

Automatic Machine Learning

How to Get Started

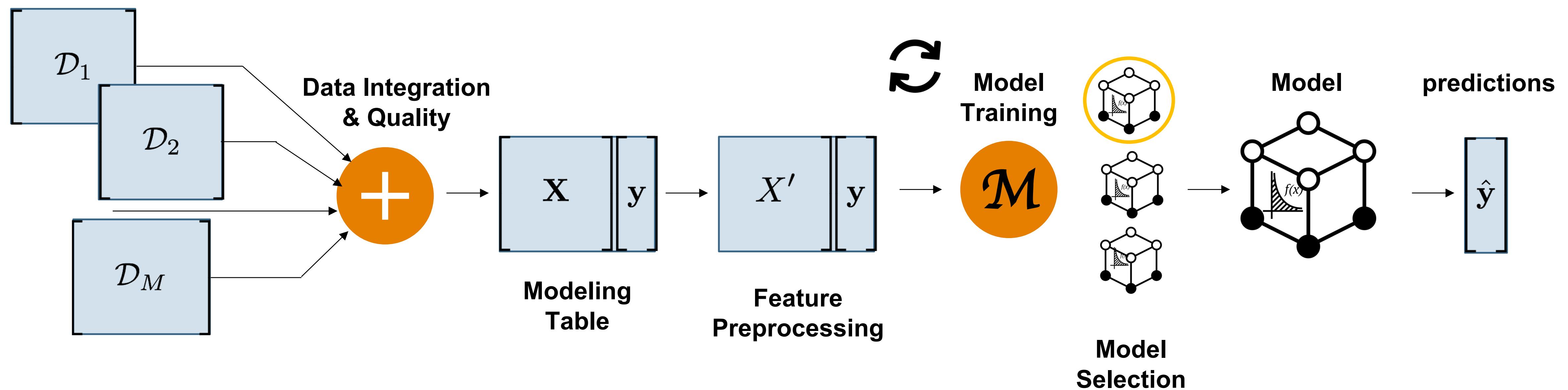
Lauren DiPerna // Data Scientist @ H2O

Auto-What?

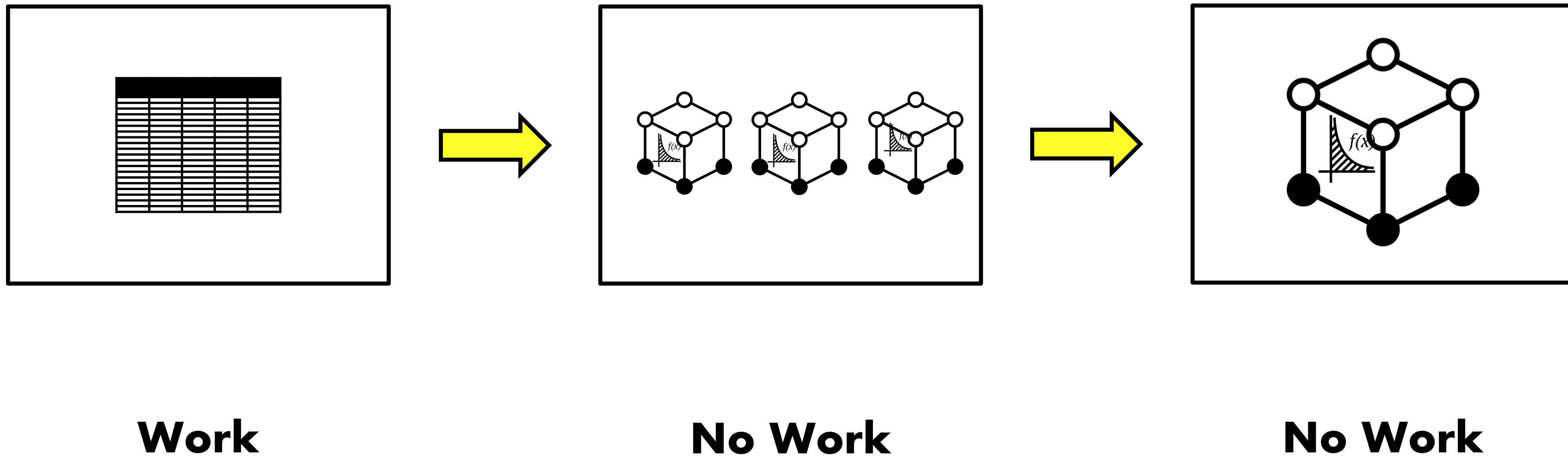
What is AutoML?



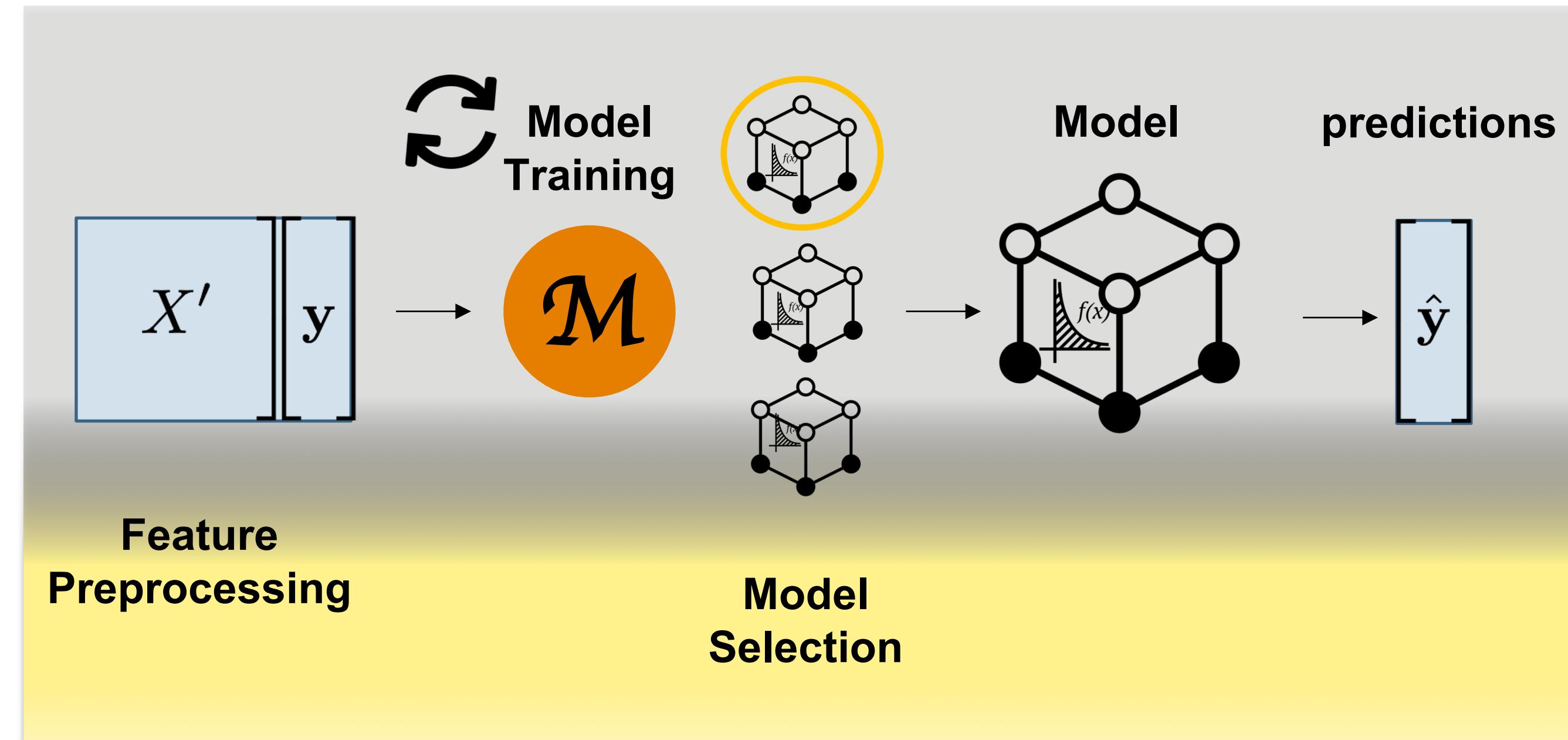
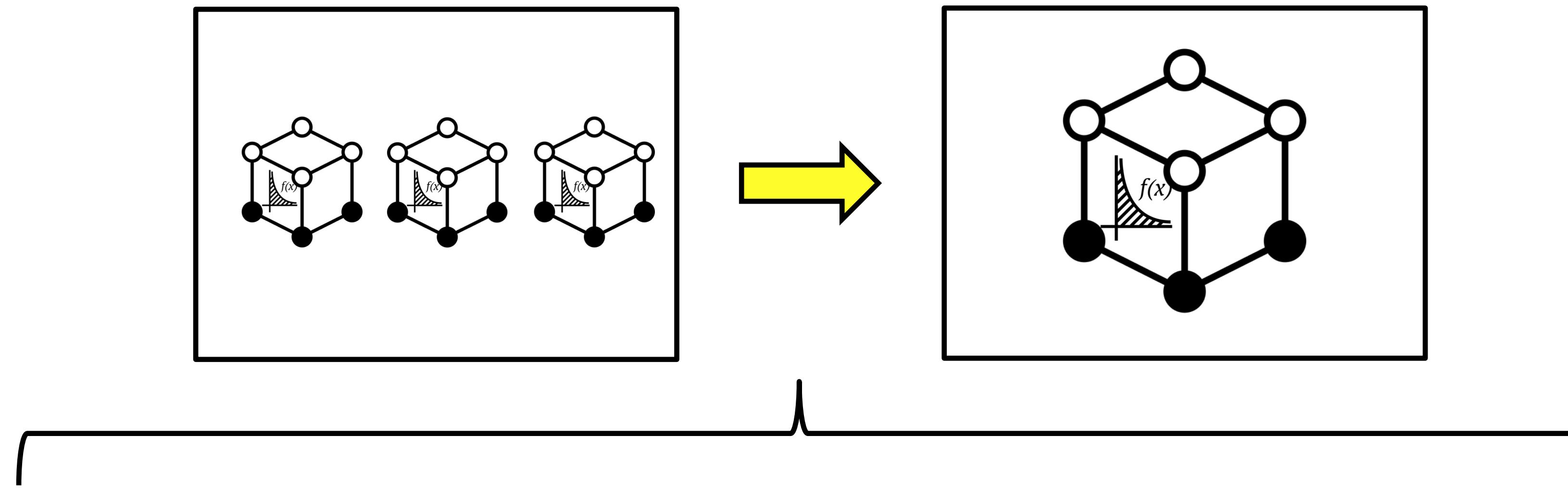
The Machine Learning Pipeline



