H_2O Deep Water

Making Deep Learning Accessible to Everyone

Jo-fai (Joe) Chow – Data Scientist at H_2O.ai

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 H_2O in 2014
- Data Scientist
  - From 2015
  - Virgin Media (UK)
  - Domino Data Lab (Silicon Valley)
  - H_2O.ai (Silicon Valley)

Agenda

- About H_2O.ai
  - Company
  - Machine Learning Platform
- Deep Learning Tools
  - TensorFlow, MXNet, Caffe, H_2O
- Deep Water
  - Motivation
  - Benefits
  - Interface
  - Learning Resources
- Conclusions

About H_2O.ai

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	• H_2O 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 • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Headquarters	Mountain View, CA

Scientific Advisory Council

Dr. Trevor Hastie

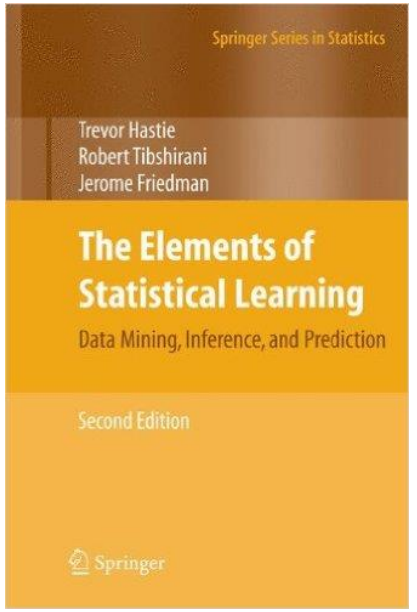
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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*

