The Making of a Real-World Moneyball

Finding Undervalued Players with H_2O, LIME and Shiny

Jo-fai (Joe) Chow

Data Science Evangelist / Community Manager

joe@h2o.ai

@matlabulous

More Info → https://bit.ly/

h2o_meetups

About Me

LONDON R
Tuesday 11th March 2014

Last Week – Paris

- **Before H_2O**
  - Water Engineer / EngD Researcher / Matlab Fan Boy
    (wonder why @matlabulous?)
  - Discovered R, Python, H_2O … never look back again
  - Data Scientist at Virgin Media (UK), Domino Data Lab (US)
- **At H_2O …**
  - Data Scientist / Evangelist /
  - Sales Engineer / Solution Architect /
  - Community Manager

  … The harsh reality of startup life …

Reminder: #360Selfie

H_2O.ai

H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: $75M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	• H2O Open Source Machine Learning (14,000 organizations) • H2O Driverless AI – Automatic Machine Learning
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

Worldwide Recognition in the H2O.ai Community

Open source community

222 OF THE FORTUNE 500 ♥ H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES

Paying Customers

CONFIDENTIAL
COMCAST
CISCO

G5
DIRECT MAILERS
Integral Ad Science
Nielsen Catalina Solutions

STANLEY BLACK & DECKER
H-E-B
Travelport
Walgreens
ebay
Booking.com
macy's

CREDIT SUISSE
WELLS FARGO
citi
deserve
experian.
RBC
EQUIFAX
MarketAxess
ING
DISCOVER
Capital One
PayPal

Nationwide
AEGON
XCEEDANCE
underwrite.ai
CONFIDENTIAL
TRANSAMERICA
PROGRESSIVE
ZURICH

starling
dun & bradstreet
opta
AI-ACADEMY
KAISER PERMANENTE
CHANGE HEALTHCARE
Australian Government IP Australia
ARMADA Health Care
ADP
aetna
pwc

HW Vendors | Marketing | Retail | Financial | Insurance | Healthcare | Advisory, Accounting

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

H2O.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

Source: Gartner (February 2018)

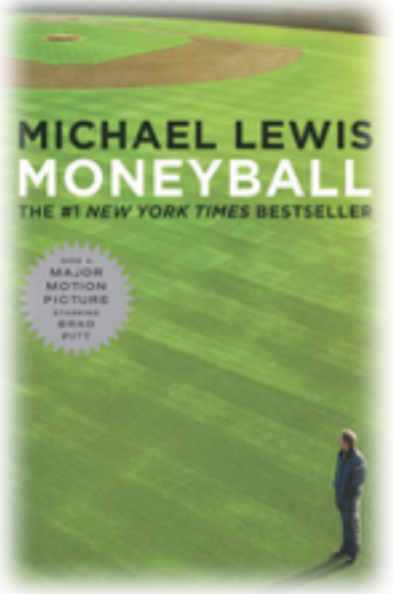
Get the Gartner Magic Quadrant [here]

