

Introduction to Machine Learning with H2O



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

Paris Machine Learning Meetup
Murex
21st September, 2016

About Me: Civil Engineer → Data Scientist

- 2005 - 2015
- Water Engineer
 - Consultant for Utilities
 - EngD Research
 - Machine learning + Water Engineering
 - ***Discovered H2O in 2014!***
- 2015 - Present
- Data Scientist
 - Virgin Media (UK)
 - Domino Data Lab (US)
 - H2O.ai (US)

Why? Long story – see bit.ly/joe_h2o_talk2

Agenda

- **This Talk (40 mins)**
 - About H2O.ai
 - Demos
 - Web & R Interface
 - Why H2O?
 - What's Next?
 - New developments
- **Second Talk (30 mins)**
 - Deep Water
 - Demos
 - H2O + mxnet
 - H2O + TensorFlow
- **Third Talk (45 mins)**
 - H2O + Spark = Sparkling Water
 - Demos
 - H2O + Spark MLlib

Tweets

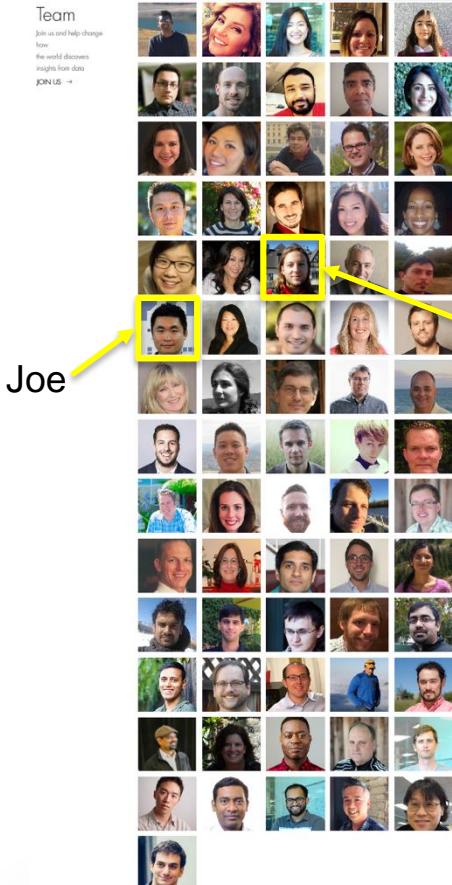
- **Paris Machine Learning Applications Group**
 - @ParisMLGroup
- **H2O.ai**
 - @h2oai
- **Murex**
 - @Work_at_Murex
 - #MurexParis
- **Joe**
 - @matlabulous

About H2O.ai



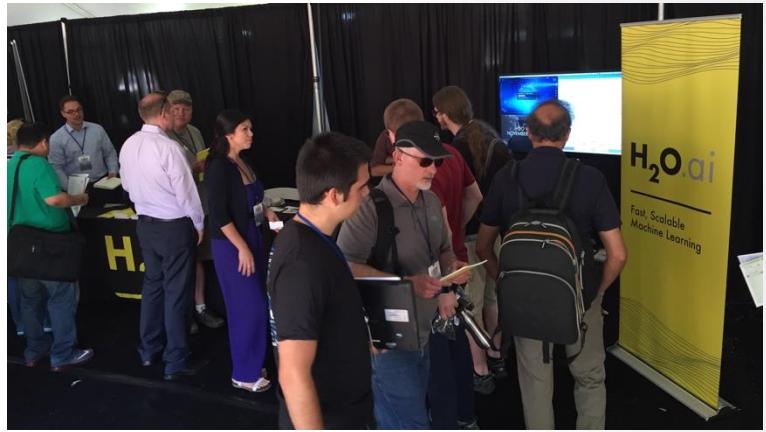
About H2O.ai

- **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



Jakub
(Kuba)

H2O.ai & Stanford University



useR! 2016 Conference at Stanford

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*



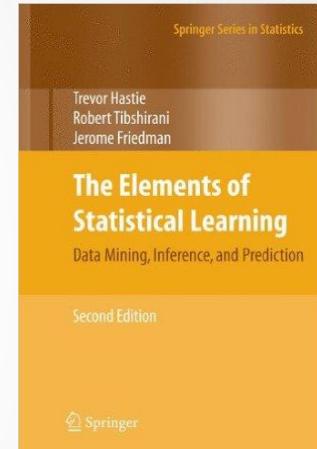
Dr. Robert Tibshirani

- 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*



H2O's Mission

Machine Learning for EVERYONE

“Complexity is your enemy. Any fool can make something complicated. It is hard to make something simple.”