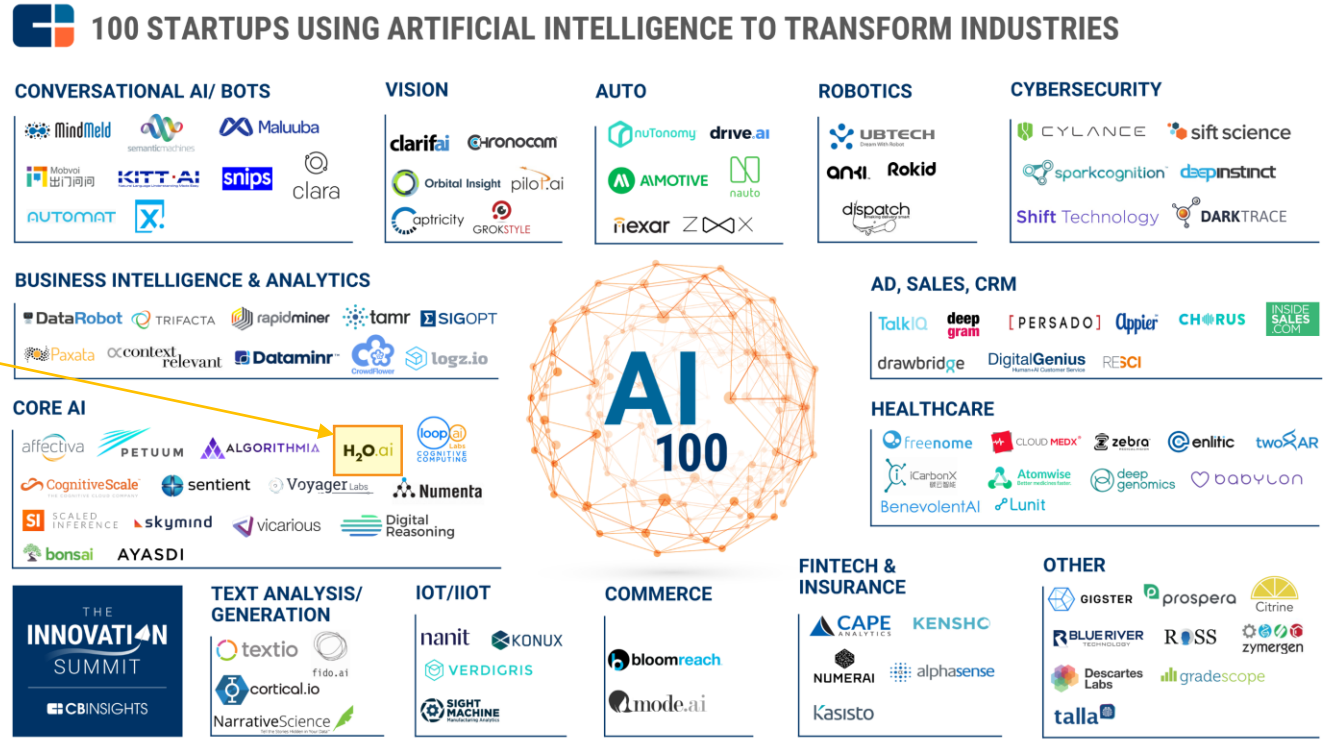


CB INSIGHTS Data Analytics & Research Ratings About us

Blog Industries Geographies VC Performance & Rankings Team Blog

JANUARY 11, 2017

AI 100: The Artificial Intelligence Startups Redefining Industries



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

H_2O Machine Learning Platform

Algorithms Overview

Supervised Learning

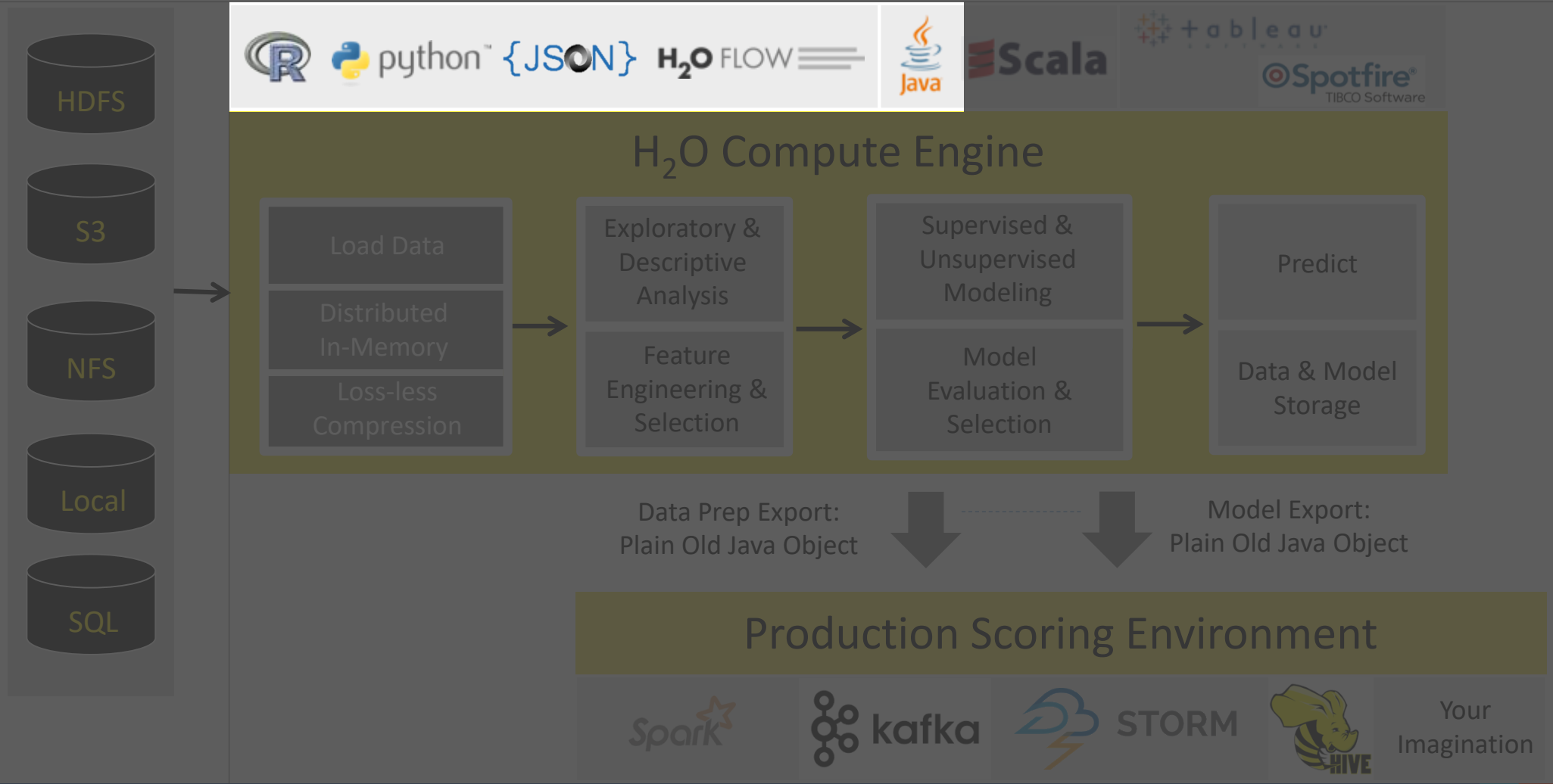
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
Deep Neural Networks	• **Deep learning**: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

