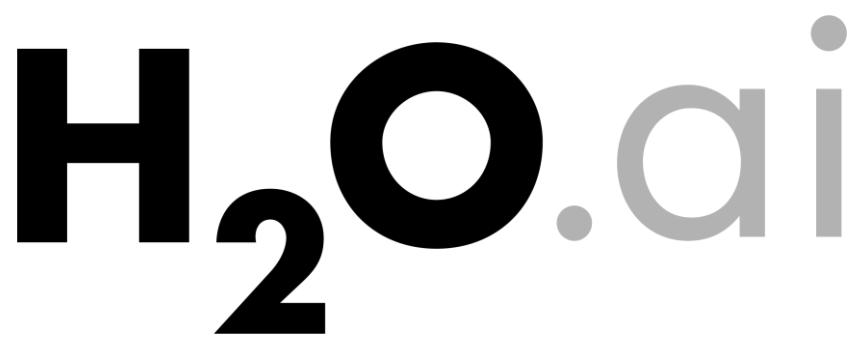


H₂O Deep Water

Making Deep Learning Accessible to Everyone



Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai
[@matlabulous](https://twitter.com/matlabulous)

ODSC London Meetup
31st January, 2017

About Me

- Civil (Water) Engineer
 - 2010 – 2015
 - Consultant (UK)
 - Utilities
 - Asset Management
 - Constrained Optimization
 - Industrial PhD (UK)
 - Infrastructure Design Optimization
 - Machine Learning + Water Engineering
 - Discovered H₂O in 2014
- Data Scientist
 - From 2015
 - Virgin Media (UK)
 - Domino Data Lab (Silicon Valley)
 - H₂O.ai (Silicon Valley)

Agenda

- About H₂O.ai
 - Company
 - Machine Learning Platform
- Deep Learning Tools
 - TensorFlow, MXNet, Caffe, H₂O
- Deep Water
 - Motivation
 - Benefits
 - Interface
 - Learning Resources
- Conclusions



About H₂O.ai

Company Overview

| | |
|---------------------|---|
| Founded | 2011 Venture-backed, debuted in 2012 |
| Products | <ul style="list-style-type: none">• H₂O Open Source In-Memory AI Prediction Engine• Sparkling Water• Steam |
| Mission | Operationalize Data Science, and provide a platform for users to build beautiful data products |
| Team | 70 employees <ul style="list-style-type: none">• Distributed Systems Engineers doing Machine Learning• World-class visualization designers |
| Headquarters | Mountain View, CA |



Scientific Advisory Council



Dr. Trevor Hastie

- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*



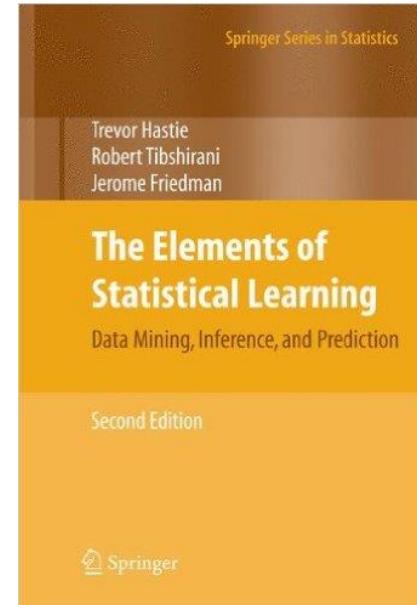
Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*



Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*



JANUARY 11, 2017

AI 100: The Artificial Intelligence Startups Redefining Industries



100 STARTUPS USING ARTIFICIAL INTELLIGENCE TO TRANSFORM INDUSTRIES

CONVERSATIONAL AI/ BOTS



VISION



AUTO



ROBOTICS



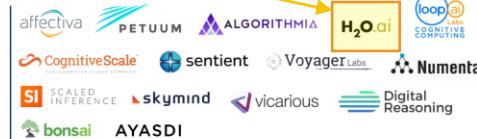
CYBERSECURITY



BUSINESS INTELLIGENCE & ANALYTICS



CORE AI



AD, SALES, CRM



HEALTHCARE



TEXT ANALYSIS/ GENERATION



IOT/IOT



COMMERCE



FINTECH & INSURANCE



OTHER



<https://www.cbinsights.com/blog/artificial-intelligence-top-startups/>

H₂O Machine Learning Platform

Algorithms Overview

Supervised Learning

Statistical Analysis

- **Generalized Linear Models:** Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**

Ensembles

- **Distributed Random Forest:** Classification or regression models
- **Gradient Boosting Machine:** Produces an ensemble of decision trees with increasing refined approximations

