使用 H_2O 进行机器学习

- 概观
  - 公司 / 人员
- 开源的H2O机器学习平台 – H2O-3
  - 概观
  - 信用卡风险示例
- 商业的H2O机器学习平台 - Driverless AI
  - 概观
  - 信用卡风险示例
- Q & A 提问

Jo-fai (Joe) Chow 周祖輝
数据科学家
joe@h2o.ai

公司简介

Founded 成立	2011年获得风险投资支持，在2012年首次亮相
Products 产品	• H_2O Open Source Machine Learning Platform 开源的H2O机器学习平台 • Sparkling Water (H_2O + Spark) • Enterprise Steam • Driverless AI (商业的H2O机器学习平台)
Mission 任务	Operationalize Data Science, and provide a platform for users to build beautiful data products 实现数据科学的操作化，为用户构建美丽的数据产品提供平台
Team 团队	75 人 • 分布式系统工程师 • 机器学习专家 • 世界级的可视化设计师
HQ 总部	Mountain View, California 加州山景城





H2O.ai

Scientific Advisory Council
科学咨询委员会

Dr. Trevor Hastie

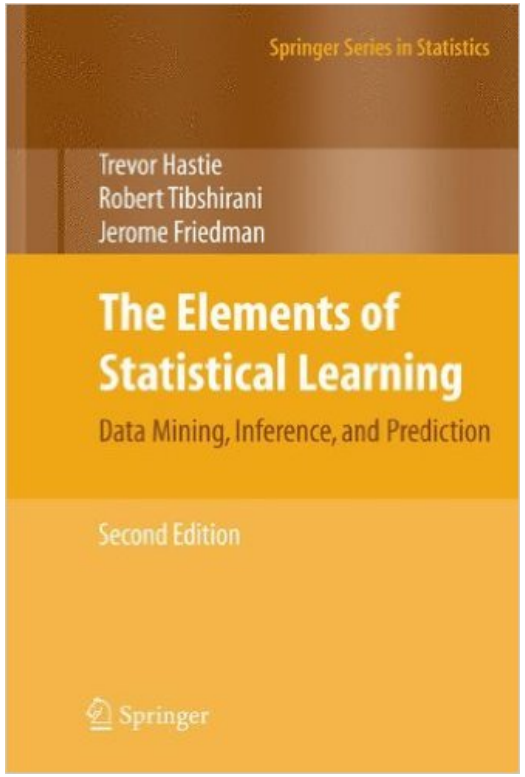
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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


Amy Vu-Tran


Angela Bartz


Anmol Bal


Apoorv Chaudhary


Arno Candel


Ashrith Barthur


Leland Wilkinson


Magnus Stensmo


Maral Mandjian


Marios Michailidis


Mark Chan


Mark Landry


Avkash Chauhan


Avni Wadhwa


Beth Payne


Branden Murray


Cari Andrews


Das Narayanan


Mateusz Dymczyk


Mathias Müller


Matt Dowle


Megan Kurka


Michal Kurka


Michal Malohlava


Dmitry Larko


Erin LeDell


Fonda Ingram


Ian Gomez


Jacqueline Scott


Jag Bining


Navdeep Gill


Nidhi Mehta


Nikhil Shekhar


Nishant Kalonia


Pasha Stetsenko


Patrick Hall


Jakub Hava


Jan Gamec


Jeff Fohl


Jeff Gambera

Jo-Fai Chow

John Schiess

Patrick Rice

Prithvi Prabhu

Ravi Purushotma

Raymond Peck

Sanjay Joshi

SriSatish Ambati

Jon Oslowski

Josephine Wang

Justin Loyola

Karen Hayrapetyan

Kerry O'Shea

Lauren DiPerna

Terone Ward

Tom Aliotti

Tom Kraljevic

Venkatesh Yadav

Vinod Iyengar

Wen Phan

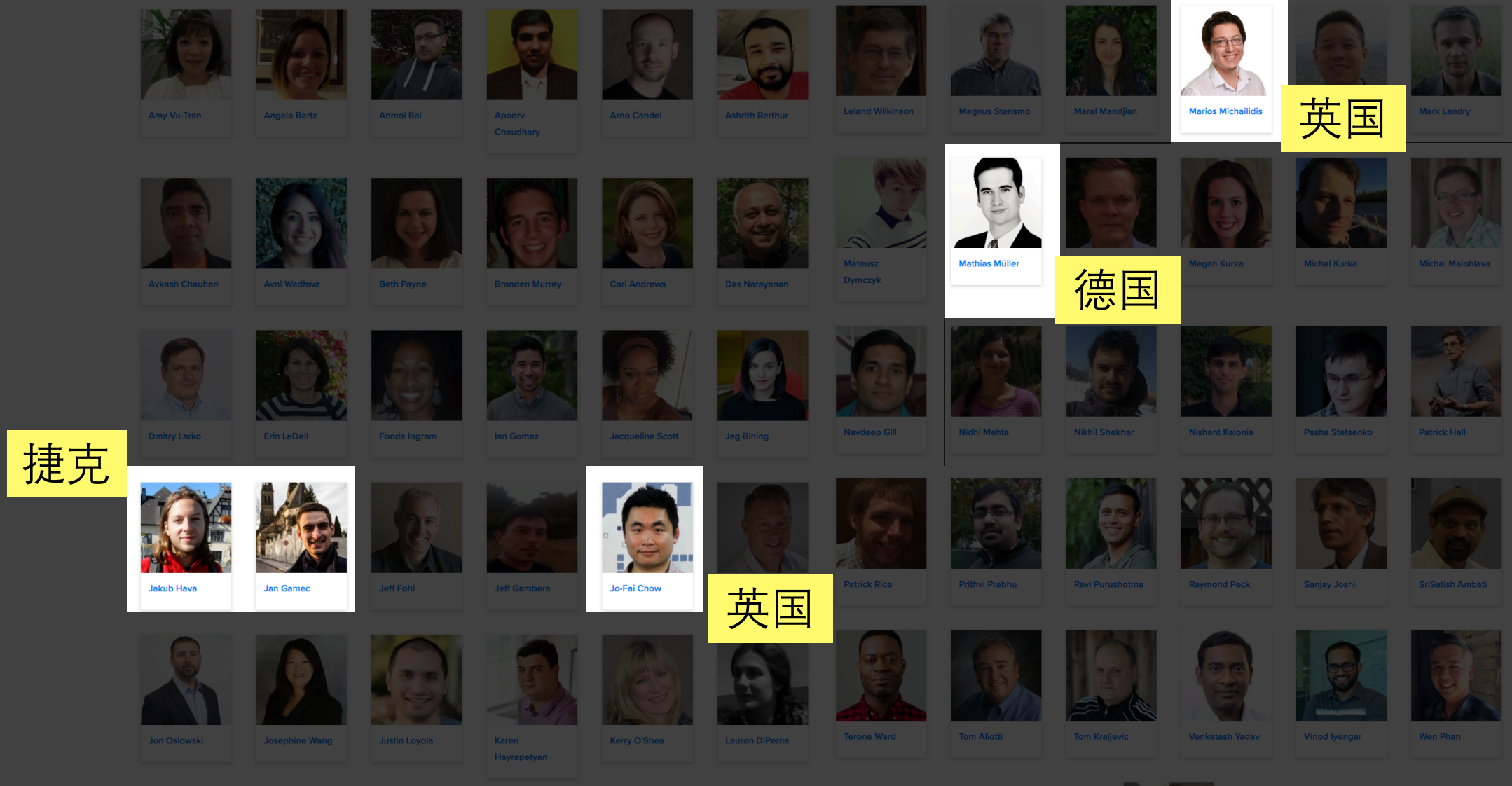
Wendy Wong

H_2O 团队

H_2O.ai

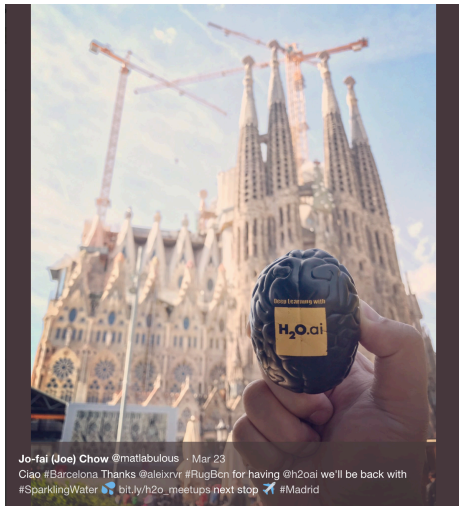
H_2O 欧洲团队(2016)

H_2O.ai



H_2O 欧洲团队(2017)

What is Joe's role at H_2O.ai 我的职责



- Data Scientist (数据科学家) /
  Sales Engineer (销售) /
  Meetup Organiser (Meetup组织者)
  (on paper)

- Unofficial Photographer of H2O.ai
  SWAG (非官方摄影师)
  (the travelling data scientist)

H_2O.ai

#AroundTheWorldWithH2Oai

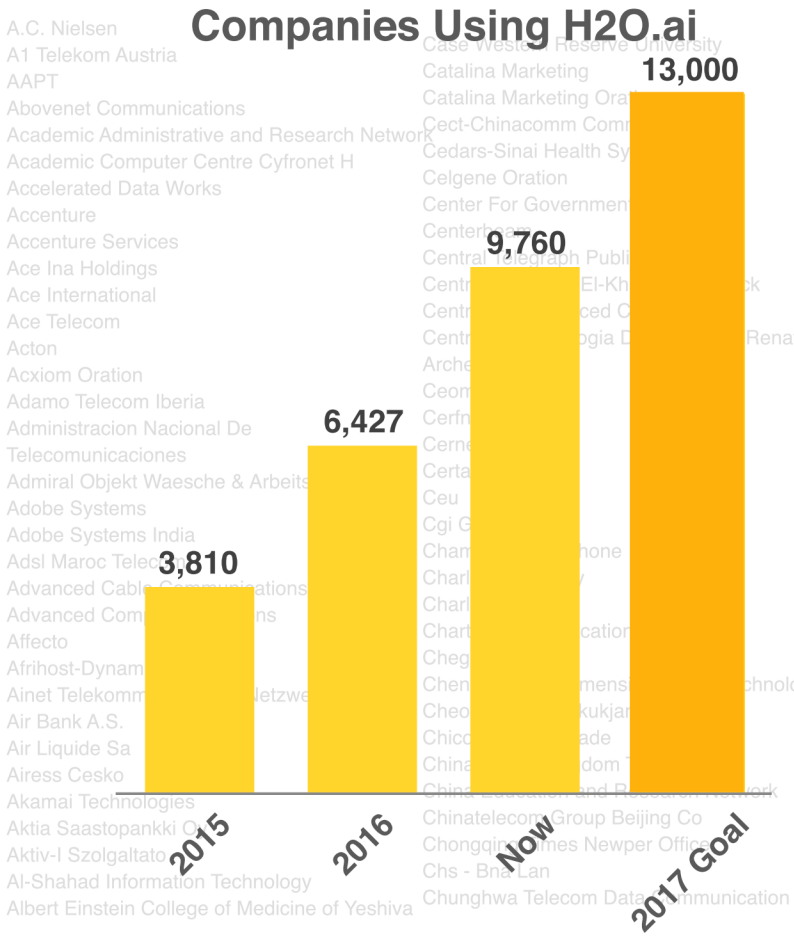
Month	Joe's H_2O Events in 2017
Jan	**London**
Feb	**London**, Warsaw, Oxford
Mar	Bay Area, **London**, Cologne, Barcelona, Madrid, Vienna
Apr	Amsterdam, Rotterdam, Poznan, **London**
May	Belgrade, Hamburg, Berlin
Jun	Amsterdam, Stockholm, Budapest, **London**, Munich, Prague
Jul	Berlin, Brussels
Aug	🏖
Sep	**London**, Dublin
Oct	Exeter, Munich, Dublin, The Hague, Amsterdam, Frankfurt
Nov	Munich, Zurich, **London**, Glasgow
Dec	Bay Area, **London**

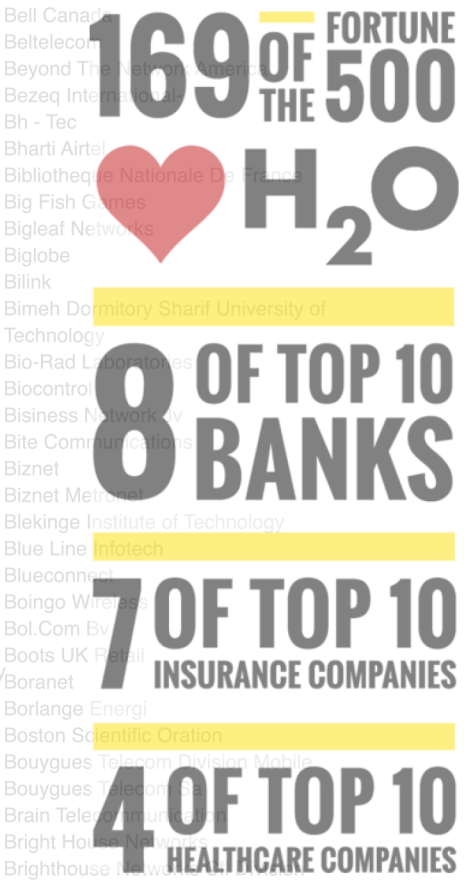
Meetups, Workshops & Conferences in 25+ Cities

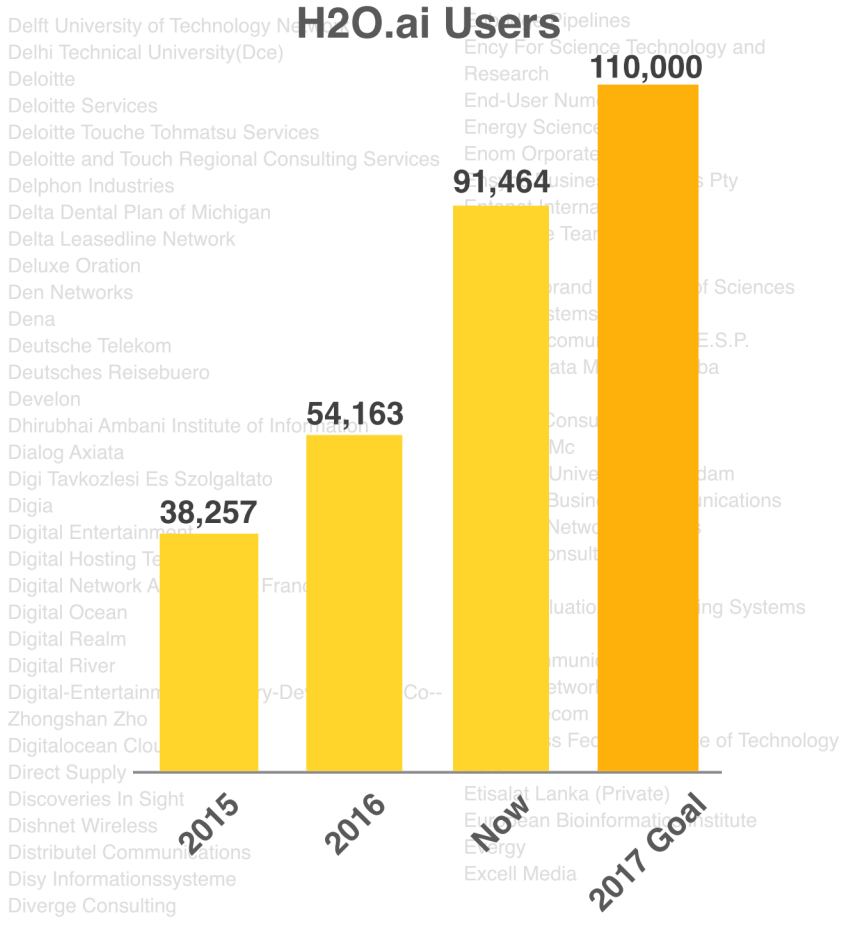
Jo-fai (Joe) Chow
Unofficial Photographer of H2O.ai Stress Ball

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

客户

"Overall customer satisfaction is very high." – Gartner

(整体客户满意度非常高)

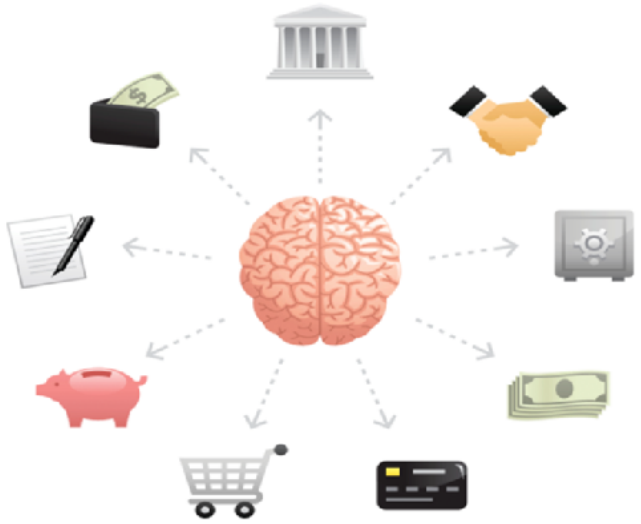
AI in Financial Services (金融服务 AI 应用)

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

Award-Winning Application (PwC + H_2O.ai) 获奖的项目

pwc | Global

PwC Global › About us › › Stories about our people, clients and communities

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

PwC @PwC Follow

Exciting night at this year's @WAI_News Awards: PwC wins 2017 Audit Innovation of the Year! pwc.to/Glai17 #taandiab17

10:15 PM - 4 Oct 2017

http://www.pwc.com/gx/en/about/stories-from-across-the-world/harnessing-the-power-of-ai-to-transform-the-detection-of-fraud-and-error.html

H_2O.ai

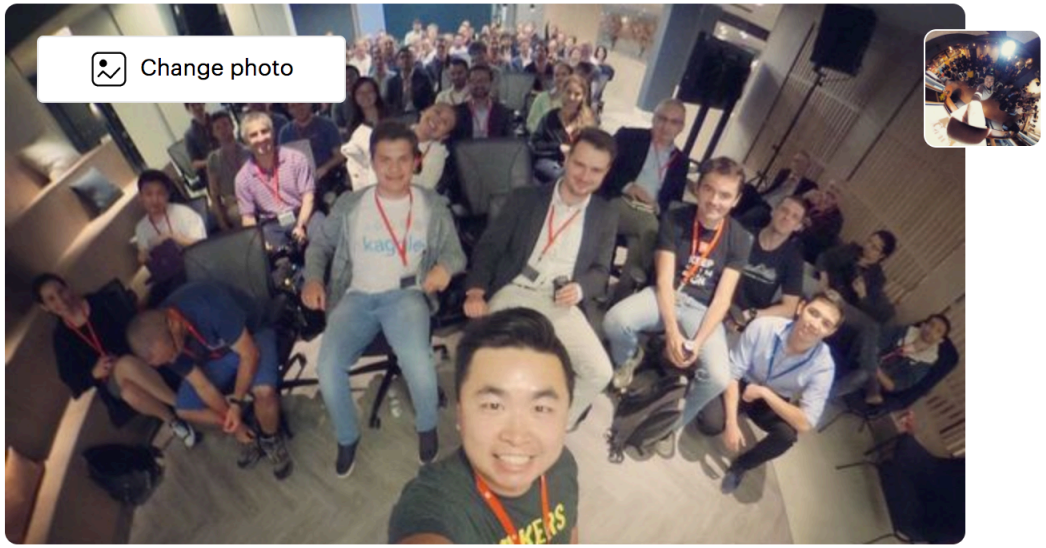
Community Expansion 社区扩展

15 Meetups a month

Meetup

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

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



London Artificial Intelligence & Deep Learning PRO

H2O Artificial Intelligence and Machine Learning -
39 groups

Location
London, United Kingdom
Members
4,184

4100+ Members

Organizers
Ian Gomez and **2 others**

Schedule

Our group | Meetups | Members | Photos | Discussions | More

Next Meetup

Next London Meetup: 12 Dec at Moody's London HQ

12 DEC

Tuesday, December 12, 2017, 6:00 PM

Interpretable Machine Learning, Tweet Classifier, H2O World Highlights and More

Hosted by Jo-fai Chow

Edit

Moody's Analytics
1 Canada Square, Canary Wharf, E14 5AB · London

Dear All, We are thrilled to see the growth of this meetup group (4100+ members right now). Let's end the year with one more exciting meetup. This time we will have speakers from Aviva, Theodo and Barclays. Many thanks to our friends from Moody's Analytics, we have a super cool venue right in the heart of Canary Wharf. They also very kindly provide food and drinks for the event. Agenda (T.B.C.): - Doors open at 6 for pizzas and drinks as usual. -

