

# Making Multimillion-Dollar ⚾ Decisions with H<sub>2</sub>O AutoML, LIME and Shiny



Jo-fai (Joe) Chow

Data Science Evangelist /  
Community Manager

joe@h2o.ai

@matlabulous

Download → [https://bit.ly/  
\*\*h2o\\_meetups\*\*](https://bit.ly/h2o_meetups)



## About this talk

- Quick Overview
  - Business problem
  - Solution and result
- Details
  - The “Moneyball” team
  - Baseball data → ML problem
  - H<sub>2</sub>O AutoML, LIME, Shiny

## You will learn ...

- How to frame “Moneyball” for machine learning.
- How to use H<sub>2</sub>O AutoML with R interface.
- How to use LIME to explain H<sub>2</sub>O models.

# About Me



Jo-fai (Joe) Chow  
@matlabulous

Good evening #Cologne 🇩🇪  
#AroundTheWorldWithH2Oai  
#CologneCathedral #Germany #twitter  
[bit.ly/2nJTxJG](http://bit.ly/2nJTxJG)



## • 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 ...
- H<sub>2</sub>O SWAG Photographer  
#AroundTheWorldWithH2Oai  
Love H<sub>2</sub>O? Get some stickers!

 Jo-fai (Joe) Chow  
@matlabulous

Thanks all for coming to my @erum2018 workshop. Here is our #360selfie. Hope you all enjoyed building @h2oai models w/ #AutoML and explaining them w/ #LIME. Looking forward to the welcome reception and #Shiny demos - totally my thing! #eRum2018 #Budapest #AroundTheWorldWithH2Oai



4:45 PM - 14 May 2018 from Budapest, Hungary

 Jo-fai (Joe) Chow  
@matlabulous

Another #FullHouse @h2oai #LondonAI #meetup tonight. Thanks @MSFTRector for the amazing venue and food! #OpenSource #Community #MVPBuzz #AroundTheWorldWithH2Oai #360Selfie 🇬🇧 cc our guest speakers @SKREDDY99 @cheukting\_ho & Josh Warwick



7:15 PM - 12 Mar 2018 from London, England

 Jo-fai (Joe) Chow  
@matlabulous

Awesome #KNIMESummit2018 #KNIMESpringSummit in #Berlin. @knime @Kurioos Marten here is our #360Selfie cc @h2oai #AroundTheWorldWithH2Oai 🇩🇪 #OpenSource #MachineLearning #Community 💪



1:54 PM - 7 Mar 2018 from Hotel Berlin



Jo-fai (Joe) Chow  
@matlabulous

@h2oai #DeepWater at #KoelnRUG #meetup thanks for having us. Slides: [github.com/h2oai/h2o-meet ...](https://github.com/h2oai/h2o-meet) #AroundTheWorldWithH2Oai #Cologne #Germany 🇩🇪



6:34 PM - 17 Mar 2017 from Cologne, Germany

 Jo-fai (Joe) Chow  
@matlabulous

Thanks @ingnl for hosting @h2oai #meetup in #Amsterdam last week. Tremendous turnout and great discussions.

#AroundTheWorldWithH2Oai #360Selfie 🇳🇱  
cc @fishnets88



7:15 AM - 26 Feb 2018 from Amsterdam, The Netherlands

 Jo-fai (Joe) Chow  
@matlabulous

Merci beaucoup Alexia, Samia & Aurelie from @lse\_dasci. We had our very first @h2oai #meetup in #Toulouse tonight. Fantastic crowd and awesome @HarryCoworking venue. We hope to see you all again in the future. Here is our #360selfie 📸 #AroundTheWorldWithH2Oai 🇫🇷



10:35 PM - 23 Apr 2018 from Toulouse, France

Reminder: #360Selfie

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

# About H2O.ai ...

Have you seen Avengers: Infinity War?

Do you know all the characters in the movie? (No spoilers - I promise)

A circular profile picture of a young woman with long brown hair and glasses, smiling.

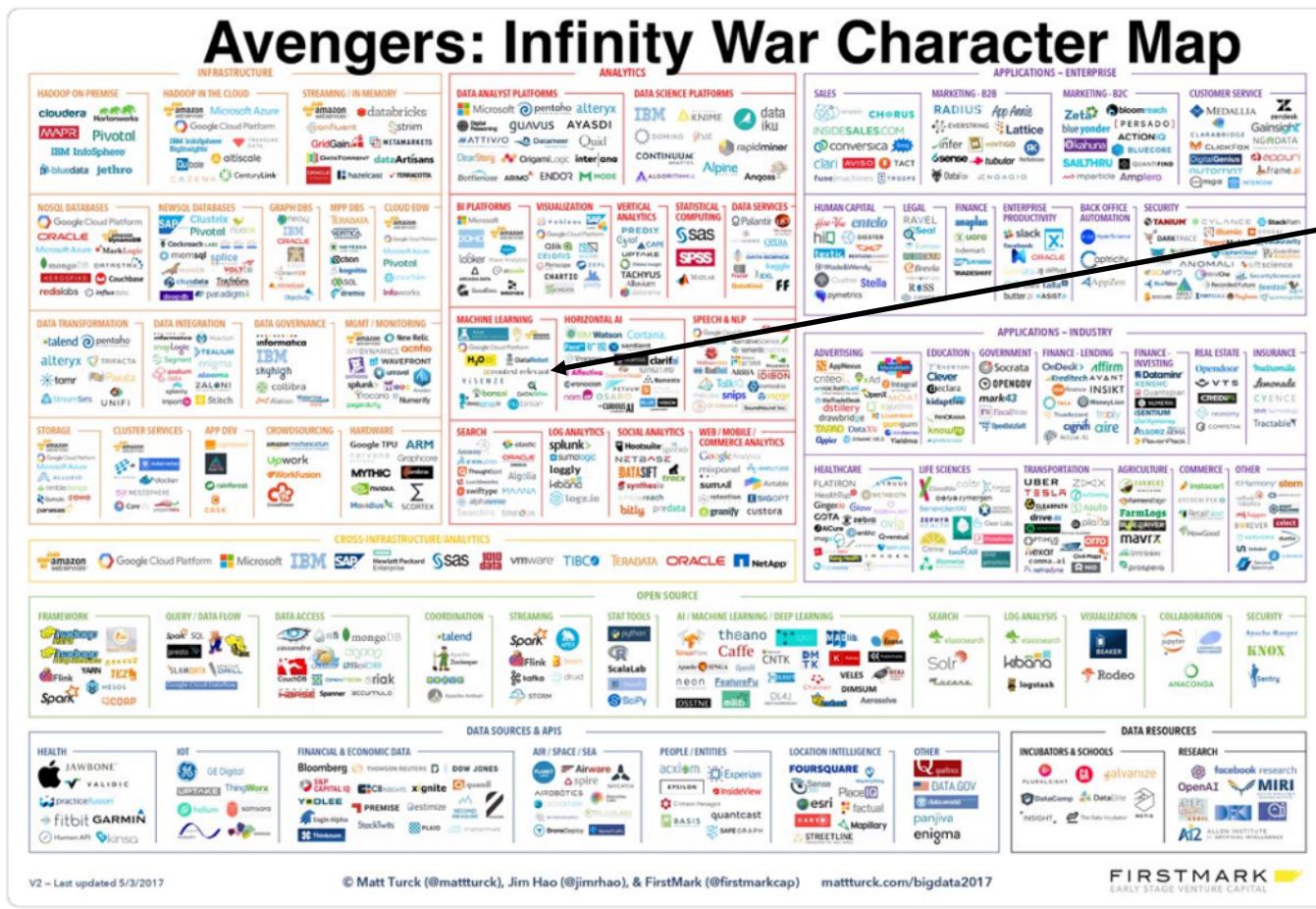
**Vicki Boykis**  
@vboykis

## Follow

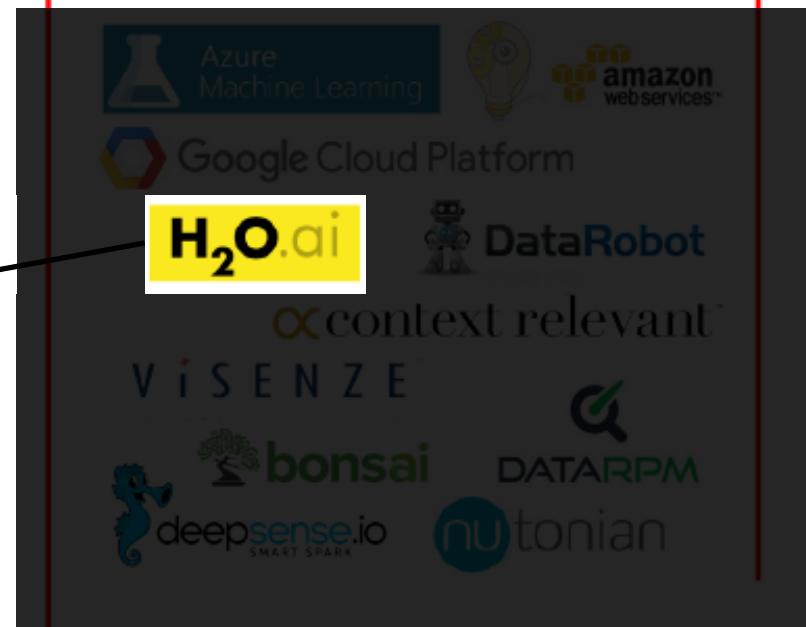
1

I made a guide for anyone who was as confused by all the characters in Infinity War as I was.

