Introduction to Machine Learning with H2O

Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai

(draft version 0.2)

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 (30 mins)**
  - About H2O.ai
  - Demos
    - Web & R Interface
  - Why H2O?
    - Our Community
    - Our Customers
  - Moving Forward
    - 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

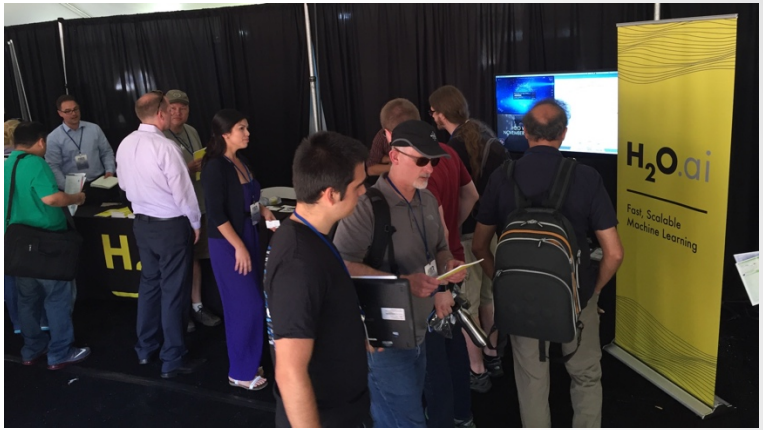
About H2O.ai

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

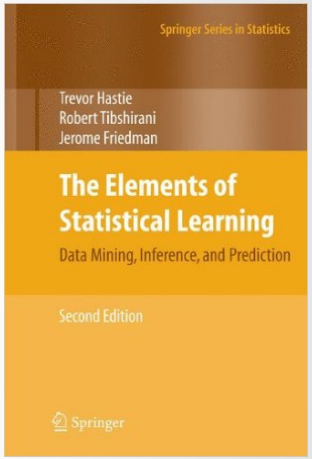
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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*

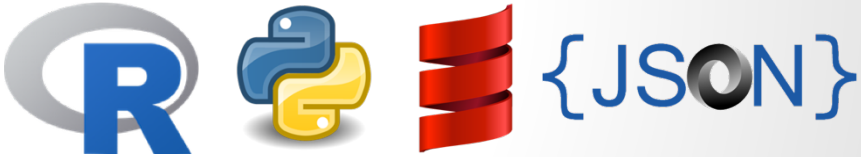
Dr. Steven Boyd

- 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 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, 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: 28.media.tumblr.com

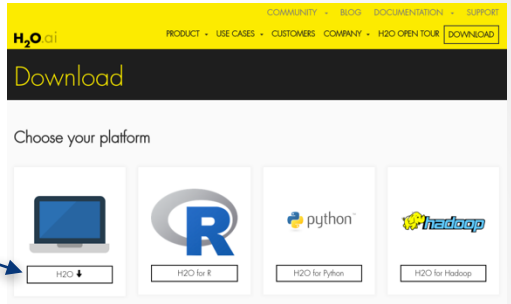
Public Dataset – Wine Quality

- **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

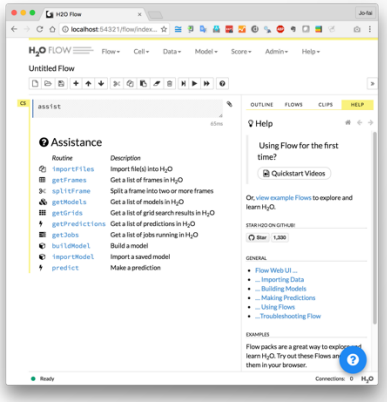
- Download and unzip **jar** from www.h2o.ai