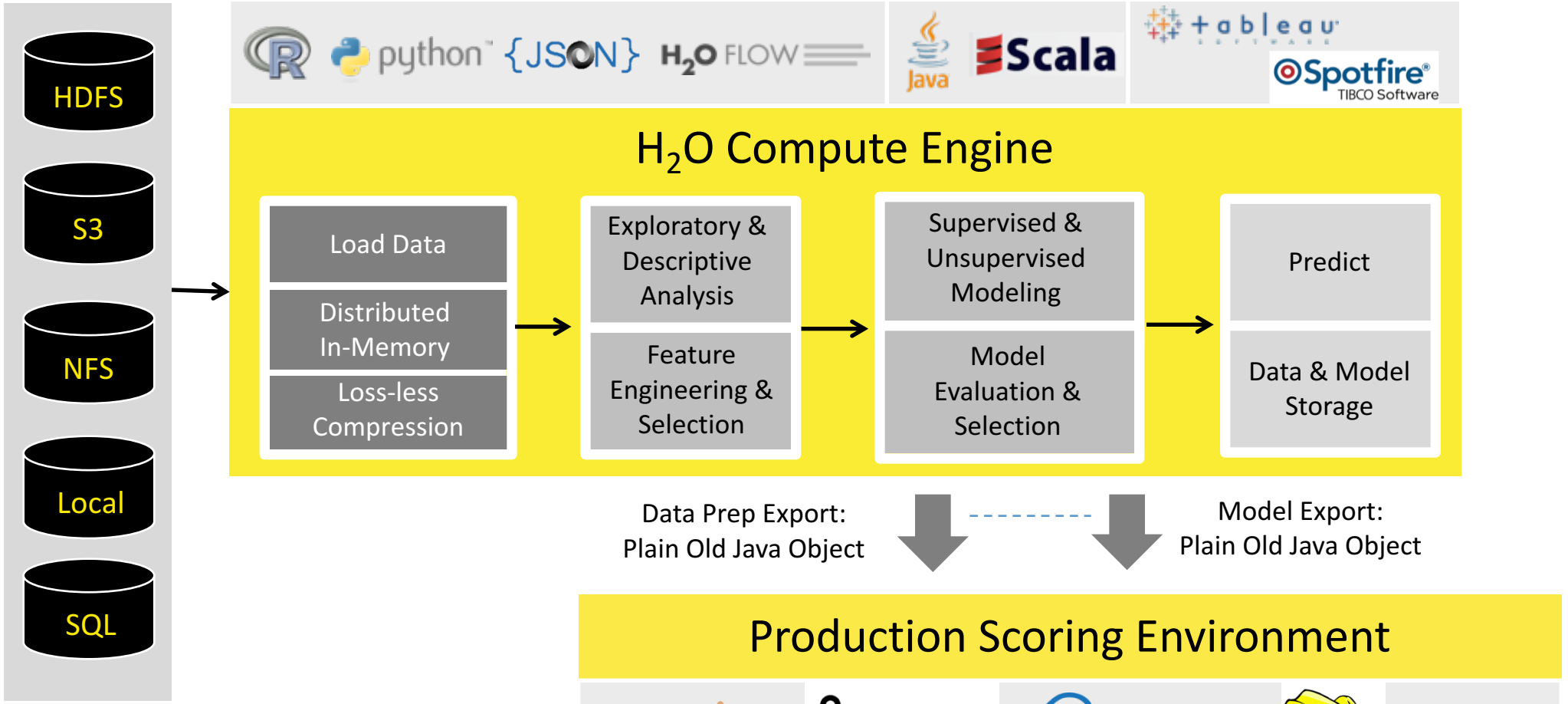
H2O is back with its flagship event, H2O World 2017.

Whether you're just getting started with H2O or you're a power user looking to expand your skill set even more, join

H_2O Machine Learning Platform

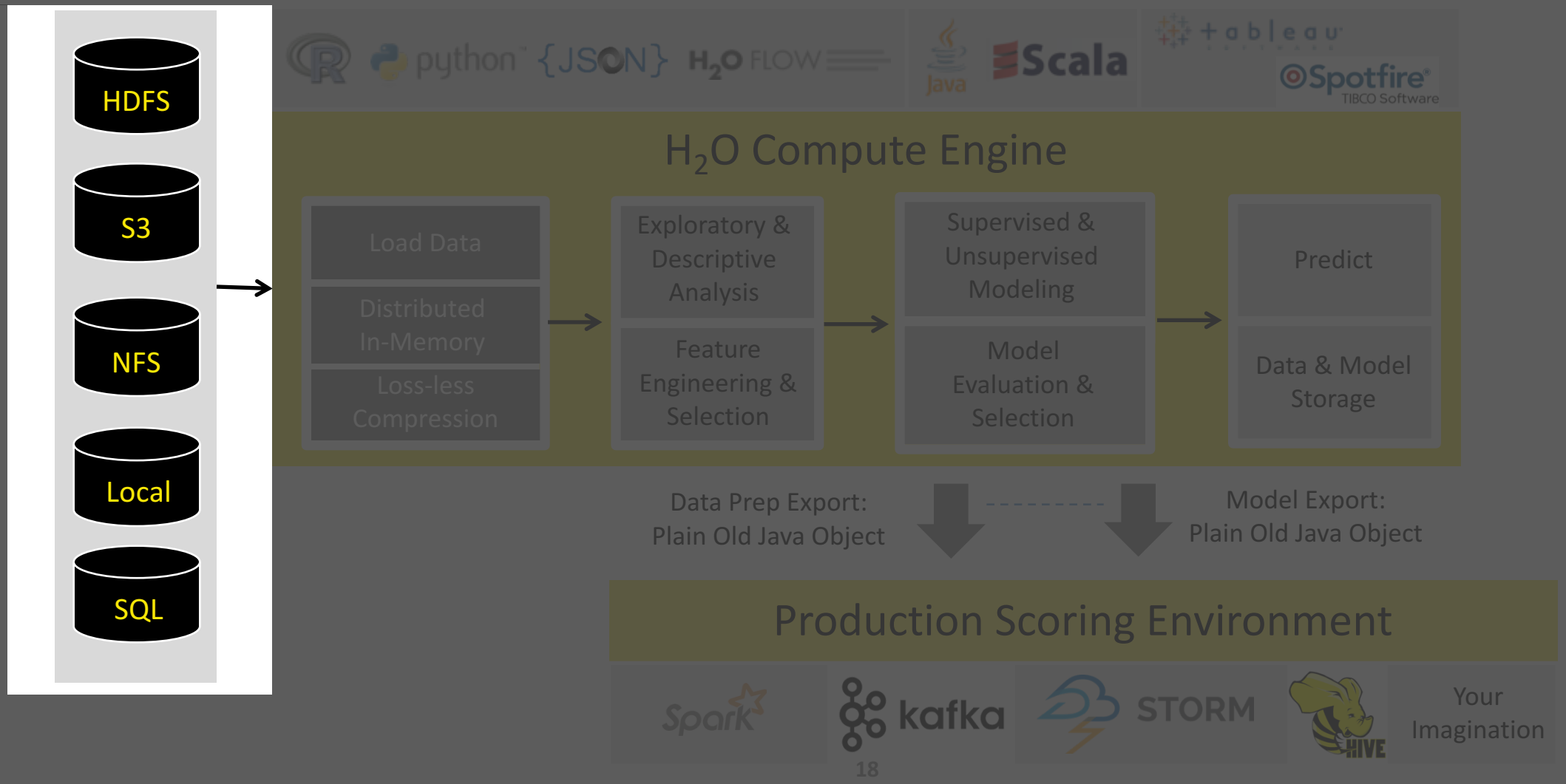
开源的 H_2O 机器学习平台

High Level Architecture 架构



High Level Architecture

从多个来源导入数据

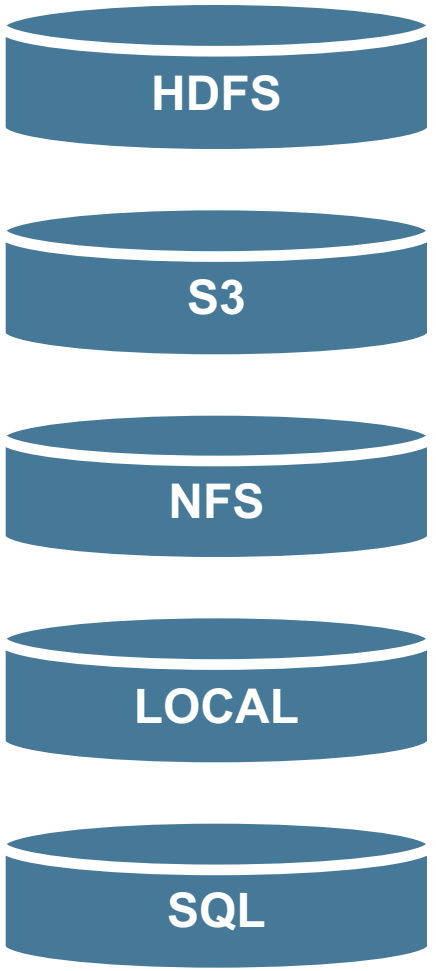


Supported Formats & Data Sources 支持的格式和数据源

9 Formats

File type or Folder of Files

- CSV
- XLS
- XLSX
- ORC*
- Hive*
- SVMLight
- ARFF
- Parquet
- Avro 1.8.0*

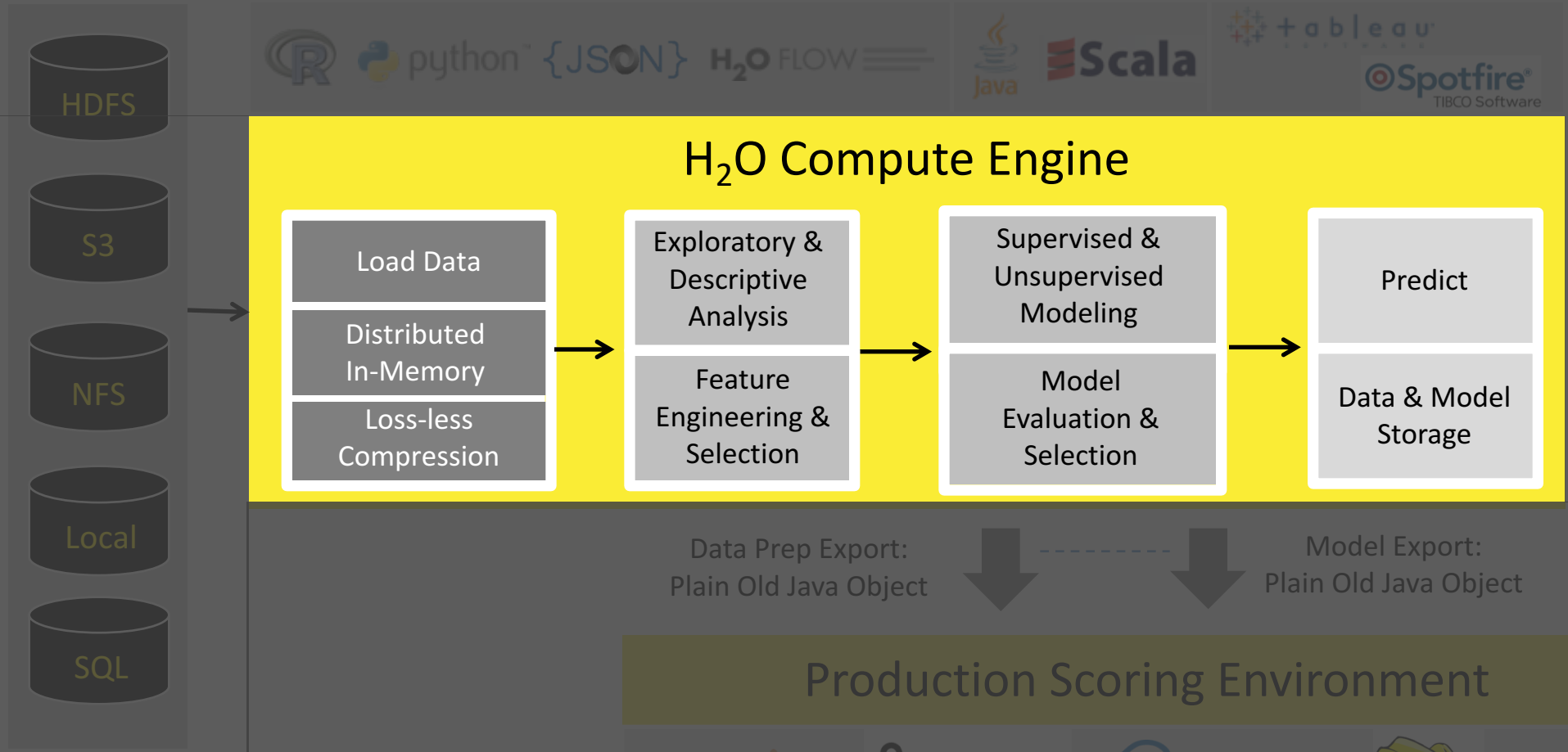


5 Sources

* 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

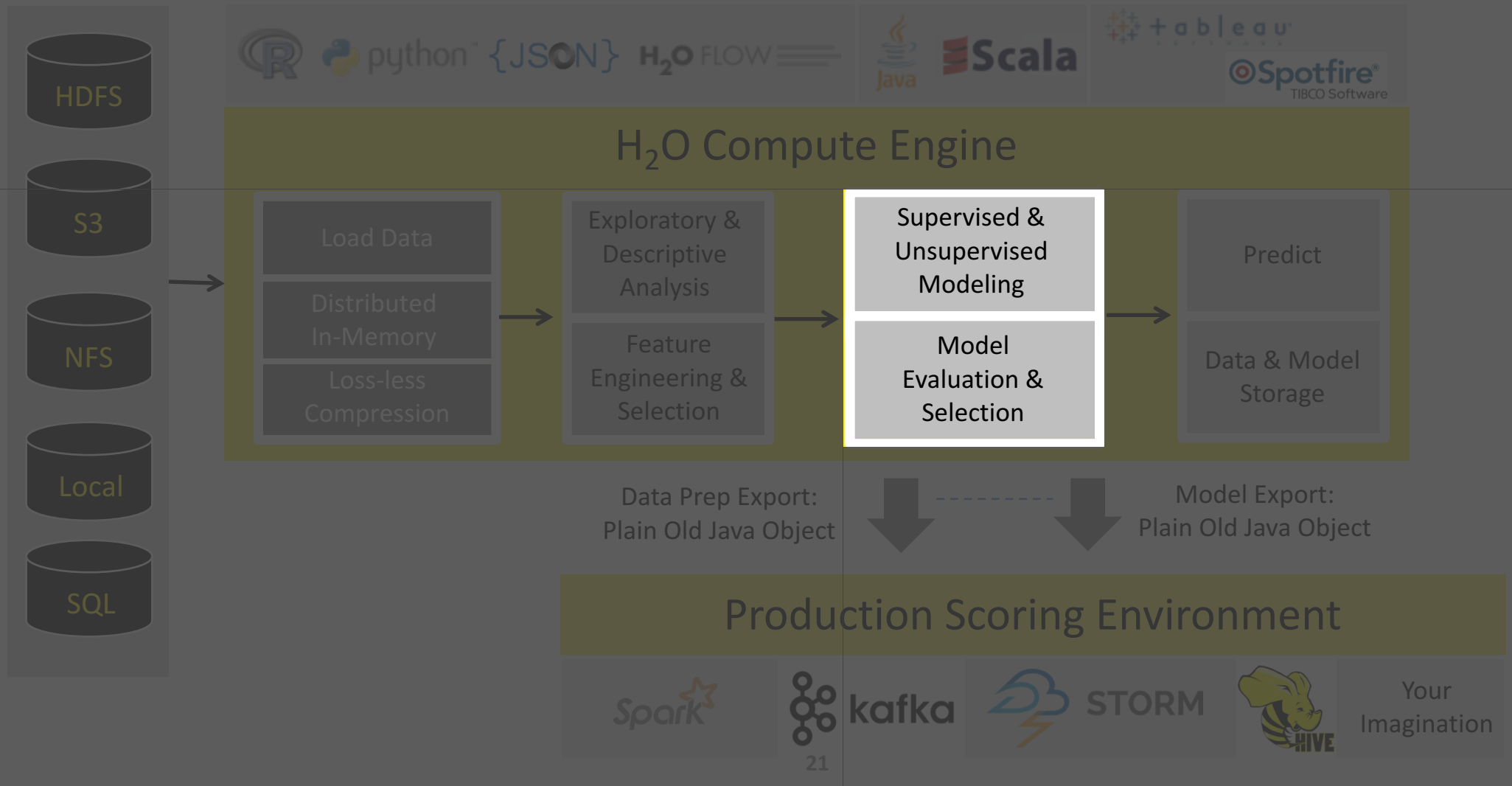
快速，可扩展和分布式计算引擎

High Level Architecture



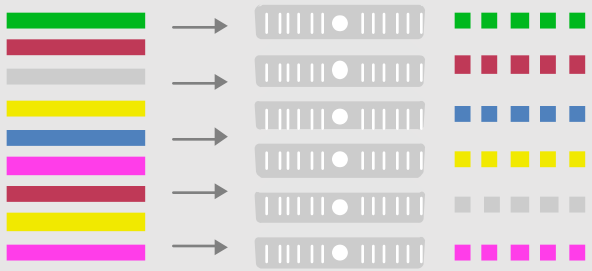
分布式算法

High Level Architecture

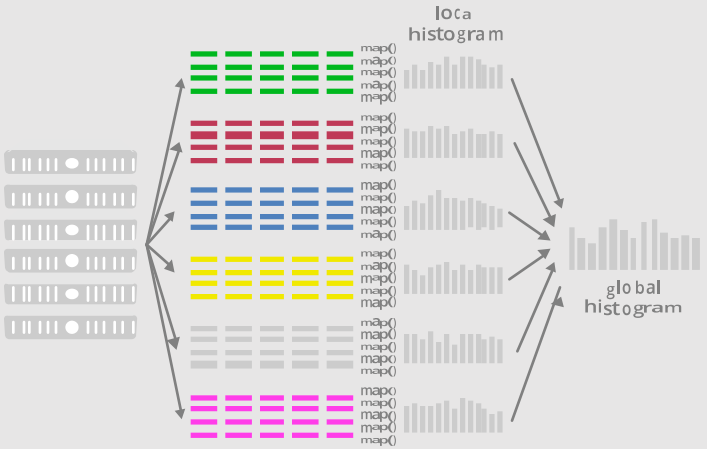


Distributed Algorithms 分布式算法

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**

Algorithms Overview 算法概述

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

Docs » AutoML: Automatic Machine Learning View page source

AutoML: Automatic Machine Learning

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software that can be used by non-experts. The first steps toward simplifying machine learning involved developing simple, unified interfaces to a variety of machine learning algorithms (e.g. H2O).

