DRIVERLESS AI

Introduction to AutoML and Machine Learning Interpretability

Jan Gamec

H2O.ai

Jan GAMEC

@JanGamec

Senior SW/ML Engineer

- Violin enthusiast
- I like hiking, climbing and generally sporting
- I like discovering new (sometimes also old) technologies

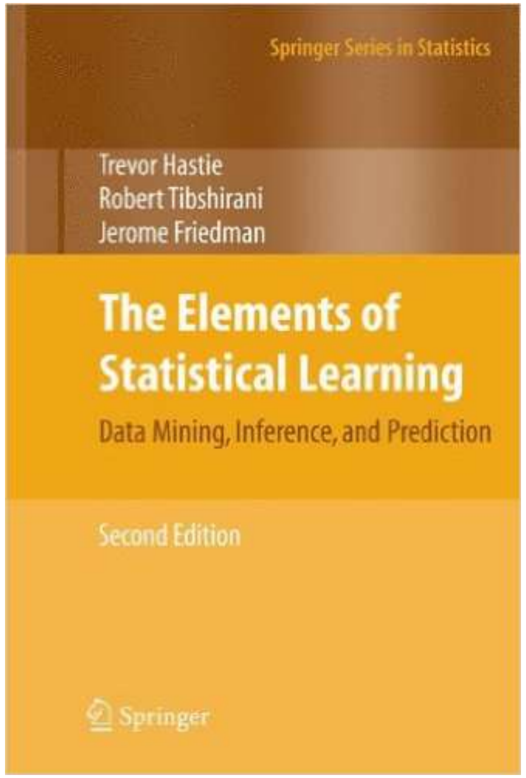
AGENDA

- H2O.ai – company and people
- Intro to DriverlessAI
- Feature engineering with DriverlessAI
- Machine Learning Interpretability with DriverlessAI
- DEMO

Company Overview

Founded	2012, Series C in Nov, 2017
Products	• Driverless AI – Automated Machine Learning • H_2O - Open Source Machine Learning Platform • H2O4GPU - Lightning Fast Machine Learning on GPUs • Sparkling Water - Integration of H_2O and Apache Spark
Mission	Democratize AI. Do Good.
Team	~130 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Offices	Mountain View, London, Prague

Scientific Advisory Council



Dr. Trevor Hastie

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

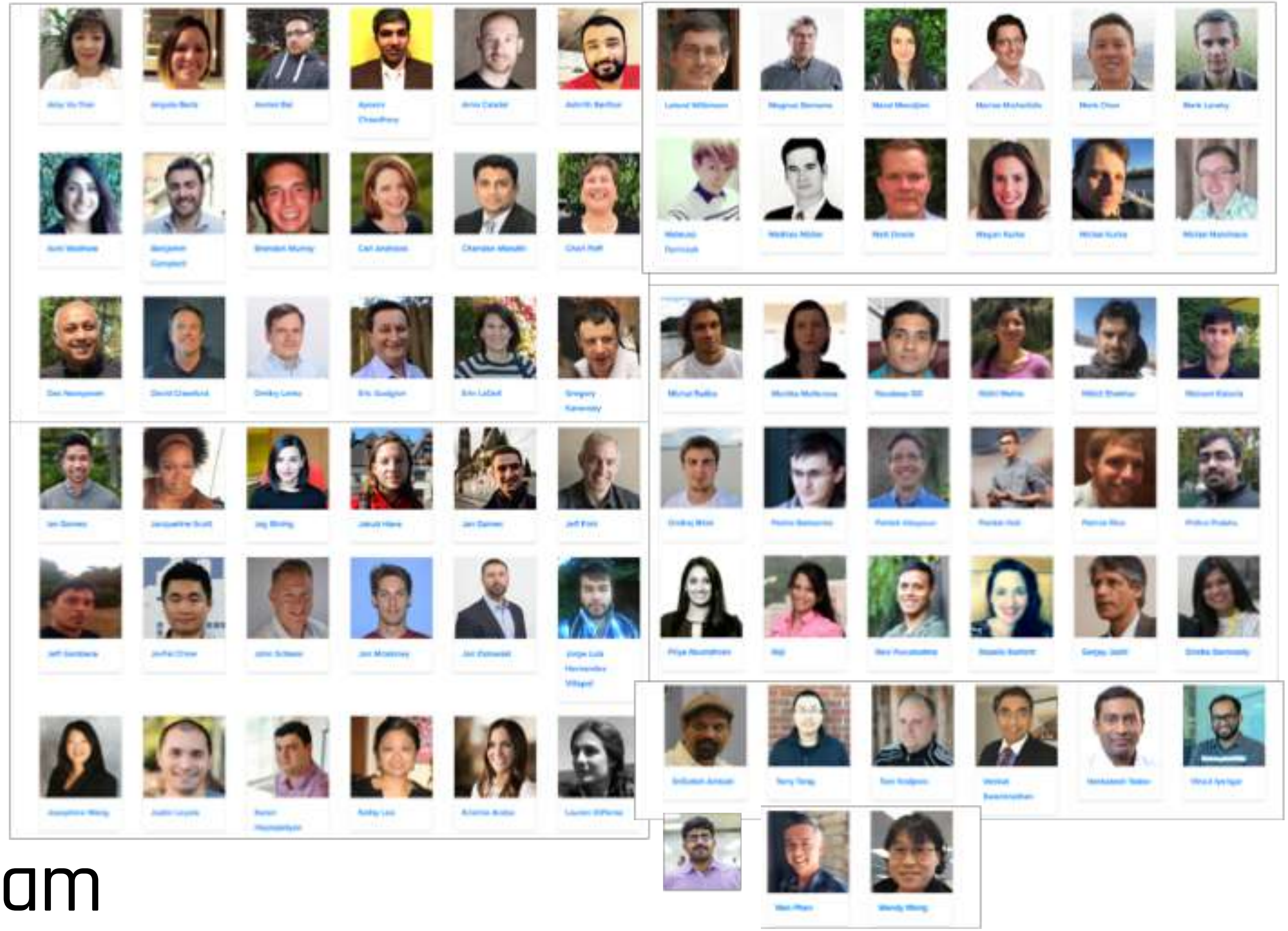
Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*

