H2O.ai

Scalable Machine Learning Using H2O

Tom Kraljevic

October 22, 2015

Las Vegas, NV @ innevation

Outline for today's talk

- About H2O (10 minutes)
- Demo of Scalability and H2O Flow (10 minutes)
- Demo of R with H2O (Loan app) (15 minutes)
- Demo of Python with H2O (Notebook) (10 minutes)
- Demo of Sparkling Water (Spark + H2O) (Ask Craig) (15 minutes)
- Q & A (up to 30 minutes)

Content for today's talk can be found at:

https://github.com/h2oai/h2o-meetups/tree/master/2015_10_22_H2O_LV

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, Deep Learning, etc.

API

Easy to use and adopt

- Written in Java – perfect for Java Programmers
- Spark and Scala integration via Sparkling Water
- 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

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

Algorithms on H_2O

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

Algorithms on H_2O

Unsupervised Learning

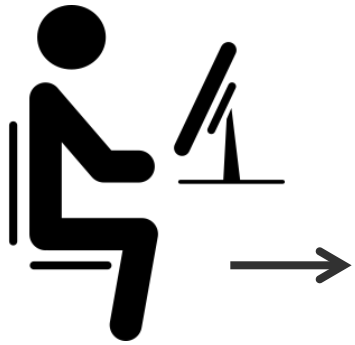
Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components
Anomaly Detection	• **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