H2O.ai

Moneyball:
The Multimillion-Dollar Business Problem

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

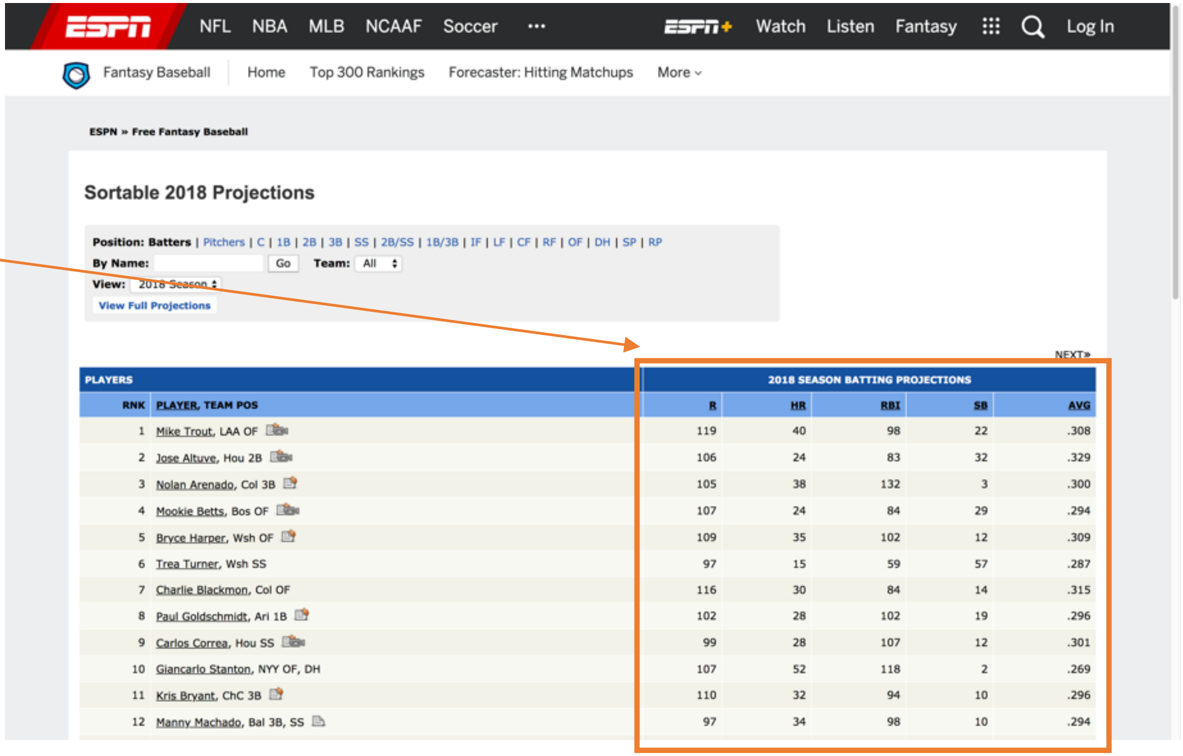
Source: Moneyball, 2011 Columbia Pictures



H2O.ai

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).


NFL NBA MLB NCAAF Soccer ··· Watch Listen Fantasy Log In
Fantasy Baseball | Home | Top 300 Rankings | Forecaster: Hitting Matchups | More
ESPN » Free Fantasy Baseball

Sortable 2018 Projections

Position: Batters | Pitchers | C | 1B | 2B | 3B | SS | 2B/SS | 1B/3B | IF | LF | CF | RF | OF | DH | SP | RP
By Name: Go Team: All
View: 2018 Season
View Full Projections

NEXT»

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

2018 SEASON BATTING PROJECTIONS

The Moneyball Team

David Kearns

PM @ IBM Data Science

Jo-Fai Chow

Data Scientist @ H_2O.ai

Ari Kaplan

Mr. Moneyball @ Aginity

IBM + aginity + H_2O.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	Tripples
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_2O AutoML

