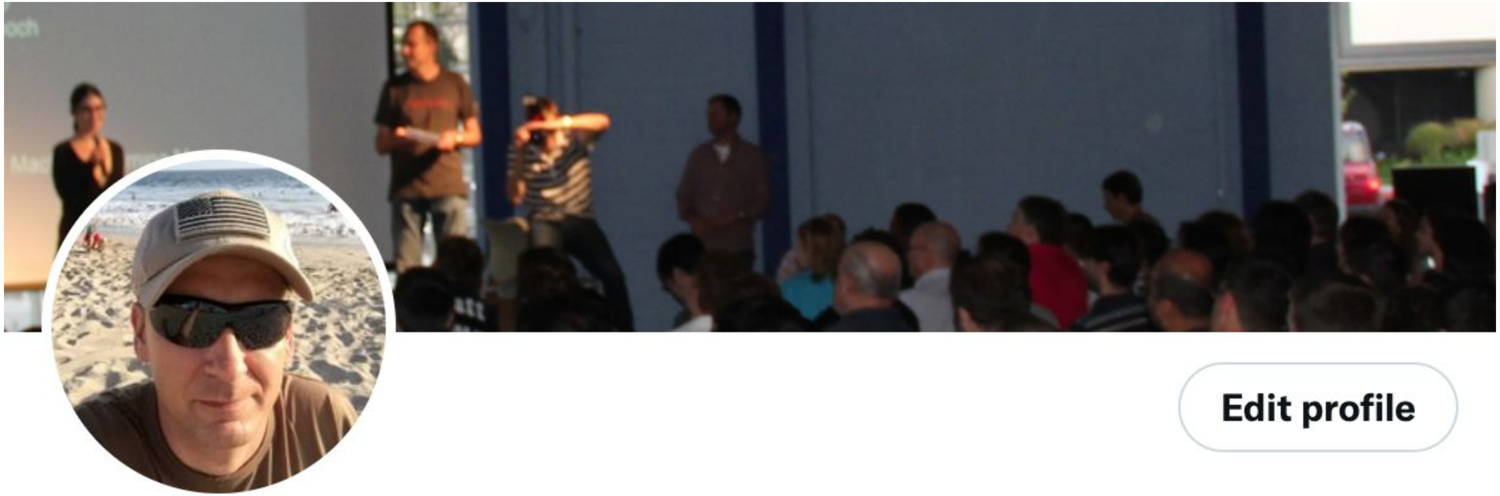
Best Algo for Tabular/Business Data? Sorry, It's Not Deep Learning…

Szilard Pafka, PhD
Chief Scientist, Epoch (USA)

Budapest ML Forum (Online)
May 2022



Edit profile

Szilard [Deeper than Deep Learning]

@SzilardPafka

physics PhD, chief (data) scientist, meetup organizer, (visiting) professor, machine learning benchmarks 🇺🇸🇨🇱🇭🇺🇪🇺

The Woodlands, Texas szilard.github.io/aboutme/

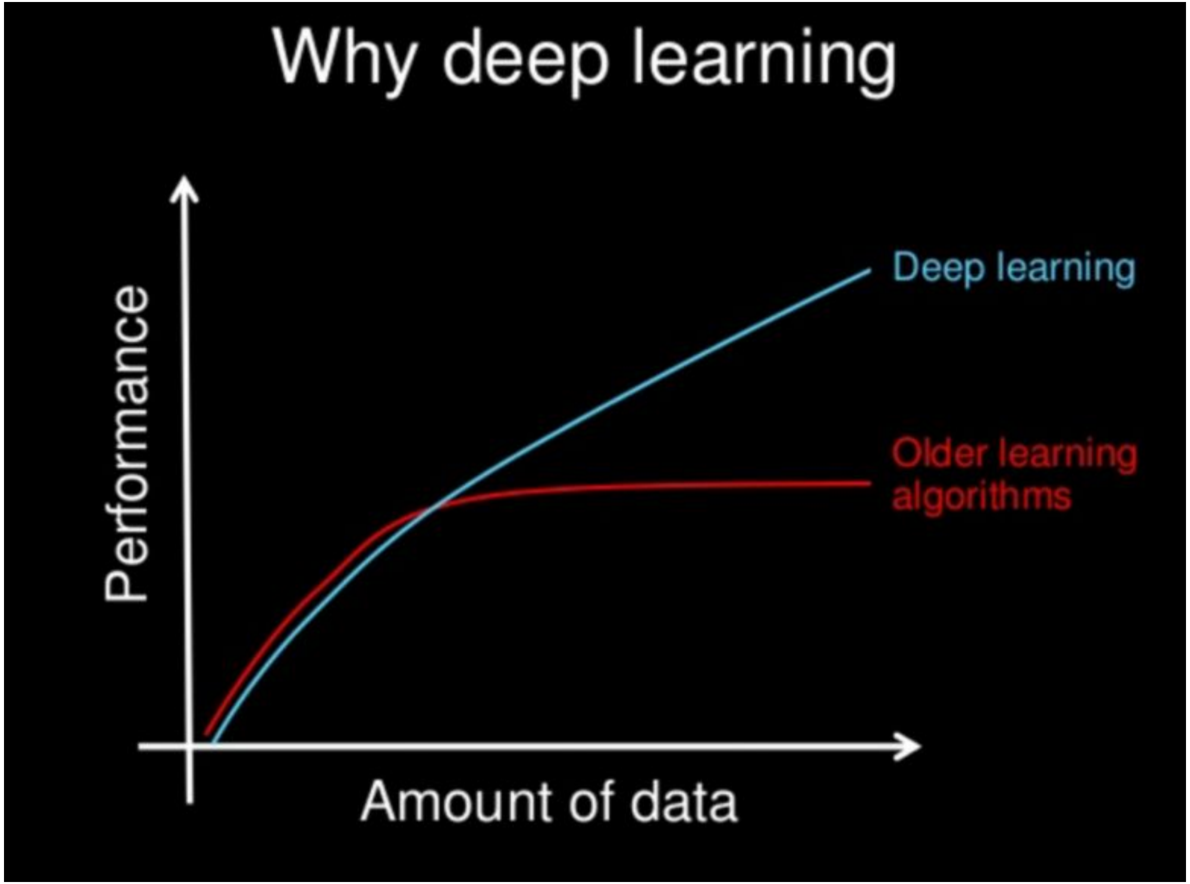
Joined February 2014

226 Following **4,736** Followers

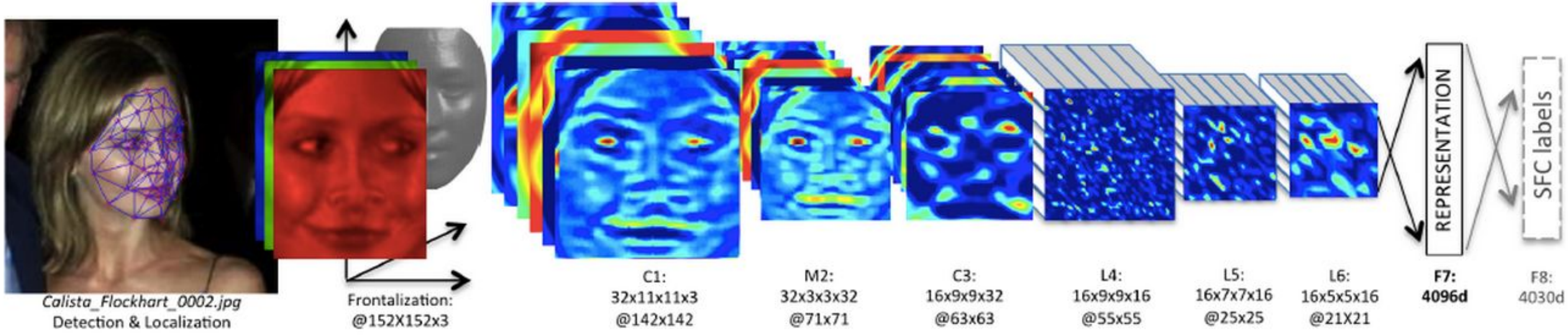
Disclaimer:

I am not representing my employer (Epoch) in this talk

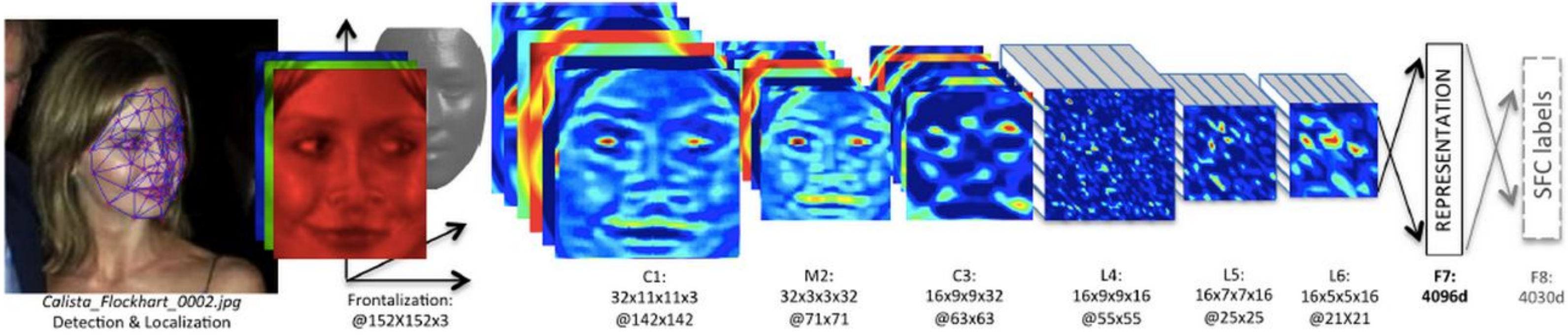
I cannot confirm nor deny if Epoch is using any of the methods, tools, results etc. mentioned in this talk

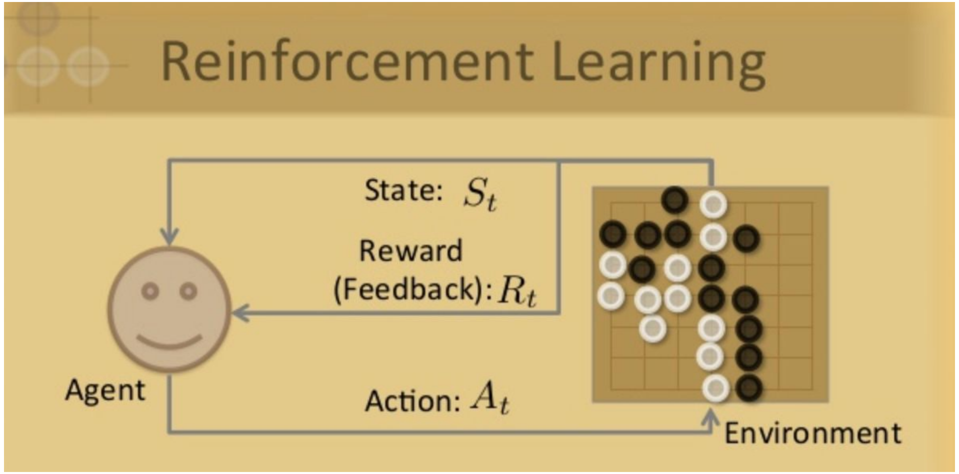


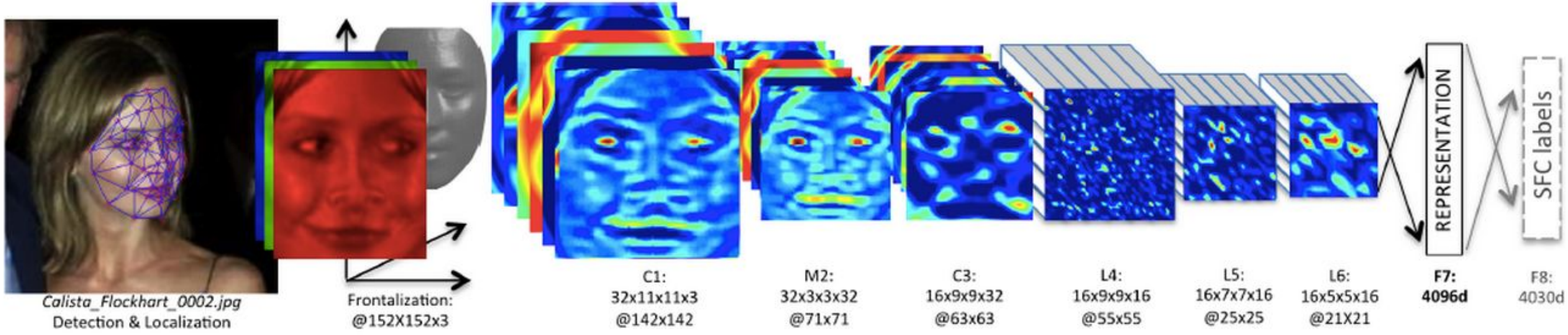
Source: Andrew Ng

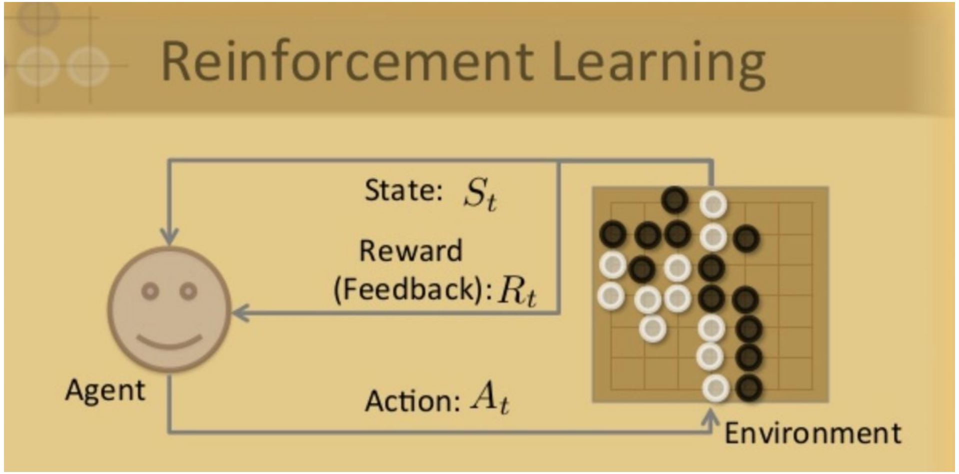


Calista_Flockhart_0002.jpg
Detection & Localization
Frontalization:
@152X152x3
C1:
32x11x11x3
@142x142
M2:
32x3x3x32
@71x71
C3:
16x9x9x32
@63x63
L4:
16x9x9x16
@55x55
L5:
16x7x7x16
@25x25
L6:
16x5x5x16
@21X21
REPRESENTATION
F7:
4096d
SFC labels
F8:
4030d

