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
- **Murex**
  - @Work_at_Murex
  - #MurexParis
- **H2O.ai**
  - @h2oai
- **Joe**
  - @matlabulous

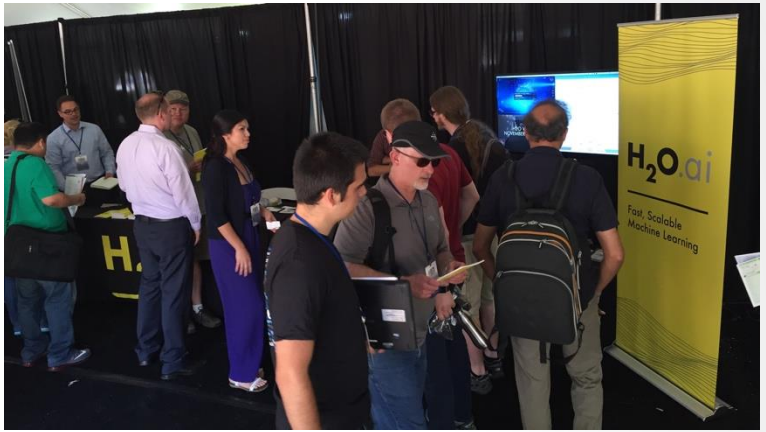
About H2O.ai

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

H2O.ai & Stanford University



Scientific Advisory Council

Dr. Trevor Hastie

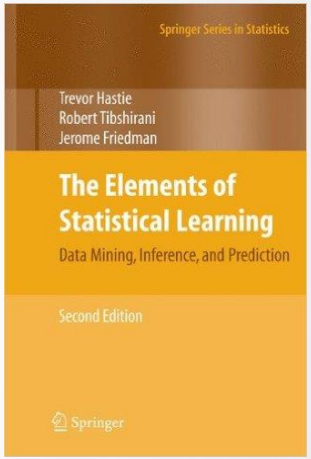
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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*



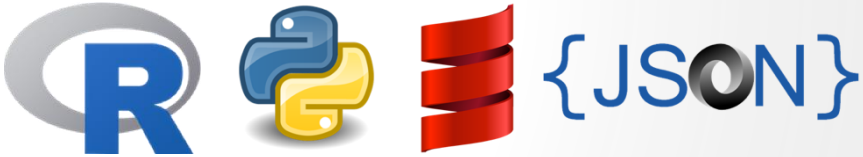
H2O's Mission

Machine Learning for EVERYONE

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

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

Web & R Demos

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

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/

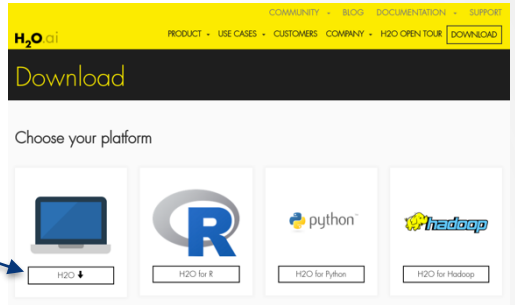
Public Dataset – Wine Quality

- **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

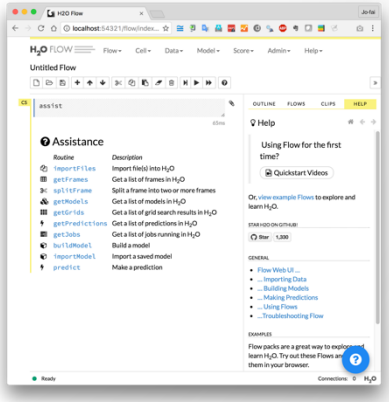
- Download and unzip **jar** from www.h2o.ai



