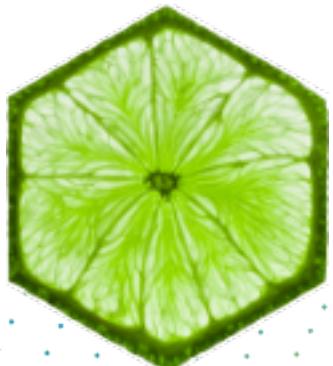


Automatic and Interpretable Machine Learning in R with H₂O and LIME



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)



Why?

- Most users/organizations can benefit from automatic machine learning pipelines.
 - Eliminate time wasted on human errors, debugging etc.
- Model interpretations is crucial for those who must explain their models to regulators or customers.

You will learn ...

- How to build high quality H₂O models (almost) automatically.
- How to explain predictions from complex H₂O models with LIME.
- **Bonus:** A real use-case that led to multimillion-dollar baseball decisions earlier this year.

AutoML + LIME + Shiny = Real Moneyball

\$20M

trade 2 weeks prior to the
season beginning



About Me

Barcelona



Berlin



Brussels

Paris

Jo-fai (Joe) Chow
@matlabulous

Milan (2016)

Thx @DataScienceMi #datasciencemilan
@h2oai bit.ly/h2o_milan_1
#AroundTheWorldWithH2Oai #t...
ift.tt/2dTKWDh



1:29 PM - 10 Oct 2016

- **Before H₂O**

- Water Engineer / EngD Researcher / Matlab Fan Boy (wonder why @matlabulous?)
- Discovered R, Python, H₂O ... never look back again
- Data Scientist at Virgin Media (UK), Domino Data Lab (US)

- **At H₂O ...**

- Data Scientist / Evangelist /
 - Sales Engineer / Solution Architect /
 - Community Manager
- ... The harsh reality of startup life ...
- **H₂O SWAG Photographer**
#AroundTheWorldWithH2Oai
Love H₂O? 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



Budapest

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



London

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 💪



Berlin

1:54 PM - 7 Mar 2018 from Hotel Berlin



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

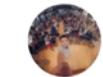
@matlabulous

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

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-moneyball-decisions-with-h2o-automl-lime-and-shiny)

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



Cologne

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

Reminder: #360Selfie

H₂O.ai

Agenda

Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H ₂ O, AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	🍕🍕🍕🍕🍕
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H ₂ O News + Q & A





Time	Topics / Tasks
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018

LIME

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

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

H2O

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

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

... and **mlbench** for datasets



Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H_2O , AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	🍕🍕🍕🍕🍕
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H_2O News + Q & A

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.

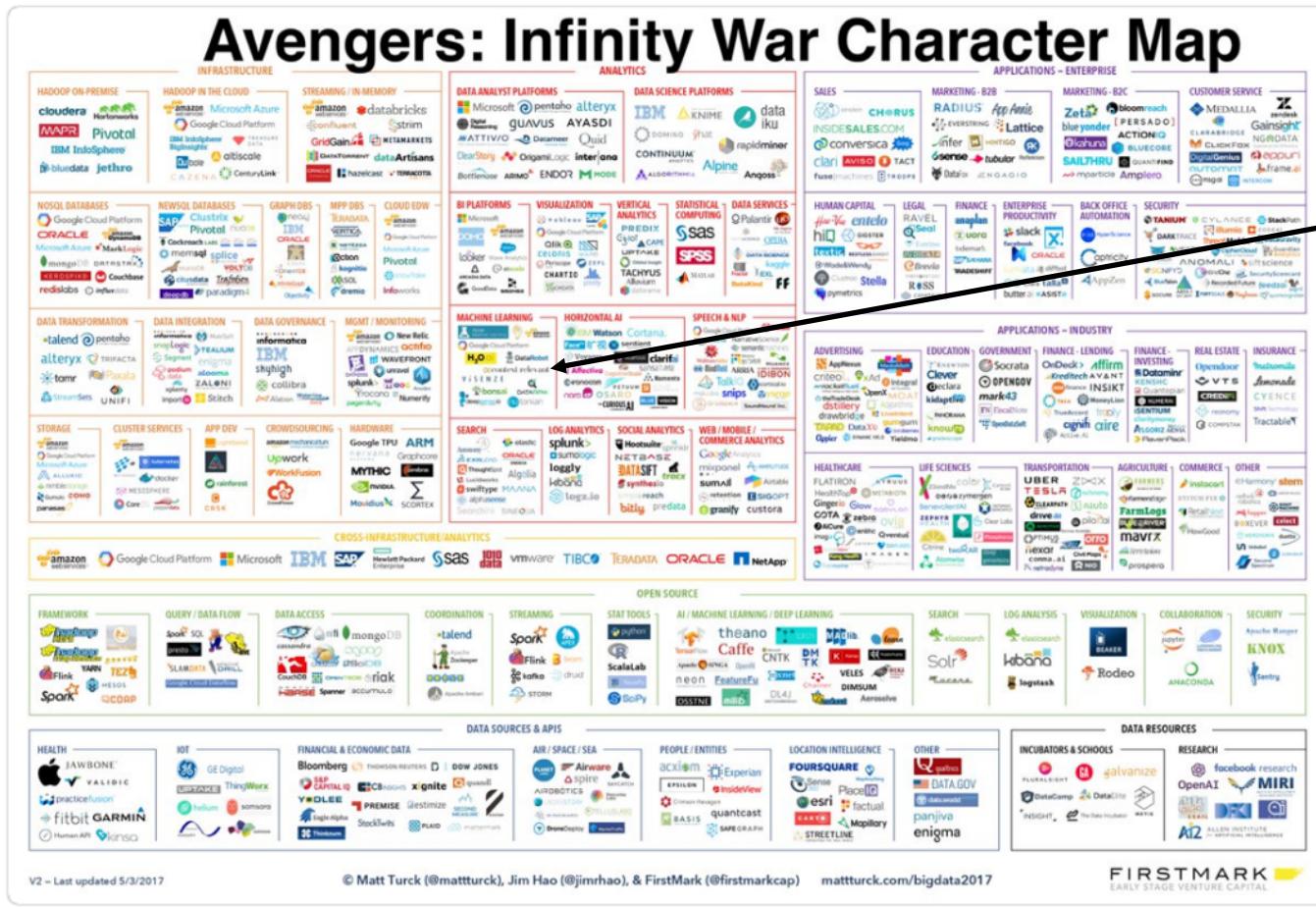
Vicki Boykis
@vboykis

Follow

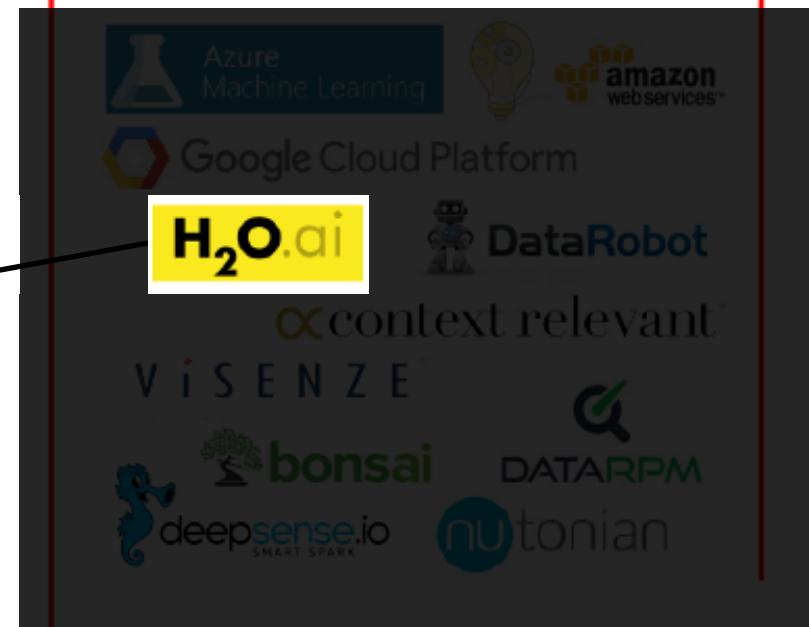
1

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

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 H₂O 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



H₂O + 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

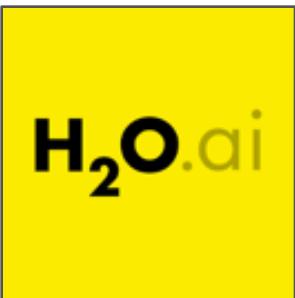
H₂O.ai

Company Overview

Founded	2012, Series C in Nov, 2017
Products	<ul style="list-style-type: none">• Driverless AI – Automated Machine Learning• H₂O Open Source Machine Learning• Sparkling Water
Mission	Democratize AI. Do Good
Team	<p>~100 employees</p> <ul style="list-style-type: none">• Distributed Systems Engineers doing Machine Learning• World-class visualization designers
Offices	Mountain View, London, Prague



H₂O 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



This Workshop

H₂O Products

H₂O.ai

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

Spark + H₂O
SPARKLING
WATER

H2O AI Open Source Engine
Integration with Spark

H₂O4GPU

Lightning Fast machine
learning on GPUs

DRIVERLESSAI

Automatic feature
engineering, machine
learning and interpretability

Steam

Secure multi-tenant H₂O clusters

H₂O.ai

Worldwide Recognition in the H2O.ai Community

Open Source
Community

222 OF FORTUNE
THE 500
 H₂O

8 OF TOP 10
BANKS

7 OF TOP 10
INSURANCE COMPANIES

4 OF TOP 10
HEALTHCARE COMPANIES

CATALINA

DIRECT
MAILERS

G5

acxiom.

poder.IO
POWER OF PREDICTION

Integral
Ad Science

Nielsen
Catalina
SOLUTIONS



CONFIDENTIAL

COMCAST

CISCO

HW
Vendors

STANLEY
BLACK & DECKER

H-E-B

Travelpoint

Walgreens

eBay

Booking.com

macy's

CREDIT SUISSE

WELLS
FARGO

CITI

deserve

experian.
RBC

EQUIFAX

MarketAxess®

ING

DISCOVER

CapitalOne™

PayPal™

ZURICH®

CONFIDENTIAL

TRANSAMERICA

PROGRESSIVE

ARMADA
Health Care

aetna®

starling

ADP

pwc

Paying Customers

opta
INFORMATION INTELLIGENCE

KAI SER PERMANENTE.

CHANGE
HEALTHCARE

ARMADA
Health Care

aetna®

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

Why H₂O?



Steph Locke
@SteffLocke

Following

My #rstats #datascience goto
IO: odbc readxl httr
EDA: DataExplorer
Prep: tidyverse
Sampling: rsample modelr
Feature Engineering: recipes
Modelling: glmnet **h2o** FFTrees
Evaluation: broom yardstick
Deployment: sqlrutils AzureML opencpu
Monitoring: flexdashboard
Docs: rmarkdown

4:29 PM - 28 Apr 2018

143 Retweets 591 Likes



10

143

591



Szilard
@DataScienceLA

Following

Friday fun: what's your favorite gradient boosting machine (GBM) library?

58% xgboost

16% lightgbm

24% h2o

2% spark mllib

127 votes • Final results

11:21 PM - 11 May 2018

9 Retweets 9 Likes



Q

3

11

9

1

✉



Tweet your reply



Arno Candel @ArnoCandel · May 12

Replies to @DataScienceLA

Did you know? H2O-3 has XGBoost integration (incl. support for GPU and distributed mode) with standalone Java scoring (MOJO) - train from Flow, R or Python. H2O AutoML and Driverless AI use XGBoost too, and @h2oai contributes to XGBoost in collaboration with @nvidia #h2o4gpu

Q

11

11

✉

Our Mission: Make Machine Learning Accessible to Everyone



Complexity is your enemy. Any fool can make something complicated. It is hard to keep things simple.

— *Richard Branson* —

AZ QUOTES

Scientific Advisory Council



Dr. Trevor Hastie

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



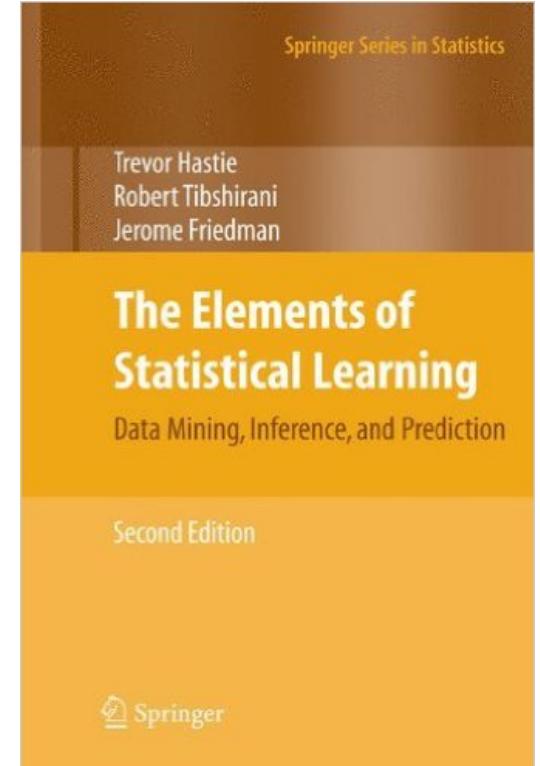
Dr. Robert Tibshirani

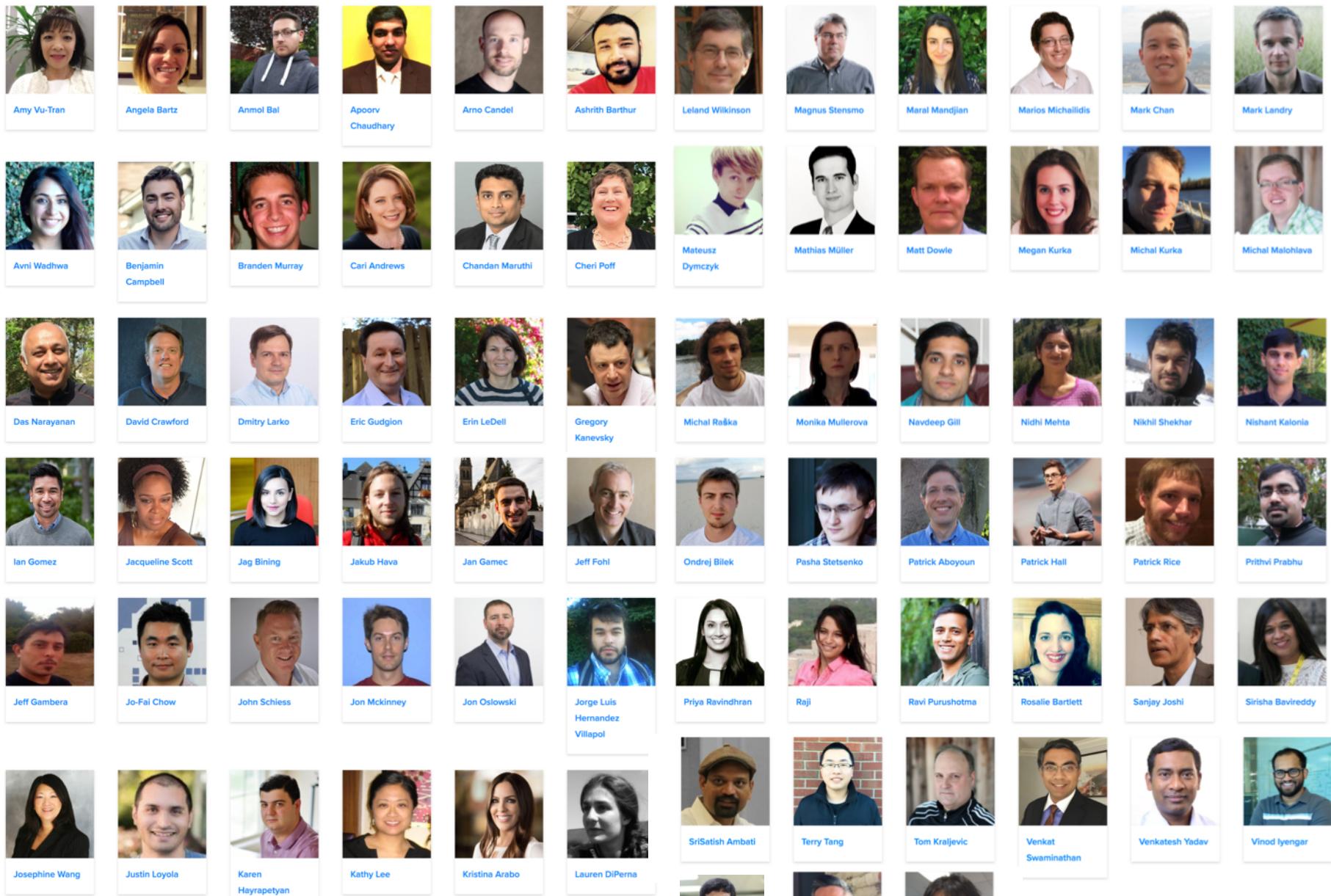
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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



