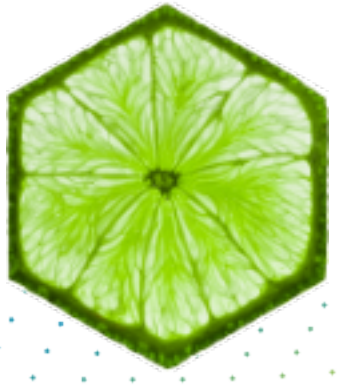
Automatic and Interpretable Machine Learning in R with H_2O and LIME



Jo-fai (Joe) Chow

Data Science Evangelist / Community Manager

joe@h2o.ai

@matlabulous

Download → https://bit.ly/

joe_eRum_2018

Milano R

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.

AutoML + LIME + Shiny = Real Moneyball

$20M

trade 2 weeks prior to the season beginning

IBM + agility + H2O.ai

About Me

Barcelona

Berlin

Brussels

Paris

Milan (2016)

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

    … The harsh reality of startup life …
  - H_2O SWAG Photographer

    #AroundTheWorldWithH2Oai

    Love H_2O? Get some stickers!

H2O.ai

Jo-fai (Joe) Chow
@matlabulous

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

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

Jo-fai (Joe) Chow
@matlabulous

Another #FullHouse @h2oai #LondonAI #meetup tonight. Thanks @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

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

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

Jo-fai (Joe) Chow
@matlabulous

Merci beaucoup Alexia, Samia & Aurelie from @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

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 …

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

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_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

Milano R

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

H2O.ai

Milano R

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

H2O.ai

About **H2O.ai** …

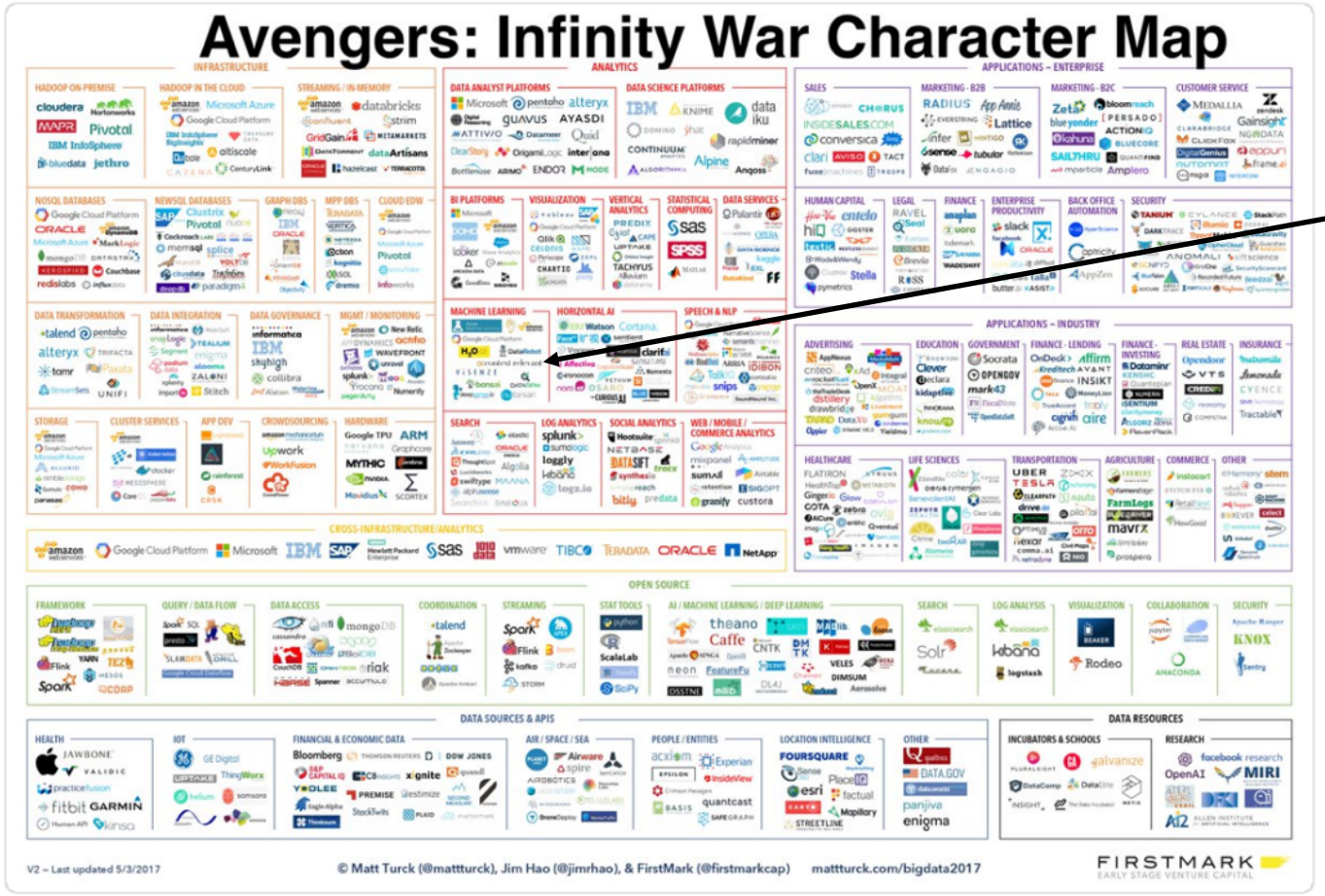
Have you seen Avengers: Infinity War?

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

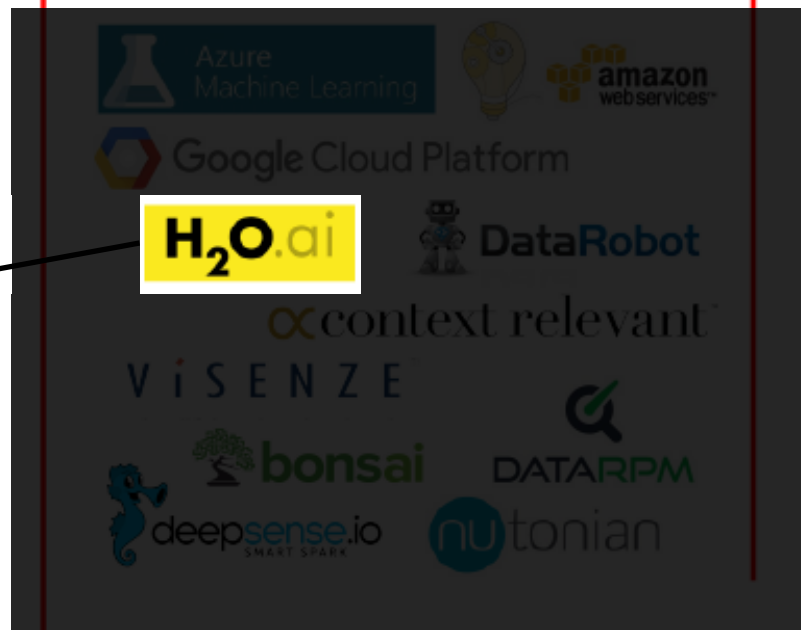
Vicki Boykis
@vboykis

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



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)

