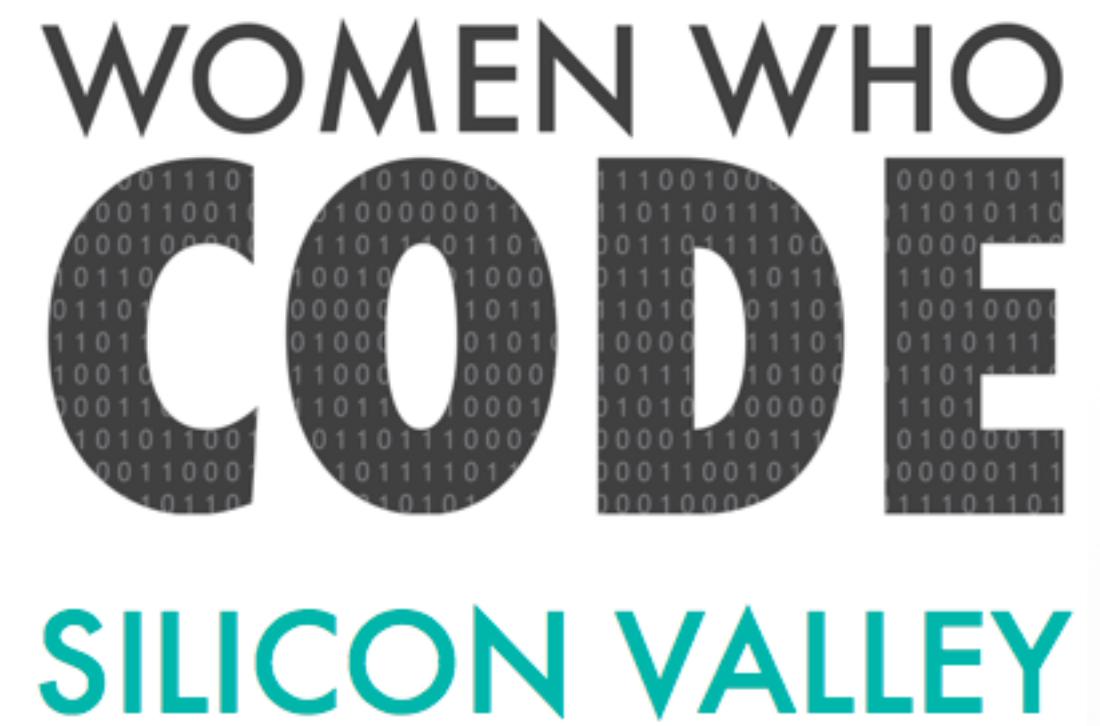


Scalable Machine Learning in R and Python with H2O



May 2016

Erin LeDell Ph.D.
Machine Learning Scientist
H2O.ai

Introduction

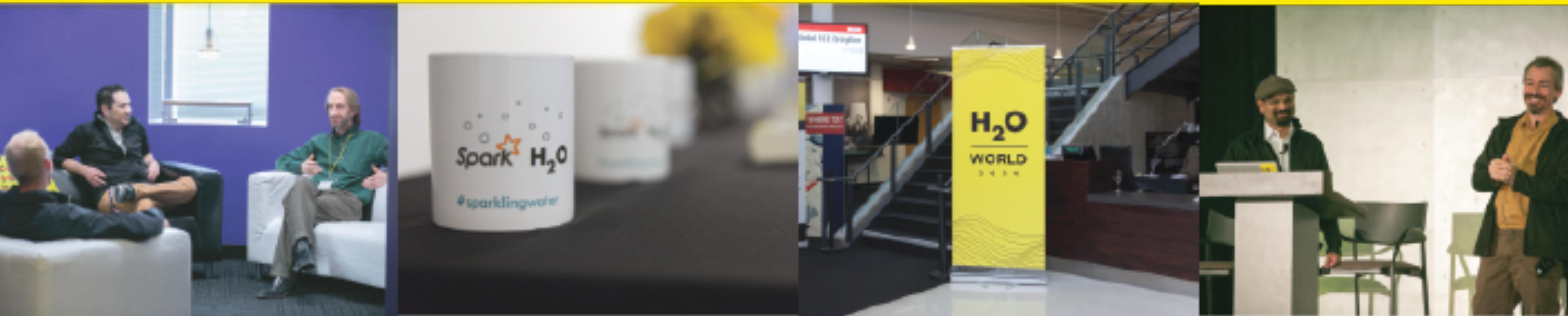
- Statistician & Machine Learning Scientist at H2O.ai in Mountain View, California, USA
- Ph.D. in Biostatistics with Designated Emphasis in Computational Science and Engineering from UC Berkeley (focus on Machine Learning)
- Worked as a data scientist at several startups



Agenda



- Who/What is H2O?
- Intro to Machine Learning
- H2O Machine Learning Platform
- H2O in R & Python
- H2O Code Tutorial



H2O.ai, the Company

- Team: 55; Founded in 2012
- Mountain View, CA
- Stanford & Purdue Math & Systems Engineers

H2O, the Platform

- Open Source Software (Apache 2.0 Licensed)
- R, Python, Scala, Java and Web Interfaces
- Distributed Algorithms that 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*



