Using H2O AutoML for Kaggle Competitions

- AutoML? Does it work?
- About H_2O AutoML
- Q & A

Jo-fai (Joe) Chow
Data Scientist at H2O.ai
joe@h2o.ai

About Me

- Civil (Water) Engineer
  - 2010 – 2015
    - Consultant (UK)
      - Utilities
      - Asset Management
      - Constrained Optimization
    - EngD (Industrial PhD) (UK)
      - Infrastructure Design Optimization
      - Machine Learning +
        Water Engineering
      - Discovered H_2O in 2014

- Data Scientist
  - 2015 – 2016
    - Virgin Media (UK)
    - Domino Data Lab (Silicon Valley)
  - 2016 – Present
    - H_2O.ai (Silicon Valley)
  - How?
    - bit.ly/joe_kaggle_story

H_2O AutoML: Does it work?

4	▲3	Juan Zhai 巻宅	0.06335...	53	2d
5	▲14	Trottefox	0.06359...	86	18h
6	▲319	cmanning	0.06362...	11	7d
7	▲90	Benchmark	0.06362...	324	13h
8	▼7	R2	0.06366...	352	2d
9	▲2	Zidmie & Kostadinov & L	0.06378...	273	9h
10	▼8	Nima Shahbazi | mcha...	0.06378...	251	12h
11	▲454	FF	0.06383...	14	9h
12	▼7	Zensemble	0.06384...	253	1d
13	▼4	KFP	0.06387...	349	18h
14	▲369	raytrace	0.06388...	47	16h
15	▲46	To Train Them Is My C...	0.06390...	66	1d
16	▲154	Batangas	0.06393...	75	13h
17	▲265	Thomas Hoffmann	0.06394...	67	13h
18	▼5	Ivonik	0.06394...	118	9h
19	▼13	Belinda Trotta	0.06394...	47	3d
20	▼5	Thomas H. Thoresen | ...	0.06397...	113	9h
21	▼1	Bierkom	0.06398...	72	20h
22	▲23	Gough	0.06398...	56	9h
23	▼13	Alpha 60	0.06400...	180	18h
24	▼12	The Slippery Appraisal...	0.06400...	363	2d
25	▼8	Dmitry Kulagin	0.06400...	27	2d
26	▲1476	Bin	0.06400...	7	1d
27	▲241	Bram Boroson	0.06400...	99	9h
28	—	ys	0.06401...	81	10h
29	new	no one	0.06401...	25	15h
30	▼9	双鸭山数据科学RUA小分...	0.06401...	175	10h
31	▲6	VincaPiggy	0.06401...	101	13h
32	▲6	Comment Allez-Vous	0.06401...	186	13h
33	▲88	mlin	0.06401...	28	9h
34	▲19	...	0.06402...	135	1d
35	▲35	proof by adverb	0.06402...	45	13h
36	▼7	Helgi	0.06402...	197	9h
37	▲90	Deal or No Deal	0.06402...	79	9h

Your Best Entry ↑

Your submission scored 0.0640257, which is an improvement of your previous score of 0.0640259. Great

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs •••

Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Zillow · 3,839 teams · 3 months to go

$1,200,000
Prize Money

Some of the H_2O Kagglers

Marios Michailidis (KazAnova)
Mathias Müller (Faron)

Dmitry Larko
… and his father

Joe … trying to catch up …
Used AutoML a lot to save time
37 out of 3839 (Top 1%)

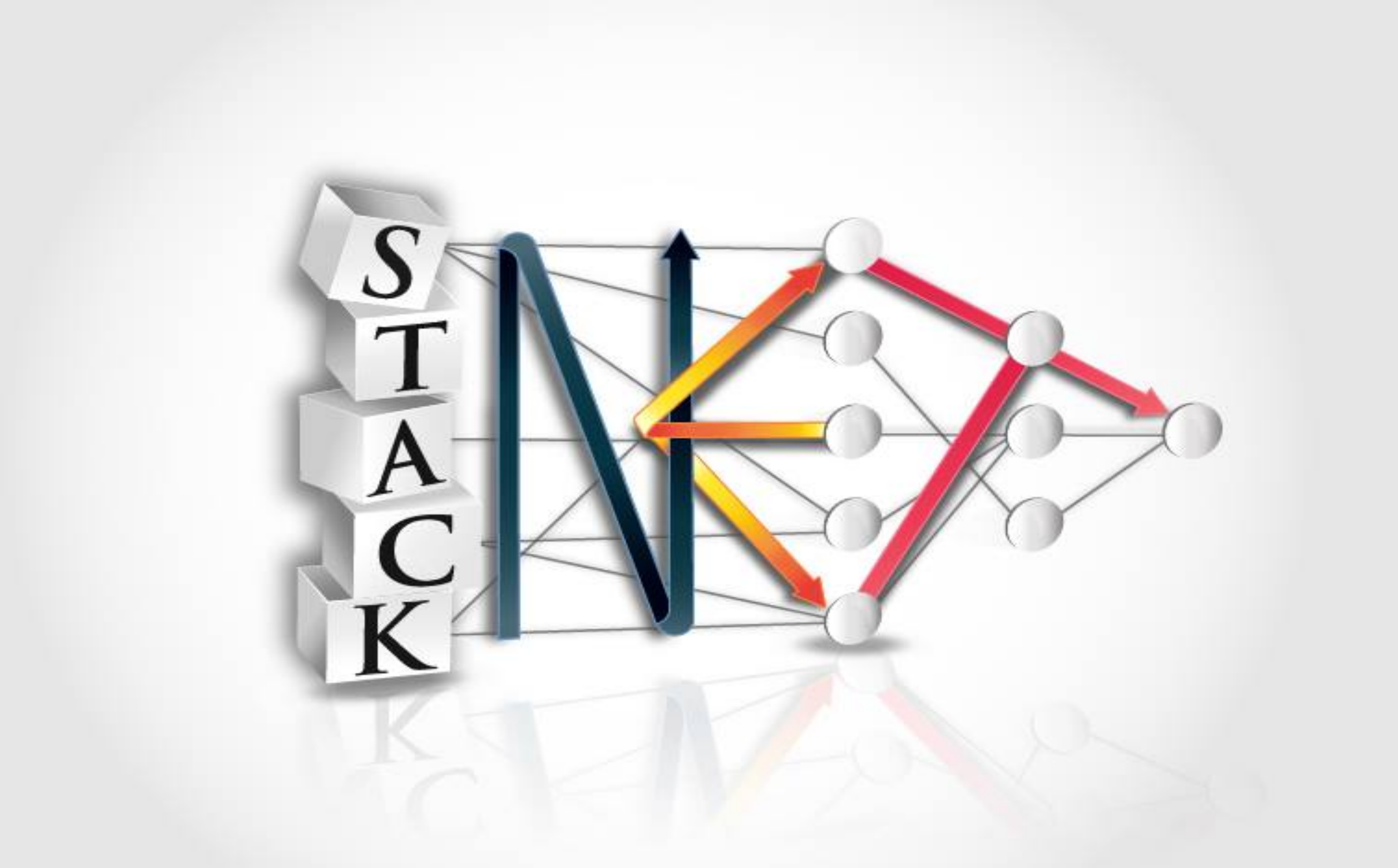
H2O.ai

Does it work with other tools?

Does it work with other tools?

YES – I used H_2O and StackNet together

H2O.ai



Introducing StackNet Meta-Modelling Framework

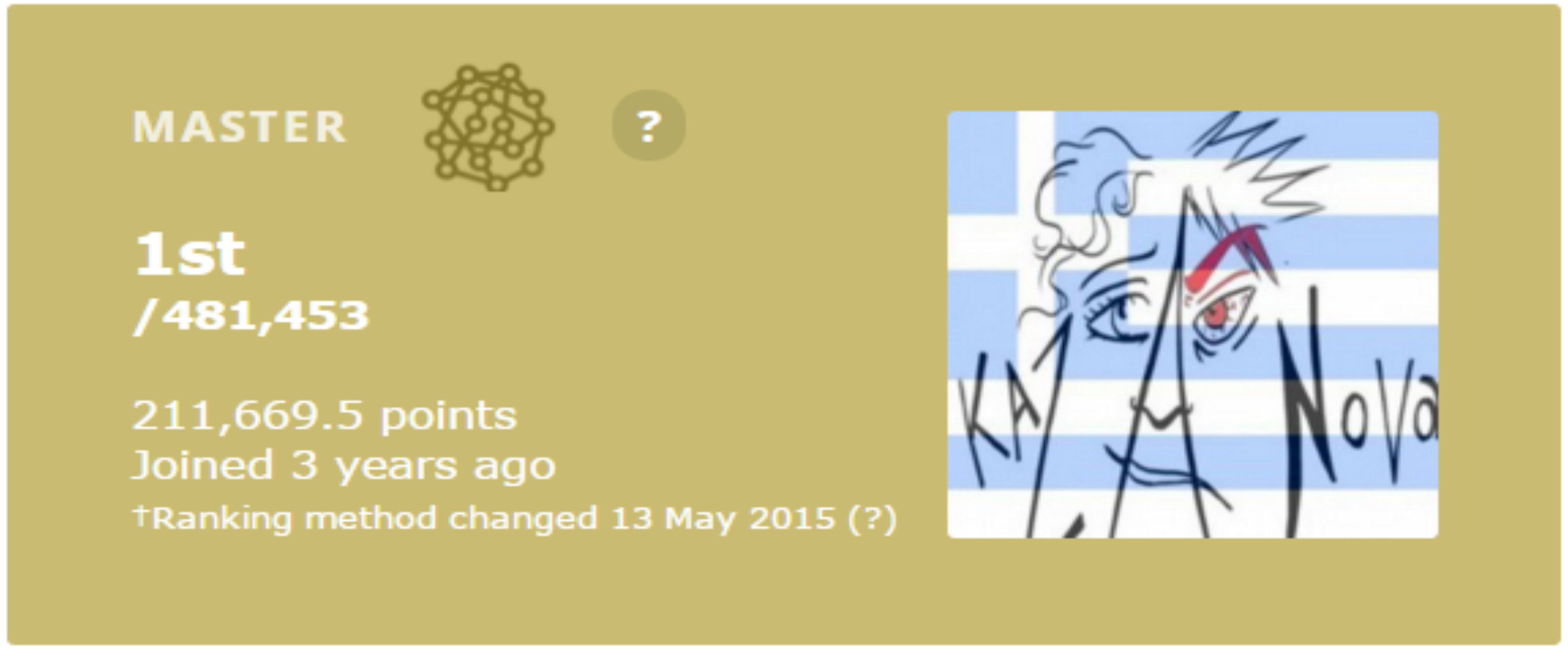
Marios Michaildis

Research Data Scientist at

Email: marios@h2o.ai

Why bother learning more about StackNet?

- It helps to **improve predictions** given the same input data
- Its is **educational** in its own way, especially in understanding Stacking.
- Compiles the **pinnacle of machine learning** into one framework-and-library.
- Has **won 2 kaggle competitions** (link A and Link B)
- Has helped many people **get top 10** results in kaggle.
- It has helped me become **kaggle #1**



About H2O AutoML

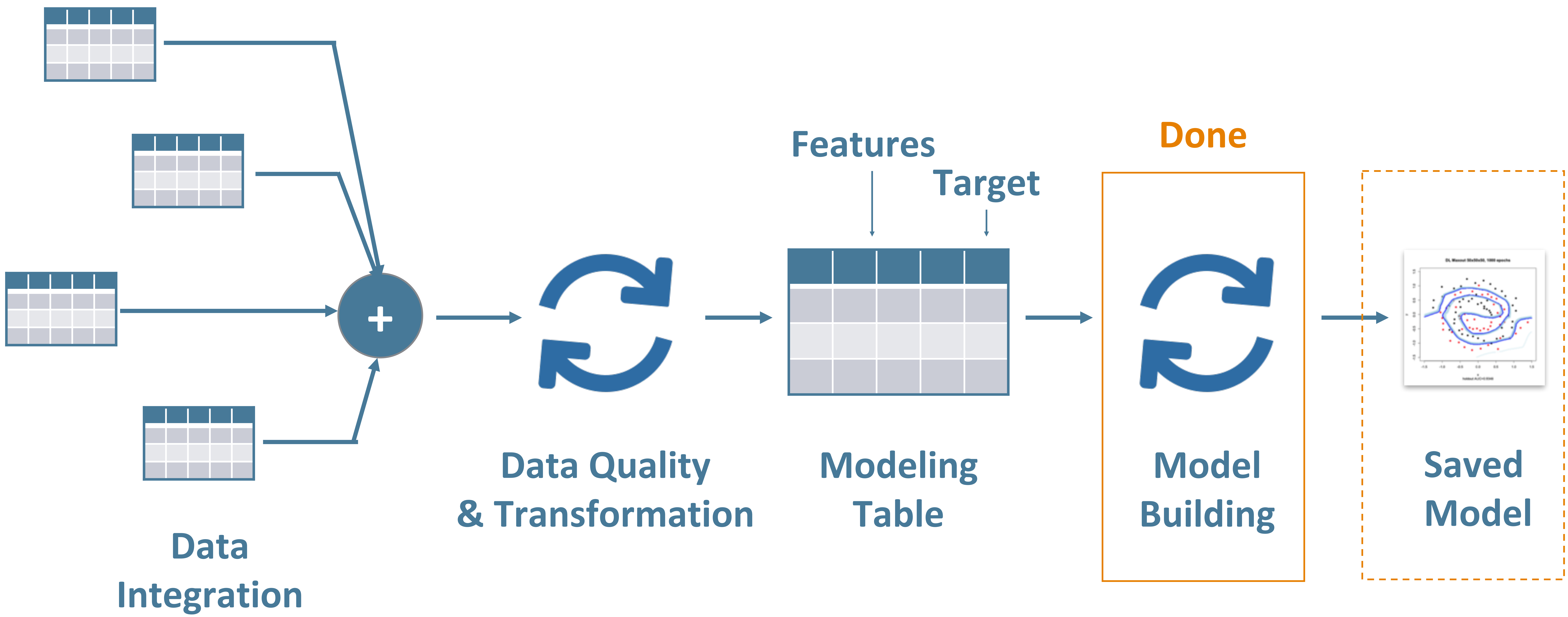
Scalable Automatic Machine Learning

Why Use AutoML?

