

Яндекс



# Машинное обучение в информационном поиске

Шаграев А. Г.

к.т.н., ст. преп. каф. ПМ МЭИ,

руководитель службы свеже-социального поиска ООО «Яндекс»

# Contents

1 | Ранжирование в Поиске

2 | Ранжирование в Я.Новостях

3 | Разнообразие поисковой выдачи

4 | Wide pFound, «Спектр»

5 | Обучение по кликам

Машинное обучение в информационном поиске

# Ранжирование в Поиске



# Ранжирование в Поиске

Как оценивать качество ранжирования?


- › Возьмём запросы, которые нам задают пользователи
- › Возьмём документы из выдачи и покажем их ассессорам
- › Ассессоры проставят каждому документу по каждому запросу метку «релевантности»
- › Вычислим какую-нибудь метрику, исходя из оценок релевантности


# Ранжирование в Поиске


**Яндекс** конституция рф ✕ 🔊 ⚙️ Найти


**ПОИСК** КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ


---


 **"Конституция Российской Федерации" (принята...)**  
Статья 23 / КонсультантПлюс    Статья 24 / КонсультантПлюс  
[Consultant.ru](#) > document/cons\_doc\_LAW\_28399/ ▾  
▪ сознавая себя частью мирового сообщества  
▪ принимаем **КОНСТИТУЦИЮ** РОССИЙСКОЙ ФЕДЕРАЦИИ.

 **Конституция Российской Федерации — Википедия**  
[ru.wikipedia.org](#) > Конституция Российской Федерации ▾  
Конститу́ция Росси́йской Федера́ции — высший нормативный правовой акт Российской Федерации. Принята народом России 12 декабря 1993 года.

 **Конституция Российской Федерации**  
Основы конституционного строя    Федеративное устройство    Глава 6  
[constitution.ru](#) ▾  
Конституция Российской Федерации. Оптическая копия официального издания. Государственная власть **РФ**.

 **Конституция Российской Федерации | Верховный Суд РФ**  
[constitution.kremlin.ru](#) ▾  
Разъяснения отдельных положений Конституции **РФ** 1993 года ... акты Президента **РФ**, ст. 90. акты, применяемые при разрешении споров: ст. 15.4, ст. 76.5.

 **Конституция Российской Федерации**  
[constitution.garant.ru](#) ▾  
Конституция **РФ** (есть англ. вариант). Акты конституционного права. История принятия, конституции СССР и РСФСР (1918-1992). Научные работы. Конституции и Уставы субъектов **РФ**.


 **Конституция Российской Федерации: Все главы и статьи**  
[kodeks.systems.ru](#) > Конституция ▾  
Кодексы и законы **РФ**. ... Конституция Российской Федерации Актуальная редакция Конституции от 21.07.2014 с изменениями, вступившими в силу с 21.07.2014.


# Ранжирование в Поиске


**Яндекс** конституция рф ✕ 🔊 ⚙️ Найти


ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ


---


 **"Конституция Российской Федерации" (принята...)**  
Статья 23 / КонсультантПлюс    Статья 24 / КонсультантПлюс  
[Consultant.ru](#) > document/cons\_doc\_LAW\_28399/ ▾  
▪ сознавая себя частью мирового сообщества  
▪ принимаем **КОНСТИТУЦИЮ** РОССИЙСКОЙ ФЕДЕРАЦИИ.

 **Конституция Российской Федерации — Википедия**  
[ru.wikipedia.org](#) > Конституция Российской Федерации ▾  
Конститу́ция Росси́йской Федера́ции — высший нормативный правовой акт Российской Федерации. Принята народом России 12 декабря 1993 года.

 **Конституция Российской Федерации**  
Основы конституционного строя    Федеративное устройство    Глава 6  
[constitution.ru](#) ▾  
Конституция Российской Федерации. Оптическая копия официального издания. Государственная власть **РФ**.

 **Конституция Российской Федерации | Верховный Суд РФ**  
[constitution.kremlin.ru](#) ▾  
Разъяснения отдельных положений Конституции РФ 1993 года ... акты Президента РФ, ст. 90. акты, применяемые при разрешении споров: ст. 15.4, ст. 76.5.

 **Конституция Российской Федерации**  
[constitution.garant.ru](#) ▾  
Конституция **РФ** (есть англ. вариант). Акты конституционного права. История принятия, конституции СССР и РСФСР (1918-1992). Научные работы. Конституции и Уставы субъектов **РФ**.

 **Конституция Российской Федерации: Все главы и статьи**  
[kodeks.systems.ru](#) > Конституция ▾  
Кодексы и законы **РФ**. ... Конституция Российской Федерации Актуальная редакция Конституции от 21.07.2014 с изменениями, вступившими в силу с 21.07.2014.

# Ранжирование в Поиске

## Метрики ранжирования: DCG

- › Пусть документы  $d_1, d_2, \dots, d_n$  в выдаче имеют релевантности  $r_1, r_2, \dots, r_n$
- › DCG = Discounted Cumulative Gain

$$DCG = \sum_{i=1}^n \frac{r_i}{\ln(i + 1)}$$



# Ранжирование в Поиске

Метрики ранжирования: DCG

$$DCG = \sum_{i=1}^n \frac{r_i}{\ln(i + 1)}$$

- › Почему логарифм?
- › Wang Y. et. al. A Theoretical Analysis of NDCG Ranking Measures (2013)

<http://proceedings.mlr.press/v30/Wang13.pdf>

# Ранжирование в Поиске

## Метрики ранжирования: pFound

- › Изобретение Яндекса: вероятностная модель пользователя
- › Пользователь просматривает выдачу сверху вниз
- › Пользователь достигает успеха на документе  $d_i$  удовлетворяет его с вероятностью  $r_i$
- › При неуспехе с вероятностью  $pBreak$  пользователь устаёт и уходит с выдачи

# Ранжирование в Поиске

Метрики ранжирования:  $pFound$

- ›  $pLook_i$  – вероятность того, что пользователь посмотрит на  $d_i$ :

$$pLook_1 = 1$$

$$pLook_{i+1} = pLook_i \cdot (1 - r_i) \cdot (1 - pBreak)$$

- ›  $pFound_i$  – вероятность того, что пользователь достигнет успеха на документе  $d_i$ :

$$pFound_i = pLook_i \cdot r_i$$

# Ранжирование в Поиске

Метрики ранжирования: *pFound*

- › Интегральный *pFound* – вероятность того, что пользователь достигнет успеха

$$pFound = \sum_{i=1}^n pFound_i = \sum_{i=1}^n pLook_i \cdot r_i$$

# Ранжирование в Поиске

## Методы обучения ранжированию

- › Метрики качества ранжирования не являются гладкими
- › Методы обучения ранжированию (list-wise) существенно сложнее, чем «точечные» методы обучения (point-wise)
- › Используются различные схемы сглаживания метрик для оптимизации

# Ранжирование в Поиске

## Ранжирование: вероятностные модели

- › Lozano J., Iruozoki E. Probabilistic Modeling on Rankings  
(2013)

[http://www.sc.ehu.es/ccwbayes/members/ekhine/tutorial\\_ranking/data/slides.pdf](http://www.sc.ehu.es/ccwbayes/members/ekhine/tutorial_ranking/data/slides.pdf)

# Ранжирование в Поиске

## Ранжирование: Plackett-Luce model

- › Пусть документы упорядочены по истинной релевантности
- › Пусть текущая формула предсказывает для них значения

$$p_1, p_2, \dots, p_n$$

# Ранжирование в Поиске

## Ранжирование: Plackett-Luce model

- › Тогда вероятность получить наилучшую перестановку равна

$$P = \frac{\exp p_1}{\sum_{i=1}^n \exp p_i} \cdot \frac{\exp p_2}{\sum_{i=2}^n \exp p_i} \cdot \dots \cdot \frac{\exp p_n}{\sum_{i=n}^n \exp p_i}$$

- › Задача обучения – оптимизировать среднее значение  $P$  по всем запросам



# Ранжирование в Поиске

Вероятностное моделирование метрик ранжирования

› Генерируем несколько перестановок документов

$$\pi_1, \pi_2, \dots, \pi_k$$

# Ранжирование в Поиске

Вероятностное моделирование метрик ранжирования

- › Генерируем несколько перестановок документов

$$\pi_1, \pi_2, \dots, \pi_k$$

- › Для каждой перестановки  $\pi_i$  определена «вероятность» её получения  $P(f, \pi_i)$ , зависящая от решающей функции  $f$ , и значение метрики качества ранжирования  $Q(\pi_i)$

# Ранжирование в Поиске

## Вероятностное моделирование метрик ранжирования

- › Генерируем несколько перестановок документов

$$\pi_1, \pi_2, \dots, \pi_k$$

- › Для каждой перестановки  $\pi_i$  определена «вероятность» её получения  $P(f, \pi_i)$ , зависящая от решающей функции  $f$ , и значение метрики качества ранжирования  $Q(\pi_i)$
- › Оптимизируем величину

$$EQ = \sum_{i=1}^k P(f, \pi_i) Q(\pi_i)$$

# Ранжирование в Поиске

## Прямая оптимизация метрик ранжирования

- › Burges C. From RankNet to LambdaRank to LambdaMART:  
an overview

<https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/>

# Ранжирование в Поиске

## Pairwise-постановка задачи

Пусть  $q_i$  – запрос, которому соответствуют векторы документов и соответствующих им релевантностей:

$$\left\langle d_i^1, d_i^2, \dots, d_i^{k_i} \right\rangle, \left\langle r_i^1, r_i^2, \dots, r_i^{k_i} \right\rangle$$

Тогда можно сказать, что целевая функция определена на парах документов:

$$y^*(d_i^j, d_i^l) = [r_i^j > r_i^l]$$

# Ранжирование в Поиске

## Pairwise-постановка задачи

Например, красиво записывается функционал потерь для логистической регрессии:

$$Q = \sum_{j=1}^{k_i} \sum_{l=1}^{k_i} \ln \left( 1 + \exp \left( - (r_i^j - r_i^l) (f(d_i^j) - f(d_i^l)) \right) \right)$$

# Ранжирование в Поиске

Pairwise-постановка задачи

$$Q = \sum_{j=1}^{k_i} \sum_{l=1}^{k_i} \ln \left( 1 + \exp \left( - (r_i^j - r_i^l) (f(d_i^j) - f(d_i^l)) \right) \right)$$

Сравните с постановкой в задаче бинарной классификации для классов  $\{+1, -1\}$

$$Q = \sum_d \ln(1 + \exp(-y^*(d) \cdot f(d)))$$

# Ранжирование в Поиске

## Pairwise-постановка задачи

Пары можно взвешивать в зависимости от того, насколько они важны

Простой способ – вес пары пропорционален разнице в значениях релевантности

Обычный pairwise-метод приводит к методу RankNet



# Ранжирование в Поиске

Pairwise-постановка задачи

Градиенты RankNet выделены чёрным

Красные градиенты – те, которых хотелось бы добиться



# Ранжирование в Поиске

Pairwise-постановка задачи

Lambda-trick: величина градиента умножается на изменение метрики ранжирования при перестановке двух элементов

# Ранжирование в Поиске

## Pairwise-постановка задачи

## Доказательство корректности метода ☺

arbitrarily tight bound by applying a one-sided Monte Carlo test: choose sufficiently many random directions in weight space, move the weights a little along each such direction, and check that  $M$  always decreases as we move away from  $w^*$ . Specifically, we choose directions uniformly at random by sampling from a spherical Gaussian. Let  $p$  be the fraction of directions that result in  $M$  increasing. Then

$$P(\text{We miss an ascent direction despite } n \text{ trials}) = (1 - p)^n$$

Let's call  $1 - P$  our confidence. If we require a confidence of 99% (i.e. we choose  $\delta = 0.01$  and require  $P \leq \delta$ ), how large must  $n$  be, in order that  $p \leq p_0$ , where we choose  $p_0 = 0.01$ ? We have

$$(1 - p_0)^n \leq \delta \rightarrow n \geq \frac{\log \delta}{\log(1 - p_0)} \quad (9)$$

which gives  $n \geq 459$  (i.e. choose 459 random directions and always find that  $M$  decreases along those directions; note that larger value of  $p_0$  would require fewer tests). In general we have confidence at least  $1 - \delta$  that  $p \leq p_0$  provided we perform at least  $n = \frac{\log \delta}{\log(1 - p_0)}$  tests.

