QCon 2015: Building Machine Learning Applications

Michal Malohlava

Amy Wang

H2O.ai

Company Overview

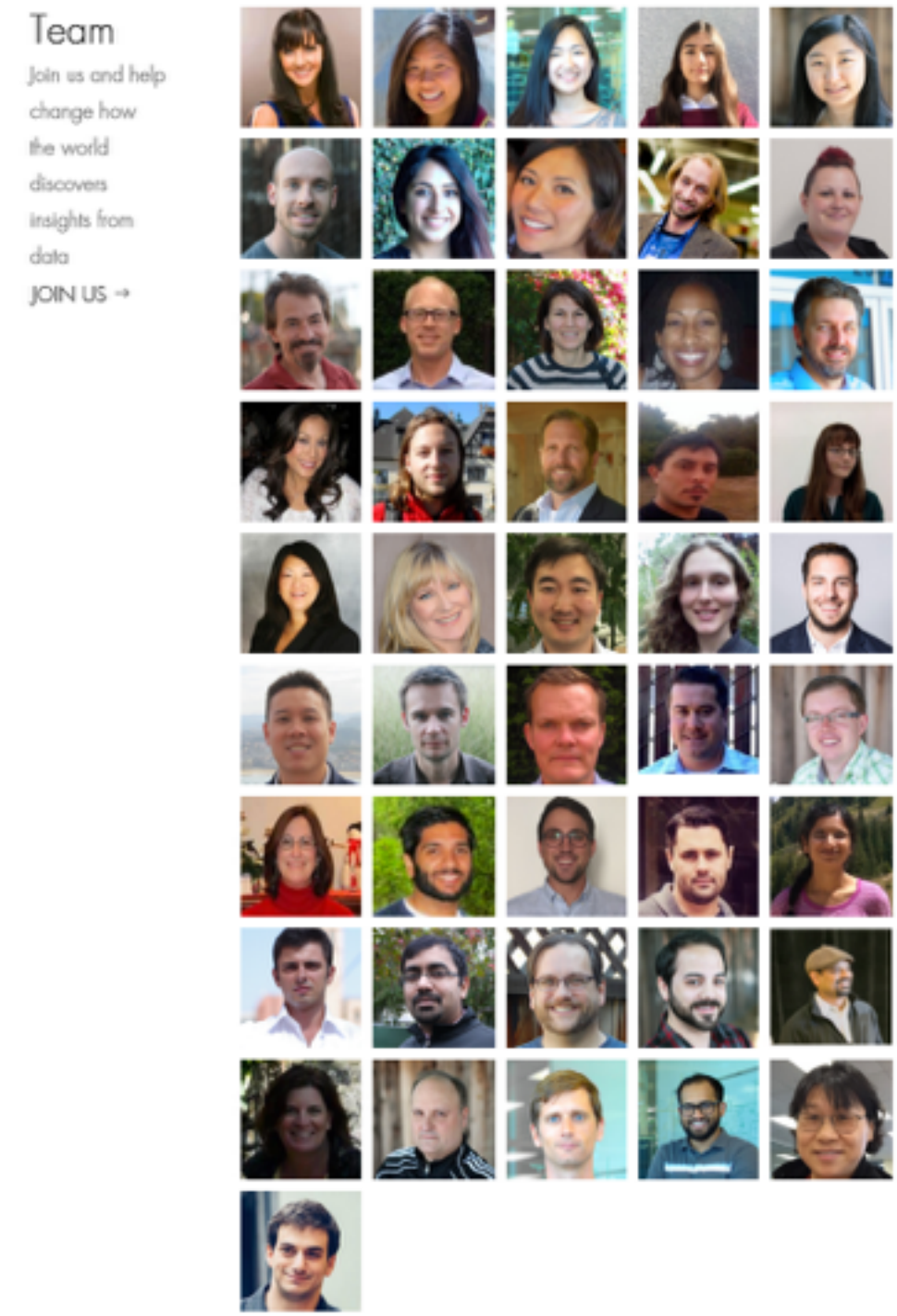
Company

- Team: 46. Founded in 2012, Mountain View, CA
- Stanford Math & Systems Engineers

Product

- Open Source Leader in Machine & Deep learning
- Ease of Use and Smarter Applications
- R, Python, Spark & Hadoop Interfaces
- Expanding Predictions to Mass Analyst markets


