# Introduction to $H_2O$

Jakub Háva
jakub@h2o.ai

Thanks for the slides to:

Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai
@matlabulous

The Amsterdam Pipeline Factory
of Data Science, Amsterdam
6th December, 2016

# About $H_2O$.ai

What exactly is $H_2O$?

# Company Overview

| | |
|---|---|
| **Founded** | 2011 Venture-backed, debuted in 2012 |
| **Products** | • H2O Open Source In-Memory AI Prediction Engine<br>• Sparkling Water<br>• Steam |
| **Mission** | Operationalize Data Science, and provide a platform for users to build beautiful data products |
| **Team** | 70 employees<br>• Distributed Systems Engineers doing Machine Learning<br>• World-class visualization designers |
| **Headquarters** | Mountain View, CA |



H2O.ai

$H_2O$ is an open source platform empowering business transformation

# Bring AI To Business Empower Transformation

# Financial Services, Insurance and Healthcare as Our Vertical Focus

# Community as Our Foundation

# Users In Various Verticals Adore $H_2O$

| Financial | Telecom | Insurance | Healthcare | Marketing |
|---|---|---|---|---|
| Capital One | COMCAST | TRANSAMERICA | aetna | macy's |
| PayPal | CISCO | ZURICH | HCA Hospital Corporation of America | MarketShare a Neustar Solution |
| | | PROGRESSIVE | KAISER PERMANENTE | Nielsen Catalina SOLUTIONS |
| | | | Walgreens | |

H2O.ai

# H2O In Action

www.h2o.ai/customers

### Capital One

Capital One uses H2O open source machine learning for various use cases.

### MarketShare


H2O predictive analytics helps boost the impact and results of digital marketing.

### Kaiser


Kaiser uses H2O machine learning to save lives.

### Zurich Insurance

Zurich turned to H2O as a strategic differentiator for commercial insurance.

### Progressive

Progressive uses H2O predictive analytics for user-based insurance.

### Comcast

Comcast uses H2O to improve customer experience.

### Hospital Corporation of America

HCA uses H2O to predict patient outcomes in real-time.

### McKesson

McKesson discusses the adoption of artificial intelligence in healthcare.

### Macy's

Macy's uses H2O for personalized site recommendations.

### Transamerica

Transamerica turns to H2O to develop a product recommendation platform for insurance.

### Paypal

Paypal turned to H2O Deep Learning for fraud detection and customer churn.

### eBay

eBay chose H2O for open source machine learning.

H2O.ai

# $H_2O$.ai Makes A Difference as an AI Platform

## Open Source


- 100% open source

## Big Data Ecosystem

hadoop

Spark

## Flexible Interface

R

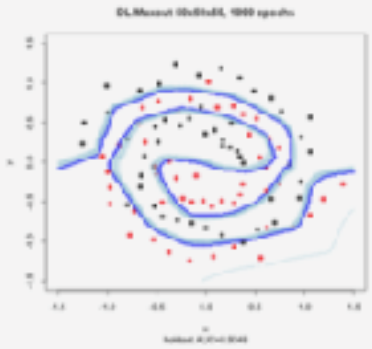
Scala

Java

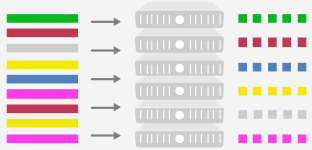
python

$H_2O$ Flow

## Smart and Fast Algorithms



## Scalability and Performance



- Distributed In-Memory Computing Platform
- Distributed Algorithms
- Fine-Grain MapReduce

## Rapid Model Deployment

- Highly portable models deployed in Java (POJO) and Model Object Optimized (MOJO)
- Automated and streamlined scoring service deployment with Rest API

REST API

## GPU Enablement

GPU

## Cloud Integration

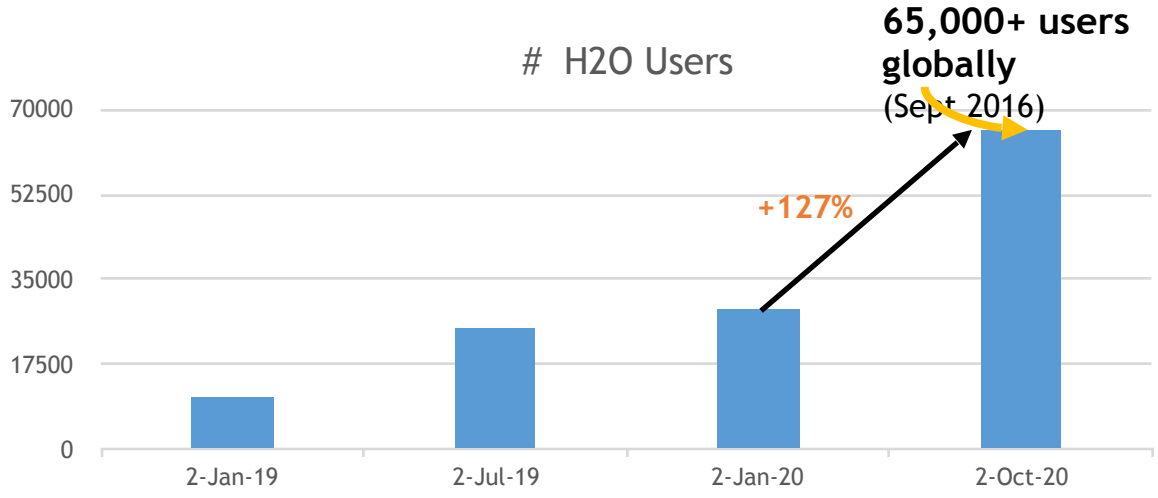
amazon web services

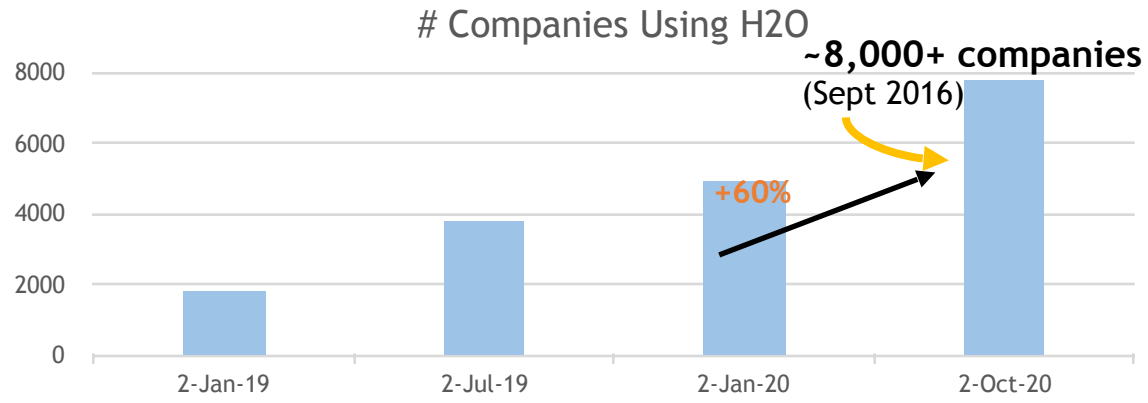
Microsoft Azure

Google

# $H_2O$ Community Growth

## Tremendous Momentum Globally




**Large User Circle**

- 65,000+ users from ~8,000 companies in 140 countries. Top 5 from:


* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

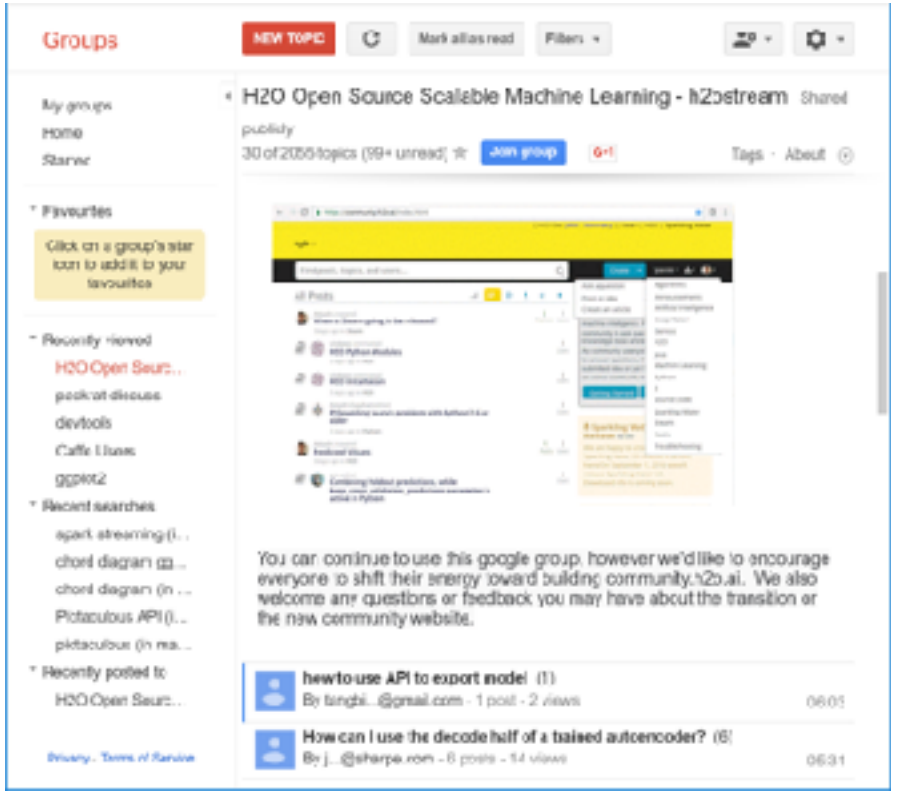


H2O.ai

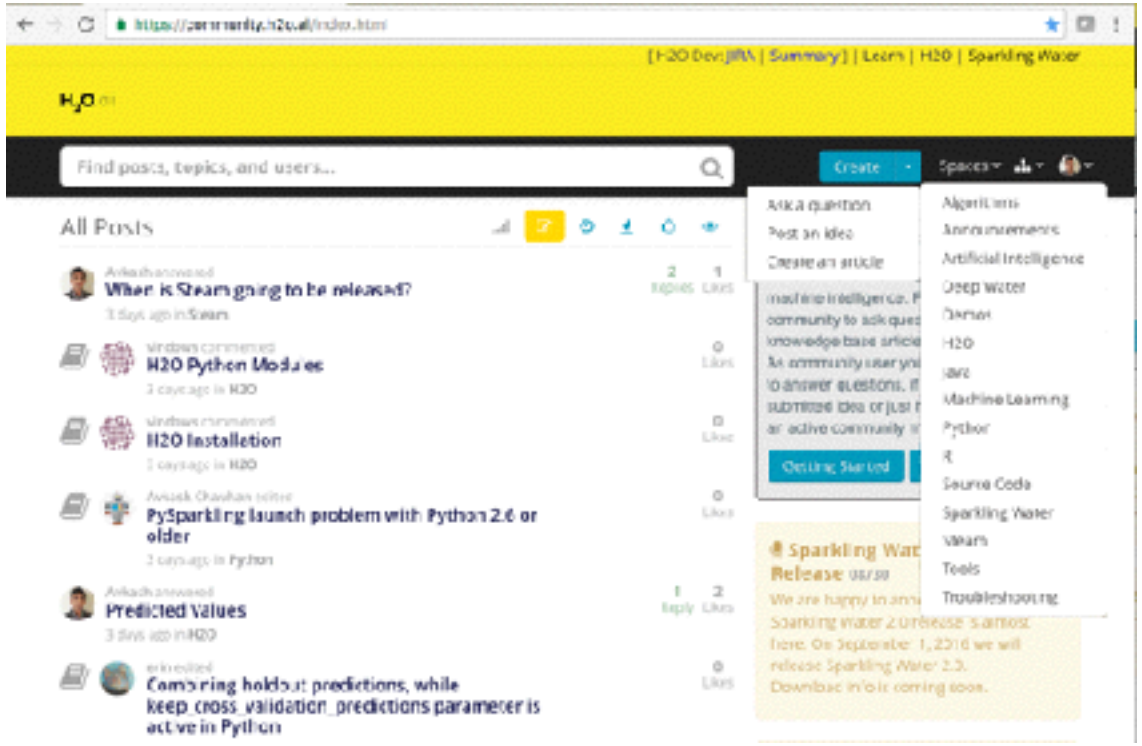
# $H_2O$ Community Support

Please try

## Google forum – h2ostream




## community.h2o.ai



# $H_2O$ for Kaggle Competitions

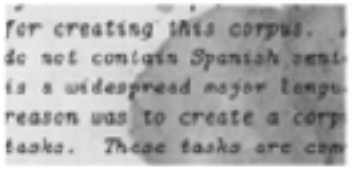
**CIFAR-10 Competition Winners: Interviews with Dr. Ben Graham, Phil Culliton, & Zygmunt Zając**

Triskelion | 01.02.2015

"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."

READ MORE

Kaggle challenge
2nd place winner
Colin Priest




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."

READ MORE

$H_2O$.ai

# $H_2O$ for Academic Research

European Journal of Operational Research

Available online 22 October 2016

In Press, Accepted Manuscript — Note to users

Innovative Applications of O.R.

Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500

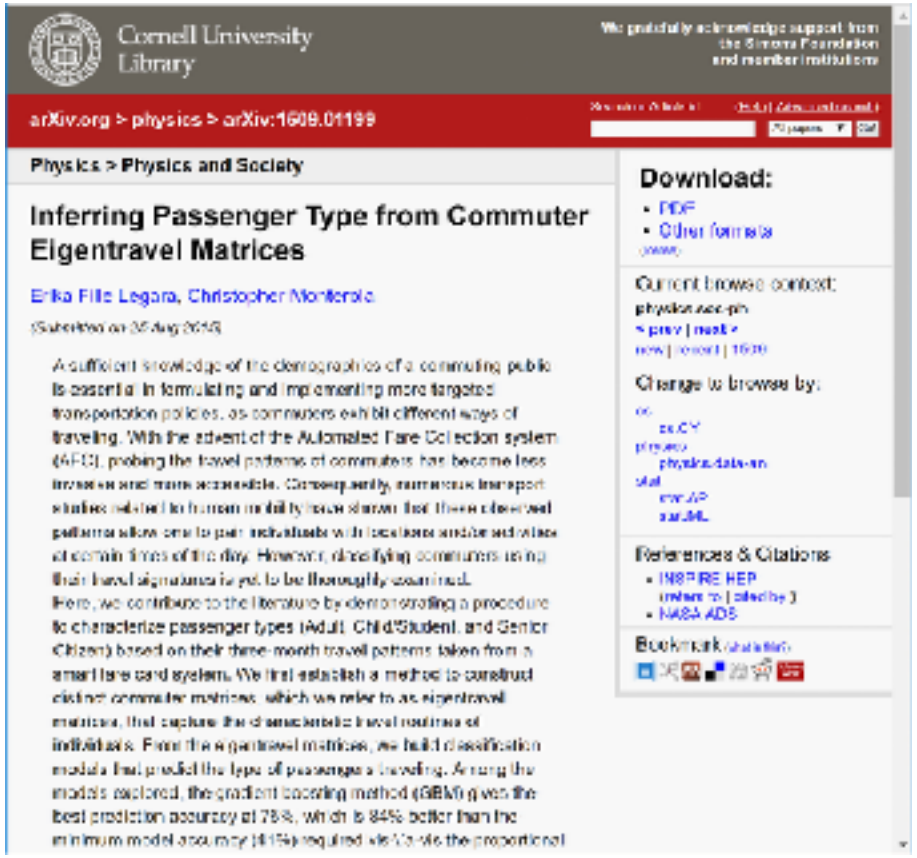
Christopher Krauss, Xuan Anh Do, Nicolas Huck

Received 15 April 2016, Revised 22 August 2016, Accepted 18 October 2016, Available online 22 October 2016

Highlights

- Latest machine learning techniques are deployed in a statistical arbitrage context.
- Deep neural networks, gradient-boosted trees, and random forests are considered.
- An equal-weighted ensemble of these techniques produces the best performance.
- Daily returns are substantial though declining over time.
- The system is especially effective at times of financial turmoil.

http://www.sciencedirect.com/science/article/pii/S0377221716308657



Cornell University Library

arXiv.org > physics > arXiv:1509.01199

Physics > Physics and Society

Inferring Passenger Type from Commuter Eigentravel Matrices

Erika Fille Legara, Christopher Monterola

Download:

- PDF
- Other formats

Current browse context:

Change to browse by:

References & Citations

- INSPIRE HEP
- NASA ADS

Bookmark

https://arxiv.org/abs/1509.01199

$H_2O$

# democratizes

artificial intelligence & big data science

# Our Open Source Products

100% Open Source. Big Data Science for Everyone!

H2O.ai

# $H_2O$.ai Offers AI Open Source Platform

## Product Suite to Operationalize Data Science with Visual Intelligence

**VI**

Visual Intelligence and UX Framework For Data Interpretation and Story Telling on top of Beautiful Data Products

**100% Open Source**

$H_2O$.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

Spark + $H_2O$ SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

# $H_2O$.ai Offers AI Open Source Platform

## Product Suite to Operationalize Data Science with Visual Intelligence

VI

Visual Intelligence and UX Framework For Data Interpretation and Story Telling on top of Beautiful Data Products

**100% Open Source**

$H_2O$.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

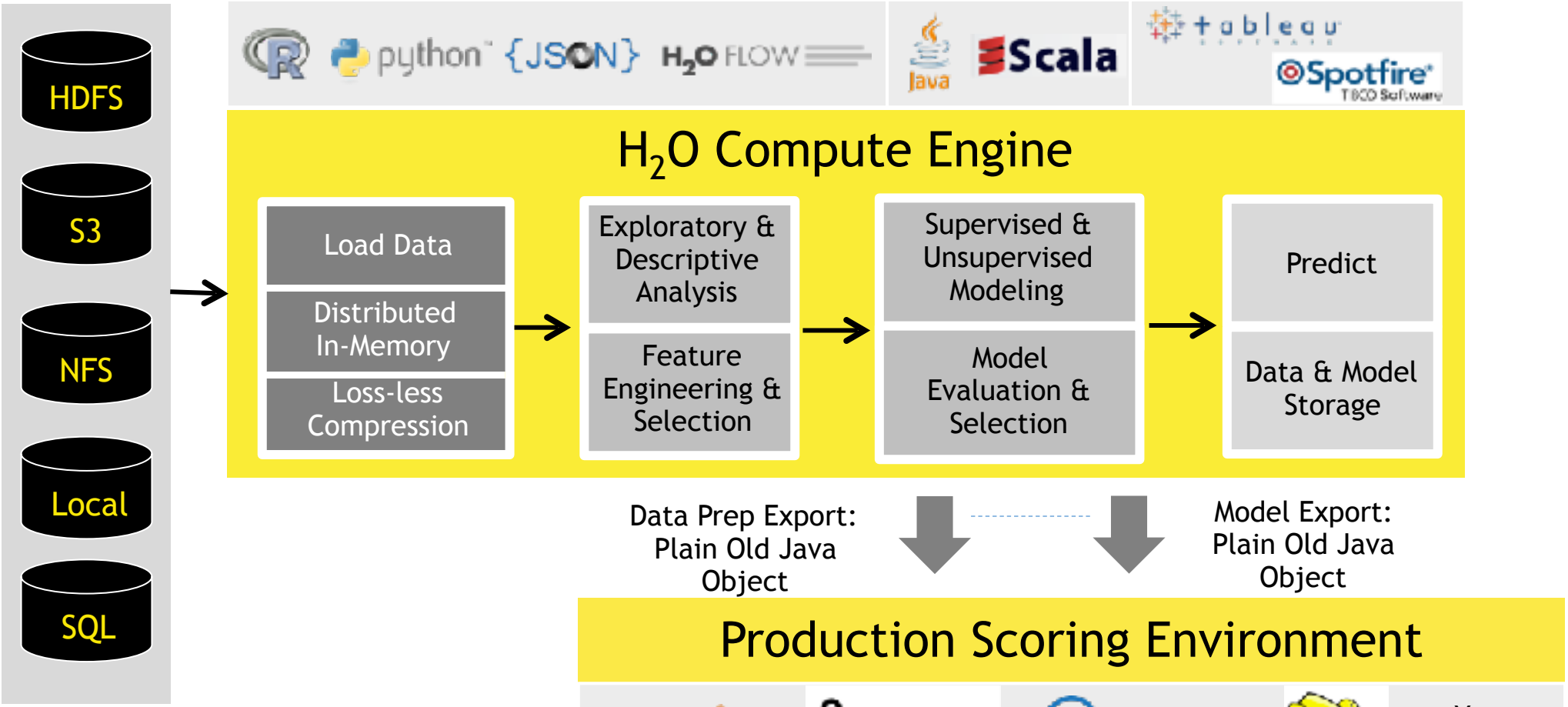
Spark + $H_2O$ SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

# High Level Architecture



# Algorithms Overview

## Supervised Learning

| | |
|---|---|
| Statistical Analysis | • **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie<br>• **Naïve Bayes** |
| Ensembles | • **Distributed Random Forest**: Classification or regression models<br>• **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 |

