



Yandex  
CatBoost

# Ранжирование в CatBoost

Иван Лыжин,  
Разработчик CatBoost

# CatBoost

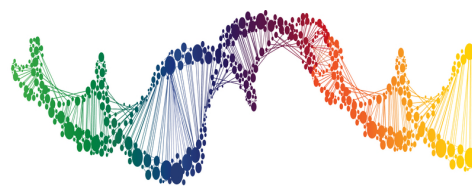


# Тип данных

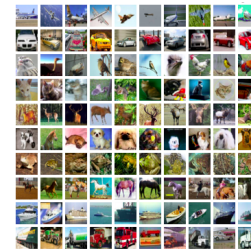
## Однородные данные



Музыка



ДНК



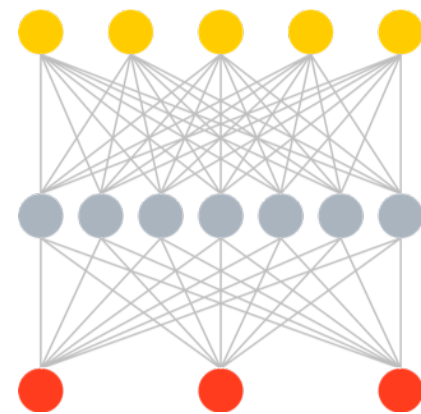
Изображения



Текст



Глубокие нейронные сети



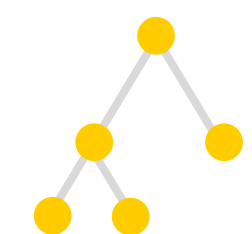
## Неоднородные данные

Сгенерированные экспертом фичи

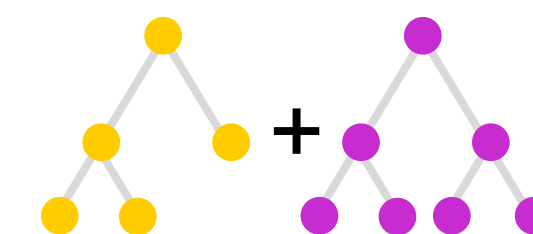
Music track length	Year	Rating	Label
2	1990	3	1
3	1950	5	0
15	1970	4	1



