Yandex

Yandex

CatBoost: Fast And Scalable Gradient Boosting On GPU

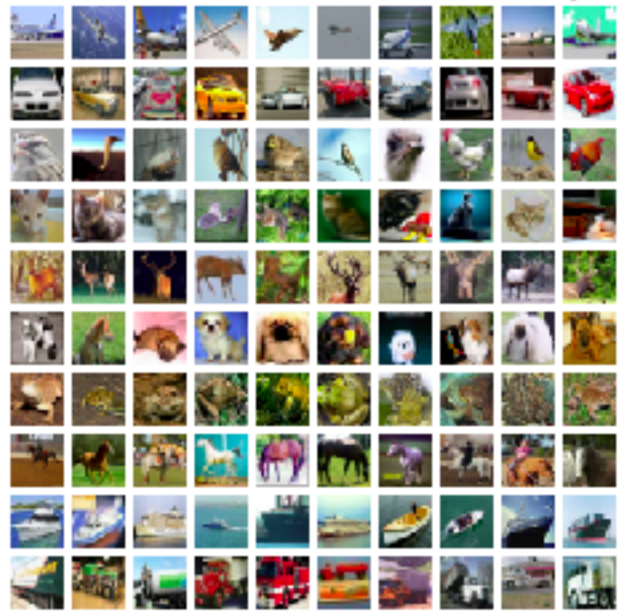
Vasily Ershov, Software Developer

Content

- More data => More profit
- CatBoost: decision trees could be done efficiently on GPU
- Benefits to users
  - GPU vs CPU
  - CatBoost vs Competitors
  - Solving real-world tasks in Yandex

Learn data

Images



CNN

Sequence

Text, DNA

RNN

Ordered features

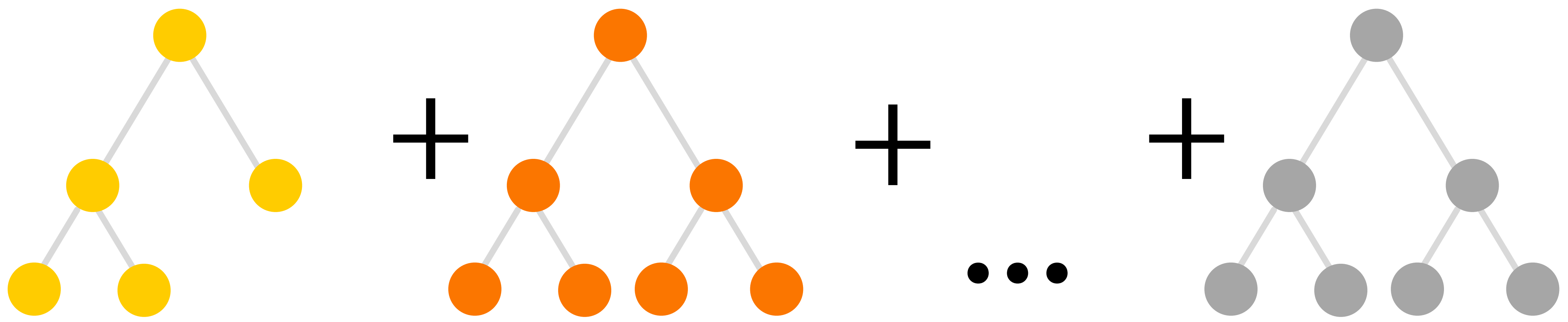
〉 Music album release year

1960 < 1970 < 1980

Gradient boosted decision trees

Categorical features

CatBoost: Categorical + Boosting



DataSet sizes

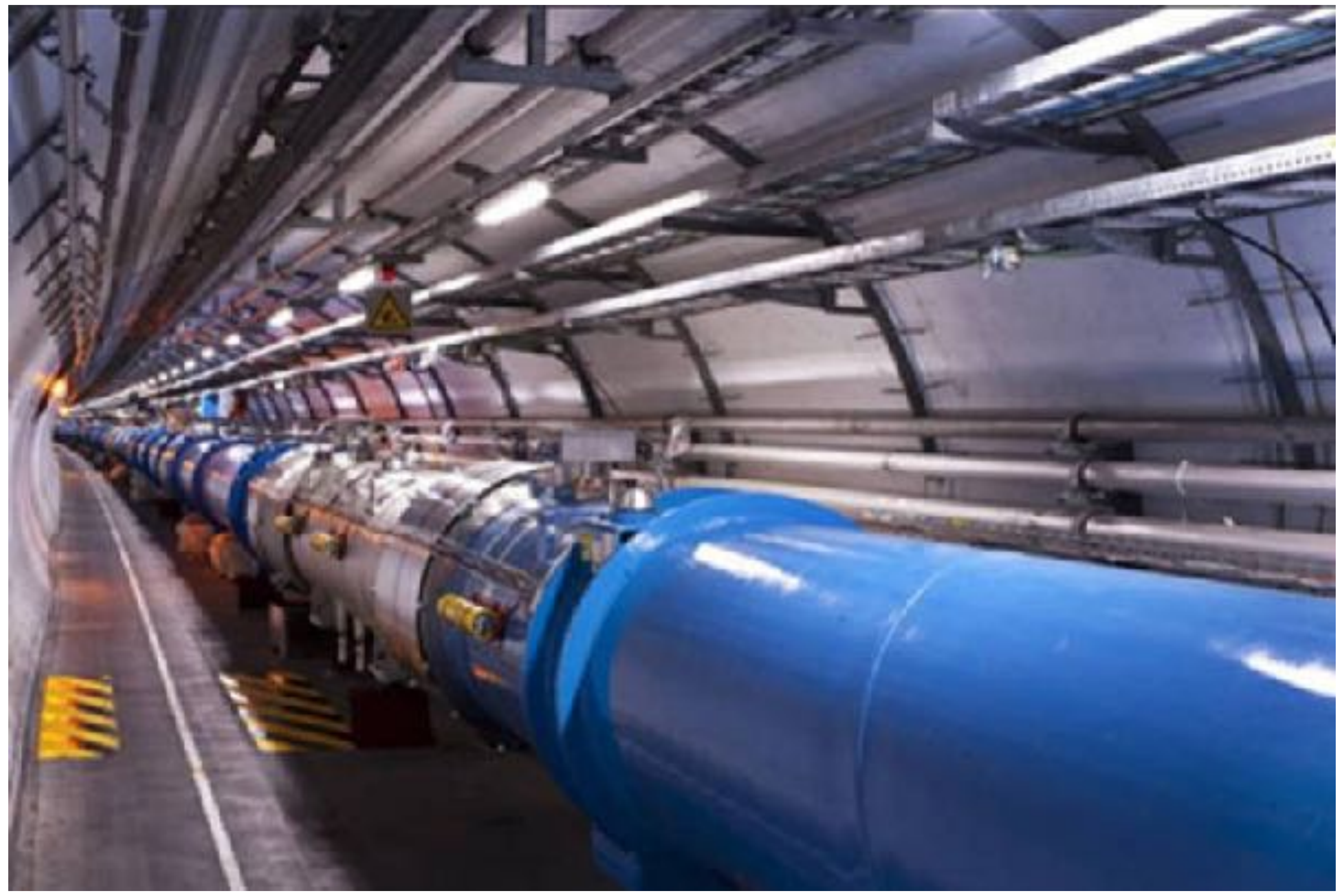
Classical research and competitions:

- 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



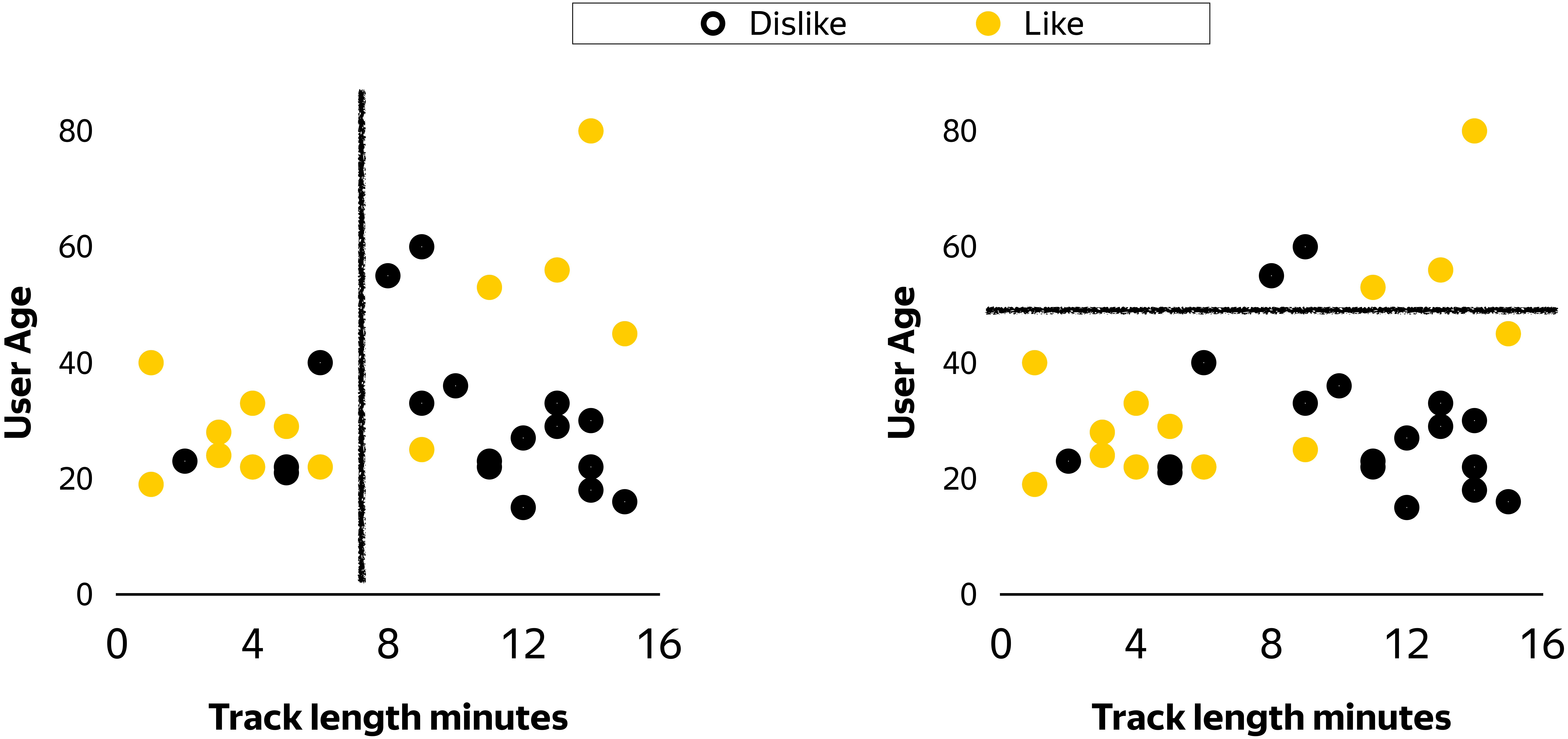
Could we use GPU for CatBoost?

GPU could efficiently handle both feature types:

- ordered: histograms computation for decision trees
- categorical: scatter/gather + radix sort + segmented primitives

