

# Automatic and Interpretable Machine Learning in R with H<sub>2</sub>O and LIME



#ODSC<sup>o</sup>

Jo-fai (Joe) Chow

Data Science Evangelist /  
Community Manager

joe@h2o.ai

@matlabulous

Download → [https://bit.ly/  
\*\*odsc2018\\_h2o\*\*](https://bit.ly/odsc2018_h2o)

# About Me

London  
12<sup>th</sup> Sep



Edinburgh  
18<sup>th</sup> Sep



Paris 19<sup>th</sup> Sep

- **Before H<sub>2</sub>O**

- Water Engineer / EngD Researcher / Matlab Fan Boy  
(wonder why  @matlabulous?)
- Discovered R, Python, H<sub>2</sub>O ... never look back again
- Data Scientist at Virgin Media (UK), Domino Data Lab (US)

- **At H<sub>2</sub>O ...**

- Data Scientist / Evangelist /
- Sales Engineer / Solution Architect /
- Community Manager  
... The harsh reality of startup life ...

**Reminder: #360Selfie**

# Agenda

Time	Topics / Tasks
9:30 – 9:35 am	Install R packages from CRAN slides/code: <a href="https://bit.ly/odsc2018_h2o">bit.ly/odsc2018_h2o</a>
9:35 – 9:50 am	Introduction ( $H_2O$ , AutoML, LIME)
9:50 – 10:15 am	Worked Examples
10:15 – 10:25 pm	Real-World Use Case: Moneyball
10:25 – 10:30 am	Other $H_2O$ News + Q & A





#ODSC

# Why?

- Most users/organizations can benefit from **automatic machine learning pipelines**.
  - Eliminate time wasted on repetitive tasks, human errors, debugging etc.
- GDPR mandates a “right to explanation” from machine learning models.
  - **model interpretations** are crucial for those who must explain their models to regulators or customers.

# You will learn ...

- How to build high quality **H<sub>2</sub>O** models (almost) automatically.
- How to explain predictions from complex **H<sub>2</sub>O** models with **LIME**.
- **Bonus:** A real use case that led to a **multimillion-dollar** baseball decision earlier this year.

# $H_2O$ AutoML + LIME for REAL !!!

led to the signing of a  
Major League Baseball (MLB) player

**\$20M**

**multi-year contract**

finalised two weeks  
before the regular season





#ODSC

Time	Topics / Tasks
9:30 – 9:35 pm	Install R packages from CRAN slides/code: <a href="https://bit.ly/odsc2018_h2o">bit.ly/odsc2018_h2o</a>

# LIME

**Reference:** <https://github.com/thomasp85/lime>

```
# Install 'lime' from CRAN
install.packages('lime')
```

# mlbench for datasets

```
install.packages('mlbench')
```

# H2O

**Reference:** <https://www.h2o.ai/download/>

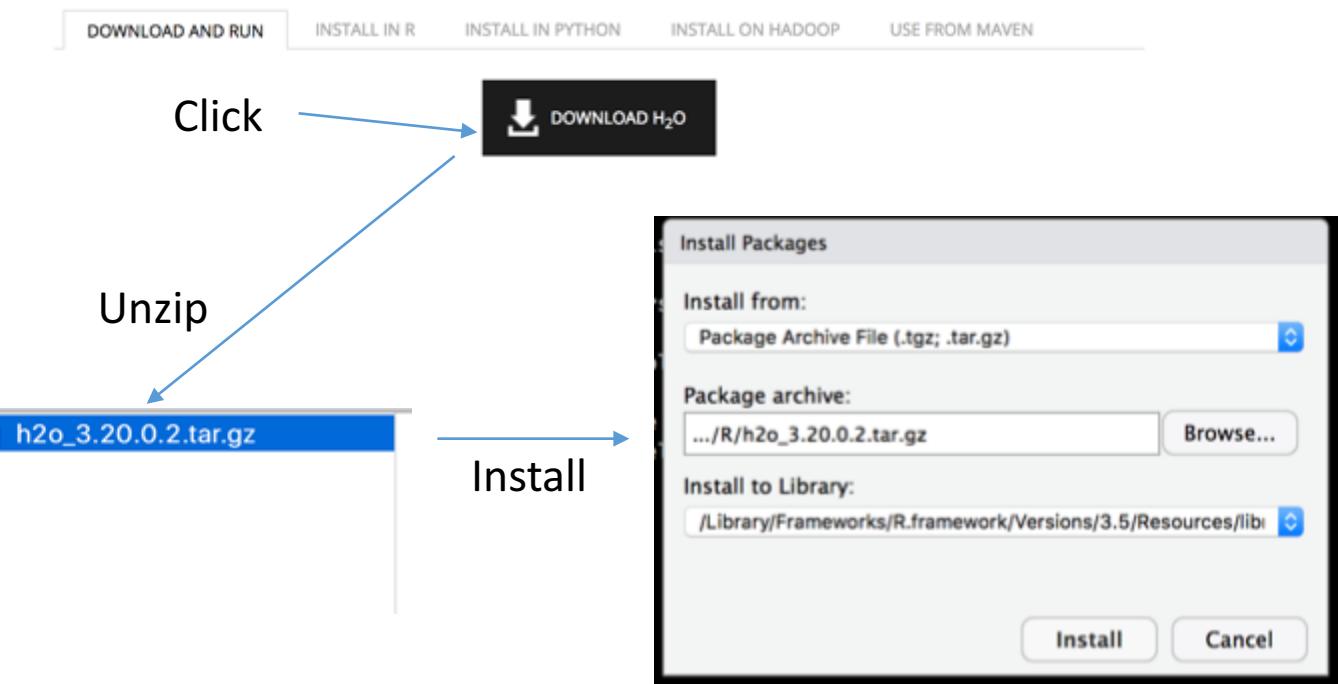
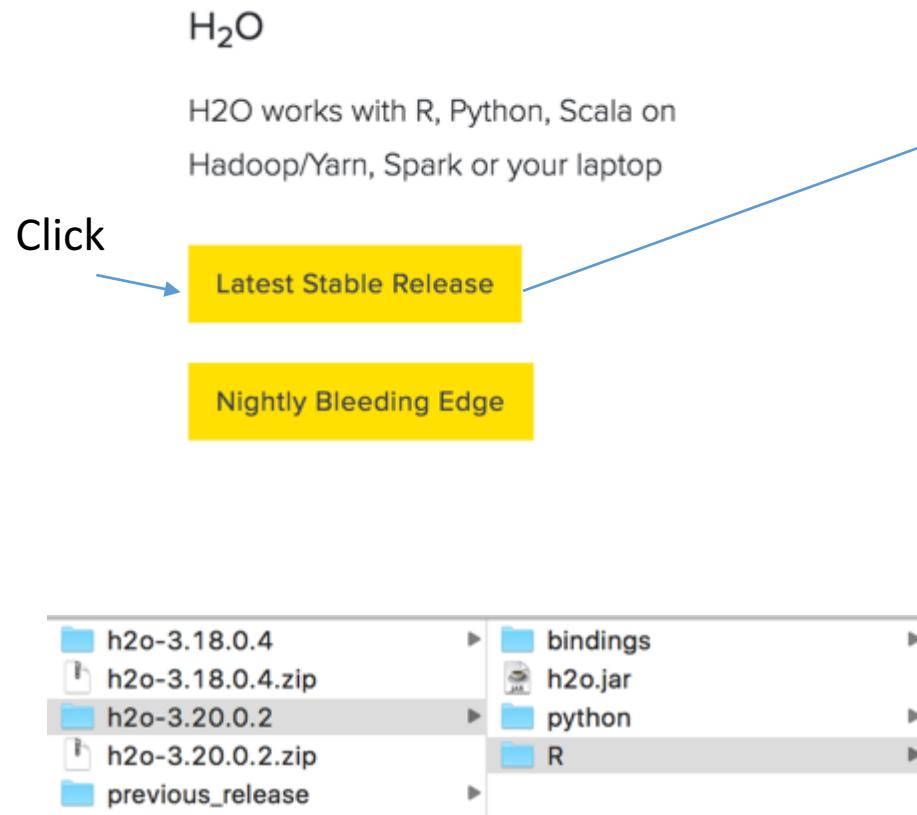
```
# Install 'h2o' from CRAN
install.packages('h2o')
```

# rmarkdown for rendering HTML

```
install.packages('rmarkdown')
```

# Install H<sub>2</sub>O using tar.gz (if needed)

Go to <https://www.h2o.ai/download/>



# H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: \$75M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	<ul style="list-style-type: none"><li>• H2O Open Source Machine Learning (14,000 organizations)</li><li>• H2O Driverless AI – Automatic Machine Learning</li></ul>
Leadership	Leader in Gartner MQ Machine Learning and Data Science Platform
Team	120 AI experts (Kaggle Grandmasters, Distributed Computing, Visualization)
Global	Mountain View, London, Prague, India



# H2O.ai Product Suite



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

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise Support subscriptions

- Enterprise software
- Built for domain users, analysts & data scientists – GUI based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

# H2O.ai Product Suite

**H<sub>2</sub>O**

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

## This Workshop: Open Source H<sub>2</sub>O Core

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise Support subscriptions

DRIVERLESSAI

Automatic feature  
engineering, machine  
learning and interpretability

- Enterprise software
- Built for domain users, analysts & data scientists – GUI based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

# H<sub>2</sub>O Flow (Web)

The screenshot shows the H2O Flow (Web) interface running in a browser window. The title bar reads "H2O Flow". The main menu bar includes "Flow", "Cell", "Data", "Model" (which is highlighted in yellow), "Score", "Admin", and "Help". A sub-menu for "Model" is open, listing various machine learning and data processing routines. The left sidebar is titled "Untitled Flow" and contains a toolbar with icons for file operations and a search bar with the text "assist". Below the toolbar is a section titled "Assistance" with a table of routines and their descriptions. The right side of the interface features tabs for "OUTLINE", "FLOWS", "CLIPS", and "HELP" (which is also highlighted in yellow). The "HELP" tab contains sections for "Using Flow for the first time?", "Quickstart Videos", "Or, view example Flows to explore and learn H2O.", "STAR H2O ON GITHUB!", and "GENERAL" and "EXAMPLES" sections with links to documentation.

Model

- Aggregator...
- Deep Learning...
- Distributed Random Forest...
- Gradient Boosting Machine...
- Generalized Linear Modeling...
- Generalized Low Rank Modeling...
- K-means...
- Naive Bayes...
- Principal Components Analysis...
- Stacked Ensemble...
- Word2Vec...
- XGBoost...

ROUTINE DESCRIPTION

importFiles	Import file(s) into H <sub>2</sub> O
getFrames	Get a list of frames in H <sub>2</sub> O
splitFrame	Split a frame into two or more
mergeFrames	Merge two frames into one
getModels	Get a list of models in H <sub>2</sub> O
getGrids	Get a list of grid search resul
getPredictions	Get a list of predictions in H <sub>2</sub> O
getJobs	Get a list of jobs running in H <sub>2</sub> O
buildModel	Build a model
runAutoML	Automatically train and tune
importModel	Import a saved model
predict	Make a prediction

Using Flow for the first time?

Quickstart Videos

Or, view example Flows to explore and learn H<sub>2</sub>O.

