Making Machine Learning Accessible to Everyone

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

Data Science Warsaw
Centrum Nowych Technologii UW
9th November, 2016

About Me: Civil Engineer → Data Scientist

- 2005 - 2015
- Water Engineer
  - Consultant for Utilities
  - Industrial PhD
    - Water Engineering + Machine Learning
    - *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
  - About H2O.ai
  - Machine Learning Demo
    - Web Interface & R
  - Why H2O?
  - What's Next?
- Coming Up …
  - Sparkling Water 2.0
  - R + H2O Showcase
  - Deep Water
    - H2O's integration with other deep learning libraries

About H2O.ai

About H2O.ai

- H2O.ai, the Company
  - Team: 80
  - Founded in 2012
  - HQ: Mountain View, California
- H2O, the Platform
  - Open Source (Apache 2.0)
  - Algorithms written in Java
    - Fast, distributed and scalable
  - Multiple interfaces to suit different users
    - Web, R, Python, Java, Scala, REST/JSON
  - Works with desktop/laptop, cloud, Spark and Hadoop

Scientific Advisory Council

Dr. Trevor Hastie

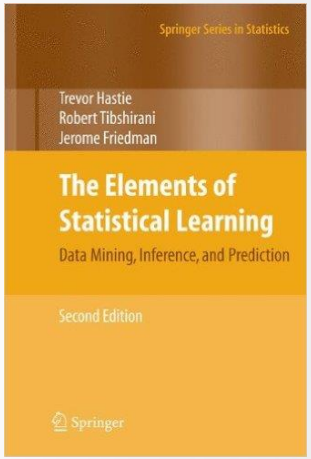
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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

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's Mission

Making Machine Learning Accessible to Everyone

Photo credit: Virgin Media

Machine Learning Demo

A Typical Machine Learning Task

- Demo
  - Dataset – MNIST
    - LeCun et al. (1999)
    - Hand-written Digits
  - Import & Explore Data
  - Build & Evaluate Models
  - Make Predictions

Photo credit: http://www.opendeep.org/v0.0.5/docs/tutorial-classifying-handwritten-mnist-images

MNIST Hand-Written Digits

- 784 Inputs
  - 28 x 28 = 784 pixels
- 1 Output
  - 0, 1, 2, 3, 4, 5, 6, 7, 8 or 9
  - Classification
- Files
  - Train (42k Records)
  - Test (28k)
- Links
  - https://www.kaggle.com/c/digit-recognizer/data