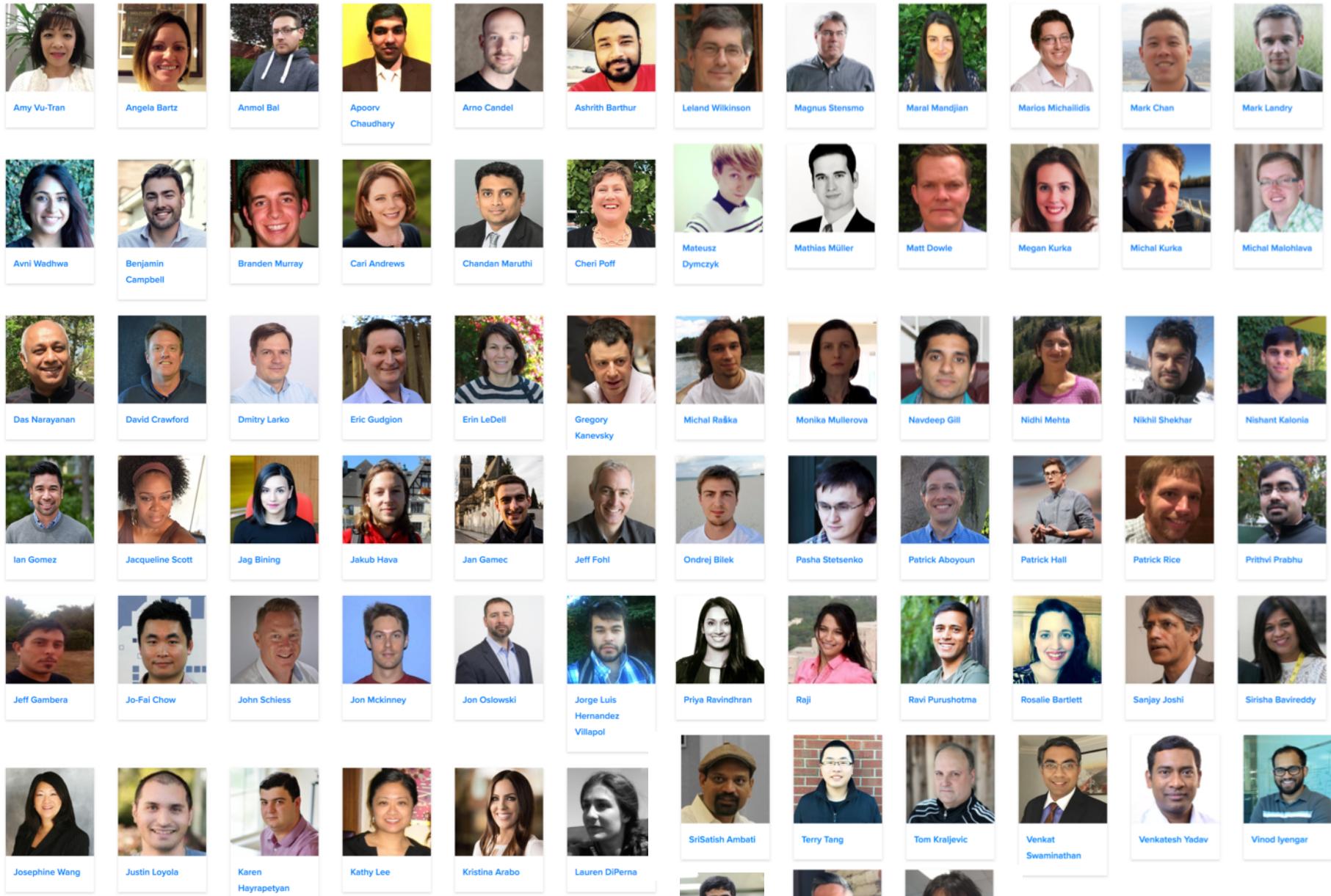
# Avengers: Infinity War Character Map



# MACHINE LEARNING



We develop  
machine learning platforms



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

# 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)

As of January 2018

© Gartner, Inc

# 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)

© Gartner, Inc

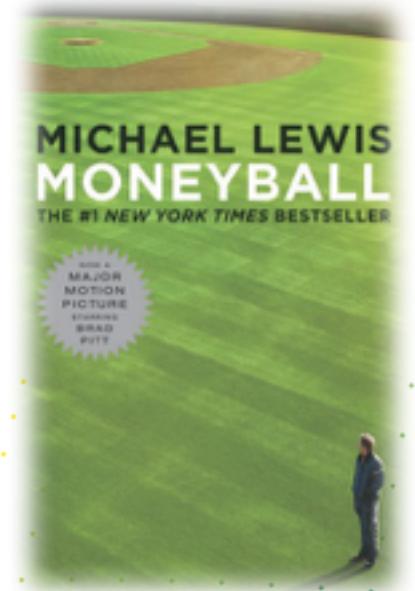
H<sub>2</sub>O.ai

# Moneyball: The Multimillion-Dollar Business Problem

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

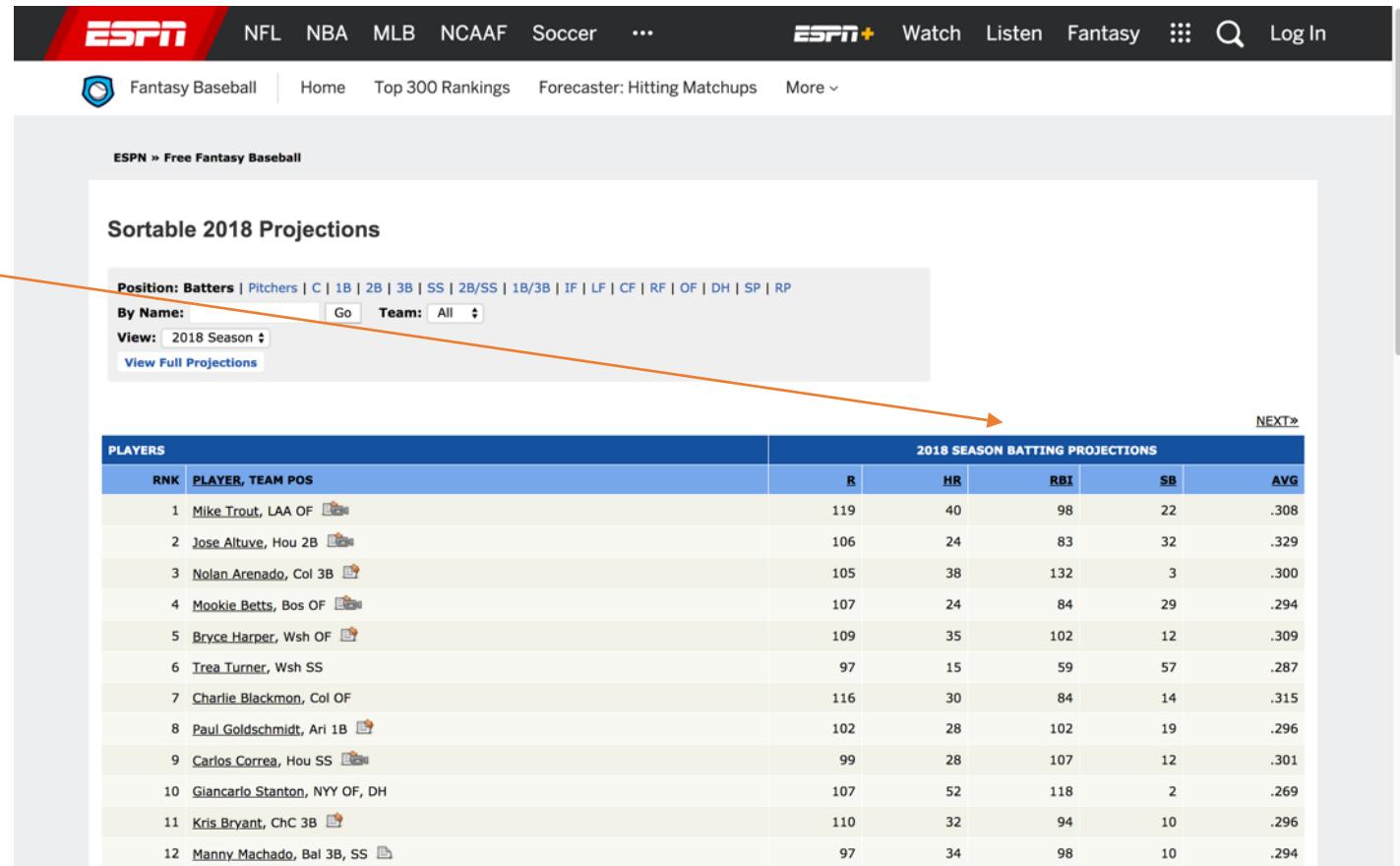


Source: Moneyball, 2011 Columbia Pictures



# 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 4 to 5-year when they sign a new contract.
- MLB teams need to evaluate players' long-term performance (i.e. > 1 year)



The screenshot shows the ESPN Fantasy Baseball website with the URL [fantasy.espn.com/free/baseball/projections?scoringType=batting](https://fantasy.espn.com/free/baseball/projections?scoringType=batting). The page title is "Sortable 2018 Projections". It features a search bar for "Position: Batters | Pitchers | C | 1B | 2B | 3B | SS | 2B/SS | 1B/3B | IF | LF | CF | RF | OF | DH | SP | RP" and filters for "By Name:" and "Team: All". A dropdown menu shows "View: 2018 Season" and a link to "View Full Projections". The main content is a table titled "PLAYERS" with columns: RNK, PLAYER, TEAM POS, R, HR, RBI, SB, and AVG. The table lists 12 players with their 2018 season batting projections. An orange arrow points from the text "Existing Forecasts (e.g. ESPN) are usually projections for the next year only." to the "View: 2018 Season" dropdown.

