Introduction to machine learning in Hydrology

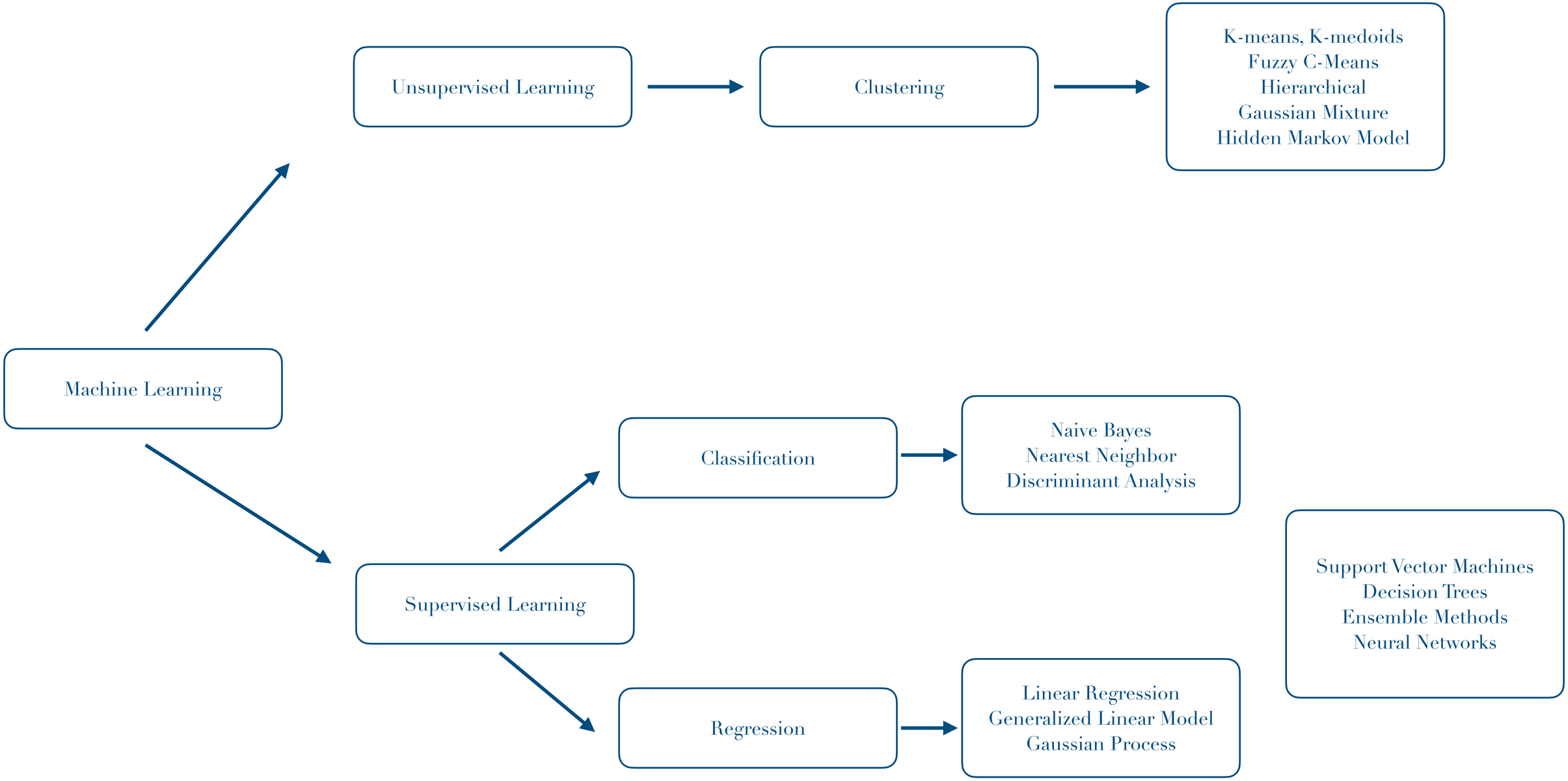
Lazaro J. Perez & Marc Berghouse

contact: lazaro.perez@dri.edu

Outline

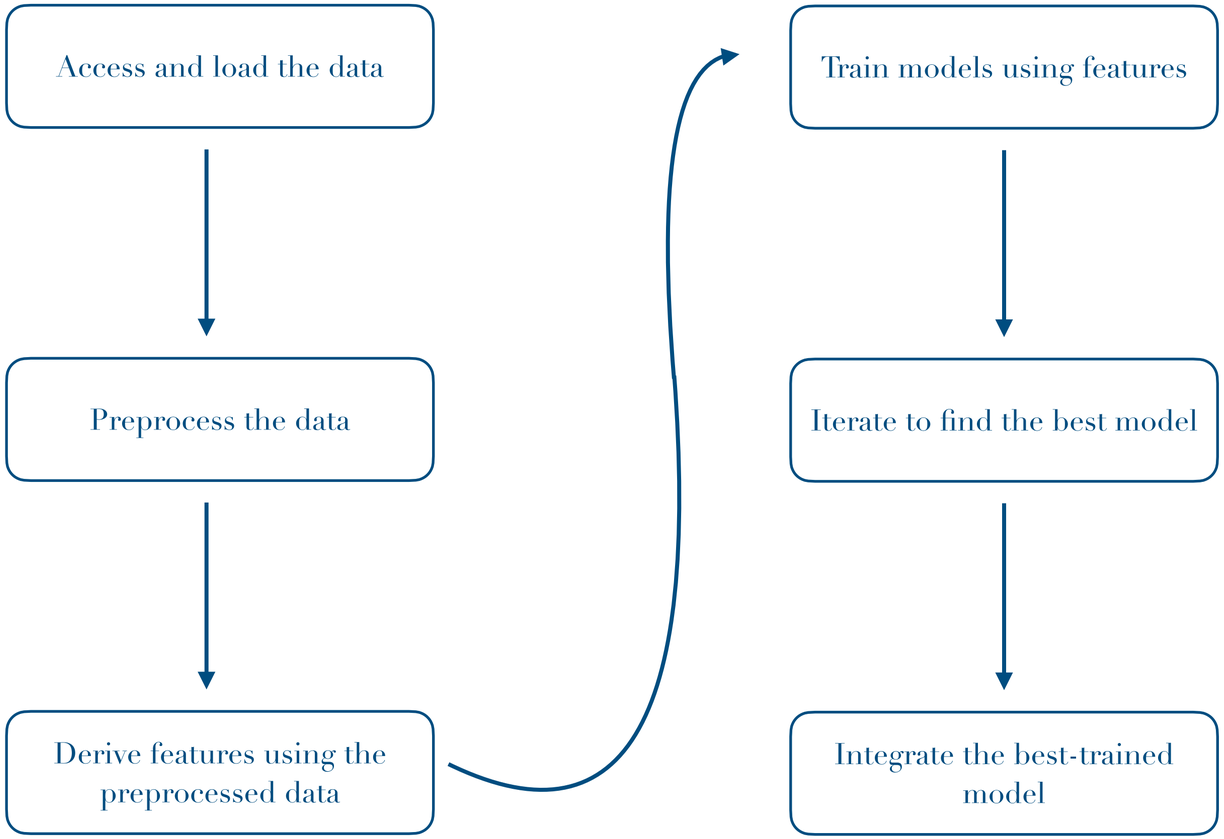
- Machine Learning in Matlab
- Classification Learner App
- Evaluating Classification Models
- Regression Learner App
- Evaluating Regression Models

