

Prepare for the Gradient Boosting tutorial

- › Download the notebook:

<http://bit.ly/2XXbZo5> OR <http://bit.ly/2NOmLsU>

- › Install the libraries:

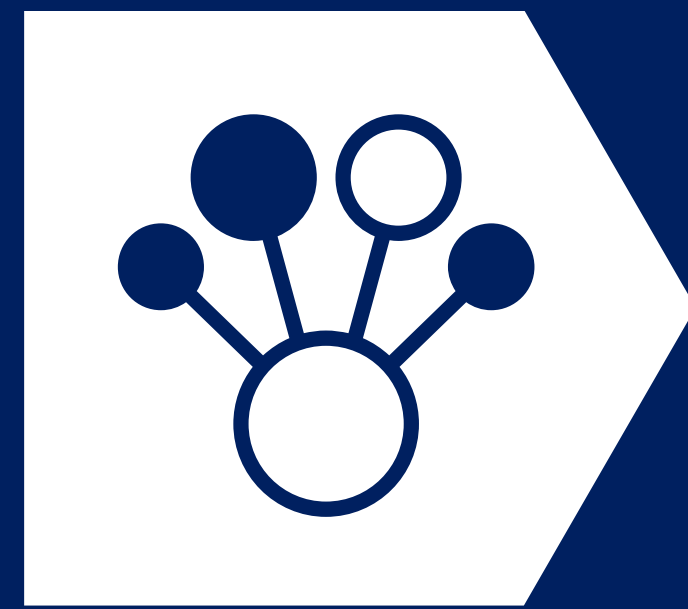
```
pip install catboost shap ipywidgets sklearn
```

```
jupyter nbextension enable --py widgetsnbextension
```

- › Download the slides:

https://github.com/catboost/catboost/tree/master/slides/2019_PyData_London

Mastering Gradient Boosting with CatBoost



CatBoost

Gradient Boosting Library

Plan

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

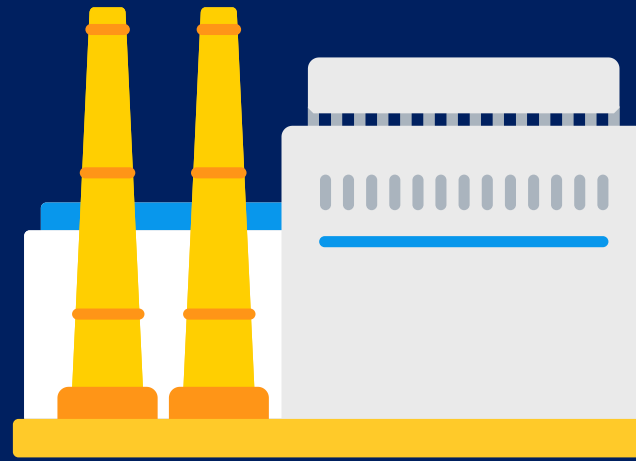
Gradient Boosting

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

Applications



Medicine



Industry



Finance

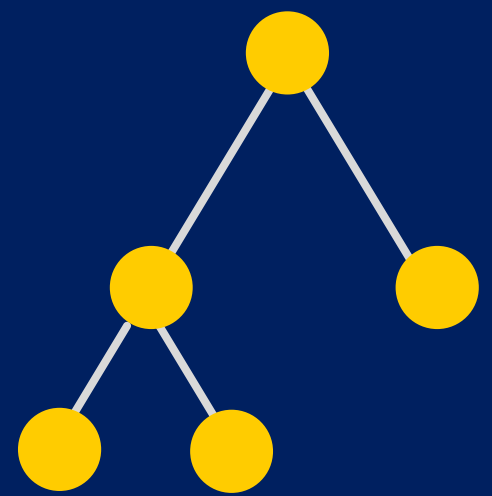


Music and video
recommendations

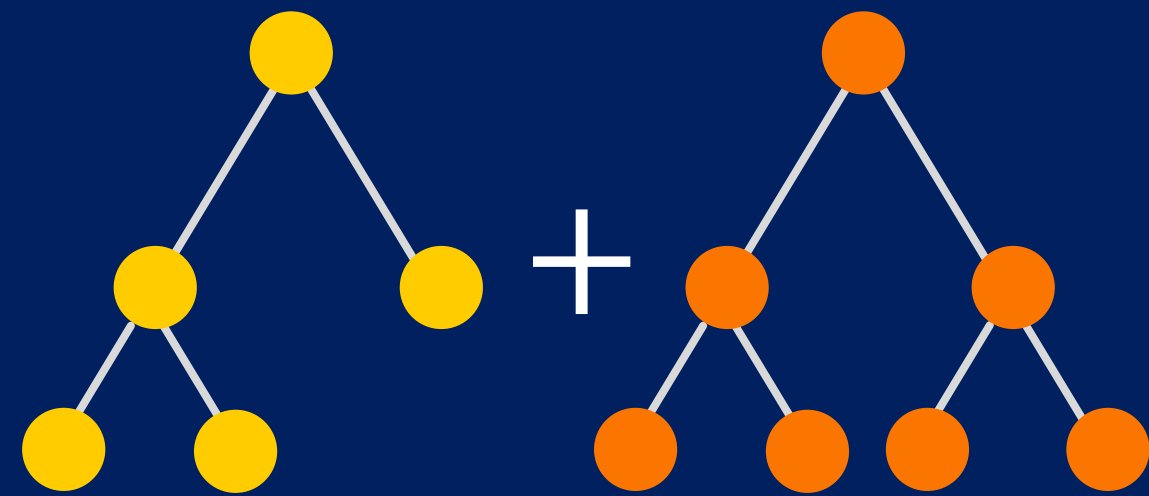


Sales prediction

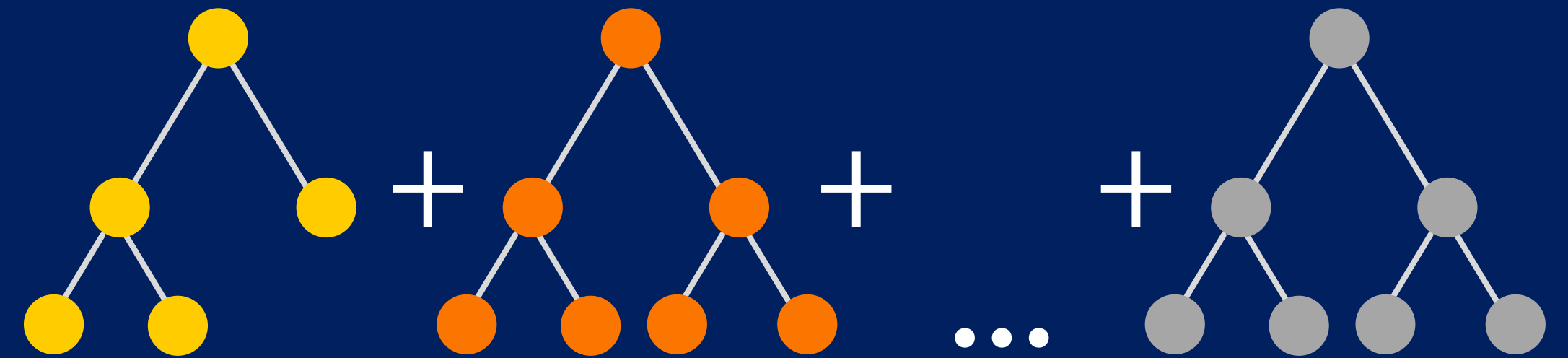