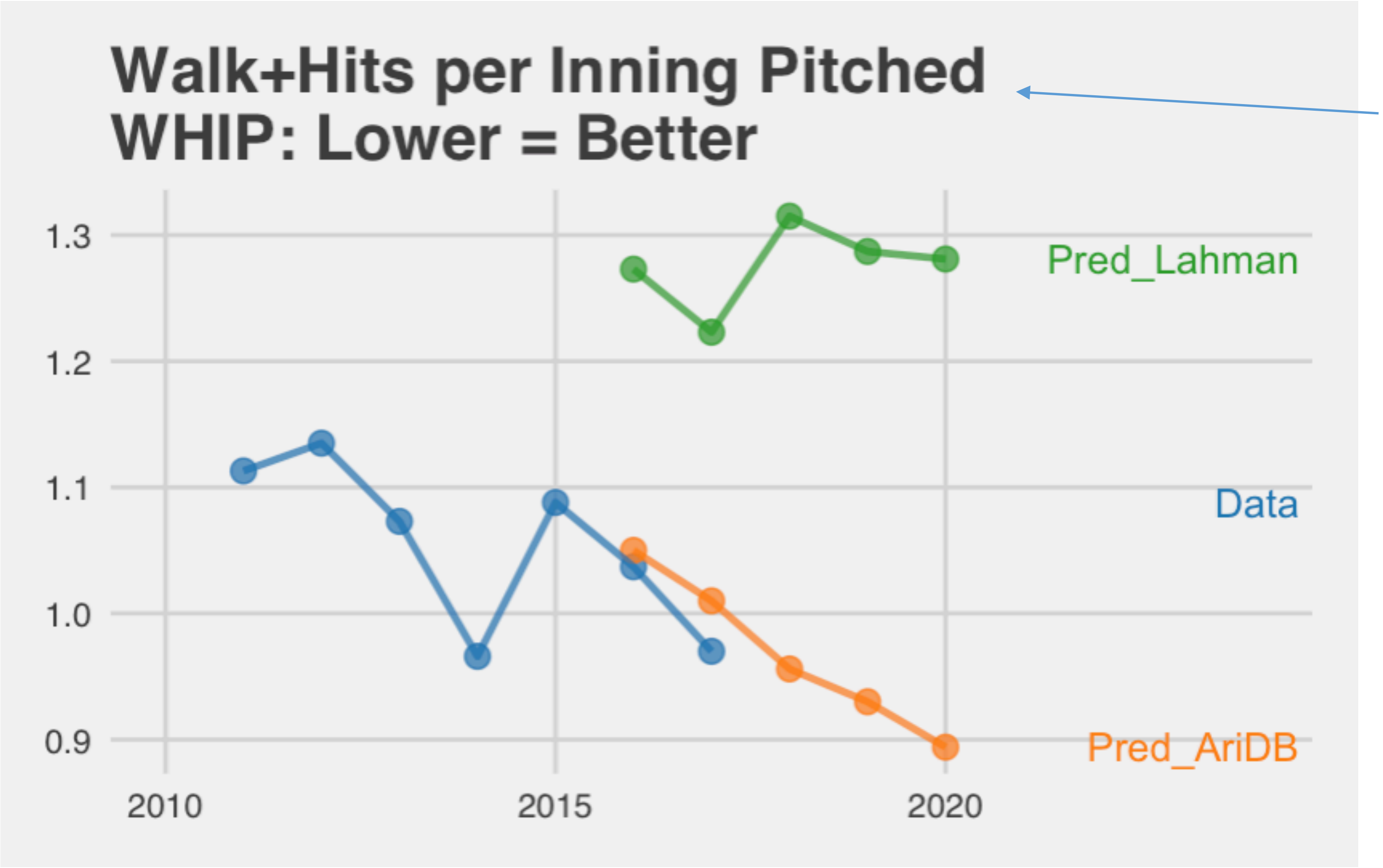
- Framed data as regression problems for performance prediction.

- Historical player performance as features.

- Used H_2O AutoML to build ensembles (linear model, random forests, gradient boosting, and deep neural networks).

