Automatic and Interpretable Machine Learning in R with H_2O and LIME

Jo-fai (Joe) Chow

Data Science Evangelist / Community Manager

joe@h2o.ai

@matlabulous

Download → https://bit.ly/

joe_eRum_2018

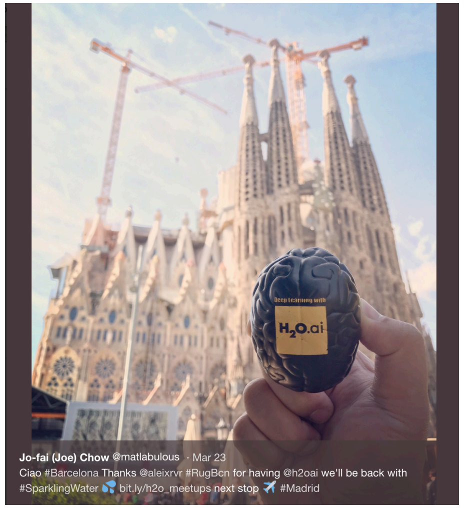
Why?

- 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 H_2O models (almost) automatically.
- How to explain predictions from complex H_2O models with LIME.
- **Bonus**: A real use-case that led to multimillion-dollar baseball decisions earlier this year.

About Me



- **Before H_2O**
  - Water Engineer / Researcher / Matlab Fan Boy @matlabulous
  - Discovered R, Python, H_2O … never look back again
  - Data Scientist at Virgin Media (UK), Domino Data Lab (US)

- **At H_2O …**
  - Data Scientist / Evangelist /
  - Sales Engineer / Solution Architect /
  - Community Manager
    (The harsh reality of startup life)
  - H_2O SWAG Photographer
    (#AroundTheWorldWithH2Oai)
  - H_2O SWAG Distributor
    (Love H_2O? Come get some stickers!)

H2O.ai

What I really do …

Jo-fai (Joe) Chow
@matlabulous

Thanks @ingnl for hosting @h2oai #meetup in #Amsterdam last week. Tremendous turnout and great discussions. #AroundTheWorldWithH2Oai #360Selfie 🇳🇱 cc @fishnets88

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

Jo-fai (Joe) Chow
@matlabulous

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

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

Jo-fai (Joe) Chow
@matlabulous

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

1:54 PM - 7 Mar 2018 from Hotel Berlin

Jo-fai (Joe) Chow
@matlabulous

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

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

Reminder: #360Selfie

Agenda

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 pm	Other H_2O News + Q & A

Time	Topics / Tasks
1:30 – 1:45 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
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 pm	Other H_2O News + Q & A

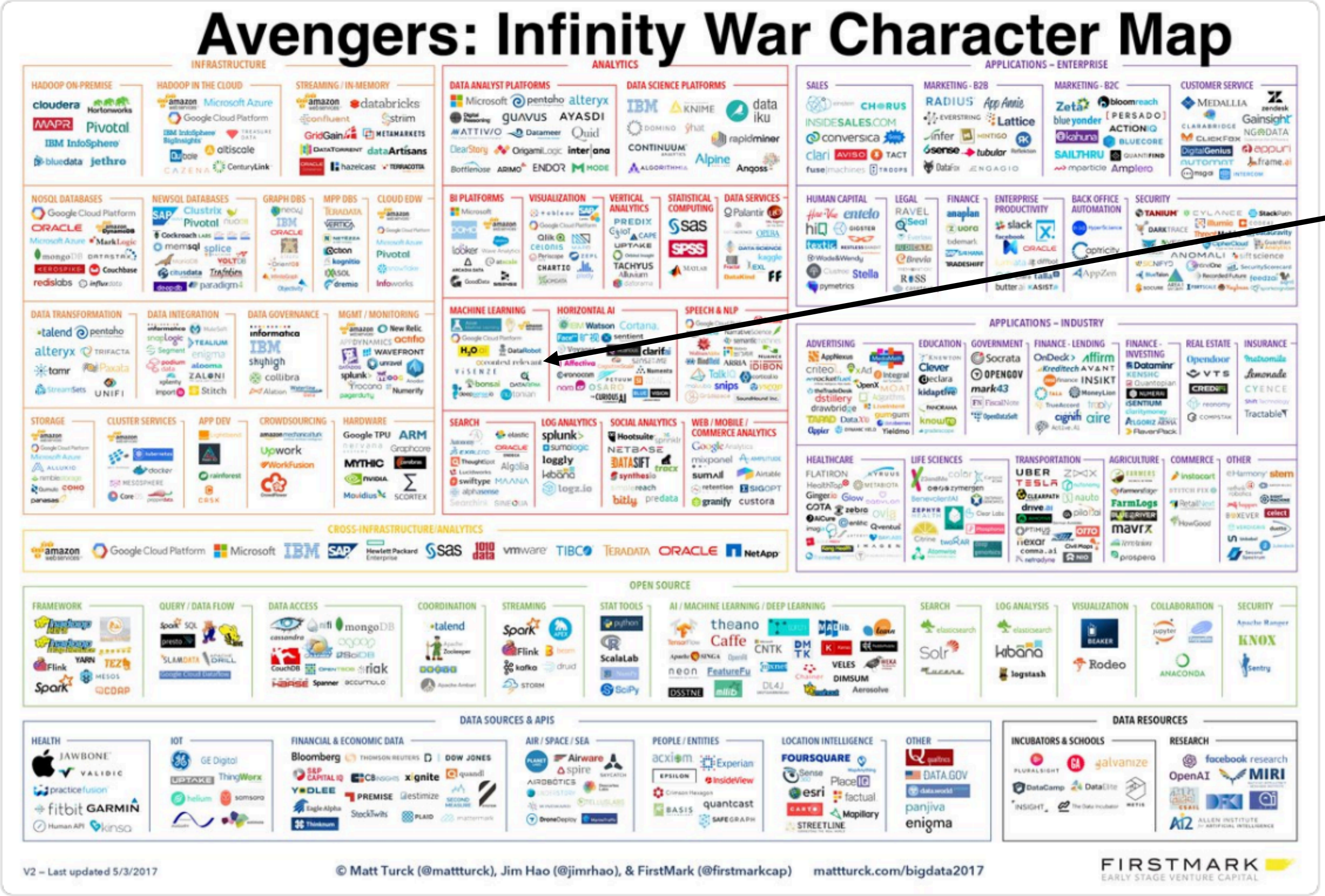
Have you seen
Avengers: Infinity War?

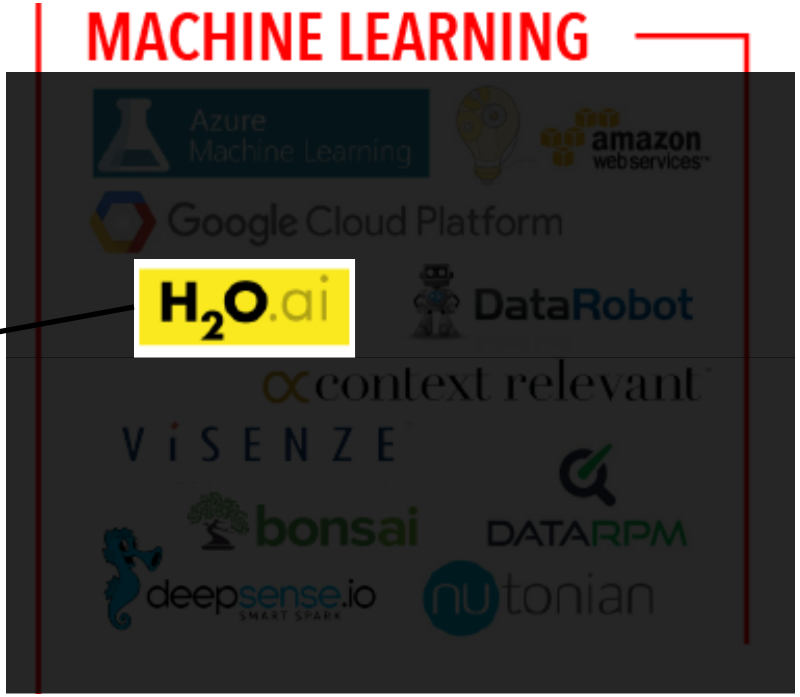
Do you know all the characters in the movie? (No spoilers - I promise)

Vicki Boykis
@vboykis

Follow

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





1:06 AM - 11 May 2018

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)

Platforms with H_2O integration

srisatish
@srisatish

Following

Replying to @BobMuenchen @knime @h2oai

@KNIME gained the ability to run @H2O.ai algorithms, so these two may be viewed as complementary, not competitors #Ecosystem #OpenSource

3:32 PM - 2 Mar 2018

H_2O + 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)

Company Overview

Founded	2012, Series C in Nov, 2017
Products	• Driverless AI – Automated Machine Learning • H_2O Open Source Machine Learning • Sparkling Water
Mission	Democratize AI. Do Good
Team	~100 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Offices	Mountain View, London, Prague

H_2O Products

H2O.ai

In-Memory, Distributed Machine Learning Algorithms with H_2O Flow GUI

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H2O4GPU

Lightning Fast machine learning on GPUs

DRIVERLESS AI

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H_2O clusters

This Workshop

H_2O Products

In-Memory, Distributed Machine Learning Algorithms with H_2O Flow GUI

H2O AI Open Source Engine Integration with Spark

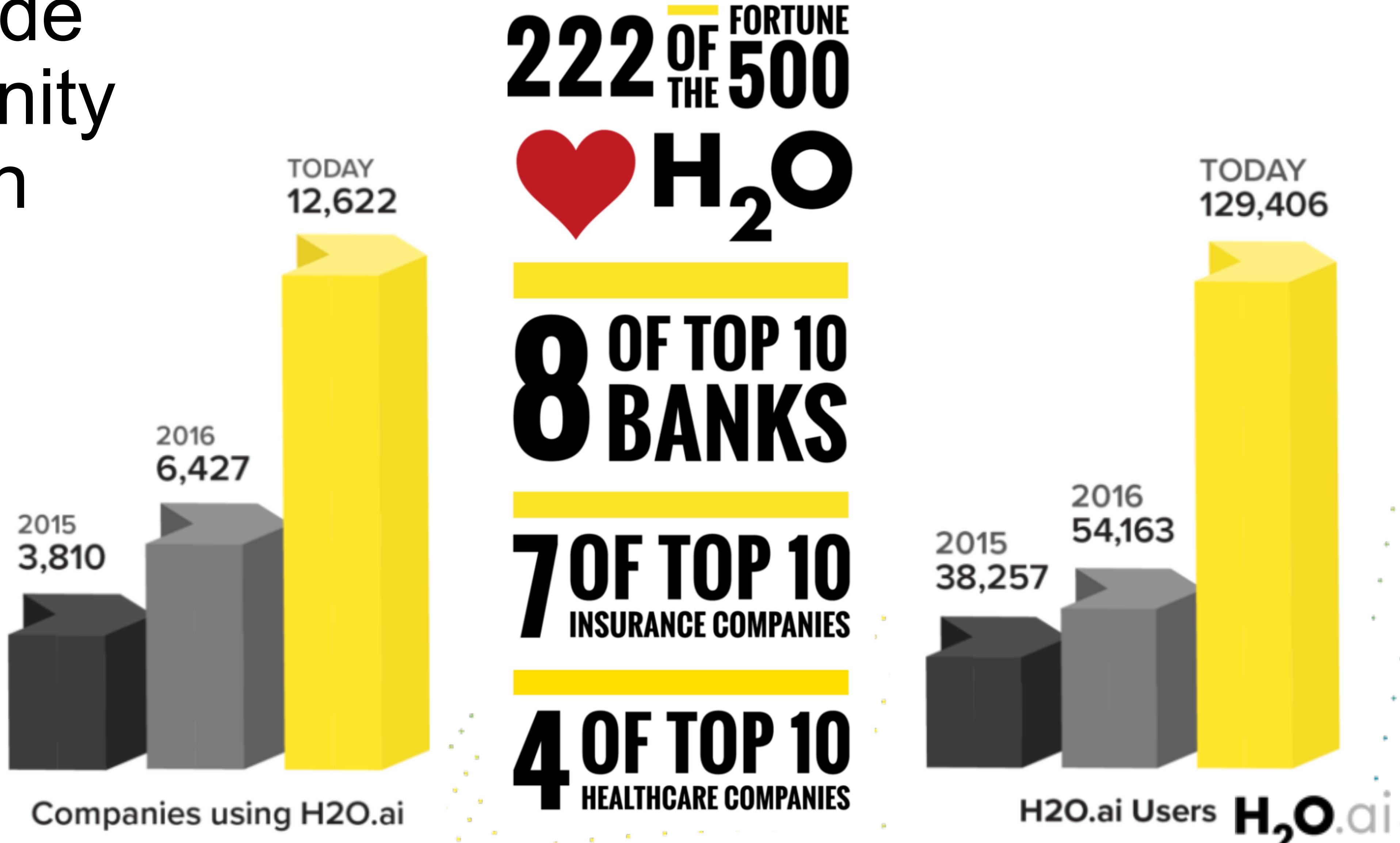
Lightning Fast machine learning on GPUs

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H_2O clusters

Worldwide Community Adoption



* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H2O.ai Solution Leadership Across Verticals

Retail
Travelport
Walgreens
ebay
Booking.com
macy's

Healthcare
CHANGE HEALTHCARE
KAISER PERMANENTE
aetna
CONFIDENTIAL

Marketing
acxiom
CATALINA
Integral Ad Science
MarketShare a Neustar Solution
Nielsen Catalina Solutions

Financial
citi
deserve
DISCOVER
experian
Capital One
EQUIFAX
MarketAxess
WELLS FARGO
RBC
PayPal

Advisory & Accounting
ADP
pwc
starling
underwrite.ai

Insurance
State Farm
AEGON
XCEEDANCE
TRANSAMERICA
PROGRESSIVE
ZURICH

Telecom
COMCAST
CONFIDENTIAL
CISCO

H2O.ai

Why H_2O?