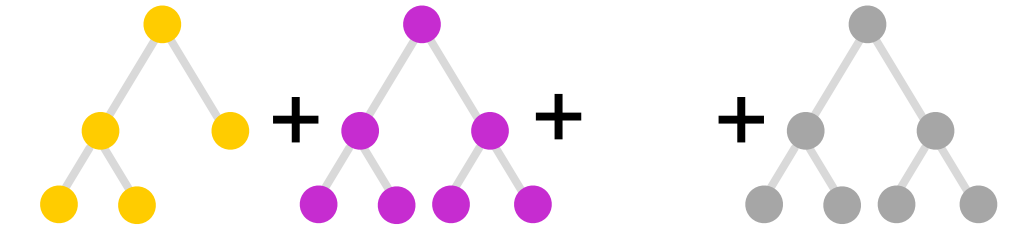
Градиентный бустинг на деревьях решений



Big error

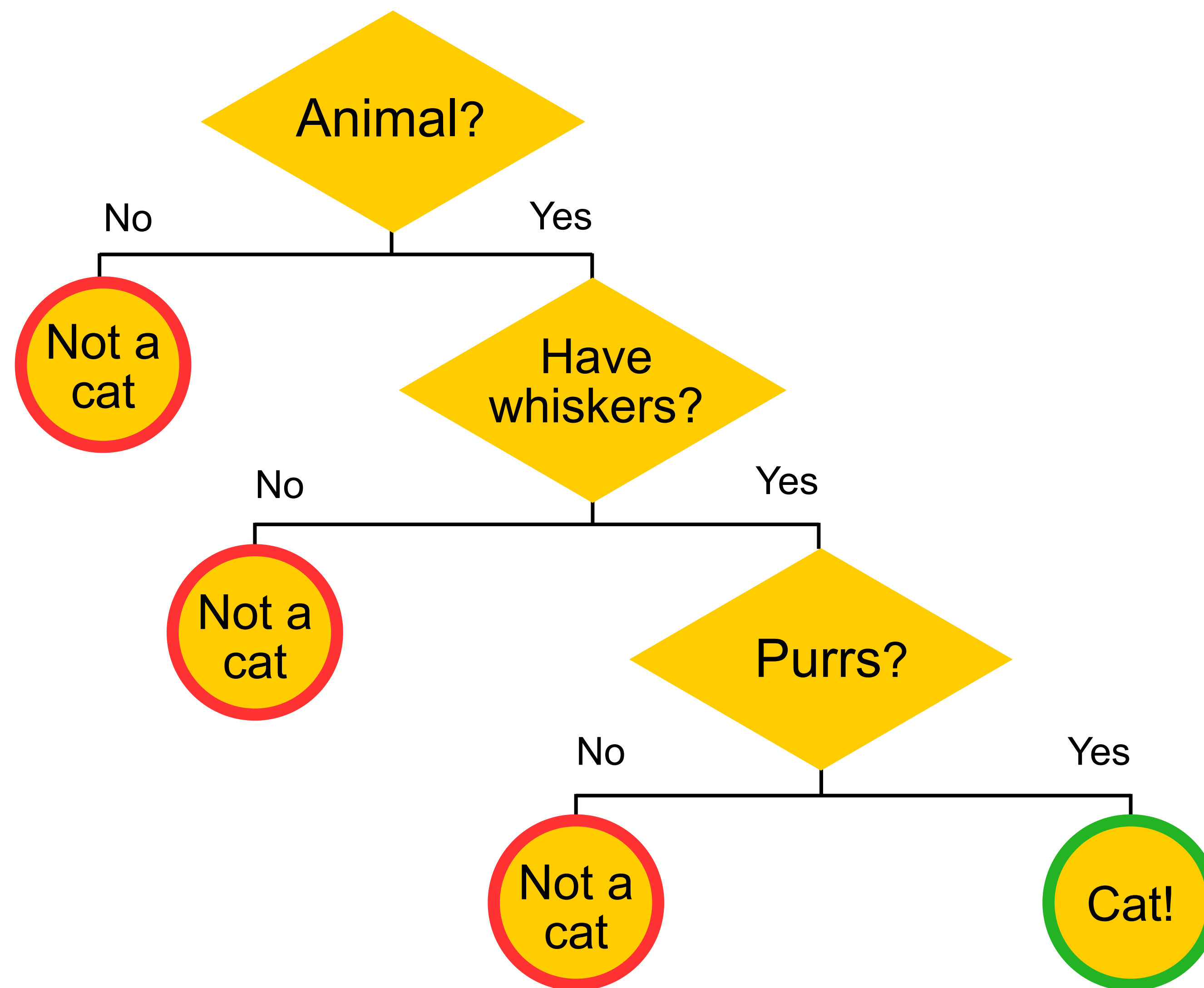


Better



Ship it

# Неоднородные данные? Деревья решений!

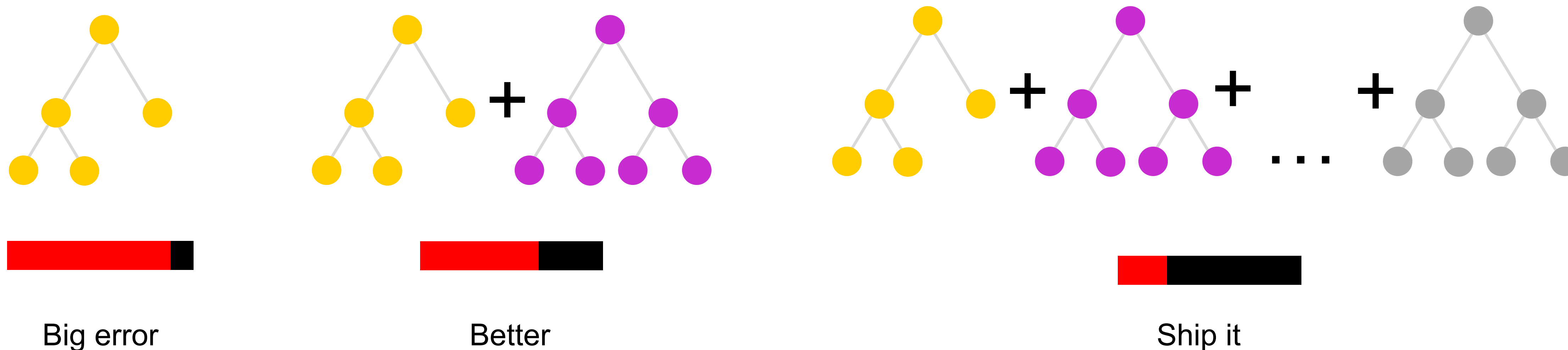


# Градиентный бустинг на деревьях решений

State-of-the-art для табличных разнородных данных

Легко использовать, простой подбор параметров

Показывает хорошие результаты как на маленьких данных, так и на задачах big data



# Библиотеки градиентного бустинга

*dmlc*  
***XGBoost***



Yandex  
CatBoost



Microsoft

LightGBM

# CatBoost в Яндексе

Yandex.Zen

Yandex.Music

Yandex.Self-Driving Cars

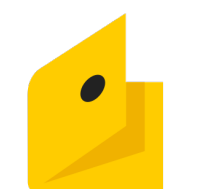
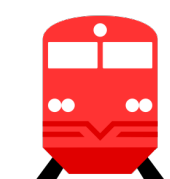
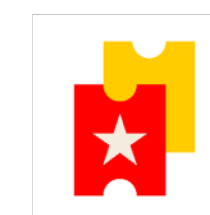
Yandex.Search

Yandex.Ads

Yandex.Weather

Yandex Alice

Практически везде!