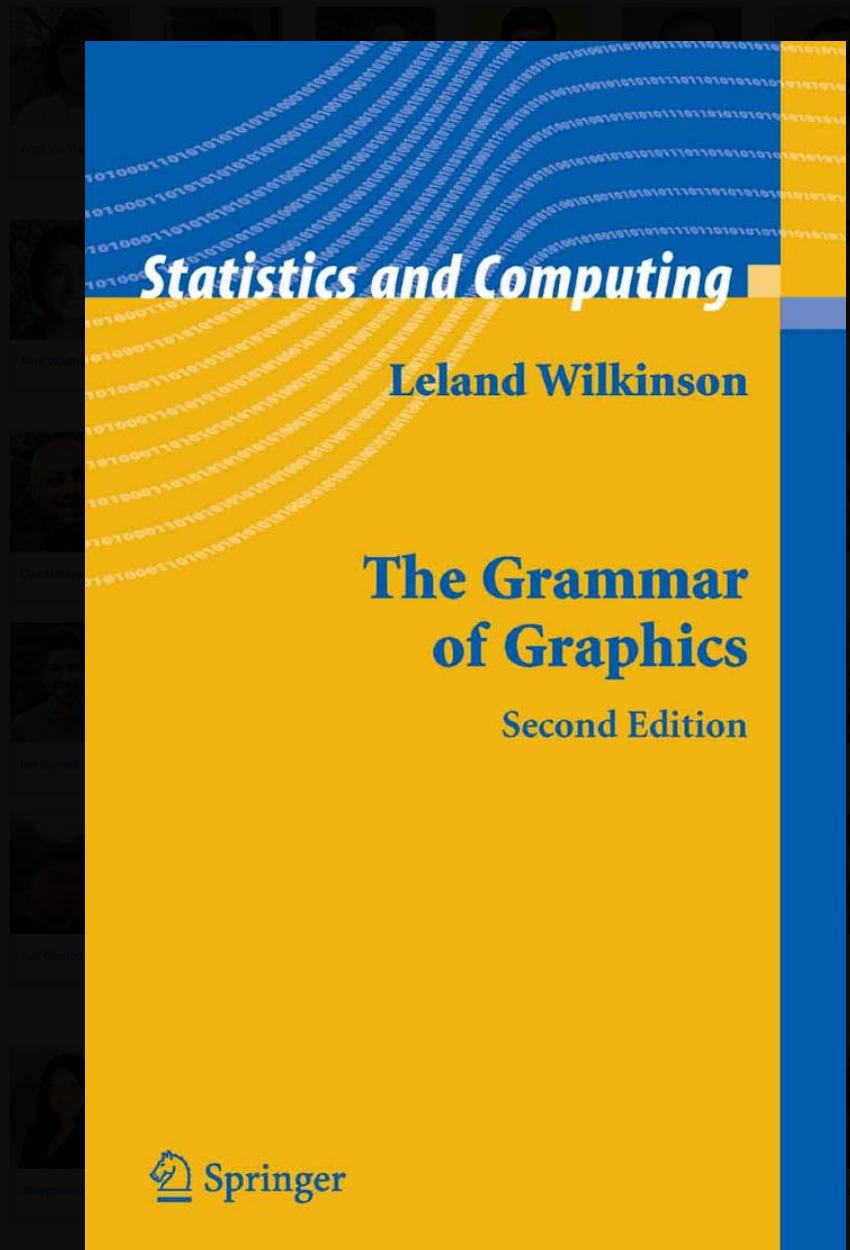
Dr. Steven Boyd

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





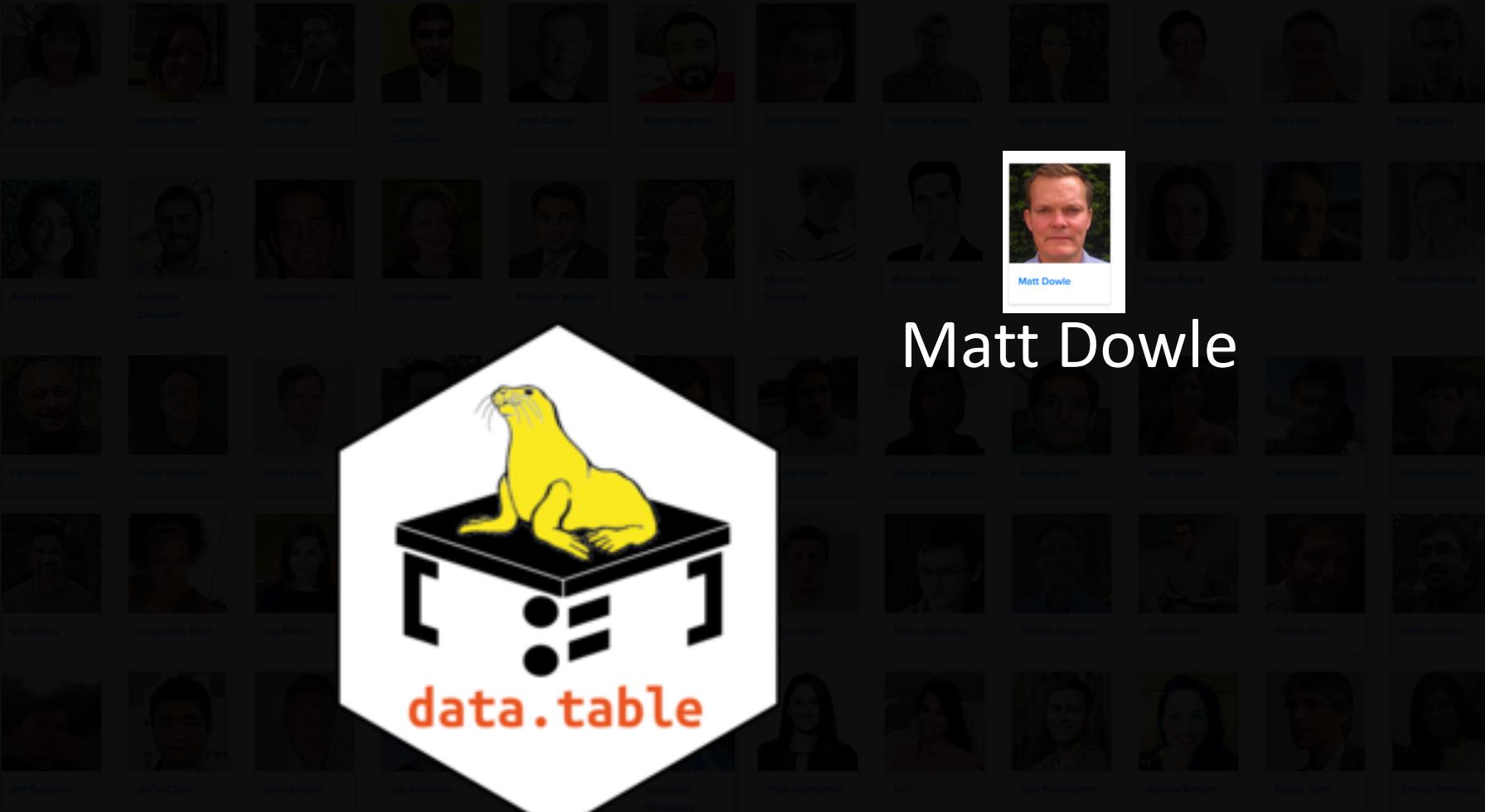
H₂O Team



Leland Wilkinson

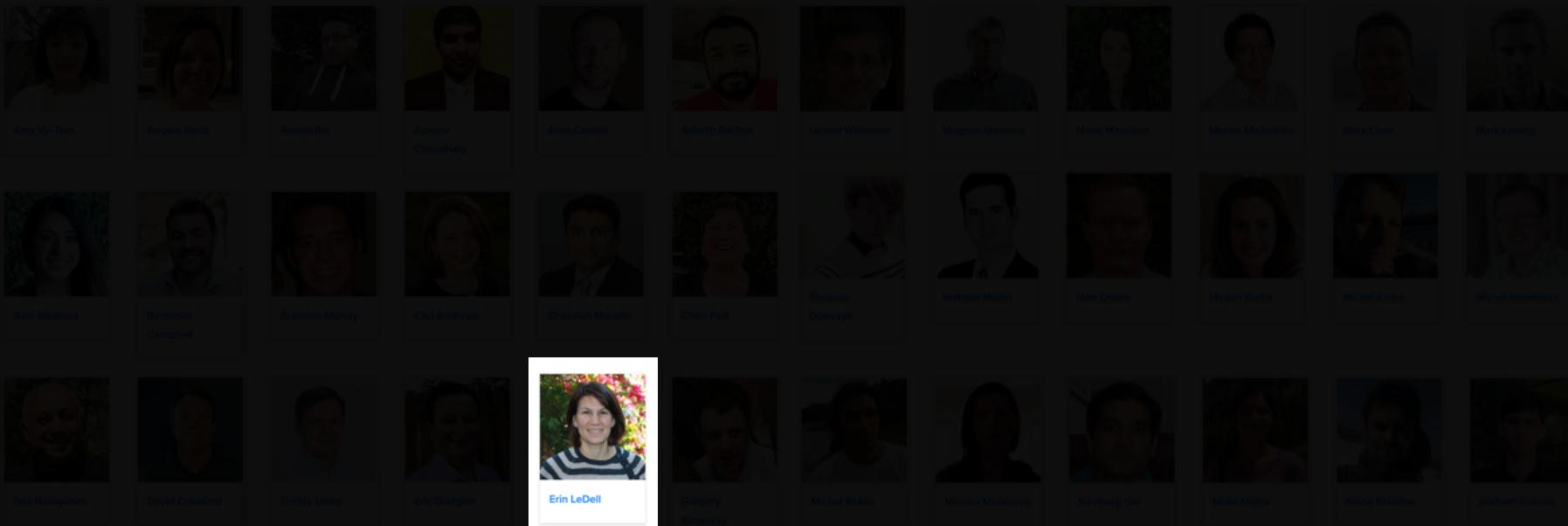
Origin of R Package `ggplot2`





Matt Dowle

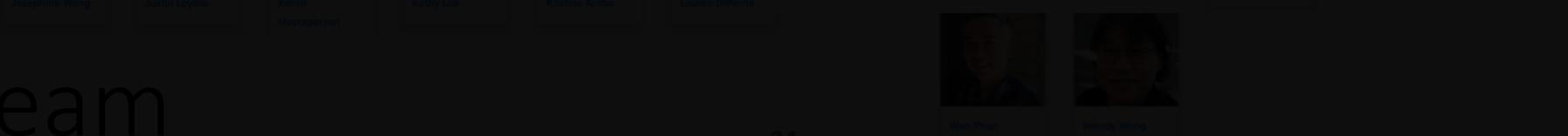
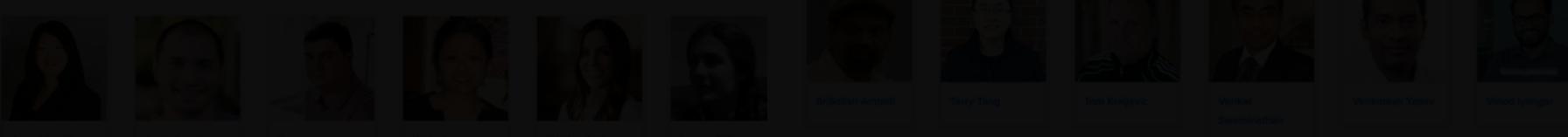
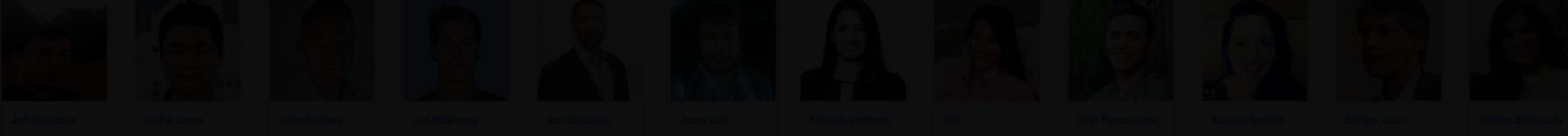
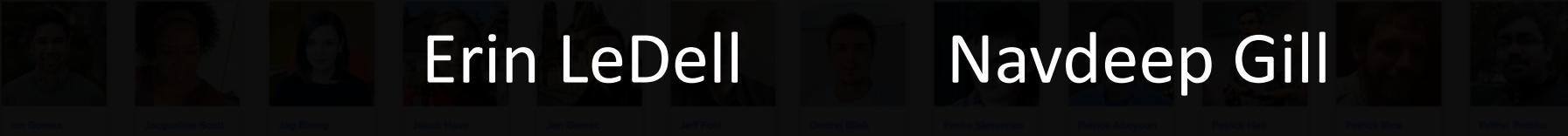
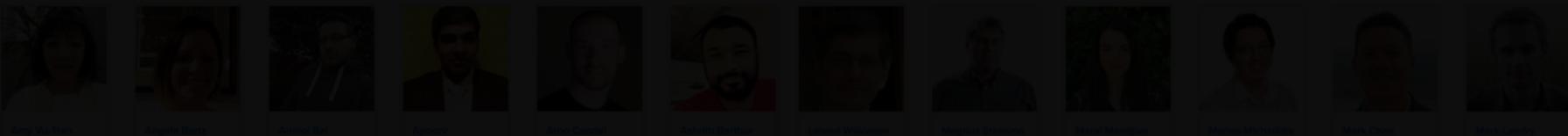
H₂O Team



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



H₂O Team



H₂O AutoML

Erin LeDell

Navdeep Gill

H₂O Team

Kaggle Grand Masters (and their Highest Rank)



