

# The Making of a Real-World Moneyball

## Finding Undervalued Players with H<sub>2</sub>O, LIME and Shiny



Jo-fai (Joe) Chow

Data Science Evangelist at H2O.ai

joe@h2o.ai / @matlabulous

# About Me

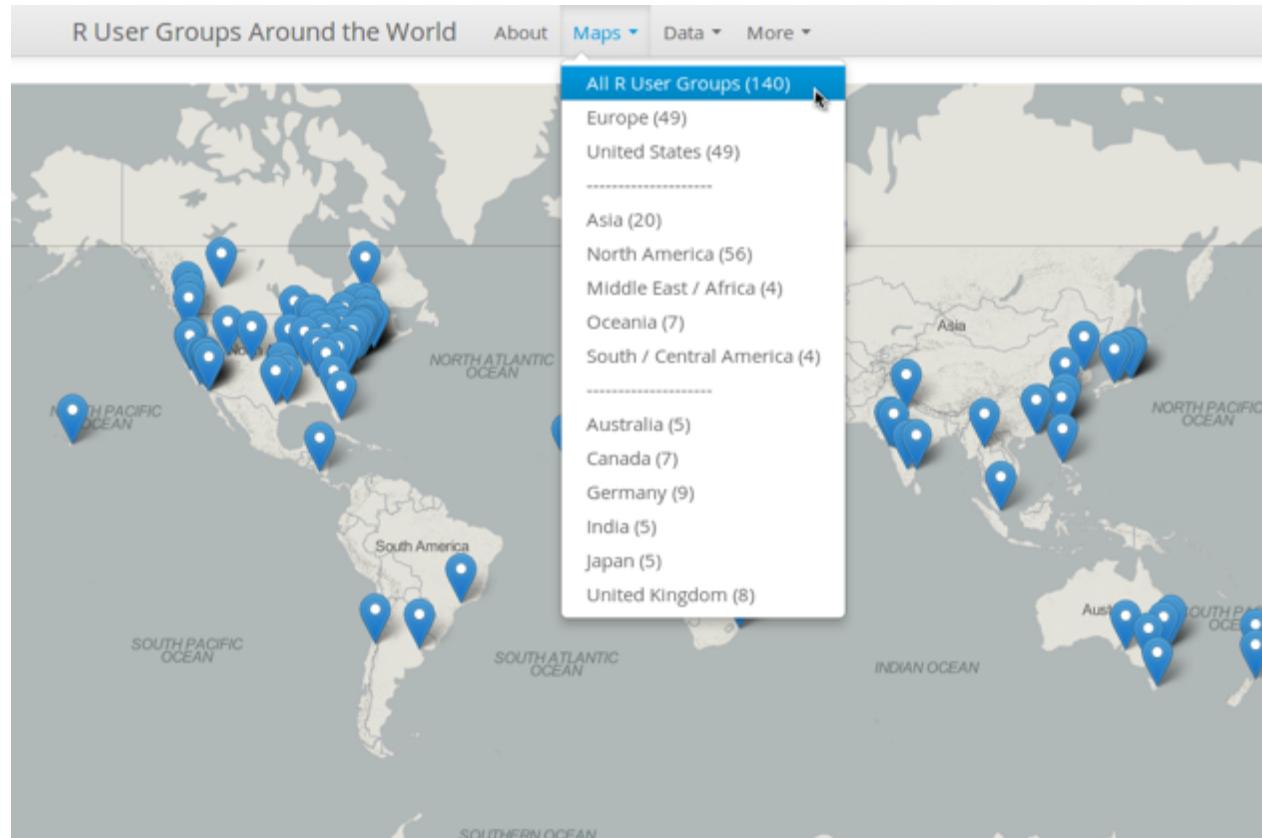


- **Before H<sub>2</sub>O**

- 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)

•

# About Me (before H<sub>2</sub>O)

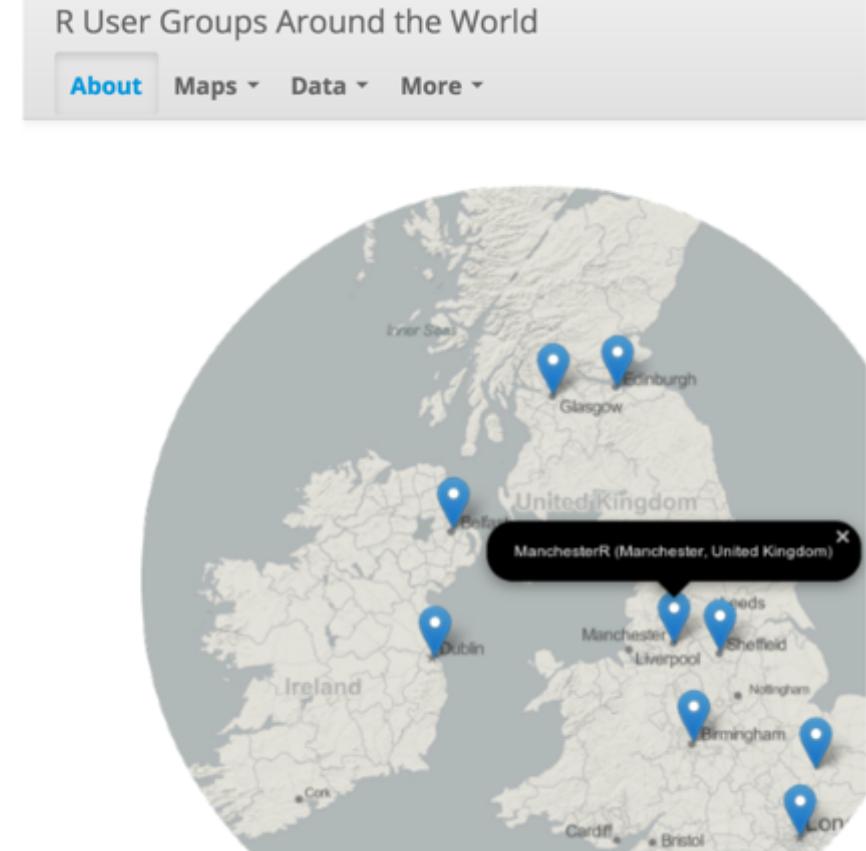


August 21, 2014

## Revolution Analytics' User Group Map Contest has a Winner

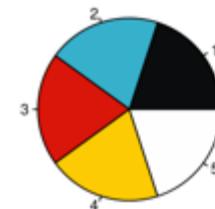
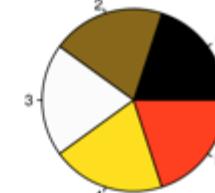
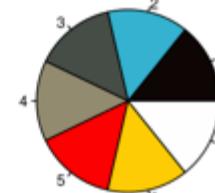
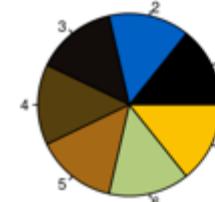
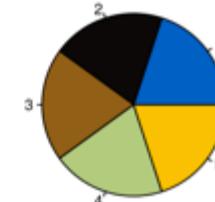
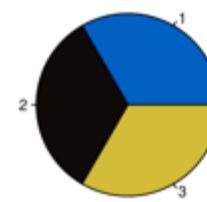
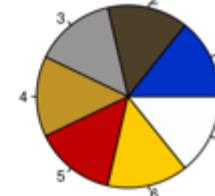
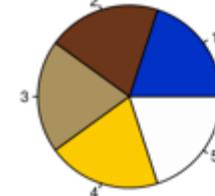
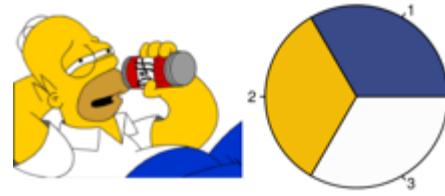
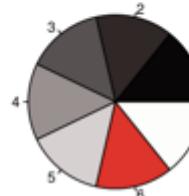
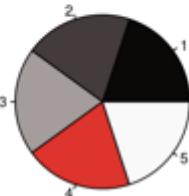
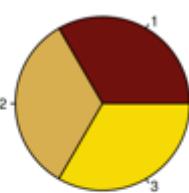
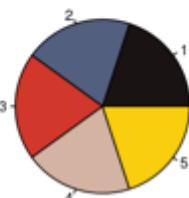
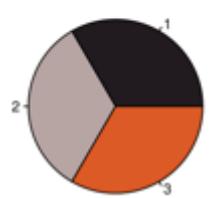
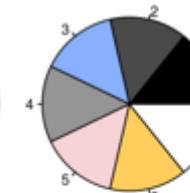
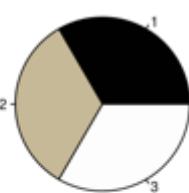
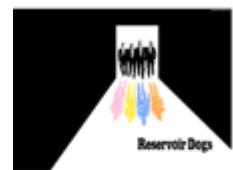
by Joseph Rickert

We are pleased to announce that [Jo-fai Chow](#) is the winner of the Revolution Analytics contest. Jo-fai's entry, which was implemented as a [Shiny project](#), may be viewed by clicking on the figure below.



<http://blog.revolutionanalytics.com/2014/08/winner-for-revolution-analytics-user-group-map-contest.html>  
<https://github.com/woobe/rugsmaps>

# About Me (before H<sub>2</sub>O)



<http://blenditbayes.blogspot.com/2014/05/towards-yet-another-r-colour-palette.html>  
<https://github.com/woobe/rPlotter>

# About Me



Paris

- **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 ...



