Driverless AI: AI to do AI

Jo-fai (Joe) Chow
Data Scientist and Community Manager at H_2O.ai
joe@h2o.ai

Driverless AI: AI to do AI

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

Company Overview

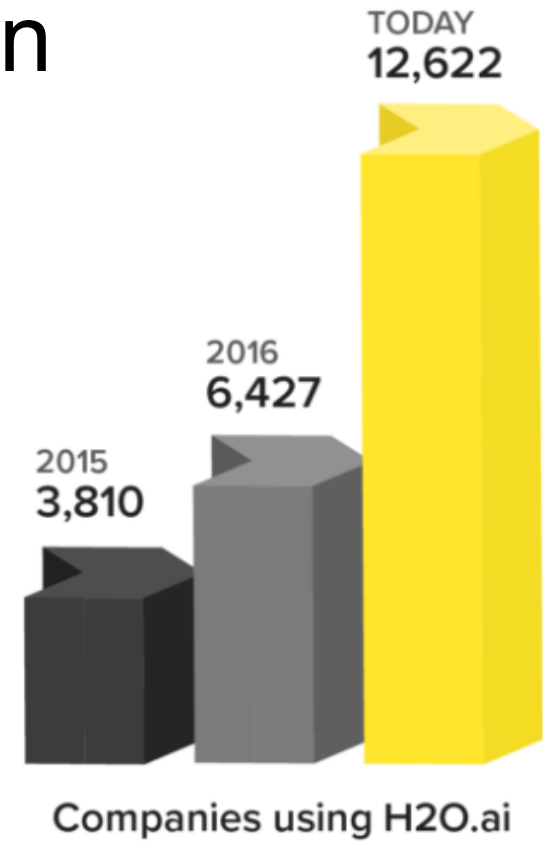
Founded	2012, Series C in Nov, 2017
Products	• Driverless AI – Automated Machine Learning • H2O Open Source Machine Learning • Sparkling Water
Mission	Democratize AI. Do Good
Team	~100 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Offices	Mountain View, London, Prague





H2O.ai

Worldwide Community Adoption

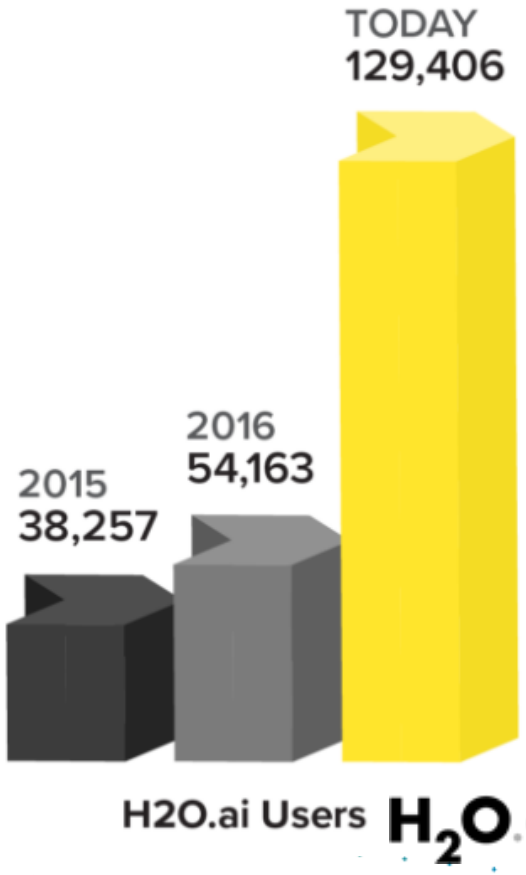


222 OF THE FORTUNE 500

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

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)

Platforms with H_2O integration

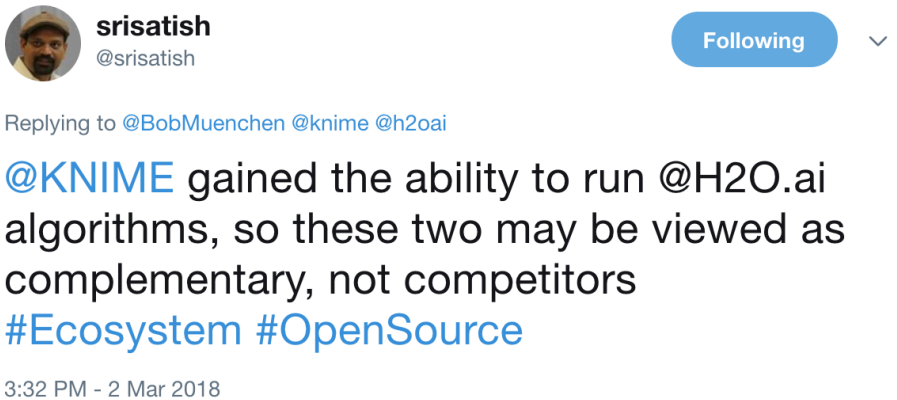


Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms

Source: Gartner (February 2018)

H2O.ai Solution Leadership Across Verticals

H2O.ai

H_2O 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

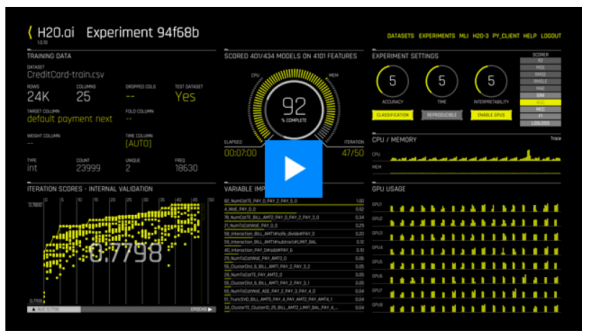
Automatic feature engineering, machine learning and interpretability

Secure multi-tenant H2O clusters

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

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: *

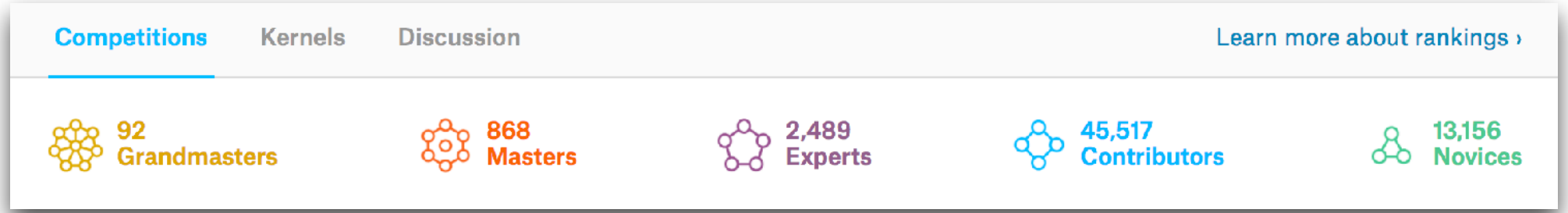
SEND ME A TRIAL LICENSE

Why Driverless AI?

