

# Machine Learning with H2O



Jo-fai (Joe) Chow  
Data Scientist  
[joe@h2o.ai](mailto:joe@h2o.ai)  
@matlabulous

Budapest Data Science Meetup  
Prezi House of Ideas  
2<sup>nd</sup> 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](http://bit.ly/joe_h2o_talk2)

# I love Prezi!

## Quantifying Green Values

No description

by Jo-fai Chow on 28 June 2013 • 998 views Tweet

Comments (0)

### Popular presentations

See more [popular](#) or the [latest](#) prezis

My favorite presentation during my EngD time.

# 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



# 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*



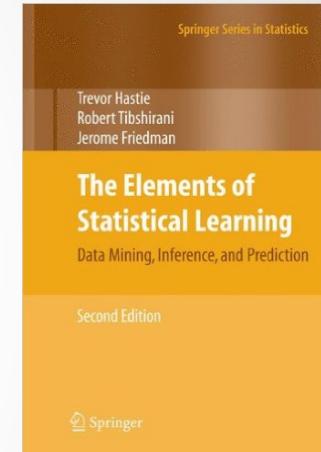
## Dr. Robert Tibshirani

- 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

See [bit.ly/joe\\_h2o\\_talk4](http://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 Flow Cell Data Model Score Admin Help

DeepLearning\_MNIST

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

Model Parameters

Scoring History - LogLoss

Scoring History - MSE

Variable Importances

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	0 / 993
1	0	1105	0	0	0	0	0	0	0	0	0	0 / 1,105

Validation Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	970	0	0	0	0	0	0	0	0	0	0	0 / 970
1	0	1125	4	0	1	2	0	3	0	0	0	0.0088 10 / 1,125

Output - Status of Neuron Layers (Predicting C795, 10-class classification, multinomial distribution, crossentropy loss, 100,810 weights/biases, 899.2 KB, 9,240,000 training samples, mini-batch size: 1)

Output - Scoring History

Output - Training Metrics

Output - Validation Metrics

Output - Validation Metrics - Top-10 Hit Ratios

Output - Variable Importances

Preview POJO

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

Scoring History - LogLoss

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	0 / 993
1	0	1105	0	0	0	0	0	0	0	0	0	0 / 1,105

Validation Metrics - Confusion Matrix Vertical: Actual; Across: Predicted

	0	1	2	3	4	5	6	7	8	9	Error	Rate
0	970	0	0	0	0	0	0	0	0	0	0	0 / 970
1	0	1125	4	0	1	2	0	3	0	0	0	0.0088 10 / 1,125

Output - Status of Neuron Layers (Predicting C795, 10-class classification, multinomial distribution, crossentropy loss, 100,810 weights/biases, 899.2 KB, 9,240,000 training samples, mini-batch size: 1)

Output - Scoring History

Output - Training Metrics

Output - Validation Metrics

Output - Validation Metrics - Top-10 Hit Ratios

Output - Variable Importances

Preview POJO

Ready

# 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](https://docs.h2o.ai)

## H2O and Sparkling Water Documentation

### Getting Started

**H2O**  
What is H2O  
Open Source License (Apache V2)  
Download H2O  
H2O User Guide  
Recent Changes

Quick Start Video - Flow Web UI  
Quick Start Video - R  
Quick Start Video - Python

**Sparkling Water**  
What is Sparkling Water?  
Open Source License (Apache V2)  
Download Sparkling Water  
Sparkling Water Booklet  
PySparkling Readme

Quick Start Video - Scala  
Quick Start Video - Python

**Questions and Answers**  
FAQ  
Inboxstream Community Forum  
Issue Tracking (JIRA)  
Gitter  
Stack Overflow  
Cross Validated

## Data Science Algorithms

### Supervised Learning

Generalized Linear Modeling (GLM)

Gradient Boosting Machine (GBM)

Deep Learning

Distributed Random Forest

Naïve Bayes

Ensembles (Stacking)