113
Grandmasters



980
Masters



3,339
Experts



46,135
Contributors



33,242
Novices

About 80,000 Kagglers

H₂O Team

H₂O.ai



Amy Vu-Tran



Angela Bartz



48th



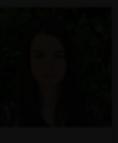
Arno Candel



Arshin Barthar



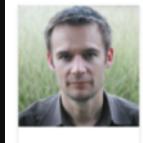
Leland Wilkinson

Magnus Stenman
Mari Mardian

Marios Michailidis



Mark Chan



Mark Landry



Aarti Wedher



Benjamin Campbell



Branden Murray



Carl Andrews



Chandan Manathil



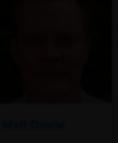
Chen Poff



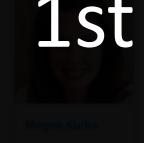
Mateusz Dymczyk



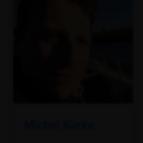
Mathias Müller



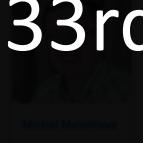
Matt Doyle



Megan Kunka



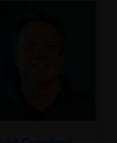
Michael Kunka



Michal Molontava



Das Narayanan



David Crawford



Dmitry Larko



Eric Gudgion



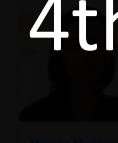
Erin LeDell



Gregory Kanovsky



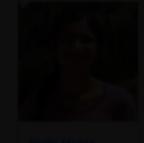
Michal Ratajka



Monika Mullerova



Navdeep Gill



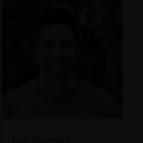
Nishi Mehta



Nisha Shukhar



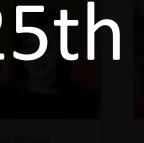
Nishant Kalra



Ian Gomez



Jacqueline Scott



Jag Birring



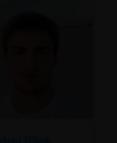
Jaihun Hong



Jen Gamec



Jeff Ford



Chandan Bhattacharya



Piotr Bojanowski



Patrick Abeyasekera



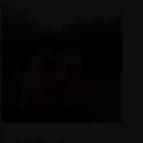
Patrick Hall



Patrick Rice



Prithviraj Dasgupta



Jeff Gambetta



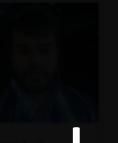
Jo-Fai Chow



Joelle Pineau



Jitendra Malik



Kristina Arias



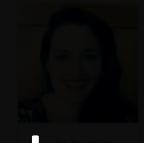
Lauren DiPerna



Srikumar Ramamurthy



Terry Ting



Tom Kraljevic



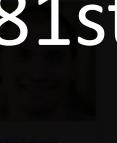
Venkatraman Venkateswaran



Virendra Tyagi



Josephine Wang



Justin Loyola



Karen Heysepian



Kathy Lee



Kristina Arias



Lauren DiPerna



13th



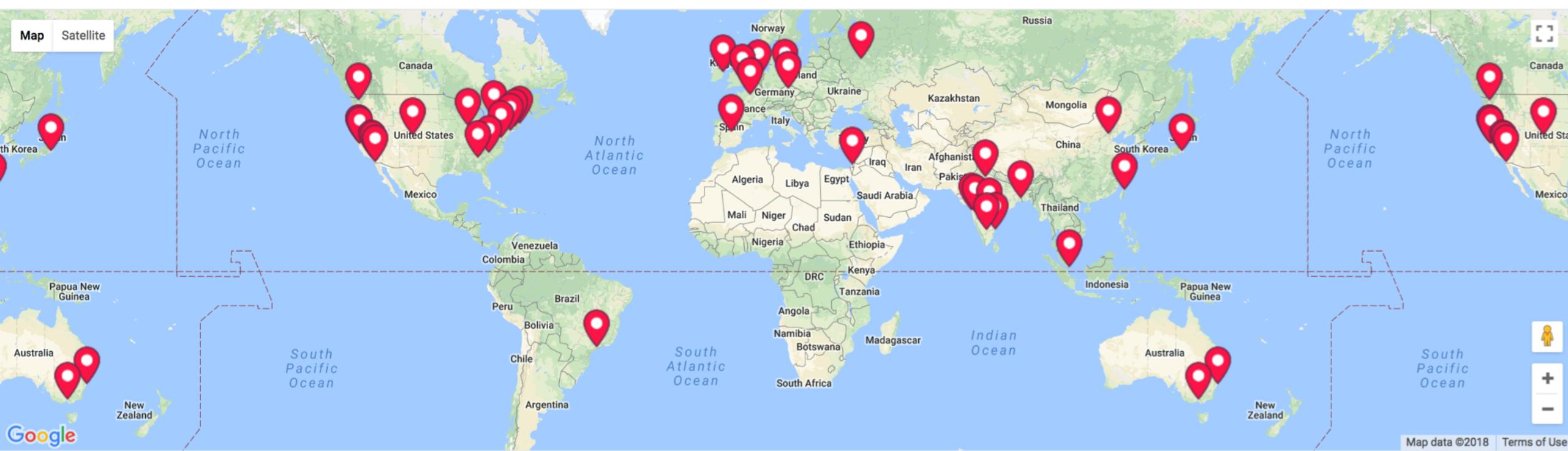
Wen Phan



Wendy Wong

H₂O TeamH₂O.ai

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



H2O Artificial Intelligence and Machine Learning

Members
78,356

Groups
39

Countries
18

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

We sponsor meetups



R-Ladies London
Location: London, United Kingdom
Members: 1,040
Organizers: Chihin Tan and 7 others



Artificial Intelligence (AI) Club for Gender Minorities!
Location: London, United Kingdom - 489 members - Public group
Organized by: Chihin Tan and 5 others

Share: [Facebook](#) [Twitter](#) [LinkedIn](#)

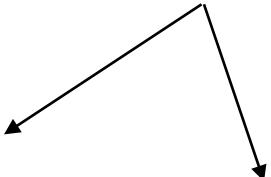
London Data Science Workshop (formerly London Kaggle Meetup)
Location: London, United Kingdom
Members: 2,783
Organizers: Alex Glaser and 6 others



Women in Kaggle
Location: London, United Kingdom
Members: 261
Organizers: Julia MacMillan and 3 others

... and more

Source: <https://www.maddycoupons.in/blog/yummy-yummy-pizza/>



We encourage diversity

Joe's Meetups	Female Speaker	Female Speaker Ratio
London Dec 2017	Kasia Kulma	1/3
Amsterdam Feb 2018	Andreea Bejinaru	1/2
London Mar 2018	Cheuk Ting Ho	1/3
London Jun 2018	Torgyn Shaikhina	1/4
Overall (so far) = 4/12 = 33.3%		

Encourage your friends/colleagues to give a talk.

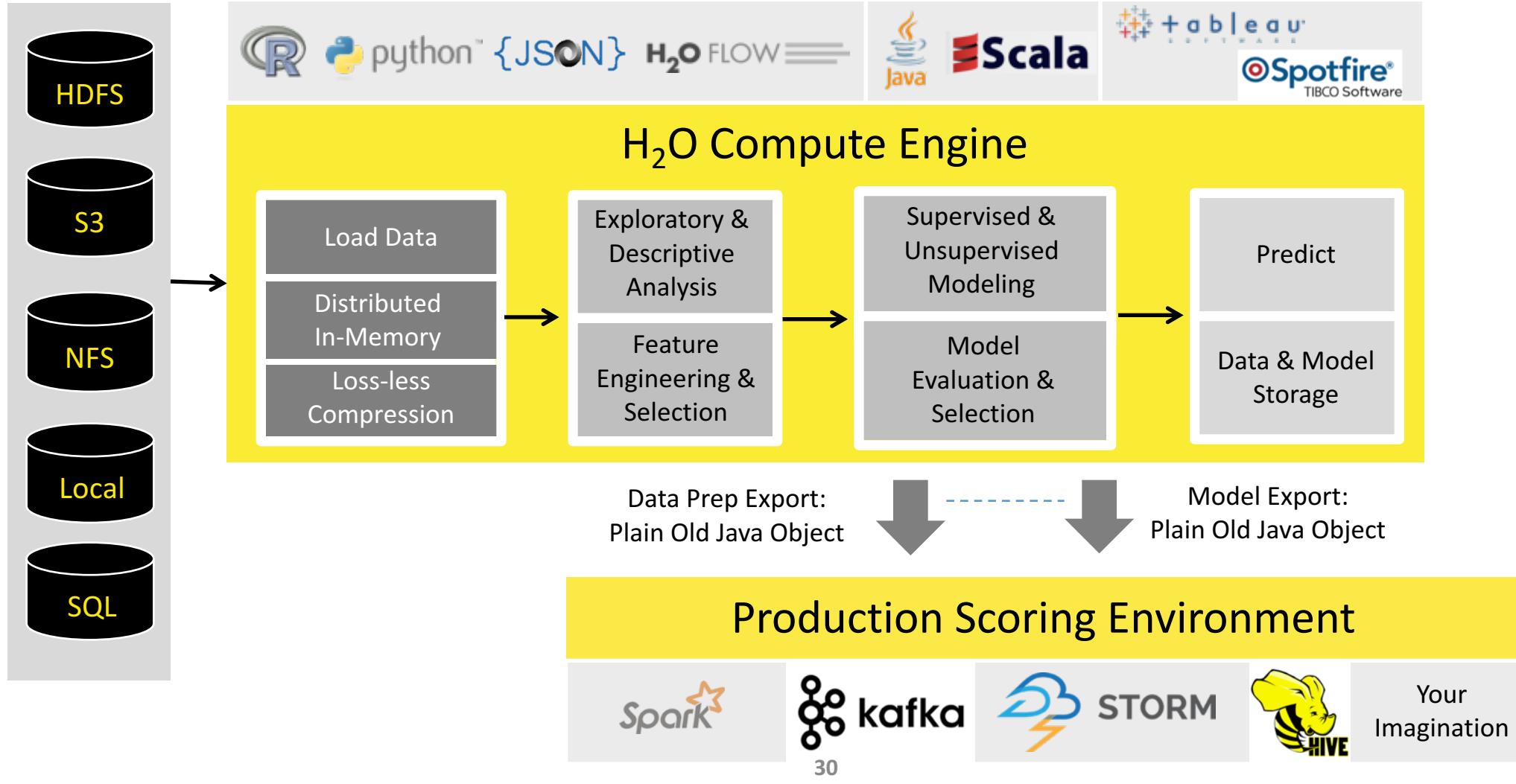
joe@h2o.ai

About H₂O AutoML

Automatic Machine Learning with H₂O

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

High Level Architecture



H₂O-3 Algorithms Overview

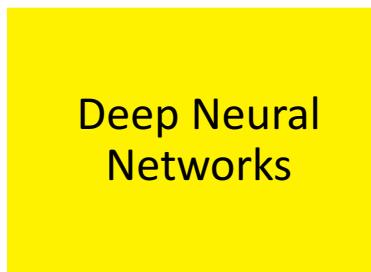
Supervised Learning



- **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

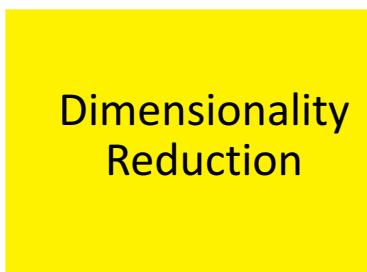


- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

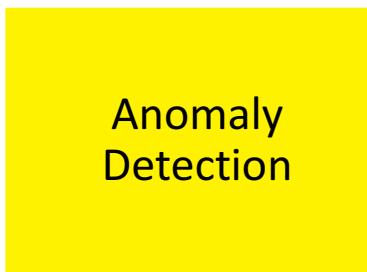
Unsupervised Learning



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

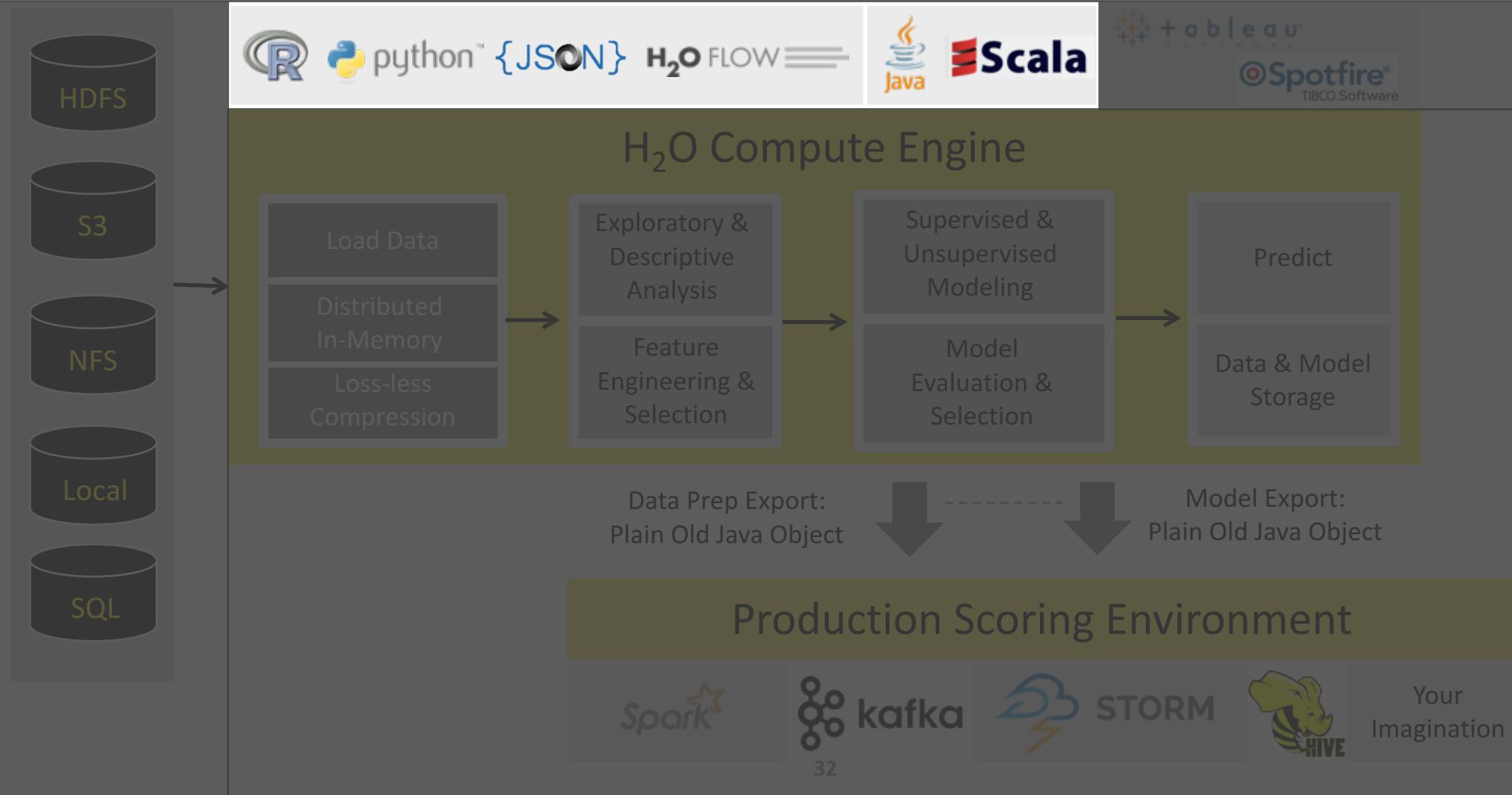


- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data



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

High Level Architecture



H₂O Flow (Web)

The screenshot shows the H2O Flow (Web) interface running in a browser window. The title bar reads "H2O Flow" and the address bar shows "localhost:54321/flow/index.html". The top navigation bar includes "H2O FLOW", "Flow", "Cell", "Data", "Model" (which is highlighted in yellow), "Score", "Admin", and "Help". A context menu is open under the "Model" dropdown, listing various machine learning models and utilities. The main workspace on the left is titled "Untitled Flow" and contains a single step labeled "assist". Below this is a table titled "Assistance" with a list of routines and their descriptions. The right side of the interface features a sidebar with sections for "OUTLINE", "FLOWS", "CLIPS", and "HELP" (which is also highlighted in yellow). The "HELP" section includes links for "Using Flow for the first time?", "Quickstart Videos", and "view example Flows". It also has sections for "GENERAL" (with links to Flow Web UI, Importing Data, Building Models, Making Predictions, Using Flows, and Troubleshooting Flow) and "EXAMPLES" (describing Flow packs and providing a link to Browse installed packs...). The bottom right corner shows "Connections: 0" and the H2O logo.

