Overview of Machine Learning and H2O.ai

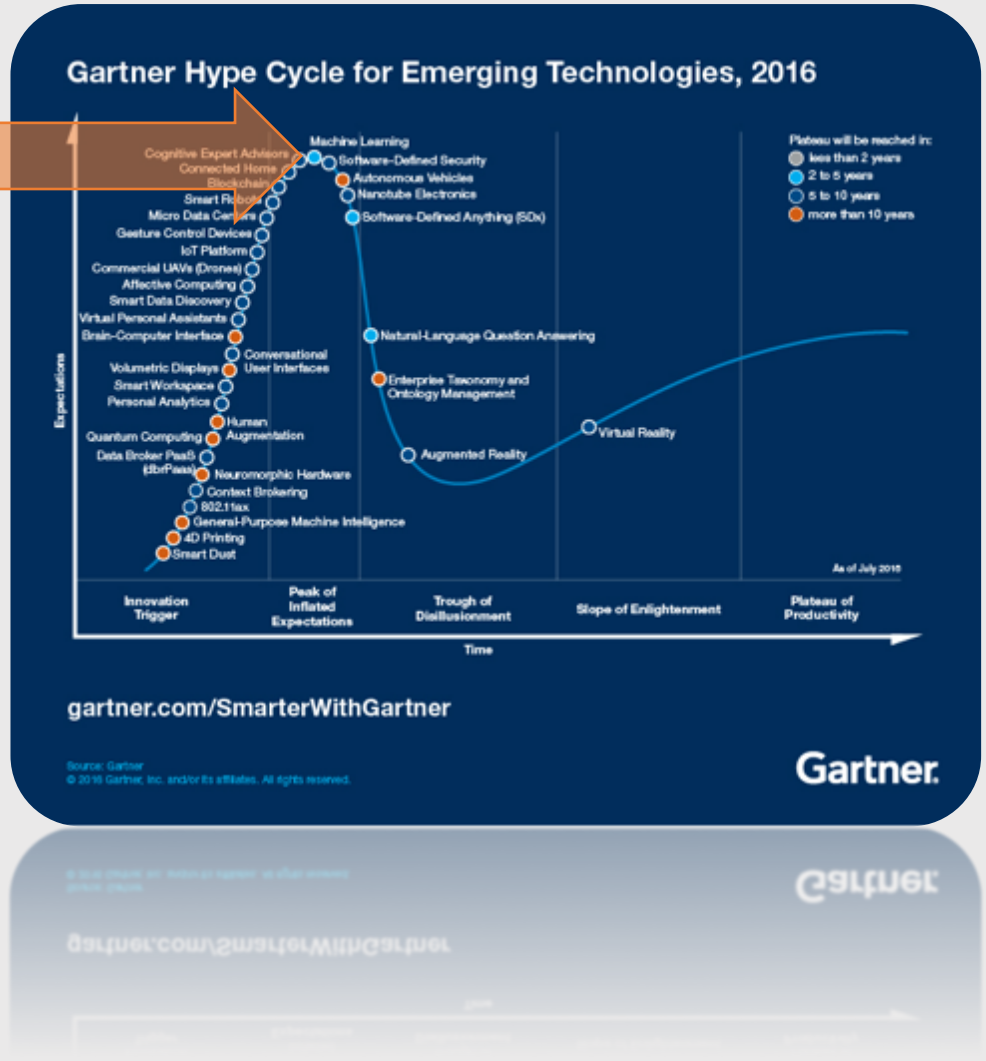
Machine Learning Overview

What is machine learning?

A field of study that gives computers the ability to learn without being explicitly programmed.