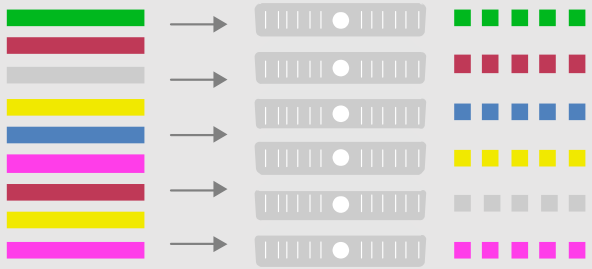
## Unsupervised Learning

| | |
|---|---|
| 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<br>• **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 |

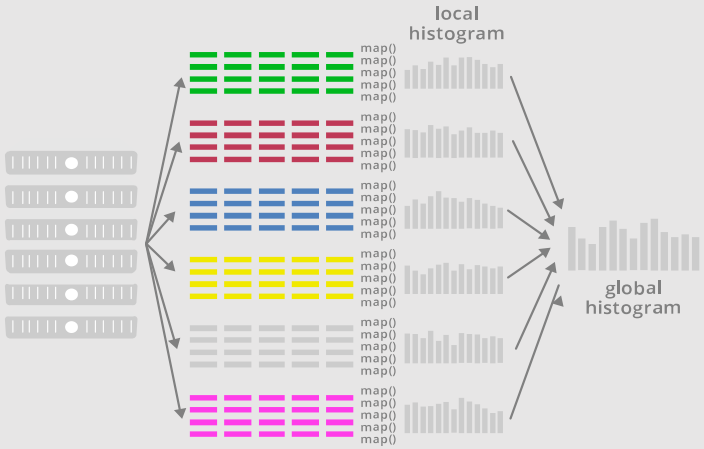
H2O.ai

# Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**




**Fine Grain Map Reduce** Illustration:
Scalable Distributed Histogram Calculation for GBM

## Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - Distributed Data Frames and columnar compression
- All algorithms are distributed in $H_2O$: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

## User Benefits

- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- In-house expertise for all algorithms

# $H_2O$ 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 |
|---|---|---|
| 116495259 | 12 | 2GB |

Job

| | |
|---|---|
| Run Time | 00:00:36.712 |
| Remaining Time | 00:00:17.188 |
| Type | Model |
| Key | deeplearning-dd2f42f7-81f7-42e8-9d98-e34437309828 |
| Description | DeepLearning |
| Status | RUNNING |
| Progress | 69% |

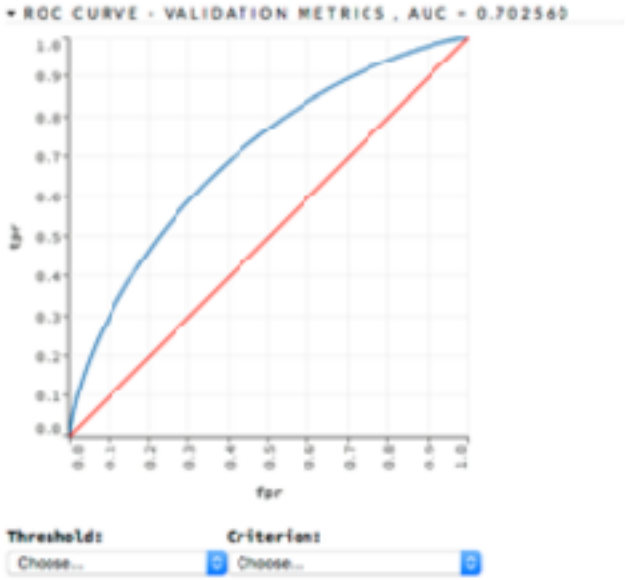
Iterations: 12. Epochs: 0.628821. Speed: 2,243,735 samples/sec. Estimated time left: 21.849 sec

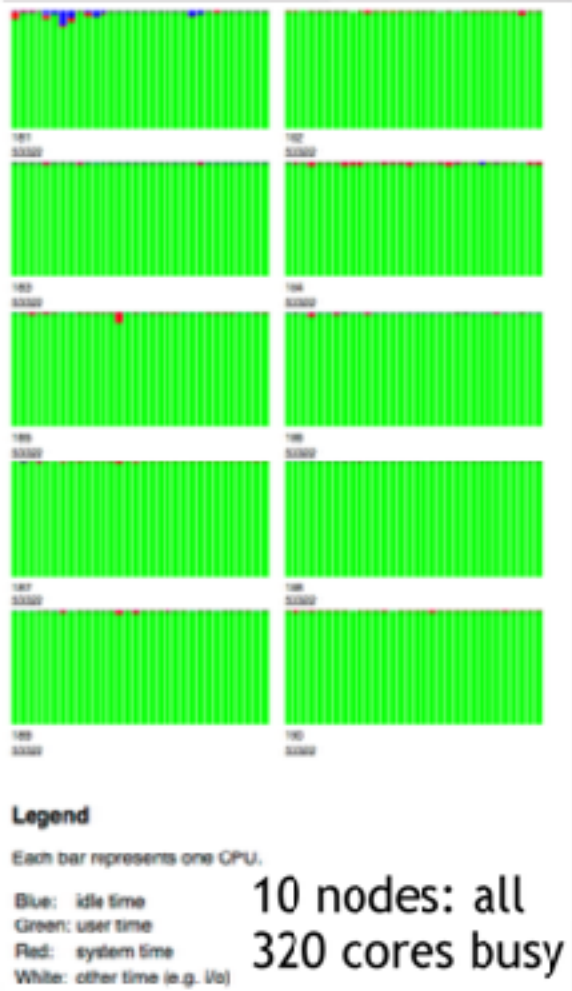
Actions: View | Cancel Job

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

OUTPUT - STATUS OF NEURON LAYERS: PREDICTING ISDEPDELAYED, 2-CLASS CLASSIFICATION, BERNOULLI DISTRIBUTION, CROSSENTROPY LOSS, 17,462 WEIGHTS/BIASES, 221.3 KB, 106,545,365 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 | 887 | Input | 0 | | | | | | | | | |
| 2 | 20 | Rectifier | 0 | 0 | 0 | 0.0403 | 0.2020 | 0 | -0.0021 | 0.2111 | -0.9139 | 1.0036 |
| 3 | 20 | Rectifier | 0 | 0 | 0 | 0.0197 | 0.0227 | 0 | -0.1033 | 0.5062 | -1.3900 | 1.0259 |
| 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 | 0.0161 | 0.0083 | 0 | 0.0741 | 0.7260 | 0.4269 | 0.2056 |