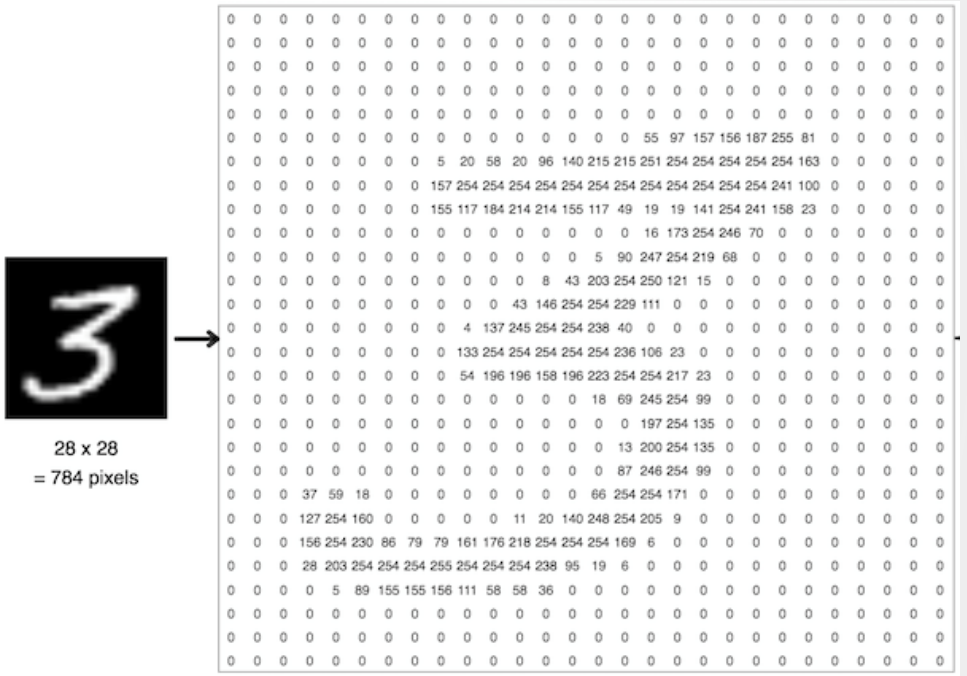


Photo credit: https://ml4a.github.io/ml4a/neural_networks/

A Simple Neural Network

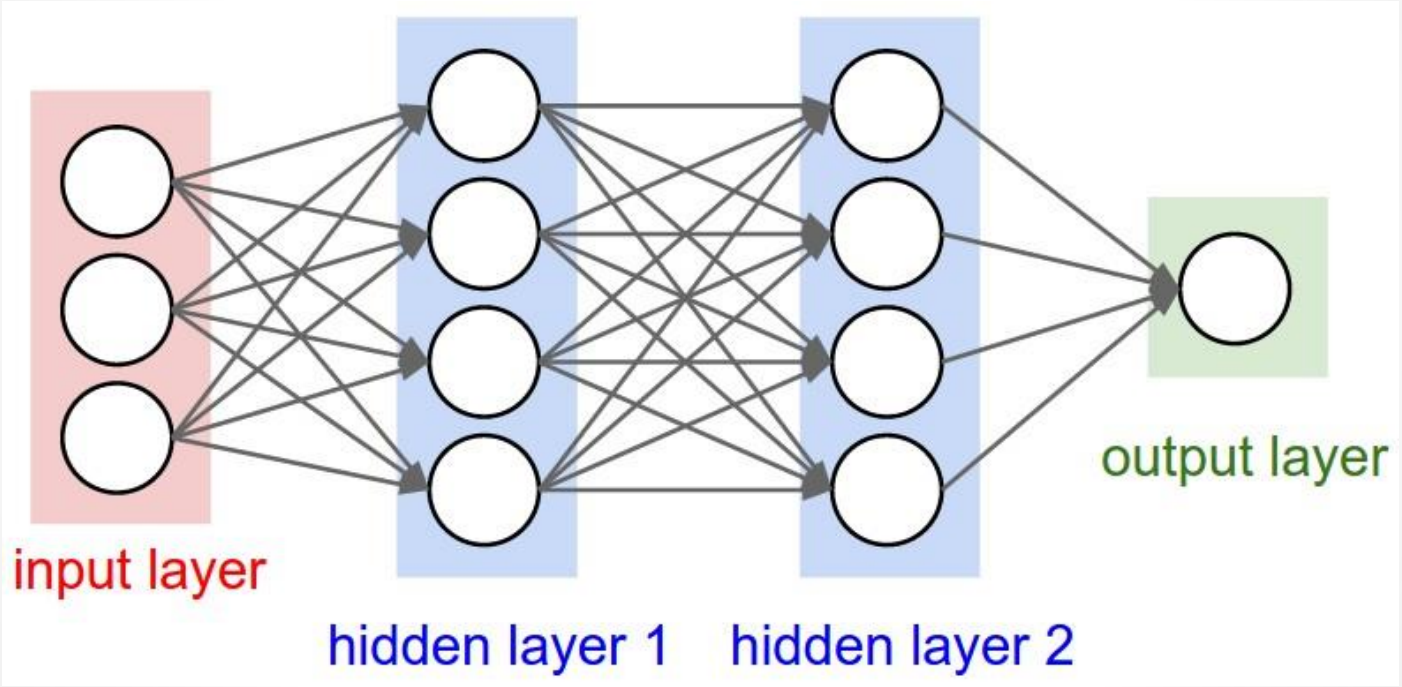


Image credit: http://cs231n.github.io/

H2O Deep Learning in Action

116M rows, 6GB CSV file
800+ predictors (numeric + categorical)

airlines_all_selected_cols.hex

Actions: View Data | Split... | Build Model... | Predict | Download | Export

Rows	Columns	Compressed Size
116695259	12	2GB

Job

Run Time 00:00:36.712
Remaining Time 00:00:17.188
Type Model
Key deeplearning-dd2f42f7-81f7-42e8-9d98-e34437309828
Description DeepLearning
Status RUNNING
Progress 69%

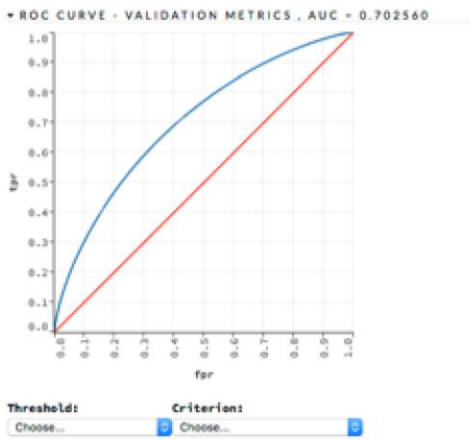
Iterations: 12. Epochs: 0.628821. Speed: 2,243,735 samples/sec. Estimated time left: 21.849 sec

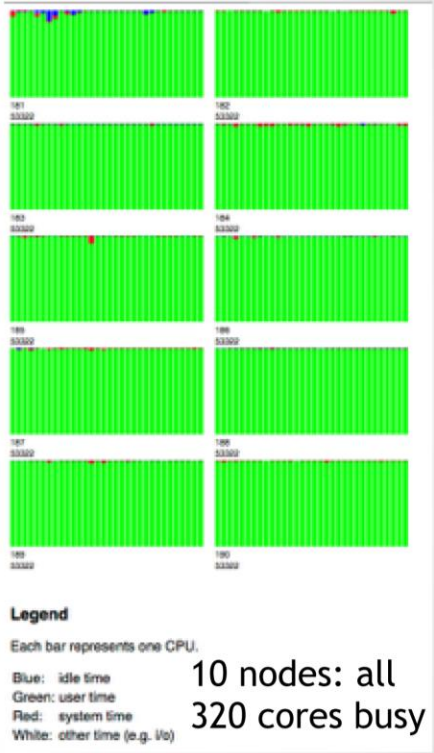
Actions: View | Cancel Job

model trained in <1 min:
2M+ samples/second

▾ OUTPUT - STATUS OF NEURON LAYERS (PREDICTING ISDEPDELAYED, 2-CLASS CLASSIFICATION, BERNOULLI DISTRIBUTION, CROSSENTROPY LOSS, 17,462 WEIGHTS/BIASES, 221.3 KB, 106,585,365 TRAINING SAMPLES, MINI-BATCH SIZE 1)

layer	units	type	dropout	l1	l2	mean_rate	rate_RMS	momentum	mean_weight	weight_RMS	mean_bias	bias_RMS
1	807	Input	0									
2	20	Rectifier	0	0	0	0.0483	0.2020	0	-0.0021	0.2111	-0.9139	1.0036
3	20	Rectifier	0	0	0	0.0197	0.0227	0	-0.1033	0.5362	-1.3908	1.5259
4	20	Rectifier	0	0	0	0.0517	0.0446	0	-0.1575	0.3068	-0.0846	0.6046
5	20	Rectifier	0	0	0	0.0761	0.0844	0	-0.0374	0.2275	-0.2647	0.2481
6	2	Softmax		0	0	0.0161	0.0083	0	0.0741	0.7260	0.4269	0.2056