STAR H<sub>2</sub>O ON GITHUB!

GENERAL

- Flow Web UI ...
- ... Importing Data
- ... Building Models
- ... Making Predictions
- ... Using Flows
- ... Troubleshooting Flow

EXAMPLES

Flow packs are a great way to explore and learn H<sub>2</sub>O. Try out these Flows and run them in your browser.

Browse installed packs...

localhost:54321/flow/index.html#

Connections: 0 H<sub>2</sub>O

# Using H<sub>2</sub>O with R and Python

The screenshot shows the RStudio Source Editor with an R script named `credit_card_example.R`. The code performs the following steps:

- Imports datasets from S3.
- Starts and connects to a local H2O cluster.
- Imports data frames from CSV files.
- Looks at datasets using `summary`.
- Defines features and target variable.
- Trains a GBM model using `h2o.gbm`.
- Prints the trained model.
- Uses the GBM model for predictions.
- (Extra) Uses H2O's AutoML to train a model.
- Prints the leaderboard.
- Uses the best model for predictions.

The screenshot shows a Jupyter Notebook titled `credit_card_example` running in Python 2. The notebook contains the following code and output:

- Attempts to start a local H2O server but fails because it is already running.
- Shows the H2O cluster status:

H2O cluster uptime:	02 secs
H2O cluster version:	3.13.0-3981
H2O cluster version age:	29 days
H2O cluster name:	H2O_from_python_jofalchow_id7iqa
H2O cluster total nodes:	1
- Imports datasets from S3.
- Shows progress bars for importing data frames from CSV files.
- Shows the summary of the data frames:

	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	-2.0
mean	165471.466667	1.85	1.55578703704	35.4053240741	-0.00523148148148	-0.122361111111	-0.15537037037037	-0.210601	-0.210601
maxs	1000000.0	6.0	3.0	79.0	8.0	8.0	8.0	8.0	8.0
sigma	128853.314839	0.779550606278	0.522505078476	0.27675421641	1.12668964211	1.2008854503	1.20727030901	1.172176	1.172176
zeros	0	9	37	0	10563	11284	11309	11905	11905
missing	0	0	0	0	0	0	0	0	0

CONFIDENTIAL

# 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*



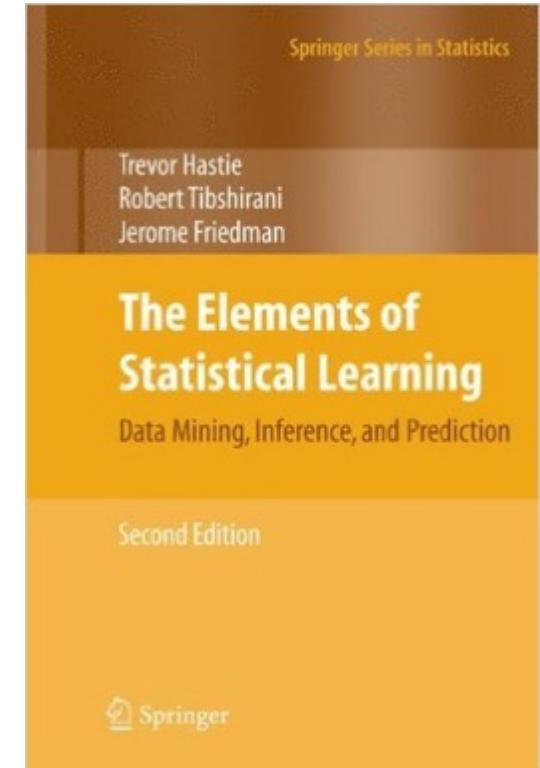
## Dr. Robert Tibshirani

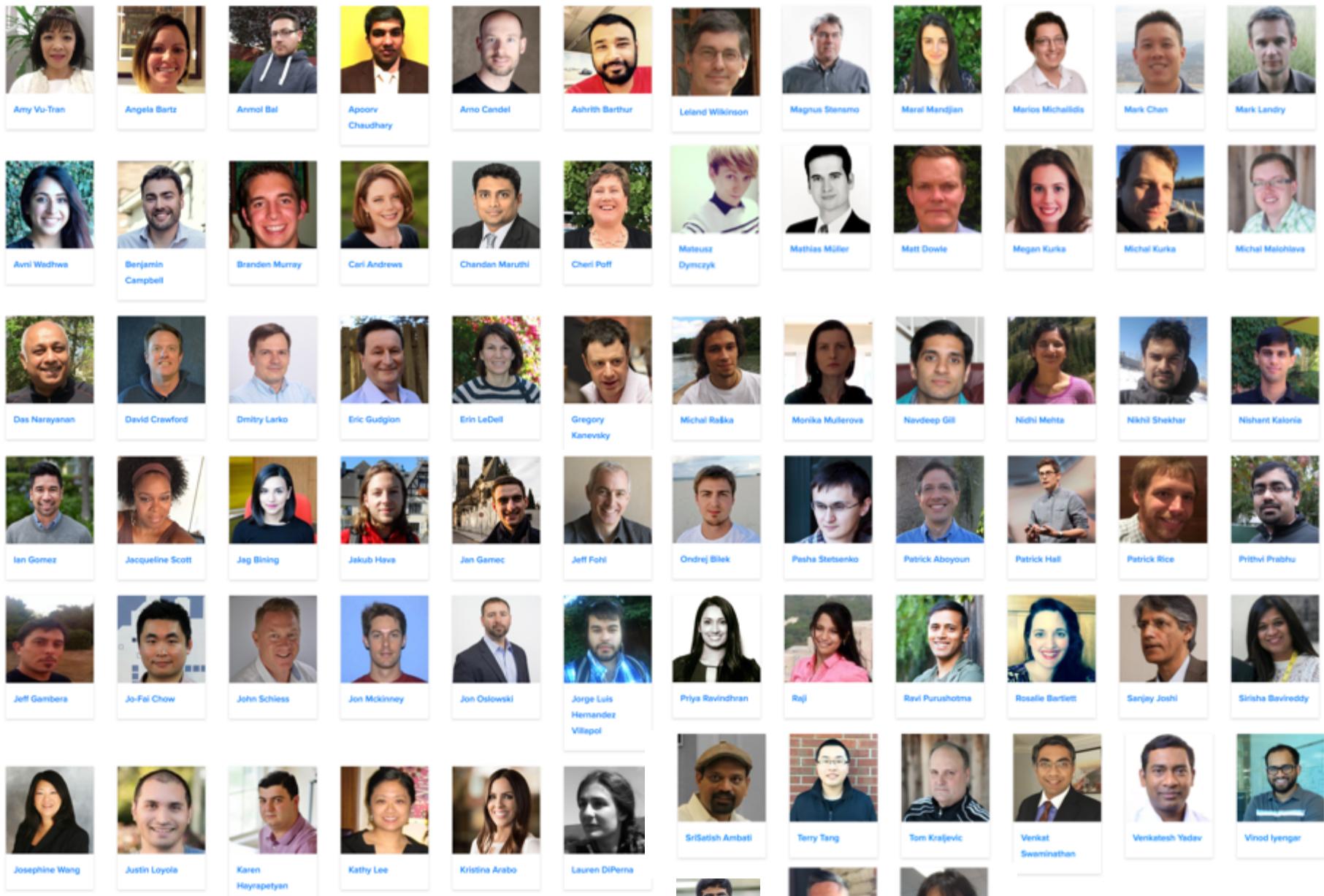
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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



