

A/Б-тесты

2 занятие

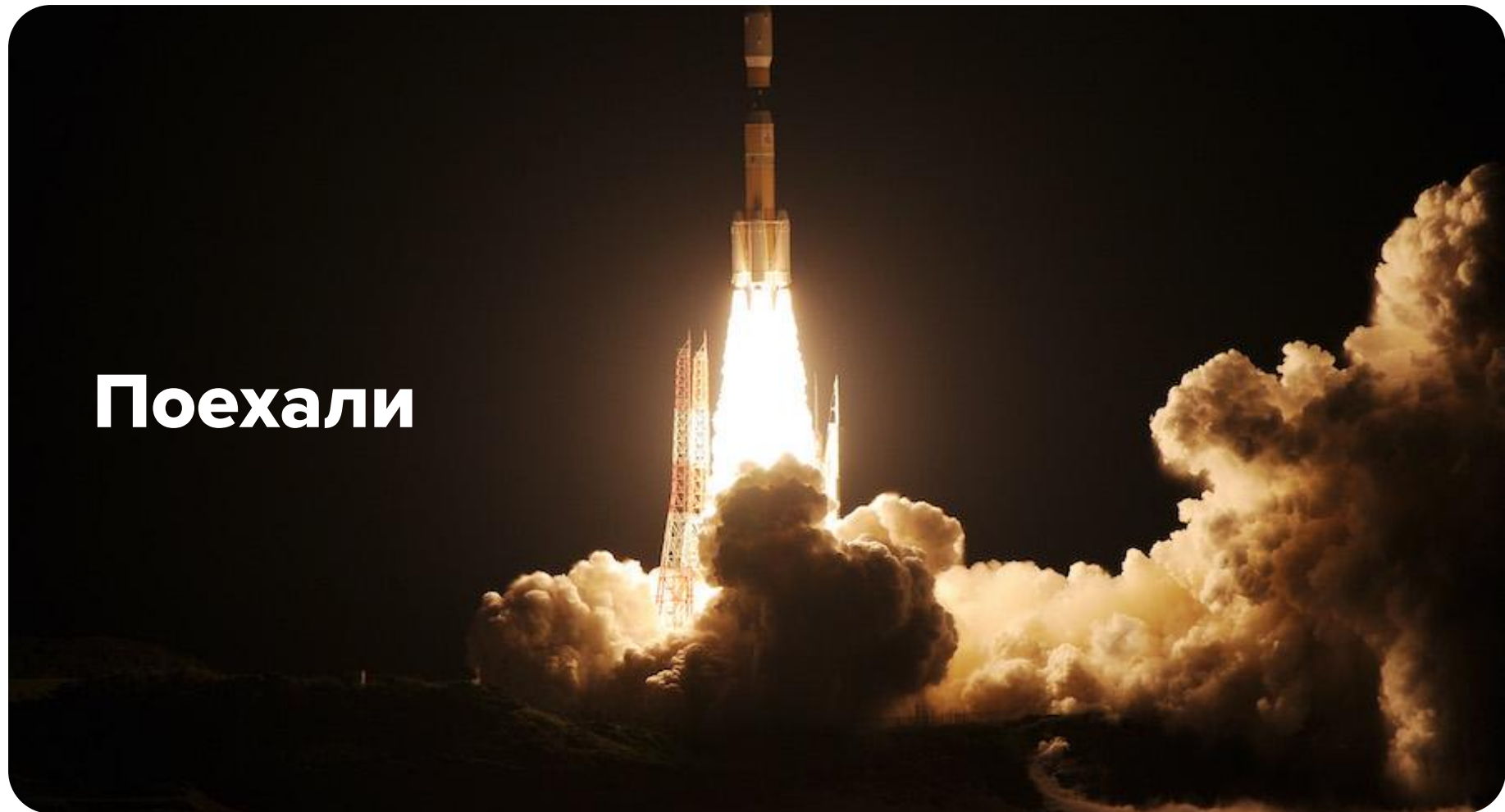
Игорь Полянский
Нетология Сентябрь 2020

Основы А/Б тестирования

Проверка связи



Поехали



О чем говорили раньше:

- **Data-informed** - использует не только данные, но и другие источники
- **Простая гипотеза** = Бизнес-гипотеза = **ЕСЛИ ... ТО ...**
- **Маркетинговая/продуктовая гипотеза** = что делаем + на кого повлияет + какой результат ожидаем + почему ожидаем такой результат
- **Приоритезация гипотез - ICE, PIE**
- **Экспресс-анализ, дерево и план гипотез**

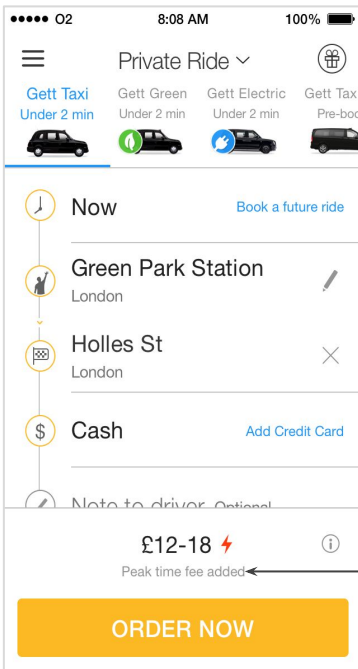


Agenda

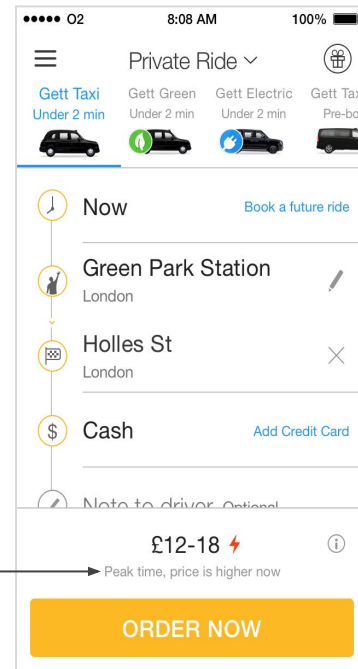
1. Что такое А/Б тесты? Можно без них?
2. Виды А/Б тестов
3. Что можно тестировать, а что нет?
4. Типы метрик для анализа теста
5. Методы агрегаций значений метрики
6. Немного теории и последствий ее незнания

**Что такое АВ
тестирование?**

A/B тестирование - это способ сравнить
несколько версий одной переменной с помощью
измерения реакции субъектов на несколько
вариаций, из которых выбирается наиболее
подходящая



VS



Разные надписи :)

**В чем важность АВ
тестов?**

С помощью АВ тестов можно делать выводы о
причинно-следственной связи

А разве без них нельзя? Разберемся...

