



# CatBoost: как и зачем обучать градиентный бустинг на GPU?

Vasily Ershov, Software Developer

# Содержание

- | Для каких данных использовать?

- | Возможности библиотеки и как использовать

- | Какой профит для пользователя?

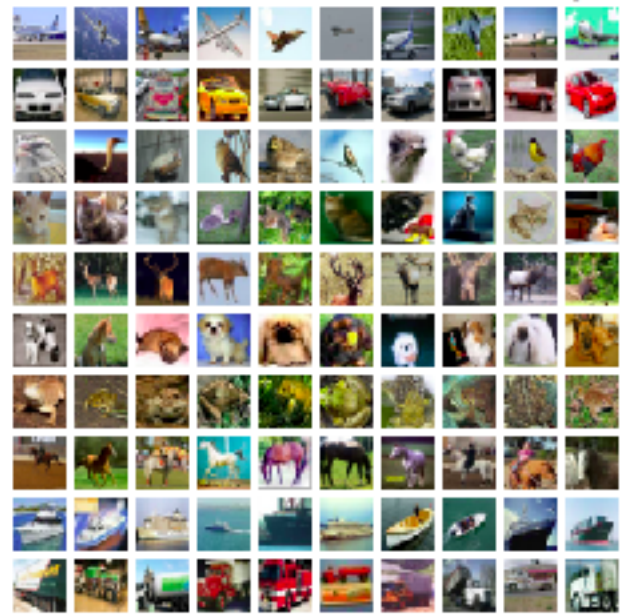
  - › GPU vs CPU

  - › CatBoost vs Competitors

  - › Real-world example

# Входные данные

Изображения



CNN

Последовательности



Текст, DNA

RNN

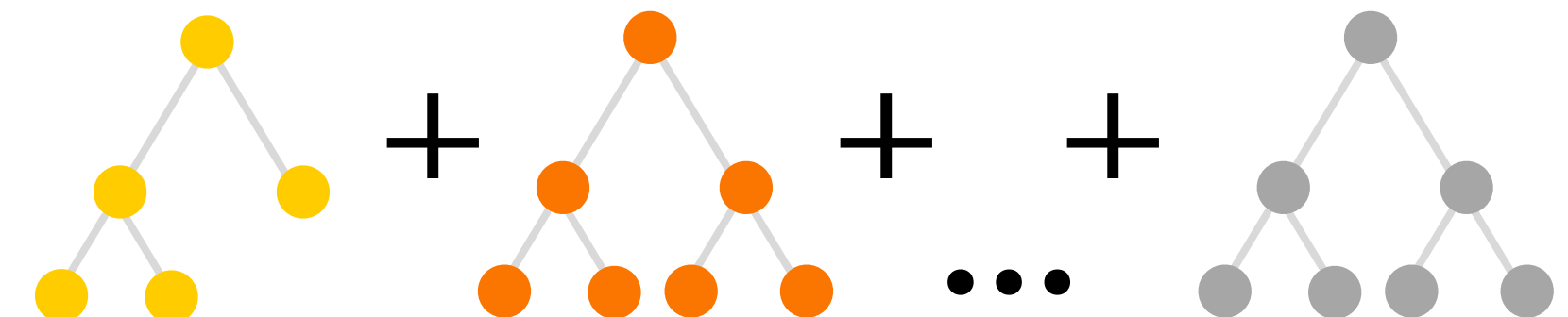