Automates Model Building Workflow

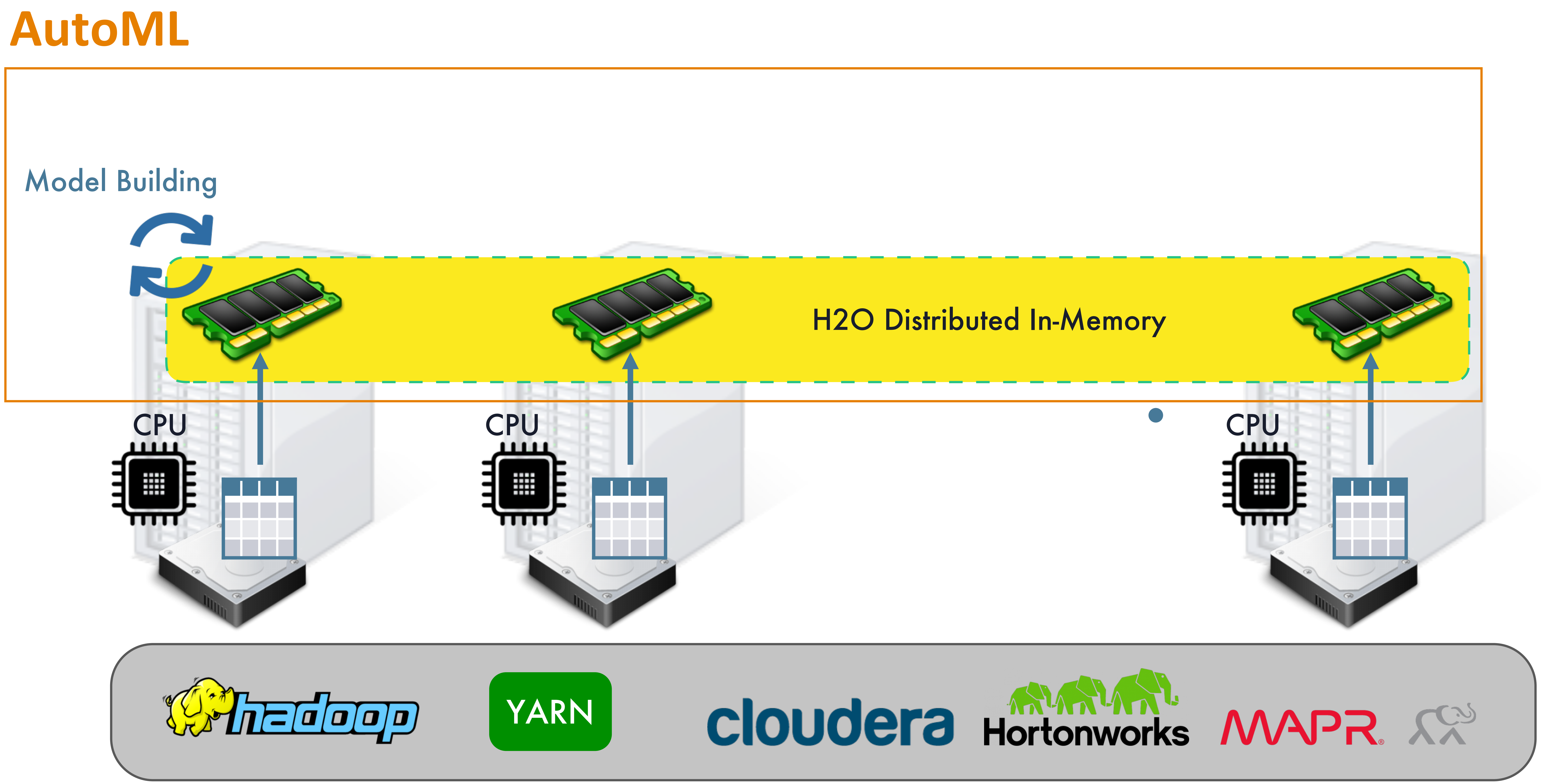
- Includes **Automatic training** & **tuning** of a large selection of candidate models
- Allows for user-specified performance metric-based **Stopping criterion** or time-limit
- Provides **Real-time monitoring** of model building progress
- Includes highly predictive **Stacked Ensembles** trained on collection of models

What is Completed?

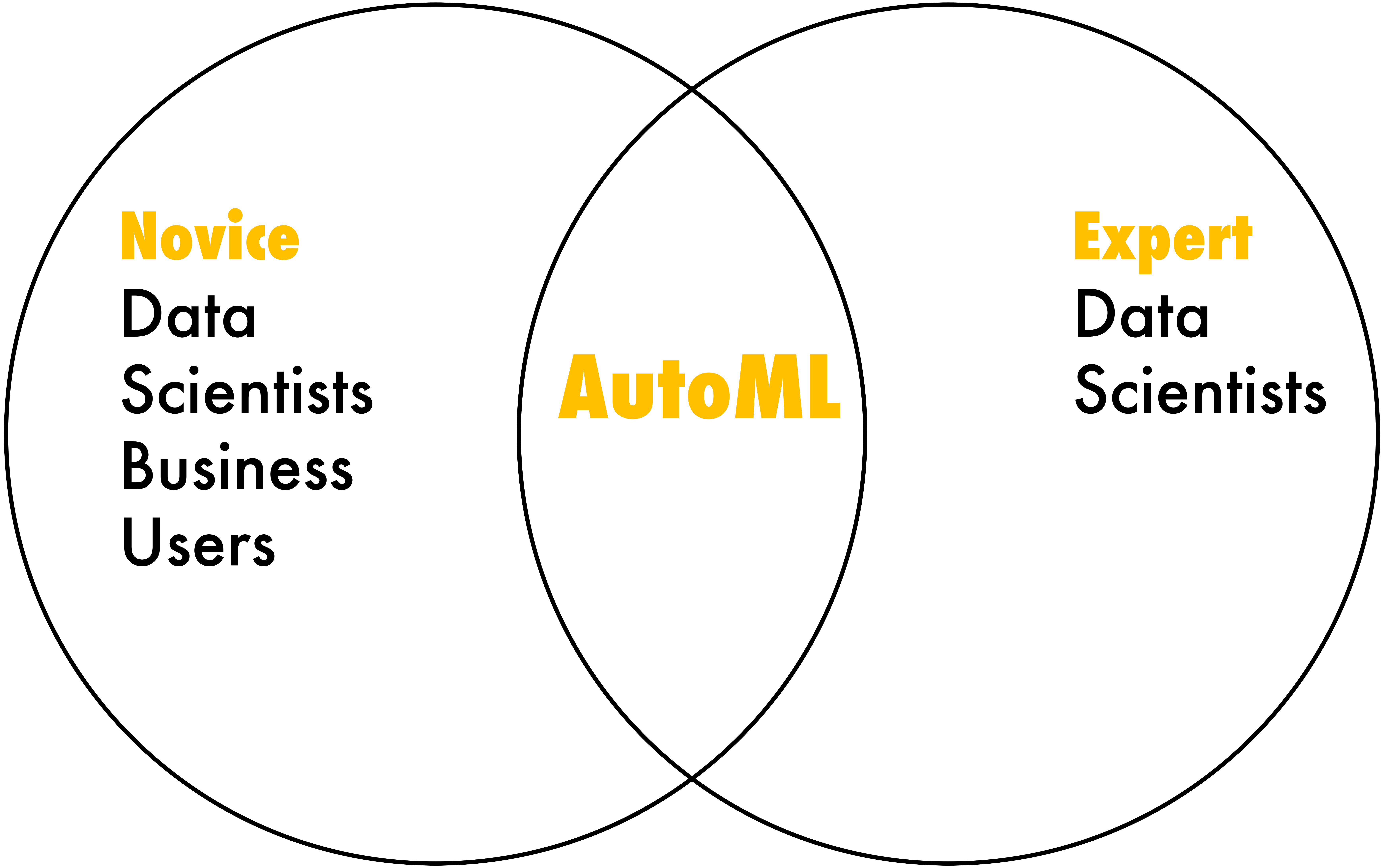


H2O.ai

Fast: Distributed & In-Memory



Who is it For?

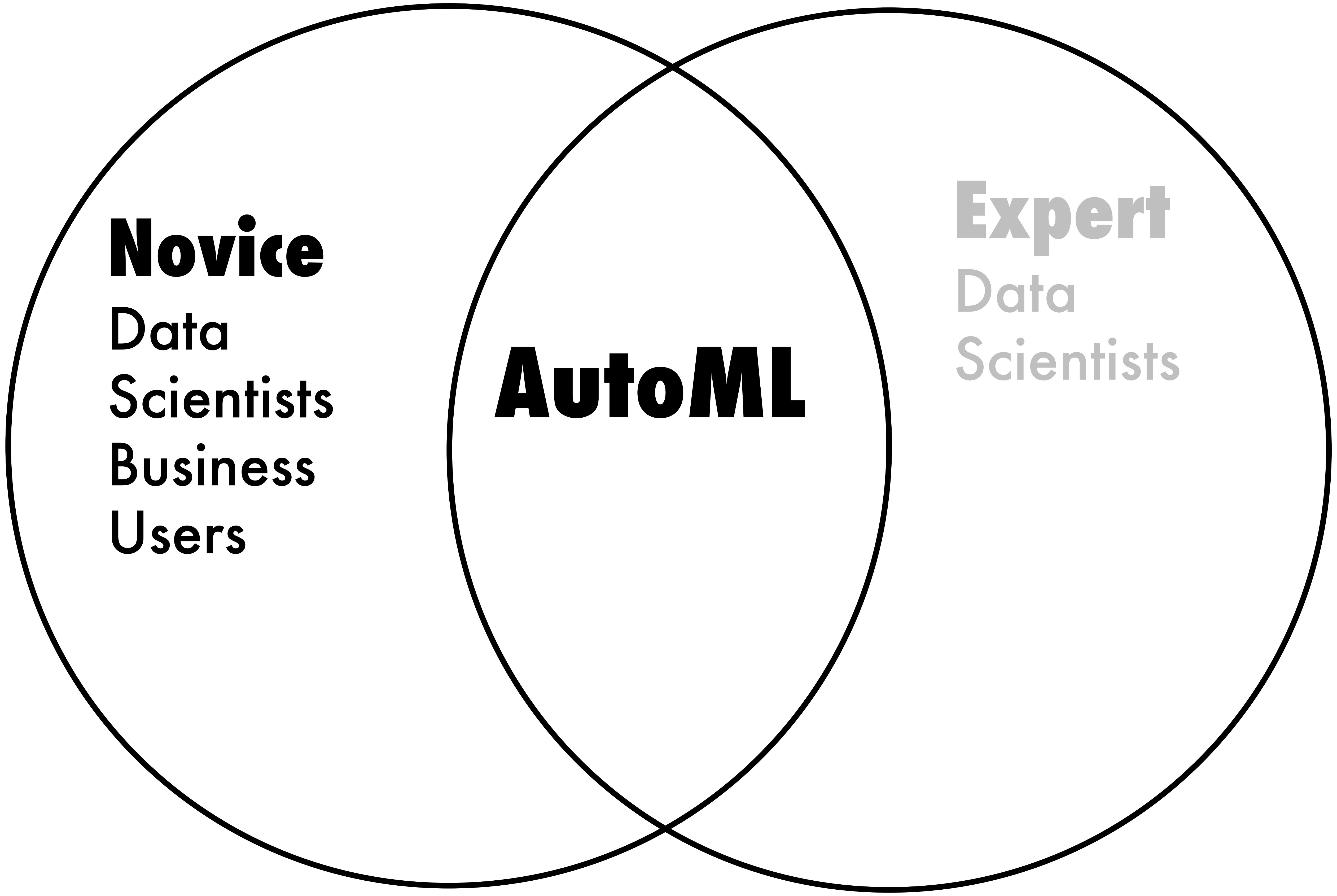


H2O.ai

Who is it For?

AUTOMATES

- **basic preprocessing**
- **model training**
- **hyperparameter tuning**
- **stacking**
- **model results table**



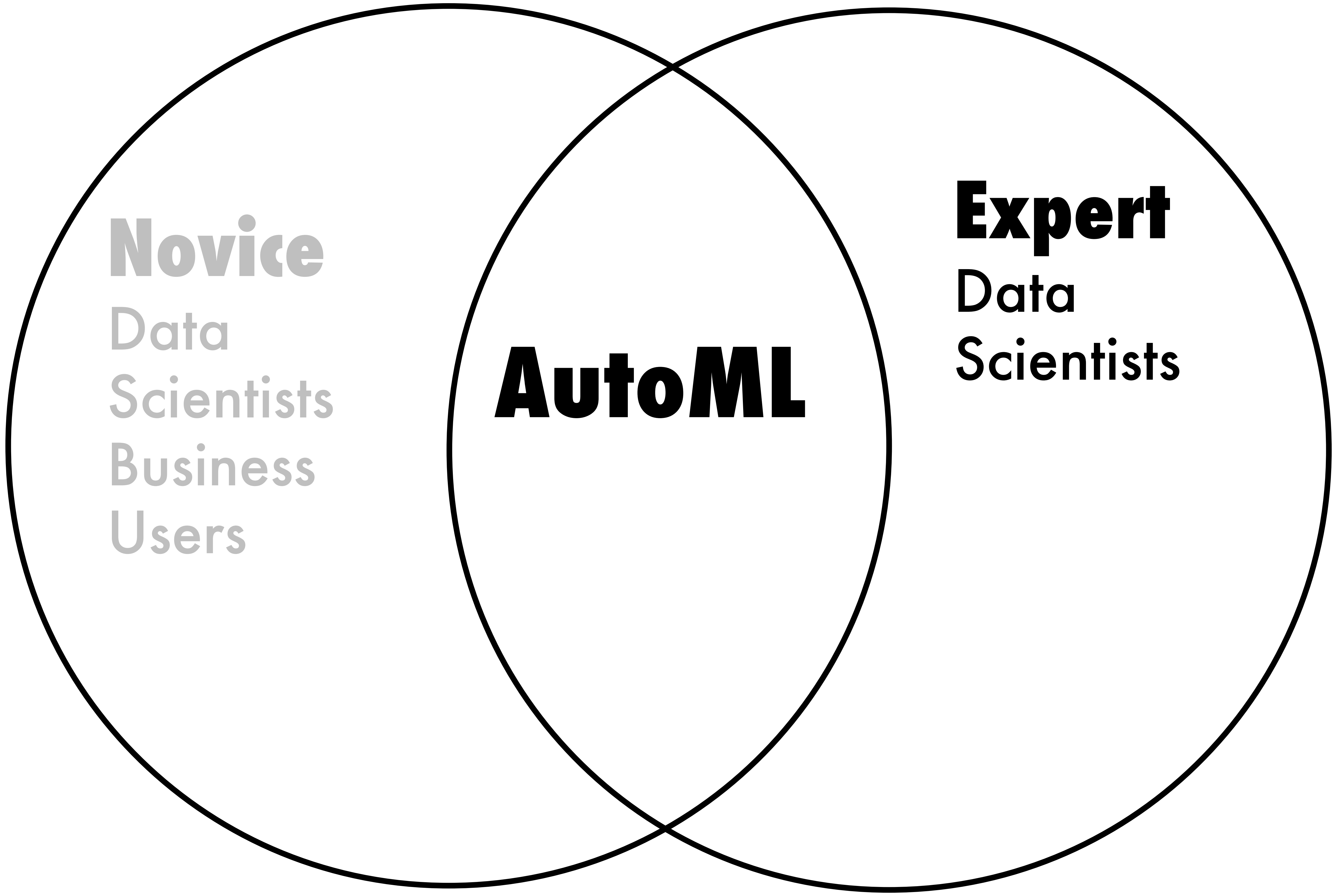
FREES TIME FOR

- data-preprocessing
- feature engineering
- model deployment

Who is it For?

