Extreme Gradient Boosting

H2O POWERED MACHINE LEARNING

FILIP WÓJCIK

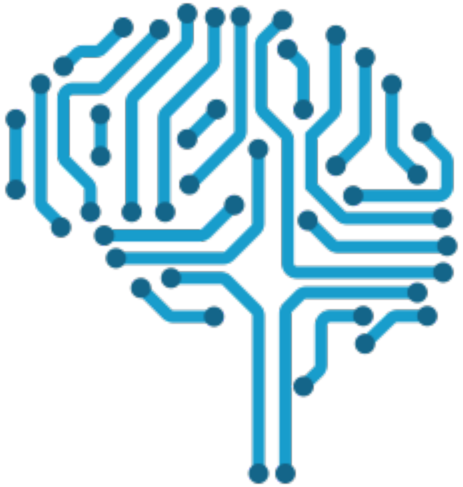
SENIOR DATA SCIENTIST, UNIVERSITY LECTURER

FILP.WOJCIK@OUTLOOK.COM

Agenda

1. Machine learning madness
   - Core machine learning technologies
   - Machine learning overview
2. Is deep learning an ultimate algorithm (hint: no!)?
   - Neural networks vs the rest of the world
   - Top machine learning algorithms
3. Xgboost overivew
   - Decision trees primer
   - Boosted trees - powerful extension to decision trees
4. H2O framework – scalable machine learning tool
   - H2O architecture
   - H2O and big data
   - DEMO!

Machine learning madness



Machine learning madness

- Data science & machine learning are gaining popularity in recent years
- Rapid boost of data science & analytical software
- Big data influence – larger volumes of data can be processed now, without any problems
- Machine learning everywhere!

Hi. I'm Cortana.

