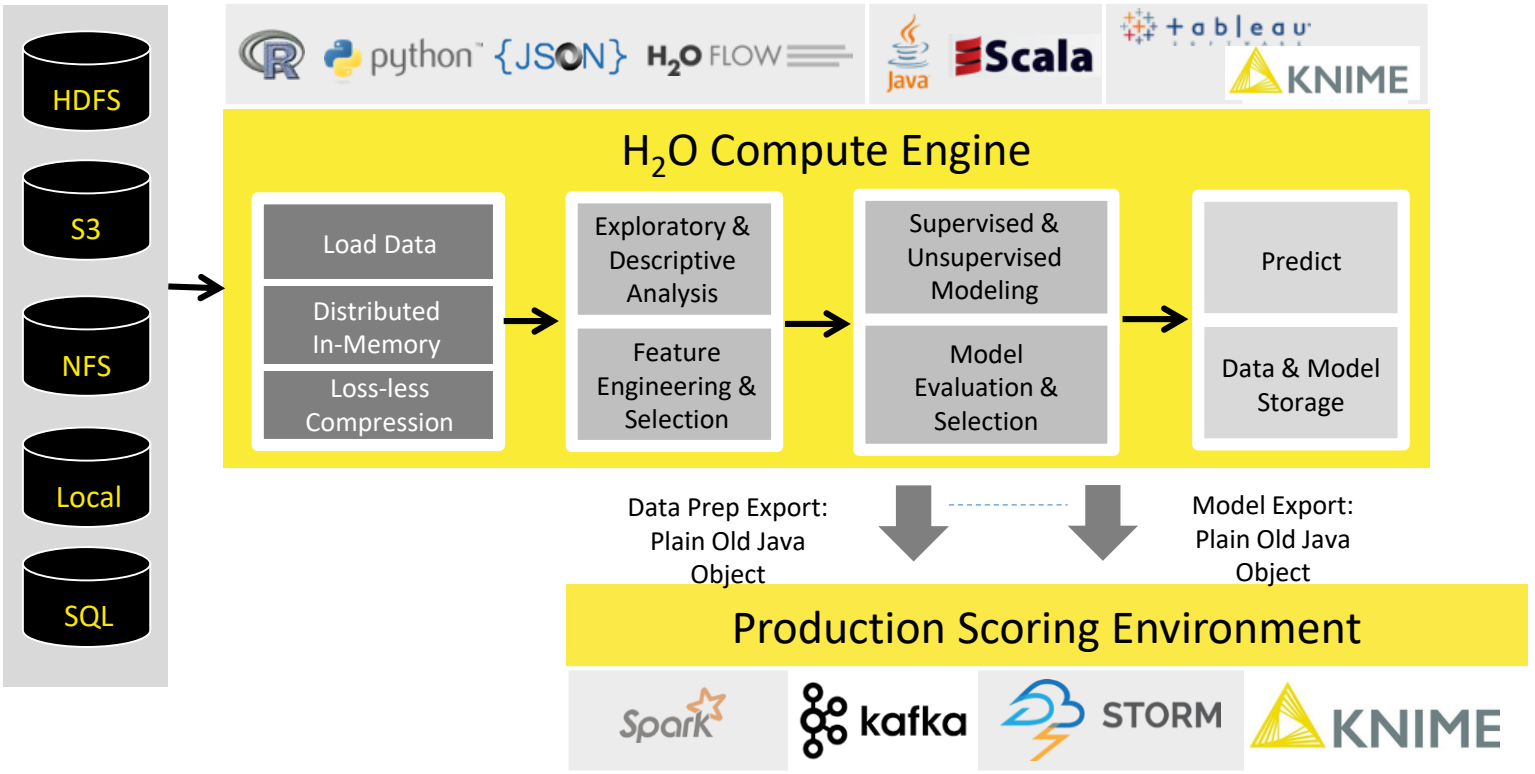
Integrating high-performance machine learning: H2O and KNIME

Mark Landry (H2O), Christian Dietz (KNIME)

Speed + Accuracy

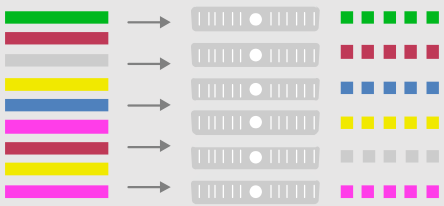
H2O: in-memory machine learning platform designed for speed on distributed systems.

High Level Architecture

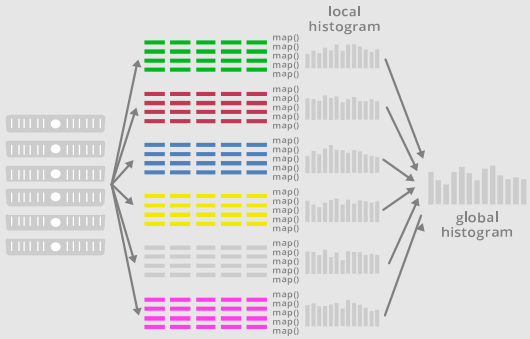


Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

Scientific Advisory Council

Dr. Trevor Hastie

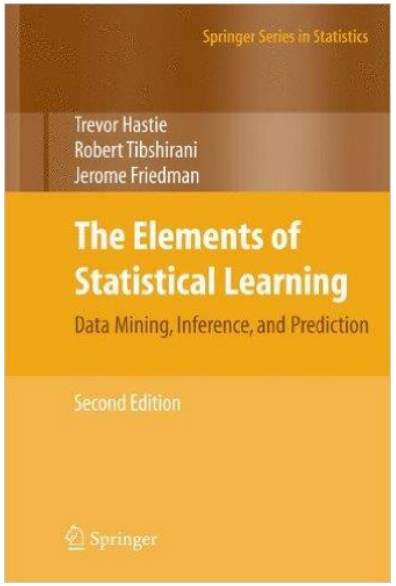
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- 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*

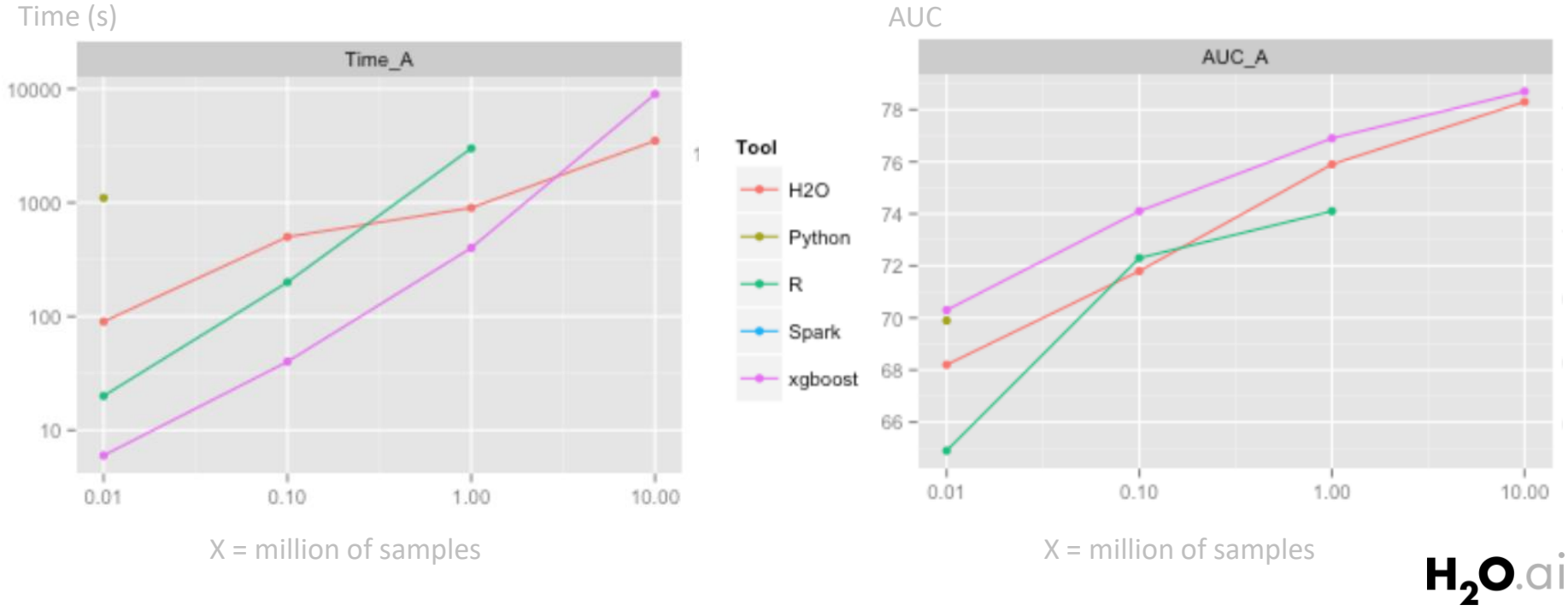
Dr. Steven Boyd

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



Machine Learning Benchmarks

(https://github.com/szilard/benchm-ml)



Gradient Boosting Machine Benchmark

(also available for GLM and Random Forest)

H2O Algorithms

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

Unsupervised Learning

Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components • **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data
Anomaly Detection	• **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

H2O in KNIME

Live Demo

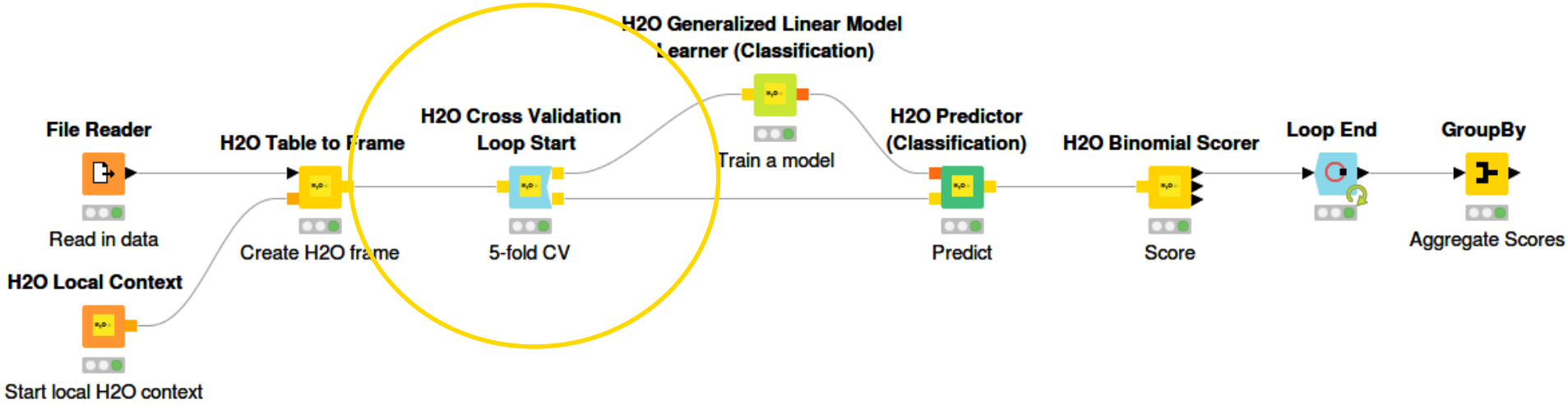
H2O in KNIME

- Offer our users high-performance machine learning algorithms from H2O in KNIME
- Allow to mix & match with other KNIME functionality
  - Data wrangling KNIME Analytics Platform functionality
  - KNIME Big-Data Connectors
  - Text Mining, Image Processing, Cheminformatics, …
  - and more!

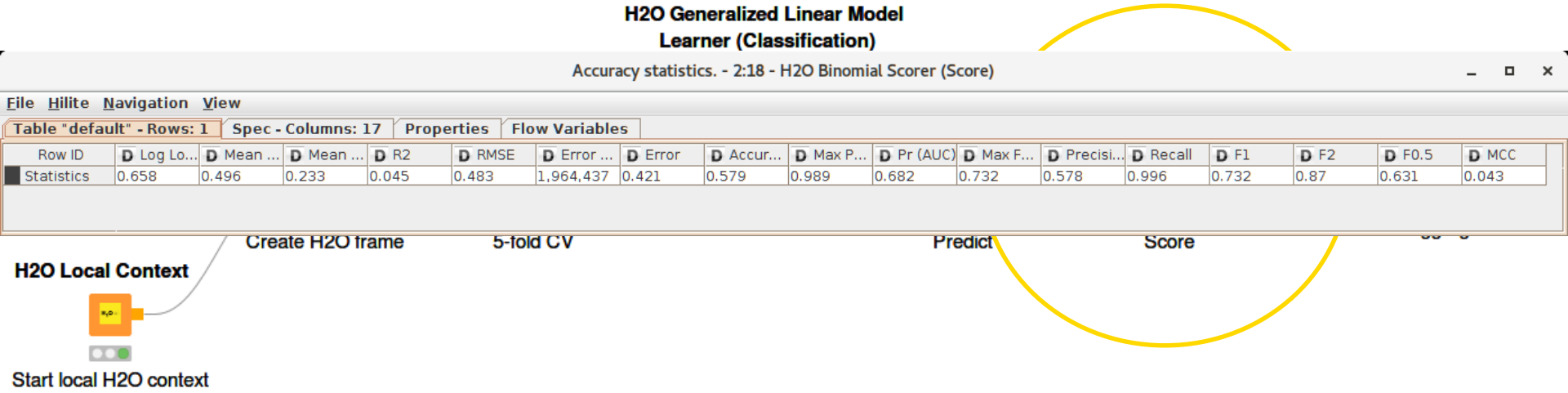
H2O in KNIME

Live Demo

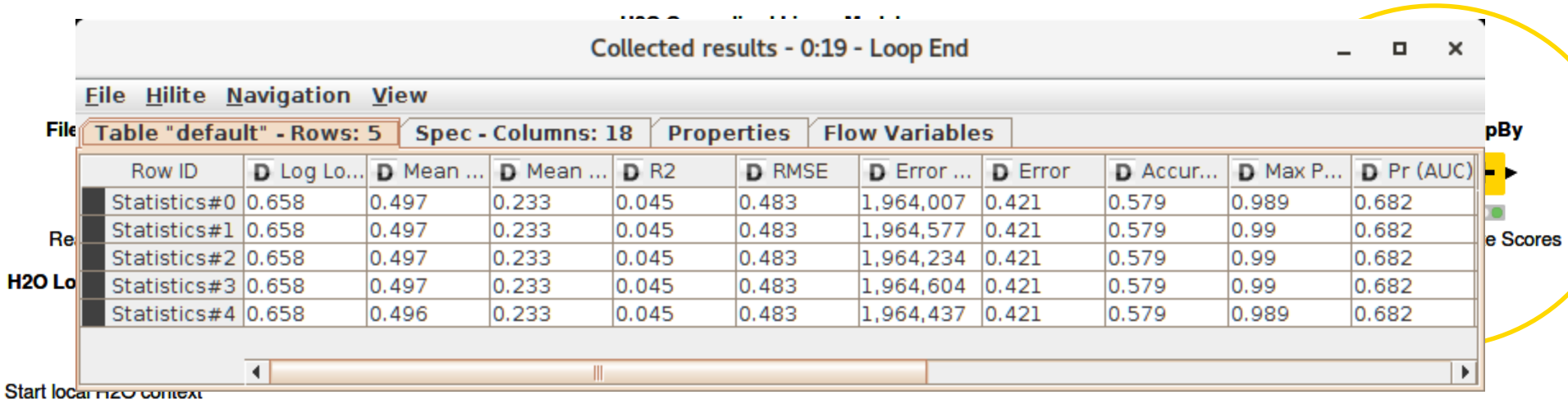
H2O in KNIME – Cross Validation



H2O in KNIME – Cross Validation

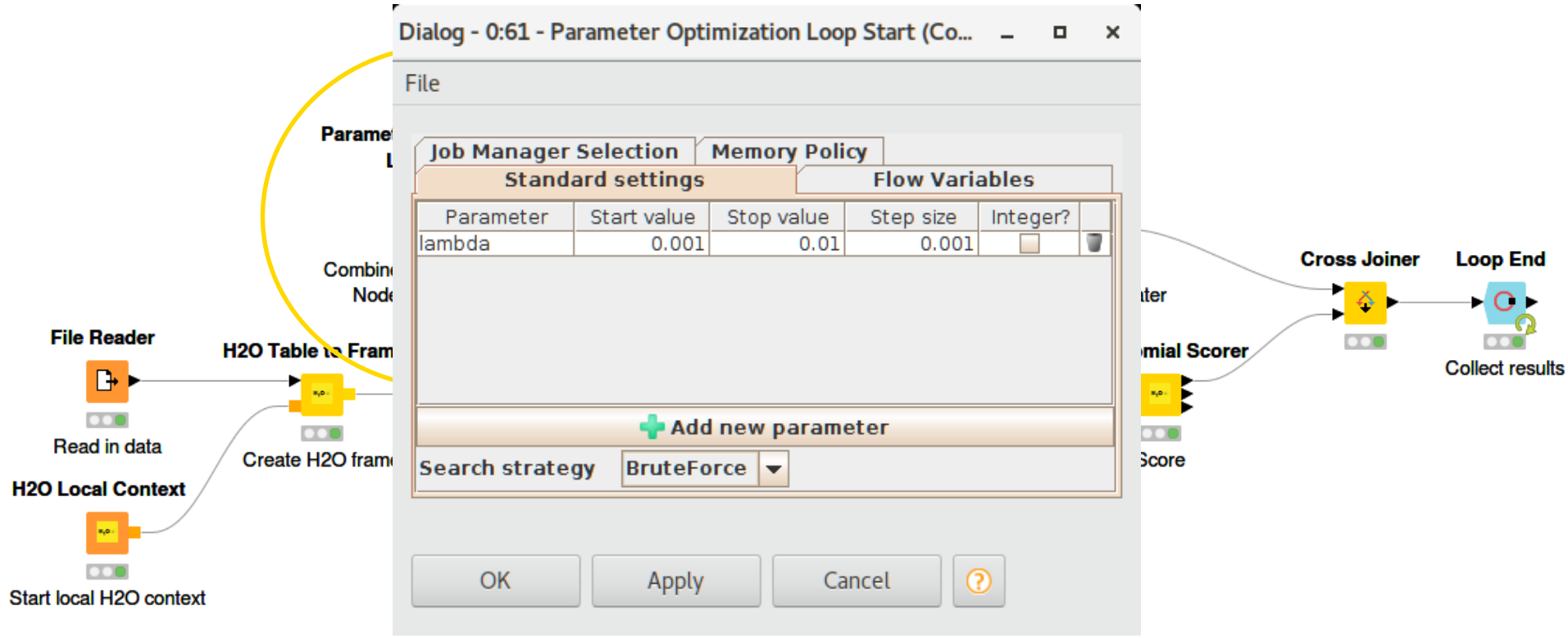


H2O in KNIME – Cross Validation

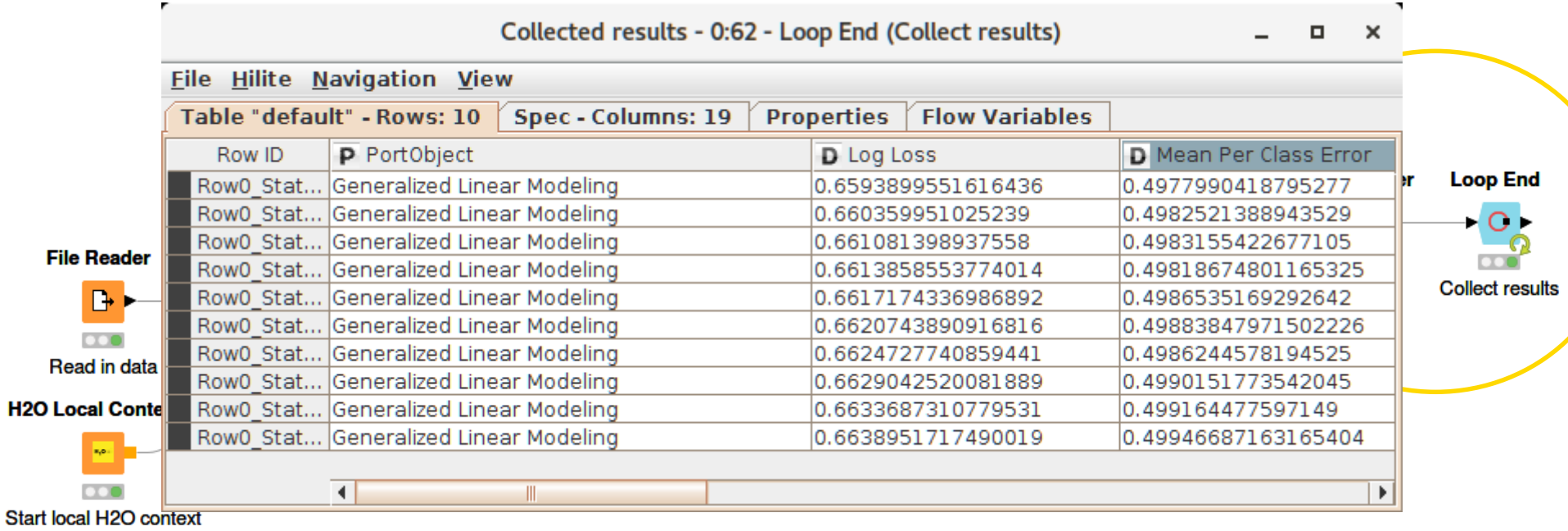


Row ID	Log Lo...	Mean ...	Mean ...	R2	RMSE	Error ...	Error	Accur...	Max P...	Pr (AUC)
Statistics#0	0.658	0.497	0.233	0.045	0.483	1,964,007	0.421	0.579	0.989	0.682
Statistics#1	0.658	0.497	0.233	0.045	0.483	1,964,577	0.421	0.579	0.99	0.682
Statistics#2	0.658	0.497	0.233	0.045	0.483	1,964,234	0.421	0.579	0.99	0.682
Statistics#3	0.658	0.497	0.233	0.045	0.483	1,964,604	0.421	0.579	0.99	0.682
Statistics#4	0.658	0.496	0.233	0.045	0.483	1,964,437	0.421	0.579	0.989	0.682

H2O in KNIME – Parameter Optimization



H2O in KNIME – Parameter Optimization



H2O in KNIME – Nodes in KNIME 3.4

H2O Gradient Boosting Machine Learner (Classification)

H2O Gradient Boosting Machine Learner (Regression)

H2O Frame Statistics

H2O Naive Bayes Learner

H2O Random Forest Learner (Classification)

H2O Multinominal Scorer

H2O Random Forest Learner (Regression)

H2O Predictor (Regression)

H2O Regression Scorer

H2O Predictor (Classification)

H2O in KNIME – What's cooking?

H2O PCA

H2O PCA Apply

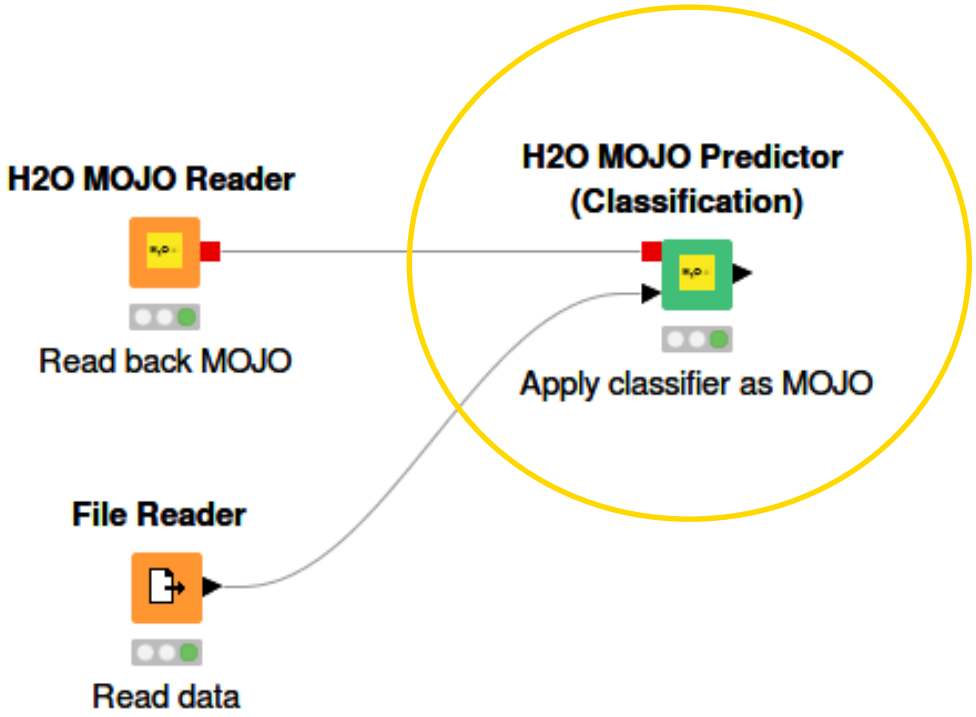
H2O PCA Compute

H2O Generalized Low Rank Models (GLRM) Missing Value Impute

H2O K-Means

H2O Cluster Assigner

H2O in KNIME – What's cooking?



Thank you!