Machine learning in Matlab

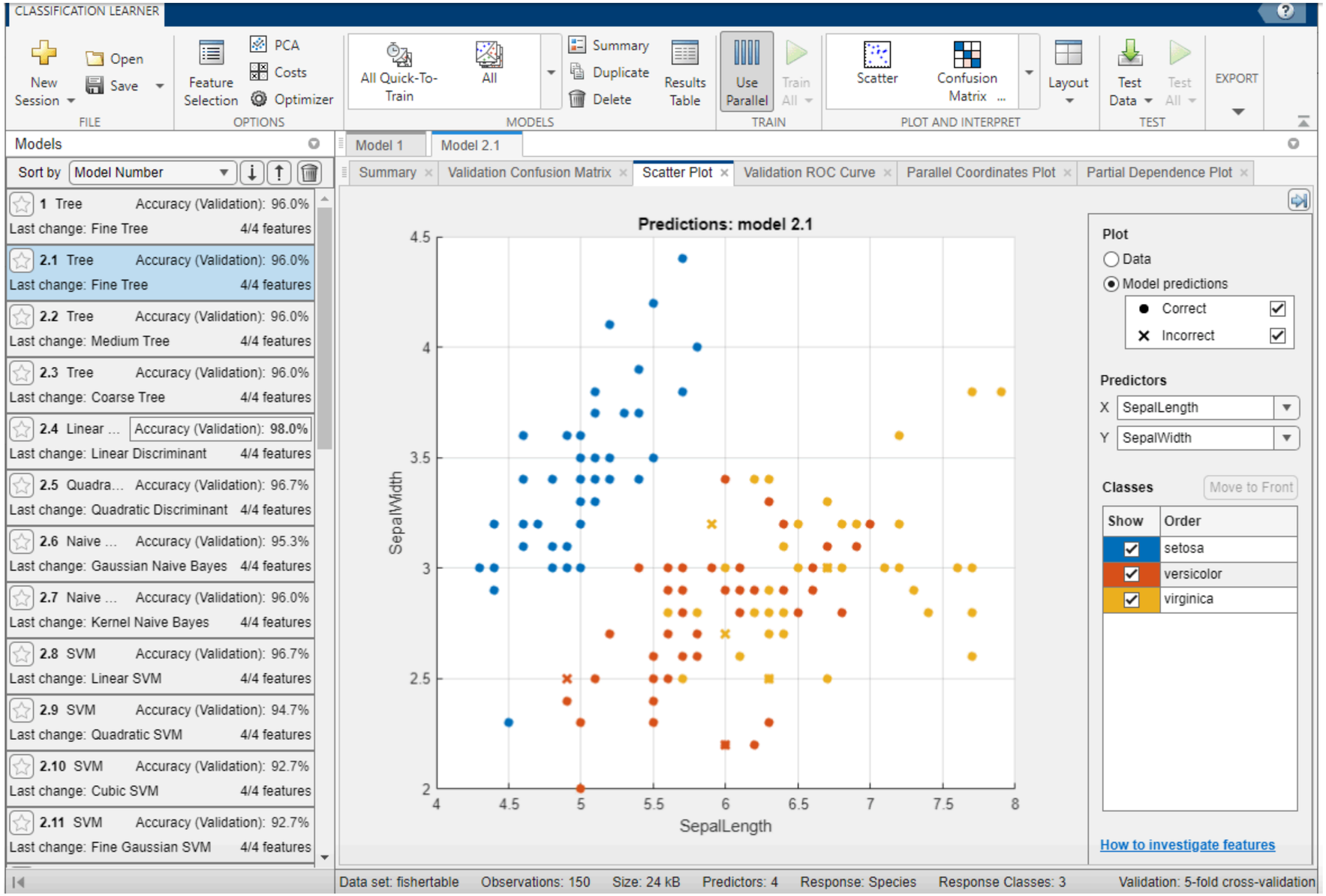


Machine learning in Matlab

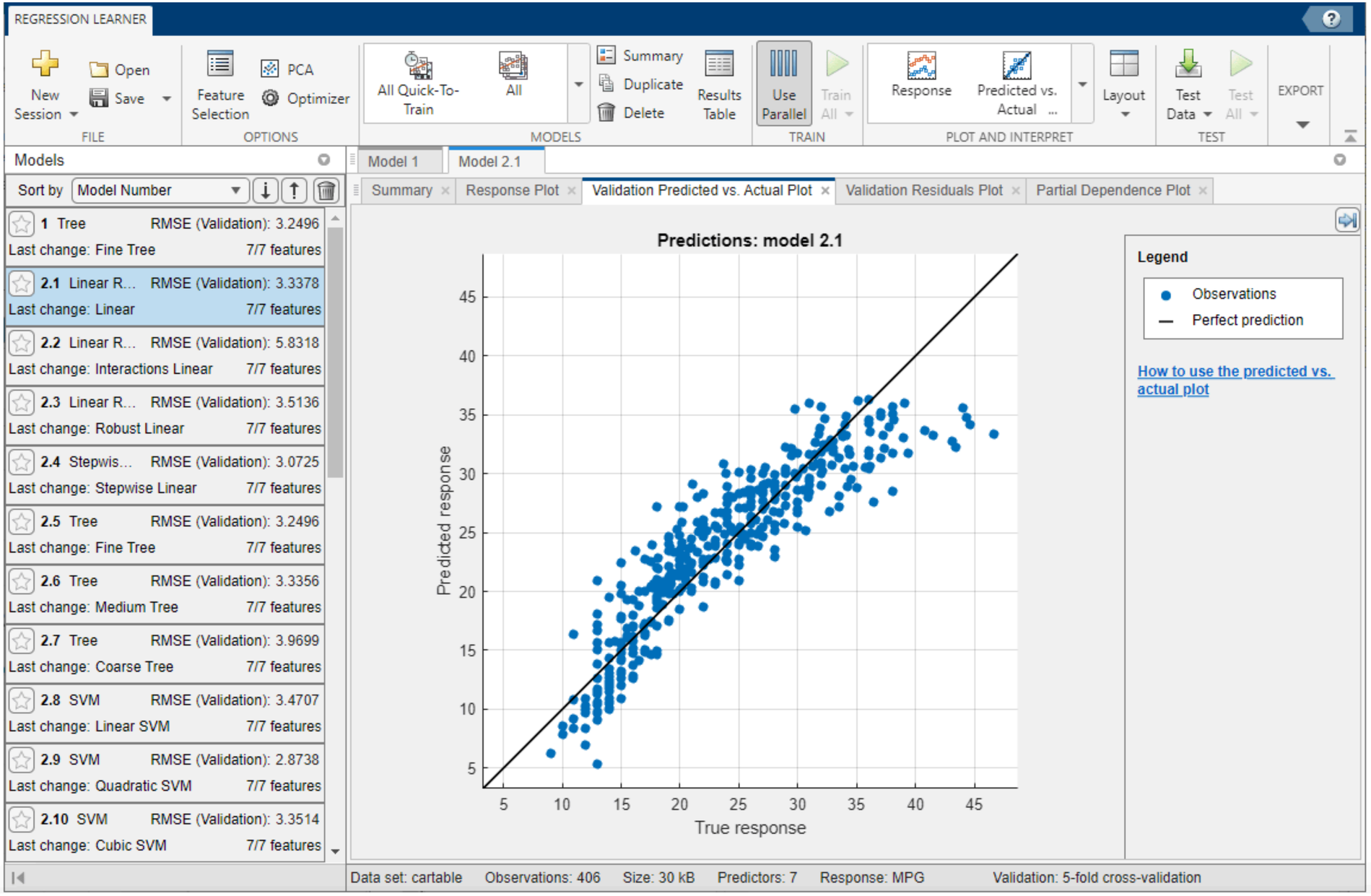
General workflow



Classification Learner



Regression Learner



High-elevation mountain regions

ARTICLE

https://doi.org/10.1038/s43247-020-00039-w OPEN

Check for updates

Significant stream chemistry response to temperature variations in a high-elevation mountain watershed

Wei Zhi[1], Kenneth H. Williams[2,3], Rosemary W. H. Carroll[2,4], Wendy Brown[2], Wenming Dong[3], Devon Kerins[1] & Li Li[1✉]