Deep Neural Networks

- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

Unsupervised Learning

Clustering

- **K-means:** Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

Dimensionality Reduction

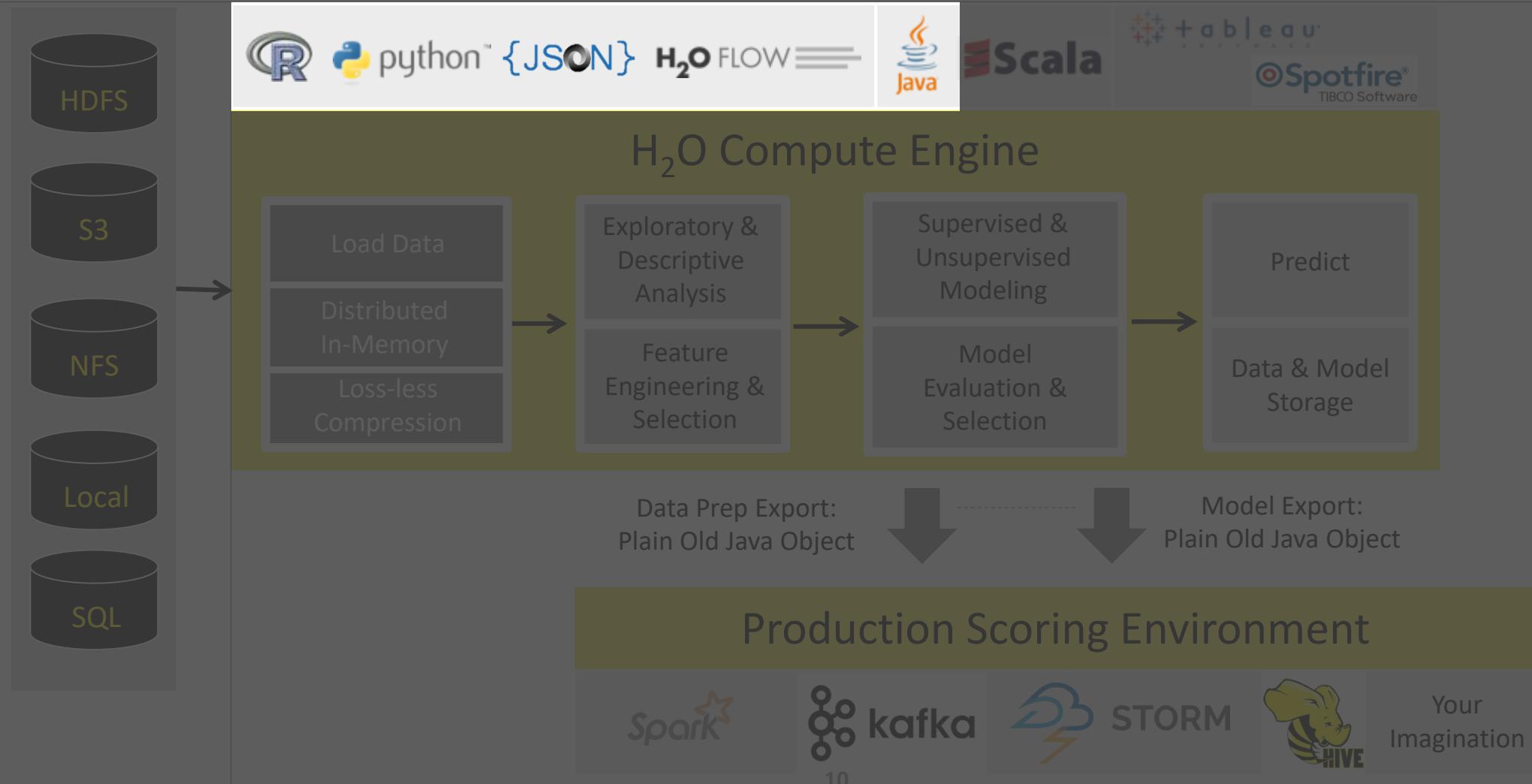
- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Anomaly Detection

- **Autoencoders:** Find outliers using a nonlinear dimensionality reduction using deep learning

High Level Architecture

Flow (Web), R, Python API
Java for computation





Flow ▾ Cell ▾ Data ▾

Model ▾ Score ▾ Admin ▾ Help ▾

Iris Demo



CS

Expression...

- Aggregator...
- Deep Learning...
- Distributed Random Forest...
- Gradient Boosting Machine... 🕒
- Generalized Linear Modeling...
- Generalized Low Rank Modeling...
- K-means...
- Naive Bayes...
- Principal Components Analysis...

- List All Models
- List Grid Search Results
- Import Model...
- Export Model...

H₂O Flow (Web) Interface



Connections: 0 H₂O

Iris Demo



Expression...

CS buildModel "drf"

192ms

Build a Model

Select an algorithm: **Distributed Random Forest** ▾

PARAMETERS

GRID?

| | | |
|-------------------------|-----------------|--|
| <i>model_id</i> | DRF-Iris-Demo | Destination id for this model; auto-generated if not specified. |
| <i>training_frame</i> | iris_from_csv ▾ | Id of the training data frame (Not required, to allow initial validation of model parameters). |
| <i>validation_frame</i> | (Choose...) | Id of the validation data frame. |
| <i>nfolds</i> | 0 | Number of folds for N-fold cross-validation (0 to disable or >= 2). |
| <i>response_column</i> | Species | Response variable column. |
| <i>ignored_columns</i> | Search... | |