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

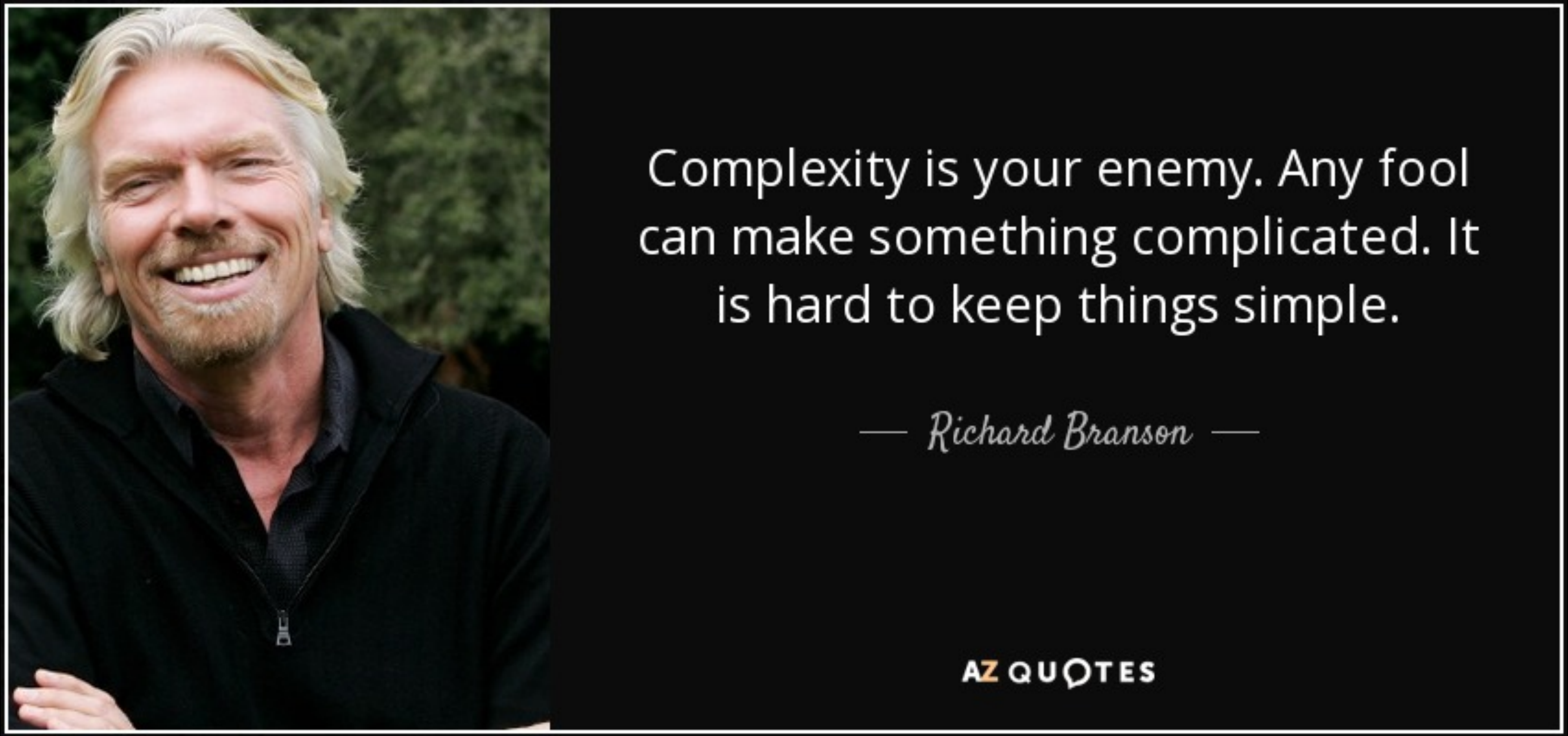
3 9 9

Tweet your reply

Arno Candel @ArnoCandel · May 12
Replying 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

11

Our Mission:
Make Machine Learning Accessible to Everyone



Scientific Advisory Council

Dr. Trevor Hastie

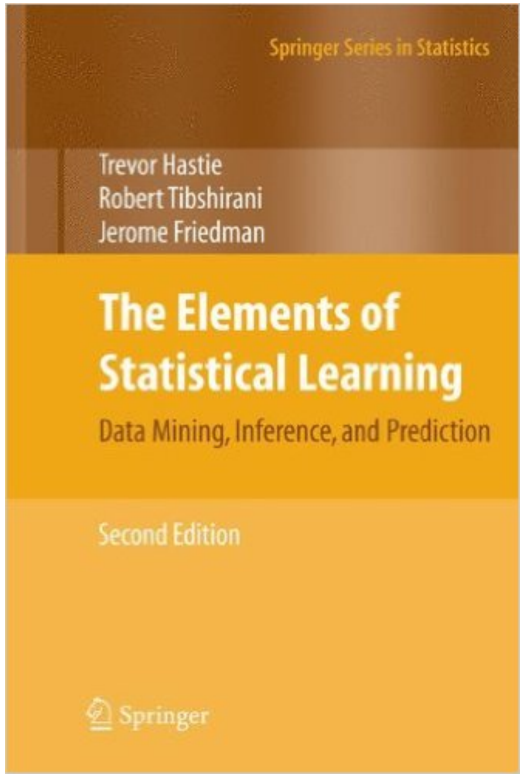
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

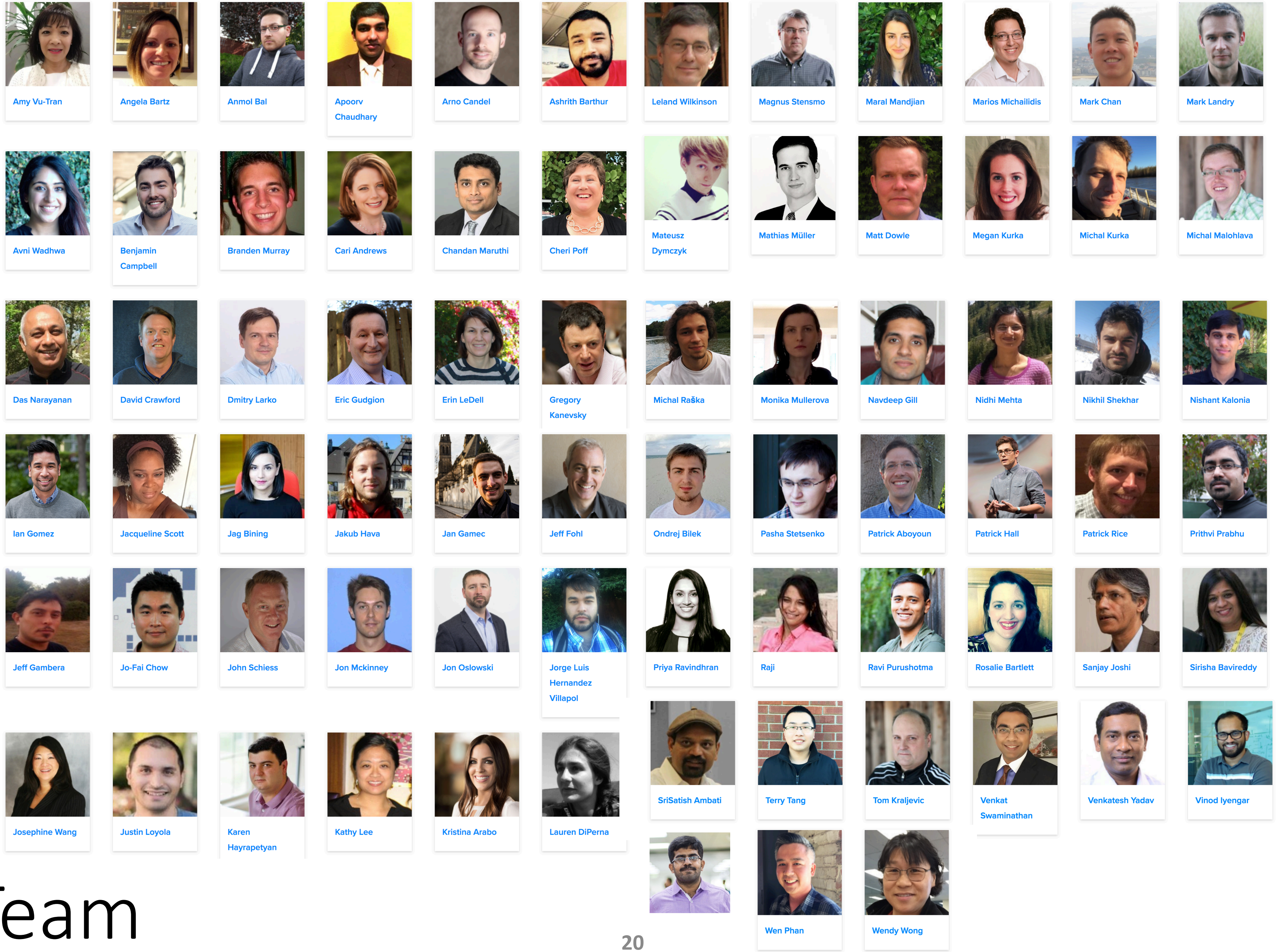
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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