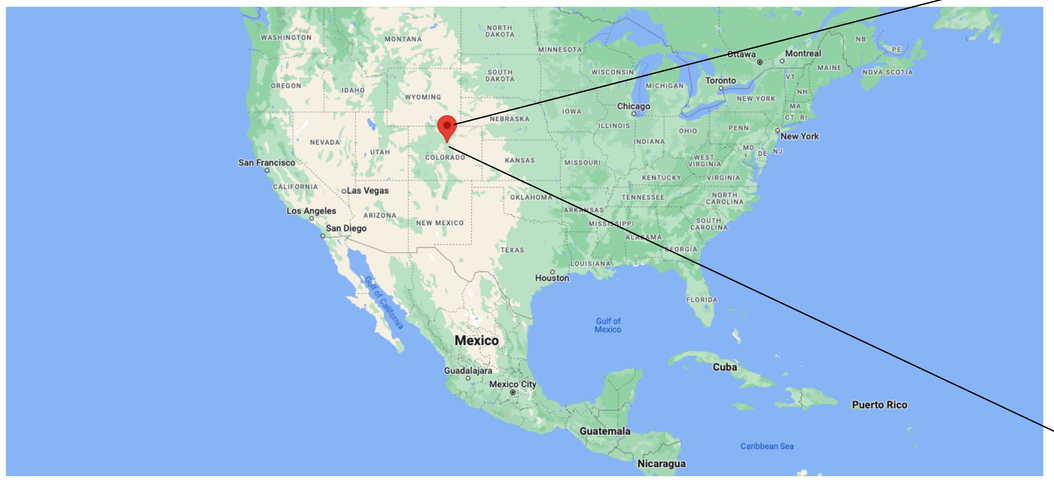
High-elevation mountain regions, central to global freshwater supply, are experiencing more rapid warming than low-elevation locations. High-elevation streams are therefore potentially critical indicators for earth system and water chemistry response to warming. Here we present concerted hydroclimatic and biogeochemical data from Coal Creek, Colorado in the central Rocky Mountains at elevations of 2700 to 3700 m, where air temperatures have increased by about 2 °C since 1980. We analyzed water chemistry every other day from 2016 to 2019. Water chemistry data indicate distinct responses of different solutes to inter-annual hydroclimatic variations. Specifically, the concentrations of solutes from rock weathering are stable inter-annually. Solutes that are active in soils, including dissolved organic carbon, vary dramatically, with double to triple peak concentrations occurring during snowmelt and in warm years. We advocate for consistent and persistent monitoring of high elevation streams to record early glimpse of earth surface response to warming.

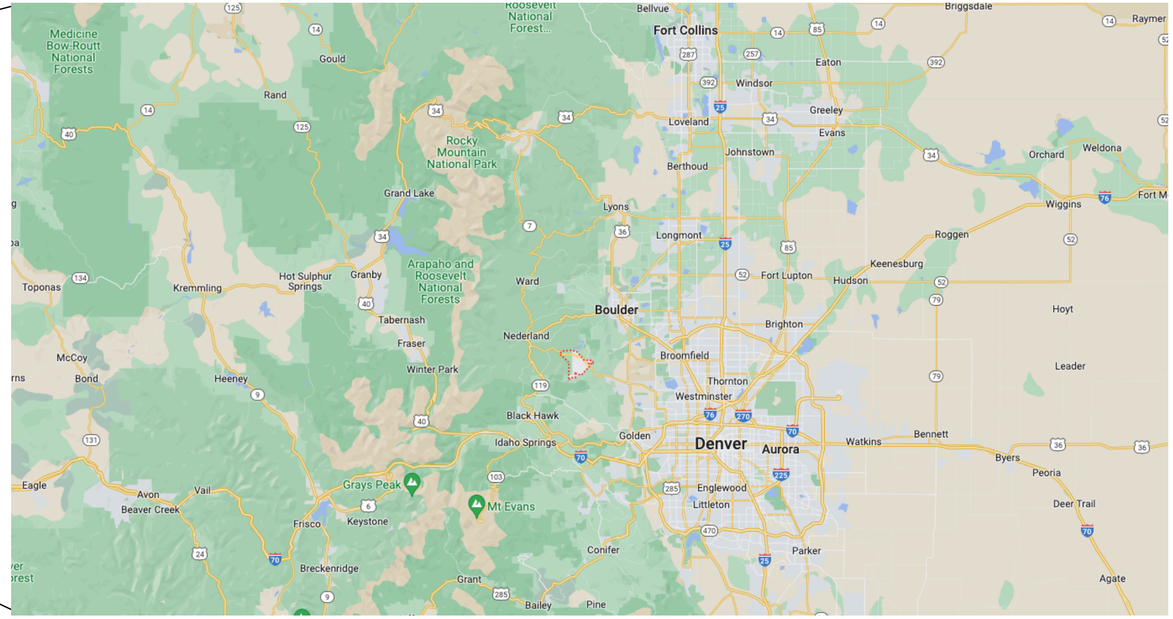
Hydrology application

High-elevation mountain regions

Coal Creek, CO: Rocky Mountains (2700-3700 m)

Temperatures have increased by about 2 °C since 1980





Hydrology application

High-elevation mountain regions

Coal Creek, CO: Rocky Mountains (2700-3700 m)

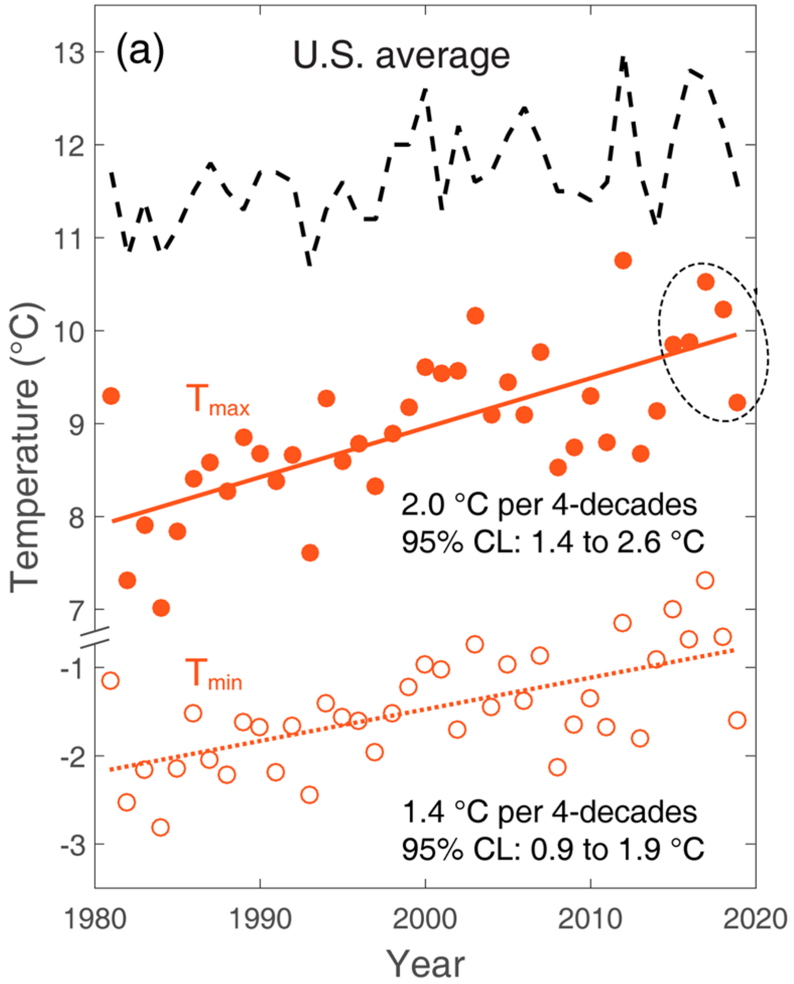
Temperatures have increased by about 2 °C since 1980

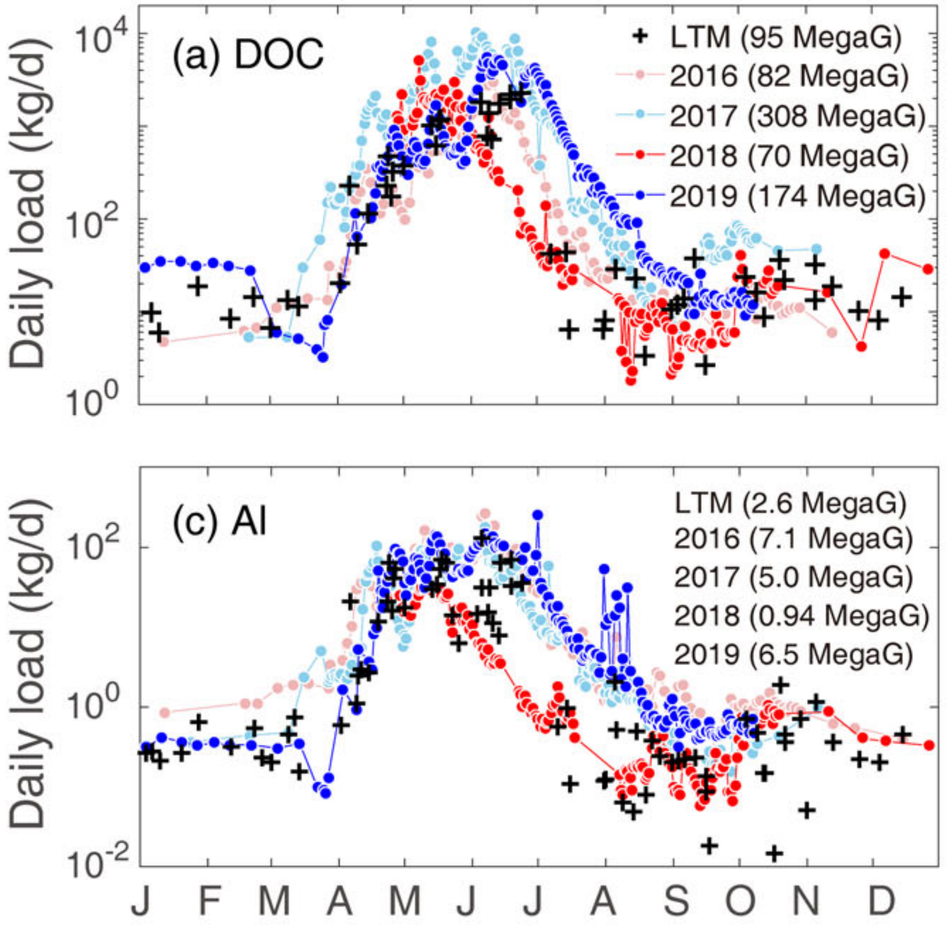
Research questions:

- Can water chemistry indicate warming?
- How does water chemistry respond to inter-annual hydro climatic variations?

Hydrology application

High-elevation mountain regions





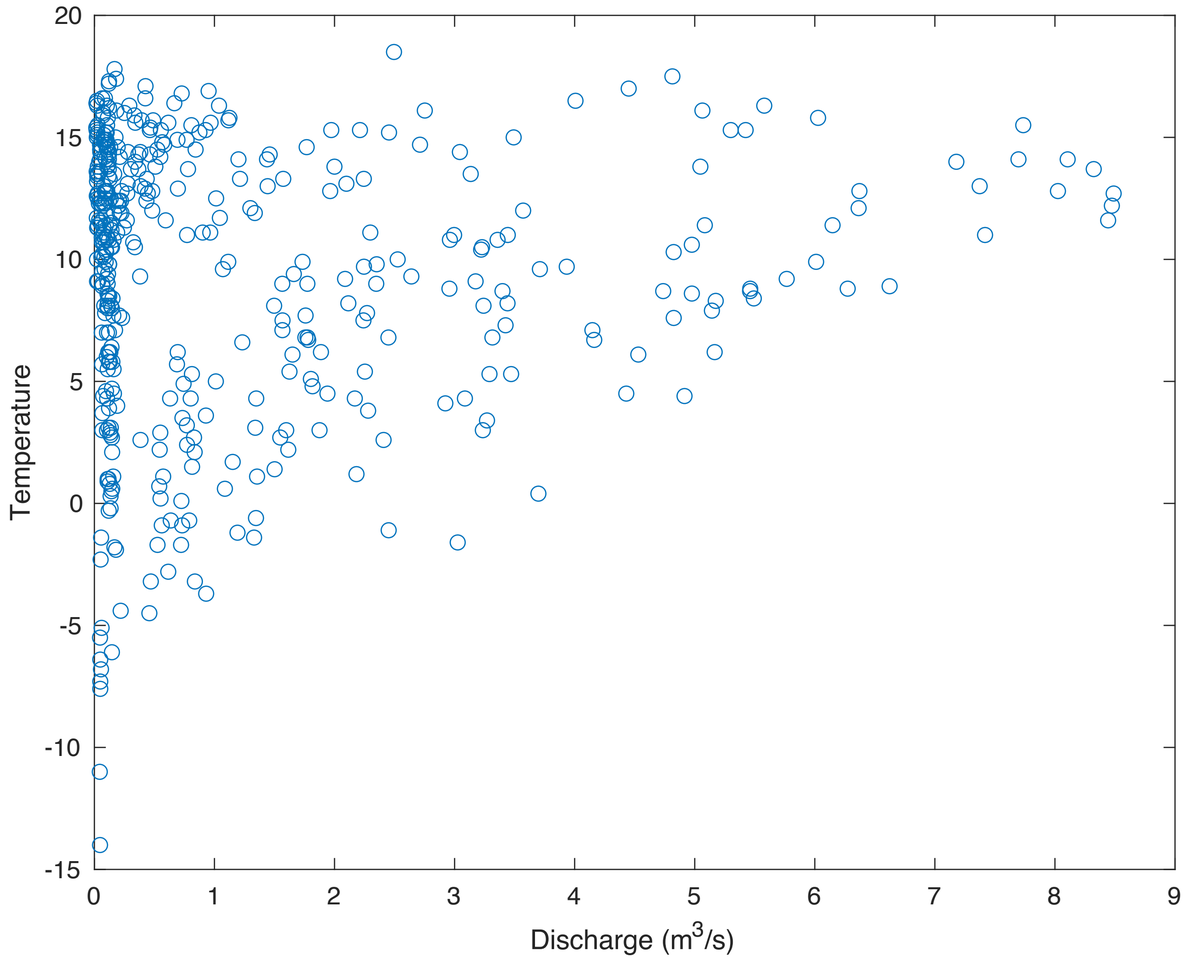
Quiz

Hypothesis

In a high-elevation mountain watershed, what is the relationship between temperature and discharge?