Photo credit: Virgin Media

H2O Platform Overview

- Core algorithms written in high performance Java.
- Fast, distributed and scalable.
- APIs available in R, Python, Scala, REST/JSON and web.
- Works with laptop, cloud, Hadoop and Spark



Current Algorithm Overview

Statistical Analysis

- Linear Models (GLM)
- Naïve Bayes

Ensembles

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

Web & R Demos

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

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

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 LondonR Talk
bit.ly/joe_h2o_talk3

H2O Demos



H2O Demos

- **Demo 1: Web Interface**
 - Public dataset
 - Import data
 - Explore data
 - Build & evaluate models
 - Make predictions
- **Demo 2: R Interface**
 - Same process using R script

Public Dataset – Wine Quality



Photo credit: <http://cs231n.github.io/convolutional-networks/>

Public Dataset – Wine Quality

- **11 Features**
 - Characteristics of wine
 - Acidity, Sugar, pH ... etc.
- **1 Output**
 - Quality (0 – 10)
 - Classification / Regression
- **4898 Records**
 - White wine
- **UCI Machine Learning Repository**
 - <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>
- **CSV**
 - <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv>

H2O Flow (Web Interface) Demo

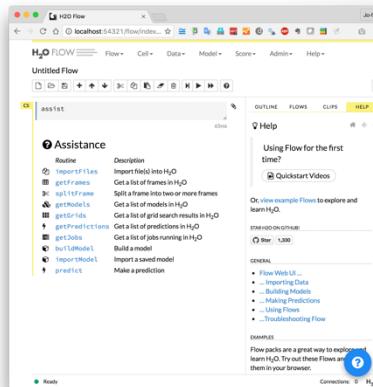
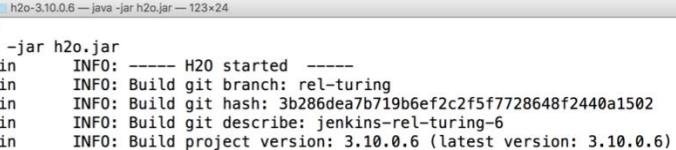
- Download and unzip jar from www.h2o.ai



- In terminal:
 - java -jar h2o.jar

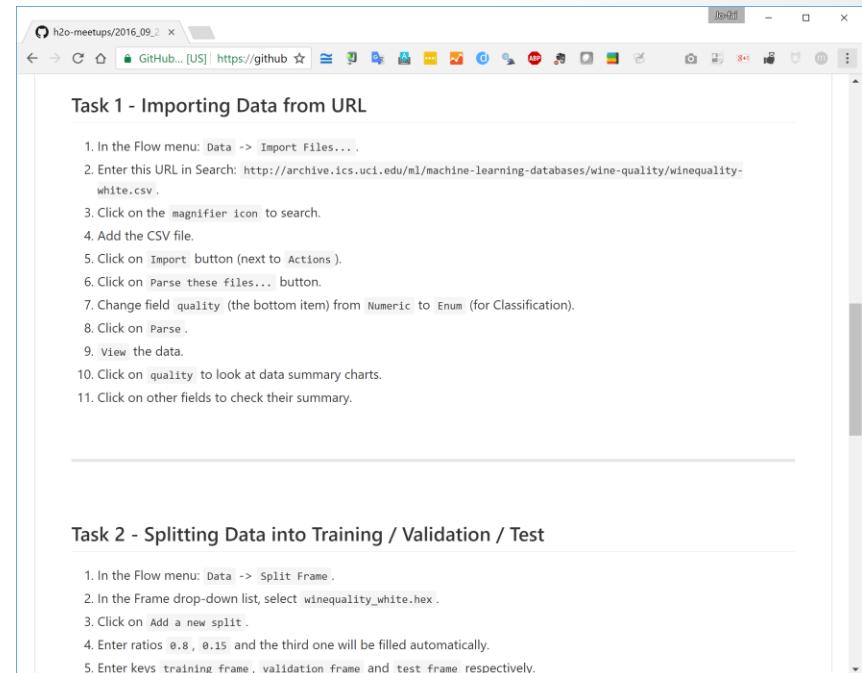
```
Jo-fais-MacBook-Pro-2:~ jofaichow$ cd h2o-3.10.0.6
Jo-fais-MacBook-Pro-2:h2o-3.10.0.6 jofaichow$ java -jar h2o.jar
09-18 13:16:13.620 192.168.0.6:54321 8620 main INFO: ----- H2O started -----
09-18 13:16:13.636 192.168.0.6:54321 8620 main INFO: Build git branch: rel-turing
09-18 13:16:13.636 192.168.0.6:54321 8620 main INFO: Build git hash: 3b286dea7b719b6ef2c2f5f7728648f2440a1502
09-18 13:16:13.636 192.168.0.6:54321 8620 main INFO: Build git describe: jenkins-rel-turing-6
09-18 13:16:13.636 192.168.0.6:54321 8620 main INFO: Build project version: 3.10.0.6 (latest version: 3.10.0.6)
```