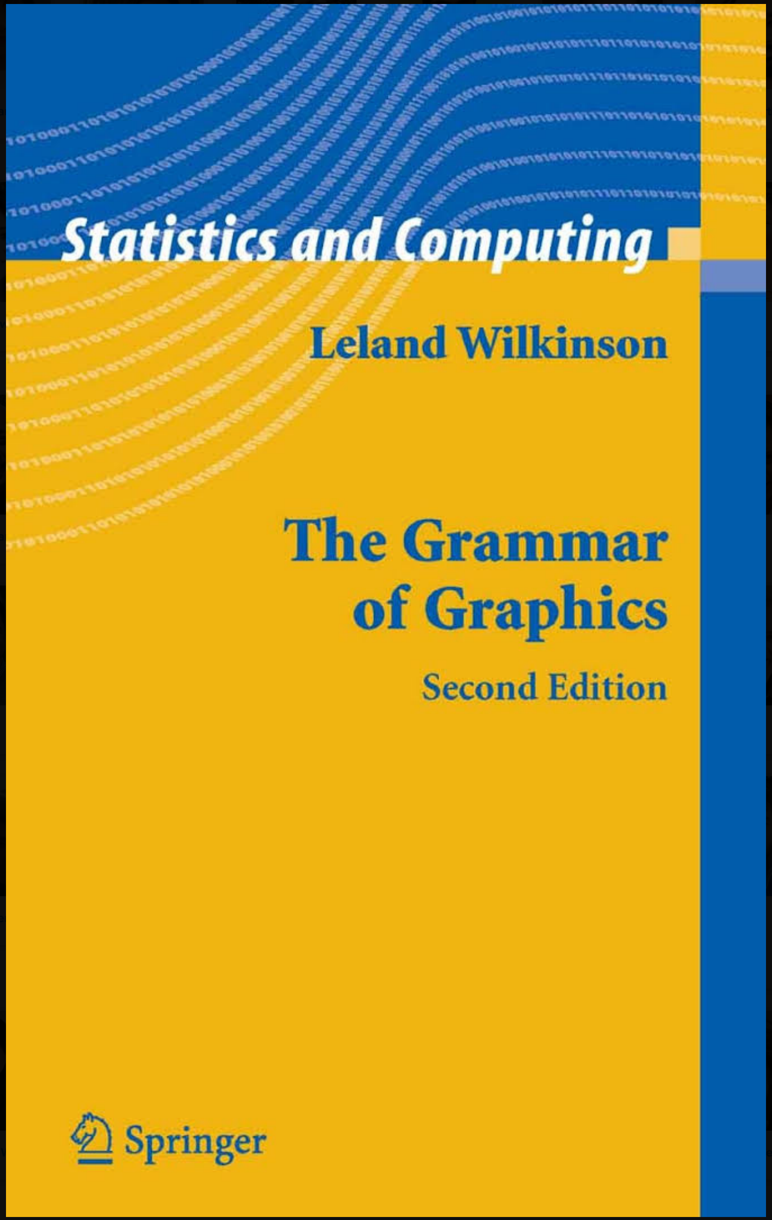
H2O.ai

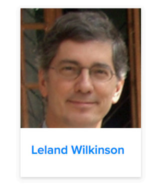


H_2O Team

H_2O.ai

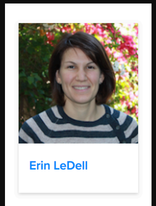
Origin of R Package `ggplot2`





Matt Dowle

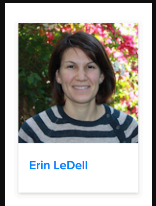
H_2O Team

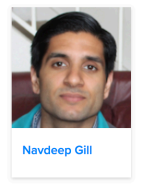


Erin LeDell, Chief ML Scientist
Women in ML/DS & R-Ladies Global

H_2O Team

H_2O AutoML

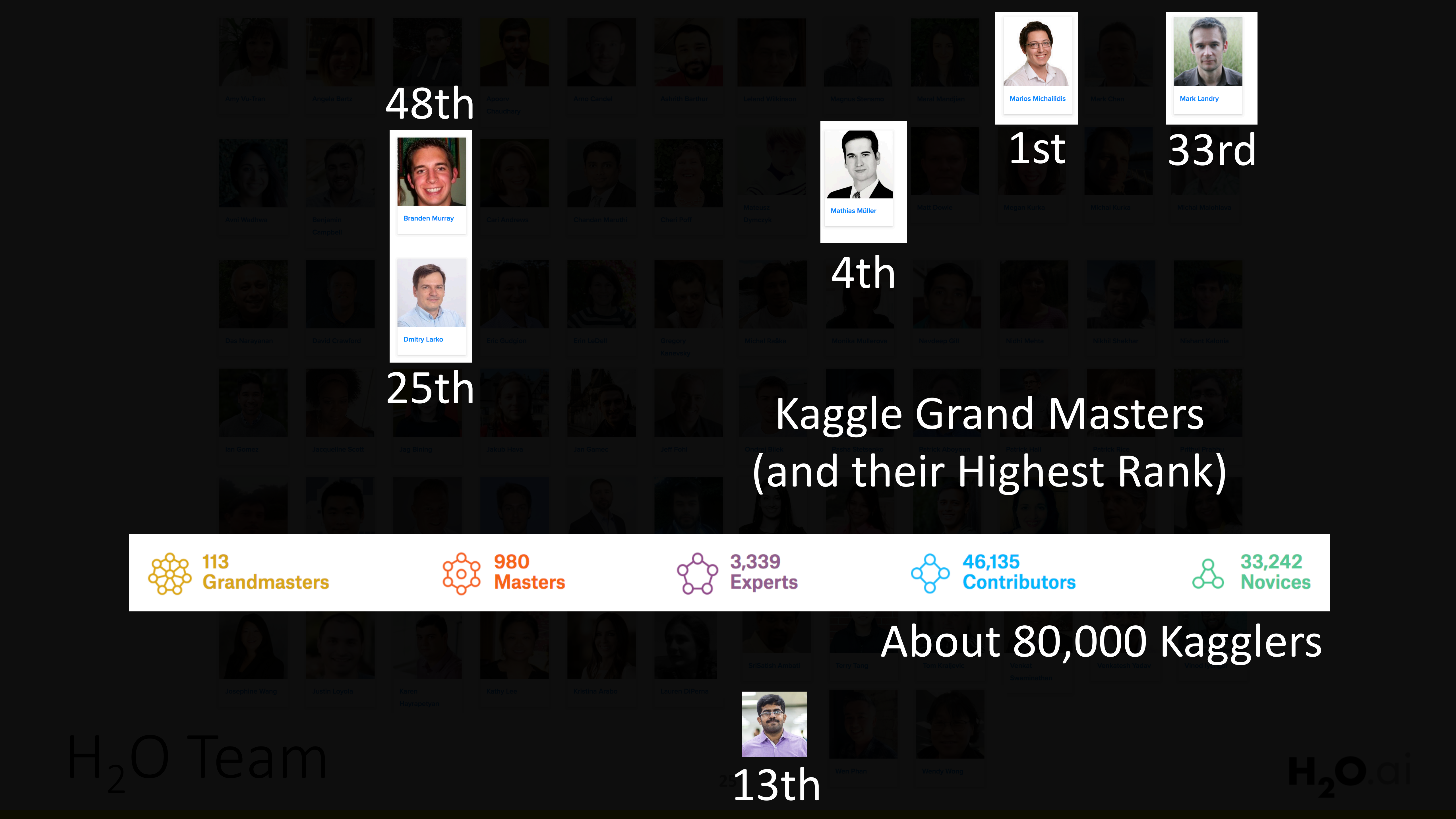




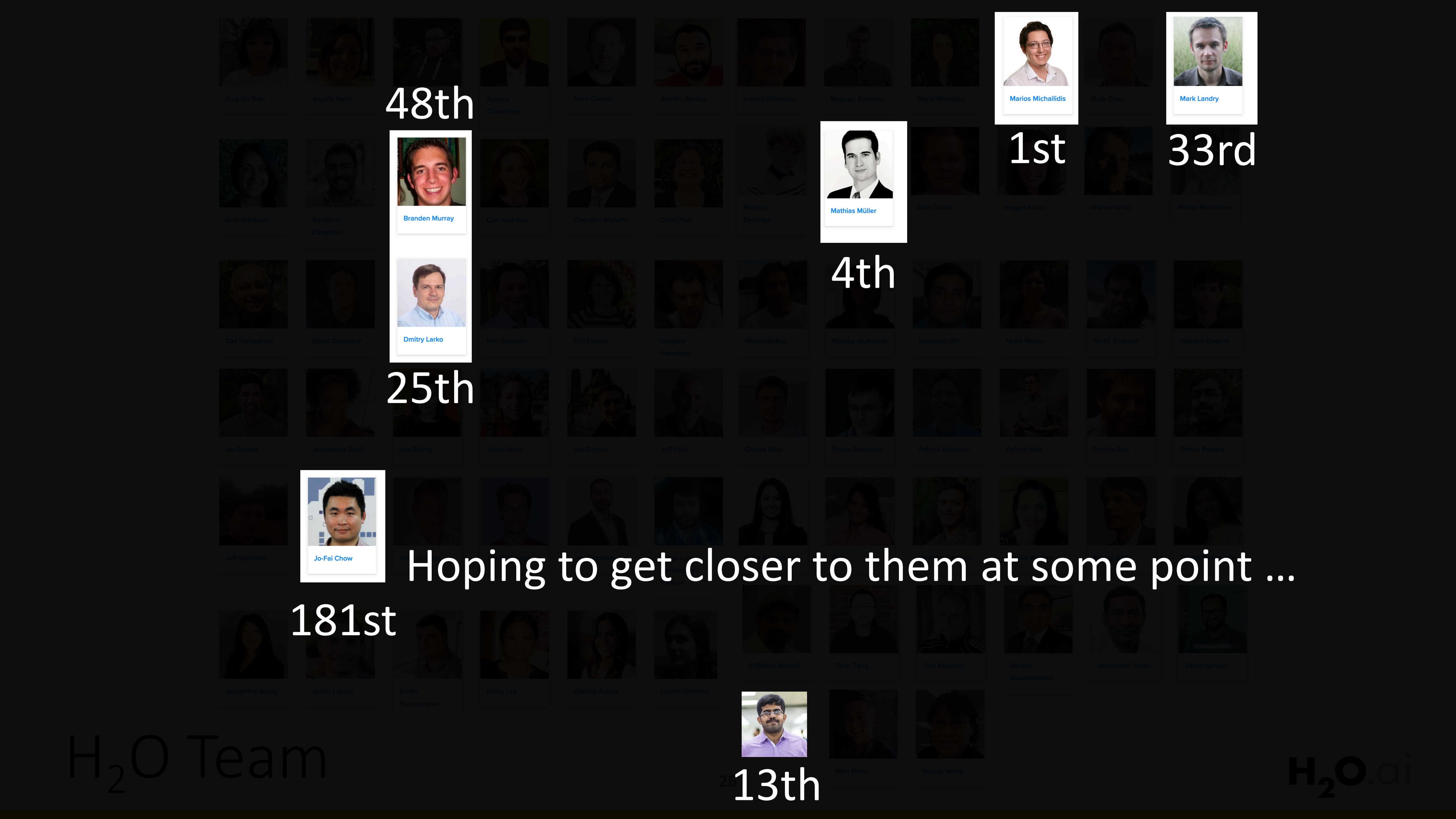
Erin LeDell

Navdeep Gill

H_2O Team

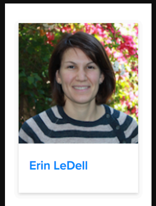


Marios Michailidis
Mark Landry
48th
Branden Murray
Mathias Müller
1st
33rd
4th
Dmitry Larko
25th
Kaggle Grand Masters
(and their Highest Rank)
113 Grandmasters
980 Masters
3,339 Experts
46,135 Contributors
33,242 Novices
About 80,000 Kagglers
H2O Team
13th

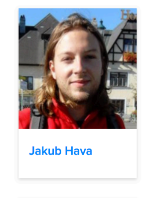


48th
Branden Murray
Dmitry Larko
25th
Mathias Müller
4th
Marios Michailidis
1st
Mark Landry
33rd
Jo-Fai Chow
181st
Hoping to get closer to them at some point …
13th
H2O Team

H_2O Team at eRum 2018



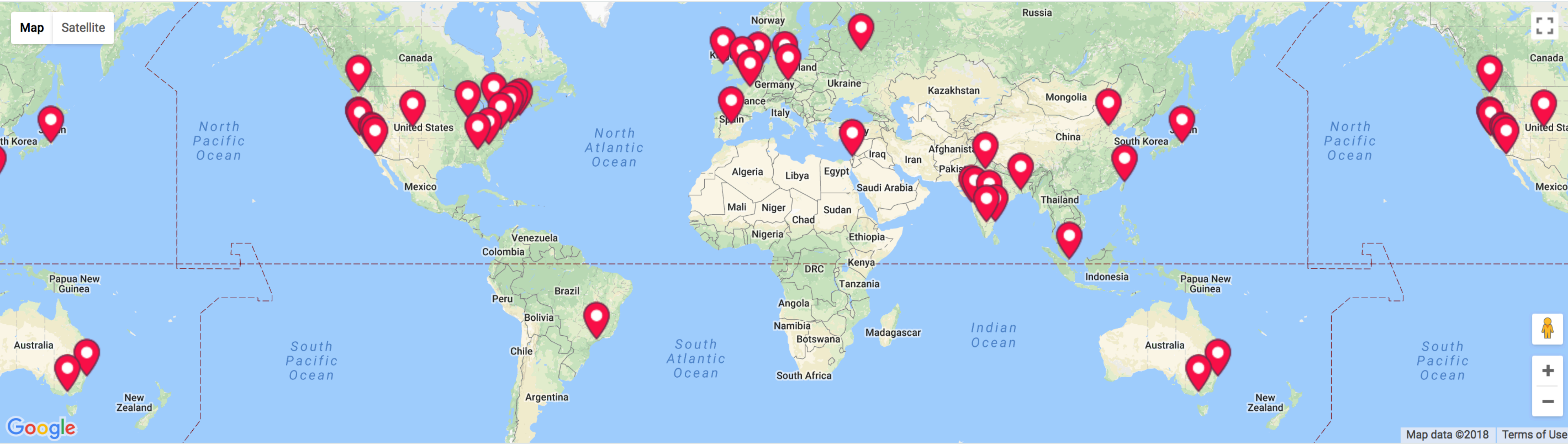
Erin LeDell – Ask her about AutoML



Jakub (or Kuba) Hava – Ask him about Sparkling Water (H_2O + Spark)

Joe Chow – Happy to trade rare H_2O swag for beers … or talk about H_2O-3, Driverless AI, R, Shiny …

Meetup | Pro



H2O Artificial Intelligence and Machine Learning

Members	Groups	Countries
78,356	39	18

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

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

We sponsor meetups

R-Ladies London
Location
London, United Kingdom
Members
1,040
Organizers
Chiin Tan and 7 others
You're a member

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

London Data Science Workshop (formerly London Kaggle Meetup)
Location
London, United Kingdom
Members
2,783
Organizers
Alex Glaser and 6 others
You're a member

Women in Kaggle
Location
London, United Kingdom
Members
261
Organizers
Julia MacMillan and 3 others
You're a member

… and more

We encourage diversity

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
Since Dec 2017 = 3/8 = 37.5%		

Encourage your friends/ colleagues to give a talk.

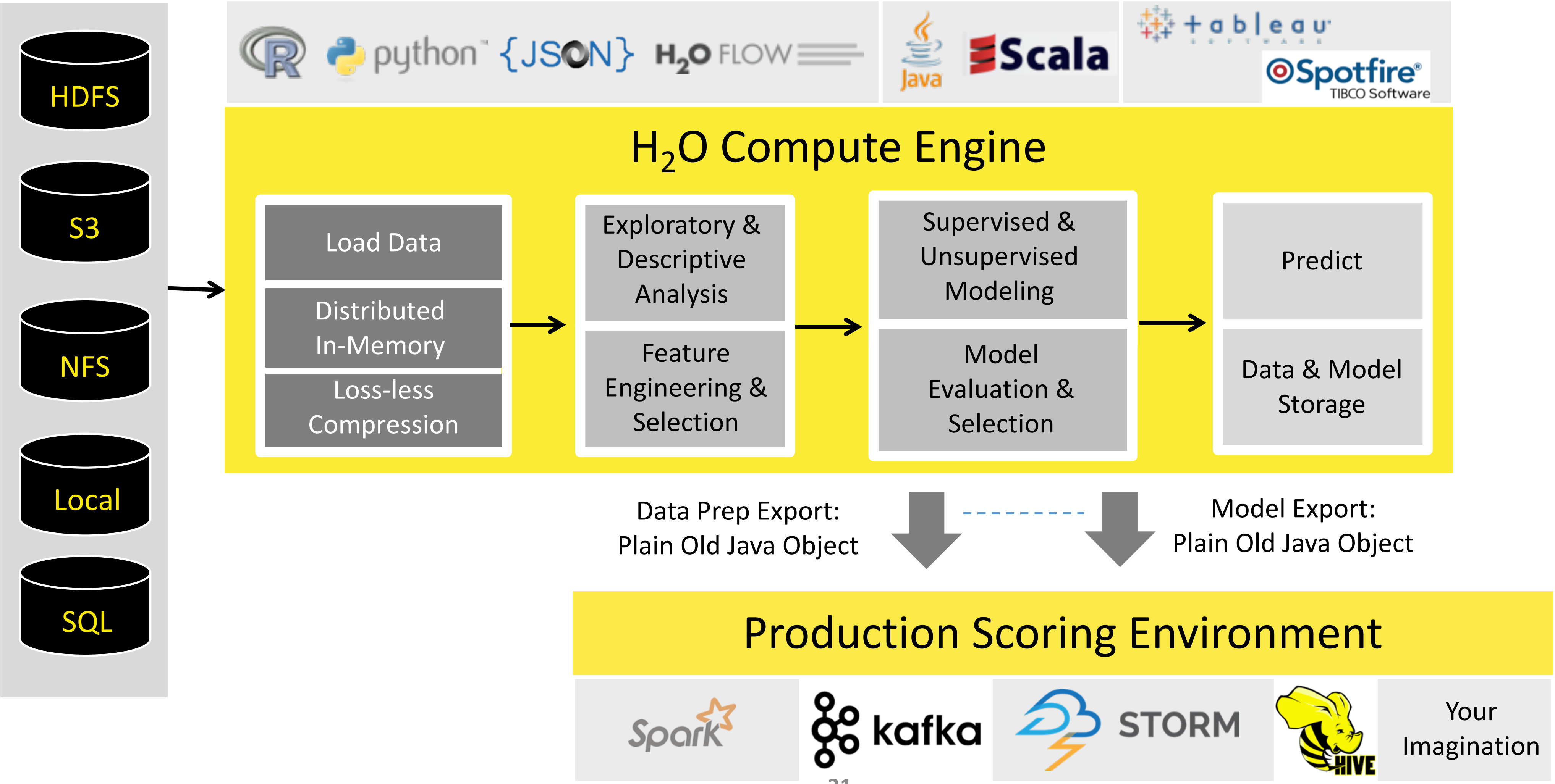
joe@h2o.ai

About H_2O AutoML

Automatic Machine Learning with H_2O

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

High Level Architecture



H_2O-3 Algorithms Overview

Supervised Learning

Statistical Analysis

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

Ensembles

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

Deep Neural Networks

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

Unsupervised Learning

Clustering

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

Dimensionality Reduction

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

Anomaly Detection

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

Multiple Interfaces

High Level Architecture

HDFS

S3

NFS

Local

SQL

R python {JSON} H2O FLOW Java Scala

tableau Spotfire TIBCO Software

H_2O Compute Engine

Load Data

Distributed In-Memory

Loss-less Compression

Exploratory & Descriptive Analysis

Feature Engineering & Selection

Supervised & Unsupervised Modeling

Model Evaluation & Selection

Predict

Data & Model Storage

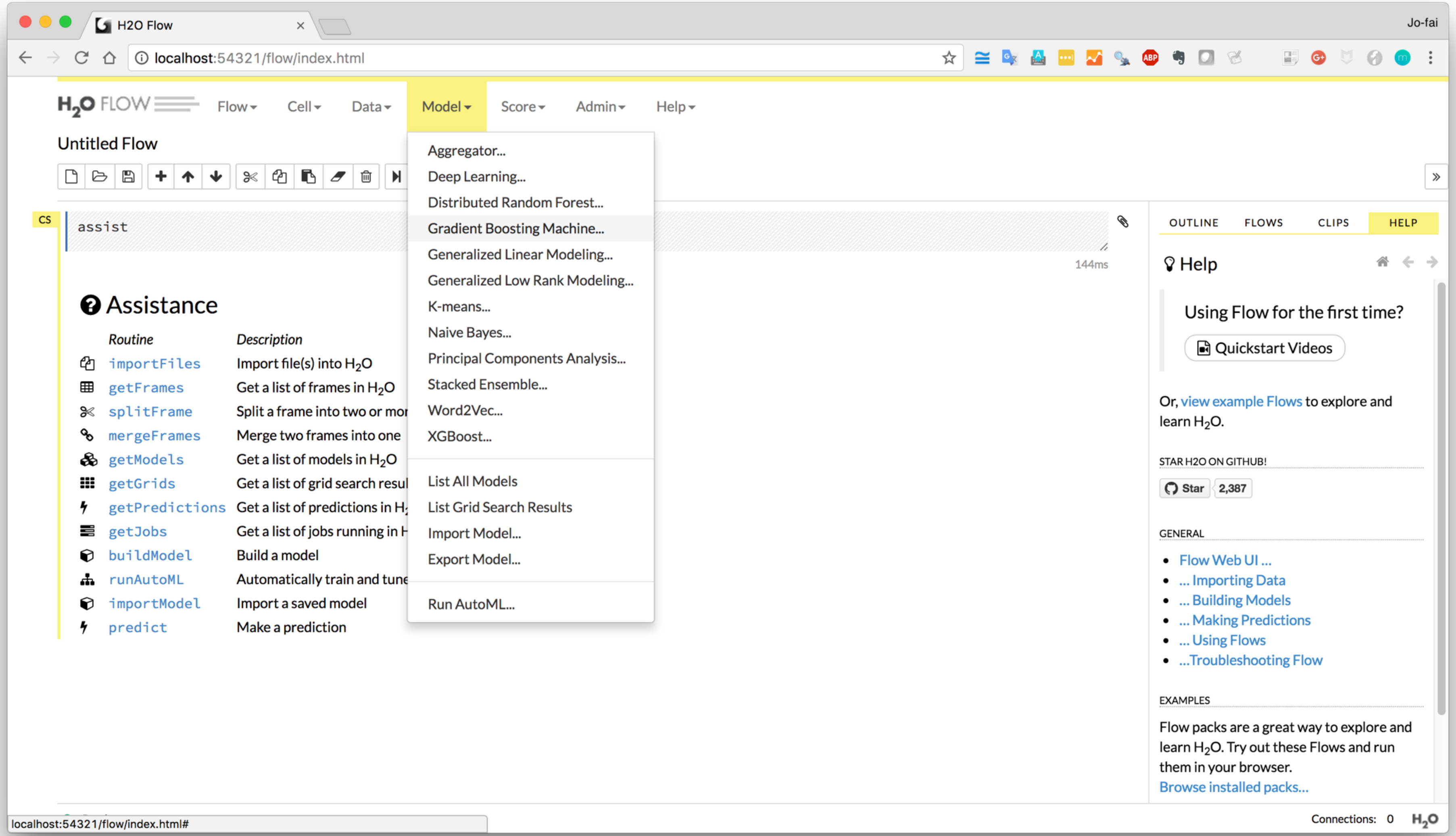
Data Prep Export: Plain Old Java Object

Model Export: Plain Old Java Object

Production Scoring Environment

Spark kafka STORM HIVE Your Imagination

H_2O Flow (Web)


H2O Flow
localhost:54321/flow/index.html
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
assist
Assistance
Routine
Description
importFiles
Import file(s) into H2O
getFrames
Get a list of frames in H2O
splitFrame
mergeFrames
Merge two frames into one
getModels
Get a list of models in H2O
getGrids
getPredictions
getJobs
buildModel
Build a model
runAutoML
importModel
Import a saved model
predict
Make a prediction
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...
List All Models
List Grid Search Results
Import Model...
Export Model...
Run AutoML...
OUTLINE
FLOWS
CLIPS
HELP
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...
Connections: 0

Using H_2O with R and Python

~/Documents/repo_h2o/sales-engineering - master - RStudio Source Editor

credit_card_example.R

```r
# Credit Card Example

# Datasets:
# https://s3.amazonaws.com/h2o-training/credit_card/credit_card_train.csv
# https://s3.amazonaws.com/h2o-training/credit_card/credit_card_test.csv

# Start and connect to a local H2O cluster
library(h2o)
h2o.init(nthreads = -1)

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

# Look at datasets
summary(df_train)
summary(df_test)

# Define features and target
features = colnames(df_test)
target = "DEFAULT_PAYMENT_NEXT_MONTH"

# Train a GBM model
model_gbm = h2o.gbm(x = features,
                    y = target,
                    training_frame = df_train,
                    seed = 1234)
print(model_gbm)

# Use GBM model for making predictions
yhat_test = h2o.predict(model_gbm, newdata = df_test)
head(yhat_test)

# (Extra) Use H2O's AutoML
aml = h2o.automl(x = features,
                 y = target,
                 training_frame = df_train,
                 max_runtime_secs = 60,
                 seed = 1234)

# Print leaderboard
print(aml@leaderboard)

# Use best model for making predictions
best_model = aml@leader
yhat_test = h2o.predict(best_model, newdata = df_test)
head(yhat_test)
```

credit_card_example

localhost:8888/notebooks/credit_card_example.ipynb

jupyter credit_card_example Last Checkpoint: 5 minutes ago (unsaved changes)