Dr. Robert Tibshirani

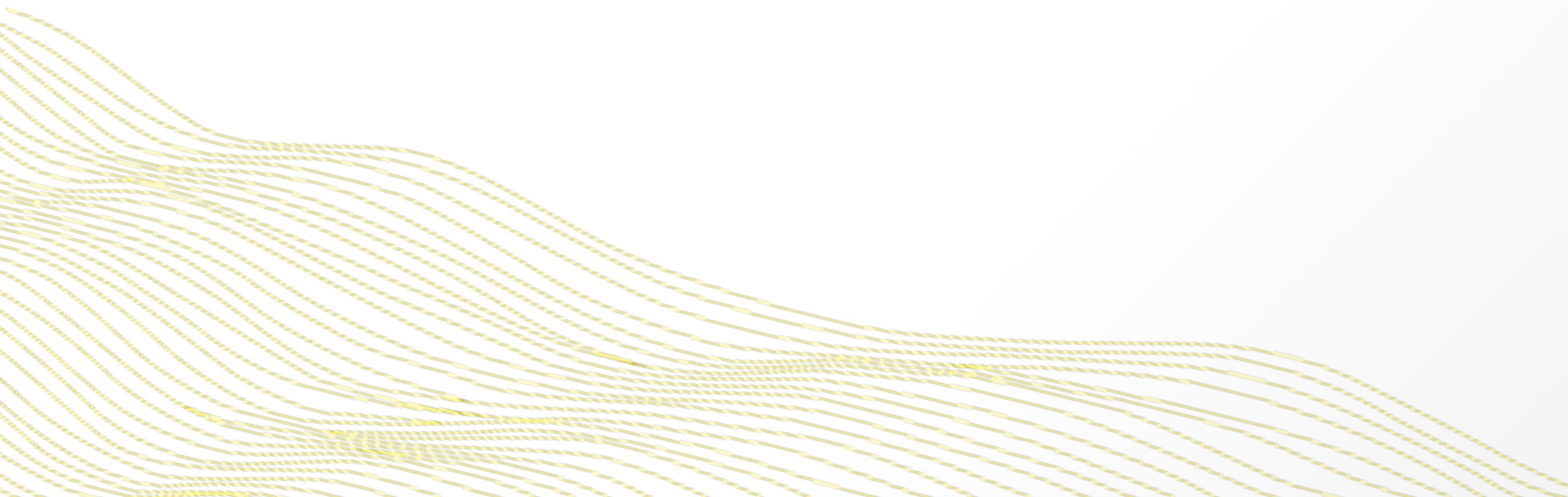
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*



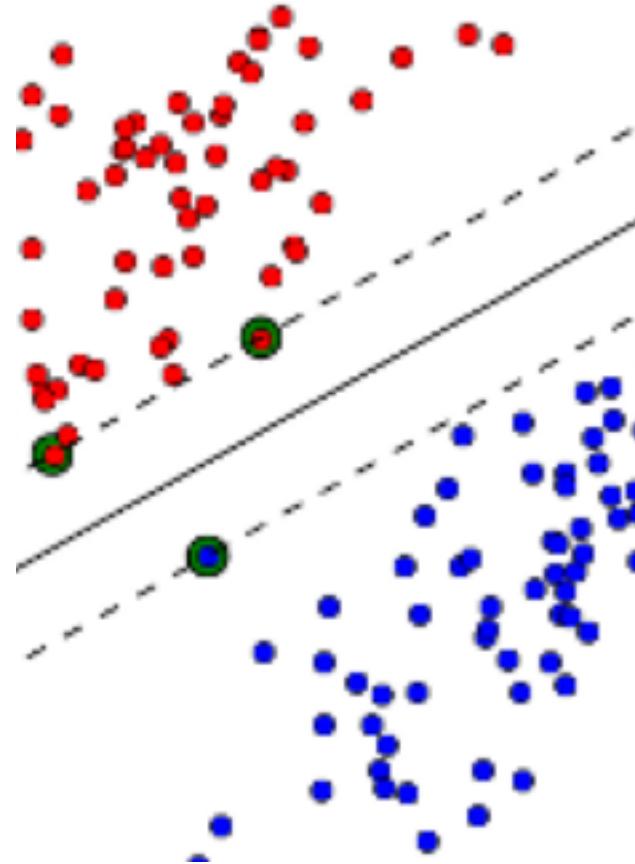
Dr. Steven Boyd

- 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*

Machine Learning



What is Machine Learning?



"Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel, 1959

Unlike rules-based systems which require a human expert to hard-code domain knowledge directly into the system, a machine learning algorithm learns how to make decisions from the data alone.

Machine Learning Tasks

Regression

- Predict a real-valued response (e.g. viral load, price)
 - Gaussian, Gamma, Poisson, etc. distributed response
 - Evaluate with MSE or R²
-

Classification

- Multi-class or binary classification
 - Ranking (e.g. Google Search results order)
 - Evaluate with Classification Error or AUC
-

