Карты, бустинг 2 стула

Алексей Натекин

Open Data Science


dmlc
XGBoost
2016

GBM план

1. Че кого
2. Че, есть чо
3. Че как
4. Ложная дилемма

GBM план

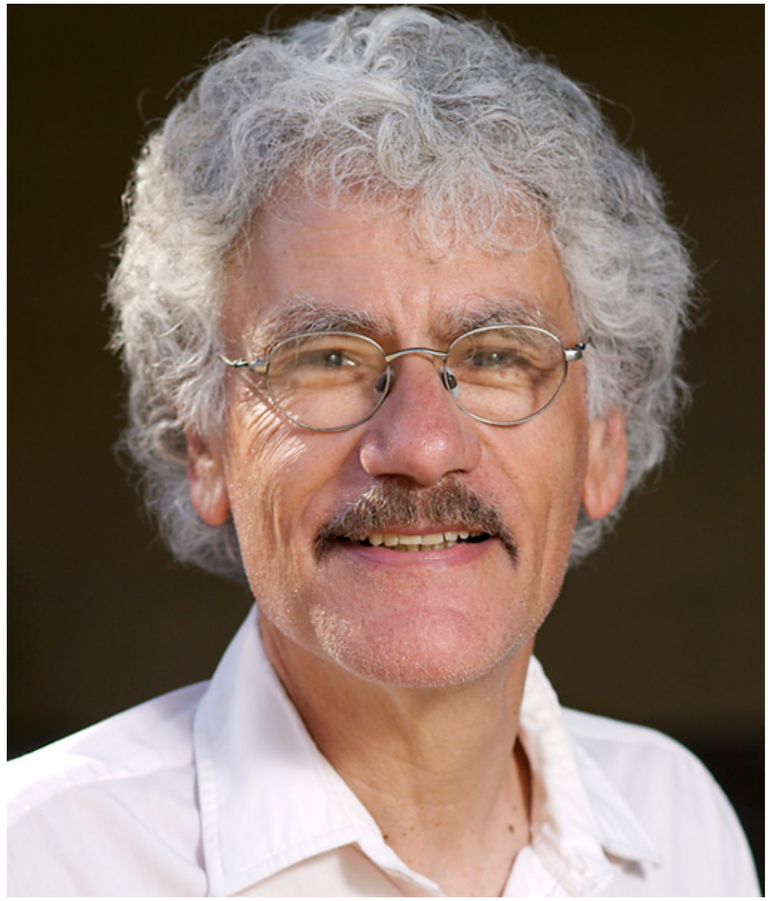
1. Че кого

2. Че, есть чо

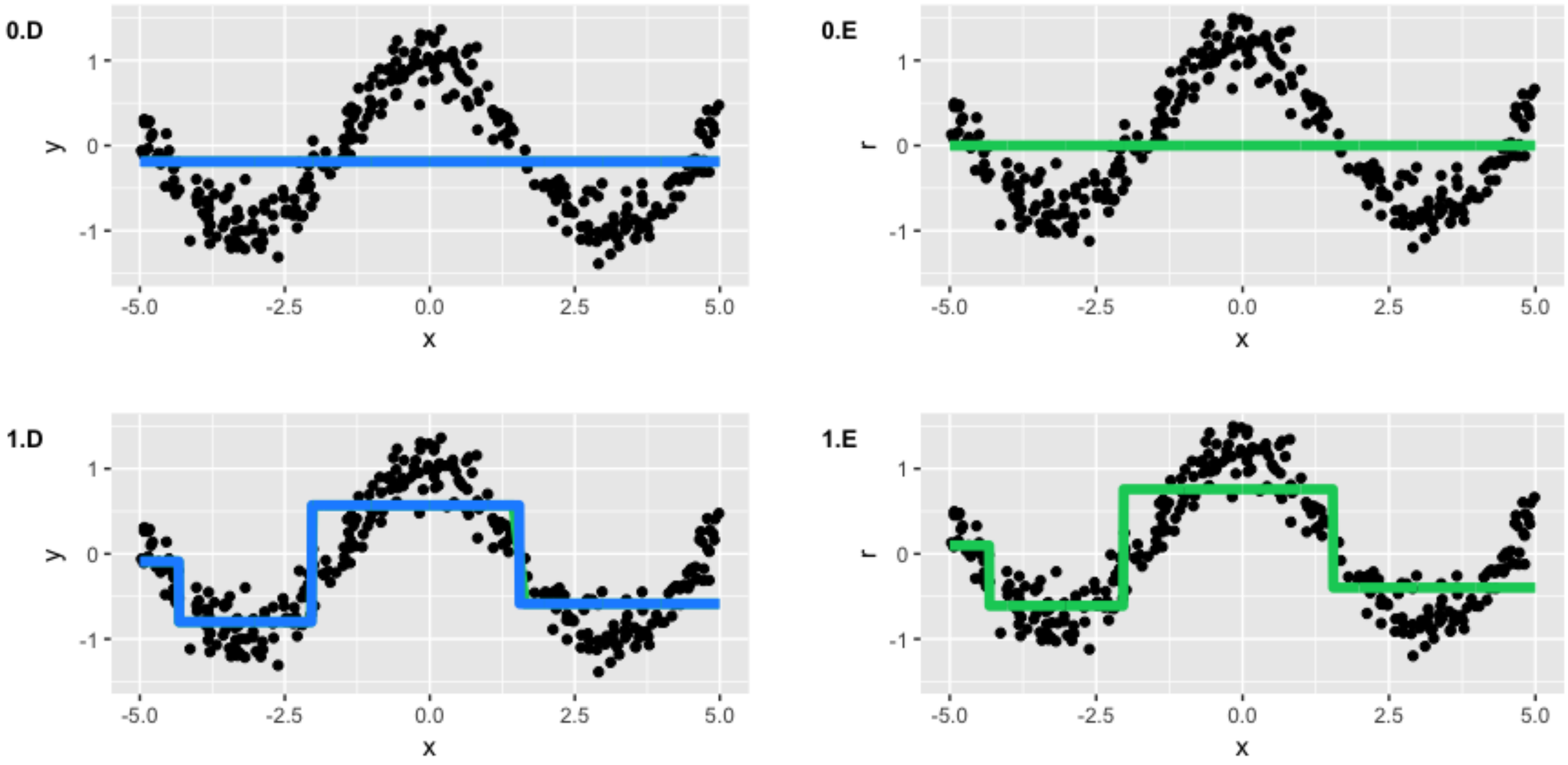
3. Че как

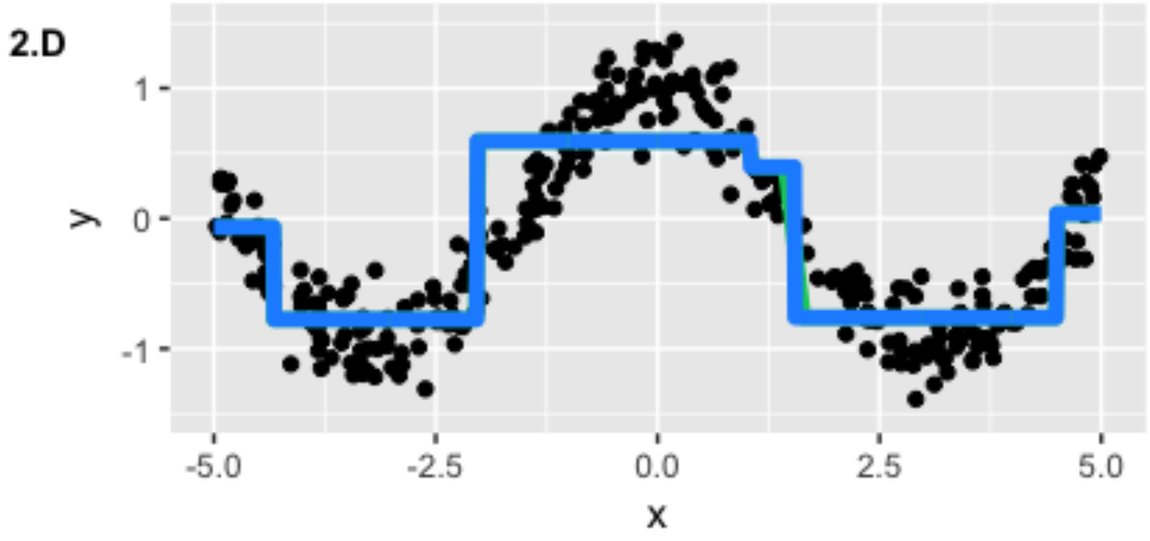
4. Ложная дилемма

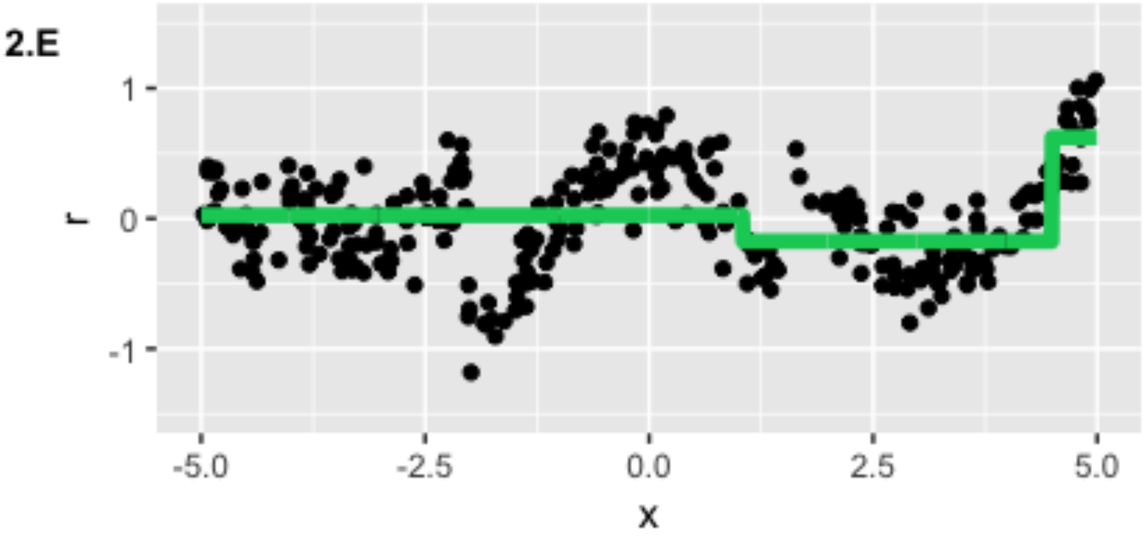