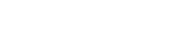
Jo-fai (Joe) Chow  
@matlabulous

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



Budapest

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



Jo-fai (Joe) Chow  
@matlabulous

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

#AroundTheWorldWithH2Oai #360Selfie 🇳🇱  
cc @fishnets88



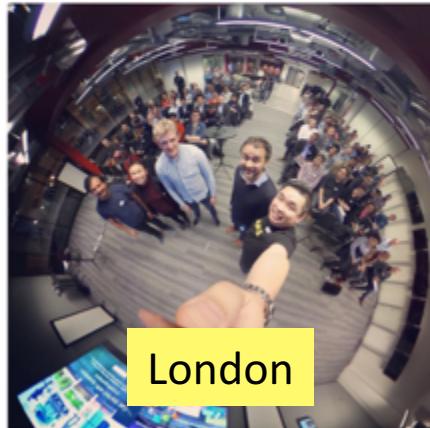
Amsterdam

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



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



London

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



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 🇫🇷



Toulouse

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



Jo-fai (Joe) Chow  
@matlabulous

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



Berlin

1:54 PM - 7 Mar 2018 from Hotel Berlin



Jo-fai (Joe) Chow  
@matlabulous

My first #Moneyball talk at #Cologne #rstats #meetup last week was well received. Thanks Jessica Peterka-Bonetta and @eyeo for having me.

Slides: [slideshare.net/JofaiChow/maki ...](https://slideshare.net/JofaiChow/making-mymillion-dollar-e3-decisions-with-h2o-automl-lime-and-shiny)

#AroundTheWorldWithH2Oai #360Selfie  
cc @h2oai @IBMDatascience @Aginity  
@DaithiOCiaran @arikaplan1



Cologne

4:56 PM - 18 Jun 2018 from Cologne, Germany

Reminder: #360Selfie

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

# 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



# 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*



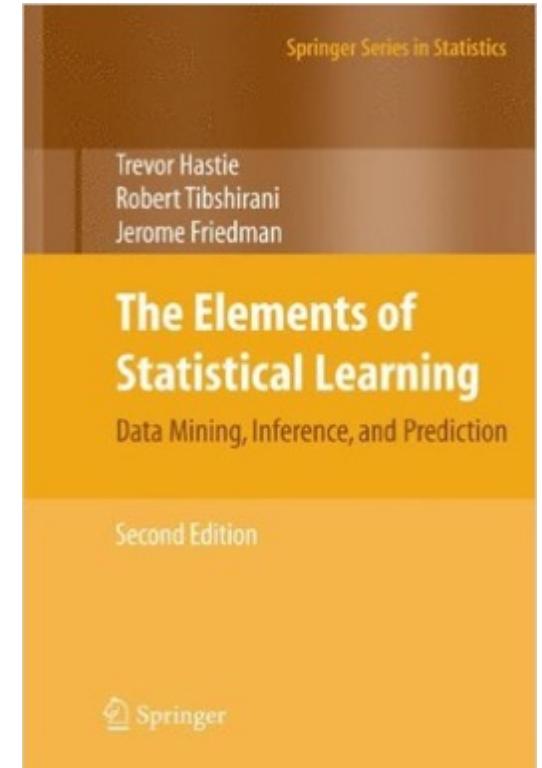
## Dr. Robert Tibshirani

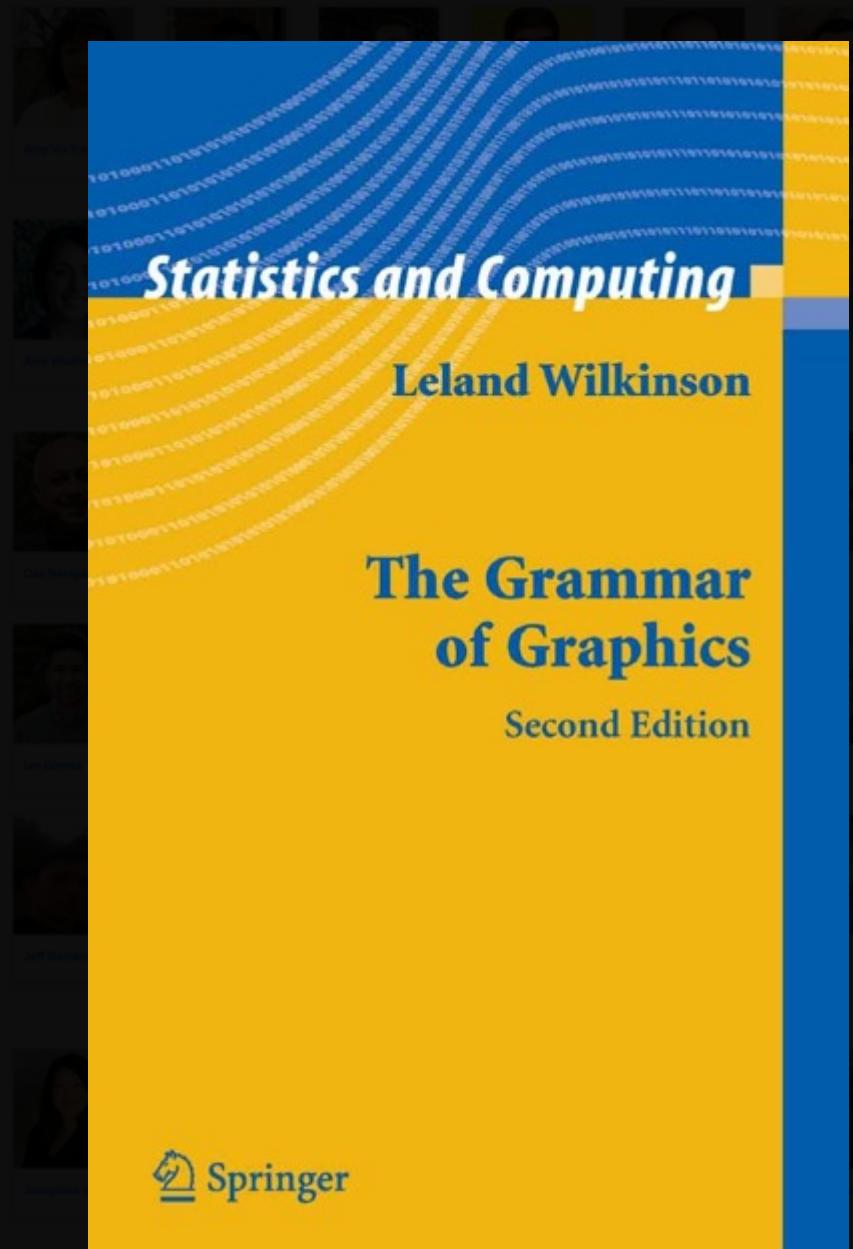
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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



## Dr. Steven Boyd

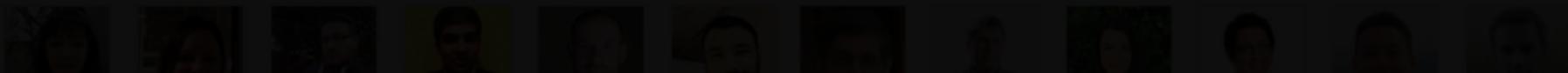
- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities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





