H_2O.ai

Scalable Machine Learning For Smarter Applications

H_2O.ai

Hank Roark
Data Scientist / Hacker
hank@h2o.ai
@hankroark
https://www.linkedin.com/in/hankroark

H2O.ai Overview

- Founded: 2011 venture-backed, debuted in 2012
- Product: H2O open source in-memory prediction engine
- Team: 34
- HQ: Mountain View, CA
- SriSatish Ambati – CEO & Co-founder (Founder Platfora, DataStax; Azul)
- Cliff Click – CTO & Co-founder (Creator Hotspot, Azul, Sun, Motorola, HP)
- Tom Kraljevic – VP of Engineering (CTO & Founder Luminix, Azul, Chromatic)

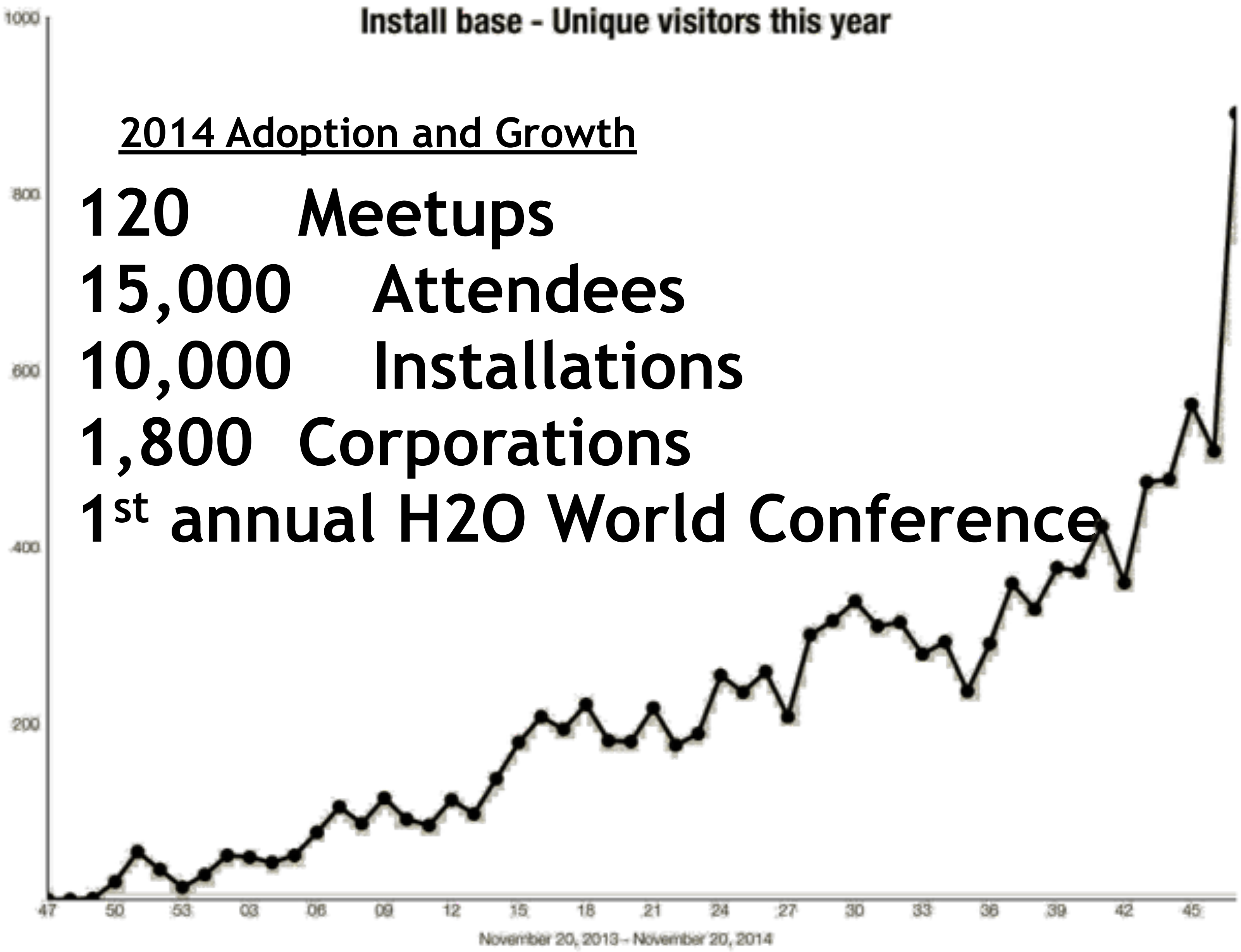
Distributed Systems Engineers Making ML Scale!

Scientific Advisory Council

Stephen Boyd
Professor of EE Engineering
Stanford University

Rob Tibshirani
Professor of Health Research and Policy, and Statistics
Stanford University

