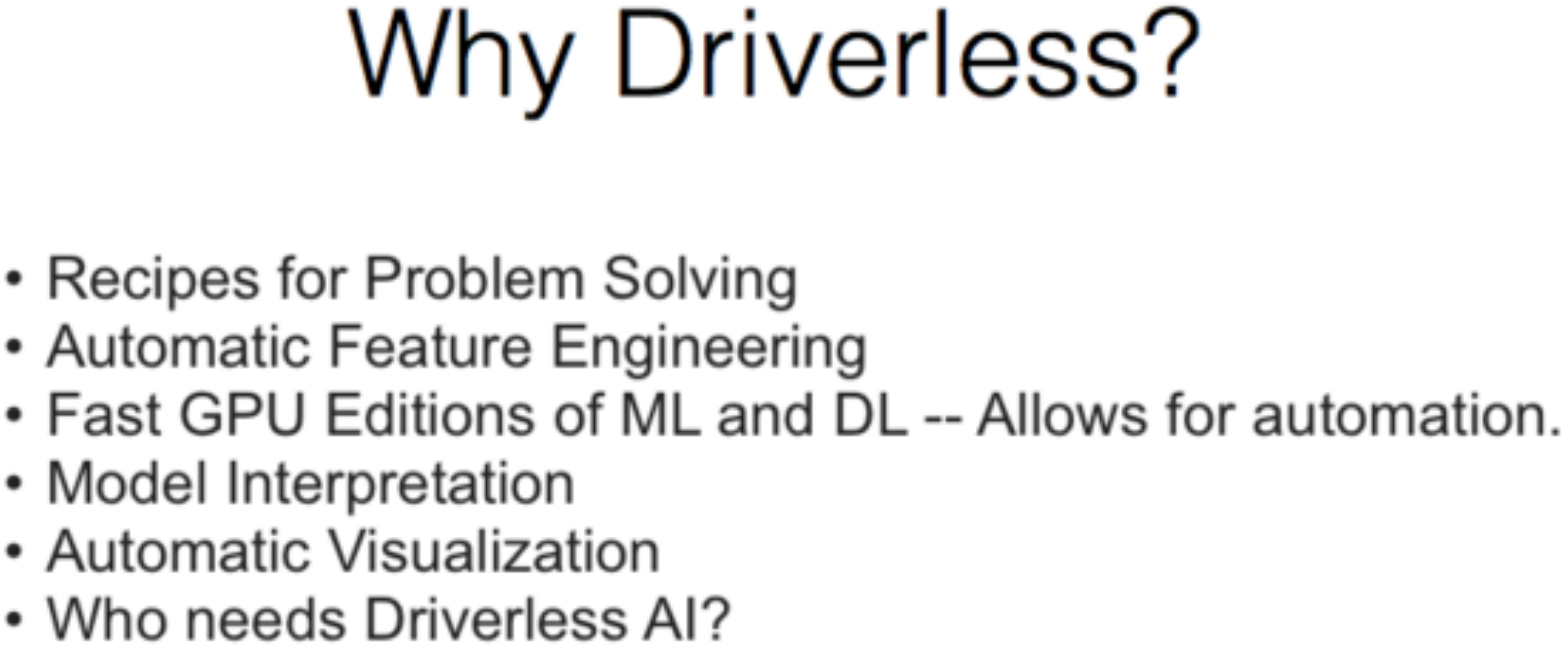
H2O Driverless AI

Webinar

H2O.ai, Inc, Jul 27 2017



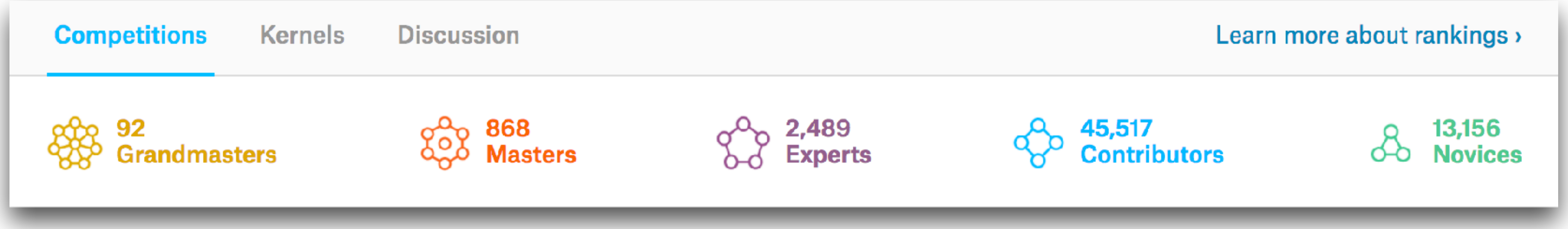
Why Driverless?

- Recipes for Problem Solving
- Automatic Feature Engineering
- Fast GPU Editions of ML and DL -- Allows for automation.
- Model Interpretation
- Automatic Visualization
- Who needs 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