AUTOMATES

- basic preprocessing
- model training
- tuning with validation
- stacking
- model results table



FREES TIME FOR

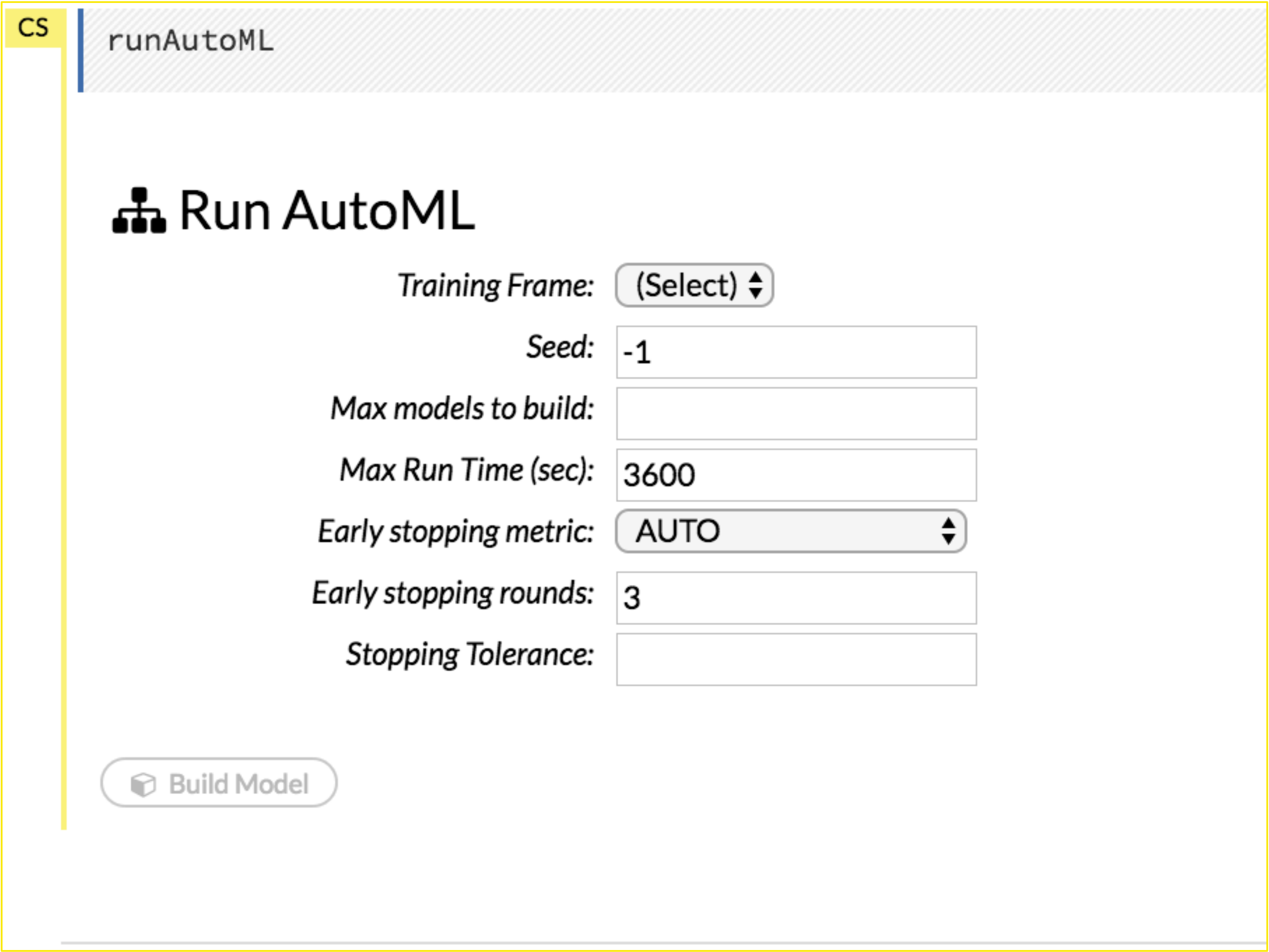
- data-preprocessing
- feature engineering
- model deployment

The Interface

Simplify Machine Learning

2 Required parameters

training frame & response



The Interface

R

```r
# Identify predictors and response
y <- "response"
x <- setdiff(names(train), y)

aml <- h2o.automl(x = x, y = y,
                  training_frame = train,
                  leaderboard_frame = test,
                  max_runtime_secs = 30)

# View the AutoML Leaderboard
lb <- aml@leaderboard
lb
```

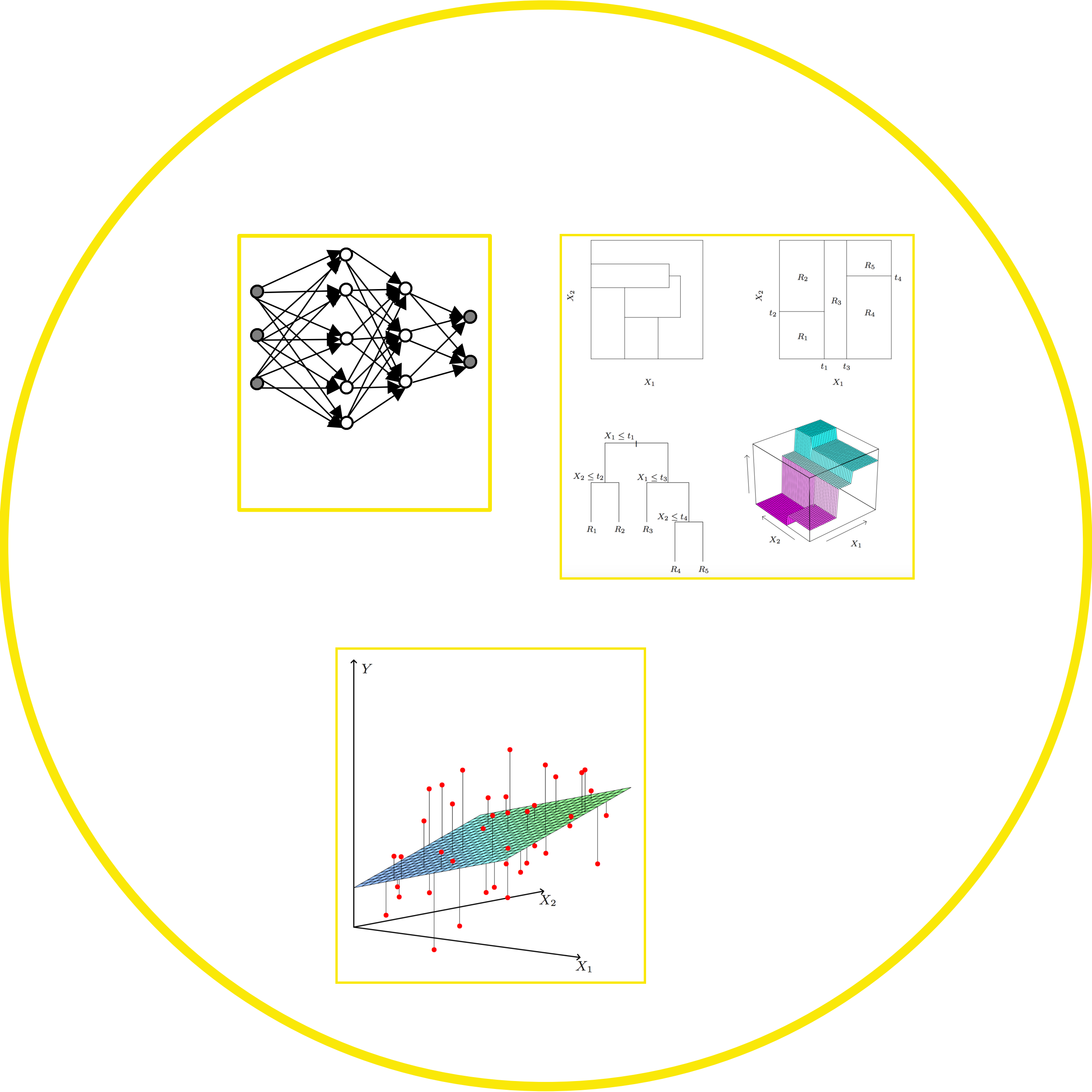
PYTHON

```python
# Identify predictors and response
x = train.columns
y = "response"
x.remove(y)

# Run AutoML for 30 seconds
aml = H2OAutoML(max_runtime_secs = 30)
aml.train(x = x, y = y,
          training_frame = train,
          leaderboard_frame = test)

# View the AutoML Leaderboard
lb = aml.leaderboard
lb
```

Behind the Scenes

Grid Search

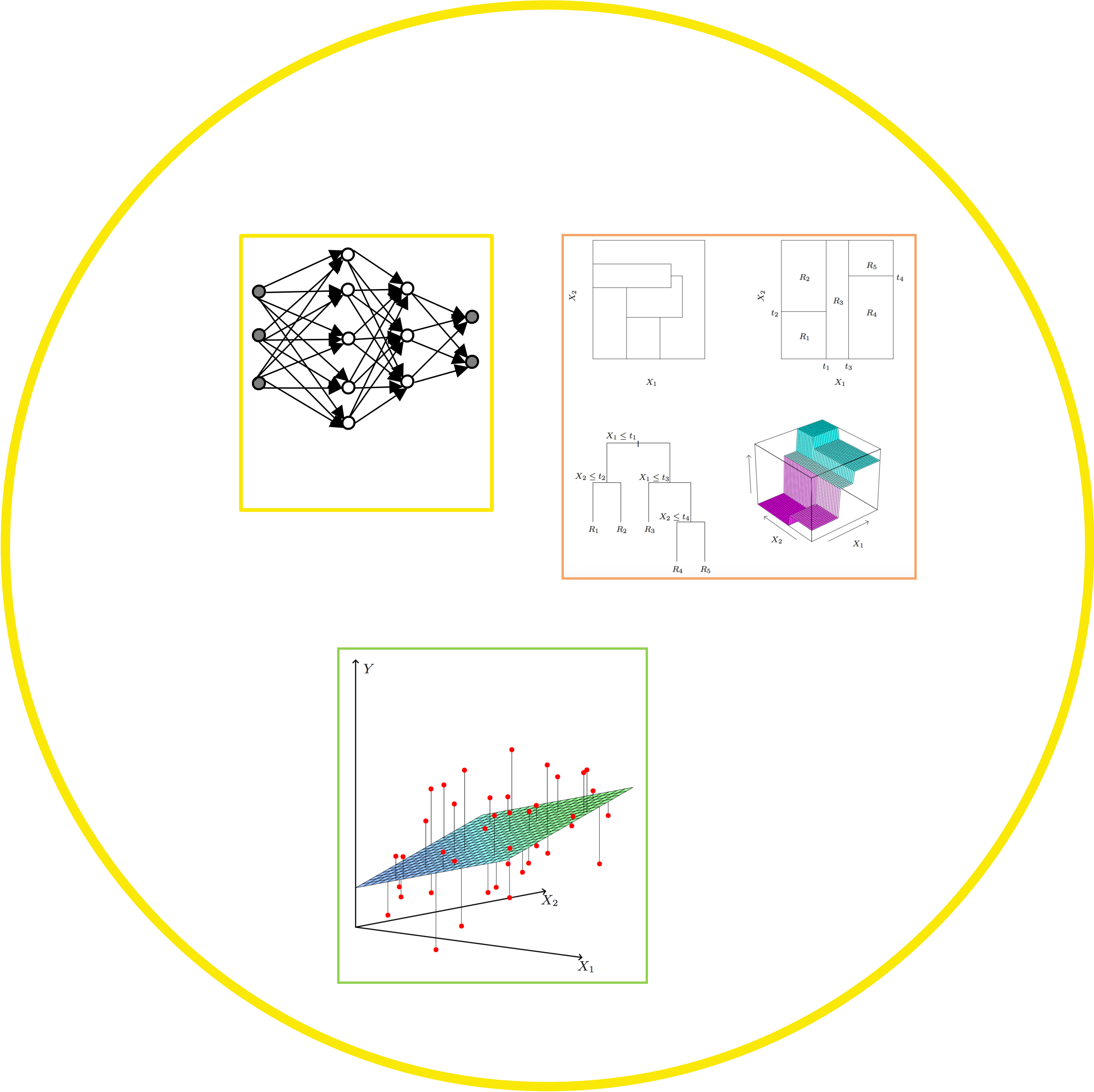
- Large selection of models
- Hyperparameter tuning
- Early Stopping



