



# Getting Your Feet Wet with Driverless AI

Michelle Gabriele Tanco

May 7<sup>th</sup>, 2019

# AGENDA

- H2O.ai and Machine Learning overview
- What is Automated Machine Learning and why do I care?
- Driverless AI in action
  - **Predicted amount:** How much will this vehicle sell for?
  - **Predicted class:** Who will default on their loan?
  - **Predicting future sales:** How much will each store make over the next year?
- What to do next

# H2O.ai Overview

H<sub>2</sub>O.ai

<b>Company</b>	Founded in Silicon Valley in 2012 Funded: \$75M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
<b>Products</b>	<ul style="list-style-type: none"><li>• H2O Open Source Machine Learning (14,000 organizations)</li><li>• H2O Driverless AI – Automatic Machine Learning</li></ul>
<b>Team</b>	130 AI experts (Expert data scientists, Kaggle Grandmasters, Distributed Computing, Visualization)
<b>Global</b>	Mountain View, London, Prague, India



## Open Source



In-memory, distributed  
machine learning algorithms  
with H2O Flow GUI



H2O AI open source engine  
integration with Spark



Lightning fast machine  
learning on GPUs

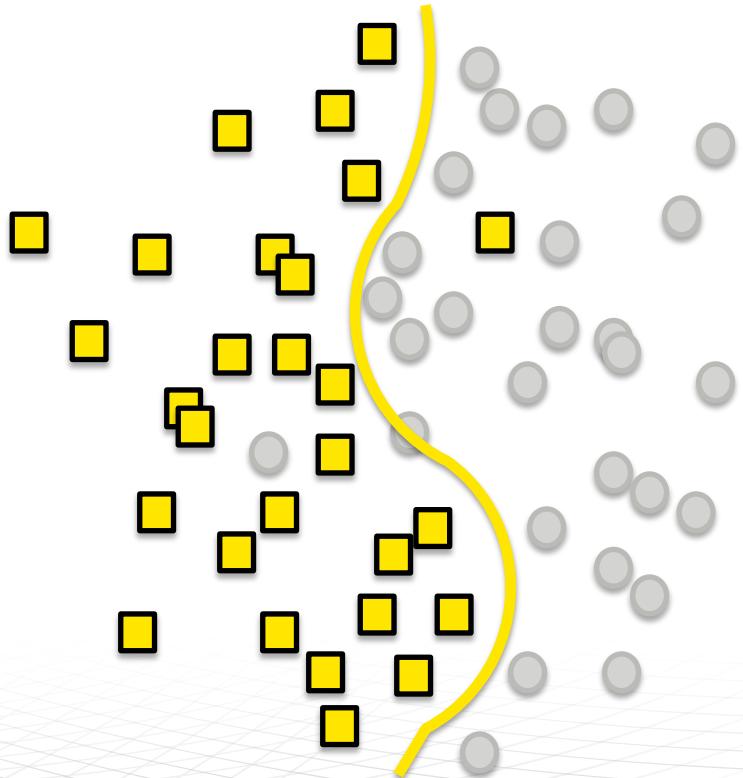
- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise support subscriptions

DRIVERLESSAI

Automatic feature engineering,  
machine learning and interpretability

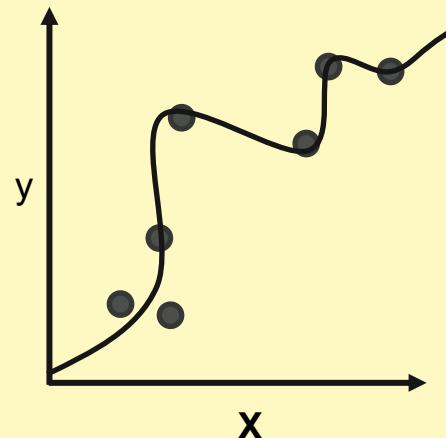
- Enterprise software
- Built for domain users, analysts and data scientists – GUI-based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

