H2O Introductory Training

Erin LeDell Ph.D.

Machine Learning Scientist, H2O.ai

May 3, 2016

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
- Written a handful of machine learning R packages

What is Data Science?

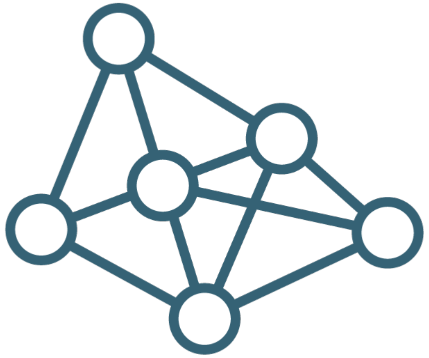
Data Science Tools?

What is Machine Learning?

What is Deep Learning?

What is Ensemble Learning?

R & Python Code Tutorial



What is Data Science?

Data Processing

- Identify a data task or prediction problem
- Collect relevant data

Problem Formulation

- Clean, transform, filter, aggregate, impute
- Convert into X and Y

Machine Learning

- Train models
- Evaluate models

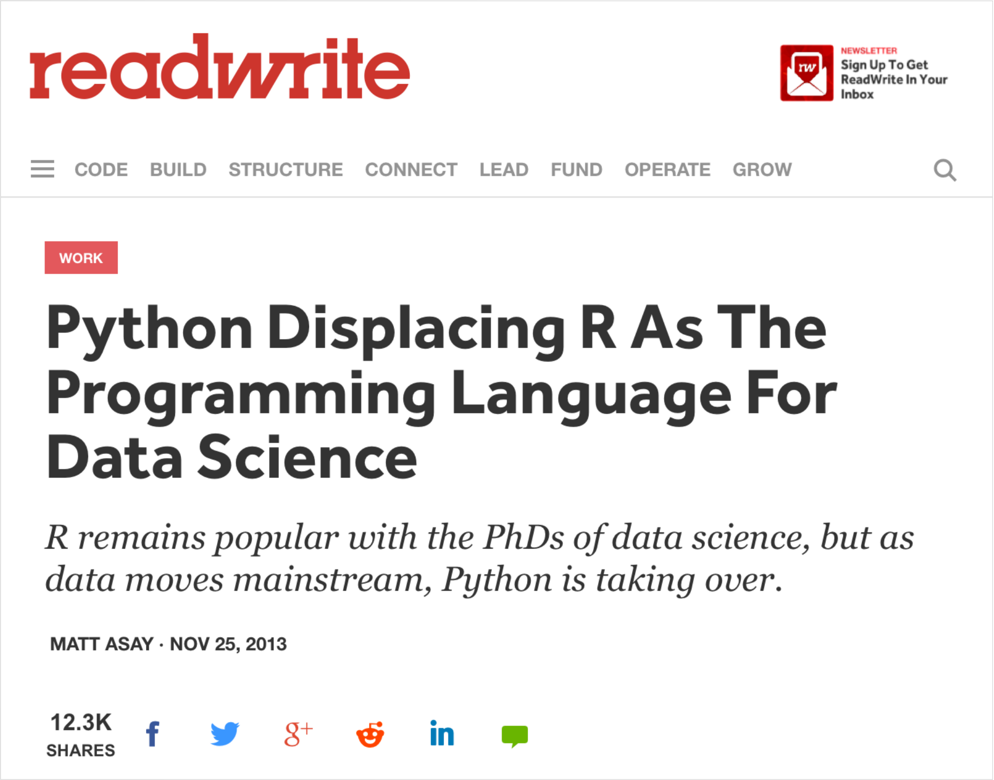
Survey of Data Scientists on LinkedIn

The number of data scientists has doubled over the last 4 years.