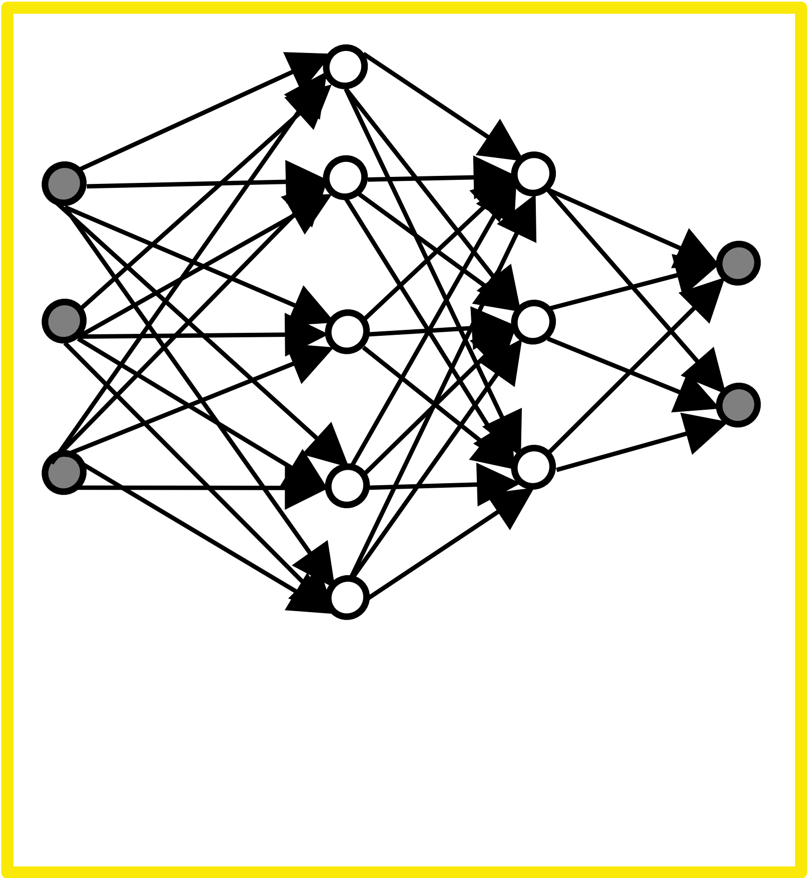
Behind the Scenes

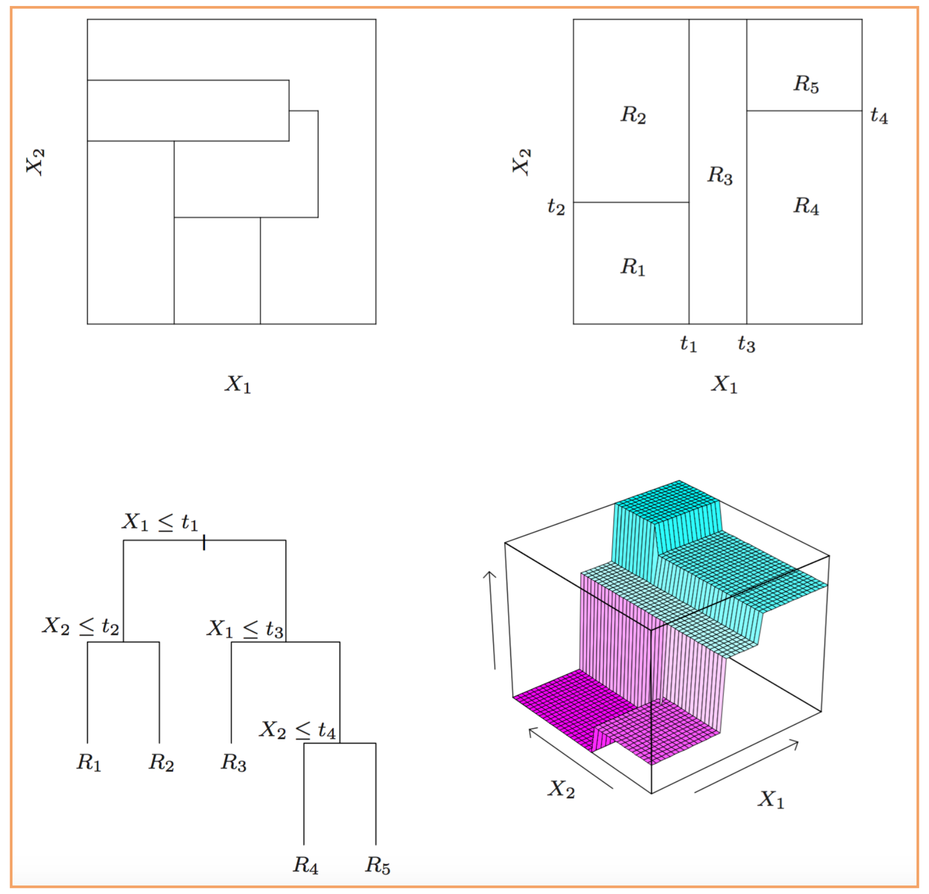
Stacked Ensemble

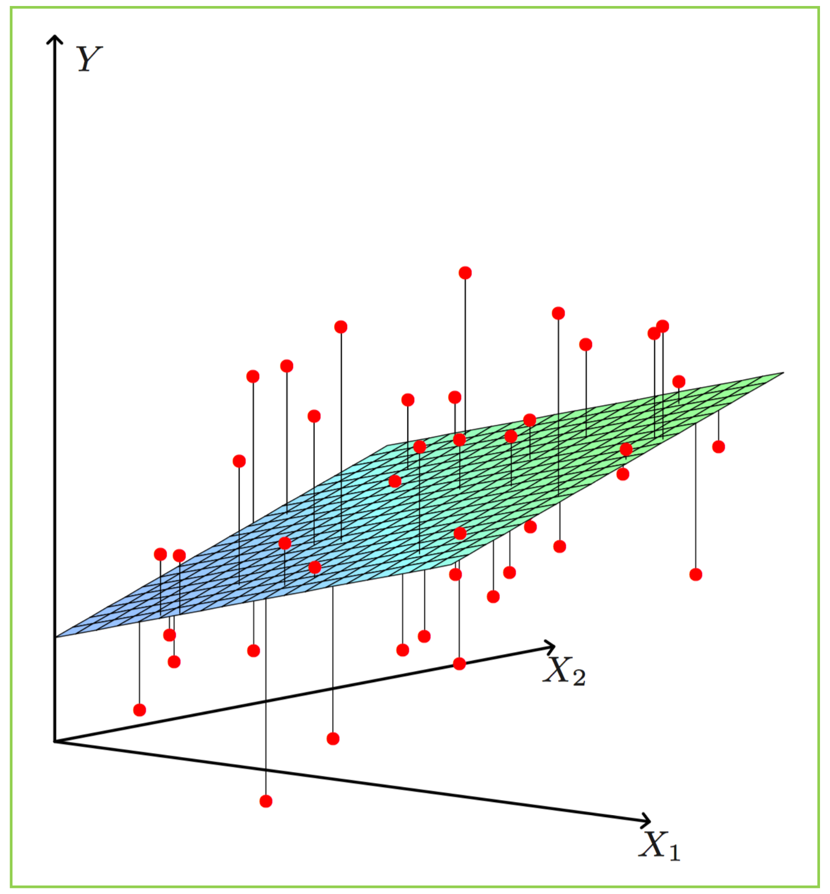
- Highly predictive ensemble trains on all the models



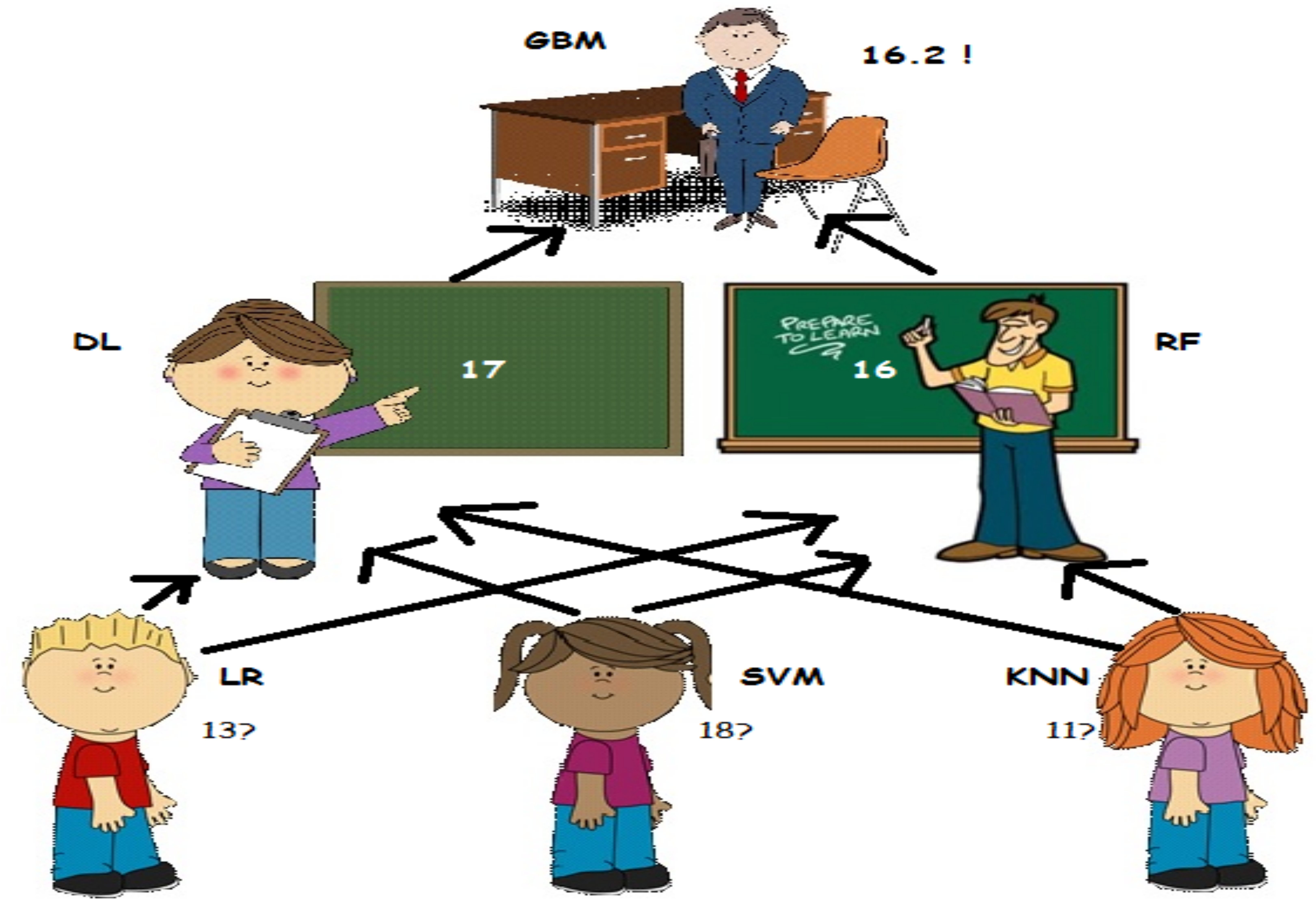
Stacking Base Learners







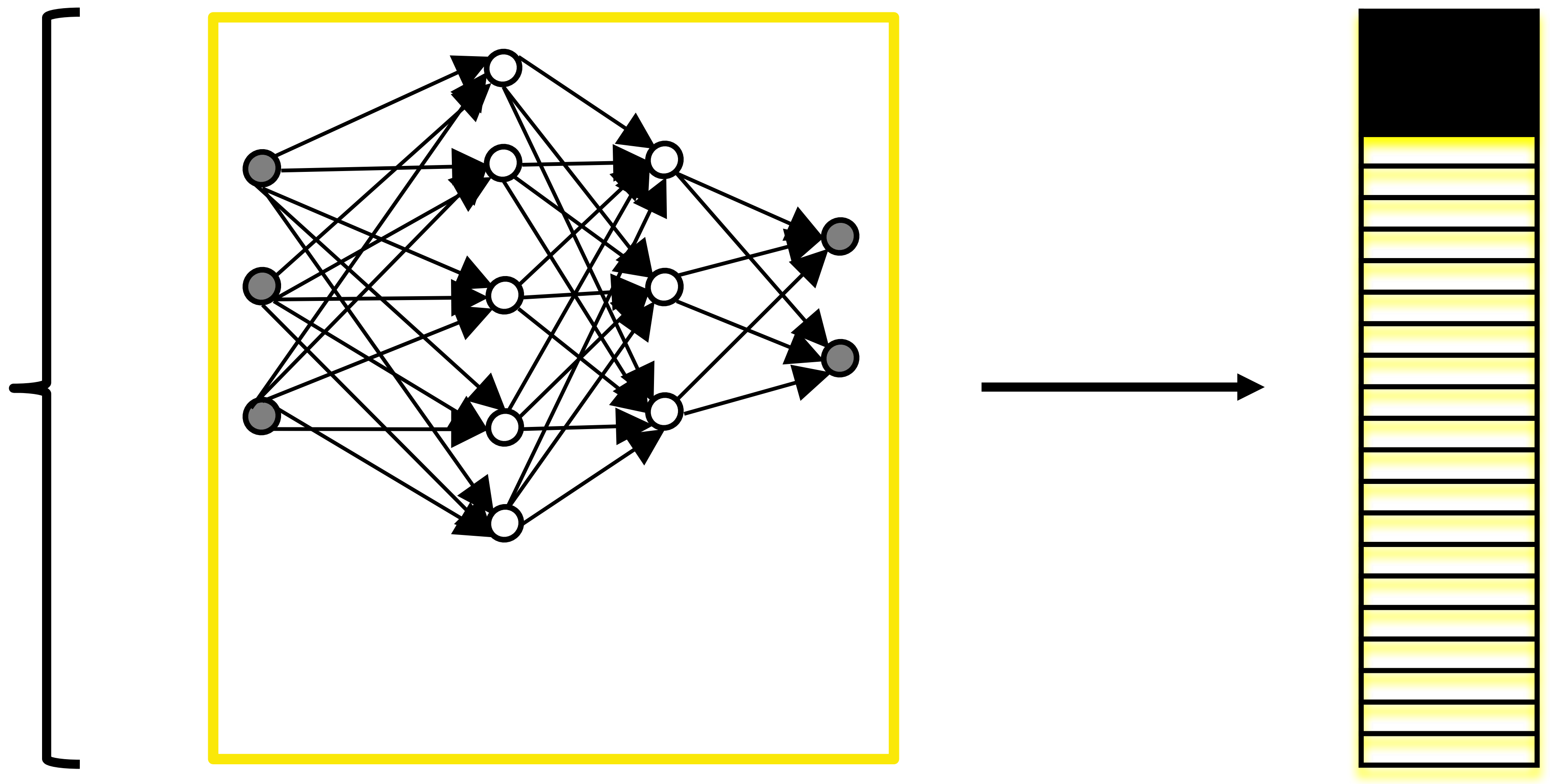
Why meta modelling?