ДЕДЫ

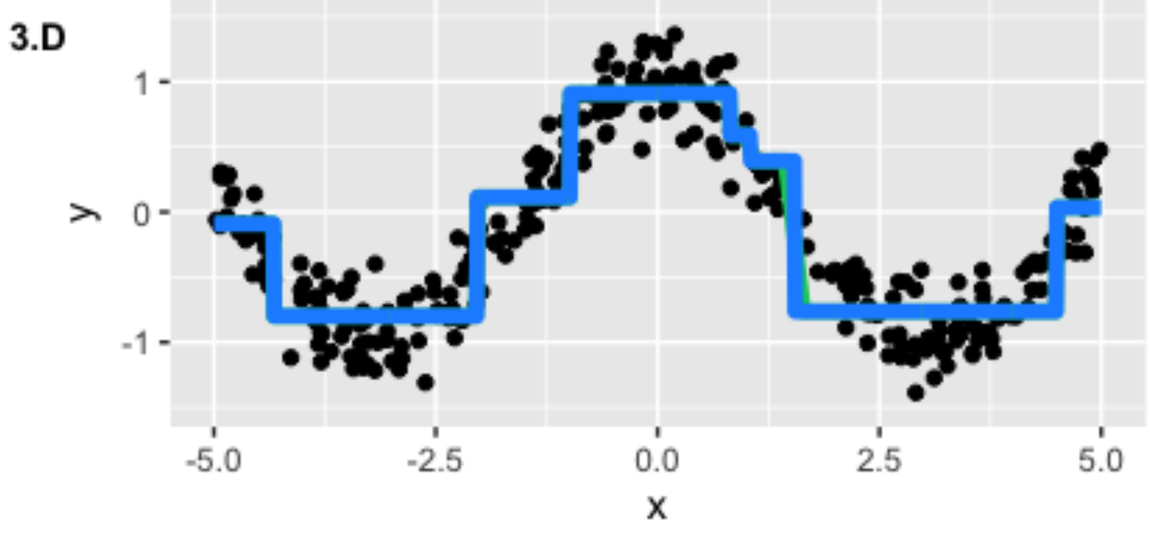


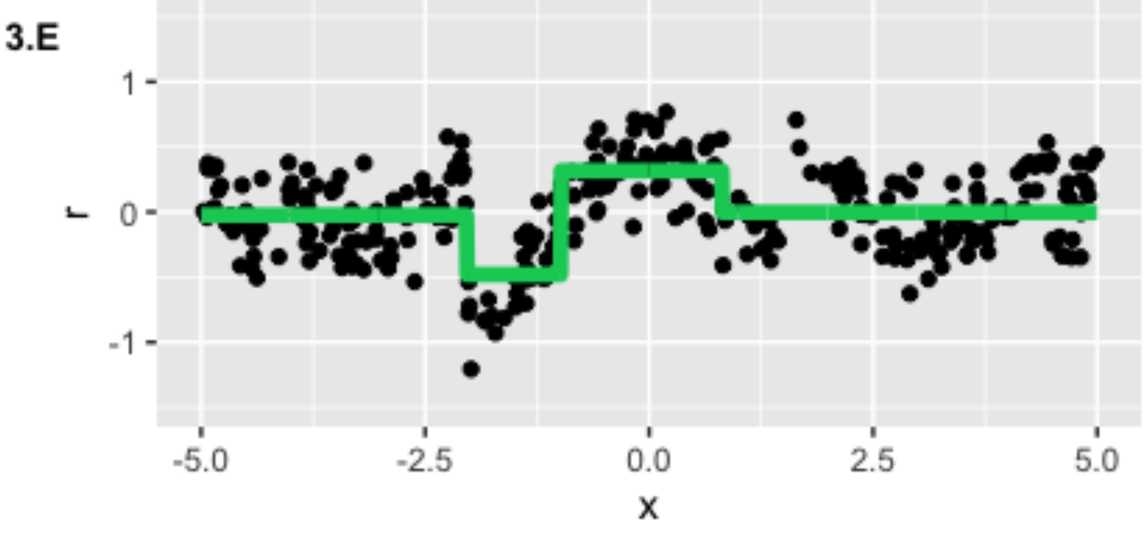
ОБУЧАЛИ

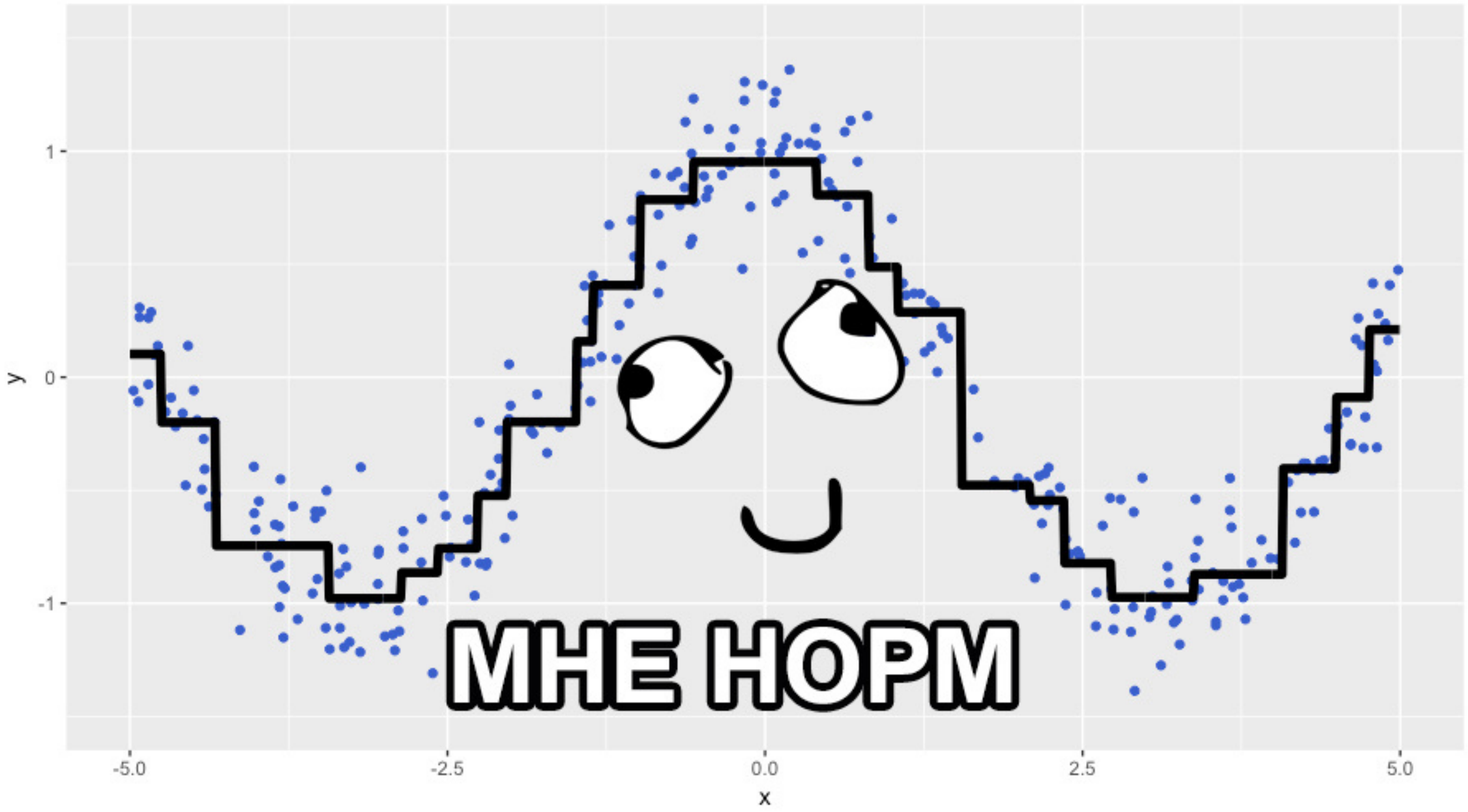

0.D
1
0
-1
y
-5.0
-2.5
0.0
2.5
5.0
x
0.E
1
0
-1
r
-5.0
-2.5
0.0
2.5
5.0
x
1.D
1
0
-1
y
-5.0
-2.5
0.0
2.5
5.0
x
1.E
1
0
-1
r
-5.0
-2.5
0.0
2.5
5.0
x


2.D
1
0
-1
y
-5.0
-2.5
0.0
2.5
5.0
x


2.E
1
0
-1
r
-5.0
-2.5
0.0
2.5
5.0
x


3.D
1