10 nodes: all
320 cores busy

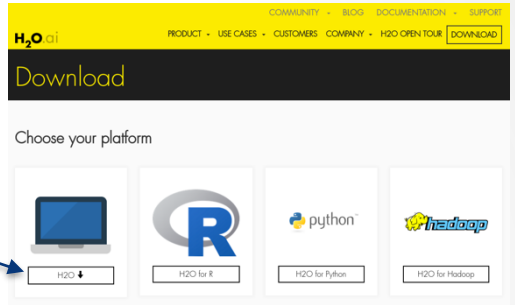
real-time, interactive
model inspection in Flow

H2O.ai Deep Learning Model

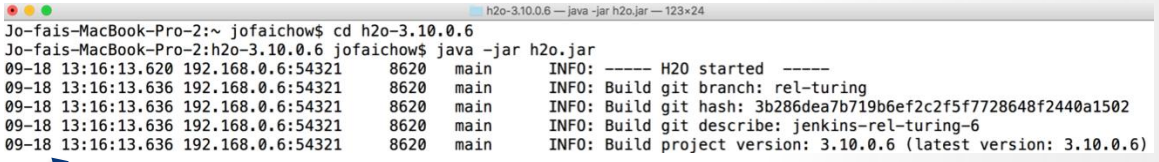
H2O Flow Demo

H2O Flow (Web Interface) Demo

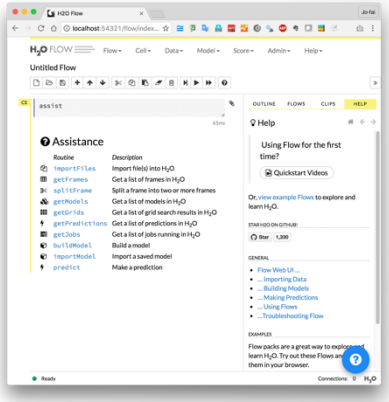
- Download and unzip **jar** from www.h2o.ai



