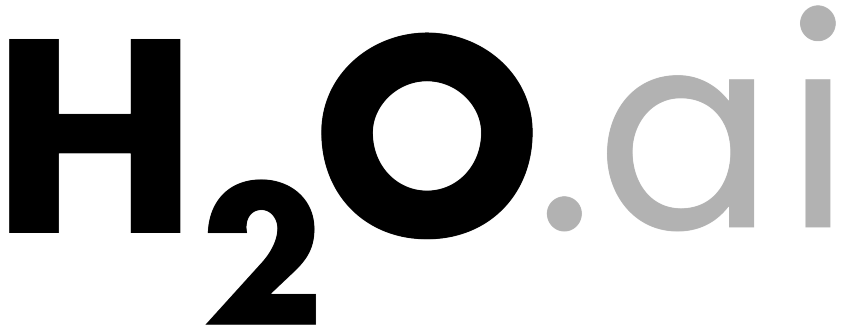
Introduction to Machine Learning with H_2O

- Introduction (20 Mins)
  - Company
  - H_2O Machine Learning Platform
- First Use Case (15 Mins)
  - Credit Card Dataset (Small CSV)
  - H_2O Quick Start (Web/R/Python)
- Second Use Case (20 Mins)
  - Freddie Mac Dataset (HDFS – 2.58 GB)
  - Higgs Boson Machine (HDFS – 7.48GB)
  - H_2O on Multi-Node Cluster
- Q & A (5 Mins)



Jo-fai (Joe) Chow
Data Scientist at H2O.ai
joe@h2o.ai

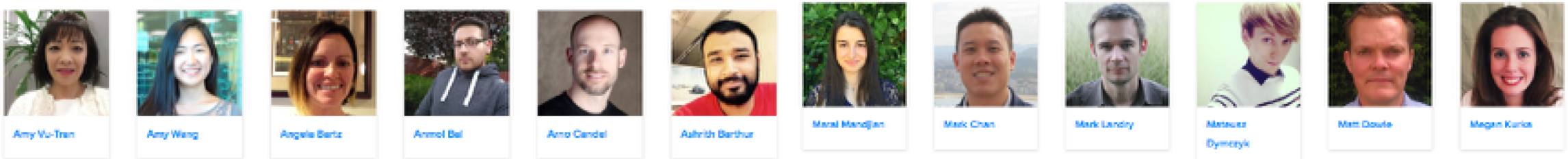
About H_2O.ai

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	• **H_2O Open Source In-Memory AI Prediction Engine** • Sparkling Water • Deep Water • Steam
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	70 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Headquarters	Mountain View, CA






Amy Vu-Tran
Amy Wang
Angela Bartz
Anmol Bal
Anna Candal
Ashrith Barthur
Matt Dowle
Megan Kurka
Avkash Chauhan
Beth Payne
Brandon Murray
Cori Andrews
Michal Kurka
Michal Malohlava
Nidhi Mehta
Nishant Kalonia
Desmond Chan
Dmitry Larko
Erin LeDell
Fonda Ingram
Ian Gomez
Jacqueline Scott
Pasha Stetsenko
Patrick Hall
Patrick Rice
Prithvi Prabhu
Ravi Purushotma
Raymond Peck
Jakub Hava
Jeff Fohl
Jeff Gambera
Jo-fai Chow
Jon Dalowski
Josephine Wang
Sebastian Vidrio
Tom Kraljevic
Justin Loyola
Kevin Kelly
Leland Wilkinson
Vinod Iyengar
Wen Phan
Wendy Wong

Joe (UK)

H2O.ai

Scientific Advisory Council

Dr. Trevor Hastie

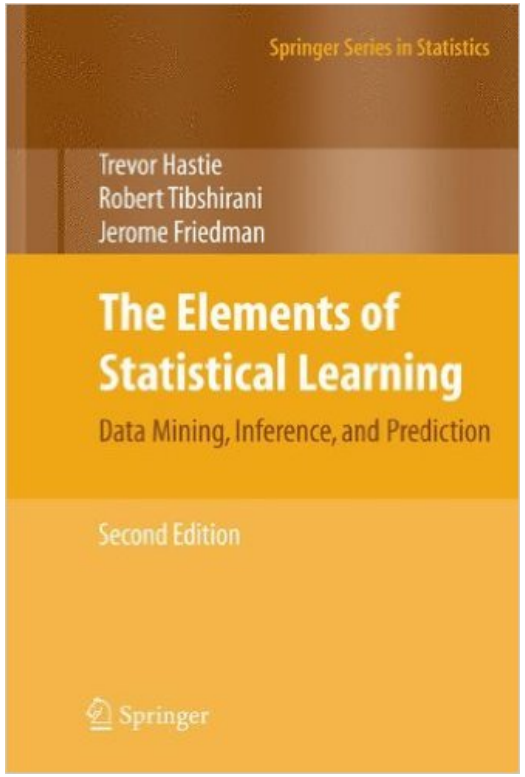
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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*

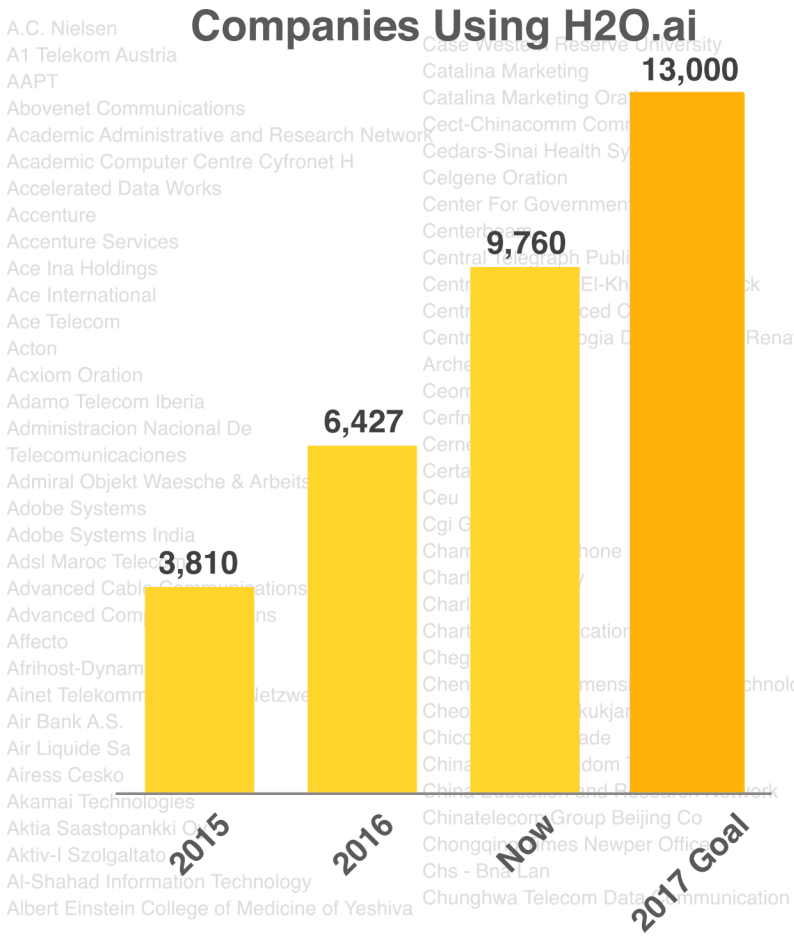
Dr. Steven Boyd

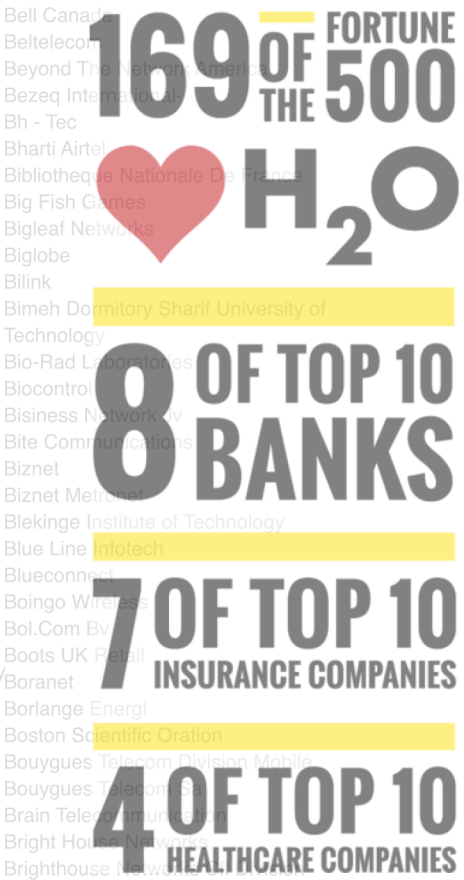
- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities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

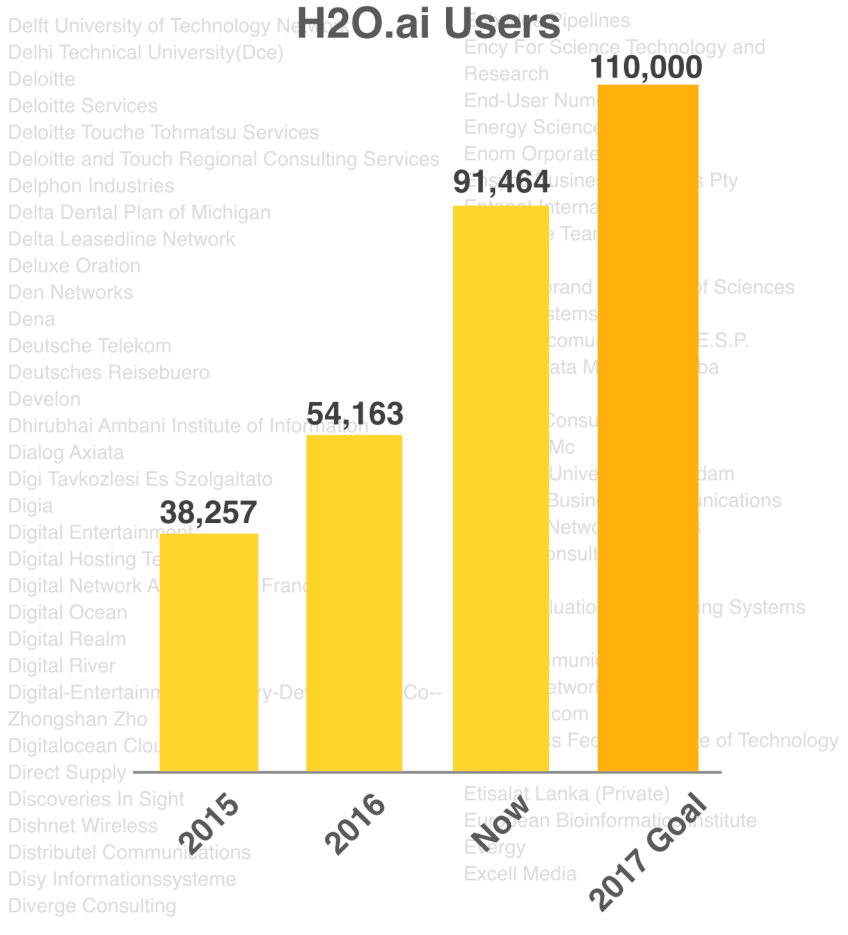


H2O.ai

H2O Community & Fortune 100 customers







Select Reference Customers:

"Overall customer satisfaction is very high." - Gartner

CapitalOne PayPal Experian WELLS FARGO ADP COMCAST at&t KAISER PERMANENTE PROGRESSIVE Nielsen Catalina Booking.com macy's H2O.ai

Select Customers

"Overall customer satisfaction is very high." - Gartner

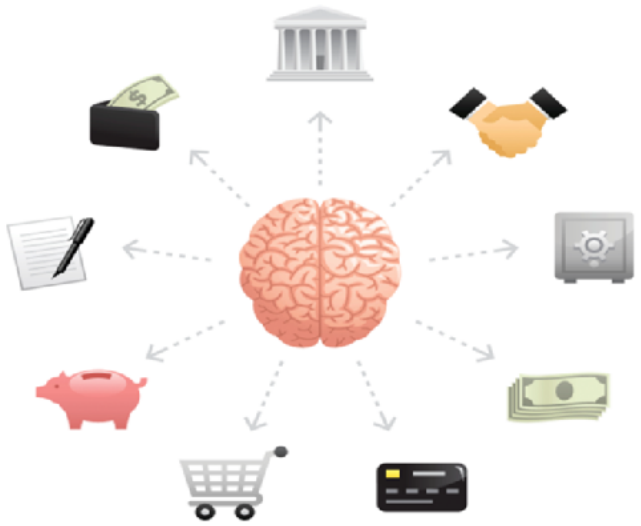
AI in Financial Services

Wholesale / Commercial Banking

- Know Your Customers (KYC)
- Anti-Money Laundering (AML)

Retail Banking

- Deposit Fraud
- Customer Churn Prediction
- Auto-Loan



Today's Use Case Examples

IT Infrastructure

- Security Cyberlake
- DoS Detection and Protection
- Master Data Management

Card/Payments Business

- Transaction Frauds
- Real-time Targeting
- **Credit Risk Scoring**
- In-Context Promotion

