

Полный пайплайн АВ тестирования

АЛЕКСАНДР САХНОВ

linkedin.com/in/amsakhnov

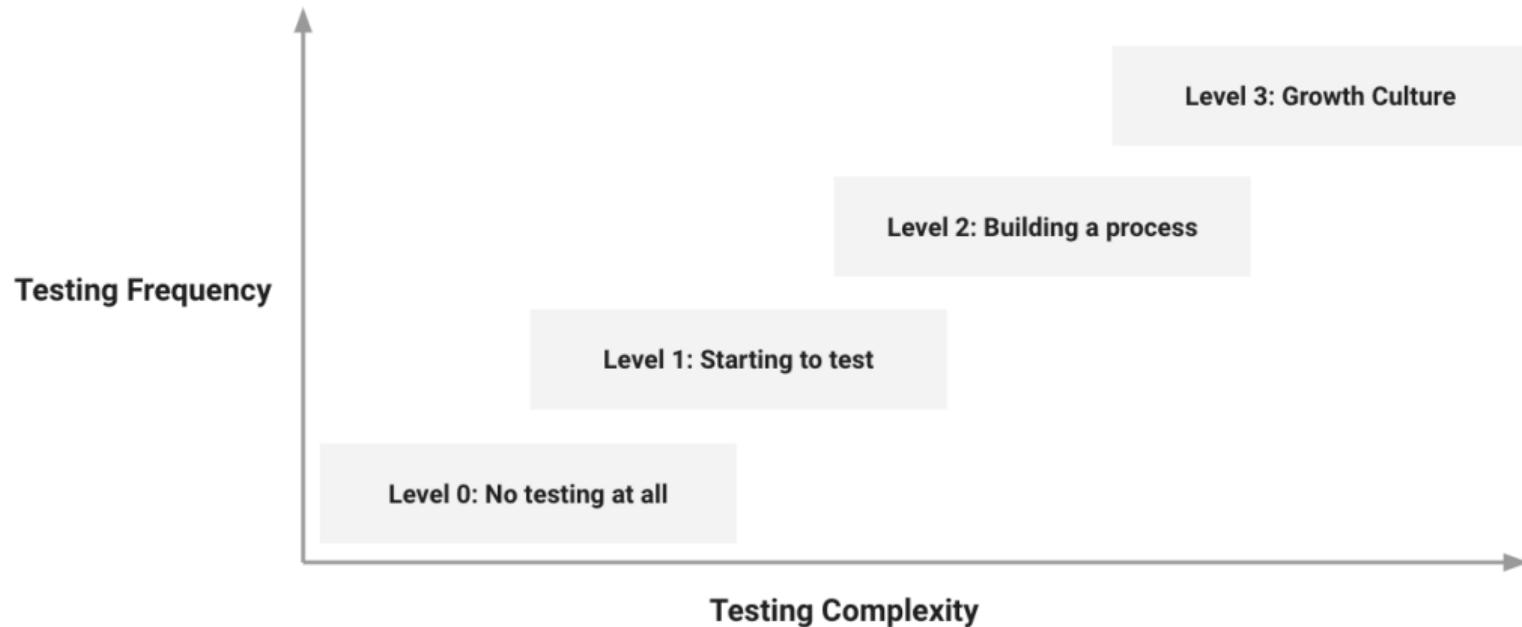
Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

- 1 Внедрение АВ тестирования в крупных компаниях
- 2 Запуск и управление экспериментами
- 3 Подготовка к запуску эксперимента
 - Методы повышения чувствительности
 - Продолжительность эксперимента
 - Техническая реализация
- 4 Мониторинги во время проведения эксперимента
 - Canary deployment
 - Парадокс Симсона
 - Проблема ранней остановки
- 5 Анализ результатов после завершения эксперимента

Уровни внедрения АБ-тестов



Уровни внедрения АВ-тестов

- Level 0** Компании, которые, возможно, только начали свою программу тестирования или все еще набирают команду, но пока не могут проводить АВ эксперименты.
- Level 1** Компании, которые проводят несвязанные тесты. Общие процессы не выстроены.
- Level 2** Команды, которые находятся на пути к построению процесса, принятого всеми соответствующими заинтересованными сторонами. Этот шаг является одним из самых важных, но наименее принятых, поскольку он требует координации между командами, которые часто изолированы. Однако уровень 2, скорее всего, является самым важным шагом к общему успеху компании. Принятие этих процессов позволяет достичь взаимного понимания инициатив и целей во время ваших тестов.
- Level 3** Технологические лидеры отрасли. Это такие компании, как Facebook, Google, Netflix или Яндекс. Эти компании обычно предполагают гармонию между несколькими потоками, что, в свою очередь, сводится к созданию культуры роста в компании. Без грамотной системы тестирования идей рост крупной компании невозможен.

Платформы автоматизации АВ-тестирования

Основные функции платформы АВ-тестирования

- Помогает быстро запускать эксперименты
- Контролирует нежелательные пересечения экспериментов
- Считает метрики, стат. тесты, визуализирует результаты

Цикл тестирования

- Заказчик (аналитик или продакт-менеджер) настраивает параметры эксперимента.
- Сплит-сервис, согласно этим параметрам, раздает клиентскому устройству нужную группу А/В.
- Действия пользователей собираются в сырье логи, которые проходят через агрегацию и превращаются в метрики.
- Метрики «прогоняются» через статистические тесты.
- Результаты визуализируются на дашбордах.

Диаграмма цикла тестирования



Автоматизация проведения экспериментов

Какие этапы проведения эксперимента можно автоматизировать?

- Создание пилота
 - Реализация новой функциональности
 - Определение параметров пилота (ошибки I и II рода, ожидаемый эффект, на кого направлен, метрики, ...)
 - Выбор оптимальной комбинации техник АВ тестирования
 - Расчёт продолжительности пилота
- Проведение пилота
 - Запуск по наличию свободных слотов
 - Постепенное развёртывание
 - Сбор и визуализация метрик
 - Обнаружение аномалий
 - Прекращение эксперимента
- Анализ результатов
 - Визуализация результатов
 - Точечные оценки, доверительные интервалы, статзначимость

Основные этапы эксперимента

Какие фазы проходит эксперимент?