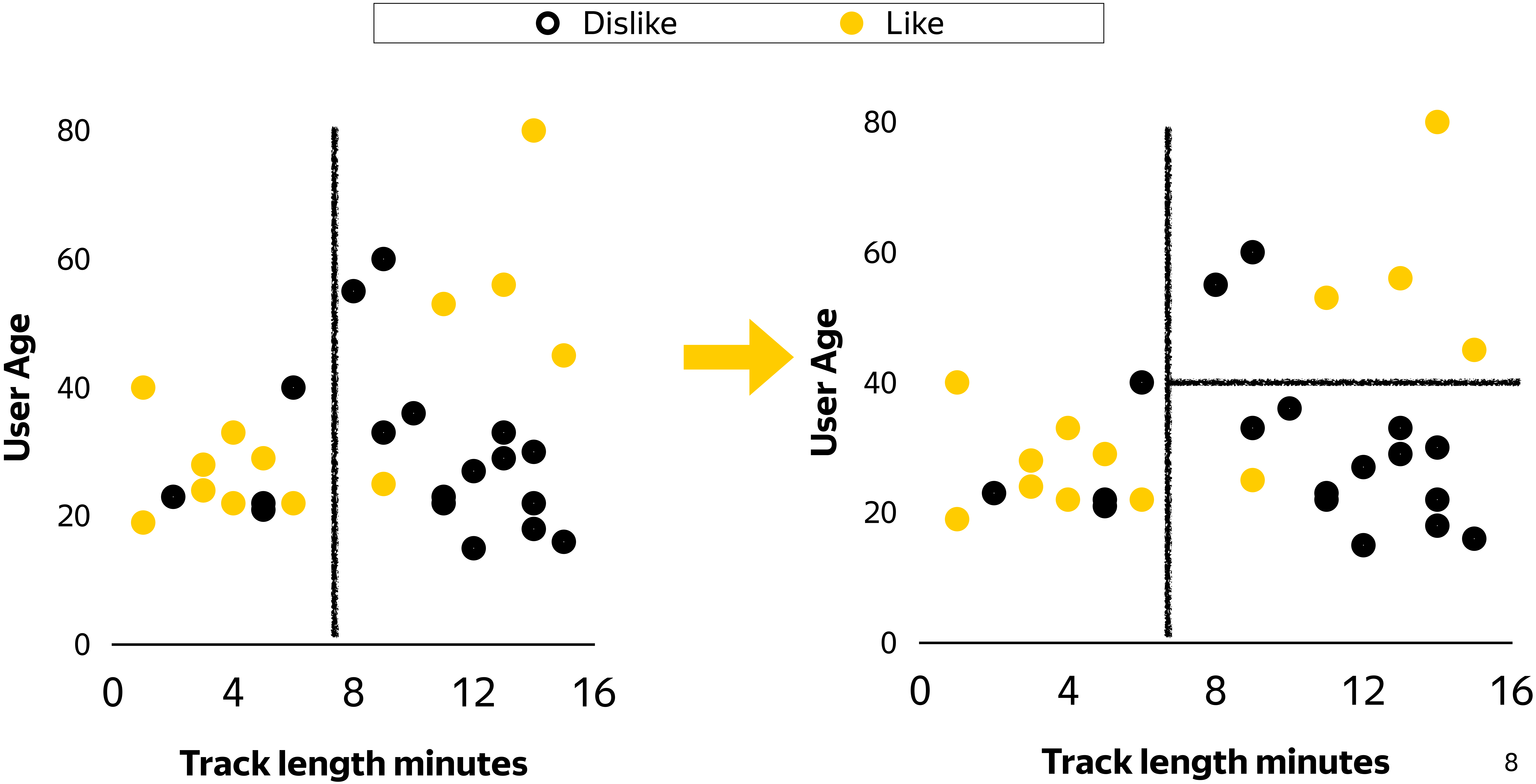
Today: only most important block to deal with ordered features

