

АБТ – часть 2

Данил Валгушев
Елена Кунакова

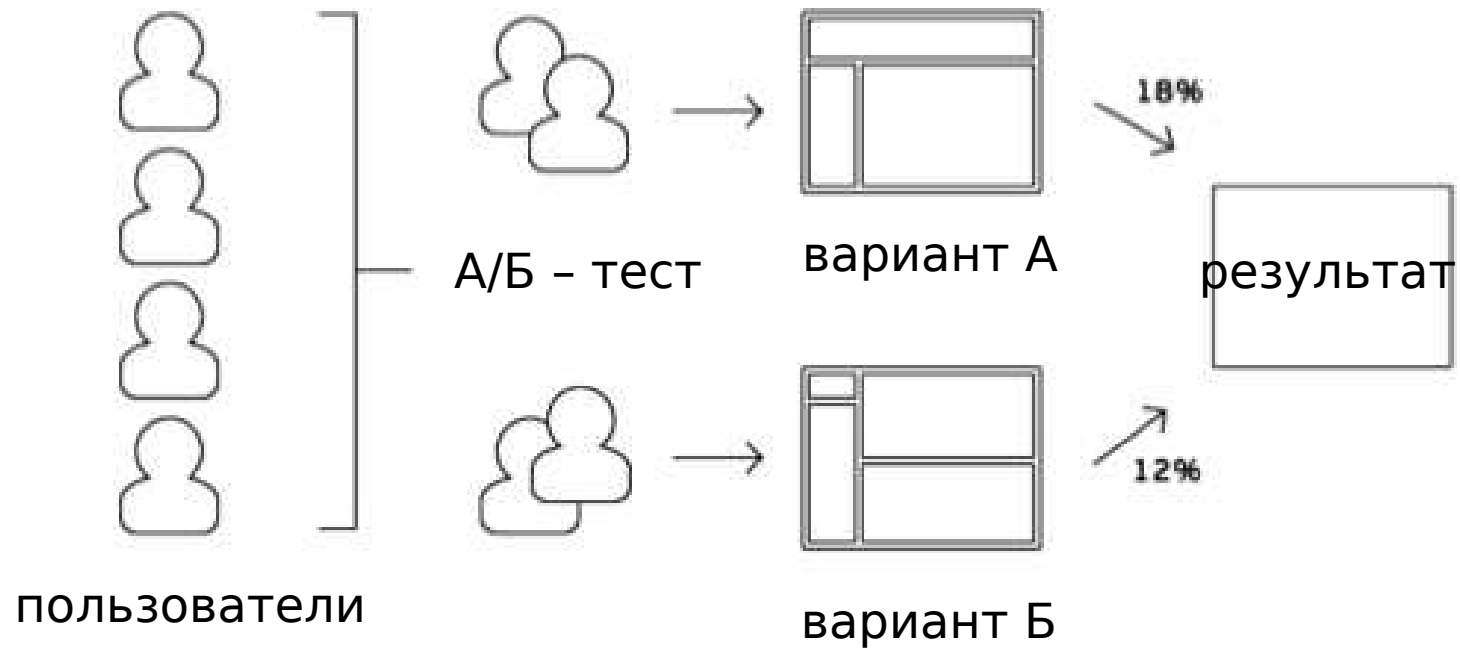
Анонс

- Обзор предыдущей лекции
 - Типы экспериментов
 - Статистические тесты и их настройка
 - Валидация метрик
-
- Эксперименты смешивания поисковых выдач
 - Обзор инструментов аналитика
 - Разбор кейсов из реальной жизни

АБТ: Repeat



АБТ: Repeat



АБТ: Repeat

- варианты разбиения

АБТ: Repeat

- варианты разбиения
- количество информации & шум

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов
- масштабы экспериментов

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов
- масштабы экспериментов
- технические сложности

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов
- масштабы экспериментов
- технические сложности
- пользовательские сложности

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов
- масштабы экспериментов
- технические сложности
- пользовательские сложности
- отладка

АБТ: Repeat

- варианты разбиения
- количество информации & шум
- типы экспериментов
- масштабы экспериментов
- технические сложности
- пользовательские сложности
- отладка
- результаты

АБТ: Типы экспериментов

- Контроль с контролем

АБТ: Типы экспериментов

- Контроль с контролем
- Заведомо плохие

АБТ: Типы экспериментов

- Контроль с контролем
- Заведомо плохие
- Обратные эксперименты

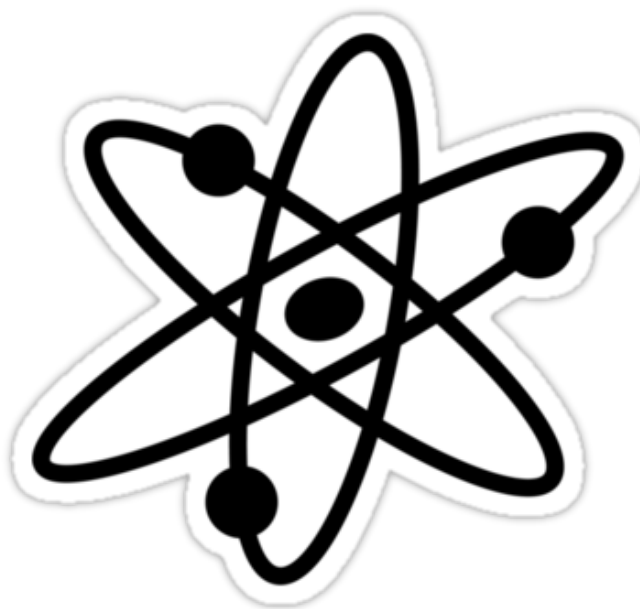
АБТ: Типы экспериментов

- Контроль с контролем
- Заведомо плохие
- Обратные эксперименты
- Технические изменения

АБТ: Типы экспериментов

- Контроль с контролем
- Заведомо плохие
- Обратные эксперименты
- Технические изменения
- Один контроль, несколько экспериментов.

АБТ: Матчасть



АБТ: Матчасть

- A & B (“контроль” и “эксперимент”)
- $M(X)$ – **случайная** величина
- $diff = M(A) - M(B)$
- $diff$ **значима? (significance)**
- насколько **достоверно? (confidence)**

АБТ: Матчасть

- $diff$ = эффект воздействия изменения + случайность
- чтобы измерить эффект, используем стат. тесты:
 - ♦ Student t-test
 - ♦ Mann-Whitney / U test / Wilcoxon
 - ♦ Wald test
 - ♦ ...
 - ♦

АБТ: Статистические тесты

- Выбираем две гипотезы H_0 и H_1
- На примере АБ-тестинга:
 - ♦ H_0 – метрики равны
 - ♦ H_1 – не равны

АБТ: Ошибки первого и второго рода

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка второго рода)
	H_1	H_0 неверно отвергнута (Ошибка первого рода)	H_0 верно отвергнута

АБТ: Ошибки первого и второго рода

- Ошибка первого рода (уровень значимости)
 - ♦ $\alpha = P(H_1 | H_0)$
- Ошибка второго рода
 - ♦ $\beta = P(H_0 | H_1)$
- Мощность критерия
 - ♦ $(1 - \beta)$

АБТ: p-value

- p-value = насколько вероятно получить наблюдаемый результат или результат “хуже” наблюдаемого, если верна H_0
- Для АБ-тестинга
 - Насколько вероятно, что разница в метриках обусловлена шумом
- Принятие решения:
 - $p\text{-value} > \alpha \rightarrow H_0$
 - $p\text{-value} \leq \alpha \rightarrow H_1$
(маловероятно, что виноват только шум)

АБТ: Постановка задачи

- Даны выборки
 - ♦ x_1, x_2, \dots, x_{n1}
 - ♦ y_1, y_2, \dots, y_{n2}
- Распределения
 - ♦ $E_x = \mu_1 \quad E_y = \mu_2$
 - ♦ $D_x = \sigma_1^2 \quad D_y = \sigma_2^2$
- Гипотезы
 - ♦ $H_0 : \mu_1 = \mu_2$
 - ♦ $H_1 : \mu_1 \neq \mu_2$

АБТ: Типы критериев

- Для одной выборки
 - ♦ Z-test
 - ♦ T-test
 - ♦ Wilcoxon
- Для двух независимых выборок
 - ♦ T-test
 - ♦ Mann-Whitney
- Для двух зависимых выборок
 - ♦ В контексте АБТ неинтересны

АБТ: T-test

- Требования
 - ♦ x и y из нормального распределения
 - ♦ Внутри каждой из выборок все элементы независимы
 - ♦ Выборки независимы друг от друга
 - ♦ $\sigma_1^2 = \sigma_2^2$ (дисперсии равны)

