Deep Learning with R and H2O

Professor Matthew Lanham
Academic Director, MS BAIM Program
Associate Director of Student Engagements,
Krenicki Center for Business Analytics & Machine Learning Center
MatthewALanham.com

Materials

You can get the slides and codes for this presentation here:

https://github.com/MatthewALanham/2019_MWDSI

Agenda

- About H2O Platform
- **h2o** R Library/API demo
- Deep Learning & AutoML (Automated Machine Learning) demo

H2O.ai is Democratizing Artificial Intelligence

- https://www.h2o.ai/ was founded in 2012
- Provide scalable architecture + distributed machine learning algorithms to tackle small and big data problems
- They claim to be focusing on automated AI and making AI more accessible to everyone
- Have 5000+ customers and there are several interesting customer stories on their website (https://www.h2o.ai/customer-stories/)
- August 2019 – received $72.5M in Series D funding

H2O Pros

- Open source (Apache 2.0 licensed)
- Well-documented and commercially supported
  - Bookmark: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html
- Easy to use (technical to sort-of-technical person)
- Scalable to big data
- Has mature architecture
- OS: Windows, Mac, Ubuntu, RHEL/CentOS
- APIs - R, Python, Scala, Java, or a Web GUI (no programming)
- Automated Model Building functionality via AutoML, "Driverless AI"
- Can train and tune a deep learning model in one line of code

H2O Pros

"Driverless AI is automatic machine learning, which brings the power of a world-class data scientists in the hands of everyone. It builds models automatically using machine learning algorithms of every kind."

Sri Ambati, H2O Co-Founder & CEO

H2O Platform

High performance learning

- Distributed (multi-core + multi-node) implementations of ML algorithms
- Core algorithms written in high performance Java

Use favorite language, environment, and easily deploy into action

- APIs available that make it easy to work with big data from you laptop using your favorite analytics language
- Meant to work anywhere – Your laptop, Hadoop, Spark, EC2, etc.
- Easily deploy models to production as pure Java code OR if you can just save your models to disk as R/Python

H2O Distributed Computing

H2O Cluster

- Multi-node cluster with shared memory model
- All computations are in memory
- Each node only sees 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