Competitions Kernels Discussion Learn more about rankings ›

92 Grandmasters

868 Masters

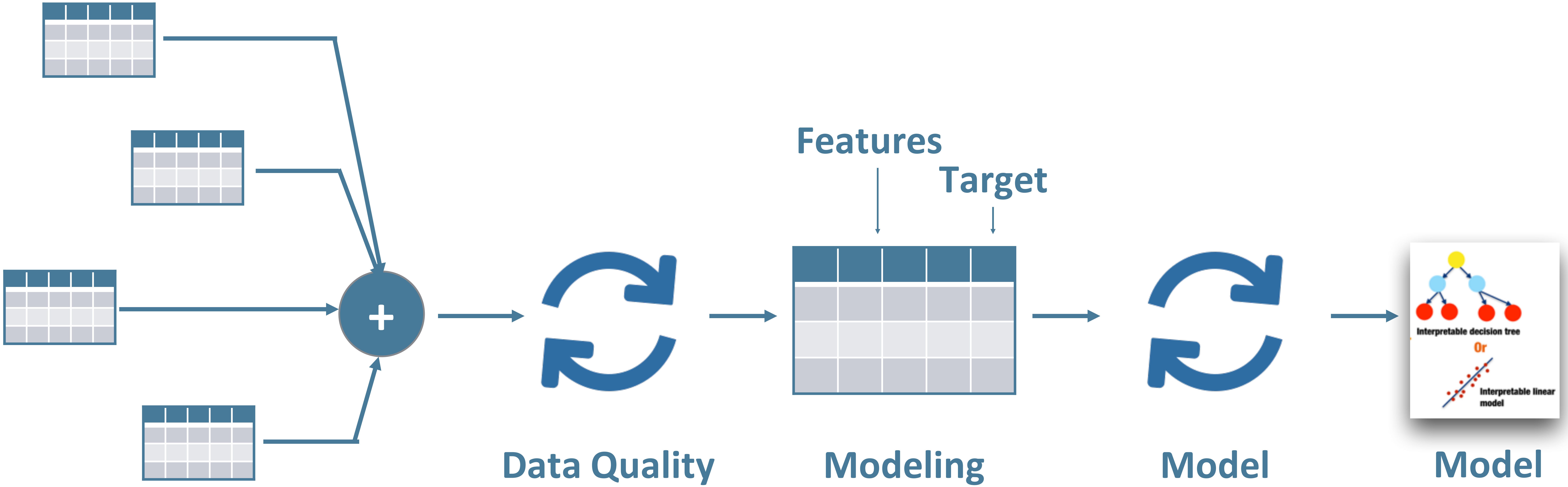
2,489 Experts

45,517 Contributors

13,156 Novices

"Confidential and property of H2O.ai. All rights reserved"

