

Driverless AI: AI to do AI

Jo-fai (Joe) Chow

Data Scientist and Community Manager at H₂O.ai

joe@h2o.ai



Driverless AI: AI to do AI

- Company and Products Overview
- PayPal Use-Case
- Demo
- Other News

H₂O

Company Overview

Founded 2012, Series C in Nov, 2017

Products

- Driverless AI – Automated Machine Learning
- H2O Open Source Machine Learning
- Sparkling Water

Mission Democratize AI. Do Good

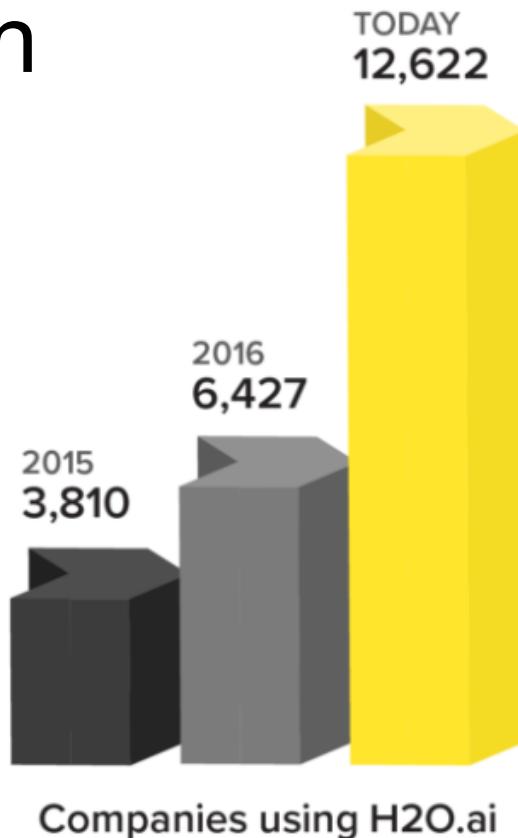
Team

- ~100 employees
- Distributed Systems Engineers doing Machine Learning
- World-class visualization designers

Offices Mountain View, London, Prague



Worldwide Community Adoption



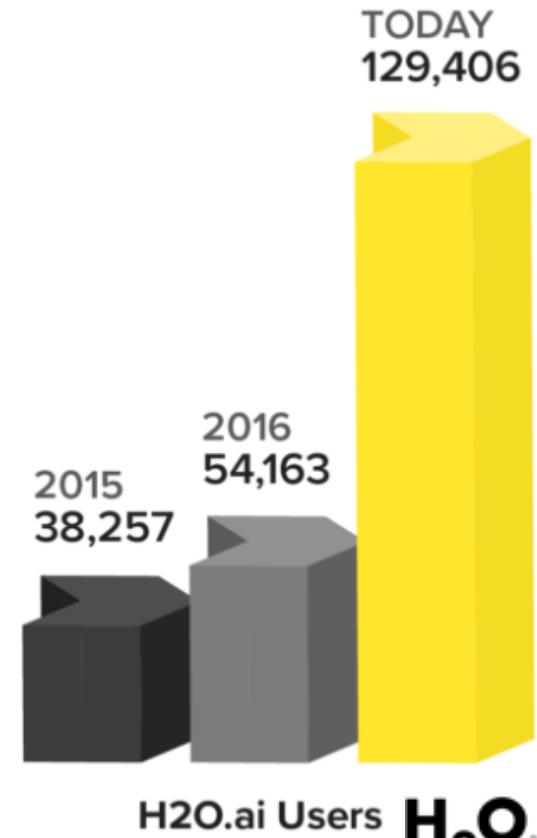
222 OF THE FORTUNE 500



8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES



* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H₂O.ai

Gartner names H2O as Leader with the most completeness of vision

- H2O.ai recognized as a **technology leader with most completeness of vision**
- H2O.ai was recognized for the mindshare, partner network and status as a **quasi-industry standard** for machine learning and AI.
- **H2O customers gave the highest overall score** among all the vendors for sales relationship and account management, customer support (onboarding, troubleshooting, etc.) and overall service and support.

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms



Source: Gartner (February 2018)

© Gartner, Inc

Platforms with H₂O integration



srisatish
@srisatish

Following

Replying to @BobMuenchen @knime @h2oai

@KNIME gained the ability to run @H2O.ai algorithms, so these two may be viewed as complementary, not competitors
#Ecosystem #OpenSource

3:32 PM - 2 Mar 2018

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms



Source: Gartner (February 2018)

© Gartner, Inc

H₂O.ai

H2O.ai Solution Leadership Across Verticals



H₂O Products



In-Memory, Distributed
Machine Learning Algorithms
with H2O Flow GUI



H2O AI Open Source Engine
Integration with Spark



Lightning Fast machine
learning on GPUs

DRIVERLESSAI

Automatic feature
engineering, machine
learning and interpretability

Steam

Secure multi-tenant H2O clusters



<https://www.h2o.ai/try-driverless-ai/>



PRODUCTS CUSTOMERS COMPANY SUPPORT

DOWNLOAD

Try Driverless AI

Driverless AI speeds up data science workflows by automating feature engineering, model tuning, ensembling and model deployment.



Request a Free 21-Day Trial

First Name: *

Last Name: *

Job Title: *

Company Name: *

Corporate Email Address: *

Phone Number: *

SEND ME A TRIAL LICENSE

Why Driverless AI?