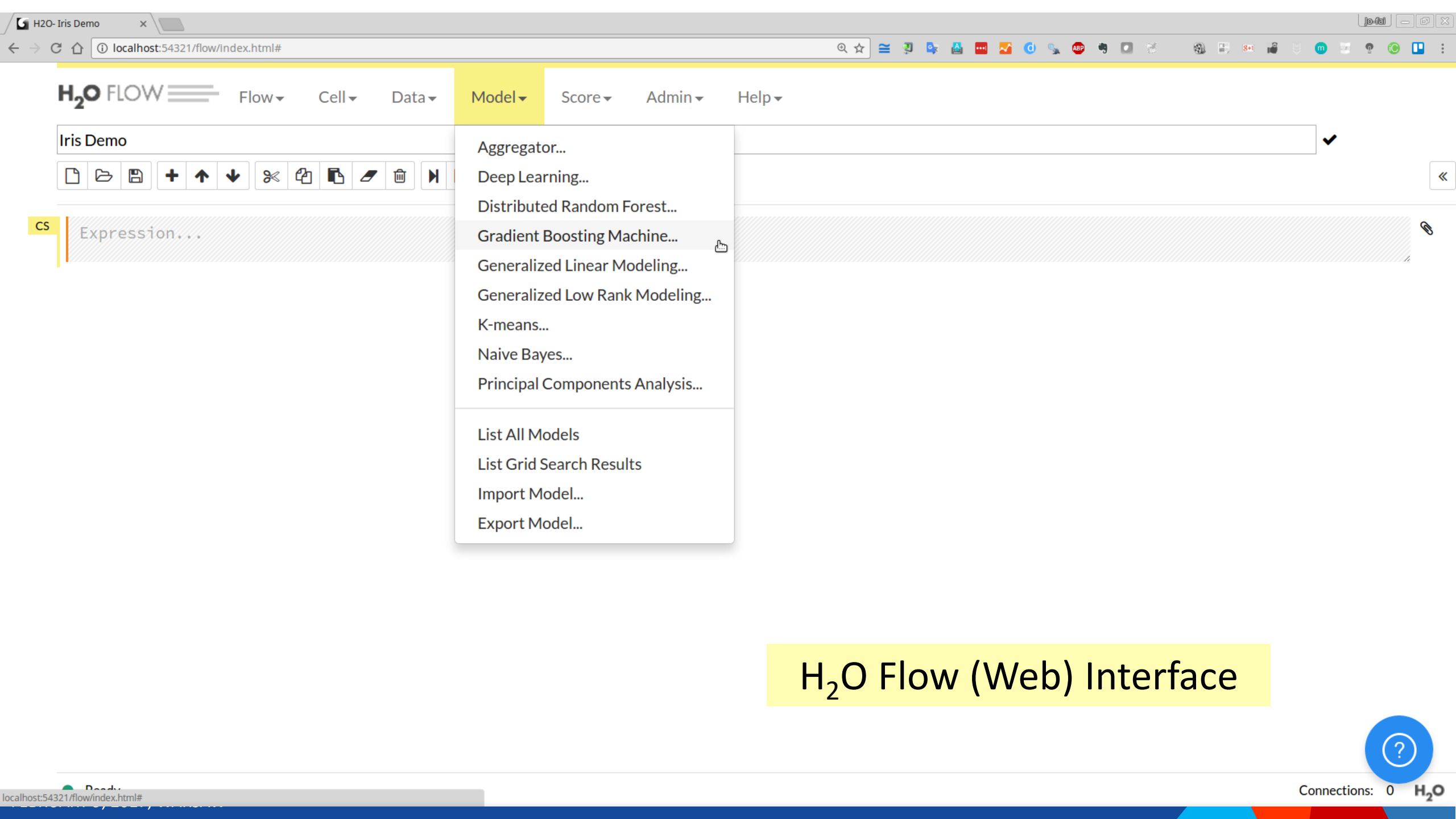
Unsupervised Learning

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

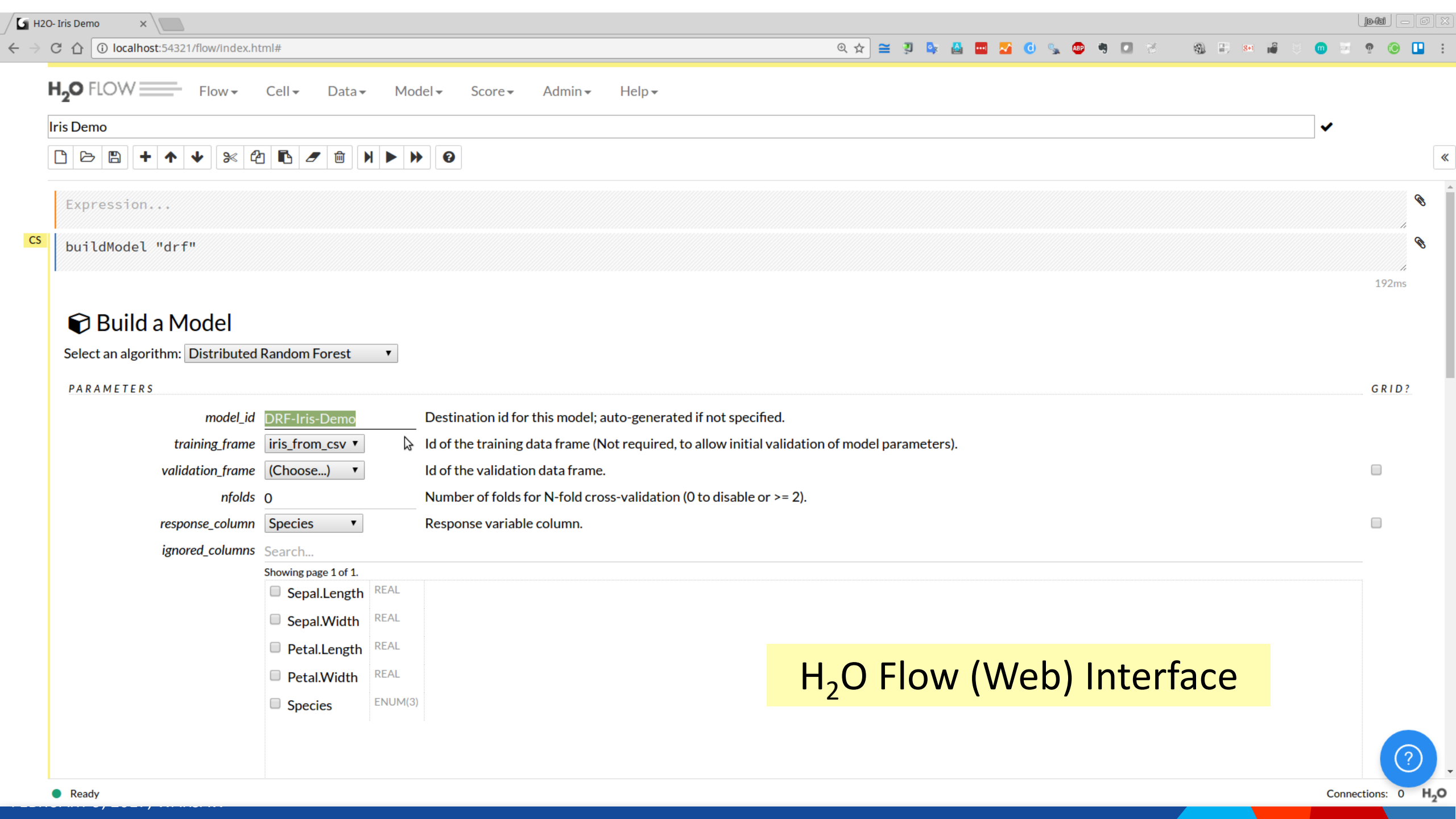
High Level Architecture

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





H_2O Flow (Web) Interface


H2O- Iris Demo
localhost:54321/flow/index.html#
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Iris Demo
Expression...
CS
buildModel "drf"
192ms
Build a Model
Select an algorithm: Distributed Random Forest
PARAMETERS
GRID?
model_id
DRF-Iris-Demo
Destination id for this model; auto-generated if not specified.
training_frame
iris_from_csv
Id of the training data frame (Not required, to allow initial validation of model parameters).
validation_frame
(Choose...)
Id of the validation data frame.
nfolds
0
Number of folds for N-fold cross-validation (0 to disable or >= 2).
response_column
Species
Response variable column.
ignored_columns
Search...
Showing page 1 of 1.
Sepal.Length REAL
Sepal.Width REAL
Petal.Length REAL
Petal.Width REAL
Species ENUM(3)
H2O Flow (Web) Interface
Ready
Connections: 0

