



> Конспект > 7 урок > Множественное тестирование, $2 + 2 = -2$

> Оглавление

> Оглавление

> Множественная проверка гипотез

- > Отдельная проверка единичных гипотез не подходит
- > Связанные и контрольные метрики
- > Пересекающиеся эксперименты

> Независимые гипотезы

- > Много гипотез
- > Обозначения
- > Групповая гипотеза
 - > Нулевая гипотеза при множественном тестировании
 - > Когда можно применять групповую гипотезу?
- > FWER
- > Метод Бонферрони
- > Недостаток метода Бонферрони
 - > Отсутствие взаимосвязи между тестами отдельных гипотез
 - > Расплата мощностью критерия
- > Распределение p-value
 - > Когда выбираем уровни значимости?
 - > Равномерное распределение
- > Метод Холма
- > Метод Бенджамини-Хохберга
 - > Описание метода Бенджамини-Хохберга
- > Сравнение методов

- > Большие выбросы
- > Один эксперимент и много метрик
- > **Зависимые гипотезы**
 - > Один эксперимент и много метрик
- > **Параллельный запуск экспериментов**
 - > Много гипотез и мало времени
 - > Параллельный запуск
 - > Одномерная и многомерная схемы запуска экспериментов
 - > Схемы параллельного запуска экспериментов
 - > Применение многомерной схемы
 - > Одномерная схема
 - > Строго изолированные эффекты
 - > Многомерная схема. Коррелированные группы
 - > Одна и та же группа в двух экспериментах
 - > Пилот и контроль в двух группах перевёрнуты
 - > Некоррелированные группы
 - > Разбиение на основе хэша
 - > Слоты
 - > Распределение слотов по экспериментам
 - > Память пользователей. Двойное хэширование
 - > Двойное хэширование
 - > Выделение слотов в многомерной схеме
 - > Тройной "посол"
 - > Три этапа хэширования
 - > Конфигурация экспериментов
 - > Требования к экспериментальным группам
 - > Эвристика размещения экспериментов
 - > Размещение в свободные слоты
 - > Максимально плотное размещение
 - > **Доля успешных экспериментов**
 - > **Дополнительные материалы**

> Множественная проверка гипотез

- Много гипотез
 - Много метрик
 - Мало пользователей
-

> Отдельная проверка единичных гипотез не подходит

Установленный уровень значимости α означает, что с вероятностью $P = \alpha$ мы будем находить эффект там, где его нет.

Если мы выбираем $\alpha = 0.05$, то в одном из 20 случаев мы получим ложноположительный результат. Если тестировать 20 не приносящих эффекта изменений, то скорее всего в каких-то из них мы "обнаружим" эффект.

Избежать тестирования множества гипотез иногда мы не можем. Например, если речь идёт о каких-то изменениях дизайна. Нам необходимо проверить сочетания в разных комбинациях. При этом крайне желательно избежать фантомных эффектов.

> Связанные и контрольные метрики

В реальных экспериментах всегда **рассматривается множество метрик**:

- Основная метрика;
- Контрольные метрики.

Изменения в этих метриках **не будут независимыми**, т.к. вызваны одними и теми же причинами — нашим экспериментом.

Проблема связанных метрик:

- Необходимо научиться оценивать результаты зависимых метрик;
 - Нужно понять, как принимать решение в том случае, если часть метрик говорят о наличии эффекта, а часть — о его отсутствии.
-

> Пересекающиеся эксперименты

Идеальный эксперимент — это эксперимент **изолированный**. Мы должны минимизировать воздействие внешних факторов.

На практике экспериментов очень много, они образуют длинные очереди. Каждый требует большого объёма пользователей. Приходится либо ждать, либо проводить эксперименты одновременно.

Один пользователь может попасть сразу в несколько экспериментов.

Как можно и как нельзя пересекать эксперименты:

- Пользователь может попасть одновременно в несколько экспериментов только при условии, что они проверяют разную функциональность: например, дизайн главной страницы и алгоритм поисковой выдачи.
 - Разбиения на A/B-группы для разных экспериментов должны производиться независимо.
-

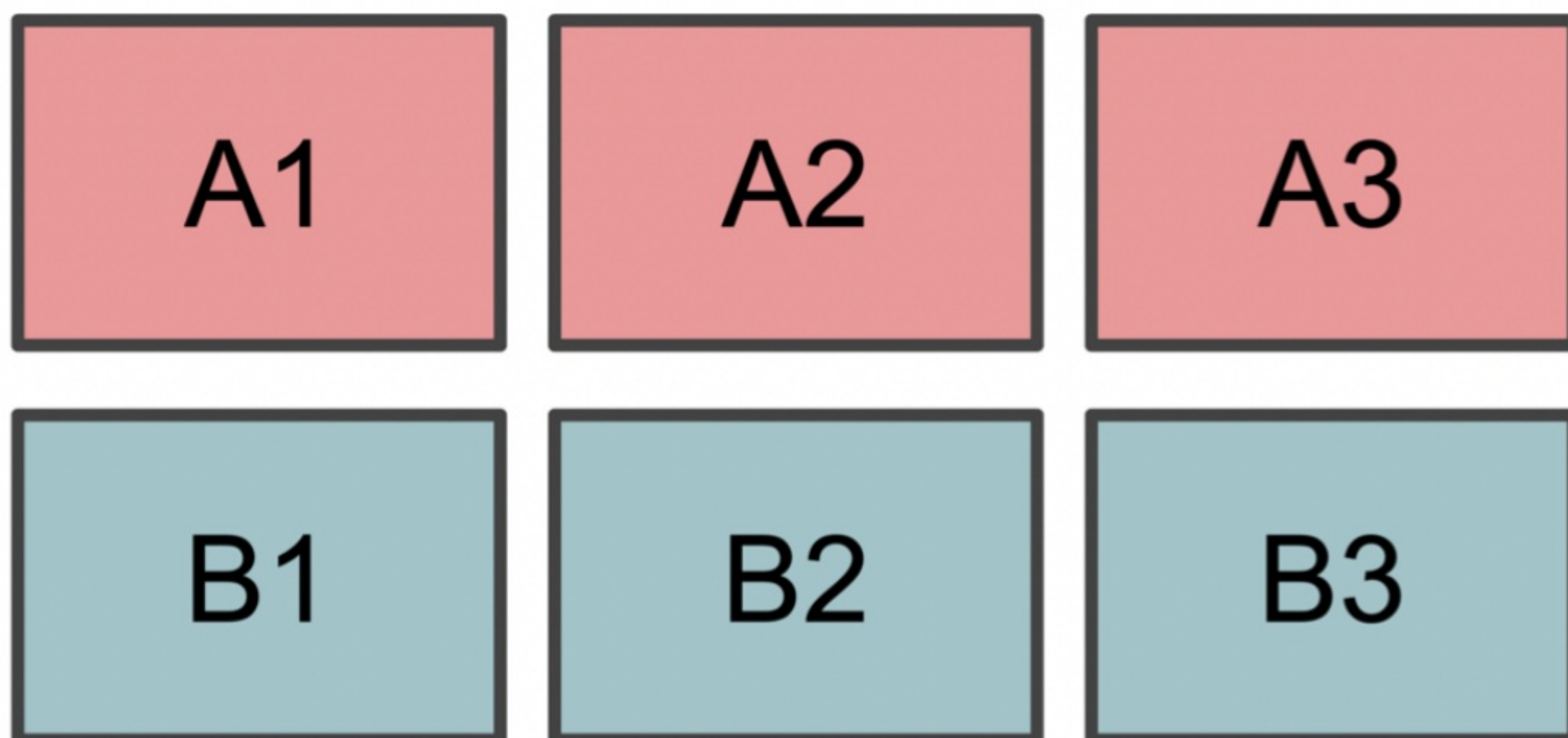
> Независимые гипотезы

> Много гипотез

Ситуация: хотим попробовать новые шрифты на сайте.

Зафиксировали параметры пилота:

- α — уровень значимости;
- β — допустимая ошибка II рода;
- ϵ — ожидаемый эффект;
- n — размеры групп.



Подобрали контрольные и пилотные группы, на каждой проводится A/B-тест

> Обозначения

	H_0 верна	H_0 неверна	Всего
Не отвергаем H_0	TN	FN	TN+FN
Отвергаем H_0	FP	TP	FP+TP
Всего	TN+FP	FN+TP	m

- m — общее число гипотез;
- TN — число истинно отрицательных результатов;
- FN — число ложно отрицательных результатов;
- FP — число ложно положительных результатов;
- TP — число истинно положительных результатов.

> Групповая гипотеза

> Нулевая гипотеза при множественном тестировании

При множественном тестировании мы хотим узнать, **есть ли** среди тестируемых нами гипотез **те, которые приводят к улучшению**.

Базовым предположением будет, что все наши изменения **не приводят** к значимым отличиям.

Альтернатива состоит в том, что **какой-то из экспериментов** действительно даёт улучшение.

> Когда можно применять групповую гипотезу?

Групповая гипотеза применяется, если у нас идут тысячи экспериментов — например, это может происходить в биоинформатике. Или если мы находимся на очень высоком уровне развития IT продукта, когда скорее всего все сливки уже сняты и большинство экспериментов уйдёт в мусор.

В этом случае мы хотим убрать большинство случайных срабатываний, и можно применить множественное тестирование.

> FWER

FWER (*family-wise error rate*) — групповая вероятность ошибки I рода.

