AB - тестирование

Сергей Юдин, Яндекс

Что такое AB-тест

https://texterra.ru › Блог ▾ Translate this page

Что такое A/B-тестирование и как его проводить - TexTerra

Oct 17, 2019 — Дайте ему название, укажите URL базовой страницы, копии которой будут тестироваться и выберите режим «**Эксперимент** А/Б».

https://tilda.education › articles-yourf... ▾ Translate this page

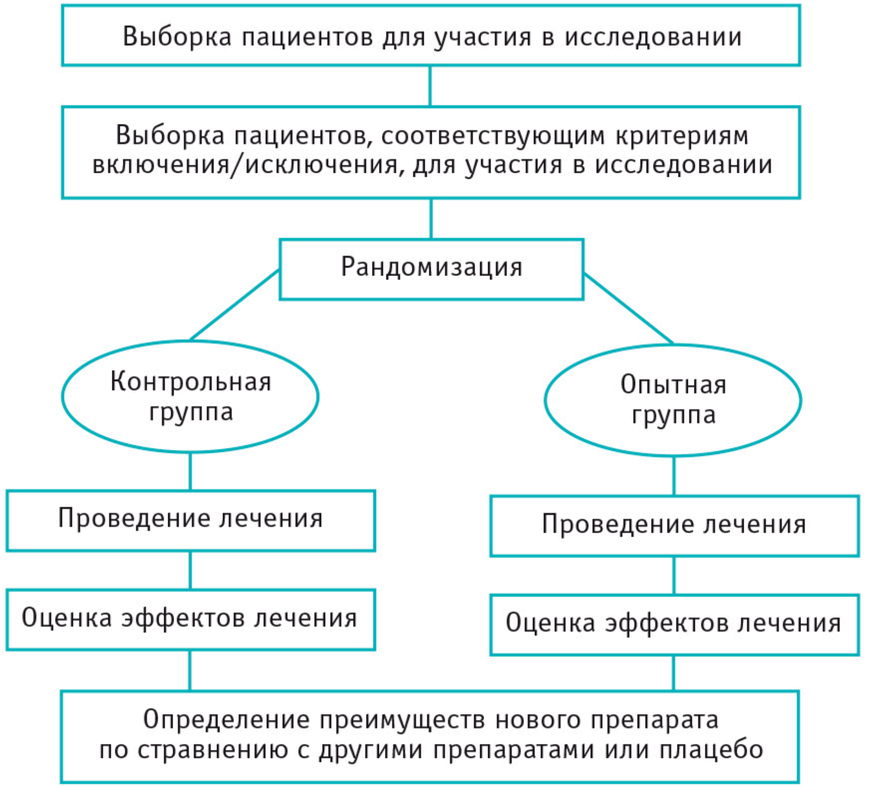
А/Б тестирование — практическое руководство

Если выбрать способ «A/B-тестирование», то в **эксперименте** будет участвовать только одна ... Источник: https://vwo.com/blog/saas-pricing-**ab**-test/.

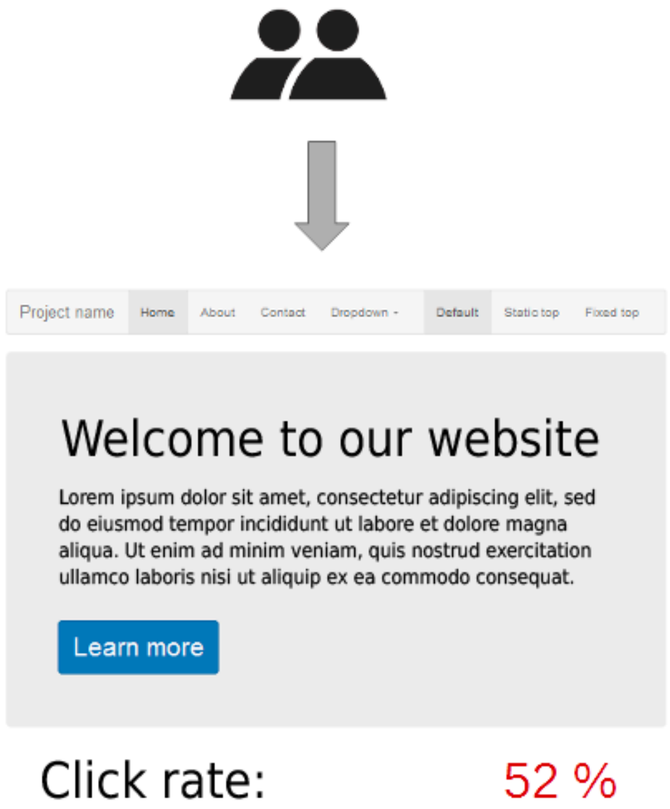
https://en.wikipedia.org/wiki/A/B_testing

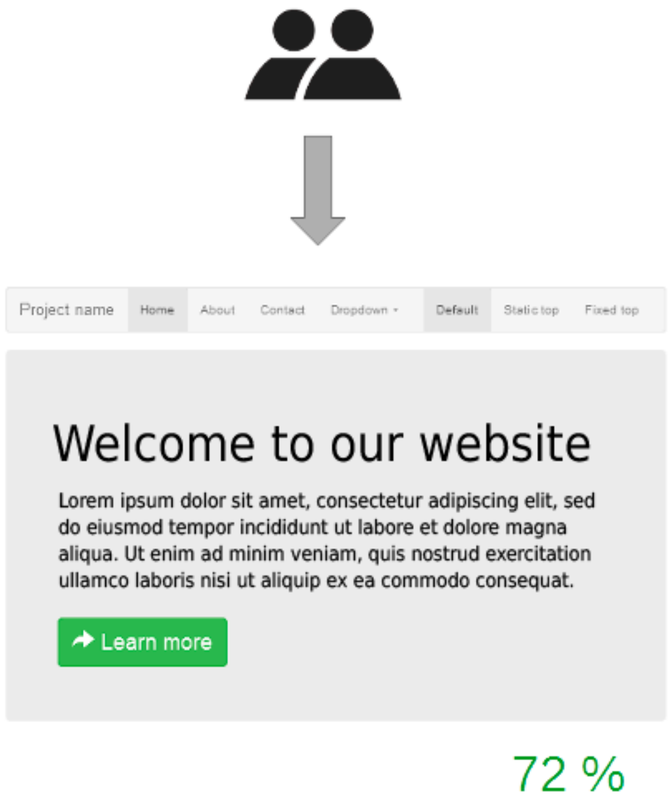
Что такое AB-тест

Двойное слепое рандомизированное плацебо-контролируемое исследование



Что такое AB-тест





https://en.wikipedia.org/wiki/A/B_testing

Что такое AB-тест

Для чего нужно?

- Оценка полезности нового внедрения
- Исследование зависимостей
- Мониторинги
- Построение целей и KPI

Из чего состоит

- Разбиение пользователей
- Конфигурация эксперимента
- Метрики
- Интерпретация результатов
- Корректность эксперимента

Про разбиение

А что сложного?