$H_2O + R$

h2o_iris_demo.R

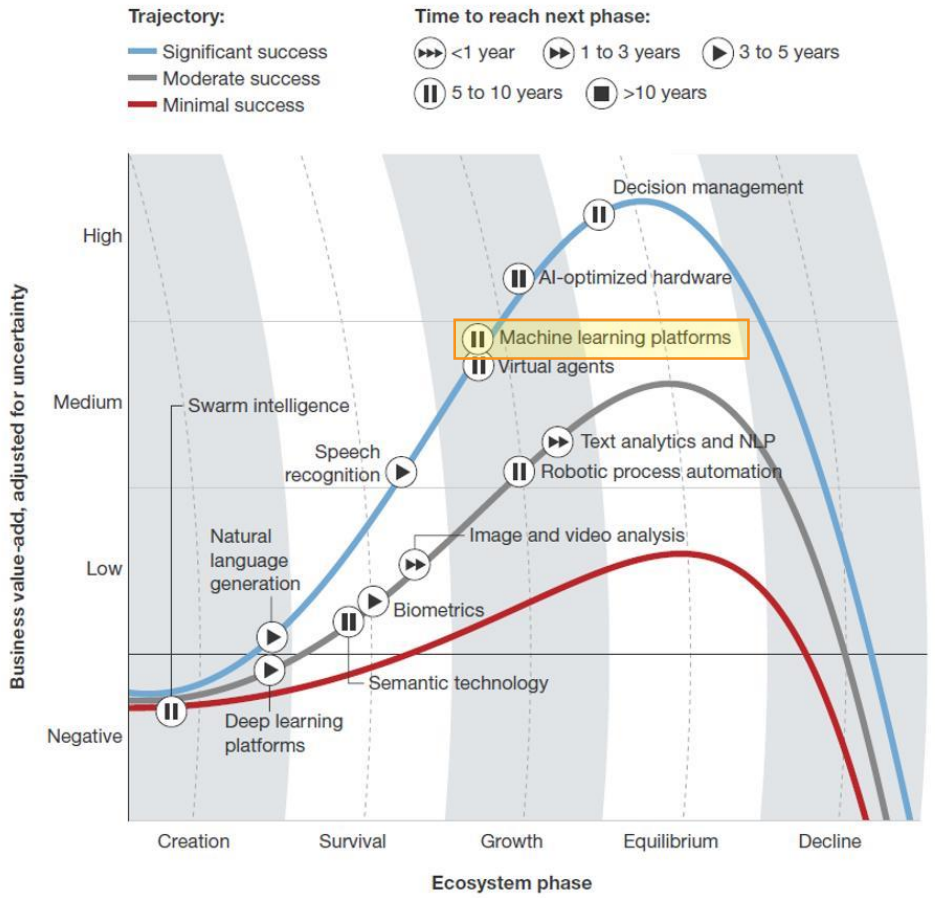
Source on Save | Run | Source

```r
# ------------------------------------------------------------------------------
# Build a simple classification model using iris dataset
# ------------------------------------------------------------------------------

# Start and connect to a local H2O cluster
library(h2o)
h2o.init(nthreads = -1)

# Import data from a R data frame
data(iris)
d_iris <- as.h2o(iris)

# Define Targets and Features
target <- "Species"
features <- setdiff(colnames(d_iris), c("Species"))

# ------------------------------------------------------------------------------
# Train a H2O Model
# ------------------------------------------------------------------------------

# Train three basic H2O models
model_drf <- h2o.randomForest(x = features,
                              y = target,
                              model_id = "iris_random_forest",
                              training_frame = d_iris)

model_gbm <- h2o.gbm(x = features,
                     y = target,
                     model_id = "iris_gbm",
                     training_frame = d_iris)

model_dnn <- h2o.deeplearning(x = features,
                              y = target,
                              model_id = "iris_deep_learning",
                              training_frame = d_iris)
```

Business Value of ML Platform

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

FIGURE 4 TechRadar™: Artificial Intelligence Technologies, Q1 '17



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.

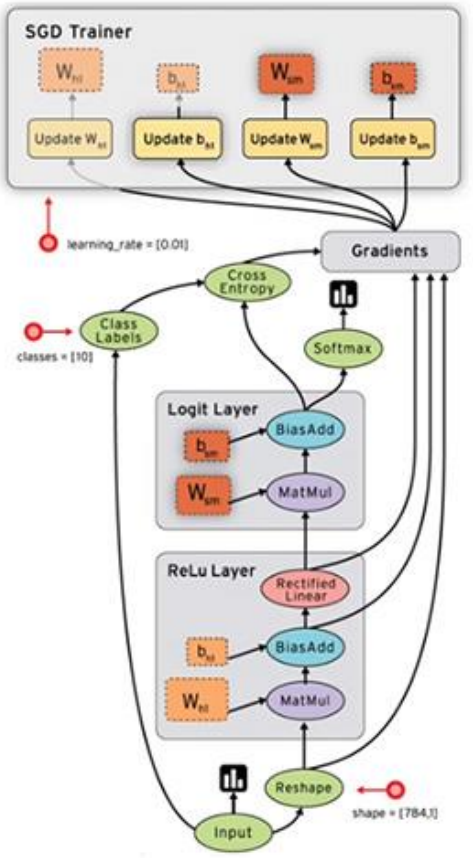
http://www.forbes.com/sites/gilpress/2017/01/23/top-10-hot-artificial-intelligence-ai-technologies/

Deep Learning Tools

TensorFlow, mxnet, Caffe and H_2O 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 …



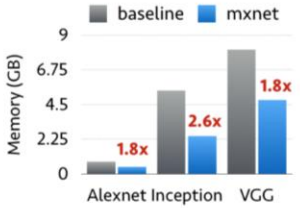
https://github.com/tensorflow/tensorflow

dmlc mxnet for Deep Learning

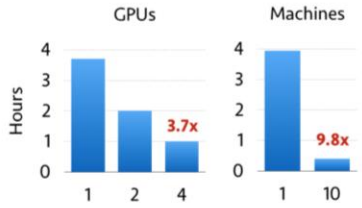
build passing | docs latest | license Apache 2.0

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

- **C**onvolution **A**rchitecture **F**or **F**eature **E**xtraction (CAFFE)
- Pure C++ / CUDA architecture for deep learning
- Command line, Python and MATLAB interface
- Model Zoo
  - Open collection of models

https://docs.google.com/presentation/d/1UeKXVgRvvxg9OUdh_UiC5G71UMscNPlvArsWER41PsU/

H_2O Deep Learning

	Supervised Learning		Unsupervised Learning	
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**	Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations	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
Deep Neural Networks	• **Deep learning**: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations	Anomaly Detection	• **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

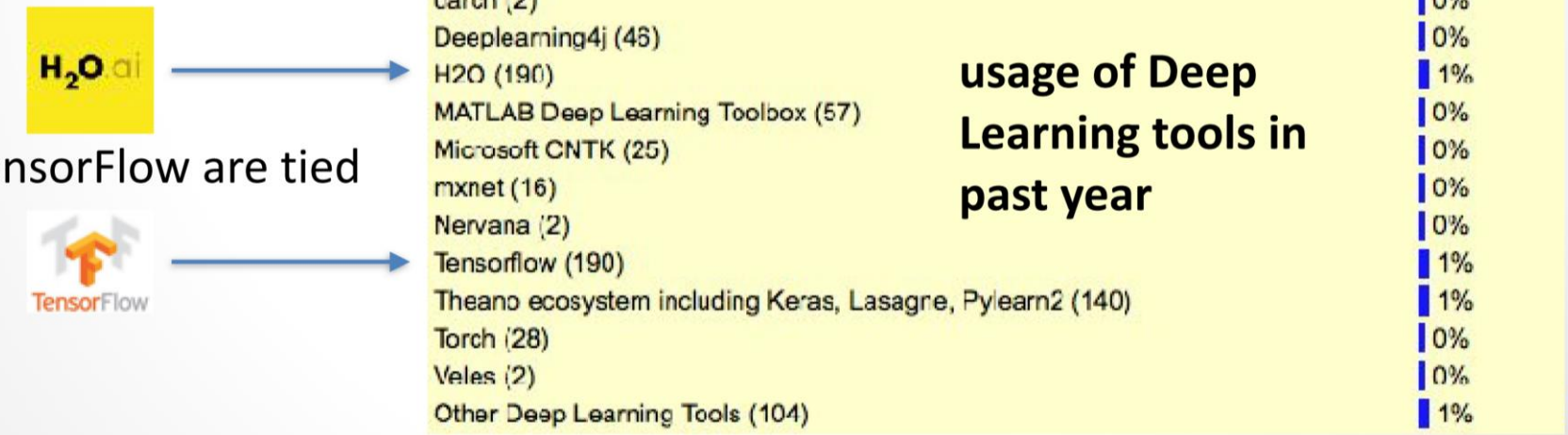
Both TensorFlow and H_2O are widely used

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




H2O.ai
Products
Industries
Customers
Community
Documentation
Download
Deep Water
H2O.ai
Caffe
mxnet
TensorFlow

TensorFlow , MXNet, Caffe and H_2O democratize the power of deep learning.