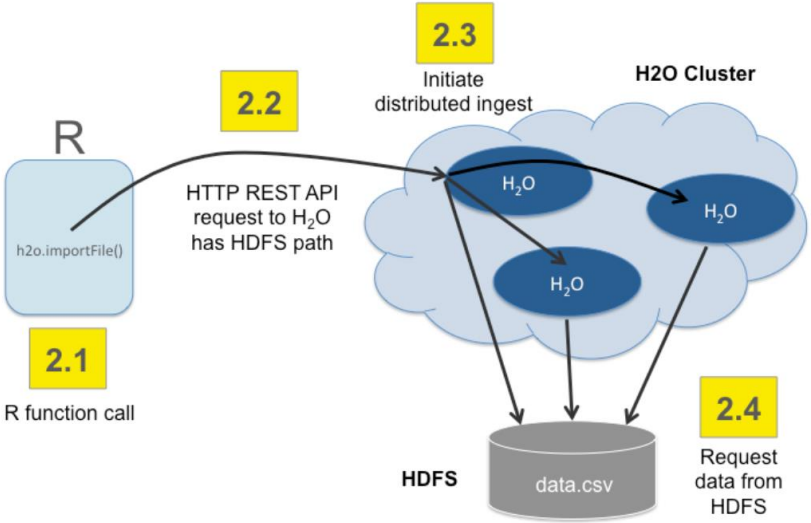
H2O Frame

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

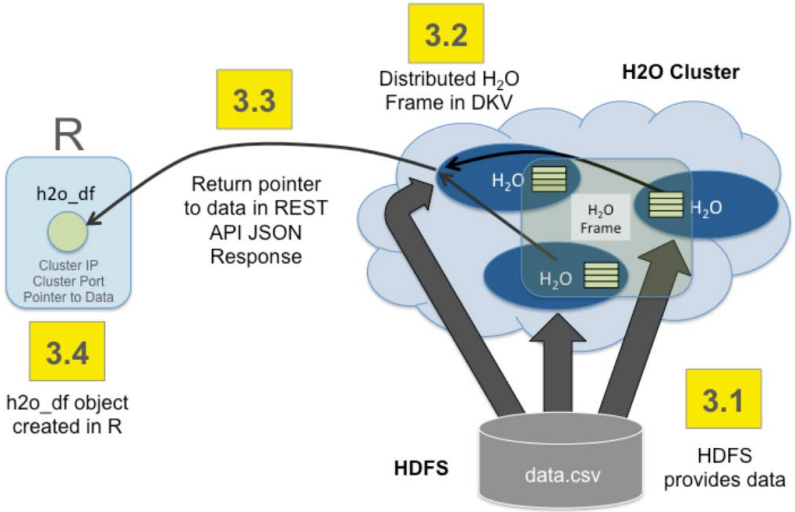
H2O Distributed Computing

Source: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/architecture.html

The R client tells the cluster to read the data



The data is returned from HDFS into a distributed H2O Frame



H2O Algorithms + Common Workflow Tasks

Supervised

- Cox Proportional Hazards (CoxPH)
- Deep Learning (Neural Networks)
- Distributed Random Forest (DRF)
- Generalized Linear Model (GLM)
- Gradient Boosting Machine (GBM)
- Naïve Bayes Classifier
- Stacked Ensembles
- Support Vector Machine (SVM)
- XGBoost

Unsupervised

- Aggregator
- Generalized Low Rank Models (GLRM)
- Isolation Forest
- K-Means Clustering
- Principal Component Analysis (PCA)

Common

- Quantiles
- Early Stopping

Miscellaneous

- Word2vec

Generic Models

- Generic Models

MOJO Models

- MOJO Models

Common workflow tasks:

- Imputation, normalization, auto one-hot encoding
- Cross-validation, grid or random search
- Variable importance, model evaluation metrics, plots

Source: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html#

H2O Installation

- You obtain **h2o** for R just like any other R package:

install.packages("h2o")

- The library is really **just an API**
- All the data is stored on the cluster (the server), not on our client. Even when the client and cluster are the same machine.

  Thus, when we want to train a model or make predictions, we first have to get the data into the H2O cluster.

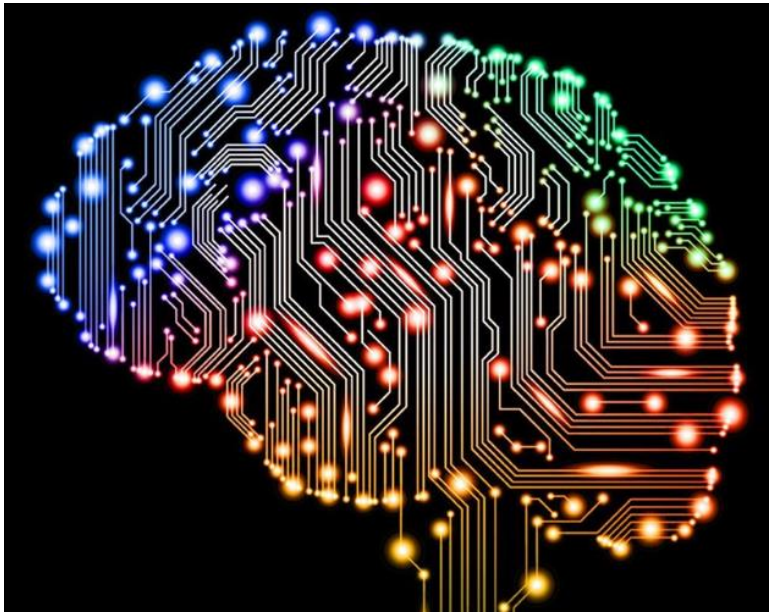
- You will need java, but most likely you already have that on your machine. If not, go to: http://www.oracle.com/technetwork/java/javase/downloads/index.html

Deep Learning

Professor Matthew Lanham
Academic Director, MS BAIM Program
Associate Director of Student Engagements,
Krenicki Center for Business Analytics & Machine Learning Center
MatthewALanham.com

What is deep learning?

Deep learning is a branch of machine learning where a multi-layered (deep) architecture is used to map the relations between inputs or observed features and the outcome.



