Productionizing H2O Models with Apache Spark

Big Data Conference Vilnius
Vilnius, Lithuania, November 28-29 2018

Spark + H_2O

SPARKLING WATER

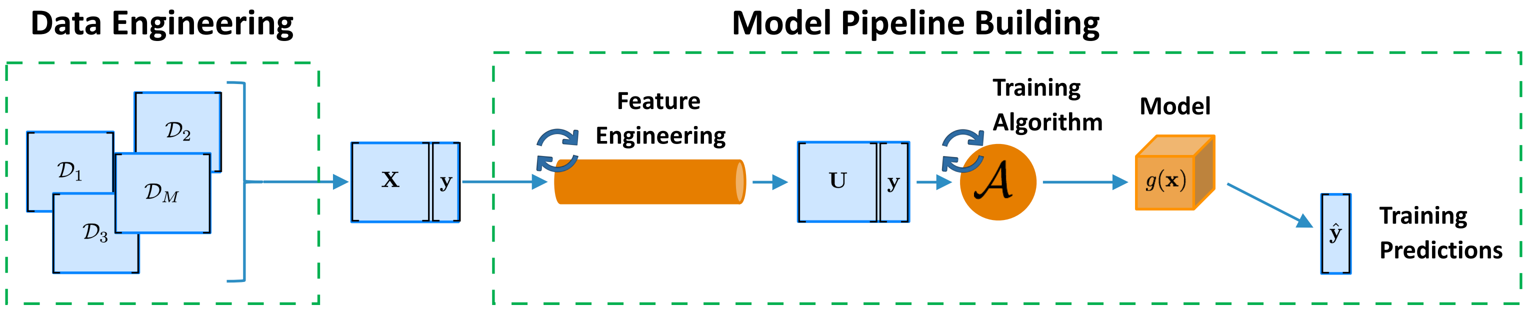
Jakub Háva,
jakub.hava@h2o.ai
https://github.com/jakubhava
https://www.linkedin.com/in/havaj/

Who are we?

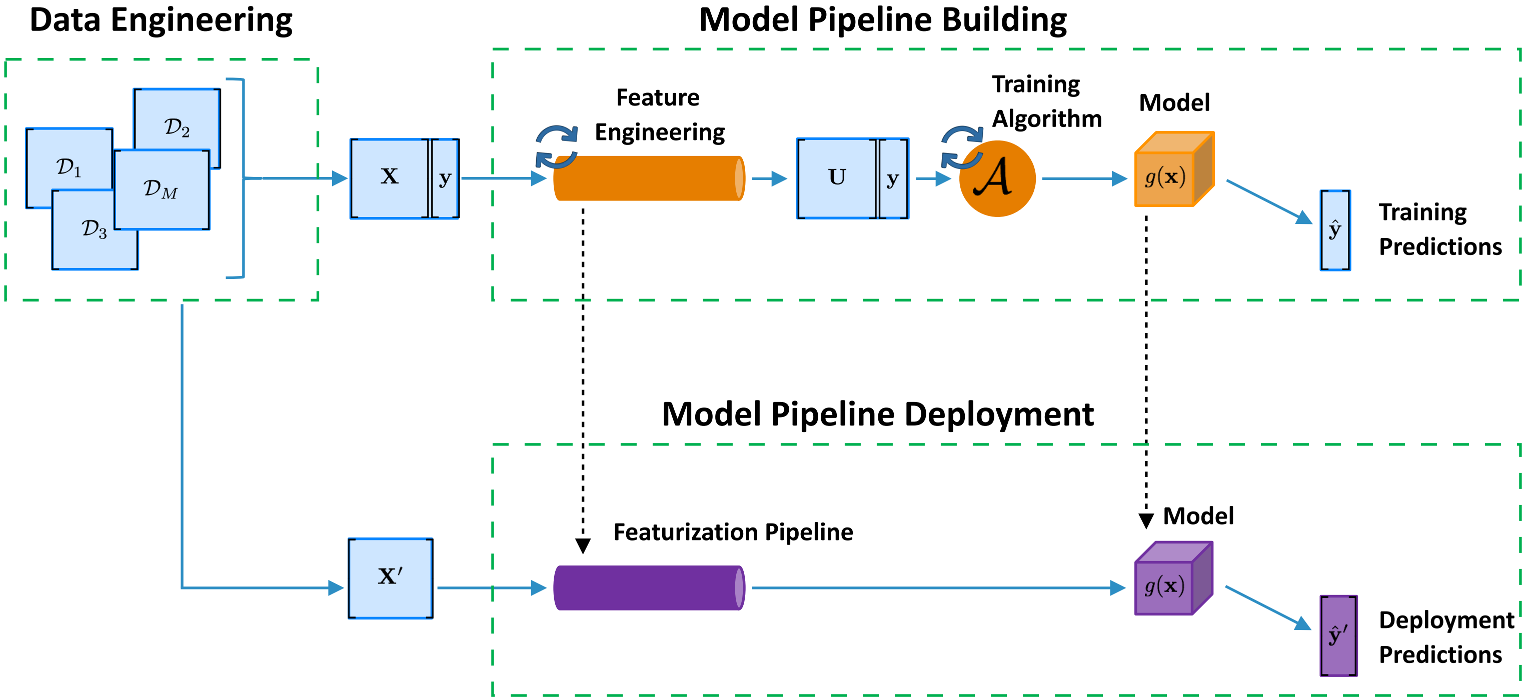
- **Kuba**
  - Senior Software engineer at H2O.ai - Core Sparkling Water
  - Master's at Charles University (CZ)
  - Implemented high-performance cluster monitoring tool for JVM based languages (JNI, JVMTI, instrumentation)
- **Michal**
  - VP of Engineering at H2O.ai
  - Creator of Sparkling Water
  - Ph.D at Charles University (CZ), PostDoc at Purdue (US)

Machine Learning (ML) Lifecycle

Basic ML Lifecycle



Basic ML Lifecycle



Example Implementations

Model Building			Model Deployment	
Data Engineering	**Feature Engineering**	**Training Algorithm**	**Deployment Pipeline**	**Model**
Spark		H2O	Spark	H2O MOJO
Spark	H2O Driverless AI		Spark	H2O Driverless AI MOJO

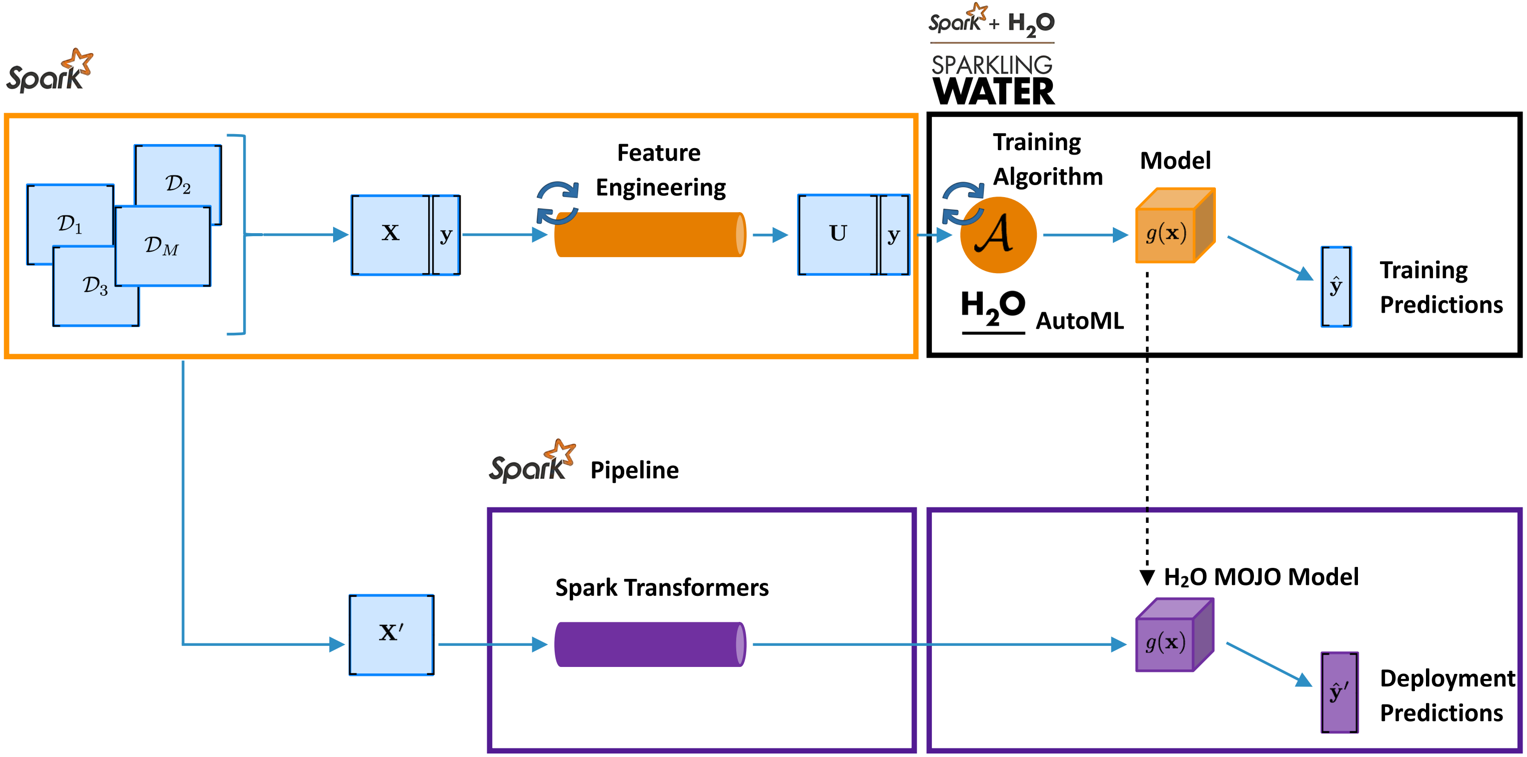
H2O + Spark = Sparkling Water

H2O + Spark

- H2O
    - Machine Learning Library
    - Distributed Algorithms
    - For ML experts

- Sparkling Water
    - Integrates H2O & Spark Ecosystems
    - Transparent for Spark users
    - Based on Spark pipelines & H2O

Basic ML Lifecycle: Sparkling Water



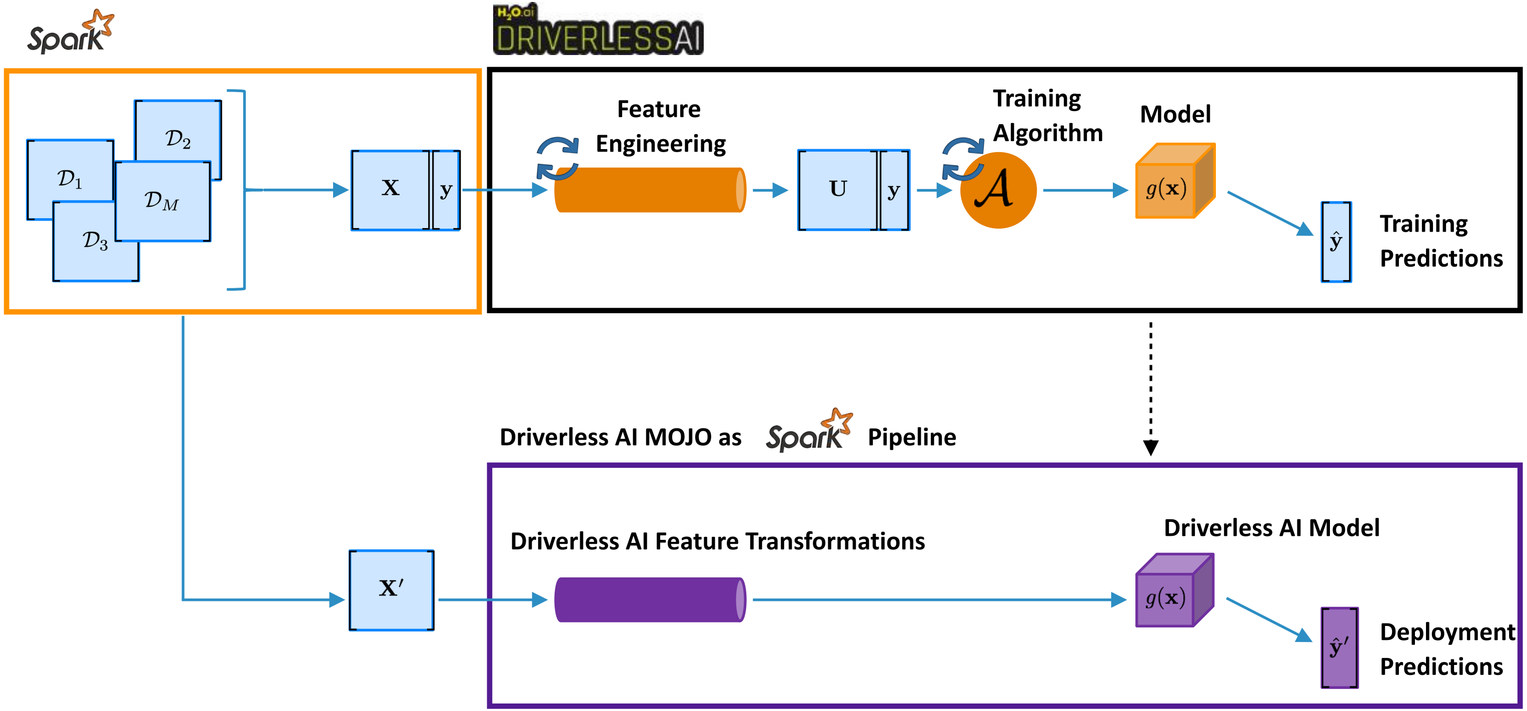
Demo:
Spark Pipeline

H2O Driverless AI

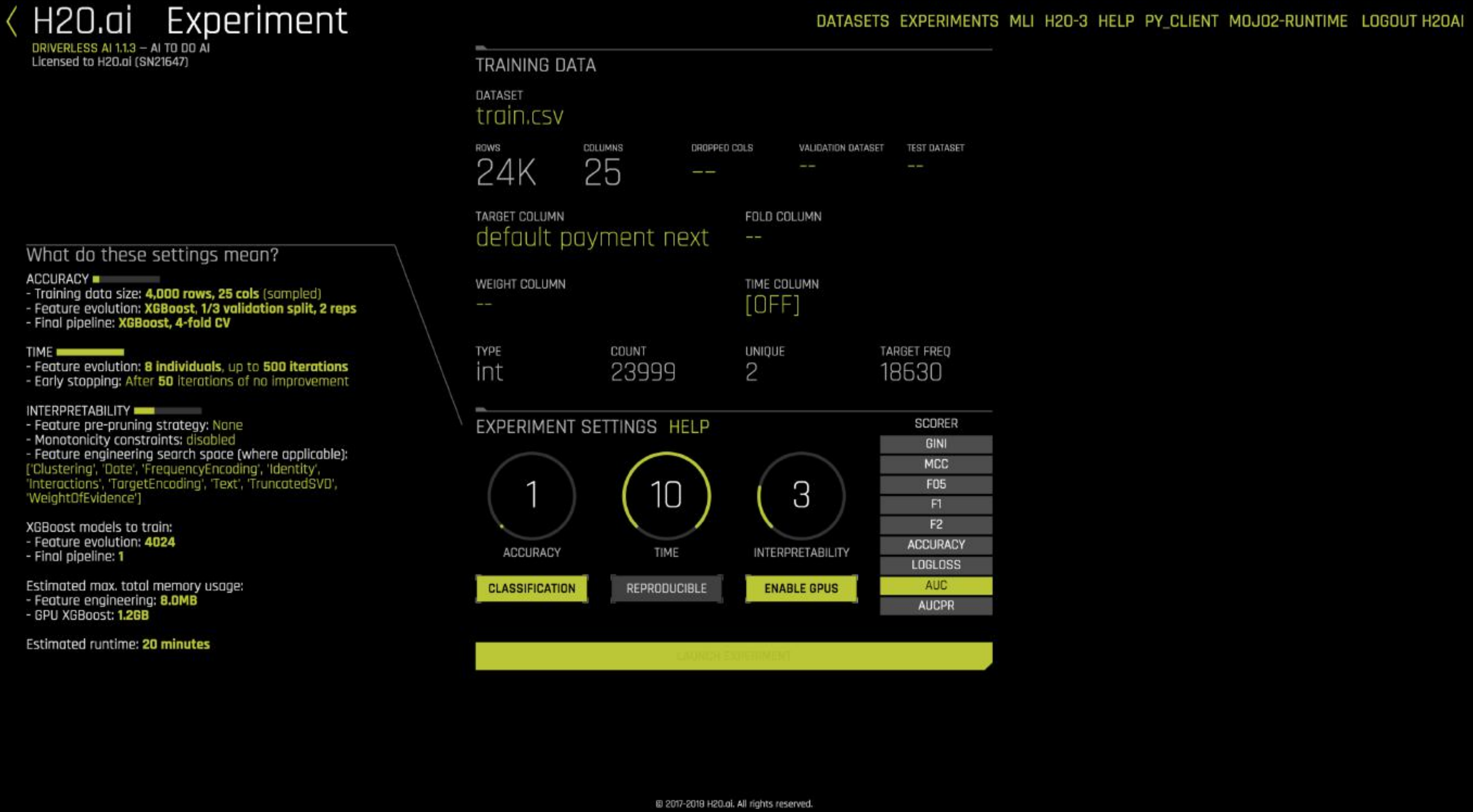
H2O Driverless AI

- What if I’m not expert ?
  - H2O Driverless AI
- H2O Driverless AI
  - No expert knowledge required
  - Automatic **Feature Engineering** & ML

Basic ML Lifecycle: Driverless AI



Demo: Driverless AI as Spark Pipeline



〈 H2O.ai Experiment

DRIVERLESS AI 1.1.3 – AI TO DO AI
Licensed to H2O.ai (SN21647)

TRAINING DATA

DATASET
train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
24K	25	--	--	--

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE	COUNT	UNIQUE	TARGET FREQ
int	23999	2	18630

What do these settings mean?

ACCURACY
- Training data size: **4,000 rows, 25 cols** (sampled)
- Feature evolution: **XGBoost, 1/3 validation split, 2 reps**
- Final pipeline: **XGBoost, 4-fold CV**

TIME
- Feature evolution: **8 individuals**, up to **500 iterations**
- Early stopping: After **50** iterations of no improvement

INTERPRETABILITY
- Feature pre-pruning strategy: None
- Monotonicity constraints: disabled
- Feature engineering search space (where applicable): ['Clustering', 'Date', 'FrequencyEncoding', 'Identity', 'Interactions', 'TargetEncoding', 'Text', 'TruncatedSVD', 'WeightOfEvidence']

XGBoost models to train:
- Feature evolution: **4024**
- Final pipeline: **1**

Estimated max. total memory usage:
- Feature engineering: **8.0MB**
- GPU XGBoost: **1.2GB**

Estimated runtime: **20 minutes**

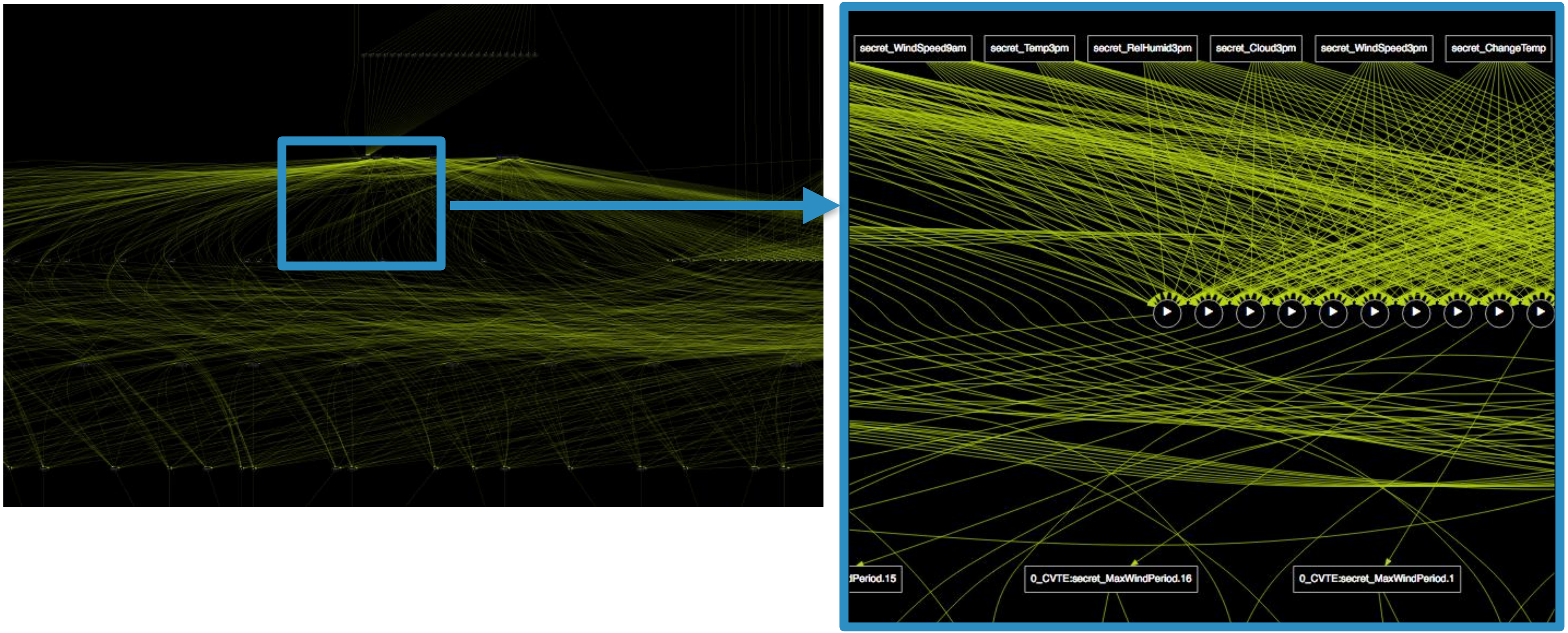
EXPERIMENT SETTINGS HELP

ACCURACY	TIME	INTERPRETABILITY
1	10	3

SCORER: AUC

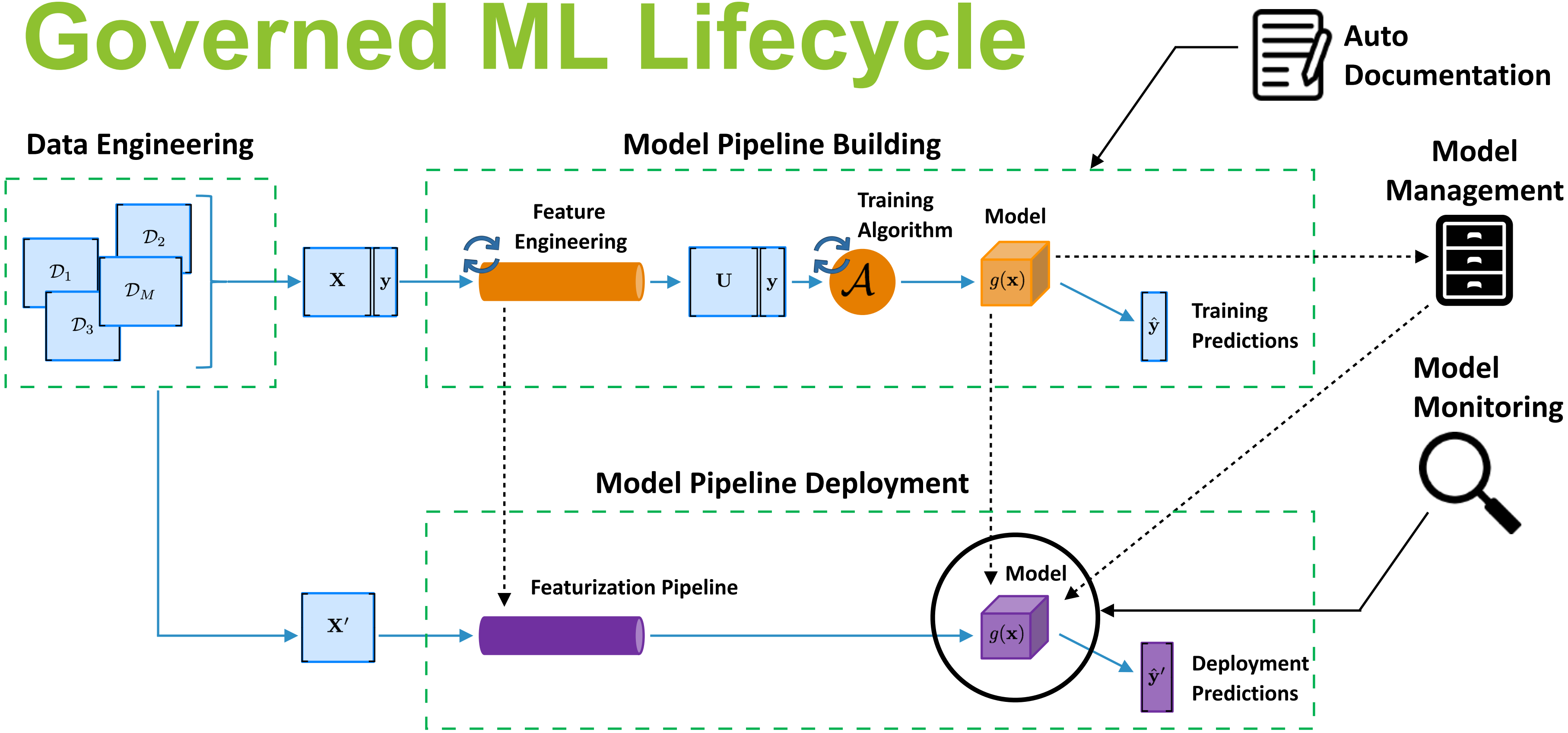
© 2017-2018 H2O.ai. All rights reserved.

Driverless AI Pipeline



Governed ML Lifecycle

Governed ML Lifecycle

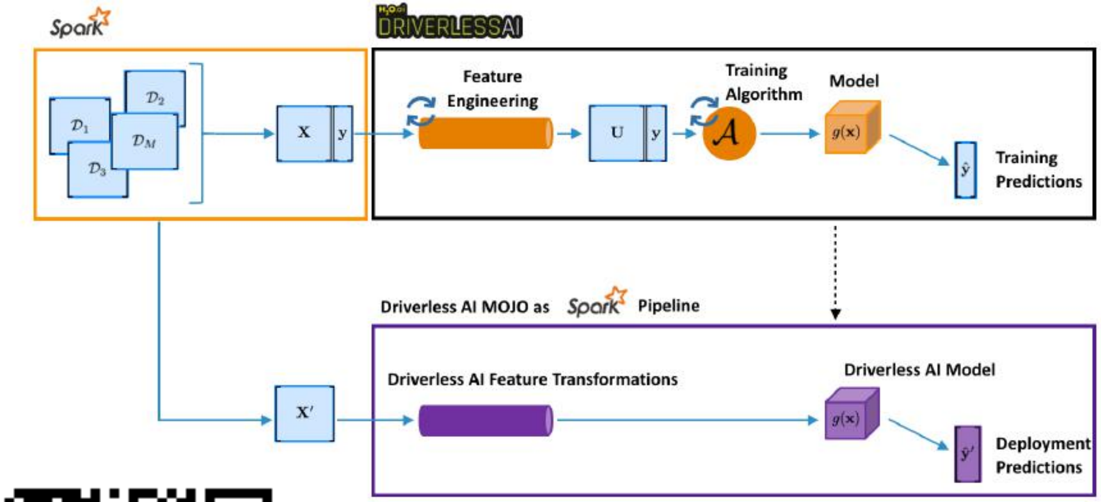


Materials

https://bit.ly/2sxowxD

Thank you!

Sparkling Water enables deployment of H2O ML models with Spark Pipelines


Spark
H2O.ai
DRIVERLESS AI
$\mathcal{D}_2$
$\mathcal{D}_1$
$\mathcal{D}_M$
$\mathcal{D}_3$
X
y
Feature Engineering
U
y
Training Algorithm
$\mathcal{A}$
Model
$g(\mathbf{x})$
$\hat{y}$
Training Predictions
Driverless AI MOJO as Spark Pipeline
Driverless AI Feature Transformations
Driverless AI Model
X'
$g(\mathbf{x})$
$\hat{y}'$
Deployment Predictions