e^xperiment $f$est

Анализ A/B-тестирований

Опыт

Мы консультируем продуктовые команды, а также ведем обучающие мероприятия по анализу онлайн и офлайн экспериментов

Ключевые клиенты

Что сегодня в программе?

Об экспериментах

День 1

Об экспериментах

ИДЕИ И ГИПОТЕЗЫ

Тезис 1

Все начинается с идеи

- Источники: анализ данных, анализ рынка, генерация
- Неподтвержденная идея является гипотезой

Тезис 2

Каждой идее – свой метод проверки

- Глубинные интервью
- Опросы
- Юзабилити-тестирования

Тезис 3

Проверке гипотезы необходимы условия

- Отсутствие влияющих факторов, кроме самого тестируемого изменения
- Репрезентативная оценка
- Точность

ЧТО ТАКОЕ АБ

e^xperiment $f$est

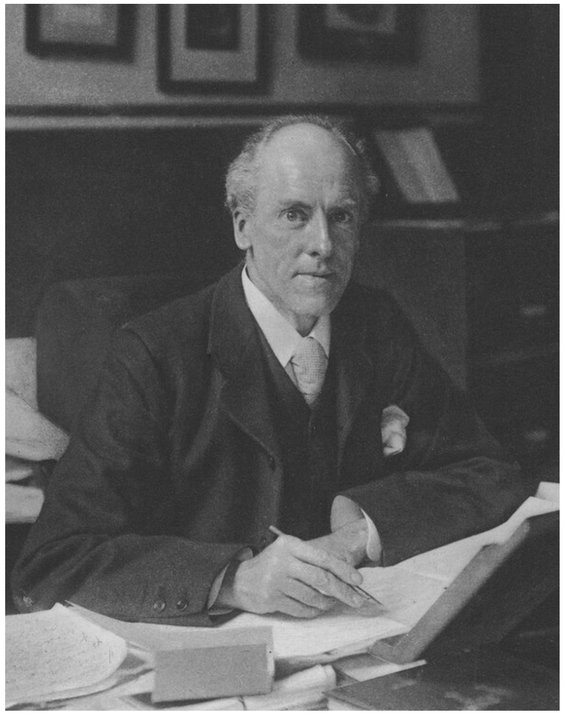
С чего все начиналось?

- **Биология, химия, медицина.** Статистика как инструмент использовалась для клинических исследований. Благодаря этим задачам математическая статистика развивалась.

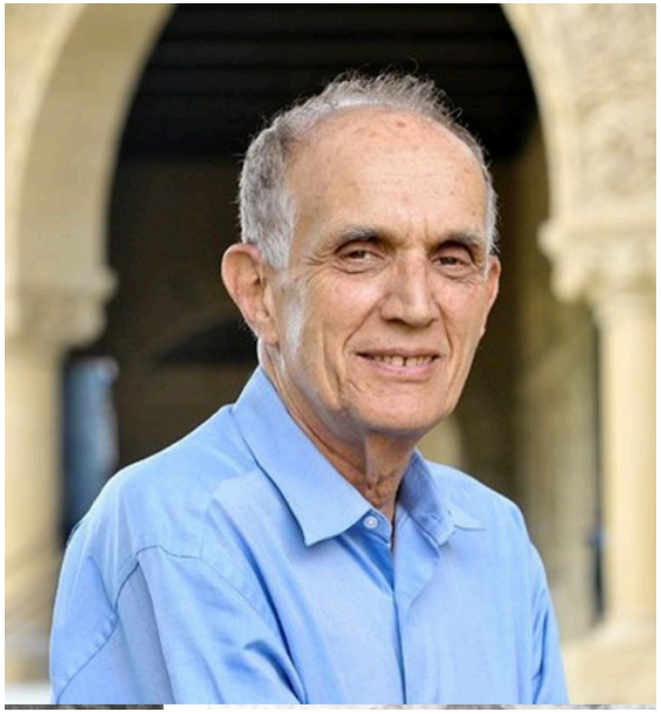
Фишер

Госет

Пирсон



Эфрон



A/B-тестирование – это где...