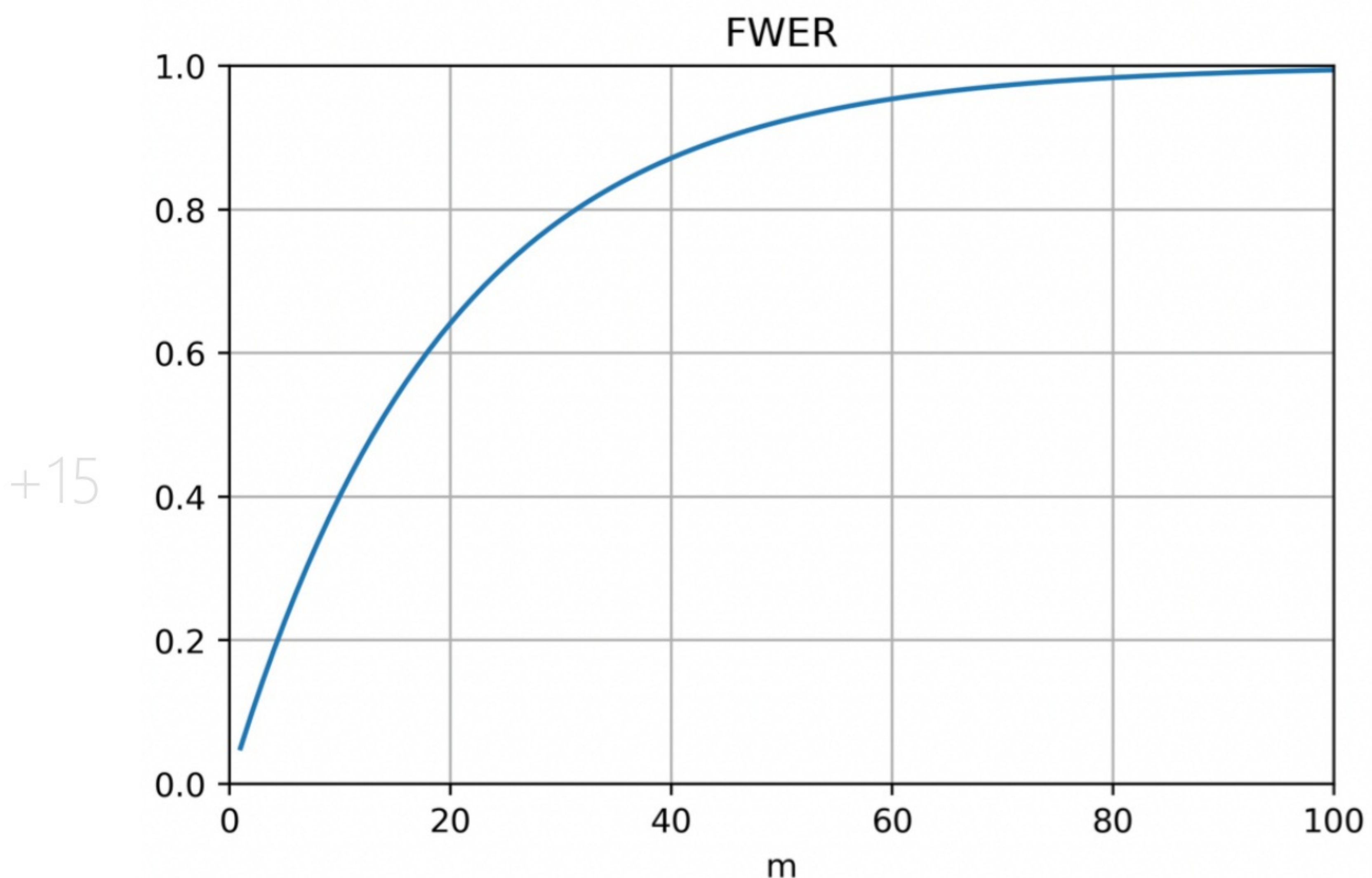
$$FWER = \mathbb{P}(FP > 0)$$

При верности нулевых гипотез и уровне значимости α для m тестов.

$$FWER = \mathbb{P}(FP > 0) = 1 - \mathbb{P}(FP = 0) = 1 - (1 - \alpha)^m$$

Для $\alpha = 0.05$

- $FWER \approx 0.1$ при $m = 2$
- $FWER \approx 0.4$ при $m = 10$



При увеличении числа гипотез мы получаем частые случайные срабатывания. Хочется **контролировать** групповой **уровень ошибок**.

> Метод Бонферрони

Хотим контролировать $FWER$ на уровне значимости α . $\alpha_1, \dots, \alpha_n$ — уровни значимости проверки гипотез.

Задача: выбрать $\alpha_1, \dots, \alpha_n$ так, чтобы $FWER \leq \alpha$.

Определим уровни значимости как $\alpha_1, \dots, \alpha_n = \frac{\alpha}{n}$, тогда $FWER \leq \alpha$

Пусть p_i — p-value i -й гипотезы, k — число верных нулевых гипотез. Тогда

$$FWER = \mathbb{P}(\cup_{i=1}^k \{p_i \leq \frac{\alpha}{n}\}) \leq \sum_{i=1}^k \mathbb{P}(p_i \leq \frac{\alpha}{n}) = \sum_{i=1}^k \frac{\alpha}{n} = \frac{k}{n} \alpha \leq \alpha$$

Замечание: Следует помнить, что уменьшая α для каждого отдельного теста, мы тем самым **сильно** уменьшаем мощность этих тестов. Растёт риск пропустить настоящий эффект.

Замечание 2: В дополнение к вышесказанному метод Бонферрони перестраховывается в отношении ошибки I рода, т.е. мощность такой статистической процедуры снижается.

> Недостаток метода Бонферрони

> Отсутствие взаимосвязи между тестами отдельных гипотез

В методе Бонферрони отдельные тесты не связаны друг с другом. Мы раз и навсегда выбрали для каждого из них новый уровень значимости и этим ограничились.

Фактически мы вместо тестирования групповой гипотезы проводим серию индивидуальных тестов на существенно более низком уровне значимости.

> Расплата мощностью критерия

За простоту метода Бонферрони мы расплачиваемся снижением мощности критерия.

Так эффект в 4σ мы можем пропустить в 30% случаев при $n = 100$ и более чем в половине случаев при $n = 1000$.

$n =$	$\Delta = 3\sigma$	$\Delta = 4\sigma$	$\Delta = 5\sigma$
1	14.92%	2.07%	0.12%
10	42.35%	11.64%	1.42%
100	68.47%	30.18%	6.44%
1000	85.44%	52.22%	17.25%

Ошибка второго рода для метода Бонферрони при $\alpha = 5\%$

> Распределение p-value

> Когда выбираем уровни значимости?

В методе Бонферрони мы **сначала** выбрали **уровни значимости** для всех тестов, а **потом** стали считать **p-value**. Но ведь можно сделать и **наоборот**! Сначала посчитать p-value, а затем выбрать уровни значимости.

> Равномерное распределение

При условии справедливости нулевой групповой гипотезы p-value равномерно распределены на $[0, 1]$. Попробуем это учесть.

Большая часть вычисленных p-value будут далеки от нуля. В фокусе нашего интереса должны оказаться те тесты, для которых полученные значения малы. Эта идея позволяет построить процедуры **с большей мощностью** критериев.

> Метод Холма

Построим вариационный ряд p-value: $p_{(1)}, \dots, p_{(m)}$.

$H_{(1)}, \dots, H_{(m)}$ — соответствующие нулевые гипотезы.

Определим уровни значимости $\alpha_i = \frac{\alpha}{m - i + 1}$

Процедура Холма:

1. Если $p_{(1)} \geq \alpha_1$, то говорим, что все нулевые гипотезы не противоречат данным, и останавливаемся. Иначе отклоняем $H_{(1)}$ и продолжаем процедуру.

2. Если $p_{(2)} \geq \alpha_2$, то говорим, что оставшиеся гипотезы не противоречат данным, и останавливаемся. Иначе отклоняем $H_{(2)}$ и продолжаем процедуру.

3. ...

Процедура обеспечивает $FWER \leq \alpha$

Метод Холма имеет большую мощность, т.к. все уровни значимости не меньше уровней значимости метода Бонферрони.

> Метод Бенджамини-Хохберга

FDR (*false discovery rate*) — средняя доля ложных отклонений.

$$FDR = \mathbb{E}\left(\frac{FP}{\max(FP + TP, 1)}\right)$$

где $FP + TP$ — число отвергнутых нулевых гипотез.

> Описание метода Бенджамини-Хохберга

Построим вариационный ряд p-value: $p_{(1)}, \dots, p_{(m)}$.

Определим уровни значимости $\alpha_i = i\alpha/m$

Процедура Бенджамини-Хохберга:

1. Если $p_{(1)} \geq \alpha_1$, то говорим, что все нулевые гипотезы не противоречат данным, и останавливаемся. Иначе отклоняем $H_{(1)}$ и продолжаем процедуру.
2. Продолжаем аналогично процедуре Холма.

Процедура обеспечивает $FDR \leq \alpha$.