Корреляция vs причинно-следственная связь

Корреляция – связь величин, при которой изменение одной из них сопутствует систематическому изменению другой.

Причинно-следственная связь – связь величин, при которой изменение одной величины напрямую влечет изменение другой.

Если между переменными есть причинно-следственная связь, то между ними точно будет корреляция. Обратное неверно.

Корреляция vs причинно-следственная связь: пример

Достоверное известно, что в большинстве стран с доступным высшим образованием продолжительность жизни тоже несколько больше.

- Доступное высшее образование коррелирует с продолжительностью жизни?
- Доступное высшее образование удлиняет жизнь?

Корреляция vs причинно-следственная связь: кейс

Достоверно известно, что в Нью-Йорке цена на билет в метро растет вместе с ростом цены на кусочек пиццы.

Объясните, как связаны эти два события?

Корреляция vs причинно-следственная связь: кейс

Достоверно известно, что в Нью-Йорке цена на билет в метро растет вместе с ростом цены на кусочек пиццы.

Объясните, как связаны эти два события?

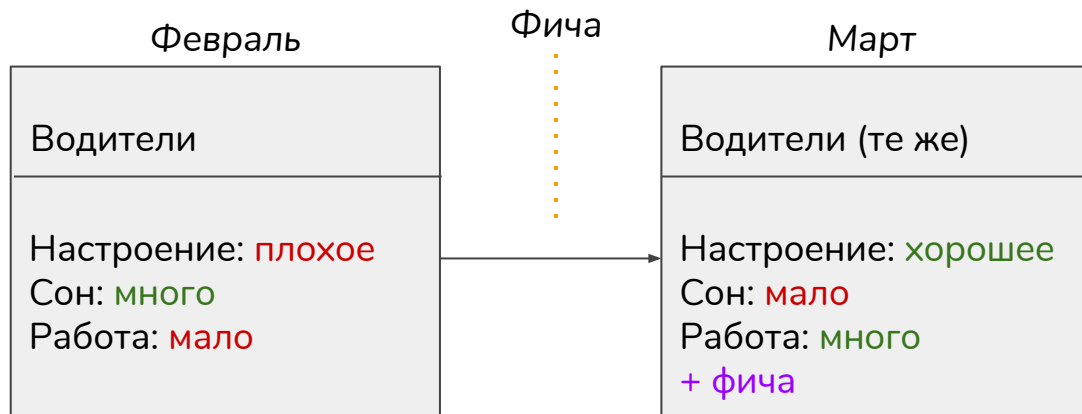
Ответ: цена пиццы коррелируют с ценой на билет в метро, но причиной колебаний на самом деле является инфляция.

Вернемся к А/Б тестам



Наблюдение

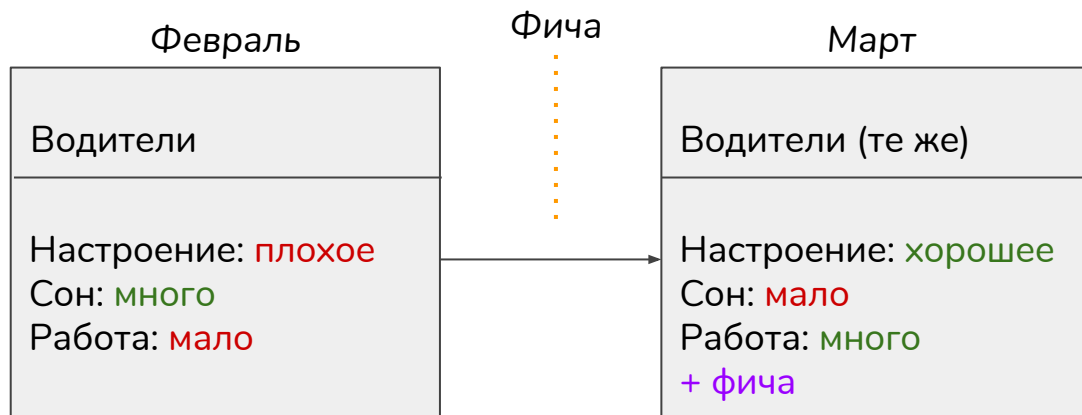
Целевая метрика: поездки



Вывод: водители в среднем стали делать больше поездок, то есть фича крутая?

Наблюдение

Целевая метрика: поездки. Но на нее влияет куча всего!



Вывод: водители в среднем стали делать больше поездок, то есть фича крутая! (нет)

Проблема: многие факторы меняются => невозможно определить причину изменения

А/Б тест

Целевая метрика: поездки. На этот раз группы различаются лишь по фиче

Водители	Водители (те же)
Настроение: хорошее Сон: много Работа: мало	Настроение: хорошее Сон: много Работа: мало + фича

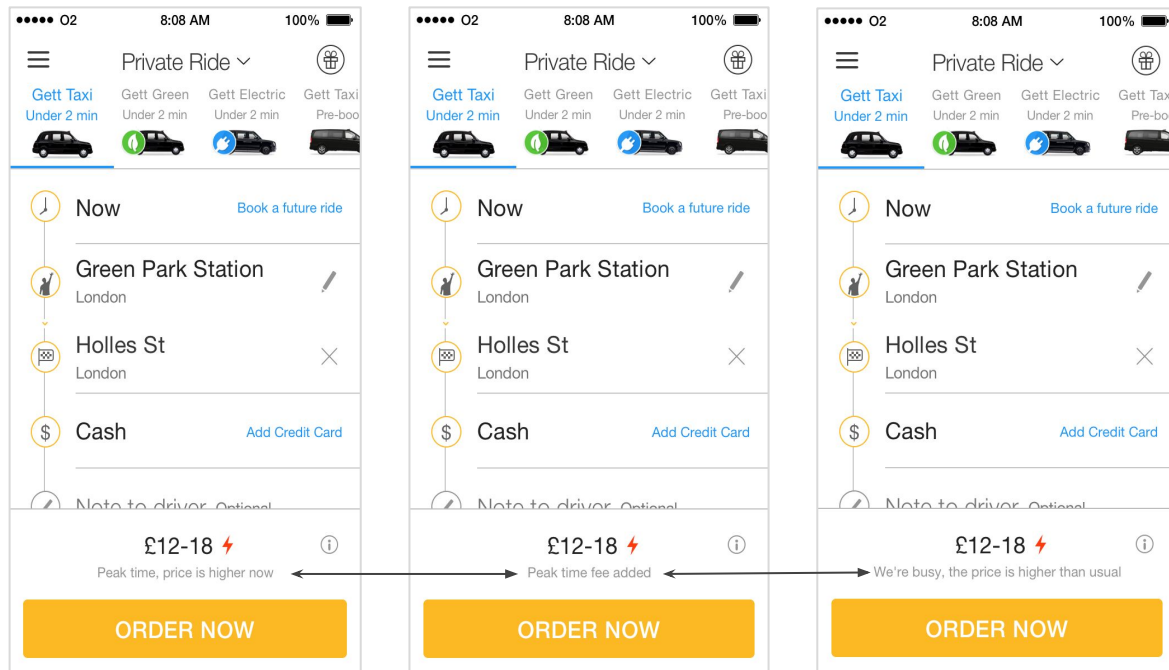
Вывод: водители в среднем стали делать больше поездок, то есть фича крутая! (да)

Проблема: устранена

**А/Б могут сравнить
только две группы?
Или нет?**

Конечно, нет. А/В тесты не ограничиваются
сравнением двух групп - можно сравнить все N.
Однако, при увеличении N, требуется больше трафика
для получения надежных результатов

A/B/C/D/n тест



*Разные надписи :)

Хочу поменять не только надпись. Так можно?