The top five skills listed by data scientists:

1. Data Analysis
2. R
3. Python
4. Data Mining
5. Machine Learning

Data Science Tools



readwrite

NEWSLETTER Sign Up To Get ReadWrite In Your Inbox

CODE BUILD STRUCTURE CONNECT LEAD FUND OPERATE GROW

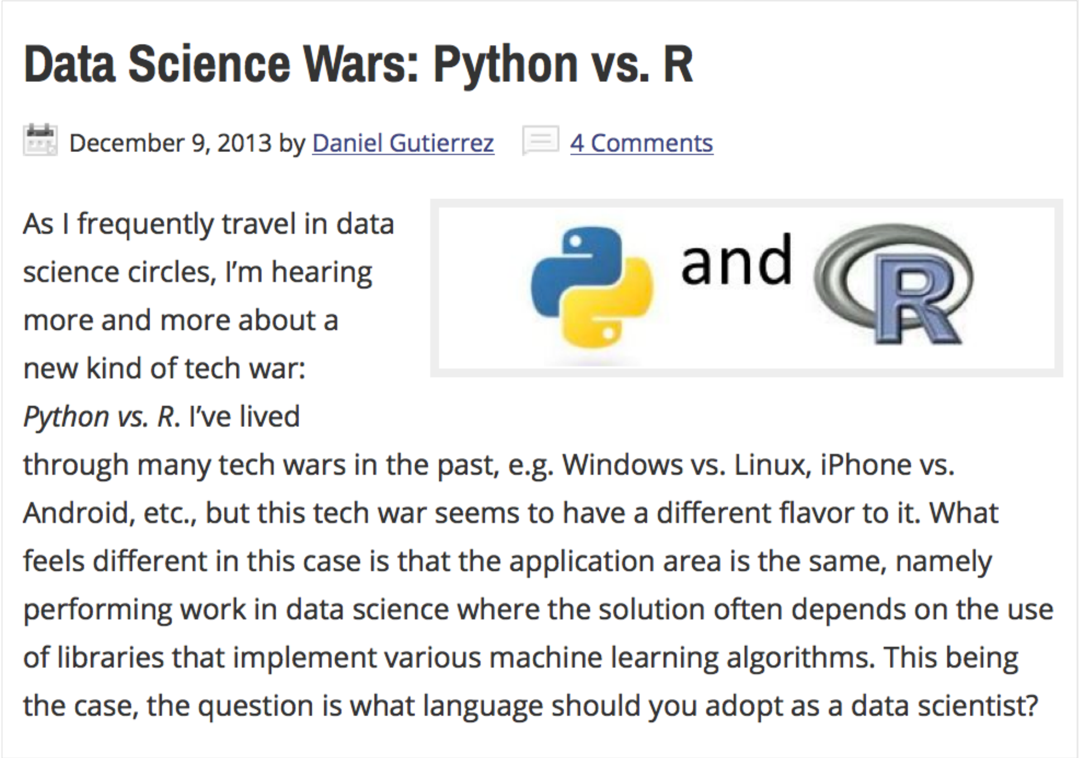
WORK

Python Displacing R As The Programming Language For Data Science

R remains popular with the PhDs of data science, but as data moves mainstream, Python is taking over.