- Проверяется два и более варианта (контроль и тест) с целью определения наиболее эффективного
- Степень эффективности «измеряется» с помощью посчитанных вероятностей ложноположительных и ложноотрицательных случаев

Решения

Ложноотрицательные

Полезные изменения упускаются из виду

Ложноположительные

Публикуем бесполезные изменения, которые не работают

Ложноположительное решение

Гипотеза: повышение недельной цены подписки с 1$ до 2$

Итог:

На первых двух днях эксперимента был зафиксирован статистически значимый результат. Продакт и аналитик приняли решение принять результат как успешный. А вот после публикации изменения на всех пользователей –ключевой показатель изменился в худшую сторону.

Ложноотрицательное решение

Проводили эксперимент 2 дня, не видели разницы, остановили

Итог:

Т.к. в эксперименте было охвачено только 2 дня, мы не учитываем поведение аудитории в остальные дни недели.

Возможно изменение имеет отложенный эффект: пользователь в понедельник попал в тестовую группу, а в пятницу принял решение

Области применения A/B:

- Эксперименты в дизайне (UI / UX)
- Тестирование нового функционала в приложении и на сайте
- Операционные эксперименты
- Оптимизация back-end'а и алгоритмов (например, ранжирование)
- Эксперименты в ценообразовании

Метрики эксперимента: уровни

Целевые

Показатели, на которые направлено изменение

Опережающие

Показатели, хорошо коррелируемые с целевыми, дающие предикт и полезны тогда, когда нет времени ждать основную метрику

Guardrail

Показатели, на которые направленно влияет изменение, но не являющиеся целевыми. Рекомендуется за ними наблюдать и на их основе в том числе принимать решение (например, каннибализация)

Пример системы уровней

Пример: e-commerce, тест нового UI корзины

Целевые

- Конверсия в покупку, средний чек, ARPU, ARPPU

Опережающие

- Добавления товара в корзину на сессию, просмотры товаров на сессию, отток чекаута, ошибки на чекауте

Guardrail

- Время от входа в корзину до ее прохождения, доля поисковых запросов из корзины, взаимодействие с рекомендательными блоками в корзине

Пример системы уровней

Пример: образовательный продукт, тест нового образовательного контента

Целевые

- Продление обучения, Средний доход на платящего пользователя (ARPPU)

Опережающие

- Интенсивность обучения, кол-во ошибок в момент обучения, частота обращений в службу поддержки, технические характеристики качества видео

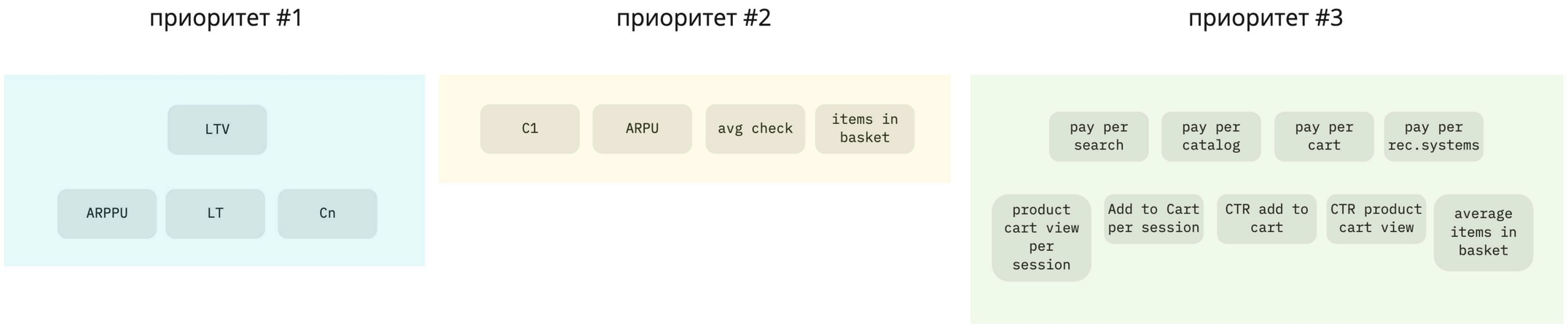
Guardrail

- Время проведенное за одним занятием, прерывание занятий, перемотка видео-контента