> Сравнение методов

	Без поправок	
	H_0 верна	H_0 не верна
Не отвергаем H_0	379	1.3
Отвергаем H_0	21	98.7

	Метод Холма	
	H_0 верна	H_0 не верна
Не отвергаем H_0	399.9	37.2
Отвергаем H_0	0.1	62.8

	Метод Бонферрони	
	H_0 верна	H_0 не верна
Не отвергаем H_0	399.9	38.6
Отвергаем H_0	0.1	61.4

	Метод Бенджамини-Хохберга	
	H_0 верна	H_0 не верна
Не отвергаем H_0	395.7	5.1
Отвергаем H_0	4.3	94.9

Сравнение методов. 500 синтетических экспериментов

Без поправок. Контролируемая доля верных нулевых гипотез отвергается. Мы получаем значительный процент ложноположительных результатов. При этом для альтернативных гипотез сохраняется мощность критерия. Если будет реальный положительный эффект, то мы его скорее всего заметим.

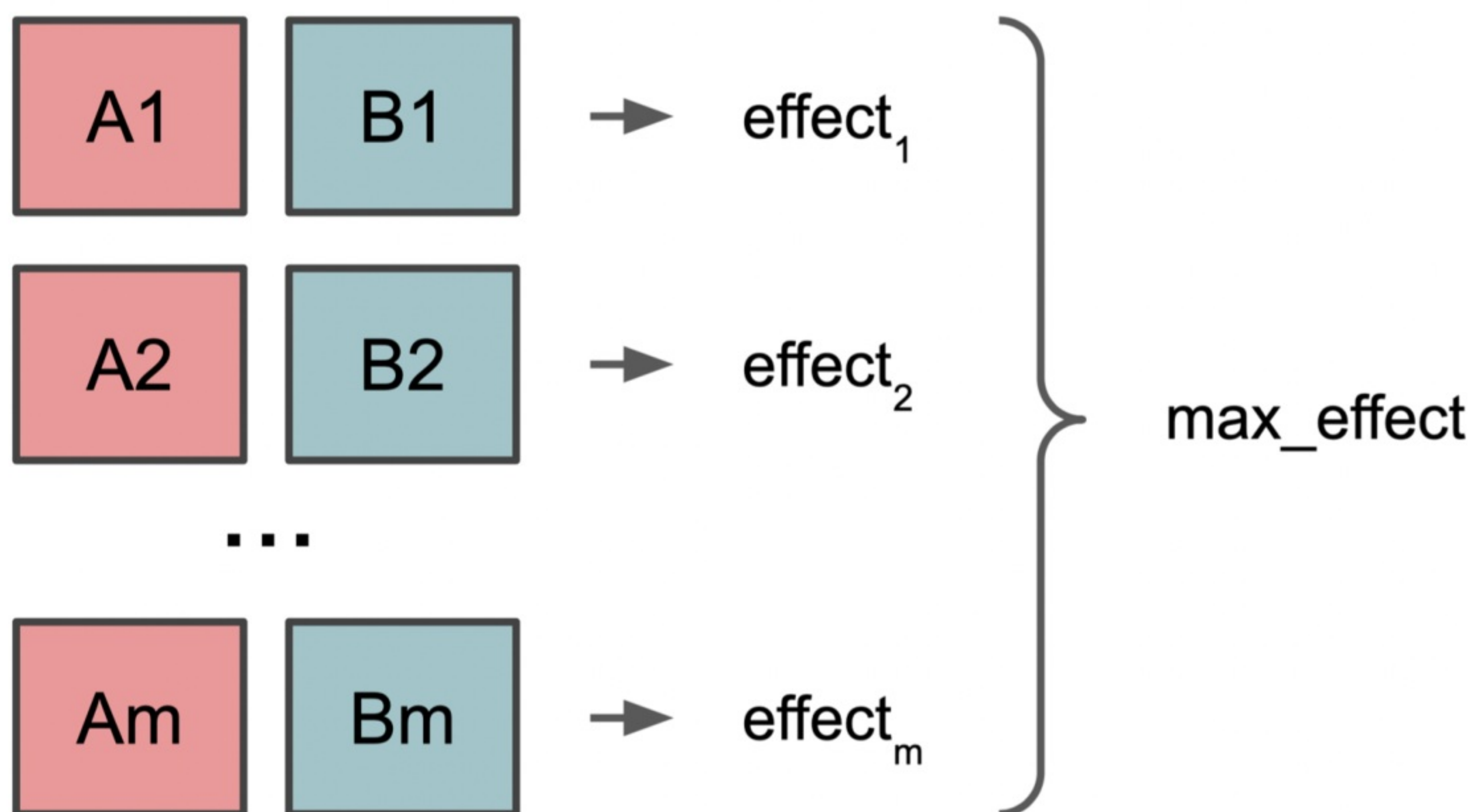
Метод Бонферрони и метод Холма. Контролируется вероятность хоть раз отвергнуть верную нулевую гипотезу. За счёт этого почти нет ложноположительных результатов. Это достигается ценой пропуска значительного числа верных альтернативных гипотез.

Метод Бенджамини-Хохберга. Контролируется FDR. Мы куда чаще, чем в предыдущих методах, отклоняем верную нулевую гипотезу. Но это происходит на контролируемом уровне. При этом значительно повышаем мощность критерия, что позволяет реже пропускать положительные эффекты.

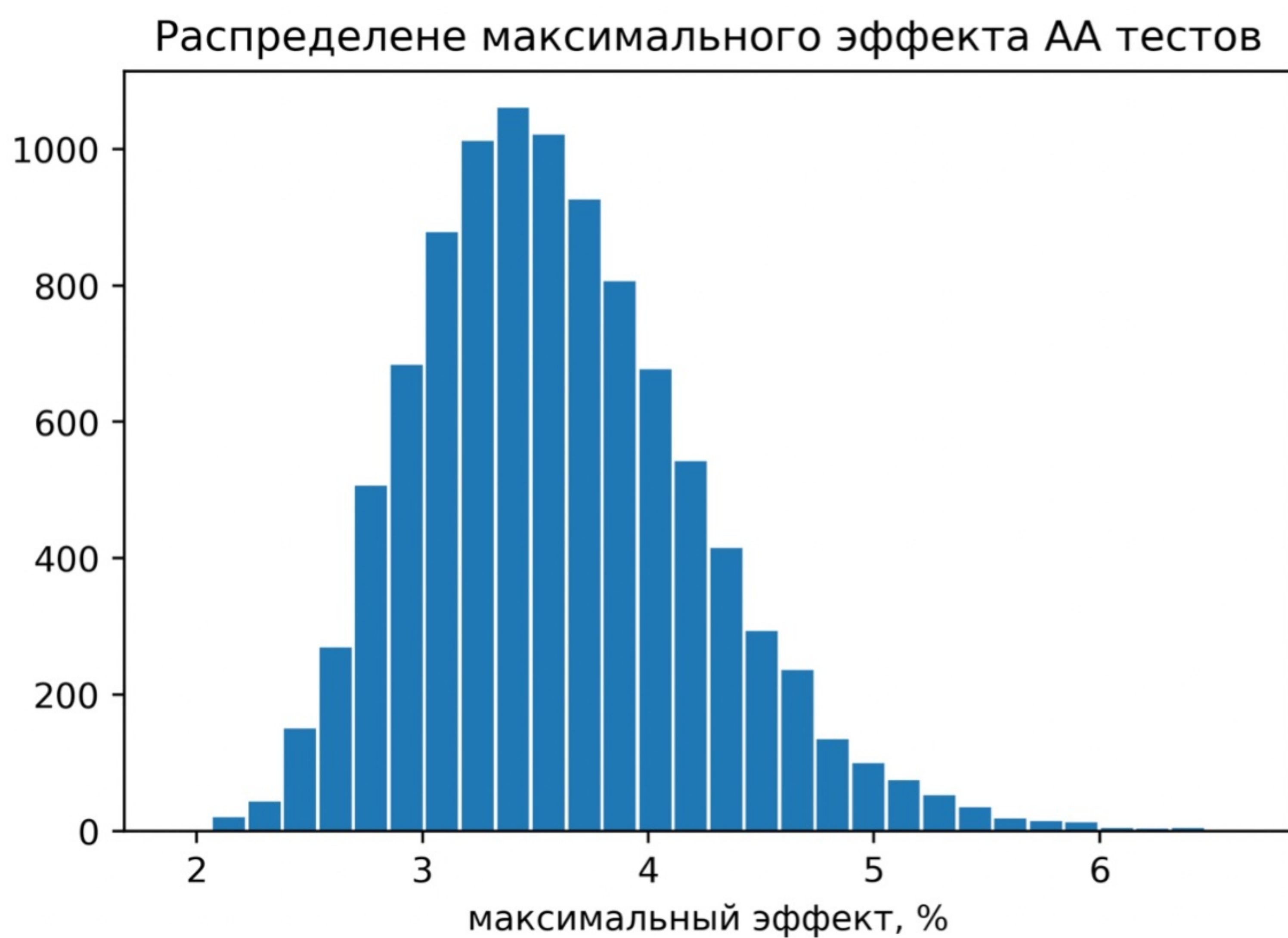
> Большие выбросы

Больше экспериментов — больше вероятность больших выбросов.

Проведём 10000 A/A-тестов, вычислим максимальный эффект. Повторим много раз, построим распределение максимальных эффектов.