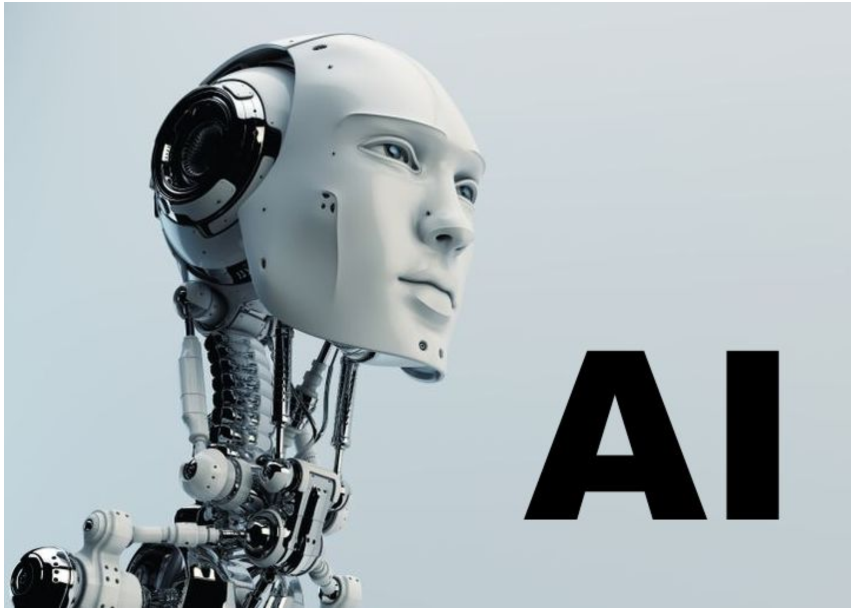

Calista_Flockhart_0002.jpg
Detection & Localization
Frontalization:
@152X152x3
C1:
32x11x11x3
@142x142
M2:
32x3x3x32
@71x71
C3:
16x9x9x32
@63x63
L4:
16x9x9x16
@55x55
L5:
16x7x7x16
@25x25
L6:
16x5x5x16
@21X21
REPRESENTATION
F7:
4096d
SFC labels
F8:
4030d

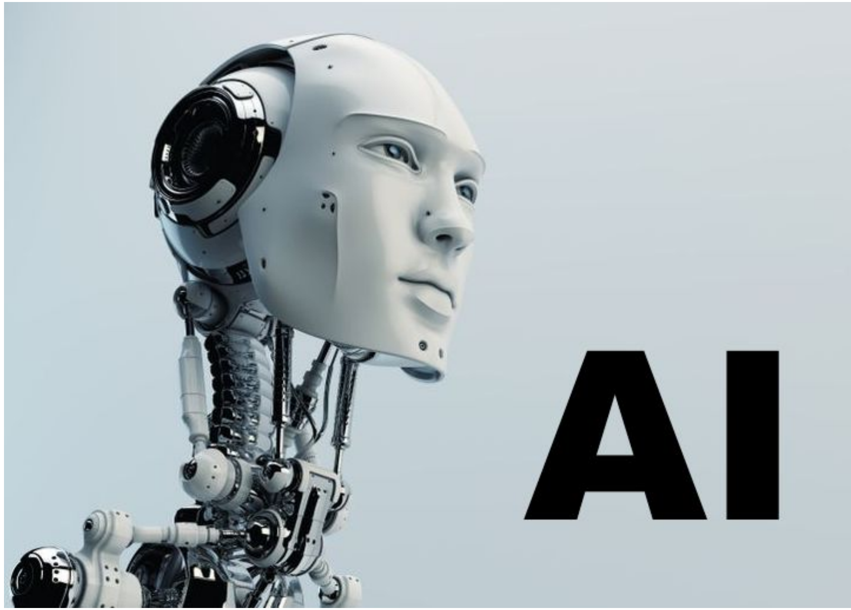

Reinforcement Learning
State: S_t
Reward
(Feedback): R_t
Agent
Action: A_t
Environment


Calista_Flockhart_0002.jpg
Detection & Localization
Frontalization:
@152X152x3
C1:
32x11x11x3
@142x142
M2:
32x3x3x32
@71x71
C3:
16x9x9x32
@63x63
L4:
16x9x9x16
@55x55
L5:
16x7x7x16
@25x25
L6:
16x5x5x16
@21X21
REPRESENTATION
F7:
4096d
SFC labels
F8:
4030d


Reinforcement Learning
State: S_t
Reward
(Feedback): R_t
Agent
Action: A_t
Environment

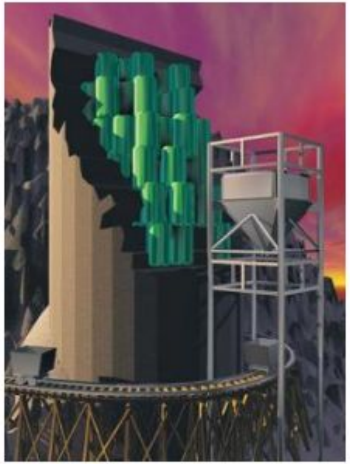

GPT-3
(414) 257 1122
122 N. Mason Street...
984574398275439...
John Henry Smith IV
392-12-11...


AI


AI

KDD-2001 San Francisco, CA August 26-29

More information

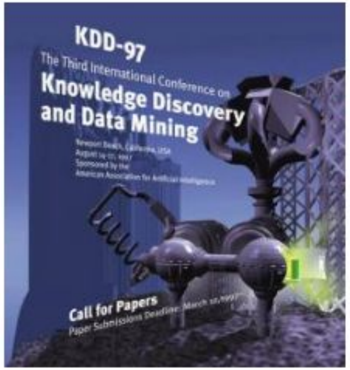


KDD-1996 Portland, OR August 2-4

More information

KDD-2002 Edmonton, AB July 23-26

More information



KDD-1997 Newport Beach, CA August 14-17

More information

KDD-1998 New York, NY August 27-31

More information

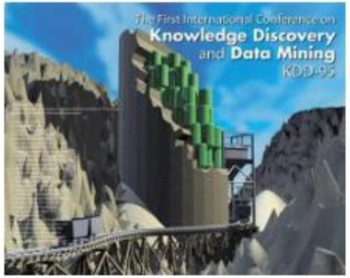
KDD-1993 workshop Washington, D.C., July 11-12

More information

KDD-1989 workshop Detroit, MI, August 20

KDD-2000 Boston, MA August 20-23

More information



KDD-1995 Montreal, QC August 20-21

More information


FRAUD
FRAUD


710
POOR
GOOD


marketing
client
delivering
customers
profit
solution
concept
creating
relationships
techniques
activities
society
competitor
production
identify
strategy
brand
value
offer
business
management
development
organization
build
activity
strong
long term
satisfaction
communicating
focus
goal
goods
variety
support
sales
view
contribute
generate
profitable
sell
term

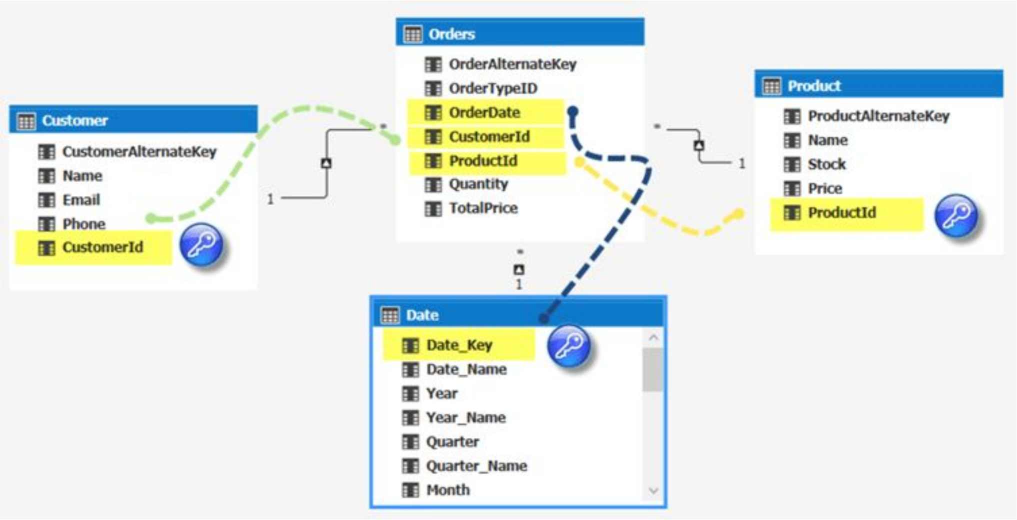

FRAUD
FRAUD


710
POOR
GOOD


marketing
client
delivering
customers
profit
solution
concept
strategy
brand
value
offer
business
management
development
organization
communicating
satisfaction
build
activity
strong
creating
relationships
techniques
society
activities

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00



Params	AUC	Time (s)	Epochs
default: `activation = "Rectifier", hidden = c(200,200)`	73.1	270	1.8
`hidden = c(50,50,50,50), input_dropout_ratio = 0.2`	73.2	140	2.7
`hidden = c(50,50,50,50)`	73.2	110	1.9
`hidden = c(20,20)`			
`hidden = c(20)`			
`RectifierWithDropout, c(200,200,20`			
`ADADELTA rho = 0.95, epsilon = 1e-`			
`rho = 0.999, epsilon = 1e-08`	73.3	270	1.9
`adaptive = FALSE` default: `rate = 0.005, decay = 1, momentum = 0`	73.0	340	1.1
`rate = 0.001, momentum = 0.5 / 1e5 / 0.99`	73.2	410	0.7
`rate = 0.01, momentum = 0.5 / 1e5 / 0.99`	73.3	280	0.9
`rate = 0.01, rate_annealing = 1e-05, momentum = 0.5 / 1e5 / 0.99`	73.5	360	1
`rate = 0.01, rate_annealing = 1e-04, momentum = 0.5 / 1e5 / 0.99`	72.7	3700	8.7
`rate = 0.01, rate_annealing = 1e-05, momentum = 0.5 / 1e5 / 0.9`	73.4	350	0.9

szilard commented Nov 27, 2015

Trying to see if DL can match RF/GBM in accuracy on the airline dataset (wh
years 2005-2006, while validation and test sets sampled disjunctly from 20
kept categorical artificially and are intentionally not encoded as ordinal varia
structure of business datasets).

szilard / **benchm-ml**

DL with h2o #28

Closed szilard opened this issue on Nov 27, 2015 · 14 comments

szilard commented on Nov 27, 2015 Owner

Trying to see if DL can match RF/GBM in accuracy on the airline dataset (where train is sampled from years 2005-2006, while validation and test sets sampled disjunctly from 2007). Also, some variables are kept categorical artificially and are intentionally not encoded as ordinal variables (to better match the structure of business datasets).

arnocandel commented on Nov 29, 2015

Yes, after a bit of tinkering, I also cannot get higher than 0.735 test set AUC. On my i7 5820k home PC:

```
system.time({
md <- h2o.deeplearning(x = Xnames, y = "dep_delayed_15min", training_frame = dx_train,
```

some feature engineering (e.g., cutting the original DepTime into 48 categorical half-hour slots). Out of 675 input neurons, only 2 are always populated with non-zero values (the two numeric features), and 673 values are mostly 0, only 6 categoricals are set to 1. That's where the inefficiency comes from. GBM/DRF are much more efficient at simply cutting up the feature space, which is was seems to be needed here.

Best,
Arno

DL with mxnet #29

Closed szilard opened this issue on Nov 27, 2015 · 3 comments

szilard commented on Nov 27, 2015 Owner

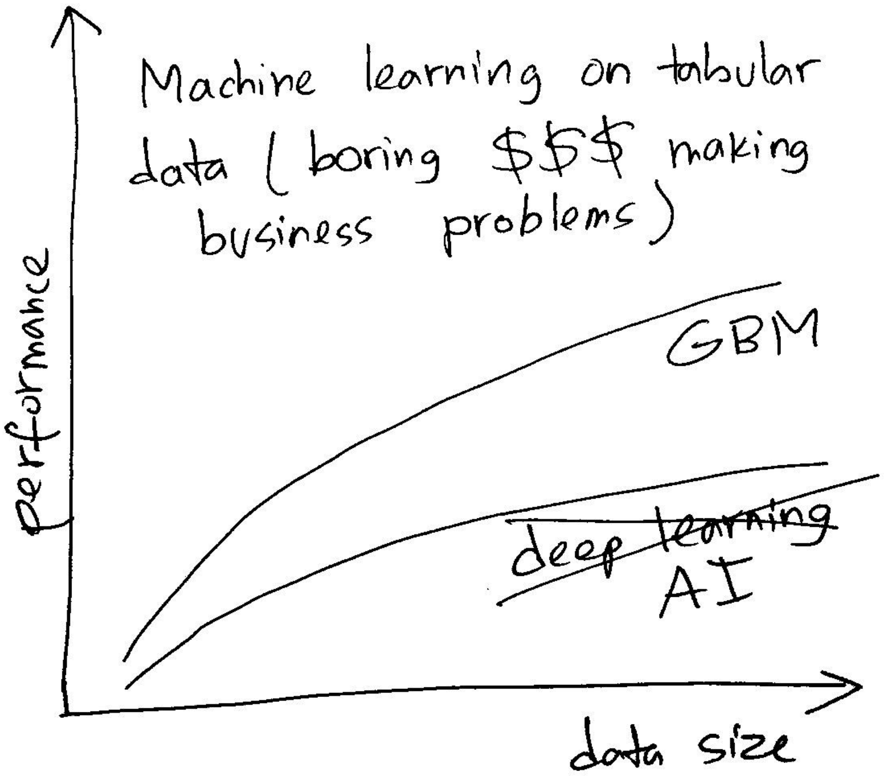
Trying to see if DL can match RF/GBM in accuracy on the airline dataset (where train is sampled from years 2005-2006, while validation and test sets sampled disjunctly from 2007). Also, some variables are kept categorical artificially and are intentionally not encoded as ordinal variables (to better match the structure of business datasets).

tqchen commented on Nov 30, 2015

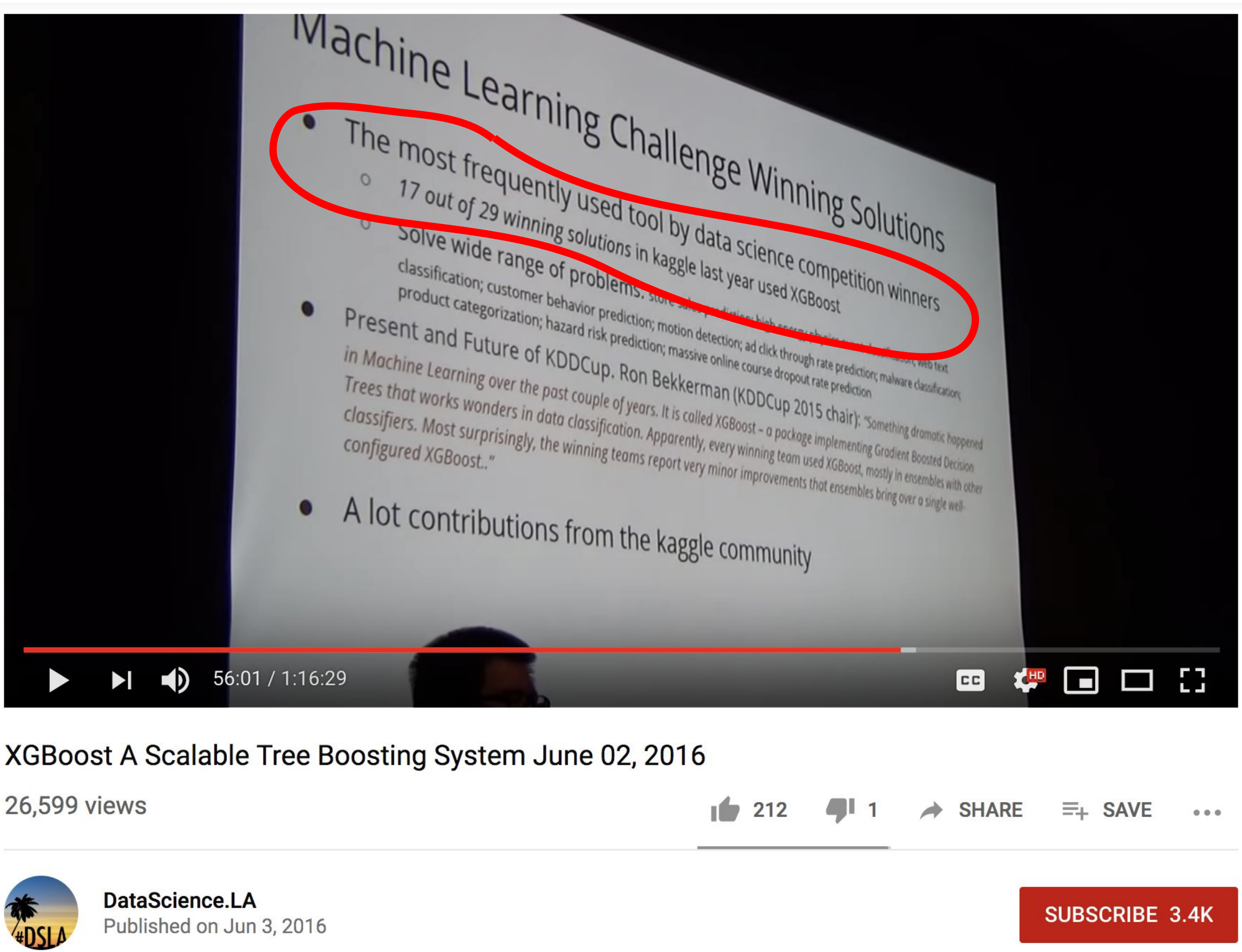
Deep nets are definitely harder to tune, if things converge too fast, try smaller learning rate, shuffle the data. Seems much of gains in the airline dataset comes from combination of categories, which deepnet may not be very good at

