QCon 2015: Building Machine Learning Applications

Michal Malohlava

Amy Wang

H2O.ai

Company Overview

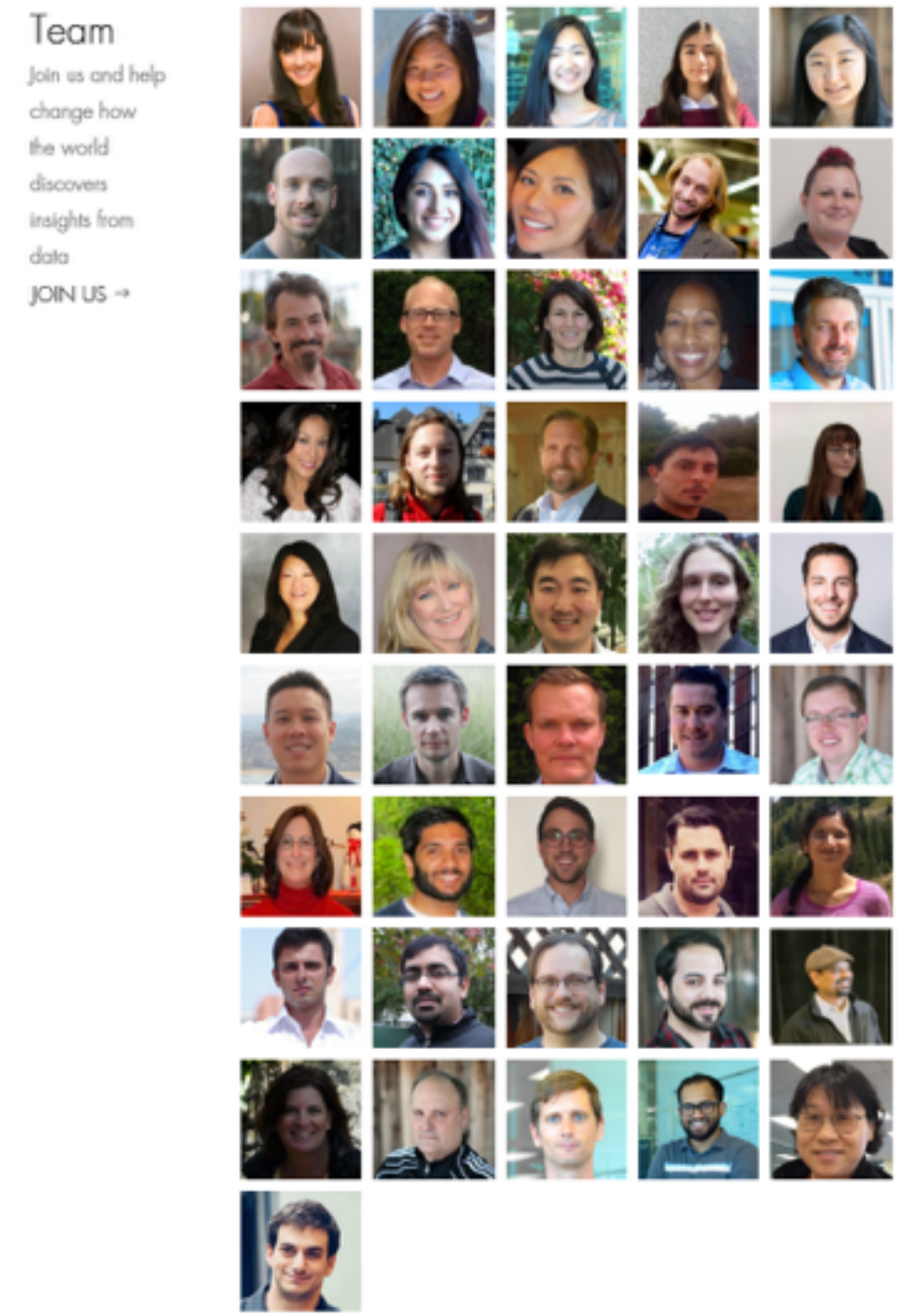
Company

