Introduction to Scalable & Automatic Machine Learning with H_2O

H_2O.ai

Agenda

- Talk 1: Introduction to H_2O
  - Company and People
  - H_2O Open Source ML Platform
  - Live Demos
- Talk 2: The Next Generation of Machine Learning on Apache Spark
  - Introduction to Sparkling Water
  - Live Demos
- Talk 3: Introduction to Driverless AI
  - Introduction to Driverless AI

About Me

- Graduated from Charles University in Prague (June 2017)
- Bachelor degree from Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics

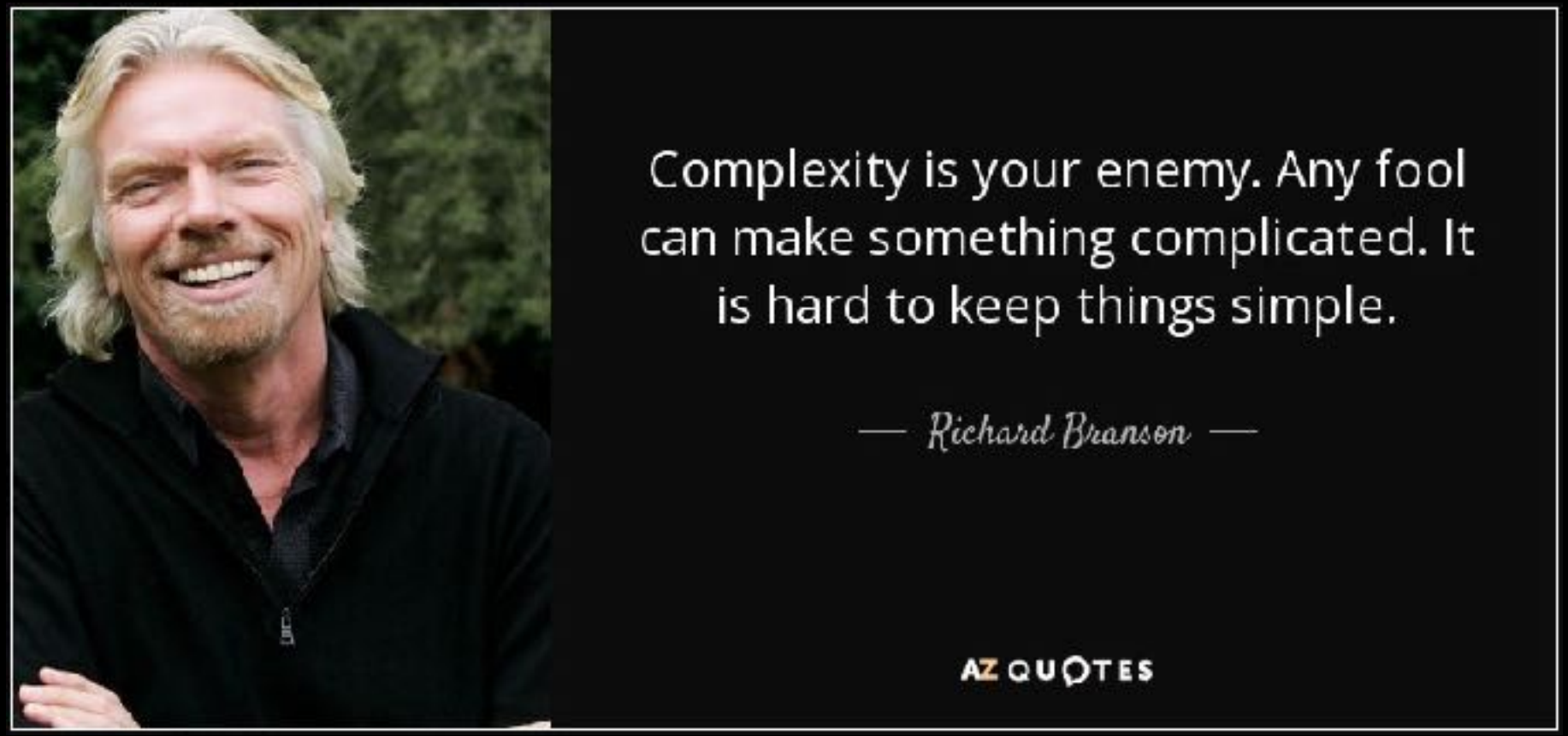
- Software Engineer
  - May 2015 - June 2017
    - Ruby on Rails and Android developer in min60 s.r.o.
  - June 2014 - February 2015
    - Ruby on Rails developer in Staffino s.r.o.
  - June 2017 - Present
    - H_2O.ai (working remotely from Bratislava)

Company Overview

Founded	2012, Series C in Nov, 2017
Products	• Driverless AI – Automated Machine Learning • H_2O - Open Source Machine Learning Platform • H2O4GPU - Lightning Fast Machine Learning on GPUs • Sparkling Water - Integration of H_2O and Apache Spark
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.ai

Our Mission:
Make Machine Learning Accessible to Everyone



Scientific Advisory Council

Dr. Trevor Hastie

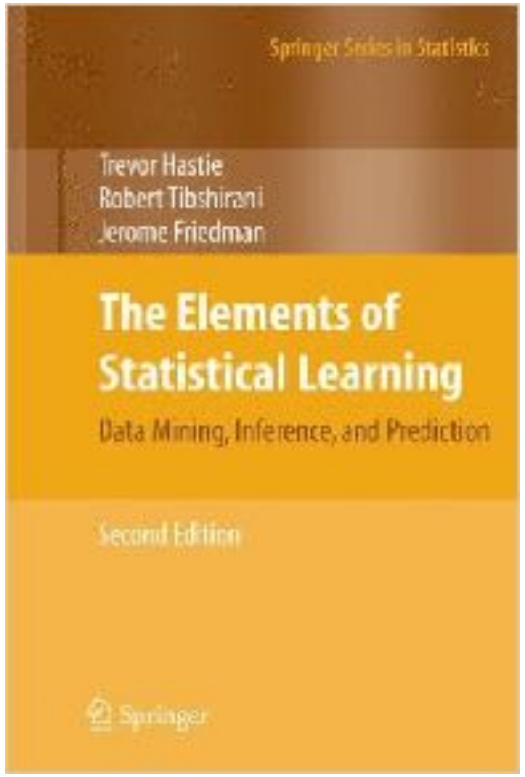
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

