

# H2O Machine Learning Use Cases



satRday #1  
MTA TTK, Budapest, Hungary  
3<sup>rd</sup> September, 2016

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

# About Me: Civil Engineer → Data Scientist

- 2005 - 2015
- Water Engineer
  - Consultant for Utilities
  - EngD Research
- 2015 - Present
- Data Scientist
  - Virgin Media (UK)
  - Domino Data Lab (US)
  - H2O.ai (US)

Why? Long story – see [bit.ly/joe\\_h2o\\_talk2](http://bit.ly/joe_h2o_talk2)

# About this Talk

- Introduction to H2O
- Machine learning use cases from our users



# About H2O.ai

- **H2O.ai, the Company**
  - Team: 80 (71 shown)
  - Founded in 2012,
  - HQ: Mountain View, California
- **H2O, the Platform**
  - Open Source (Apache 2.0)
  - R, Python, Scala, Java and Web Interfaces
  - Distributed Algorithms that Scale to Big Data
  - Works with Laptop, Hadoop & Spark



# Current Algorithm Overview

satRday  
Workshop

## Statistical Analysis

- Linear Models (GLM)
- Naïve Bayes

## Ensembles

- Random Forest
- Distributed Trees
- Gradient Boosting Machine
- R Package - Stacking / Super Learner

## Deep Neural Networks

- Multi-layer Feed-Forward Neural Network
- Auto-encoder
- Anomaly Detection
- Deep Features

## Clustering

- K-Means

## Dimension Reduction

- Principal Component Analysis
- Generalized Low Rank Models

## Solvers & Optimization

- Generalized ADMM Solver
- L-BFGS (Quasi Newton Method)
- Ordinary Least-Square Solver
- Stochastic Gradient Descent

## Data Munging

- Scalable Data Frames
- Sort, Slice, Log Transform

# H2O Interfaces – R, Python & Others

- R

```
1 # Load H2O R package
2 library(h2o)
3
4 # Initialize and Connect to H2O
5 h2o.init()
```

- Resources - [docs.h2o.ai](https://docs.h2o.ai)

## H2O and Sparkling Water Documentation

### Getting Started

**H2O**  
What is H2O  
Open Source License (Apache V2)  
Download H2O  
H2O User Guide  
Recent Changes

Quick Start Video - Flow Web UI  
Quick Start Video - R  
Quick Start Video - Python

**Sparkling Water**  
What is Sparkling Water?  
Open Source License (Apache V2)  
Download Sparkling Water  
Sparkling Water Booklet  
PySparkling Readme

Quick Start Video - Scala  
Quick Start Video - Python

**Questions and Answers**  
FAQ  
H2OCommunity Forum  
Issue Tracking (JIRA)  
GitHub  
Stack Overflow  
Cross Validated

## Data Science Algorithms

### Supervised Learning

Generalized Linear Modeling (GLM)

Gradient Boosting Machine (GBM)

Deep Learning

Distributed Random Forest

Naïve Bayes

Ensembles (Stacking)

Tutorial Booklet Reference

### Unsupervised Learning

Generalized Low Rank Models (GLRM)

K-Means Clustering

Principal Components Analysis (PCA)

Tutorial Reference

Tutorial Reference

Tutorial Reference

## Languages

### R

Quick Start Video - R  
R Package Docs  
R Booklet  
Examples and Demos  
R FAQs  
Migrating from H2O-2

Quick Start Video - Python  
Python Module Docs  
Python Booklet

Examples and Demos  
Python API  
PySparkling Readme

### Java

Pojo Model Javadoc  
H2O Core Javadoc  
H2O Algorithms Javadoc

Sparkling Water API

Sparkling Water Scaladoc  
H2O Scaladoc

- Python

```
1 # Import H2O Python module
2 import h2o
3
4 # Initialize and Connect to H2O
5 h2o.init()
```

# H2O R Code Example

```
42 # Train a GBM with default values  
43 model_gbm ← h2o.gbm(x = features,  
44                      y = target,  
45                      training_frame = h2o_df_boston)  
46  
47 # First look  
48 print(model_gbm)
```

Slide & Code for Workshop:

[bit.ly/h2o\\_budapest\\_workshop](http://bit.ly/h2o_budapest_workshop)

# H2O Interfaces – Web (H2O Flow)

H2O FLOW Flow Cell Data Model Score Admin Help

DeepLearning\_MNIST

Model ID: deeplearning-d5c35043-8929-441a-9a23-dc44b06b519f  
Algorithm: Deep Learning  
Actions: Refresh Predict Download POJO Export Inspect Delete

