# Метрики в A/B-тестировании

Курс «Продуктовые метрики»

# Как выбрать метрику?

**Требования:**

- Измеримая (measurable)
- Устойчивая (stable)
- Объяснимая (attributable)
- Чувствительная (sensitive)

Kohavi, Tang & Xu — Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (2020)

# Как выбрать метрику?

**OEC** (Overall Evaluation Criterion) — количественная мера, отражающая основную цель эксперимента.

**Типы метрик в эксперименте:**

- Таргет (goal) — метрика успеха организации
- Драйверы (driver / proxy) — влияющие на успех
- Guardrail / контр-метрики — метрики-ограничители, которые сигнализируют о нарушении

+

- Качество данных (data quality) — метрики валидности эксперимента
- Диагностики (diagnostic) / информативные / debug — дополнительная гранулярность, контекст

Kohavi, Tang & Xu — Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (2020)

# Как выбрать метрику?

**Правила выбора:**

- Таргет (goal):
  - Понятность
  - Устойчивость
- Драйверы:
  - Сонаправленность с таргет-метрикой
  - Управляемость
  - Чувствительность
  - Устойчивость к манипуляциям

Kohavi, Tang & Xu — Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (2020)

# Как выбрать метрику?

## Как на практике?

**Метрик-сет:**

- Таргет-метрика (P0)
- Метрики-драйверы (P1) — при дизайне эксперимента мы считаем «таргет»-метрикой именно драйвер, то есть H0 чаще всего формулируется для них
- Информативные гранулярные метрики, дающие контекст (P1-P2-P3)
- Guardrails (P3-P4; P1-P2 других команд, если есть риск каннибализации)
- Debug (P3-P4)
- Качество данных (метрики свойств эксперимента)

Не забываем, что при увеличении числа метрик в эксперименте появляется проблема множественного тестирования. Самый простой вариант решения — снижать alpha.

# Как выбрать метрику?

**Пример:** Тестирование push-уведомления о брошенной корзине в e-commerce приложении.

**Метрик-сет:**

- Таргет-метрики: GMV per User, Orders per User
- Метрики-драйверы: CR from Cart to Purchase
- Информативные: AOV, CTR, Orders per User by Category
- Guardrails: Unsubscribe Rate, Uninstall Rate, Errors, User Complaints
- Debug: Delivery Rate
- Качество данных: Balance Test / Control

# Как выбрать метрику?

## Как принять решение о раскатке?

Среди ключевых метрик эксперимента (таргет + драйверы + guardrails):

- все метрики значимы положительно / отрицательно — катим / не катим;
- стат значима лишь одна и эффект положительный — катим;
- стат значима лишь одна и эффект отрицательный — не катим;
- все не стат значимы — не катим, думаем о перезапуске (повышение чувствительности, изменение дизайна);
- часть значима положительно, часть отрицательно — принимаем решение на основе трейд оффа.

Kohavi, Tang & Xu — Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (2020)

# Как выбрать метрику?

## Как принять решение о раскатке?

Для целей A/B-тестирования можно разработать взвешенную метрику, чтобы проще принимать решения в последнем кейсе.
Пример OEC для Email-рассылок в Amazon:

$$OEC = \frac{\Delta\text{Revenue} - \left(s \times \text{Unsubscribe\_Lifetime\_Loss}\right)}{\text{Number\_of\_Users}}$$

- Δ Revenue — инкрементальная выручка от email рассылки в результате эксперимента
- s — число пользователей, отписавшихся от получения писем
- Unsubscribe_lifetime_loss — оценка потери выручки от отсутствия коммуникаций с юзером в email
- Number_of_Users — число пользователей в тестовой группе

Kohavi, Tang & Xu — Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (2020)

# Поиск прокси-метрики

# Поиск прокси-метрики

**Прокси-метрика** — краткосрочный индикатор, предсказывающий долгосрочный эффект таргет-метрики.

**Требования:**

- Чувствительность — позволяет выявить стат значимый эффект за более короткий период, чем требуется для таргета.
- Сонаправленность (Label Agreement) — изменения прокси сонаправлены с изменениями таргета, поэтому можно принимать решение о раскатке, опираясь на неё.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**Единого алгоритма поиска прокси-метрики не существует.**

Но есть набор проверок, на которых можно построить алгоритм внутри компании:

- Binary Sensitivity (бинарная чувствительность)
- MSE с таргет-метрикой
- Корреляция с таргет-метрикой
- Proxy Score (или Label Agreement)

Для поиска прокси-метрик необходимо иметь накопленный корпус экспериментов. Мы оцениваем способность прокси предсказывать таргет именно по эффектам в экспериментах, а не по динамике самих временных рядов.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**Binary Sensitivity** — доля экспериментов, в которых метрика показала статистически значимый эффект.

