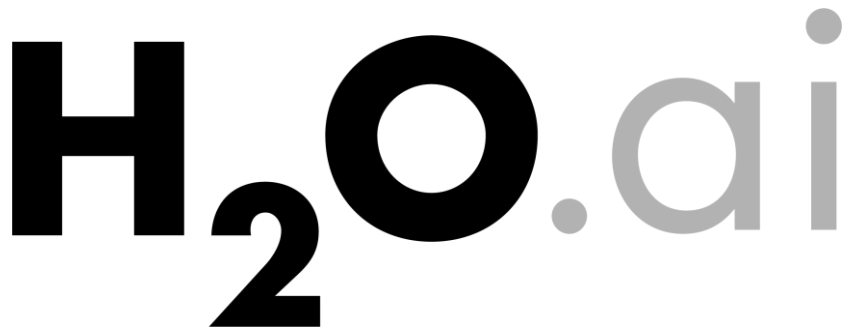
Machine Learning with H_2O



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

@matlabulous

R Addicts Paris Meetup

1st December, 2016

Agenda

- Introduction to H_2O
  - About H_2O.ai
- Our Open Source Products
  - Overview
  - H_2O Platform
  - Steam
  - Live Demo
    - H2O + R + Web + Steam

H_2O.ai

About Me

- Civil (Water) Engineer
  - 2010 – 2015
  - Consultant (UK)
    - Utilities
    - Asset Management
    - Constrained Optimization
  - Industrial PhD (UK)
    - Infrastructure Design Optimization
    - Machine Learning + Water Engineering
    - Discovered H2O in 2014
- Data Scientist
  - From 2015
  - Virgin Media (UK)
  - Domino Data Lab (Silicon Valley)
  - H_2O.ai (Silicon Valley)

H_2O.ai

I ♥ R

Jo-fai Chow

woobe

Civil Engineer turned Data Scientist

H2O.ai
United Kingdom
jofai.chow@gmail.com
http://www.jofaichow.co.uk/
Joined on Aug 10, 2012

Organizations

Overview | Repositories 41 | Stars 372 | Followers 118 | Following 30

Popular repositories

Customize your pinned repositories

blenditbayes
Code used in my blog "Blend it like a Bayesian!"
R ★ 73 81

deepr
An R package to streamline the training, fine-tuning and predicting processes for deep learning based on 'darch' and 'deepnet'.
R ★ 40 14

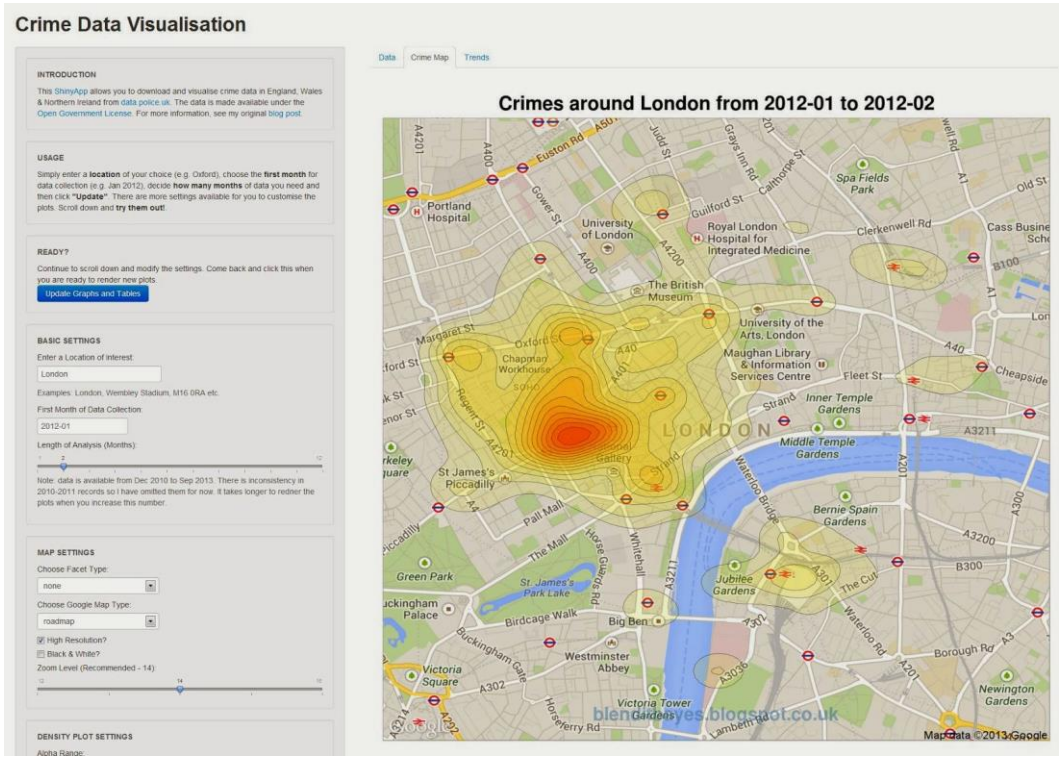
rPlotter
Wrapper functions that make plotting in R a lot easier for beginners.
R ★ 29 4

