H_2O.ai

Machine Learning for Smarter Applications

Tom Kraljevic
January 28, 2015
Jacksonville, FL

Outline for today's talk

- About H2O.ai and H2O (10 minutes)
- H2O in Big Data Environments (10 minutes)
- How H2O Processes Data (10 minutes)
- Building Smarter Apps with H2O (20 minutes)
- (Demo) Storm App w/ Streaming Predictions (10 minutes)
- Q & A (20 minutes)

Content for today's talk can be found at:

https://github.com/h2oai/h2o-meetups/tree/master/2015_01_28_MLForSmarterApps

H2O.ai Overview

- Founded: 2011 venture-backed, debuted in 2012
- Product: H2O open source in-memory prediction engine
- Team: 30
- 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!

H$_2$O.ai
Machine Intelligence

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

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

Algorithms on H_2O

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

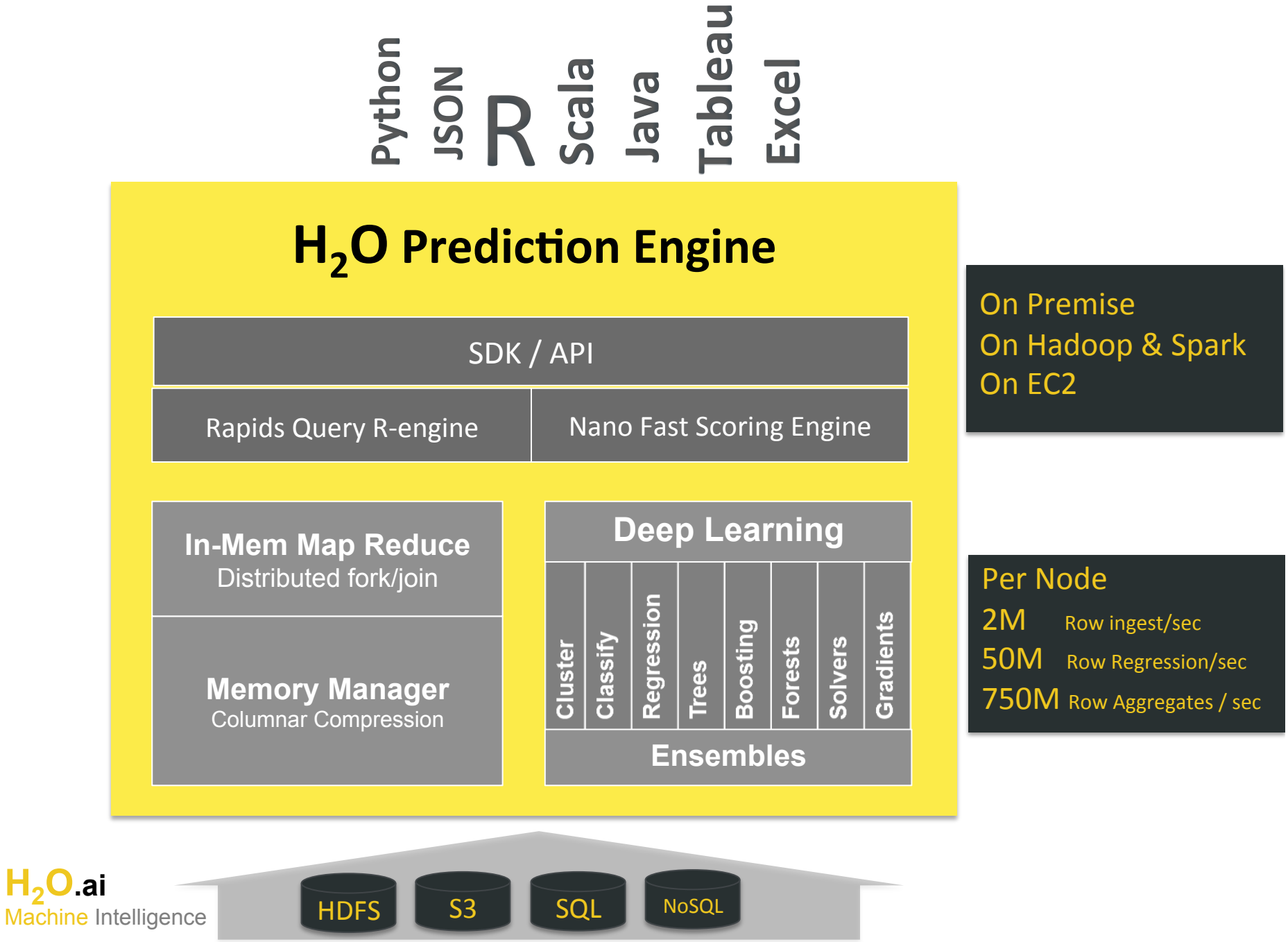
Deep Neural Networks

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

Algorithms on H_2O

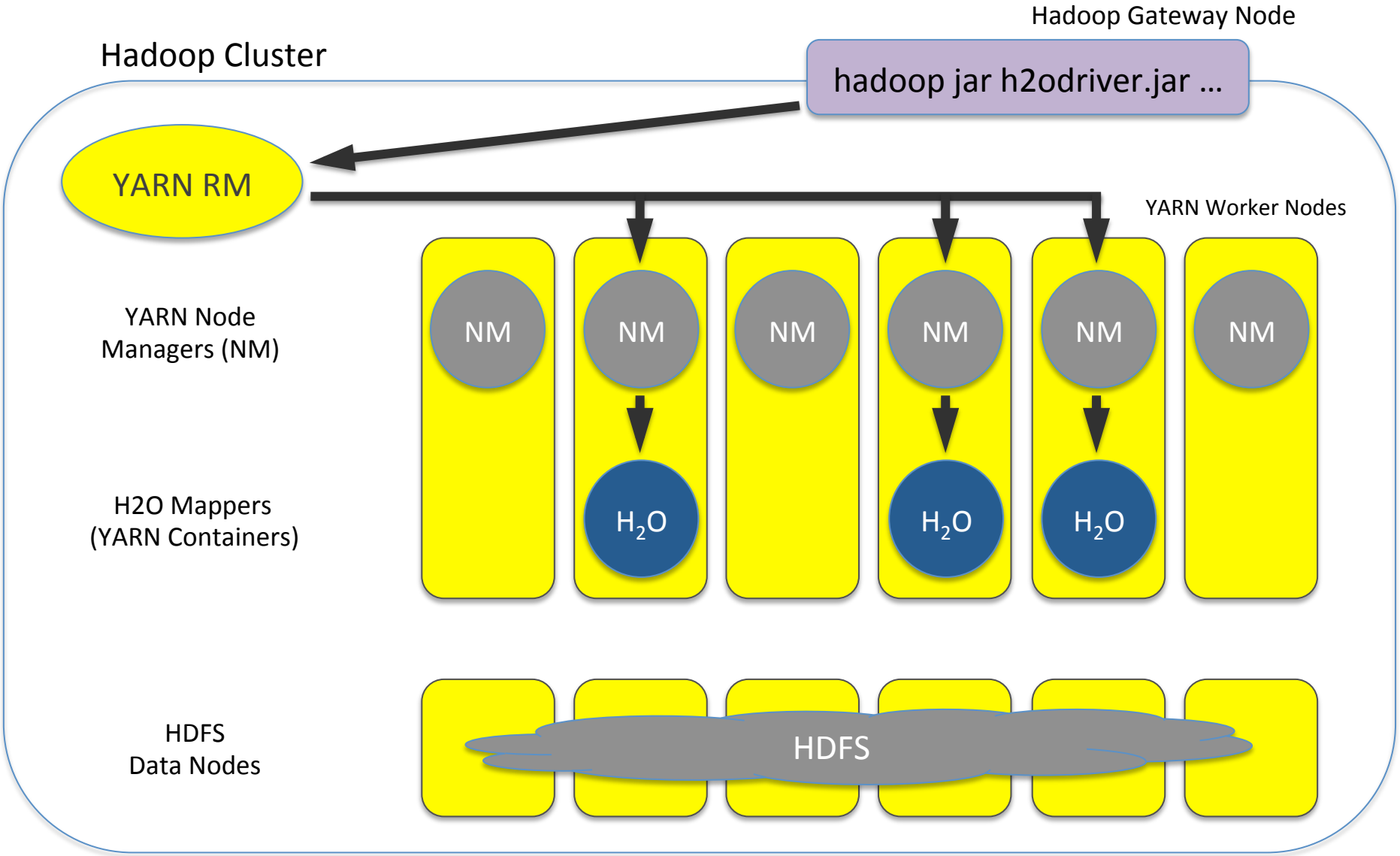
Unsupervised Learning

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

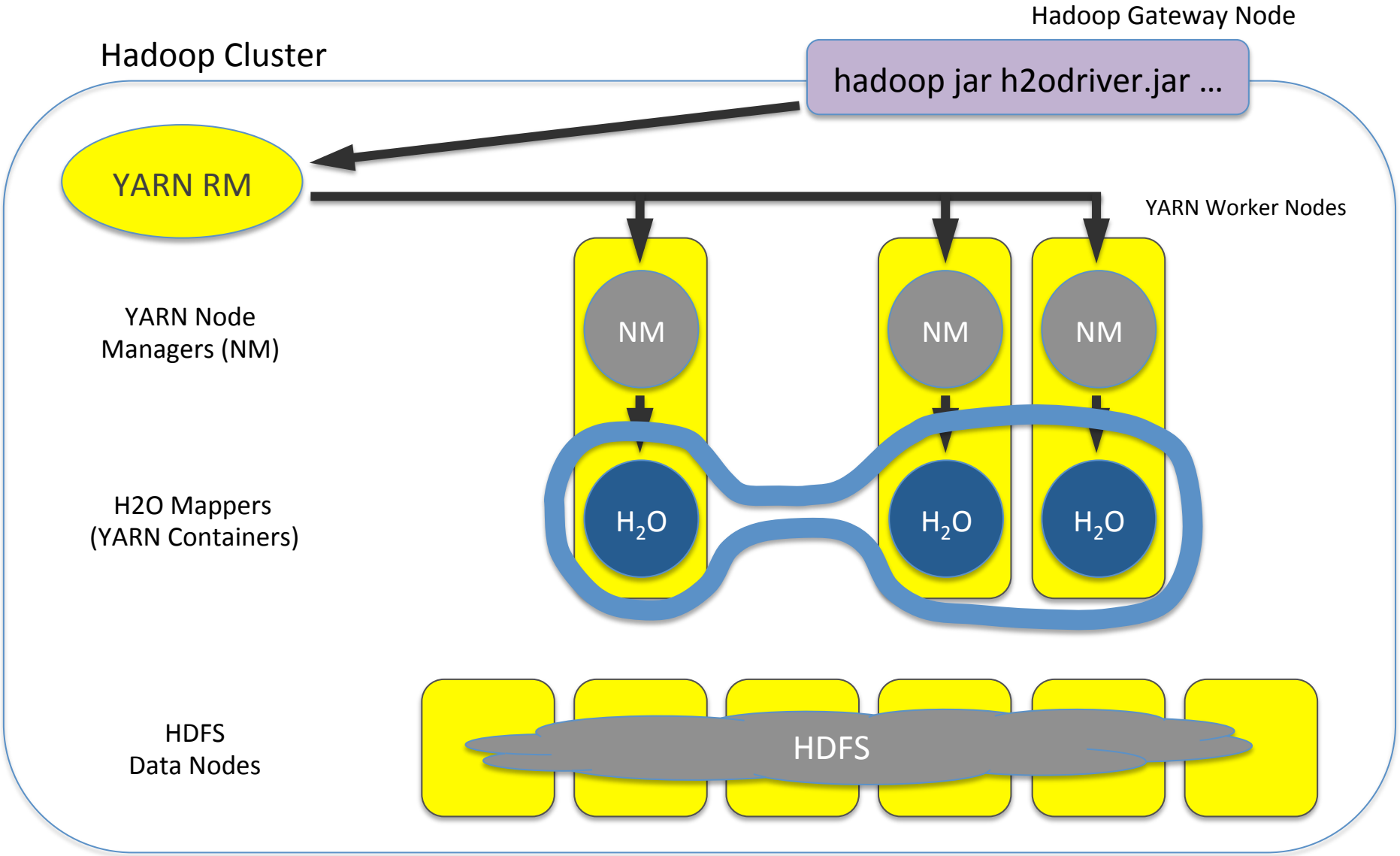

Python
JSON
R
Scala
Java
Tableau
Excel
H2O Prediction Engine
SDK / API
Rapids Query R-engine
Nano Fast Scoring Engine
In-Mem Map Reduce
Distributed fork/join
Memory Manager
Columnar Compression
Deep Learning
Cluster
Classify
Regression
Trees
Boosting
Forests
Solvers
Gradients
Ensembles
On Premise
On Hadoop & Spark
On EC2
Per Node
2M Row ingest/sec
50M Row Regression/sec
750M Row Aggregates / sec
HDFS
S3
SQL
NoSQL
H2O.ai
Machine Intelligence