-- Arthur Samuel, 1959

H2O.ai



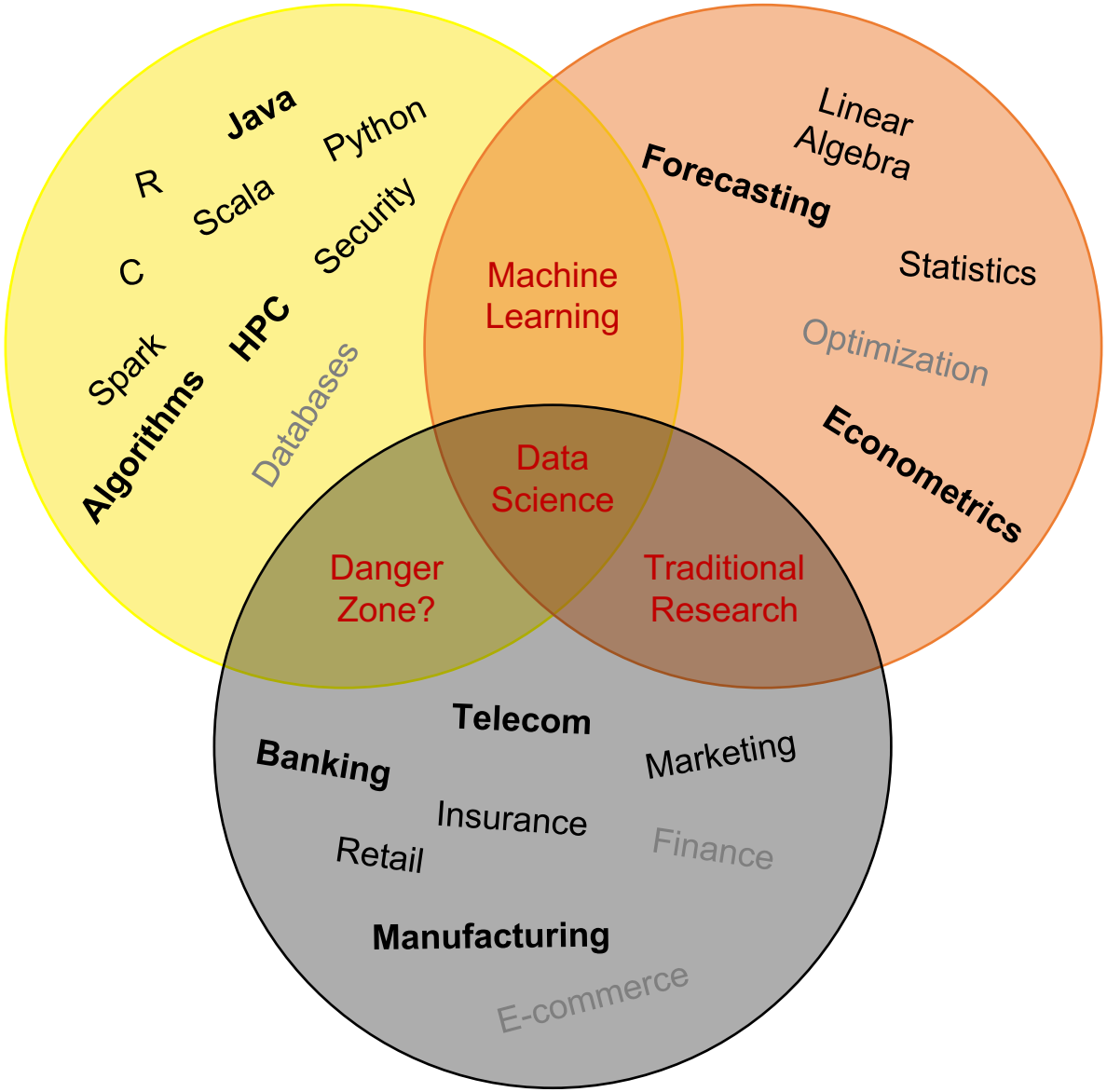
Why now?