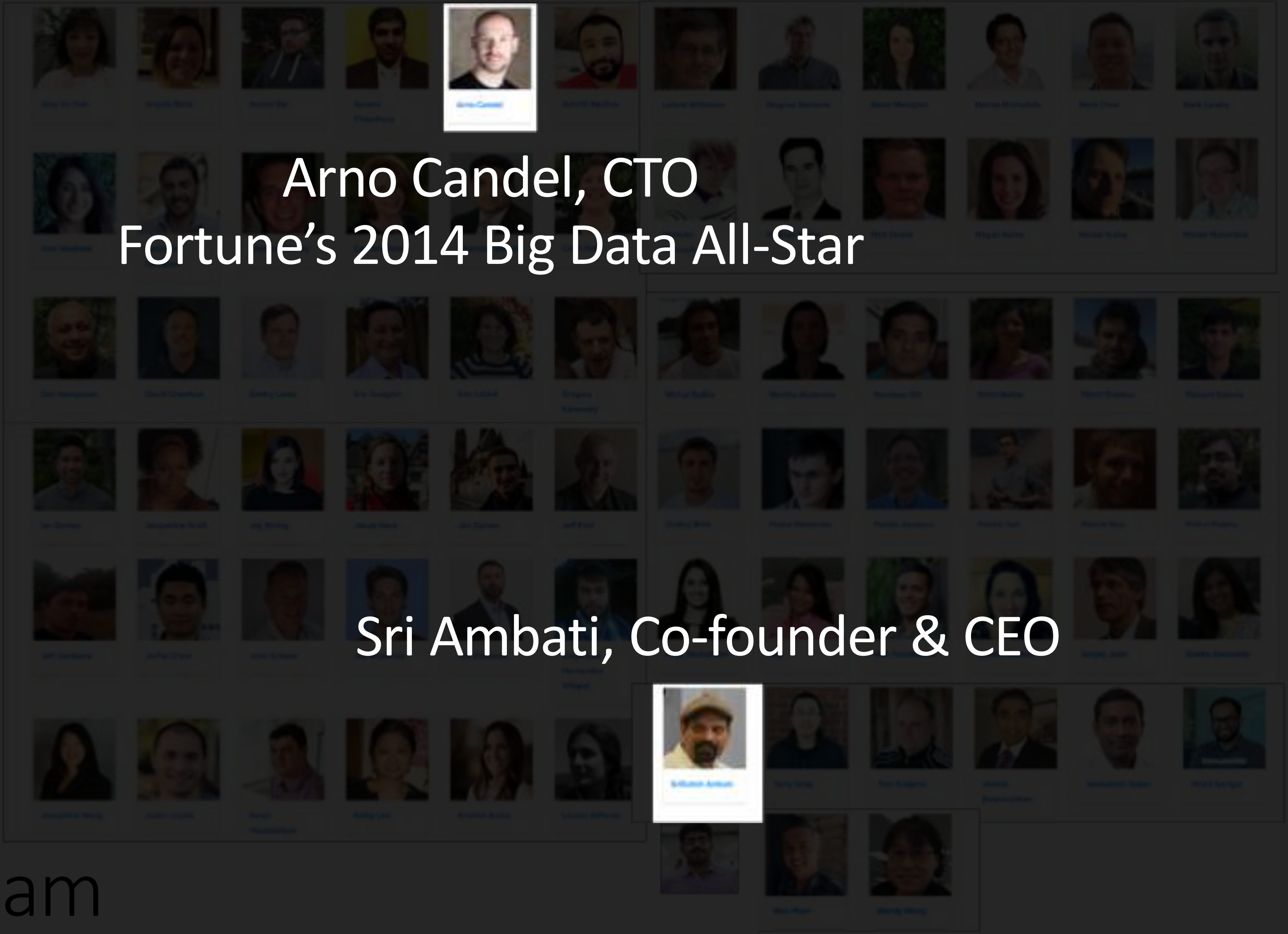
Dr. Steven Boyd

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

H2O.ai



H_2O Team



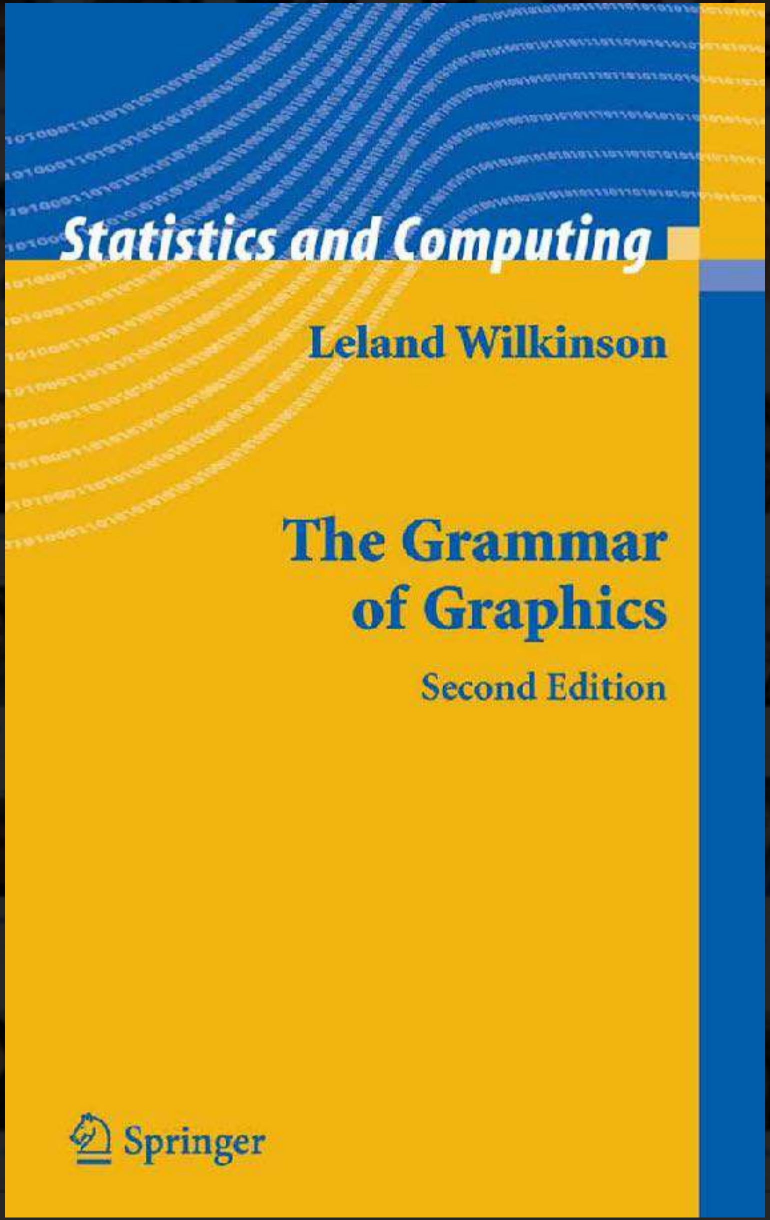
Arno Candel, CTO

Fortune’s 2014 Big Data All-Star