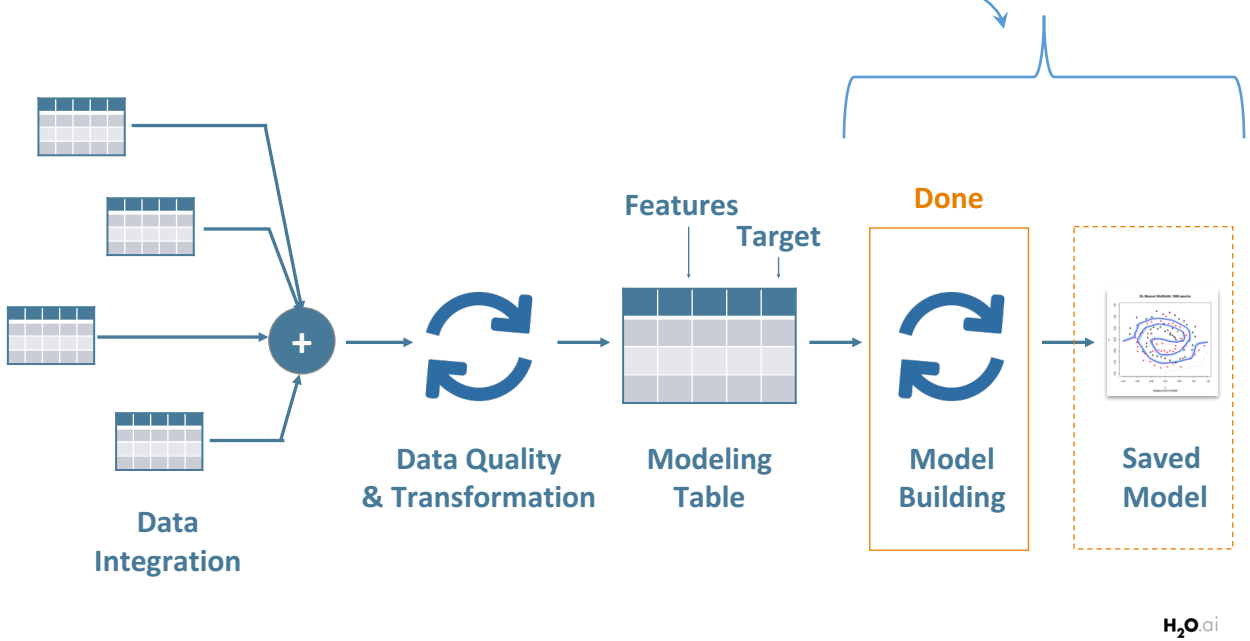
Although H2O has made it easy for non-experts to experiment with machine learning, there is still a fair bit of knowledge and background in data science that is required to produce high-performing machine learning models. Deep Neural Networks in particular are notoriously difficult for a non-expert to tune properly. In order for machine learning software to truly be accessible to non-experts, we have designed an easy-to-use interface which automates the process of training a large selection of candidate models. H2O's AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.

H2O's AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit. The user can also use a performance metric-based stopping criterion for the AutoML process rather than a specific time constraint. Stacked Ensembles will be automatically trained on the collection individual models to produce a highly predictive ensemble model which, in most cases, will be the top performing model in the AutoML Leaderboard. Stacked ensembles are not yet available for multiclass classification problems, so in that case, only singleton models will be trained.

AutoML Interface

The H2O AutoML interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time constraint, a maximum number of models constraint, and early stopping parameters.

自动机器学习



H2O.ai

AutoML Output

自动机器学习

The AutoML object includes a “leaderboard” of models that were trained in the process, ranked by a default metric based on the problem type (the second column of the leaderboard). In binary classification problems, that metric is AUC, and in multiclass classification problems, the metric is mean per-class error. In regression problems, the default sort metric is deviance. Some additional metrics are also provided, for convenience.

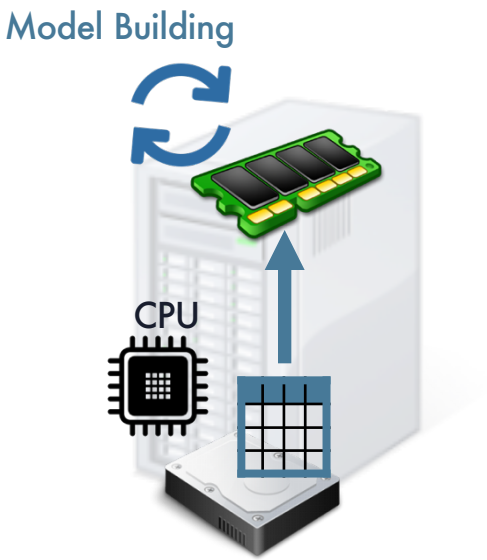
Here is an example leaderboard for a binary classification task:

model_id	auc	logloss
StackedEnsemble_0_AutoML_20170605_212658	0.776164	0.564872
GBM_grid_0_AutoML_20170605_212658_model_2	0.75355	0.587546
DRF_0_AutoML_20170605_212658	0.738885	0.611997
GBM_grid_0_AutoML_20170605_212658_model_0	0.735078	0.630062
GBM_grid_0_AutoML_20170605_212658_model_1	0.730645	0.67458
XRT_0_AutoML_20170605_212658	0.728358	0.629296
GLM_grid_0_AutoML_20170605_212658_model_1	0.685216	0.635137
GLM_grid_0_AutoML_20170605_212658_model_0	0.685216	0.635137

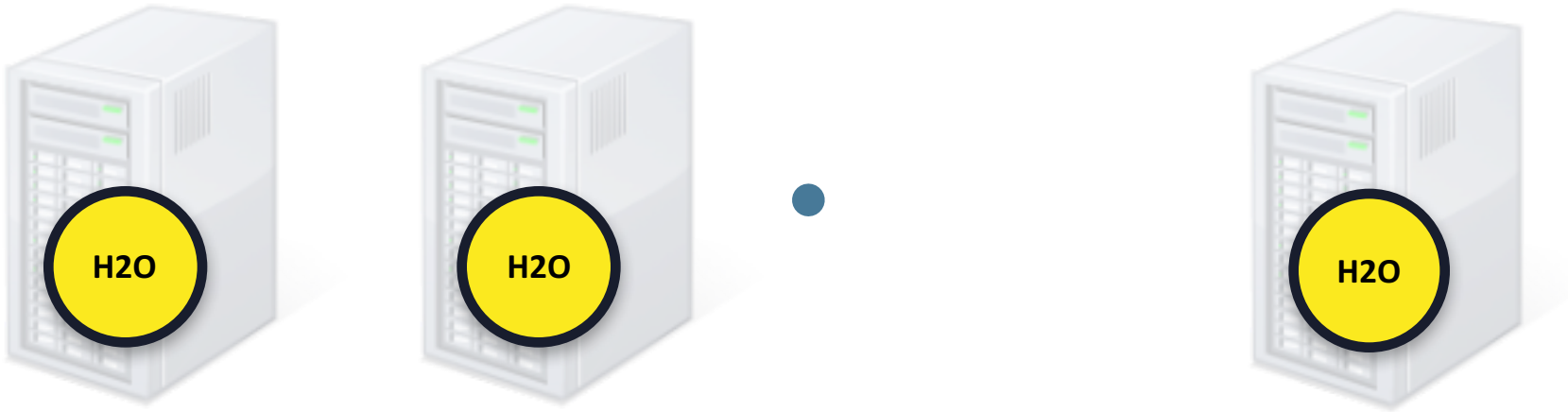
H2O.ai

H_2O Core 核心

H_2O Core 核心

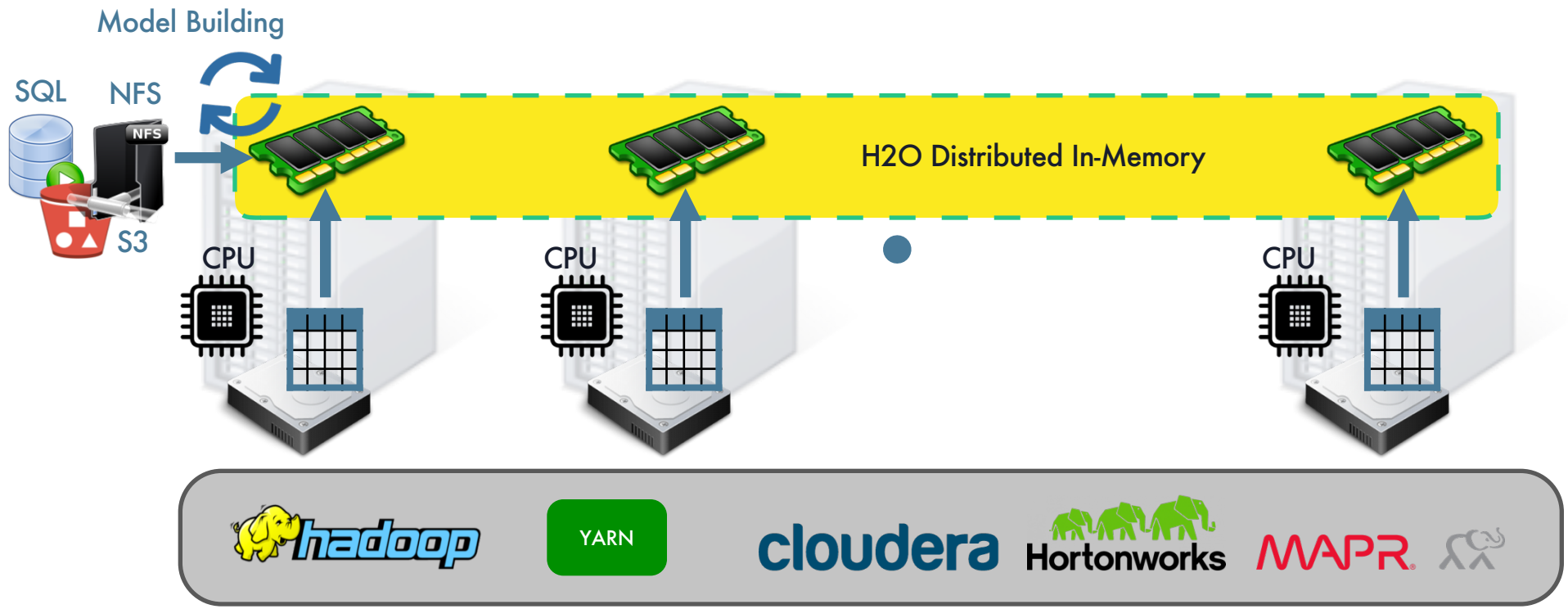


H_2O Core 核心

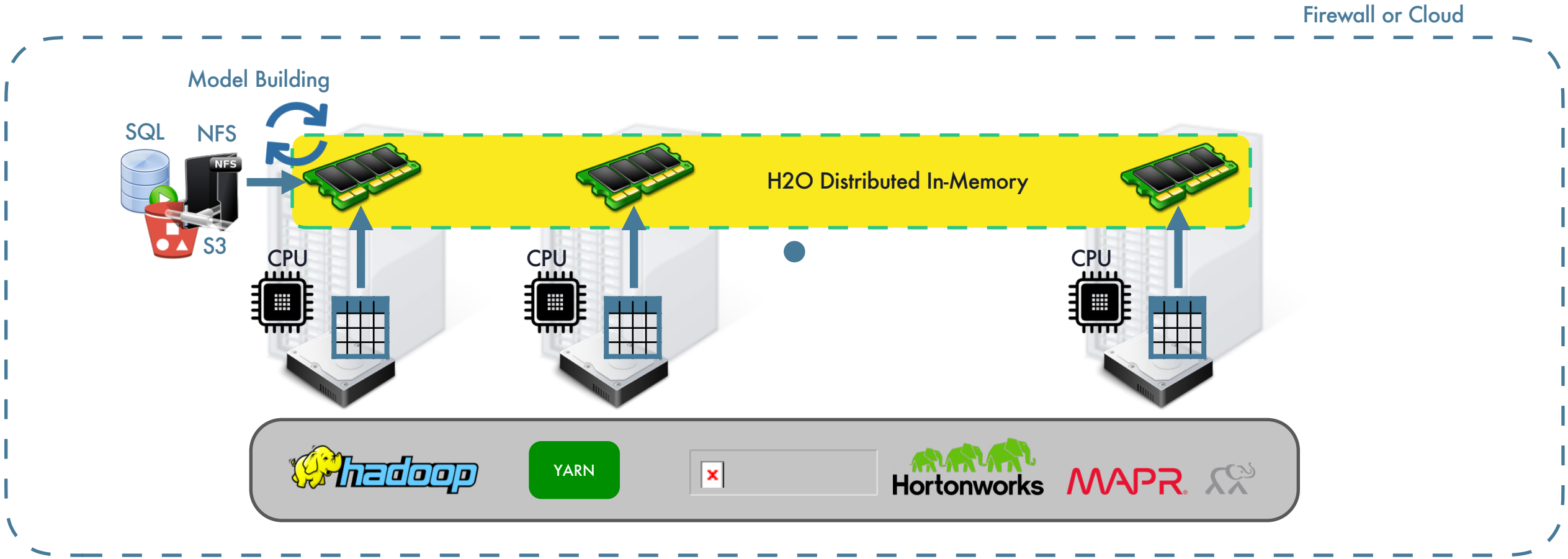


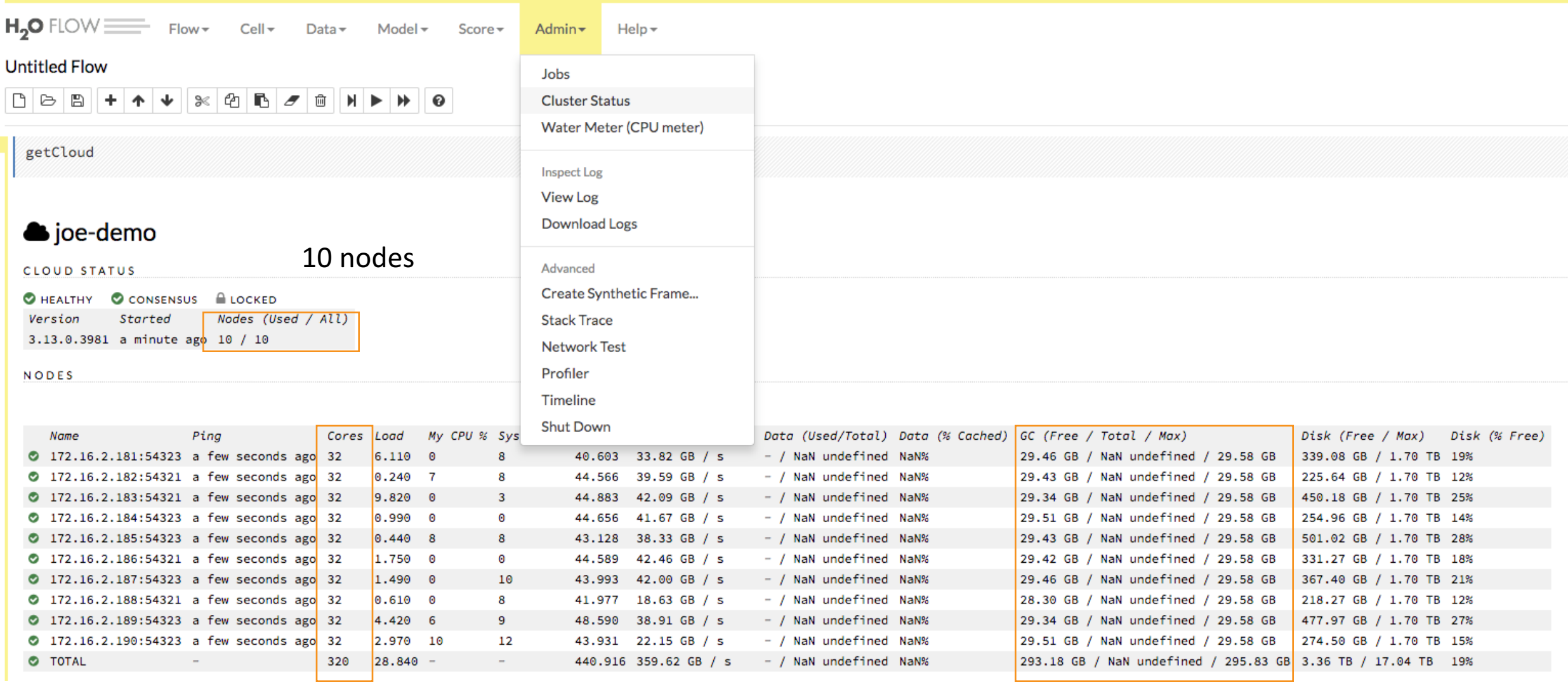
H_2O Core 核心

H_2O Core 核心


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

H_2O Core 核心


Firewall or Cloud
Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
Hortonworks
MAPR

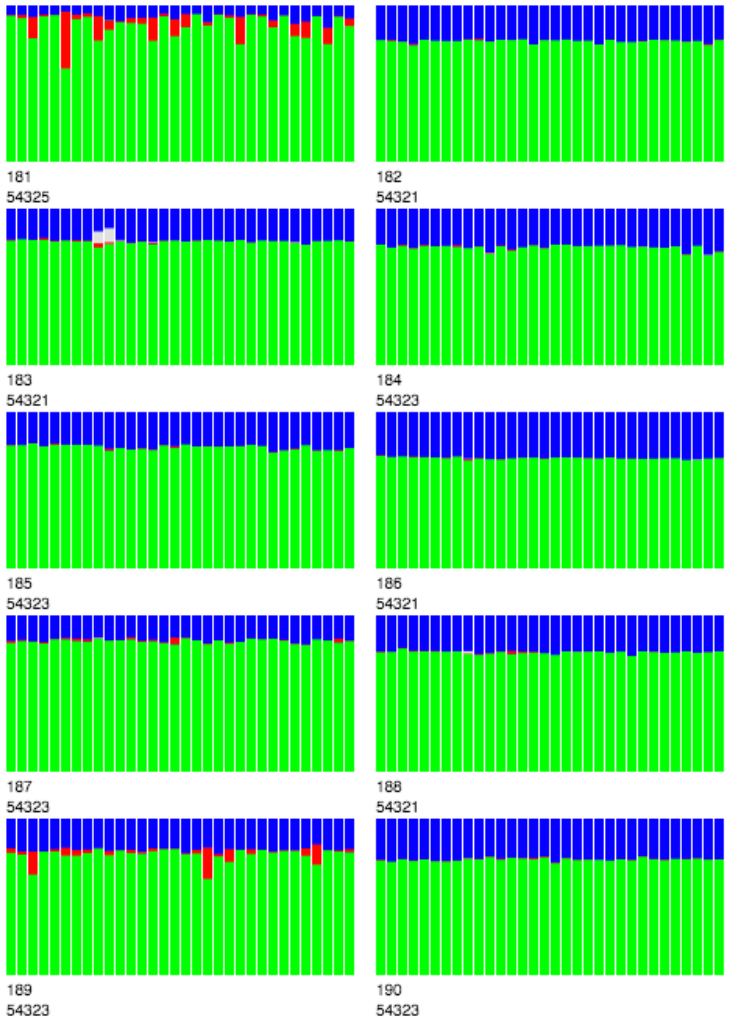


10 x 32 = 320
CPU核

10 x 29.6 = 296
GB 记忆体

H_2O Water Meter (CPU 功率监视器)