- Web browser:
 - localhost:54321



H2O Web Interface – Live Demo

- For online audience
 - Go to this GitHub repo
bit.ly/h2o_paris_1
 - Go to sub-folder
[demo_01_flow](#)
 - Following the procedures in
[README.md](#)



The screenshot shows a web browser window titled "h2o-meetups/2016.09.2". The address bar shows "GitHub... [US] https://github". The main content area displays two tasks:

Task 1 - Importing Data from URL

1. In the Flow menu: Data -> Import Files....
2. Enter this URL in Search: <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv>.
3. Click on the magnifier icon to search.
4. Add the CSV file.
5. Click on Import button (next to Actions).
6. Click on Parse these files... button.
7. Change field quality (the bottom item) from Numeric to Enum (for Classification).
8. Click on Parse.
9. View the data.
10. Click on quality to look at data summary charts.
11. Click on other fields to check their summary.

Task 2 - Splitting Data into Training / Validation / Test

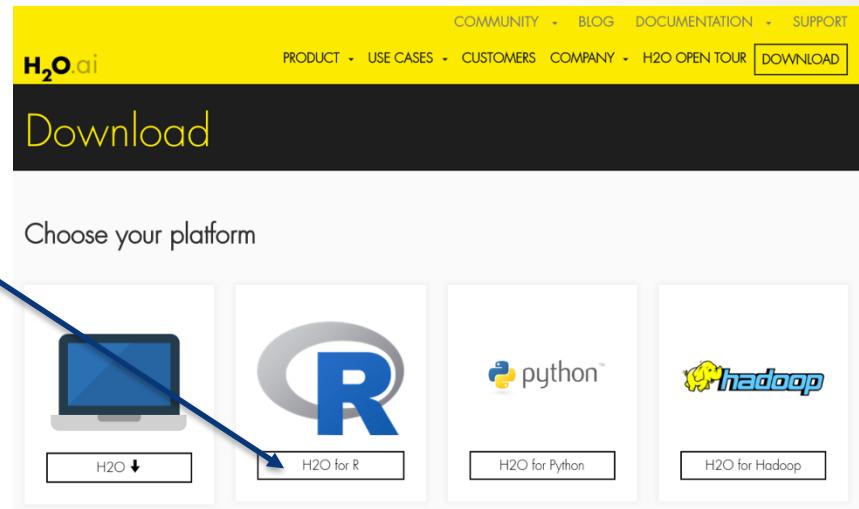
1. In the Flow menu: Data -> Split Frame.
2. In the Frame drop-down list, select winequality_white.hex.
3. Click on Add a new split.
4. Enter ratios 0.8, 0.15 and the third one will be filled automatically.
5. Enter keys training_frame, validation_frame and test_frame respectively.