## Origin of R Package `ggplot2`





Amy Vi Tran

Angela Benz

Alfred Bai

Alyson  
Cheadle

Ann Cardel

Aaron Berliner

Laurie Wilkinson

Magda Stevens

Mark Morrison

Monica Morrison

Mike Chan

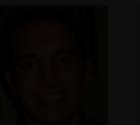
Mike Lavelle



Kent Wachter



Benjamin Campbell



Warren Murray



Carl Andrews



Charles Marchi



Cheri Poff

Michael  
Olynyk

Matt Dowle



Maggie Kuhn



Michael Parker



Michael Morrison



Dan Hargrave



David Crawford



Dorothy Lamko



Eric Danner



Jay Birrell



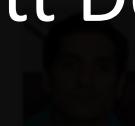
Jeff Gammie



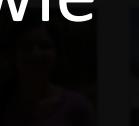
Jo Fel Chow



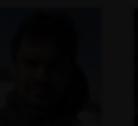
John Schaefer



Jon McNaull



Kristin Morrison



Kristin Morrison



Kristin Morrison



Ian Gammie



Jacqueline Scott



Jay Birrell



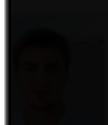
Jeff Gammie



Jo Fel Chow



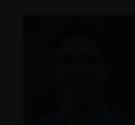
John Schaefer



Jon McNaull



Kristin Morrison



Kristin Morrison



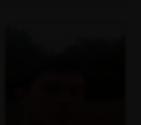
Kristin Morrison



Kristin Morrison



Kristin Morrison



Josephine Wang



Justin Loyola

Karen  
Heyneke

Kathy Lee



Kristina Andrus



Lauren Difesa



Michael Parker



Tony Wong



Tim Keeling



Venkat Venkatesan



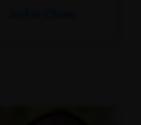
Venkat Venkatesan



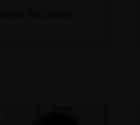
Michael Parker



Amy Vi Tran



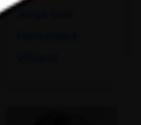
Angela Benz



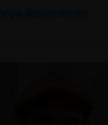
Alfred Bai

Alyson  
Cheadle

Ann Cardel



Aaron Berliner



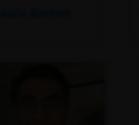
Laurie Wilkinson



Magda Stevens



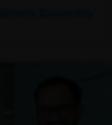
Mark Morrison



Monica Morrison

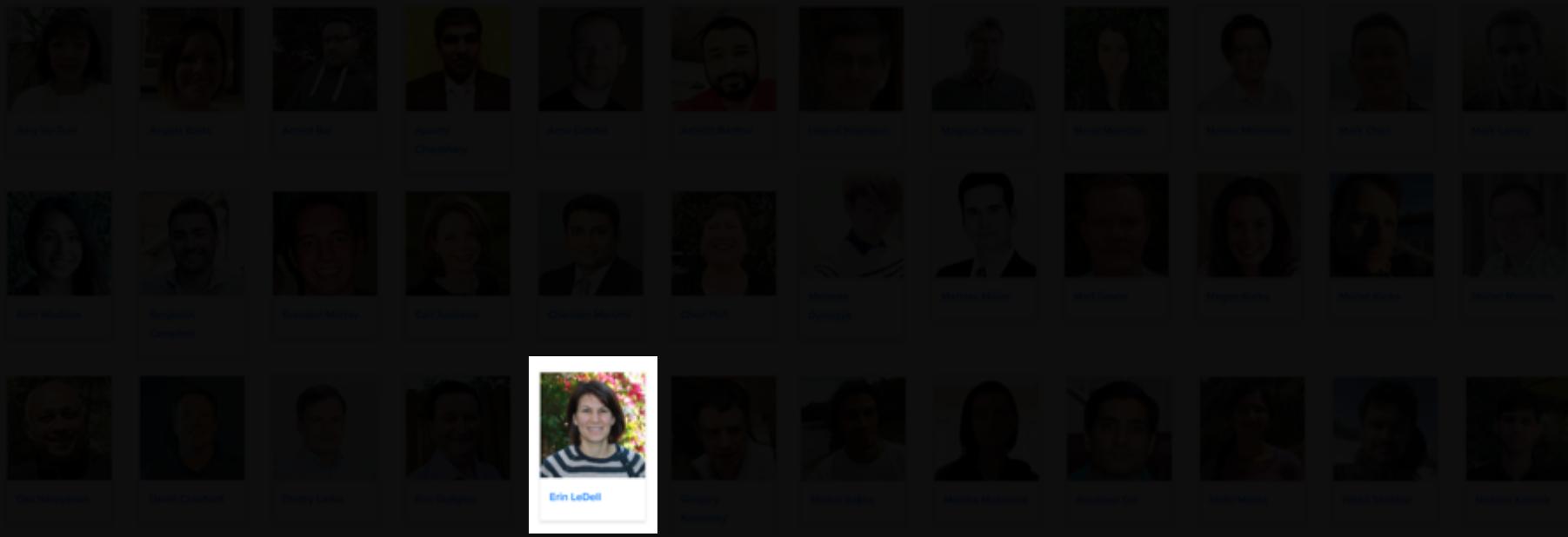


Mike Chan



Mike Lavelle

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



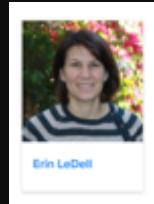
Erin LeDell, Chief ML Scientist  
Women in ML/DS & R-Ladies Global



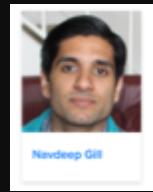
H<sub>2</sub>O Team

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

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



Erin LeDell

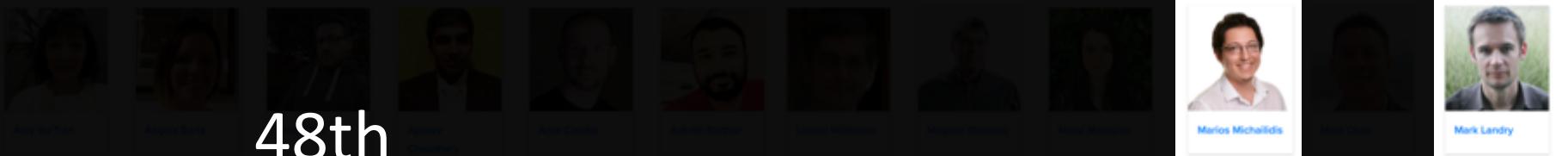


Navdeep Gill

Erin LeDell

Navdeep Gill

H<sub>2</sub>O Team



48th

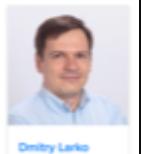


1st

33rd



25th



4th

## Kaggle Grand Masters (and their Highest Rank)



113  
Grandmasters



980  
Masters



3,339  
Experts



46,135  
Contributors



33,242  
Novices

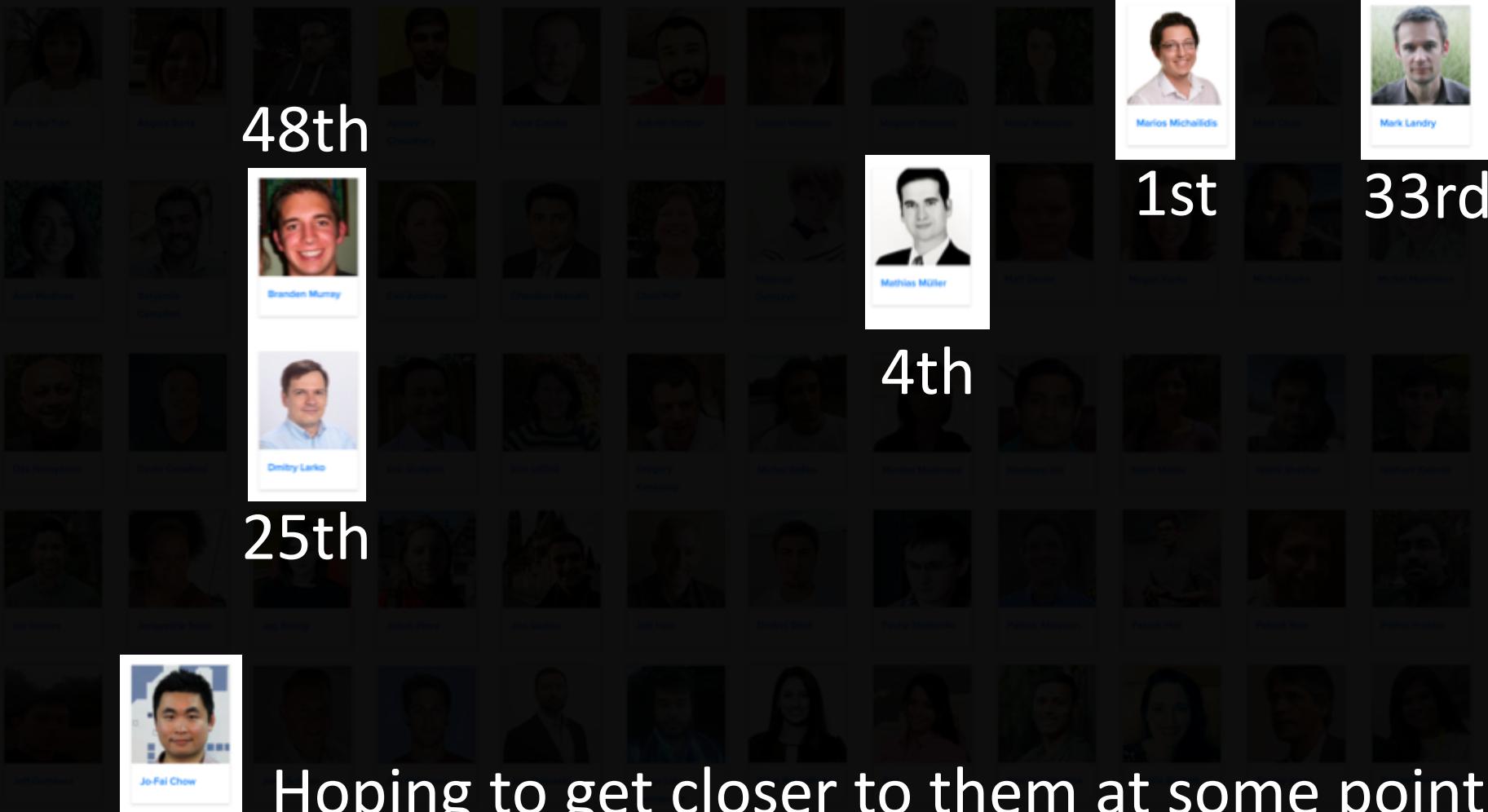
About 80,000 Kagglers

H<sub>2</sub>O Team



13th

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



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

H<sub>2</sub>O Team

13th

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

# 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

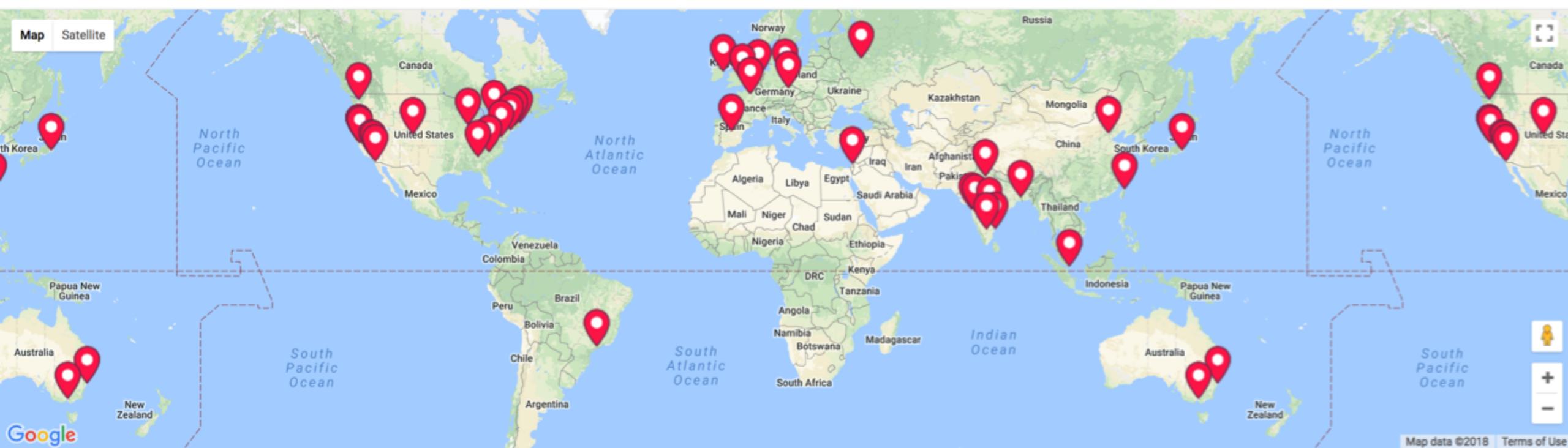
# H2O.ai is a **Leader** in the 2018 Gartner Data Science and Machine Learning Platforms Magic Quadrant