- 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*

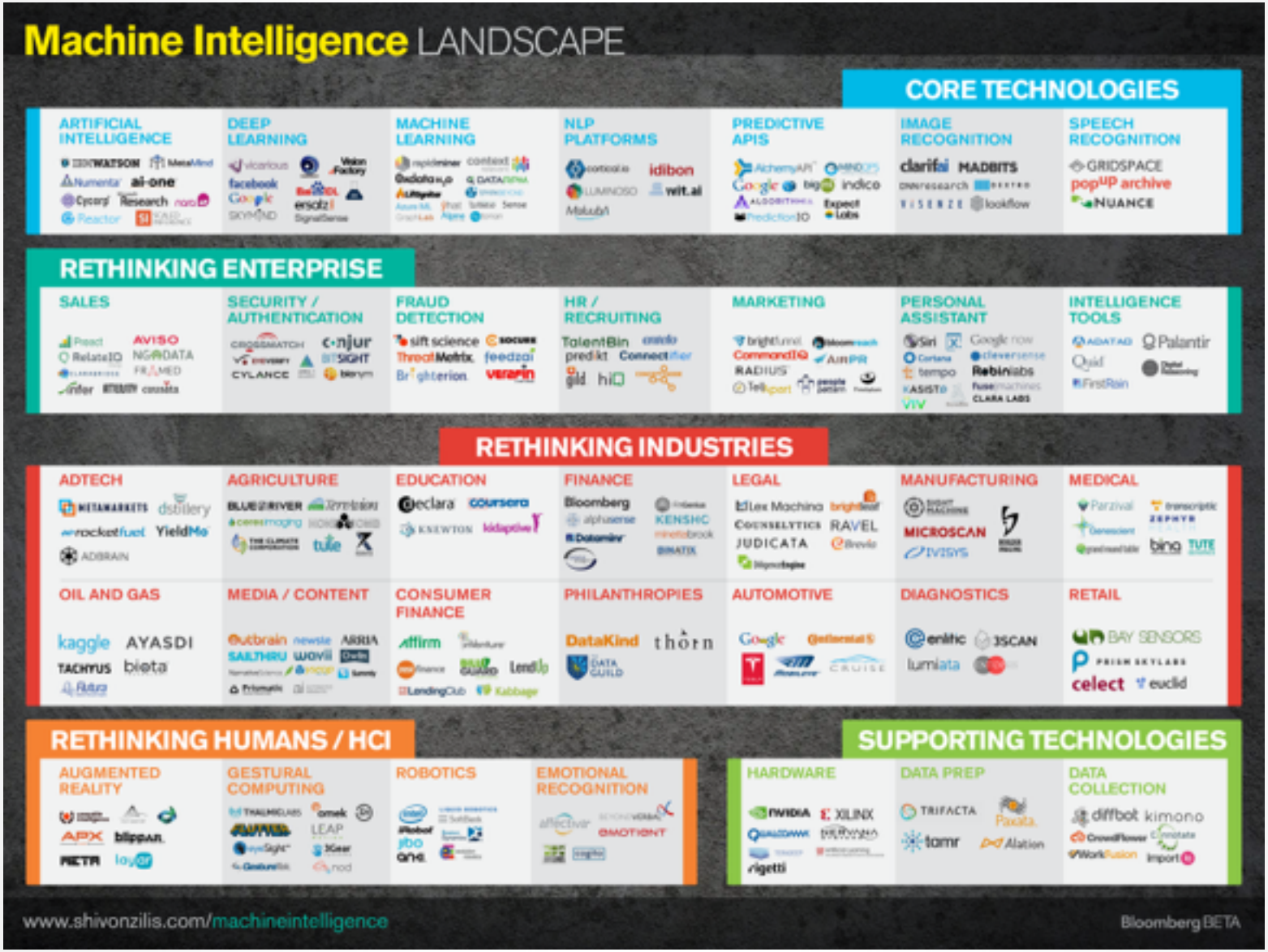
Dr. Stephen Boyd

- 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!