Clustering

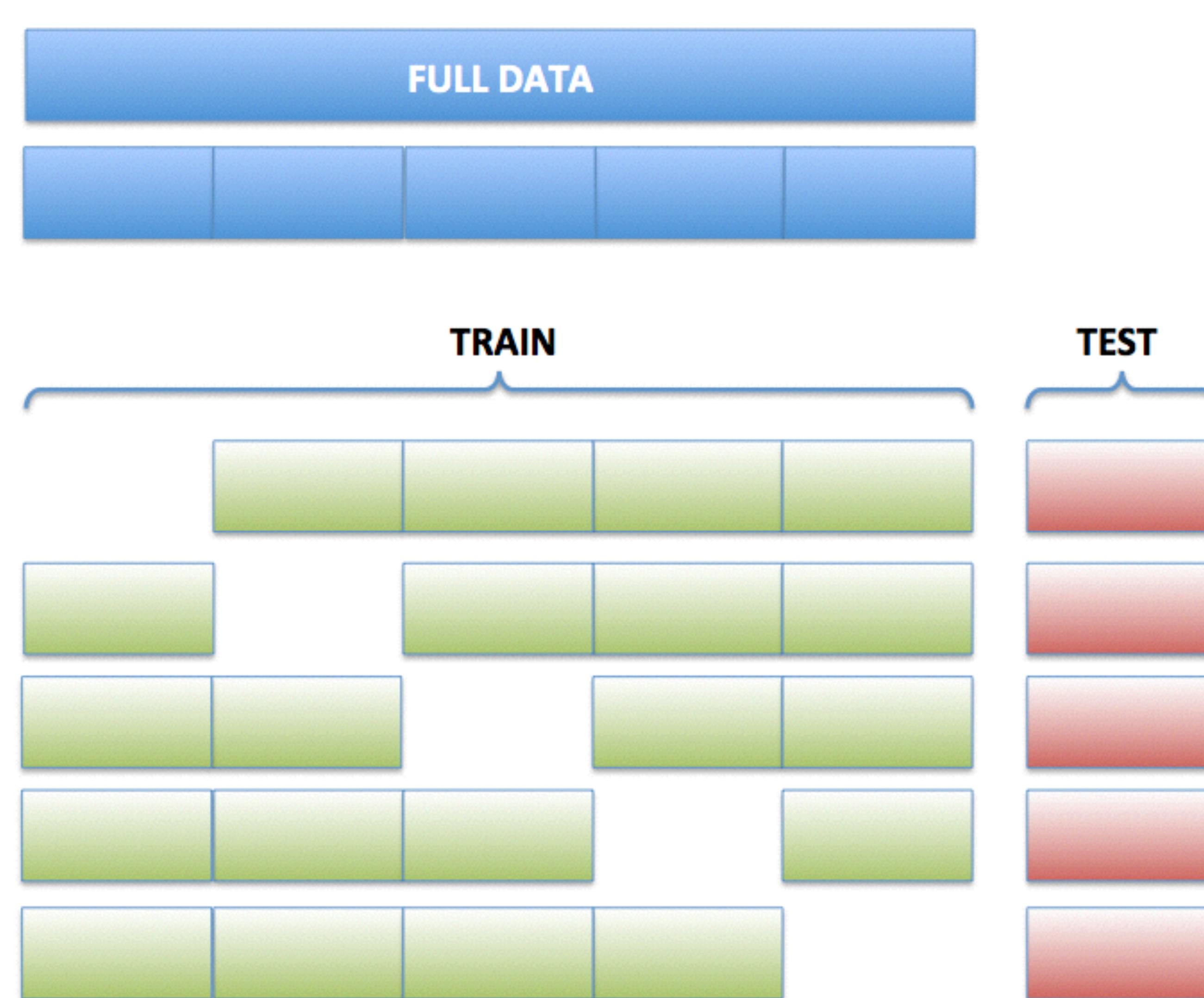
- Unsupervised learning (no training labels)
- Partition the data; identify clusters or sub-populations
- Evaluate with AIC, BIC or Total Sum of Squares

Train, Validation and Test Set

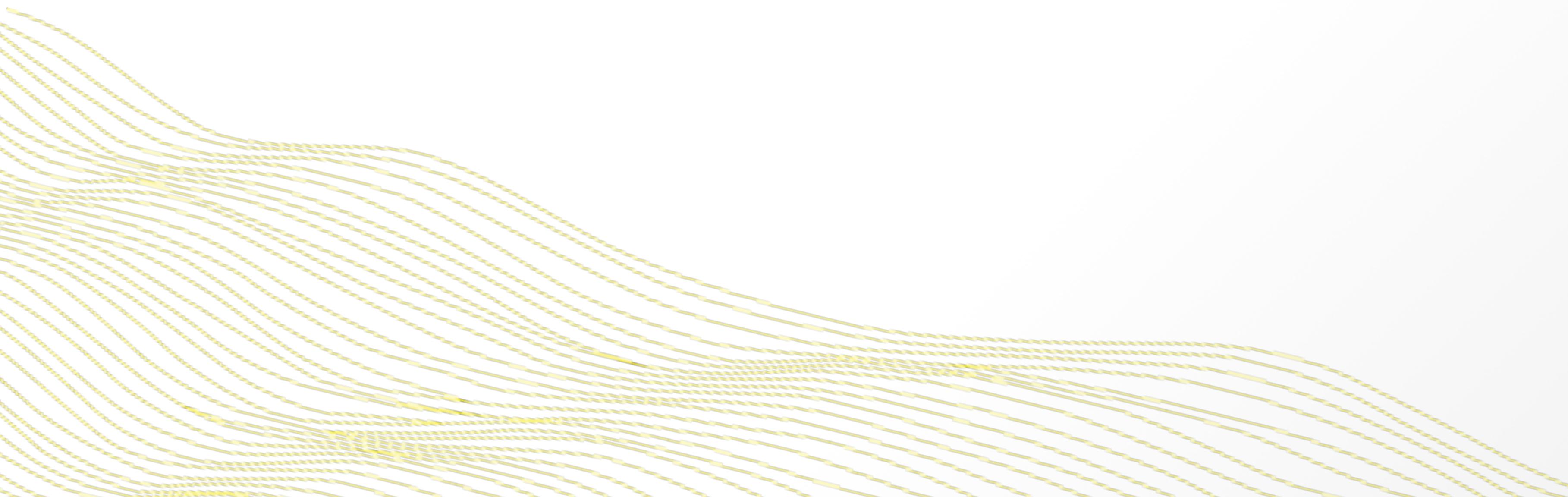


- If you plan on doing any model tuning, you should split your dataset into three parts: Train, Validation and Test
- There is no general rule for how you should partition the data and it will depend on how strong the signal in your data is, but an example could be:
50% Train, 25% Validation and 25% Test
- The validation set is used strictly for model tuning and the test set is used to make a final estimate of the generalization error.