- 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](#)



## H2O Artificial Intelligence and Machine Learning

Members  
78,356

Groups  
39

Countries  
18

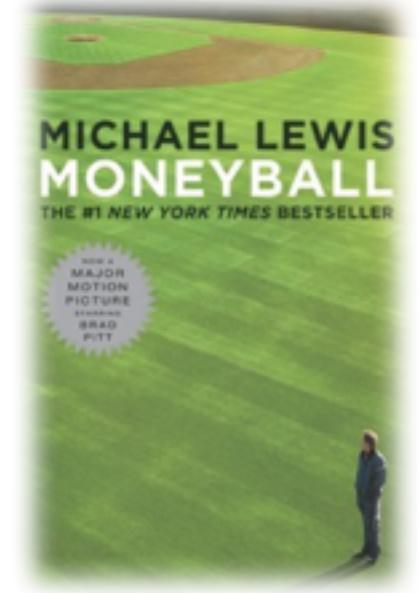
<https://www.meetup.com/pro/h2oai/>

# 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 that is a sub-navigation for Fantasy Baseball with links to Home, Top 300 Rankings, Forecaster: Hitting Matchups, and More. The main content area is titled "Sortable 2018 Projections" under "ESPN = Free Fantasy Baseball". It has filters for Position (Batters, Pitchers, etc.), 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 several statistics. To the right of this table is a larger table titled "2018 SEASON BATTING PROJECTIONS" with columns for R, HR, RBI, SB, and AVG. This second table is also highlighted with an orange border. A blue banner at the bottom of the page reads "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

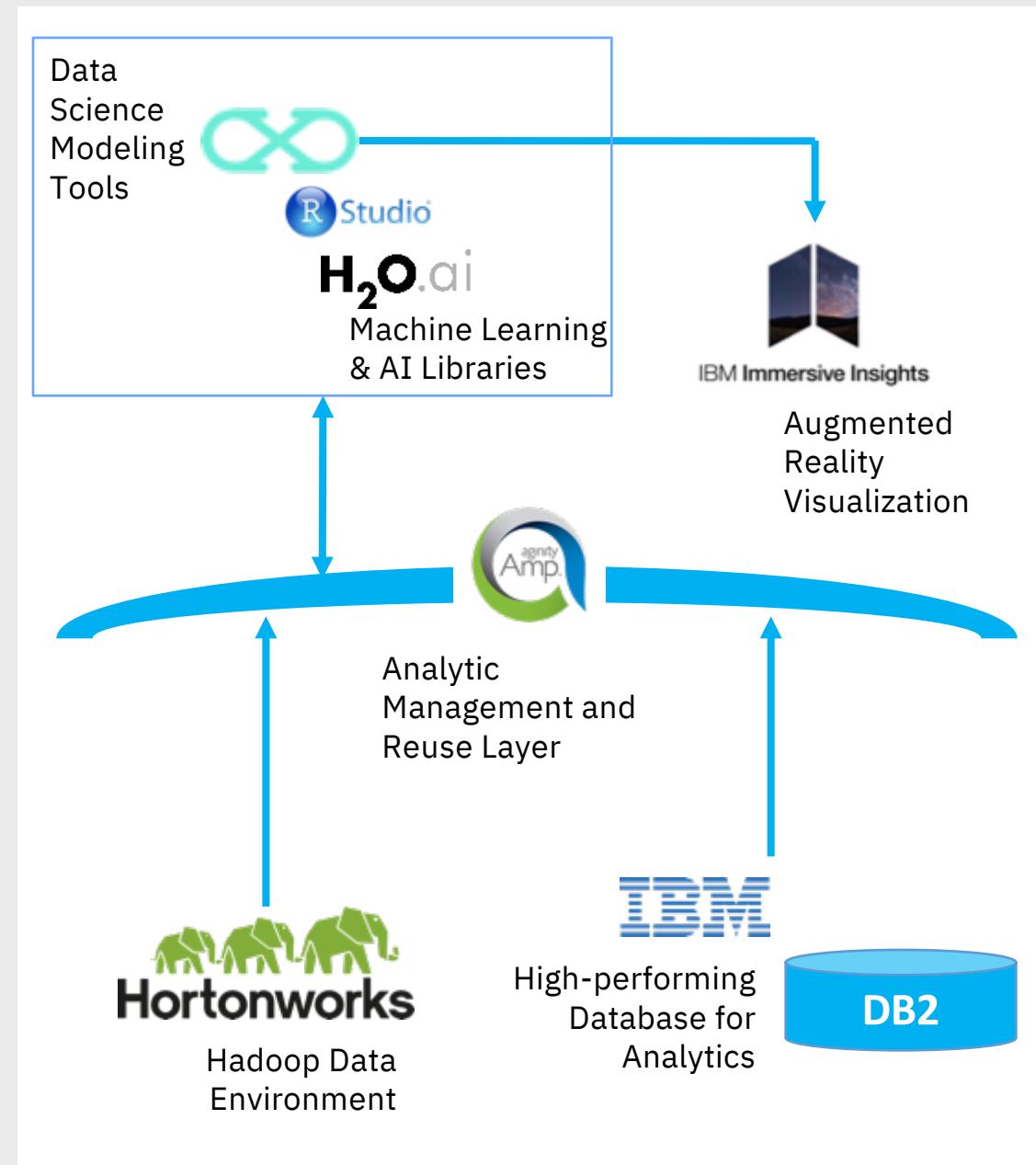
A portrait of Ari Kaplan, a bald man with glasses, wearing a dark suit and a yellow tie. He is smiling. The Aginity logo is in the bottom left corner of the photo frame.

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

# Enterprise Solution

## The Workflow

1. Data loaded into the databases
2. Connected diverse data sources to Amp
3. Amp used to create derived attributes and publish them and data to DSX and H<sub>2</sub>O
4. DSX and H<sub>2</sub>O to build and tweak statistical and machine learning models
5. Visualizations tested in Immersive Insights
6. Steps 4 and 5 repeated to get settled data
7. Statistical and machine learning models saved in Amp
8. Data exported to Immersive Insights for final visualizations



# 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



# Framing the Business Problem for Machine Learning

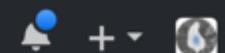
Code on GitHub (without Ari's proprietary data)

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



Search or jump to...

Pull requests Issues Marketplace Explore



woobe / moneyball

Unwatch ▾ 2

Star 1

Fork 0

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Settings

Moneyball Demo (Public Version)

Add topics

7 commits

Apache-2.0

More Info → [github.com/woobe/  
moneyball](https://github.com/woobe/moneyball)