Sri Ambati, Co-founder & CEO

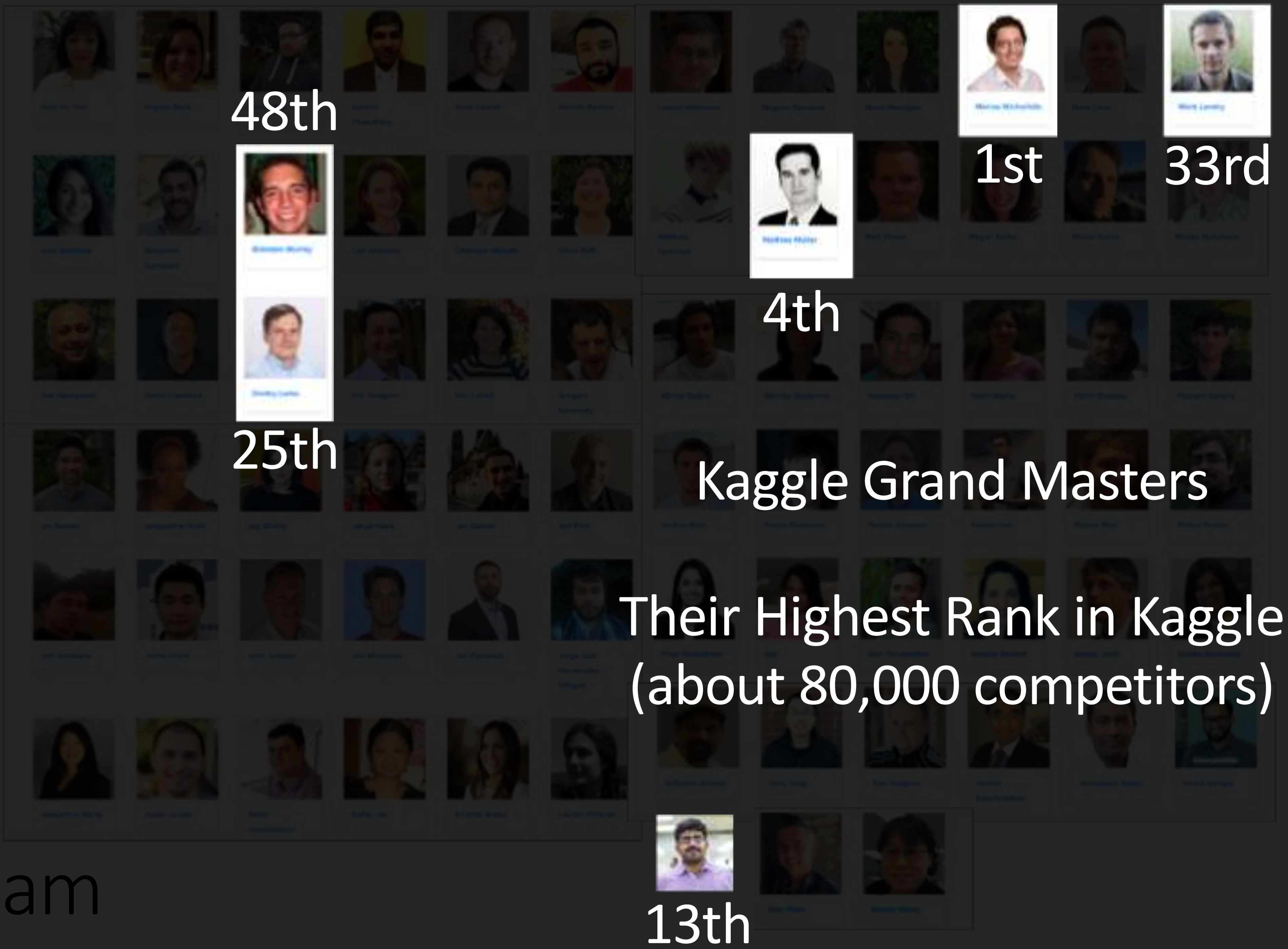
H_2O Team

Origin of R Package `ggplot2`



ggplot2

H_2O Team


48th
1st
33rd
4th
25th
Kaggle Grand Masters
Their Highest Rank in Kaggle
(about 80,000 competitors)
13th