Gradient boosting



Loss

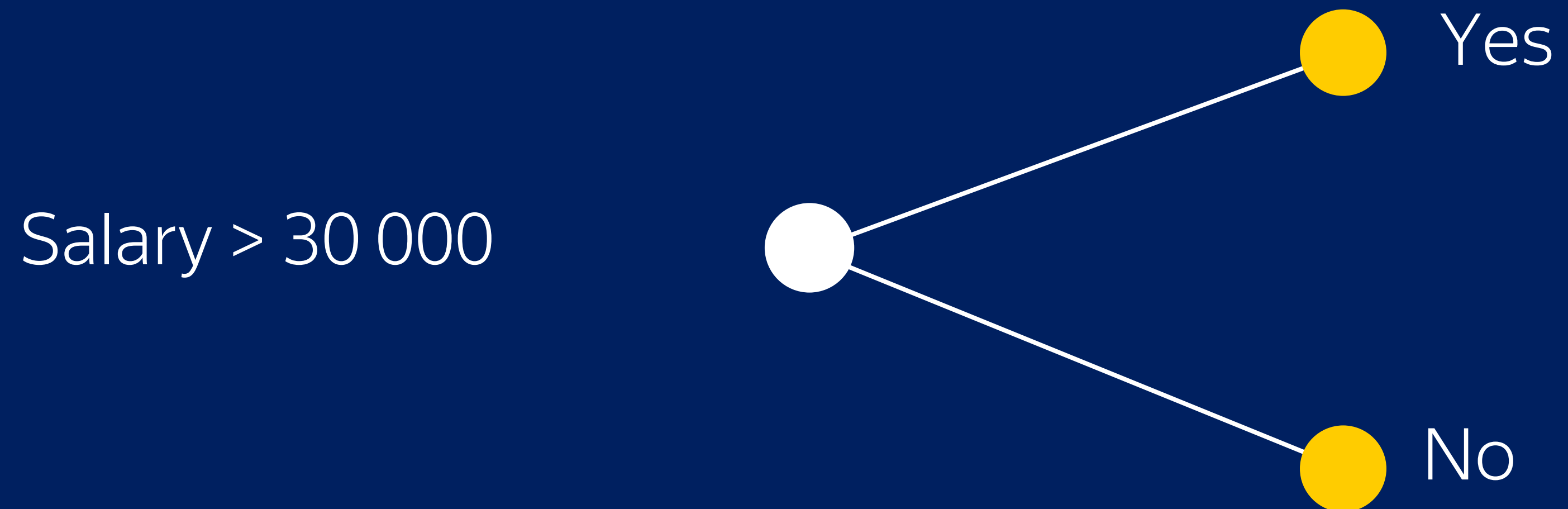


Loss



Loss

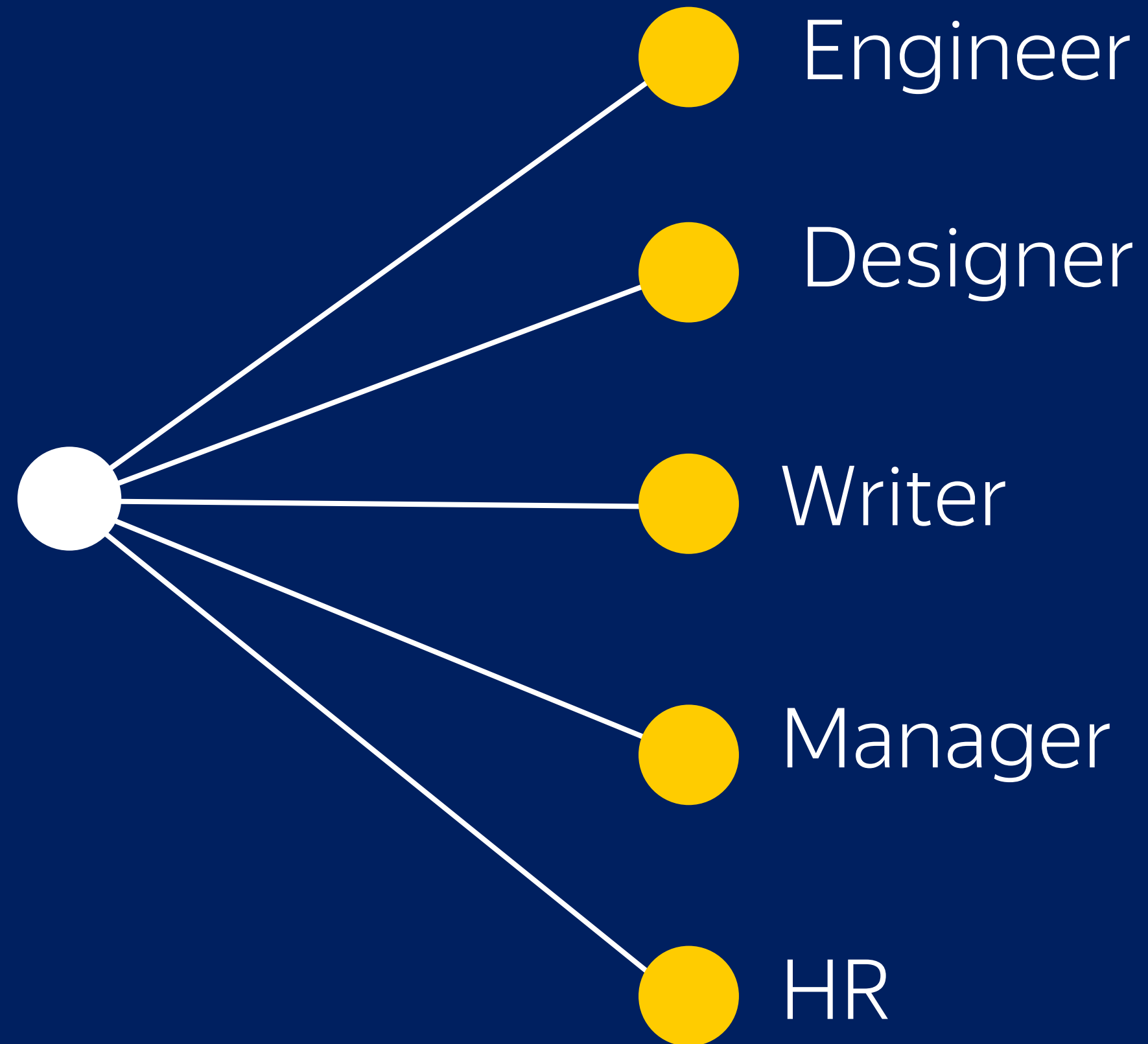
Numerical features



Categorical features

Categorical data

Occupation



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%

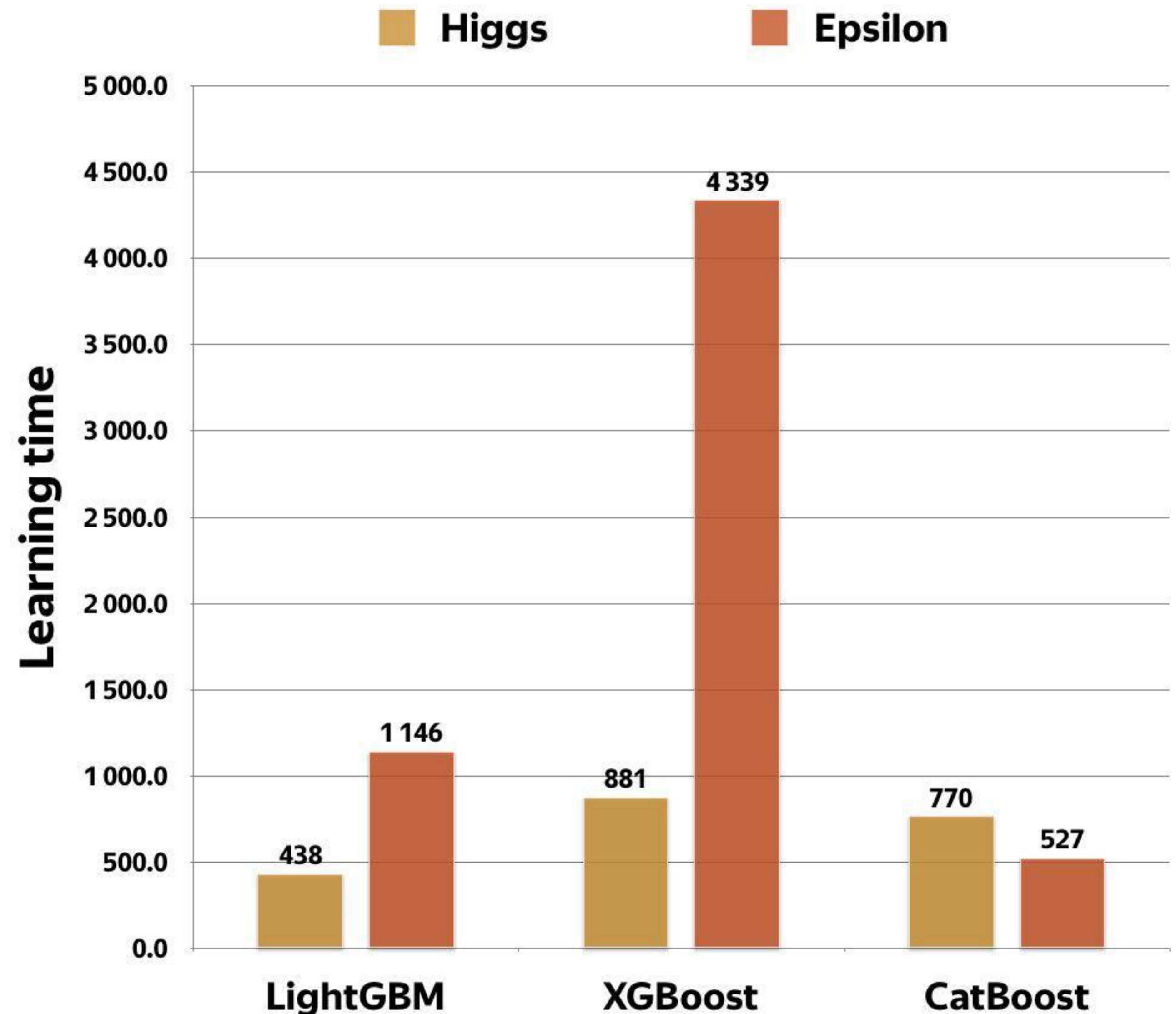