- In terminal:
  - **java -jar h2o.jar**

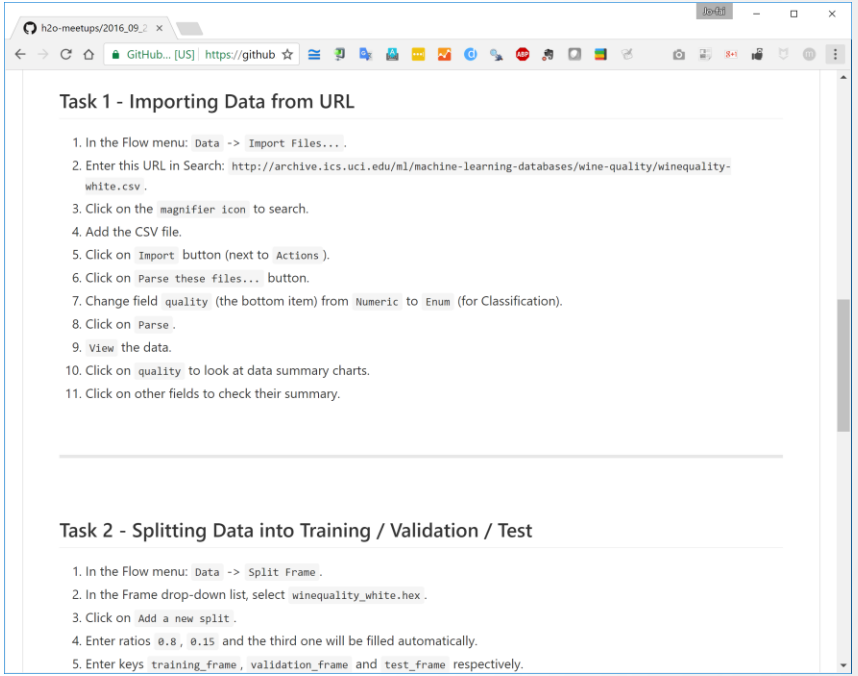
```
h2o-3.10.0.6 — java -jar h2o.jar — 123×24
Jo-fais-MacBook-Pro-2:~ jofaichow$ cd h2o-3.10.0.6
Jo-fais-MacBook-Pro-2:h2o-3.10.0.6 jofaichow$ java -jar h2o.jar
09-18 13:16:13.620 192.168.0.6:54321     8620   main      INFO: ----- H2O started  -----
09-18 13:16:13.636 192.168.0.6:54321     8620   main      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**