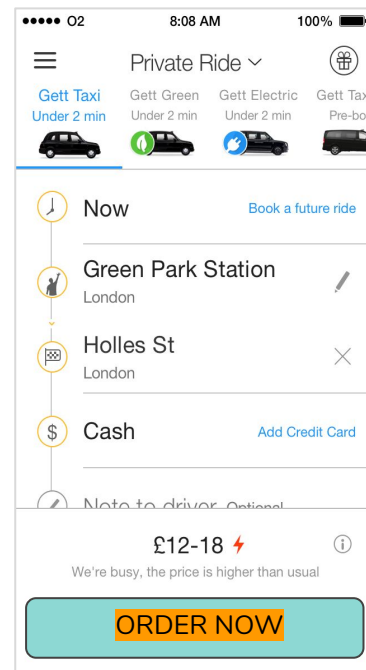
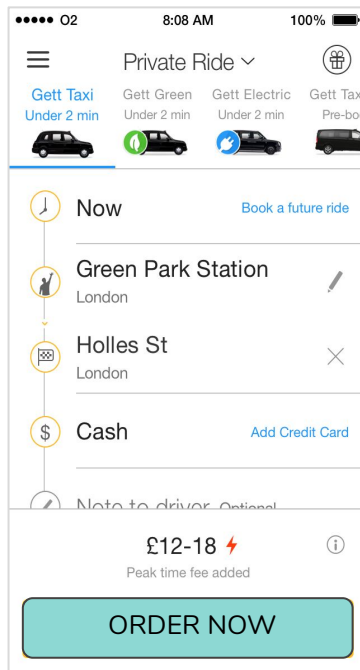
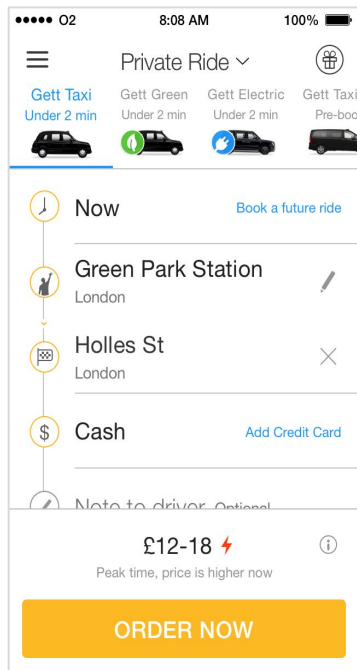
Многомерное тестирование

Позволяет протестировать **группу** изменений.

Например, если бы в прошлом примере мы меняли не только надпись, но и цвет кнопки “Order now”.

Таким образом, мы могли бы найти оптимальное **сочетание цвета кнопки “Order now” и текста** над ней.

Многомерное тестирование: пример



*Каждый вариант содержит уникальное сочетание текста и оформления кнопки "Order now"



QUESTION TIME

**Что можно и нужно А/Б
тестировать, а что нет?**

Поможет ли А/В тест?


Поможет

- “Это или то?”

Не поможет

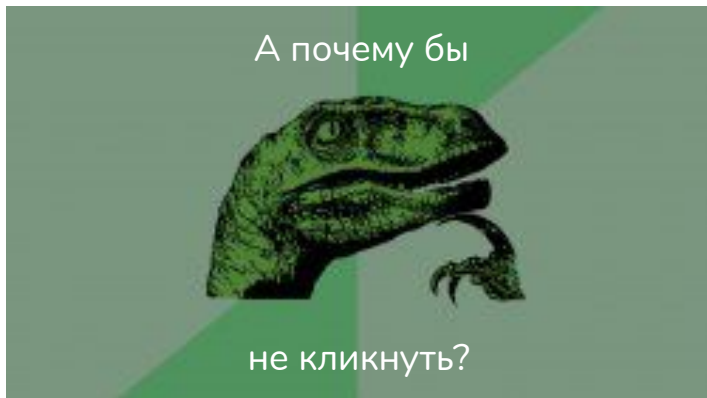
- Принципиально новый опыт*
- Качество продукта
- Долгосрочное изменение

Это или то?

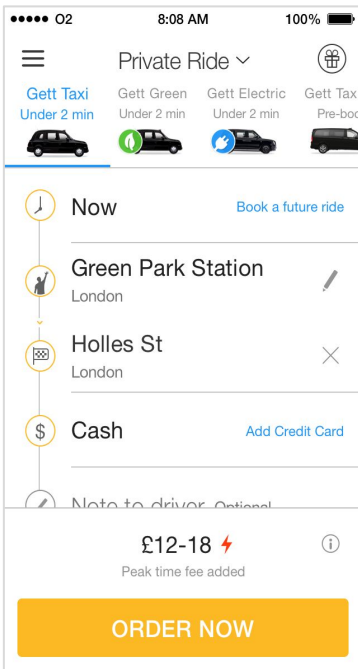
Несколько вариаций, из которых нужно выбрать
лучшую? Однозначно А/В тест 

Принципиально новый опыт

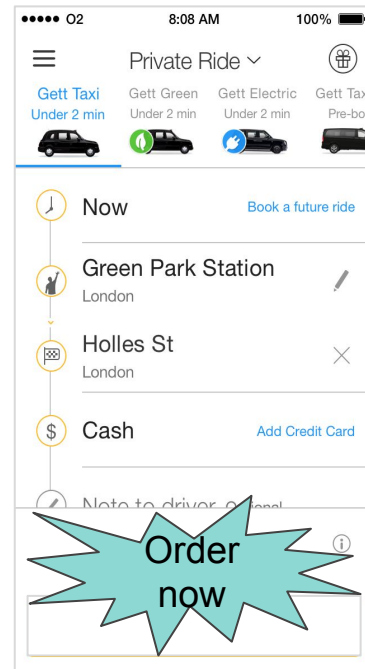
1. Юзеры могут совершать целевое действие просто потому, что фича сильно привлекает внимание (например, яркая кнопка)
2. Юзеры могут не любить изменения



Принципиально новый опыт: пример



VS



Любая новая “видимая” фича - новый опыт?



