

Introduction to Machine Learning with H2O



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

The Big Data League Data Hackathon
Stamford Bridge, Home of Chelsea FC
7th September, 2016

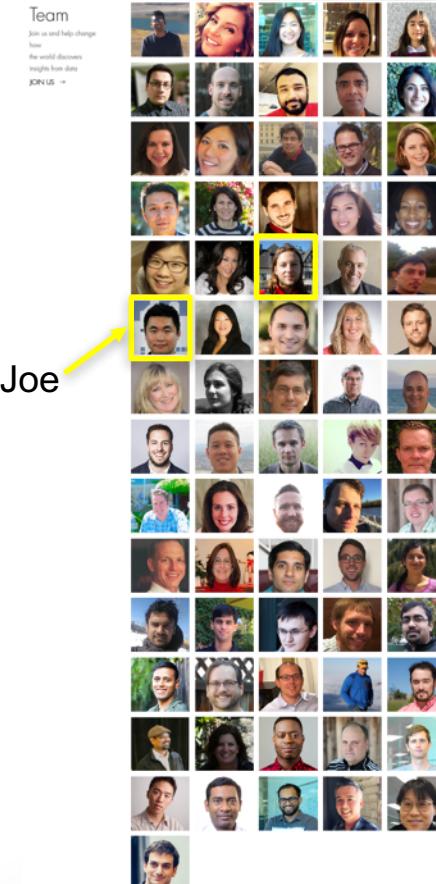
About Me: Civil Engineer → Data Scientist

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

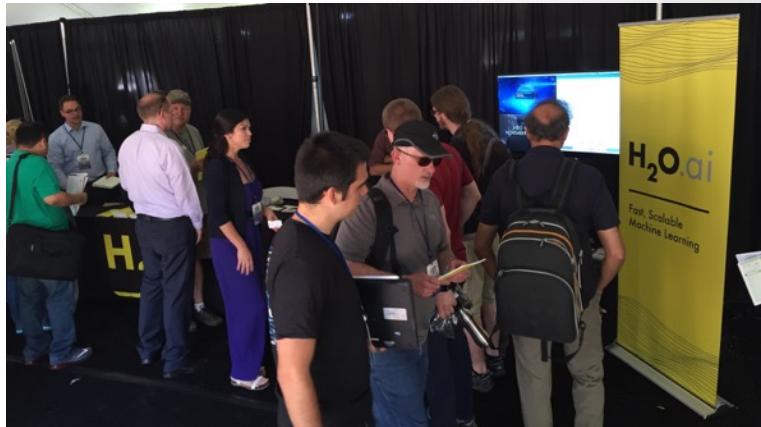
Why? Long story – see bit.ly/joe_h2o_talk2

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



H2O.ai & Stanford University



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*



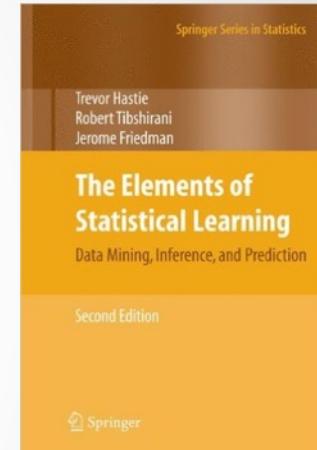
Dr. Robert Tibshirani

- 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*



Current Algorithm Overview

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

Joe's Strata Hadoop
London Talk
bit.ly/joe_h2o_talk4

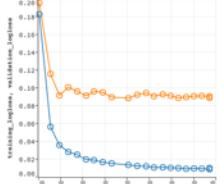
H2O Interfaces – Web (H2O Flow)

H2O FLOW Flow Cell Data Model Score Admin Help

DeepLearning_MNIST

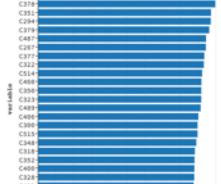
Model
Model ID: deeplearning-d5c35043-8929-441a-9a23-dc44b06b519f
Algorithm: Deep Learning
Actions: Refresh Predict Download POJO Inspect Delete