- Data, computers, and algorithms are commodities
- Unstructured data
- Increasing competition in business

H2O.ai

Estimating a model for inference	Training a model for prediction
What happened? Why?	What will happen?
Assumptions, parsimony, interpretation	Predictive accuracy, production deployment
Linear models, statistics	Machine learning
Models tend to be static	Many models can evolve elegantly

H2O.ai


Java
Python
R
Scala
Security
C
HPC
Spark
Algorithms
Databases
Linear
Algebra
Forecasting
Statistics
Optimization
Econometrics
Machine
Learning
Data
Science
Danger
Zone?
Traditional
Research
Telecom
Banking
Marketing
Insurance
Retail
Finance
Manufacturing
E-commerce

H2O.ai

1. There is no perfect language.

If someone claims to have the perfect programming language, he is either a fool or a salesman or both.

-- Bjarne Stroustrup

2. There is no perfect algorithm.

Algorithms that search for an extremum of a cost function perform exactly the same when averaged over all possible cost functions.

-- D.H. Wolpert

3. Doing things right is always hard.

Developing and deploying ML systems is relatively fast and cheap, but maintaining them over time is difficult and expensive.

-- Google,
Hidden Technical Debt in Machine Learning Systems

H_2O.ai Overview

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	• **H2O**: In-Memory AI Prediction Engine • **Sparkling Water**: Spark Integration • **Steam**: Deployment engine • **Deep Water**: Deep Learning
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	70 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Headquarters	Mountain View, CA




H2O.ai

H2O.ai Offers AI Open Source Platform

Product Suite to Operationalize Data Science

100% Open Source

In-Memory, Distributed Machine Learning Algorithms with Speed and Accuracy

State-of-the-art Deep Learning on GPUs with TensorFlow, MXNet or Caffe with the ease of use of H2O

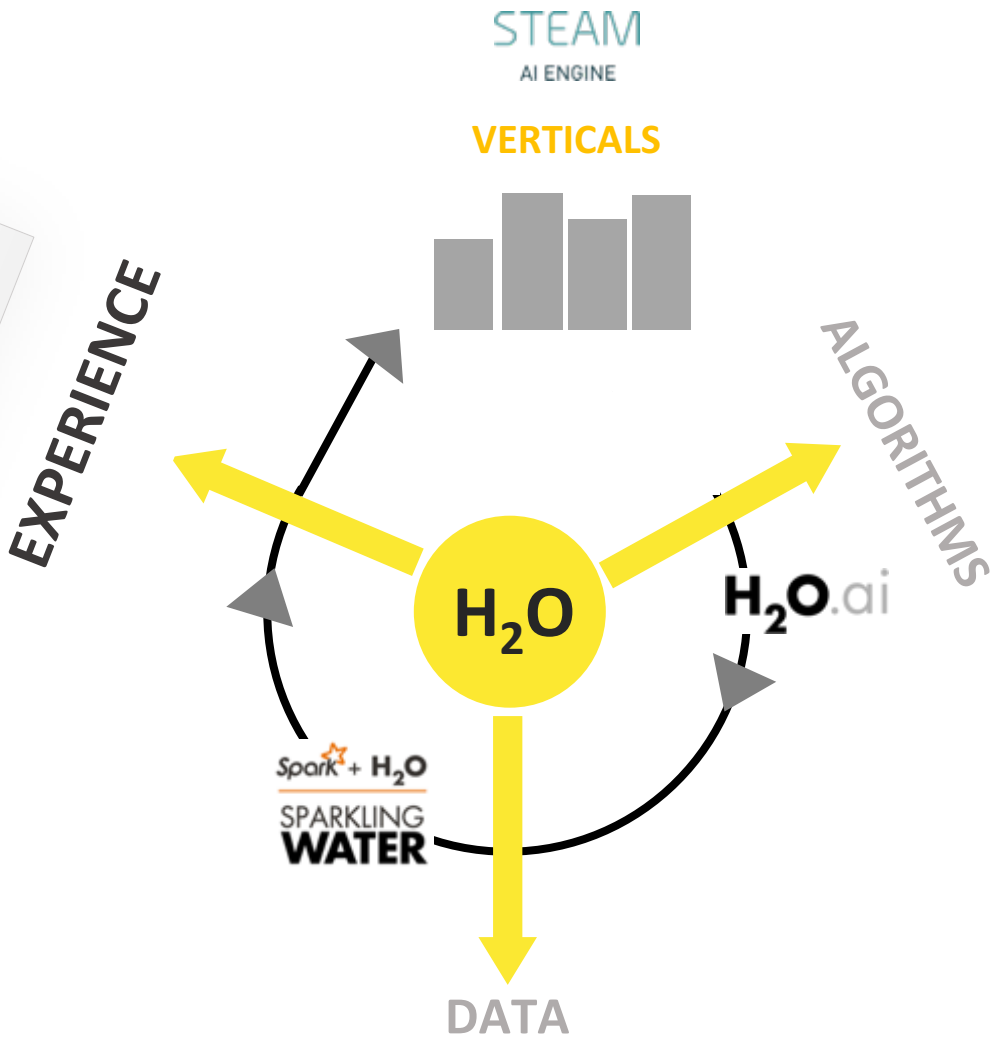
H2O Integration with Spark. Best Machine Learning on Spark.