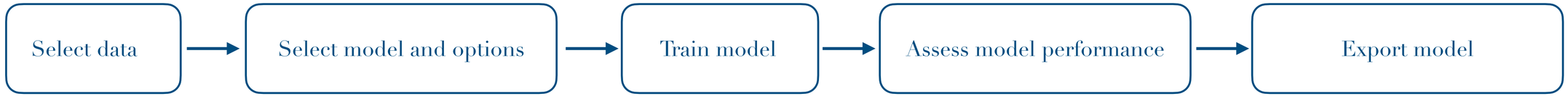
Hydrology application

High-elevation mountain regions



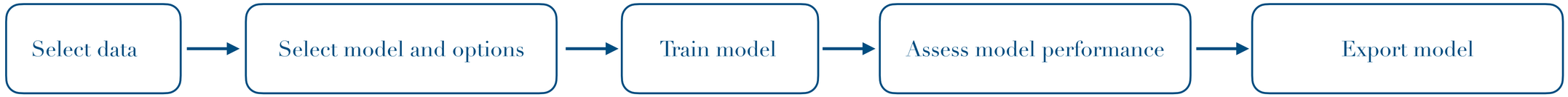
Classification Learner

General workflow

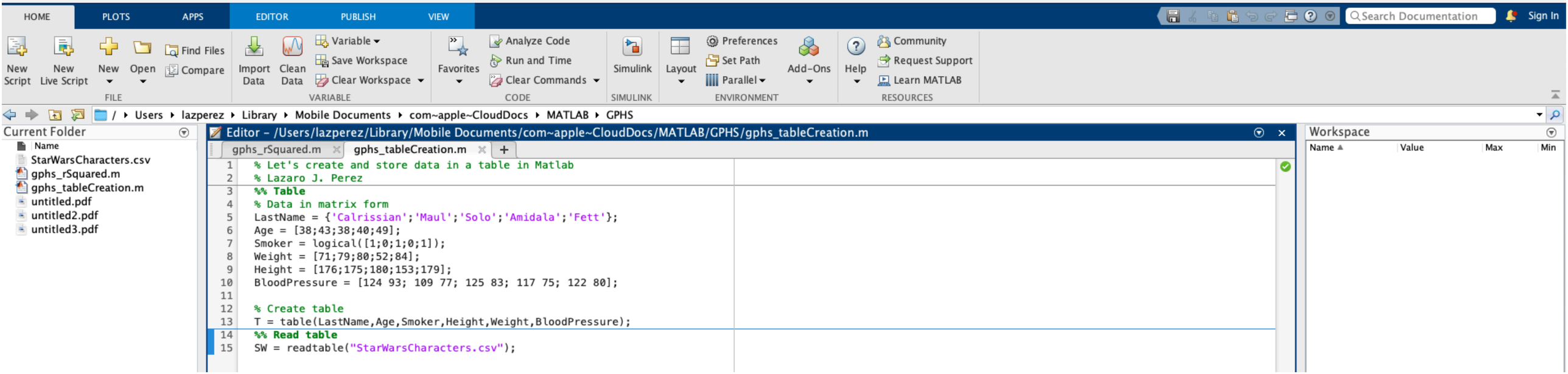


Classification Learner

General workflow

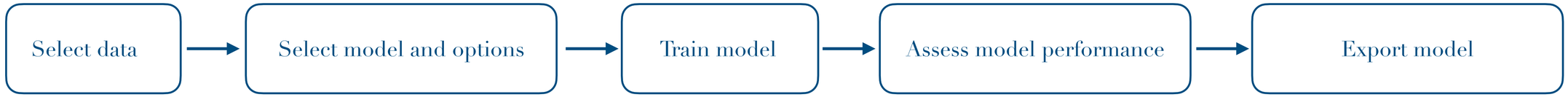


Select data

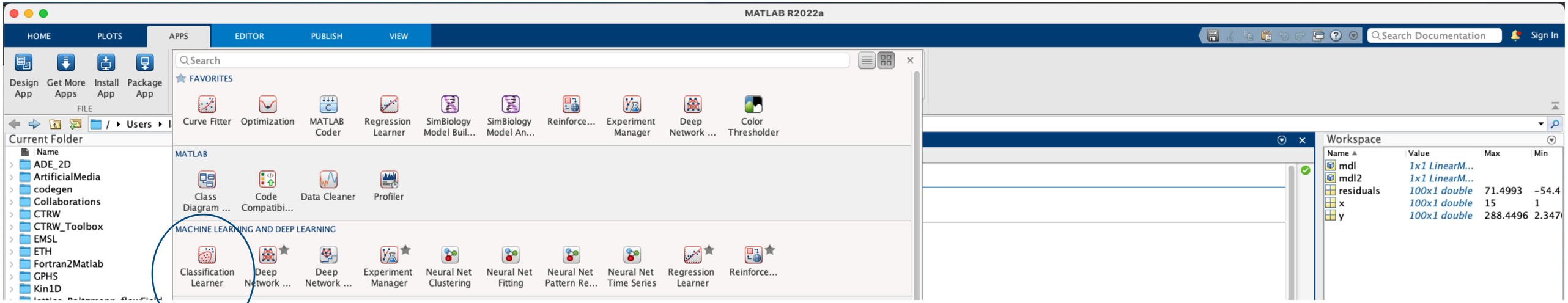


Classification Learner

General workflow



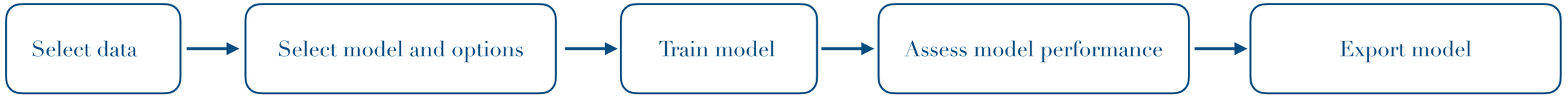
Select data