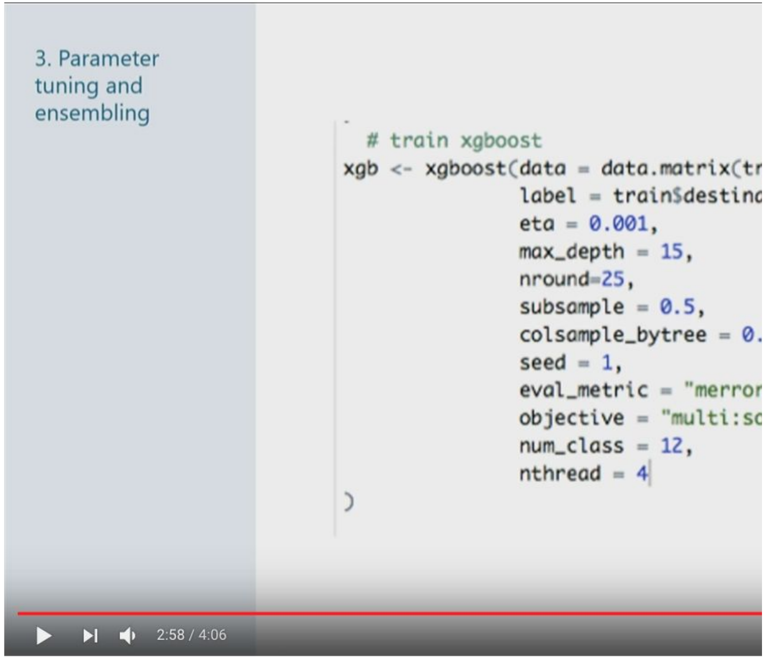
Szilard [Deeper than Deep Learning] @DataScienceLA · 2 Nov 2016

Can anyone **beat** GBMs with deep learning (ahem, AI) on the airline dataset (or generally tabular/business data)? github.com/szilard/benchm…



kaggle


Machine Learning Challenge Winning Solutions
The most frequently used tool by data science competition winners
17 out of 29 winning solutions in kaggle last year used XGBoost
Solve wide range of problems
Present and Future of KDDCup. Ron Bekkerman (KDDCup 2015 chair)
A lot contributions from the kaggle community
56:01 / 1:16:29
XGBoost A Scalable Tree Boosting System June 02, 2016
26,599 views
212
1
SHARE
SAVE
DataScience.LA
Published on Jun 3, 2016
SUBSCRIBE 3.4K

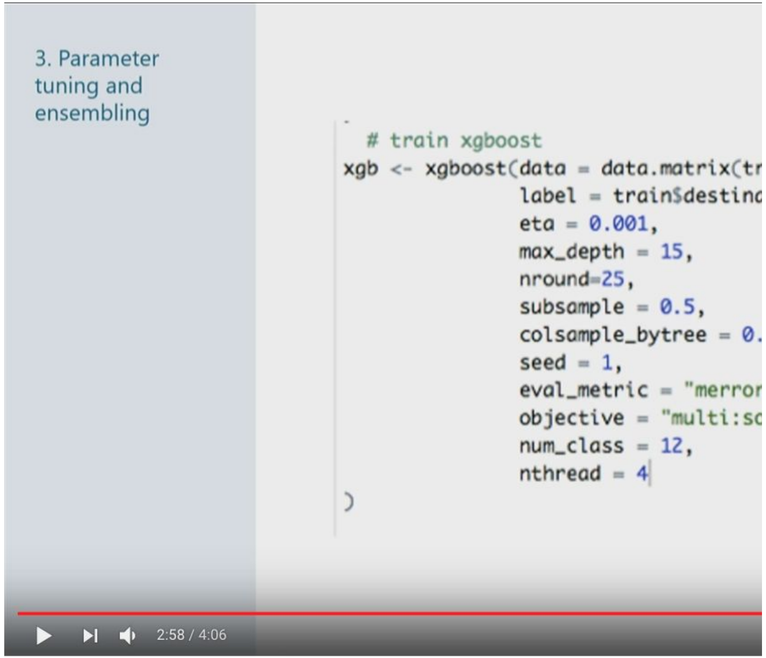


What Kaggle has learned from almost a million data scientists - Anthony Goldbloom

18,153 views

O'Reilly

Published on May 25, 2017



What Kaggle has learned from almost a million data scientists - Anthony Goldbloom

18,153 views

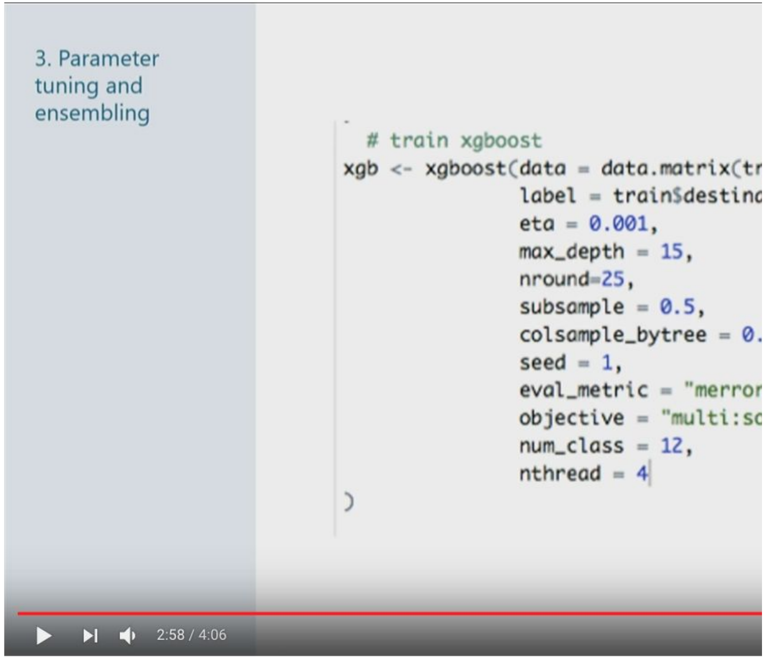
O'Reilly
Published on May 25, 2017

Gilberto Titericz • 1st · 4mo •••

Data Scientist at NVIDIA Rapids

In my experience GBMs are, by far, the best for tabular structured data.

Like · 35 | Reply



What Kaggle has learned from almost a million data scientists - Anthony Goldbloom

18,153 views

O'Reilly
Published on May 25, 2017

Gilberto Titericz • 1st 4mo •••
Data Scientist at NVIDIA Rapids

In my experience GBMs are, by far, the best for tabular structured data.

Like · 35 | Reply

Bojan Tunguz @tunguz · Apr 5

There are two kinds of people in the World.

1. Those who are using XGBoost for **tabular** data
2. Those who will use XGBoost for **tabular** data

11 9 152

Szilard [Deeper than Deep Learning]
@SzilardPafka

Best algo for tabular data? (most often)

Gradient Boosted Trees	**76%**
Neural Nets / Deep Learn.	2%
Other	22%

50 votes · Final results

2:42 PM · Feb 22, 2022 · Twitter Web App

Szilard [Deeper than Deep Learning]
@SzilardPafka

Best algo for tabular data? (most often)

Gradient Boosted Trees	**76%**
Neural Nets / Deep Learn.	2%
Other	22%

50 votes · Final results

2:42 PM · Feb 22, 2022 · Twitter Web App

Szilard Pafka
physics PhD, chief (data) scientist, meetup organizer, (visiting) professor, ...
1mo ·

Best algo for tabular data? (most often)
You can see how people vote. Learn more

Gradient Boosted Trees	92%
Neural Nets / Deep Learning	3%
Other	6%

72 votes · Poll closed

Szilard [Deeper than Deep Learning]
@SzilardPafka

Best algo for tabular data? (most often)

Gradient Boosted Trees	**76%**
Neural Nets / Deep Learn.	2%
Other	22%

50 votes · Final results

2:42 PM · Feb 22, 2022 · Twitter Web App

Szilard Pafka
physics PhD, chief (data) scientist, meetup organizer, (visiting) professor, ...
1mo ·

Best algo for tabular data? (most often)
You can see how people vote. Learn more

Gradient Boosted Trees	92%
Neural Nets / Deep Learning	3%
Other	6%

72 votes · Poll closed

Bojan Tunguz, Ph.D. Author 1mo
Machine Learning at NVIDIA | Physicist | Quadruple Kaggle Gran...

Szilard Pafka Unfortunately it's still far from being mainstream. But some of us are working hard on getting it there.

MLP [188]	TabTransformer [99]
DeepFM [14]	NODE [6]
DeepGBM [52]	DNFNet [43]
RLN [54]	STG [189]
TabNet [5]	NAM [190]
VIME [67]	SAINT [9]

MLP [188]

DeepFM [14]

DeepGBM [52]

RLN [54]

TabNet [5]

VIME [67]

TabTransformer [99]

NODE [6]

DNFNet [43]

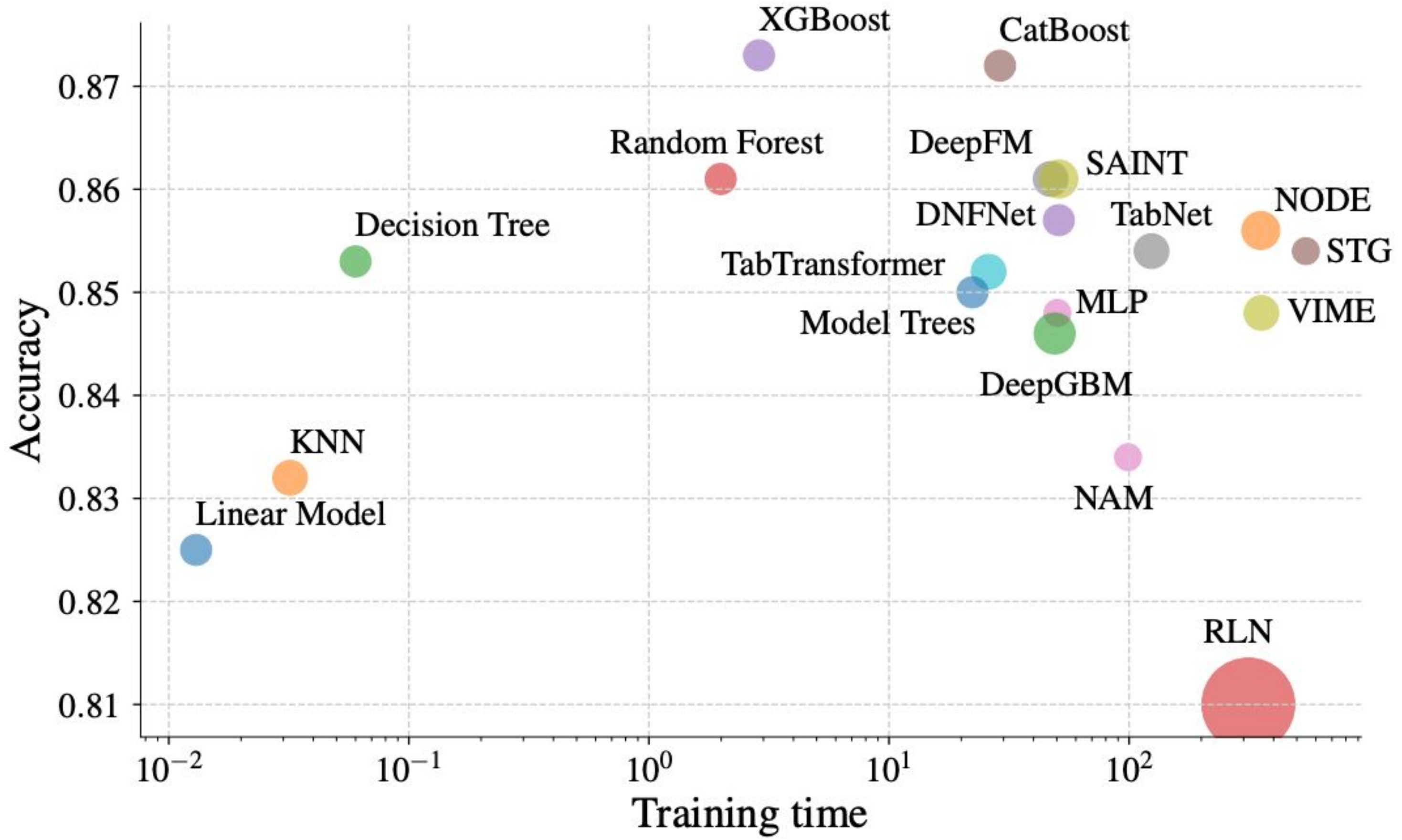
STG [189]

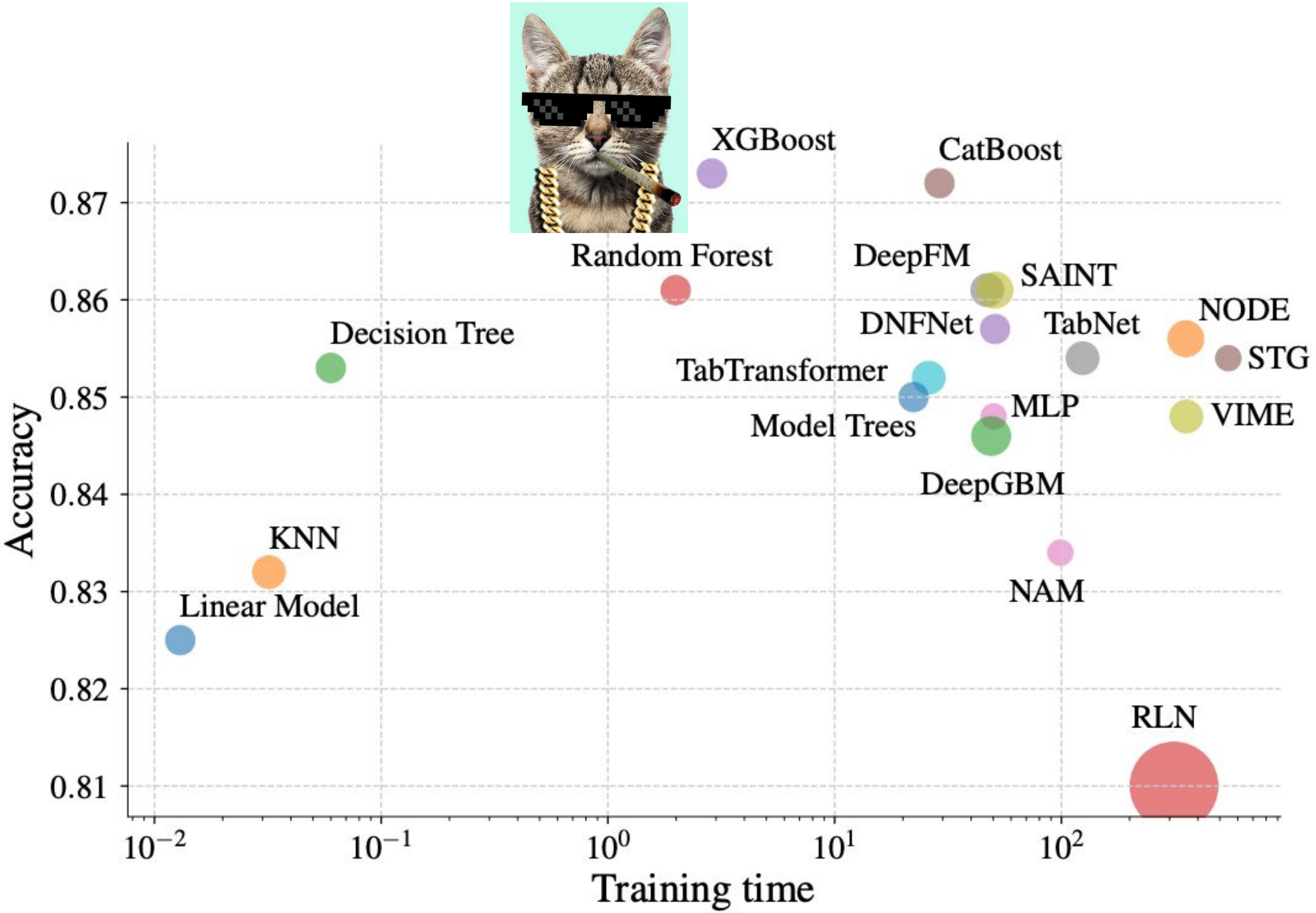
NAM [190]