PLAYERS		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, Wsh OF 	109	35	102	12	.309
6	Trea Turner, Wsh 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, ChC 3B 	110	32	94	10	.296
12	Manny Machado, Bal 3B, SS 	97	34	98	10	.294

# Our Solution

- Open data – Lahman Database.
- Proprietary data from Ari Kaplan – our real Moneyball guy.
- Framed data for ML.
- Used H<sub>2</sub>O AutoML to make predictions for next three years.
- Created a Shiny app for quick navigation.
- Ari used the app to look at predictions for some free agents. He found an undervalued player and recommended that player to his team.
- The rest is history.



# In case you're wondering... final project result

\$20M

trade 2 weeks prior to the  
season beginning



# Live Demo

**Moneyball Demo**

- [Introduction](#)
- [Results \(Pitching\)](#) Demo
- [Results \(Batting\)](#)
- [About Us](#)
- [YouTube](#)

**Pitching Performance: Player Stats (Up to 2017) and Projection (2018-2020)**

Name	Team (2017)	Weight	Height	Bats	Throws	Birth Country	Birth Year	Debut Year
Chris Sale	BOS	180	78	L	L	USA	1989	2010

**Notes:**

1. Training Period: 2010 to 2015.
2. Validation Period: 2016 and 2017.
3. Projection Period: 2018 to 2020.

[Charts](#) [Table](#) [Explanation \(ERA\)](#) [Explanation \(AVG\)](#) [Explanation \(WHIP\)](#)

**Earned Run Average ERA: Lower = Better**

2010 2015 2020

**Average Allowed AVG: Lower = Better**

2010 2015 2020

**Walk+Hits per Inning Pitched WHIP: Lower = Better**

2010 2015 2020

**Green: Predictions based on Lahman only**

**Orange: Predictions based on AriDB + Lahman**

**Moneyball Demo**

- [Introduction](#)
- [Results \(Pitching\)](#) Demo
- [About Us](#)
- [YouTube](#)

**Pitching Performance: Player Stats (Up to 2017) and Projection (2018-2020)**

Name	Team (2017)	Weight	Height	Bats	Throws	Birth Country	Birth Year	Debut Year
Chris Sale	BOS	180	78	L	L	USA	1989	2010

**Notes:**

1. Training Period: 2010 to 2015.
2. Validation Period: 2016 and 2017.
3. Projection Period: 2018 to 2020.

[Charts](#) [Table](#) [Explanation \(ERA\)](#) [Explanation \(AVG\)](#) [Explanation \(WHIP\)](#)

Data	Year	ERA (Historical Data)	ERA (Predictions based on Ari_DB)	ERA (Predictions based on Lahman)	AVG (Historical Data)	AVG (Predictions based on Ari_DB)	AVG (Predictions based on Lahman)	WHIP (Historical Data)	WHIP (Predictions based on Ari_DB)	WHIP (Predictions based on Lahman)
Training	2011	2.790			0.203			1.13		
Training	2012	3.000			0.236			1.35		
Training	2013	3.070			0.230			1.07		
Training	2014	2.170			0.206			0.86		
Training	2015	3.410			0.233			1.68		
Validation	2016	3.340	3.060	3.890	0.227	0.229	0.251	1.037	1.050	1.273
Validation	2017	2.900	2.950	3.470	0.208	0.225	0.251	0.970	1.010	1.223
Prediction	2018		2.910	3.610		0.214	0.242		0.956	1.315
Prediction	2019		2.720	3.820		0.210	0.234		0.930	1.287
Prediction	2020		2.620	4.100		0.203	0.242		0.894	1.281

**Moneyball Demo**

- [Introduction](#)
- [Results \(Pitching\)](#) Demo
- [About Us](#)
- [YouTube](#)

**Pitching Performance: Player Stats (Up to 2017) and Projection (2018-2020)**

Name	Team (2017)	Weight	Height	Bats	Throws	Birth Country	Birth Year	Debut Year
Chris Sale	BOS	180	78	L	L	USA	1989	2010

**Notes:**

1. Training Period: 2010 to 2015.
2. Validation Period: 2016 and 2017.
3. Projection Period: 2018 to 2020.

[Charts](#) [Table](#) [Explanation \(ERA\)](#) [Explanation \(AVG\)](#) [Explanation \(WHIP\)](#)

**Chris Sale - ERA 2018 Projection**

0.20 0.21 0.22 0.23 0.24 0.25

**Chris Sale - AVG 2018 Projection**

0.90 0.91 0.92 0.93 0.94 0.95

**Chris Sale - WHIP 2018 Projection**

0.80 0.81 0.82 0.83 0.84 0.85

# More Details

- Moneyball team
- How to frame the Moneyball problem for automatic machine learning
- Tools: H<sub>2</sub>O AutoML, LIME, Shiny ...

# The Moneyball Team



**David Kearns**

PM @ IBM Data Science



**Ari Kaplan**

Mr. Moneyball @ Aginity



**Jo-Fai Chow**

Data Scientist @ H<sub>2</sub>O.ai



# Ari Kaplan – the Real Moneyball

- The real characters in the movie (Billy Beane and Paul DePodesta) did not want to work with Hollywood.
- The filmmaker interviewed Ari instead and created the Paul character based on Ari's real-life story.
- Ari happens to work at Aginity so we have a real "Moneyball" for this project.



# A Proof-of-Concept Demo for IBM Think Conference Talk



IBM Data Science @IBMDatascience

Following

@DaithiOCiaran @arikaplan1 & @matlabulous are smoothly passing the mic back & forth to talk about their joint Moneyball project to a jam-packed room. #think2018 #machinelearning @Aginity @h2oai #DSX

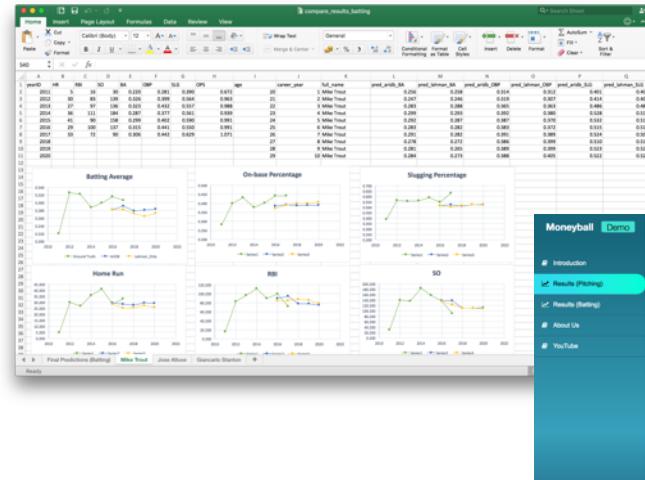


9:04 PM - 22 Mar 2018



# From PoC Demo to Real Moneyball

- **March 19** – AutoML Predictions finalized.  
Initial presentation in Excel.
- **March 20** – Version 1 of Shiny app. Ari used to app to validate some players he had in mind and recommended one player to his team.
- **March 21** – Multimillion-dollar contract finalized.
- **March 22** – Moneyball presentation at IBM Think



# Framing the Business Problem for Machine Learning

Code on GitHub (without Ari's proprietary data)

<https://github.com/woobe/moneyball>

# Lahman Data

Player's information

birthYear	birthMonth	birthDay	birthCountry	birthState	birthCity					
1991	8	7	USA	NJ	Vineland					
nameFirst	nameLast	nameGiven	weight	height	bats	throws	debut	finalGame	retroID	bbrefID
Mike	Trout	Michael Nelson	235	74	R	R	2011-07-08	2017-10-01	troum001	troutmi01