H2O in Big Data Environments

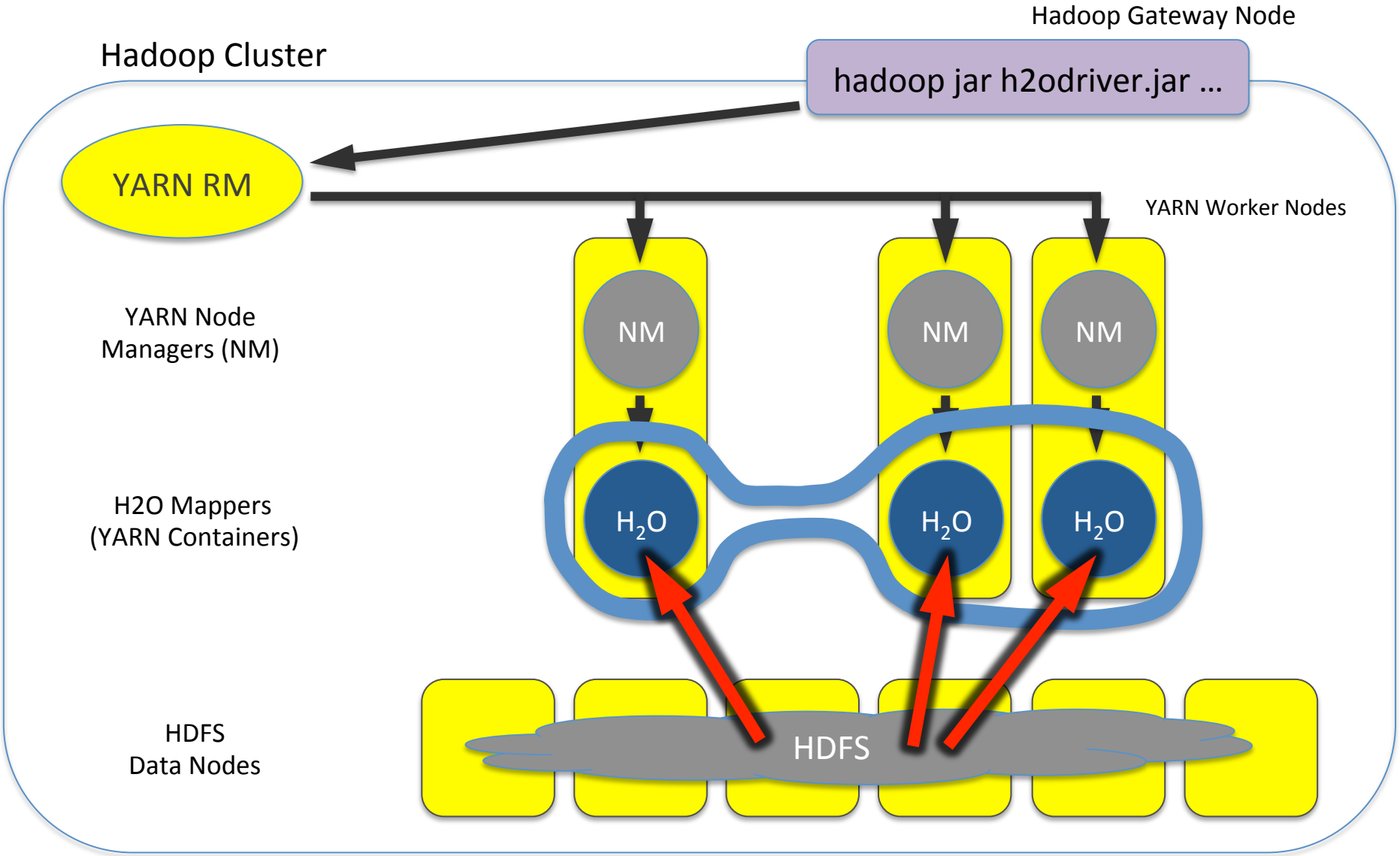
H2O on YARN Deployment



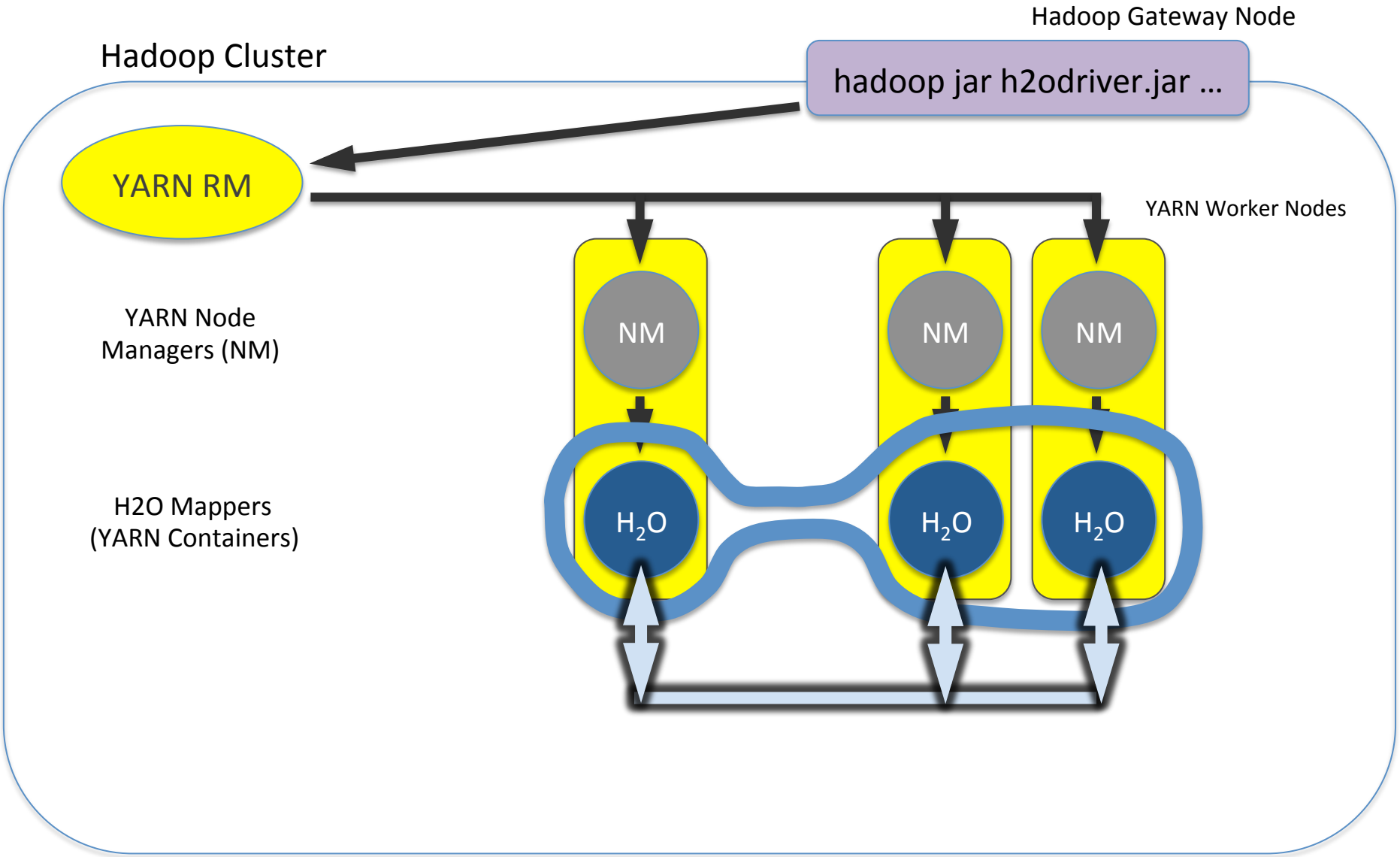
Now You Have an H2O Cluster



Read Data from HDFS *Once*

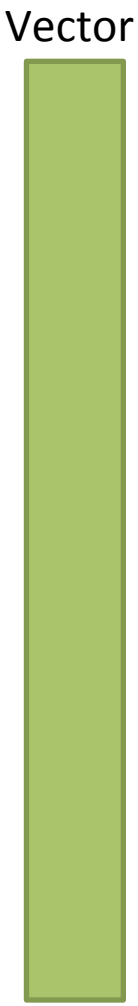


Build Models *in-Memory*

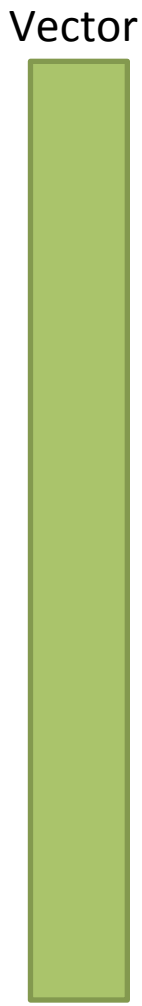


How H2O Processes Data

Distributed Data Taxonomy



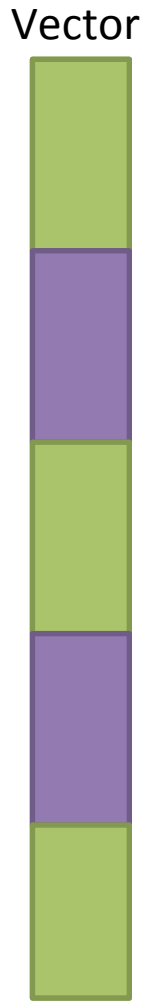
Distributed Data Taxonomy



The vector may be very large
(billions of rows)

- Stored as a compressed column (often 4x)