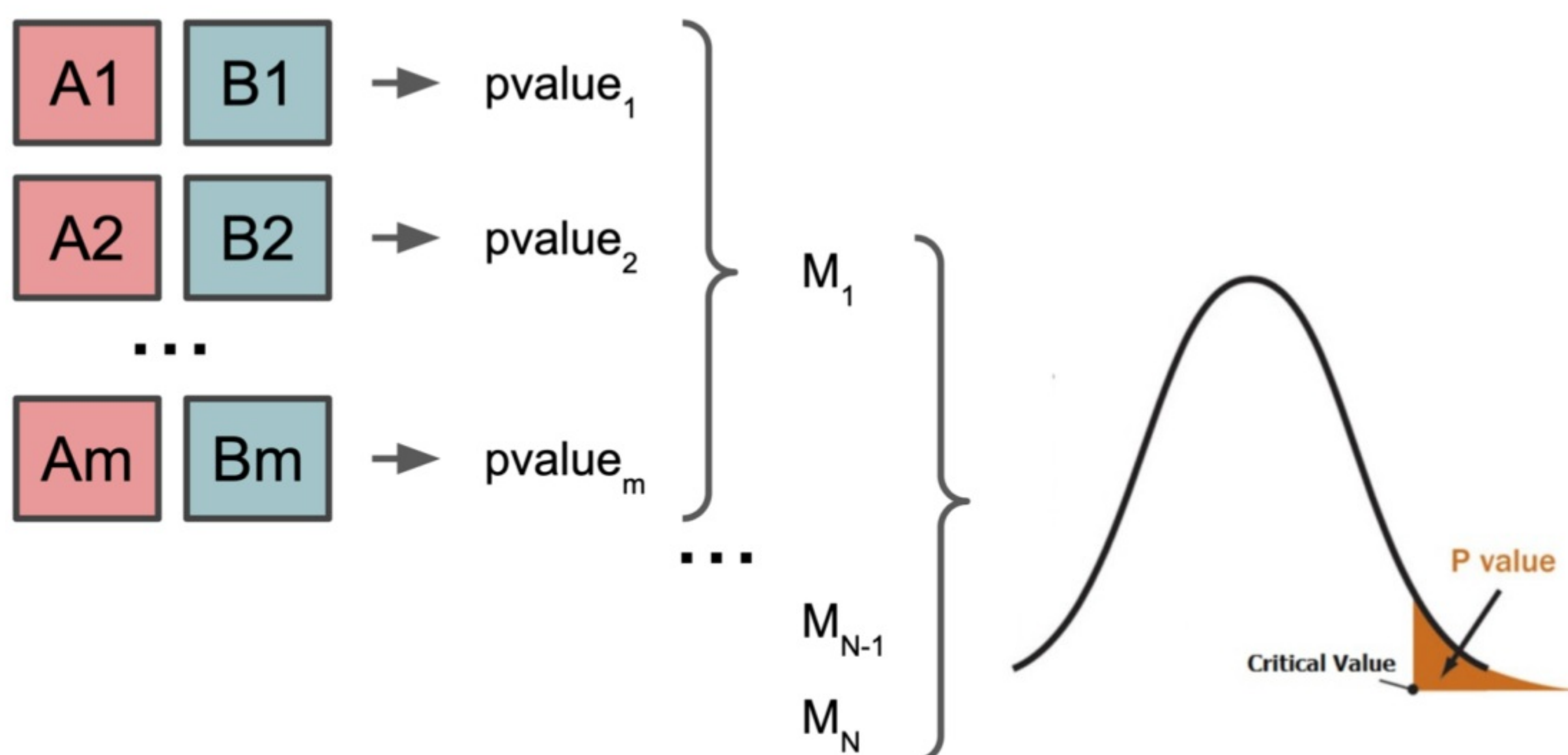
- Случайные выбросы могут быть больше ожидаемых эффектов.
- Нельзя отличить выброс от реального эффекта.



Нужно **осознанно выбирать гипотезы** для эксперимента, а не перебирать вручную множество параметров.

> Один эксперимент и много метрик

Провели 100 экспериментов, 7 показало значимый эффект.



Вопрос: случайность или есть реальное улучшение?

Мы знаем, что в отсутствие эффекта:

1. p-value равномерно распределено;
2. Число покрасившихся тестов должно иметь биномиальное распределение.

Это позволяет понять, отклонились ли мы от ожидаемых значений. В отсутствии эффекта отклоняться мы не должны.

> Зависимые гипотезы

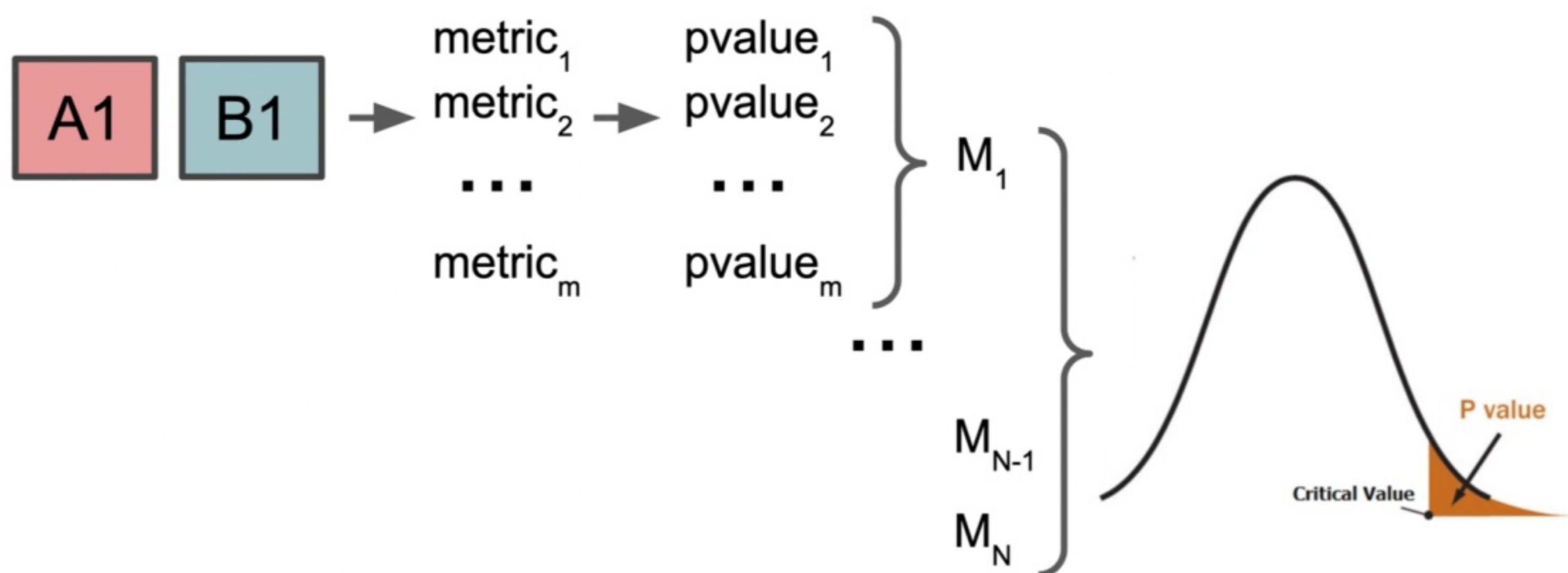
> Один эксперимент и много метрик

Покрасилась половина вспомогательных метрик.

Проблема: метрики сильно взаимосвязаны.

Вопрос: случайность или есть реальное улучшение?

КУПЛЕНО НА
SKLADCHIK.COM



Пусть при вычислении метрик и расчёте p-value для каждой оказалось, что многие значимы. Выглядит как множественная проверка. При применении поправок может оказаться так, что ни один эффект с ними не значим — парадокс. Поэтому в таком случае не рекомендуется использовать стандартные поправки.

1. Определим метрику M , характеризующую вспомогательные метрики: — количество прокрасившихся метрик; — среднее значение p-value.
2. Оценим распределение этой метрики, предполагая, что эффектов нет.
3. Вычислим $p\text{-value}(M)$ метрики по полученному распределению.

> Параллельный запуск экспериментов

> Много гипотез и мало времени

Любая крупная компания живёт в условиях следующих ограничений:

- Много гипотез по улучшению показателей;
- Скорее всего улучшения незначительны, а значит, для их детектирования нужно много пользователей и времени;
- Если хорошие гипотезы долго стоят в очереди, то компания теряет деньги.

> Параллельный запуск

Эксперименты запускаются только на малой доле пользователей. Возникает соблазн провести сразу несколько экспериментов.

Важно: эксперименты не должны оказывать влияние друг на друга.

> Одномерная и многомерная схемы запуска экспериментов

> Схемы параллельного запуска экспериментов

Одобрённые эксперименты попадают в очередь ожидания. Как сократить время нахождения в очереди? Можно проводить сразу несколько экспериментов.

