

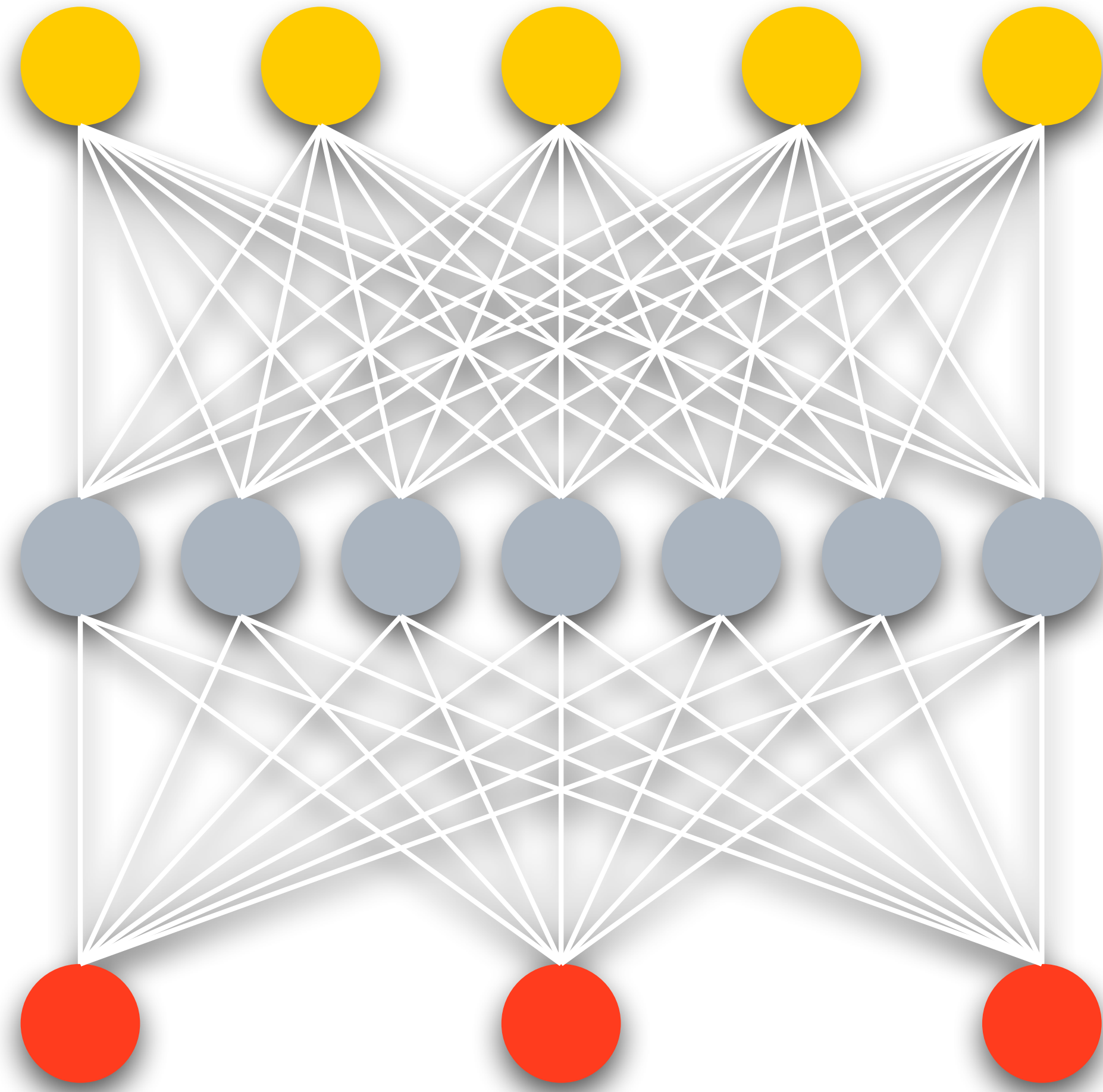
Яндекс

Yandex

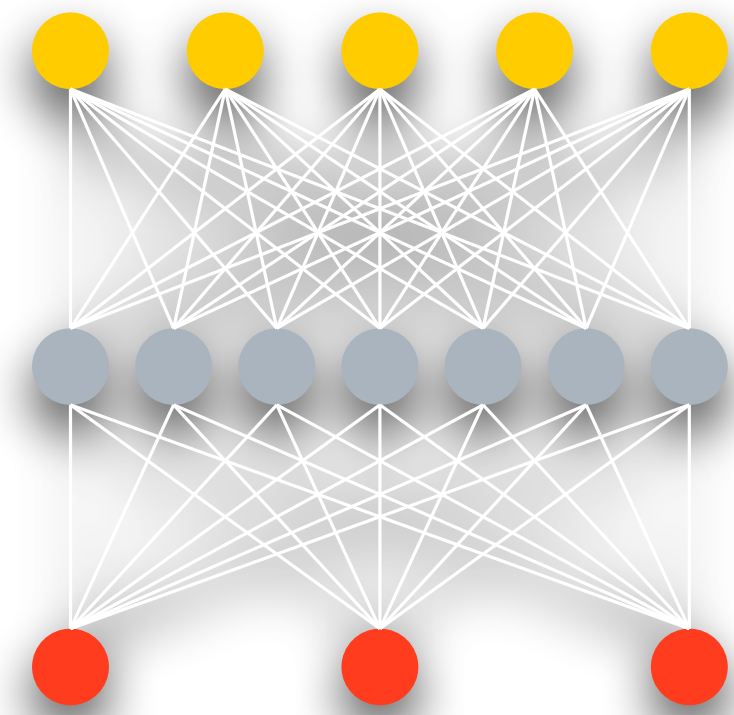
CatBoost — gradient boosting library

Vasily Ershov, Software Developer

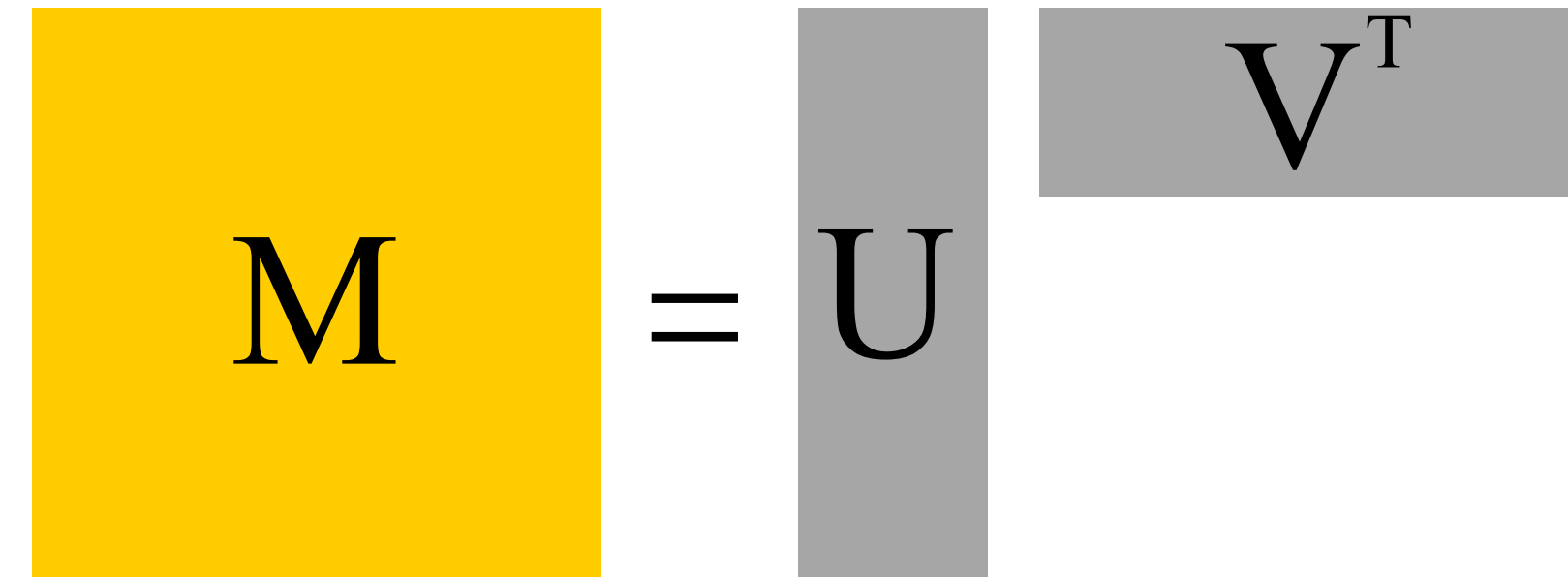
Машинное обучение = AI?



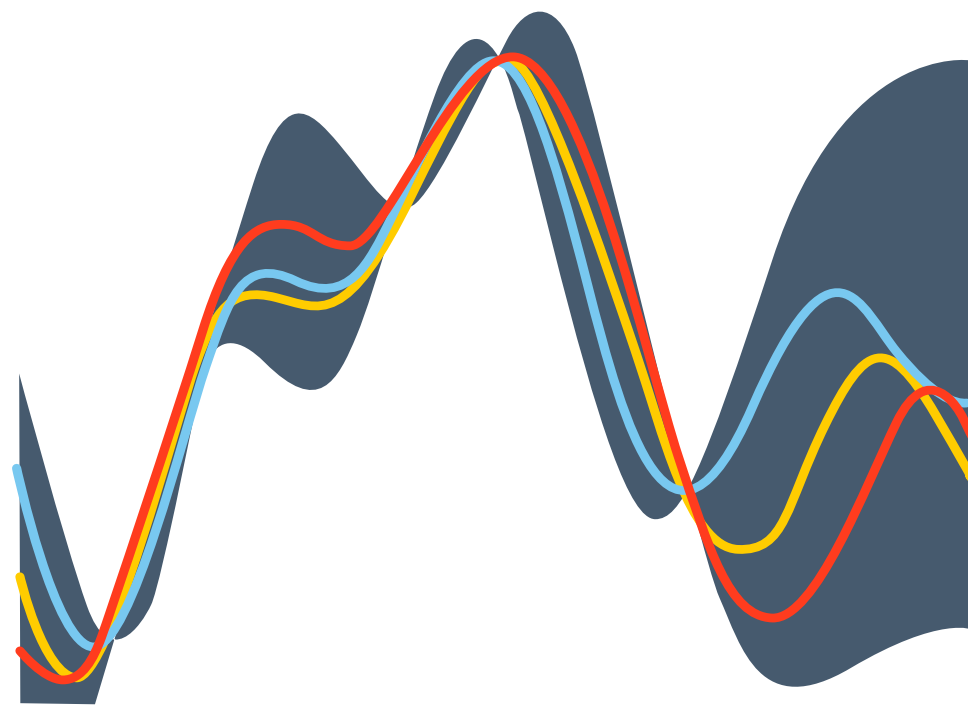
Машинное обучение ≠ AI



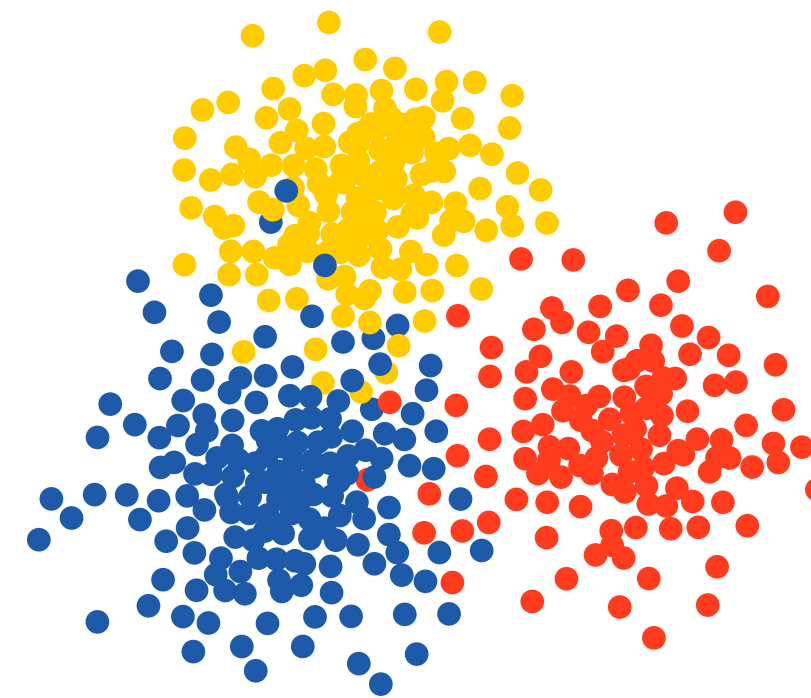
Нейронные сети

$$M = U V^T$$


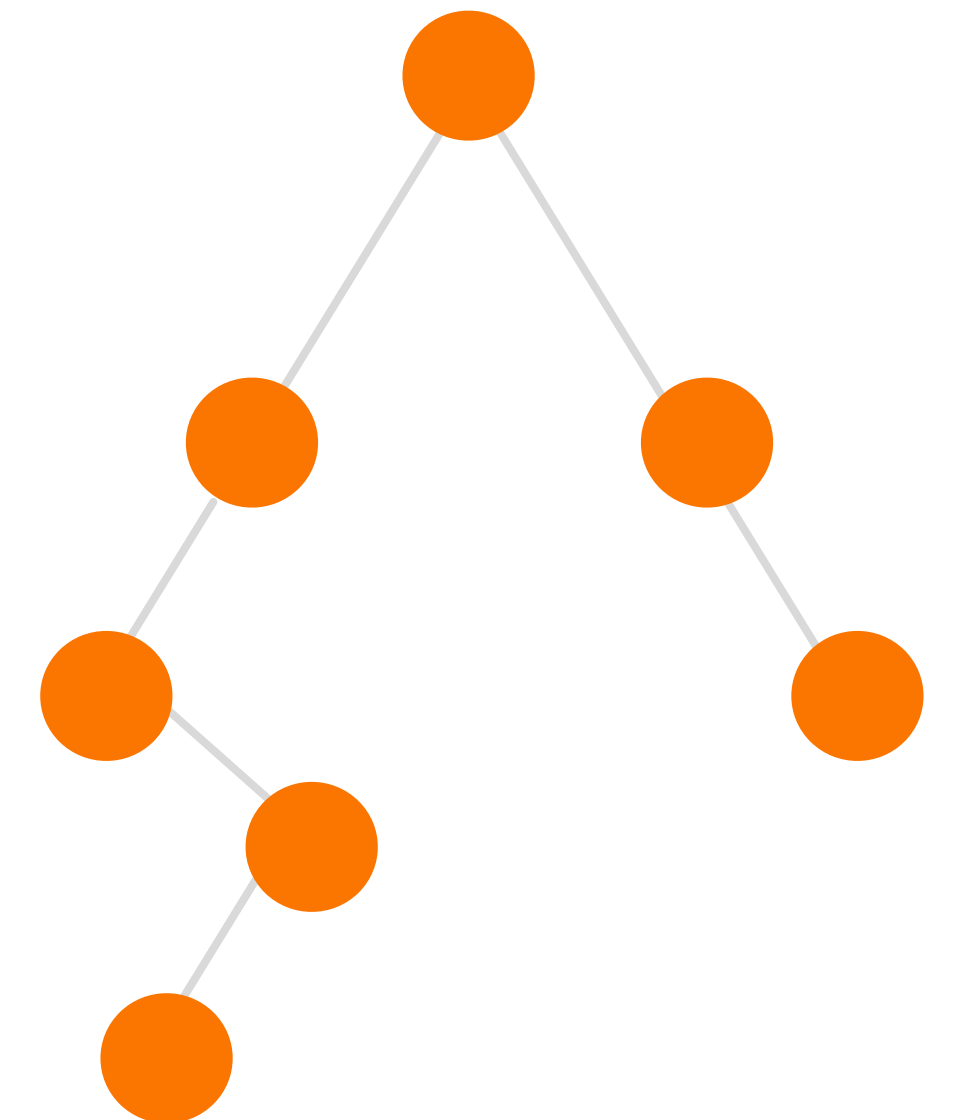
Факторизация матриц



Гауссовские процессы



Кластеризация



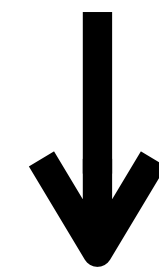
Деревья решений

Обучение с учителем

$$x \in X, y \in Y$$

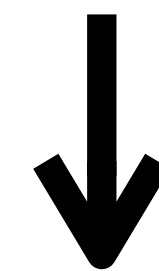
Предположения

$$f : X \rightarrow Y$$



Данные для обучения

$$(x_1, y_1), \dots, (x_n, y_n)$$

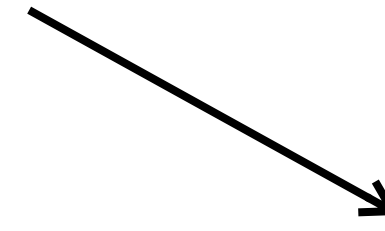
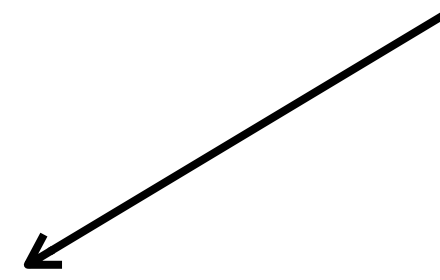