1. **Генерация идей.** На этом этапе нужно придумать что мы хотим поменять и по какой причине это окажет влияние на метрики. На этой фазе генерируется множество идей.
2. **Описание и приоритезация.** Надо транслировать наши инсайты в проверяемые гипотезы. Определить для каждого эксперимента к какому слою он относится и расставить приоритеты.
3. **Подготовка и запуск.** Для каждого эксперимента мы должны установить ограничения, указать ключевые и вспомогательные метрики и определить размер и продолжительность необходимые для достижения статистической значимости.
4. **Анализ результатов.** Нужно обработать результаты эксперимента. Проконтролировать отсутствие ошибок, проанализировать изменение метрик и уровни значимости полученных результатов.
5. **Принятие бизнес-решения.** Достаточен ли полученный результат для внедрения? Можем раскатывать изменения или отвергаем гипотезу? Или нужно дополнительное исследование?
6. **Обобщение опыта.** Все эксперименты (удачные и неудачные) объединяем в общую базу. Время от времени можно перепроверять результаты и отслеживать общий прогресс.

Описание эксперимента

Конфигурируем эксперимент без знаний программирования

В крупных компаниях проводится очень много экспериментов. Часто ими могут заниматься аналитики без глубоких знаний программирования.

Хорошо спроектированный фреймворк тестирования позволяет провести все настройки не погружаясь в код. Так в Airbnb, Авто и других компаниях используются YAML-конфиги.

YAML-конфиг

YAML позволяет в человекочитаемом формате задать:

- Срез для эксперимента
- Параметры вносимых изменений
- Время начала и продолжительность
- Интересующие нас метрики

Работу по запуску эксперимента хороший фреймворк возьмет на себя.

```

1 |label:                                # snake_case | используется разработчиками в коде
2 title:                                 # в произвольной форме
3 jira_task:                            # ссылка на jira
4
5 team:                                 # например BuyerX
6
7 - platforms:                          # лишние удалить
8   - Avito.ru
9   - m.Avito.ru
10  - iOS
11  - Android
12
13 - features:
14 -   old_trash:
15     is_control: True                 # должна быть ровно одна контрольная фича
16     # description:
17     # weight:
18     new_beautiful_design:           # по умолчанию = 1
19
20 start_time:                          # например 2018-02-20 14:00:00
21 traffic_percent:                    # от 0.01 до 100
22
23 ab_days:                            # например 7

```

Зона ответственности аналитика

Что должен определить аналитик до начала эксперимента

Нужно определить и зафиксировать следующие параметры

- Гипотеза и метрики
- Ожидаемый размер эффекта
- Допустимые ошибки I и II рода
- Способ подбора групп
- Набор методов для повышения чувствительности
- Размер групп
- Алгоритм принятия решений по результатам пилота
- Техническая реализация

На основе этих данных пишется YAML-конфиг или производится ручной запуск эксперимента.

Перед запуском эксперимента

Выбор параметров позволяющих принять бизнес-решение

Необходимо обсудить какой минимальный размер эффекта позволит принять решение о внедрении изменений и выбрать допустимые уровни ошибок I и II рода. Это ключевые параметры, от которых отталкиваемся при дальнейшей разработке дизайна эксперимента. Нужно зафиксировать алгоритм принятия решения.

Размер и продолжительность эксперимента

Эксперимент — затратное мероприятие. Мы хотим провести его быстро, но не потерять статзначимость.

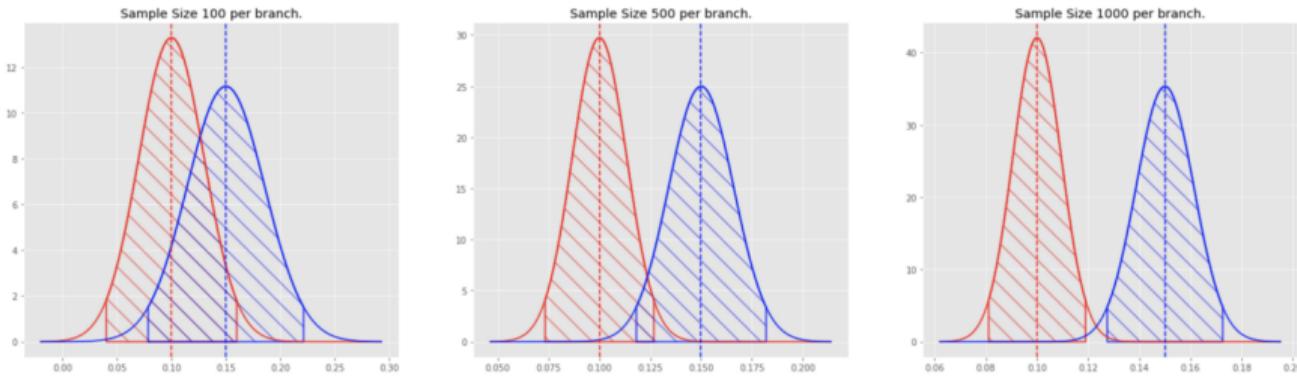
Для этого необходимо использовать методы повышения чувствительности. На основе выбранных параметров нужно определить необходимый размер и продолжительность эксперимента.

Техническая реализация

Нужно проконтролировать, что технически всё готово к проведению эксперимента. Есть нужная функциональность, пишутся логи и нет никаких ошибок.

Эта работа либо выполняется самостоятельно, либо контролируется её выполнение командой, отвечающей за техническую подготовку эксперимента.

Методы повышения чувствительности



О вреде дисперсии

$$n > \frac{[\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)]^2 (\sigma_X^2 + \sigma_Y^2)}{\varepsilon^2}$$

Дисперсия в данных снижает чувствительность. Чтобы получить статистически значимый результат нам нужно брать больше данных.

Методы повышения чувствительности

Снижение дисперсии позволяет уменьшить размер теста. Снижение достигается за счет:

- Удаления аномальных данных
- Стратификации
- CUPED
- Применения ML моделей

Поиск и удаление аномалий

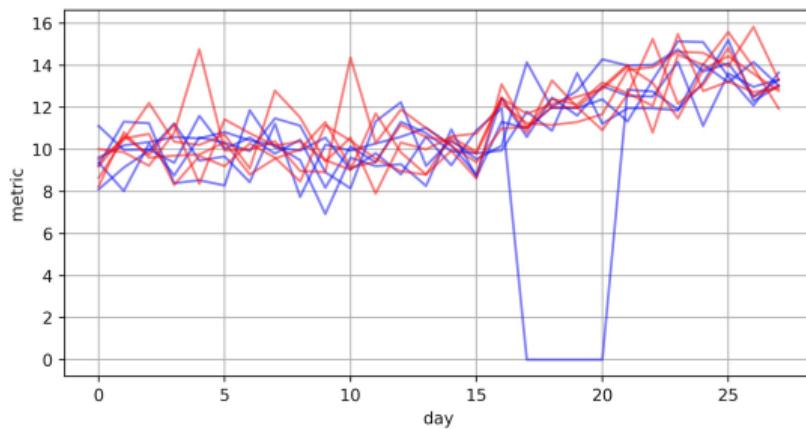
Выбросы в данных увеличивают дисперсию и снижают чувствительность теста.