Decision tree: classification

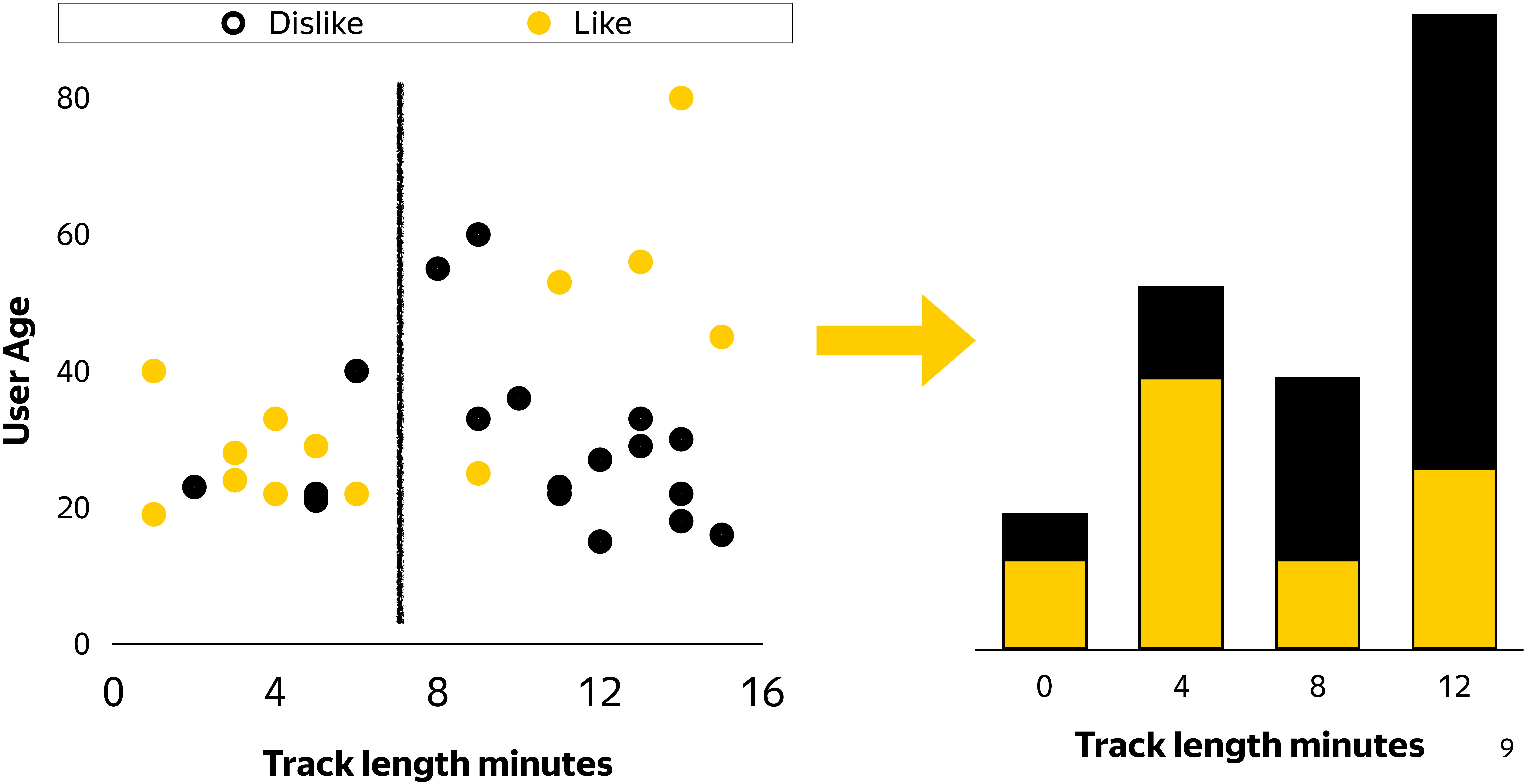


Best split: Track length minutes > 8

Decision tree: classification

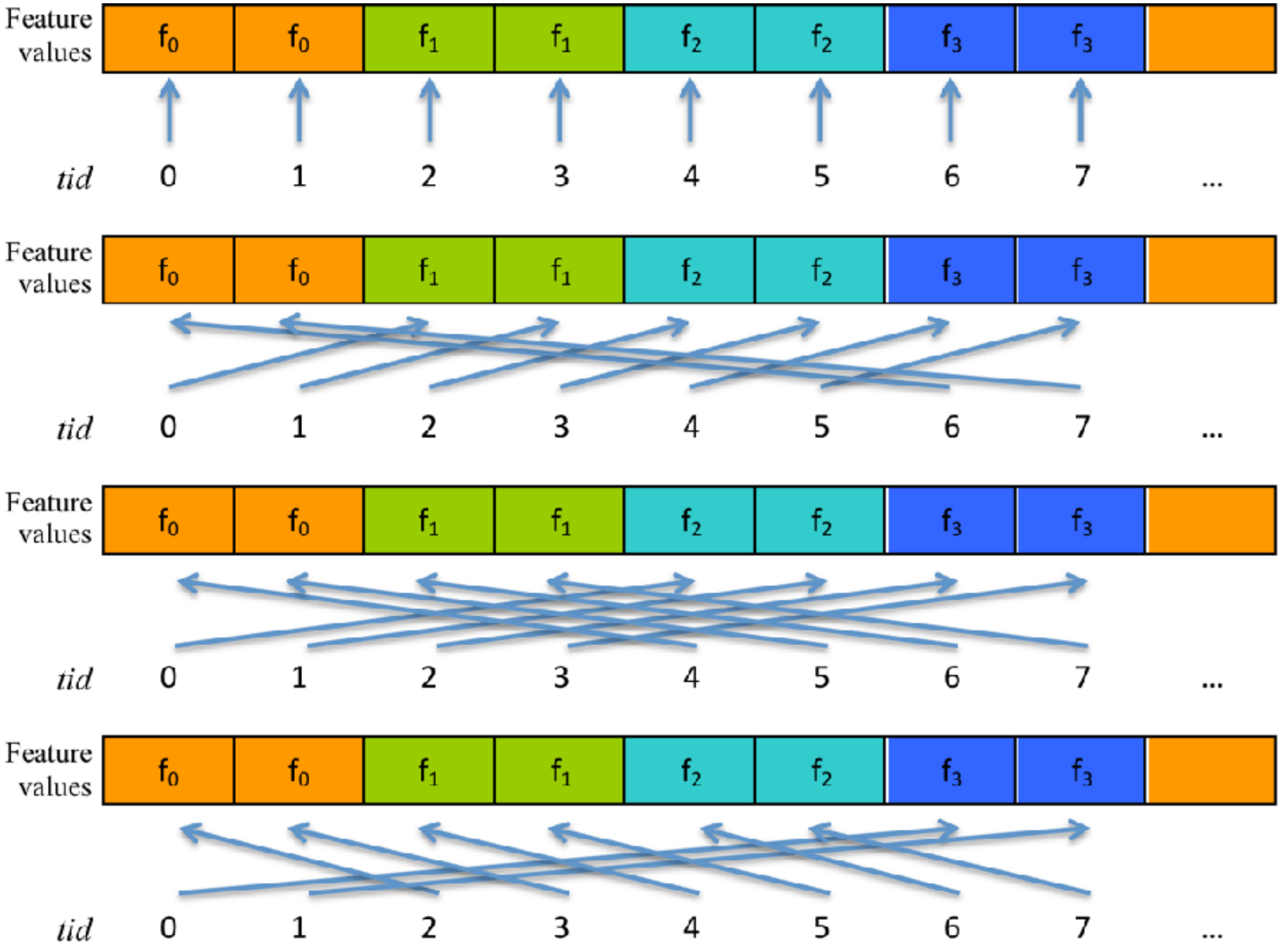


Decision tree: classification



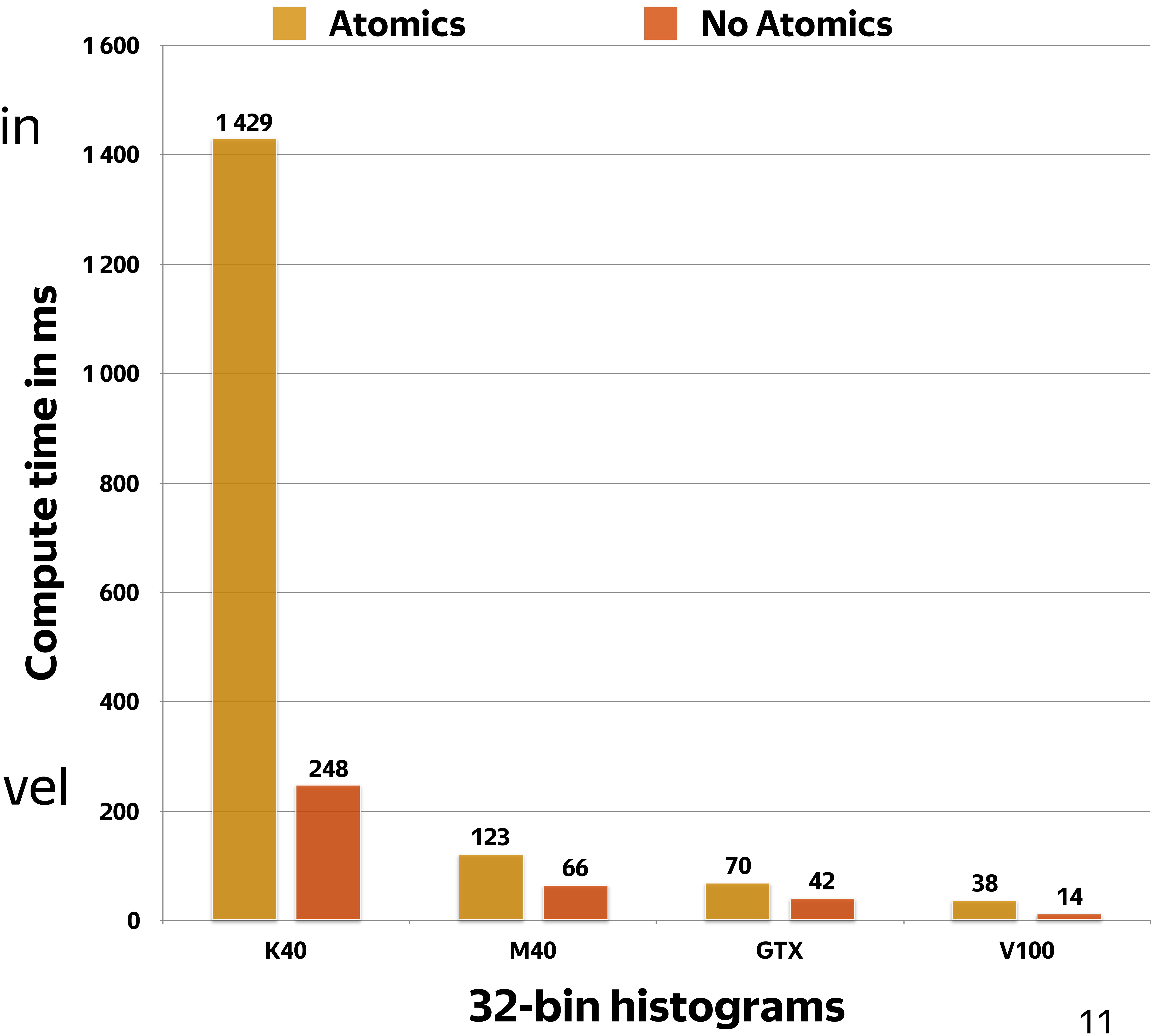
Histograms on GPU

- Aggregation in fast shared memory
- Layout to avoid bank conflicts
- No atomics: no need to hardware support
- Trade-off: Occupancy vs Atomics
  - 384 threads
  - 48KB shared memory
- CatBoost is open-source:
  - feel free to ask questions
  - or just read our code


Feature values
f_0 f_0 f_1 f_1 f_2 f_2 f_3 f_3
tid
0 1 2 3 4 5 6 7 ...
Feature values
f_0 f_0 f_1 f_1 f_2 f_2 f_3 f_3
tid
0 1 2 3 4 5 6 7 ...
Feature values
f_0 f_0 f_1 f_1 f_2 f_2 f_3 f_3
tid
0 1 2 3 4 5 6 7 ...
Feature values
f_0 f_0 f_1 f_1 f_2 f_2 f_3 f_3
tid
0 1 2 3 4 5 6 7 ...