Model Parameters

Scoring History - LogLoss

Scoring History - MSE

Variable Importances

Training Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	993	0	0	0	0	0	0	0	0	0	0	0 / 993
1	0	1105	0	0	0	0	0	0	0	0	0	0 / 1,105

Validation Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	970	0	0	0	0	0	0	0	0	0	0	0 / 970
1	0	1125	4	0	1	2	0	3	0	0	0	0.0008 10 / 1,125

Output

Status of Neuron Layers (Predicting C795, 10-class classification, multinomial distribution, crossentropy loss, 100,810 weights/biases, 899.2 KB, 9,240,000 training samples, mini-batch size: 1)

Scoring History

Training Metrics

Validation Metrics

Validation Metrics - Top 10 Hit Ratios

Validation Metrics - Top 10 Hit Ratios

Variable Importances

Preview POJO

Ready

H2O FLOW Flow Cell Data Model Score Admin Help

DeepLearning\_MNIST

OUTLINE FLOWS CLIPS HELP

Help

examples

- GBM\_Example.flow
- DeepLearning\_MNIST.flow
- GLM\_Example.flow
- DRF\_Example.flow
- K-Means\_Example.flow
- Million\_Songs.flow
- KDDCup2009\_Churn.flow
- QuickStartVideos.flow
- Airlines\_Delay.flow
- GBM\_Airlines\_Classification.flow
- GBM\_GridSearch.flow
- RandomData\_Benchmark\_Small.flow

Scoring History - LogLoss

Training Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	993	0	0	0	0	0	0	0	0	0	0	0 / 993
1	0	1105	0	0	0	0	0	0	0	0	0	0 / 1,105

Validation Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	970	0	0	0	0	0	0	0	0	0	0	0 / 970
1	0	1125	4	0	1	2	0	3	0	0	0	0.0008 10 / 1,125

Output

Status of Neuron Layers (Predicting C795, 10-class classification, multinomial distribution, crossentropy loss, 100,810 weights/biases, 899.2 KB, 9,240,000 training samples, mini-batch size: 1)

Scoring History

Training Metrics

Validation Metrics

Validation Metrics - Top 10 Hit Ratios

Validation Metrics - Top 10 Hit Ratios

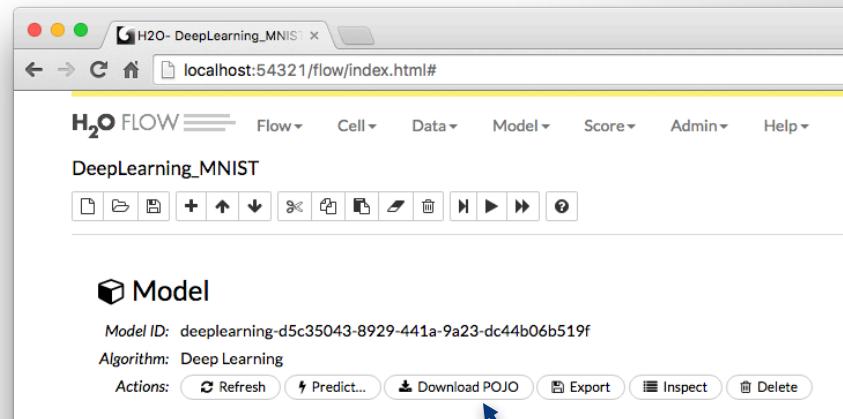
Variable Importances

Preview POJO

Ready

# Export Plain Old Java Object (POJO)

The screenshot shows the H2O Flow interface with the title "DeepLearning\_MNIST". The main area displays the generated Java code for a "DeepLearningModel" named "deeplearning\_d5c35043\_8929\_441a\_9a23\_dc44b06b519f". The code includes imports for java.util.Map, hex.gm.GemModel, and hex.gm.annotations.ModelPojo; a class definition for "DeepLearningModel"; and a static final class "NORMALUL" that implements java.io.Serializable. The "NORMALUL" class contains a static final double[] "VALUES" with 26 elements, each with a value like 0.1838291371915183. Below the code, a note says: "How to download, compile and execute: mkdir tmpdir cd tmpdir curl -O http://127.0.0.1:54321/h2o-genmodel.jar > h2o-genmodel.jar curl -O http://127.0.0.1:54321/H2olets.java</DeepLearning-d5c35043\_8929\_441a\_9a23\_dc44b06b519f> > deeplearning.java javac -cp h2o-genmodel.jar -D Xmx2g -D XX:MaxPermSize=128m deeplearning\_d5c35043\_8929\_441a\_9a23\_dc44b06b519f (Note: Try java argument -XX:+PrintCompilation to show runtime JIT compiler behavior.)". The bottom status bar says "Ready" and "Connections: 0".