In [2]:

```python
# 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_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av
  JVM stdout: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av/h2o_jofaichow_started_from_python.out
  JVM stderr: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpFdP3Av/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_id7iqa
H2O cluster total nodes:	1

In [3]:

```python
# 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%
```

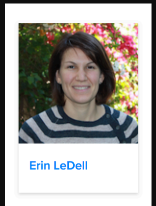
In [4]:

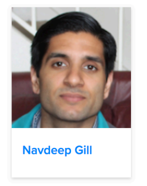
```python
# Look at datasets
df_train.summary()
df_test.summary()
```

	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	-0.15537037037	-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	0

H_2O.ai

H_2O AutoML



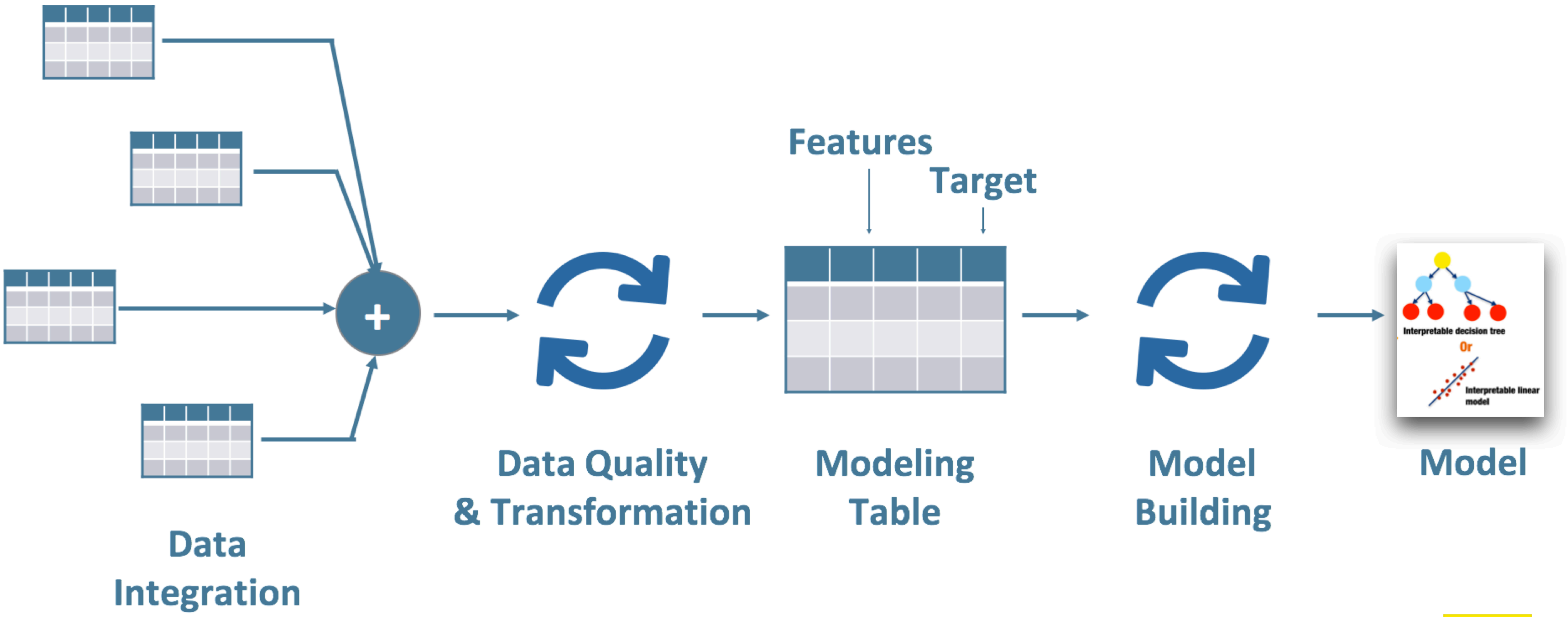


Erin LeDell

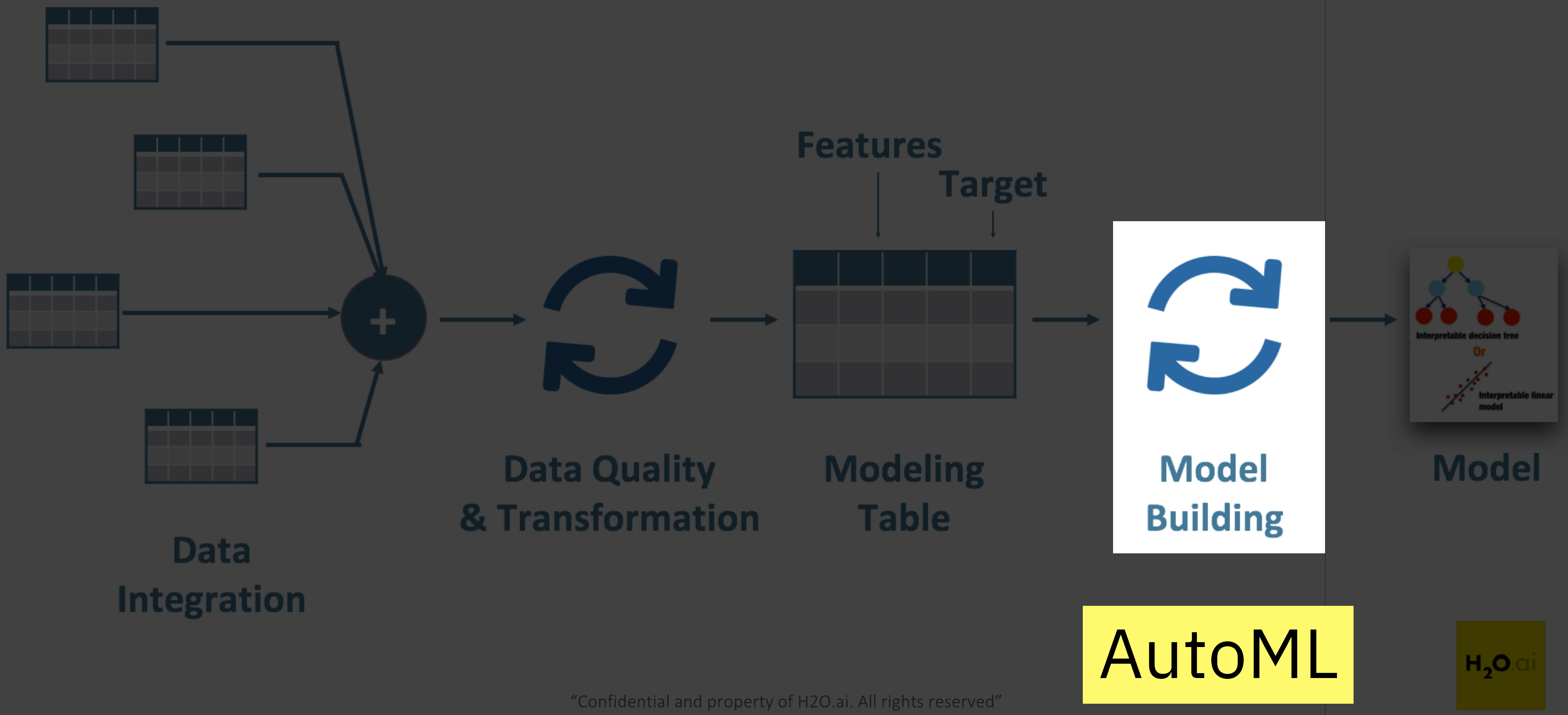
Navdeep Gill

H_2O Team

Typical Enterprise Machine Learning Workflow



Typical Enterprise Machine Learning Workflow



AutoML Interface

The H2O AutoML interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time constraint or limit on the number of total models trained.

In both the R and Python API, AutoML uses the same data-related arguments, `x`, `y`, `training_frame`, `validation_frame`, as the other H2O algorithms. Most of the time, all you'll need to do is specify the data arguments. You can then configure values for `max_runtime_secs` and/or `max_models` to set explicit time or number-of-model limits on your run.

Required Parameters

Required Data Parameters

- y: This argument is the name (or index) of the response column.
- training_frame: Specifies the training set.

Required Stopping Parameters

One of the following stopping strategies (time or number-of-model based) must be specified. When both options are set, then the AutoML run will stop as soon as it hits one of either of these limits.

- max_runtime_secs: This argument controls how long the AutoML run will execute for. This defaults to 3600 seconds (1 hour).
- **max_models**: Specify the maximum number of models to build in an AutoML run, excluding the Stacked Ensemble models. Defaults to `NULL/None`.

AutoML Output

The AutoML object includes a “leaderboard” of models that were trained in the process, including the 5-fold cross-validated model performance (by default). The number of folds used in the model evaluation process can be adjusted using the `nfolds` parameter. If the user would like to score the models on a specific dataset, they can specify the `leaderboard_frame` argument, and then the leaderboard will show scores on that dataset instead.

The models are ranked by a default metric based on the problem type (the second column of the leaderboard). In binary classification problems, that metric is AUC, and in multiclass classification problems, the metric is mean per-class error. In regression problems, the default sort metric is deviance. Some additional metrics are also provided, for convenience.

Here is an example leaderboard for a binary classification task:

model_id	auc	logloss
StackedEnsemble_AllModels_0_AutoML_20171121_012135	0.788321	0.554019
StackedEnsemble_BestOfFamily_0_AutoML_20171121_012135	0.783099	0.559286
GBM_grid_0_AutoML_20171121_012135_model_1	0.780554	0.560248
GBM_grid_0_AutoML_20171121_012135_model_0	0.779713	0.562142
GBM_grid_0_AutoML_20171121_012135_model_2	0.776206	0.564970
GBM_grid_0_AutoML_20171121_012135_model_3	0.771026	0.570270
DRF_0_AutoML_20171121_012135	0.734653	0.601520
XRT_0_AutoML_20171121_012135	0.730457	0.611706
GBM_grid_0_AutoML_20171121_012135_model_4	0.727098	0.666513
GLM_grid_0_AutoML_20171121_012135_model_0	0.685211	0.635138

About Machine Learning Interpretability

LIME (Local Interpretable Model-Agnostic Explanations)

… and more

Acknowledgement

- **Marco Tulio Ribeiro**: Original LIME Framework and Python package

- **Thomas Lin Pedersen**: LIME R package
- **Matt Dancho**: LIME + H2O AutoML example + LIME R package improvement

- **Kasia Kulma**: LIME + H2O AutoML example

Why Should I Trust Your Model?

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

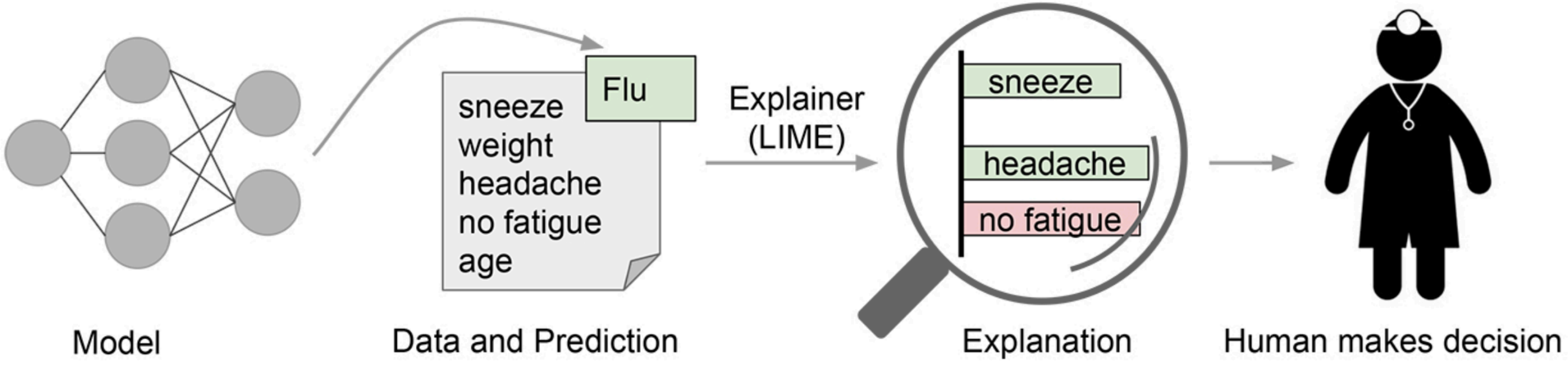


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

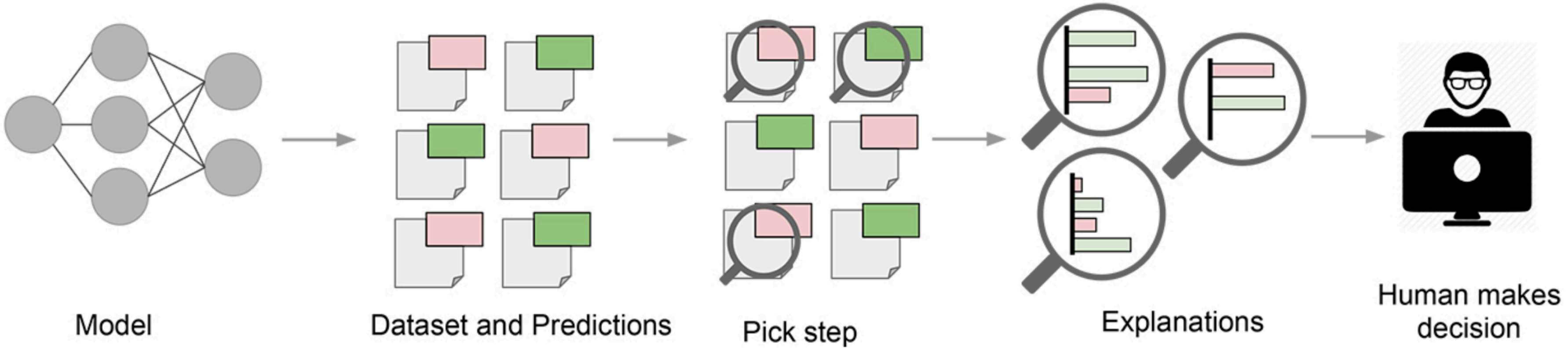


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

Local Interpretable Model-Agnostic Explanations LIME - How does it work?

Theory

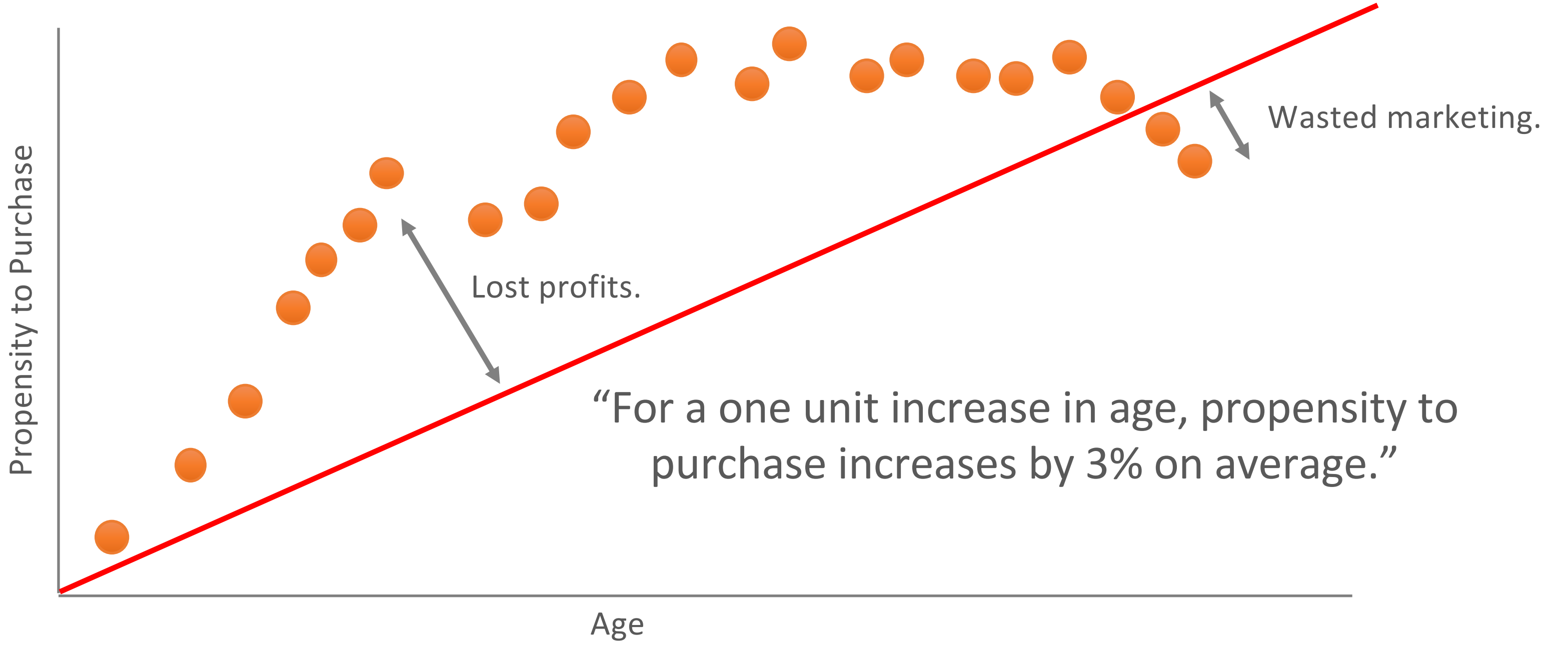