АБТ: T-test

$$\bar{X}_1 = \frac{\sum X_1}{n_1} \quad \tilde{s}_1^2 = \frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1}}{n_1 - 1}$$

$$\bar{X}_2 = \frac{\sum X_2}{n_2} \quad \tilde{s}_2^2 = \frac{\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}}{n_2 - 1}$$

n1 = n2

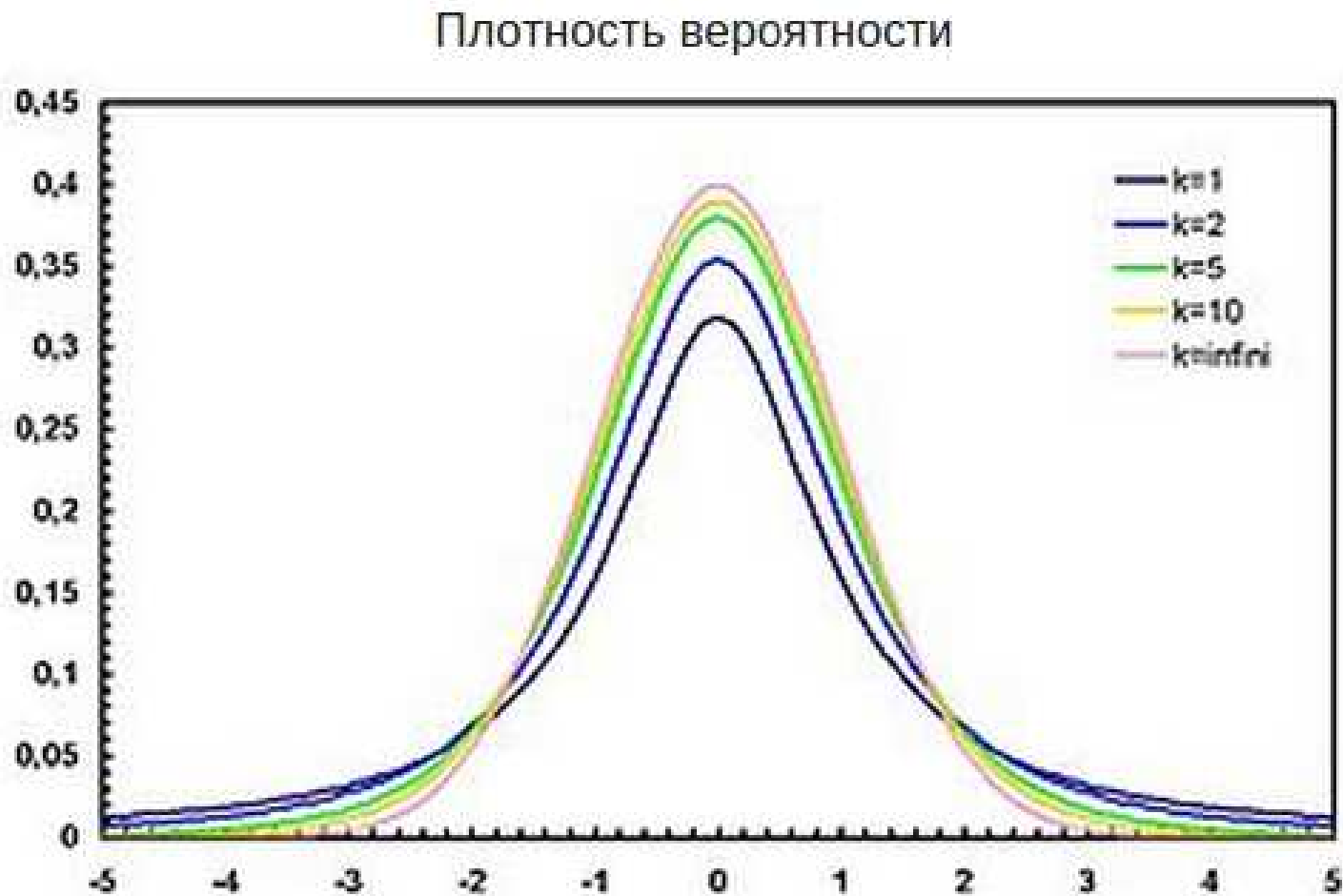
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\tilde{s}_1^2}{n_1} + \frac{\tilde{s}_2^2}{n_2}}}$$

n1 ≠ n2

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)\tilde{s}_1^2 + (n_2 - 1)\tilde{s}_2^2}{n_1 + n_2 - 2} \right] \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

АБТ: T-test

- Распределение Стьюдента

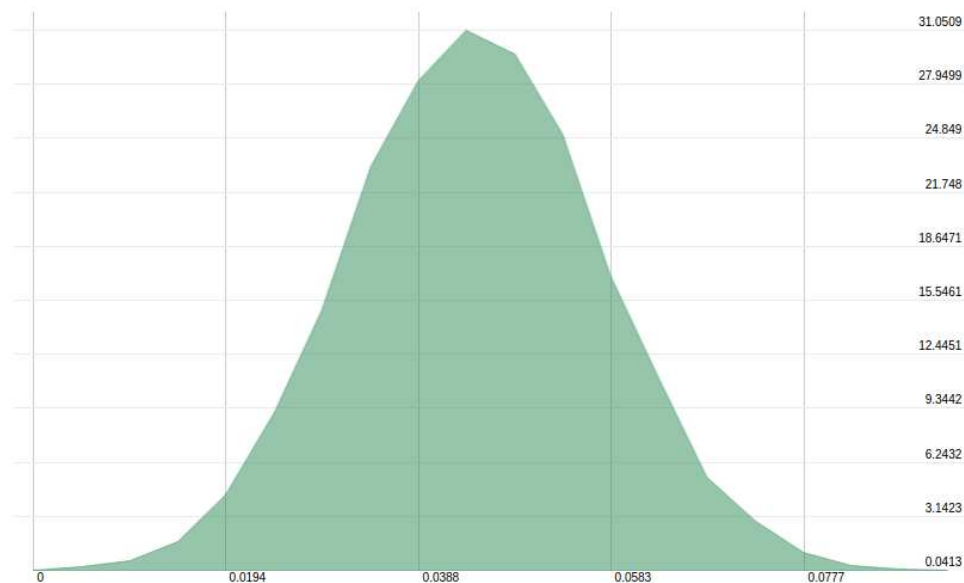


АБТ: Mann–Whitney U Test

- Требования
 - ♦ x и y из одинакового распределения (с точностью до сдвига)
 - ♦ Внутри каждой из выборок все элементы независимы
 - ♦ Выборки независимы друг от друга
- Следствие
 - ♦ $\sigma_1^2 = \sigma_2^2$ (дисперсии равны)

АБТ: Bootstrap

- Не делает предположений о форме распределения
- Простая идея, много вычислений
- Позволяет оценить распределение некоторой функции от случайной выборки



АБТ: Статистические тесты

- Как откалибровать
 - ♦ Найти баланс между α и β (ошибками первого и второго рода)
 - ♦ На контрольных должны принимать H_0
 - ♦ На заведомо плохих – H_1
- Проверяем правильность срабатывания для каждой метрики

АБТ: Статистические тесты

- Литература
 - *David J. Sheskin* “Handbook of parametric and nonparametric statistical procedures”

АБТ: Валидация метрик



АБТ: Валидация метрик


Хотим измерять качество поиска с помощью метрик.

Какие метрики говорят о качестве?

КОТИКИ

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ

Только **Котики** - YouTube
[youtube.com > user/onlykotiki](https://youtube.com/user/onlykotiki)
Только **Котики**. ПодписатьсяПодписка оформленаОтменить подписку. ... Пользователь
Только **Котики** добавил видео 4 года назад. 1:51. Следующее.

