

H2O Around the World

London Kaggle Meetup



Strata Hadoop



Chelsea Football Club



PyData
Amsterdam



useR! 2016
Stanford



satRdays
Budapest

H2O Around the World



Paris Machine Learning Meetup (last week)

Szilard Pafka – Chief Data Scientist at Epoch

- Sziland's talks / blog posts about H2O:
 - ML Benchmark
 - Intro to ML with H2O
 - H2O Scoring
 - Tweets



Also @h2oai on monster 2TB RAM 128 cores
EC2 X1 #bigdata #machinelearning
#datascience twitter.com/DataScienceLA/ ...

Index	Model	Accuracy (%)	Training Time (hrs)	Testing Time (hrs)	RAM (GB)	Cores	Notes
1	Random Forest	58.1%	33	11	65	11	
2	Random Forest	48.1%	34	11	62	11	
3	Random Forest	48.1%	35	11	60	11	
4	Random Forest	48.1%	36	11	60	11	
5	Random Forest	43.9%	37	11	68	11	
6	Random Forest	43.9%	38	11	69	11	
7	Random Forest	45.2%	39	11	70	11	
8	Random Forest	43.9%	40	11	72	11	
9	Random Forest	43.6%	41	11	74	11	
10	Random Forest	42.7%	42	11	73	11	
11	Random Forest	54.1%	43	11	75	11	
12	Random Forest	46.5%	44	11	76	11	
13	Random Forest	44.9%	45	11	77	11	
14	Random Forest	47.1%	46	11	78	11	
15	Random Forest	44.2%	47	11	79	11	
16	Random Forest	47.1%	48	11	80	11	
17	Random Forest	48.7%	49	11	81	11	
18	Random Forest	43.6%	50	11	82	11	
19	Random Forest	42.3%	51	11	83	11	
20	Random Forest	43.3%	52	11	84	11	
21	Random Forest	44.6%	53	11	85	11	
22	Random Forest	45.2%	54	11	86	11	
23	Random Forest	41.9%	55	11	87	11	
24	Random Forest	44.6%	56	11	88	11	
25	Random Forest	44.6%	57	11	89	11	
26	Random Forest	43.7%	58	11	90	11	
27	Random Forest	45.5%	59	11	91	11	
28	Random Forest	45.5%	60	11	92	11	
29	Random Forest	43.4%	61	11	93	11	
30	Random Forest	41.4%	62	11	94	11	
31	Random Forest	43.3%	63	11	95	11	
32	Random Forest	40.6%	64	11	96	11	

23814/19675286
0/846
Tasks: 44, S25 thr, 1870 kthr; 3 running
Load average: 70.10 33.24
Uptime: 00:36:08

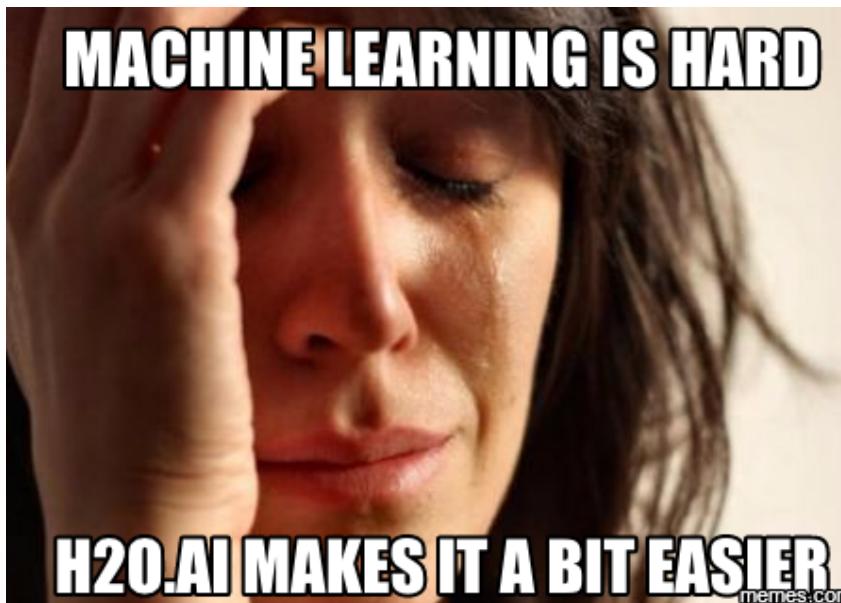
RETWEETS 10 LIKES 28

5:55 PM - 24 Jun 2016

3

Szilard Pafka – Why H2O?