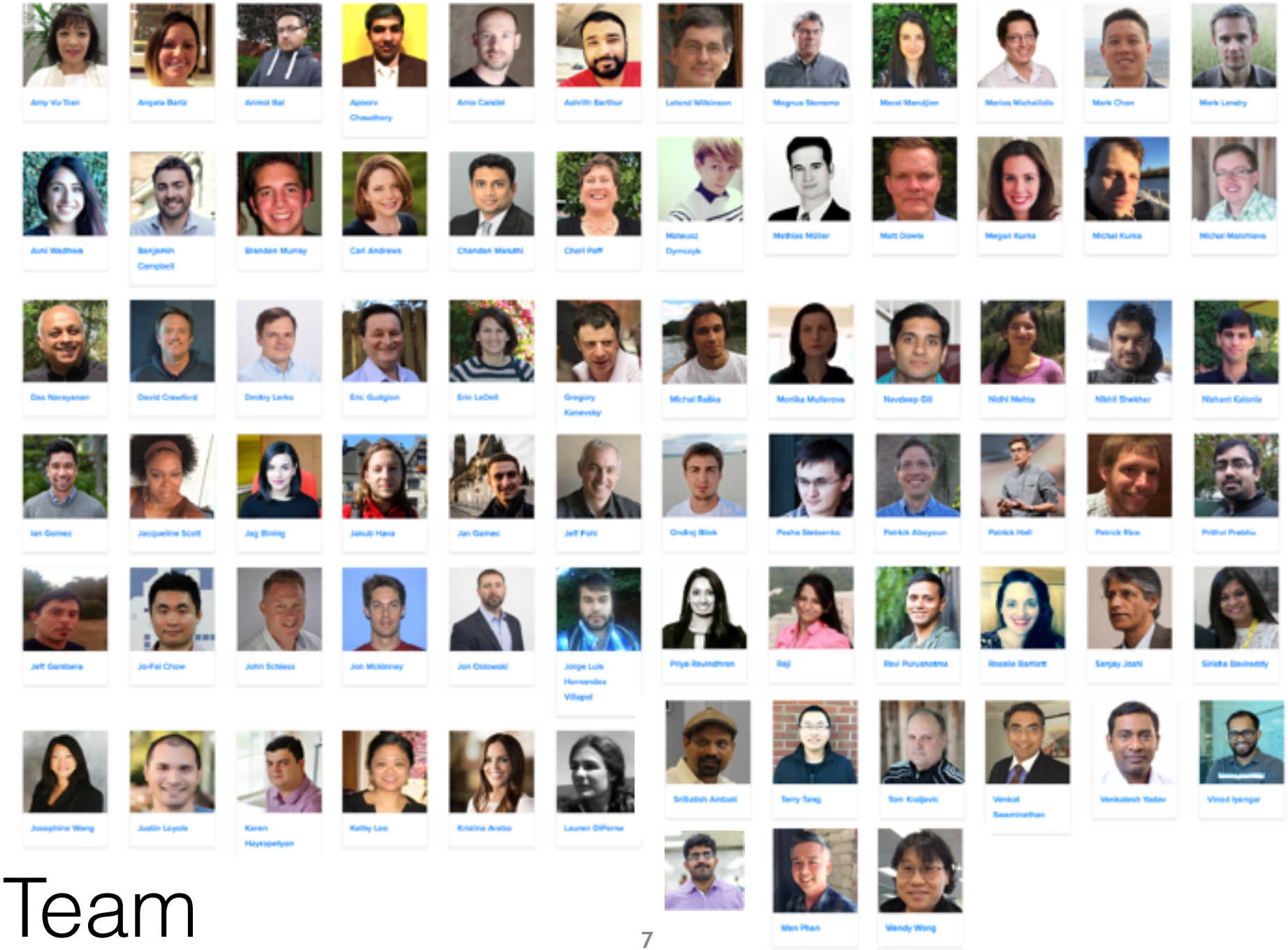
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction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

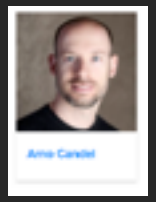


H2O.ai



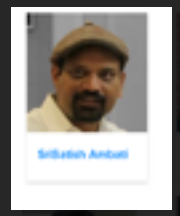
H_2O Team

H2O.ai



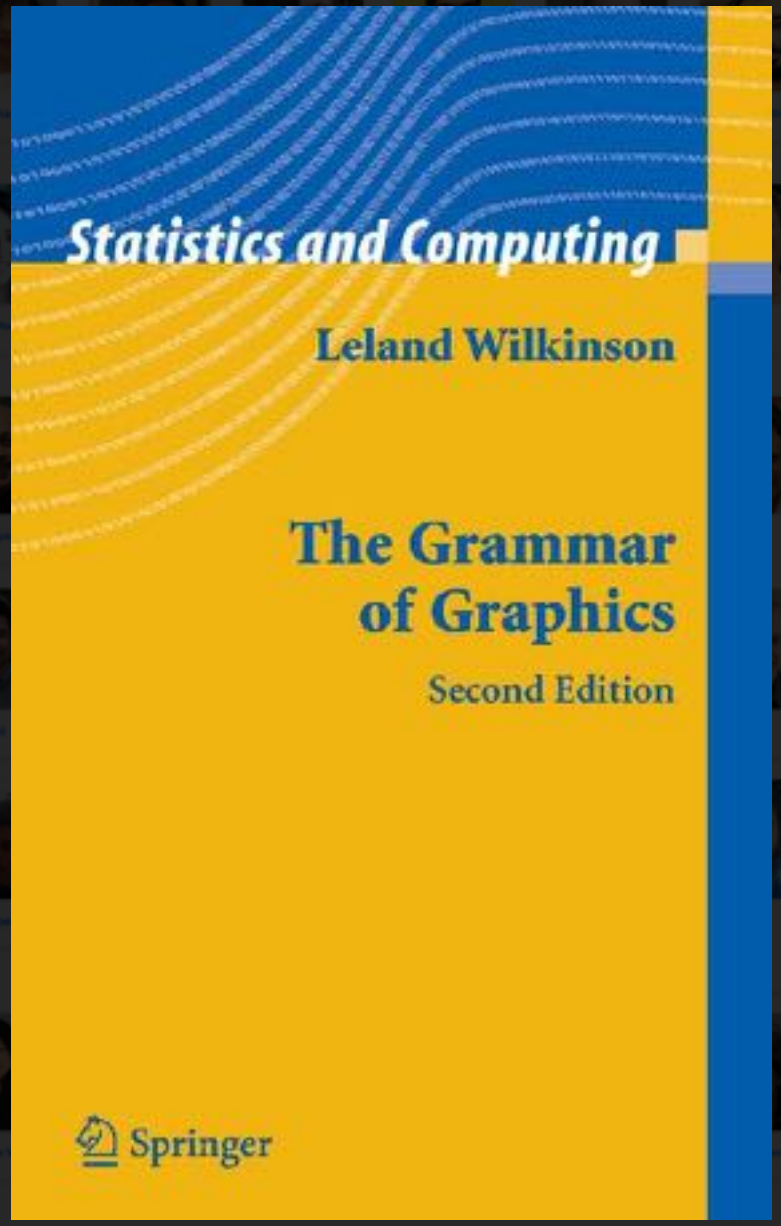
Arno Candel, CTO
Fortune's 2014 Big Data All-Star