Typical Enterprise Machine Learning Workflow



Features

Target

Data Integration

Data Quality & Transformation

Modeling Table

Model Building

Model

Typical Enterprise Machine Learning Workflow

Data Integration

Data Quality & Transformation

Features

Target

Modeling Table

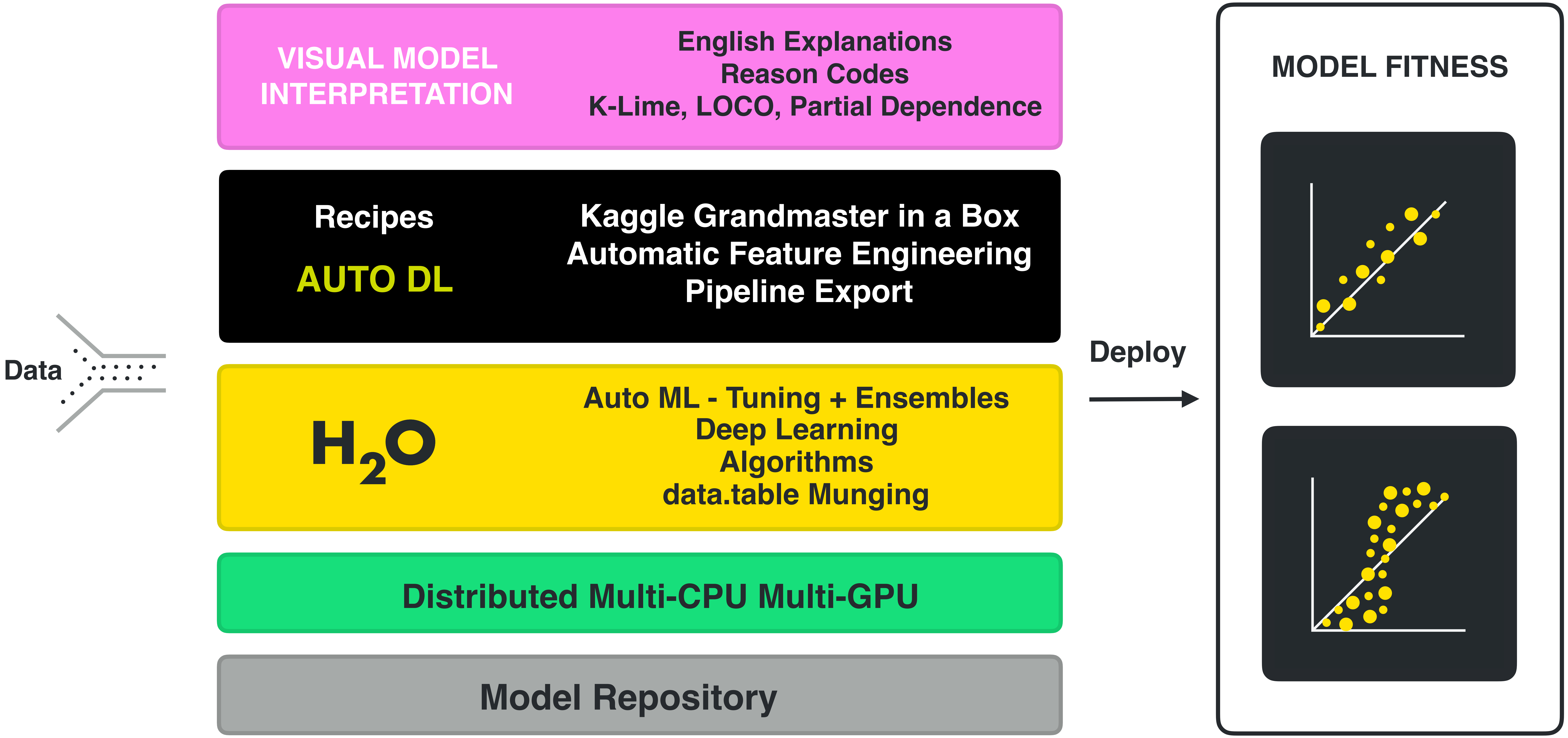
Model Building

Model

Interpretable decision tree
Or
Interpretable linear model

Driverless AI

Driverless AI



Data

VISUAL MODEL INTERPRETATION

English Explanations
Reason Codes
K-Lime, LOCO, Partial Dependence

Recipes

AUTO DL

Kaggle Grandmaster in a Box
Automatic Feature Engineering
Pipeline Export

H_2O

Auto ML - Tuning + Ensembles
Deep Learning
Algorithms
data.table Munging

Distributed Multi-CPU Multi-GPU

Model Repository

Deploy

MODEL FITNESS

3 Pillars

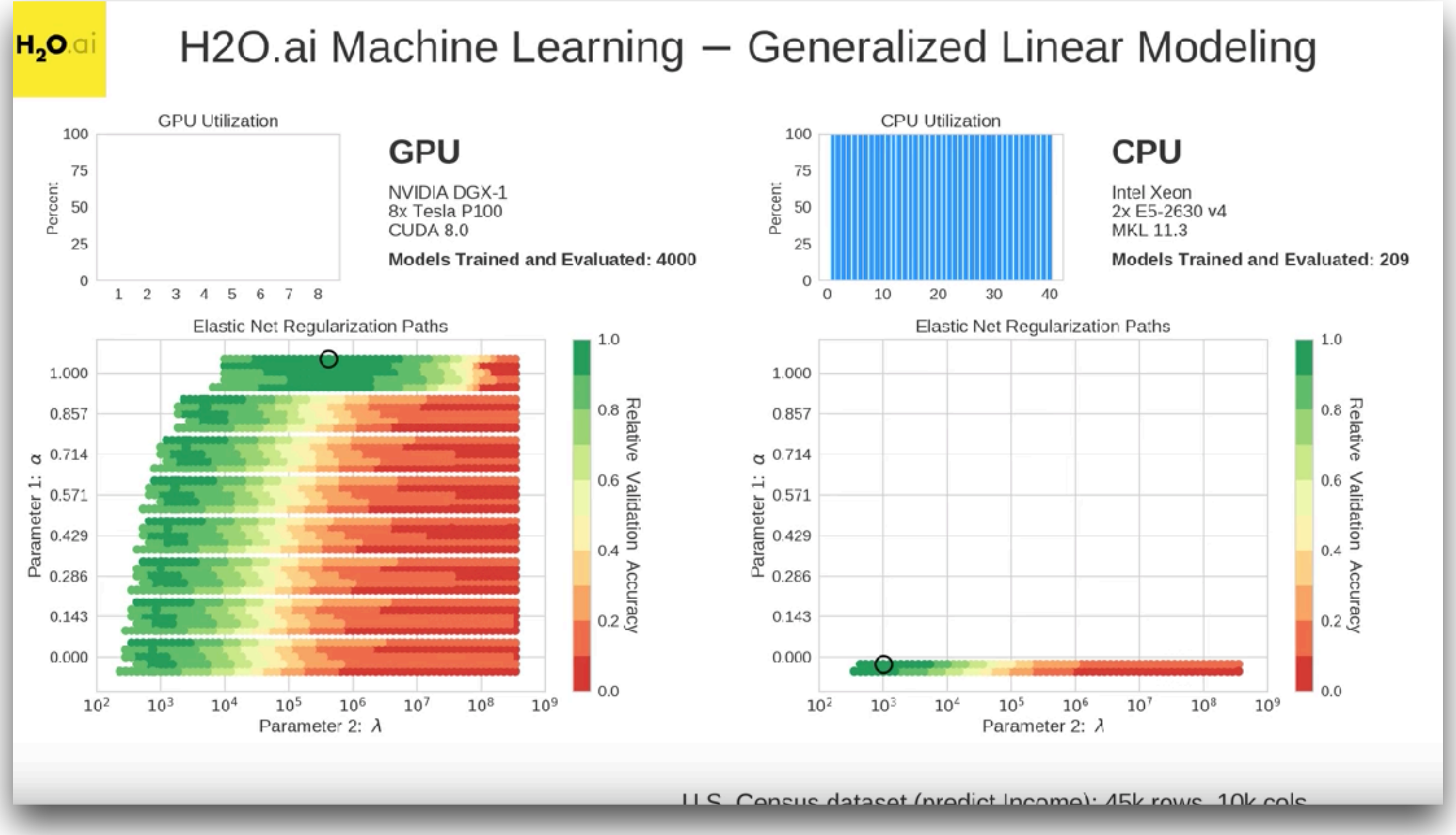
Speed

Accuracy

Interpretability

Speed

- GPU acceleration to achieve up to 40x speedups vs CPU
- Multi GPU - XGBoost, GLM, K-Means and more
- Achieve best performance in shortest time