Operationalize and Streamline Model Building, Training and Deployment Automatically and Elastically

H_2O.ai Now Focused On **Experience**

Beyond Algorithms and Data

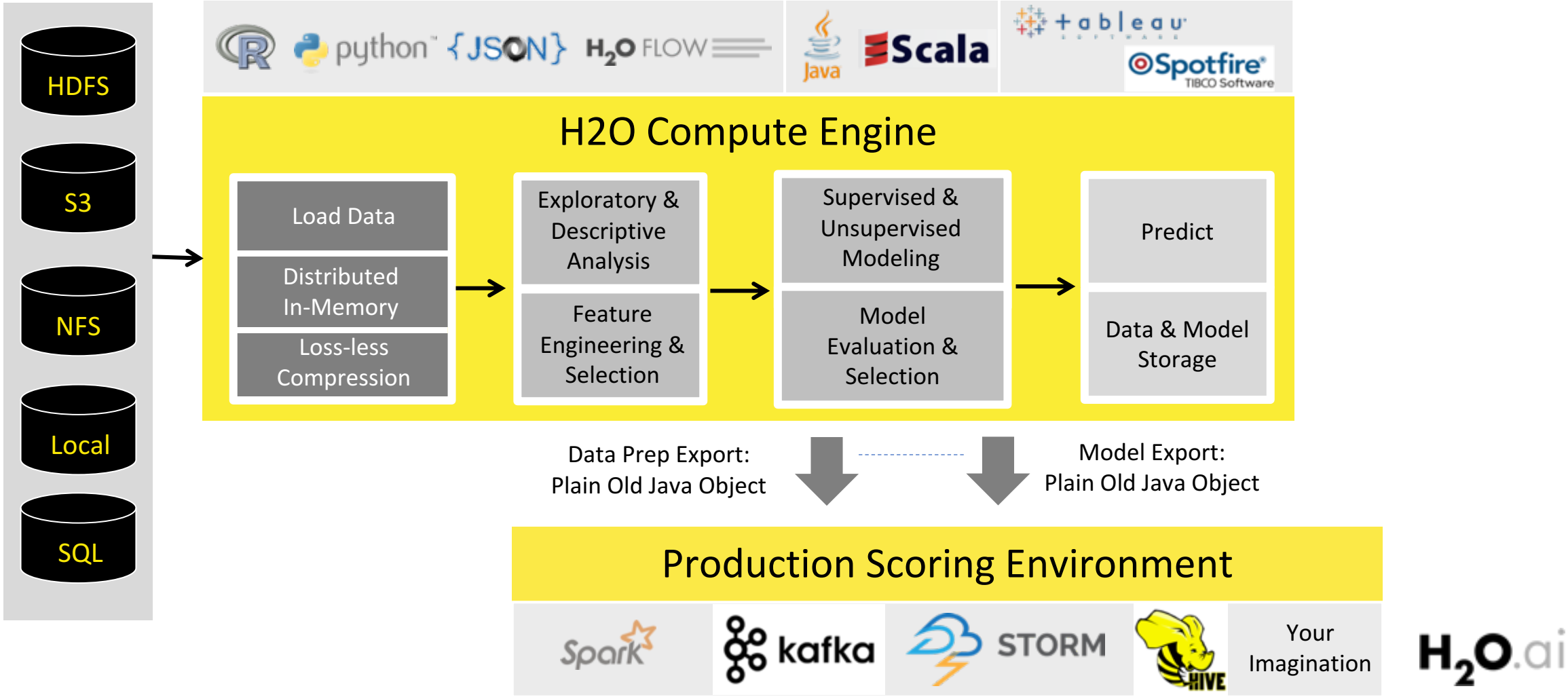
- **H_2O Flow**
  Single web-based Document for code execution, text, mathematics, plots and rich media
- **R, Python, Spark APIs**
  Advanced, scalable ML in the language of your choice
- **H_2O Steam**
  Elastic ML & Auto ML
  Operationalize Data Science