SAINT [9]

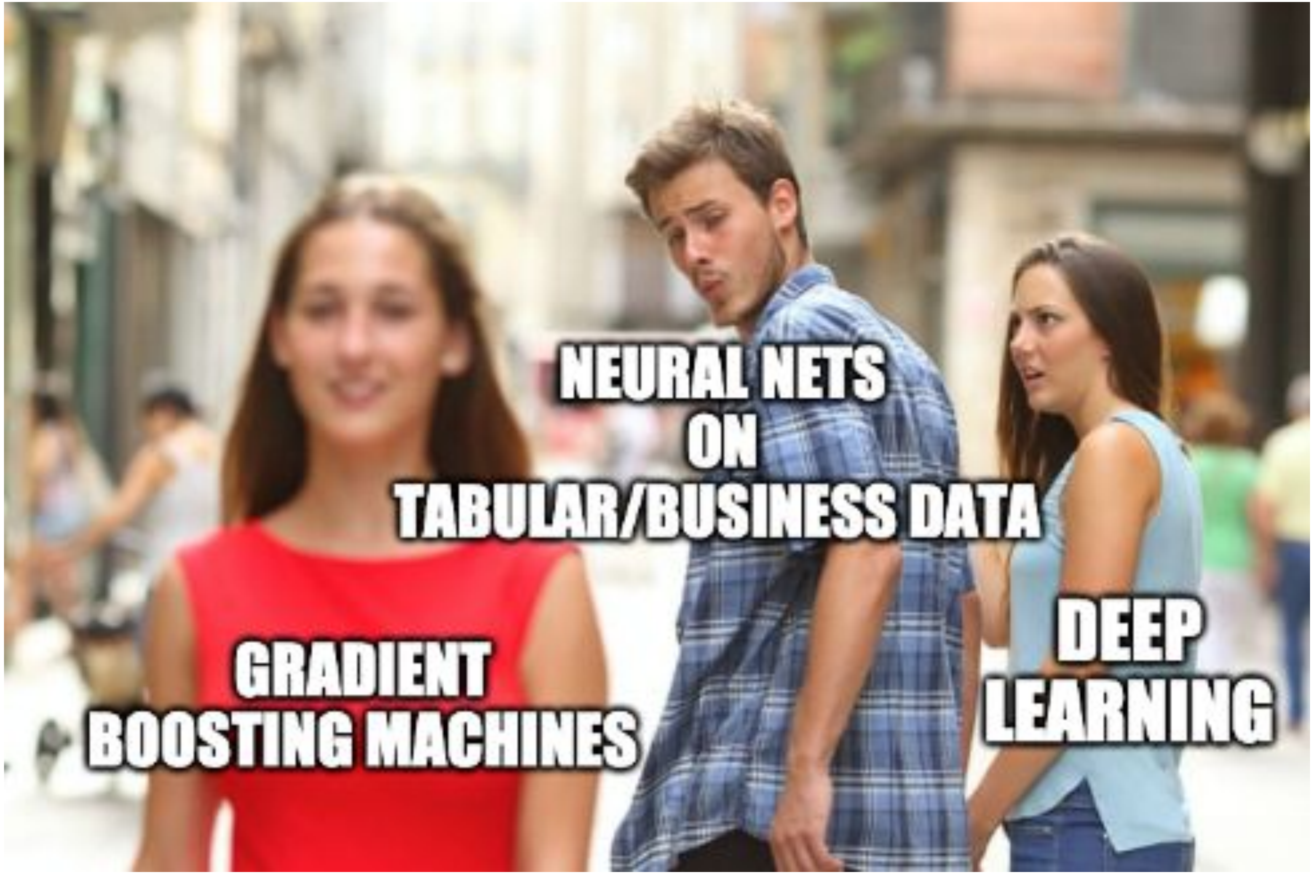
Deep Neural Networks and Tabular Data: A Survey

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug,
Martin Pawelczyk and Gjergji Kasneci


XGBoost
CatBoost
Random Forest
DeepFM
SAINT
NODE
DNFNet
TabNet
STG
Decision Tree
TabTransformer
MLP
VIME
Model Trees
DeepGBM
KNN
NAM
Linear Model
RLN
0.87
0.86
0.85
0.84
0.83
0.82
0.81
Accuracy
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Training time



XGBoost
CatBoost
Random Forest
DeepFM
SAINT
DNFNet
TabNet
NODE
STG
Decision Tree
TabTransformer
MLP
VIME
Model Trees
DeepGBM
KNN
NAM
Linear Model
RLN
Accuracy
0.87
0.86
0.85
0.84
0.83
0.82
0.81
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Training time


NEURAL NETS
ON
TABULAR/BUSINESS DATA
GRADIENT
BOOSTING MACHINES
DEEP
LEARNING

MODEL	1ST	2ND
BST-DT	0.580	0.228
RF	0.390	0.525
BAG-DT	0.030	0.232
SVM	0.000	0.008
ANN	0.000	0.007
KNN	0.000	0.000
BST-STMP	0.000	0.000
DT	0.000	0.000
LOGREG	0.000	0.000
NB	0.000	0.000

AVG	1ST	2ND
RF	0.727	0.207
ANN	0.053	0.172
BSTDT	0.059	0.228
SVM	0.043	0.195
LR	0.089	0.132
BAGDT	0.002	0.012
KNN	0.023	0.045
BSTST	0.004	0.009
PRC	0	0
NB	0	0

An Empirical Comparison of Supervised Learning Algorithms

http://www.cs.cornell.edu/~alexn/papers/empirical.icml06.pdf

An Empirical Evaluation of Supervised Learning in High Dimensions

http://lowrank.net/nikos/pubs/empirical.pdf

MODEL	1ST	2ND
BST-DT	0.580	0.228
RF	0.390	0.525
BAG-DT	0.030	0.232
SVM	0.000	0.008
ANN	0.000	0.007
KNN	0.000	0.000
BST-STMP	0.000	0.000
DT	0.000	0.000
LOGREG	0.000	0.000
NB	0.000	0.000

AVG	1ST	2ND
RF	0.727	0.207
ANN	0.053	0.172
BSTDT	0.059	0.228
SVM	0.043	0.195
LR	0.089	0.132
BAGDT	0.002	0.012
KNN	0.023	0.045
BSTST	0.004	0.009
PRC	0	0
NB	0	0

gbm_1.5-3.tar.gz	2005-10-07 22:49 249K
gbm_1.5-5.tar.gz	2006-01-21 12:58 249K
gbm_1.5-7.tar.gz	2006-04-18 11:58 254K
gbm_1.5.tar.gz	2005-05-09 22:56 250K
gbm_1.6-1.tar.gz	2007-06-14 08:29 257K

An Empirical Comparison of Supervised Learning Algorithms

http://www.cs.cornell.edu/~alexn/papers/empirical.icml06.pdf

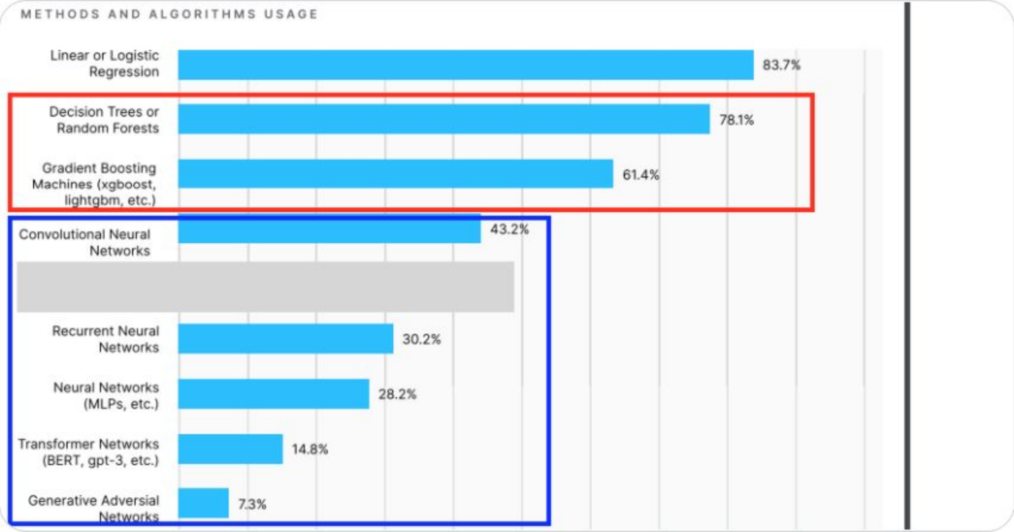
An Empirical Evaluation of Supervised Learning in High Dimensions

http://lowrank.net/nikos/pubs/empirical.pdf

randomForest_4.5-12.tar.gz	2005-06-21 09:36 80K
randomForest_4.5-15.tar.gz	2005-09-22 19:35 81K
randomForest_4.5-16.tar.gz	2006-01-24 10:21 81K
randomForest_4.5-18.tar.gz	2006-12-10 16:07 67K
randomForest_4.5-19.tar.gz	2007-10-16 20:38 67K

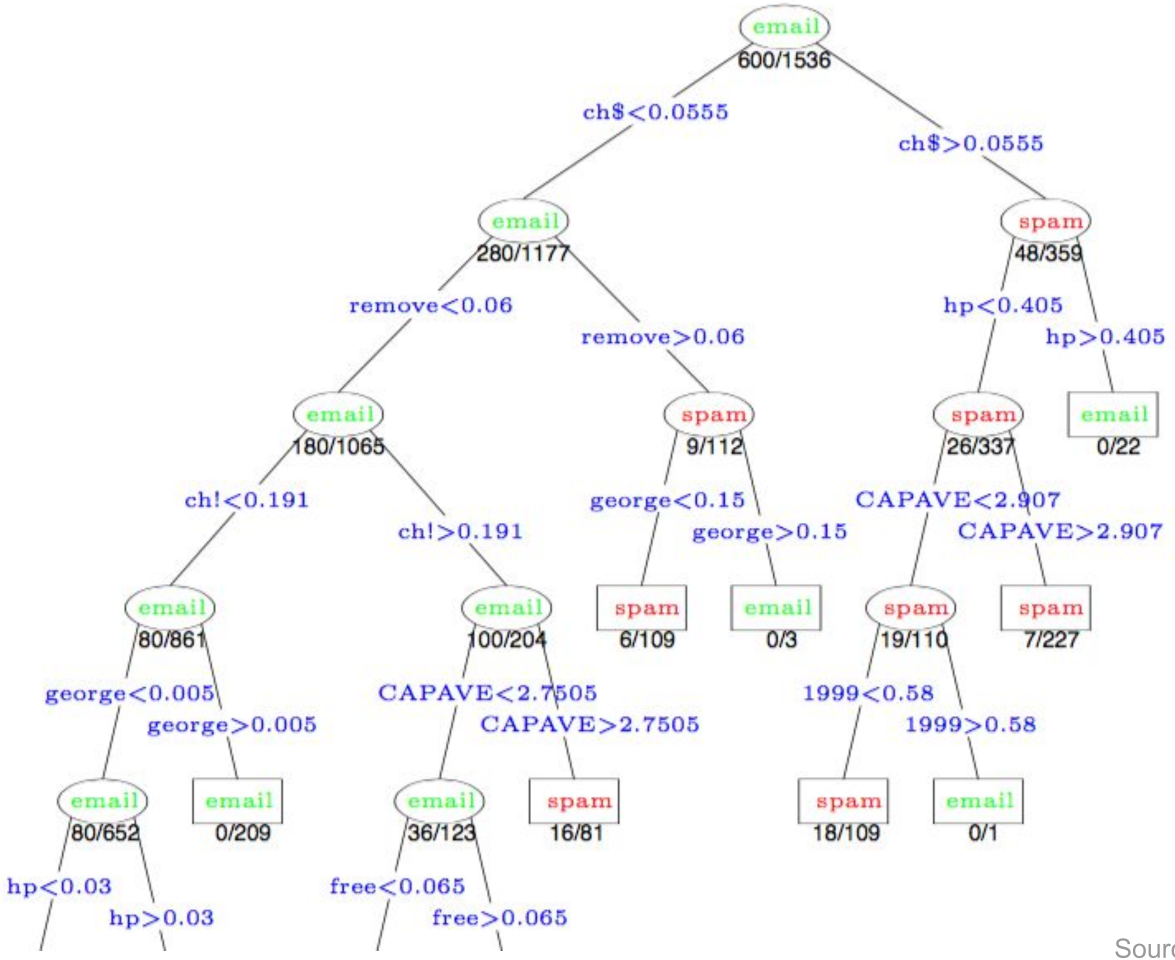
Szilard [Deeper than Deep Learning]
@DataScienceLA

Let's just note that at the end of 2020 gradient boosting (GBMs) and random forests still beat neural networks (deep learning and all that shit) hands down (results from the 2020 Kaggle survey kaggle.com/kaggle-survey-...). GBM/RF ~85% vs NN (any) ~50%. I won a 5-year long bet 🎉🤓




WORLD'S
BEST
BOSS


WORLD'S
BEST
BOSS



Source: Hastie etal, ESL 2ed

Algorithm 10.1 *AdaBoost.M1.*

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.

2. For $m = 1$ to M:

   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.

   (b) Compute

$$\mathrm{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$

   (c) Compute $\alpha_m = \log((1 - \mathrm{err}_m)/\mathrm{err}_m)$.

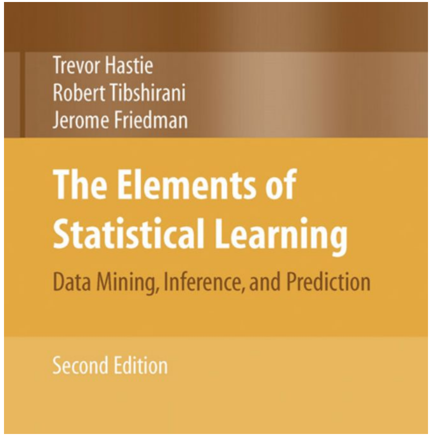
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.

3. Output $G(x) = \mathrm{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

Source: Hastie etal, ESL 2ed

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm},\ j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.