H2O.ai Machine Learning – Generalized Linear Modeling

GPU

NVIDIA DGX-1
8x Tesla P100
CUDA 8.0

Models Trained and Evaluated: 4000

CPU

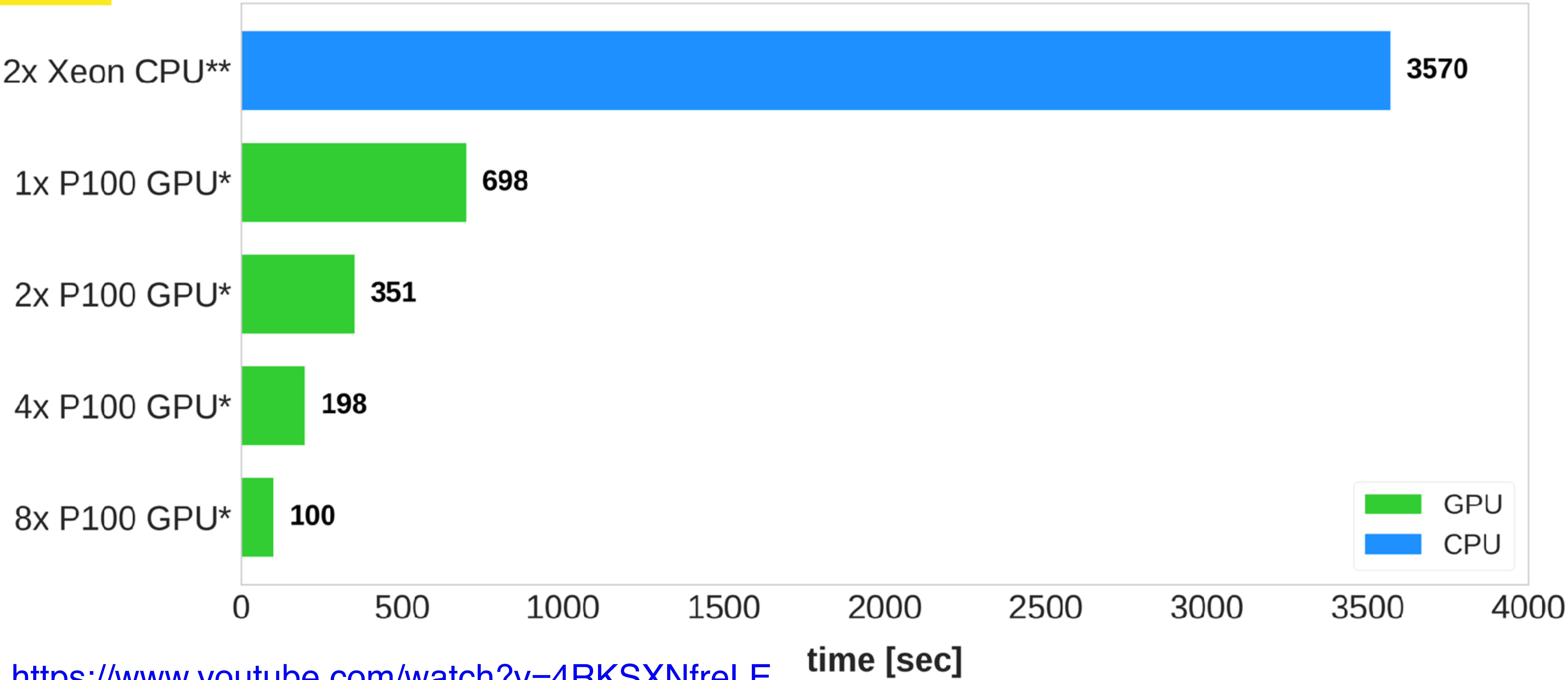
Intel Xeon
2x E5-2630 v4
MKL 11.3

Models Trained and Evaluated: 209

U.S. Census dataset (predict Income): 45k rows, 10k cols

H2O.ai Machine Learning – Generalized Linear Modeling

Time to Train and Evaluate 4000 Models



https://www.youtube.com/watch?v=4RKSXNfreLE

http://github.com/h2oai/perf/

*NVIDIA DGX-1, **Dual Intel Xeon E5-2630 v4

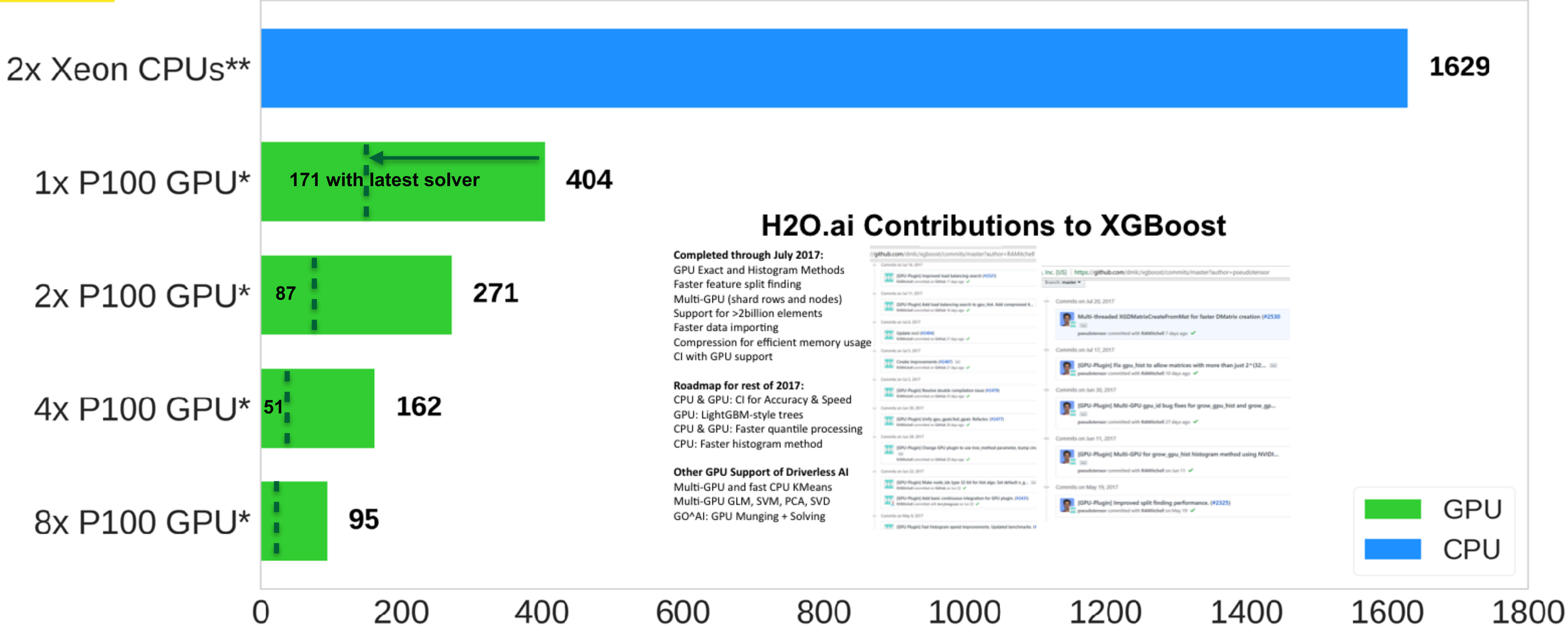
U.S. Census dataset (predict Income): 45k rows, 10k cols

Elastic Net Model Parameters: 5-fold cross-validation, $\alpha = \{\frac{i}{7}, i = 0 \ldots 7\}$, full λ-search

H2O.ai

H2O.ai Machine Learning – Gradient Boosting Machine

Time to Train 16 H2O XGBoost Models



H2O.ai Contributions to XGBoost

Completed through July 2017:
GPU Exact and Histogram Methods
Faster feature split finding
Multi-GPU (shard rows and nodes)
Support for >2billion elements
Faster data importing
Compression for efficient memory usage
CI with GPU support

Roadmap for rest of 2017:
CPU & GPU: CI for Accuracy & Speed
GPU: LightGBM-style trees
CPU & GPU: Faster quantile processing
CPU: Faster histogram method

Other GPU Support of Driverless AI
Multi-GPU and fast CPU KMeans
Multi-GPU GLM, SVM, PCA, SVD
GO^AI: GPU Munging + Solving

//github.com/dmlc/xgboost/commits/master?author=RAMitchell

Commits on Jul 16, 2017
[GPU-Plugin] Improved load balancing search (#2521)
RAMitchell committed on GitHub 11 days ago ✓

Commits on Jul 11, 2017
[GPU-Plugin] Add load balancing search to gpu_hist. Add compressed it...
RAMitchell committed on GitHub 16 days ago ✓

Commits on Jul 6, 2017
Update nccl (#2494)
RAMitchell committed on GitHub 21 days ago ✓

Commits on Jul 5, 2017
Cmake improvements (#2487)
RAMitchell committed on GitHub 21 days ago ✓

Commits on Jul 2, 2017
[GPU-Plugin] Resolve double compilation issue (#2479)
RAMitchell committed on GitHub 25 days ago ✓

Commits on Jun 30, 2017
[GPU-Plugin] Unify gpu_gpair/bst_gpair. Refactor. (#2477)
RAMitchell committed on GitHub 28 days ago ✓

Commits on Jun 28, 2017
[GPU-Plugin] Change GPU plugin to use tree_method parameter, bump cm...
RAMitchell committed on GitHub 29 days ago ✓