0
-1
y
-5.0
-2.5
0.0
2.5
5.0
x


3.E
1
0
-1
r
-5.0
-2.5
0.0
2.5
5.0
x

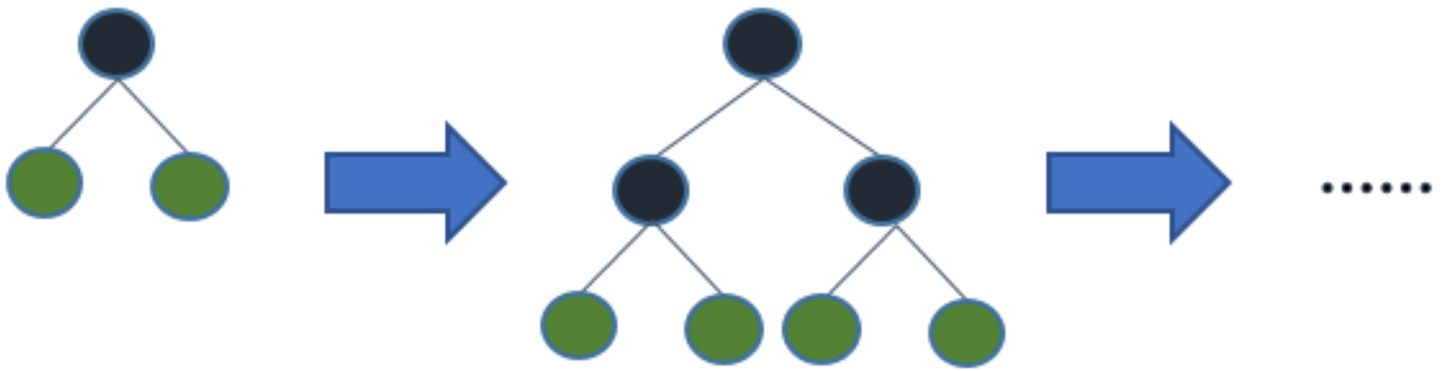

1
0
-1
y
-5.0
-2.5
0.0
2.5
5.0
x
MHE HOPM

0. Строим начальное приближение

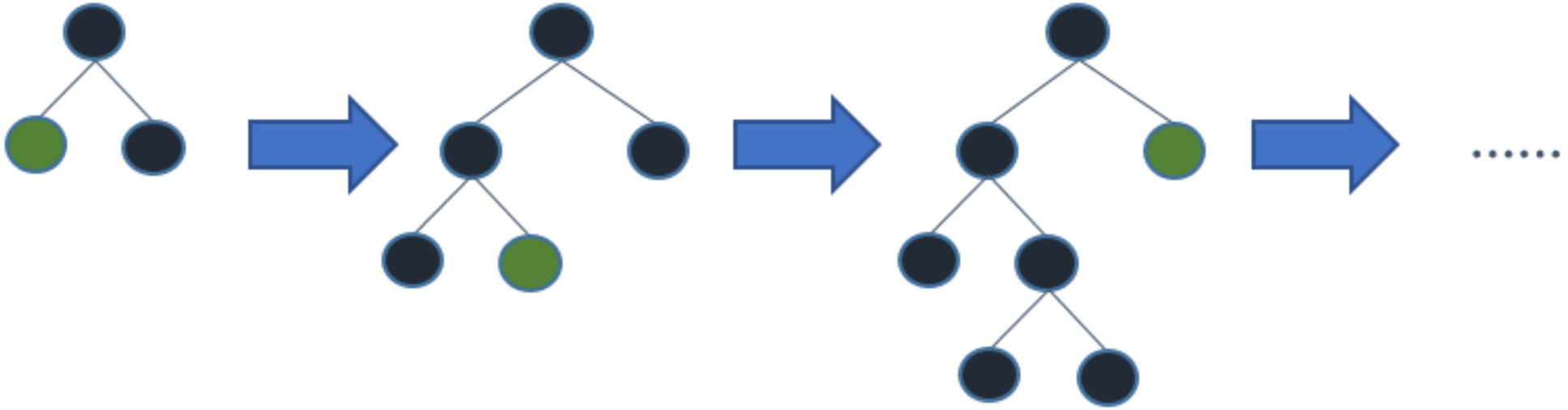
1. Считаем текущий Loss
2. Строим очередное дерево **параллелится**
3. Подбираем дереву всякие коэффициенты
4. Добавляем дерево к ансамблю

...