Demo of H2O Flow

H2O and R / Python

Reading Data from HDFS into H2O with R

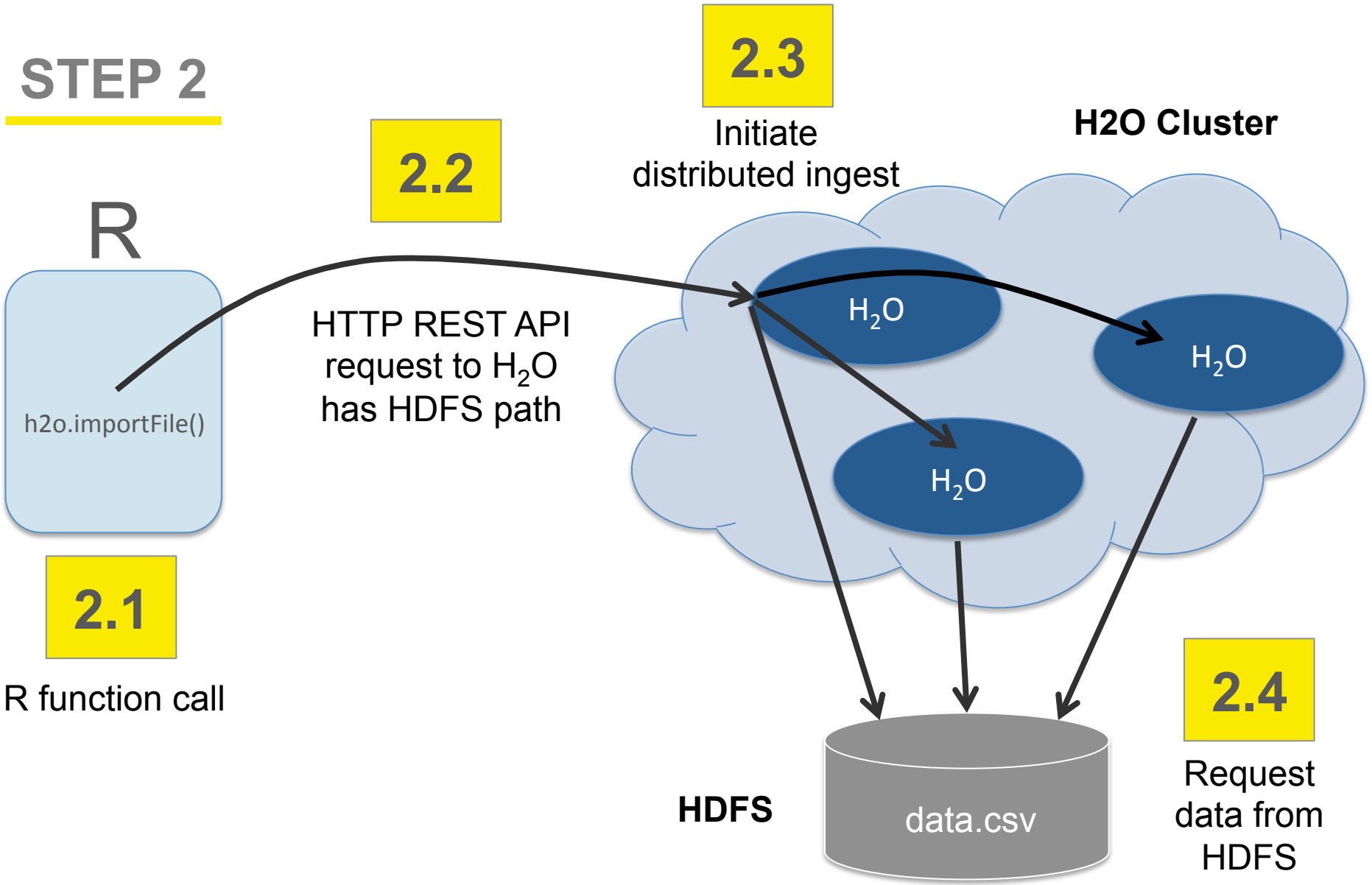