Spark + H2O

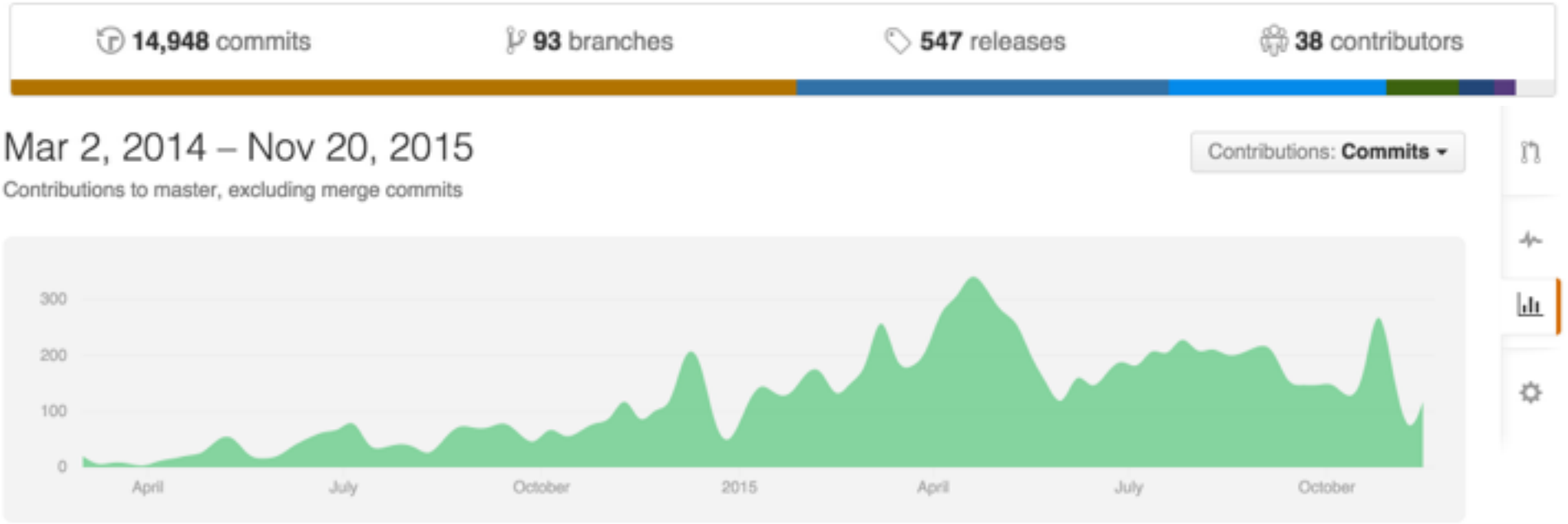
SPARKLING WATER

Smarter Machine Learning Applications

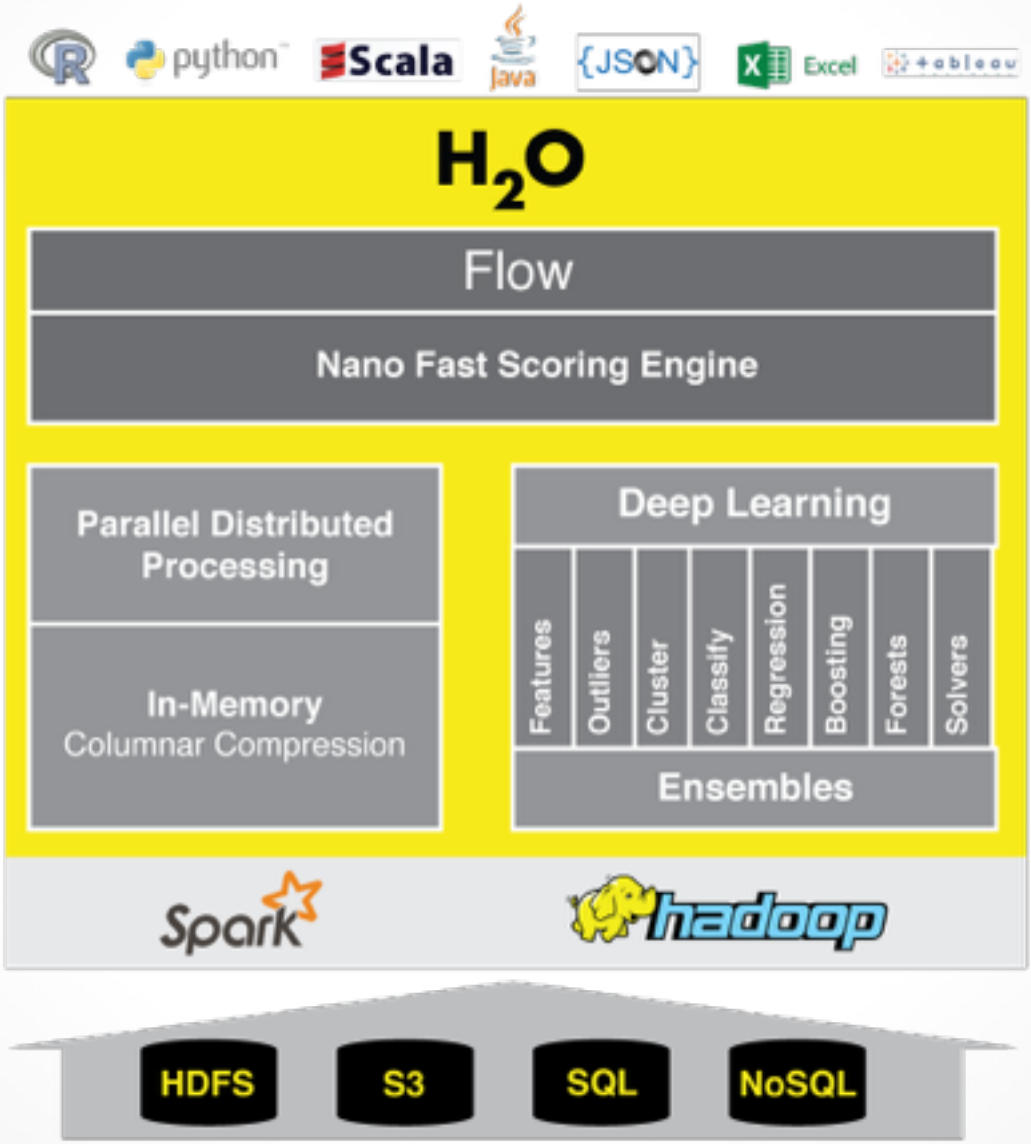
Table of Contents

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