- Szilard's Summary Slide



H2O Customers

Brendan Herger
Data Scientist
Capital One

"We evaluated a large number of hard and soft metrics. H2O just scored really well with all of these areas, particularly relative to a lot of the machine learning frameworks that are available at the moment."

Pawan Divakarla
Data and Analytics Business Leader
Progressive Insurance

"H2O is like an enabler in how people are thinking about the data and how they want to use the data, and that's come in very handy for some of our data scientists and advanced analytic users. Now they have the right toolset that they can use on the data."

Edward Agarwala
Data Scientist
Progressive Insurance

"H2O has gradient memory performance, even on single nodes. It will utilize all of the cores, even on single nodes. It has all of the latest and greatest algorithms, including statistic GBM, including random forest, including GLM, and it has things like the weights has the different loss functions"

Prateem Mandal
Technical Lead Architect
MarketShare

Check out the videos - www.h2o.ai

H2O Community Support

Google forum – h2osteam

Groups

New Topic

Mark all as read

Filters

Groups

My groups

Home

Starred

Favourites

Click on a group's star icon to add it to your favourites

Recently viewed

- H2O Open Sour...
- packrat-discuss
- devtools
- Caffe Users
- ggplot2

Recent searches

- spark streaming (i...)
- chord diagram gg...
- chord diagram (in ...)
- Pictaculous API (i...)
- pictaculous (in ma...

Recently posted to

- H2O Open Sour...

Privacy - Terms of Service

H2O Open Source Scalable Machine Learning - h2osteam Shared publicly 30 of 2055 topics (99+ unread) Join group G+ Tags About

All Posts

When is Steam going to be released? 3 days ago in Steam

H2O Python Modules 3 days ago in H2O

H2O Installation 3 days ago in H2O

PySparkling launch problem with Python 2.6 or older 3 days ago in Python

Predicted Values 3 days ago in H2O

Combining holdout predictions, while keep_cross_validation_predictions parameter is active in Python 3 days ago in Python

You can continue to use this google group, however we'd like to encourage everyone to shift their energy toward building community.h2o.ai. We also welcome any questions or feedback you may have about the transition or the new community website.

how to use API to export model (1) By tangb...@gmail.com - 1 post - 2 views 06:03

How can I use the decode half of a trained autoencoder? (6) By j...@sharpe.com - 6 posts - 14 views 05:31

community.h2o.ai

Find posts, topics, and users...

Create

Spaces

Ask a question

Post an idea

Create an article

Algorithms

Announcements

Artificial Intelligence

Deep Water

Demos

H2O

Java

Machine Learning

Python

R

Source Code

Sparkling Water

Steam

Tools

Troubleshooting

H2O Dev JIRA | Summary | Learn | H2O | Sparkling Water

H2O.ai

All Posts

When is Steam going to be released? 3 days ago in Steam

H2O Python Modules 3 days ago in H2O

H2O Installation 3 days ago in H2O

PySparkling launch problem with Python 2.6 or older 3 days ago in Python

Predicted Values 3 days ago in H2O

Combining holdout predictions, while keep_cross_validation_predictions parameter is active in Python 3 days ago in Python