# Use Cases

# Telenor – Why H2O?

## WHAT WERE THE MAIN ASPECTS WE VALUED IN A ML SOLUTION IN 2015?

	R	Spark ML 1.3	Radoop	H <sub>2</sub> O
Easy to work with	:(	:(	:)	:)
Cost efficient	:)	:)	:(	:)
Provide the methods that we normally use	:)	:(	:)	:)
Easy-to-use real-time capability	:(	:(	:)	:)
Can utilize our hardware setup	:(	:)	:)	:)



# Telenor – Use Case

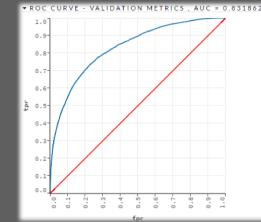
## USE CASE: CHURN MODELLING

Modelling on 100 GB of data.

- 40 M rows
- 551 columns

Results:

- GBM
- AUC → 0.83
- LIFT top 1% → 20



# H2O in Brazil



Rommel Carvalho  
@rommelncln



Estimating students performance in order to decrease student's probability of failing with @h2oai #BrasilDigital



Rommel Carvalho  
@rommelncln



@h2oai and after parameter tuning the results got a lot better! 93% AUC



# R + H2O + Spark at EARL 2016

EARL 2016 London

About    Agenda    Speakers    Sponsors & Exhibitors    Workshops    Venue    Register    Contact    [EARL Boston](#)



Vincent Warmerdam |  GoDataDriven

I'm a data scientist at GoDataDriven in Amsterdam where I help companies get better at being a company using data. I'm also a preferred Rstudio training partner and co-chair of PyData Amsterdam. I also have a blog that is doing well over at koaning.io.



## ML for SparkR: just add water

The SparkR project allows more scalable tools for R users that work well with the Rstudio stack as well as the Hadoop stack. The ML capabilities however are somewhat lacking in features.

In this small talk I'll demo how to provision SparkR with H2O allowing more algorithms to the R users of Spark. I'll discuss the pros and cons of this approach and demo a few shiny features and use cases of the H2O stack.

Some of the themes I'll discuss;

- ease of provisioning
- works on python stack as well
- your model will be saved as a \*.jar which makes it easy for deployment
- grid search is a feature
- deep models on spark

# Our Customers

The screenshot shows a grid of four testimonial cards. Each card features a portrait of a customer, their name, title, and company, followed by a 'WATCH VIDEO' button.

- 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."
- 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."
- 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."
- Prateem Mandal**  
Technical Lead Architect  
MarketShare  
"H2O gave us the capability of what we call big modeling, which means that the amount of data that math can ingest and process is now not limited by the capability of a single node...there is no limit to scaling in H2O"

The screenshot shows a grid of four testimonial cards. Each card features a portrait of a customer, their name, title, and company, followed by a 'WATCH VIDEO' button.

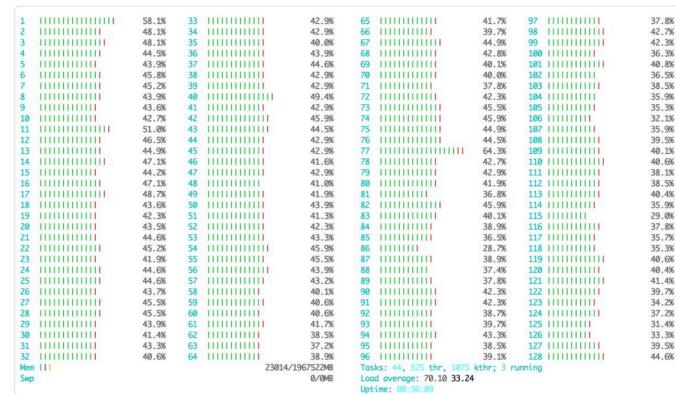
- Brendan Herger**  
Data Scientist  
Capital One  
"We evaluated a large number of hard and soft metrics. H2O just scored really well with all of these areas, particularly relative to a lot of the machine learning frameworks that are available at the moment."
- Edward Agarwala**  
Data Scientist  
Progressive 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."
- 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."
- Prateem Mandal**  
Technical Lead Architect  
MarketShare  
"H2O gave us the capability of what we call big modeling, which means that the amount of data that math can ingest and process is now not limited by the capability of a single node...there is no limit to scaling in H2O"

# Szilard Pafka

- Budapest DS Meetup
  - Big thanks to Szilard Pafka
    - Intro to ML with H2O
  - [Link \(video\)](#)
  - [Link \(Slides\)](#)