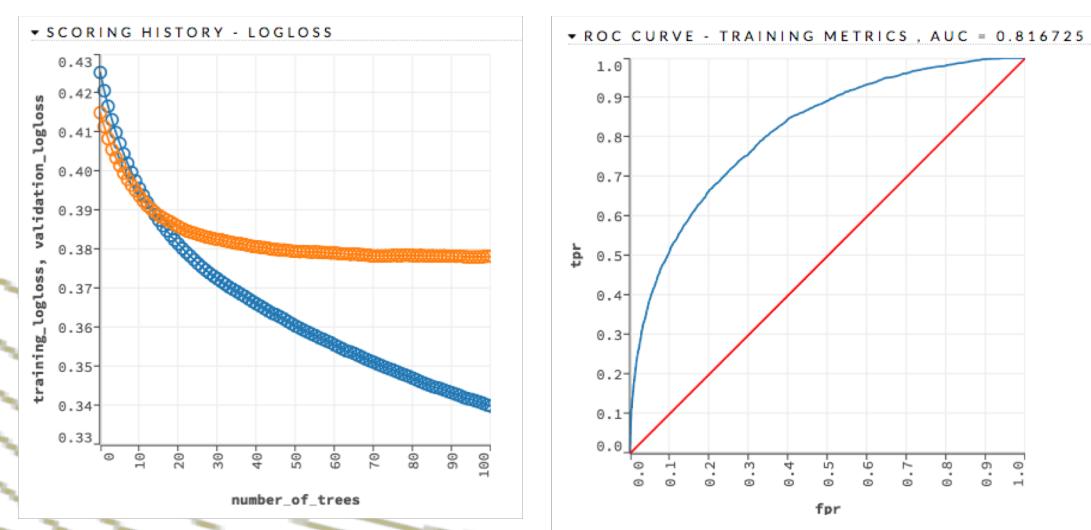
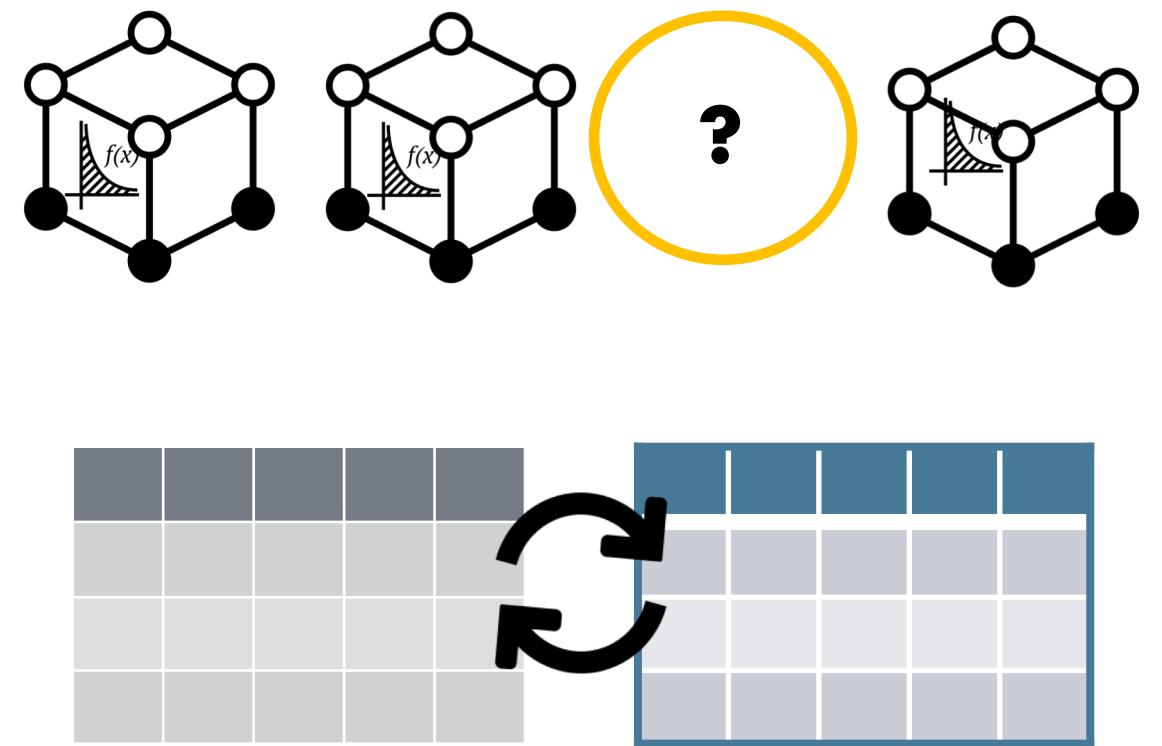
What is Automatic Machine Learning?



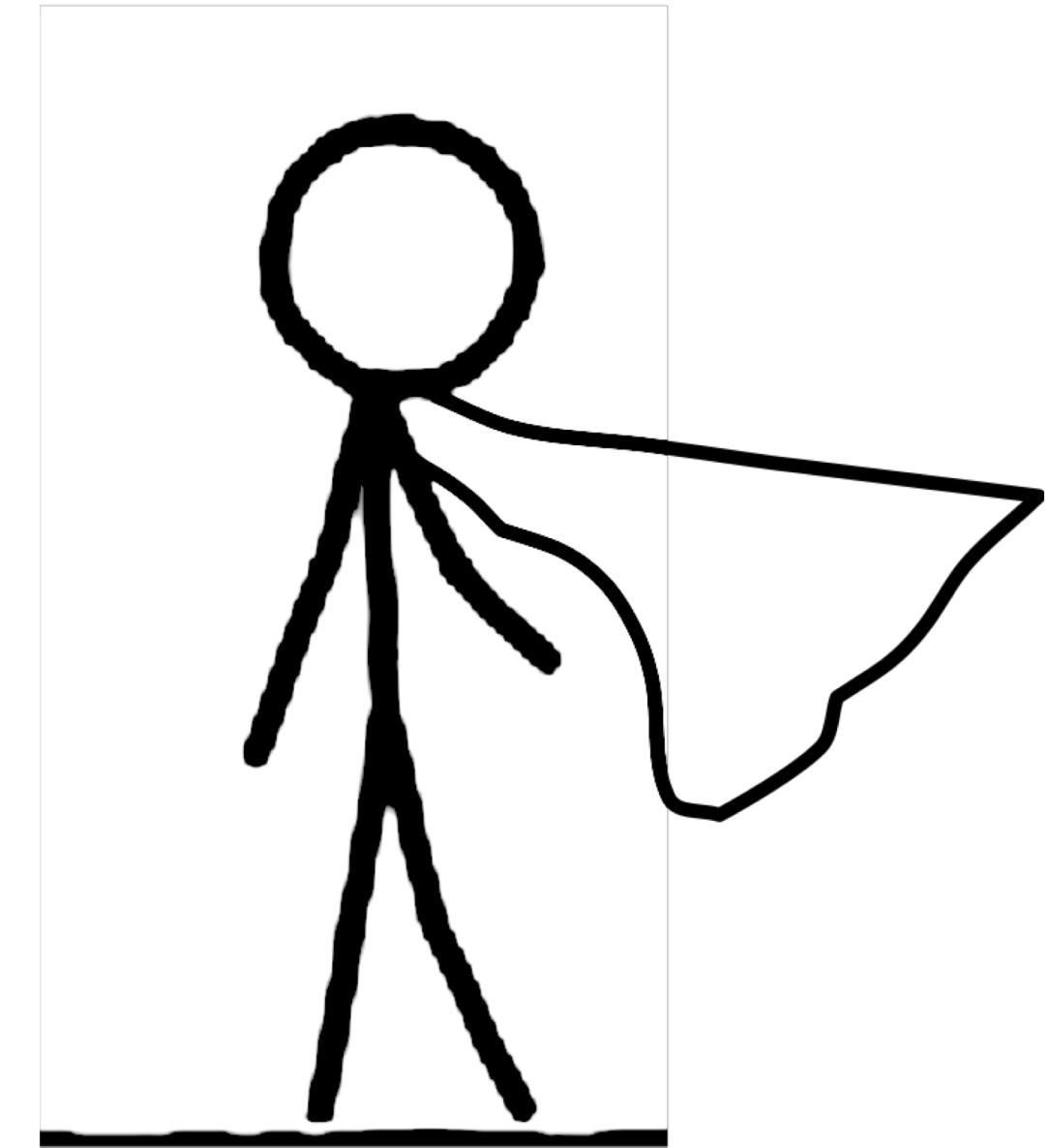
What is Automatic Machine Learning?



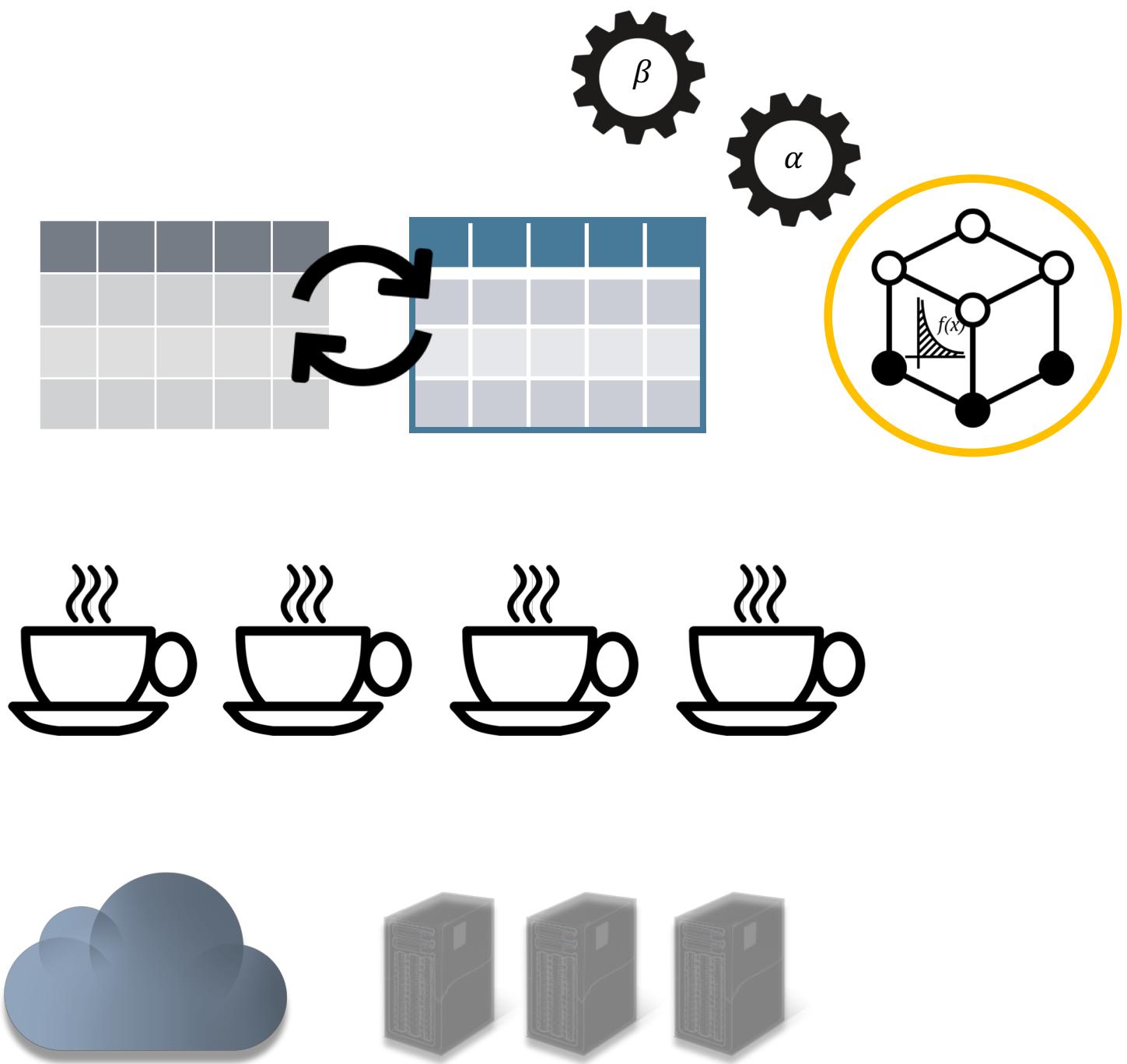
Why Try it?