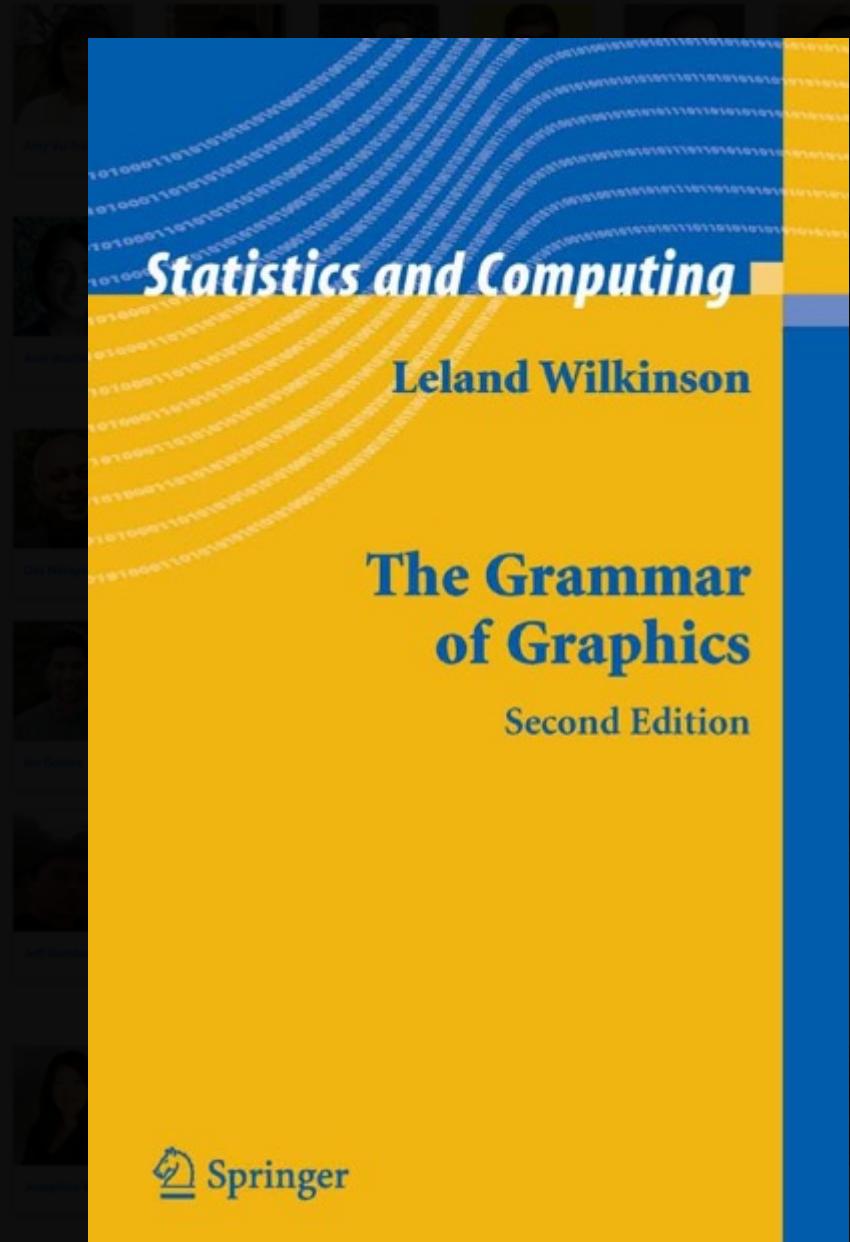
## Dr. Steven Boyd

- 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*





# H<sub>2</sub>O Team



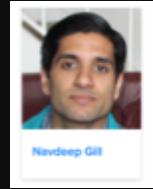


H<sub>2</sub>O Team

# H<sub>2</sub>O AutoML



Erin LeDell

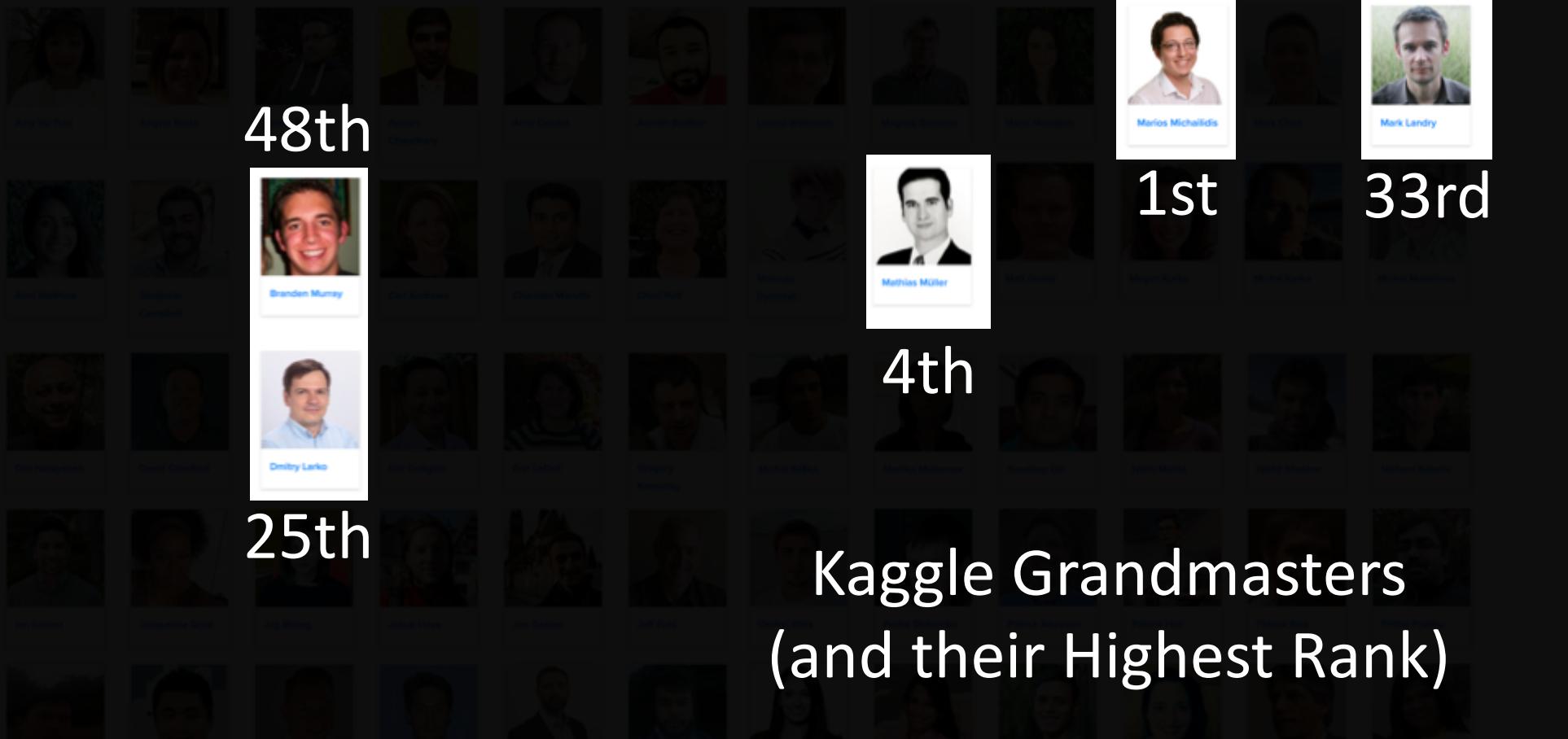


Navdeep Gill

Erin LeDell

Navdeep Gill

H<sub>2</sub>O Team



## Kaggle Grandmasters (and their Highest Rank)

 113  
Grandmasters

 980  
Masters

 3,339  
Experts

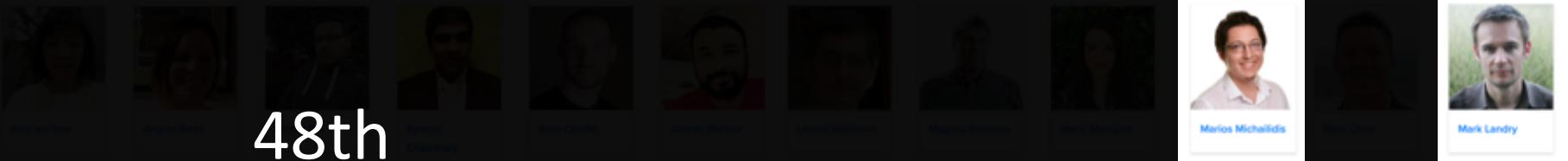
 46,135  
Contributors

 33,242  
Novices

About 80,000 Kagglers

H<sub>2</sub>O Team

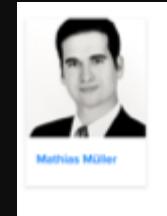
H<sub>2</sub>O.ai



25th



181st



4th

Marios Michalidis

1st

Mark Landry

33rd

13th

H<sub>2</sub>O Team

H<sub>2</sub>O.ai

Hoping to get closer to them at some point ...

# Worldwide Recognition in the **H2O.ai** Community

Open source  
community

**222** OF FORTUNE  
THE 500  
 **H<sub>2</sub>O**

**8** OF TOP 10  
BANKS

**7** OF TOP 10  
INSURANCE COMPANIES

**4** OF TOP 10  
HEALTHCARE COMPANIES

Paying Customers



*"H2O.ai's reference customers gave it the highest overall score for sales relationship and overall service and support" - Gartner MQ 2018*

**H<sub>2</sub>O.ai**

# H2O.ai is a **Leader** in the 2018 Gartner Data Science and Machine Learning Platforms Magic Quadrant

- Technology leader with most completeness of vision
- Recognized for the mindshare, partner network and status as a **quasi-industry standard** for machine learning and AI
- H2O.ai 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



Get the  
Gartner  
Magic  
Quadrant  
[here](#)

# Platforms with H<sub>2</sub>O 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



H<sub>2</sub>O + KNIME Talk  
at KNIME Summit  
Mar 2017

1:54 PM - 7 Mar 2018 from Hotel Berlin

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



Source: Gartner (February 2018)