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.

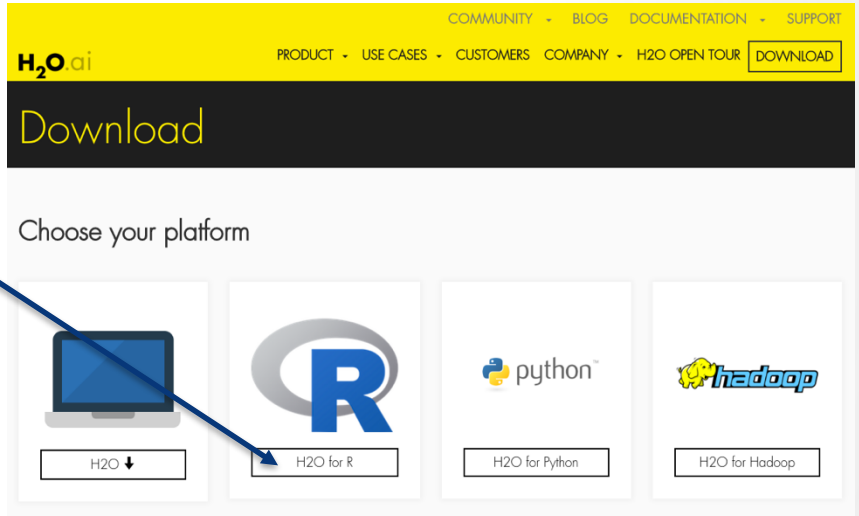
Task 2 - Splitting Data into Training / Validation / Test

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

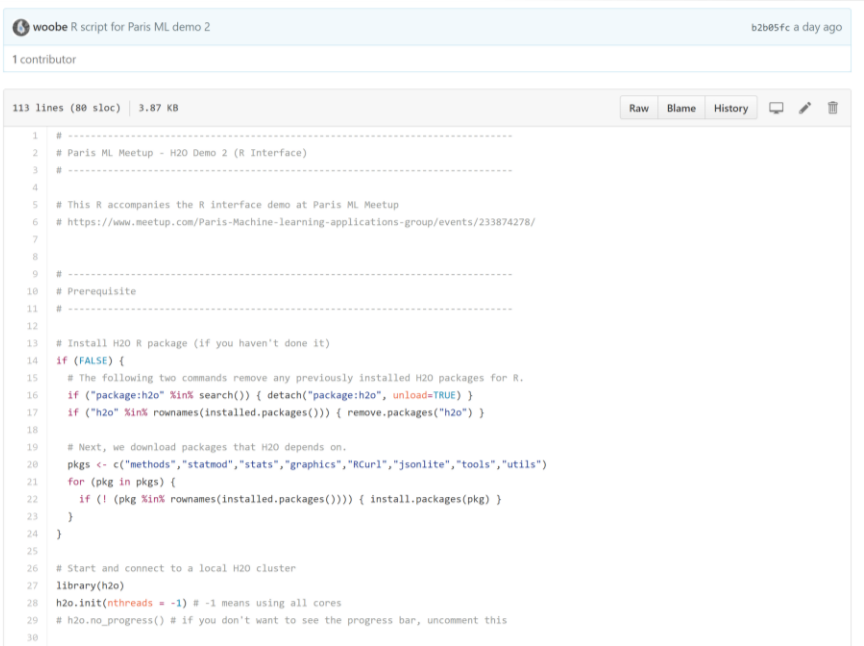
H2O R Package Demo

- **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**.



H2O + R Live Demo

Other H2O Interfaces

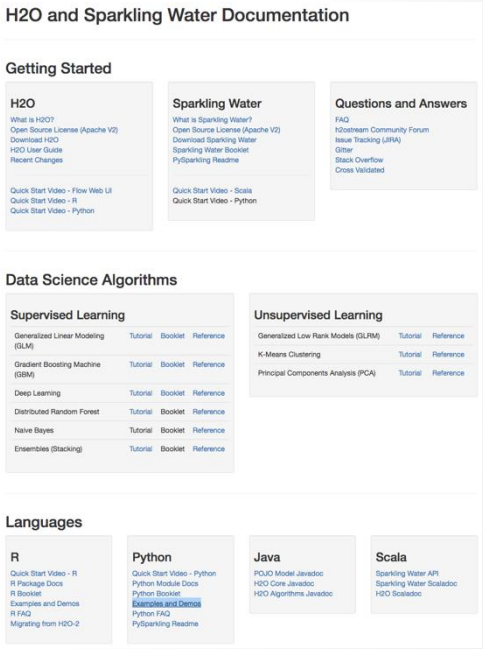
- R