Deep Water

H_2O.ai

High Level Architecture



Intro to Machine Learning
Algos

Algorithms on H_2O

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

Stacking

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

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

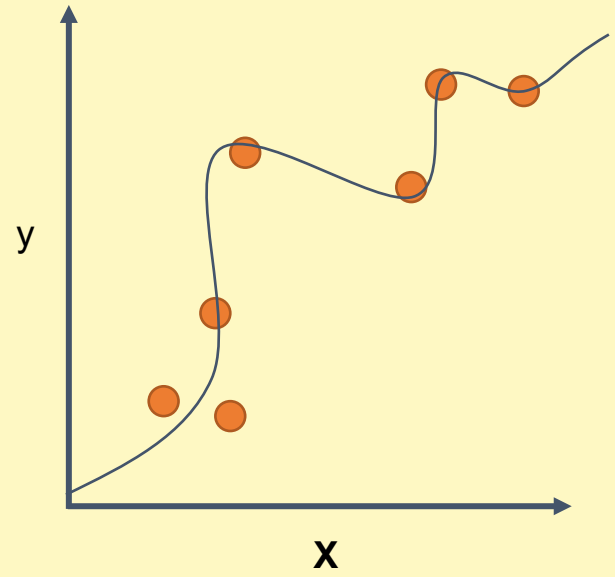
Anomaly Detection

Term Embeddings

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

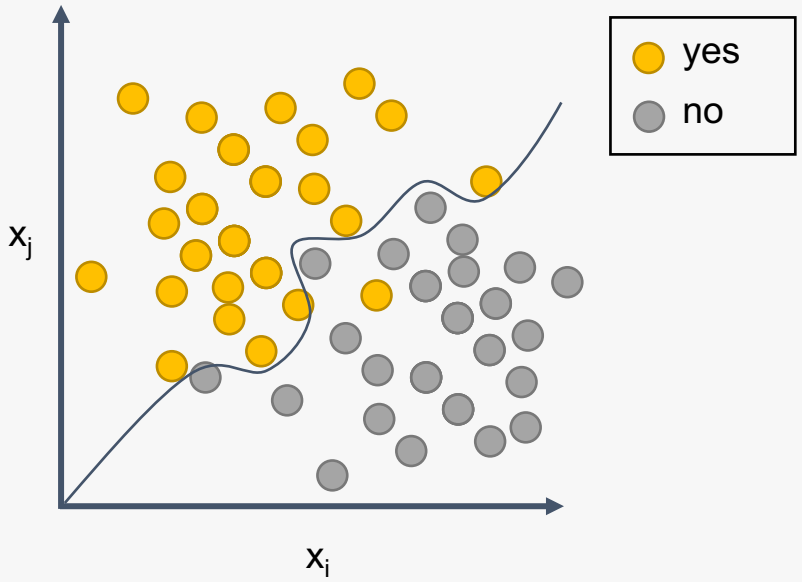
Supervised Learning

Regression:
How much will a customers spend?