# Ранжирование в Поиске

## Оптимизация кликовых метрик

- › Релевантность – не единственный аспект качества поисковой системы
- › Можно получить много сигнала для обучения, если анализировать поведение пользователей

# Ранжирование в Поиске

## Оптимизация кликовых метрик

- › Joachims T. Optimizing Search Engines using Chlickthrough Data (2002)

[http://www.cs.cornell.edu/people/tj/publications/joachims\\_02c.pdf](http://www.cs.cornell.edu/people/tj/publications/joachims_02c.pdf)

# Ранжирование в Поиске

## Оптимизация кликовых метрик

- › Скачиваются результаты нескольких поисковых систем
- › Результаты демонстрируются пользователям, собираются клики
- › Обучается модель, предсказывающая, какие документы привлекут большее количество кликов

# Ранжирование в Поиске

$d_1$

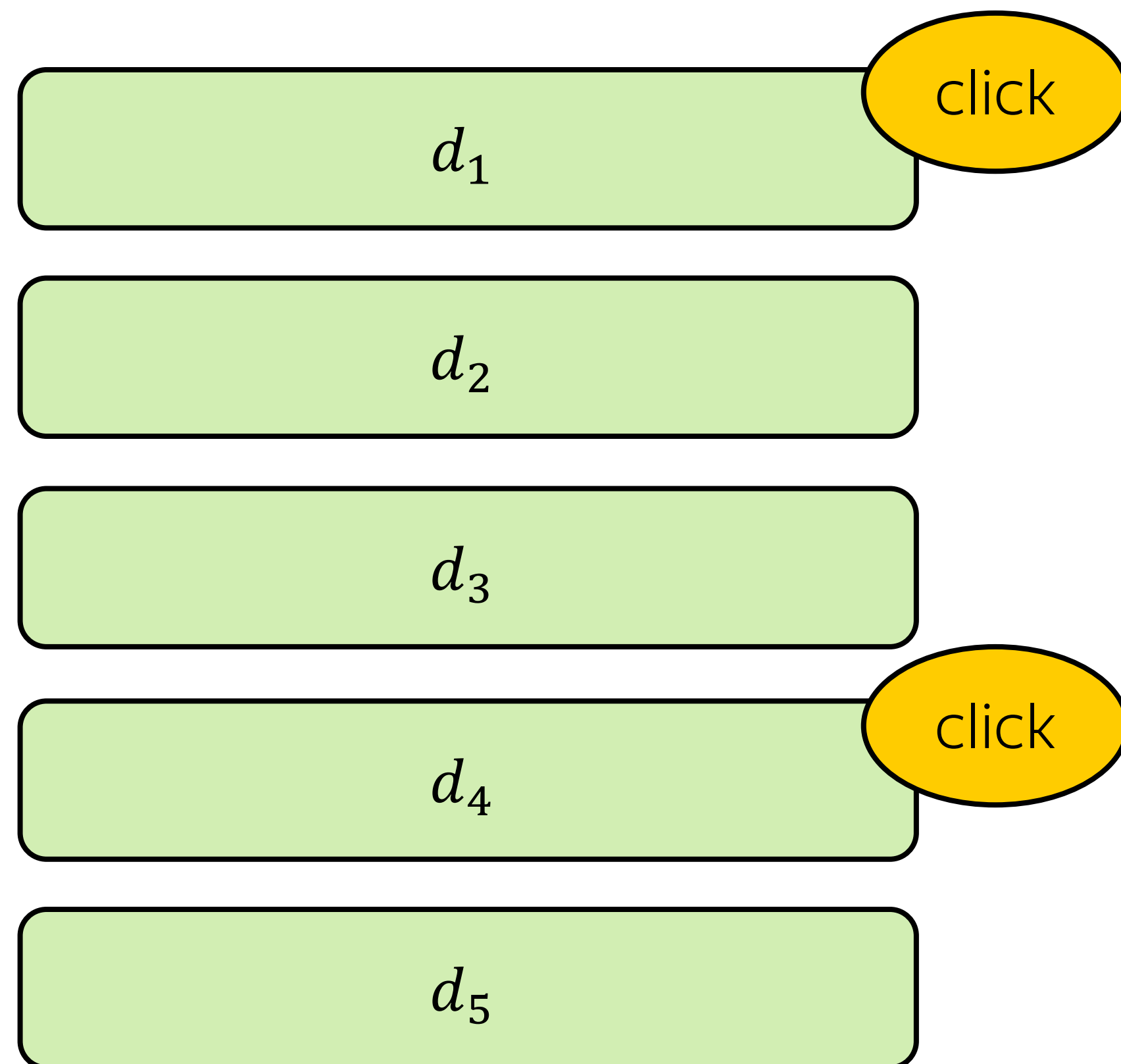
$d_2$

$d_3$

$d_4$

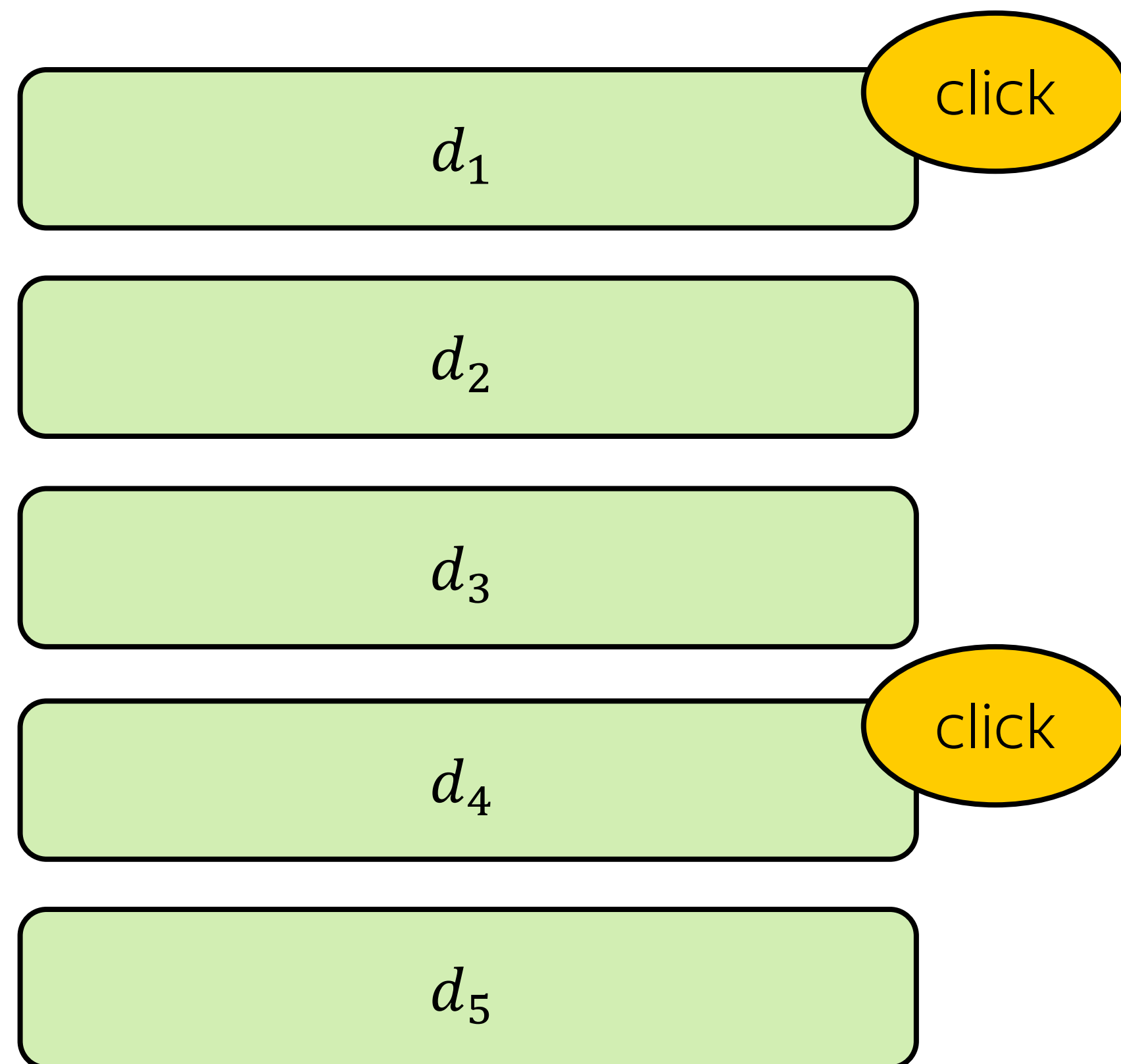
$d_5$

# Ранжирование в Поиске





# Ранжирование в Поиске



Примеры для обучения

$$d_2 \succ d_1$$

$$d_4 \succ d_1$$

$$d_4 \succ d_3$$

# Ранжирование в Поиске

## Клики как способ сравнивать поисковые системы

- › Chapelle O., Joachims T., Radlinski F., Yue Y. Large-scale validation and analysis of interleaved search evaluation (2012)

[http://www.cs.cornell.edu/people/tj/publications/chapelle\\_etal\\_12a.pdf](http://www.cs.cornell.edu/people/tj/publications/chapelle_etal_12a.pdf)

- › Chuklin A., Schuth A., Hofmann K., Serdyukov P., de Rijke M. Evaluating Aggregated Search Using Interleaving (2013)

<https://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/chuklin-evaluating-2013.pdf>

# Ранжирование в Поиске

## Клики как способ сравнивать поисковые системы

- › Balanced interleaving. Thorsten Joachims. Evaluating Retrieval Performance using Clickthrough Data (2002)

[https://www.cs.cornell.edu/people/tj/publications/joachims\\_02b.pdf](https://www.cs.cornell.edu/people/tj/publications/joachims_02b.pdf)

- › Team draft interleaving (TDI). Radlinski F., Kurup M., Joachims T. How does clickthrough data reflect retrieval quality (2008)

[http://www.cs.cornell.edu/People/tj/publications/radlinski\\_etal\\_08b.pdf](http://www.cs.cornell.edu/People/tj/publications/radlinski_etal_08b.pdf)

Машинное обучение в информационном поиске

# Я.Новости: ранжирование



# Я.Новости: ранжирование

- › Кластеризация новостей – пример обучения без учителя, в процессе которого нужно решать множество задач обучения с учителем

# Я.Новости: ранжирование

- › Кластеризация новостей – пример обучения без учителя, в процессе которого нужно решать множество задач обучения с учителем
- › Ранжирование новостей – пример обучения с учителем, в котором учителя на самом деле нет

# Я.Новости: ранжирование

- › Основной результат – топ-5 новостей на главной странице Яндекса

Сейчас в СМИ в Москве 11 октября, среда 09 15

- Руководство Каталонии подписало декларацию о независимости
- Чешский премьер раскритиковал президента Земана за слова о Крыме
- «Муж в порядке»: жена Хворостовского опровергла сообщения о его смерти
- Украинские военные объяснили попадание на территорию РФ бойца ВСУ
- Трамп прокомментировал слова о возможности третьей мировой войны