- 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**

H2O Interfaces – R, Python & Others

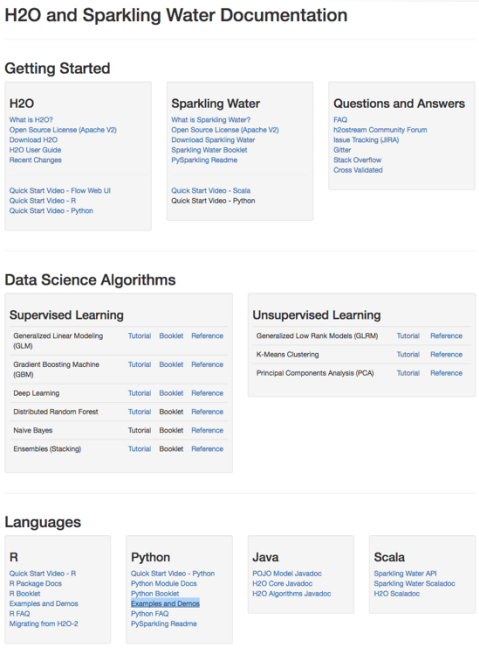
- R

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

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

- Python

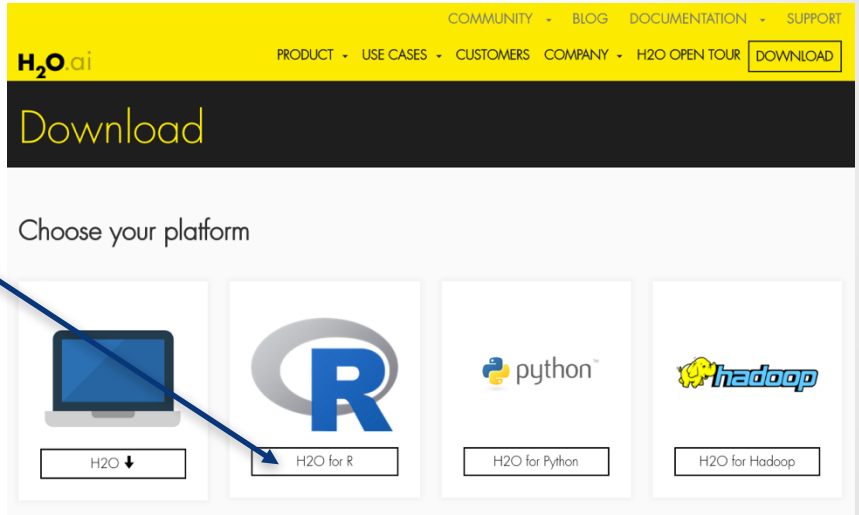
```python
# Import H2O Python module
import h2o

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