Восстановленная

зависимость

$$\hat{f} : X \rightarrow Y$$

Входные данные



Неструктурированные данные

Структурированные данные

Neural networks

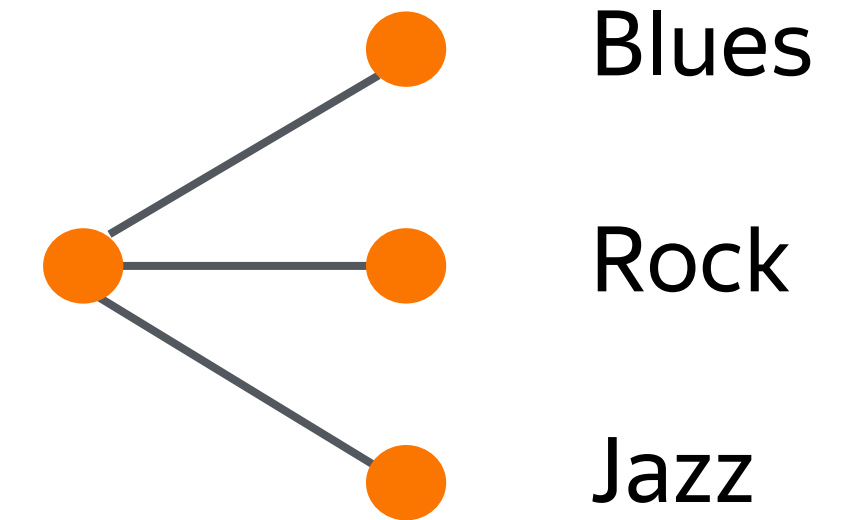
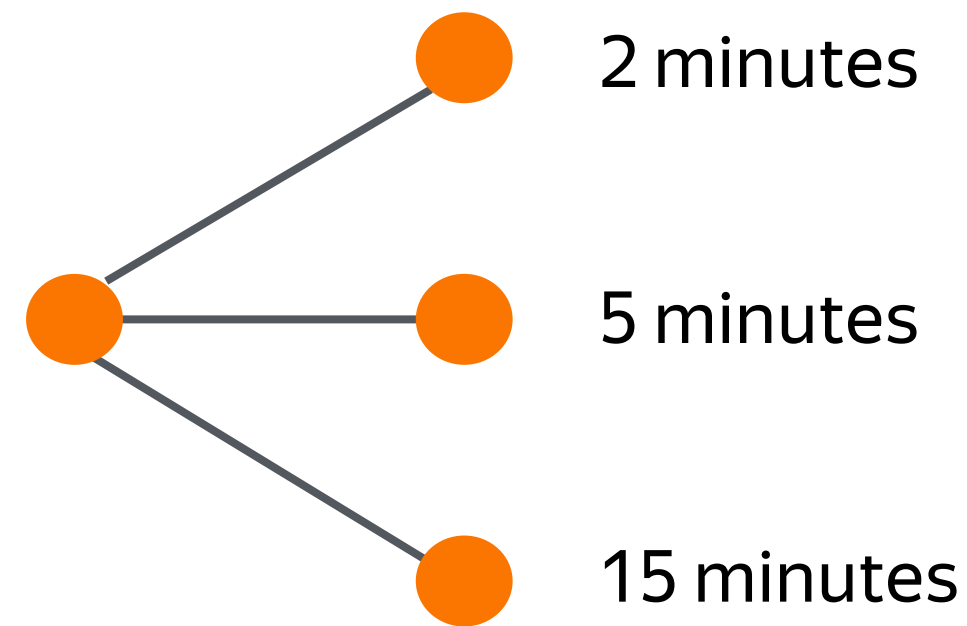
«Числовые» признаки

Категориальные признаки



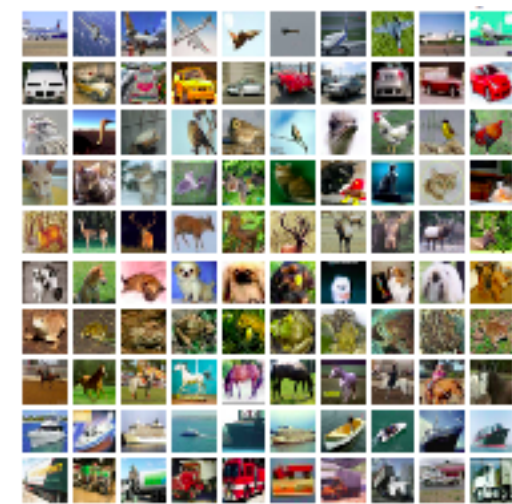
DNA

Текст



Длина песни

Жанр песни



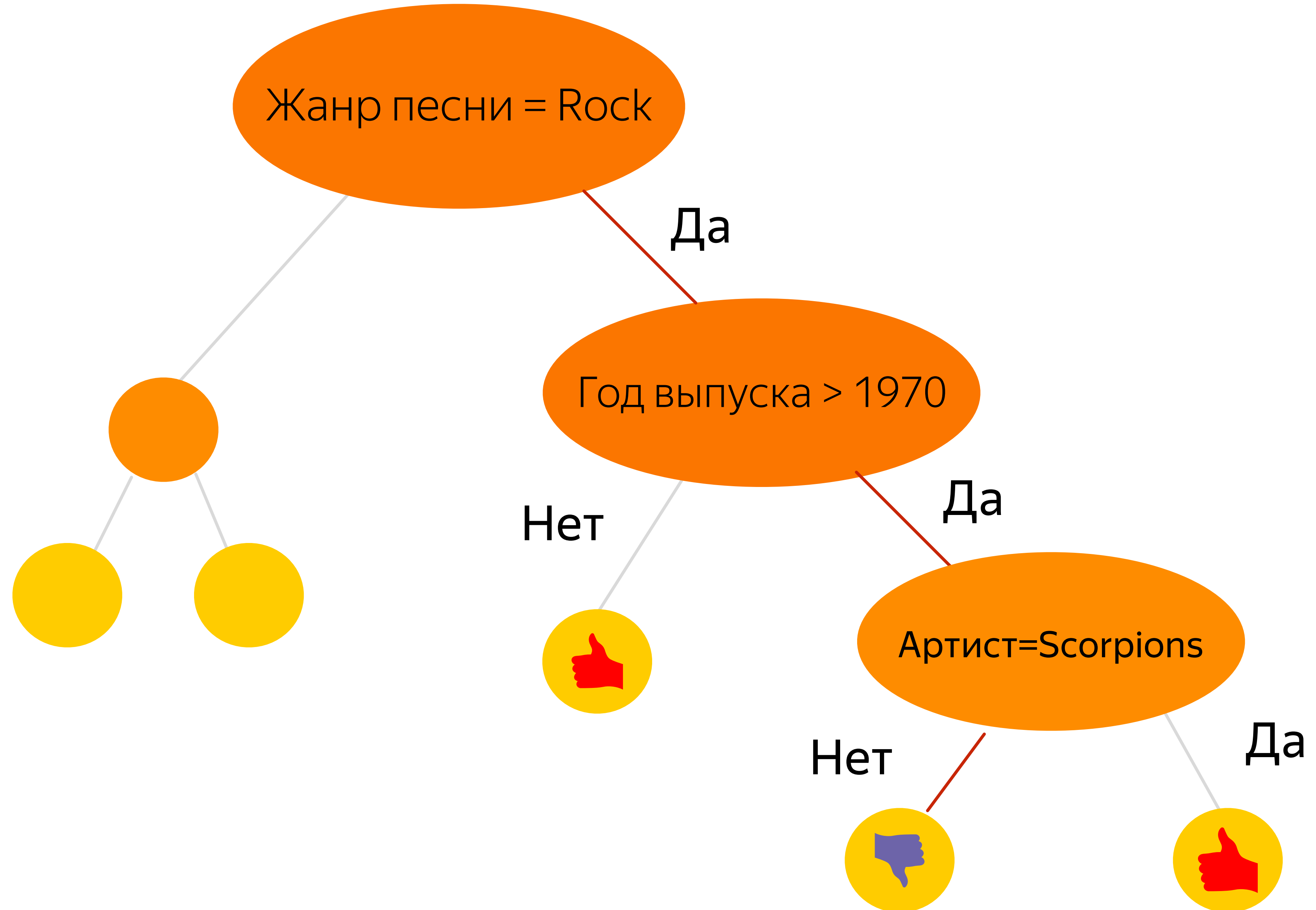
Изображения

Музыка

Длина песни	Год выпуска	Рейтинг
2	1990	3
3	1950	5
15	1970	4

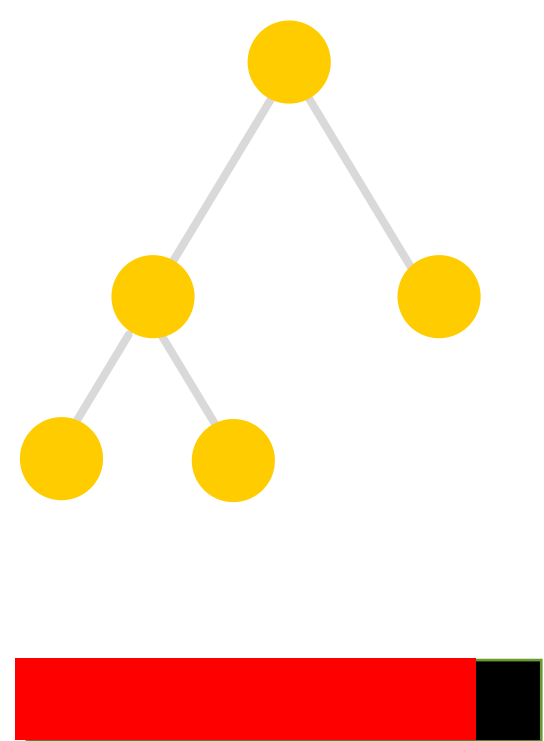
Жанр	Исполнитель
Rock	Scorpions
Jazz	Louis Armstrong
Blues	B.B.King

Деревья решений

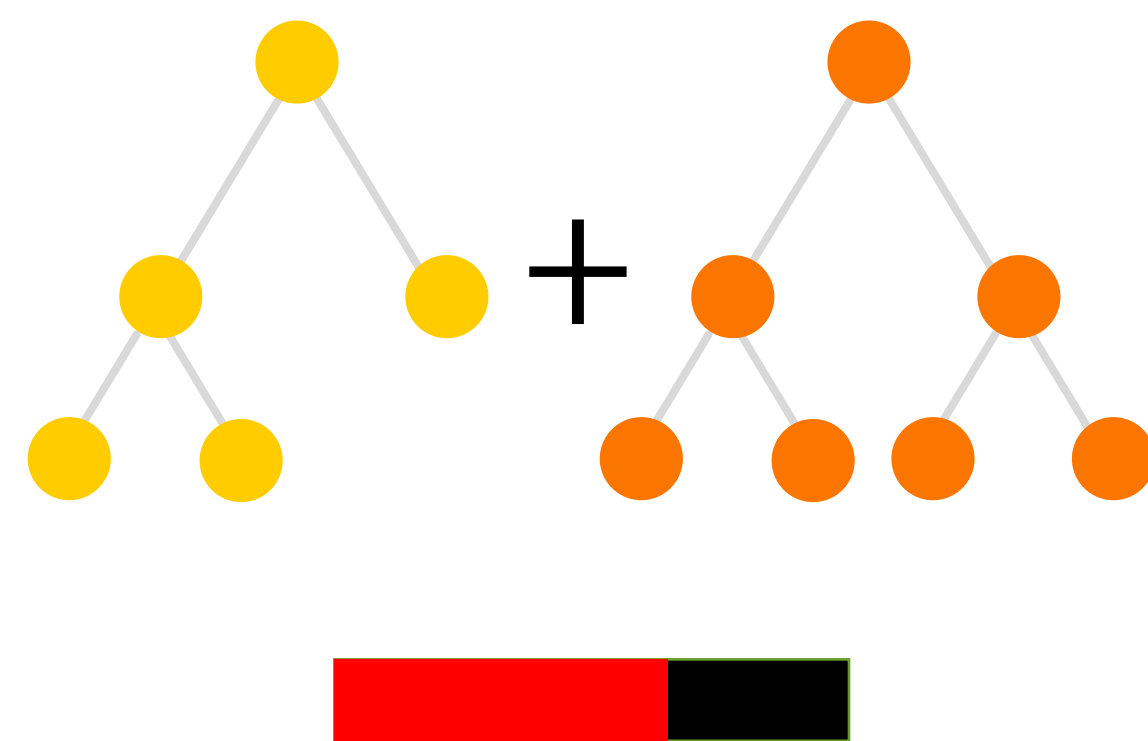


Boosting на деревьях решений

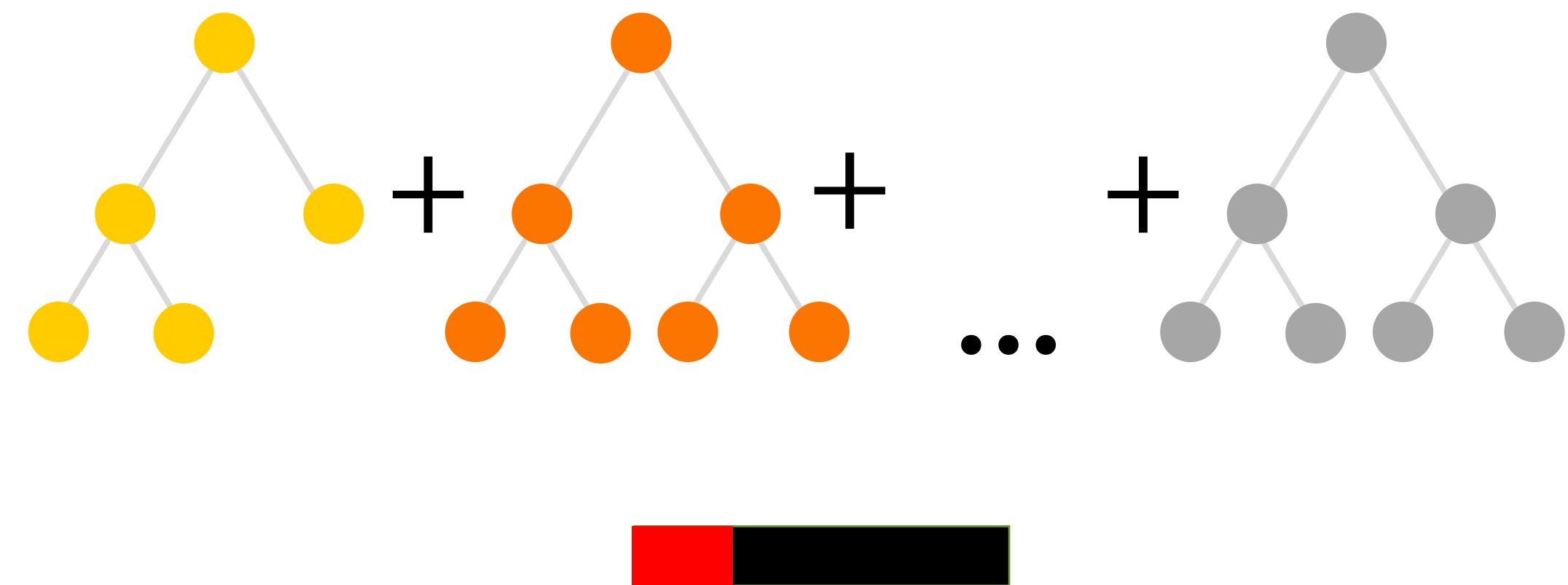
- State-of-the-art quality на структурированных данных
- Прост в использовании
- Работает на небольших объемах данных, а также легко масштабируется на «Big data problems»



Большая ошибка



Стало лучше



Можно в production

Main Boosting libraries

dmlc
XGBoost



Yandex
CatBoost



Microsoft
LightGBM

CatBoost

catboost / catboost

Unwatch 166 Star 3,083 Fork 405

Code Issues 74 Pull requests 0 Insights

CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R <https://catboost.yandex>

machine-learning decision-trees gradient-boosting gbm gbd python r kaggle gpu-computing catboost tutorial

categorical-features distributed gpu coreml opensource data-science big-data

2,475 commits 8 branches 23 releases 73 contributors Apache-2.0

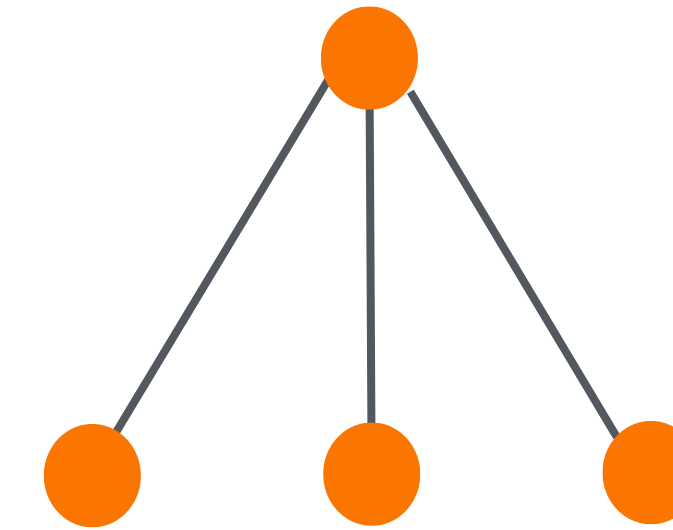
Branch: master New pull request Create new file Upload files Find file Clone or download

Sergey Preis New ymake Latest commit 9043c6b an hour ago

build New ymake an hour ago

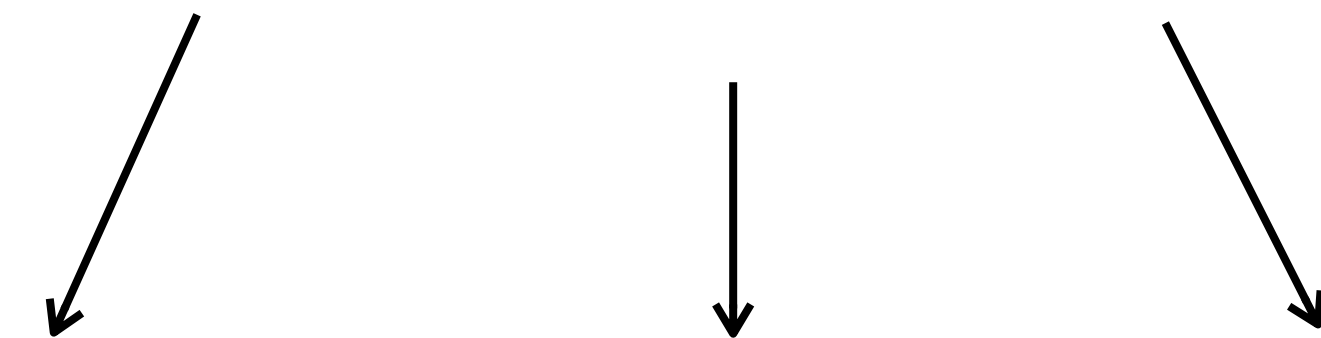
Возможности CatBoost'a

Поддержка категориальных данных



Классический препроцессинг (one-hot-encoding)