H_2O platform democratizes 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 H_2O

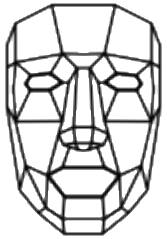
One Interface - GPU Enabled - Significant Performance Gains

Inherits All H_2O Properties in Scalability, Ease of Use and Deployment

TensorFlow Caffe H_2O.ai mxnet

Deep Water

H_2O integrates with existing GPU backends for significant performance gains

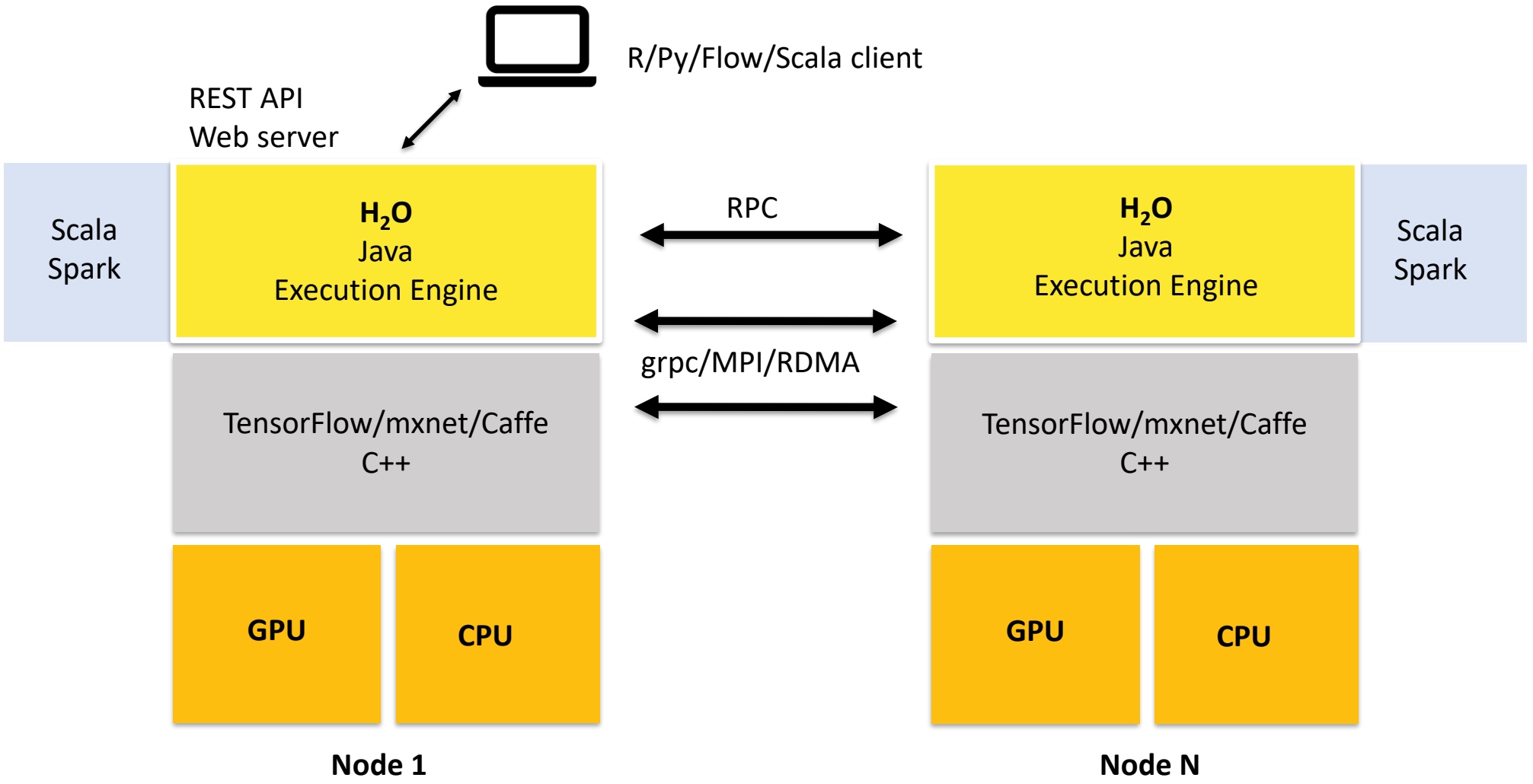


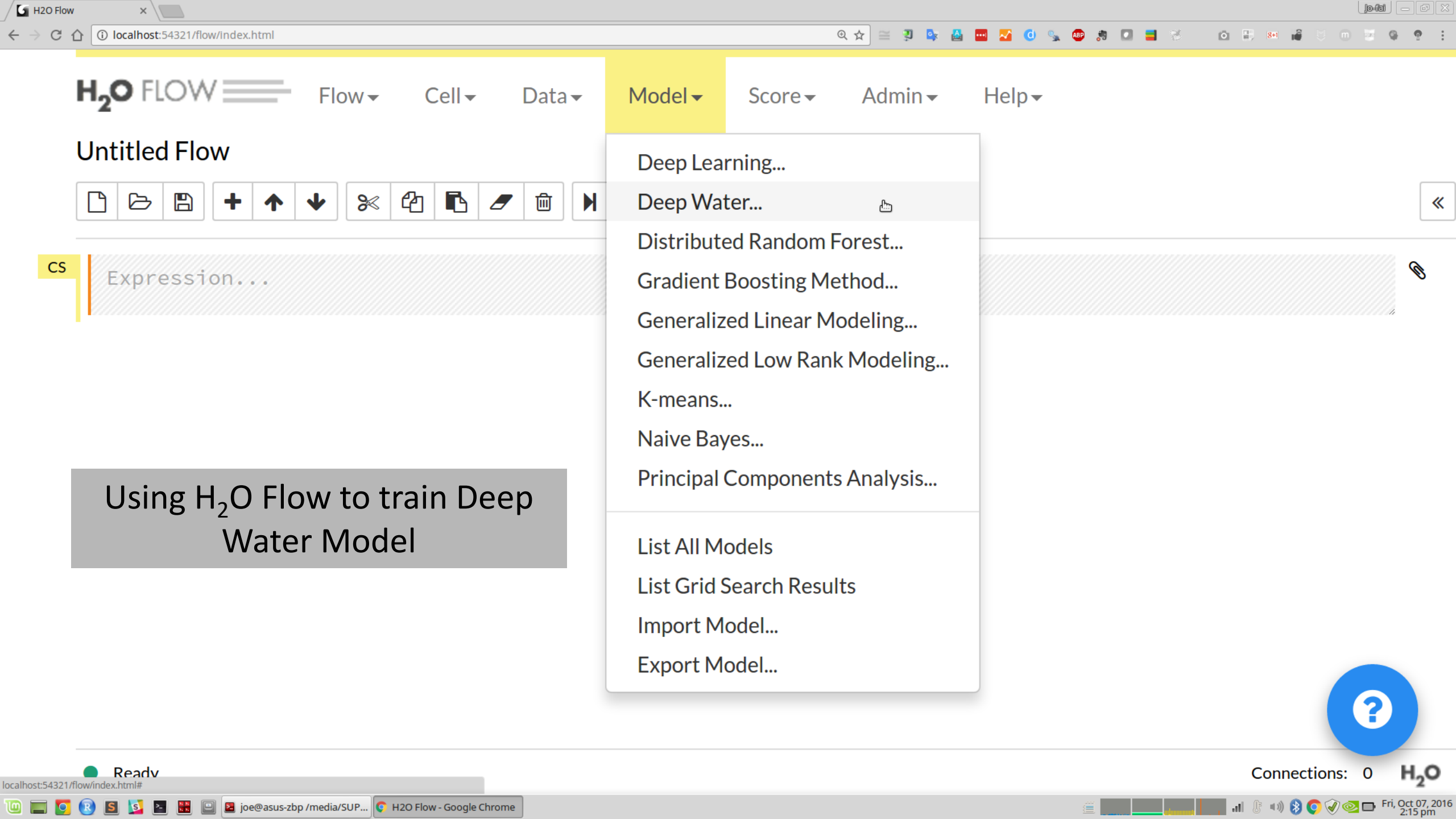
Convolutional Neural Networks enabling **Image, video, speech recognition**

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

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

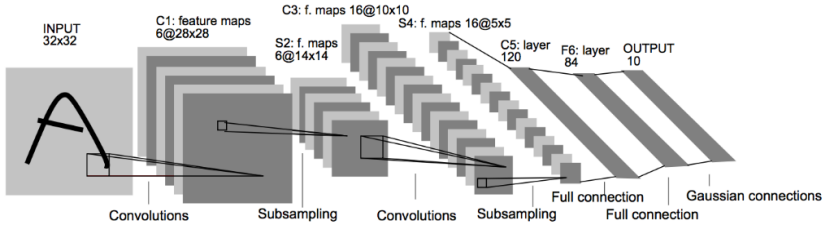
Deep Water Architecture