Siri

ASSISTANT

TESLA

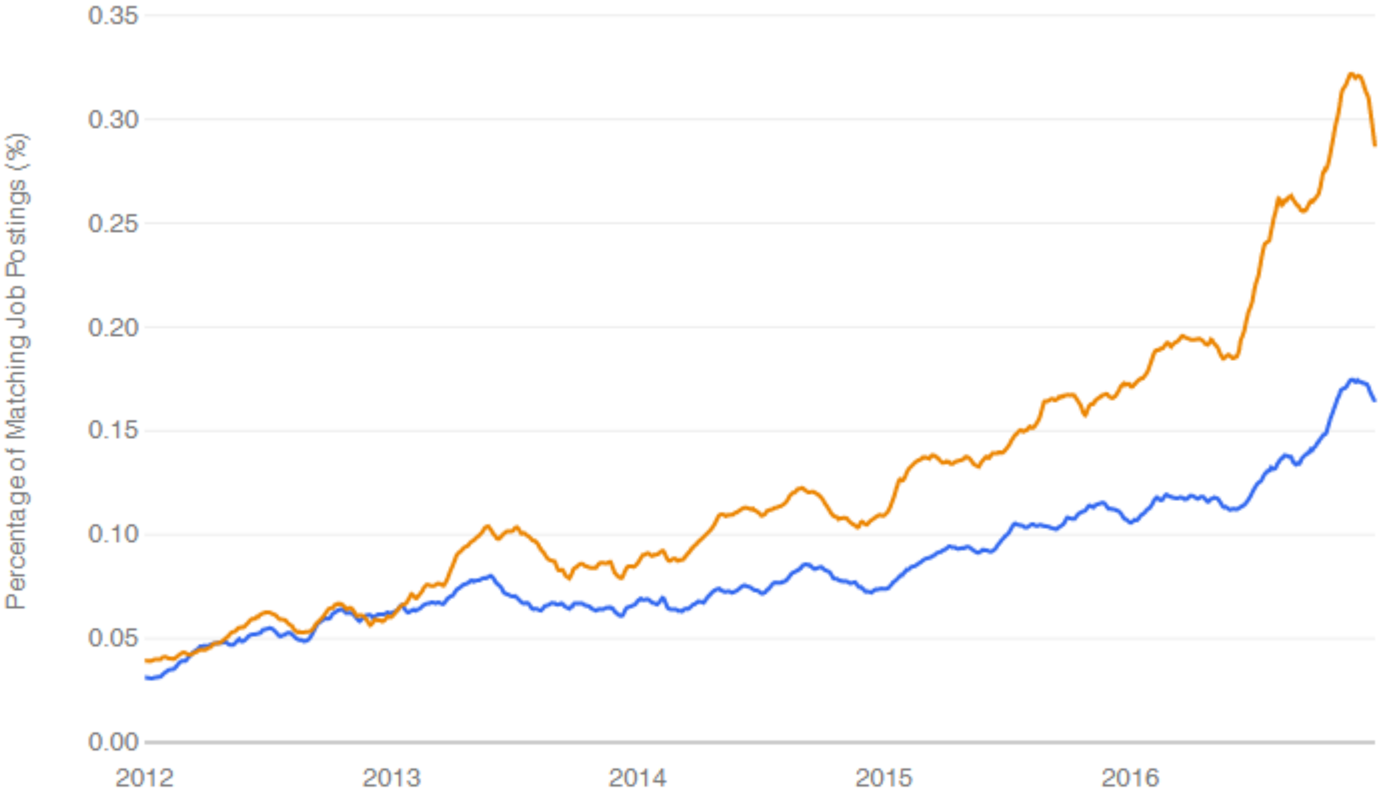
NETFLIX

Machine learning madness

Machine learning job positions count

R jobs

Python jobs

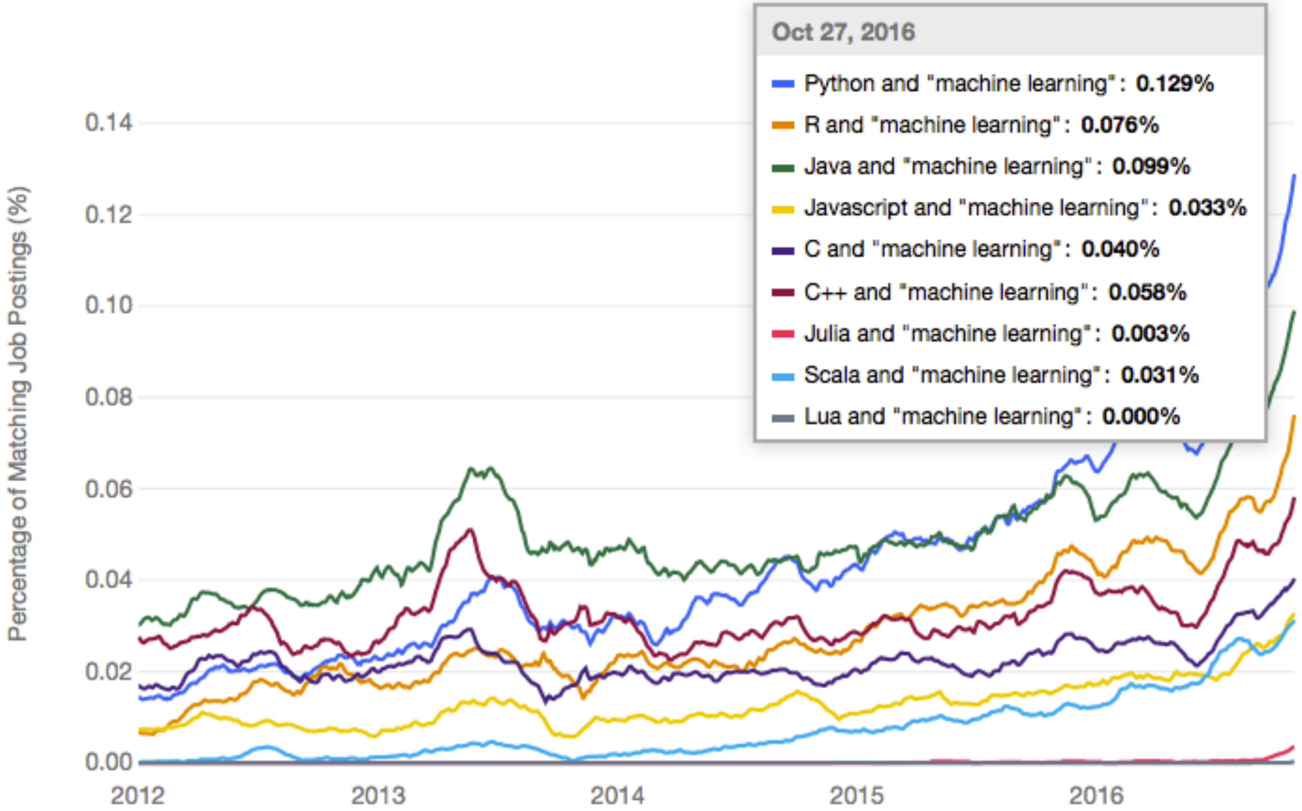


Source: r4stat, The Popularity of Data Science Software by Robert A. Muenchen, http://r4stats.com/articles/popularity/