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

Predictive Modelling – H_2O AutoML

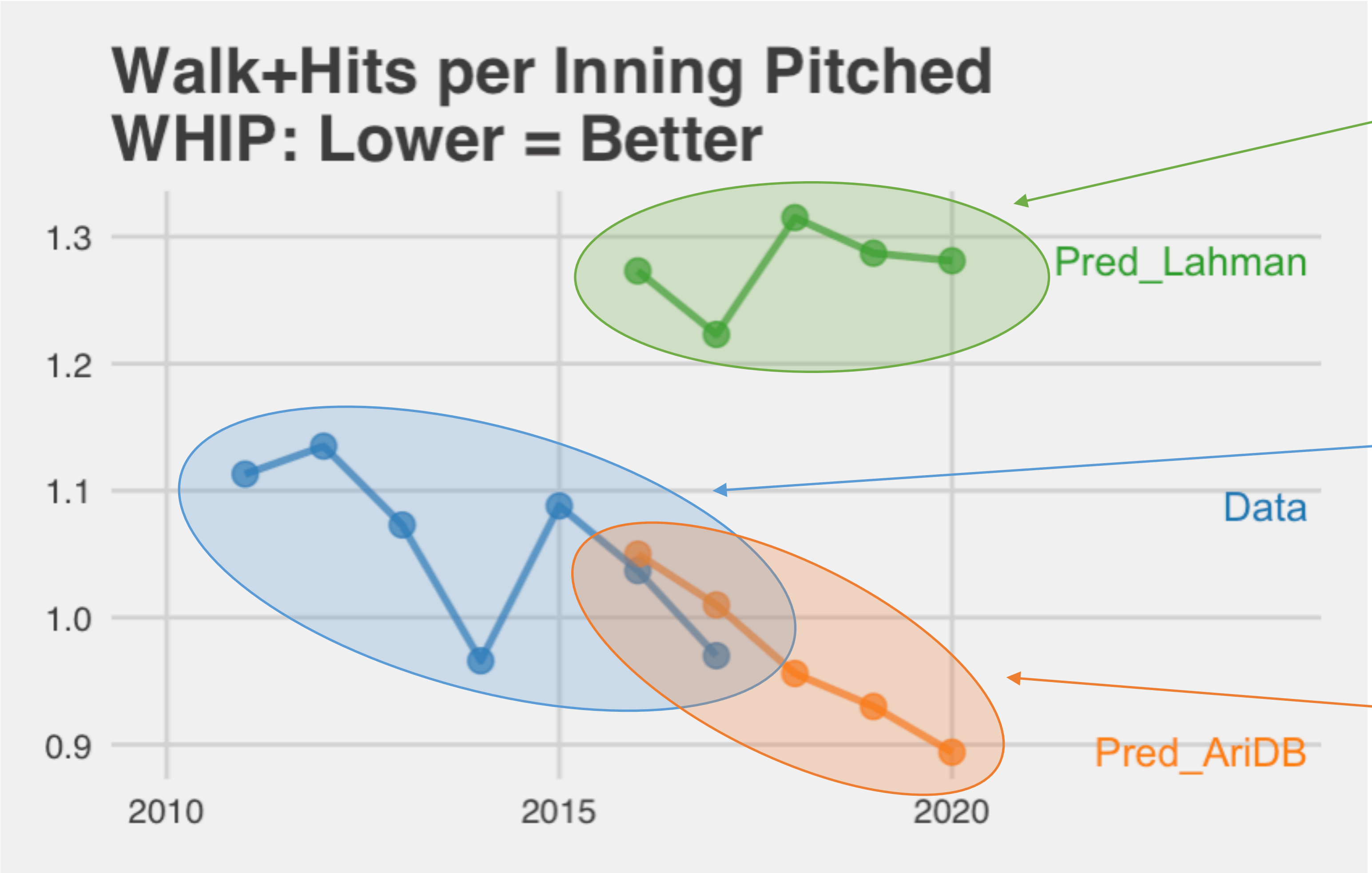


One of Many Targets
(e.g. Home Runs, Batting Average)