Рассчитывается для каждой метрики (и для прокси, и для таргет) по формуле:

$$\text{Binary Sensitivity}(\text{metric}) = \frac{|\{\text{experiments where metric is significant}\}|}{|\{\text{all experiments}\}|}$$

Обратите внимание: Binary Sensitivity ничего не говорит о сонаправленности, она лишь помогает оценить, получаем ли мы для кандидатов в прокси стат значимые результаты чаще, чем для таргета.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**MSE (Mean Squared Error)** показывает, насколько величина эффекта прокси отклоняется от эффекта таргета в разных экспериментах.

Рассчитывается попарно для каждой прокси и таргета по формуле:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta^{rel}_{\text{proxy},i} - \Delta^{rel}_{\text{target},i}\right)^2 \text{, где} \quad \Delta^{rel} = \frac{\bar{x}_{test} - \bar{x}_{control}}{\bar{x}_{control}}$$

Меньшее значение MSE означает, что прокси ближе к таргету по оценке эффекта. Применяется только вместе с другими метриками для комплексной оценки.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**Корреляция** измеряет силу и направление связи между эффектами прокси и таргета в исторических экспериментах.

Рассчитывается попарно для каждой прокси и таргета по формуле:

$$\rho = \frac{\sum_{i=1}^{N} \left(\Delta^{rel}_{\text{proxy},i} - \overline{\Delta^{rel}_{\text{proxy}}}\right)\left(\Delta^{rel}_{\text{target},i} - \overline{\Delta^{rel}_{\text{target}}}\right)}{\sqrt{\sum_{i=1}^{N} \left(\Delta^{rel}_{\text{proxy},i} - \overline{\Delta^{rel}_{\text{proxy}}}\right)^2} \cdot \sqrt{\sum_{i=1}^{N} \left(\Delta^{rel}_{\text{target},i} - \overline{\Delta^{rel}_{\text{target}}}\right)^2}}$$ , где $$\Delta^{rel} = \frac{\bar{x}_{test} - \bar{x}_{control}}{\bar{x}_{control}}$$

В примере корреляция Пирсона, но можно использовать любую другую.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**Proxy Score** показывает, насколько прокси-метрика согласуется с таргет-метрикой в исторических экспериментах.

Рассчитывается попарно для каждой прокси и таргета по формуле:

$$\text{Proxy Score} = \frac{\text{Detections} - \text{Mistakes}}{\#\{\text{experiments where target is significant}\}}$$

- Detections — число экспериментов, где прокси и таргет значимы и сонаправлены по направлению эффекта.
- Mistakes — число экспериментов, где прокси значима, но не совпадает с таргетом по направлению.

Значение Proxy Score лежит в диапазоне от −1 до 1. Чем ближе к 1, тем надежнее прокси отражает таргет.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

**Матрица согласия-несогласия знаков** — это вспомогательный инструмент. Она визуализирует совпадения / расхождения по знаку эффекта в исторических экспериментах.

| | | North Star Metric<br>Long-Term Effect | | |
|---|---|---|---|---|
| | | **negative** | **neutral** | **positive** |
| **Proxy Metric**<br>Short-Term Effect | **negative** | 34 | 203 | 3 |
| | **neutral** | 15 | 571 | 5 |
| | **positive** | 0 | 93 | 76 |

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)

# Поиск прокси-метрики

## Принятие решения

- Оценка идет сразу по всем метрикам: Binary Sensitivity, MSE и/или корреляция, Proxy Score.
- Если результаты согласуются, прокси можно рассматривать как кандидата.
- Если есть противоречия — прокси не надежна.

## Предостережения

- Убедитесь, что прокси действительно нужны.
- Сначала попробуйте улучшить дизайн эксперимента.
- Выбирайте прокси со здравым смыслом.
- Валидируйте и мониторьте прокси.

Dmitriev & Wu — Measuring Metrics (Microsoft, 2016)
Richardson, Zito, Greaves & Soriano — Pareto Optimal Proxy Metrics (2023)