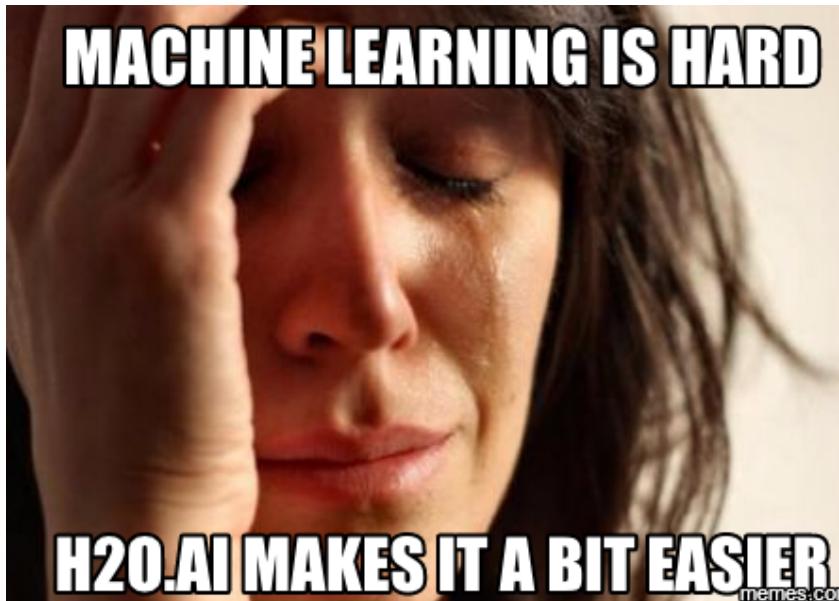


Also [@h2oai](#) on monster 2TB RAM 128 cores  
EC2 X1 #bigdata #machinelearning  
[#datascience](#) [twitter.com/DataScienceLA/...](#)



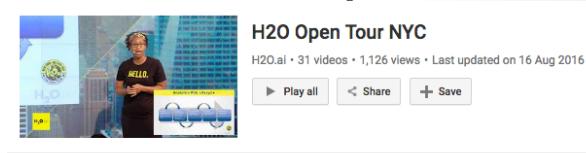
# Szilard – Why H2O?

- Szilard's Summary Slide



# H2O is Evolving

- Advanced data munging
  - Visual ML
  - Deep Water
    - H2O + TensorFlow, mxnet, Theano, Caffe ...
    - GPU
  - Steam
- YouTube Playlist

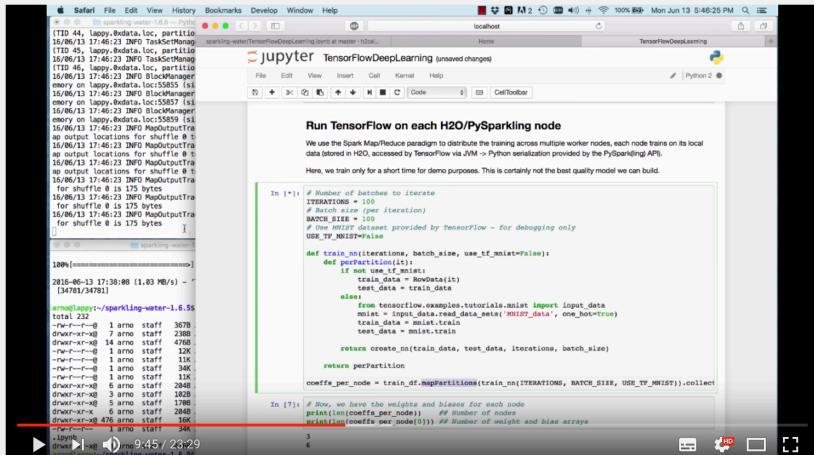


**H2O Open Tour NYC**  
H2O.ai • 31 videos • 1,126 views • Last updated on 16 Aug 2016  
▶ Play all   < Share   + Save

Rank	Video Title	Uploader
1	Migrating from Closed Source to Open Source with Ken Sanford & Fonda Ingram	by H2O.ai
2	H2O Open Tour: NYC - Opening Keynote From CEO Sri Ambati	by H2O.ai
3	Advancements in H2O with Arno Candel	by H2O.ai
4	Steam Product Demo with Bill Gallmeister	by H2O.ai
5	Advanced Munging in H2O with Matt Dowle	by H2O.ai
6	Sparkling Water 2.0 with Tom Kraljevic	by H2O.ai
7	Visual Machine Learning with Tony Chu	by H2O.ai