Commits on Jun 22, 2017
[GPU-Plugin] Make node_idx type 32 bit for hist algo. Set default n_g...
RAMitchell committed on GitHub on Jun 22 ✓
[GPU-Plugin] Add basic continuous integration for GPU plugin. (#2431)
RAMitchell committed with terrytangyuan on Jun 22 ✓

Commits on May 8, 2017
[GPU Plugin] Fast histogram speed improvements. Updated benchmarks. (#

, Inc. [US] | https://github.com/dmlc/xgboost/commits/master?author=pseudotensor
Branch: master

Commits on Jul 20, 2017
Multi-threaded XGDMatrixCreateFromMat for faster DMatrix creation (#2530
pseudotensor committed with RAMitchell 7 days ago ✓

Commits on Jul 17, 2017
[GPU-Plugin] Fix gpu_hist to allow matrices with more than just 2^(32...
pseudotensor committed with RAMitchell 10 days ago ✓

Commits on Jun 30, 2017
[GPU-Plugin] Multi-GPU gpu_id bug fixes for grow_gpu_hist and grow_gp...
pseudotensor committed with RAMitchell 27 days ago ✓

Commits on Jun 11, 2017
[GPU-Plugin] Multi-GPU for grow_gpu_hist histogram method using NVIDI...
pseudotensor committed with RAMitchell on Jun 11 ✓

Commits on May 19, 2017
[GPU-Plugin] Improved split finding performance. (#2325)
pseudotensor committed with RAMitchell on May 19 ✓

time [sec]

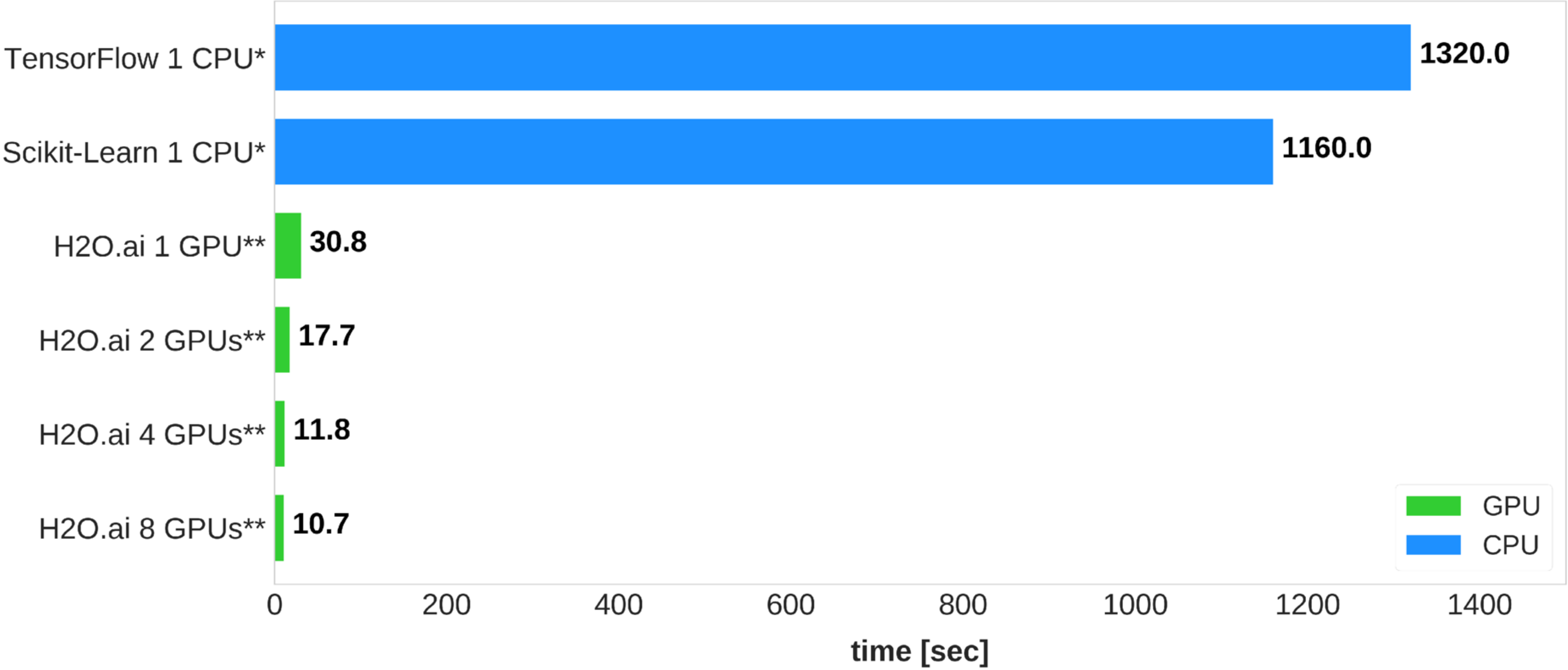
https://www.youtube.com/watch?v=NkeSDrifJdg

http://github.com/h2oai/perf/

*NVIDIA DGX-1, **Dual Intel Xeon E5-2630 v4
Higgs dataset (binary classification): 1M rows, 29 cols; max_depth: {6,8,10,12}, sample_rate: {0.7,0.8,0.9,1.0}

H2O.ai Machine Learning – k-Means Clustering

Time to run 1000 Lloyds iterations for k=1000 clusters



Kaggle Homesite Home Insurance Claims Predictions Dataset (261k rows, 298 cols)
k-Means Clustering (Lloyds), random initialization, 1000 centroids, 1000 iterations
Hardware: *Intel i7 5820k (6-core), **NVIDIA Tesla P100 (DGX-1)