Trevor Hastie
Professor of Statistics
Stanford University



2014 Adoption and Growth

120 Meetups

15,000 Attendees

10,000 Installations

1,800 Corporations

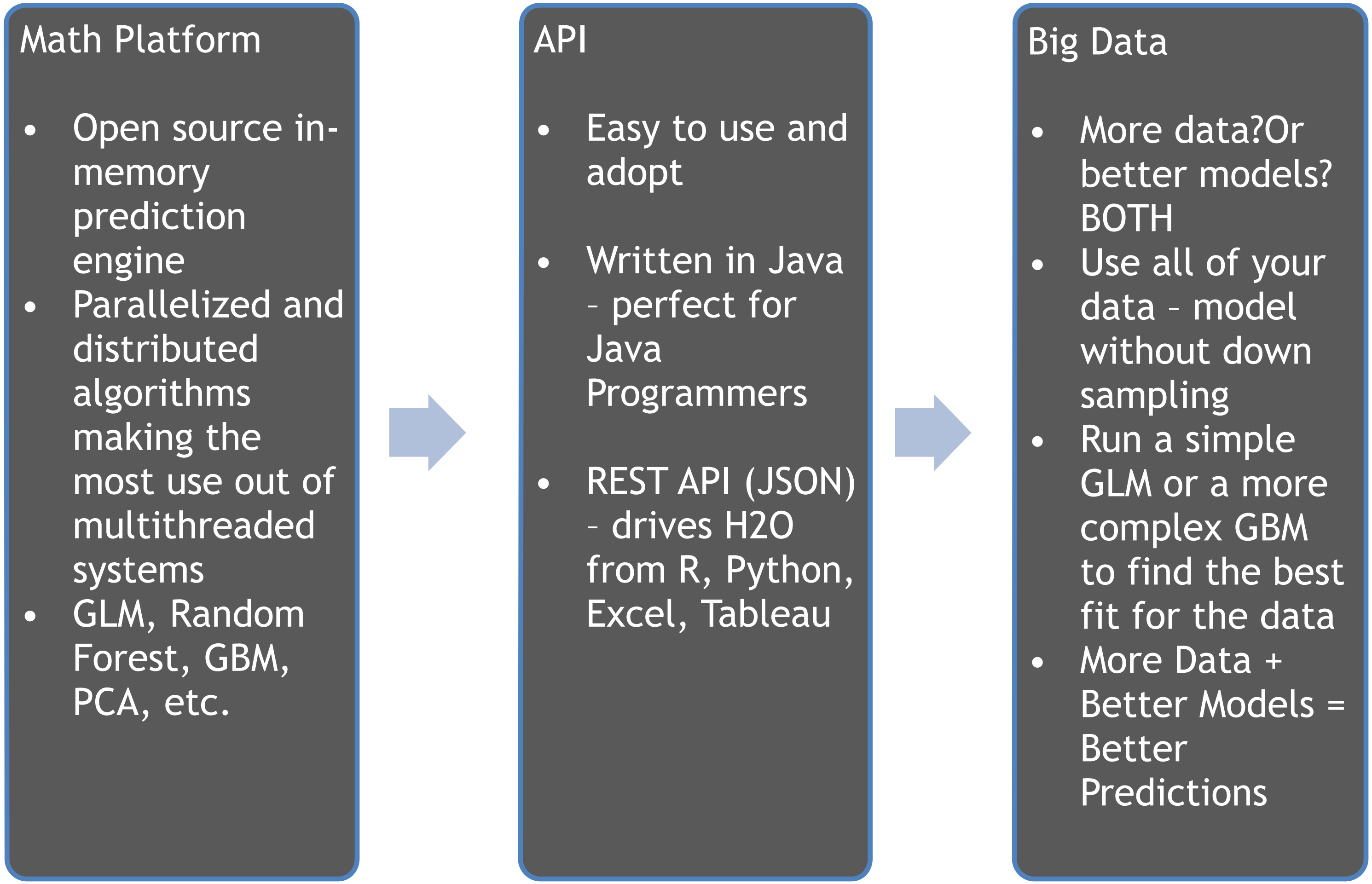
1st annual H2O World Conference

weeks

H_2O.ai

Machine Intelligence

What is H2O?



Math Platform

- Open source in-memory prediction engine
- Parallelized and distributed algorithms making the most use out of multithreaded systems
- GLM, Random Forest, GBM, PCA, etc.