Нужно случайно всех пользователей разбить на 2 группы

Про разбиение

Вопросы:

- Что произойдет с пользователем, который вернулся через несколько дней?
- Как проводить несколько экспериментов одновременно?
- Как можно детектировать, что разбиение плохое?

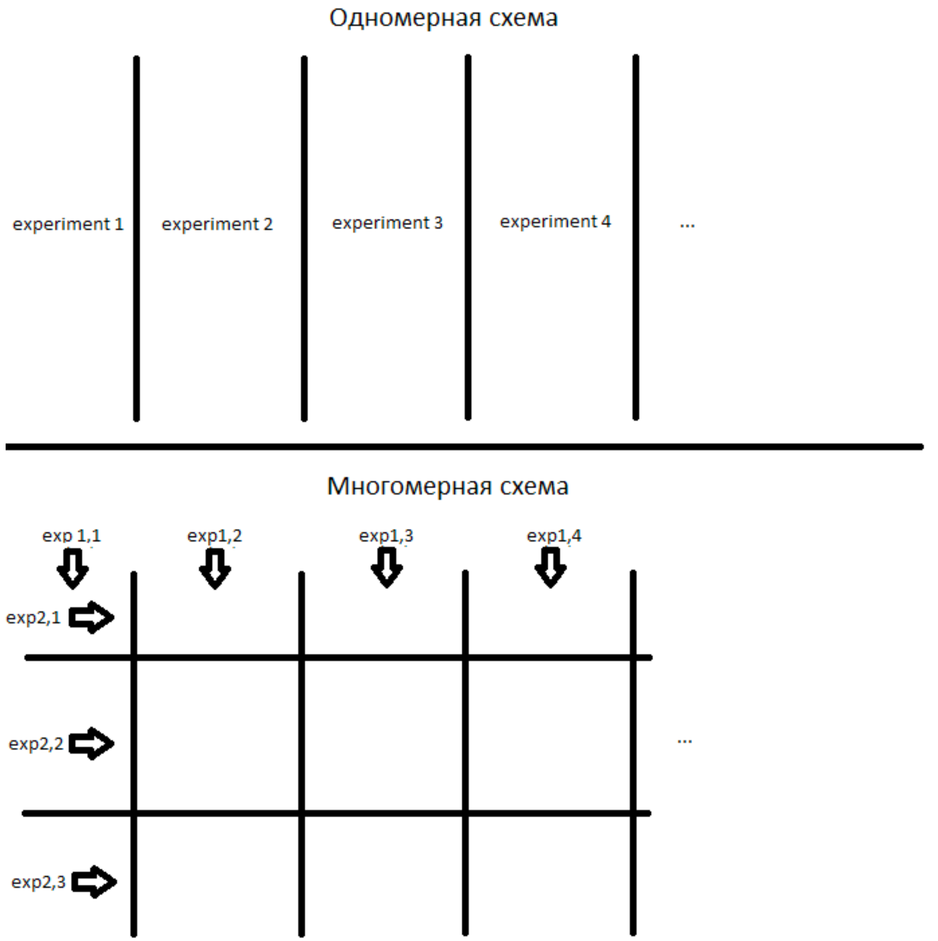
Про разбиение

Виды:

- По пользователям
- По визитам
- По действиям

Виды экспериментов

- Одномерный / многомерный
- Прямой / обратный
- Временный / вечный
- AA, AB



Конфигурация

- Размер выборок
- Длительность
- Срез

Метрики

- Чувствительность
- Шум
- Интерпретация
- Иерархия

Значимость

Эксперимент как проверка гипотезы:

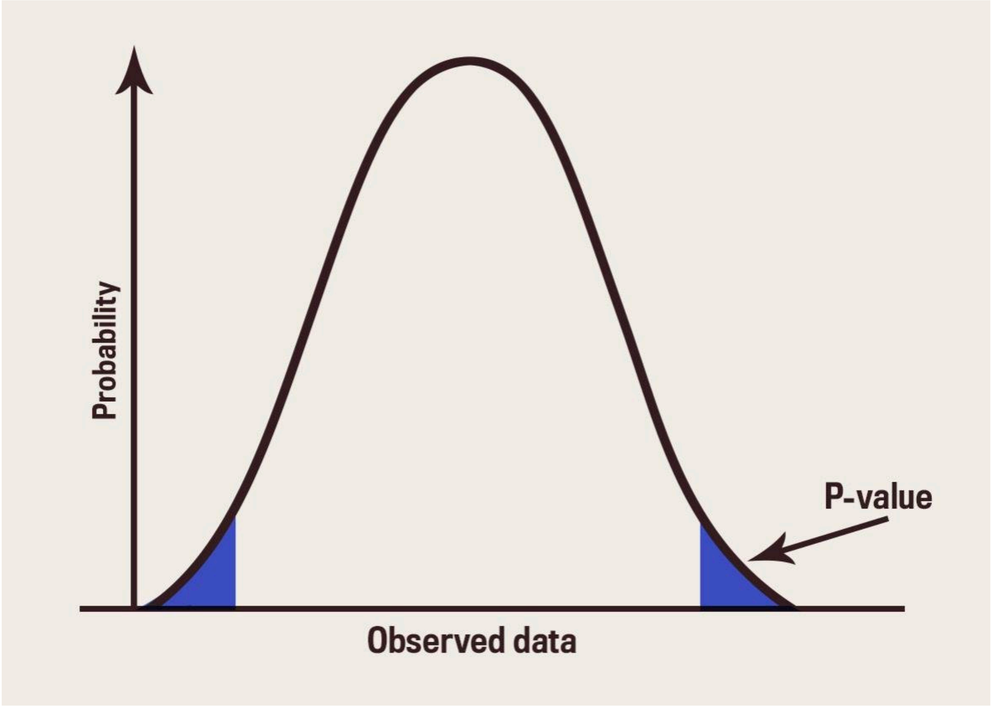
- **Нулевая гипотеза** (H_0) – утверждение о параметре генеральной совокупности (параметрах генеральных совокупностей) или распределении, которое необходимо проверить.
- **Альтернативная гипотеза** (H_A) – утверждение, противоположное нулевой гипотезе. Выдвигается, но не проверяется.

  Все гипотезы можно разделить на двусторонние (ненаправленные) и односторонние (направленные).