The H2O.ai Team



Scientific Advisory Council

Dr. Trevor Hastie

- PhD in Statistics, Stanford University
- John A. Overdeck Professor of Mathematics, 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*
- 108,404 citations (via Google Scholar)

Dr. Rob Tibshirani

- PhD in Statistics, Stanford University
- Professor of Statistics and Health Research and Policy, Stanford University
- COPPS Presidents' Award recipient
- 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*

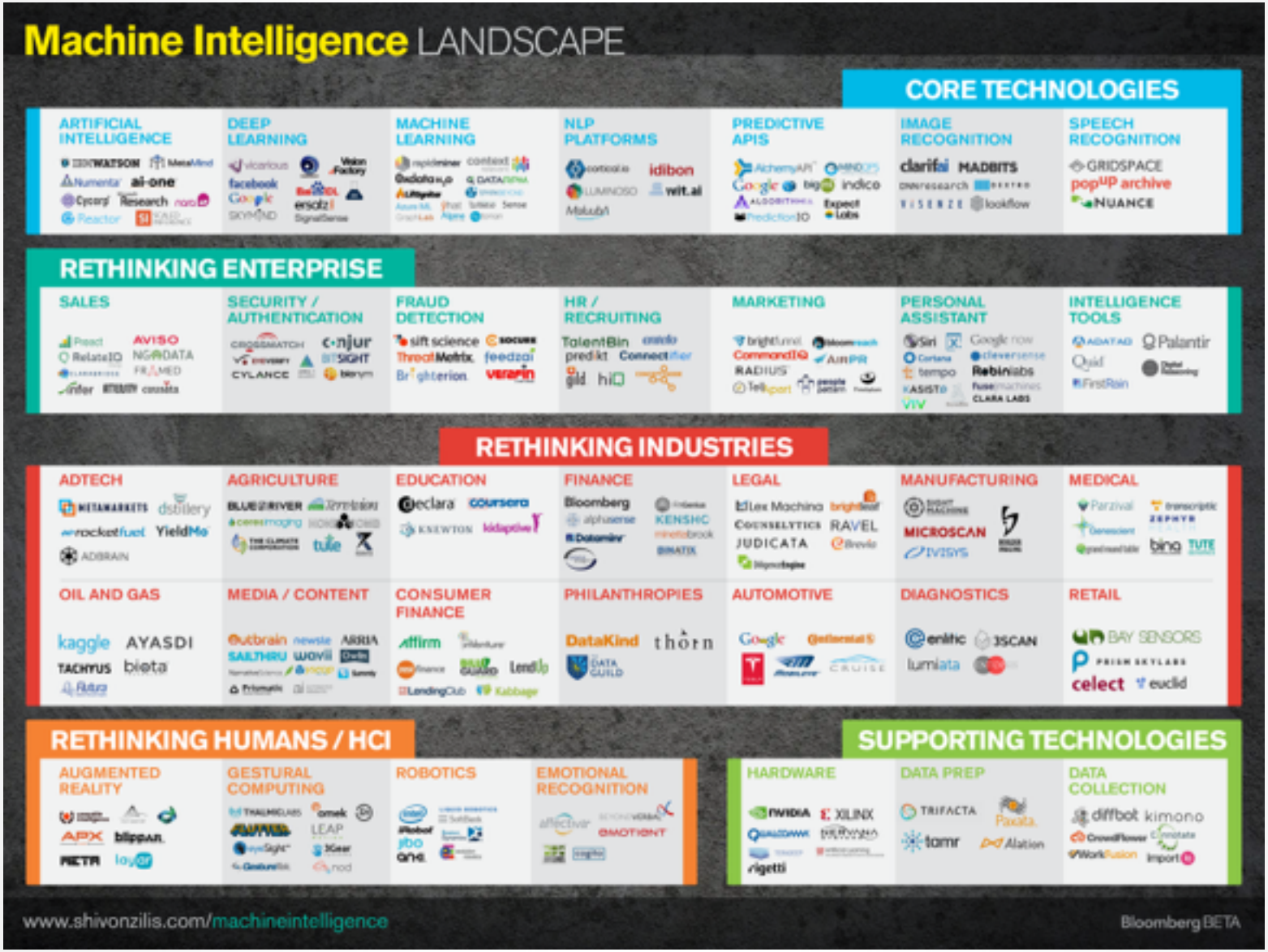
Dr. Stephen Boyd

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

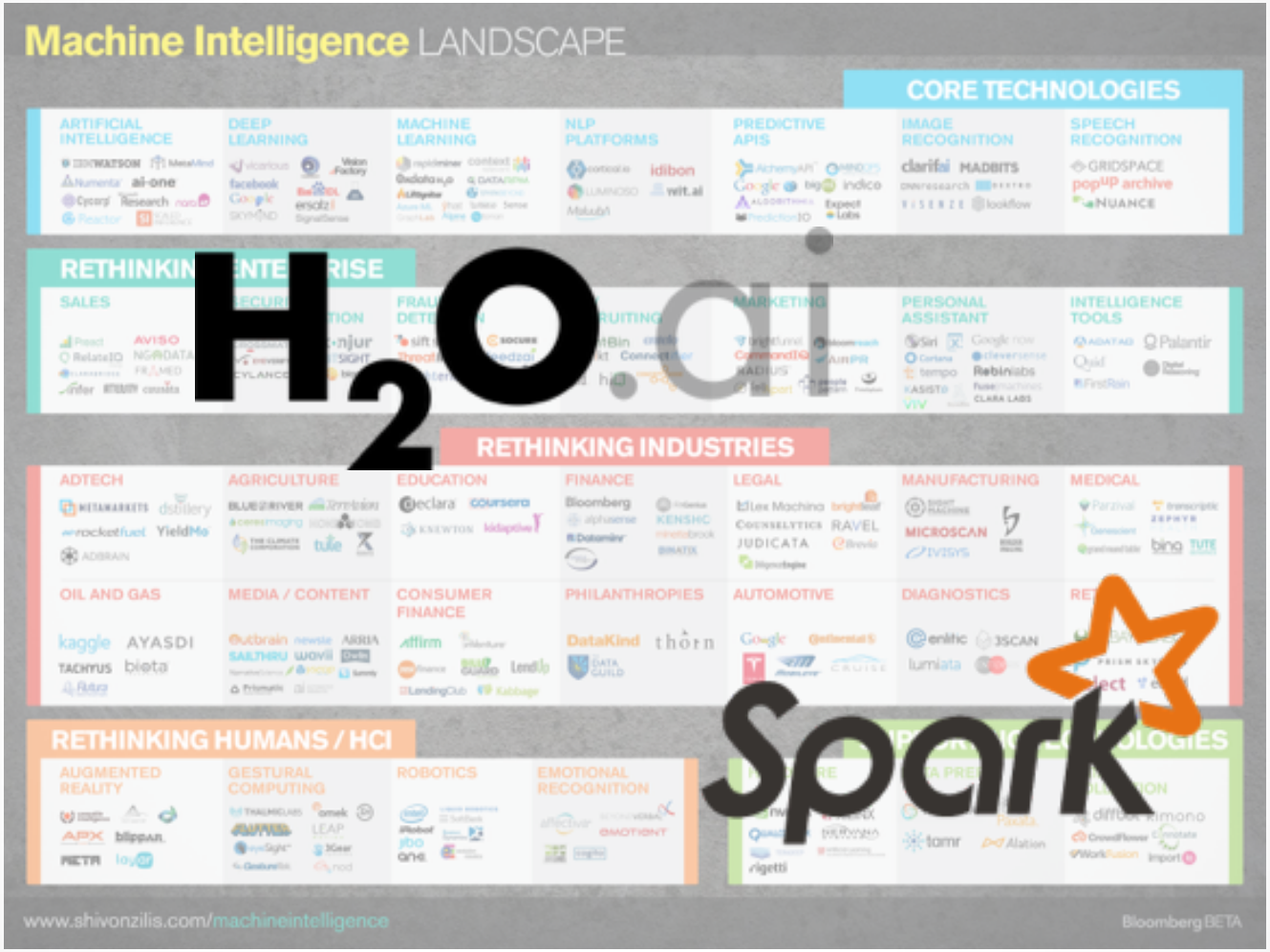
The Goal of Today

- Learn about H2O
- Learn about Spark
- Learn how to use Sparkling Water
- Build data products or smarter applications using Sparkling Water

Build an application with ... ?



... with Spark and H2O!