Player's past performance (batting in this case)

playerID	yearID	stint	teamID	lgID	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	
95484	troutmi01	2011	1	LAA	AL	40	123	20	27	6	0	5	16	4	0	9	30	0	2	0	1	2
96904	troutmi01	2012	1	LAA	AL	139	559	129	182	27	8	30	83	49	5	67	139	4	6	0	7	7
98308	troutmi01	2013	1	LAA	AL	157	589	109	190	39	9	27	97	33	7	110	136	10	9	0	8	8
99744	troutmi01	2014	1	LAA	AL	157	602	115	173	39	9	36	111	16	2	83	184	6	10	0	10	6
101226	troutmi01	2015	1	LAA	AL	159	575	104	172	32	6	41	90	11	7	92	158	14	10	0	5	11
102712	troutmi01	2016	1	LAA	AL	159	549	123	173	32	5	29	100	30	7	116	137	12	11	0	5	5
104195	troutmi01	2017	1	LAA	AL	114	402	92	123	25	3	33	72	22	4	94	90	15	7	0	4	8

# Lahman Data Framed as a ML problem

yearID	teamID	lgID	weight	height	bats	throws	birthYear	birthCountry	birthState	birthCity	age	career_year
2011	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	20	1
2012	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	21	2
2013	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	22	3
2014	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	23	4
2015	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	24	5
2016	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	25	6
2017	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	26	7
2018	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	27	8
2019	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	28	9
2020	LAA	AL	235	74	R	R	1991	USA	NJ	Vineland	29	10

Player  
Attributes

last1_HR	last2_HR	last3_HR	last4_HR	last5_HR	avg_last2_HR	avg_last3_HR	avg_last4_HR	avg_last5_HR
NA	NA	NA	NA	NA	Nan	Nan	Nan	Nan
5	NA	NA	NA	NA	5.0	5.00000	5.00000	5.00000
30	5	NA	NA	NA	17.5	17.50000	17.50000	17.50000
27	30	5	NA	NA	28.5	20.66667	20.66667	20.66667
36	27	30	5	NA	31.5	31.00000	24.50000	24.50000
41	36	27	30	5	38.5	34.66667	33.50000	27.80000
29	41	36	27	30	35.0	35.33333	33.25000	32.60000
33	29	41	36	27	31.0	34.33333	34.75000	33.20000
33	33	29	41	36	33.0	31.66667	34.00000	34.40000
33	33	33	29	41	33.0	33.00000	32.00000	33.80000

One of the Targets

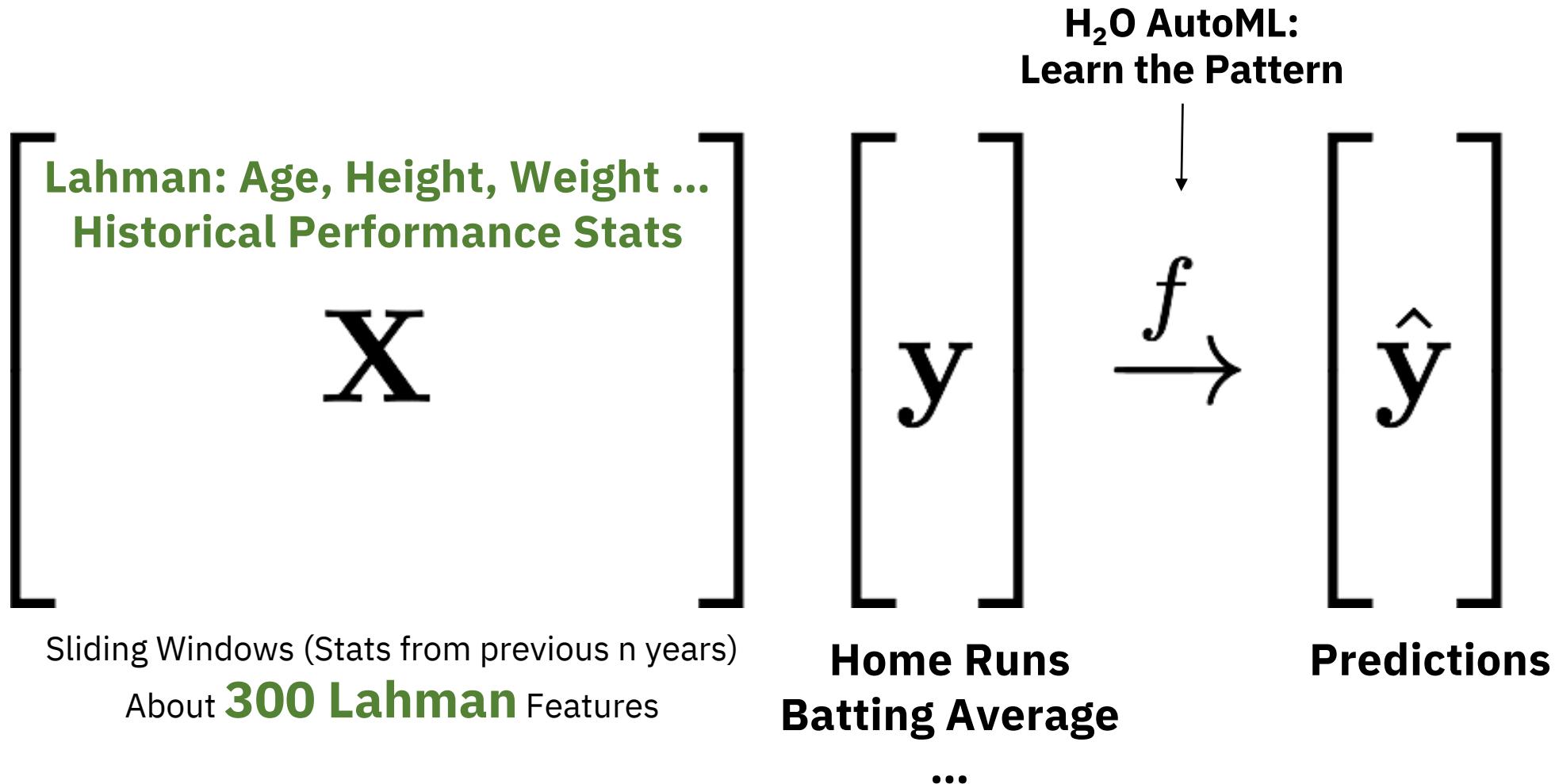
yearID	HR
2011	5
2012	30
2013	27
2014	36
2015	41
2016	29
2017	33
2018	NA
2019	NA
2020	NA

Training  
Validation  
Forecast

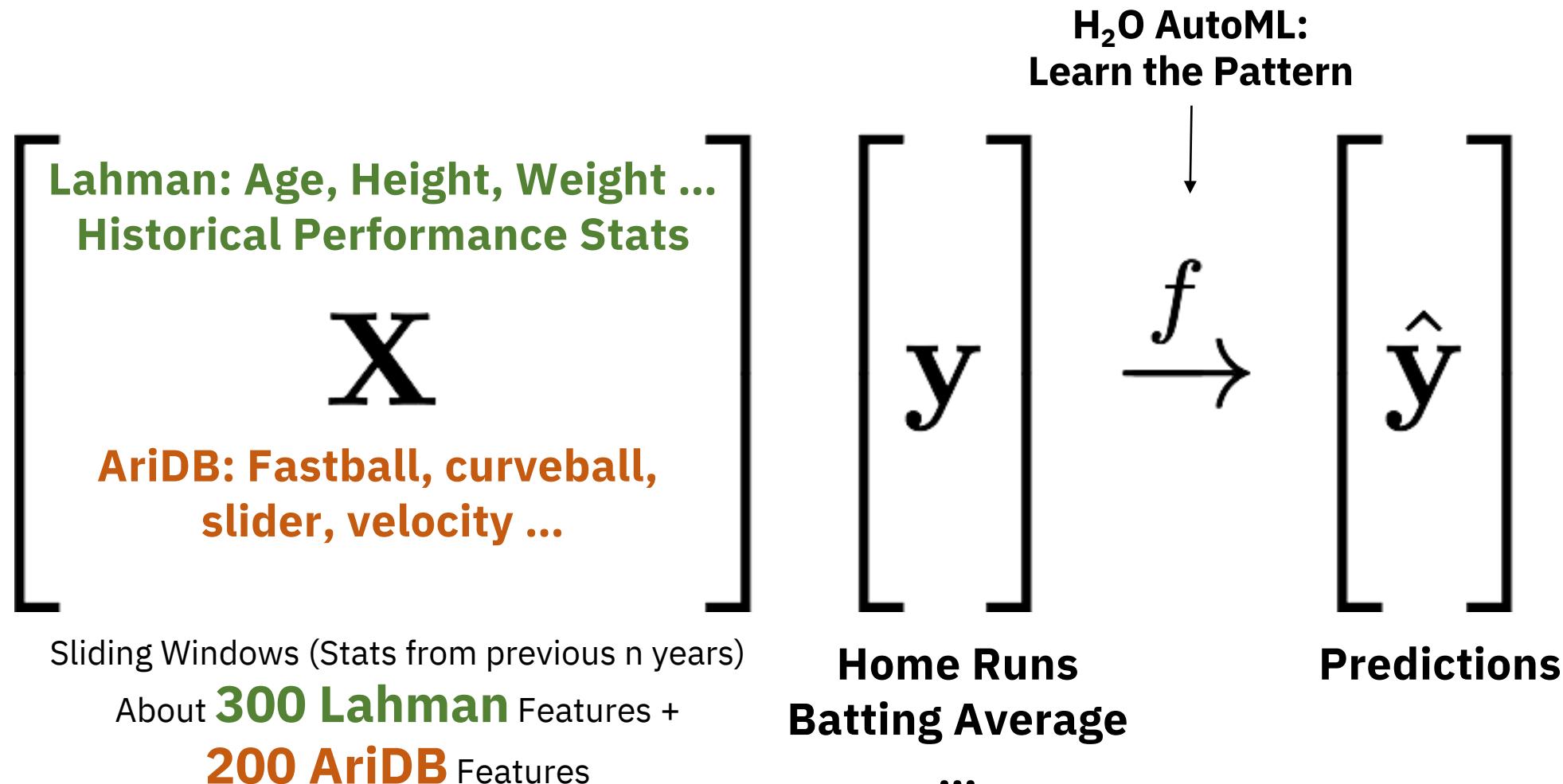
No data. Used 2017 value. Not perfect (a quick hack).