# Я.Новости: ранжирование

- › Обучение ранжирования по ручной разметке невозможно: у нас нет редакционной политики
- › Обучение по кликам невозможно: «жёлтые» новости побеждают
- › Обучение по публикациям в СМИ невозможно из-за «обратной связи»: они быстро начинают писать о том, что попало в топ-5



# Я.Новости: ранжирование

## Обучение ранжированию: принципы

- › Нужно брать пользовательский сигнал
- › Сигнал нужно очищать от обратной связи (не учитывать переходы из Я.Новостей) и накруток

# Я.Новости: ранжирование

Обучение ранжированию: принципы

› Возьмём чистые данные от пользователей

# Я.Новости: ранжирование

## Обучение ранжированию: принципы

- › Возьмём чистые данные от пользователей
- › Будем предсказывать для каждой новости внимание, которое она получит в ближайшие часы

# Я.Новости: ранжирование

## Обучение ранжированию: принципы

- › Возьмём чистые данные от пользователей
- › Будем предсказывать для каждой новости внимание, которое она получит в ближайшие часы
- › Те новости, что мы решили поместить в топ, мы должны поместить как можно раньше

# Я.Новости: ранжирование

## Обучение ранжированию: принципы

- › Принимаем алгоритм по CTR новостей, контролируя желтизну, «происшественность», разнообразие ~~и отзывы в соцсетях~~
- › Используем факторы, которые позволяют бороться с накрутками

Машинное обучение в информационном поиске

# Разнообразие поисковой выдачи



# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

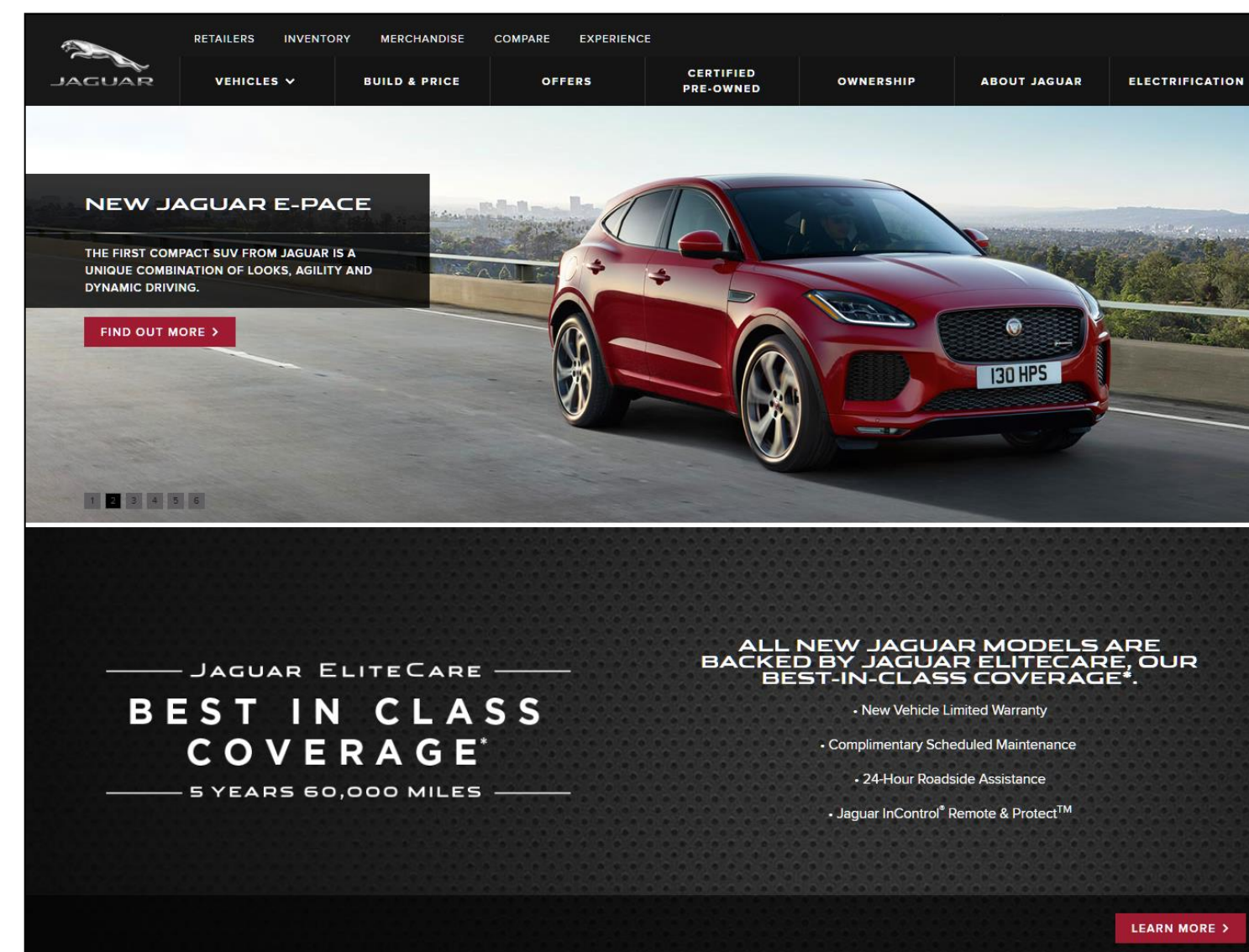
› Например, запрос «Ягуар»



# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

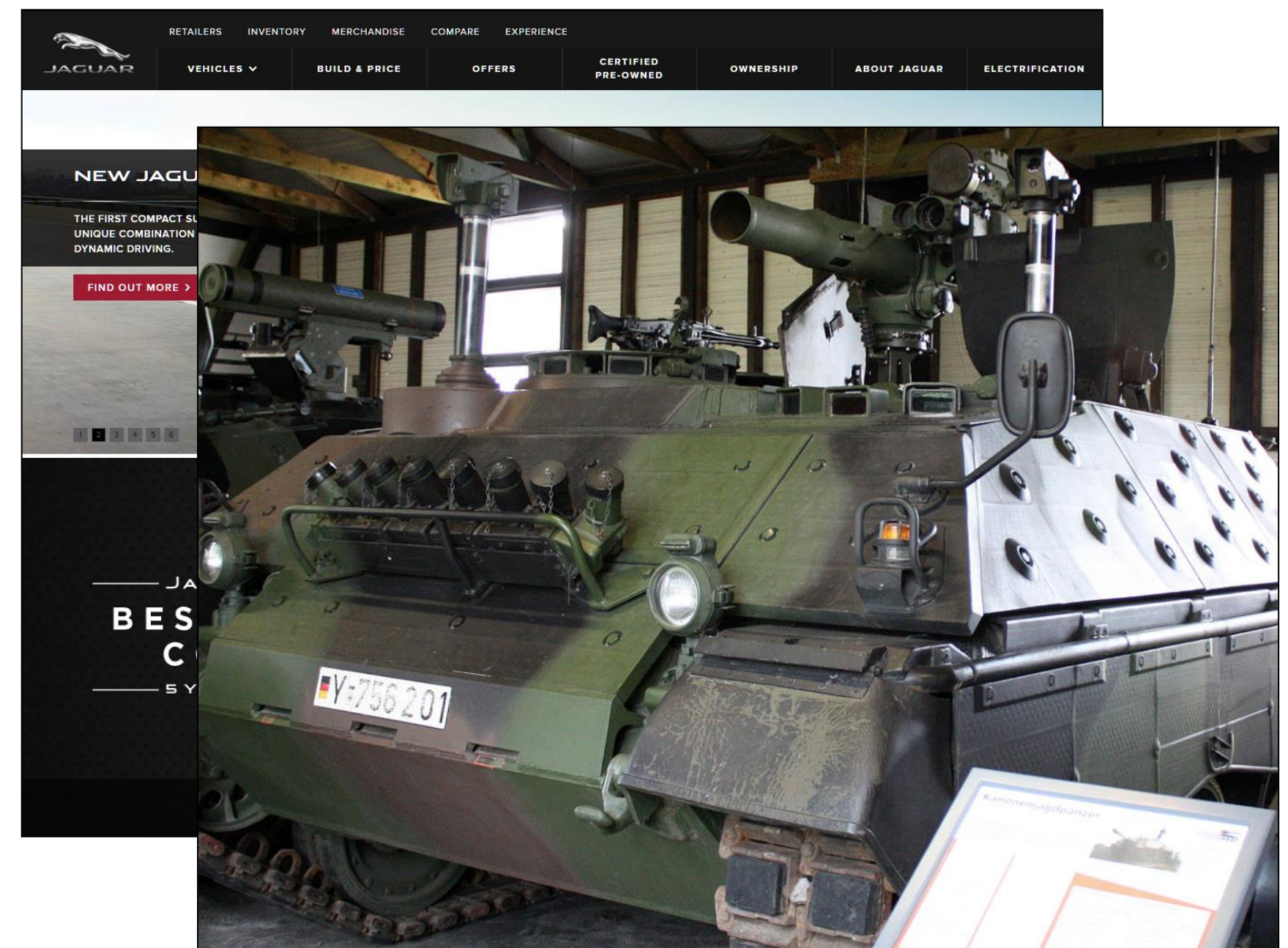
- › Например, запрос «Ягуар»
- › Автомобиль?



# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

- › Например, запрос «Ягуар»
- › Автомобиль?
- › Танк?