- LIME approximates model locally as logistic or linear model
- Repeats process many times
- Outputs features that are most important to local models

Outcome

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

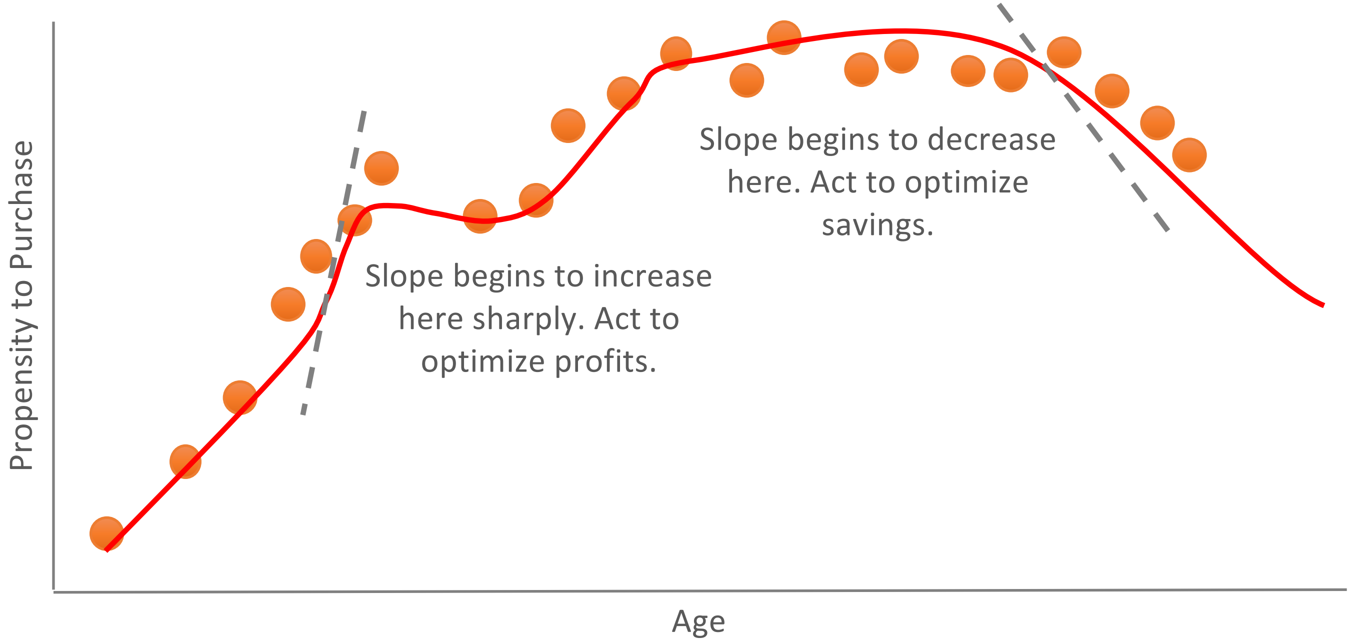
Linear Models

Exact explanations for approximate models.



Machine Learning

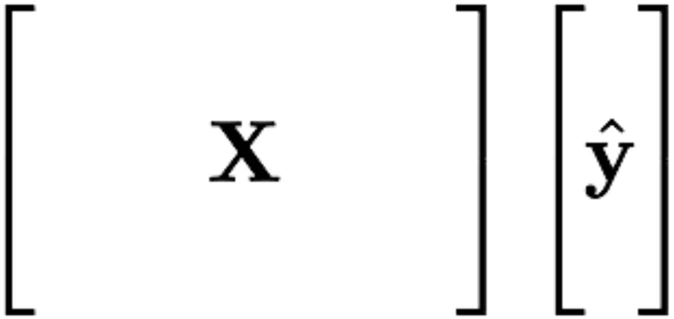
Approximate explanations for exact models.



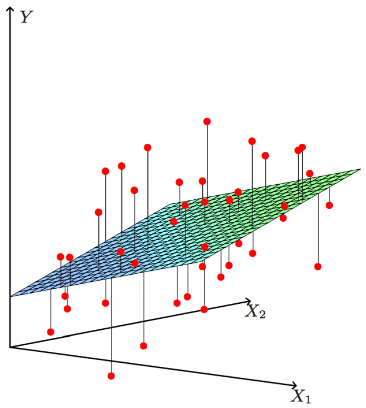
H2O.ai

… there are more techniques!

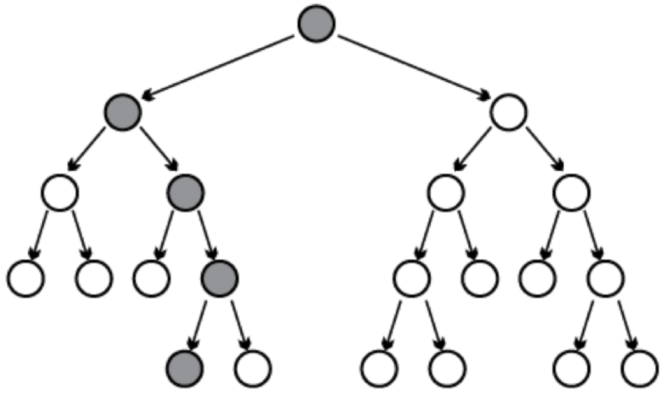
Machine Learning Interpretability



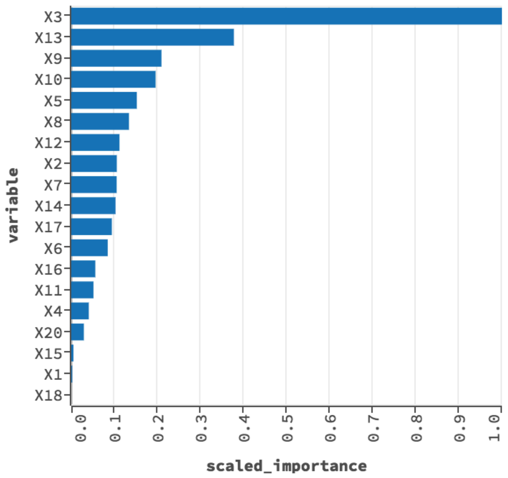
Local Models



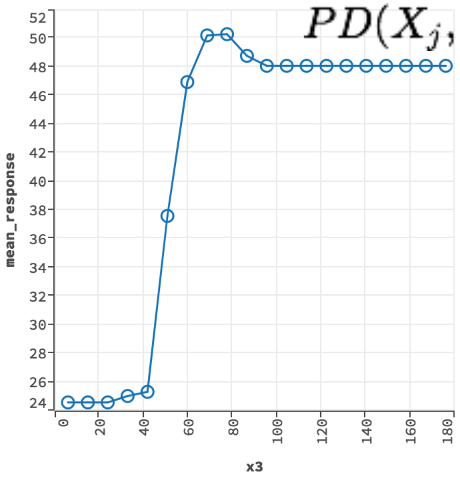
Surrogate Model



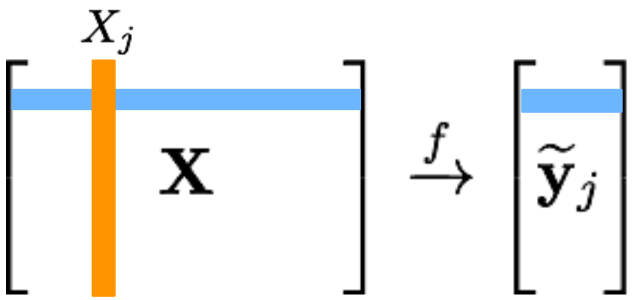
Variable Importance



Partial Dependence



$$PD(X_j, f) = g(\tilde{\mathbf{y}}_j)$$



Workshop day (May 14, 2018 -- Monday)

Earlier Today ...

Room	N15 Aud A (150 seats)	N15 Aud B (150 seats)	N15 103 (75 seats)	N15 101 (50 seats)	N15 106 (40 seats)	N15 203 (25 seats)	N15 202 (20 seats)
8:00 – 8:30	Registration						
9:00 – 10:00	Efficient R programming	DALEX: Descriptive mAchine Learning EXplanations	Clean R code - how to write it and what will the benefits be	Building an Interpretable NLP model to classify tweets	Geocomputation with R	Graphs: A datastructure to query	Forwards Package Development Workshop for Women
10:30	Coffee break						
11:00 – 12:00	Efficient R programming	DALEX: Descriptive mAchine Learning EXplanations	Clean R code - how to write it and what will the benefits be	Building an Interpretable NLP model to classify tweets	Geocomputation with R	Graphs: A datastructure to query	Forwards Package Development Workshop for Women
12:30 – 13:00	Lunch break						
13:30 – 14:30	Deep Learning with Keras for R	Automatic and Interpretable Machine Learning in R with H2O and LIME	The beauty of data manipulation with data.table	Building a package that lasts	Building a pipeline for reproducible data screening and quality control	Plotting spatial data in R	Forwards Package Development Workshop for Women
15:00	Coffee break						
15:30 – 16:30	Deep Learning with Keras for R	Automatic and Interpretable Machine Learning in R with H2O and LIME	The beauty of data manipulation with data.table	Building a package that lasts	Building a pipeline for reproducible data screening and quality control	Plotting spatial data in R	Forwards Package Development Workshop for Women
...	Walk from CEU to Akvarium Klub						

This Workshop

Erin LeDell
@ledell

Following

List of #MachineLearning model interpretability pkgs in R:
lime, ShapleyR, live, xgboostExplainer, breakDown, DALEX #eRum2018

github.com/thomasp85/lime
github.com/redichh/Shaple…
mi2datalab.github.io/live/
github.com/AppliedDataSci …
pbiecek.github.io/breakDown/
pbiecek.github.io/DALEX/

AppliedDataSciencePartners/xgboostExplainer
xgboostExplainer - An R package that makes xgboost models fully interpretable
github.com

8:34 AM - 14 May 2018

4 Retweets 4 Likes

1 4 4

Tweet your reply

Erin LeDell @ledell · 10m
One more (for the randomForest package):
mi2datalab.github.io/randomForestEx…

O'REILLY® Ideas Learning Platform Conferences Shop

ON OUR RADAR AI BUSINESS DATA DESIGN ECONOMY JUPYTER OPERATIONS SECURITY SOFTWARE ARCHITECTURE SOFTWARE ENGINEERI SEE ALL

DATA SCIENCE

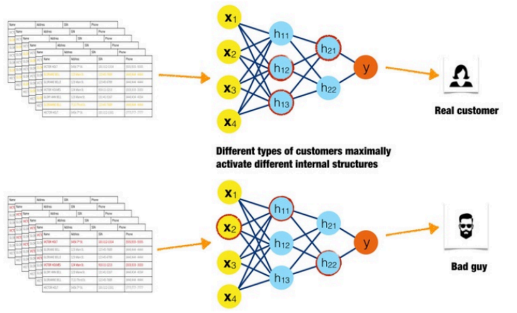
Ideas on interpreting machine learning

Mix-and-match approaches for visualizing data and interpreting machine learning models and results.

By Patrick Hall, Wen Phan, and SriSatish Ambati. March 15, 2017

Check out the "Data Science & Machine Learning" sessions at the Strata Data Conference in London, May 21-24, 2018.

You've probably heard by now that machine learning algorithms can use big data to predict whether a donor will give to a charity, whether an infant in a NICU will develop sepsis, whether a customer will respond to an ad, and on and on. Machine learning can even drive cars and predict elections. ... Err, wait. Can it? I believe it can, but these recent high-profile hiccups should leave everyone who works with data (big or not) and machine learning algorithms asking themselves some very hard questions: do I understand my data? Do I understand the model and answers my machine learning algorithm is giving me? And do I trust these answers?



Inputs activating different neurons in a neural network. (source: Image courtesy of Patrick Hall and the h2o.ai team, used with permission)