H_2O Team

H_2O Products

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

H2O AI Open Source Engine Integration with Spark

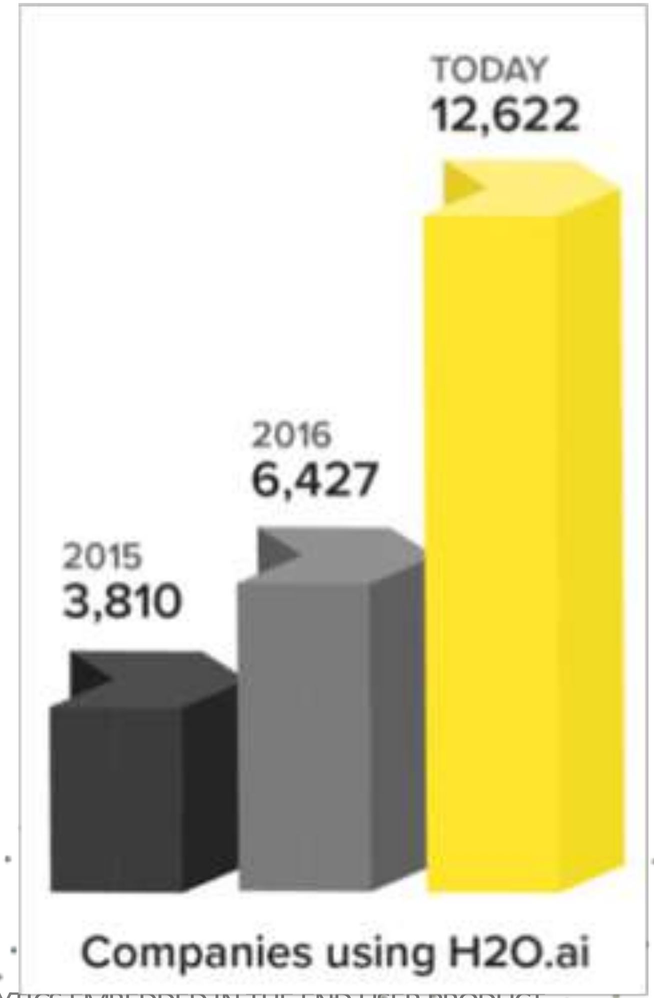
Lightning Fast machine learning on GPUs

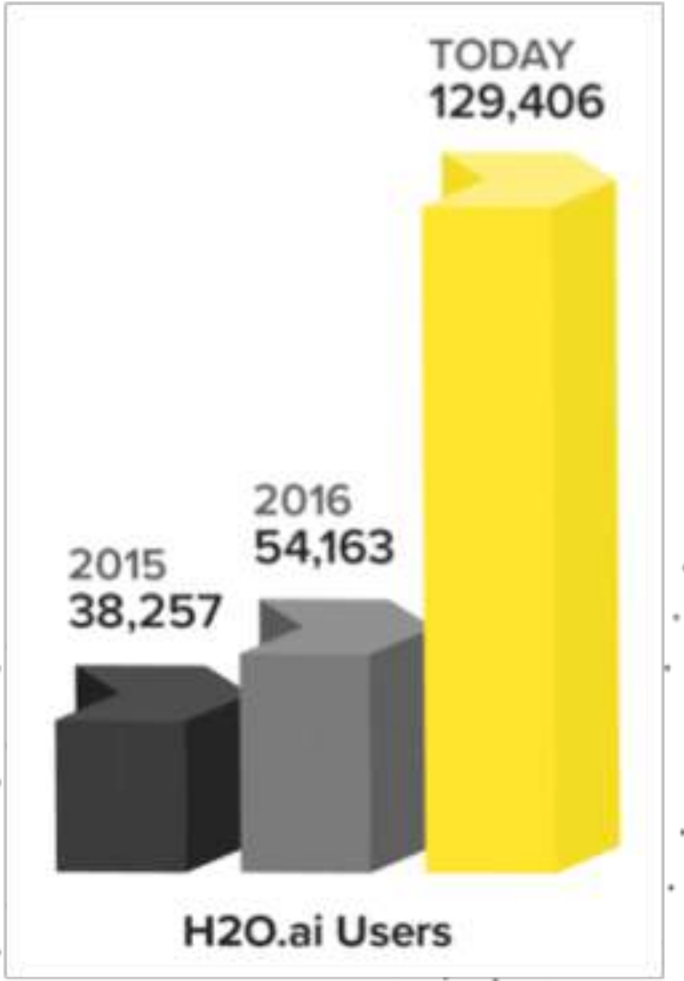
Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

Worldwide Community Adoption





* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H2O.ai Solution Leadership Across Verticals

Retail
- Travelport
- Walgreens
- ebay
- Booking.com
- macy's

Healthcare
- CHANGE HEALTHCARE
- KAISER PERMANENTE
- aetna
- CONFIDENTIAL

Marketing
- acxiom
- CATALINA
- Integral Ad Science
- MarketShare
- Nielsen Catalina Solutions

Financial
- citi
- deserve
- DISCOVER
- experian
- Capital One
- EQUIFAX
- MarketAxess
- WELLS FARGO
- RBC
- PayPal