Чем точнее будем находить выбросы, тем чувствительнее будет тест.

Причины появления выбросов

- Закрытие магазина
- Большая закупка (свадьба, поход)
- WEB crawler
- Сбой в системе сбора данных

Важно при удалении аномалий понимать по какой причине эта аномалия произошла. Иначе вместо аномалий можно отфильтровать эффект.

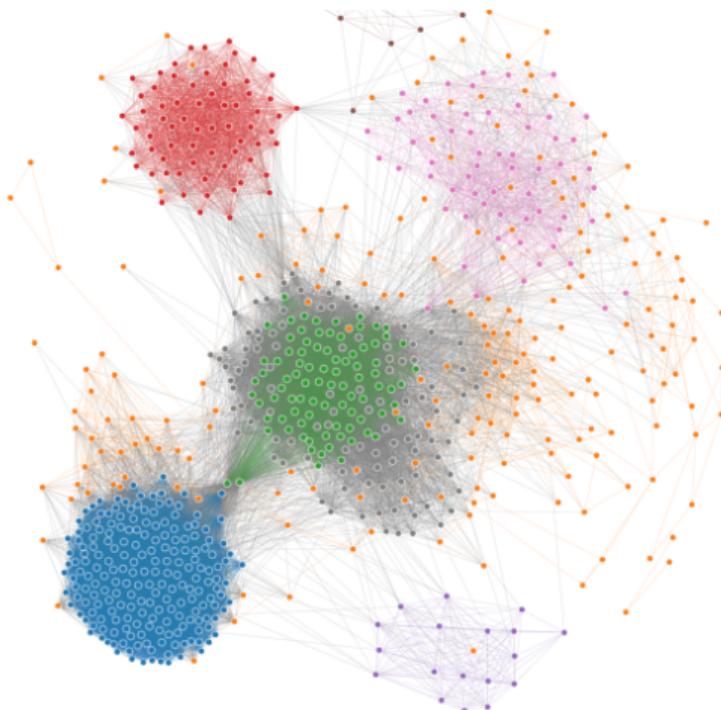


Кластеризация и стратификация

Дисперсия стратифицированного сэмплирования меньше дисперсии случайного сэмплирования на

$$\frac{1}{n} \sum_{k=1}^K w_k (\mu_k - \mu)^2$$

Если построить кластеризацию пользователей по каким-то признакам (возраст, регион проживания и т.п.) такую, что значения метрики в каждом кластере сильно различаются, то это помогает снизить дисперсию для перекошенных выборок.



CUPED

CUPED

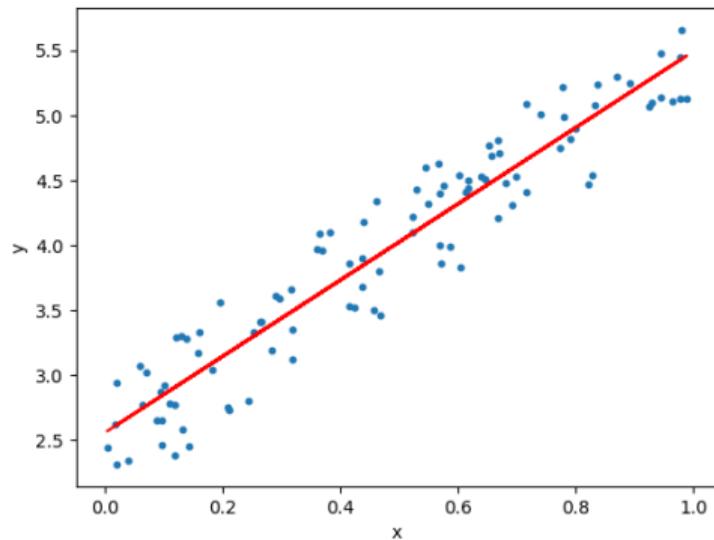
Использование исторических данных позволяет учесть и удалить факторы изменчивости, оставшиеся за рамками нашего внимания.

$$Z = Y - \theta(X - \mathbb{E}X)$$

Мы оставляем только основной эффект, что снижает дисперсию.

Прогнозирование ковариаты

Вместо исторических данных можно брать прогноз метрики на текущий период. Чем точнее прогноз X , тем чувствительнее будет тест.



Auto ML

Применение ML

Машинное обучение может помочь в кластеризации пользователей, прогнозировании значений или обработке пропусков. Всё это так же приводит к снижению дисперсии.

Auto ML

На основе предложенных техник можно собрать систему автоматического подбора оптимальной модели. Определять какие методы и в каком объеме необходимо использовать для каждого из экспериментов.

Нужно соблюдать баланс. Чем сложнее модели, тем точнее предсказание, но тем больше шансов переобучиться.



Sample Size

Расчет необходимого объема данных

На основе значения минимального размера эффекта ε , который мы хотим детектировать, и установленных размеров ошибок I и II рода (α и β) мы можем рассчитать объем необходимых данных

$$n > \frac{[\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)]^2 (\sigma_X^2 + \sigma_Y^2)}{\varepsilon^2}.$$

При этом дисперсии σ_X^2 , σ_Y^2 мы имеем возможность оценить только по историческим данным. С учетом всех наших алгоритмов повышения чувствительности.

Оставляем небольшой запас

Надо помнить, что на реальных данных во время эксперимента мы можем получить дисперсию больше. Размер теста окажется заниженным.

Чтоб побороть эту проблему, можно выбрать размер теста несколько больше. Например, увеличить его на 20%.

Продолжительность эксперимента

Какими параметрами мы управляем?

Для сбора данных мы можем управлять двумя параметрами:

- Доля пользователей, попадающих в эксперимент
- Продолжительность эксперимента

Совместно они определят объем поступающих к нам данных.

Как набрать необходимый объем данных?

Эксперимент на группах в 1% продолжительностью 5 недель даст такой же объем данных, что эксперимент на 5% группах продолжительностью в 1 неделю (при некоторых условиях регулярности).