10 nodes: all
320 cores busy

real-time, interactive
model inspection in Flow

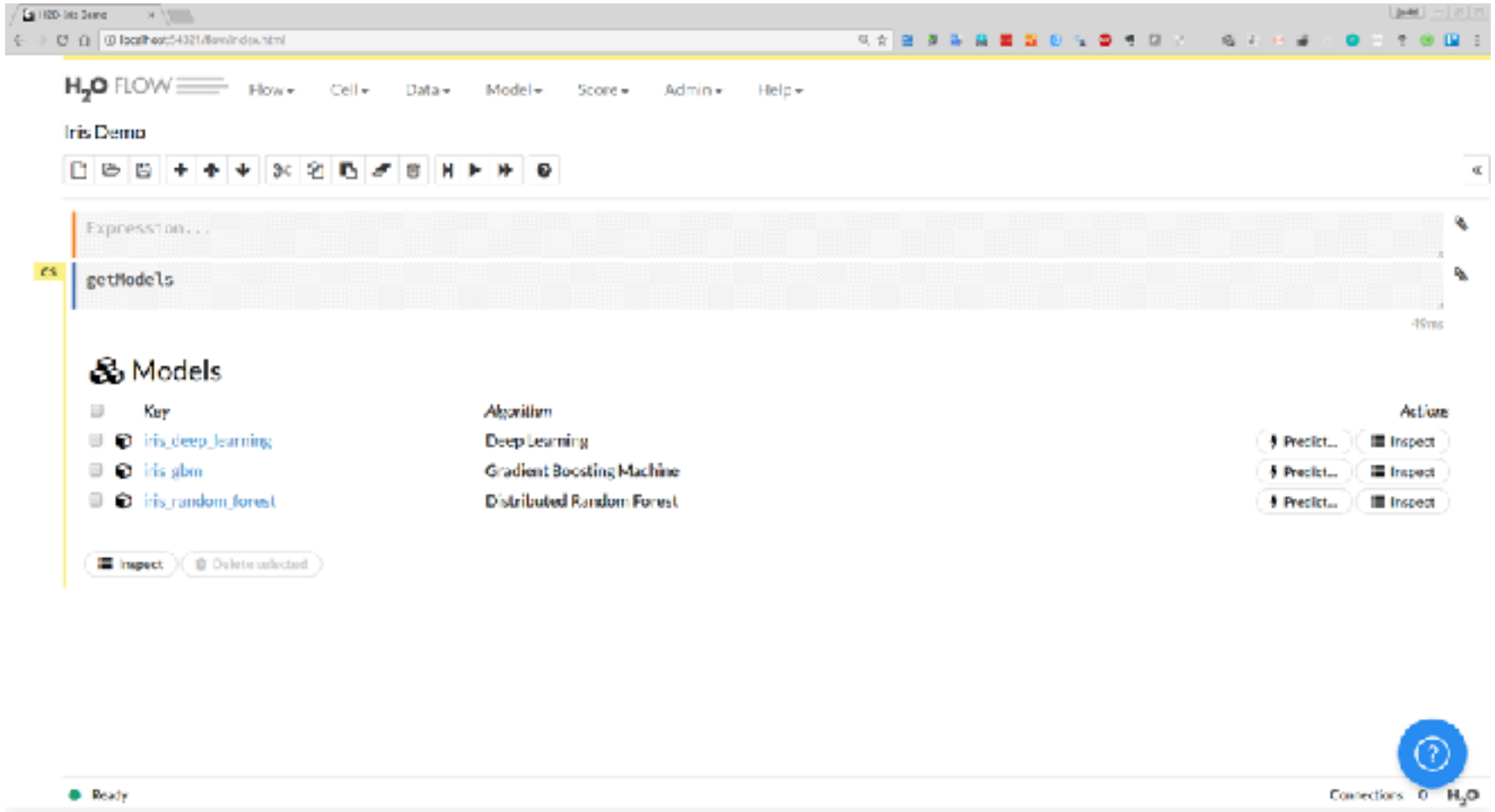
$H_2O$.ai **Deep Learning Model**

# $H_2O$ + R

```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)
```

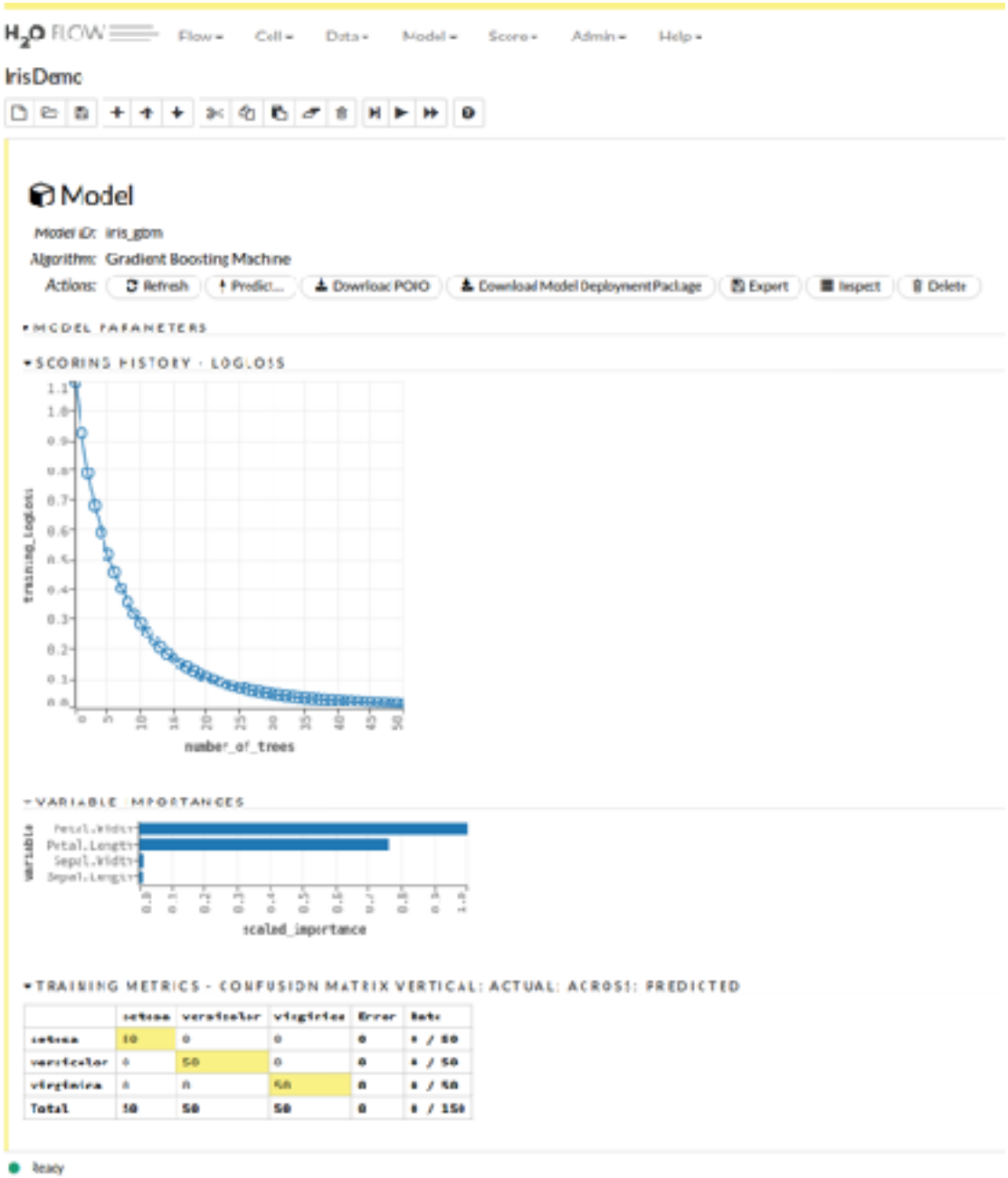
Please try

# $H_2O$ Flow (Web)




H2O.ai

# $H_2O$ Flow

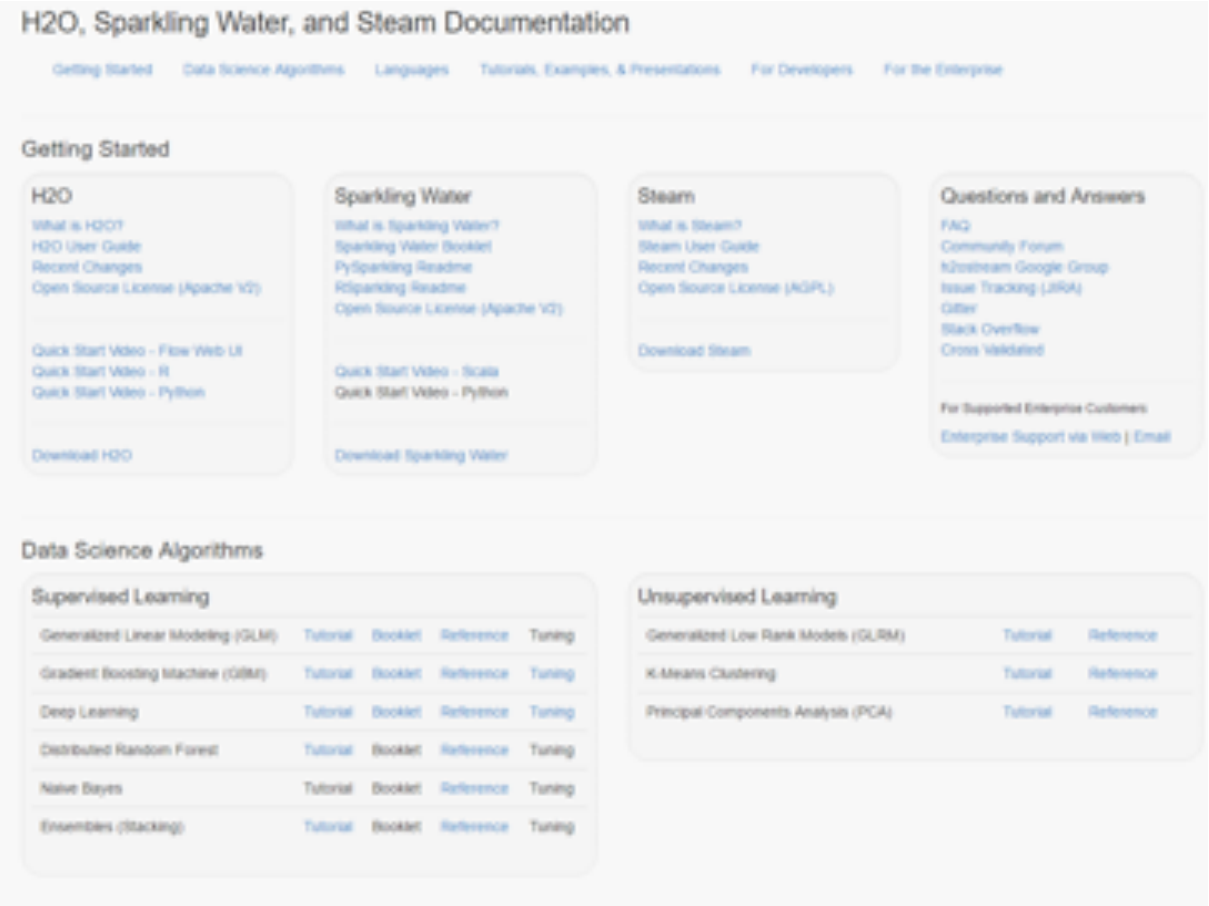





H2O.ai

# Key Learning Resources

- Help Documentations
  - docs.h2o.ai
- Meetups
  - bit.ly/h2o_meetups
- YouTube Channel
  - bit.ly/h2o_youtube



# $H_2O$.ai Offers AI Open Source Platform

## Product Suite to Operationalize Data Science with Visual Intelligence

VI

Visual Intelligence and UX Framework For Data Interpretation and Story Telling on top of Beautiful Data Products

**100% Open Source**

$H_2O$.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

Spark + $H_2O$ SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

# 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

usage of Deep Learning tools in past year

| | |
|---|---|
| Deep Learning Tools & Platforms: Caffe (63) | 0% |
| Cuda-convnet (22) | 0% |
| Convnet.js (8) | 0% |
| carch (2) | 0% |
| Deeplearning4j (46) | 0% |
| H2O (190) | 1% |
| MATLAB Deep Learning Toolbox (57) | 0% |
| Microsoft CNTK (25) | 0% |
| mxnet (16) | 0% |
| Nervana (2) | 0% |
| Tensorflow (190) | 1% |
| Theano ecosystem including Keras, Lasagne, Pylearn2 (140) | 1% |
| Torch (28) | 0% |
| Veles (2) | 0% |
| Other Deep Learning Tools (104) | 1% |

TensorFlow democratizes the power of deep learning.

$H_2O$ democratizes artificial intelligence & big data science.

There are other open source libraries like MXNet and Caffe too.

Let's have a party, this will be fun!