Advisory & Accounting
- ADP
- pwc
- starling
- !underwrite.ai

Insurance
- State Farm
- AEGON
- XCEEDANCE
- TRANSAMERICA
- PROGRESSIVE
- ZURICH

Telecom
- COMCAST
- CONFIDENTIAL
- CISCO

H2O.ai

DRIVERLESS AI

Driverless AI Overview

H2O.ai

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

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H2O4GPU

Lightning Fast machine learning on GPUs

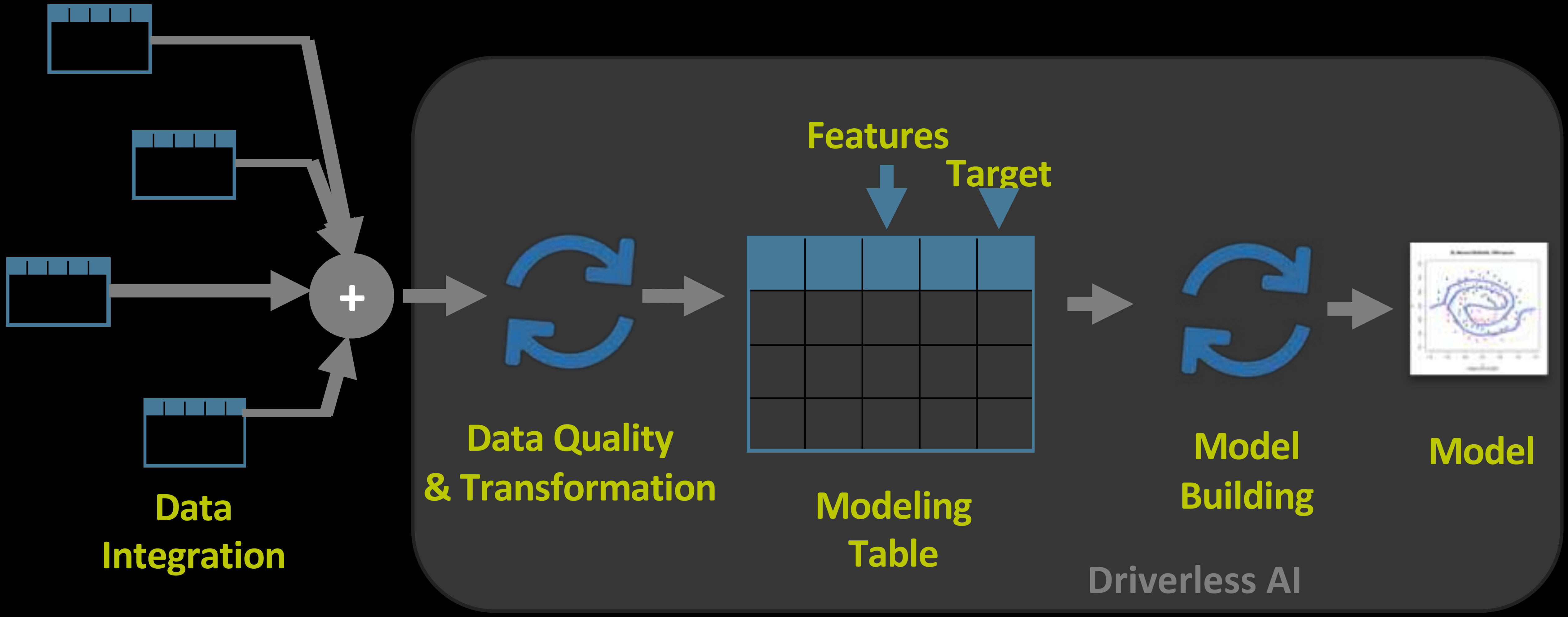
DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

Typical Enterprise Machine Learning Workflow



DRIVERLESSAI Workflow

Data Processing

Model Tuning

Feature Engineering

Final Model Training

Scoring Pipeline

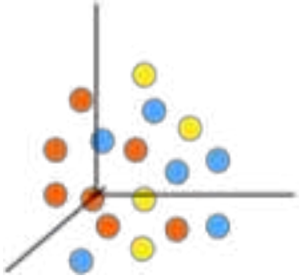
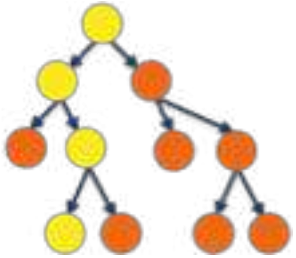
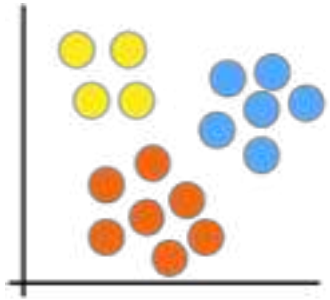
3 Pillars of Driverless AI