As of January 2018  
© Gartner, Inc

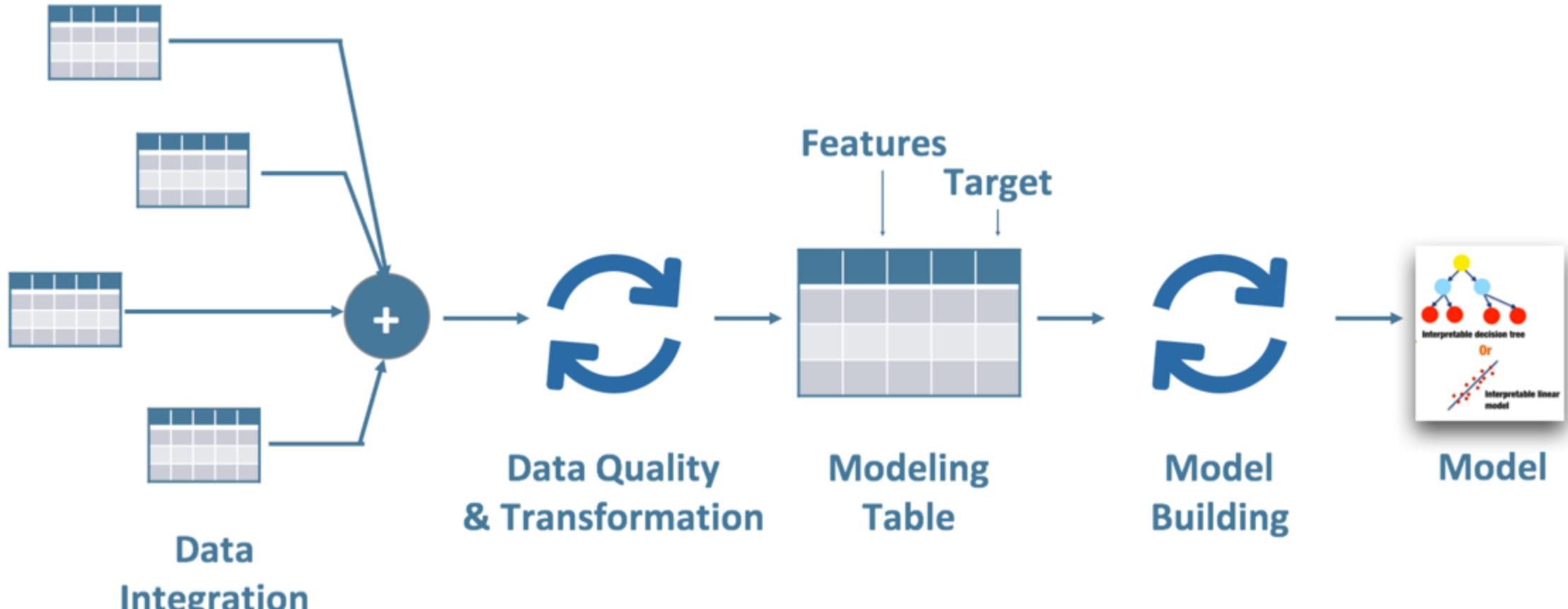
H<sub>2</sub>O.ai

# H<sub>2</sub>O AutoML

Automatic Machine Learning with H<sub>2</sub>O

<http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>

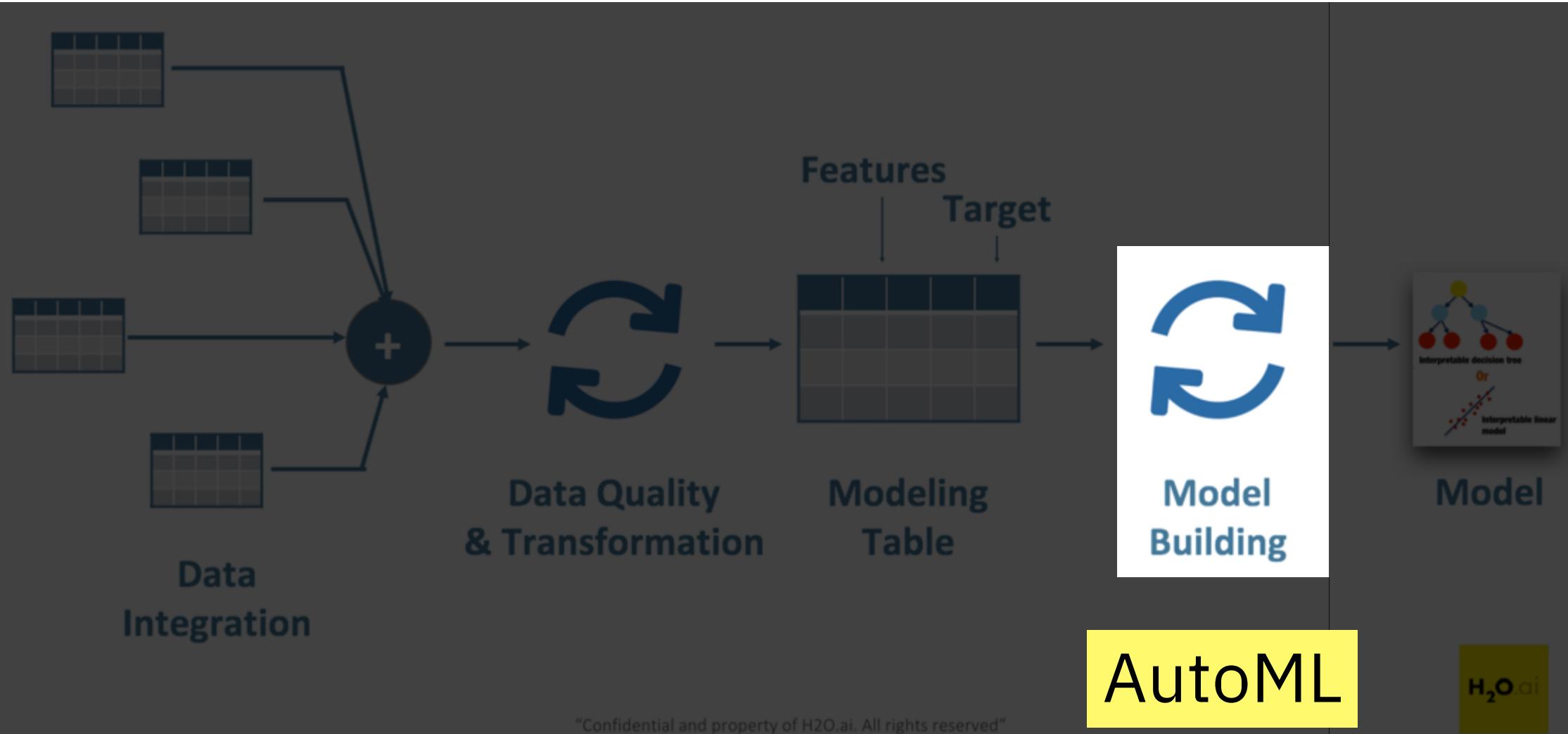
# Typical Enterprise Machine Learning Workflow



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



# Typical Enterprise Machine Learning Workflow



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

# H<sub>2</sub>O-3 Supervised Algorithms in AutoML

## 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

### 3.3 H2O AutoML: Multiple H2O Models + Stacked Ensemble

```
# Train multiple H2O models with H2O AutoML
# Stacked Ensembles will be created from those H2O models
# You tell H2O ...
#   1) how much time you have and/or
#   2) how many models do you want
# Note: H2O deep learning algo on multi-core is stochastic
model_automl = h2o.automl(x = features,
                           y = target,
                           training_frame = h_train,
                           nfolds = 5,           # Cross-Validation
                           max_runtime_secs = 120, # Max time
                           max_models = 100,      # Max no. of models
                           stopping_metric = "RMSE", # Metric to optimize
                           project_name = "my_automl",
                           exclude_algos = NULL,    # If you want to exclude any algo
                           seed = n_seed)
```

## 3.4 AutoML Leaderboard

```
model_automl@leaderboard
```

```
##                                     model_id
## 1 StackedEnsemble_BestOfFamily_0_AutoML_20180514_100853
## 2 DeepLearning_grid_0_AutoML_20180514_100853_model_1
## 3 StackedEnsemble_AllModels_0_AutoML_20180514_100853
## 4 GBM_grid_0_AutoML_20180514_100853_model_1
## 5 GBM_grid_0_AutoML_20180514_100853_model_3
## 6 GBM_grid_0_AutoML_20180514_100853_model_2
##   mean_residual_deviance      rmse       mae      rmsle
## 1          8.850089 2.974910 2.017611 0.1422685
## 2          9.119150 3.019793 2.124464 0.1572535
## 3          9.537272 3.088247 2.030157 0.1397577
## 4         11.299723 3.361506 2.176395 0.1501403
## 5         11.535687 3.396423 2.181420 0.1510191
## 6         11.737661 3.426027 2.208042 0.1524836
##
## [20 rows x 5 columns]
```

# H<sub>2</sub>O Documentation

[Getting Started & User Guides](#) | [Q & A](#) | [Algorithms](#) | [Languages](#) | [Tutorials, Examples, & Presentations](#) | [API & Developer Docs](#) | [For the Enterprise](#)

## Getting Started & User Guides

  Open Source |   Commercial

**H<sub>2</sub>O**

What is H<sub>2</sub>O?  
**H<sub>2</sub>O User Guide** (Main docs)  
H<sub>2</sub>O Book (O'Reilly)  
Recent Changes  
Open Source License (Apache V2)

Quick Start Video - Flow Web UI  
Quick Start Video - R  
Quick Start Video - Python

[Download H<sub>2</sub>O](#)

**Sparkling Water**

What is Sparkling Water?  
**Sparkling Water User Guide** 2.3 2.2 2.1  
Sparkling Water Booklet  
RSparkling Readme  
PySparkling User Guide 2.3 2.2 2.1  
Recent Changes 2.3 2.2 2.1  
Open Source License (Apache V2)

Quick Start Video - Scala

[Download Sparkling Water](#)

**Driverless AI**

What is Driverless AI?  
Driverless AI User Guide [HTML](#) [PDF](#)  
Recent Changes  
Driverless AI Booklet  
MLI with Driverless AI Booklet

Quick Start Video - Downloading Driverless AI  
Quick Start Video - Launching an Experiment  
Driverless AI Webinars

[Download Driverless AI](#)

**H2O4GPU (alpha)**

H2O4GPU Readme  
Open Source License (Apache V2)

[Download H2O4GPU](#)

**URL: [docs.h2o.ai](#)**

# LIME

## Local Interpretable Model- Agnostic Explanations

# Acknowledgement

- **Marco Tulio Ribeiro:** Original LIME Framework and Python package 
- **Thomas Lin Pedersen:** LIME R package 
- **Matt Dancho:** LIME + H2O AutoML example + LIME R package improvement 
- **Kasia Kulma:** LIME + H2O AutoML example 

