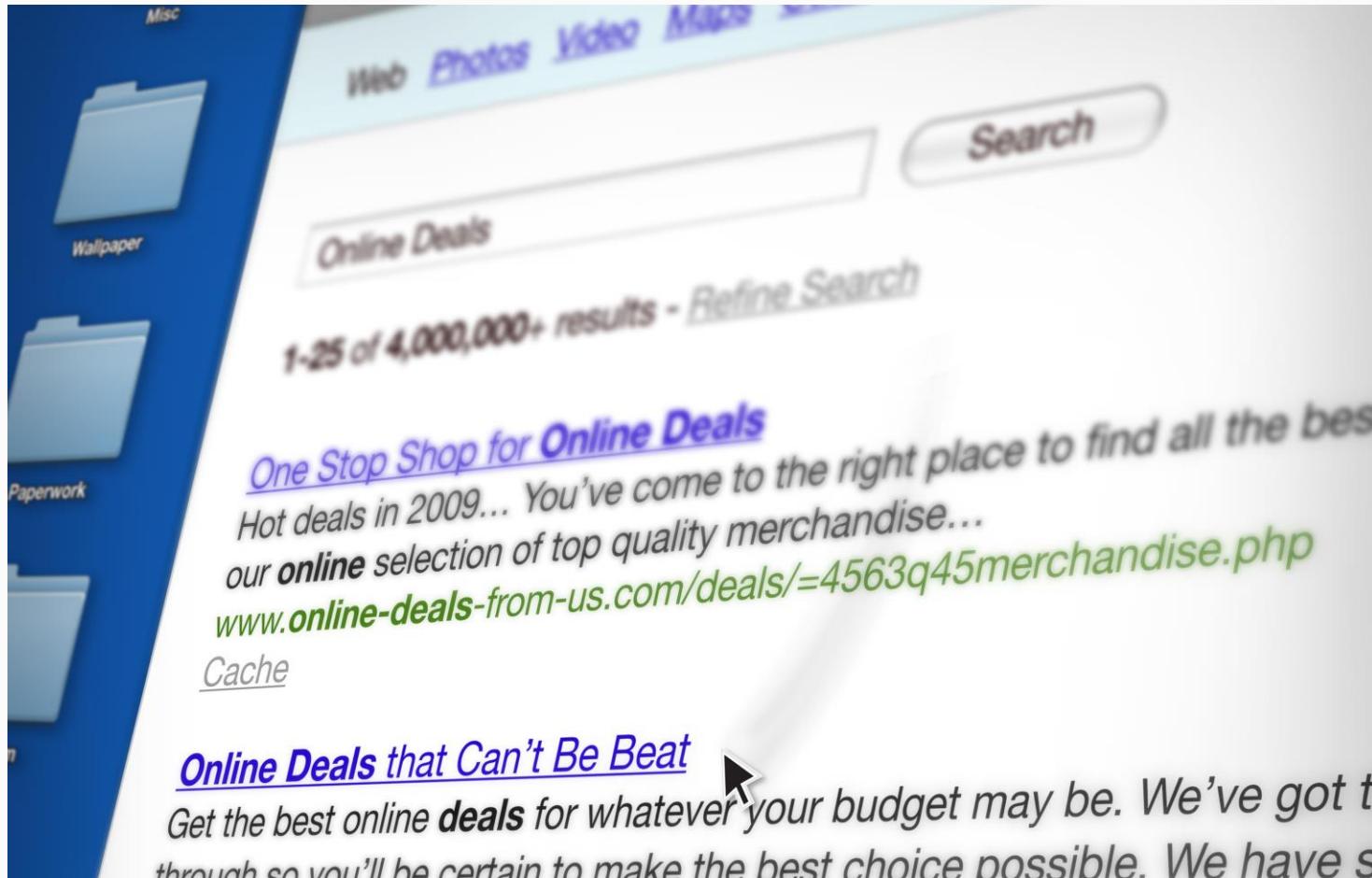
- Количество показов блоков с подсказками
- Сколько позиций внутри блока
- Клики по позициям
- Клики вне блока для его закрытия
- Как быстро находят ответ **с** нашей подсказкой
- Как быстро находят ответ **без** подсказки
- Есть ли переформулировка запроса после подсказки
- ...



Все метрики полезно смотреть в абсолютных значениях и в нормированных

Пример онлайн-метрики

→ Качество результатов в выдаче



Пример онлайн-метрики



Что можем замерять?

- По каким запросам результат в принципе есть

Пример онлайн-метрики



Что можем замерять?

- По каким запросам результат в принципе есть
- Соответствие документов запросу

Пример онлайн-метрики



Что можем замерять?

- По каким запросам результат в принципе есть
- Соответствие документов запросу
- Насколько документы полезны и решают задачи пользователя

Пример онлайн-метрики



Что можем замерять?

- По каким запросам результат в принципе есть
- Соответствие документов запросу
- Насколько документы полезны и решают задачи пользователя
- Позиции хороших и плохих документов относительно друг друга
- ...

Пример онлайн-метрики



Что можем замерять?

- По каким запросам результат в принципе есть
- Соответствие документов запросу
- Насколько документы полезны и решают задачи пользователя
- Позиции хороших и плохих документов относительно друг друга
- ...



Нужно репрезентативное множество примеров, которые отражают весь поток данных



По результатам на множестве судим обо всем потоке

Использование метрик

- Для всех элементов сервиса создаем разнообразные метрики онлайн и офлайн

Использование метрик

- Для всех элементов сервиса создаем разнообразные метрики онлайн и офлайн
- Полезны как простые метрики, которые легко можно вырастить, так и более сложные, рост которых означает значимое улучшение качества

Использование метрик

- Для всех элементов сервиса создаем разнообразные метрики онлайн и офлайн
- Полезны как простые метрики, которые легко можно вырастить, так и более сложные, рост которых означает значимое улучшение качества
- Любые изменения в сервисе должны быть отражены в изменениях метрик

Использование метрик

- Для всех элементов сервиса создаем разнообразные метрики онлайн и офлайн
- Полезны как простые метрики, которые легко можно вырастить, так и более сложные, рост которых означает значимое улучшение качества
- Любые изменения в сервисе должны быть отражены в изменениях метрик
- Метрики должны отражать заведомо хорошие и заведомо плохие изменения и не противоречить продуктовому смыслу

Использование метрик

- Для всех элементов сервиса создаем разнообразные метрики онлайн и офлайн
- Полезны как простые метрики, которые легко можно вырастить, так и более сложные, рост которых означает значимое улучшение качества
- Любые изменения в сервисе должны быть отражены в изменениях метрик
- Метрики должны отражать заведомо хорошие и заведомо плохие изменения и не противоречить продуктовому смыслу
- Ортогональные сигналы позволяют взглянуть на продукт с разных сторон

Проверка гипотез



План



Как формулировать и проверять гипотезы



Ошибки первого и второго рода — как можно неправильно интерпретировать данные



Как можно неправильно сравнить два числа

Пример гипотезы

→ Пусть у нас есть интернет-блог

Пример гипотезы

- Пусть у нас есть интернет-блог
- На страницах есть медиа-контент
 - Страницы грусятся долго

Пример гипотезы

- Пусть у нас есть интернет-блог
- На страницах есть медиа-контент
 - Страницы грусятся долго
- Придумали, как ускорить загрузку

Пример гипотезы

- Пусть у нас есть интернет-блог
- На страницах есть медиа-контент
 - Страницы грусятся долго
- Придумали, как ускорить загрузку
 - Замерили старый и новый варианты

Пример гипотезы

До: [2.1, 3.5, 1.6, 3.6, 4.0]

После: [1.0, 2.0, 1.5, 3.4, 0.8]

Пример гипотезы

До: [2.1, 3.5, 1.6, 3.6, 4.0]

После: [1.0, 2.0, 1.5, 3.4, 0.8]



Время загрузки стало меньше

Пример гипотезы

До: [2.1, 3.5, 1.6, 3.6, 4.0]

После: [1.0, 2.0, 1.5, 3.4, 0.8]

→ Время загрузки стало меньше

→ Стало ли лучше пользователю?

Пример гипотезы

До: [2.1, 3.5, 1.6, 3.6, 4.0]

После: [1.0, 2.0, 1.5, 3.4, 0.8]



Время загрузки стало меньше



~~Стало ли лучше пользователю?~~



Насколько значимо изменение времени?

Пример гипотезы

До: [2.1, 3.5, 1.6, 3.6, 4.0]

После: [1.0, 2.0, 1.5, 3.4, 0.8]

→ Время загрузки стало меньше

→ ~~Стало ли лучше пользователю?~~

→ Насколько значимо изменение времени?



→ На основании этих данных сказать сложно

Формулирование гипотезы

→ Сформулируем две гипотезы

- H_0 : изменения незначительны
- $H_1: \sim H_0$ — изменения значимы

Формулирование гипотезы

→ Сформулируем две гипотезы

- H_0 : изменения незначительны
- $H_1: \sim H_0$ — изменения значимы

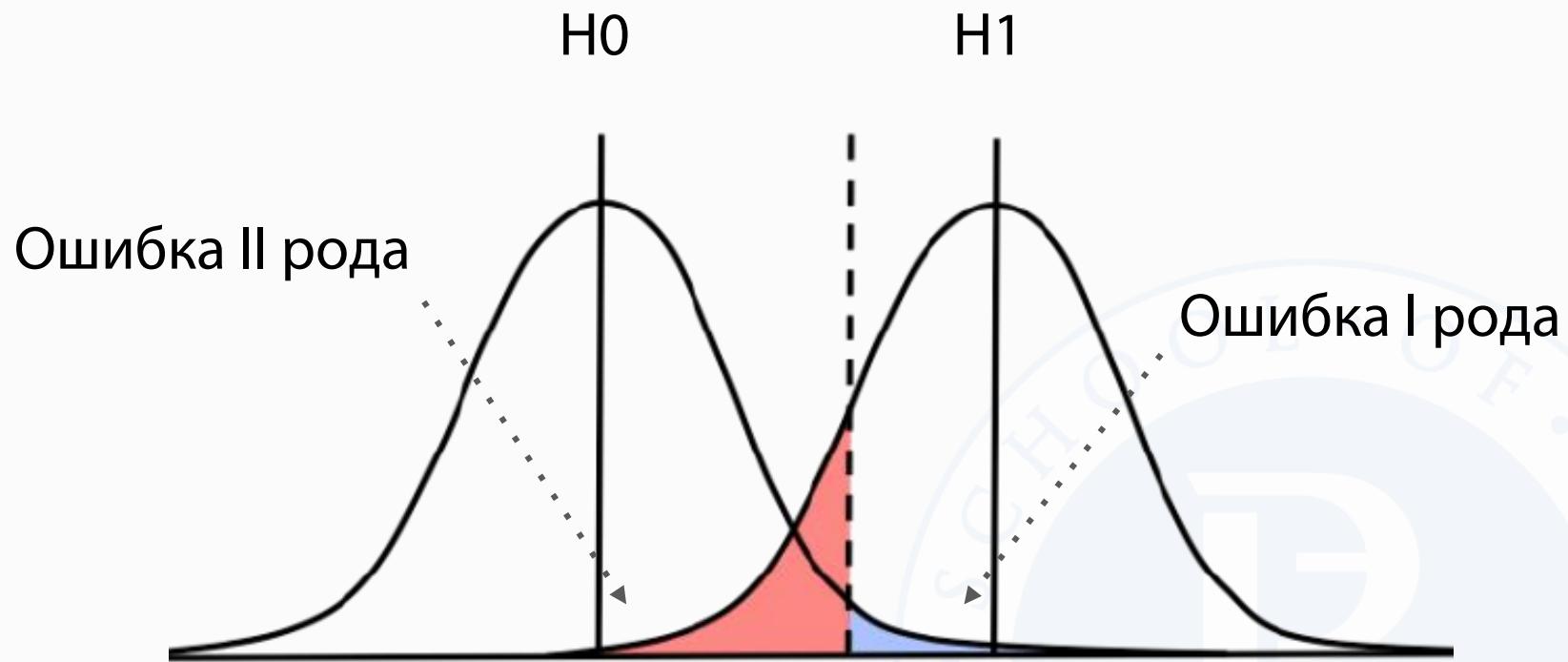
→ Вопрос, на который хотим ответить:

- Какова вероятность **получить отклонения** текущие или более выраженные при условии, что **изменений не было?** — p-value

Ошибки I и II рода

	H_0 верна	H_0 неверна
Принимаем H_0	Верно	Ошибка II рода (β)
Отвергаем H_0	Ошибка I рода (α)	Верно

Ошибки I и II рода



P-value

→ P-value — вероятность получить отклонения **текущие** или **более выраженные** при условии, что изменений **не было**

P-value

- P-value — вероятность получить отклонения **текущие** или **более выраженные** при условии, что изменений не было
- Получили **p-value=0.04** при пороге достоверности 0.05

P-value

- P-value — вероятность получить отклонения **текущие** или **более выраженные** при условии, что изменений не было
- Получили **p-value=0.04** при пороге достоверности 0.05
 - P-value ничего не говорит о причинно-следственной связи

P-value

- P-value — вероятность получить отклонения **текущие** или **более выраженные** при условии, что изменений не было
- Получили **p-value=0.04** при пороге достоверности 0.05
 - P-value ничего не говорит о причинно-следственной связи
 - P-value не означает вероятность случайно получить определенный результат

P-value

- P-value — вероятность получить отклонения **текущие** или **более выраженные** при условии, что изменений **не было**
- Получили **p-value=0.04** при пороге достоверности 0.05
 - P-value ничего не говорит о причинно-следственной связи
 - P-value не означает вероятность случайно получить определенный результат
 - P-value не говорит о том, что бы было, если бы мы получили более высокие значения (например, 0.06)

Проверка гипотезы

1. Формулируем основную и альтернативную гипотезы

Проверка гипотезы

1. Формулируем основную и альтернативную гипотезы
2. По реальным данным строим распределение нужной величины (эмпирическое значение)

Проверка гипотезы

1. Формулируем основную и альтернативную гипотезы
2. По реальным данным строим распределение нужной величины (эмпирическое значение)
3. Для случая верности H_0 строим теоретическое распределение этого же значения

Проверка гипотезы

1. Формулируем основную и альтернативную гипотезы
2. По реальным данным строим распределение нужной величины (эмпирическое значение)
3. Для случая верности H_0 строим теоретическое распределение этого же значения
4. Решаем, принять или отклонить H_0

Проверка гипотезы

1. Формулируем основную и альтернативную гипотезы
 2. По реальным данным строим распределение нужной величины (эмпирическое значение)
 3. Для случая верности H_0 строим теоретическое распределение этого же значения
 4. Решаем, принять или отклонить H_0
- Быстро посчитать значимость отклонений помогают статистические критерии

Как сравнить два числа?

→ Хотим получить детектор лиц людей

Как сравнить два числа?

- Хотим получить детектор лиц людей
- Сделали две модели, сравниваем по F-мере
 - I модель — 0.86
 - II модель — 0.89

Как сравнить два числа?

- Хотим получить детектор лиц людей
- Сделали две модели, сравниваем по F-мере
 - I модель — 0.86
 - II модель — 0.89
- Вторая лучше?

Как сравнить два числа?

- Хотим получить детектор лиц людей
- Сделали две модели, сравниваем по F-мере
 - I модель — 0.86
 - II модель — 0.89
- Вторая лучше?
- Хотим сравнить именно **модели**, а **не итоговые числа**

Как сравнить два числа?



Почему значения могут не отражать действительность?

Как сравнить два числа?



Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме

Как сравнить два числа?

→ Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме
- Различные начальные условия

Как сравнить два числа?

→ Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме
- Различные начальные условия
- Проверочные данные смещенные

Как сравнить два числа?

→ Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме
- Различные начальные условия
- Проверочные данные смещенные

→ Проверим гипотезу, что одна модель лучше другой

Как сравнить два числа?



Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме
- Различные начальные условия
- Проверочные данные смещенные



Проверим гипотезу, что одна модель лучше другой

- Сгенерируем наблюдения от обеих моделей

Как сравнить два числа?



Почему значения могут не отражать действительность?

- Не повезло с рандомизацией в алгоритме
- Различные начальные условия
- Проверочные данные смещенные



Проверим гипотезу, что одна модель лучше другой

- Сгенерируем наблюдения от обеих моделей
- Оценим статистическую значимость различий

Как сравнить два числа?

- Модель I: [0.86, 0.90, 0.89, 0.88]
- Модель II: [0.89, 0.87, 0.91, 0.89]

Как сравнить два числа?

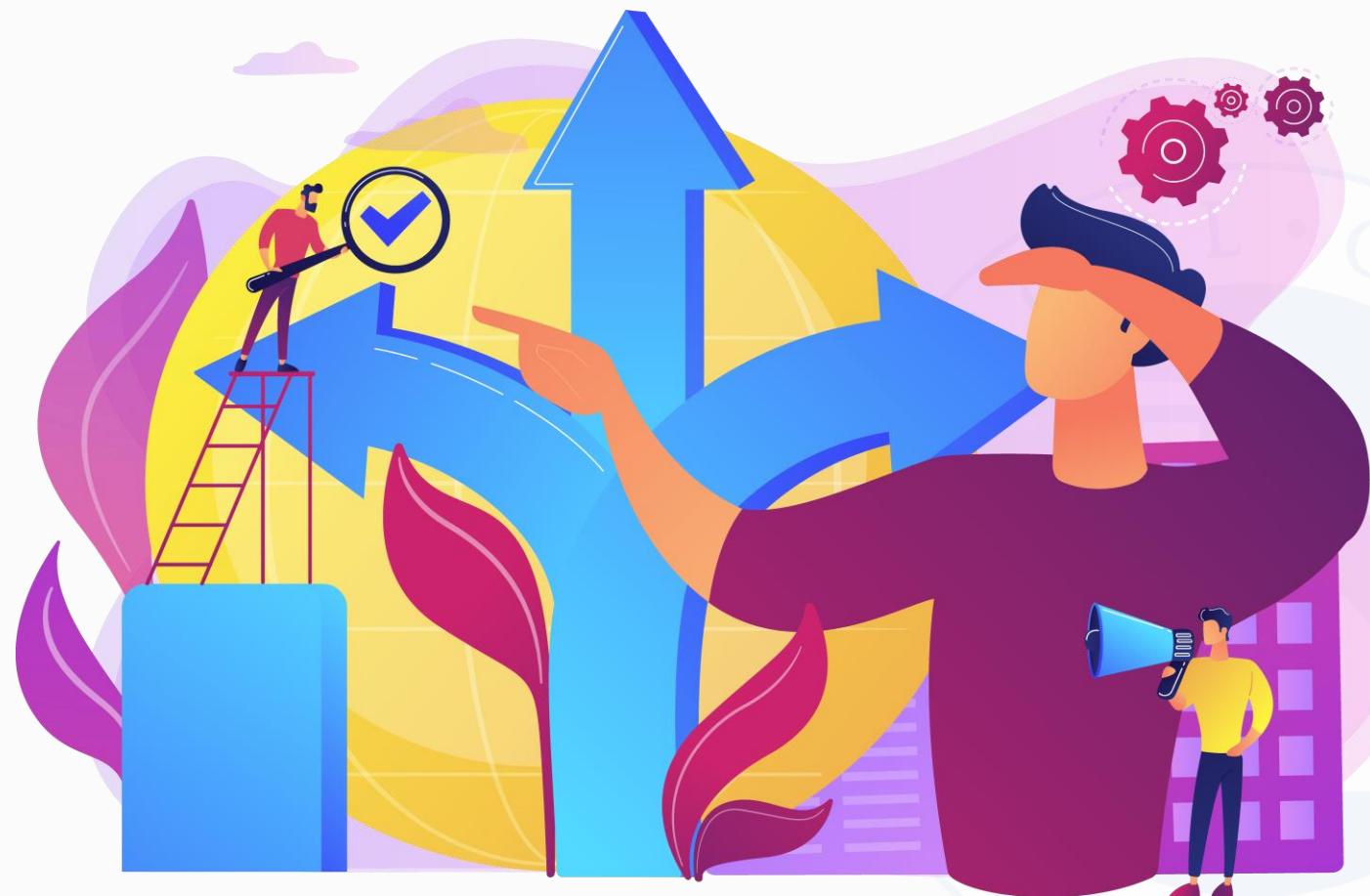
- Модель I: [0.86, 0.90, 0.89, 0.88]
- Модель II: [0.89, 0.87, 0.91, 0.89]

→ По таким данным уже сложно сказать, какая модель лучше



Для чего проверять гипотезы?

→ Чтобы делать обоснованные выводы!