H2O Flow Live Demo



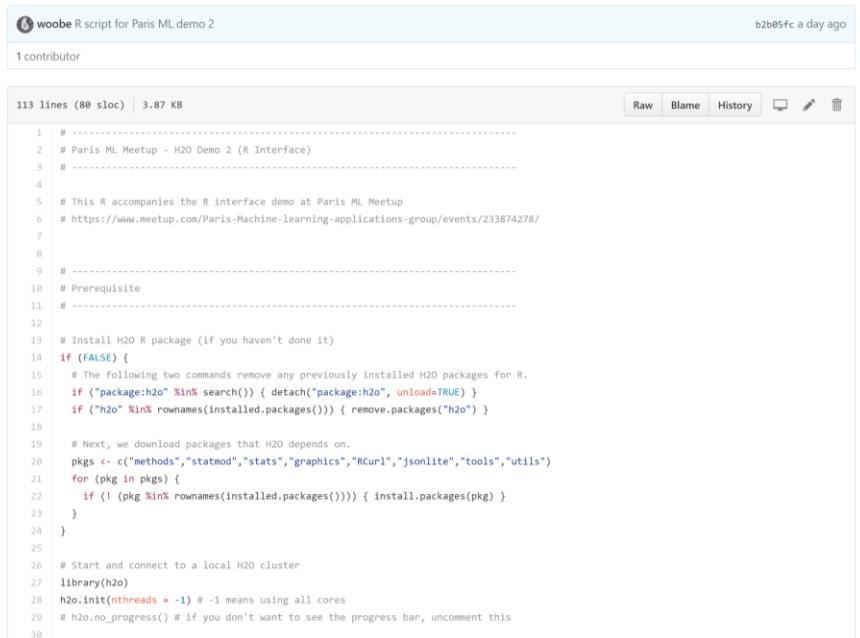
H2O R Package Demo

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



H2O R Interface – Live Demo

- For online audience
 - Go to this GitHub repo
bit.ly/h2o_paris_1
 - Go to sub-folder
[demo_02_r](#)
 - Download and run R scripts.



The screenshot shows a GitHub repository page for a file named 'woobe.R'. The repository has 1 contributor and was last updated a day ago. The file contains 113 lines of R code. The code is used to install the H2O package, remove any previously installed H2O packages, download dependencies, and start a local H2O cluster. It includes comments explaining the steps and provides links to the Paris ML Meetup and the event details.

```
1 # -----
2 # Paris ML Meetup - H2O Demo 2 (R Interface)
3 # -----
4 #
5 # This R accompanies the R interface demo at Paris ML Meetup
6 # https://www.meetup.com/Paris-Machine-learning-applications-group/events/233874278/
7 #
8 #
9 # -----
10 # Prerequisite
11 # -----
12 #
13 # Install H2O R package (if you haven't done it)
14 if (FALSE) {
15   # The following two commands remove any previously installed H2O packages for R.
16   if ("package:h2o" %in% search()) { detach("package:h2o", unload=TRUE) }
17   if ("h2o" %in% rownames(installed.packages())) { remove.packages("h2o") }
18 #
19   # Next, we download packages that H2O depends on.
20   pkgs <- c("methods","statmod","stats","graphics","RCurl","jsonlite","tools","utils")
21   for (pkg in pkgs) {
22     if (! (pkg %in% rownames(installed.packages()))) { install.packages(pkg) }
23   }
24 }
25 #
26 # Start and connect to a local H2O cluster
27 library(h2o)
28 h2o.init(nthreads = -1) # -1 means using all cores
29 # h2o.no_progress() # if you don't want to see the progress bar, uncomment this
```

H2O + R Live Demo



Other H2O Interfaces

- R

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

- Python

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

- Resources - docs.h2o.ai

The screenshot shows the main documentation page for H2O and Sparkling Water. It features three main sections: 'Getting Started' (with links for H2O and Sparkling Water), 'Data Science Algorithms' (with sub-sections for Supervised Learning and Unsupervised Learning), and 'Languages' (with sub-sections for R, Python, Java, and Scala). Each section contains links to Tutorials, Booklets, and References.