Platforms with H_2O integration

H_2O + KNIME Talk at KNIME Summit Mar 2017

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





H2O.ai

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

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

Secure multi-tenant H_2O clusters

Worldwide Recognition in the H2O.ai Community

Open Source Community

222 OF FORTUNE 500

H2O

8 OF TOP 10 BANKS

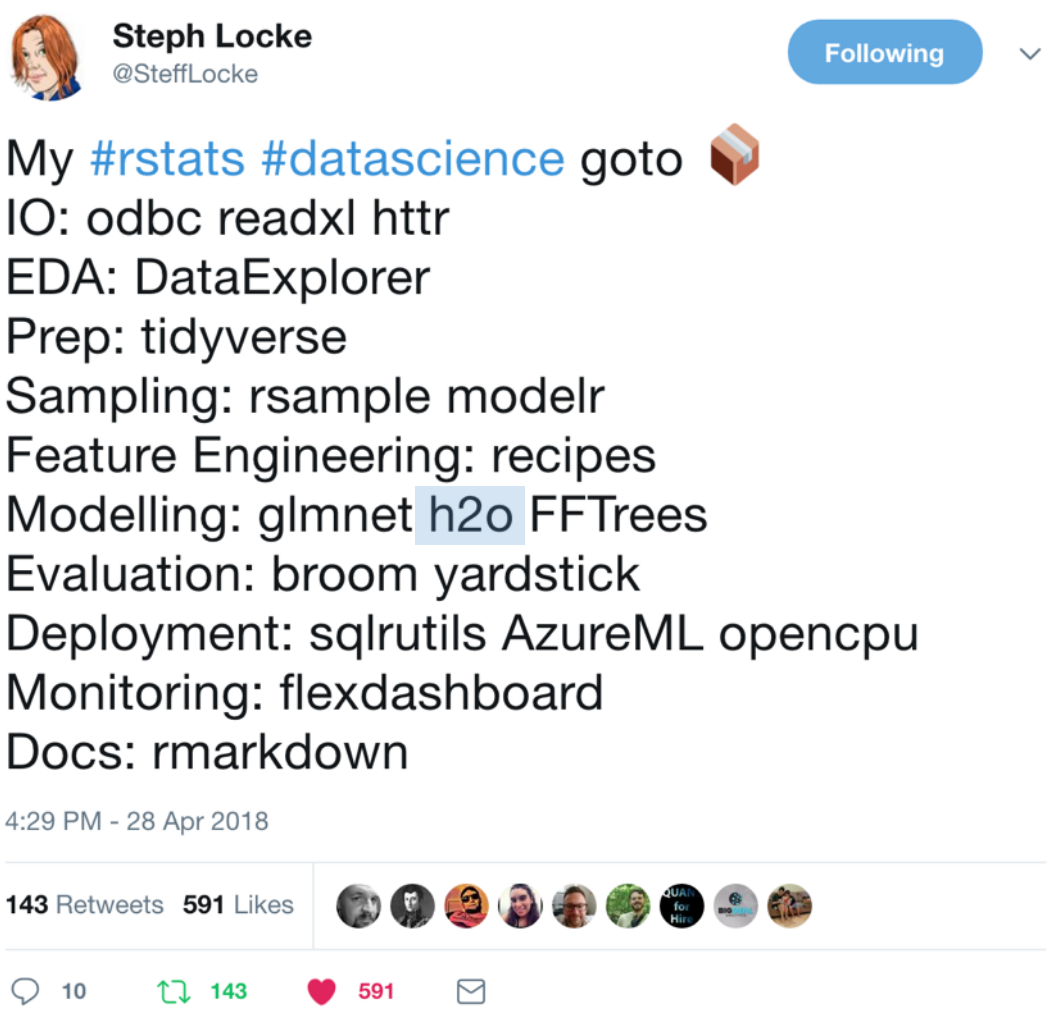
7 OF TOP 10 INSURANCE COMPANIES

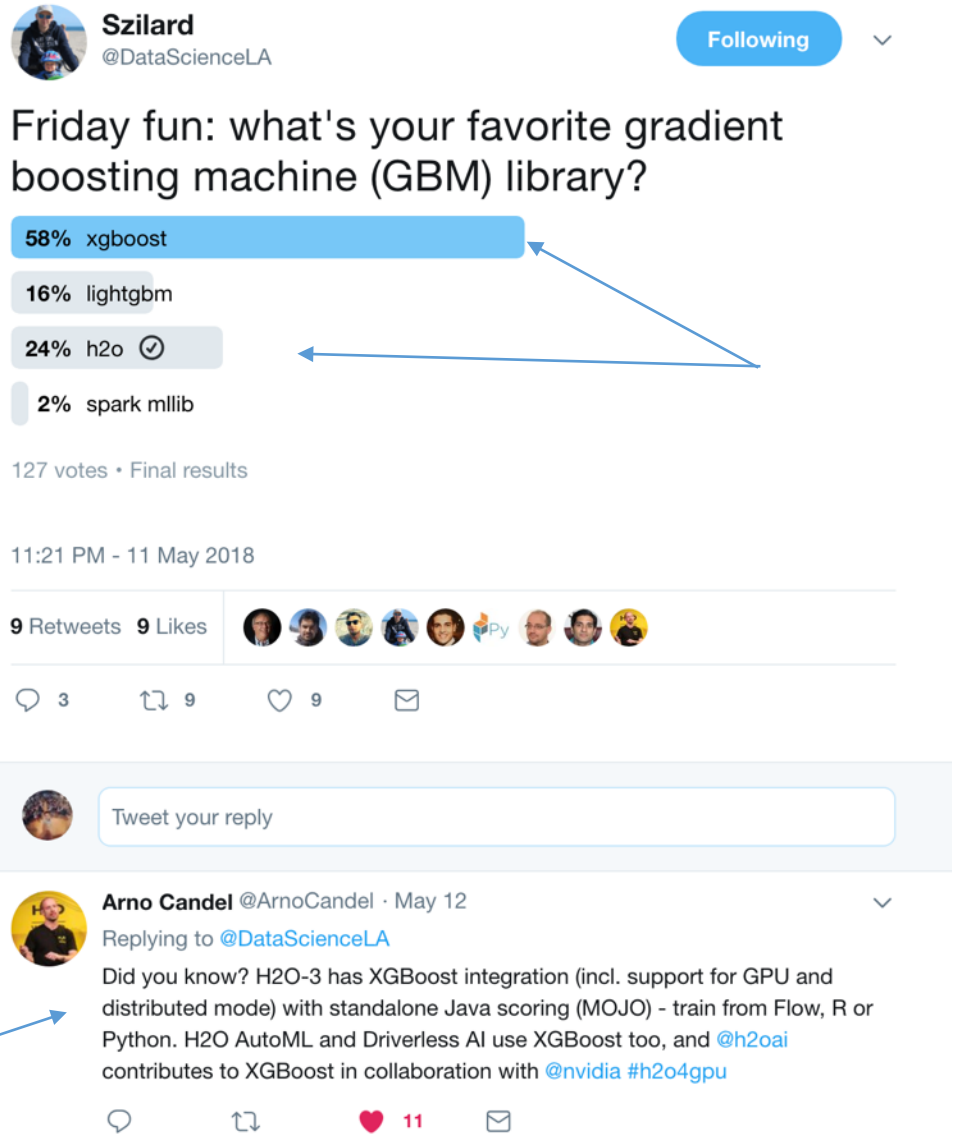
4 OF TOP 10 HEALTHCARE COMPANIES

Paying Customers

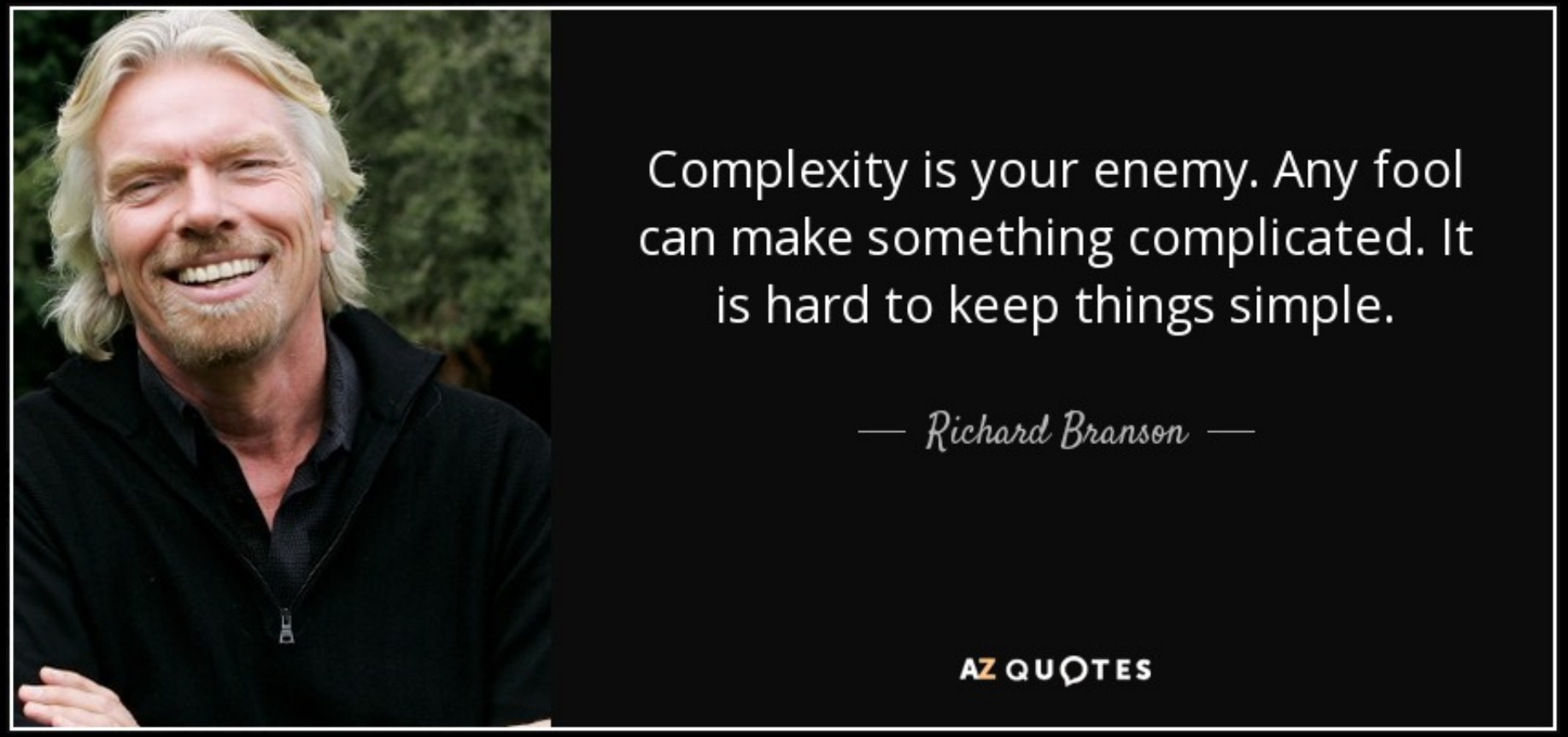
"H2O.ai's reference customers gave it the highest overall score for sales relationship and overall service and support" - Gartner MQ 2018

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

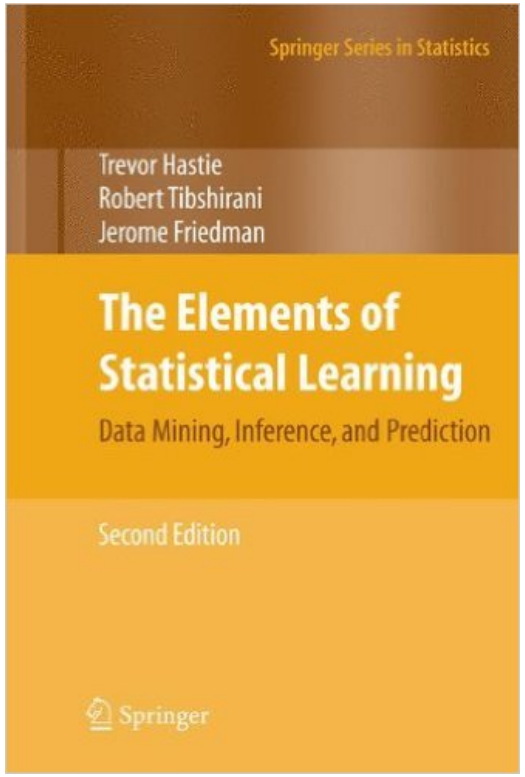
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

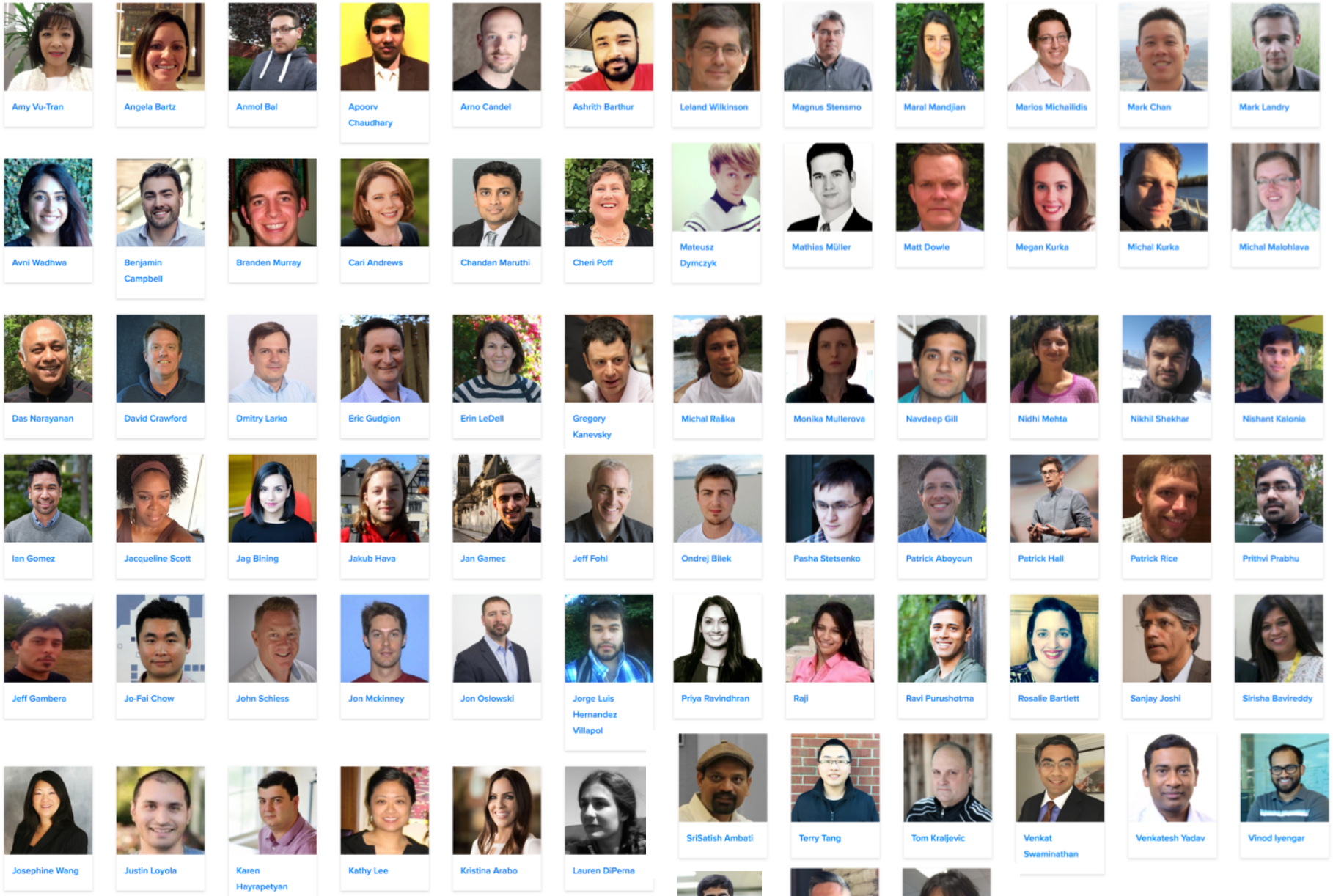
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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

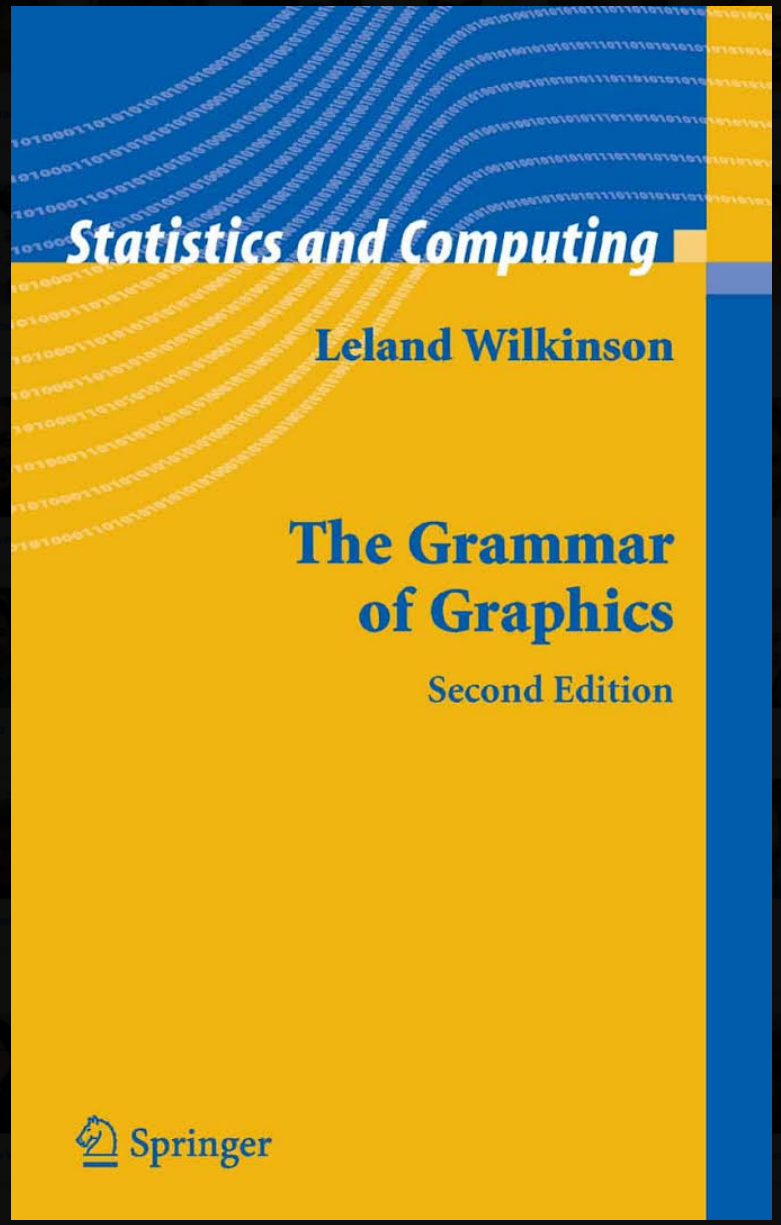


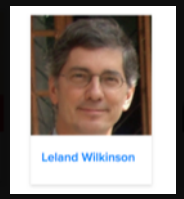
H2O.ai



H_2O Team

H_2O.ai





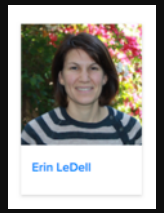
Origin of R Package `ggplot2`

Matt Dowle

data.table

H2O Team

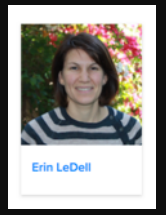
H2O.ai



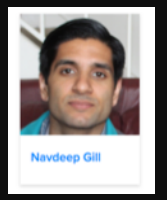
Erin LeDell, Chief ML Scientist
Women in ML/DS & R-Ladies Global

H_2O Team

H_2O AutoML



Erin LeDell



Navdeep Gill

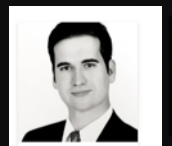
H_2O Team

48th

Branden Murray

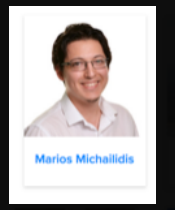
Dmitry Larko

25th



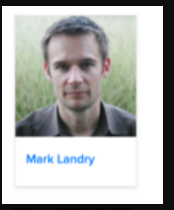
Mathias Müller

4th



Marios Michailidis

1st



Mark Landry

33rd

Kaggle Grand Masters (and their Highest Rank)

113 Grandmasters	980 Masters	3,339 Experts	46,135 Contributors	33,242 Novices

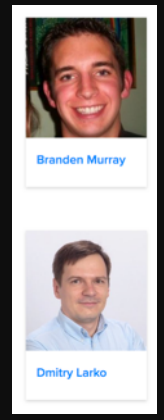
About 80,000 Kagglers

13th

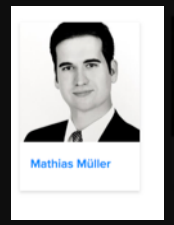
H_2O Team

H_2O.ai

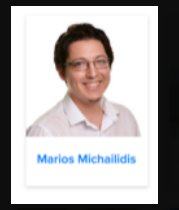
48th



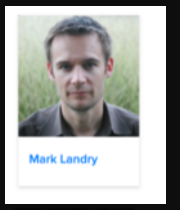
25th



4th



1st



33rd

181st

Hoping to get closer to them at some point …

13th

H_2O Team

H_2O.ai

H2O Artificial Intelligence and Machine Learning

Members	Groups	Countries
78,356	39	18

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

We sponsor meetups

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

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

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:

Women in Kaggle

Location
London, United Kingdom

Members
261

Organizers
Julia MacMillan and 3 others

You're a member

… and more

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.