Community Expansion

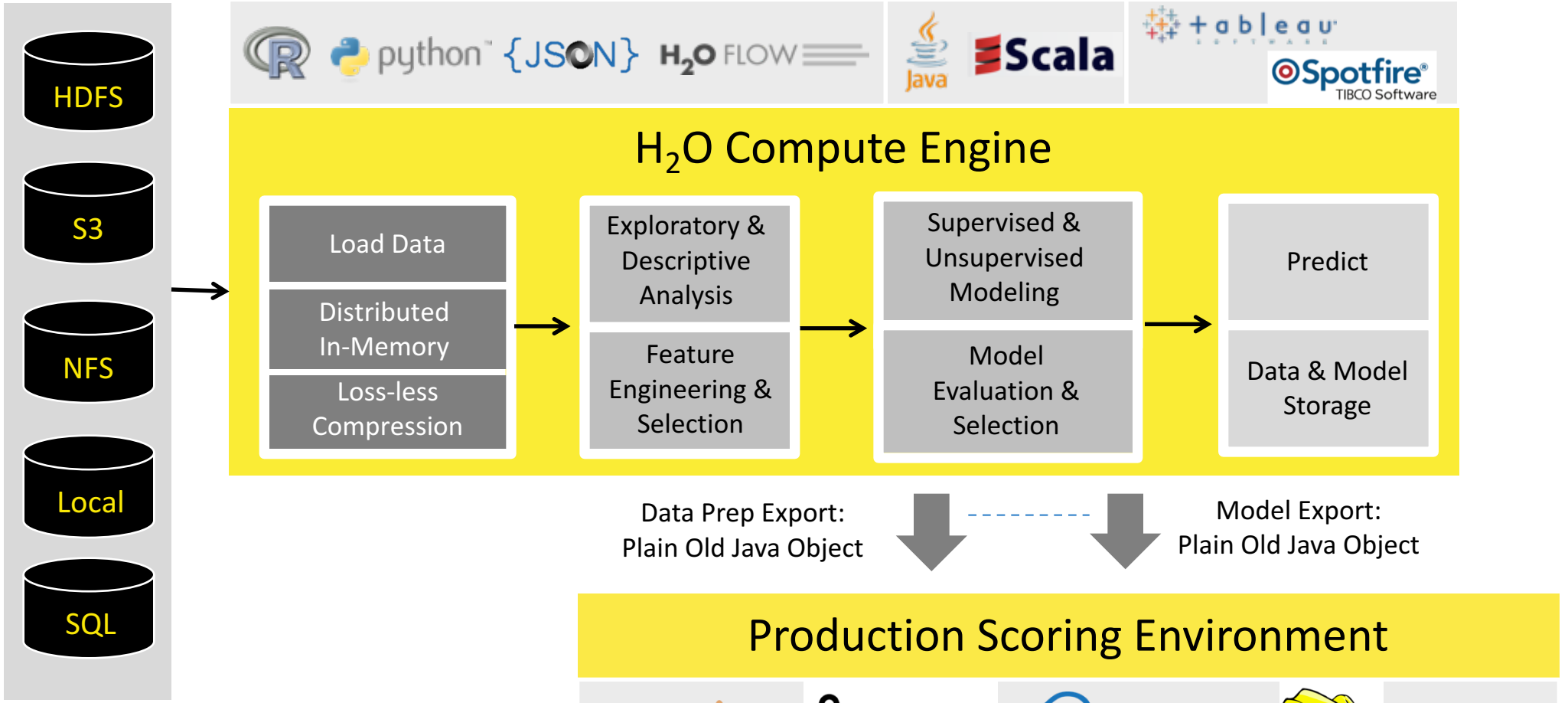
Meetup

15 Meetups a month

7th Sep - London
12th Sep - Dublin

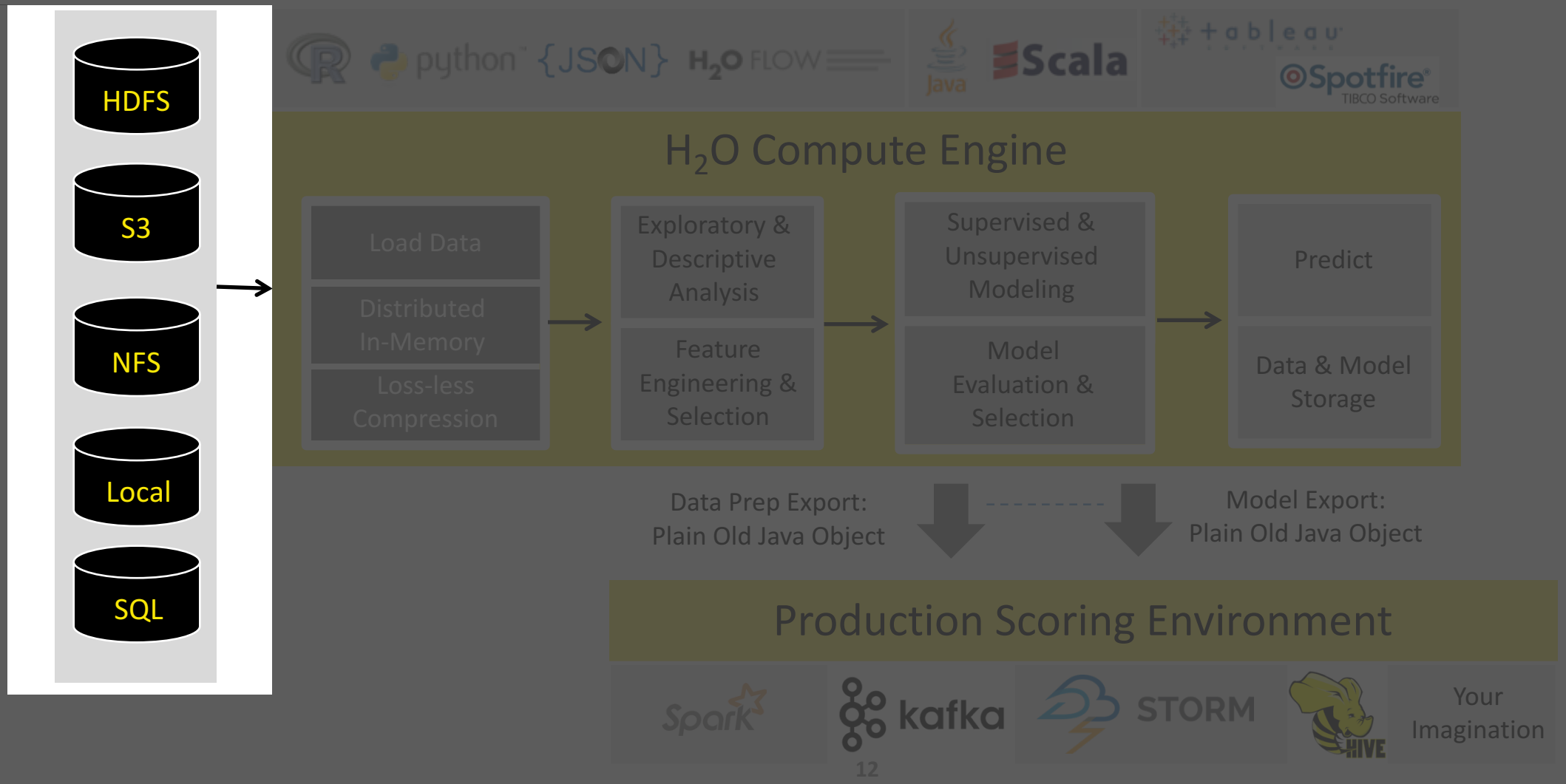
H_2O Machine Learning Platform

High Level Architecture



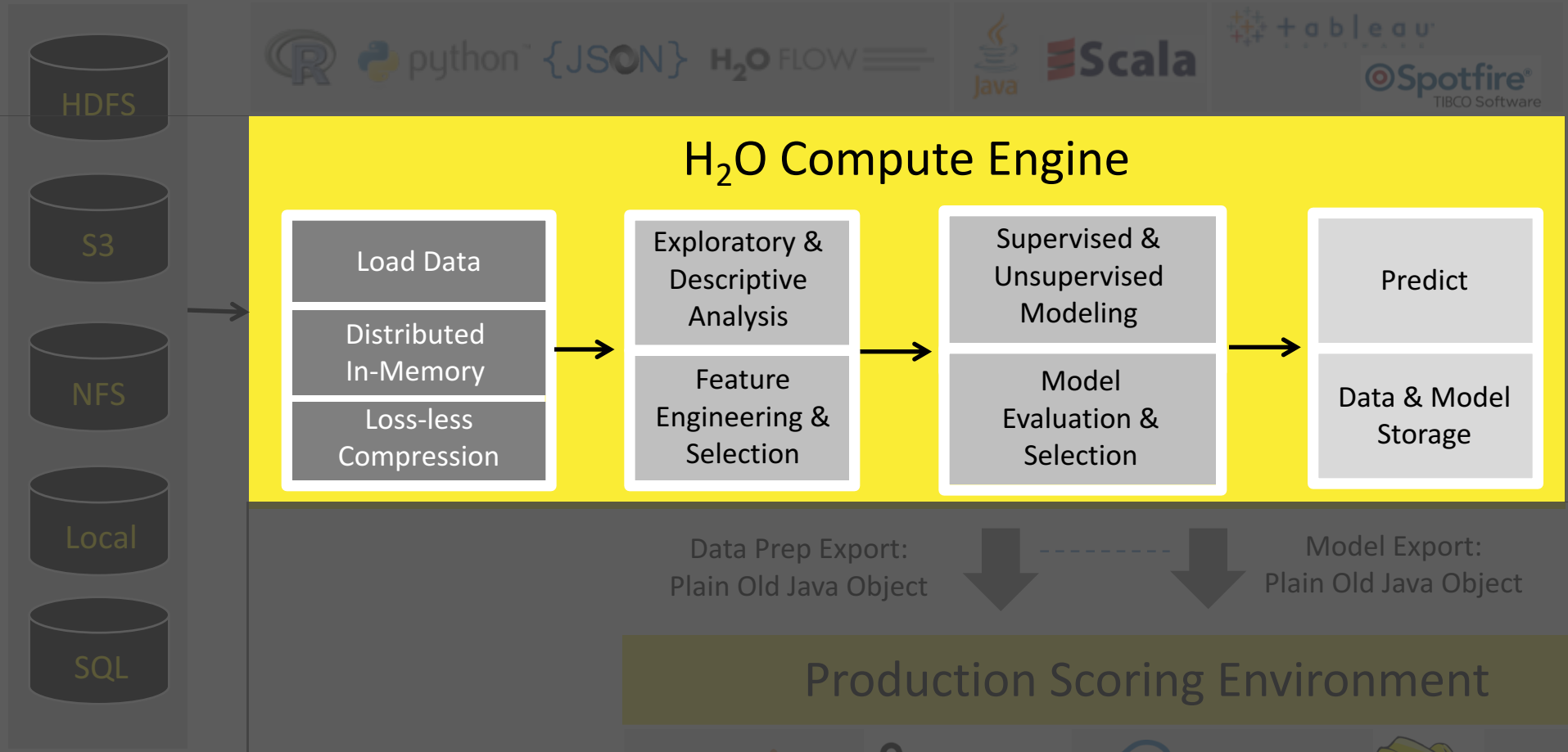
High Level Architecture

Import Data from Multiple Sources



High Level Architecture

Fast, Scalable & Distributed Compute Engine Written in Java



Algorithms Overview

Supervised Learning

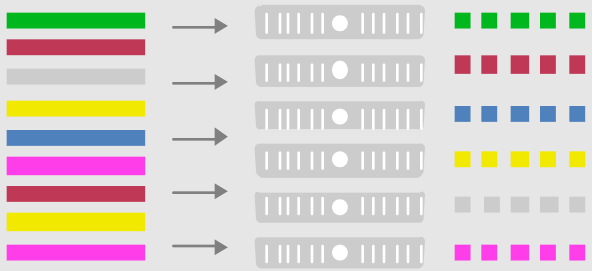
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
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

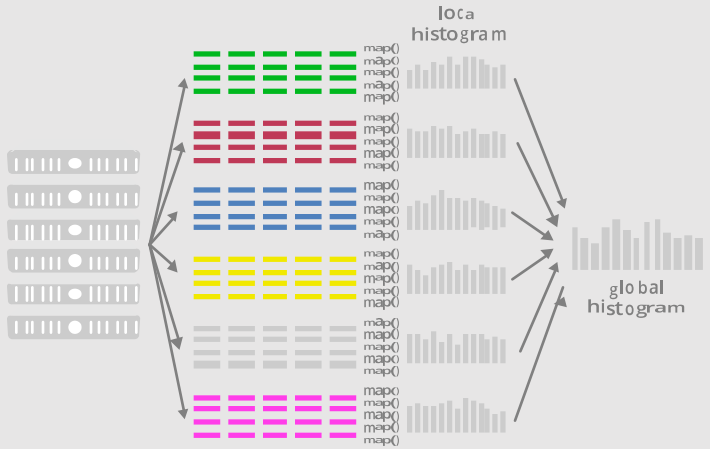
Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components • **Generalized Low Rank Models:** extend the 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

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



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

Advantageous Foundation

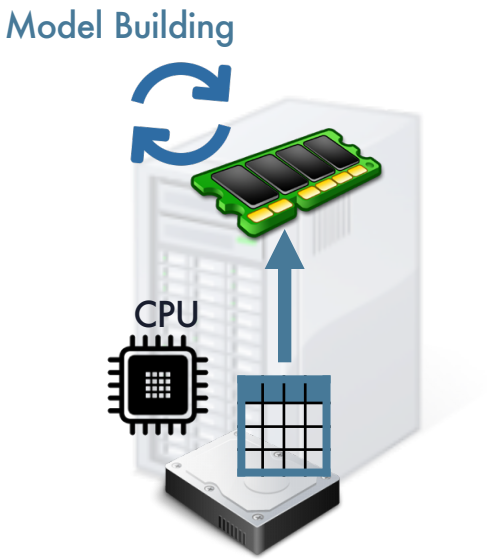
- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

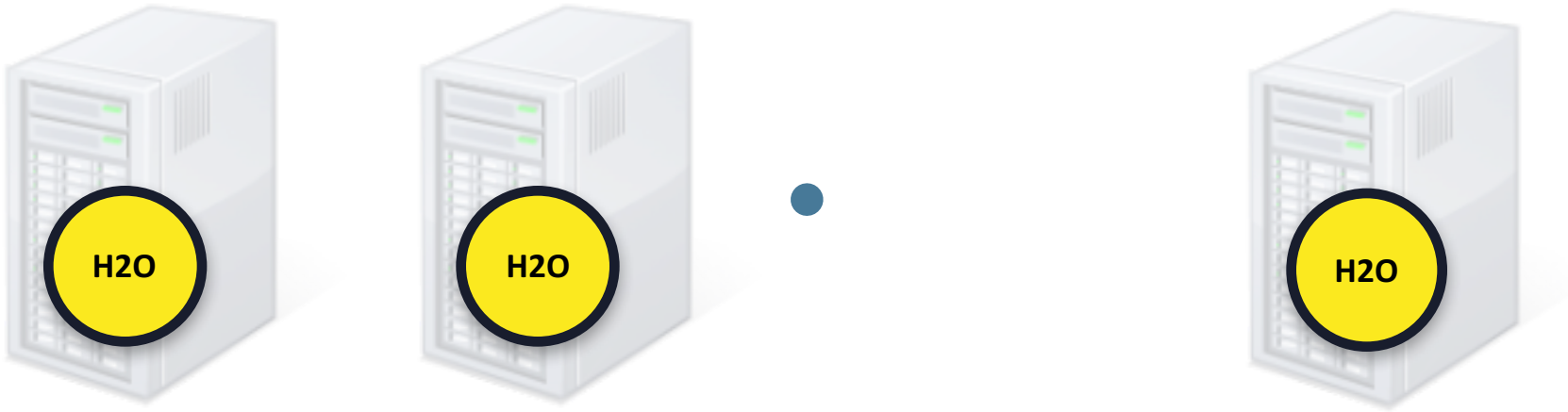
- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