Shortage of Data Scientists

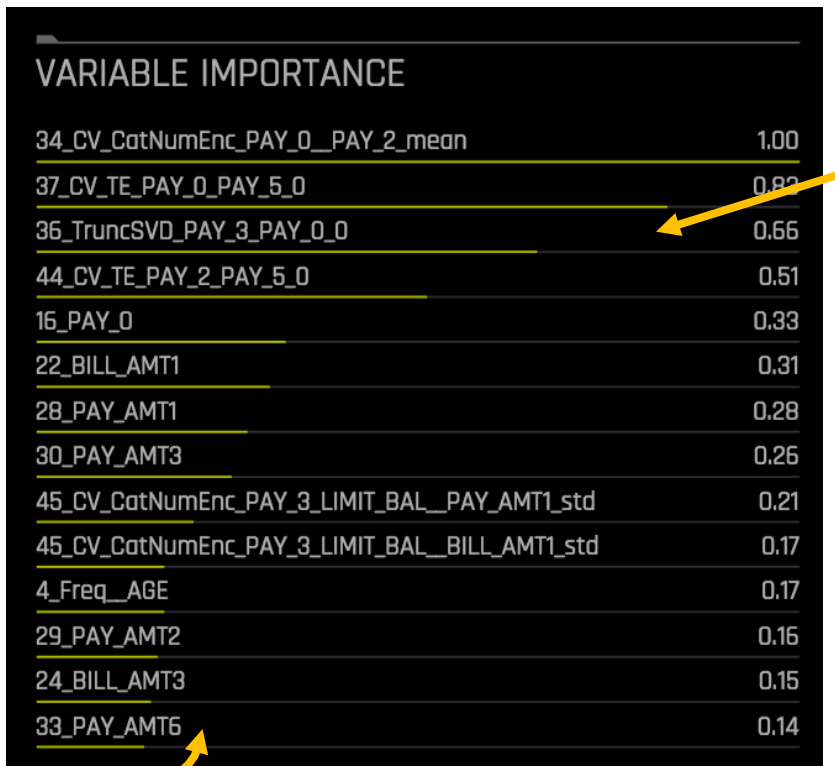
"The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts"

–McKinsey Prediction for 2018



H2O.ai

Auto Feature Generation
Kaggle Grand Master Out of the Box

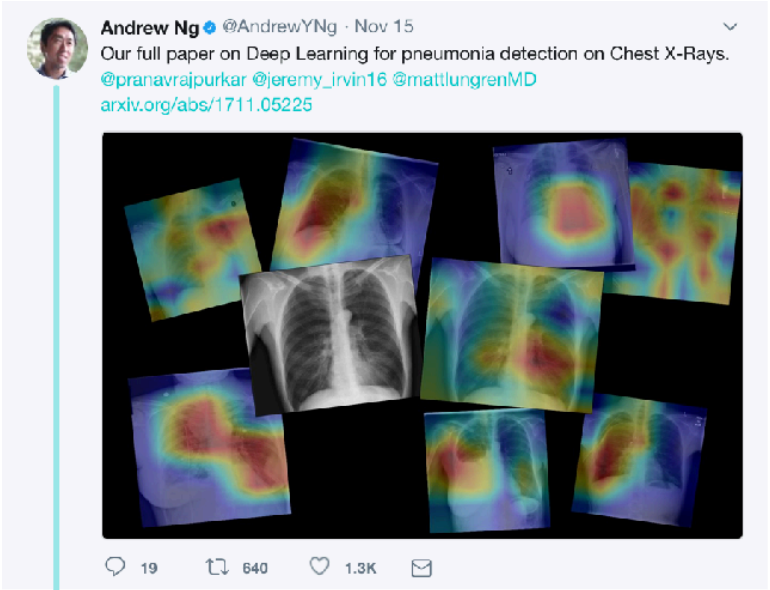


Generated Features

Original Features

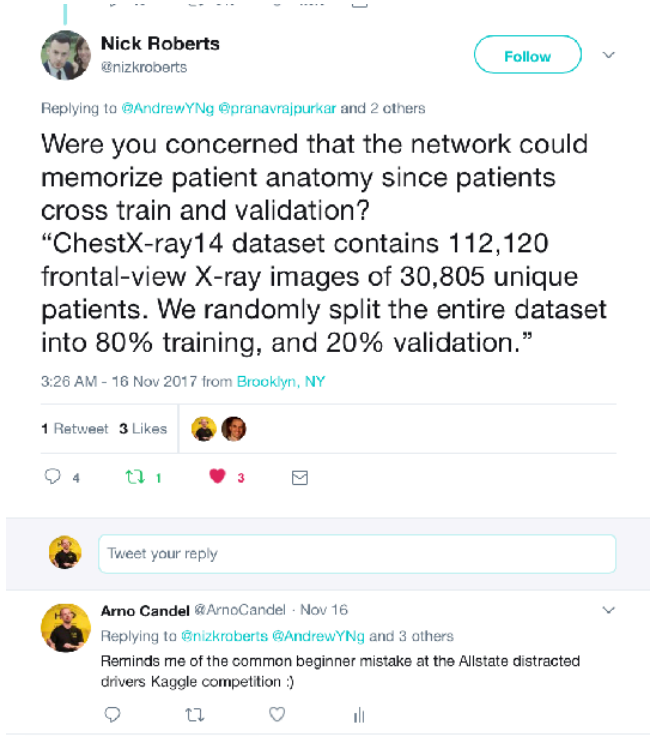
Feature Transformations

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



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.



