Scalable Machine Learning in R with H2O

useR! Stanford
June 2016

Navdeep Gill M.S.
H_2O.ai

Introduction

- Hacker/Data Scientist @ H2O.ai
- Experience:
  - Behavioral/Cognitive Neuroscience: 3 years
  - Predictive Analytics/Data Science: 2 years
  - Software Development: 1 year
  - R user: 4 years
- Education:
  - M.S. Computational Statistics @ CSU East Bay
  - B.S./B.A. Statistics, Mathematics, and Psychology @ CSU East Bay

- Brief overview of H2O
- H2O Platform
- H2O in R
  - Demo
- H2O R API in AWS EC2
  - Demo

H2O.ai

H2O Company

- Team: 50. Founded in 2012, Mountain View, CA
- Stanford Math & Systems Engineers

H2O Software

- Open Source Software
- Ease of Use via Web Interface
- R, Python, Scala, Spark & Hadoop Interfaces
- Distributed Algorithms Scale to Big Data


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*
- 108,404 citations (via Google Scholar)

Dr. Rob Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- COPPS Presidents' Award recipient
- 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. Stephen Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, Convex Optimization
- Co-author, Linear Matrix Inequalities in System and Control Theory
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*

H2O Overview

Speed Matters!

- Time is valuable
- In-memory is faster
- Distributed is faster
- High speed AND accuracy

No Sampling

- Scale to big data
- Access data links
- Use all data without sampling

Interactive UI

- Web-based modeling with H2O Flow
- Model comparison

Cutting-Edge Algorithms

- Suite of cutting-edge machine learning algorithms
- Deep Learning & Ensembles
- NanoFast Scoring Engine

Current Algorithm Overview

Statistical Analysis

- Linear Models (GLM)
- Naïve Bayes

Ensembles

- Random Forest
- Distributed Trees
- Gradient Boosting Machine
- R Package - Super Learner Ensembles

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 Components

H2O Cluster

- Multi-node cluster with shared memory model.
- All computations in memory.
- Each node sees only some rows of the data.
- No limit on cluster size.

Distributed Key Value Store

- Objects in the H2O cluster such as data frames, models and results are all referenced by key.
- Any node in the cluster can access any object in the cluster by key.

H2O Frame

- Distributed data frames (collection of vectors).
- Columns are distributed (across nodes) arrays.
- Each node must be able to see the entire dataset (achieved using HDFS, S3, or multiple copies of the data if it is a CSV file).

Data in H2O

Highly Compressed

- We read data fully parallelized from: HDFS, NFS, Amazon S3, URLs, URIs, CSV, SVMLight.
- Data is highly compressed (about 2-4 times smaller than gzip).

Speed

- Memory bound, not CPU bound.
- If data accessed linearly, as fast as C or Fortran.
- Speed = data volume / memory bandwidth
- ~50GB / sec (varies by hardware).

Data Shape

- Table width: <1k fast, <10k works, <100k slow
- Table length: Limited only by memory
- We have tested 10's of billions of rows (TBs)