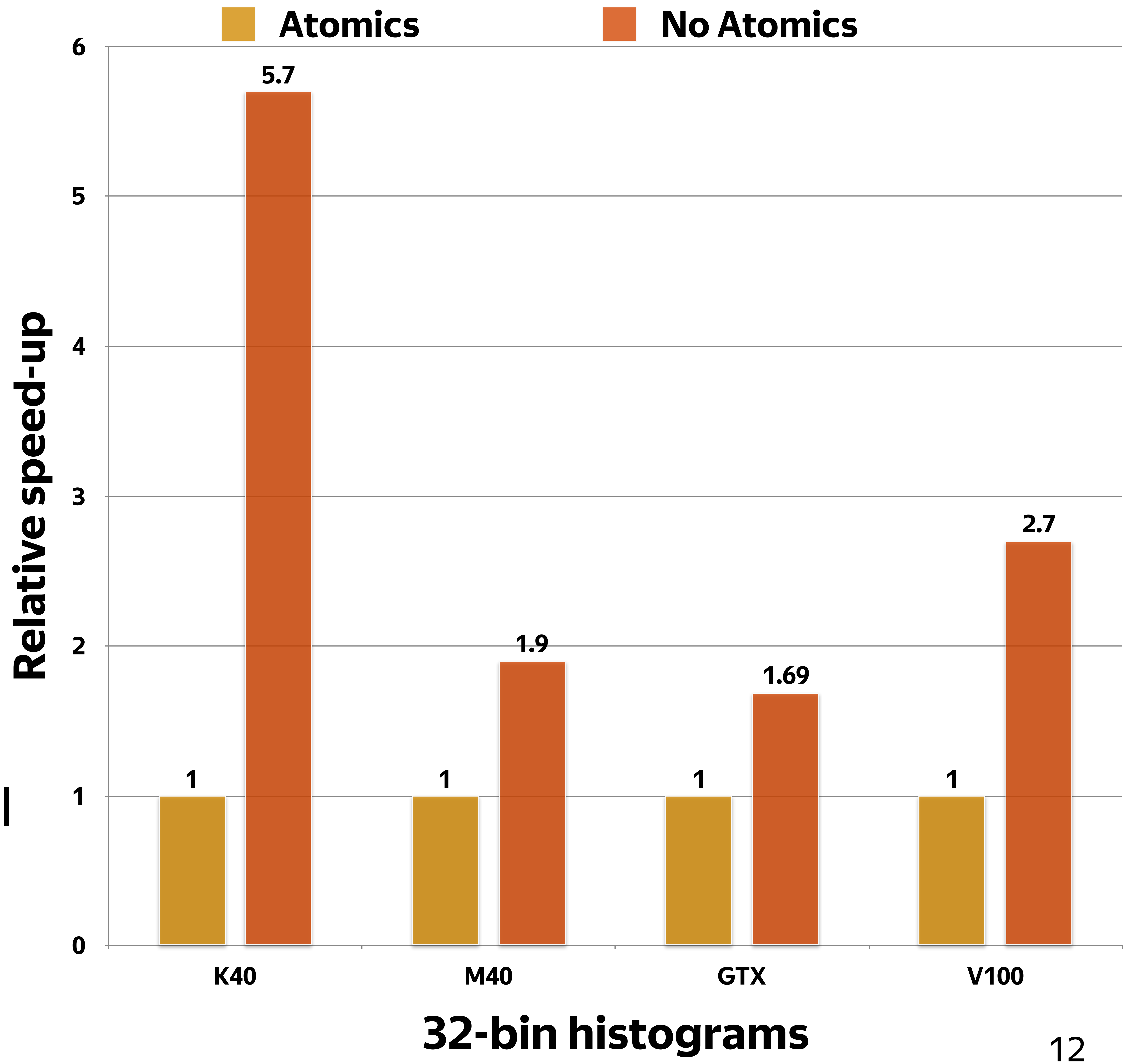
Avoid atomics

- Experiment: increase occupancy in exchange for atomic operations
  - K40: 19% => 38%
  - M40, 1080Ti, V100: 38% => 75%
- Hardware:
  - K40, M40, 1080Ti, V100
- Result (Maxwell and later):
  - x1.5-x3 performance for first level histograms
  - x1.25-x2 faster training time



Avoid atomics

- Experiment: increase occupancy in exchange for atomic operations
  - K40: 19% => 38%
  - M40, 1080Ti, V100: 38% => 75%
- Hardware:
  - K40, M40, 1080Ti, V100
- Result (Maxwell and later):
  - x1.5-x3 performance for first level histograms
  - x1.25-x2 faster training time



Benchmarks

GPU vs CPU

Hardware

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

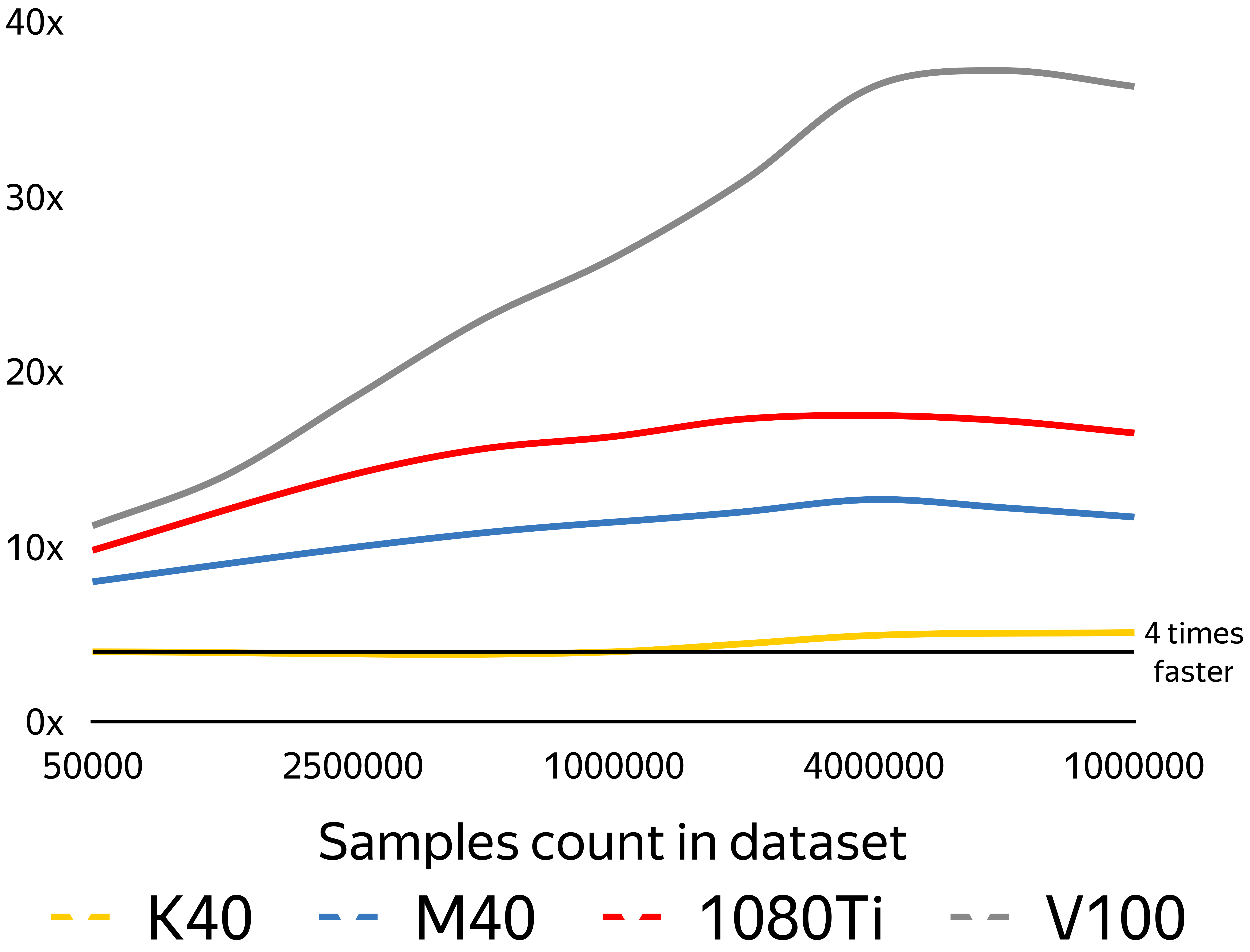
Dataset

- ≈800 features

Price:

- 2xIntel Xeon E5-2660v4: ≈3000$ (amazon.com)
- Titan V: 3000$

GPU relative speed-up for different sample count



4 times faster

Comparison with competitors

Parameters

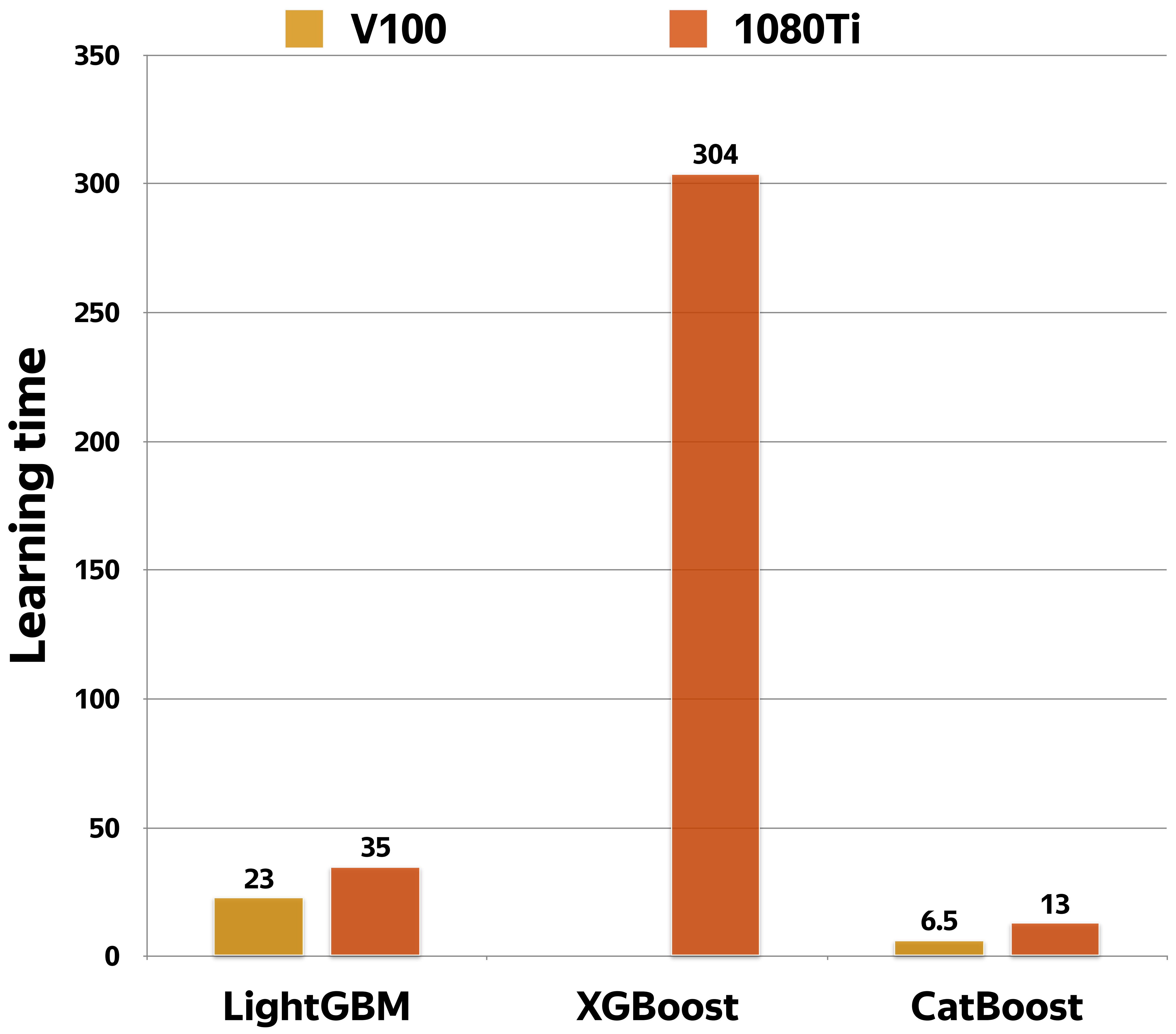
- 32 bins, 64 leaves, 200 iterations

Dataset

- ≈800 features
- 4M samples

XGBoost + V100?

- XGBoost 0.7 crashed with “Illegal Memory Access”; previous (working) revision doesn’t support Volta



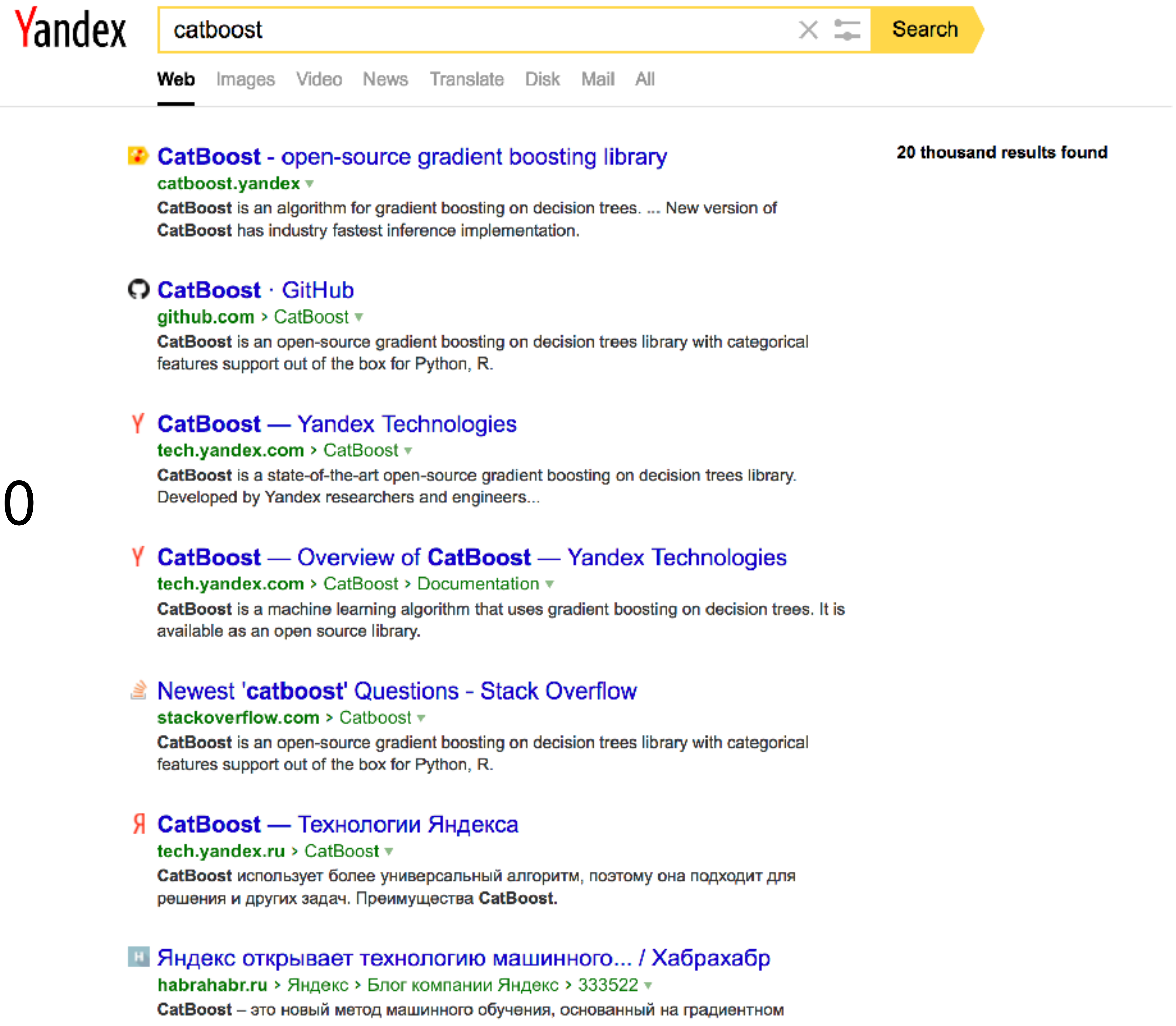
Quality?