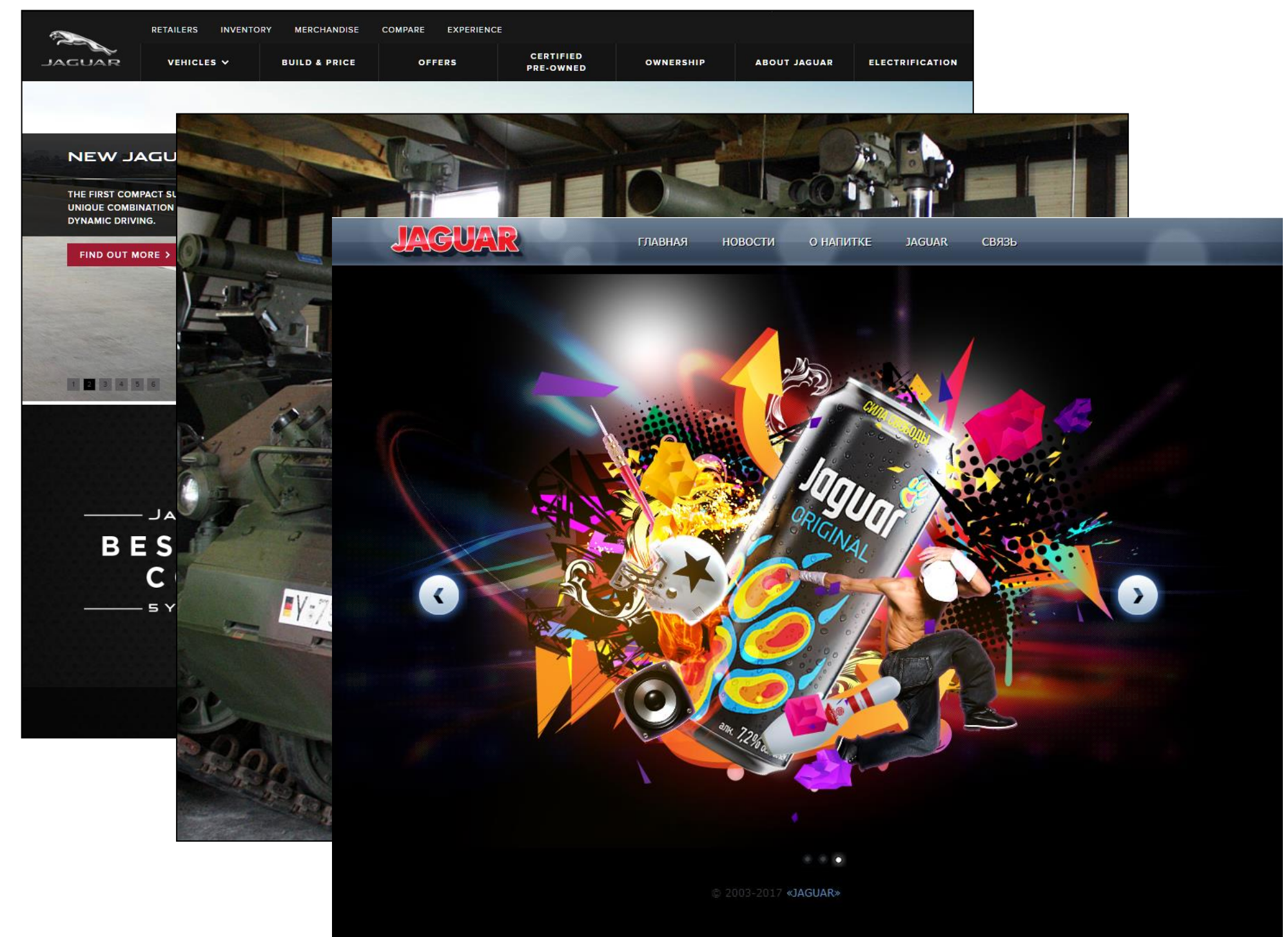




# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

- › Например, запрос «Ягуар»
- › Автомобиль?
- › Танк?
- › Напиток?



# Разнообразие поисковой выдачи

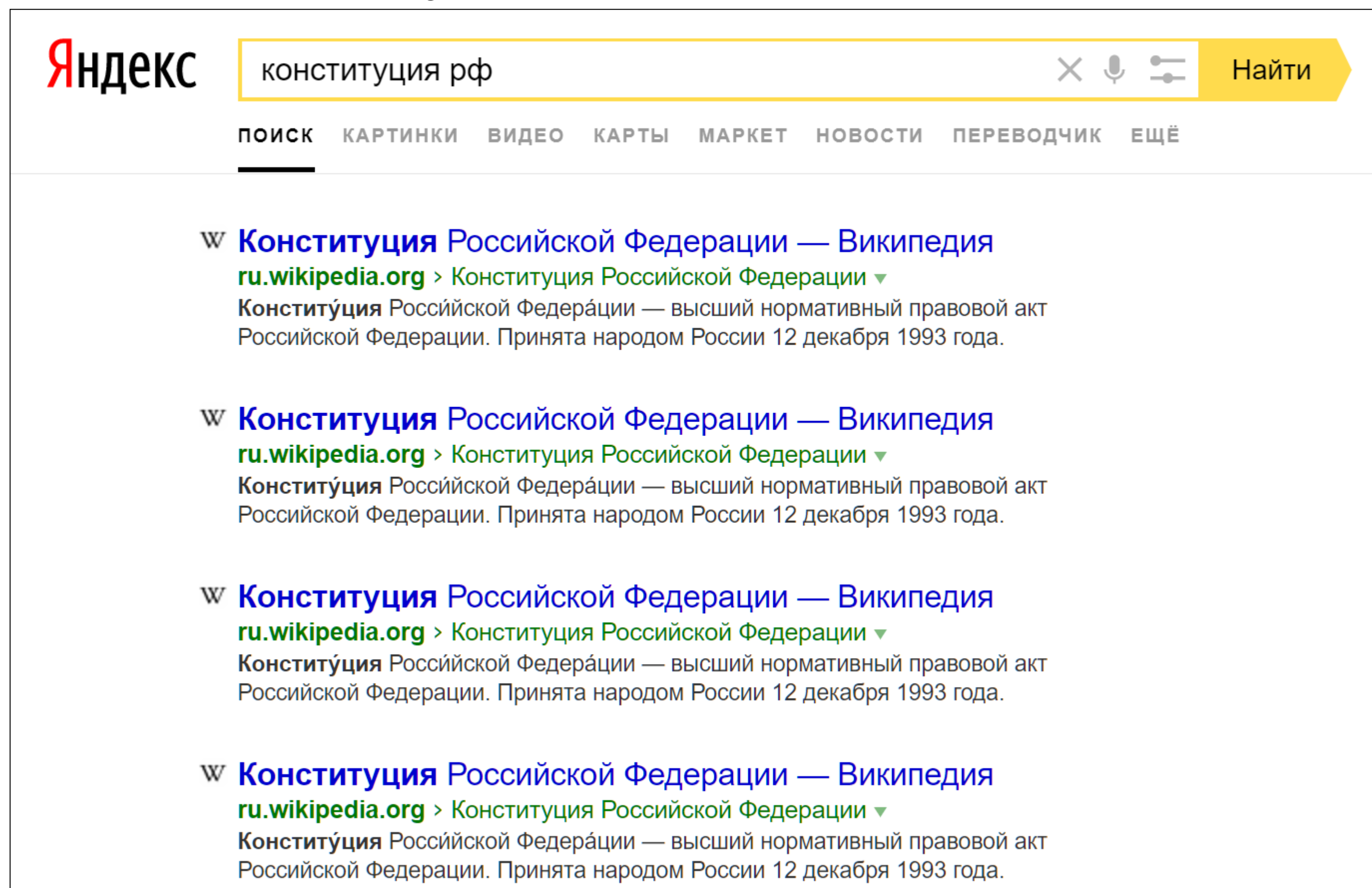
Проблема: неоднозначные запросы

- › С т.з. обычных метрик, выгодно замостить выдачу самым релевантным документом

# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы

- › С т.з. обычных метрик, выгодно замостить выдачу самым релевантным документом

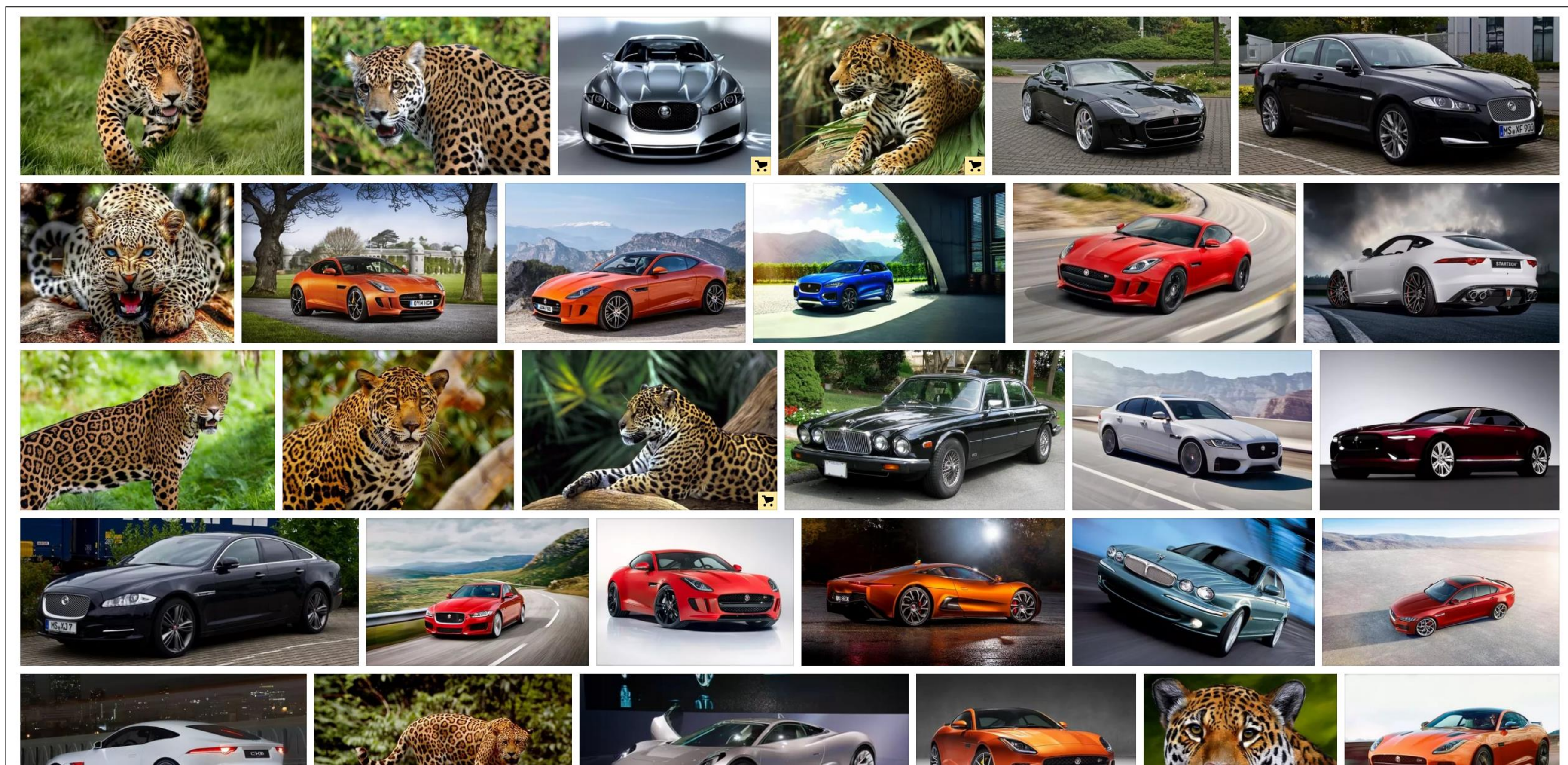




# Разнообразие поисковой выдачи

Проблема: неоднозначные запросы


› Разнообразие позволяет сделать так





# Разнообразие поисковой выдачи

## Проблема #2: разные типы информации

Найти


ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ

Отображение куки


**Кошка** — Википедия  
[ru.wikipedia.org](https://ru.wikipedia.org) > Кошка ▾  
Ко́шка, или домашняя ко́шка (лат. *Félis silvestris catus*), — домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов». С зоологической точки зрения домашняя кошка — млекопитающее семейства...

Купить **кошек** и котят из питомника и частные объявления...  
Котята Мейн кун Британский кот Кекс и Абрикос  
[avito.ru](https://avito.ru) > Объявления > Кошки ▾  
Объявления о продаже **кошек** и котят в **Москве** на Avito. ... Сегодня 20:33. В избранное. Котята коты и кошки.