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)

CheXNet (ours)
0.8209
0.9048
0.8831
0.7204
0.8618
0.7766
0.7632
0.8932
0.7939
0.8932
0.9260
0.8044
0.8138
0.9387

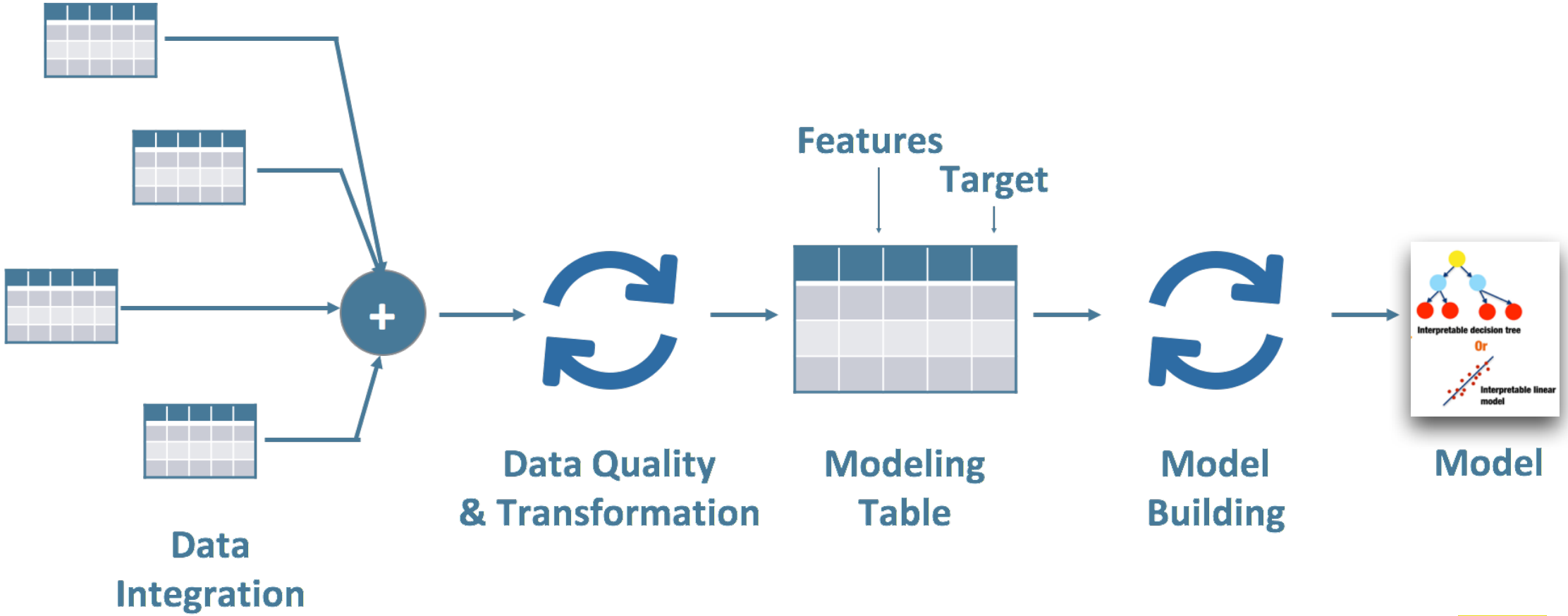
CheXNet (ours)
0.8094
0.9248
0.8638
0.7345
0.8676
0.7802
0.7680
0.8887
0.7901
0.8878
0.9371
0.8047
0.8062
0.9164

Correction

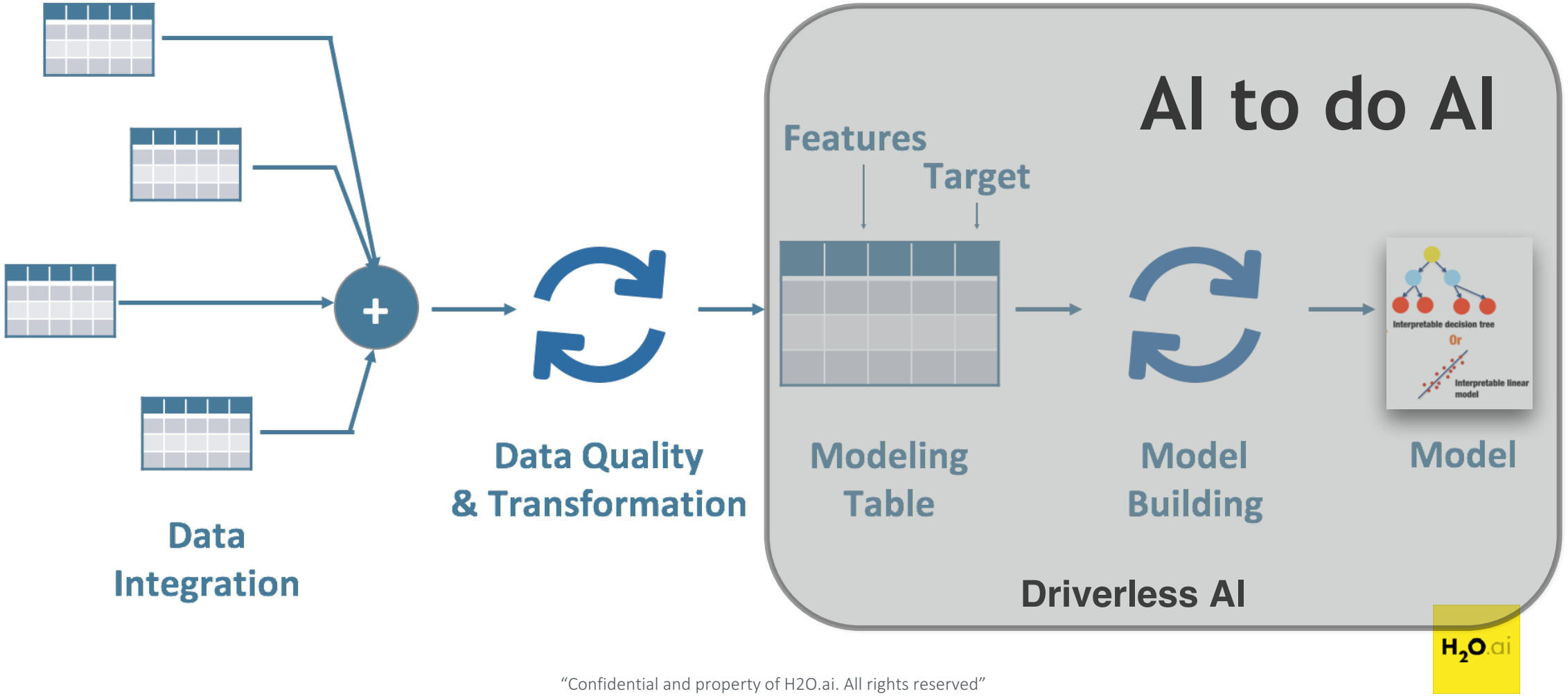
We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images 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

