Scalable Automatic Machine Learning with H2O

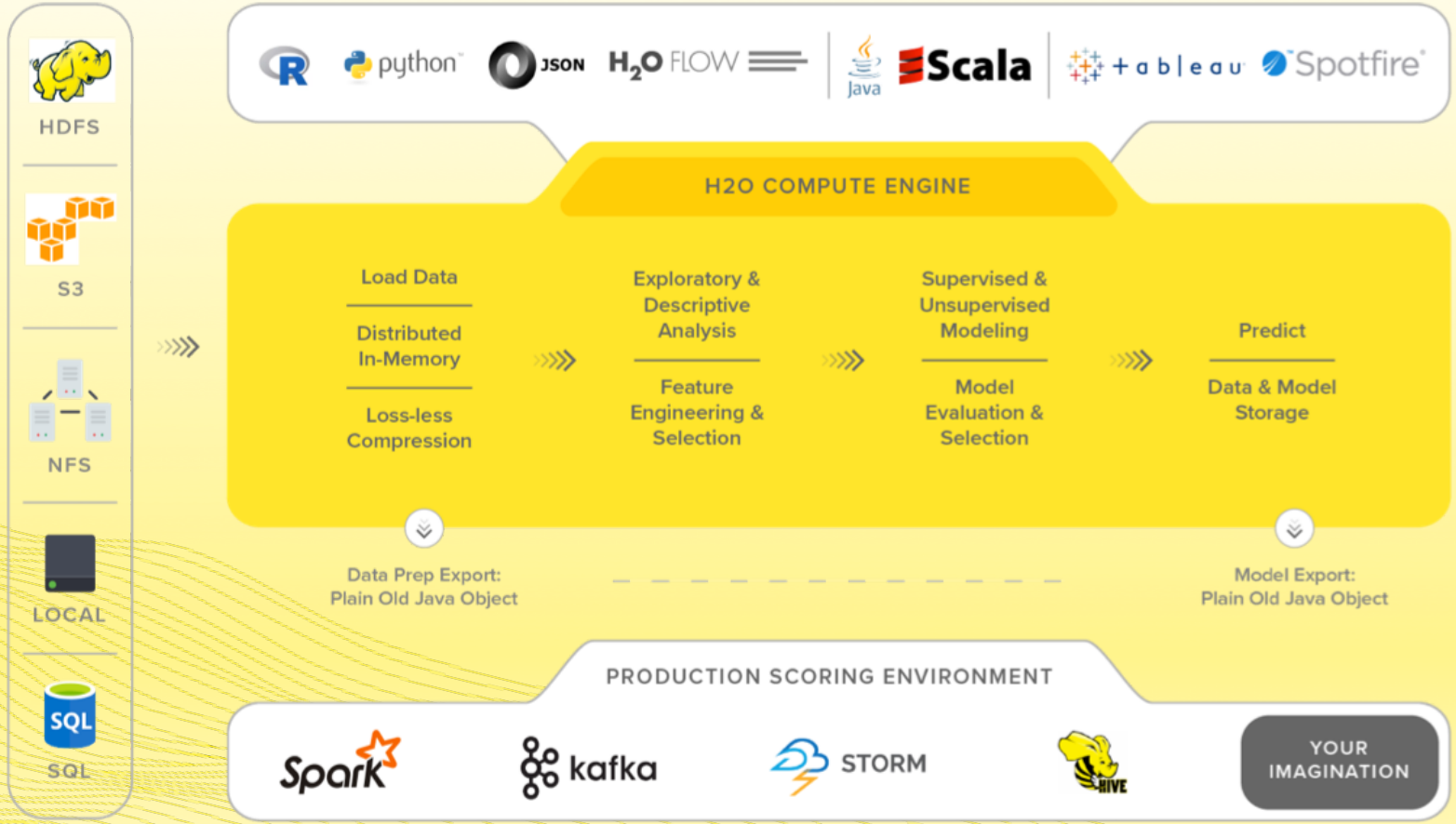
Parul Pandey
Data Science Evangelist, H2O.ai