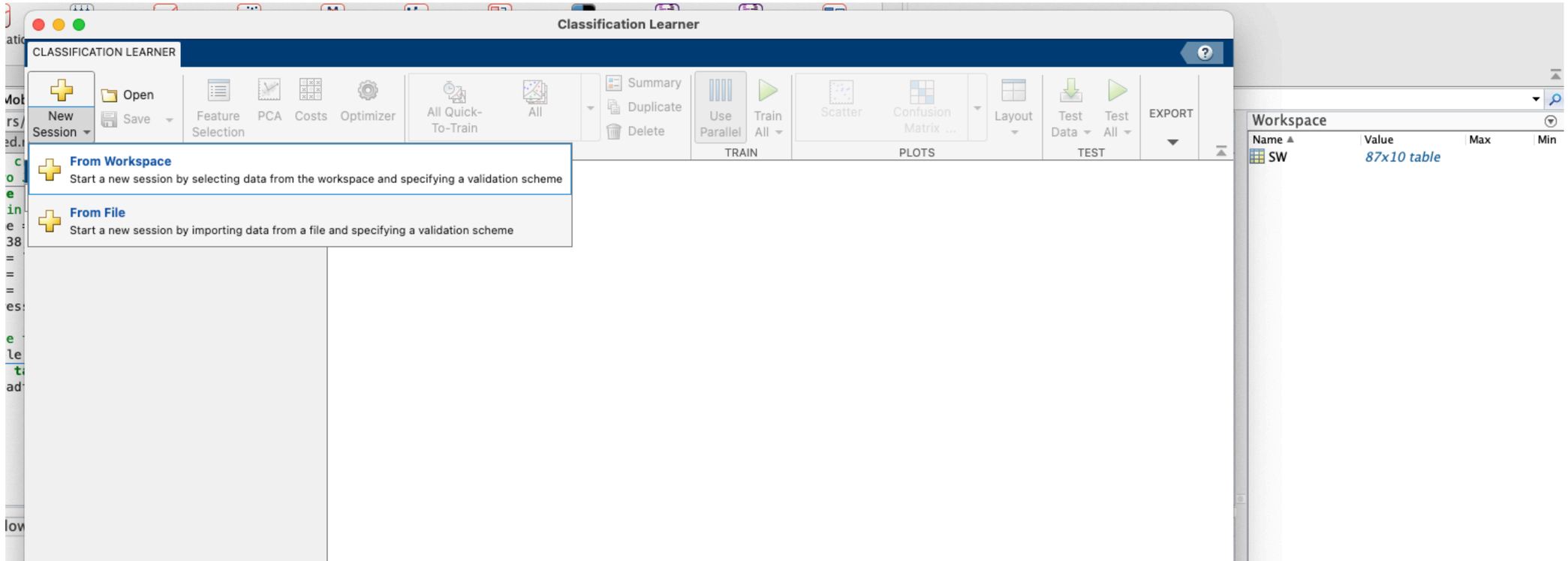
Classification learner

Classification Learner

General workflow



Select data

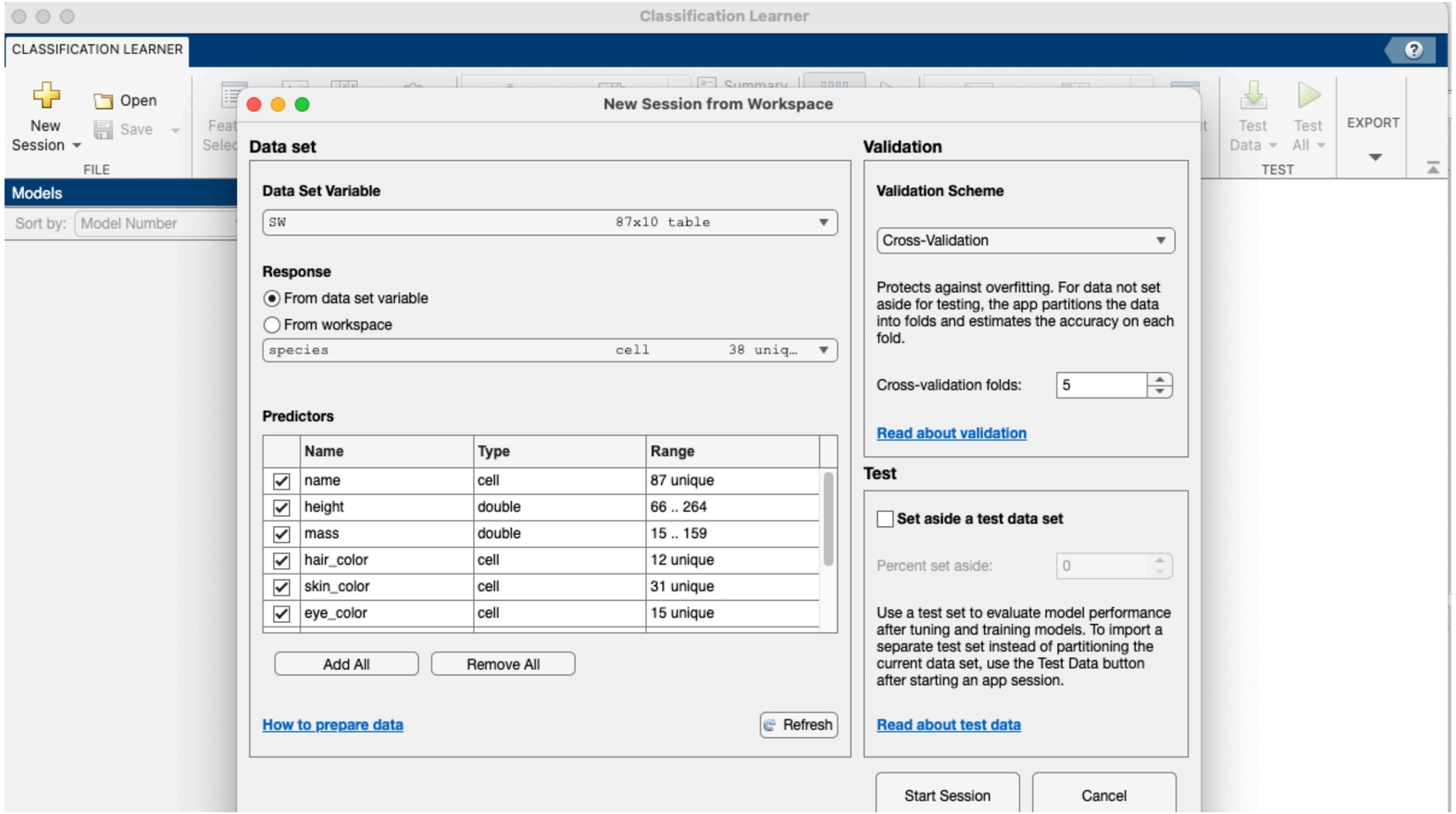


Classification Learner

General workflow

Select data → Select model and options → Train model → Assess model performance → Export model

Select data

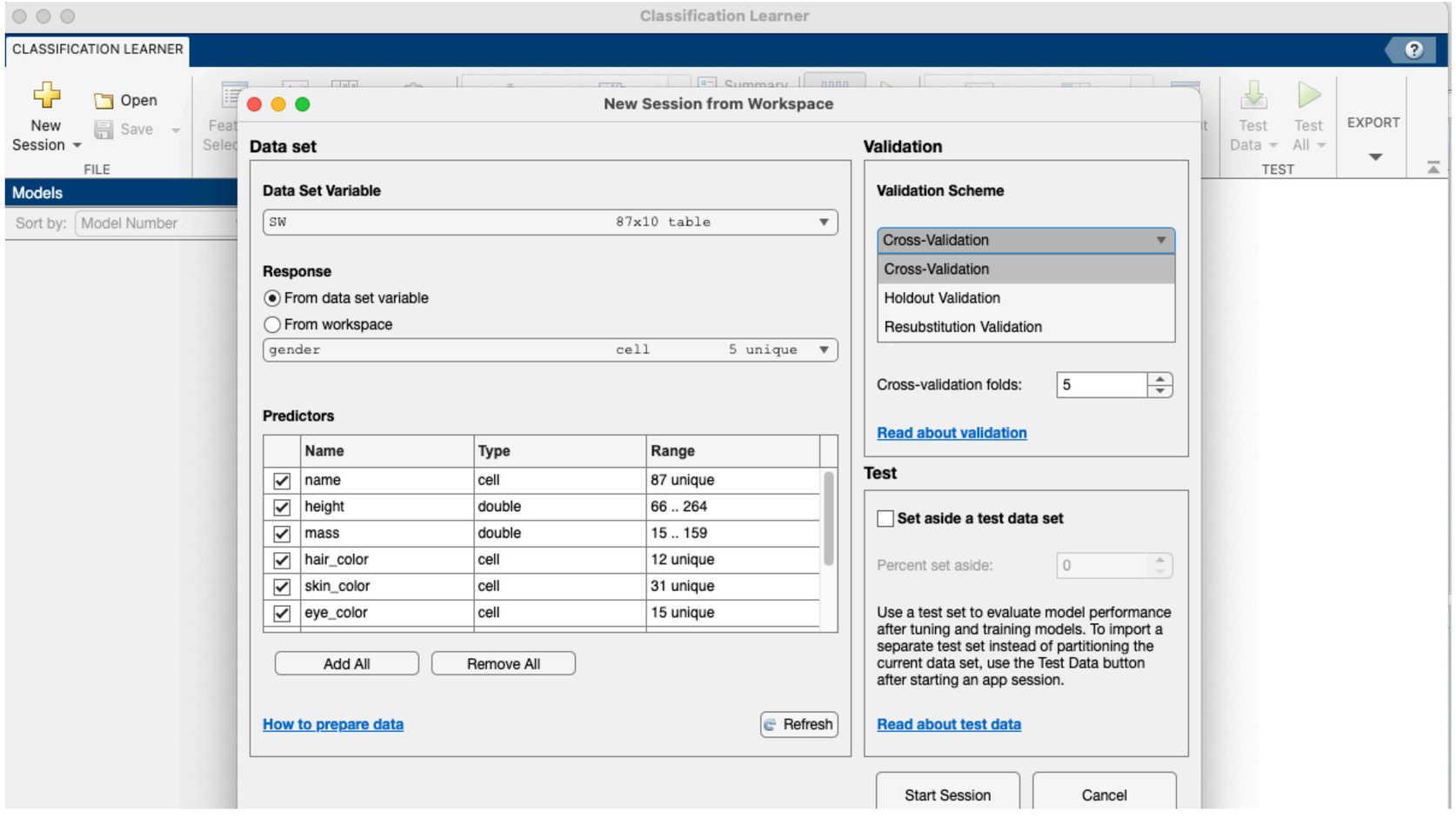


Classification Learner

General workflow

Select data → Select model and options → Train model → Assess model performance → Export model