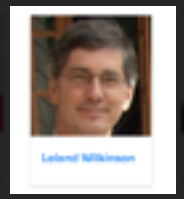
Sri Ambati, Co-founder & CEO



H_2O Team

H_2O.ai





Origin of R Package `ggplot2`

ggplot2

H_2O Team

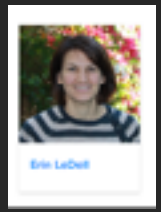
H_2O.ai

Matt Dowle

data.table

H_2O Team

H2O.ai

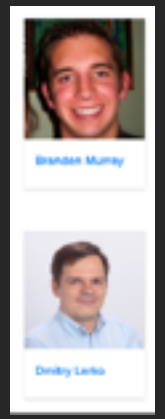


Erin LeDell, Chief ML Scientist
Women in ML/DS & R-Ladies
Global

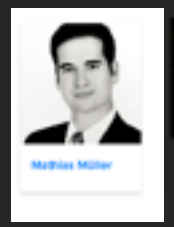
H_2O Team

H_2O.ai

48th

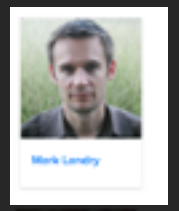


25th



4th

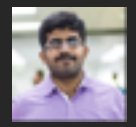
1st



33rd

Kaggle Grand Masters

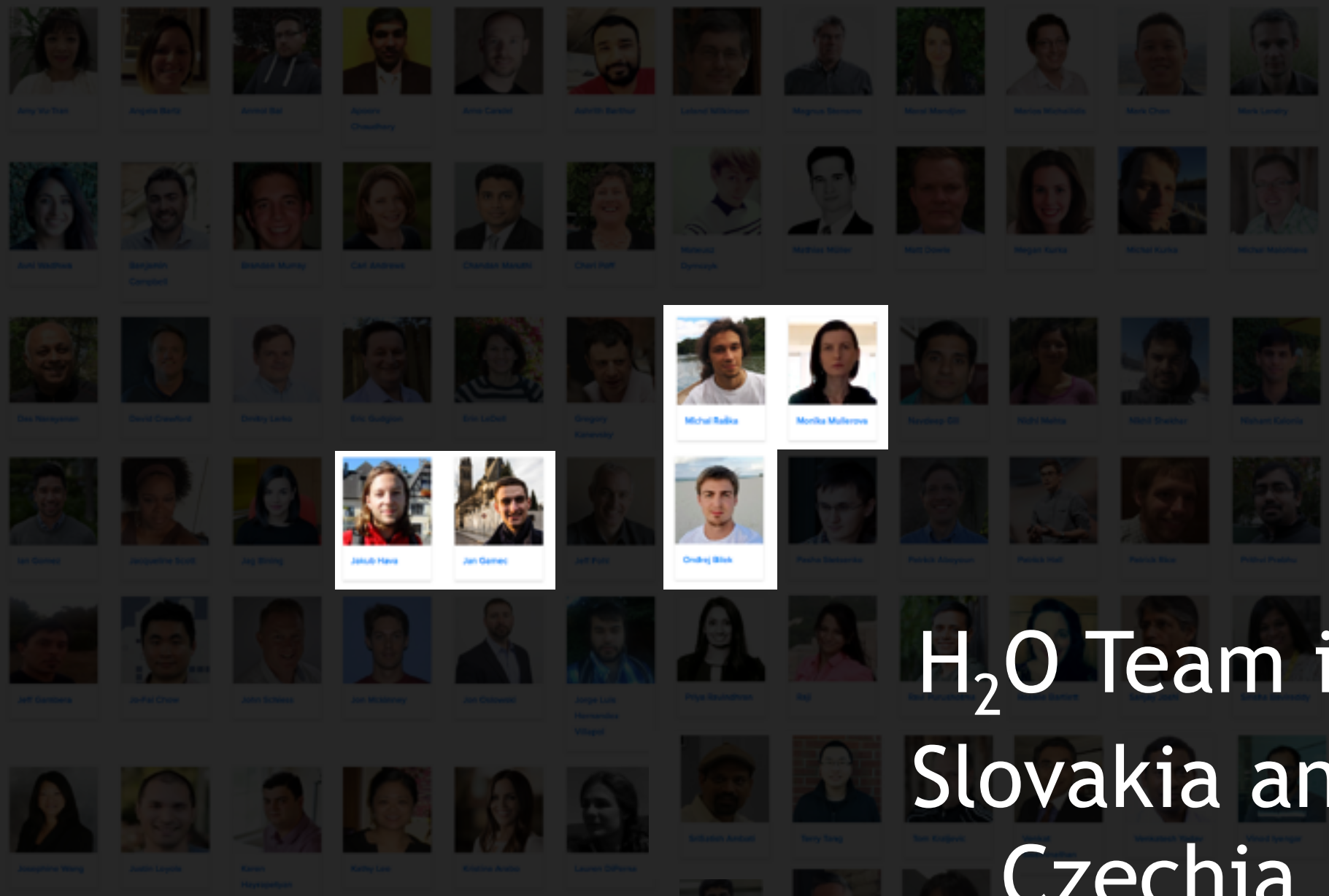
Their Highest Rank in Kaggle
(about 80,000 competitors)



13th

H_2O Team

H2O.ai



H_2O Team in Slovakia and Czechia

H_2O Team

H2O.ai

H_2O Products

H_2O.ai

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

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H_2O4GPU

Lightning Fast machine learning on GPUs

DRIVERLESSAI

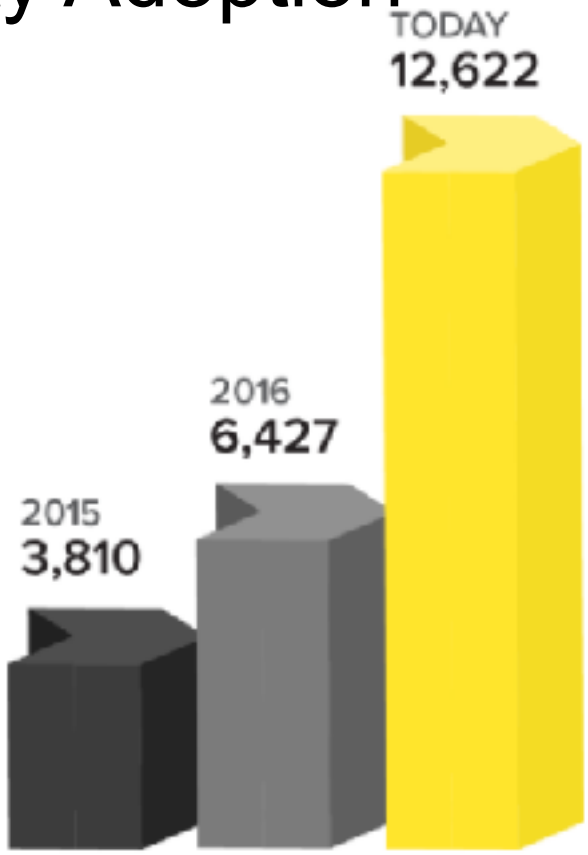
Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