- **Одномерная схема.** Каждый пользователь попадает не более чем в один эксперимент. Лучшая схема, но возможна только при небольшом числе экспериментов.
 - **Многомерная схема.** Пользователь может попасть сразу в несколько экспериментов. Разумеется, пересекающиеся эксперименты не должны конфликтовать друг с другом.
-

> Применение многомерной схемы

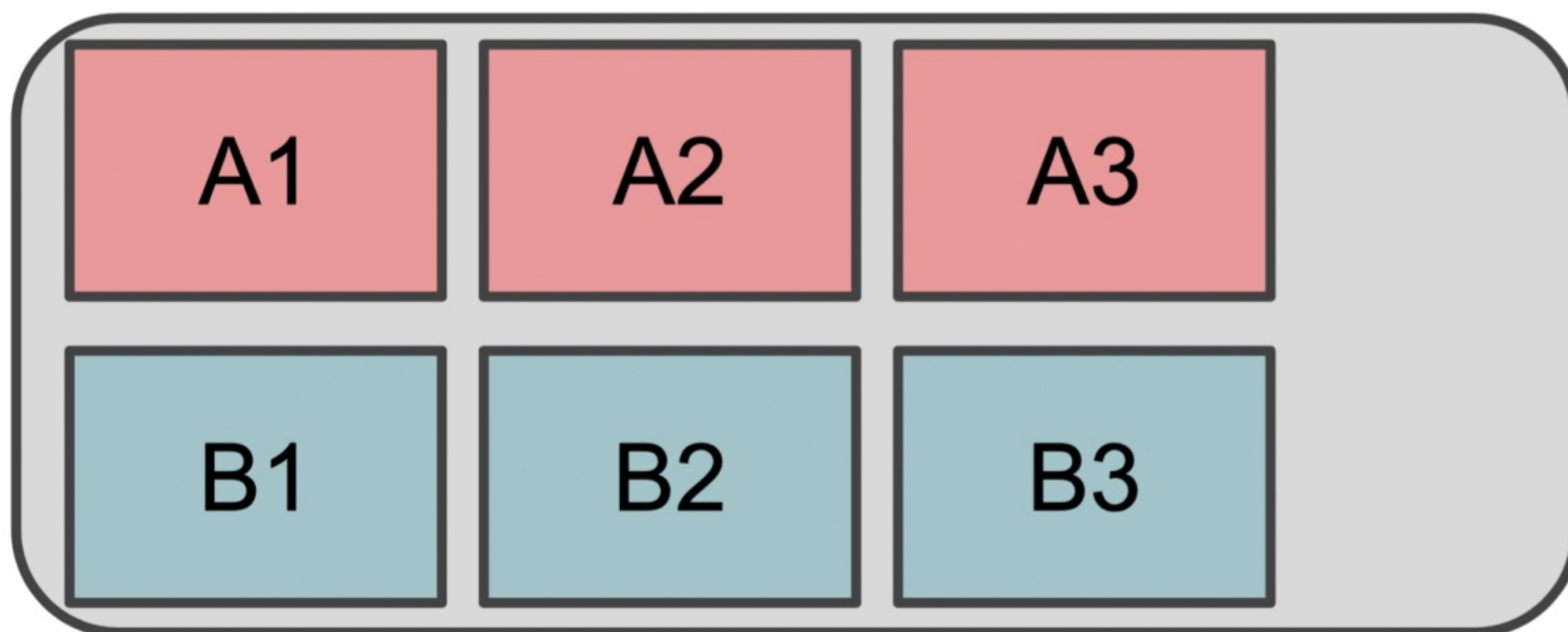
Многомерная схема позволяет проводить одновременно значительно больше экспериментов. При этом пересекающиеся эксперименты должны оставаться независимыми.

Обычно они относятся либо к разным сервисам, либо к разным аспектам качества одного сервиса.

> Одномерная схема

> Строго изолированные эффекты

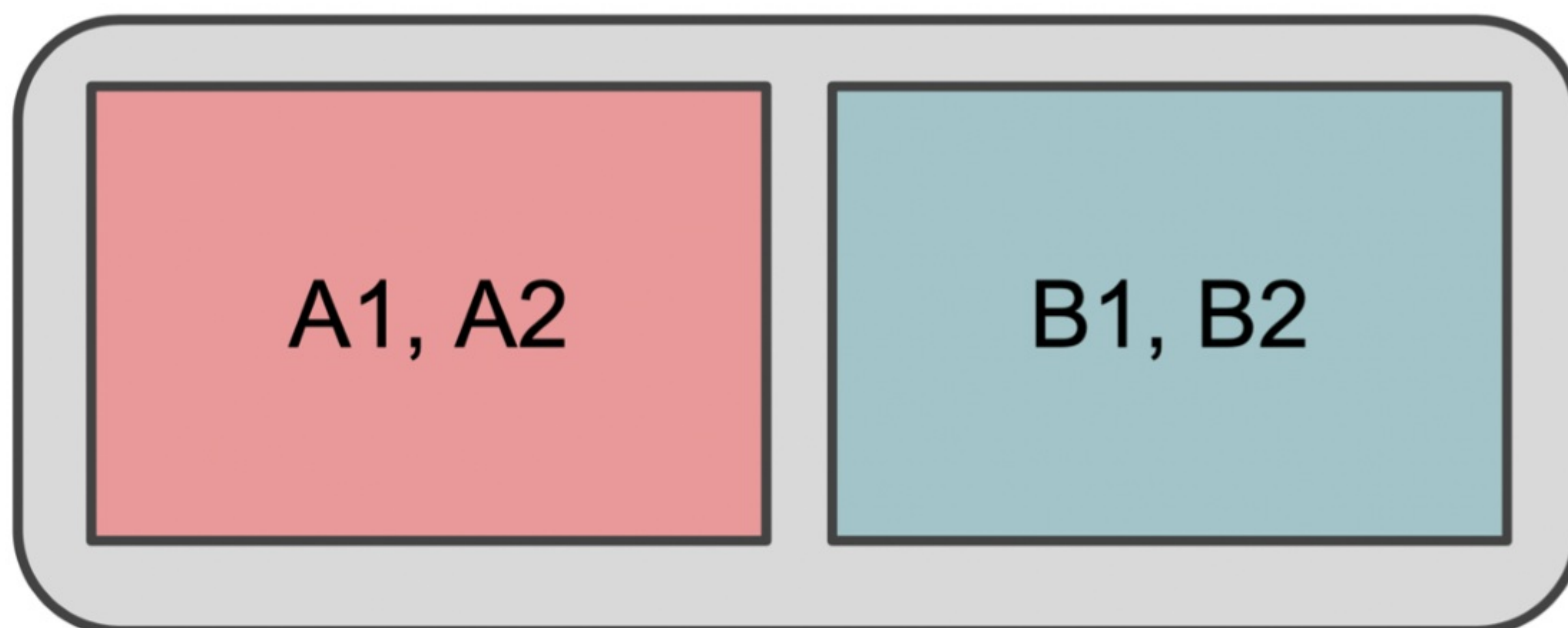
Если для всех экспериментов удастся подобрать пилотные и контрольные группы так, что каждый пользователь участвует не более чем в одном эксперименте, то мы получим строго изолированные результаты.



Но экспериментов много, а пользователей мало. Чаще всего отдельные пользователи одновременно участвуют в нескольких экспериментах.

> Многомерная схема. Коррелированные группы

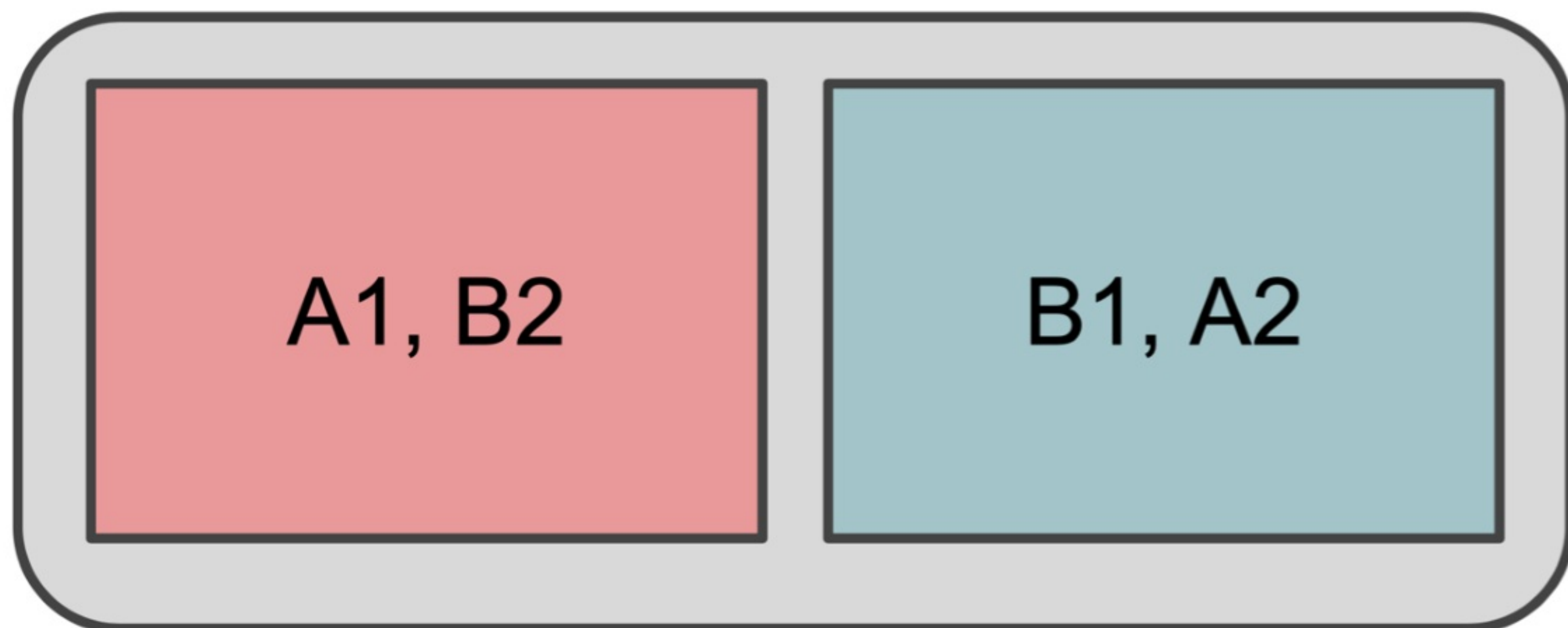
> Одна и та же группа в двух экспериментах



Если одна и та же группа пользователей будет пилотной в двух экспериментах:

- Незначимые эффекты могут сложиться и стать статзначимыми; получим два ложных срабатывания.
- Если в одном эксперименте есть эффект, а в другом нет, то мы не сможем понять, что нам дало выигрыш.