Метрики: типы

Доли	регистрации, удержание на 7 день [0,1,1,0,0,0,1,1,1,0,1,0,1]
Непрерывные	время в сек./мин./т.п., чек в рублях [1123.32, 324.4, 823.21, 924.91]
Отношения	поездок на водителя, кликов на сессии, цена за 1000 показов [10/123 = 0.081, 4129.2/12488 = 0.33, 1/100 = 0.01]

Приоритет метрик



miro

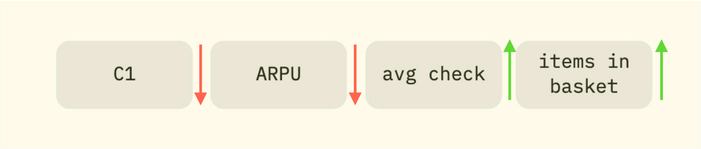
Приоритет метрик

Системе координат метрик – позволяет лучше и быстрее принимать решения, когда метрик много. Если нет понимания приоритета – принять решение достаточно сложно.

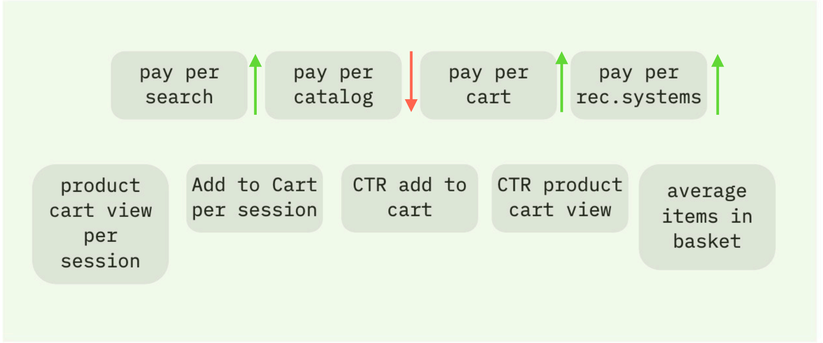
Приоритет метрик

приоритет #1

приоритет #2



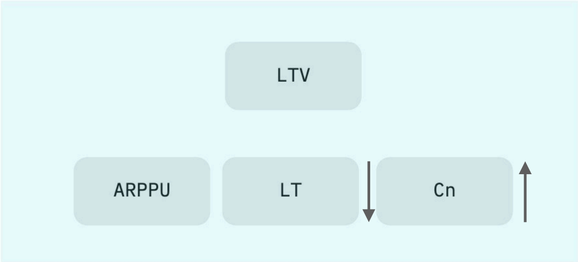
приоритет #3



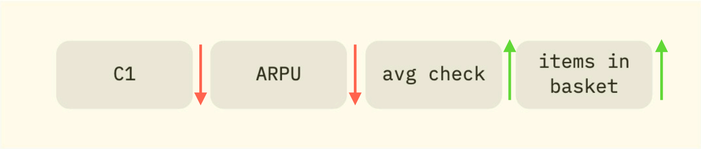
miro

Приоритет метрик

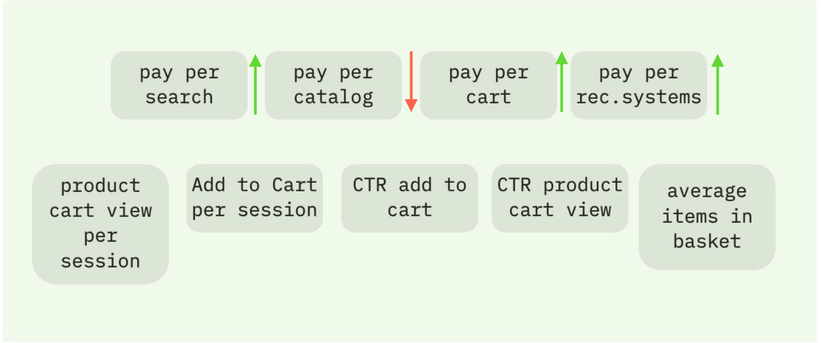
приоритет #1



приоритет #2



приоритет #3



miro

Приоритет метрик

$ важнее **<3**

<3 важнее **$**

На разных стадиях развития продукта система метрик может значительно меняться – это динамичный артефакт, который должен постоянно версионировать.

