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)

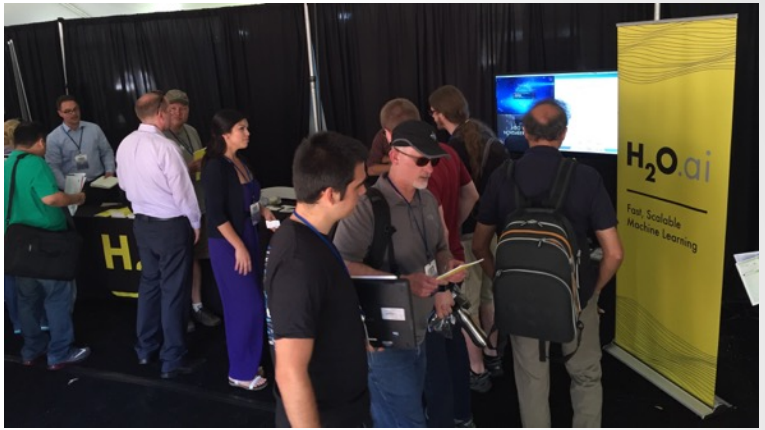
Why? Long story – see bit.ly/joe_h2o_talk2

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

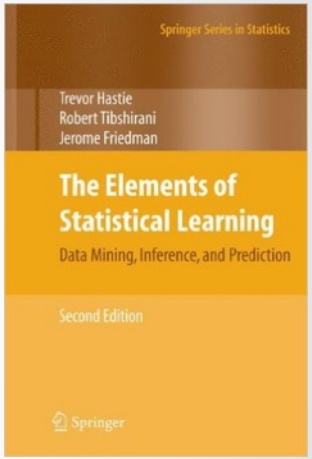
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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

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