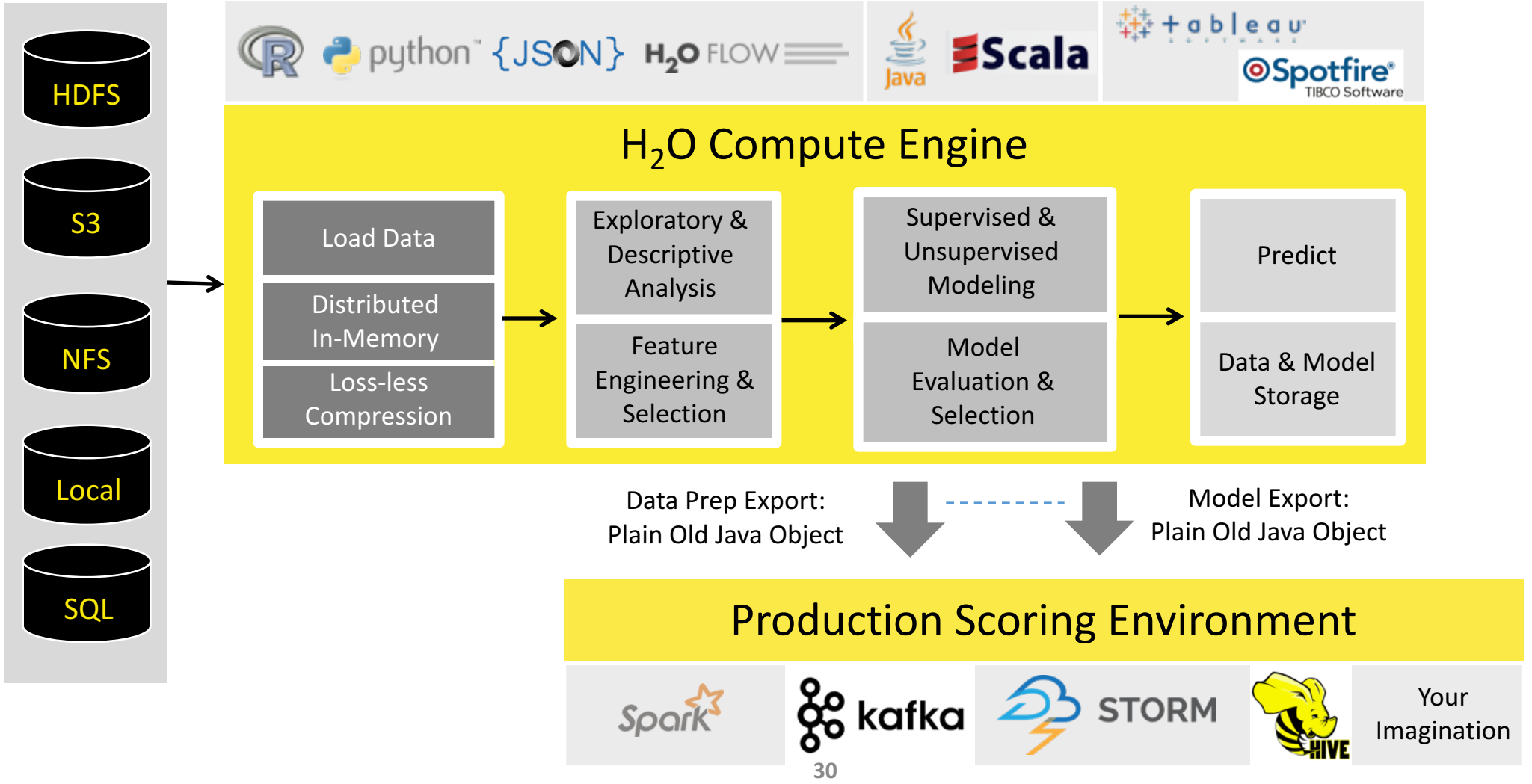
joe@h2o.ai

About 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

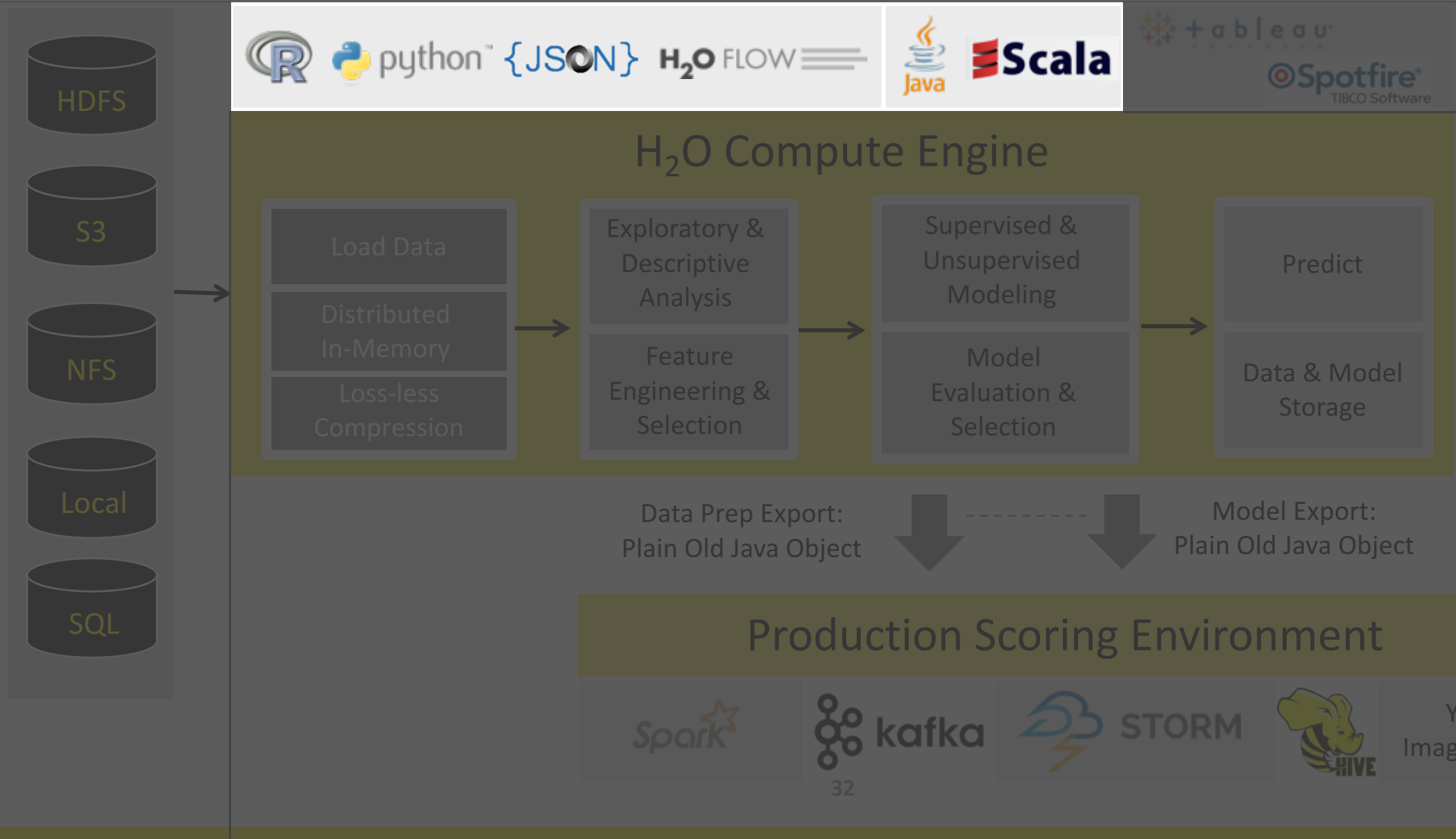
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
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

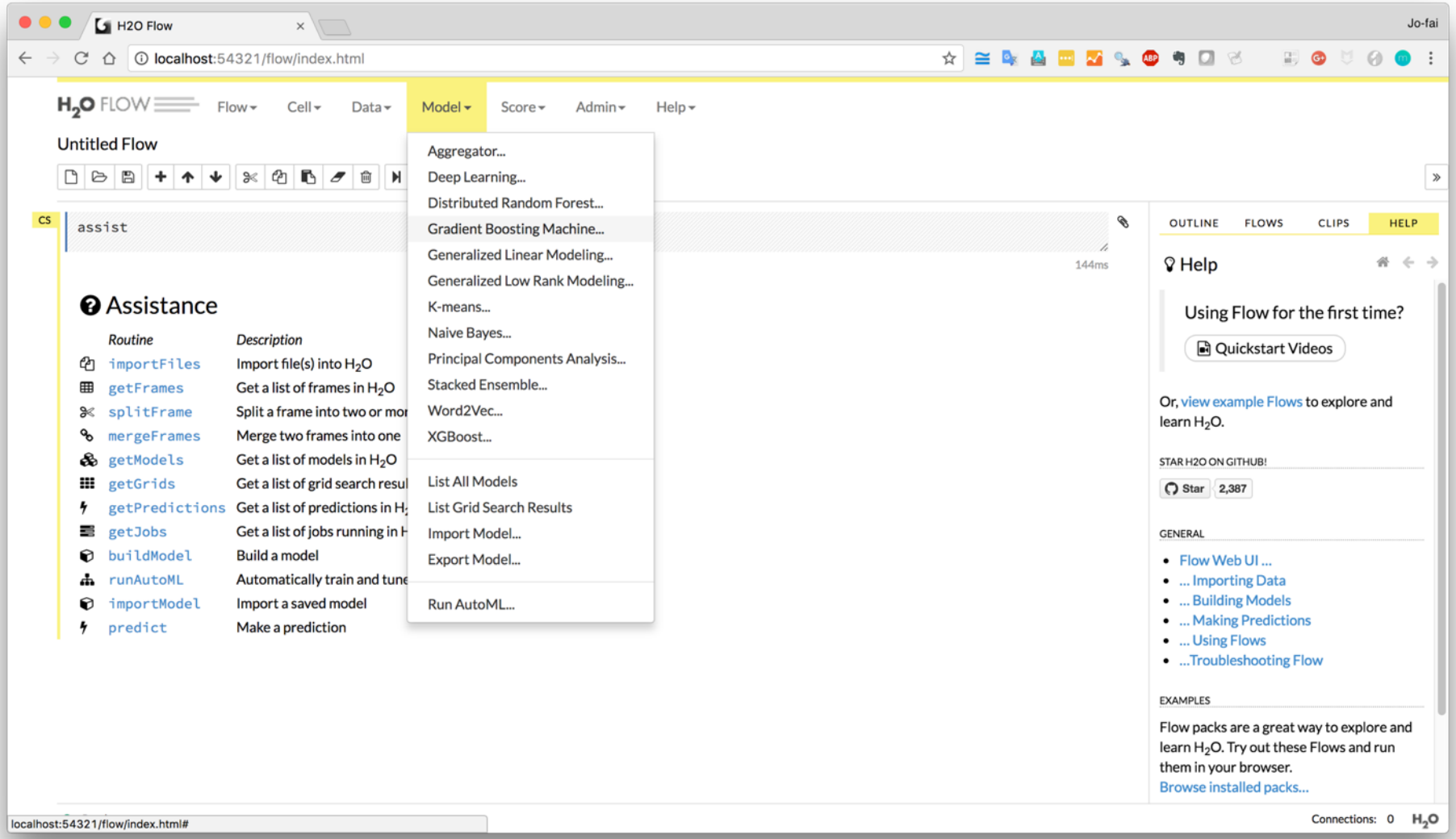
Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components • **Generalized Low Rank Models:** extend the 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