Edit

Branch: master ▾

New pull request

Create new file

Upload files

Find file

Clone or download ▾

woobe Added descriptions

Latest commit d630812 2 days ago

cache\_data

Raw data from Lahman database

3 days ago

.gitignore

Initial commit

22 days ago

LICENSE

Initial commit

22 days ago

README.md

Added descriptions

2 days ago

step\_1\_data\_munging.R

Data munging for Lahman data only

3 days ago

step\_2\_model\_pitching.R

H2O AutoML Model Building Scripts

2 days ago

step\_3\_model\_batting.R

H2O AutoML Model Building Scripts

2 days ago

README.md

# 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



# 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')
```

# 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

# Creating Consistent and Reusable Analytic Assets Managed by Aginity Amp

The screenshot shows the Aginity Amp application interface. On the left, there's a sidebar with navigation links: Home, New, Browse, Jobs, and Settings. The main workspace is titled "Money Ball" and contains a search bar and a tree view of analytic assets. One asset is selected: "AriDB2012\_batter\_derived\_attributes\_sa". The right side of the screen displays the configuration details for this asset.

**AriDB2012\_batter\_derived\_attributes\_sa**

Select Analytic: [/AriDB2012\\_batter\\_derived\\_attributes/aridb2012\\_batter\\_derived\\_output](#)

**\$ Frame Parameters**

NAME	DATATYPE	PROMPT FOR INPUT?	VALUES
------	----------	-------------------	--------

**SCHEMA**   **RELATIONSHIPS**   **PREVIEW**   **Test**

COLUMN NAME	LABEL	DATA TYPE	DESCRIPTION	PRIMARY KE
batter	batter	INTEGER		<input type="checkbox"/>
BA_fb_over_93	BA_fb_over_93	DOUBLE		<input type="checkbox"/>
AB_fb_over_93	AB_fb_over_93	LONG		<input type="checkbox"/>
H_fb_over_93	H_fb_over_93	LONG		<input type="checkbox"/>
TB_fb_over_93	TB_fb_over_93	LONG		<input type="checkbox"/>
HR_fb_over_93	HR_fb_over_93	LONG		<input type="checkbox"/>
BA_fastball_under_93	BA_fastball_under_93	DOUBLE		<input type="checkbox"/>
AB_fastball_under_93	AB_fastball_under_93			<input type="checkbox"/>
H_fb_under_93	H_fb_under_93			<input type="checkbox"/>
TB_fb_under_93	TB_fb_under_93			<input type="checkbox"/>
HR_fb_under_93	HR_fb_under_93			<input type="checkbox"/>
BA_slider	BA_slider			<input type="checkbox"/>
AB_slider	AB_slider			<input type="checkbox"/>
H_slider	H_slider			<input type="checkbox"/>
TB_slider	TB_slider			<input type="checkbox"/>
HR_slider	HR_slider			<input type="checkbox"/>
BA_curve	BA_curve			<input type="checkbox"/>

**OUTPUT SCHEMA**

- batter
- BA\_fb\_over\_93
- AB\_fb\_over\_93
- H\_fb\_over\_93
- TB\_fb\_over\_93
- HR\_fb\_over\_93
- BA\_fastball\_under\_93
- AB\_fastball\_under\_93
- H\_fb\_under\_93
- TB\_fb\_under\_93
- HR\_fb\_under\_93
- BA\_slider
- AB\_slider
- H\_slider
- TB\_slider
- HR\_slider
- BA\_curve

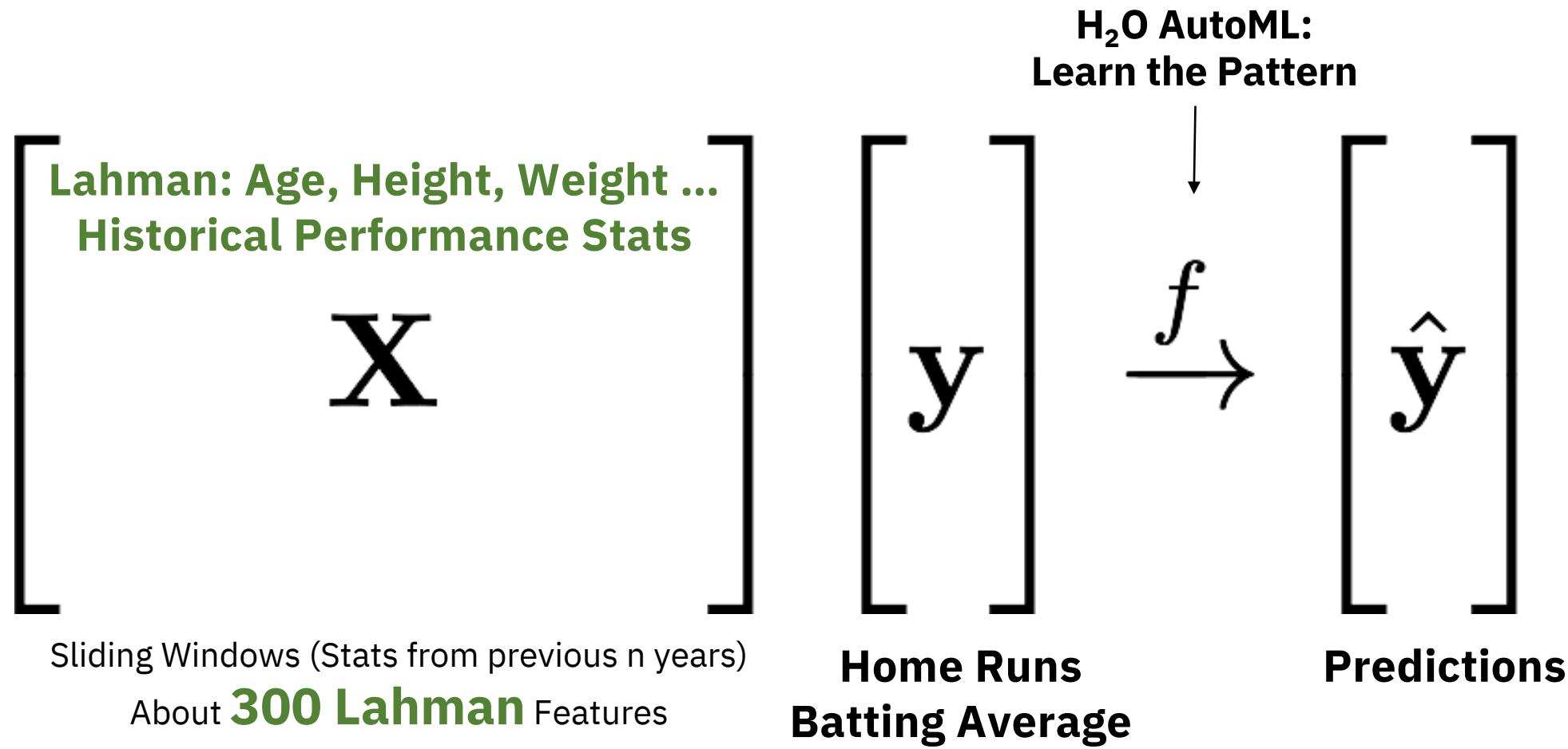
More **Save**

# Creating Consistent and Reusable Analytic Assets Managed by Aginity Amp

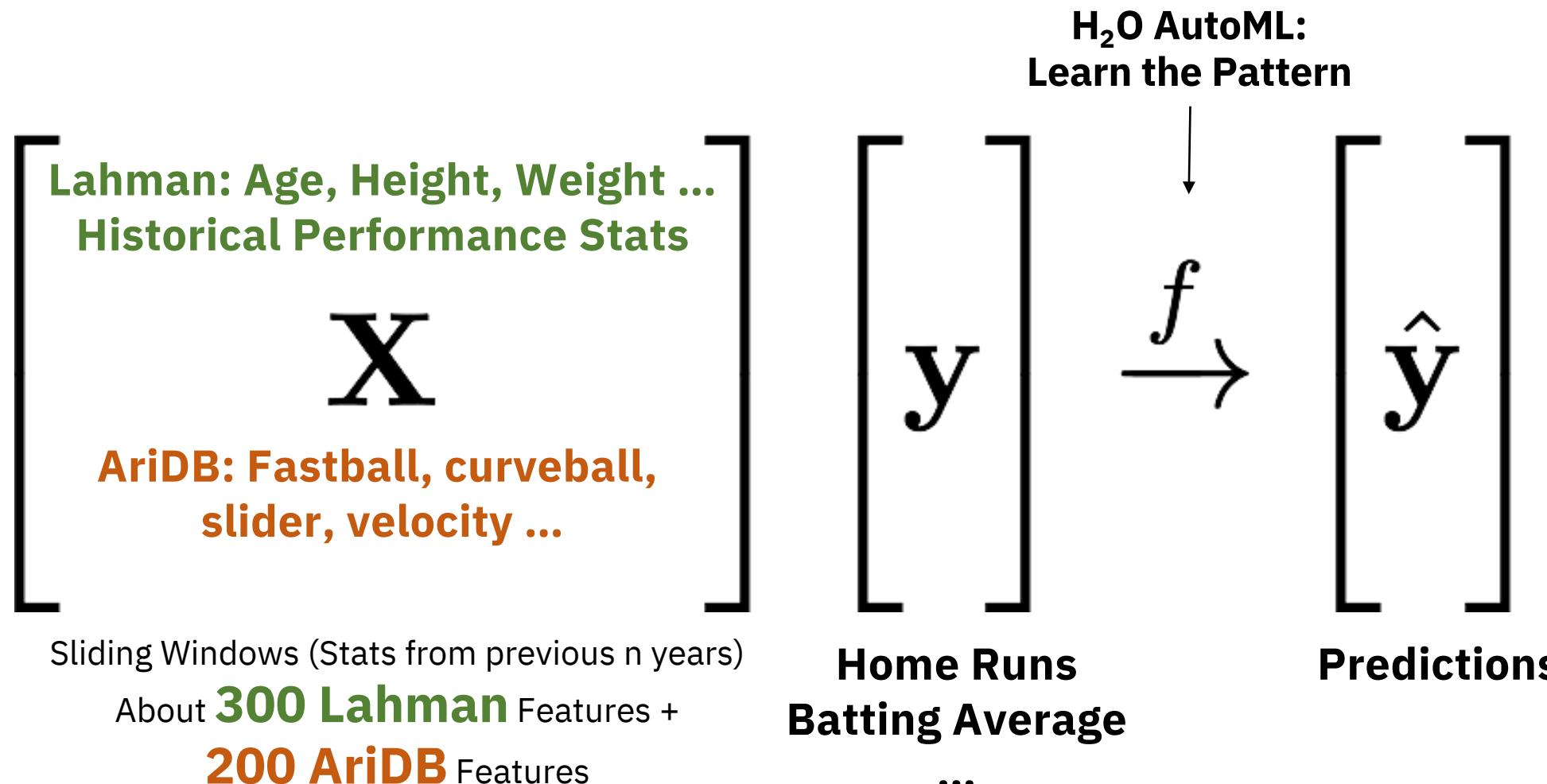
The screenshot shows the Aginity Amp application interface. On the left, the navigation sidebar includes icons for Home, New, Browse, Jobs, and Settings, along with a search bar and a workspace dropdown set to "Main Workspace". The main workspace, titled "Money Ball", contains a list of analytic assets under the "Analytics" category, such as "AriDB2012\_batter\_derived\_attributes", "AriDB2012\_pitcher\_derived\_attributes", and "AriDB2013\_batter\_derived\_attributes". The central panel displays the details for the selected asset, "AriDB2012\_batter\_derived\_attributes". It shows the "INPUT SCHEMA" section with a "Parameters" dropdown set to "AriDB... ./AriDB20...". The code editor on the right contains the SQL query:

```
%sql
create or replace temporary view AriDB2012_batter_derived_output as
  select batter,
  round(sum(if(Event in ('Single','Double','Triple','Home Run') and
  end_ab_flag='1' and start_speed>=93.0 and pitch_type in ('FF',
  'FT','FA'),1,0)) / sum(if(Event in ('Single','Double','Triple'
  , 'Home Run','Batter Interference','Bunt groundout','Bunt pop out',
  'Double Play','Fielders Choice','Fielders Choice Out','Flyout',
  'Forceout','Grounded Into DP','Groundout','Lineout','Pop Out',
  'Strikeout','Strikeout - DP','Triple Play') and end_ab_flag='1'
  and start_speed>=93.0 and pitch_type in ('FF','FT','FA'),1,0)),3)
  as BA_fb_over_93,
  sum(if(Event in ('Single','Double','Triple','Home Run','Batter
  Interference','Bunt groundout','Bunt pop out','Double Play',
  'Fielders Choice','Fielders Choice Out','Flyout','Forceout',
  'Grounded Into DP','Groundout','Lineout','Pop Out','Strikeout',
  'Strikeout - DP','Triple Play') and end_ab_flag='1' and
  start_speed>=93.0 and pitch_type in ('FF','FT','FA'),1,0)) as
  AB_fb_over_93,
  sum(if( spray_type='H' and end_ab_flag='1' and start_speed>=93.0 and
  pitch_type in ('FF','FT','FA'),1,0)) as H_fb_over_93,
  sum(if( spray_type='H' and end_ab_flag='1' and start_speed>=93.0 and
  pitch_type in ('FF','FT','FA'),if(event='Home Run',4,0)+if(event
  ='Triple',3,0)+if(event='Double',2,0)+if(event='Single',1,0),0))
  as TB_fb_over_93,
  sum(if(event='Home Run' and end_ab_flag='1' and start_speed>=93.0 and
  pitch_type in ('FF','FT','FA'),1,0)) as HR_fb_over_93,
  round(sum(if(Event in ('Single','Double','Triple','Home Run') and
  end_ab_flag='1' and start_speed<93.0 and pitch_type in ('FF','FT',
  'FA'),1,0)) / sum(if(Event in ('Single','Double','Triple','Home
  Run','Batter Interference','Bunt groundout','Bunt pop out',
  'Double Play','Fielders Choice','Fielders Choice Out','Flyout',
  'Forceout','Grounded Into DP','Groundout','Lineout','Pop Out',
  'Strikeout','Strikeout - DP','Triple Play') and end_ab_flag='1'
  and start_speed<93.0 and pitch_type in ('FF','FT','FA'),1,0)),3)
```

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



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



# 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).

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

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

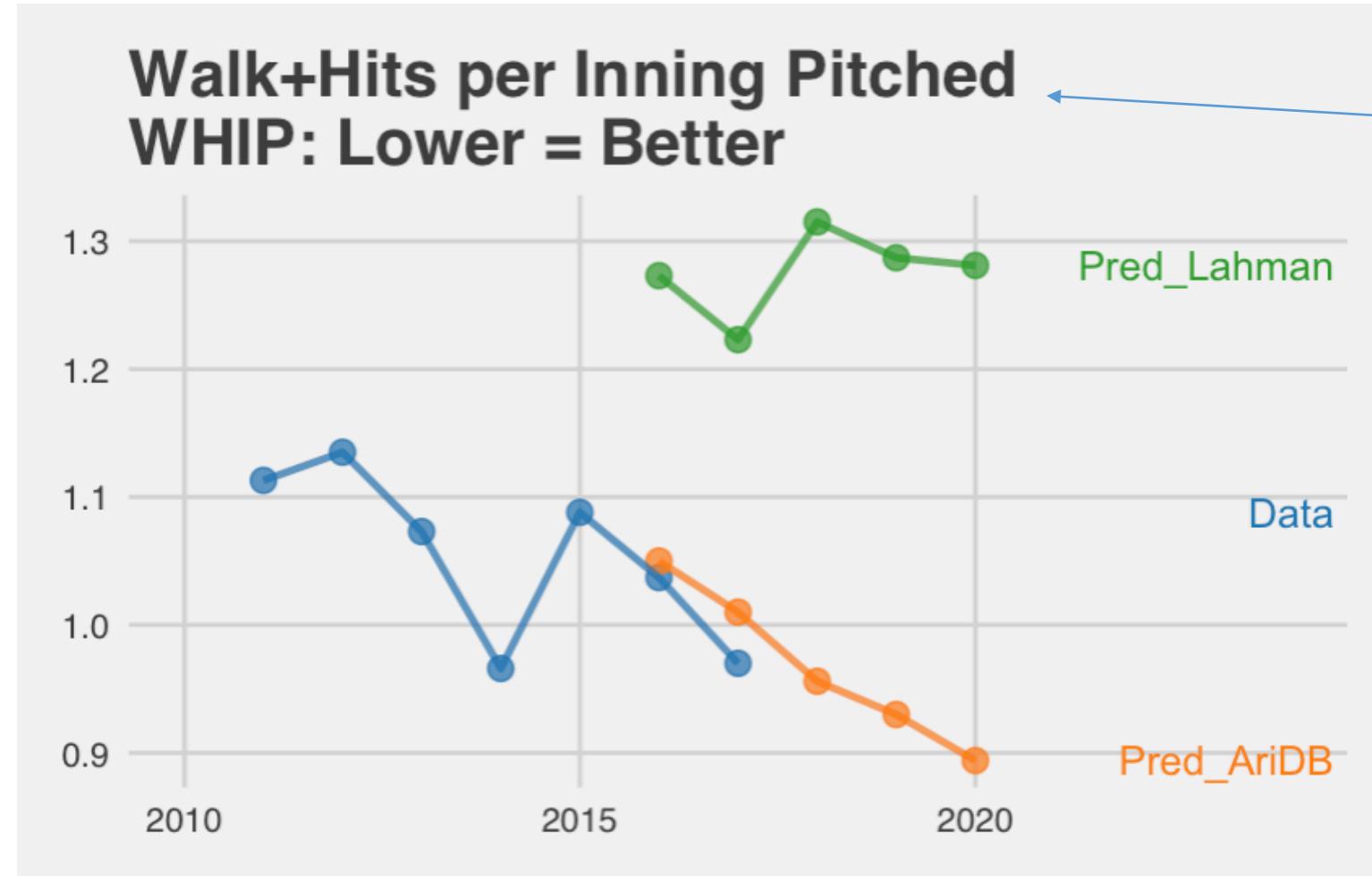
MSE: 0.00246453
RMSE: 0.04964404
MAE: 0.03335875
RMSLE: 0.04124294
Mean Residual Deviance : 0.00246453
```

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]

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

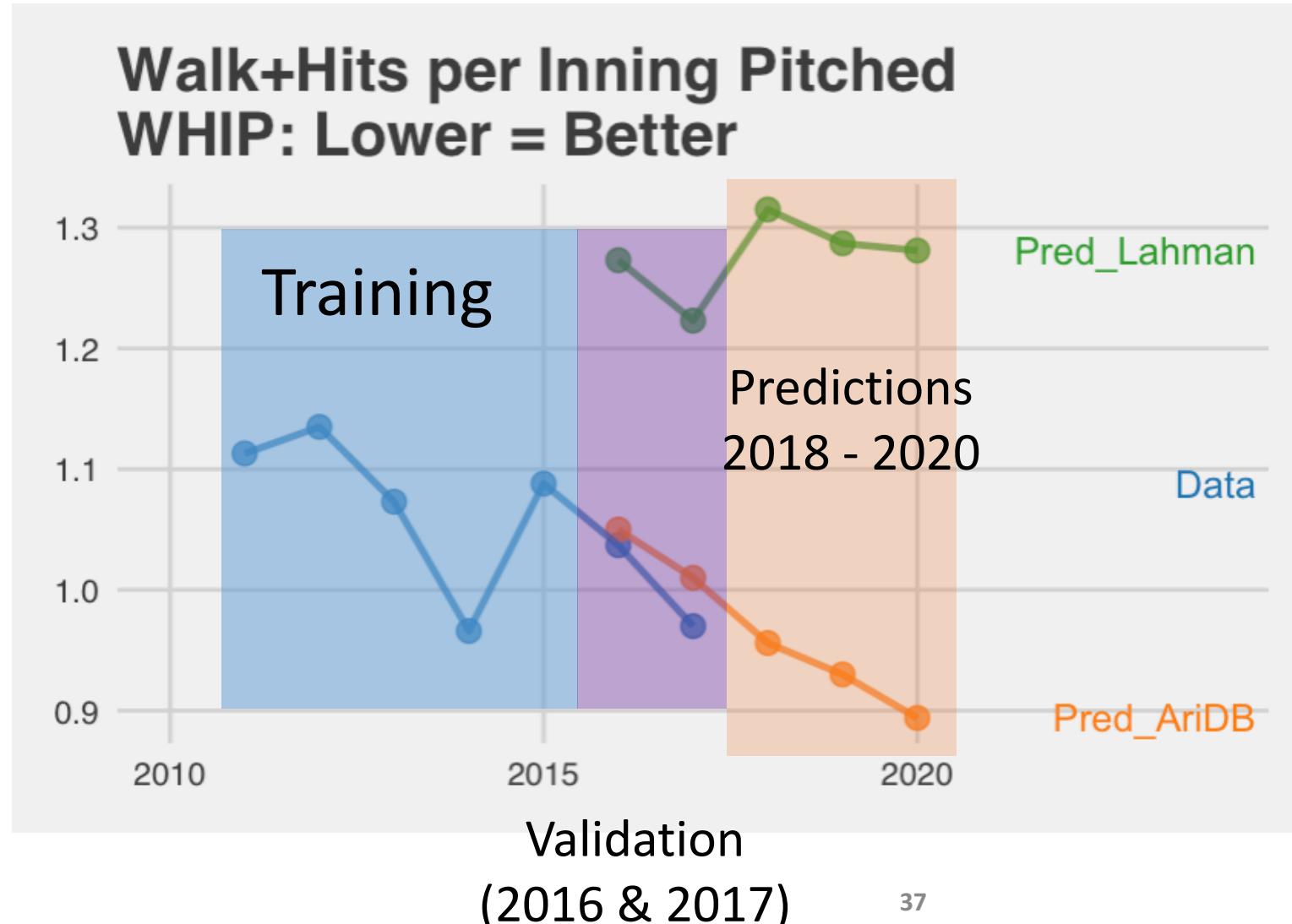


One of Many Targets  
(e.g. Home Runs, Batting Average)