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


H2O.ai

Predictive Modelling – H_2O AutoML

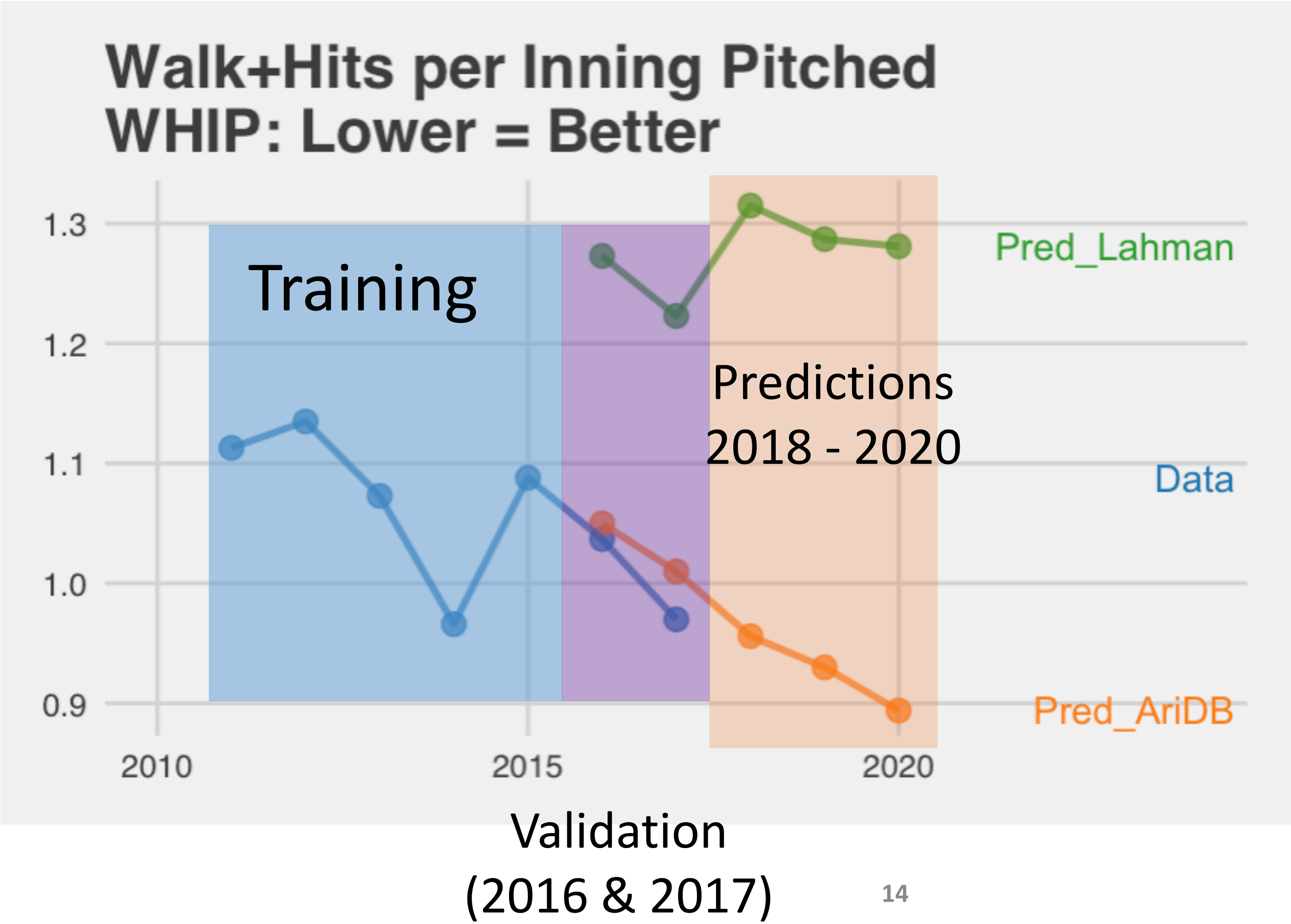


Results from models based on Lahman data only

Historical player performance data

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

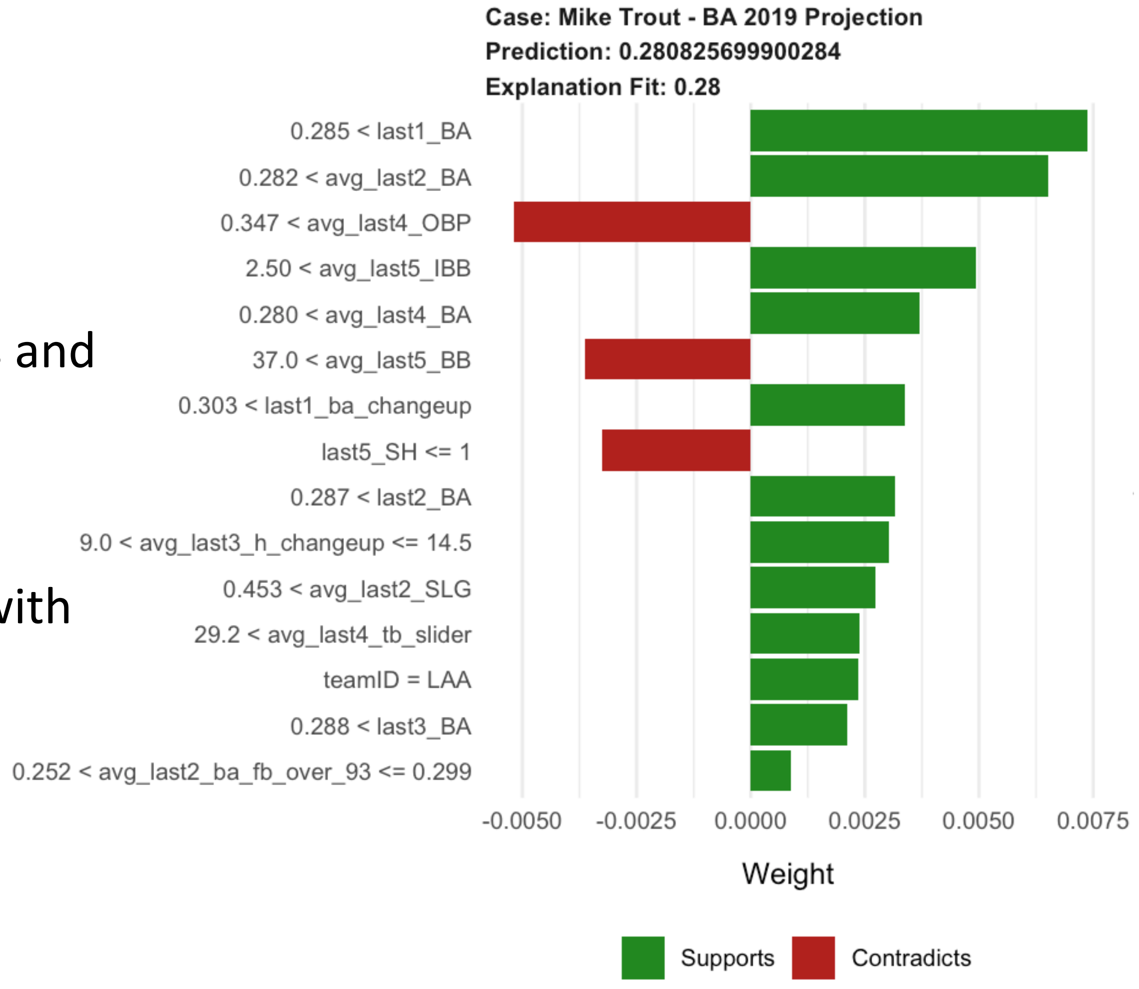
Predictive Modelling – H_2O AutoML




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

Explaining the Predictions