По идее, да. Можно исключать первые несколько дней из последующего анализа, но мы так обычно не делаем



Почему?



Потому что:

1. Данных не всегда много
2. Юзеры могут использовать продукт с **низкой частотностью**. В таком случае **невозможно убрать все первичные взаимодействия** юзеров с новой функцией

Качество продукта

A/B тест не сможет дать прямой ответ на вопрос,
комфортно ли пользоваться продуктом в настоящий
момент :)

Тут помогут качественные исследования

Долгосрочное изменение

1. За достаточно большой промежуток времени **поведение** юзера **может измениться**
2. **Вряд ли** вы **будете ждать** год, чтобы измерить долгосрочный retention

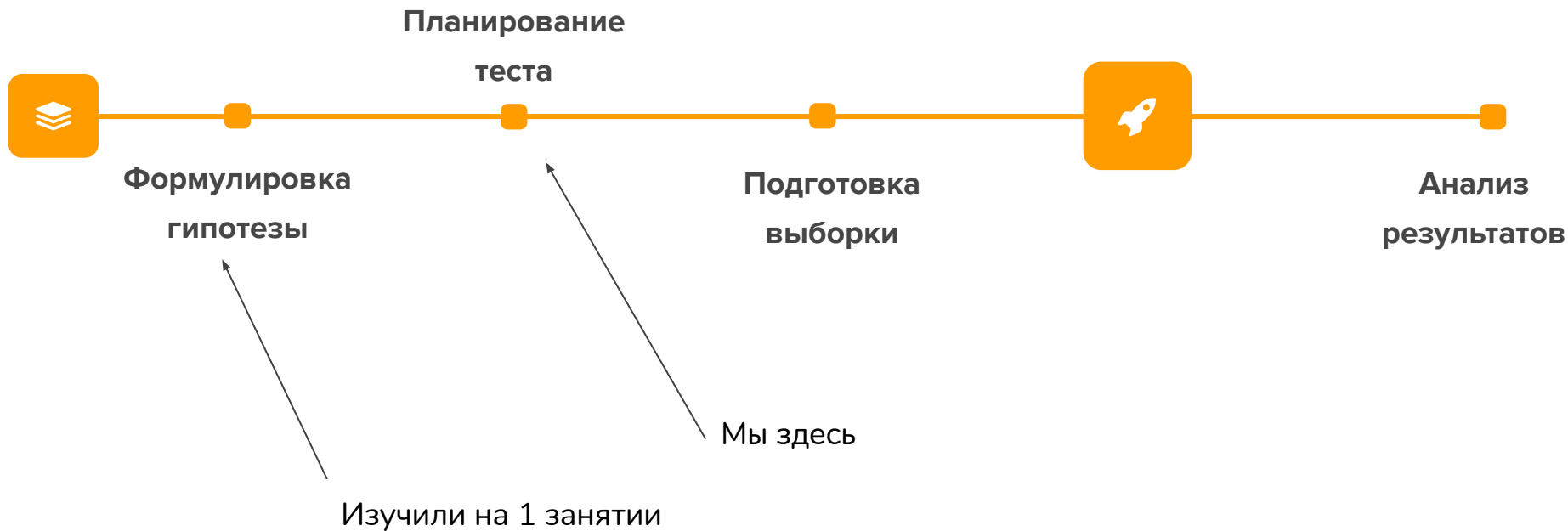




QUESTION TIME

Путь к структурированному A/B тесту

Повторим общий процесс



Итак, как спланировать тест?



Шаг 1: убедиться, что мы понимаем бизнес-цели

Прежде чем работать с гипотезой (или даже ее формулировать) **стоит ответить** на следующий **вопрос:**

- Хорошо ли определена цель?

И только **потом** на **эти:**

- Есть ли баги в продукте, которые могут повлиять на наши метрики?
- Есть ли различия между сегментами юзеров (Android/iOS, Chrome/Safari, итд)? Если есть, они могут повлиять на вывод по тесту (в случае, если в одной из групп будет больше доля пользователей iOS, например)

А давайте поменяем цвет кнопки “Order Now”?

Давайте, но:

1. Чего мы хотим добиться?
2. Поможет ли это достичь OKRs*?



Допустим, цель определена



Подбор метрик: классификация

Выделяют два класса метрик:

- **Первичные:** метрики, позволяющие детально отследить взаимодействие юзера с продуктом, а также высокоуровневые метрики для отслеживания общего состояния бизнеса
- **Вторичные:** метрики, позволяющие убедиться в отсутствии негативного эффекта у фичи. Например, количество юзеров, совершивших какое-либо действие (позже покажу наглядно)

Подбор метрик

Для каждого теста мы в **Gett** ® подбираем:

1. **Минимум** одну первичную метрику

Которая должна **улучшиться** в тестовой группе

2. **Минимум** одну вторичную метрику

Которая должна **не ухудшиться** в тестовой группе

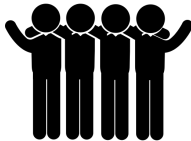
Может, обойдемся первичными метриками?

Нет

Важность вторичных метрик

Первичная метрика: среднее количество поездок

Вторичная метрика: количество водителей, совершивших хотя бы одну поездку



Водители

Группа А (контроль)

Группа Б (тест)

Метрика

Среднее кол-во поездок

5

70

Первичная

Важность вторичных метрик

Первичная метрика: среднее количество поездок

Вторичная метрика: количество водителей, совершивших хотя бы одну поездку



	Группа А (контроль)	Группа Б (тест)	Метрика
Среднее кол-во поездок	5	70	Первичная
Активные водители	80	2	Вторичная
Всего водителей	100	100	

Заметим, что **почти все водители перестали пользоваться приложением в группе Б**, несмотря на высокую активность в среднем



QUESTION TIME

Шаг 2: разбираемся с данными

Данные логируются в
виде events/logs?