Typical Enterprise Machine Learning Workflow

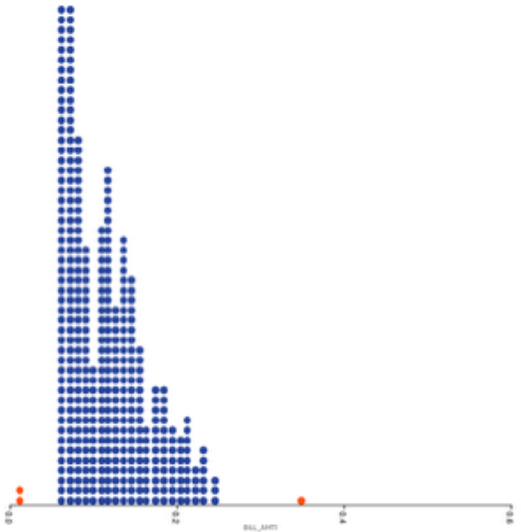


Typical Enterprise Machine Learning Workflow



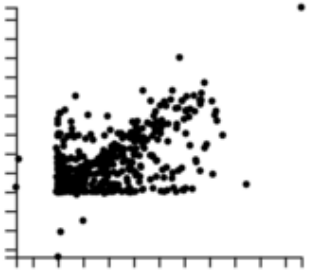
Automatic Visualization

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

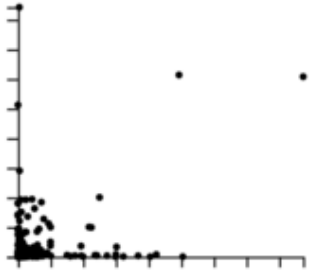


H2O.ai

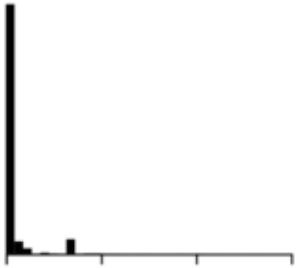
CLUMPY SCATTERPLOTS



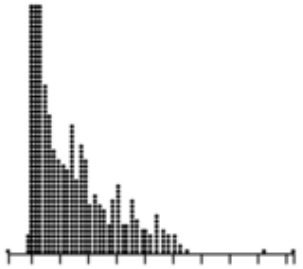
MONOTONIC SCATTERPLOTS



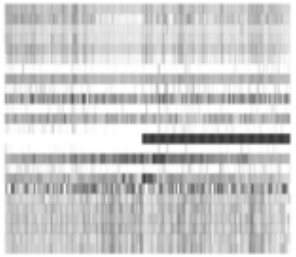
SPIKEY HISTOGRAMS



OUTLIERS



HEATMAPS



H2O.ai

3 Pillars

Speed

Accuracy

Interpretability

H2O.ai

Speed

H_2O

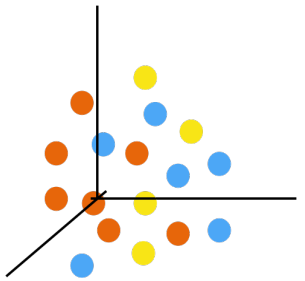
H2O4GPU Algorithms

5X

GLM

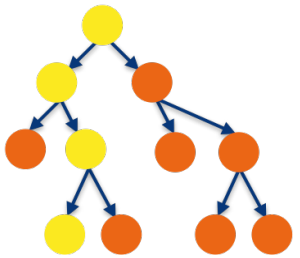
5X

SVD



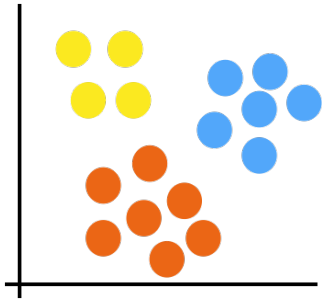
10X

XGBoost



40X

K-means