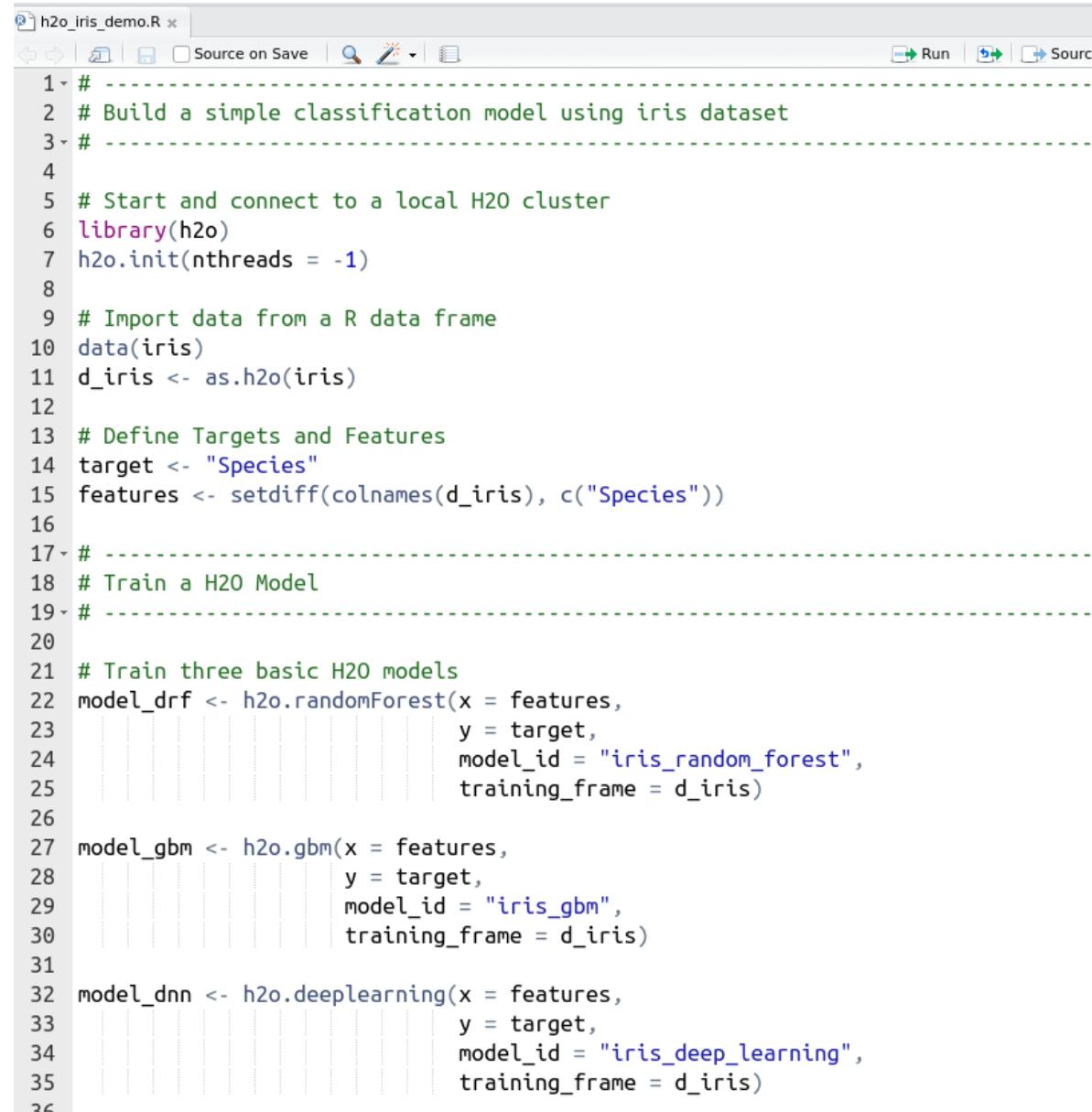
Showing page 1 of 1.

| | |
|---------------------------------------|---------|
| <input type="checkbox"/> Sepal.Length | REAL |
| <input type="checkbox"/> Sepal.Width | REAL |
| <input type="checkbox"/> Petal.Length | REAL |
| <input type="checkbox"/> Petal.Width | REAL |
| <input type="checkbox"/> Species | ENUM(3) |