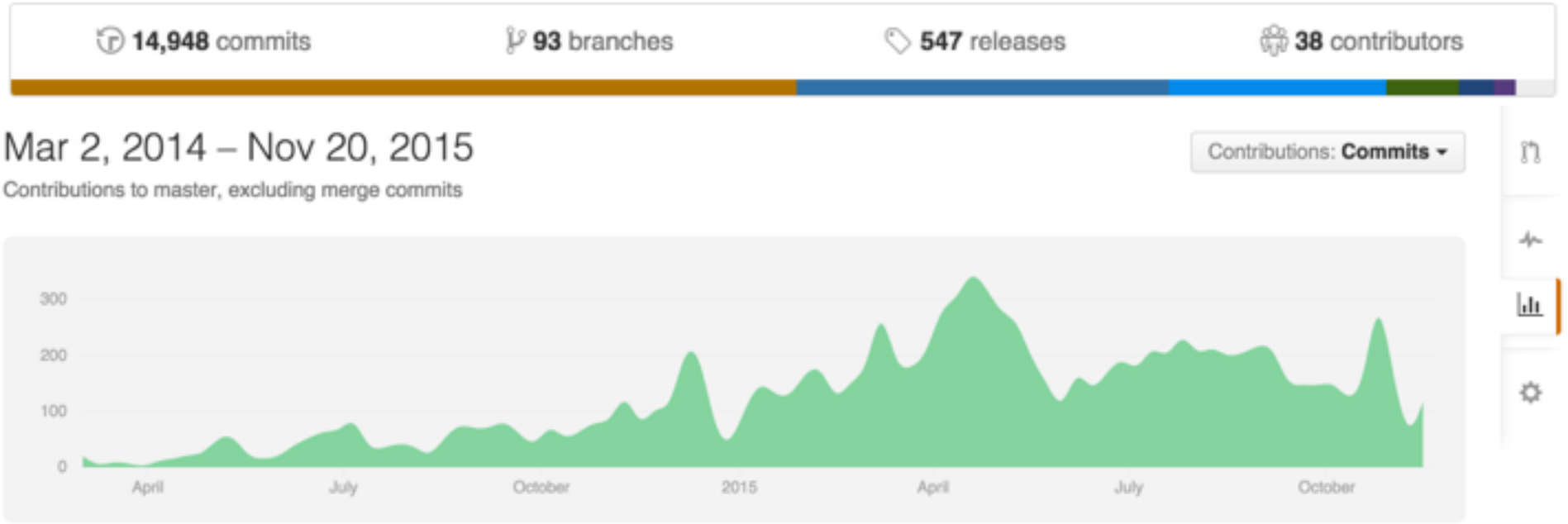


Smarter Machine Learning Applications

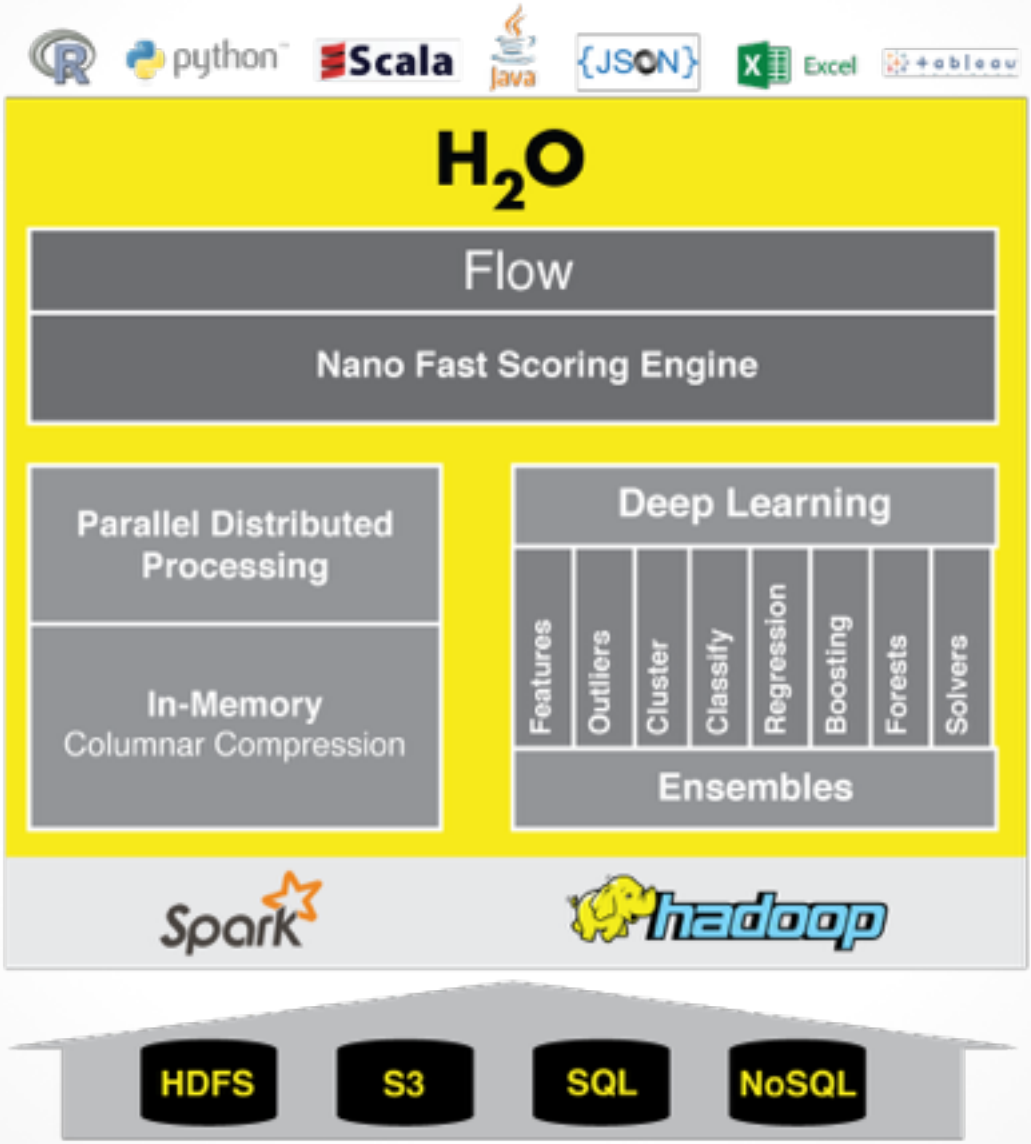
Table of Contents

What is H2O?

- Open source distributed execution platform
- User-friendly R, Python, Java, and Scala APIs for data transformation based on Data Frames
- Library of production ready machine learning algorithms
- New releases with new features and bug fixes available at rapid fire pace



What is H2O?


python
Scala
Java
{JSON}
Excel
+ableau
H2O
Flow
Nano Fast Scoring Engine
Parallel Distributed Processing
In-Memory
Columnar Compression
Deep Learning
Features
Outliers
Cluster
Classify
Regression
Boosting
Forests
Solvers
Ensembles
Spark
hadoop
HDFS
S3
SQL
NoSQL

Algorithms on H2O

Supervised Learning

Statistical Analysis

- Generalized Linear Models : Binomial, Gaussian, Gamma, Poisson and Tweedie
- Cox Proportional Hazards Models
- 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