Algorithms on H_2O-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

H2O.ai

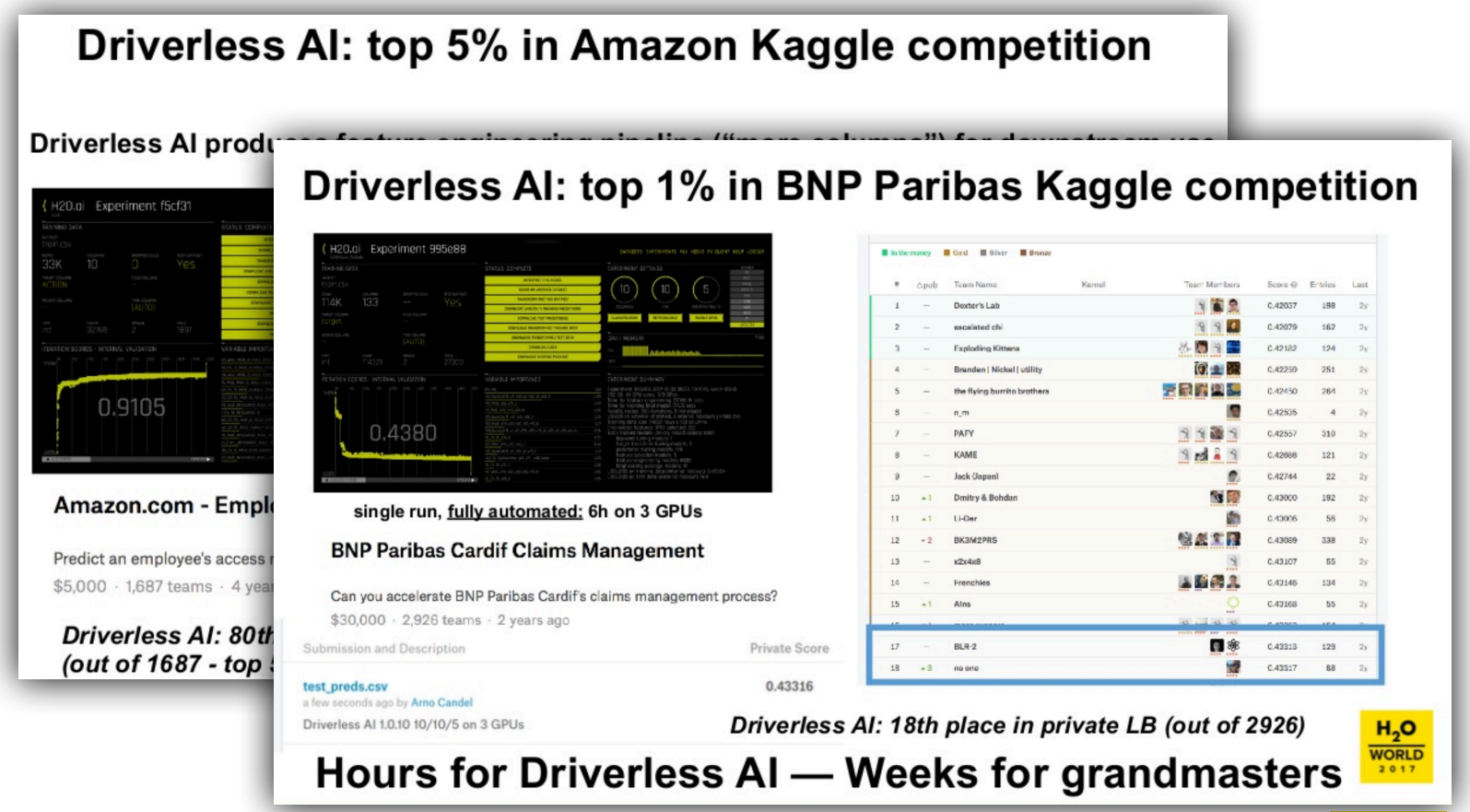
Accuracy

What's the "Secret" to Accuracy?

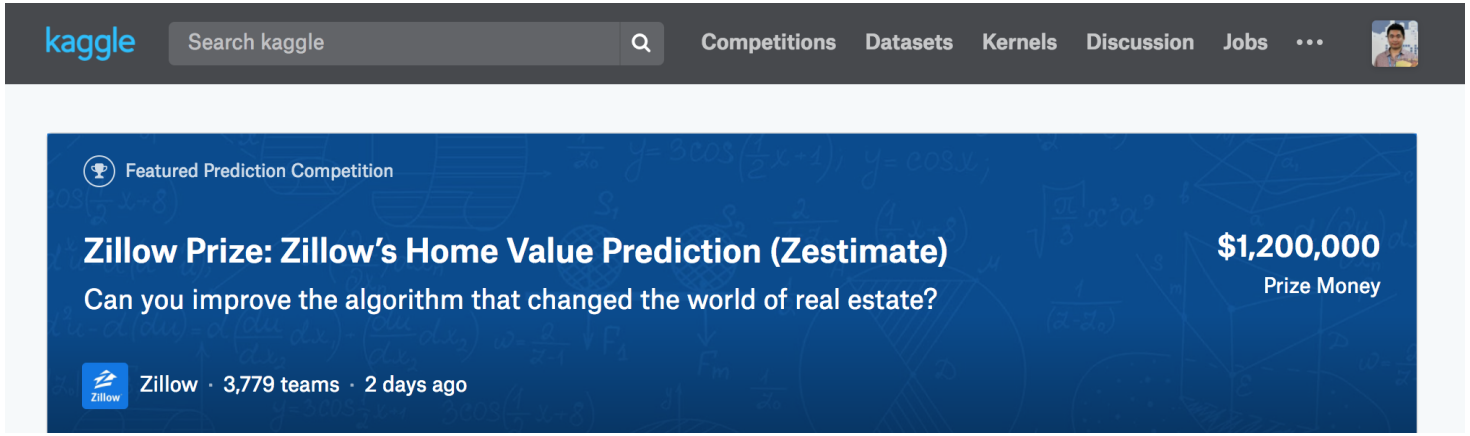
Key	Challenges
Feature Engineering	Time consuming & requires expert knowledge
Domain Knowledge	Requires years to experience in a domain to gain it
Hyperparameter Tuning	Complex and requires expert knowledge
Experimentation	Takes too long to run thousands of experiments
Debugging	Debugging models requires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

Accuracy

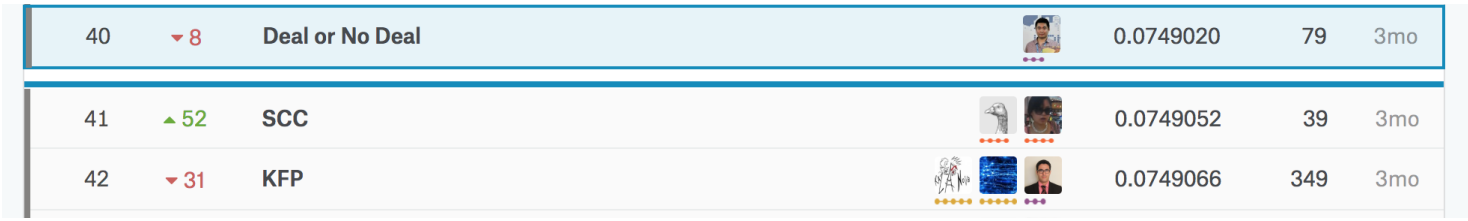
- Automatic feature engineering to increase accuracy - AlphaGo for AI
- Automatic Kaggle Grandmaster recipes in a box for solving wide variety of use-cases
- Automatic machine learning to find and tune the right ensemble of models