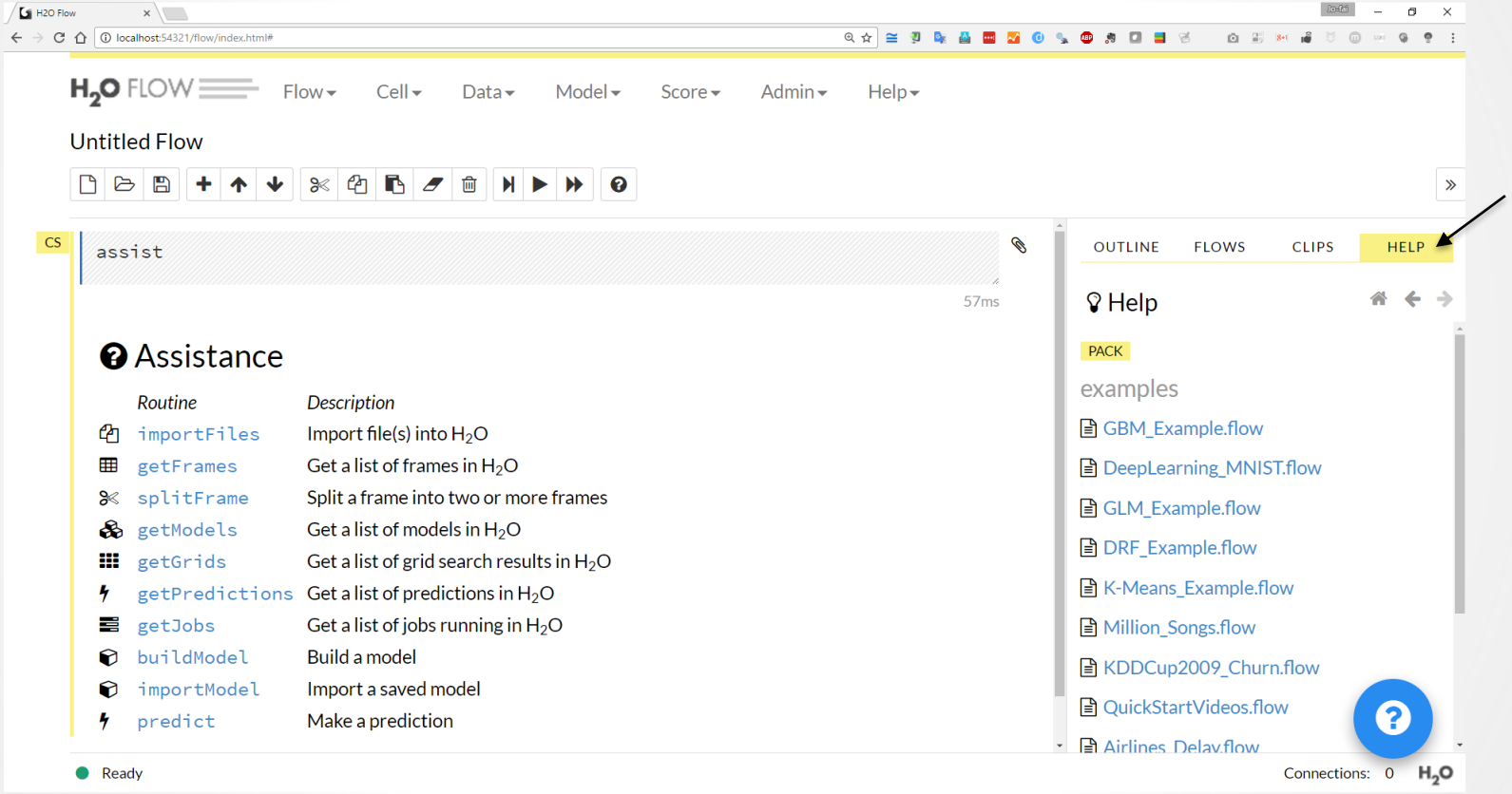
- In terminal:
  - java -jar h2o.jar



- Web browser:
  - localhost:54321



H2O Flow Examples



H2O Flow Demo

- Tasks:
  - Import data
  - Visualize data
  - Split data
    - 80% training
    - 20% validation
  - Build models