Past  
Performance  
Sliding  
Windows  
+  
Other  
Stats

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



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



# Live Demo

**Moneyball Demo**

Introduction  
Results (Pitching) **Results (Batting)**  
About Us  
YouTube

**Pitching Performance: Player Stats (Up to 2017) and Projection (2018-2020)**

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

Select a Player: Chris Sale

Name	Team (2017)	Weight	Height	Bats	Throws	Birth Country	Birth Year	Debut Year
Chris Sale	BOS	180	78	L	L	USA	1989	2010

Notes:  
1. Training Period: 2010 to 2015.  
2. Validation Period: 2016 and 2017.  
3. Projection Period: 2018 to 2020.

Charts Table Explanation (ERA) Explanation (AVG) Explanation (WHIP)

**Earned Run Average ERA: Lower = Better**

**Average Allowed AVG: Lower = Better**

**Walk+Hits per Inning Pitched WHIP: Lower = Better**

Green: Predictions based on Lahman only

Orange: Predictions based on AriDB + Lahman

Year	Data	Pred_Lahman	Pred_AriDB
2010	2.8	2.8	2.8
2011	3.0	3.0	3.0
2012	3.0	3.0	3.0
2013	2.6	2.6	2.6
2014	3.4	3.4	3.4
2015	3.3	3.3	3.3
2016	3.0	3.0	3.0
2017	2.9	2.9	2.9
2018	2.8	2.8	2.8
2019	2.7	2.7	2.7
2020	2.6	2.6	2.6

Year	Data	Pred_Lahman	Pred_AriDB
2010	0.21	0.21	0.21
2011	0.23	0.23	0.23
2012	0.22	0.22	0.22
2013	0.20	0.20	0.20
2014	0.23	0.23	0.23
2015	0.22	0.22	0.22
2016	0.21	0.21	0.21
2017	0.22	0.22	0.22
2018	0.21	0.21	0.21
2019	0.20	0.20	0.20
2020	0.19	0.19	0.19

Year	Data	Pred_Lahman	Pred_AriDB
2010	1.1	1.1	1.1
2011	1.1	1.1	1.1
2012	1.0	1.0	1.0
2013	0.9	0.9	0.9
2014	1.0	1.0	1.0
2015	1.1	1.1	1.1
2016	1.2	1.2	1.2
2017	1.25	1.25	1.25
2018	1.3	1.3	1.3
2019	1.25	1.25	1.25
2020	1.2	1.2	1.2

# H<sub>2</sub>O AutoML and LIME Materials for eRum Workshop

---

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



Jo-fai (Joe) Chow  
Data Science Evangelist /  
Community Manager

joe@h2o.ai

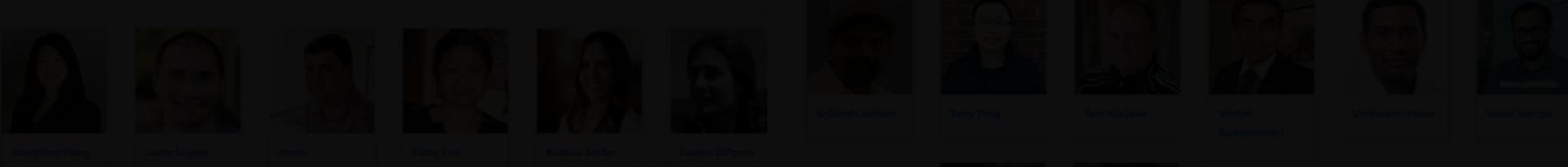
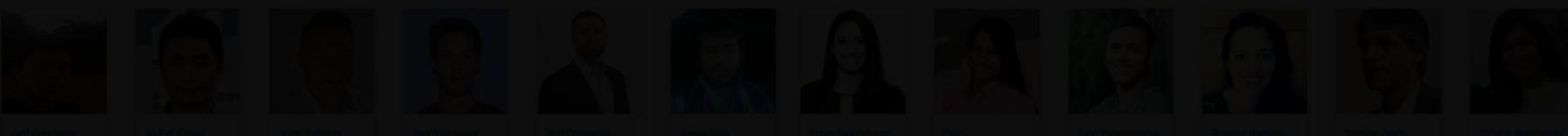
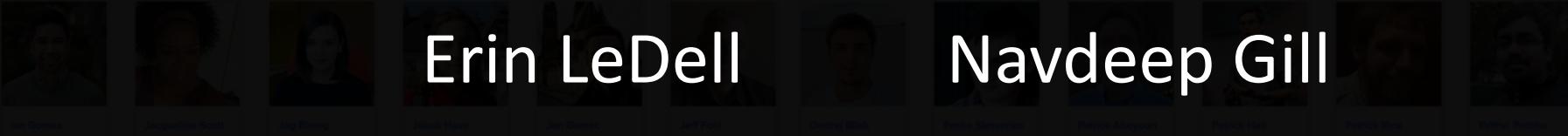
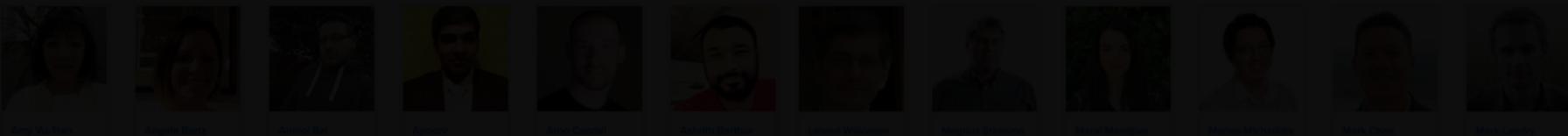
@matlabulous

Download → [https://bit.ly/  
\*\*joe\\_eRum\\_2018\*\*](https://bit.ly/joe_eRum_2018)

# About 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>



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

Erin LeDell

Navdeep Gill

H<sub>2</sub>O Team

# H<sub>2</sub>O-3 Algorithms Overview

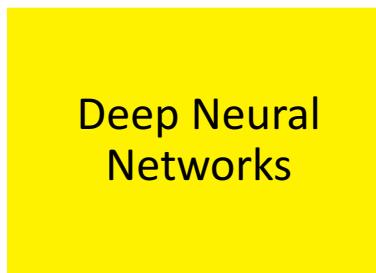
## Supervised Learning



- **Generalized Linear Models:** Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**



- **Distributed Random Forest:** Classification or regression models
- **Gradient Boosting Machine:** Produces an ensemble of decision trees with increasing refined approximations