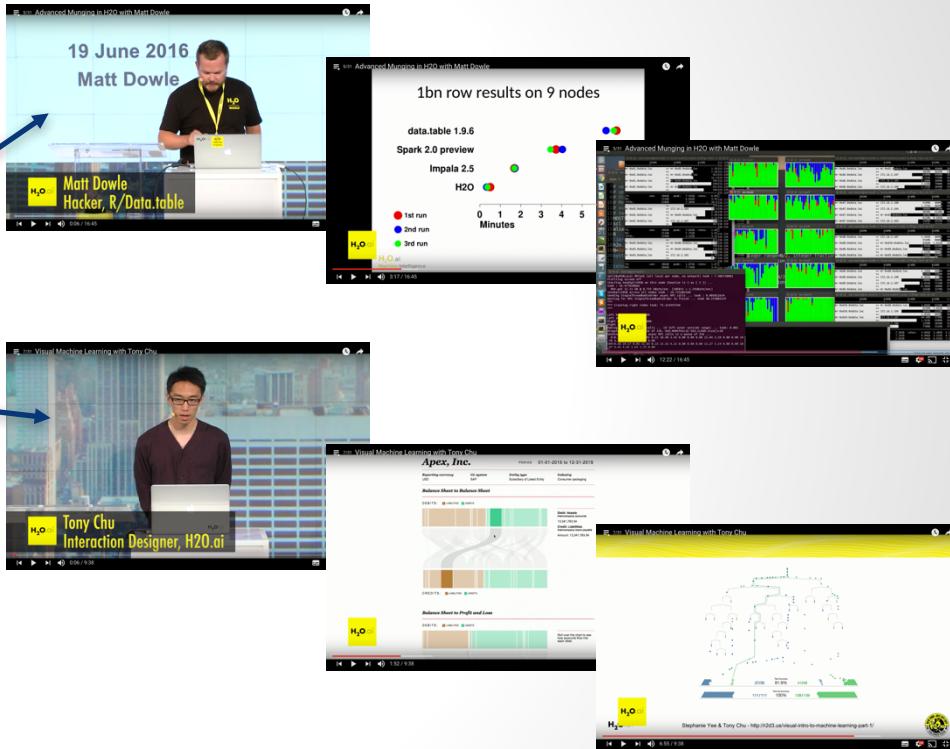
Sparkling Water Release 08/30

We are happy to announce that Sparkling Water 2.0 release is almost here. On September 1, 2016 we will release Sparkling Water 2.0. Download info is coming soon.

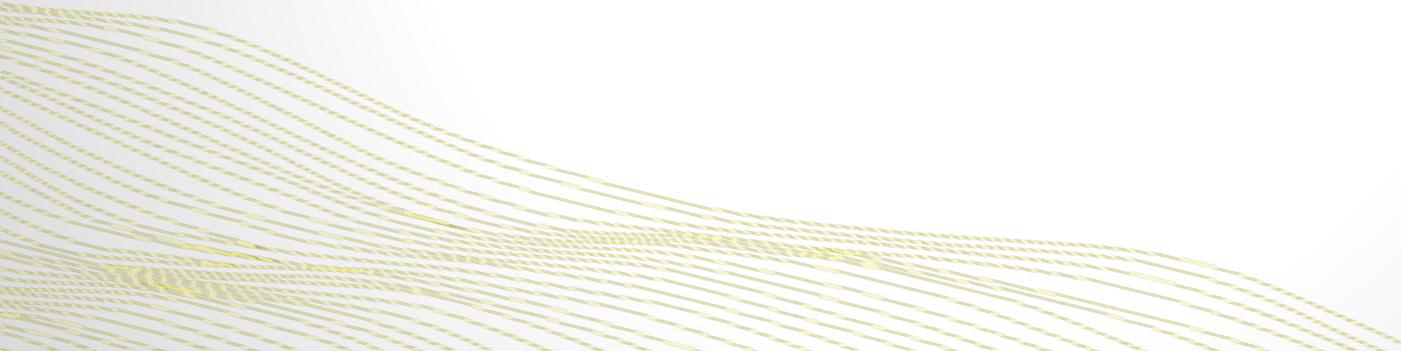
H2O is Evolving

- H2O Open Tour NYC
YouTube Playlist

- Advanced data munging
- Visual ML
- Deep Water (2nd talk)
- Sparkling Water (3rd talk)
- Steam
 - New data science platform



Why Deep Water?



H2O Deep Learning in Action

116M rows, 6GB CSV file
800+ predictors (numeric + categorical)

airlines_all_selected_cols.hex

Actions: View Data Split... Build Model... Predict Download Export

Rows	Columns	Compressed Size
116695259	12	2GB



Job

Run Time 00:00:36.712

Remaining Time 00:00:17.188

Type Model

Key Q_ddeeplearning-dd2f42f7-81f7-42e8-9d98-e34437309828

Description DeepLearning

Status RUNNING

Progress 69%

Iterations: 12. Epochs: 0.628821. Speed: 2,243,735 samples/sec. Estimated time left: 21.849 sec

Actions View Cancel Job

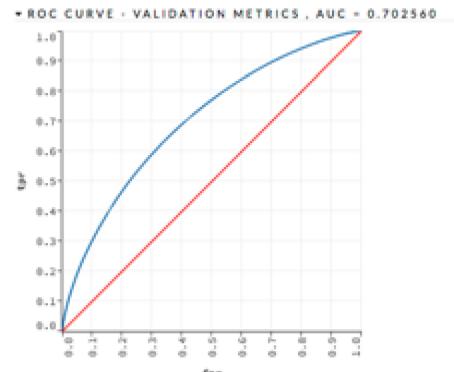
model trained in <1 min:
2M+ samples/second