Logloss

Speed

- › Training on CPU
- › Training on GPU
- › Prediction speed

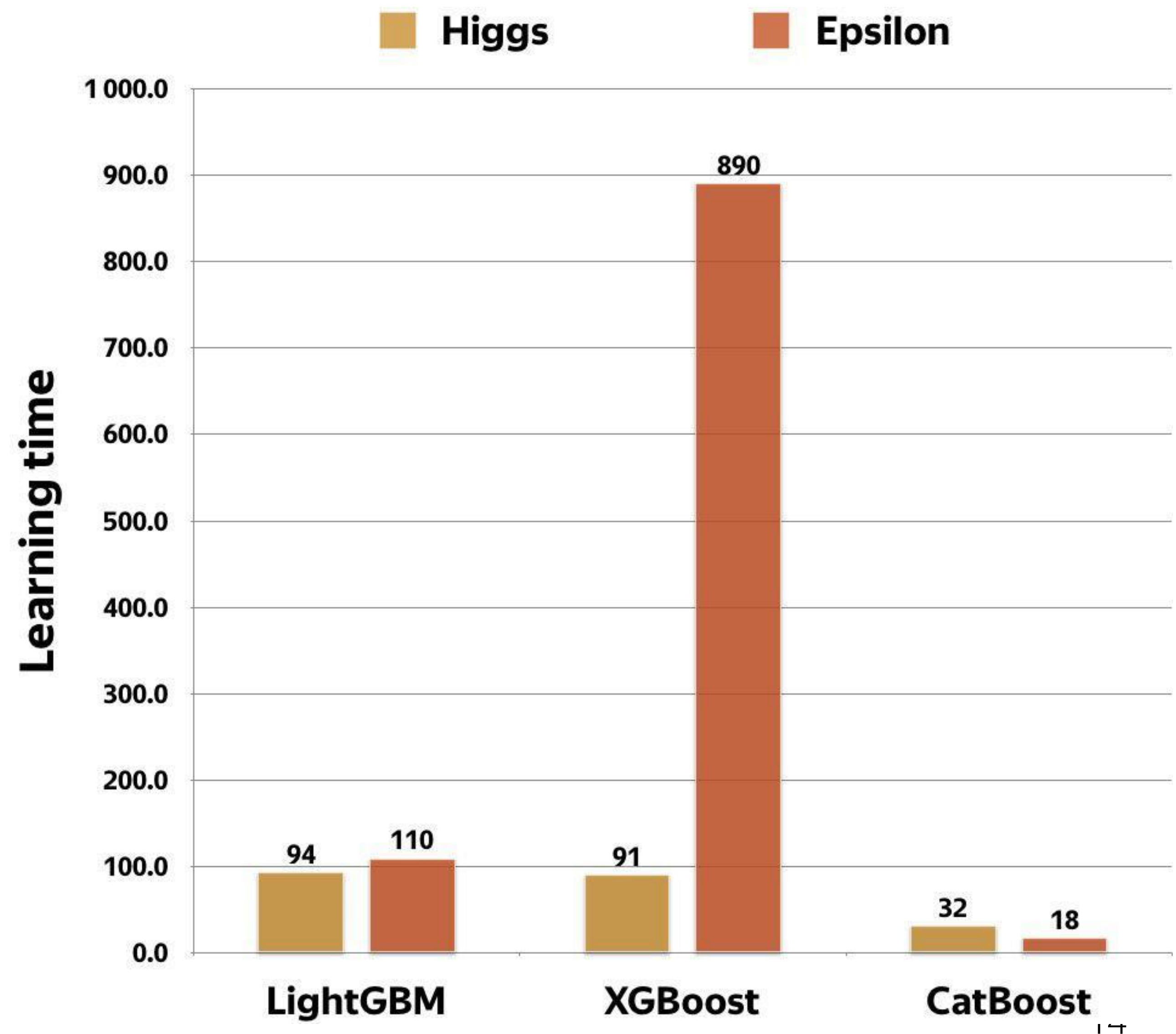
CPU: Comparison with other libraries

- Parameters:
128 bins, 64 leafs, 400 iterations
- Higgs:
800 features, 4M samples
- Epsilon:
2000 features, 400K samples