H₂O

Shortage of Data Scientists

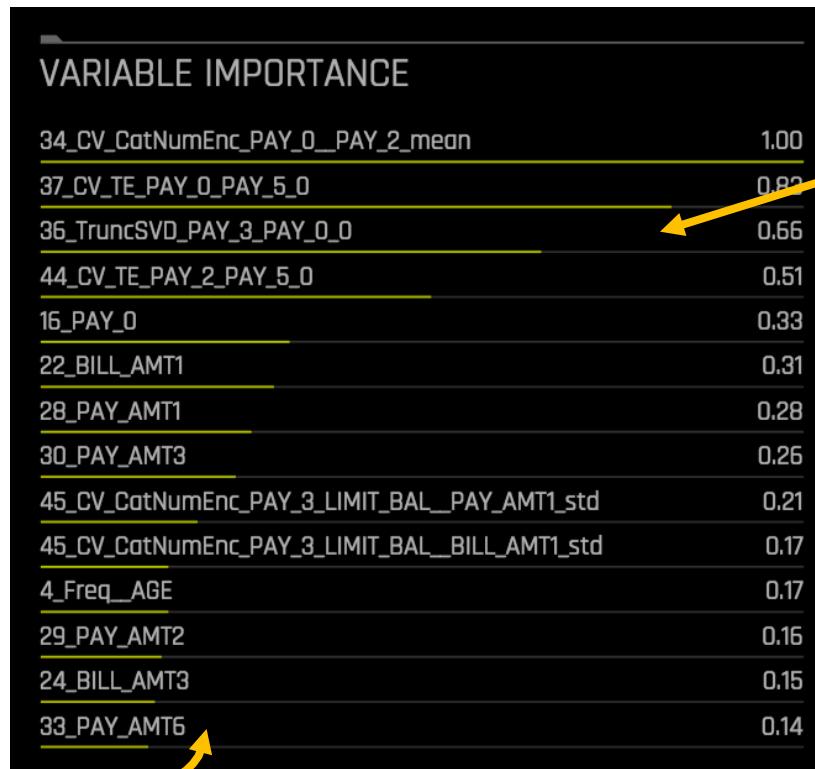
“The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts”

—McKinsey Prediction for 2018



Auto Feature Generation

Kaggle Grand Master Out of the Box



Generated Features

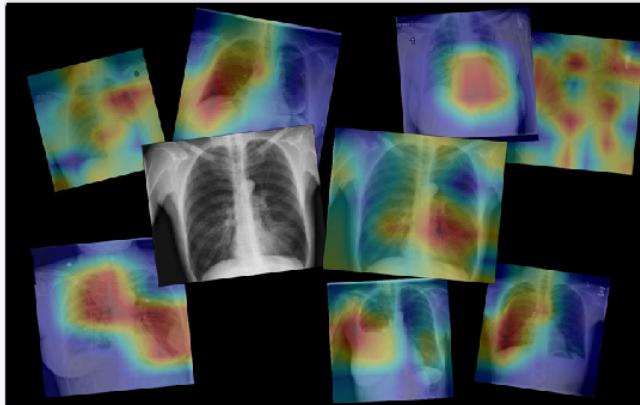
Feature Transformations

- Automatic Text Handling
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD
- Clustering and more



Andrew Ng @AndrewYNg · Nov 15

Our full paper on Deep Learning for pneumonia detection on Chest X-Rays.
@pranavrajpurkar @jeremy_irvin16 @mattlungrenMD
arxiv.org/abs/1711.05225



19 640 1.3K

Mistake

We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples for the pneumonia detection task. We randomly split the entire dataset into 80% training, and 20% validation.



Nick Roberts
@nizkroberts

Follow

Replies to @AndrewYNg @pranavrajpurkar and 2 others

Were you concerned that the network could memorize patient anatomy since patients cross train and validation?

"ChestX-ray14 dataset contains 112,120 frontal-view X-ray images of 30,805 unique patients. We randomly split the entire dataset into 80% training, and 20% validation."

3:26 AM - 16 Nov 2017 from Brooklyn, NY

1 Retweet 3 Likes



4 1 3

Tweet your reply



Arno Candel @ArnoCandel · Nov 16

Replies to @nizkroberts @AndrewYNg and 3 others

Reminds me of the common beginner mistake at the Allstate distracted drivers Kaggle competition :)

1 1 1

CheXNet (ours)	CheXNet (ours)
0.8209	0.8094
0.9048	0.9248
0.8831	0.8638
0.7204	0.7345
0.8618	0.8676
0.7766	0.7802
0.7632	0.7680
0.8932	0.8887
0.7939	0.7901
0.8932	0.8878
0.9260	0.9371
0.8044	0.8047
0.8138	0.8062
0.9387	0.9164

Automation needed to avoid human error

Submission history

From: Pranav Rajpurkar [view email]

[v1] Tue, 14 Nov 2017 17:58:50 GMT (16273kb,D)

[v2] Sat, 25 Nov 2017 04:21:27 GMT (321kb,D)