Коты. Котики. Котята. — ВКонтакте
[vk.com > kotomanechka](https://vk.com/kotomanechka)
 Перейдите на страницу пользователя, чтобы посмотреть публикации или отправить сообщение.
О себе: Котомания

котики — 5 млн видео
[Яндекс.Видео > котики](https://yandex.ru/video/search/?text=котики)

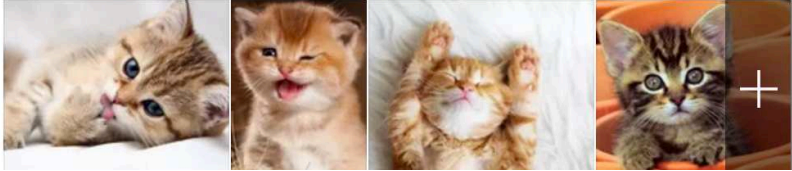
Котики
myvi.ru *

Приколы **Котики** и Кошечки Смотреть РЖАЧ Подборка...
youtube.com

Милые **котики**)
myvi.ru *

смешные **котики**
youtube.com

котики — смотрите картинки
[yandex.ru/images > котики](https://yandex.ru/images/search/?text=котики)



АБТ: Валидация метрик

Клики

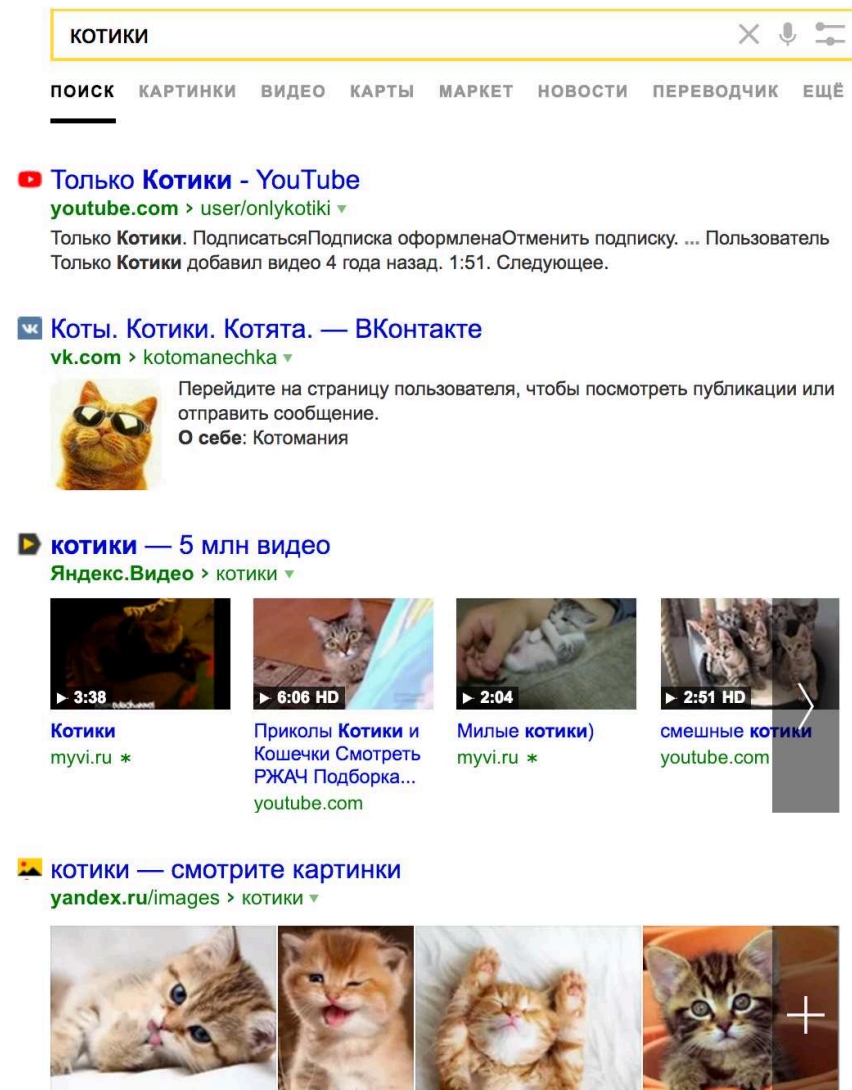
- общее количество кликов
- кликов на запрос
- кликов на документ
- доля некликнутых

Dwell time клика:

- длинные клики ($dw > 30c$)
- короткие клики ($dw < 15c$)

Utility vs Effort

- длинные и короткие клики
- длинных кликов на запрос
- доля длинных кликов



АБТ: Валидация метрик

Мадонна
Найти


поиск
КАРТИНКИ
ВИДЕО
КАРТЫ
МАРКЕТ
НОВОСТИ
ПЕРЕВОДЧИК
ЕЩЁ

▼ **Мадонна (певица) — Википедия**
[ru.wikipedia.org > Мадонна \(певица\)](#) ▼
Мадо́нна Луиза Чикко́не (англ. **Madonna Louise Ciccone**, род. 16 августа 1958, Бей-Сити, Мичиган, США) — американская певица, автор-исполнитель, музыкант…

🔥 **Слушайте бесплатно — мадонна**
[music.yandex.ru > Madonna](#)
 Все песни на Яндекс.Музыке: «Music», «You'll See», «Like A Virgin» и другие.

📻 **Мадонна: 6873 песни скачать бесплатно в mp3 и слушать...**
[zf.fm > Madonna](#) ▼
 Будущая королева поп-музыки — **Мадонна Луиза Чиконе** родилась в семье франко-канадки и итало-американца 16 августа 1958года.

📰 **Мадонна — новости**



Мадонна устроила джем-сейшн со своими п...
[gordona.uom.com](#) 7 ноя 2017
 Мадонна: Джем-сейшн с младшими Чиконе Фото: madonna / Instagram.




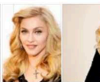





Мадонна, Боно и Кира Найтли засветились в "райском..."
[24tv.ua](#) 6 ноя 2017

Среди фигурантов Paradise Papers оказались Мадонна и сол...
[news.liga.net](#) 6 ноя 2017

📖 **Мадонна (Madonna) биография, фото — новости о певице...**
[uznayvse.ru > znamenitosti/biografiya-madonna.html](#) ▼
Мадонна (Madonna). Настоящее имя: Луиза Вероника Чикконе. День рождения:

Мадонна


Певица

Американская поп-певица, автор песен, музыкант, танцовщица и музыкальный продюсер. Переехав в Нью-Йорк в 1978 году ради карьеры в танцевальной труппе, Мадонна сначала стала участницей рок-групп, а потом успешной сольной исполнительницей и автором песен. [Википедия](#)

Родилась: 16 августа 1958 г. (59 лет), Бей-Сити, Мичиган, США
В браке с: [Гай Ричи](#) (2000-2008 гг.), [Шон Пенн](#) (1985-1989 гг.)
Партнёры: Карлос Леон (но 1998 г.), [Брахим Заибат](#)
Дети: Лопа Леон
Родители: [Мадонна Луиза Чикконе](#), [Сиэльмо Чикконе](#)
Награды: Премия «Грэмми» за лучшую песню к кинофильму, телевизионному или другому визуальному представлению, [...]»

Популярные треки





Music Album Version

Современные поисковые системы позволяют получать ответы прямо на серпе

сколько калорий в банане

✕





поиск

картинки

видео

карты

маркет

новости

переводчик

ещё

Отображение куки

дата крещения руси

✕ 🎤 🔗

поиск

картинки

видео

карты

маркет

новости

переводчик

ещё

Банан > Пищевая ценность

На 100 г продукта

▼

96 ккал

Белки


Жиры

Углеводы

1,5 г

0,5 г

21 г



АБТ: Валидация метрик

Что такое «задача пользователя» ?

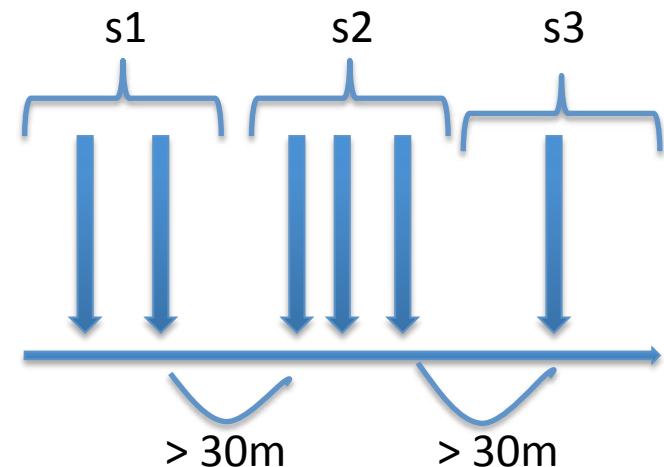
- 1 запрос != 1 задача
- запросы → сессии

Способы деления на сессии

- по 30-минутному интервалу
- по схожести текстов запросов
- по интервалам отсутствия с учетом пользовательских действий
- ...