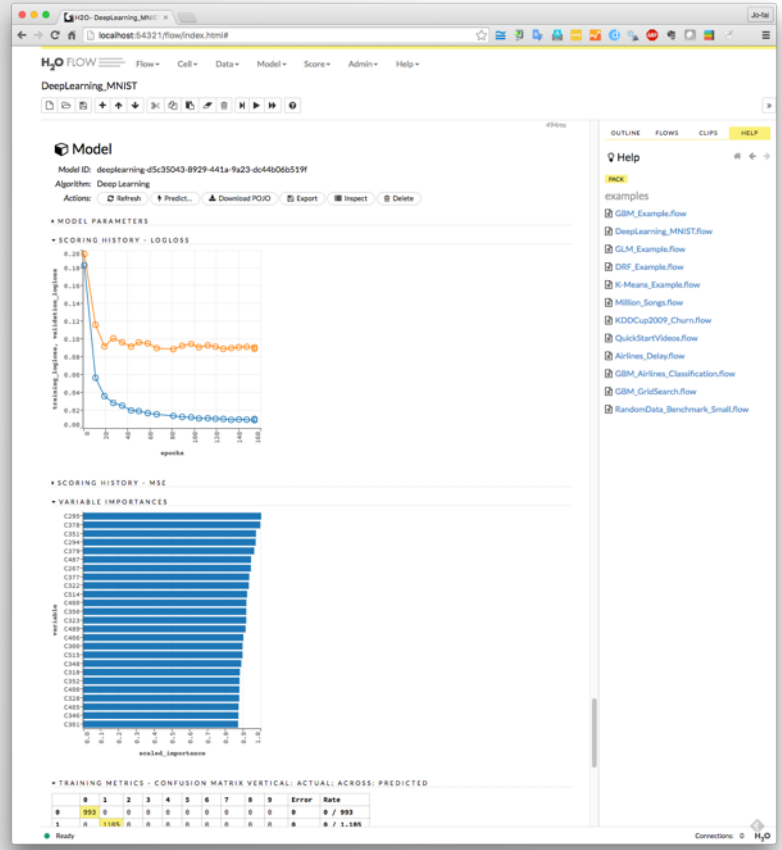
Solvers & Optimization

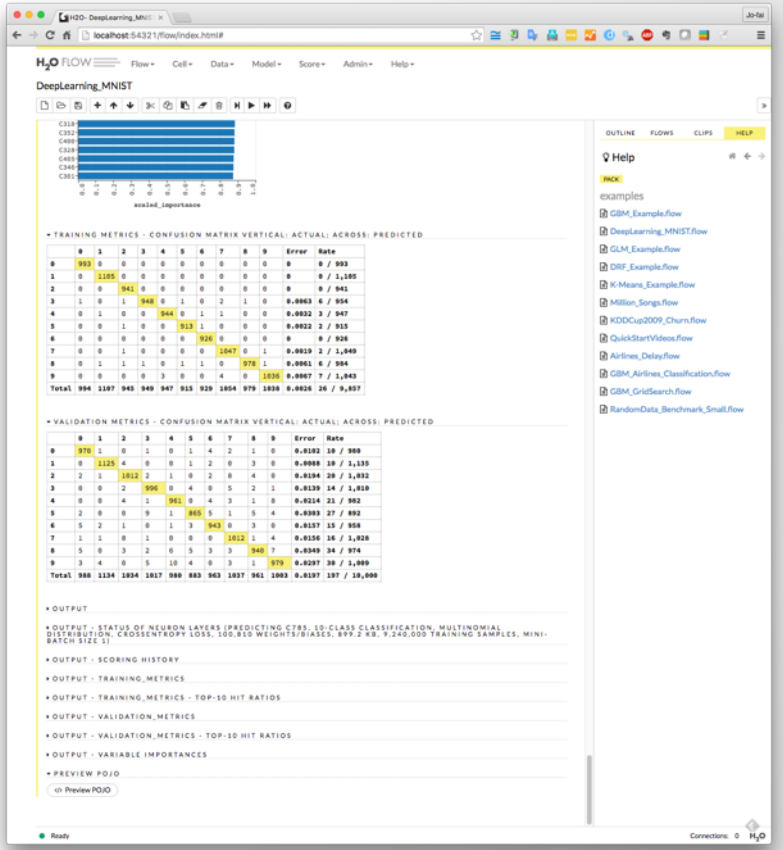
- Generalized ADMM Solver
- L-BFGS (Quasi Newton Method)
- Ordinary Least-Square Solver
- Stochastic Gradient Descent

Data Munging

- Scalable Data Frames
- Sort, Slice, Log Transform

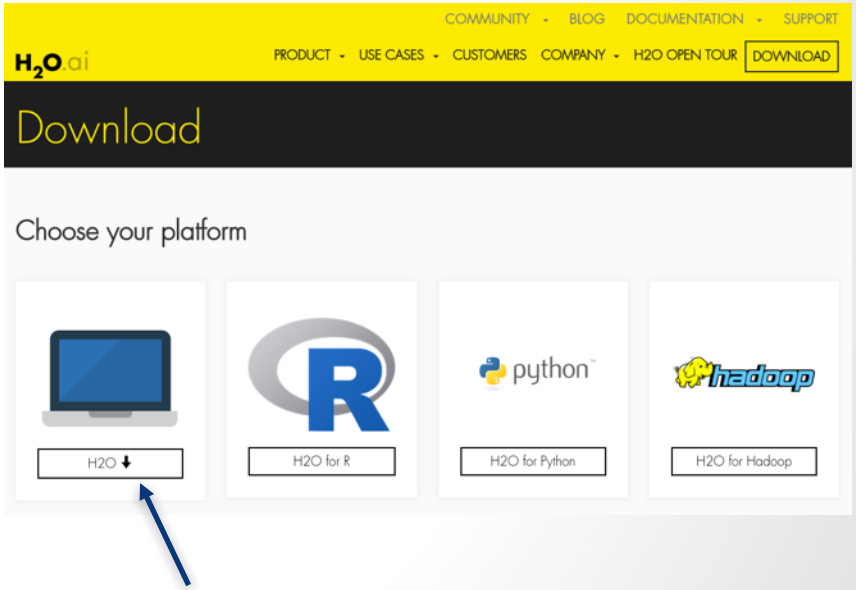
H2O Interfaces – Web (H2O Flow)





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