H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
CS
Expression...
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...
Using H2O Flow to train Deep Water Model
Ready
Connections: 0
H2O
localhost:54321/flow/index.html#
joe@asus-zbp /media/SUP...
H2O Flow - Google Chrome
Fri, Oct 07, 2016
2:15 pm

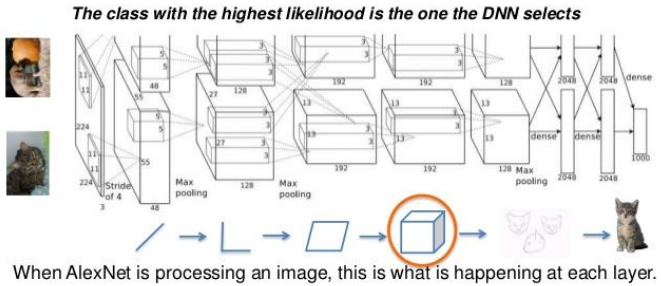
Available Networks in Deep Water

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



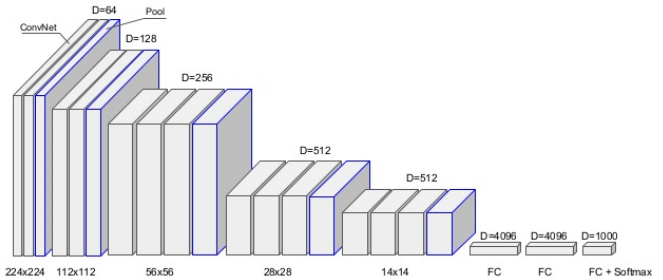
CNN called LeNet by Yann LeCun (1998)

AlexNet (Krizhevsky et al. 2012)

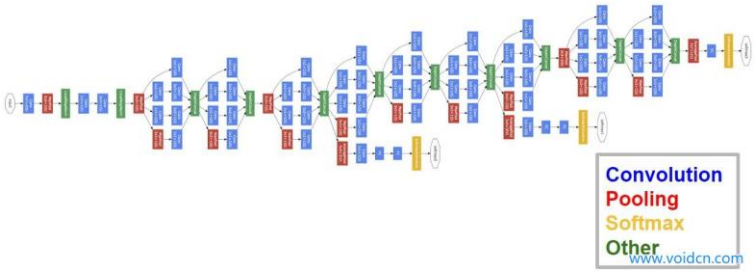


When AlexNet is processing an image, this is what is happening at each layer.

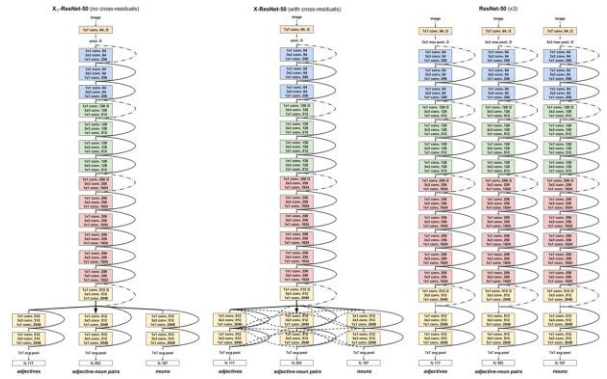
Classical CNN topology - VGGNet (2013)