- Access as Java primitives with
  on-the-fly decompression
- Support fast Random access
- Modifiable with Java memory semantics

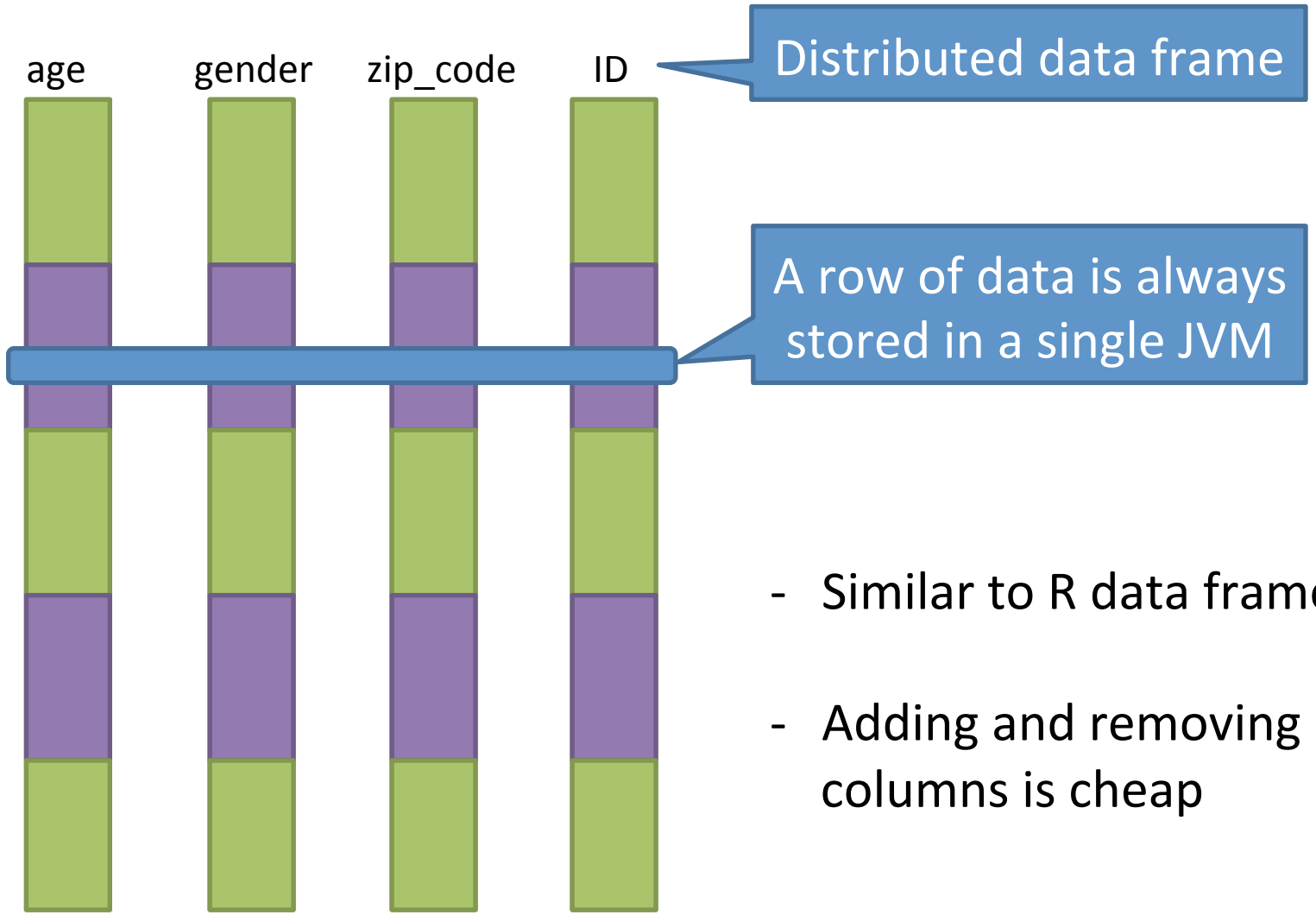
Distributed Data Taxonomy



Large vectors must be distributed over multiple JVMs

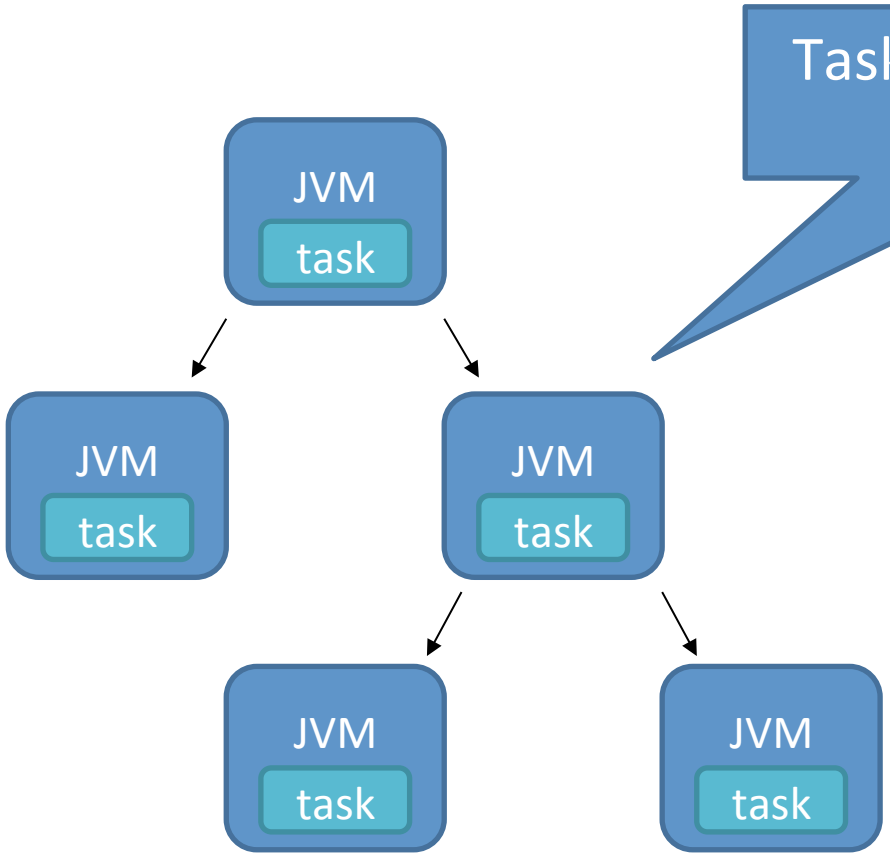
- Vector is split into chunks
- Chunk is a unit of parallel access
- Each chunk ~ 1000 elements
- Per-chunk compression
- Homed to a single node
- Can be spilled to disk
- GC very cheap