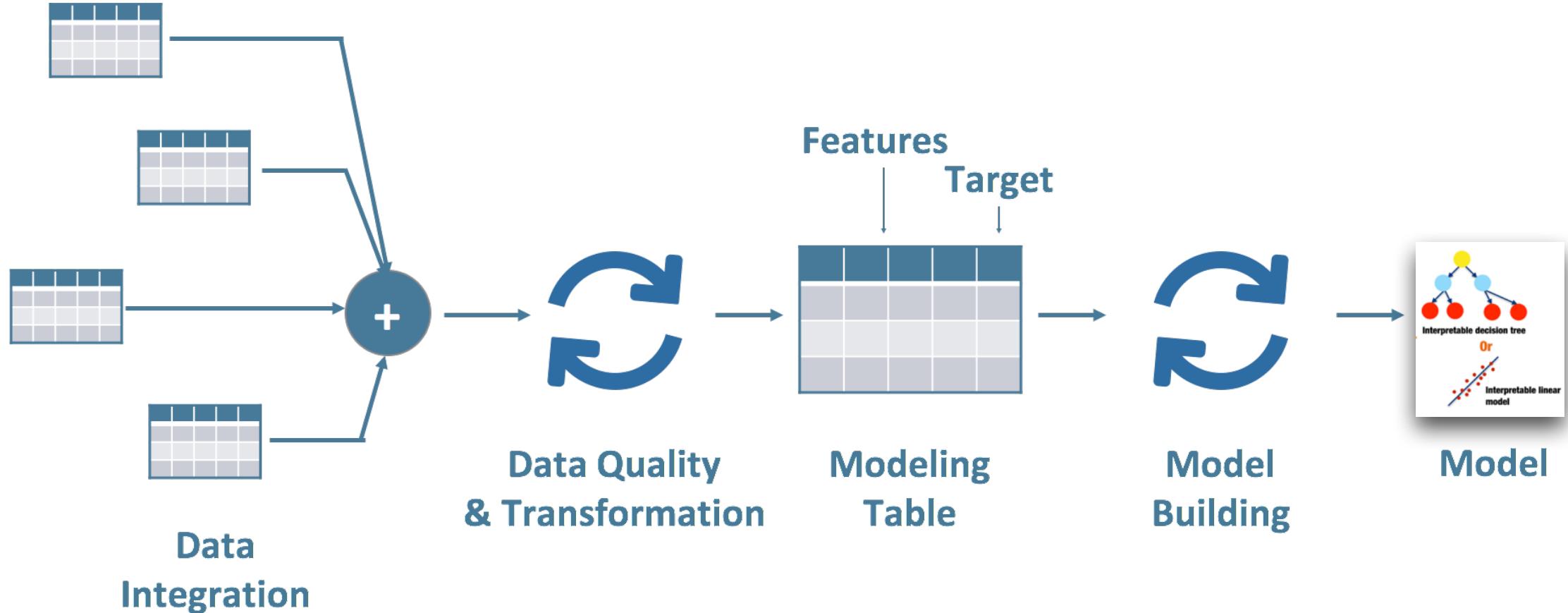
Correction

We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples. For the pneumonia detection task, we randomly split the dataset into training (28744 patients, 98637 images), validation (1672 patients, 6351 images), and test (389 patients, 420 images). There is no patient overlap between the sets.