Статистические тесты



План



Что такое статистический тест



Примеры статистических тестов



Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение
- Определить уровень значимости

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение
- Определить уровень значимости
- Получить наблюдения в эксперименте

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение
- Определить уровень значимости
- Получить наблюдения в эксперименте
- Сравнить ожидаемые и наблюдаемые значения

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение
- Определить уровень значимости
- Получить наблюдения в эксперименте
- Сравнить ожидаемые и наблюдаемые значения
- Посчитать p-value

Статистические тесты

→ Тесты позволяют ответить на вопрос, есть ли статистически значимая разница между наблюдениями

→ Алгоритм:

- Определить ожидаемое значение
- Определить уровень значимости
- Получить наблюдения в эксперименте
- Сравнить ожидаемые и наблюдаемые значения
- Посчитать p-value
- Принять или отклонить нулевую гипотезу

Выбор теста

→ У каждого теста есть условия применимости

Выбор теста

- У каждого теста есть условия применимости
 - Распределение в данных

Выбор теста



У каждого теста есть условия применимости

- Распределение в данных
- Количество выборок

Выбор теста



У каждого теста есть условия применимости

- Распределение в данных
- Количество выборок
- Зависимость данных в выборках

Выбор теста



У каждого теста есть условия применимости

- Распределение в данных
- Количество выборок
- Зависимость данных в выборках
- Количество данных

Выбор теста

→ У каждого теста есть условия применимости

- Распределение в данных
- Количество выборок
- Зависимость данных в выборках
- Количество данных
- Абсолютные значения в данных

Выбор теста

- У каждого теста есть условия применимости
- Распределение в данных
 - Количество выборок
 - Зависимость данных в выборках
 - Количество данных
 - Абсолютные значения в данных
 - Тип данных — категориальный или количественный

Критерий Хи-квадрат

→ Эксперимент — проверка «правильности» игральной кости

	Наблюдение О	Ожидание Е
1	8	13
2	12	13
3	13	13
4	9	13
5	12	13
6	24	13
Итого	78	78

Критерий Хи-квадрат

$$\chi_n^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

Критерий Хи-квадрат

$$\chi_n^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi_6^2 = \frac{(8 - 13)^2}{13} + \dots + \frac{(24 - 13)^2}{13} = 12$$

Критерий Хи-квадрат

- Уровень значимости — 0.05
- 5 степеней свободы

Критерий Хи-квадрат

→ Уровень значимости — 0.05

→ 5 степеней свободы

Число степеней свободы k	Уровень значимости а			
	0,01	0,025	0,05	0,95
1	6,6	5	3,8	0,0039
2	9,2	7.4	6	0,103
3	11,3	9.4	7,8	0,352
4	13,3	11,1	9,5	0,711
5	15,1	12.8	11,1	1,15

→ Наше значение — 11.1

Критерий Хи-квадрат

→ Уровень значимости — 0.05

→ 5 степеней свободы

Число степеней свободы k	Уровень значимости а			
	0,01	0,025	0,05	0,95
1	6,6	5	3,8	0,0039
2	9,2	7.4	6	0,103
3	11,3	9.4	7,8	0,352
4	13,3	11,1	9,5	0,711
5	15,1	12.8	11,1	1,15

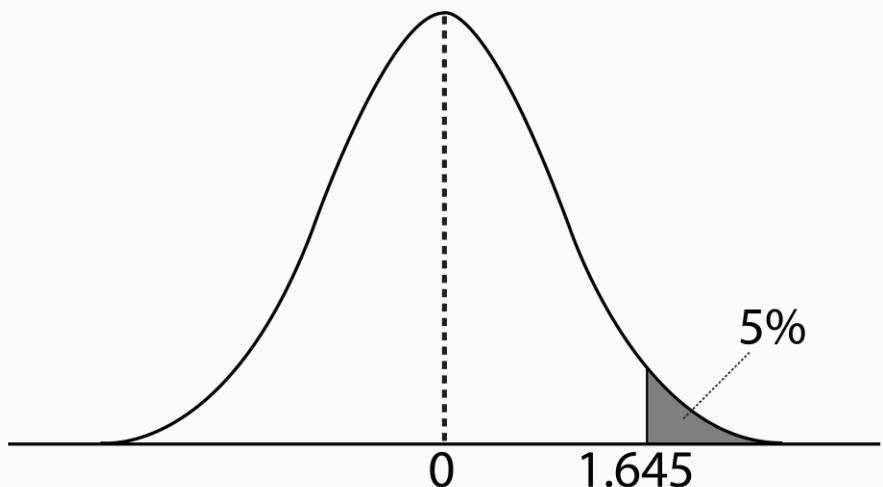
→ Наше значение — 11.1

→ $12 > 11.1$, принимаем решение о «неправильности» игральной кости

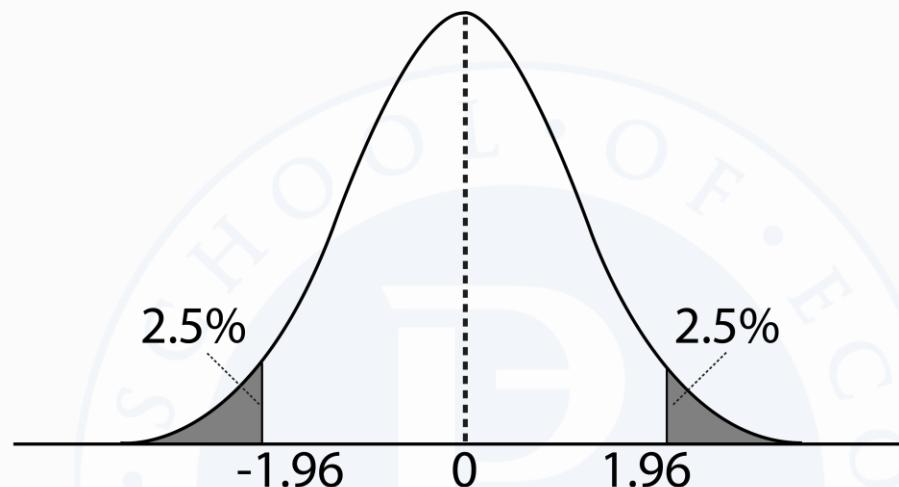
Направленность теста



Тест может быть односторонним или двусторонним



(a) One-tailed test

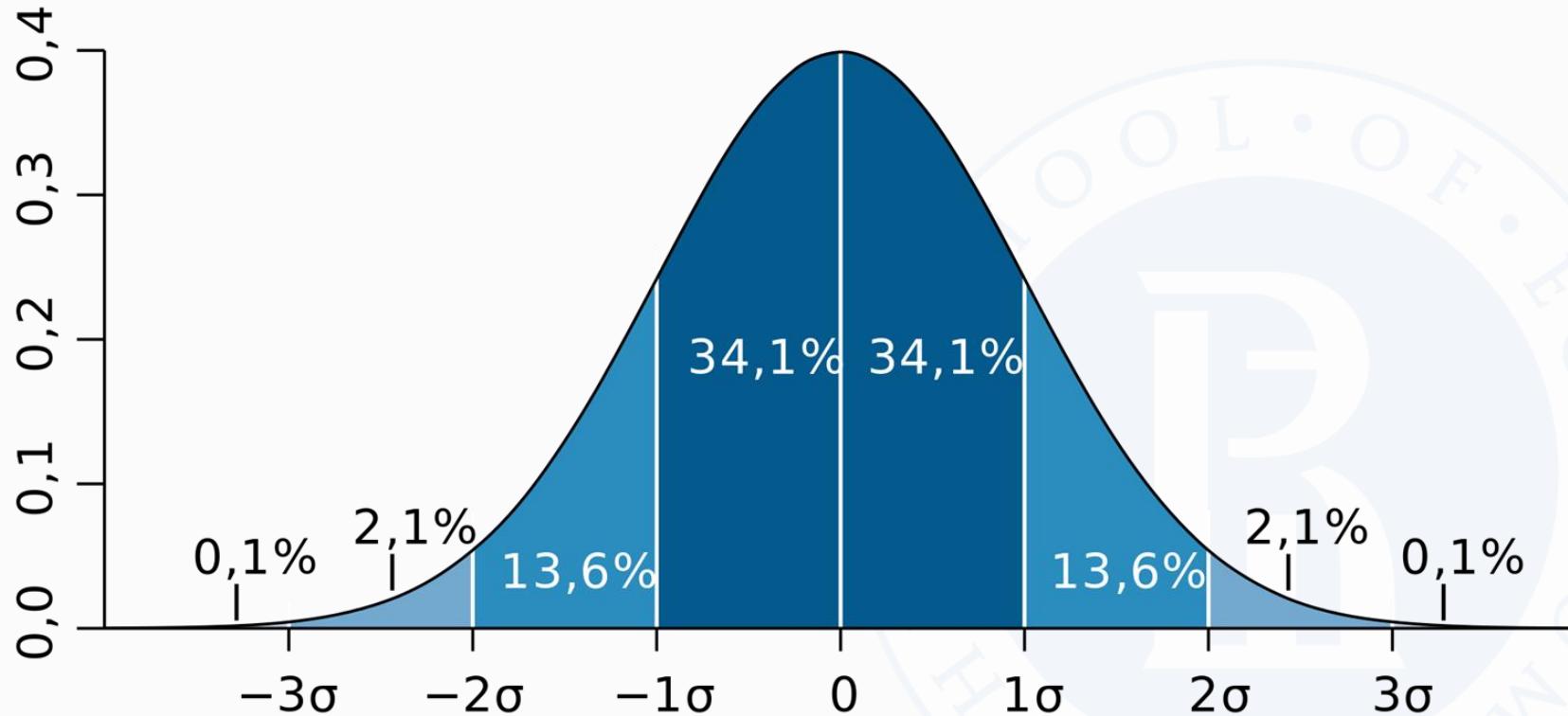


(b) Two-tailed test

Популярные тесты

→ T-test — нормальное распределение

→ Mann-Whitney U-test — непараметрический



Онлайн-метрики. Введение



План



Что такое онлайн-метрика



Сбор данных



Иерархия метрик



Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

Онлайн-метрика

- Метрика, посчитанная по пользовательским активностям
- Какие могут быть активности?
 - Загрузки страниц

Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

→ Какие могут быть активности?

- Загрузки страниц
- Показы элементов на странице
- Переходы между страницами

Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

→ Какие могут быть активности?

- Загрузки страниц
- Показы элементов на странице
- Переходы между страницами
- Время взаимодействия с элементом

Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

→ Какие могут быть активности?

- Загрузки страниц
- Показы элементов на странице
- Переходы между страницами
- Время взаимодействия с элементом
- Покупки

Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

→ Какие могут быть активности?

- Загрузки страниц
- Показы элементов на странице
- Переходы между страницами
- Время взаимодействия с элементом
- Покупки
- Технические счетчики

Онлайн-метрика

→ Метрика, посчитанная по пользовательским активностям

→ Какие могут быть активности?