Speed

Accuracy

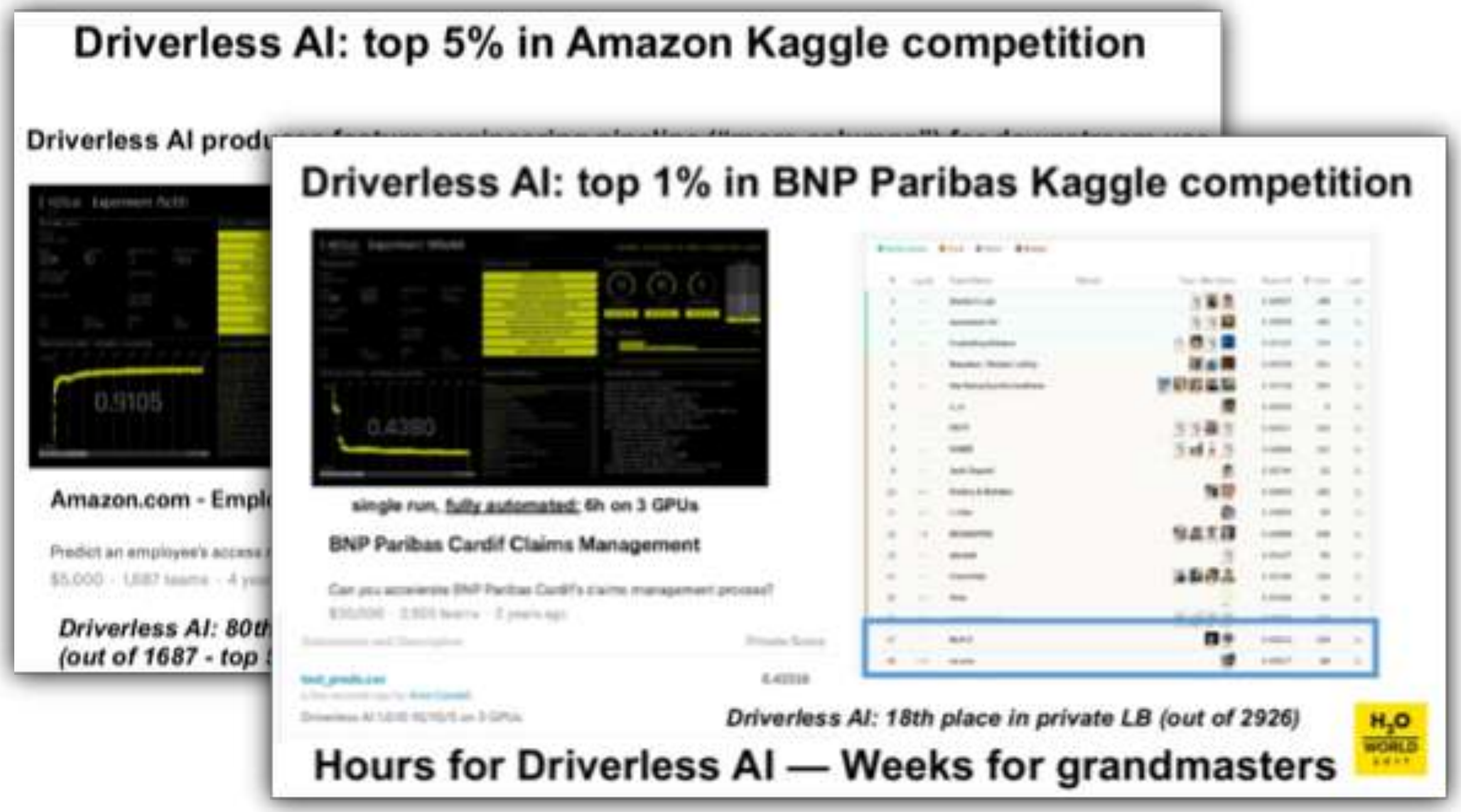
Interpretability

Speed == H2O4GPU Algorithms

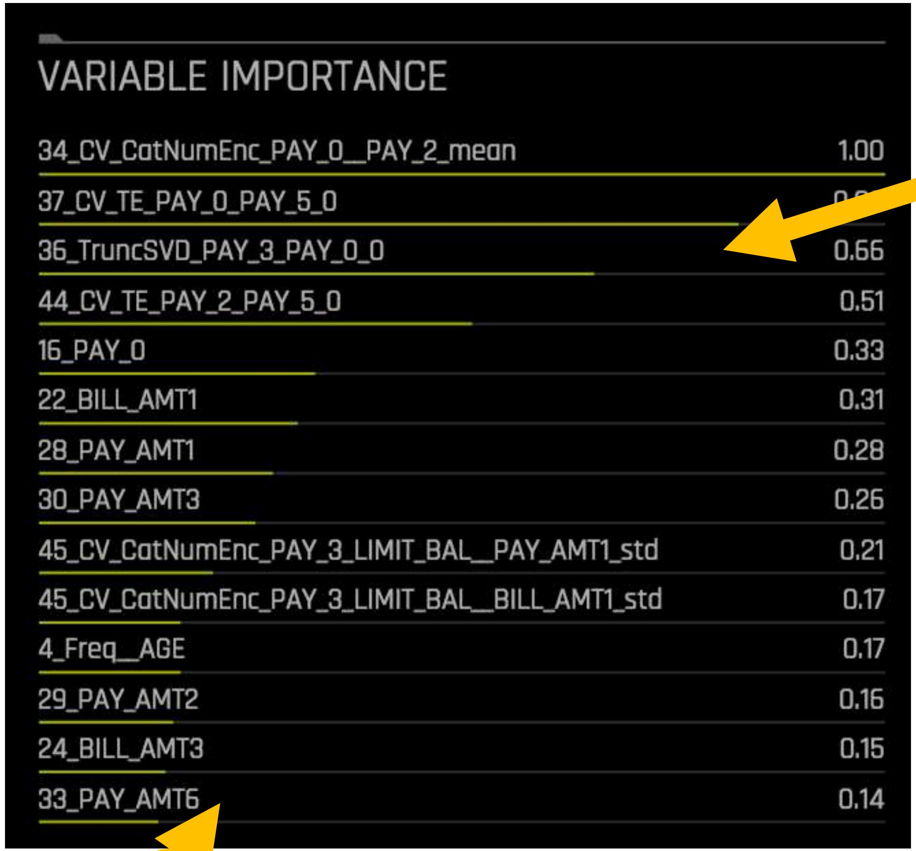
5X	5X	10X	40X
GLM	SVD	XGBoost	K-means
			

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



Auto Feature Generation
Kaggle Grand Master Out of the Box



Generated Features

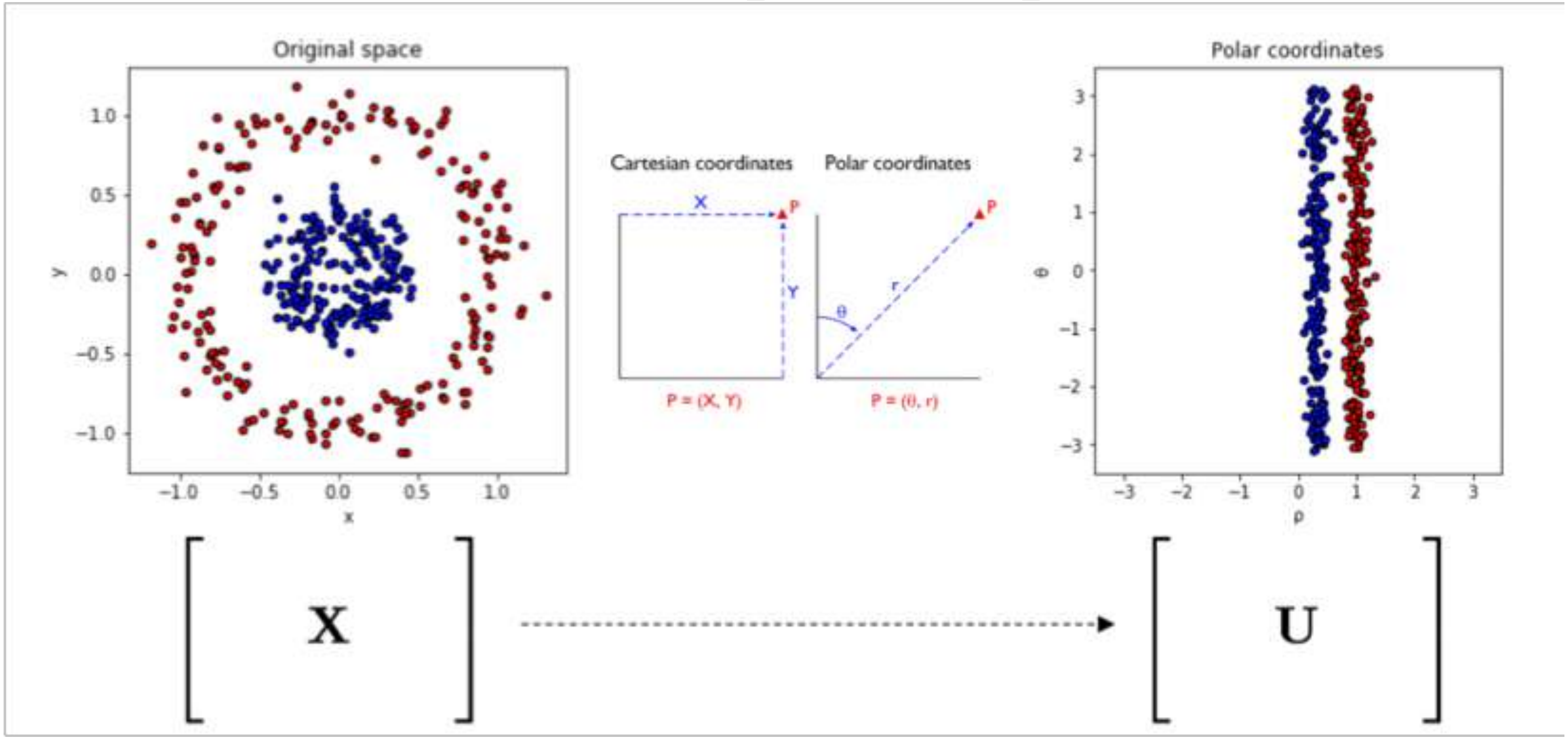
Original Features

Feature Transformations