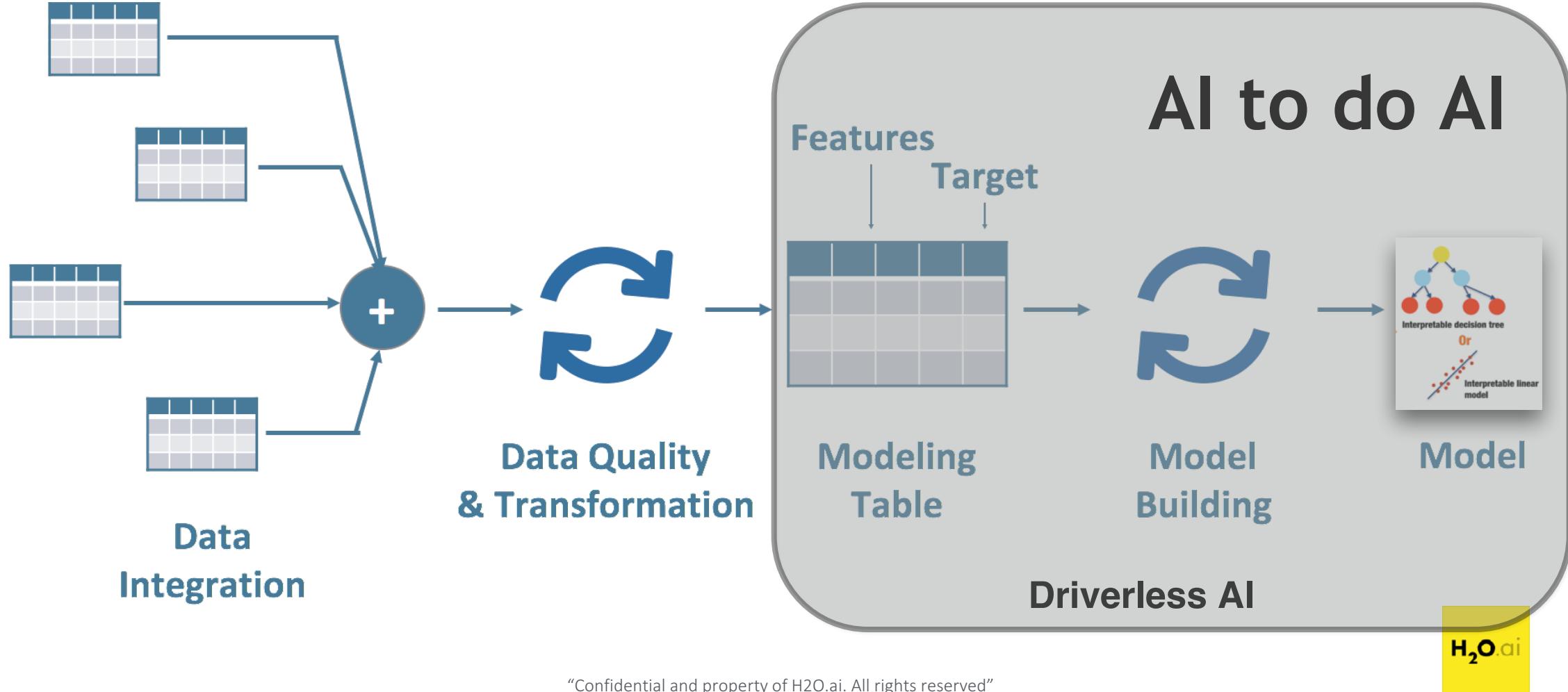
Driverless AI: AI to do AI

H₂O

Typical Enterprise Machine Learning Workflow



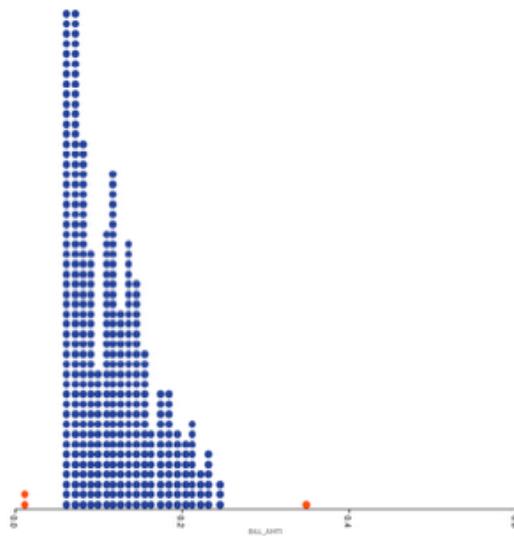
Typical Enterprise Machine Learning Workflow



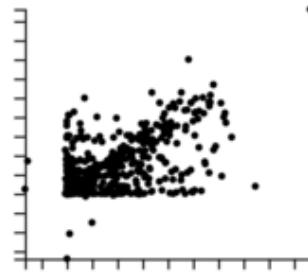
Automatic Visualization

H2O.ai

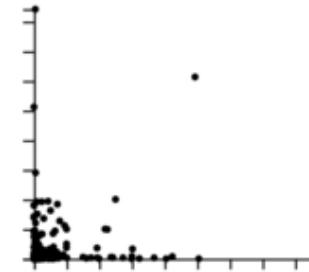
Automatic Scagnostics and other visualizations to generate the most relevant visualizations for each dataset



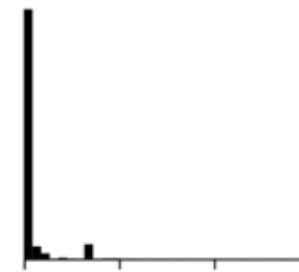
CLUMPY SCATTERPLOTS



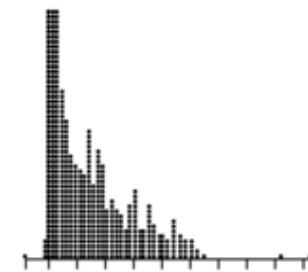
MONOTONIC SCATTERPLOTS



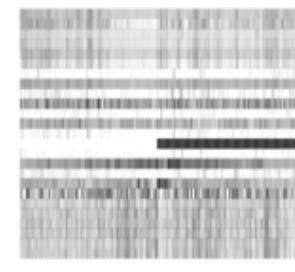
SPIKEY HISTOGRAMS



OUTLIERS



HEATMAPS



3 Pillars



Speed



Accuracy



Interpretability

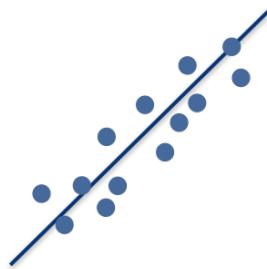
Speed

H₂O

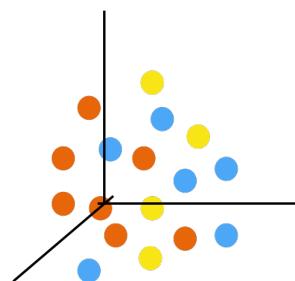
H2O4GPU Algorithms



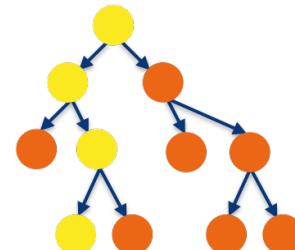
GLM



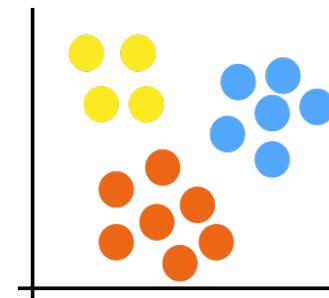
SVD



XGBoost



K-means



Algorithms on H₂O-3 (CPU)

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

Algorithms on H2O4GPU

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

Accuracy

H₂O

What's the “Secret” to Accuracy?

Key	Challenges
Feature Engineering	Time consuming & requires expert knowledge
Domain Knowledge	Requires years to experience in a domain to gain it
Hyperparameter Tuning	Complex and requires expert knowledge
Experimentation	Takes too long to run thousands of experiments
Debugging	Debugging models requires deep understanding of data and model

Accuracy

- Automatic feature engineering to increase accuracy - AlphaGo for AI
- Automatic Kaggle Grandmaster recipes in a box for solving wide variety of use-cases
- Automatic machine learning to find and tune the right ensemble of models

Driverless AI: top 5% in Amazon Kaggle competition

Driverless AI produced features and trained a "single run, fully automated" ensemble



Amazon.com - Employee Access Prediction

Predict an employee's access to sensitive data
\$5,000 · 1,687 teams · 4 years ago

Driverless AI: 80th place in private LB (out of 1687 - top 5%)

Driverless AI: top 1% in BNP Paribas Kaggle competition



single run, fully automated: 6h on 3 GPUs

BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?

\$30,000 · 2,926 teams · 2 years ago

Submission and Description Private Score

test_preds.csv
a few seconds ago by Arno Candel
Driverless AI 1.0.10 10/10/5 on 3 GPUs

Driverless AI: 18th place in private LB (out of 2926)

Hours for Driverless AI — Weeks for grandmasters



H2O.ai

Driving the Driverless AI

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ...

Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Zillow · 3,779 teams · 2 days ago

\$1,200,000 Prize Money

Competition Round One (Top 100 to Next Round)

40	▼ 8	Deal or No Deal		0.0749020	79	3mo
41	▲ 52	SCC		0.0749052	39	3mo
42	▼ 31	KFP		0.0749066	349	3mo

Finished above my H2O Kaggle Grandmasters colleagues

Search kaggle Competitions Datasets Kernels Discussion Jobs ...

Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Zillow · 3779 teams · 2 days ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Your most recent submission

Name: combine_027.csv	Submitted: 3 months ago	Wait time: 0 seconds	Execution time: 32 seconds	Score: 0.064025
-----------------------	-------------------------	----------------------	----------------------------	-----------------

Jump to your position on the Leaderboard ▾

Public Leaderboard Private Leaderboard

The private leaderboard is calculated with approximately 40% of the test data. This competition has completed. This leaderboard reflects the final standings.

#	Δ/Δp	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲ 9	Zensemble			0.0749061	253	3mo
2	▲ 1	Juan Zhai			0.0742108	53	3mo
3	▲ 133	Silogram-2			0.0749162	111	3mo
4	▲ 17	Alpha 60			0.0749059	180	3mo
5	▲ 418	Jack (Japan)			0.0744219	10	3mo
6	▲ 63	zhonglian			0.0745138	63	3mo
7	▲ 6	dset / alchohai			0.0746033	449	3mo
8	▲ 2	R2			0.0745701	302	3mo
9	▲ 1	Nirna Shahzadi			0.0745746	251	3mo
10	▲ 8	Zhihi Wang			0.0745842	95	3mo
11	▲ 257	Vistor S.D			0.0745902	61	3mo
12	▲ 204	alfe10			0.0747250	250	3mo
13	—	To Train Them Is My Cause			0.0747457	66	3mo
14	▲ 538	VV950713			0.0747700	69	3mo
15	▲ 2	Belinda Trotta			0.0747839	47	3mo
16	—	ivonik			0.0747903	118	3mo
17	▲ 3	Gough			0.0747993	56	3mo
18	▲ 6	raytrace			0.0748019	47	3mo
19	▲ 1384	Dr. Knape			0.0748106	10	3mo
20	▲ 16	Trottefex			0.0748134	86	3mo
21	▲ 27	gooren			0.0748159	101	3mo
22	▲ 5	no one			0.0748159	25	3mo
23	▲ 6	mlin			0.0748179	28	3mo
24	▲ 5	Bierkem			0.0748237	72	3mo
25	▲ 10	Thomas Hoffmann			0.0748328	67	3mo
26	▲ 3	Dmitry Kulagin			0.0748457	27	3mo
27	▲ 414	Commander Keen			0.0748479	35	3mo
28	▲ 21	Zidme & Kostadinov & L			0.0748479	273	3mo
29	▲ 11	Zillow			0.0748574	454	3mo
30	▲ 1	Heigl			0.0748629	197	3mo
31	▲ 177	anonymouseus			0.0748652	11	3mo
32	▲ 4	刘伟山(1981年1月1日-1995年1月1日)			0.0748730	7	3mo
33	▲ 120	Jun Wen			0.0748847	61	3mo
34	▲ 27	The Ox			0.0748888	45	3mo
35	▲ 5	prod by adverb			0.0748907	113	3mo
36	▲ 10	Thomas H. Thoreen Kjetil A...			0.0748955	136	3mo
37	▲ 40	Pauo Pinto			0.0748944	136	3mo
38	▲ 115	sloston			0.0748955	65	3mo
39	▲ 30	FF			0.0748983	14	3mo
40	▲ 8	Deal or No Deal			0.0749020	79	3mo
41	▲ 62	SCC			0.0749052	39	3mo
42	▲ 31	KIP			0.0749066	349	3mo
43	▲ 17	ys			0.0749078	81	3mo
44	▲ 19	hbo			0.0749165	57	3mo
45	▲ 66	Dominic Brassandra			0.0749197	136	3mo
46	▲ 18	russianscube			0.0749202	130	3mo
47	▲ 23	LuccelBr			0.0749223	13	3mo
48	▲ 90	JavierBlez			0.0749244	244	3mo
49	▲ 55	ywlkm			0.0749260	40	3mo
50	▲ 167	三毛真美呀			0.0749329	72	3mo