MODEL PARAMETERS

SCORING HISTORY - LOGLOSS


epoch	validation_logloss
0	0.20
10	0.15
20	0.10
30	0.08
40	0.07
50	0.06
60	0.055
70	0.05
80	0.045
90	0.04
100	0.04

SCORING HISTORY - MSE

VARIABLE IMPORTANCES


variable	scaled_importance
C291	0.15
C292	0.14
C293	0.13
C294	0.12
C295	0.11
C296	0.10
C297	0.09
C298	0.08
C299	0.07
C300	0.06
C301	0.05
C302	0.04
C303	0.03
C304	0.02
C305	0.01
C306	0.005
C307	0.002
C308	0.001
C309	0.0005
C310	0.0002
C311	0.0001
C312	0.00005

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 / 993
1	0	1185	0	0	0	0	0	0	0	0	0 / 1,185
2	0	0	941	0	0	0	0	0	0	0	0 / 941
3	1	0	946	0	1	0	2	1	0	0	0.0063 / 6 / 954
4	0	2	0	0	944	0	1	1	0	0	0.0052 / 2 / 957
5	0	0	1	0	0	913	1	0	0	0	0.0052 / 2 / 935
6	0	0	0	0	0	0	926	0	0	0	0 / 936
7	0	0	1	0	0	0	0	1047	0	1	0.0019 / 2 / 1,049
8	0	1	1	0	1	1	0	0	978	1	0.0019 / 6 / 994
9	0	0	0	3	0	0	4	0	1039	7	0.0007 / 7 / 1,043
Total	994	1187	945	946	947	915	929	1054	979	1038	0.0026 / 26 / 9,957

OUTPUT

OUTPUT - STATUS OF NEURON LAYERS (PREDICTING C785, 10-CLASS CLASSIFICATION, MULTINOMIAL DISTRIBUTION, CROSSENTROPY LOSS, 100,810 WEIGHTS/BIASES, 899.2 KB, 9,240,000 TRAINING SAMPLES, MINIBATCH SIZE 32)

OUTPUT - SCORING HISTORY

OUTPUT - TRAINING_METRICS

OUTPUT - TRAINING_METRICS - TOP-10 HIT RATIOS

OUTPUT - VALIDATION_METRICS

OUTPUT - VALIDATION_METRICS - TOP-10 HIT RATIOS

OUTPUT - VARIABLE IMPORTANCES

PREVIEW POJO

Connections: 0 H2O

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

scaled_importance

TRAINING METRICS - CONFUSION MATRIX VERTICAL: ACTUAL; ACROSS: PREDICTED

	0	1	2	3	4	5	6	7	8	9	Error Rate
0	970	1	0	1	0	1	2	1	0	0	0.0126 / 10 / 989
1	0	1125	4	0	0	1	2	0	3	0	0.0086 / 10 / 1,139
2	0	1	1812	2	1	0	2	0	4	0	0.0159 / 29 / 1,032
3	0	0	2	996	0	4	0	5	2	1	0.0139 / 14 / 1,010
4	0	0	4	1	951	0	4	3	1	0	0.0232 / 23 / 982
5	0	2	0	0	9	1	865	5	1	4	0.0380 / 27 / 892
6	5	2	1	0	1	3	943	0	3	0	0.0157 / 15 / 958
7	1	1	0	2	0	0	0	1012	1	4	0.0150 / 16 / 1,028
8	0	3	0	2	6	5	3	3	946	7	0.0349 / 34 / 974
9	3	4	0	5	10	4	0	0	1	979	0.0229 / 36 / 1,069
Total	988	1134	1034	1017	989	983	943	1037	961	1003	0.0197 / 17 / 10,000