# Why Should I Trust Your Model?



System that performs behaviour but you don't know how it works

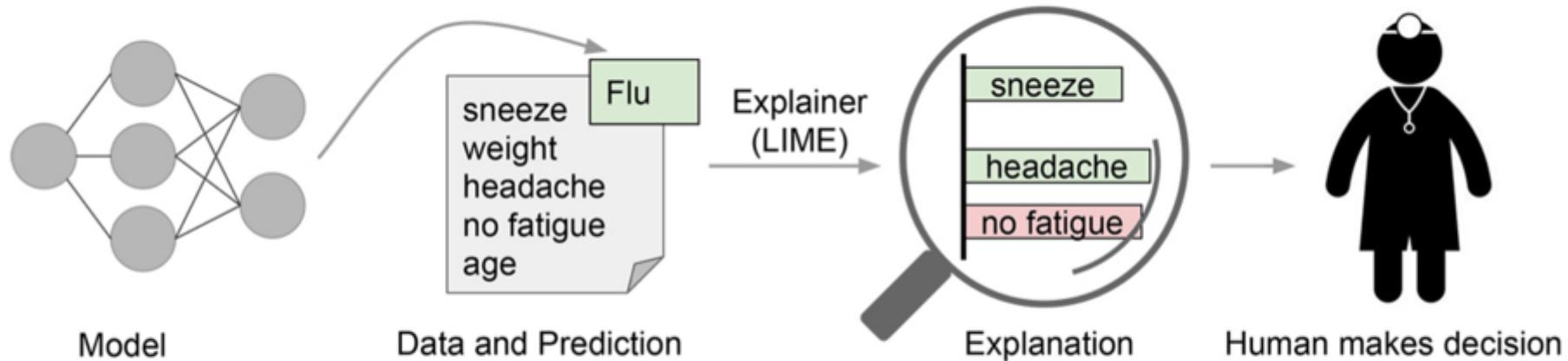


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

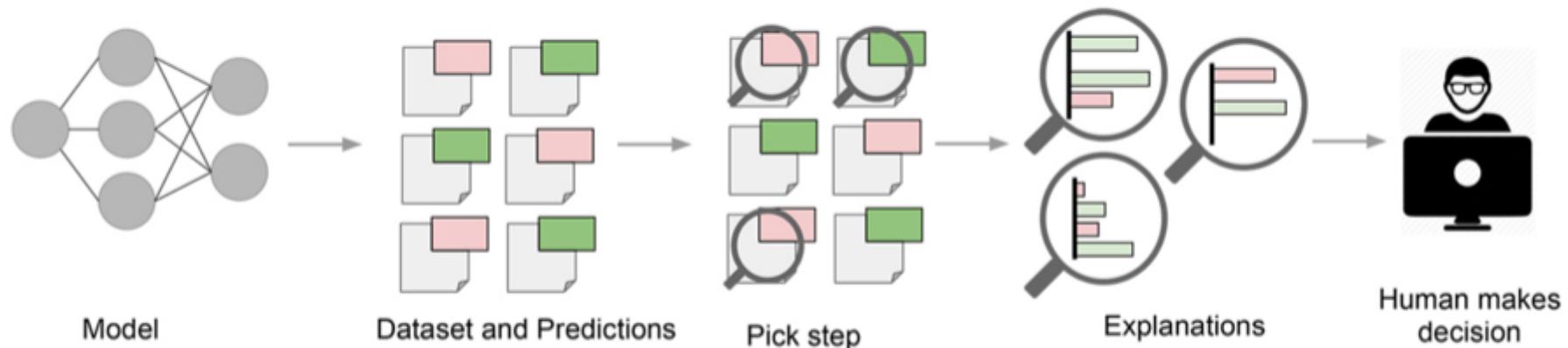


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

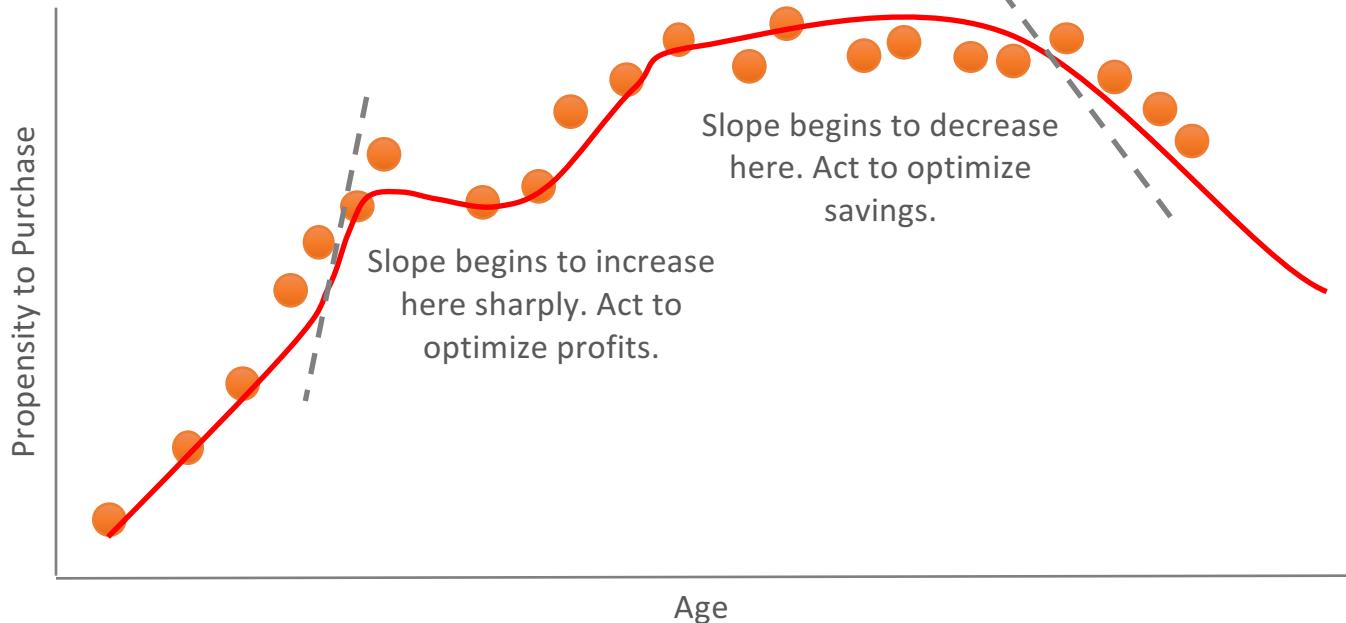
## Linear Models

*Exact explanations for approximate models.*



## Machine Learning

*Approximate explanations for exact models.*



# Local Interpretable Model-Agnostic Explanations

## LIME - How does it work?

### Theory

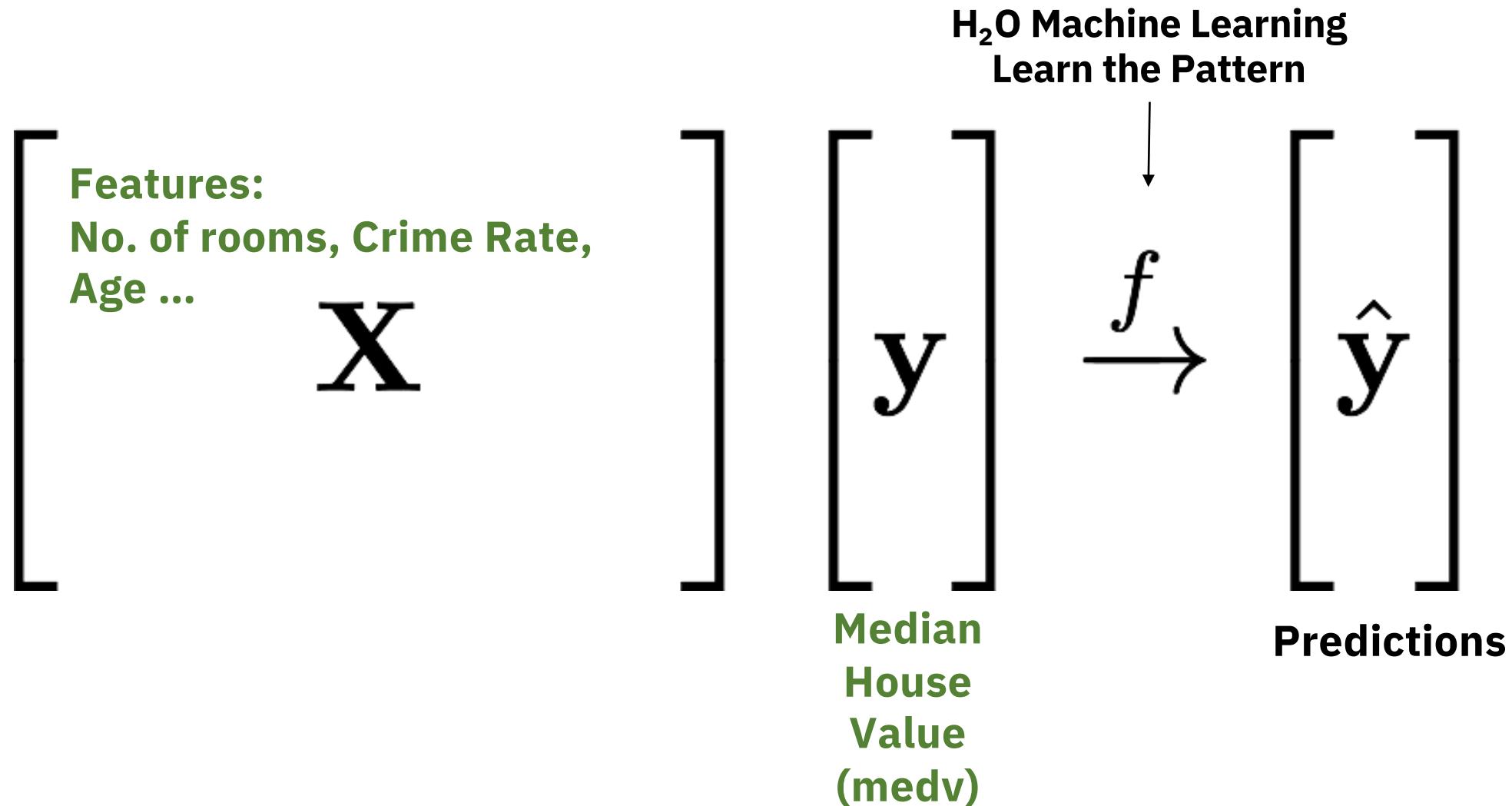
- LIME approximates model locally as logistic or linear model
- Repeats process many times
- Outputs features that are most important to local models

### Outcome

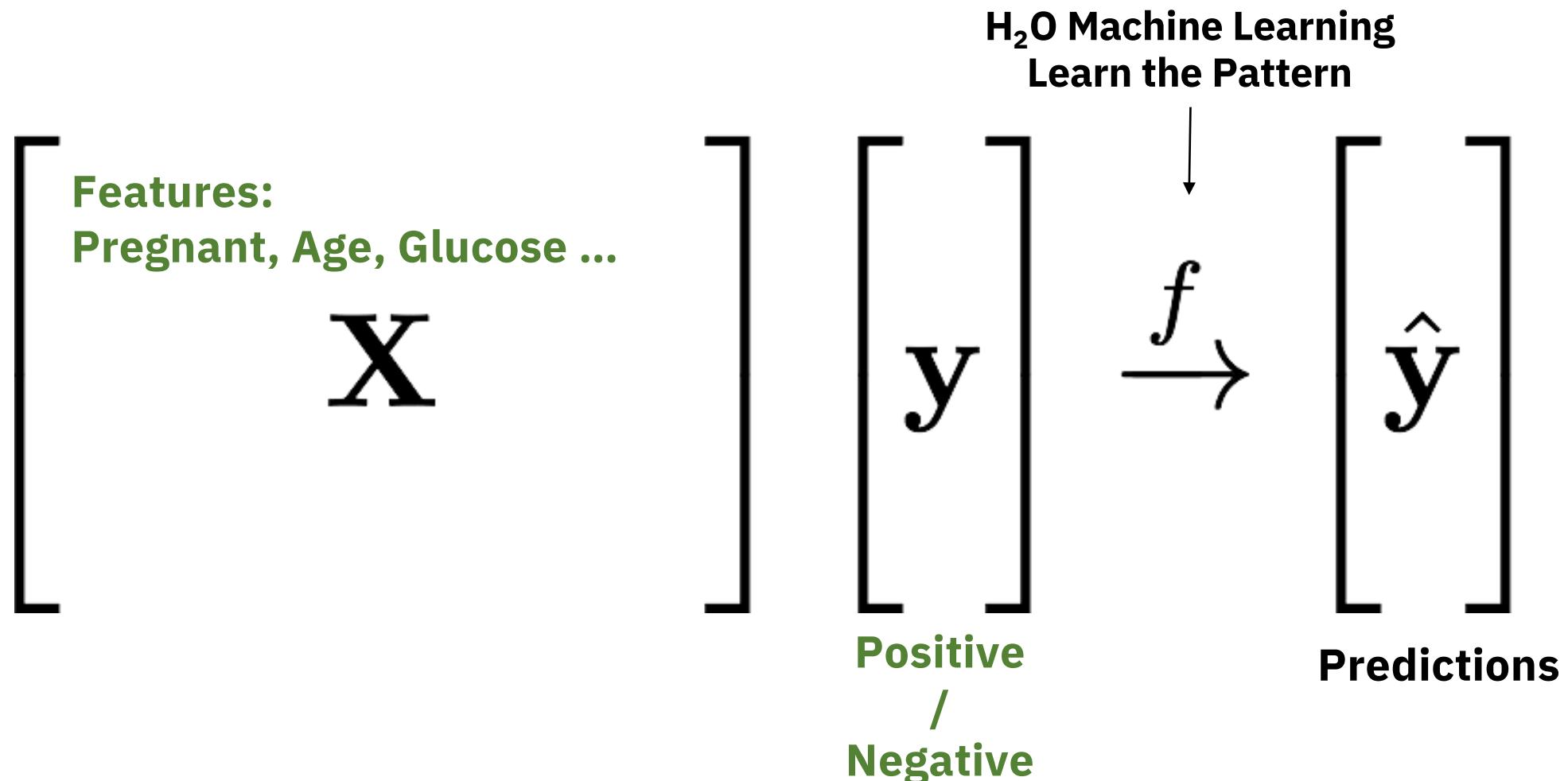
- Approximate reasoning
- Complex models can be interpreted
  - Neural nets, Random Forest, Ensembles etc.

# Worked Examples

# Learning from **Boston Housing** Data



# Learning from **Diabetes** Data



# Quick Recap



#ODSC

# Why?

- Most users/organizations can benefit from **automatic machine learning pipelines**.
  - Eliminate time wasted on repetitive tasks, human errors, debugging etc.
- GDPR mandates a “right to explanation” from machine learning models.
  - **model interpretations** are crucial for those who must explain their models to regulators or customers.

# You will learn ...

- How to build high quality **H<sub>2</sub>O** models (almost) automatically.
- How to explain predictions from complex **H<sub>2</sub>O** models with **LIME**.
- **Bonus:** A real use case that led to a **multimillion-dollar** baseball decision earlier this year.

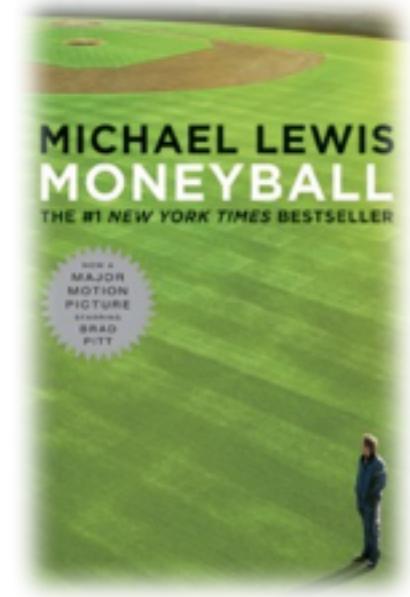
# Quick Recap

# Moneyball: The Multimillion-Dollar Business Problem

The quest to find the most undervalued players  
(before other teams notice them)



Source: Moneyball, 2011 Columbia Pictures



# The Real Business Problem in Major League Baseball (MLB)

- Existing Forecasts (e.g. ESPN) are usually projections for the **next year only**.
- MLB players usually consider terms for 3 to 5 years when they sign a new contract.
- MLB teams need to consider players' **long-term performance** (i.e. > 1 year).