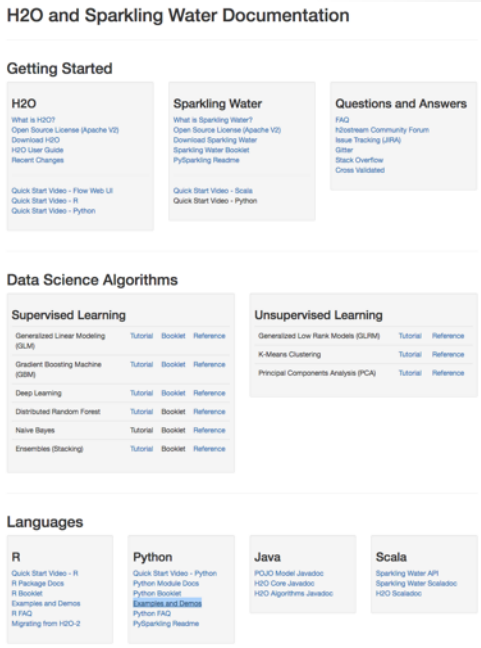
- R

```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



H2O R Code Example

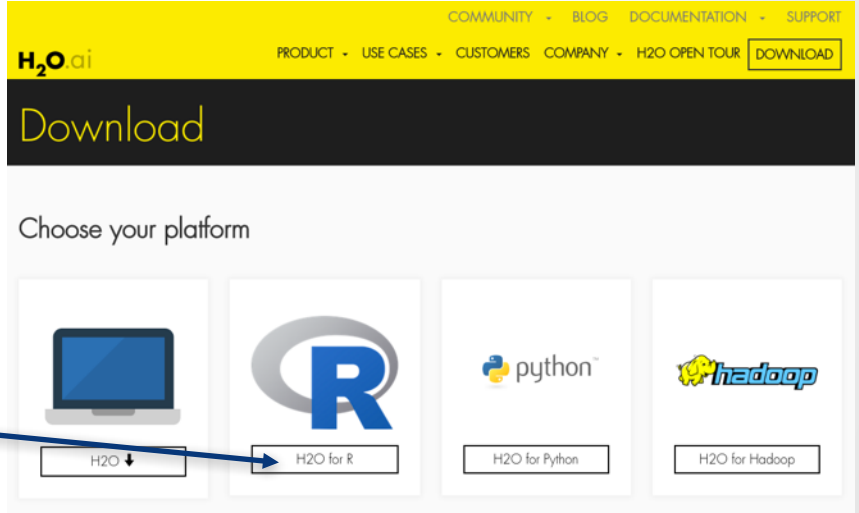
```
42  # Train a GBM with default values
43  model_gbm ← h2o.gbm(x = features,
44                      y = target,
45                      training_frame = h2o_df_boston)
46
47  # First look
48  print(model_gbm)