Small Cape



Big Cape

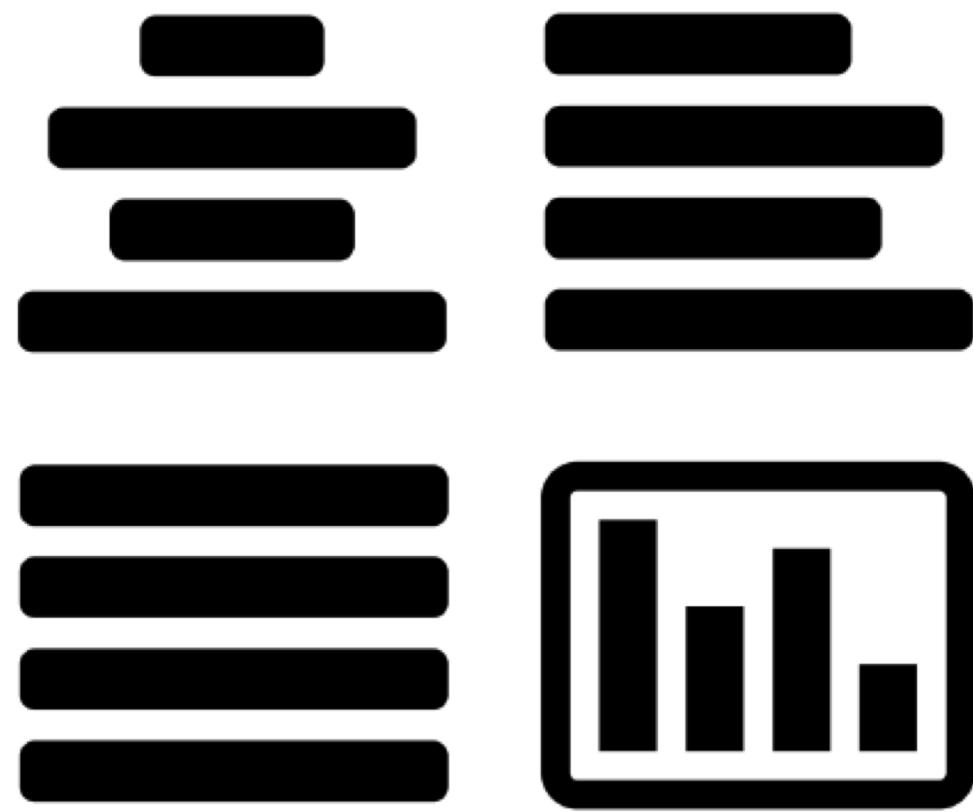


Automatic Machine Learning

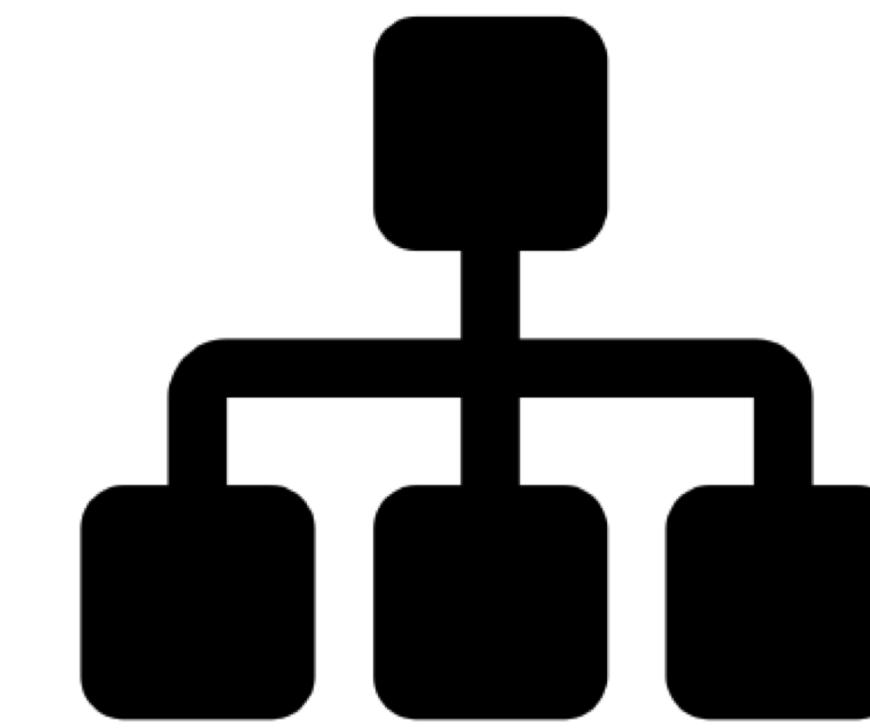
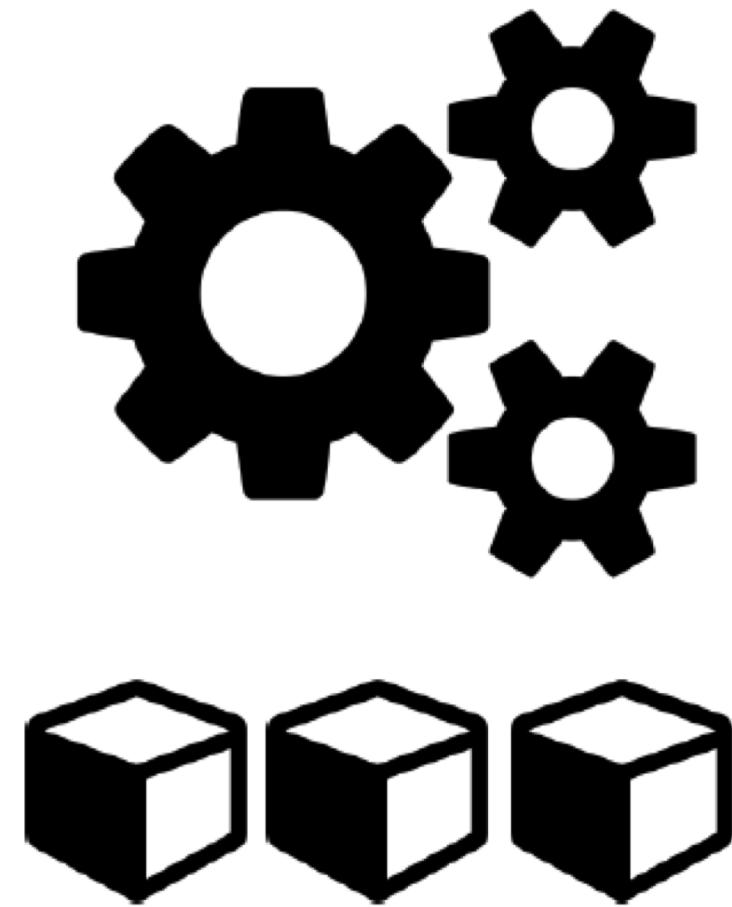
The Aspects of AutoML



Model Generation



Data Prep

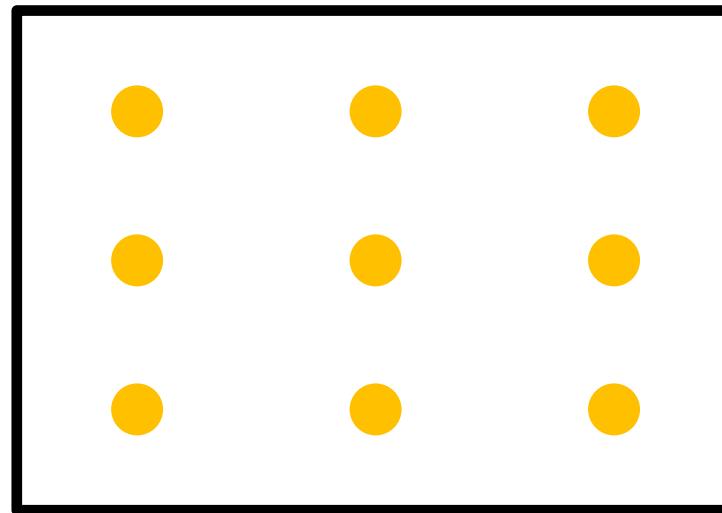


Ensembles

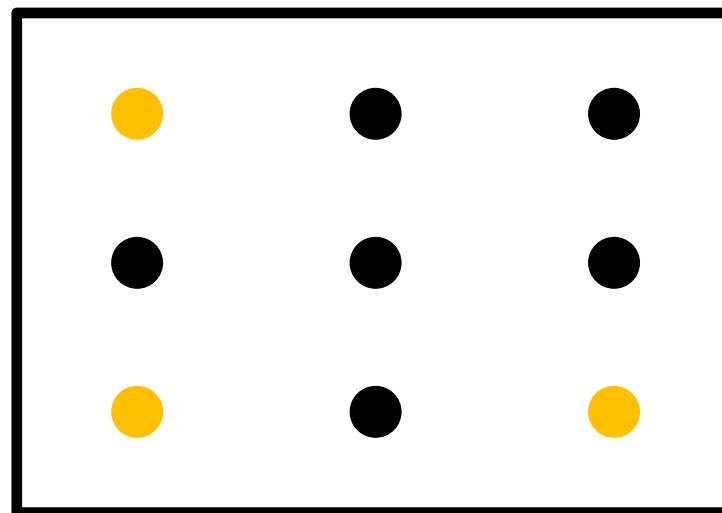
Data Prep

- **Format Data:**
 - Imputation, one-hot encoding, standardization
- **Choose Columns:**
 - Feature selection, feature extraction (e.g. PCA)
- **Advanced Encoding:**
 - Count, Label, Target encoding of categorical features