# CatBoost в мире

Классификация частиц  
в CERN

Предсказание места назначения в  
Careem taxi service

Ранжирование отелей  
в Aviasales

Защита от ботов в CloudFlare

Медицинские исследования  
в University of NSW Sydney

Netflix

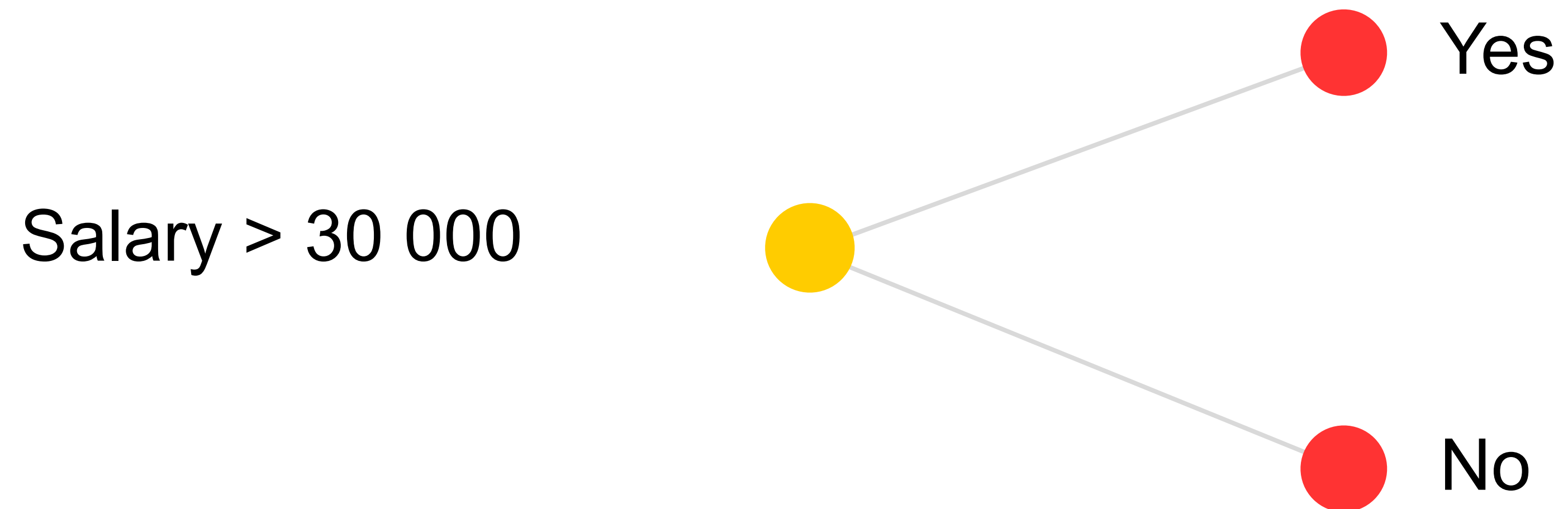
Соревнования на Kaggle



**NETFLIX**



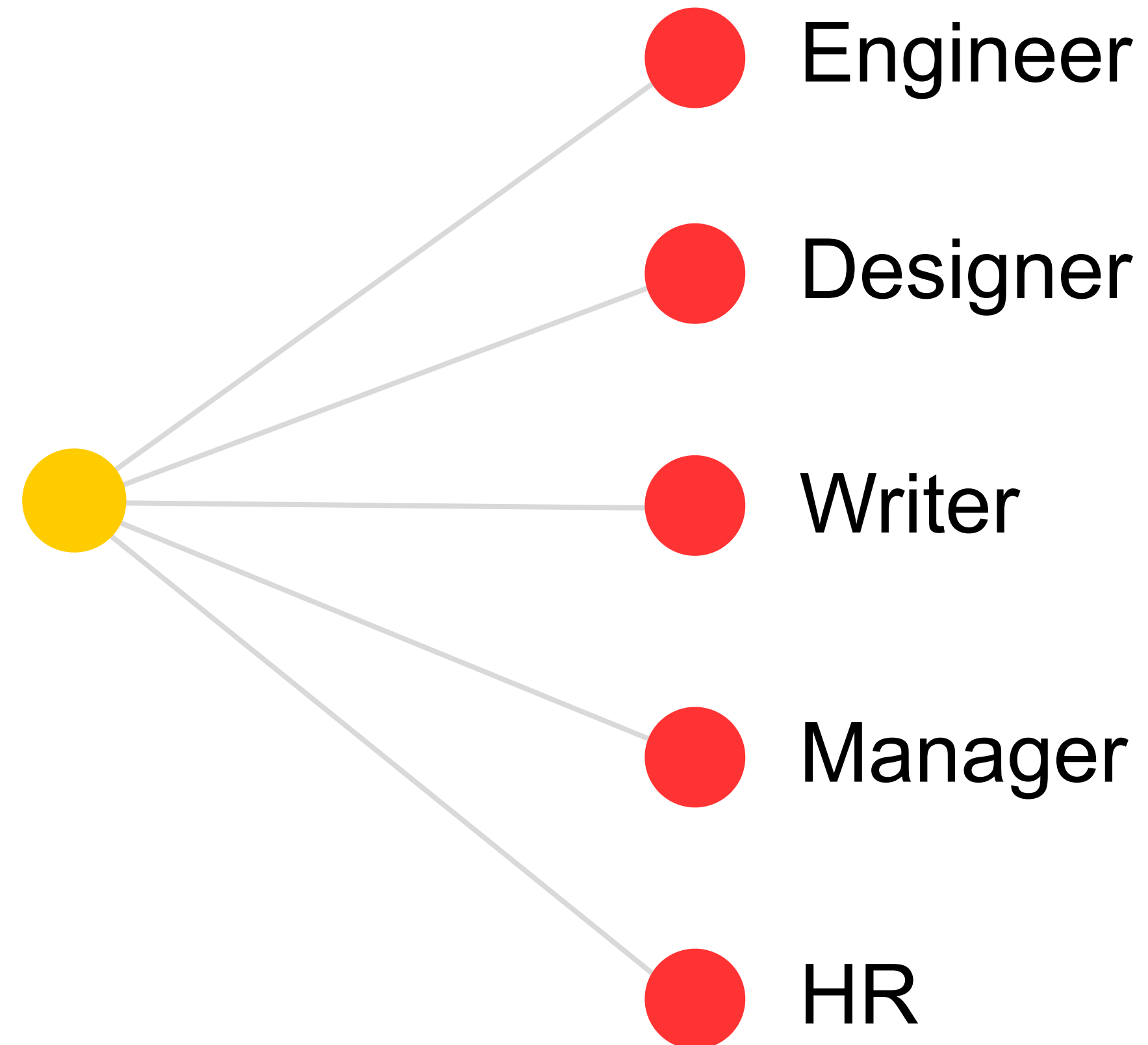
# Числовые фичи



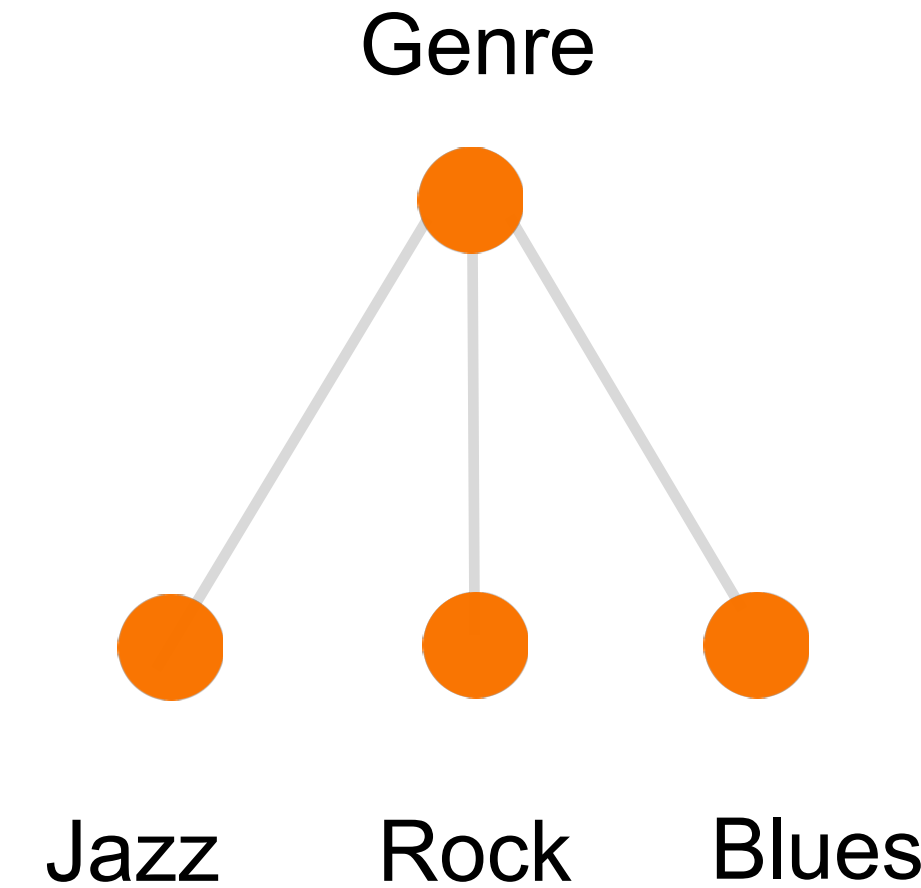
# Категориальные фичи

Categorical data

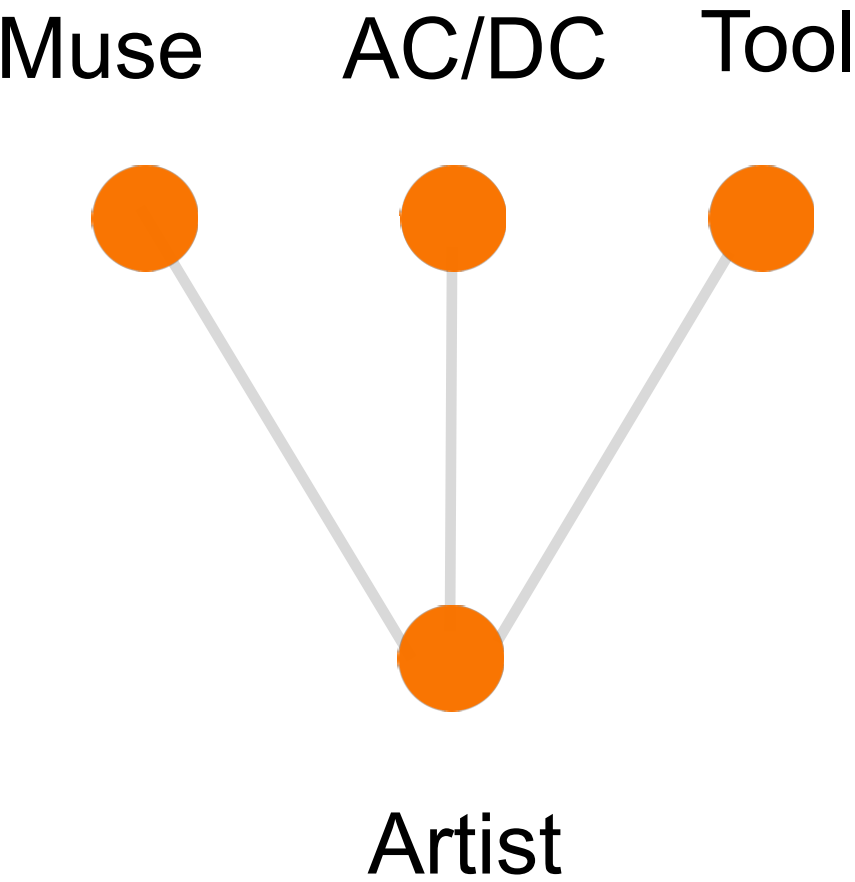
Occupation



# Обработка категориальных данных



One-hot encoding



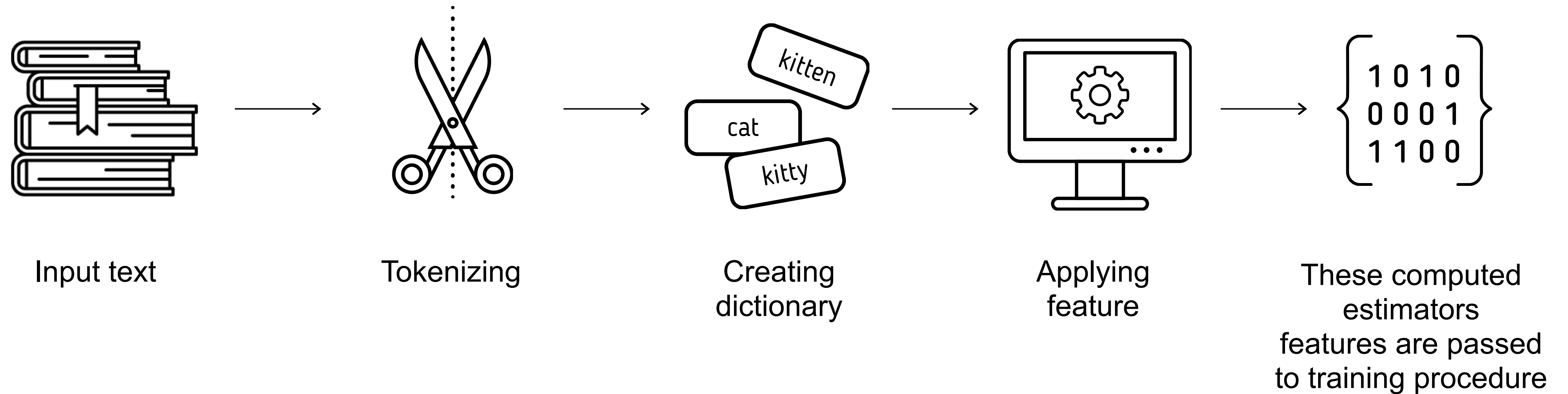
Statistics based on category

↓  
Category-based

↓  
Label based

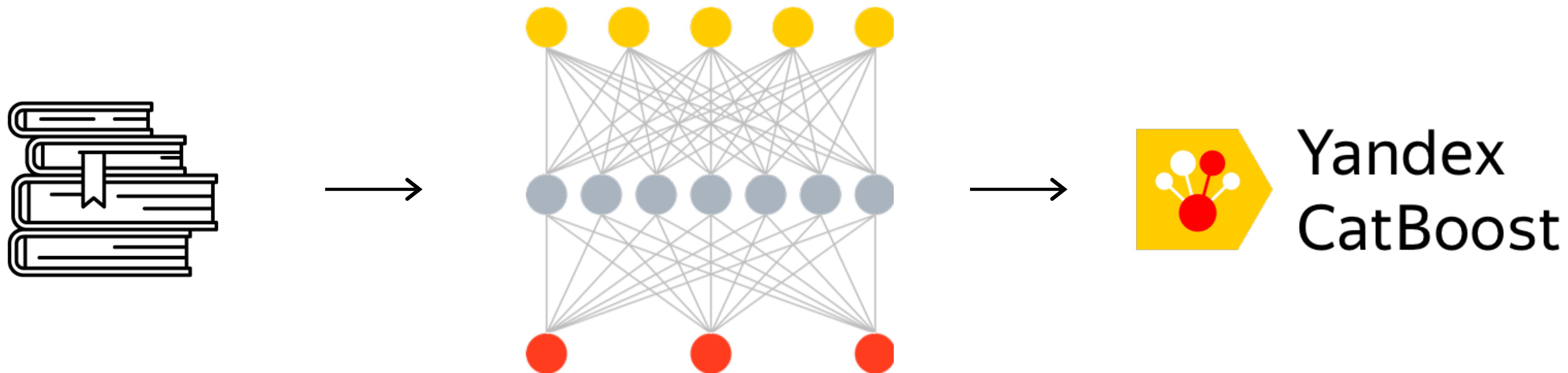
Greedy search for combinations

# Текстовые фичи



\* Пока только для задач классификации

# Embeddings - stay tuned



# Достоинства CatBoost

- Отличное качество с дефолтными параметрами

- Поддержка категориальных и текстовых данных

- Скорость обучения и применения

- Инструменты для анализа модели

# Ранжирование



# Яндекс.Поиск

## Задача

- › Предсказание порядка документов в поисковой выдаче