H_2O Core

H_2O Core



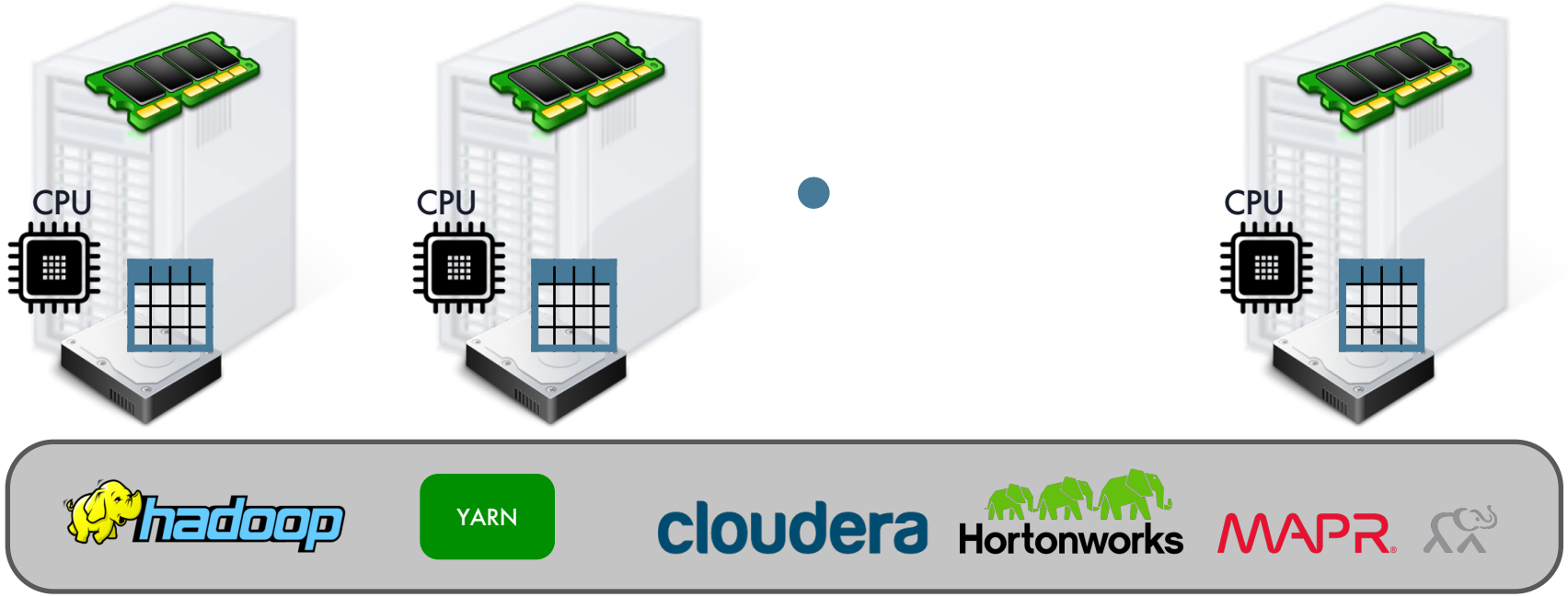
H_2O Core



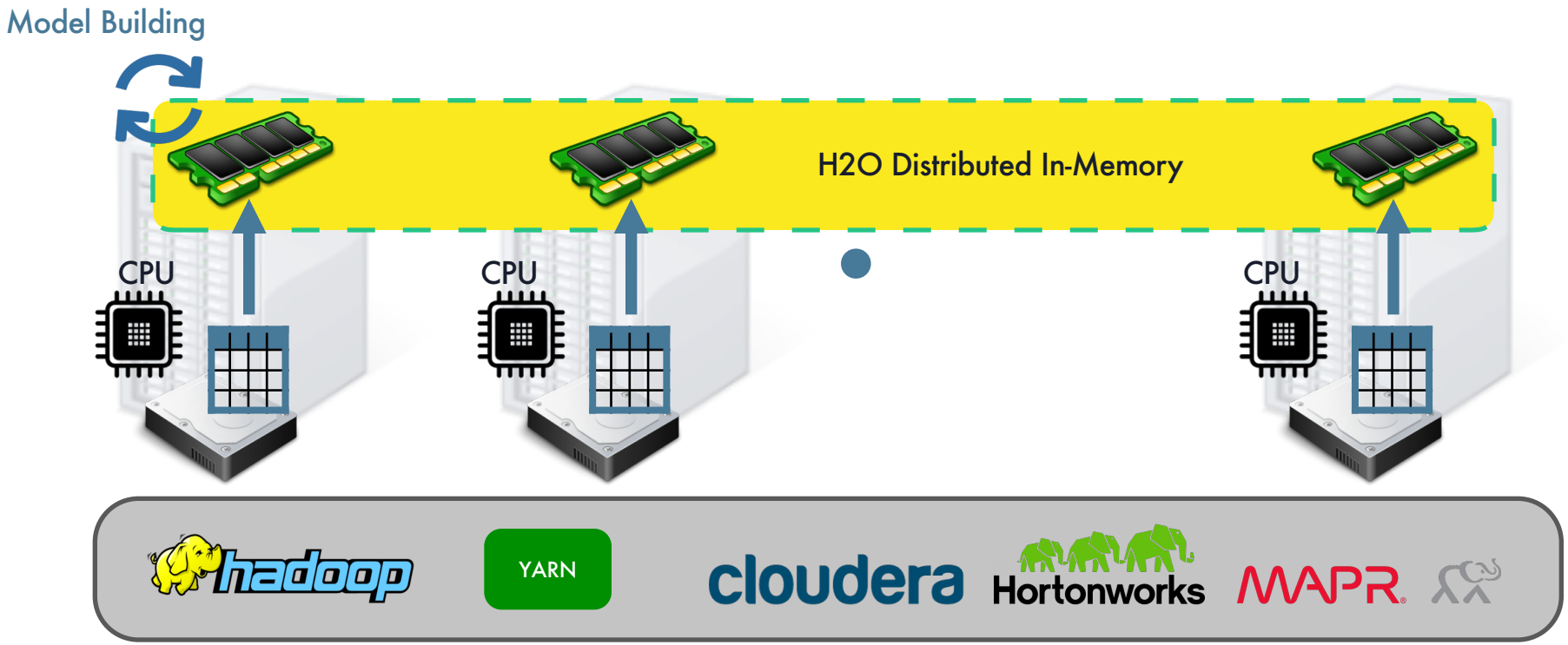
H_2O Core

H_2O Core

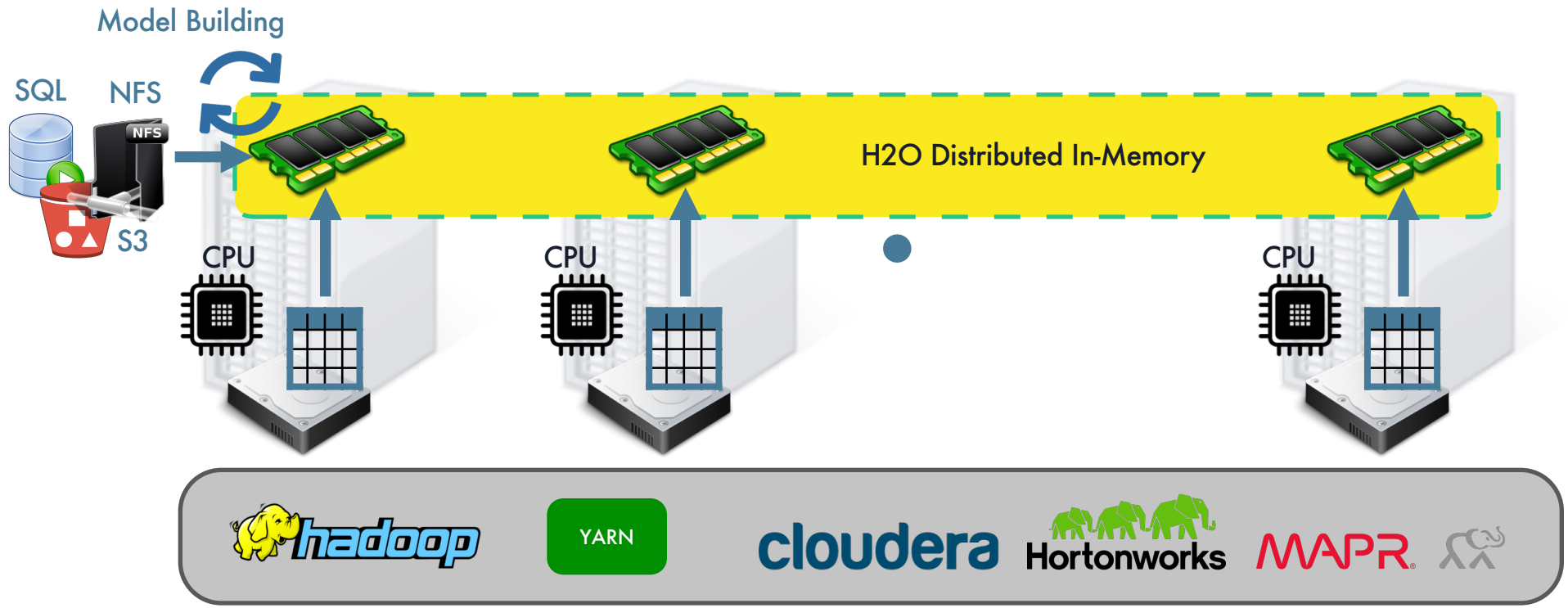
H_2O Core



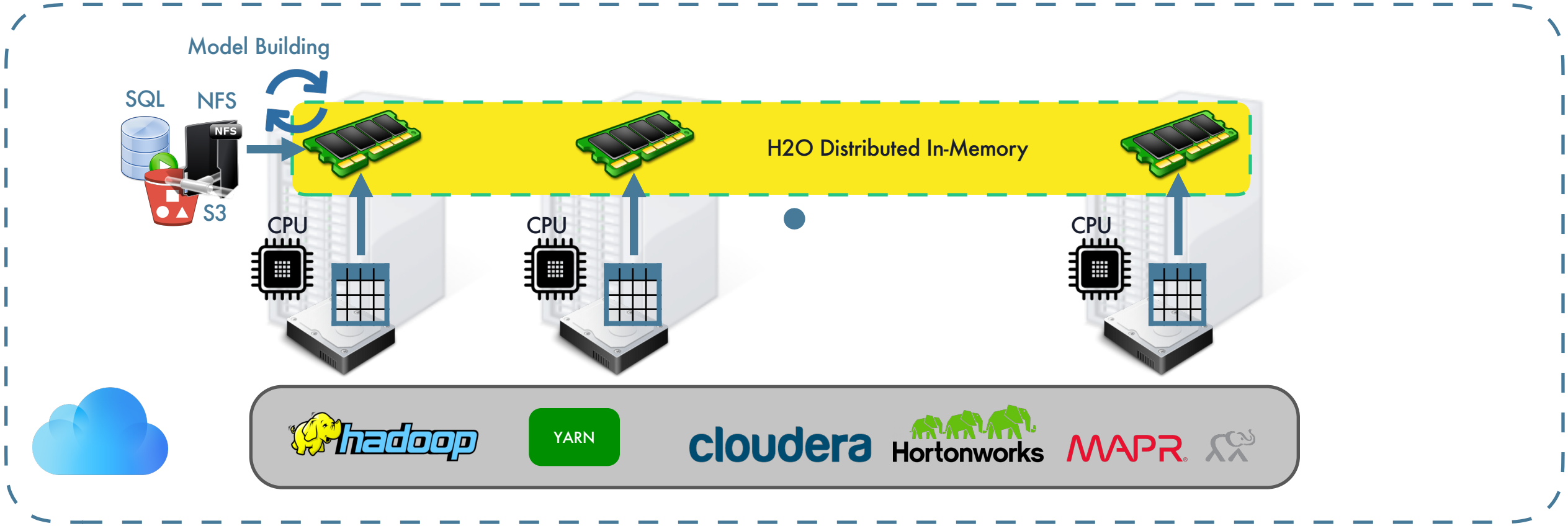
H_2O Core


Model Building
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

H_2O Core

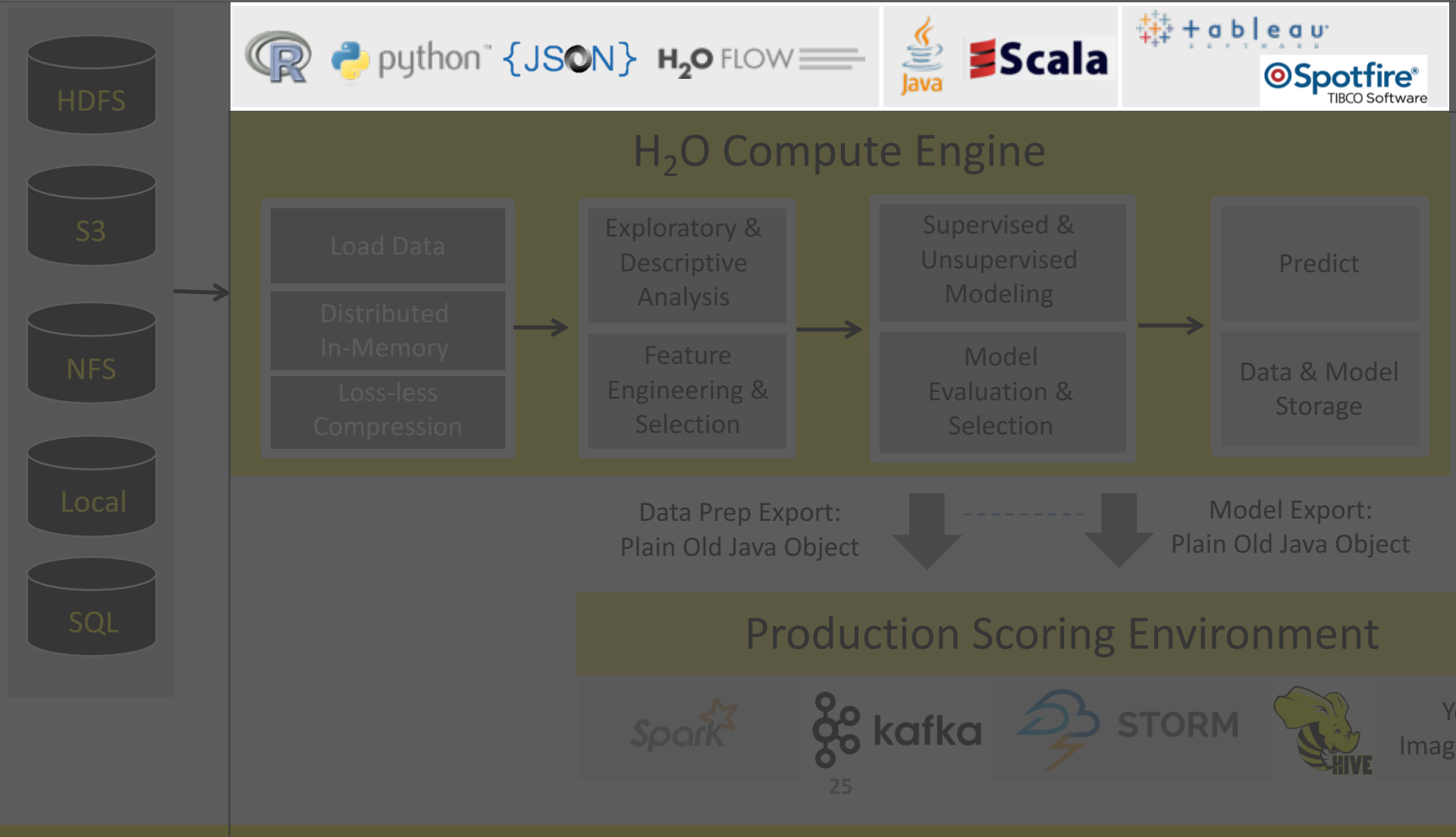

Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

H_2O Core

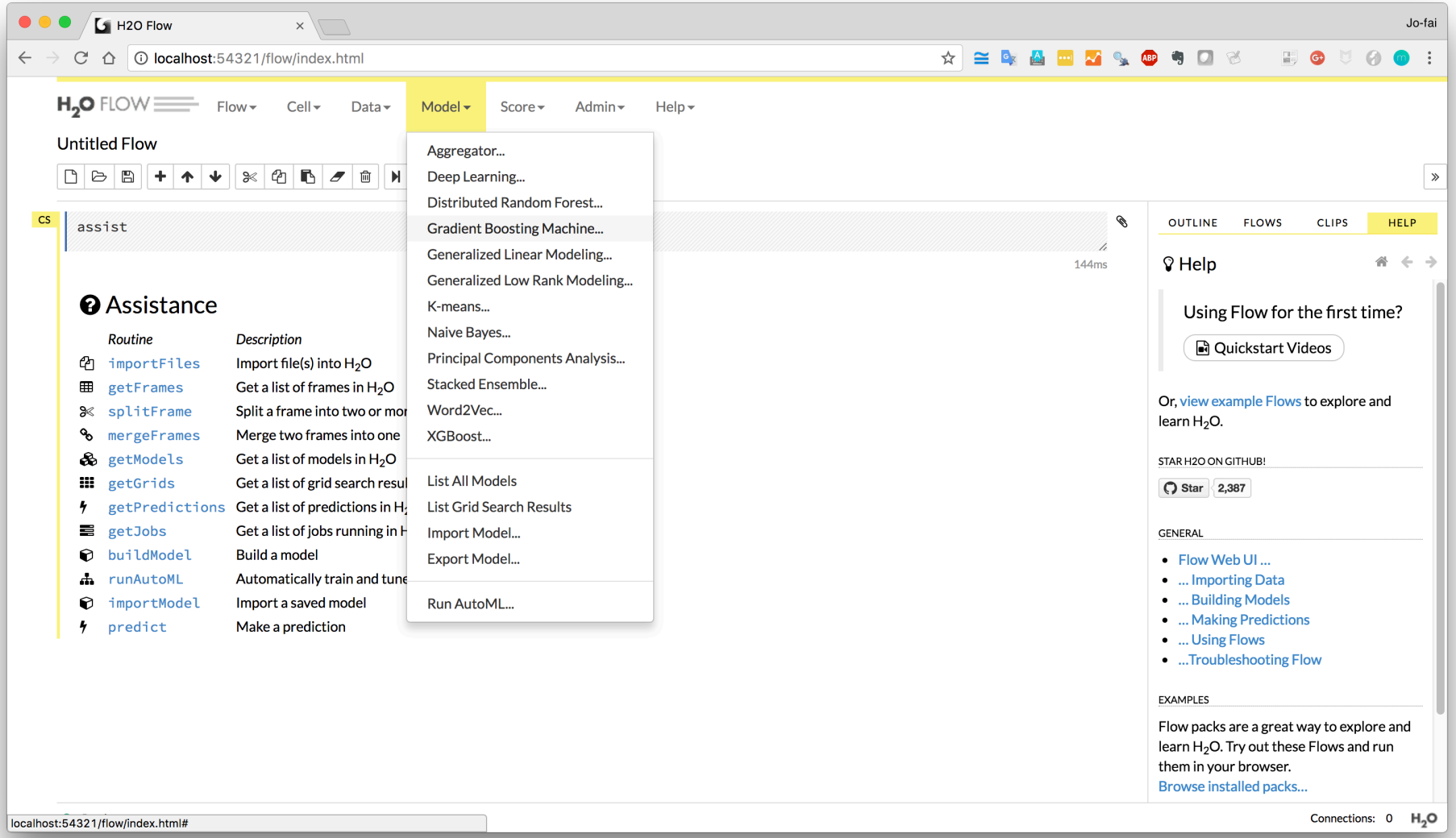

Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

High Level Architecture

Multiple Interfaces



H_2O Flow (Web)



H2O.ai

H_2O + R

```
# ------------------------------------------------------------
# Train a H2O Model
# ------------------------------------------------------------

# Train three basic H2O models
model_drf <- h2o.randomForest(x = features,
                              y = target,
                              model_id = "iris_random_forest",
                              training_frame = d_iris)

model_gbm <- h2o.gbm(x = features,
                     y = target,
                     model_id = "iris_gbm",
                     training_frame = d_iris)

model_dnn <- h2o.deeplearning(x = features,
                              y = target,
                              model_id = "iris_deep_learning",
                              training_frame = d_iris)
```