MATT ASAY · NOV 25, 2013

12.3K SHARES



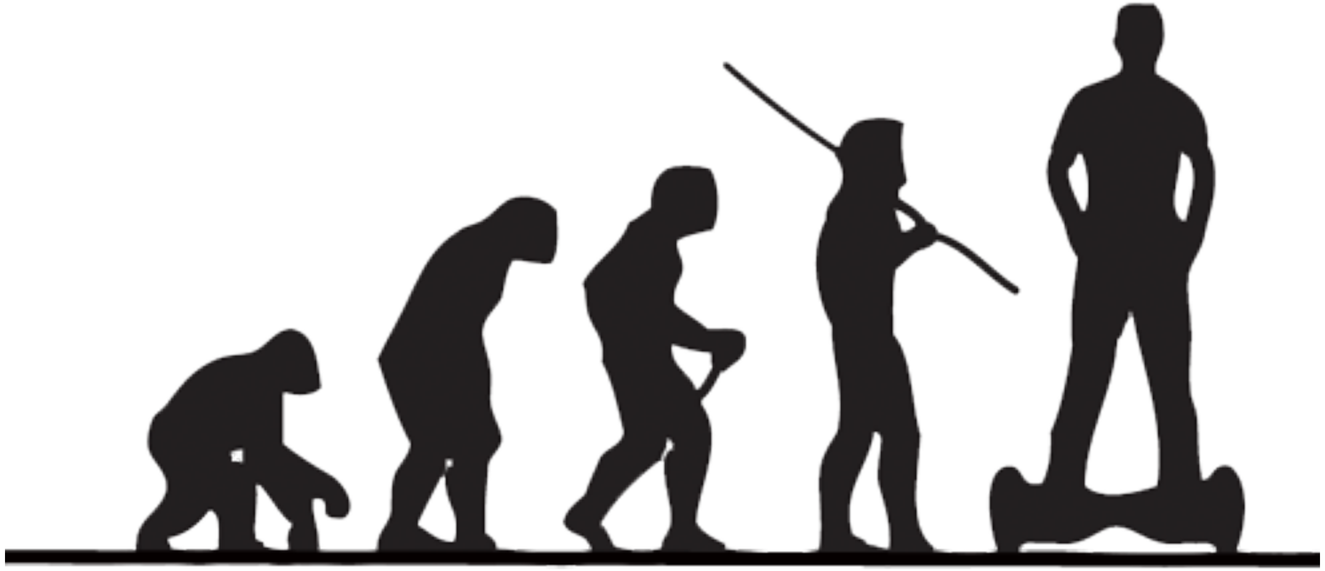
Data Science Wars: Python vs. R

December 9, 2013 by Daniel Gutierrez 4 Comments

As I frequently travel in data science circles, I'm hearing more and more about a new kind of tech war: *Python vs. R*. I've lived through many tech wars in the past, e.g. Windows vs. Linux, iPhone vs. Android, etc., but this tech war seems to have a different flavor to it. What feels different in this case is that the application area is the same, namely performing work in data science where the solution often depends on the use of libraries that implement various machine learning algorithms. This being the case, the question is what language should you adopt as a data scientist?

2013 was the year of the data science "language wars."