```

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

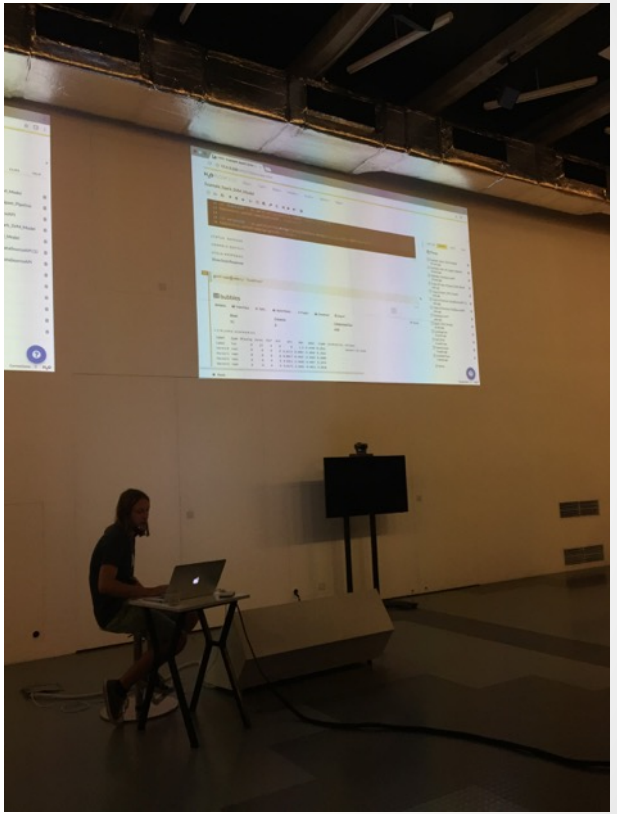


Using H2O Flow & R Together

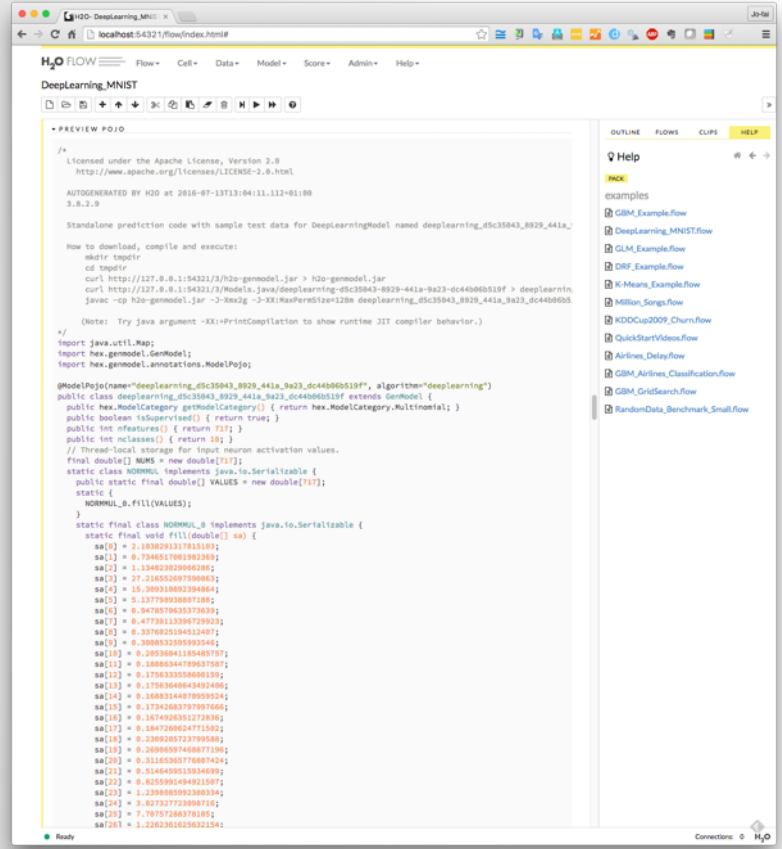
- Demo

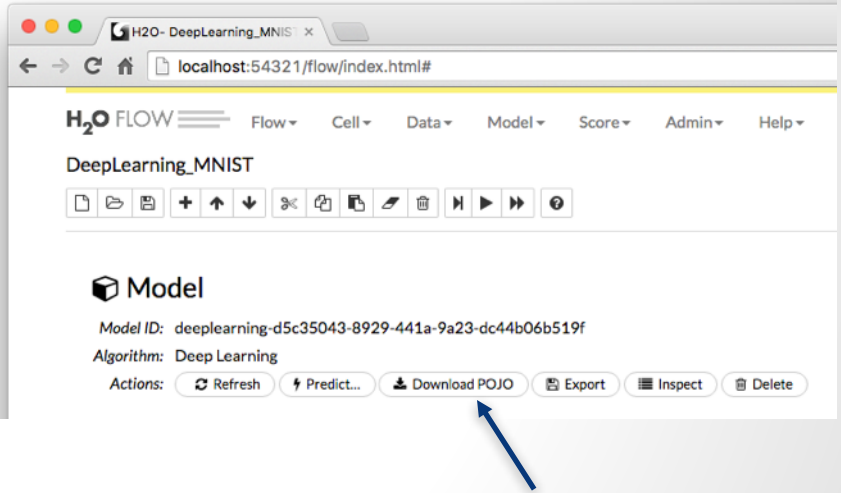
Sparkling Water 2.0

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



Export Plain Old Java Object (POJO)

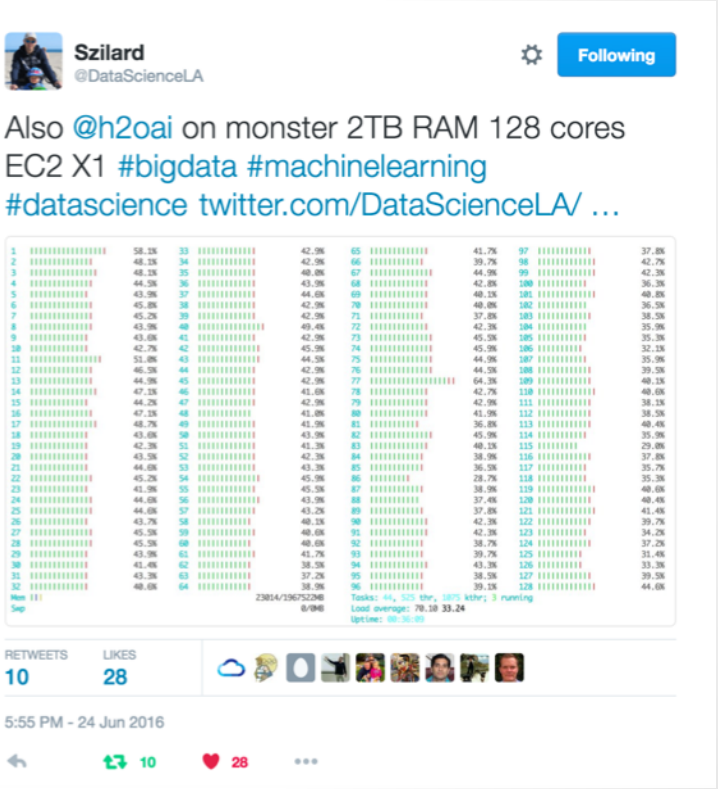




Why H2O?