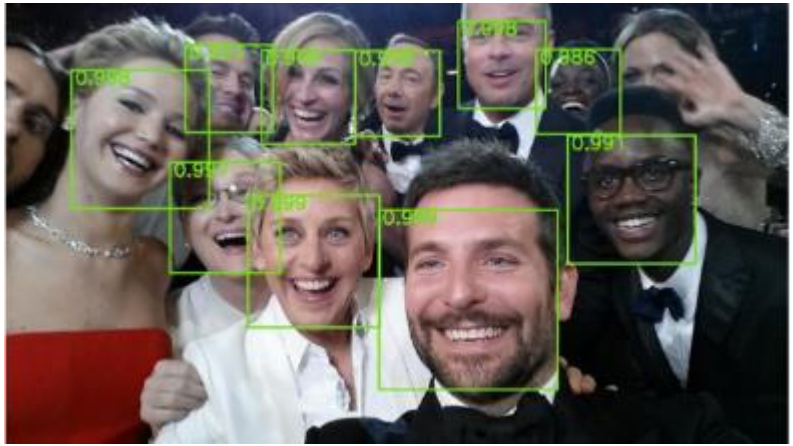
This deep architecture makes deep learning particularly suitable for handling a large number of features.

Also, features tend to be generated as part of the learning algorithm, rather than feature creation as a separate step.

What is deep learning?

Deep learning has been proven effective in the fields of

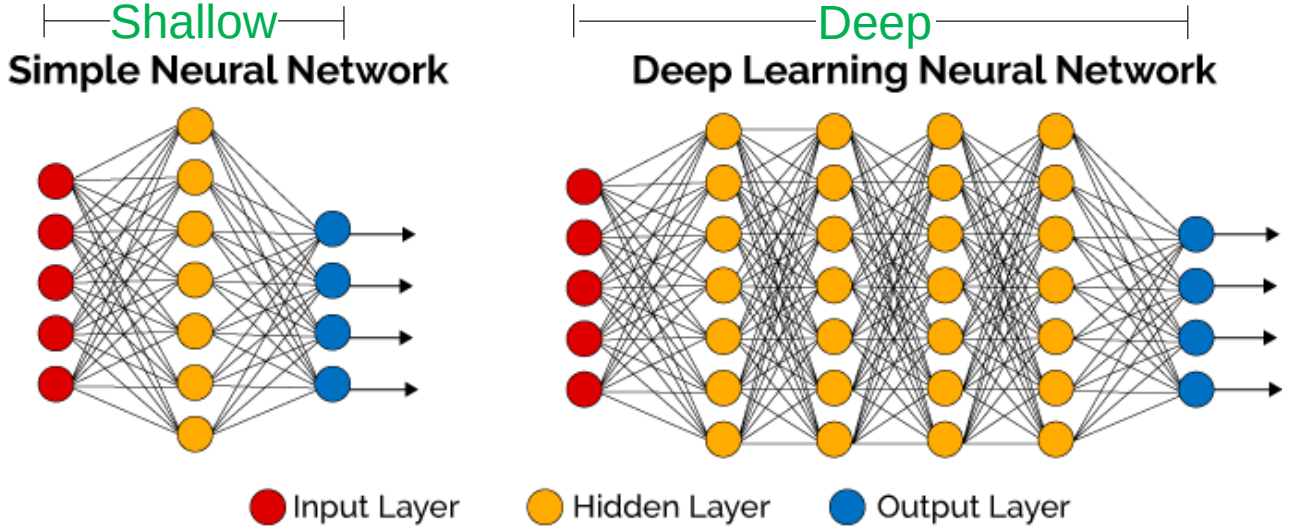
- image recognition and
- natural language processing



It is now even being used in more traditional modeling endeavors such as demand forecasting.