https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	**Regression Example** **\examples\regression_...Rmd**
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 pm	Other H_2O News + Q & A

Learning from **Boston Housing** Data

H_2O Machine Learning
Learn the Pattern

Features:
No. of rooms, Crime Rate, Age ...

$$\left[\quad \mathbf{X} \quad\right] \left[\, \mathbf{y} \,\right] \xrightarrow{f} \left[\, \hat{\mathbf{y}} \,\right]$$

Median House Value (medv)

Predictions

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example \examples\classification_...Rmd
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 pm	Other H_2O News + Q & A

Learning from **Diabetes** Data

H_2O Machine Learning
Learn the Pattern

Features:
Pregnant, Age, Glucose ...

$$\left[\mathbf{X}\right] \left[\mathbf{y}\right] \xrightarrow{f} \left[\hat{\mathbf{y}}\right]$$

Positive / Negative

Predictions

@h2oai @matlabulous

#eRum2018 #AutoML #LIME

Please come back at 3:30pm

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

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 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 …

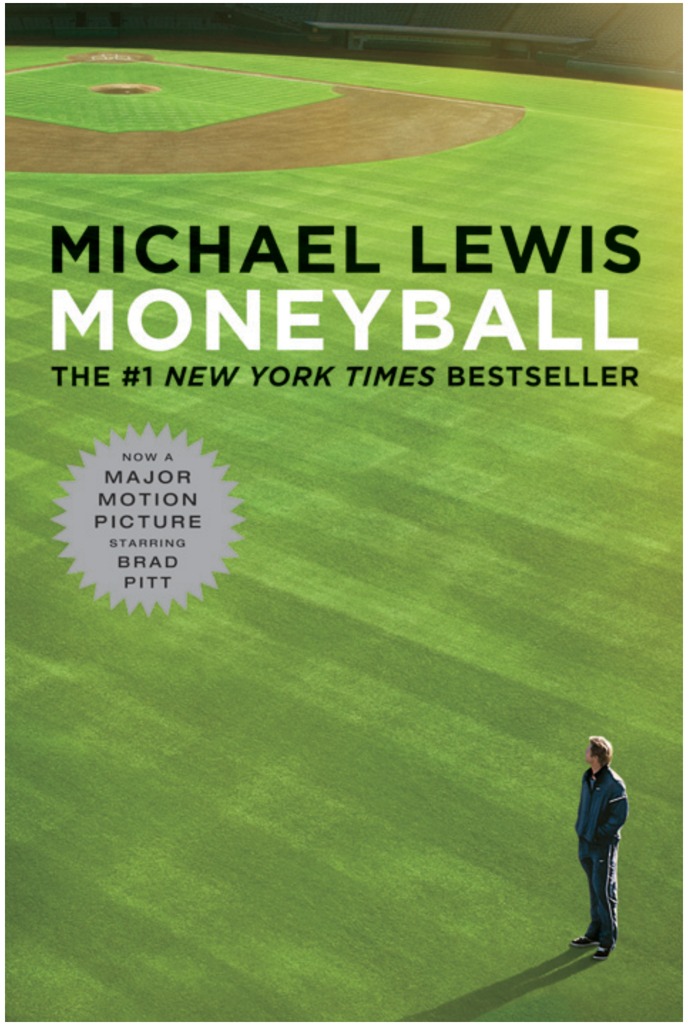
- How to build high quality H_2O models (almost) automatically.
- How to explain predictions from complex H_2O models with LIME.
- **Bonus**: A real use-case that led to multimillion-dollar baseball decisions earlier this year.

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	**Real Use-Case: Moneyball**
4:15 – 4:30 pm	Other H_2O News + Q & A

Making Multimillion-Dollar Decisions with H_2O AutoML, LIME and Shiny

My journey to a real Moneyball application

About Moneyball



Billy Beane

Peter Brand
(based on Paul DePodesta)

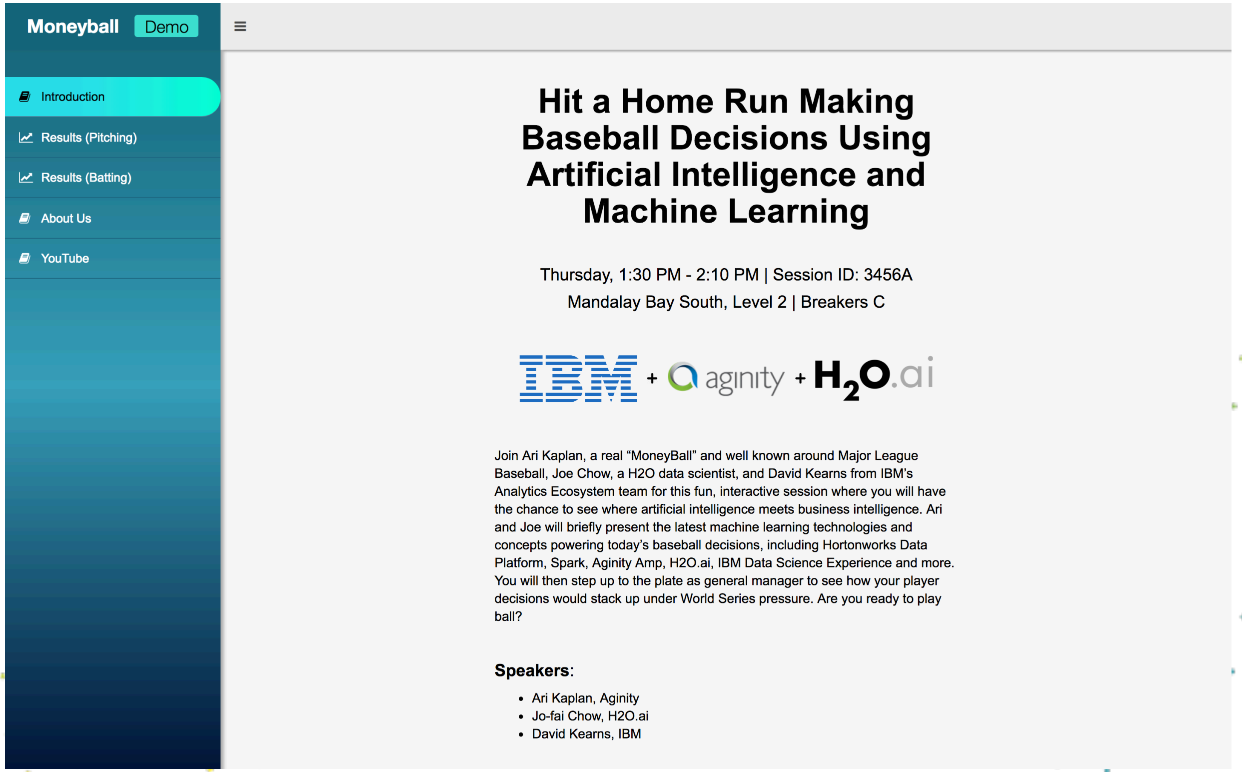
Ari Kaplan – the Real ”Moneyball” Guy

- The real characters in the movie (Billy Beane and Paul DePodesta) did not want to work with Hollywood.

- The filmmaker interviewed Ari instead and created the Paul DePodesta character based on Ari’s real-life story.

- Ari happens to work at Aginity so we have a real ”Moneyball” guy for this demo.

A Proof-of-Concept Demo for IBM Think Conference

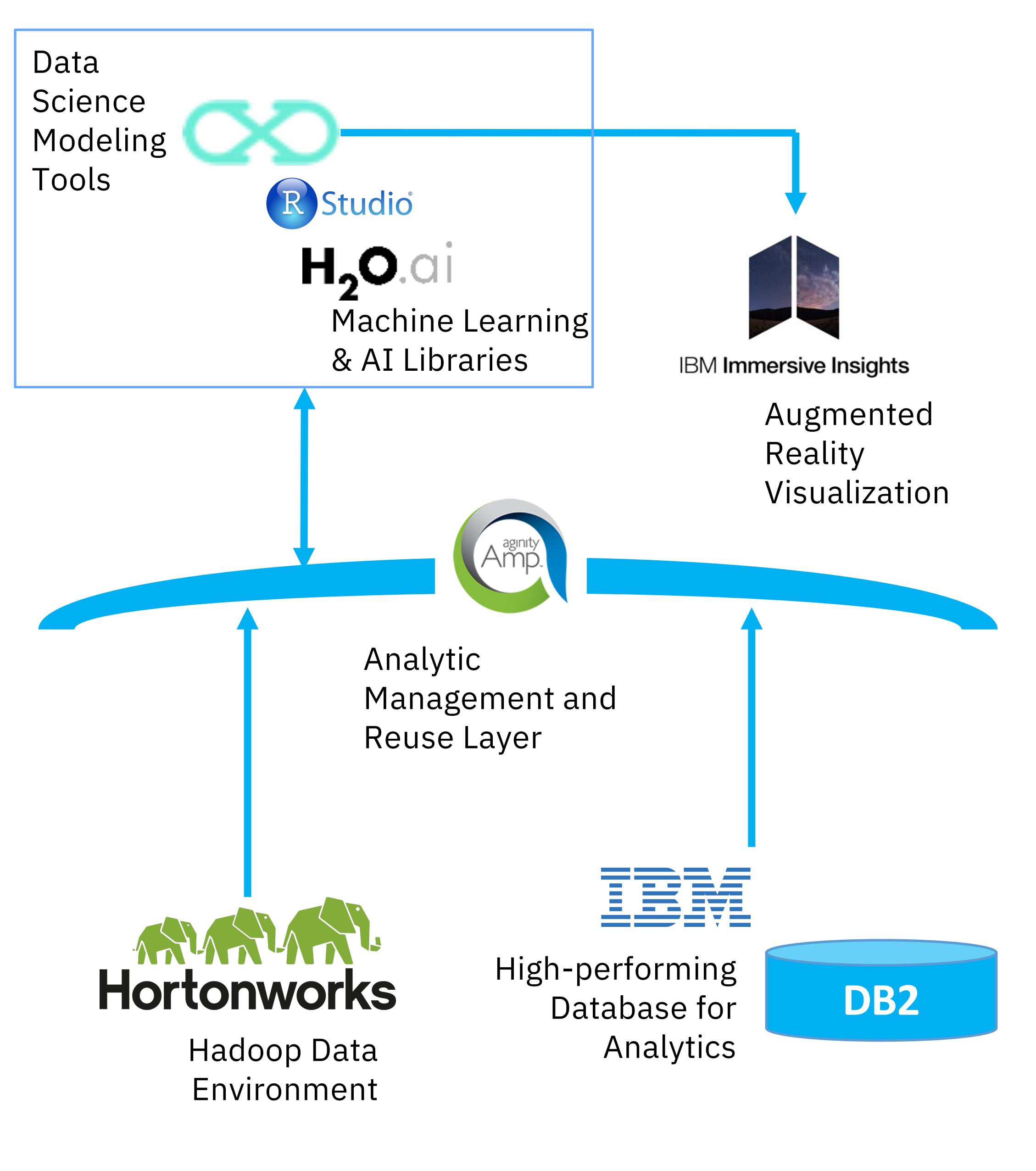


The Architecture

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_2O
4. DSX and H_2O 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



Approach One: Learning from **Lahman** only

H_2O AutoML:
Learning the Pattern

Lahman: Age, Height, Weight …
Historical Performance Stats

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

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

Home Runs
Batting Average
…

Predictions

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

H_2O AutoML:
Learn the Pattern

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

$$\left[\mathbf{X}\right] \left[\mathbf{y}\right] \xrightarrow{f} \left[\hat{\mathbf{y}}\right]$$

AriDB: Fastball, curveball, slider, velocity ...

Sliding Windows (Stats from previous n years)

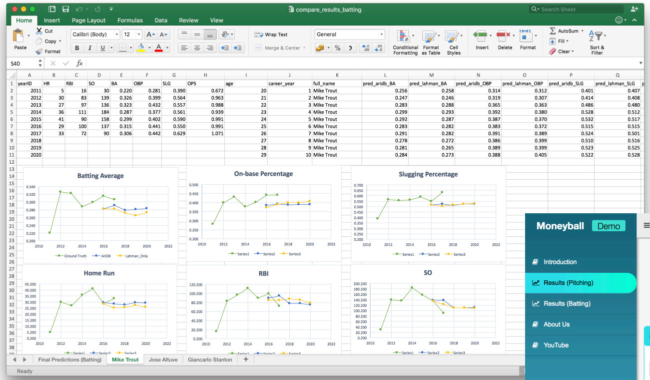
About **300 Lahman** Features +
200 AriDB Features

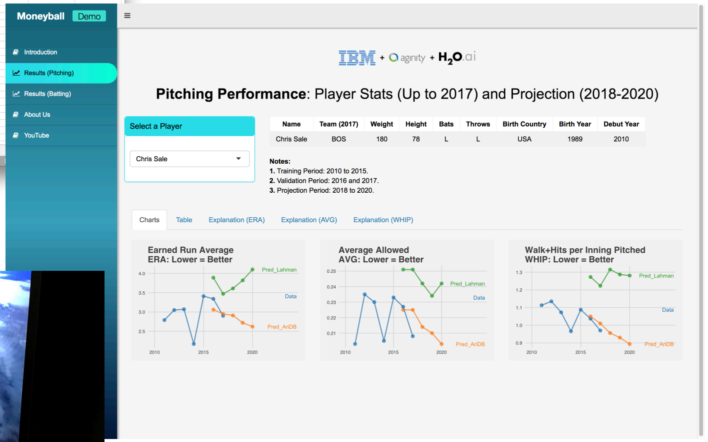
Home Runs
Batting Average
...

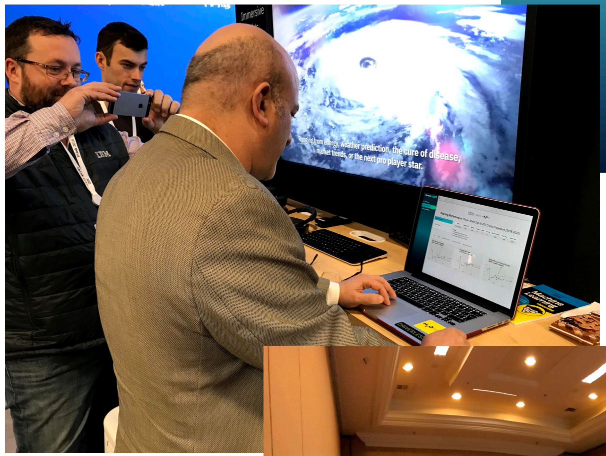
Predictions

Timeline

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

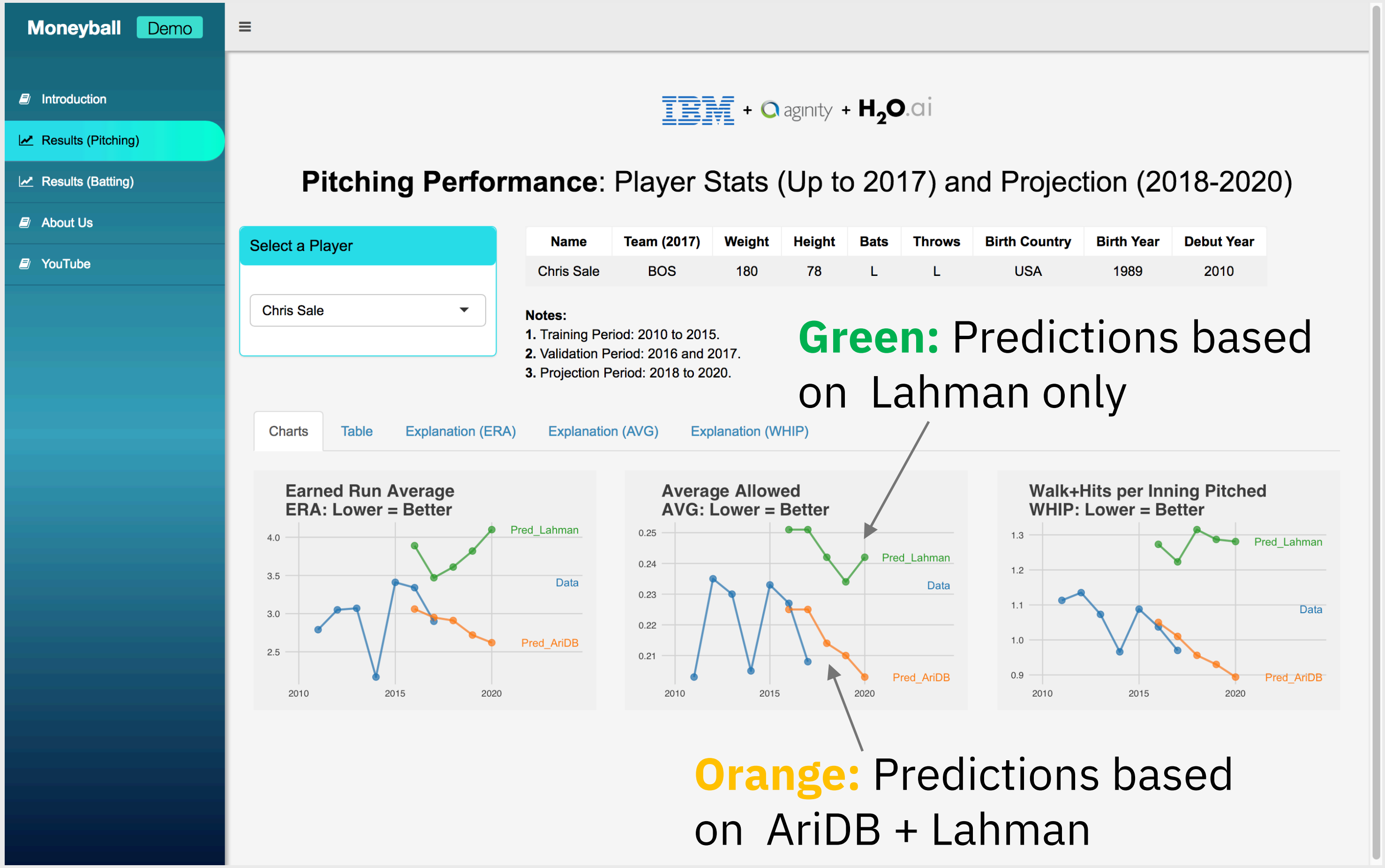


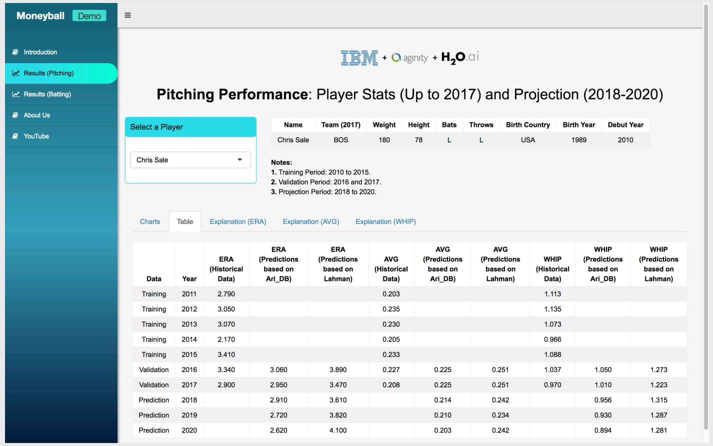


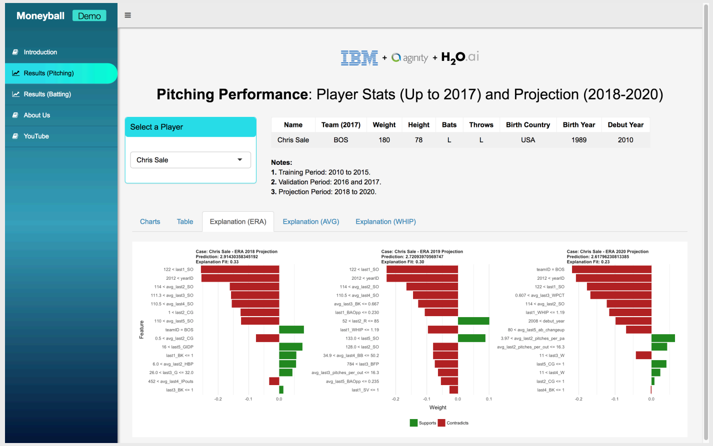


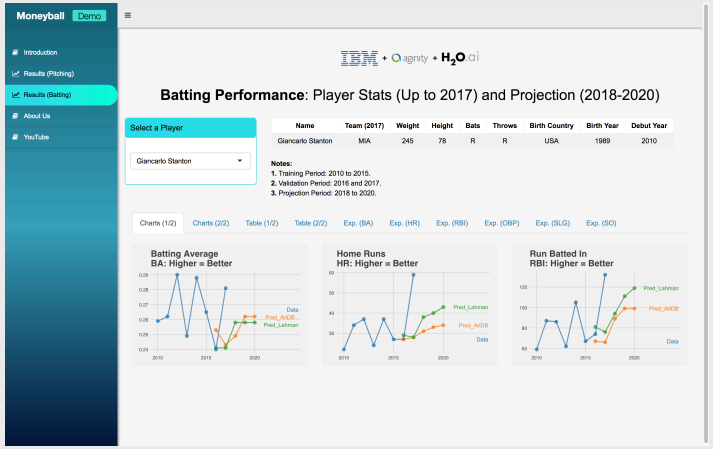
Presentation

Shiny App









Acknowledgement

IBM Data Science @IBMDataScience — Following

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

9:04 PM - 22 Mar 2018

H_2O.ai

Time	Topics / Tasks
1:30 – 1:45 pm	Install **h2o**, **lime**, **mlbench** from CRAN slides/code: bit.ly/**joe_eRum_2018**
1:45 – 2:00 pm	Introduction (H_2O, AutoML, LIME)
2:00 – 2:30 pm	Regression Example
2:30 – 3:00 pm	Classification Example
3:00 – 3:30 pm	☕🍰🍪
3:30 – 3:45 pm	Quick Recap
3:45 – 4:15 pm	Real Use-Case: Moneyball
4:15 – 4:30 pm	Other H_2O News + Q & A

H_2O Products

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

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

Lightning Fast machine learning on GPUs

DRIVERLESS AI

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

Algorithms on H2O-3 (CPU)

Supervised Learning

Statistical Analysis

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

Ensembles

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

Deep Neural Networks

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