```r
# Load H2O R package
library(h2o)

# Initialize and Connect to H2O
h2o.init()
```

- Python

```python
# Import H2O Python module
import h2o

# Initialize and Connect to H2O
h2o.init()
```

- Resources - docs.h2o.ai



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

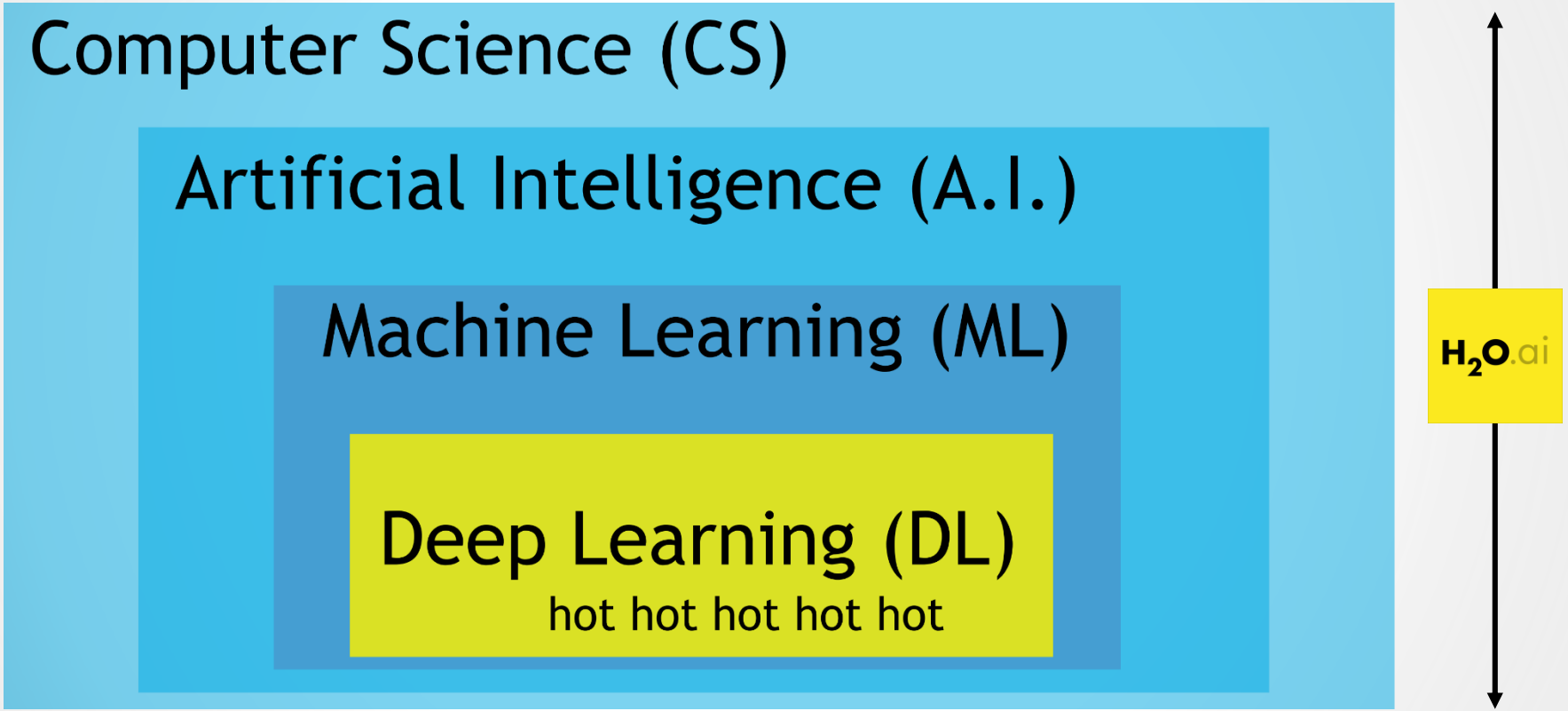
Why H2O?

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



H2O's Mission

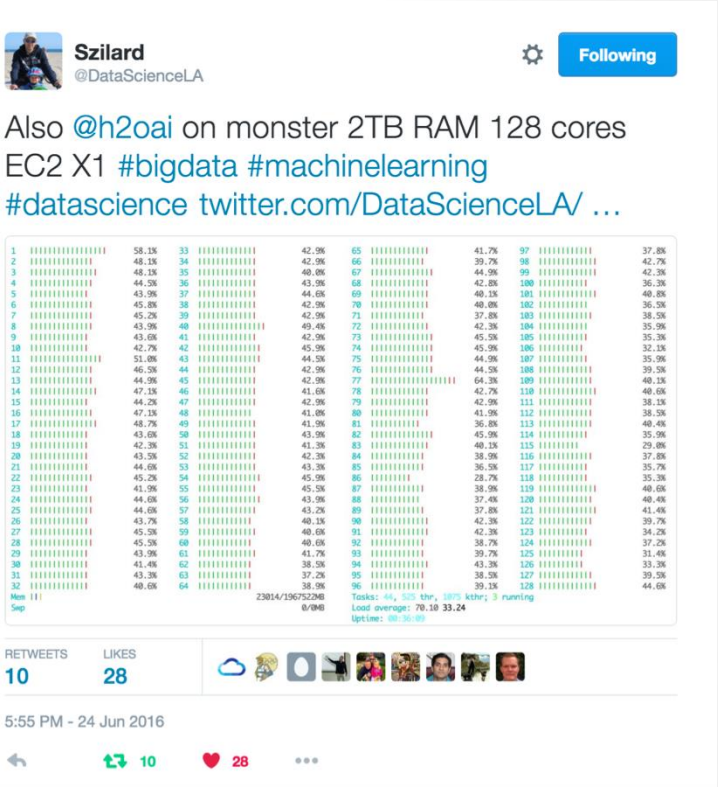