Deep Neural Networks

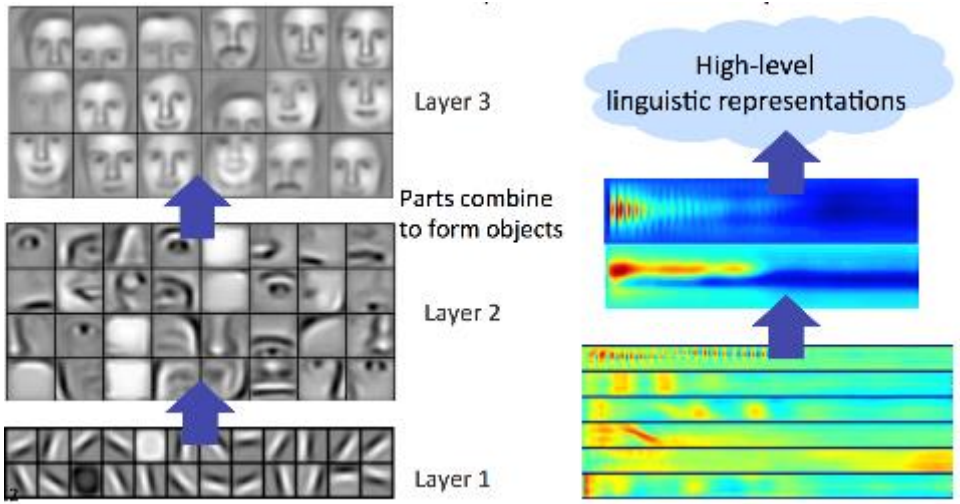
- The simplest definition of a **deep neural network (DNN)** is that it is a neural net with multiple hidden layers.
- Often such neural nets are referred to as a having a deep architecture, because there is a lot going on (many hidden layers). While, simple neural nets are referred to as being “shallow” (b/c one hidden layer).



Deep Neural Network pros

- Multiple hidden layers allow for a more sophisticated built-up from the simple elements to more complex ones.
- In the simple neural net, that considered predicting if an image was a circle or square, a deep neural net might have many circles and squares combined to form other more advanced shapes.

Example: Because the shallow NN cannot build more advanced shapes from basic pieces, in order to prove equal accuracy compared to the DNN, it must represent each unique object.



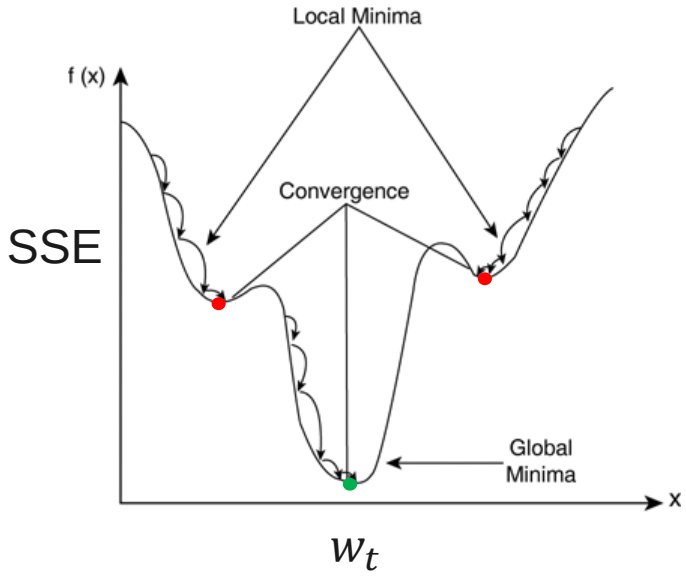
Deep Neural Network cons

One of the challenges in training DNNs is how to efficiently learn the weights.

- The models are often complex and local minima abound, making the optimization problem a challenging one.

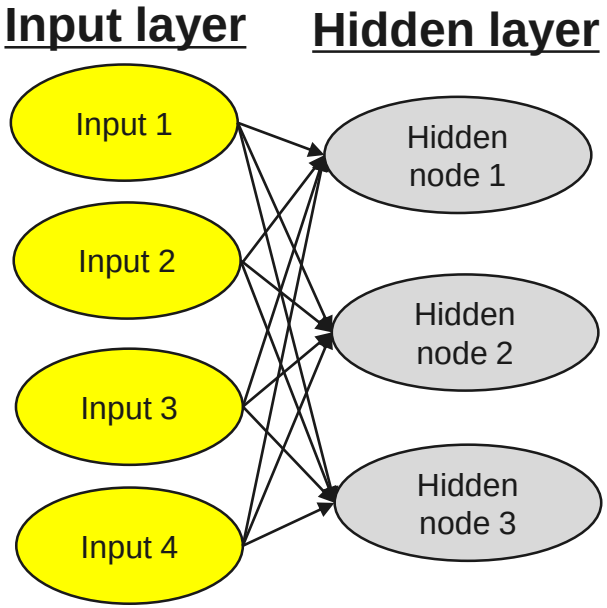
- In 2006, **Deep Belief Networks (DBNs)**, allowed modelers to train one layer at a time.