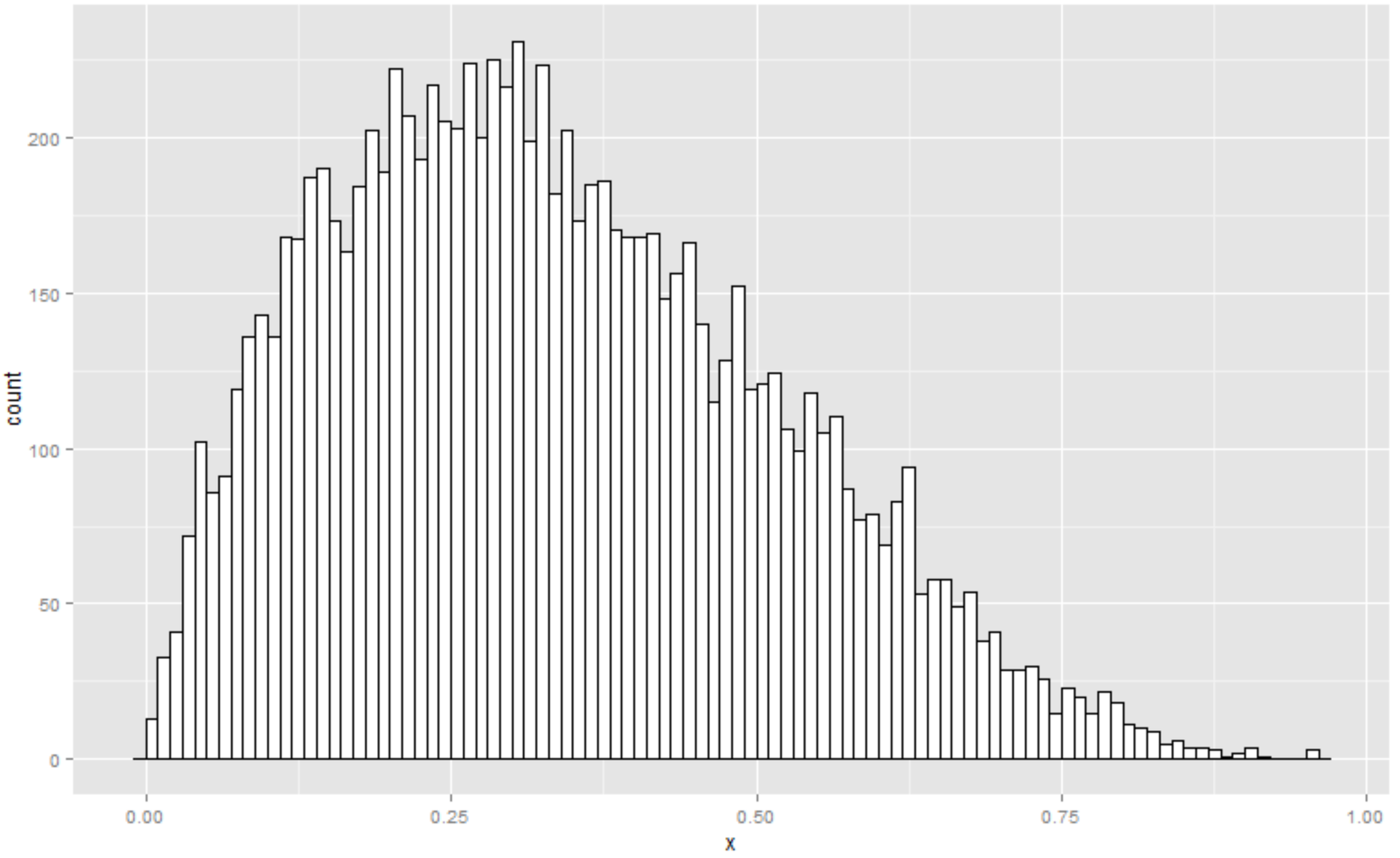
Profit



Level-wise tree growth



Leaf-wise tree growth


count
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
x

GBM план

1. Че кого
2. Че, есть чо
3. Че как
4. Ложная дилемма

Such boosting

much learning

wow

scikit learn

Spark MLlib

R CRAN

H2O.ai

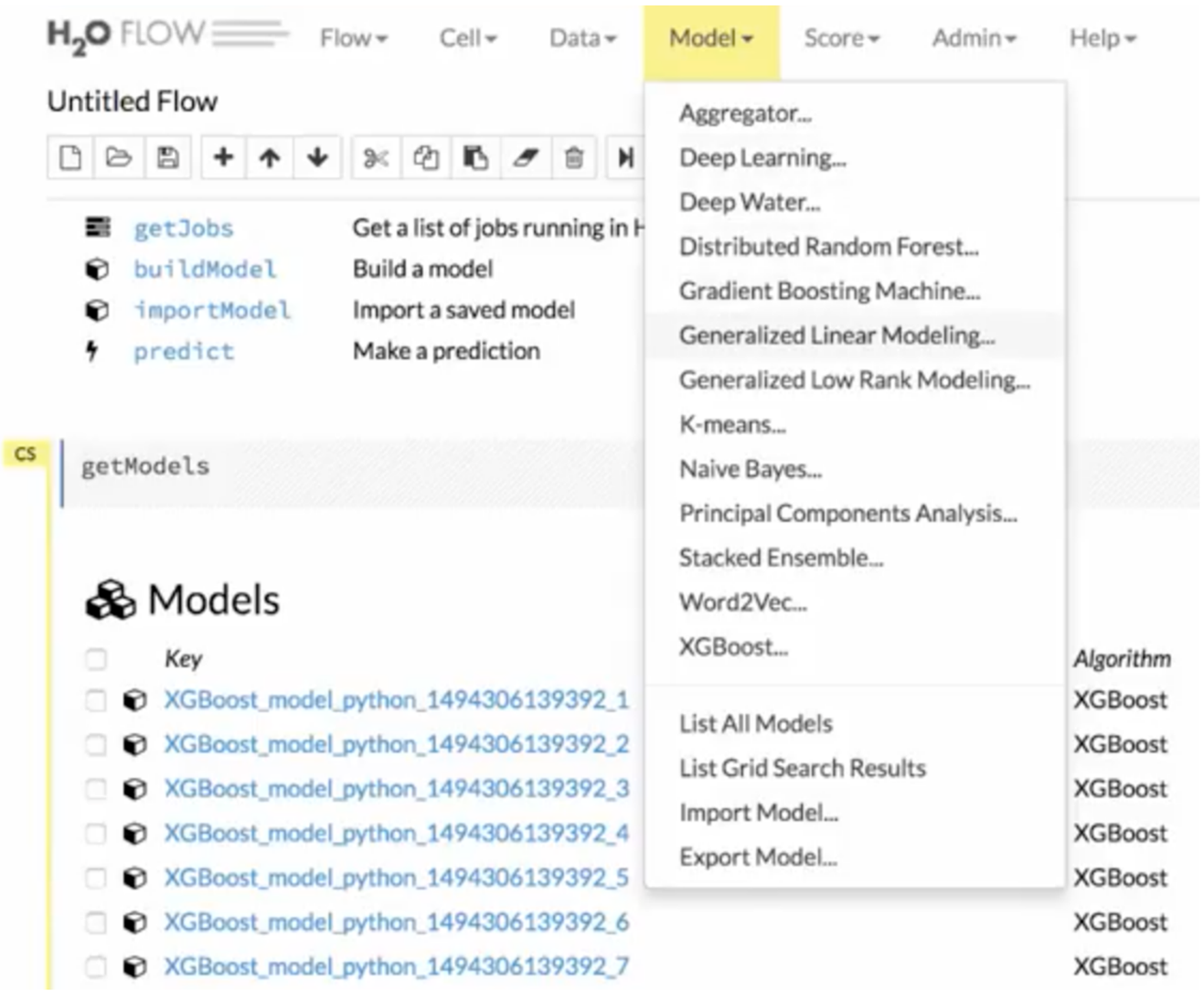
dmlc XGBoost

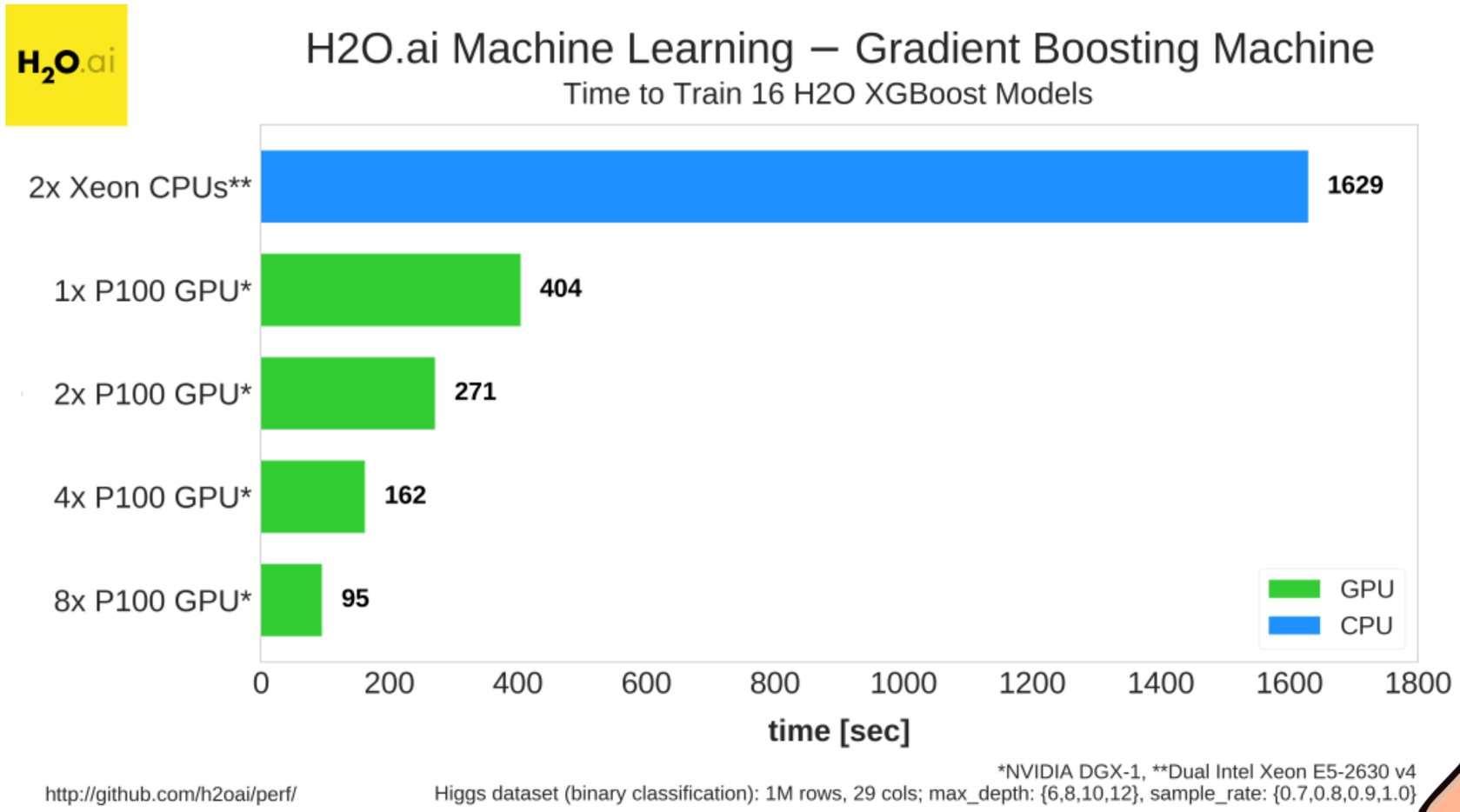
Microsoft LightGBM

Yandex CatBoost

H2O.ai	dmlc XGBoost	Microsoft LightGBM	Yandex CatBoost
Hist	Hist	Hist	Hist?
Level-wise	Level-wise	Leaf-wise	~Level-wise

H2O.ai	dmlc XGBoost	Microsoft LightGBM	Yandex CatBoost
Hist	Hist	Hist	Hist?
Level-wise	Level-wise	Leaf-wise	~Level-wise
GPU promised	Single GPU	Single GPU	Single GPU


H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
getJobs
Get a list of jobs running in H
buildModel
Build a model
importModel
Import a saved model
predict
Make a prediction
CS
getModels
Models
Key
Algorithm
XGBoost_model_python_1494306139392_1