- Загрузки страниц
- Показы элементов на странице
- Переходы между страницами
- Время взаимодействия с элементом
- Покупки
- Технические счетчики

→ Из всех сигналов нужно собрать единую метрику, по которой можно судить о качестве пользовательского опыта

Блоки на странице

The image shows the Netflix homepage with several UI elements highlighted by red boxes:

- Top Bar:** Includes the Netflix logo, a search bar, and navigation links for Home, TV Shows, Movies, Latest, and My List.
- Profile Area:** Shows a user profile icon with a blue background and a red notification badge indicating 4 notifications.
- Main Content Area:** Features a large banner for the TV series "LUCIFER". The banner includes:
 - A "Play" button with a play icon.
 - A "More Info" button with an info icon.
 - A rating indicator showing "18+".
- Popular on Netflix Sidebar:** A horizontal row of thumbnails for popular shows:
 - LUCIFER (shirtless man)
 - THE UMBRELLA ACADEMY (monkey)
 - how i met your mother (two men)
 - modern family (couple)
 - VIKINGS (viking warrior)

Иерархия метрик

Высокоуровневые



Высокоуровневые метрики

→ Самые важные показатели

- Количество уникальных пользователей за день
- Количество сессий на пользователя

Высокоуровневые метрики

→ Самые важные показатели

- Количество уникальных пользователей за день
- Количество сессий на пользователя

→ Изменения отражаются на бизнес-показателях компании

Высокоуровневые метрики

→ Самые важные показатели

- Количество уникальных пользователей за день
- Количество сессий на пользователя

→ Изменения отражаются на бизнес-показателях компании

→ Долговременный накопительный эффект

Высокоуровневые метрики

→ Самые важные показатели

- Количество уникальных пользователей за день
- Количество сессий на пользователя

→ Изменения отражаются на бизнес-показателях компании

→ Долговременный накопительный эффект

→ Стабильные значения

Иерархия метрик

Высокоуровневые



Прокси-метрики

Прокси-метрики

- Метрики, агрегирующие качество пользовательских сценариев использования сервиса

Прокси-метрики



Метрики, агрегирующие качество пользовательских сценариев использования сервиса

- Отражают качество всего сервиса, при этом их можно «прокрасить»

Прокси-метрики

- Метрики, агрегирующие качество пользовательских сценариев использования сервиса
 - Отражают качество всего сервиса, при этом их можно «прокрасить»
 - Улучшение прокси-метрик должно вести за собой улучшение высокоуровневых метрик

Прокси-метрики

- Метрики, агрегирующие качество пользовательских сценариев использования сервиса
 - Отражают качество всего сервиса, при этом их можно «прокрасить»
 - Улучшение прокси-метрик должно вести за собой улучшение высокоуровневых метрик
 - Учитывают взаимодействие разных компонент между собой

Прокси-метрики

- Метрики, агрегирующие качество пользовательских сценариев использования сервиса
 - Отражают качество всего сервиса, при этом их можно «прокрасить»
 - Улучшение прокси-метрик должно вести за собой улучшение высокоуровневых метрик
 - Учитывают взаимодействие разных компонент между собой
 - Можно растягивать по метрике, улучшая отдельные компоненты

Прокси-метрики

- Метрики, агрегирующие качество пользовательских сценариев использования сервиса
 - Отражают качество всего сервиса, при этом их можно «прокрасить»
 - Улучшение прокси-метрик должно вести за собой улучшение высокоуровневых метрик
 - Учитывают взаимодействие разных компонент между собой
 - Можно растягивать по метрике, улучшая отдельные компоненты
 - Непонятно, как построить идеальную прокси-метрику

Прокси-метрики

→ Пример:

```
p_metric = 1.0 * total_view_time_minutes  
          - 0.5 * n_clicks  
          - 0.2 * n_reloads
```

Прокси-метрики

→ Пример:

```
p_metric = 1.0 * total_view_time_minutes  
        - 0.5 * n_clicks  
        - 0.2 * n_reloads
```

→ Бонус за время просмотра

→ Штраф за лишние клики

→ Штраф за перезагрузку страницы

Прокси-метрики

→ Пример:

```
p_metric = 1.0 * total_view_time_minutes  
          - 0.5 * n_clicks  
          - 0.2 * n_reloads
```

→ Проблемы:

- Как подбирать веса?

Прокси-метрики

→ Пример:

```
p_metric = 1.0 * total_view_time_minutes  
          - 0.5 * n_clicks  
          - 0.2 * n_reloads
```

→ Проблемы:

- Как подбирать веса?
- Какие еще сигналы можно учесть?

Прокси-метрики

→ Пример:

```
p_metric = 1.0 * total_view_time_minutes  
          - 0.5 * n_clicks  
          - 0.2 * n_reloads
```

→ Проблемы:

- Как подбирать веса?
- Какие еще сигналы можно учесть?
- В каких случаях метрика не будет работать?

Иерархия метрик



Метрики отдельных элементов



Замеряют конкретный аспект сервиса



Метрики отдельных элементов

- Замеряют конкретный аспект сервиса
- Не учитывают связи компонент между собой



Метрики отдельных элементов

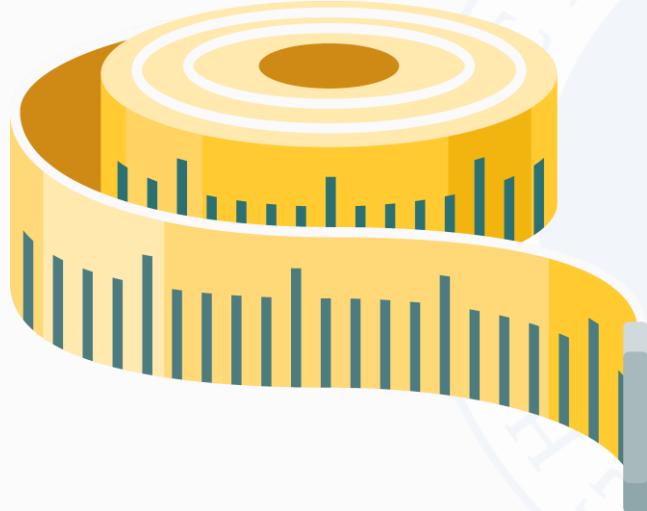
- Замеряют конкретный аспект сервиса
- Не учитывают связи компонент между собой
- Легко «накрутить»



Что мерить?

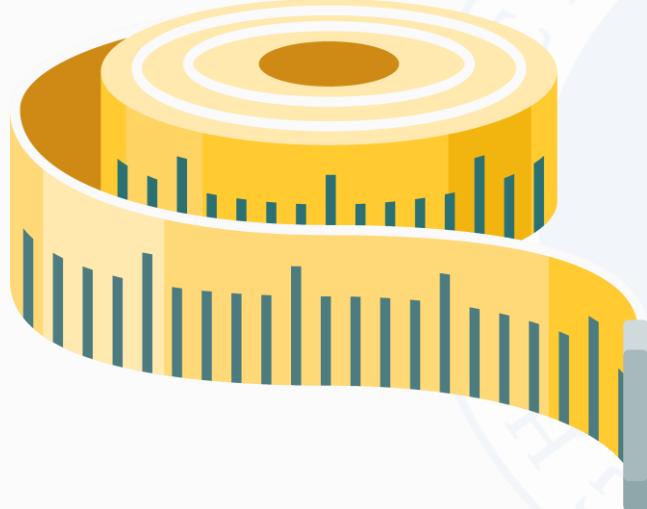


Главные показатели — высокоуровневые метрики



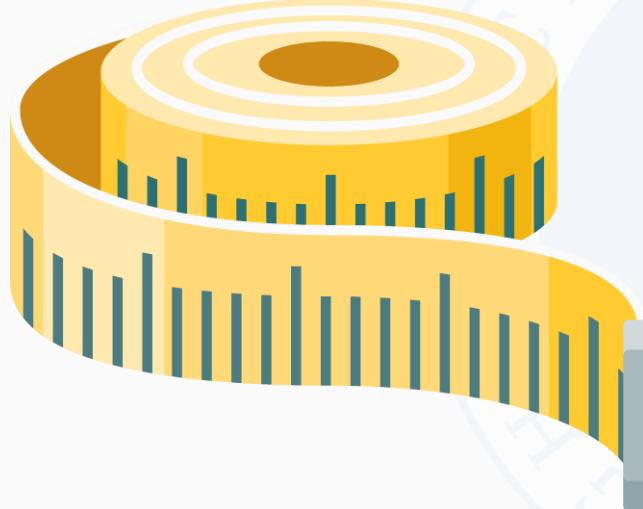
Что мерить?

- Главные показатели — высокоуровневые метрики
- Прокси-метрики позволяют измерять небольшие улучшения, которые не отражаются на высокоуровневых



Что мерить?

- Главные показатели — высокоуровневые метрики
- Прокси-метрики позволяют измерять небольшие улучшения, которые не отражаются на высокоуровневых
- Низкоуровневые метрики помогают локализовать проблему



A/B-тестирование



A/B-тест

→ Тестируем новую функциональность на части пользователей

- Одной группе показываем текущий вариант
- Другой группе — новый

→ Если результаты положительные, можно применять новый функционал на всех пользователях



Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

- Ожидаемый результат ⇒ все хорошо
- Не ожидаемый результат ⇒ поиск причин

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

- Ожидаемый результат ⇒ все хорошо
- Не ожидаемый результат ⇒ поиск причин
 - Оцениваем **несколько** изменений сразу

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

→ Не ожидаемый результат ⇒ поиск причин

- Оцениваем **несколько** изменений сразу
- Неправильные **метрики**

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

→ Не ожидаемый результат ⇒ поиск причин

- Оцениваем **несколько** изменений сразу
- Неправильные **метрики**
- Не учтены **особенности** сервиса

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

→ Не ожидаемый результат ⇒ поиск причин

- Оцениваем **несколько** изменений сразу
- Неправильные **метрики**
- Не учтены **особенности** сервиса
- Неравномерная **аудитория**

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

→ Не ожидаемый результат ⇒ поиск причин

- Оцениваем **несколько** изменений сразу
- Неправильные **метрики**
- Не учтены **особенности** сервиса
- Неравномерная **аудитория**
- Ошибка в **данных**

Ожидания от эксперимента

Фиксируем ожидания по изменениям метрик

→ Ожидаемый результат ⇒ все хорошо

→ Не ожидаемый результат ⇒ поиск причин

- Оцениваем **несколько** изменений сразу
- Неправильные **метрики**
- Не учтены **особенности** сервиса
- Неравномерная **аудитория**
- Ошибка в **данных**
- Ошибка в **параметрах** теста

Параметры эксперимента



Какие пользователи:

- Случайное разбиение
- Новые пользователи тоже должны участвовать

Параметры эксперимента



Какие пользователи:

- Случайное разбиение
- Новые пользователи тоже должны участвовать



Какие платформы:

- Десктоп, мобильные

Параметры эксперимента



Какие пользователи:

- Случайное разбиение
- Новые пользователи тоже должны участвовать



Какие платформы:

- Десктоп, мобильные



Какой процент пользователей:

- Ограничение объема на все эксперименты

Параметры эксперимента



Какие пользователи:

- Случайное разбиение
- Новые пользователи тоже должны участвовать



Какие платформы:

- Десктоп, мобильные



Какой процент пользователей:

- Ограничение объема на все эксперименты



Сколько дней:

- Баланс между скоростью и достоверностью результатов
- Сезонность

Проведение эксперимента

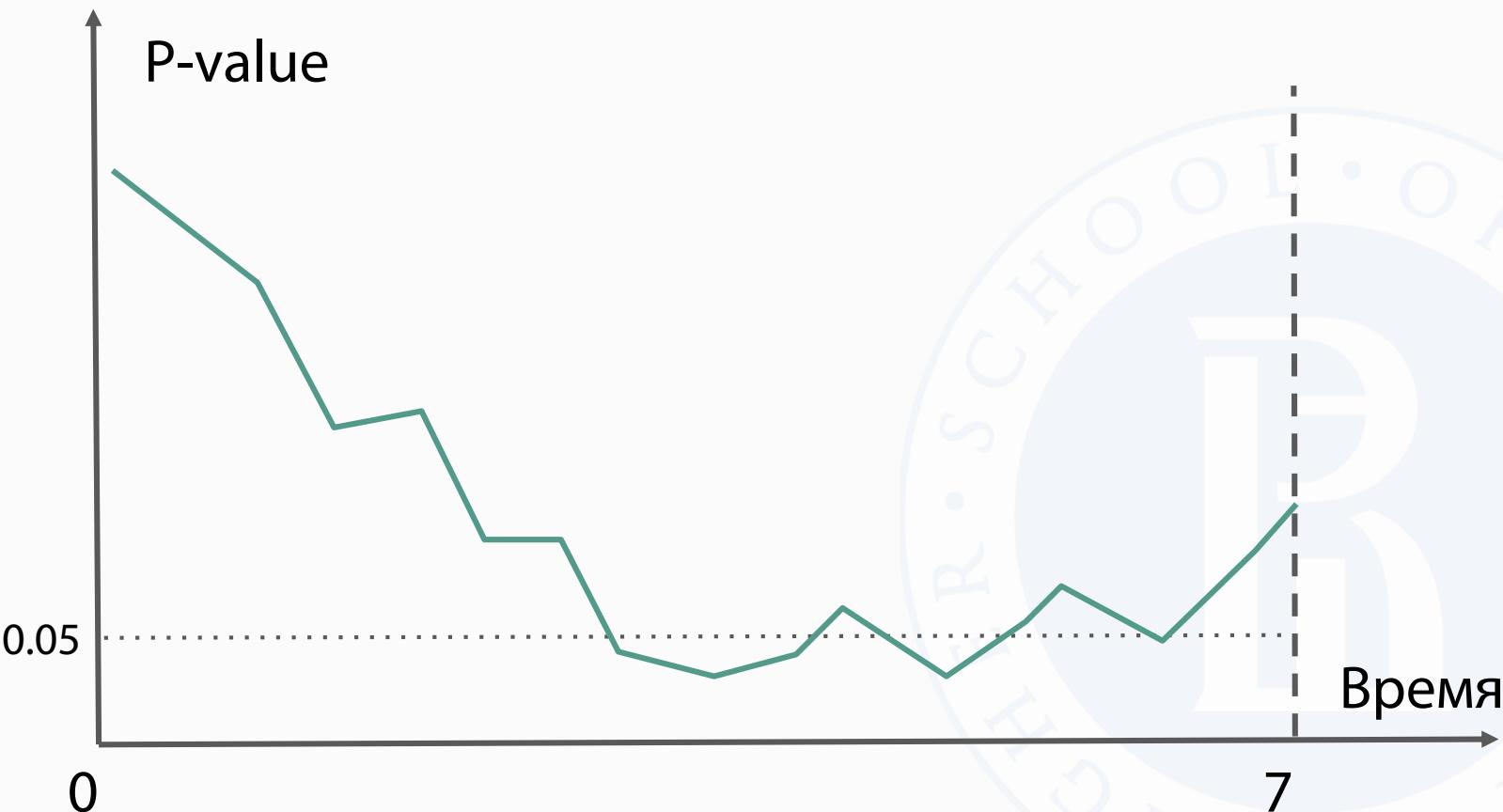
- В случае ошибки тест можно остановить

Проведение эксперимента

- В случае ошибки тест можно остановить
- Получить результат раньше конца теста нельзя

Проведение эксперимента

- В случае ошибки тест можно остановить
- Получить результат раньше конца теста нельзя



Анализ результатов



Посчитать все доступные метрики



Анализ результатов

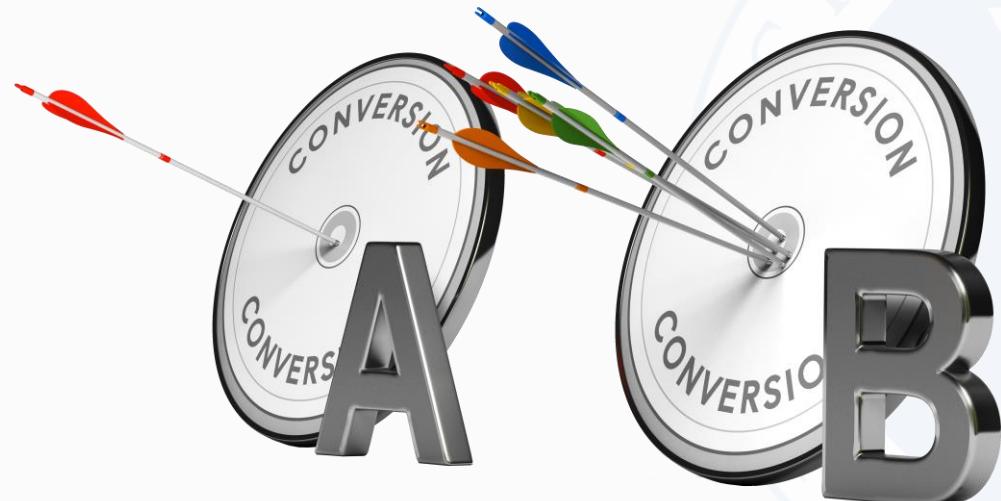
- Посчитать все доступные метрики
- Посмотреть на стабильность результатов

Анализ результатов

- Посчитать все доступные метрики
- Посмотреть на стабильность результатов
- Сопоставить результаты с ожиданиями

Анализ результатов

- Посчитать все доступные метрики
- Посмотреть на стабильность результатов
- Сопоставить результаты с ожиданиями
- Принять решение о том, какая выборка лучше



Подбор онлайн-метрик



План



A/A-тест



Как подбирать метрики



Разладки и мониторинг

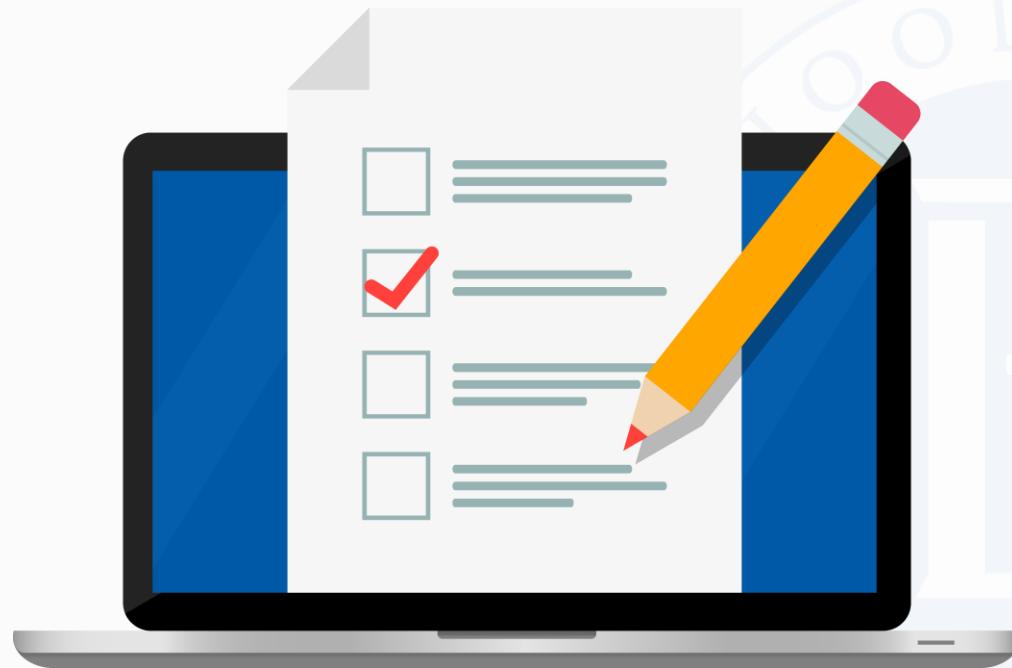


Прокси-метрики

```
p_metric = 1.0 * total_view_time_minutes  
        - 0.5 * n_clicks  
        - 0.2 * n_reloads
```