H_2O + Python

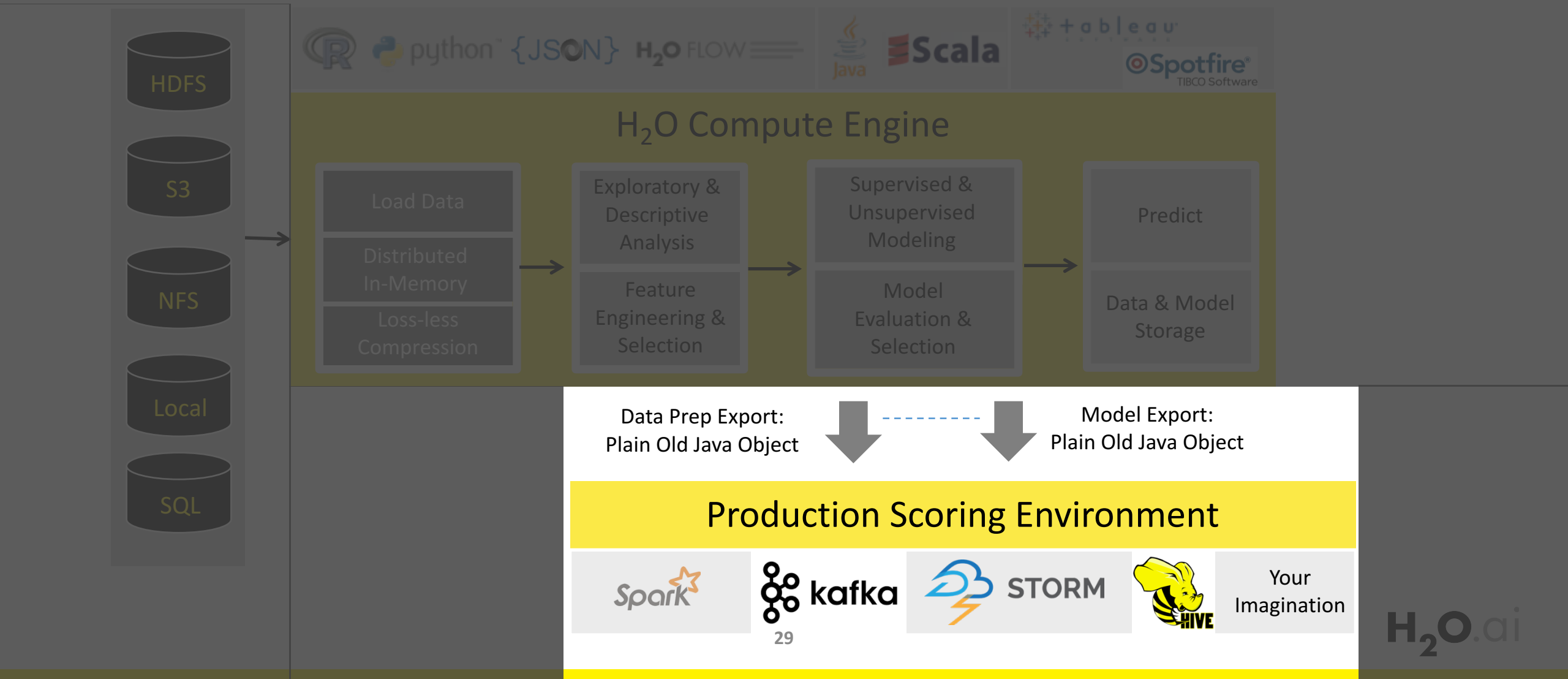
Gradient Boosting Machines

```
# Build a Gradient Boosting Machines (GBM) model with default settings

# Import the function for GBM
from h2o.estimators.gbm import H2OGradientBoostingEstimator

# Set up GBM for regression
# Add a seed for reproducibility
gbm_default = H2OGradientBoostingEstimator(model_id = 'gbm_default', seed = 1234)

# Use .train() to build the model
gbm_default.train(x = features,
                  y = 'quality',
                  training_frame = wine_train)
```

```
gbm Model Build progress: |████████████████████████████████████████████████| 100%
```

High Level Architecture

Export Standalone Models for Production



docs.h2o.ai

Getting Started & User Guides

H_2O

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 H_2O

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

Steam

What is Steam?
Steam User Guide
Recent Changes
Open Source License (AGPL)

Download Steam

Deep Water (preview)

Deep Water Readme
Deep Water Booklet
Deep Water AMI Guide
Deep Water Docker Image
Open Source License (Apache V2)

Launch Deep Water AMI
(choose p2.xlarge)

Q & A

FAQ
Issue Tracking (JIRA)
Stack Overflow
h2ostream Google Group
Gitter
Cross Validated

For Supported Enterprise Customers

Enterprise Support Web | Email

Algorithms

Supervised Learning

Generalized Linear Modeling (GLM)	Tutorial	Booklet	Reference	Tuning
Gradient Boosting Machine (GBM)	Tutorial	Booklet	Reference	Tuning
Deep Learning	Tutorial	Booklet	Reference	Tuning
Distributed Random Forest	Tutorial	Booklet	Reference	Tuning
Naive Bayes	Tutorial	Booklet	Reference	Tuning
Stacked Ensembles	Tutorial	Booklet	Reference	Tuning
XGBoost	Tutorial	Booklet	Reference	Tuning

Unsupervised Learning

Generalized Low Rank Models (GLRM)	Tutorial	Reference
K-Means Clustering	Tutorial	Reference
Principal Components Analysis (PCA)	Tutorial	Reference

Miscellaneous

Word2vec	Tutorial	Reference

TOP

H_2O.ai

Use Case 1

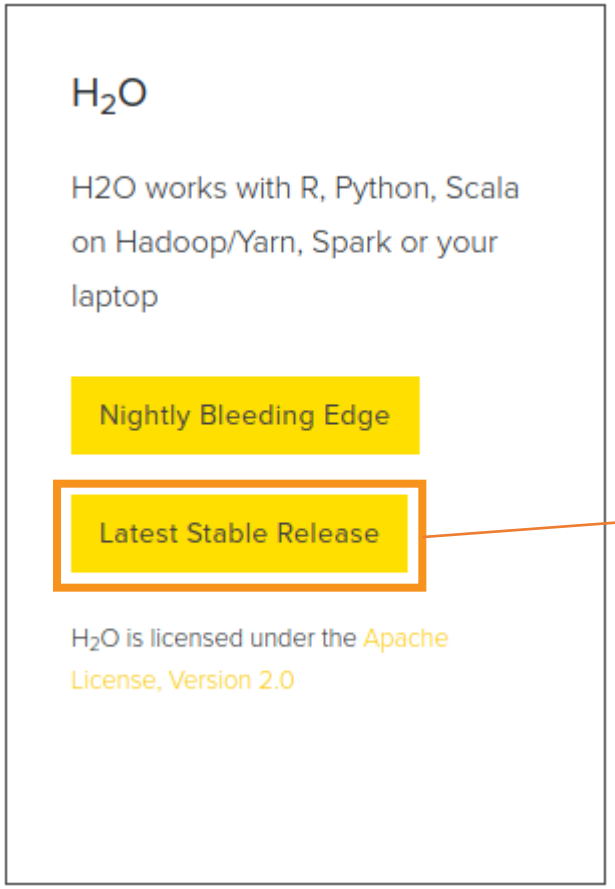
- Install and start H_2O (Web/R/Python)
- Machine Learning Basics
- Import Credit Card dataset and train a model to predict default payments

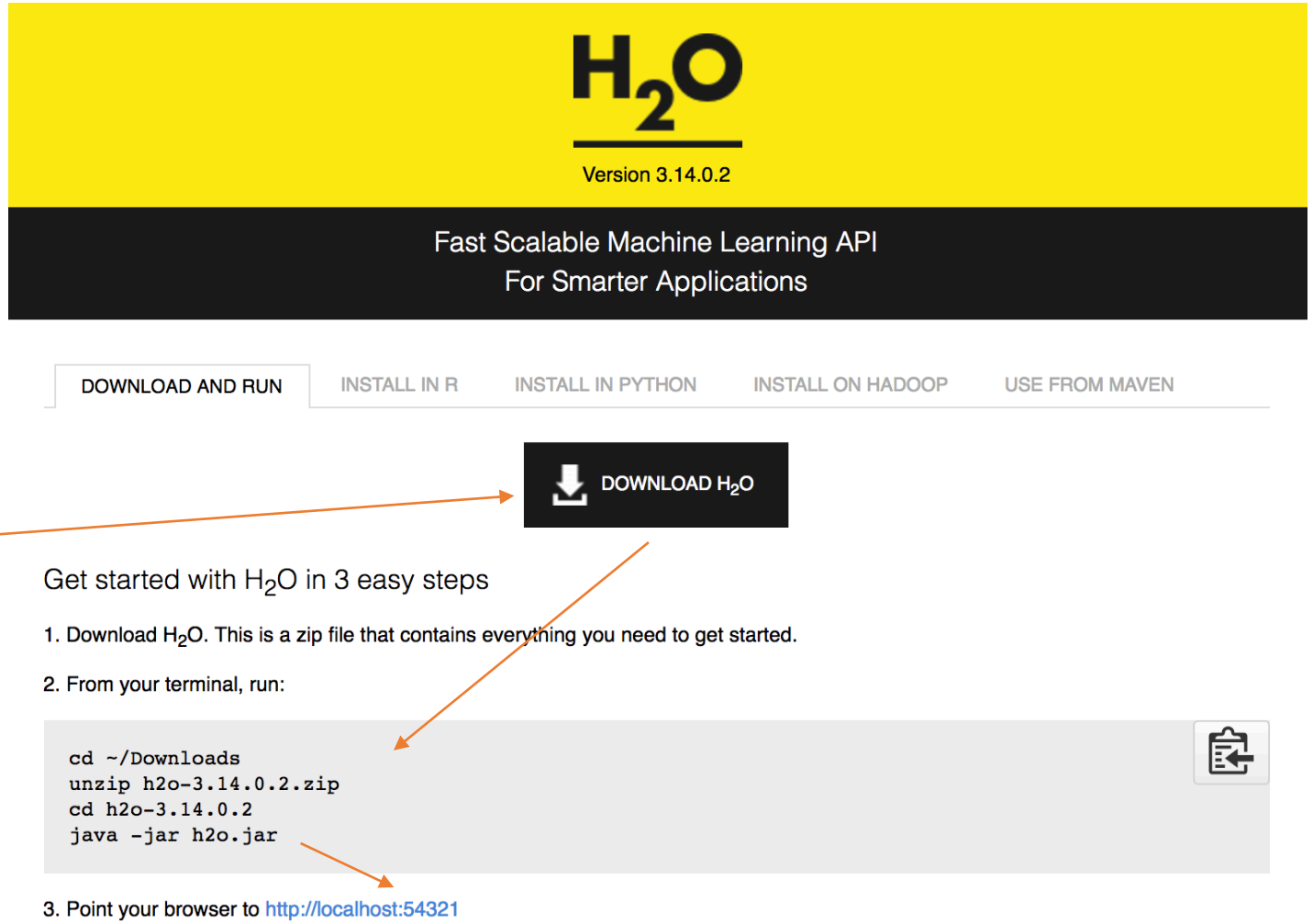
H_2O Quick Start

Note: please contact IT support for files/folders access permission

www.h2o.ai/download

Prerequisite: Java version 1.7+

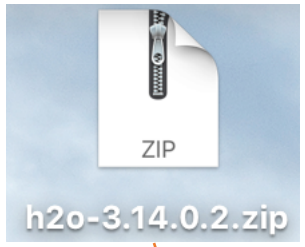


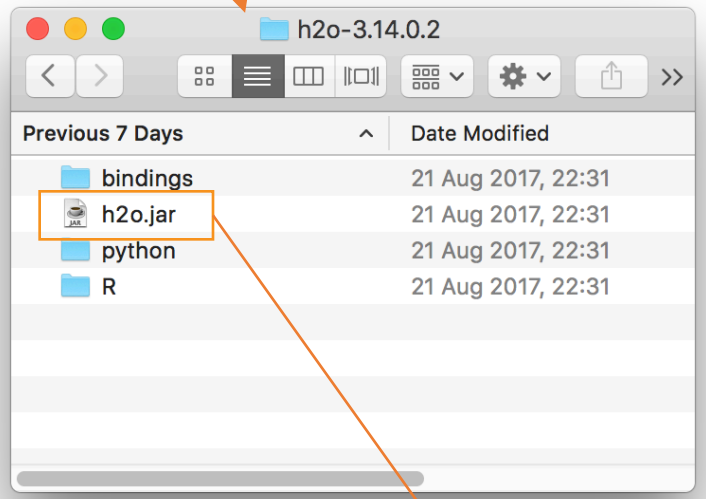


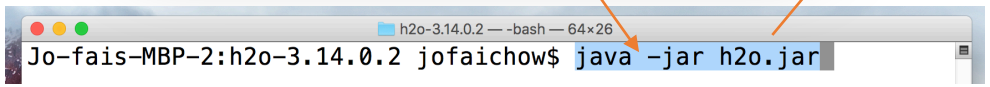
H2O.ai

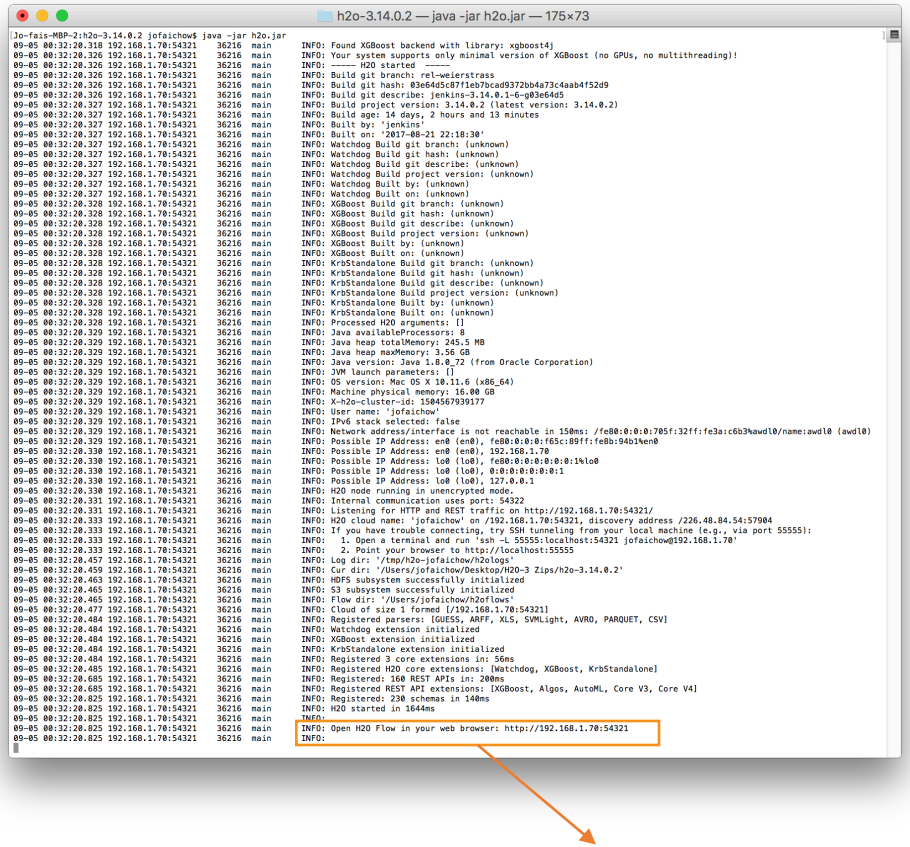
Note: please contact IT support for files/folders access permission

Install and Start H_2O Flow (Web Interface)