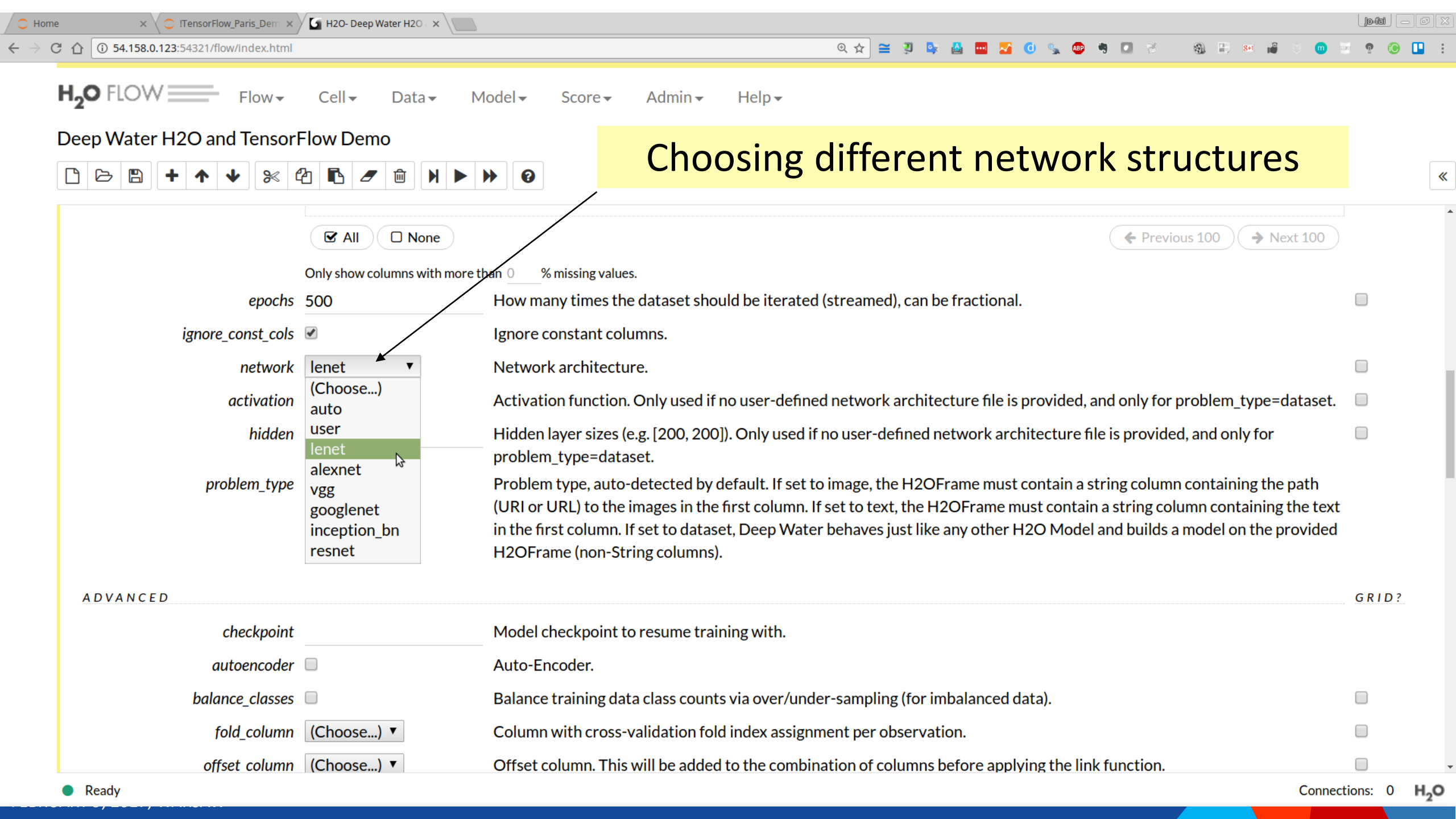


GoogLeNet



ResNet




Home
ITensorFlow_Paris_Dem
H2O- Deep Water H2O
54.158.0.123:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Deep Water H2O and TensorFlow Demo
Choosing different network structures
All
None
Previous 100
Next 100
Only show columns with more than 0 % missing values.
epochs
500
How many times the dataset should be iterated (streamed), can be fractional.
ignore_const_cols
Ignore constant columns.
network
lenet
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).
(Choose...)
auto
user
lenet
alexnet
vgg
googlenet
inception_bn
resnet
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
(Choose...)
Column with cross-validation fold index assignment per observation.
offset_column
(Choose...)
Offset column. This will be added to the combination of columns before applying the link function.
Ready
Connections: 0

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 H_2O 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

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

Branch: master ▾ h2o-3 / examples / deeplearning / notebooks / Create new file | Upload files | Find file | History

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.ip...	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_incep...	Add credit card default risk model, update other notebooks.	3 months ago
deeplearning_cat_dog_mouse_lenet....	Add credit card default risk model, update other notebooks.	3 months ago
deeplearning_cat_dog_mouse_lenet...	Add back model.plot() and scoring history.	3 months ago
deeplearning_cifar10_vgg.ipynb	Rename notebooks.	3 months ago
deeplearning_credit_card_default_ri...	changing the name of deeplearning_credit_card_default_risk_prediction...	11 days ago
deeplearning_ensemble_boston_ho...	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...	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.ip...	Add missing file.	3 months ago
deeplearning_tensorflow_cat_dog_...	Add tensorflow example (#529)	2 months ago
deeplearning_tensorflow_mnist.ipynb	Added MNIST example for TensorFlow	a month ago