## Тип задачи: ранжирование

## Фичи датасета:

- › Классические фичи (PageRank, BM25 и др.)
- › Выходы нейронных сетей

## CatBoost:

- › YetiRankPairwise лосс
- › Распределённое обучение GPU
- › Смесь моделей
- › Анализ важности фич
- › Отбор фич
- › Анализ ранжирования

The screenshot shows the Yandex search engine interface. The search bar at the top contains the text 'catboost'. Below the search bar, there are tabs for 'Web', 'Images', 'Video', 'News', 'Translate', 'Disk', 'Mail', and 'All'. The 'Web' tab is selected. The search results show 20 thousand results found. The first result is 'CatBoost - open-source gradient boosting library' from catboost.yandex. The second result is 'CatBoost · GitHub' from github.com. The third result is 'CatBoost — Yandex Technologies' from tech.yandex.com. The fourth result is 'CatBoost — Overview of CatBoost — Yandex Technologies' from tech.yandex.com. The fifth result is 'Newest 'catboost' Questions - Stack Overflow' from stackoverflow.com. The sixth result is 'CatBoost — Технологии Яндекса' from tech.yandex.ru. The seventh result is 'Яндекс открывает технологию машинного...' from habrahabr.ru.

Yandex catboost X Search

Web Images Video News Translate Disk Mail All

**CatBoost - open-source gradient boosting library** 20 thousand results found  
catboost.yandex  
CatBoost is an algorithm for gradient boosting on decision trees. ... New version of CatBoost has industry fastest inference implementation.

**CatBoost · GitHub**  
github.com > CatBoost  
CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.

**CatBoost — Yandex Technologies**  
tech.yandex.com > CatBoost  
CatBoost is a state-of-the-art open-source gradient boosting on decision trees library. Developed by Yandex researchers and engineers...

**CatBoost — Overview of CatBoost — Yandex Technologies**  
tech.yandex.com > CatBoost > Documentation  
CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. It is available as an open source library.

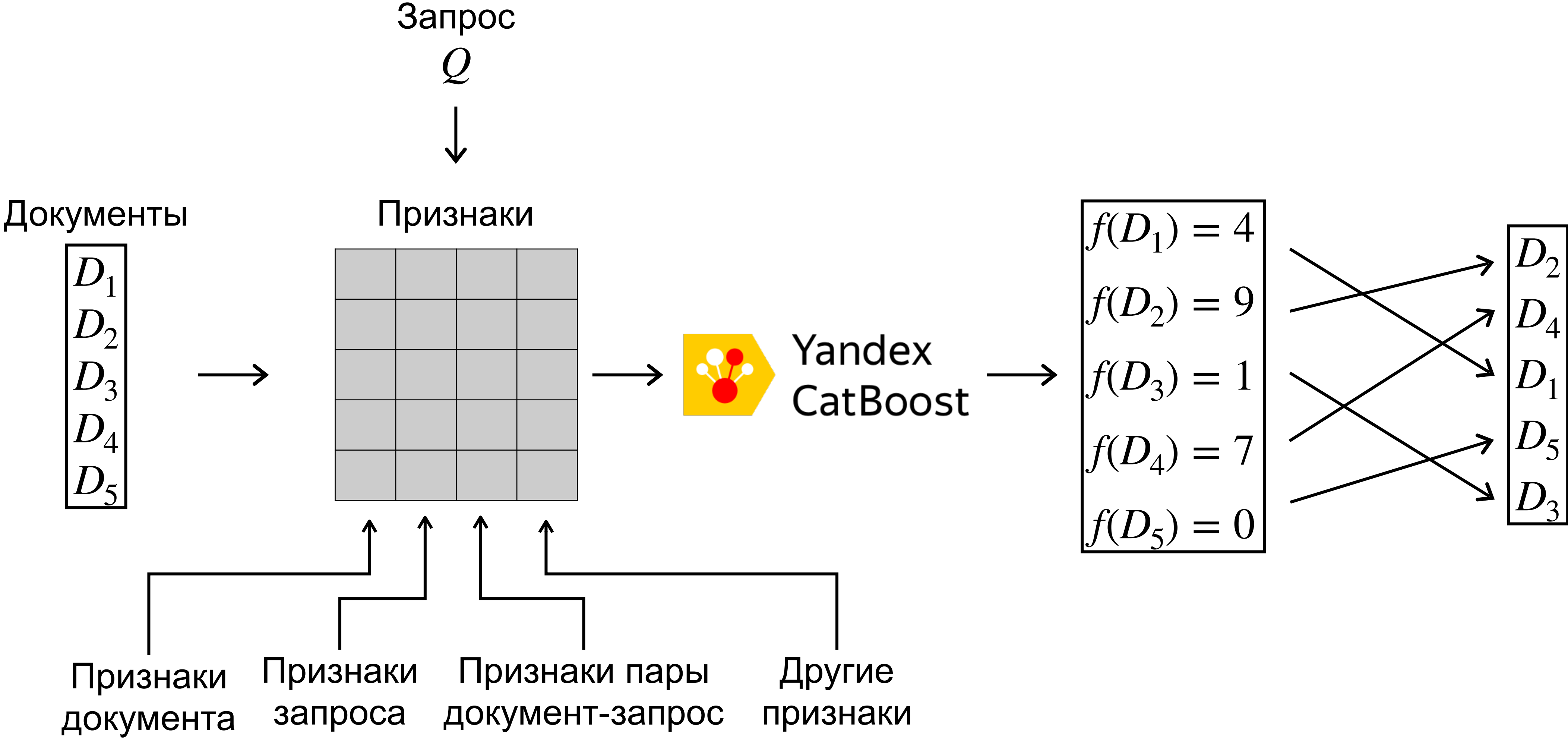
**Newest 'catboost' Questions - Stack Overflow**  
stackoverflow.com > Catboost  
CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.