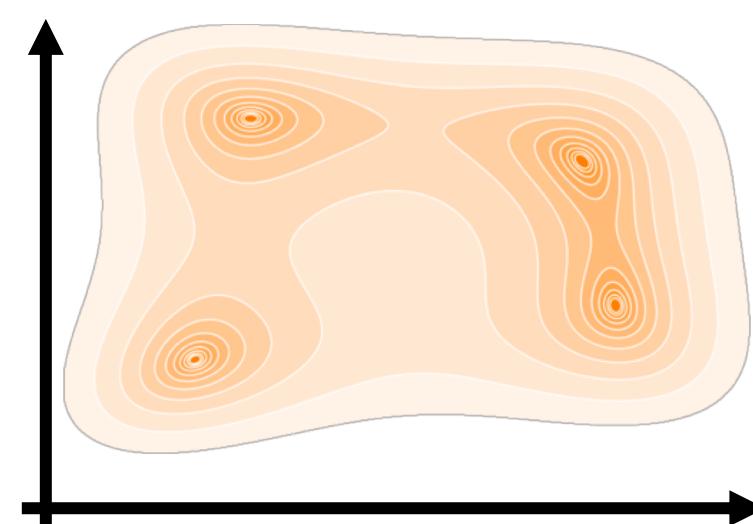
Model Generation



Cartesian grid search

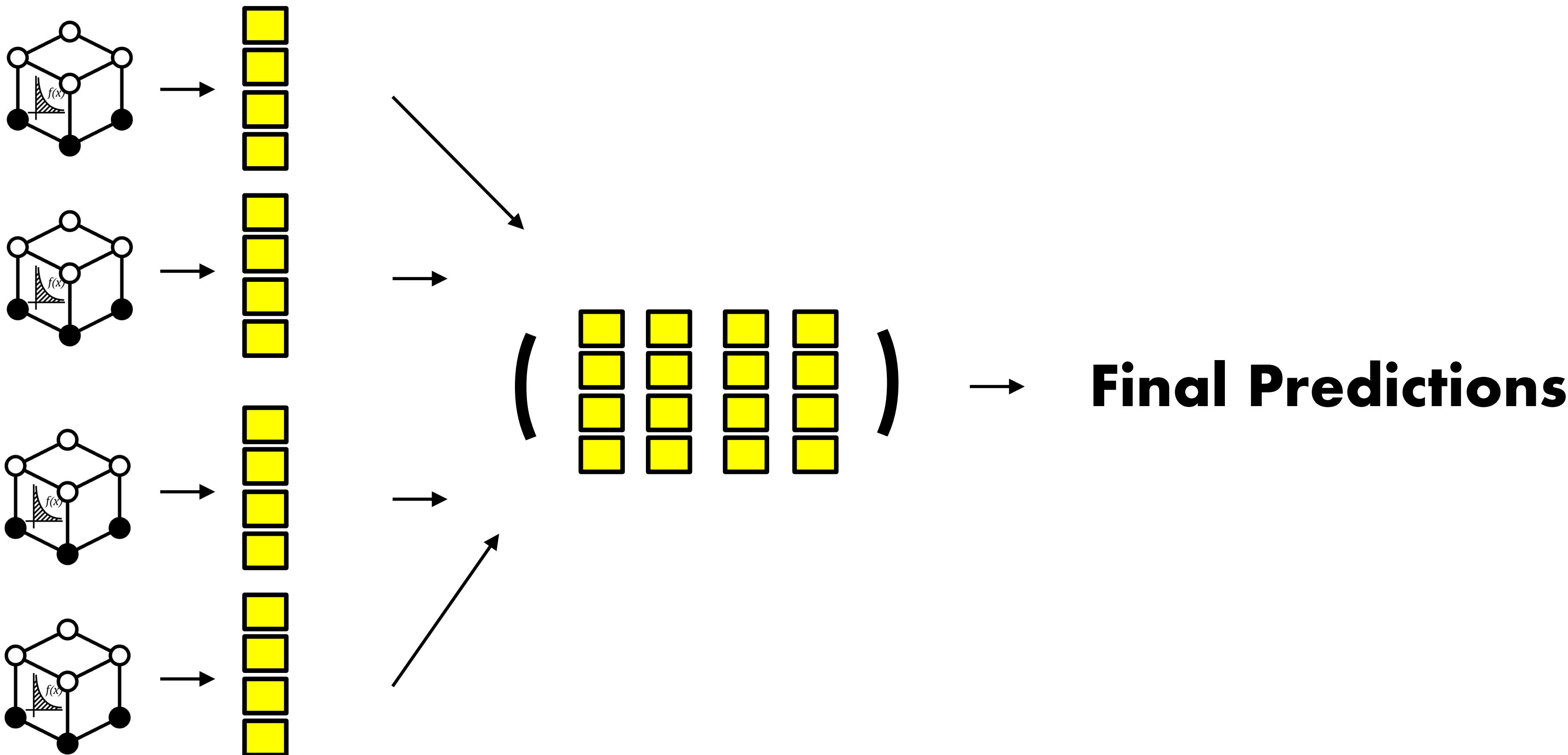


Random grid search



Bayesian Hyperparameter Optimization

Ensembles



Ensembles

Ensembles often out-perform individual models

- Bagging
- Boosting
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

What is H2O?

A Machine Learning Platform

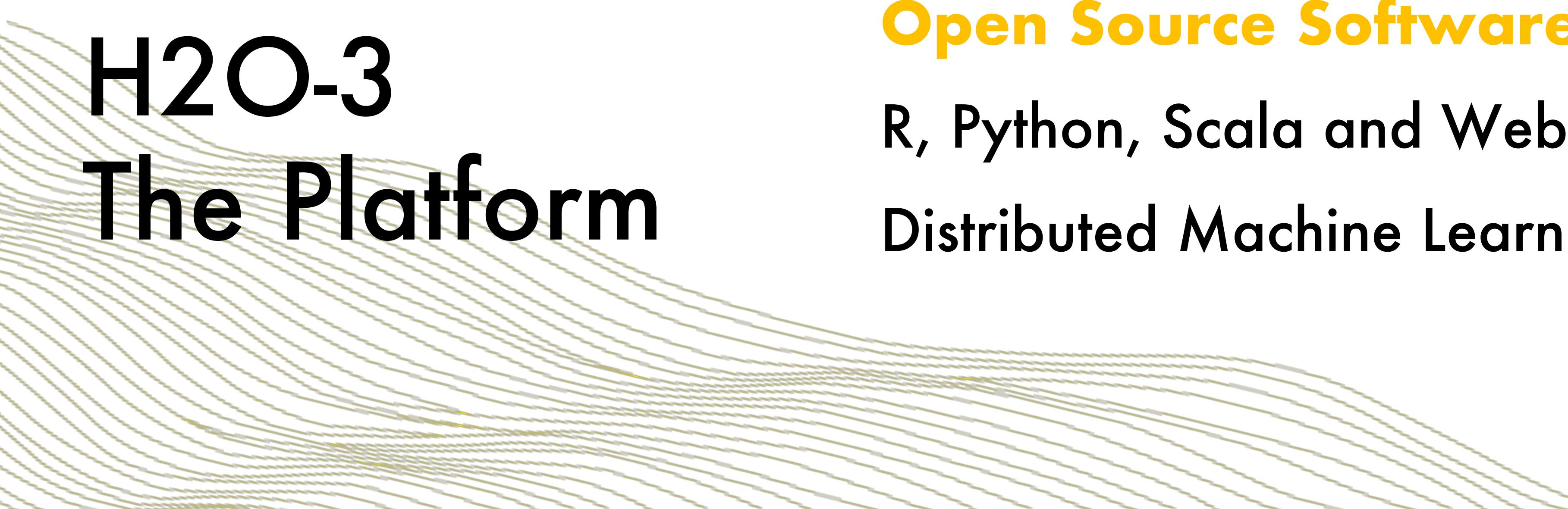


H2O.ai The Company

Founded in 2012
Advised by Stanford Professors:
Hastie, Tibshirani & Boyd

Headquarters: Mountain View, California, USA

H2O-3 The Platform



Open Source Software
R, Python, Scala and Web Interfaces
Distributed Machine Learning Algorithms

What is H2O-3?



What is H2O-3?



`pip install h2o`

What is H2O-3?