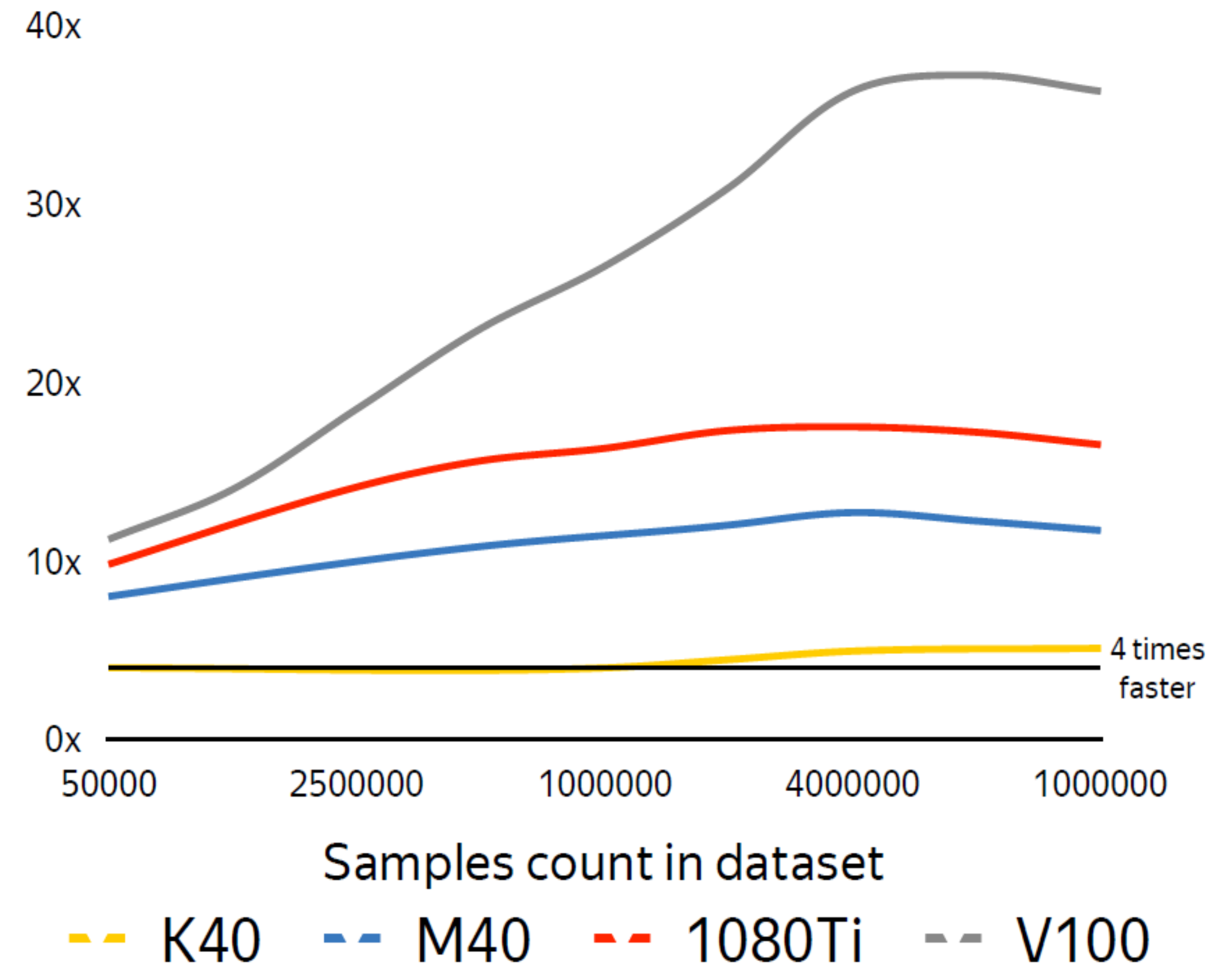
GPU: Comparison with other libraries

- Parameters:
128 bins, 64 leafs, 400 iterations
- Higgs:
800 features, 4M samples
- Epsilon:
2000 features, 400K samples