k-fold Cross-validation



H2O Platform



H2O Platform Overview

- Distributed implementations of cutting edge ML algorithms.
- Core algorithms written in high performance Java.
- APIs available in R, Python, Scala, REST/JSON.
- Interactive Web GUI.



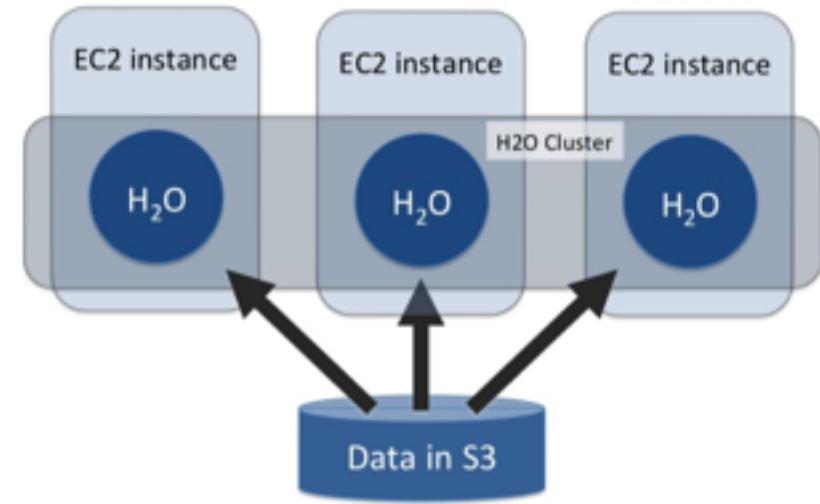
H2O Platform Overview

- Write code in high-level language like R (or use the web GUI) and output production-ready models in Java.
- To scale, just add nodes to your H2O cluster.
- Works with Hadoop, Spark and your laptop.