Machine Learning for EVERYONE

Photo credit: Virgin Media

H2O Community

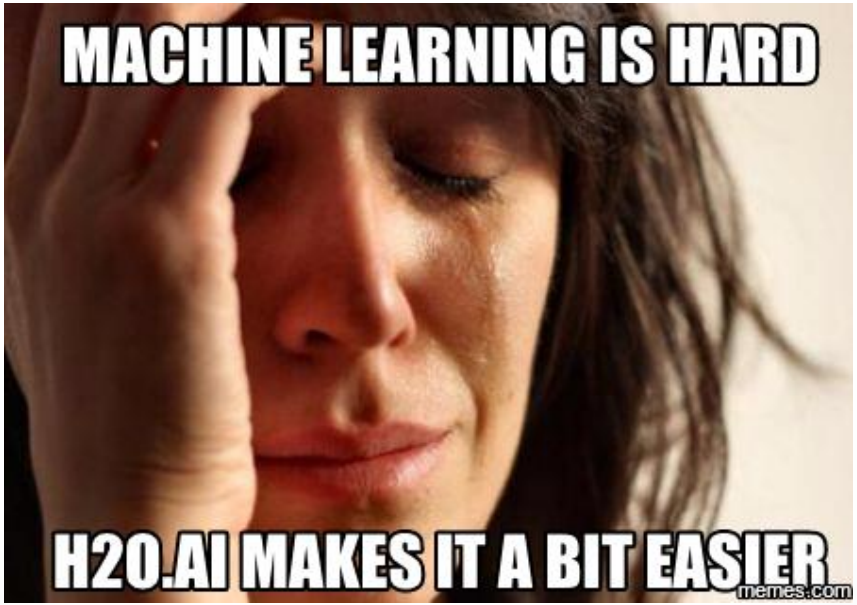
Szilard Pafka – Chief Data Scientist at Epoch

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



Szilard Pafka – Why H2O?

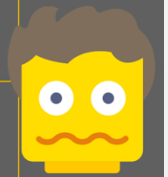
- Szilard's Summary Slide



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_2O
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

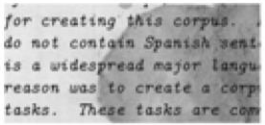