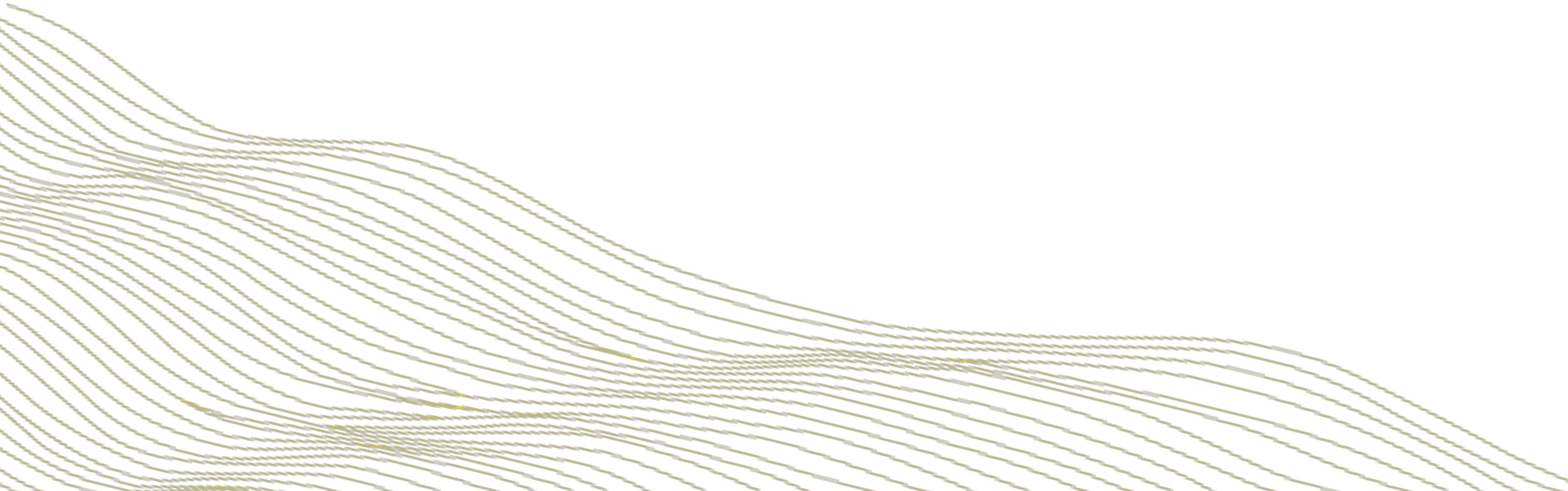
```
conda install -c h2oai h2o
```

What is H2O-3?

```
install.packages("h2o")
```

H2O's AutoML

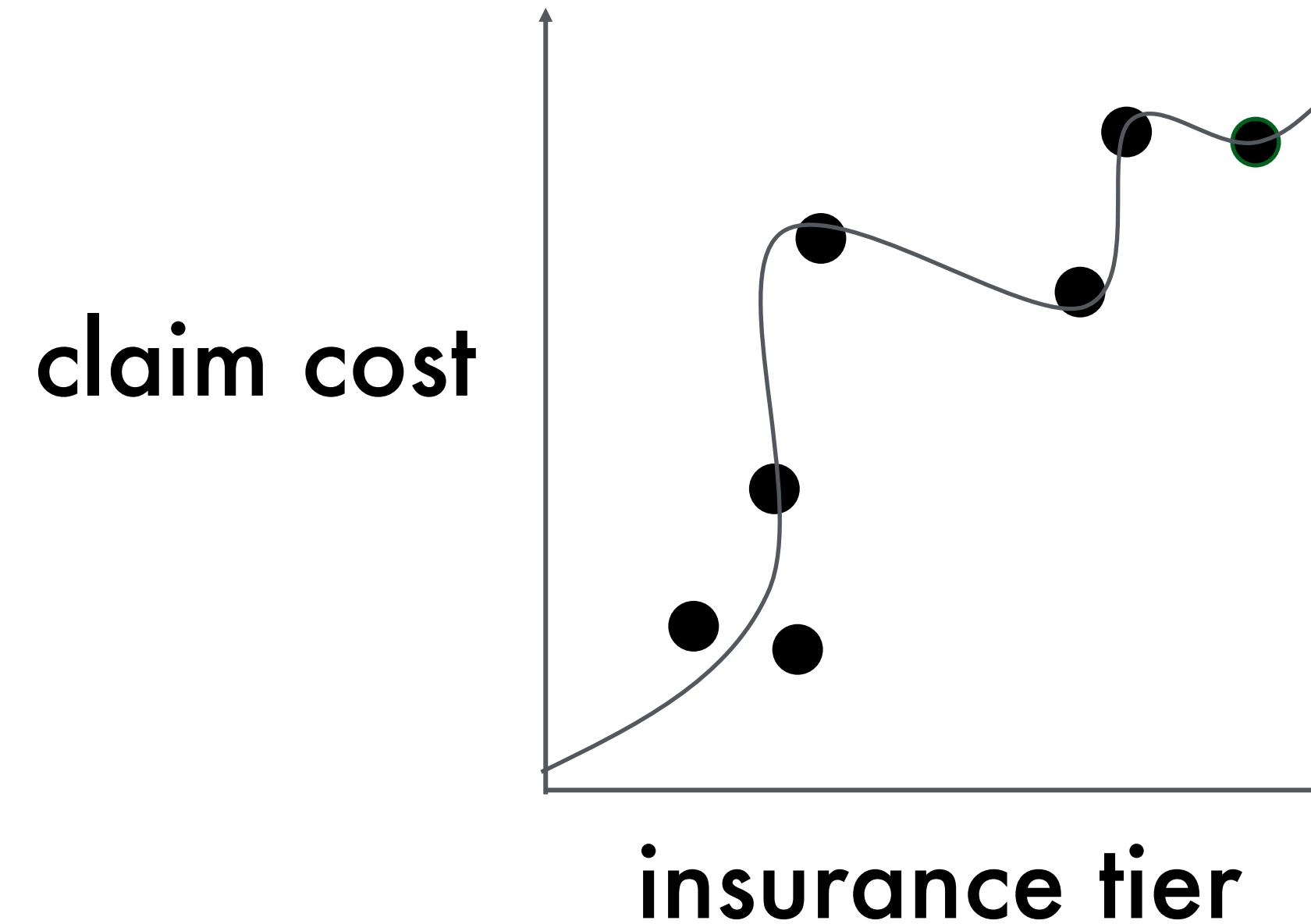
Scalable Automatic Machine Learning



AutoML: Supervised Learning

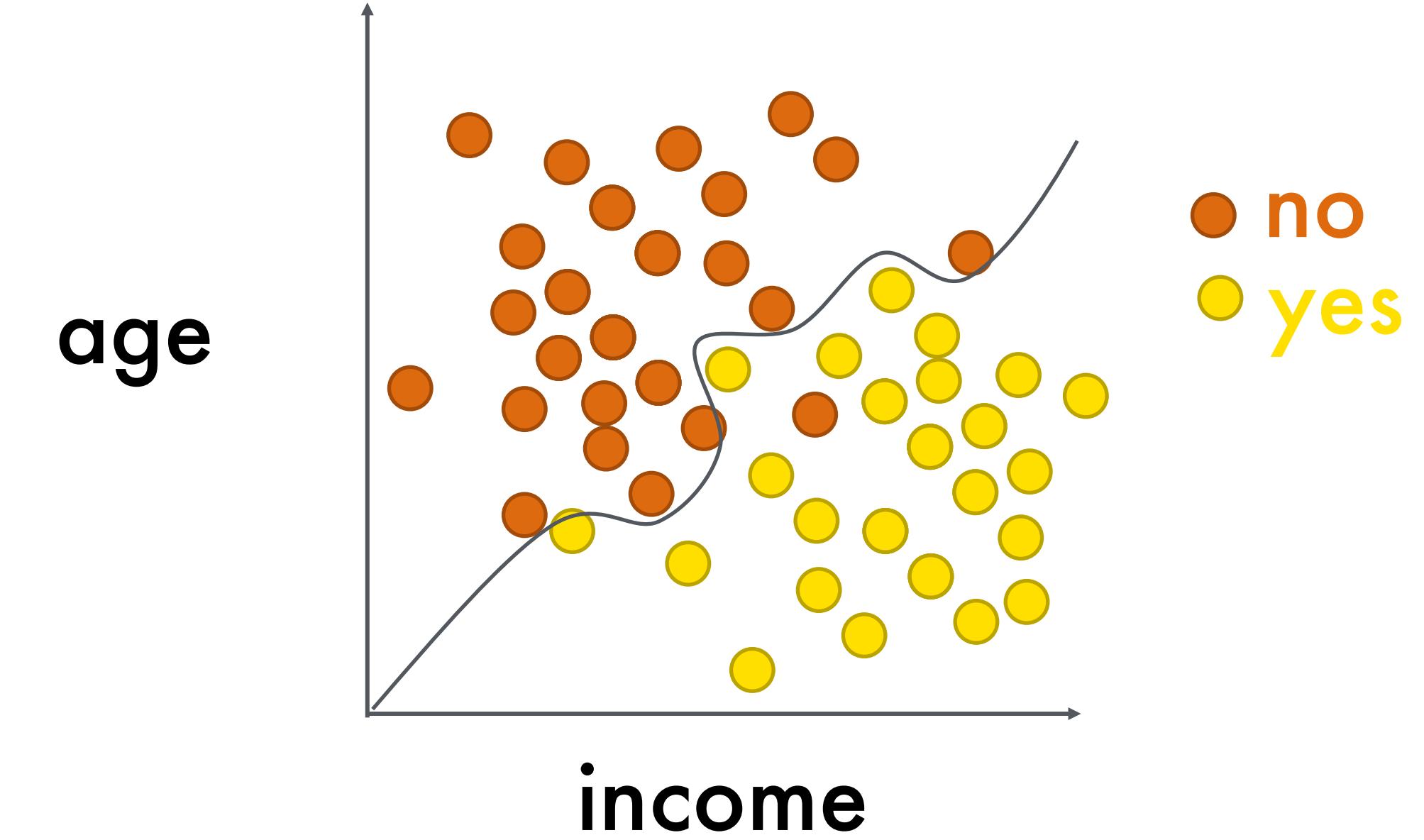
Regression:

How much will a claim cost?



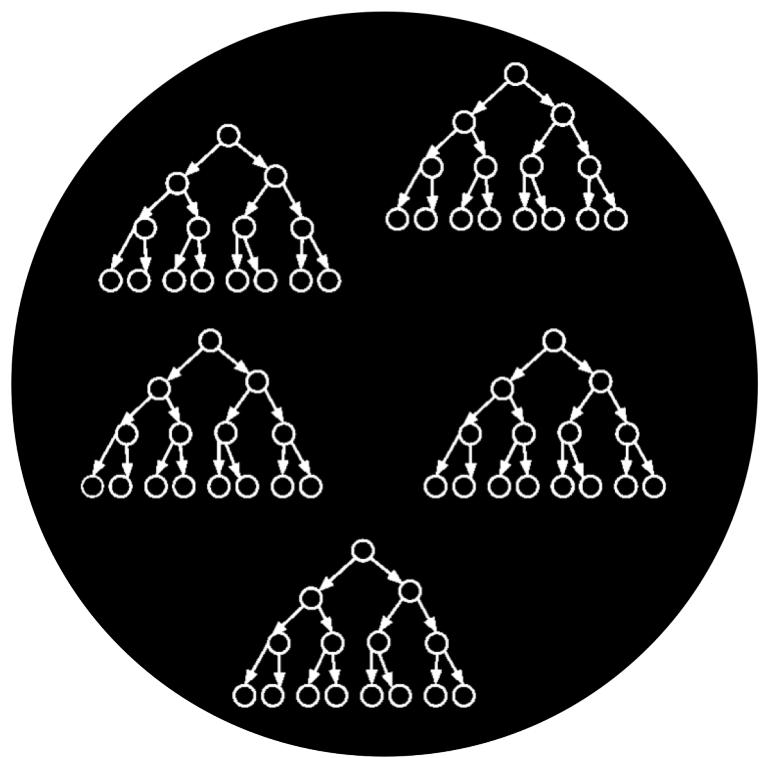
Classification:

Will a borrower pay off their loan?