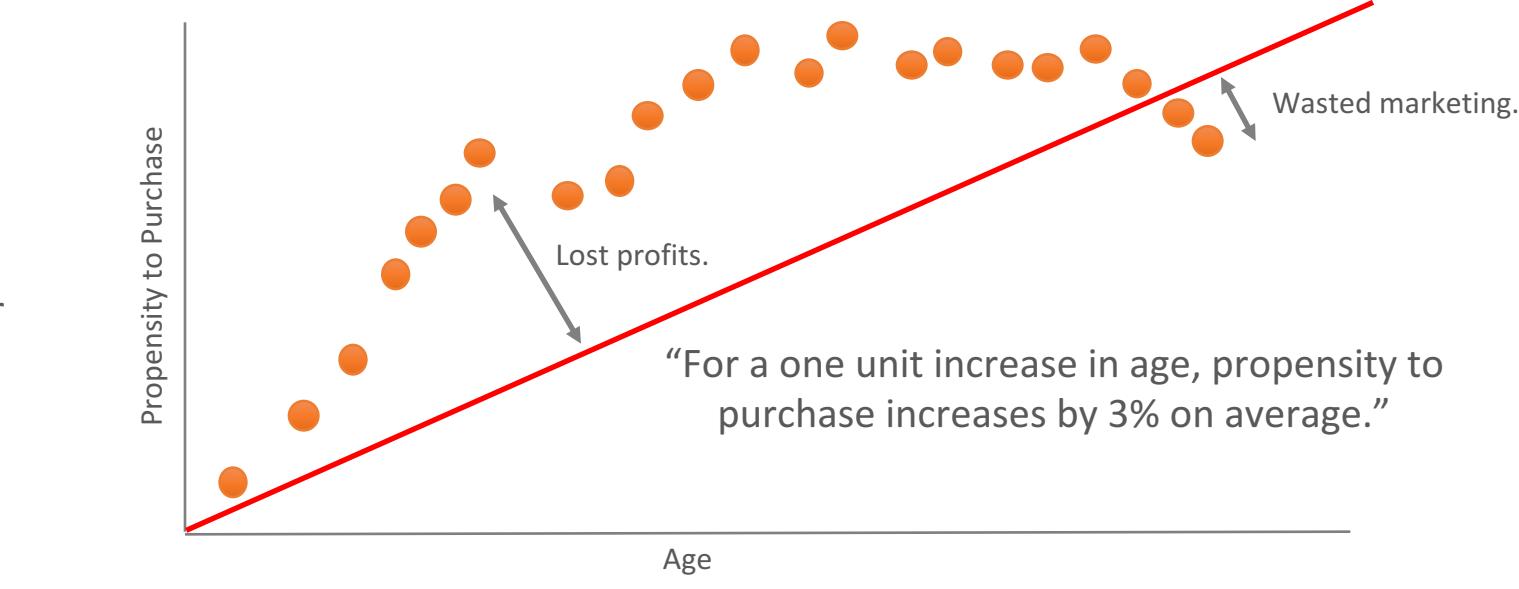
Machine Learning Interpretability

An approximate interpretation of an accurate model
is better than an exact interpretation of an approximate model

H₂O

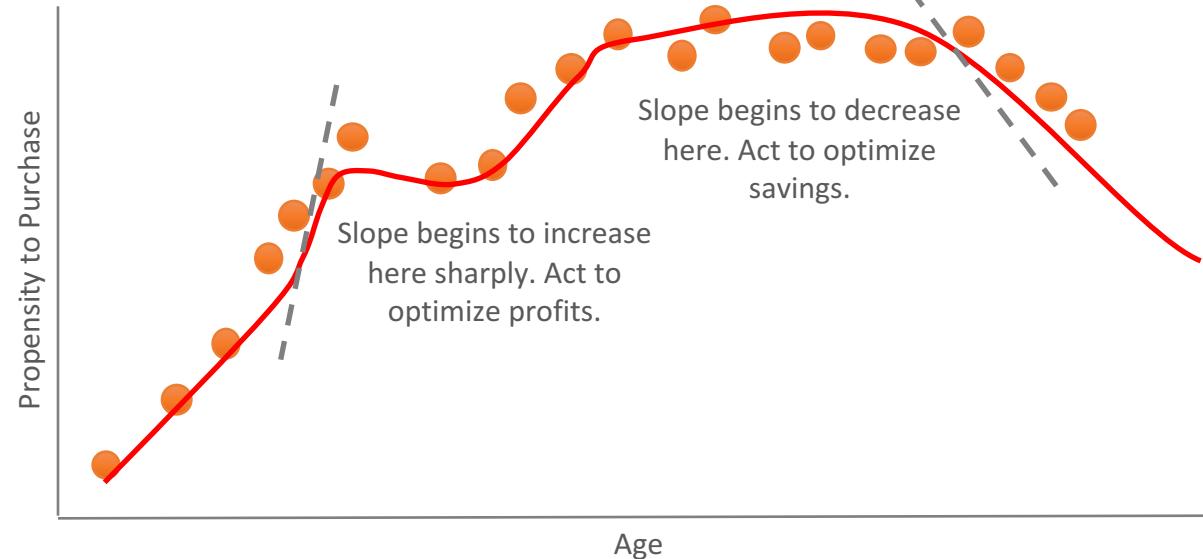
Linear Models

Exact explanations for
approximate models.



Machine Learning

Approximate explanations
for **exact** models.