Agenda

- H2O Platform
- Automatic Machine Learning (AutoML)
- H2O AutoML Overview
- Demo

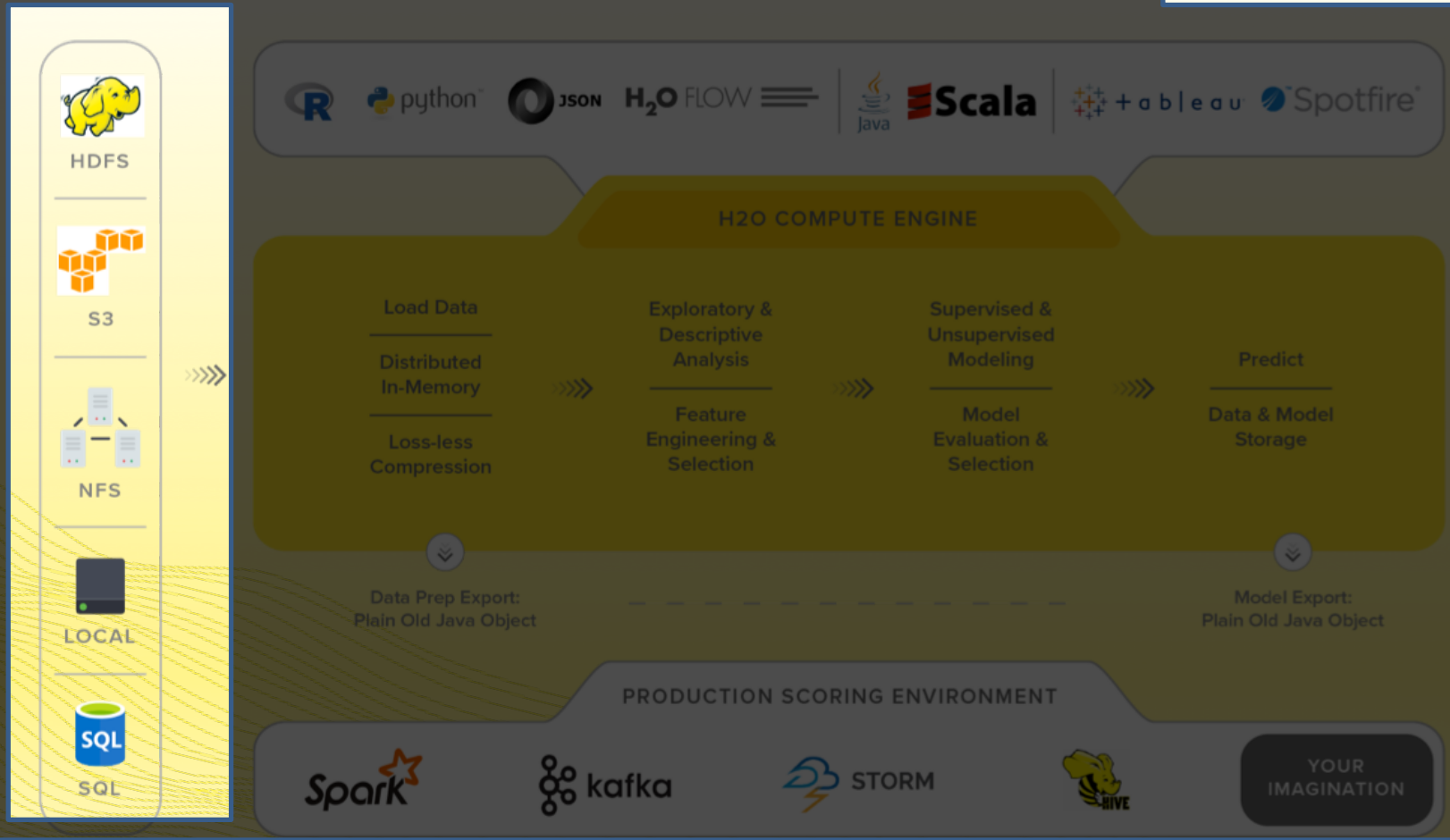
H2O Platform

High Level Architecture

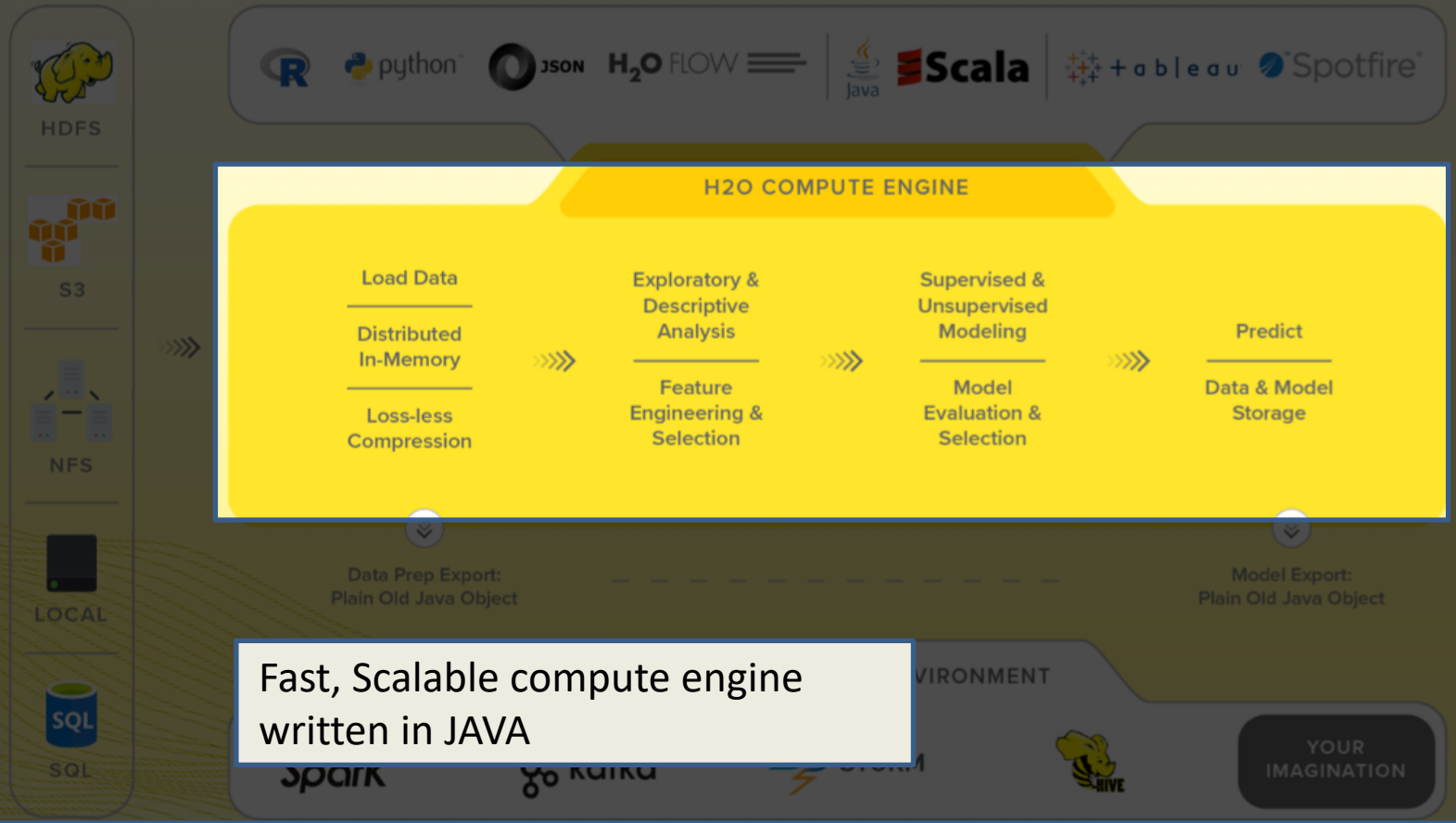