Порядковые признаки

› Music album release year

1960 < 1970 < 1980

Gradient boosted  
decision trees

**CatBoost: Categorical + Boosting**



Категориальные  
признаки

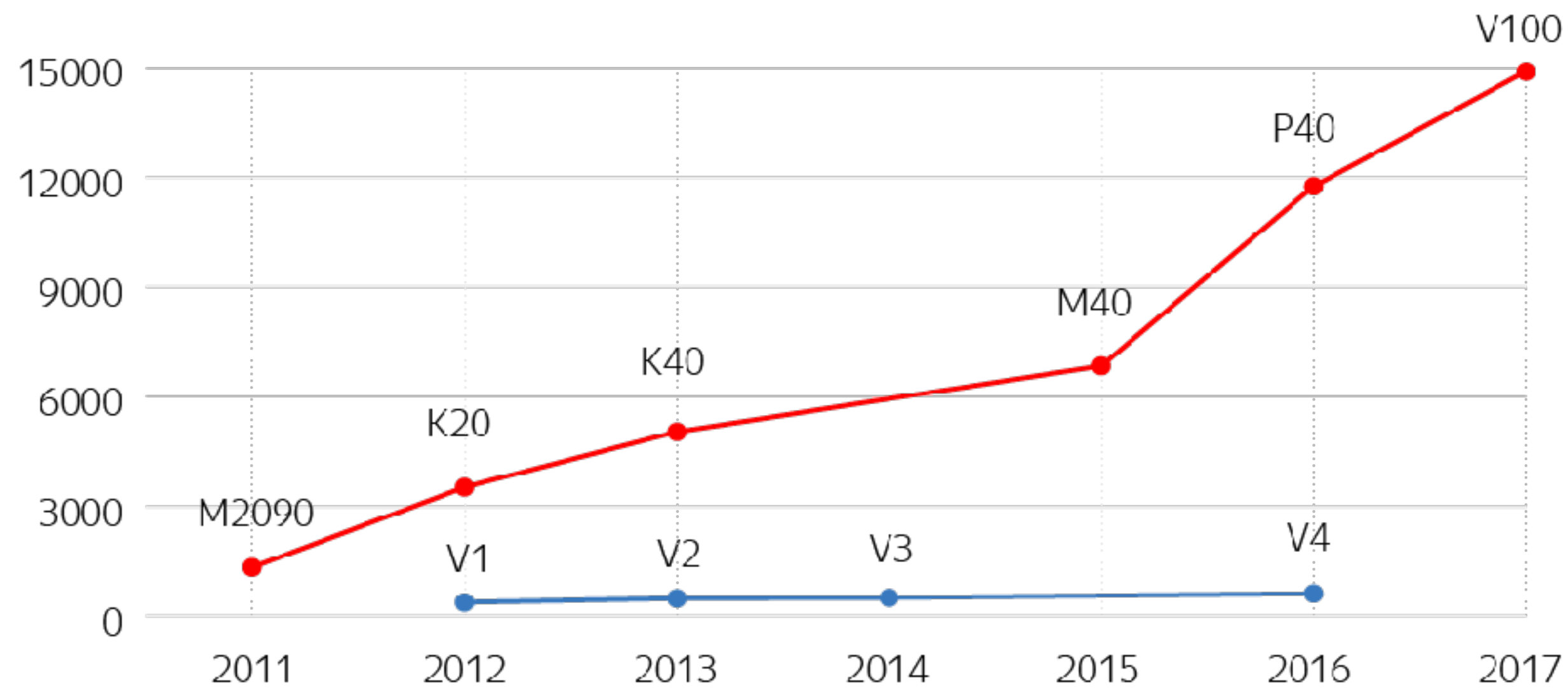


# Бустинг в индустрии

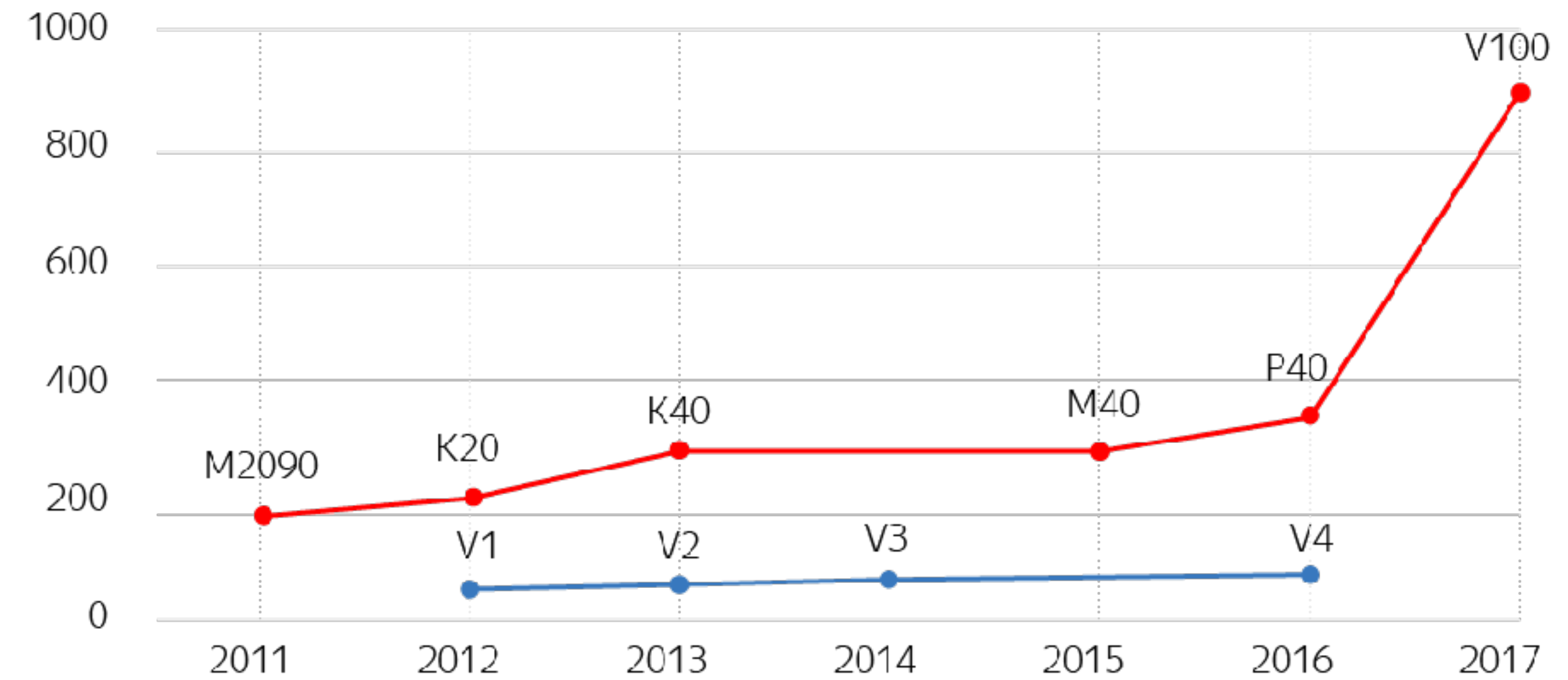
- | Больше данных => ~~выше качество~~ больше денег
- | Больше деревьев => ~~выше качество~~ больше денег
- | Быстрее обучение => ~~больше экспериментов~~ больше денег