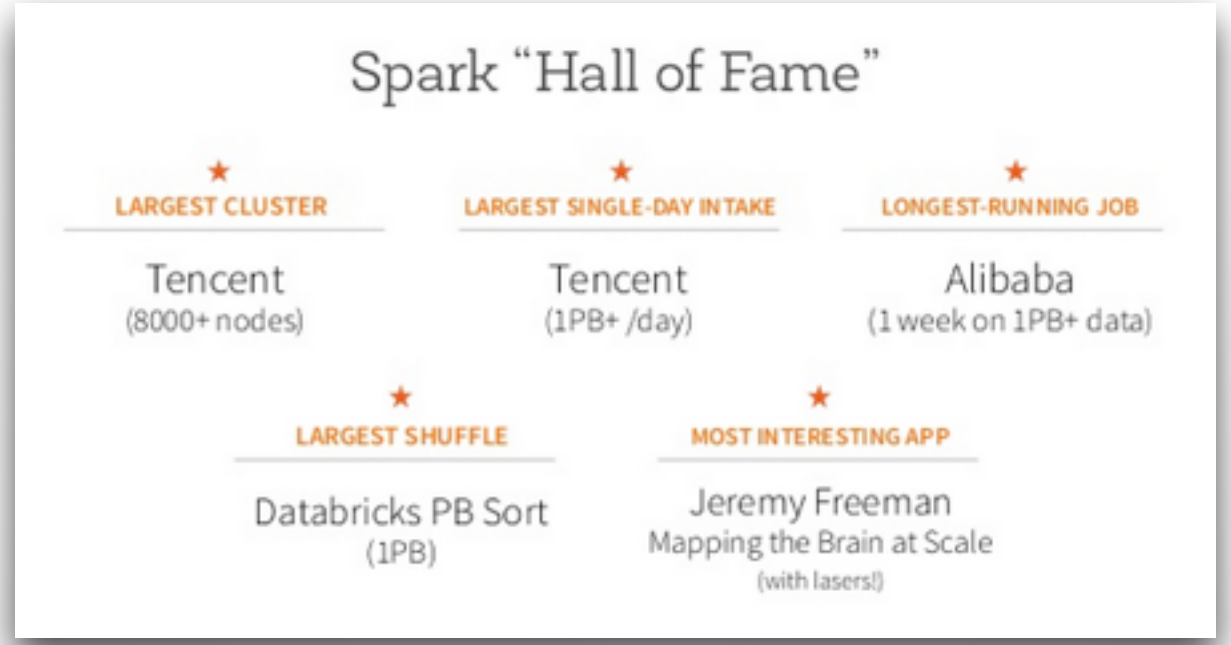
Algorithms on H2O

Unsupervised Learning

Clustering	• K-means : Partitions observations into k clusters/ groups of the same spatial size
Dimensionality Reduction	• Principal Component Analysis : Linearly transforms correlated variables to independent components
Anomaly Detection	• Autoencoders: Find outliers using a nonlinear dimensionality reduction using deep learning

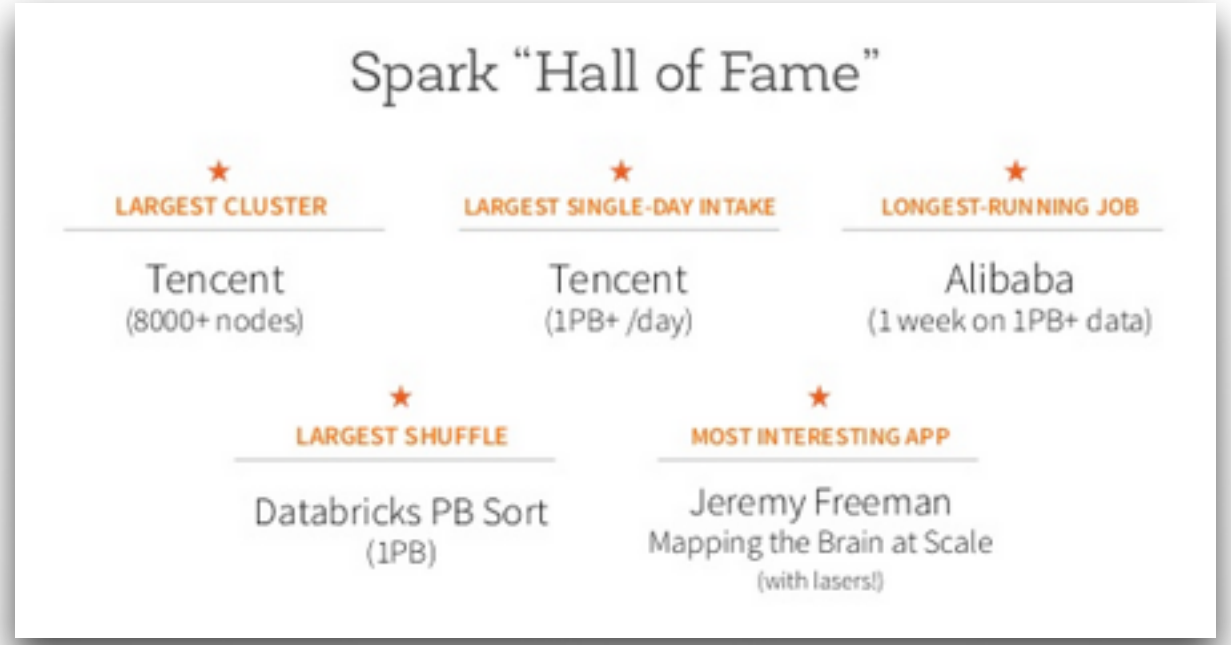
What is Spark?