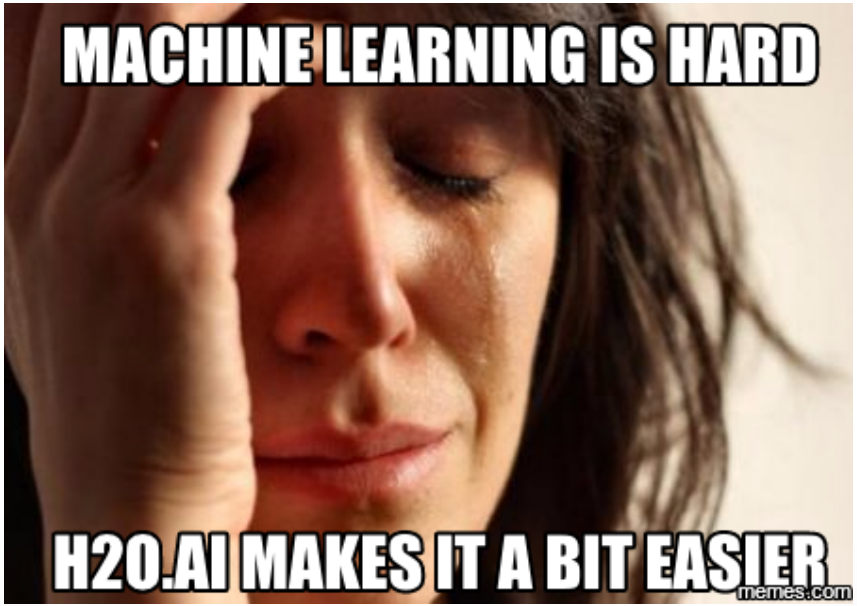
Szilard Pafka – Chief Data Scientist at Epoch

- 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?

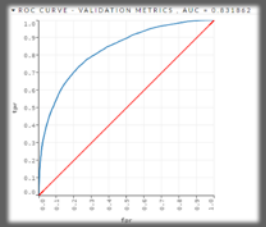
	R	Spark ML 1.3	Radoop	H_2O
Easy to work with	😐	☹	☺	☺
Cost efficient	☺	☺	☹	☺
Provide the methods that we normally use	☺	😐	☺	☺
Easy-to-use real-time capability	☹	☹	☺	☺
Can utilize our hardware setup	☹	☺	☺	☺

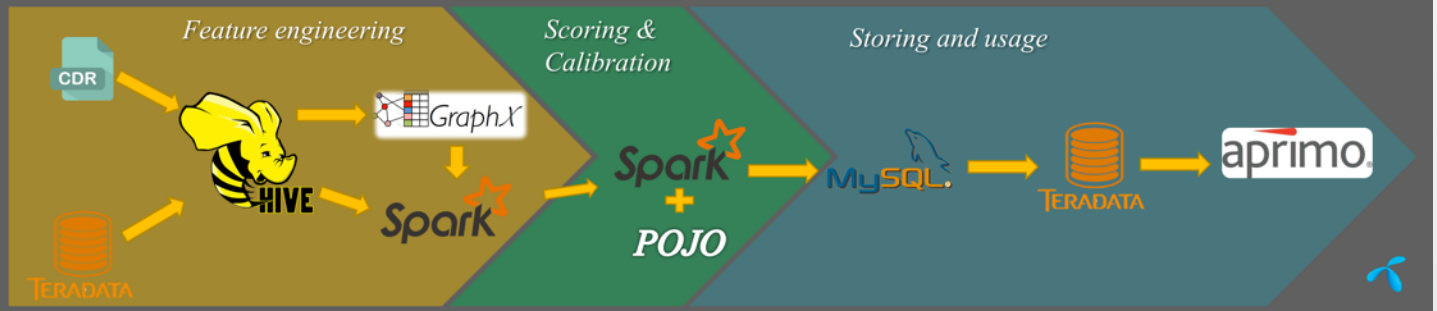
Telenor – Use Case

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.