H_2O.ai

Worldwide Community Adoption



Companies using H2O.ai

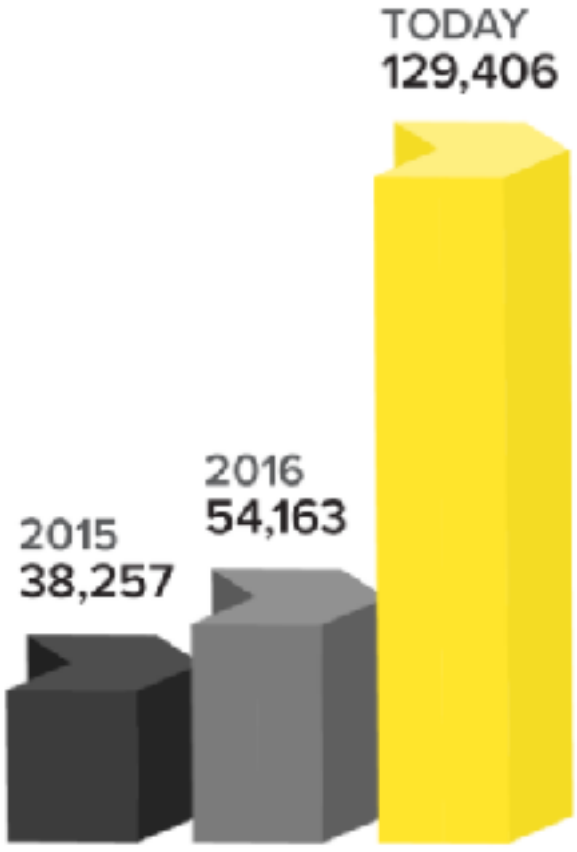
222 OF THE FORTUNE 500

H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES



H2O.ai Users

* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H2O.ai Solution Leadership Across Verticals

H2O.ai

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

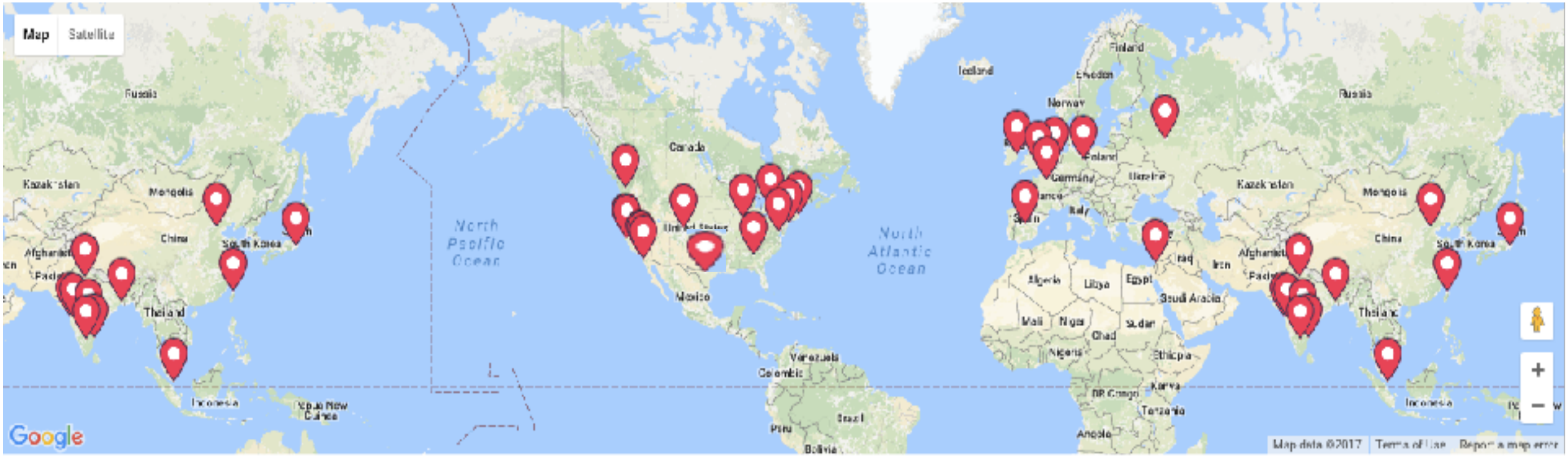
H_2O + KNIME Talk at KNIME Summit Mar 2017

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms

Source: Gartner (February 2018)