H2O.ai

# 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**

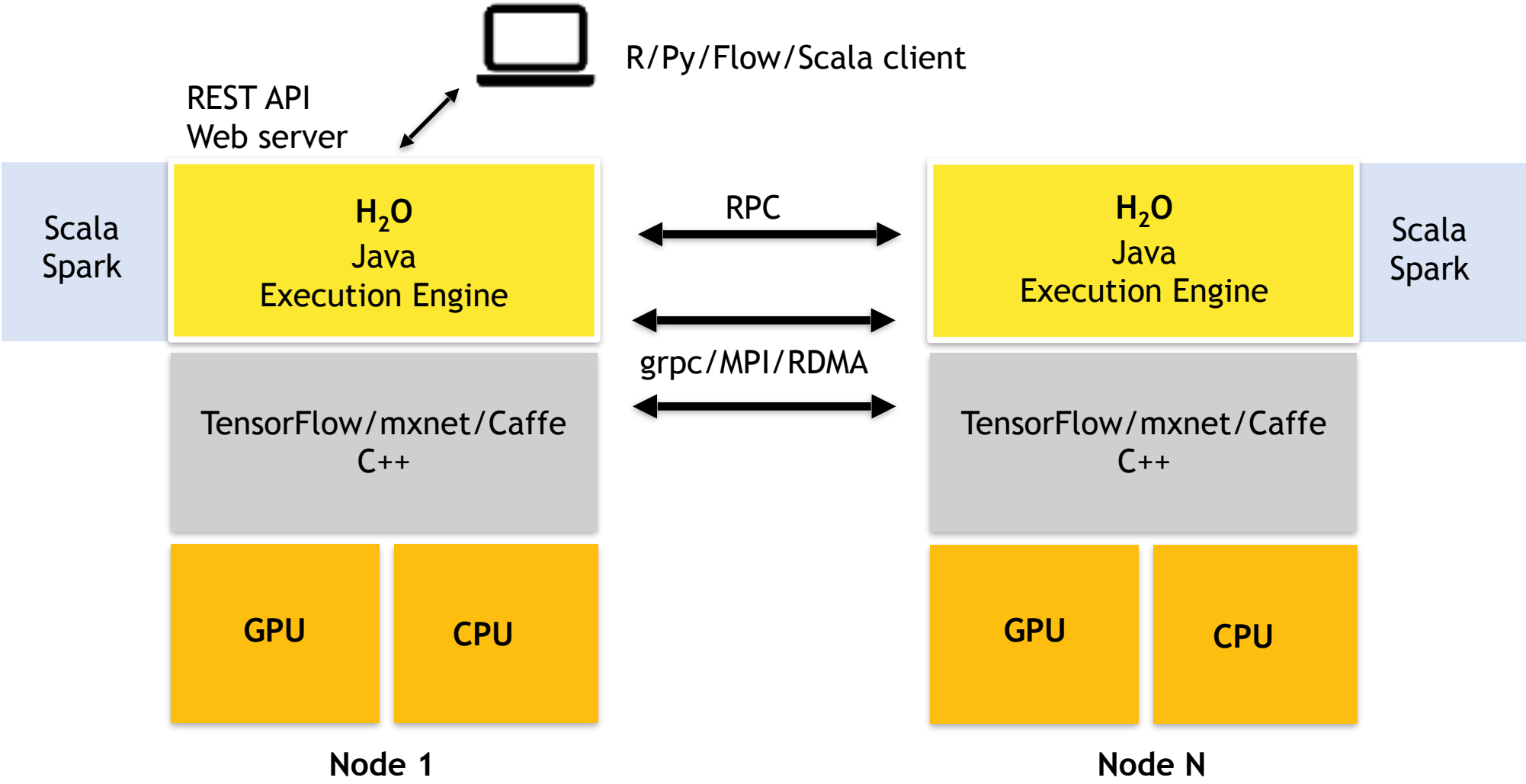
Convolutional Neural Networks enabling **Image, video, speech recognition**

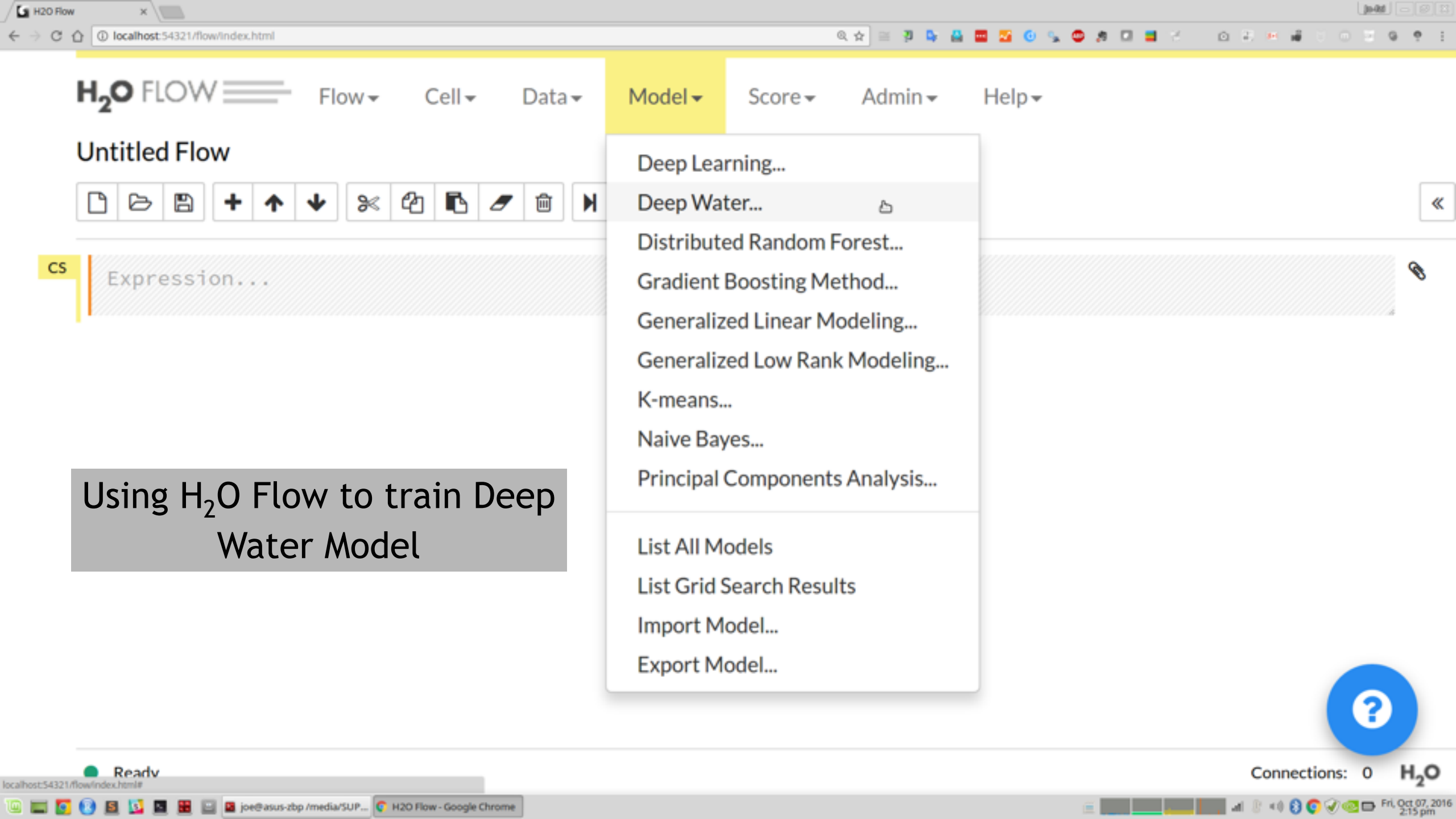
Hybrid Neural Network Architectures enabling **speech to text translation, image captioning, scene parsing** and more


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

$H_2O$.ai

# Deep Water Architecture






Using $H_2O$ Flow to train Deep Water Model

# Same H2O R/Python Interface

To build a LeNet image classification model in H2O, simply specify network = "lenet":

```
model <- h2o.deepwater(x=path, y=response,
                       training_frame=df, epochs=50,
                       learning_rate=1e-3, network = "lenet")
model
```

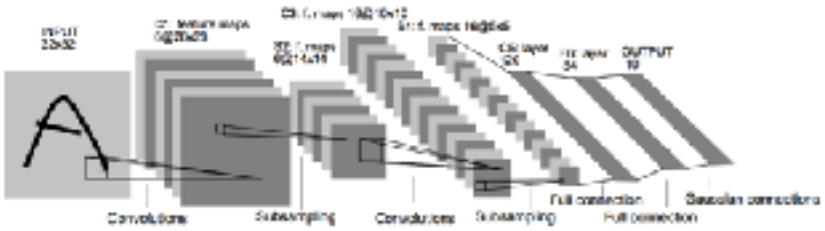
```
  |======================================================================| 100%
Model Details:
==============

H2OMultinomialModel: deepwater
Model ID:  DeepWater_model_R_1477378362430_2
Status of Deep Learning Model: lenet, 1.6 MB, predicting C2, 3-class classification, 14,336 training sampl
es, mini-batch size 32
  input_neurons     rate momentum
1          2352 0.000986 0.990000


H2OMultinomialMetrics: deepwater
** Reported on training data. **
** Metrics reported on full training frame **

Training Set Metrics:
=====================

Extract training frame with `h2o.getframe("cat_dog_mouse.hex_sid_95f8_1")`
MSE: (Extract with `h2o.mse`) 0.131972
RMSE: (Extract with `h2o.rmse`) 0.3520386
Logloss: (Extract with `h2o.logloss`) 0.4176429
```