VALIDATION METRICS - CONFUSION MATRIX VERTICAL: ACTUAL; ACROSS: PREDICTED

	0	1	2	3	4	5	6	7	8	9	Error Rate
0	970	1	0	1	0	1	2	1	0	0	0.0126 / 10 / 989
1	0	1125	4	0	0	1	2	0	3	0	0.0086 / 10 / 1,139
2	0	1	1812	2	1	0	2	0	4	0	0.0159 / 29 / 1,032
3	0	0	2	996	0	4	0	5	2	1	0.0139 / 14 / 1,010
4	0	0	4	1	951	0	4	3	1	0	0.0232 / 23 / 982
5	0	2	0	0	9	1	865	5	1	4	0.0380 / 27 / 892
6	5	2	1	0	1	3	943	0	3	0	0.0157 / 15 / 958
7	1	1	0	2	0	0	0	1012	1	4	0.0150 / 16 / 1,028
8	0	3	0	2	6	5	3	3	946	7	0.0349 / 34 / 974
9	3	4	0	5	10	4	0	0	1	979	0.0229 / 36 / 1,069
Total	988	1134	1034	1017	989	983	943	1037	961	1003	0.0197 / 17 / 10,000

OUTPUT

OUTPUT - STATUS OF NEURON LAYERS (PREDICTING C785, 10-CLASS CLASSIFICATION, MULTINOMIAL DISTRIBUTION, CROSSENTROPY LOSS, 100,810 WEIGHTS/BIASES, 899.2 KB, 9,240,000 TRAINING SAMPLES, MINIBATCH SIZE 32)

OUTPUT - SCORING HISTORY

OUTPUT - TRAINING_METRICS

OUTPUT - TRAINING_METRICS - TOP-10 HIT RATIOS

OUTPUT - VALIDATION_METRICS

OUTPUT - VALIDATION_METRICS - TOP-10 HIT RATIOS

OUTPUT - VARIABLE IMPORTANCES

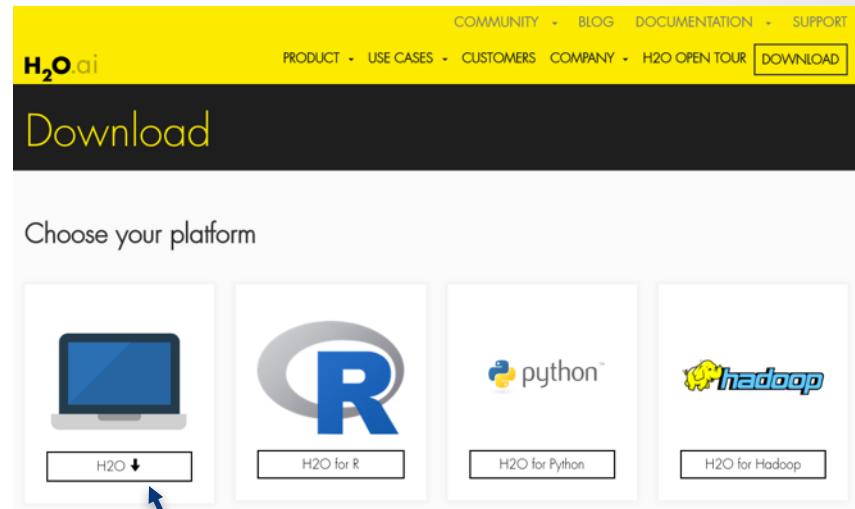
PREVIEW POJO

Connections: 0 H2O

Ready

H2O Flow (Web Interface) Demo

- Download zip from
www.h2o.ai
- Unzip
- Start a local cluster
 - java –jar h2o.jar
- Go to "localhost:54321"



H2O Flow (Web Interface) Demo

- Demo

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

The screenshot shows the main navigation menu of the H2O documentation site. It includes sections for "Getting Started" (with links to H2O, Sparkling Water, and Questions and Answers), "Data Science Algorithms" (Supervised Learning and Unsupervised Learning), and "Languages" (R, Python, Java, and Scala). Each section contains links to various tutorials, booklets, and reference documents.

- Python

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

This screenshot focuses on the "Languages" section of the documentation. It lists four supported languages: R, Python, Java, and Scala. Each language has a corresponding box containing links to quick start videos, package documentation, and other resources.