H_2O.ai

Community Expansion



88,286	43	52	48	18
members	interested	Meetups	cities	countries

Find out more: **www.h2o.ai/community/**

H_2O Machine Learning Platform

H_2O Products

H_2O.ai

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

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H_2O4GPU

Lightning Fast machine learning on GPUs

DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

Steam

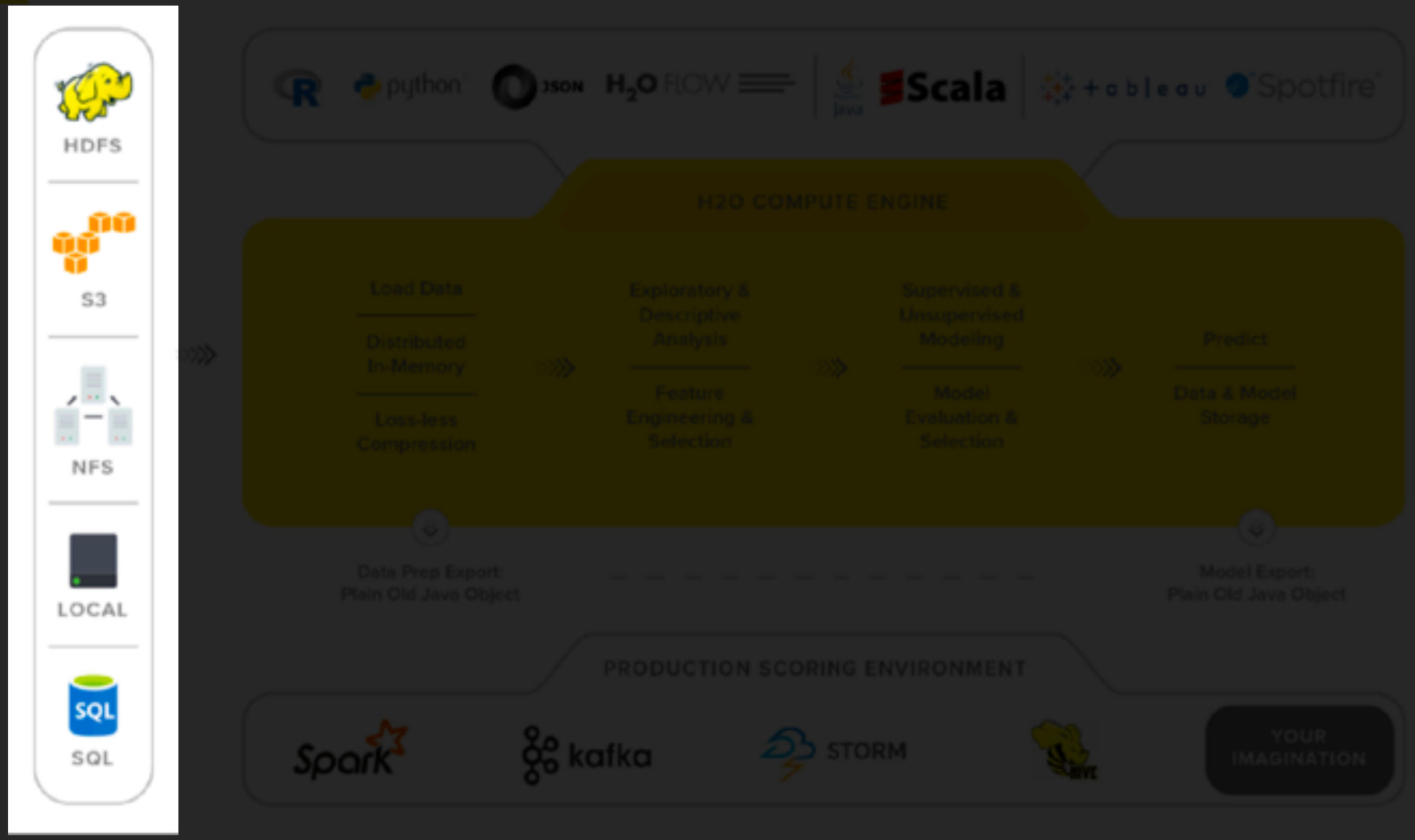
Secure multi-tenant H2O clusters

H_2O.ai

High Level Architecture

High Level Architecture

Import Data from Multiple Sources

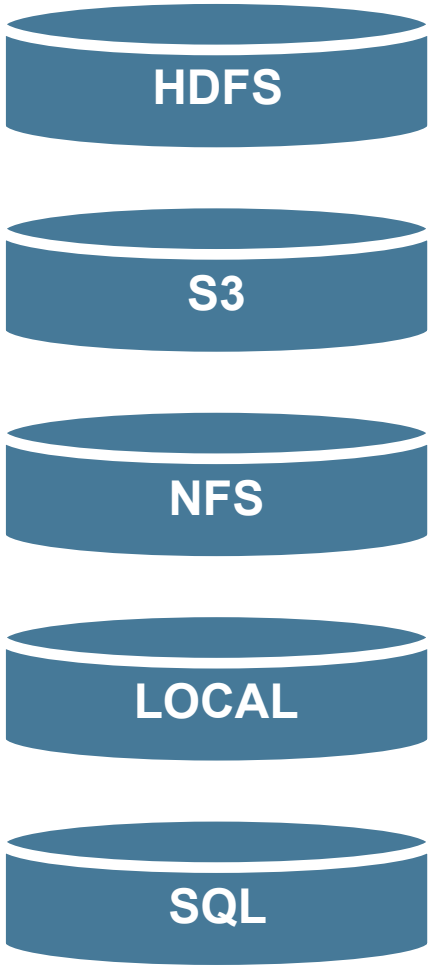


Supported Formats & Data Sources