Select data

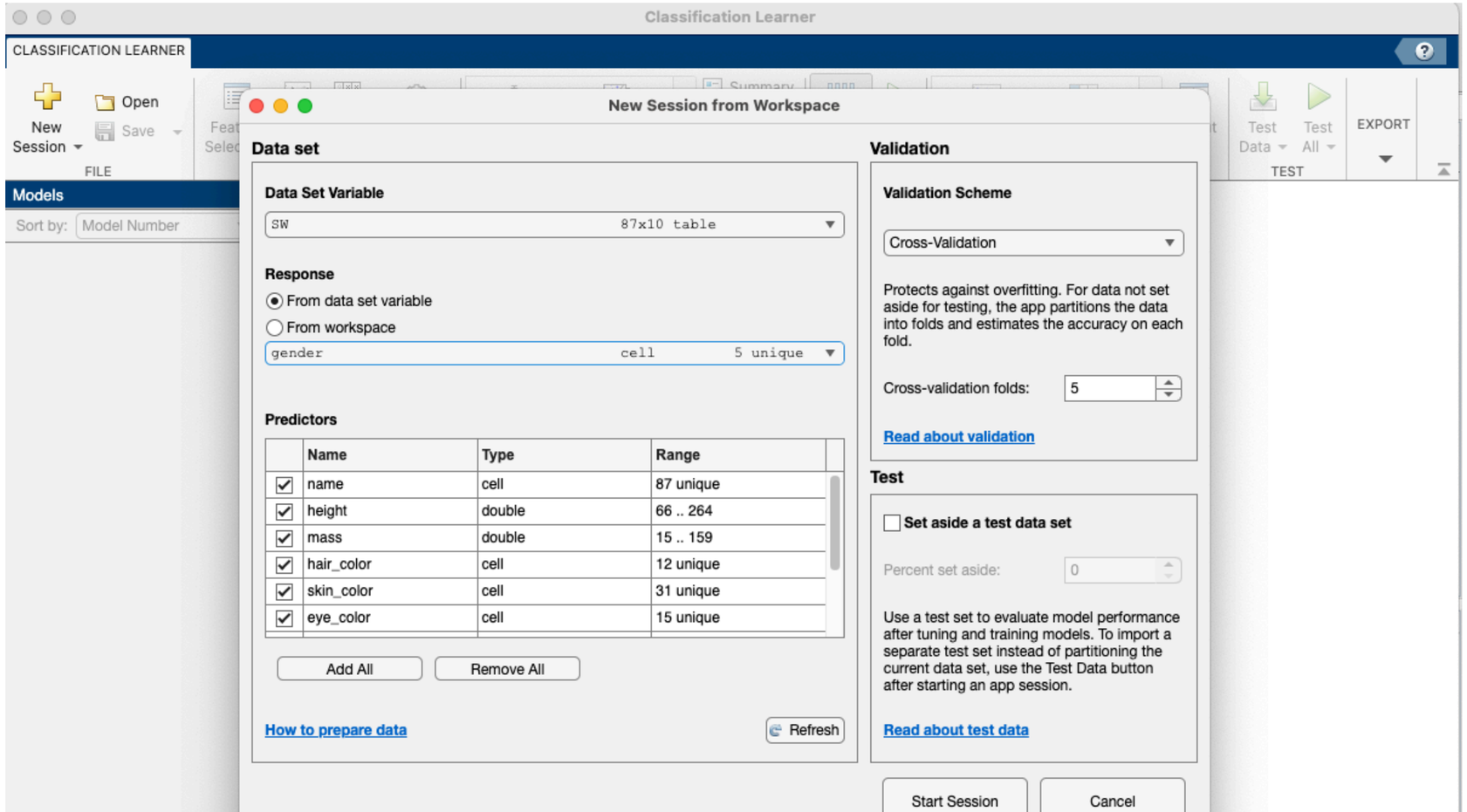


Classification Learner

General workflow

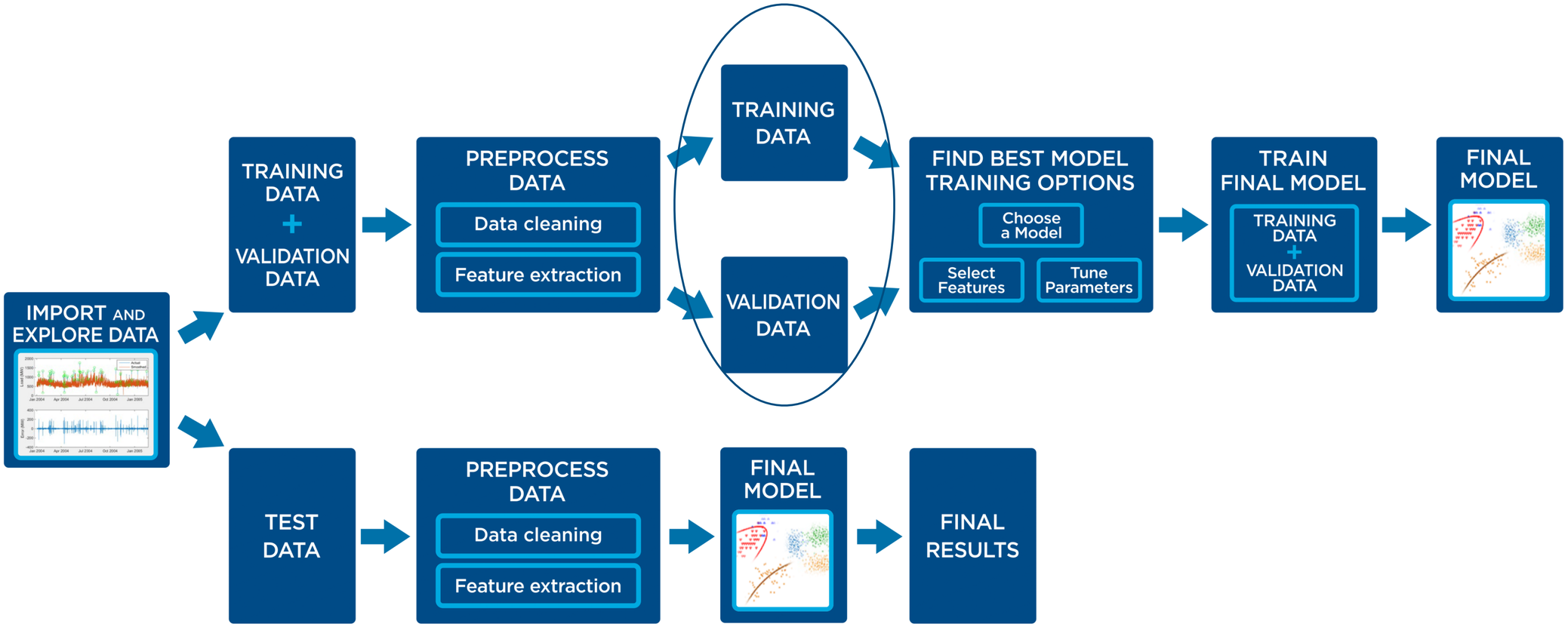
Select data → Select model and options → Train model → Assess model performance → Export model

Select data



Classification Learner

Validation



Classification Learner

Training data

This type of data builds up the machine learning algorithm. The training data feed the algorithm input data, corresponding to an expected output

The model evaluates the data repeatedly to learn more about the data's behavior and then adjusts itself to serve its intended purpose

Validation data

During training, validation data infuses new data into the model that it hasn't been evaluated before

Validation data provides the first test against unseen data, allowing data scientists to evaluate how well the model makes predictions based on the new data

Classification Learner

Test data

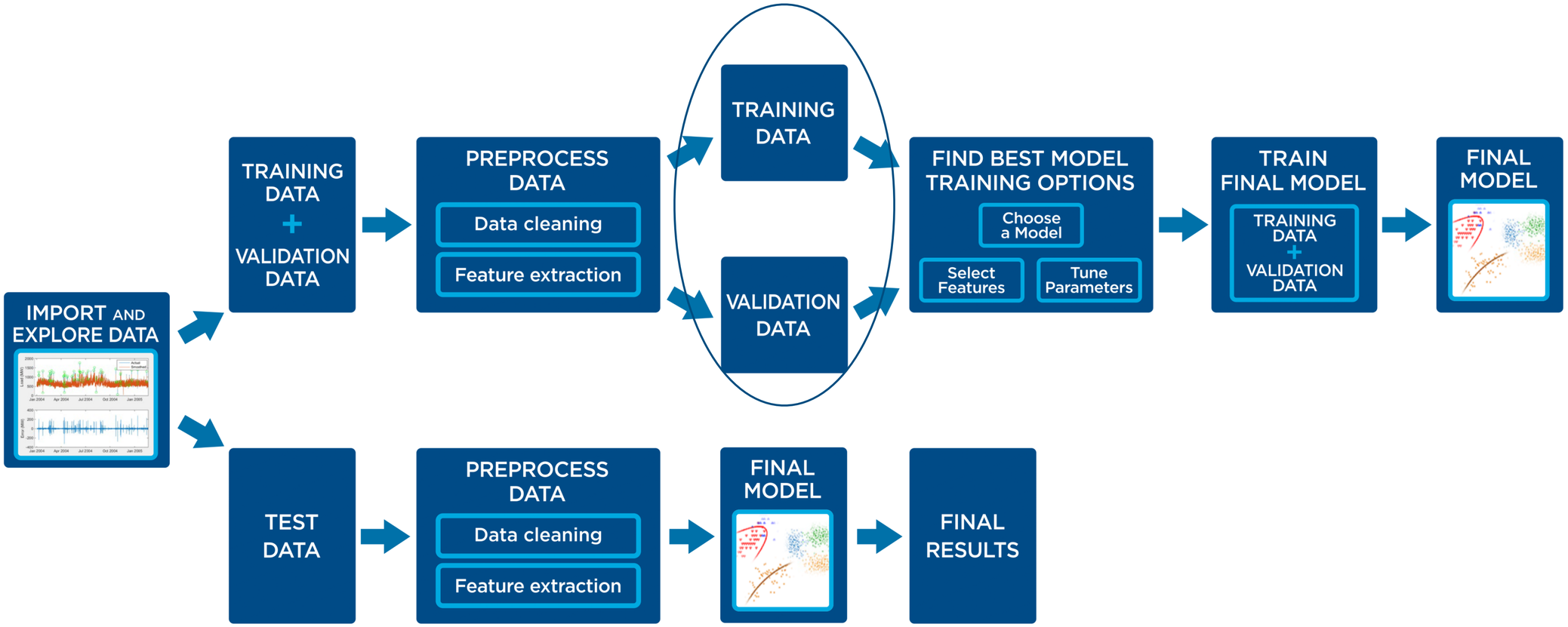
After the model is built, testing data once again validates that it can make accurate predictions

If training and validation data include labels to monitor the performance metrics of the model, the testing data should be unlabeled

Test data provides a final, real-world check of an unseen dataset to confirm that the ML algorithm was trained effectively.