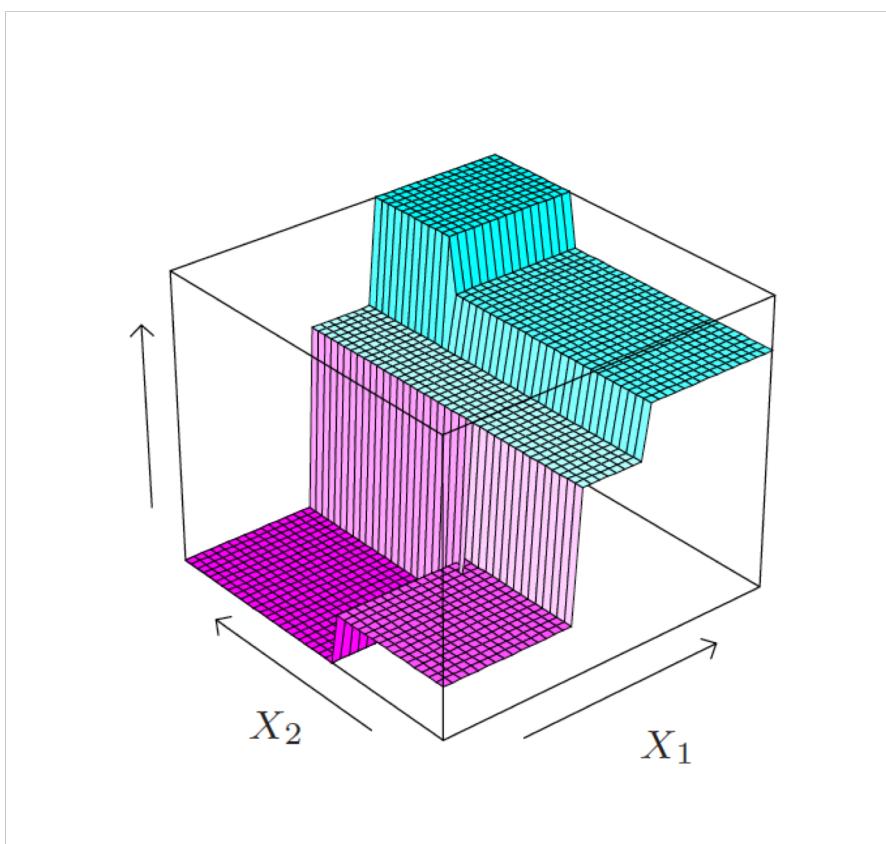


H2O AutoML: Available Algos

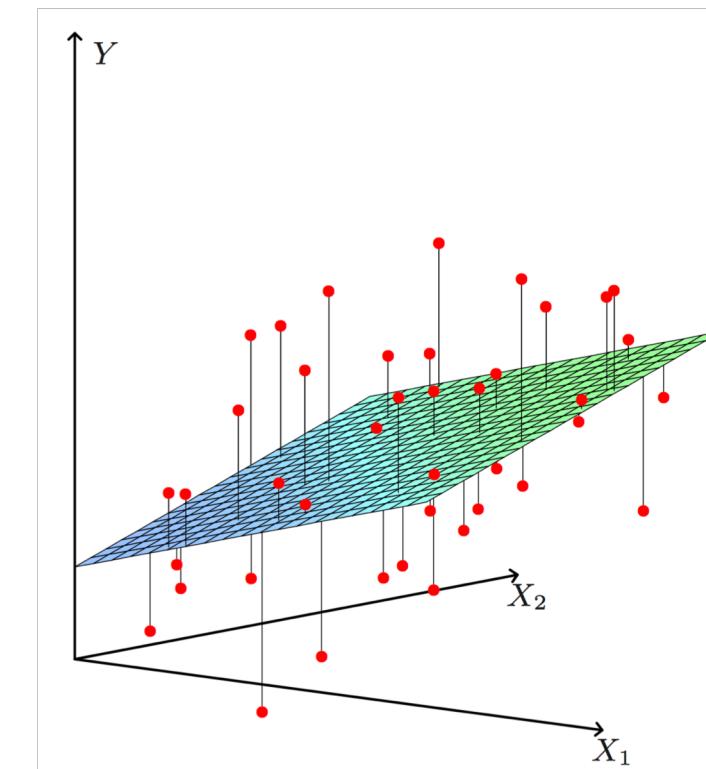
DRF
XRT



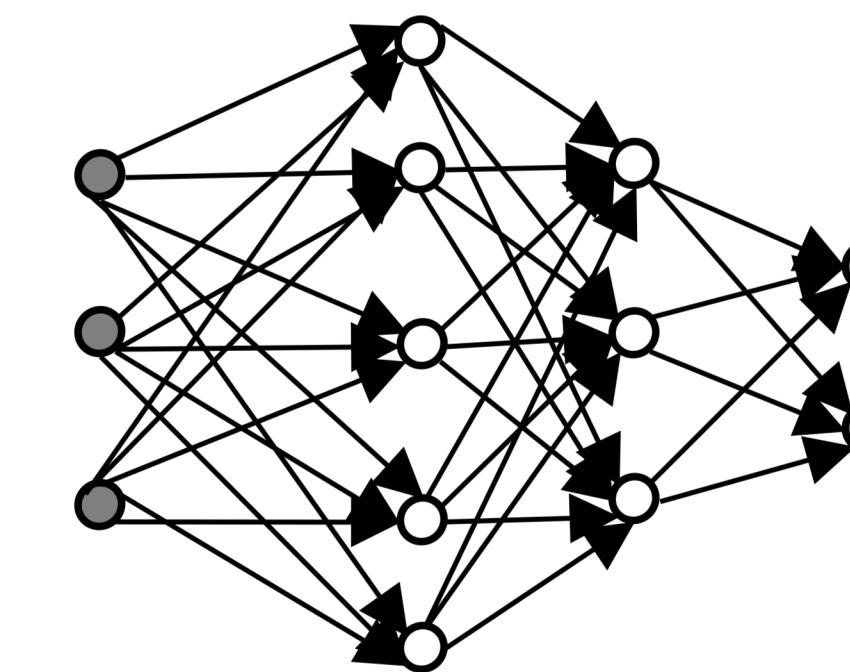
GBM
XGBoost



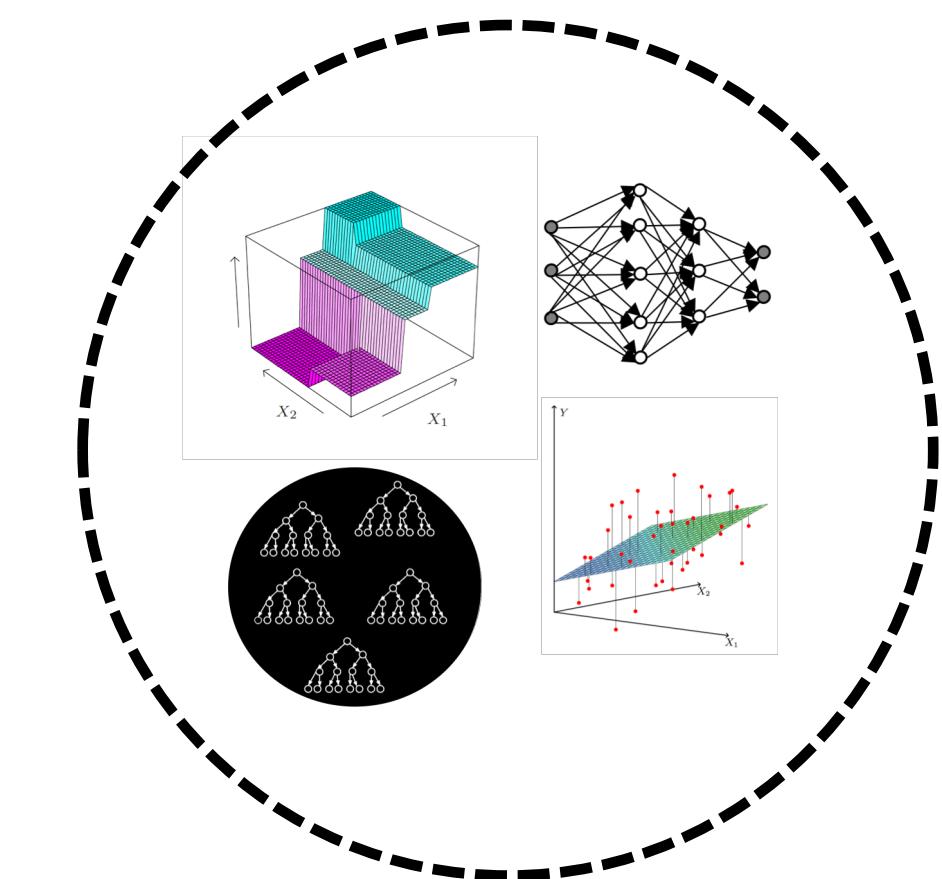
GLM



DNN



**Stacked
Ensemble**



AutoML: Data Preprocessing

Inherits from **H2O Algorithm**

Imputation

NA
10
50

→

30
10
50

Standardization

30
10
50

→

0
-1
1

One-hot Encoding

dog
cat
dog

→

1
0
1

0
1
0

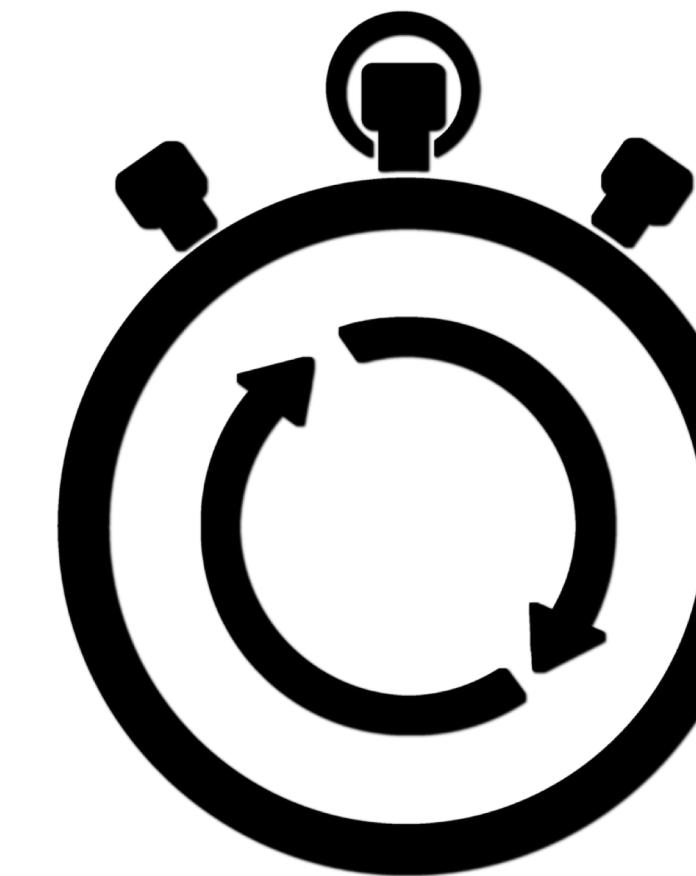
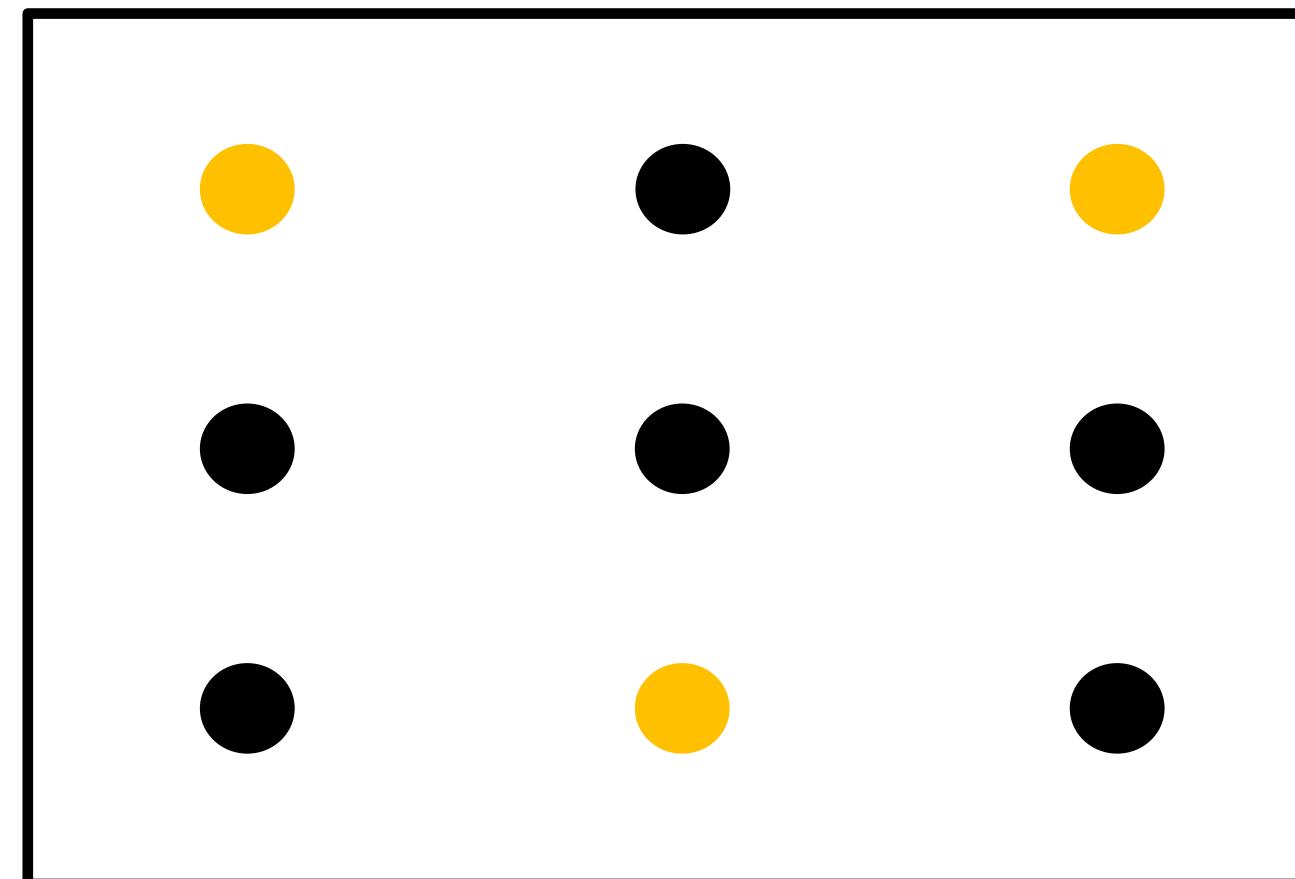
Label Encoder

dog
cat
dog

→

2
1
2

AutoML: Model Generation

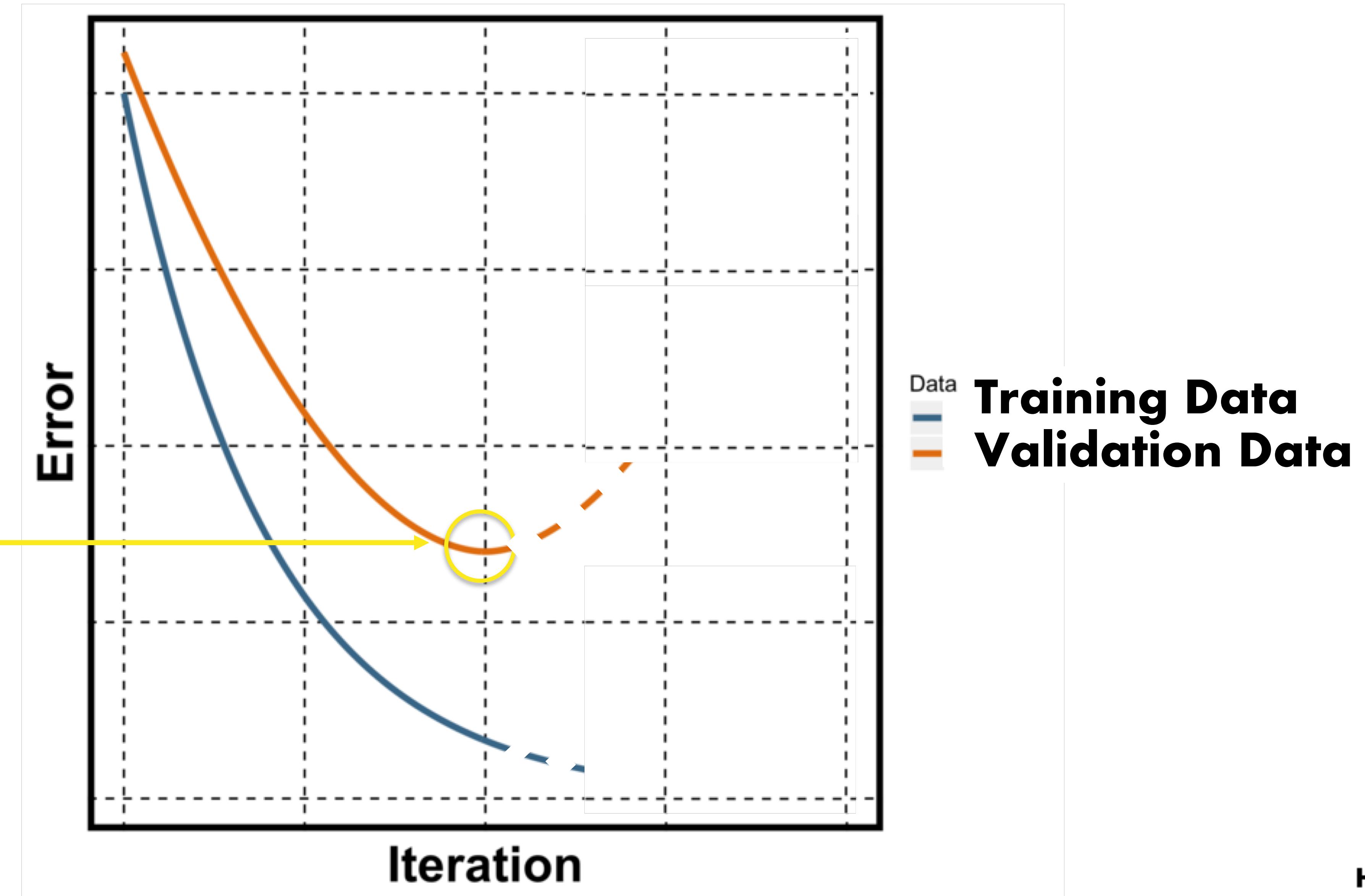


```
1037 void defaultSearchGBM(Key<Grid> gridKey) {
1038     Algo algo = Algo.GBM;
1039     WorkAllocations.Work work = workAllocations.getAllocation(algo, JobType.HyperparamSearch);
1040     if (work == null) return;
1041
1042     GBMParameters gbmParameters = new GBMParameters();
1043     setCommonModelBuilderParams(gbmParameters);
1044     gbmParameters._score_tree_interval = 5;