Machine learning madness

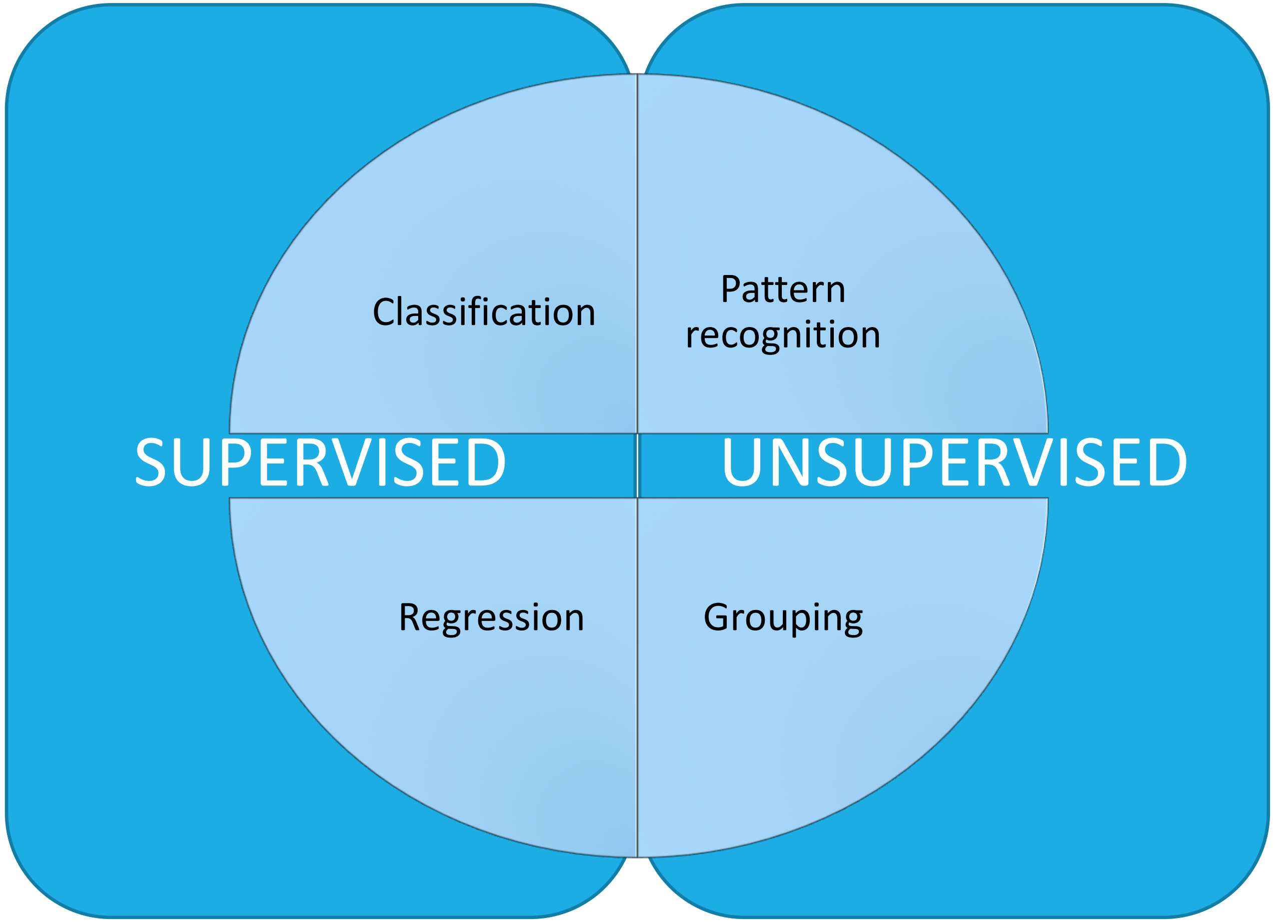
Machine learning core technologies

R
Spark
H_2O.ai

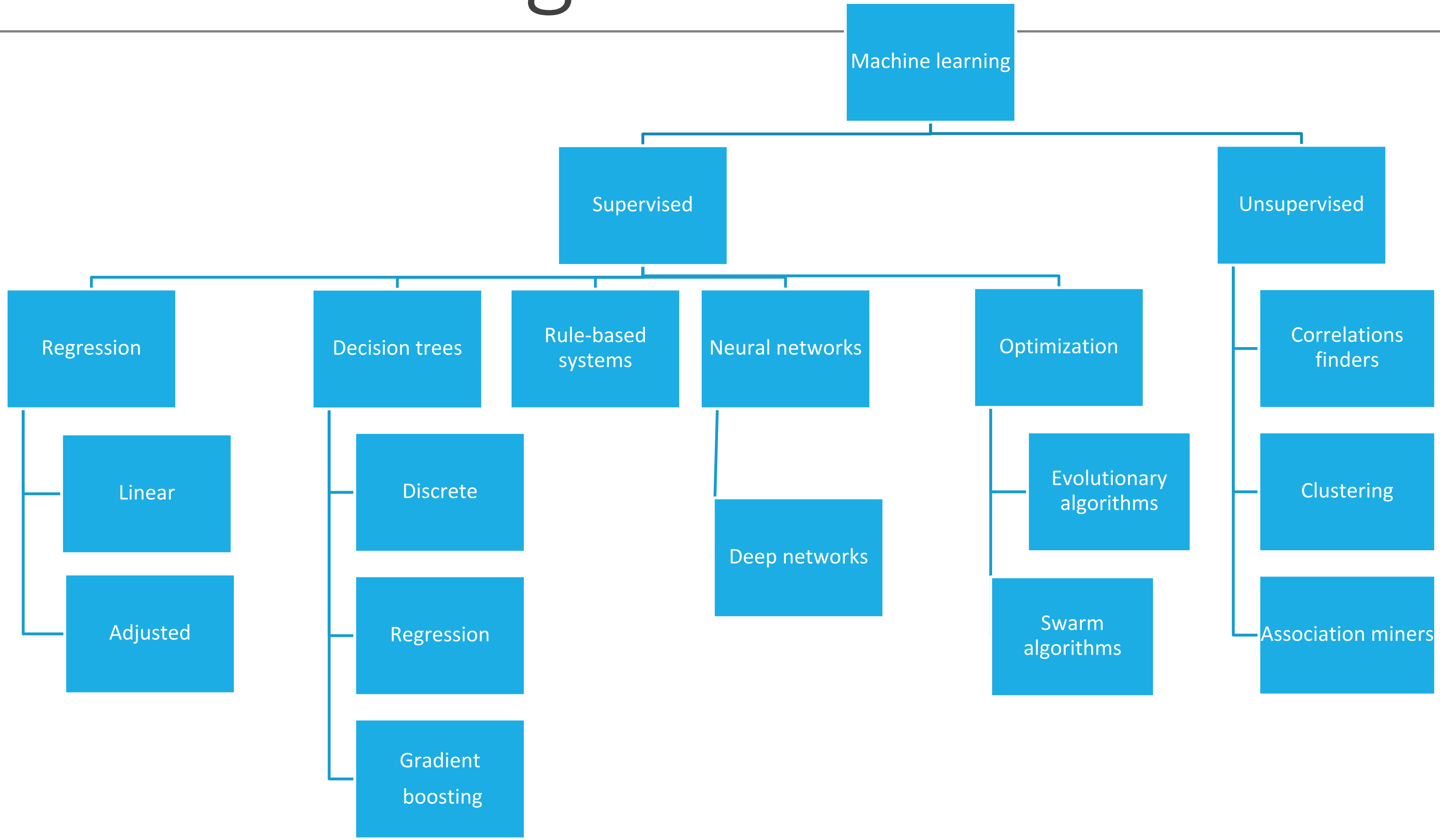


Source: KDNuggets, Jean-Francois Puget, 2017 The Most Popular Language For Machine Learning and Data Science Is…, http://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html

Machine learning madness



Machine learning madness



Machine learning

Supervised

Unsupervised

Regression

Linear

Adjusted

Decision trees

Discrete

Regression

Gradient boosting

Rule-based systems

Neural networks

Deep networks

Optimization

Evolutionary algorithms

Swarm algorithms

Correlations finders

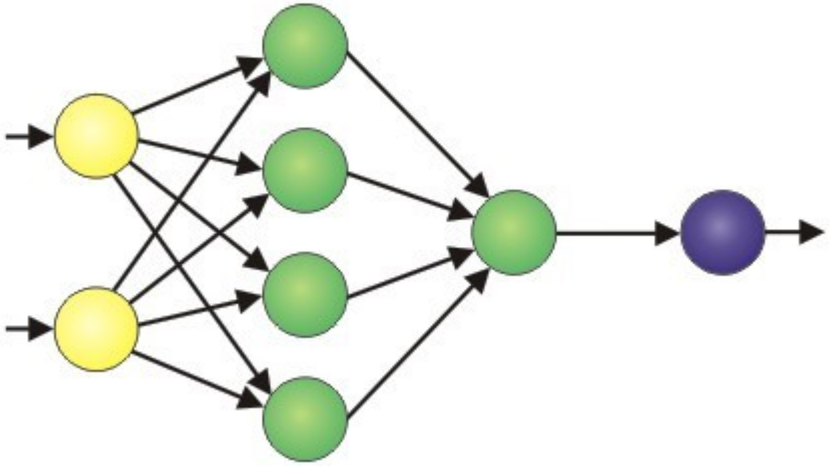
Clustering

Association miners

ACHIEVEMENT UNLOCKED

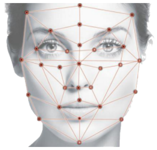
Ready for anything.

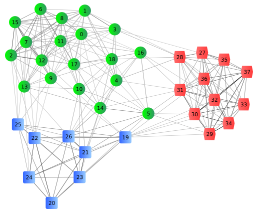
Is deep learning an ultimate algorithm?

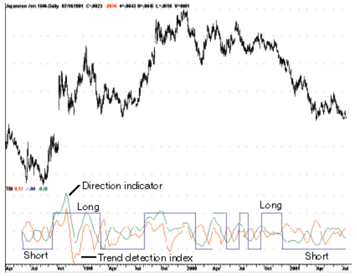


Is deep learning an ultimate algorithm?

- Deep learning seems to be the most hyped ML algorithm nowadays
- Many analysts try to use it to solve ALL kinds of problems
- In reality deep learning is great tool for:
  - ✓ Pattern recognition in images
  - ✓ Motion detection
  - ✓ Sentiment analysis
  - ✓ Classification in fuzzy contexts
  - ✓ Trend identification







Is deep learning an ultimate algorithm?

No free lunch theorem!

No learning algorithm has an inherent superiority over other learning algorithms for all problems.

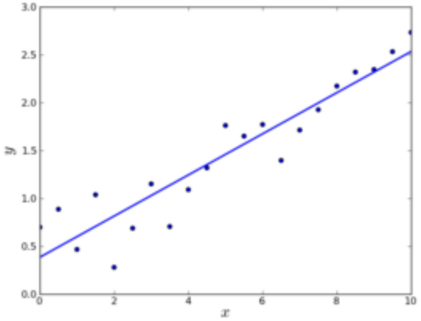
Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization”

Is deep learning an ultimate algorithm?

Top multi-purpose machine learning algorithms

1. Linear regression

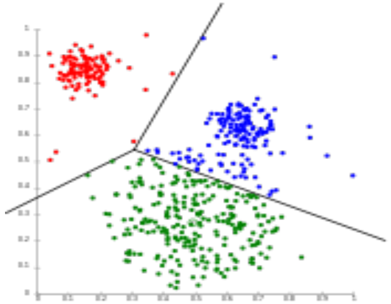
2. Logistic regression



3. Decision trees

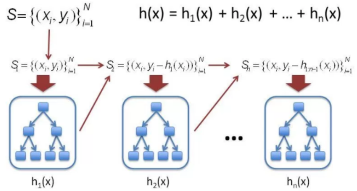
4. Random forests

5. K-means clustering



6. Support Vector Machines

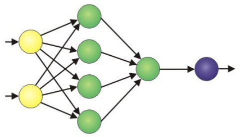
7. Gradient tree boosting


$S = \{(x_i, y_i)\}_{i=1}^{N}$
$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$
$S_1 = \{(x_i, y_i)\}_{i=1}^{N}$
$S_2 = \{(x_i, y_i - h_1(x_i))\}_{i=1}^{N}$
$S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^{N}$
…
$h_1(x)$
$h_2(x)$
$h_n(x)$

8. Naive Bayes

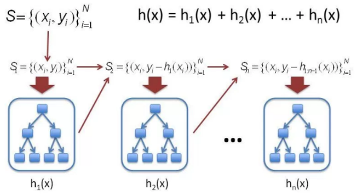
$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$$

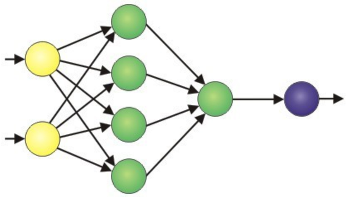
9. Neural networks



Source:
UpX, Namartha Peddi, 10 most popular machine learning algorithms, https://upxacademy.com/10-popular-machine-learning-algorithms/
Data Science Central, Top 10 most popular machine learning algorithms, http://www.datasciencecentral.com/profiles/blogs/top-10-machine-learning-algorithms

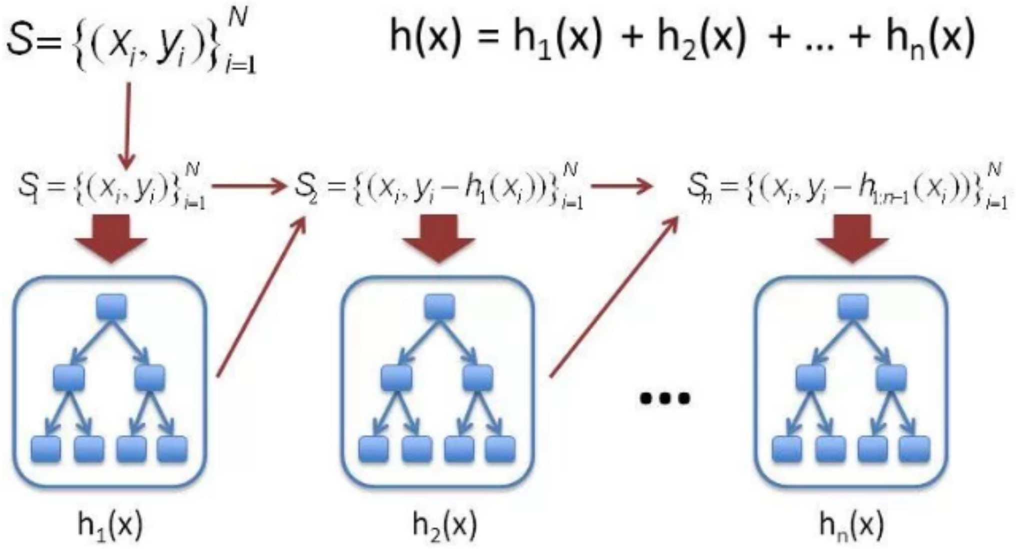
Is deep learning an ultimate algorithm?


$S = \{(x_i, y_i)\}_{i=1}^{N}$
$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$
$S_1 = \{(x_i, y_i)\}_{i=1}^{N}$
$S_2 = \{(x_i, y_i - h_1(x_i))\}_{i=1}^{N}$
$S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^{N}$
…
$h_1(x)$
$h_2(x)$
$h_n(x)$



	Xgboost	Deep learning / ANN
Building complexity	Fast to design & train	Require carefull tuning
Data types	Tabular/structured data	Unstructured data/pictures/speech/etc.
Volumes of data	High to moderate	Small to INSANE
Mathematical explanation	Gradient function	Black - box

Xgboost – an overview


$S=\{(x_i, y_i)\}_{i=1}^{N}$
$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$
$S_1=\{(x_i, y_i)\}_{i=1}^{N}$
$S_2=\{(x_i, y_i - h_1(x_i))\}_{i=1}^{N}$
$S_n=\{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^{N}$
...
$h_1(x)$
$h_2(x)$
$h_n(x)$

Xgboost – an overview

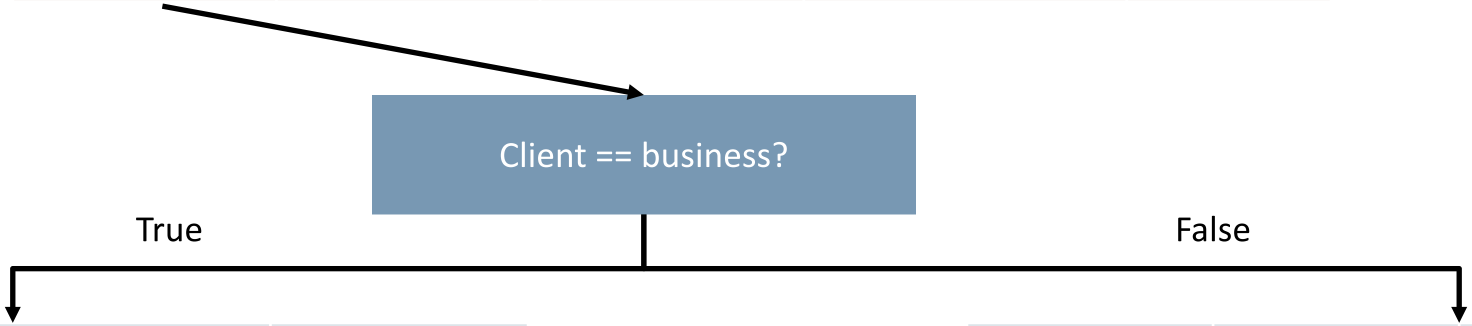
- Based on decision trees & random forests
- Using boosting procedure – learning on previous mistakes:
  - Putting more emphasis on wrongly classified examples
  - Training new classifiers one-by-one, with instruction to „watch out” for particular errors
- Can be used to both tasks:
  - Classification
  - Regression
- Both tasks are based on assigning numerical score, which corresponds to decision certaninty level (classification) and/or raw numeric result

Xgboost – an overview

client	hotel	addons	money_spent	offer
business	Hilton	trip	40,000	deluxe
business	Hilton	full board	38,000	deluxe
business	Hilton	trip	40,000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1,500	premium

Value	Count	%
Deluxe	3	0.5
Basic	2	0.333
Premium	1	0.16666

client	hotel	addons	money_spent	offer
business	Hilton	trip	40,000	deluxe
business	Hilton	full board	38,000	deluxe
business	Hilton	trip	40,000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1,500	premium



Client == business?

True

False

hotel	addons	money_spent	offer
Hilton	trip	40,000	deluxe
Hilton	full board	38,000	deluxe
Hilton	trip	40,000	deluxe

hotel	addons	money_spent	offer
Meta	none	800	basic
Meta	meal	900	basic
Meta	spa	1,500	premium

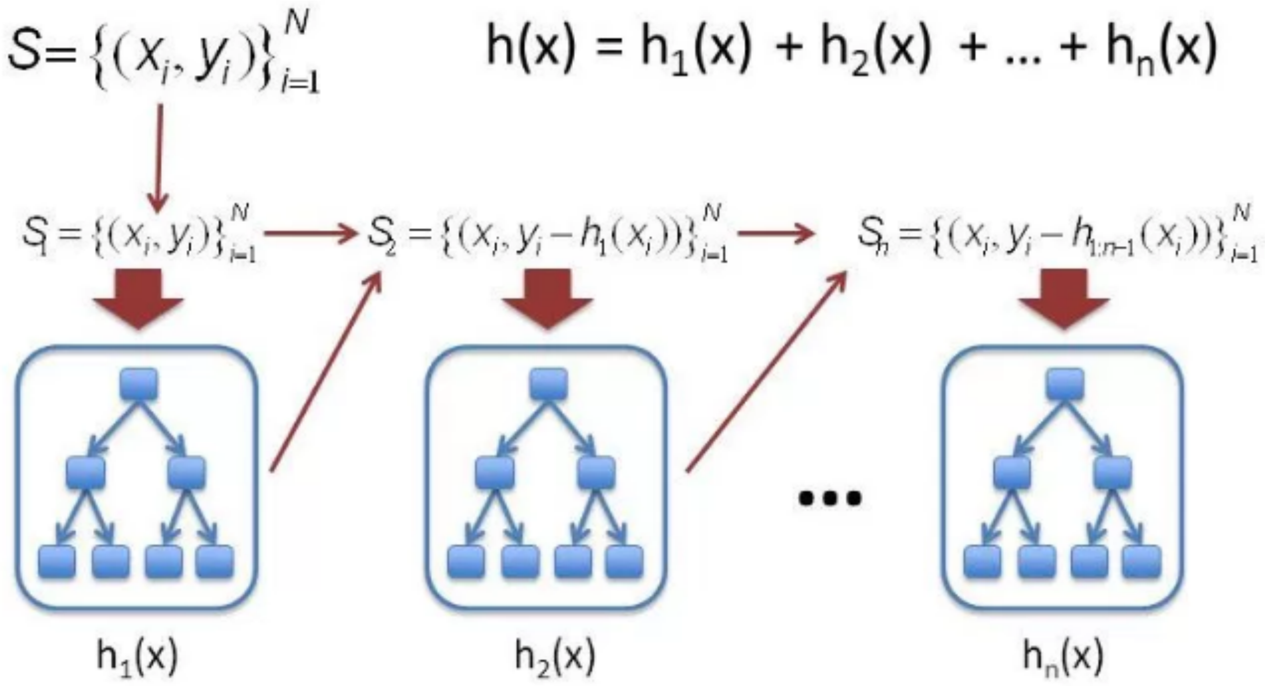
Xgboost – an overview

- Multiple decision trees, each learning from mistakes of its predecessors
- Each tree receives slightly randomized dataset (different columns, resampled rows), to get rid of noise
- Steps:

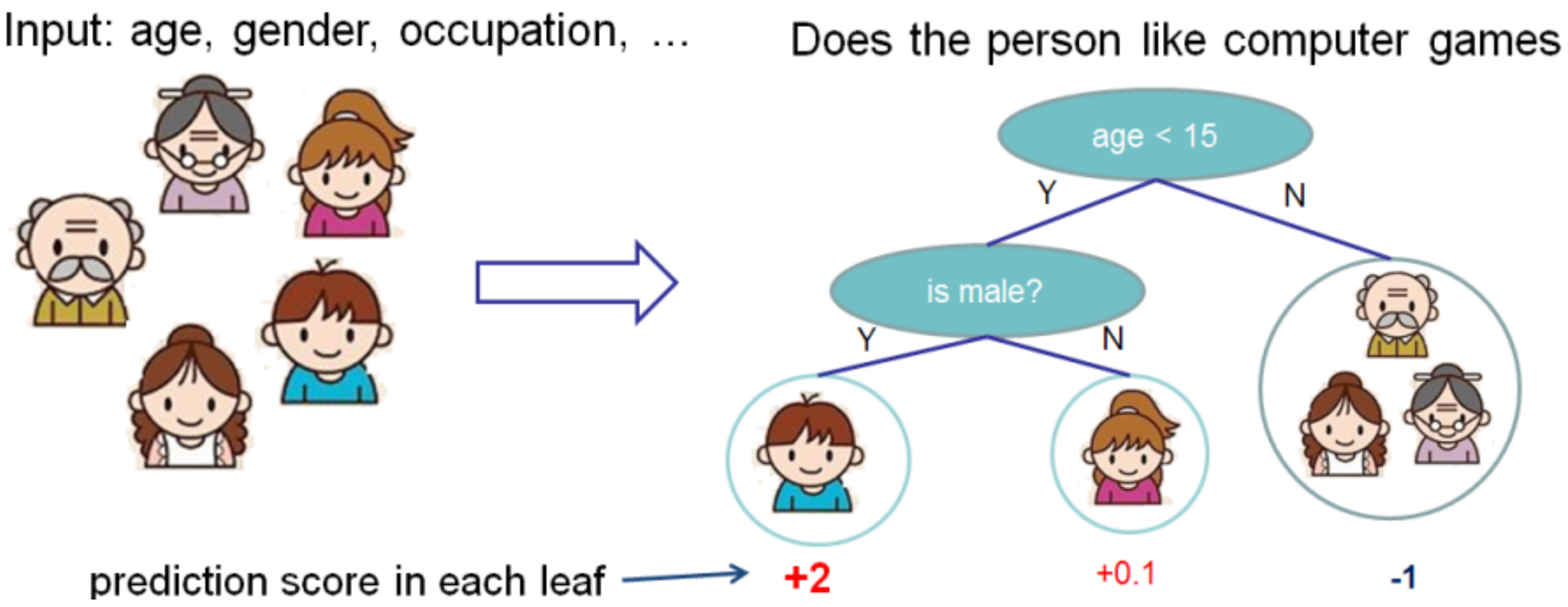
1. Fit initial model - simple prediction, e.g. mean: $F_1(x) = \hat{y}$
2. Calculate error magnitude for each data point: $h_1(x) = y - F_1(x)$
3. Create a new model, which will correct errors of its predecessor: $F_2(x) = F_1(x) + h_1(x)$
4. Continue until error rates are small enough or until reaching tree limit

$$F(x) = F_1(x) \rightarrow F_2(x) \rightarrow \cdots \rightarrow F_n(x)$$

Xgboost – an overview

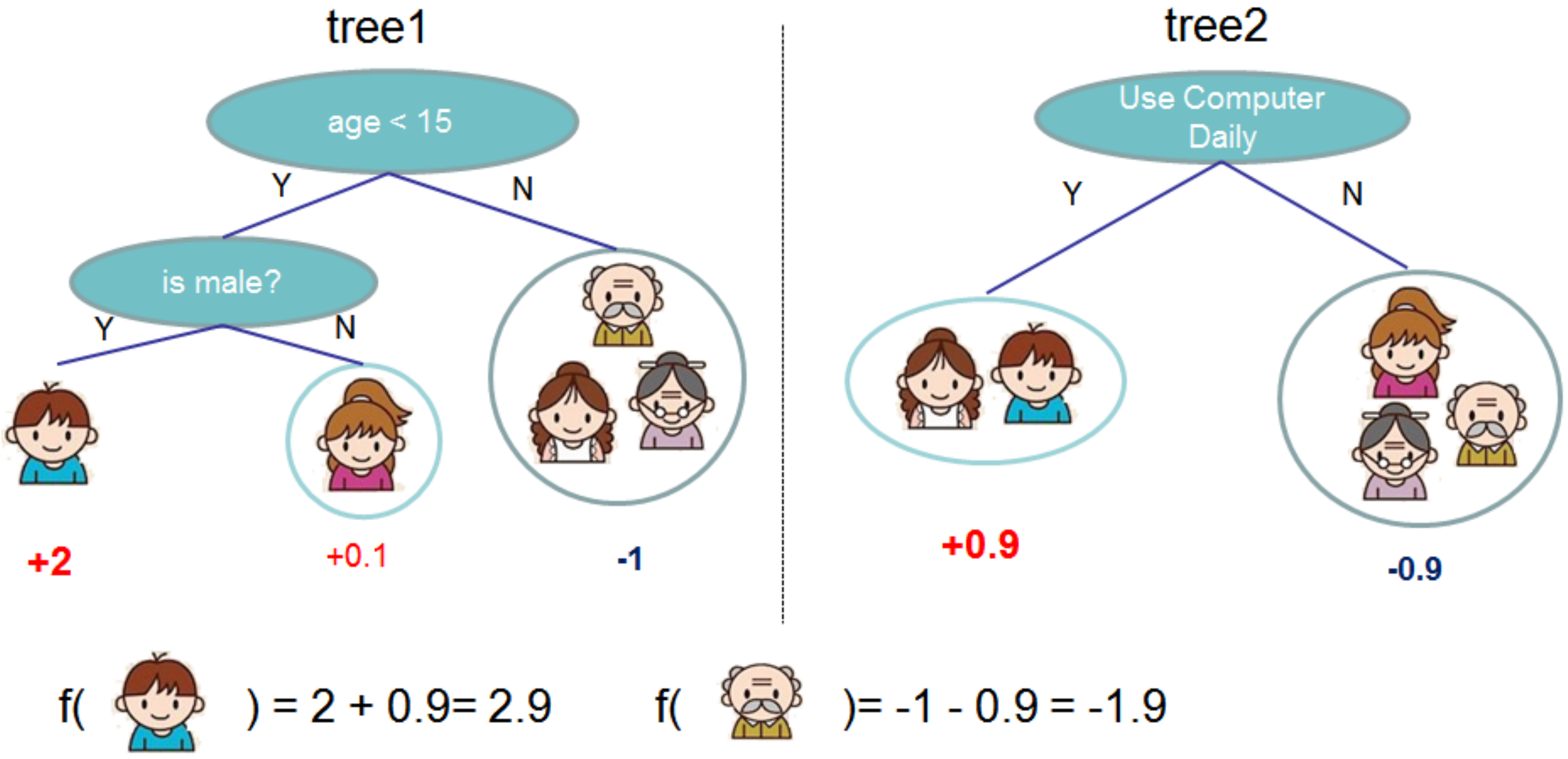

$S = \{(x_i, y_i)\}_{i=1}^{N}$
$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$
$S_1 = \{(x_i, y_i)\}_{i=1}^{N}$
$S_2 = \{(x_i, y_i - h_1(x_i))\}_{i=1}^{N}$
$S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^{N}$
…
$h_1(x)$
$h_2(x)$
$h_n(x)$

Xgboost – an overview


Input: age, gender, occupation, ...
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

Source: official Xgboost documentation. http://xgboost.readthedocs.io/en/latest/model.html

Xgboost – an overview



Source: official Xgboost documentation. http://xgboost.readthedocs.io/en/latest/model.html

Xgboost – an overview

- Additional performance tuning tricks:
  - Using random set of columns for every tree – prevents overfitting
  - Using random set of examples for every tree – prevents overfitting
  - Using different tree depths – prevents overfitting ☺

H2O.AI and XGBoost

H2O.AI and XGBoost

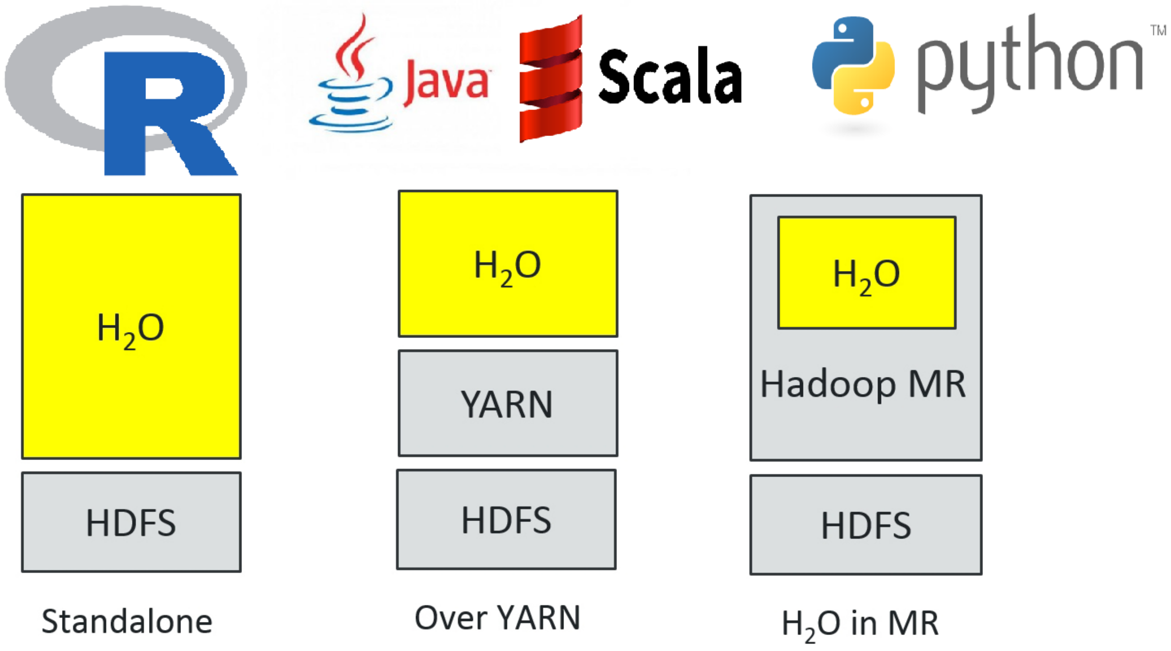
- H2O AI is
  - Open-source
  - Fast
  - Scalable
  - In-memory

  processing engine, equipped with predefined set of machine learning models

- Big-data ready and optimized

  - Special data structures (hex)
  - Highly compressed
  - Lazy operations (like in Apache Spark)
  - Immutable, distributed structures

H2O.AI and XGBoost



H_2O

HDFS

Standalone

H_2O

YARN

HDFS

Over YARN

H_2O

Hadoop MR

HDFS

H_2O in MR

Cluster

••Multi-node clusters
••Data distributed across nodes
••No limit on numer of nodes ☺

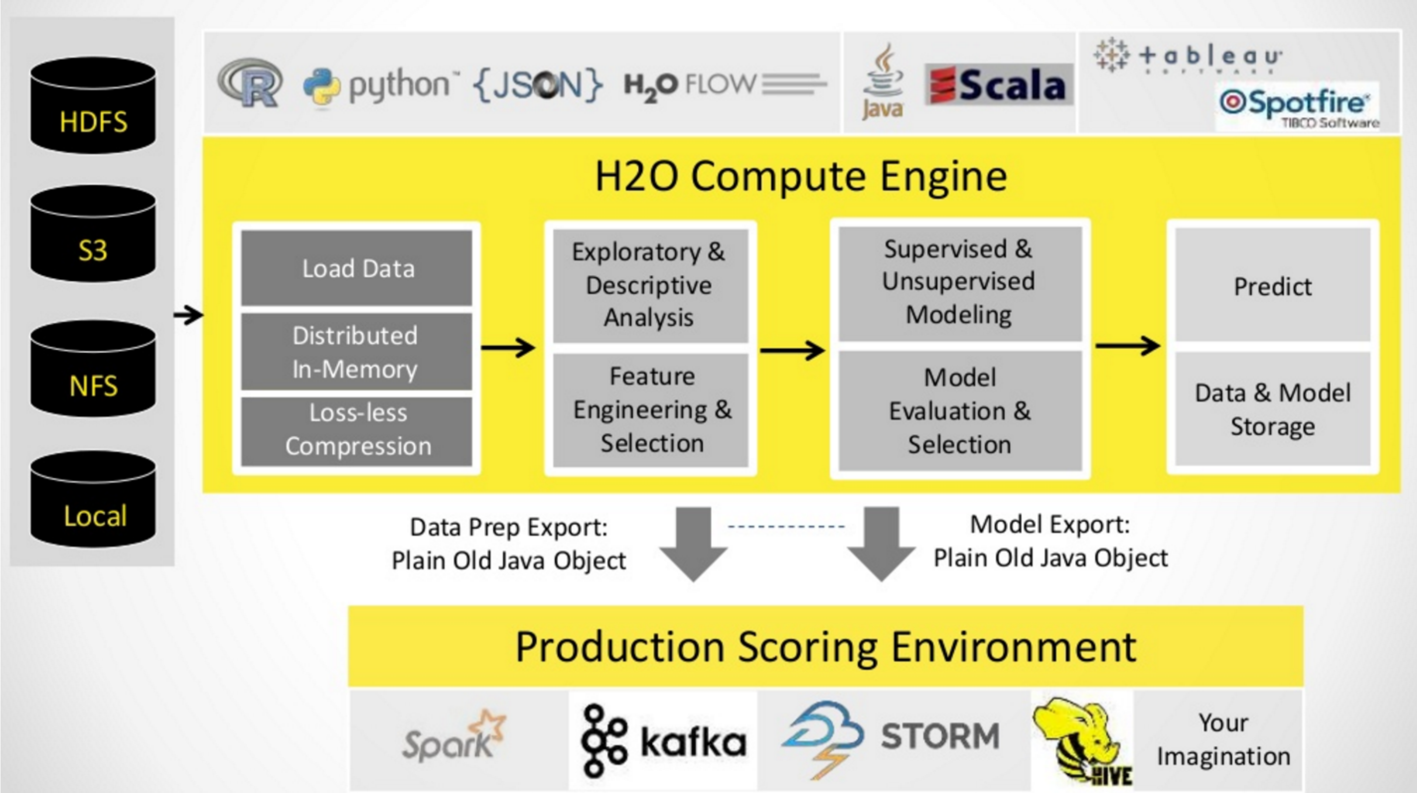
Key-value pairs data

••Datasets referenced by key
••Columnary datastore
••SQL-like structures

Multi-language

••Native implementation in Java
••Interfaces for Python/Scala/R

H2O.AI and XGBoost


HDFS
S3
NFS
Local
python
{JSON}
H2O FLOW
Java
Scala
tableau
Spotfire
TIBCO Software
H2O Compute Engine
Load Data
Distributed In-Memory
Loss-less Compression
Exploratory & Descriptive Analysis
Feature Engineering & Selection
Supervised & Unsupervised Modeling
Model Evaluation & Selection
Predict
Data & Model Storage
Data Prep Export: Plain Old Java Object
Model Export: Plain Old Java Object
Production Scoring Environment
Spark
kafka
STORM
HIVE
Your Imagination

Source: official H2O documentation, http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html

H2O.AI and XGBoost

DEMO!