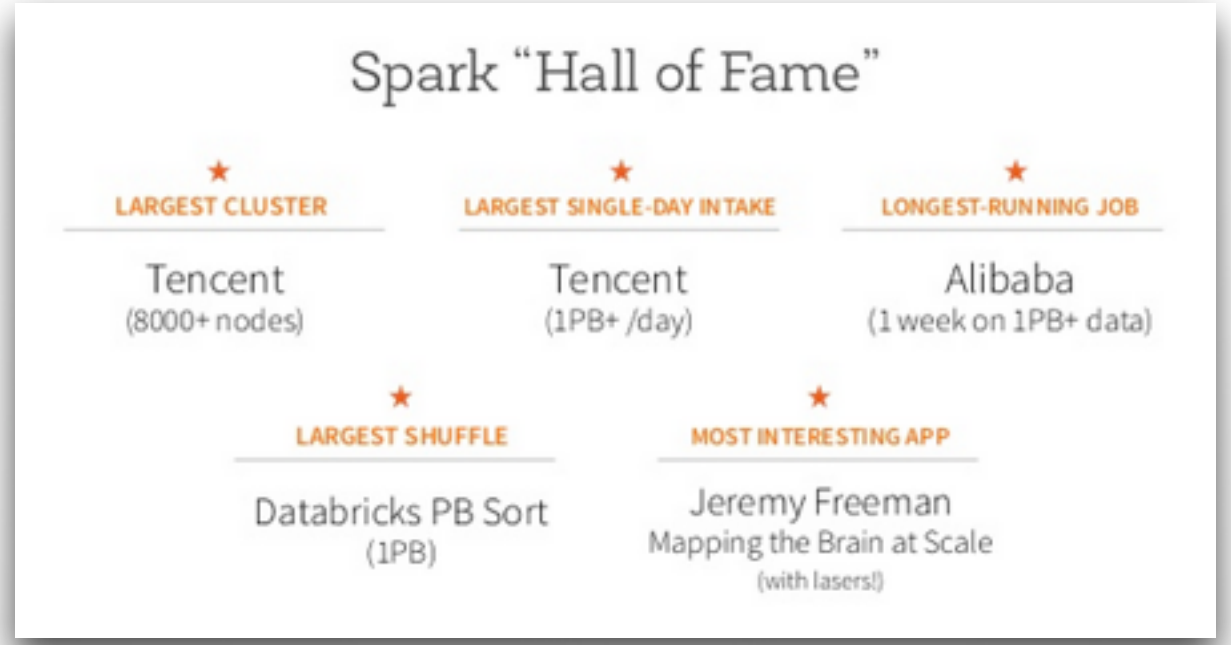
Algorithms on H2O

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



Spark + H_2O

SPARKLING WATER

What is Sparkling Water?