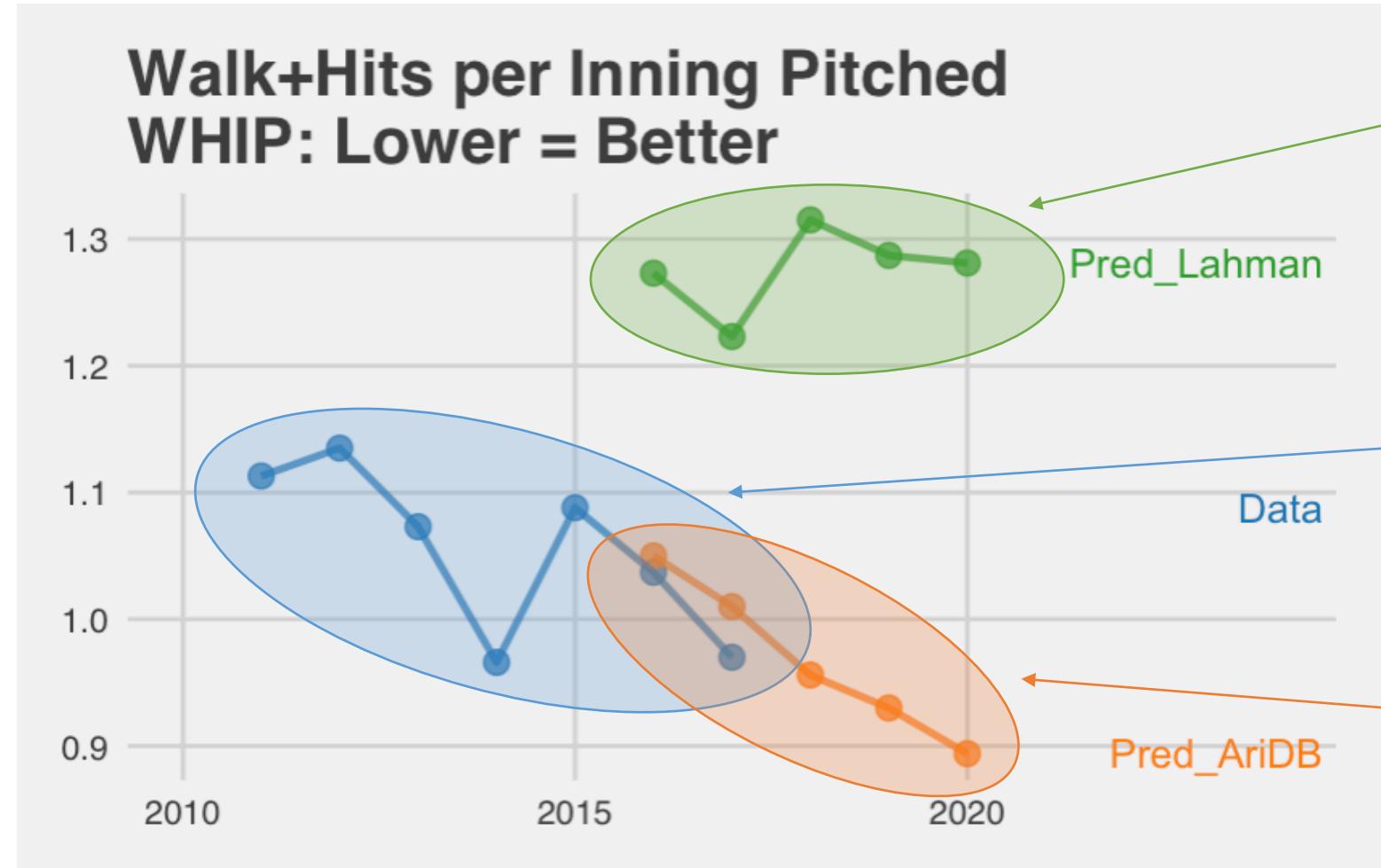
```
# Install 'h2o' from CRAN  
install.packages('h2o')
```

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



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

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



Results from models based on Lahman data only

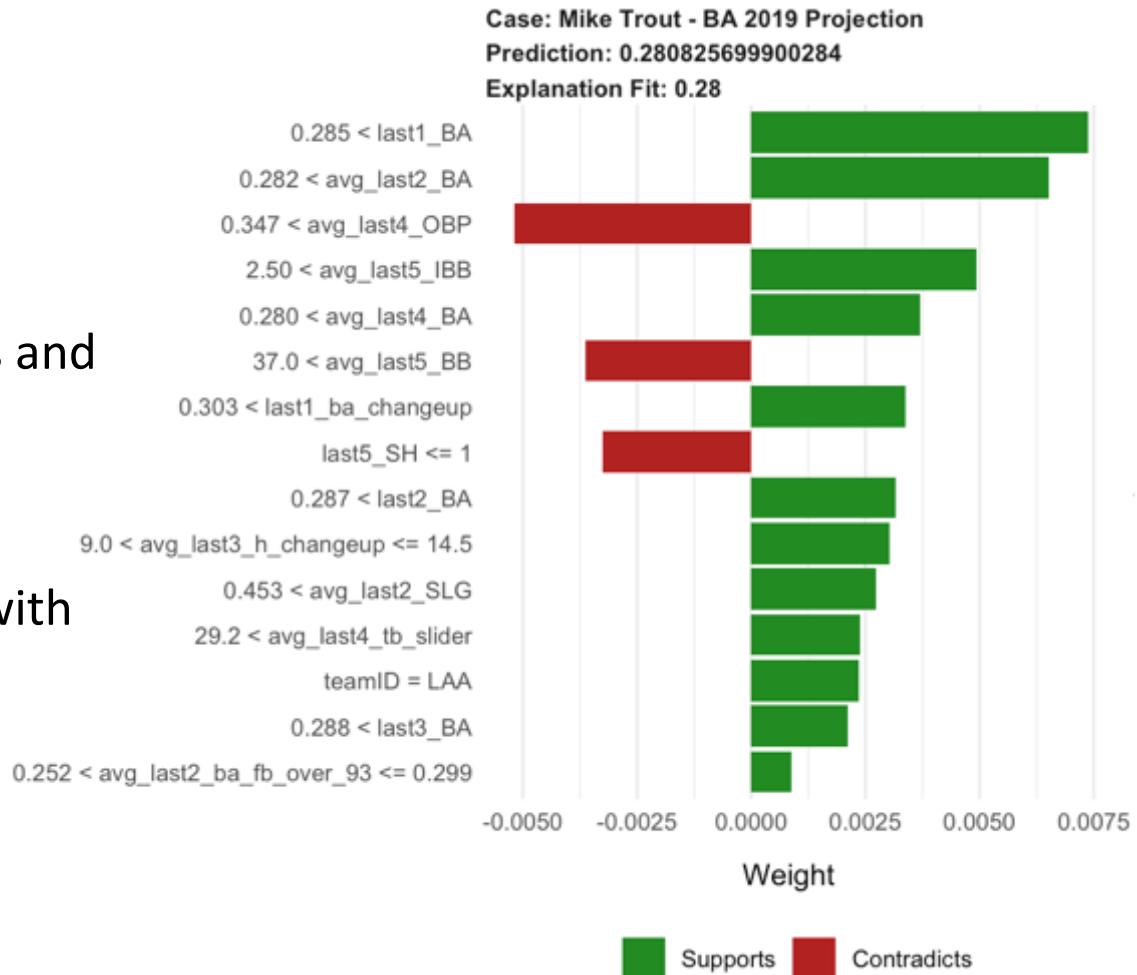
Historical player performance data

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

# Explaining the Predictions

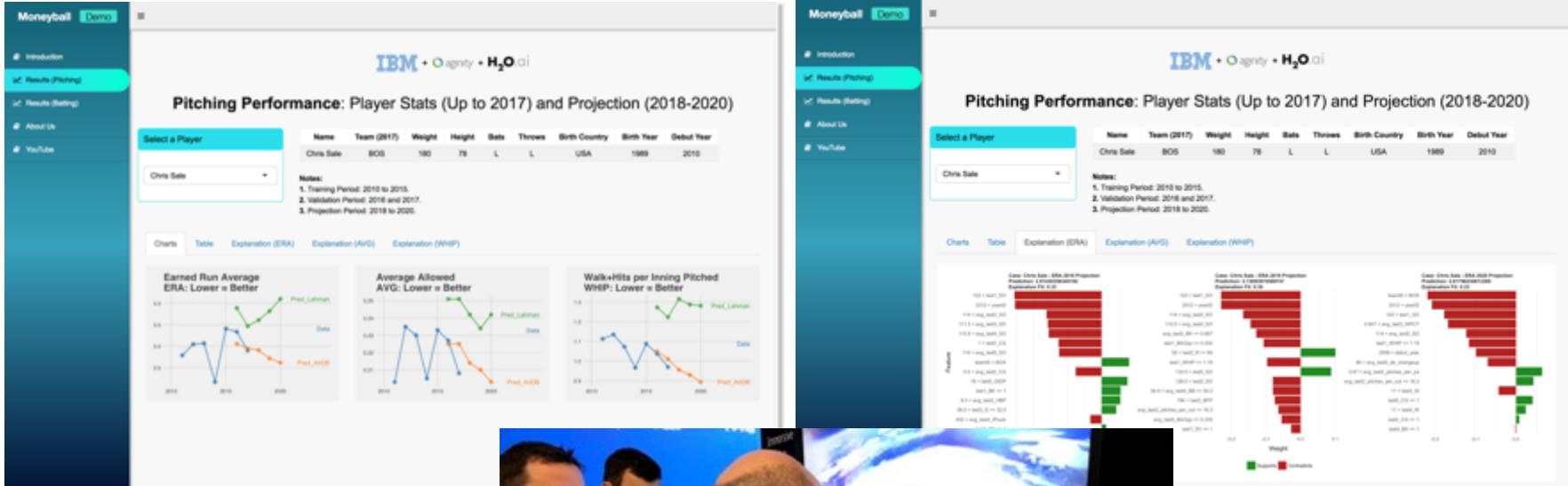
## LIME – Local Interpretable Model-agnostic Explanations