Distributed H2O Frame

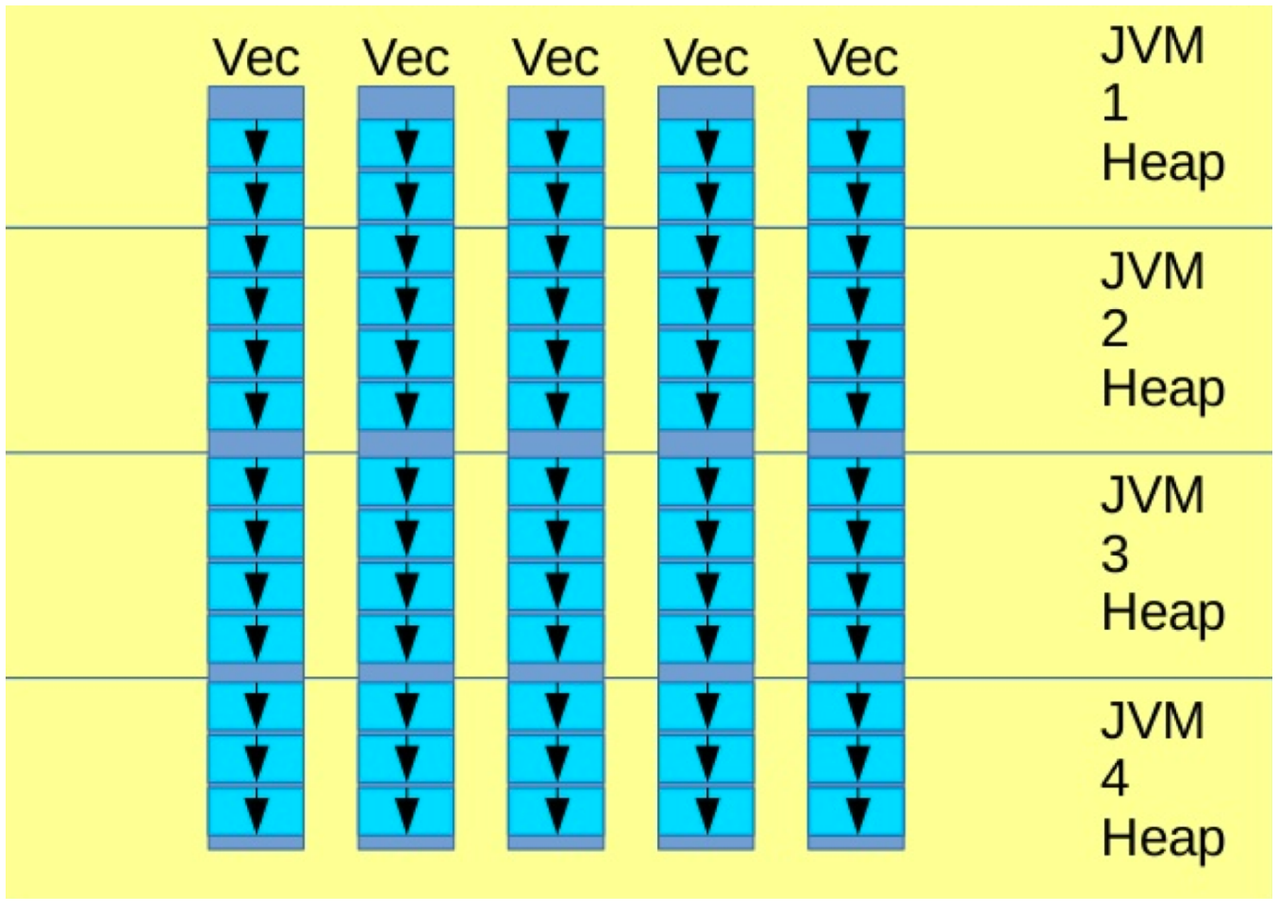


Diagram of distributed arrays. An “H2O Frame” is a collection of distributed arrays.

H2O in R

h2o R package

Requirements

- The only requirement to run the “h2o” R package is R >=3.1.0 and Java 7 or later.
- Linux, OS X and Windows.

Installation

- The easiest way to install the “h2o” R package is to install directly from CRAN.
- Latest version: `http://h2o.ai/download`

Design

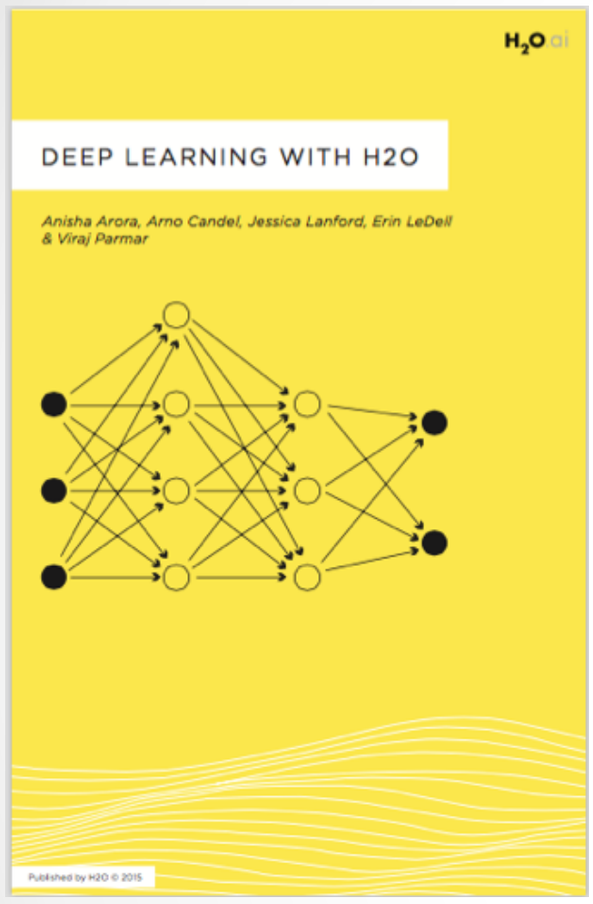
- No computation is ever performed in R.
- All computations are performed (in highly optimized Java code) in the H2O cluster and initiated by REST calls from R.