Source: Hastie etal, ESL 2ed

- R packages
- Python scikit-learn
- Vowpal Wabbit
- H2O
- xgboost
- Spark MLlib
- a few others

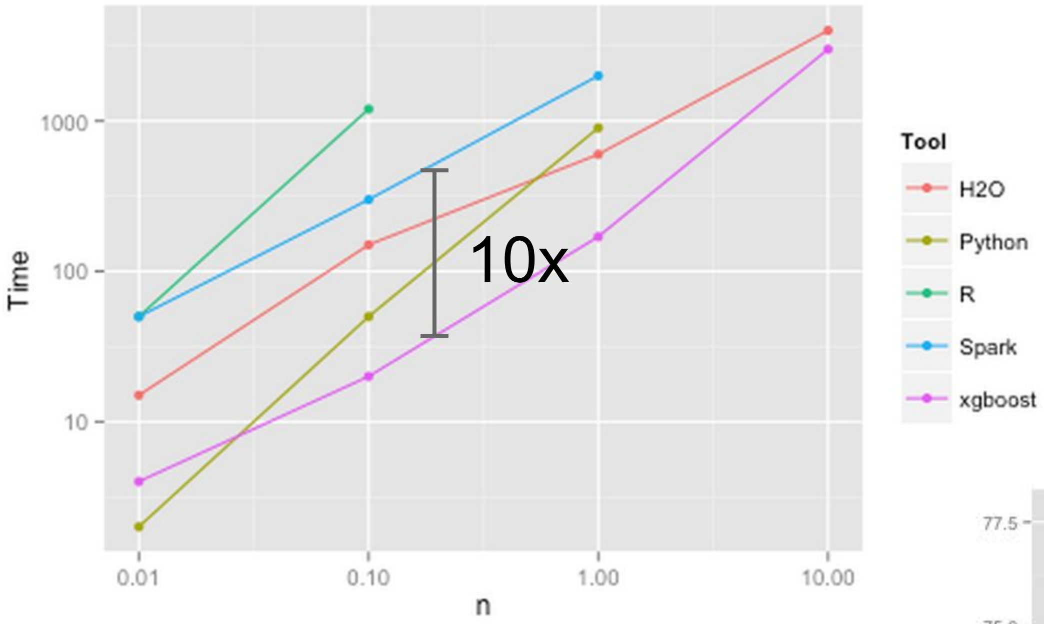
- R packages
- Python scikit-learn
- Vowpal Wabbit
- H2O
- xgboost
- Spark MLlib
- a few others

szilard / **benchm-ml**

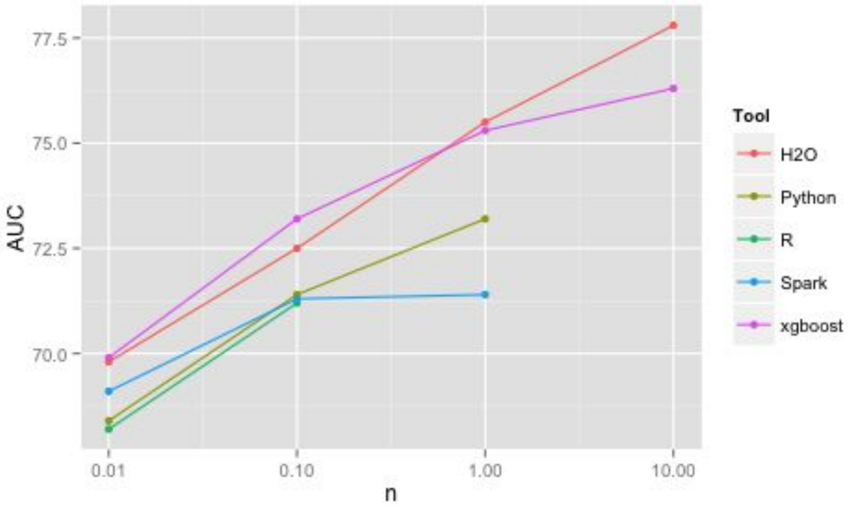
Star 1,203

Simple/limited/incomplete benchmark

(2015-)



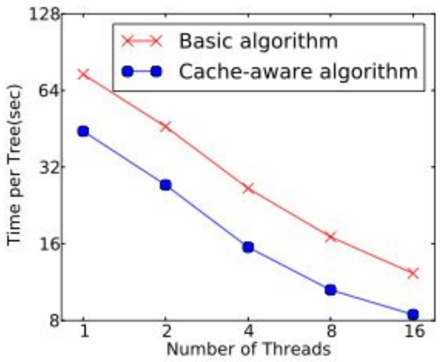
szilard / benchm-ml



arXiv.org > cs > arXiv:1603.02754

Computer Science > Learning

XGBoost: A Scalable Tree Boosting System

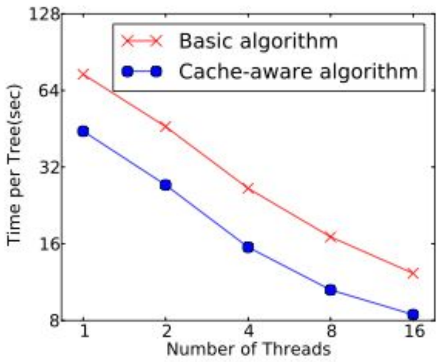


H2O.ai

arXiv.org > cs > arXiv:1603.02754

Computer Science > Learning

XGBoost: A Scalable Tree Boosting System



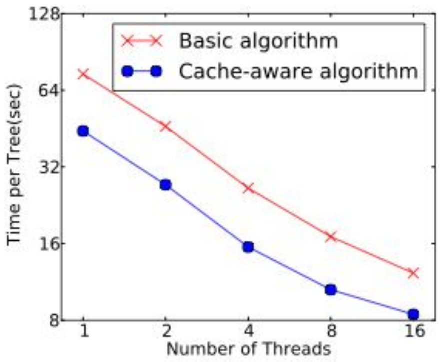
H_2O.ai

2015

arXiv.org > cs > arXiv:1603.02754

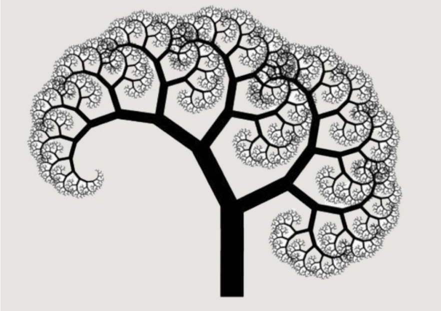
Computer Science > Learning

XGBoost: A Scalable Tree Boosting System



H2O.ai

Microsoft / LightGBM



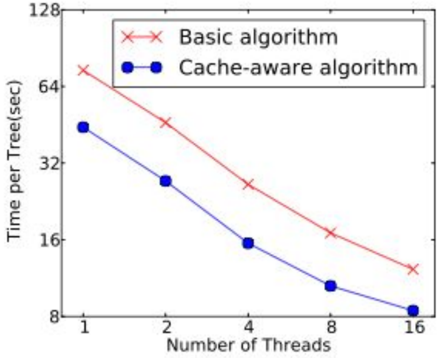
CatBoost

2017

arXiv.org > cs > arXiv:1603.02754

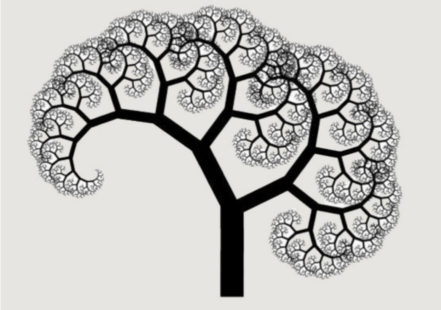
Computer Science > Learning

XGBoost: A Scalable Tree Boosting System



H_2O.ai

Microsoft / LightGBM



CatBoost

https://cran.r-project.org/web/pa

xgboost: Extreme Gradient Boosting

https://cran.r-project.org/

h2o: R Interface for H2O

PyData

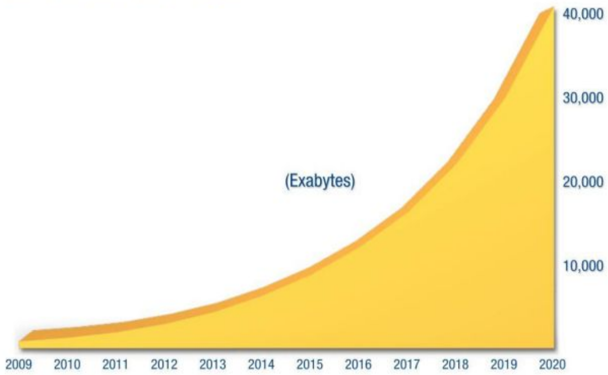
APACHE
Spark™

Szilard [Deeper than Deep Learning]
@DataScienceLA

Why not using Spark for training gradient boosting machines/boosted trees (GBM/GBDT)? Because it's >100x slower and uses >100x more RAM compared to top libraries such as xgboost or lightgbm. 100 f[redacted]ing times worse 🤡🤡🤡 See details in my talk here youtube.com/watch?v=qjuizR...

Figure 1

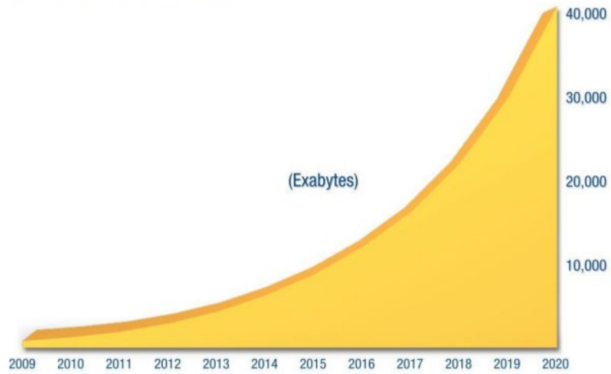
The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Figure 1

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

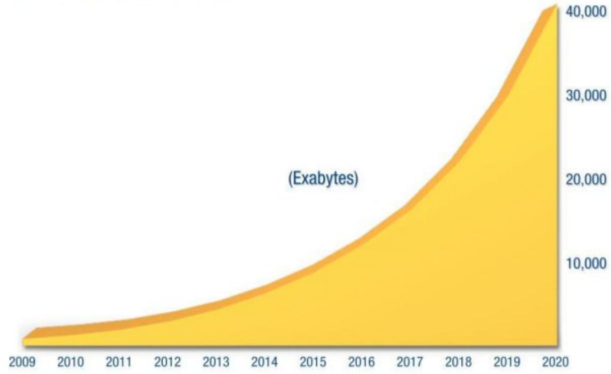
Hadley Wickham
@hadleywickham

Following

"It takes a big man to admit his data is small" — @jcheng

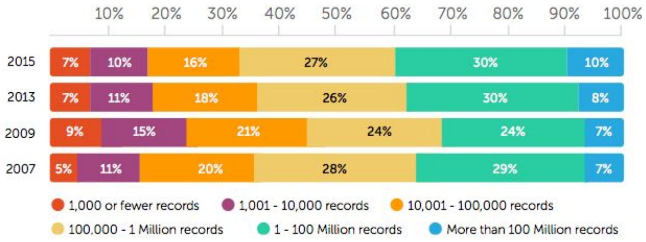
Figure 1

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

TYPICAL SIZE OF DATASETS



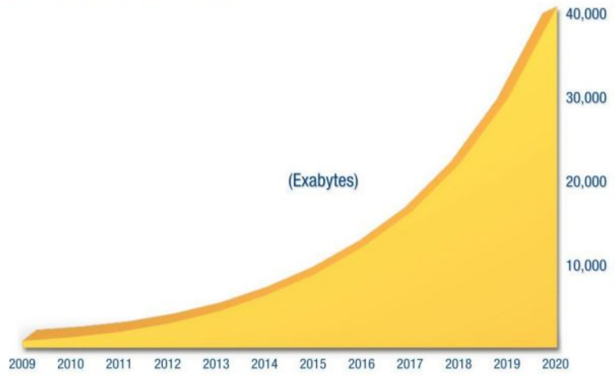
Hadley Wickham
@hadleywickham

Following

"It takes a big man to admit his data is small" — @jcheng

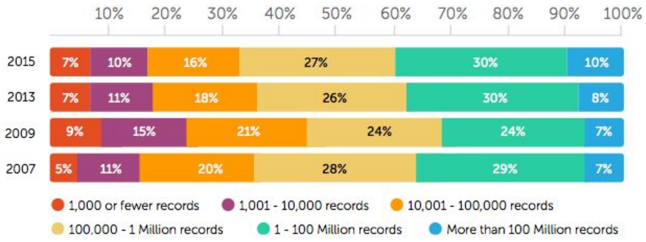
Figure 1

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

TYPICAL SIZE OF DATASETS

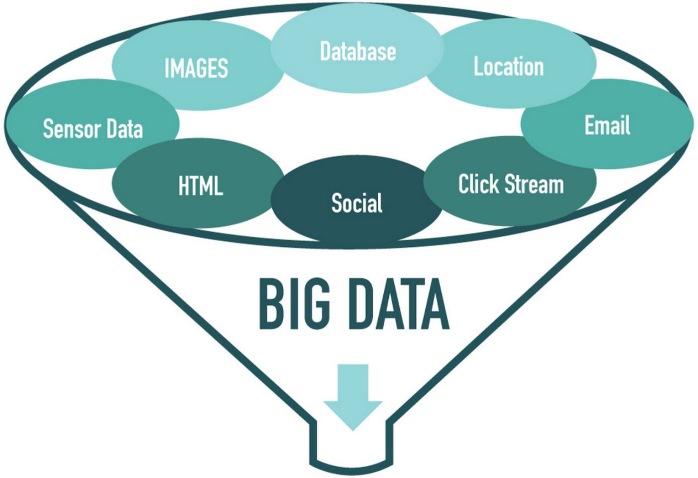


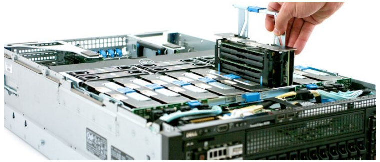
17 Rexer Analytics

Hadley Wickham
@hadleywickham

Following

"It takes a big man to admit his data is small" — @jcheng





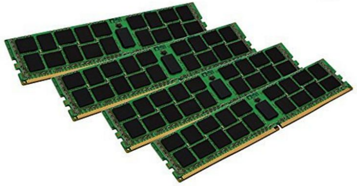
Kingston Technology Value RAM 128GB Kit (4x32GB) 2133MHz DDR4 ECC Reg CL15 (KVR21R15D4K4/128)

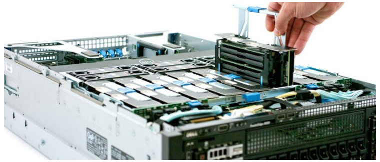
by Kingston Technology

Be the first to review this item

Was: ~~$743.99~~

Price: $743.96 & **FREE Shipping**. Details





Kingston Technology Value RAM 128GB Kit (4x32GB) 2133MHz DDR4 ECC Reg CL15 (KVR21R15D4K4/128)

by Kingston Technology

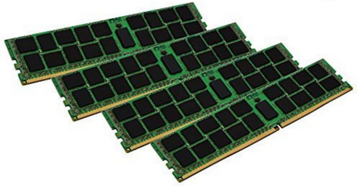
Be the first to review this item

Was: ~~$743.99~~

Price: $743.96 & **FREE Shipping**. Details

amazon web services

Model	vCPU	Mem (GiB)	
r3.8xlarge	32	244	(2015)
x1e.32xlarge	128	3,904	
u-12tb1.metal	448	12	**(TiB)**



Kingston Technology Value RAM 128GB Kit (4x32GB) 2133MHz DDR4 ECC Reg CL15 (KVR21R15D4K4/128)

by Kingston Technology

Be the first to review this item

Was: ~~$743.99~~

Price: $743.96 & **FREE Shipping**. Details

amazon web services

Model	vCPU	Mem (GiB)
r3.8xlarge	32	244
x1e.32xlarge	128	3,904
u-12tb1.metal	448	12 **(TiB)**

(2015)

Szilard [Deeper than Deep Learning]
@DataScienceLA

How much RAM do you have on the server/desktop/laptop you are most commonly using to train machine learning models?