- 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



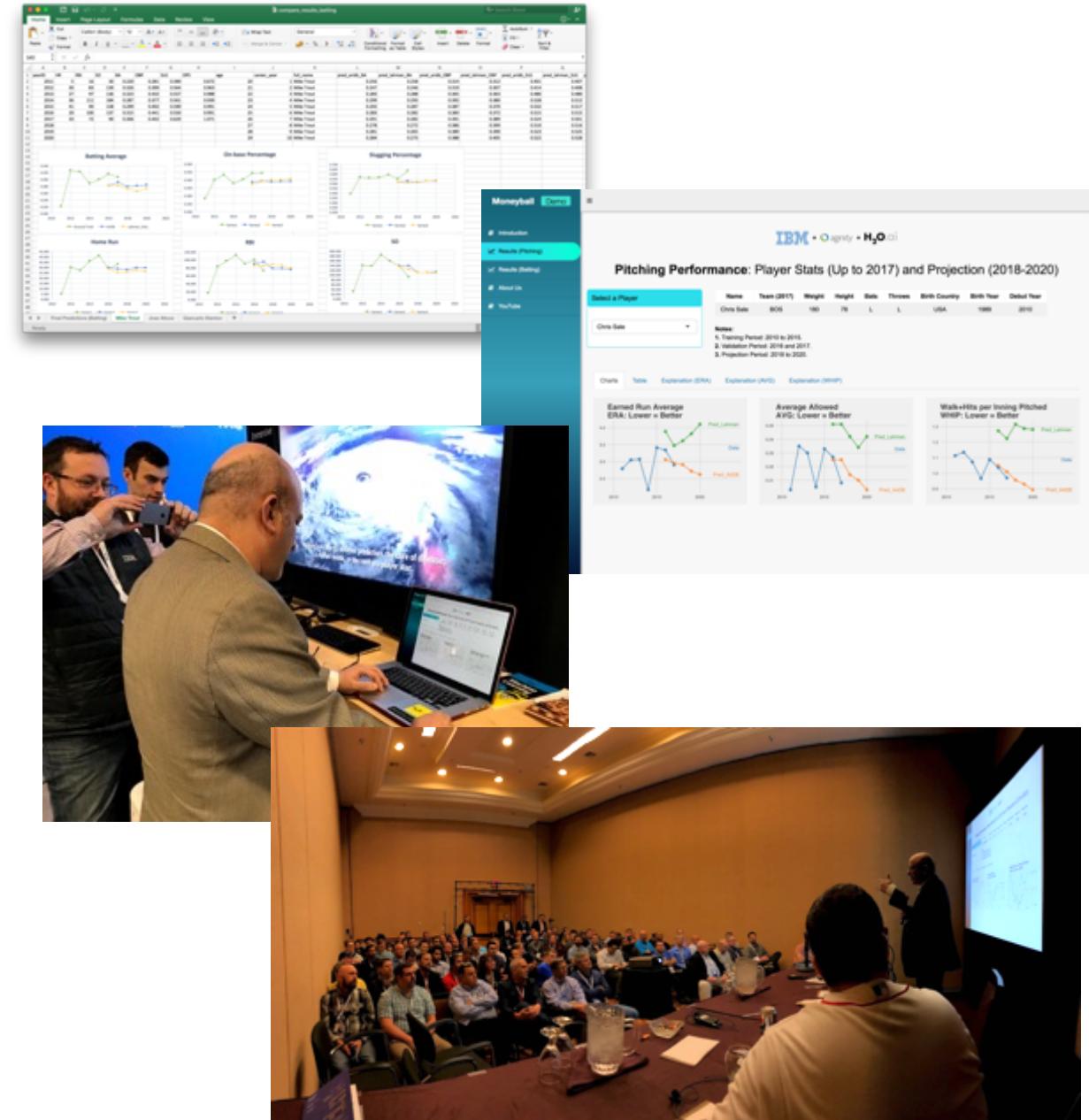
Live Demo  
on my laptop



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

# From Toy Demo to Real Moneyball

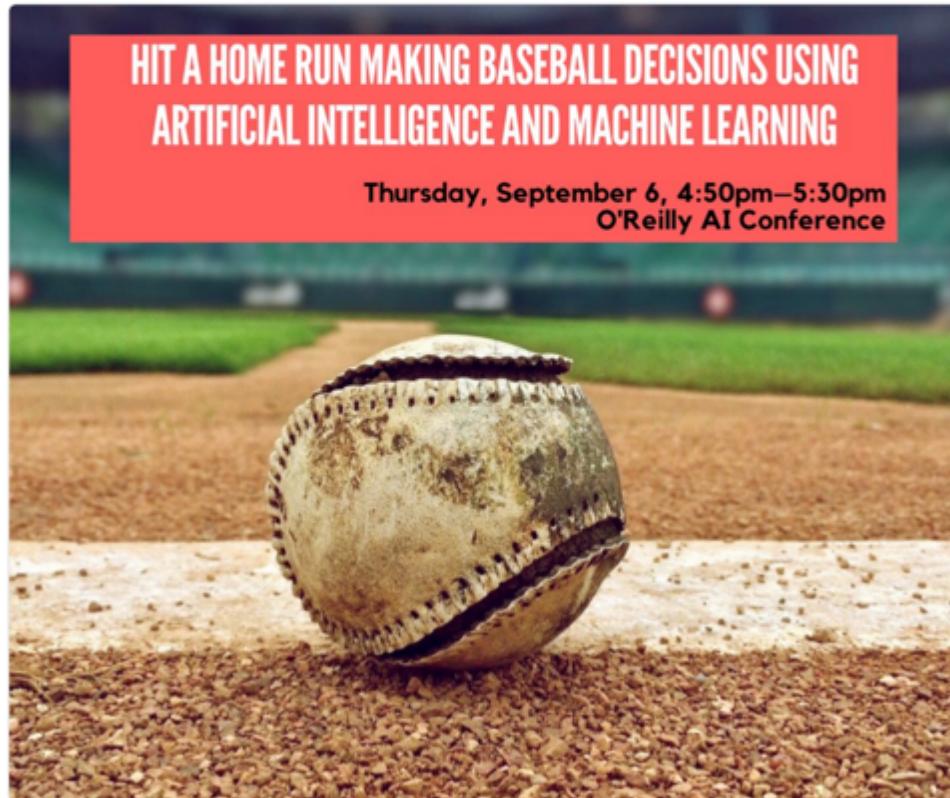
- **March 19** – AutoML Predictions finalized.  
Initial presentation in Excel.
- **March 20** – Version 1 of Shiny app. Ari used the app to validate some players he had in mind and recommended one player to his team.
- **March 21** – Multimillion-dollar contract finalized.
- **March 22** – Moneyball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





H2O.ai  
@h2oai

Attending O'Reilly #AI Conference? Learn how to hit a home run using #artificialintelligence and #machinelearning from @Ledell and former MLB Moneyball analyst @arikaplan1: [conferences.oreilly.com/artificial-int ...](http://conferences.oreilly.com/artificial-int ...)



8:39 PM - 24 Aug 2018

Following ▾



David Kearns  
@DaithiOCiaran

Following ▾

Today is the day #TheAIConf. Want to hear how #MachineLearning can be applied to #Moneyball. Join myself, @arikaplan1 @chriscoad and @ledell @IBMDatascience @h2oai @Aginity 4.50pm Location: Imperial A



3:29 PM - 6 Sep 2018

# $H_2O$ .ai in Munich this Week

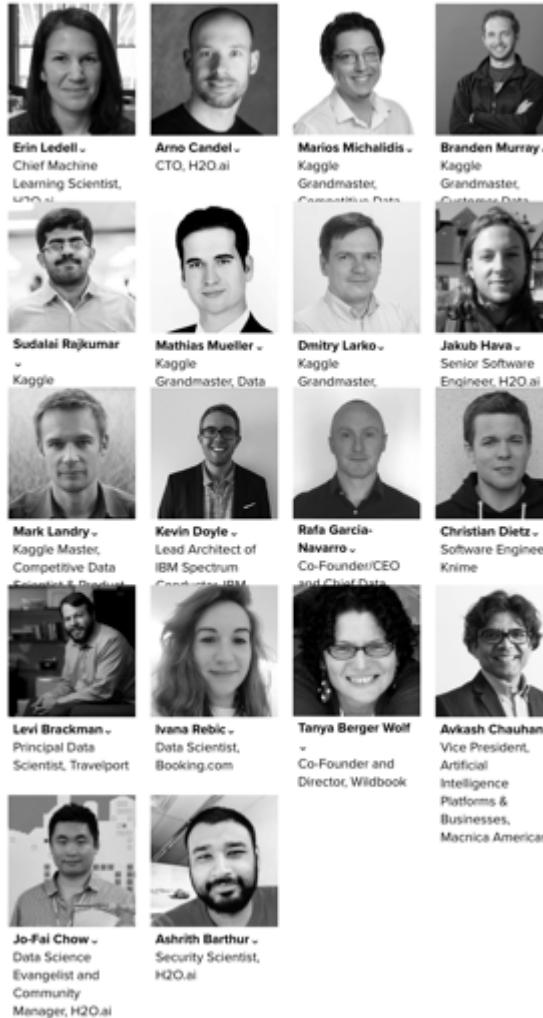
- **Today (8<sup>th</sup> Oct)**
  - Applied R Munich Meetup (Thanks!)
- **Tuesday (9<sup>th</sup> Oct)**
  - IBM PowerAI Munich Meetup
- **Wednesday + Thursday (10 and 11<sup>th</sup> Oct)**
  - GTC Europe (Booth E.03)

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



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

Session Speakers



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

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

# Thanks!

- e.on
- Applied R Munich
- More Info, Code, and Slides
  - [bit.ly/  
h2o\\_meetups](https://bit.ly/h2o_meetups)
- Contact
  - [joe@h2o.ai](mailto:joe@h2o.ai)
  - [@matlabulous](https://twitter.com/matlabulous)
  - [github.com/woobe](https://github.com/woobe)