H₂O Flow (Web) Interface



H₂O + R

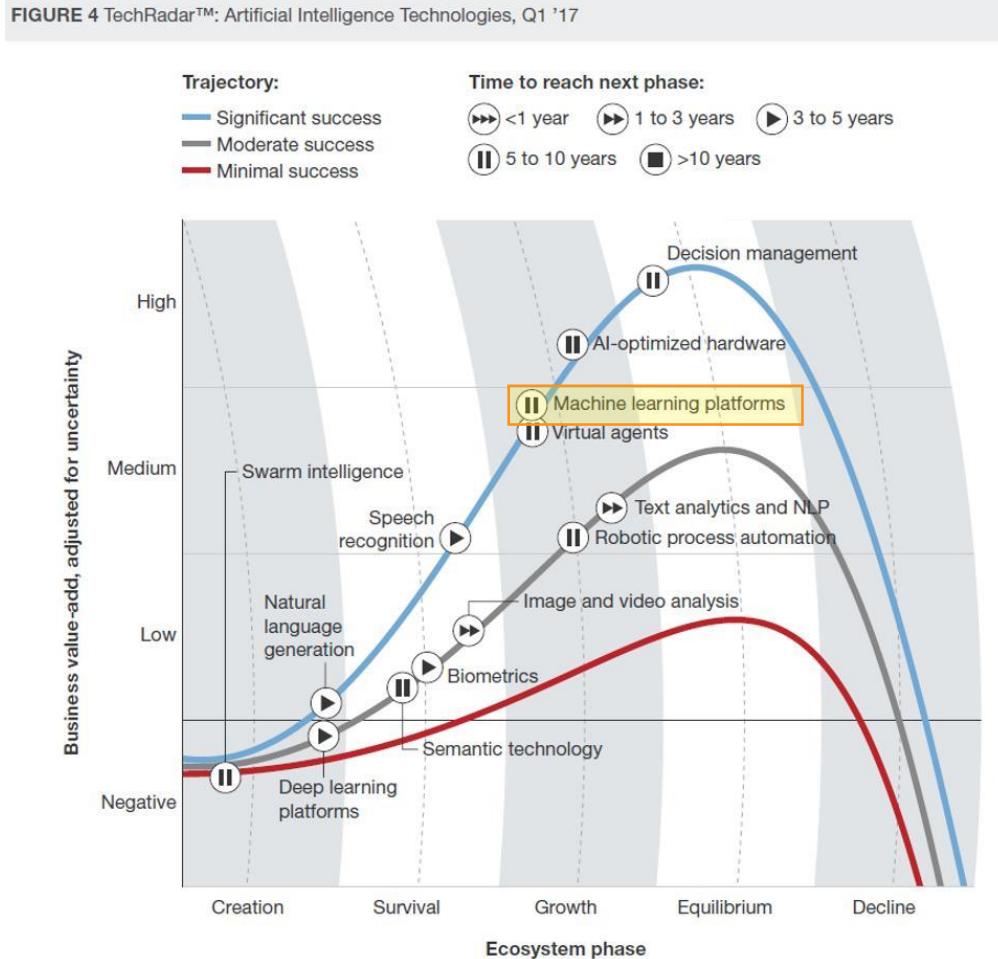


The screenshot shows an RStudio interface with a file named "h2o_iris_demo.R" open. The code in the editor is as follows:

```
1 # -----
2 # Build a simple classification model using iris dataset
3 # -----
4
5 # Start and connect to a local H2O cluster
6 library(h2o)
7 h2o.init(nthreads = -1)
8
9 # Import data from a R data frame
10 data(iris)
11 d_iris <- as.h2o(iris)
12
13 # Define Targets and Features
14 target <- "Species"
15 features <- setdiff(colnames(d_iris), c("Species"))
16
17 # -----
18 # Train a H2O Model
19 # -----
20
21 # Train three basic H2O models
22 model_drf <- h2o.randomForest(x = features,
23                                y = target,
24                                model_id = "iris_random_forest",
25                                training_frame = d_iris)
26
27 model_gbm <- h2o.gbm(x = features,
28                        y = target,
29                        model_id = "iris_gbm",
30                        training_frame = d_iris)
31
32 model_dnn <- h2o.deeplearning(x = features,
33                                y = target,
34                                model_id = "iris_deep_learning",
35                                training_frame = d_iris)
36
```

Business Value of ML Platform

Top 10 Hot A.I. Technologies (Forbes 2017)



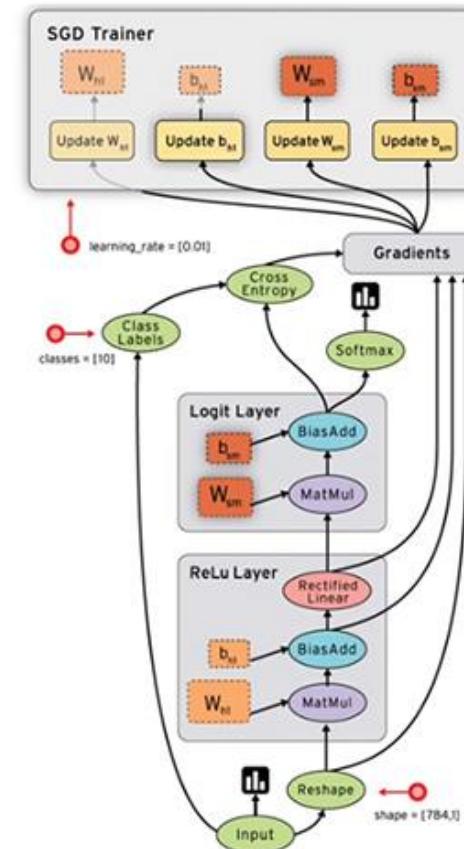
4. Machine Learning Platforms: Providing algorithms, APIs, development and training toolkits, data, as well as computing power to design, train, and deploy models into applications, processes, and other machines. Currently used in a wide range of enterprise applications, mostly involving prediction or classification. Sample vendors: Amazon, Fractal Analytics, Google, H2O.ai, Microsoft, SAS, Skytree.

Deep Learning Tools

TensorFlow, mxnet, Caffe and H₂O Deep Learning

TensorFlow

- Open source machine learning framework by Google
- Python / C++ API
- TensorBoard
 - Data Flow Graph Visualization
- Multi CPU / GPU
 - v0.8+ distributed machines support
- Multi devices support
 - desktop, server and Android devices
- Image, audio and NLP applications
- **HUGE** Community
- Support for Spark, Windows ...



<https://github.com/tensorflow/tensorflow>

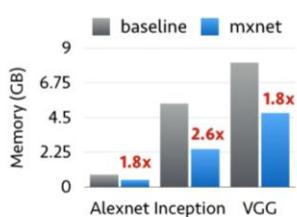
dmlc mxnet for Deep Learning

build passing docs latest license Apache 2.0

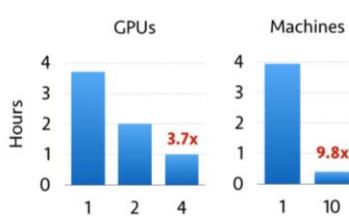
Portable



Efficient



Scalable



MXNet is a deep learning framework designed for both *efficiency* and *flexibility*. It allows you to *mix* the *flavours* of symbolic programming and imperative programming to *maximize* efficiency and productivity. In its core, a dynamic dependency scheduler that automatically parallelizes both symbolic and imperative operations on the fly. A graph optimization layer on top of that makes symbolic execution fast and memory efficient. The library is portable and lightweight, and it scales to multiple GPUs and multiple machines.