Tutorial Booklet Reference

### Unsupervised Learning

Generalized Low Rank Models (GLRM)

K-Means Clustering

Principal Components Analysis (PCA)

Tutorial Reference

Tutorial Reference

Tutorial Reference

## Languages

### R

Quick Start Video - R  
R Package Docs  
R Booklet  
Examples and Demos  
R FAQs  
Migrating from H2O-2

Python  
Quick Start Video - Python  
Python Module Docs  
Python Booklet  
**Examples and Demos**  
Python FAQs  
PySparkling Readme

Java  
POJO Model JavaDoc  
H2O Core JavaDoc  
H2O Algorithms JavaDoc

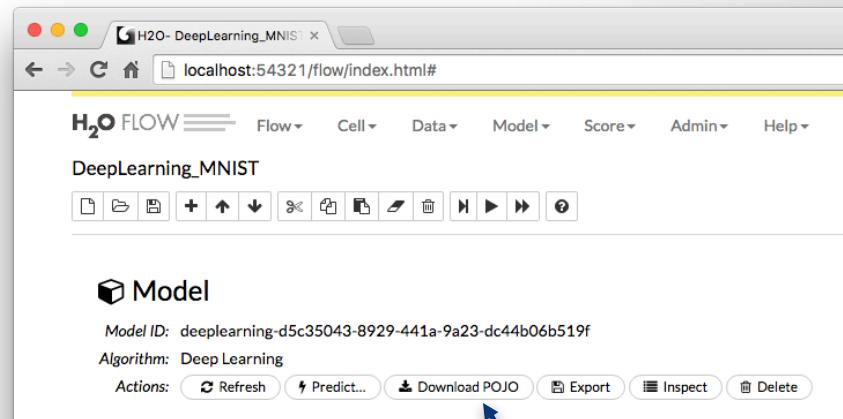
Scala  
Sparkling Water API  
Sparkling Water Scaladoc  
H2O Scaladoc

- Python

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

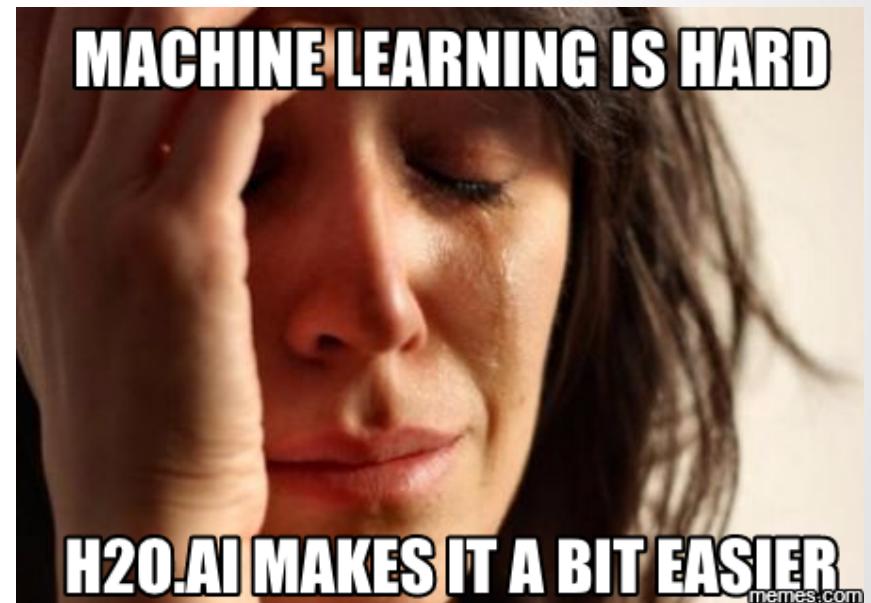
# Export Plain Old Java Object (POJO)