rCrimemap
This is the next generation of CrimeMap!
R ★ 22 9

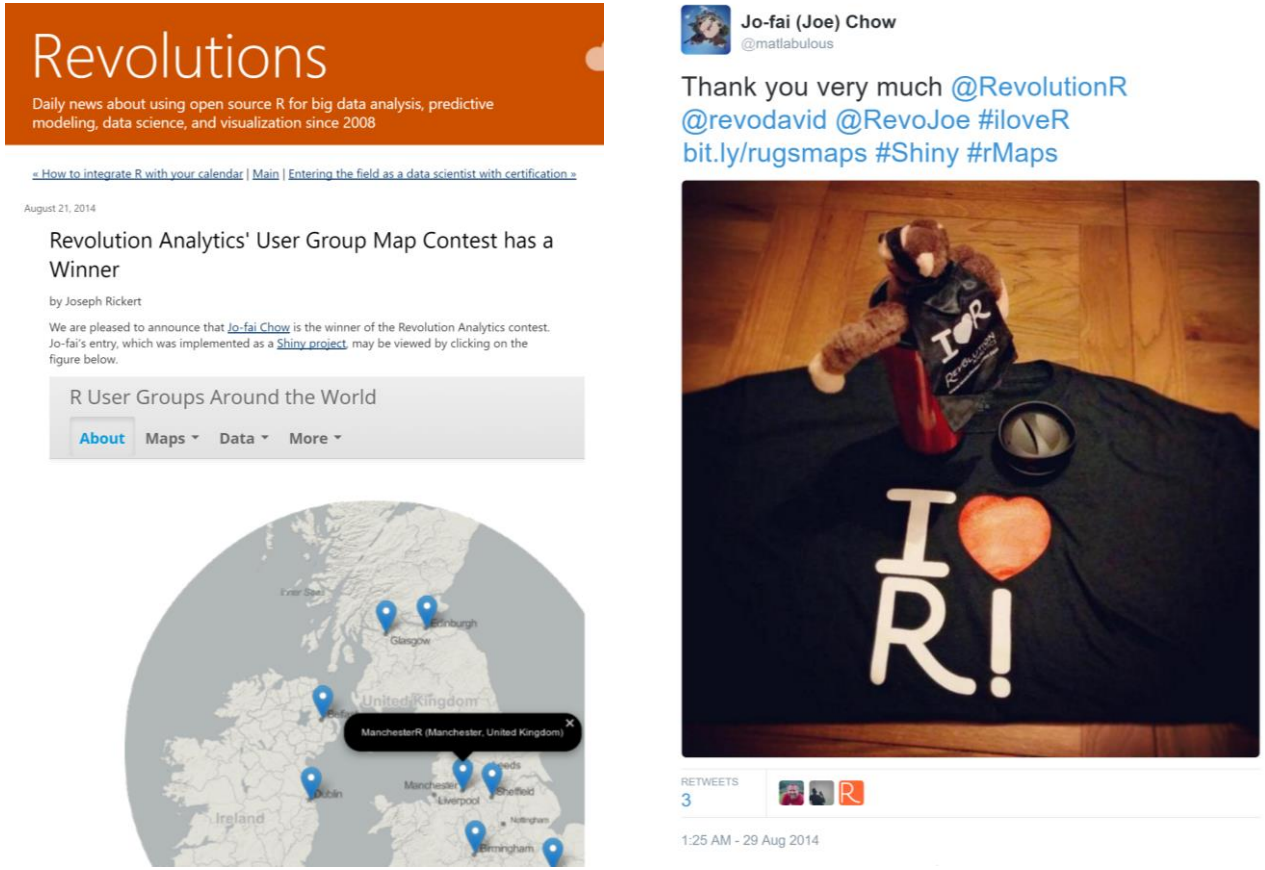
rugsmaps
This app is my submission to the visualization contest held by Revolution Analytics.
R ★ 19 18