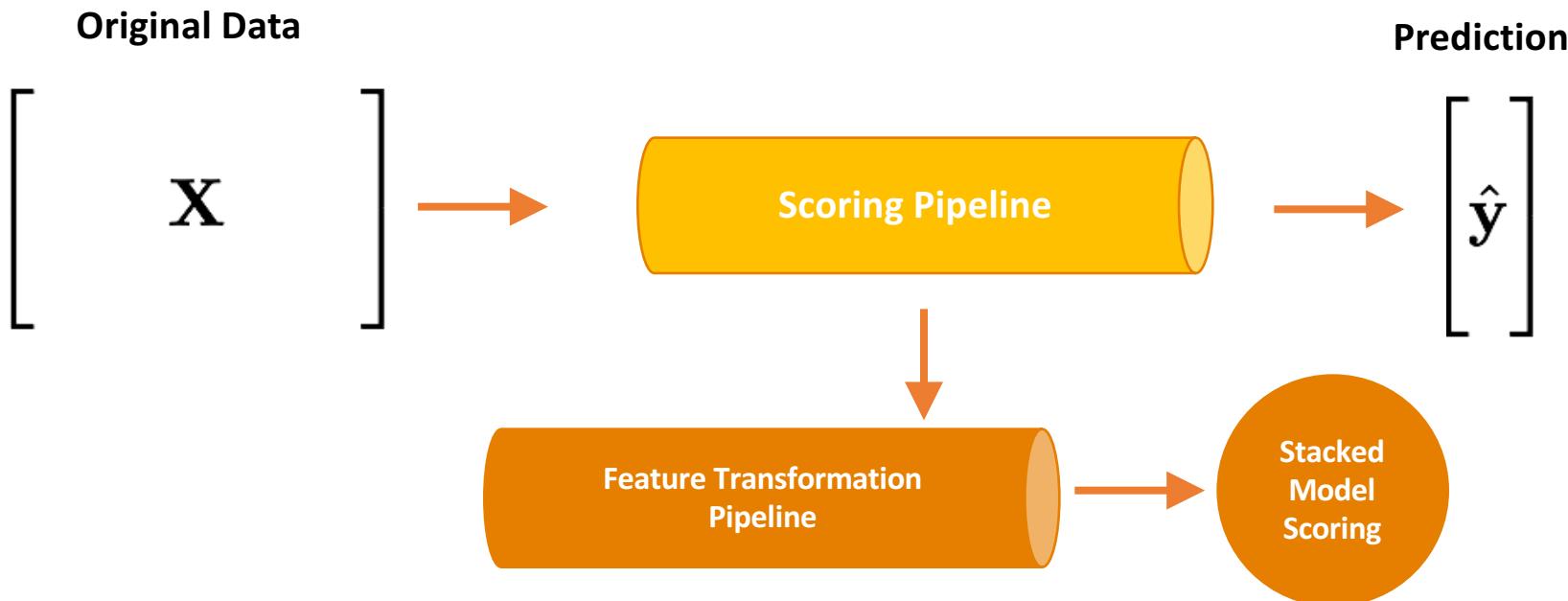


Deployment

H₂O

Scoring Pipelines

- Scoring pipelines are independent of Driverless AI
- Contain the feature engineering logic and final model scoring



Scoring Pipeline Options

Scoring Option	Description	Use When you Want..
Python Scoring Pipeline	Python module	<ul style="list-style-type: none">• Minimum dependencies• Everything done in Python on a single machine
Python Scoring Service	Hosts the Python module as an HTTP or TCP service	<ul style="list-style-type: none">• To host the scoring module as an HTTP or TCP service• To invoke the Python scoring module from languages other than Python• To invoke the Python scoring module from another computer
MOJO Scoring Pipeline	Java object	<ul style="list-style-type: none">• To score in real time• MOJO offers the fastest latency but is still in alpha state

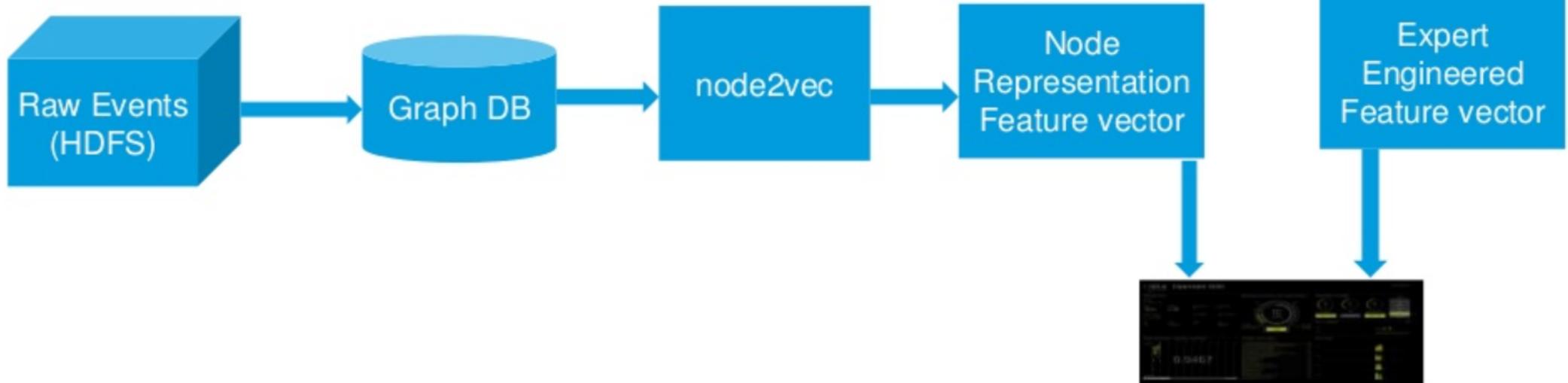
PayPal Use-Case

H₂O

DRIVERLESS AI AT PAYPAL



Human Expert



Driverless AI
(Feature Engineering +
Model Training)

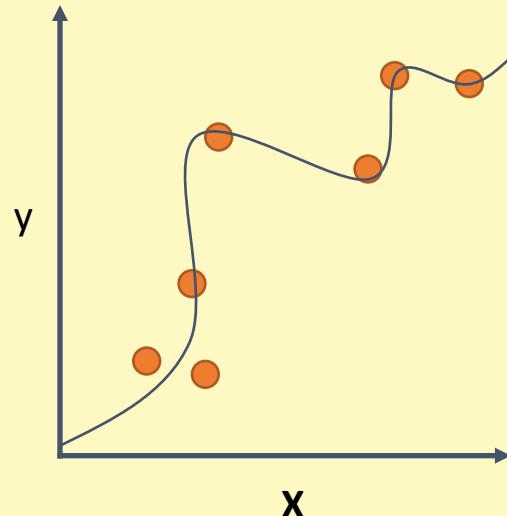
Demo

H₂O

Supervised Learning

Regression:

How much will a customer spend?

