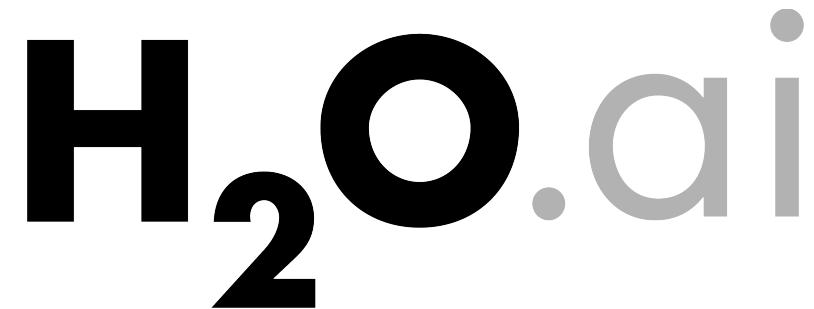


H₂O Training (20th Feb, 2018)

- Agenda
 - Sign in / Setup Equipment
 - Introduction
 - Machine Learning Basics
 - Hands-on (Web/R/Python)
 - Regression
 - Classification
 - Clustering
 - Driverless AI demo
 - Q&A / Open Discussion
- Lunch Break (12:00 – 13:00)



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

@matlabulous

About Me

- Civil (Water) Engineer
 - 2010 – 2015
 - Consultant (UK)
 - Utilities
 - Asset Management
 - Constrained Optimization
 - EngD (Industrial PhD) (UK)
 - Infrastructure Design Optimization
 - Machine Learning + Water Engineering
 - Discovered H₂O in 2014
 - Data Scientist
 - 2015 – 2016
 - Virgin Media (UK)
 - Domino Data Lab (Silicon Valley)
 - 2016 – Present
 - H₂O.ai (Silicon Valley)
 - How?
 - bit.ly/joe_kaggle_story

What is Joe's role at H₂O.ai?



- Data Scientist
(the job title ...)
- Sales Engineer / Conference & Meetup Speaker / Community Manager
(hard truth about tech startup ...)
- Unofficial Photographer of H₂O.ai SWAG
(the travelling data scientist)
- H₂O.ai SWAG EMEA Distributor
(please help yourself)

About H₂O.ai

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	<ul style="list-style-type: none">• H₂O Open Source In-Memory AI Prediction Engine• Sparkling Water (H₂O + Spark)• Enterprise Steam• Driverless AI
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	<p>85+ employees</p> <ul style="list-style-type: none">• Distributed Systems Engineers doing Machine Learning• World-class visualization designers
Headquarters	Mountain View, CA



THE H₂O.ai TEAM



Scientific Advisory Council



Dr. Trevor Hastie

- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*



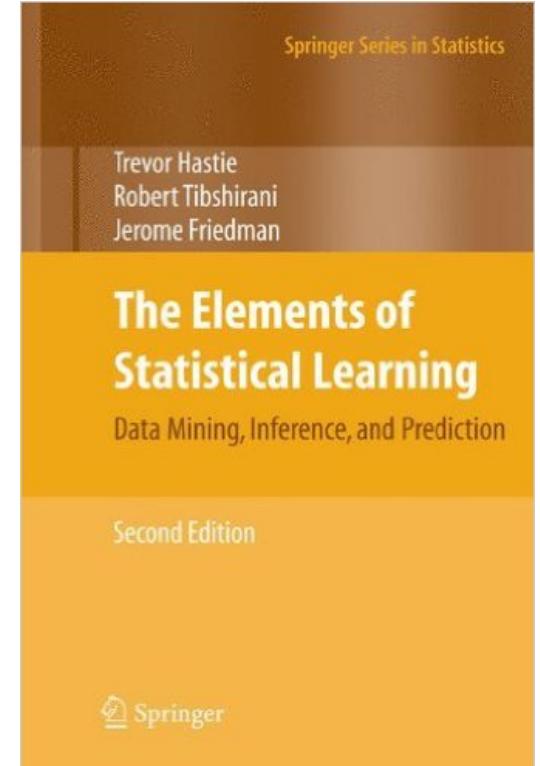
Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*



Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*





wenphan
@wenphan

Following

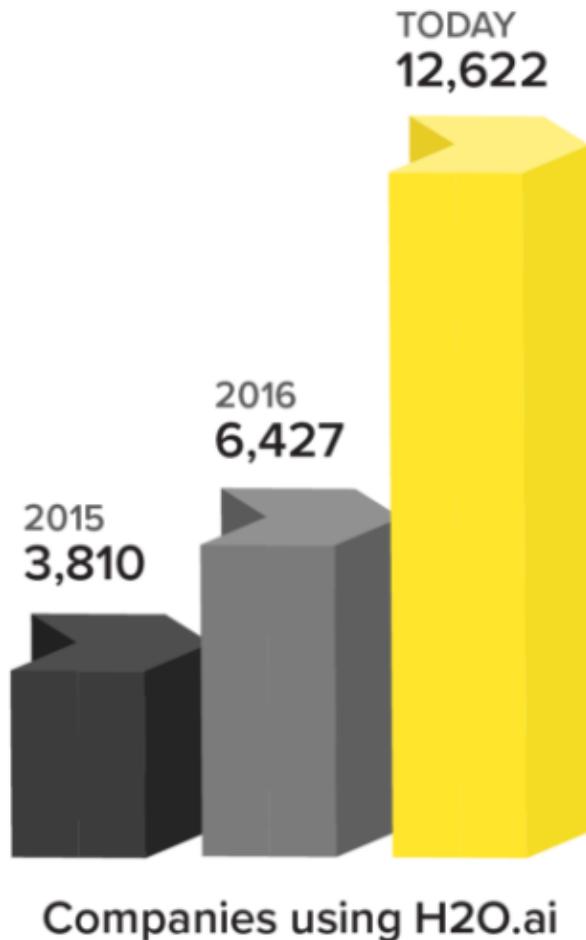


So much brain power in one place:
[@ArnoCandel](#) and Stanford profs. Boyd,
Tibs, and Hastie. Hacking algos at [@h2oai](#)
HQ



Arno (CTO)

12,000+ Companies use H2O — World Wide Community Adoption

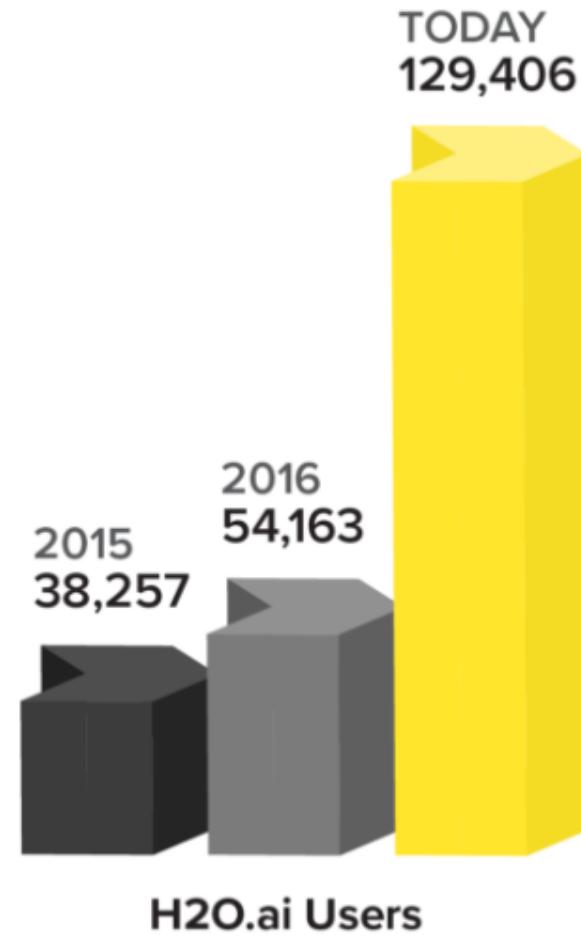


222 FORTUNE
OF THE 500
♥ **H₂O**

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES



H2O.ai Select Paying Customers



“Overall customer satisfaction is very high.” - Gartner

Harnessing the power of AI to transform the detection of fraud and error

Setting the scene

PwC has invested significantly in pioneering the use of artificial intelligence for the audit and has partnered with H2O.ai, a leading Silicon Valley-based AI company.

Following 18 months of development, the first outcome of this partnership is PwC's GL.ai, the first module of PwC's Audit.ai - a revolutionary bot that does what humans can't. Its AI analyses billions of different data points in seconds and applies judgement to detect anomalies in general ledger transactions.



"The reason this is such a brilliant tool is the ability to look at different risks in context at the same time. For example, it would be uneconomical for an auditor to look at every single user's pattern of activity and decide what was unusual. With GL.ai, the algorithms do it for us."

Laura Needham partner, PwC UK



Follow

Exciting night at this year's @WAI_News Awards: PwC wins 2017 Audit Innovation of the Year! pwc.to/Glaia17 #taandiab17



10:15 PM - 4 Oct 2017

Check
out our
website
h2o.ai



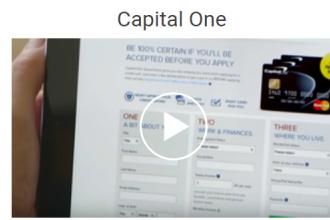
Various data leaders discuss the transformative impact of H2O AI for ADP.



What data products mean and why H2O keeps this industry leader relevant.



See how Progressive uses H2O predictive analytics for User-based Insurance (UBI).



Capital One uses H2O machine learning for various use cases.



H2O predictive analytics helps boost the impact and results of digital marketing.



Kaiser uses H2O machine learning to save lives.



Zurich turned to H2O as a strategic differentiator for commercial insurance.



Comcast uses H2O to improve customer experience.



McKesson discusses the adoption of artificial intelligence in healthcare.



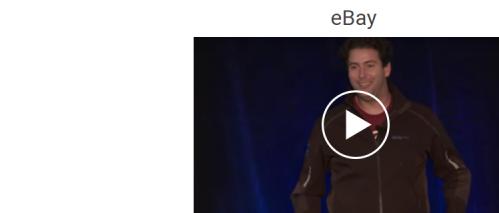
Macy's uses H2O for personalized site recommendations.



Transamerica turns to H2O to develop an insurance recommendation platform.



Paypal turned to H2O Deep Learning for fraud detection and customer churn.



eBay chose H2O for open source machine learning.



Cisco uses H2O to build a scalable model factory to improve sales and marketing.



H2O helps the country's largest TV behavior analytics company optimize ad performance.

Partners



NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI -- the next era of computing -- with the GPU acting as the brain of computers, robots, and self-driving cars that can perceive and understand the world.

[Website](#)



Founded in 1975, Microsoft (Nasdaq "MSFT") is the worldwide leader in software, services, devices and solutions that help people and businesses realize their full potential. You can launch Sparkling Water on Azure HDInsights with just a few clicks to build your data science pipeline on the cloud.

[Website](#) | [Documentation](#)



Watson Data Platform enables AI-powered decision making by simplifying and automating the development and operationalization of new insights. It enables unprecedented levels of collaboration amongst data savvy professionals.



Cloudera delivers a modern platform for data management and analytics, helping businesses solve their most challenging problems with data.



Anaconda is the leading open data science platform powered by Python. Continuum Analytics is the creator and driving force behind Anaconda. We put superpowers into the hands of people who are changing the world.



The MapR Converged Data Platform integrates enormous power of Hadoop and Spark with global event streaming, real-time database capabilities, and enterprise storage.



Kensu's mission is to lift Data Science to the Enterprise level focusing on the production environment and the maintenance across time.

[Website](#)



SigOpt is the optimization platform that amplifies your research. SigOpt takes any research pipeline and tunes it, right in place.

[Website](#)



Nimbix is the leading provider of purpose-built cloud computing for machine learning, AI and HPC applications. Powered by JARVICE™, the Nimbix Cloud provides high-performance software as a service, dramatically speeding up data processing for Energy, Life Sciences, Manufacturing, Media and Analytics applications.

[Website](#)



Databricks provides a just-in-time data platform, to simplify data integration, real-time experimentation, and robust deployment of production applications.



Hortonworks drives actionable intelligence with Connected Data Platforms that maximize the value of all data—data-in-motion and data-at-rest.



Minio is an object storage server built for cloud application developers and devops.



DataScience.com pairs data expertise with powerful tools to help businesses unlock the value in their data. Enabling data science for every business.

[Website](#)

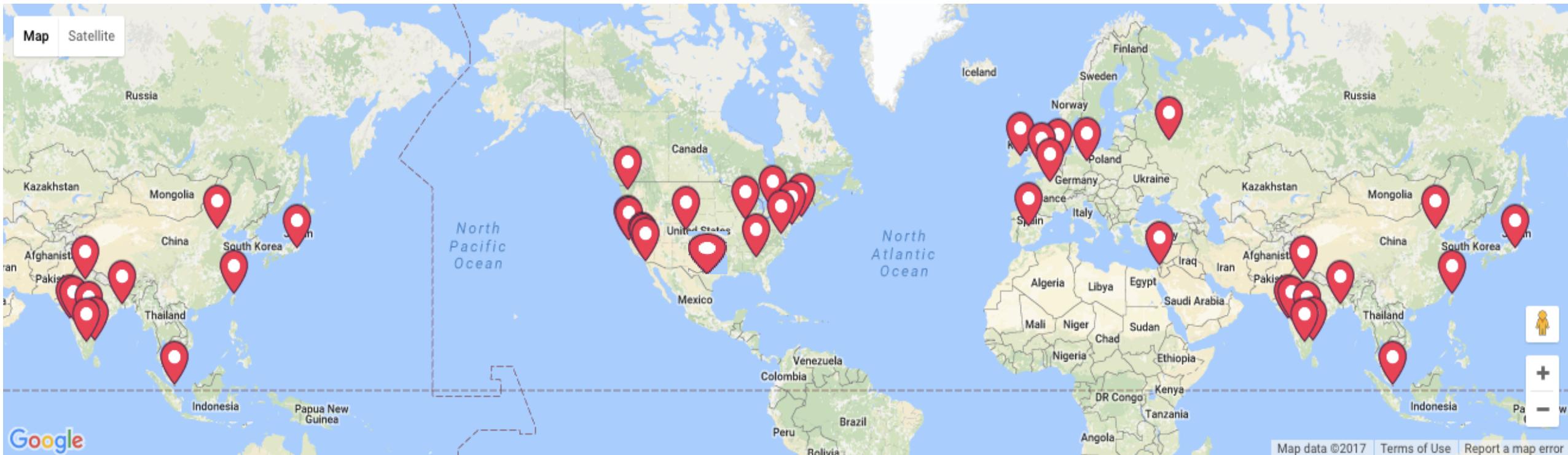


MapD makes queries faster and creates a fluid and immersive data exploration experience that removes the disconnect between an analyst and their data. Making extracting insight from data effortless and lightning fast.

[Website](#)

Community Expansion

15 Meetups a month



66,843
members 33
interested 50
Meetups 45
cities 18
countries

Find out more: www.h2o.ai/community/

22
FEB

Thursday, February 22, 2018

External Registration Required: Amsterdam AI & Deep Learning Meetup at ING

Hosted by [Jo-fai Chow](#)From [Amsterdam Artificial Intelligence & Deep Learning](#)

You're going

 [Share](#) [Tweet](#) [Invite](#)

Details

 [Organizer tools ▾](#)

Many thanks to ING for hosting our next event in Amsterdam.

Note: EXTERNAL REGISTRATION (Eventbrite) is required. Please use the following link to register. You must provide your real name and email address. Please also bring your ID for security check (just in case).

<https://www.eventbrite.com/e/amsterdam-ai-deep-learning-meetup-at-ing-tickets-42783160585>

Thursday, February 22, 2018

5:45 PM to 8:15 PM

[Add to calendar](#)

Needs a location

Agenda:

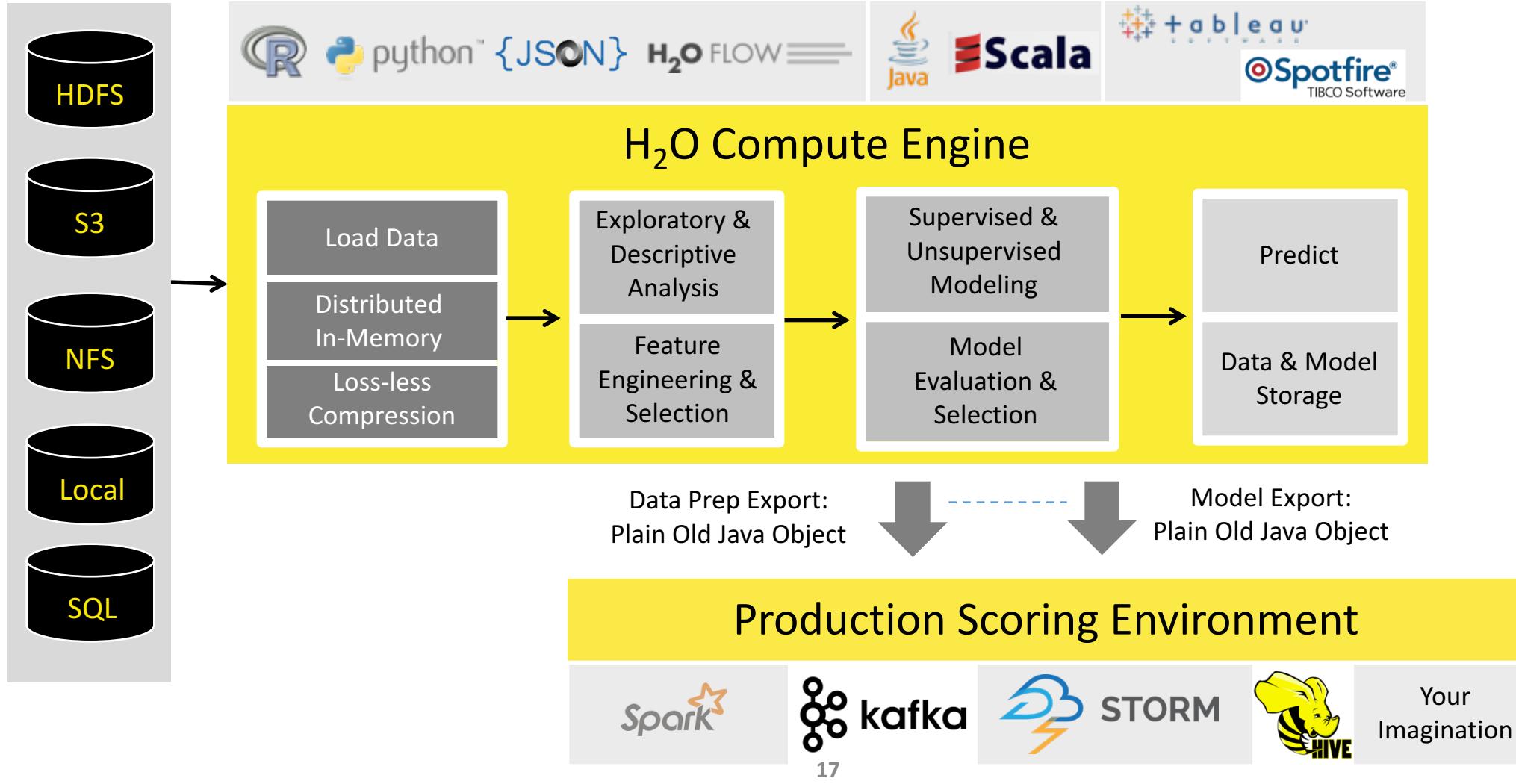
17:50 - 18:10 - Doors open, pizzas, drinks and networking

18:10 - 18:20 - Introduction

18:20 - 18:45 - Talk 1: Anomaly Detection in Finance using Isolation Forest by Andreea Bejinaru

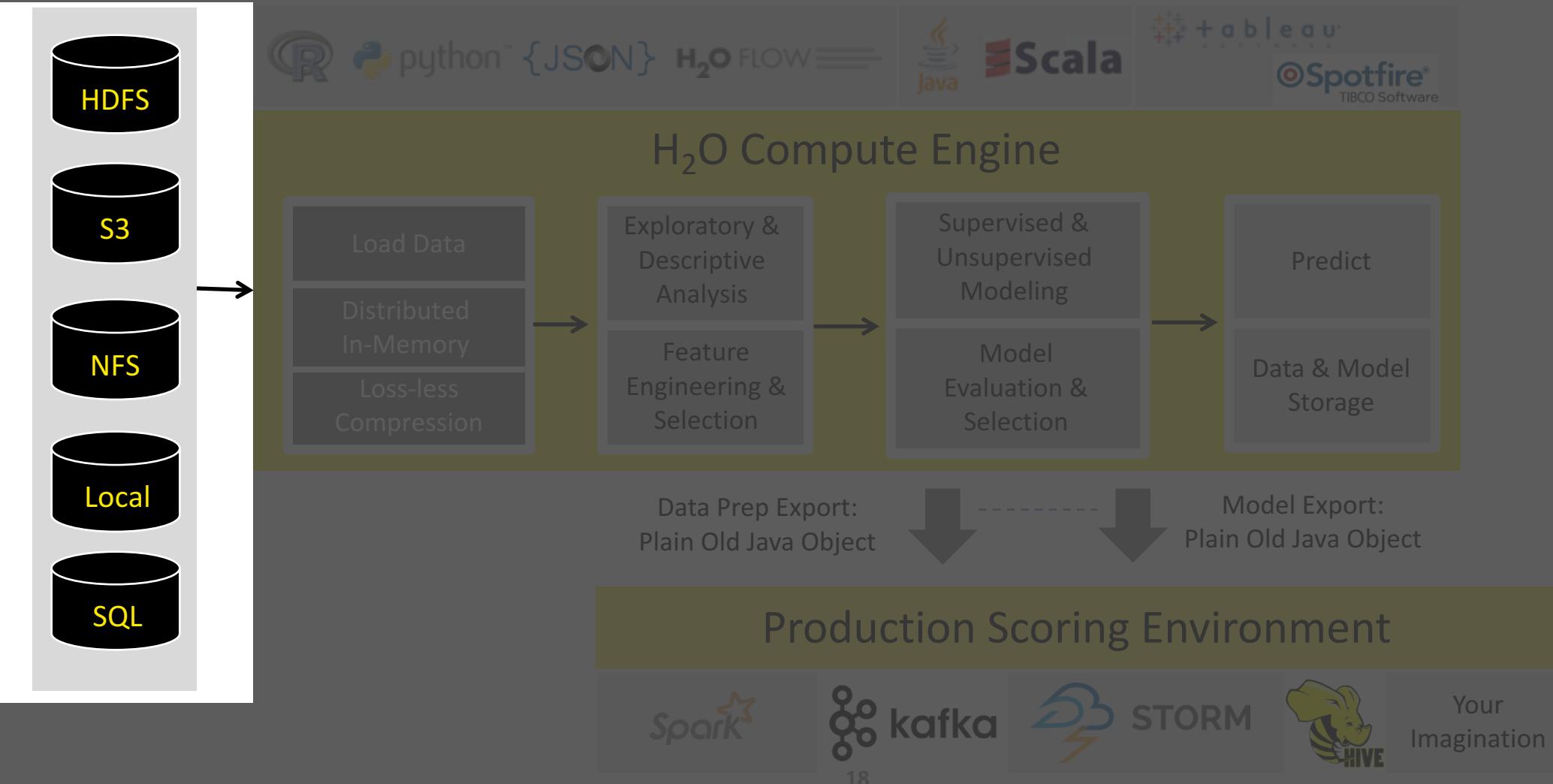
H₂O Machine Learning Platform

High Level Architecture

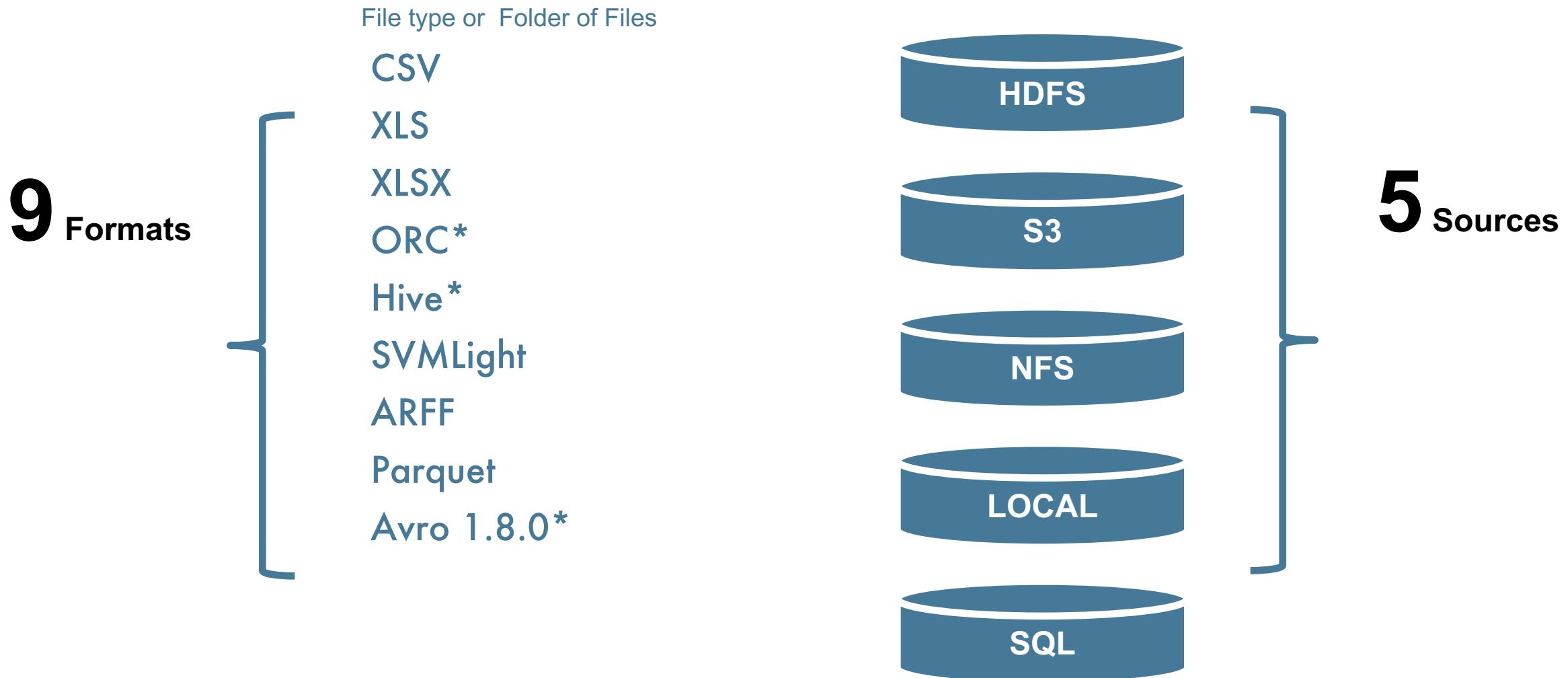


High Level Architecture

Import Data from
Multiple Sources



Supported Formats & Data Sources



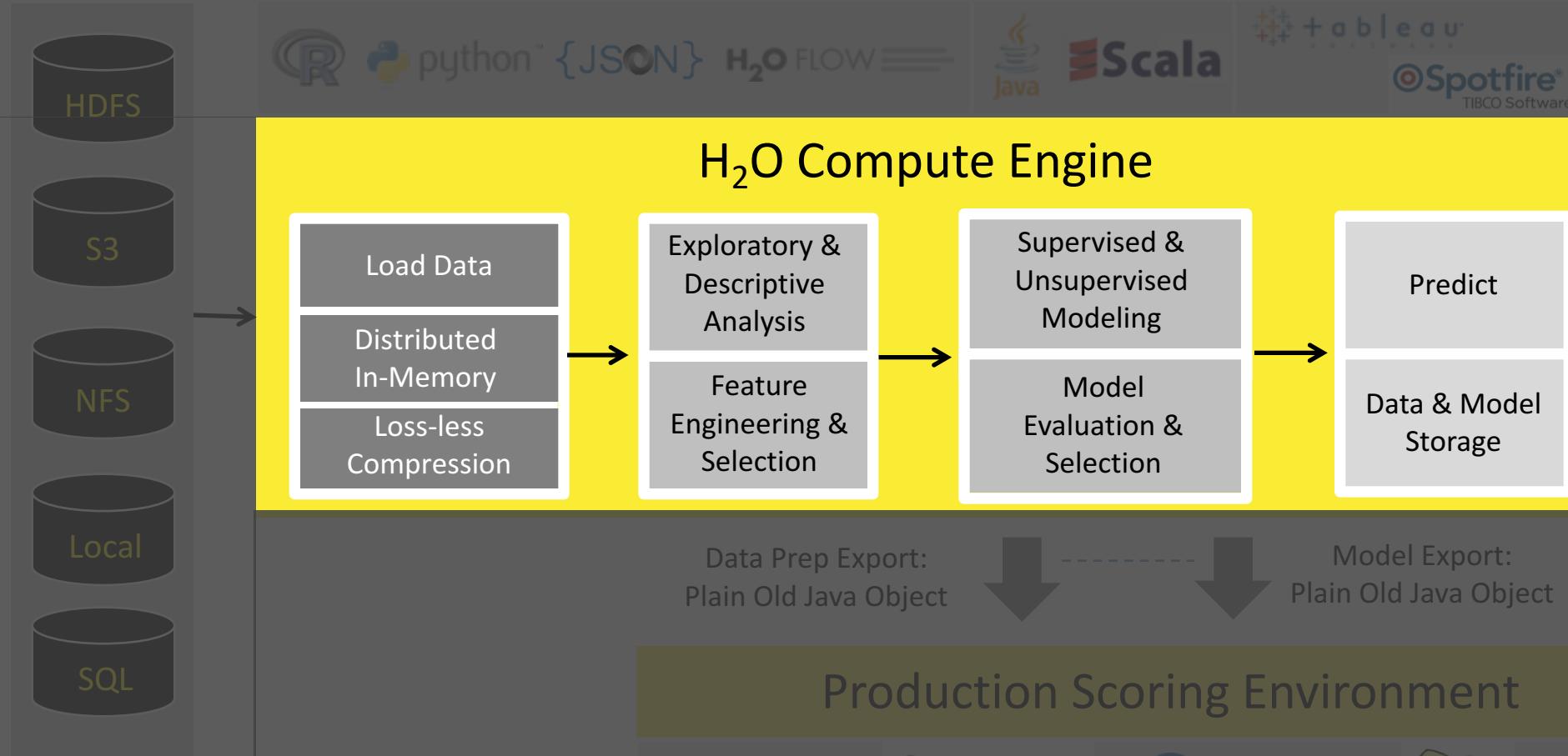
* 1. only if H2O is running as a Hadoop job

* 2. Hive files that are saved in ORC format

* 3. without multi-file parsing or column type modification

High Level Architecture

Fast, Scalable & Distributed
Compute Engine Written in
Java

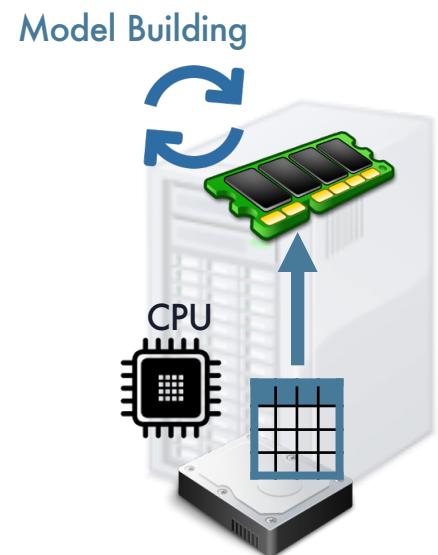


Your
Imagination

H₂O Core



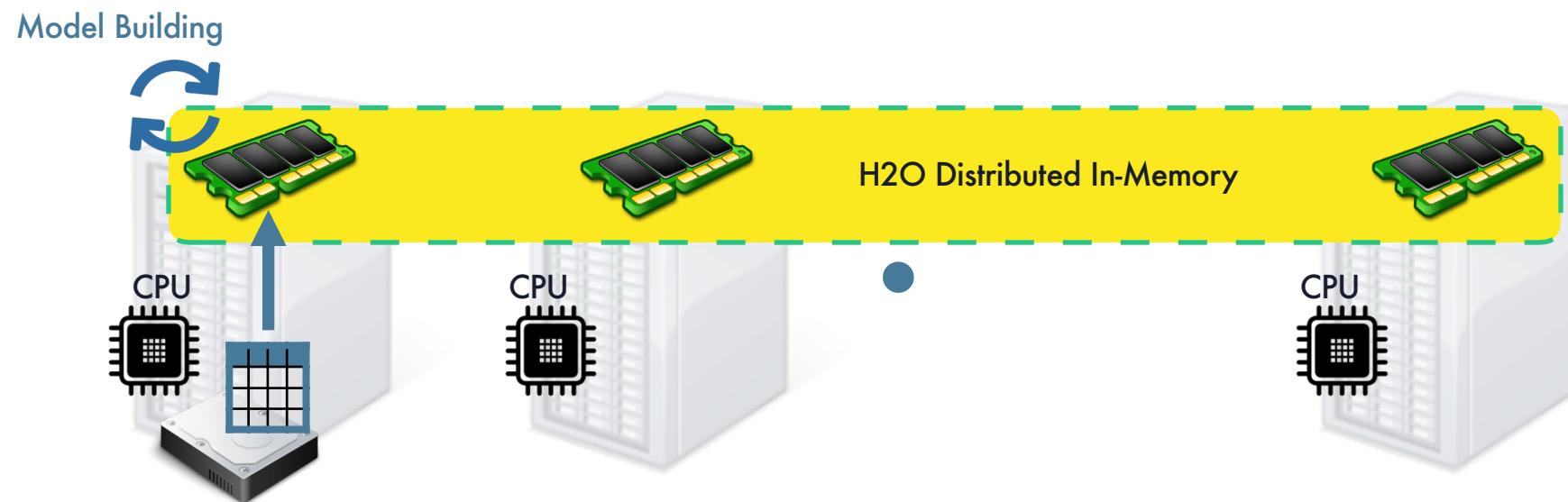
H₂O Core



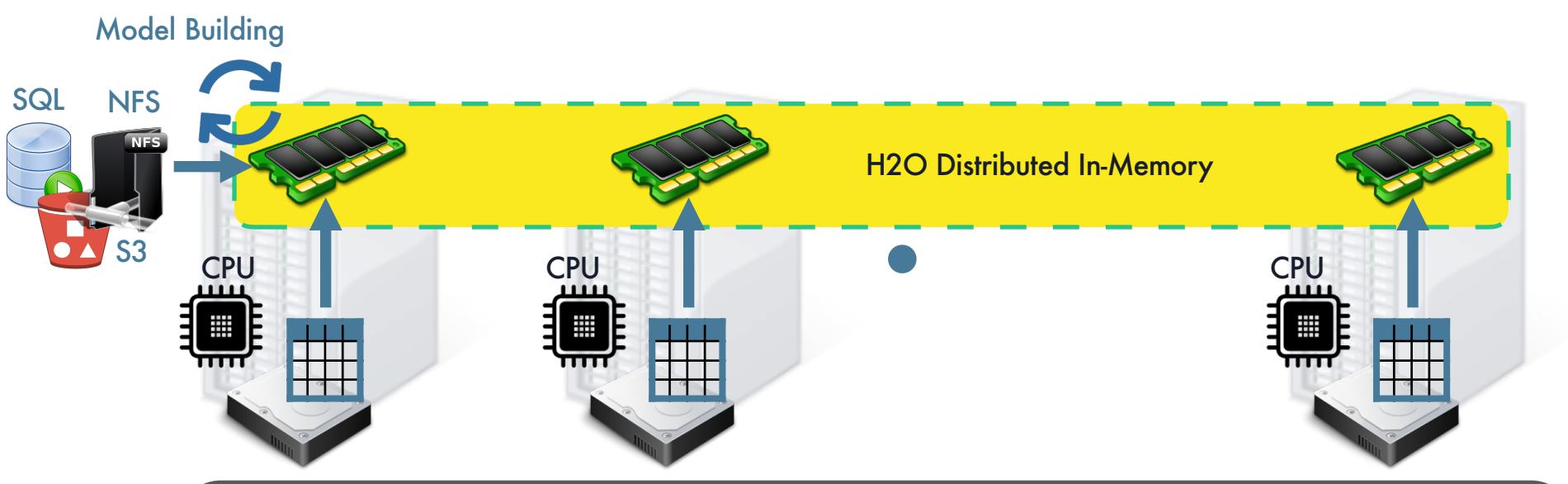
H₂O Core



H₂O Core



H₂O Core



YARN

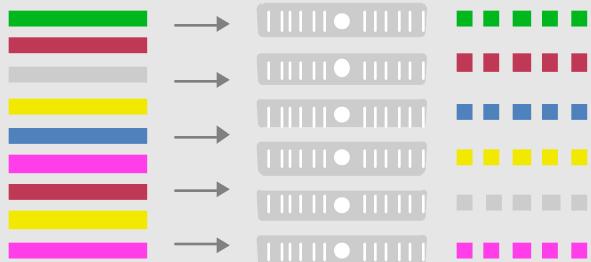
cloudera

Hortonworks

MAPR

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable
Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H₂O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- “Out-of-box” functionalities for all algorithms (**NO MORE SCRIPTING**) and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

Algorithms Overview

Supervised Learning

Statistical Analysis

- **Generalized Linear Models:** Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**

Ensembles

- **Distributed Random Forest:** Classification or regression models
- **Gradient Boosting Machine:** Produces an ensemble of decision trees with increasing refined approximations

Deep Neural Networks

- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

Unsupervised Learning

Clustering

- **K-means:** Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

Dimensionality Reduction

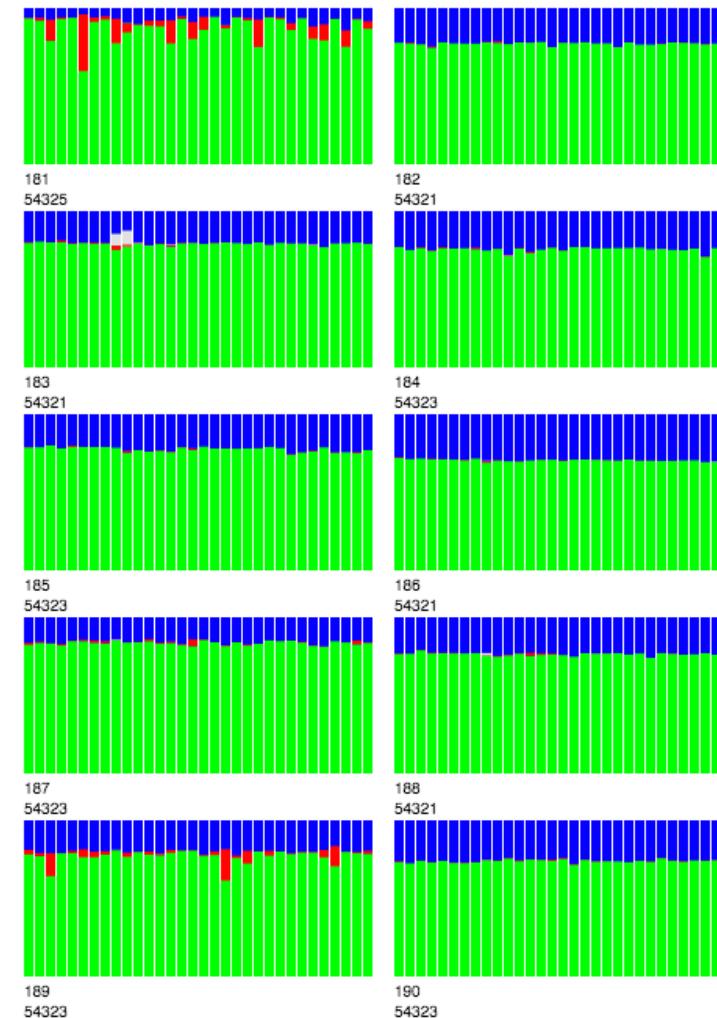
- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Anomaly Detection

- **Autoencoders:** Find outliers using a nonlinear dimensionality reduction using deep learning

H₂O Water Meter (CPU Monitor)

10 x 32 = 320 Cores



Legend

Each bar represents one CPU.

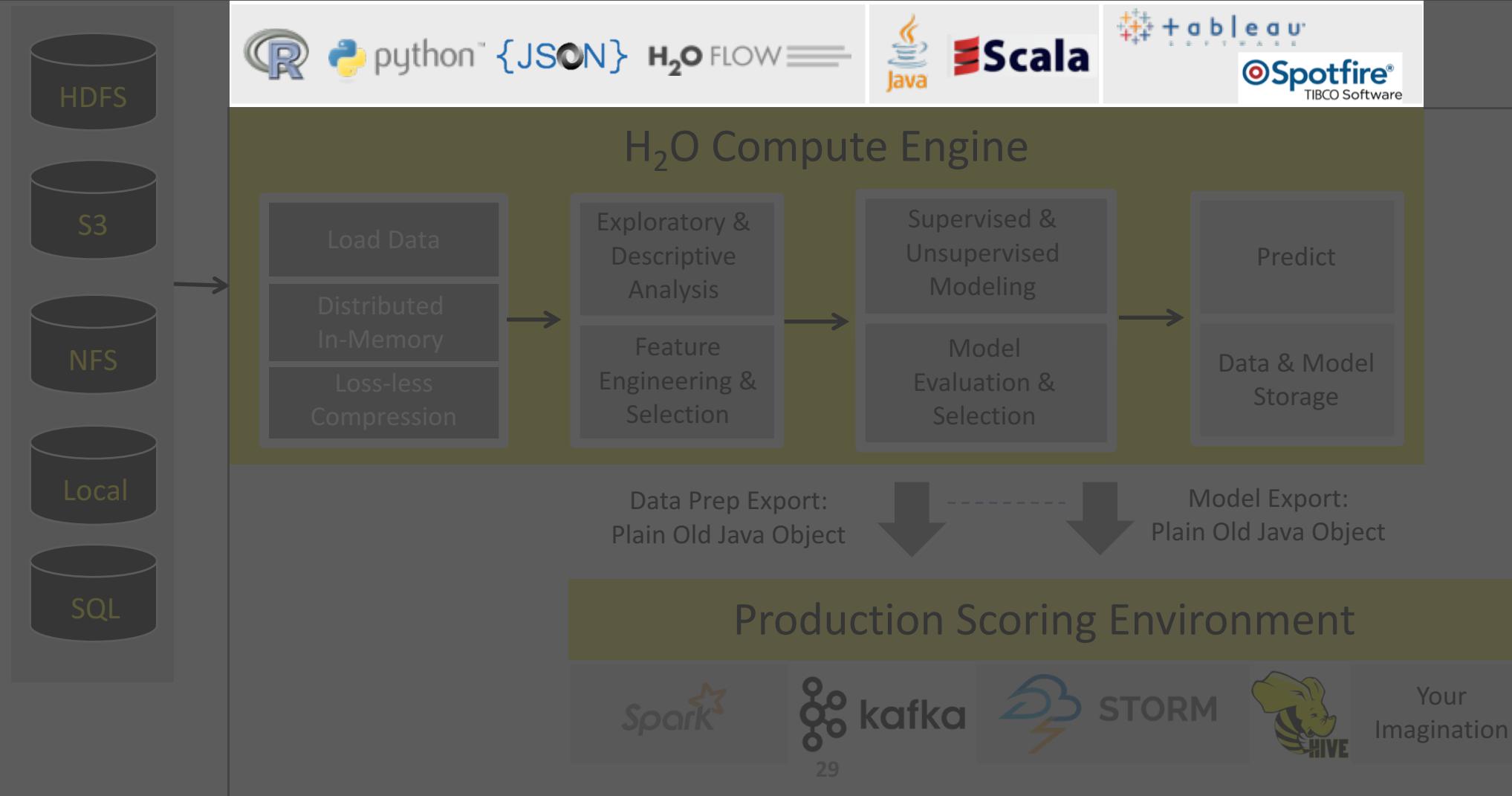
Blue: idle time

Green: user time

Red: system time

White: other time (e.g. i/o)

High Level Architecture



H₂O Flow (Web)

The screenshot shows the H2O Flow (Web) interface running in a web browser. The title bar reads "H2O Flow" and the address bar shows "localhost:54321/flow/index.html". The top navigation bar includes "Flow", "Cell", "Data", "Model" (which is highlighted in yellow), "Score", "Admin", and "Help". A toolbar below the navigation bar contains various icons for file operations like import, export, and search.

The main workspace is titled "Untitled Flow" and contains a single step labeled "assist". To the left of the workspace is a sidebar titled "Assistance" which lists various H2O routines with their descriptions:

Routine	Description
<code>importFiles</code>	Import file(s) into H2O
<code>getFrames</code>	Get a list of frames in H2O
<code>splitFrame</code>	Split a frame into two or more
<code>mergeFrames</code>	Merge two frames into one
<code>getModels</code>	Get a list of models in H2O
<code>getGrids</code>	Get a list of grid search results
<code>getPredictions</code>	Get a list of predictions in H2O
<code>getJobs</code>	Get a list of jobs running in H2O
<code>buildModel</code>	Build a model
<code>runAutoML</code>	Automatically train and tune
<code>importModel</code>	Import a saved model
<code>predict</code>	Make a prediction

A context menu is open over the "assist" step, showing options such as Aggregator..., Deep Learning..., Distributed Random Forest..., Gradient Boosting Machine..., Generalized Linear Modeling..., Generalized Low Rank Modeling..., K-means..., Naive Bayes..., Principal Components Analysis..., Stacked Ensemble..., Word2Vec..., and XGBoost... .

The right side of the interface features a "HELP" panel with sections for "Using Flow for the first time?", "Quickstart Videos", "Or, view example Flows to explore and learn H2O.", "STAR H2O ON GITHUB!", "GENERAL" (with links to Flow Web UI, Importing Data, Building Models, Making Predictions, Using Flows, and Troubleshooting Flow), and "EXAMPLES" (describing Flow packs and providing a link to Browse installed packs...). The bottom right corner shows "Connections: 0" and the H2O logo.

Interface – R and Python

The screenshot shows the RStudio Source Editor window with the file `credit_card_example.R` open. The code is a script for training a GBM model on a credit card dataset. It includes imports, data loading from S3, model training, predictions, and a brief look at datasets.

```
~/Documents/repo_h2o/sales-engineering - master - RStudio Source Editor
credit_card_example.R
Source on Save Run Source Cell Toolbar
1 # Credit Card Example
2
3 # Datasets:
4 # https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv
5 # https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv
6
7 # Start and connect to a local H2O cluster
8 library(h2o)
9 h2o.init(nthreads = -1)
10
11 # Import datasets from s3
12 df_train = h2o.importFile("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv")
13 df_test = h2o.importFile("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv")
14
15 # Look at datasets
16 summary(df_train)
17 summary(df_test)
18
19 # Define features and target
20 features = colnames(df_test)
21 target = "DEFAULT_PAYMENT_NEXT_MONTH"
22
23 # Train a GBM model
24 model_gbm = h2o.gbm(x = features,
25   y = target,
26   training_frame = df_train,
27   seed = 1234)
28 print(model_gbm)
29
30 # Use GBM model for making predictions
31 yhat_test = h2o.predict(model_gbm, newdata = df_test)
32 head(yhat_test)
33
34 # (Extra) Use H2O's AutoML
35 aml = h2o.automl(x = features,
36   y = target,
37   training_frame = df_train,
38   max_runtime_secs = 60,
39   seed = 1234)
40
41 # Print leaderboard
42 print(aml@leaderboard)
43
44 # Use best model for making predictions
45 best_model = aml@leader
46 yhat_test = h2o.predict(best_model, newdata = df_test)
47 head(yhat_test)
48
49
```

The screenshot shows a Jupyter Notebook interface with the notebook `credit_card_example.ipynb` open. The notebook contains Python code for connecting to a local H2O cluster, importing datasets, and summarizing them. The output cell shows the successful connection to the H2O server and the summary statistics for the datasets.

In [2]:

```
# Start and connect to a local H2O cluster
import h2o
h2o.init(nthreads = -1)

Checking whether there is an H2O instance running at http://localhost:54321.... not found.
Attempting to start a local H2O server...
Java Version: java version "1.8.0_72"; Java(TM) SE Runtime Environment (build 1.8.0_72-b15); Java HotSpot(TM) 64-Bit Server VM (build 25.72-b15, mixed mode)
Starting server from /Users/jofaichow/anaconda/lib/python2.7/site-packages/h2o/backend/bin/h2o.jar
Ice root: /var/folders/4z/p7yt7_4n4fjijlyg6g4qfbw000gn/T/tmpPdP3Av
JVM stdout: /var/folders/4z/p7yt7_4n4fjijlyg6g4qfbw000gn/T/tmpPdP3Av/h2o_jofaichow_started_from_python.out
JVM stderr: /var/folders/4z/p7yt7_4n4fjijlyg6g4qfbw000gn/T/tmpPdP3Av/h2o_jofaichow_started_from_python.err
Server is running at http://127.0.0.1:54321
Connecting to H2O server at http://127.0.0.1:54321... successful.
```

H2O cluster uptime:	02 secs
H2O cluster version:	3.13.0.3981
H2O cluster version age:	29 days
H2O cluster name:	H2O_from_python_jofaichow_id7qa
H2O cluster total nodes:	1

In [3]:

```
# Import datasets from s3
df_train = h2o.import_file("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv")
df_test = h2o.import_file("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv")
```

Parse progress: | 100%
Parse progress: | 100%

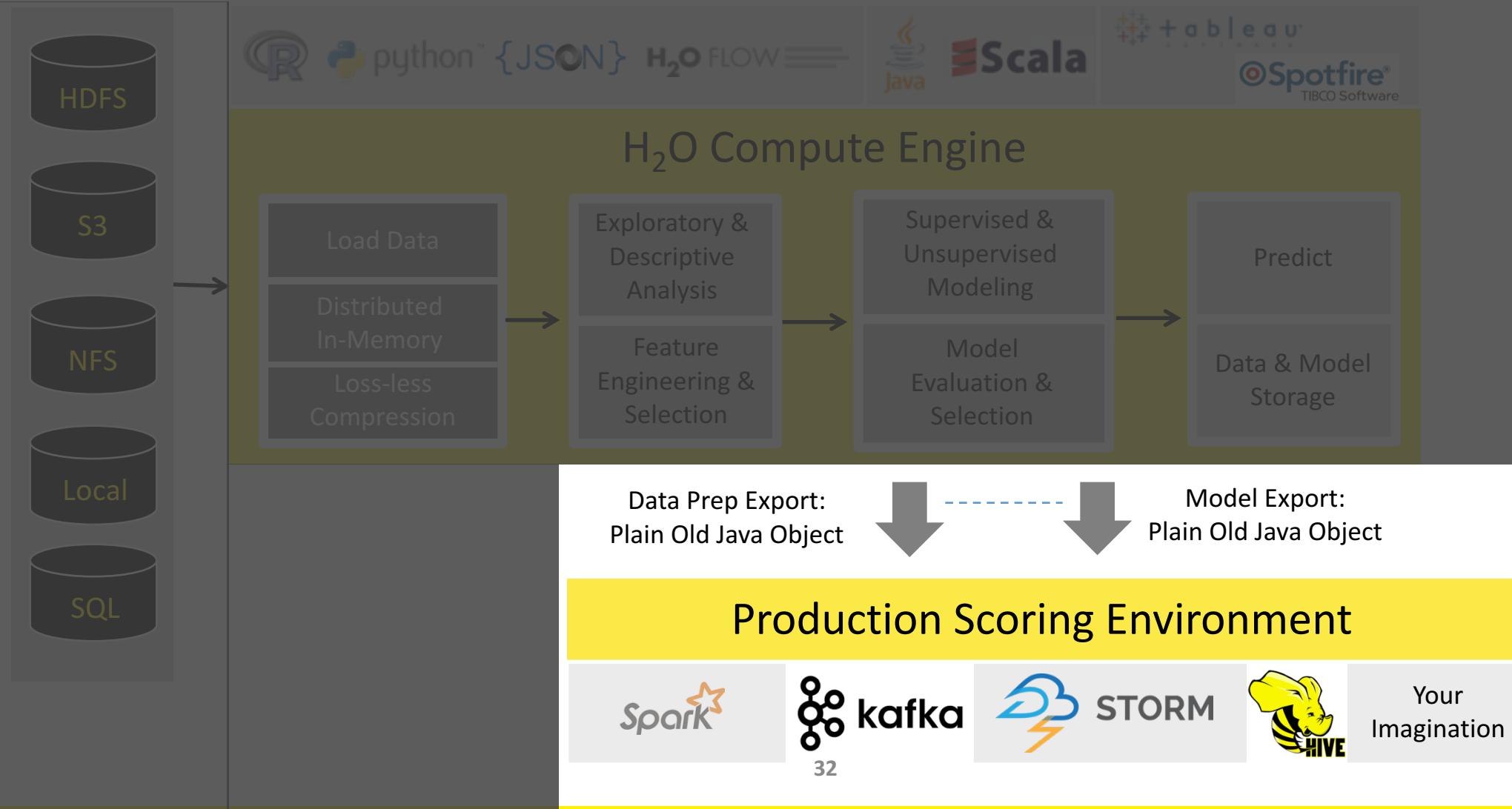
In [4]:

```
# Look at datasets
df_train.summary()
df_test.summary()
```

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4
type	int	enum	int	int	int	int	int	int	int
mins	10000.0		0.0	0.0	21.0	-2.0	-2.0	-2.0	-2.0
mean	165471.466667		1.85	1.55578703704	35.4053240741	-0.00523148148148	-0.122361111111	-0.15537037037	-0.210601
maxs	1000000.0		6.0	3.0	79.0	8.0	8.0	8.0	8.0
sigma	128853.314839		0.779559696278	0.522505078476	9.27675421641	1.12668964211	1.20086854503	1.20727030901	1.172176
zeros	0		9	37	0	10563	11284	11309	11905
missing	0		0	0	0	0	0	0	0

High Level Architecture

Export Standalone Models
for Production



H₂O Documentation

[Getting Started & User Guides](#) | [Q & A](#) | [Algorithms](#) | [Languages](#) | [Tutorials, Examples, & Presentations](#) | [API & Developer Docs](#) | [For the Enterprise](#)

Getting Started & User Guides

 Open Source |  Commercial

H₂O

What is H₂O?
[H₂O User Guide](#) (Main docs)
H₂O Book (O'Reilly)
Recent Changes
Open Source License (Apache V2)

Quick Start Video - Flow Web UI
Quick Start Video - R
Quick Start Video - Python

[Download H₂O](#)

Sparkling Water

What is Sparkling Water?
Sparkling Water Booklet
PySparkling Readme 2.0 | 2.1 | 2.2
RSparkling Readme
Open Source License (Apache V2)

Quick Start Video - Scala

[Download Sparkling Water](#)

Driverless AI

What is Driverless AI?
Driverless AI User Guide [HTML](#) [PDF](#)
Driverless AI Booklet
MLI with Driverless AI Booklet

Driverless AI Webinars

[Download Driverless AI](#)

H₂O4GPU (alpha)

H₂O4GPU Readme
Open Source License (Apache V2)

[Download H₂O4GPU](#)

URL: [docs.h2o.ai](#)