H2O R Code Example

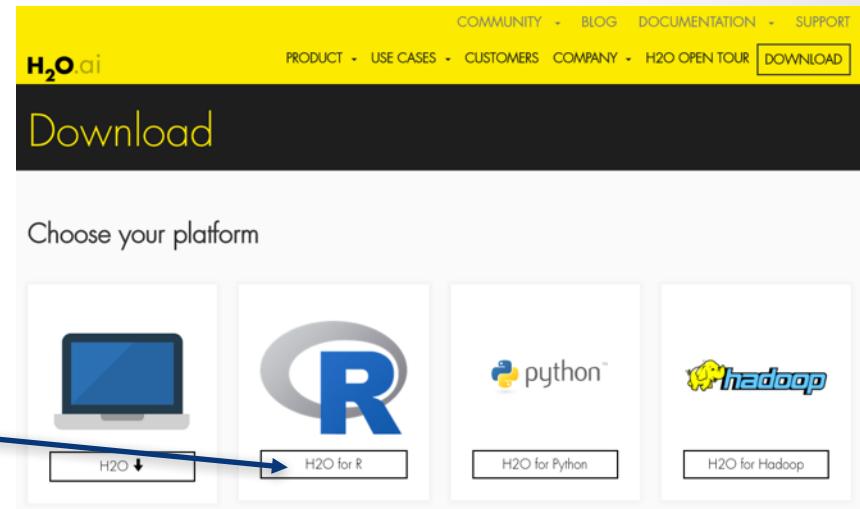
```
42 # Train a GBM with default values  
43 model_gbm ← h2o.gbm(x = features,  
44                      y = target,  
45                      training_frame = h2o_df_boston)  
46  
47 # First look  
48 print(model_gbm)
```

Slide & Code for Workshop:

bit.ly/h2o_budapest_workshop

H2O R Package Demo

- Install from CRAN
 - `install.packages("h2o")`
- Install latest stable
 - See instructions on webpage

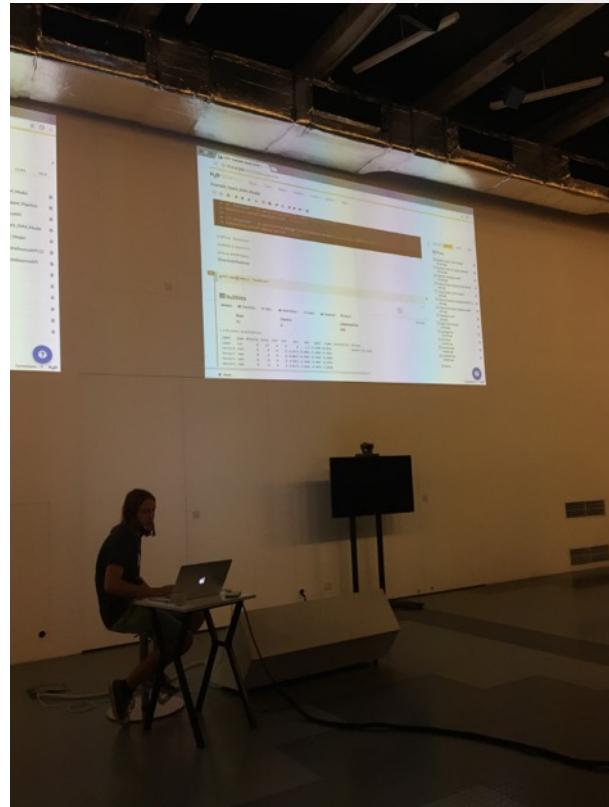


Using H2O Flow & R Together

- Demo

Sparkling Water 2.0

- H₂O + Spark =
Sparkling Water
- See slides from a
recent meetup



Export Plain Old Java Object (POJO)