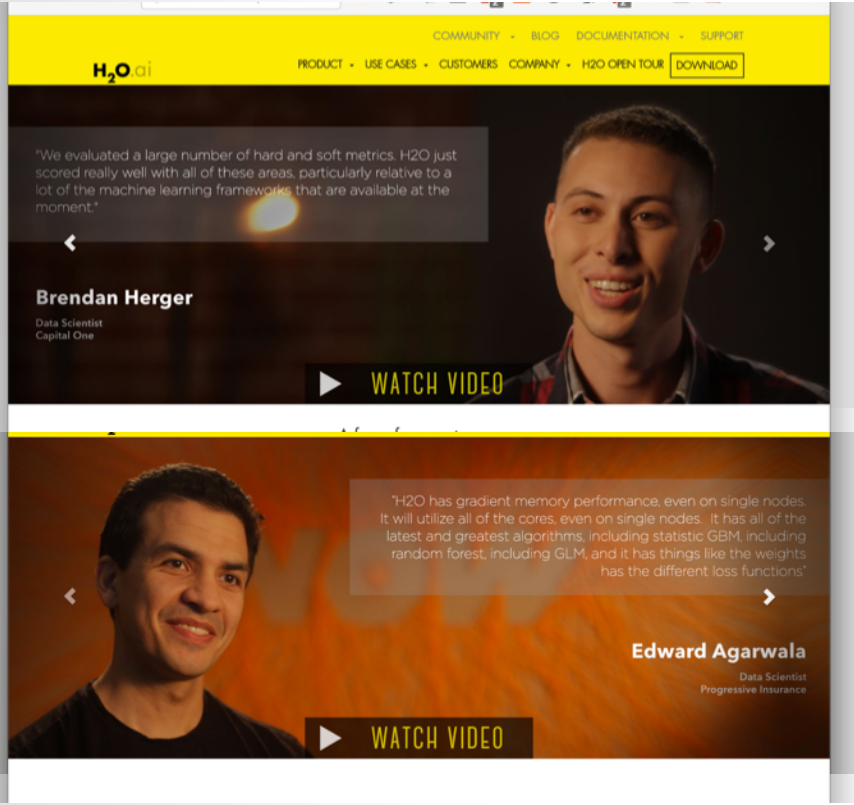
In this small talk I'll demo how to provision SparkR with H20 allowing more algorithms to the R users of Spark. I'll discuss the pros and cons of this approach and demo a few shiny features and use cases of the H20 stack.

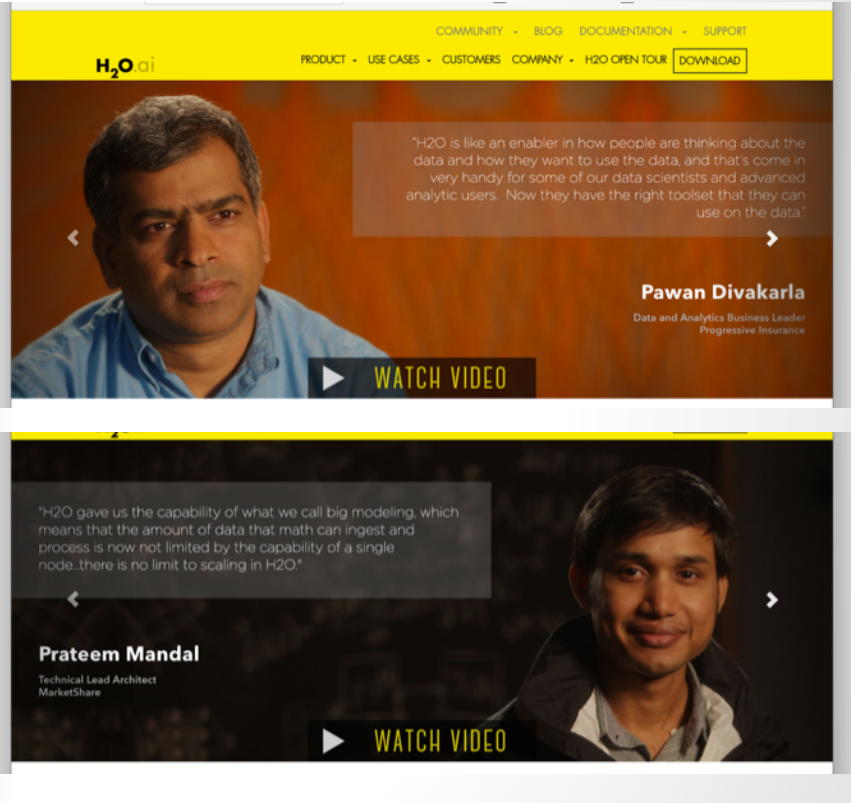
Some of the themes I'll discuss;

- 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



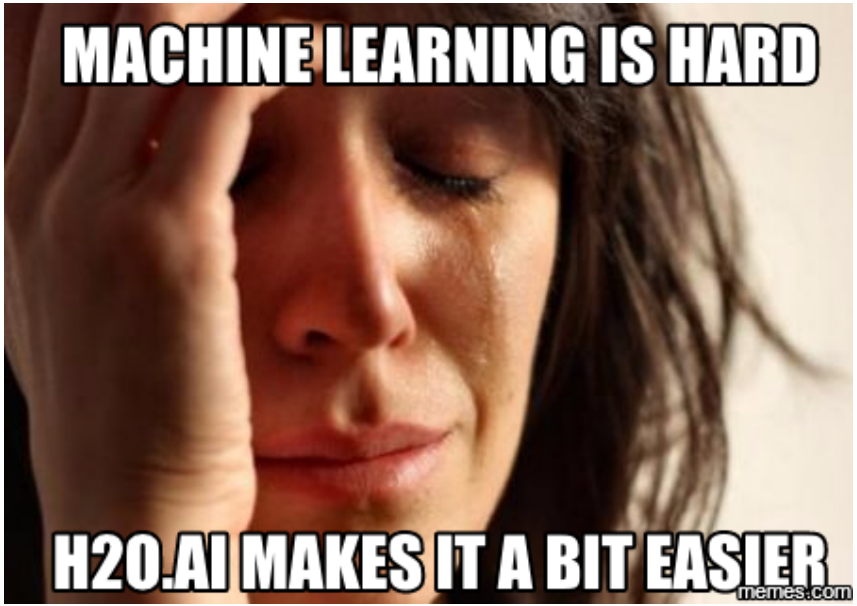


Check out the videos - **www.h2o.ai**