> Пилот и контроль в двух группах перевёрнуты

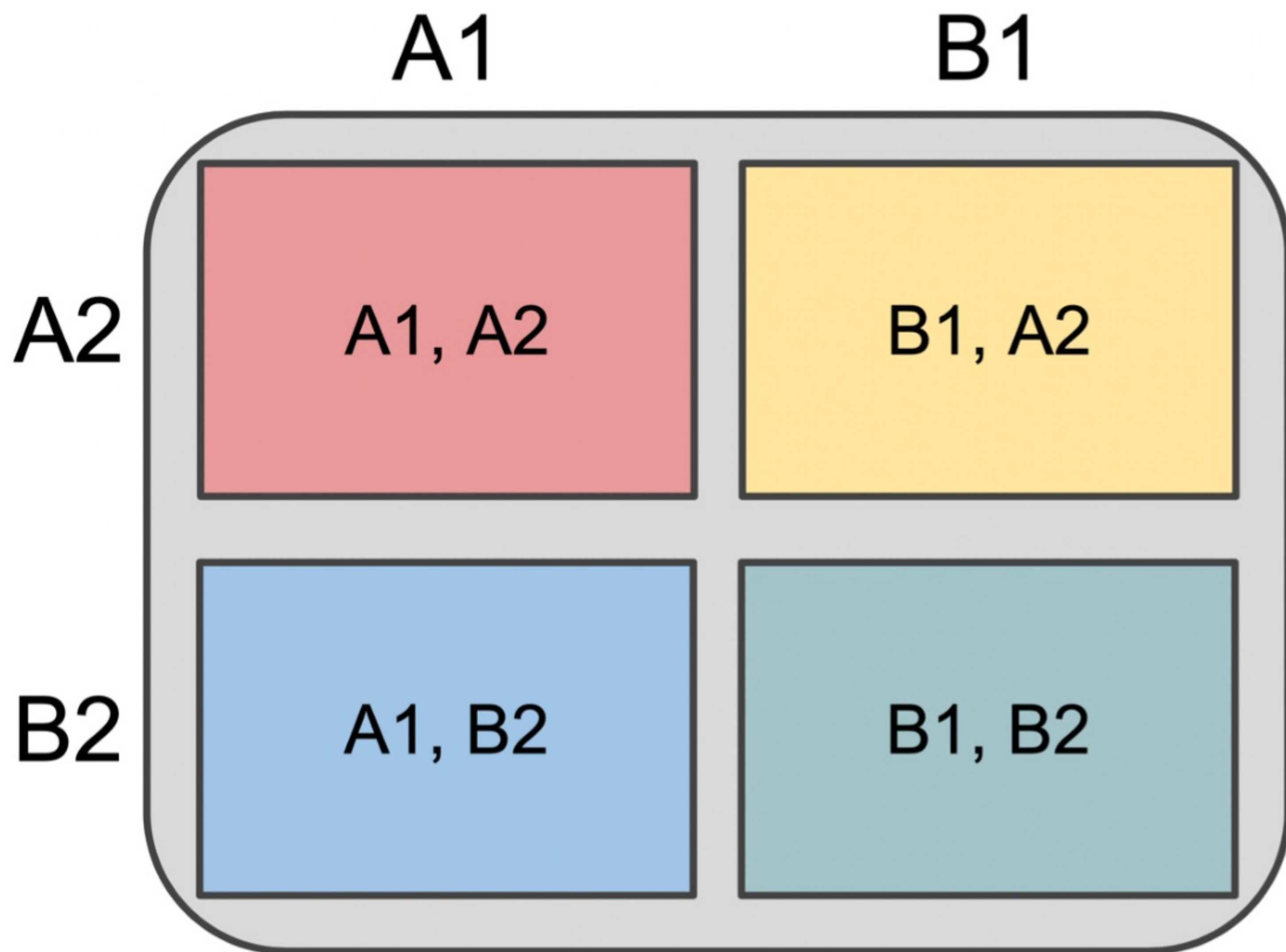


Если **пилот и контроль в двух группах перевёрнуты**:

- Если два значимых эффекта будут вычтены друг из друга, то можем получить незначимый эффект; получим два пропуска эффекта.
- Как и раньше, нельзя сделать вывод, что сыграло главную роль.

> **Некоррелированные группы**

+15



Необходимо, чтобы пользователи, участвующие в нескольких экспериментах, были одинаково представлены в группах A и B как в одном, так и в другом эксперименте.

Такое разбиение называется **ортогональным**.

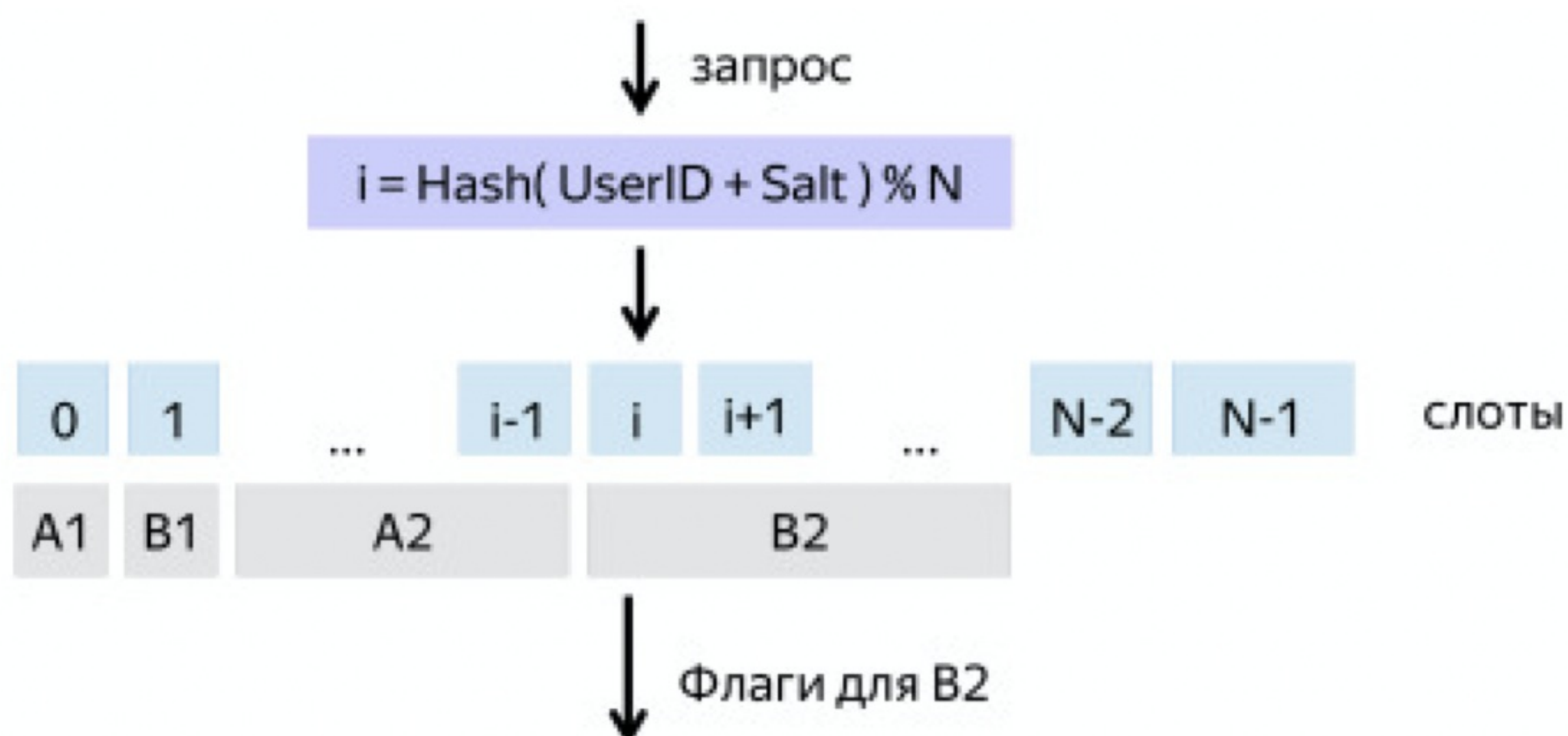
Если в каком-то эксперименте возникнет перекос между группами A и B , то мы получим смещённые результаты.

> Разбиение на основе хэша

> **Слоты**

Мы можем взять хэш от UID пользователя и найти остаток по модулю N .

Хорошая функция хэширования обеспечит случайное разбиение пользователей на N слотов.