Приоритет метрик

долгие метрики VS *быстрые метрики*

У каждой метрики есть «окно» – пользователь не сразу принимает решение о том, чтобы совершить желанное действие для продукта. Важно искать «быстрые» метрики, которые зависимы к «долгим» и на основе этого менять приоритет и всю иерархию.

Типы экспериментов

Классические: A/B

A/A

A/B/C/...

TDI (team draft interleaving)

Diff-in-Diff

Synthetic control

A/B

Чем полезен?

Измерить эффект от изменения

Ключевые особенности

- Каждая группа эксперимента видит свой вариант
- Группы независимы
- Группы взяты из одной ГС
- Распределение может быть неравномерным

A/A

Чем полезен?

- Проверить сплит-систему
- Выбрать гомогенные группы

Ключевые особенности

- Группы независимы
- Группы взяты из одной ГС
- Часто используется для симуляций

A/B/C/...

Чем полезен?

Тот же A/B, только проверяется от 2
и более изменений

Ключевые особенности

- Группы независимы
- Группы взяты из одной ГС
- Сопряжена с проблемой множественной проверки гипотез

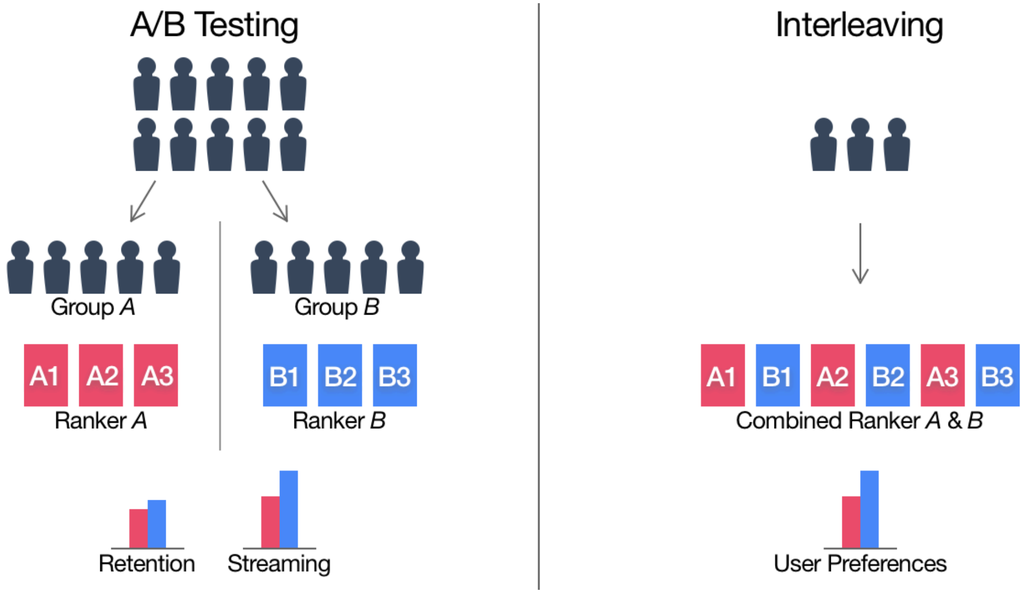
TDI (team draft interleaving)

Чем полезен?