<32GB	**50.7%**
32-127 GB	33.1%
128GB-1TB	14.1%
>1TB	2.1%

142 votes · Final results

Szilard [Deeper than Deep Learning]
@DataScienceLA

I wish my #machinelearning worked... ("both" is not a choice 😃) #bigdata #datascience #rstats #pydata cc @h2o @databricks @cloudera @kaggle

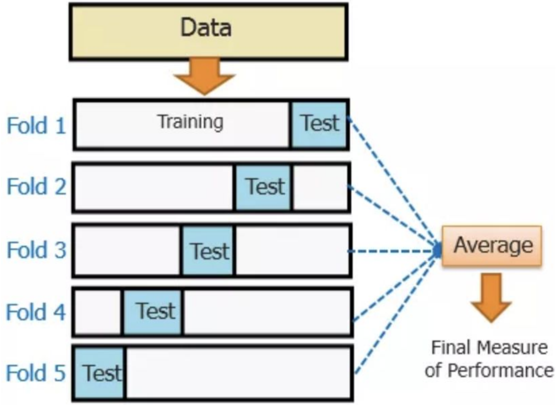
on 10x bigger data	9.6%
10x faster	**70.2%**
I don't care about either	20.2%

104 votes · Final results

8:48 AM · Aug 3, 2017 · Twitter Web Client


16
MAHLE


16


Data
Fold 1
Training
Test
Fold 2
Test
Fold 3
Test
Fold 4
Test
Fold 5
Test
Average
Final Measure
of Performance


Hyperparameter
tuning

szilard / GBM-perf

(2017-)

```
git clone https://github.com/szilard/GBM-perf.git
cd GBM-perf/cpu
sudo docker build -t gbmperf_cpu .
sudo docker run --rm gbmperf_cpu
```

Szilard
@DataScienceLA

Friday fun: what's your favorite gradient boosting machine (GBM) library?

58% xgboost

16% lightgbm

24% h2o

2% spark mllib

127 votes • Final results

3:21 PM - 11 May 2018

Szilard
@DataScienceLA

Friday fun: what's your favorite gradient boosting machine (GBM) library?

58% xgboost

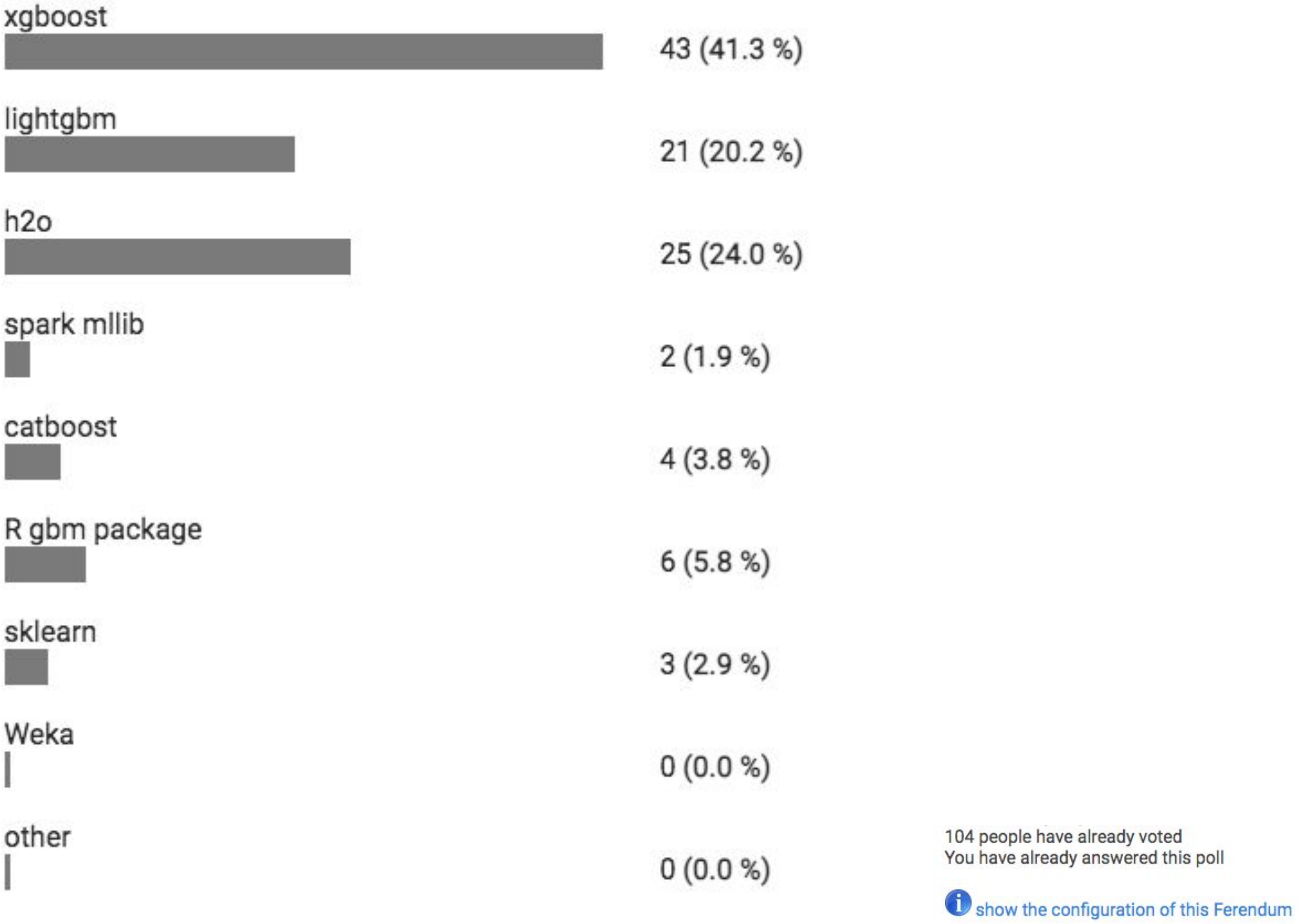
16% lightgbm

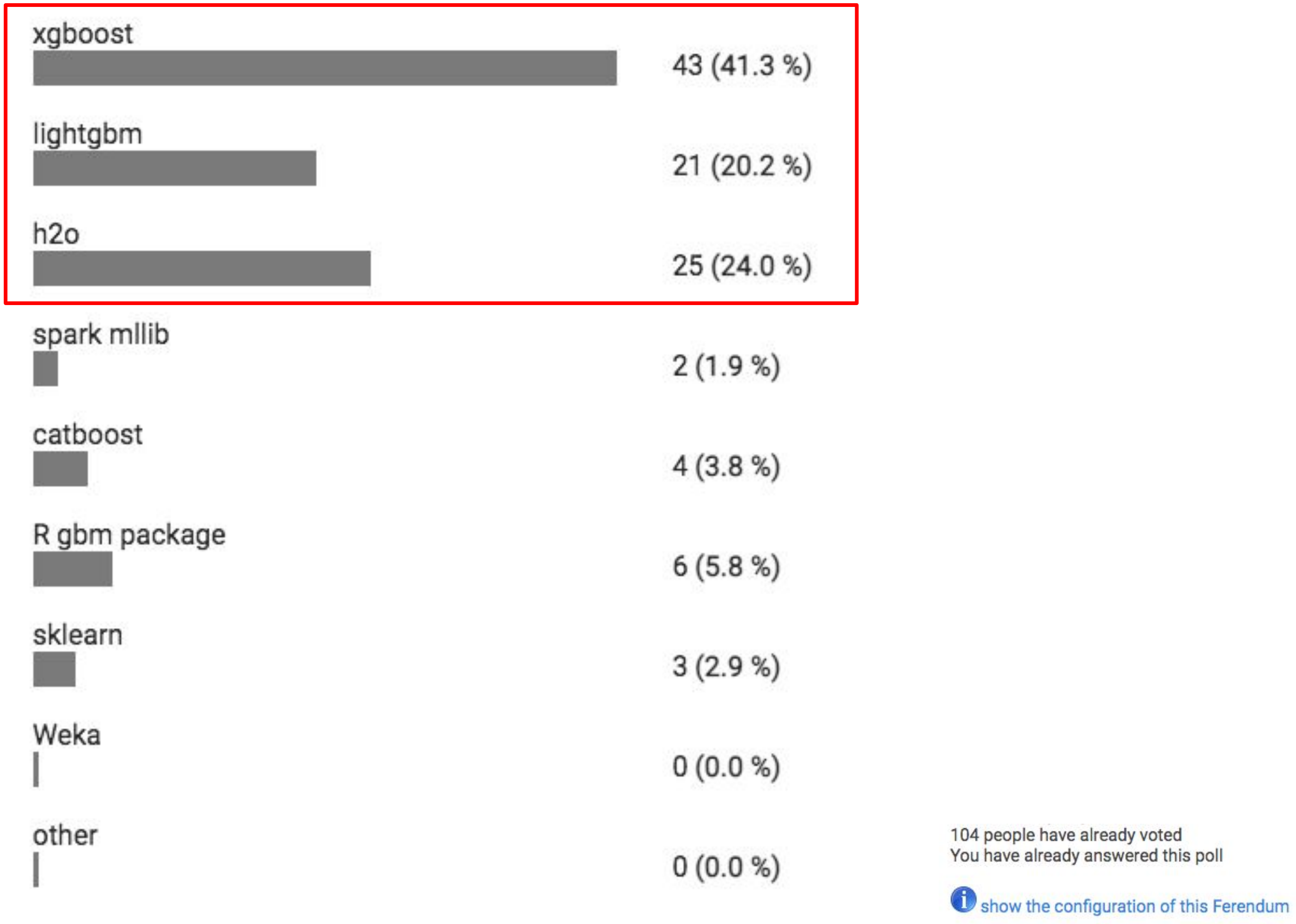
24% h2o

2% spark mllib ← **no-one is using this crap**

127 votes • Final results

3:21 PM - 11 May 2018


xgboost
43 (41.3 %)
lightgbm
21 (20.2 %)
h2o
25 (24.0 %)
spark mllib
2 (1.9 %)
catboost
4 (3.8 %)
R gbm package
6 (5.8 %)
sklearn
3 (2.9 %)
Weka
0 (0.0 %)
other
0 (0.0 %)
104 people have already voted
You have already answered this poll
show the configuration of this Ferendum


xgboost
43 (41.3 %)
lightgbm
21 (20.2 %)
h2o
25 (24.0 %)
spark mllib
2 (1.9 %)
catboost
4 (3.8 %)
R gbm package
6 (5.8 %)
sklearn
3 (2.9 %)
Weka
0 (0.0 %)
other
0 (0.0 %)
104 people have already voted
You have already answered this poll
show the configuration of this Ferendum

Szilard [Deeper than Deep Learning]
@DataScienceLA

What gradient boosting machine(GBM) library have you been using the most in 2020? (4 options, for others please reply to tweet)

xgboost	**53.5%**
lightgbm	26.7%
h2o	10.9%
catboost	8.9%

570 votes · Final results

10:59 AM · Sep 9, 2020 · Twitter Web App

r4.8xlarge (32 cores, but run on physical cores only/no hyperthreading) with software as of 2021-01-14:

Tool	Time[s] 100K	Time[s] 1M	Time[s] 10M	AUC 1M	AUC 10M
h2o	12	15	90	0.762	0.776
xgboost	**0.6**	**3.5**	40	0.748	0.754
lightgbm	2.6	4.2	**20**	0.765	0.792
catboost	3.8	10	80	0.734	0.735

r4.8xlarge (32 cores, but run on physical cores only/no hyperthreading) with software as of 2021-01-14:

Tool	Time[s] 100K	Time[s] 1M	Time[s] 10M	AUC 1M	AUC 10M
h2o	12	15	90	0.762	0.776
xgboost	**0.6**	**3.5**	40	0.748	0.754
lightgbm	2.6	4.2	**20**	0.765	0.792
catboost	3.8	10	80	0.734	0.735

p3.2xlarge (1 GPU, Tesla V100) with software as of 2021-01-15 (and CUDA 11.0):

Tool	Time[s] 100K	Time[s] 1M	Time[s] 10M	AUC 1M	AUC 10M
h2o xgboost	6.4	14	45	0.749	0.756
xgboost	3.6	6.5	**11**	0.748	0.756
lightgbm	7	10	42	0.767	0.792
catboost	**1.8**	**4.6**	37	0.732 ?!	0.736 ?!

100M records and RAM usage

CPU (m5.12xlarge):

Tool	time [s]	AUC	RAM train [GB]
h2o	520	0.775	8
xgboost	510	0.751	15
lightgbm	**310**	0.774	**5**
catboost	3360	0.723 ?!	140

UPDATE 2020-09-08:

Tool	time [s]	AUC	RAM train [GB]
catboost	930	0.736	50

100M records and RAM usage

GPU (Tesla V100):

Tool	time [s]	AUC	GPU mem [GB]	extra RAM [GB]
h2o xgboost	270	0.755	4	30
xgboost	**80**	0.756	6	**0**
lightgbm	400	0.774	3	6
catboost	crash (OOM)		>16	14

UPDATE 2020-09-08:
catboost still crashes out-of-memory

CPU (m5.12xlarge):

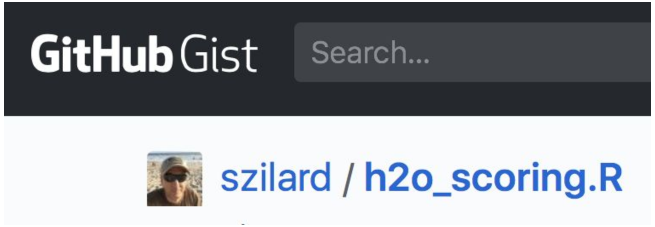
Tool	time [s]	AUC	RAM train [GB]
h2o	520	0.775	8
xgboost	510	0.751	15
lightgbm	**310**	0.774	**5**
catboost	3360	0.723 ?!	140

UPDATE 2020-09-08:

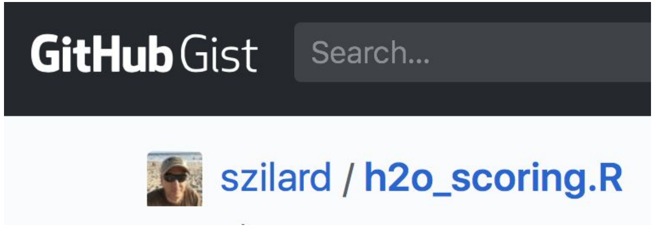
Tool	time [s]	AUC	RAM train [GB]
catboost	930	0.736	50

```
## exporting model for scoring

h2o.download_mojo(md_rf, path = "./h2o")
```

```
## building prediction service

# (need jetty-runner.jar ROOT.war from Steam)
java -jar jetty-runner.jar ROOT.war

curl -X POST --form mojo=@h2o_RF.zip --form jar=@h2o-genmodel.jar \
        localhost:8080/makewar > h2o_RF_MOJO.war
```



```
## exporting model for scoring

h2o.download_mojo(md_rf, path = "./h2o")
```