The screenshot shows the ESPN Fantasy Baseball interface. At the top, there's a navigation bar with the ESPN logo, NFL, NBA, MLB, NCAAF, Soccer, and other links. Below the navigation is a search bar and a login link. The main content area is titled "Sortable 2018 Projections" and "Position: Batters". It includes filters for "By Name" and "View: 2018 Season". A large orange arrow points from the text "Existing Forecasts (e.g. ESPN) are usually projections for the next year only." to this section. Below the filters is a table titled "PLAYERS" with columns for RANK, PLAYER, TEAM POS, and various statistics like R, HR, RBI, SB, and AVG. An orange box highlights the "2018 SEASON BATTING PROJECTIONS" table, which lists 12 players with their projected stats. A blue banner at the bottom reads "2018 SEASON BATTING PROJECTIONS".

2018 SEASON BATTING PROJECTIONS						
RNK	PLAYER, TEAM POS	R	HR	RBI	SB	AVG
1	Mike Trout, LAA OF	119	40	98	22	.308
2	Jose Altuve, Hou 2B	106	24	83	32	.329
3	Nolan Arenado, Col 3B	105	38	132	3	.300
4	Mookie Betts, Bos OF	107	24	84	29	.294
5	Bryce Harper, Wash OF	109	35	102	12	.309
6	Trea Turner, Wash SS	97	15	59	57	.287
7	Charlie Blackmon, Col OF	116	30	84	14	.315
8	Paul Goldschmidt, Ari 1B	102	28	102	19	.296
9	Carlos Correa, Hou SS	99	28	107	12	.301
10	Giancarlo Stanton, NYY OF, DH	107	52	118	2	.269
11	Kris Bryant, Chi 3B	110	32	94	10	.296
12	Manny Machado, Bal 3B, SS	97	34	98	10	.294

# The Moneyball Team



IBM

**David Kearns**  
PM @ IBM Data Science

A portrait of David Kearns, a man with dark hair and a beard, wearing a pink striped shirt. He is standing in front of a wall with the letters "TH" and "NC" visible. The IBM logo is in the bottom left corner of the photo frame.

H<sub>2</sub>O

**Jo-Fai Chow**  
Data Scientist @ H<sub>2</sub>O.ai

A portrait of Jo-Fai Chow, a man with dark hair, wearing a red and white checkered shirt. He is standing in front of a blue and white graphic background featuring stylized buildings and data points. The H<sub>2</sub>O.ai logo is in the bottom left corner of the photo frame.

Aginity

**Ari Kaplan**  
Mr. Moneyball @ Aginity

A portrait of Ari Kaplan, a bald man with glasses, wearing a dark suit and a yellow tie. He is smiling. The Aginity logo is in the bottom left corner of the photo frame.

IBM + Aginity + H<sub>2</sub>O.ai

# Baseball Player Performance Data

- Open data – **Lahman** Database.
- Proprietary data (**AriDB**) from Ari Kaplan – our real Moneyball guy.
- Enriched Lahman data with Ari's Data – Final dataset for predictive modelling



# Lahman Database

<http://www.seanlahman.com/baseball-archive/statistics/>

Attribute	Description
playerID	Player ID code
yearID	Year player was born
G	Games
AB	At Bats
R	Runs
H	Hits
2B	Doubles
3B	Triples
HR	Homeruns
SO	Strike Outs
IBB	Intentional Walks
SF	Sacrifice flies

# Ari's Database

- Private database containing 5 years of data
- Pitch-by-pitch play for each MLB game:
  - Pitch type, top speed, end speed, spin rate, x, y, z coordinates, batter result etc.

Attribute	Description
Pitch_Type	Two - character code of type of pitch. FF=fastball, CU=curveball, SL=slider, etc.
Spin_rate	Spin of the pitch in rotations per minute. One of the top fields for a feature...the theory is the more spin the harder it is to hit.
Start_speed	The velocity of the pitch in mph (when it leaves the hand, which is the measure used for tv).
End_speed	The velocity of the pitch when it arrives at the plate
Z0	Feet off the ground when the pitch is released.
Spray_x	When ball is hit into play, this is the x - coordinate of where it is hit/picked up by a fielder
Spray_y	When ball is hit into play, this is the y - coordinate of where it is hit/picked up by a fielder
Spray_des	Classification of type of hit: pop out, flyout, groundout, hit, error

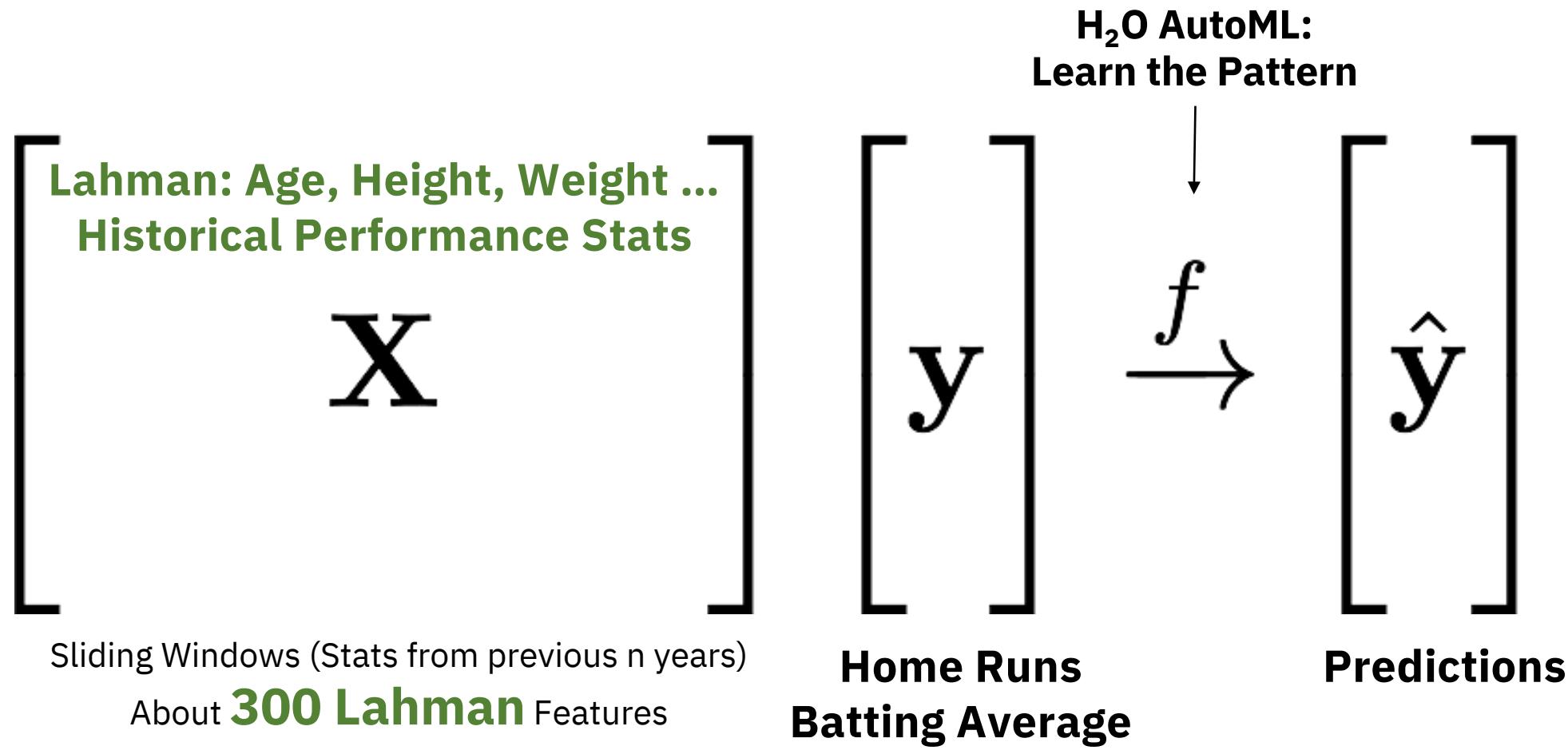
# Predictive Modelling – H<sub>2</sub>O AutoML

- Framed data as regression problems for performance prediction.
- Historical player performance as features.
- Used H<sub>2</sub>O AutoML to build ensembles (linear model, random forests, gradient boosting, and deep neural networks).

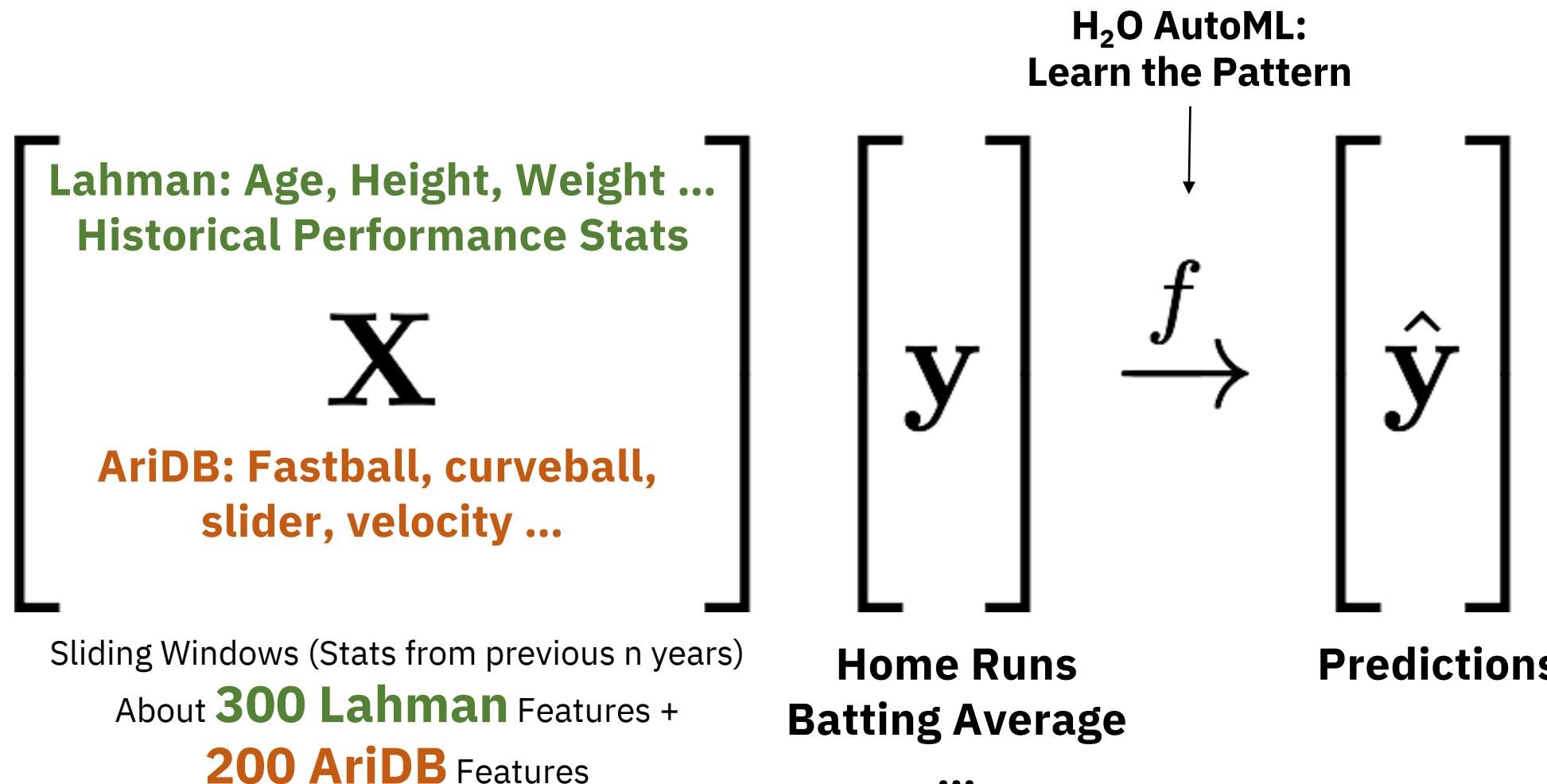


```
# Install 'h2o' from CRAN  
install.packages('h2o')
```