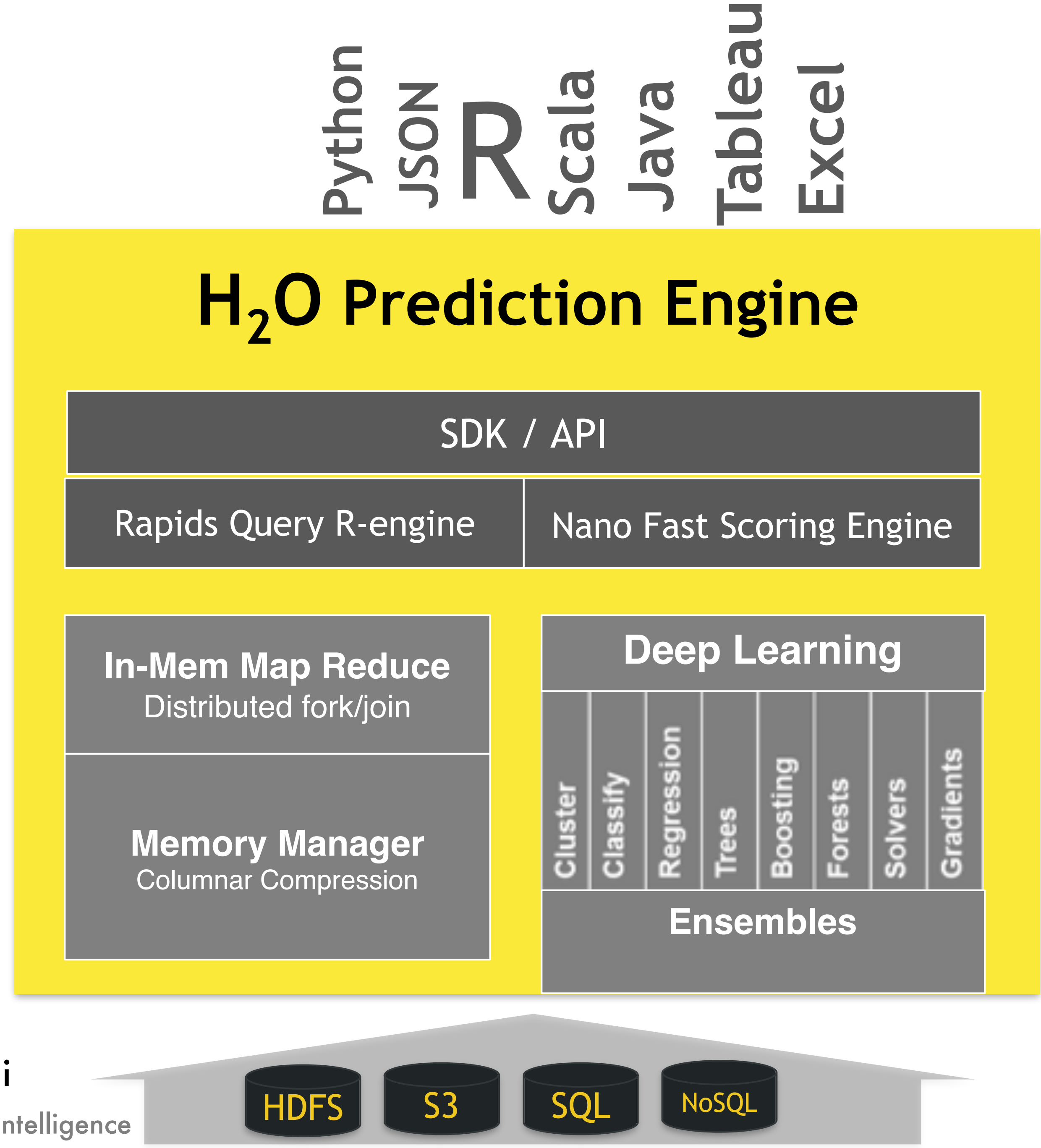
API

- Easy to use and adopt
- Written in Java - perfect for Java Programmers
- REST API (JSON) - drives H2O from R, Python, Excel, Tableau

Big Data

- More data?Or better models? BOTH
- Use all of your data - model without down sampling
- Run a simple GLM or a more complex GBM to find the best fit for the data
- More Data + Better Models = Better Predictions

H_2O.ai
Machine Intelligence



Python
JSON
R
Scala
Java
Tableau
Excel

H_2O Prediction Engine

SDK / API

Rapids Query R-engine

Nano Fast Scoring Engine

In-Mem Map Reduce
Distributed fork/join

Memory Manager
Columnar Compression

Deep Learning

Cluster
Classify
Regression
Trees
Boosting
Forests
Solvers
Gradients

Ensembles

HDFS
S3
SQL
NoSQL

On Premise
On Hadoop & Spark
On EC2

Per Node

2M	Row ingest/sec
50M	Row Regression/sec
750M	Row Aggregates / sec

H_2O.ai
Machine Intelligence

Algorithms on H2O

Supervised Learning

Statistical Analysis

- **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Cox Proportional Hazards Models**
- **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

Algorithms on H2O

Unsupervised Learning

Clustering

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

Dimensionality Reduction

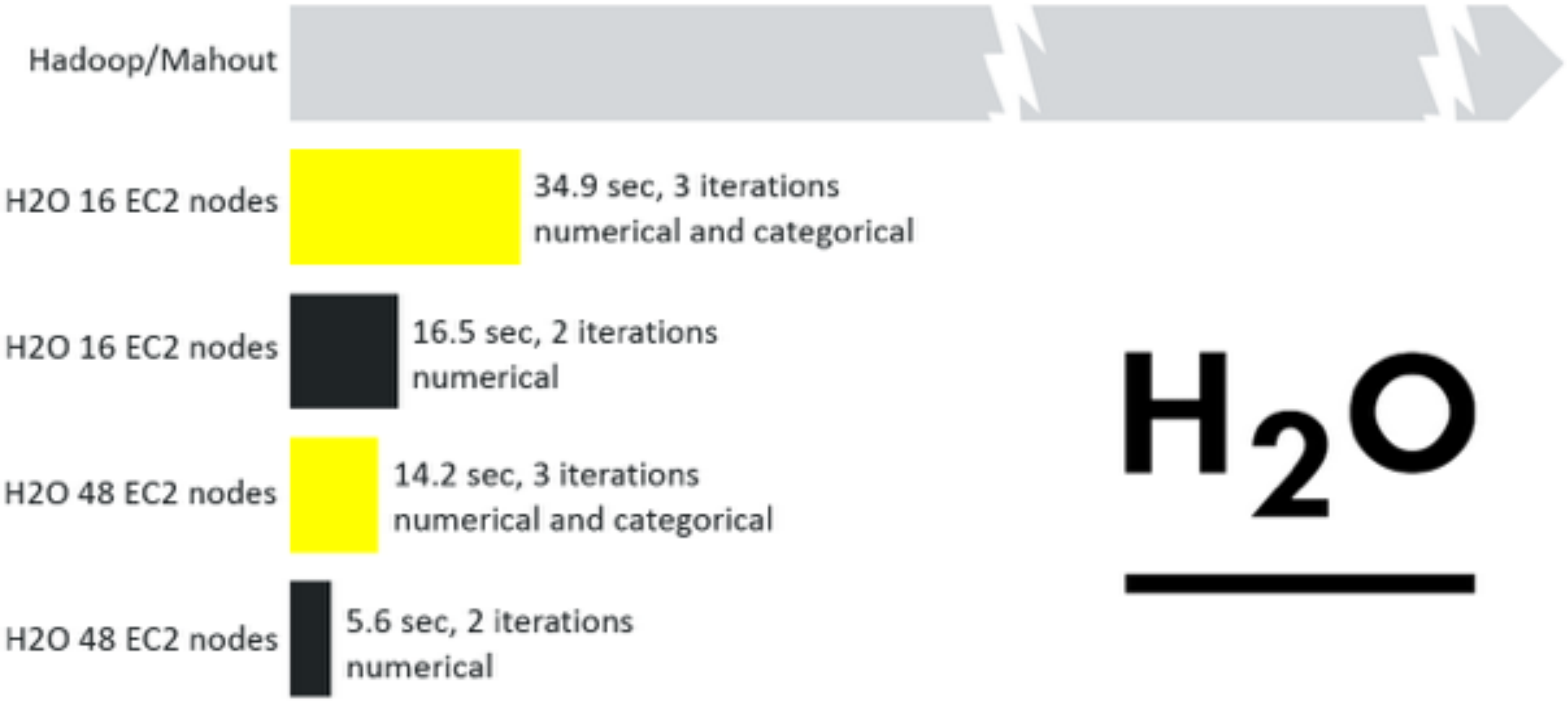
- **Principal Component Analysis**: Linearly transforms correlated variables to independent components