Unsupervised Learning

Clustering

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

Dimensionality Reduction

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

Anomaly Detection

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

Algorithms on H_2O4GPU (more to come)

Supervised Learning

Statistical Analysis

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

Ensembles

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

Deep Neural Networks

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

Unsupervised Learning

Clustering

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

Dimensionality Reduction

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

Anomaly Detection

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

H2O.ai

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 (**M**odel **O**b**J**ect, **O**ptimized). 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

Thanks!

- Organizers & Sponsors

Budapest
eRum 2018

Platinum

RStudio

Gold

MANGO SOLUTIONS

QUANTIDE knowledge from data

Silver

H2O.ai

emarsys

Open Analytics

BUDAPEST BI FORUM

Bronze

WLOG SOLUTIONS

Jumping Rivers

R consortium

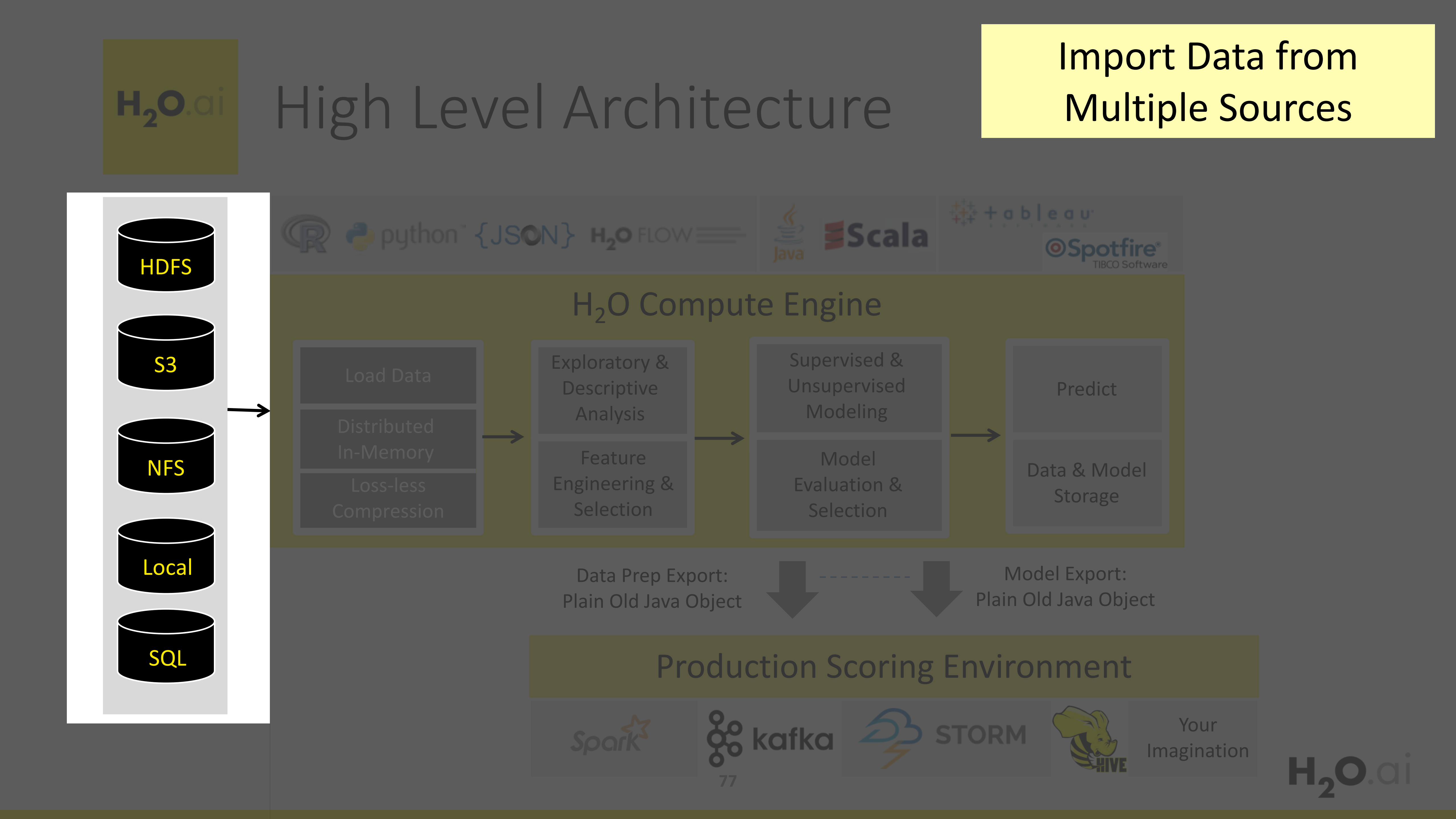
Shift

R-bloggers

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

H2O.ai

Appendix



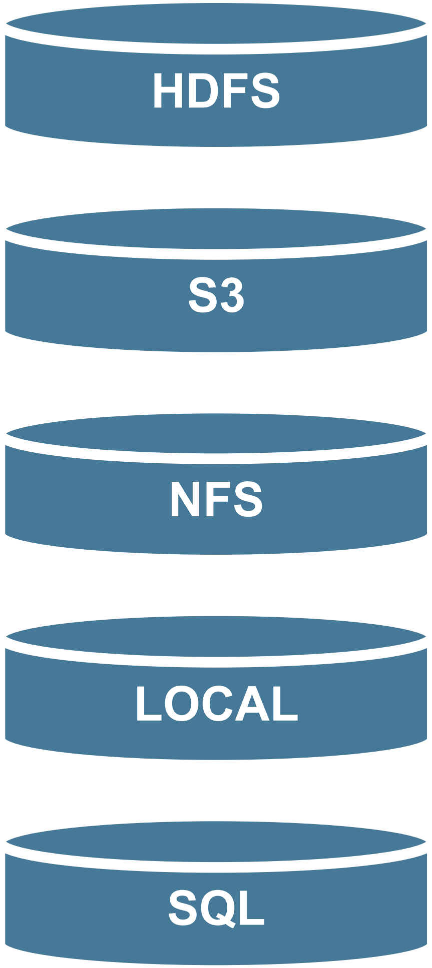
Import Data from Multiple Sources
H2O.ai
High Level Architecture
HDFS
S3
NFS
Local
SQL
python
{JSON}
H2O FLOW
Java
Scala
tableau
Spotfire
TIBCO Software
H2O Compute Engine
Load Data
Distributed In-Memory
Loss-less Compression
Exploratory & Descriptive Analysis
Feature Engineering & Selection
Supervised & Unsupervised Modeling
Model Evaluation & Selection
Predict
Data & Model Storage
Data Prep Export: Plain Old Java Object
Model Export: Plain Old Java Object
Production Scoring Environment
Spark
kafka
STORM
HIVE
Your Imagination
H2O.ai

Supported Formats & Data Sources

9 Formats

File type or Folder of Files

- CSV
- XLS
- XLSX
- ORC*
- Hive*
- SVMLight
- ARFF
- Parquet
- Avro 1.8.0*



5 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

HDFS
S3
NFS
Local
SQL

R · python · {JSON} · H_2O FLOW · Java · Scala · tableau · Spotfire TIBCO Software

H_2O Compute Engine

- Load Data
- Distributed In-Memory
- Loss-less Compression

→

- Exploratory & Descriptive Analysis
- Feature Engineering & Selection

→

- Supervised & Unsupervised Modeling
- Model Evaluation & Selection

→

- Predict
- Data & Model Storage

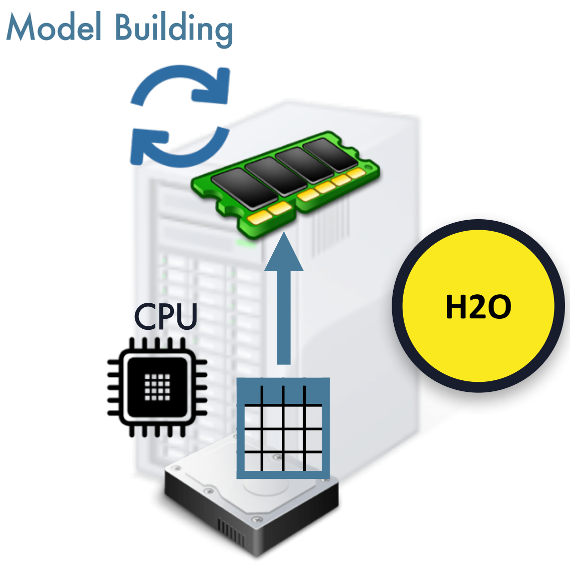
Data Prep Export: Plain Old Java Object

Model Export: Plain Old Java Object

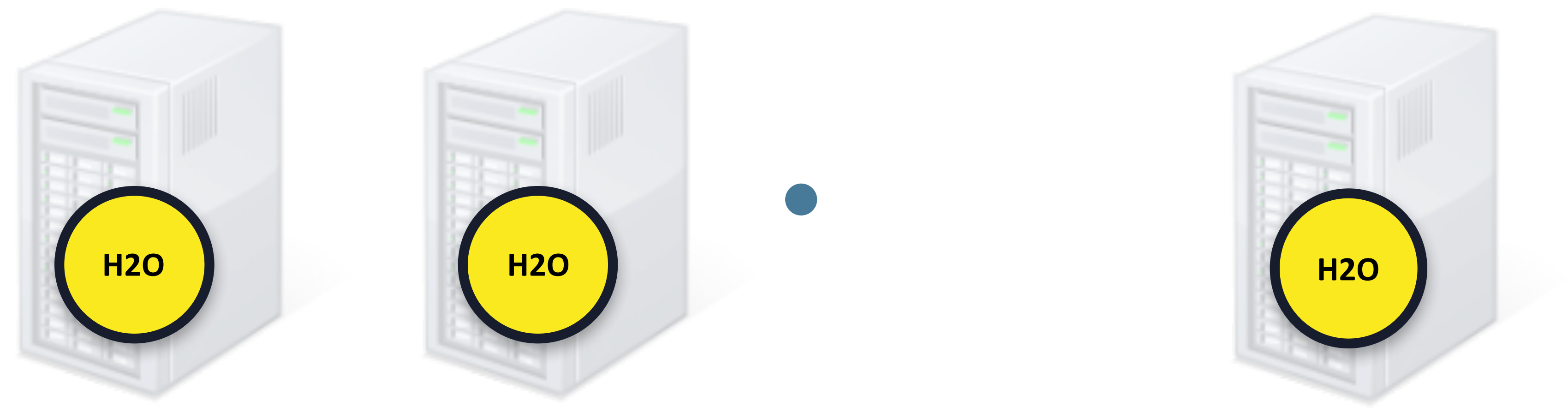
Production Scoring Environment

Spark · kafka · STORM · HIVE · Your Imagination

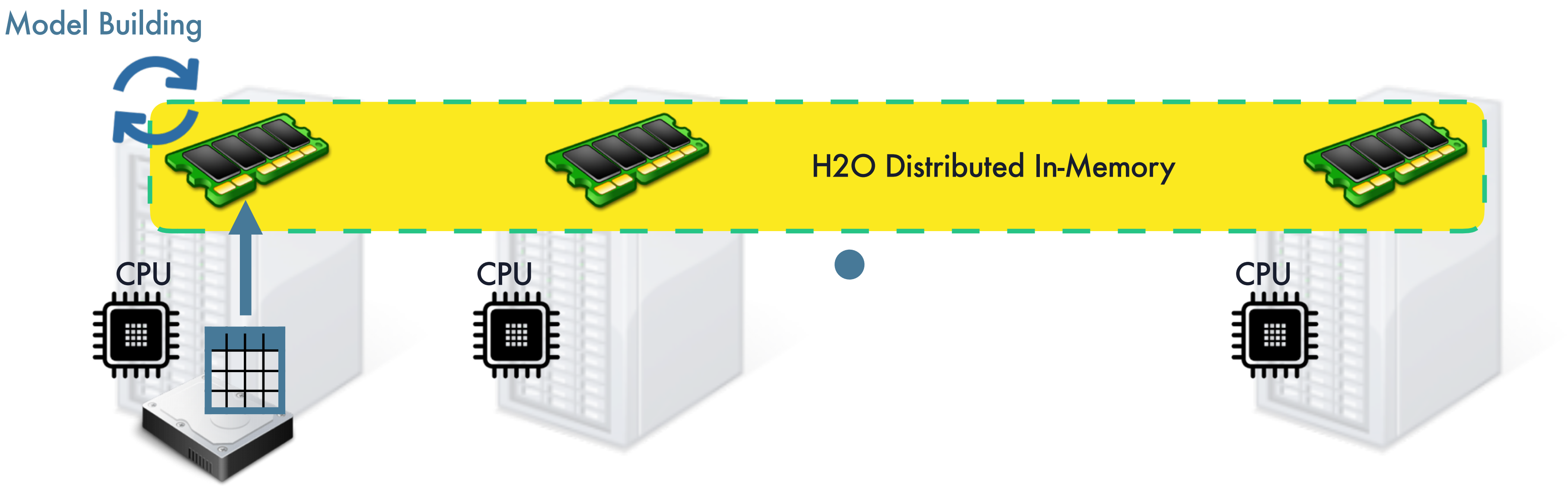
H_2O Core



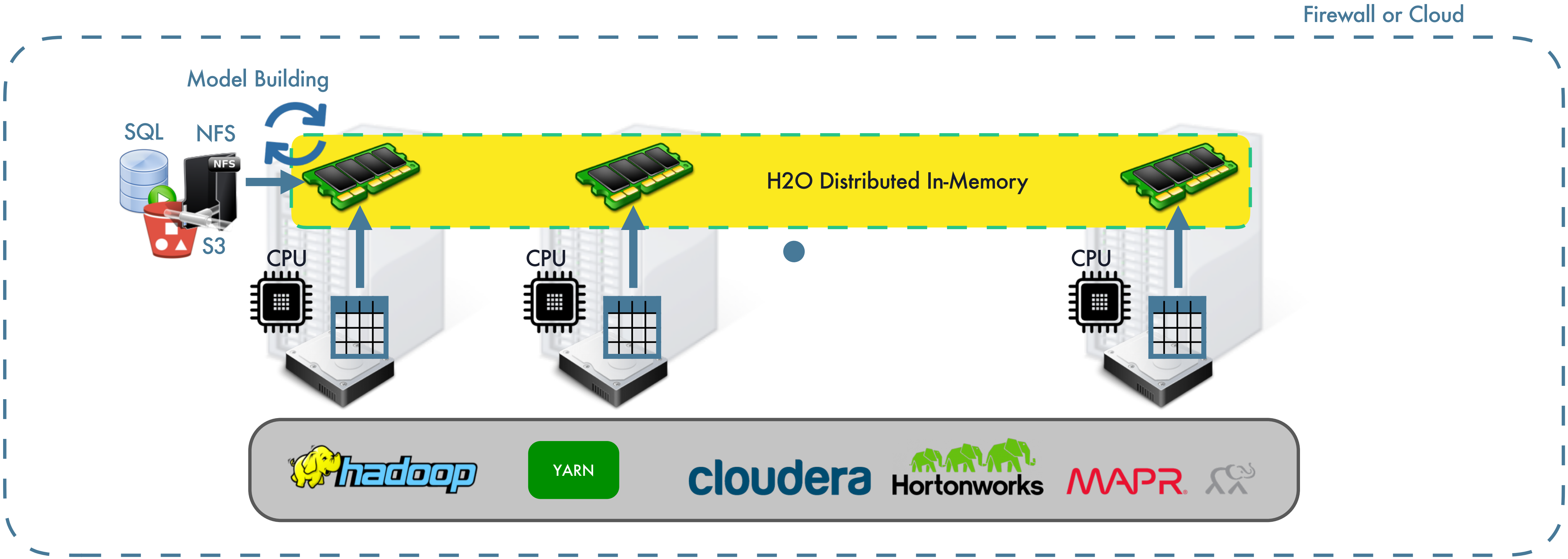
H_2O Core



H_2O Core



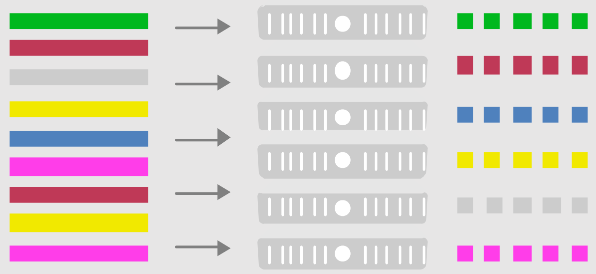
H_2O Core



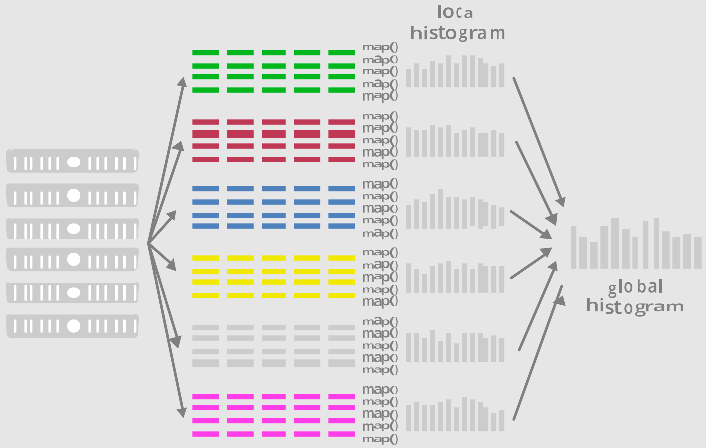
Firewall or Cloud
Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



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

Advantageous Foundation

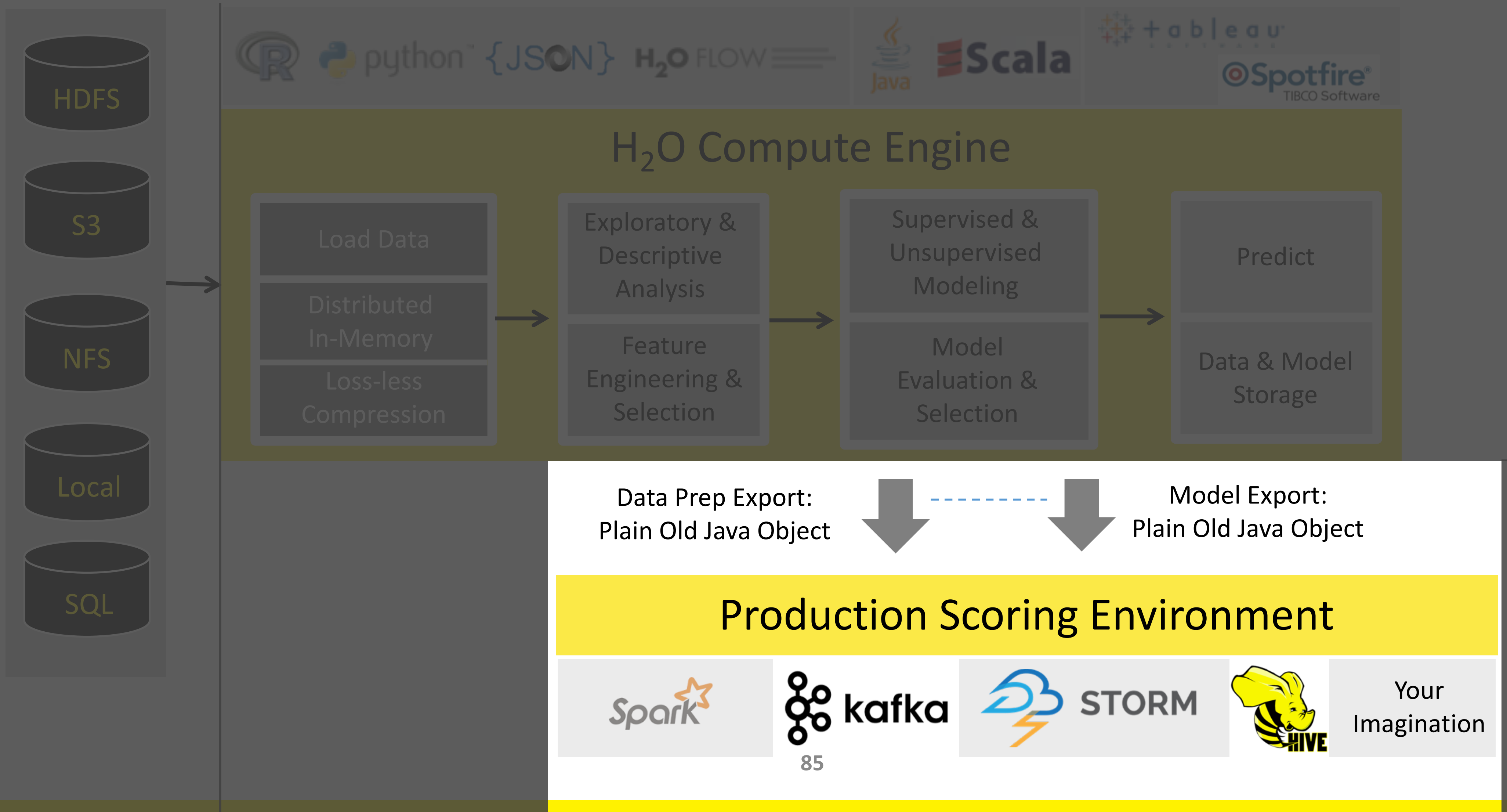
- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in 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.ai

Products | Industries | Customers | Community | Partners | Download

H_2O Documentation

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

Getting Started & User Guides

Open Source | Commercial

H_2O

- 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 H_2O

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:
H_2O 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**

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

Overview

Description

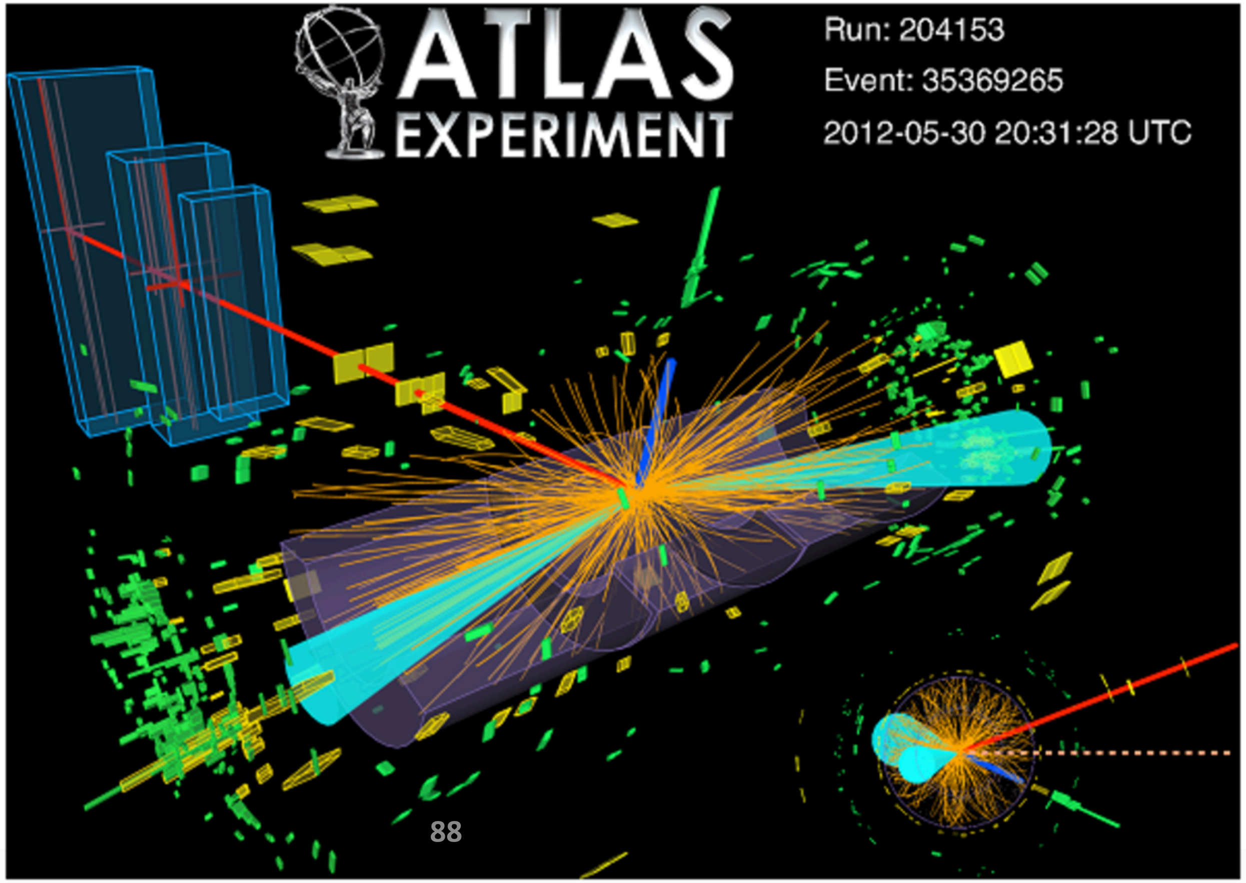
Evaluation

Prizes

About The Sponsors

Timeline

Winners



Learning from Higgs Boson Machine Data

11M Rows

Sensors (Detector) Data

Learn the Pattern

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

28 Features

Historical Outcome
Is it a Higgs Particle (Yes/No)

Predicted Outcome

Raw Data Size: **7.48 GB**

11M Rows

Size (Raw): **7.48 GB**
Compressed: **2.00 GB** (≈ 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	.	.

H2O FLOW   Flow ▾   Cell ▾   Data ▾   Model ▾   Score ▾   Admin ▾   Help ▾

Untitled Flow

- Jobs
- Cluster Status
- Water Meter (CPU meter)

Inspect Log
- View Log
- Download Logs

Advanced
- Create Synthetic Frame...
- Stack Trace
- Network Test
- Profiler
- Timeline
- Shut Down

CS