XGBoost_model_python_1494306139392_2
XGBoost_model_python_1494306139392_3
XGBoost_model_python_1494306139392_4
XGBoost_model_python_1494306139392_5
XGBoost_model_python_1494306139392_6
XGBoost_model_python_1494306139392_7
XGBoost
Aggregator...
Deep Learning...
Deep Water...
Distributed Random Forest...
Gradient Boosting Machine...
Generalized Linear Modeling...
Generalized Low Rank Modeling...
K-means...
Naive Bayes...
Principal Components Analysis...
Stacked Ensemble...
Word2Vec...
XGBoost...
List All Models
List Grid Search Results
Import Model...
Export Model...


H2O.ai Machine Learning – Gradient Boosting Machine
Time to Train 16 H2O XGBoost Models
2x Xeon CPUs**
1x P100 GPU*
2x P100 GPU*
4x P100 GPU*
8x P100 GPU*
1629
404
271
162
95
GPU
CPU
0
200
400
600
800
1000
1200
1400
1600
1800
time [sec]
*NVIDIA DGX-1, **Dual Intel Xeon E5-2630 v4
Higgs dataset (binary classification): 1M rows, 29 cols; max_depth: {6,8,10,12}, sample_rate: {0.7,0.8,0.9,1.0}
http://github.com/h2oai/perf/