- DBMs are a type of deep neural network where multiple hidden layers and connections between layers (but not within).



RBMs to DBN

Deep Belief Networks (DBNs) essentially have the same definition of a NN of what is known as a **Restricted Boltzmann Machine (RBM),** which is a network having one input layer and one hidden layer.



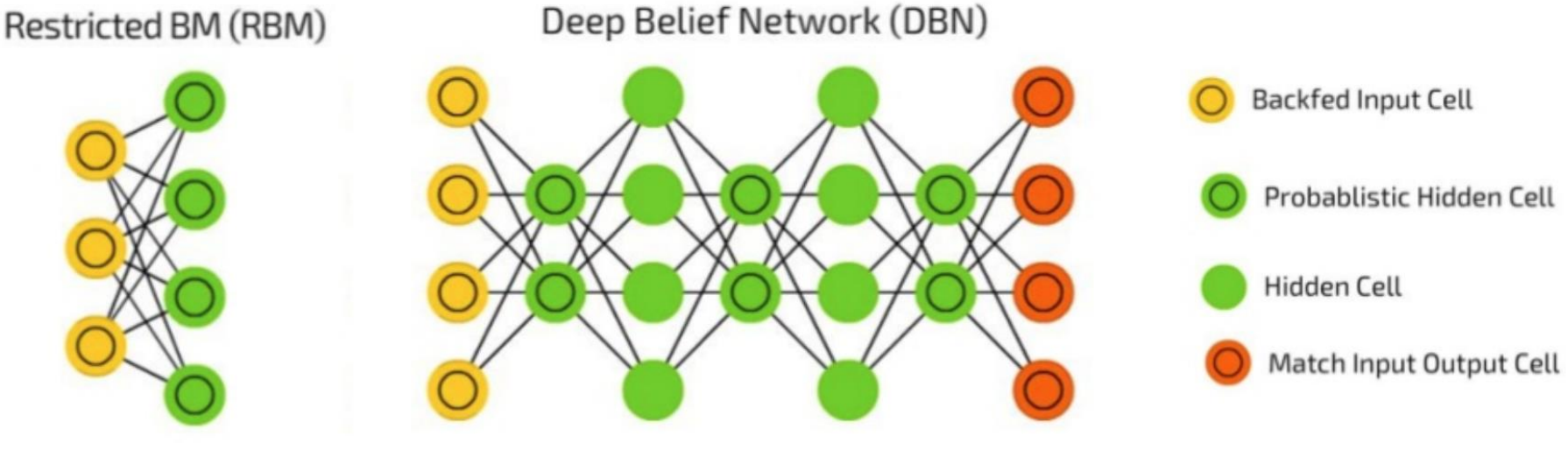
The restriction of no connections within a layer is valuable as it allows for much faster training algorithms to be used (e.g. contrastive divergence algorithm).

If several RBMs are stacked together, they can form a DBN..

Essentially, the DBN can be trained in a series of RBMs.

RBMs to DBN

The first RBM layer is trained and used to transform raw data into hidden neurons, which are then treated as a new set of inputs into a second RBM, and the process is repeated until all layers have been trained.



AutoML

Professor Matthew Lanham
Academic Director, MS BAIM Program
Associate Director of Student Engagements,
Krenicki Center for Business Analytics & Machine Learning Center
MatthewALanham.com