**CatBoost — Технологии Яндекса**  
tech.yandex.ru > CatBoost  
CatBoost использует более универсальный алгоритм, поэтому она подходит для решения и других задач. Преимущества CatBoost.

**Яндекс открывает технологию машинного... / Хабрахабр**  
habrahabr.ru > Яндекс > Блог компании Яндекс > 333522  
CatBoost — это новый метод машинного обучения, основанный на градиентном



# Задача ранжирования



# Обучающая выборка

GroupId

1
1
1
2
2
2
2
3
3
3

+

Признаки


+

Оценка  
релевантности

0.6
0.9
1
0.6
0.1
0
0.4
1
0.4
0.2

сгруппирован!

# Метрики качества ранжирования

DCG/NDCG

PFound

PrecisionAt

RecallAt

FilteredDCG

AverageGain

MAP

AUC

PairAccuracy

# DCG/NDCG

$$DCG = \sum_{i=1}^n \frac{2^{y(d_i)} - 1}{\log_2(i + 1)}$$

Высокая релевантность → большой вклад

Низкая позиция → меньший вклад

$$NDCG = \frac{DCG}{maxDCG}$$

Нормализованный DCG проще сравнивать

# PFound

$$PFound = \sum_{i=1}^n y(d_i) * decay^{i-1} * \prod_{j<i} (1 - y(d_j))$$

Вероятность найти ответ  
в текущем документе

Вероятность продолжить  
поиск после просмотра  
предыдущих документов

Вероятность не найти  
ответ среди предыдущих  
документов

# PFound

catboost

Web

Images

Video


News

Translate

Disk


Mail

All

 **CatBoost** - open-source gradient boosting library


[catboost.yandex](#)

CatBoost is an algorithm for gradient boosting on decision trees. ... New version of CatBoost has industry fastest inference implementation.

 **CatBoost** · GitHub


[github.com](#) > CatBoost

CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.

 **CatBoost** — Yandex Technologies

[tech.yandex.com](#) > CatBoost

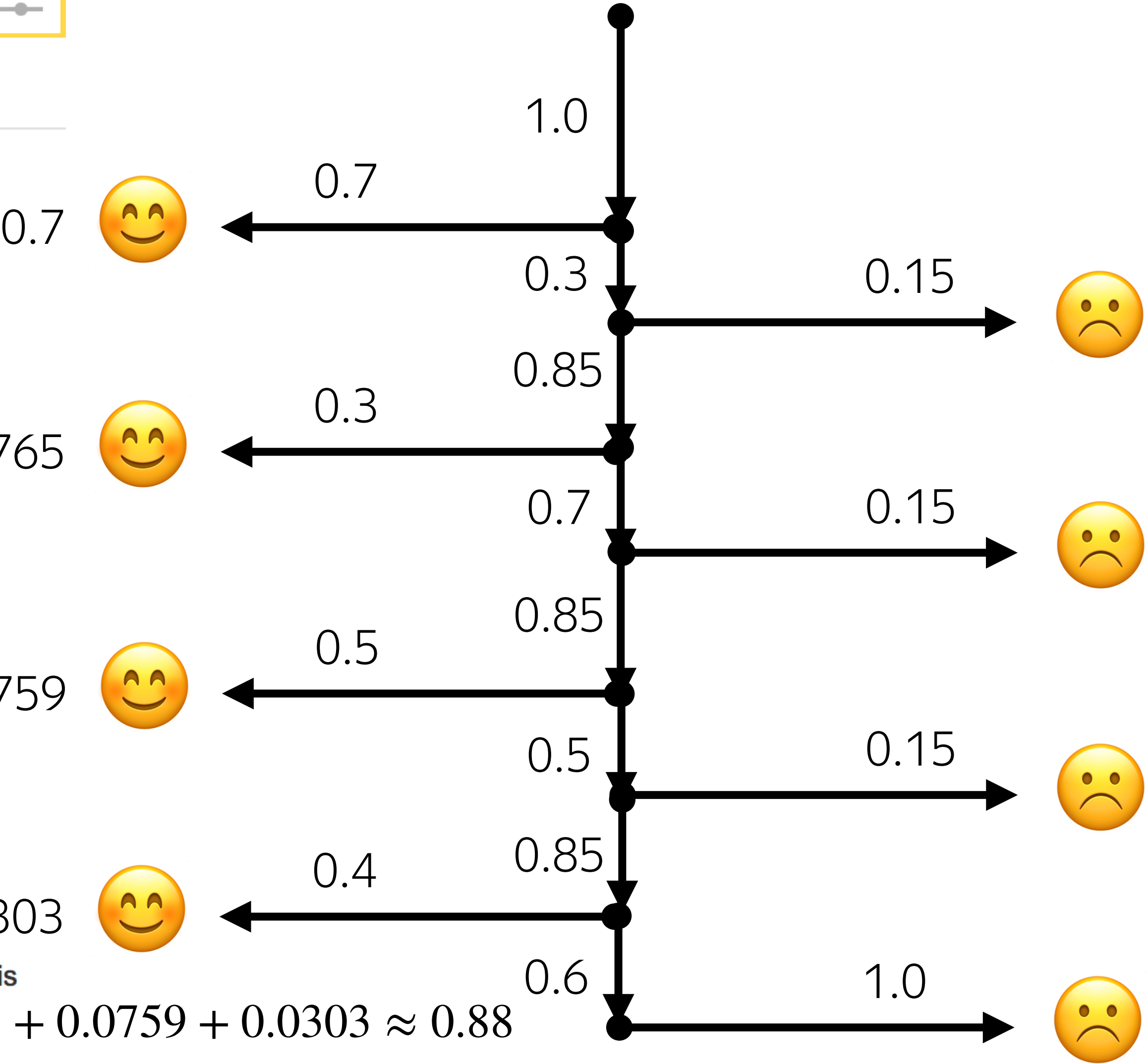
CatBoost is a state-of-the-art open-source gradient boosting on decision trees library. Developed by Yandex researchers and engineers...