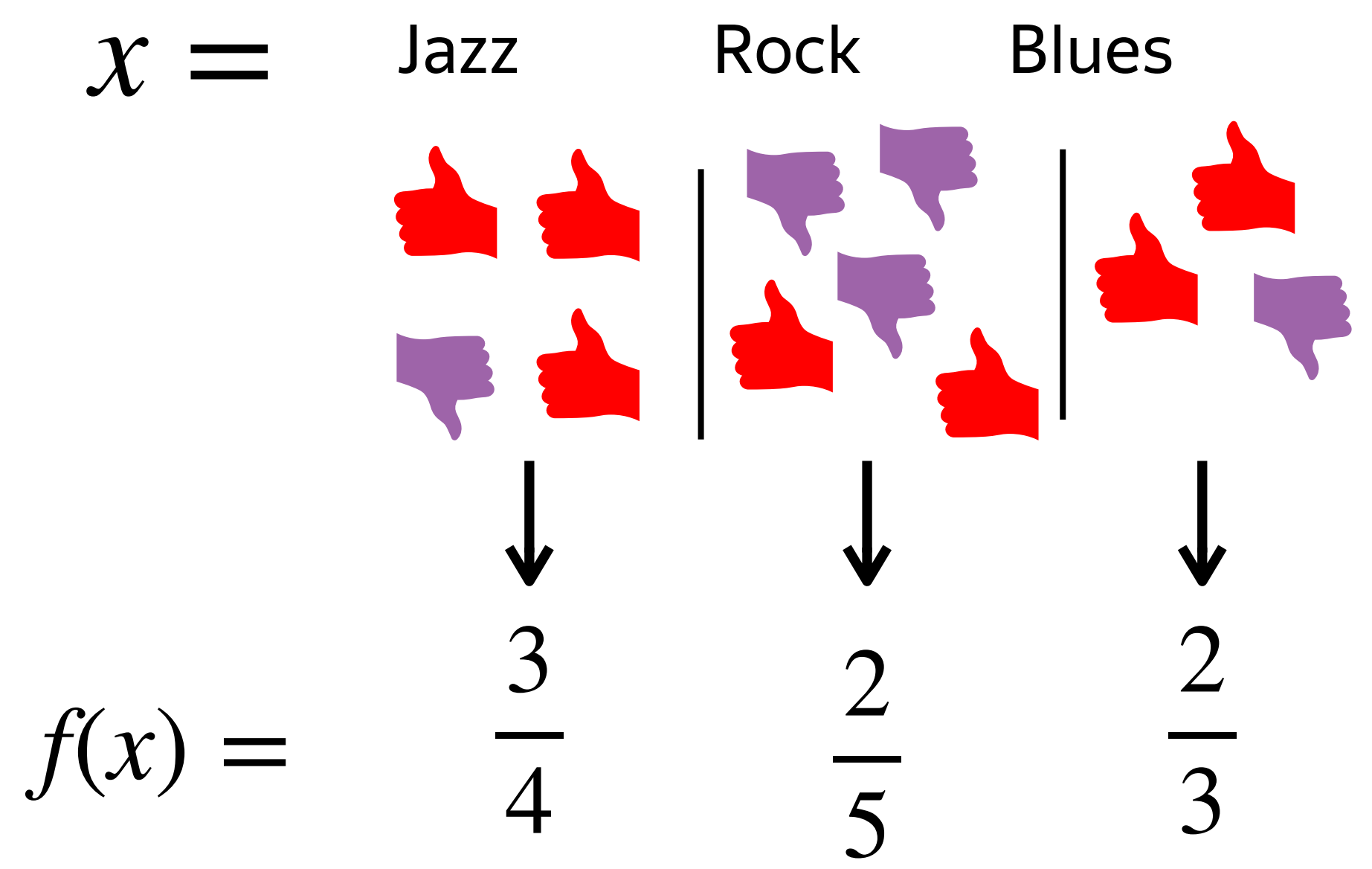
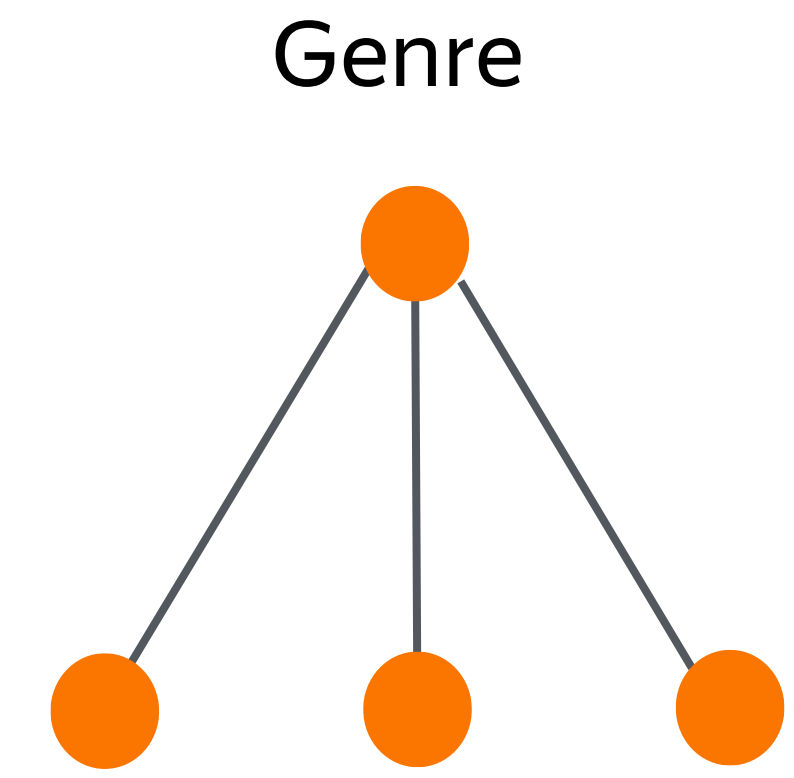
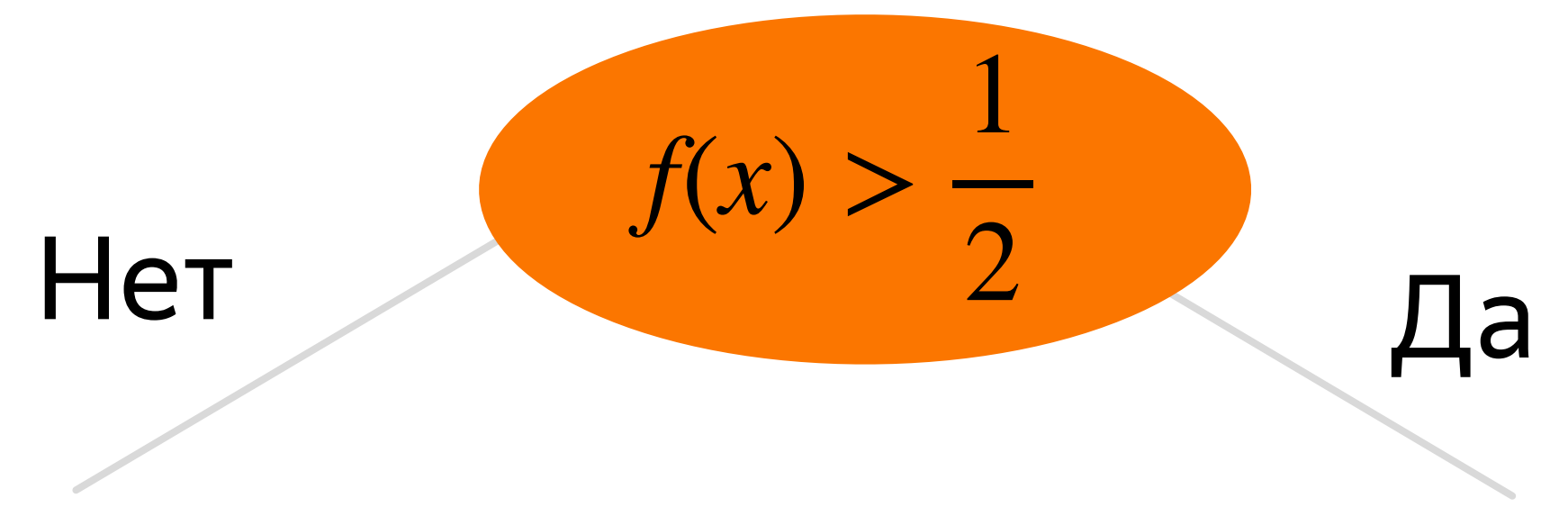
Jazz Rock Blues



Значение признака	value=jazz	value=rock	value=blues
Jazz	1	0	0
Blues	0	0	1
Rock	0	1	0

Возможности CatBoost'a

- Поддержка категориальных данных
 - Классический препроцессинг (one-hot-encoding)
 - Подход к работе с категориальными признаками на основе автоматического вычисления статистик по категориальным признакам

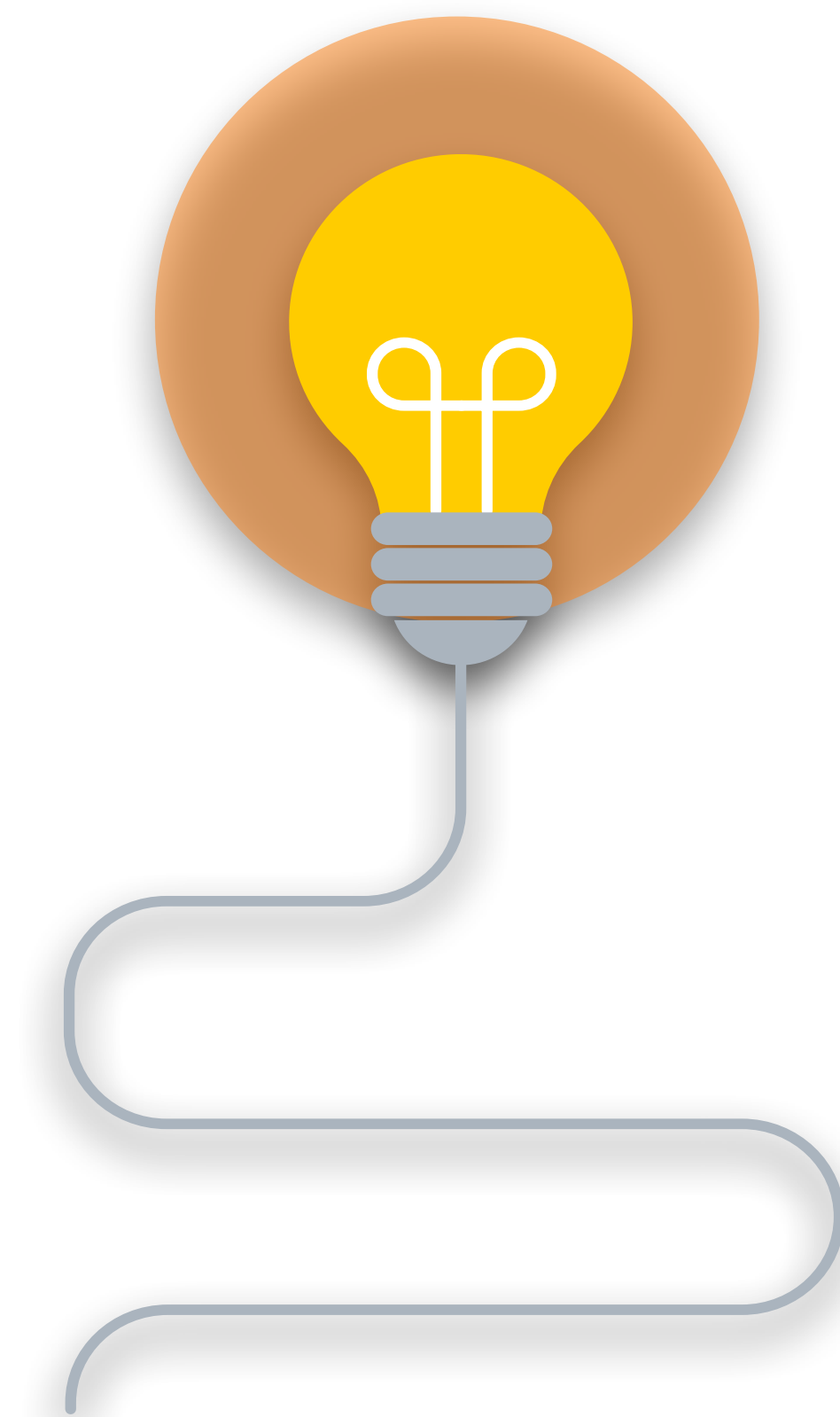


Возможности CatBoost'a

Поддержка категориальных данных

Высокое качество

- › Новая схема boosting'a, позволяющая избегать переобучения (ordered boosting)
- › Полезные эвристики, основанные на богатом опыте применения boosting'a: мы обучали ансамбли из деревьев решения до того, как это стало модным :)



Возможности CatBoost'a

Поддержка категориальных данных

Высокое качество

Широкий спектр решаемых задач

1

Регрессия

2

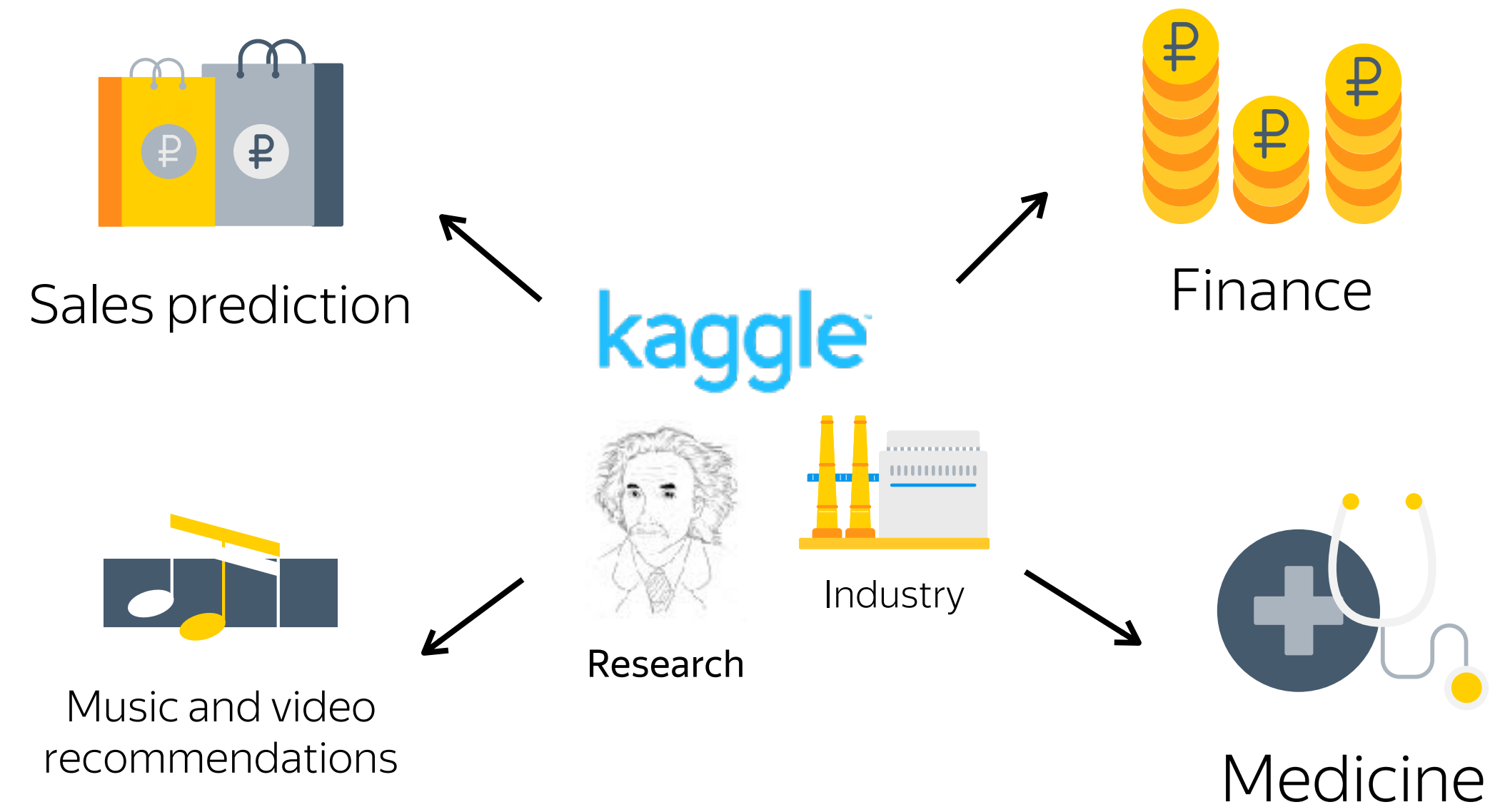
Классификация и
мультиклассификация

3

Ранжирование

4

Специализированные режимы



Возможности CatBoost'a

Поддержка категориальных данных

1

Регрессия

Высокое качество

Широкий спектр решаемых задач

2

Классификация и
мультиклассификация

Ищем вещественно-значную функцию

3

Ранжирование

$$f: X \rightarrow \mathbb{R}$$

4

Специализированные режимы

Использование в Яндексе: погода

Что хотим?

› Предсказать температуру

Тип задачи: регрессия

Данные для обучения

› Физические модели погоды

› Online-данные с датчиков

› Данные об температуре за последние n лет



Возможности CatBoost'a

Поддержка категориальных данных

1

Регрессия

Высокое качество

Широкий спектр решаемых задач

2

**Классификация и
мультиклассификация**

Пытаемся предсказать один из m
классов объекта

3

Ранжирование

$$f: X \rightarrow \{0, 1, \dots, m\}$$

4

Специализированные режимы

Использование в Яндексе: погода

Что хотим?