**кошки** — 11 тыс. видео  
[Яндекс.Видео](https://yandex.ru/video) > кошки ▾




▶ 12:37 HD  
Смешные кошки 2017.  
[myvi.ru](https://myvi.ru)  
19 сентября  
Видео партнёра



▶ 5:31 HD  
Кошки и котята лучшие моменты.  
[youtube.com](https://youtube.com)





▶ 3:57  
Смешные кошки 2017.  
[myvi.ru](https://myvi.ru)  
17 сентября  
Видео партнёра



▶ 7:33 HD  
Смешные кошки new подборка 1 2015 (HD)  
[youtube.com](https://youtube.com)

→ **Породы кошек** с фотографиями и названиями.  
[mydata.com](https://mydata.com) > Породы кошек


### Кошка




Домашнее животное, одно из наиболее популярных «животных-компаньонов». С зоологической точки зрения домашняя кошка - млекопитающее семейства кошачьих отряда хищных. [Википедия](#)

Семейство: кошачьи (Felidae)  
Вид: Кошки (Felis)


### Смотрите также




Манчкин




Ориентал... кошка




Шартрез



Русская голубая кошка



Шотландс... вислоухая кошка



Мэнкс

55

Блендер

Wide pFound





# Wide pFound

## Развитие поисковых метрик: Wide pFound

- › Предполагаем, что пользователь мог иметь в виду один из множества «интентов»  $I = \{I_1, \dots, I_m\}$
- › Примеры интентов: машины, картинки, видео, новости, животные...
- › Каждый интент  $I_i$  имеет некоторую вероятность  $p(I_i)$  и порождает собственное распределение релевантностей на документах  $r: I \times D \rightarrow [0,1]$

# Wide pFound

## Развитие поисковых метрик: Wide pFound

- › Тогда для каждого интента  $I_i$  по отдельности можно вычислить соответствующий ему  $pFound(I_i)$

$$wpFound = \sum_{i=1}^m p(I_i) \cdot pFound(I_i)$$

- › Как вычислять вероятности интентов?

# Wide rFound для Свежести

Пример:

›  $freshIP = 0.4$

›  $webIP = 0.6$

› Векторы релевантностей:

$[0.2, 0.18, 0.16, 0.15, 0.14, \dots]$

# Wide pFound для Свежести

```
1 import re
2
3 webRelevances = [0.20, 0.18, 0.16, 0.15, 0.14, 0.13, 0.12, 0.11, 0.10, 0.09, 0.08]
4 freshRelevances = [0.20, 0.18, 0.16, 0.15, 0.14, 0.13, 0.12, 0.11, 0.10, 0.09, 0.08]
5
6 webIP = 0.6
7 freshIP = 0.4
8
9 wpFound = 0.
10
11 webPFound = 0.
12 freshPFound = 0.
13
14 webPLook = 1.
15 freshPLook = 1.
16
17 webPos = 0
18 freshPos = 0
19
```

# Wide pFound для Свежести

```
20 for i in range(10):
21     freshAddition = freshIP * freshPLook * freshRelevances[freshPos]
22     webAddition = webIP * webPLook * webRelevances[freshPos]
23
24     parts = map(str, ["", freshPLook, webPLook, freshPFound, webPFound])
25
26     if freshAddition > webAddition:
27         parts[0] = 'FRESH'
28         parts += ['0', str(freshRelevances[freshPos])]
29
30         wpFound += freshAddition
31         freshPFound += freshPLook * freshRelevances[freshPos]
32
33         webPLook *= 0.85
34         freshPLook *= 0.85 * (1 - freshRelevances[freshPos])
35
36         freshPos += 1
37     else:
38         parts[0] = 'WEB'
39         parts += [str(webRelevances[freshPos]), '0']
40
41         wpFound += webAddition
42         webPFound += webPLook * webRelevances[webPos]
43
44         webPLook *= 0.85 * (1 - webRelevances[freshPos])
45         freshPLook *= 0.85
46
47         webPos += 1
48
49     print re.sub('\.', ',', '\t'.join(parts))
50
```

# Wide rFound для Свежести

Яндекс

акции в дикси

✕ 🔊 ⚙

Найти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ

Д

**Дикси – Большой выбор низких цен.**

[dixy.ru](#) > [akcii/have-time/](#) ▾

Акции партнеров. Вместе мы делаем мир лучше. ... © Дикси 2005-2017. Все права защищены.

Р

**Каталог Дикси - Акции сегодня - с 6 по 12 ноября 2017...**

Архив С 22 по 28 мая С 29 мая по 4 июня С 15 по 21 мая

[proshoper.ru](#) > [Акции](#) > [Дикси](#) > [Москва](#) ▾

Каталог акций «Дикси» — Москва. с 6 по 12 ноября 2017. ... Сезонные каталоги акций в Дикси сегодня

♥

**Дикси каталог с 13 по 19 ноября 2017. - Акции в Москве...**

[moskidka.ru](#) > [news.php?extend.27865](#) ▾

Дикси акции. Очередной каталог товаров магазина Дикси, с 20 по 26 ноября 2017 года, действующий в Москве и Московской области, будет опубликован на нашем... **позавчера**

■

**Дикси - Акции и скидки супермаркетов Москвы**

[skidkaonline.ru](#) > [Москва](#) > [Дикси](#) ▾

Акции супермаркета Дикси обновляются регулярно, а об их начале и окончании можно узнать, посмотрев на картинку товара.

♥

**Дикси - Акции в Москве и Подмосковье.**

[moskidka.ru](#) > [plugins/tagcloud/tagcloud.php?Дикси](#) ▾

Дикси акции: Каталог Дикси с товарами по акции выходит еженедельно и ... Дикси акции Москва: Хотите получить в магазине Дикси второй товар со скидкой?

■

**16 Ноября - Дикси - Акции и скидки супермаркетов...**

[skidkaonline.ru](#) > [smolensk/16-11-2017/shops/diksi/](#) ▾

Акции супермаркета Дикси обновляются регулярно, а об их начале и окончании можно узнать, посмотрев на картинку товара. **15 часов назад**

🛒

**Акции Дикси - Скидки в супермаркетах Москвы и Области**

[skidkimarket.ru](#) > [Каталог-Акции-Дикси.php](#) ▾

Каталог Дикси: акции в Москве можно разделить на две большие категории. Это каталоги о «предложение недели» и «уникальная акция».

📖

**Каталог: скидки и акции в «Дикси» с 13 по 19 ноября 2017**

[wikishopping.ru](#) > [news/entry/katalog...aksii-v-diksi...](#) ▾

Каталог акций и скидок в «Дикси» с 13 по 19 ноября 2017. Новые предложения недели, скидки по акции «1+1» и сотни других товаров по низким ценам. **позавчера**

★★★★★ 5/5 - 1 оценка

# Wide pFound для Свежести

Документ	fpLook	wpLook	fpFound	wpFound	wRel	fRel
<b>WEB</b>	1,0000	1,0000	0,0000	0,0000	<b>0,2000</b>	0,0000
<b>WEB</b>	0,8500	0,6800	0,0000	0,2000	<b>0,1800</b>	0,0000
<b>FRESH</b>	0,7225	0,4624	0,0000	0,3224	0,0000	<b>0,2000</b>
<b>WEB</b>	0,4913	0,3930	0,1445	0,3224	<b>0,1600</b>	0,0000
<b>FRESH</b>	0,4176	0,2739	0,1445	0,3853	0,0000	<b>0,1800</b>
<b>WEB</b>	0,2911	0,2329	0,2197	0,3853	<b>0,1500</b>	0,0000
<b>WEB</b>	0,2474	0,1663	0,2197	0,4202	<b>0,1400</b>	0,0000
<b>FRESH</b>	0,2103	0,1187	0,2197	0,4435	0,0000	<b>0,1600</b>
<b>WEB</b>	0,1502	0,1009	0,2533	0,4435	<b>0,1300</b>	0,0000
<b>FRESH</b>	<b>0,1276</b>	<b>0,0729</b>	<b>0,2533</b>	<b>0,4566</b>	0,0000	<b>0,1500</b>

# Wide pFound для Свежести

Документ	fpLook	wpLook	fpFound	wpFound	wRel	fRel
<b>WEB</b>	1,0000	1,0000	0,0000	0,0000	<b>0,2000</b>	0,0000
<b>WEB</b>	0,8500	0,6800	0,0000	0,2000	<b>0,1800</b>	0,0000
<b>WEB</b>	0,7225	0,4624	0,0000	0,3224	<b>0,1600</b>	0,0000
<b>WEB</b>	0,6141	0,3144	0,0000	0,3964	<b>0,1500</b>	0,0000
<b>WEB</b>	0,5220	0,2138	0,0000	0,4435	<b>0,1400</b>	0,0000
<b>WEB</b>	0,4437	0,1454	0,0000	0,4735	<b>0,1300</b>	0,0000
<b>WEB</b>	0,3771	0,0989	0,0000	0,4924	<b>0,1200</b>	0,0000
<b>WEB</b>	0,3206	0,0672	0,0000	0,5042	<b>0,1100</b>	0,0000
<b>WEB</b>	0,2725	0,0457	0,0000	0,5116	<b>0,1000</b>	0,0000
<b>WEB</b>	<b>0,2316</b>	<b>0,0311</b>	<b>0,0000</b>	<b>0,5162</b>	<b>0,0900</b>	0,0000



# Wide pFound для Свежести

Документ	fpLook	wpLook	fpFound	wpFound	wRel	fRel
<b>FRESH</b>	1,0000	1,0000	0,0000	0,0000	0,0000	<b>0,2000</b>
<b>FRESH</b>	0,6800	0,8500	0,2000	0,0000	0,0000	<b>0,1800</b>
<b>FRESH</b>	0,4740	0,7225	0,3224	0,0000	0,0000	<b>0,1600</b>
<b>FRESH</b>	0,3384	0,6141	0,3982	0,0000	0,0000	<b>0,1500</b>
<b>FRESH</b>	0,2445	0,5220	0,4490	0,0000	0,0000	<b>0,1400</b>
<b>FRESH</b>	0,1787	0,4437	0,4832	0,0000	0,0000	<b>0,1300</b>
<b>FRESH</b>	0,1322	0,3771	0,5065	0,0000	0,0000	<b>0,1200</b>
<b>FRESH</b>	0,0989	0,3206	0,5223	0,0000	0,0000	<b>0,1100</b>
<b>FRESH</b>	0,0748	0,2725	0,5332	0,0000	0,0000	<b>0,1000</b>
<b>FRESH</b>	<b>0,0572</b>	<b>0,2316</b>	<b>0,5407</b>	<b>0,0000</b>	0,0000	<b>0,0900</b>

# Wide pFound для Свежести

- › До 2015 года fresh intent probability подбирался только по ассессорским оценкам
- › До 2015 года все колдунщики подмешивались по wPfound

Блендер

«Спектр»



# Разнообразие поисковой выдачи

Технология «Спектр» (2010)

Плахов А. Поисковая технология «Спектр»

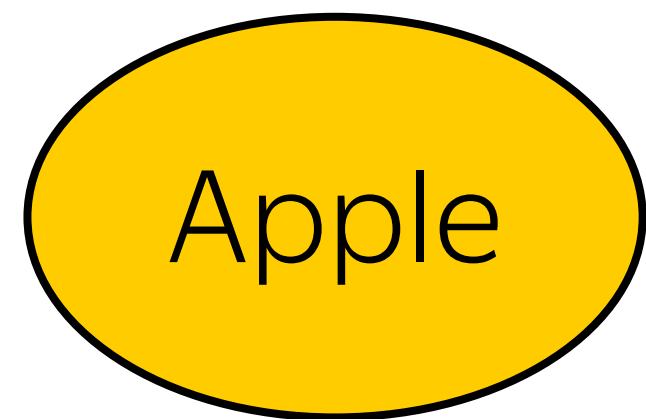
<https://events.yandex.ru/lib/talks/12/>