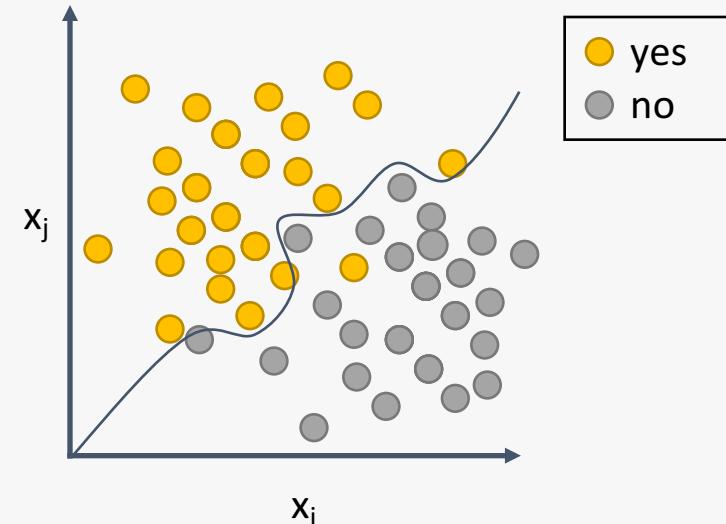


H₂O algos:

Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:

Will a customer churn?



H₂O algos:

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

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
...

Other News

H₂O



London Artificial Intelligence & Deep Learning

PRO

H2O Artificial Intelligence and Machine Learning - 39 groups

Location

London, United Kingdom

Members

5,851



Organizers

Ian Gomez and 1 other

Schedule

...



Hit a Home Run Making Baseball Decisions Using Artificial Intelligence and Machine Learning

Thursday, 1:30 PM - 2:10 PM | Session ID: 3456A

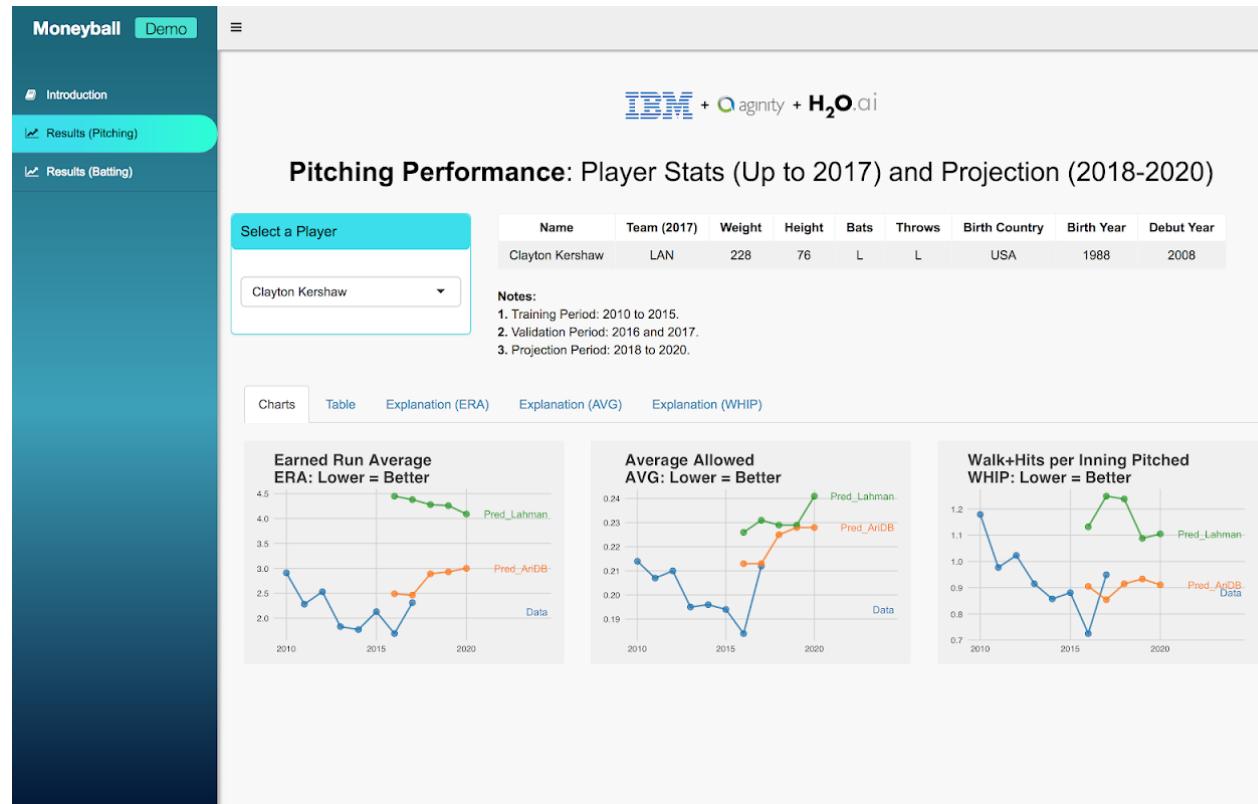
Mandalay Bay South, Level 2 | Breakers C



Join Ari Kaplan, a real "MoneyBall" and well known around Major League Baseball, Joe Chow, a H2O data scientist, and David Kearns from IBM's Analytics Ecosystem team for this fun, interactive session where you will have the chance to see where artificial intelligence meets business intelligence. Ari and Joe will briefly present the latest machine learning technologies and concepts powering today's baseball decisions, including Hortonworks Data Platform, Spark, Aginity Amp, H2O.ai, IBM Digital Science Experience and more. You will then step up to the plate as general manager to see how your player decisions would stack up under World Series pressure. Are you ready to play ball?

Speakers:

- Ari Kaplan, Aginity
- Jo-fai Chow, H2O.ai
- David Kearns, IBM



Thank you!

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

For more information, go to www.h2o.ai