Model

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

ROUTINE

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

OUTLINE FLOWS CLIPS HELP

Using Flow for the first time?

Quickstart Videos

Or, view example Flows to explore and learn H₂O.

STAR H₂O ON GITHUB!

Star 2,387

GENERAL

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

EXAMPLES

Flow packs are a great way to explore and learn H₂O. Try out these Flows and run them in your browser.

Browse installed packs...

localhost:54321/flow/index.html#

Connections: 0 H₂O

Using H₂O with R and Python

The screenshot shows the RStudio Source Editor window with the file `credit_card_example.R` open. The code is an R script for a credit card example, demonstrating the use of the `h2o` package. It imports datasets from S3, starts a local H2O cluster, trains a GBM model, and prints the leaderboard. The code is well-structured with comments explaining each step.

```
1 # Credit Card Example
2
3 # Datasets:
4 # https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv
5 # https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv
6
7 # Start and connect to a local H2O cluster
8 library(h2o)
9 h2o.init(nthreads = -1)
10
11 # Import datasets from s3
12 df_train = h2o.importFile("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv")
13 df_test = h2o.importFile("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv")
14
15 # Look at datasets
16 summary(df_train)
17 summary(df_test)
18
19 # Define features and target
20 features = colnames(df_test)
21 target = "DEFAULT_PAYMENT_NEXT_MONTH"
22
23 # Train a GBM model
24 model_gbm = h2o.gbm(x = features,
25                      y = target,
26                      training_frame = df_train,
27                      seed = 1234)
28 print(model_gbm)
29
30 # Use GBM model for making predictions
31 yhat_test = h2o.predict(model_gbm, newdata = df_test)
32 head(yhat_test)
33
34 # (Extra) Use H2O's AutoML
35 aml = h2o.automl(x = features,
36                   y = target,
37                   training_frame = df_train,
38                   max_runtime_secs = 60,
39                   seed = 1234)
40
41 # Print leaderboard
42 print(aml@leaderboard)
43
44 # Use best model for making predictions
45 best_model = aml@leaderboard[[1]]
46 yhat_test = h2o.predict(best_model, newdata = df_test)
47 head(yhat_test)
48
49
```

The screenshot shows a Jupyter Notebook interface with the notebook `credit_card_example.ipynb` open. The notebook contains Python code for a credit card example, which includes starting a local H2O cluster, importing datasets, and summarizing them. The output of the first cell shows the Java version and the successful connection to the H2O server. The second cell shows the import progress of the datasets. The third cell displays the summary statistics for the datasets, including the number of rows and columns, and the fourth cell shows the detailed summary for the `df_train` dataset.

```
In [2]: # Start and connect to a local H2O cluster
import h2o
h2o.init(nthreads = -1)

Checking whether there is an H2O instance running at http://localhost:54321.... not found.
Attempting to start a local H2O server...
Java Version: java version "1.8.0_72"; Java(TM) SE Runtime Environment (build 1.8.0_72-b15); Java HotSpot(TM) 64-Bit Server VM (build 25.72-b15, mixed mode)
Starting server from /Users/jofaichow/anaconda/lib/python2.7/site-packages/h2o/backend/bin/h2o.jar
Ice root: /var/folders/4z/p7yt7_4n4fjijiy6g4qfbw000gn/T/tmpPdP3Av
JVM stdout: /var/folders/4z/p7yt7_4n4fjijiy6g4qfbw000gn/T/tmpPdP3Av/h2o_jofaichow_started_from_python.out
JVM stderr: /var/folders/4z/p7yt7_4n4fjijiy6g4qfbw000gn/T/tmpPdP3Av/h2o_jofaichow_started_from_python.err
Server is running at http://127.0.0.1:54321
Connecting to H2O server at http://127.0.0.1:54321... successful.

H2O cluster uptime: 02 secs
H2O cluster version: 3.13.0.3981
H2O cluster version age: 29 days
H2O cluster name: H2O_from_python_jofaichow_id7qa
H2O cluster total nodes: 1

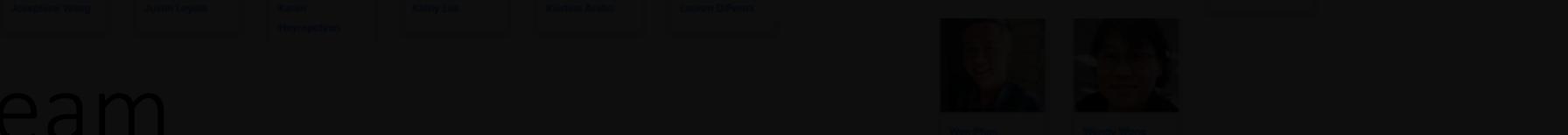
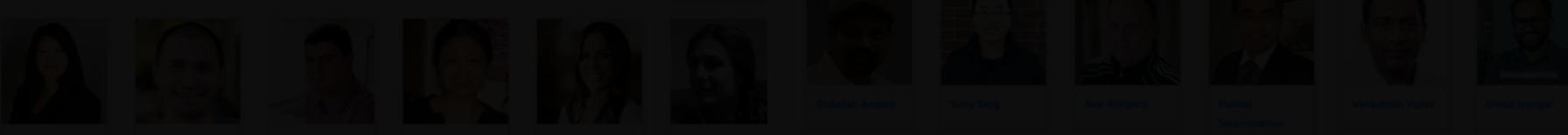
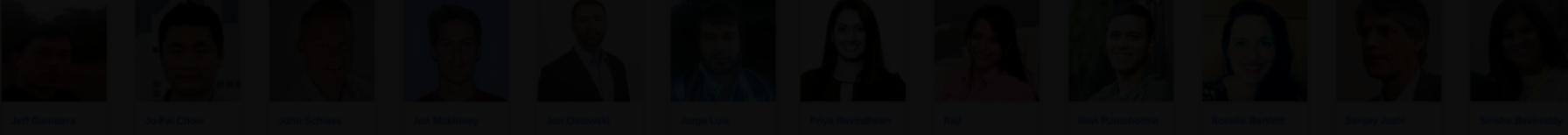
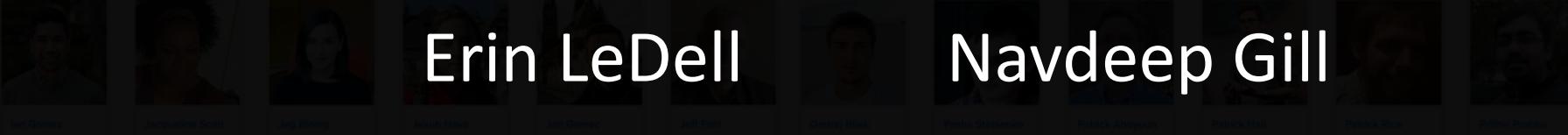
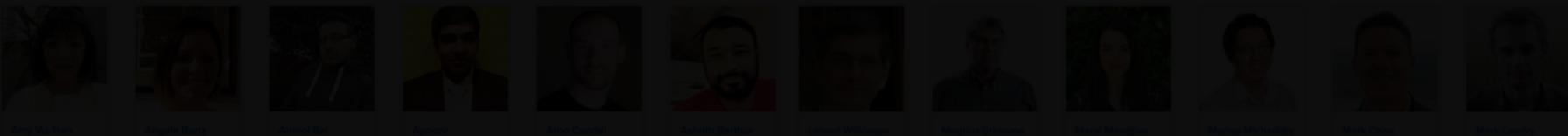
In [3]: # Import datasets from s3
df_train = h2o.import_file("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv")
df_test = h2o.import_file("https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv")

Parse progress: |██████████| 100%
Parse progress: |██████████| 100%

In [4]: # Look at datasets
df_train.summary()
df_test.summary()


```

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4
type	int	enum	int	int	int	int	int	int	int
mins	10000.0		0.0	0.0	21.0	-2.0	-2.0	-2.0	-2.0
mean	165471.466667		1.85	1.55578703704	35.4053240741	-0.00523148148148	-0.122361111111	-0.15537037037037	-0.210601
maxs	1000000.0		6.0	3.0	79.0	8.0	8.0	8.0	8.0
sigma	128853.314839		0.779559696278	0.522505078476	9.27675421641	1.12668964211	1.20086854503	1.20727030901	1.172176
zeros	0		9	37	0	10563	11284	11309	11905
missing	0		0	0	0	0	0	0	0



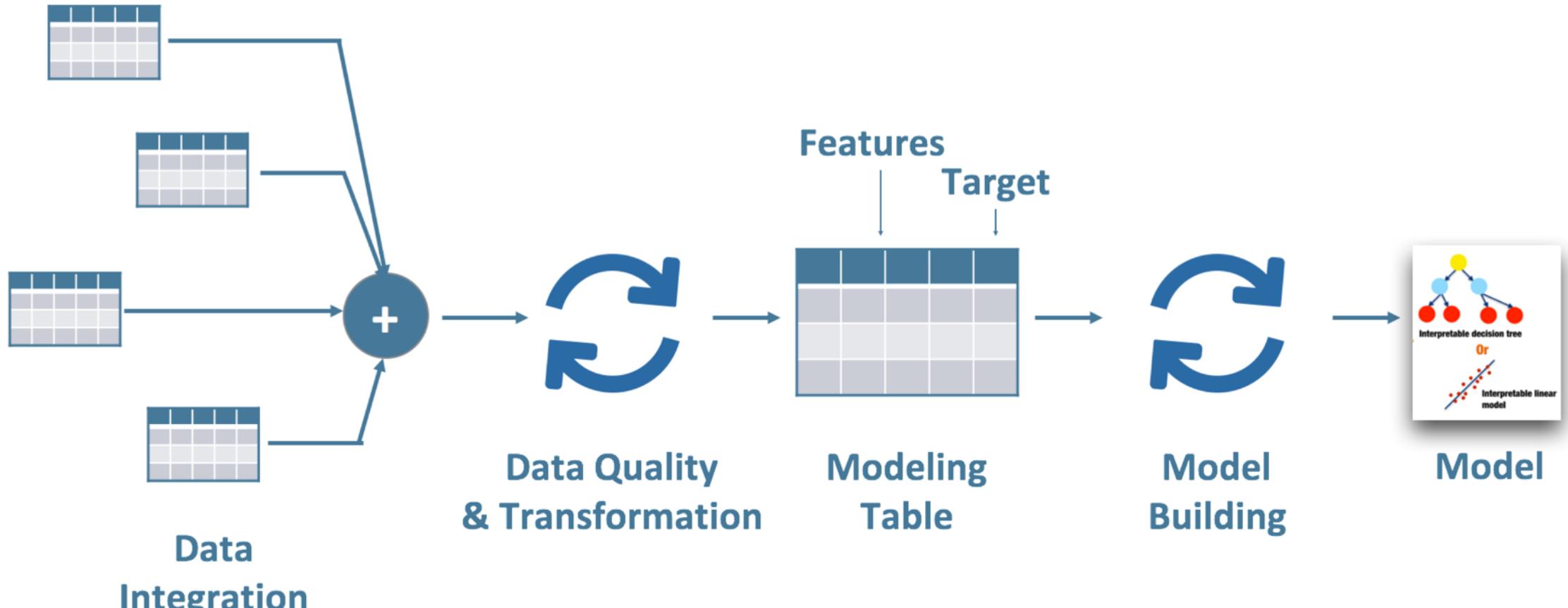
H₂O AutoML

Erin LeDell

Navdeep Gill

H₂O Team

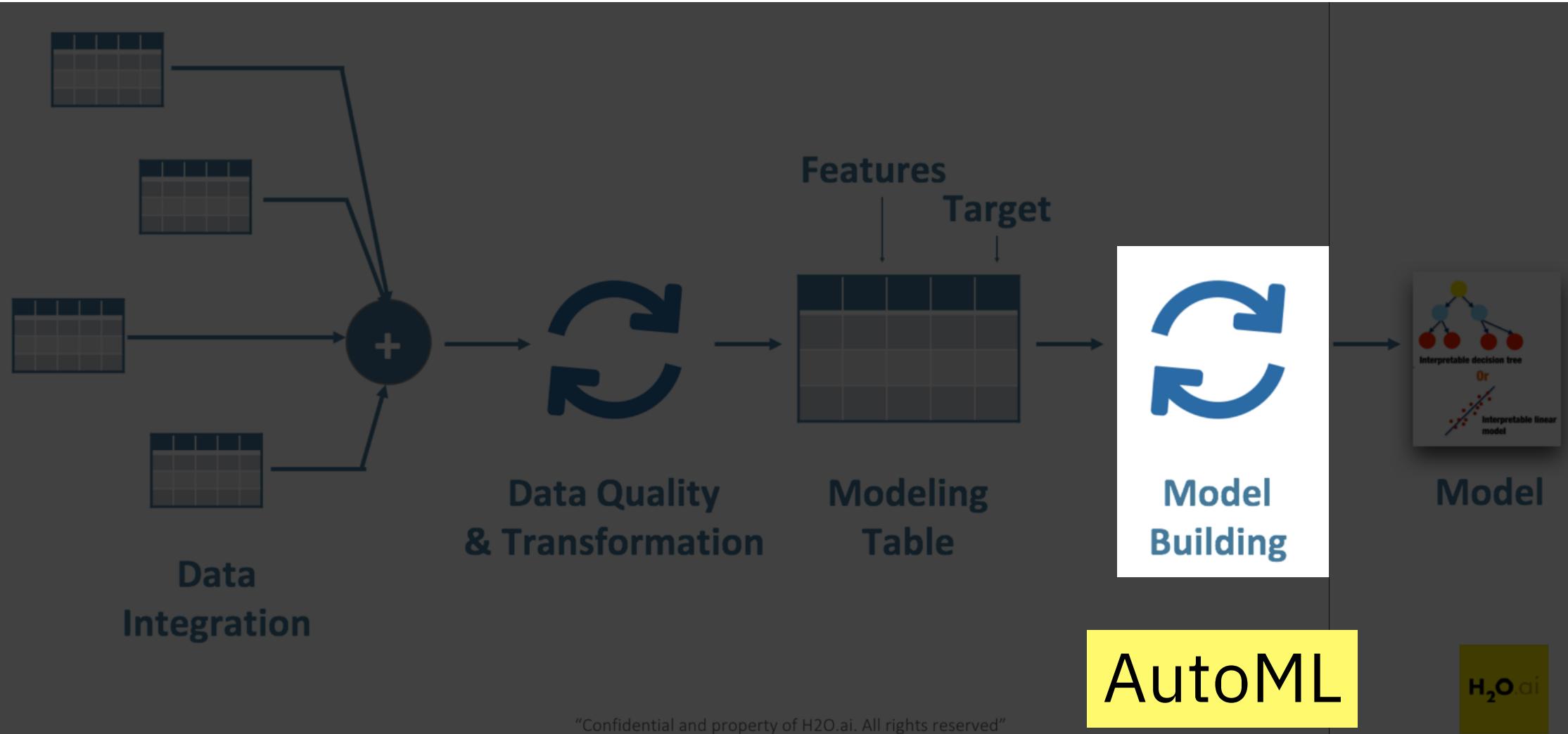
Typical Enterprise Machine Learning Workflow