 **CatBoost** — Overview of CatBoost — Yandex Technologies

[tech.yandex.com](#) > CatBoost > Documentation

CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. It is available as an open source library.

$$PFound = 0.7 + 0.0765 + 0.0759 + 0.0303 \approx 0.88$$





# PrecisionAt

Yandex

catboost

Web

Images

Video


News

Translate


Disk


Mail

All





**CatBoost** - open-source gradient boosting library  
[catboost.yandex](#) ▾  
CatBoost is an algorithm for gradient boosting on decision trees. ... New version of CatBoost has industry fastest inference implementation.







**CatBoost** · GitHub  
[github.com](#) > CatBoost ▾  
CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.







**CatBoost** — Yandex Technologies  
[tech.yandex.com](#) > CatBoost ▾  
CatBoost is a state-of-the-art open-source gradient boosting on decision trees library. Developed by Yandex researchers and engineers...







**CatBoost** — Overview of CatBoost — Yandex Technologies  
[tech.yandex.com](#) > CatBoost > Documentation ▾  
CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. It is available as an open source library.







**Newest 'catboost' Questions - Stack Overflow**  
[stackoverflow.com](#) > Catboost ▾  
CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.






**CatBoost** — Технологии Яндекса  
[tech.yandex.ru](#) > CatBoost ▾  
CatBoost использует более универсальный алгоритм, поэтому она подходит для решения и других задач. Преимущества CatBoost.





**Яндекс открывает технологию машинного... / Хабрахабр**  
[habrahabr.ru](#) > Яндекс > Блог компании Яндекс > 333522 ▾  
CatBoost – это новый метод машинного обучения, основанный на градиентном



$$Precision@k = \frac{|Relevant|}{k} = \frac{5}{7}$$

# Ранжирующие режимы в CatBoost

QueryRMSE

QueryCrossEntropy

QuerySoftMax

StochasticRank

PairLogit

PairLogitPairwise

YetiRank

YetiRankPairwise



# QueryRMSE

$N$  - число объектов

$t_i$  - истинная релевантность для объекта

$a_i$  - предсказание для объекта

RMSE: 
$$\mathcal{L} = \sqrt{\frac{1}{N} \sum_{i=1}^N (a_i - t_i)^2}$$

RMSE:  $a_i \longrightarrow t_i$

$a_i$	$t_i$
1	7
2	8
3	9

$RMSE=10.4$

Вычтем среднее из таргета и аппрокса

$$a_i - \frac{\sum_{j \in group(i)} a_j}{groupSize(i)} \longrightarrow t_i - \frac{\sum_{j \in group(i)} t_j}{groupSize(i)}$$

# QueryRMSE

$$\mathcal{L} = \sqrt{\frac{1}{N} \sum_{group \in G} \sum_{i \in group} (a_i - t_i - A(group))^2}$$

$$A(group) = \frac{1}{|group|} \sum_{j \in group} (a_j - t_j)$$

$a_i$	$t_i$
1	7
2	8
3	9

QueryRMSE=0

$N$  - число объектов

$t_i$  - истинная метка для объекта

$a_i$  - предсказание для объекта

# Пары вместо оценки релевантности

Идентификатор  
группы/запроса

1
1
1
2
2
2
2
3
3
3

+

Признаки


+

Пары  
winner-loser

2	0
1	2
3	5
5	6
4	3
4	5
7	8
9	7

сгруппирован!

# PairLogit

$$P(i \succ j) = \sigma(a_i - a_j) = \frac{e^{a_i}}{e^{a_i} + e^{a_j}}$$

$$\mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right)$$

$P$  — набор пар (победитель и проигравший)

$p, n$  — индексы победившего и проигравшего объекта

Мы хотим правильно классифицировать все пары

Пары объектов задаются пользователем или генерируются перед обучением на основании оценок релевантности

# PairLogit и PairLogitPairwise

$$\mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right) + \text{Поэлементная оценка листьев} = \text{PairLogit}$$

$$\mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right) + \text{Попарная оценка листьев} = \text{PairLogitPairwise}$$

# YetiRank

$$P(i \succ j) = \sigma(a_i - a_j) = \frac{e^{a_i}}{e^{a_i} + e^{a_j}} \quad \mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right)$$

$P$  — набор пар (победитель и проигравший)

$p, n$  — индексы победившего и проигравшего объекта

Мы хотим правильно классифицировать все пары

Пары объектов, по которым происходит оптимизация, генерируются на каждой итерации на основании текущих предсказаний модели.