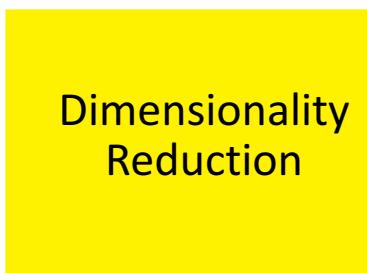


- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

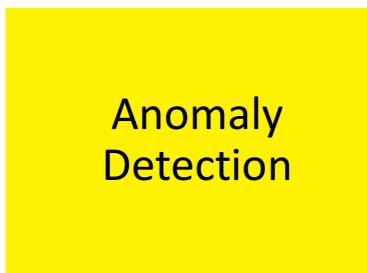
## Unsupervised Learning



- **K-means:** Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

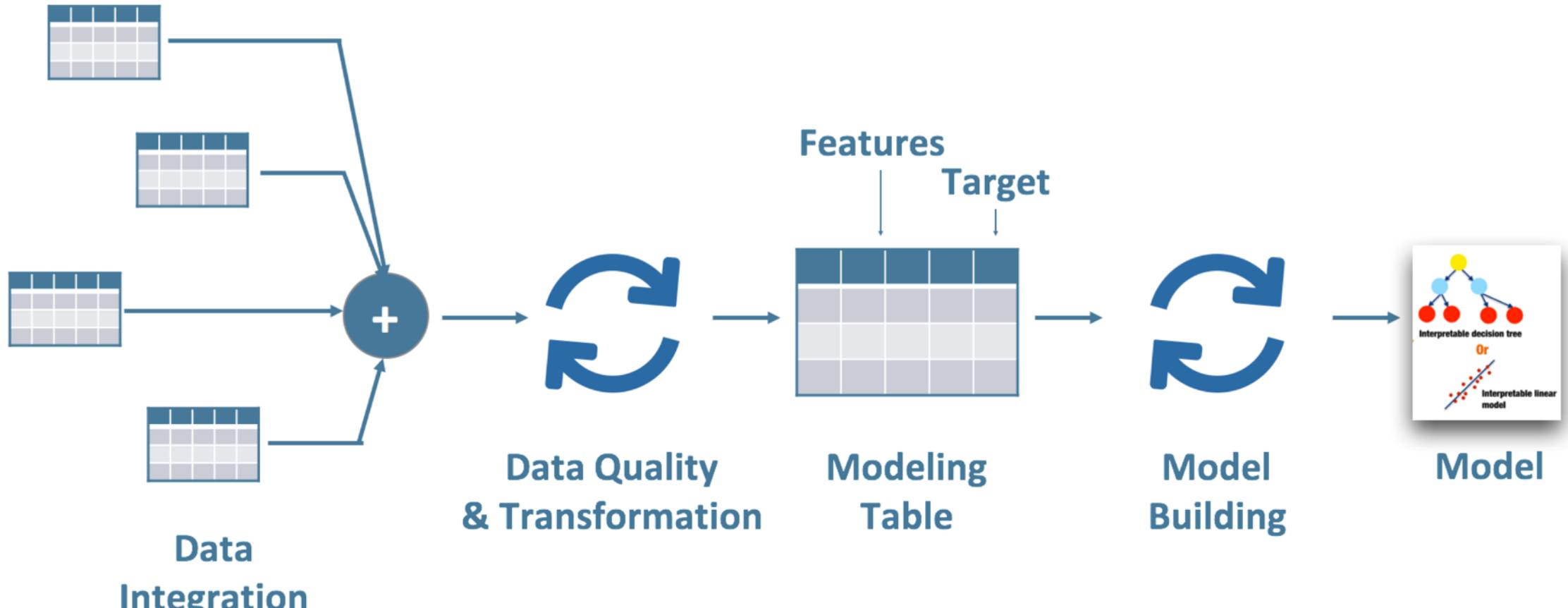


- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

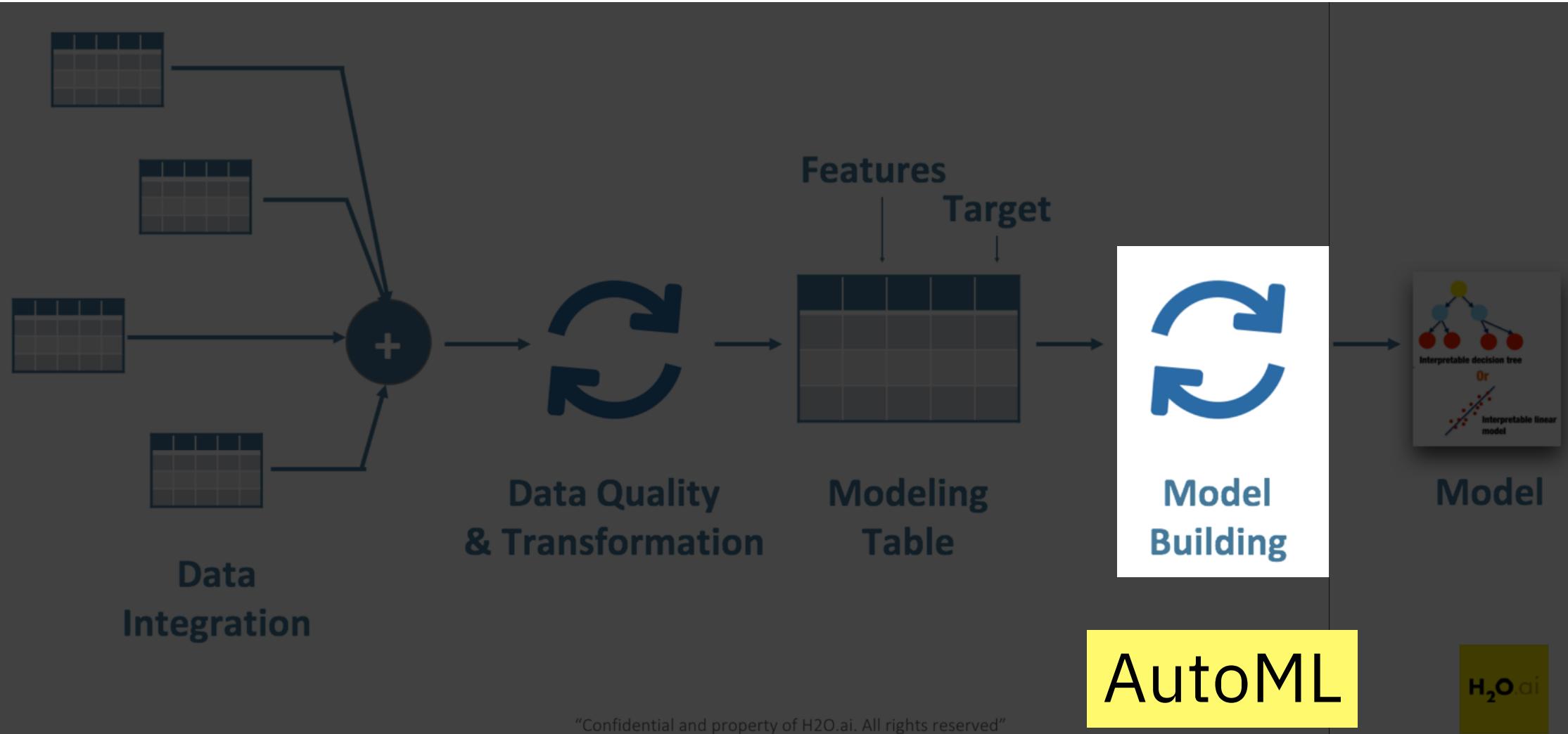


- **Autoencoders:** Find outliers using a nonlinear dimensionality reduction using deep learning

# Typical Enterprise Machine Learning Workflow



# Typical Enterprise Machine Learning Workflow



# AutoML Code

```
# H2O AutoML with Lahman only
automl_lahman = h2o.automl(x = features,
                            y = targets[n_target],
                            training_frame = h_train,
                            validation_frame = h_valid,
                            max_models = 10, # increase this to allow more models
                            max_runtime_secs = 120, # increase this to allow more time
                            stopping_metric = "RMSE",
                            stopping_rounds = 3,
                            seed = n_seed,
                            exclude_algos = c("DeepLearning"), # you can exclude any algo
                            project_name = paste0("AutoML_Lahman", targets[n_target]))
```

# AutoML Results

```
H2OResgressionMetrics: stackedensemble
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

MSE: 0.00246453
RMSE: 0.04964404
MAE: 0.03335875
RMSLE: 0.04124294
Mean Residual Deviance : 0.00246453
```

Slot "leaderboard":

		model_id	mean_residual_deviance	rmse	mae	rmsle
1	StackedEnsemble_BestOfFamily_0_AutoML_20180615_040834		0.002465	0.049644	0.033359	0.041243
2	StackedEnsemble_AllModels_0_AutoML_20180615_040834		0.002467	0.049669	0.033367	0.041265
3	GLM_grid_0_AutoML_20180615_040834_model_0		0.002480	0.049802	0.033560	0.041401
4	GBM_grid_0_AutoML_20180615_040834_model_4		0.002486	0.049856	0.033707	0.041373
5	GBM_grid_0_AutoML_20180615_040834_model_2		0.002564	0.050638	0.034346	0.042008
6	GBM_grid_0_AutoML_20180615_040834_model_1		0.002569	0.050684	0.034261	0.042022

[12 rows x 5 columns]

# About Machine Learning Interpretability

LIME (Local Interpretable Model-Agnostic Explanations)