High Level Architecture

Multiple Interfaces

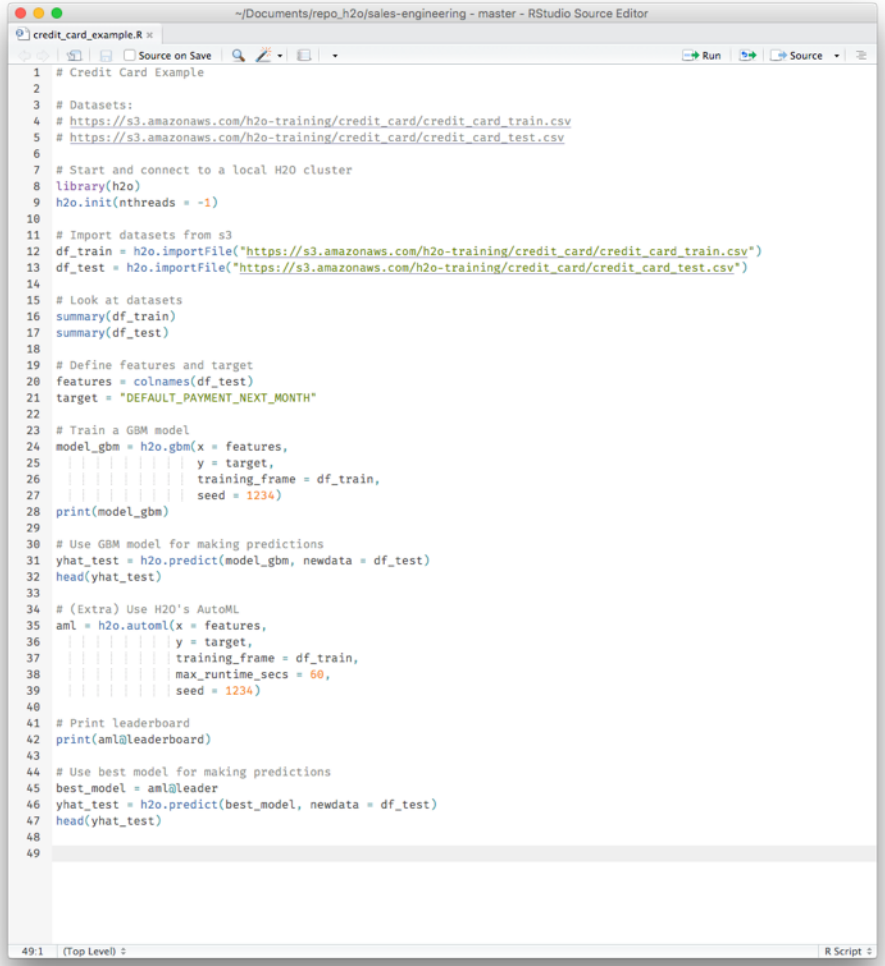


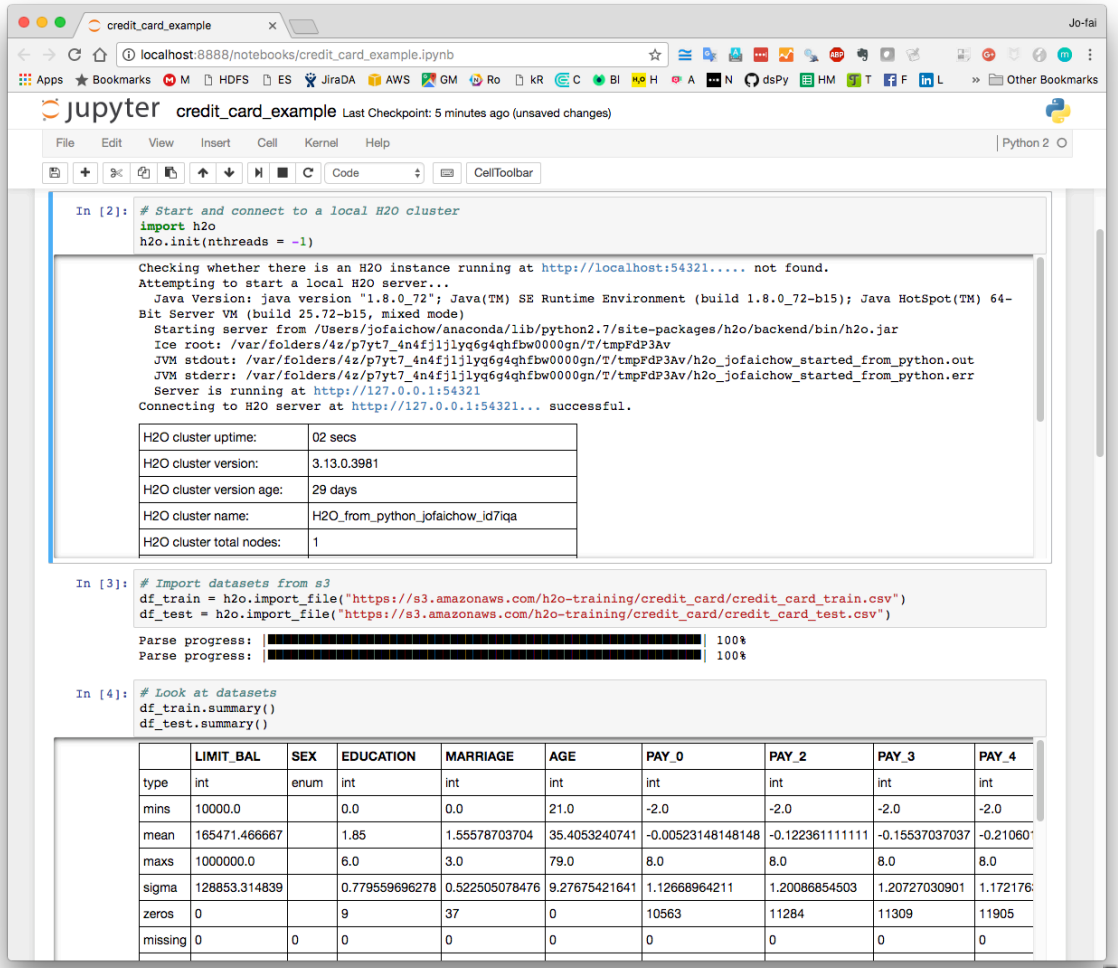
H_2O Flow (Web)