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

H2O.ai

Automatic Feature Engineering



Why feature engineering is hard

- Feature transformation (like target encoding) can introduce leakage when applied wrong
- Usually requires domain knowledge
- Time consuming

Key elements of feature engineering

Target transformations:

- Fix skewed distributions and improve model fit
- *Log(y),Sqrt(y)*,…

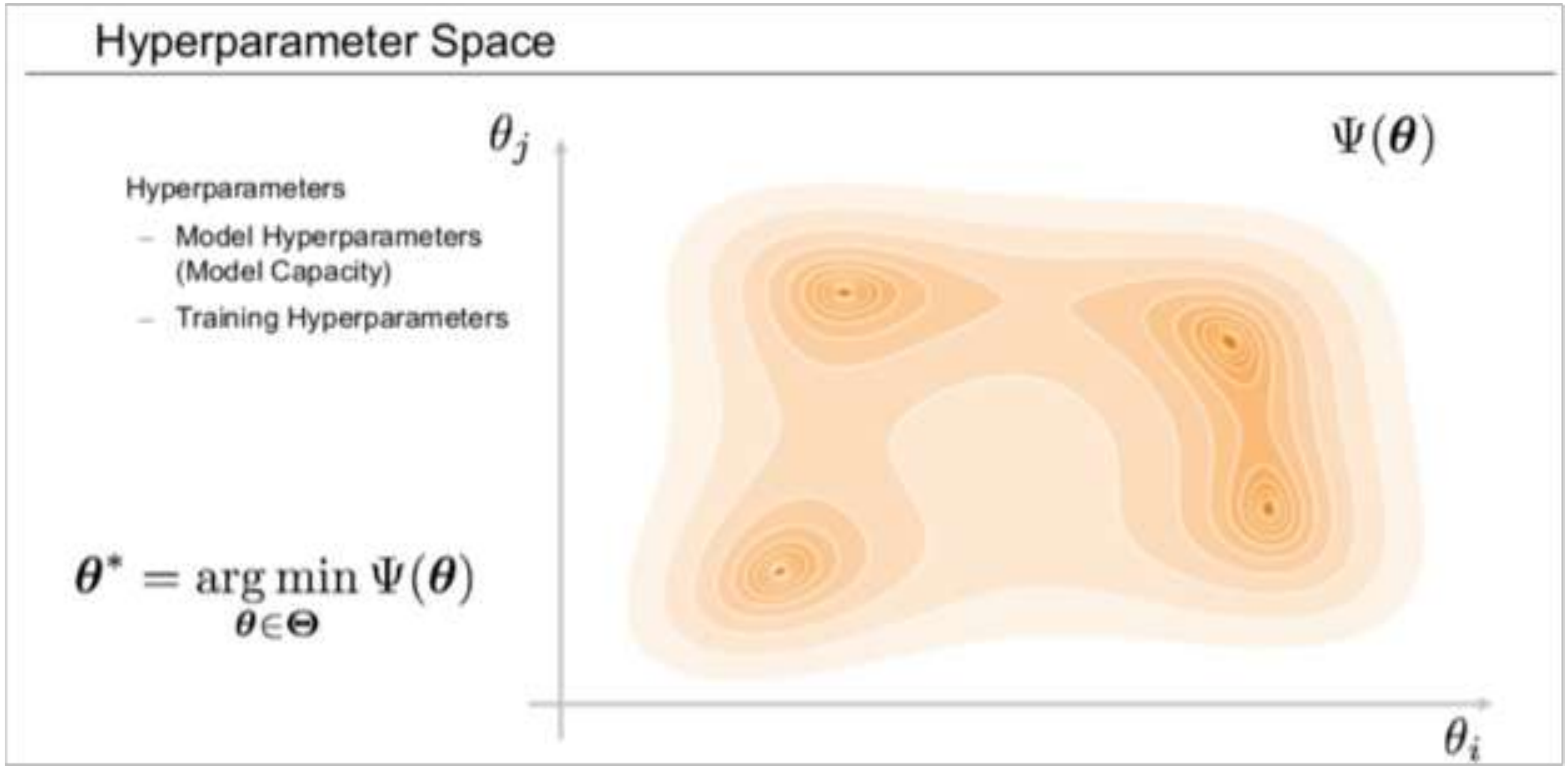
Label encoding:

- Turn categoricals into numericals (most ML tools work with numbers as their input)
- *One-hot encoding*, *Frequency encoding*, *Target mean encoding*
- Binning of the numericals
- Dimensionality reduction (*PCA*, *SVD*) and clustering (*Dist. To cluster mean*)

Feature Interactions & Extraction:

- $y = x_1^2 + x_2^2 - x_1 + 1$ => Adding x_1^2 and x_2^2 will improve model performance

Automatic finding and tuning of models



- Genetic algorithm based approach
- Grid search

Machine Learning Interpretability

Complexity of learned functions:

- Linear, monotonic
- Nonlinear, monotonic
- Nonlinear, non-monotonic

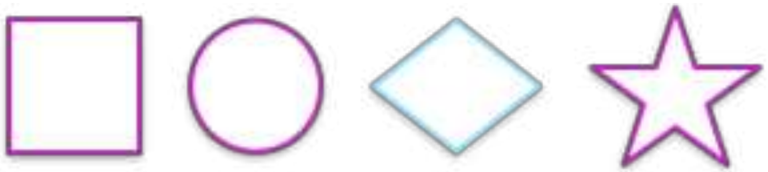
Scope of interpretability:

Global vs. local

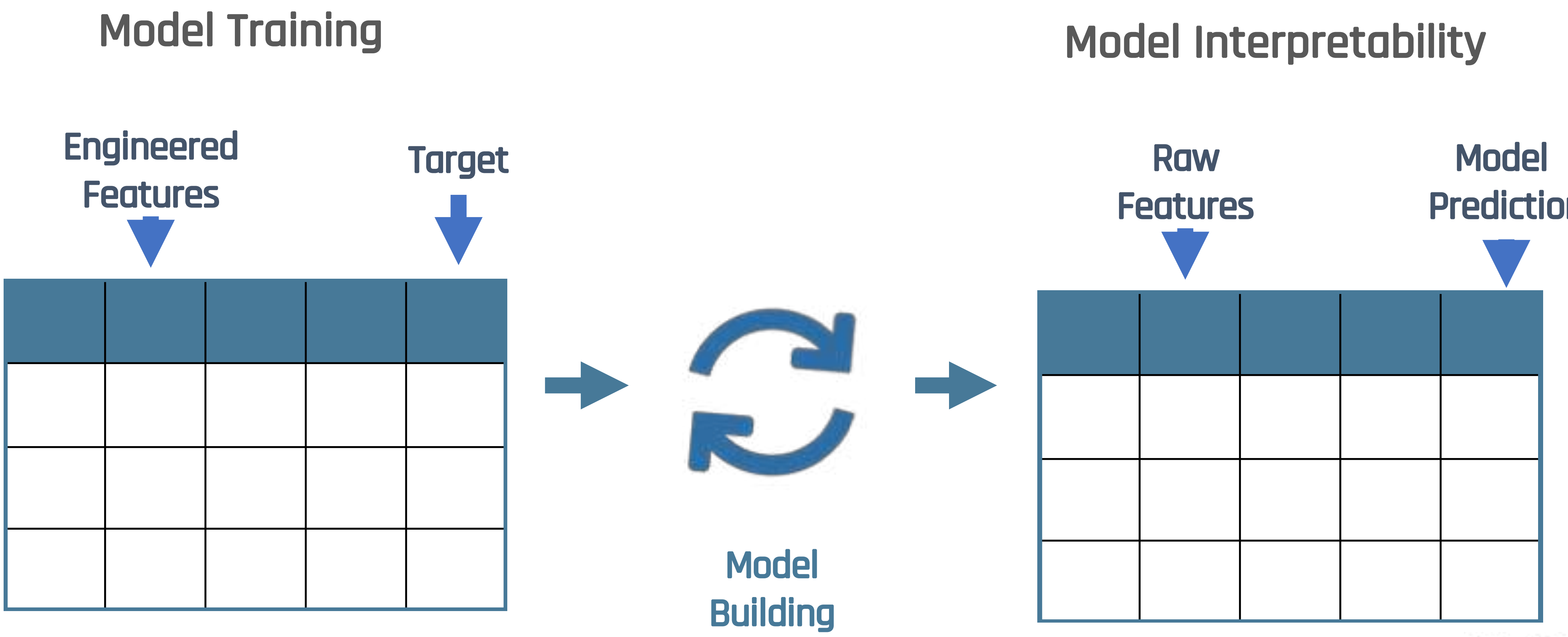
Enhancing trust and understanding: the mechanisms and results of an interpretable model should be both transparent AND dependable.

Application domain:

Model-agnostic vs. model-specific



How interpretability works?!



H2O.ai

The Solution

DEFAULT	TE_PAY_1	LIMIT_BAL	TE_EDUCATION
YES	0.53	$20,000	0.32
NO	0.19	$90,000	0.32
NO	0.15	$50,000	0.19

1. Train a complex machine learning model on generated features

Prob DEFAULT	PAY_1	LIMIT_BAL	EDUCATION
81%	Missed 2 Mo	$20,000	university
32%	Up to Date	$90,000	university
21%	Up to Date	$50,000	graduate

2. Train a complex machine learning model on raw features and the predicted target values of our original model

DEMO

Credit Card Default Dataset

- **USE CASE:**
  - Probability of default for Credit Card Loans

- **FEATURES**
  - **Default payment next month** (True/False)
  - LIMIT_BAL - Credit limit in dollars
  - EDUCATION, SEX, MARRIAGE, AGE
  - PAY_0 - Was a payment received in current month?
  - PAY_2 - Was a payment received 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
  - …

WALMART weekly sales

- **USE CASE:**
  - Predict weekly sales of Walmart stores/departments
- **FEATURES**
  - **Weekly Sales**
  - STORE, DEPT – ID of store/department
  - DATE – Time column
  - Temperature, Fuel Cost – Characterize the region
  - Markdown 1-5 – promotions in stores
  - isHoliday – Was the holiday on the date
  - …

Q?

10X