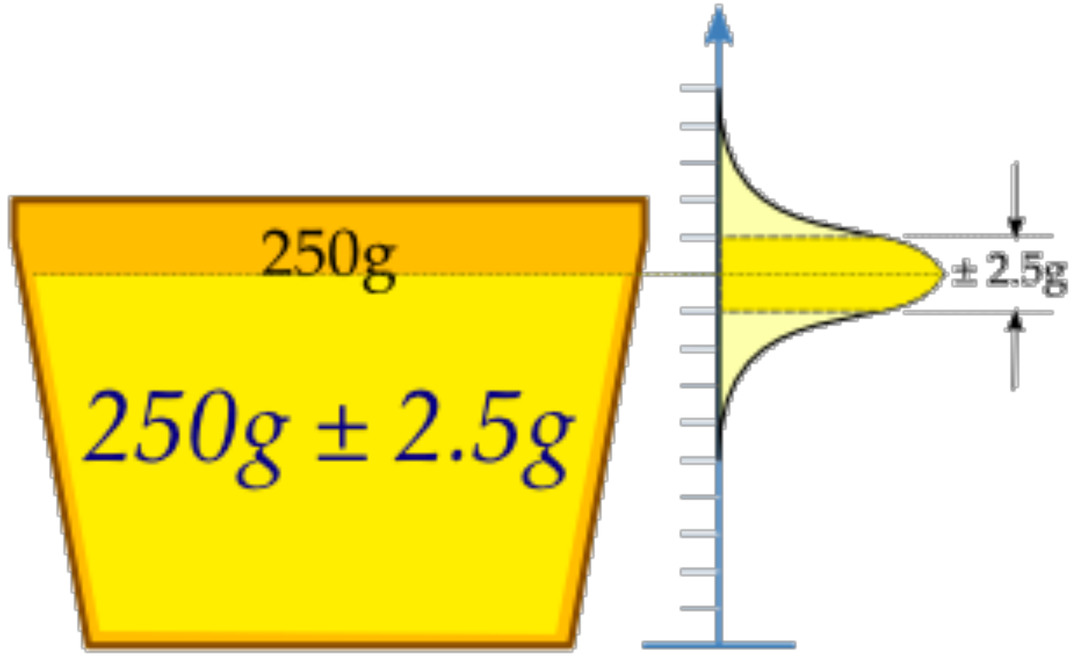
Двусторонние альтернативы	**Односторонние альтернативы**
$(H_A : p \neq 0.5)$	- левосторонние $(H_A : p < 0.5)$ - правосторонние $(H_A : p > 0.5)$

Значимость

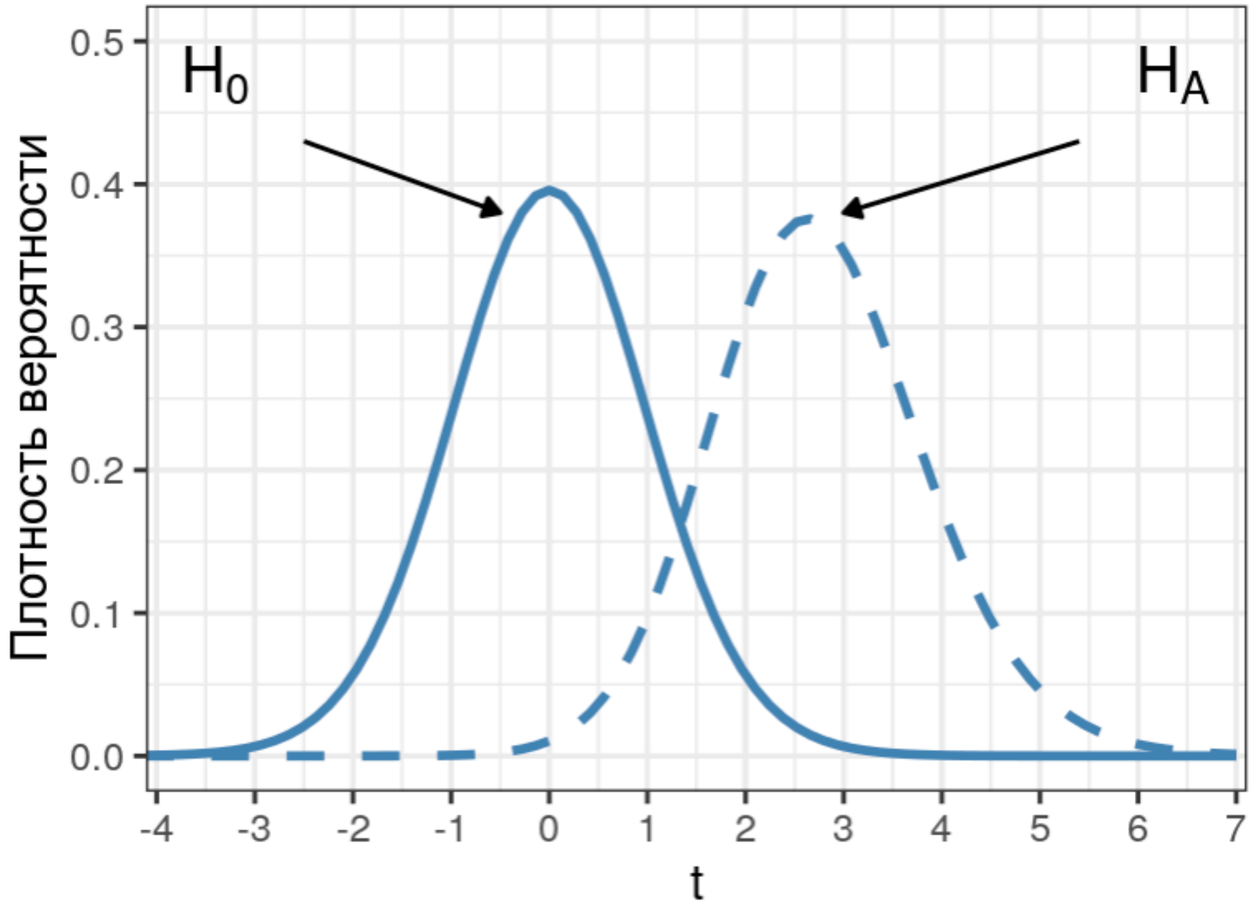
Статистический критерий — правило, которое позволяет делать вывод о том, стоит ли на основе имеющихся данных отвергать нулевую гипотезу или нет.
P-значение (P-value) — вероятность ошибки, при условии, что нулевая гипотеза верна.
Значимое изменение — изменение метрики при котором принимается альтернативная гипотеза о неравенстве средних