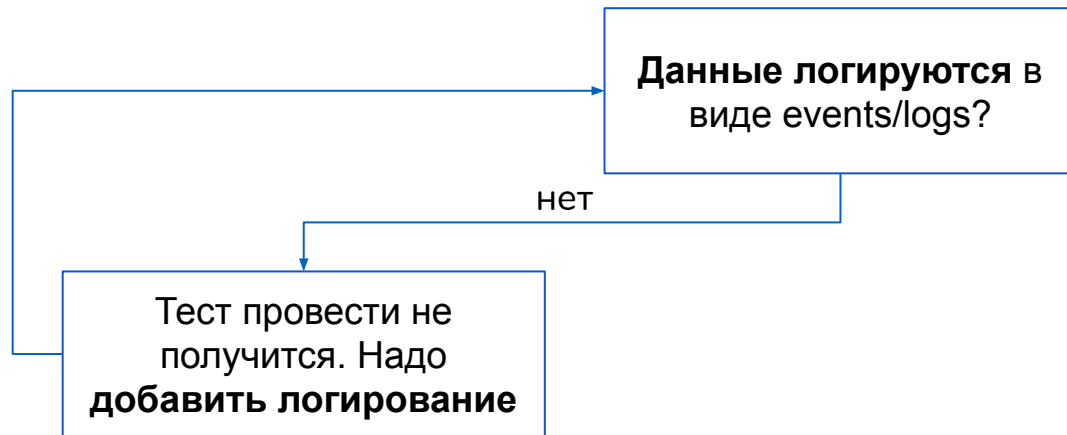
Что такое event?

Event - запись информации о каком-либо событии в продукте.

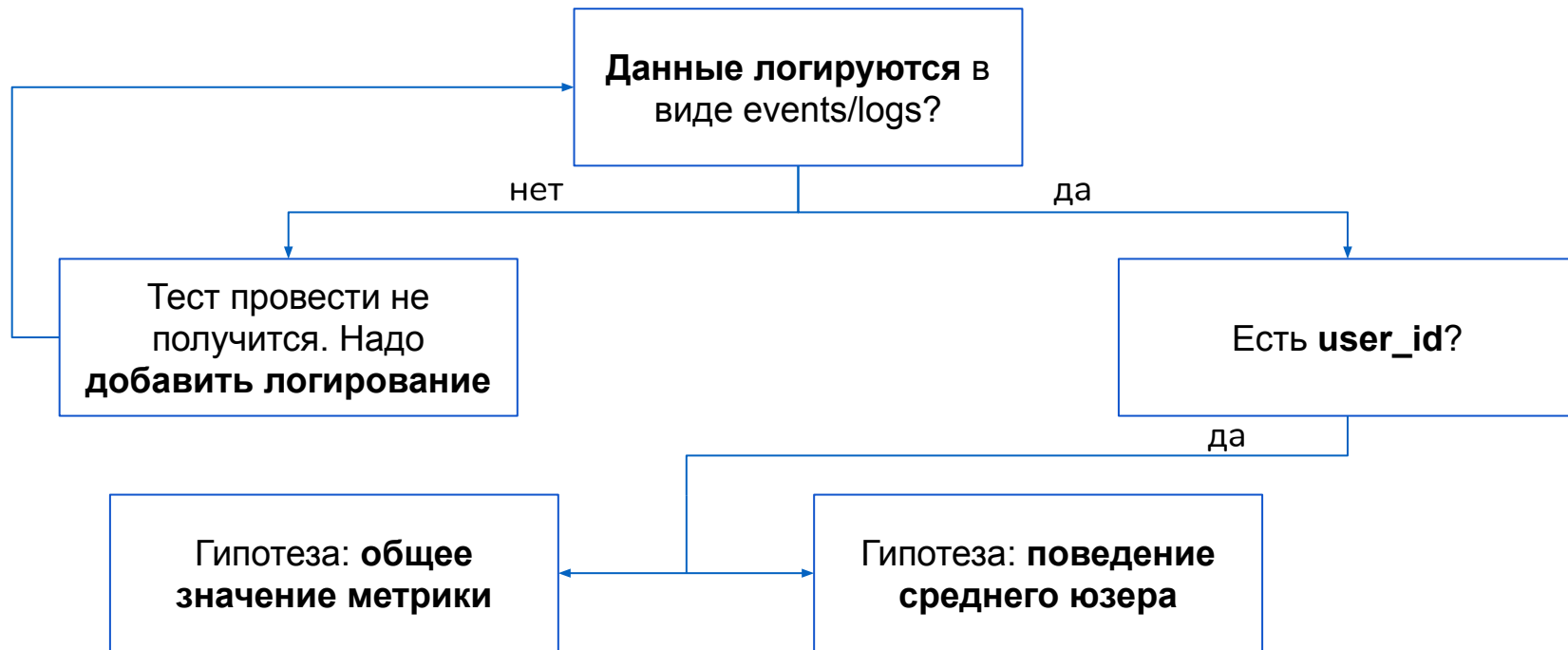
Примеры:

1. Пользователь открыл приложение (содержит информацию о user_id)
2. Пользователь нажал “заказать” (содержит информацию о user_id)
3. Изменение surge в зависимости от плотности заказов (НЕ содержит информацию о user_id)

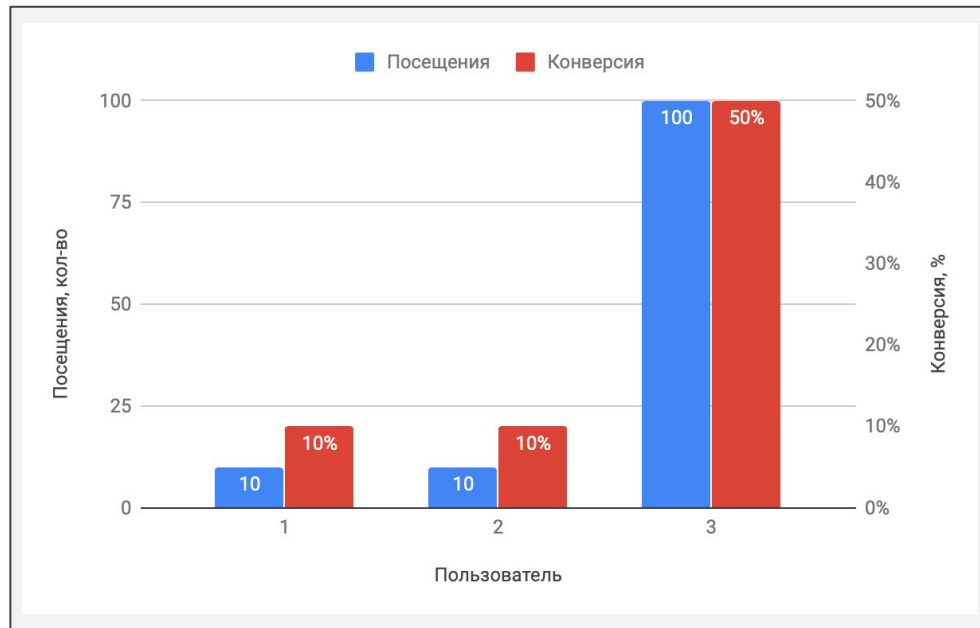
Шаг 2: разбираемся с данными



Шаг 2: разбираемся с данными



“Поведение среднего юзера” vs “общее значение метрики”



Общая конверсия:
 $(1+1+50) / (10+10+100) = 43\%$

Средняя конверсия:
 $(1/10 + 1/10 + 50/100) / 3 = 23\%$

“Поведение среднего юзера” vs “общее значение метрики”: кейс 1

Представьте, что ваша **цель - максимизировать продажи** со страницы. Каждый юзер может зайти на страницу и совершить покупку неограниченное количество раз.