9 Formats

File type or Folder of Files

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



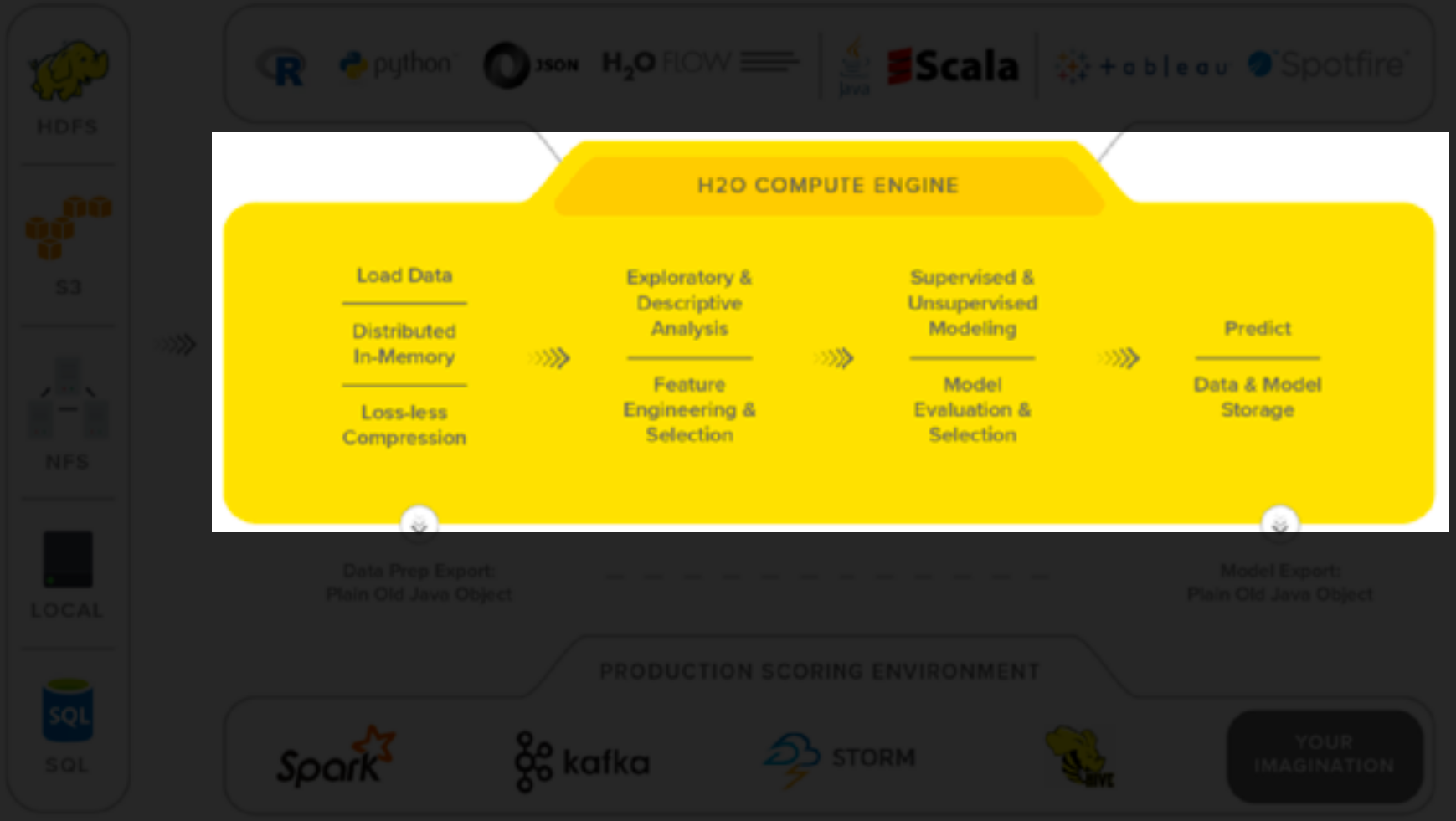
5 Sources

* 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

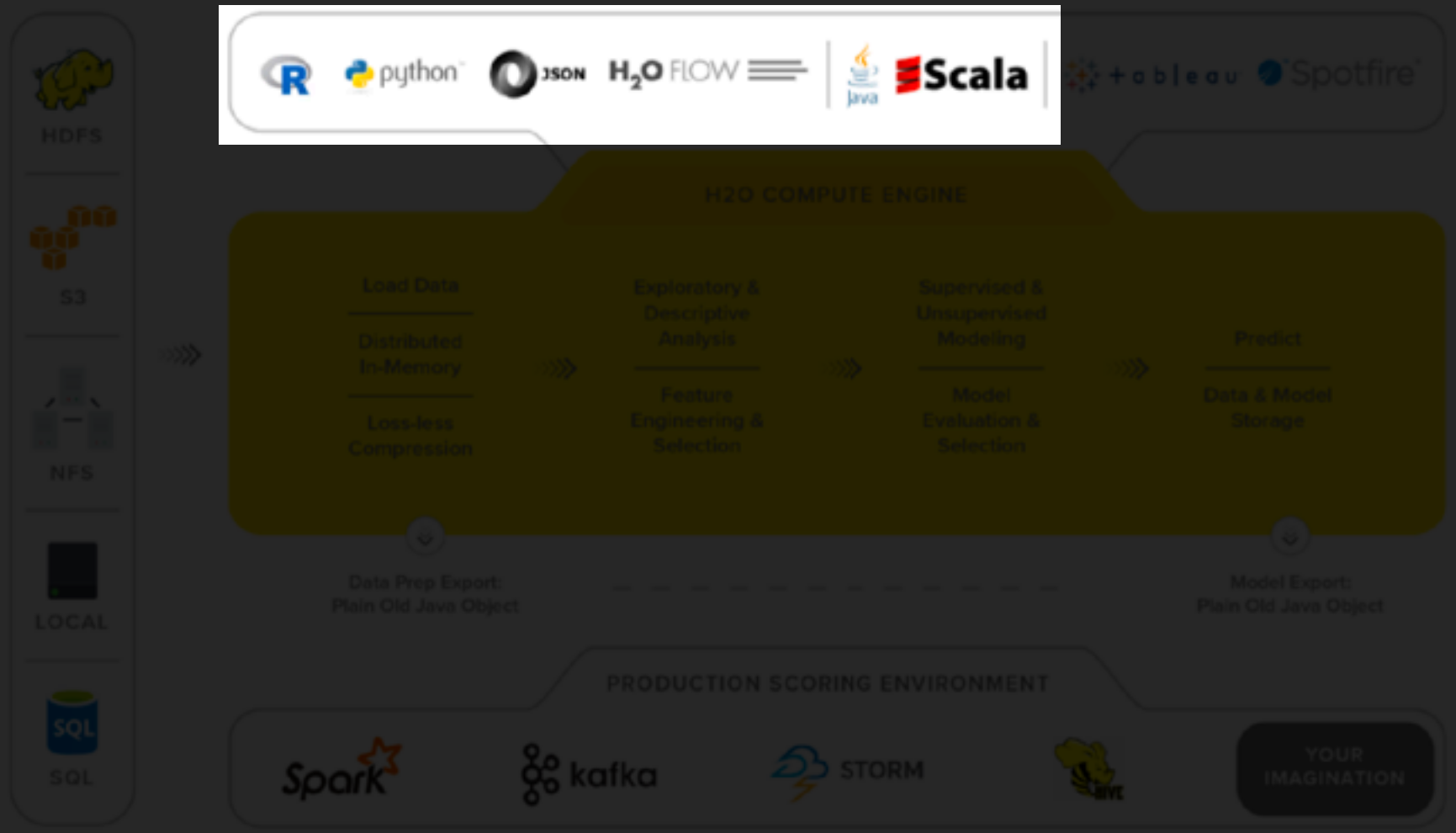
High Level Architectur

Fast, Scalable & Distributed Compute Engine Written in Java



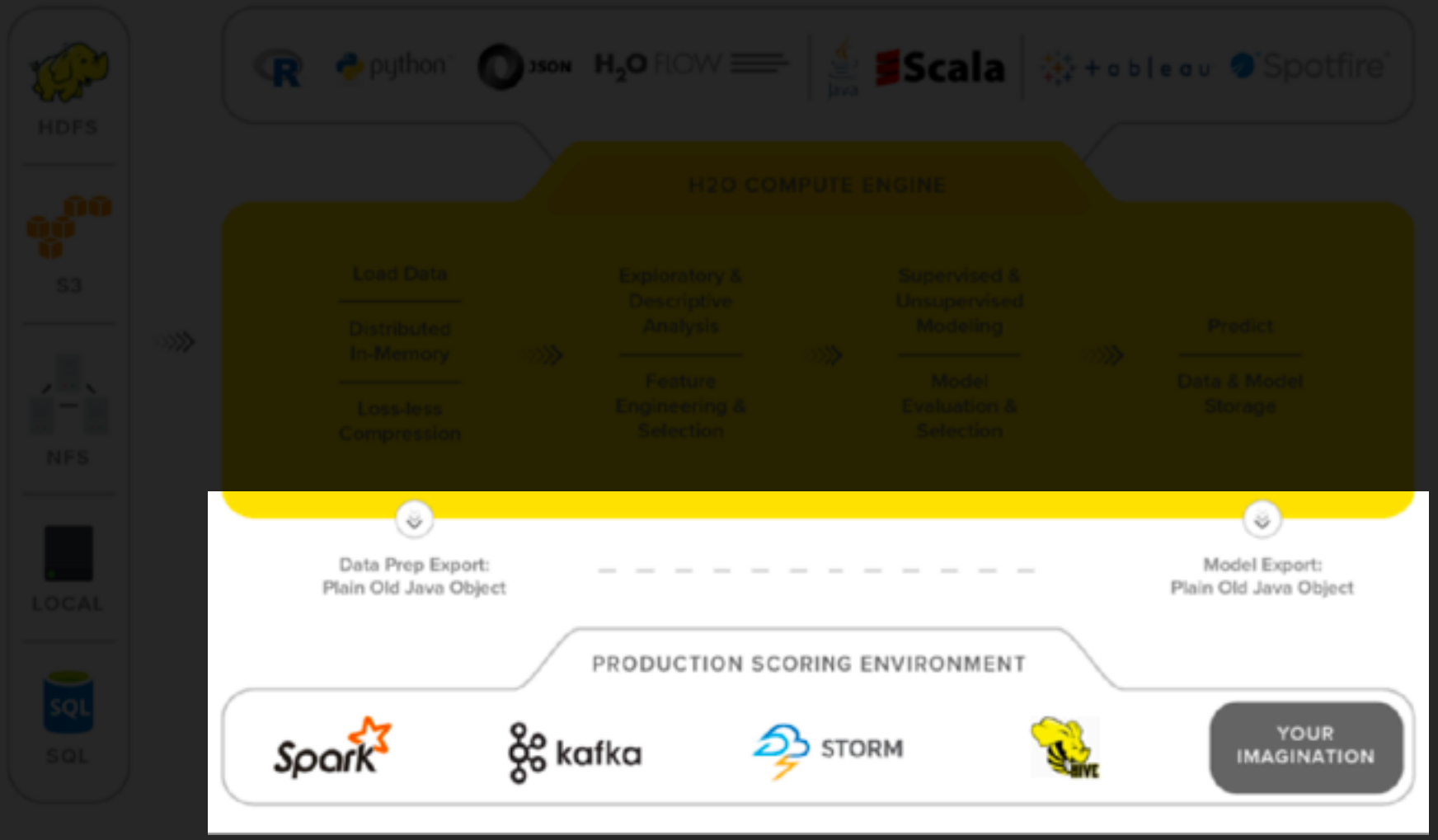
Multiple Interfaces

High Level Architecture



High Level Architecture

Export Standalone Models for Production

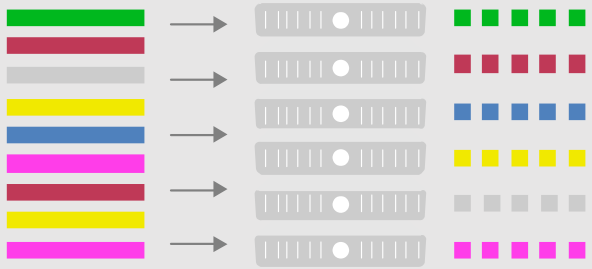


High Level Architecture

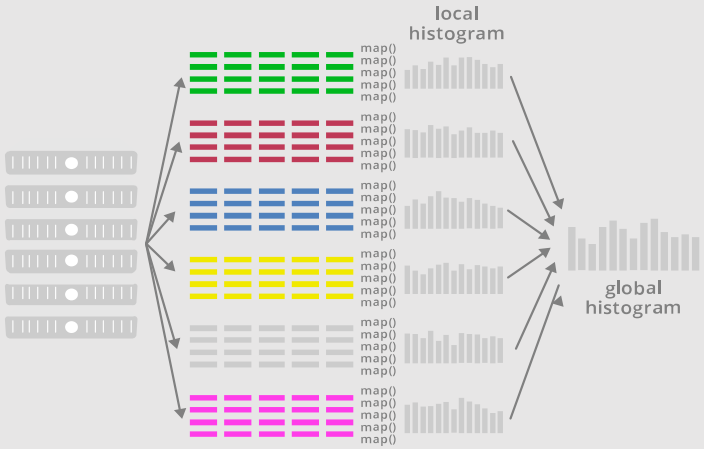
H_2O - Open Source Machine Learning Platform