# What is Machine Learning?

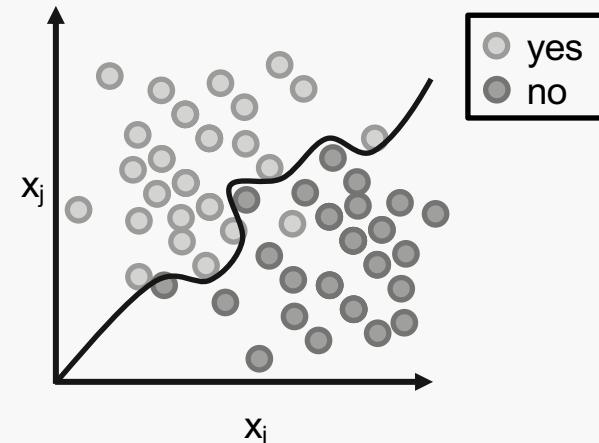


# Supervised Learning

**Regression:**  
How much will a customer spend?



**Classification:**  
Will a customer churn?



# Stacked Ensembles

$$\begin{bmatrix} y \end{bmatrix} \approx \begin{bmatrix} \hat{y}_l \end{bmatrix} = f_l \left( \begin{bmatrix} x_1 \dots x_p \end{bmatrix} \right)$$

$$\begin{bmatrix} y \end{bmatrix} \approx g \left( \begin{bmatrix} \hat{y}_1 \dots \hat{y}_L \end{bmatrix} \right)$$

- Engineer original feature set
- Train L models using strong learners
- Control overfitting using k-fold CV

- Stack these predicted values to form new feature set
- Train meta learner on stacked predictions

# Preventing Over Fitting

## Training Set vs. Test Set

- Partition the original data (randomly) into a training set and a test set. (e.g. 70/30)
- Train a model using the training set and evaluate performance (a single time) on the test set.



## K-fold Cross-validation

- Train and test K models using separate folds.
- Average the model performance over the K test sets
- Report cross-validated metrics.