H2O.ai

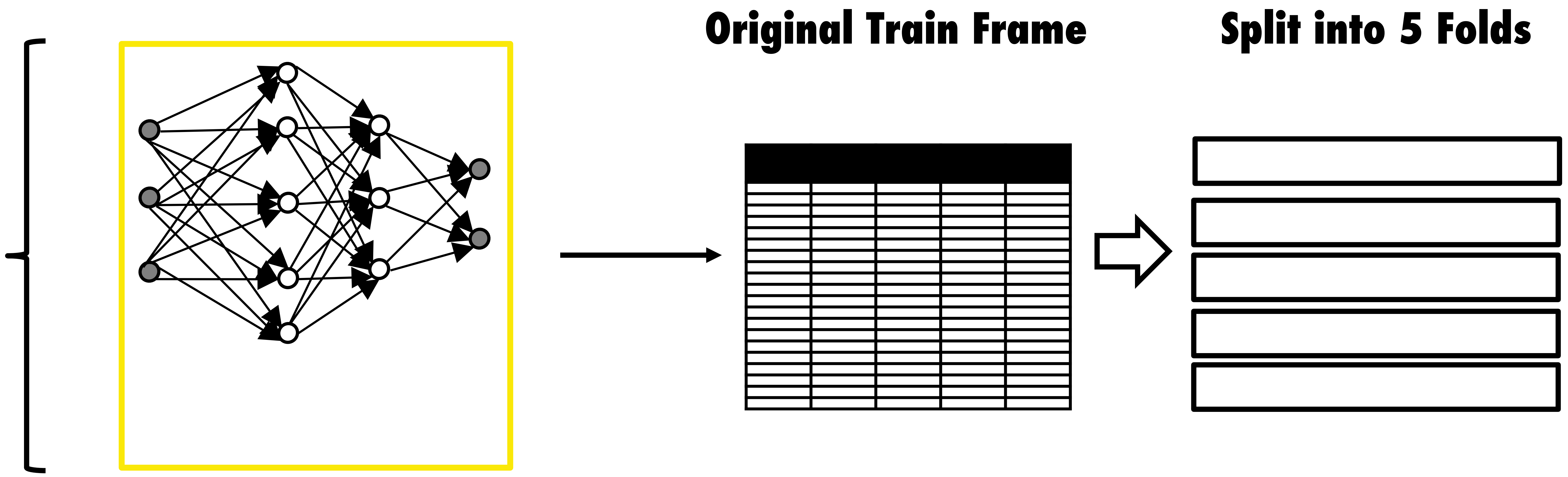
Base Learner Results

CV Prediction Results Column



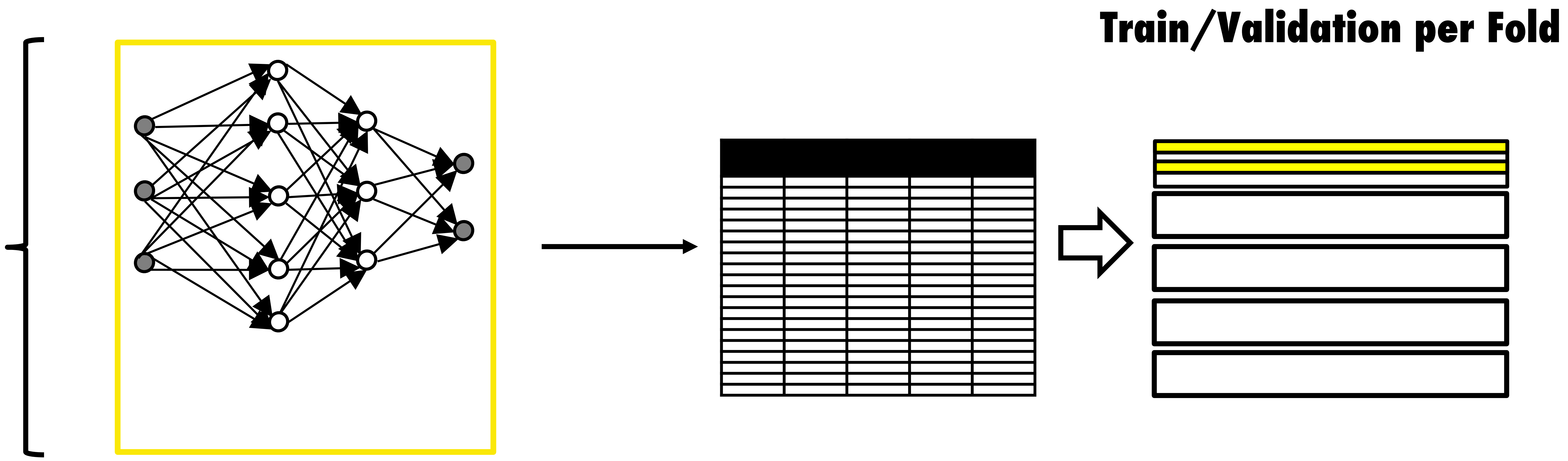
Get CV Prediction Column

Split Dataset



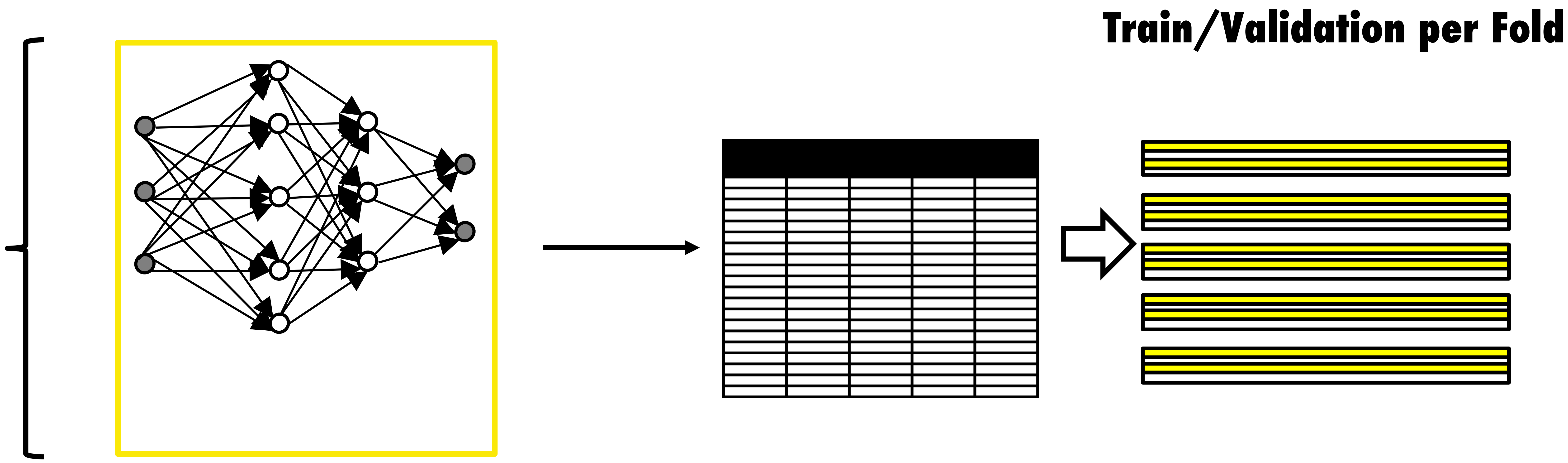
Behind the Scenes

Split into Train and Valid

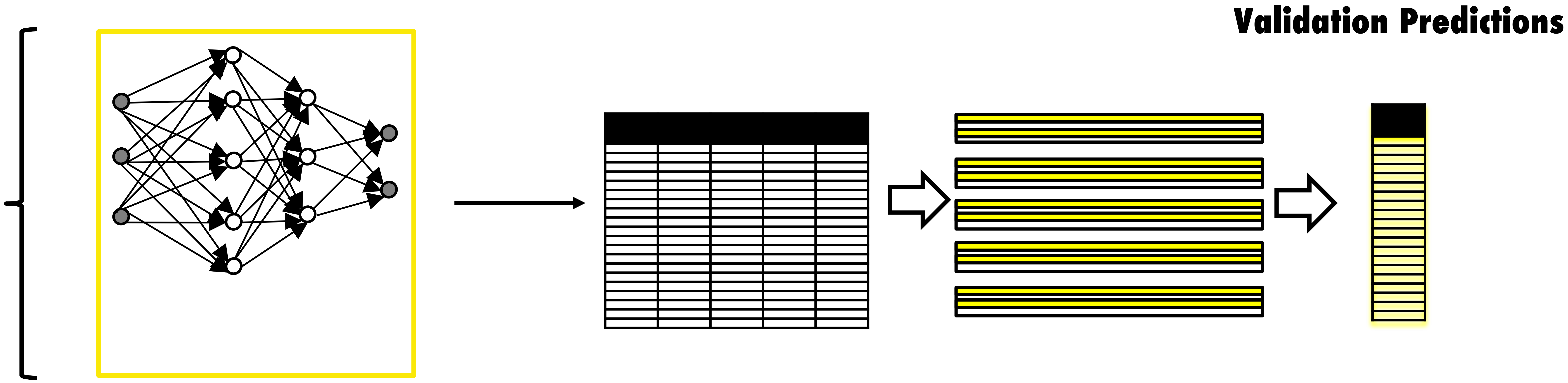


Behind the Scenes

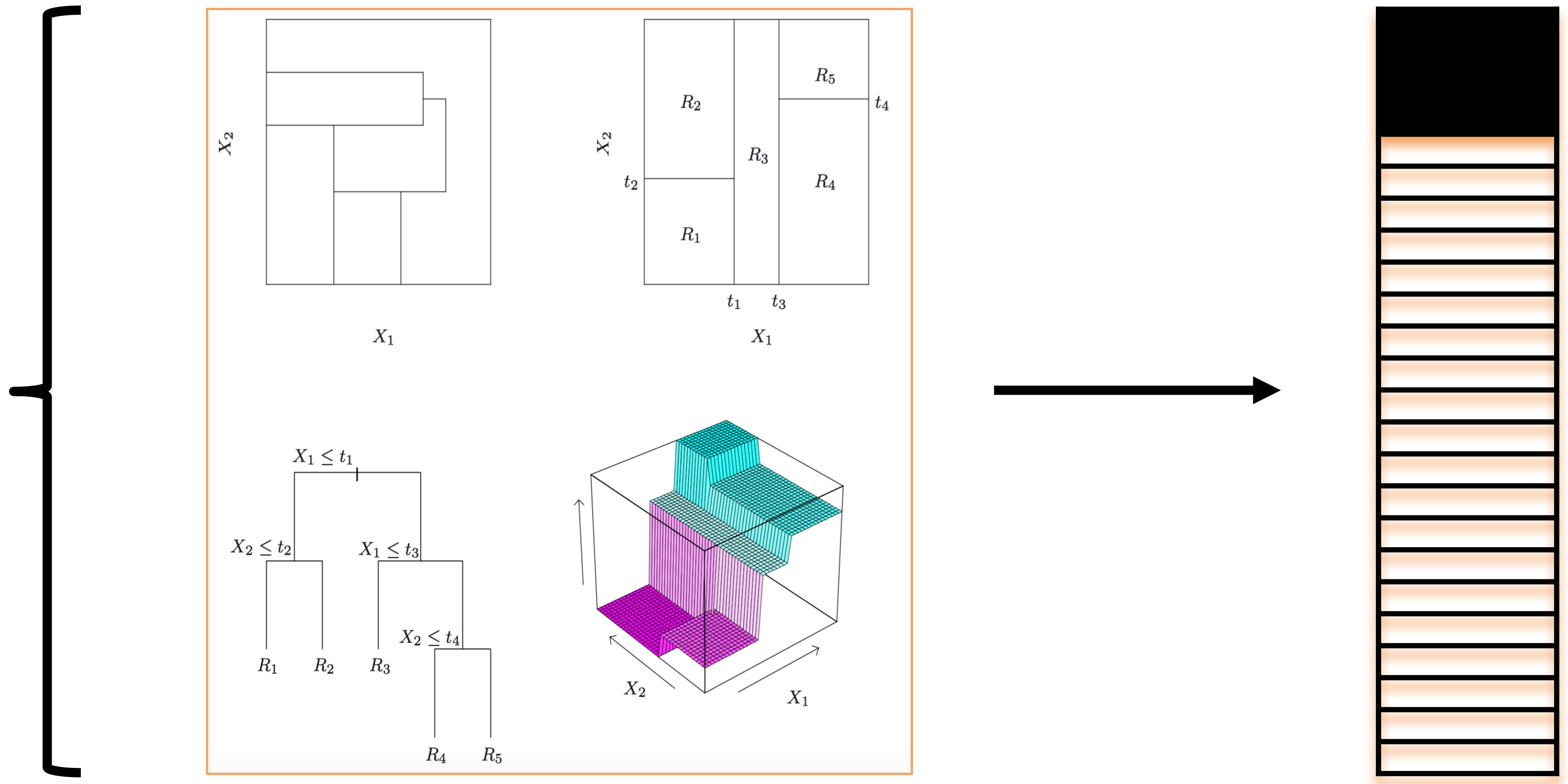
Split into Train and Valid per Fold



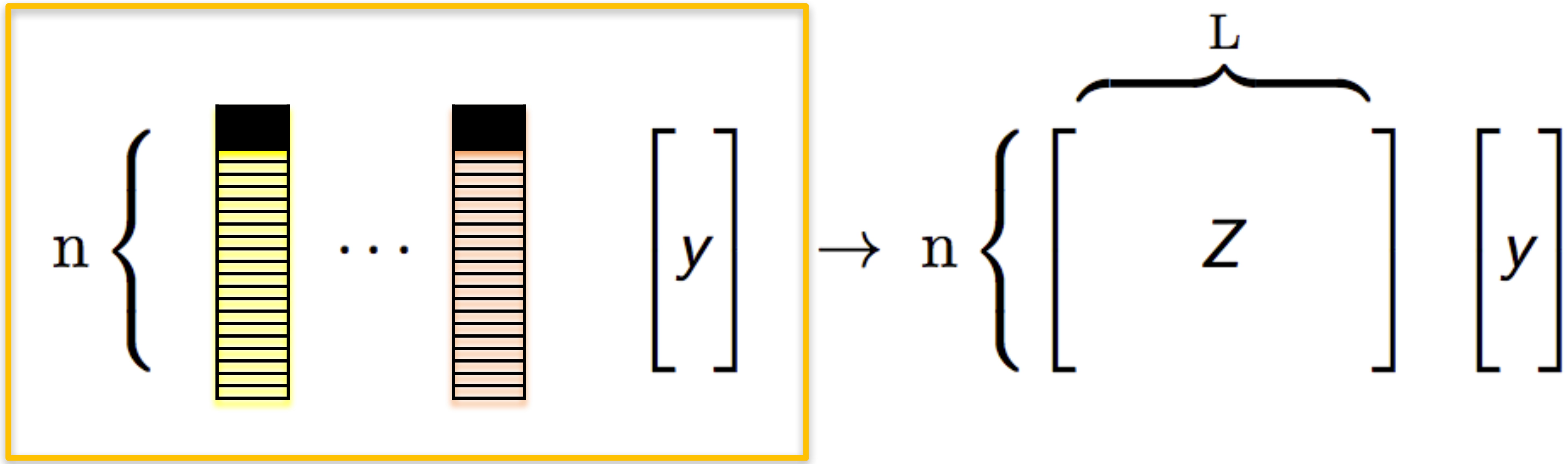
Form Prediction Column



Prediction Results Column



Stacking: Level-One Data



- Collect the predicted values from *k*-fold CV that was performed on each of the L base learners