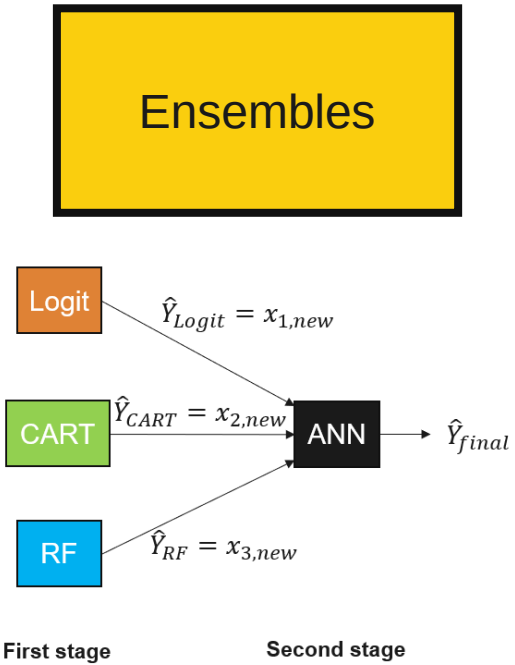
AutoML

AutoML is the path of least resistance for finding a competitive predictive model.

Data Preparation	Model Generation	Ensembles
• Imputation • One-hot encoding • Standardization • Label/Target encoding • Feature selection	• Logistic regression? • Tree? • Neural network? • Hyperparameter tuning • Generalizability/early stopping	• Focus on predictive performance? • Stacking/meta-modeling? • Ensemble selection

AutoML

- The goal of AutoML is to achieve competitive prediction.
- Ensembling
  - Combining multiple learners together.
  - Many different ways to do this.
  - Can perform well if you have decent base learners and the models have uncorrelated errors
- Stacking tries to find the optimal combination of base learners via a meta-model.
- AutoML uses Random Grid Search (GBMs, GLMs, etc.) with Stacked Ensembles
- Provides a **Leaderboard** output of the top models. Demo…



H2O Demo in RStudio

Professor Matthew Lanham
Academic Director, MS BAIM Program
Associate Director of Student Engagements,
Krenicki Center for Business Analytics & Machine Learning Center
MatthewALanham.com

Demo

See *h2o.R* script

H2O Flow

Professor Matthew Lanham
Academic Director, MS BAIM Program
Associate Director of Student Engagements,
Krenicki Center for Business Analytics & Machine Learning Center
MatthewALanham.com

H2O Flow

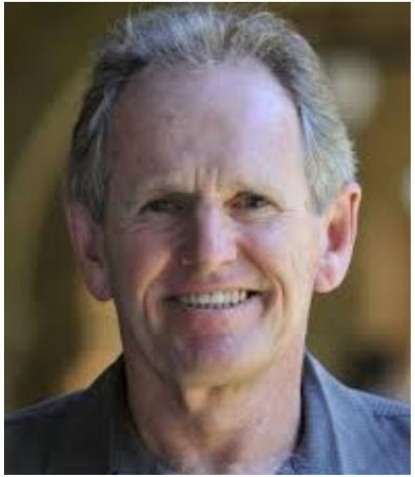
- Flow is the name of the web GUI that is part of H2O. It is just another client, making the same web service calls to the H2O backend that the R client is making.
- You can do the following:
  - *View data you have uploaded through your client*
  - *Upload data directly*
  - *View models you have created through your R client (and those currently being created)*
  - *Create models directly*
  - *View and run predictions you have generated through your client*
- If using your RStudio Desktop, open a browser and go here to see your flow
  - http://127.0.0.1:54321/flow/index.html
- If using RStudio Server, the link will be slightly modified (example):
  - http://rstudio.scholar.rcac.purdue.edu:54321/flow/index.html

Future investigations

- Testing performance on large scale datasets
- Importing non-native h2o models into the workflow
- Integrating models into applications outside of H2O

> “H2O.ai provides impressively scalable implementations of many of the important machine learning tools in a user-friendly environment. Allowing for free academic use sets a generous example for commercial software developers — it is also the way forward in the era of open-source software.”

Trevor J. Hastie
John A. Overdeck Professor of Mathematical Sciences
Professor of Statistics
Professor of Biomedical Data Science
Department of Statistics
Stanford University
USA



References and Resources

You can get the slides and codes for this presentation here:

https://github.com/MatthewALanham/2019_MWDSI

- Candel, A., Parmar, V., LeDell, E., and Arora, A. (Sept 2019). Deep Learning with H2O. http://h2o:ai/resources.
- https://github.com/DarrenCook/h2o
- https://github.com/h2oai/h2o-meetups/tree/master/2018_09_05_SF_Meetup_AutoML