... and more

# 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

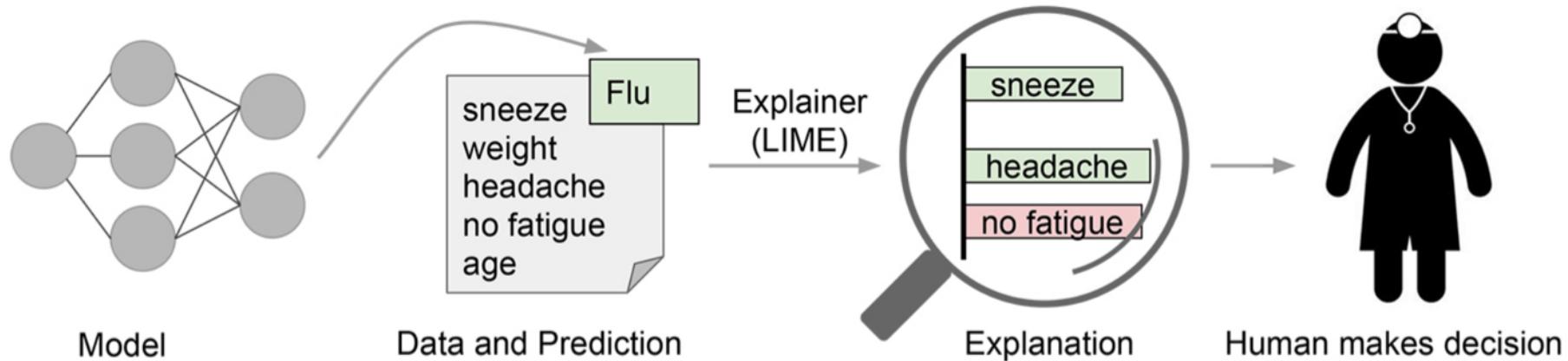


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

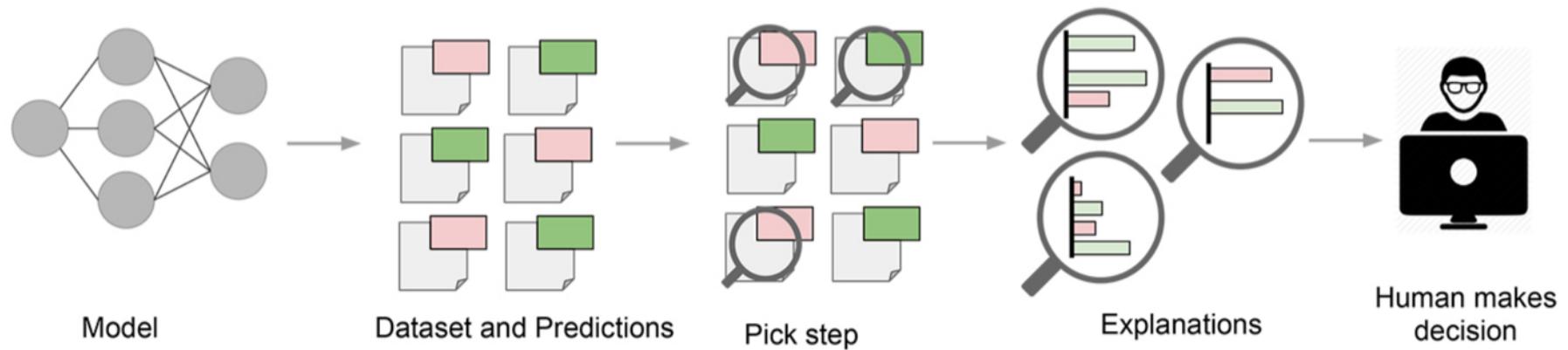


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# 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.

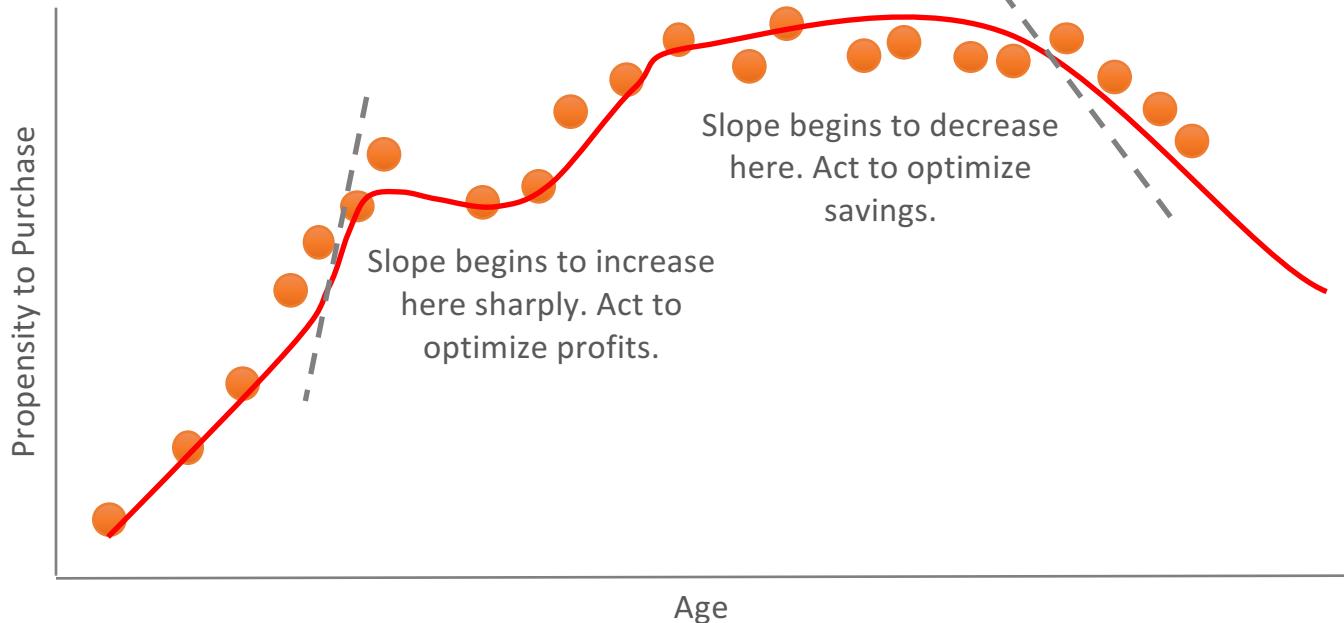
## Linear Models

*Exact explanations for approximate models.*



## Machine Learning

*Approximate explanations for exact models.*



## 4 Explain the Model

### 4.1 Step 1: Create an `explainer`

```
explainer = lime::lime(x = as.data.frame(h_train[, features]),
                      model = model_automl@leader)
```

### 4.2 Step 2: Turn `explainer` into `explanations`

```
# Extract one sample (change `1` to any row you want)
d_samp = as.data.frame(h_test[1, features])
```

```
# Assign a specific row name (for better visualization)
row.names(d_samp) = "Sample 1"
```

```
# Create explanations
explanations = lime::explain(x = d_samp,
                             explainer = explainer,
                             n_permutations = 5000,
                             feature_select = "auto",
                             n_features = 13) # Look top x features
```

### 4.3 Look at Explanations (Bar Chart)

```
lime::plot_features(explanations, ncol = 1)
```

# Live Demo

**Moneyball Demo**

Introduction Results (Pitching) Results (Batting) About Us YouTube

**Select a Player**

Chris Sale

**Notes:**

1. Training Period: 2010 to 2015.
2. Validation Period: 2016 and 2017.
3. Projection Period: 2018 to 2020.

Charts Table Explanation (ERA) **Explanation (AVG)** Explanation (WHIP)

Case: Chris Sale - ERA 2018 Projection  
Prediction: 2.91430358345192  
Explanation Fit: 0.33

122 < last1\_SO  
2012 < yearID  
114 < avg\_last2\_SO  
111.3 < avg\_last3\_SO  
110.5 < avg\_last4\_SO  
1 < last2(CG)  
110 < avg\_last5\_SO  
teamID = BOS  
0.5 < avg\_last2(CG)  
16 < last5(GIDP)  
last1\_BK <= 1  
6.0 < avg\_last2(HBP)  
26.0 < last3\_G <= 32.0  
452 < avg\_last4(IPouts)  
last3\_BK <= 1

Case: Chris Sale - ERA 2019 Projection  
Prediction: 2.72093970569747  
Explanation Fit: 0.30

122 < last1\_SO  
2012 < yearID  
114 < avg\_last2\_SO  
110.5 < avg\_last3\_SO  
avg\_last1\_BK <= 0.667  
last1\_BAOpp <= 0.230  
52 < last2\_R <= 85  
last1\_WHIP <= 1.19  
133.0 < last5\_SO  
34.9 < avg\_last4\_BB <= 50.2  
784 < last3\_BFP  
avg\_last3\_pitches\_per\_out <= 16.3  
avg\_last5\_BAOpp <= 0.235  
last1\_SV < 1

Case: Chris Sale - ERA 2020 Projection  
Prediction: 2.61796230813385  
Explanation Fit: 0.23

teamID = BOS  
2012 < yearID  
122 < last1\_SO  
0.607 < avg\_last3\_WPCT  
114 < avg\_last2\_SO  
last1\_WHIP <= 1.19  
2008 < debut\_year  
80 < avg\_last5\_ab\_changeup  
3.97 < avg\_last2\_pitches\_per\_pa  
avg\_last2\_pitches\_per\_out <= 16.3  
11 < last3\_W  
last3\_BK <= 1  
11 < last4\_W  
last2(CG) <= 1  
last4\_BK <= 1

Feature: last1\_SO, avg\_last2\_SO, avg\_last3\_SO, avg\_last4\_SO, avg\_last5\_SO, teamID, last2(CG), last5(GIDP), last1\_BK, avg\_last2(HBP), last3\_G, avg\_last4(IPouts), last3\_BK, last1\_BAOpp, last2\_R, last1\_WHIP, last5\_SO, avg\_last4\_BB, last3\_BFP, avg\_last3\_pitches\_per\_out, avg\_last5\_BAOpp, last1\_SV, avg\_last2\_WPCT, avg\_last2\_pitches\_per\_pa, avg\_last2\_pitches\_per\_out, last3\_W, last4\_W, last2(CG), last4\_BK