- › Предсказать тип погоды (осадки, ветер, etc)

Тип задачи: мультиклассификация

Данные для обучения

- › Физические модели погоды
- › Online-данные с датчиков
- › Данные об облачности за последние n лет



Особенности CatBoost'a

Поддержка категориальных данных

1

Регрессия

Высокое качество

Широкий спектр решаемых задач

2

Классификация и
мультиклассификация

Функция, по значению которой
можно сортировать объекты
некоторым оптимальным образом

3

Ранжирование

$$f : X \rightarrow \mathbb{R}$$

4

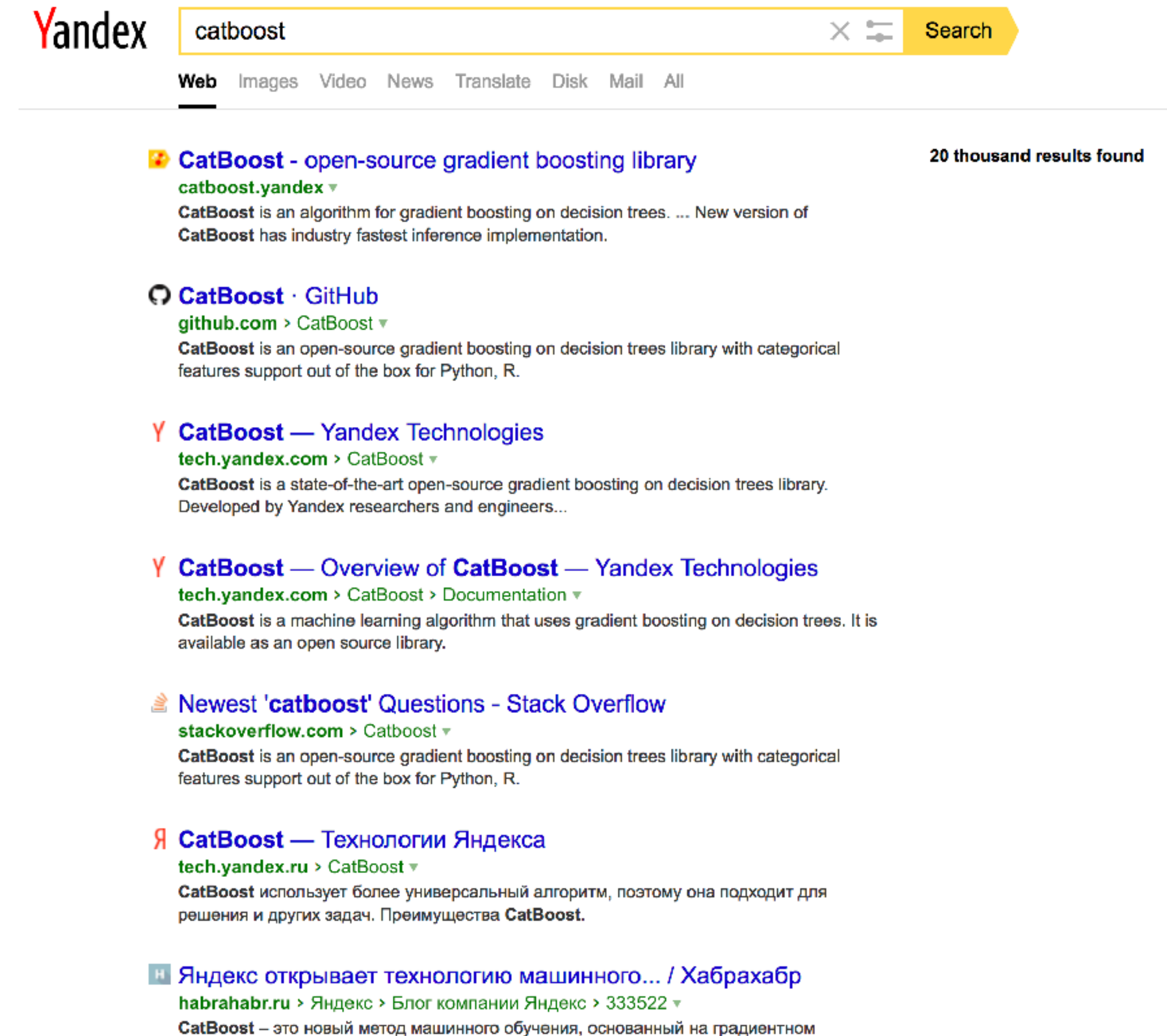
Специализированные режимы

Использование в Яндексе: поиск

Тип задачи: ранжирование

Данные для обучения

- › Запрос
- › Признаки на основе документов
- › Ручная разметка документов на релевантные и не релевантные



The screenshot shows a Yandex search interface with the query 'catboost' entered in the search bar. The search bar includes a 'Search' button and a 'Web' tab selected. Below the search bar, there are navigation links for 'Images', 'Video', 'News', 'Translate', 'Disk', 'Mail', and 'All'. The search results are displayed in a list format, with the first result being '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... / Хабрахабр' from 'habrahabr.ru'. The search results are sorted by relevance, and the total number of results found is 20 thousand.

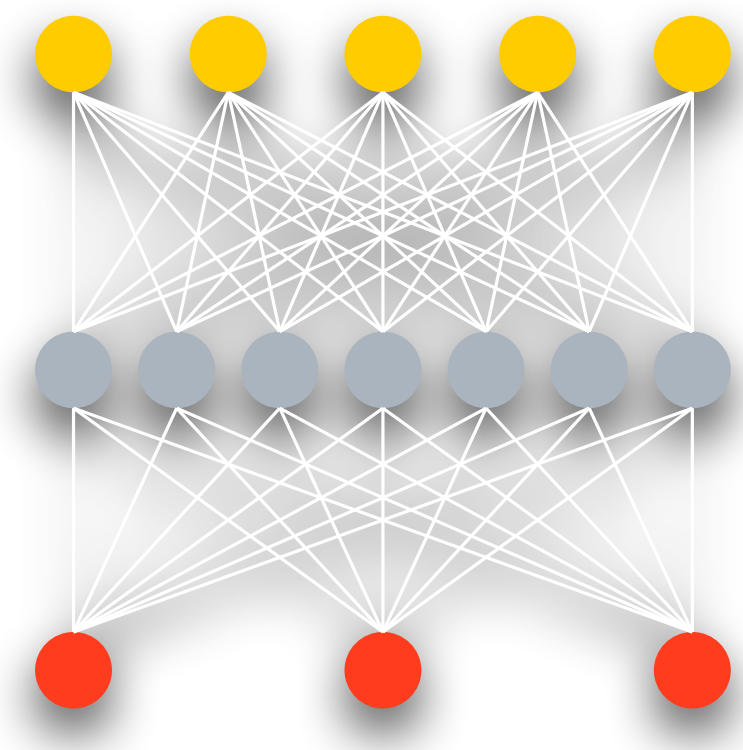
Yandex X ⇄ Search

Web Images Video News Translate Disk Mail All

20 thousand results found

- 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

Использование в Яндексе: поиск

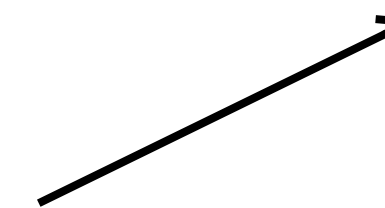
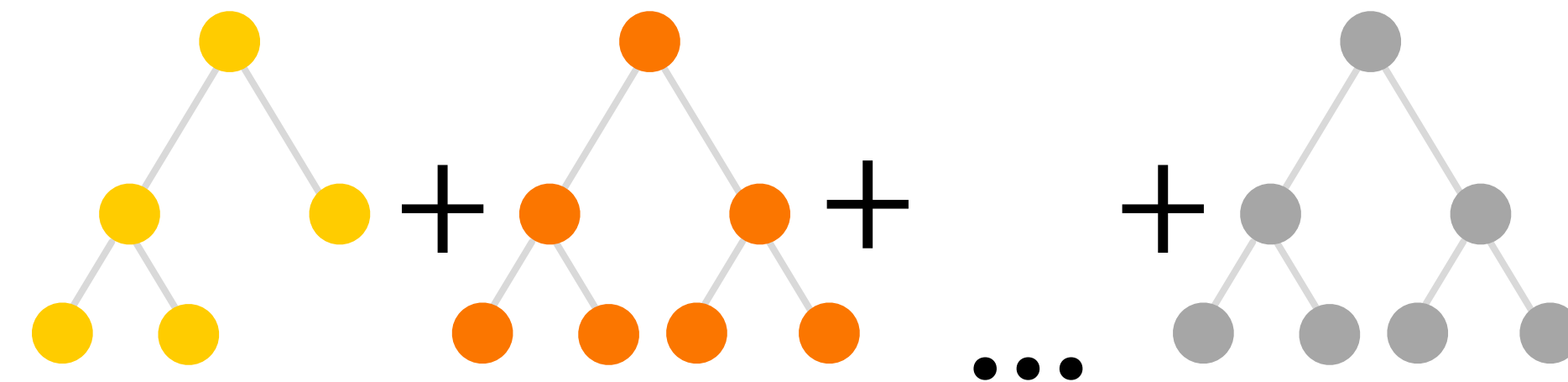
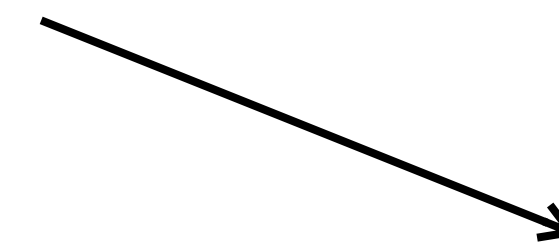


Обработка текста

Обработка картинок



Признаки, разработанные
людьми
(PageRank, VM25, etc)



Boosting эффективно
комбинирует нейронные сети
и «человеческий интеллект»

Возможности CatBoost'a

Поддержка категориальных данных

1

Регрессия

Высокое качество

Широкий спектр решаемых задач

2

Классификация и
мультиклассификация

Более сложные
специализированные режимы

3

Ранжирование

4

Специализированные режимы

Использование в Яндексе: реклама

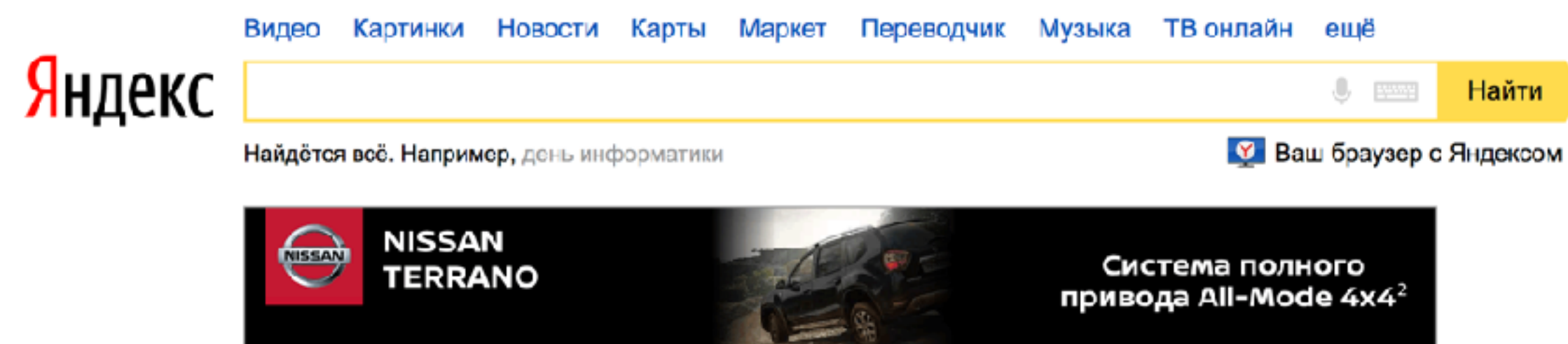
Что хотим?

- › Показать баннер, который понравится пользователю

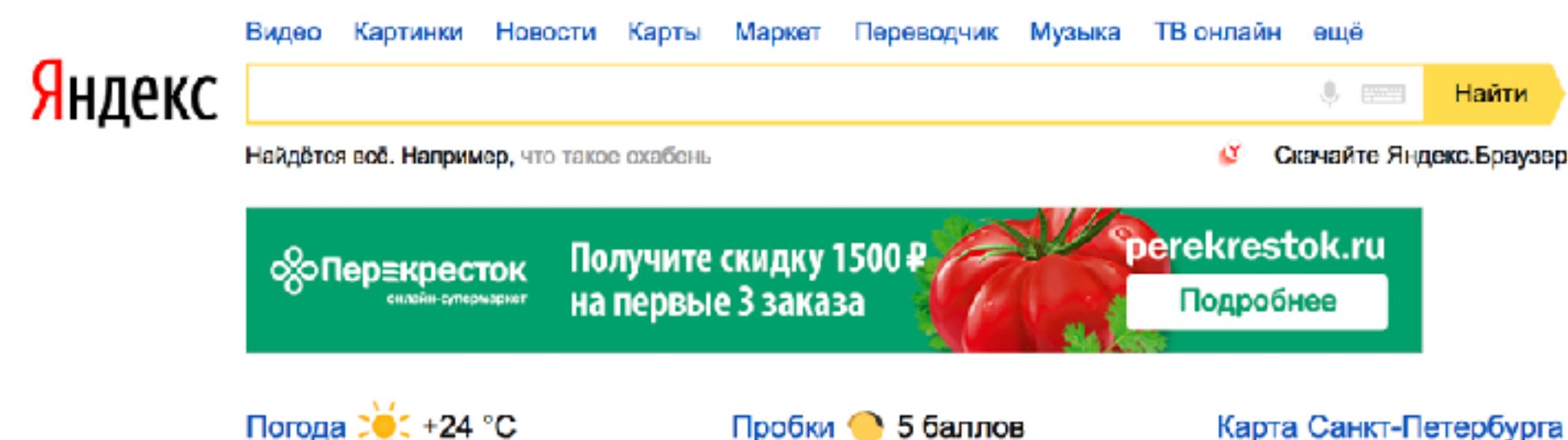
Тип задачи: смесь ранжирование и классификации