Сессионные метрики

- длина сессии
- запросов на сессию, кликов на сессию
- среднее время между сессиями



АБТ: Валидация метрик

Разные сценарии поиска

- поиск нужного сайта (ранжирование)
- поиск ответа (факты, карточки, врезки)
- поиск «интересного» (карусельки, рекомендации)

Вместе с «фильмы про ученых» ищут:

фильмы про научные эксперименты [ivi.ru](#)

игры разума фильм 2002

вселенная стивена хокинга фильм 2015

проблеск гениальности фильм 2008

фильмы про гениев

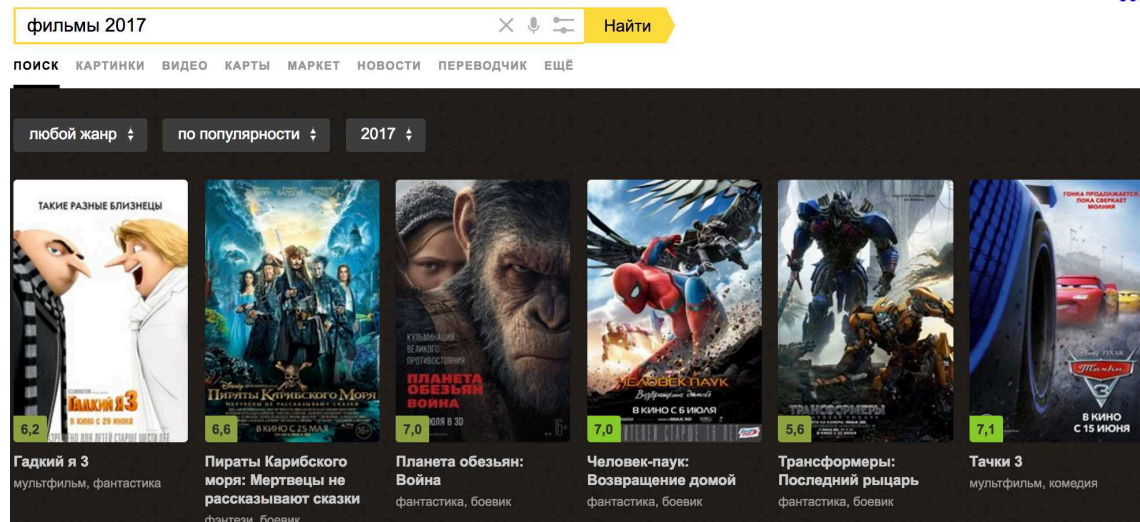
девять дней одного года фильм 1962

заражение вирус смерти фильм 2006

машина времени фильм 2002

дьявольский доктор z фильм 1966

письма мертвого человека фильм 1986



* не всегда сессия короткая

* не всегда клики длинные

АБТ: Валидация метрик

Высокоуровневые метрики

- SpU (sessions per user)
- AT (absence time)
- search engine switch
- деньги

Высокоуровневые метрики

- хороши в качестве kpi
 - не всегда их стат значимое изменение в эксперименте
- говорит о качестве сервиса

АБТ: Валидация метрик

Как сделать новую метрику

- согласованность
- проверка на ухудшающем изменении
- проверка в обратном эксперименте
- АА-тесты
- множественное тестирование

АБТ: Валидация метрик

Пусть тестируемое изменение затрагивает 1% от всех визитов.

И каждый пользователь становится на 1% счастливее.

Тогда метрика должна заметить 0.01% изменений.

АБТ: Валидация метрик

— чувствительность метрик

АБТ: Валидация метрик

- чувствительность метрик
- улучшение чувствительности

АБТ: Валидация метрик

- система растёт, метрик становится все больше – нужна структура

АБТ: Валидация метрик

- система растет, метрик становится все больше – нужна структура
- иерархия метрик
- приемочные
 - влияние на решения
 - полное доверие

АБТ: Валидация метрик

- система растет, метрик становится все больше – нужна структура
- иерархия метрик
- приемочные
- метрики контроля
 - контроль валидности эксперимента
 - являются блокерами

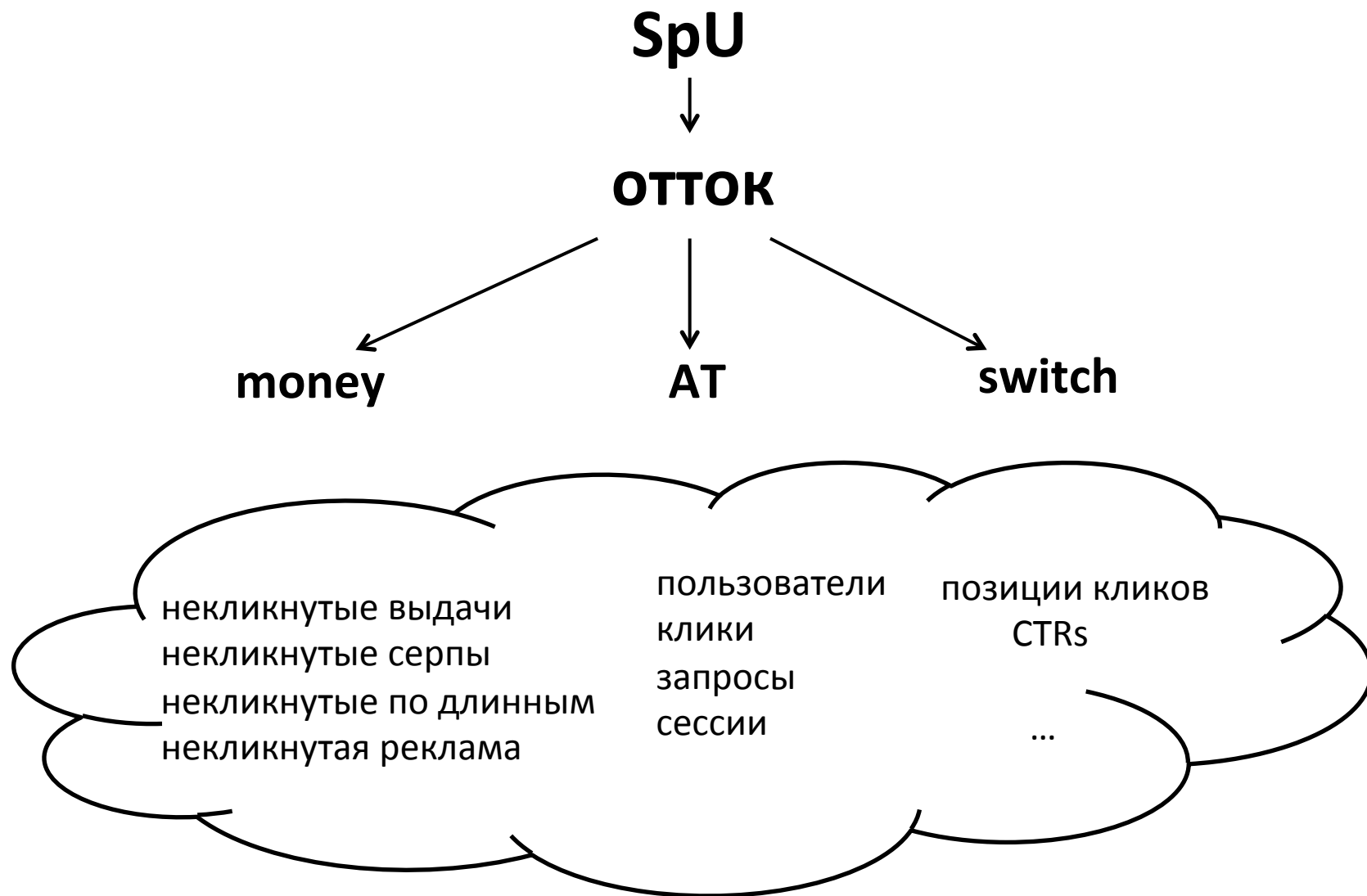
АБТ: Валидация метрик

- система растет, метрик становится все больше – нужна структура
- иерархия метрик
- приемочные
- метрики контроля
- информационные
 - свойства различных объектов
 - межобъектное взаимодействие
 - отладка