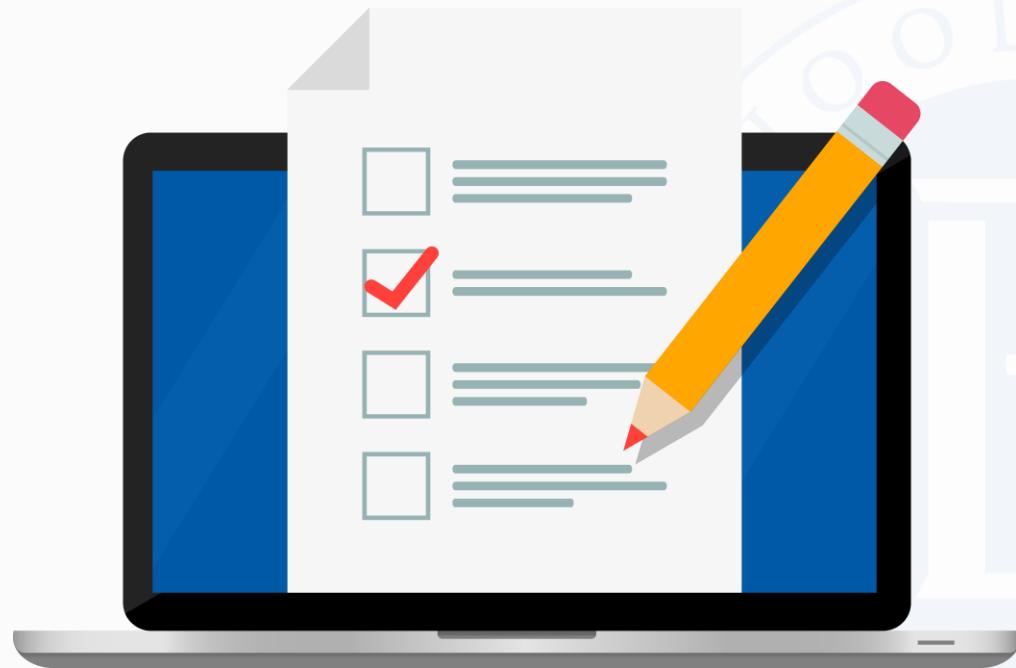
A/A-тест

→ Сравнивается одинаковая функциональность



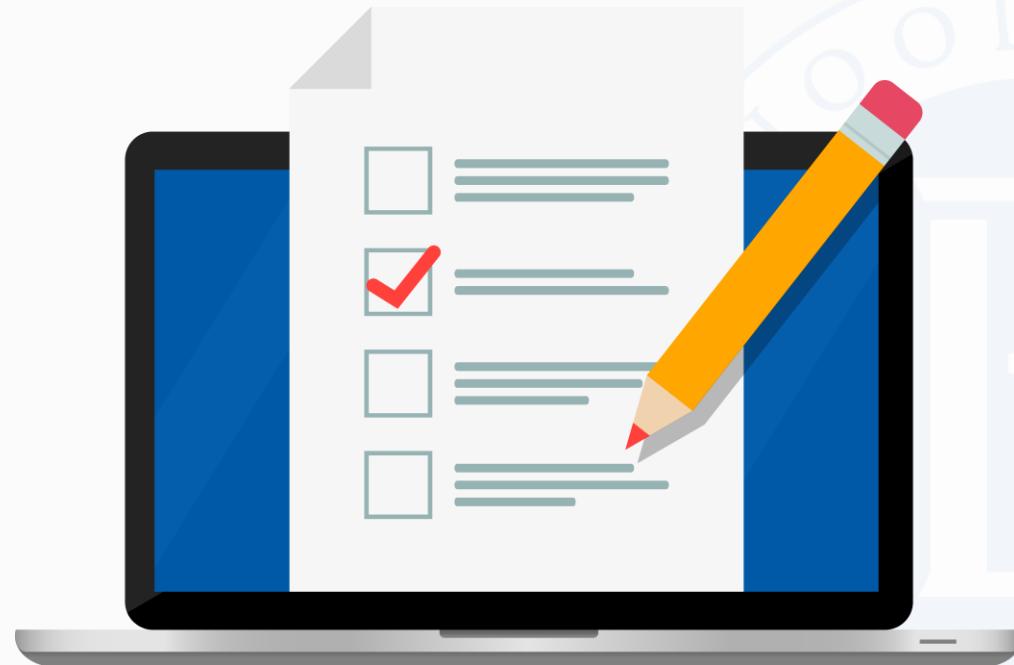
A/A-тест

- Сравнивается одинаковая функциональность
- «Красится» в доле случаев на заданном уровне значимости



A/A-тест

- Сравнивается одинаковая функциональность
- «Красится» в доле случаев на заданном уровне значимости
- Тест для проверки корректности



Как подбирать метрики



Идеальная метрика:

- Всегда показывает правильный результат
- Всегда красится

Как подбирать метрики



Идеальная метрика:

- Всегда показывает правильный результат
- Всегда красится



Такой метрики не существует 😞

Как подбирать метрики



Идеальная метрика:

- Всегда показывает правильный результат
- Всегда красится



Такой метрики не существует 😞



Задача — приблизиться к идеальной метрике

Как подбирать метрики



Идеальная метрика:

- Всегда показывает правильный результат
- Всегда красится



Такой метрики не существует 😞



Задача — приблизиться к идеальной метрике

- Задать метрику продуктовым решением
- Подобрать метрику на основе данных

Набор экспериментов

Нужны данные об уже проведенных экспериментах:

- Суть эксперимента

Набор экспериментов

Нужны данные об уже проведенных экспериментах:

- Суть эксперимента
- Все залогированные данные

Набор экспериментов

Нужны данные об уже проведенных экспериментах:

→ Суть эксперимента

→ Все залогированные данные

→ Вердикт

- Зеленый
- Красный
- Серый

Набор экспериментов

Нужны данные об уже проведенных экспериментах:

- Суть эксперимента
- Все залогированные данные
- Вердикт
 - Зеленый
 - Красный
 - Серый
- Уверенность вердикта

Что ждем от новой метрики

1. Прохождение A/A-теста



Что ждем от новой метрики

1. Прохождение A/A-теста
2. Правильная прокраска экспериментов с известным вердиктом — мощность

Что ждем от новой метрики

1. Прохождение A/A-теста
2. Правильная прокраска экспериментов с известным вердиктом — мощность
3. Любая прокраска экспериментов с неизвестным вердиктом — чувствительность

Что ждем от новой метрики

1. Прохождение A/A-теста
2. Правильная прокраска экспериментов с известным вердиктом — **мощность**
3. Любая прокраска экспериментов с неизвестным вердиктом — **чувствительность**



Действует приоритетность пунктов 1, 2, 3

Что ждем от новой метрики

1. Прохождение A/A-теста
2. Правильная прокраска экспериментов с известным вердиктом — **мощность**
3. Любая прокраска экспериментов с **неизвестным** вердиктом — **чувствительность**

→ Действует приоритетность пунктов 1, 2, 3

→ Перебирая параметры, получаем наилучший вариант метрики

Мониторинг

→ Используем онлайн-метрики не только для А/В-тестов, но и для постоянного мониторинга качества

Мониторинг

- Используем онлайн-метрики не только для А/В-тестов, но и для постоянного мониторинга качества
- По каждому показателю определяем математическое ожидание

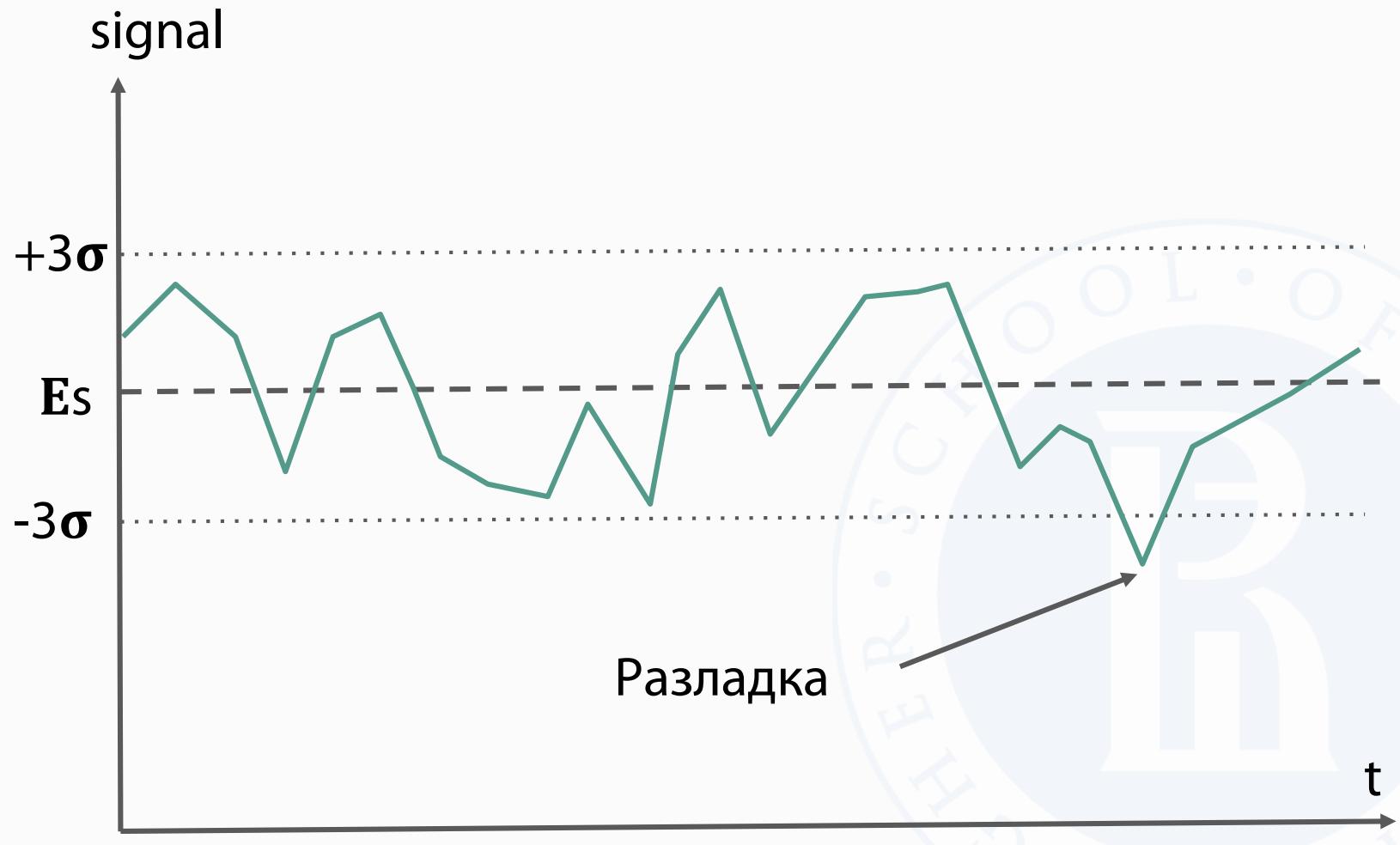
Мониторинг

- Используем онлайн-метрики не только для А/В-тестов, но и для постоянного мониторинга качества
- По каждому показателю определяем математическое ожидание
- Сильное отклонение от мат. ожидания — **возможная** проблема сервиса

Мониторинг

- Используем онлайн-метрики не только для А/В-тестов, но и для постоянного мониторинга качества
- По каждому показателю определяем математическое ожидание
- Сильное отклонение от мат. ожидания — **возможная** проблема сервиса
- Учет сезонности

Мониторинг



Онлайн-метрики

- Важный элемент для оценки сервиса
- Используются как в А/В-тестах, так и в мониторингах
- Метрики можно подбирать на основе реальных данных

Оффлайн-метрики



План



Что такое онлайн-метрика



Подготовка набора данных



Краудсорсинг



Метрики без разметки



Оффлайн-метрика



Метрика, посчитанная по размеченному набору данных

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Зачем нужен онлайн?



Проверяем гипотезы **не** на пользователях



Зачем нужен онлайн?

- Проверяем гипотезы **не** на пользователях
- Быстрое время получения результата

Зачем нужен онлайн?

- Проверяем гипотезы **не** на пользователях
- Быстрое время получения результата
- Легче учитывать продуктовые требования

Откуда брать данные?



Сами данные:

- Открытые источники:
[Kaggle](#), [Google Dataset Search](#), [Sklearn.Datasets](#)
- Сcrapинг

Откуда брать данные?



Сами данные:

- Открытые источники:

Kaggle, Google Dataset Search, Sklearn.Datasets

- Сcrapинг



Целевая переменная (таргет):

- Открытые источники

- Сcrapинг

Откуда брать данные?



Сами данные:

- Открытые источники:

Kaggle, Google Dataset Search, Sklearn.Datasets

- Сcrapинг



Целевая переменная (таргет):

- Открытые источники

- Сcrapинг

- Применить готовую модель к данным

Откуда брать данные?