Данные для обучения

- › История поиска пользователей
- › История кликов пользователей



ИЛИ



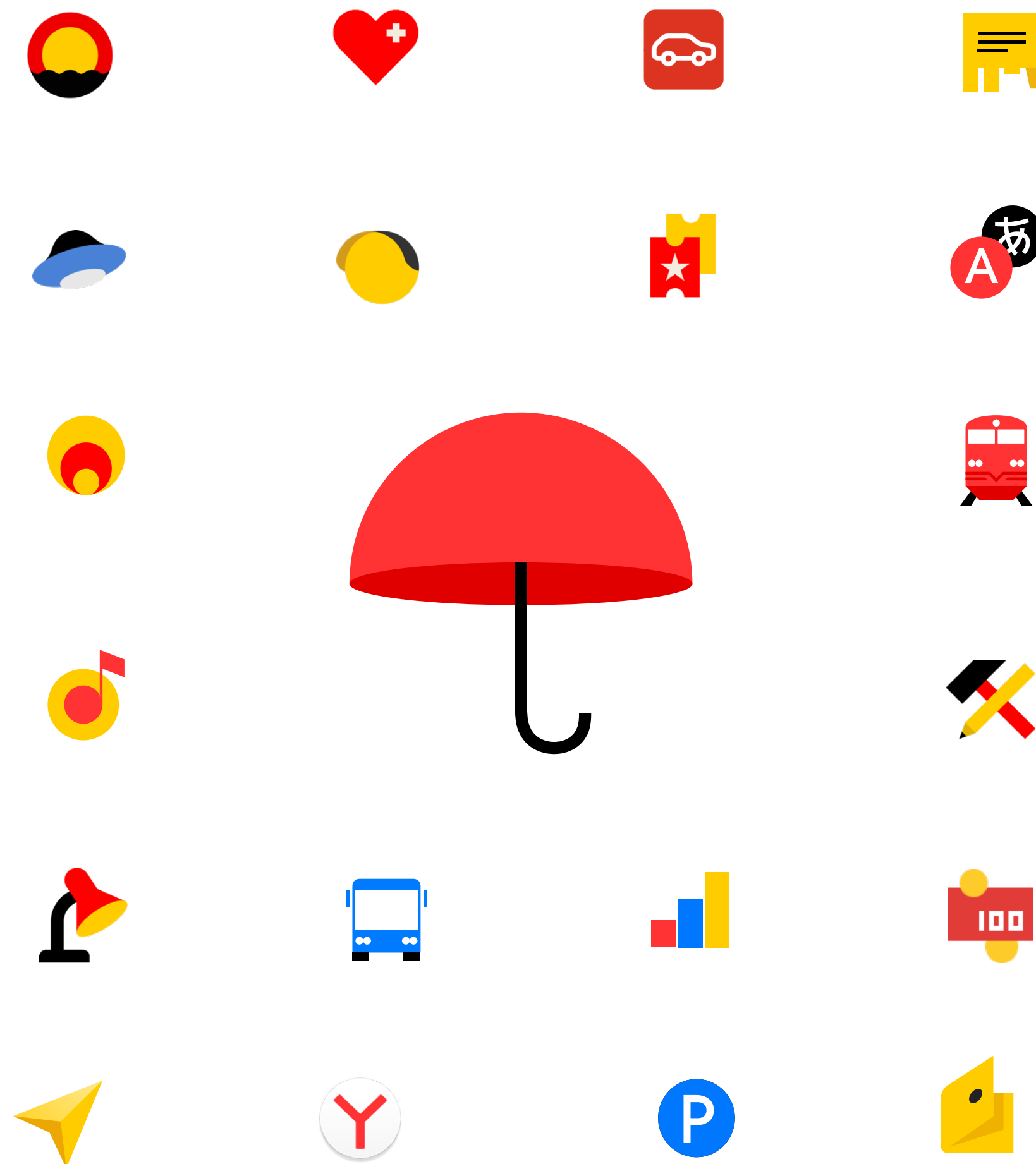
Использование в Яндексе

Алиса: ранжирование

Дзен: ранжирование, рекомендации

Музыка: рекомендации

И многие другие



Возможности CatBoost'a

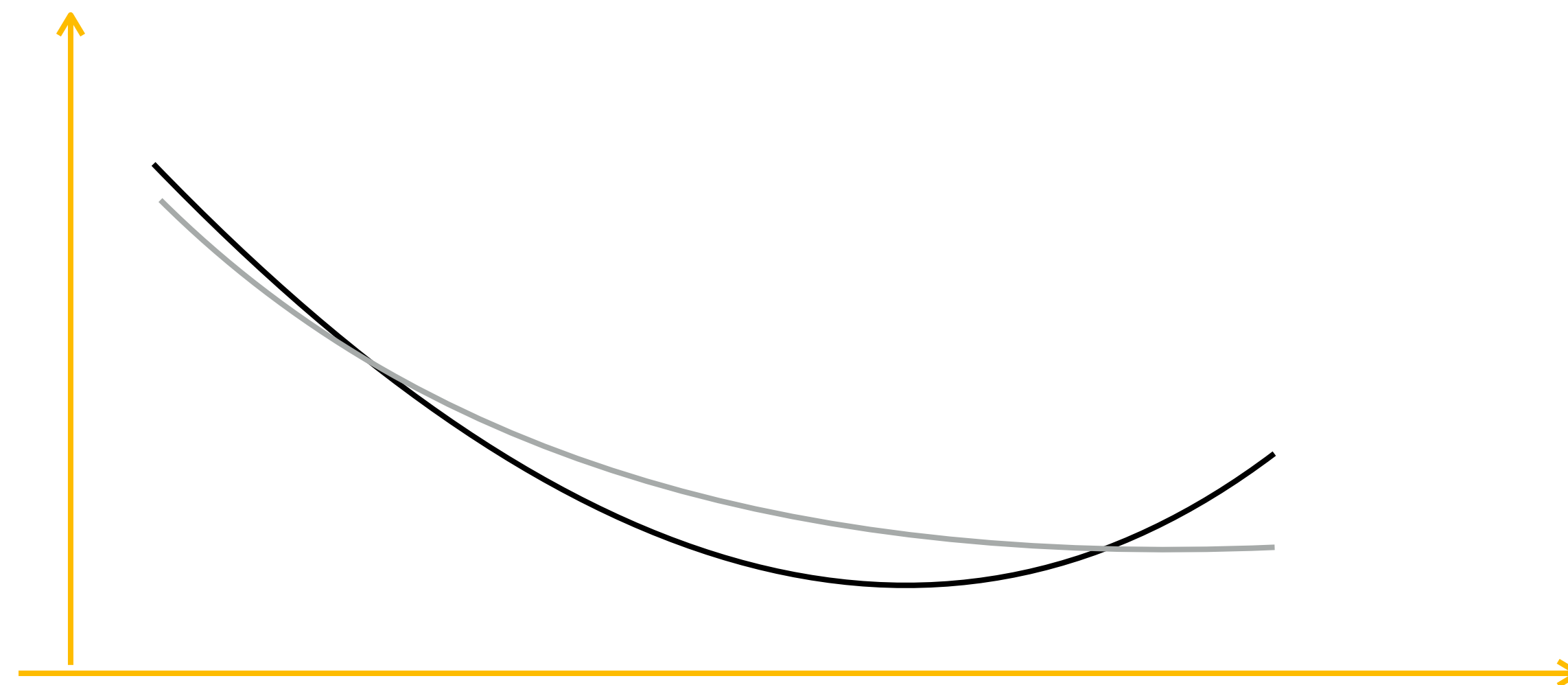
Поддержка категориальных данных

Высокое качество

Широкий спектр решаемых задач

Удобно пользоваться

› Встроенная аналитика



Графики ошибок + интеграция с
tensorboard, jupyter

Поиск важных примеров

Влияние признаков

Возможности CatBoost'a

Поддержка категориальных данных

Высокое качество

Широкий спектр решаемых задач

Удобно пользоваться

› Встроенная аналитика

› Хорошие гиперпараметры алгоритмов



Подбор параметров?

Можно, но работаем и без него хорошо

Возможности CatBoost'a

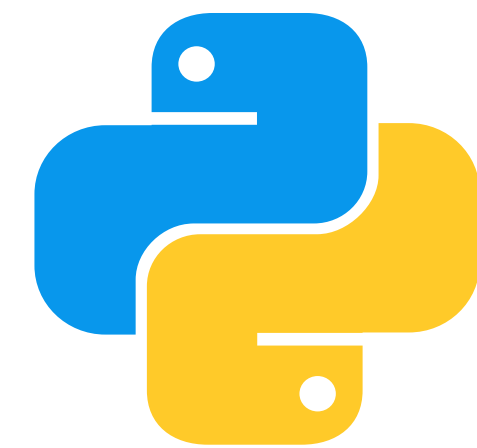
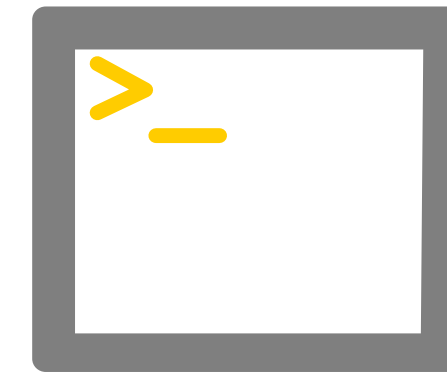
Поддержка категориальных данных

Высокое качество

Широкий спектр решаемых задач

Удобно пользоваться

- › Встроенная аналитика
- › Хорошие гиперпараметры алгоритмов
- › Удобные библиотеки и документация к ним



Пример использования

Оцененная релевантность документа

ID запроса

мета-информация

Признаки

GroupId	Relevance	URL	SubgroupId	F0	F1	F2	F3	F4	F5	...	F39	F40	F41	F42	F43	F
10	0.00	http://www.chtivo.ru/chtivo=3&bkid=776986.htm	21	0.345548	0.248034	0.085307	1	1	0	...	0.002241	0	0	0	0	
10	0.00	http://www.rukomos.ru	4	0.424438	0.164384	0.572649	1	1	0	...	0.001410	0	0	0	0	
10	0.00	http://www.vhlam.ru/pressa.html	73	0.317637	0.076506	0.263320	1	1	0	...	0.000621	0	0	0	0	
10	0.07	http://advtime.ru/online/ruki/ruki.html	2	0.257727	0.021591	0.171299	1	1	1	...	0.000173	0	0	0	0	
10	0.07	http://db.kvadroom.ru	190	0.744063	0.000000	0.624348	1	0	0	...	0.000000	0	0	0	0	

Входная таблица

Пример использования: обучение

```
from catboost import Pool
from catboost import CatBoostRegressor

train_pool = Pool('/home/noxoomo/ranking_pool/features.txt',
                  column_description='/home/noxoomo/ranking_pool/pool.cd')
test_pool = Pool('/home/noxoomo/ranking_pool/featuresTest.txt',
                 column_description='/home/noxoomo/ranking_pool/pool.cd')

model = CatBoostRegressor(
    thread_count=16,
    iterations=1000,
    learning_rate=0.2,
    logging_level="Silent",
    border_count=32,
    loss_function="YetiRankPairwise")

model.fit(train_pool, eval_set=test_pool, plot=True)
```


Возможности CatBoost'a

Поддержка категориальных данных

Высокое качество

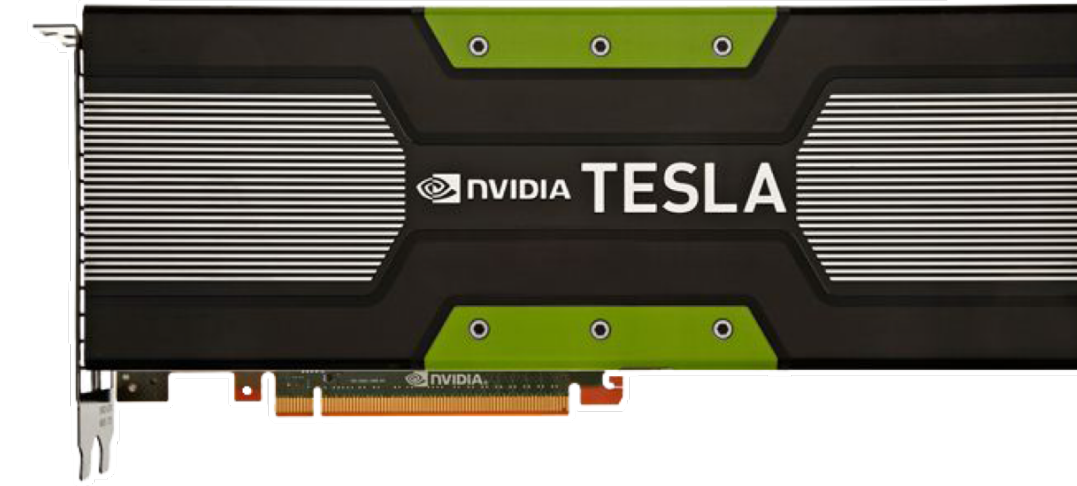
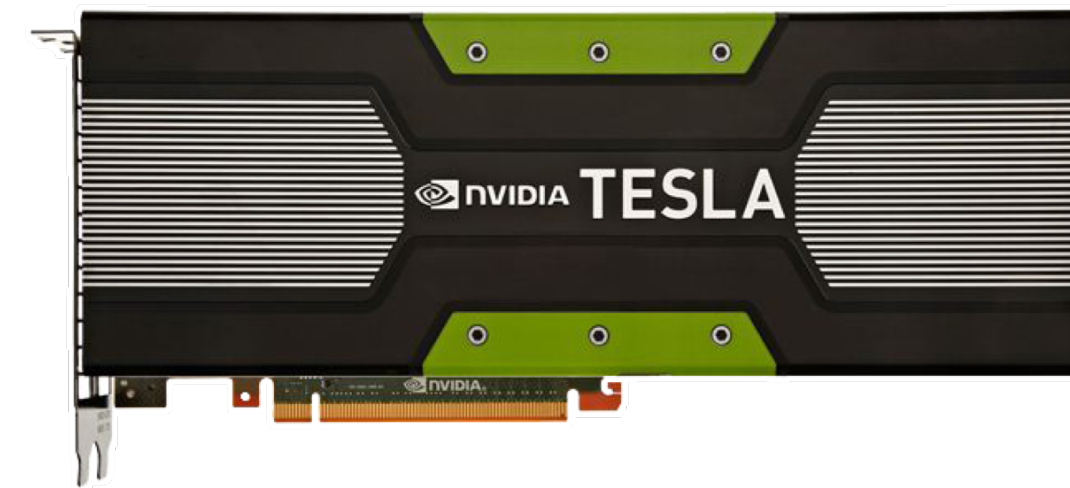
Удобно пользоваться

Широкий спектр решаемых задач

Производительность и масштабируемость

› Обучение

› Применение (deploy)