- Tasks:
  - Evaluate models
  - Use models for predictions
  - Grid search

H2O + R Demo

Other H2O Interfaces

- R

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

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

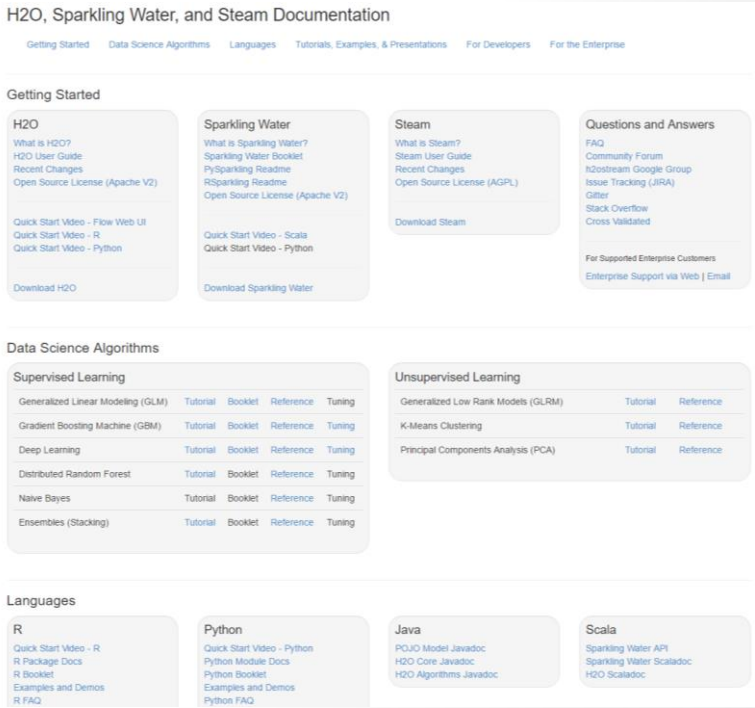
- Python

```python
# Import H2O Python module
import h2o

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

Key Resources

- docs.h2o.ai



Advanced Features

- Advanced Features
  - Hyperparameters Tuning
  - Model Stacking
  - Saving/Loading Models
  - Export Plain Old Java Object (POJO)

- Key Resources
  - docs.h2o.ai

- Joe's Previous H2O Talks
  - bit.ly/joe_h2o_talk3
  - bit.ly/h2o_budapest_1
  - bit.ly/h2o_paris_1
  - bit.ly/h2o_milan_1

Why H2O?


H2O.ai
Products
Industries
Customers
Community
Documentation
Download
Stories of AI Transformation

H2O In Action

www.h2o.ai/customers

Capital One

Capital One uses H2O open source machine learning for various use cases.

MarketShare

H2O predictive analytics helps boost the impact and results of digital marketing.

Kaiser

Kaiser uses H2O machine learning to save lives.

Zurich Insurance

Zurich turned to H2O as a strategic differentiator for commercial insurance.

Progressive

Progressive uses H2O predictive analytics for user-based insurance.

Comcast

Comcast uses H2O to improve customer experience.

Hospital Corporation of America

HCA uses H2O to predict patient outcomes in real-time.

McKesson

McKesson discusses the adoption of artificial intelligence in healthcare.

Macy's

Macy's uses H2O for personalized site recommendations.

Transamerica

Transamerica turns to H2O to develop a product recommendation platform for insurance.

Paypal

Paypal turned to H2O Deep Learning for fraud detection and customer churn.

eBay

eBay chose H2O for open source machine learning.

H2O for Kaggle Competitions

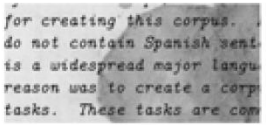