STEP 1



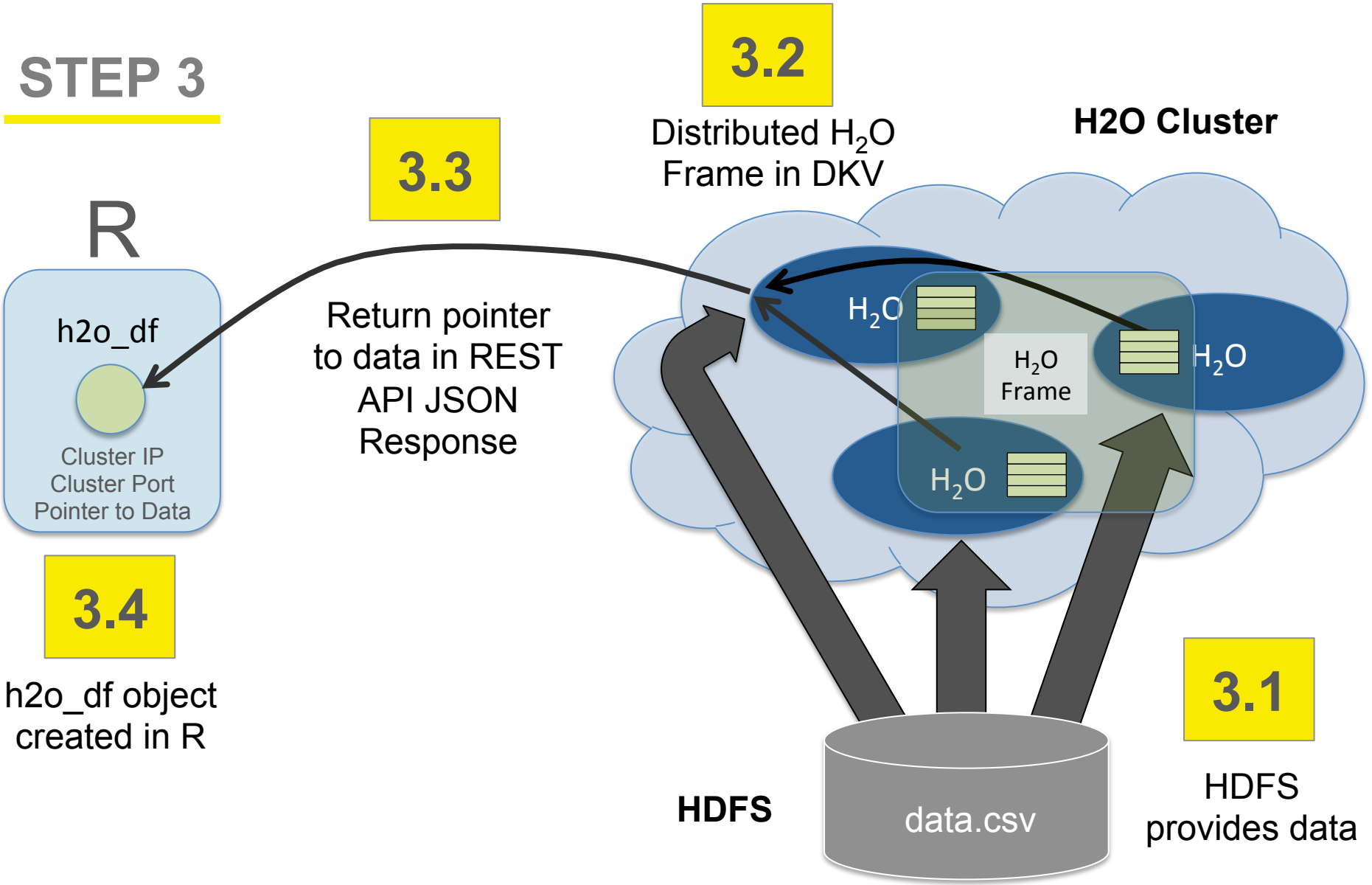
h2o_df = h2o.importFile("hdfs://path/to/data.csv")