Доверительный интервал



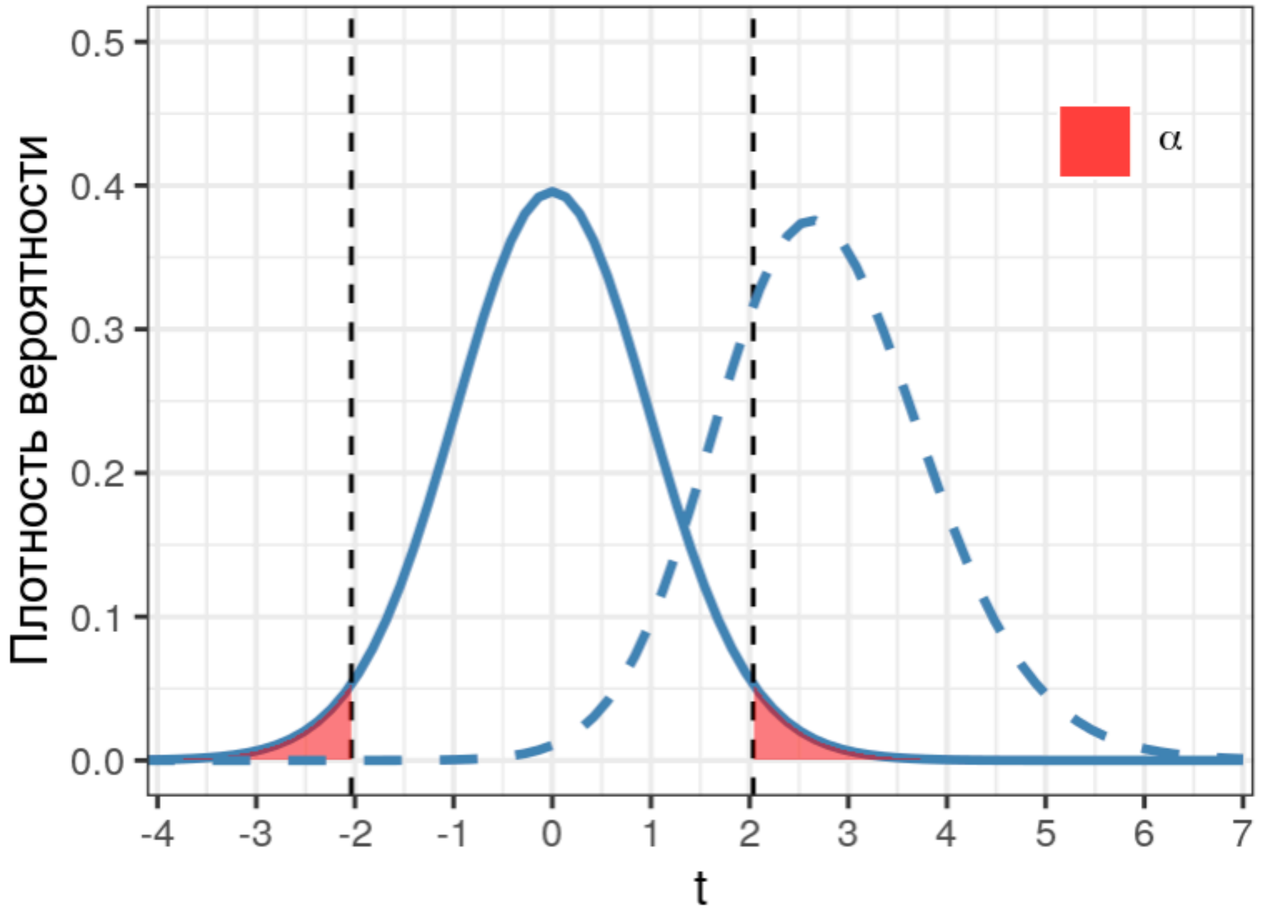
Доверительный интервал



$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

Доверительный интервал



Значимость

Критерии:

- T-test
- Mann–Whitney

Z-test / T-test

Z-test

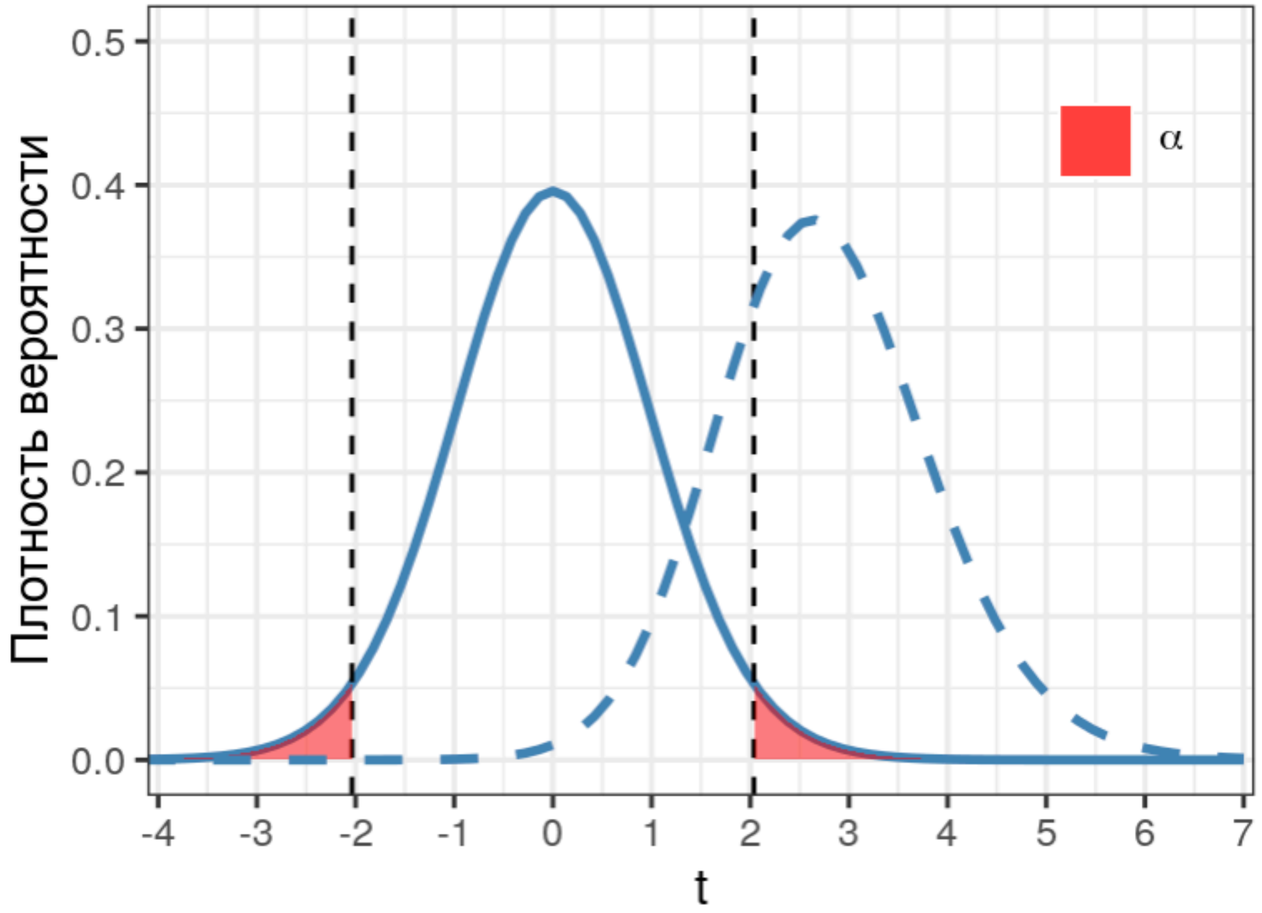
$$z_{\overline{X}} = \frac{\overline{X} - m_{H_o}}{\sigma/\sqrt{\mathrm{n}}}$$