rApps
Repository for my R (Shiny) web applications.
R ★ 16 35

I ♥ R

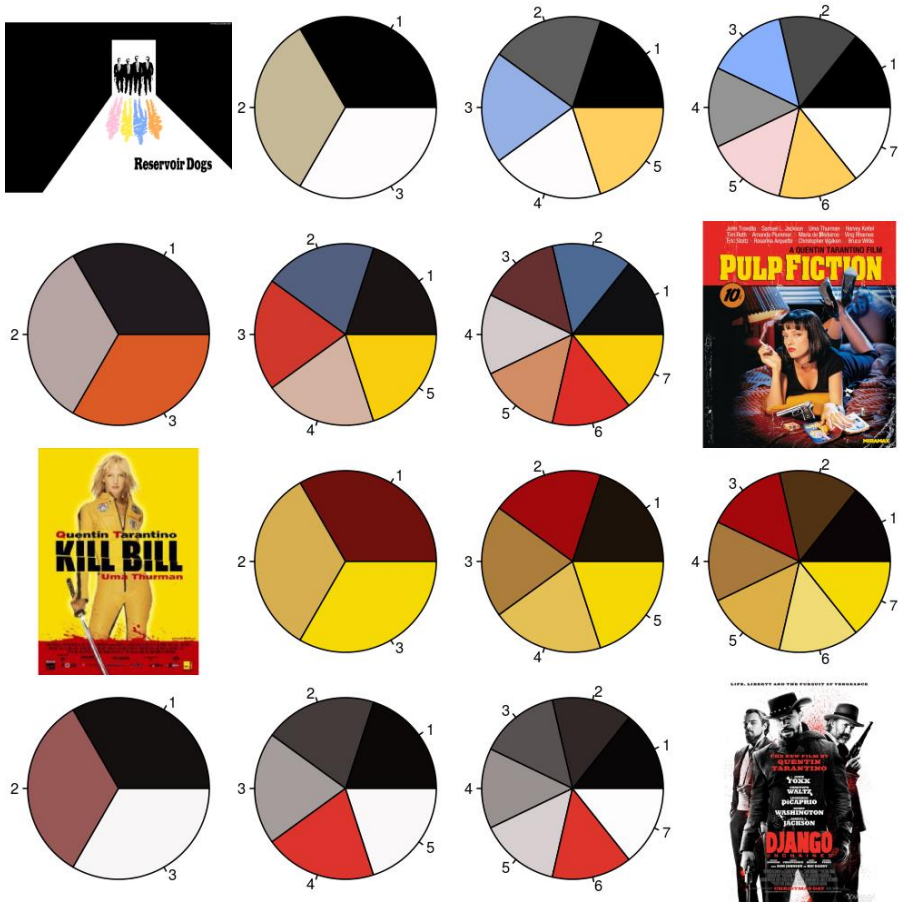


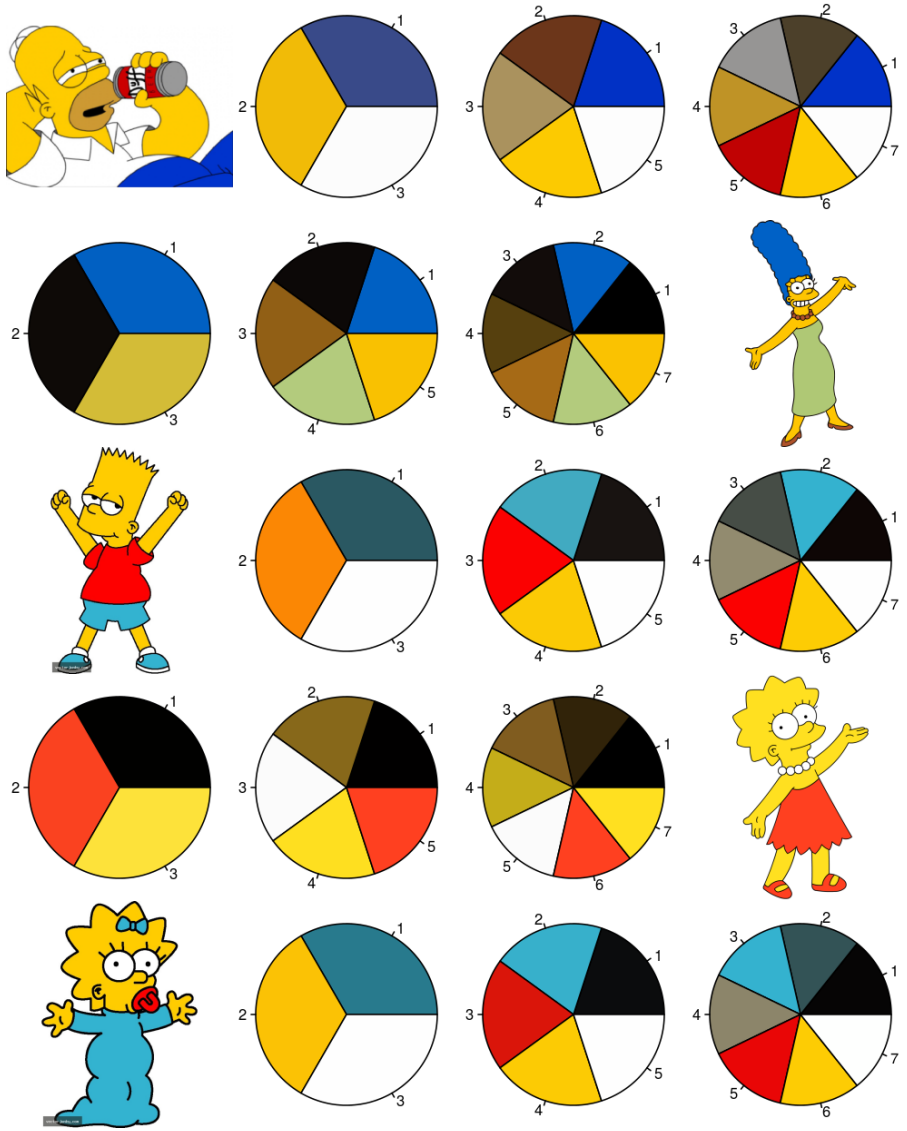
My First Data Viz & Shiny App Experience
CrimeMap (2013)



Revolution Analytics' Data Viz Contest
RUGSMAPS (2014)

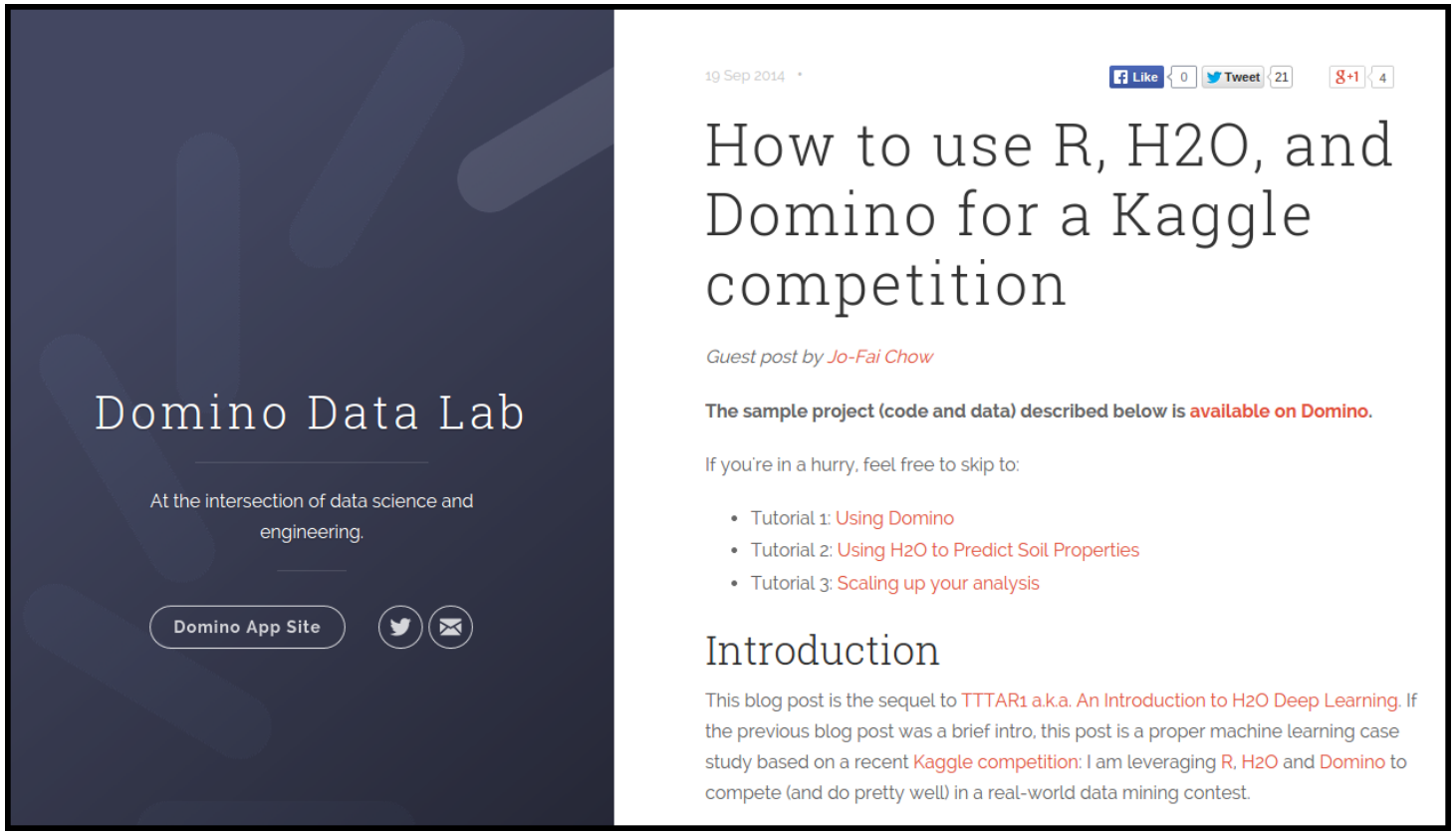
I ♥ R





Developing R Packages for Fun
rPlotter (2014)

I ♥ R



R + H2O + Domino for Kaggle

Guest Blog Post for Domino & H_2O (2014)

About H_2O.ai

What exactly is H_2O?