Изменение в ранжированных списках

Ключевые особенности

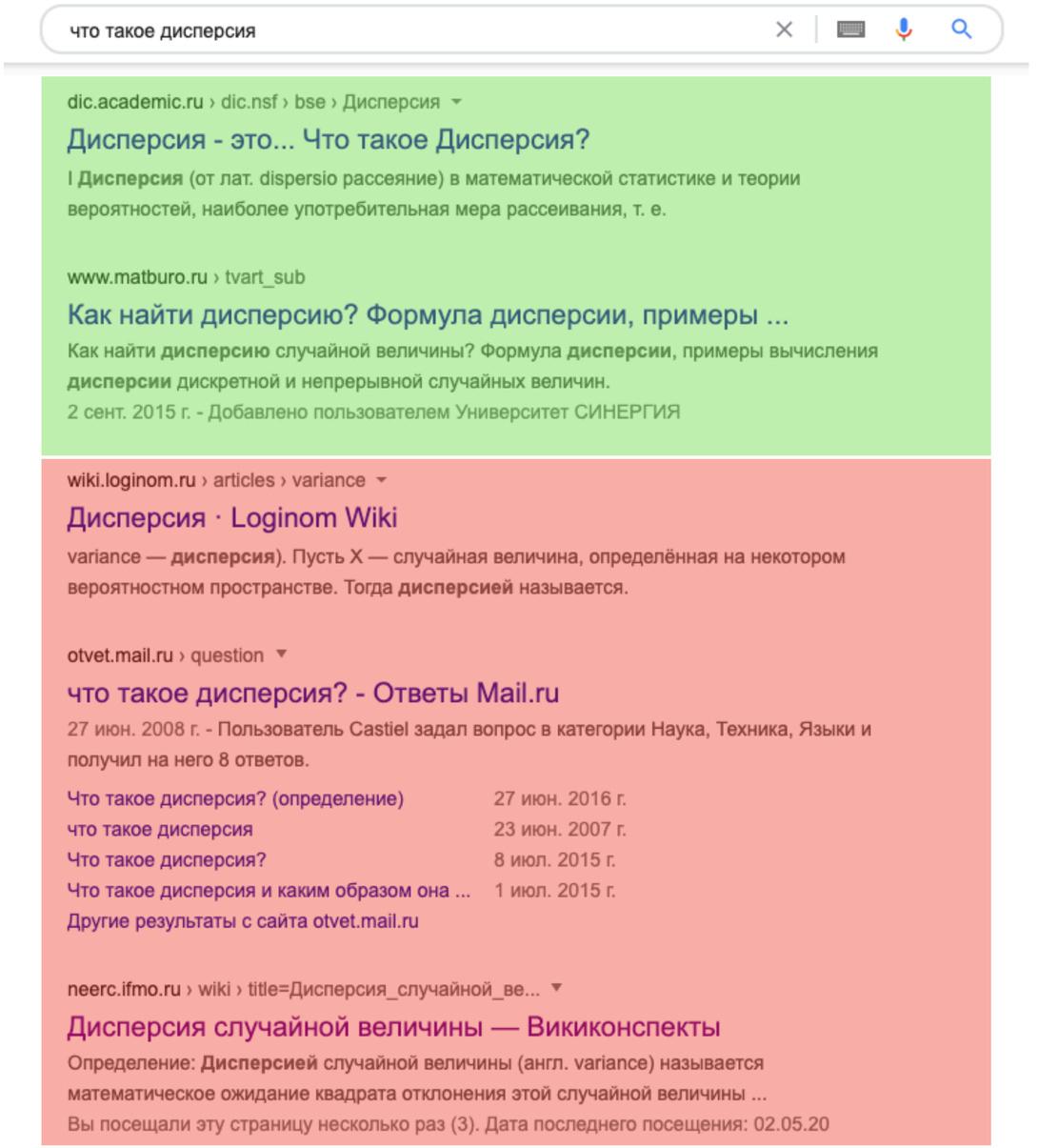
- Один пользователь видит сразу несколько вариантов
- Чаще всего используется в поиске и рекомендациях
- Выборки зависимы - что накладывает особенности



TDI

Задача: протестировать два поисковых алгоритма

Метод TDI позволяет показать пользователю результат выдачи пользовательского запроса с применением двух алгоритмов. Пользователь решает кликом, какой же алгоритм дал релевантный ответ. Разделения на варианты здесь нет



Diff-in-Diff

Чем полезен?