LIME – Local Interpretable Model-agnostic Explanations

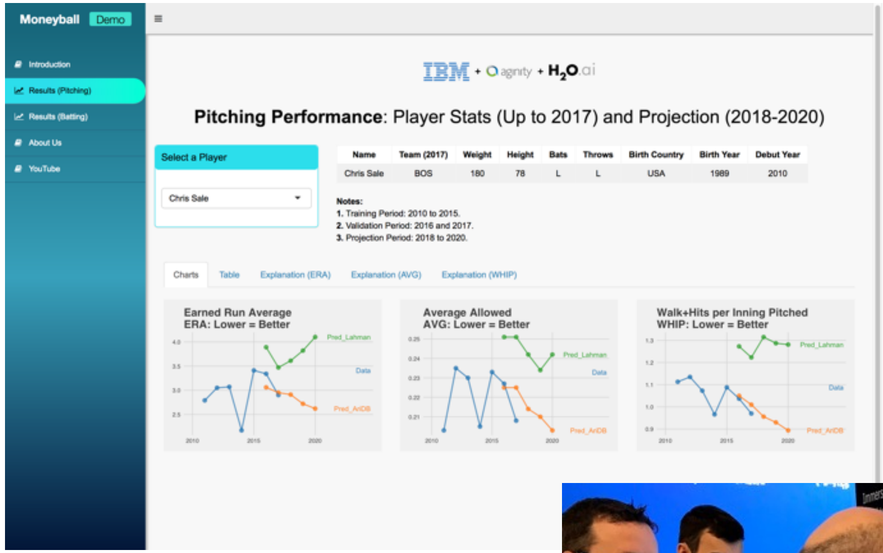
- Approximate reasoning of complex ML models (ensembles).
- Most important attributes and their contributions to the predictions.
- Ari validated the models with his 30+ years of baseball domain knowledge.
- He trusted the models.

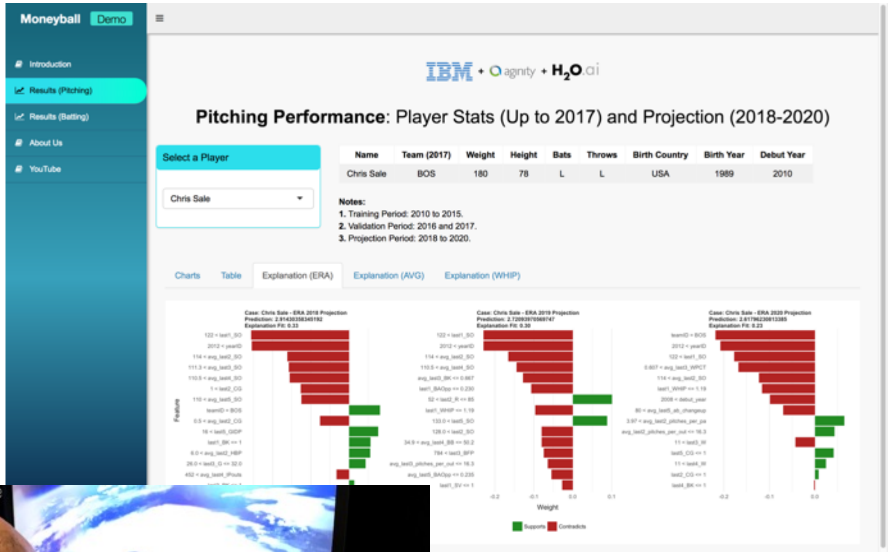



```r
# Install 'lime' from CRAN
install.packages('lime')
```