Anomaly Detection

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

H2O Billion Row Machine Learning Benchmark

GLM Logistic Regression

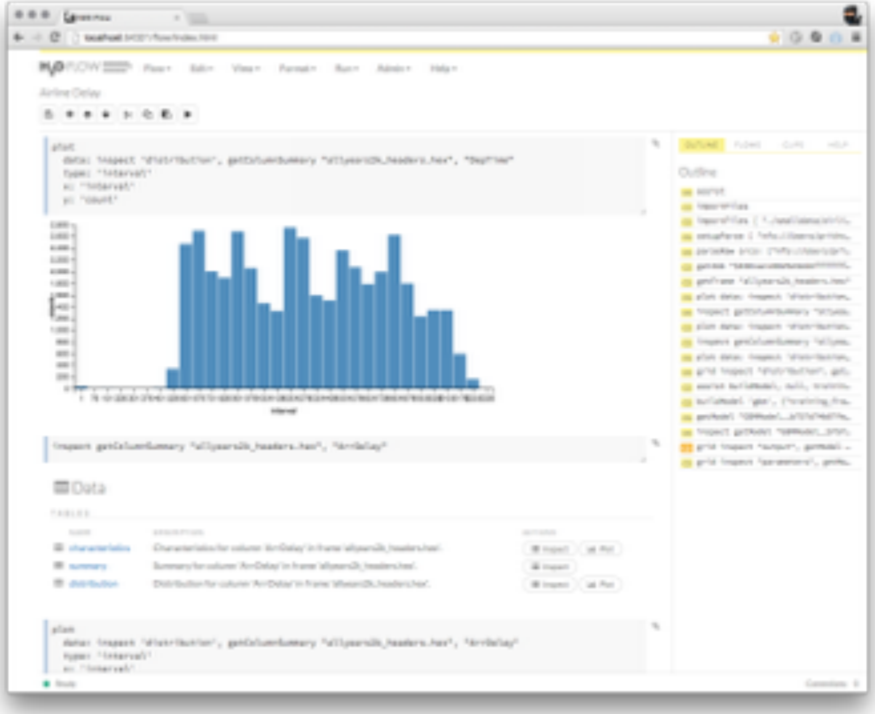


Compute Hardware: AWS EC2 c3.2xlarge - 8 cores and 15 GB per node, 1 GbE interconnect

Airline Dataset 1987-2013, 42 GB CSV, 1 billion rows, 12 input columns, 1 outcome column
9 numerical features, 3 categorical features with cardinalities 30, 376 and 380

H2O Flow

- A Web-based interactive computing environment for Big Data Machine Learning
- New web interface of H2O
- Model comparisons
- Mixed environment for
  - Coffeescript
  - Text & Markdown
  - Charts & Visualization (more to come)
  - R/Spark/Python code (coming soon)
  - Mathematics Equations (coming soon)
  - Video & Rich Media



```
plot
  data: inspect 'distribution', getColumnSummary "allyears2k_headers.hex", "DepTime"
  type: 'interval'
  x: 'interval'
  y: 'count'
```

```
inspect getColumnSummary "allyears2k_headers.hex", "ArrDelay"
```