GBM план

1. Че кого

2. Че, есть чо

3. Че как

4. Ложная дилемма

1. Ubuntu 16.04

2. Intel 7850K

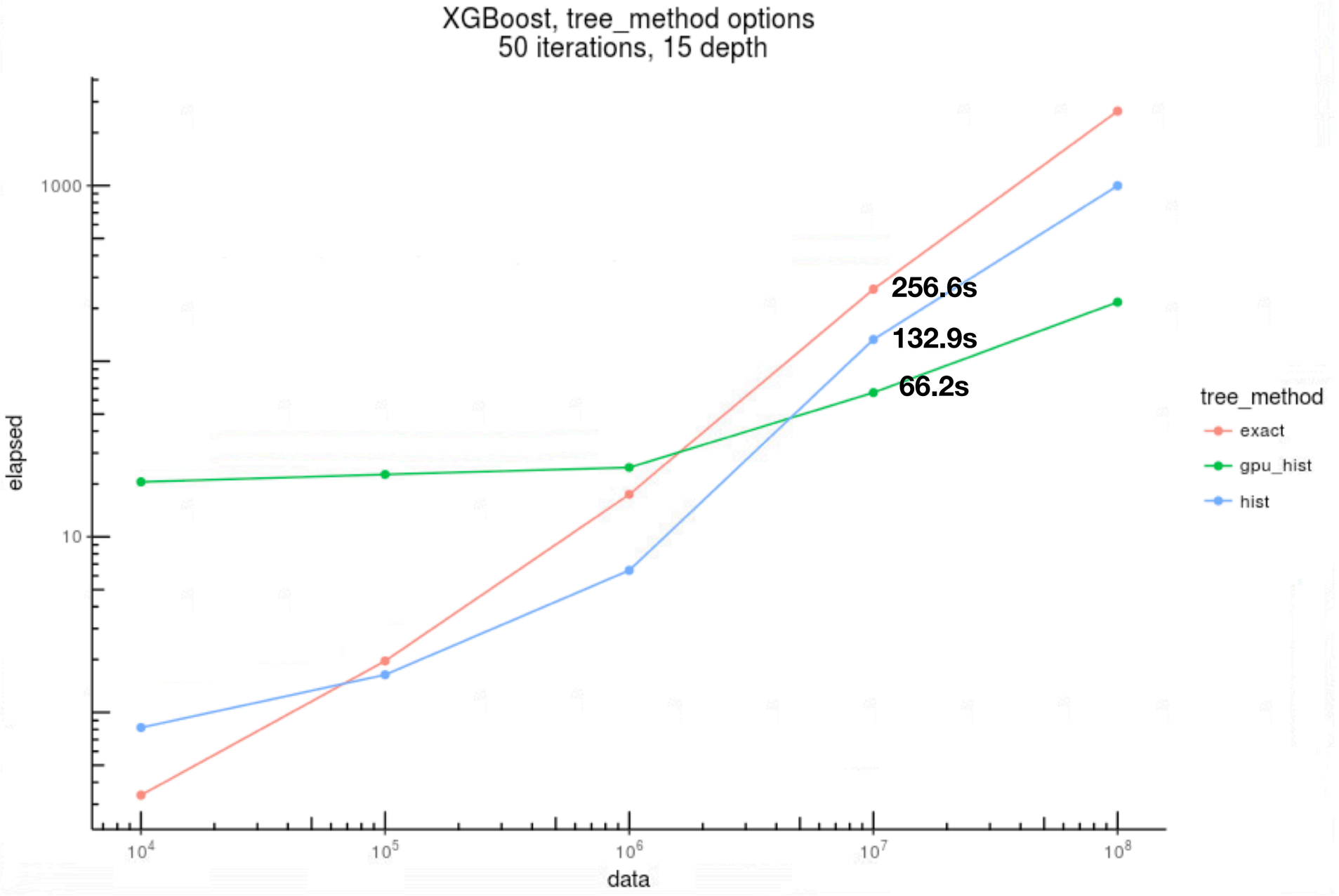
3. 1080 Ti, Asus

4. XGBoost 0.6.4.7

5. LightGBM v2

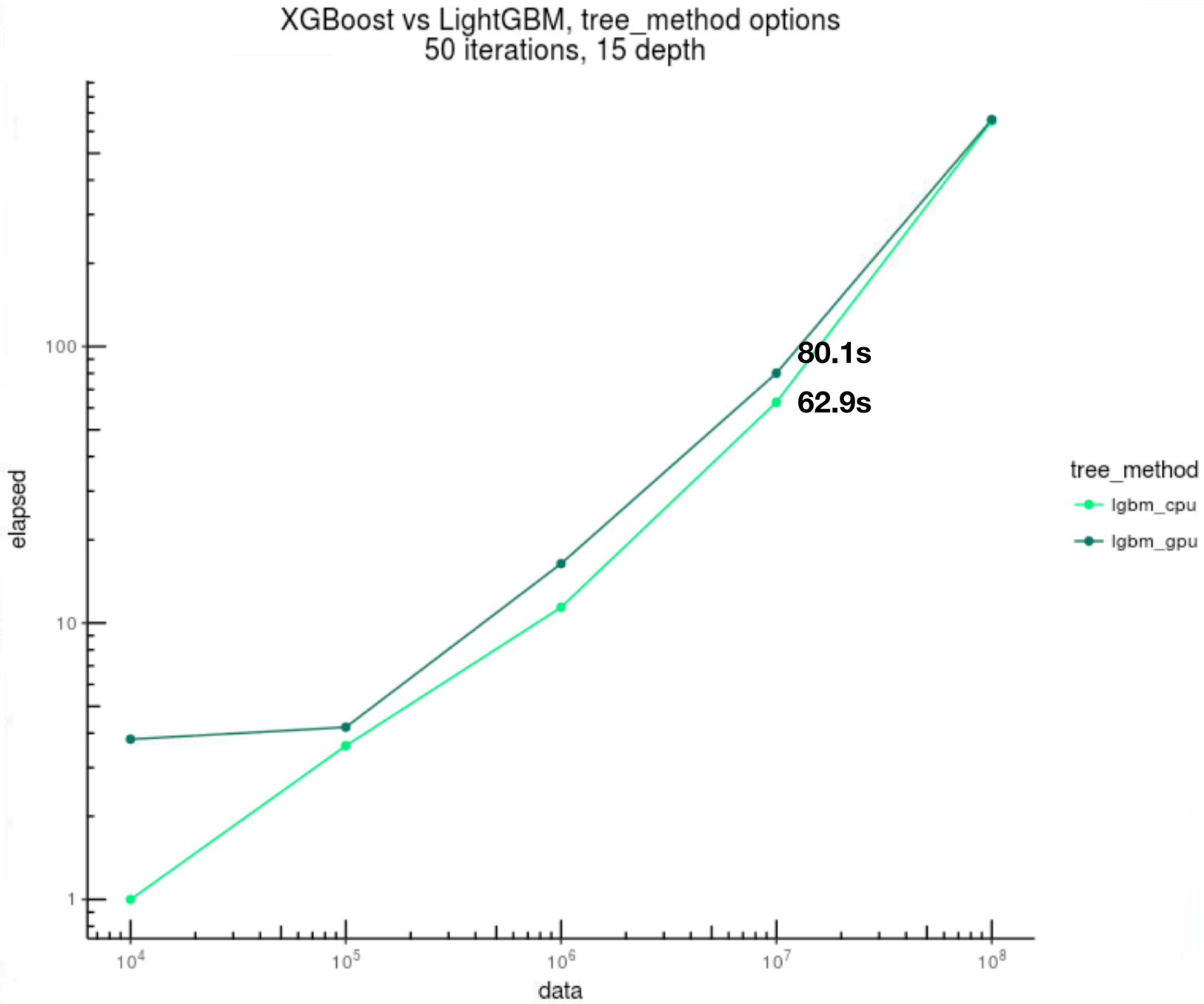
Airlines dataset
2005, 2006 - train
2007 - test