H2O.ai

Putting Everything Together – Moneyball Shiny App





Live Demo
on my laptop

In case you're wondering … final project result

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

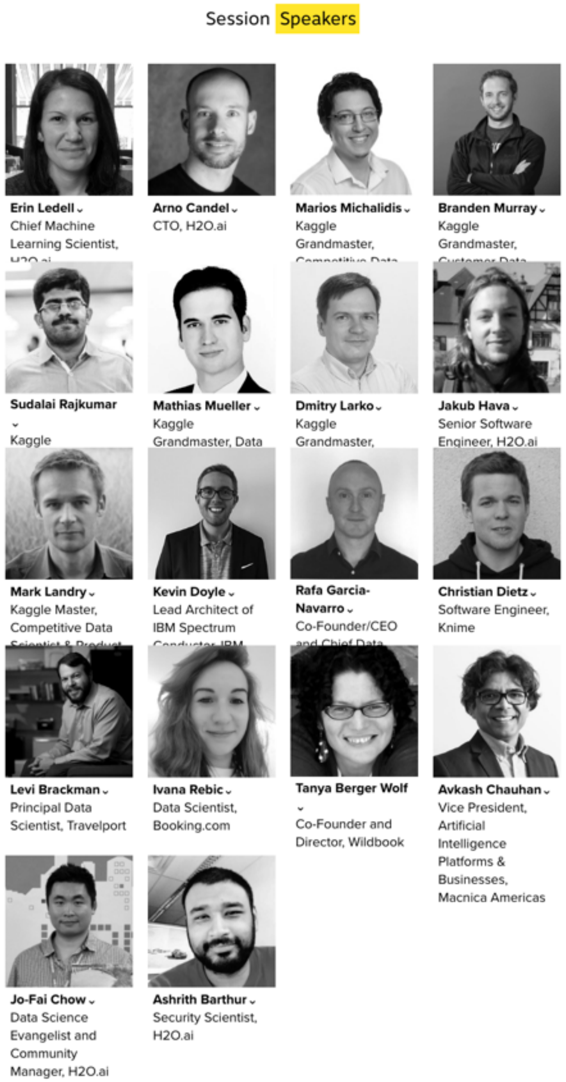
$20M

multi-year contract

finalised two weeks
before the regular season

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

29th & 30th Oct, London



More real-world use cases

+

All H_2O Kaggle Grandmasters

+

Hands-on Training

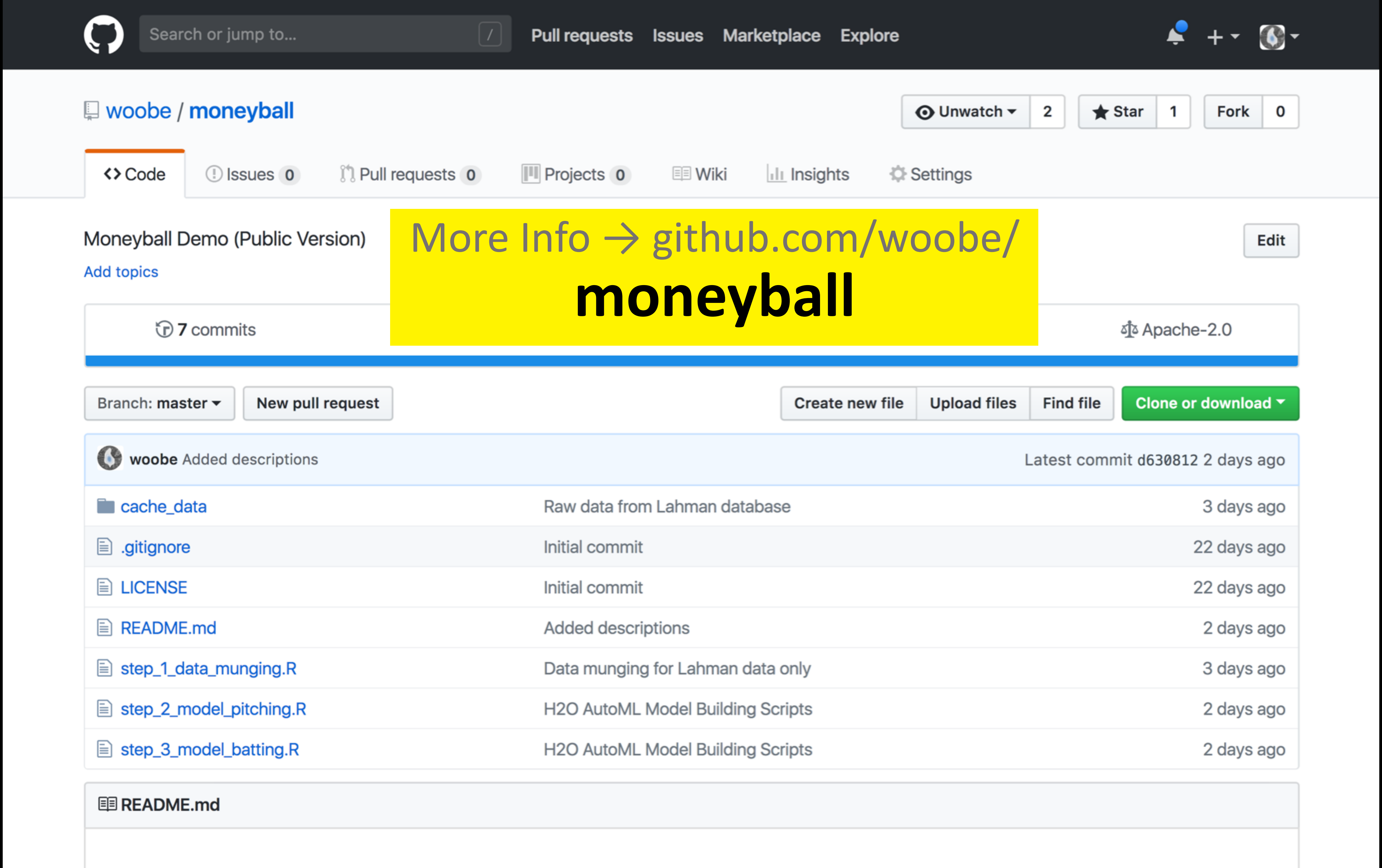
H2O.ai

Thanks!

- More Info, Code, and Slides
  - bit.ly/h2o_meetups
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe

Appendix

More Info → github.com/woobe/ **moneyball**



Approach One: Learning from **Lahman** only

H_2O AutoML:
Learn the Pattern

Lahman: Age, Height, Weight ...
Historical Performance Stats

$$\begin{bmatrix} \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{y} \end{bmatrix} \xrightarrow{f} \begin{bmatrix} \hat{\mathbf{y}} \end{bmatrix}$$

Sliding Windows (Stats from previous n years)
About **300 Lahman** Features

Home Runs
Batting Average
...

Predictions

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

H_2O AutoML: Learn the Pattern

Lahman: Age, Height, Weight …
Historical Performance Stats

$$\begin{bmatrix} \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{y} \end{bmatrix} \xrightarrow{f} \begin{bmatrix} \hat{\mathbf{y}} \end{bmatrix}$$