1045     gbmParameters._histogram_type = SharedTreeParameters.HistogramType.AUTO;
1046
1047     Map<String, Object[]> searchParams = new HashMap<>();
1048     searchParams.put("_ntrees", new Integer[]{10000});
1049     searchParams.put("_max_depth", new Integer[]{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17});
1050     searchParams.put("_min_rows", new Integer[]{1, 5, 10, 15, 30, 100});
1051     searchParams.put("_learn_rate", new Double[]{0.001, 0.005, 0.008, 0.01, 0.05, 0.08, 0.1, 0.5, 0.8});
1052     searchParams.put("_sample_rate", new Double[]{0.50, 0.60, 0.70, 0.80, 0.90, 1.00});
1053     searchParams.put("_col_sample_rate", new Double[]{0.4, 0.7, 1.0});
1054     searchParams.put("_col_sample_rate_per_tree", new Double[]{0.4, 0.7, 1.0});
1055     searchParams.put("_min_split_improvement", new Double[]{1e-4, 1e-5});
1056
1057     Job<Grid> gbmJob = hyperparameterSearch(gridKey, work, gbmParameters, searchParams);
1058     pollAndUpdateProgress(Stage.ModelTraining, "GBM hyperparameter search", work, this.job(), gbmJob);
1059 }
```