Distributed Data Taxonomy



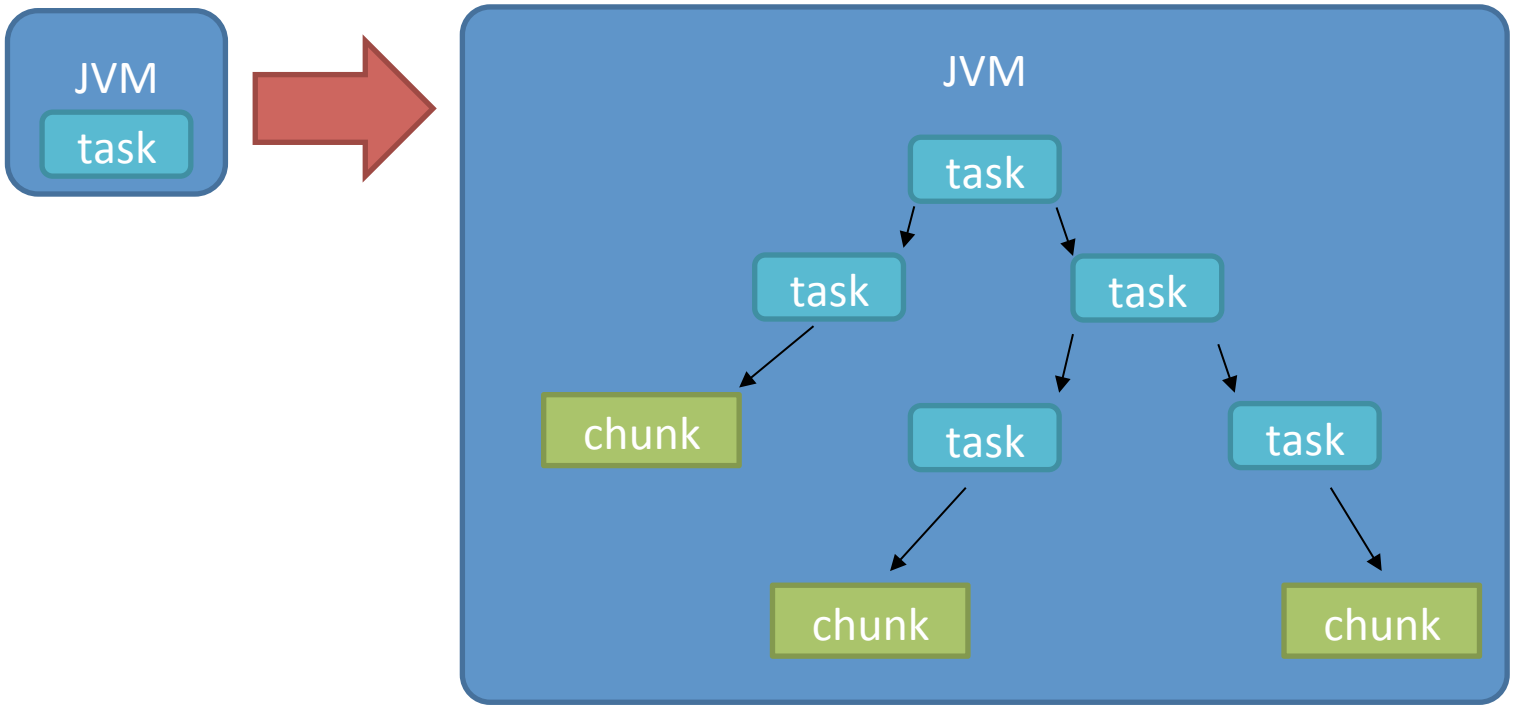
- Similar to R data frame
- Adding and removing columns is cheap

Distributed Fork/Join



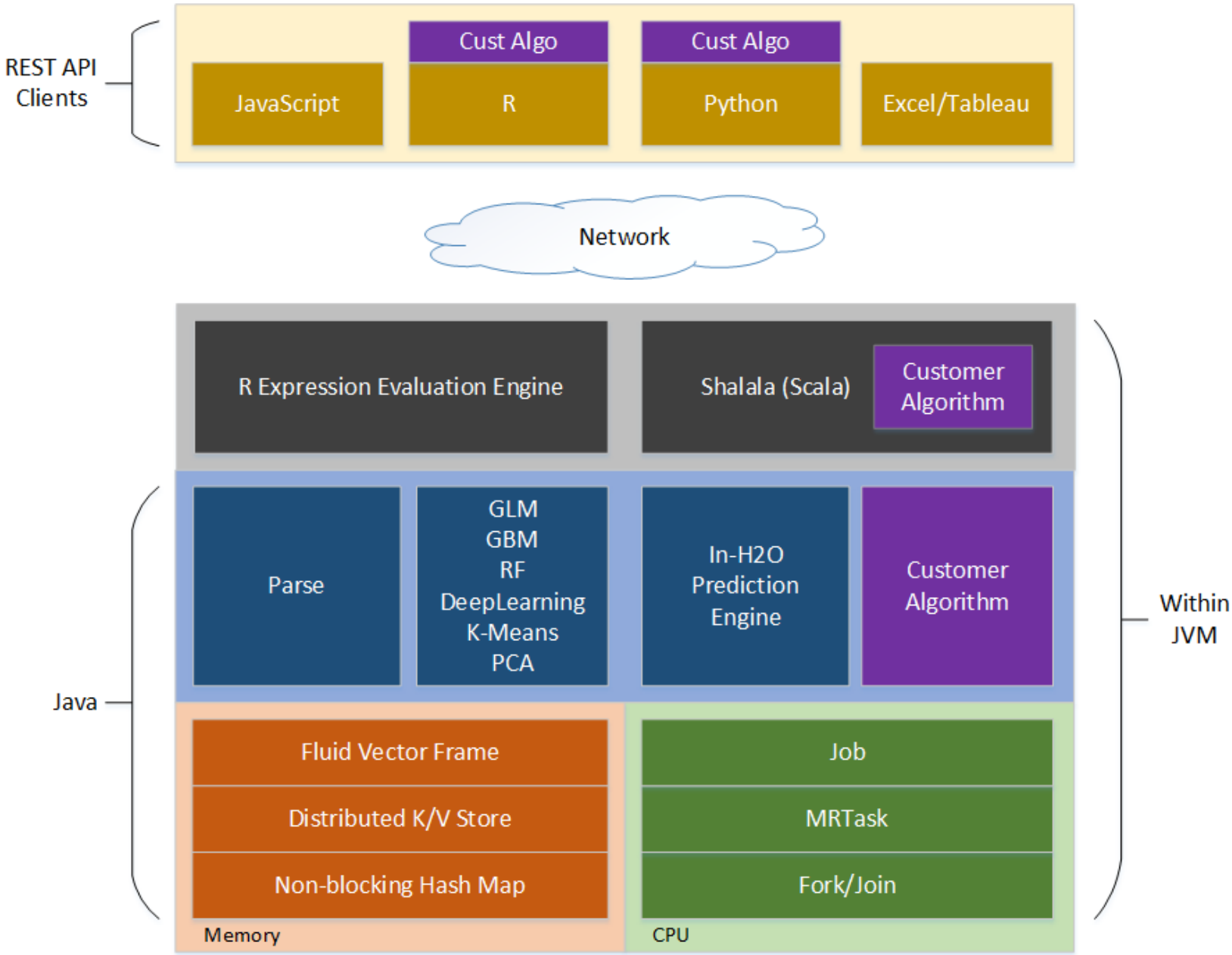
- Results are reduced at each inner node
- Returns with a single result when all subtasks done