Getting Started	
H2O	Tutorial Booklet Reference
Sparkling Water	Tutorial Booklet Reference
Questions and Answers	FAQs, H2O Community Forum, Issue Tracking (JIRA), GitHub, Back Overflow, Cross Validated

Data Science Algorithms	
Supervised Learning	Tutorial Booklet Reference
Unsupervised Learning	Tutorial Booklet Reference

Languages	
R	Tutorial Booklet Reference
Python	Tutorial Booklet Reference
Java	Tutorial Booklet Reference
Scala	Tutorial Booklet Reference

More Advanced Topics

- **Advanced Features:**
 - Hyperparameters Tuning
 - Model Stacking
 - Saving/Loading Models
 - Export Plain Old Java Object (POJO)
- **Resources:**
 - bit.ly/h2o_budapest_1
 - bit.ly/joe_h2o_talk3
 - docs.h2o.ai

Why H2O?



Quote from our CEO

“Software is eating the world and artificial intelligence (AI) is eating software, as traditional rules-based models no longer cut it in today’s rapidly changing world.”

— Sri Ambarti, CEO and Co-founder, H2O.ai

H2O Overview

Computer Science (CS)

Artificial Intelligence (A.I.)

Machine Learning (ML)

Deep Learning (DL)

hot hot hot hot hot



H2O's Mission

Machine Learning for EVERYONE

“Complexity is your enemy. Any fool can make something complicated. It is hard to make something simple.”



Photo credit: Virgin Media

H2O Community

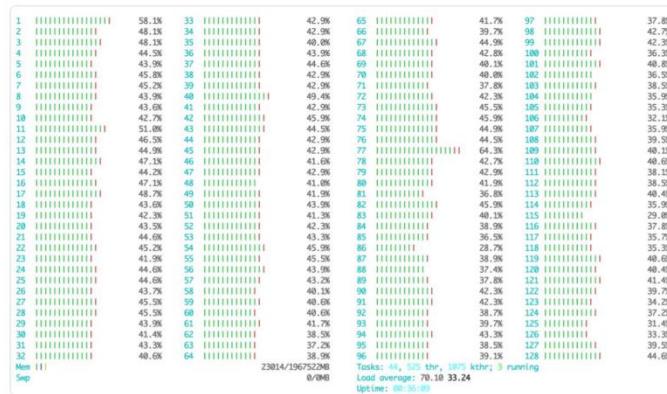


Szilard Pafka – Chief Data Scientist at Epoch

- Sziland's talks / blog posts about H2O:
 - ML Benchmark
 - Intro to ML with H2O
 - H2O Scoring
 - Tweets

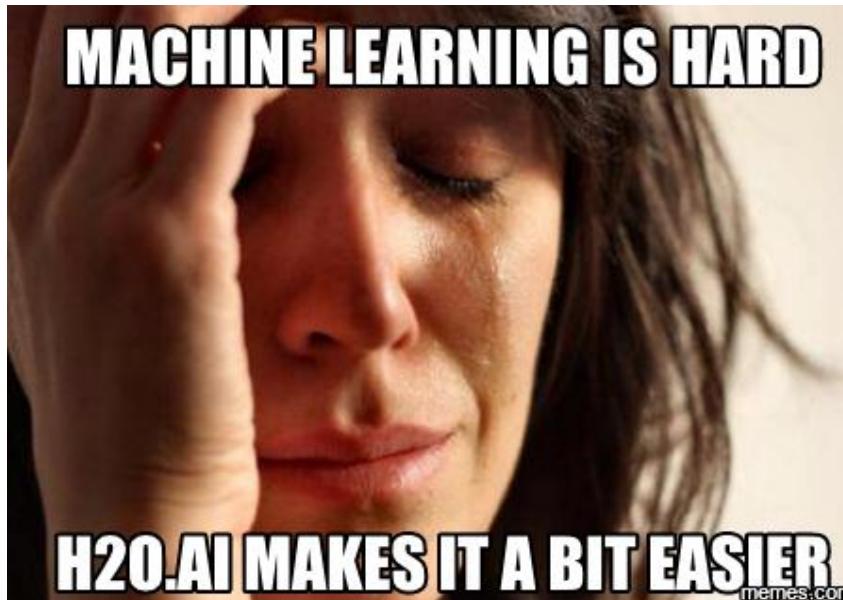


Also @h2oai on monster 2TB RAM 128 cores
EC2 X1 #bigdata #machinelearning
#datascience twitter.com/DataScienceLA/ ...