АБТ: Валидация метрик



АБТ: Валидация метрик



АБТ: Инструменты



АБТ: Инструменты

— простой интерфейс

АБТ: Инструменты

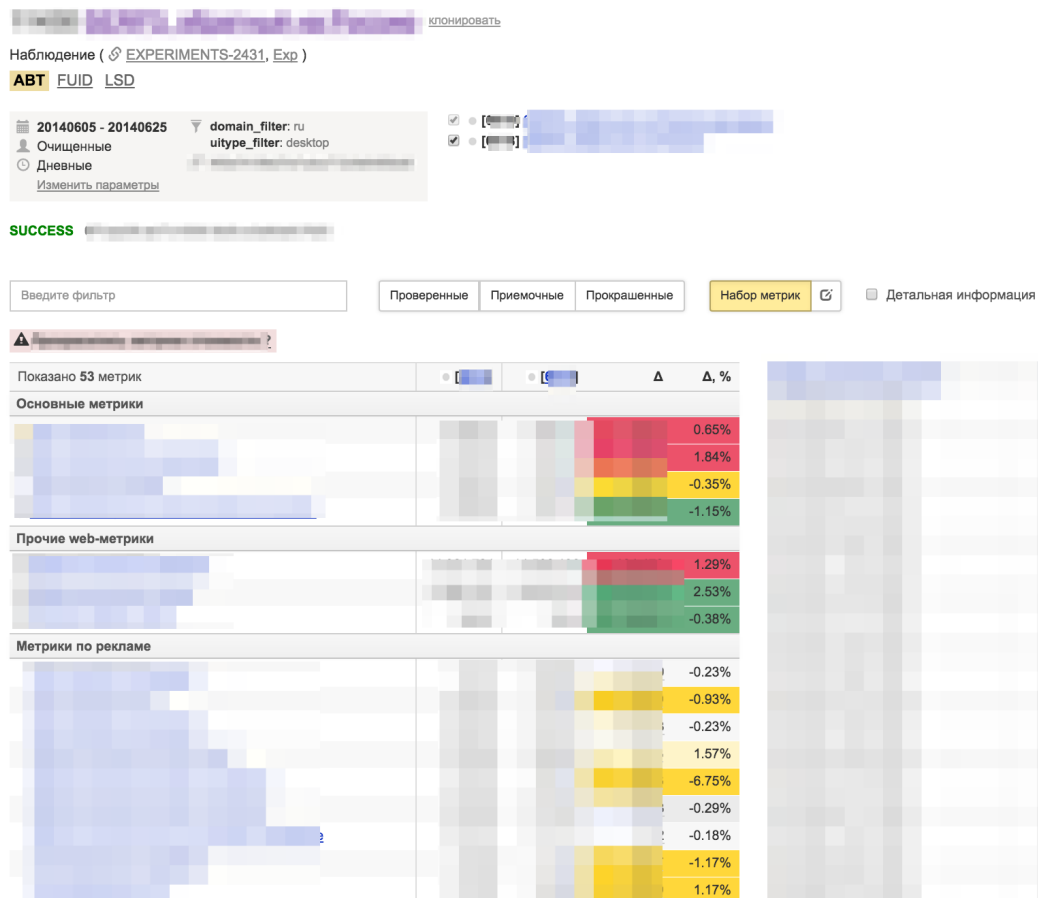
- простой интерфейс
- расчет метрик и стат. критериев «из коробки»

АБТ: Инструменты

- простой интерфейс
- расчет метрик и стат. критериев «из коробки»
- развитие и поддержка

АБТ: Инструменты

— вьюер экспериментов



АБТ: Инструменты

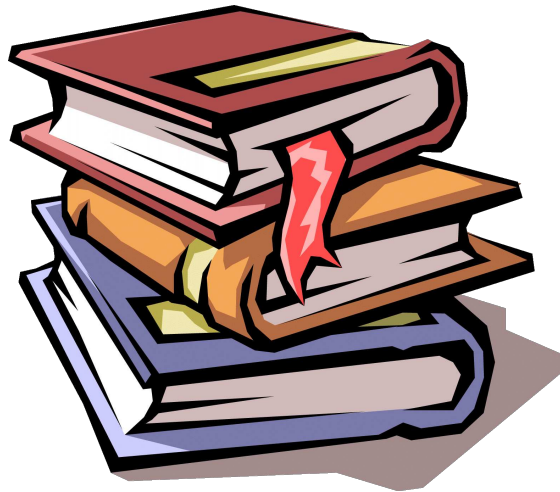
- вьюер экспериментов
- вьюер срезов

✖		выбрать метрики					
✖ ✖		+					
BCE	✖	71.5017 >	70.9688	7 710 095 >	7 702 389	8 307 646 <	8 403 672
	✖	65.3178 >	65.1422	47 669 >	47 656	34 176 <	34 810
	✖	70.1254 >	69.8048	746 463 >	744 413	738 057 <	743 944
	✖	77.9007 <	78.7728	41 527 <	43 869	29 132 <	30 411
	✖	62.3041 >	61.8986	2 910 071 <	2 913 312	3 463 323 <	3 501 703
	✖	88.0167 >	86.5654	918 628 <	923 710	739 330 <	762 515
не	✖	69.8212 >	69.3465	6 791 467 >	6 778 679	7 568 316 <	7 641 157
	✖	81.3009 <	82.9178	29 109 <	30 750	22 249 <	23 140
	✖	85.9098 <	86.5043	43 078 >	41 973	35 736 >	35 207
	✖	86.3133 <	87.6213	35 833 >	35 566	27 877 >	27 545
	✖	74.7101 >	74.2445	139 907 >	139 008	98 219 <	99 772

АБТ: Инструменты

- вьюер экспериментов
- вьюер срезов
- стенд метрик
 - подбор экспериментов по условиям
 - расчет новых метрик на старых экспериментах
 - валидация и калибровка новых метрик

АБТ: Stories



АБТ: STORIES

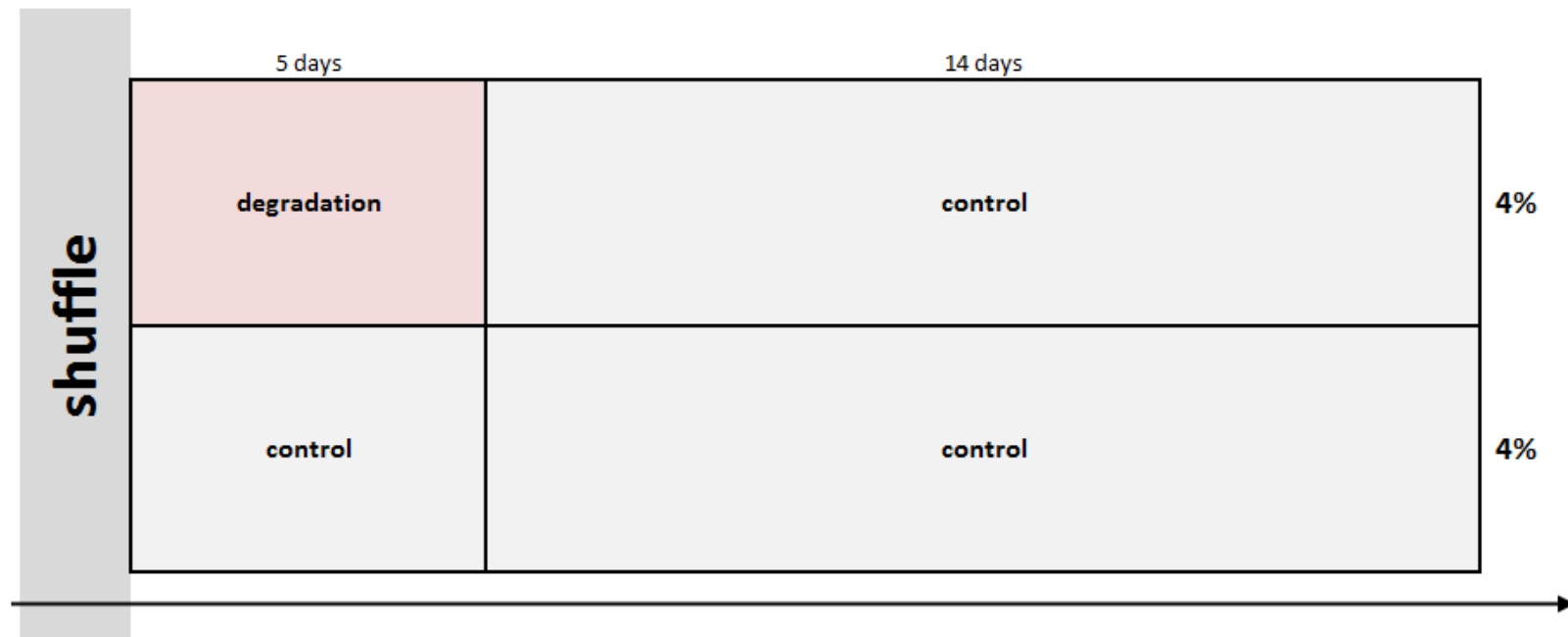
Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий

АБТ: STORIES

Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий
- 5 дней ухудшение + 14 дней отдыха



АБТ: STORIES

Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий
- 5 дней ухудшение + 14 дней отдыха
- последующий АА-тест **значимо
прокрасился**

АБТ: STORIES

Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий
- 5 дней ухудшение + 14 дней отдыха
- последующий АА-тест **значимо прокрасился**
- баннерная слепота?