> **Распределение слотов по экспериментам**

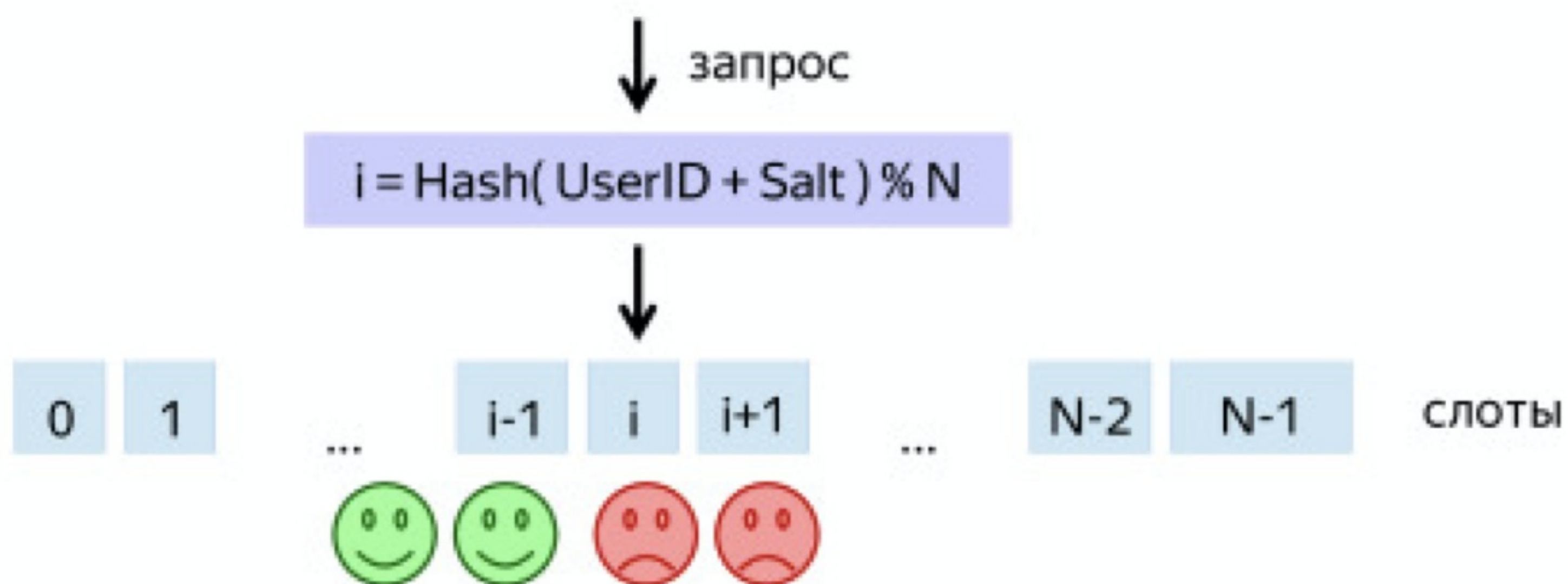
Полученные слоты можно отправить в эксперименты.

Система A/B-тестирования может распределить эти слоты в соответствии с заданными параметрами экспериментов в одномерной схеме. При этом гарантируется отсутствие пересечений.

> **Память пользователей. Двойное хэширование**

Если перед нами работал эксперимент, в котором группе *A* было хорошо, а *B* чуть хуже, то пользователи соответствующих слотов начинают вести себя по-разному.

Когда включаем свой эксперимент, возникает смещение: *A* и *B* находятся в неравных условиях.

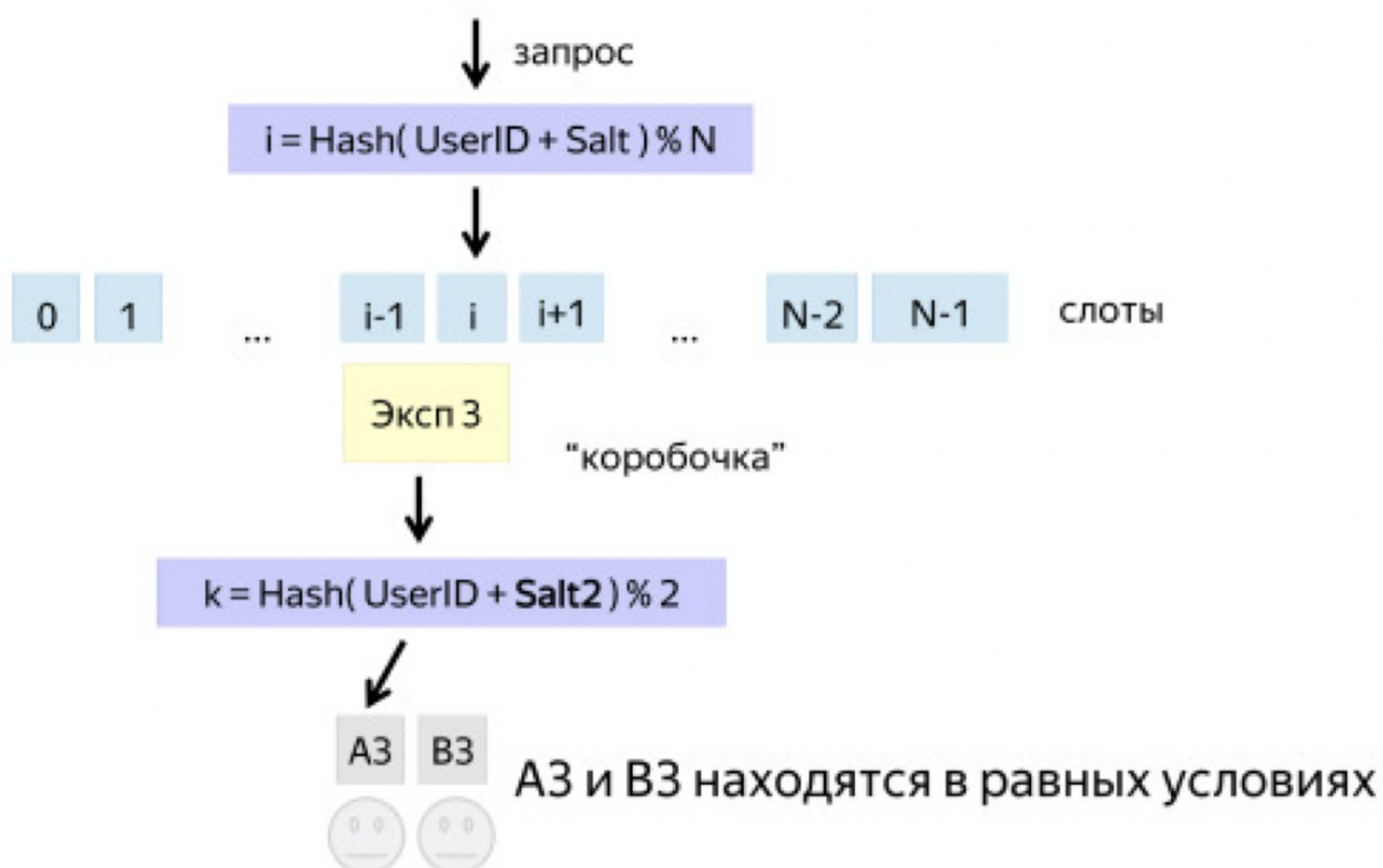


Например, у пользователей изначально разный опыт работы с платформой

> Двойное хэширование

Нам нужно избежать корреляции групп A и B между экспериментами.

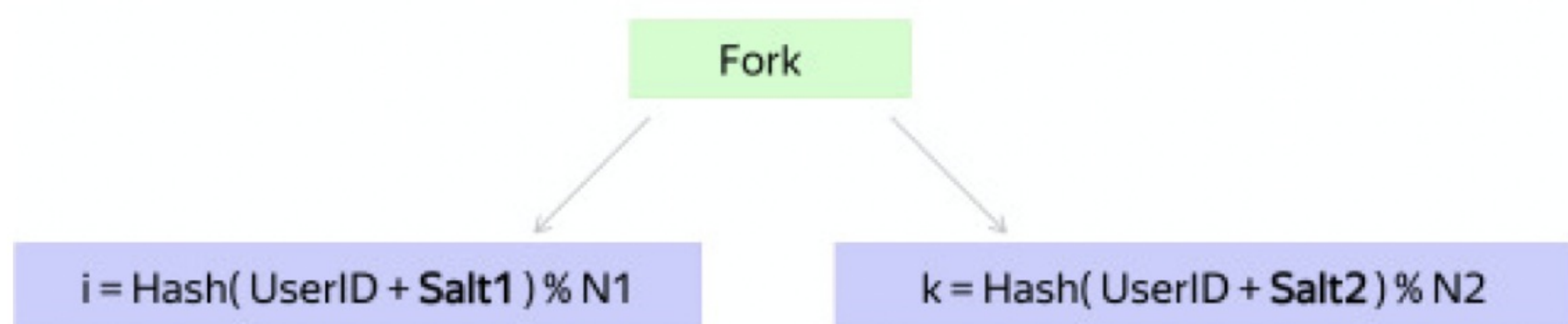
Ортогональности можно достичь, если ввести дополнительное хэширование внутри слотов, попавших в один эксперимент. По этому хэшу мы распределяем пользователей на группы A и B . **Не забывайте про подсаживание!**



> Выделение слотов в многомерной схеме

Обычно у нас есть независимые ветки изменений: дизайн, алгоритм ранжирования, работа клиентской службы и т.п. Можно полагать, что изменения в разных ветках не будут влиять друг на друга, а конфликты могут возникать только внутри одной ветки.

Когда измерения ведутся в разных ветках, они имеют разные значения *Salt1* и *Salt2*. Это обеспечивает независимость разбиения. Внутри одной ветки соль будет совпадать.



