Prepare for the Gradient Boosting tutorial

- Download the notebook:

https://drive.google.com/file/d/1iTM_TixmwWREdTB830BASap0P7aCLrX1/view
OR http://bit.ly/2GXIysG OR
git clone https://github.com/catboost/tutorials

cd events/2019_odsc_east

- Install the libraries:

pip install catboost shap ipywidgets sklearn

jupyter nbextension enable --py widgetsnbextension

OPEN
DATA
SCIENCE
CONFERENCE

#ODSC

@ODSC

Boston | April 30 - May 4, 2019

#ODSC

BOSTON
APR 30 – MAY 3

Mastering Gradient Boosting with CatBoost

Anna Veronika Dorogush
Head of CatBoost Team,
Yandex

CatBoost

Gradient Boosting Library

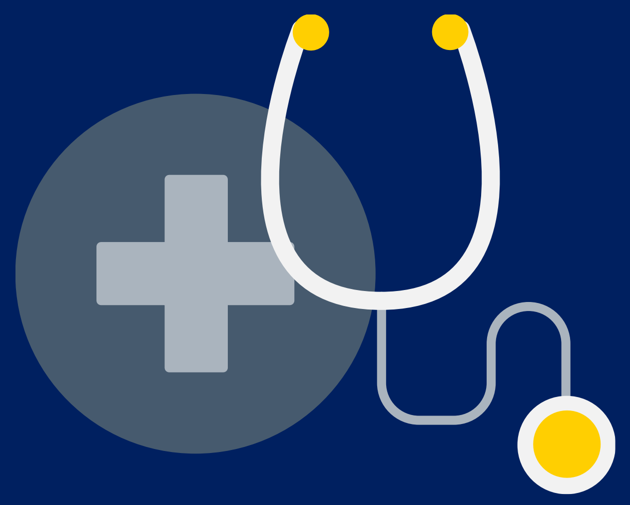
Plan

- Intro to Gradient Boosting
- Intro to CatBoost and benchmarks
- Tutorial
- Next releases