# H2O Integration with Other Libraries

- H2O + TensorFlow demo



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell contains Python code for training a TensorFlow model on H2O data. The second cell shows the output of the training process, which includes logs from multiple worker nodes and a progress bar indicating the completion of the training.

```
In [1]: # Number of batches to iterate
ITERATIONS = 100
# Batch size (per iteration)
BATCH_SIZE = 100
# MNIST dataset provided by TensorFlow - for debugging only
USER_TV_MNIST=True

def train_nn(iterations, batch_size, use_tf_mnist=False):
    del iterations, batch_size
    if not use_tf_mnist:
        train_data = RowDataList()
        test_data = RowDataList()
    else:
        from tensorflow.examples.tutorials.mnist import input_data
        mnist = input_data.read_data_sets('MNIST_data', one_hot=True)
        train_data = mnist.train
        test_data = mnist.train

    return create_nn(train_data, test_data, iterations, batch_size)

def create_nn(train_data, test_data, iterations, batch_size):
    coeffs_per_node = train_data.mapPartitions(lambda x: len(x)).collect()

    In [7]: # Now, we have the weights and biases for each node
print(len(coeffs_per_node)) # Number of nodes
print(len(coeffs_per_node[0])) # Number of weight and bias arrays
```

Output of the code:

```
[2016-06-13 17:38:00 (1.03 MB/s) - (3478/3478)]
```

```
INFO:sparkling-water-1.6.55
total 225
-rw-r--r-- 0 1 arno staff 367B
drwxr-xr-x 0 1 arno staff 238B
drwxr-xr-x 0 1 arno staff 144B
-rw-r--r-- 0 1 arno staff 12K
-rw-r--r-- 0 1 arno staff 11K
-rw-r--r-- 0 1 arno staff 34K
-rw-r--r-- 0 1 arno staff 11K
drwxr-xr-x 0 1 arno staff 100B
drwxr-xr-x 0 3 arno staff 182B
drwxr-xr-x 0 6 arno staff 290B
drwxr-xr-x 0 476 arno staff 16K
-rw-r--r-- 0 1 arno staff 34K
```

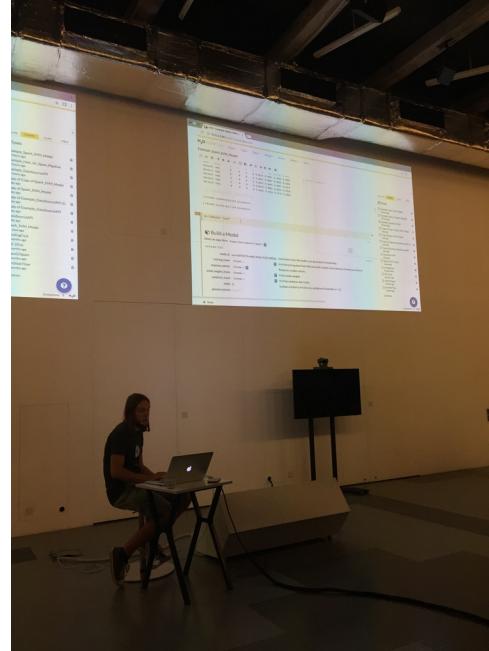
Bottom status bar: 1.03 MB/s, 9:45 / 23:29

H2O TensorFlow Deep Learning Demo

H2O.ai  Subscribed 1,860

1,769 views

- H2O + MLLib SVM demo



# We're Hiring!

- Check it out
- [www.h2o.ai/careers/](http://www.h2o.ai/careers/)

## Open Positions

Location	Position
Mountain View, CA	<a href="#">UI Engineers</a>
Mountain View, CA	<a href="#">Algorithm Engineers</a>
Mountain View, CA	<a href="#">Solutions Architect, Data Engineering</a>
Mountain View, CA	<a href="#">Distributed Systems Platform Engineer</a>
Mountain View, CA	<a href="#">Customer Support Manager</a>
Mountain View, CA	<a href="#">Quality Engineer</a>
Multiple	<a href="#">Program Manager</a>
Multiple	<a href="#">Solutions Data Scientist</a>
Multiple	<a href="#">Data Journalist</a>

To apply, send your resume to [careers@h2o.ai](mailto:ccareers@h2o.ai)

# Thanks!

- Sponsors & Organizers
  - satRdays
  - Budapest Data Science Meetup
- Contact
  - [joe@h2o.ai](mailto:joe@h2o.ai)
  - [@matlabulous](https://twitter.com/matlabulous)
  - [github.com/woobe](https://github.com/woobe)
- Resources
  - [docs.h2o.ai](https://docs.h2o.ai)
- Slides and Code
  - [github.com/h2oai/h2o-meetups](https://github.com/h2oai/h2o-meetups)