> Тройной "посол"

Самые важные изменения мы хотим тестировать на выборках, не затронутых никакими экспериментами. Для этого можно держать отдельную группу пользователей.

> Три этапа хэширования

Можно выделить три этапа разбиения пользователей (каждый со своим этапом подсаливания):

- **Exclusive**. Выделяем 10% трафика, который предназначен для самых важных экспериментов. Эти пользователи не участвуют ни в каких других тестированиях. Соль единая для всех экспериментов.
- **Layer**. Внутри каждой ветки/слоя мы выделяем свою соль, которая обеспечит ортогональность с другими слоями. Эта соль остаётся постоянной для всех экспериментов в одном слое.
- **Shuffle**. Соль для перемешивания внутри слотов между экспериментами. Для каждого эксперимента она уникальна.

> Конфигурация экспериментов

> Требования к экспериментальным группам

Каждый эксперимент имеет набор ограничений:

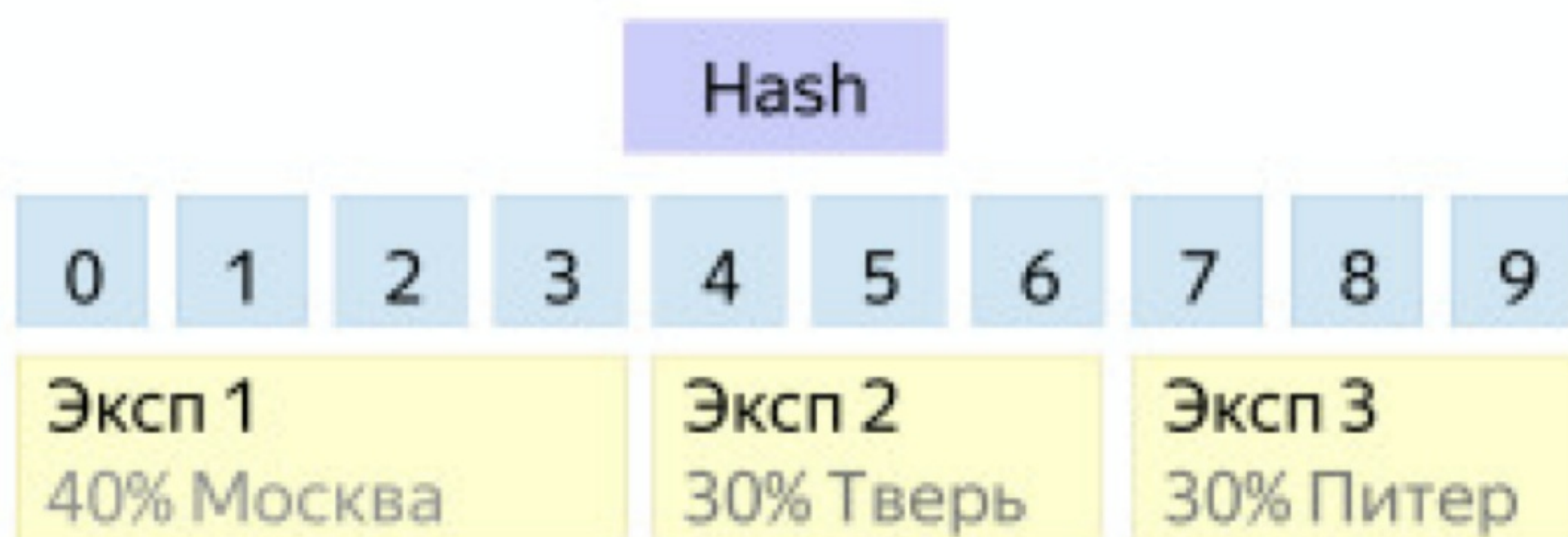
- Размер групп;
- Регионы участия;
- Браузеры;
- Мобильные платформы и т.п.

Необходимо оптимально разместить их с учётом этих ограничений.

> Эвристика размещения экспериментов

> Размещение в свободные слоты

Если размещать все маленькие эксперименты в свободные слоты, то, когда придёт большой эксперимент, ему уже может не хватить места, т.к. он будет перекрываться со всеми остальными.

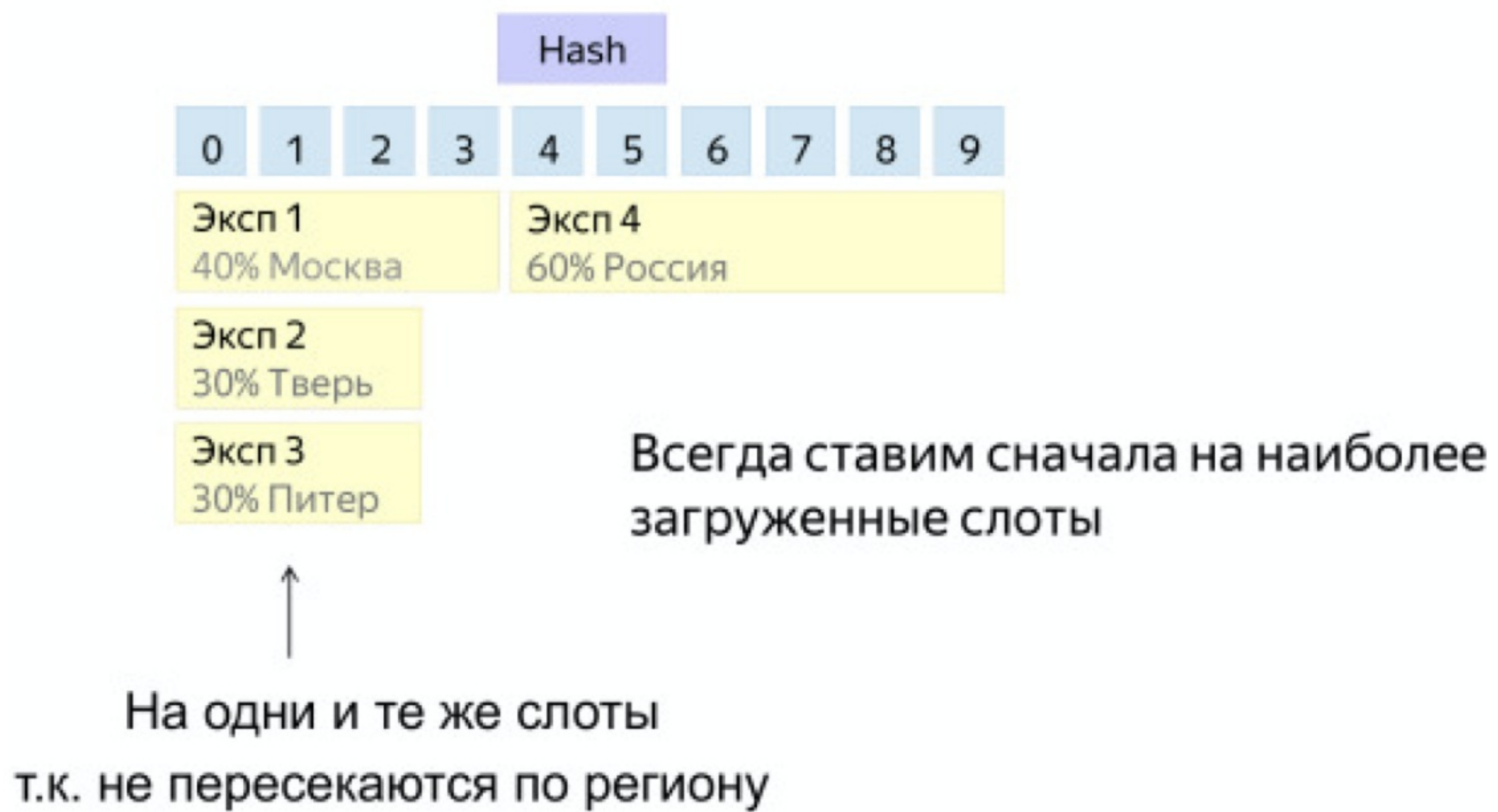


Не поместился!

Эксп 4
60% Россия

> Максимально плотное размещение

Достаточно хорошо работает простая эвристика. Размещая новый эксперимент, нужно стараться выбрать уже используемые слоты, которые не имеют пересечений с нашим экспериментом. Когда придёт большой эксперимент на широкие ограничения, нам нужно, чтобы для него осталось место.



+15

> Доля успешных экспериментов

Полезно ретроспективно смотреть долю успешных экспериментов, а также распределение p-value для A/A-тестов:



стат. значимые
отличия

Допустим, доля составила 10%



реальный эффект

случайные выбросы

При уровне значимости 0.05 и равномерном распределении p-value

Возможно, какую-то долю "успешных" экспериментов можно прогнать ещё раз.

Большое количество **необдуманных** экспериментов (в большей мере, чем обдуманных) приводит к:

- Большим случайным выбросам;
 - Пропуску полезных изменений;
 - Коллизиям при множественном тестировании.
-

> **Дополнительные материалы**

1. [Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing](#)
2. [A sharper Bonferroni procedure for multiple tests of significance](#)
3. [A Simple Sequentially Rejective Multiple Test Procedure](#)
4. [Modified Sequentially Rejective Multiple Test Procedures](#)
5. [Множественная проверка гипотез](#)
6. [Множественные эксперименты: теория и практика](#)
7. [Google. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation](#)
8. [Как устроено A/B-тестирование в Авито](#)
9. [Как у нас устроено A/B-тестирование. Лекция Яндекса](#)