CIFAR-10 Competition Winners: Interviews with Dr. Ben Graham, Phil Culliton, & Zygmunt Zając

Triskelion | 01.02.2015

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."

Kaggle challenge 2nd place winner Colin Priest



Completed • Knowledge • 161 teams

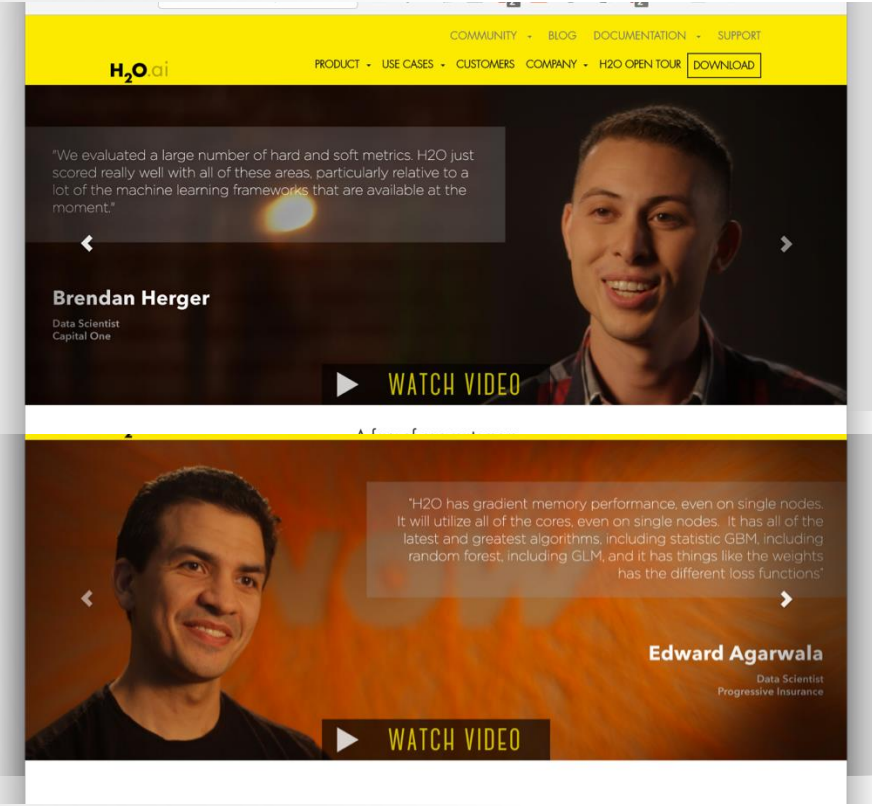
Denoising Dirty Documents

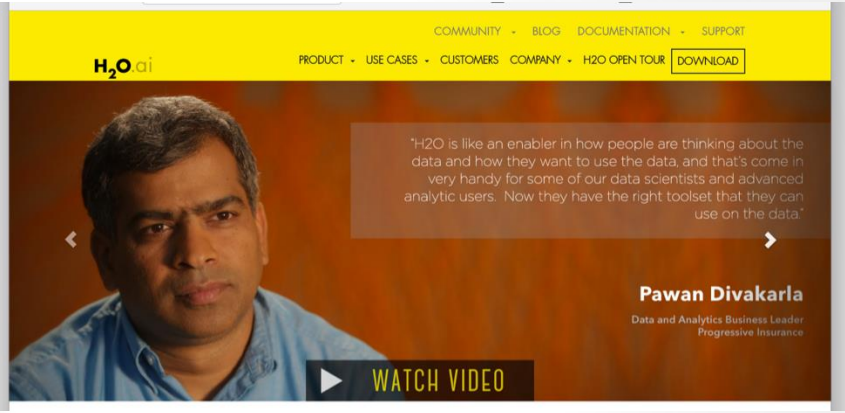
Mon 1 Jun 2015 – Mon 5 Oct 2015 (3 months ago)