Но важно помнить, что два таких эксперимента не будут эквивалентны. У каждого подхода есть свои плюсы и минусы.

Сезонность

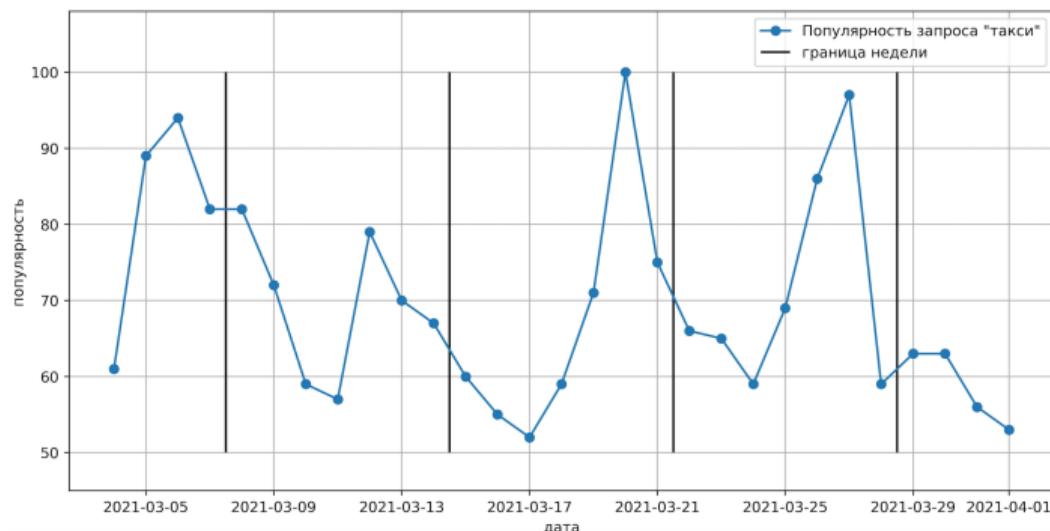
Поведение пользователей может иметь сезонность. Изменения эксперимента могут влиять на разные периоды сезона по-разному.

Замечание

Лучше включать целое число сезонов в период проведения пилота. Нужно следить, чтобы в период проведения пилота не попадали нетипичные дни.

Чтобы исключить недельную сезонность можно выбрать продолжительность 7 или 14 дней.

Рис.: Популярность запроса "такси" в Google Trends



Техническая реализация

Подготовка эксперимента

- реализация функциональности самого изменения
- параметризация конфигураций
- алгоритм распределения пользователей по группам
- расчёт метрик
- возможность экстренного прекращения эксперимента

Логирование

- Сбор логов с клиента и сервера
- Агрегация логов
- Запись в хранилище

Мониторинг

- Общее состояние системы
- Метрики пилота
- Автоматическая детекция аномалий

Техническая реализация

Подготовка эксперимента

- реализация функциональности самого изменения
- параметризация конфигураций
- алгоритм распределения пользователей по группам
- расчёт метрик
- возможность экстренного прекращения эксперимента