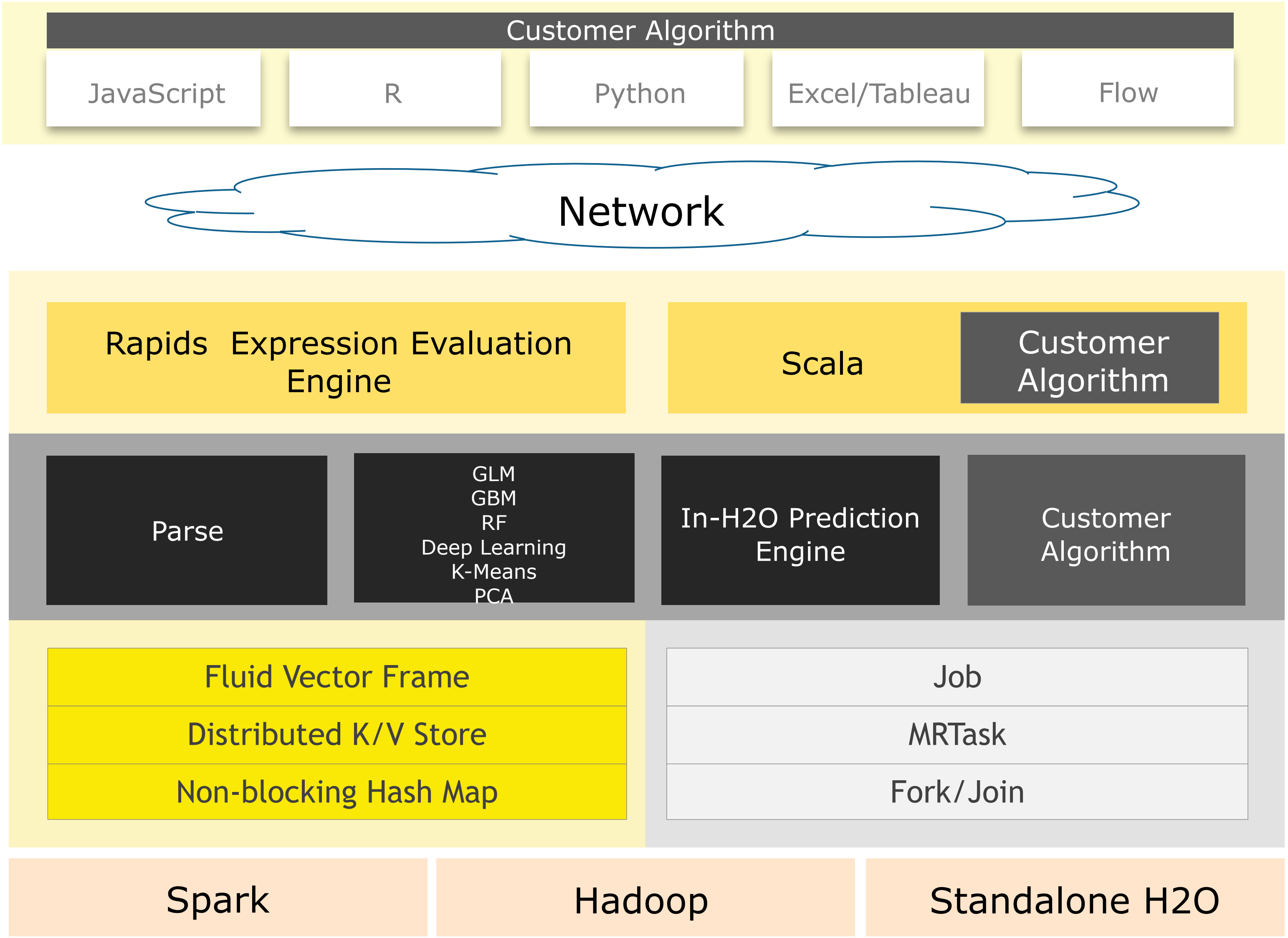
Data

TABLES

NAME	DESCRIPTION	ACTIONS
characteristics	Characteristics for column 'ArrDelay' in frame 'allyears2k_headers.hex'.	Inspect Plot
summary	Summary for column 'ArrDelay' in frame 'allyears2k_headers.hex'.	Inspect
distribution	Distribution for column 'ArrDelay' in frame 'allyears2k_headers.hex'.	Inspect Plot

```
plot
  data: inspect 'distribution', getColumnSummary "allyears2k_headers.hex", "ArrDelay"
  type: 'interval'
  x: 'interval'
```