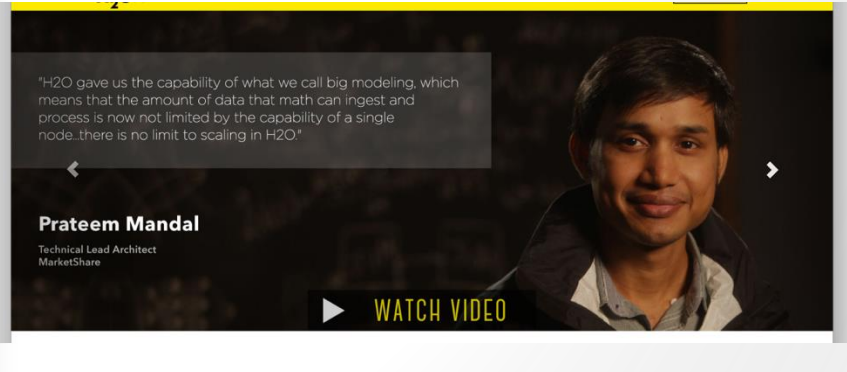
READ MORE

"For my final competition submission I used an ensemble of models, including 3 deep learning models built with R and h2o."

H2O Customers



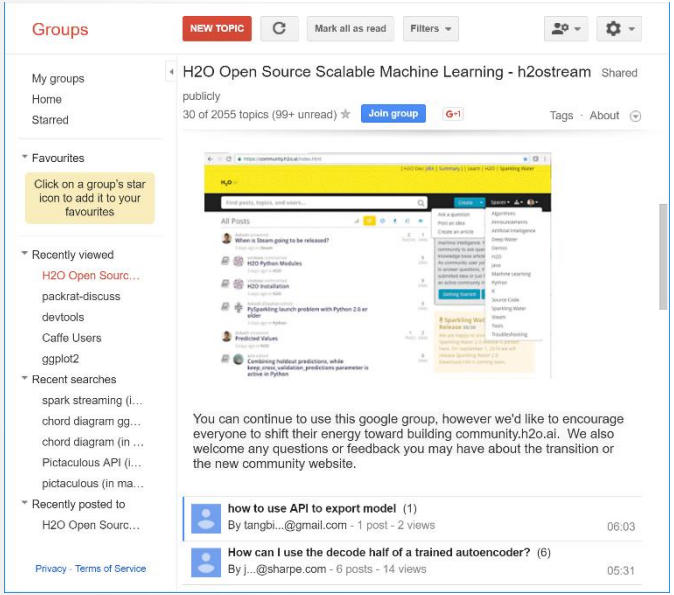




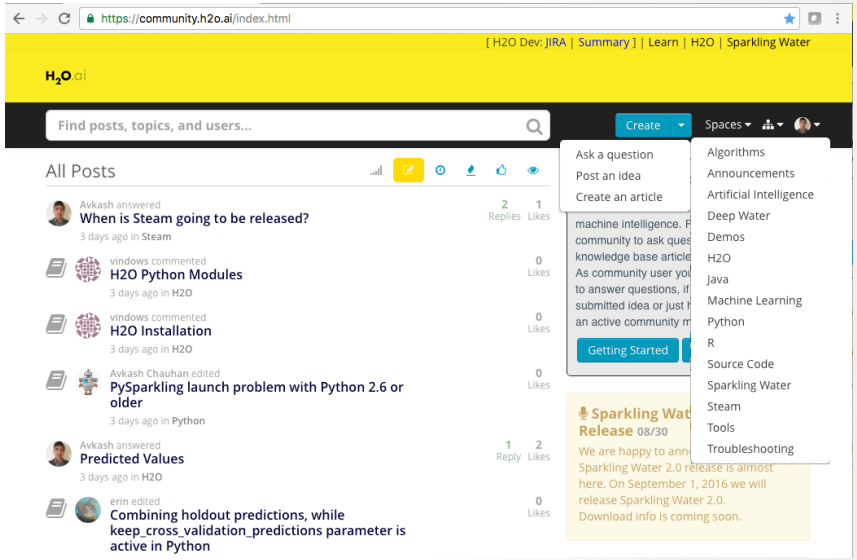
Check out the videos - **www.h2o.ai**

H2O Community Support

Google forum – **h2osteam**

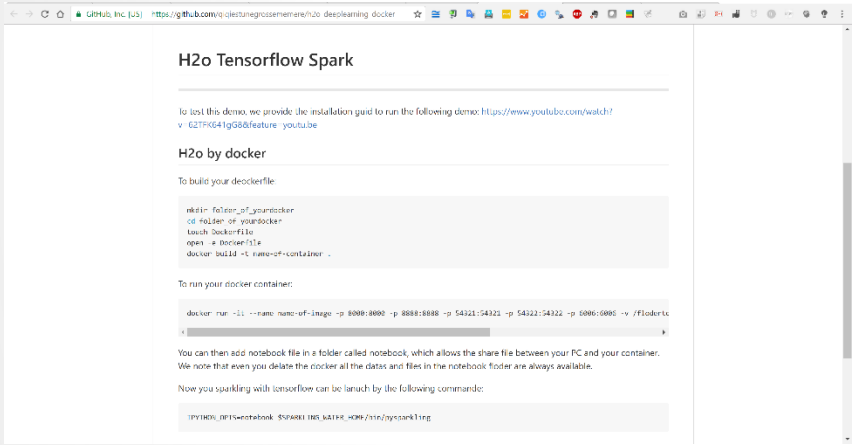


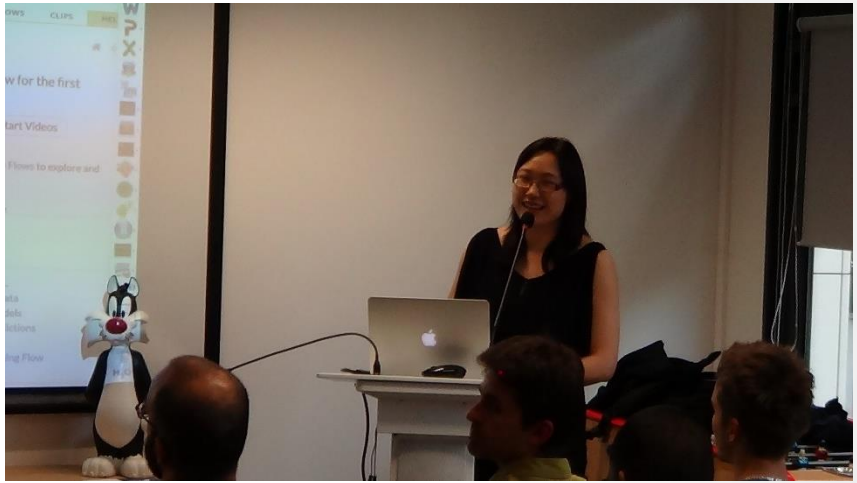
community.h2o.ai