The screenshot shows the H2O FLOW interface with the "DeepLearning_MNIST" project selected. On the left, there is a "PREVIEW POJO" section containing Java code for a DeepLearningModel. The code includes imports for java.util.Map, hex.gennmodel.GenModel, and hex.gennmodel.annotations.ModelPojo; and defines a ModelPojo named "deeplearning_d5c35043_8929_441a_9a23_dc44b06b519f". The class extends GenModel and implements hex.ModelCategory. It has methods for isSupervised, getClassNames, and fill. The fill method contains a large list of numerical values representing weights or biases for a neural network, indexed from 0 to 281.

```
/*
 Licensed under the Apache License, Version 2.0
 http://www.apache.org/licenses/LICENSE-2.0.html
 AUTOGENERATED BY H2O at 2016-07-13T13:04:11.112+01:00
 3.8.2.9

 Standalone prediction code with sample test data for DeepLearningModel named deeplearning_d5c35043_8929_441a_9a23_dc44b06b519f.

 (Note: Try java arguments -XX:+PrintCompilation to show runtime JIT compiler behavior.)

 */
import java.util.Map;
import hex.gennmodel.GenModel;
import hex.gennmodel.annotations.ModelPojo;

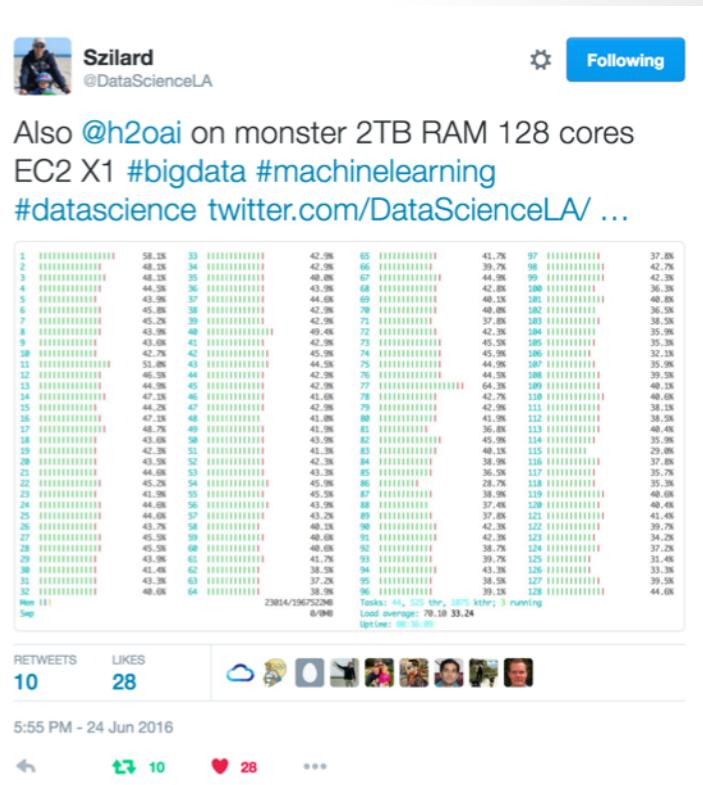
@ModelPojo(name="deeplearning_d5c35043_8929_441a_9a23_dc44b06b519f", algorithm="deeplearning")
public class deeplearning_d5c35043_8929_441a_9a23_dc44b06b519f extends GenModel {
    public hex.ModelCategory getModelCategory() { return hex.ModelCategory.Multinomial; }
    public boolean isSupervised() { return true; }
    public String[] getClassNames() { return null; }
    public int getClassCount() { return 10; }
    // Thread-local storage for input neuron activation values.
    final double[] MUX = new double[717];
    static final RandomNumberGenerator random; implements java.io.Serializable {
        public static final float double1 = new double[717];
        static {
            NORMAL_0.fill(double1);
        }
        static final class NORMAL_0 implements java.io.Serializable {
            static final void fill(double[] sa) {
                sa[0] = 2.303093179181183;
                sa[1] = 0.734017891382369;
                sa[2] = 0.170810382180111;
                sa[3] = 27.2165208759963;
                sa[4] = 15.38931809238464;
                sa[5] = 0.137789038807128;
                sa[6] = 0.100000000000000;
                sa[7] = 0.4770811339672993;
                sa[8] = 0.3370925194312987;
                sa[9] = 0.300032399939346;
                sa[10] = 0.295388411384054;
                sa[11] = 0.295388411384054;
                sa[12] = 0.1795333535660159;
                sa[13] = 0.1795333535660159;
                sa[14] = 0.160000000000000;
                sa[15] = 0.160000000000000;
                sa[16] = 0.1670920303277030;
                sa[17] = 0.184720624771582;
                sa[18] = 0.2392920572379958;
                sa[19] = 0.2392920572379958;
                sa[20] = 0.3116530577680742;
                sa[21] = 0.3116530577680742;
                sa[22] = 0.8253991494921597;
                sa[23] = 1.2396880592380334;
                sa[24] = 1.2396880592380334;
                sa[25] = 1.2396880592380334;
                sa[26] = 7.797572688781051;
                sa[27] = 1.2262361623632154;
            }
        }
    }
}
```