Distributed Fork/Join

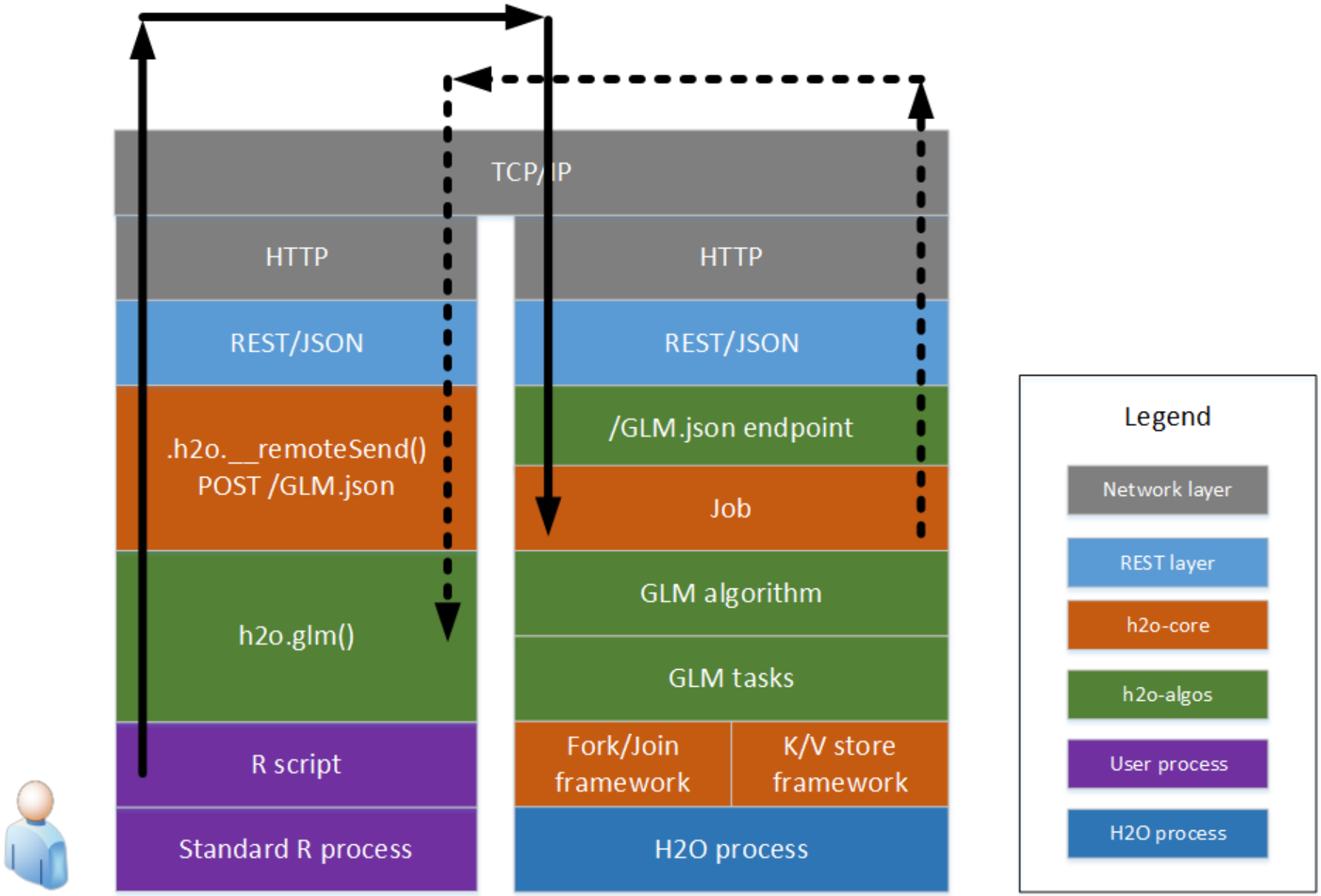


On each node the task is parallelized using Fork/Join

H2O Software Stack



R Script Starting H2O GLM



H2O.ai

CISCO

Propensity to Buy modeling factory

What is a Propensity To Buy model?

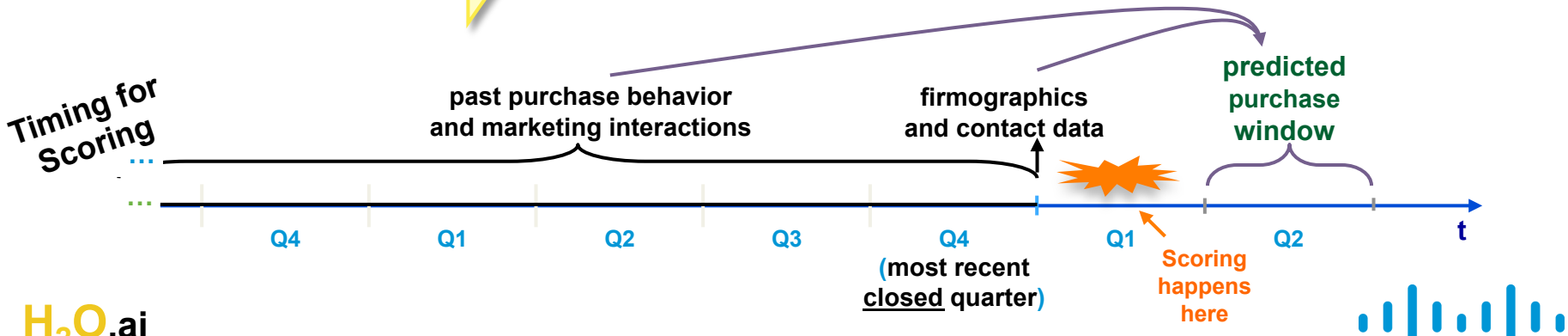
A Propensity To Buy model (P2B) is a statistical model that tries to predict whether a certain company will buy a certain product in a given time frame in the future.

Historical
- Firmographics
- Past Purchase Behavior
- Contacts (# and types)
- Marketing Interactions
- Cust Sat surveys
- Macroeconomic indicators
- Purchase / Non Purchase

Latest
- Firmographics
- Past Purchase Behavior
- Contacts (# and types)
- Marketing Interactions
- Customer satis. surveys
- Macroeconomic indicators

Build Statistical Model

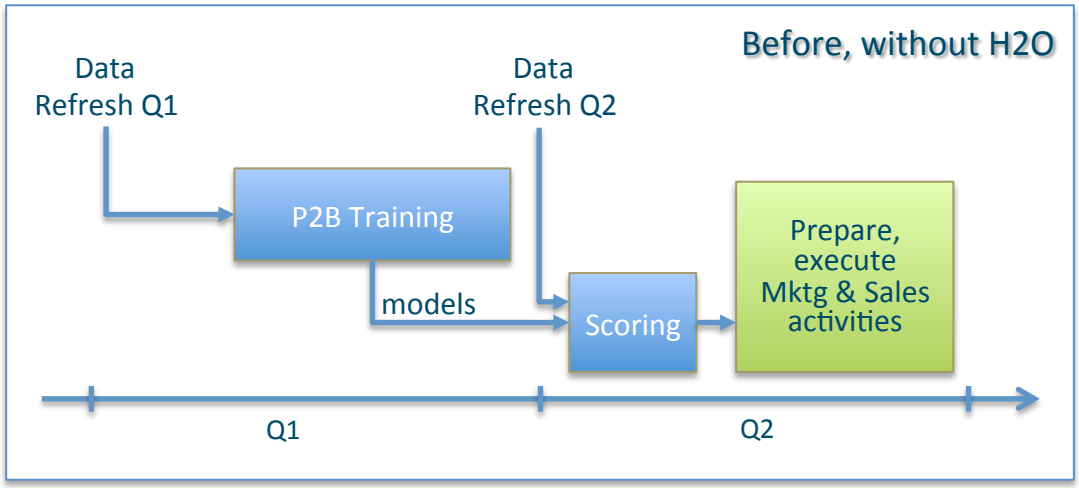
- **SCORE: probability that a company will buy a specific technology in the next quarter**
- **VALUE: bookings amount that Cisco will likely see IF the company in fact buys the technology**



H2O.ai
Machine Intelligence

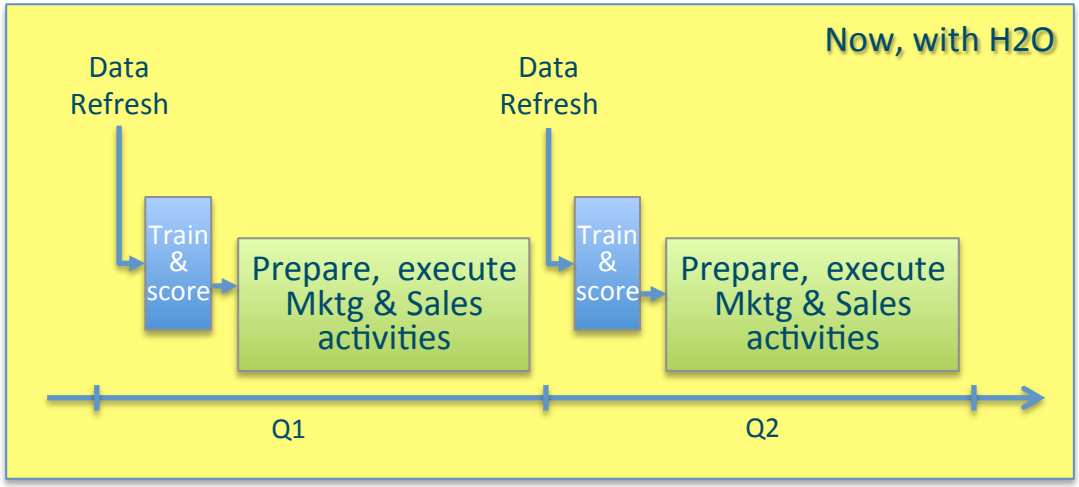
CISCO

H2O Business Benefit with P2B



Without H2O:

- Models needed to be prepared in advance, not to delay scoring
- More time preparing models, less time left for using the scores in the sales activities



With H2O:

- Newer buying patterns incorporated immediately into models
- Scores are published sooner
  - More time for planning and executing activities

H2O.ai
Machine Intelligence

CISCO

Results & Lessons Learned

Improvements

- P2B factory is 15x faster with H2O
- Quicker techniques for simpler problems, deeper for harder ones (grid searches!)
- Ensembles improved accuracy and stability of models significantly

Lessons Learned

- Even with few nodes speed improvement over traditional data mining tools is substantial
- H2O becomes really powerful and robust when combined with R
- Rely on H2O's extremely responsive support

H2O.ai

Fraud prevention using Deep Learning

Fraud Prevention at PayPal

Transaction Level	• Employs state of the art machine learning and statistical models to flag fraudulent behavior upfront • More sophisticated algorithms after transaction complete
Account Level	• Monitor activity to identify abusive behavior • Frequent payments, suspicious profile changes
Network Level	• Monitor account to account interaction • Frequent transfer of money from sever accounts to one central account suspicious

H2O.ai
Machine Intelligence

PayPal

Fraud Prevention at PayPal

Why Deep Learning?