10 x 32 = 320 Cores



Legend

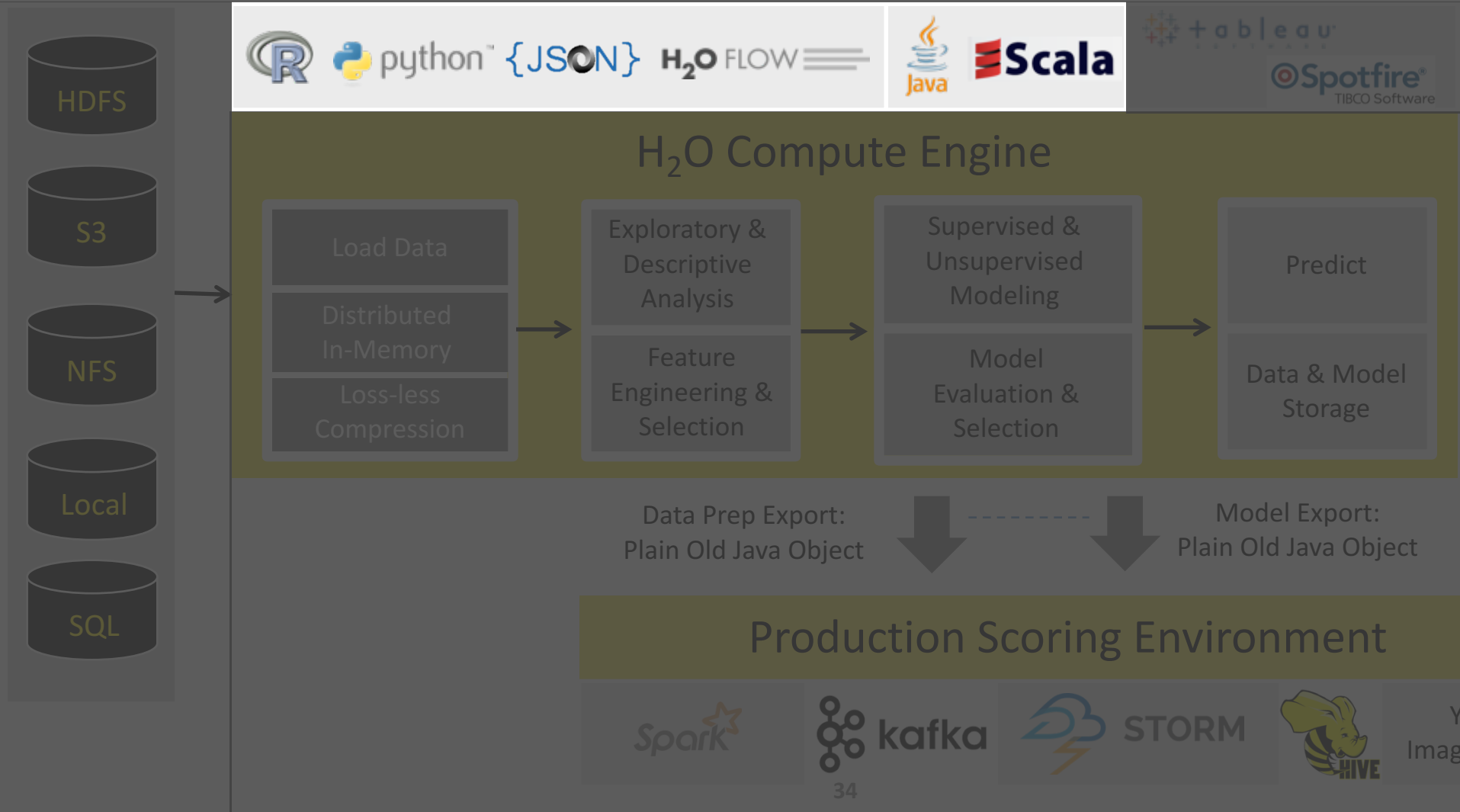
Each bar represents one CPU.

Blue: idle time
Green: user time
Red: system time
White: other time (e.g. i/o)

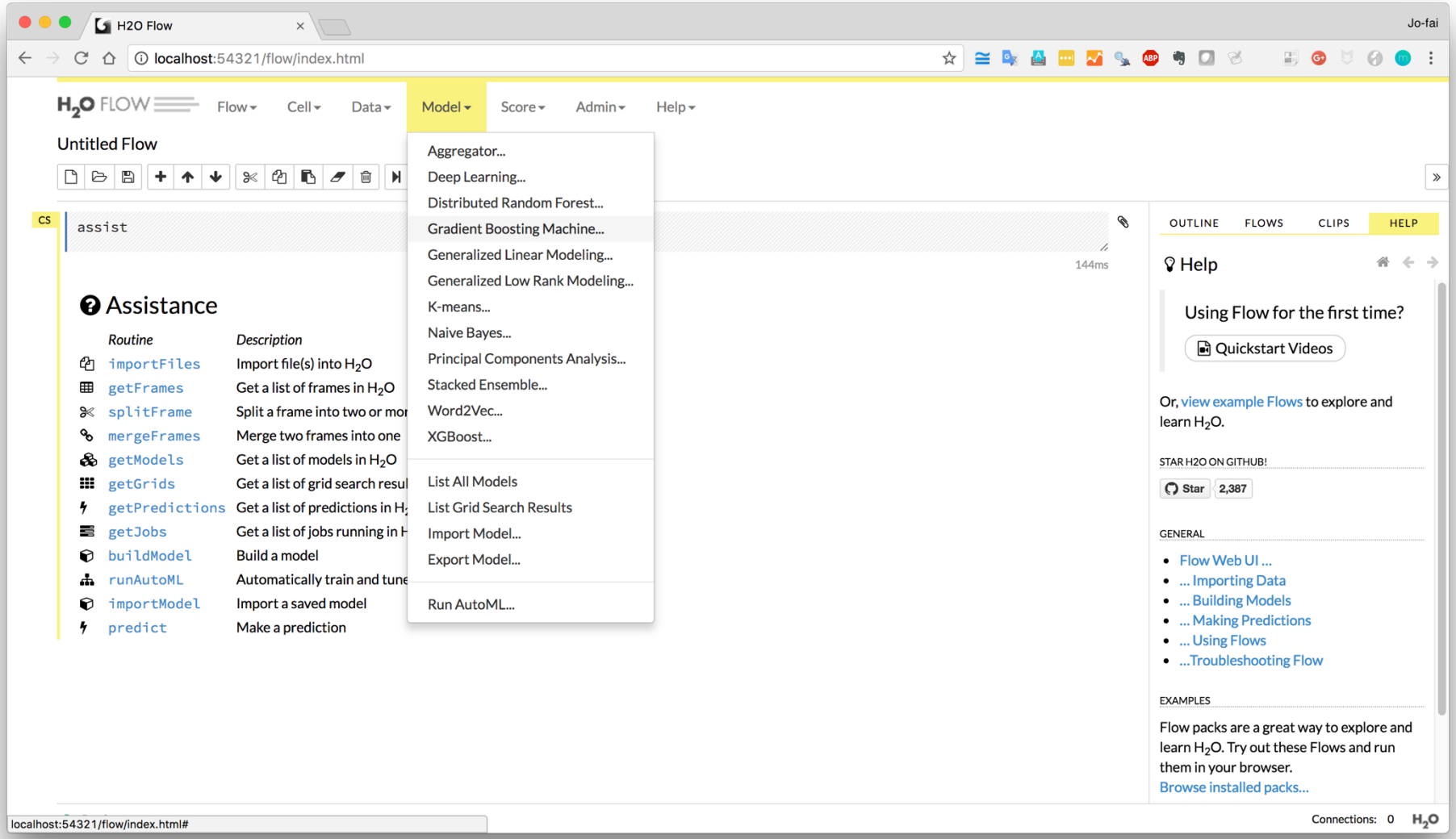
H2O.ai

High Level Architecture

多介面



H_2O Flow (网页介面)



H2O.ai

H_2O - R / Python 介面

~/Documents/repo_h2o/sales-engineering - master - RStudio Source Editor

credit_card_example.R

```r
# Credit Card Example

# Datasets:
# https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv
# https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv

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

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

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

# Define features and target
features = colnames(df_test)
target = "DEFAULT_PAYMENT_NEXT_MONTH"

# Train a GBM model
model_gbm = h2o.gbm(x = features,
                    y = target,
                    training_frame = df_train,
                    seed = 1234)
print(model_gbm)

# Use GBM model for making predictions
yhat_test = h2o.predict(model_gbm, newdata = df_test)
head(yhat_test)

# (Extra) Use H2O's AutoML
aml = h2o.automl(x = features,
                 y = target,
                 training_frame = df_train,
                 max_runtime_secs = 60,
                 seed = 1234)

# Print leaderboard
print(aml@leaderboard)

# Use best model for making predictions
best_model = aml@leader
yhat_test = h2o.predict(best_model, newdata = df_test)
head(yhat_test)
```

49:1 (Top Level) R Script

credit_card_example

localhost:8888/notebooks/credit_card_example.ipynb

jupyter credit_card_example Last Checkpoint: 5 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