Когда нет возможности поделить пользователей
на группы в один момент времени

Ключевые особенности

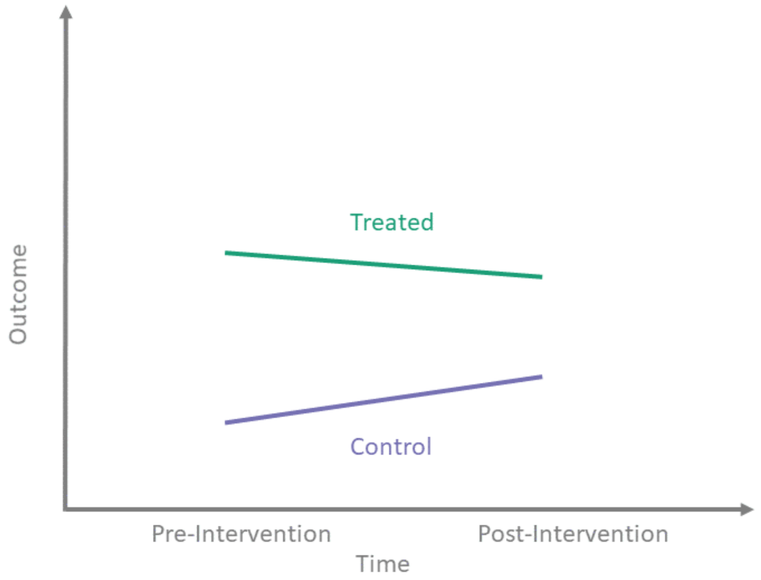
- Группы зависимы и разнесены во времени
- Один из типов регрессий

Diff-in-Diff

Экономическое изменение в государстве

Одна из групп подвержена воздействию, или участвует в некоторой программе, во втором периоде, но не в первом. Вторая группа не подвержена воздействию ни в одном из периодов.

Метод устраняет смещение при сравнении исходов в опытной и контрольной группах только во втором периоде, которое может быть следствием постоянных различий между этими группами



Synthetic control

Чем полезен?

Не чувствителен к социальным (сетевым) эффектам

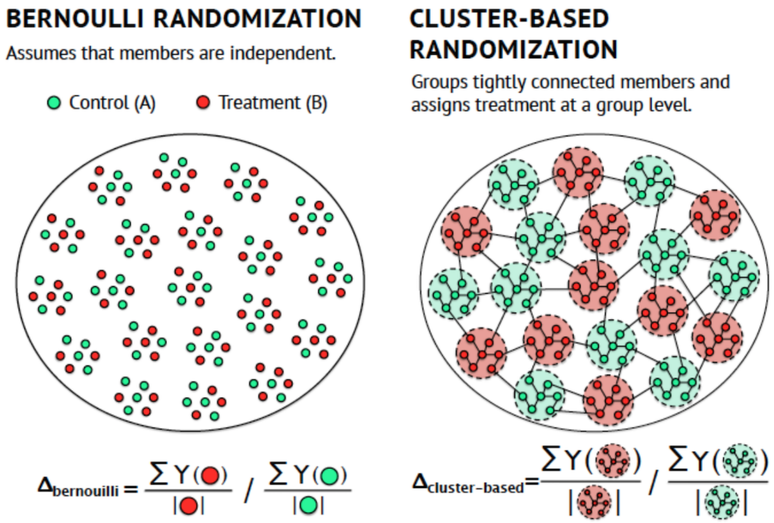
Ключевые особенности

- Группы отделены друг от друга географически или физически
- Группы схожим по описательным статистикам, но находятся далеко друг от друга
- Контроль регулярно версионирует

Synthetic control

Мы хотим провести эксперимент в социальной сети – дать возможность пользователям отправить анимированные смайлики в сообщениях.

В подобных продуктах есть большая особенность – пользователи общаются между собой. И общение пользователей из групп А и Б могут оказывать сильное влияние на исход всего эксперимента, т.к. в процессе кто-то



Самое важное в Synthetic control – это корректно выбрать группы для его формирования.

Важно, чтобы группы до получения «влияния» не отличались друг от друга по описательным статистикам и были репрезентативны друг другу.
Формирование групп не является разовым процессом, скорее перманентным поиском близких выборок для проведения эксперимента

Варианты для контроля №2

Варианты для контроля №1

Как устроено АБ-тестирование в продуктовых командах

- Каждый большой продукт имеет свои особенности как с точки зрения бизнеса так и с точки зрения метрик. Эти особенности накладывают определенные ограничения и дают творческий простор для развития методологии экспериментов

Uber

Коммуникация водителей между собой

- Водители между собой общаются
- Общие водителей создает сетевой эффект, который влияет на исход эксперимента
- Отслеживать факт общения по событиям крайне проблематично. Нет триггера, что коммуникация началась

Uber

Вечный контроль (моделируемый)

- Избавление от сетевых эффектов по средствам разделения пользователей географически
- Используется на экспериментах для водительских мотиваций
- Вынужденная особенность продукта