- Resources - docs.h2o.ai



H2O R 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**.

More Advanced Topics

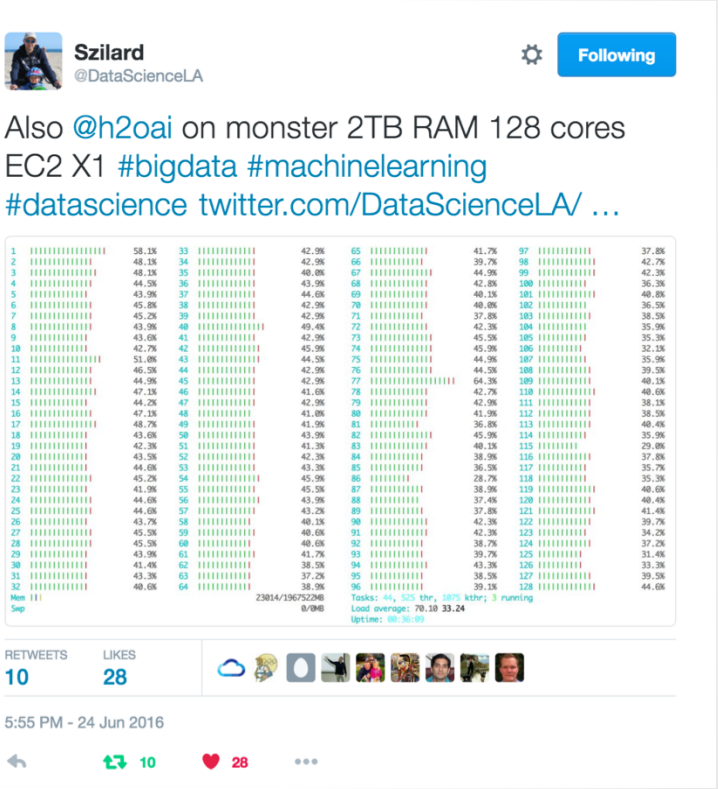
- Hyperparameters Optimization
- Model Stacking
- Export Plain Old Java Object (POJO)

- Resources:
  - <T.B.A.>

Why H2O?

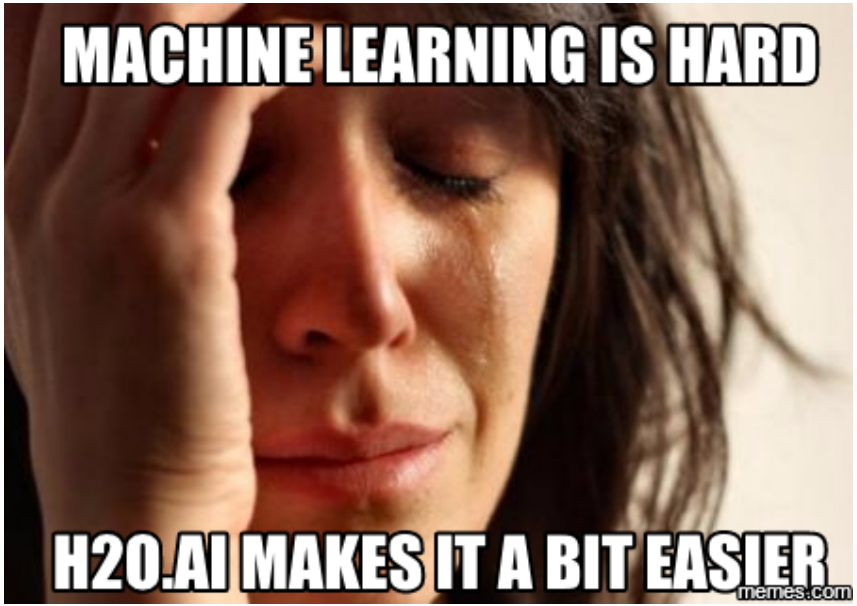
Szilard Pafka – Chief Data Scientist at Epoch

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