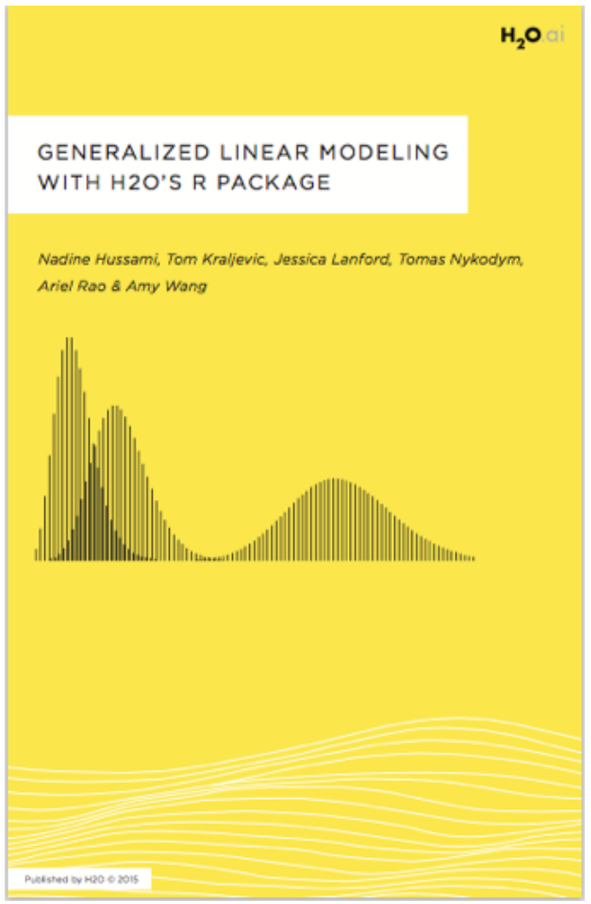
DEMO!

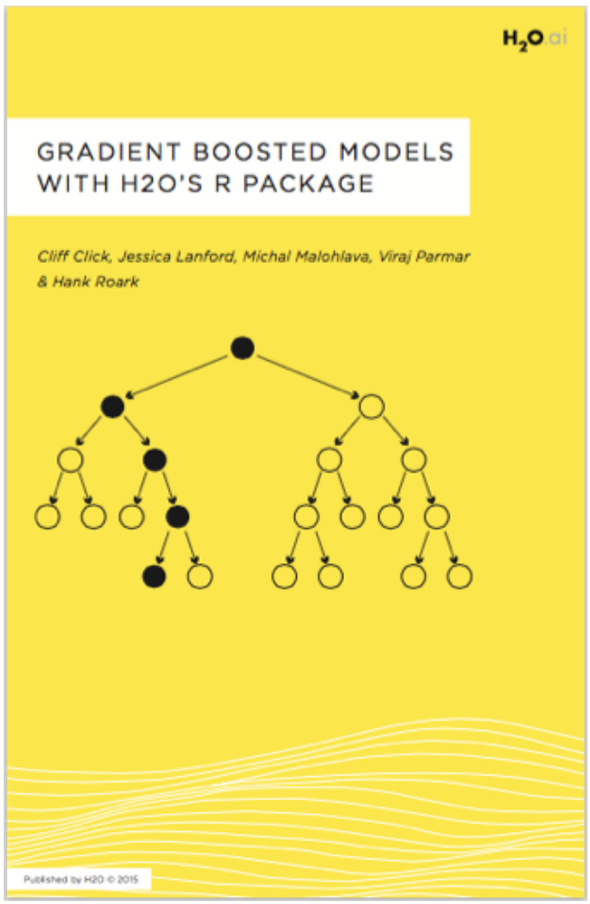
Where to learn more?