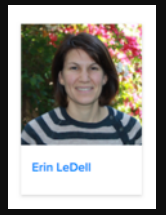
H2O.ai

Using H_2O with R and Python

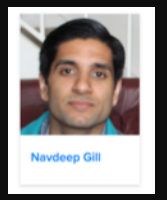




H_2O AutoML



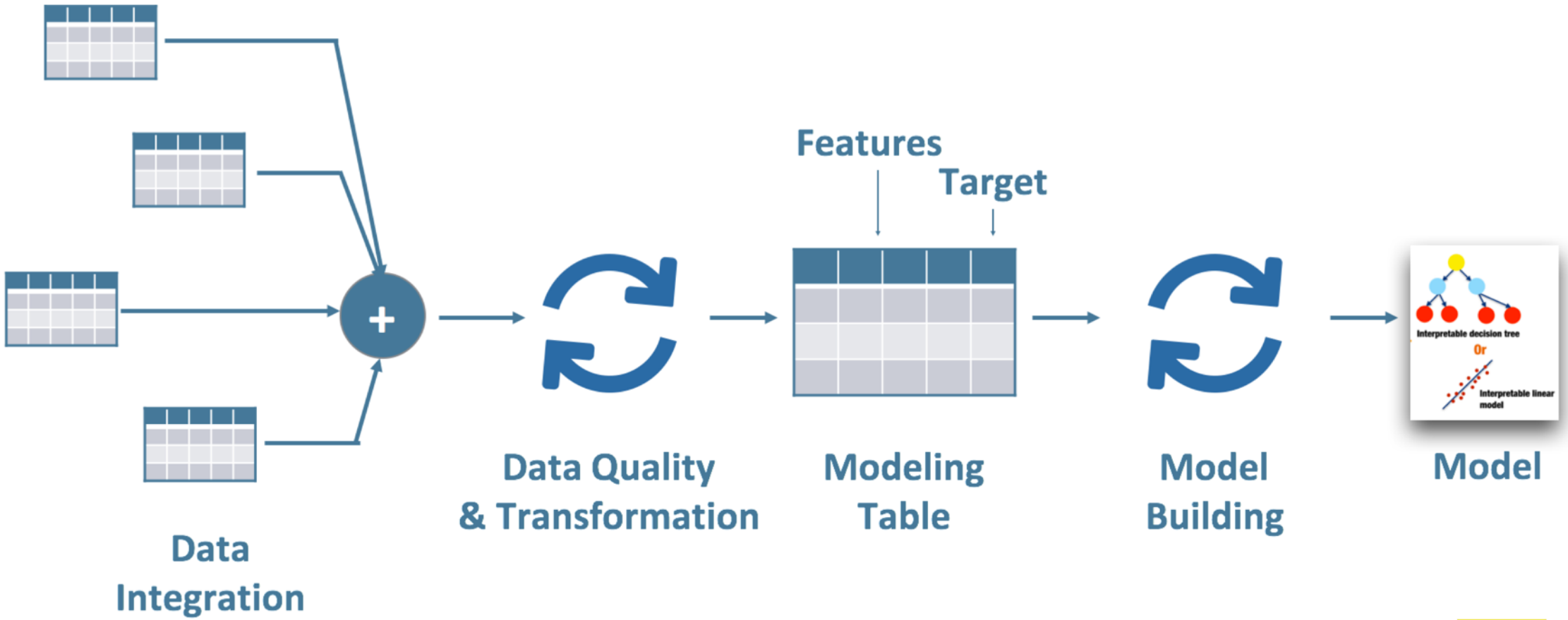
Erin LeDell



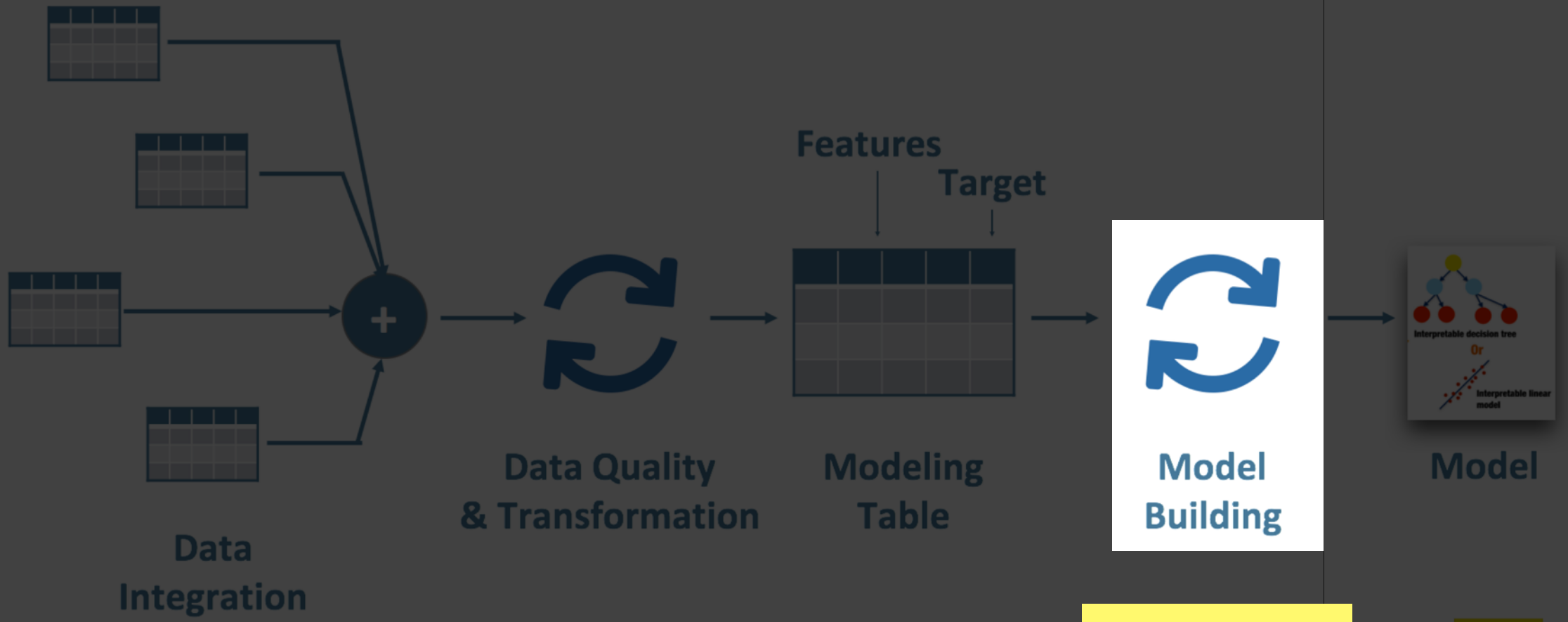
Navdeep Gill

H_2O Team

Typical Enterprise Machine Learning Workflow



Typical Enterprise Machine Learning Workflow



AutoML

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

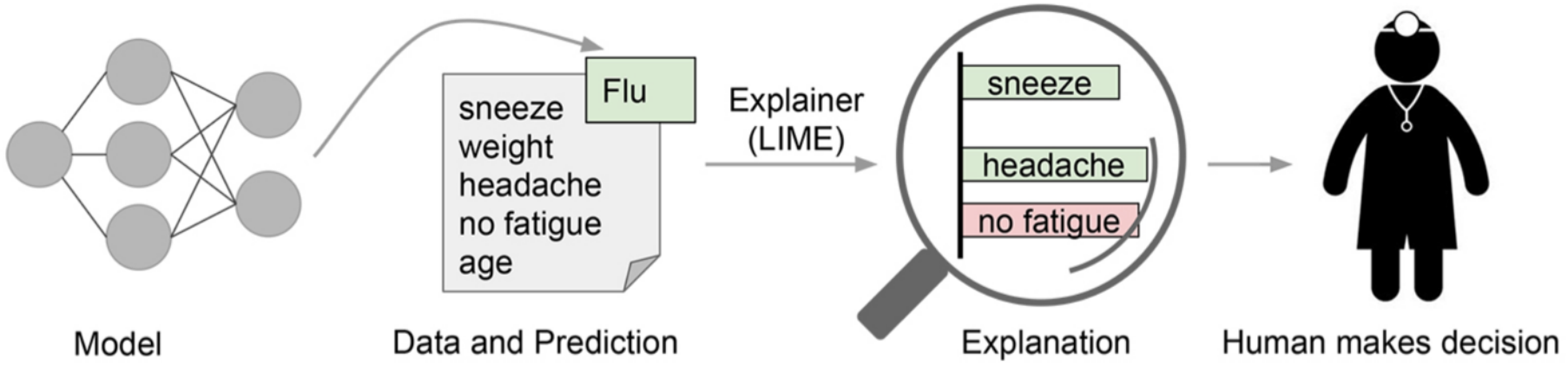


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

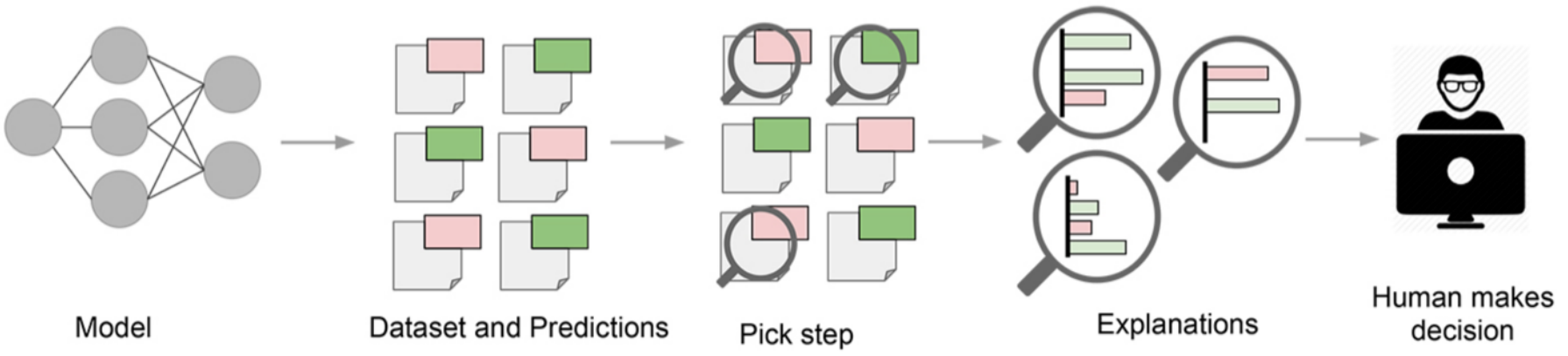
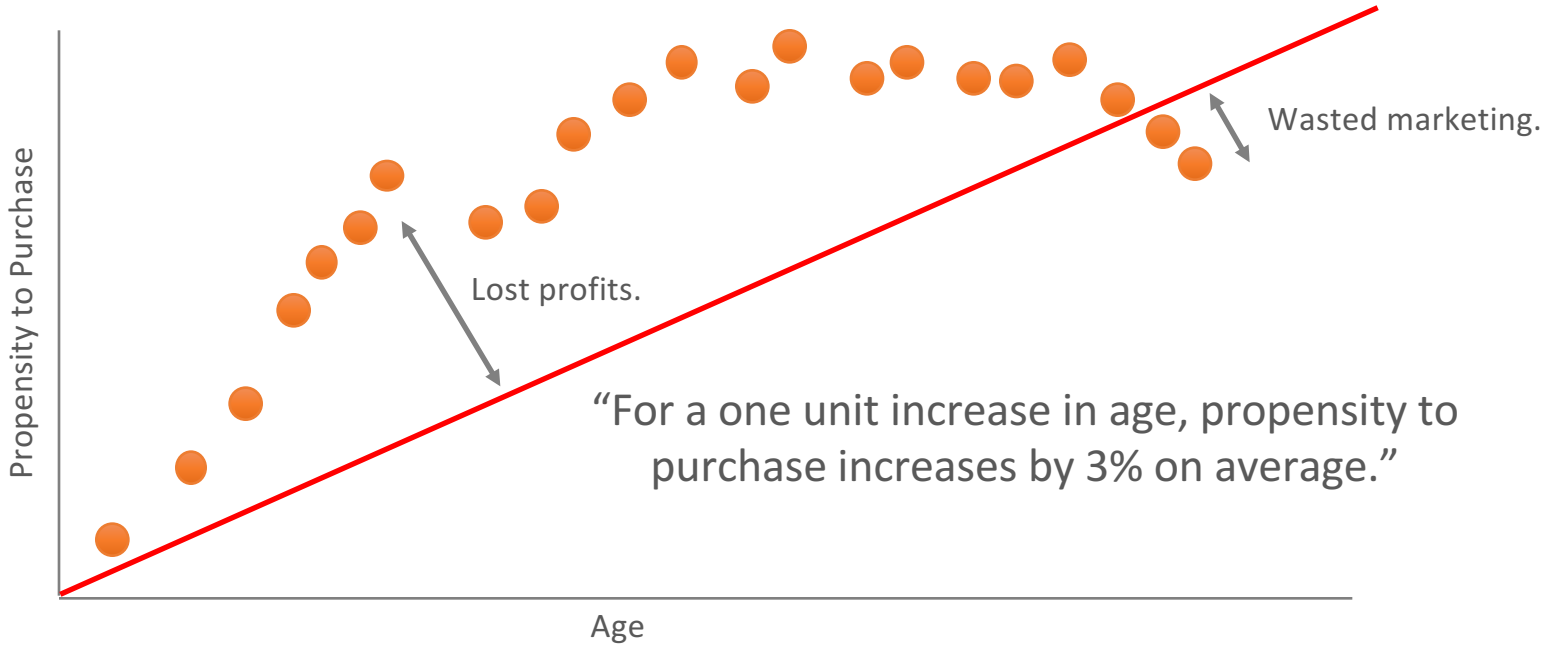


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

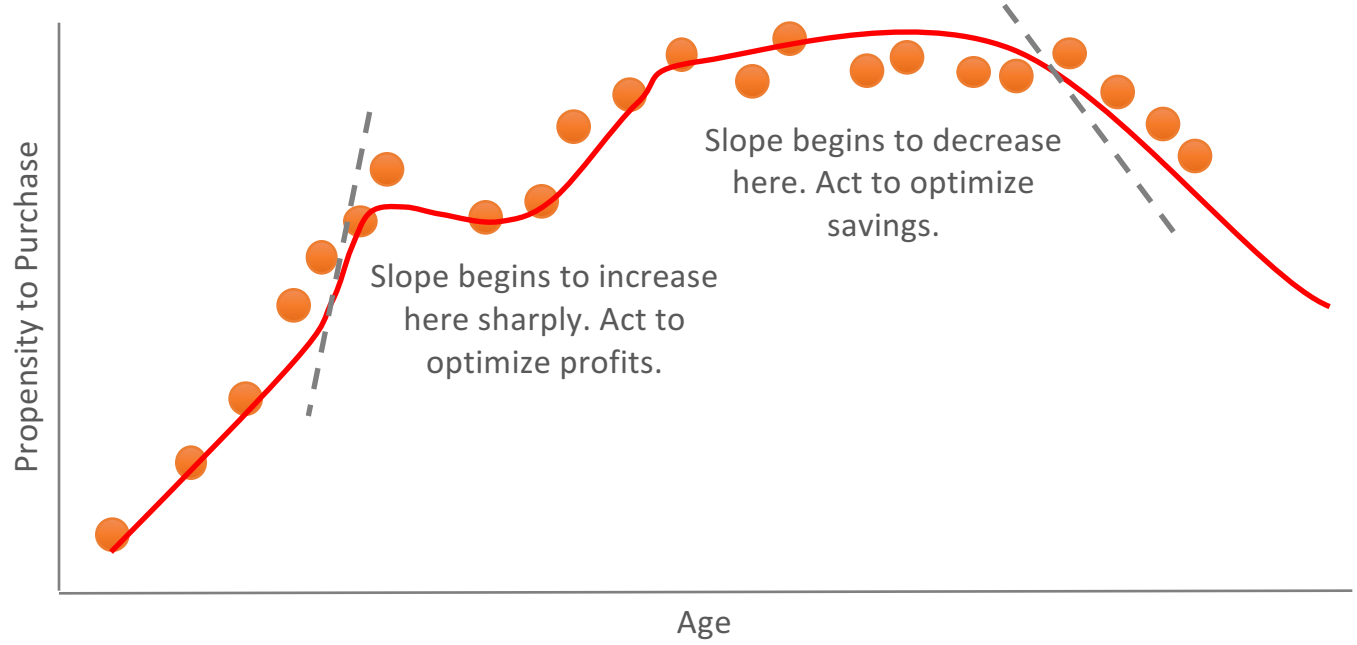
Linear Models

Exact explanations for approximate models.



Machine Learning

Approximate explanations for exact models.



H2O.ai

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

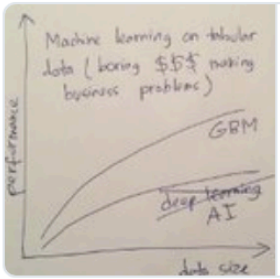
Outcome

- 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 … (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

5:40 PM - 21 Jun 2018

awesome-machine-learning-*interpretability* awesome

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

- 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

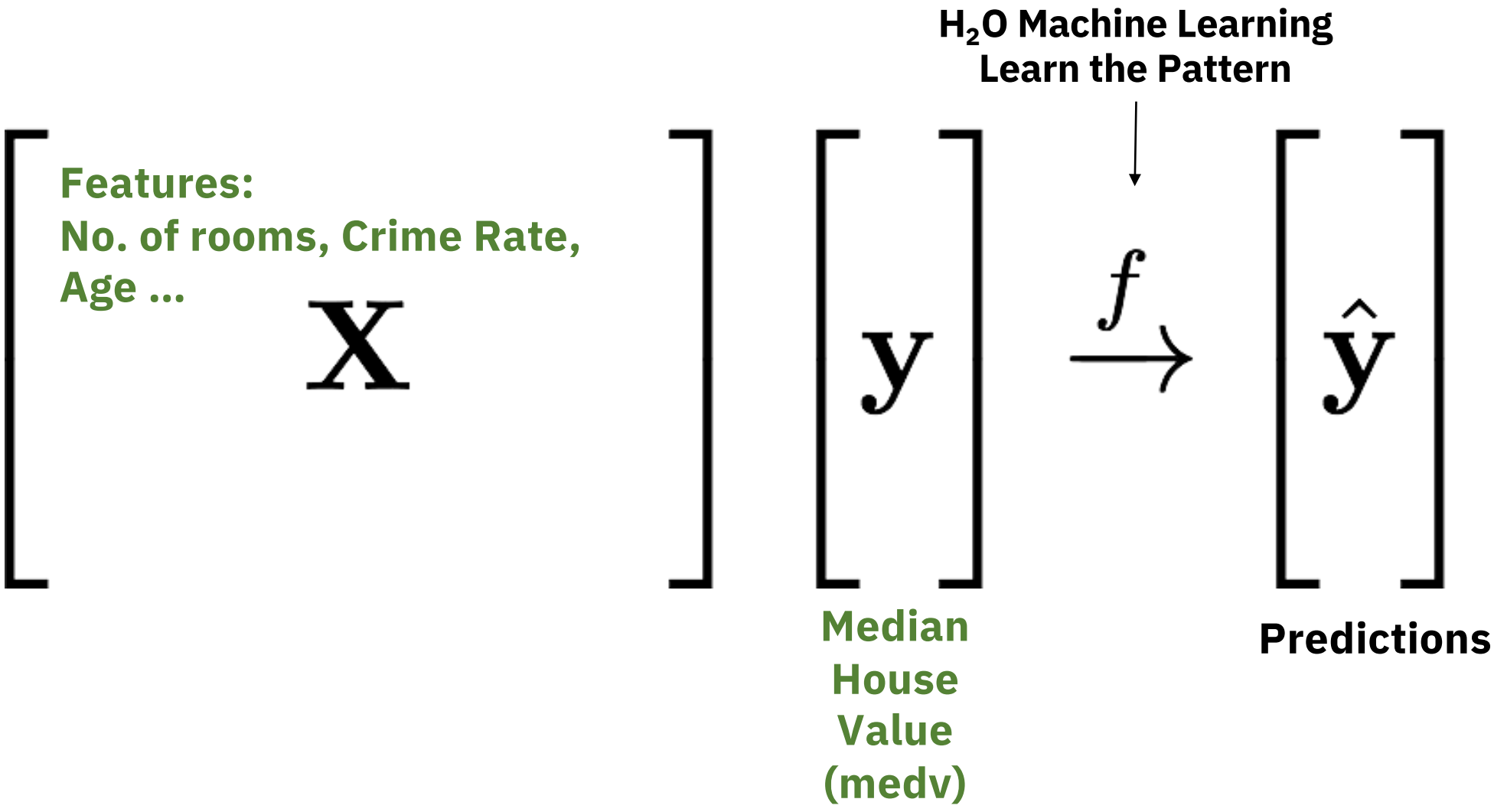
H2O.ai

Milano R

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 \examples\regression_...Rmd
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

H2O.ai

Learning from **Boston Housing** Data

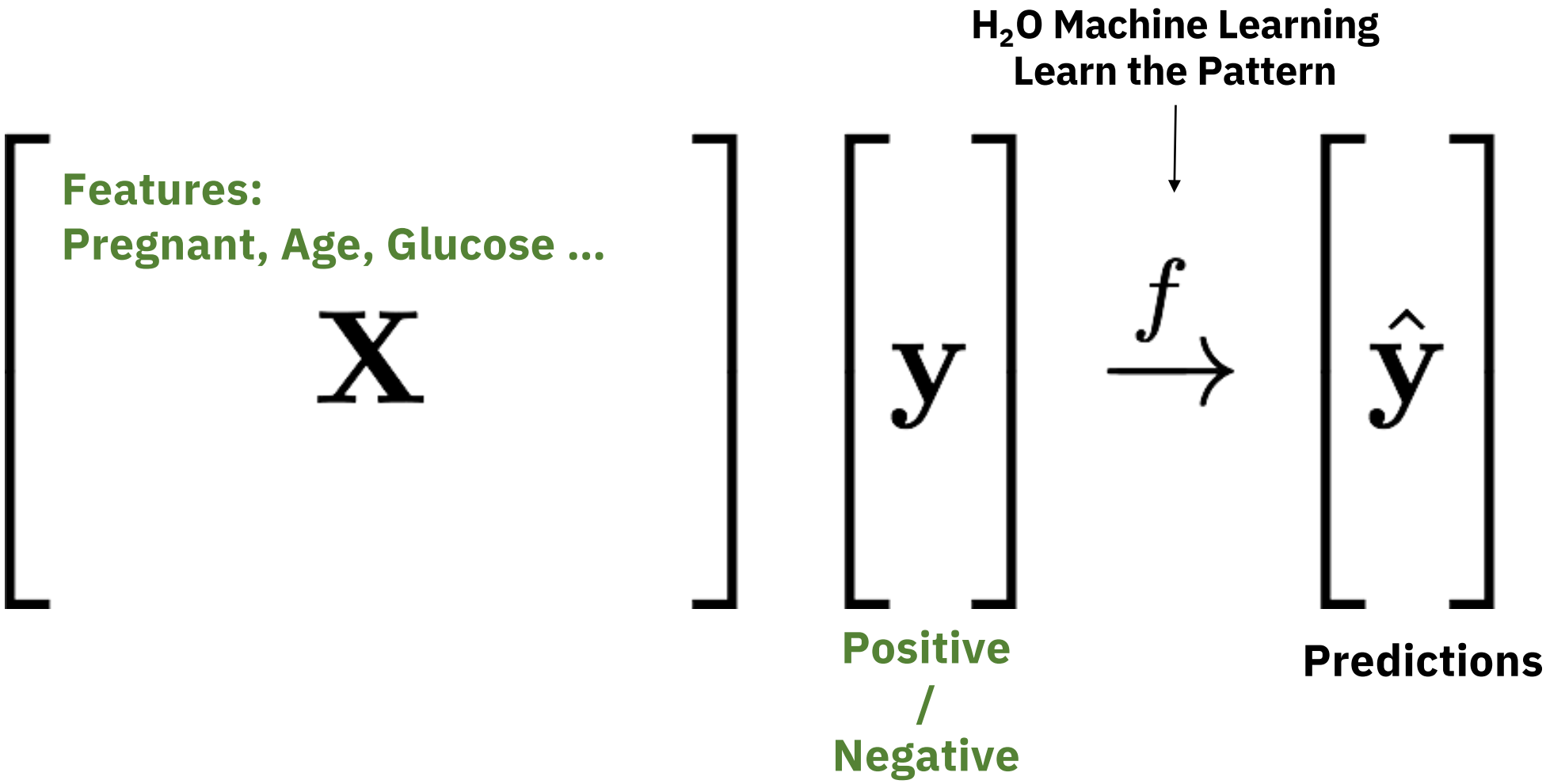


Milano R

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 \examples\classification_...Rmd
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.ai

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

H2O.ai

Milano R

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?