H_2O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:
Will a customer make a purchase? Yes or No

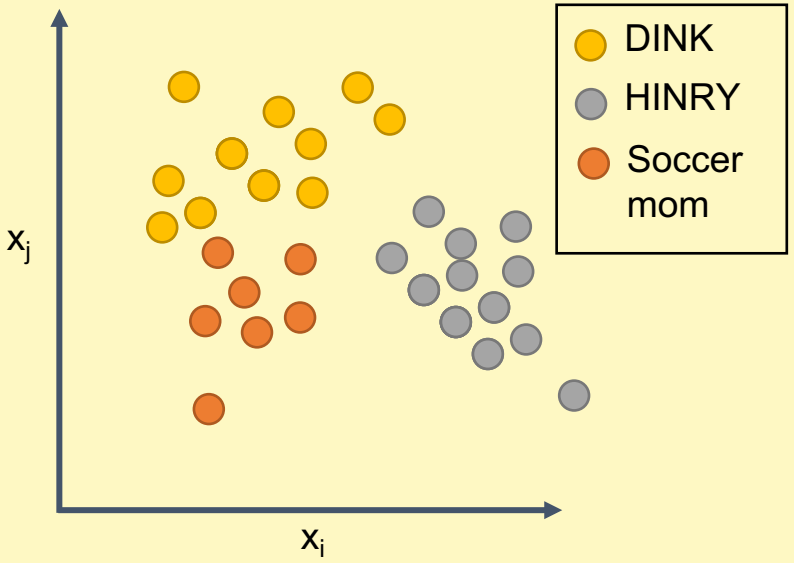


H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Unsupervised Learning

Clustering:

Grouping rows – e.g. creating groups of similar customers

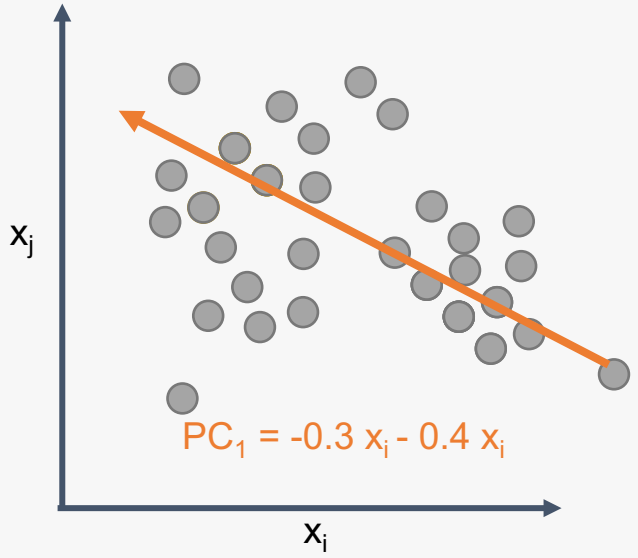


H_2O algos:

k – means

Feature extraction:

Grouping columns – Create a small number of new representative dimensions

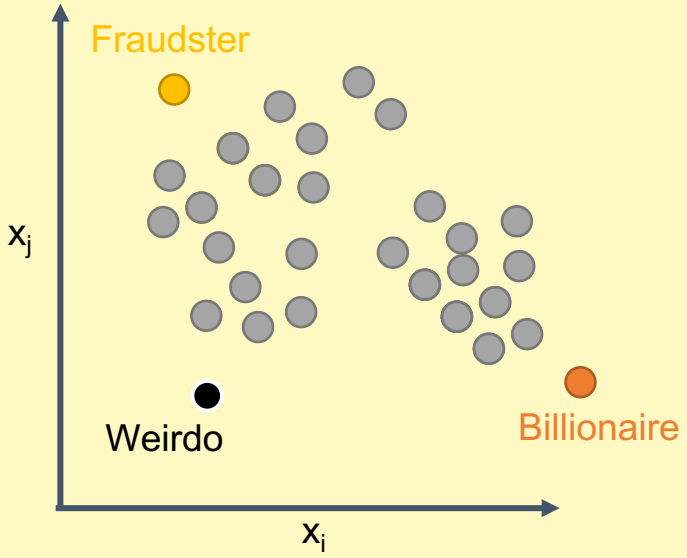


H_2O algos:

Principal components
Generalized low rank models
Autoencoders
Word2Vec

Anomaly detection:

Detecting outlying rows - Finding high-value, fraudulent, or weird customers



H_2O algos:

Principal components
Generalized low rank models
Autoencoders

H_2O.ai

H2O.ai

	Usage	Recommendations	Problems
Penalized Linear Models	• Regression • Classification	• Creates interpretable models with super-fast training time • Nonlinear and interaction terms to be specified manually • Can extrapolate beyond training data domain • Select the correct target distribution • Few hyperparameters to tune	• NAs • Outliers/influential points • Strongly correlated inputs • Rare categorical levels in new data
Naïve Bayes	• Classification	• Nonlinear and interaction terms should be specified by users	• Linear independence assumption • Often less accurate than more sophisticated classifiers • Rare categorical levels in new data
Random Forest	• Regression • Classification	• Builds accurate models without overfitting • Few hyperparameters to tune • Requires less data prep • Great for implicitly modeling interactions	• Difficulty extrapolating beyond training data domain • Can be difficult to interpret • Rare categorical levels in new data
Gradient Boosting Machines	• Regression • Classification	• Builds accurate models without overfitting (often more accurate than random forest) • Requires less data prep • Great for implicitly modeling interactions	• Many hyperparameters • Difficulty extrapolating beyond training data domain • Can be difficult to interpret • Rare categorical levels in new data
Neural Networks (Deep learning & MLP)	• Regression • Classification	• Great for modeling interactions in fully connected topologies • Can extrapolate beyond training data domain • Deep learning architectures best-suited for pattern recognition in images, videos, and sound	• NAs • Overfitting • Outliers/influential points • Long training times • Difficult to interpret • Many hyperparameters • Strongly correlated inputs • Rare categorical levels in new data

	Usage	Recommendations	Problems
k **- means**	• Clustering	• Great for creating Gaussian, non-overlapping, roughly equally sized clusters • The number of clusters can be unknown	• NAs • Outliers/influential points • Strongly correlated inputs • Cluster labels sensitive to initialization • Curse of dimensionality
Principal Components Analysis	• Feature extraction • Dimension reduction • Anomaly detection	• Great for extracting a number <= ***N*** of linear, orthogonal features from i.i.d. numeric data • Great for plotting extracted features in a reduced-dimensional space to analyze data structure, e.g. clusters, hierarchy, sparsity, outliers	• NAs • Outliers/influential points • Categorical inputs
Generalized Low Rank Models	• Feature extraction • Dimension reduction • Anomaly detection • Matrix completion	• Great for extracting linear features from mixed data • Great for plotting extracted features in a reduced-dimensional space to analyze data structure, e.g. clusters, hierarchy, sparsity, outliers • Great for imputing NAs	• Outliers/influential points
Autoencoders (Neural Networks)	• Feature extraction • Dimension reduction • Anomaly detection	• Great for extracting a number of nonlinear features from mixed data • Great for plotting extracted features in a reduced dimensional space to analyze structure, e.g. clusters, hierarchy, sparsity, outliers	• NAs • Overtraining • Outliers/influential points • Long training times • Many hyperparameters • Strongly correlated inputs • Rare categorical levels in new data
Word2Vec	• Highly representative feature extraction from text	• Great for extracting highly representative, context sensitive term embeddings (e.g. numerical vectors) from text • Great for text preprocessing prior to further supervised or unsupervised analysis	• Many Hyperparameters • Overtraining • Specifying term weightings prior to training • Long training times