H2O.ai

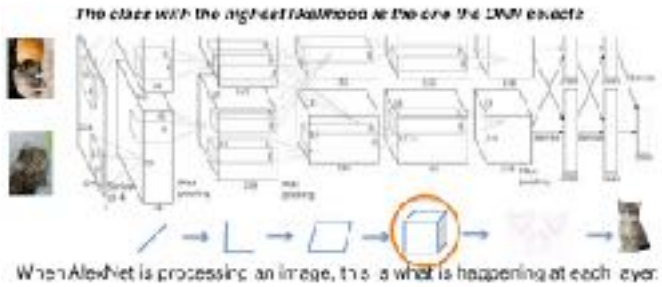
# Available Networks in Deep Water

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

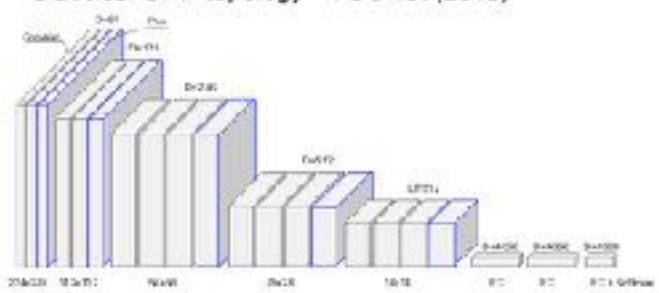


CNN called LeNet by Yann LeCun (1998)

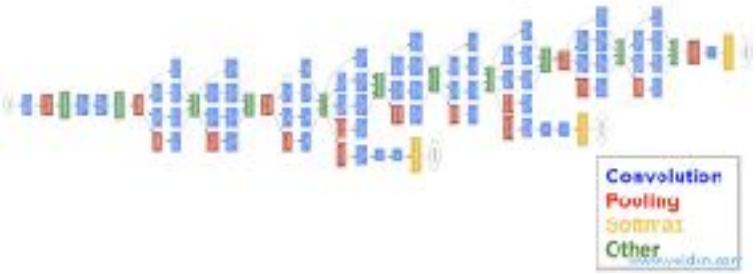
AlexNet (Krizhevsky et al. 2012)



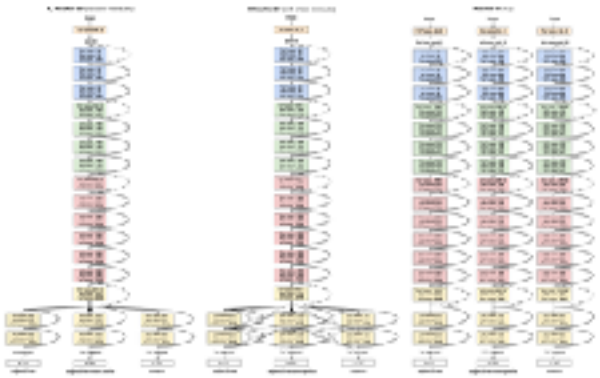
Classical CNN topology - VGGNet (2013)



GoogLeNet

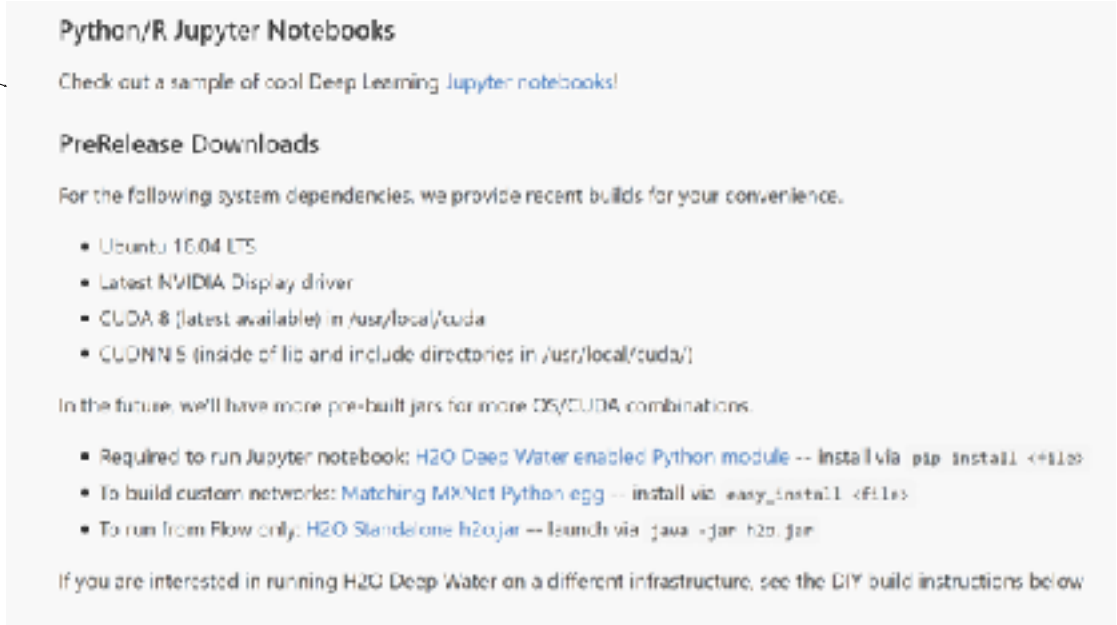



ResNet



# Want to try Deep Water?

- Build it
  - github.com/h2oai/deepwater
  - Ubuntu 16.04
  - CUDA 8
  - cuDNN 5
  - …
- Pre-built Amazon Machine Images (AMIs)
  - Info to be confirmed



### Python/R Jupyter Notebooks

Check out a sample of cool Deep Learning Jupyter notebooks!

### PreRelease Downloads

For the following system dependencies, we provide recent builds for your convenience.

- Ubuntu 16.04 LTS
- Latest NVIDIA Display driver
- CUDA 8 (latest available) in /usr/local/cuda
- CUDNN 5 (inside of lib and include directories in /usr/local/cuda/)

In the future, we'll have more pre-built jars for more OS/CUDA combinations.

- Required to run Jupyter notebook: H2O Deep Water enabled Python module -- install via `pip install <file>`
- To build custom networks: Matching MXNet Python egg -- install via `easy_install <file>`
- To run from Flow only: H2O Standalone h2o.jar -- launch via `java -jar h2o.jar`

If you are interested in running H2O Deep Water on a different infrastructure, see the DIY build instructions below

# $H_2O$'s Mission


## Making Machine Learning Accessible to Everyone

*Photo credit: Virgin Media*

H2O.ai