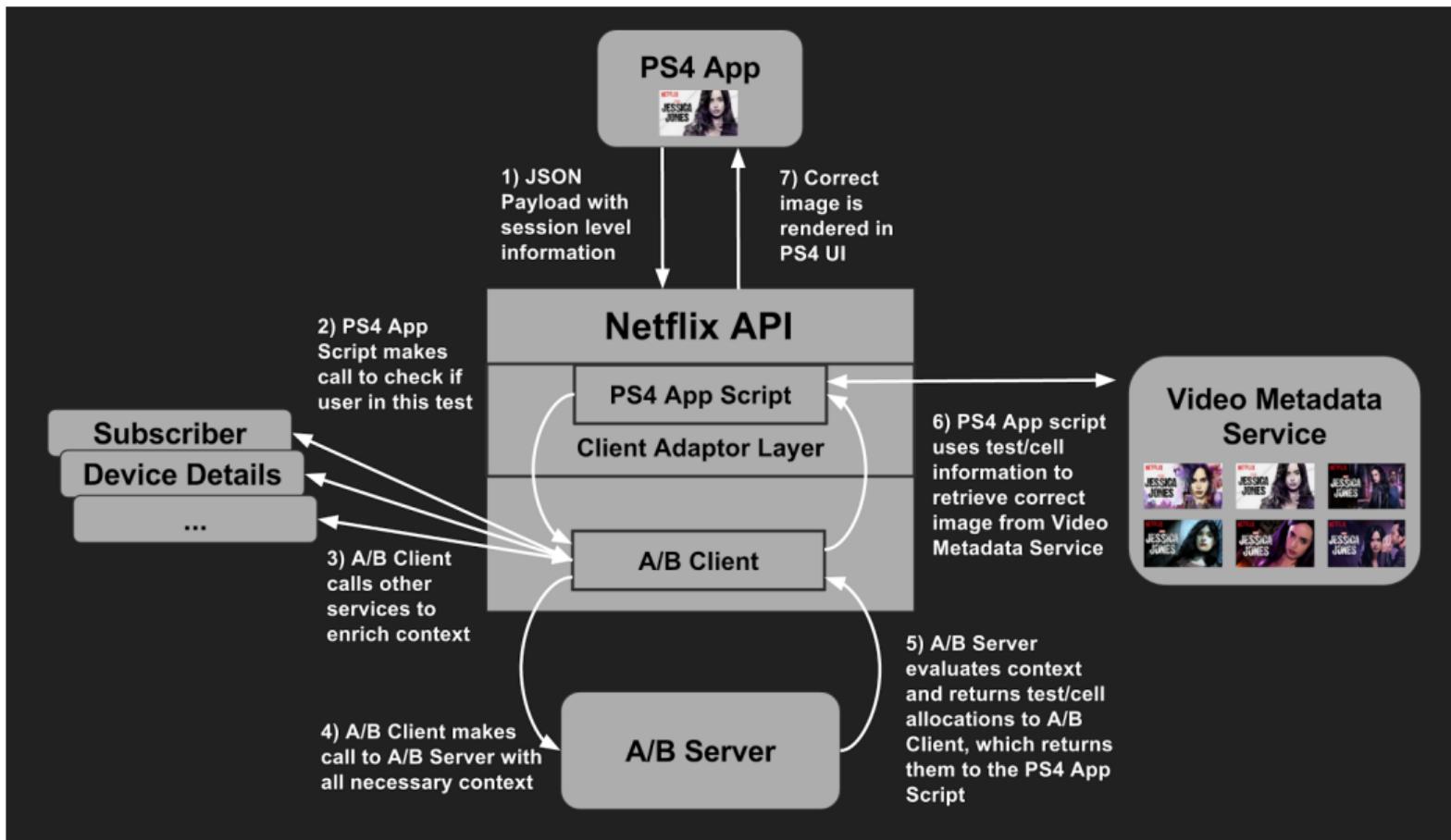
Логирование

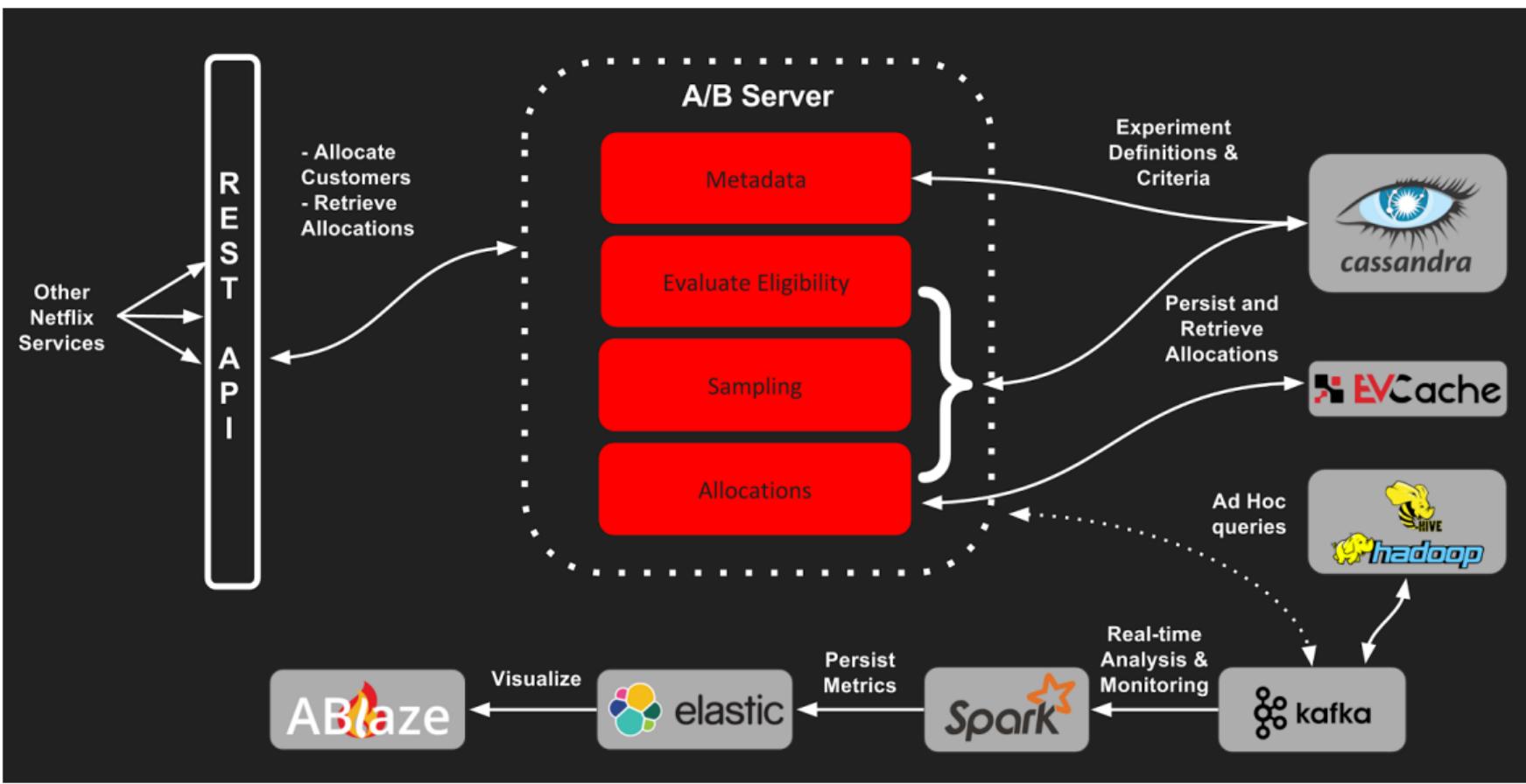
- Сбор логов с клиента и сервера
- Агрегация логов
- Запись в хранилище

Мониторинг

- Общее состояние системы
- Метрики пилота
- Автоматическая детекция аномалий

Тестирование!





Date Range

02/12/2016 - 03/24/2016

Analyst

e.g. flast

Category

Impact Areas

Algorithm, ARO, Billboard, ...

Platforms

Properties

Regions

Legend ⓘ

Today

Real-Time Placeholder

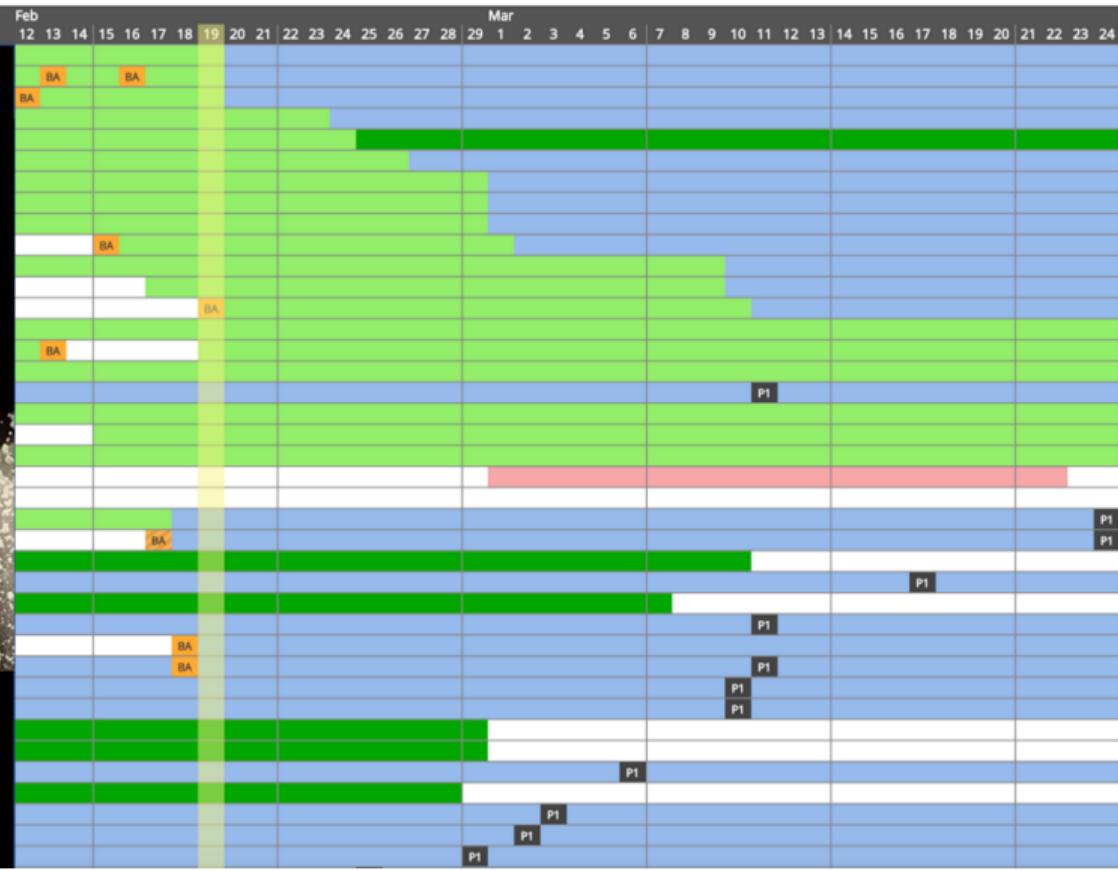
Real-Time Alloc Running

Post-Alloc Analysis

Analyzing

Period Analysis

Batch Allocate



Мониторинги во время проведения эксперимента

Зачем нужны мониторинги?

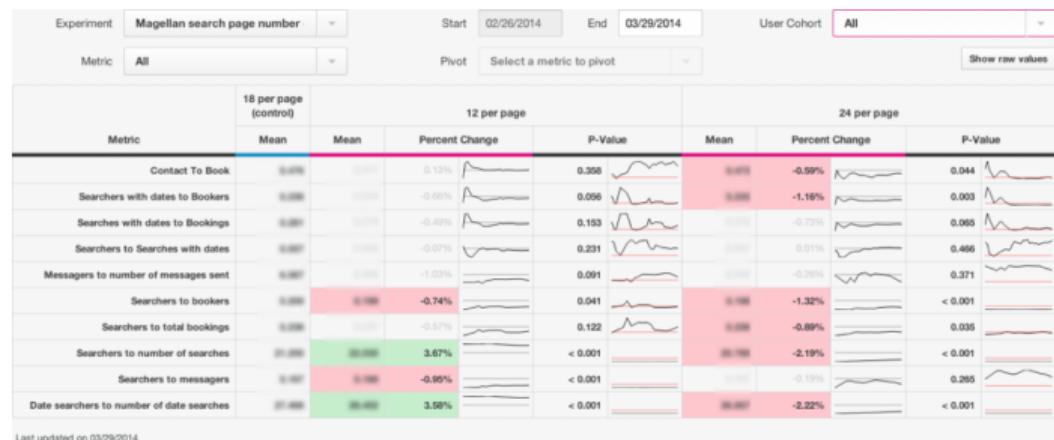
Много экспериментов — много потенциальных проблем. Как минимум, нужны мониторинги по срыву счетчиков. Чтобы не выяснилось через 2 месяца, что были проблемы в логировании.

Для расчета многих метрик требуется значительное время дни или даже недели. Для онлайн-мониторинга такие метрики не подойдут.

Онлайн-метрики

- Нужны быстрые мониторинги по сырьем данным
- Отслеживание ключевых показателей
- Метрики, сигнализирующие, что в эксперименте нет проблем

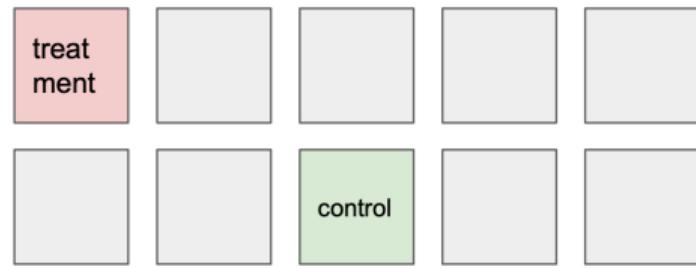
Остальное обработаем позже.



Провальные тесты

Запуск пилота

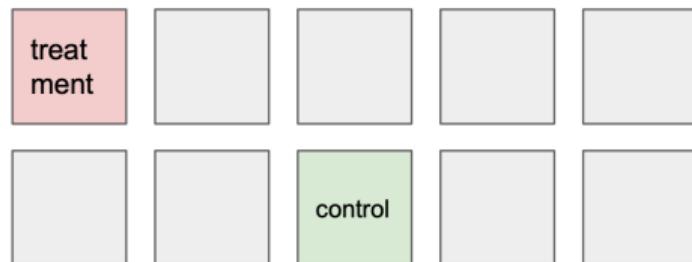
Решили провести эксперимент на 10% трафика продолжительностью 21 день.



Провальные тесты

Запуск пилота

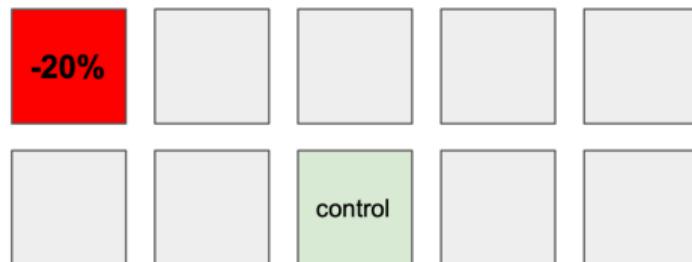
Решили провести эксперимент на 10% трафика продолжительностью 21 день.



На следующий день метрика упала на 20%.

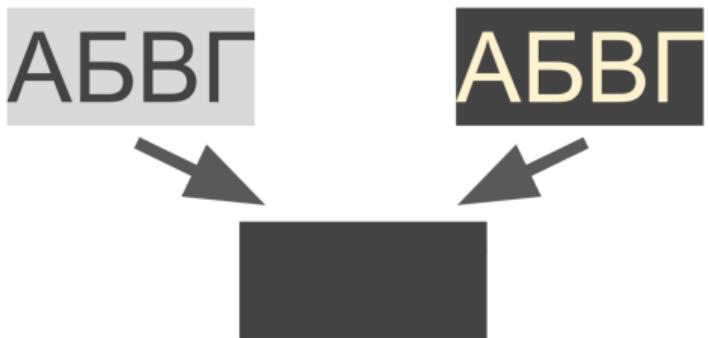
Наши потери составляют 2% за сутки!