# CPU vs GPU

## Peak GFLOPS



## Peak memory bandwidth



— GPU — CPU (Intel E5-2690)



# Объемы данных

Classical research and competitions:

› Higgs: 28 features, 11M samples, 7GB, 2014

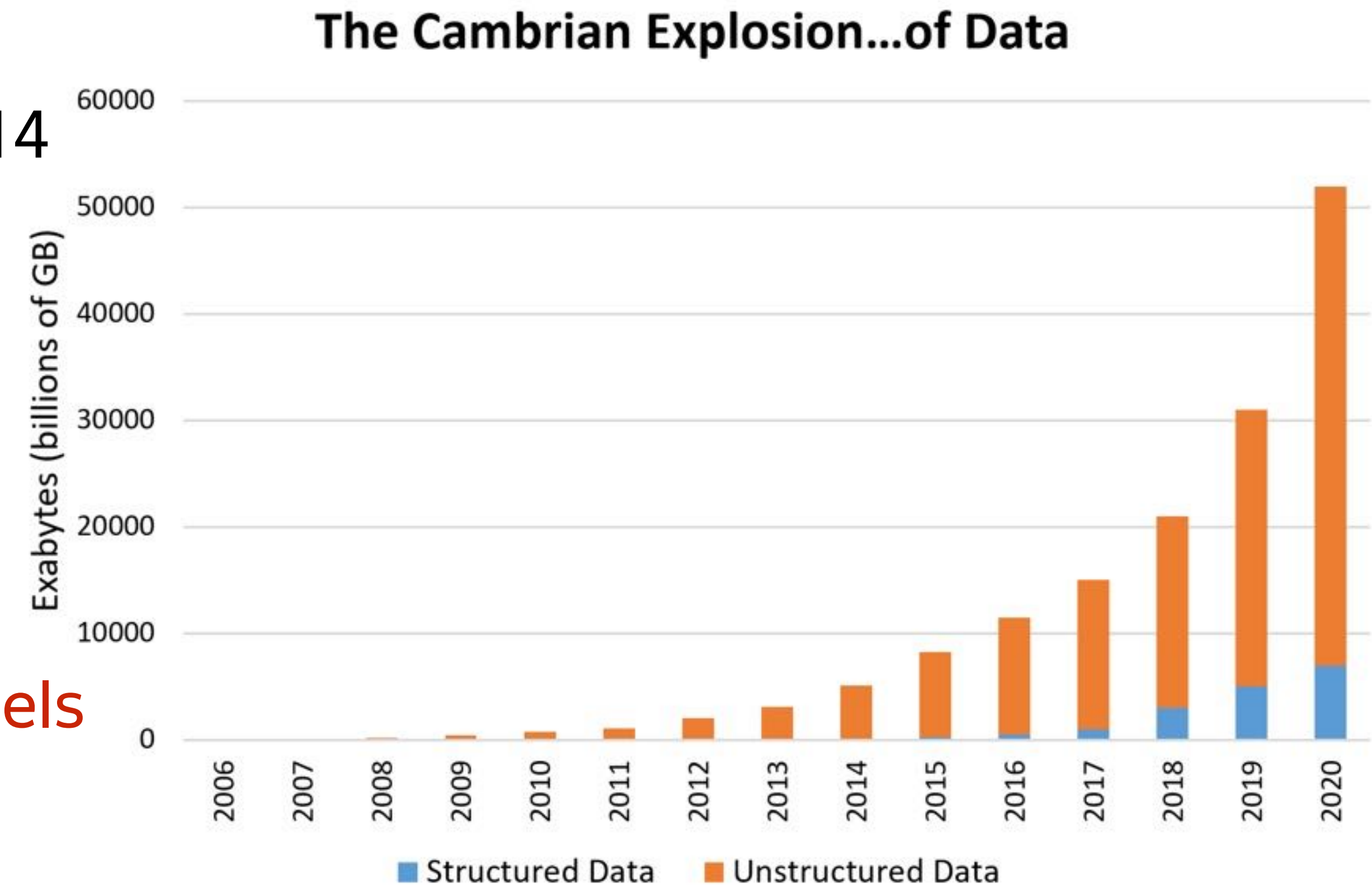
› 500MB GPU Memory, 1 GPU

Modern research and production:

› Yandex: 100GB is small

› 8 GPU, 24 GB per each for production models

CERN: as much data as you want



# CatBoost

# CatBoost: boosting + categorical

High-performance, CPU and GPU versions (MultiGPU, CUDA)

Удобная библиотека для python: `pip install catboost`

## CPU

```
1 catboost = CatBoostClassifier(iterations=1000,  
2                               learning_rate=0.02,  
3                               loss_function='Logloss')  
4 catboost.fit(ds)
```

## GPU

```
1 catboost = CatBoostClassifier(iterations=1000,  
2                               learning_rate=0.02,  
3                               task_type="GPU",  
4                               devices = [0, 2],  
5                               loss_function='Logloss')  
6 catboost.fit(ds)
```

Быстрый inference

Встроенная аналитика (“сила” признаков, графики ошибок, etc)

Поддержка категориальных признаков



# Порядковые (числовые) признаки

- Для деревьев не нужны 32-bit float в качестве признаков:
  - › “Равномерная” дискретизация на  $n$  частей, например, на основе квантилей распределений
  - › `feature_border_count`, `feature_border_type`
- Специализация вычислительных блоков под  $<2$ ,  $<16$ ,  $<32$ ,  $<64$ ,  $<128$ ,  $<255$ 
  - › 128 по-умолчанию, 32 удачный трейд-офф скорость/качество
  - › При  $<16$  в 2 раза меньший расход GPU RAM
  - › Для бинарного признака достаточно 1 бита на наблюдение

# Категориальные признаки

## One-hot encoding

- › **Никогда не делайте его вручную!!!**

- › `one_hot_max_size`

## Статистики на основе категориальных факторов

- › Зависящие от метки: оценки “вероятности успеха”

- › Не зависящие от метки: “частота категории”

## Жадный подбор комбинаций признаков

- › `gpu_cat_features_storage`, `max_ctr_complexity`

## Специальные техники для борьбы с переобучением

# Качество: LogLoss на открытых датасетах

	CatBoost	LightGBM		XGBoost		H2O	
Adult	0.269741	0.276018	+ 2.33 %	0.275423	+ 2.11%	0.275104	+ 1.99%
Amazon	0.137720	0.163600	+ 18.79 %	0.163271	+ 18.55%	0.162641	+ 18.09%
Appet	0.071511	0.071795	+ 0.40 %	0.071760	+ 0.35%	0.072457	+ 1.32%
Click	0.390902	0.396328	+ 1.39 %	0.396242	+ 1.37%	0.397595	+ 1.71%
Internet	0.208748	0.223154	+ 6.90 %	0.225323	+ 7.94%	0.222091	+ 6.39%
Kdd98	0.194668	0.195759	+ 0.56 %	0.195677	+ 0.52%	0.195395	+ 0.37%
Kddchurn	0.231289	0.232049	+ 0.33 %	0.233123	+ 0.79%	0.232752	+ 0.63%
Kick	0.284793	0.295660	+ 3.82 %	0.294647	+ 3.46%	0.294814	+ 3.52%

Подробное описание экспериментов на GitHub

# Benchmarks

# GPU vs CPU

## Hardware

- › Dual-Socket Intel Xeon E5-2660v4 as baseline
- › Several modern GPU as competitors