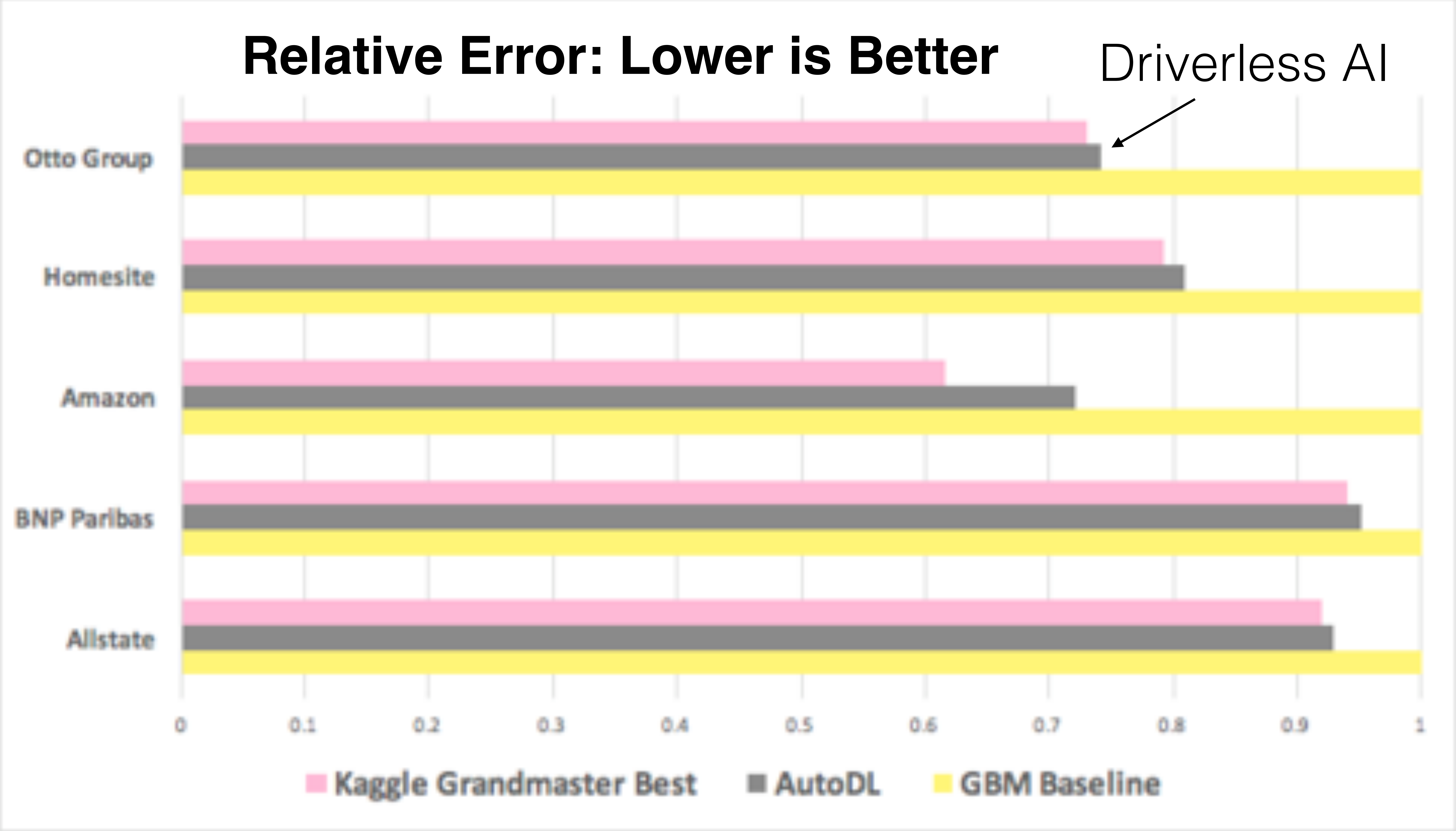
http://github.com/h2oai/perf/

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



Relative Error: Lower is Better

Driverless AI

Preliminary results - untuned, single model

What’s the “Secret” to Data Science?

“Coming up with features is difficult, time-consuming, requires expert knowledge. “Applied machine learning” is basically feature engineering.”

–Andrew Ng

“… some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.”

–Pedro Domingos

“Good data preparation and feature engineering is integral to better predictions.”

–Kazanova, H2O.ai Kaggle Grandmaster #2 (former #1)

Why is it so difficult to be a good Data Scientist?

Data matters: need access to business-relevant data and need domain knowledge about how features interact with each other

Data Science Recipes

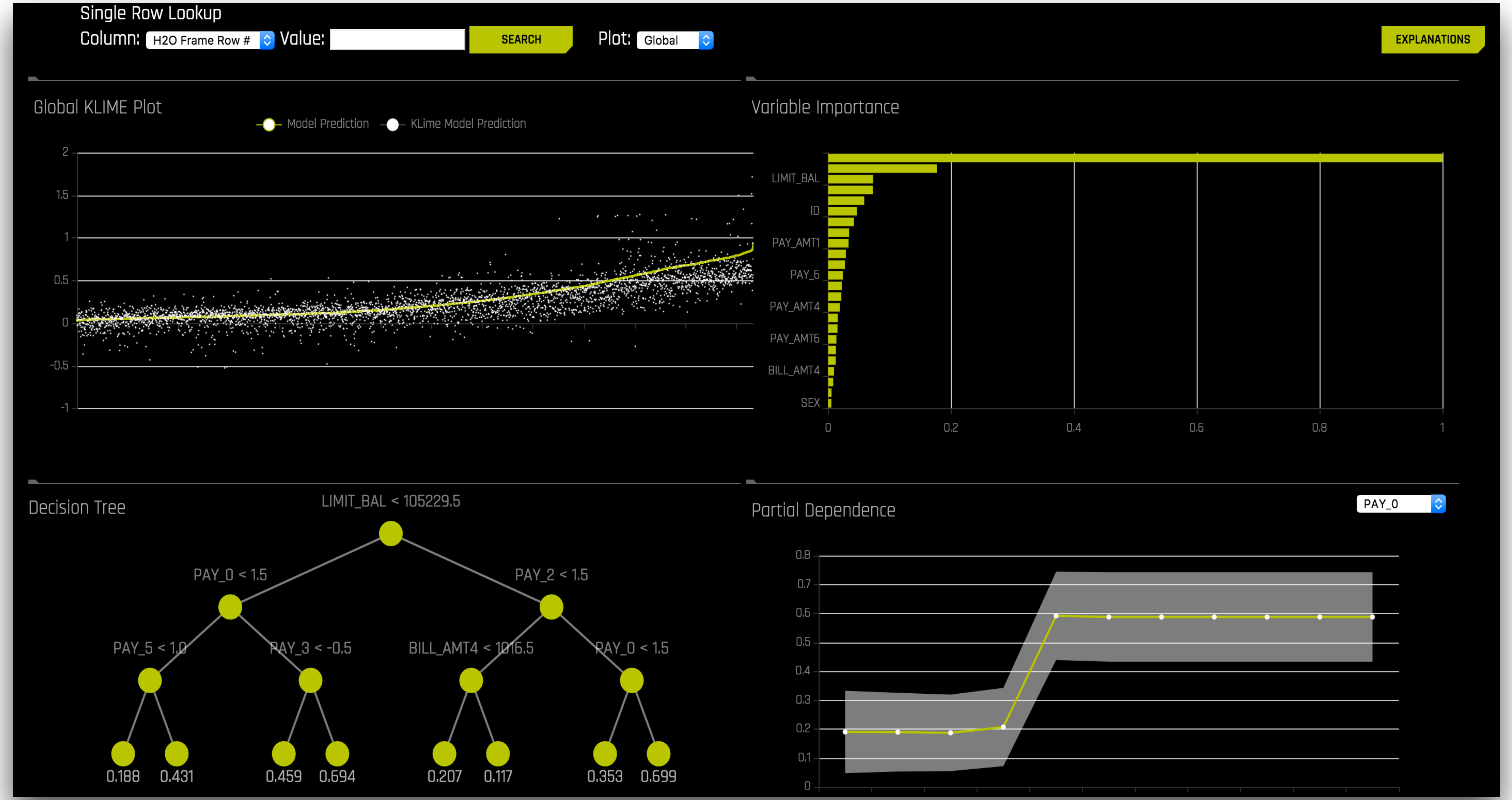
Powerful feature transformations (like target encoding) can lead to overfit models if done wrong, need strong math & statistics skills