“Confidential and property of H2O.ai. All rights reserved”



Typical Enterprise Machine Learning Workflow



3.3 H2O AutoML: Multiple H2O Models + Stacked Ensemble

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

3.4 AutoML Leaderboard

```
model_automl@leaderboard
```

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

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

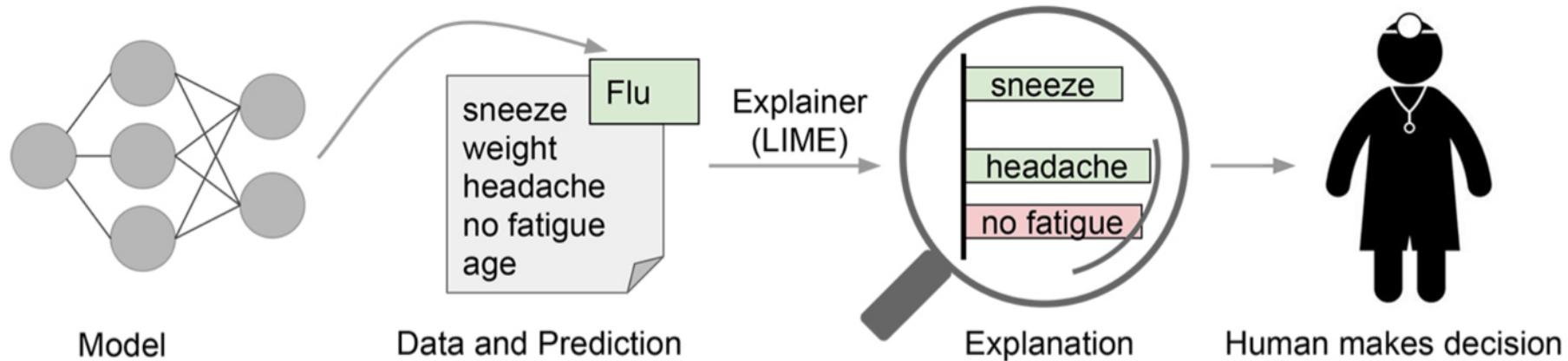


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

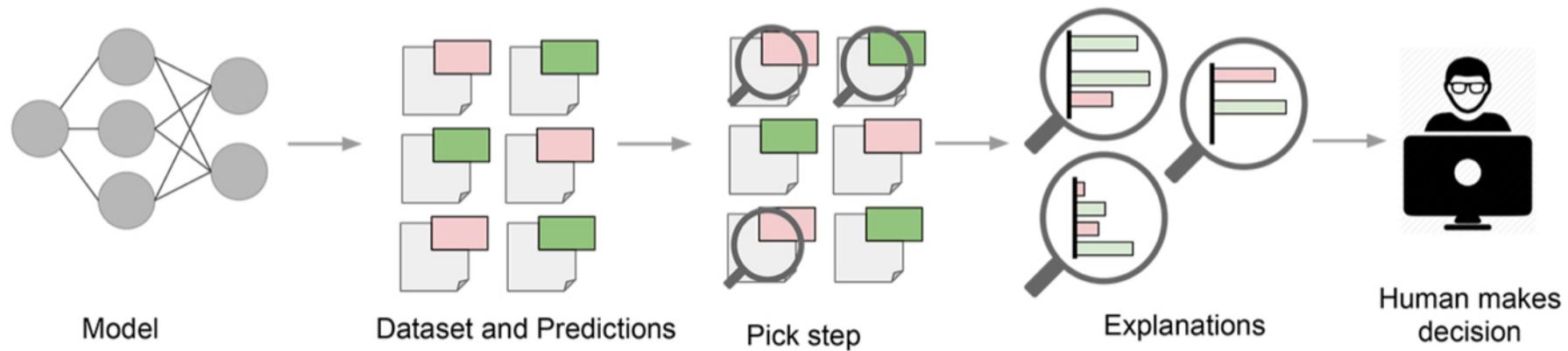


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

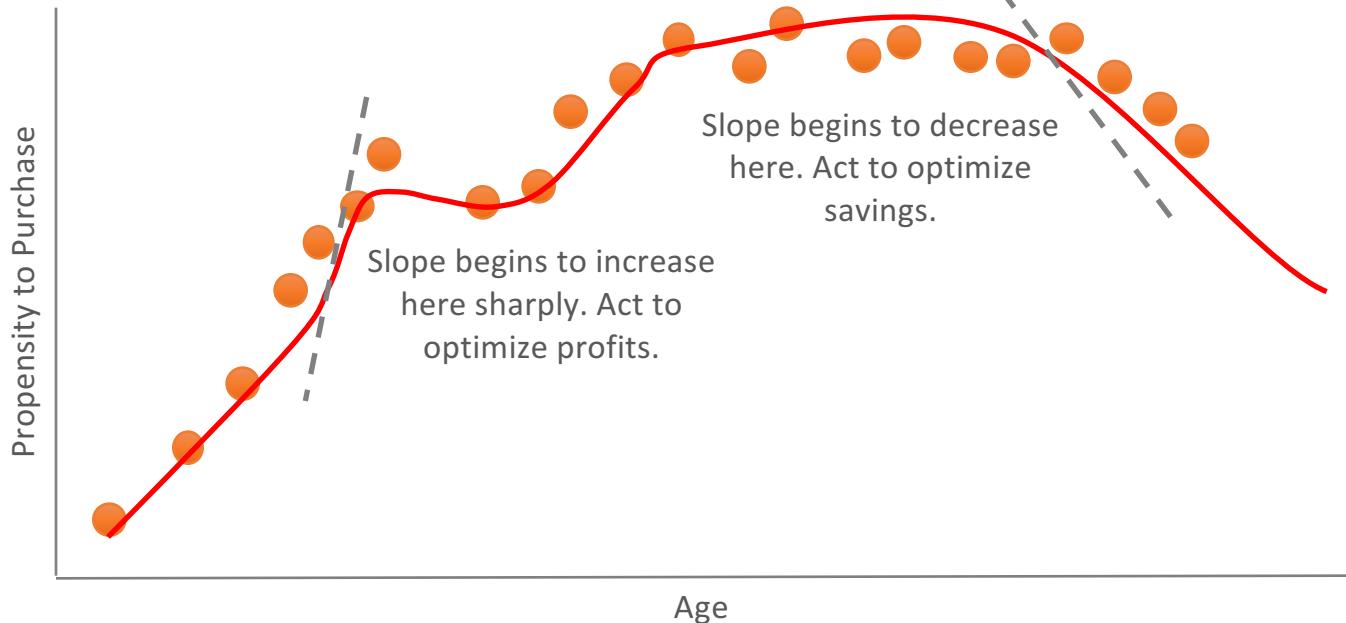
Linear Models

Exact explanations for approximate models.



Machine Learning

Approximate explanations for exact models.



Local Interpretable Model-Agnostic Explanations

LIME - How does it work?

Theory

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

Outcome

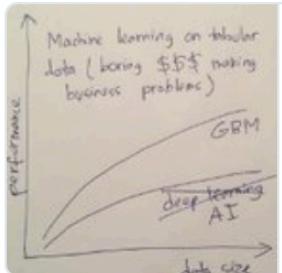
- Approximate reasoning
- Complex models can be interpreted
 - Neural nets, Random Forest, Ensembles etc.



Patrick Hall
@jpatrickhall

Following

I made an #awesome list for #MachineLearning interpretability (#XAI, #FATML). Help me update and curate it! [github.com/jphall663/awes ...](https://github.com/jphall663/awesome-machine-learning-interpretability) (And if an #awesome list for this already exists please point me there ... I will merge). #DataScience #rstats #Python



jphall663/awesome-machine-learning-interpretability

awesome-machine-learning-interpretability - A curated list of awesome machine learning interpretability resources.

[github.com](https://github.com/jphall663/awesome-machine-learning-interpretability)

5:40 PM - 21 Jun 2018

awesome-machine-learning-interpretability

A curated list of awesome machine learning interpretability resources.

If you want to contribute to this list (*and please do!*) read over the [contribution guidelines](#), send a pull request, or contact me @jpatrickhall.

Table of Contents

- [Comprehensive Software Examples and Tutorials](#)
- [Interpretability and Fairness Software Packages](#)
 - [Python](#)
 - [R](#)
- [Free Books](#)
- [Other Interpretability and Fairness Lists](#)
- [Review Papers](#)
- [Whitebox Modeling Packages](#)
 - [Python](#)
 - [R](#)

Comprehensive Software Examples and Tutorials

- [Getting a Window into your Black Box Model](#)
- [IML](#)
- [Interpretable Machine Learning with Python](#)

Interpretability and Fairness Software Packages

Python

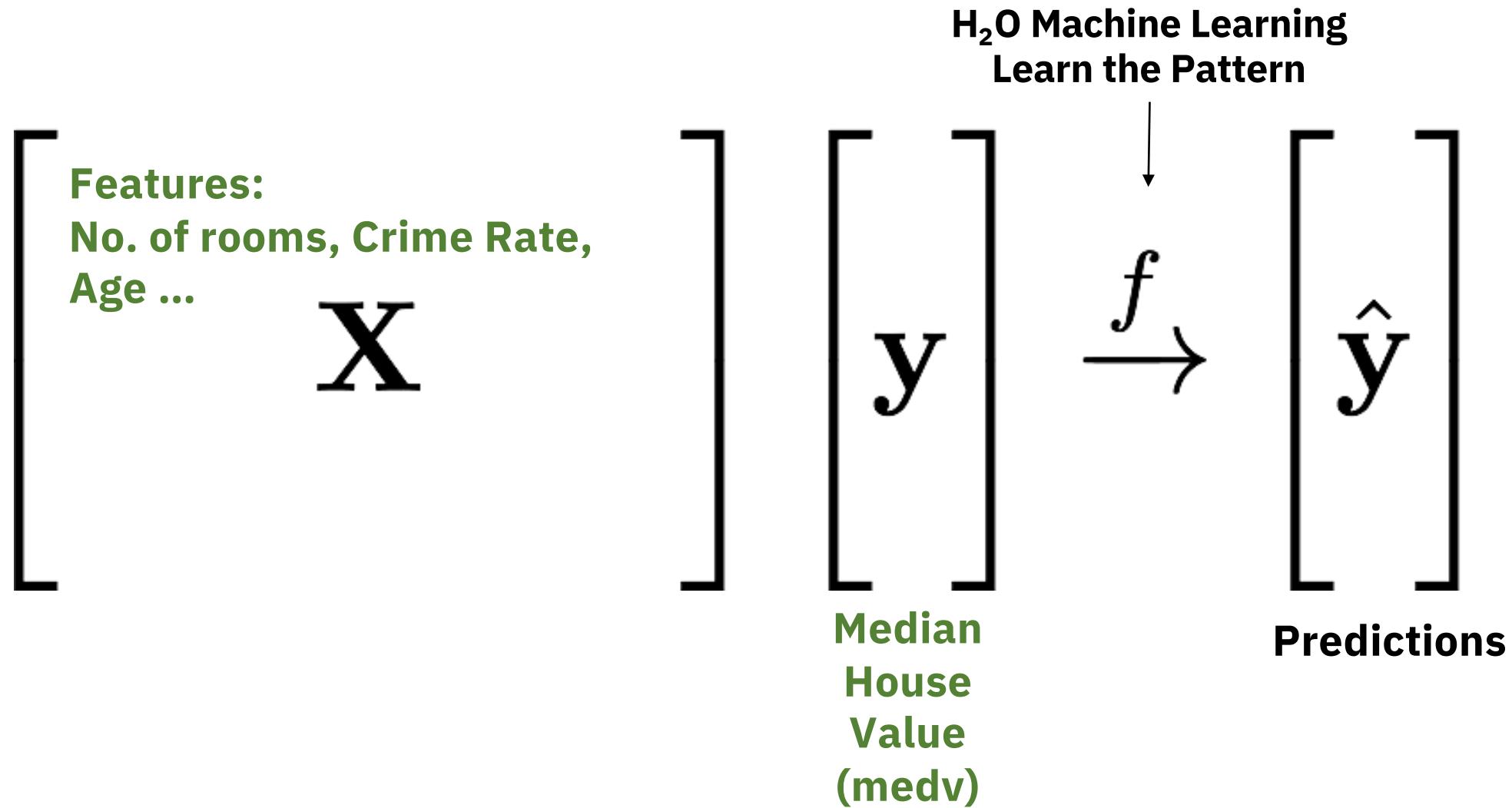
- [aequitas](#)
- [anchor](#)
- [eli5](#)
- [fairml](#)
- [lime](#)
- [PDPbox](#)
- [PyCEbox](#)
- [shap](#)
- [Skater](#)

<https://github.com/jphall663/awesome-machine-learning-interpretability>



Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H_2O , AutoML, LIME)
8:00 – 8:30 pm	Regression Example <code>\examples\regression_...Rmd</code>
8:30 – 9:00 pm	
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H_2O News + Q & A

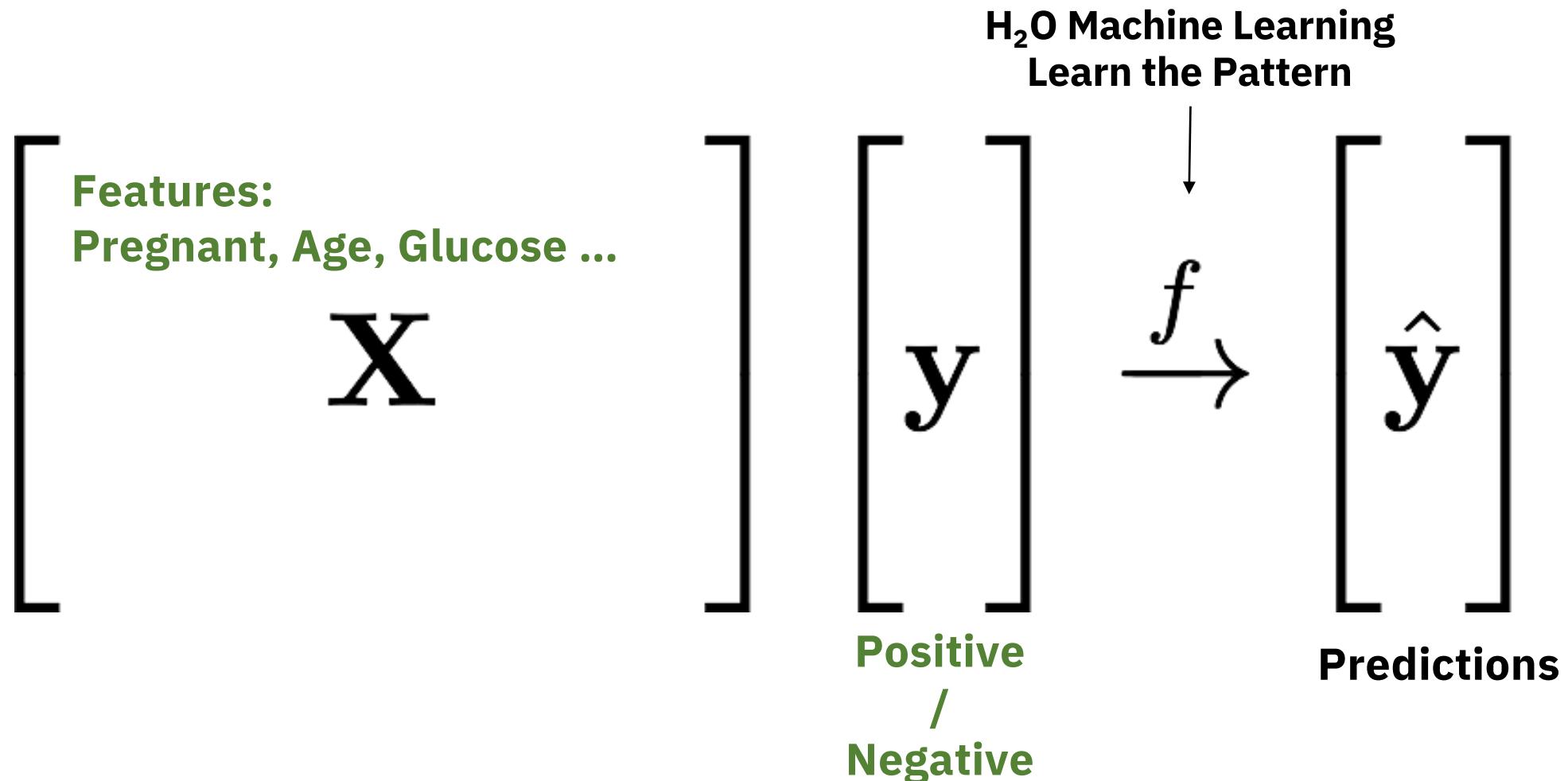
Learning from **Boston Housing** Data





Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H_2O , AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	
9:00 – 9:20 pm	Classification Example <code>\examples\classification_...Rmd</code>
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H_2O News + Q & A

Learning from **Diabetes** Data





@h2oai @matlabulous

#AutoML #LIME

Pizza Time 8:30 – 9:00 pm

Late to the party? Download → bit.ly/joe_eRum_2018



Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H_2O , AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H_2O News + Q & A



Why?