AriDB: Fastball, curveball, slider, velocity …

Sliding Windows (Stats from previous n years)
About **300 Lahman** Features +
200 AriDB Features

Home Runs
Batting Average
…

Predictions

H2O.ai

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)

d_bat_raw

Filter

	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

Player Attributes

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

Past Performance Sliding Windows + Other Stats

One of the Targets

last1_HR	last2_HR	last3_HR	last4_HR	last5_HR	avg_last2_HR	avg_last3_HR	avg_last4_HR	avg_last5_HR	yearID	HR	
NA	NA	NA	NA	NA	NaN	NaN	NaN	NaN	2011	5	Training
5	NA	NA	NA	NA	5.0	5.00000	5.00000	5.00000	2012	30	
30	5	NA	NA	NA	17.5	17.50000	17.50000	17.50000	2013	27	
27	30	5	NA	NA	28.5	20.66667	20.66667	20.66667	2014	36	
36	27	30	5	NA	31.5	31.00000	24.50000	24.50000	2015	41	
41	36	27	30	5	38.5	34.66667	33.50000	27.80000	2016	29	Validation
29	41	36	27	30	35.0	35.33333	33.25000	32.60000	2017	33	
33	29	41	36	27	31.0	34.33333	34.75000	33.20000	2018	NA	Forecast
33	33	29	41	36	33.0	31.66667	34.00000	34.40000	2019	NA	
33	33	33	29	41	33.0	33.00000	32.00000	33.80000	2020	NA	

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

H2O.ai

H_2O 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]))
```

H_2O AutoML Results

```
H2ORegressionMetrics: 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]
```