# Разнообразие поисковой выдачи

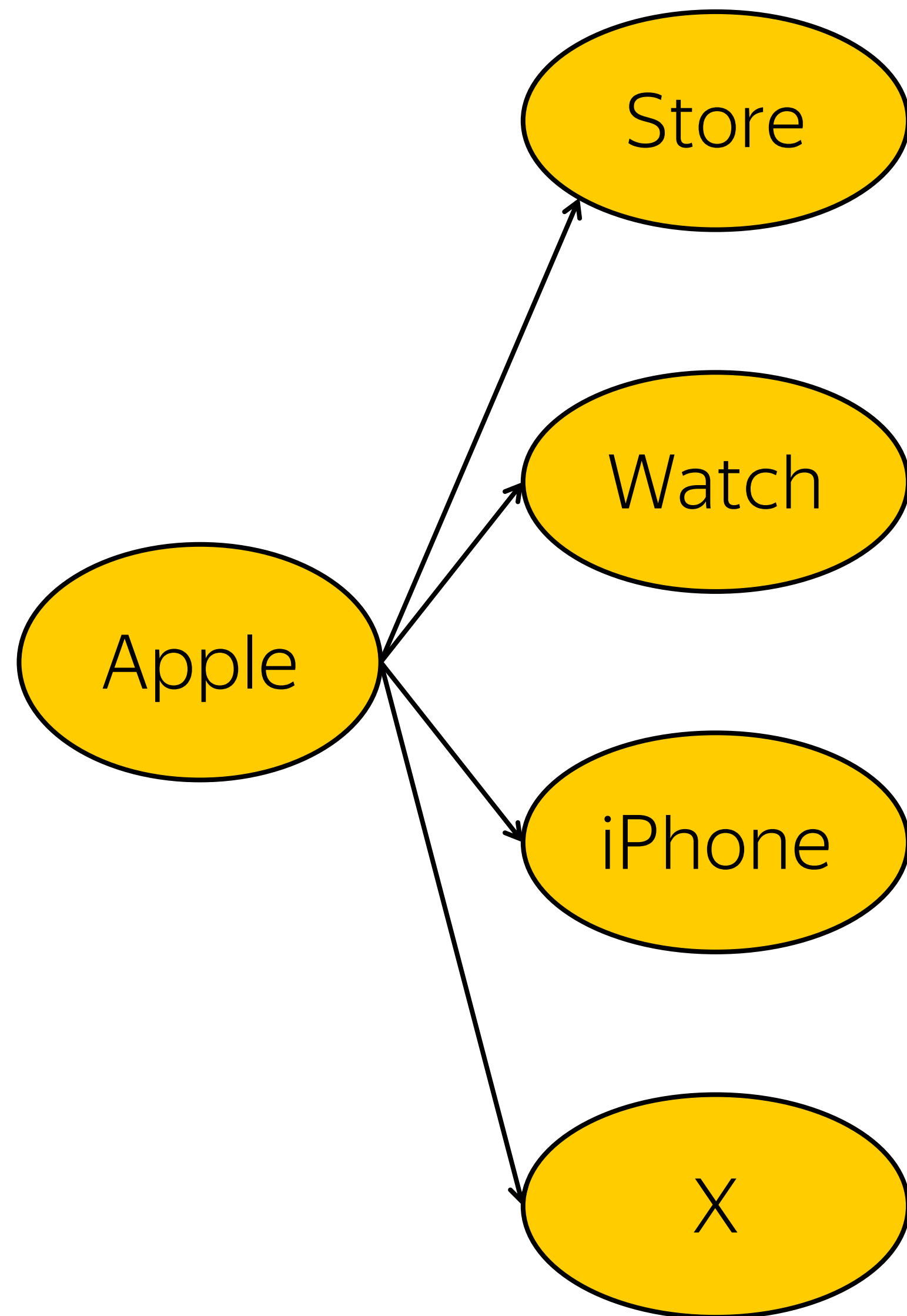
## Технология «Спектр»

- › Будем определять, что мог иметь в виду пользователь, по продолжениям введённого запроса
- › Научимся классифицировать продолжения по различным тематикам
- › Тематики станут нашими «интентами»
- › Вероятности будем определять по частоте соответствующих продолжений запросов