- Most users/organizations can benefit from automatic machine learning pipelines.
 - Eliminate time wasted on human errors, debugging etc.
- Model interpretations is crucial for those who must explain their models to regulators or customers.

You will learn ...

- How to build high quality H₂O models (almost) automatically.
- How to explain predictions from complex H₂O models with LIME.
- **Bonus:** A real use-case that led to multimillion-dollar baseball decisions earlier this year.



Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H ₂ O, AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	A row of five small, stylized pizza icons.
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H ₂ O News + Q & A



Jo-fai Chow

Data Science Evangelist & Community Manager at H2O.ai

2h

...

One of my best data science projects so far - a real **#Moneyball** with **Ari Kaplan** and **David Kearns** that led to a multi-year **#MLB** contract. Check out this **IBM + Aginity + H2O.ai** collaboration <https://lnkd.in/epJVD9t>

Making Multimillion-Dollar ⚾ Decisions with H₂O AutoML, LIME and Shiny



Jo-fai (Joe) Chow
Data Science Evangelist /
Community Manager
joe@h2o.ai
[@matlabulous](https://twitter.com/matlabulous)

Download → https://bit.ly/h2o_meetups

[SlideShare Link](#)



Ari Kaplan

Principal at Aginity

2d

Artificial intelligence in MLB scouting: thanks to **#ibmanalytics** for the interview with me and David Kearns:



Making Data Simple: Hit a home run using AI & machine learning
ibmbigdatahub.com

[Podcast Link](#)



Time	Topics / Tasks
7:00 – 7:30 pm	Welcome + Data Hack Italia
7:30 – 7:45 pm	Install h2o , lime , mlbench from CRAN slides/code: bit.ly/joe_eRum_2018
7:45 – 8:00 pm	Introduction (H_2O , AutoML, LIME)
8:00 – 8:30 pm	Regression Example
8:30 – 9:00 pm	A row of five small, stylized pizza icons.
9:00 – 9:20 pm	Classification Example
9:20 – 9:30 pm	Quick Recap
9:30 – 9:45 pm	Real Use-Case: Moneyball
9:45 – 10:00 pm	Other H_2O News + Q & A

H₂O Products



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



H2O AI Open Source Engine
Integration with Spark



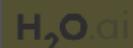
Lightning Fast machine
learning on GPUs

DRIVERLESSAI

Automatic feature
engineering, machine
learning and interpretability

Steam

Secure multi-tenant H2O clusters

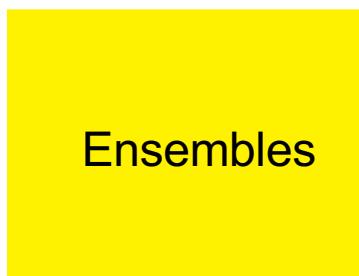


Algorithms on H₂O-3 (CPU)

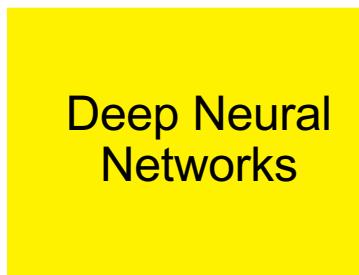
Supervised Learning



- 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

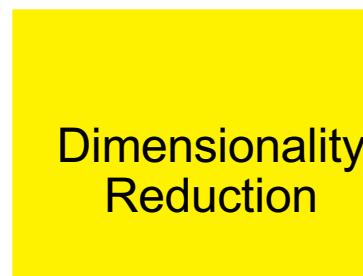


- Deep learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

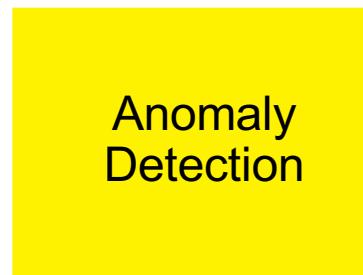
Unsupervised Learning



- 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



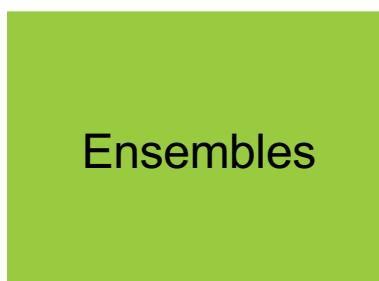
- Autoencoders: Find outliers using a nonlinear dimensionality reduction using deep learning

Algorithms on H₂O4GPU (more to come)

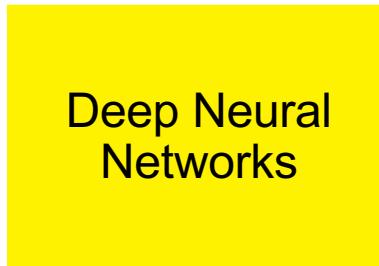
Supervised Learning



- 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

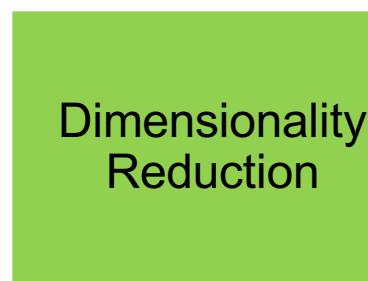


- Deep learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

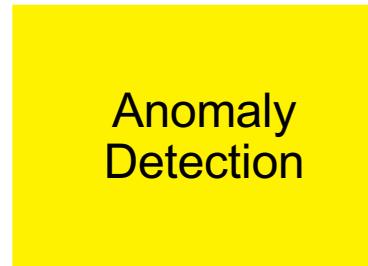
Unsupervised Learning



- 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

Coming Up

Date	City	Talk / Workshop
25 June	Milan	H ₂ O + LIME Workshop
28 June	Rome	H ₂ O Intro + Use Cases
Mid-July	London	Next London Meetup (T.B.C.)
Mid-Oct	London	H ₂ O World London
13-14 Nov	London	Big Data LDN Keynote by Sri (CEO) + Meetup



Thanks!

- Organizers & Sponsors

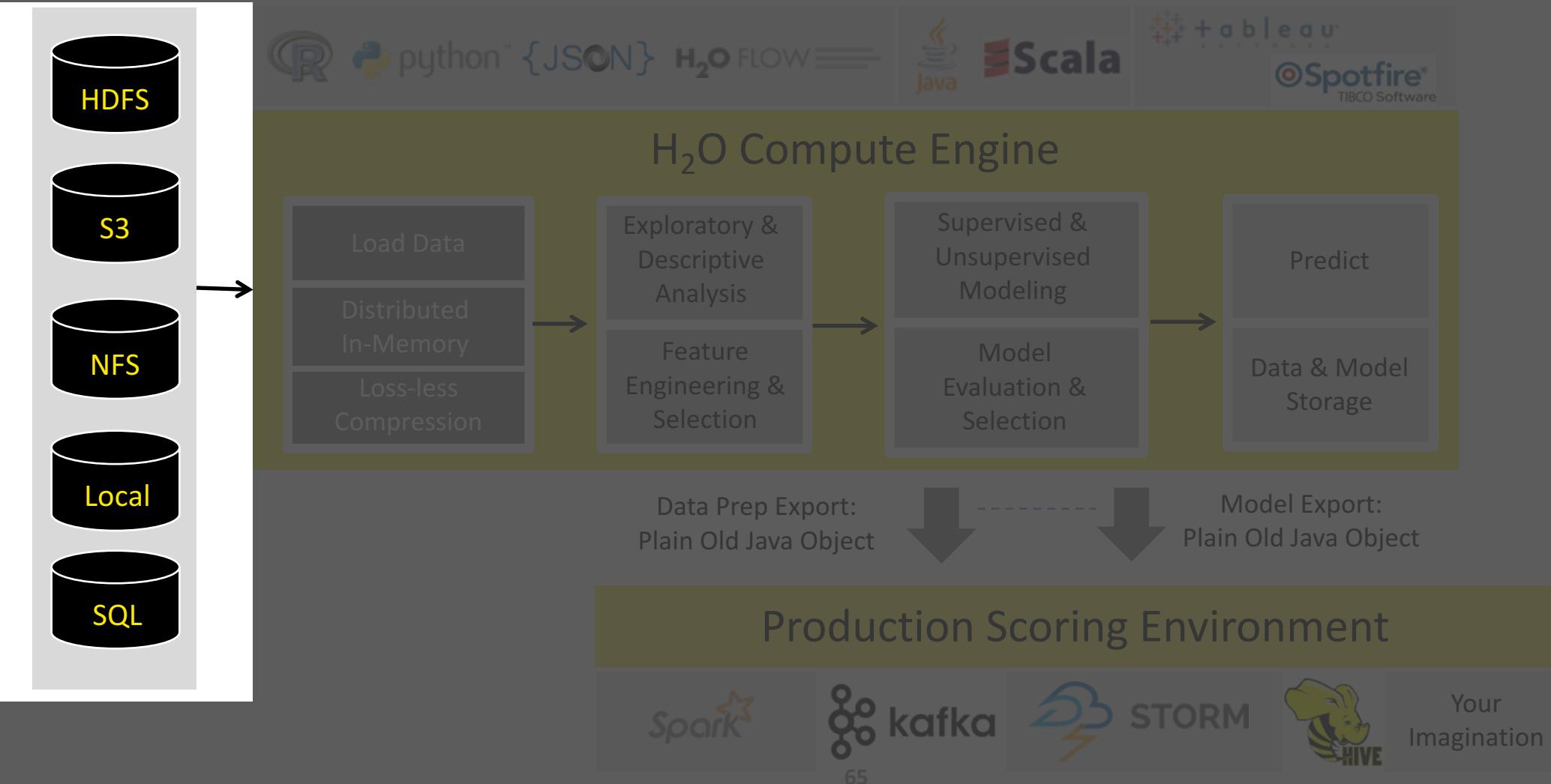


- Code, Slides & Documents
 - bit.ly/h2o_meetups
 - bit.ly/joe_eRum_2018
 - docs.h2o.ai
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe
- Please search/ask questions on **Stack Overflow**
 - Use the tag `h2o` (not h2 zero)