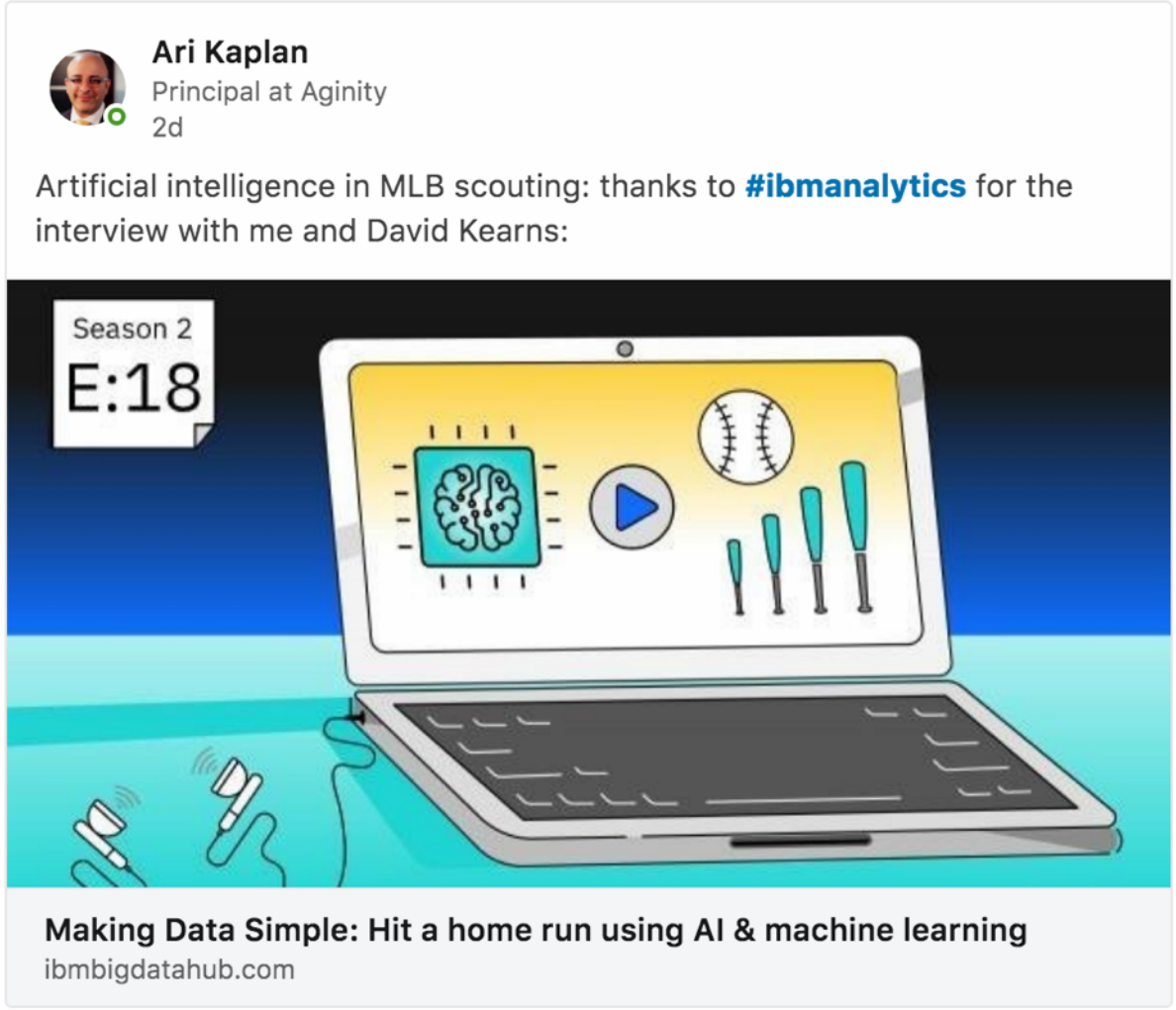
- Most users/organizations can benefit from automatic machine learning pipelines.
  - Eliminate time wasted on human errors, debugging etc.
- Model interpretations is crucial for those who must explain their models to regulators or customers.

You will learn …

- How to build high quality 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
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

SlideShare [Link]



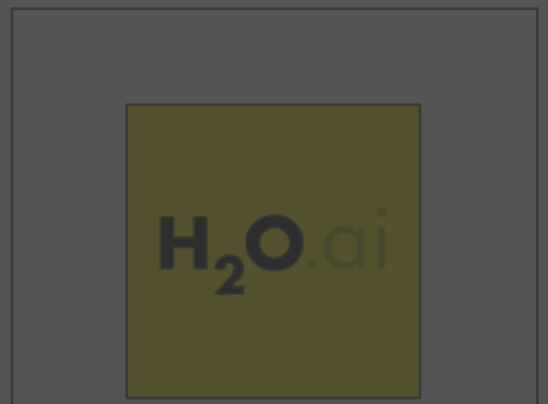
Podcast [Link]

Milano R

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

H2O.ai

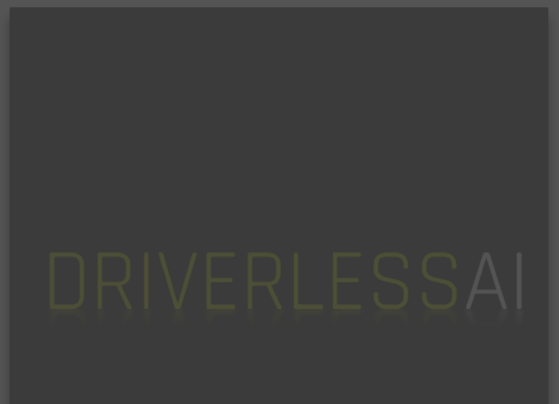
H_2O Products



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

H2O AI Open Source Engine Integration with Spark

Lightning Fast machine learning on GPUs



Automatic feature engineering, machine learning and interpretability

Secure multi-tenant H2O clusters

H2O.ai

Algorithms on H_2O-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

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.

blog.h2o.ai

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.

Coming Up

Date	City	Talk / Workshop
25 June	Milan	H_2O + LIME Workshop
28 June	Rome	H_2O Intro + Use Cases
Mid-July	London	Next London Meetup (T.B.C.)
Mid-Oct	London	H2O World London
13-14 Nov	London	Big Data LDN Keynote by Sri (CEO) + Meetup

Thanks!

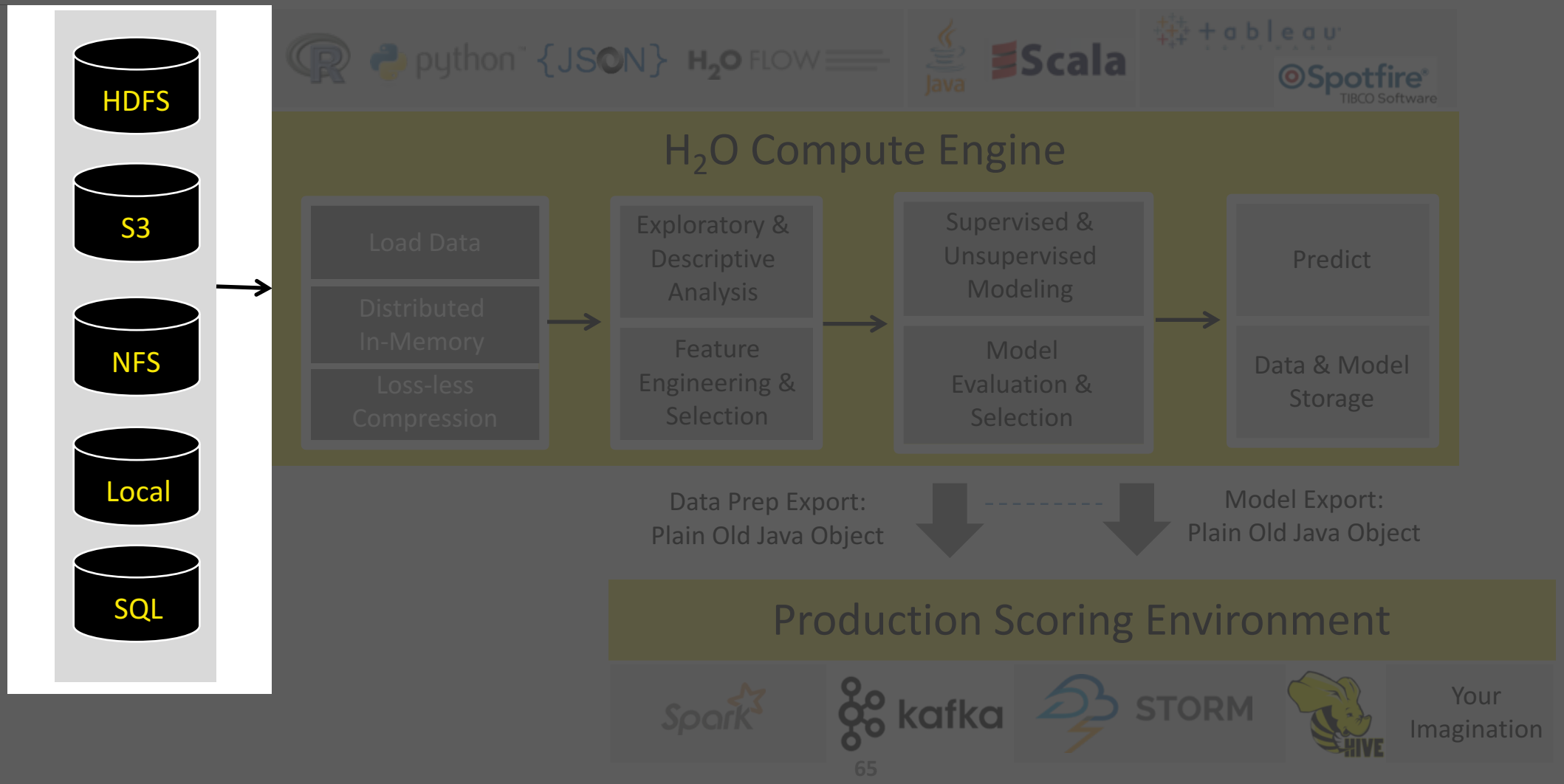
- Organizers & Sponsors

- Code, Slides & Documents
  - bit.ly/h2o_meetups
  - bit.ly/joe_eRum_2018
  - docs.h2o.ai
- Contact
  - joe@h2o.ai
  - @matlabulous
  - 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