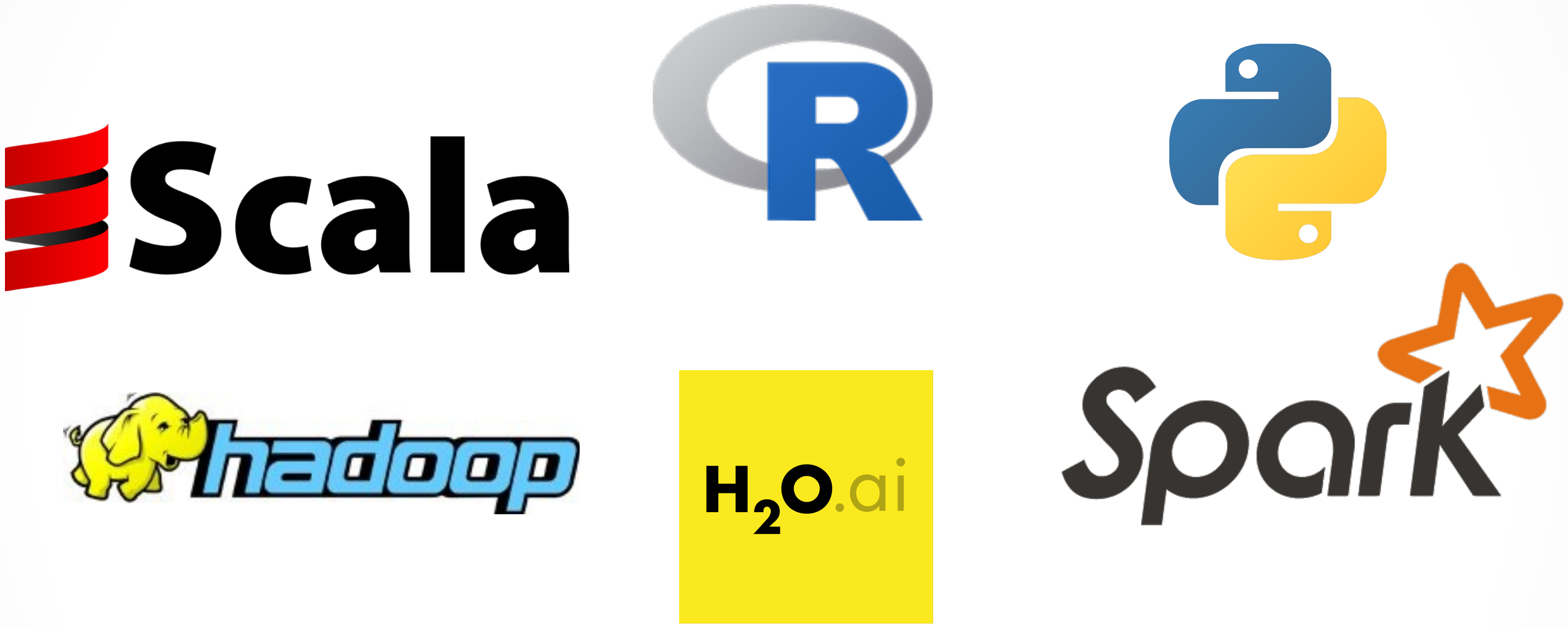
Data Science Tools



In 2016, we have evolved beyond this…

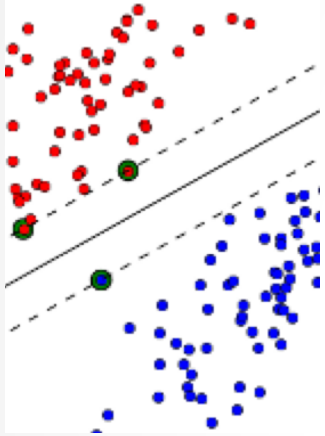
We are too busy doing actual data science!

Data Science Tools



We are headed toward language agnostic data science, where friendly APIs connect to powerful data processing engines.

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 R^2
- Super Learner finds the optimal combination of a combination of a collection of base learning algorithms.