AB тесты могут оказаться очень дорогими.



Возможные причины

- Несовместимость с другими экспериментами
- Ошибка в технической реализации
- Гипотеза оказалась неверной

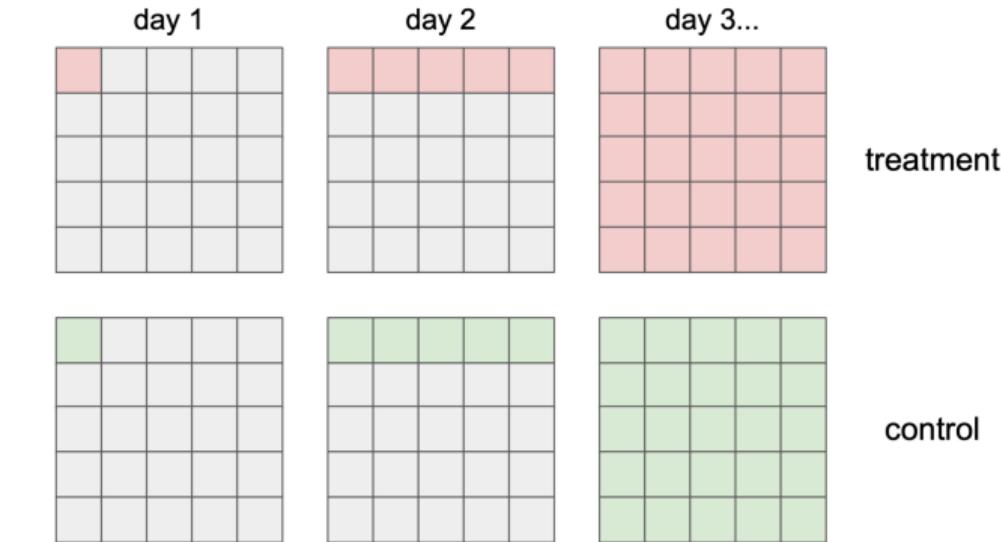


```
35 | group = hash(gender) % 2
36 | return group
37 |
```

Canary deployment

Постепенное внедрение

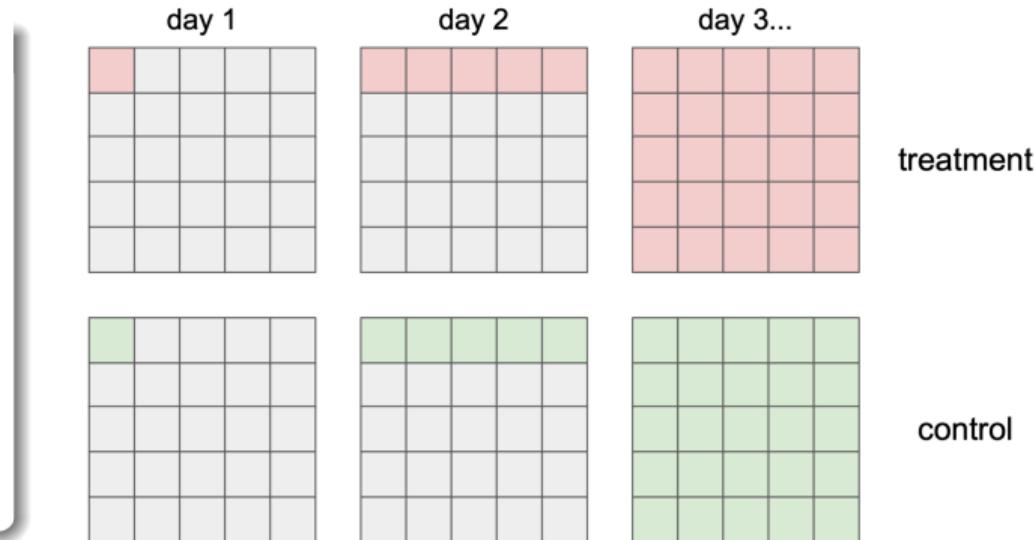
- В первый день раскатываемся на 1% пользователей;
- Если всё в порядке, на следующий день раскатываемся на 10%;
- Если всё нормально, раскатываемся на всех пользователей.



Canary deployment

Постепенное внедрение

- В первый день раскатываемся на 1% пользователей;
- Если всё в порядке, на следующий день раскатываемся на 10%;
- Если всё нормально, раскатываемся на всех пользователей.



Пример

Если в первый день увидели огромное падение метрик, то воздействовали негативно лишь на 1% pilotной группы. То есть наши **потери в 100 раз меньше**.

Парадокс Симсона

Пользователей не хватает

- Раскатываем постепенно
- Используем "многоруких бандитов"

		день 1	день 2	всего
A	Пользователи	1000	1000	2000
	Конверсии	400	100	500
	%	40%	10%	25%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

Парадокс Симсона

Пользователей не хватает

- Раскатываем постепенно
- Используем "многоруких бандитов"

		день 1	день 2	всего
A	Пользователи	1000	1000	2000
	Конверсии	400	100	500
	%	40%	10%	25%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

		день 1	день 2	всего
A	Пользователи	100	1000	1100
	Конверсии	40	100	140
	%	40%	10%	12.7%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

Парадокс Симсона

Пользователей не хватает

- Раскатываем постепенно
- Используем "многоруких бандитов"

		день 1	день 2	всего
A	Пользователи	1000	1000	2000
	Конверсии	400	100	500
	%	40%	10%	25%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

		день 1	день 2	всего
A	Пользователи	100	1000	1100
	Конверсии	40	100	140
	%	40%	10%	12.7%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