Сами данные:

- Открытые источники:

Kaggle, Google Dataset Search, Sklearn.Datasets

- Сcrapинг



Целевая переменная (таргет):

- Открытые источники

- Сcrapинг

- Применить готовую модель к данным

- Объемлющий набор данных

Откуда брать данные?



Сами данные:

- Открытые источники:

Kaggle, Google Dataset Search, Sklearn.Datasets

- Сcrapинг



Целевая переменная (таргет):

- Открытые источники

- Сcrapинг

- Применить готовую модель к данным

- Объемлющий набор данных

- Разметить самостоятельно

Откуда брать данные?



Сами данные:

- Открытые источники:
[Kaggle](#), [Google Dataset Search](#), [Sklearn.Datasets](#)
- Сcrapинг



Целевая переменная (таргет):

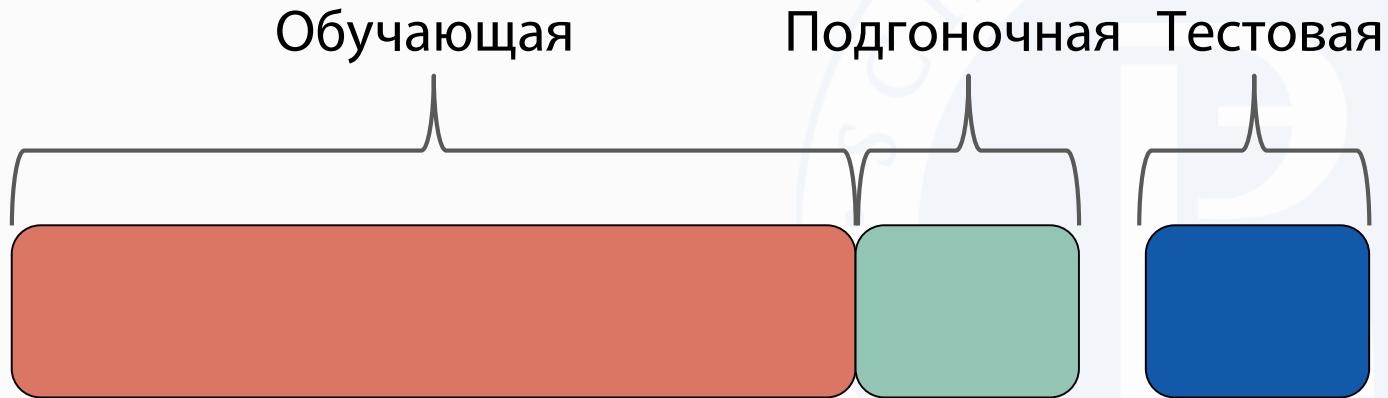
- Открытые источники
- Сcrapинг
- Применить готовую модель к данным
- Объемлющий набор данных
- Разметить самостоятельно
- Краудсорсинг

Разделение на выборки



Тренировочная, подгоночная, тестовая

- Обучаем на тренировочной
- Проверяем на подгоночной
- Итоговый замер на тестовой
- Задача — предсказать сигнал на основе признаков



Формулировка таргета

→ Чем сложнее таргет, тем лучше

Формулировка таргета

- Чем сложнее таргет, тем лучше
- Особенности простого сигнала:
 - На простой таргет легко переобучиться
 - Простой таргет ограничивает рост

Формулировка таргета

→ Чем сложнее таргет, тем лучше

→ Особенности простого сигнала:

- На простой таргет легко переобучиться
- Простой таргет ограничивает рост

→ Особенности сложного сигнала:

- Может быть недостижим
- Потребуются промежуточные прокси-сигналы

Примеры сигналов

Хотим собрать галерею с красивыми картинками



Примеры сигналов

Хотим собрать галерею с красивыми картинками

→ «Хороший» сигнал:

- Качество картинки
- Эстетическая привлекательность картинки

Примеры сигналов

Хотим собрать галерею с красивыми картинками

→ «Хороший» сигнал:

- Качество картинки
- Эстетическая привлекательность картинки

→ «Плохой» сигнал:

- Размер картинки
- Цветность картинки

Краудсорсинг



Люди могут помочь собрать нужный сигнал

Яндекс Толока 

Краудсорсинг

ⓘ <https://toloka.yandex.ru/task/1143149/eeb70807-b16d-43b1-93f1-3da244d67894>

Толока

Задания

В работе ①

Мой профиль

Сообщения ①

Форум

0,00 \$ / 0

,00 \$ Привлекательность картинки для видео в выдаче

Гадкий Я 2: Мини-фильм - Миньоны: Страховочные колеса 2013



Эта картинка:

- ① Хорошая
- ③ Плохая

- ② Нормальная
- ④ Не загрузилась

Действительно ли нужна разметка?



Действительно ли нужна разметка?

- Разметку можно не собирать заранее, а получать в режиме реального времени



Действительно ли нужна разметка?

- Разметку можно не собирать заранее, а получать в режиме реального времени
 - Внешние сервисы — например, поисковик



Действительно ли нужна разметка?

- Разметку можно не собирать заранее, а получать в режиме реального времени
- Внешние сервисы — например, поисковик
 - Краудсорсинг



Разметка через внешний сервис

- У нас есть интернет-издание про инвестиции

Разметка через внешний сервис

- У нас есть интернет-издание про инвестиции
- Используем поисковик как внешний сервис

Разметка через внешний сервис

→ У нас есть интернет-издание про инвестиции

→ Используем поисковик как внешний сервис

→ Чем выше мы показываемся в результатах выдачи, тем лучше



Разметка через внешний сервис

- Соберем набор запросов по нашей тематике
 - Как инвестировать на бирже
 - ETF и БПИФ
 - Брокерский счет
 - ...

Разметка через внешний сервис

- Соберем набор запросов по нашей тематике
 - Как инвестировать на бирже
 - ETF и БПИФ
 - Брокерский счет
 - ...

- Скачаем выдачу в поиске по этим запросам

Разметка через внешний сервис



Соберем набор запросов по нашей тематике

- Как инвестировать на бирже
- ETF и БПИФ
- Брокерский счет
- ...



Скачаем выдачу в поиске по этим запросам



Метрика — сумма позиций нашего издания на выдаче по всем запросам:

- Чем меньше, тем лучше
- Если нашего сайта вообще нет, ставим «бесконечность»

Разметка через внешний сервис

→ Проблемы:

- Как сформировать набор запросов?



Разметка через внешний сервис



Проблемы:

- Как сформировать набор запросов?
- Почему мы уверены, что чем выше, тем лучше?



Разметка через внешний сервис



Проблемы:

- Как сформировать набор запросов?
- Почему мы уверены, что чем выше, тем лучше?
- Смешивается качество нашего ресурса и сайтов конкурентов



Разметка через внешний сервис



Проблемы:

- Как сформировать набор запросов?
- Почему мы уверены, что чем выше, тем лучше?
- Смешивается качество нашего ресурса и сайтов конкурентов
- Можно «накрутить» через кликбейт



Разметка через краудсорсинг



То же издание про инвестиции



Разметка через краудсорсинг

→ То же издание про инвестиции

→ Пользователи пишут статьи

Разметка через краудсорсинг

- То же издание про инвестиции
- Пользователи пишут статьи
- Каждый день на главной странице отбираем лучшие

Разметка через краудсорсинг

- То же издание про инвестиции
- Пользователи пишут статьи
- Каждый день на главной странице отбираем лучшие
- Как оценить качество контента на главной странице?

Разметка через краудсорсинг

- То же издание про инвестиции
- Пользователи пишут статьи
- Каждый день на главной странице отбираем лучшие
- Как оценить качество контента на главной странице?
- Попросить людей поставить оценку статьям

Разметка через краудсорсинг

- То же издание про инвестиции
- Пользователи пишут статьи
- Каждый день на главной странице отбираем лучшие
- Как оценить качество контента на главной странице?
- Попросить людей поставить оценку статьям
- Метрика — сумма оценок всех статей на главной

Разметка через краудсорсинг



Проблемы:

- Как вообще составить задание?

Разметка через краудсорсинг



Проблемы:

- Как вообще составить задание?
- Не бесплатно

Разметка через краудсорсинг



Проблемы:

- Как вообще составить задание?
- Не бесплатно
- Оценки субъективны

Разметка через краудсорсинг



Проблемы:

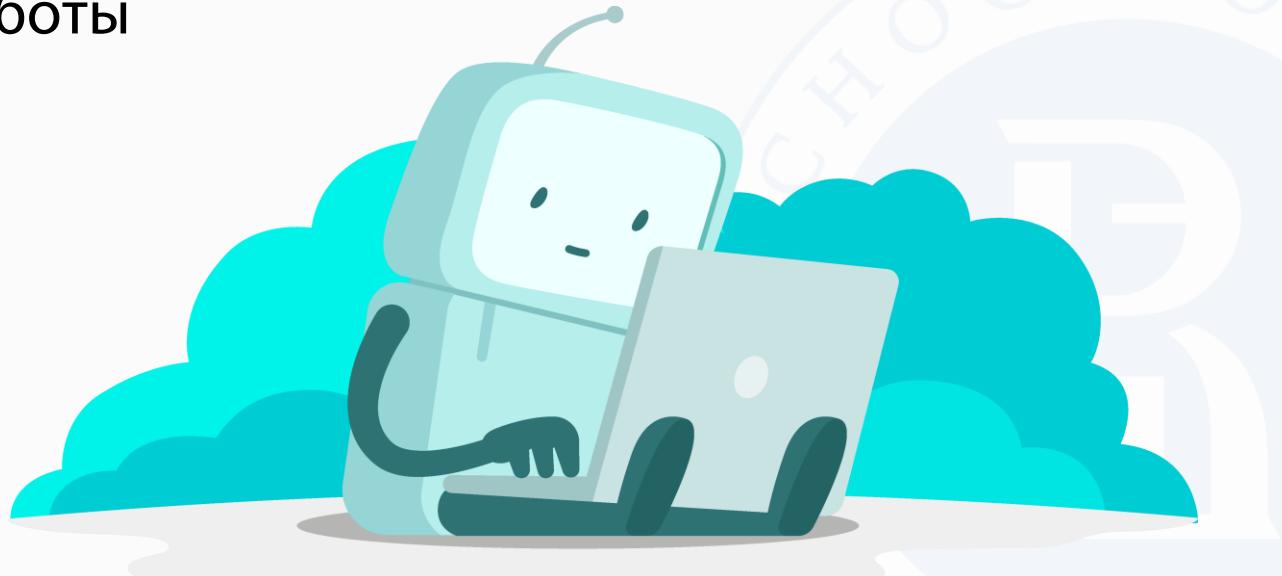
- Как вообще составить задание?
- Не бесплатно
- Оценки субъективны
- Не все будут выполнять задание качественно

Разметка через краудсорсинг



Проблемы:

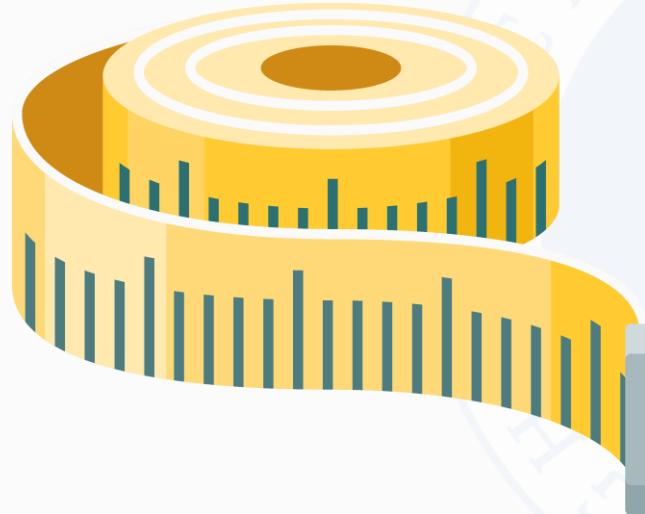
- Как вообще составить задание?
- Не бесплатно
- Оценки субъективны
- Не все будут выполнять задание качественно
- Роботы



Как работать с онлайн-метриками

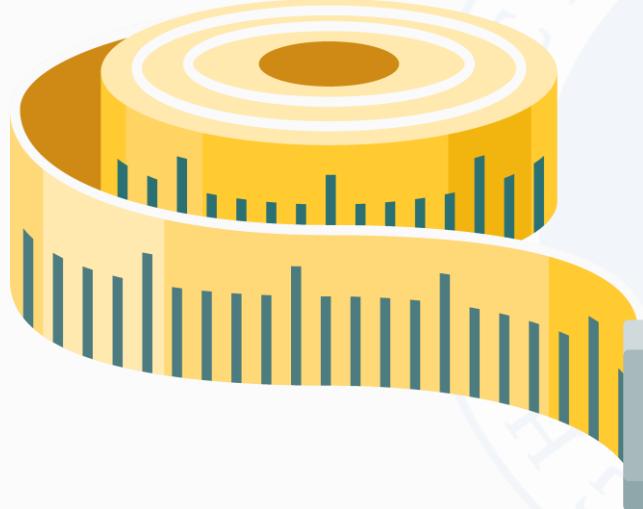


Строим метрики



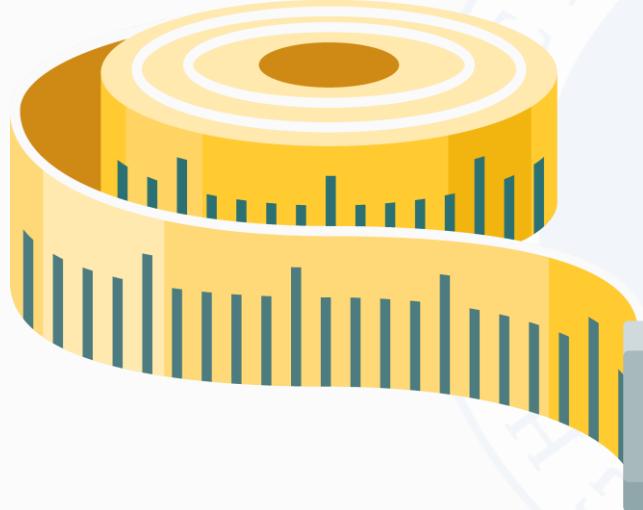
Как работать с онлайн-метриками

- Строим метрики
- Ищем точки роста по замерам на открытом наборе данных



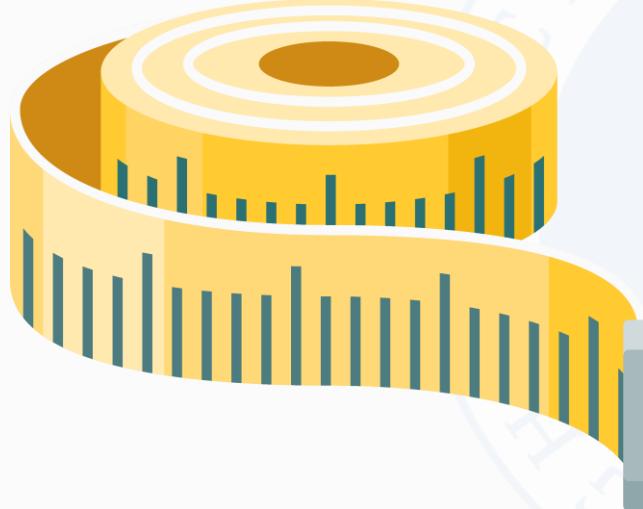
Как работать с онлайн-метриками

- Строим метрики
- Ищем точки роста по замерам на открытом наборе данных
- Проверяем качество на закрытом наборе



Как работать с онлайн-метриками

- Строим метрики
- Ищем точки роста по замерам на открытом наборе данных
- Проверяем качество на закрытом наборе
- Получаем инсайты и улучшаем метрики



Продуктовая сторона



План



Области применения метрик, их расхождения



Соизмеримость метрик и реального качества сервиса



«Красные метрики» — это плохо?



Рост метрик или улучшение продукта

Область применения метрик



Проблемы оффлайна:

- Нерепрезентативность потока данных

Область применения метрик

- Проблемы оффлайна:
 - Нерепрезентативность потока данных
 - Различия продуктового видения и пользовательских сигналов

Область применения метрик

- Проблемы онлайн:
 - Нерепрезентативность потока данных
 - Различия продуктового видения и пользовательских сигналов
 - «Накрутка» компонент метрики

Область применения метрик

- Проблемы онлайн:
 - Нерепрезентативность потока данных
 - Различия продуктового видения и пользовательских сигналов
 - «Накрутка» компонент метрики
 - Персонализация

Область применения метрик

- Проблемы онлайн:
 - Объединение различных сценариев использования сервиса

Область применения метрик



Проблемы онлайн:

- Объединение различных сценариев использования сервиса
- Одновременный учет технических и пользовательских сигналов

Область применения метрик



Проблемы онлайн:

- Объединение различных сценариев использования сервиса
- Одновременный учет технических и пользовательских сигналов
- Долгое время получения результатов

Область применения метрик



Проблемы онлайна:

- Объединение различных сценариев использования сервиса
- Одновременный учет технических и пользовательских сигналов
- Долгое время получения результатов
- Ухудшающие эксперименты

Область применения метрик

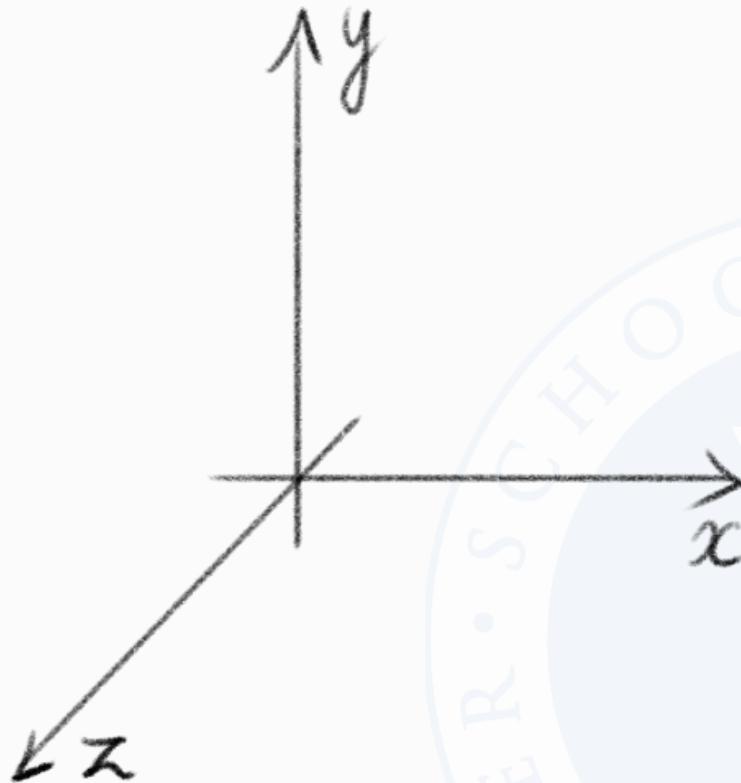
→ Любую метрику можно «сломать»



Сонаправленность типов метрик



Разные типы метрик могут не иметь корреляции



Соnаправленность типов метрик

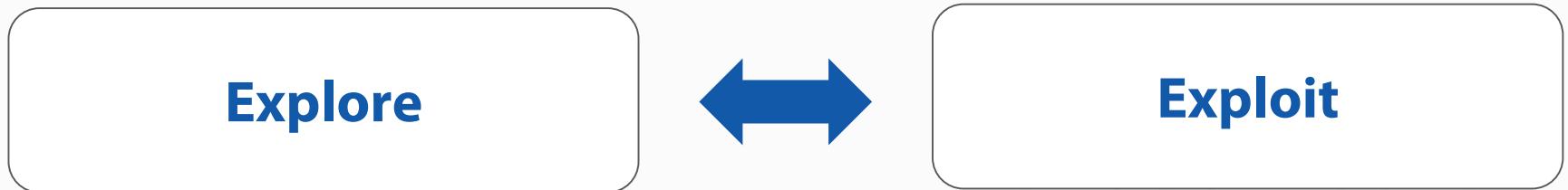
→ Онлайн и офлайн-метрики дополняют друг друга



«Красные» метрики — это плохо?

Метрика	A	B	diff	p-value
Просмотры	130,570	121,601	↓↓ -8,931	0.003
Клики	8,819	7,105	↓ -1,714	0.049

Explore & Exploit



- Поиск новых сигналов
- Рост по текущим метрикам

Взгляд со стороны разных команд



Разработчики

- Метрика — инструмент замера **качества инфраструктуры**
- Цель — повышать значение метрики

Взгляд со стороны разных команд



Разработчики

- Метрика — инструмент замера **качества инфраструктуры**
- Цель — повышать значение метрики



Продуктовые менеджеры

- Метрика — инструмент замера **качества продукта**
- Цель — повышать значение метрики

Взгляд со стороны разных команд



Разработчики

- Метрика — инструмент замера **качества инфраструктуры**
- Цель — повышать значение метрики



Продуктовые менеджеры

- Метрика — инструмент замера **качества продукта**
- Цель — повышать значение метрики



Аналитики

- Метрика — честная оценка **качества взаимодействия** пользователя с сервисом
- Цель — улучшать **качество метрики**

Заключение



Заключение



Для понимания качества продукта нужны данные



На основе данных можно считать разнообразные метрики



Изменения в продукте можно проверять в А/Б-тестах



Математическая статистика поможет сделать обоснованный вывод



Метрики должны помогать развивать продукт и не противоречить здравому смыслу