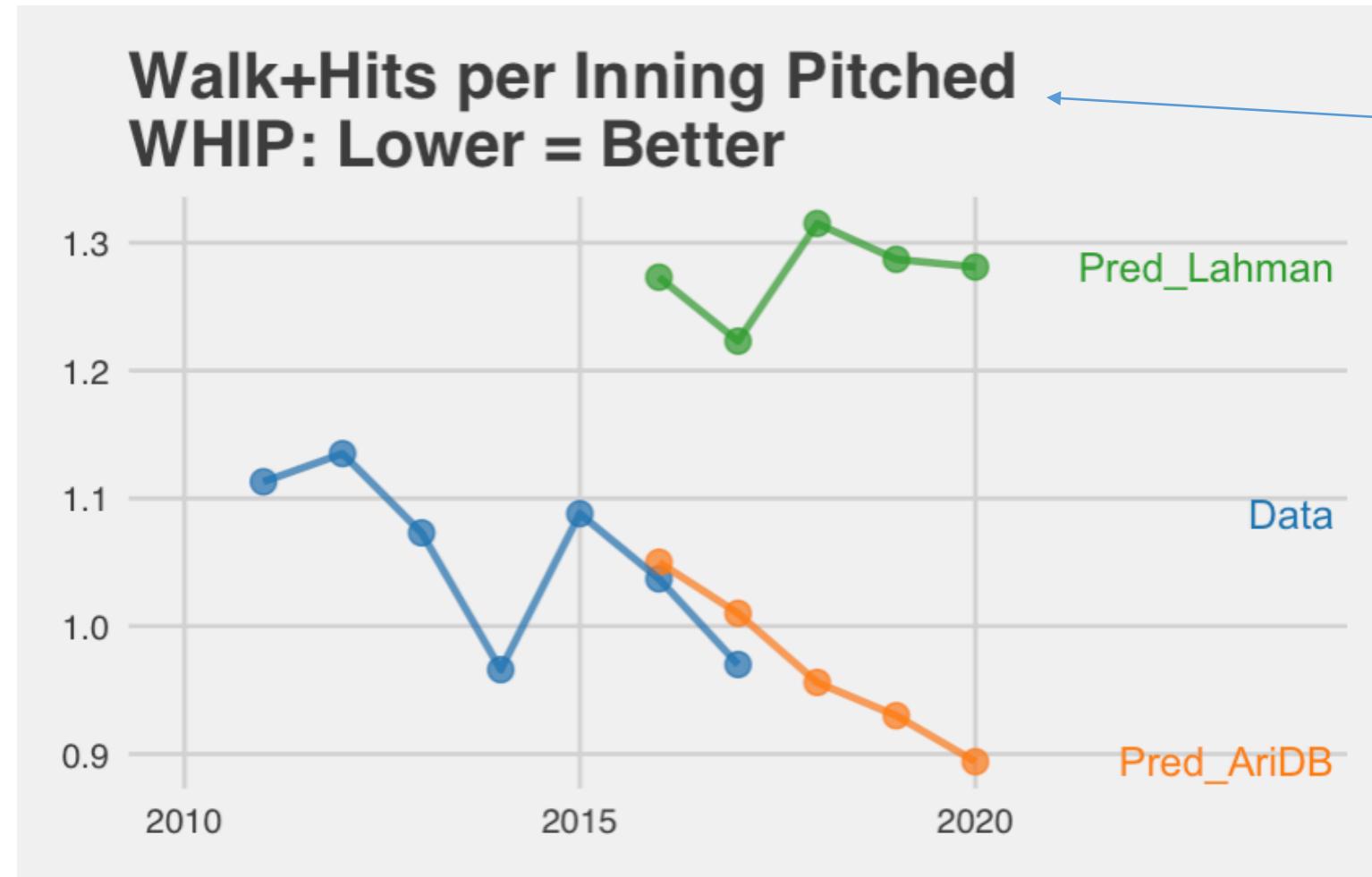
# Approach One: Learning from **Lahman** only



# Approach Two: Learning from **Lahman** & **AriDB**



# Predictive Modelling – H<sub>2</sub>O AutoML

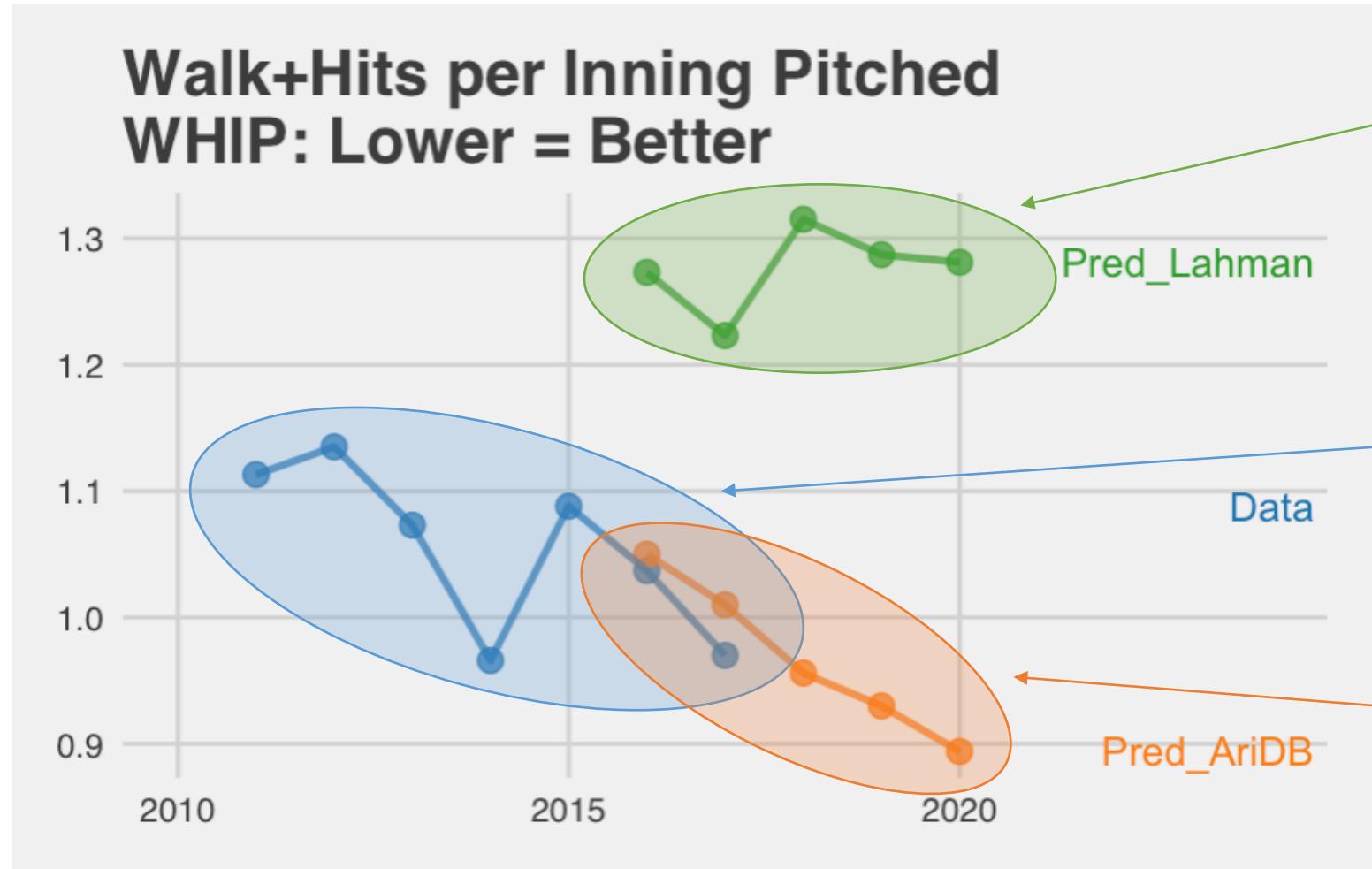


One of Many Targets  
(e.g. Home Runs, Batting Average)



```
# Install 'h2o' from CRAN  
install.packages('h2o')
```