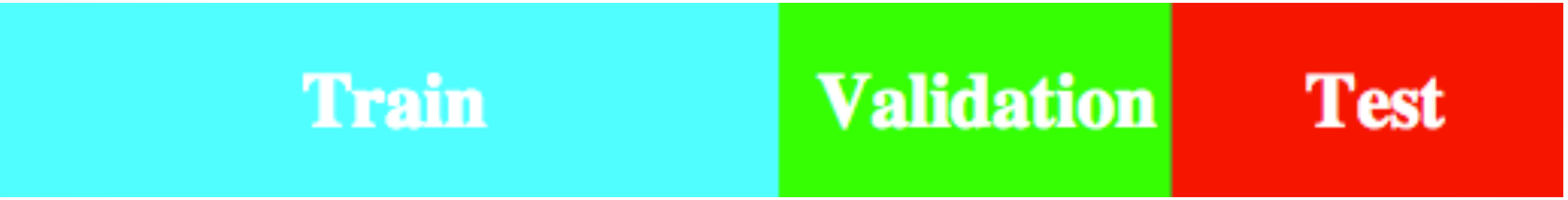
Classification

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

Clustering

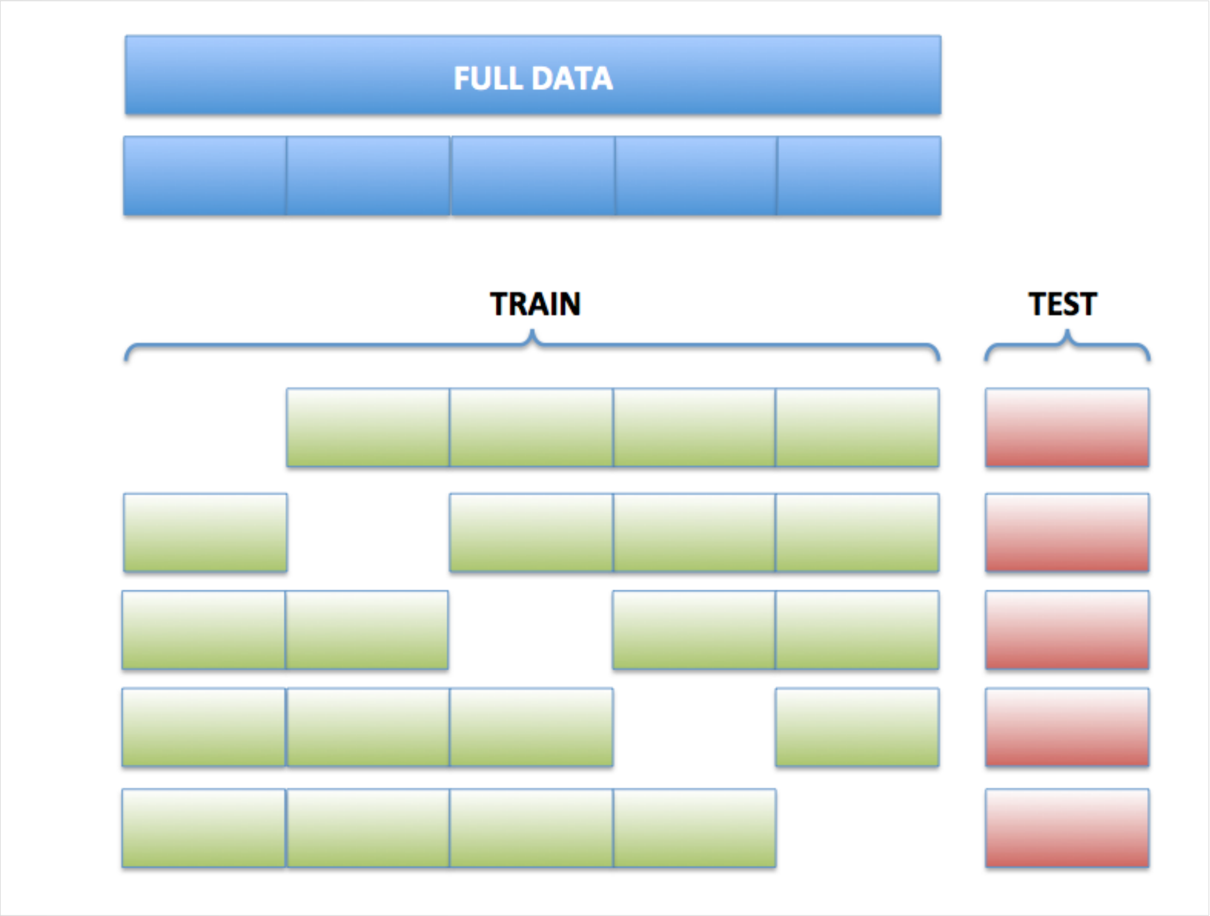
- Unsupervised learning (no training labels)
- Partition the data; identify clusters or sub-populations
- Evaluate with AIC, BIC or Total Sum of Squares

Train, Validation and Test Set



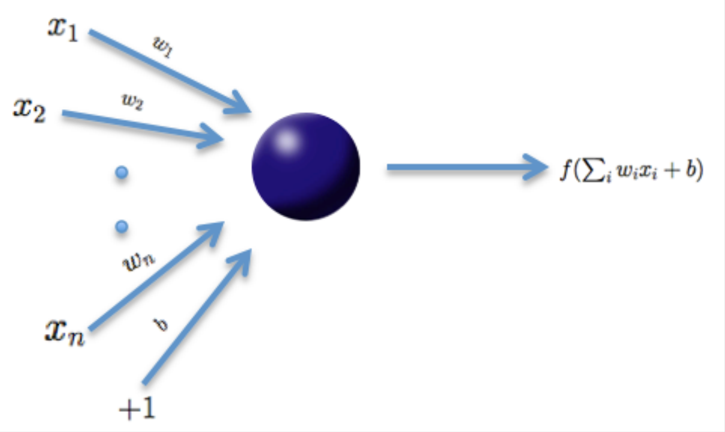
- If you plan on doing any model tuning, you should split your dataset into three parts: Train, Validation and Test
- There is no general rule for how you should partition the data and it will depend on how strong the signal in your data is, but an example could be:
  50% Train, 25% Validation and 25%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

K-fold Cross-validation



- K-fold Cross-validation (CV) is used to evaluate the performance of machine learning algorithms.
- CV will give you the most “mileage” on your training data.
- Performance metrics are averaged across k folds.