- Categorical: state-of-the-art
- Ordered: comparable or better

See benchmarks on our GitHub

GPU Gradient boosting usage in Yandex

- Proprietary (old) version of CatBoost
- Ranking formulas:
  - CPU: 75 hours on 100 machines
  - GPU: 7-9 hours on 1 machine with 8P40
- Management:
  - More money => More data



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.

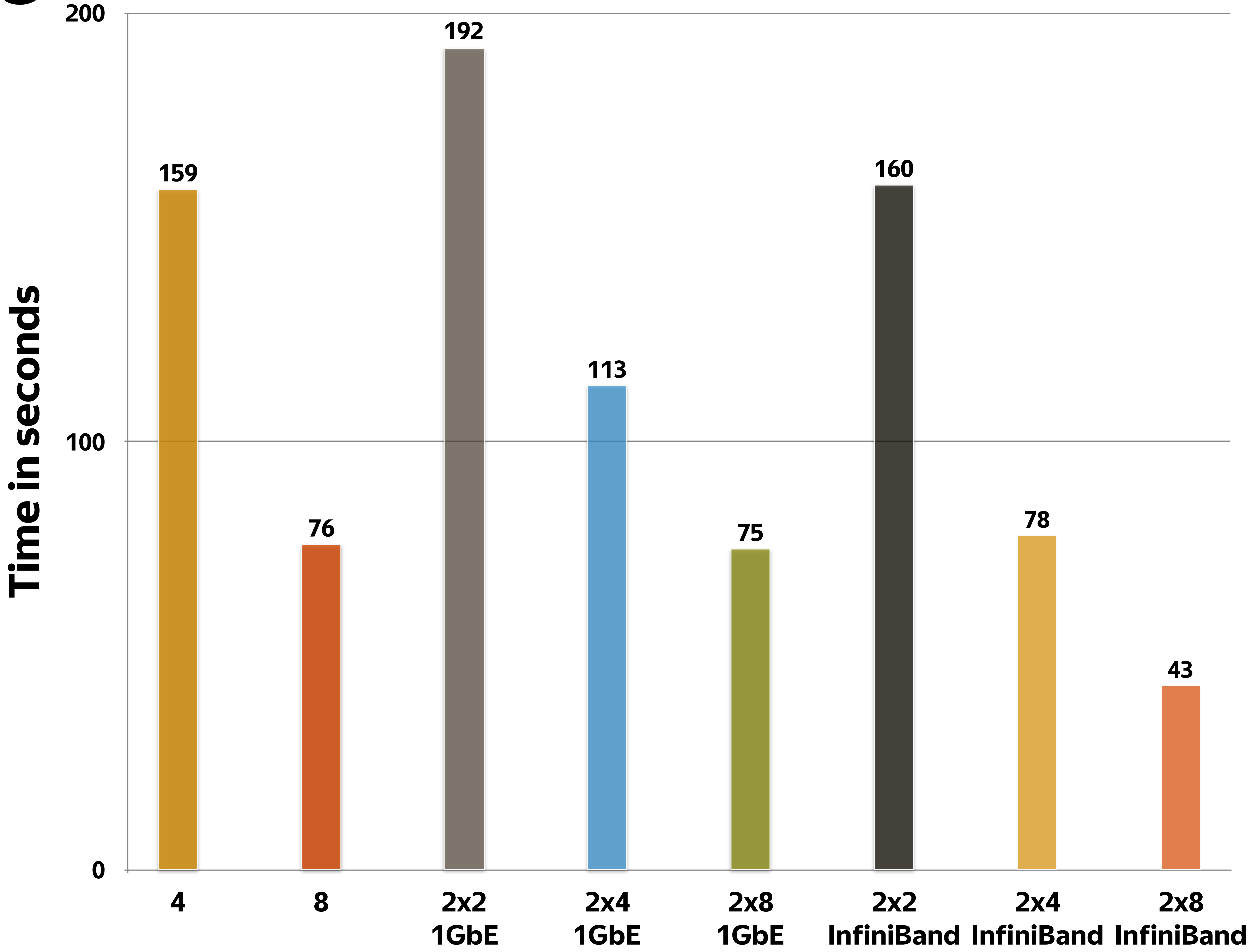
Newest '**catboost**' Questions - Stack Overflow
stackoverflow.com › Catboost
CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R.

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

Beyond one machine

- First open-source distributed GBDT on GPU
- Could be used even on 1GB/s ethernet, if you have enough data
- Learn time speed-up with fast interconnection like Mellanox InfiniBand



Thank You!

For more information:

https://catboost.yandex

Vasily Ershov
Software developer

✉ noxoomo@yandex-team.ru

📱 +7 921 332 45 71