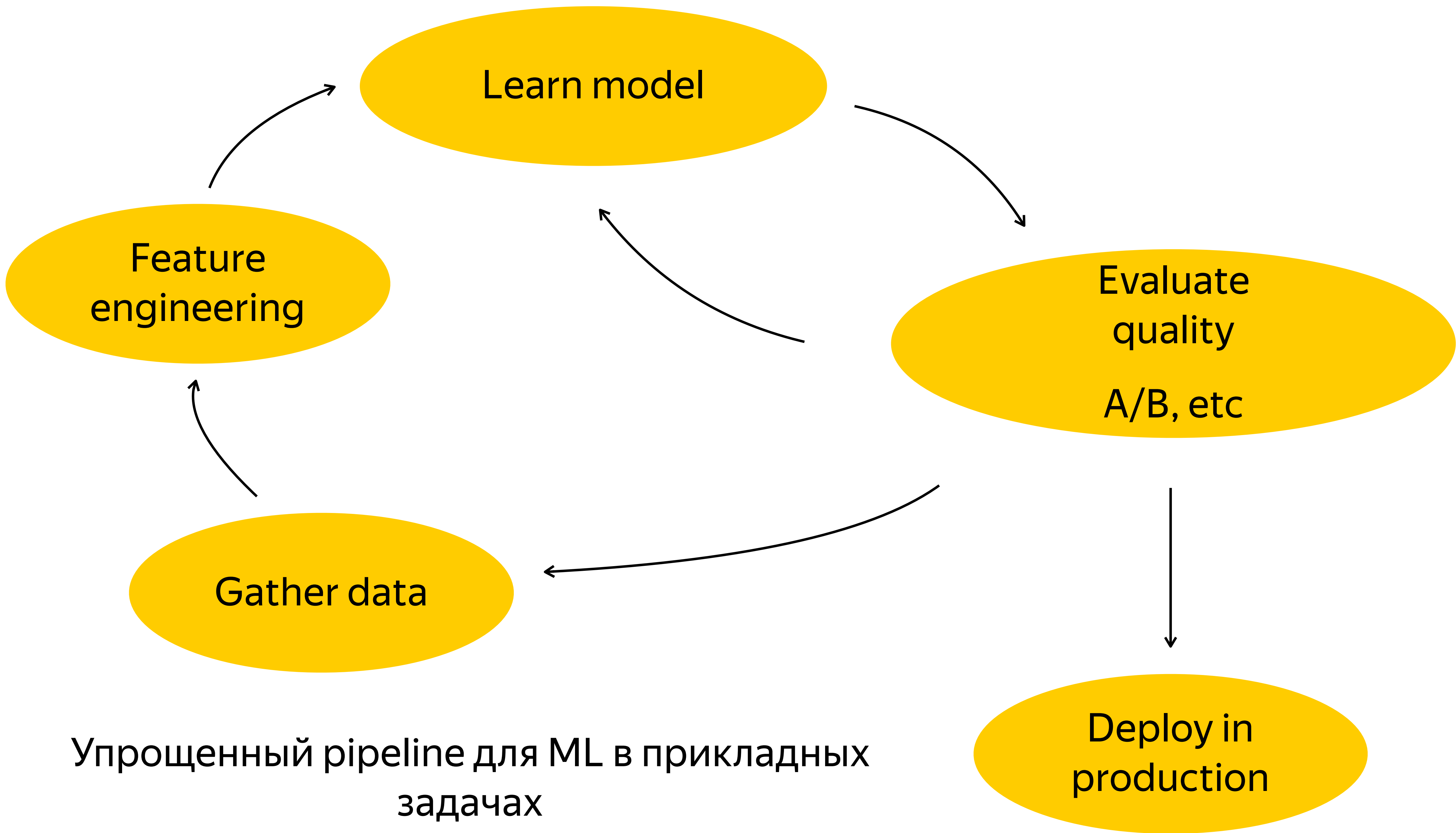


GPU and MultiGPU

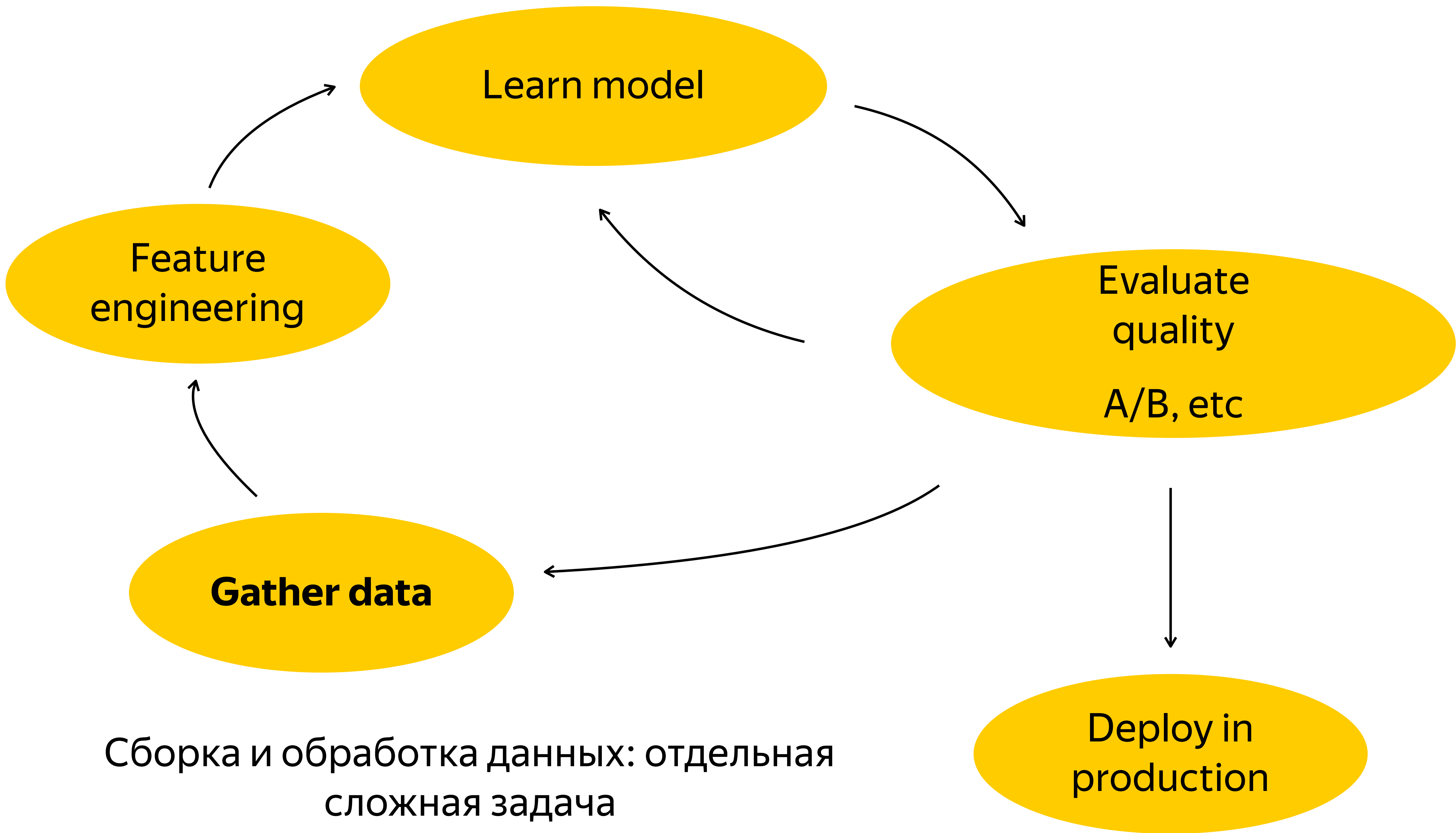


Distributed CPU

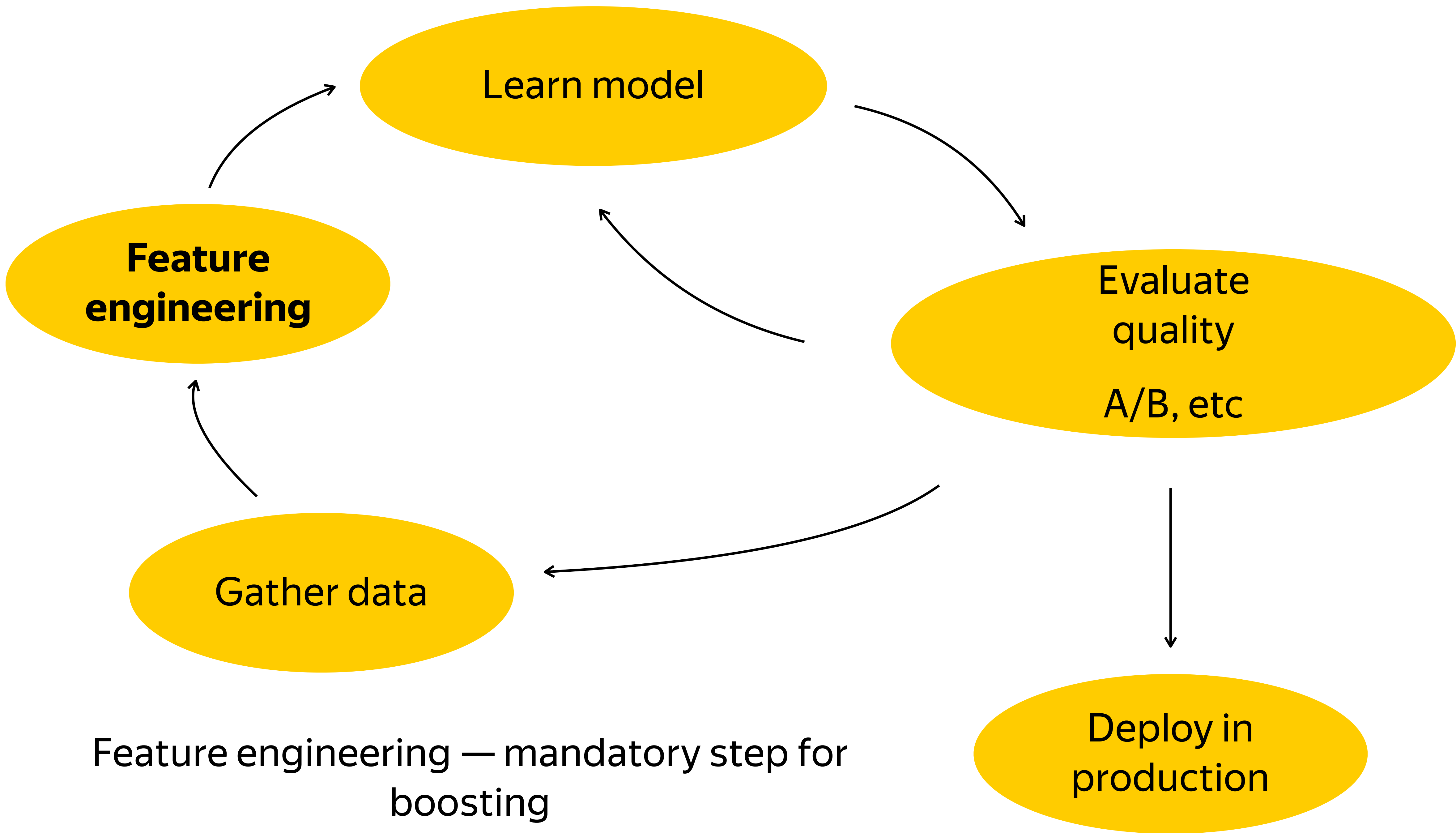
CatBoost: решаем прикладные задачи в индустрии



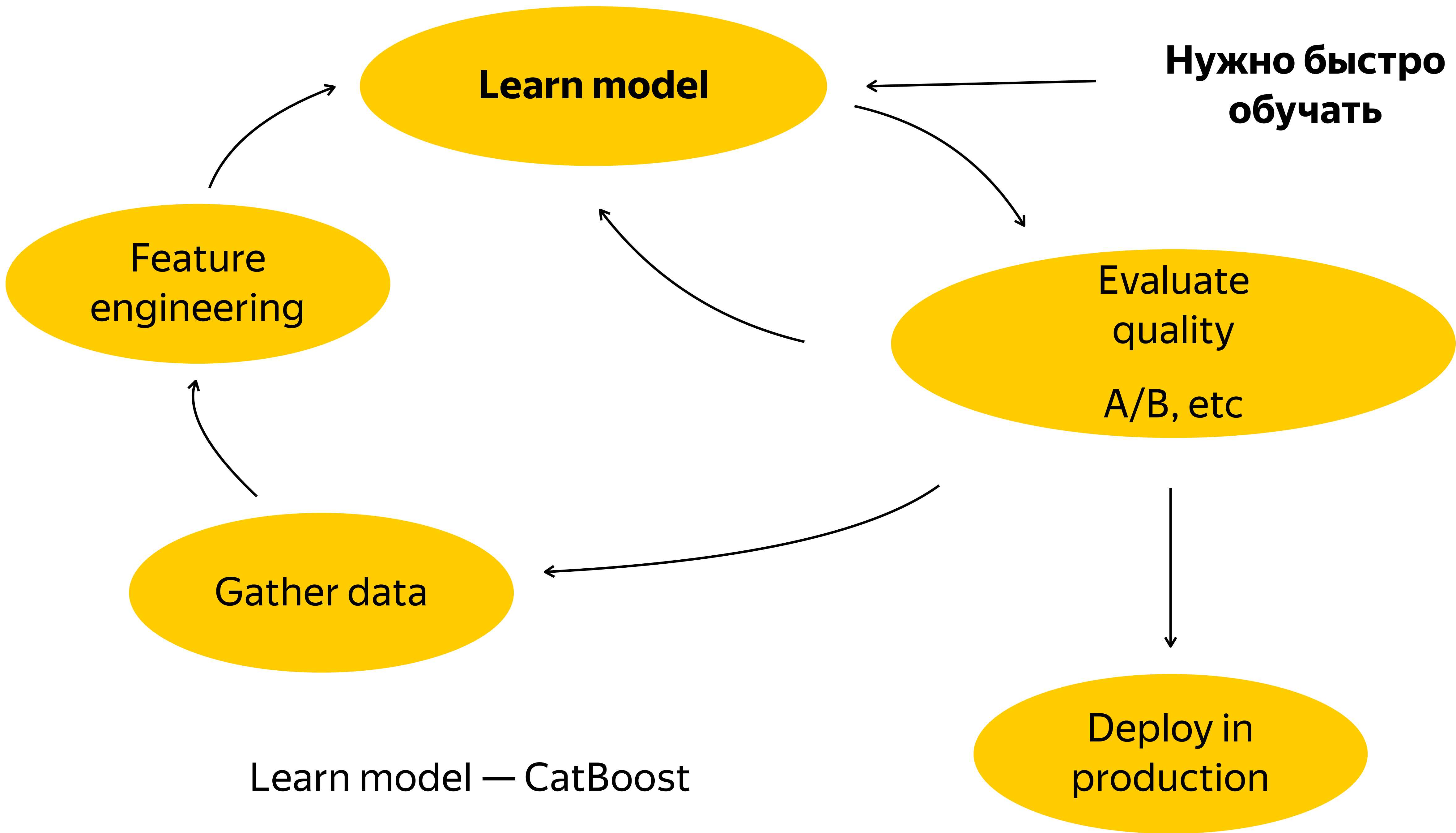
Упрощенный pipeline для ML в прикладных задачах

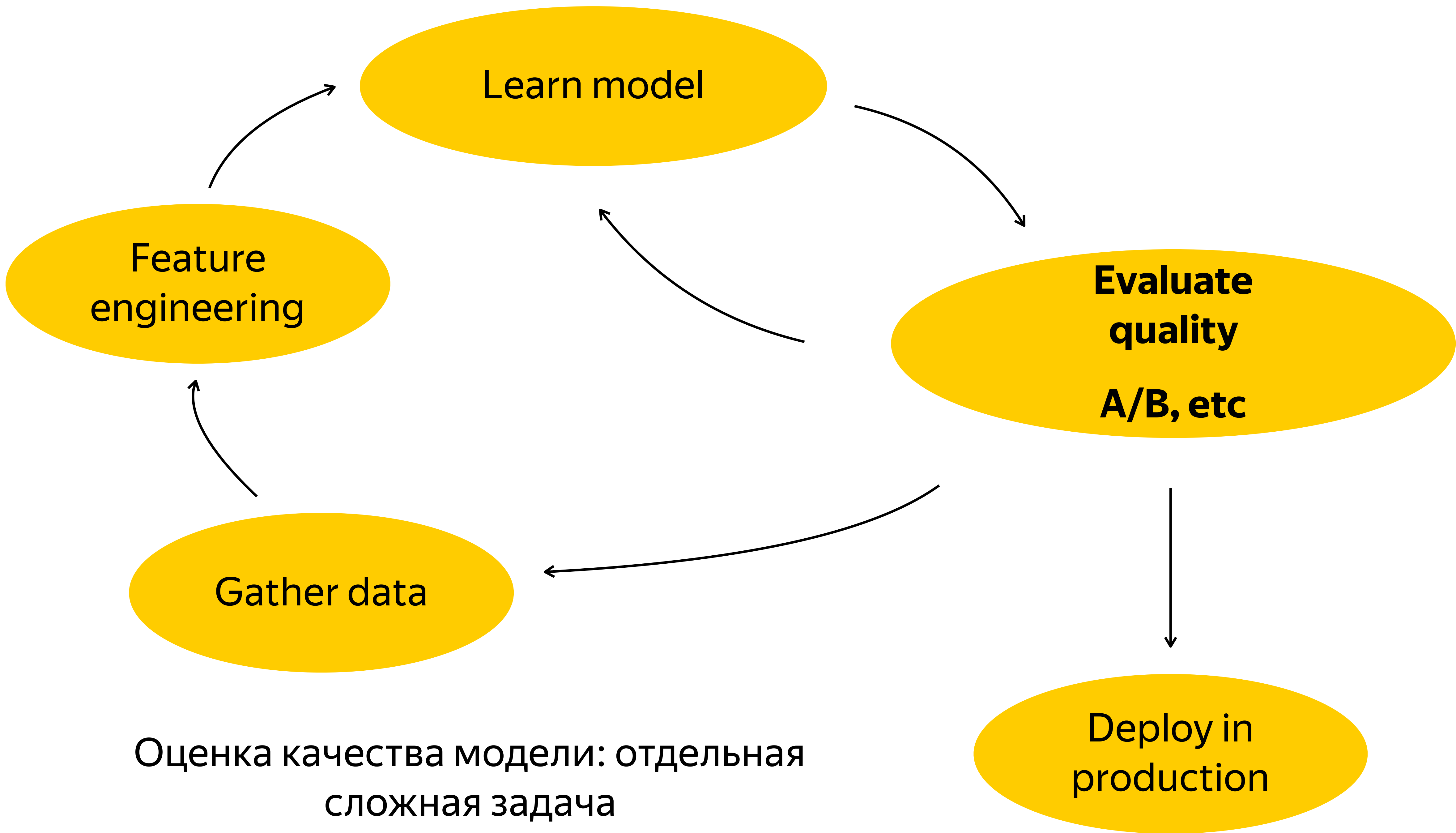


Сборка и обработка данных: отдельная сложная задача

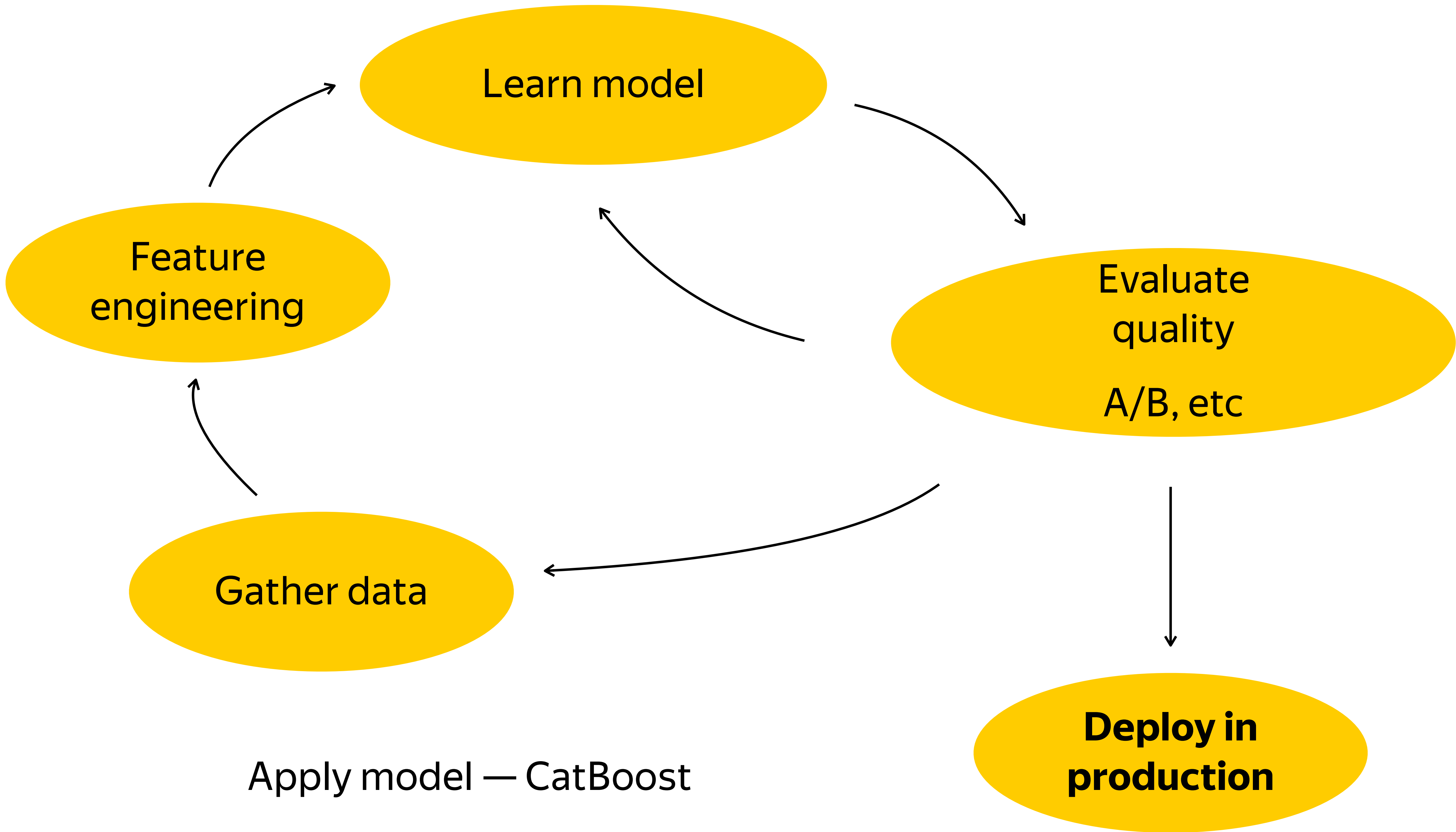


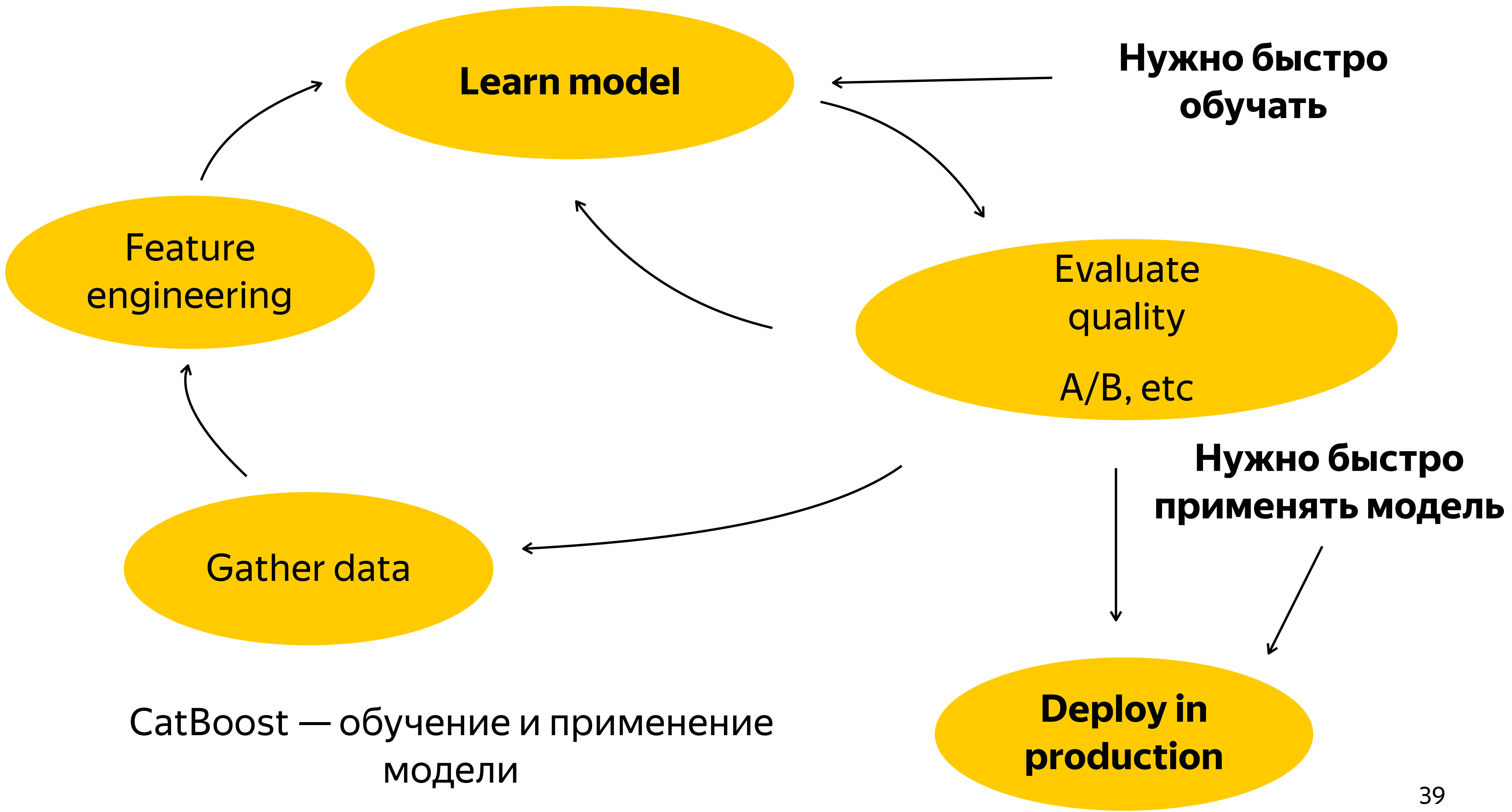
Feature engineering — mandatory step for boosting