T-test

$$t = \frac{\overline{X} - m}{s_X/\sqrt{n}}$$

$$s_X^2 = \sum_{t=1}^{n} (X_t - \overline{X})^2/(n-1)$$

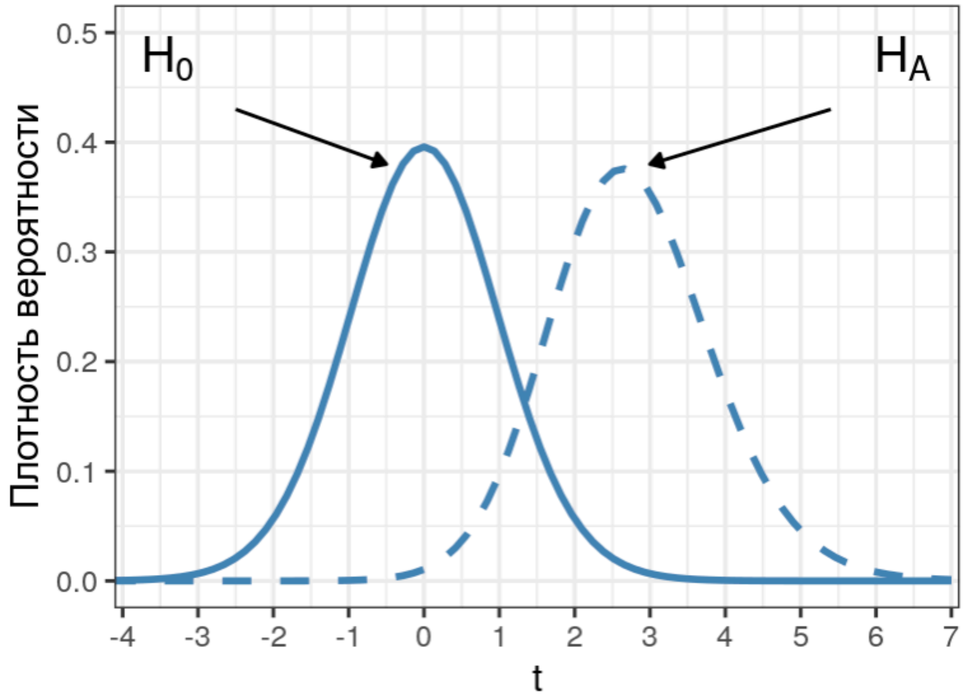
Z-test / T-test



Z-test / T-test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

$$s^2 = \frac{\sum_{t=1}^{n} (X_t - \overline{X})^2}{n - 1}$$



Z-test / T-test

Условия:

- Выборки независимы
- Либо достаточно много данных, либо нормальное распределение выборок

Mann–Whitney

X	0,288	0,782	0,894	0,812	0,850	0,532	0,207	0,408
Y	0,753	0,668	0,307	0,426	0,542	0,534	0,518	1,068

Mann–Whitney

Значение	0,522	0,569	0,577	0,584	0,681	0,729	0,775	0,860	0,870	0,876	0,925	0,953	0,957	1,056	1,094	1,098
Выборка	X	X	X	X	X	X	Y	X	Y	Y	X	Y	Y	Y	Y	Y
Ранг	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ранги X	1	2	3	4	5	6		8			11					
Ранги Y							7		9	10		12	13	14	15	16

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$$

$$U = min\{U_1, U_2\}$$

Mann–Whitney

Critical Values for the Mann-Whitney U-Test

Level of significance: 5% (P = 0.05)

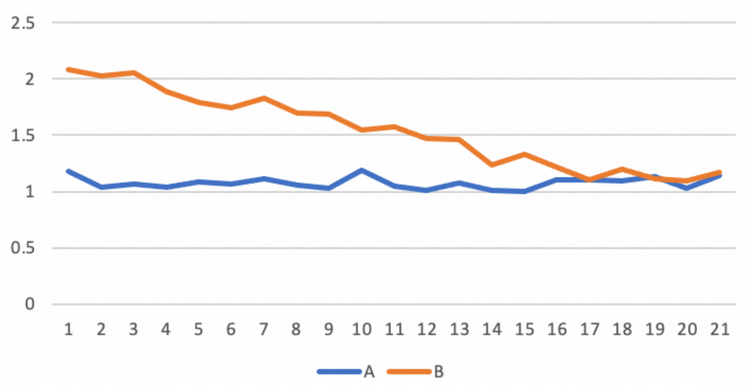
Size of the largest sample (n_2)

Size of the smallest sample (n_1)	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	13	13
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	17	18	19	20	21	22	23
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23	24	25	27	28	29	30	32	33
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30	32	33	35	37	38	40	42	43
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54
8				13	15	17	19	22	24	26	29	31	34	36	38	41	43	45	48	50	53	55	57	60	62	65
9					17	20	23	26	28	31	34	37	39	42	45	48	50	53	56	59	62	64	67	70	73	76
10						23	26	29	33	36	39	42	45	48	52	55	58	61	64	67	71	74	77	80	83	87
11							30	33	37	40	44	47	51	55	58	62	65	69	73	76	80	83	87	90	94	98
12								37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109
13									45	50	54	59	63	67	72	76	80	85	89	94	98	102	107	111	116	120
14										55	59	64	67	74	78	83	88	93	98	102	107	112	118	122	127	131
15											64	70	75	80	85	90	96	101	106	111	117	122	125	132	138	143
16												75	81	86	92	98	103	109	115	120	126	132	138	143	149	154
17													87	93	99	105	111	117	123	129	135	141	147	154	160	166
18														99	106	112	119	125	132	138	145	151	158	164	171	177
19															113	119	126	133	140	147	154	161	168	175	182	189
20																127	134	141	149	156	163	171	178	186	193	200
21																	142	150	157	165	173	181	188	196	204	212
22																		158	166	174	182	191	199	207	215	223
23																			175	183	192	200	209	218	226	235

Разбор эксперимента 1

Эксперимент: проверяли некоторое внедрение на долгом периоде в 3 недели. По результатам увидели ухудшение некоторой метрики

Гипотеза: несмотря на начальное ухудшение пользователи привыкли и можно выкатывать



Вопрос: правда ли пользователи привыкли? Как это проверить?

Разбор эксперимента 2

Эксперимент: хотим выкатить улучшение продукта, завели эксперимент на 2 недели

Гипотеза: по первым 4 дням видим улучшение метрик

Вопрос: можно ли останавливать и выкатывать?

Разбор эксперимента 3

Эксперимент: хотим выкатить формулу с некоторым порогом, завели эксперимент на 2 недели

Гипотеза: по первым 4 дням видим ухудшение метрик

Вопрос: можно ли в процессе эксперимента поменять порог для формулы?