H2O Software Stack



Customer Algorithm

JavaScript

R

Python

Excel/Tableau

Flow

Network

Rapids Expression Evaluation Engine

Scala

Customer Algorithm

Parse

GLM
GBM
RF
Deep Learning
K-Means
PCA

In-H2O Prediction Engine

Customer Algorithm

Fluid Vector Frame

Distributed K/V Store

Non-blocking Hash Map

Job

MRTask

Fork/Join

Spark

Hadoop

Standalone H2O

Reading Data from HDFS into H2O with R

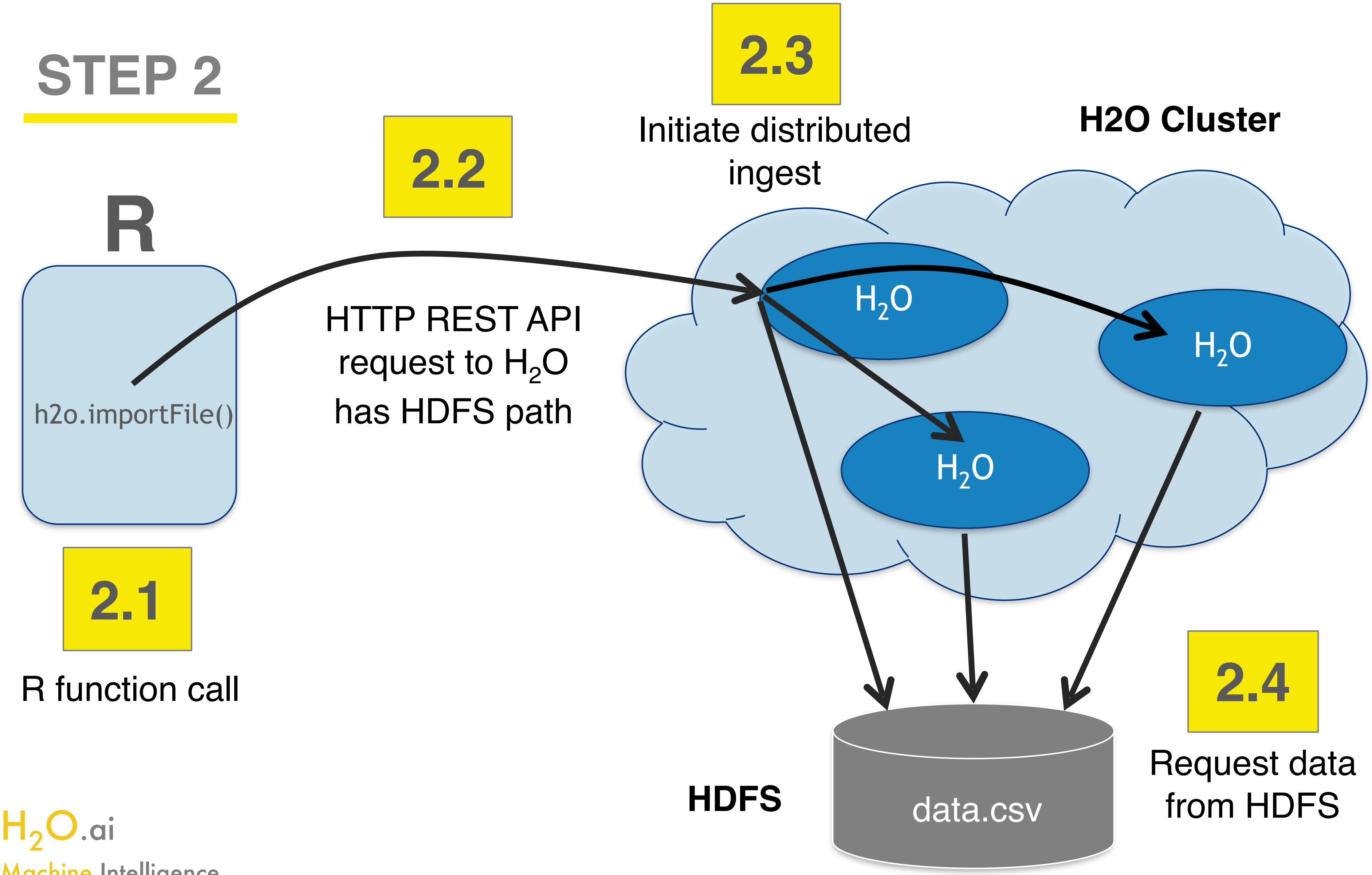
STEP 1

```
h2o_df = h2o.importFile(“hdfs://path/to/data.cs
```

R user

Reading Data from HDFS into H2O with R

STEP 2



R

h2o.importFile()