Оценка качества модели: отдельная сложная задача





CatBoost — обучение и применение модели

Gradient Boosting в индустрии

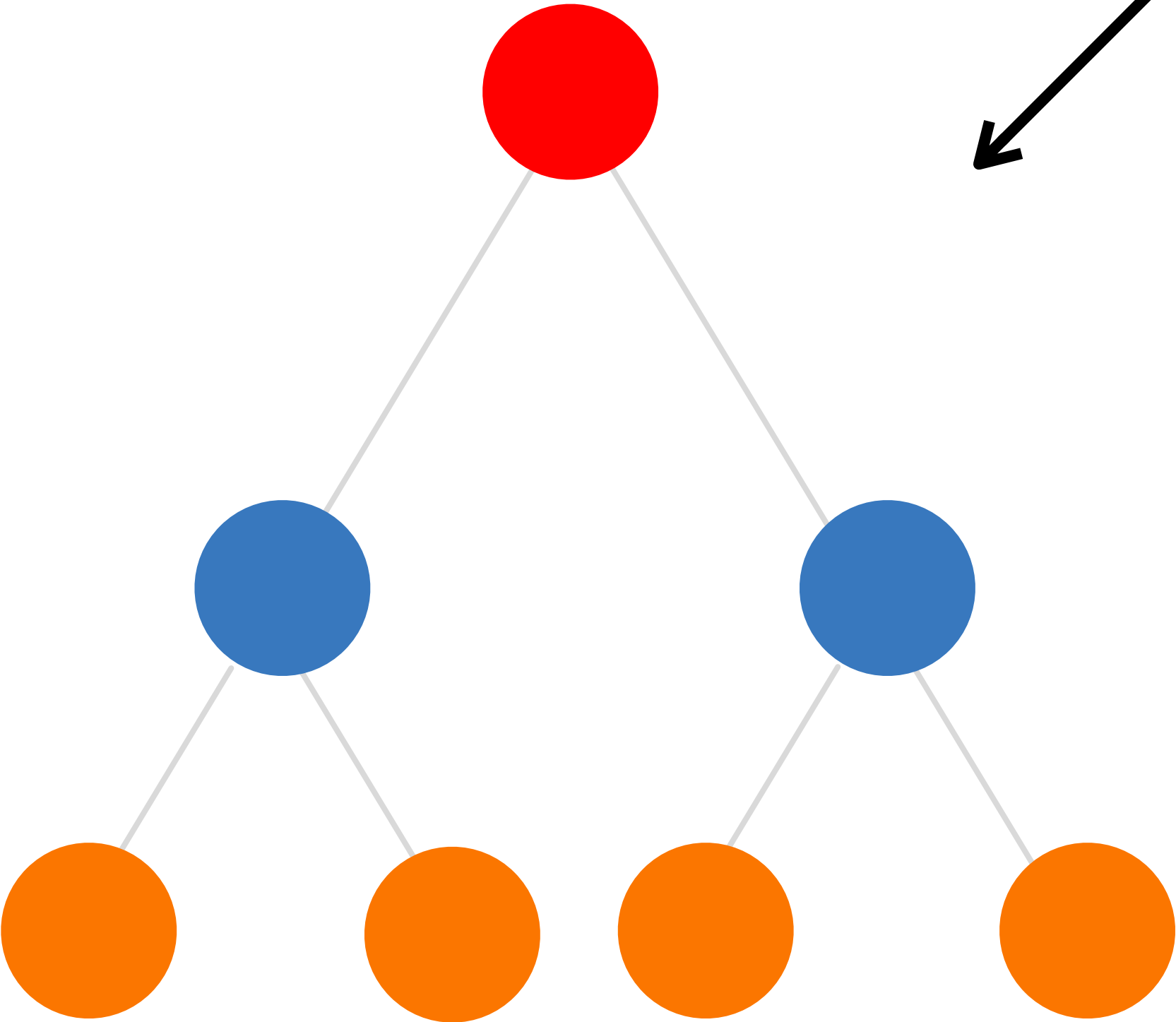
Больше данных => ~~выше качество~~ больше денег

Больше деревьев => ~~выше качество~~ больше денег

Быстрее обучение => ~~больше экспериментов~~ больше денег

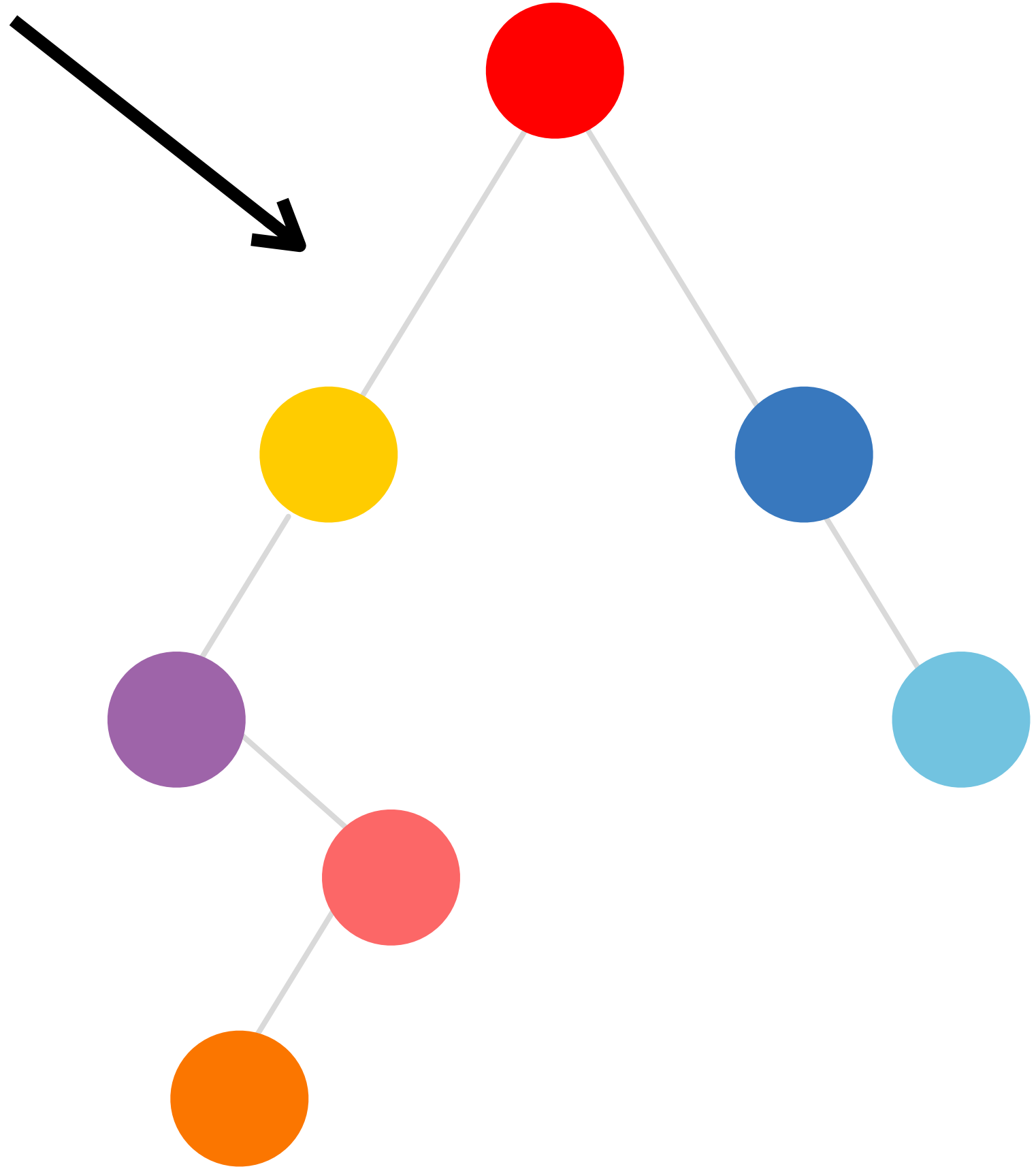
Быстро применять?

Можно



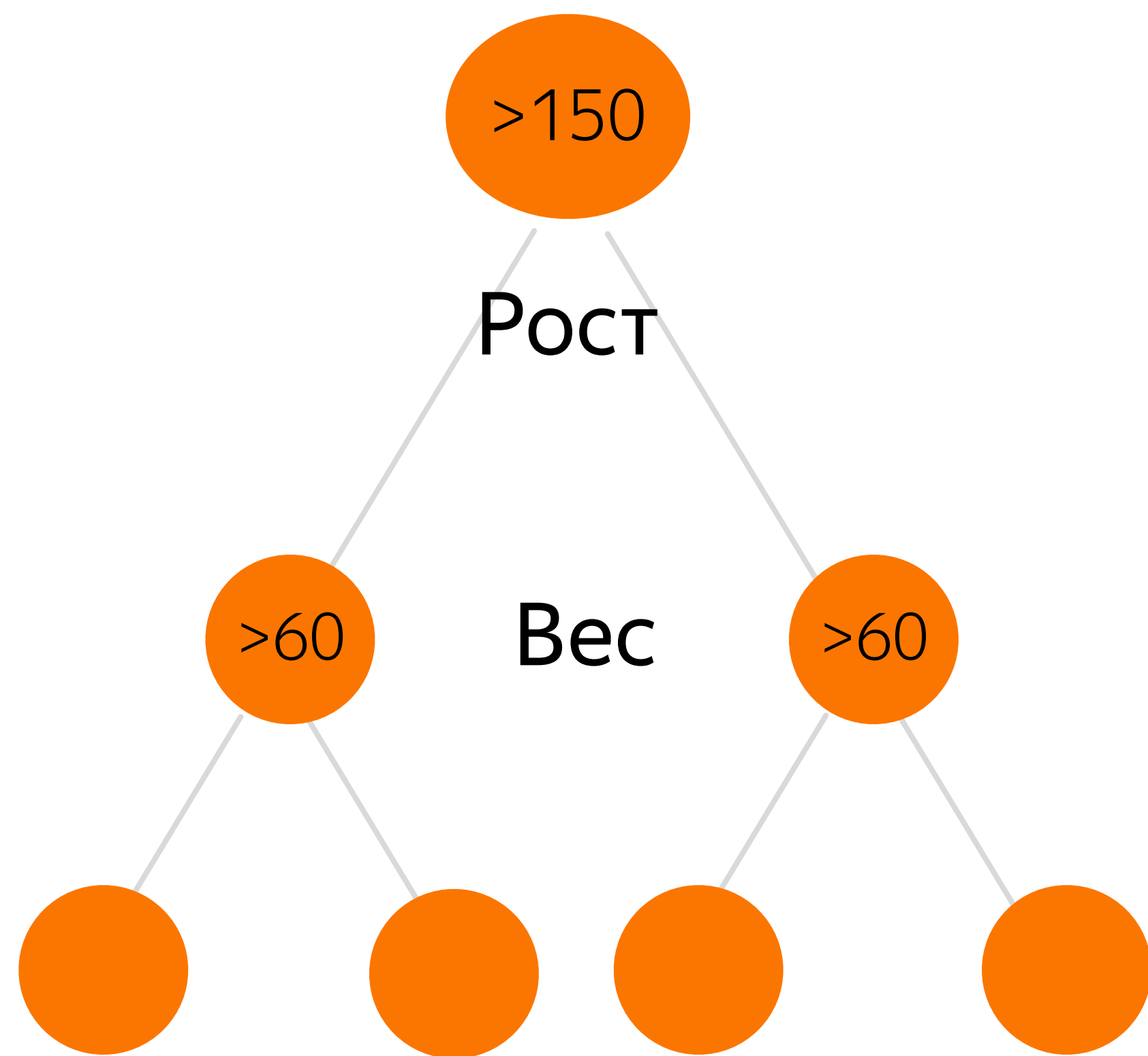
Симметричные деревья,
CatBoost

Нельзя



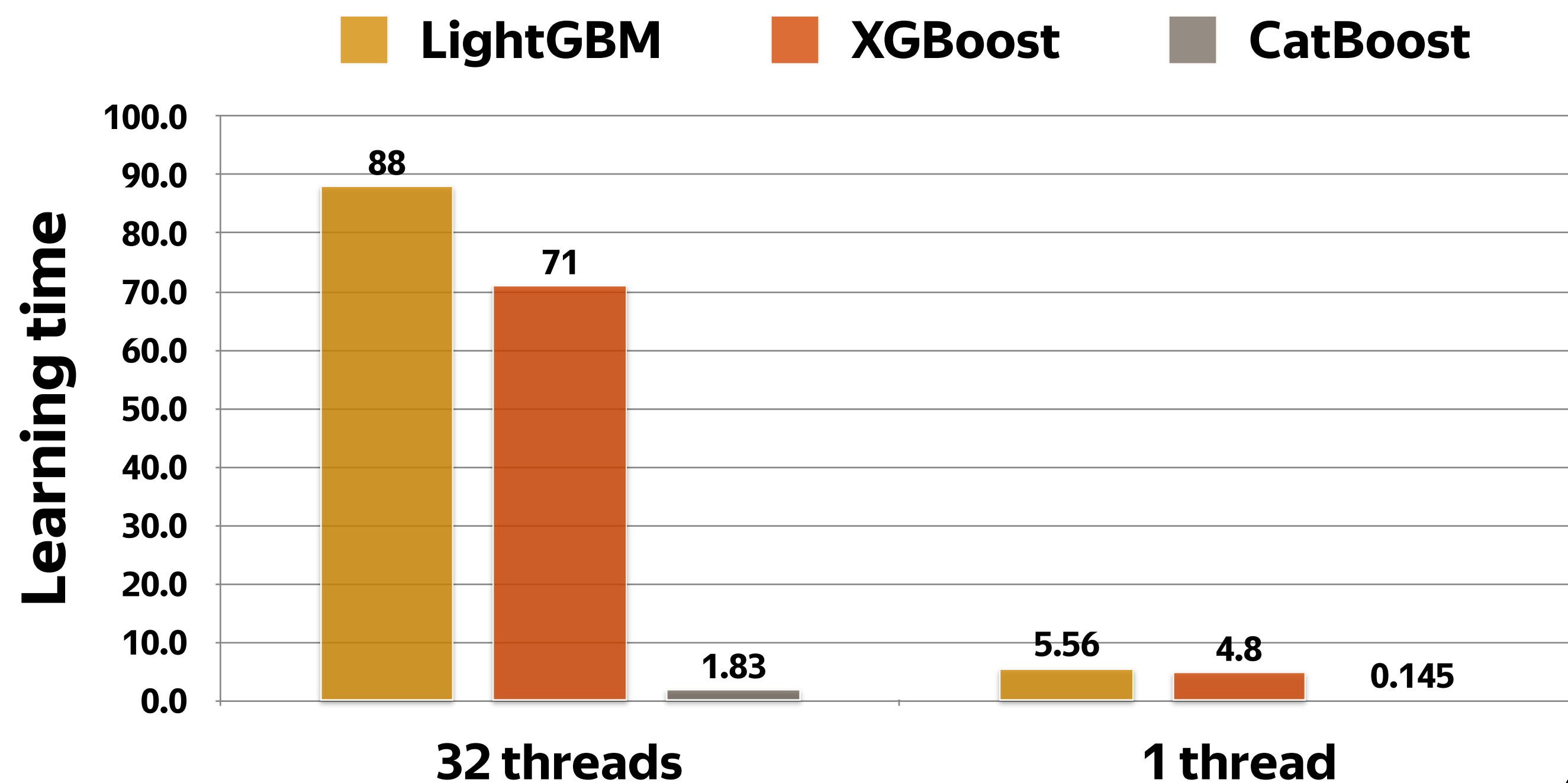
Классические деревья,
Все остальные библиотеки

Быстро применять?