Szilard Pafka – Why H2O?

- Szilard's Summary Slide



Telenor (Telecom in Hungary) – 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	:(:)	:)	:)



H2O for Kaggle

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

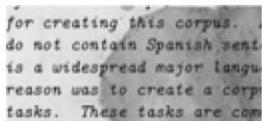
Triskelion | 01.02.2015

[READ MORE](#)

Kaggle challenge 2nd place winner Colin Priest

[READ MORE](#)

“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.”



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.”



H2O Customers

The image displays four separate video player interfaces arranged in a 2x2 grid. Each interface includes a customer's name, title, and company, along with a quote 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 video player has a dark background with a blurred image of the customer. A grey text box contains their testimonial, and a yellow 'WATCH VIDEO' button is at the bottom right.

Check out the videos - www.h2o.ai

H2O Community Support

Google forum – h2osteam

The screenshot shows the Google forum interface for the group 'h2osteam'. The left sidebar includes links for 'Groups', 'My groups', 'Home', 'Starred', 'Favourites' (with a note to click the star icon), 'Recently viewed' (including 'H2O Open Sour...', 'sparkr-d...'), 'Recent searches' (including 'spark streaming (i...'), and 'Recently posted to H2O Open Sour...'). The main content area displays a post from 'h2o Open Source Scalable Machine Learning - h2ostream' with 30 topics and 99+ unread messages. Below it is a preview of the community.h2o.ai website.

community.h2o.ai

The screenshot shows the community.h2o.ai website. The top navigation bar includes links for 'H2O Dev', 'JIRA', 'Summary', 'Learn', 'H2O', and 'Sparkling Water'. The sidebar on the right lists categories like 'Algorithms', 'Announcements', 'Artificial Intelligence', etc. The main content area shows a list of posts under 'All Posts':

- Avkash answered **When is Steam going to be released?** 3 days ago in Steam
- windows commented **H2O Python Modules** 3 days ago in H2O
- windows commented **H2O Installation** 3 days ago in H2O
- Avkash Chauhan edited **PySparkling launch problem with Python 2.6 or older** 3 days ago in Python
- Avkash answered **Predicted Values** 3 days ago in H2O
- erin edited **Combining holdout predictions, while keep_cross_validation_predictions parameter is active in Python**

A yellow banner at the bottom right announces the 'Sparkling Water Release 0.8/30'.

H2O Community in Paris

Your local H2O expert is here – Jiqiong (Ji) Qiu

The screenshot shows a GitHub repository page for 'H2o Tensorflow Spark'. The page includes sections for 'H2o by tensorflow' and 'H2o by docker'. It provides instructions for building a Dockerfile and running a Docker container. A command-line example for building a Docker image is shown:

```
cd $HOME/your_docker
cd $HOME/your_docker
touch Dockerfile
open -e Dockerfile
docker build -t name-of-container .
```

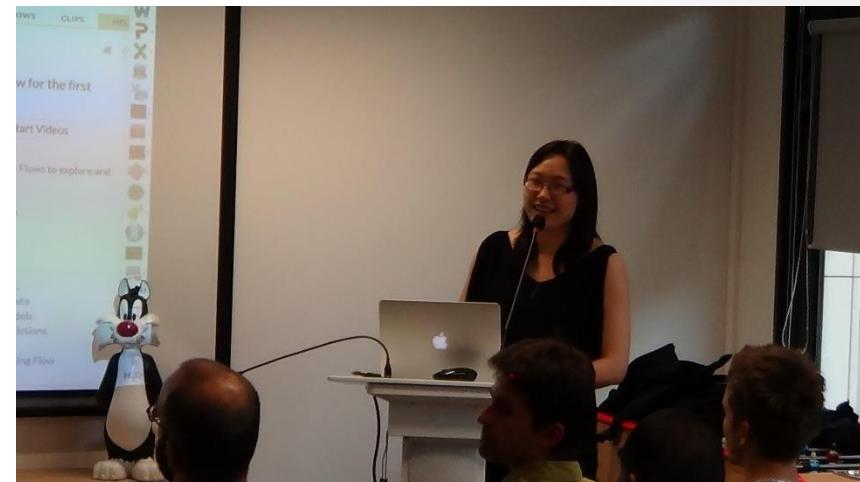
Instructions for running the Docker container are provided, along with a note about sharing files between the host and container via a notebook folder.

```
docker run -it --name name-of-image -p 8080:8080 -p 8888:8888 -p 54321:54321 -p 54322:54322 -p 6086:6086 -v /flodetc
```