The screenshot shows the H2O Flow interface with the title "DeepLearning\_MNIST". The main area displays the generated Java code for a "DeepLearningModel" named "deeplearning\_d5c35043\_8929\_441a\_9a23\_dc44b06b519f". The code includes imports for java.util.Map, hex.gm.GemModel, and hex.gm.annotations.ModelPojo; a class definition for "DeepLearningModel"; and a static final class "NORMALUL" that implements java.io.Serializable. The "NORMALUL" class contains a static final double[] "VALUES" with 26 elements, each with a value like 0.1838291371915183. Below the code, a note says: "How to download, compile and execute: mkdir tmpdir cd tmpdir curl -O http://127.0.0.1:54321/h2o-genmodel.jar > h2o-genmodel.jar curl -O http://127.0.0.1:54321/H2olets.java</DeepLearning-d5c35043\_8929\_441a\_9a23\_dc44b06b519f> > deeplearning.java javac -cp h2o-genmodel.jar -D Xmx2g -D XX:MaxPermSize=128m deeplearning\_d5c35043\_8929\_441a\_9a23\_dc44b06b519f (Note: Try java argument -XX:+PrintCompilation to show runtime JIT compiler behavior.)". The bottom status bar says "Ready" and "Connections: 0".



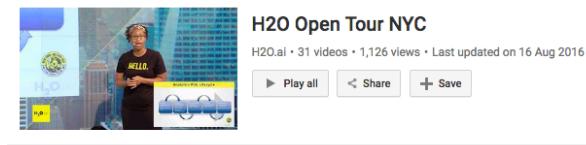
# Why H2O?

- Previous Budapest DS Meetup
  - Big thanks to Szilard Pafka
    - Intro to ML with H2O
  - [Link \(video\)](#)
  - [Link \(Slides\)](#)
- Szilard's Summary Slide



# H2O is Evolving

- Advanced data munging
  - Visual ML
  - Deep Water
    - H2O + TensorFlow, mxnet, Theano, Caffe ...
    - GPU
  - Steam
- YouTube Playlist



**H2O Open Tour NYC**  
H2O.ai • 31 videos • 1,126 views • Last updated on 16 Aug 2016  
▶ Play all   < Share   + Save

Rank	Video 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

# H2O at satRday #1

- **Hands-on Workshop (0800)**
  - H2O R package
  - Scalable machine learning
  - Model training, tuning and stacking
- **Machine Learning Use Cases (1600)**
  - How others use H2O for their projects

# Three More H2O Talks

- Zsolt Toth
  - Rapid Miner + H2O Integration (20 mins)
- Norbert Liki
  - Data Mining with H2O at Telenor (20 mins)
- Jakub (Kuba) Háva
  - Sparkling Water 2.0 (35 mins)

# We're Hiring!

- Check it out
- [www.h2o.ai/careers/](http://www.h2o.ai/careers/)

## Open Positions

Location	Position
Mountain View, CA	<a href="#">UI Engineers</a>
Mountain View, CA	<a href="#">Algorithm Engineers</a>
Mountain View, CA	<a href="#">Solutions Architect, Data Engineering</a>
Mountain View, CA	<a href="#">Distributed Systems Platform Engineer</a>
Mountain View, CA	<a href="#">Customer Support Manager</a>
Mountain View, CA	<a href="#">Quality Engineer</a>
Multiple	<a href="#">Program Manager</a>
Multiple	<a href="#">Solutions Data Scientist</a>
Multiple	<a href="#">Data Journalist</a>

To apply, send your resume to [careers@h2o.ai](mailto:careers@h2o.ai)

# Thanks!

- Prezi
  - Organizers
    - Budapest DS Meetup
    - satRdays
  - Szilard
  - H2O.ai
- 
- Contact
    - [joe@h2o.ai](mailto:joe@h2o.ai)
    - [@matlabulous](https://twitter.com/matlabulous)
    - [github.com/woobe](https://github.com/woobe)
  - Resources
    - [docs.h2o.ai](https://docs.h2o.ai)
  - Slides and Code
    - [github.com/h2oai/h2o-meetups](https://github.com/h2oai/h2o-meetups)