2.1

R function call

2.2

HTTP REST API request to H_2O has HDFS path

2.3

Initiate distributed ingest

H2O Cluster

2.4

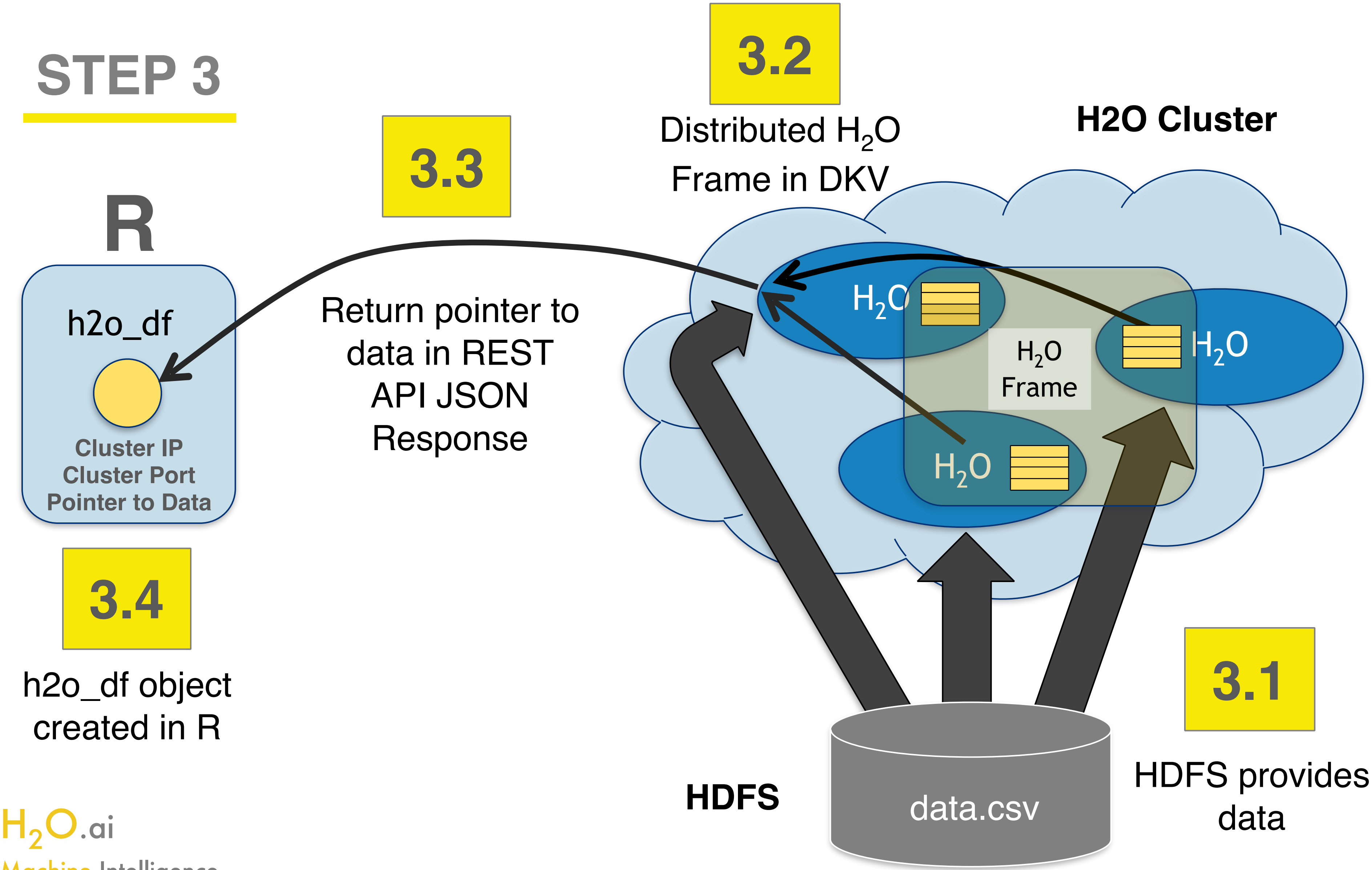
Request data from HDFS

HDFS

data.csv

Reading Data from HDFS into H2O with R

STEP 3



3.1

HDFS provides data

3.2

Distributed H_2O Frame in DKV

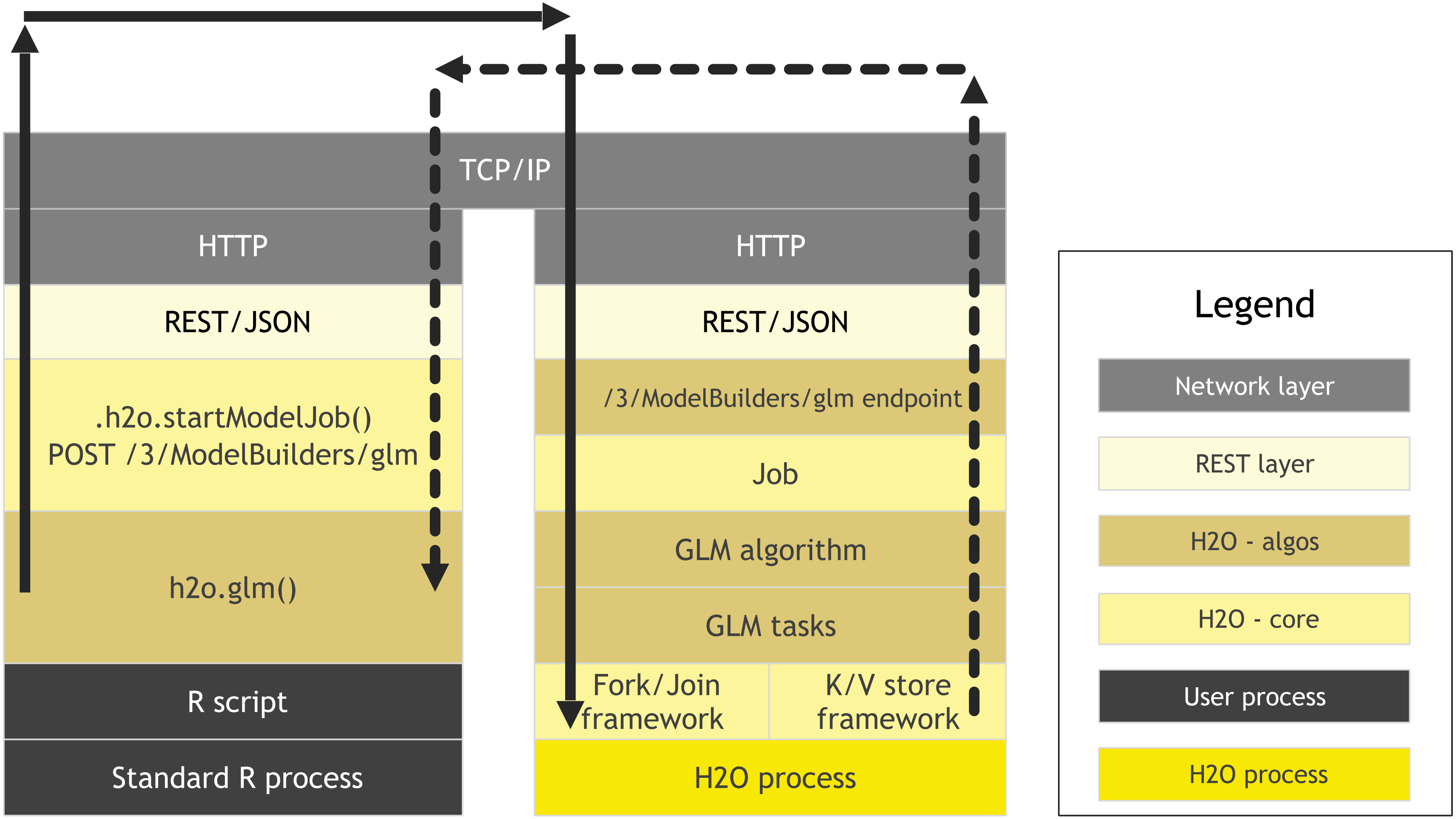
3.3

Return pointer to data in REST API JSON Response

3.4

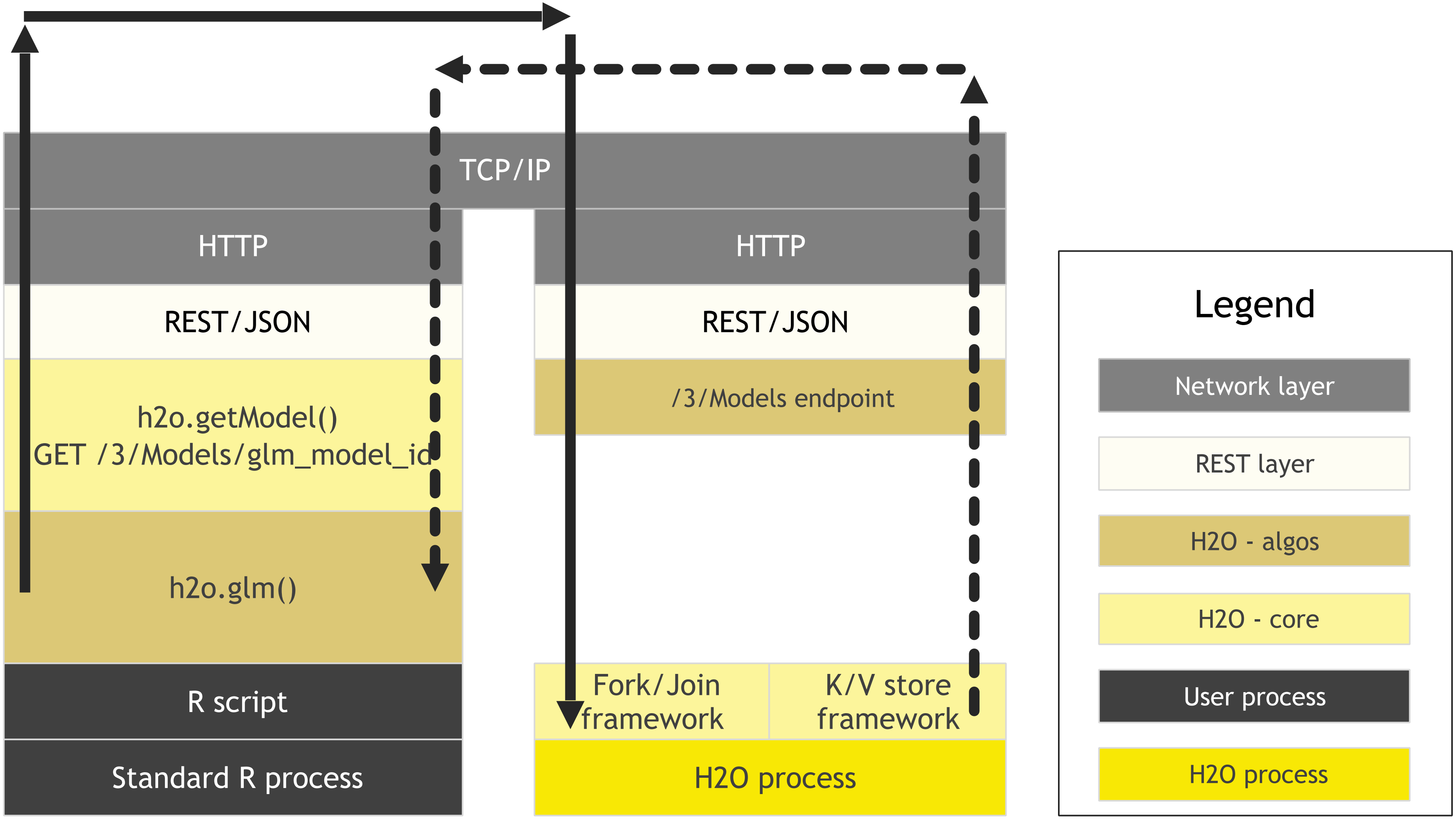