CIFAR-10 Competition Winners: Interviews with Dr. Ben Graham, Phil Culliton, & Zygmunt Zając

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

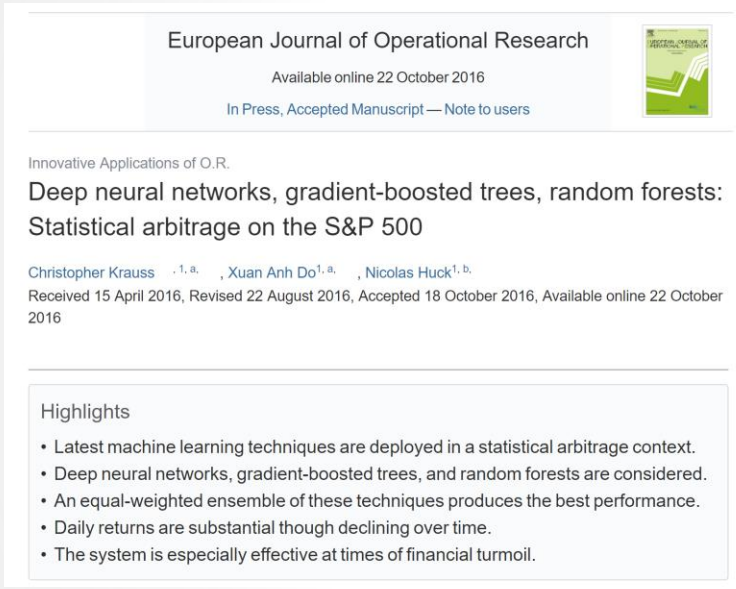
Denoising Dirty Documents

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 for Academic Research



European Journal of Operational Research

Available online 22 October 2016

In Press, Accepted Manuscript — Note to users

Innovative Applications of O.R.

Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500

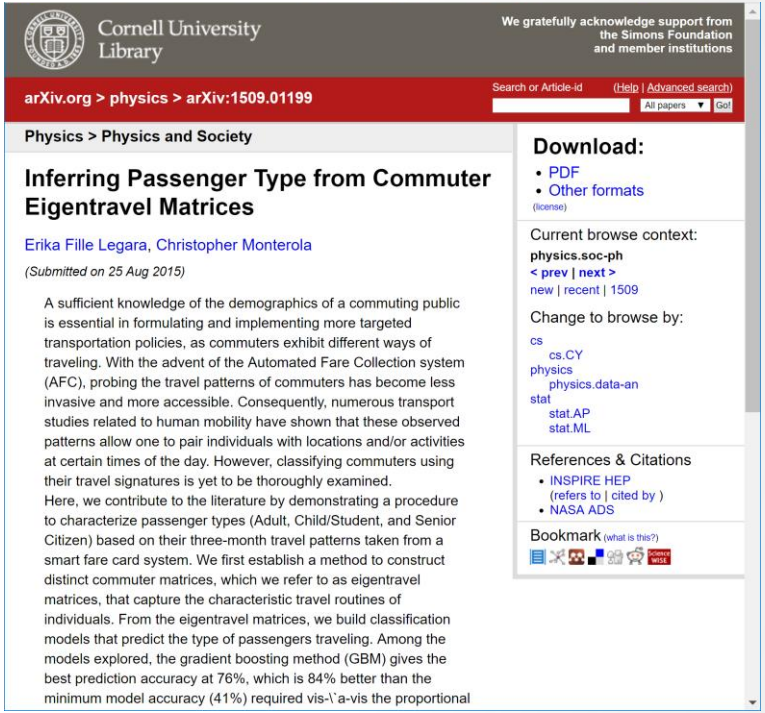
Christopher Krauss [1, a], Xuan Anh Do [1, a], Nicolas Huck [1, b]

Received 15 April 2016, Revised 22 August 2016, Accepted 18 October 2016, Available online 22 October 2016

Highlights

- Latest machine learning techniques are deployed in a statistical arbitrage context.
- Deep neural networks, gradient-boosted trees, and random forests are considered.
- An equal-weighted ensemble of these techniques produces the best performance.
- Daily returns are substantial though declining over time.
- The system is especially effective at times of financial turmoil.

http://www.sciencedirect.com/science/article/pii/S0377221716308657



Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > physics > arXiv:1509.01199

Search or Article-id (Help | Advanced search) All papers Go!

Physics > Physics and Society

Inferring Passenger Type from Commuter Eigentravel Matrices

Erika Fille Legara, Christopher Monterola

(Submitted on 25 Aug 2015)

A sufficient knowledge of the demographics of a commuting public is essential in formulating and implementing more targeted transportation policies, as commuters exhibit different ways of traveling. With the advent of the Automated Fare Collection system (AFC), probing the travel patterns of commuters has become less invasive and more accessible. Consequently, numerous transport studies related to human mobility have shown that these observed patterns allow one