MXNet is also more than a deep learning project. It is also a collection of *blue prints and guidelines* for building deep learning system, and interesting insights of DL systems for hackers.

MXNet now chosen by Amazon as Deep Learning Framework

By Geneva Clark | 2016-11-24

19 0

Share this magazine



Amazon has announced that it has chosen MXNet as its deep learning framework of choice for its web services(AWS). Amazon extensively uses machine learning in areas like fraud detection, abusive review detection, and book classification. Amazon also uses it in application areas such as text and speech recognition, autonomous drones etc...

<https://github.com/dmlc/mxnet>

<https://www.zeolearn.com/magazine/amazon-to-use-mxnet-as-deep-learning-framework>

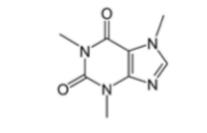
Caffe

- Convolution Architecture For Feature Extraction (CAFFE)
- Pure C++ / CUDA architecture for deep learning
- Command line, Python and MATLAB interface
- Model Zoo
 - Open collection of models

DIY Deep Learning for Vision: a Hands-On Tutorial with Caffe



| | Maximally accurate | Maximally specific |
|----------|--------------------|--------------------|
| espresso | 2.23192 | |
| coffee | 2.19914 | |
| beverage | 1.93214 | |
| liquid | 1.89367 | |
| fluid | 1.85519 | |



caffe.berkeleyvision.org



github.com/BVLC/caffe



Evan Shelhamer, Jeff Donahue, Jon Long,
Yangqing Jia, and Ross Girshick

Look for further
details in the
outline notes



H₂O Deep Learning

Supervised Learning

Statistical Analysis

- **Generalized Linear Models:** Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**

Ensembles

- **Distributed Random Forest:** Classification or regression models
- **Gradient Boosting Machine:** Produces an ensemble of decision trees with increasing refined approximations

Deep Neural Networks

- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

Unsupervised Learning

Clustering

- **K-means:** Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

Dimensionality Reduction

- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Anomaly Detection

- **Autoencoders:** Find outliers using a nonlinear dimensionality reduction using deep learning

Both TensorFlow and H₂O are widely used

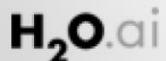
The usage of Hadoop/Big Data tools grew to 39%, up from 29% in 2015 (and 17% in 2014), driven by Apache Spark, MLlib (Spark Machine Learning Library) and H2O.

See also

- KDnuggets interview with Spark Creator Matei Zaharia
- KDnuggets interview with Arno Candel, H2O.ai on How to Quick Start Deep Learning with H2O

<http://www.kdnuggets.com>

H2O and TensorFlow are tied



Deep Water

H₂O.ai Caffe  mxnet  TensorFlow

TensorFlow, **MXNet**, **Caffe** and **H₂O**
democratise the power of deep learning.

H₂O platform democratises artificial
intelligence & big data science.

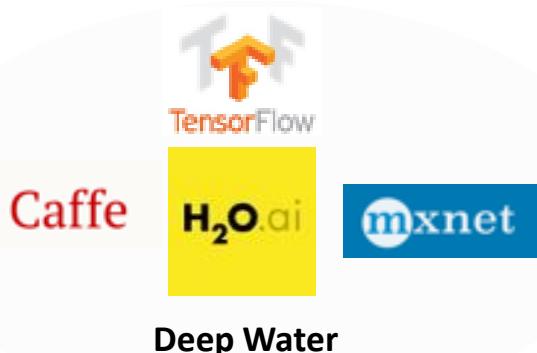
There are other open source deep learning libraries like Theano and Torch too.
Let's have a party, this will be fun!

Deep Water

Next-Gen Distributed Deep Learning with H₂O

One Interface - GPU Enabled - Significant Performance Gains

Inherits All H₂O Properties in Scalability, Ease of Use and Deployment



H₂O integrates with existing **GPU** backends
for **significant performance gains**