Provides

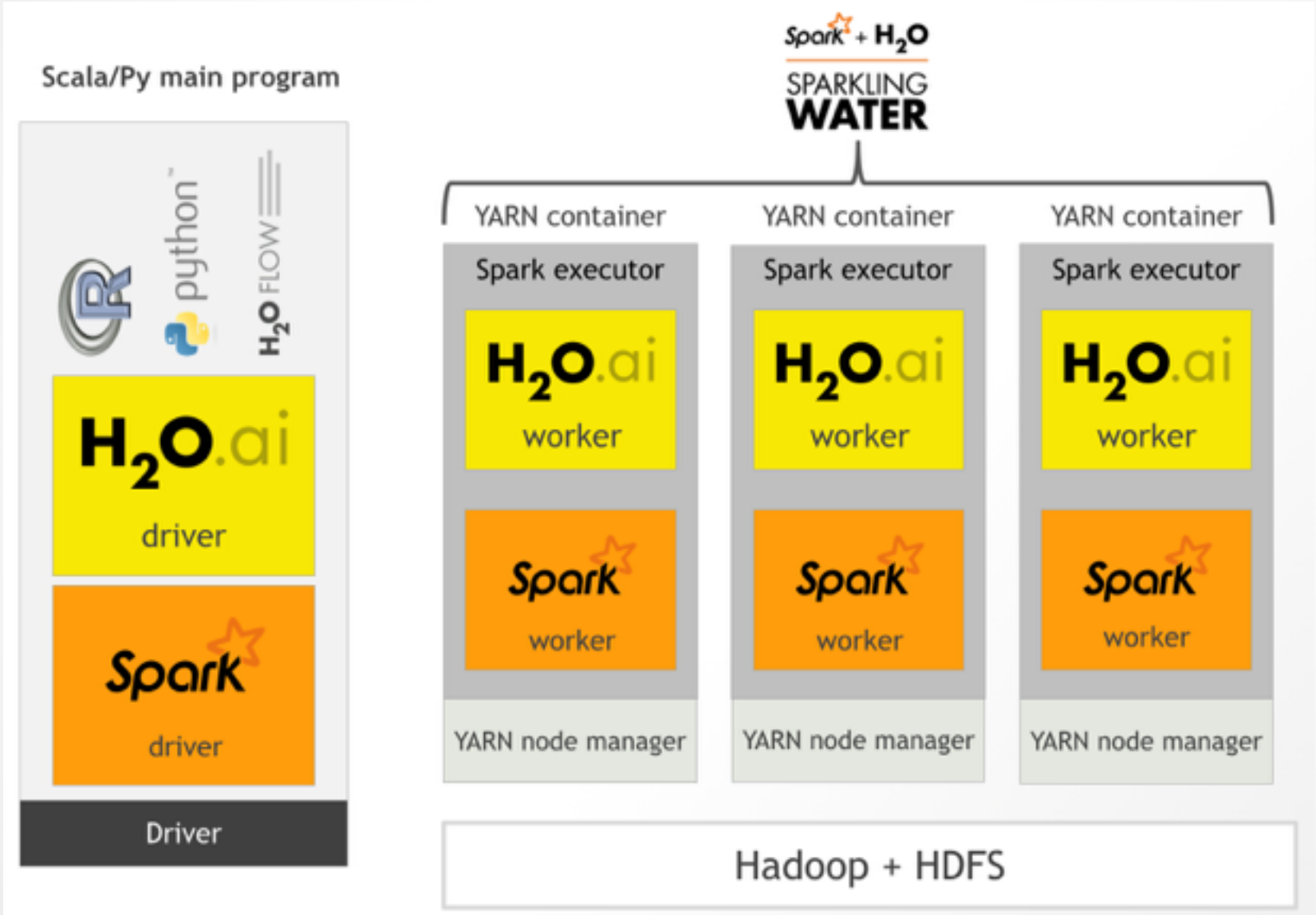
Transparent integration of H2O with Spark ecosystem