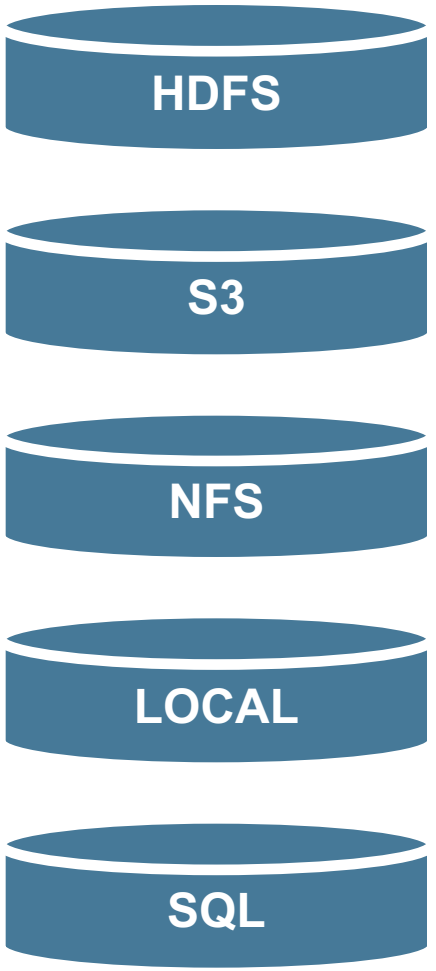


Supported Formats & Data Sources

9 Formats

File type or Folder of Files

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



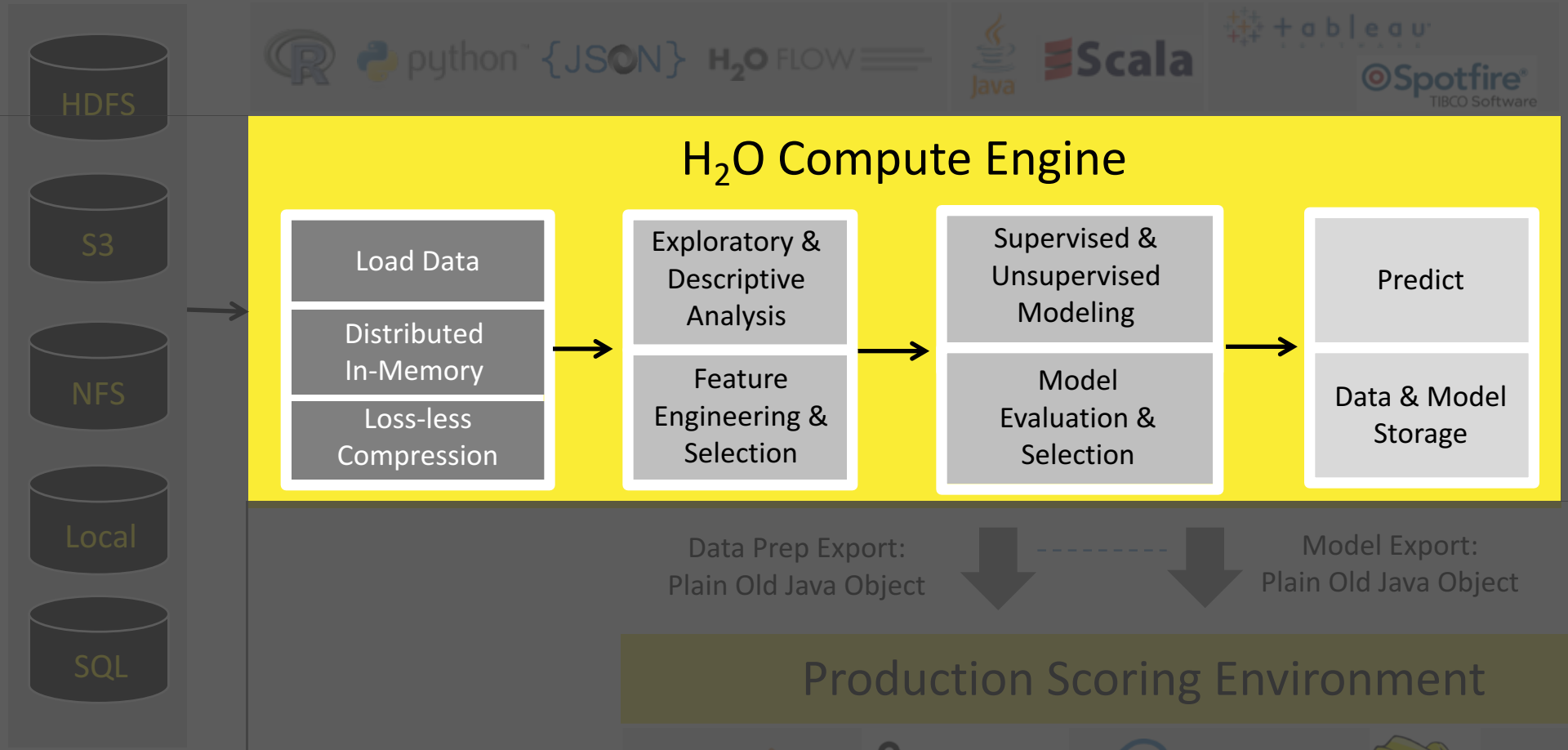
5 Sources

* 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

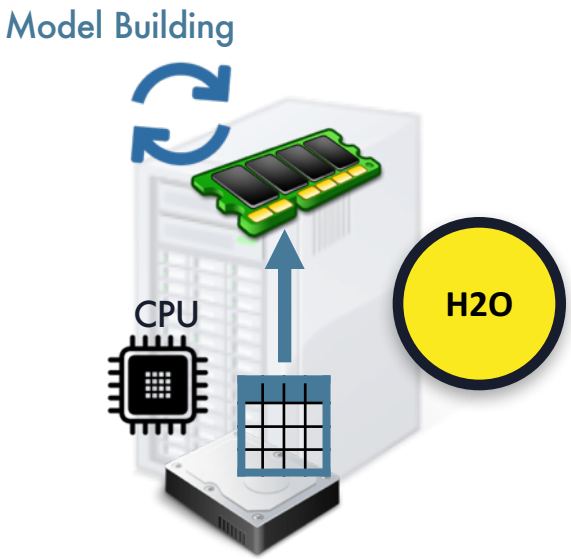
H2O.ai

High Level Architecture

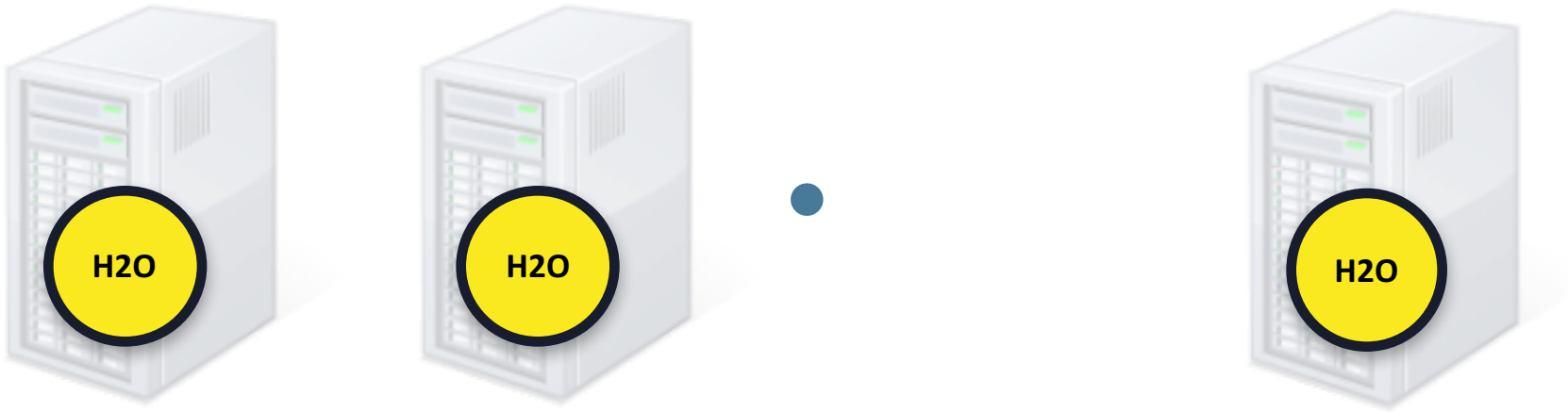
Fast, Scalable & Distributed Compute Engine Written in Java



H_2O Core

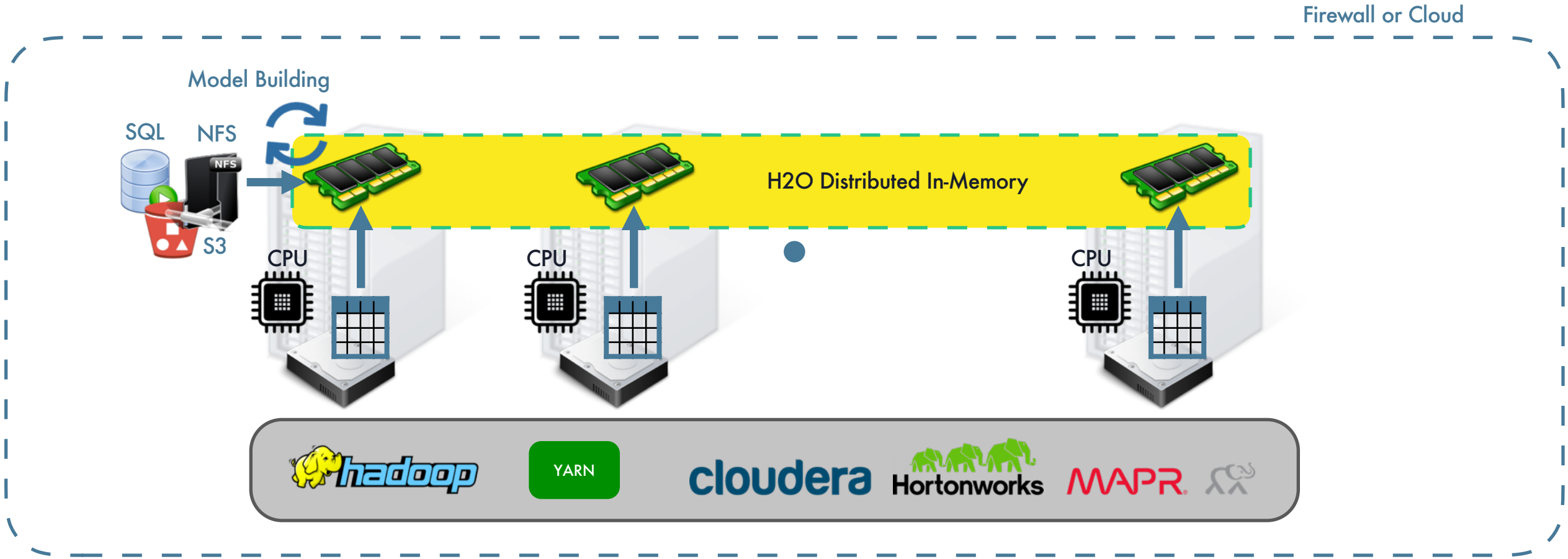


H_2O Core



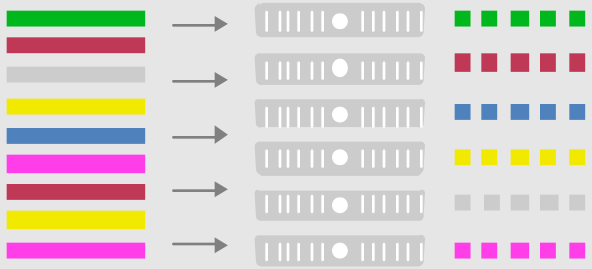
H_2O Core

H_2O Core

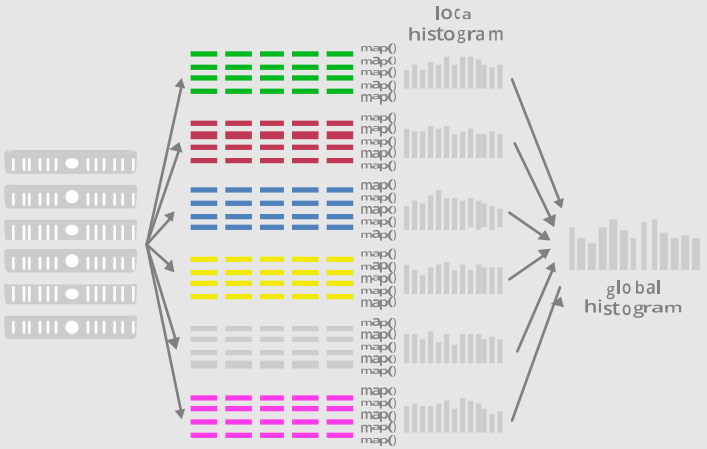

Firewall or Cloud
Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



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

Advantageous Foundation

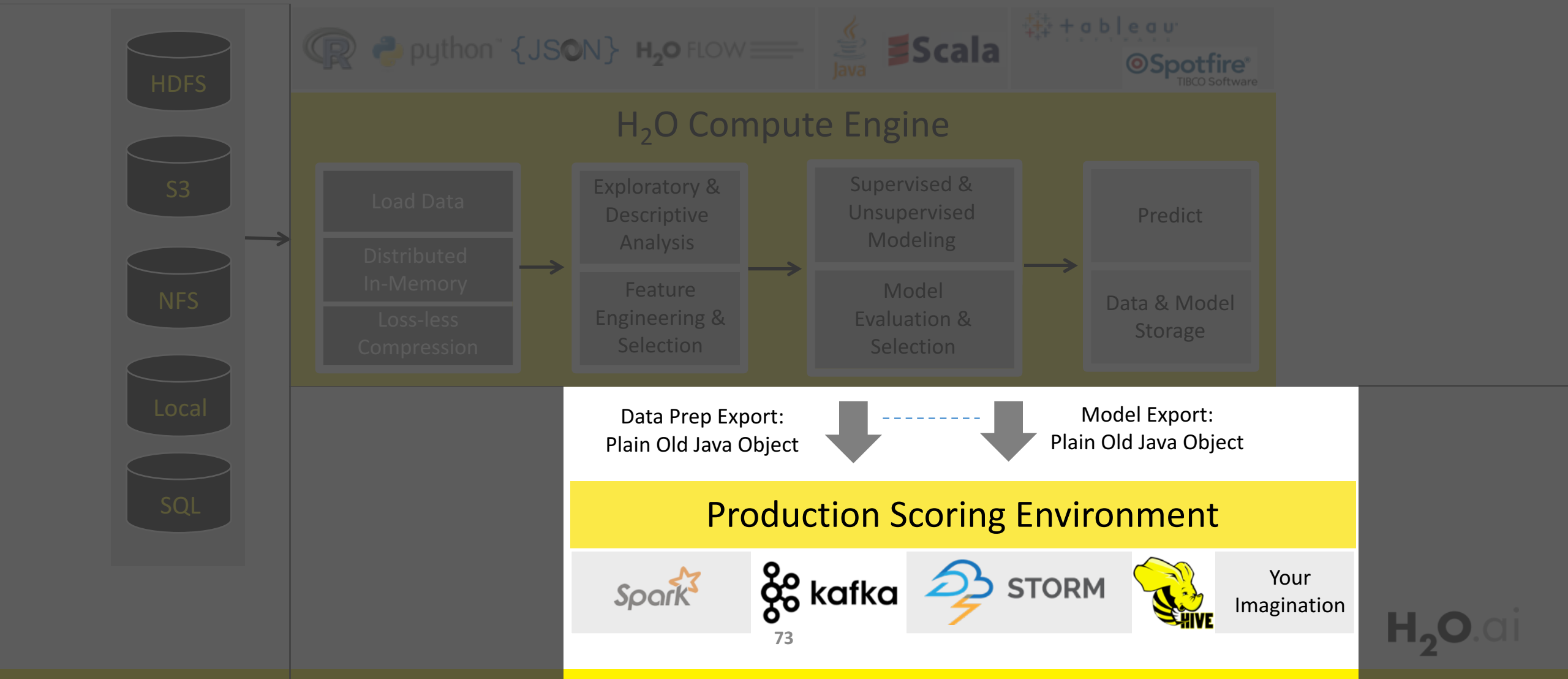
- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in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

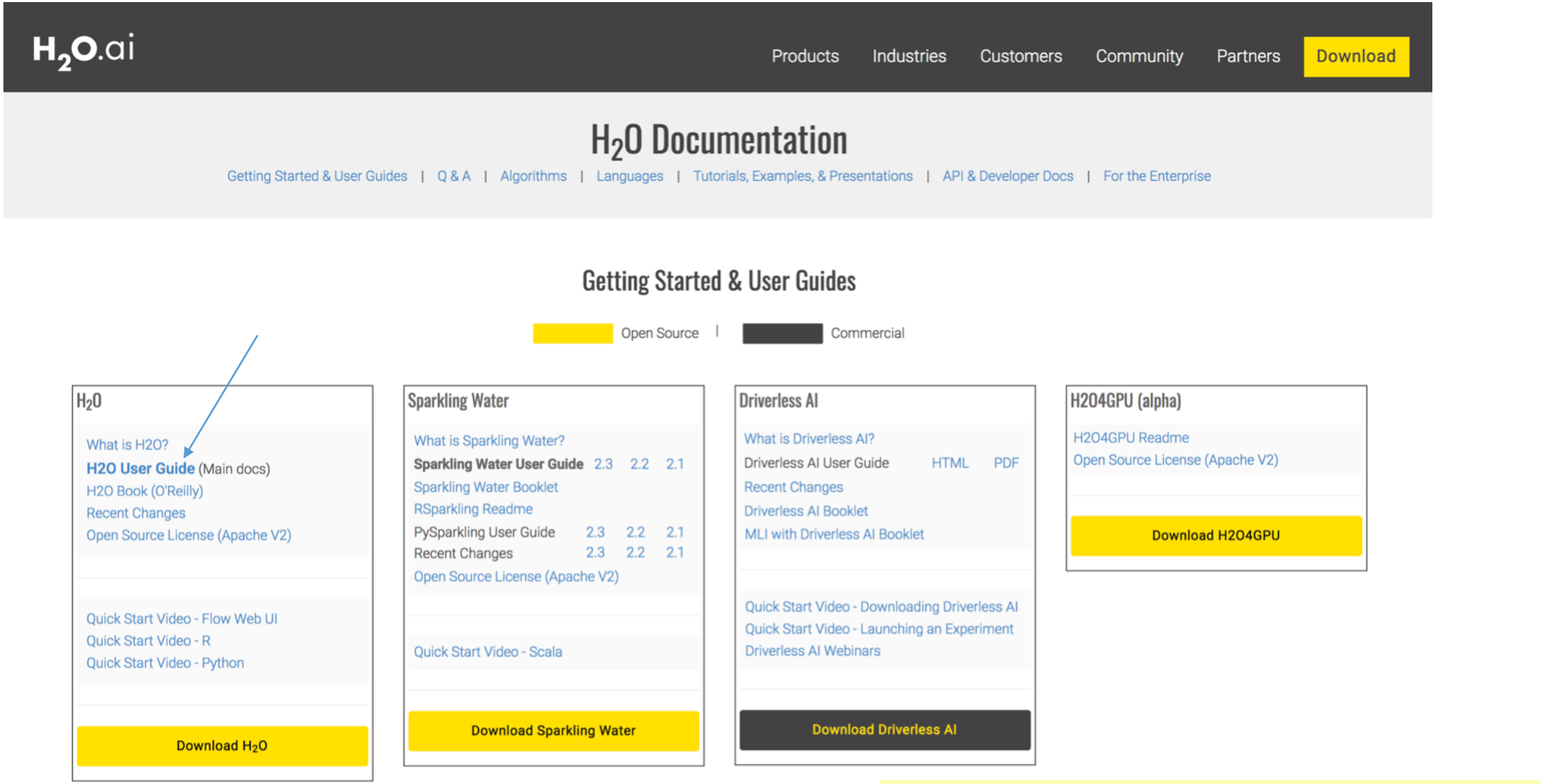
User Benefits

- "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





URL: **docs.h2o.ai**

H2O.ai

Demo: H_2O on a 320-Core Hadoop Cluster

(Web Interface)