High Level Architecture

Import Data from multiple Data Sources

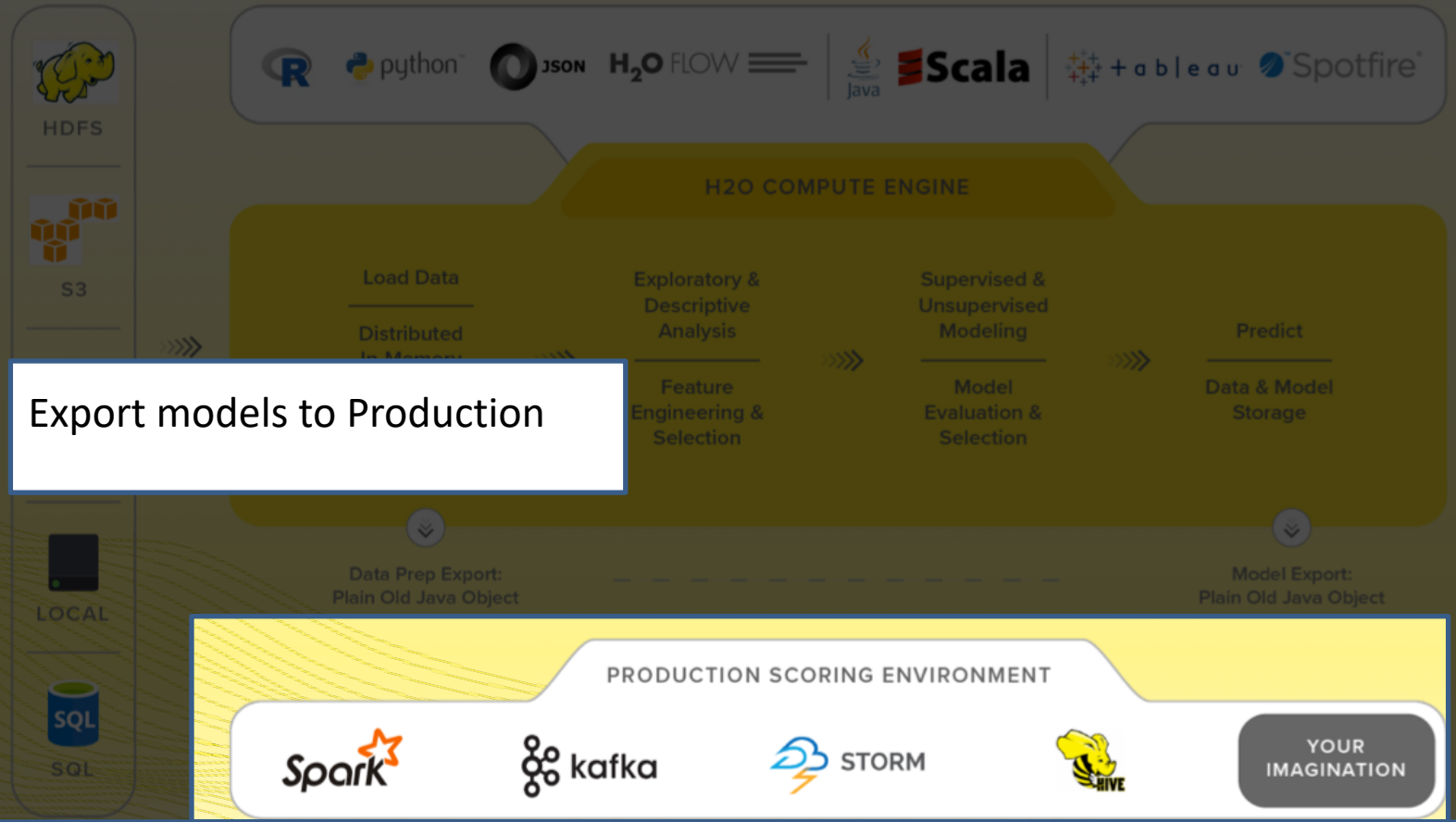


High Level Architecture



Fast, Scalable compute engine written in JAVA

High Level Architecture



Export models to Production

H2O Machine Learning Methods

H2O.ai

Supervised Learning

Statistical Analysis

- **Penalized Linear Models**: Super-fast, super-scalable, and interpretable
- **Naïve Bayes:** Straightforward linear classifier