- Unearth low-level complex abstractions
- Learn complex, highly varying functions not present in the training examples
- Widely employed for image/video processing and object recognition

Why H2O?

- Highly scalable
- Superior performance
- Flexible deployment
- Works seamlessly with other big data frameworks
- Simple interface

Experiment

- Dataset
  - 160 million records
  - 1500 features (150 categorical)
  - 0.6TB compressed in HDFS
- Infrastructure
  - 800 node Hadoop (CDH3) cluster
- Decision
  - Fraud/not-fraud

H2O.ai
Machine Intelligence

PayPal

Conclusions using H2O

Deep Learning with H2O is beneficial for payment fraud prevention

- Network architecture- 6 layers with 600 neurons each performed the best
- Activation function
  - RectifierWithDropout performed the best
- Improved performance with limited feature set & a deep network
  - 11% improvement with a third of the original feature set, 6 hidden layers, 600 neurons each
- Robust to temporal variations

H_2O.ai

Bordeaux Wines

Unsupervised Learning & Bordeaux Wine

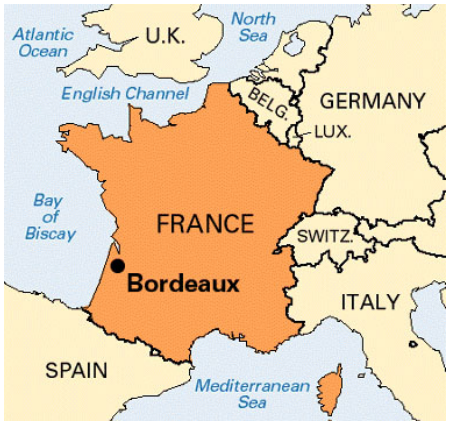
- A personal and expen$$$ive obsession
- Alex Tellez - h2o.ai

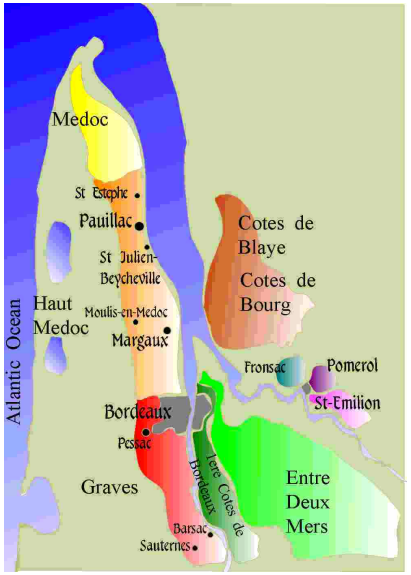
H2O.ai

BORDEAUX WINE

Largest wine-growing region in France

+ 700 Million bottles of wine produced / year !





Some years better than others: Great ($$$) vs. Typical ($)

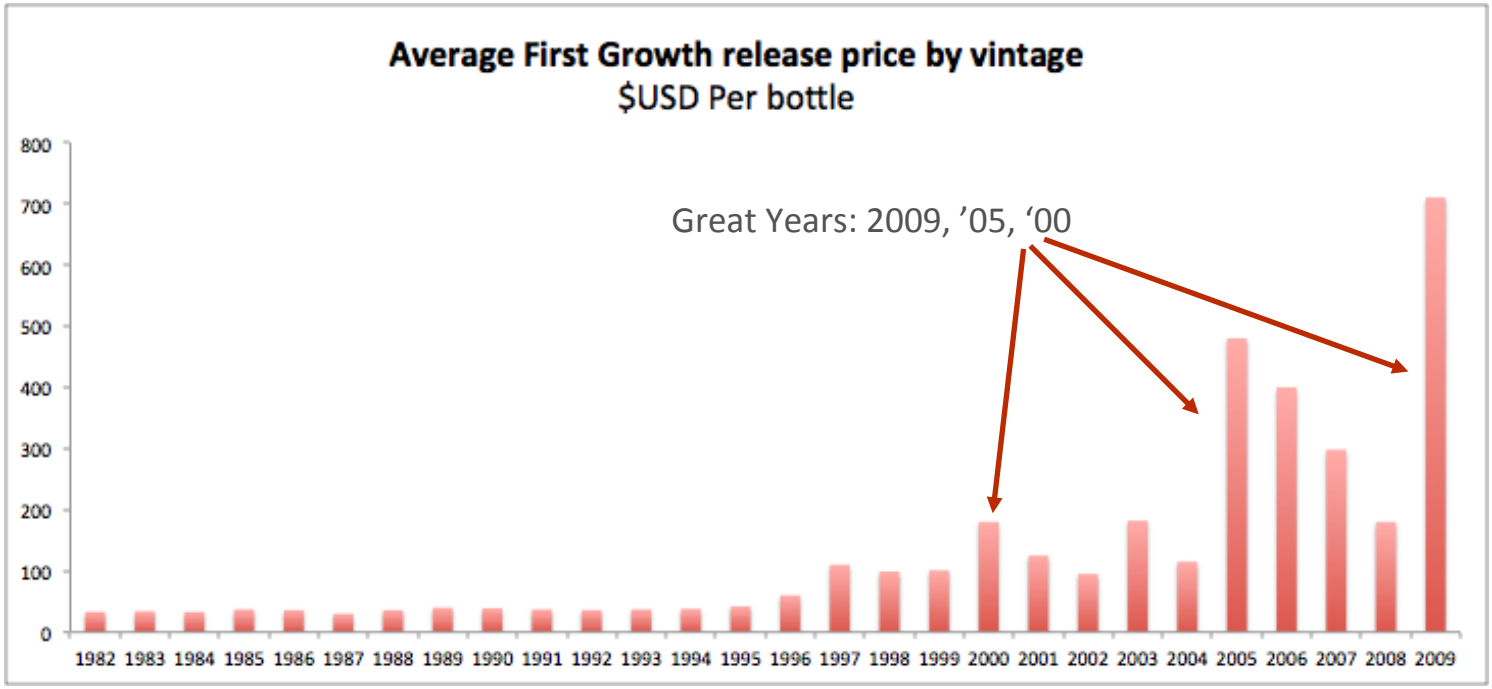
Last Great years: 2010, 2009, 2005, 2000

H2O.ai

'EN PRIMEUR'

While wine is still barreled, purchasers can 'invest' in the wine *before* bottling and official public release.

Advantage: Wines may be considerably cheaper during *'en primeur'* period than @ bottling.



H2O.ai

GREAT VS. TYPICAL VINTAGE?

Question:

Can we study weather patterns in Bordeaux leading up to harvest to identify 'anomalous' weather years >> correlates to Great ($$$) vs. Typical ($) Vintage?

The Bordeaux Dataset (1952 - 2014 Yearly)

Amount of Winter Rain (Oct > Apr of harvest year)

Average Summer Temp (Apr > Sept of harvest year)

Rain during Harvest (Aug > Sept)

Years since last Great Vintage

H$_2$O.ai

AUTOENCODER + ANOMALY DETECTION

In Steps:

1) Train deep autoencoder to learn 'typical' vintage weather pattern
2) Append 'great' vintage year weather data to original dataset
3) *IF great* vintage year weather data does NOT match learned weather pattern, autoencoder will produce high reconstruction error (MSE)

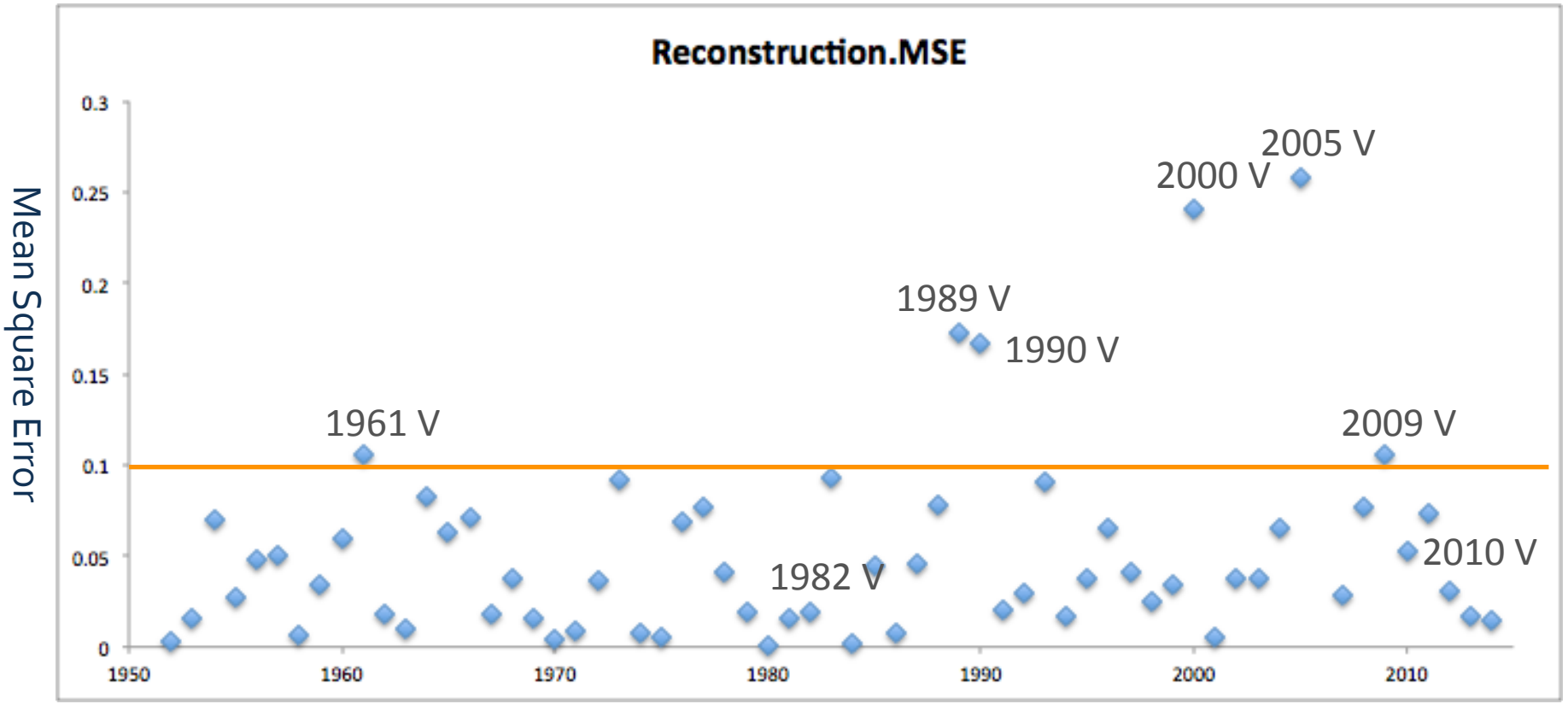
Goal:

'en primeur of en primeur'

- Can we use weather patterns to identify anomalous years >> indicates great vintage quality?

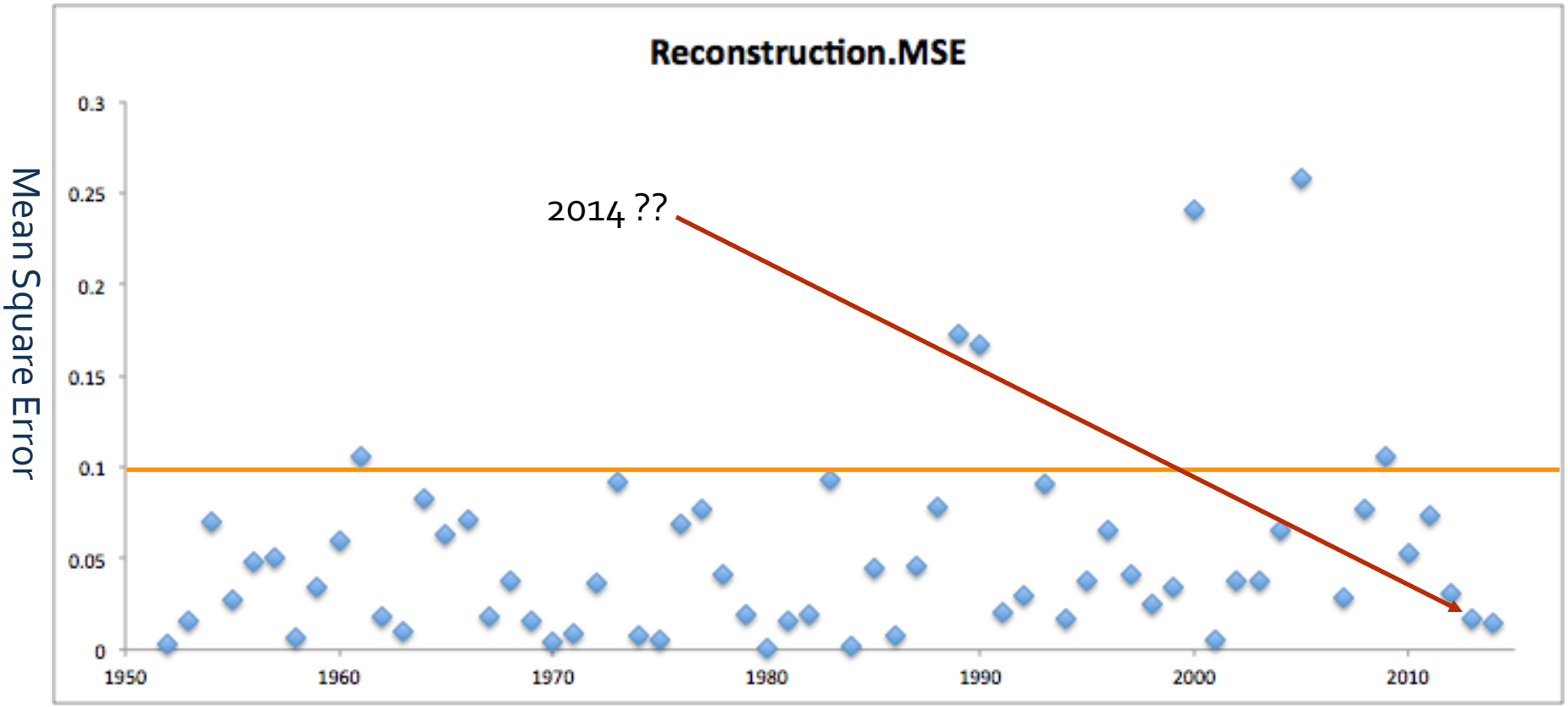
H2O.ai

RESULTS (MSE > 0.10)



H2O.ai

2014 BORDEAUX??



H2O.ai

ROBERT PARKER JR.

“The world’s most prized palette” - Financial Times

H2O.ai

DESCRIBING WINES

=

…wet forest floor
…decaying animal skin
…cat pee
…grandmother's closet
…barnyardy (AKA sh*t)
…asian plum sauce
…*chewy tannins*

Can we use Parker reviews to recommend wines we would li

H2O.ai

WINE RECOMMENDATIONS

I like: 2000 Cheval Blanc (Red)

Robert Parker Tasting Notes:

"…sweet nose of menthol, melted licorice, boysenberry, blueberry, creme de cassis…sweet tannins w/ hints of coffee & earth"

So I'm at BevMo…now what?

H$_2$O.ai

APPROACH

In Steps:

1. Collect hundreds of Robert Parker reviews of wine

2. Create bag-of-words dataset & scale word counts between 0 & 1 (~ probability a particular word occurs in a wine review)

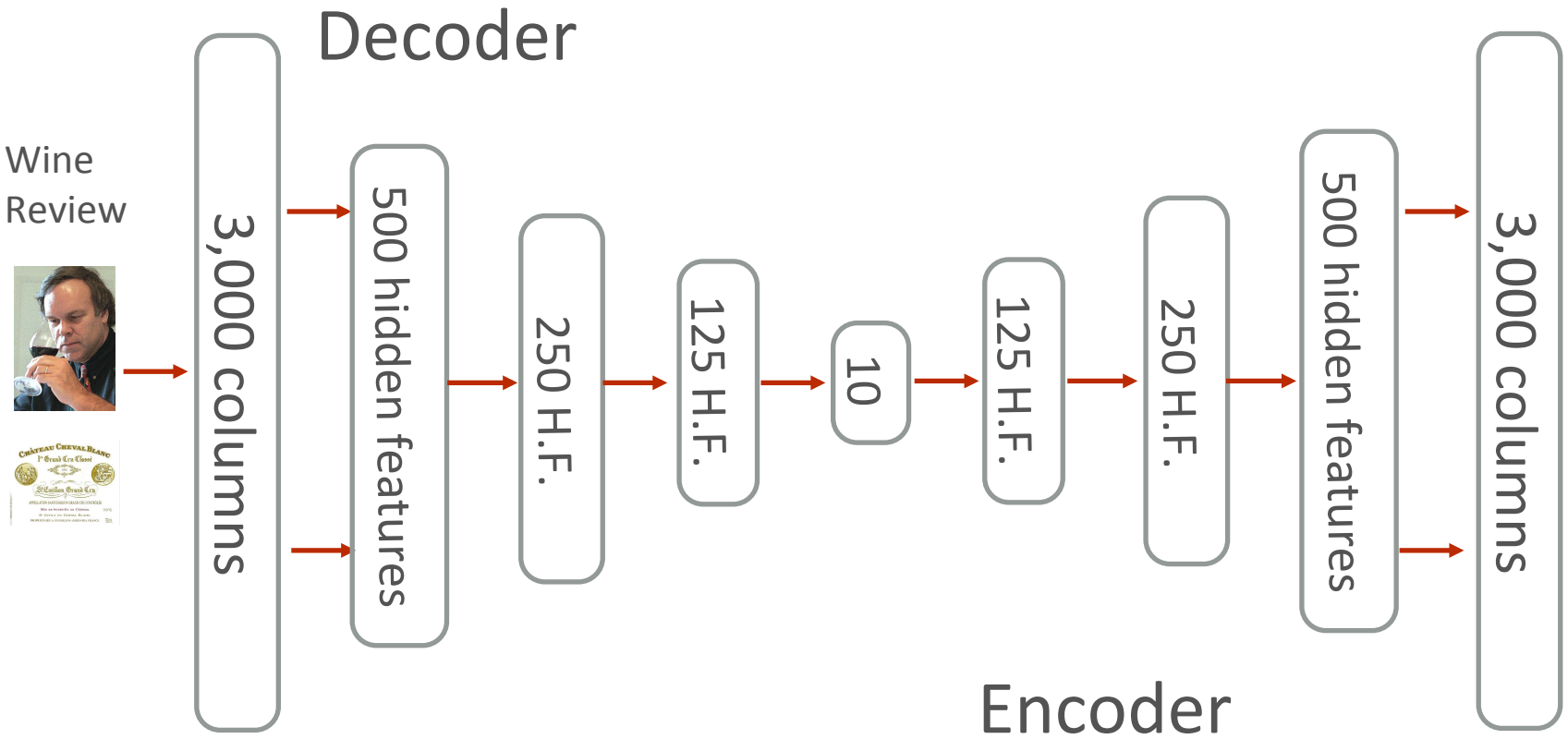
3. Build deep autoencoder from sparse bag-of-words matrix >> reduce to 10 'deep features'

4. Take Euclidian distance of wine I like (2000 Cheval Blanc) to other wines >> Information Retrieval

5. 'Nearby' wine review-vectors = recommendations!