АБТ: STORIES

Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий
- 5 дней ухудшение + 14 дней отдыха
- последующий АА-тест **значимо прокрасился**
- медленное восстановление

АБТ: STORIES

Фильтры:			99.0 99.5 99.9		
Эксперименты: Контроль 4% в эксклюзиве с формулой 7656, контроль - проверка влияния экспериментов			99.0 99.5 99.9		
99.0 99.5 99.9					
Основные метрики			WaldTest	Дельта	Дельта, %
1	28.4222	28.4134	0.76	-0.0088 ± 0.0557	-0.03%
	32.6993	32.6942	0.14	-0.0051 ± 0.0576	-0.02%
	77.8955	77.7454	99.15	-0.1501 ± 0.0488	-0.19%
	23.8542	23.7952	37.80	-0.0589 ± 0.0571	-0.25%
1	4.934.463	4.927.479	16.27	-6.984 ± 12519	-0.14%
	6.115.628	6.128.515	30.24	12.887 ± 16148	0.21%
	1.8799	1.8911	100.00	0.0112 ± 0.0019	0.60%
	2.0494	2.0631	100.00	0.0137 ± 0.0021	0.67%
1	49.6232	49.5229	64.37	-0.1004 ± 0.0732	-0.20%
	20.1074	20.2173	97.26	0.1099 ± 0.0376	0.55%
	12.5369	12.6888	100.00	0.1519 ± 0.0277	1.21%
	9.0194	9.1428	100.00	0.1234 ± 0.0246	1.37%
1	6.8087	6.9353	99.94	0.1267 ± 0.0300	1.86%
	5.1802	5.2798	100.00	0.0996 ± 0.0176	1.92%
	4.1296	4.2086	100.00	0.0791 ± 0.0175	1.91%
	3.4227	3.4824	99.97	0.0597 ± 0.0140	1.74%
1	2.9984	3.0452	99.21	0.0468 ± 0.0146	1.56%
	2.8472	2.9057	99.98	0.0585 ± 0.0139	2.05%
	50.9736	50.6439	99.99	-0.3296 ± 0.0795	-0.65%
	22.2279	22.3545	98.77	0.1266 ± 0.0401	0.57%
1	13.7644	13.9149	100.00	0.1505 ± 0.0303	1.09%
	10.5686	10.6597	99.77	0.0911 ± 0.0272	0.86%
	7.5748	7.7162	99.99	0.1414 ± 0.0301	1.87%
	5.7005	5.8048	100.00	0.1043 ± 0.0184	1.83%
1	4.5153	4.5863	99.99	0.071 ± 0.018	1.57%
	3.6971	3.7472	98.92	0.05 ± 0.02	1.35%
	3.1717	3.2126	96.85	0.0409 ± 0.0151	1.29%
	2.9072	2.9638	99.99	0.0566 ± 0.0127	1.95%
1	75.1252	74.9263	100.00	-0.1989 ± 0.0372	-0.26%
	2.927	2.9488	100.00	0.0218 ± 0.0034	0.75%

АБТ: STORIES

Фильтры:			99.0 99.5 99.9		
Эксперименты: Контроль 4% в эксклюзиве с формулой 7656, контроль - проверка влияния экспериментов			99.0 99.5 99.9		
			99.0 99.5 99.9		
Основные метрики			WaldTest	Дельта	Дельта, %
	28.3548	28.1627	98.40	-0.1922 ± 0.0672	-0.68%
	32.5124	32.2941	99.68	-0.2183 ± 0.0678	-0.67%
	78.0228	78.1403	79.14	0.1175 ± 0.0663	0.15%
	24.1692	23.9881	97.55	-0.1812 ± 0.0644	-0.75%
	3.545.501	3.518.091	98.16	-27.410 ± 9869	-0.77%
	4.414.487	4.404.845	27.58	-9.642 ± 12486	-0.22%
	1.8656	1.8746	99.99	0.009 ± 0.002	0.48%
	2.0317	2.0393	99.45	0.0077 ± 0.0024	0.38%
	49.5836	49.6209	10.75	0.0373 ± 0.0802	0.08%
	20.0346	20.0893	35.40	0.0548 ± 0.0584	0.27%
	12.4693	12.5514	74.02	0.0821 ± 0.0510	0.66%
	6.5829	6.6863	99.98	0.1034 ± 0.0252	1.57%
	4.9727	5.0592	99.99	0.0865 ± 0.0222	1.74%
	3.3062	3.37	99.95	0.0638 ± 0.0176	1.93%
	2.8993	2.9699	100.00	0.0705 ± 0.0161	2.43%
	2.7797	2.8343	99.95	0.0546 ± 0.0148	1.97%
	13.9883	14.0175	10.74	0.0292 ± 0.0625	0.21%
	7.448	7.5482	99.96	0.1002 ± 0.0281	1.35%
	5.5811	5.6745	99.94	0.0934 ± 0.0237	1.67%
	4.4032	4.4844	99.99	0.0812 ± 0.0200	1.84%
	3.0931	3.1723	100.00	0.0793 ± 0.0164	2.56%
	2.8501	2.9148	99.98	0.0647 ± 0.0156	2.27%
	75.5511	75.3644	99.99	-0.1867 ± 0.0410	-0.25%
	2.8925	2.9109	100.00	0.0184 ± 0.0038	0.64%
	26.2325	26.0548	98.63	-0.1777 ± 0.0640	-0.68%
	30.4235	30.2162	99.75	-0.2073 ± 0.0659	-0.68%
	13.1424	14.3403	100.00	-0.1993 ± 0.0372	-0.26%
	2.927	2.9488	100.00	0.0218 ± 0.0034	0.75%

АБТ: STORIES

Фильтры:			99.0 99.5 99.9		
Эксперименты: Контроль 4% в эксклюзиве с формулой 7656, контроль - проверка влияния экспериментов			99.0 99.5 99.9		
99.0 99.5 99.9			99.0 99.5 99.9		
Основные метрики			WaldTest	Дельта	Дельта, %
	27.0718	27.0375	15.60	-0.0343 ± 0.0565	-0.13%
	31.8443	31.7884	38.85	-0.0559 ± 0.0574	-0.18%
	23.2328	23.1791	35.01	-0.0537 ± 0.0552	-0.23%
	39.773	39.6654	77.64	-0.1076 ± 0.0585	-0.27%
	4 120 906	4 108 163	44.25	$-12 743 \pm 11020$	-0.31%
	5 345 424	5 338 016	9.86	$-7 408 \pm 14807$	-0.14%
	1.82	1.8243	91.59	0.0042 ± 0.0019	0.23%
	2.0017	2.0057	81.68	0.004 ± 0.002	0.20%
	24.7098	24.6141	38.82	-0.0957 ± 0.0966	-0.39%
	51.7244	51.7709	28.68	0.0465 ± 0.0548	0.09%
	45.8828	45.9437	48.55	0.061 ± 0.053	0.13%
	53.9088	53.8958	1.18	-0.013 ± 0.081	-0.02%
	21.6305	21.6571	17.15	0.0266 ± 0.0428	0.12%
	12.4015	12.4955	98.80	0.0941 ± 0.0309	0.76%
	8.8711	8.9017	49.45	0.0306 ± 0.0266	0.34%
	6.6197	6.6763	95.38	0.0565 ± 0.0218	0.85%
	4.9651	5.027	98.81	0.0619 ± 0.0202	1.25%
	3.9511	3.9872	87.84	0.0361 ± 0.0170	0.91%
	3.2883	3.3249	91.37	0.0365 ± 0.0155	1.11%
	2.8749	2.9037	84.78	0.0287 ± 0.0151	1.00%
	2.7454	2.788	98.76	0.0425 ± 0.0146	1.55%
	55.3202	55.1861	63.21	-0.1341 ± 0.0876	-0.24%
	24.0719	24.1333	57.27	0.0614 ± 0.0494	0.25%
	13.8694	13.9585	94.29	0.089 ± 0.034	0.64%
	10.5588	10.5683	5.49	0.0095 ± 0.0271	0.09%
	7.4407	7.5148	99.55	0.0741 ± 0.0234	1.00%
	5.5545	5.5966	90.00	0.0421 ± 0.0203	0.76%
	4.3955	4.4301	84.37	0.0345 ± 0.0180	0.79%
	3.5623	3.6041	93.21	0.0419 ± 0.0169	1.18%
	3.0874	3.1094	66.68	0.022 ± 0.014	0.71%
	2.6255	2.6554	66.68	0.022 ± 0.014	0.71%

АБТ: STORIES

Побочные эффекты

- эксклюзив, 4%/4%, ухудшающий
- 5 дней ухудшение + 14 дней отдыха
- последующий АА-тест **значимо прокрасился**
- медленное восстановление
- ~1 месяц приготовлений, ~1 месяц анализа