Random Grid Search

AutoML: Early Stopping

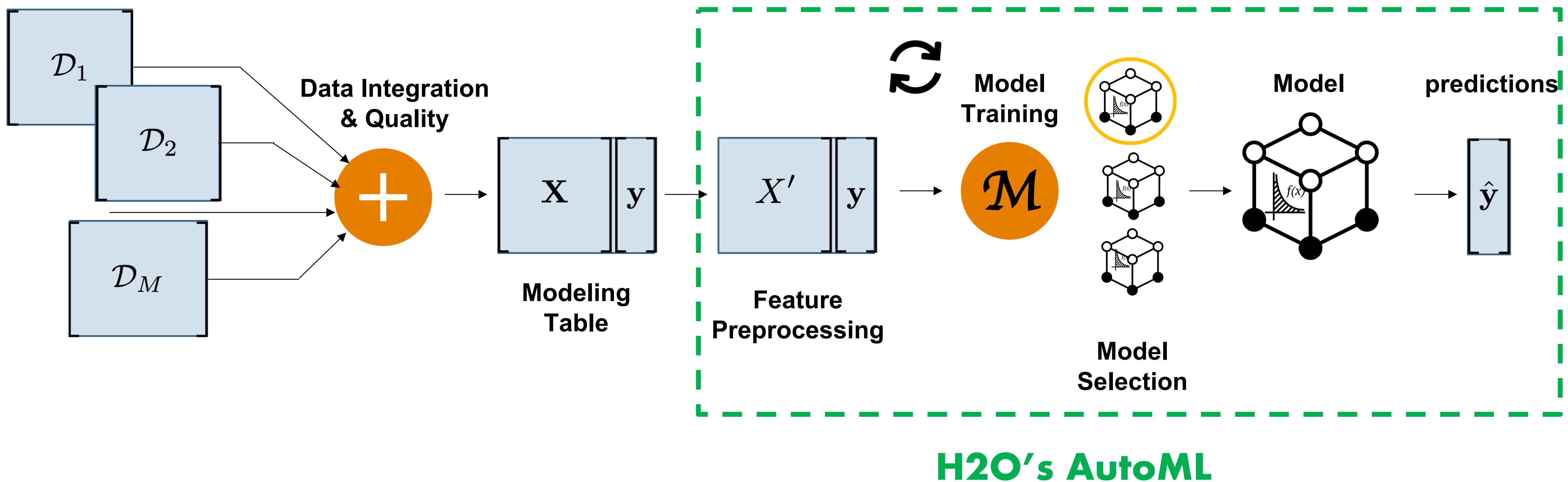
**The
Sweet
Spot**



Summary: H2O's AutoML

1. Imputation, Categorical Encoding & Standardization
2. Random Grid Search
3. Early Stopping for Individual Models
4. 2 Stacked Ensembles
5. Auto-selects Winning Model

The Machine Learning Pipeline



The Interface: Web UI

2 Required parameters training frame & response

The screenshot shows the H2O FLOW web interface with the title "AutoML in Flow". The main section is titled "Run AutoML" and contains the following configuration fields:

- Project Name: [Input field]
- Training Frame: [Select dropdown] (Select)
- Balance classes: [checkbox]
- Exclude these algorithms:
 - [checkbox] GLM
 - [checkbox] DRF
 - [checkbox] GBM
 - [checkbox] XGBoost
 - [checkbox] DeepLearning
 - [checkbox] StackedEnsemble
- Max models to build: [Input field]
- Max Run Time (sec): 3600
- Early stopping metric: AUTO
- Leaderboard sort metric: AUTO
- Early stopping rounds: 3
- Early stopping tolerance: [Input field]
- nfold: 5
- Keep cross-validation predictions:
- Keep cross-validation models:
- Keep cross-validation fold assignment:
- Seed: -1
- Checkpoints path: [Input field]

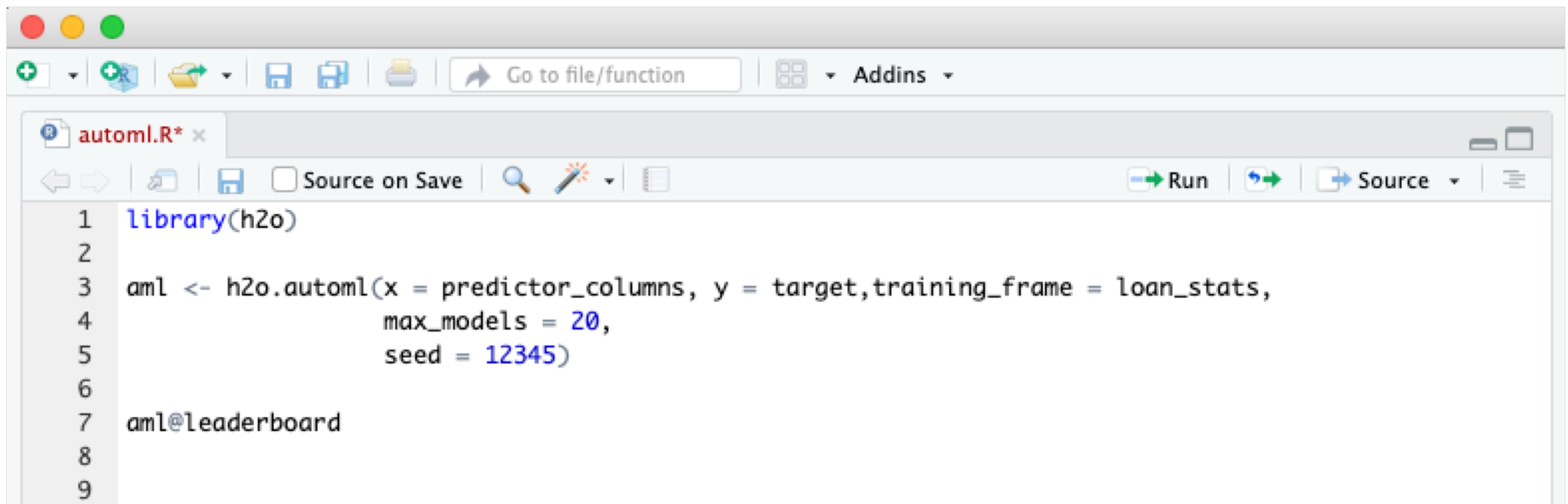
The Interface: Python

```
from h2o.automl import H2OAutoML  
  
automl = H2OAutoML(max_runtime_secs = 60, seed = 12345)  
  
automl.train(x = predictor_columns, y = target, training_frame = loan_stats)
```

```
automl.leaderboard
```

model_id	auc	logloss
StackedEnsemble_AllModels_0_AutoML_20180113_105834	0.720685	0.385048
StackedEnsemble_BestOfFamily_0_AutoML_20180113_105834	0.720319	0.385191
GLM_grid_0_AutoML_20180113_105834_model_0	0.71506	0.384576
GBM_grid_0_AutoML_20180113_105834_model_0	0.709894	0.387613
GBM_grid_0_AutoML_20180113_105834_model_1	0.703851	0.388938
GBM_grid_0_AutoML_20180113_105834_model_2	0.699872	0.391217
DRF_0_AutoML_20180113_105834	0.690366	0.406982
XRT_0_AutoML_20180113_105834	0.685729	0.39767

The Interface: R



The screenshot shows the RStudio interface with an R script file open. The file is titled "automl.R*" and contains the following code:

```
1 library(h2o)
2
3 aml <- h2o.automl(x = predictor_columns, y = target, training_frame = loan_stats,
4                     max_models = 20,
5                     seed = 12345)
6
7 aml@leaderboard
8
9
```

The RStudio toolbar at the top includes standard file operations (New, Open, Save, Print) and a "Go to file/function" search bar. Below the toolbar, the "Addins" menu is visible. The main workspace shows the R code with syntax highlighting. The bottom right corner of the window has the H2O.ai logo.

Additional Resources

Driverless AI H2O-3 Sparkling Water H2O4GPU Enterprise Steam Puddle Additional Resources

H2O-3

The H2O open source platform works with R, Python, Scala on Hadoop/Yarn, Spark, or your laptop.

H2O is licensed under the [Apache License, Version 2.0](#)

Prior releases

End User Documentation

- [H2O User Guide](#)
- [Recent Changes](#)
- [H2O README](#)
- [H2O Book \(O'Reilly\)](#)

Videos

- [Quick Start with Flow Web UI](#)
- [Quick Start with Python](#)
- [Quick Start with R](#)

Algorithms

Supervised Learning

AutoML	Tutorial	Booklet	Reference	Tuning
Cox Proportional Hazards (CoxPH)	Tutorial	Booklet	Reference	Tuning
Deep Learning (DL)	Tutorial	Booklet	Reference	Tuning
Distributed Random Forest (DRF)	Tutorial	Booklet	Reference	Tuning
Generalized Linear Modeling (GLM)	Tutorial	Booklet	Reference	Tuning
Gradient Boosting Machine (GBM)	Tutorial	Booklet	Reference	Tuning
Naive Bayes	Tutorial	Booklet	Reference	Tuning
Stacked Ensembles	Tutorial	Booklet	Reference	Tuning
XGBoost	Tutorial	Booklet	Reference	Tuning

Unsupervised Learning

Aggregator	Tutorial	Reference
Generalized Low Rank Models (GLRM)	Tutorial	Reference
K-Means Clustering	Tutorial	Reference
Isolation Forest	Tutorial	Reference
Principal Component Analysis (PCA)	Tutorial	Reference

Miscellaneous

Word2Vec	Tutorial	Reference
----------	--------------------------	---------------------------

H2O.ai
3.22.1.4

Search docs

Welcome to H2O 3

Quick Start Videos

Cloud Integration

Downloading & Installing H2O

Starting H2O

Getting Data into Your H2O Cluster

Data Manipulation

Algorithms

Cross-Validation

Variable Importance

Grid (Hyperparameter) Search

Checkpointing Models

Performance and Prediction

AutoML: Automatic Machine Learning

AutoML Interface

AutoML Output

Saving and Loading a Model

Docs » AutoML: Automatic Machine Learning

[View page source](#)

AutoML: Automatic Machine Learning

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software that can be used by non-experts. The first steps toward simplifying machine learning involved developing simple, unified interfaces to a variety of machine learning algorithms (e.g. H2O).

Although H2O has made it easy for non-experts to experiment with machine learning, there is still a fair bit of knowledge and background in data science that is required to produce high-performing machine learning models. Deep Neural Networks in particular are notoriously difficult for a non-expert to tune properly. In order for machine learning software to truly be accessible to non-experts, we have designed an easy-to-use interface which automates the process of training a large selection of candidate models. H2O's AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.

H2O's AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit. **Stacked Ensembles** – one based on all previously trained models, another one on the best model of each family – will be automatically trained on collections of individual models to produce highly predictive ensemble models which, in most cases, will be the top performing models in the AutoML

<http://docs.h2o.ai>

H₂O.ai

Thanks!