to pair individuals with locations and/or activities at certain times of the day. However, classifying commuters using their travel signatures is yet to be thoroughly examined. Here, we contribute to the literature by demonstrating a procedure to characterize passenger types (Adult, Child/Student, and Senior Citizen) based on their three-month travel patterns taken from a smart fare card system. We first establish a method to construct distinct commuter matrices, which we refer to as eigentravel matrices, that capture the characteristic travel routines of individuals. From the eigentravel matrices, we build classification models that predict the type of passengers traveling. Among the models explored, the gradient boosting method (GBM) gives the best prediction accuracy at 76%, which is 84% better than the minimum model accuracy (41%) required vis-\`a-vis the proportional

Download:

- PDF
- Other formats

(license)

Current browse context: physics.soc-ph

< prev | next >

new | recent | 1509

Change to browse by:

cs

cs.CY

physics

physics.data-an

stat

stat.AP

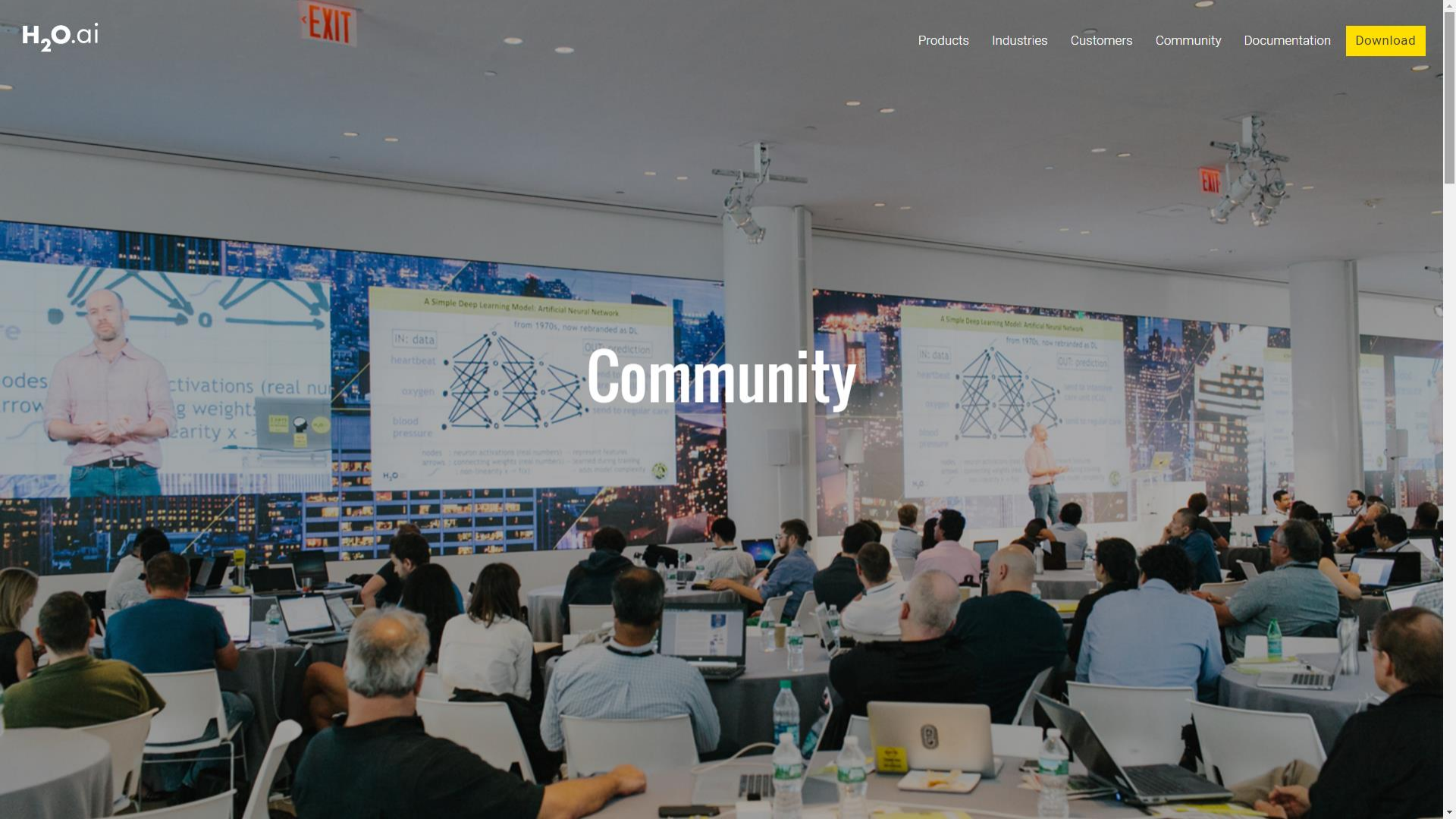
stat.ML

References & Citations

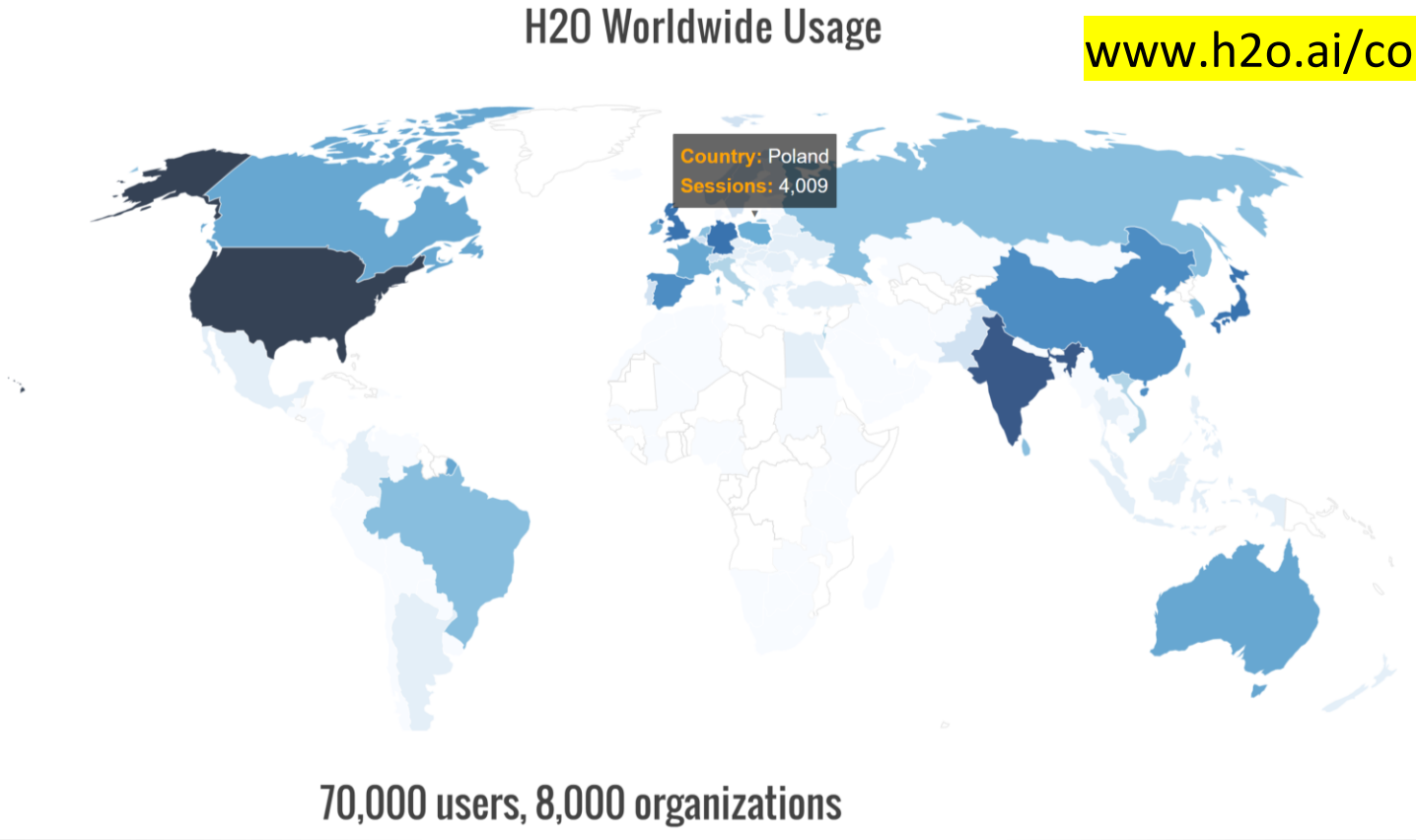
- INSPIRE HEP (refers to | cited by)
- NASA ADS

Bookmark (what is this?)

https://arxiv.org/abs/1509.01199


H2O.ai
Products
Industries
Customers
Community
Documentation
Download
EXIT
A Simple Deep Learning Model: Artificial Neural Network
from 1970s, now rebranded as DL
IN: data
heartbeat
oxygen
blood pressure
OUT: prediction
Community

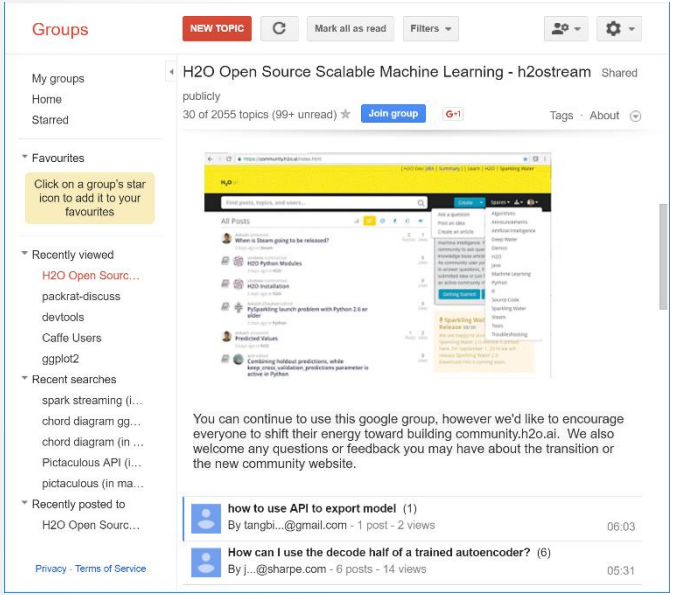
H2O Worldwide Usage



www.h2o.ai/community

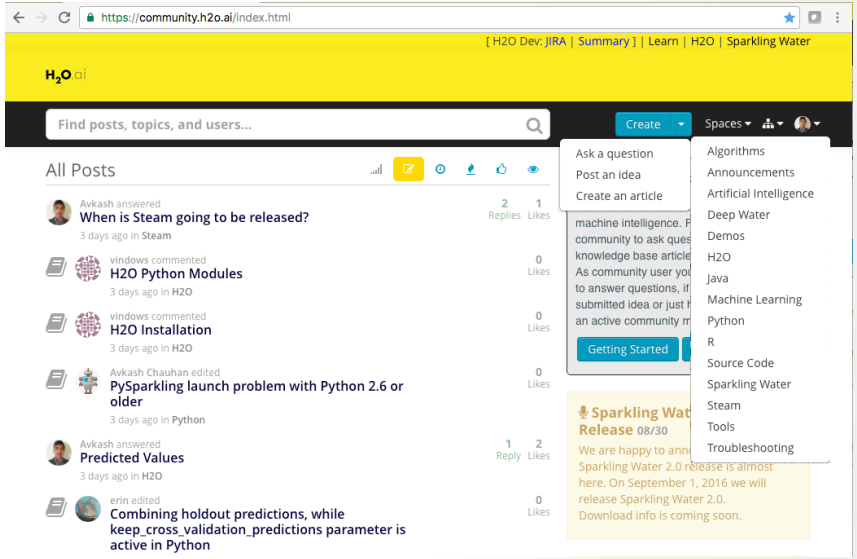
H2O Community Support

Google forum – h2osteam



Please try

community.h2o.ai



#AroundTheWorldWithH2Oai

London Kaggle Meetup

Strata Hadoop London

Chelsea FC

Big Data London

PyData Amsterdam

useR! 2016 Stanford

satRdays Budapest

Paris ML Meetup