Decision Tree Ensembles

- **Distributed Random Forest**: Easy-to-use tree-bagging ensembles
- **Gradient Boosting Machine**: Highly tunable tree-boosting ensembles
- **eXtreme Gradient Boosting**: Popular XGBoost algorithm in H2O

Stacking

- **Stacked Ensemble**: Combine multiple types of models for better predictions

AutoML

- **Automatic Machine Learning**: Automated exploration of supervised learning approaches

Unsupervised Learning

Clustering

- **K-means**: Partitions observations into similar groups; automatically detects number of groups

Dimensionality Reduction

- **Principal Component Analysis**: Transforms correlated variables to independent components
- **Generalized Low Rank Models:** Extends the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Aggregator

- **Aggregator:** Efficient, advanced sampling that creates smaller data sets from larger data sets

Neural Networks

Multilayer Perceptron

Deep Learning

- **Deep neural networks**: Multi-layer feed-forward neural networks for standard data mining tasks
- **Convolutional neural networks:** Sophisticated architectures for pattern recognition in images, sound, and text

Anomaly Detection

Term Embeddings

- **Autoencoders**: Find outliers using a nonlinear dimensionality reduction technique
- **Word2vec:** Generate context-sensitive numerical representations of a large text corpus

H2O Machine Learning Features

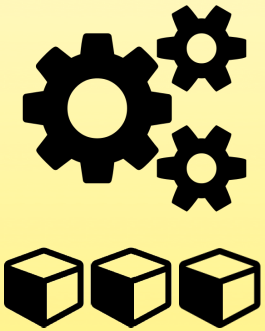
- Supervised & unsupervised machine learning algos
- Imputation, normalization & auto one-hot-encoding
- Automatic early stopping
- Cross-validation, grid search & random search
- Variable importance, model evaluation metrics, plots

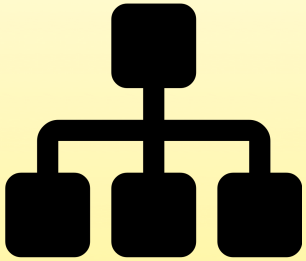
Intro to Automatic Machine Learning

Aspects of Automatic Machine Learning

Data Prep

Model Generation

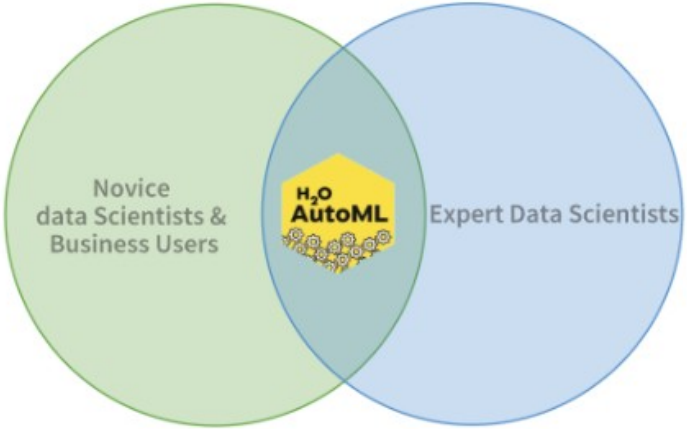




Ensembles

Who is it for?