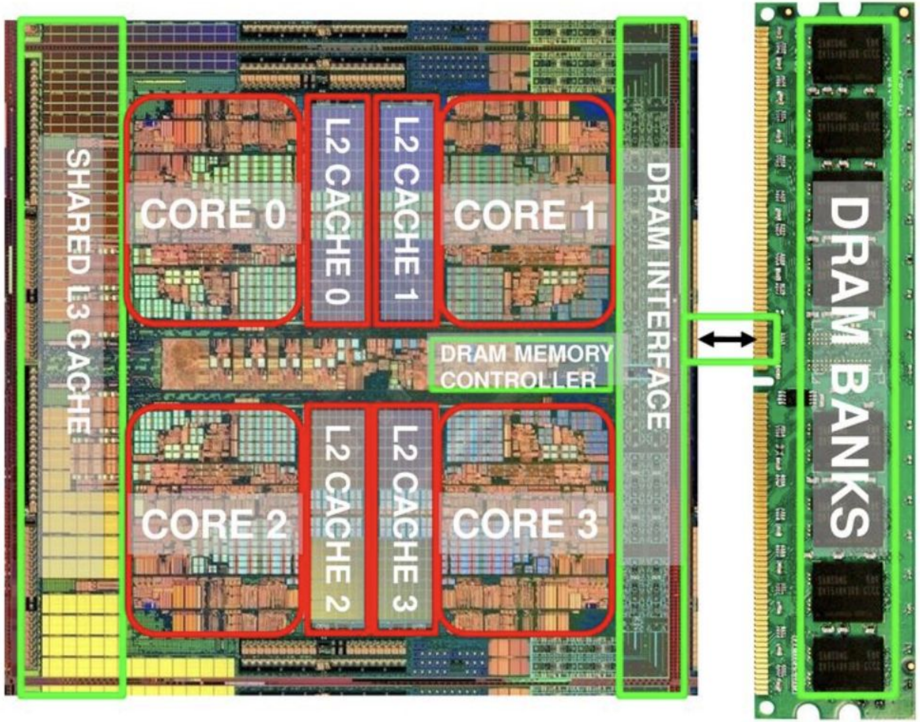

```
## building prediction service

# (need jetty-runner.jar ROOT.war from Steam)
java -jar jetty-runner.jar ROOT.war

curl -X POST --form mojo=@h2o_RF.zip --form jar=@h2o-genmodel.jar \
        localhost:8080/makewar > h2o_RF_MOJO.war
```

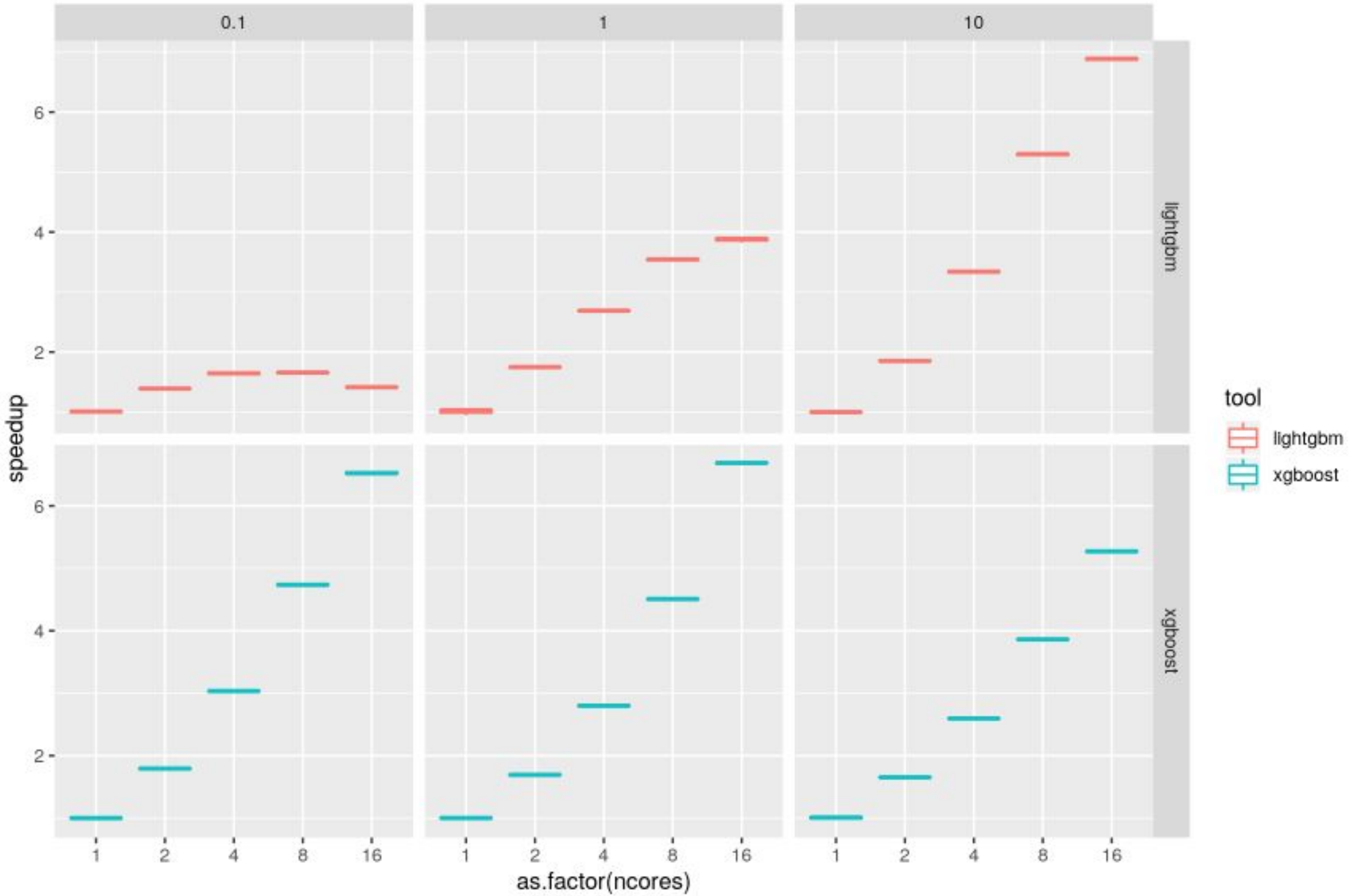
```
## run prediction service

java -jar jetty-runner.jar --port 20000 h2o_RF_MOJO.war
```

H2O.ai

```
## score via REST API

time curl "http://localhost:20000/predict?Month=c-8&DayofMonth=c-21&Da
# (fast scoring needs JVM to warm up with a few requests)
```

szilard / GBM-perf





0.1
1
10
lightgbm
xgboost
speedup
6
4
2
1
2
4
8
16
as.factor(ncores)
tool
lightgbm
xgboost

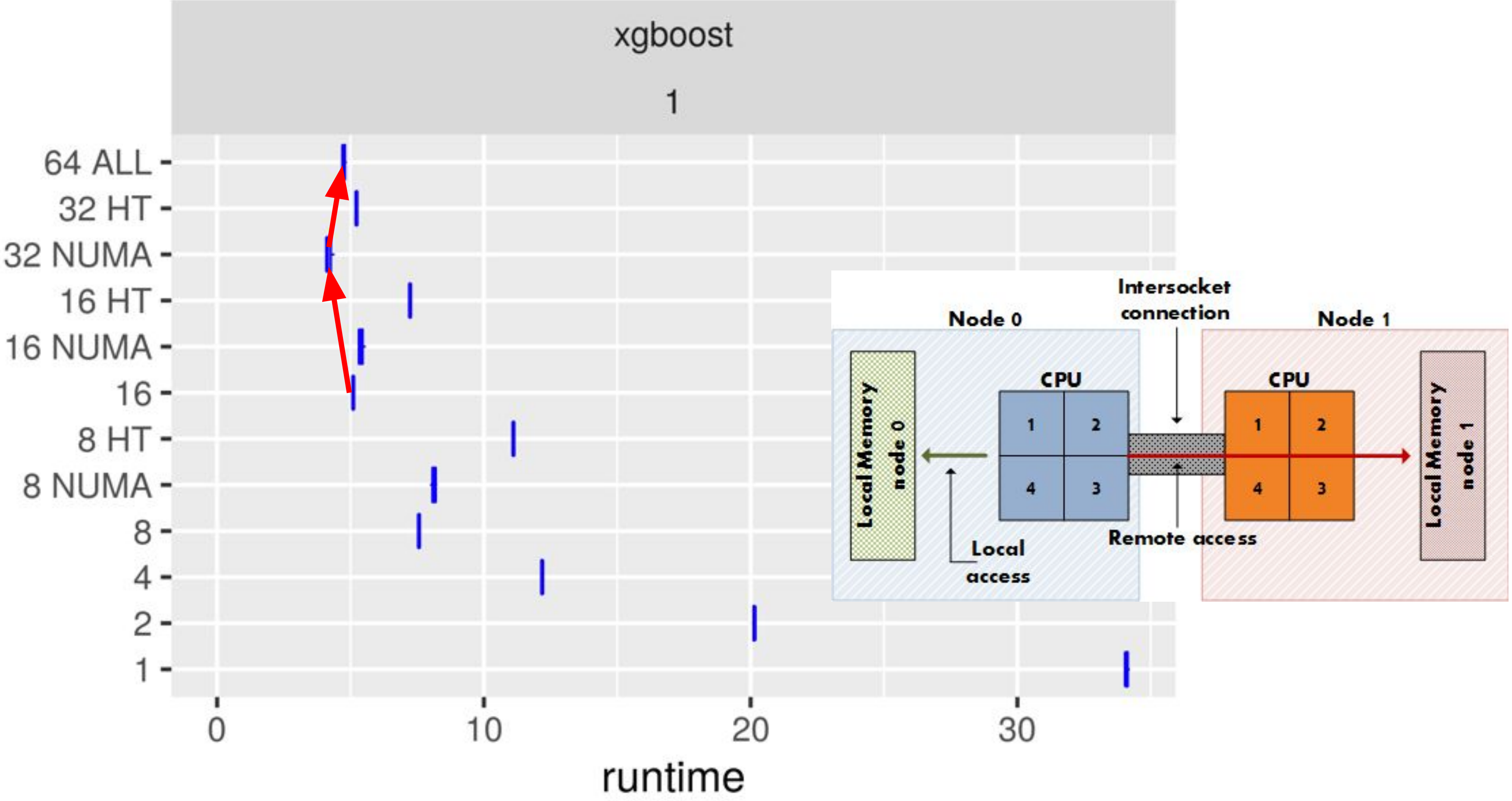
Speedup from 1 core to 16:

data size	h2o	xgboost	lightgbm	catboost
0.1M	3x	6.5x	1.5x	3.5x
1M	8x	6.5x	4x	6x
10M	24x	5x	7.5x	8x

Speedup from 1 core to 16:

data size	h2o	xgboost	lightgbm	catboost
0.1M	3x	6.5x	1.5x	3.5x
1M	8x	6.5x	4x	6x
10M	24x	5x	7.5x	8x

was **2.5x** before 2020


xgboost
1
64 ALL
32 HT
32 NUMA
16 HT
16 NUMA
16
8 HT
8 NUMA
8
4
2
1
0
10
20
30
runtime
Intersocket connection
Node 0
Node 1
CPU
CPU
Local Memory node 0
Local Memory node 1
1
2
4
3
1
2
4
3
Local access
Remote access

szilard / ml-x1

Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki

xgboost CPU % usage patterns #1

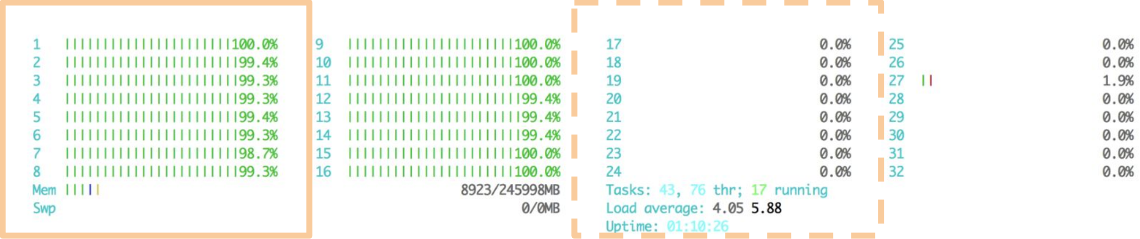
Open szilard opened this issue on Nov 6, 2016 · 4 comments

szilard commented on Nov 6, 2016 • edited

r3.8xlarge: CPU1 0-7 (and 16-23 hyperthread pairs), CPU2 8-15

CPU 1

```
taskset -c 0-7 Rscript xgb.R 8 &
taskset -c 8-15 Rscript xgb.R 8
```



szilard / ml-x1

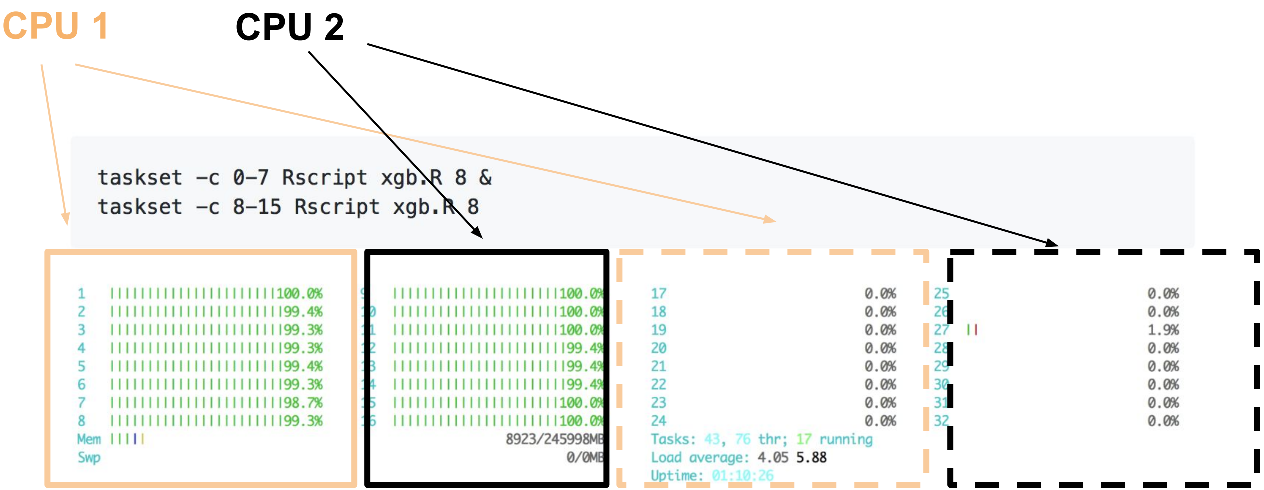
<> Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki

xgboost CPU % usage patterns #1

Open **szilard** opened this issue on Nov 6, 2016 · 4 comments

szilard commented on Nov 6, 2016 • edited

r3.8xlarge: CPU1 0-7 (and 16-23 hyperthread pairs), CPU2 8-15



szilard / ml-x1

Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki

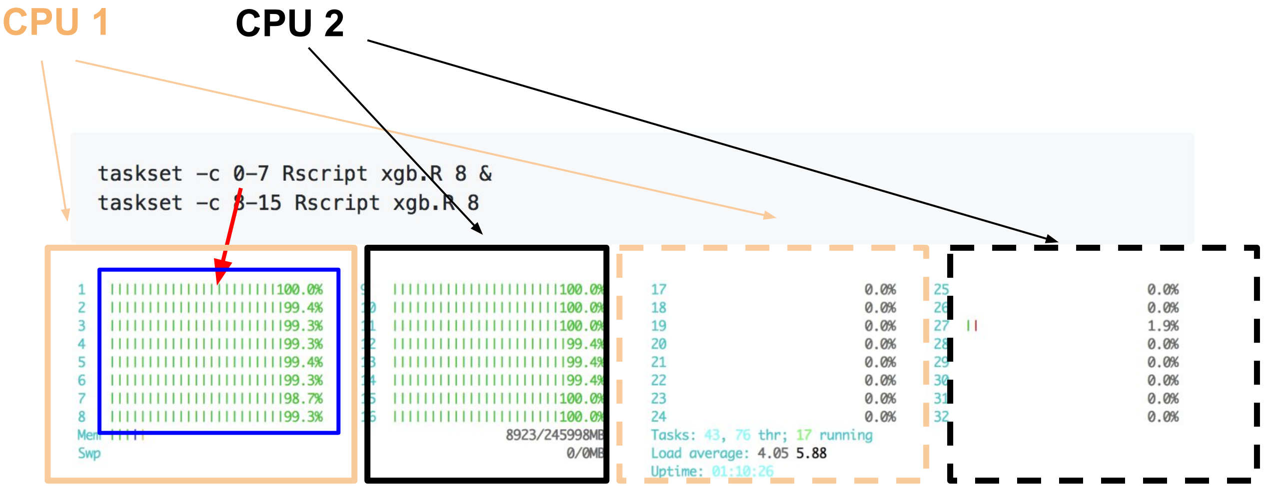
xgboost CPU % usage patterns #1

Open szilard opened this issue on Nov 6, 2016 · 4 comments

szilard commented on Nov 6, 2016 • edited

r3.8xlarge: CPU1 0-7 (and 16-23 hyperthread pairs), CPU2 8-15

```
taskset -c 0-7 Rscript xgb.R 8 &
taskset -c 8-15 Rscript xgb.R 8
```