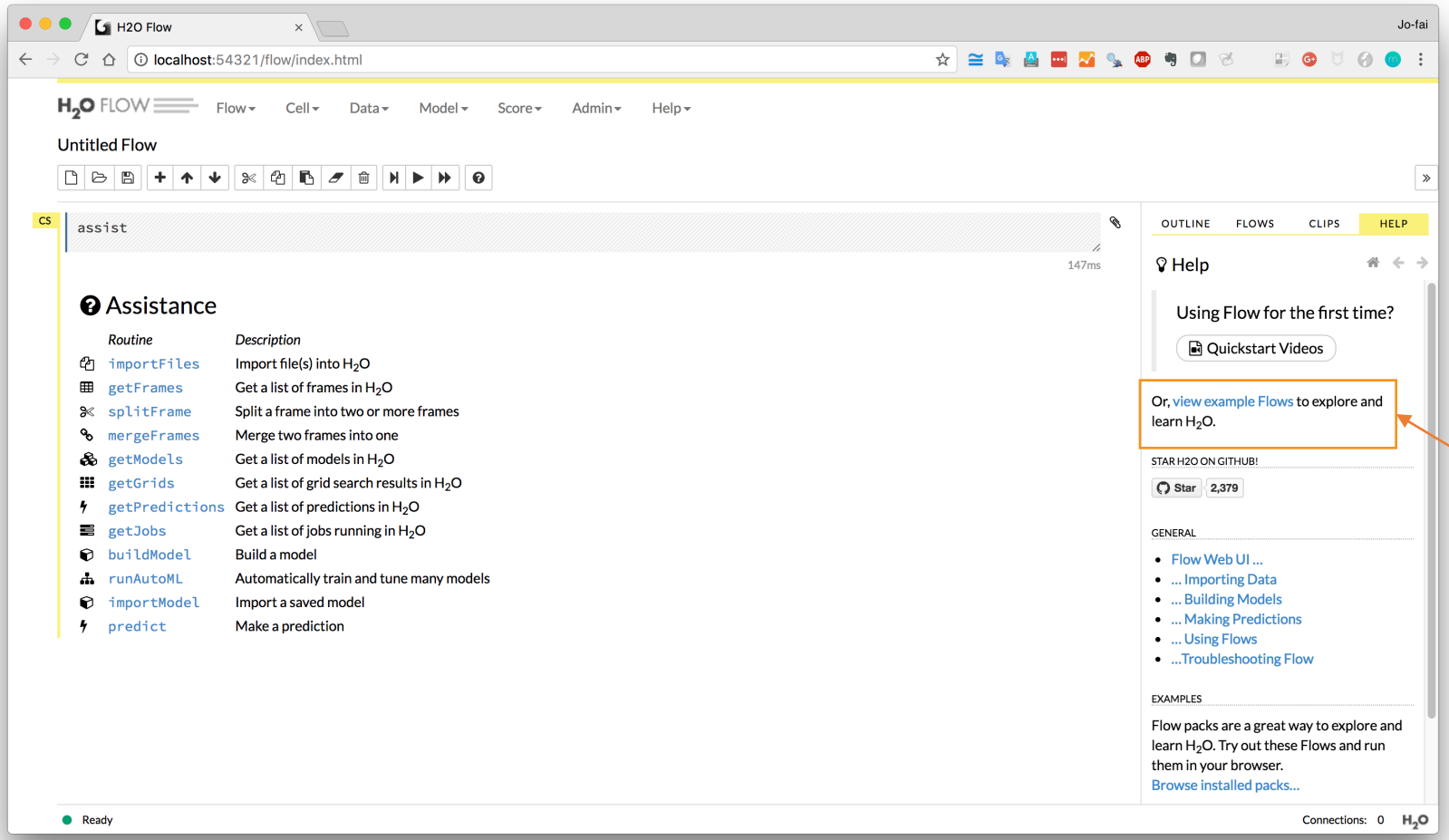


INFO: Open H2O Flow in your web browser: http://192.168.1.70:54321

H_2O.ai

Note: please contact IT support for files/folders access permission

H_2O Flow (Web Interface)

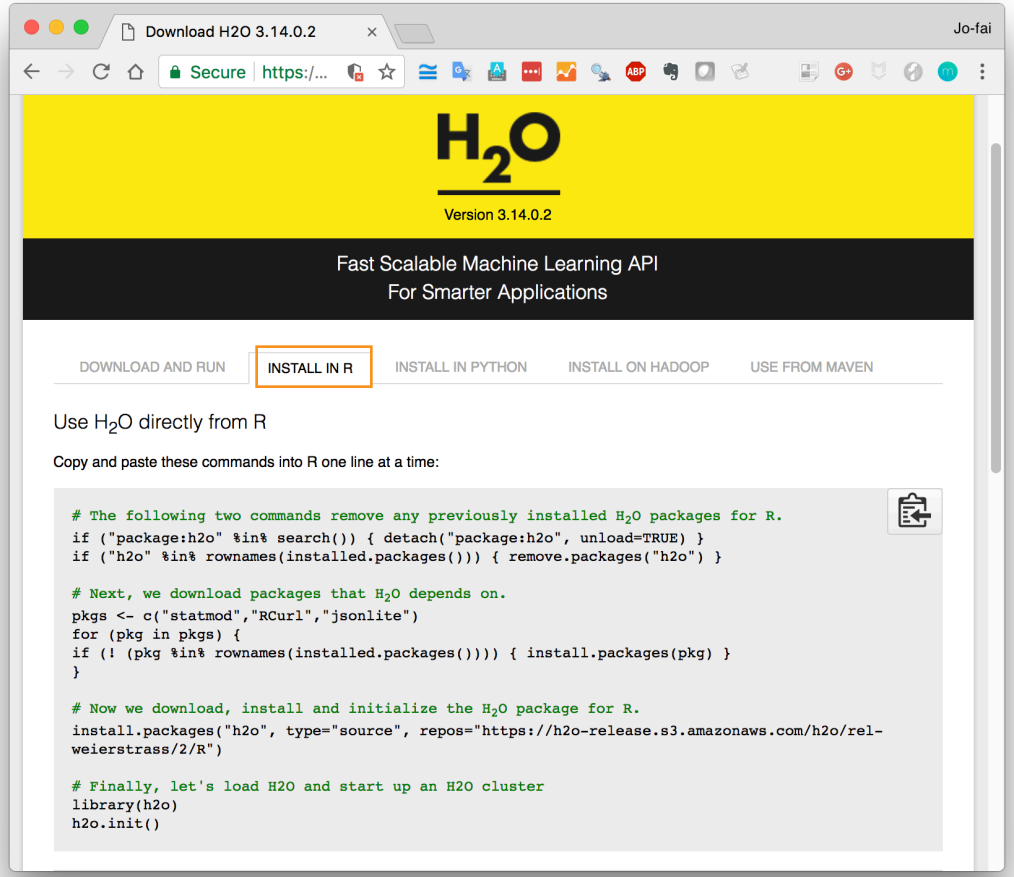


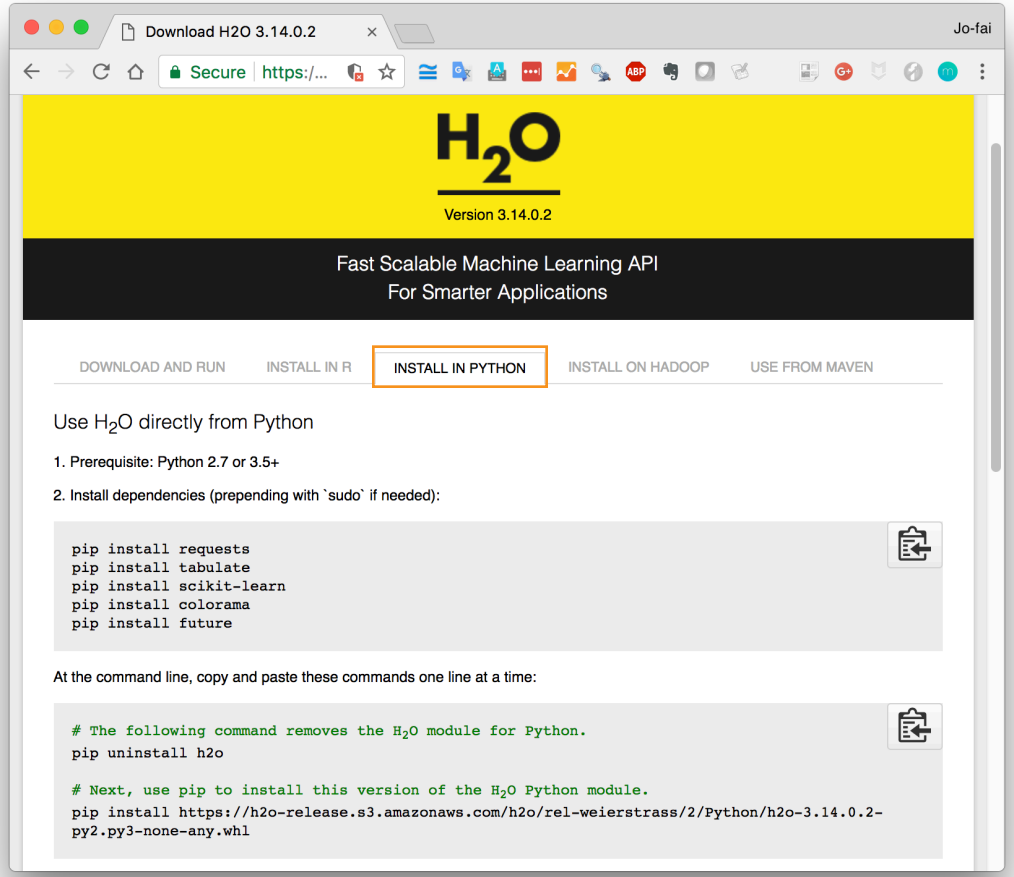
More Examples

H2O.ai

Note: please contact IT support for files/folders access permission

Install and Start H_2O in R / Python

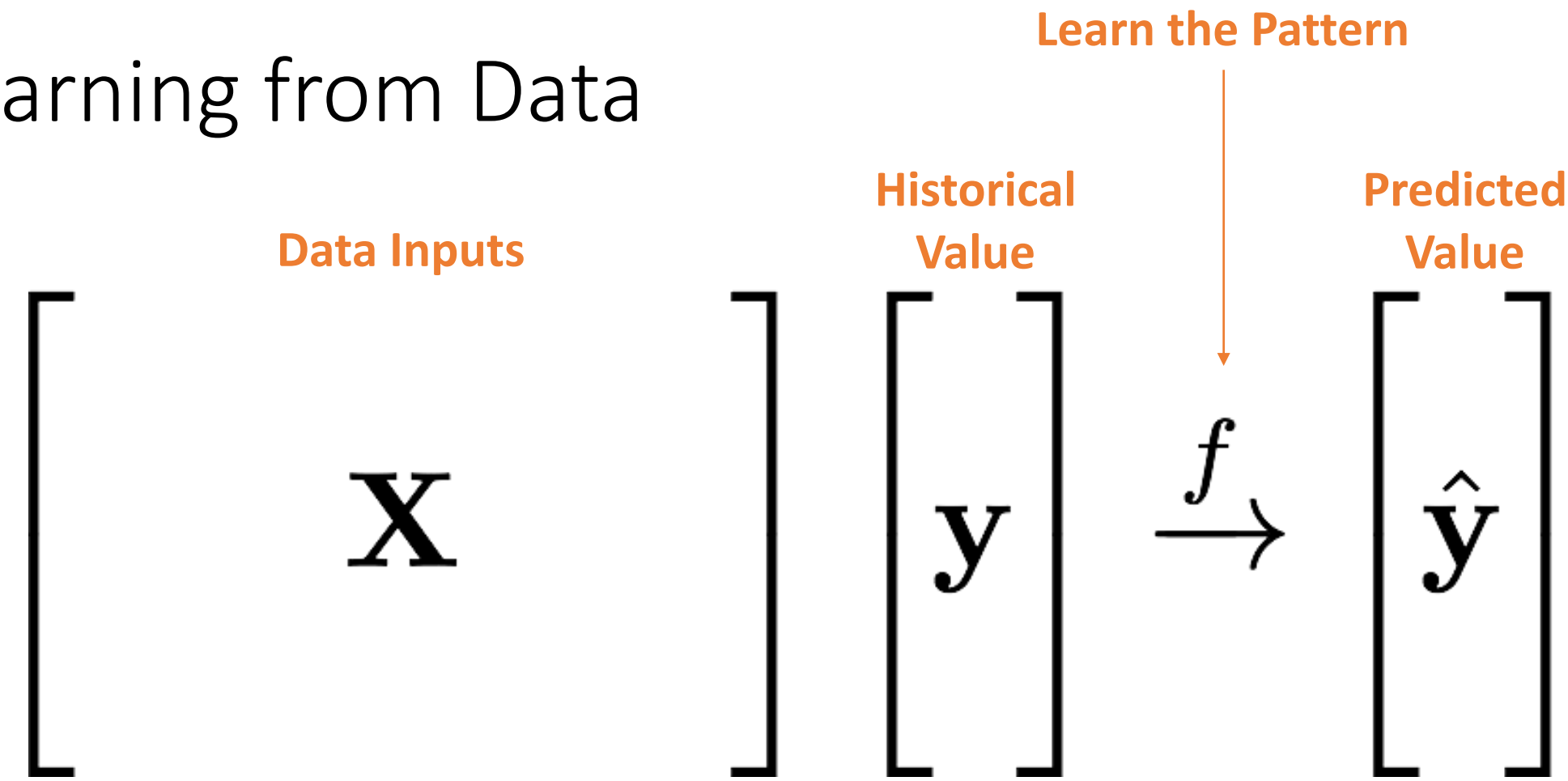




H2O.ai

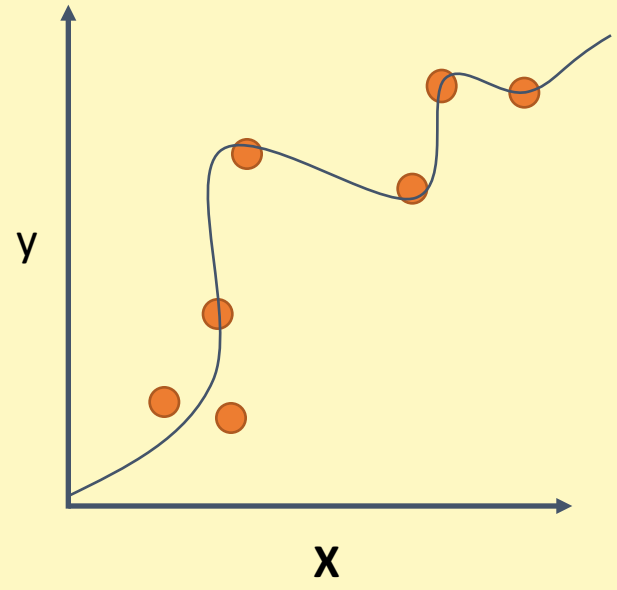
Machine Learning Basics

Learning from Data



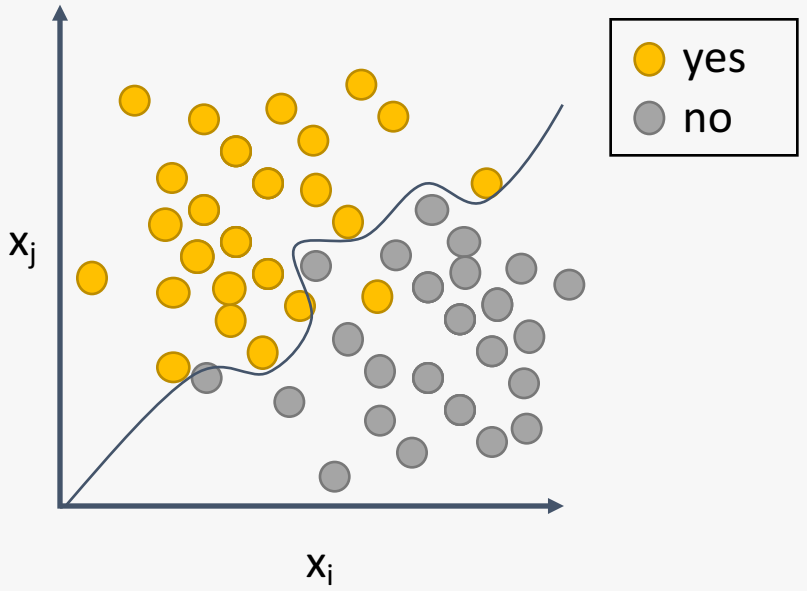
Supervised Learning

Regression:
How much will a customers spend?