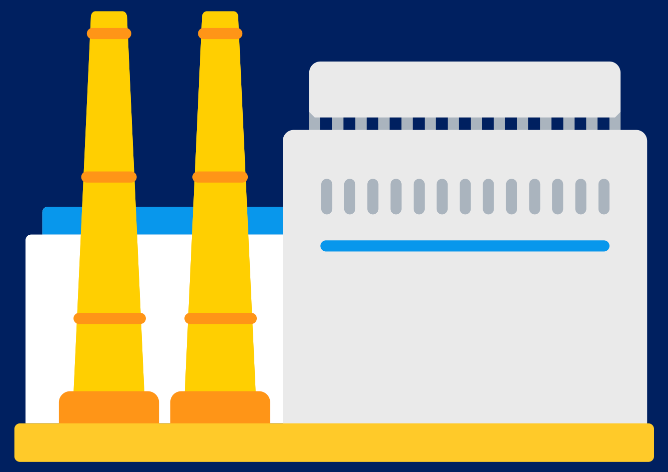
Gradient Boosting

- Best solution for heterogeneous data
- Easy to use
- Works well for small data

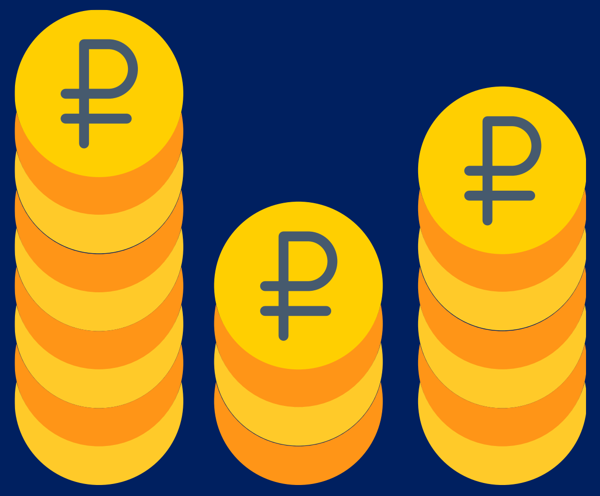
Applications



Medicine



Industry

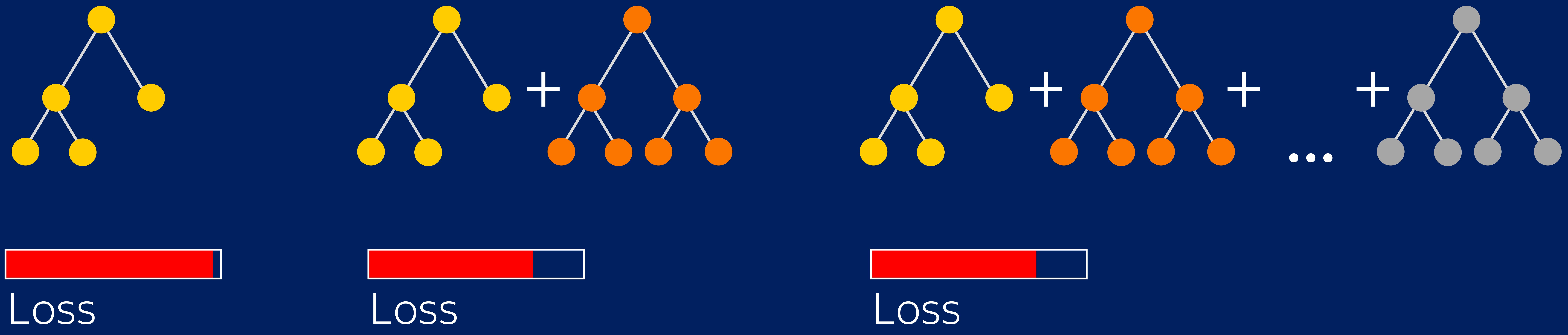


Finance

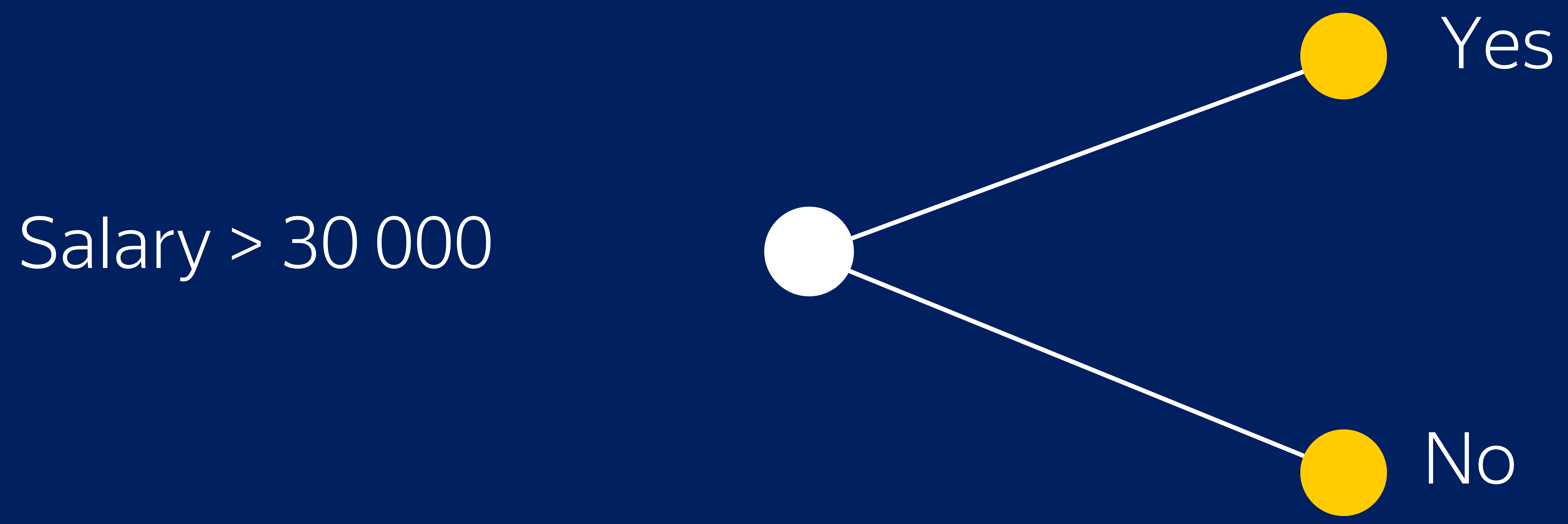