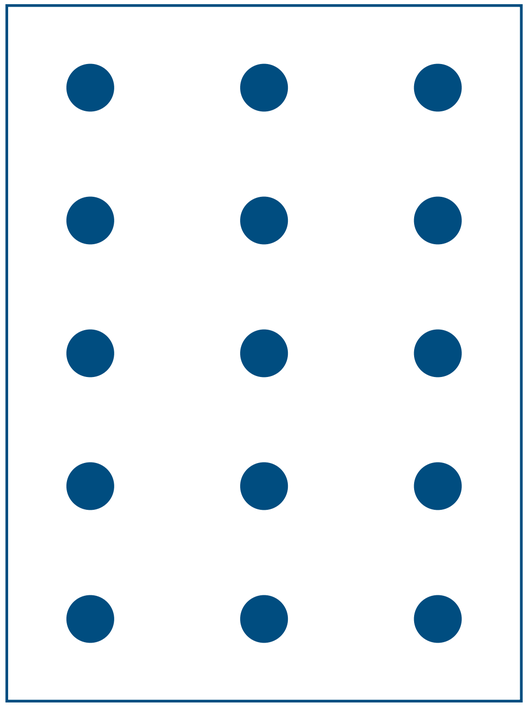
Classification Learner

Validation



Classification Learner

Validation



Training data

Validation data

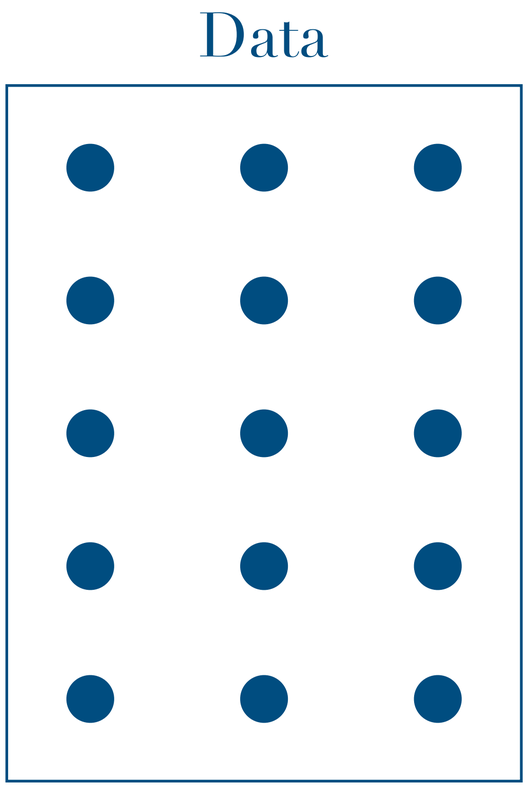
Validation

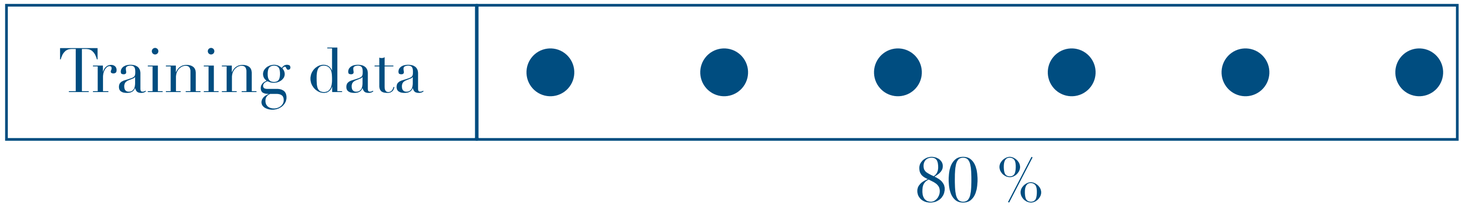
- Hold out
- K-fold Cross Validation
- Resubstitution Validation

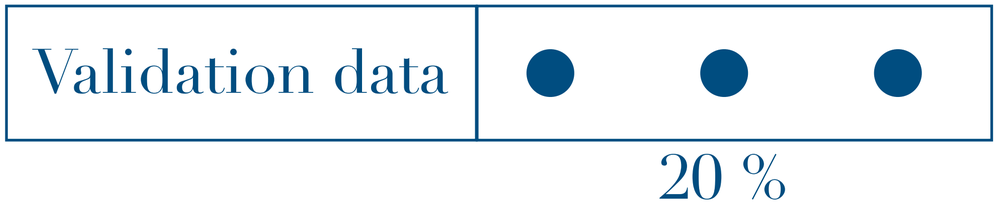
Classification Learner

Validation

- Hold out



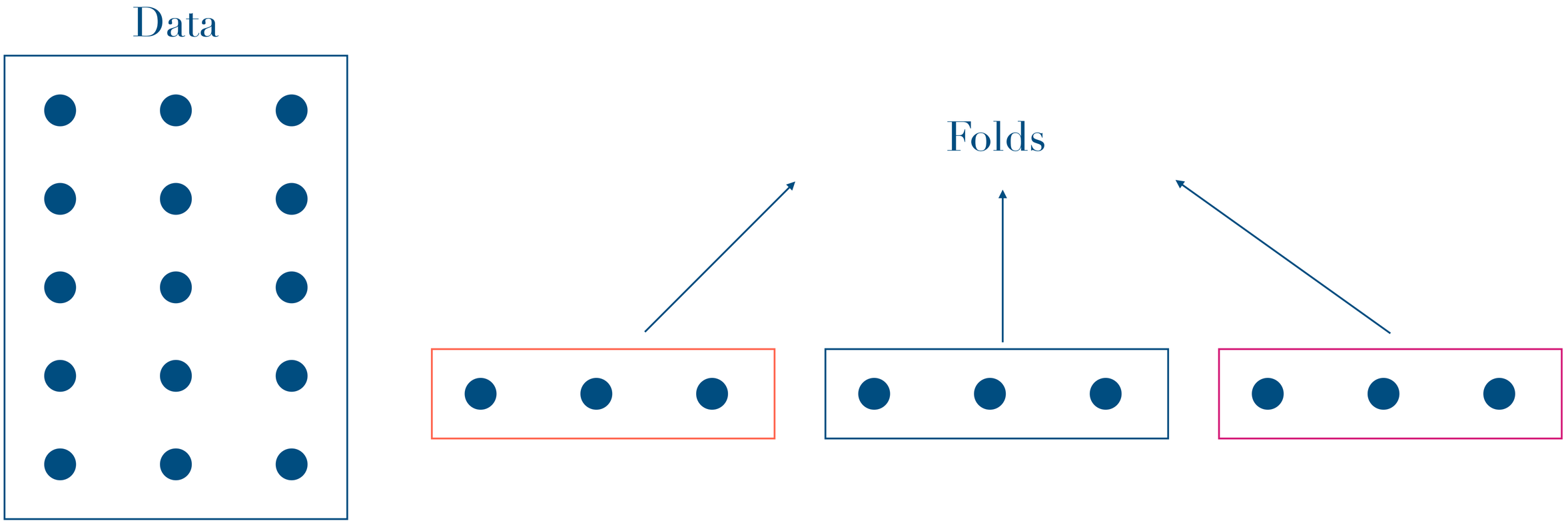




Classification Learner

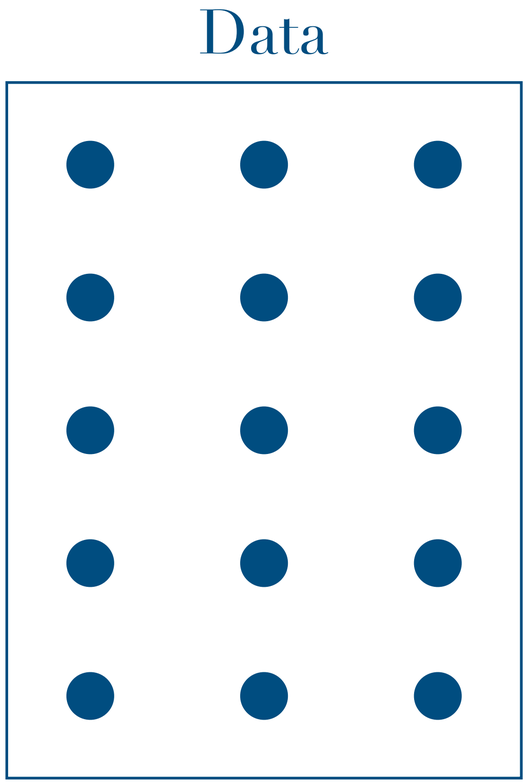
Validation

- K-fold Cross Validation



Classification Learner

Validation



- Resubstitution Validation

All data is used for training the model and the error rate is evaluated based on outcome vs. actual value from the same training data set