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	• H2O Open Source In-Memory AI Prediction Engine • Sparkling Water • Steam
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	70 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Headquarters	Mountain View, CA




H2O.ai

H_2O is an open source platform empowering business transformation

Bring AI To Business Empower Transformation

Financial Services, Insurance and Healthcare as Our Vertical Focus

Community as Our Foundation

Users In Various Verticals Adore H_2O

H2O.ai

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.ai

H_2O.ai Makes A Difference as an AI Platform

Open Source

- 100% open source

Big Data Ecosystem

hadoop

Spark

Flexible Interface

R

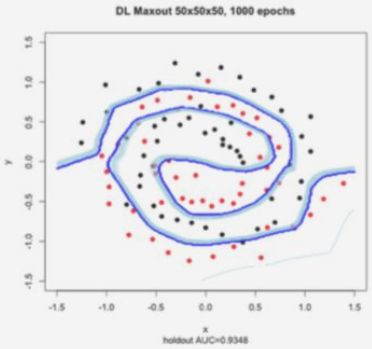
Scala

Java

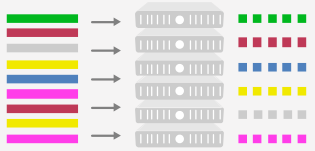
python

H_2O Flow

Smart and Fast Algorithms



Scalability and Performance



- Distributed In-Memory Computing Platform
- Distributed Algorithms
- Fine-Grain MapReduce

Rapid Model Deployment

- Highly portable models deployed in Java (POJO) and Model Object Optimized (MOJO)
- Automated and streamlined scoring service deployment with Rest API

REST API

GPU Enablement

GPU

Cloud Integration

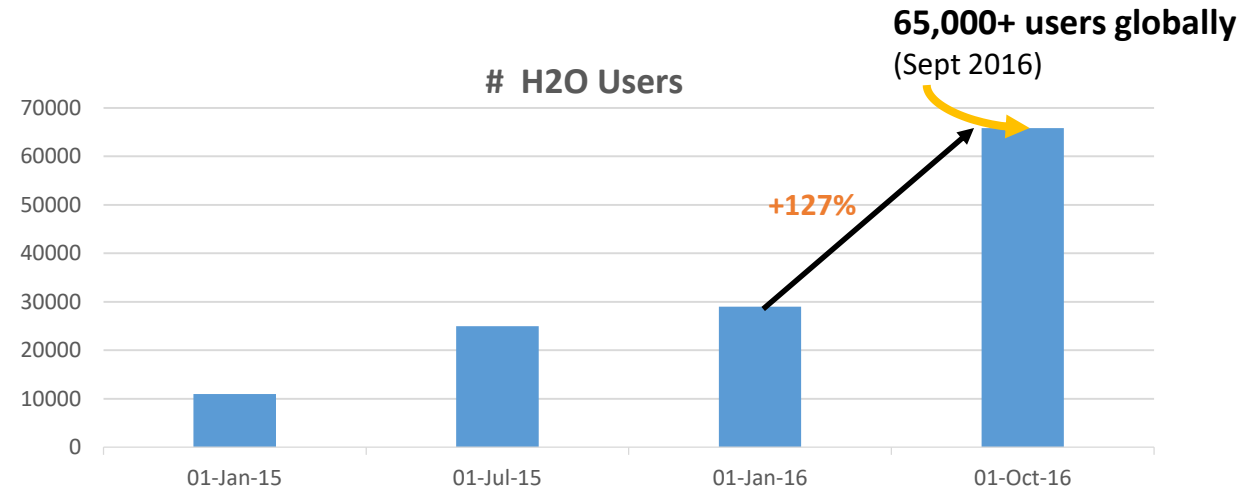
amazon web services™

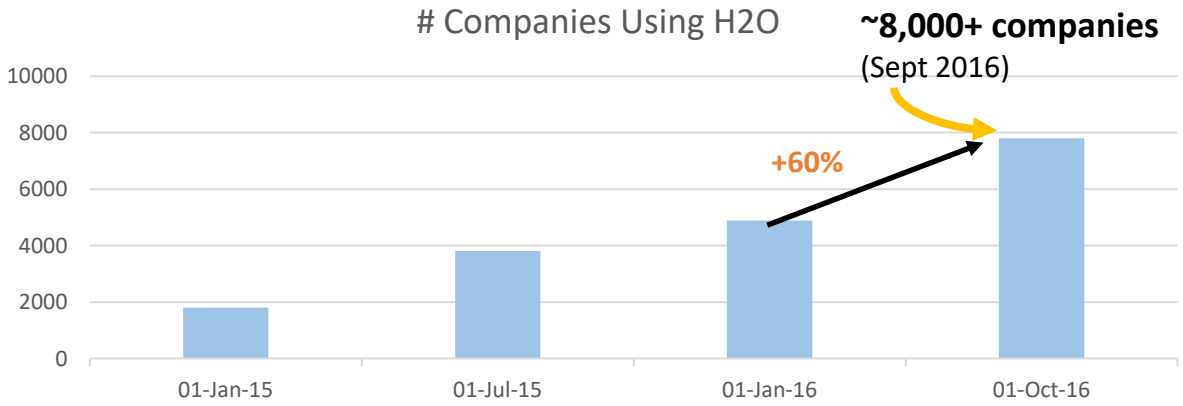
Microsoft Azure

Google

H_2O Community Growth

Tremendous Momentum Globally





Large User Circle

- 65,000+ users from ~8,000 companies in 140 countries. Top 5 from:

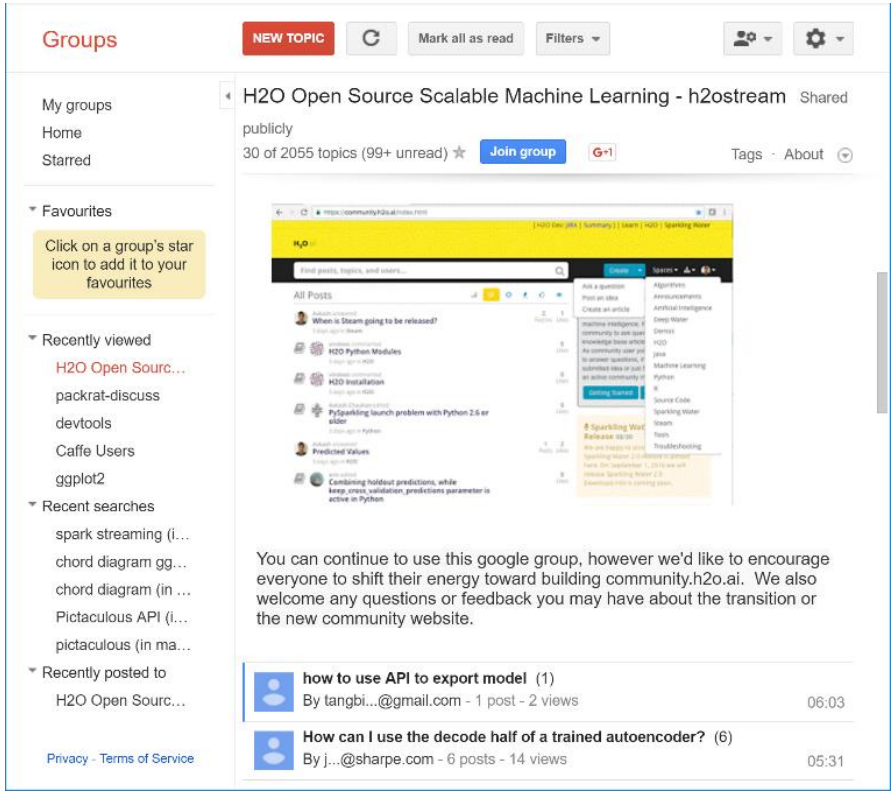
1. United States
2. India
3. Japan
4. Germany
5. United Kingdom

* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H2O.ai

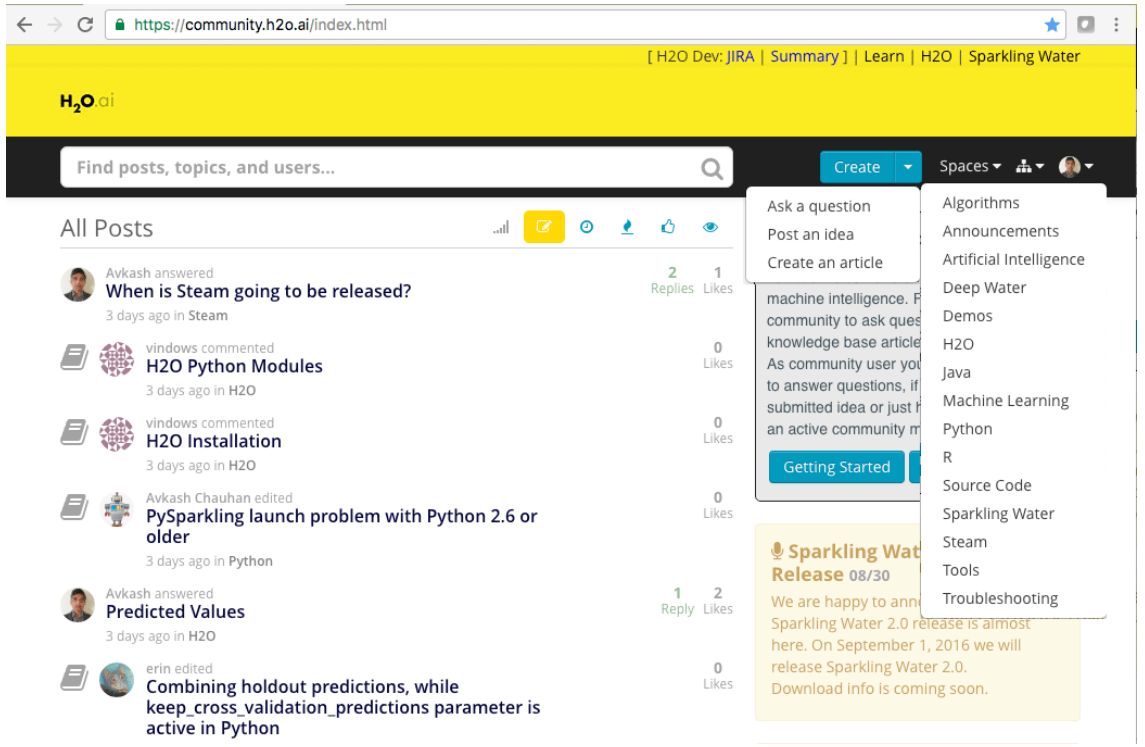
H_2O Community Support

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

H2O.ai

H_2O for Kaggle Competitions

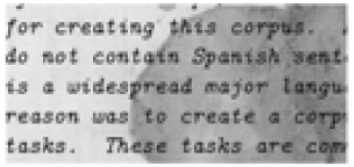
CIFAR-10 Competition Winners: Interviews with Dr. Ben Graham, Phil Culliton, & Zygmunt Zając

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

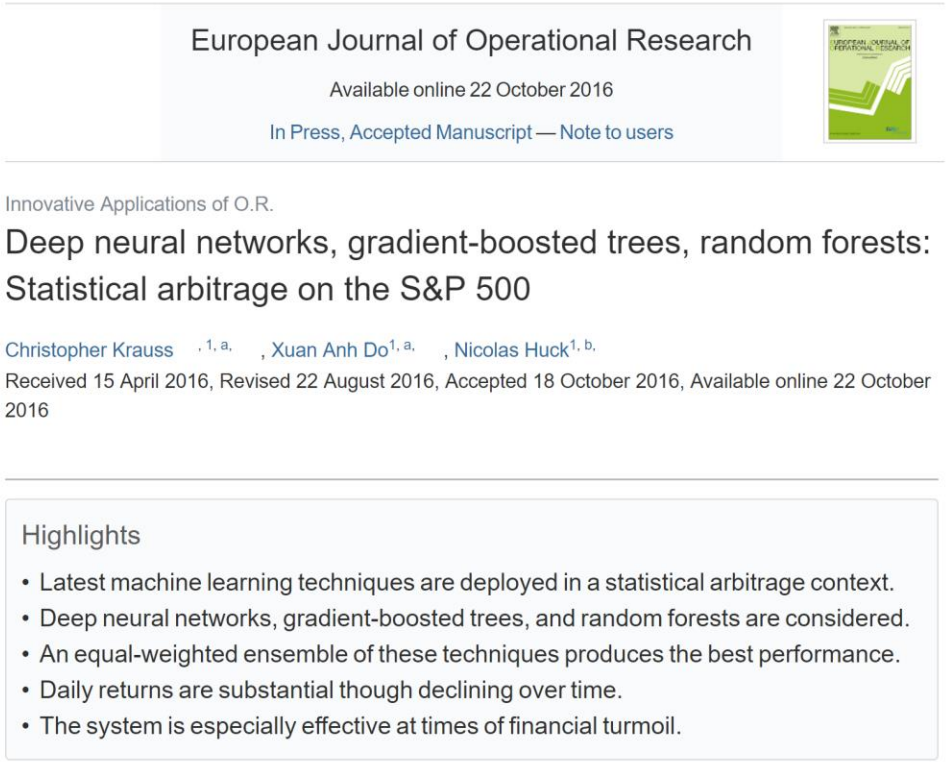
Mon 1 Jun 2015 – Mon 5 Oct 2015 (3 months ago)