R user

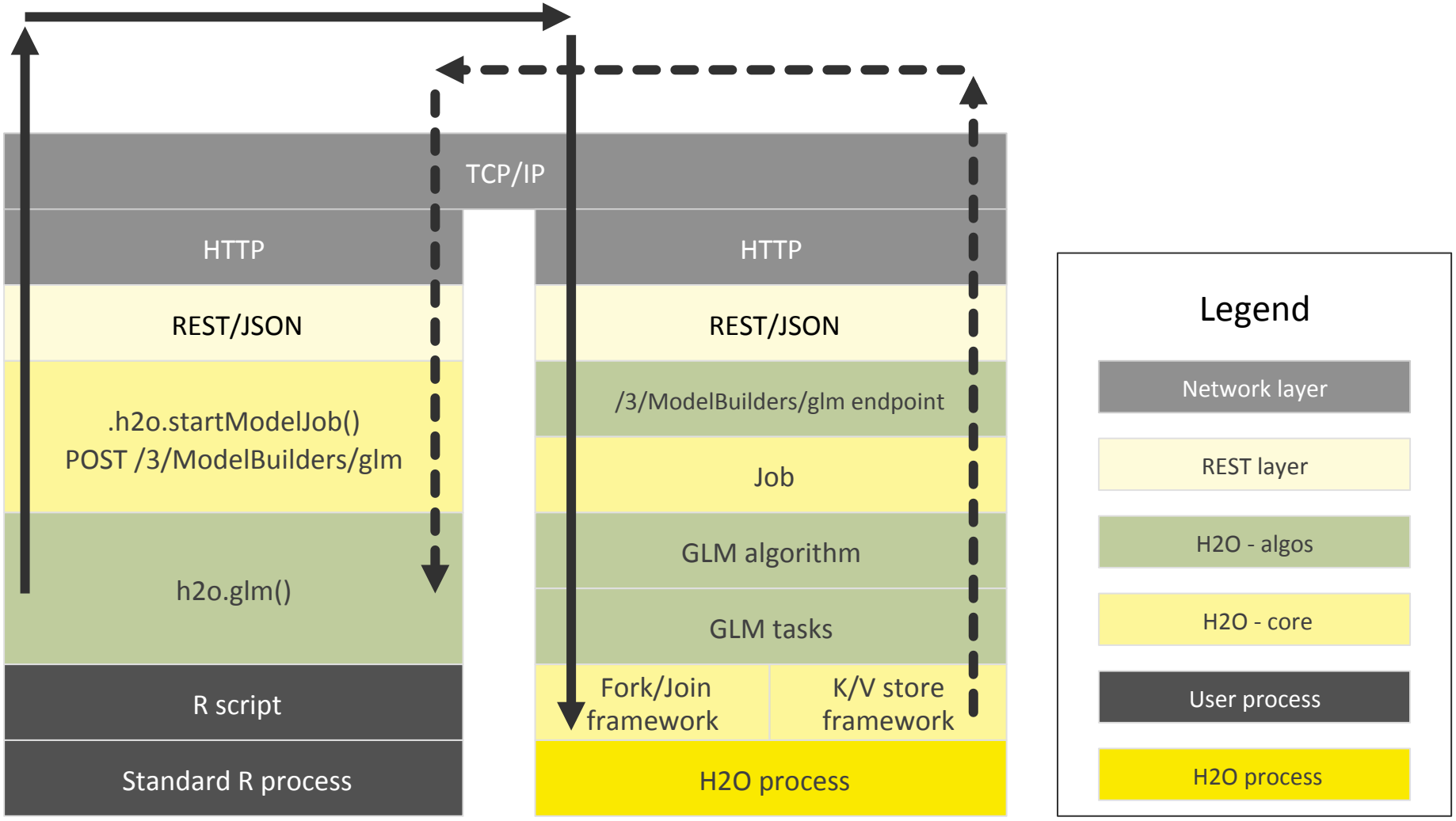
Reading Data from HDFS into H2O with R



Reading Data from HDFS into H2O with R



R Script Starting H2O GLM