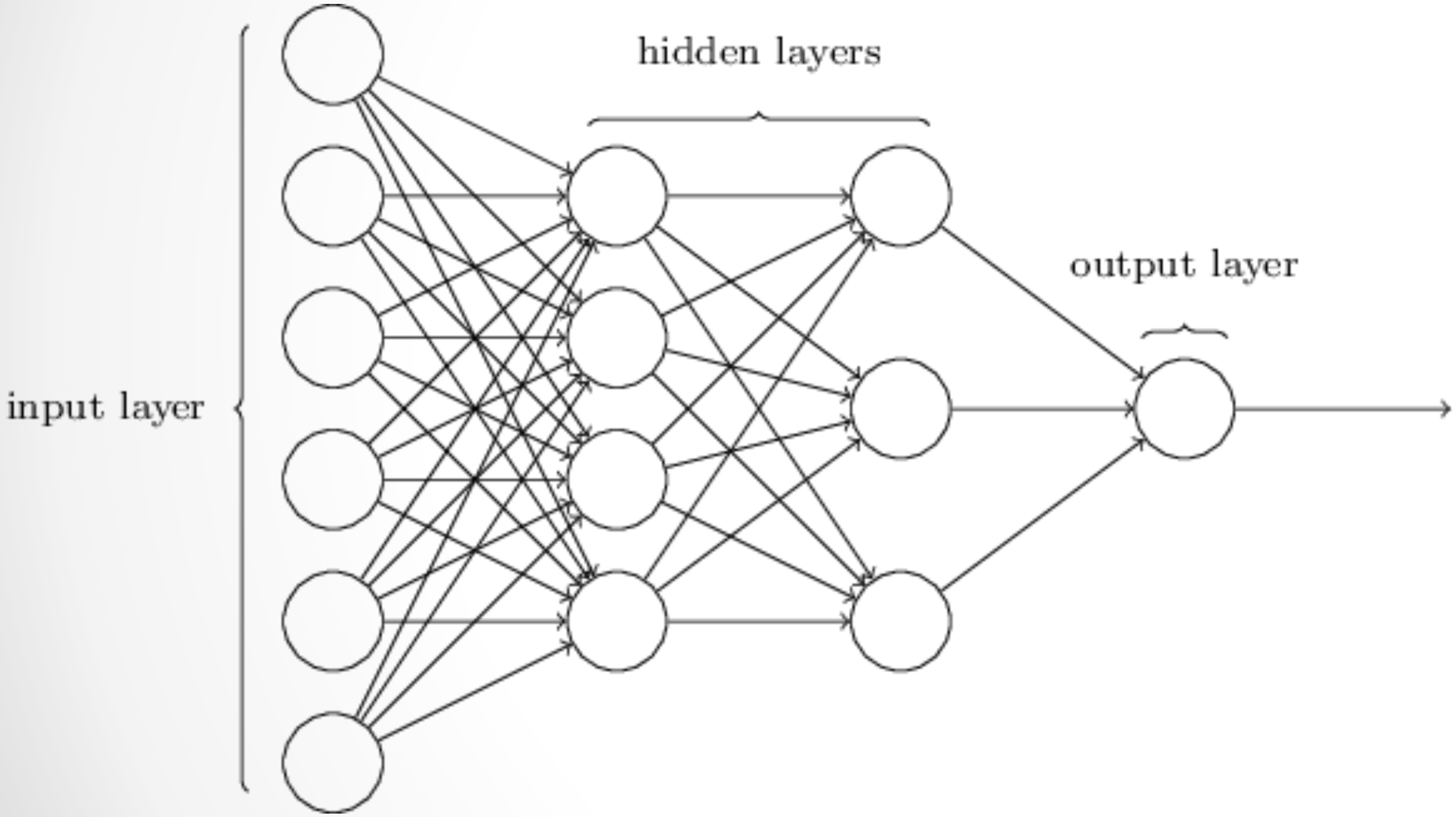
What is Deep Learning?



> ″A branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, composed of multiple non-linear transformations.″
>
> — Wikipedia (2016)

- Deep neural networks have more than one hidden layer in their architecture. That’s why they are called “deep” neural networks.
- Very useful for complex input data such as images, video, audio.

What is Deep Learning?



- Deep learning architectures, specifically artificial neural networks (ANNs) have been around since 1980.
- However, there were breakthroughs in training techniques that lead to their recent resurgence in the mid 2000's.
- Combined with modern computing power, they are quite effective.

What is Ensemble Learning?

> "Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms."
>
> — Wikipedia (2016)

- Random Forests and Gradient Boosting Machines (GBM) are both ensembles of decision trees.
- Stacking, or Super Learning, is technique for combining various learners into a single, powerful learner using a second-level metalearning algorithm.