READ MORE

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

H_2O.ai

H_2O 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

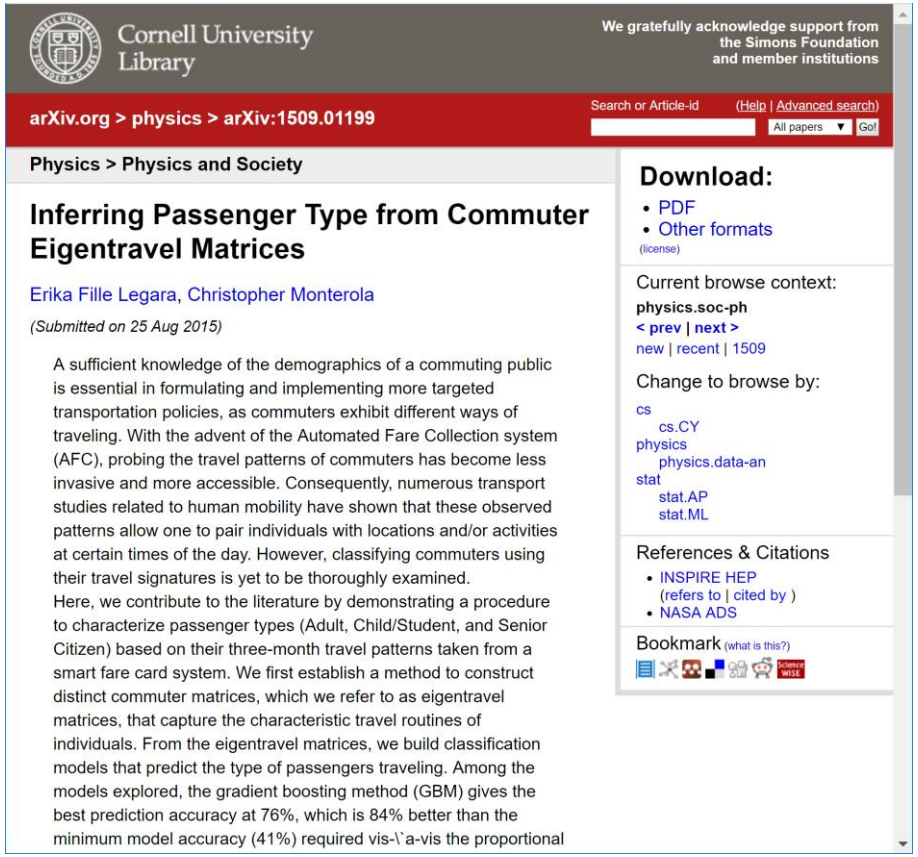
Christopher Krauss [1, a], Xuan Anh Do [1, a], Nicolas Huck [1, b]

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

H_2O

democratizes

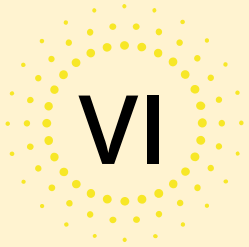
artificial intelligence & big data science

Our Open Source Products

100% Open Source. Big Data Science for Everyone!

H_2O.ai Offers AI Open Source Platform

Product Suite to Operationalize Data Science with Visual Intelligence



Visual Intelligence and UX Framework For Data Interpretation and Story Telling on top of Beautiful Data Products

100% Open Source

H_2O.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

Spark + H_2O

SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

H_2O.ai Offers AI Open Source Platform

Product Suite to Opera… Visual Intelligence

This Talk + Live Demos

… Framework For Data Interpretation and
…autiful Data Products

100% Open Source

H_2O.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

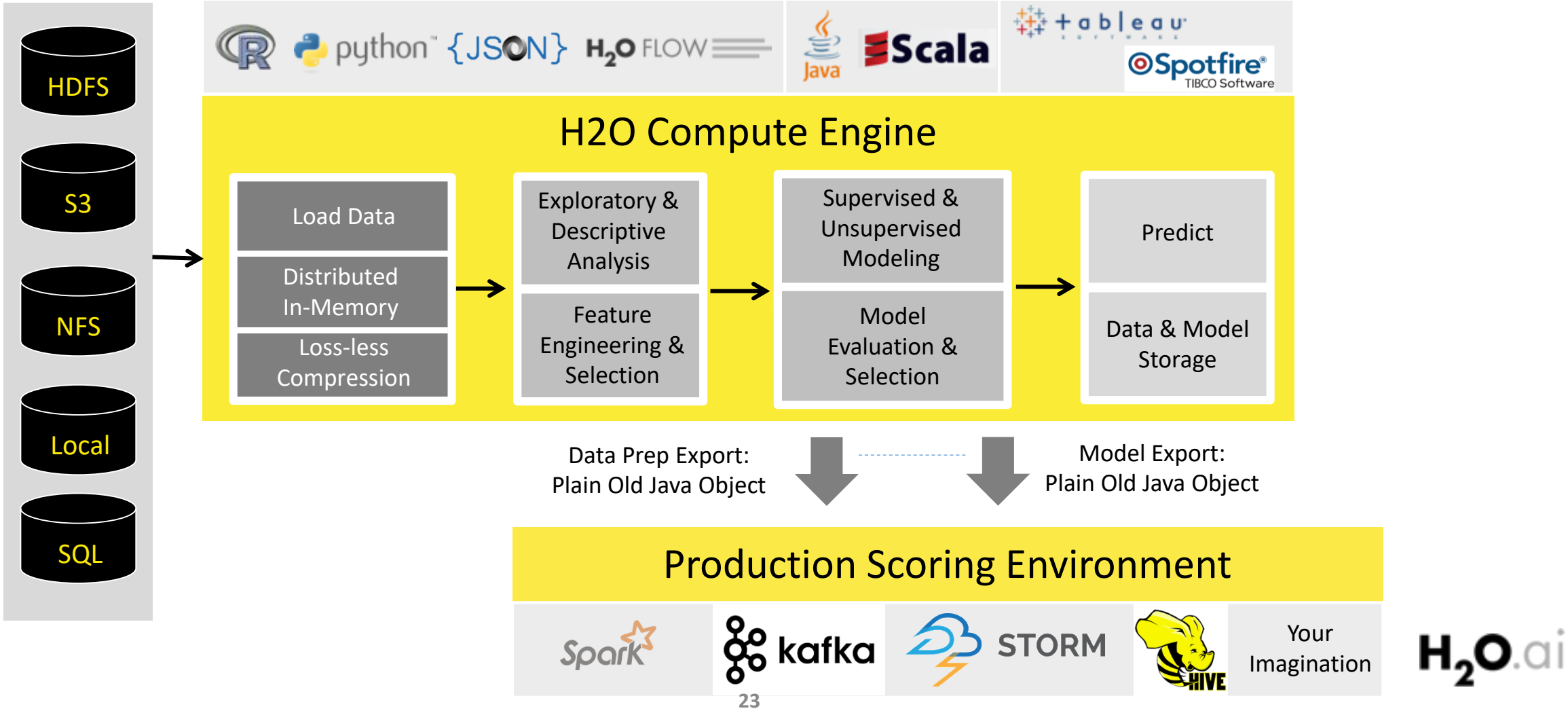
Spark + H_2O SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