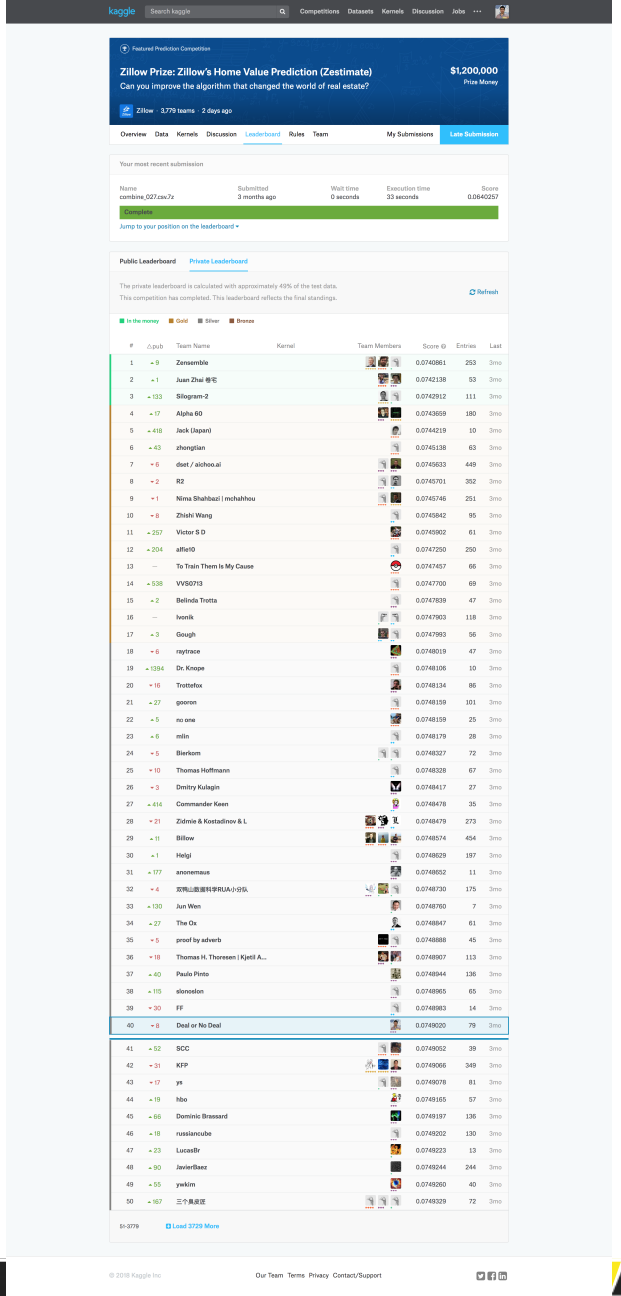
Driving the Driverless AI



Competition Round One (Top 100 to Next Round)



Finished above my H2O Kaggle Grandmasters colleagues

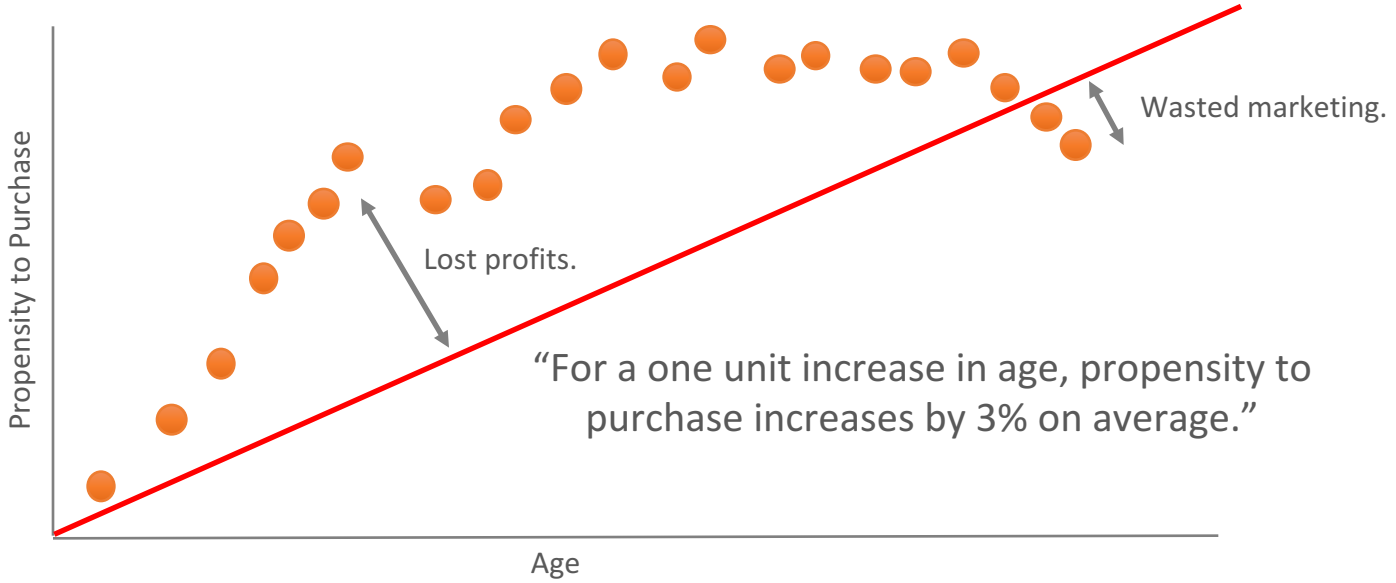


Machine Learning Interpretability

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

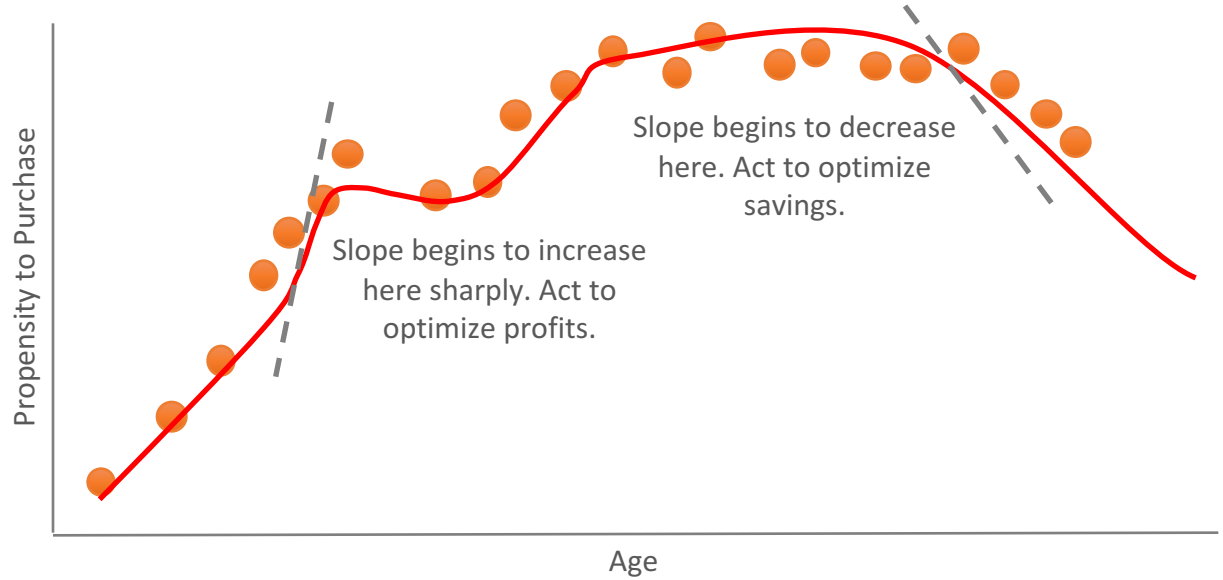
Linear Models

Exact explanations for ***approximate*** models.