# YetiRank и YetiRankPairwise

$$\mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right)$$

+

Поэлементная  
оценка листьев

Генерация пар

=

**YetiRank**

$$\mathcal{L} = -\frac{1}{|P|} \sum_{p,n \in P} \log \left( \frac{e^{a_p}}{e^{a_p} + e^{a_n}} \right)$$

+

Попарная  
оценка листьев

Генерация пар

=

**YetiRankPairwise**

# QuerySoftMax

$$\mathcal{L} = - \sum_{group \in G} \sum_{i \in group} \frac{t_i}{\sum_{j \in group} t_j} \log \left( \frac{e^{a_i}}{\sum_{j \in group} e^{a_j}} \right)$$

$t_i$  - истинная метка для объекта

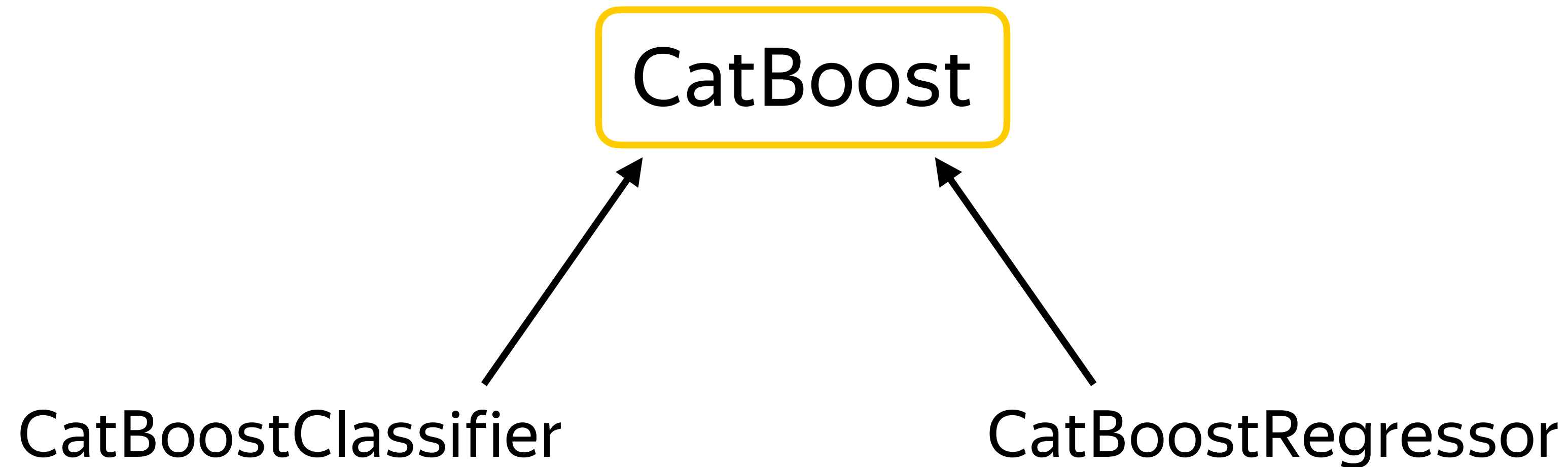
$a_i$  - предсказание для объекта

Хорошо подходит для решения задачи top-1

Обычно в группе один объект с меткой 1, а остальные - 0.



# Python



```
from catboost import Pool, CatBoost
from catboost.datasets import msrank_10k

train, _ = msrank_10k()

train_y, train_group_id, train_X = train[0], train[1], train.drop(columns=[0, 1])
train_y /= max(train_y)
train_pool = Pool(train_X, train_y, group_id=train_group_id)

model = CatBoost(dict(loss_function='YetiRankPairwise', iterations=10))
model.fit(train_pool)
```

# Feature Importance

## PredictionValuesChange

- Показывает влияние каждого признака на изменение предсказаний модели

- По-умолчанию для классификации и регрессии

- Плохо подходит для ранжирования

## LossFunctionChange

- Показывает влияние каждого признака на изменение значения функции потерь

- По-умолчанию для ранжирования

- Может применяться для классификации и регрессии

```
model.get_feature_importance(train_pool, type=EFstrType.LossFunctionChange)
```

# Советы



# Советы

Не нужно задавать большую глубину деревьев. Максимум 8, лучше 6. Иначе обучение будет очень долгим. (Особенно **YetiRankPairwise** и **PairLogitPairwise**).

Для **YetiRankPairwise** лучше ставить большой **learning\_rate** - он медленно сходится, но дает самые лучшие результаты.

# Советы - квантизация

В режимах `YetiRankPairwise` и `PairLogitPairwise` число бордеров при дискретизации фичей также достаточно сильно влияет на скорость (`--border-count`).

Для особо важных фичей можно увеличить количество бордеров (`--per-float-feature-quantization`).

# Советы - оценка значений в листьях

Для ускорения можно попробовать уменьшить число шагов по градиенту (`--leaf-estimation-iterations 1`).

Если режимы `YetiRankPairwise` и `PairLogitPairwise` учатся долго, то попробуйте заменить их на `YetiRank` и `PairLogit`.

# Советы - GroupId

- | В режимах **PairLogit** и **PairLogitPairwise** нужно разделять объекты на группы (задавать **GroupId**), так как внутри все параллелится именно по ним.

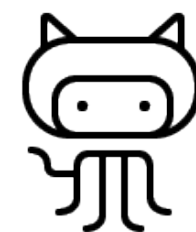
# Вопросы?

**Иван Лыжин**

Разработчик CatBoost



[catboost.ai](https://catboost.ai)



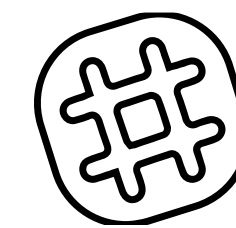
[github.com/catboost](https://github.com/catboost)



[twitter.com/CatBoostML](https://twitter.com/CatBoostML)



[t.me/catboost\\_en](https://t.me/catboost_en), [t.me/catboost\\_ru](https://t.me/catboost_ru)



[ods.ai](https://ods.ai) => slack (40k people community)  
=> tool\_catboost chanel