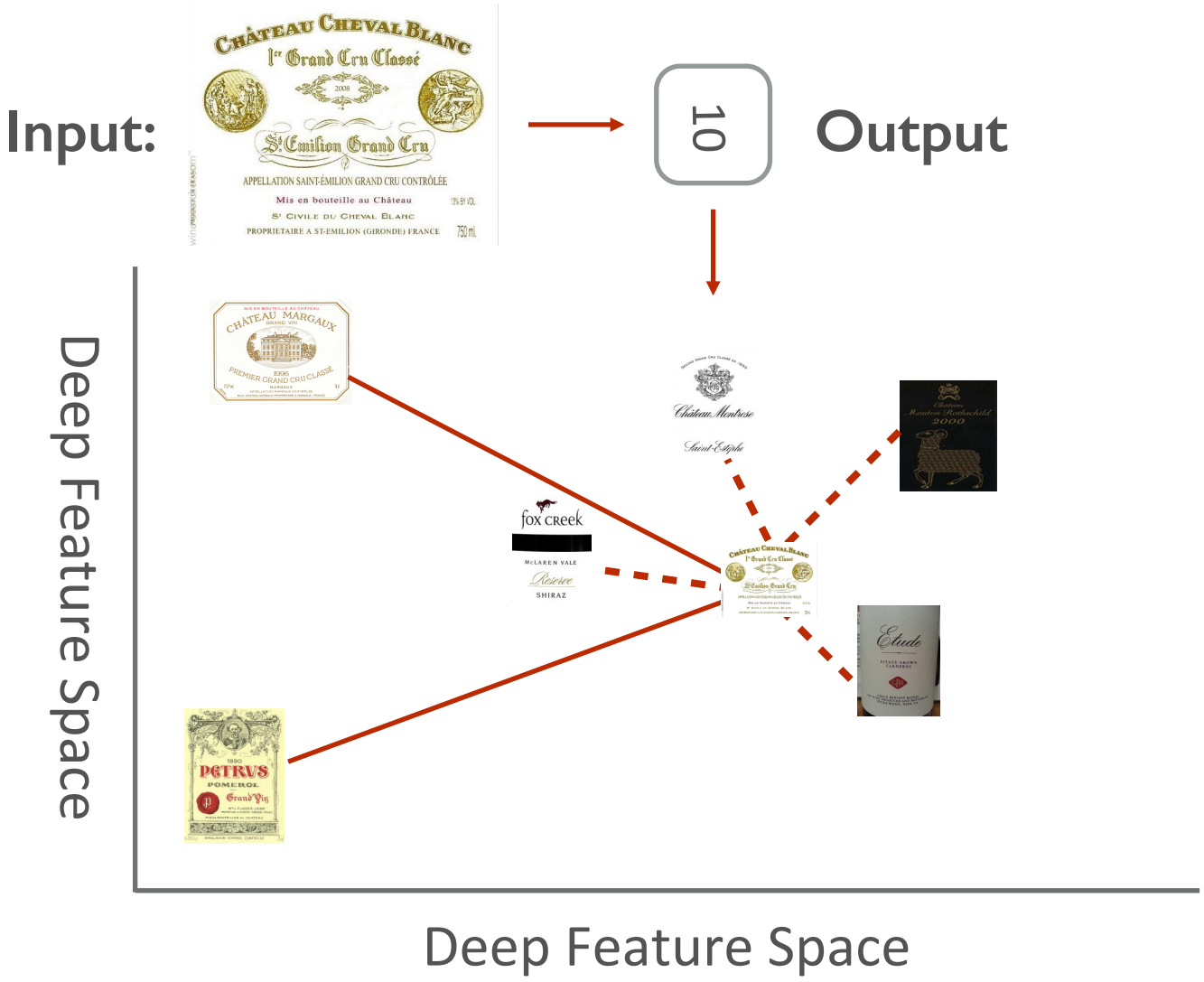
H_2O.ai

(HELLA) DEEP AUTOENCODER



H2O.ai

EUCLIDIAN DISTANCE



RESULTS

I like: 2000 Cheval Blanc

"…sweet nose of menthol, melted licorice, boysenberry, blueberry, creme de cassis…sweet tannins w/ hints of coffee & earth"

Similar Wines:

"… aromas of licorice, damp forest floor, blueberries & currant"

"…loads of cassis, coffee, earth & chocolatey notes"

"…massive blackberry, mulberry fruit intermixed with earth, crushed rocks"

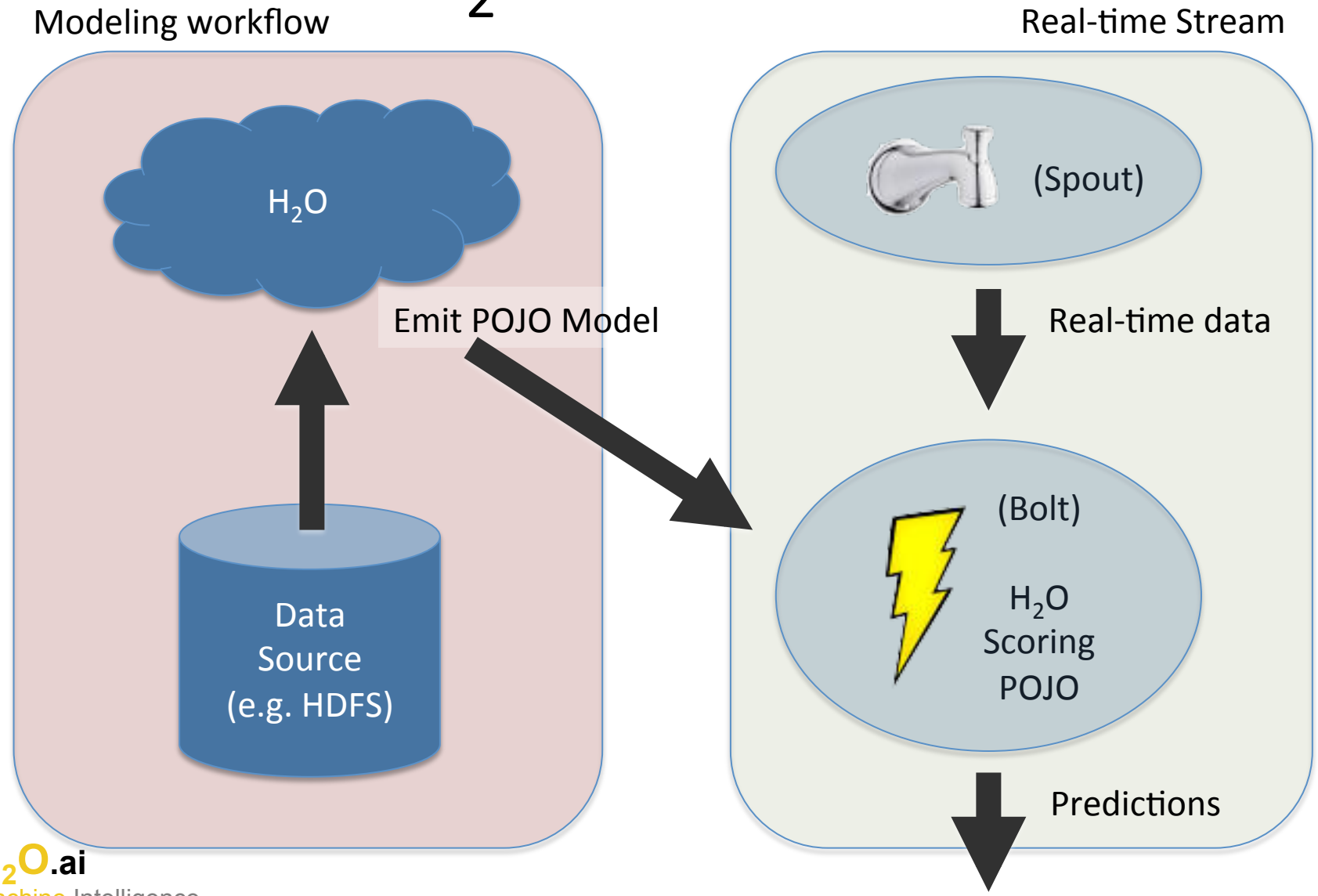
"…ethereal bouquet of menthol, coffee, wet stones, black cherries & blackberries."

H2O.ai

H_2O.ai

Real-time Streaming Predictions with H2O and Storm

H_2O on Storm



H_2O.ai
Machine Intelligence

Q & A

Thanks for attending!

Content for today's talk can be found at:

https://github.com/h2oai/h2o-meetups/tree/master/2015_01_28_MLForSmarterApps