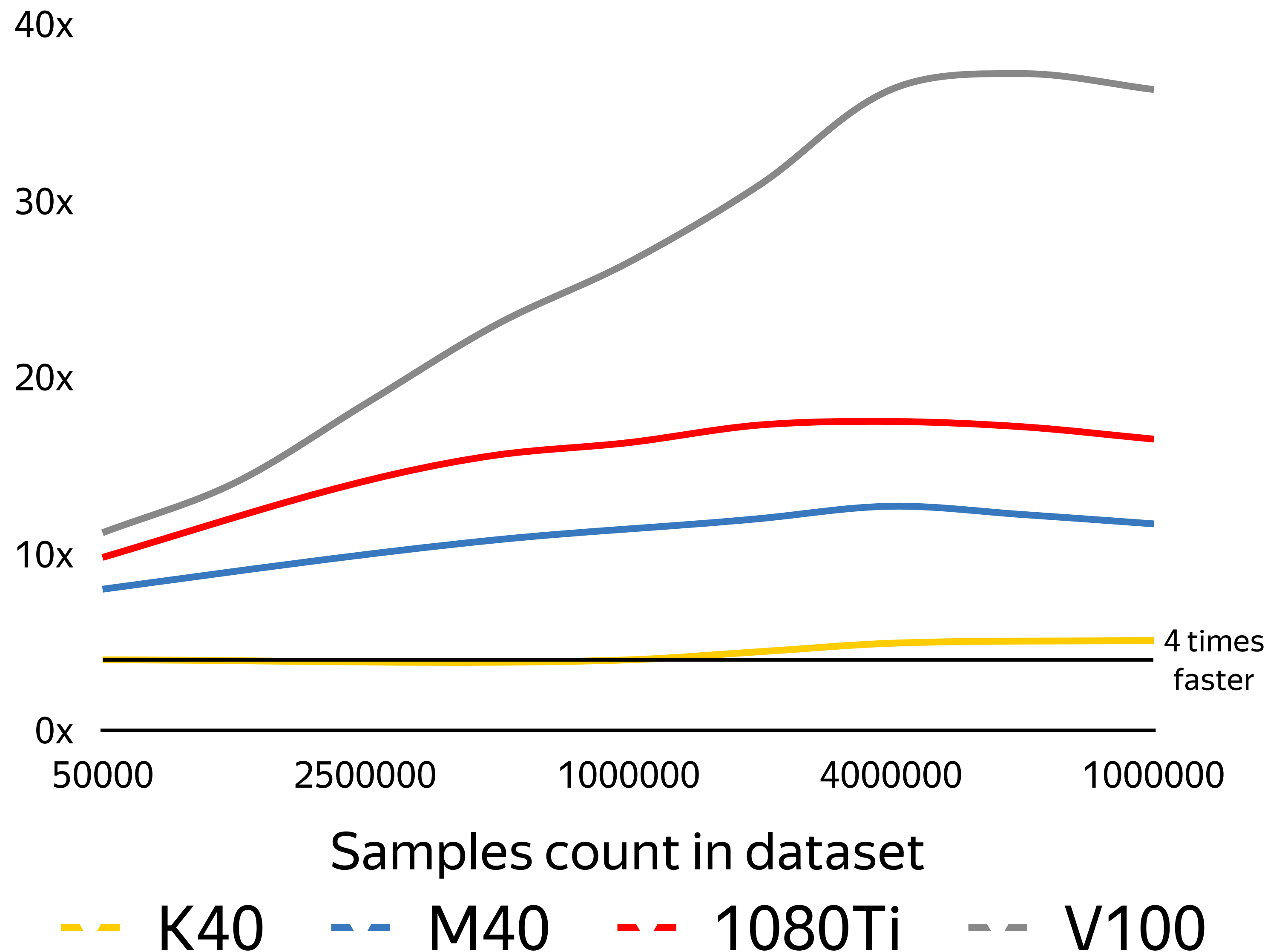
## Dataset

- ›  $\approx 800$  float features

## Price:

- › 2x Intel Xeon E5-2660v4:  $\approx 3000\$$  (amazon.com)
- › Titan V: 3000\$

## Относительное ускорение GPU по сравнению с CPU



# Сравнение с конкурентами