# Machine Learning is Applicable to Every Industry



Insurance



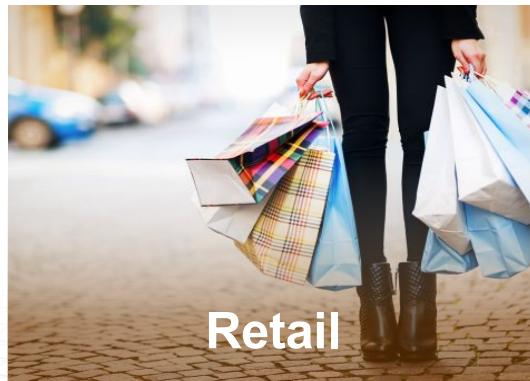
Manufacturing



Financial Services



Healthcare

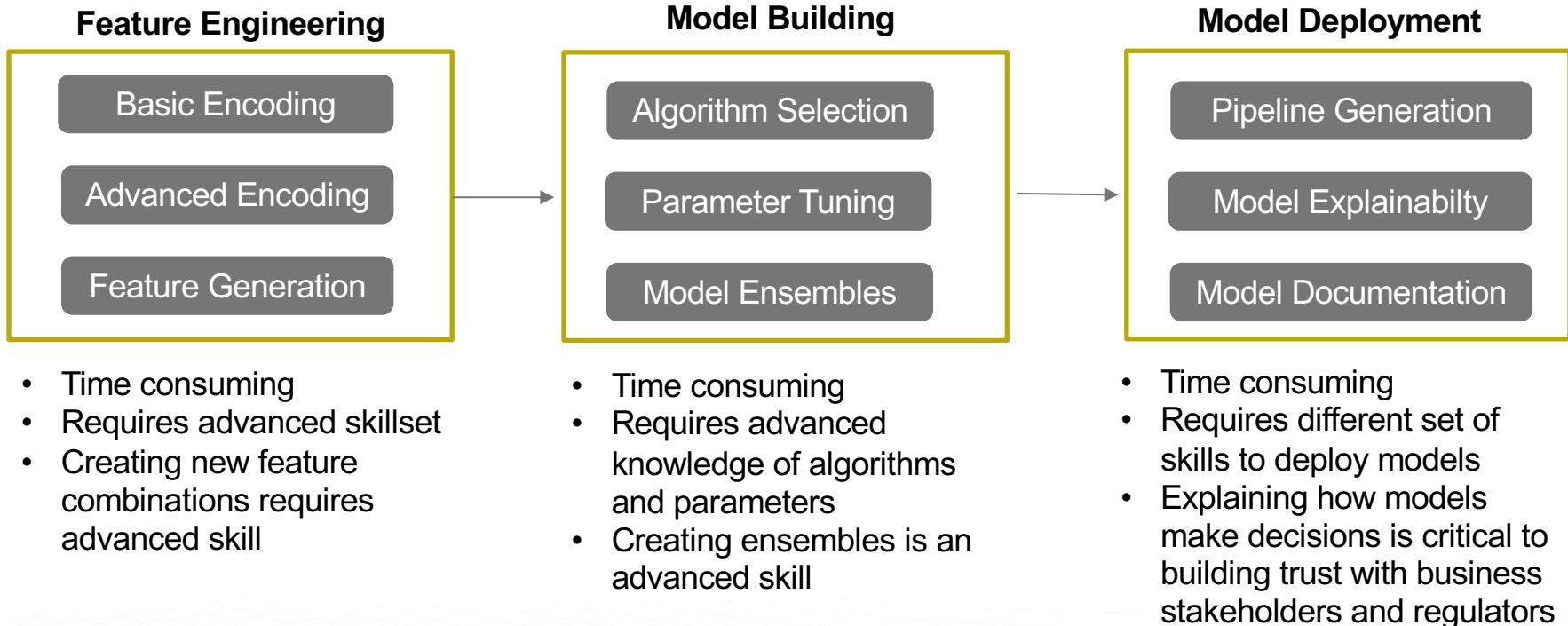


Retail



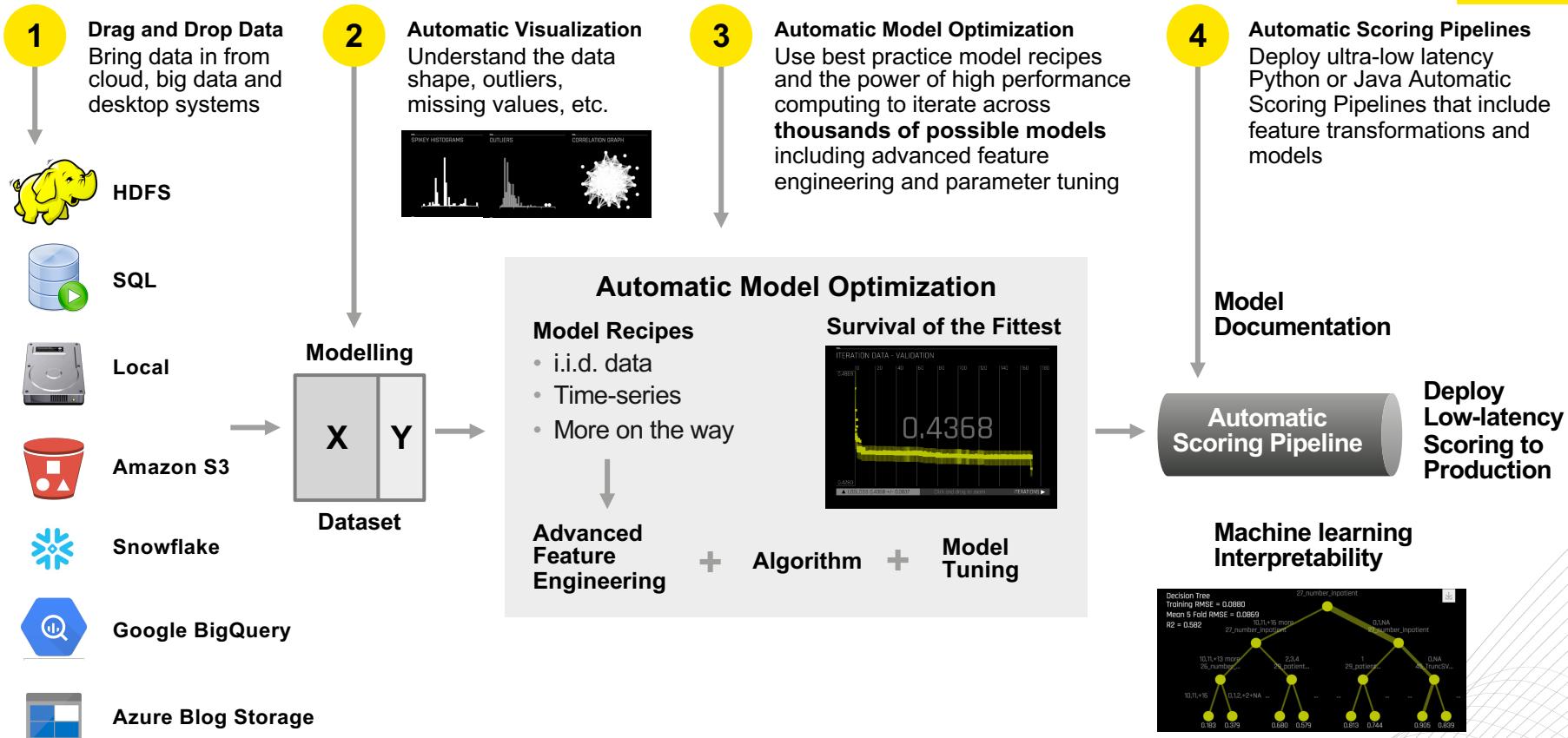
Ad Tech / MarTech

# Challenges in Model Development



**The entire process is highly iterative and can take weeks or months to develop a single production-ready model.**

# H2O Driverless AI – How it Works



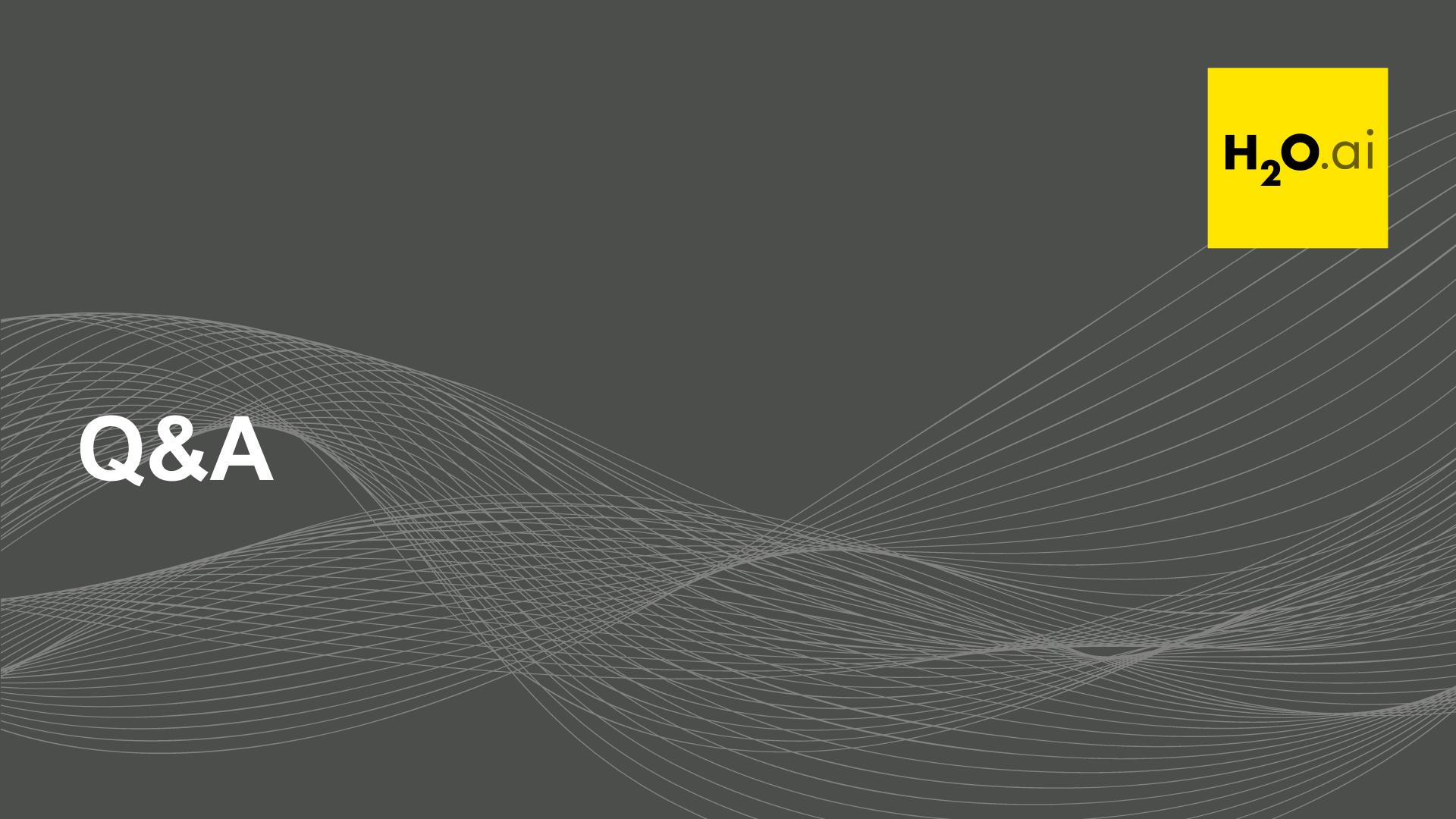
# Driverless AI Demo

- Ingesting and viewing a dataset
- Using AutoViz to understand data and relationships
- Running an experiment to build the optimal predictive model
- Understanding the model globally, and the reasoning behind each prediction
- Deploying to production
- Running Driverless AI from python or R



# 5 Steps to get started with H2O Driverless AI

- Join the [community slack](#) for Questions/Answers, tips and news
- Get a 21-day Driverless AI trial license:
  - [Request a 21-day license](#)
  - [Cloud Test Drive](#) – two hour session with no install required
- [Install Driverless AI](#)
  - To a local machine or server
  - In any major cloud
- Do [the tutorials](#) – guided introduction with sample data
- Learn more from:
  - [H2O World SF session replays](#)
  - [H2O Driverless AI docs](#)



**H<sub>2</sub>O.ai**

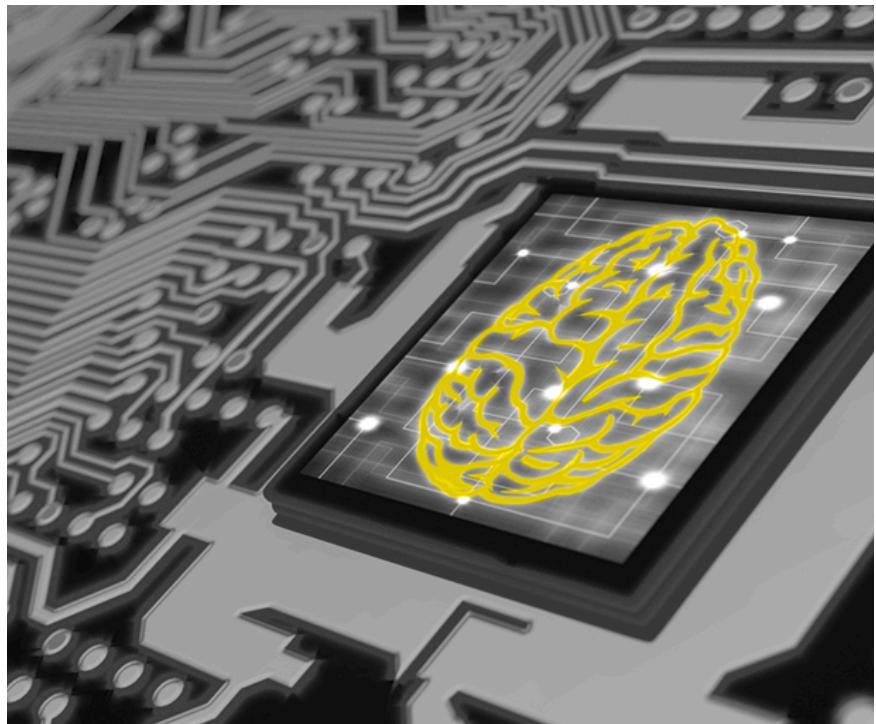
# Q&A



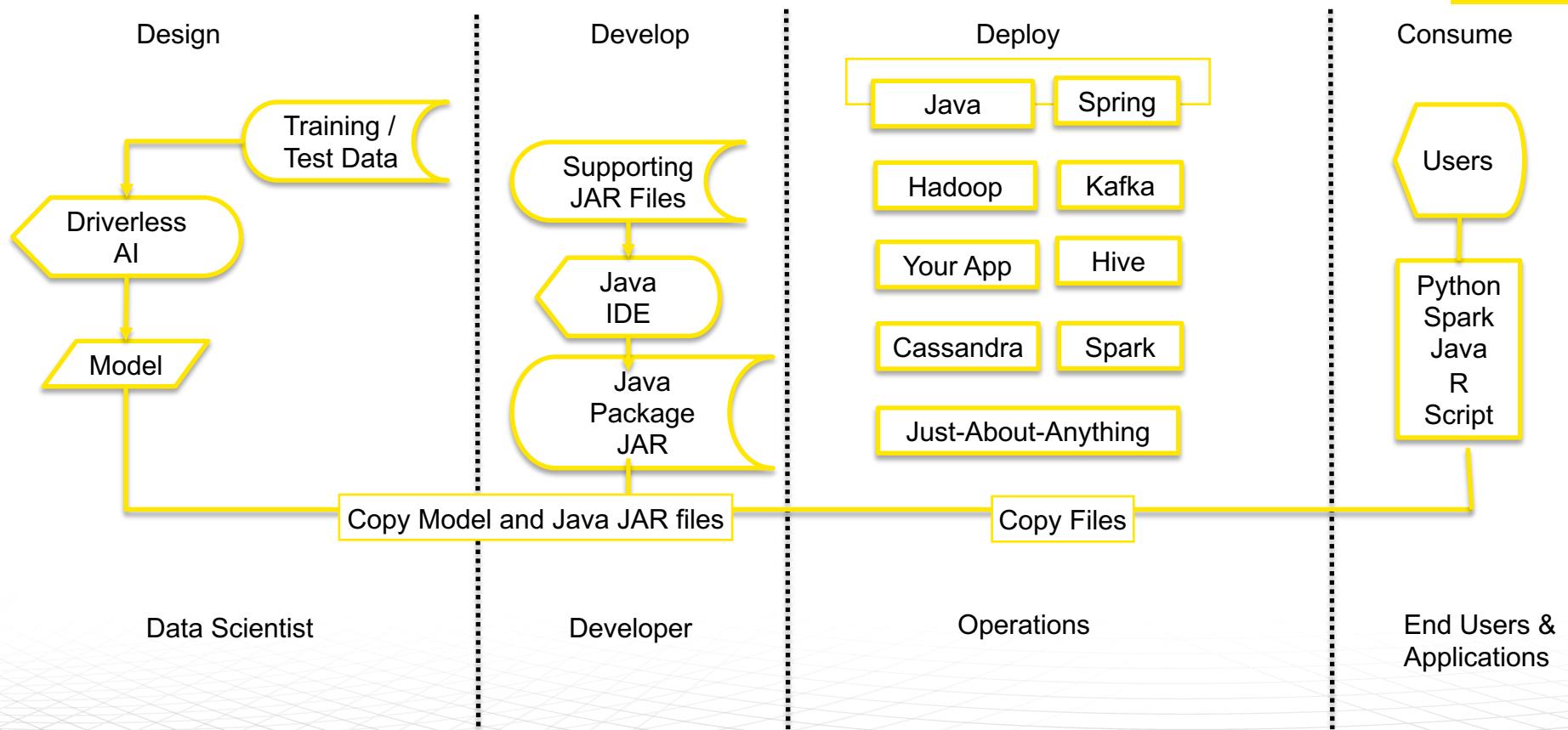
# Appendix: DAI

# Easy Deployment for Low-latency Models

- Stand-alone scoring pipeline that is easy for IT to deploy and manage including all feature transformations
- Easy to update when a new model version is available
- Optimized the scoring code for complex (big data) models
- Streamlined scoring code to deploy on any device on the edge, mobile, etc.
- Really fast (millisecond) to satisfy today's real-time apps
- REST Server to further automate deployment



# Life Cycle Flow Diagram





# Appendix: H2O-3

# H2O Open Source AI Platform

H<sub>2</sub>O.ai

## Open Source



100% open source

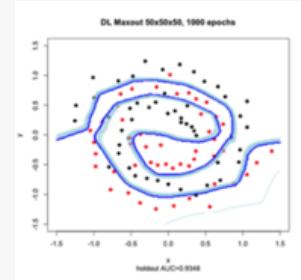
## Big Data Ecosystem



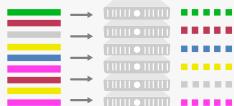
## Flexible Interface



## Smart and Fast Algorithms



## Scalability and Performance



- Distributed in-memory computing platform
- Distributed algorithms
- Fine-grain MapReduce

## Rapid Model Deployment

- Highly portable models deployed in Java (POJO) and Model Object Optimized (MOJO)
- Automated and streamlined scoring service deployment with Rest API



## GPU Enablement



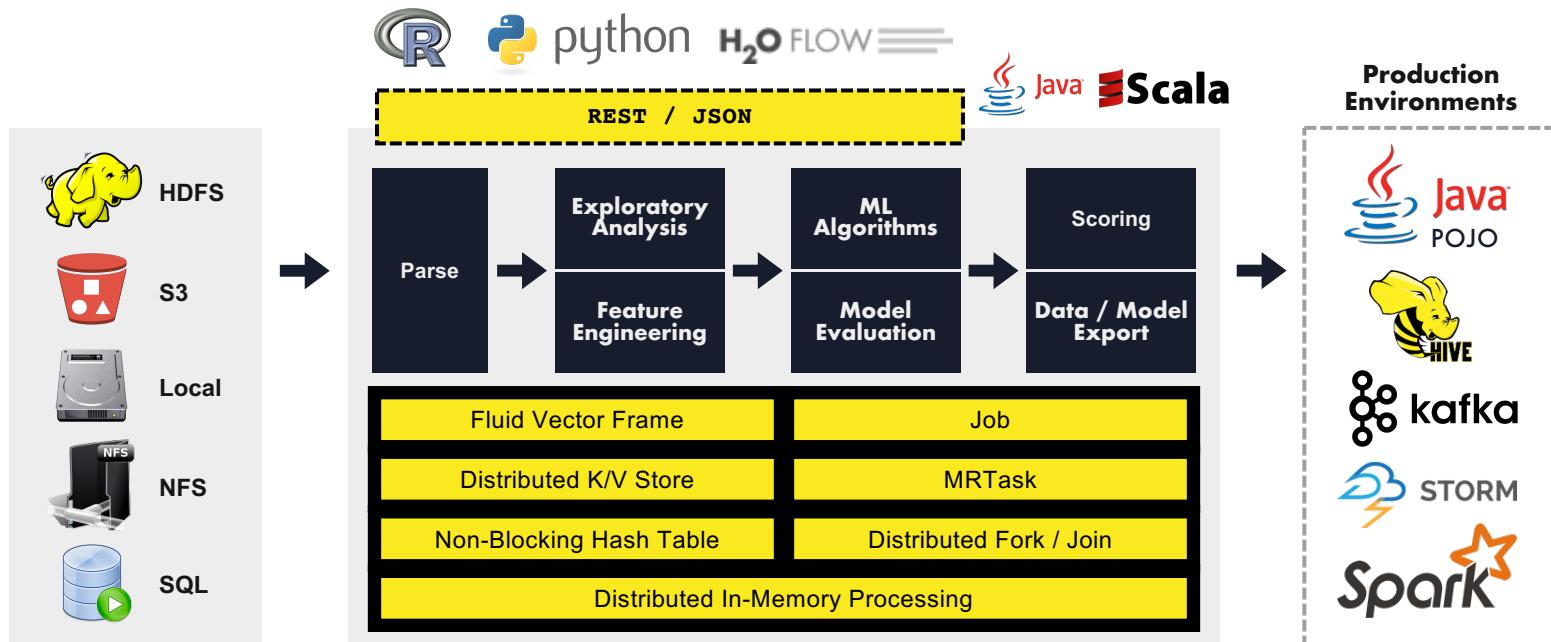
**H2O4GPU**

## Cloud Integration



# H2O.ai High Level Architecture

H<sub>2</sub>O.ai



cloudera  
MAPR

Hortonworks

Spark + H<sub>2</sub>O  
SPARKLING  
WATER