Stacking: Level-One Data

$$\boxed{n\left\{\begin{bmatrix} \\ p_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ p_L \\ \\ \end{bmatrix} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}\right.} \rightarrow n\left\{\begin{bmatrix} \\ \overbrace{\qquad Z \qquad}^{L} \\ \\ \end{bmatrix} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}\right.$$

- Collect the predicted values from *k*-fold CV that was performed on each of the L base learners

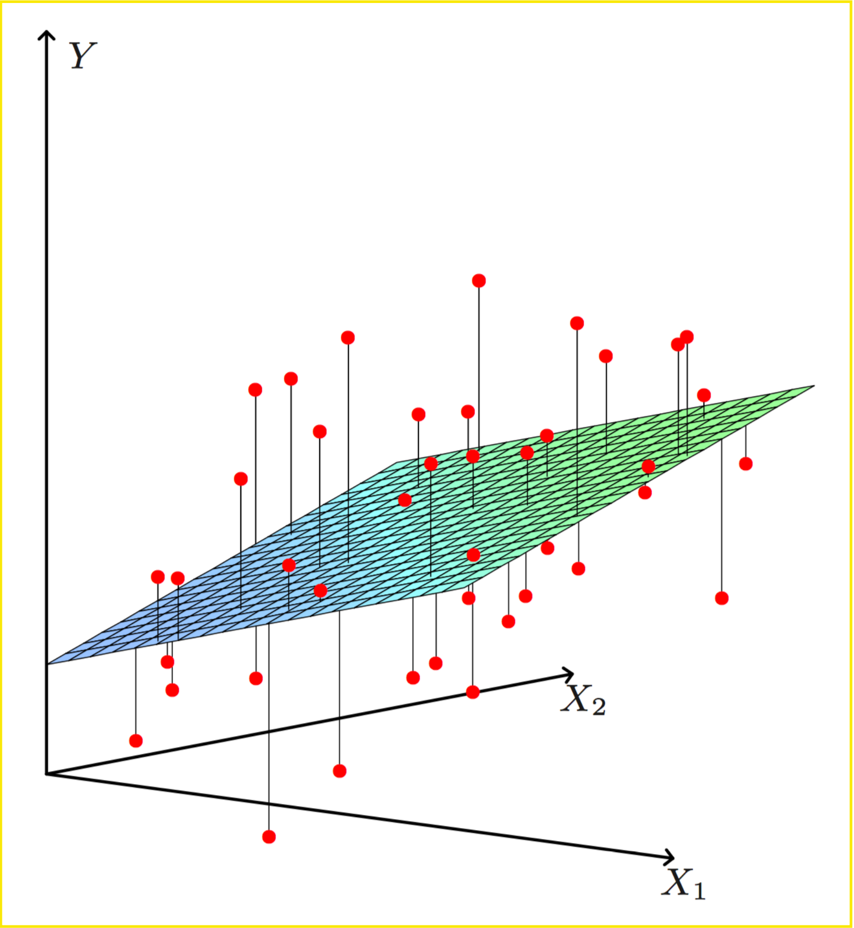
Stacking: Level-One Data

$$n\left\{\begin{bmatrix} \\ p_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ p_L \\ \\ \end{bmatrix} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}\right. \rightarrow \boxed{n\left\{\begin{bmatrix} \overbrace{\qquad\qquad}^{L} \\ Z \\ \\ \end{bmatrix}\right.} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}$$

- Collect the predicted values from *k*-fold CV that was performed on each of the L base learners
- Column-bind ("stack") these prediction vectors together to form a new design matrix, Z

Stacking: Level-One Data

$$n\left\{\begin{bmatrix} p_1 \end{bmatrix} \cdots \begin{bmatrix} p_L \end{bmatrix} \begin{bmatrix} y \end{bmatrix}\right. \rightarrow n\left\{\overbrace{\begin{bmatrix} \quad\quad Z \quad\quad \end{bmatrix}}^{L} \begin{bmatrix} y \end{bmatrix}\right.$$



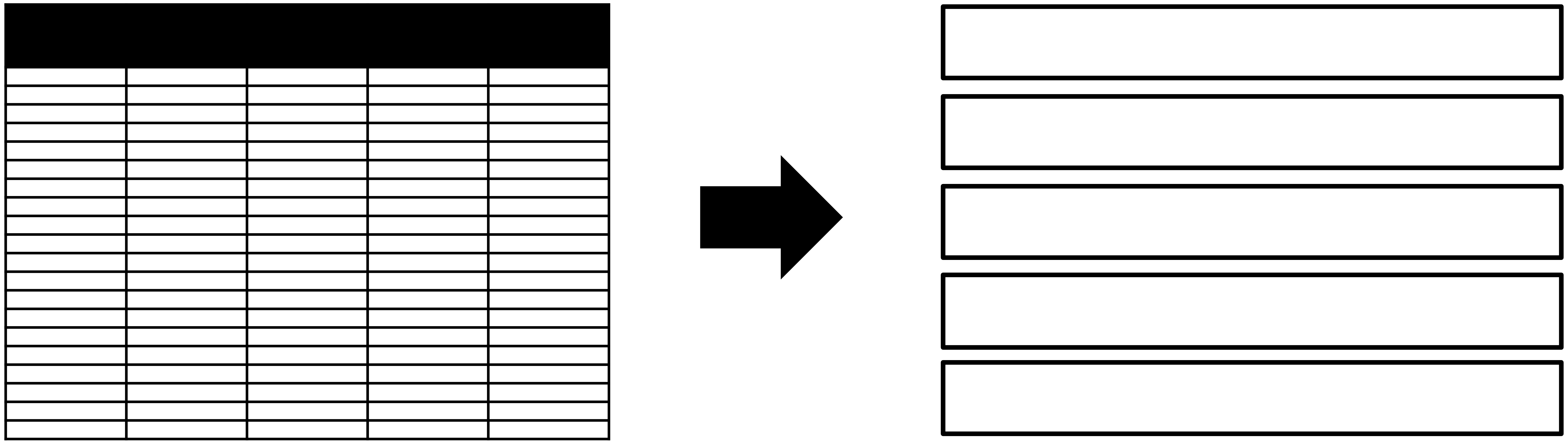
- Collect the predicted values from *k*-fold CV that was performed on each of the L base learners
- Column-bind ("stack") these prediction vectors together to form a new design matrix, Z
- Train the metalearner (currently a GLM) using Z, y