- n_rounds = 50
- max_leaves = 255
- max_depth = 15
- max_bin = 63



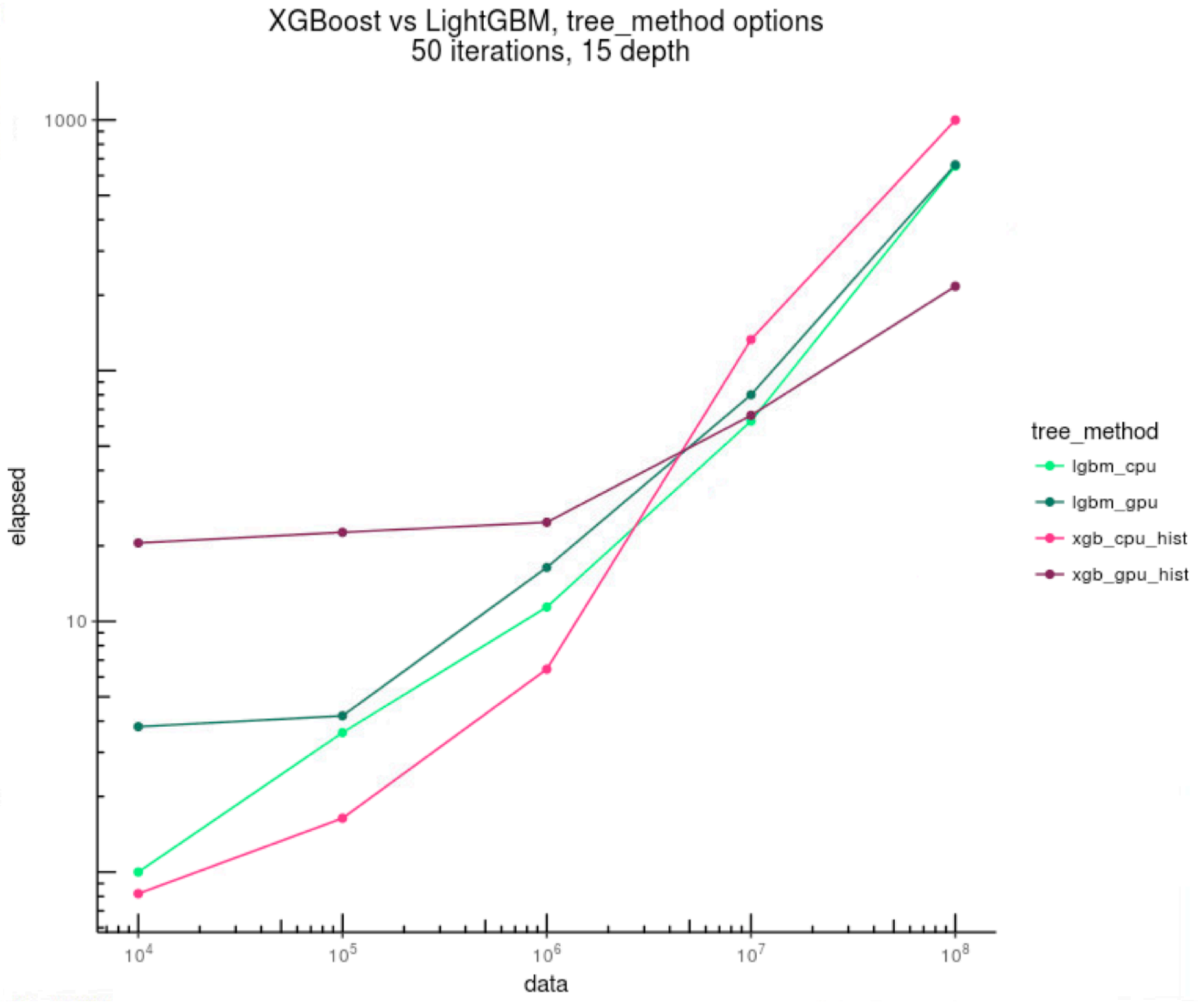
Airlines dataset
2005, 2006 - train
2007 - test

- n_rounds = 50
- max_leaves = 255
- max_depth = 15
- max_bin = 63



Airlines dataset
2005, 2006 - train
2007 - test

- n_rounds = 50
- max_leaves = 255
- max_depth = 15
- max_bin = 63



Airline subsample size	Lib	CPU (50 rounds)		GPU (50 rounds)		CPU (200 rounds)		GPU (200 rounds)		CPU (500 rounds)		GPU (500 rounds)	
		training time	AUC	training time	AUC	training time	AUC	training time	AUC	training time	AUC	training time	AUC
10,000	xgb	**0.139s**	0.725	4.099s	0.742	**0.867s**	0.729	**3.063s**	**0.746**	1.287s	**0.720**	**3.605s**	**0.732**
	xgb_hist	0.544s	0.735	1.510s	0.742	4.791s	**0.736**	4.250s	**0.746**	5.432s	0.715	7.456s	0.722
	lgb	0.141s	**0.737**	**0.985s**	**0.743**	2.149s	0.724	3.924s	0.743	**1.093s**	0.718	6.124s	0.726
100,000	xgb	0.887s	0.775	3.773s	0.776	**3.433s**	0.795	**8.715s**	0.794	8.288s	**0.801**	**14.523s**	**0.804**
	xgb_hist	2.590s	0.786	**3.126s**	0.789	15.108s	0.795	17.139s	**0.795**	32.416s	0.792	33.020s	0.796
	lgb	**0.720s**	**0.788**	4.716s	**0.790**	4.400s	**0.796**	14.481s	0.794	**7.575s**	0.795	30.175s	0.796
1,000,000	xgb	10.509s	0.788	15.416s	0.786	51.787s	0.820	59.977s	0.821	154.981s	0.833	128.406s	0.834
	xgb_hist	6.088s	**0.806**	**7.484s**	0.805	23.949s	**0.831**	29.357s	0.833	50.610s	0.837	55.009s	0.839
	lgb	**2.282s**	**0.806**	8.312s	**0.806**	**10.663s**	**0.831**	**26.191s**	**0.834**	**21.502s**	**0.838**	**42.278s**	**0.840**
10,000,000	xgb	143.350s	0.791	153.138s	0.790	565.631s	0.828	526.385s	0.826	1579.905s	0.844	1392.087s	0.843
	xgb_hist	34.149s	0.791	64.802s	**0.808**	124.467s	0.840	159.932s	0.840	229.613s	0.854	238.149s	0.854
	lgb	**28.167s**	**0.808**	**34.269s**	0.807	**111.330s**	**0.841**	**92.459s**	**0.841**	**162.910s**	**0.855**	**124.260s**	**0.855**
100,000,000	xgb	1732.920s	0.791	-	-	6971.948s	0.828	-	-	-*	-*	-	-
	xgb_hist	304.908s	0.808	427.252s	0.808	1008.774s	0.841	1141.762s	0.840	1958.587s	0.856	2098.376s	0.856
	lgb	**252.798s**	**0.809**	**307.032s**	**0.809**	**969.227s**	**0.842**	**506.922s**	**0.841**	**1610.885s**	**0.857**	**977.676s**	**0.857**

Benchmark of XGBoost, XGBoost hist and LightGBM training time and AUC for different data sizes and rounds. Same as before, XGBoost in GPU for 100 million rows is not shown due to an out of memory (-). In XGBoost for 100 million rows and 500 rounds we stopped the computation after 5 hours (-). The best training time and the highest AUC for each sample size are in boldface text.*

https://blogs.technet.microsoft.com/machinelearning/2017/07/25/lessons-learned-benchmarking-fast-machine-learning-algorithms/