Transparent use of H2O data structures and algorithms with Spark API

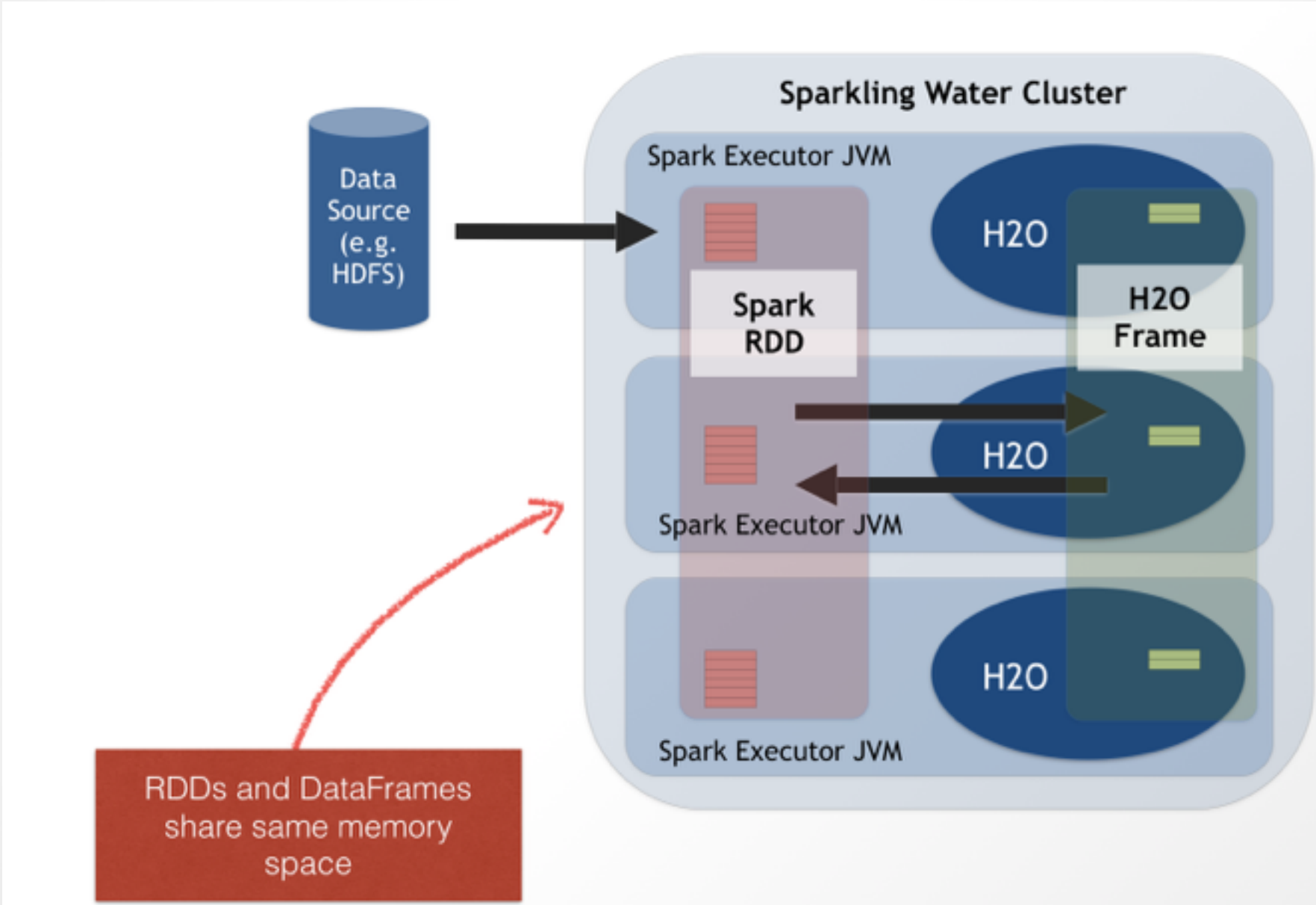
Excels in existing Spark workflows requiring advanced Machine Learning algorithms

Platform for building Smarter ML Applications

What is Sparkling Water?



Spark and H2O Data Sharing



Let's get started with installations!