H_2O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:
Will a customer make a purchase? Yes or No

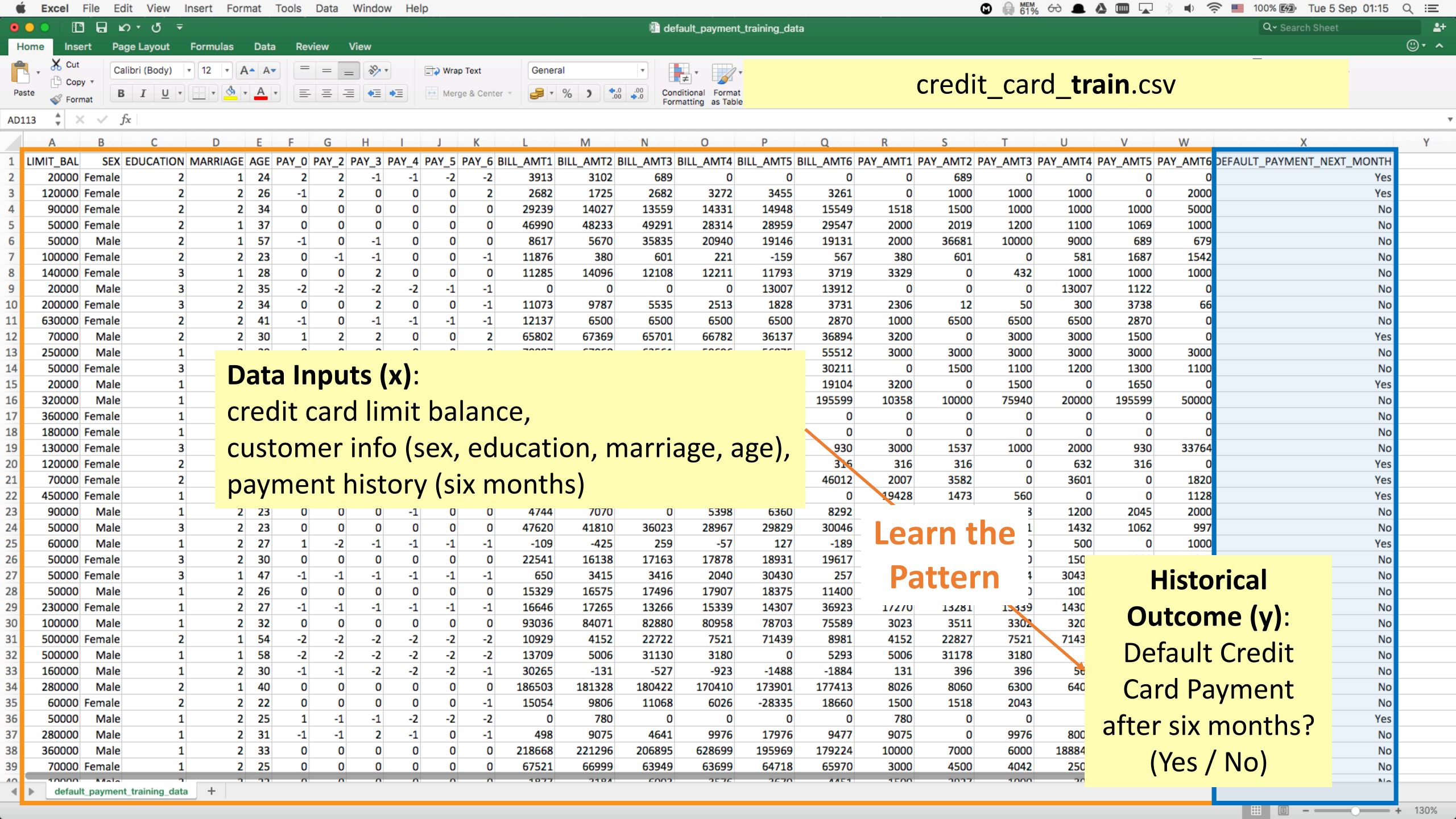


H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

H_2O.ai

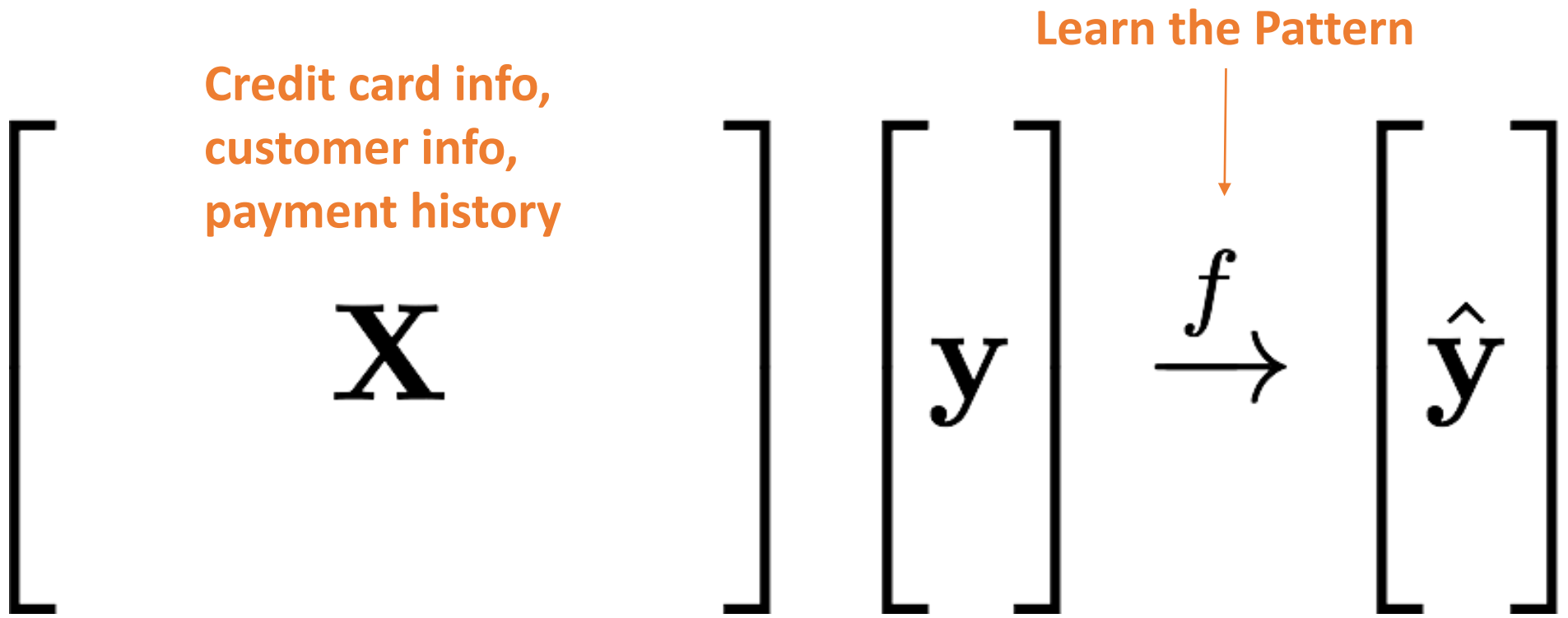
Credit Card Dataset

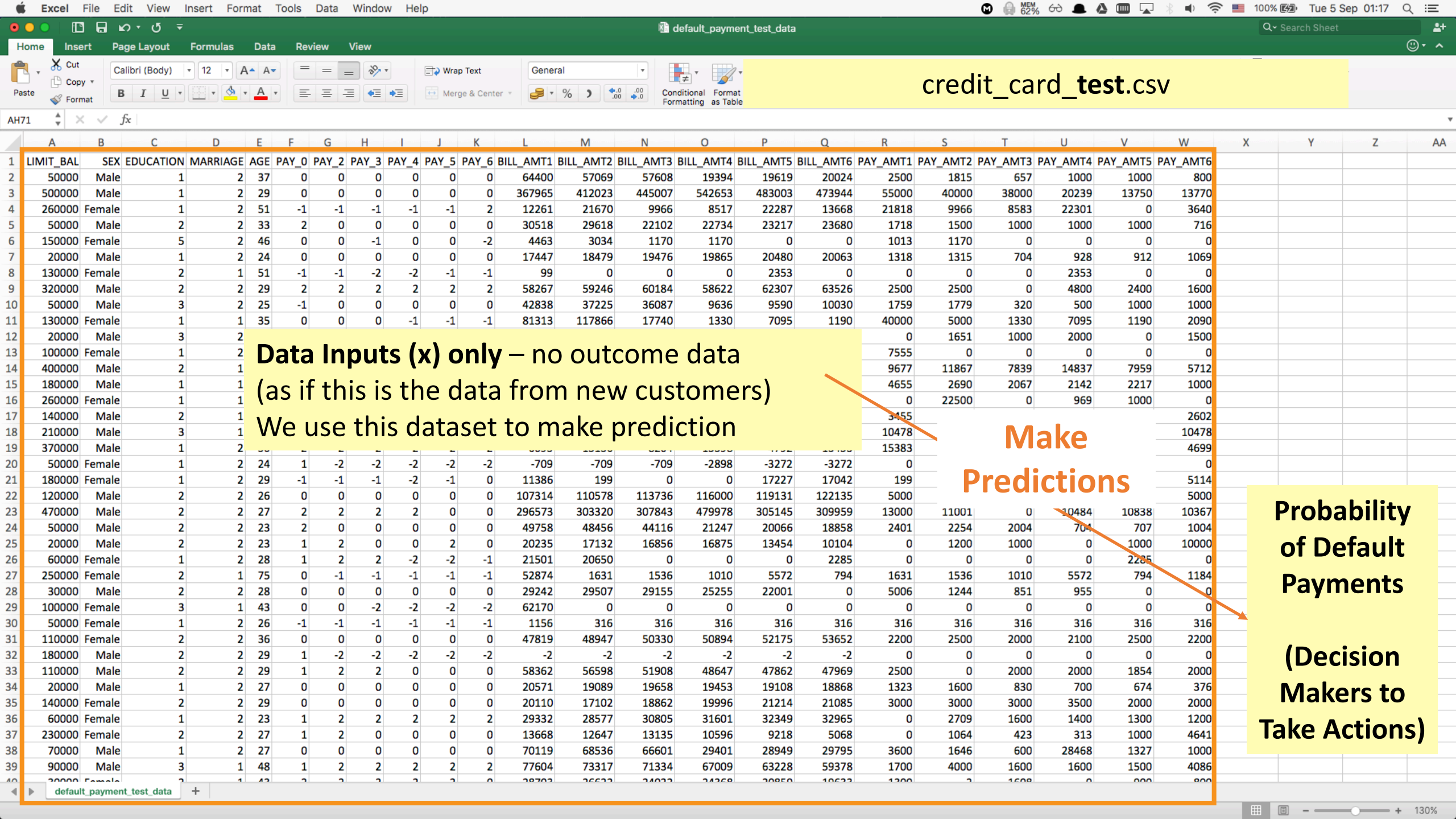
- Links:
  - https://github.com/woobe/h2o_tutorials/raw/master/datasets/credit_card_train.csv
  - https://github.com/woobe/h2o_tutorials/raw/master/datasets/credit_card_test.csv



LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	DEFAULT_PAYMENT_NEXT_MONTH
20000	Female	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	Yes
120000	Female	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	Yes
90000	Female	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	No
50000	Female	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	No
50000	Male	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	No
100000	Female	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	No
140000	Female	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	No
20000	Male	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	No
200000	Female	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	No
630000	Female	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	No
70000	Male	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	Yes
250000	Male	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	No
50000	Female	3	1	23	0	0	0	0	0	0	49764	49867	48137	48005	47989	30211	0	1500	1100	1200	1300	1100	No
20000	Male	1	2	28	0	0	0	2	0	0	18010	17428	18338	17905	18467	19104	3200	0	1500	0	1650	0	Yes
320000	Male	1	1	35	-1	-1	-1	-1	-1	-1	317	2362	632	1200	1800	195599	10358	10000	75940	20000	195599	50000	No
360000	Female	1	2	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	No
180000	Female	1	2	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	No
130000	Female	3	2	30	0	0	0	0	0	0	2331	1663	1256	1290	1315	930	3000	1537	1000	2000	930	33764	No
120000	Female	2	2	25	0	0	0	0	0	0	316	316	316	316	316	316	316	316	0	632	316	0	Yes
70000	Female	2	2	24	0	0	0	0	0	0	46012	46012	46012	46012	46012	46012	2007	3582	0	3601	0	1820	Yes
450000	Female	1	2	28	0	0	0	0	0	0	0	0	0	0	0	0	19428	1473	560	0	0	1128	Yes
90000	Male	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	3			1200	2045	2000	No
50000	Male	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1			1432	1062	997	No
60000	Male	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0			500	0	1000	Yes
50000	Female	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	0			150			No
50000	Female	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	4			3043			No
50000	Male	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	0			100			No
230000	Female	1	2	27	-1	-1	-1	-1	-1	-1	16646	17265	13266	15339	14307	36923	17270	13281	15339	1430			No
100000	Male	1	2	32	0	0	0	0	0	0	93036	84071	82880	80958	78703	75589	3023	3511	3302	320			No
500000	Female	2	1	54	-2	-2	-2	-2	-2	-2	10929	4152	22722	7521	71439	8981	4152	22827	7521	7143			No
500000	Male	1	1	58	-2	-2	-2	-2	-2	-2	13709	5006	31130	3180	0	5293	5006	31178	3180				No
160000	Male	1	2	30	-1	-1	-2	-2	-2	-1	30265	-131	-527	-923	-1488	-1884	131	396	396	58			No
280000	Male	2	1	40	0	0	0	0	0	0	186503	181328	180422	170410	173901	177413	8026	8060	6300	640			No
60000	Female	2	2	22	0	0	0	0	0	-1	15054	9806	11068	6026	-28335	18660	1500	1518	2043				No
50000	Male	1	2	25	1	-1	-1	-2	-2	-2	0	780	0	0	0	0	780	0	0				Yes
280000	Male	1	2	31	-1	-1	2	-1	0	-1	498	9075	4641	9976	17976	9477	9075	0	9976	800			No
360000	Male	1	2	33	0	0	0	0	0	0	218668	221296	206895	628699	195969	179224	10000	7000	6000	18884			No
70000	Female	1	2	25	0	0	0	0	0	0	67521	66999	63949	63699	64718	65970	3000	4500	4042	250			No

Learning from Credit Card Data



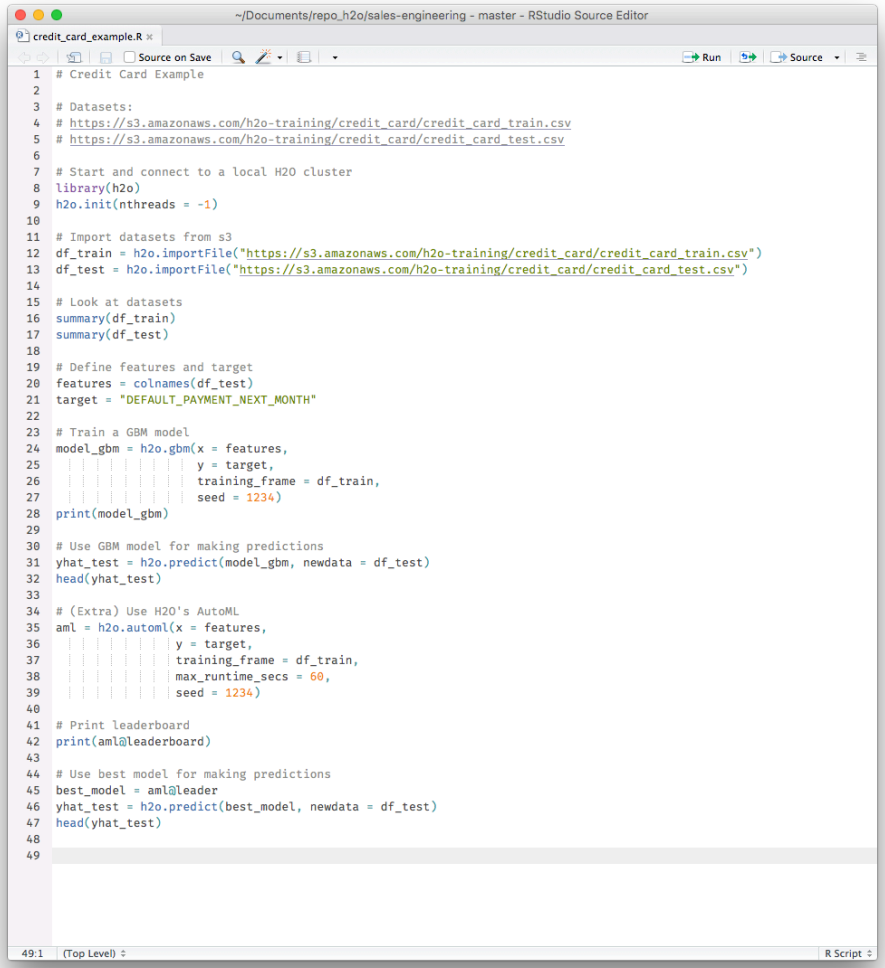


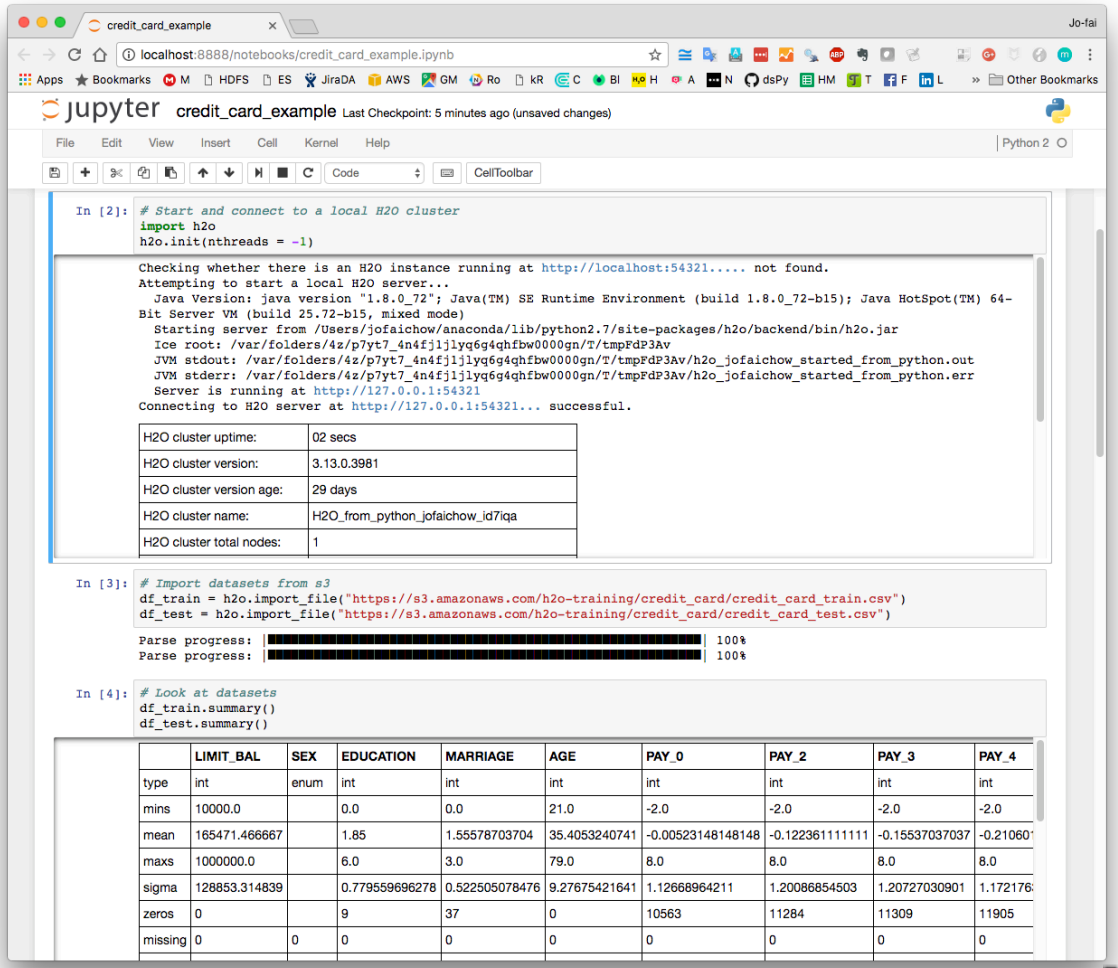
LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
50000	Male	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800
500000	Male	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770
260000	Female	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640
50000	Male	2	2	33	2	0	0	0	0	0	30518	29618	22102	22734	23217	23680	1718	1500	1000	1000	1000	716
150000	Female	5	2	46	0	0	-1	0	0	-2	4463	3034	1170	1170	0	0	1013	1170	0	0	0	0
20000	Male	1	2	24	0	0	0	0	0	0	17447	18479	19476	19865	20480	20063	1318	1315	704	928	912	1069
130000	Female	2	1	51	-1	-1	-2	-2	-1	-1	99	0	0	0	2353	0	0	0	0	2353	0	0
320000	Male	2	2	29	2	2	2	2	2	2	58267	59246	60184	58622	62307	63526	2500	2500	0	4800	2400	1600
50000	Male	3	2	25	-1	0	0	0	0	0	42838	37225	36087	9636	9590	10030	1759	1779	320	500	1000	1000
130000	Female	1	1	35	0	0	0	-1	-1	-1	81313	117866	17740	1330	7095	1190	40000	5000	1330	7095	1190	2090
20000	Male	3	2														0	1651	1000	2000	0	1500
100000	Female	1	2														7555	0	0	0	0	0
400000	Male	2	1														9677	11867	7839	14837	7959	5712
180000	Male	1	1														4655	2690	2067	2142	2217	1000
260000	Female	1	1														0	22500	0	969	1000	0
140000	Male	2	1														3455					2602
210000	Male	3	1														10478					10478
370000	Male	1	2														15383					4699
50000	Female	1	2	24	1	-2	-2	-2	-2	-2	-709	-709	-709	-2898	-3272	-3272	0					0
180000	Female	1	2	29	-1	-1	-1	-2	-1	0	11386	199	0	0	17227	17042	199					5114
120000	Male	2	2	26	0	0	0	0	0	0	107314	110578	113736	116000	119131	122135	5000					5000
470000	Male	2	2	27	2	2	2	2	0	0	296573	303320	307843	479978	305145	309959	13000	11001	0	10484	10838	10367
50000	Male	2	2	23	2	0	0	0	0	0	49758	48456	44116	21247	20066	18858	2401	2254	2004	704	707	1004
20000	Male	2	2	23	1	2	0	0	2	0	20235	17132	16856	16875	13454	10104	0	1200	1000	0	1000	10000
60000	Female	1	2	28	1	2	2	-2	-2	-1	21501	20650	0	0	0	2285	0	0	0	0	2285	0
250000	Female	2	1	75	0	-1	-1	-1	-1	-1	52874	1631	1536	1010	5572	794	1631	1536	1010	5572	794	1184
30000	Male	2	2	28	0	0	0	0	0	0	29242	29507	29155	25255	22001	0	5006	1244	851	955	0	0
100000	Female	3	1	43	0	0	-2	-2	-2	-2	62170	0	0	0	0	0	0	0	0	0	0	0
50000	Female	1	2	26	-1	-1	-1	-1	-1	-1	1156	316	316	316	316	316	316	316	316	316	316	316
110000	Female	2	2	36	0	0	0	0	0	0	47819	48947	50330	50894	52175	53652	2200	2500	2000	2100	2500	2200
180000	Male	2	2	29	1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0
110000	Male	2	2	29	1	2	2	0	0	0	58362	56598	51908	48647	47862	47969	2500	0	2000	2000	1854	2000
20000	Male	1	2	27	0	0	0	0	0	0	20571	19089	19658	19453	19108	18868	1323	1600	830	700	674	376
140000	Female	2	2	29	0	0	0	0	0	0	20110	17102	18862	19996	21214	21085	3000	3000	3000	3500	2000	2000
60000	Female	1	2	23	1	2	2	2	2	2	29332	28577	30805	31601	32349	32965	0	2709	1600	1400	1300	1200
230000	Female	2	2	27	1	2	0	0	0	0	13668	12647	13135	10596	9218	5068	0	1064	423	313	1000	4641
70000	Male	1	2	27	0	0	0	0	0	0	70119	68536	66601	29401	28949	29795	3600	1646	600	28468	1327	1000
90000	Male	3	1	48	1	2	2	2	2	2	77604	73317	71334	67009	63228	59378	1700	4000	1600	1600	1500	4086

Live Demo 1

- Import and explore credit card dataset
- Train a gradient boosting model and make predictions
- Train many models using AutoML and make predictions using the best model

Using H_2O with R and Python



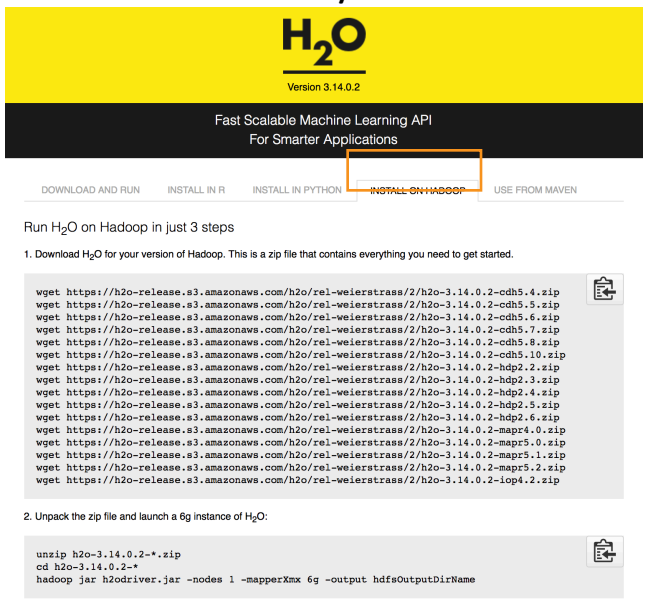


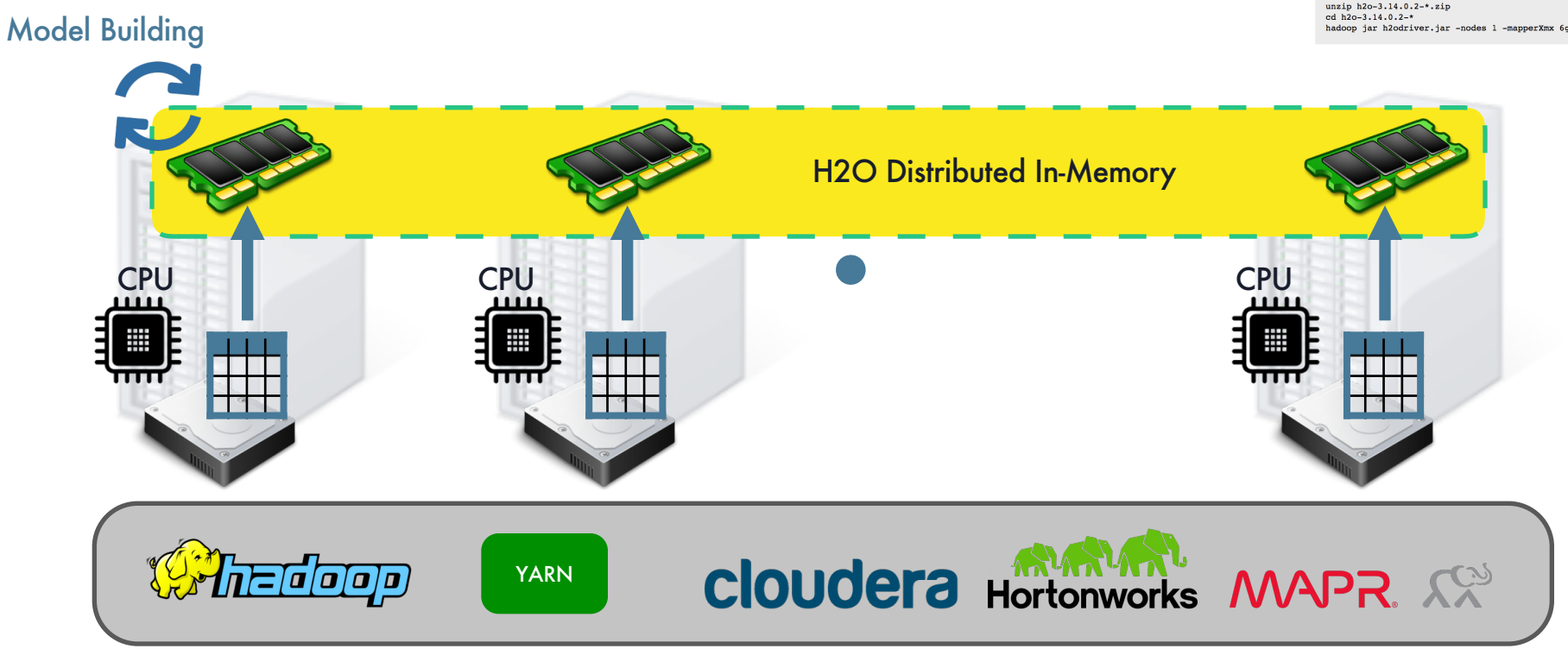
Use Case 2

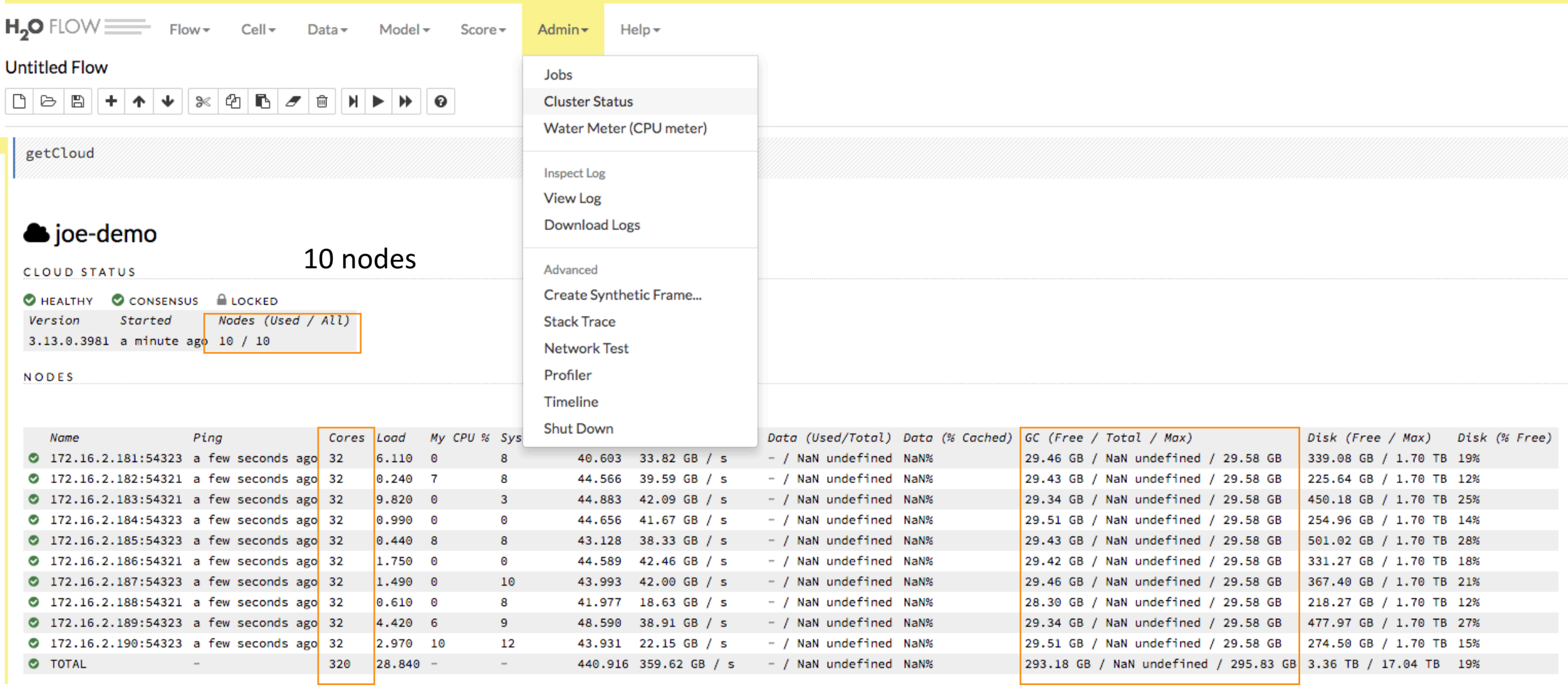
- Import large datasets directly from HDFS
- Train models on a multi-node cluster

Running H_2O on Hadoop

www.h2o.ai/download





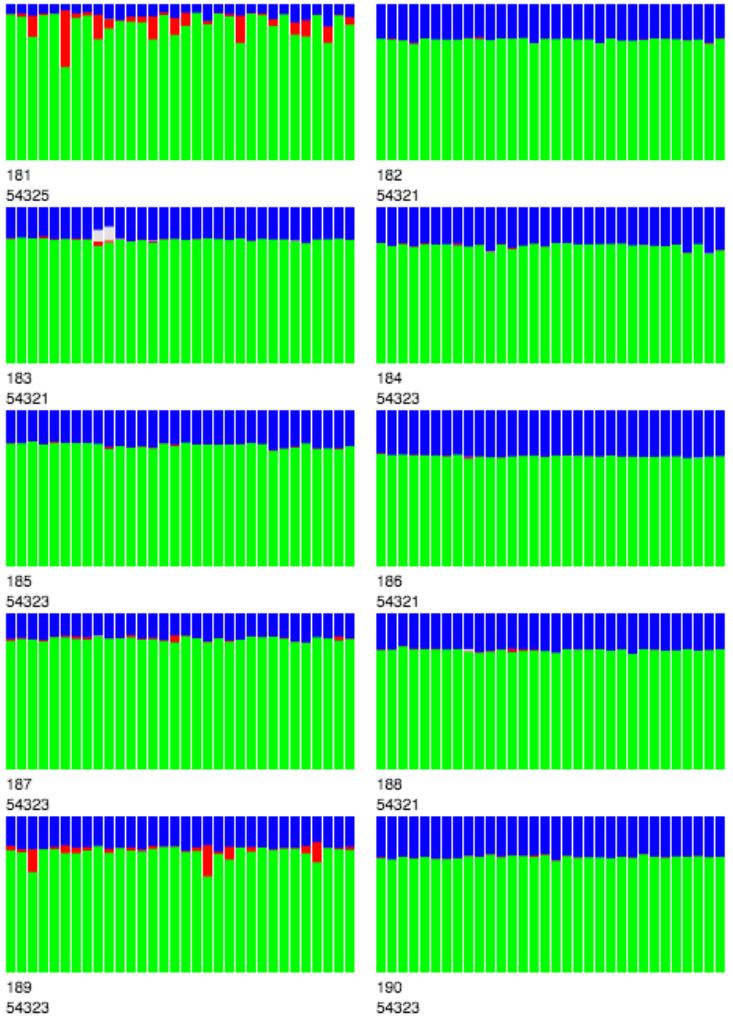


10 x 32 =
320 Cores

10 x 29.6 = 296
GB Memory

H2O Water Meter (CPU Monitor)

10 x 32 = 320 Cores



Use Case 2 – Freddie Mac Loans

http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html

Single Family Loan-Level Datas

www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html

Freddie Mac Home | Single-Family | Multifamily | Capital Markets | Renters, Buyers, & Owners | Search

Freddie Mac
We make home possible®

PERSPECTIVES | RESEARCH | BLOG | MEDIA ROOM | ABOUT | OUR BUSINESSES

Mortgage Rates | Insight | Outlook | Consumer Research | Indices | Additional Datasets

Single Family Loan-Level Dataset

As part of a larger effort to increase transparency, Freddie Mac is making available loan-level credit performance data on a portion of fully amortizing fixed-rate mortgages that the company purchased or guaranteed from 1999 to 2016.

The availability of this data will help investors build more accurate credit performance models in support of ongoing risk sharing initiatives highlighted by our regulator, the Federal Housing Finance Agency in the 2017 conservatorship scorecard [PDF]

The dataset covers approximately 24.5 million fixed-rate mortgages (including HARP loans) originated between January 1, 1999 and June 30, 2016. Monthly loan performance data, including credit performance information up to and including property disposition, is being disclosed through December 31, 2016. Specific credit performance information in the dataset includes voluntary prepayments and loans that were Foreclosure Alternatives and REOs. Specific actual loss data in the dataset includes net sales proceeds, MI recoveries, non-MI recoveries, expenses, current deferred UPB, and due date of last paid installment.

The information in the historical dataset is unaudited and subject to change. Freddie Mac cannot guarantee the dataset is complete or error free. Read our disclaimer. The historical dataset is not to be construed as securities disclosure.

In addition, Freddie Mac requires a licensing agreement [PDF] for commercial redistribution of the data in its Single-Family Loan-Level Dataset. Use of the dataset continues to be free for non-commercial, academic/research and for limited use, subject to the applicable terms and conditions.

Resources

Additional Datasets

SINGLE-FAMILY LOAN-LEVEL DATASET

QUARTERLY REFINANCE STATISTICS

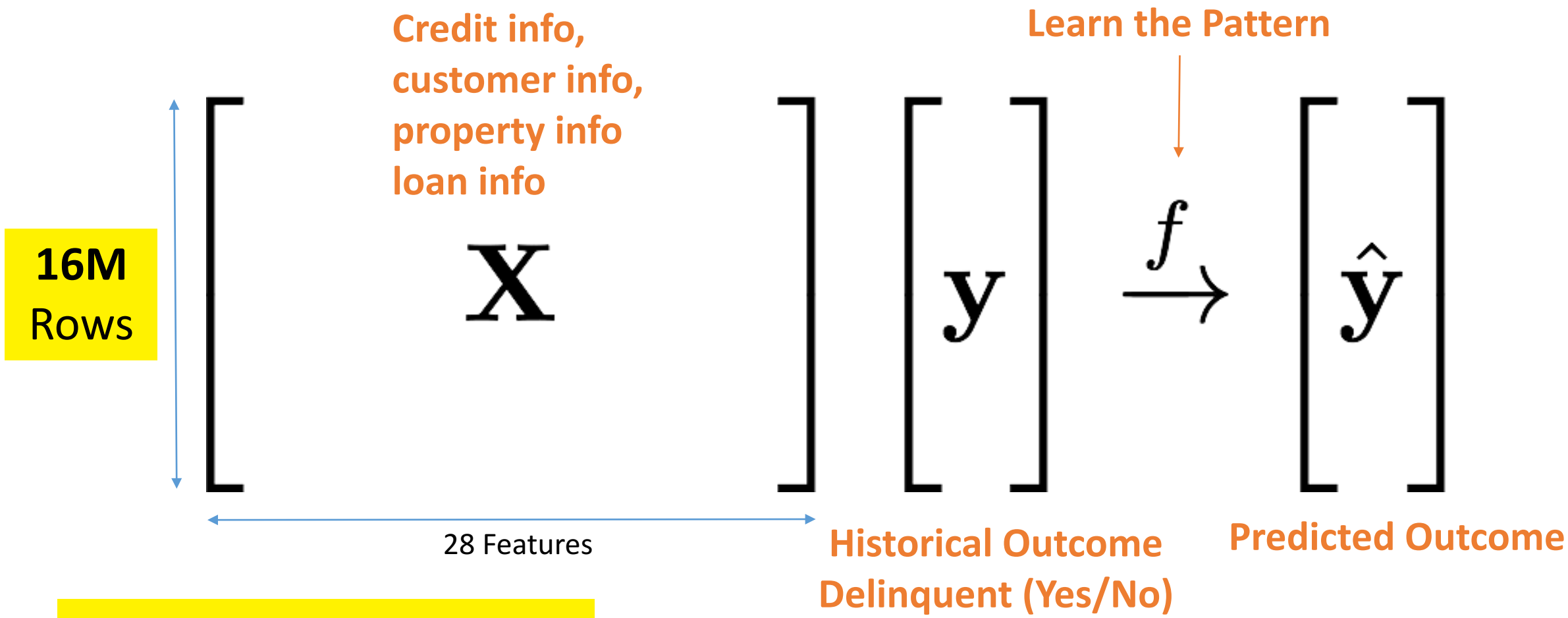
FEDERAL COST OF FUNDS INDEX

Authorized Licensees and Distributors

1010data INSIGHT. NOW.

CoreLogic

Learning from Freddie Mac (Loans) Data



Raw Data Size: **2.56 GB**

16M Rows

Size (Raw): **2.58 GB**
Compressed: **0.77 GB** (≈ 30% of Raw)

freddie_mac_loans

Actions: View Data | Split... | Build Model... | Predict | Download | Export | Delete

Rows	Columns	Compressed Size
16366856	29	769MB

COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
credit_score	int	36902	0	0	0	300.0	850.0	731.9424	55.1077	·	Convert to enum
first_payment_date	int	0	0	0	0	199901.0	201609.0	200527.8795	391.7542	·	Convert to enum
first_time_homebuyer_flag	enum	3975291	10846414	0	0	0	1.0	0.1247	0.3304	2	Convert to numeric
maturity_date	int	0	0	0	0	200309.0	205405.0	203359.6542	493.3003	·	Convert to enum
msa	int	2327122	0	0	0	10180.0	49740.0	30497.4987	11351.0992	·	Convert to enum
mortgage_insurance_percentage	int	233246	13064396	0	0	0	55.0	4.7115	10.1318	·	Convert to enum
number_of_units	int	194	0	0	0	1.0	4.0	1.0251	0.2063	·	Convert to enum
occupancy_status	enum	0	730645	0	0	0	2.0	·	·	3	Convert to numeric
cltv	int	933	0	0	0	6.0	199.0	72.9324	17.0772	·	Convert to enum
dti	int	294607	0	0	0	1.0	65.0	33.7514	11.7248	·	Convert to enum
original_upb	int	0	0	0	0	5000.0	1403000.0	181878.1577	98894.4838	·	Convert to enum
original_loan-to-value_(ltv)	int	844	0	0	0	6.0	105.0	71.6317	16.6764	·	Convert to enum
original_interest_rate	real	0	0	0	0	2.2500	13.9500	6.0189	1.1074	·	·
channel	enum	0	679430	0	0	0	3.0	·	·	4	Convert to numeric
prepayment_penalty_mortgage_flag	enum	105587	16240319	0	0	0	1.0	0.0013	0.0359	2	Convert to numeric

H2O.ai

Use Case 2 – Higgs Boson

Higgs Boson Machine Learning Challenge

Use the ATLAS experiment to identify the Higgs boson

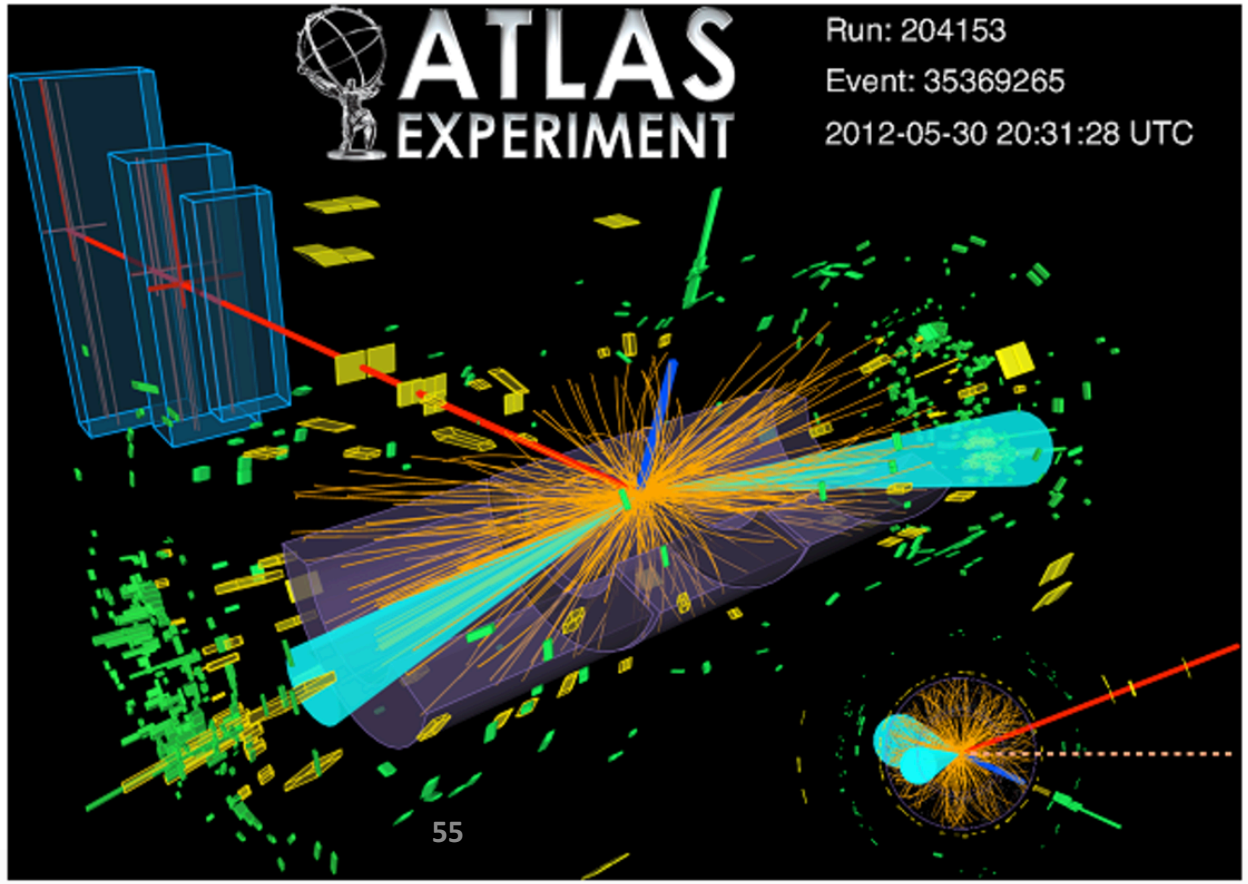
$13,000 · 1,785 teams · 3 years ago

Overview | Data | Discussion | Leaderboard | Rules | Team | My Submissions | Late Submission

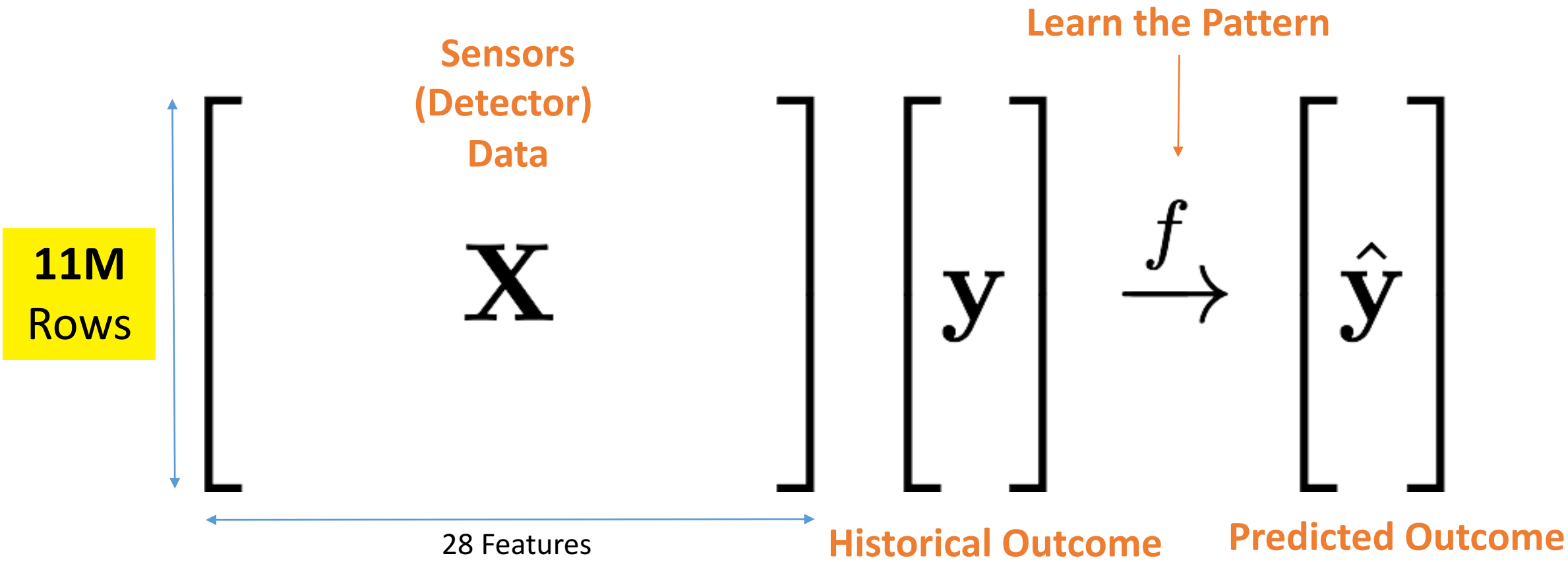
https://www.kaggle.com/c/higgs-boson

Overview

- Description
- Evaluation
- Prizes
- About The Sponsors
- Timeline
- Winners



Learning from Higgs Boson Machine Data



Raw Data Size: **7.48 GB**

11M Rows

Size (Raw): **7.48 GB**
Compressed: **2.00 GB** (≈ 27% of Raw)

HIGGS.hex

Actions: View Data | Split... | Build Model... | Predict | Download | Export

Rows	Columns	Compressed Size
11000000	29	2GB

COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
C1	enum	0	5170877	0	0	0	1.0	0.5299	0.4991	2	Convert to numeric
C2	real	0	0	0	0	0.2747	12.0989	0.9915	0.5654	·	·
C3	real	0	0	0	0	-2.4350	2.4349	-0.0	1.0088	·	·
C4	real	0	0	0	0	-1.7425	1.7432	-0.0	1.0063	·	·
C5	real	0	0	0	0	0.0002	15.3968	0.9985	0.6000	·	·
C6	real	0	0	0	0	-1.7439	1.7433	0.0	1.0063	·	·
C7	real	0	0	0	0	0.1375	9.9404	0.9909	0.4750	·	·
C8	real	0	0	0	0	-2.9697	2.9697	-0.0	1.0093	·	·
C9	real	0	0	0	0	-1.7412	1.7415	0.0	1.0059	·	·
C10	real	0	5394611	0	0	0	2.1731	1.0	1.0278	·	·
C11	real	0	0	0	0	0.1890	11.6471	0.9927	0.5000	·	·
C12	real	0	0	0	0	-2.9131	2.9132	-0.0	1.0093	·	·
C13	real	0	0	0	0	-1.7424	1.7432	-0.0	1.0062	·	·
C14	real	0	5523912	0	0	0	2.2149	1.0	1.0494	·	·
C15	real	0	0	0	0	0.2636	14.7090	0.9923	0.4877	·	·
C16	real	0	0	0	0	-2.7297	2.7300	0.0	1.0087	·	·
C17	real	0	0	0	0	-1.7421	1.7429	0.0	1.0063	·	·
C18	real	0	6265240	0	0	0	2.5482	1.0	1.1937	·	·
C19	real	0	0	0	0	0.3654	12.8826	0.9861	0.5058	·	·
C20	real	0	0	0	0	-2.4973	2.4980	-0.0	1.0077	·	·

H2O.ai

Use Case 2 Live Demo

- Import large datasets directly from HDFS
- Train models on a multi-node cluster

Q & A