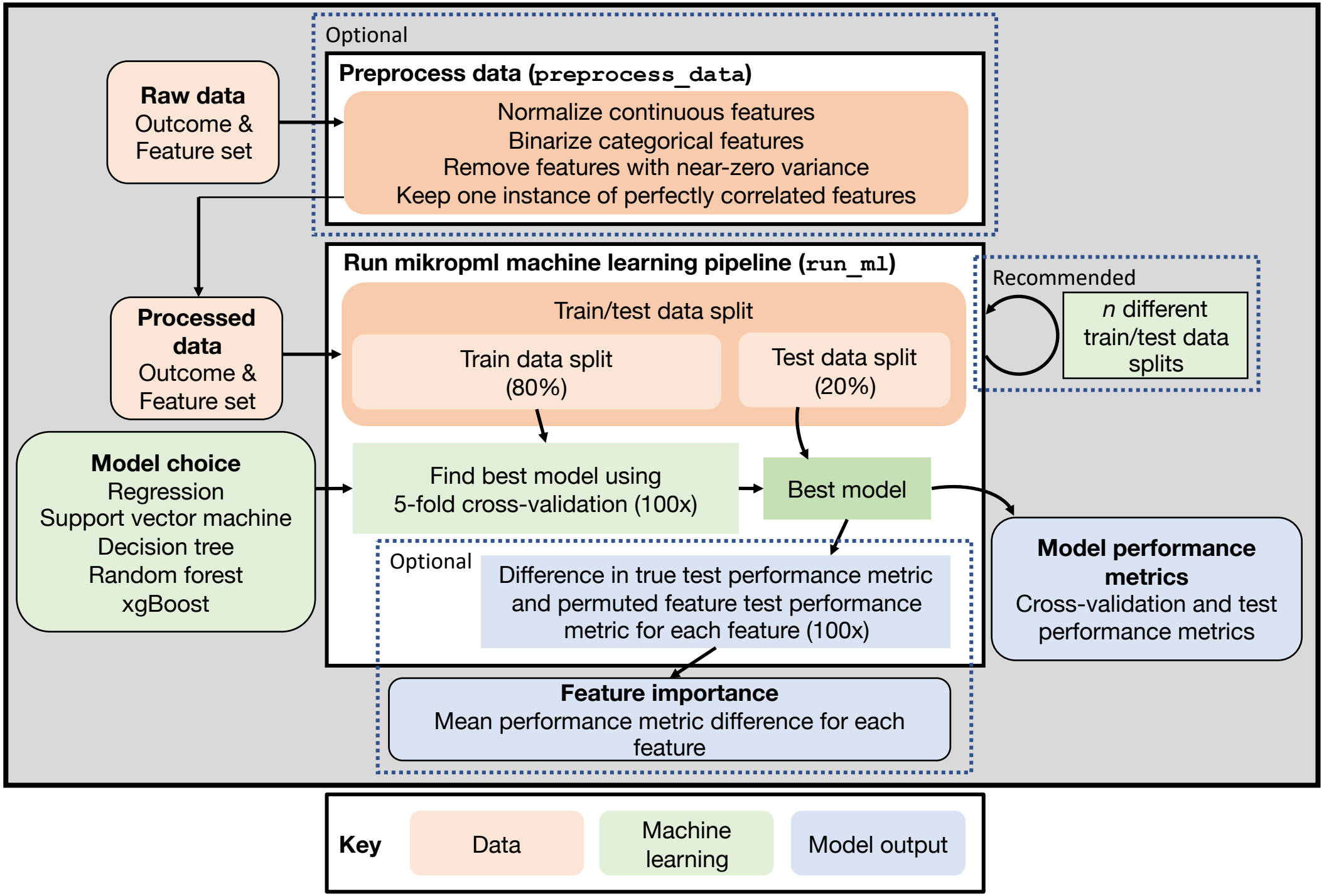
Optional
Preprocess data (preprocess_data)
Normalize continuous features
Binarize categorical features
Remove features with near-zero variance
Keep one instance of perfectly correlated features
Raw data
Outcome &
Feature set
Processed data
Outcome &
Feature set
Run mikropml machine learning pipeline (run_ml)
Train/test data split
Train data split
(80%)
Test data split
(20%)
Recommended
n different
train/test data
splits
Model choice
Regression
Support vector machine
Decision tree
Random forest
xgBoost
Find best model using
5-fold cross-validation (100x)
Best model
Optional
Difference in true test performance metric
and permuted feature test performance
metric for each feature (100x)
Model performance
metrics
Cross-validation and test
performance metrics
Feature importance
Mean performance metric difference for each
feature
Key
Data
Machine
learning
Model output