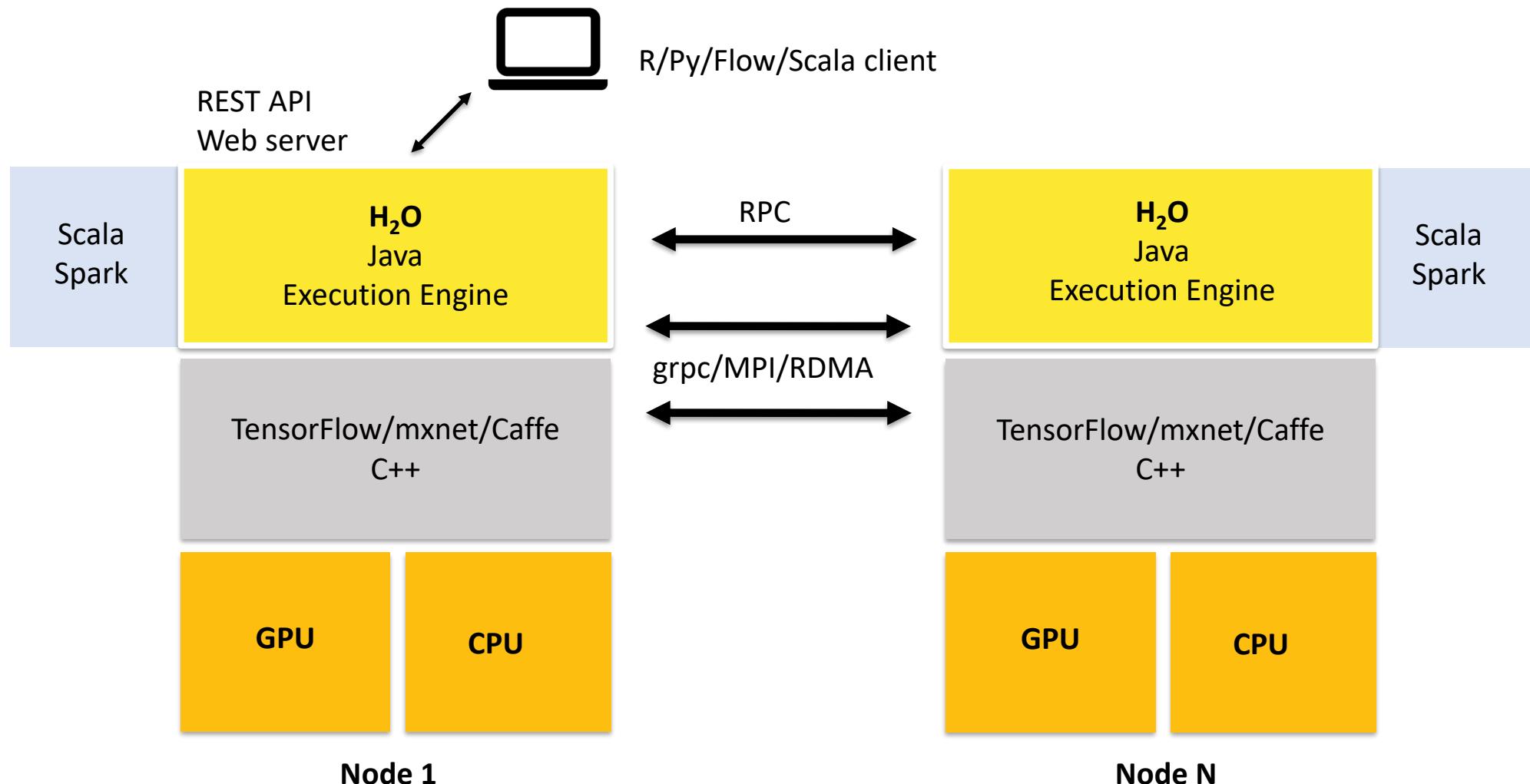
Convolutional Neural Networks enabling
Image, video, speech recognition



Recurrent Neural Networks
enabling **natural language processing, sequences, time series**, and more

Hybrid Neural Network Architectures
enabling **speech to text translation, image captioning, scene parsing** and more

Deep Water Architecture



H₂O FLOW

Flow ▾

Cell ▾

Data ▾

Model ▾

Score ▾

Admin

Help ▾

Untitled Flow



CS

Expression...

Using H₂O Flow to train Deep Water Model

- Deep Learning...
Deep Water... 
 - Distributed Random Forest...
 - Gradient Boosting Method...
 - Generalized Linear Modeling...
 - Generalized Low Rank Modeling...
 - K-means...
 - Naive Bayes...
 - Principal Components Analysis...

 - List All Models
 - List Grid Search Results
 - Import Model...
 - Export Model...



Ready

Ready

localhost:54321/flow/index.html#

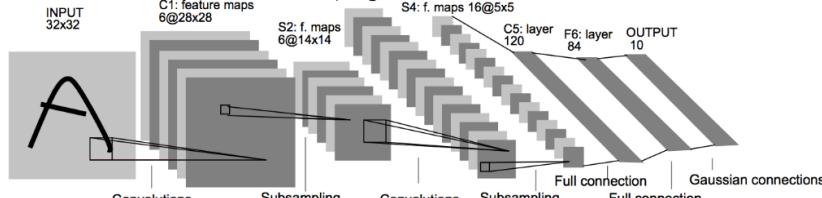
Connections: 0

H_2O



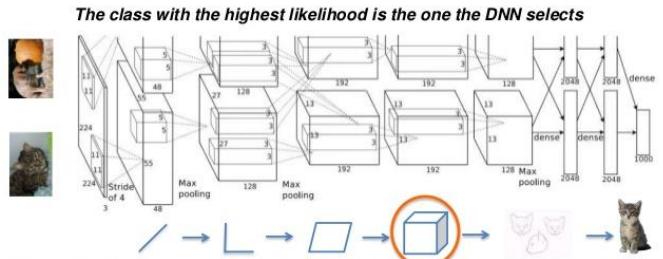
Available Networks in Deep Water

- LeNet
- AlexNet
- VGGNet
- Inception (GoogLeNet)
- ResNet (Deep Residual Learning)
- Build Your Own



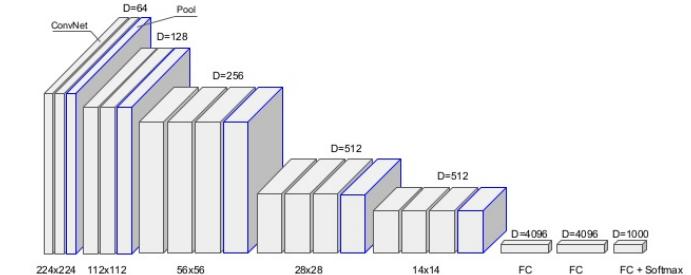
CNN called LeNet by Yann LeCun (1998)

AlexNet (Krizhevsky et al. 2012)

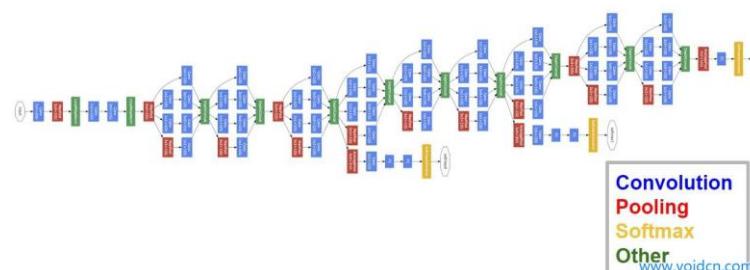


When AlexNet is processing an image, this is what is happening at each layer.