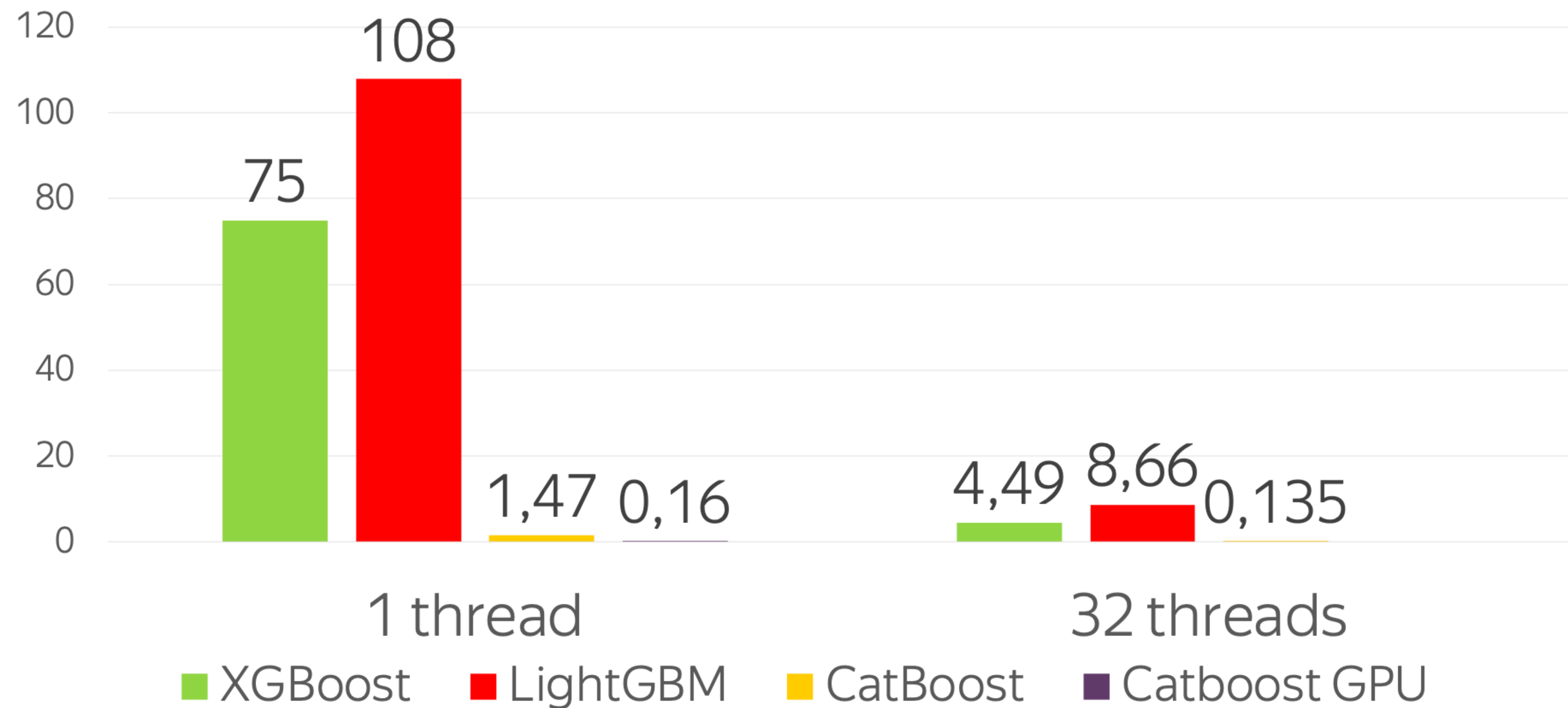


CPU vs GPU

- 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/2XXbZo5> OR <http://bit.ly/2NOmLsU>

- › 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

- catboost.ai
- twitter.com/CatBoostML
- Telegram:

t.me/catboost_en

- Slack:

[ods.ai => slack \(30k people community\) => #tool_catboost](#)

- Slides:

https://github.com/catboost/catboost/tree/master/slides/2019_PyData_London

Questions?

Anna Veronika Dorogush

Head of CatBoost team