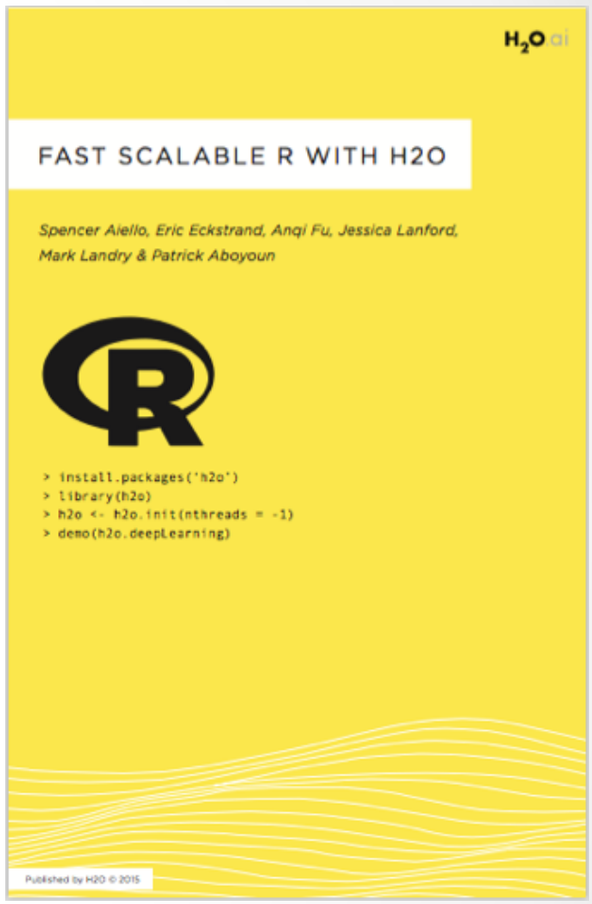
- H2O Online Training (free): http://learn.h2o.ai
- H2O Slidedecks: http://www.slideshare.net/0xdata
- H2O Video Presentations: https://www.youtube.com/user/0xdata
- H2O Community Events & Meetups: http://h2o.ai/events
- Machine Learning & Data Science courses: http://coursebuffet.com

H2O Booklets







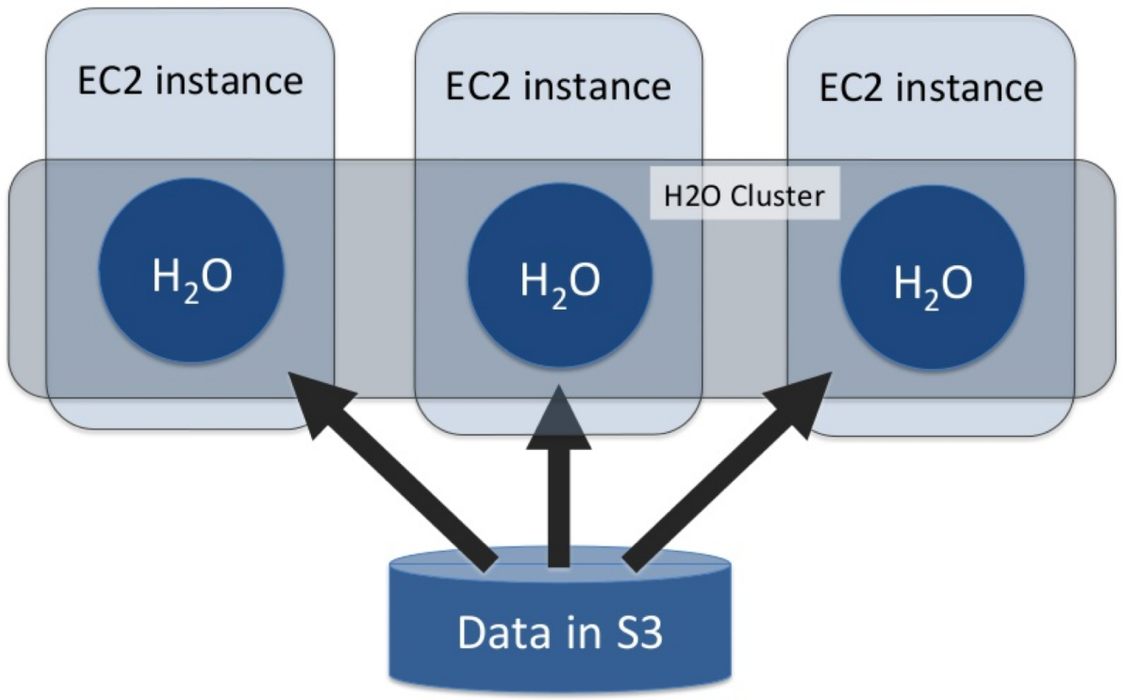


https://github.com/h2oai/h2o-3/tree/master/h2o-docs/src/booklets/v2_2015/PDFs/online

H2O R API in AWS EC2

Amazon EC2

H2O on Amazon EC2



H2O can easily be deployed on an Amazon EC2 cluster. The GitHub repository contains example scripts that help to automate the cluster deployment.

DEMO!

Thank you!

@Navdeep_Gill_ on Twitter
navdeep-G on GitHub
navdeep@h2o.ai

Slides available at: https://github.com/navdeep-G/useR-scalable-ml-h2o-tutorial/tree/master/presentation

Link to Demos: https://github.com/navdeep-G/useR-scalable-ml-h2o-tutorial/tree/master/scripts