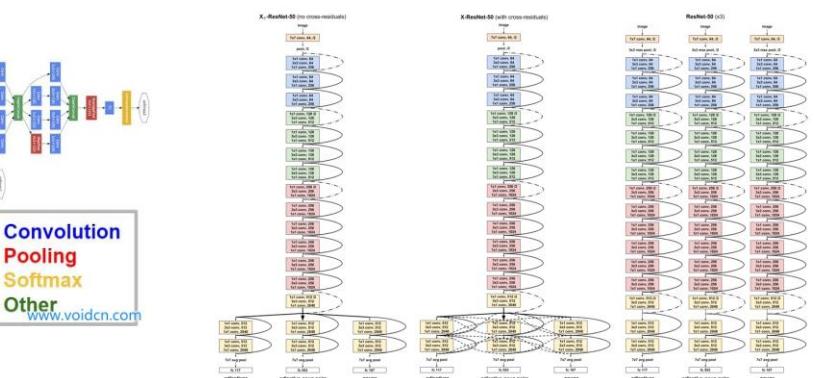
Classical CNN topology - VGGNet (2013)



GoogLeNet



ResNet



Deep Water H2O and TensorFlow Demo



Choosing different network structures

All

None

Only show columns with more than % missing values.

epochs

How many times the dataset should be iterated (streamed), can be fractional.

ignore_const_cols

Ignore constant columns.

network

Network architecture.

activation

Activation function. Only used if no user-defined network architecture file is provided, and only for problem_type=dataset.

hidden

Hidden layer sizes (e.g. [200, 200]). Only used if no user-defined network architecture file is provided, and only for problem_type=dataset.

problem_type

Problem type, auto-detected by default. If set to image, the H2OFrame must contain a string column containing the path (URI or URL) to the images in the first column. If set to text, the H2OFrame must contain a string column containing the text in the first column. If set to dataset, Deep Water behaves just like any other H2O Model and builds a model on the provided H2OFrame (non-String columns).

ADVANCED

GRID ?

checkpoint

Model checkpoint to resume training with.

autoencoder

Auto-Encoder.

balance_classes

Balance training data class counts via over/under-sampling (for imbalanced data).

fold_column

Column with cross-validation fold index assignment per observation.

offset_column

Offset column. This will be added to the combination of columns before applying the link function.

Unified Interface for TF, MXNet and Caffe

```
: model = H2ODeepWaterEstimator(epochs      = 500,
                               network       = "lenet",
                               image_shape  = [28,28], ## provide image size
                               channels     = 3,
                               backend       = "tensorflow",
                               model_id     = "deepwater_tf_simple")

model.train(x = [0], # file path e.g. xxx/xxx/xxx.jpg
            y = 1, # label cat/dog/mouse
            training_frame = frame)

model.show()

deepwater Model Build progress: |██████████| 100%
Model Details
=====
H2ODeepWaterEstimator : Deep Water
Model Key: deepwater_tf_simple
```

Change backend to
“mxnet”, “caffe” or “auto”

Easy Stacking with other H₂O Models

Model Stacking

Now we have three different models, we are ready to carry out model stacking.

```
In [47]: # Create a list to include all the models for stacking  
models <- list(model_dw, model_gbm, model_drf)
```

```
In [48]: # Define a metalearner (one of the H2O supervised machine learning algorithms)  
metalearner <- "h2o.glm.wrapper"
```

```
In [49]: # Use h2o.stack() to carry out metalearning  
stack <- h2o.stack(models = models,  
                    response_frame = h_train$medv,  
                    metalearner = metalearner)
```

```
[1] "Metalearning"
```

```
In [50]: # Finally, we evaluate the predictive performance on the ensemble as well as individual models.  
h2o.ensemble_performance(stack, newdata = h_test)
```

```
Base learner performance, sorted by specified metric:  
learner      MSE  
1 h2o_deepwater 8.377644  
2      h2o_gbm 8.106541  
3      h2o_drf 7.443517
```

```
H2O Ensemble Performance on <newdata>:  
-----
```

```
Family: gaussian
```

```
Ensemble performance (MSE): 5.80436983051916
```

Ensemble of Deep Water, Gradient Boosting
Machine & Random Forest models

H₂O, Sparkling Water, Steam, & Deep Water Documentation

[Getting Started](#)[Data Science Algorithms](#)[Languages](#)[Tutorials, Examples, & Presentations](#)[For Developers](#)[For the Enterprise](#)

docs.h2o.ai

Getting Started



H₂O

[What is H₂O?](#)
[H₂O User Guide](#)
[H₂O Book \(O'Reilly\)](#)
[Recent Changes](#)
[Open Source License \(Apache V2\)](#)

[Quick Start Video - Flow Web UI](#)
[Quick Start Video - R](#)
[Quick Start Video - Python](#)

[Download H₂O](#)

Sparkling Water

[What is Sparkling Water?](#)
[Sparkling Water Booklet](#)
[PySparkling Readme 2.0 | 1.6](#)
[RSparkling Readme](#)
[Open Source License \(Apache V2\)](#)

[Quick Start Video - Scala](#)
[Quick Start Video - Python](#)

[Download Sparkling Water](#)

Steam

[What is Steam?](#)
[Steam User Guide](#)
[Recent Changes](#)
[Open Source License \(AGPL\)](#)

[Download Steam](#)

Deep Water (preview)

[Deep Water Readme](#)
[Deep Water AMI Guide](#)
[Open Source License \(Apache V2\)](#)

[Launch Deep Water AMI
\(choose g2.2xlarge\)](#)

Q & A

[FAQ](#)
[Community Forum](#)
[h2ostream Google Group](#)
[Issue Tracking \(JIRA\)](#)
[Gitter](#)
[Stack Overflow](#)
[Cross Validated](#)

For Supported Enterprise Customers
[Enterprise Support Web | Email](#)



mstensmo changing the name of deeplearning_credit_card_default_risk_prediction...