Weight: -0.2, -0.1, 0.0, 0.1

Supports Contradicts

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

### Pitching Performance: Player Stats (Up to 2017) and Projection (2018-2020)

Name	Team (2017)	Weight	Height	Bats	Throws	Birth Country	Birth Year	Debut Year
Chris Sale	BOS	180	78	L	L	USA	1989	2010

# Other H<sub>2</sub>O News

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



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



H2O AI Open Source Engine  
Integration with Spark



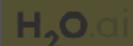
Lightning Fast machine  
learning on GPUs

DRIVERLESSAI

Automatic feature  
engineering, machine  
learning and interpretability

# Steam

Secure multi-tenant H2O clusters

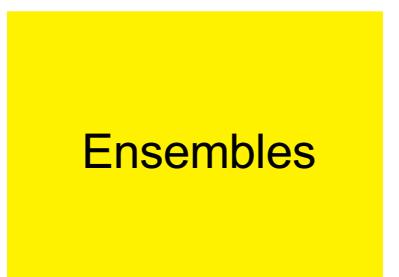


# Algorithms on H<sub>2</sub>O-3 (CPU)

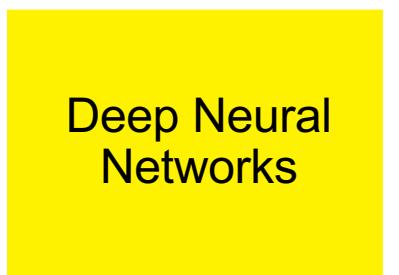
## Supervised Learning



- Generalized Linear Models: Binomial, Gaussian, Gamma, Poisson and Tweedie
- Naïve Bayes



- Distributed Random Forest: Classification or regression models
- Gradient Boosting Machine: Produces an ensemble of decision trees with increasing refined approximations

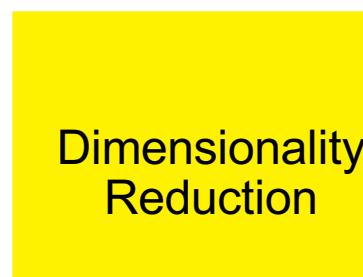


- Deep learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

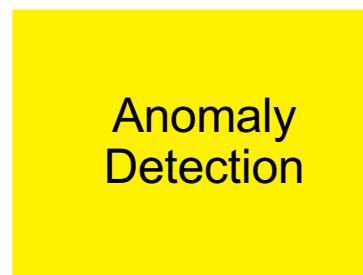
## Unsupervised Learning



- K-means: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k



- 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



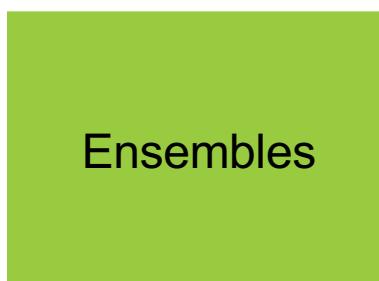
- Autoencoders: Find outliers using a nonlinear dimensionality reduction using deep learning

# Algorithms on H<sub>2</sub>O4GPU (more to come)

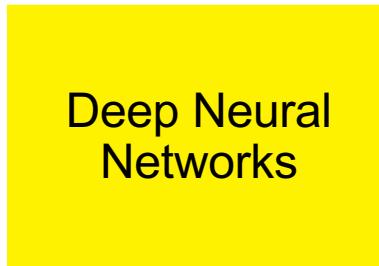
## Supervised Learning



- Generalized Linear Models: Binomial, Gaussian, Gamma, Poisson and Tweedie
- Naïve Bayes



- Distributed Random Forest: Classification or regression models
- Gradient Boosting Machine: Produces an ensemble of decision trees with increasing refined approximations

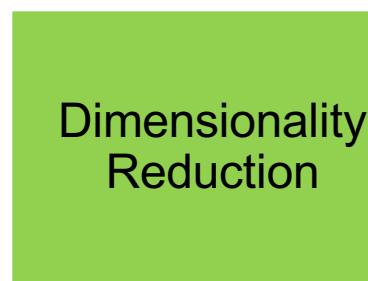


- Deep learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

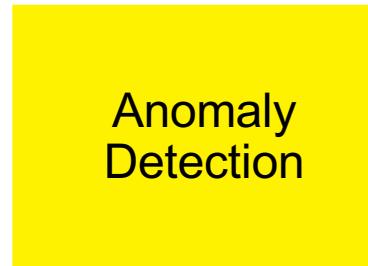
## Unsupervised Learning



- K-means: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k



- 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



- Autoencoders: Find outliers using a nonlinear dimensionality reduction using deep learning

# H2O4GPU now available in R

BY ERIN LEDELL ON MARCH 27, 2018 – 0 COMMENTS

In September, H2O.ai released a new open source software project for GPU machine learning called [H2O4GPU](#). The initial release (blog post [here](#)) included a Python module with a scikit-learn compatible API, which allows it to be used as a drop-in replacement for scikit-learn with support for GPUs on selected (and ever-growing) algorithms. We are proud to announce that the same collection of GPU algorithms is now available in R, and the `h2o4gpu` R package is available on [CRAN](#).



<https://github.com/h2oai/h2o4gpu>

# From Kaggle Grand Masters' Recipes to Production Ready in a Few Clicks

BY JO-FAI CHOW ON MAY 9, 2018 – 0 COMMENTS – EDIT

## Introducing Accelerated Automatic Pipelines in H2O Driverless AI

At H2O, we work really hard to make machine learning fast, accurate, and accessible to everyone. With H2O Driverless AI, users can leverage years of world-class, [Kaggle Grand Masters](#) experience and our GPU-accelerated algorithms ([H2O4GPU](#)) to produce top quality predictive models in a fully automatic and timely fashion.

In our most recent release (version 1.1), we are going one step further to streamline the deployment process with MOJO (Model ObjEcT, Optimized). Inherited from our popular H2O-3 platform, MOJO is a highly optimized, low-latency scoring engine that is easily embeddable in any Java environment. With automatic pipeline generation in Driverless AI, users can go from automatic machine learning to production ready in just a few clicks. This blog post illustrates the usage of MOJO in Driverless AI with a simple example.

### Easing the Pain Points in a Machine Learning Workflow

In a typical enterprise machine learning workflow, there are many things that could go wrong due to human errors, bad data science practices, different tools/infrastructure, incompatible code, lack of testing, versioning, communication and so on.

blog.h2o.ai

19  
JUN

Tuesday, June 19, 2018

## June #LondonAI: XAI, Neural Style Transfer & e-Learning (External Reg Required)

Hosted by [Jo-fai Chow](#)From [London Artificial Intelligence & Deep Learning](#)Public group 

### Details

We have a new venue for the meetup!

I love this community! Not long after I sent out the group email about the emergency change of venue, many of you reached out to me and offered help. I really appreciate this :)

Thanks to Michael James from Yoox-Net-A-Porter, we will host our meetup at their Tech Hub in White City.

Please register with your full name and email for security purposes. Here is the Eventbrite link:

<https://www.eventbrite.com/e/june-londonai-meetup-tickets-46242661044>

#### Agenda:

- 6:00 to 6:30pm Pizza and Drinks (please come early if possible)
- Introduction by H2O.ai (5 mins)
- Introduction by Yoox-Net-A-Porter Cognitive Commerce Team (15 mins)

#### Tech Talks:

- Explainable Artificial Intelligence (XAI) by Torgyn Shaikhina (20 mins)
- Neural Style Transfer by Ambroise Laurent (20 mins)
- Applying AI/ML to e-Learning by Shabbir Mookhtiar (20 mins)

You're going

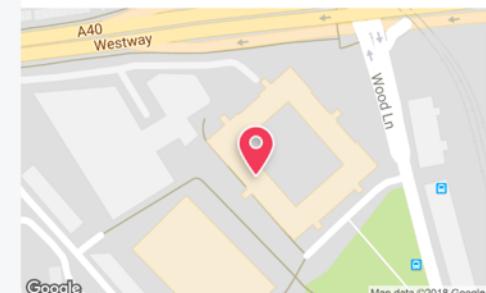


Share:    

### Organizer tools

 Tuesday, June 19, 2018  
6:00 PM to 9:00 PM  
[Add to calendar](#)

 YOOX NET A PORTER GROUP  
Tech Hub, 2nd Floor, Building 6 (The Mediaworks Building), Wood Lane, White City Place, London, W12 7TU · London



**London Meetup**  
**June 19 (Next Tuesday)**

# Danke!

- Organizers & Sponsors



- Code, Slides & Documents

- [bit.ly/h2o\\_meetups](http://bit.ly/h2o_meetups)
- [bit.ly/joe\\_eRum\\_2018](http://bit.ly/joe_eRum_2018)
- [docs.h2o.ai](http://docs.h2o.ai)

- Contact

- [joe@h2o.ai](mailto:joe@h2o.ai)
- [@matlabulous](https://twitter.com/matlabulous)
- [github.com/woobe](https://github.com/woobe)

- Please search/ask questions on  
**Stack Overflow**

- Use the tag `h2o` (not h2 zero)