АБТ: STORIES

Обновление интерфейса Поиска

- крупное изменение, **-0.4% запросов**

АБТ: STORIES

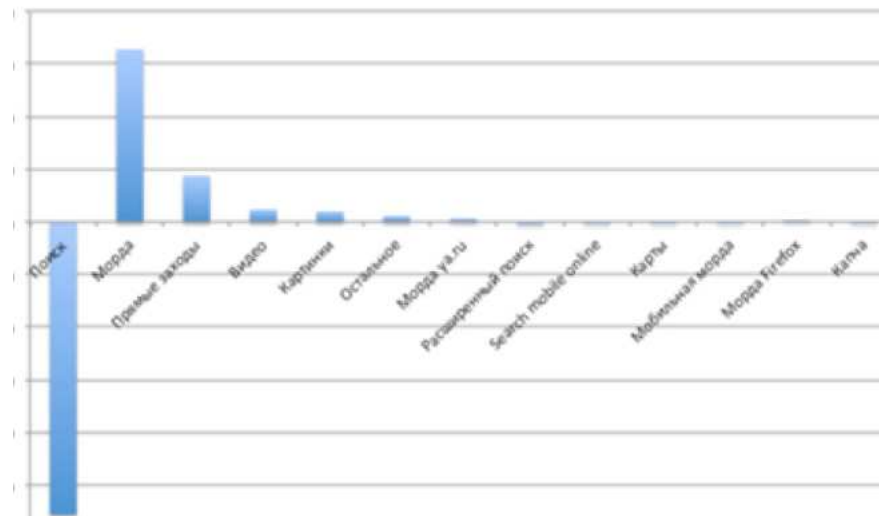
Обновление интерфейса Поиска

- крупное изменение, -0.4% запросов
- множество гипотез
 - пользователи раздражены
 - кэширование аяксовых запросов
 - технические ошибки

АБТ: STORIES

Обновление интерфейса Поиска

- крупное изменение, -0.4% запросов
- множество гипотез
- заметили изменение в портальной навигации



АБТ: STORIES

Обновление интерфейса Поиска

- крупное изменение, -0.4% запросов
- множество гипотез
- заметили изменение в порталной навигации
- клик по колдунщику открывал сайт в той же вкладке
- починили сломанную привычку

АБТ: STORIES

Обновление интерфейса Поиска

- крупное изменение, -0.4% запросов
- множество гипотез
- заметили изменение в порталной навигации
- клик по колдунщику открывал сайт в той же вкладке
- починили сломанную привычку
- ~4 месяца аналитики

АБТ: STORIES

Сниппетный угар

- 2 дня, 324 эксперимента **X** 0.25%

АБТ: STORIES

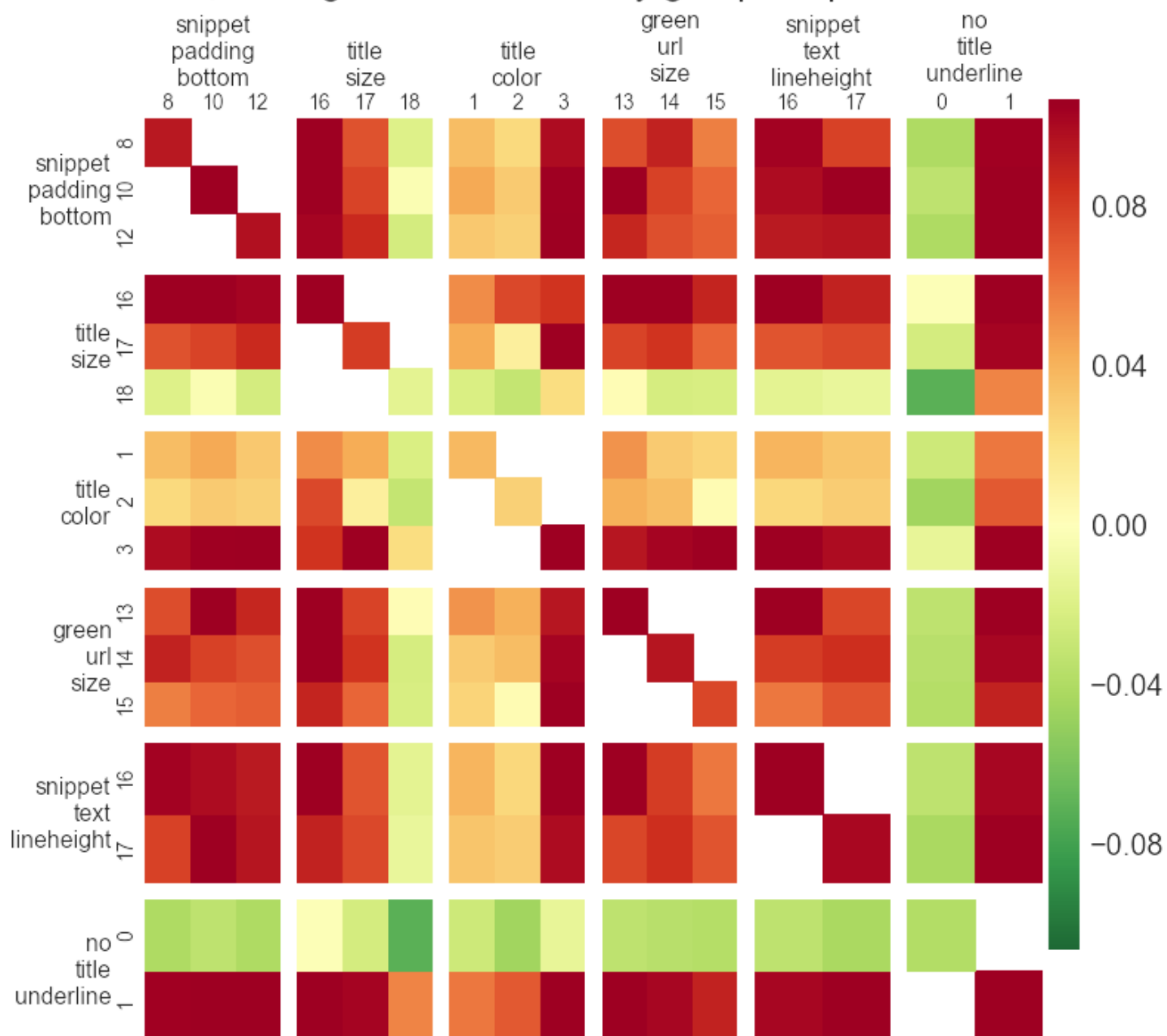
Сниппетный угар

- 2 дня, 324 эксперимента **X** 0.25%
- комбинаторный перебор:
 - размер & отступы гринурла и текста сниппета
 - размер, цвет, подчеркивание тайтлов

Visual variables and their values:

- snippet padding bottom: 8, 10, 12
- title size: 16, 17, 18
- title color: 1, 2, 3
- green url size: 13, 14, 15
- snippet text lineheight: 16, 17
- no title underline: 0, 1

Correlation scale: -0.08 to 0.08.



АБТ: STORIES

Сниппетный угар

- 2 дня, 324 эксперимента **X** 0.25%
- комбинаторный перебор
- +1.5-2% кликов :)

АБТ: STORIES

Сниппетный угар

- 2 дня, 324 эксперимента **X** 0.25%
- комбинаторный перебор
- +1.5-2% кликов :)
- ~1 месяц настройки, ~4 месяца аналитики

TDI

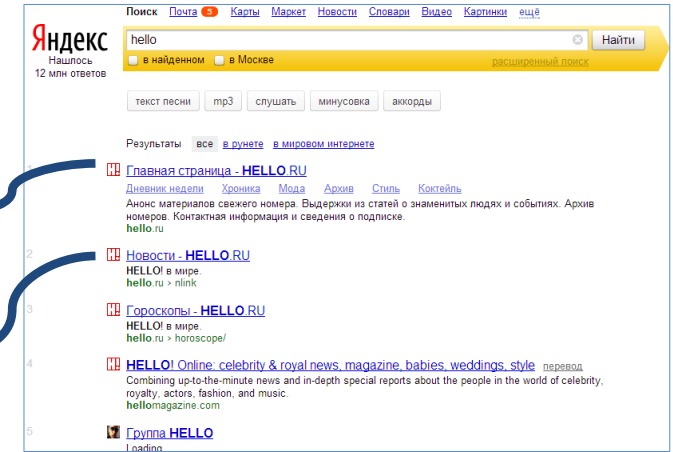
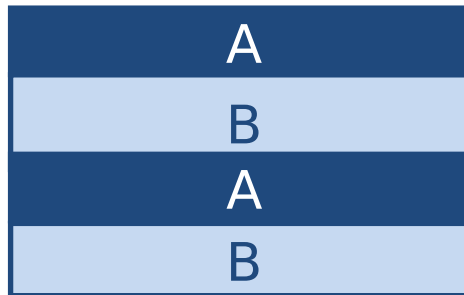
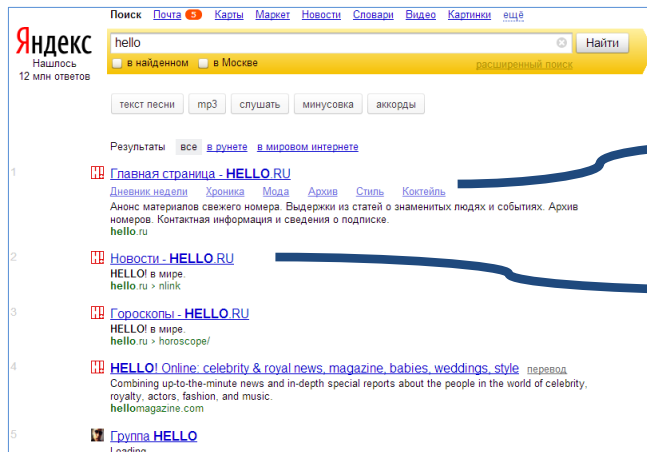
Смешивание поисковых выдач