...

Latest commit 5568350 11 days ago

| | | |
|---|--|--------------|
| .. | | |
| images | Add cat/dog/mouse lenet example. | 3 months ago |
| README.md | Update README.md | 2 months ago |
| deeplearning_anomaly_detection.ipynb | Update notebooks, introduce local paths to ~/h2o-3/ | 3 months ago |
| deeplearning_benchmark_mnist.ipynb | Update lenet test to remove all. Update MNIST benchmark with comments. | 3 months ago |
| deeplearning_cat_dog_mouse_inception.ipynb | Add credit card default risk model, update other notebooks. | 3 months ago |
| deeplearning_cat_dog_mouse_lenet.ipynb | Add credit card default risk model, update other notebooks. | 3 months ago |
| deeplearning_cat_dog_mouse_lenet.ipynb | Add back model.plot() and scoring history. | 3 months ago |
| deeplearning_cifar10_vgg.ipynb | Rename notebooks. | 3 months ago |
| deeplearning_credit_card_default_risk.ipynb | changing the name of deeplearning_credit_card_default_risk_prediction... | 11 days ago |
| deeplearning_ensemble_boston_housing.ipynb | Ensemble demo using GBM, DRF and Deep Water (#676) | 17 days ago |
| deeplearning_grid_iris.ipynb | Add two new notebooks: Lenet for R and iris grid for python | 3 months ago |
| deeplearning_grid_iris_R.ipynb | Update R py notebook. | 3 months ago |
| deeplearning_image_reconstruction.ipynb | Update notebooks, introduce local paths to ~/h2o-3/ | 3 months ago |
| deeplearning_mnist_convnet.ipynb | Update notebooks, introduce local paths to ~/h2o-3/ | 3 months ago |
| deeplearning_mnist_introduction.ipynb | Add missing file. | 3 months ago |
| deeplearning_tensorflow_cat_dog.ipynb | Add tensorflow example (#529) | 2 months ago |
| deeplearning_tensorflow_mnist.ipynb | Added MNIST example for TensorFlow | a month ago |

Conclusions

Project “Deep Water”

- H₂O + TF + MXNet + Caffe
 - a powerful combination of widely used open source machine learning libraries.
- All Goodies from H₂O
 - inherits all H₂O properties in scalability, ease of use and deployment.
- Unified Interface
 - allows users to build, stack and deploy deep learning models from different libraries efficiently.

- 100% Open Source
 - the party will get bigger!



Deep Water – Current Contributors



Fabrizio Milo



Cyprien Noel



Qiang Kou



Arno Candel



Caffe



H₂O.ai

Thanks!

- Organizers & Sponsors
 - H₂O.ai (Pizzas)
 - Dataiku (Venue)
 - ODSC
- Code, Slides & Documents
 - bit.ly/h2o_meetups
 - bit.ly/h2o_deepwater
 - docs.h2o.ai
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe



H₂O.ai

Making Machine Learning
Accessible to Everyone

Photo credit: Virgin Media



Got a tip? [Let us know.](#)

Follow Us [f](#) [i](#) [t](#) [g](#) [in](#) [g+](#) [r](#)

News ▾ Video ▾ Events ▾ Crunchbase

[Message Us](#)

[Search](#)



10TH ANNUAL CRUNCHIES AWARDS Final 2 Days To Save On Crunchies Tickets [Get Yours Now ▶](#)

Water

Software

deep learning

H2O.ai

Artificial Intelligence

Popular Posts



Doug shows you how to get rid of Amazon Fresh totes
3 days ago



Facebook will give some longer videos a boost in the News Feed
3 days ago



Qualcomm reaffirms it will continue to supply Apple during its legal dispute
4 days ago



GM and Honda partner to mass produce

H2O's Deep Water puts deep learning in the hands of enterprise users

Posted Jan 26, 2017 by [John Mannes \(@JohnMannes\)](#)



To complement existing offerings like Sparkling Water and Steam, [H2O.ai is releasing Deep Water](#), a new tool to help businesses make deep learning a part of everyday operations.

Deep Water will open up new possibilities for the TensorFlow, MXNet and Caffe communities to engage with H2O.ai. This also means that the GPU is set to become a greater part of business operations for the entire Fortune 500, not just tech companies.

Crunchbase

Matter



Steam



NEWSLETTER SUBSCRIPTIONS

[The Daily Crunch](#)

Get the top tech stories of the day delivered to your inbox

[TC Weekly Roundup](#)

Get a weekly recap of the biggest tech stories

[Crunchbase Daily](#)

The latest startup funding announcements

[SUBSCRIBE](#)

<https://techcrunch.com/2017/01/26/h2os-deep-water-puts-deep-learning-in-the-hands-of-enterprise-users/>