## Параметры

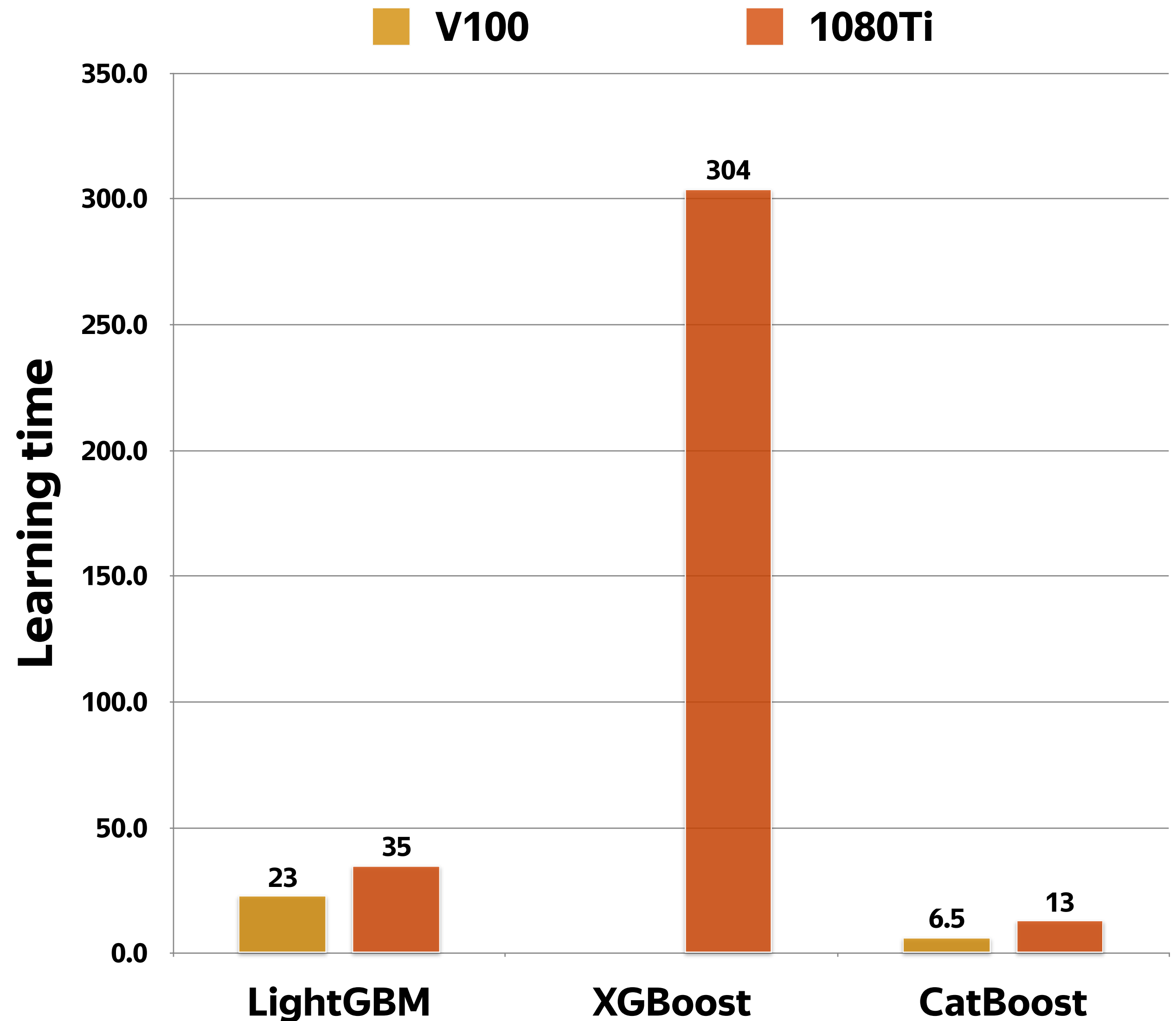
- › 32 bins, 64 leaves, 200 iterations

## Датасет:

- › ≈800 вещественных признаков
- › 4М наблюдений

## XGBoost + V100?

- › XGBoost 0.7 crashed with “Illegal Memory Access”; до 0.7 не умеет Volta, зато работает