High Level Architecture



Algorithms Overview

Supervised Learning

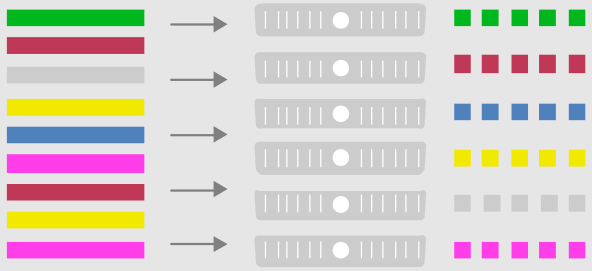
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
Deep Neural Networks	• **Deep learning**: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

Unsupervised Learning

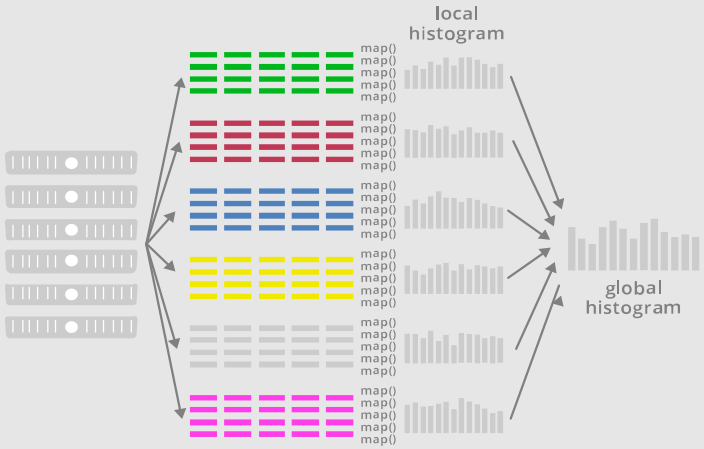
Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components • **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data
Anomaly Detection	• **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

H_2O 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%

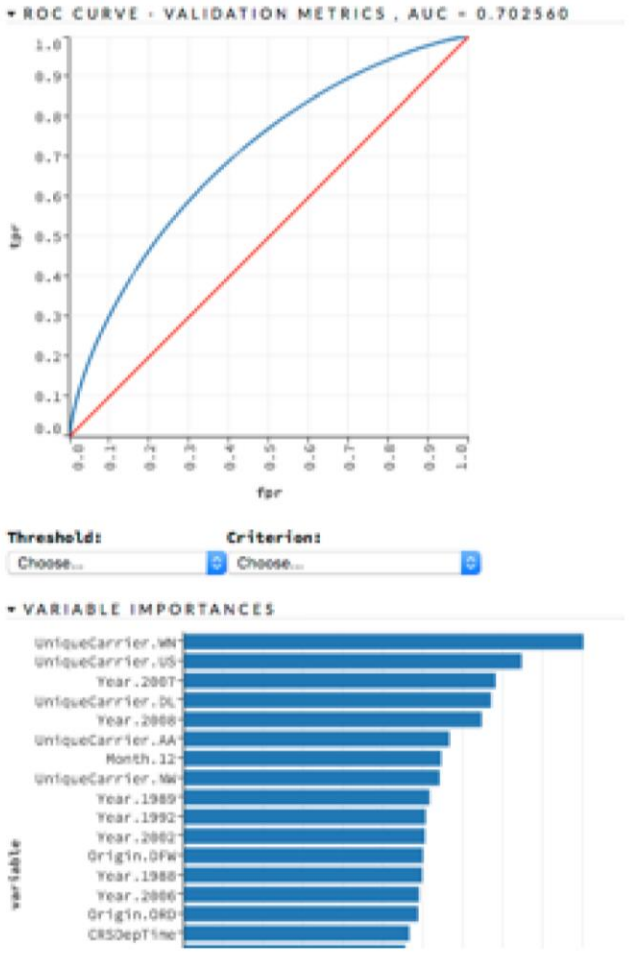
Iterations: 12. Epochs: 0.628821. Speed: 2,243,735 samples/sec. Estimated time left: 21.849 sec

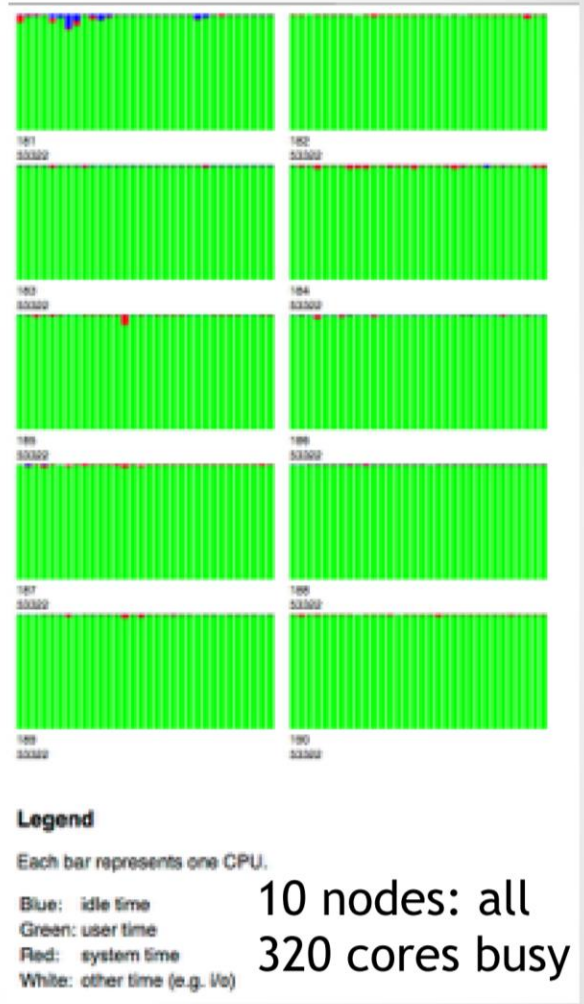
Actions: View | Cancel Job

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	0.0493	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	-1.3988	1.5259
4	20	Rectifier	0	0	0	0.0517	0.0446	0	-0.1575	0.3068	-0.6846	0.6846
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