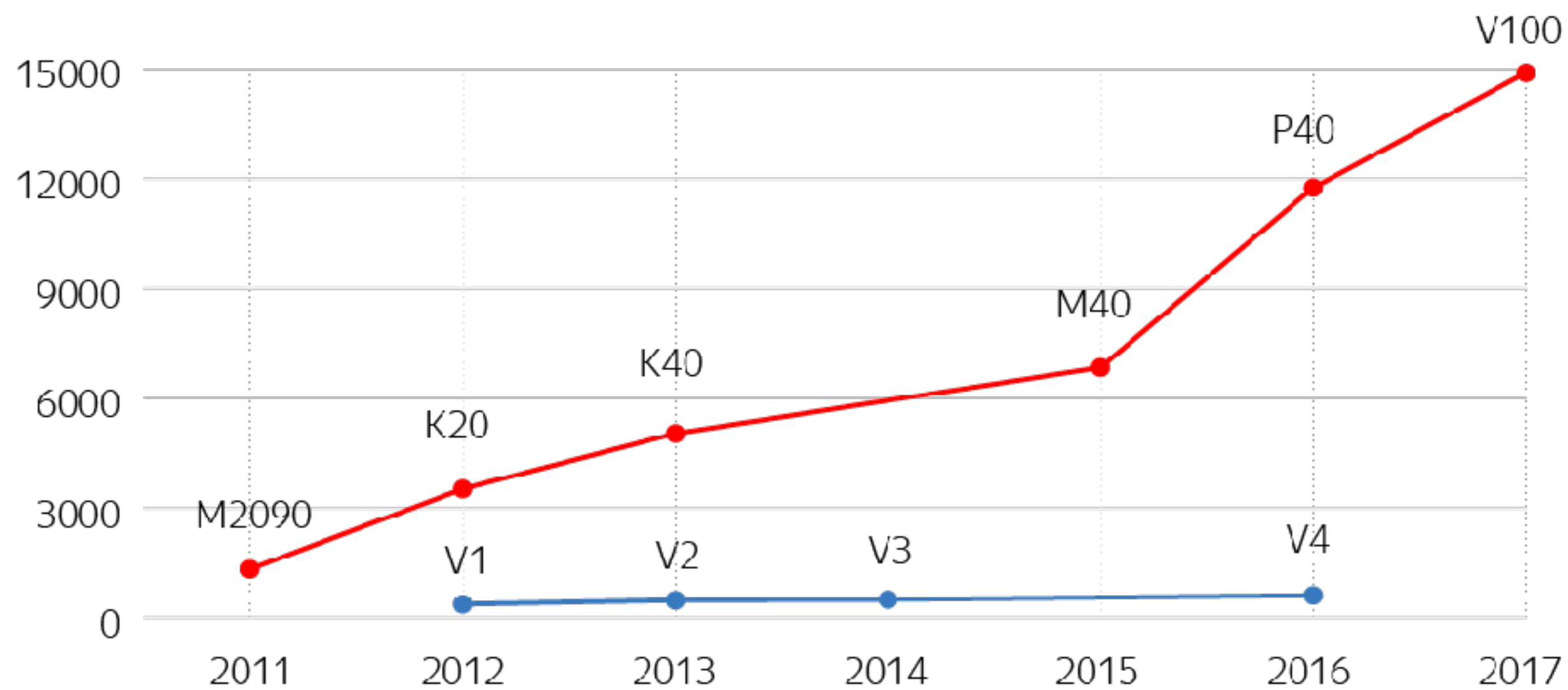
Симметричные деревья,
CatBoost

- Несколько сравнений + look-up в таблицу
- Можно использовать SSE
- За счет наших деревьев: в 20 раз быстрее конкурентов

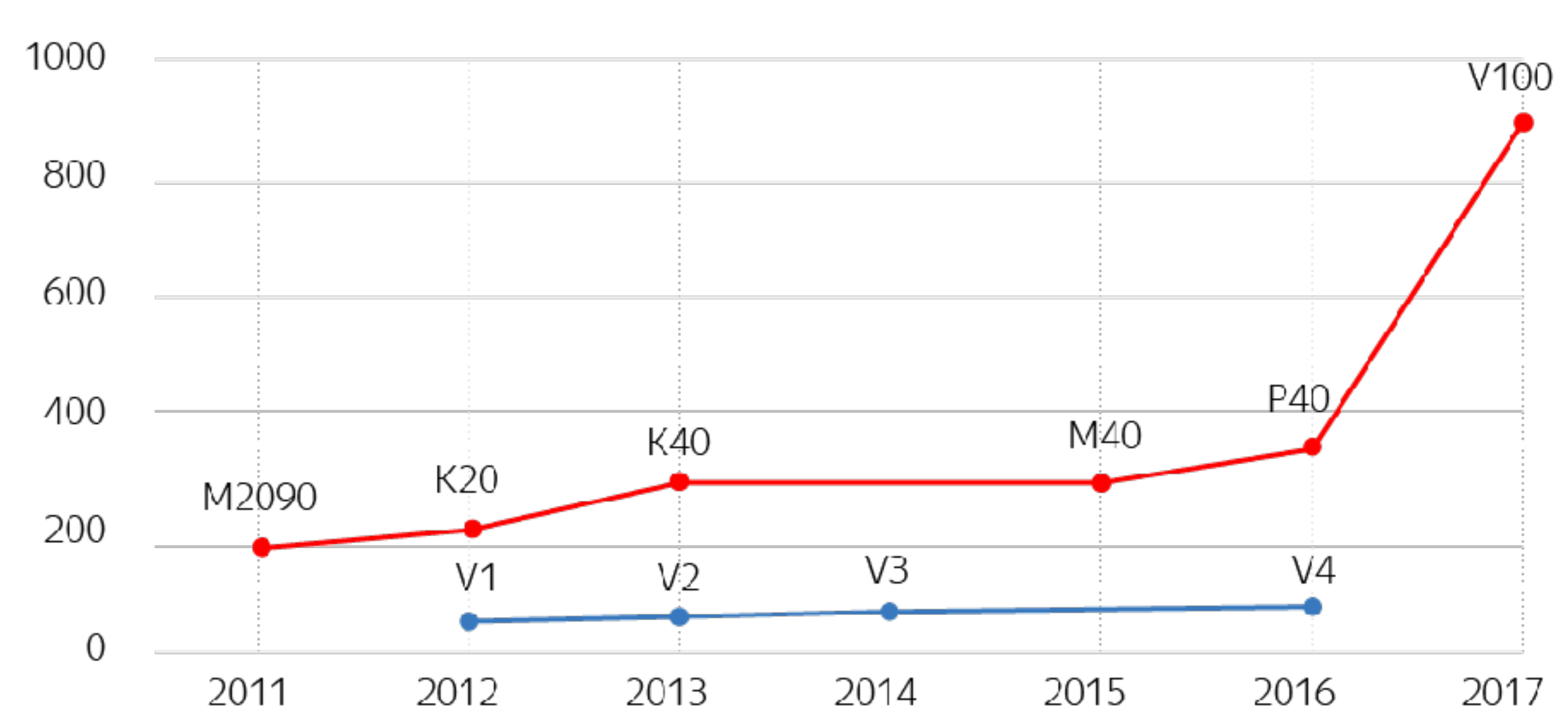


CPU vs GPU

Peak GFLOPS

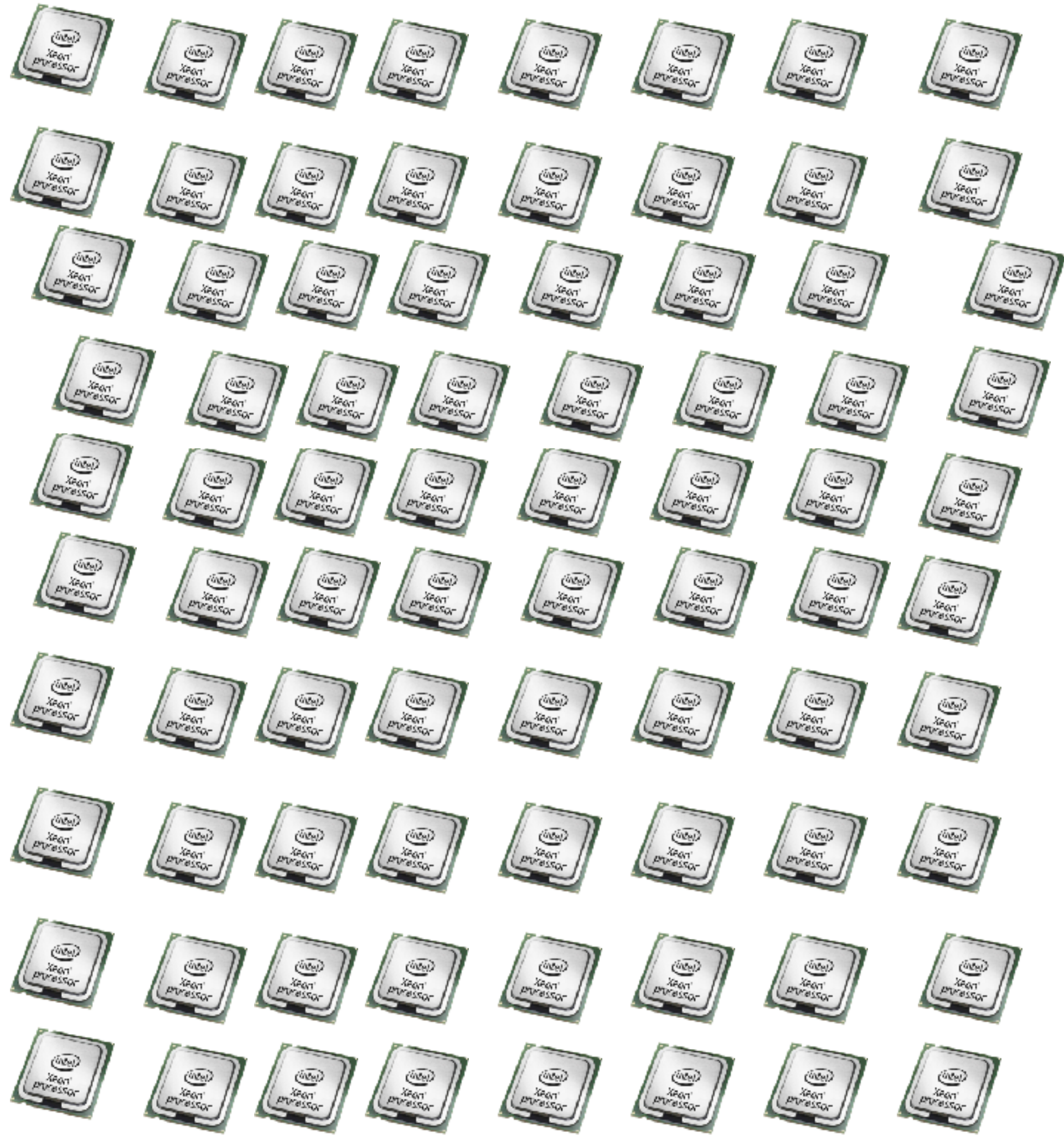


Peak memory bandwidth



GPU CPU (Intel E5-2690)

Производительность для CatBoost'a

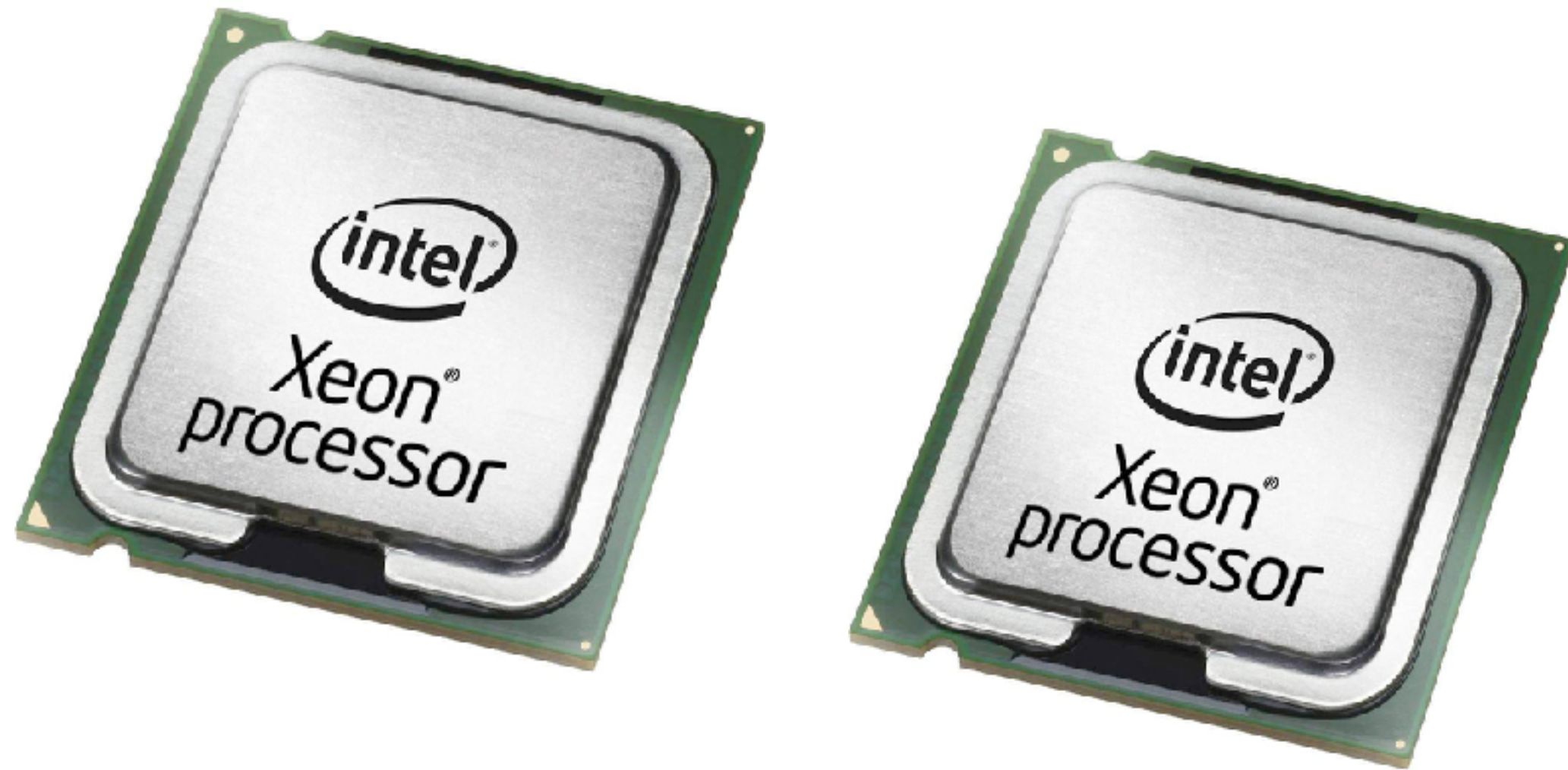


80x Intel Xeon E5-2660v4



NVIDIA Titan V

Цена



2x Intel Xeon E5-2660v4
≈ 3000\$
(amazon.com)

≈



Titan V
≈ 3000
(nvidia.com)

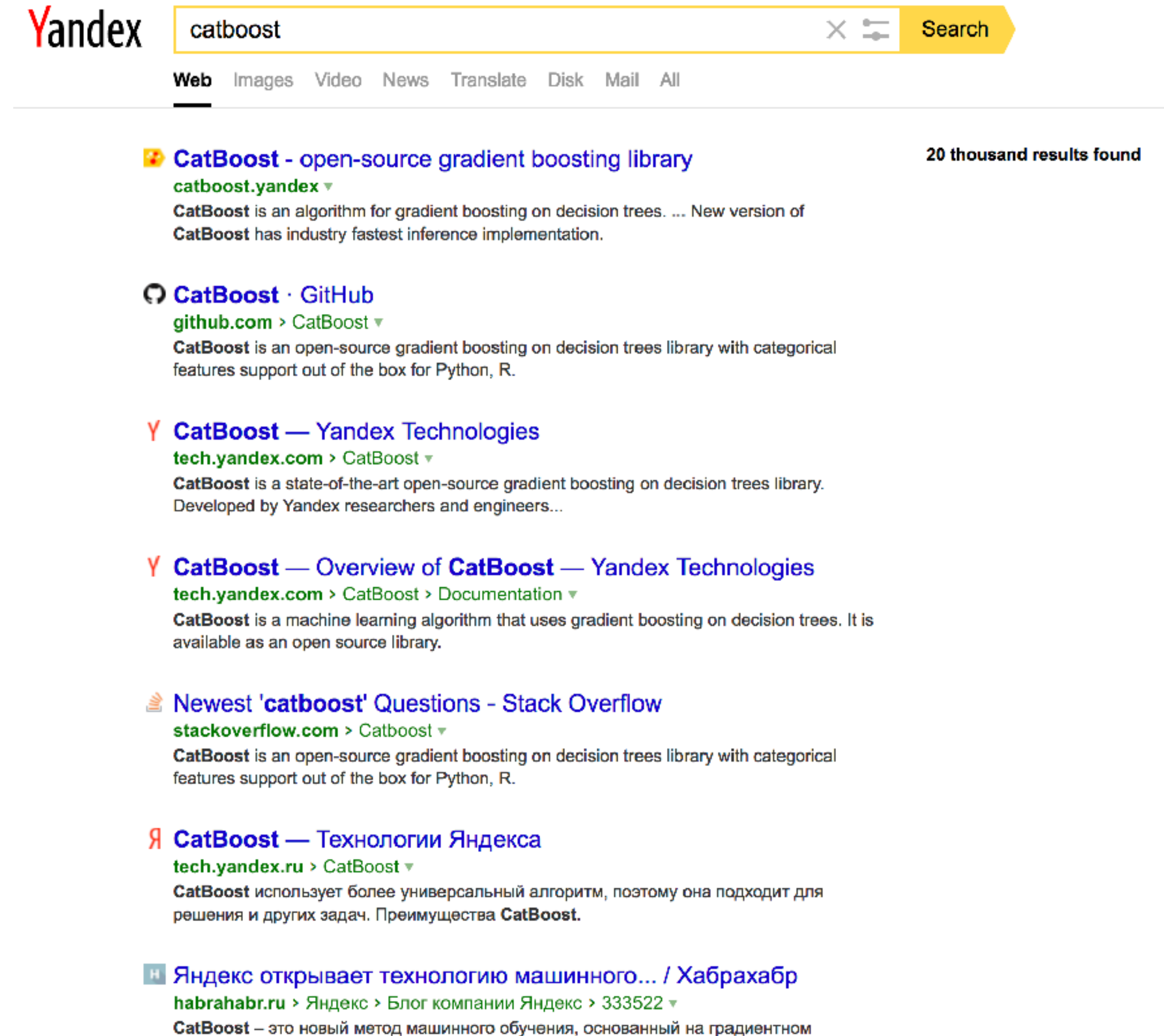
Использование в Яндексе: поиск

Время обучения CPU

- › 75 часов, 100 машин, 16-ядерный сервер

Время обучения GPU

- › 7-9 часов, один сервер с 8 Tesla P40



The screenshot shows a Yandex search interface with the query 'catboost' entered in the search bar. The search bar includes a 'Search' button and a 'Web' tab selected. Below the search bar, there are navigation links for 'Images', 'Video', 'News', 'Translate', 'Disk', 'Mail', and 'All'. The search results are displayed in a list format, with a total of '20 thousand results found' indicated on the right. 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 ✕ ↻ Search

[Web](#) [Images](#) [Video](#) [News](#) [Translate](#) [Disk](#) [Mail](#) [All](#)

20 thousand results found

- 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

Сравнение с конкурентами

Скорость обучения

- › Сравнимая с конкурентами скорость обучения на CPU
- › GPU версия в 3-20 раз быстрее конкурентов в зависимости от объема данных и режима обучения

Скорость применение

- › В десятки раз быстрее при одинаковом количестве деревьев

Качество

- › State-of-the-art на категориальных признаках
- › Сравнимое для вещественных

Спасибо за внимание!

Подробнее:

<https://catboost.ai>

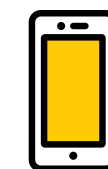


Vasily Ershov

Software developer



noxoomo@yandex-team.ru



+7 921 332 45 71



github.com/catboost