К вам приходит младший аналитик Вова за советом.

Хочет провести А/Б тест для какого-то изменения дизайна.

Спрашивает, на что лучше смотреть: на среднюю поюзерную конверсию или на общее значение конверсии?

Что вы ответите Вове?

“Поведение среднего юзера” vs “общее значение метрики”: кейс 1

Представьте, что ваша **цель - максимизировать продажи** со страницы. Каждый юзер может зайти на страницу и совершить покупку неограниченное количество раз.

К вам приходит младший аналитик Вова за советом.

Хочет провести А/Б тест для какого-то изменения дизайна.

Спрашивает, на что лучше смотреть: на среднюю поюзерную конверсию или на общее значение конверсии?

Что вы ответите Вове?

Поскольку цель - продажи, то не так важно, стала ли какая-то группа пользователей реже совершать конверсию. Важно ли общее значение. Поэтому ответ:

общее значение метрики

“Поведение среднего юзера” vs “общее значение метрики”: кейс 2

Представьте, что ваша **цель** - **изменить поведение водителей** в лучшую сторону, чтобы они принимали бОльшую долю заказов.

К вам приходит младший аналитик Вова за советом.

Хочет провести А/Б тест для какого-то изменения дизайна приложения.

Спрашивает, на что лучше смотреть: на среднюю поюзерную конверсию или на общее значение конверсии?

Что вы ответите Вове?

“Поведение среднего юзера” vs “общее значение метрики”: кейс 2

Представьте, что ваша **цель** - **изменить поведение водителей** в лучшую сторону, чтобы они принимали бОльшую долю заказов.

К вам приходит младший аналитик Вова за советом.

Хочет провести А/Б тест для какого-то изменения дизайна приложения.

Спрашивает, на что лучше смотреть: на среднюю поюзерную конверсию или на общее значение конверсии?

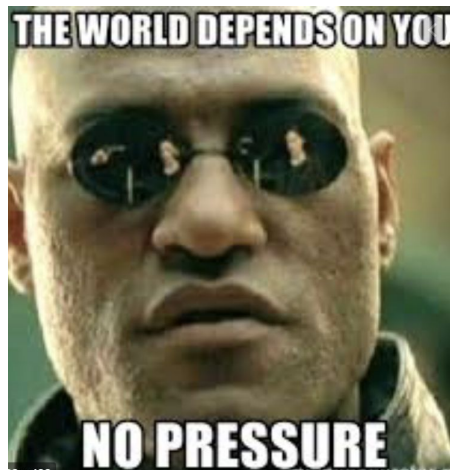
Что вы ответите Вове?

Поскольку цель - поведение водителей, то важно, чтобы в среднем каждый водитель принимал бОльшую долю заказов. Поэтому ответ:

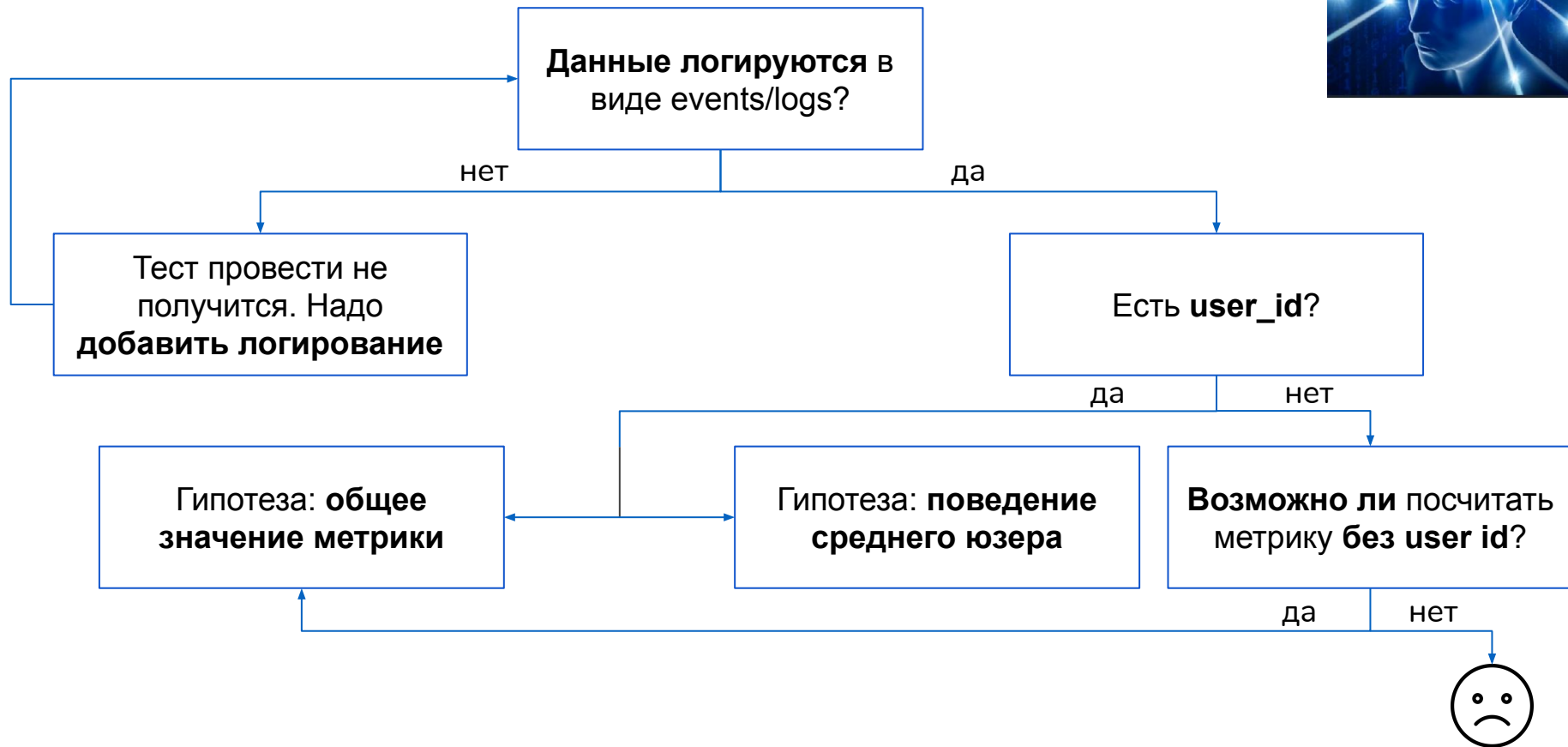
поведение среднего юзера

“Поведение среднего юзера” vs “общее значение метрики”: вывод

Все зависит от ваших целей!



