Applying Machine Learning using H2O.

Ashrith Barthur, PhD
Security Scientist

Overview

- Introduction to H2O
- H2O architecture
- Agenda
- Problem Definition
- Setup of H2O
- Data Analysis
- Results

Company Overview

Company	• Team: 80. 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

H_2O.ai
Machine Intelligence

Executive Team

Sri Satish Ambati

CEO & Co-founder

DataStax

Tom Kraljevic

VP of Engineering

Abrizio, Intel

Arno Candel

Chief Architect

HPC, CERN

Board of Directors

- Jishnu Bhattacharjee // Nexus Ventures
- Ash Bhardwaj // Flextronics

Scientific Advisory Council

- Trevor Hastie
- Stephen Boyd
- Rob Tibshirani

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, *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*

What is H2O?

Math Platform Open source in-memory prediction engine

- Parallelized and distributed algorithms making the most use out of multithreaded systems
- GLM, Random Forest, GBM, PCA, etc.

API Easy to use and adopt

- Written in Java – perfect for Java Programmers
- REST API (JSON) – drives H2O from R, Python, Excel, Tableau

Big Data More data? Or better models? BOTH

- Use all of your data – model without down sampling
- Run a simple GLM or a more complex GBM to find the best fit for the data
- More Data + Better Models = Better Predictions

Algorithms on H2O

Supervised Learning

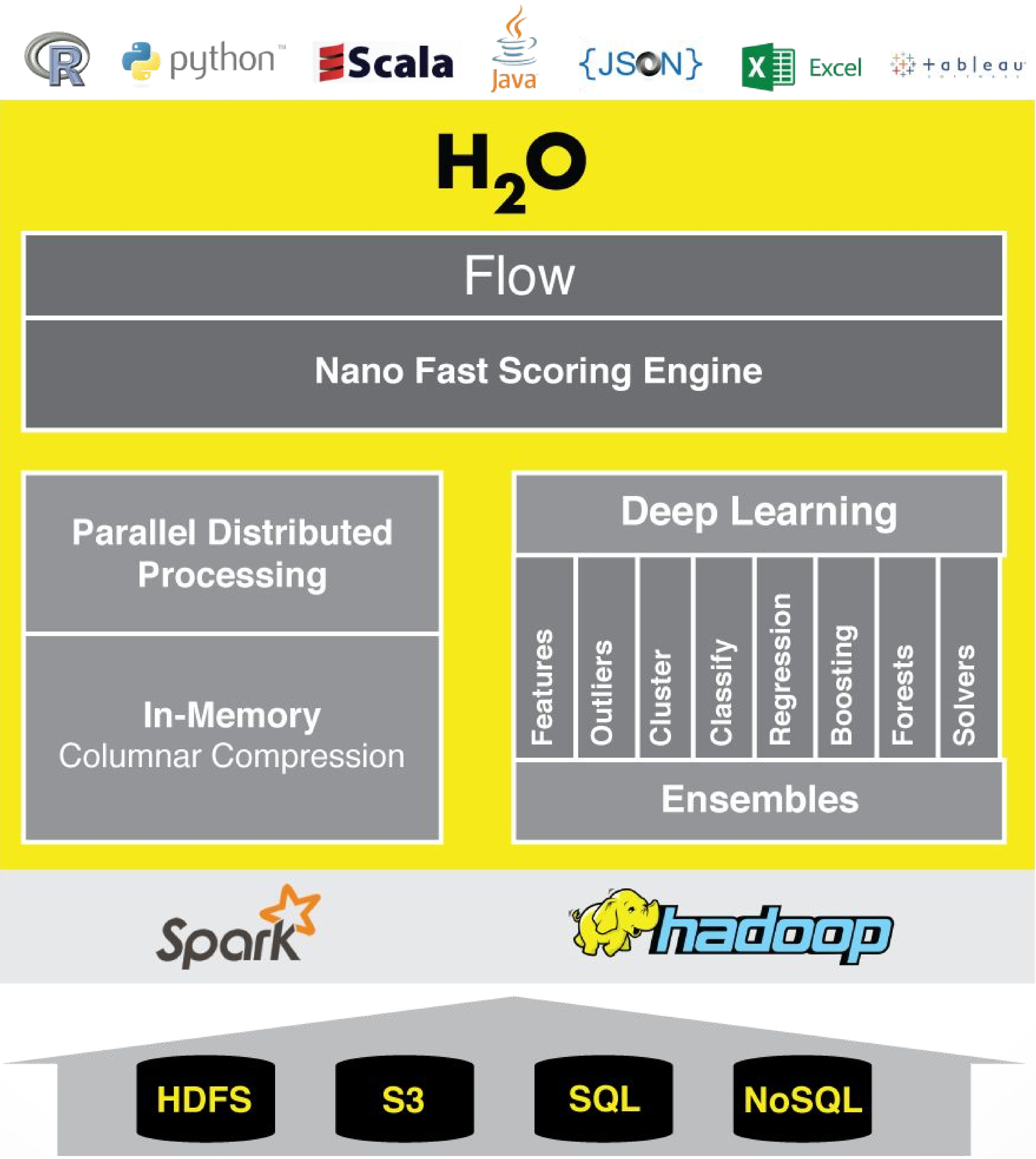
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
Deep Neural Networks	• **Deep learning**: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