szilard / ml-x1

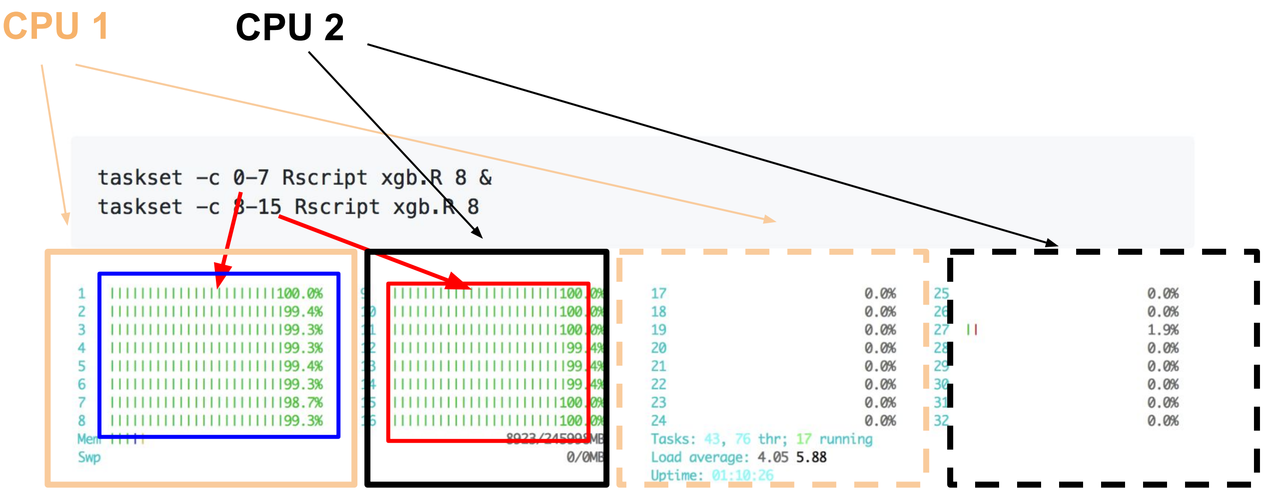
Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki

xgboost CPU % usage patterns #1

Open szilard opened this issue on Nov 6, 2016 · 4 comments

szilard commented on Nov 6, 2016 • edited

r3.8xlarge: CPU1 0-7 (and 16-23 hyperthread pairs), CPU2 8-15



Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are training machine learning models on CPU, how many CPU cores are you most commonly using?

1-4	36.5%
5-16	**40%**
17-64	13%
>64	10.4%

115 votes · Final results

12:08 PM · Sep 21, 2020 · Twitter Web App

szilard / GBM-perf

r4.8xlarge (32 cores, but run on physical cores only/no hyperthreading) with software as of 2021-01-14:

Tool	Time[s] 100K	Time[s] 1M	Time[s] 10M	AUC 1M	AUC 10M
h2o	12	15	90	0.762	0.776
xgboost	**0.6**	**3.5**	40	0.748	0.754
lightgbm	2.6	4.2	**20**	0.765	0.792
catboost	3.8	10	80	0.734	0.735

p3.2xlarge (1 GPU, Tesla V100) with software as of 2021-01-15 (and CUDA 11.0):

Tool	Time[s] 100K	Time[s] 1M	Time[s] 10M	AUC 1M	AUC 10M
h2o xgboost	6.4	14	45	0.749	0.756
xgboost	3.6	6.5	**11**	0.748	0.756
lightgbm	7	10	42	0.767	0.792
catboost	**1.8**	**4.6**	37	0.732 ?!	0.736 ?!

100M records and GPU memory usage

GPU (Tesla V100):

Tool	time [s]	AUC	GPU mem [GB]	extra RAM [GB]
h2o xgboost	270	0.755	4	30
xgboost	**80**	0.756	6	**0**
lightgbm	400	0.774	3	6
catboost	crash (OOM)		>16	14

UPDATE 2020-09-08:
catboost still crashes out-of-memory

Szilard @DataScienceLA · May 16

If you are using gradient boosting machines(GBM), are you running it (training) on GPUs or CPUs?

2018

7% GPU

93% CPU

55 votes • Final results

Szilard @DataScienceLA · May 16

If you are using gradient boosting machines(GBM), are you running it (training) on GPUs or CPUs?

2018

7% GPU

93% CPU

55 votes • Final results

Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you using (training) them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

86% CPU

14% GPU

69 votes • Final results

2019

11:39 AM - 4 May 2019

Szilard @DataScienceLA · May 16

If you are using gradient boosting machines(GBM), are you running it (training) on GPUs or CPUs?

2018

7%	GPU
93%	CPU

55 votes • Final results

Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you training them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

CPU	**61.5%**
GPU	38.5%

104 votes · Final results

2020

Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you using (training) them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

86%	CPU
14%	GPU

69 votes • Final results

2019

11:39 AM - 4 May 2019

Szilard @DataScienceLA · May 16

If you are using gradient boosting machines(GBM), are you running it (training) on GPUs or CPUs?

2018

7% GPU

93% CPU

55 votes • Final results

Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you using (training) them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

86% CPU

14% GPU

69 votes • Final results

2019

11:39 AM - 4 May 2019

Szilard [Deeper than Deep Learning]
@DataScienceLA

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you training them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

CPU	**61.5%**
GPU	38.5%

2020

104 votes · Final results

Szilard Pafka
Chief Data Scientist
3d • Edited •

POLL: If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you training them most often on the CPU or a GPU? #xgboost #lightgbm #h2oai #catboost #apachespark #mllib #sklearn

If you are using gradient boosting machines (GBM)/boosted trees (GBDT) are you training them most often on the CPU or a GPU?

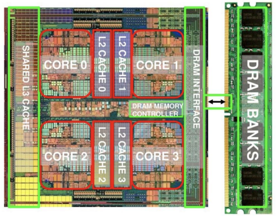
You can see how people vote. Learn more

CPU	78%
GPU	22%

32 votes • Poll closed

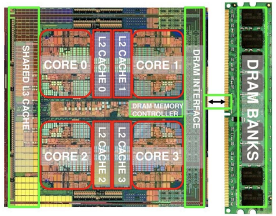
2020


AND THE
WINNER IS...

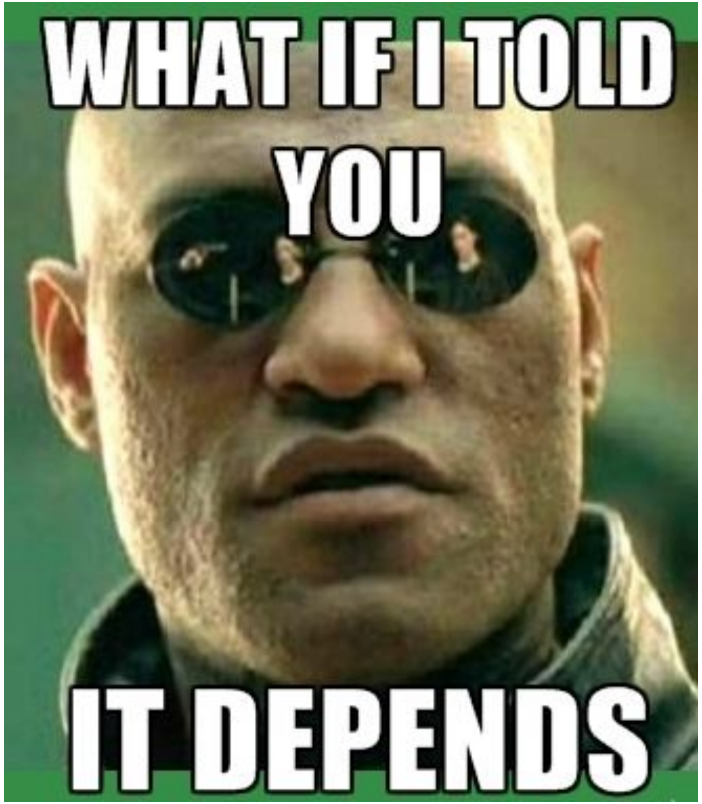

SHARED L3 CACHE
CORE 0
L2 CACHE 0
L2 CACHE 1
CORE 1
DRAM MEMORY CONTROLLER
CORE 2
L2 CACHE 2
L2 CACHE 3
CORE 3
DRAM INTERFACE
DRAM BANKS


TESLA


AND THE
WINNER IS...


SHARED L3 CACHE
CORE 0
L2 CACHE 0
L2 CACHE 1
CORE 1
DRAM MEMORY CONTROLLER
CORE 2
L2 CACHE 2
L2 CACHE 3
CORE 3
DRAM INTERFACE
DRAM BANKS


TESLA


WHAT IF I TOLD YOU
IT DEPENDS

	xgboost	lightgbm	h2o	catboost
easy R install	CRAN	CRAN	java+CRAN	devtools+pre-binary
maintained	yes	yes	yes	yes
preprocessing	1-hot	1-hot/categ int	not needed	categ internal
new cats scoring	no	no	yes	no
early stopping	yes	yes	yes	yes
speed (CPU)	fastest	fastest	slow (small data)	slow
GPU supported	yes	yes	via xgboost	yes, but mem usage
speed GPU	fastest	slow	indirectly/slow	slow on larger data
REST scoring	no	no	yes	no
other algos	RF	RF	RF/GLM/NN	none
best for	Kaggle	Kaggle	prod/real-time	Kaggle

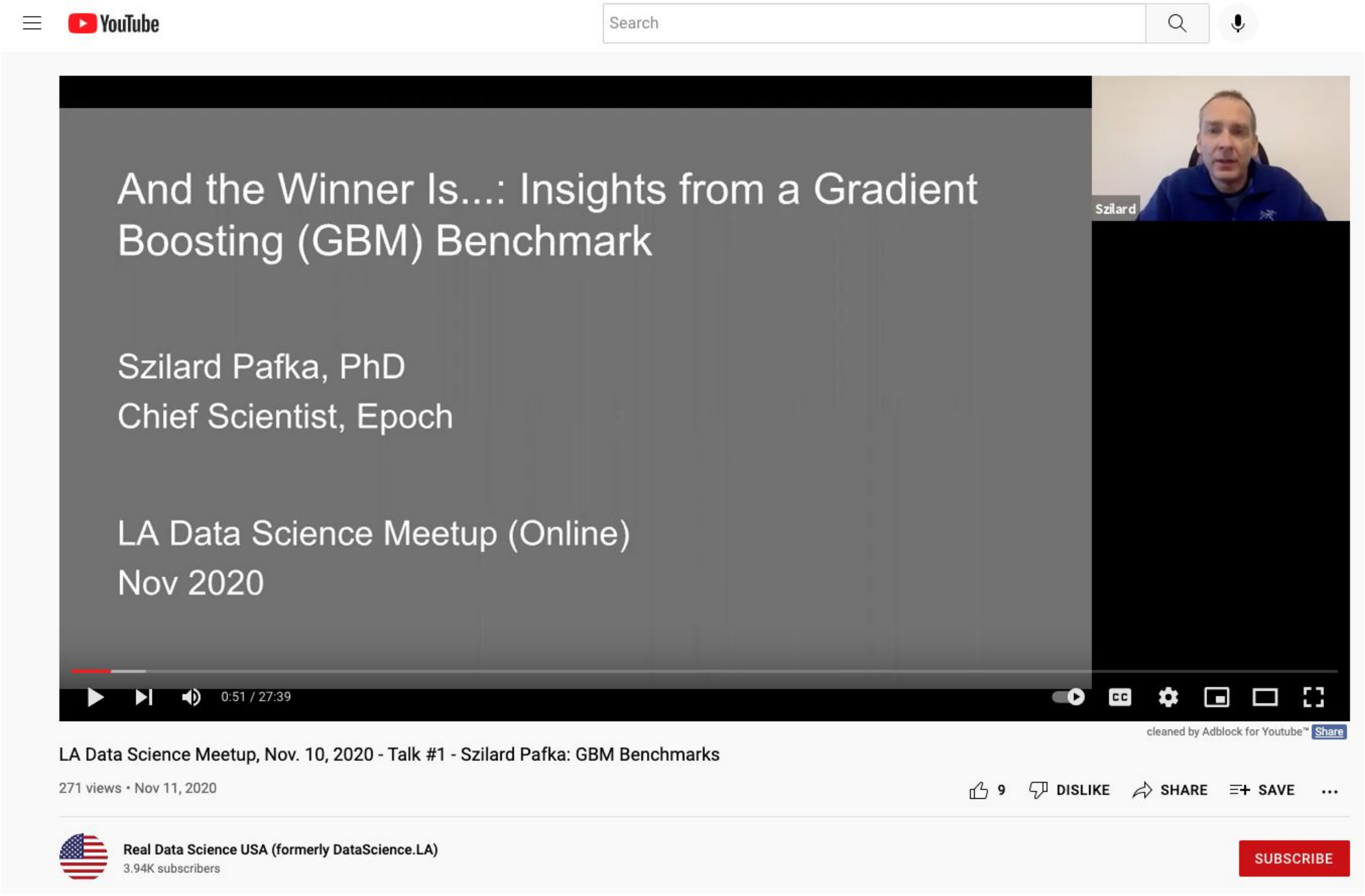

TABULAR DATA YOU HAVE
GBM USE YOU MUST

spafka@gmail.com

@DataScienceLA

linkedin.com/in/szilard

github.com/szilard


YouTube
Search
And the Winner Is...: Insights from a Gradient Boosting (GBM) Benchmark
Szilard Pafka, PhD
Chief Scientist, Epoch
LA Data Science Meetup (Online)
Nov 2020
Szilard
0:51 / 27:39
cleaned by Adblock for Youtube™ Share
LA Data Science Meetup, Nov. 10, 2020 - Talk #1 - Szilard Pafka: GBM Benchmarks
271 views • Nov 11, 2020
9
DISLIKE
SHARE
SAVE
Real Data Science USA (formerly DataScience.LA)
3.94K subscribers
SUBSCRIBE

More:

szilard / benchm-ml Star 1,203

szilard / teach-data-science-UCLA-master-appl-stats

szilard / teach-ML-CEU-master-bizanalytics

szilard / GBM-intro

szilard / GBM-workshop

szilard / ML-scoring

szilard / GBM-perf

szilard / GBM-tune

szilard / GBM-multicore

GitHubGist Search...

szilard / h2o_scoring.R

szilard / ML_with_H2O.R

Bojan Tunguz @tunguz · Mar 28

When you find out your intern used NNs on **tabular** data.

27 40 552