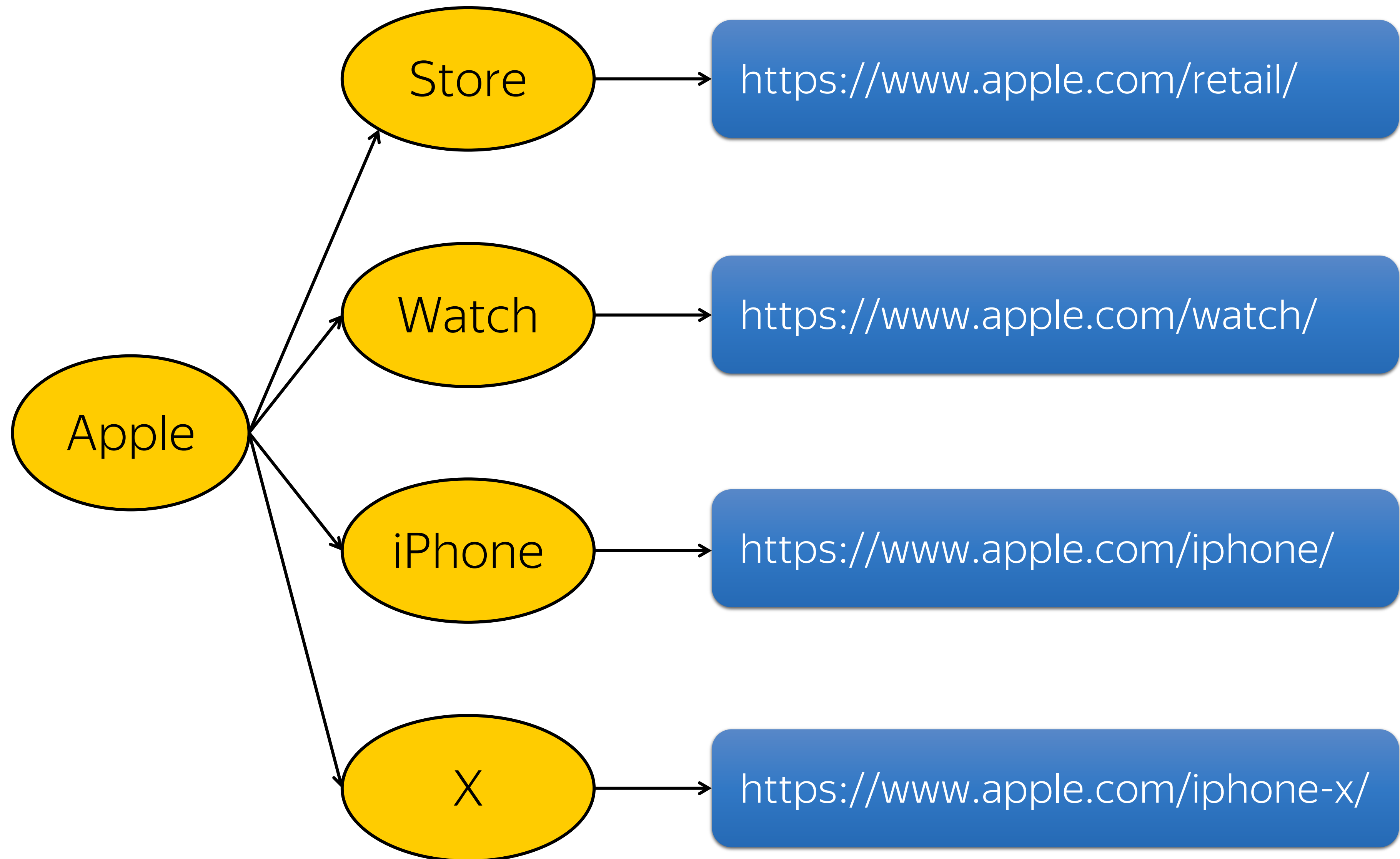
# Разнообразие поисковой выдачи



# Разнообразие поисковой выдачи

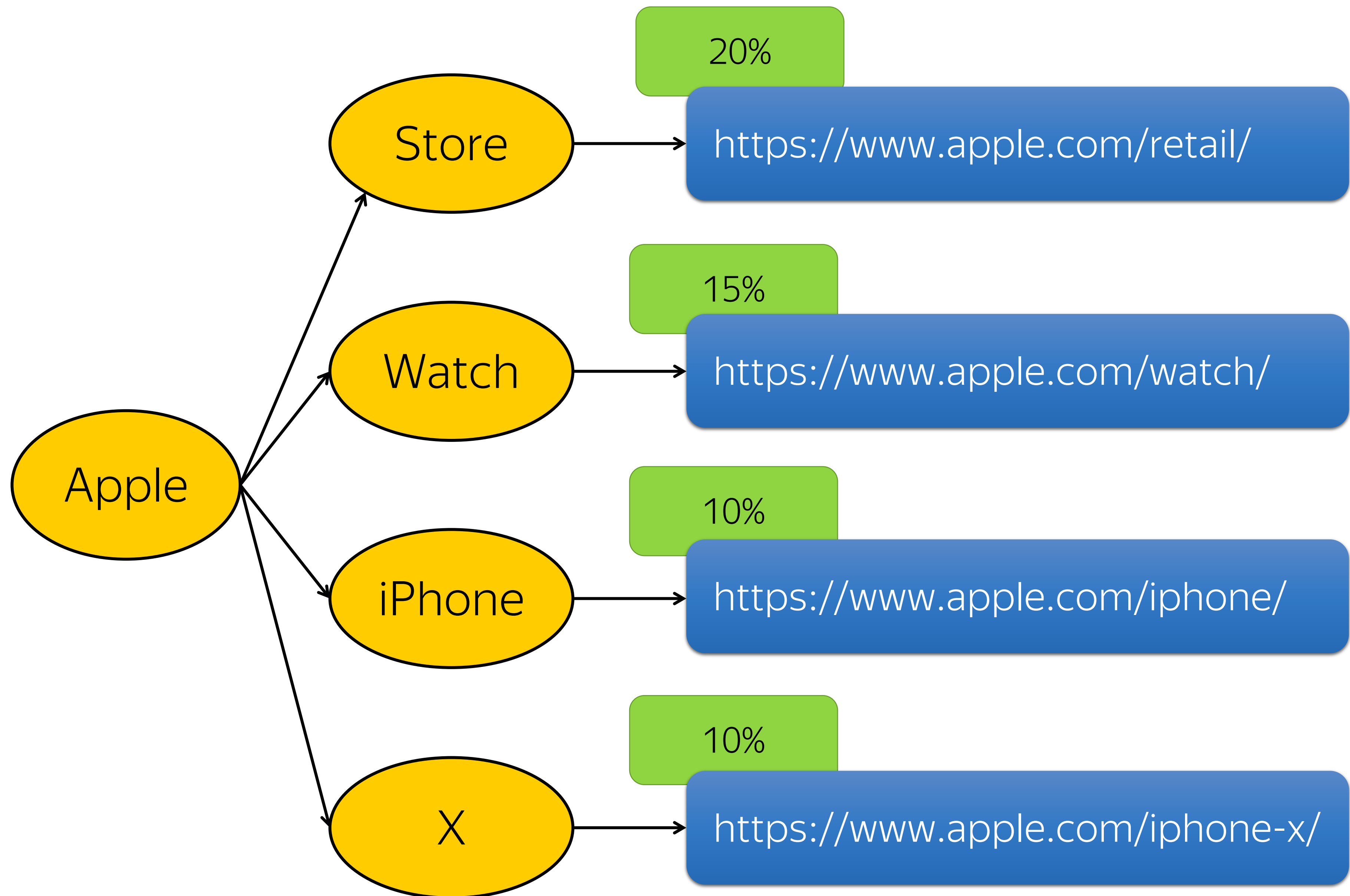


# Разнообразие поисковой выдачи





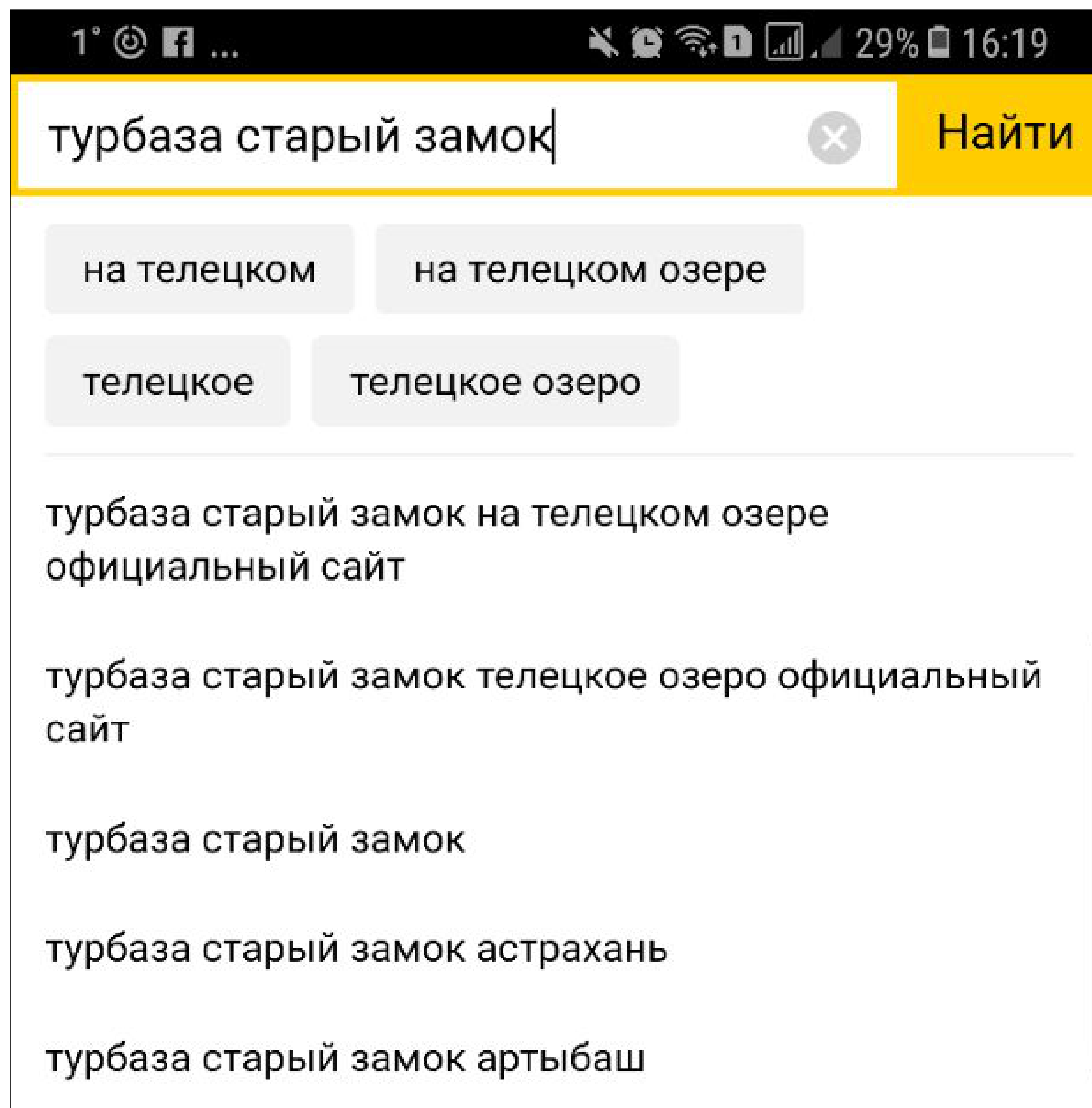
# Разнообразие поисковой выдачи



# Разнообразие поисковой выдачи

турбаза старый замок **алтай**  
турбаза старый замок **астрахань**  
турбаза старый замок **на телецком**  
турбаза старый замок **святогорск**  
турбаза старый замок **телецкое**  
турбаза старый замок **телецкое озеро**  
турбаза старый замок **телецкое адрес**  
турбаза старый замок **телецкое телефон**

# Разнообразие поисковой выдачи



Блендер

# Обучение по кликам



# Обучение по кликам

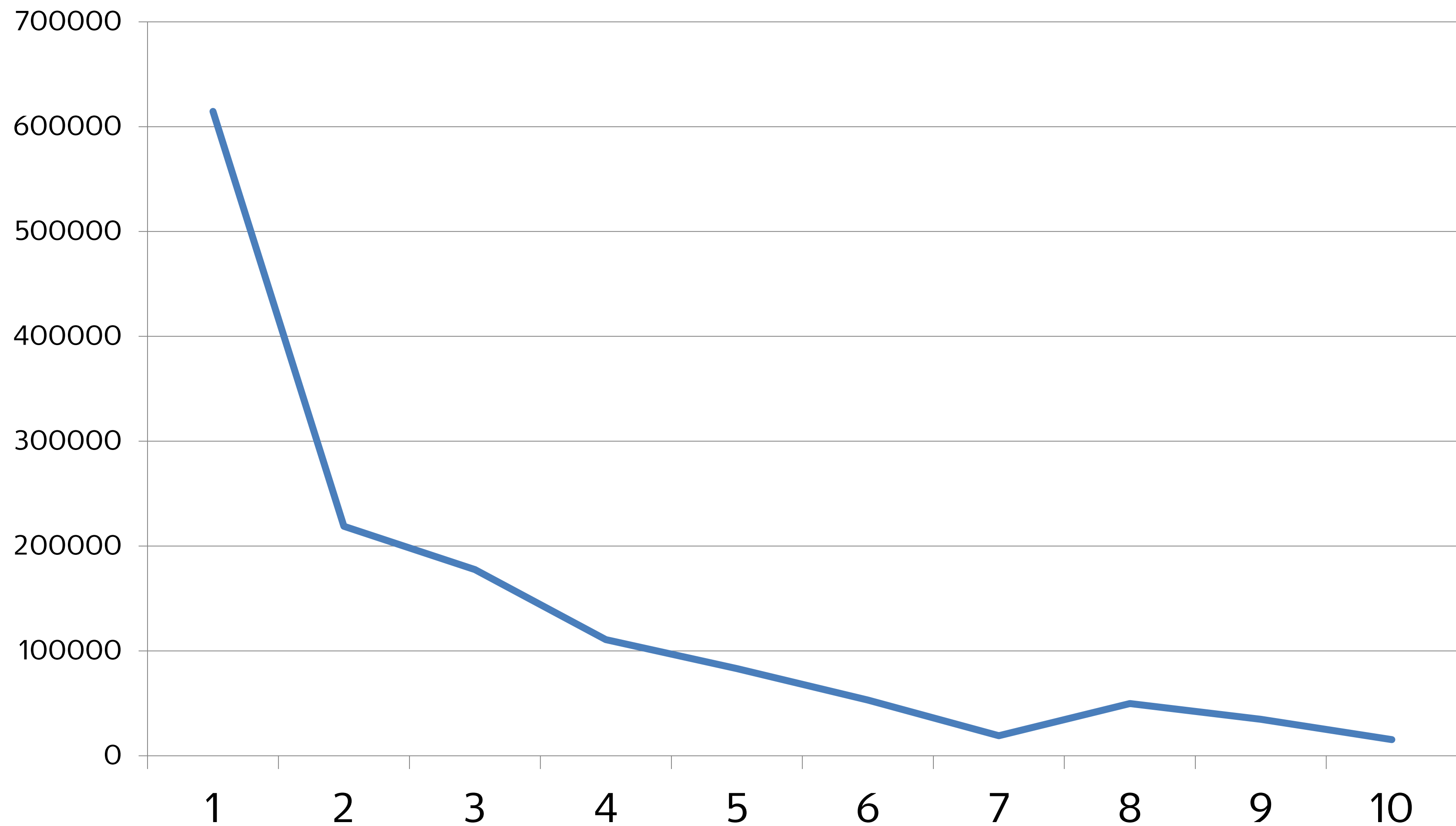
## Обучение разнообразия по кликам

Пусть есть задача: найти оптимальную позицию для некоторого конкретного колдунщика (например, картинок)

«Качество» позиции определяется некоторой SERP-wide метрикой. Например, общим количеством кликов

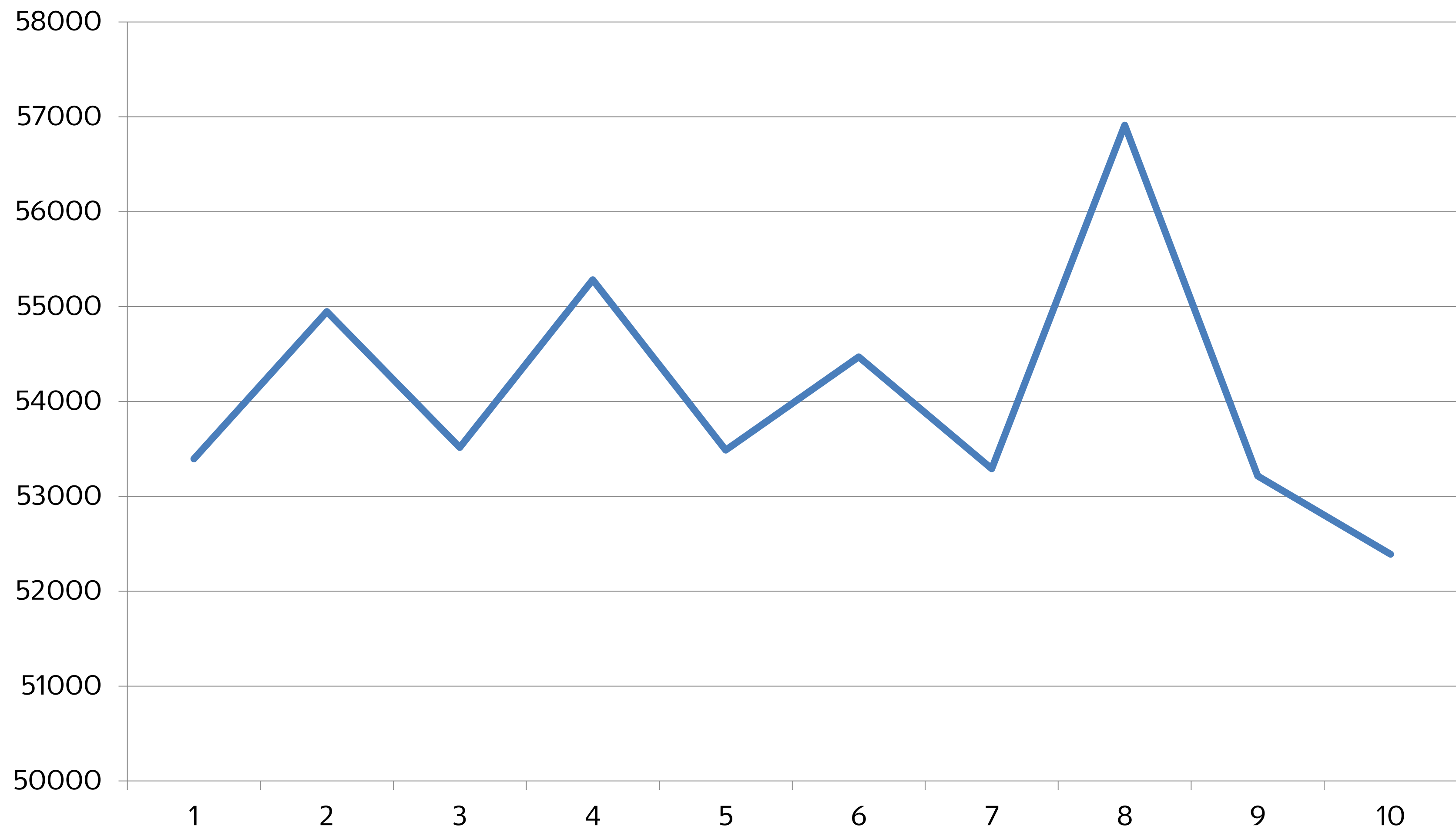
# Обучение по кликам

Количество взаимодействий с колдунщиком в зависимости от позиции



# Обучение по кликам

**Количество кликов по выдаче в зависимости от  
позиции колдунщика**



# Обучение по кликам

Обучение разнообразия по кликам

Получаем задачу «многоруких контекстных бандитов»



# Обучение по кликам

Обучение разнообразия по кликам

Получаем задачу «многоруких контекстных бандитов»

Бандиты – позиции для колдунщика

# Обучение по кликам

Обучение разнообразия по кликам

Получаем задачу «многоруких контекстных бандитов»

Бандиты – позиции для колдунщика

Контекст – поисковые факторы

# Обучение по кликам

Обучение разнообразия по кликам

Получаем задачу «многорукых контекстных бандитов»

Бандиты – позиции для колдунщика

Контекст – поисковые факторы

Решающая функция:

$$f(q) = \arg \max_p g(p, features(p))$$

# Обучение по кликам

Обучение разнообразия по кликам

Решающая функция:

$$f(q) = \arg \max_p g(p, features(p))$$

Сложность заключается в том, что выборку мы умеем собирать для функции  $g$ , а не для функции  $f$ !