https://github.com/h2oai/h2o-3/tree/master/examples/deeplearning/notebooks

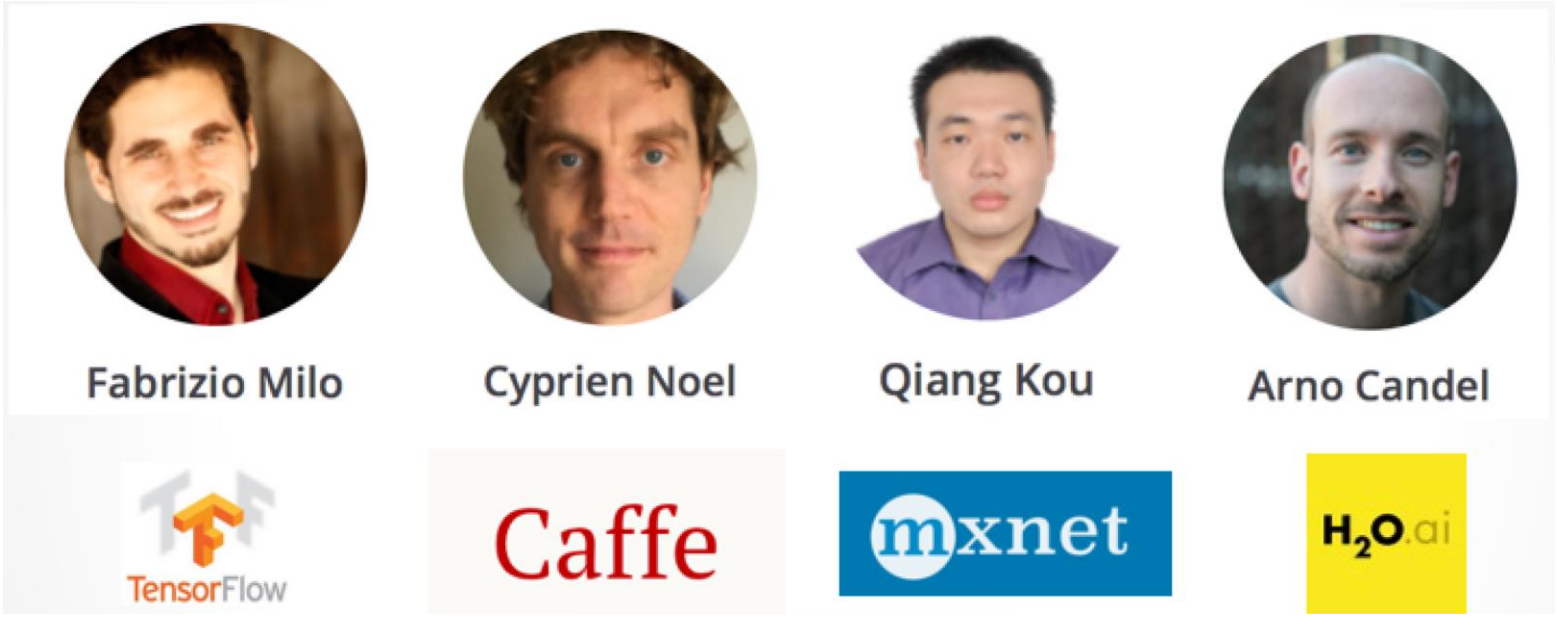
Conclusions

Project “Deep Water”

- H_2O + TF + MXNet + Caffe
  - a powerful combination of widely used open source machine learning libraries.
- All Goodies from H_2O
  - inherits all H_2O 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



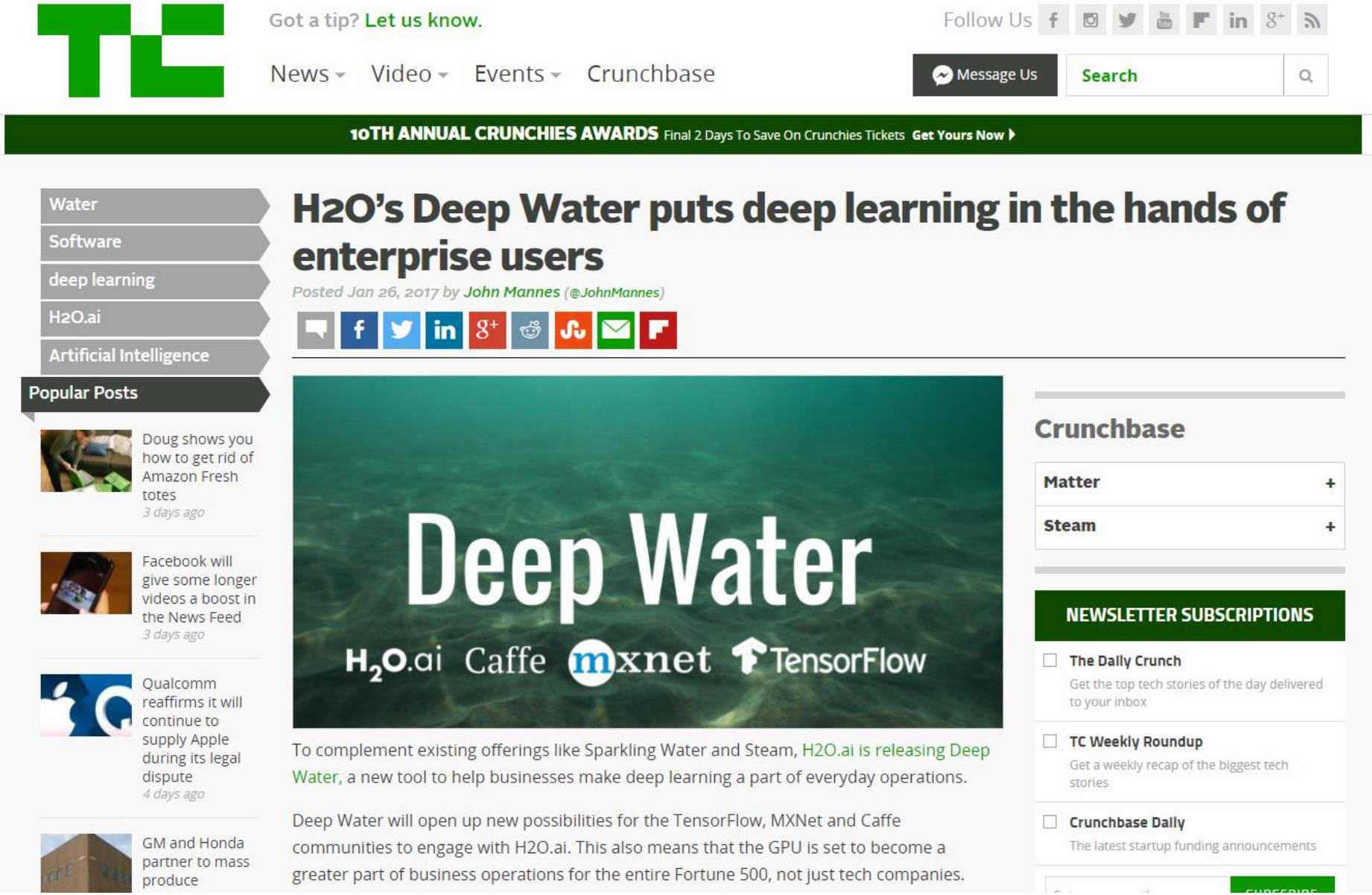
Thanks!

- Organizers & Sponsors
  - Adam Kawa
  - Karolina Seliga
- Code, Slides & Documents
  - bit.ly/h2o_meetups
  - bit.ly/h2o_deepwater
  - docs.h2o.ai
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe

Photo credit: Virgin Media

H2O.ai

Making Machine Learning Accessible to Everyone



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