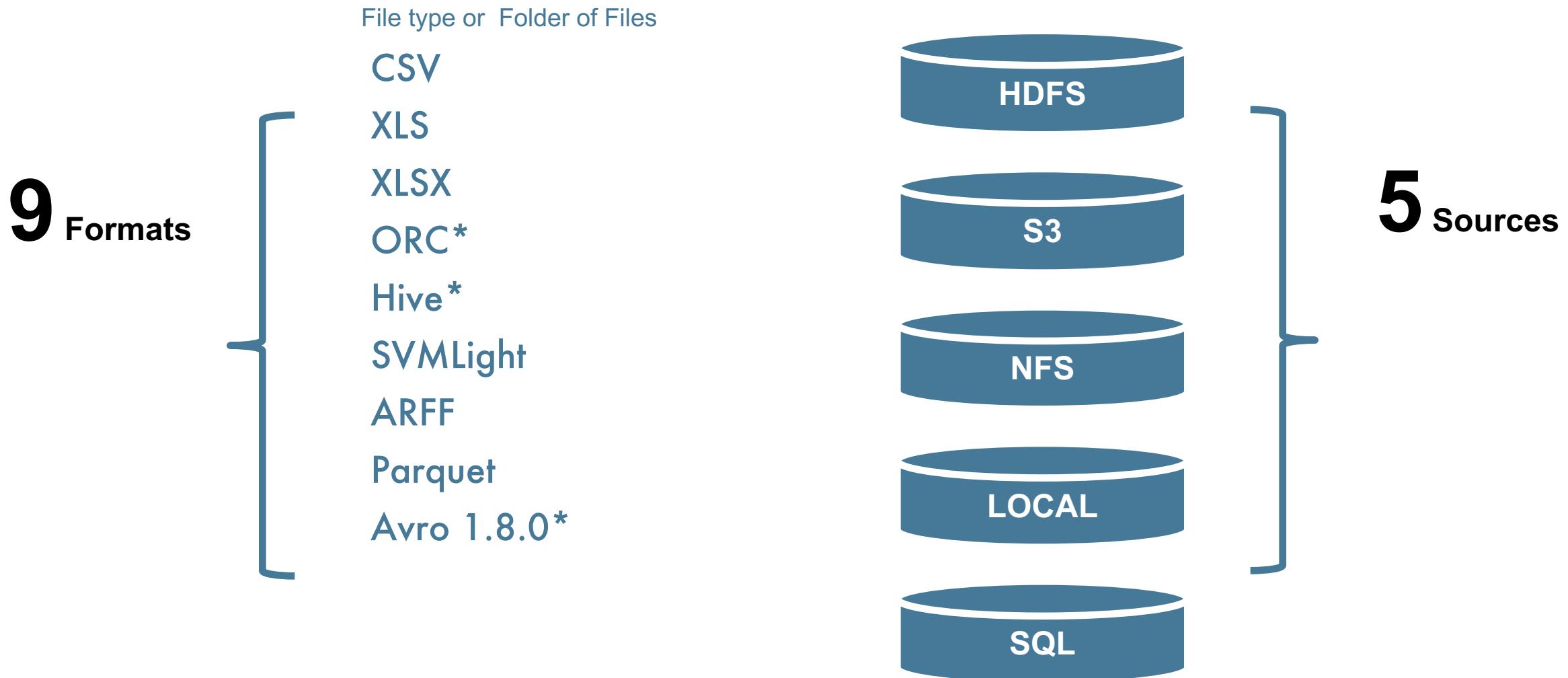
Appendix

High Level Architecture

Import Data from
Multiple Sources



Supported Formats & Data Sources



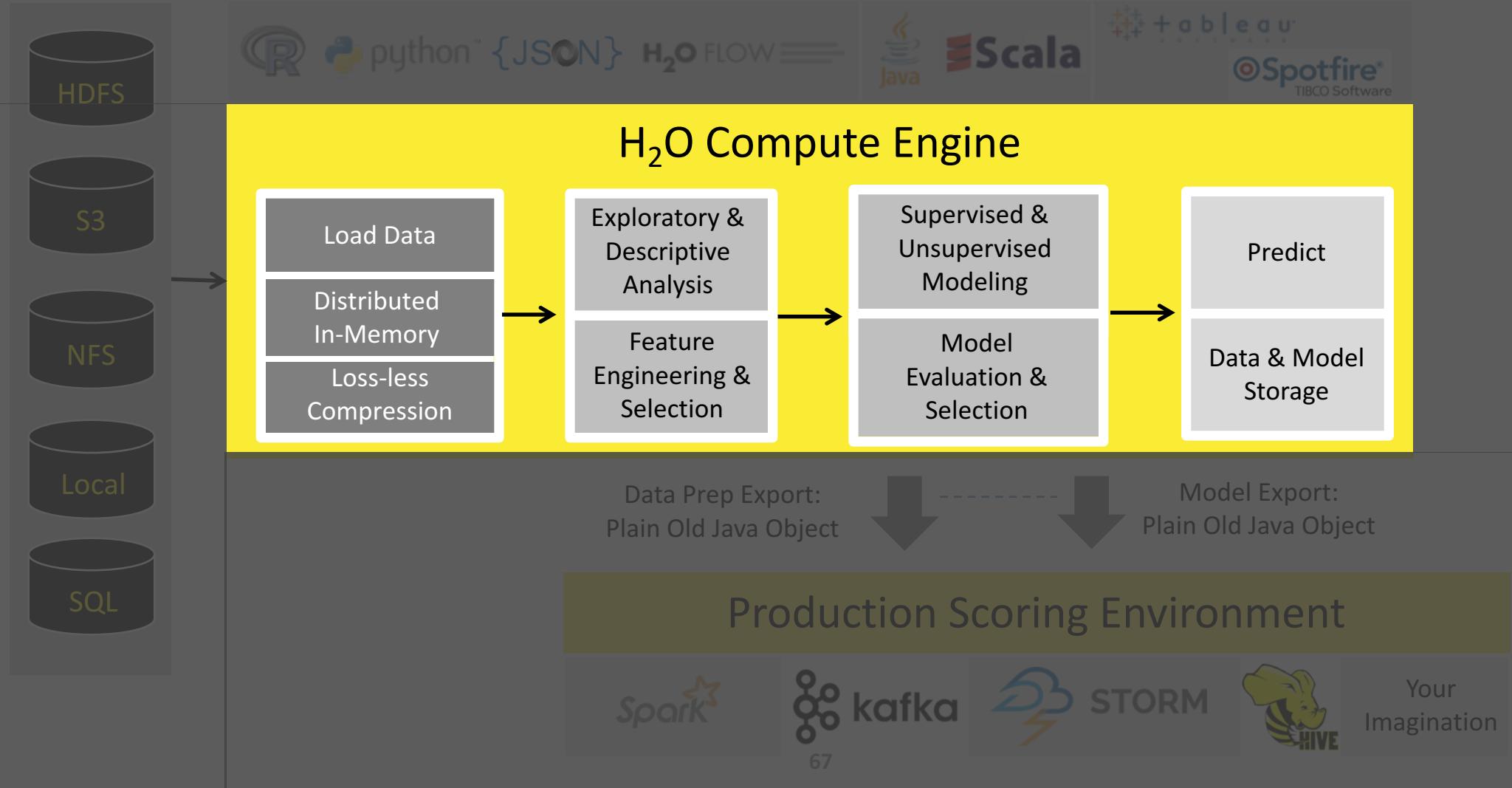
* 1. only if H2O is running as a Hadoop job

* 2. Hive files that are saved in ORC format

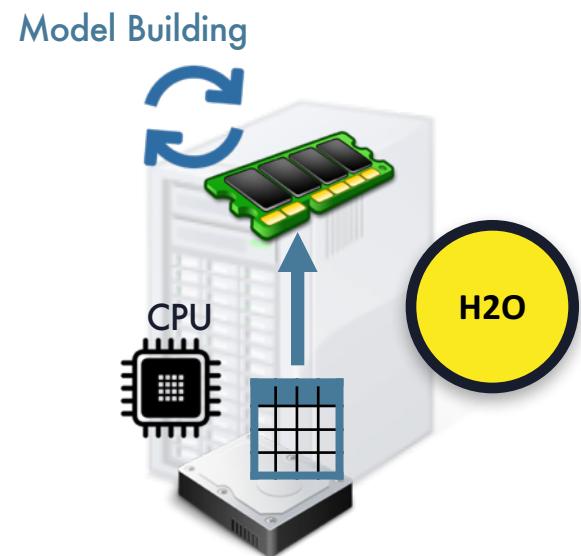
* 3. without multi-file parsing or column type modification

High Level Architecture

Fast, Scalable & Distributed
Compute Engine Written in
Java



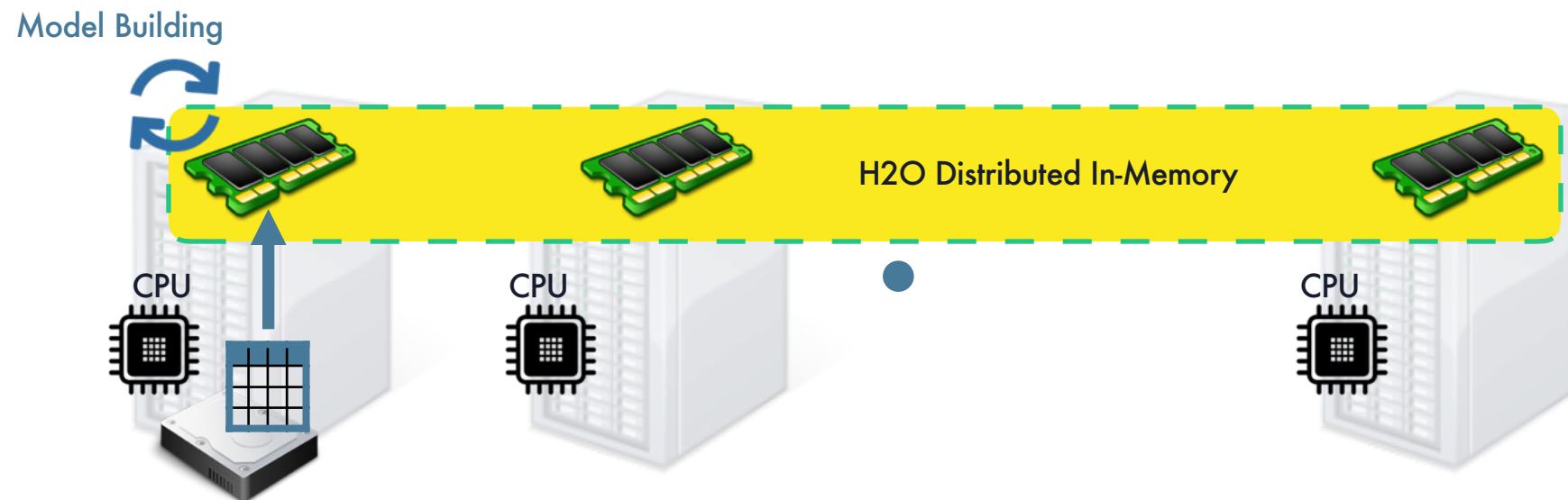
H₂O Core



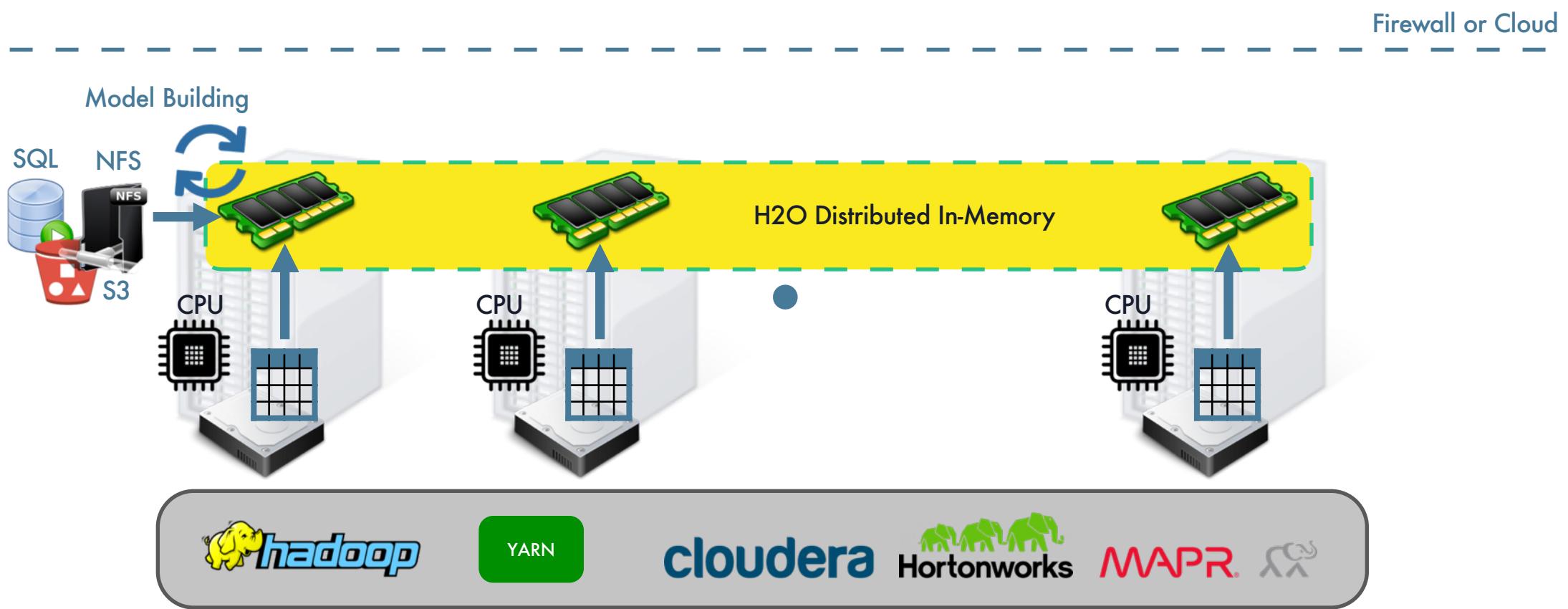
H₂O Core



H₂O Core

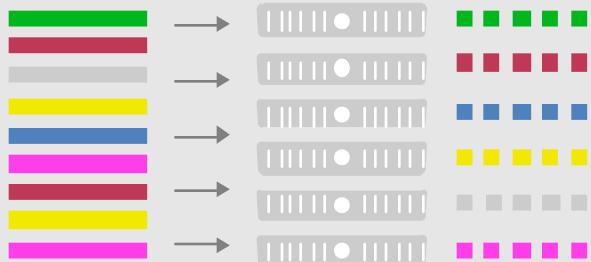


H₂O Core



Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable
Distributed Histogram Calculation for GBM

Advantageous Foundation

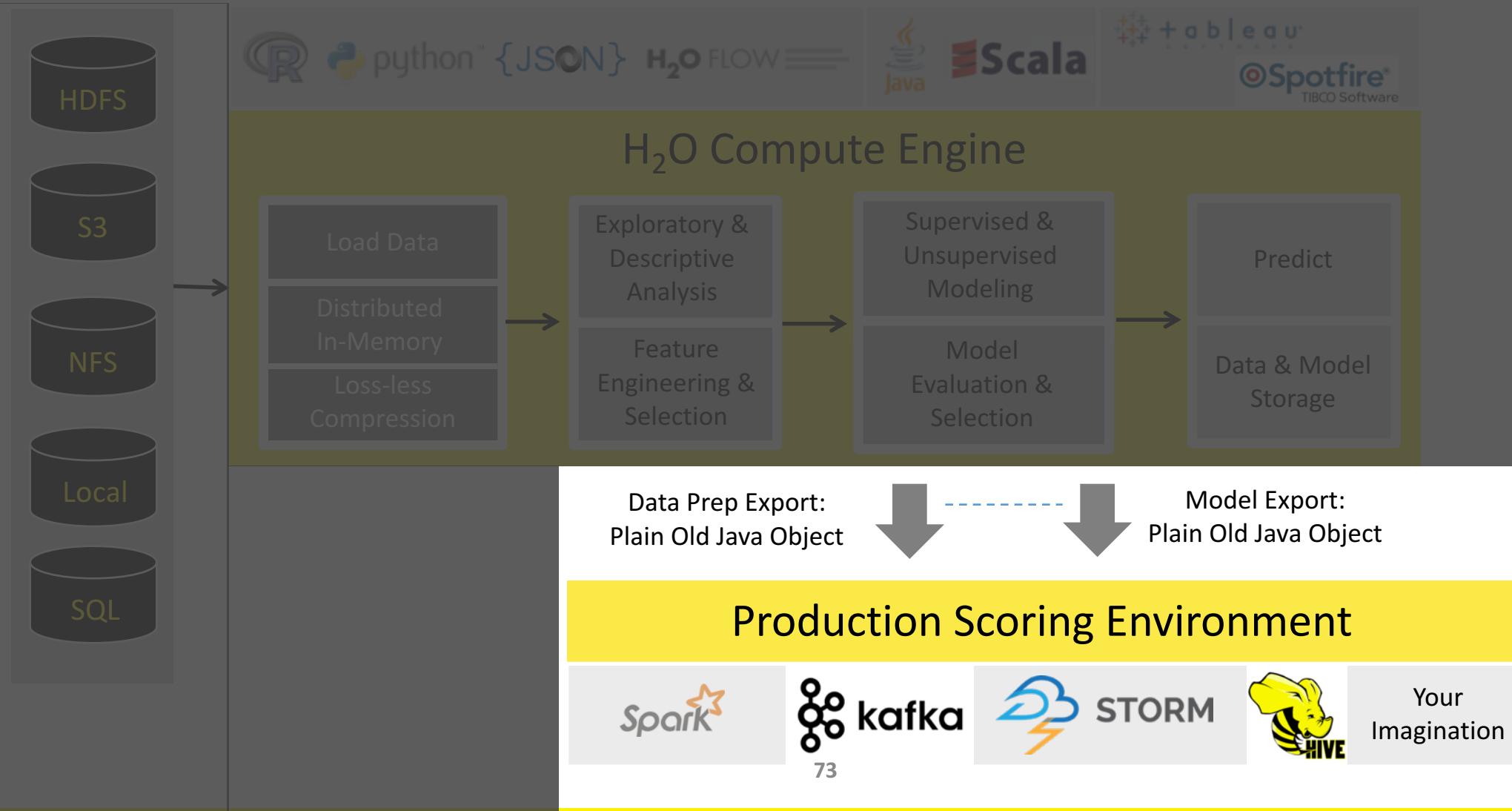
- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H₂O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- “Out-of-box” functionalities for all algorithms (**NO MORE SCRIPTING**) and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