What's Next?

Recap – H2O's Mission

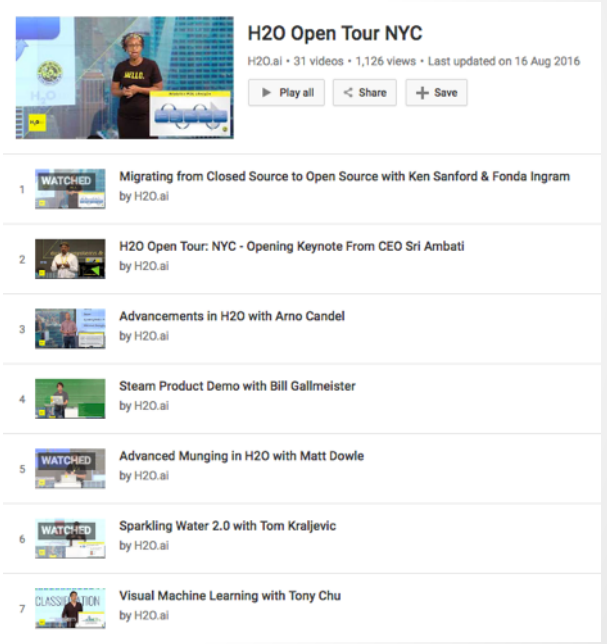
- Szilard's Summary Slide



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



Thanks!

- La Fosse Associates
- Chelsea FC
- Our Users:
  - Szilard Pafka (Epoch)
  - Liki Norbert (Telenor)
  - Vincent Warmerdam (GoDataDriven)

- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe
- Resources
  - docs.h2o.ai
- Slides and Code
  - github.com/h2oai/h2o-meetups