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

Accuracy with Speed and Scale



H_2O.ai
Machine Intelligence

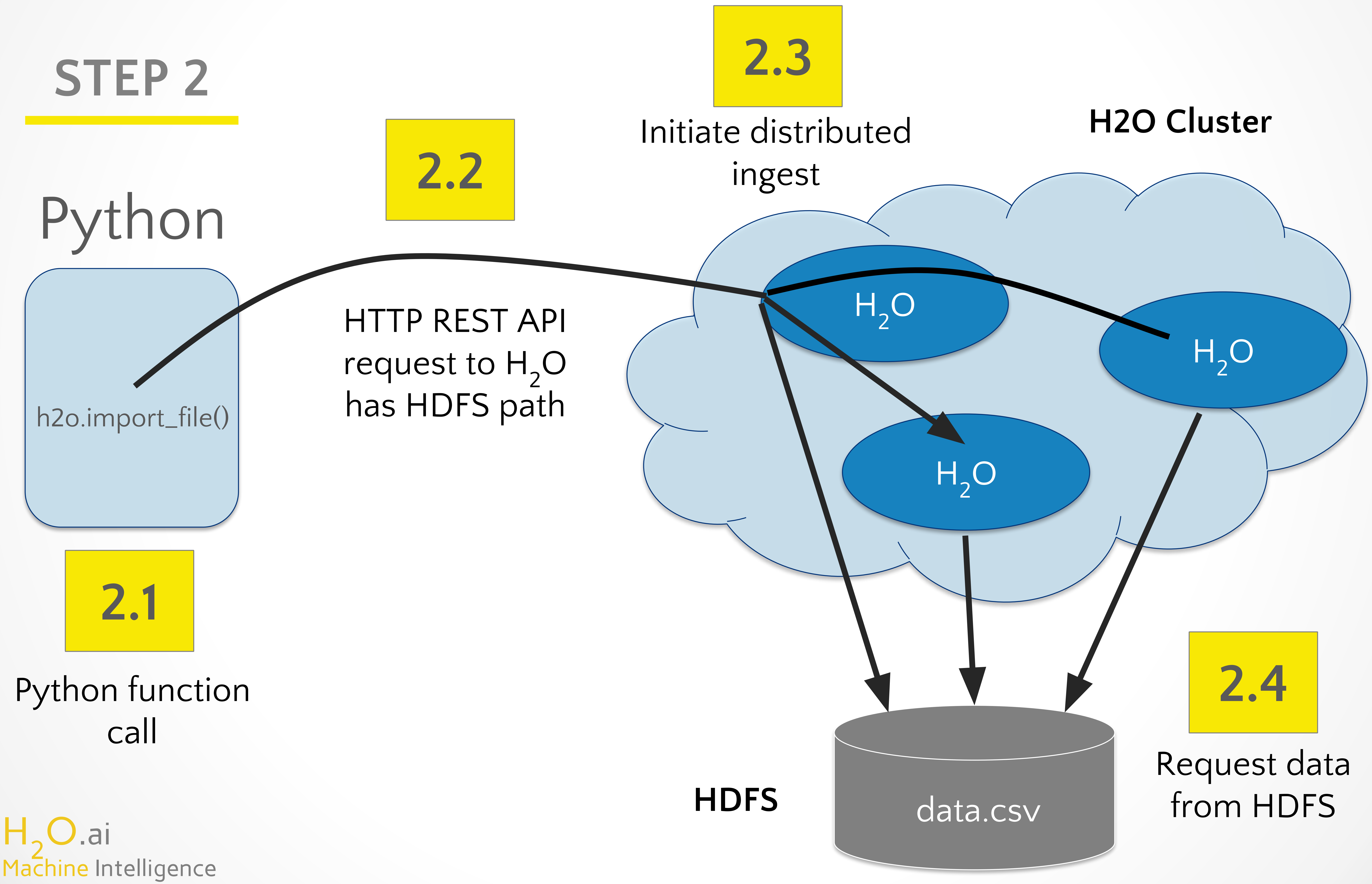
Reading Data into H2O with Python

STEP 1

Python user

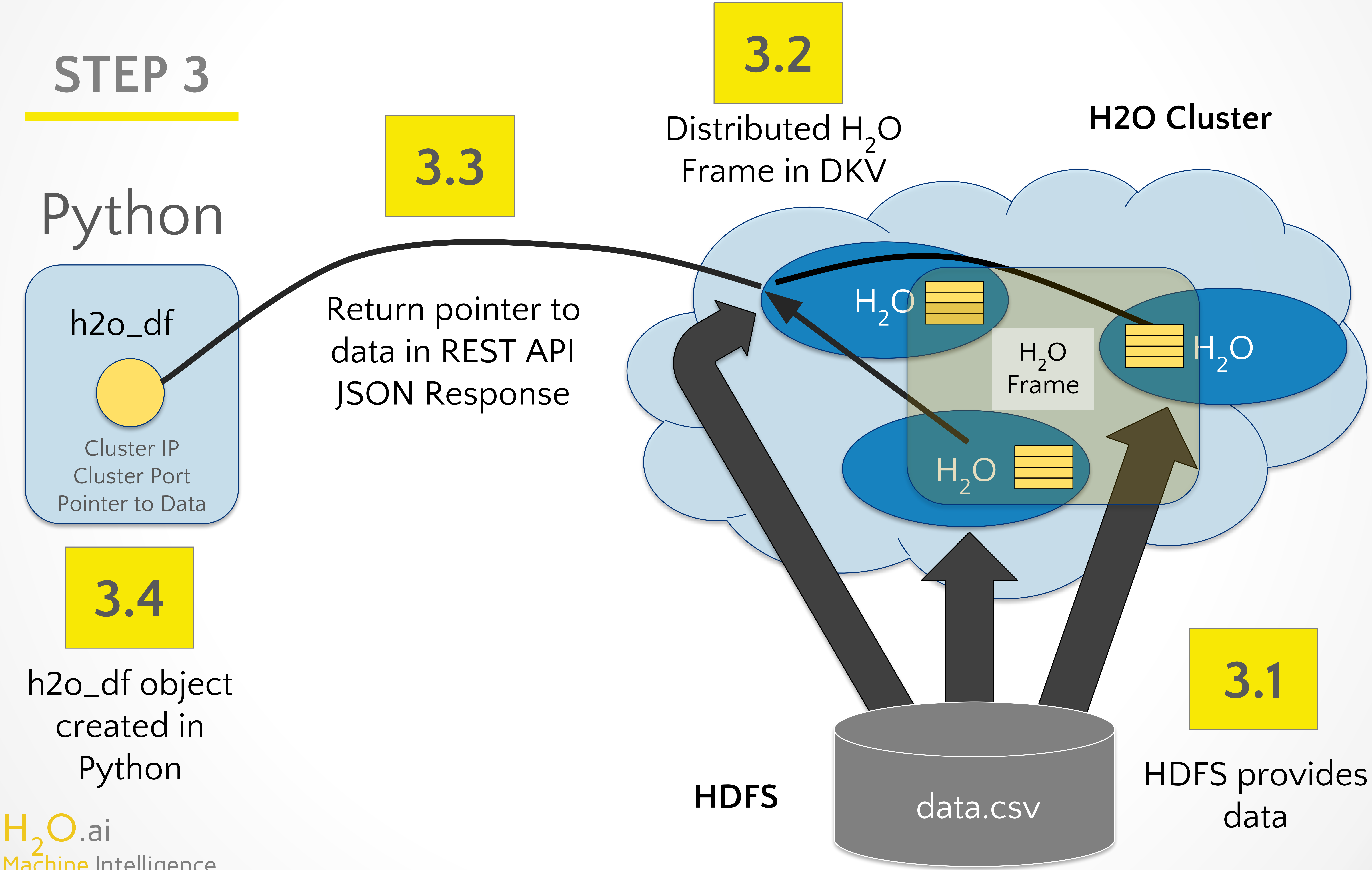
```
h2o_df = h2o.import_file("../data/allyears2k.csv")
```

Reading Data from HDFS into H2O with Python



H2O.ai
Machine Intelligence

Reading Data from HDFS into H2O with Python



H_2O.ai
Machine Intelligence

Agenda

- Understand the architecture of H2O
- Use H2O with Python through the REST API.
- Shape the data for analysis
- Model the data
- Read the results

Problem Definition

Sorting a large pool of system event logs by their source.

Problem Breakdown

- Every small and large company stores aggregated system event logs.
- Log storage is cheap and almost a requirement for compliance.
- Identifying source of logs from aggregated system logs is tough due to multiple devices, and networks.
- Breaking down these logs into their **respective source bins** is important for **security event analysis**.
- We use **machine learning** to show how this can be done.

Sample Data

<30>Feb 23 10:02:24 10.14.3.101 **dhcpd**[10039]: **balanced pool** 81df410 10.6.28.0/23 total 495 free 278 backup 186 lts 46 max-misbal 70
<166>2016-02-23T13:02:24.363Z server1.example.com **Vpxa**: [42764B90 verbose 'VpxaHalCnxHostagent' opID=WFU-e1f63e34] [WaitForUpdatesDone] Received callback
<166>2016-02-23T13:02:24.363Z server1.example.com **Vpxa**: [42764B90 verbose 'hostdvm' opID=WFU-e1f63e34] [VpxaHalVmHostagent] 4357: GuestInfo changed 'guest.disk'
<30>Feb 23 10:02:24 10.0.2.108 dhcpd[2331]: DHCPACK to **10.0.0.58** (**4c:34:88:fa:f9:a1**) via eth2
<30>Feb 23 10:02:24 10.0.2.108 dhcpd[2331]: DHCPACK to 10.0.0.58 (4c:34:88:fa:f9:a1) via eth2

Study of Logs

- The logs have a data format
- There are IP addresses in the logs
- There are also MAC addresses in the logs
- Logs from same source have the same word(s). Hence, similar sized words.
- Prival or Priority value is a valuable information to identify if the event is coming from a system daemon or other sources.
- Prival also tells us the facility and severity of the source. (Refer RFC 5424)

Features

- Priority Value
- Date Format
- Number of Characters
- Number of IP addresses
- Number of MAC addresses
- Fraction of dictionary words by size.

Labeling

Supervised approach

Requires model to understand response

Labeling Example

```
my_dict = {
'rsa.ims.authn':'securid.txt',
'\%ASA' :'ciscoasa.txt',
'\[acc\]':'bigip-vpn.txt',
'from\ssensor':'airmagnet.txt',
'CPPM\_':'arubanetworks.txt',
'Vpxa' :'vmware.txt',
'SecureSphere' :'waf.txt'
}
```

Data for Modeling

30;5;138;2;0;network.txt;0.4;0.2;0.2;0.0;0.2;0.0;0.0;0.0;0.0;0.0;0.0;0.0

166;4;166;0;0;vmware.txt;0.0;0.0;0.0;1.0;0.0;0.0;0.0;0.0;0.0;0.0;0.0;0.0

166;4;173;0;0;vmware.txt;0.333333333333;0.333333333333;0.0;0.333333333333;0.0;0.0;0.0
;0.0;0.0;0.0;0.0;0.0

Acknowledgements

- Hanif Zachary, CapitalOne
- Avni Wadhwa, H2O.ai
- Mark Chan, H2O.ai

Thank You

Questions?