# Обучение по кликам

## Стратегии получения выборки

1. Total random: колдунщик каждый раз показывается на абсолютно случайной позиции
2. Local random: колдунщик показывается на случайной позиции, которая на 1-2 отличается от позиции в продакшене
3. Accurate random: позиция случайная, вероятность появления на определённой позиции зависит от позиции в продакшене

# Offline replay

Пусть есть наборы  $\{\langle q_i, pos_i, f_i, p_i \rangle\}_{i=1}^n$

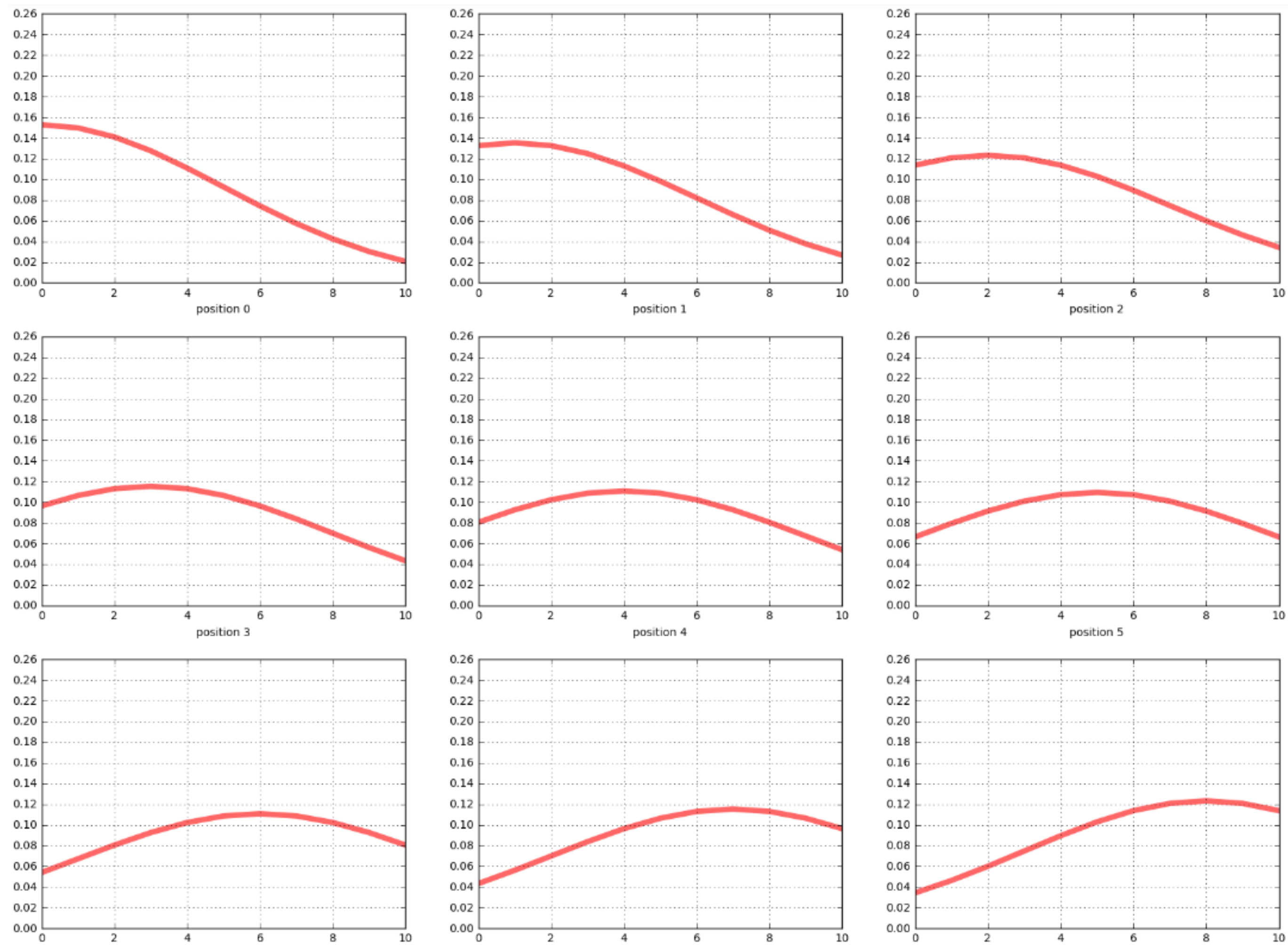
- ›  $q_i$  – функция качества (e.g. количество кликов по выдаче)
- ›  $pos_i$  – позиция колдунщика
- ›  $f_i$  – поисковые факторы для запроса
- ›  $p_i$  – вероятность показа именно на этой позиции

# Offline replay

- ›  $p_i = \frac{1}{10}$  – «глобальный» рандом
- ›  $p_i = s(F(f_i), pos_i)$  – «локальный»/«аккуратный» рандом,  
здесь  $F(f_i)$  – позиция, предсказанная продакшен-  
формулой

# Offline replay

## «Аккуратный» рандом: пример





# Offline replay

Пусть есть новая формула  $F'$

Оценка её качества:

$$EQ(F') = \frac{\sum_{i=1}^n [pos_i = F'(f_i)] \cdot \frac{q_i}{p_i}}{\sum_{i=1}^n \frac{[pos_i = F'(f_i)]}{p_i}}$$

Проблема – статистическая значимость принимаемых решений

# Offline replay

## Counterfactual replay

- › Swaminathan A., Joachims T. The Self-Normalized Estimator for Counterfactual Learning

<http://papers.nips.cc/paper/5748-the-self-normalized-estimator-for-counterfactual-learning.pdf>

# Offline replay

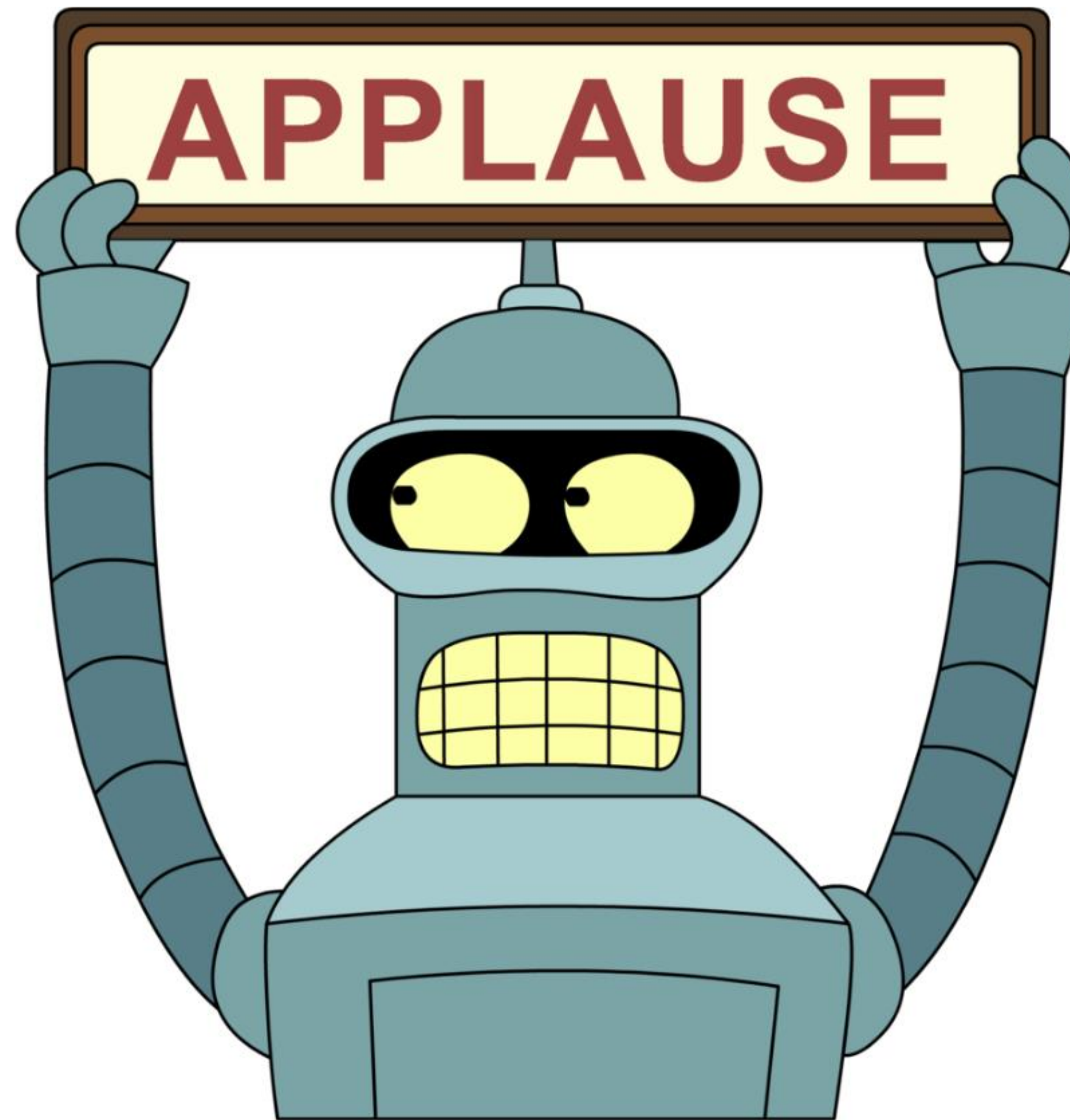
## Counterfactual replay

Для каждой решающей функции  $g$  и фактического показа  $p_i$  можно определить «вероятность»  $P(g, p_i)$  получения позиции  $p_i$  при помощи функции  $g$

С этим показом также связан некоторый reward  $Q_i$

Тогда задача обучения – оптимизация:

$$\sum_{i=1}^n P(g, p_i) Q_i \rightarrow \max$$



Алексей Шаграев



<https://www.facebook.com/ashagraev>



<https://vk.com/shagraev>



<https://habrahabr.ru/users/ashagraev/posts/>