h2o_df object created in R

R Script Starting H2O GLM



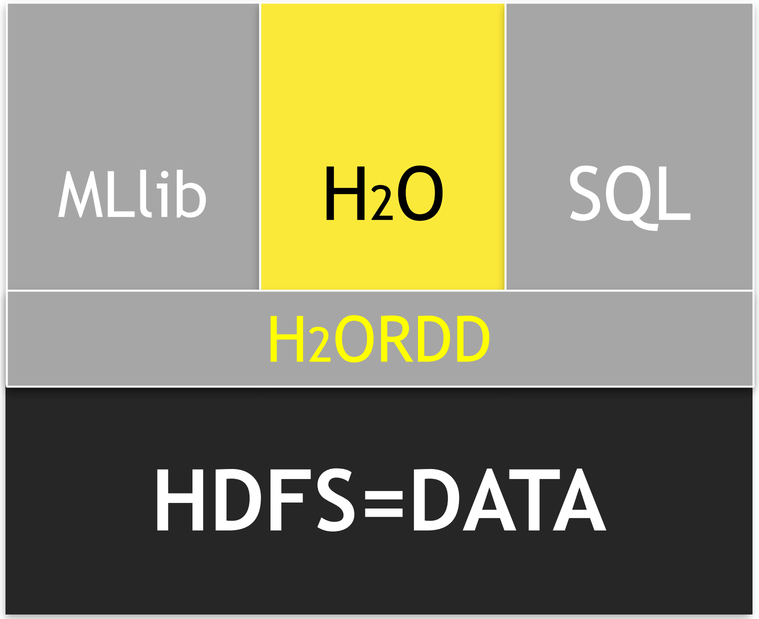
H_2O.ai
Machine Intelligence

R Script Retrieving H2O GLM Result



H_2O.ai
Machine Intelligence

H_2O – Sparkling Water



MLlib

H_2O

SQL

H_2ORDD

HDFS=DATA

In-Memory	Big Data, Columnar
ML	*100x faster Algos*
R	*CRAN, API, fast engine*
API	*Spark API, Java MM*
Community	*Devs, Data Science*

H_2O.ai

Scalable Machine Learning
For Smarter Applications