Need to run thousands of experiments to reach robust conclusions, need computer science skills and access to compute hardware

GPUs, Automation

Interpretability

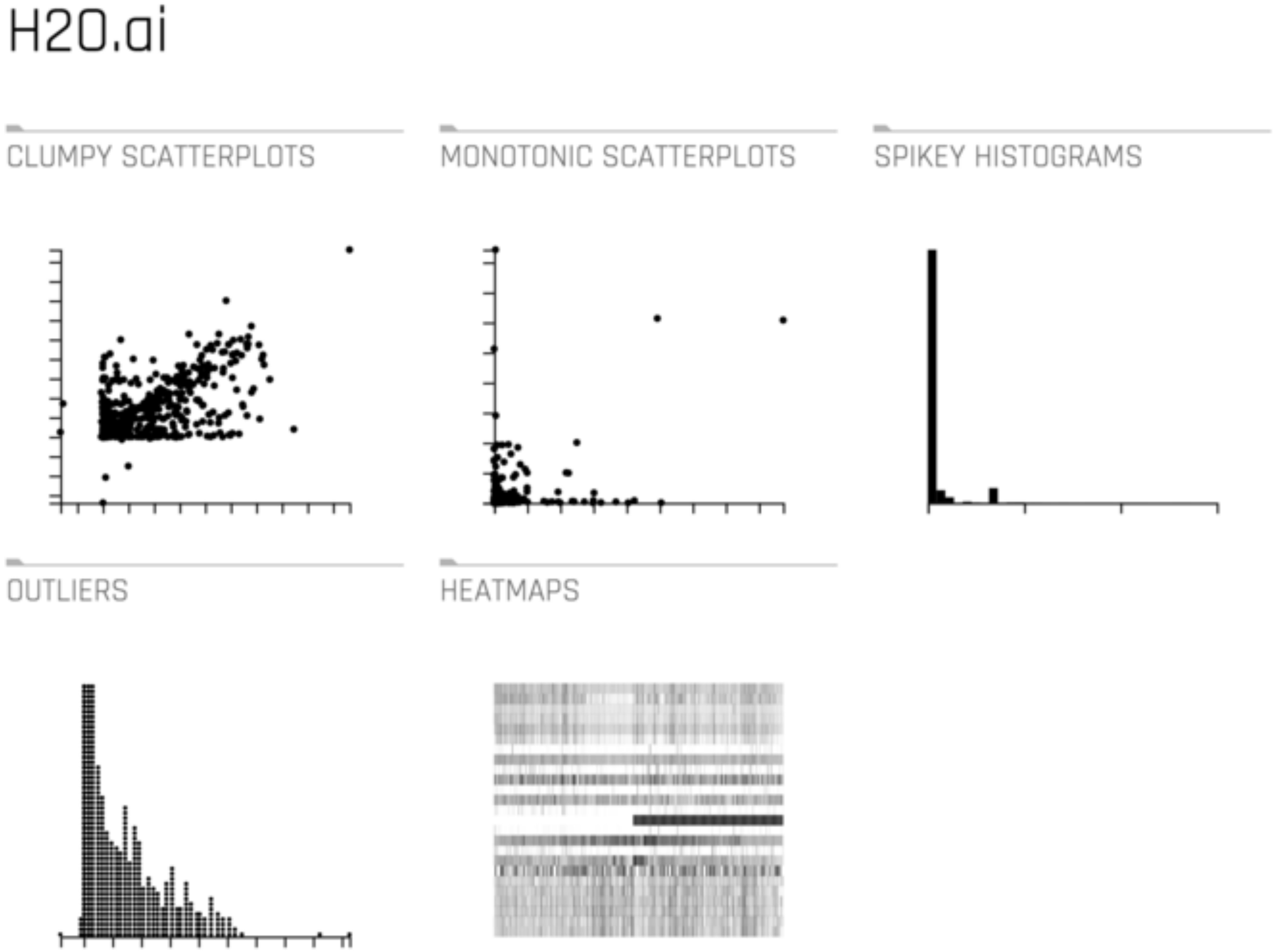
- Interpretability for debugging, not just for regulators
- Get reason codes and model interpretability in plain english
- K-Lime, LOCO, partial dependence and more