You can then add notebook file in a folder called notebook, which allows the share file between your PC and your container. We note that even you delete the notebook all the data and files in the notebook folder are always available.

Now you sparkling with tensorflow can be launch by the following command:

```
ipython notebook --SPARKING_WAIT=1000 --no-browser
```



Docker for H2O + TensorFlow demo

H2O in Europe UK & Europe

London Kaggle Meetup



Strata Hadoop



Chelsea Football Club



PyData
Amsterdam



useR! 2016
Stanford



satRdays
Budapest

What's Next?



Recap: H2O's Mission

Machine Learning for EVERYONE

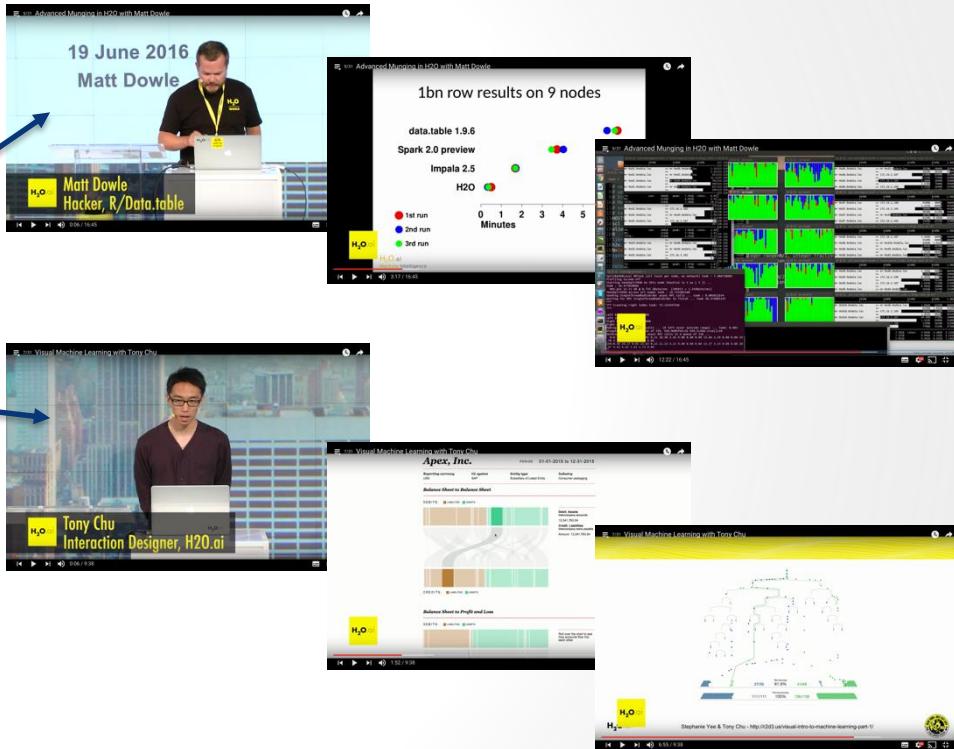
“Complexity is your enemy. Any fool can make something complicated. It is hard to make something simple.”



Photo credit: Virgin Media

H2O is Evolving

- H2O Open Tour NYC YouTube Playlist
 - Advanced data munging
 - Visual ML
 - Deep Water (2nd talk)
 - Sparkling Water (3rd talk)
 - Steam
 - New data science platform



Merci Beaucoup!

- Jiqiong (Ji) Qiu
- Franck Bardol
- Igor Carron
- Slides & Code
 - bit.ly/h2o_paris_1
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe