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_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av
  JVM stdout: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av/h2o_jofaichow_started_from_python.out
  JVM stderr: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av/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_id7iqa
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%
```

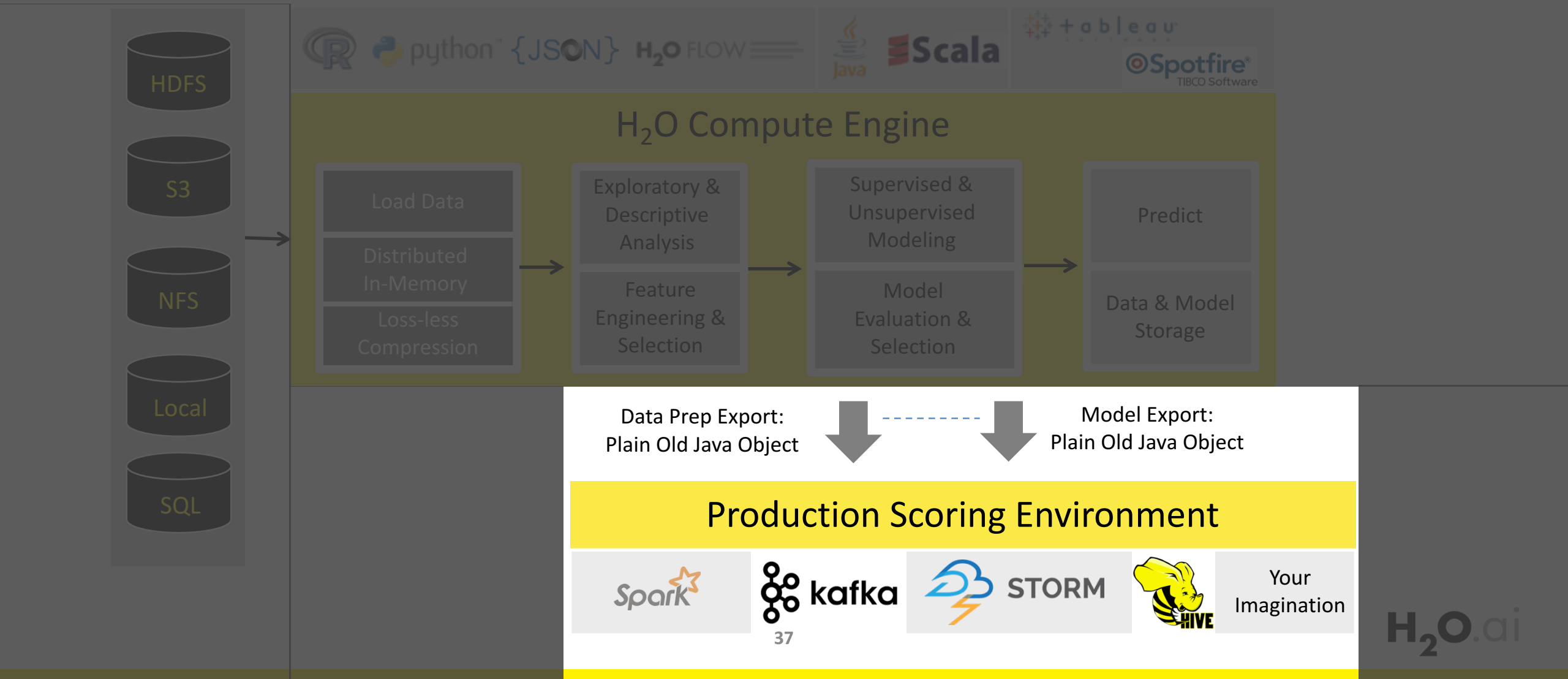
In [4]:
```
# Look at datasets
df_train.summary()
df_test.summary()
```

	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	-0.21060
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	1.17217
zeros	0		9	37	0	10563	11284	11309	11905
missing	0	0	0	0	0	0	0	0	0

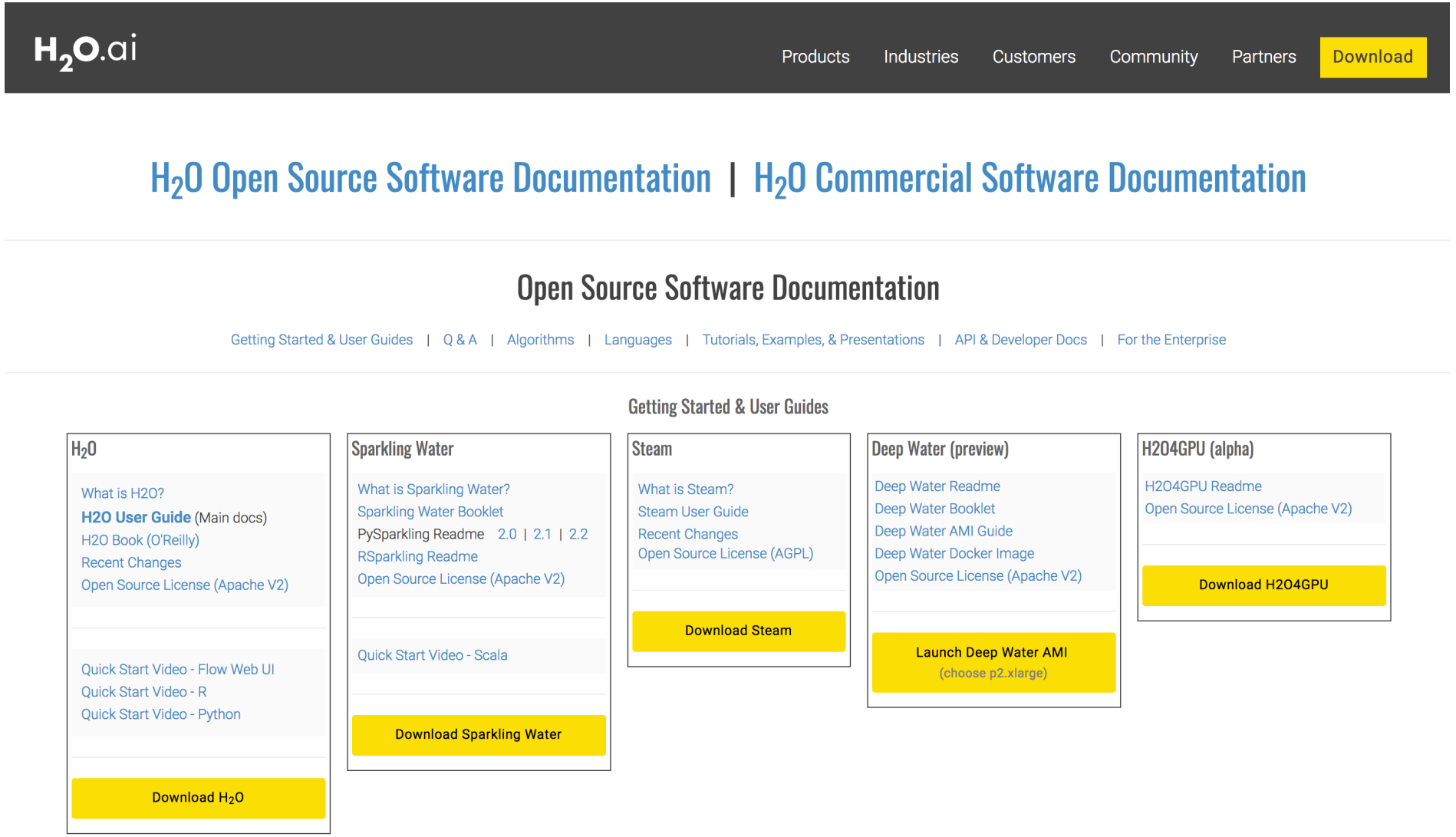
H_2O.ai

High Level Architecture

导出生产的独立模型



docs.h2o.ai

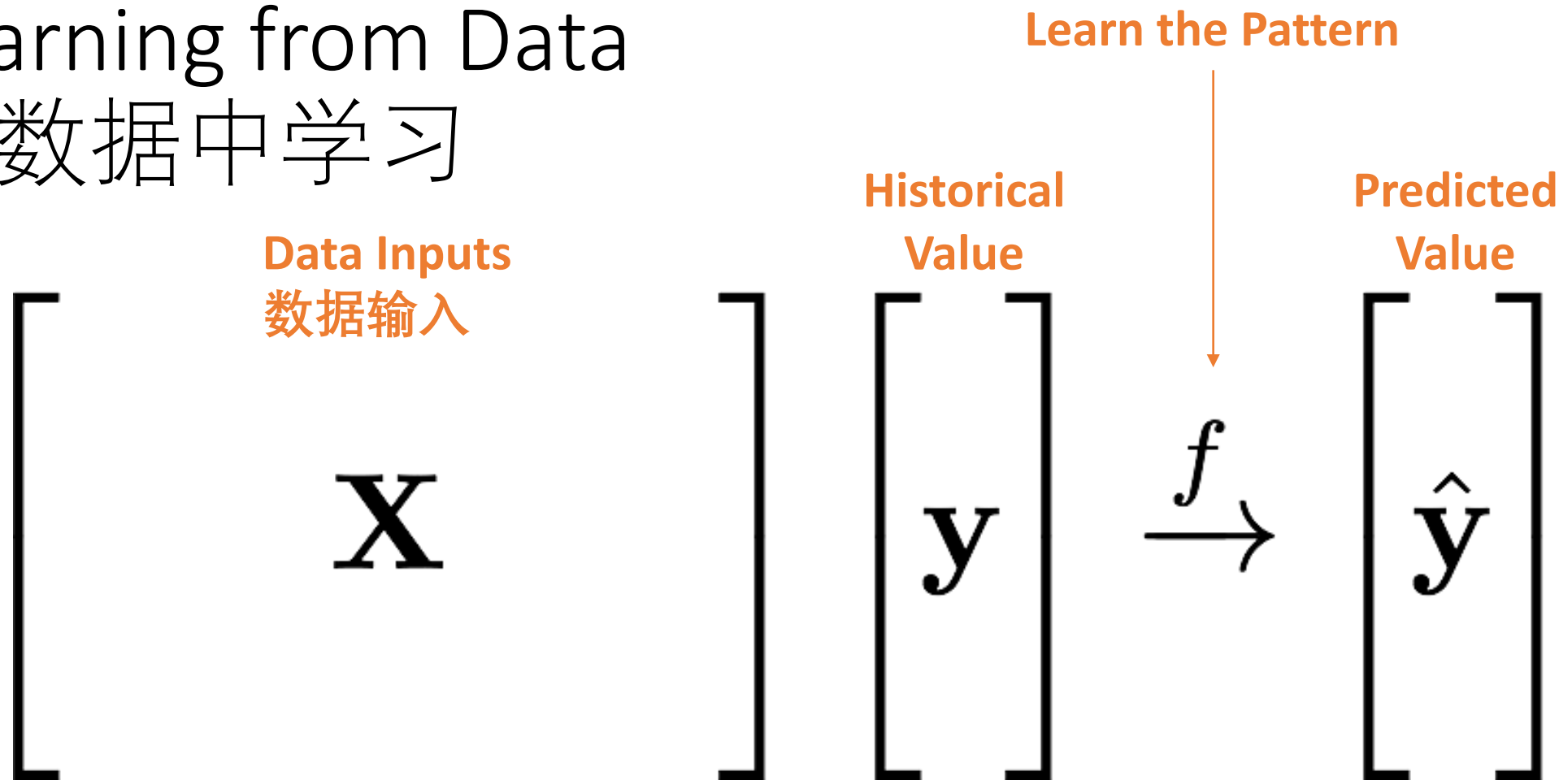


H2O.ai

Credit Card Demo

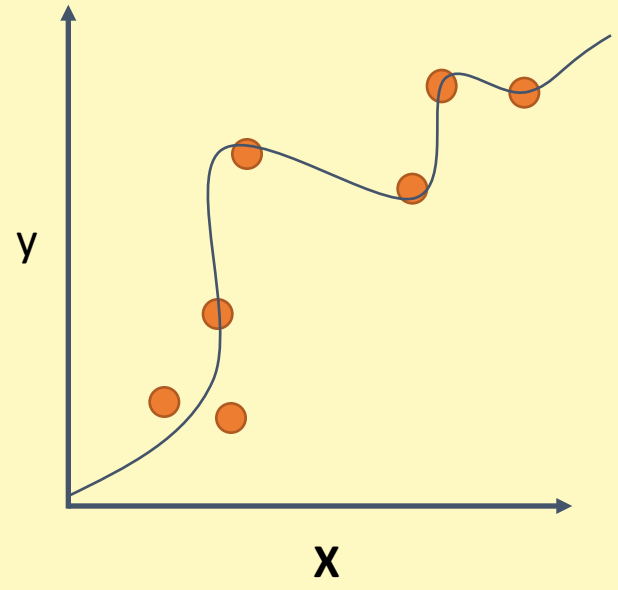
信用卡风险演示

Learning from Data
从数据中学习



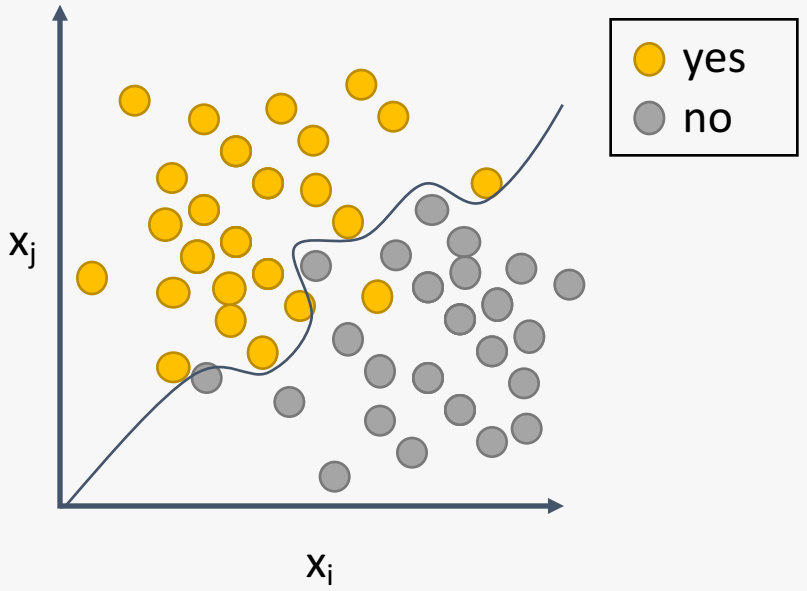
Supervised Learning 监督学习

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

Demo Introduction

/ **Use Case**: Probability of Default for Credit Card Loans

/ **Features**

- **default.payment.next.month**: Did the next loan payment default (1=True, 0=False)
- **LIMIT_BAL**: Credit limit in (NT) dollars
- **SEX**, **EDUCATION**, **MARRIAGE**, **AGE**
- **PAY_0**: Was a payment received in the current month?
- **PAY_2**: Was a payment received in the 2 months ago?

…

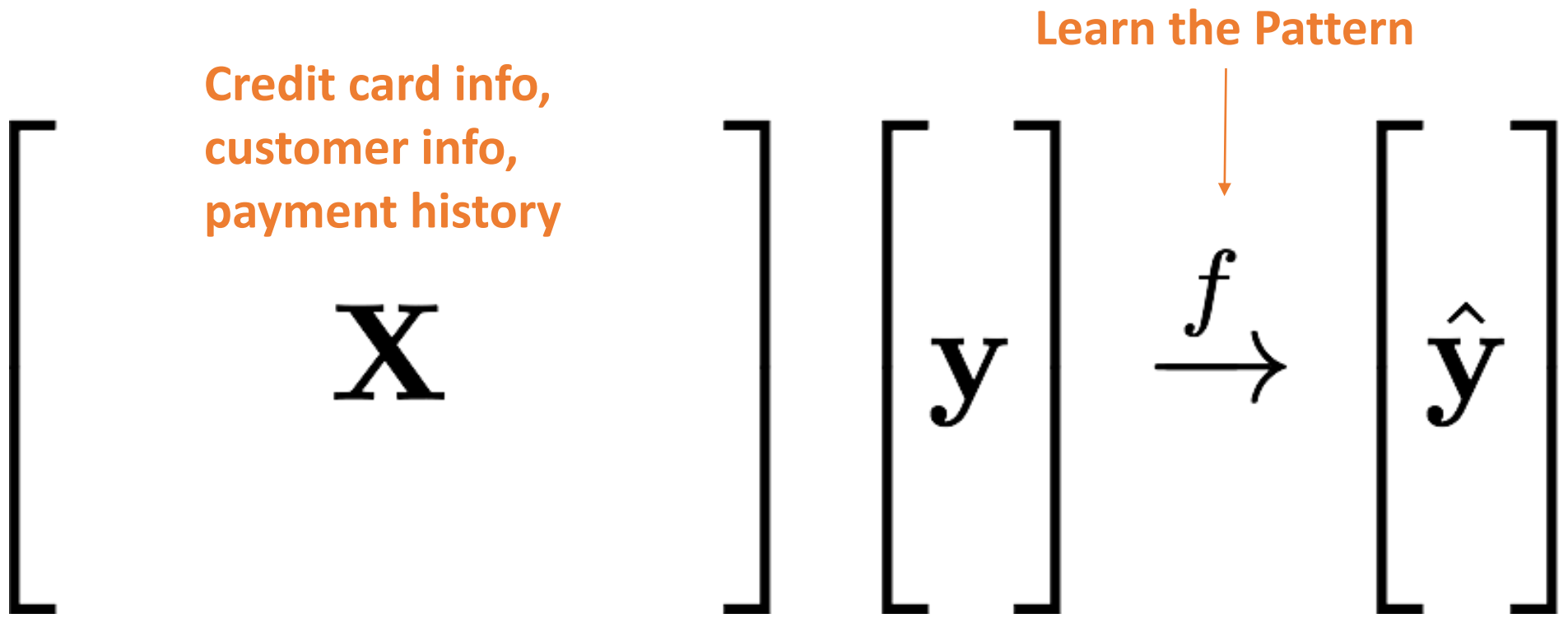
- **BILL_AMT1**: Amount of bill statement in 1 month ago
- **BILL_AMT2**: Amount of bill statement in 2 months ago

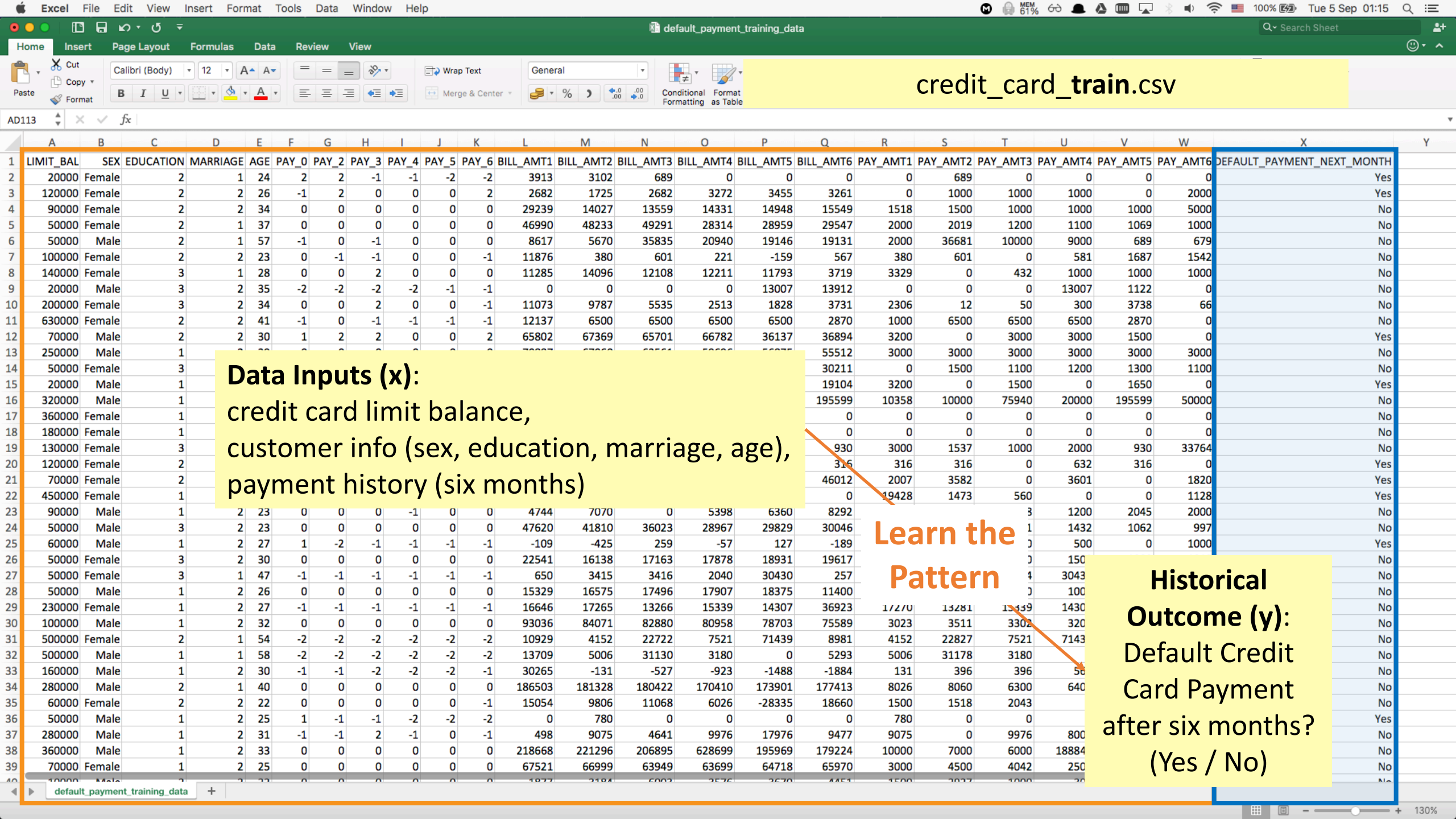
…

- **PAY_AMT1**: Amount of previous payment 1 month ago
- **PAY_AMT2**: Amount of previous payment 2 months ago

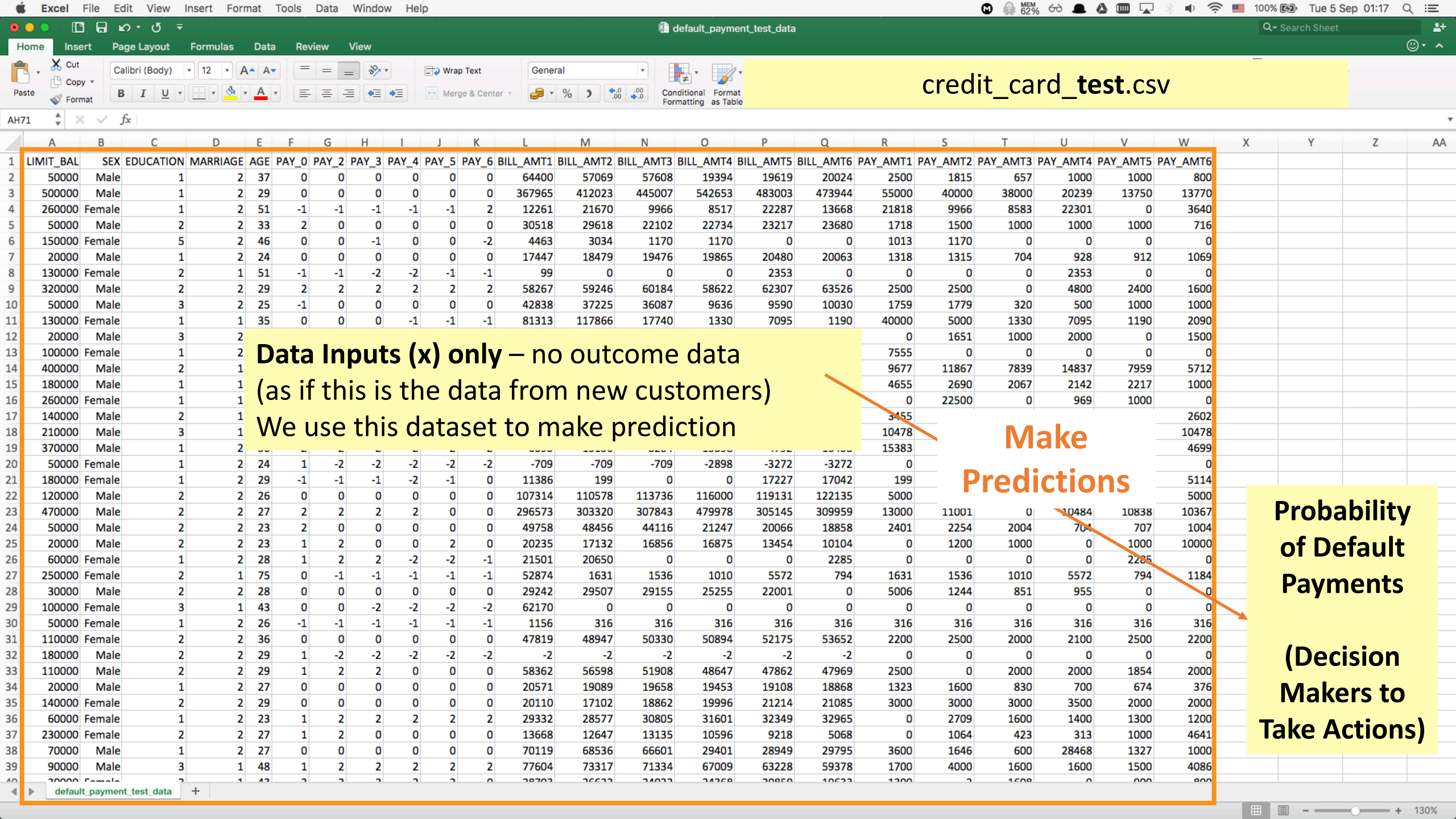
…

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	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



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

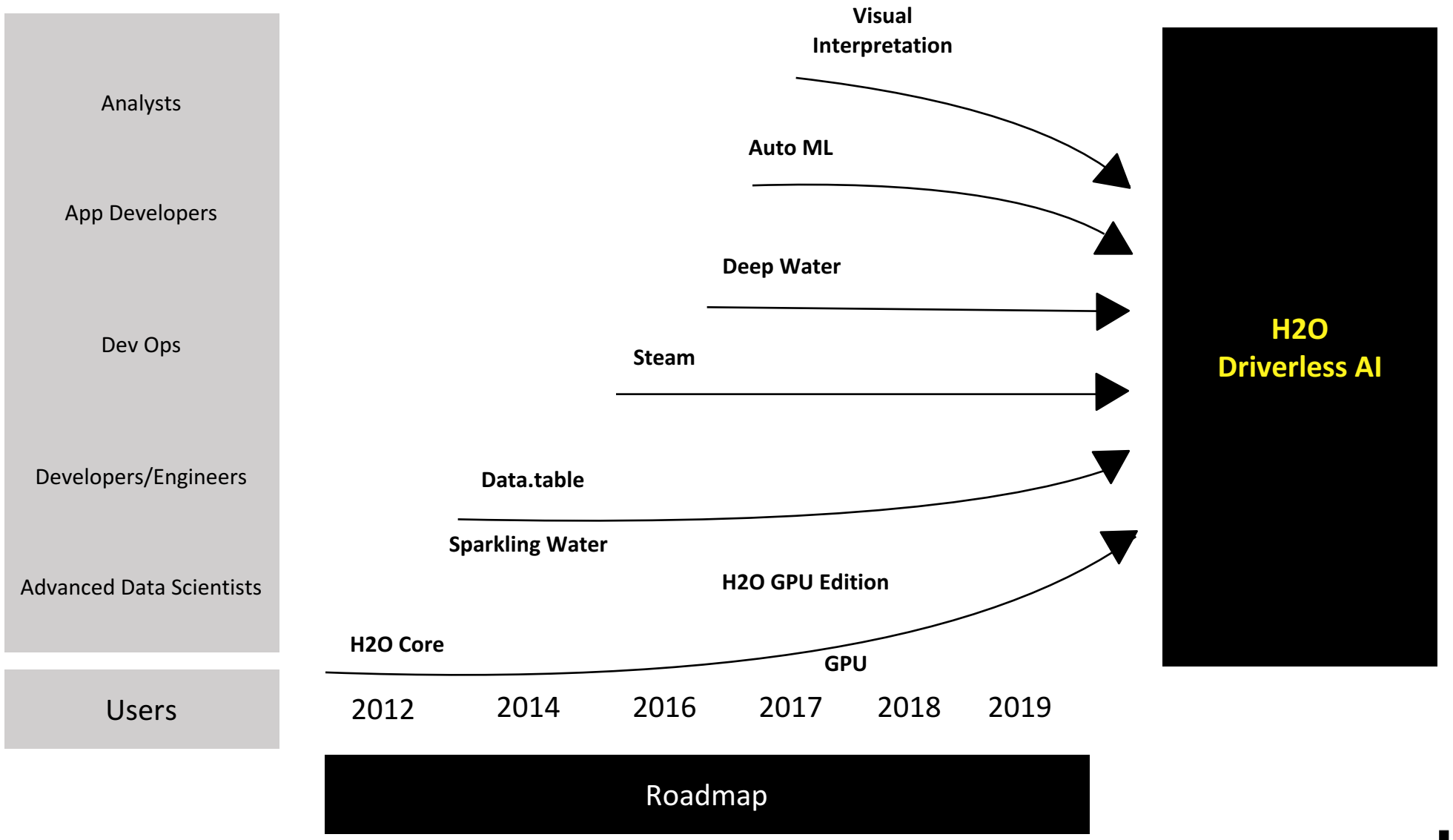
Live Demo

- 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

Driverless AI

Using AI to do AI 用人工智能制造人工智能

H_2O AI Platform Timeline



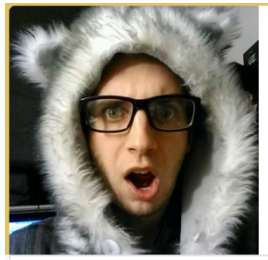
Kaggle Masters at H_2O.ai

Dmitry Larko

Sr. Data Scientist at H2O.ai
San Francisco Bay Area, CA, United States
Joined 5 years ago · last seen in the past day
https://h2o.ai

Competitions Grandmaster

Home | Competitions (34) | Discussion (32) | Organizations (2) | Followers (4) | Contact User | Follow User



Branden Murray

Data Scientist at h2o.ai
California, United States
Joined 3 years ago · last seen in the past day

Competitions Grandmaster

Home | Competitions (24) | Kernels (5) | Discussion (250) | Organizations (1) | Followers (11) | Contact User | Follow User

mlandry

Mountain View, CA, United States
Joined 5 years ago · last seen a day ago

Competitions Master

Home | Competitions (66) | Kernels (12) | Discussion (123) | Organizations (1) | Followers (17) | Contact User | Follow User

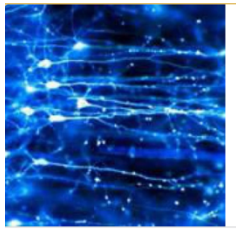
Μαριος Μιχαηλιδης KazAnova

Data Scientist at H2O ai
Volos, Greece
Joined 4 years ago · last seen in the past day
http://www.kazanovaforanalytics.com/

Competitions Grandmaster

Home | Competitions (93) | Kernels (5) | Discussion (468) | Organizations (1) | Followers (72) | Contact User | Follow User

Competitions Grandmaster			Kernels Contributor			Discussion Master		
Current Rank	Highest Rank					Rank		
2 of 58,377	1		Unranked			2 of 30,803		
23	23	21	0	0	0	30	36	233



Faron

Data Scientist at H2O.ai
Deutschland
Joined 3 years ago · last seen a day ago
http://kagglenizer.com/

Followers 402

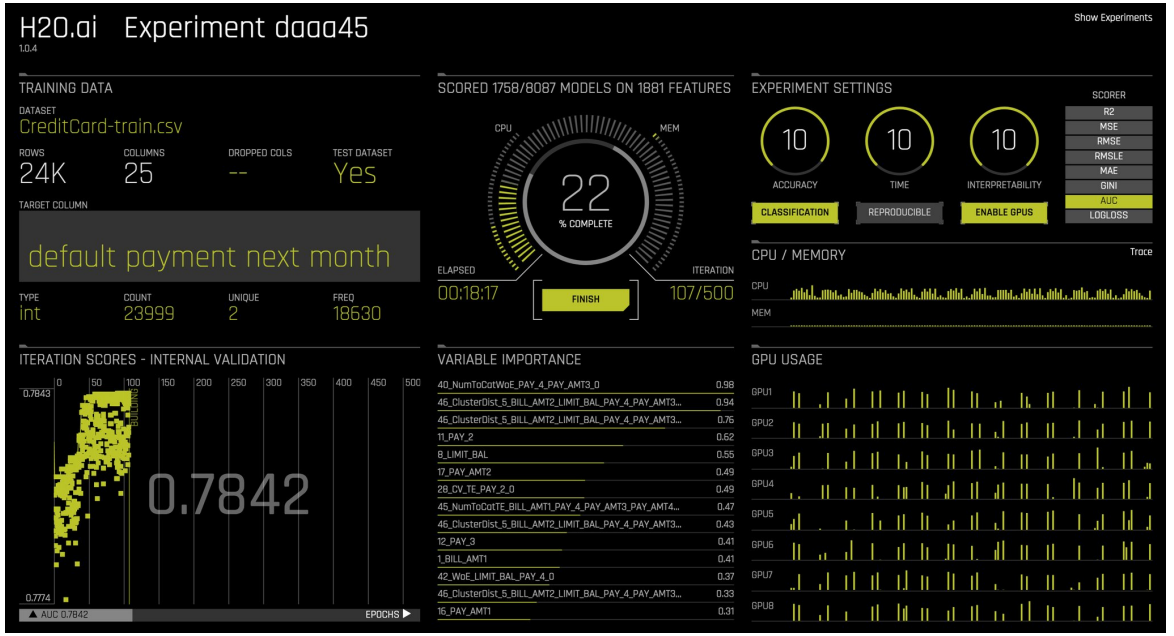
Competitions Grandmaster

Home | Competitions (26) | Kernels (9) | Discussion (377) | Organizations (1) | Followers (402) | Contact User | Follow User

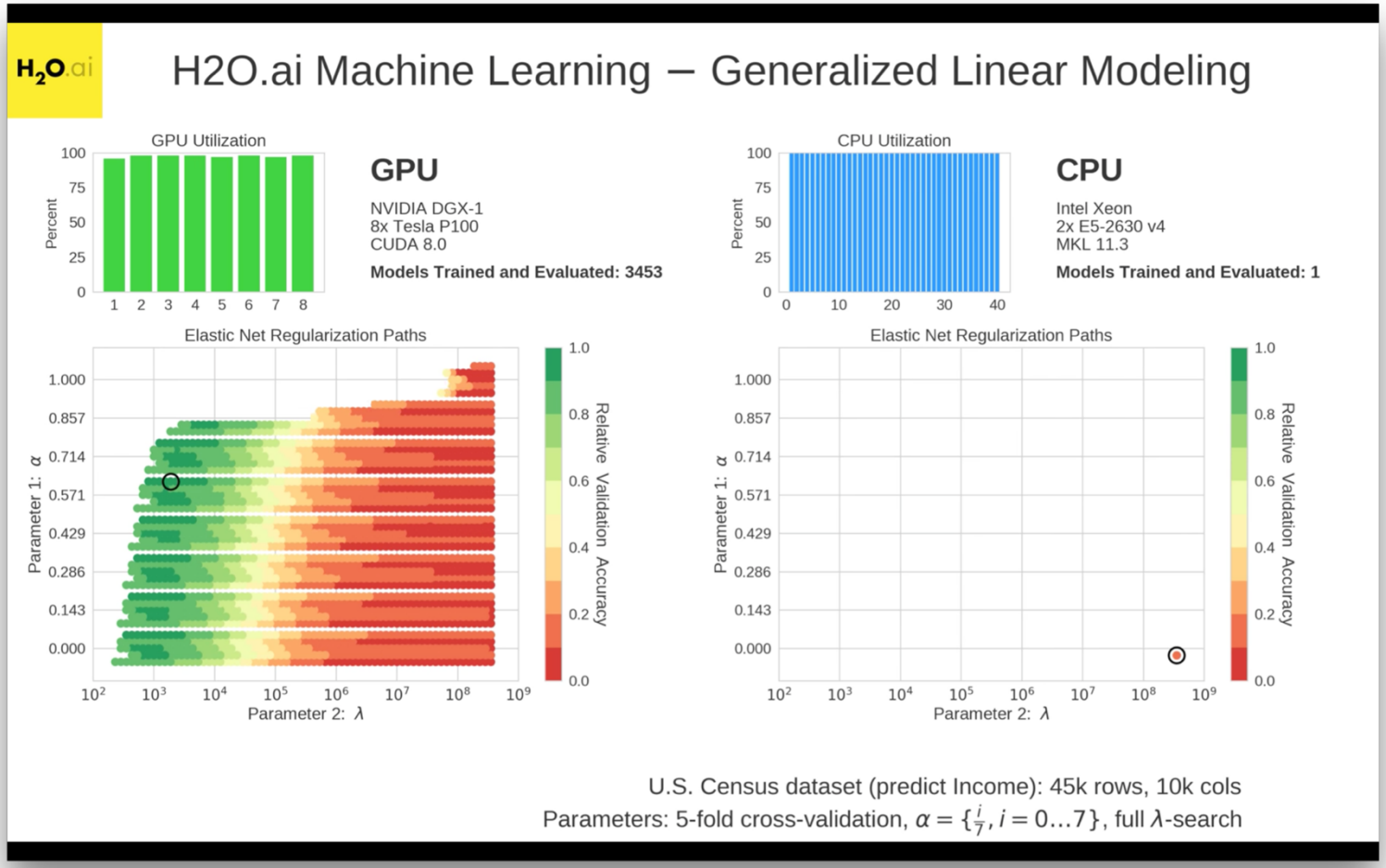
Competitions Grandmaster		Kernels Expert		Discussion Master	
Current Rank	Highest Rank	Current Rank	Highest Rank	Current Rank	Highest Rank
5 of 66,219	4	24 of 108,021	8	12 of 40,552	6

The Product: Driverless AI

- Kaggle Grandmasters in a Box
  - A solution to the shortage of data scientists
- Fast
  - H_2O Algorithms optimised for GPUs
- Accurate
  - Auto Feature Engineering
  - Auto Model Tuning / Selection / Ensemble
- Interpretable
  - Machine Learning Interpretation
- Production Ready
  - Export pipeline as a standalone Python package



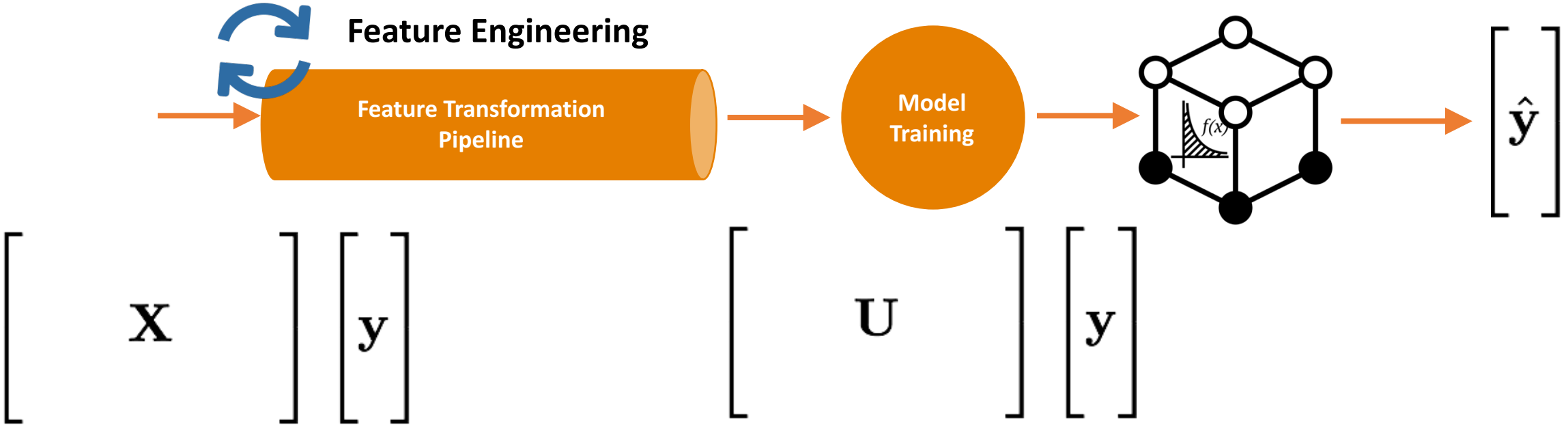
Fast: H_2O Algorithms on GPUs



3400+ models were trained on GPUs by the time the first model completed training on CPUs

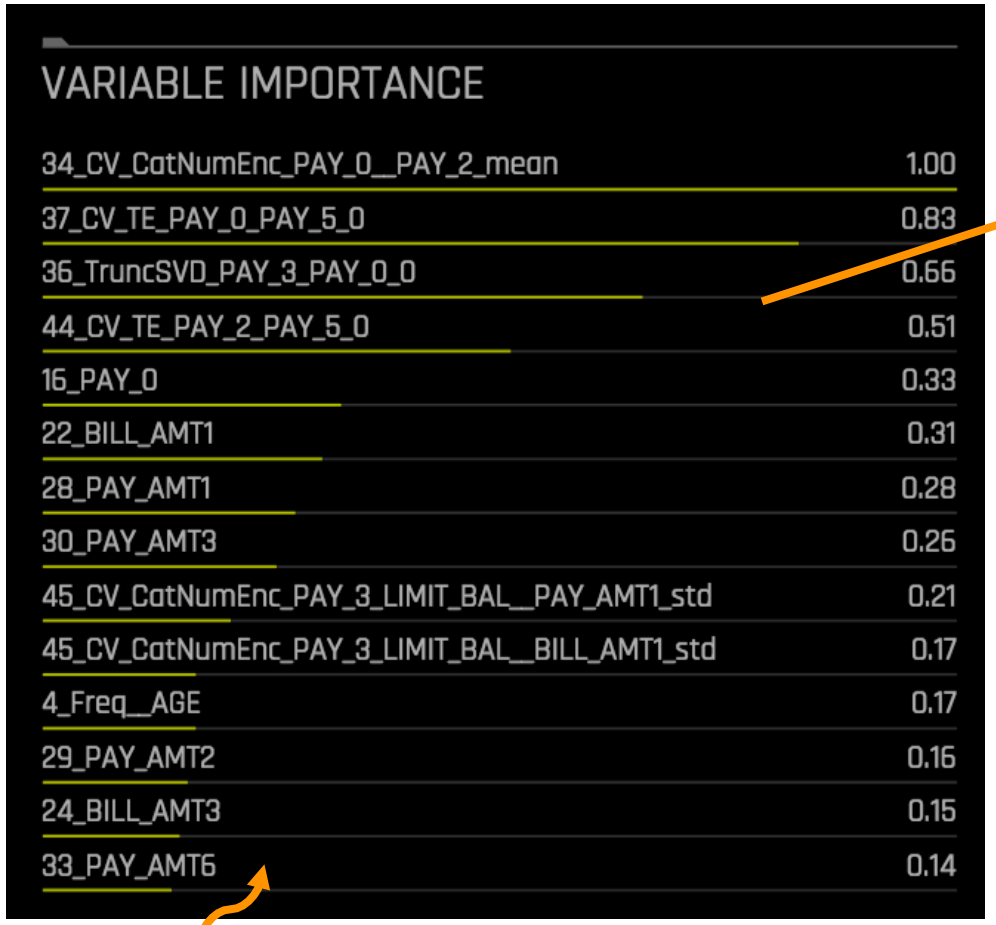
https://www.youtube.com/watch?v=LrC3mBNG7WU

Accurate: Auto Feature Engineering



H2O.ai

Accurate: Auto Feature Engineering



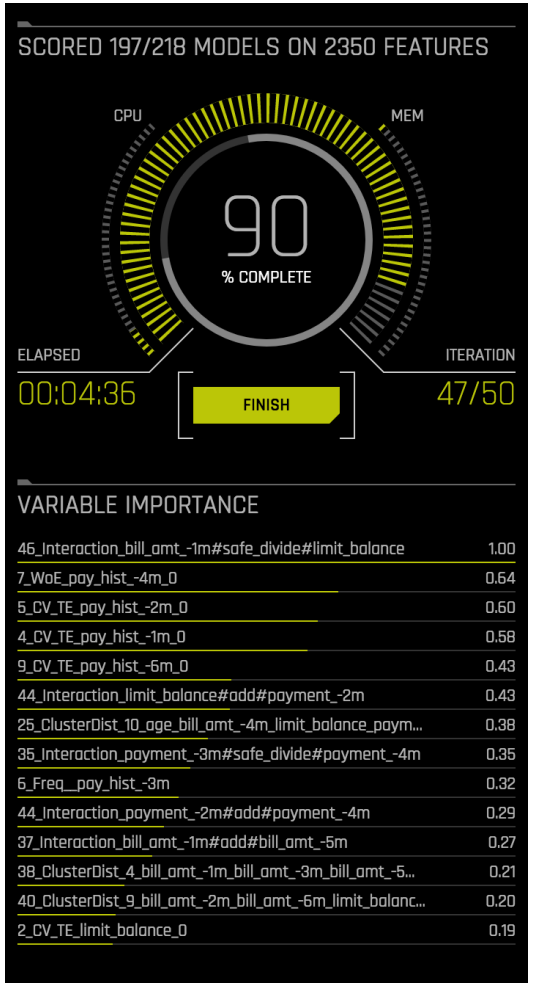
Generated Features

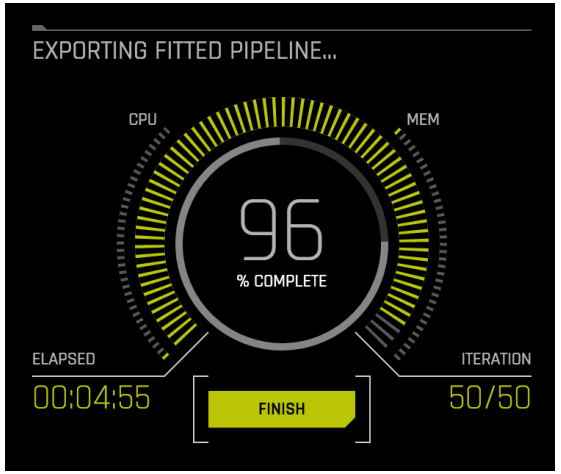
Original Features

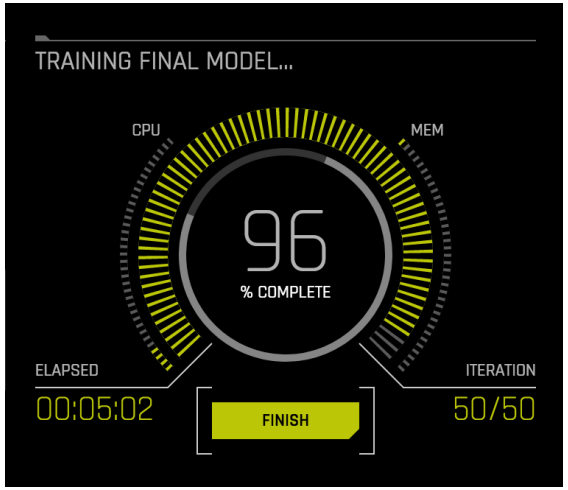
Feature Transformations

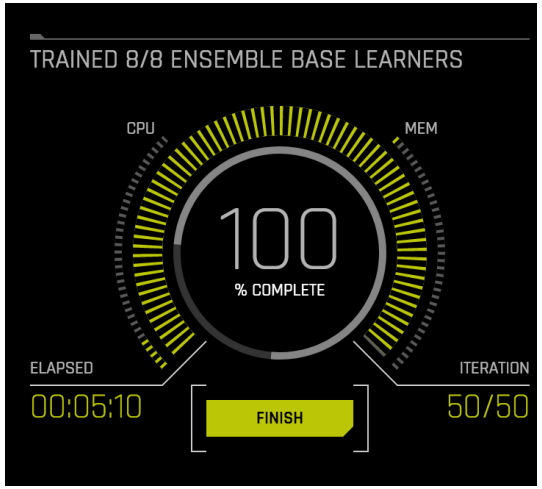
- Cross Validation Categorical Encoding
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD and More

Accurate: Auto Model Tuning/Selection/Ensemble



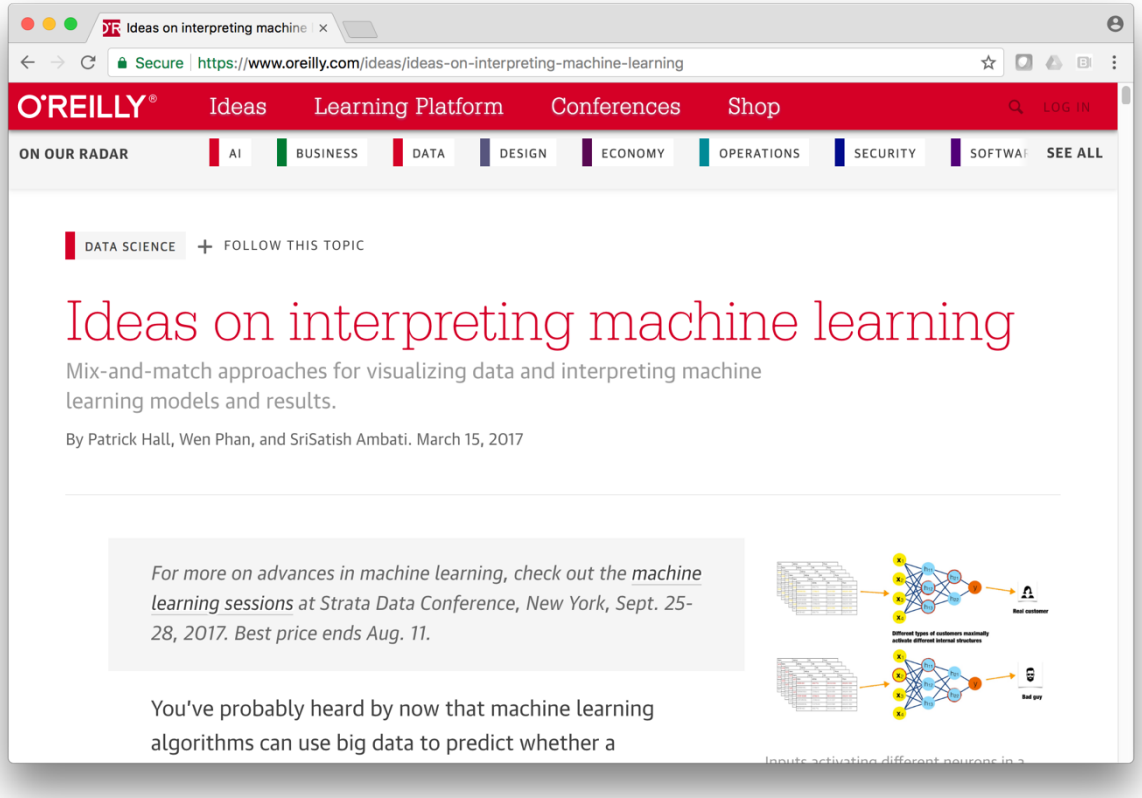




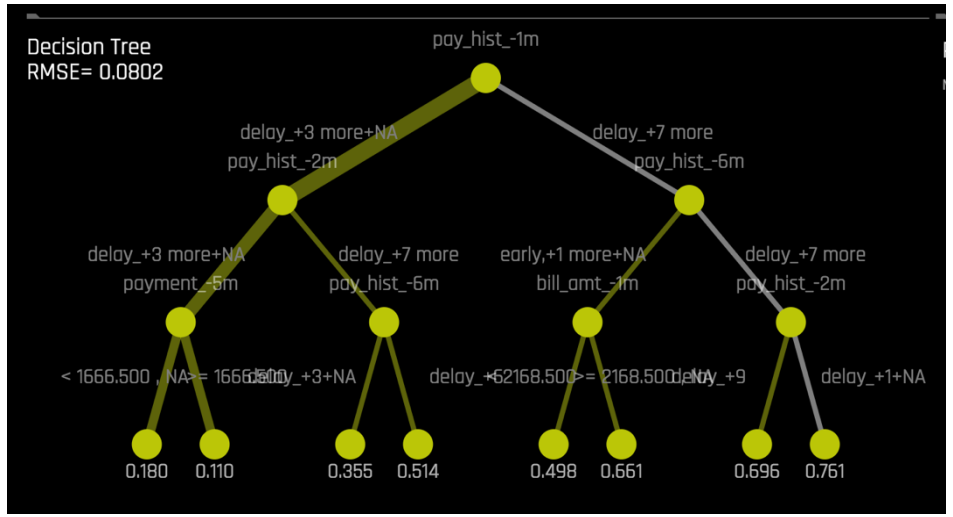


H2O.ai

Interpretable: Explainable Machine Learning



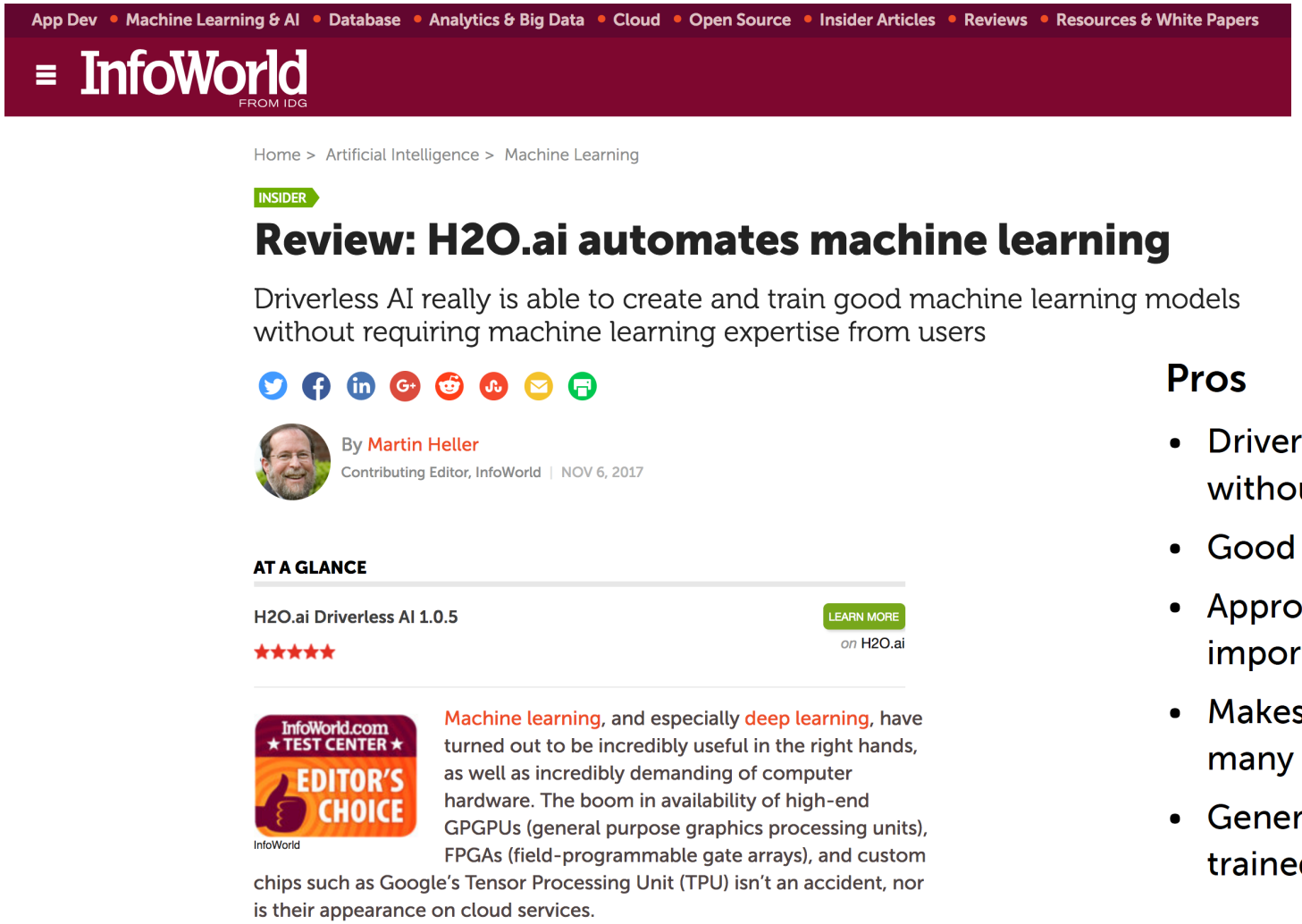
https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning



Local Reason Codes

k-LIME Local Attributions	Variable	with value	is associated with **default**	
Top Positive Local Attributions				
	pay_hist_-1m	delay_2m	increase of	0.35
	pay_hist_-4m	delay_7m	increase of	0.1
	pay_hist_-2m	delay_2m	increase of	0.06
	pay_hist_-5m	delay_7m	increase of	0.05
	age	26.000	increase of	0.02
	pay_hist_-3m	delay_7m	increase of	0.01
	sex	male	increase of	0.01
	bill_amt_-6m	2400.000	increase of	0
	bill_amt_-3m	2400.000	increase of	0

External Review



App Dev • Machine Learning & AI • Database • Analytics & Big Data • Cloud • Open Source • Insider Articles • Reviews • Resources & White Papers

InfoWorld FROM IDG

Home > Artificial Intelligence > Machine Learning

INSIDER

Review: H2O.ai automates machine learning

Driverless AI really is able to create and train good machine learning models without requiring machine learning expertise from users

By Martin Heller
Contributing Editor, InfoWorld | NOV 6, 2017

AT A GLANCE

H2O.ai Driverless AI 1.0.5

★★★★★

LEARN MORE on H2O.ai

InfoWorld.com ★TEST CENTER★ EDITOR'S CHOICE
InfoWorld

Machine learning, and especially deep learning, have turned out to be incredibly useful in the right hands, as well as incredibly demanding of computer hardware. The boom in availability of high-end GPGPUs (general purpose graphics processing units), FPGAs (field-programmable gate arrays), and custom chips such as Google's Tensor Processing Unit (TPU) isn't an accident, nor is their appearance on cloud services.

Pros

- Driverless AI is able to create and train good models without requiring user expertise
- Good integration with Nvidia GPUs (K80 and above)
- Approximate linear models help to explain important factors in a decision
- Makes quick work of generating and evaluating many models
- Generates and exports prediction pipeline for trained model

https://www.infoworld.com/article/3236048/machine-learning/review-h2oai-automates-machine-learning.html

www.h2o.ai/driverless-ai-download/

Docker Image

H2O Driverless AI

Thank you very much for your interest with our new and exciting Driverless AI product. This product leverages GPU Machines with a focus on Auto Feature Engineering, Model Interpretability, and Automatic Data Visualization.

DOWNLOAD DRIVERLESS AI | DRIVERLESS AI DOCUMENTATION

Don't have a registration key? Apply here to try Driverless AI.

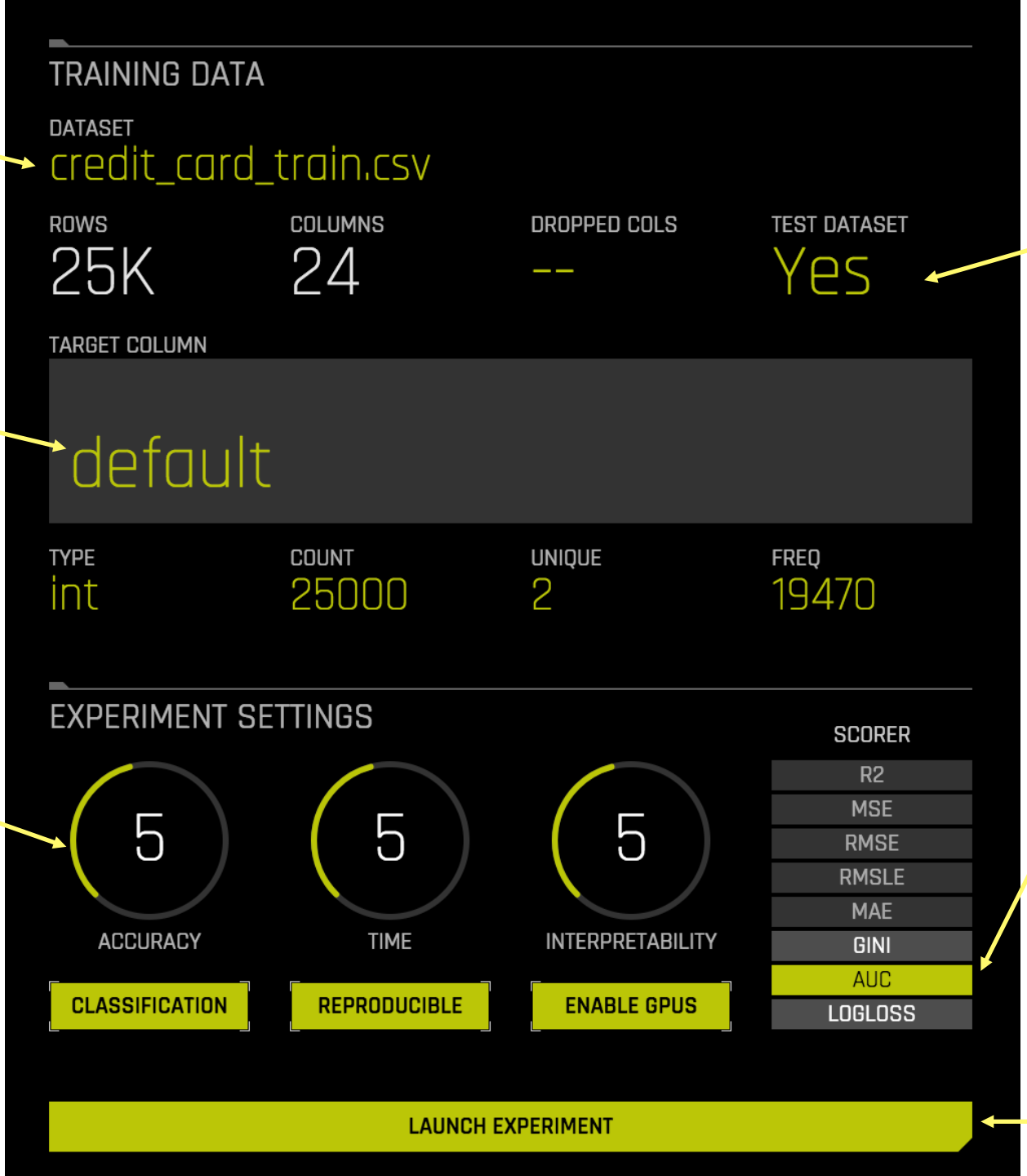
Here are some other resources to help you get started:

Using Driverless AI Booklet
Machine Learning Interpretability with H2O Driverless AI Booklet
Driverless AI Data Sheet
Webinars recently delivered on Driverless AI

30-Day Free Trial

H2O.ai

Demo: Credit Card Default Risk



Training Data

Target

Simple Settings

Optional Test Set

Select Metric

That's it!

H2O.ai Experiment b1569b

1.0.5

Show Experiments

TRAINING DATA

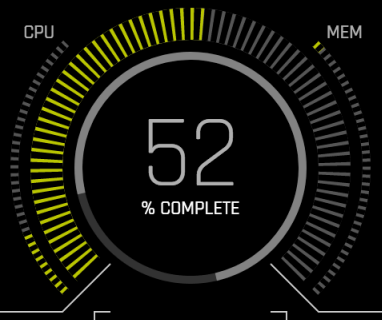
DATASET
credit_card_train.csv

ROWS	COLUMNS	DROPPED COLS	TEST DATASET
25K	24	--	Yes

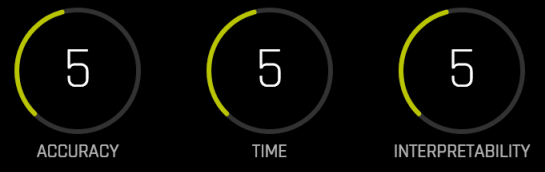
TARGET COLUMN
default

TYPE	COUNT	UNIQUE	FREQ
int	25000	2	19470

SCORED 113/218 MODELS ON 1720 FEATURES



EXPERIMENT SETTINGS

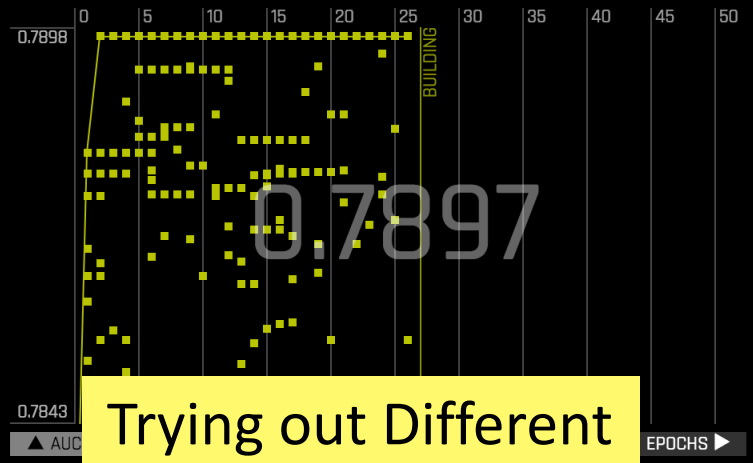


SCORER
R2
MSE
RMSE
RMSLE
MAE
GINI
AUC
LOGLOSS

CPU / MEMORY

Trace

ITERATION SCORES - INTERNAL VALIDATION



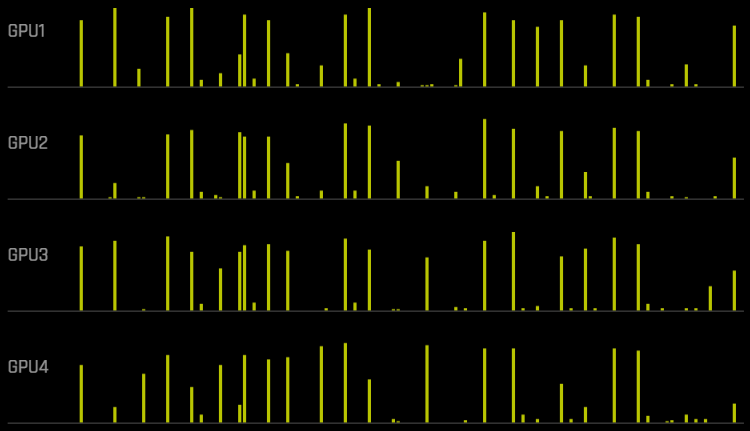
Trying out Different Sets of Features

VARIABLE IMPORTANCE

46_Interaction_bill_amt_-1m#safe_divide#limit_balance	1.00
7_WoE_pay_hist_-4m_0	0.78
5_CV_TE_pay_hist_-2m_0	0.72
4_CV_TE_pay_hist_-1m_0	0.68
25_ClusterDist_10_age_bill_amt_-4m_limit_balance_paym...	0.50
44_Interaction_payment_-2m#add#payment_-4m	0.42
44_Interaction_limit_balance#add#payment_-2m	0.40
8_WoE_pay_hist_-5m_0	0.40
6_Freq_pay_hist_-3m	0.37
9_CV_TE_pay_hist_-6m_0	0.36
38_ClusterDist_4_bill_amt_-1m_bill_amt_-3m_bill_amt_-5...	0.34
40_ClusterDist_9_bill_amt_-2m_bill_amt_-6m_limit_balanc	0.32
35_In	0.29
19_pa	0.25

Auto Feature Engineering

GPU USAGE



Using both CPUs & GPUs

Options for Interpretation, Scoring, Logs, Pipeline

H2O.ai Experiment b1569b

1.0.5

Show Experiments

TRAINING DATA

DATASET
credit_card_train.csv

ROWS	COLUMNS	DROPPED COLS	TEST DATASET
25K	24	--	Yes

TARGET COLUMN

default

TYPE	COUNT	UNIQUE	FREQ
int	25000	2	19470

STATUS: COMPLETE

- INTERPRET THIS MODEL
- SCORE ON ANOTHER DATASET
- DOWNLOAD (HOLDOUT) TRAINING PREDICTIONS
- DOWNLOAD TEST PREDICTIONS
- DOWNLOAD TRANSFORMED TRAINING DATA
- DOWNLOAD TRANSFORMED TEST DATA
- DOWNLOAD LOGS
- DOWNLOAD SCORING PACKAGE

EXPERIMENT SETTINGS

ACCURACY	TIME	INTERPRETABILITY
5	5	5

CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

SCORER: R2, MSE, RMSE, RMSLE, MAE, GINI, AUC, LOGLOSS

CPU / MEMORY

Trace

CPU

MEM

ITERATION SCORES - INTERNAL VALIDATION

VARIABLE IMPORTANCE

46_Interaction_bill_amt_-1m#safe_divide#limit_balance	1.00
7_WoE_pay_hist_-4m_0	0.64
5_CV_TE_pay_hist_-2m_0	0.60
4_CV_TE_pay_hist_-1m_0	0.58
9_CV_TE_pay_hist_-6m_0	0.43
44_Interaction_limit_balance#add#payment_-2m	0.43
25_ClusterDist_10_age_bill_amt_-4m_limit_balance_paym...	0.38
35_Interaction_payment_-3m#safe_divide#payment_-4m	0.35
6_Freq__pay_hist_-3m	0.32
44_Interaction_payment_-2m#add#payment_-4m	0.29
37_Interaction_bill_amt_-1m#add#bill_amt_-5m	0.27
38_ClusterDist_4_bill_amt_-1m_bill_amt_-3m_bill_amt_-5...	0.21
40_ClusterDist_9_bill_amt_-2m_bill_amt_-6m_limit_balanc...	0.20
2_CV_TE_limit_balance_0	0.19

GPU USAGE

GPU1

GPU2

GPU3

GPU4

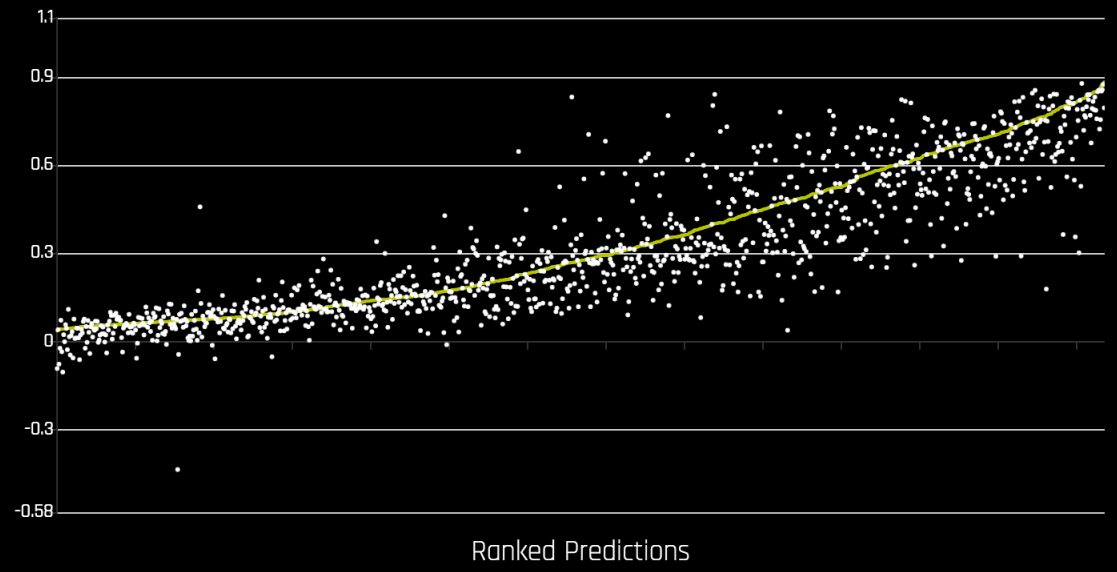
Single Row Lookup

Column: H2O Frame Row # Value: SEARCH Plot: Global

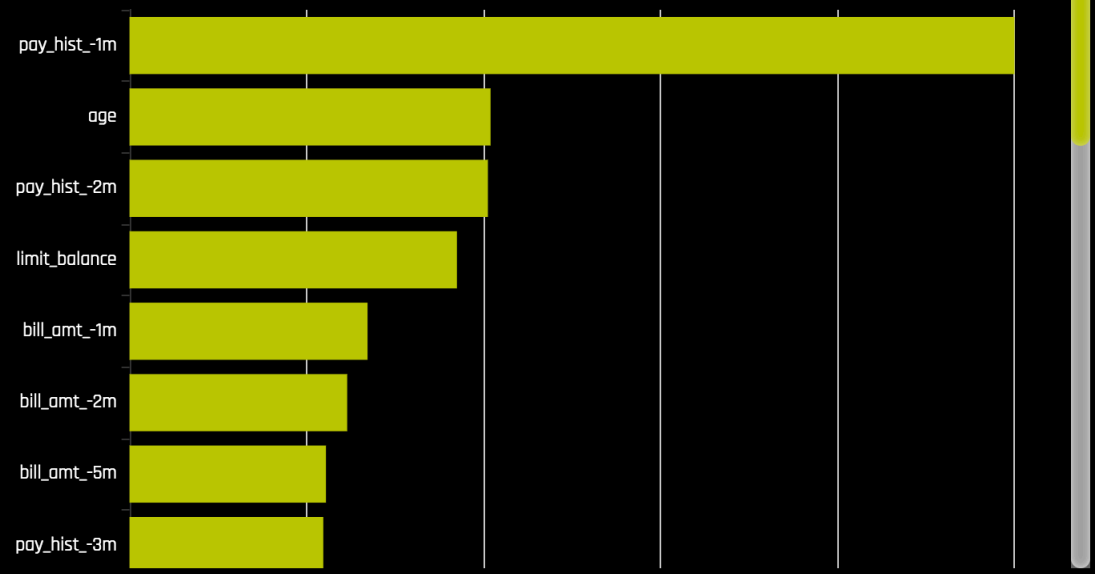
EXPERIMENT EXPLANATIONS

Global Interpretable Model Explanation Plot

Model Prediction k-LIME Model Prediction Actual Target

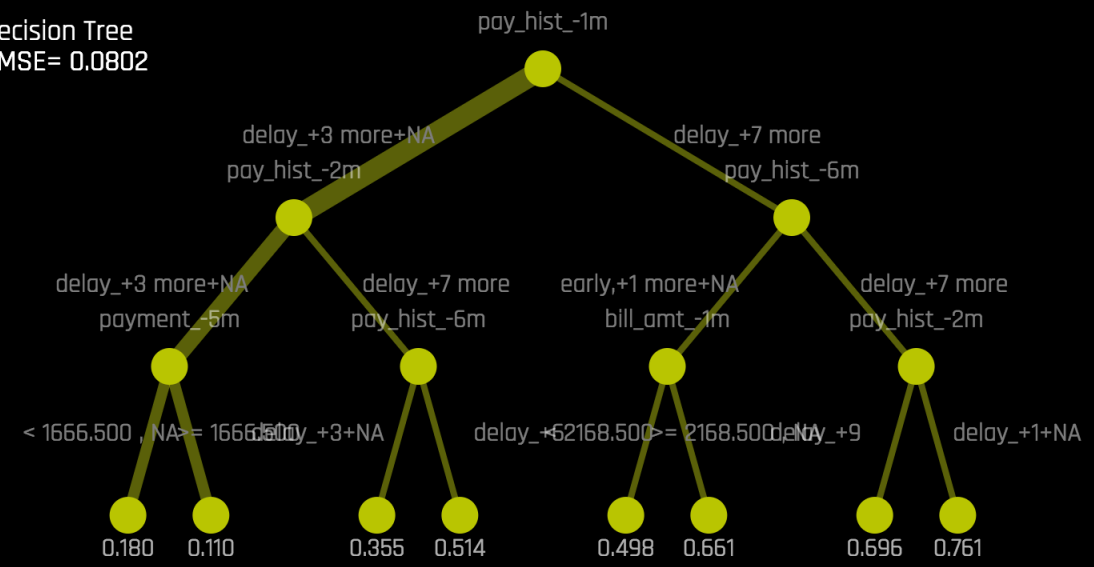


Variable Importance



Decision Tree

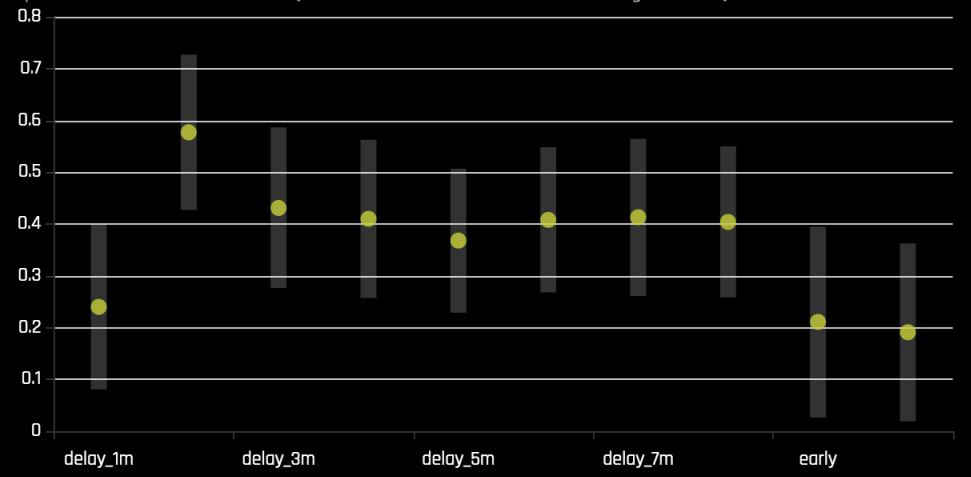
RMSE= 0.0802



Partial Dependence

pay_hist_-1m

Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)



Single Row Lookup

Column: H2O Frame Row # Value: 4850 SEARCH Plot: Cluster0

EXPERIMENT EXPLANATIONS

Cluster 0 k-LIME Plot

Model Prediction k-LIME Model Prediction Actual Target

1.1 0.9 0.3 0 -0.21

4850

k-LIME Prediction: 0.839

Model Prediction: 0.84

Actual: 1.0

Reason Codes:

pay_hist_-1m: 0.348

pay_hist_-4m: 0.096

pay_hist_-2m: 0.061

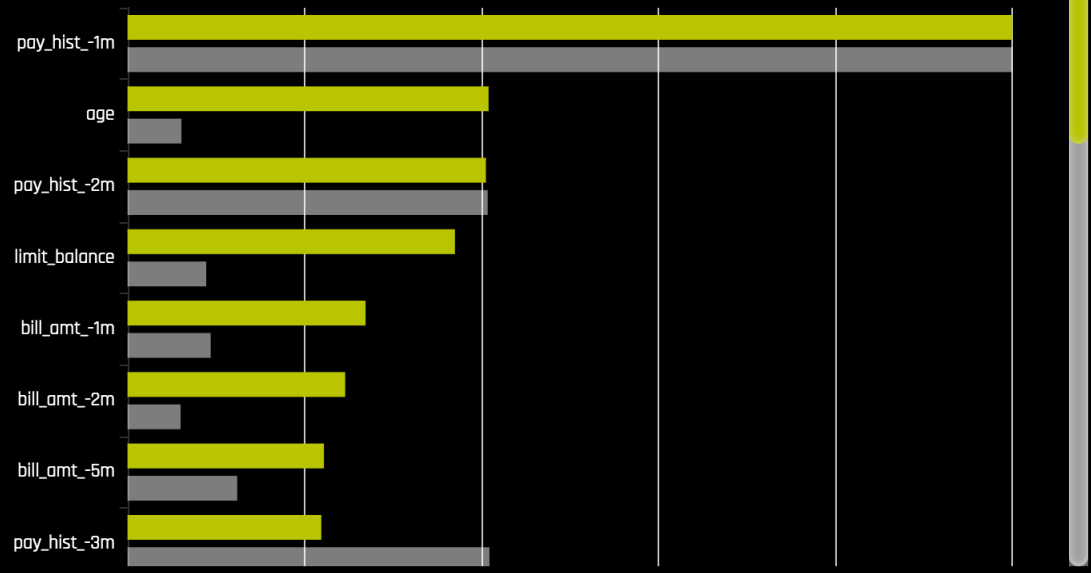
pay_hist_-5m: 0.049

age: 0.023

Ranked Predictions

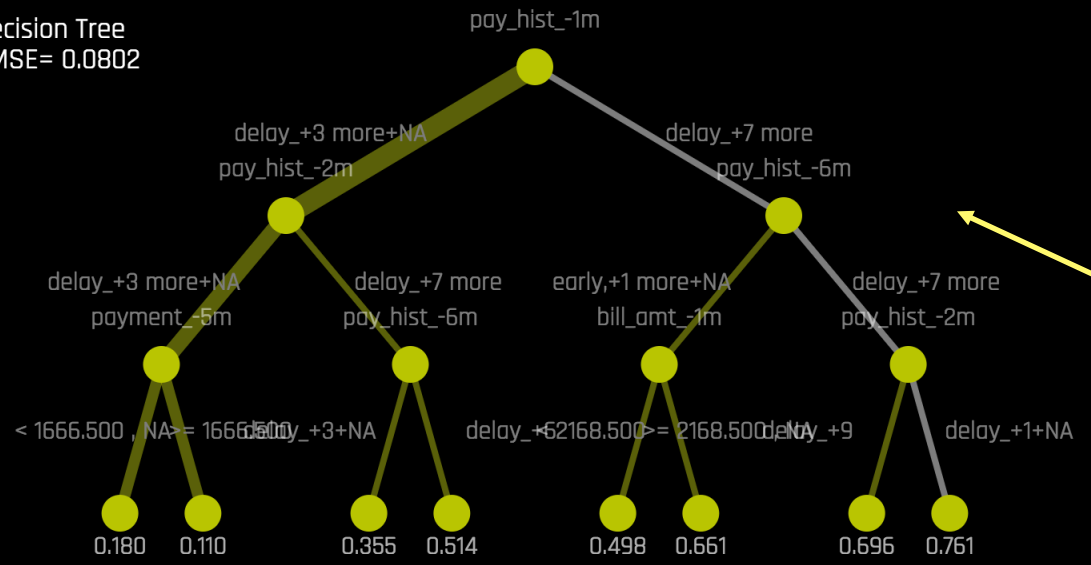
Zoom into individual records

Variable Importance



Decision Tree

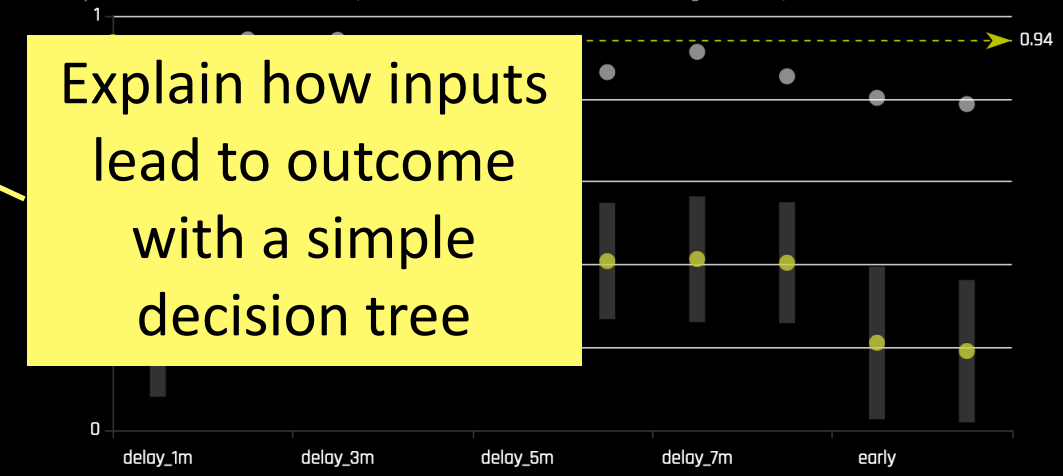
RMSE= 0.0802



Partial Dependence

pay_hist_-1m

Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)



Explain how inputs lead to outcome with a simple decision tree

Actual and Predicted Values

default (Actual)	1
Model Prediction Value	0.8403
k-LIME Prediction Value	0.8392
k-LIME Prediction Accuracy	99.9%

Local Reason Codes

k-LIME Local Attributions	Variable	with value	is associated with default	
Top Positive Local Attributions				
	pay_hist_-1m	delay_2m	increase of	0.35
	pay_hist_-4m	delay_7m	increase of	0.1
	pay_hist_-2m	delay_2m	increase of	0.06
	pay_hist_-5m	delay_7m	increase of	0.05
	age	26.000	increase of	0.02
	pay_hist_-3m	delay_7m	increase of	0.01
	sex	male	increase of	0.01
	bill_amt_-6m	2400.000	increase of	0
	bill_amt_-3m	2400.000	increase of	0

Skipped 6 additional attributions, click to view all ...

Turn decision tree into reason codes

Download Pipeline

```
ubuntu@ip-10-10-0-226:~/scorer$ ls -l
total 86684
-rw-r--r-- 1 ubuntu ubuntu       15 Nov  8 06:56 client_requirements.txt
-rw-r--r-- 1 ubuntu ubuntu  1510256 Nov  8 06:56 datatable-0.2.2+master.301.noomp-cp36-cp36m-linux_x86_64.whl
drwxrwxr-x 6 ubuntu ubuntu     4096 Nov  8 07:01 env
-rw-r--r-- 1 ubuntu ubuntu     6738 Nov  8 06:55 example_client.py
-rw-r--r-- 1 ubuntu ubuntu     9577 Nov  8 06:55 example.py
-rw-r--r-- 1 ubuntu ubuntu    16591 Nov  8 06:56 features.txt
-rw-r--r-- 1 ubuntu ubuntu 46735145 Nov  8 06:56 h2o4gpu-0.0.4-py36-none-any.whl
-rw-r--r-- 1 ubuntu ubuntu 14778303 Nov  8 06:56 h2oaicore-1.0.5-cp36-cp36m-linux_x86_64.whl
-rw-r--r-- 1 ubuntu ubuntu    10660 Nov  8 06:56 README.txt
-rw-r--r-- 1 ubuntu ubuntu      308 Nov  8 06:56 requirements.txt
-rwxr-xr-x 1 ubuntu ubuntu      305 Nov  8 06:55 run_example.sh
-rw-r--r-- 1 ubuntu ubuntu     6219 Nov  8 06:55 run_http_client.sh
-rwxr-xr-x 1 ubuntu ubuntu      520 Nov  8 06:55 run_http_server.sh
-rwxr-xr-x 1 ubuntu ubuntu      428 Nov  8 06:55 run_tcp_client.sh
-rwxr-xr-x 1 ubuntu ubuntu      519 Nov  8 06:55 run_tcp_server.sh
-rw-r--r-- 1 ubuntu ubuntu 25622313 Nov  8 06:57 scoring_1988a3_20171108045709_7ea33-1.0.0-py3-none-any.whl
-rw-r--r-- 1 ubuntu ubuntu     2722 Nov  8 06:55 scoring.thrift
-rw-r--r-- 1 ubuntu ubuntu     9613 Nov  8 06:55 server.py
-rw-r--r-- 1 ubuntu ubuntu       29 Nov  8 06:56 server_requirements.txt
drwxrwxr-x 2 ubuntu ubuntu     4096 Nov  8 07:02 tmp
ubuntu@ip-10-10-0-226:~/scorer$
```