Szilard Pafka – Why H2O?

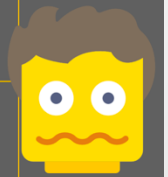
- Szilard's Summary Slide



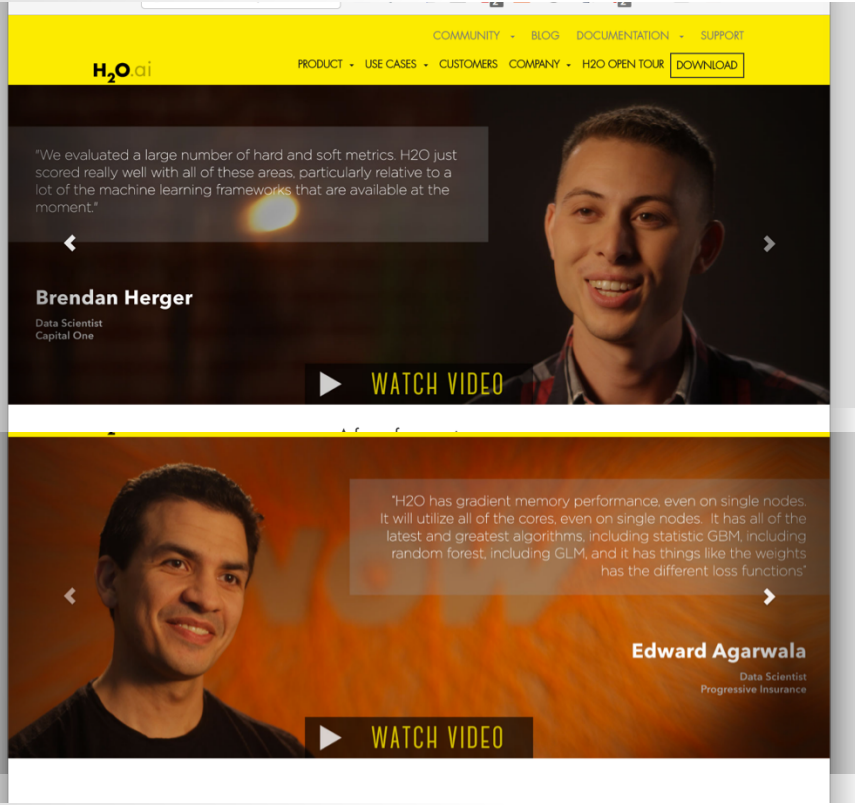
Telenor – Why H2O?

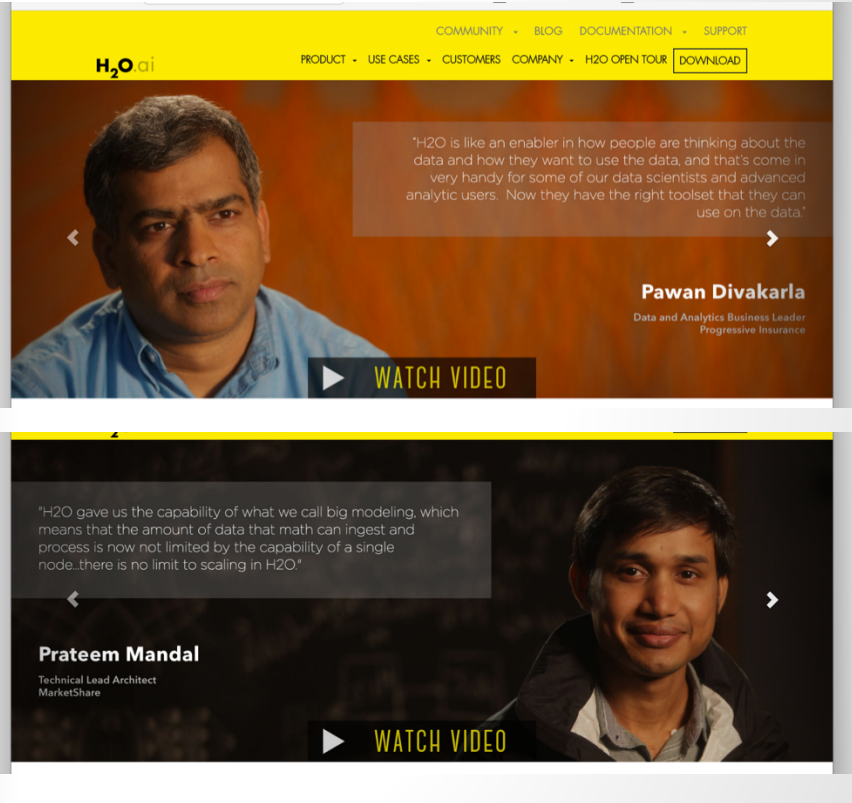
WHAT WERE THE MAIN ASPECTS WE VALUED IN A ML SOLUTION IN 2015?

	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	☹	☺	☺	☺



H2O Customers



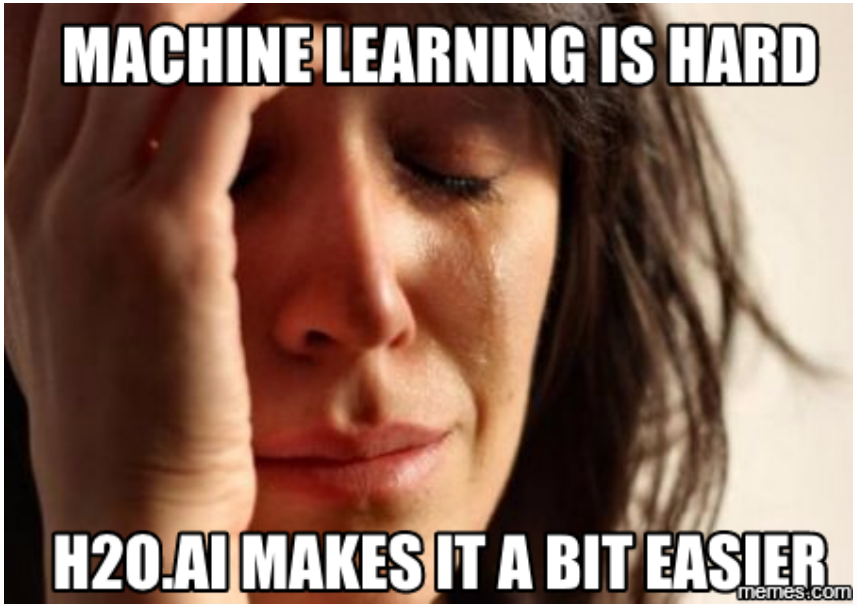