No Free Lunch

"Even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience.”

— David Hume (1711-1776)

- No general purpose algorithm to solve all problems.
- No right answer on optimal data preparation.
- Some algorithms may have such strong biases that they can only learn certain kinds of functions.

H2O R & Python Code Tutorial

http://tinyurl.com/h2o-chicago

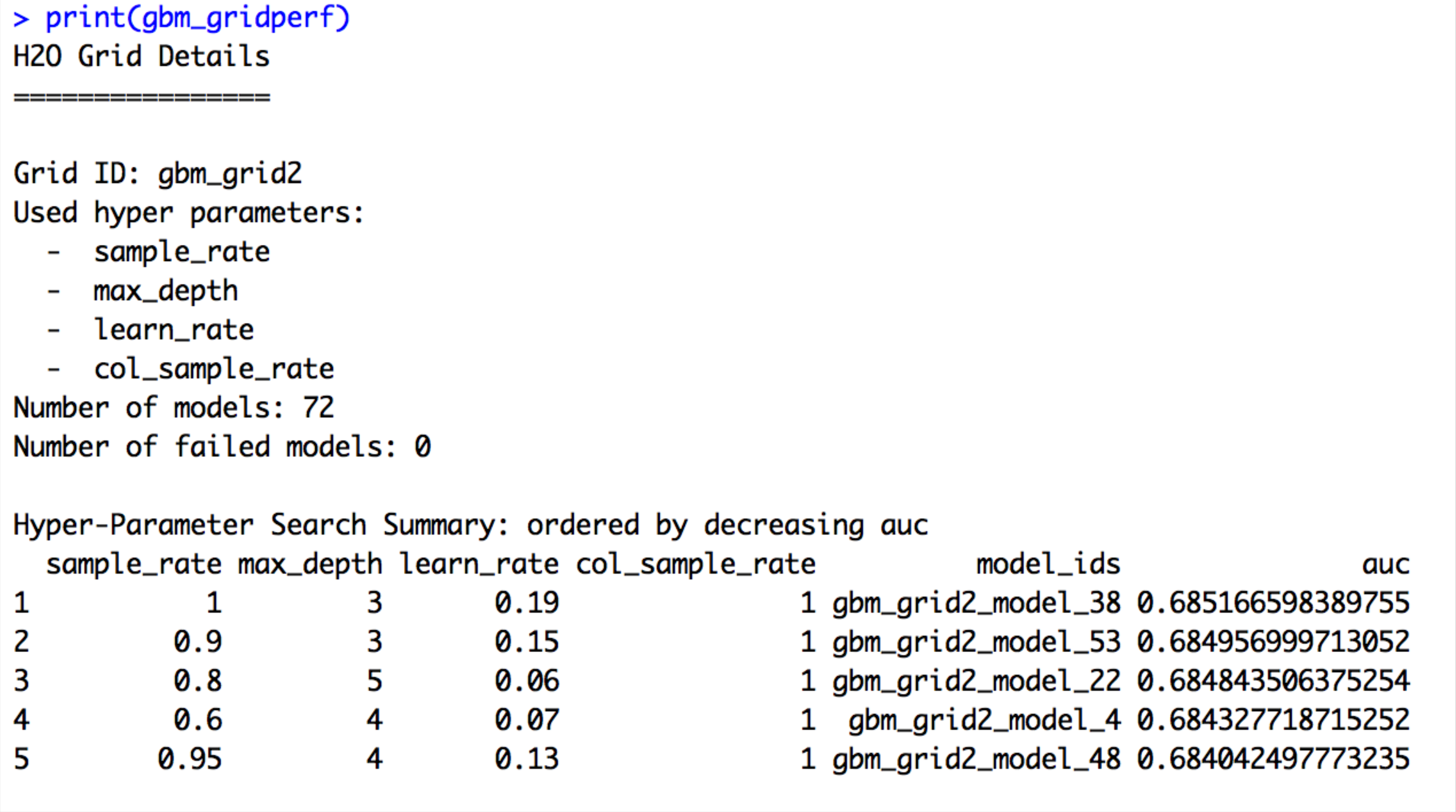
Tutorial: Intro to H2O Algorithms

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