Machine Learning

Approximate explanations for ***exact*** models.

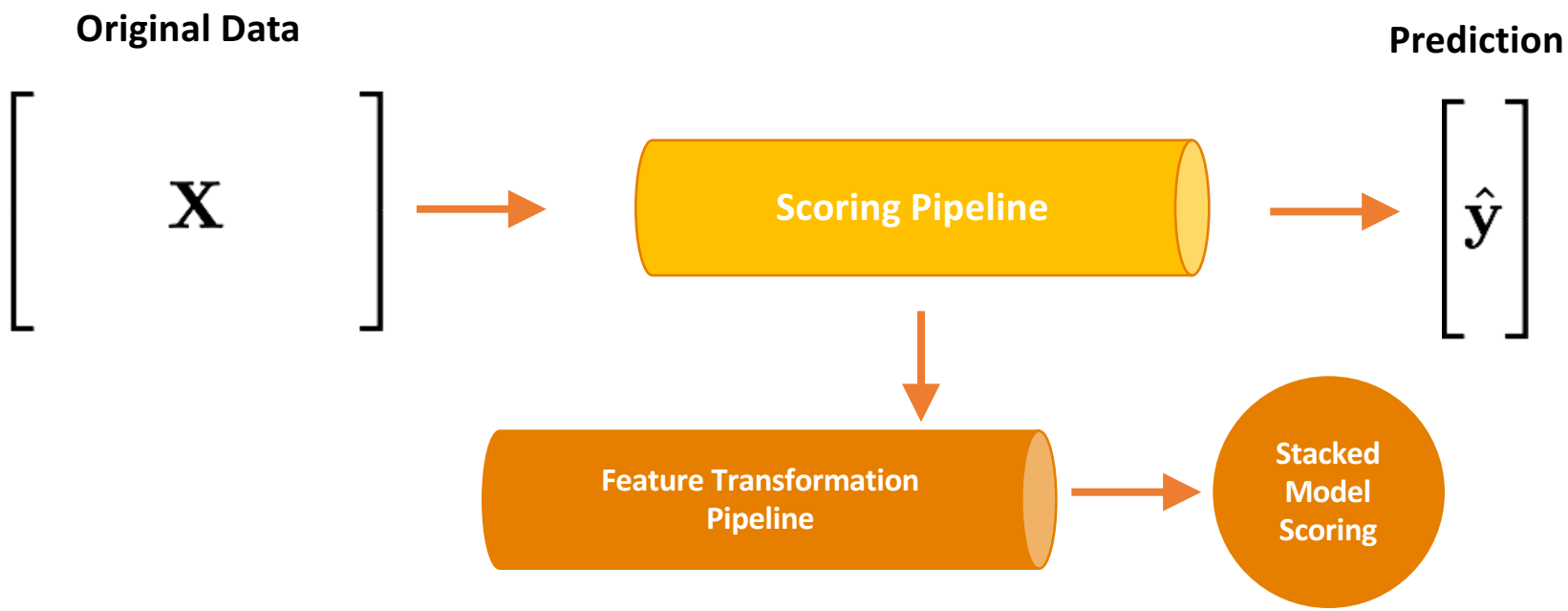


H2O.ai

Deployment

Scoring Pipelines

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



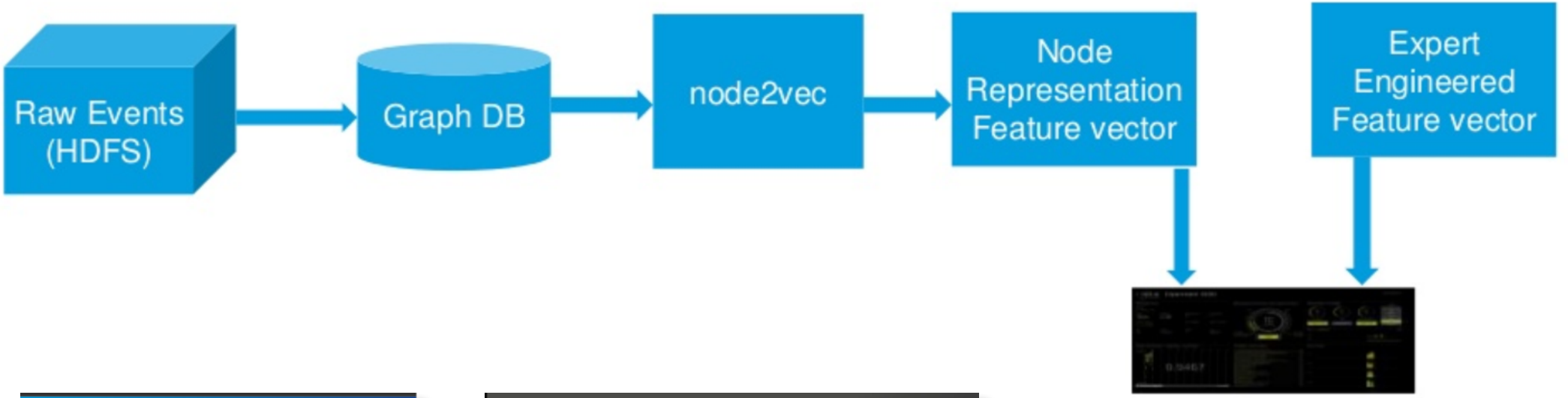
Scoring Pipeline Options

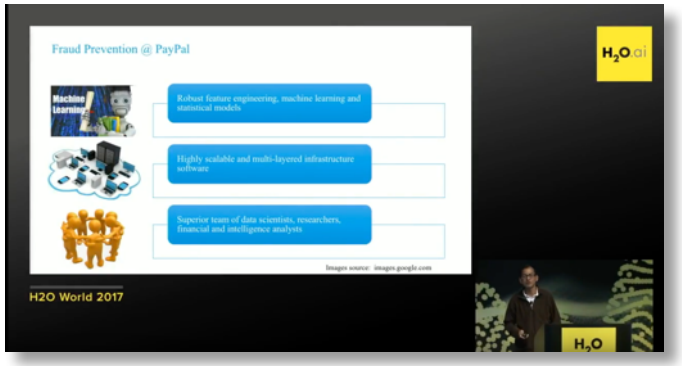
Scoring Option	Description	Use When you Want..
Python Scoring Pipeline	Python module	• Minimum dependencies • Everything done in Python on a single machine
Python Scoring Service	Hosts the Python module as an HTTP or TCP service	• To host the scoring module as an HTTP or TCP service • To invoke the Python scoring module from languages other than Python • To invoke the Python scoring module from another computer
MOJO Scoring Pipeline	Java object	• To score in real time • MOJO offers the fastest latency but is still in alpha state