Appendix

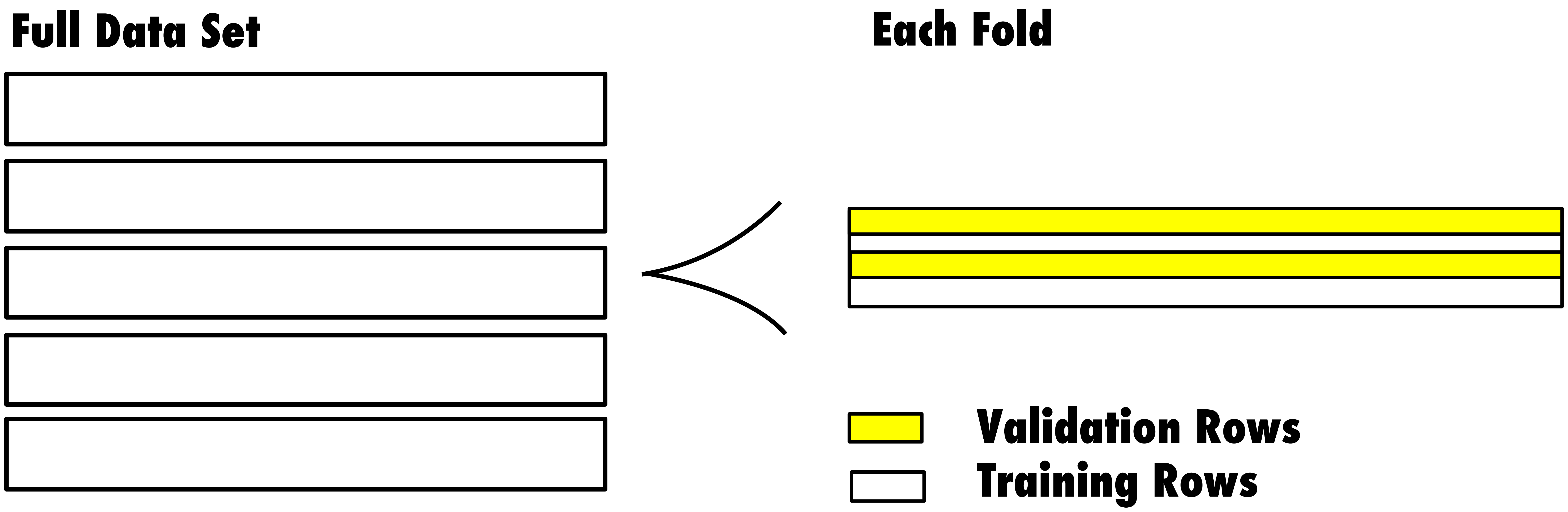
5-fold Cross Validation

Full Data Set

Split into 5 Folds

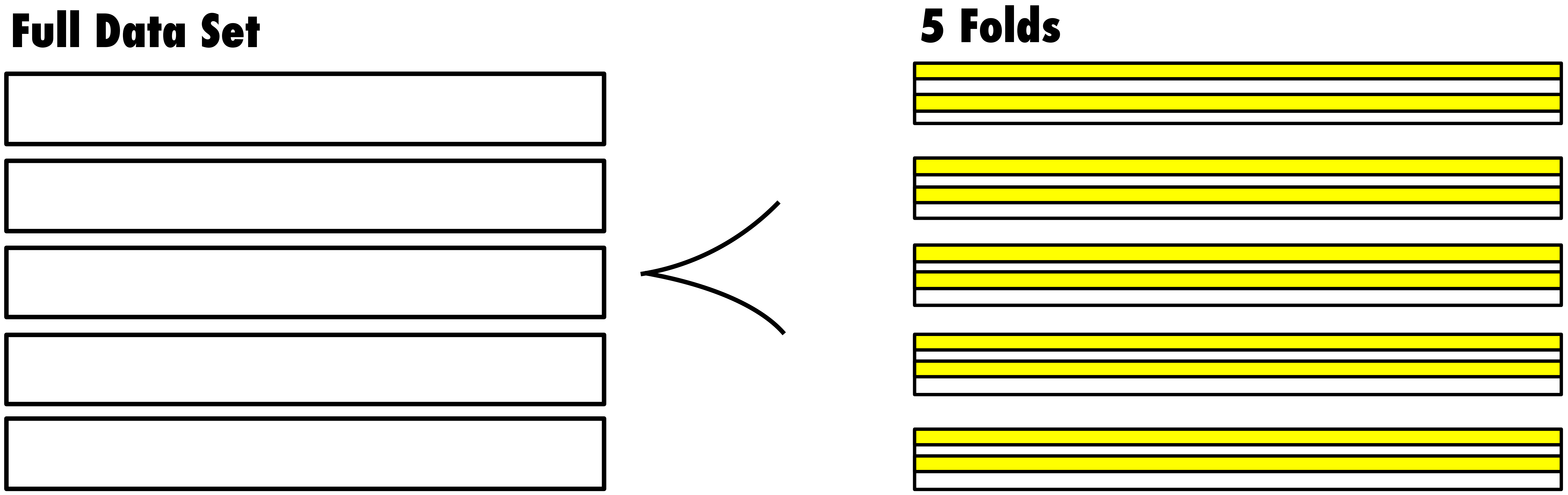


5-fold Cross Validation

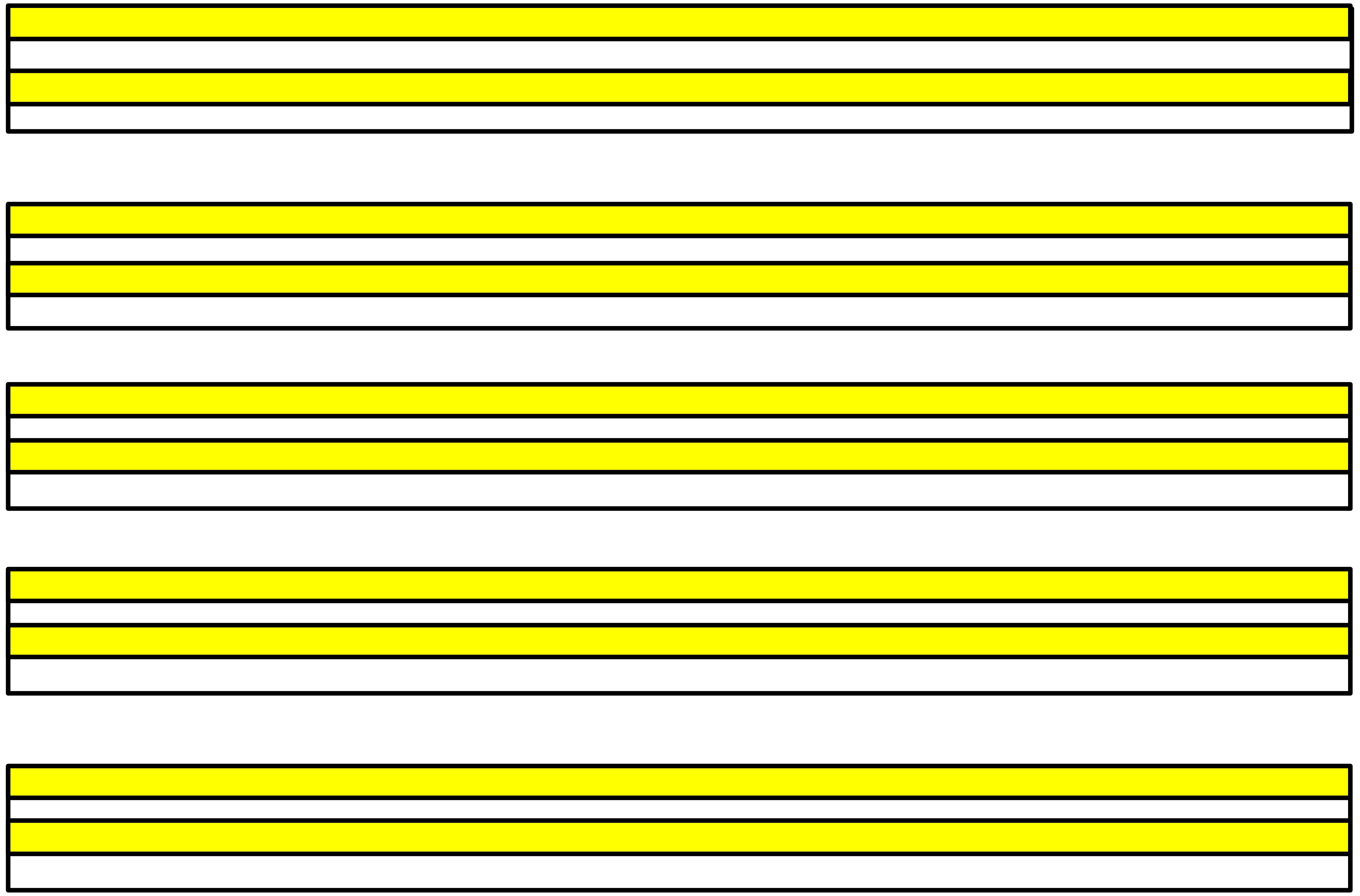


H2O.ai

5-fold Cross Validation



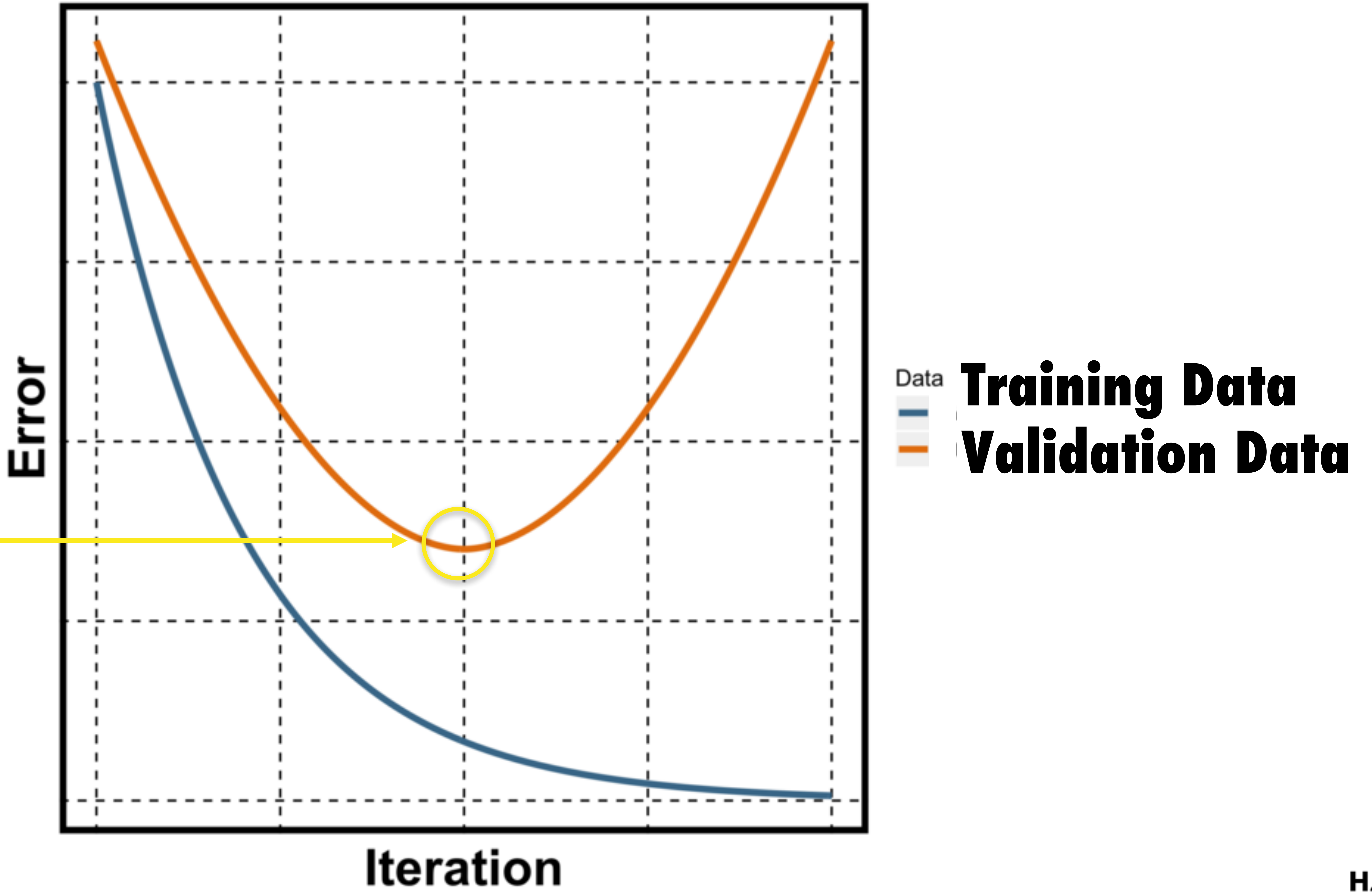
5-fold Cross Validation



Early stopping happens within each fold

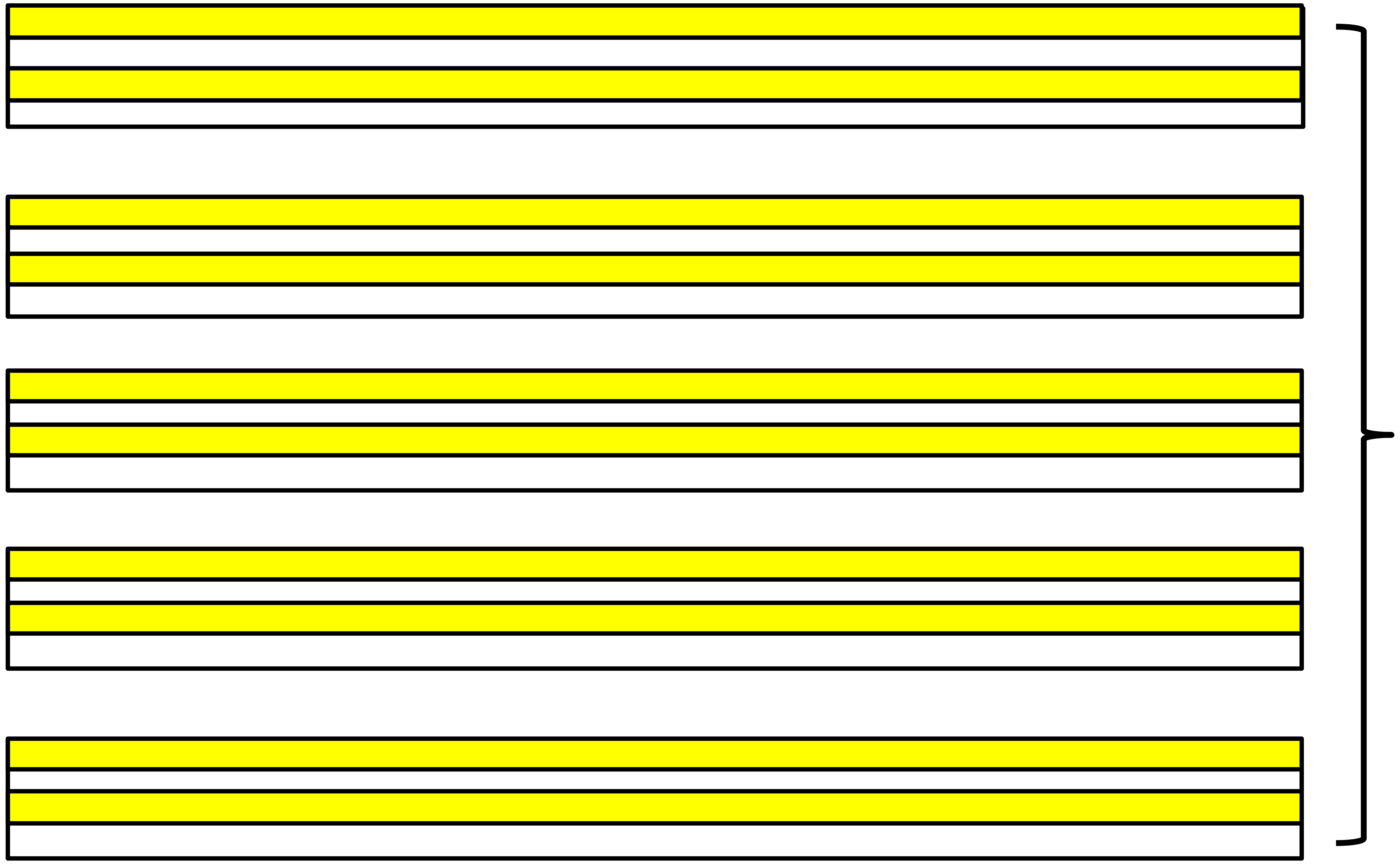
Early Stopping

The Sweet Spot



H2O.ai

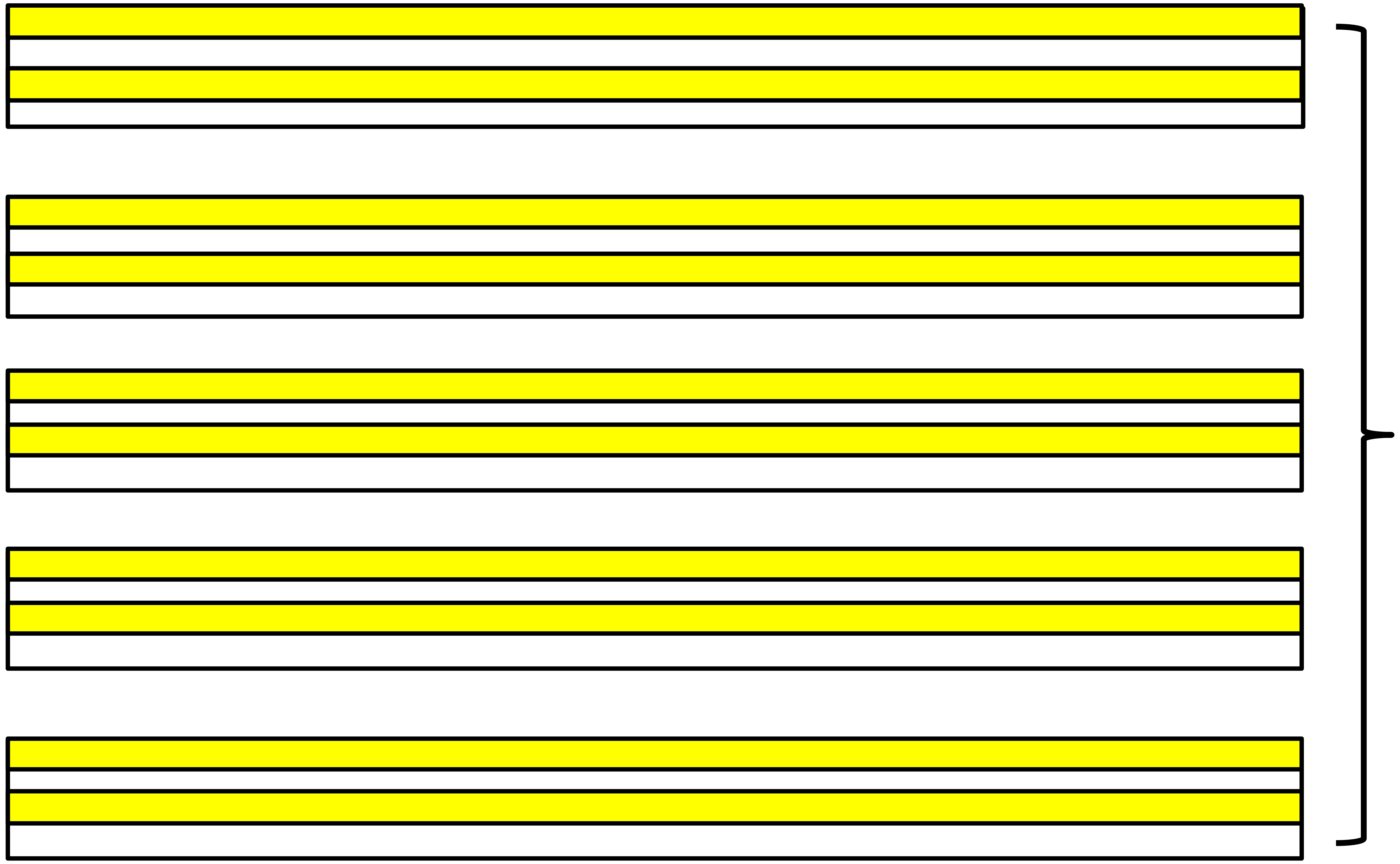
5-fold Cross Validation



GBM

Average number of trees are used to train on 100% of the training data

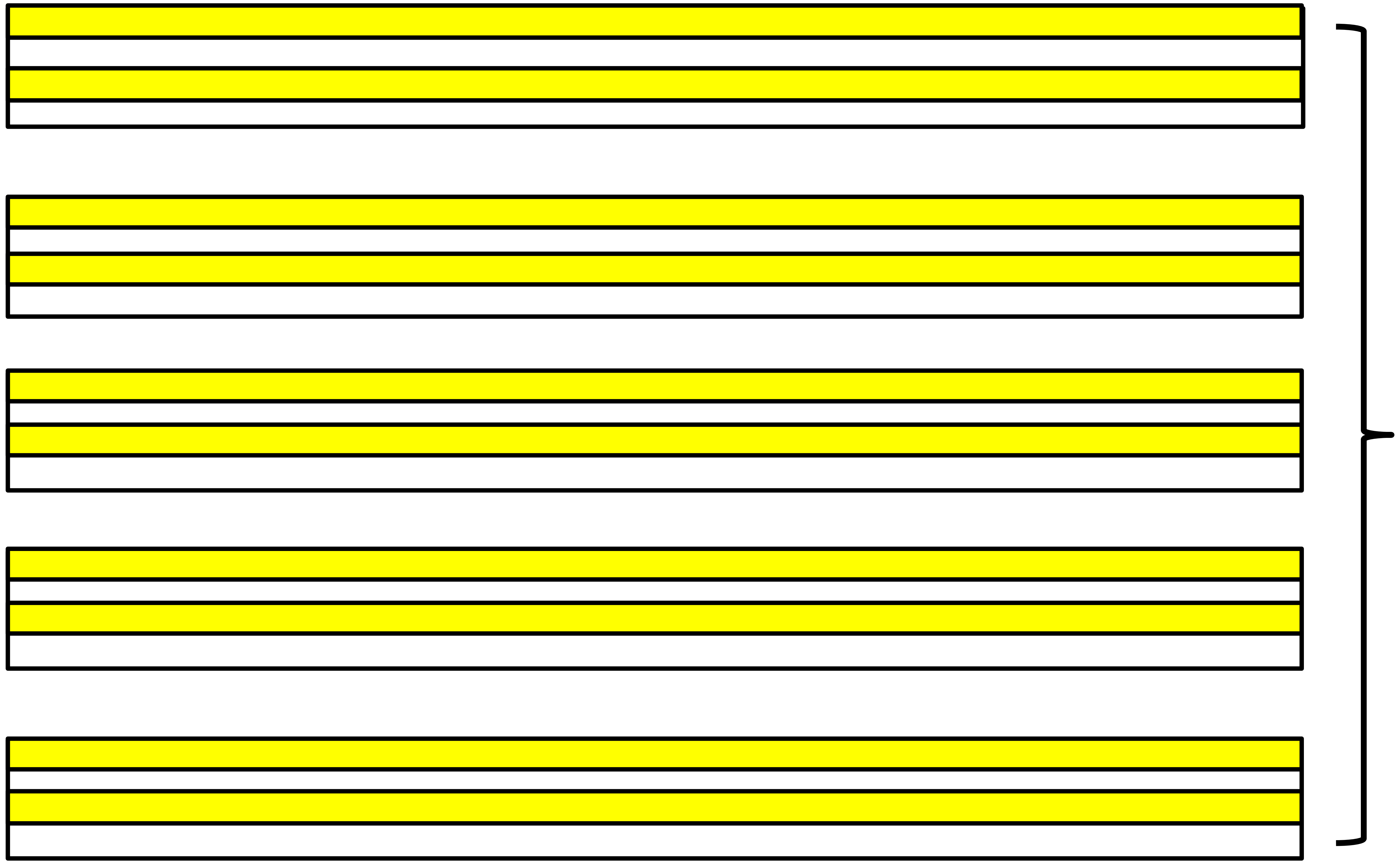
5-fold Cross Validation



DL

Average number of epochs are used to train on 100% of the training data

5-fold Cross Validation

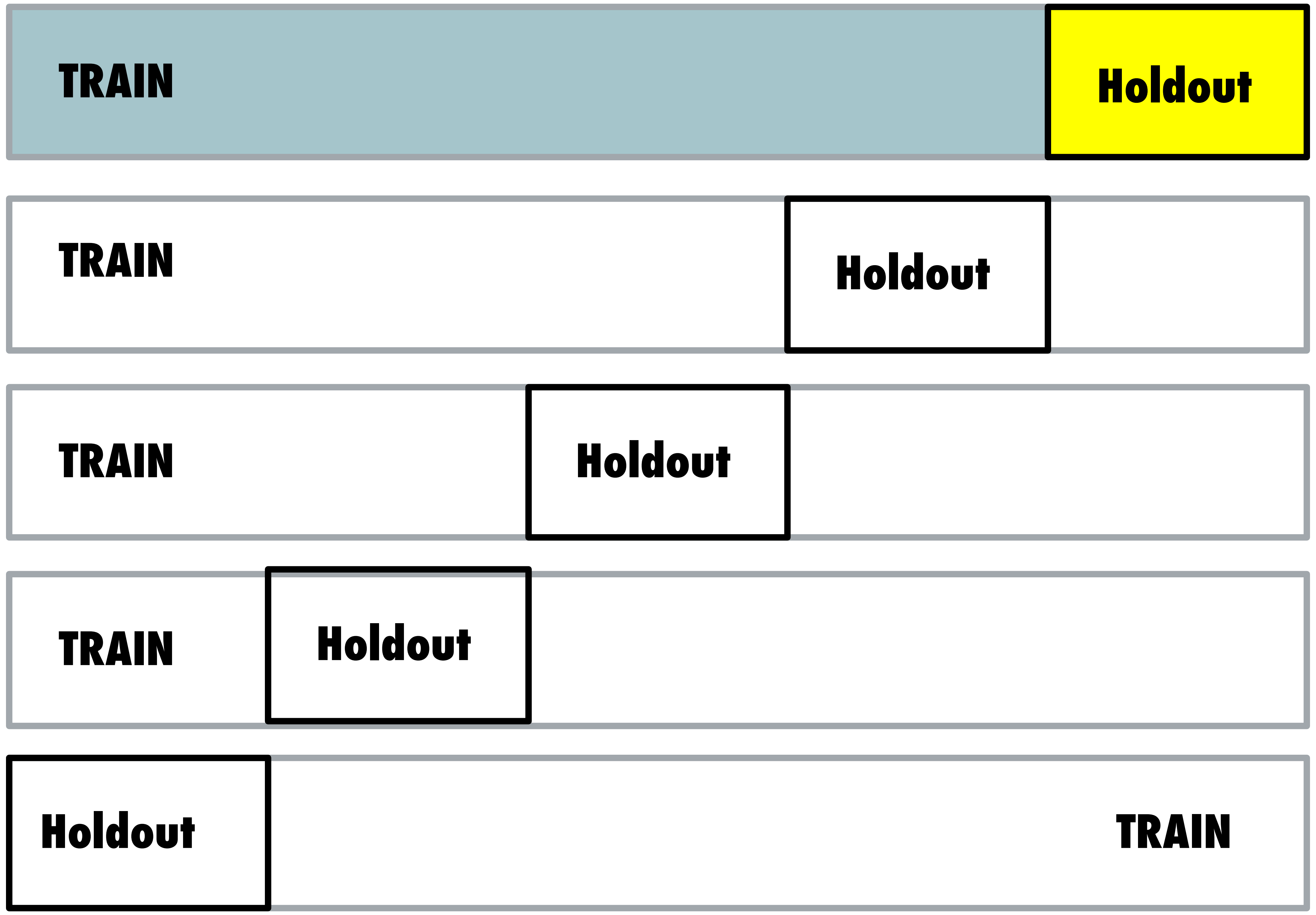


GLM