# Сравнение с конкурентами: learn on toy datasets

## Конфигурация

- › Defaults, 64 leaves, 400 iterations
- › GPU: 1080Ti

## Higgs (classification)

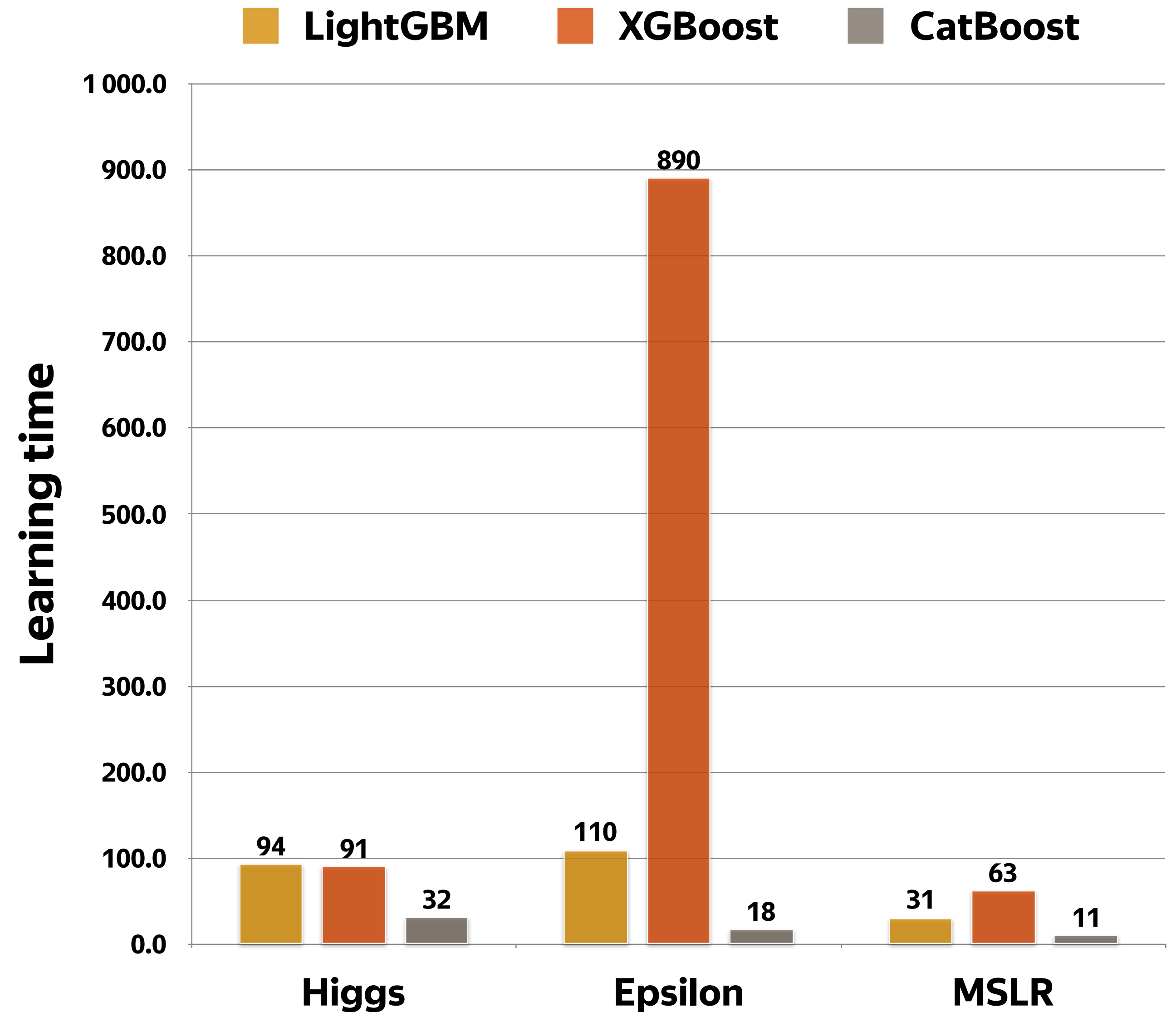
- › 28 float features, 11M samples

## MSLR (regression)

- › 136 float features, 3M samples

## Epsilon (classification)

- › 2000 float features, 400k samples



# Сравнение с конкурентами: inference

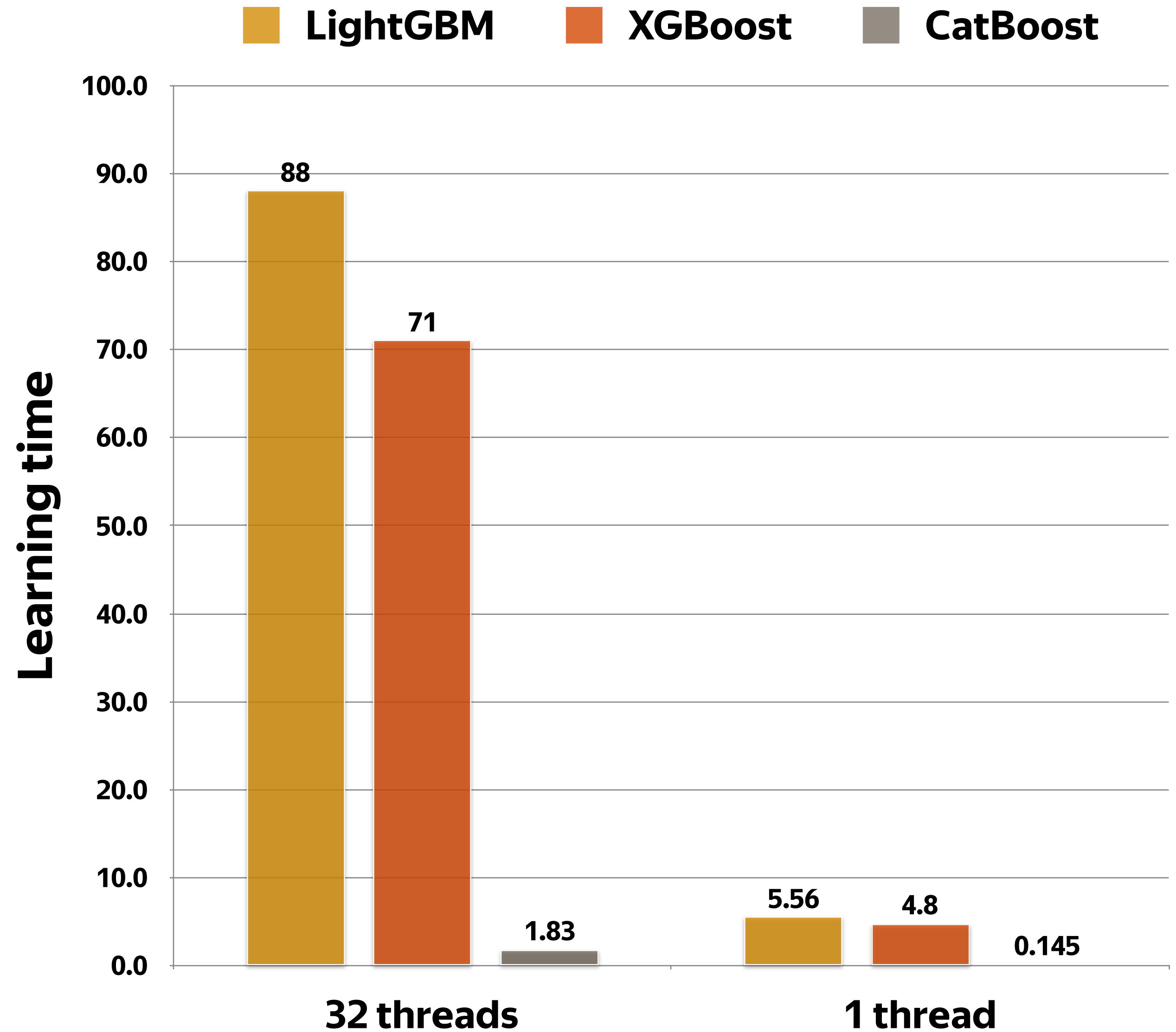
- 8к деревьев

- › 64 листа

- Epsilon

- › 2000 вещественных признаков

- › 100к наблюдений



# Использование в Яндексе

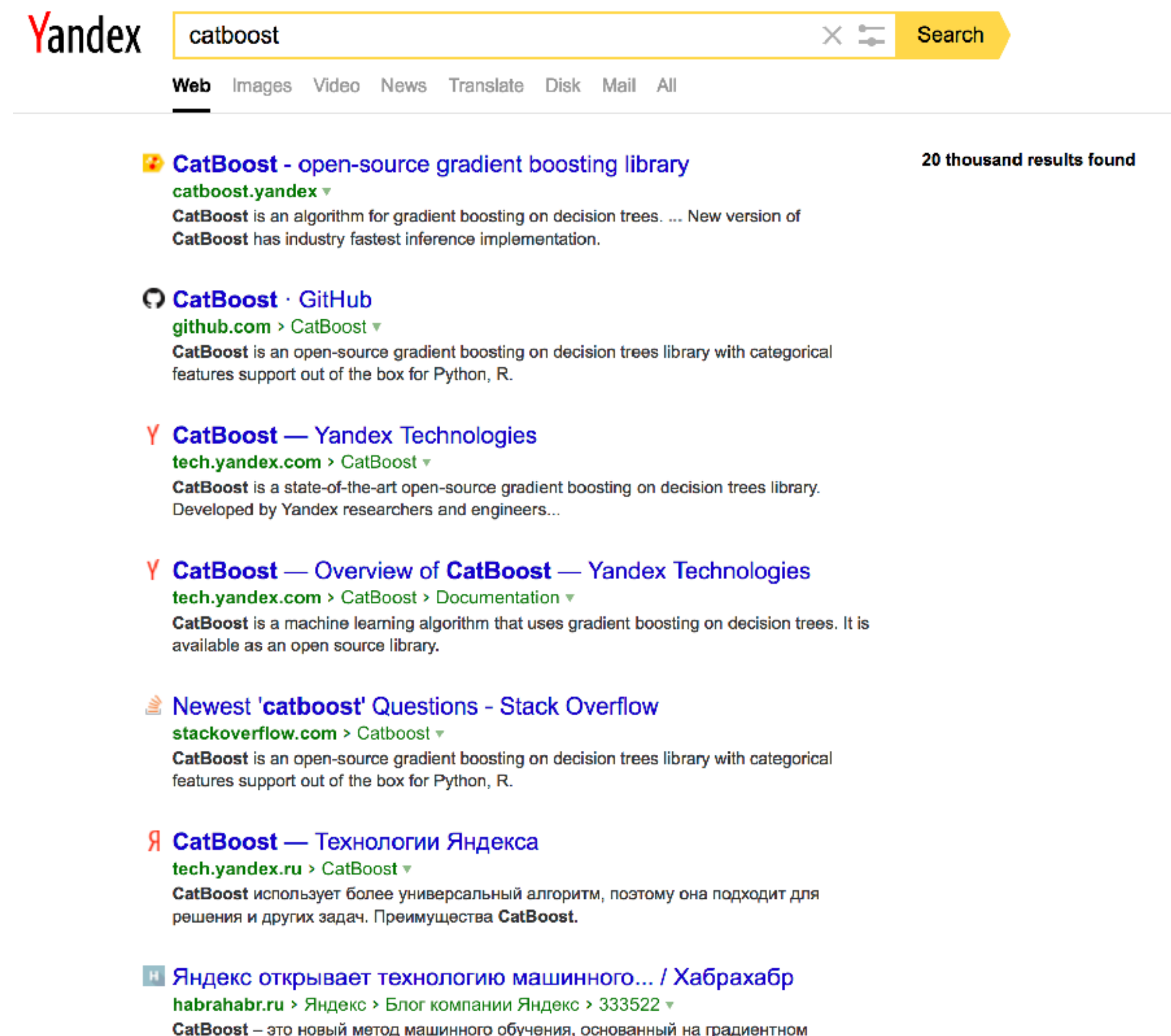
Пока используем MatrixNet,  
но активно переходим на CatBoost

Формулы ранжирования:

- › CPU: 75 часов, 100 машин, 32 ядра
- › GPU: 7-9 часов, сервер с 8P40

Больше денег => больше данных

- › CatBoost первый распределенный open-source GBDT на GPU



The screenshot shows a Yandex search interface with the query 'catboost' entered in the search bar. The search results are displayed below the navigation tabs (Web, Images, Video, News, Translate, Disk, Mail, All). The results include:

- CatBoost - open-source gradient boosting library** (catboost.yandex) with 20 thousand results found. Description: CatBoost is an algorithm for gradient boosting on decision trees. ... New version of CatBoost has industry fastest inference implementation.
- CatBoost · GitHub** (github.com > CatBoost). Description: 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). Description: 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). Description: 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). Description: 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). Description: 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). Description: CatBoost – это новый метод машинного обучения, основанный на градиентном

# Спасибо за внимание!

Подробнее:

<https://catboost.yandex>



Vasily Ershov  
Software developer



noxoomo@yandex-team.ru



+7 921 332 45 71