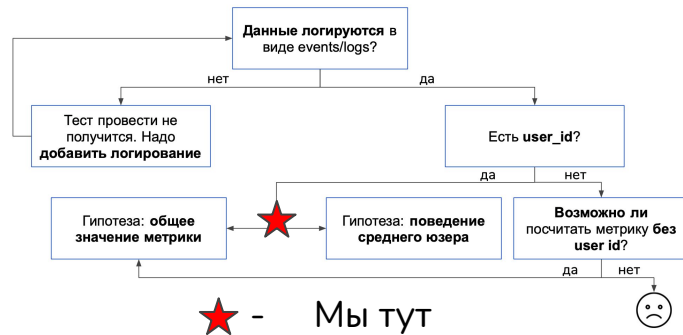
Шаг 2: разбираемся с данными



**Понимая, что одной метрикой можно тестировать
разные гипотезы, пройдемся по дереву еще раз**

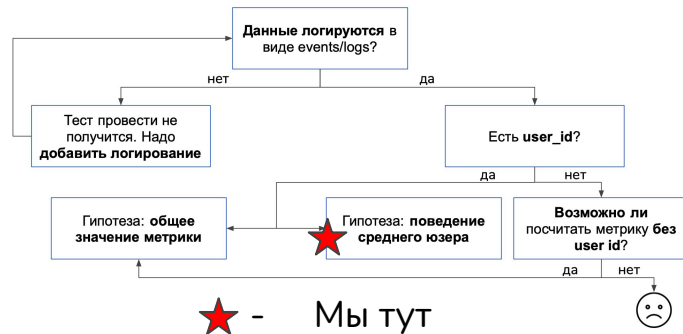


Логирование: ДА
User_id: ДА



Гипотеза: поведение среднего юзера изменилось в лучшую сторону (1)
ИЛИ
Общее значение метрики изменилось в лучшую сторону (2)

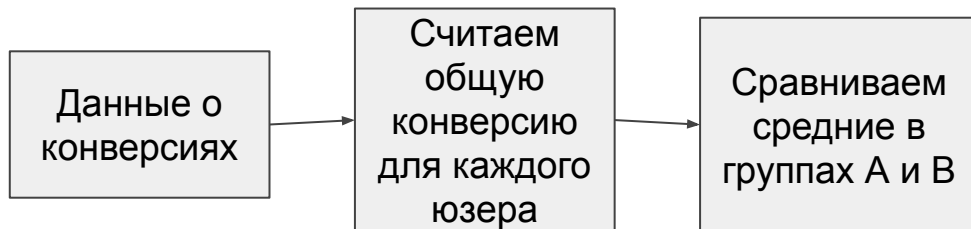
Логирование: ДА
User_id: ДА (1)



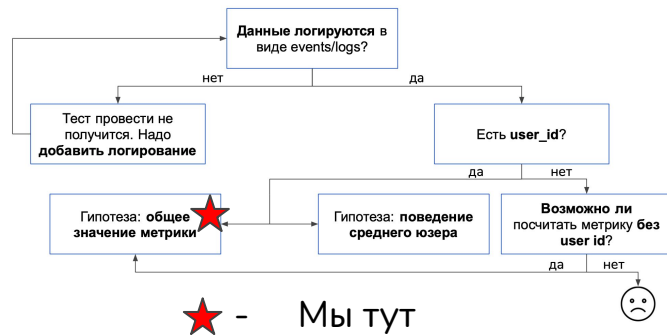
Гипотеза: поведение среднего юзера изменилось в лучшую сторону (1)

Бизнес-цель: среднестатистический юзер должен изменить поведение

Считаем значение метрики для каждого юзера, а затем сравниваем средние по группам



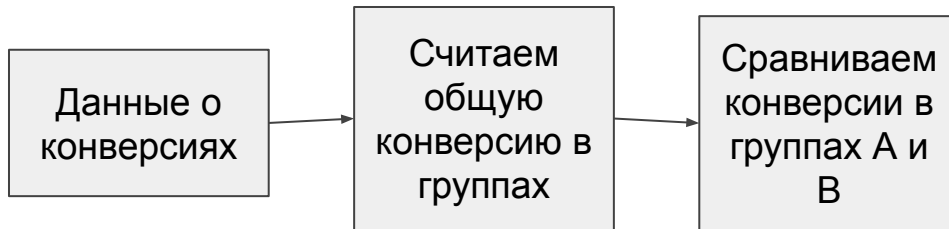
Логирование: ДА
User_id: ДА (2)



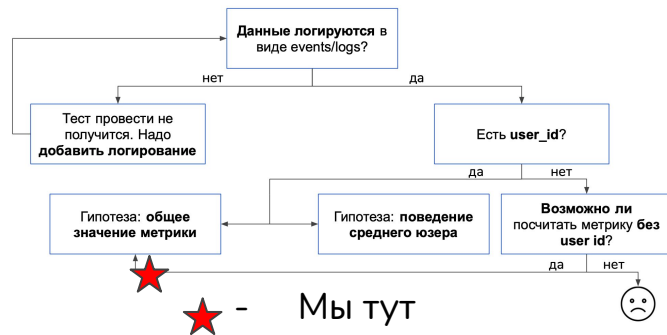
Общее значение метрики изменилось в лучшую сторону (2)

Бизнес-цель: общее значение метрики должно вырасти (даже если отдельным группам пользователей стало хуже)

Считаем общее значение метрики по группам, а затем сравниваем средние по группам



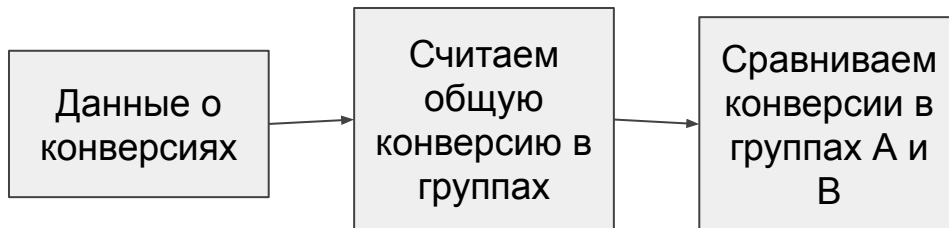
Логирование: ДА
User_id: НЕТ



Гипотеза: общее значение метрики изменилось в лучшую сторону

Бизнес-цель: общее значение метрики должно вырасти (даже если отдельным группам пользователей стало хуже)

Считаем общее значение метрики по группам, а затем сравниваем средние по группам





QUESTION TIME

Немного теории

Статистическая единица

Стат. единица - это тот, на ком будет проводиться **тестирование** (тот элемент, для кого решается, отправить ее в тестовую/контрольную группу)