# Predictive Modelling – H<sub>2</sub>O AutoML

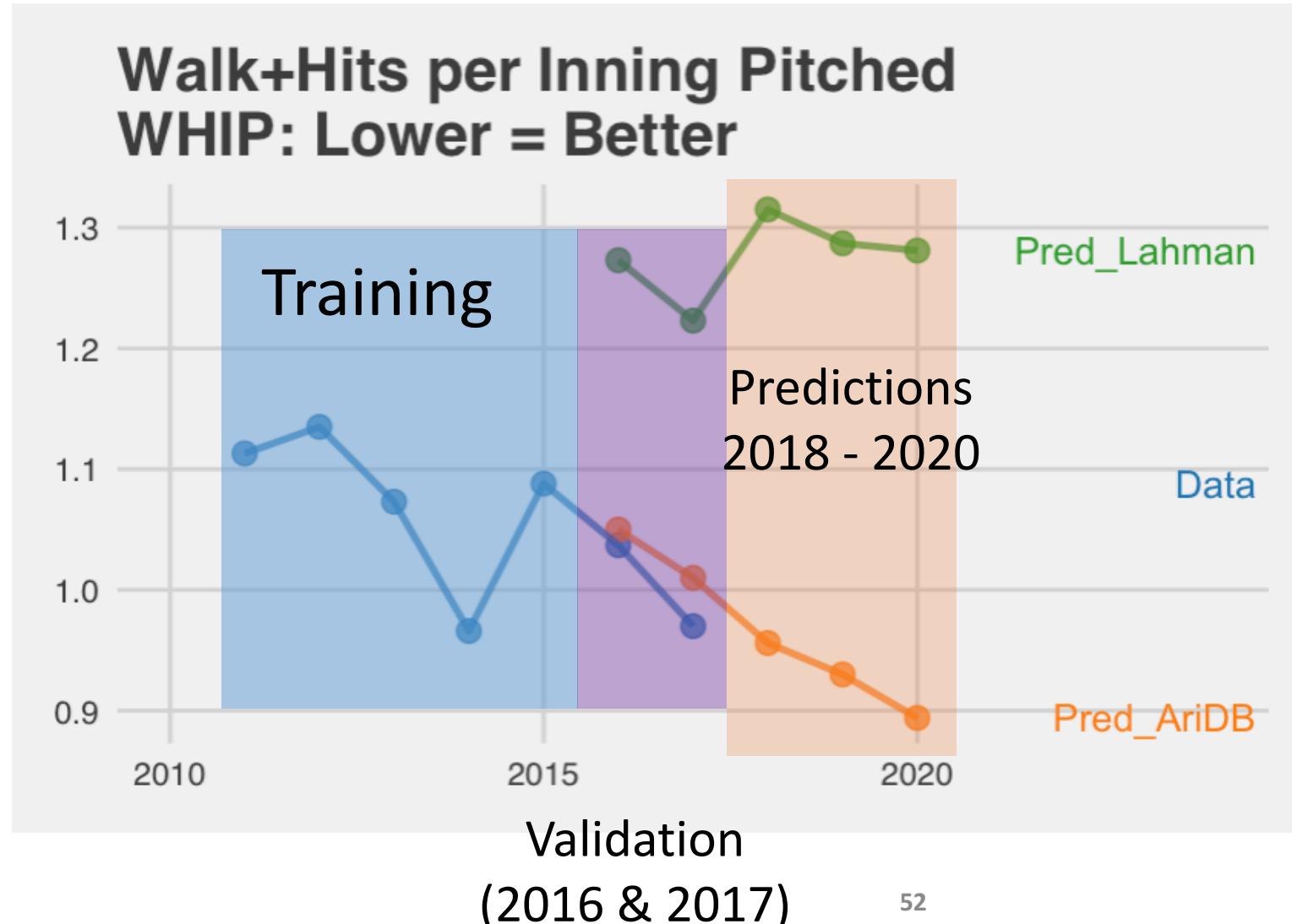


Results from models based on Lahman data only

Historical player performance data

Results from models based on final dataset (Lahman + AriDB)

# Predictive Modelling – H<sub>2</sub>O AutoML

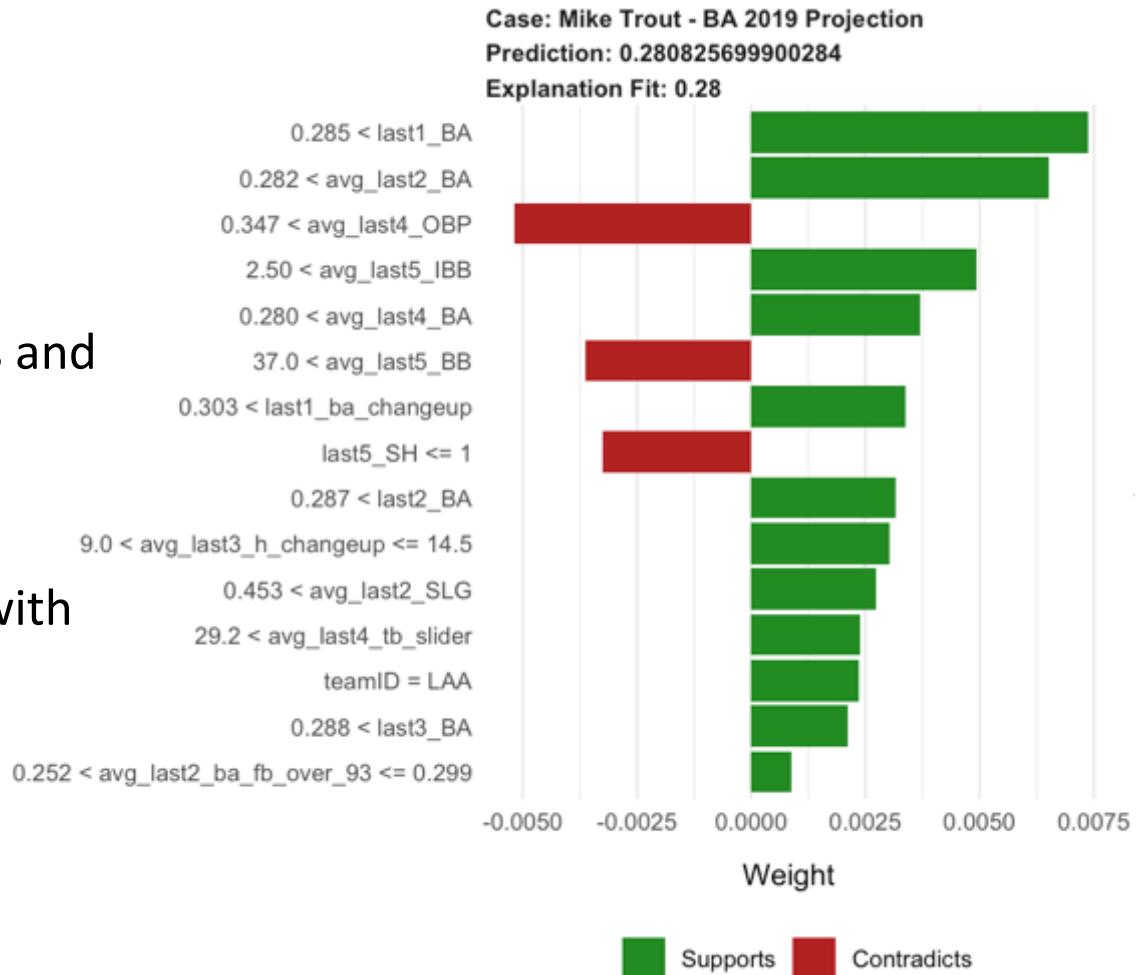


```
# Install 'h2o' from CRAN
install.packages('h2o')
```

# Explaining the Predictions

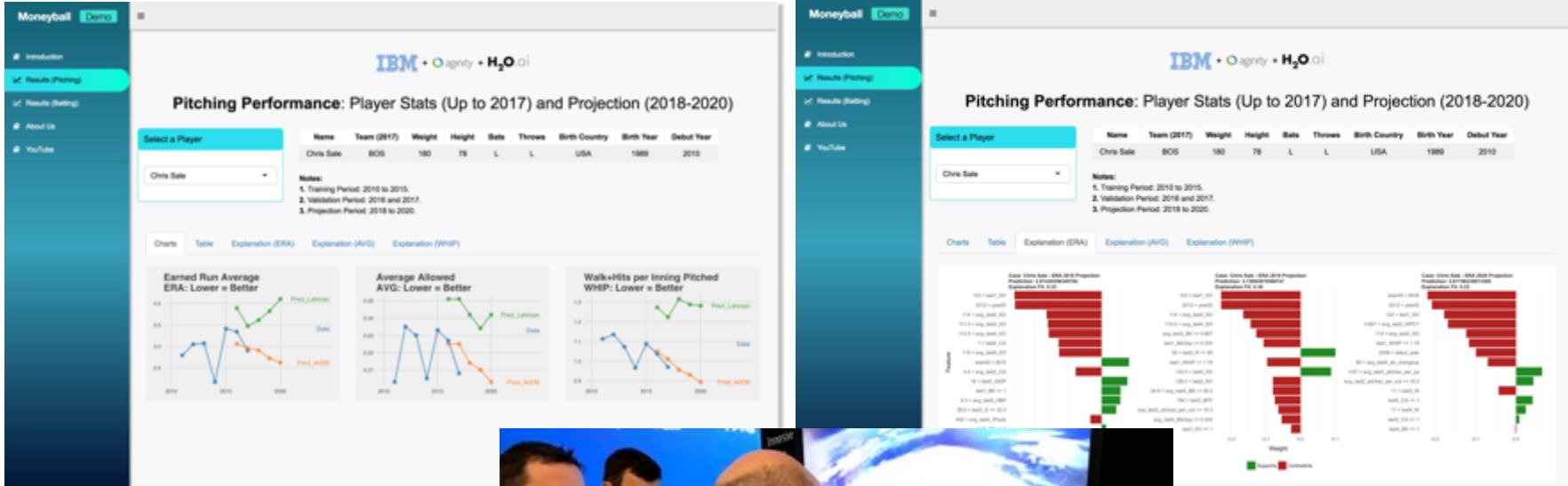
## LIME – Local Interpretable Model-agnostic Explanations

- Approximate reasoning of complex ML models (ensembles).
- Most important attributes and their contributions to the predictions.
- Ari validated the models with his 30+ years of baseball domain knowledge.
- He trusted the models.



```
# Install 'lime' from CRAN  
install.packages('lime')
```

# Putting Everything Together – Moneyball Shiny App



Live Demo

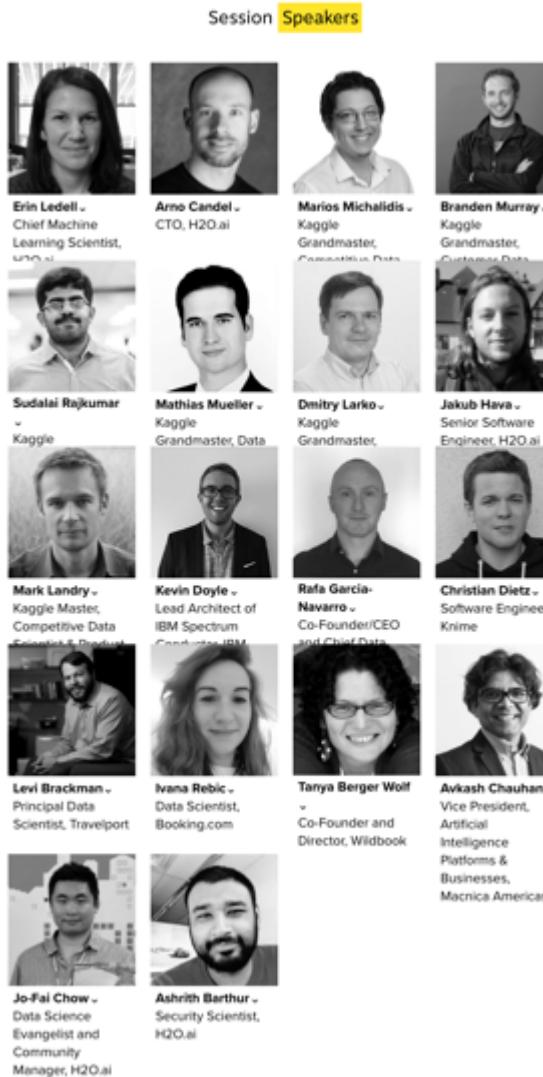


H<sub>2</sub>O.ai

If you want to hear the Moneyball story from Ari and David ...



29<sup>th</sup> & 30<sup>th</sup> Oct, London



More real-world use cases  
+  
All H<sub>2</sub>O Kaggle Grandmasters  
+  
Hands-on Training

**H<sub>2</sub>O.ai**

# Thanks!



- More Info, Code, and Slides
  - [bit.ly/  
odsc2018\\_h2o](https://bit.ly/odsc2018_h2o)
- Contact
  - [joe@h2o.ai](mailto:joe@h2o.ai)
  - [@matlabulous](https://twitter.com/matlabulous)
  - [github.com/woobe](https://github.com/woobe)