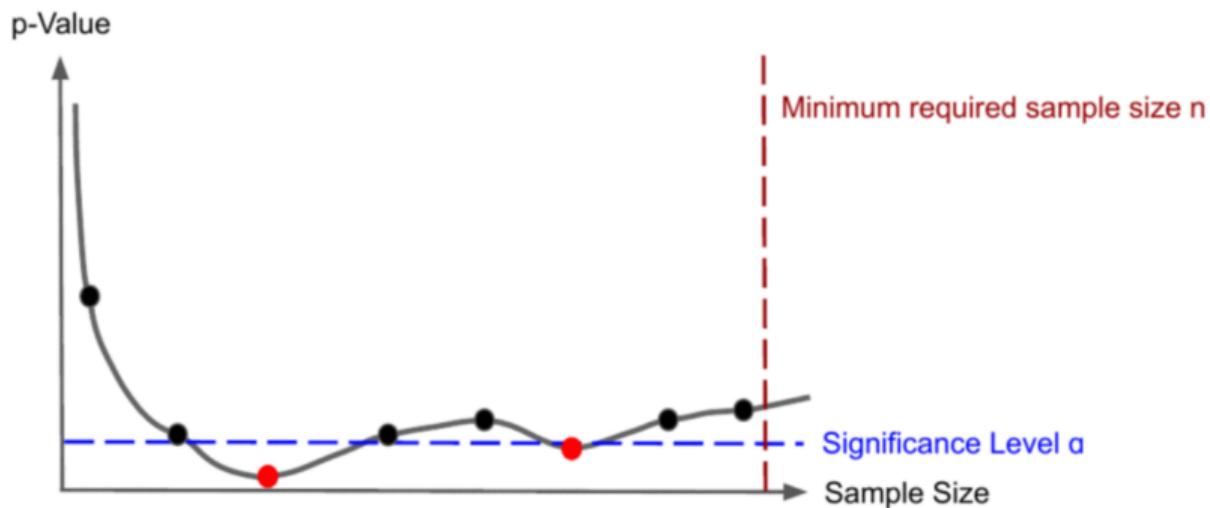
- Нужно внимательно работать со случаем, когда размеры групп меняются.
- Контрольная и экспериментальная группы должны быть одного размера!

Проблема ранней остановки

Условия ранней остановки

Мы можем следить за ходом эксперимента, но нельзя принимать решения до его завершения.

Все условия ранней остановки должны быть учтены в дизайне эксперимента. Если мы будем использовать что-то иное, то сломаем эксперимент.



После пилота

- Сбор и агрегация данных
- Построение точечной оценки и оценка значимости с помощью методов, зафиксированных до начала пилота:
 - Статистический критерий
 - Стратификация
 - CUPED
 - Линеаризация
 - Множественное тестирование
 - Последовательное тестирование
 - ...
- Анализ полученных данных
- Принятие решения о внедрении пилотируемых изменений
- Сбор инсайтов и генерация новых гипотез
- Сохранение всех результатов

Анализ результатов

Анализ результатов

- Графики метрик эксперимента. Эффект новизны, выбросы и другие аномалии.
- Инсайты в разных срезах.
- Статистическая значимость.

Принятие решения

- Критерии определённые до эксперимента.
- Результаты анализа.
- Продуктовые соображения.

AAB тест

Алгоритм

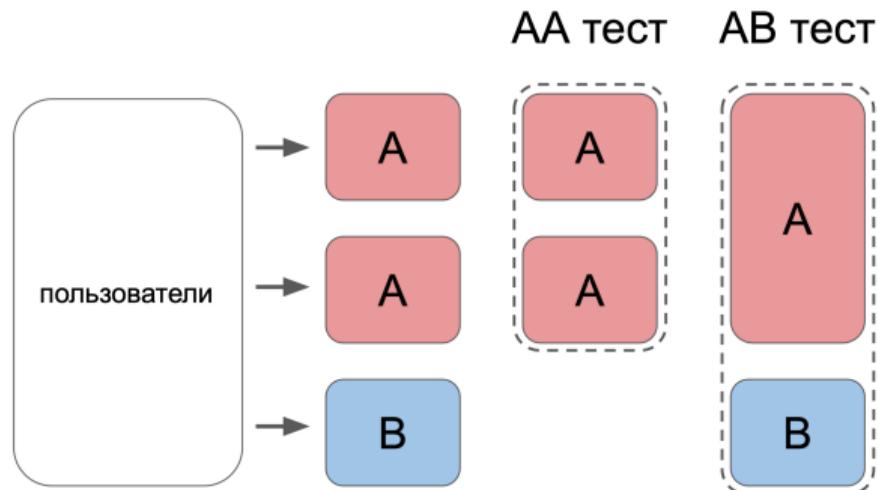
- Выделяем группы A/A/B.
- Проводим эксперимент.
- Если AA тест показал значимые отличия, то отклоняем тест.
- Иначе проверяем наличие значимого эффекта на AB тесте.

Преимущества

Выше точность теста.

Недостатки

Нужно больше данных.



Резюме: полный пайплайн проведения АВ теста

1. До запуска эксперимента

- Сформулировать гипотезу для тестирования
- Определить минимальный эффект, уровень значимости и мощность теста
- Выбрать методы повышения чувствительности
- Рассчитать необходимый объем данных, размер и продолжительность теста
- Провести техническую подготовку эксперимента

Резюме: полный пайплайн проведения АВ теста

1. До запуска эксперимента

- Сформулировать гипотезу для тестирования
- Определить минимальный эффект, уровень значимости и мощность теста
- Выбрать методы повышения чувствительности
- Рассчитать необходимый объем данных, размер и продолжительность теста
- Провести техническую подготовку эксперимента

2. Во время проведения эксперимента

- Следить за мониторингами корректности проведения эксперимента
- Если дизайн эксперимента предусматривает возможность ранней остановки, то отслеживать наступление таких условий
- Больше ничего не трогать

Резюме: полный пайплайн проведения АВ теста

1. До запуска эксперимента

- Сформулировать гипотезу для тестирования
- Определить минимальный эффект, уровень значимости и мощность теста
- Выбрать методы повышения чувствительности
- Рассчитать необходимый объем данных, размер и продолжительность теста
- Провести техническую подготовку эксперимента

2. Во время проведения эксперимента

- Следить за мониторингами корректности проведения эксперимента
- Если дизайн эксперимента предусматривает возможность ранней остановки, то отслеживать наступление таких условий
- Больше ничего не трогать

3. После завершения эксперимента

- Обработка полученных результатов
- Оценка эксперимента в строгом соответствии с ранее утвержденным дизайном
- Внедрение изменений или отказ от них

Дополнительные материалы

Ссылки для самостоятельного изучения

1. It's All A/Bout Testing: The Netflix Experimentation Platform
2. Airbnb. Experiment Reporting Framework
3. Как устроено А/В-тестирование в Авито
4. Как у нас устроено А/Б-тестирование. Лекция Яндекса
5. Scribd's A/B Test Framework
6. Step by Step Process for Planning an A/B Test