H2O Community in Paris

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





Docker for H2O + TensorFlow demo

H2O Around the World

London Kaggle Meetup

Strata Hadoop

Chelsea Football Club

PyData
Amsterdam

useR! 2016
Stanford

satRdays
Budapest

What’s Next?

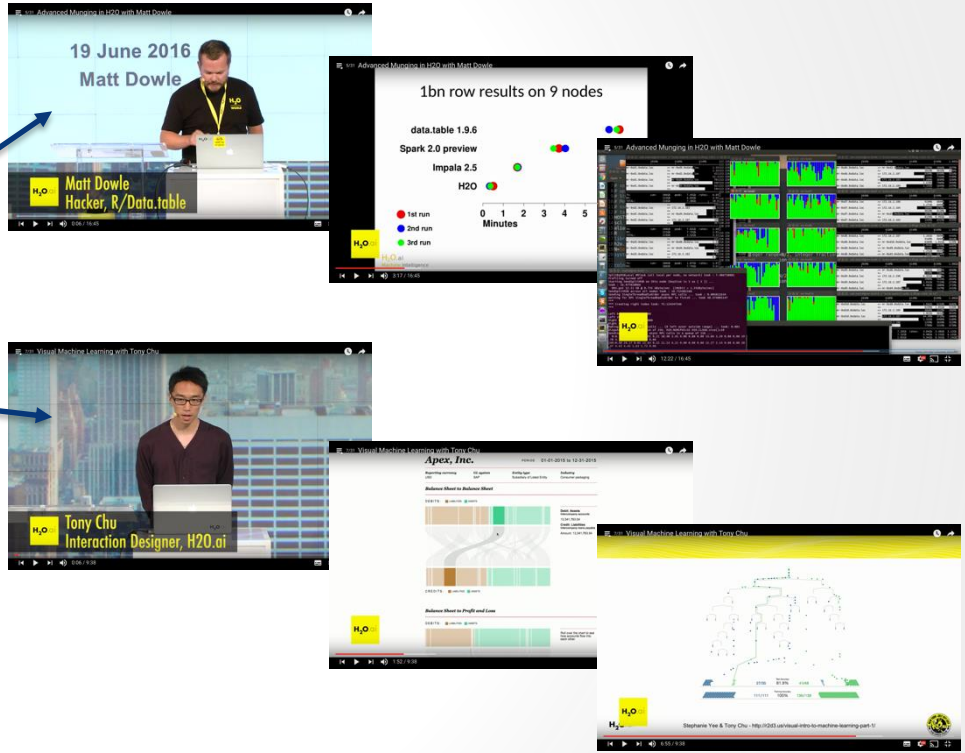
Recap: H2O's Mission

Machine Learning for EVERYONE

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
  - github.com/woobe