Смешивание выдач

A

I

B



Пользователь

Даже не
подозревает, что
выдача получена
смешиванием

Как смешать выдачи?



1. ?

Как смешать выдачи?

A
B
B
A
B
A
A
B
B
A

1. Из 10-ти документов ровно 5 из A, ровно 5 из B

Как смешать выдачи?

А
В
В
А
В
А
А
В
В
А

1. Из 10-ти документов ровно 5 из А, ровно 5 из В
2. Кликаю произвольно – в среднем получаю одинаковое количество кликов по А и В

Как смешать выдачи?

А
В
В
А
В
А
А
В
В
А

1. Из 10-ти документов ровно 5 из А, ровно 5 из В
2. Кликаю произвольно – в среднем получаю одинаковое количество кликов по А и В
3. На каждой позиции документы из А и В появляются с равной вероятностью

Как смешать выдачи?

А
В
В
А
В
А
А
В
В
А

1. Из 10-ти документов ровно 5 из А, ровно 5 из В
2. Кликаю произвольно – в среднем получаю одинаковое количество кликов по А и В
3. На каждой позиции документы из А и В появляются с равной вероятностью
4. Сохранен порядок документов из исходных выдач

Team-Draft Interleaving (TDI)

Есть такой широко применяемый алгоритм смешивания.

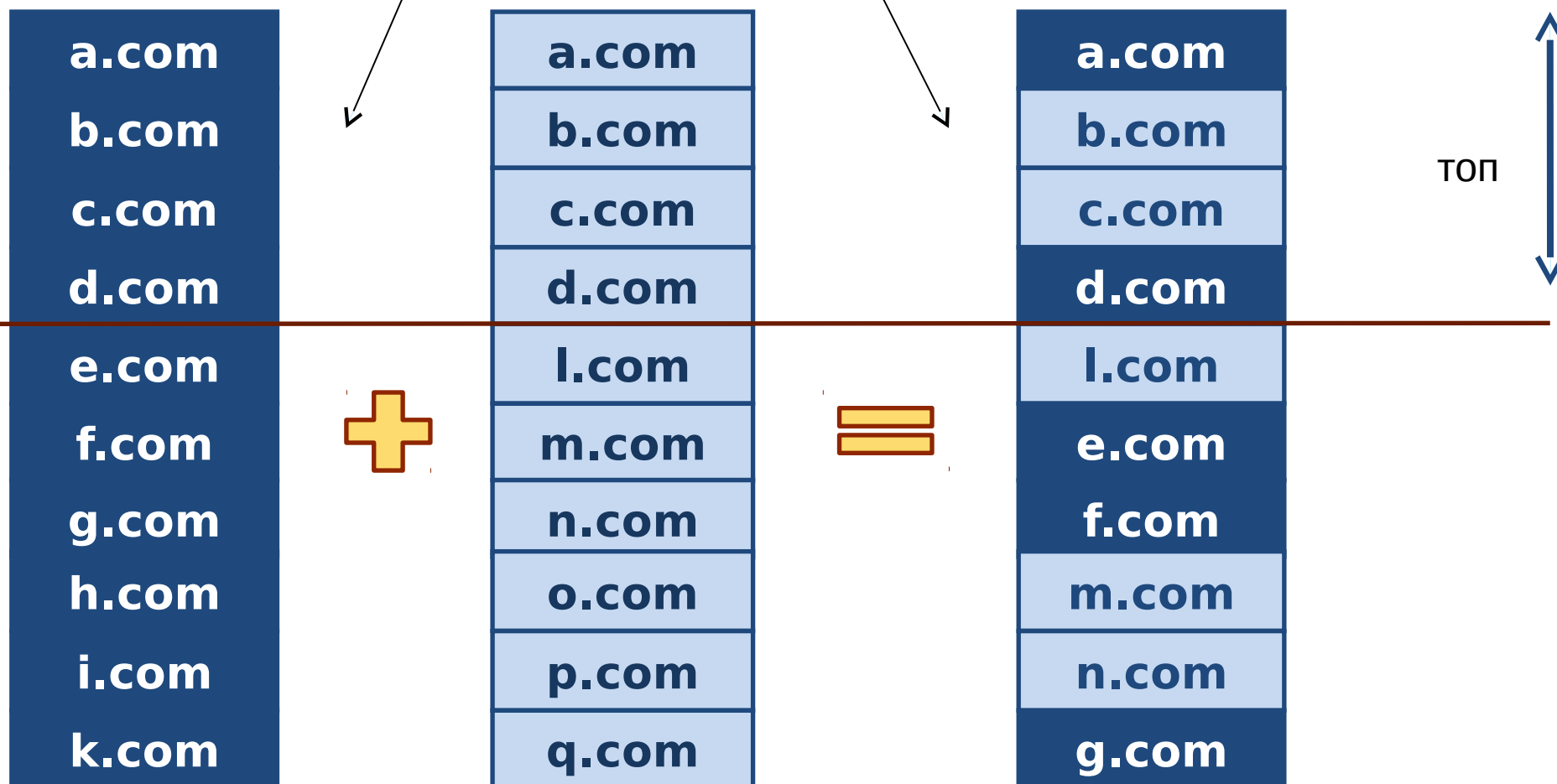
Отвечает всем требованиям.

ALGORITHM 2: Team-Draft Interleaving

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow ()$; $TeamA \leftarrow \emptyset$; $TeamB \leftarrow \emptyset$;
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do** *if not at end of A or B*
 if $(|TeamA| < |TeamB|) \vee$
 $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
 $k \leftarrow \min_i \{i : A[i] \notin I\}$ *top result in A not yet in I*
 $I \leftarrow I + A[k]$; *append it to I*
 $TeamA \leftarrow TeamA \cup \{A[k]\}$ *clicks credited to A*
 else
 $k \leftarrow \min_i \{i : B[i] \notin I\}$ *top result in B not yet in I*
 $I \leftarrow I + B[k]$ *append it to I*
 $TeamB \leftarrow TeamB \cup \{B[k]\}$ *clicks credited to B*
 end if
end while
Output: Interleaved ranking I , $TeamA$, $TeamB$

Общий топ

Инвариантен к смешиванию



Метрика



Метрика равна:

Ψ = разница кликов по **B** и **A**

← На этом примере:

$\Psi = ?$

Метрика



Метрика равна:

Ψ = разница кликов по **B** и **A**

← На этом примере:

$$\Psi = 1 - 2 = -1$$

Пользователь неявно предпочел выдачу **A**

TDI. Улучшение чувствительности

- Не учитывать клики по общему топу
- Не учитывать повторные клики по урлу
- Не учитывать общие документы
- Считать постранично
- Взвешенные клики

Olivier Chapelle et al. "Large-Scale Validation and Analysis of Interleaved Search Evaluation"

TDI

- Достоинства:
 - Чувствительнее, чем АБ-тестинг
- Недостатки:
 - Пользователи оценивают не релевантность, а свое ожидание релевантности
 - Оценивается не только ранжирование, но и сниппеты

Литература

- Ronny Kohavi <http://www.exp-platform.com/>
- Olivier Chapelle et al. "Large-Scale Validation and Analysis of Interleaved Search Evaluation"
- Alex Deng et al. "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data"
- А.Н. Ширяев "Вероятность"
- David J. Sheskin "Handbook of parametric and nonparametric statistical procedures"
- Г.И. Ивченко, Ю.И. Медведев. "Математическая статистика"

АБТ: Вопросы?