Single Row Lookup

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

Automatic Visualization



H2O.ai

CLUMPY SCATTERPLOTS

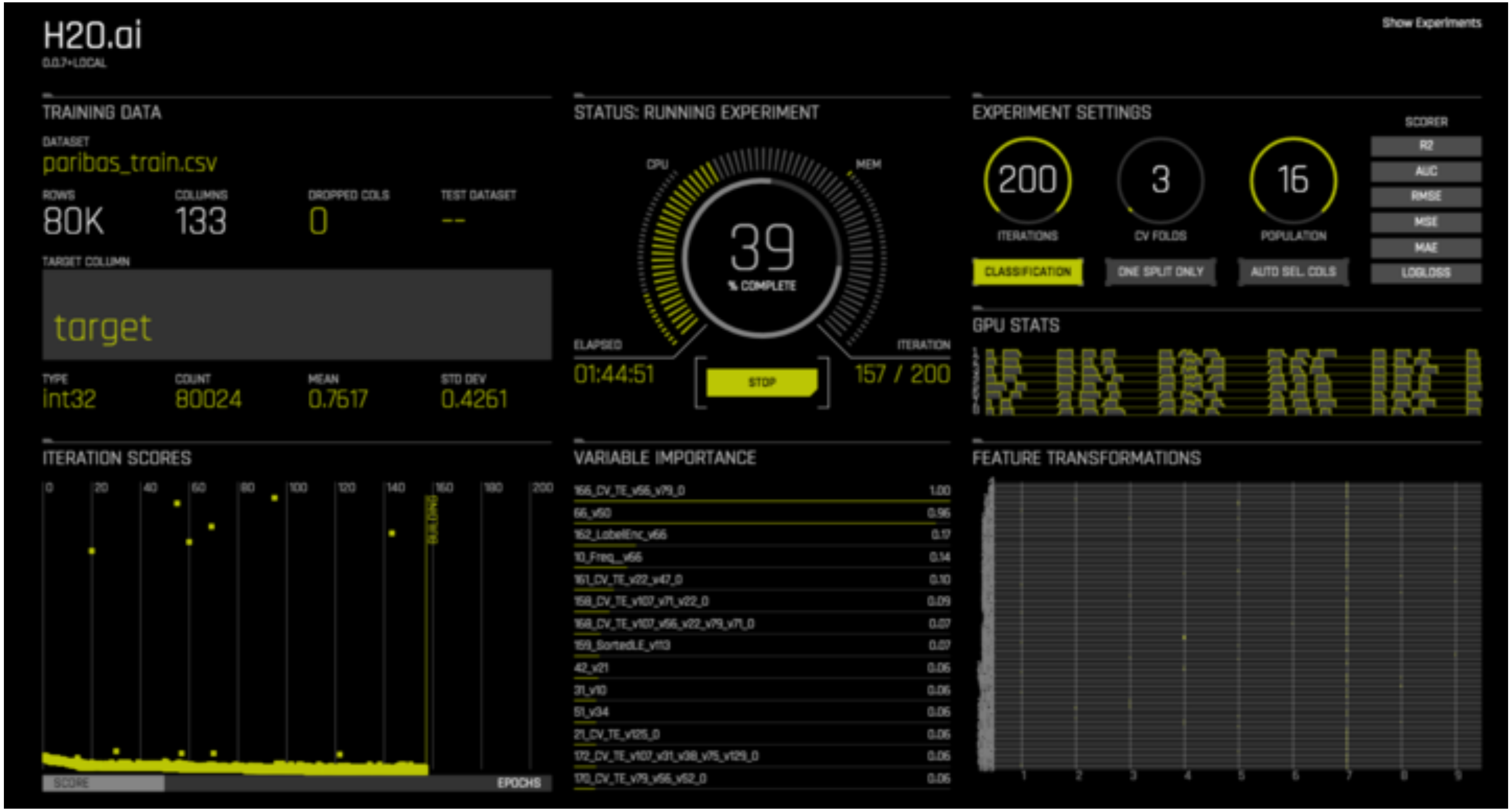
MONOTONIC SCATTERPLOTS

SPIKEY HISTOGRAMS

OUTLIERS

HEATMAPS

Next: Live Demo!



TRAINING DATA

DATASET paribas_train.csv

ROWS 80K · COLUMNS 133 · DROPPED COLS 0 · TEST DATASET --

TARGET COLUMN target

TYPE int32 · COUNT 80024 · MEAN 0.7617 · STD DEV 0.4261

STATUS: RUNNING EXPERIMENT

ELAPSED 01:44:51

ITERATION 157 / 200

EXPERIMENT SETTINGS

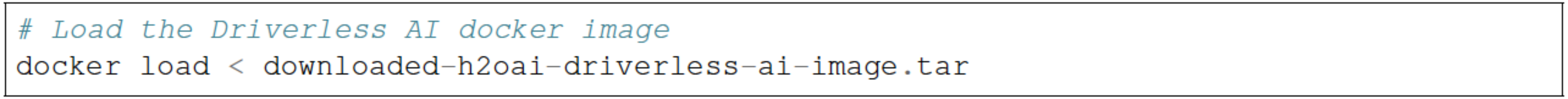
200 ITERATIONS · 3 CV FOLDS · 16 POPULATION

ITERATION SCORES

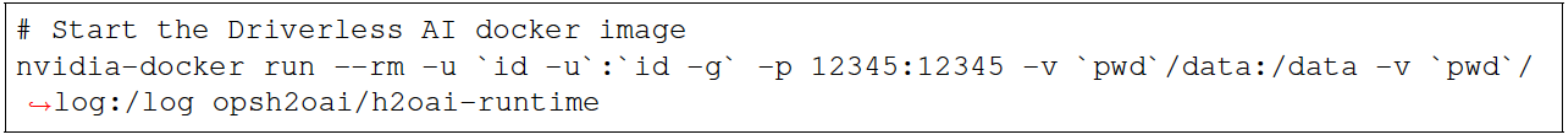
VARIABLE IMPORTANCE

166_CV_TE_v56_v79_0	1.00
66_v50	0.96
162_LabelEnc_v66	0.17
10_Freq__v66	0.14
161_CV_TE_v22_v47_0	0.10
158_CV_TE_v107_v71_v22_0	0.09
168_CV_TE_v107_v56_v22_v79_v71_0	0.07
159_SortedLE_v113	0.07
42_v21	0.06
31_v10	0.06
51_v34	0.06
21_CV_TE_v125_0	0.06
172_CV_TE_v107_v31_v38_v75_v129_0	0.06
170_CV_TE_v79_v56_v62_0	0.06

Installation and Start



```
# Load the Driverless AI docker image
docker load < downloaded-h2oai-driverless-ai-image.tar
```



```
# Start the Driverless AI docker image
nvidia-docker run --rm -u `id -u`:`id -g` -p 12345:12345 -v `pwd`/data:/data -v `pwd`/
 ↪log:/log opsh2oai/h2oai-runtime
```

The Team Behind Driverless AI

Marios Michailidis

Dmitry Larko

Branden Murray

Mark Landry

AutoDL

Matt Dowle

Pasha Stetsenko

Datatable

Patrick Hall

Mark Chan

Navdeep Gill

MLI

Arno Candel

Jon Mckinney

Rory Mitchell

GPU Acceleration

Leland Wilkinson

Prithvi Prabhu

Justin Loyola

AutoViz

H2O.ai

Upcoming Webinars

- Machine Learning Interpretability - Patrick Hall - **Aug 17**
- AutoDL - Dmitry Larko - **Aug 24**
- Automatic Visualization - **Sep (TBD)**