The screenshot shows the "Model" section of the H2O FLOW interface for the "DeepLearning_MNIST" project. It displays the Model ID as "deeplearning-d5c35043-8929-441a-9a23-dc44b06b519f" and the Algorithm as "Deep Learning". Below this, there are several action buttons: Refresh, Predict, Download POJO (which has an arrow pointing to it), Export, Inspect, and Delete.

Why H₂O?

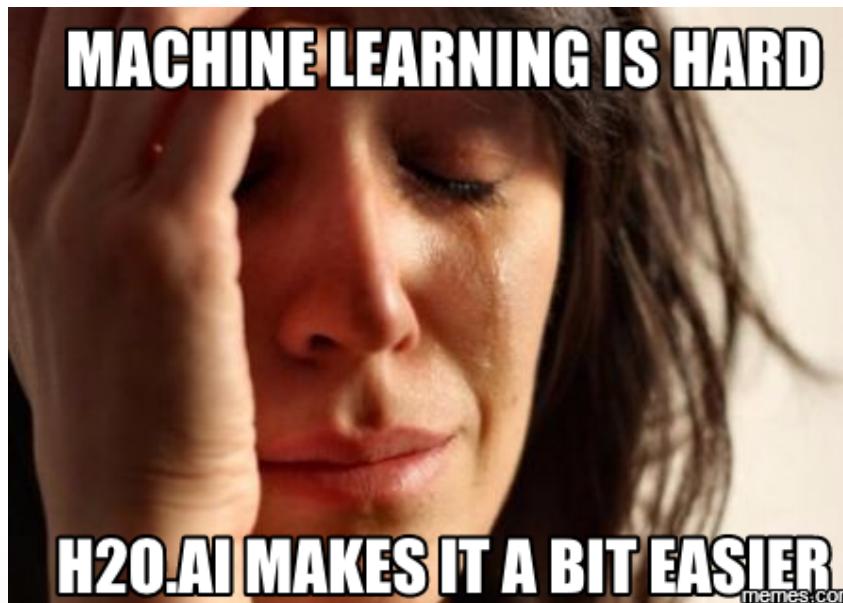
Szilard Pafka – Chief Data Scientist at Epoch

- Budapest DS Meetup
 - Szilard gave a talk
 - Intro to ML with H2O
 - User perspective
 - [Link \(video\)](#)
 - [Link \(Slides\)](#)



Szilard Pafka – Why H2O?

- Szilard's Summary Slide



Telenor – Why H2O?

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

	R	Spark ML 1.3	Radoop	H ₂ 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

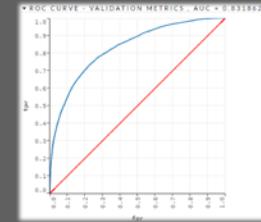
USE CASE: CHURN MODELLING

Modelling on 100 GB of data.

- 40 M rows
- 551 columns

Results:

- GBM
- AUC → 0.83
- LIFT top 1% → 20



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



#EARL2016

Our Customers

The image displays four separate video testimonial snippets arranged in a 2x2 grid. Each snippet features a different customer's face and a 'WATCH VIDEO' button.