* OUTPUT - STATUS OF NEURON LAYERS (PREDICTING ISDELAYED, 2-CLASS CLASSIFICATION, BERNoulli DISTRIBUTION, CROSSENTROPY LOSS, 17,462 WEIGHTS/BIASES, 221.3 KB, 106,585,365 TRAINING SAMPLES, MINI-BATCH SIZE 1)

layer	units	type	dropout	l1	l2	mean_rate	rate_rms	momentum	mean_weight	weight_rms	mean_bias	bias_rms
1	897	Input	0									
2	20	Rectifier	0	0	0	0.0493	0.2028	0	-0.0021	0.2111	-0.9139	1.0036
3	20	Rectifier	0	0	0	0.0197	0.0227	0	-0.1033	0.5362	-1.3968	1.5259
4	20	Rectifier	0	0	0	0.0517	0.0446	0	-0.1575	0.3068	-0.0846	0.0046
5	20	Rectifier	0	0	0	0.0761	0.0844	0	-0.0374	0.2275	-0.2647	0.2481
6	2	Softmax	0	0	0	0.0161	0.0083	0	0.0741	0.7268	0.4269	0.2056

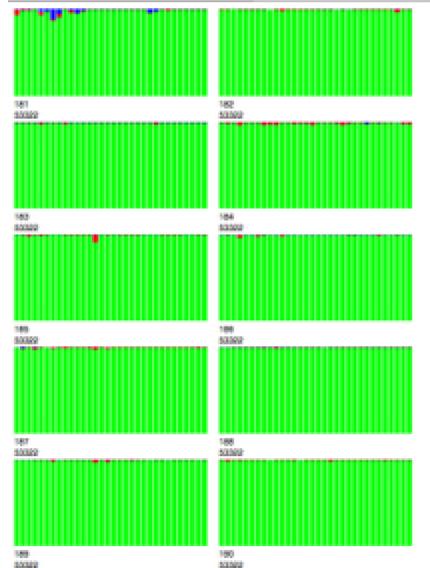
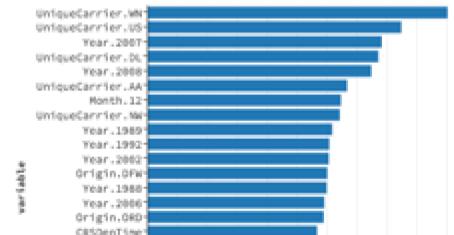
H₂O.ai

Deep Learning Model



Threshold: Choose... Criterion: Choose...

VARIABLE IMPORTANCES



Legend

Each bar represents one CPU.

Blue: idle time

Green: user time

Red: system time

White: other time (e.g. Vo)

10 nodes: all 320 cores busy



real-time, interactive
model inspection in Flow

H2O Deep Learning Community Quotes

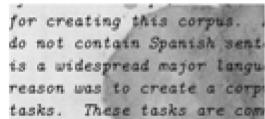
CIFAR-10 Competition
Winners: Interviews with Dr.
Ben Graham, Phil Culliton, &
Zygmunt Zajac
Triskelion | 01.02.2015

[READ MORE](#)

“I did really like H2O’s deep learning implementation in R, though - the interface was great, the back end extremely easy to understand, and it was scalable and flexible. Definitely a tool I’ll be going back to.”

Kaggle challenge
2nd place winner
Colin Priest

[READ MORE](#)



Completed • Knowledge • 161 teams

Denoising Dirty Documents

Mon 1 Jun 2015 – Mon 5 Oct 2015 (3 months ago)

“For my final competition submission I used an ensemble of models, including 3 deep learning models built with R and h2o.”

H2O Deep Learning Community Quotes



Arno Candel @ArnoCandel · Jun 29

Great use of H2O #DeepLearning for 3rd place at #Kaggle Homesite challenge! github.com/mpearmain/home... (see R scripts)

```
dl.model <- h2o.deeplearning(  
  # data specifications  
  x = xrange, y = max(xrange)+1, training_frame = x0.hex,  
  autoencoder = FALSE,  
  # network structure: activation and geometry  
  activation = "RectifierWithDropout",  
  hidden = c(size1, size2), epochs = 25,  
  input_dropout_ratio = 0.05, hidden_dropout_ratios = c(0.05, 0.02),  
  # parameters of the optimization process  
  rho = 0.99, epsilon = 1e-08, rate = 0.005,  
  rate_annealing = 1e-06, rate_decay = rate_dec, momentum_start = 0.5,  
  l1 = 0, l2 = 0, loss = c("CrossEntropy")  
)
```

18 33 ...



Completed • \$20,000 • 1,764 teams

Homesite Quote Conversion

Mon 9 Nov 2015 – Mon 8 Feb 2016 (5 months ago)

Dashboard

This competition has completed. This leaderboard reflects the final standings.