H2O Distributed Computing

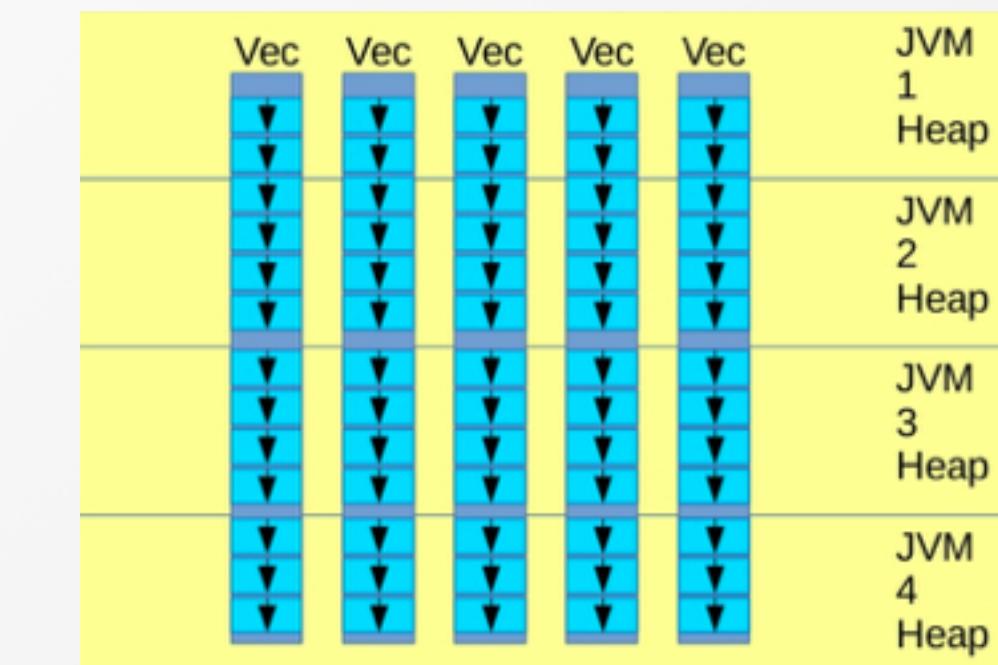
H2O Cluster



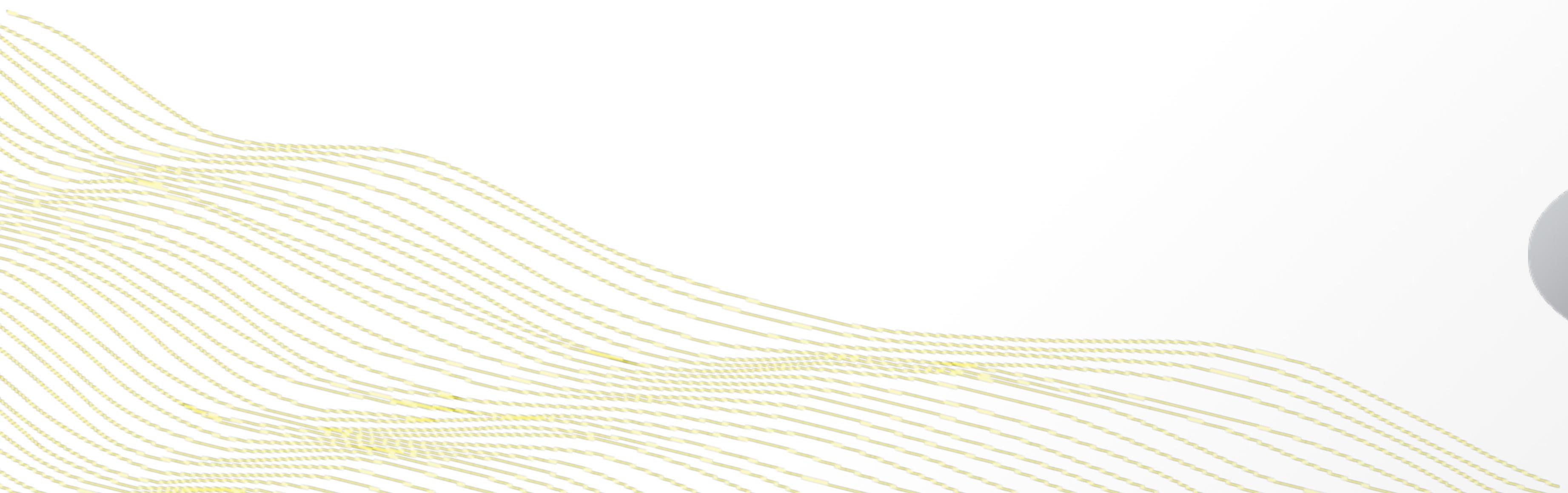
- Multi-node cluster with shared memory model.
- All computations in memory.
- Each node sees only some rows of the data.
- No limit on cluster size.

H2O Frame

- Distributed data frames (collection of vectors).
- Columns are distributed (across nodes) arrays.
- Works just like R's `data.frame` or Python Pandas `DataFrame`



H2O in R and Python



h2o R Package



Installation

- Java 7 or later; R 3.1 and above; Linux, Mac, Windows
- The easiest way to install the h2o R package is CRAN.
- Latest version: <http://www.h2o.ai/download/h2o/r>

Design

All computations are performed in highly optimized Java code in the H2O cluster, initiated by REST calls from R.

h2o Python Module



Installation

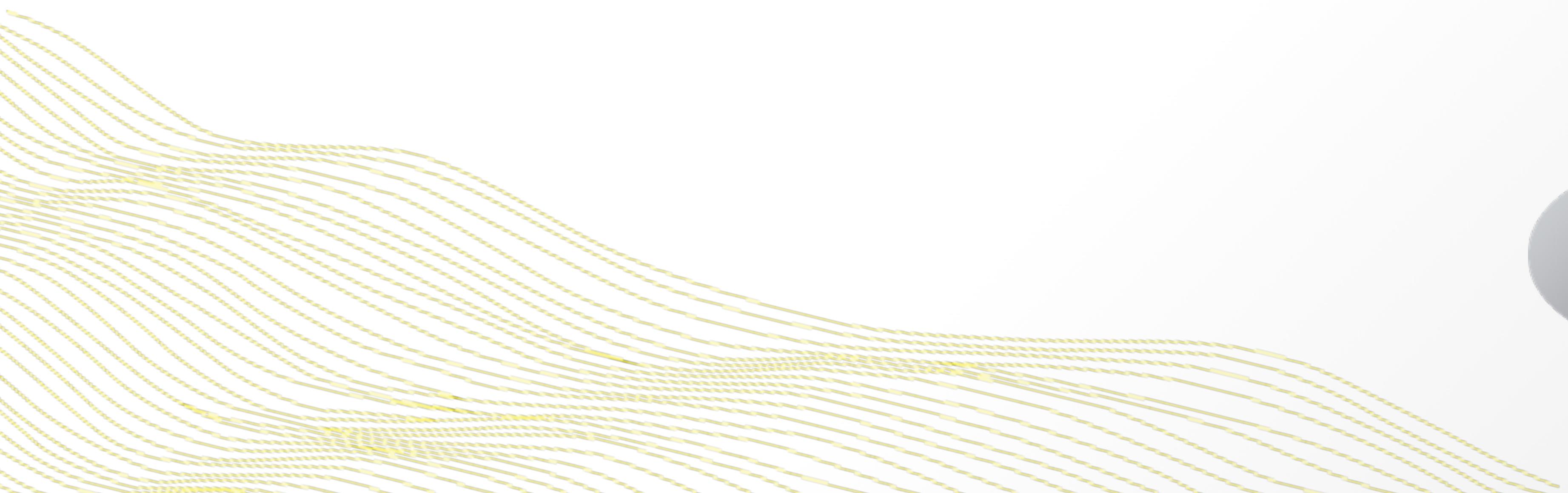
- Java 7 or later; Python 2.7, 3.5; Linux, Mac, Windows
- The easiest way to install the h2o Python module is PyPi.
- Latest version: <http://www.h2o.ai/download/h2o/python>

Design

All computations are performed in highly optimized Java code in the H2O cluster, initiated by REST calls from R.

H2O R & Python Tutorial

<http://tinyurl.com/wwcsv-h2o>



Tutorial: Intro to H2O Algorithms

The “Intro to H2O” tutorial introduces five popular supervised machine learning algorithms in the context of a binary classification problem.