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



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

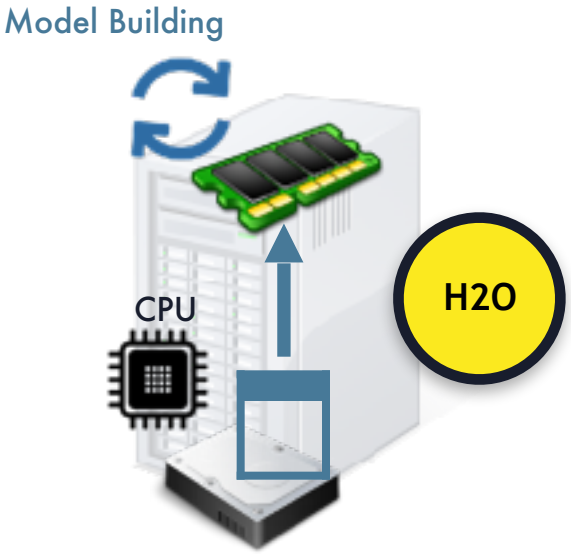
Advantageous Foundation

- 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

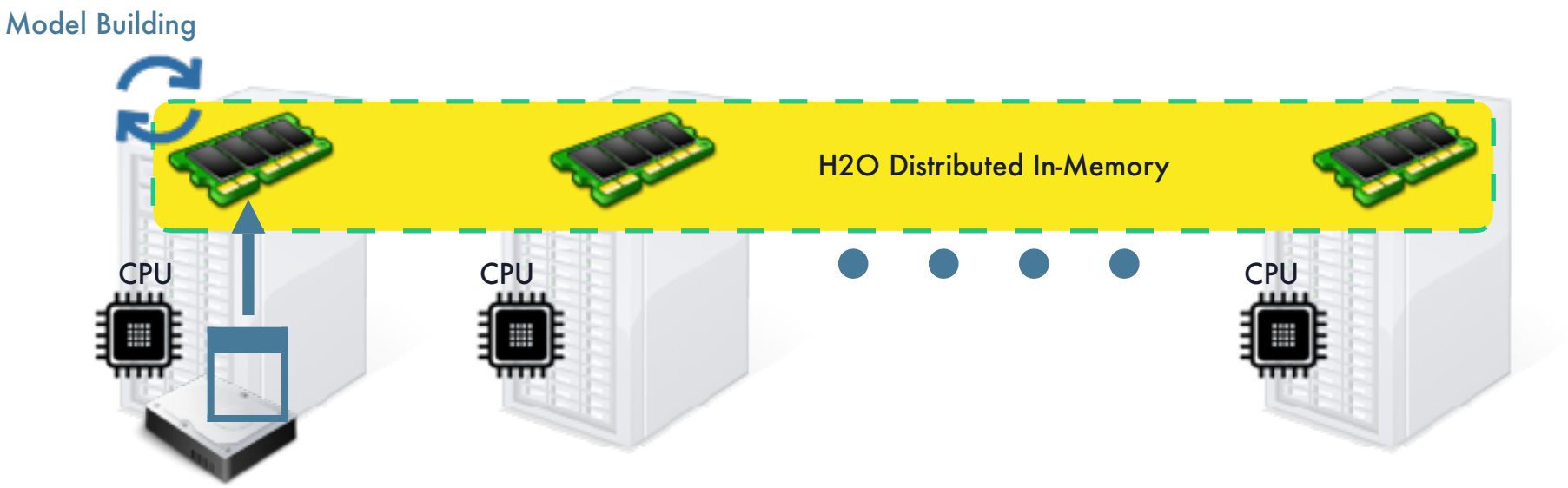
H_2O Core



H2O.ai

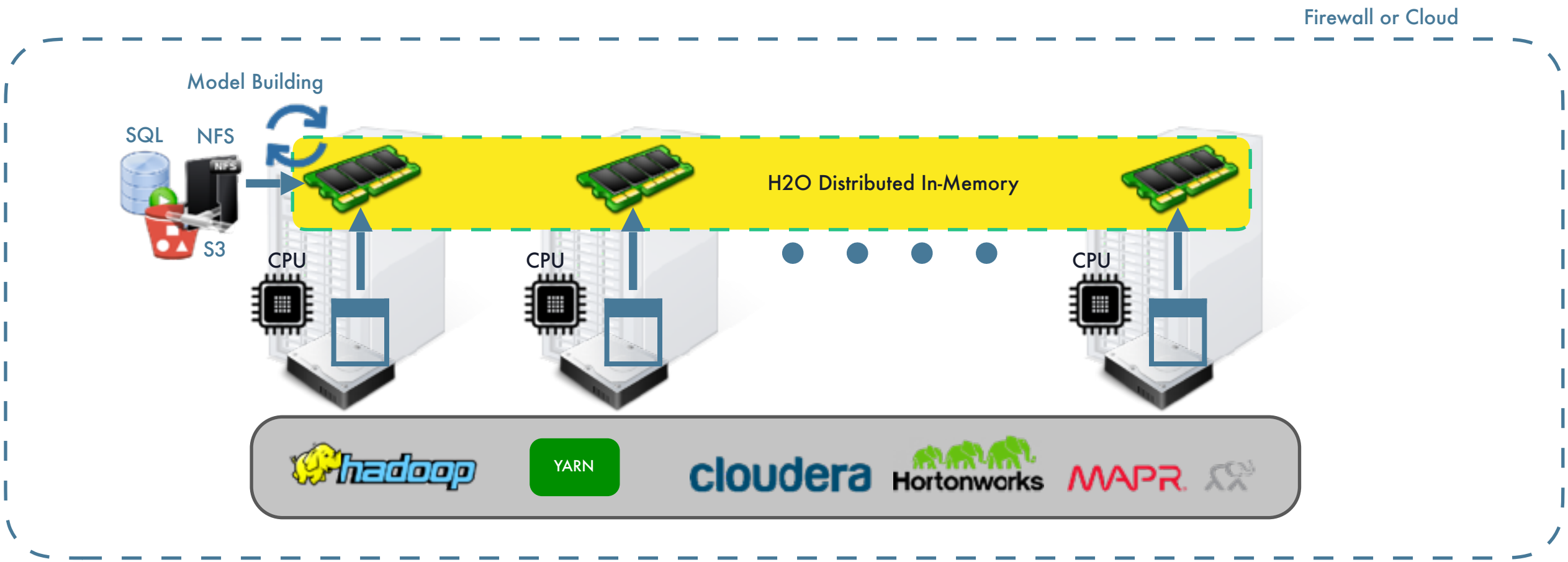
H_2O Core

H_2O Core



H2O.ai

H_2O Core



H2O.ai

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 earning**: 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

Downloading H_2O

H_2O

H2O works with R, Python, Scala on Hadoop/Yarn, Spark or your laptop

Latest Stable Release

Nightly Bleeding Edge

H_2O in the Cloud

aws

Microsoft Azure

H_2O is licensed under the Apache License, Version 2.0

H_2O

Version 3.18.0.10

Fast Scalable Machine Learning API
For Smarter Applications

DOWNLOAD AND RUN | INSTALL IN R | INSTALL IN PYTHON | INSTALL ON HADOOP | USE FROM MAVEN

DOWNLOAD H_2O

Get started with H_2O in 3 easy steps

1. Download H_2O. This is a zip file that contains everything you need to get started.
2. From your terminal, run:

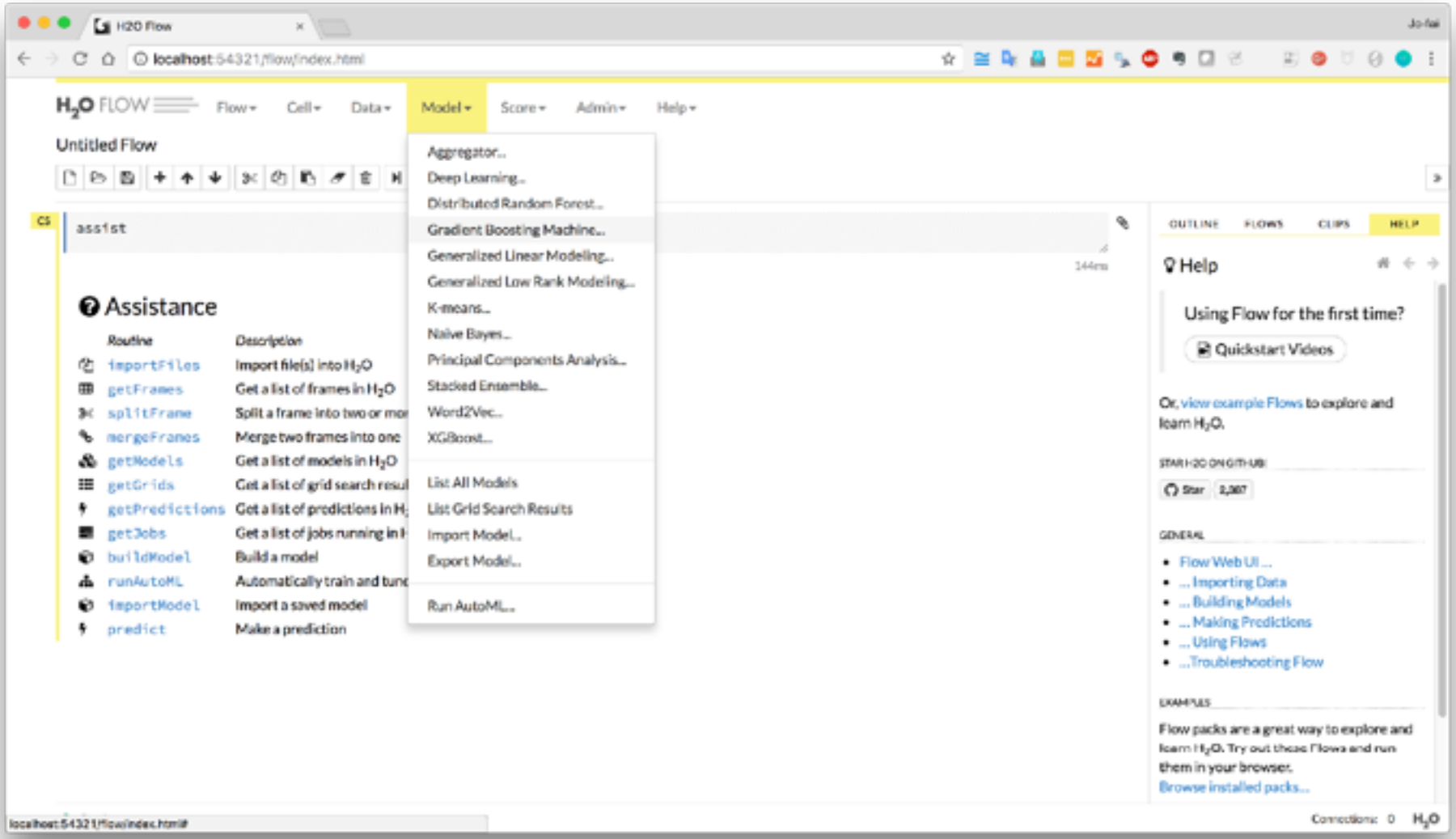
```
cd ~/Downloads
unzip h2o-3.18.0.10.zip
cd h2o-3.18.0.10
java -jar h2o.jar
```

3. Point your browser to http://localhost:54321

h2o.ai/download/

H_2O.ai

H_2O Flow and Python Client – First Demo



H2O.ai

H2O.ai

Products | Industries | Customers | Community | Partners | Download

H2O Documentation

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

Getting Started & User Guides

Open Source | Commercial

H2O

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

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

Download H2O

Sparkling Water

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

Quick Start Video - Scala

Download Sparkling Water

Driverless AI

What is Driverless AI?
Driverless AI User Guide HTML PDF
Recent Changes
Driverless AI Booklet
MLI with Driverless AI Booklet

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

Download Driverless AI

H2O4GPU (alpha)

H2O4GPU Readme
Open Source License (Apache V2)

Download H2O4GPU

docs.h2o.ai

H2O.ai

Thanks!

- Organisers & Sponsors
  - GapData Institute
  - PyData Bratislava
- H_2O's Mission
  - Democratize AI
  - Make Machine Learning Accessible to Everyone
- Code, Slides & Documents
  - bit.ly/h2o_meetups
  - docs.h2o.ai
- Contact
  - michalr@h2o.ai
  - github.com/michal-raska
- Please search/ask questions on **Stack Overflow**
  - Use the tag `**h2o**` (not h2 zero)

H_2O.ai