```
getCloud
```

joe-demo

10 nodes

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*			*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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	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	- / NaN undefined	NaN%	293.18 GB / NaN undefined / 295.83 GB	3.36 TB / 17.04 TB	19%

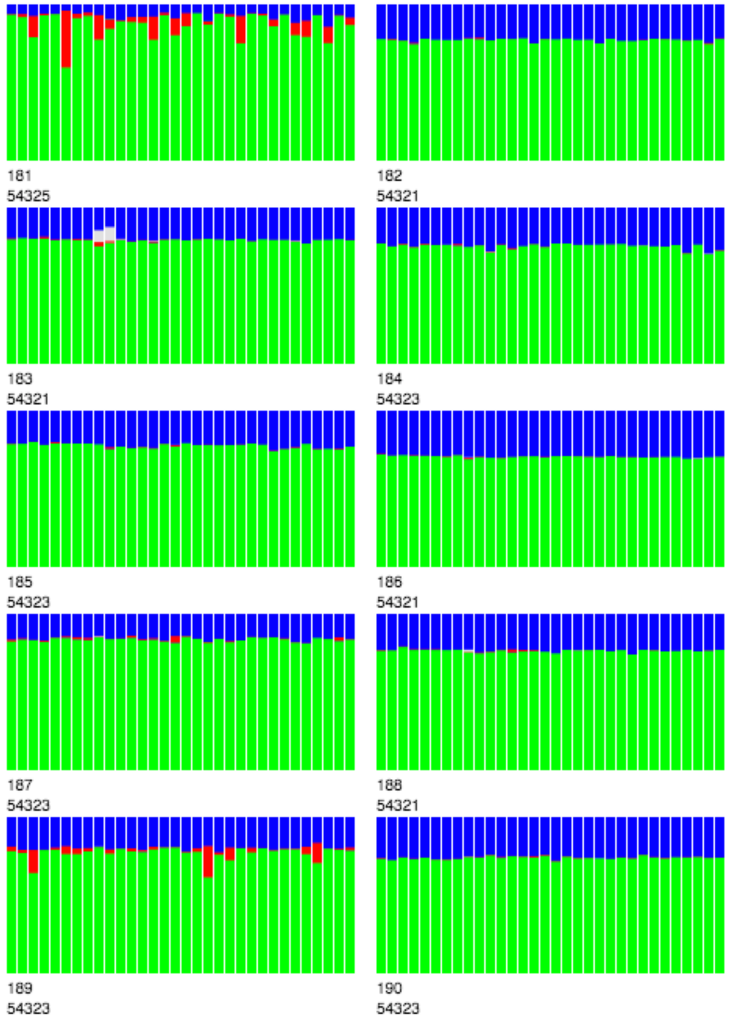
10 x 32 =
320 Cores

10 x 29.6 = 296
GB Memory

H2O.ai

H_2O 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)