Может быть идентифицирована (иметь user_id) **или анонимна** (ивент)

Выбор генеральной совокупности

Генеральная совокупность - это выборка стат. единиц, из которых набираются тестовая и контрольная группы

Пример: юзеры, сделавшие не менее 3 поездок за прошлый месяц

Выбор генеральной совокупности: действия

Если ожидается, что изменение повлияет на конкретный сегмент, из него и нужно набирать группы. В противном случае эффект будет размыт

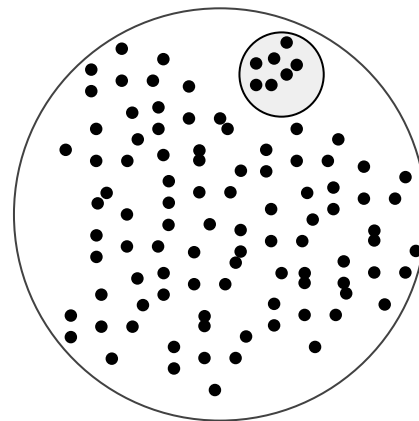
Выбор генеральной совокупности: действия

Если ожидается, что изменение повлияет на конкретный сегмент, из него и нужно набирать группы. В противном случае эффект будет размыт

Обозначения:

● - юзер

Большой круг - все юзеры, маленький - целевая группа



Плохой выбор генеральной совокупности: последствия 1

Младший аналитик Вова хочет оценить эффект от новой фичи для **юзеров, которые пользуются приложением > 1 года**.

Вова взял **всех** пользователей, разбил их на группы. Приходит к вам с выводом, что значимых улучшений нет.

Что вы скажете Вове?

Плохой выбор генеральной совокупности: последствия 1

Младший аналитик Вова хочет оценить эффект от новой фичи для **юзеров, которые пользуются приложением > 1 года**.

Вова взял **всех** пользователей, разбил их на группы. Приходит к вам с выводом, что значимых улучшений нет.

Что вы скажете Вове?

Ответ: Вова выбрал неподходящую генеральную совокупность. Незачем было оценивать значение метрики по всем юзерам, если фича влияет не на всех.

Плохой выбор генеральной совокупности: последствия 2

При проведении сразу нескольких тестов одновременно важно **не допускать пересечений тестовых групп**.

В противном случае, эффекты от разных фичей могут смешаться



Плохой выбор генеральной совокупности: последствия 2: пример

Предисловие: допустим, водителю показывается заказ (оффер) на протяжении 5 сек. За это время он может его либо принять, либо отклонить, либо проигнорировать. Если оффер был принят, водитель может потом его отменить.

Фича 1: даем водителям из тестовой группы 10 сек (вместо 5) на раздумья.

Ожидаем 1: увеличение доли принятых заказов (больше времени прочитать детали)

Фича 2: увеличиваем штраф за отмену заказа со 100 до 200 руб (в реальности штрафов нет)

Ожидание 2: уменьшение отмен, лучше юзер experience

По факту: какие результаты вы ожидаете увидеть по первой фиче?

Плохой выбор генеральной совокупности: последствия 2: пример

Предисловие: допустим, водителю показывается заказ (оффер) на протяжении 5 сек. За это время он может его либо принять, либо отклонить, либо проигнорировать. Если оффер был принят, водитель может потом его отменить.

Фича 1: даем водителям из тестовой группы 10 сек (вместо 5) на раздумья.

Ожидаем 1: увеличение доли принятых заказов (больше времени прочитать детали)

Фича 2: увеличиваем штраф за отмену заказа со 100 до 200 руб (в реальности штрафов нет)

Ожидание 2: уменьшение отмен, лучше юзер experience

По факту: какие результаты вы ожидаете увидеть по первой фиче?

Ответ: никакие :)

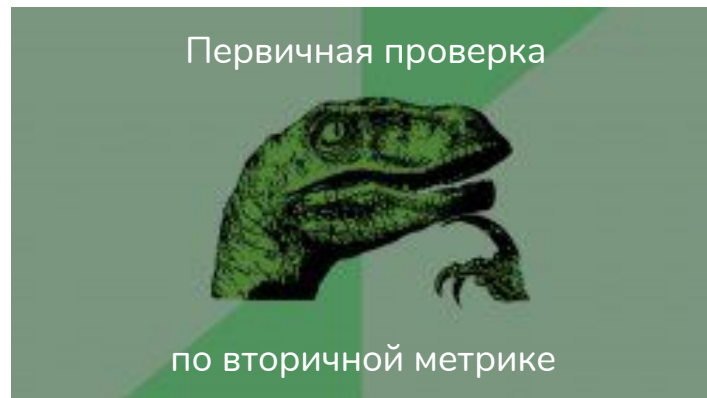
Фича 1 увеличивает долю принятых заказов, а фича 2 - понижает, так как водители принимают осторожнее

Анализ результатов

Первичная проверка



Стоит убедиться, что в каждой из групп тест прошел “по плану”,
оценив вторичные метрики



Проведение статистических тестов

Нет значимости?

Возможно, поможет сегментация:

- a) “Новые” vs “старые” юзеры
- b) Мобильные vs Desktop
- c) Chrome vs Safari vs Firefox
- d) Демография (возрастная группа, пол, локация)

Построение выводов

Вопросы для подведения итогов

- Значимые ли результаты?
- Логично ли изменение? (вдруг баг)
- Бизнес-вывод: даже если метрика осталась на прежнем уровне, возможно, стоит оставить изменение, если оно потенциально имеет долгосрочный эффект* (например, удобство -> retention), или просто добавляет юзерам комфорта

*чтобы это понять, требуется качественное исследование



QUESTION TIME

Практическое задание (домашка)

Практическое задание (45 мин)

- Пройдите опросник по результатам лекции (feedback + цель на курс)
- Выберите топ 5 гипотез из 1 дз
- Продумайте, какие А/Б тесты можно провести, чтобы проверить эти гипотезы:
Для каждого А/Б теста четко сформулируйте:
 - 1) тестируемое изменение
 - 2) бизнес-цель
 - 3) выбранную генеральную совокупность
 - 4) гипотезу (какие метрики должны измениться). Укажите первичные и вторичные метрики
 - 5) метод агрегации значения метрики (поведение среднего юзера vs общее значение метрики)
- Оформите результат в виде таблички (см. след слайд)

Практическое задание: формат

	Гипотеза	Тестируемое изменение	Бизнес-цель	Генеральная совокупность	Первичные метрики	Вторичные метрики	Метод агрегации метрик
1							
2							
3							
4							
5							

**Спасибо за
внимание!
Вы молодцы!**

Вопросы?

Спасибо за внимание