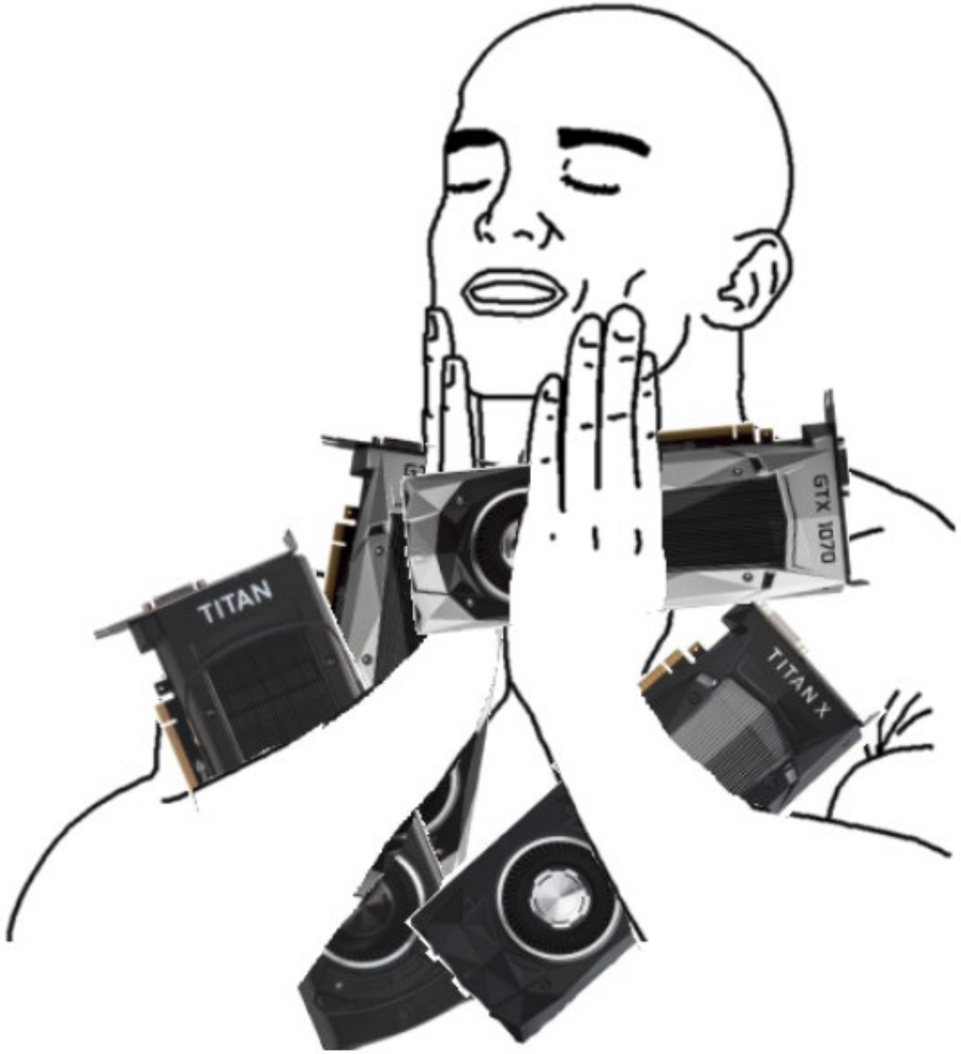
GBM план

1. Че кого

2. Че, есть чо

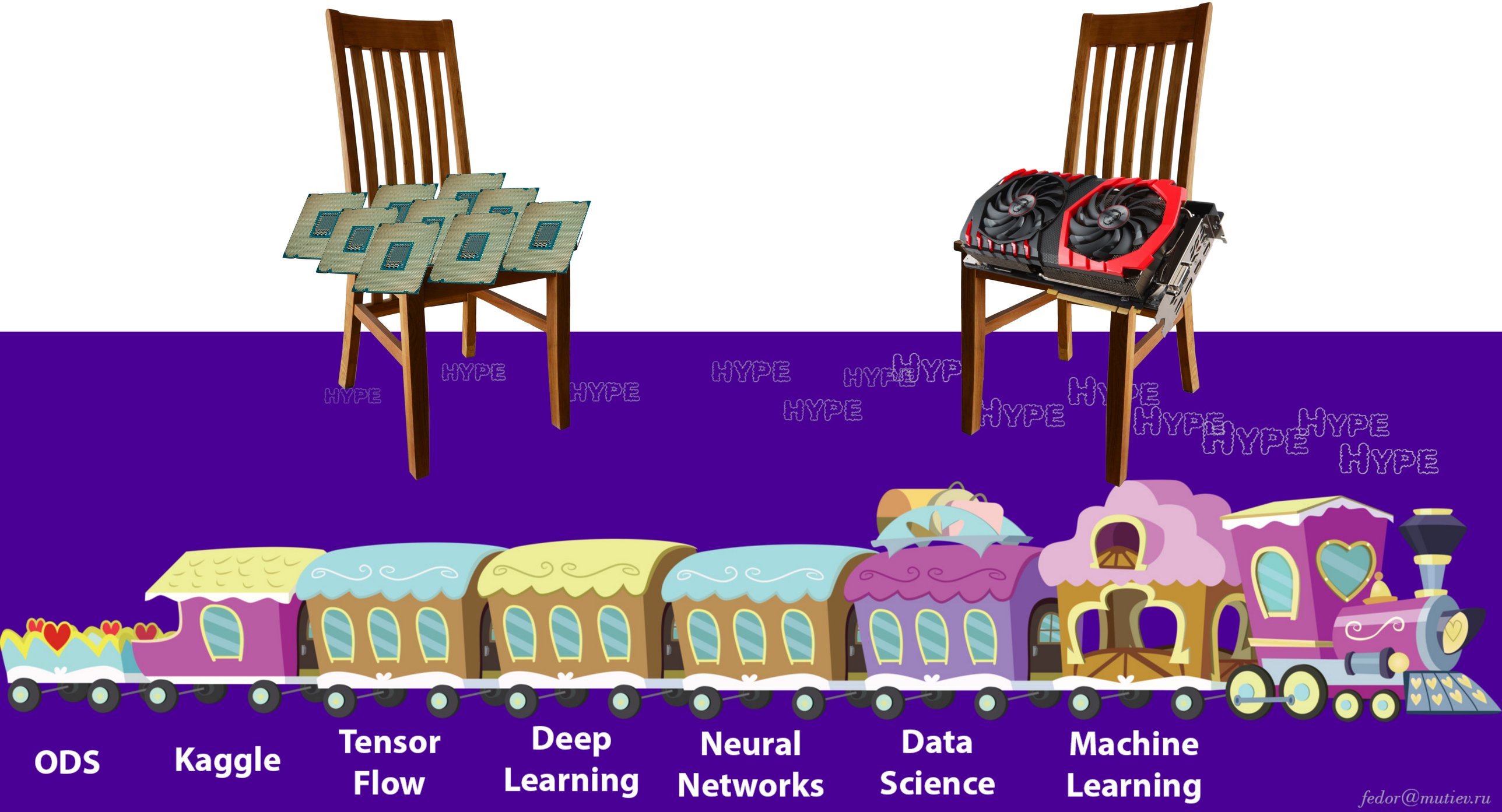
3. Че как

4. Ложная дилемма




Microsoft
LightGBM


dmlc
XGBoost


HYPE
ODS
Kaggle
Tensor Flow
Deep Learning
Neural Networks
Data Science
Machine Learning
fedor@mutiev.ru

Хозяйке на заметку

1. CPU версии и так очень эффективны
2. Лучше купите проц(ы) мощнее
   особенно если данных очень много
3. Деревья на GPU - не лучшая идея

Спасибо!

ods.ai
@natekin