Simple Spam Detector

OR

Detect spam text messages

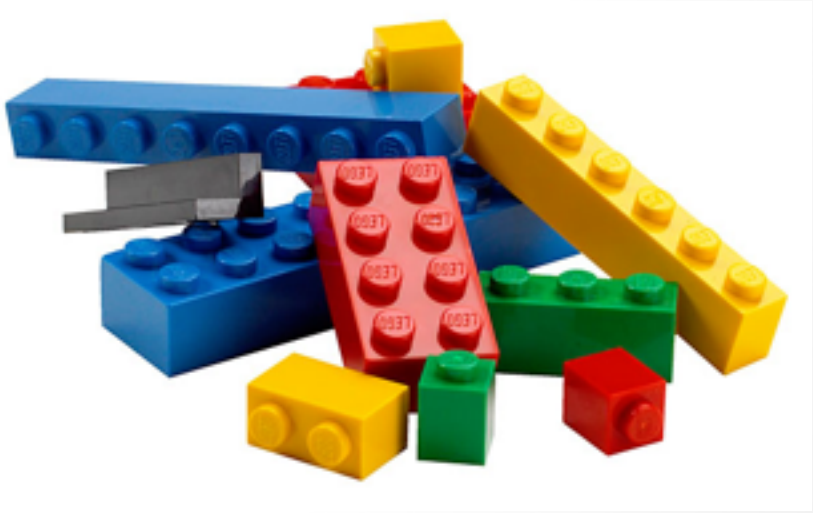
Data example

	A	B
1	ham	Ok... But they said i've got wisdom teeth hidden inside n mayb need 2 remove.
2	ham	U thk of wat to eat tonight.
3	ham	I dunno until when... Lets go learn pilates...
4	spam	Someonone you know is trying to contact you via our dating service! To find out who it could be call from your mobile or landline 09064015307 BOX334SK38ch
5	ham	Ok c u then.
6	spam	URGENT! We are trying to contact U. Todays draw shows that you have won a £800 prize GUARANTEED. Call 09050003091 from land line. Claim C52. Valid12hrs only
7	spam	Not heard from U4 a while. Call 4 rude chat private line 01223585334 to cum. Wan 2C pics of me gettin shagged then text PIX to 8552. 2End send STOP 8552 SAM xxx
8	ham	staff.science.nus.edu.sg/~phyhcmk/teaching/pc1323
9	ham	Thank god they are in bed!
10	ham	Hey tmr meet at bugis 930 ?
11	spam	You are a winner you have been specially selected to receive £1000 cash or a £2000 award. Speak to a live operator to claim call 087123002209am-7pm. Cost 10p
12	spam	URGENT! Your Mobile No. was awarded £2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM
13	spam	Loan for any purpose £500 - £75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or text back 'help'
14	ham	Haha... Sounds crazy, dunno can tahan anot...
15	spam	You have won ?spam 000 cash or a ?2,000 prize! To claim, call09050000327
16	ham	Sorry i din lock my keypad.
17	ham	Thanx but my birthday is over already.
18	spam	FREE for 1st week! No1 Nokia tone 4 ur mobile every week just txt NOKIA to 8077 Get txting and tell ur mates. www.getzed.co.uk POBox 36504 W45WQ 16+ norm150p/tone
19	spam	Congratulations - Thanks to a good friend U have WON the £2,000 Xmas prize. 2 claim is easy, just call 08712103738 NOW! Only 10p per minute. BT-national-rate
20	ham	Me n him so funny...
21	spam	pdate_Now - Double mins and 1000 txts on Orange tariffs. Latest Motorola, SonyEricsson & Nokia & Bluetooth FREE! Call MobileUpd8 on 08000839402 or call2optout/!YHL
22	ham	Ok...
23	ham	Yup no more already... Thanx 4 printing n handing it up.
24	ham	Anything lor. Juz both of us lor.
25	ham	It's é only $140 ard...É rest all ard $180 at least...Which is é price 4 é 2 bedrm ($900)
26	ham	Oh oh... Den muz change plan liao... Go back have to yan jiu again...
27	ham	Ok lor then we go tog lor...
28	ham	Okay lor... Wah... like that def they wont let us go... Haha... What did they say in the terms and conditions?
29	ham	Dunno lei... I thk mum lazy to go out... I neva ask her yet...
30	ham	THATS ALRITE GIRL, U KNOW GAIL IS NEVA WRONG!!TAKE CARE SWEET AND DONT WORRY.C U L8TR HUN!LOVE Yaxxx

ML Workflow

Goal: For a given text message identify if it is spam or not

1. Extract data
2. Transform, tokenize messages
3. Build Tf-IDF model
4. Create and evaluate Deep Learning model
5. Use the model to detect spam



Ask Craigs(list) Applications

Task: Predict the job category from a Craigslist Ad Title

Jun 8 Account Executive - (berkeley) map [x]

Jun 8 Motorsports Automotive Mechanic - (scotts valley) map [x]

Jun 8 Technology Support Specialist - (downtown / civic / van ness) [x]

Jun 8 Assistant Program Director-San Jose Place - (SOMA / south beach) img [x]

Jun 8 FRONT OFFICE MANAGER - (DOWNTOWN-CIVIC CENTER) [x]

Jun 8 Customer Service Representative - (San Francisco, CA) img [x]

Jun 8 ♛Organizers Needed: Protect CA Water NOW♛ - (oakland downtown) img [x]

Jun 8 Manager of OEM Business Development - (San Francisco, CA) img [x]