Data Science Milan

What’s Next?

H2O is Evolving

- New Features
  - Advanced Data Munging
  - Visual ML
  - Auto ML
  - Steam
  - Deep Water
  - Sparkling Water 2.0
- Find Out More
  - H2O YouTube Playlist
    - Open Tour New York
    - Open Tour Dallas
    - PyData DC 2016

H2O's Mission

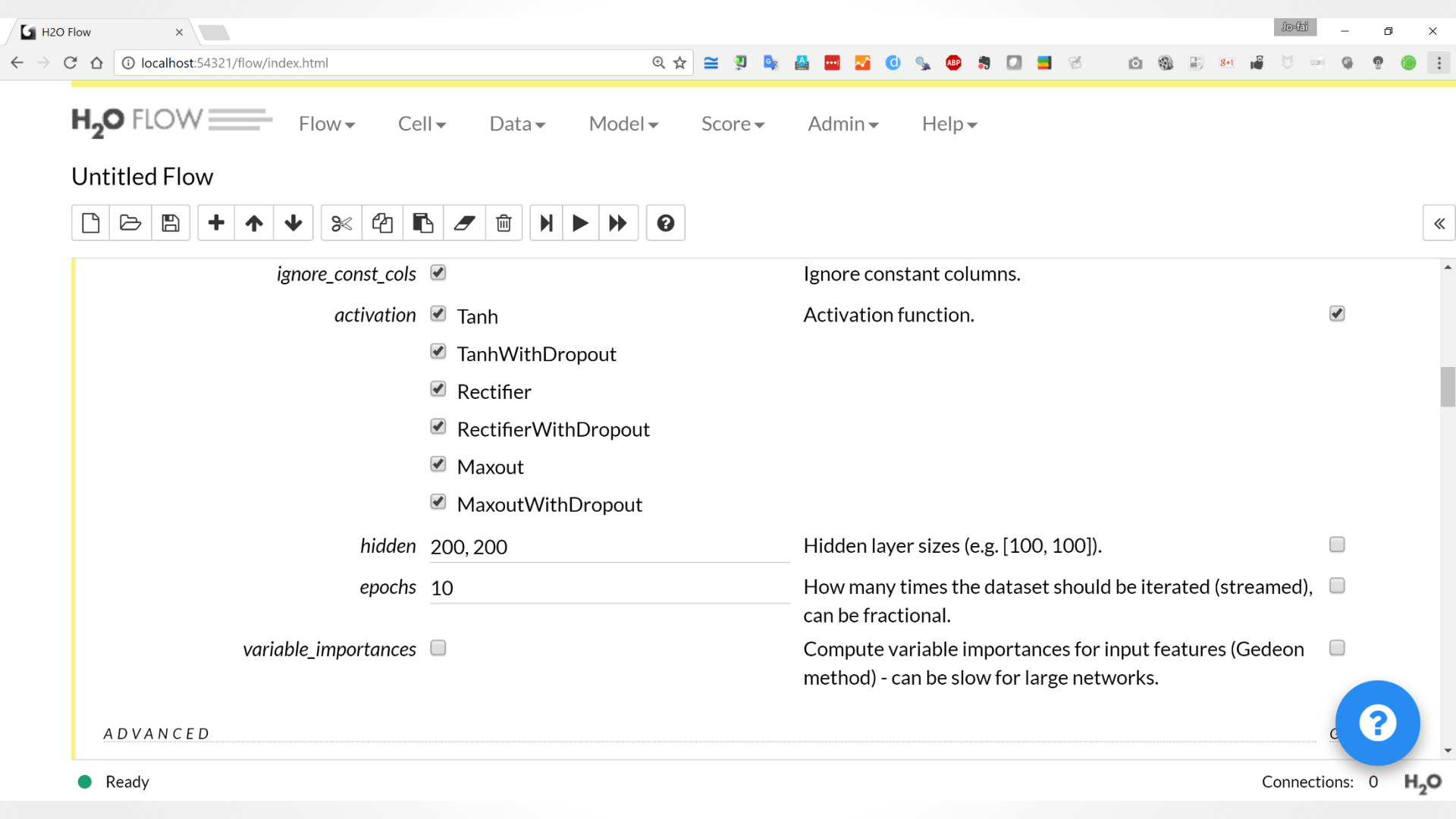
Making Machine Learning Accessible to Everyone

Photo credit: Virgin Media

Dziękuję bardzo!

- Data Science Warsaw
  - Dominik Batorski
  - Wit Jakuczun
- Slides & Code
  - bit.ly/h2o_warsaw_1
- Resources
  - github.com/h2oai/h2o-meetups
  - www.h2o.ai
  - docs.h2o.ai
- Contact
  - joe@h2o.ai
  - @matlabulous

Extra Slides


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
ignore_const_cols
Ignore constant columns.
activation
Tanh
TanhWithDropout
Rectifier
RectifierWithDropout
Maxout
MaxoutWithDropout
Activation function.
hidden
200, 200
Hidden layer sizes (e.g. [100, 100]).
epochs
10
How many times the dataset should be iterated (streamed), can be fractional.
variable_importances
Compute variable importances for input features (Gedeon method) - can be slow for large networks.
ADVANCED
Ready
Connections: 0

Untitled Flow

108ms

Grid Search

Inspect Grid Summary | Inspect Scoring History | Inspect Models | Delete selected

Key	Actions
grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_1	Predict... Inspect
grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_0	Predict... Inspect
grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_2	Predict... Inspect

CS

```
grid inspect 'summary', getGrid "grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9"
```

101ms

activation	model_ids	logloss
RectifierWithDropout	grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_1	0.18719295946954645
TanhWithDropout	grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_0	0.3456047733117182
MaxoutWithDropout	grid-deac795a-8df1-4f04-b0c3-7c7cb15dd9f9_model_2	0.656868528098143