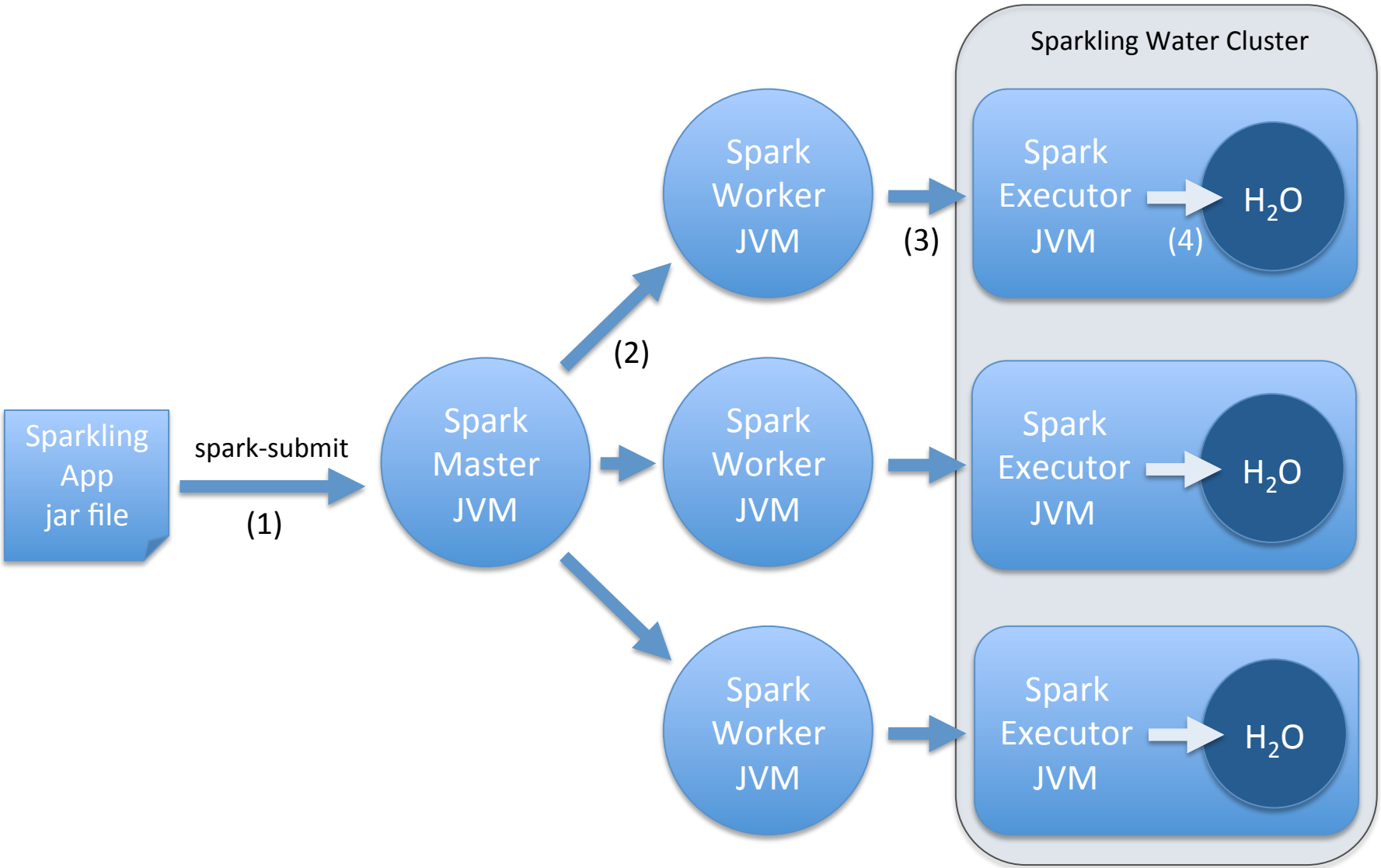
Demo of Consumer Loan App

Demo of iPython Notebook

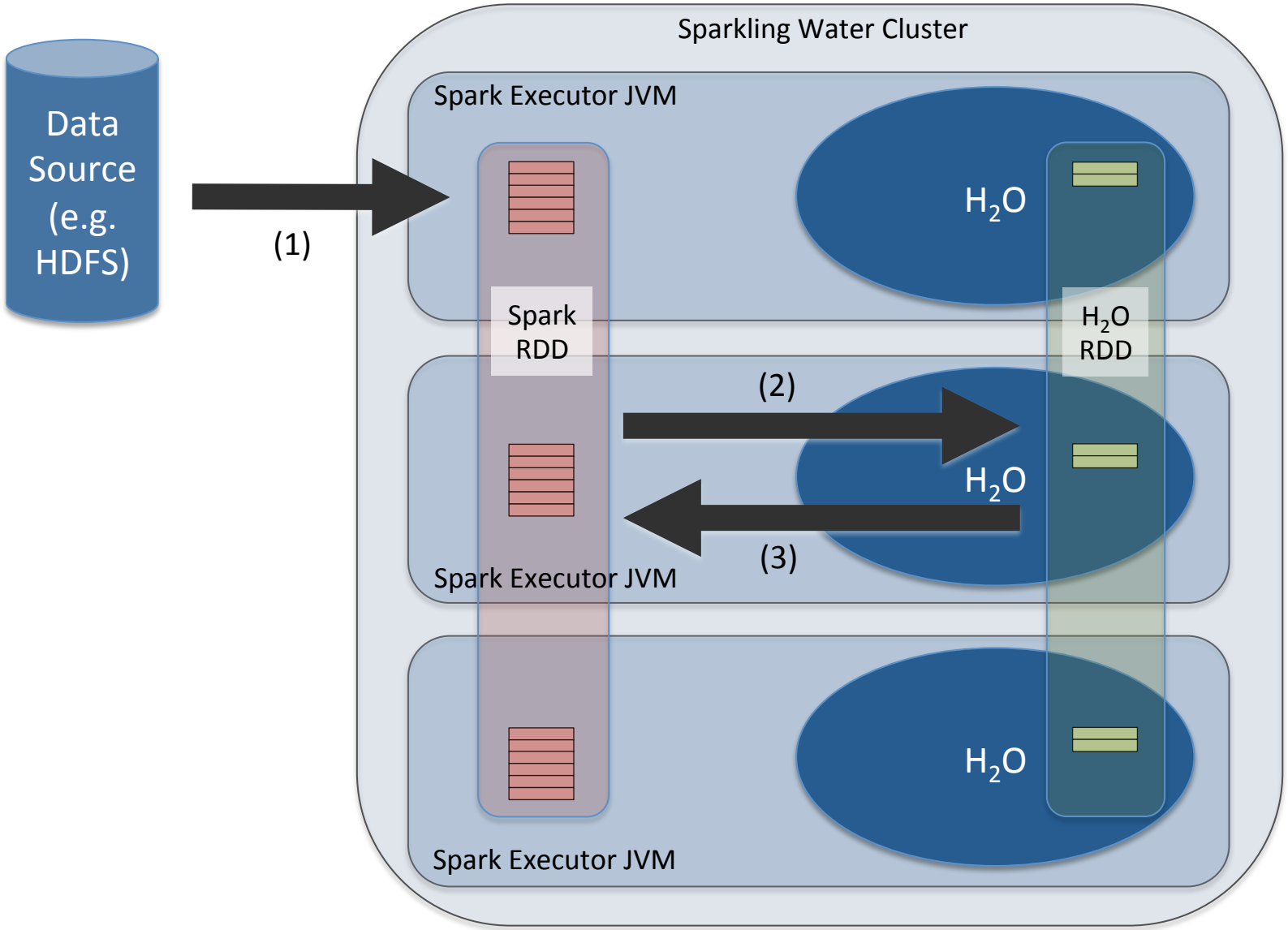
H_2O.ai

Sparkling Water
(H2O and Spark)