Best Lambda from all folds is used to train on 100% of the training data

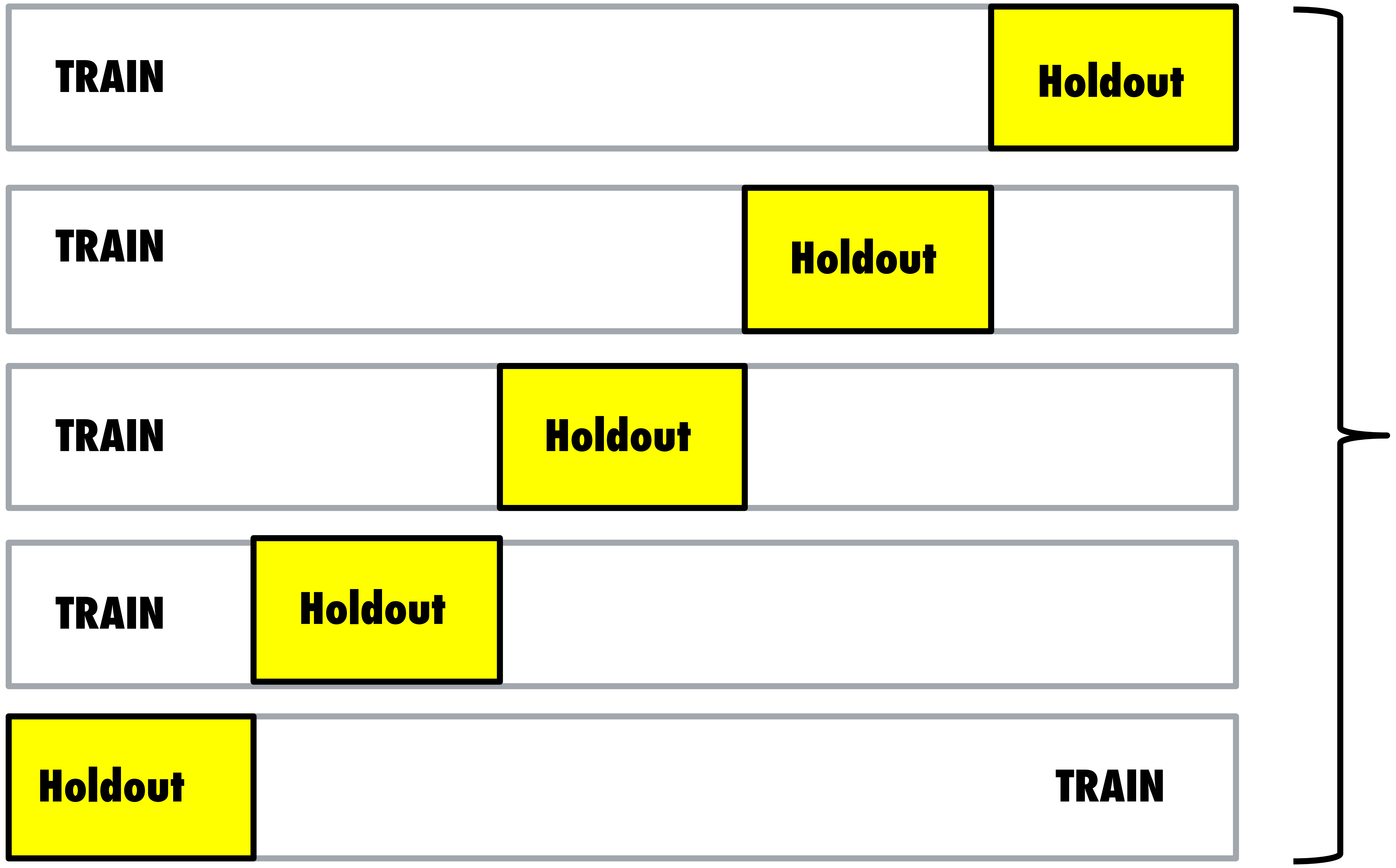
5-fold Cross Validation

Each Fold Uses its Holdout for Early Stopping



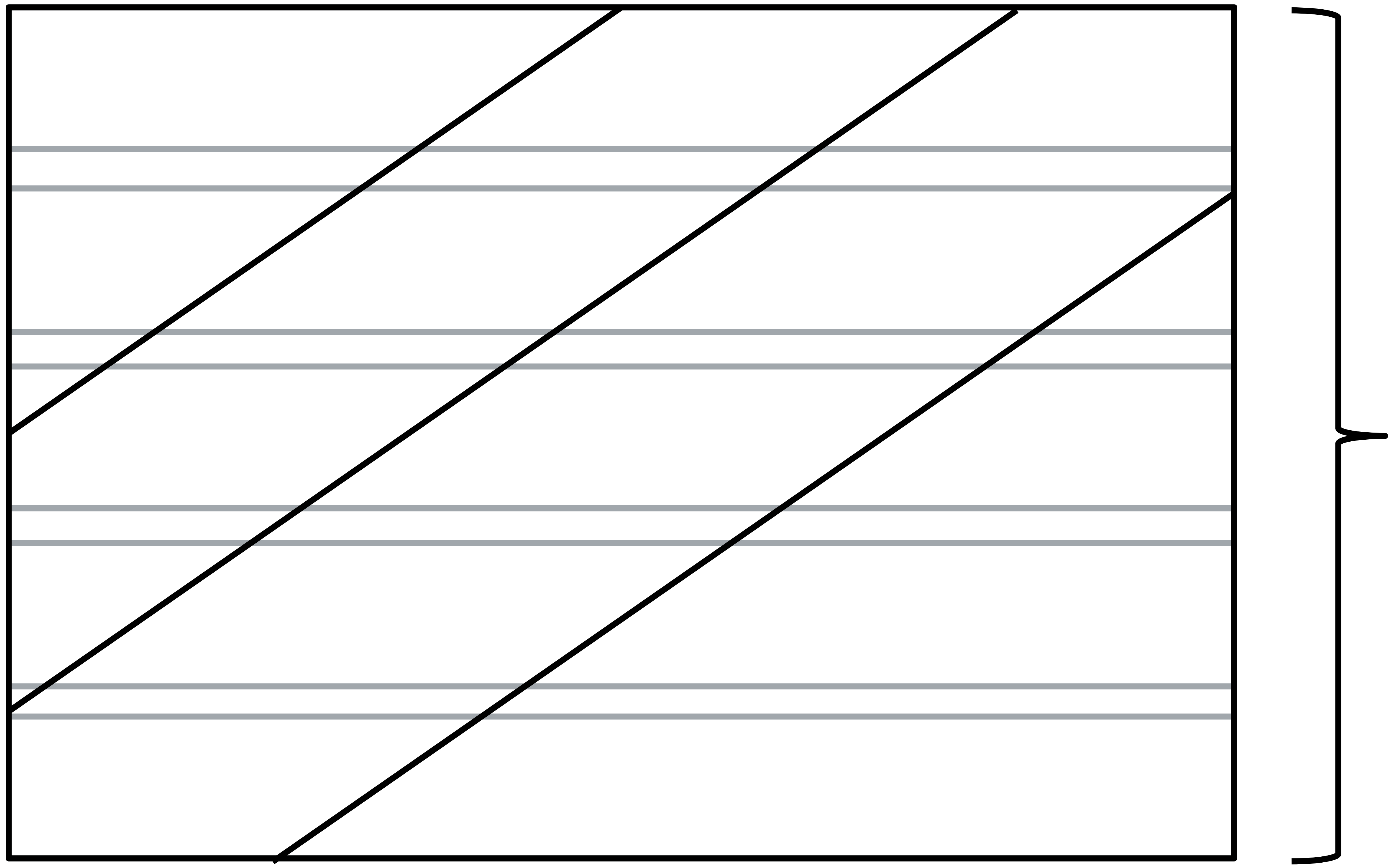
H_2O.ai

5-fold Cross Validation



Average number of trees are used to train on 100% of the training data

5-fold Cross Validation



Average number of trees are used to train on 100% of the training data **– The Model You Get Back**

Auto-Splits

User provides: **Training Frame**

Train is Split: **70%** Train, **15%** Valid, **15%** Leaderboard

Auto-Splits

User provides: **Training** & **Validation Frames**

Valid is Split: **50%** Valid, **50%** Leaderboard

Auto Splits

User provides: **Training**, **Validation** & **Leaderboard Frames**

Data is Left as is

H2O.ai