```
ubuntu@ip-10-10-0-226:~/scorer$ cat requirements.txt
numpy==1.13.1
pandas==0.19.2
scikit-learn==0.19.0
scipy==0.19.0
requests==2.13.0
pycrypto==2.6.1
hashids==1.2.0
datatable-0.2.2+master.301.noomp-cp36-cp36m-linux_x86_64.whl
h2o4gpu-0.0.4-py36-none-any.whl
h2oaicore-1.0.5-cp36-cp36m-linux_x86_64.whl
scoring_1988a3_20171108045709_7ea33-1.0.0-py3-none-any.whl
ubuntu@ip-10-10-0-226:~/scorer$
```

Standalone Python Package

```
# Create a singleton Scorer instance.
# For optimal performance, create a Scorer instance once, and
#
scorer = Scorer()



#
# To score one row at a time, use the Scorer.score() method:
#

print('---------- Score Row ----------')
print(scorer.score([
    336625,  # limit_balance
    'female',  # sex
    'high_school',  # education
    'single',  # marriage
    41,  # age
    'delay_2m',  # pay_hist_-1m
    'delay_7m',  # pay_hist_-2m
    'delay_5m',  # pay_hist_-3m
    'delay_7m',  # pay_hist_-4m
    'on_time',  # pay_hist_-5m
    'delay_7m',  # pay_hist_-6m
    780288,  # bill_amt_-1m
    781180,  # bill_amt_-2m
    677013,  # bill_amt_-3m
    417082,  # bill_amt_-4m
    -16183,  # bill_amt_-5m
    -165806,  # bill_amt_-6m
    684022,  # payment_-1m
    450296,  # payment_-2m
    125650,  # payment_-3m
    109434,  # payment_-4m
```

Q & A