Higgs challenge

Higgs Boson Machine Learning Challenge

Use the ATLAS experiment to identify the Higgs boson

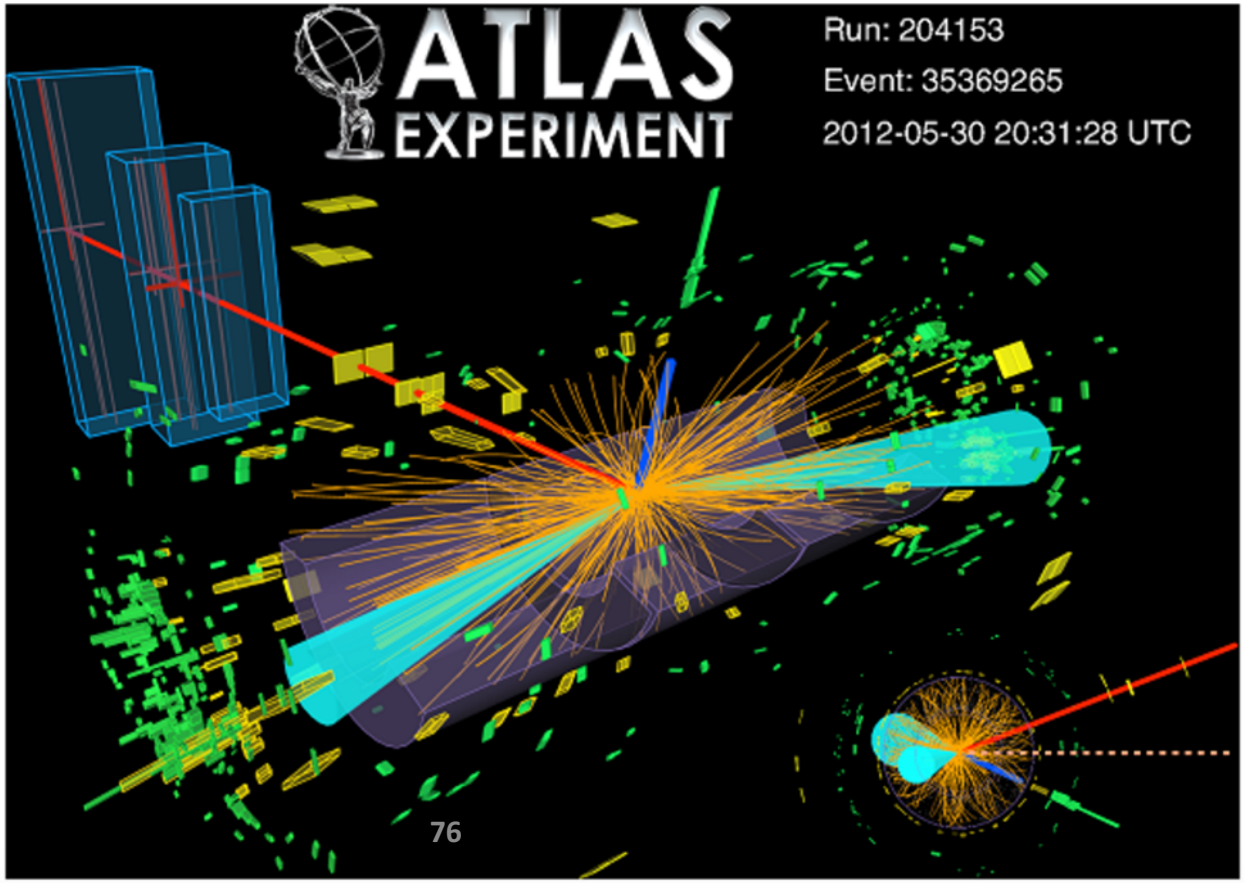
$13,000 · 1,785 teams · 3 years ago

Overview | Data | Discussion | Leaderboard | Rules | Team | My Submissions | Late Submission

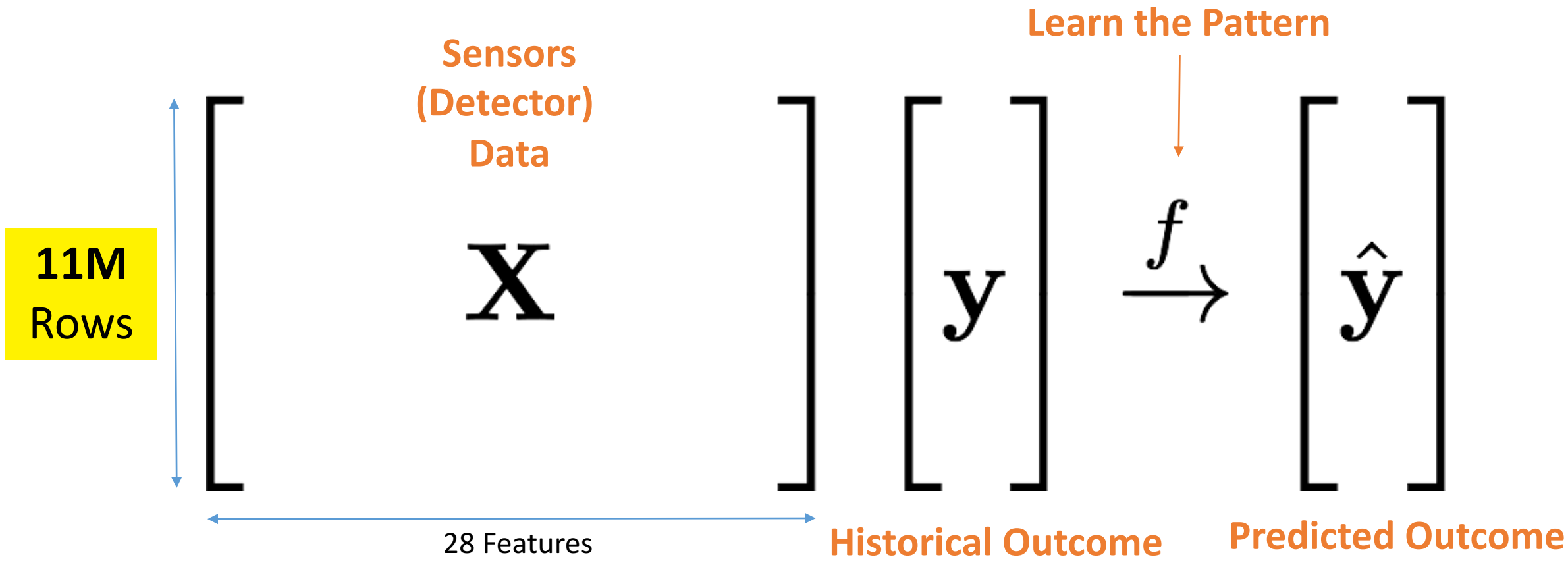
https://www.kaggle.com/c/higgs-boson

Overview

- Description
- Evaluation
- Prizes
- About The Sponsors
- Timeline
- Winners



Learning from Higgs Boson Machine Data



Raw Data Size: **7.48 GB**

11M Rows

Size (Raw): **7.48 GB**
Compressed: **2.00 GB** (≈ 27% of Raw)

HIGGS.hex

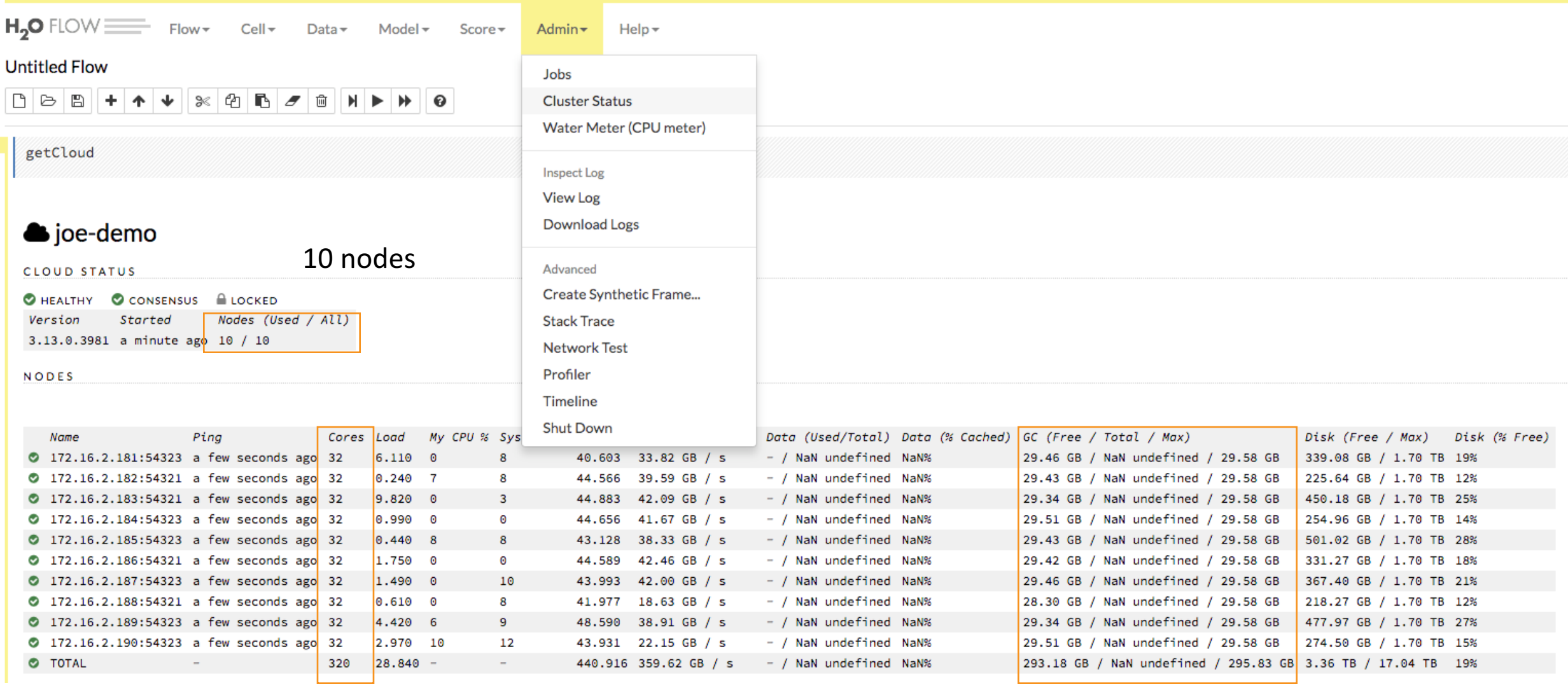
Actions: View Data | Split... | Build Model... | Predict | Download | Export

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.ai

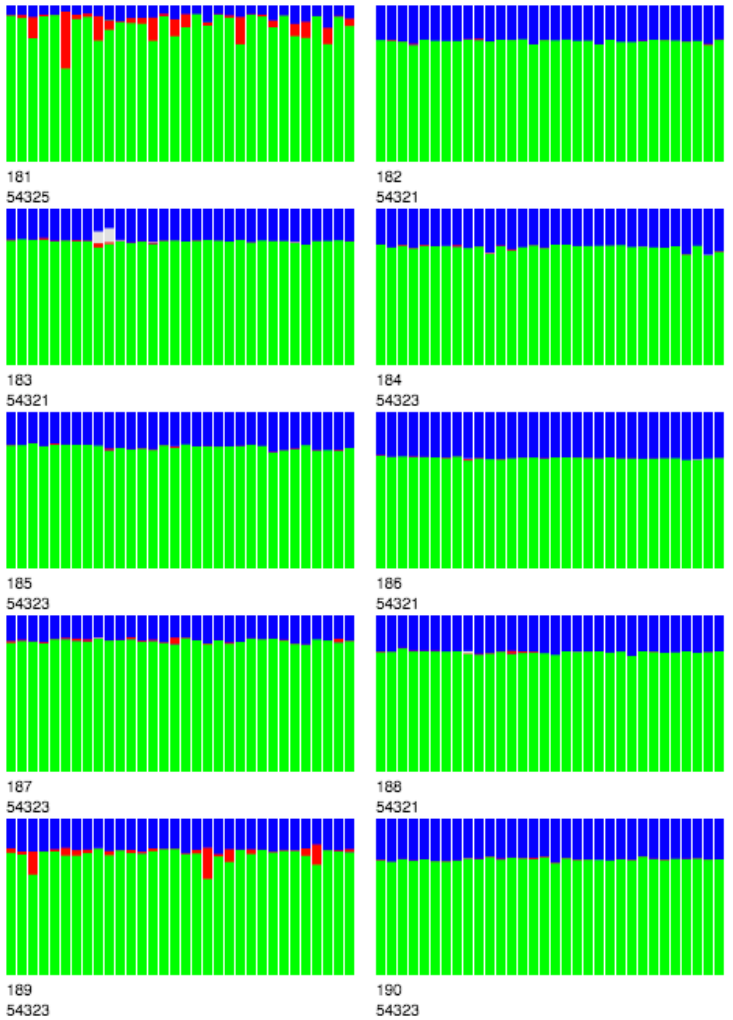


10 x 32 =
320 Cores

10 x 29.6 = 296
GB Memory

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)