Check out the videos - **www.h2o.ai**

What’s Next?

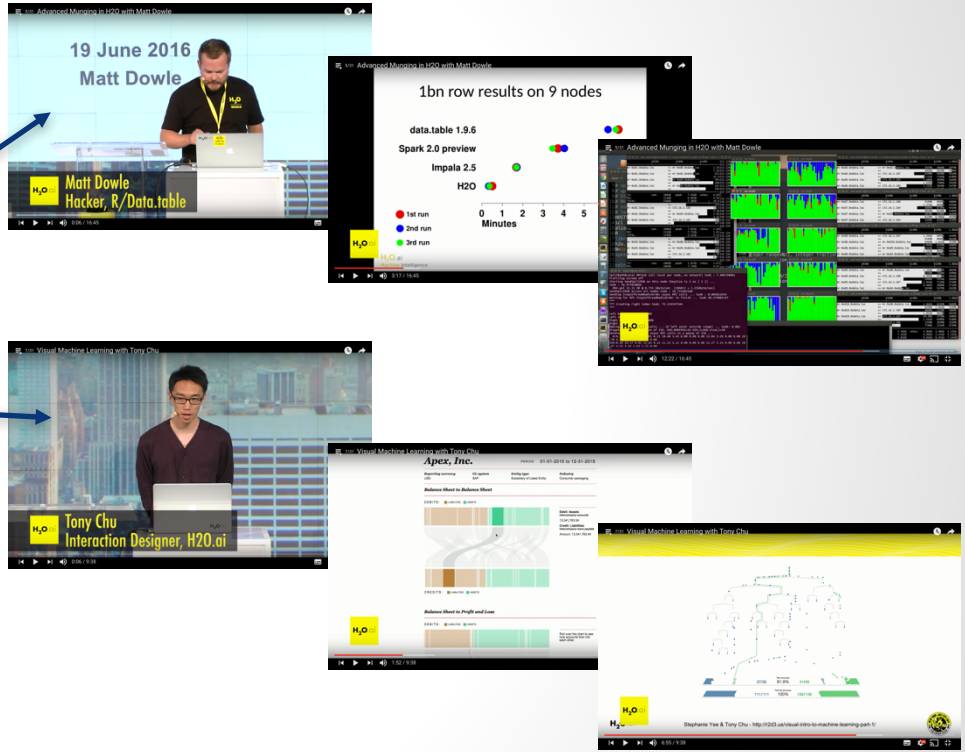
Recap – H2O's Mission

- Szilard's Summary Slide



H2O is Evolving

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



Thanks!

- Paris ML Meetup
  - Jiqiong (Ji) Qiu
  - Franck Bardol
  - Igor Carron
- Murex

- Slides & Code
  - bit.ly/h2o_paris_1
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe