H2O Training

1. Introduction to ML
2. Introduction to H2O
3. H2O Data Munging
4. Machine Learning Concepts
5. H2O Algorithms
6. Productionalizing Models

H2O Machine Learning Methods

Supervised Learning

Statistical Analysis

- **Penalized Linear Models**: Super-fast, super-scalable, and interpretable
- **Naïve Bayes:** Straightforward linear classifier

Decision Tree Ensembles

- **Distributed Random Forest**: Easy-to-use tree-bagging ensembles
- **Gradient Boosting Machine**: Highly tunable tree-boosting ensembles
- **eXtreme Gradient Boosting**: Popular XGBoost algorithm in H2O

Stacking

- **Stacked Ensemble**: Combine multiple types of models for better predictions

AutoML

- **Automatic Machine Learning**: Automated exploration of supervised learning approaches

Unsupervised Learning

Clustering

- **K-means**: Partitions observations into similar groups; automatically detects number of groups

Dimensionality Reduction

- **Principal Component Analysis**: Transforms correlated variables to independent components
- **Generalized Low Rank Models:** Extends the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Aggregator

- **Aggregator:** Efficient, advanced sampling that creates smaller data sets from larger data sets

Neural Networks

Multilayer Perceptron

Deep Learning

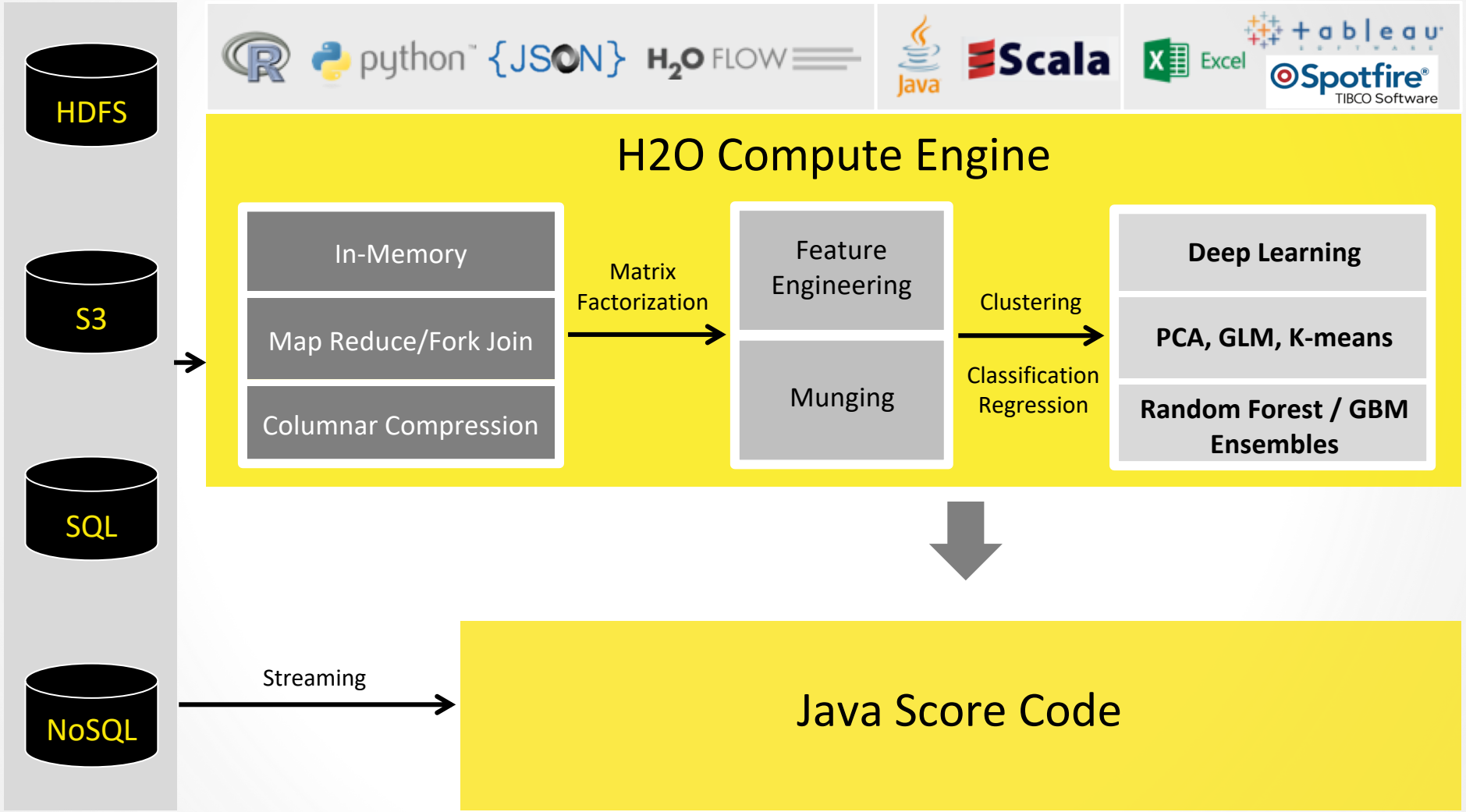
- **Deep neural networks**: Multi-layer feed-forward neural networks for standard data mining tasks
- **Convolutional neural networks:** Sophisticated architectures for pattern recognition in images, sound, and text

Anomaly Detection

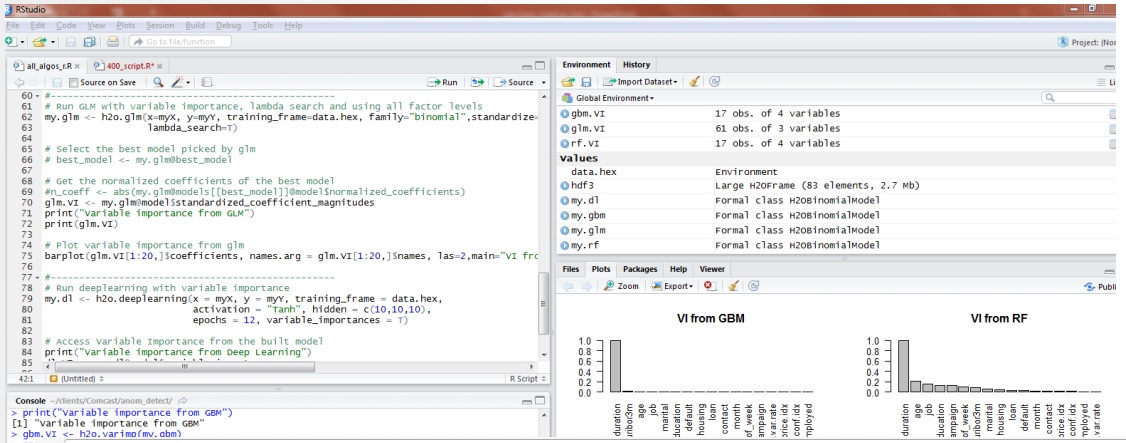
Term Embeddings

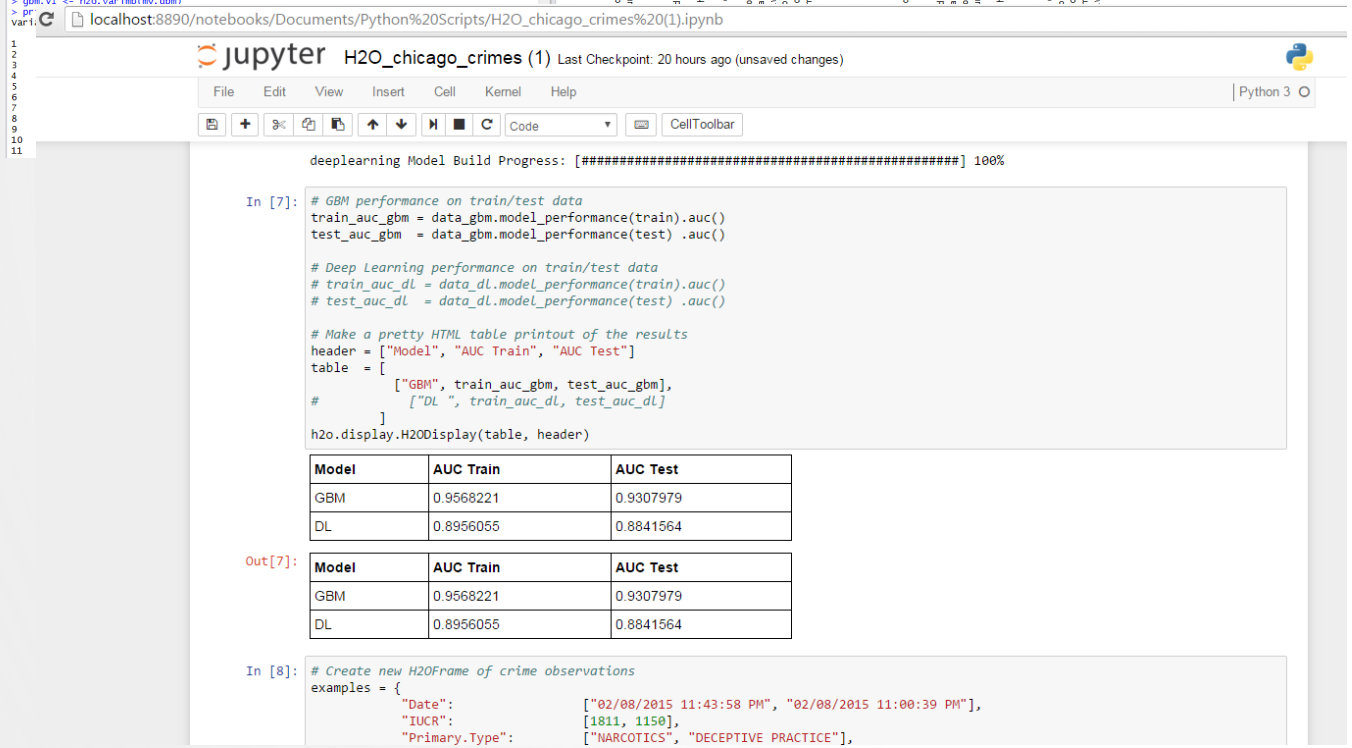
- **Autoencoders**: Find outliers using a nonlinear dimensionality reduction technique
- **Word2vec:** Generate context-sensitive numerical representations of a large text corpus

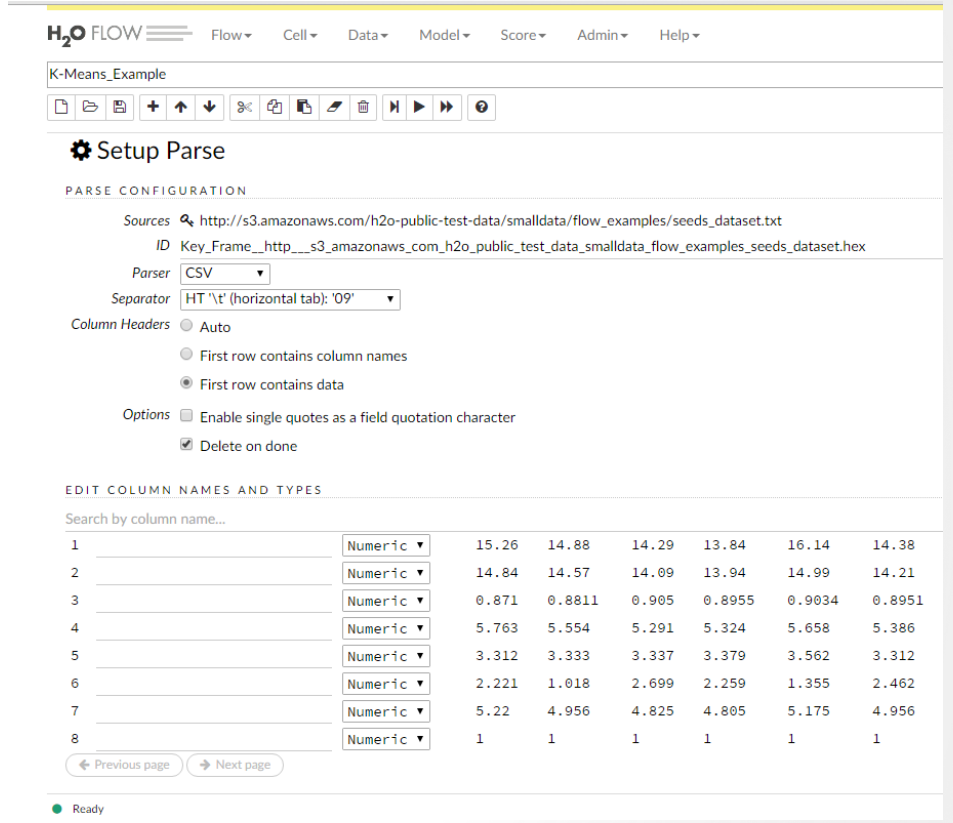
Data and Client Agnostic



R, Python and Flow







Python Interface Overview

Action	Pandas or scikit-learn	H2O
Reading data	`pandas.read_csv(data_path)`	`h2o.import_file(data_path)`
Summarizing data	`pandas_frame.describe()`	`h2o_frame.describe()`
Summary statistics	`pandas_frame.mean()`	`h2o_frame.mean()`
Combining rows	`pandas.concat(list[frame1,frame2])`	`h2o_frame.rbind(h2o_frame2)`
Combining columns	`pandas.concat(list[frame1,frame2],axis = 1)`	`h2o_frame.cbind(h2o_frame2)`
Data selection	`pandas_frame[:, :]`	`h2o_frame[:, :]`
Transforming columns	`np.log(pandas_frame[x])` `np.sqrt(pandas_frame[x])`	`h2o_frame[x].log()` `h2o_frame[x].sqrt()`
Building Random Forest	`model = RandomForestClassifier(n_estimators = 100)` `model = model.fit(x_frame, y_frame)`	`model = H2ORandomForestClassifier(n_trees = 100)` `model = model.train(x, y, train_frame)`
Model Prediction	`model.predict`	`model.predict`
Model Metrics	`metrics.auc`	`metrics = model.model_performance(frame)` `metrics.auc()`

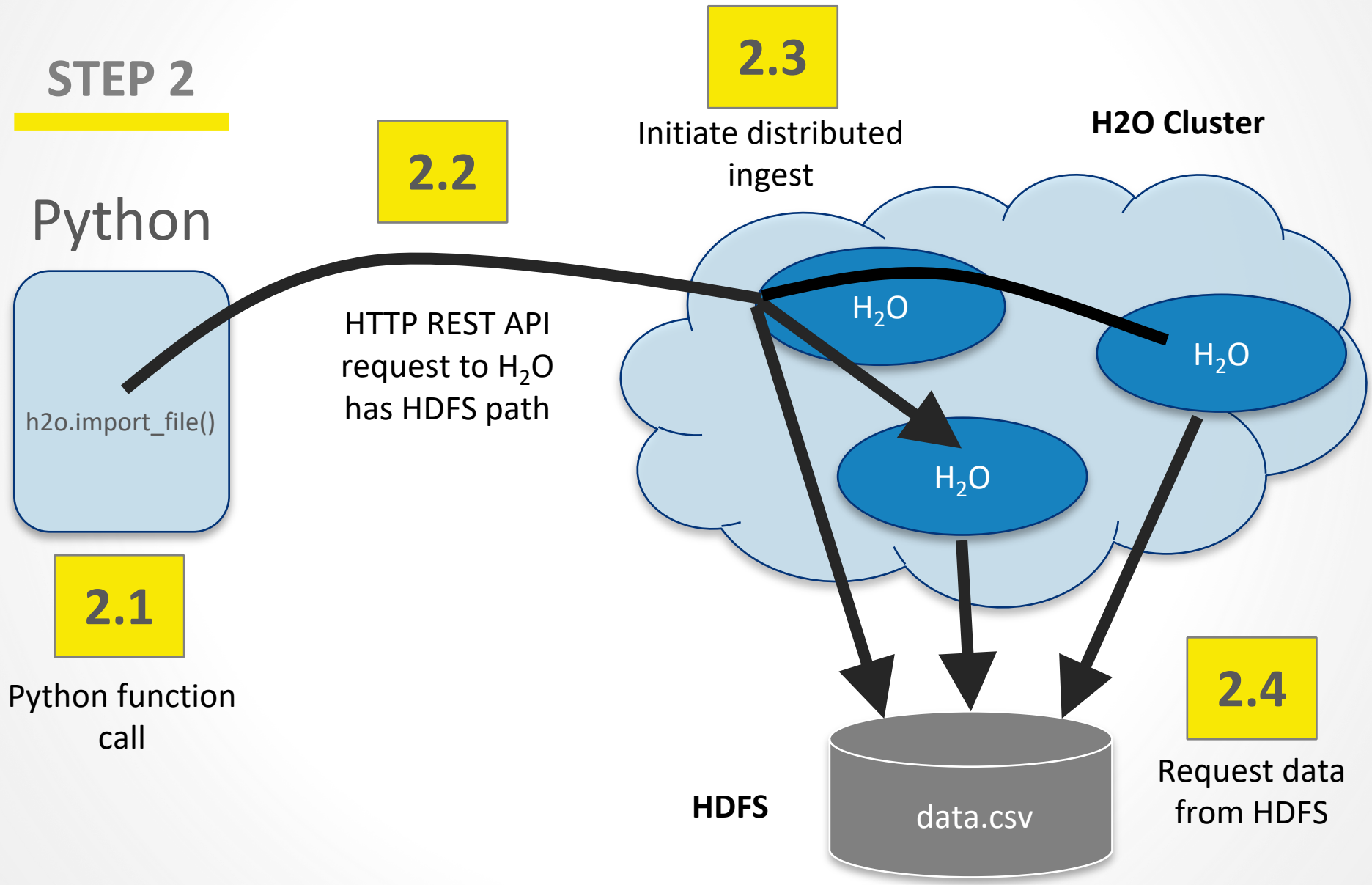
Reading Data into H2O with Python

STEP 1

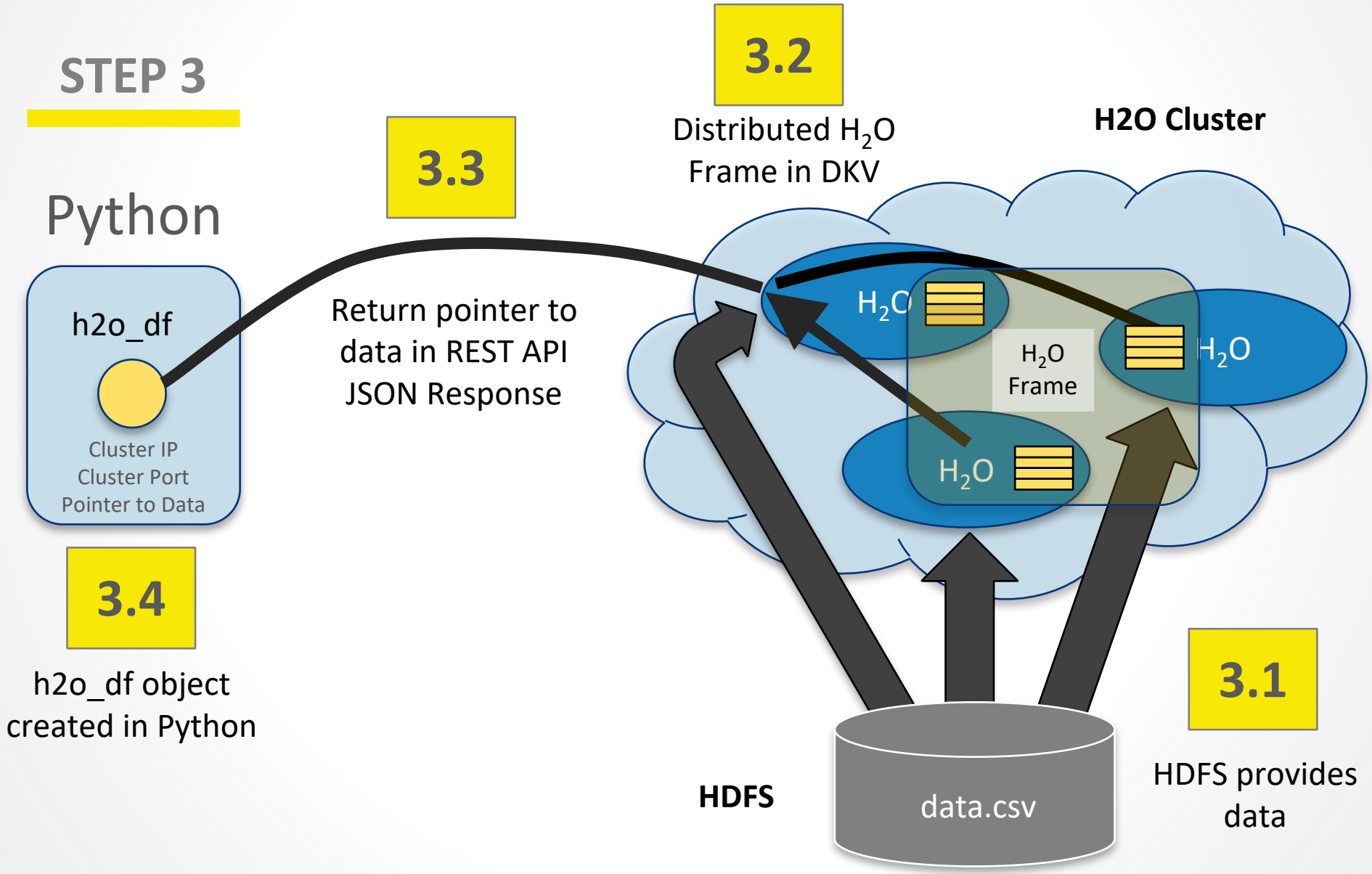
Python user

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

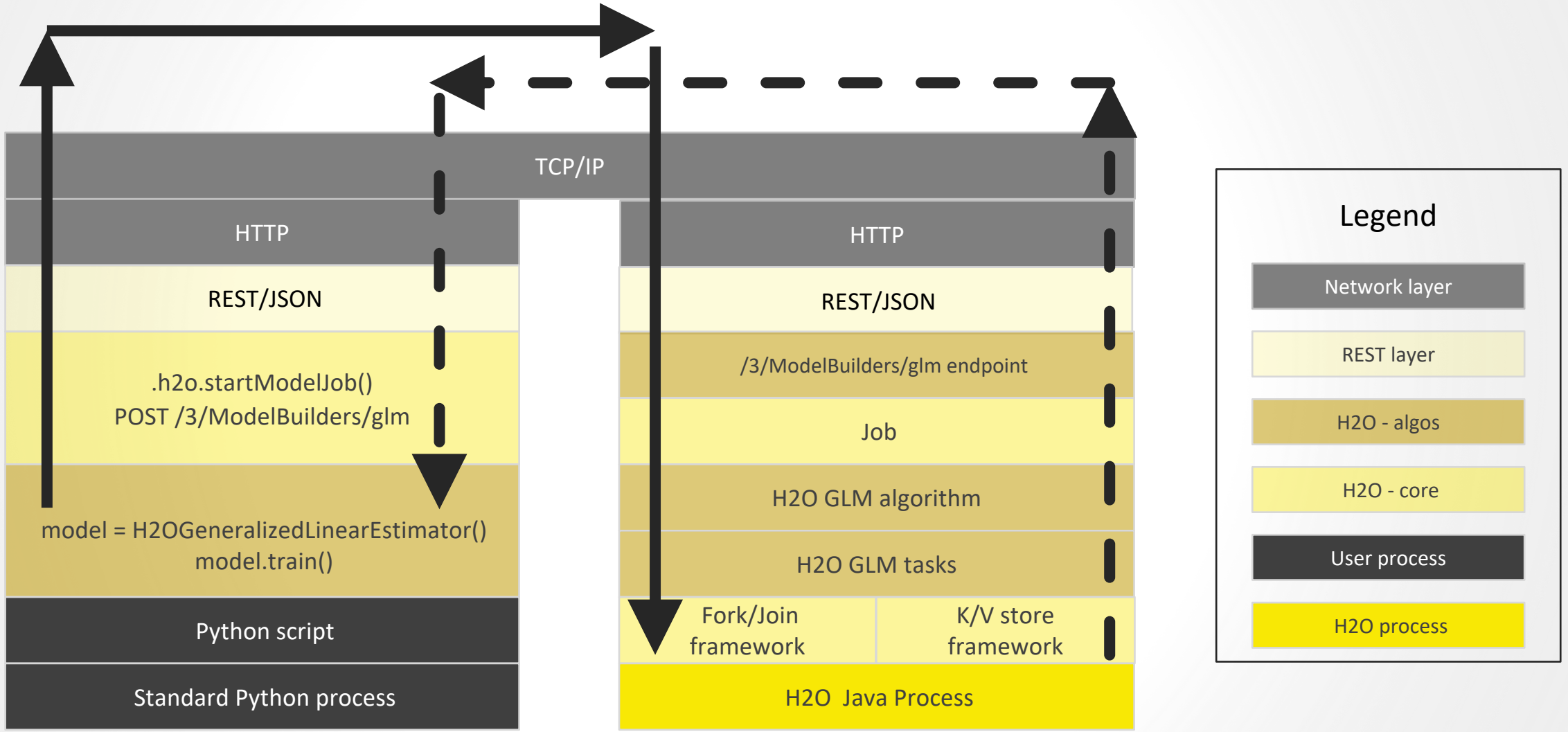
Reading Data into H2O with Python



Reading Data into H2O with Python



Training GLM from Python



Retrieving GLM Result from Python