PayPal Use-Case

DRIVERLESS AI AT PAYPAL

Human Expert

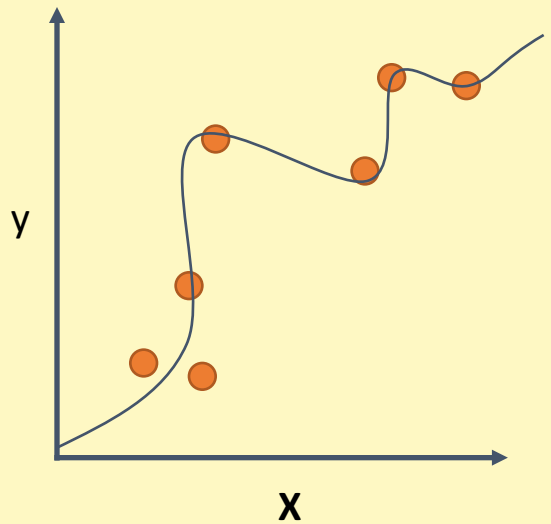




Demo

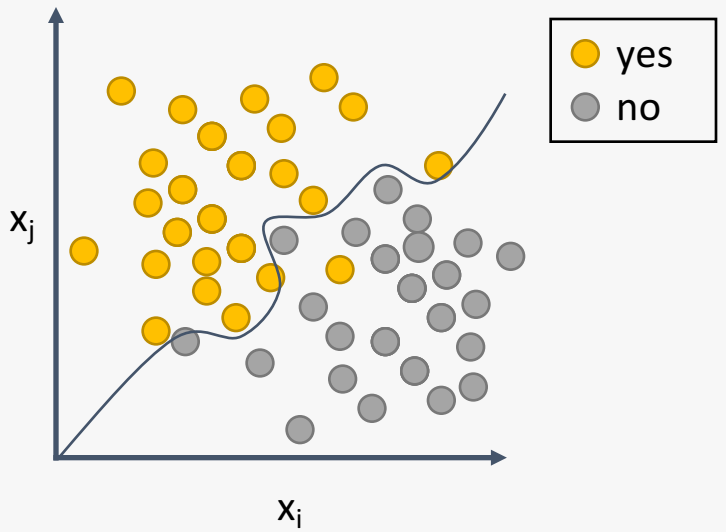
Supervised Learning

Regression:
How much will a customer spend?



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

Classification:
Will a customer churn?



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

…

Other News

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

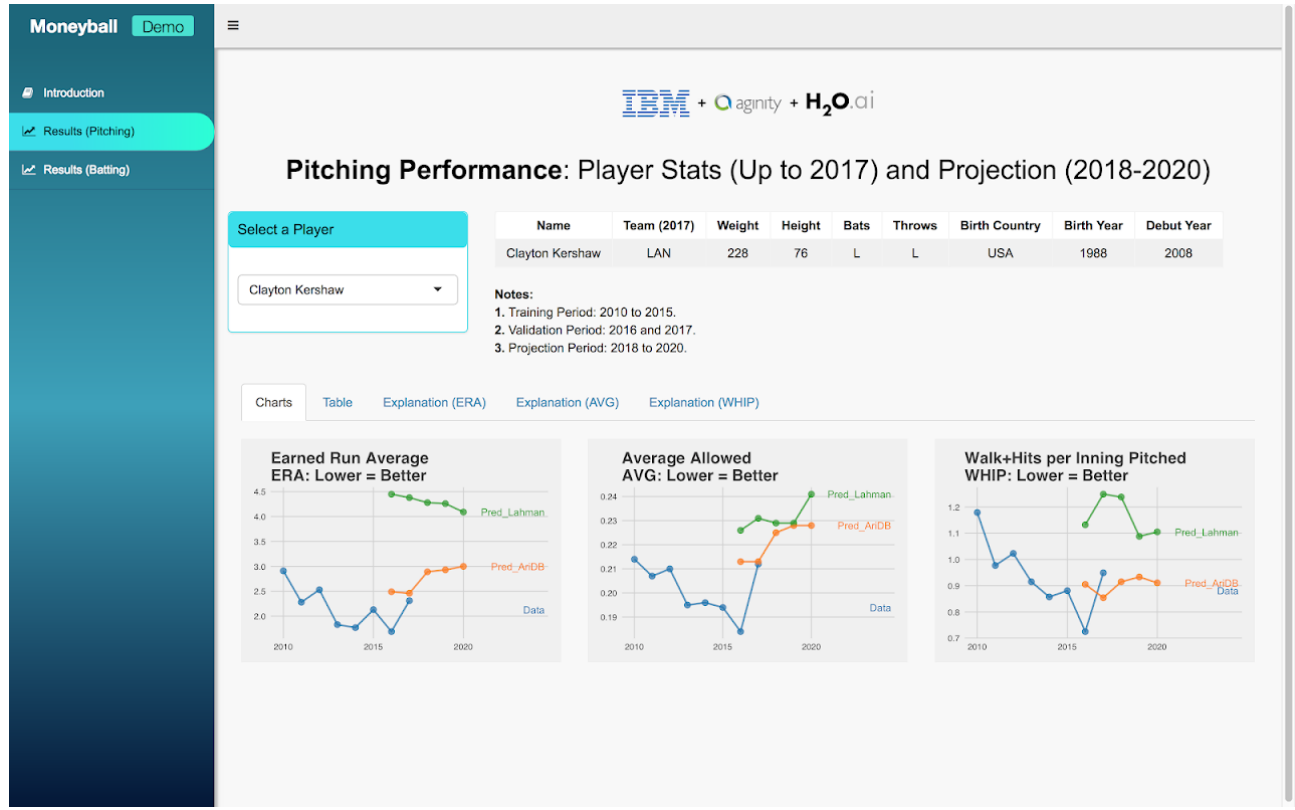
Mandalay Bay South, Level 2 | Breakers C

IBM + aginity + H_2O.ai

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 H_2O.ai
joe@h2o.ai

For more information, go to www.h2o.ai