- **AUTOMATES**
  - Basic Preprocessing
  - Model Training
  - Model Tuning with Validation
  - Stacking
  - Model's Results table



- **FREES TIME FOR**
  - Data Preprocessing
  - Feature Engineering
  - Model Deployment

H2O's AutoML

H2O AutoML

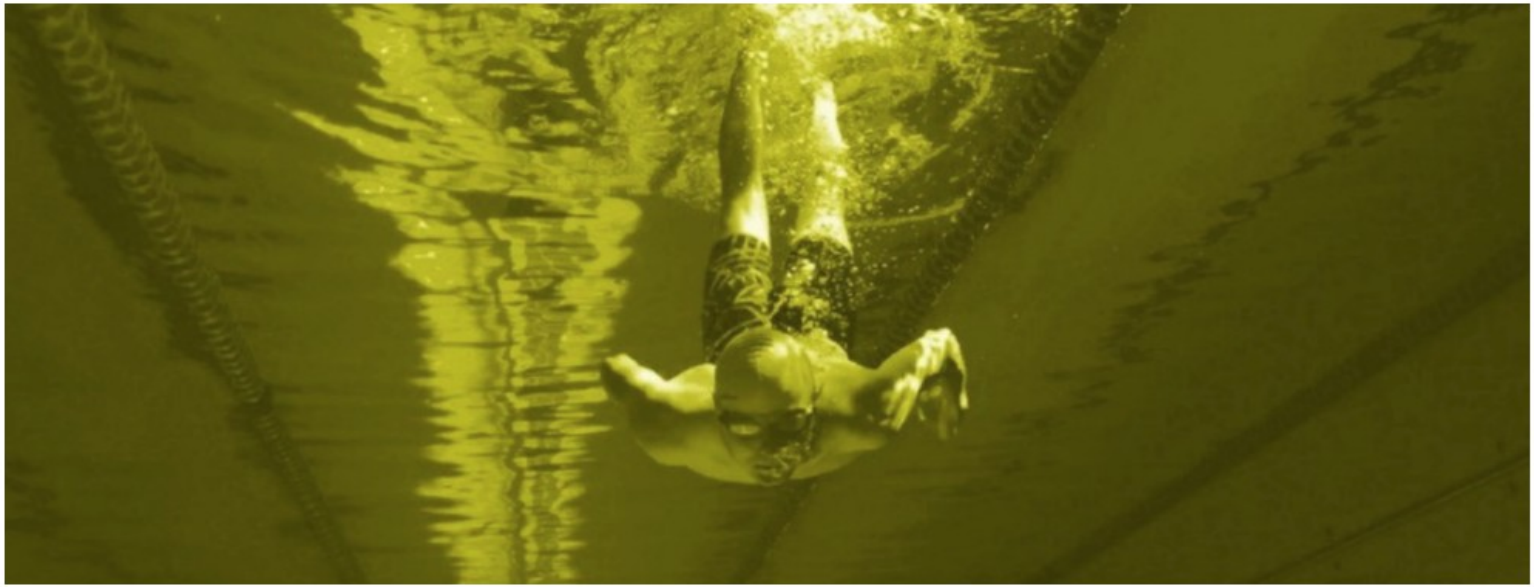
- Basic **data pre-processing**
- Trains a **Random grid of algorithms**
- Individual models are tuned using **cross-validation**.
- Two **Stacked Ensembles** are trained ("All Models" ensemble & a lightweight "Best of Family" ensemble)
- Returns a sorted "**Leaderboard**" of all models.

H2O AutoML Tutorial

October 16th, 2019

A Deep Dive into H2O's AutoML

RSS Share **Category:** AutoML, H2O, Technical



https://www.h2o.ai/blog/a-deep-dive-into-h2os-automl/

By: Parul Pandey

Learn H2O AutoML!

- Docs: https://tinyurl.com/h2o-automl-docs
- Tutorials: https://tinyurl.com/h2o-automl-tutorials
- Blog: A Deep dive into H2O's AutoML

Thank you!


pandeyparul


parulpandeyindia


Medium : @parulnith


WiMLDS Hyderabad : http://wimlds.org/about-the-hyderabad-team/