- Open source distributed execution platform.
- User-friendly API for data transformation based on RDD
- Platform Components - SQL, MLLib, Text Mining
- Multitenancy
- Large and Active Community with over 700 contributors

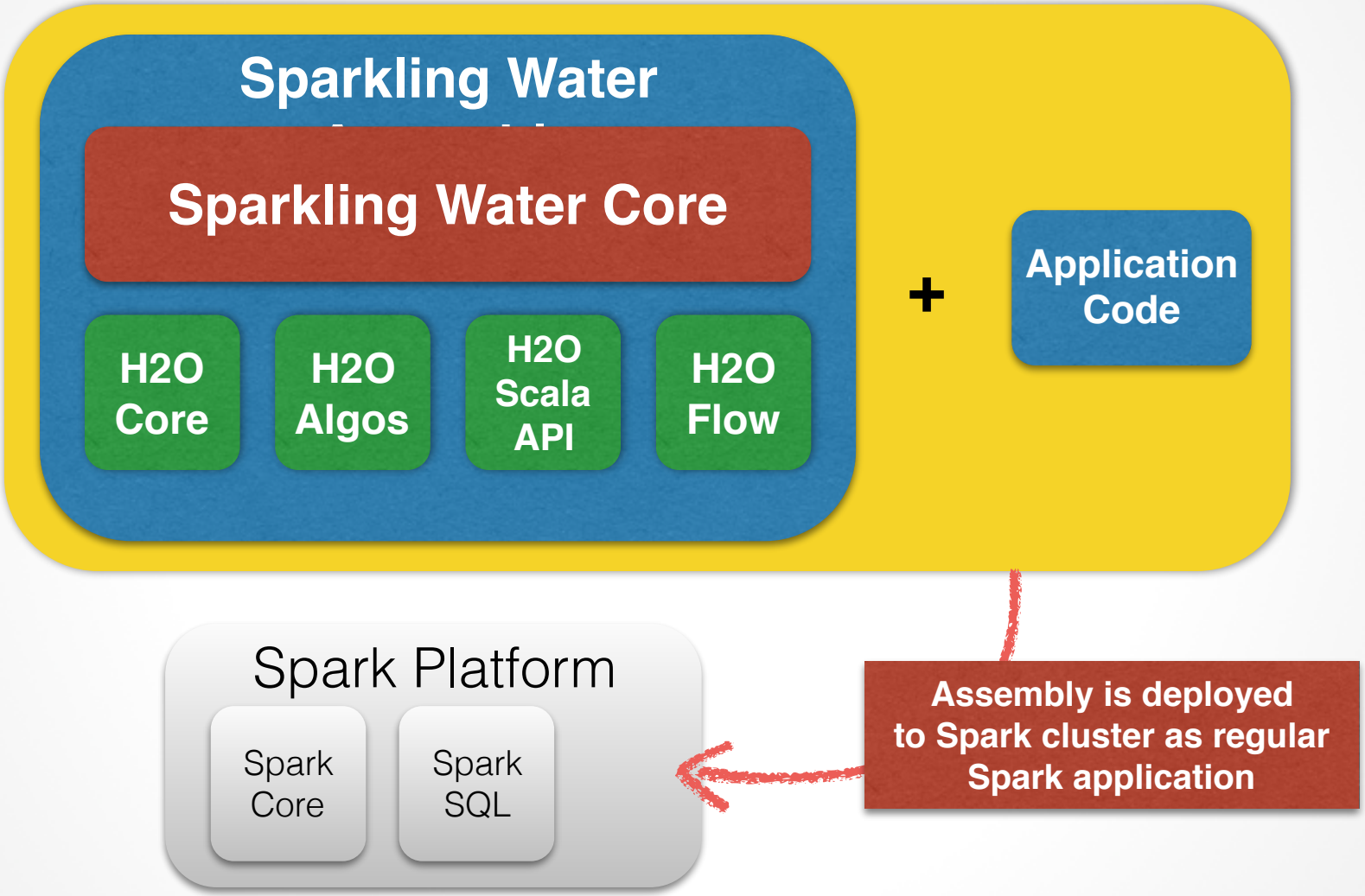


What is Spark?

- Open source distributed execution platform.
- User-friendly API for data transformation based on RDD
- Platform Components - SQL, MLLib, Text Mining
- Multitenancy
- Large and Active Community with over 700 contributors



What is Sparkling Water?



Let's get started with installations!