Sparkling Water Application Life Cycle



Sparkling Water Data Distribution



Demo of Ask Craig App

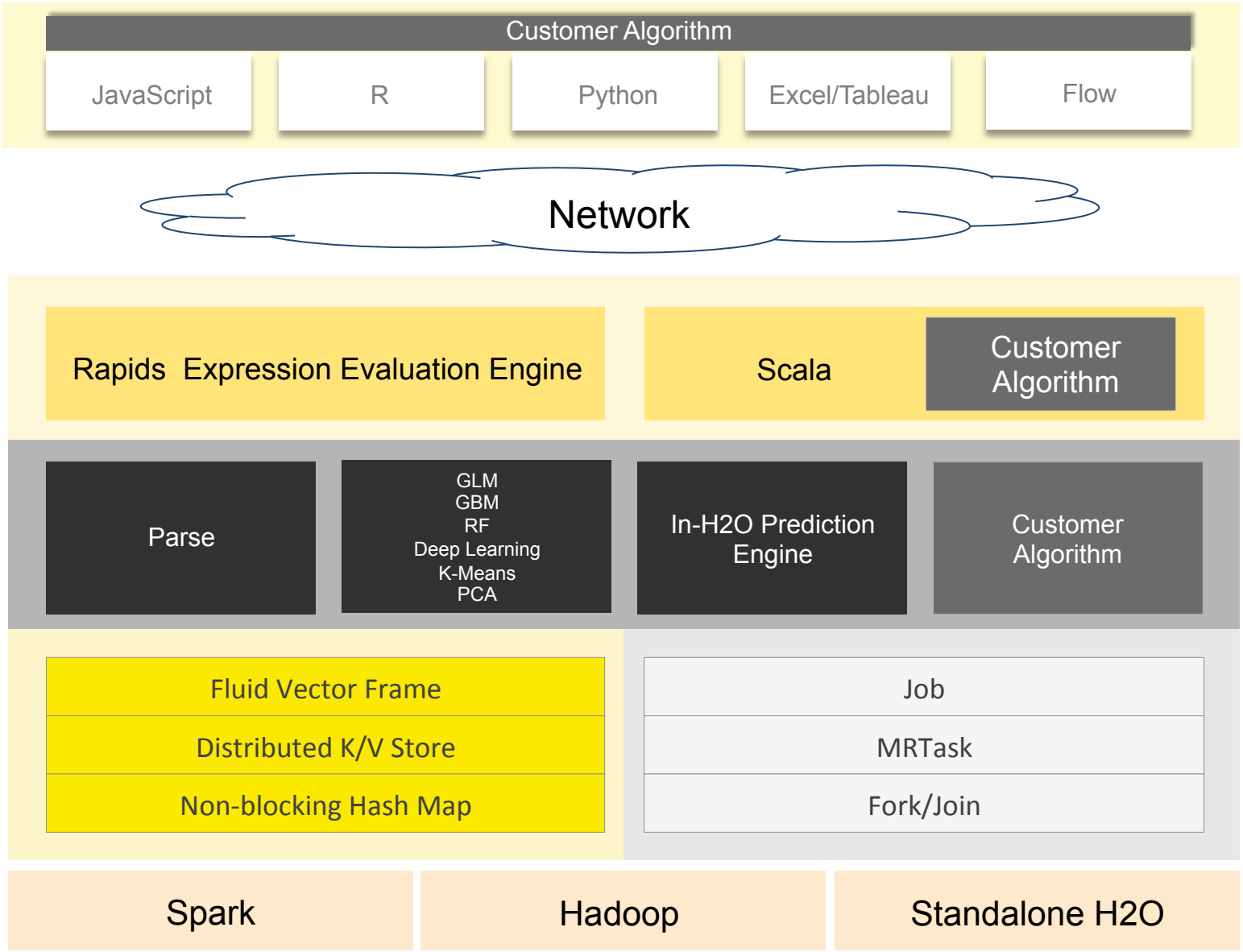
Q & A

Thanks for attending!

Content for today's talk can be found at:

https://github.com/h2oai/h2o-meetups/tree/master/2015_10_22_H2O_LV

H2O Software Stack



R Script Retrieving H2O GLM Result