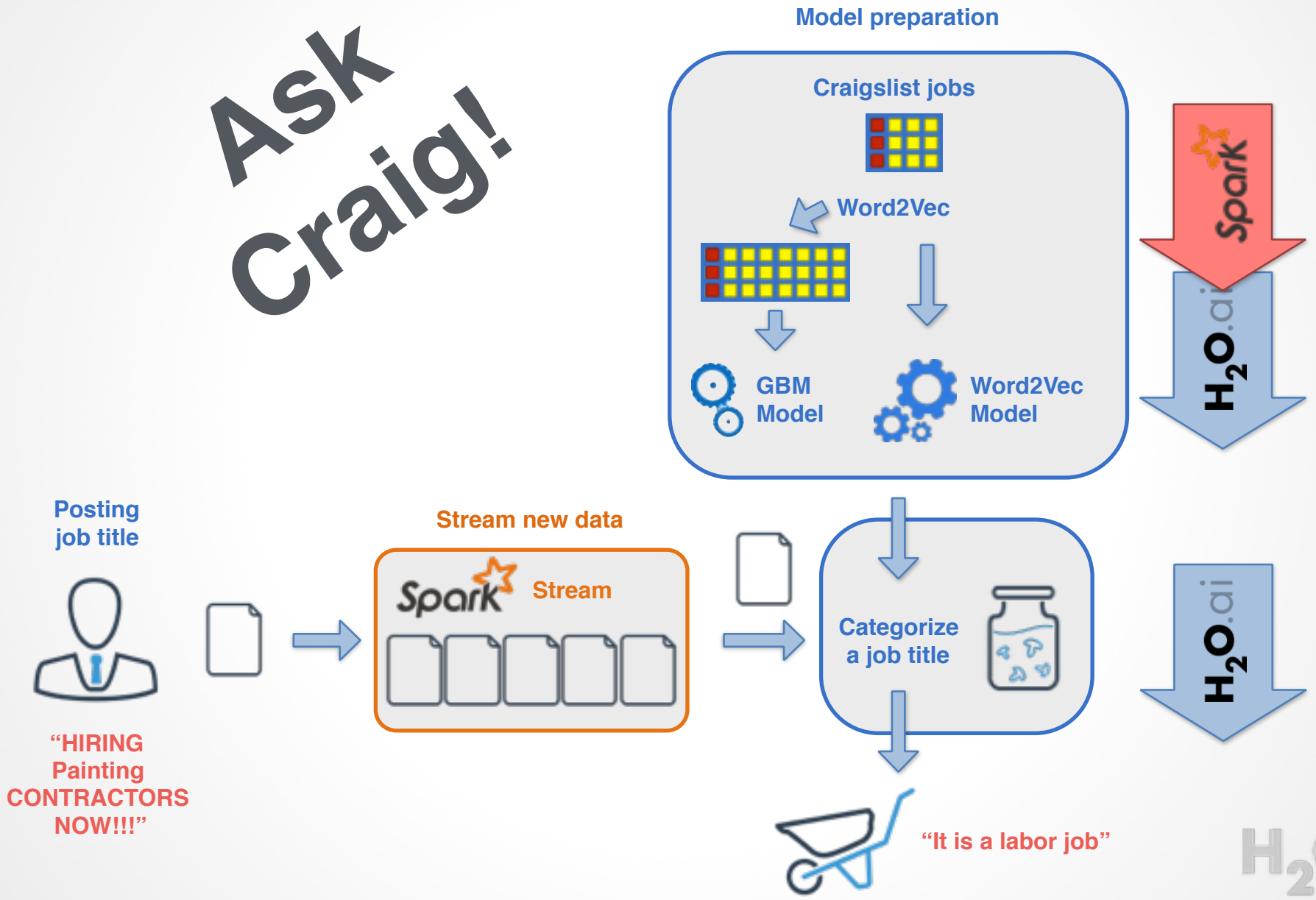
Jun 8 Seeking Rockstar EA/OM for Innovative Startup - (Palo Alto) pic [x]

Jun 8 Residential Counselor - Day & Swing Shifts - (SOMA / south beach) [x]

jobs

accounting+finance
admin / office
arch / engineering
art / media / design
biotech / science
business / mgmt
customer service
education
food / bev / hosp
general labor

Ask Craig!


Model preparation
Craigslist jobs
Word2Vec
GBM Model
Word2Vec Model
Spark
H2O.ai
Posting job title
"HIRING Painting CONTRACTORS NOW!!!"
Stream new data
Spark
Stream
Categorize a job title
H2O.ai
"It is a labor job"


ENOUGH TALK
SHOW ME A DEMO!!!
memegenerator.net