Music and video recommendations

Sales prediction

Gradient boosting

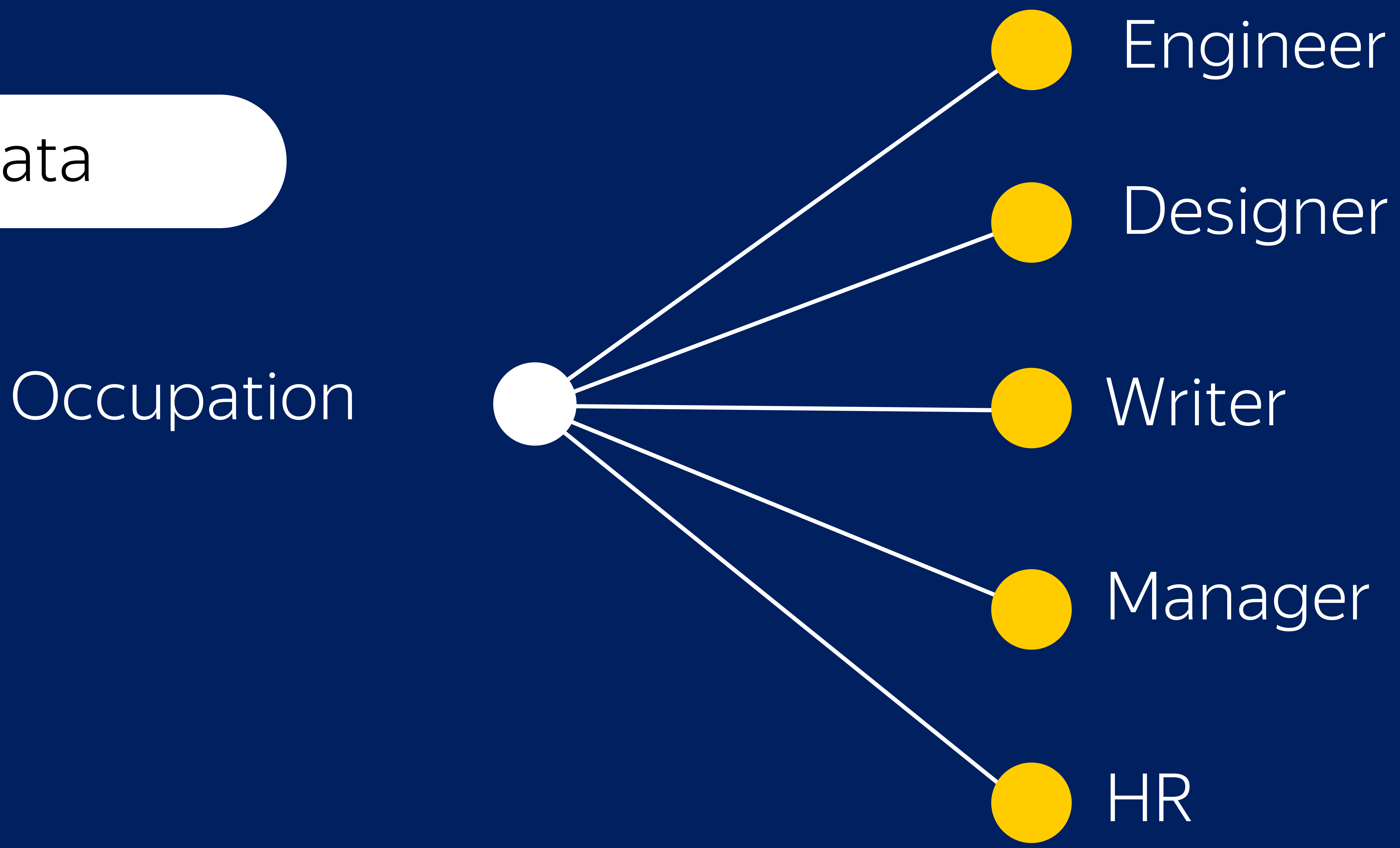


Numerical features



Categorical features

Categorical data



CatBoost advantages

- Good quality with default parameters
- Sophisticated categorical features support
- Model analysis tools