The training module demonstrates how to train models and evaluating model performance on a test set.

- Generalized Linear Model (GLM)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)
- Deep Learning (DL)
- Naive Bayes (NB)

Tutorial: Grid Search for Model Selection

```
> print(gbm_gridperf)
H2O Grid Details
=====
Grid ID: gbm_grid2
Used hyper parameters:
- sample_rate
- max_depth
- learn_rate
- col_sample_rate
Number of models: 72
Number of failed models: 0

Hyper-Parameter Search Summary: ordered by decreasing auc
  sample_rate max_depth learn_rate col_sample_rate      model_ids          auc
1           1         3       0.19  1 gbm_grid2_model_38 0.685166598389755
2           0.9       3       0.15  1 gbm_grid2_model_53 0.684956999713052
3           0.8       5       0.06  1 gbm_grid2_model_22 0.684843506375254
4           0.6       4       0.07  1 gbm_grid2_model_4   0.684327718715252
5           0.95      4       0.13  1 gbm_grid2_model_48 0.684042497773235
```

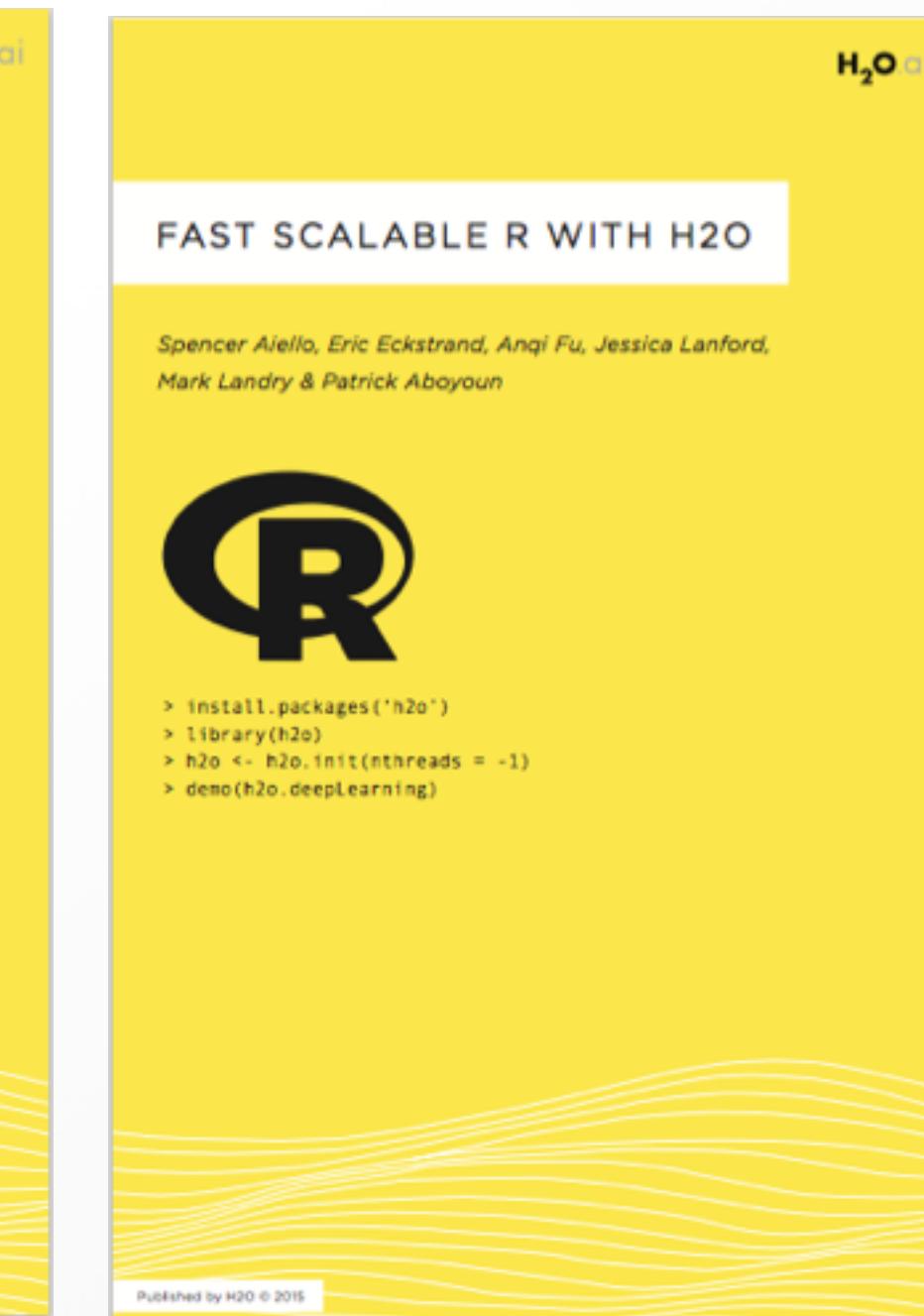
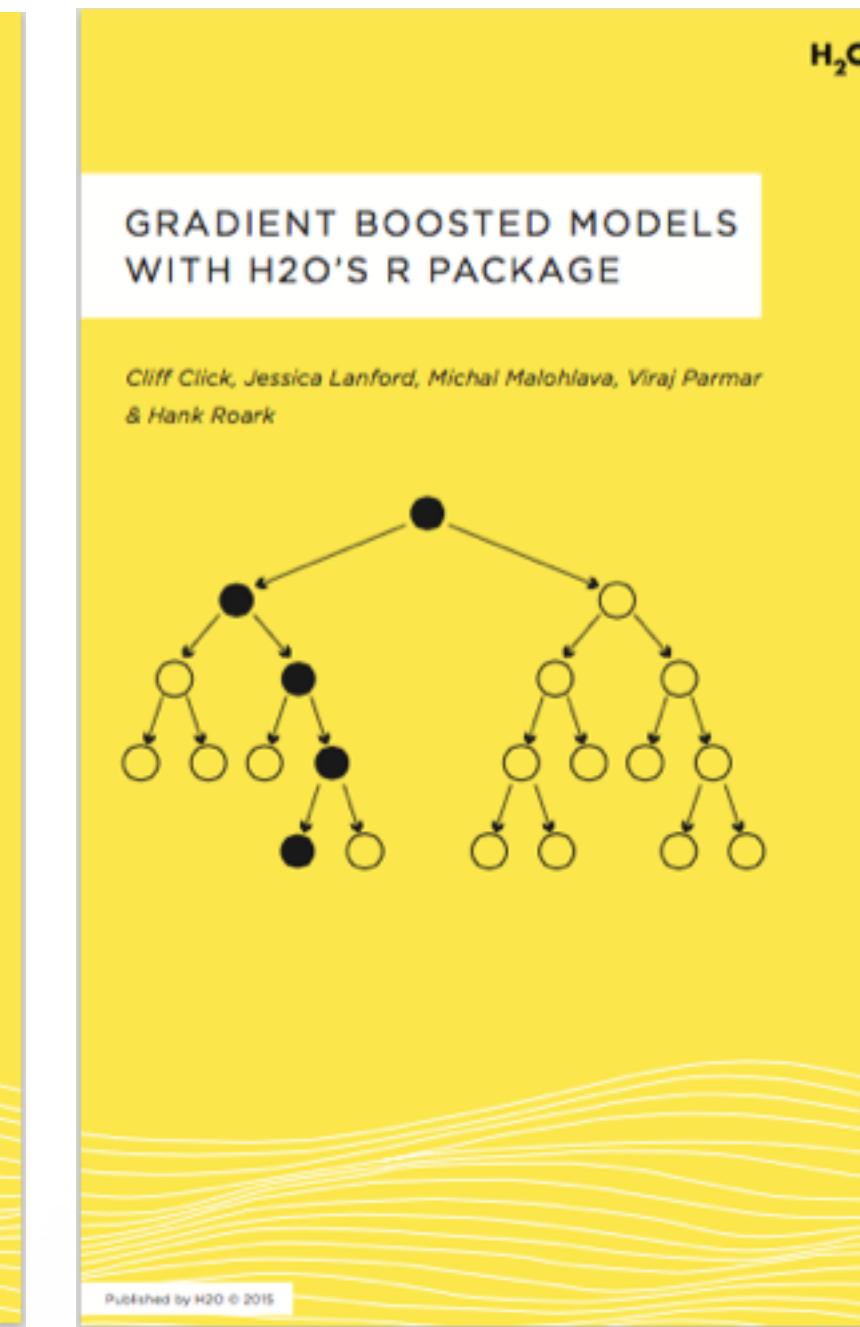
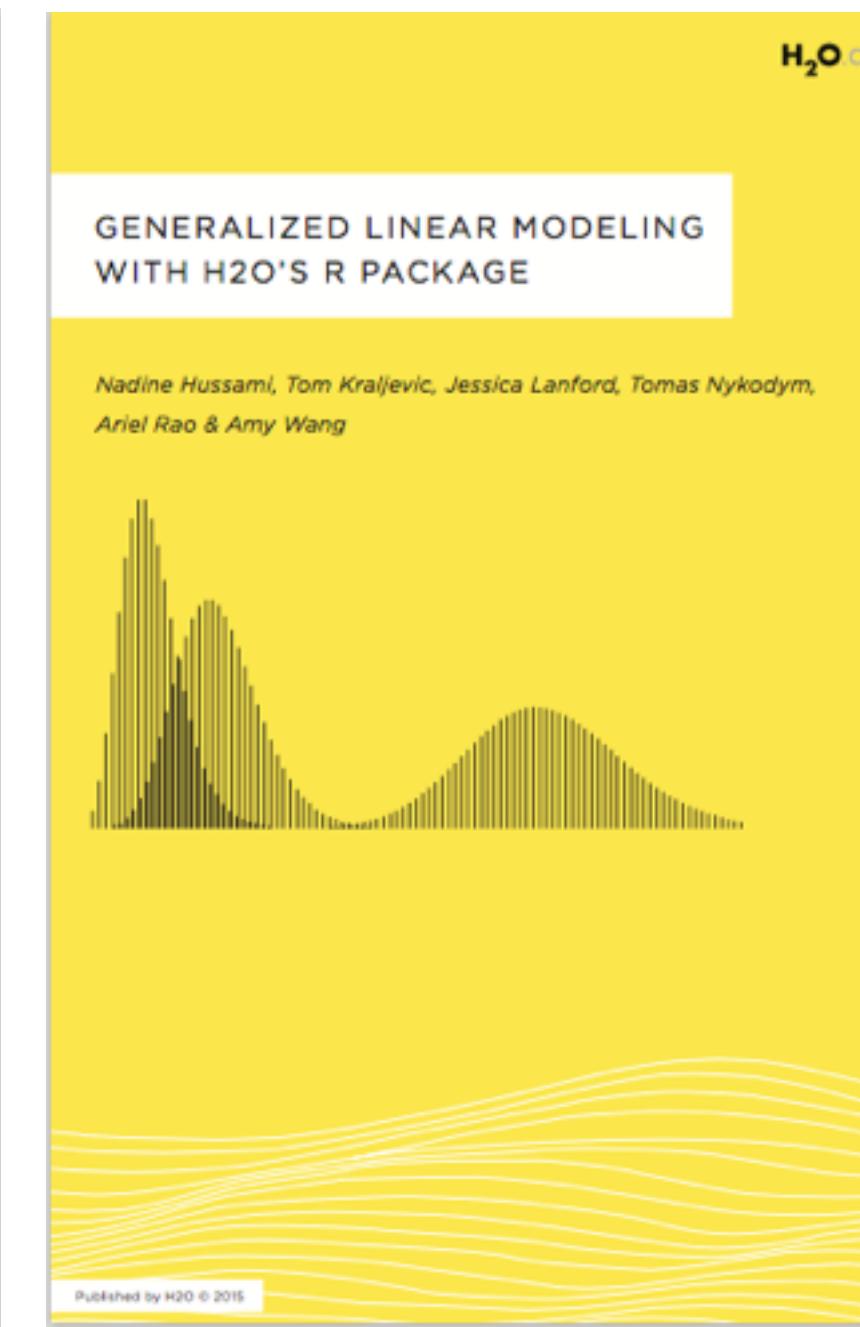
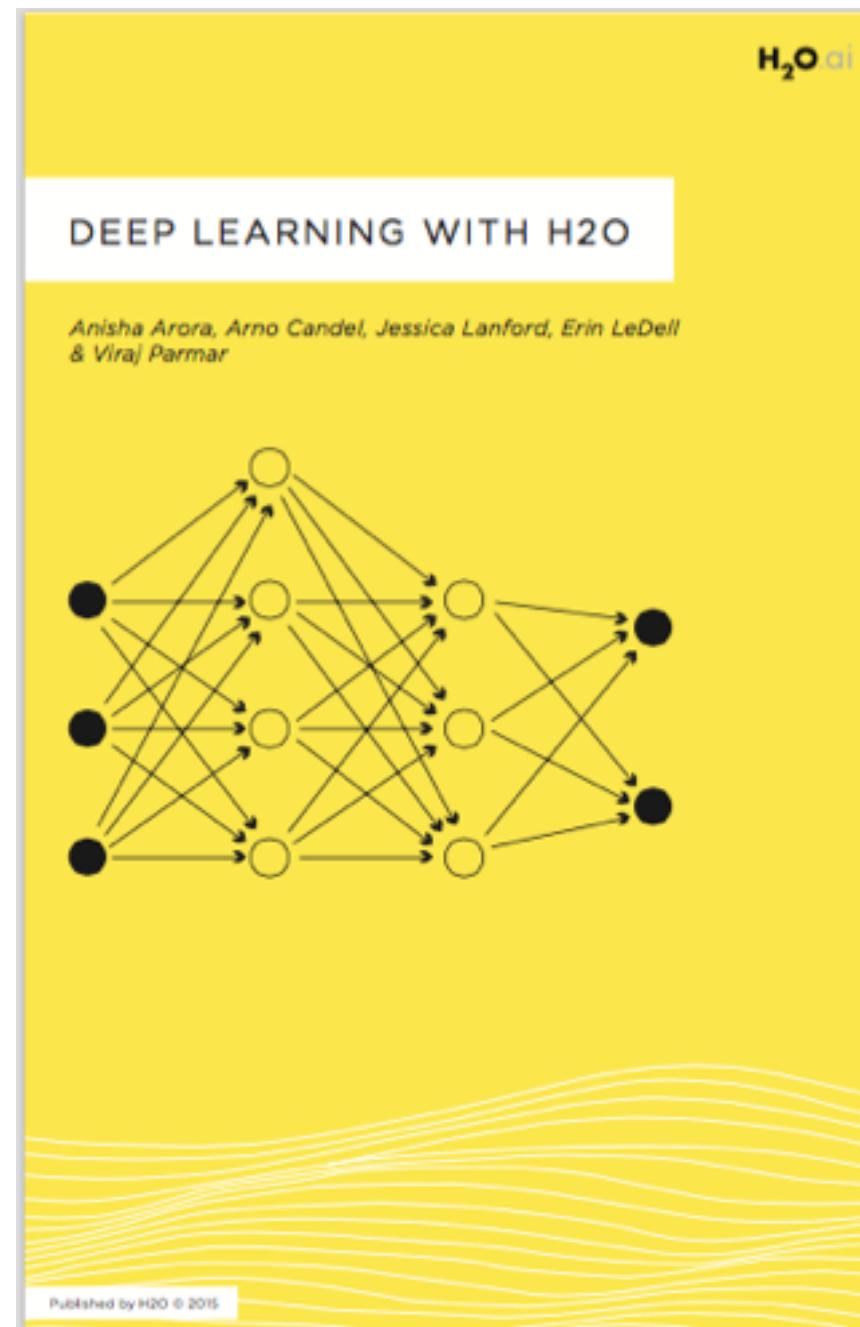
The second training module demonstrates how to find the best set of model parameters for each model using Grid Search.

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



<http://tinyurl.com/h2o-github-booklets>

Thank you!

@ledell on Github, Twitter
erin@h2o.ai

<http://www.stat.berkeley.edu/~ledell>