Варианты для контроля №2

NETFLIX

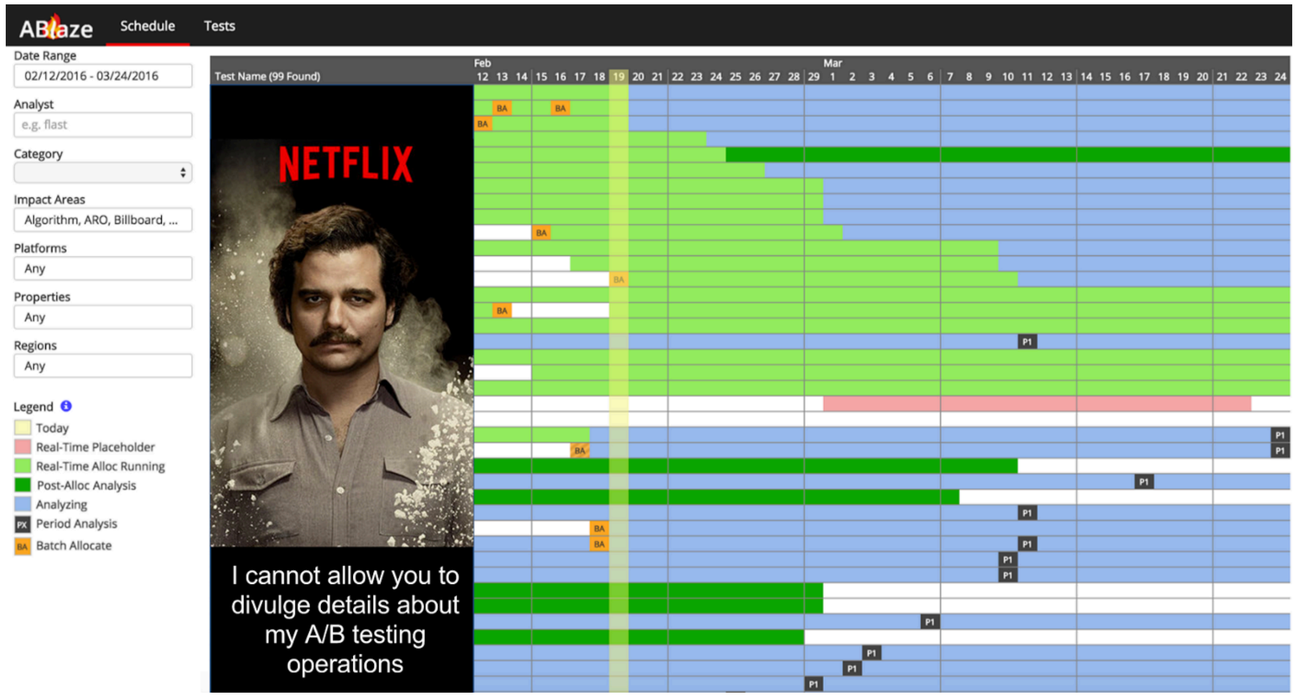
Увеличение кол-во экспериментов на ограниченном трафике

- Культура компании обязывает все делать через A/B
- Трафик ограничен – команд много
- Ограничения на a/b создают проблемы для развития продукта

NETFLIX

Платформа экспериментов, внутренние инструменты и R&D

- Платформа – инструмент автоматизации a/b и развития R&D
- Внутренние инструменты ускоряют процесс анализа сложных экспериментов
- R&D позволяет находить новые методы ускорения a/b тестов



Эксперимент может запустить каждый

- Эксперимент может запустить любой член команды на платформе
- Не каждая гипотеза чем то подтверждена
- Регулярно тестируются «гипотезы-пустышки»

Жесткая система модерации

- Каждый эксперимент модерируется
- Чтобы запустить эксперимент на платформе – нужно доказать его ценность и приложить основания для гипотезы
- Запустить эксперимент – сложно

- Эксперименты – не просто разделение трафика или базовая статистика, это процесс с полноценным R&D
- Эксперименты – это методология управления гипотезами и результатами
- Чем больше продукт, тем больше людей выделяется под задачи экспериментов

e^xperiment $f$est

Мирмахмадов Искандер
Черемисинов Виталий