Разбор эксперимента 4

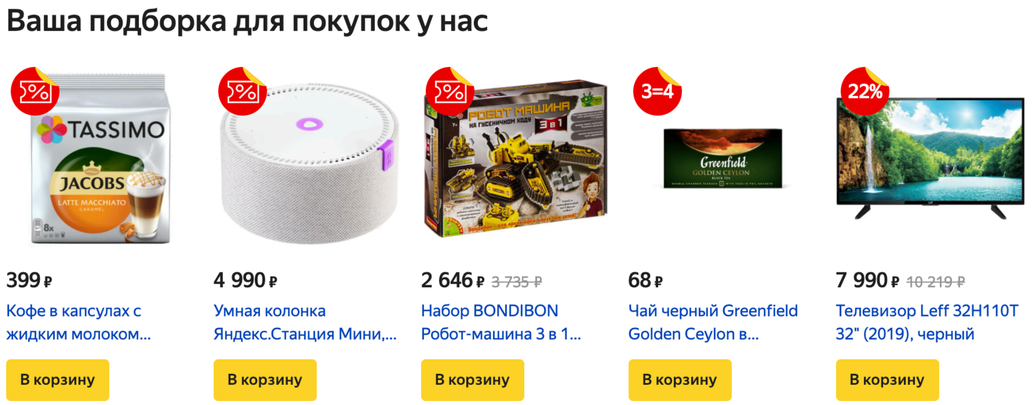
Эксперимент: разработали классное улучшение, ожидаем роста денег для продукта. Завели эксперимент на неделю, чтобы это подтвердить

Гипотеза: по итогам эксперимента деньги растут, но значимость всего 0.9

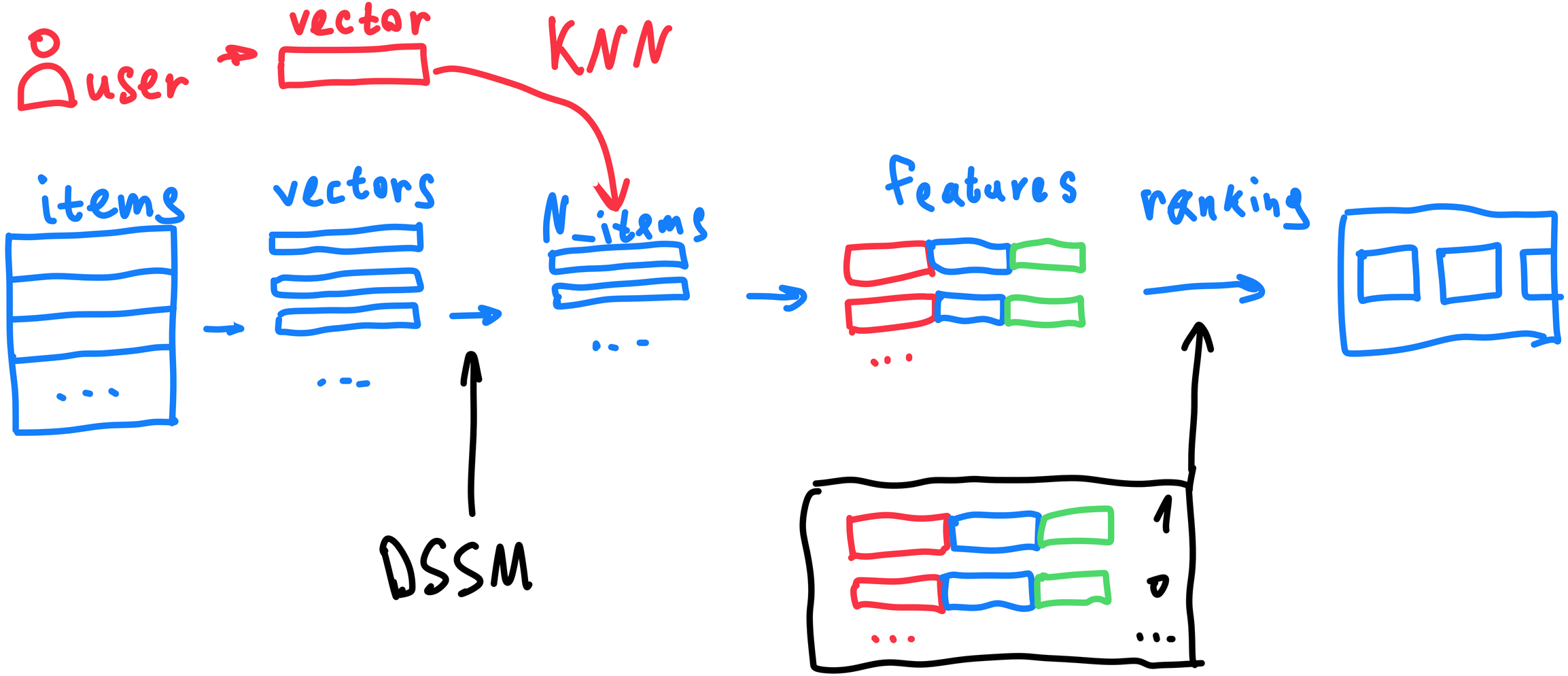
Вопрос: можно ли считать, что деньги выросли?

Аналитика ML-продукта

- Хотим внедрить новое ML-ранжирование рекомендаций товаров
- Находимся в ситуации, когда этот элемент уже есть на сайте