Algorithm comparison

	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%

Logloss

Speed

- Training on CPU
- Training on GPU
- Prediction speed

CPU: Comparison with other libraries

- Parameters:

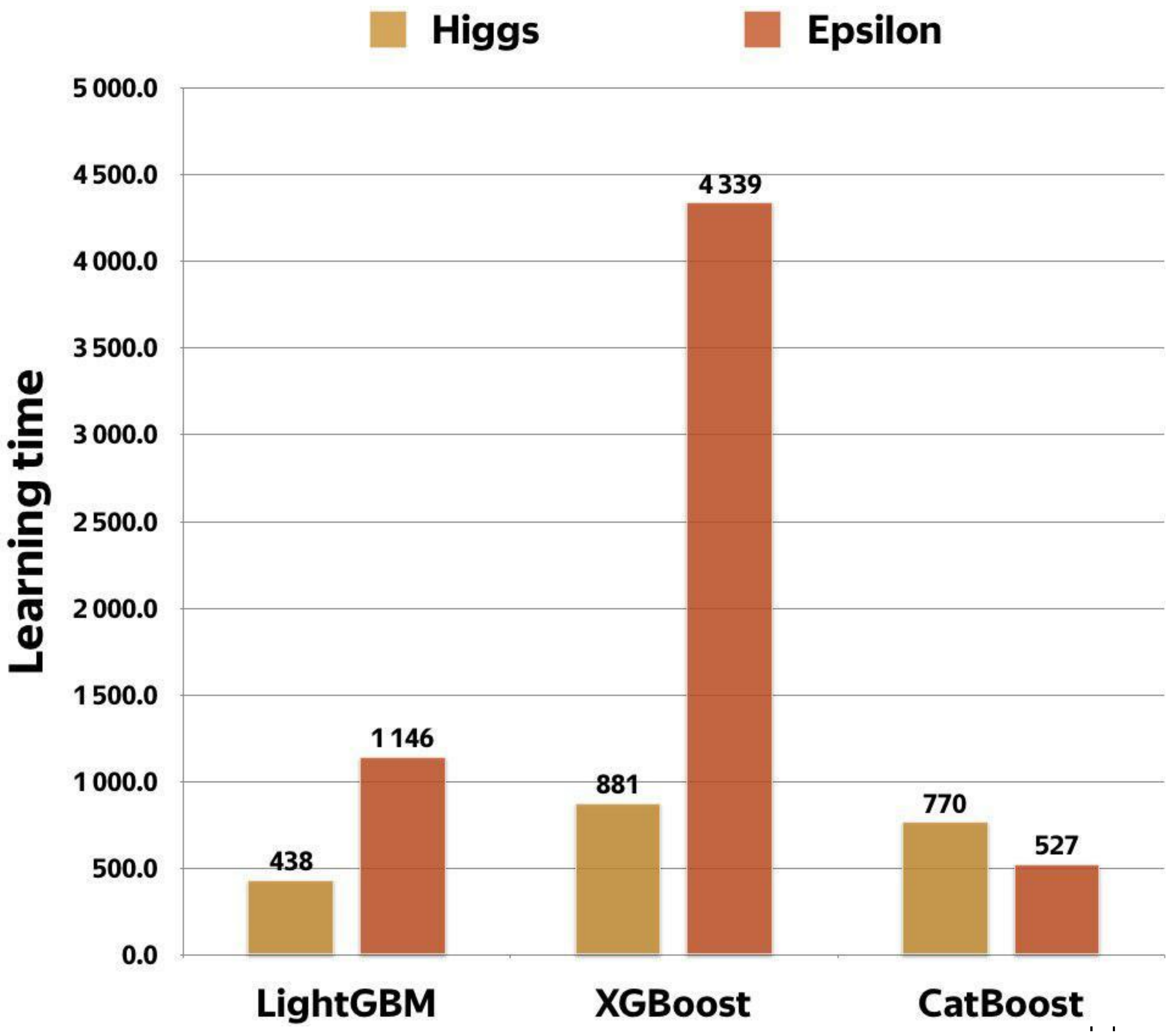
128 bins, 64 leafs, 400 iterations

- Higgs:

800 features, 4M samples

- Epsilon:

2000 features, 400K samples



GPU: Comparison with other libraries

- Parameters:

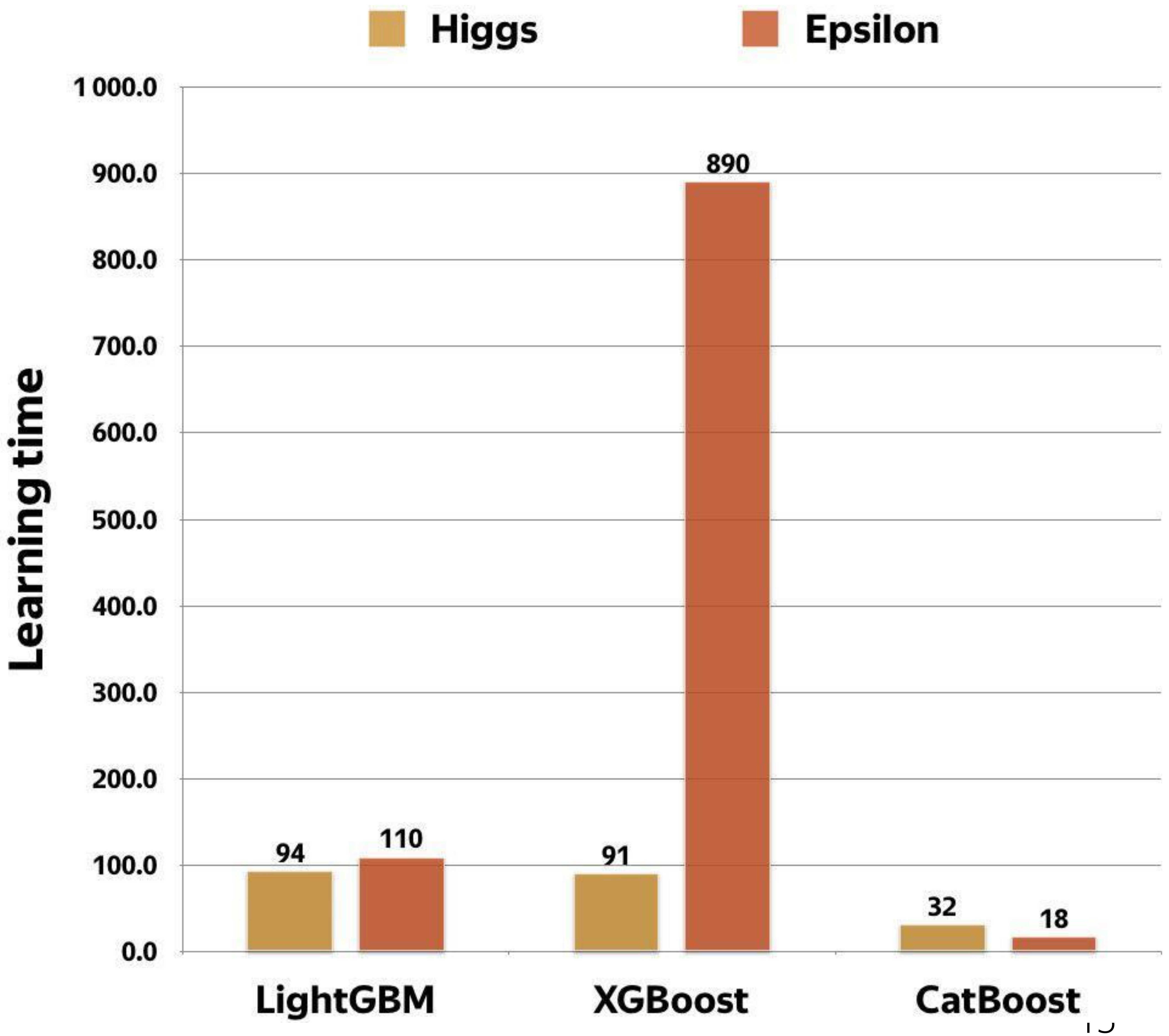
128 bins, 64 leafs, 400 iterations

- Higgs:

800 features, 4M samples

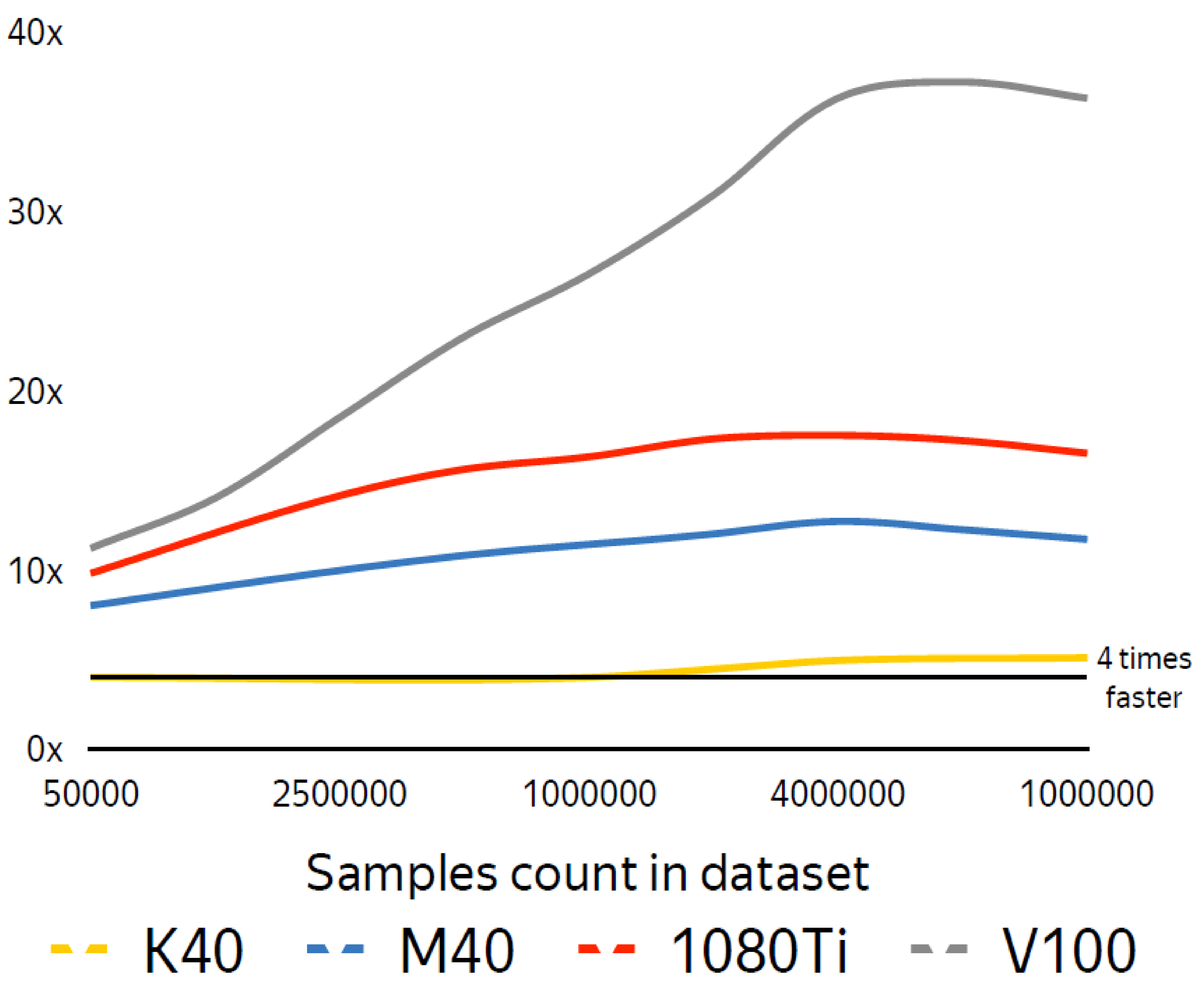
- Epsilon:

2000 features, 400K samples

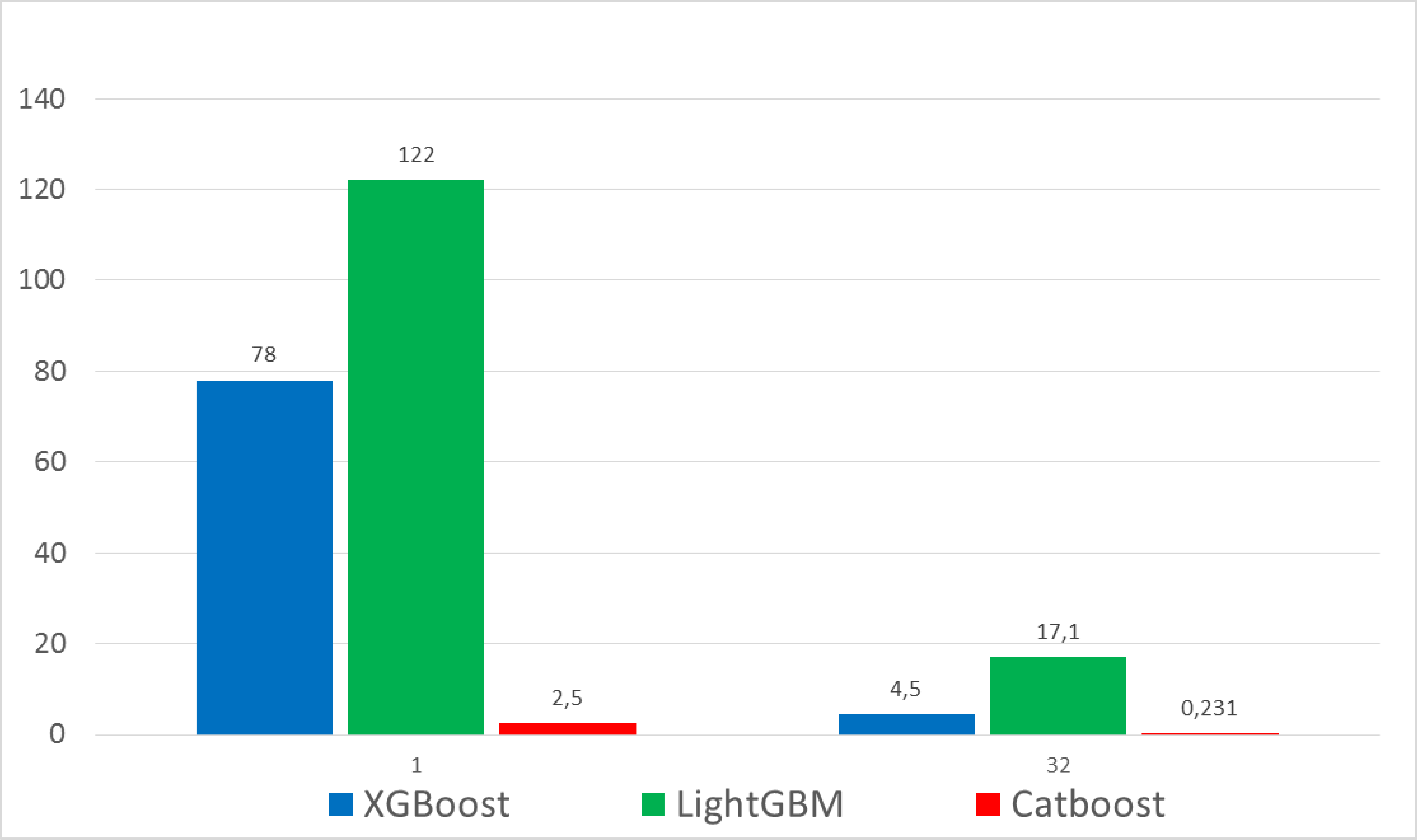


CPU vs GPU

- Dual-Socket Intel Xeon E5-2660v4 as baseline
- Several modern GPU as competitors
- Dataset: 800 features



Prediction time



Tutorial data

- Download the notebook:

http://bit.ly/2GXIysG

- Install the libraries:

pip install catboost shap ipywidgets sklearn

jupyter nbextension enable --py widgetsnbextension

Coming soon

- Sparse data support
- New types of features
- New methods for model and analysis
- More metrics
- Training speedups
- Applying CatBoost in new programming languages

- catboost.ai
- github.com/catboost
- twitter.com/CatBoostML
- t.me/catboost_en, t.me/catboost_ru
- ods.ai => slack (30k people community) => tool_catboost chanel
- forms.yandex.ru/surveys/10011699

Questions?

Anna Veronika Dorogush

Head of CatBoost team