H_2O.ai **Deep Learning Model**

Steam

DATA

BUSINESS INSIGHTS

Steam

Automated Platform to Build and Scale Smart Data Products

Smart Data Product

Smart Data Product

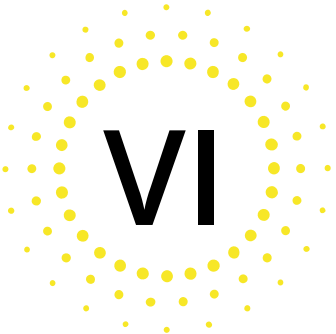
Smart Data Product

Smart Data Product

AI – Machine Learning

Automation

Scalability



Visual Intelligence

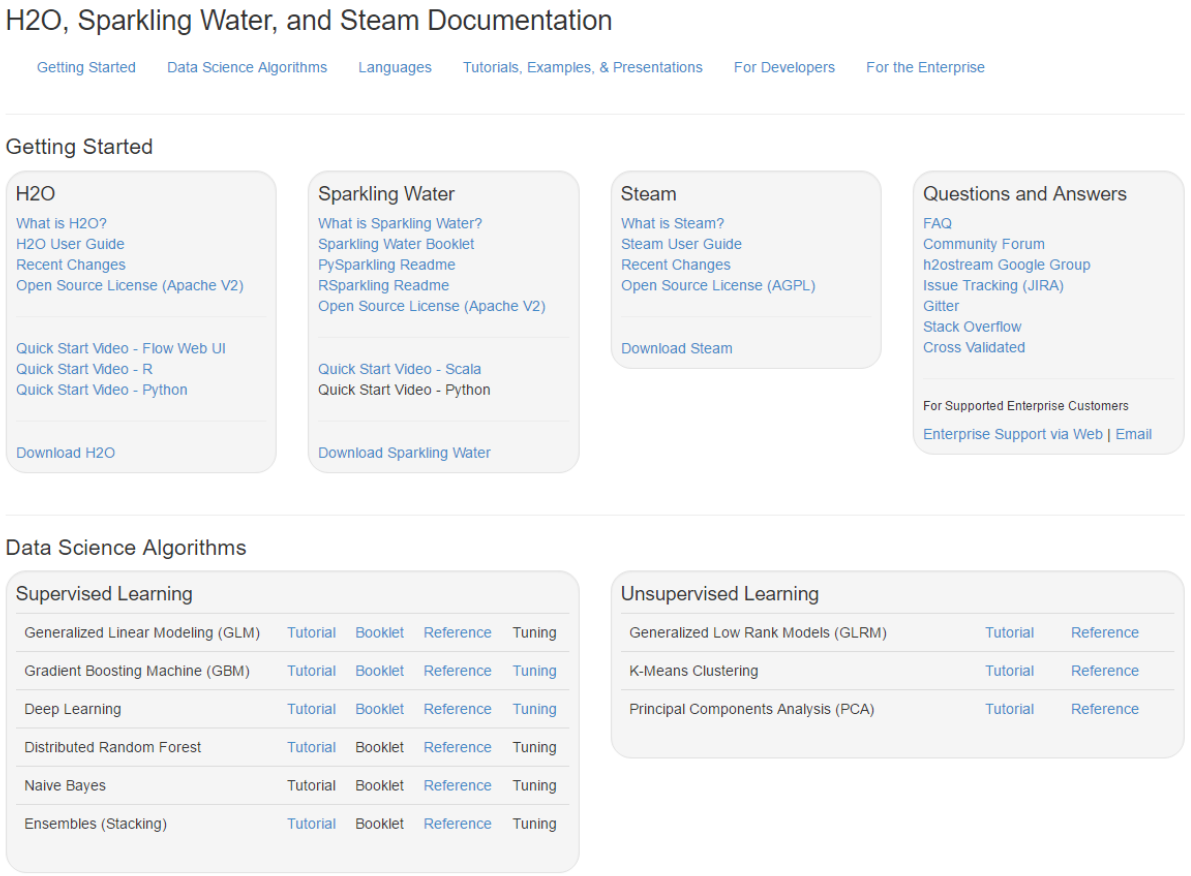
H2O.ai

H2O in Action:
H_2O + R + Web + Steam

Quick Demo (5 mins)

Key Learning Resources

- Help Documentations
  - docs.h2o.ai
- Meetups
  - bit.ly/h2o_meetup
- YouTube Channel
  - bit.ly/h2o_youtube



H2O, Sparkling Water, and Steam Documentation

Getting Started | Data Science Algorithms | Languages | Tutorials, Examples, & Presentations | For Developers | For the Enterprise

Getting Started

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

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

Download H2O

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

Quick Start Video - Scala
Quick Start Video - Python

Download Sparkling Water

Steam
What is Steam?
Steam User Guide
Recent Changes
Open Source License (AGPL)

Download Steam

Questions and Answers
FAQ
Community Forum
h2ostream Google Group
Issue Tracking (JIRA)
Gitter
Stack Overflow
Cross Validated

For Supported Enterprise Customers
Enterprise Support via Web | Email

Data Science Algorithms

Supervised Learning

Generalized Linear Modeling (GLM)	Tutorial	Booklet	Reference	Tuning
Gradient Boosting Machine (GBM)	Tutorial	Booklet	Reference	Tuning
Deep Learning	Tutorial	Booklet	Reference	Tuning
Distributed Random Forest	Tutorial	Booklet	Reference	Tuning
Naive Bayes	Tutorial	Booklet	Reference	Tuning
Ensembles (Stacking)	Tutorial	Booklet	Reference	Tuning

Unsupervised Learning

Generalized Low Rank Models (GLRM)	Tutorial	Reference
K-Means Clustering	Tutorial	Reference
Principal Components Analysis (PCA)	Tutorial	Reference

TensorFlow

PARIS

rs AI Open Source Platform

Operationalize Data Science with Visual Intelligence

Meetup Talk
Yesterday
bit.ly/h2o_meetups

3rd talk this evening
by Jakub

Visual Intelligence and U
Story Telling on top of Be
nd

100% Open Source

H2O.ai

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

Deep Water

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

Spark + H2O

SPARKLING WATER

H2O Integration with Spark. Best Machine Learning on Spark.

Steam

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

H_2O's Mission

Making Machine Learning Accessible to Everyone

Photo credit: Virgin Media

Merci beaucoup!

- Organizers & Sponsors
  - Diane, Vincent, Barthelemy, François, Julie and Timeri
  - NUMA
- Code, Slides & Documents
  - bit.ly/h2o_meetups
  - docs.h2o.ai
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe

H2O.ai