Архитектура рекомендаций

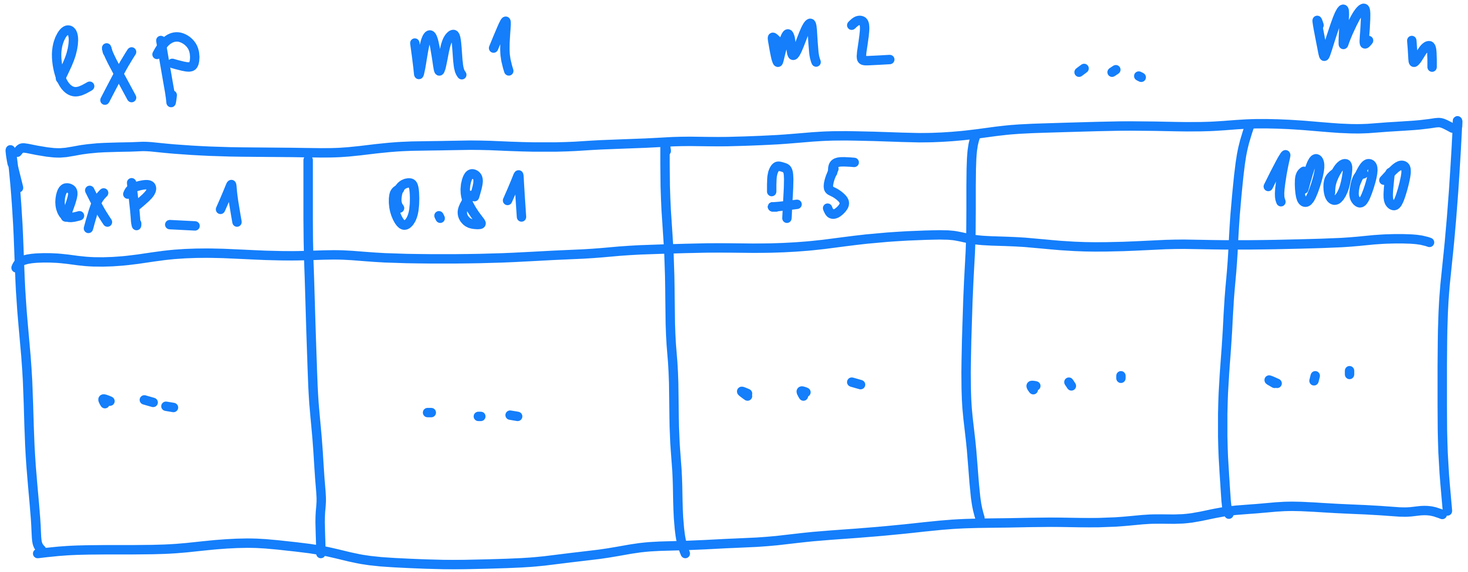


Иерархия метрик

- Выручка
- Средний чек / Число купивших пользователей
- Выручка проданных товаров, через наш элемент
- CTR элемента
- Оффлайн метрики ранжирования
- ROC AUC по валидационной выборке

Иерархия метрик

- Как проверить правильность иерархии?



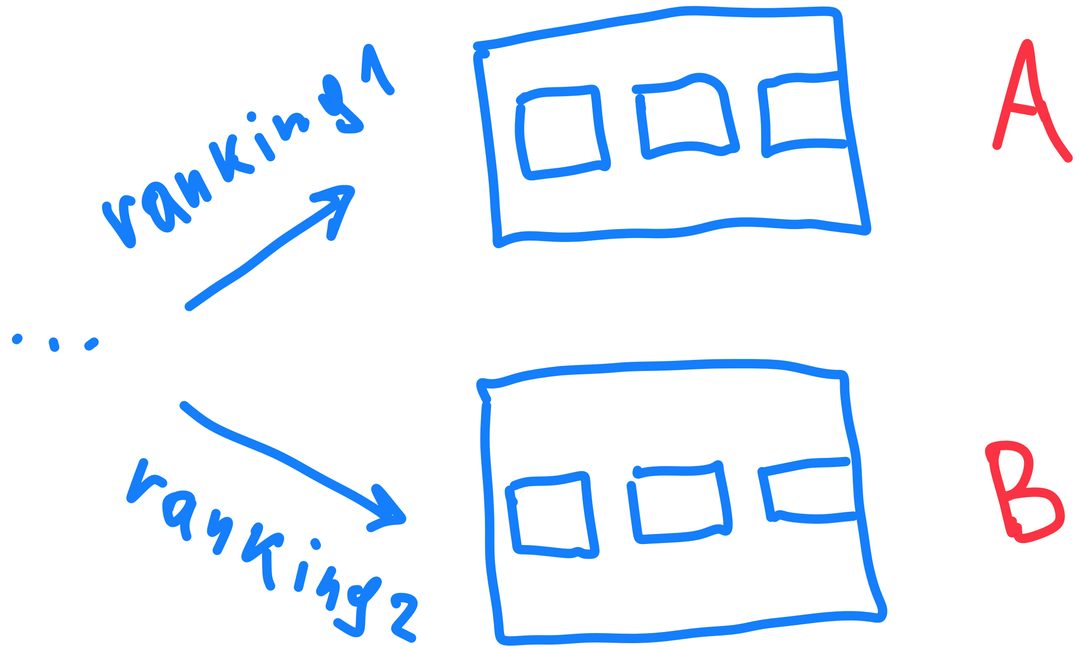
$$m_1 = \sum_{i=2}^{n} c_i m_i + C_0$$

Применение AB

- Эксперимент перед внедрением новой формулы
- Вечный эксперимент с отключением
- Вечный эксперимент со старой формулой
- Эксперимент со «случайными рекомендациями»
- Ухудшающие эксперименты

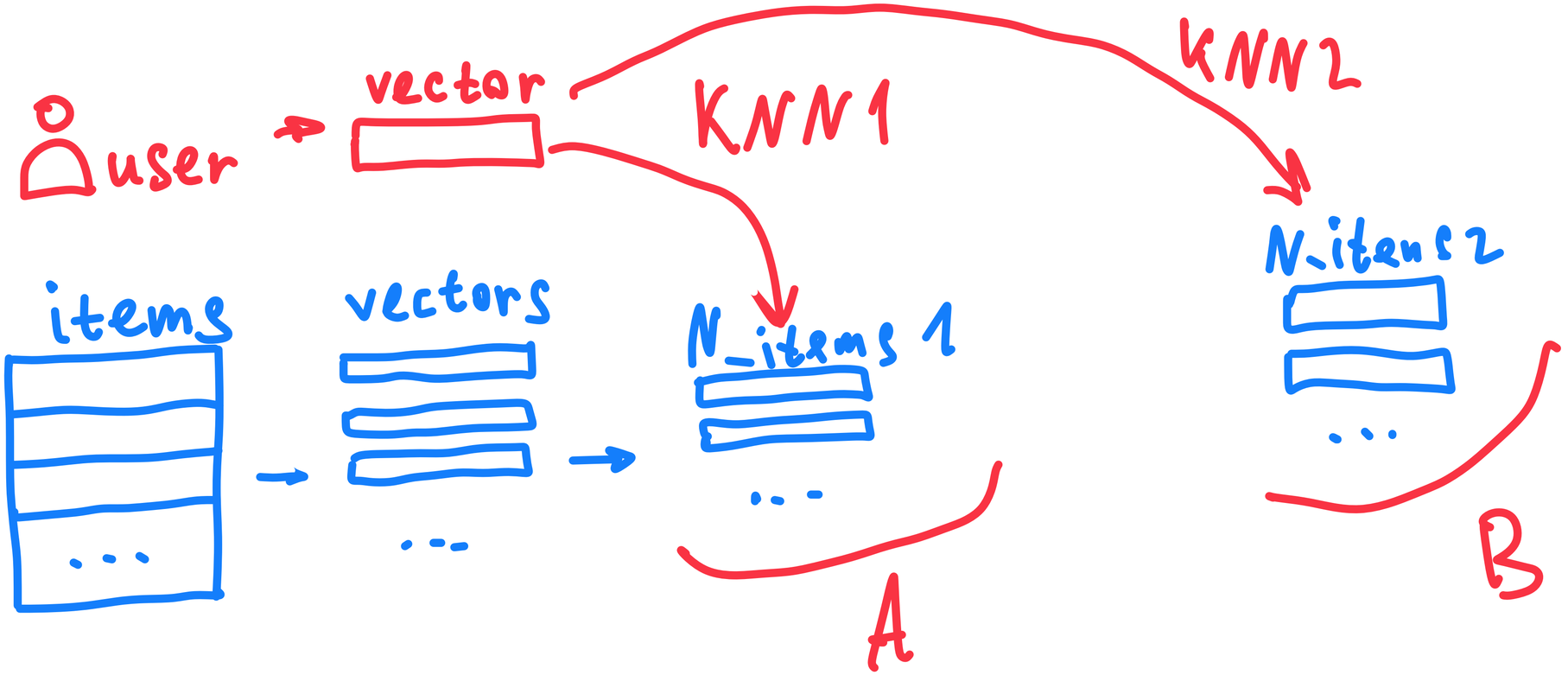
Применение AB

- Эксперимент перед внедрением новой формулы



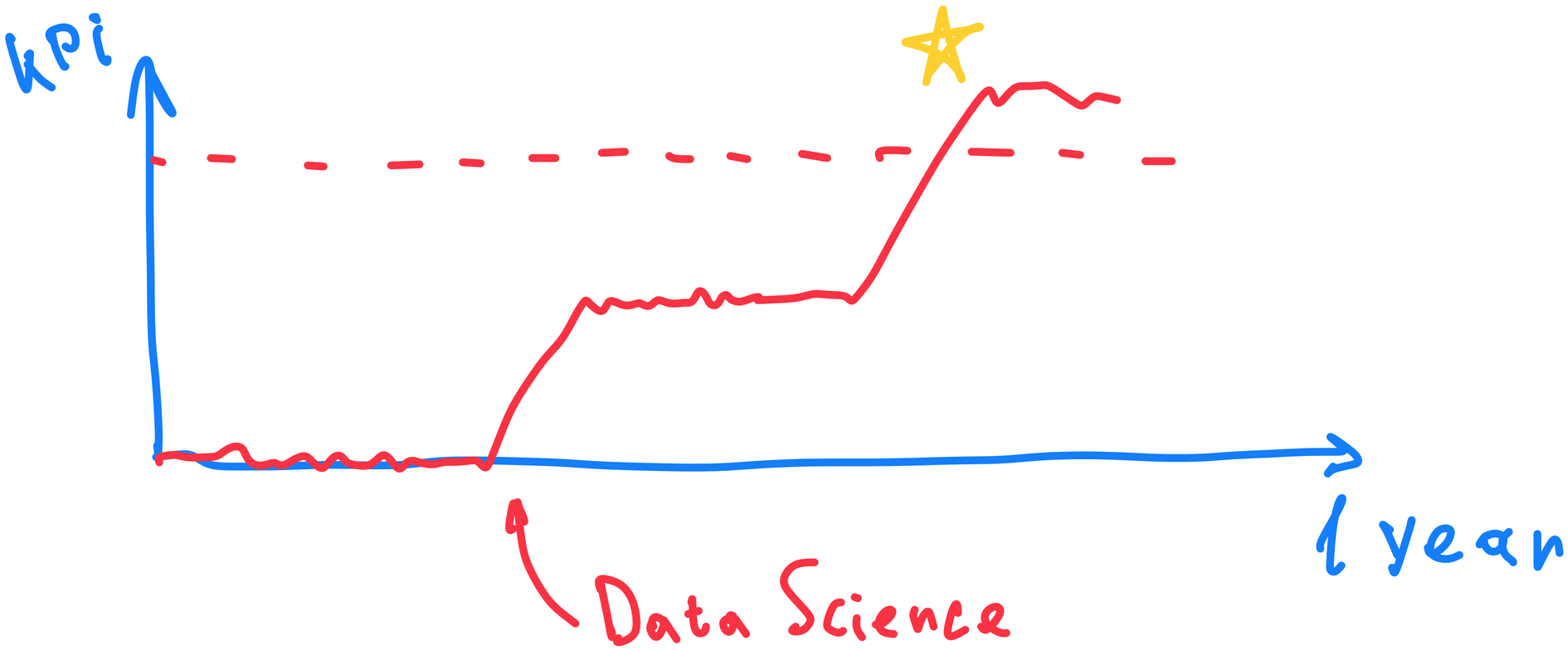
Применение AB

- Вечный эксперимент с отключением



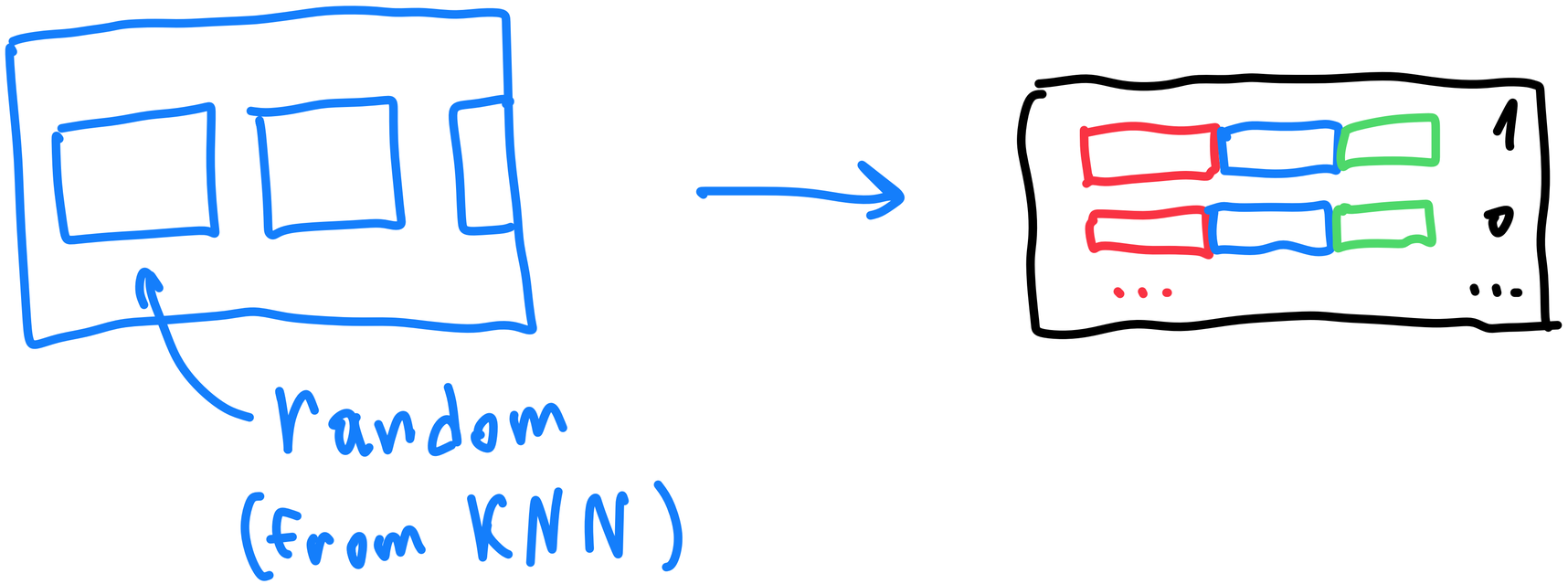
Применение AB

- Вечный эксперимент со старой формулой



Применение AB

- Эксперимент со «случайными рекомендациями»



Применение AB

- Ухудшающие эксперименты