- Brendan Herger**
Data Scientist
Capital One
- Pawan Divakarla**
Data and Analytics Business Leader
Progressive Insurance
- Edward Agarwala**
Data Scientist
Progressive Insurance
- Prateem Mandal**
Technical Lead Architect
MarketShare

Each testimonial includes a quote from the customer:

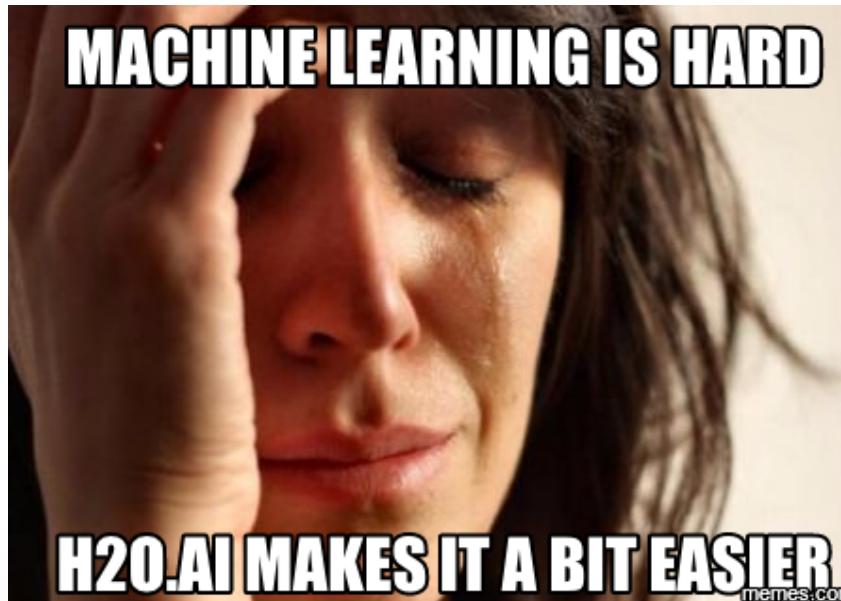
- "We evaluated a large number of hard and soft metrics. H2O just scored really well with all of these areas, particularly relative to a lot of the machine learning frameworks that are available at the moment."
- "H2O is like an enabler in how people are thinking about the data and how they want to use 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."
- "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."
- "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"

Check out the videos - www.h2o.ai

What's Next?

Recap – H2O’s Mission

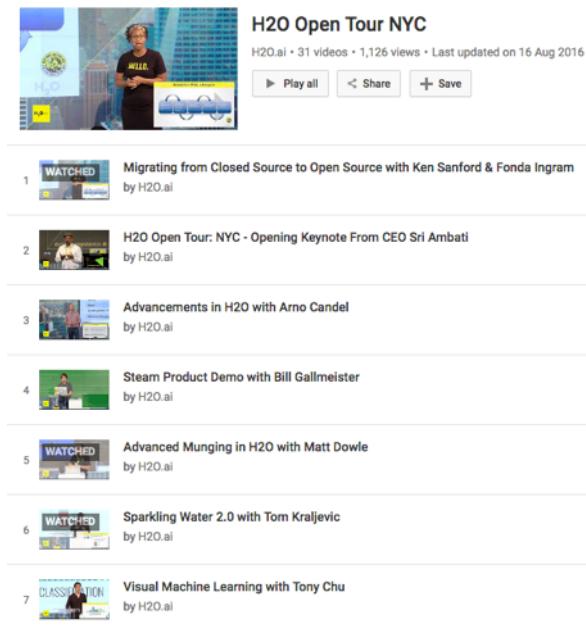
- Szilard’s Summary Slide



H2O is Evolving

- Advanced data munging
- Visual ML
- Deep Water
 - H2O + TensorFlow, mxnet, Theano, Caffe ...
 - GPU
- Steam
 - User-friendly and advanced data science platform

- [YouTube Playlist](#)



The screenshot shows a YouTube playlist page for 'H2O Open Tour NYC'. The header includes the title, the channel 'H2O.ai', the number of videos (31), views (1,126), and the last update date (16 Aug 2016). Below the header are buttons for 'Play all', 'Share', and 'Save'. The playlist contains seven entries, each with a thumbnail, a title, and the 'WATCHED' status indicator.

Rank	Thumbnail	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

Thanks!

- La Fosse Associates
- Chelsea FC
- Our Users:
 - Szilard Pafka (Epoch)
 - Liki Norbert (Telenor)
 - Vincent Warmerdam
(GoDataDriven)
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe
- Resources
 - docs.h2o.ai
- Slides and Code
 - github.com/h2oai/h2o-meetups