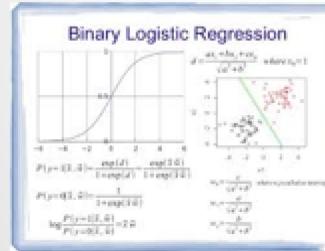
#	Rank	Team Name	model uploaded	in the money	Score	Entries	Last Sub
1	—	KazAnova Faron clobber	‡	*	0.97024	350	Mon,
2	—	Frenchies	‡	*	0.97019	241	Mon,
3	11	New Model Army CAD & QuY	‡	*	0.97002	234	Mon,
		• Konrad Banachewicz • Mike Pearmain • Charles-Abner DADI • quentin.y					

Deep Water: Best Open-Source Deep Learning

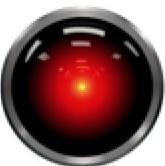
Enterprise Deep Learning for Business Transformation

Deep Water = THE Deep Learning Platform	H2O integrates the top open-source DL tools	
Native GPU support	  is up to 100x faster than	
Enterprise Ready	Easy to train and deploy, interactive, scalable, etc. Flow, R, Python, Spark/Scala, Java, REST, POJO, Steam	
New Big Data Use Cases (previously impossible or difficult in H2O)	Image - social media, manufacturing, healthcare, ... Video - UX/UI, security, automotive, social media, ... Sound - automotive, security, call centers, healthcare, ... Text - NLP, sentiment, security, finance, fraud, ... Time Series - security, IoT, finance, e-commerce, ...	

More Data is Better! — Images, Video, Text, Logs, Streams, ...



Example: Fraud Prediction



Google



Today



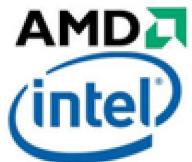
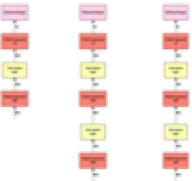
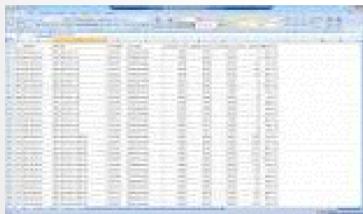
Tomorrow

Better Data — Better Models — Better Results

Deep Water opens the Floodgates for state-of-the-art Deep Learning

H2O Deep Learning: simple multi-layer neural networks

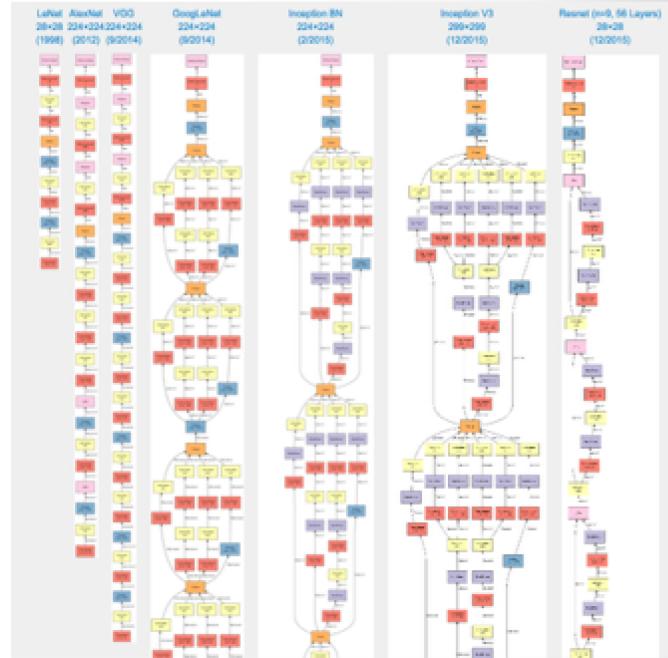
1-5 layers
MBs/GBs of data



Limited to business analytics,
statistical models (CSV data)

Deep Water: deep complex networks

5-1000 layers
GBs/TBs of data



Large networks for big data
(e.g. image 1000x1000x3 -> 3m inputs)

Current Contributors (more H2O.ai folks joining soon)



Fabrizio Milo



Cyprien Noel



Qiang Kou



Arno Candel



Caffe



This repository

mxnet



Deep Water Demos

- **H2O + mxnet**
 - Dataset:
 - Cat / Dog / Mouse
 - H2O Python interface
 - mxnet GPU backend
 - Train a LeNet (CNN) model
 - Explore model in Flow
- **H2O + TensorFlow**
 - Dataset:
 - MNIST hand-written digits
 - Sparkling Water + Jupyter Notebook
 - Convert TensorFlow model into H2O
 - Explore models in Flow

For Online Audience

- **H2O + mxnet**
 - bit.ly/h2o_paris_1
 - **demo_03_mxnet**
- **H2O + TensorFlow**
 - bit.ly/h2o_paris_1
 - **demo_04_tensorflow**