High Level Architecture

Export Standalone Models
for Production



H₂O Documentation

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

Getting Started & User Guides

 Open Source |  Commercial

H₂O

What is H₂O?
H₂O User Guide (Main docs)
H₂O Book (O'Reilly)
Recent Changes
Open Source License (Apache V2)

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

[Download H₂O](#)

Sparkling Water

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

Quick Start Video - Scala

[Download Sparkling Water](#)

Driverless AI

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

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

[Download Driverless AI](#)

H₂O4GPU (alpha)

H₂O4GPU Readme
Open Source License (Apache V2)

[Download H₂O4GPU](#)

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

Demo: H₂O on a 320-Core Hadoop Cluster

(Web Interface)



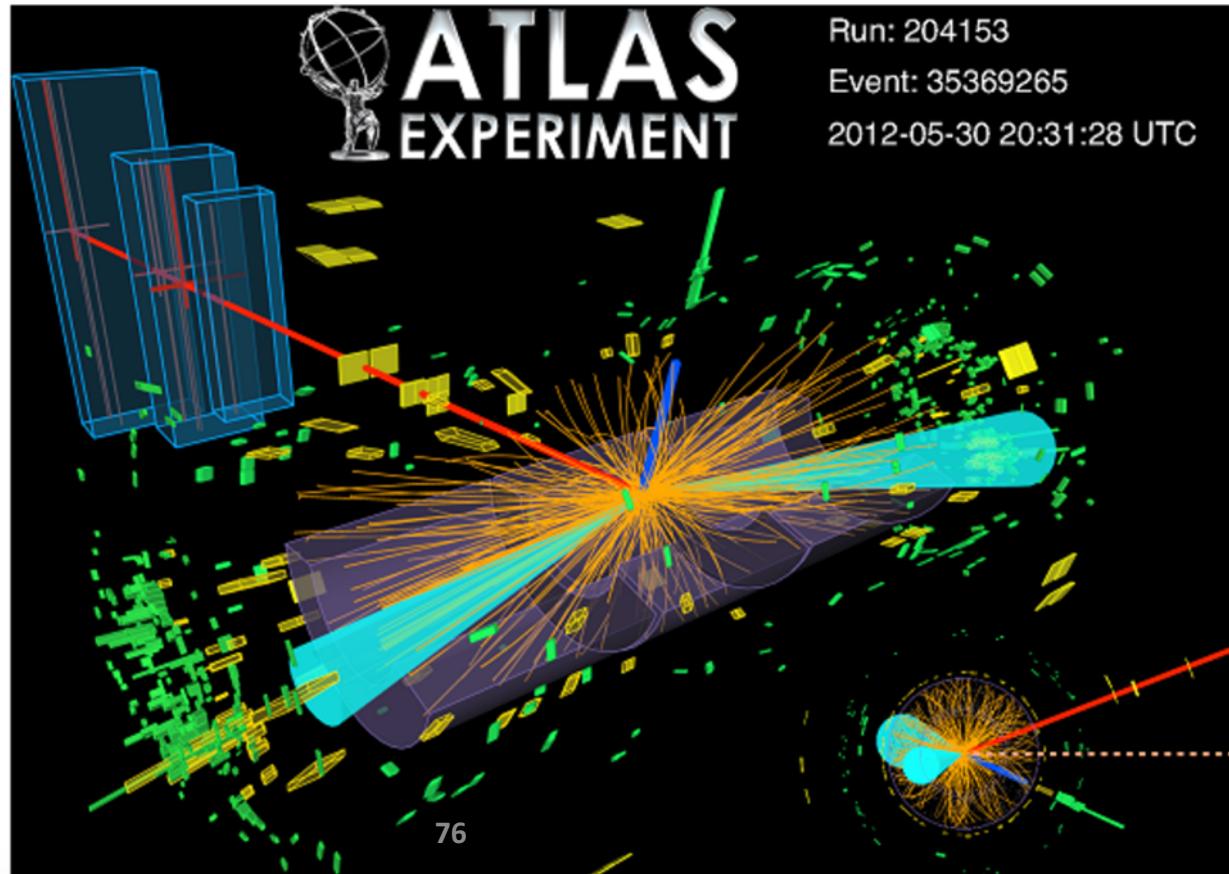
Higgs Boson Machine Learning Challenge

Use the ATLAS experiment to identify the Higgs boson

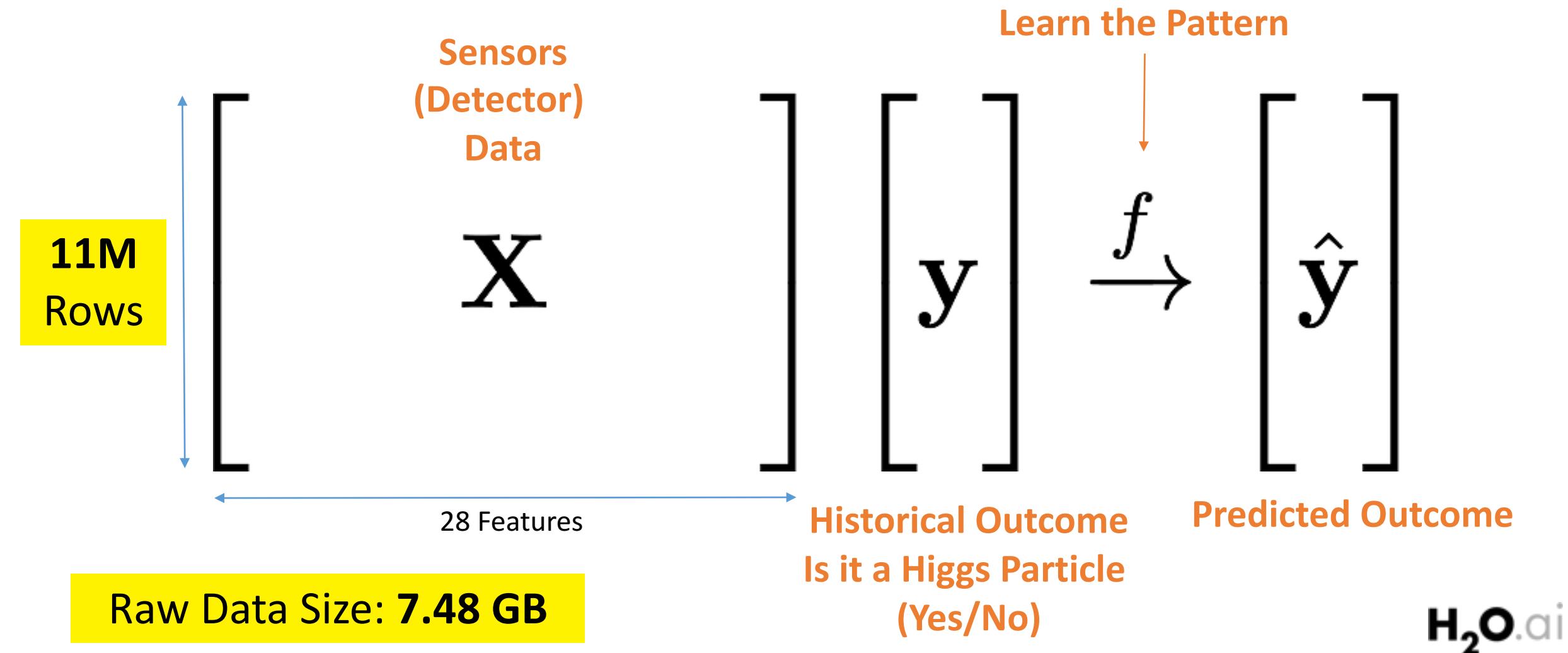
\$13,000 · 1,785 teams · 3 years ago

[Overview](#)[Data](#)[Discussion](#)[Leaderboard](#)[Rules](#)[Team](#)[My Submissions](#)[Late Submission](#)[Overview](#)

<https://www.kaggle.com/c/higgs-boson>

[Description](#)[Evaluation](#)[Prizes](#)[About The Sponsors](#)[Timeline](#)[Winners](#)

Learning from Higgs Boson Machine Data



11M Rows**Size (Raw): 7.48 GB****Compressed: 2.00 GB (\approx 27% of Raw)**

HIGGS.hex

Actions:

View Data

Split...

Build Model...

Predict

Download

Export

Rows	Columns	Compressed Size
11000000	29	2GB

▼ COLUMN SUMMARIES

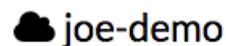
label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
C1	enum	0	5170877	0	0	0	1.0	0.5299	0.4991	2	Convert to numeric
C2	real	0	0	0	0	0.2747	12.0989	0.9915	0.5654
C3	real	0	0	0	0	-2.4350	2.4349	-0.0	1.0088
C4	real	0	0	0	0	-1.7425	1.7432	-0.0	1.0063
C5	real	0	0	0	0	0.0002	15.3968	0.9985	0.6000
C6	real	0	0	0	0	-1.7439	1.7433	0.0	1.0063
C7	real	0	0	0	0	0.1375	9.9404	0.9909	0.4750
C8	real	0	0	0	0	-2.9697	2.9697	-0.0	1.0093
C9	real	0	0	0	0	-1.7412	1.7415	0.0	1.0059
C10	real	0	5394611	0	0	0	2.1731	1.0	1.0278
C11	real	0	0	0	0	0.1890	11.6471	0.9927	0.5000
C12	real	0	0	0	0	-2.9131	2.9132	-0.0	1.0093
C13	real	0	0	0	0	-1.7424	1.7432	-0.0	1.0062
C14	real	0	5523912	0	0	0	2.2149	1.0	1.0494
C15	real	0	0	0	0	0.2636	14.7090	0.9923	0.4877
C16	real	0	0	0	0	-2.7297	2.7300	0.0	1.0087
C17	real	0	0	0	0	-1.7421	1.7429	0.0	1.0063
C18	real	0	6265240	0	0	0	2.5482	1.0	1.1937
C19	real	0	0	0	0	0.3654	12.8826	0.9861	0.5058
C20	real	0	0	0	0	-2.4973	2.4980	-0.0	1.0077

Untitled Flow



CS

getCloud



CLOUD STATUS

HEALTHY	CONSENSUS	LOCKED
Version	Started	Nodes (Used / All)
3.13.0.3981	a minute ago	10 / 10

NODES

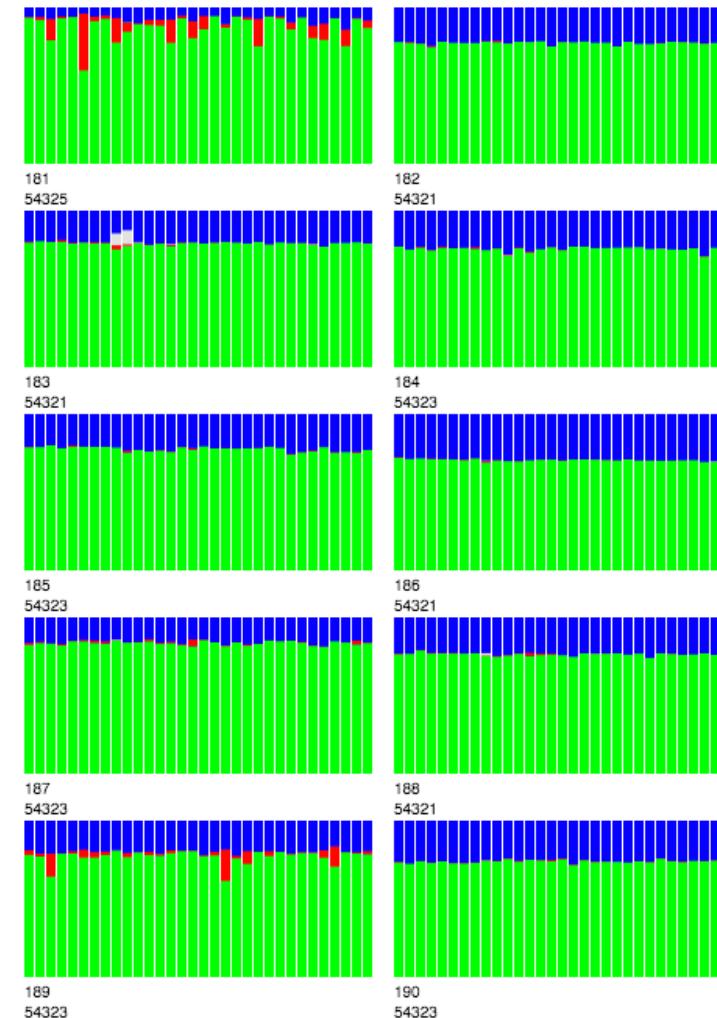
Name	Ping	Cores	Load	My CPU %	Sys	Shut Down	Data (Used/Total)	Data (% Cached)	GC (Free / Total / Max)	Disk (Free / Max)	Disk (% Free)
✓ 172.16.2.181:54323	a few seconds ago	32	6.110	0	8	-	40.603	33.82 GB / s	29.46 GB / NaN undefined / 29.58 GB	339.08 GB / 1.70 TB	19%
✓ 172.16.2.182:54321	a few seconds ago	32	0.240	7	8	-	44.566	39.59 GB / s	29.43 GB / NaN undefined / 29.58 GB	225.64 GB / 1.70 TB	12%
✓ 172.16.2.183:54321	a few seconds ago	32	9.820	0	3	-	44.883	42.09 GB / s	29.34 GB / NaN undefined / 29.58 GB	450.18 GB / 1.70 TB	25%
✓ 172.16.2.184:54323	a few seconds ago	32	0.990	0	0	-	44.656	41.67 GB / s	29.51 GB / NaN undefined / 29.58 GB	254.96 GB / 1.70 TB	14%
✓ 172.16.2.185:54323	a few seconds ago	32	0.440	8	8	-	43.128	38.33 GB / s	29.43 GB / NaN undefined / 29.58 GB	501.02 GB / 1.70 TB	28%
✓ 172.16.2.186:54321	a few seconds ago	32	1.750	0	0	-	44.589	42.46 GB / s	29.42 GB / NaN undefined / 29.58 GB	331.27 GB / 1.70 TB	18%
✓ 172.16.2.187:54323	a few seconds ago	32	1.490	0	10	-	43.993	42.00 GB / s	29.46 GB / NaN undefined / 29.58 GB	367.40 GB / 1.70 TB	21%
✓ 172.16.2.188:54321	a few seconds ago	32	0.610	0	8	-	41.977	18.63 GB / s	28.30 GB / NaN undefined / 29.58 GB	218.27 GB / 1.70 TB	12%
✓ 172.16.2.189:54323	a few seconds ago	32	4.420	6	9	-	48.590	38.91 GB / s	29.34 GB / NaN undefined / 29.58 GB	477.97 GB / 1.70 TB	27%
✓ 172.16.2.190:54323	a few seconds ago	32	2.970	10	12	-	43.931	22.15 GB / s	29.51 GB / NaN undefined / 29.58 GB	274.50 GB / 1.70 TB	15%
✓ TOTAL	-	320	28.840	-	-	-	440.916	359.62 GB / s	293.18 GB / NaN undefined / 295.83 GB	3.36 TB / 17.04 TB	19%

$$10 \times 32 = \\ 320 \text{ Cores}$$

$$10 \times 29.6 = 296 \\ \text{GB Memory}$$

H₂O Water Meter (CPU Monitor)

10 x 32 = 320 Cores



Legend

Each bar represents one CPU.

Blue: idle time

Green: user time

Red: system time

White: other time (e.g. i/o)