MACHINE LEARNING

kyle bradbury, ph.d.

Goal: Recognize patterns and trends in and make predictions on data, using algorithms

> "Field of study that gives computers the ability to learn without being explicitly programmed"
>
> ~Arthur Samuel (artificial intelligence pioneer)

clarifying **DEFINITIONS**

Machine Learning
Pattern Recognition
} Synonyms

Data Mining → Knowledge discovery from large ("Big") data sets

Artificial Intelligence → "intelligence" exhibited by machines or software

Data Analytics
Data Science
Business Intelligence
} Catchall for any analysis of data, big or small

characteristics of **BIG DATA**

- Volume – quantity of existing data
- Velocity – streaming data
- Variety – data forms (numbers, text, multimedia)
- Veracity – uncertainty of data quality

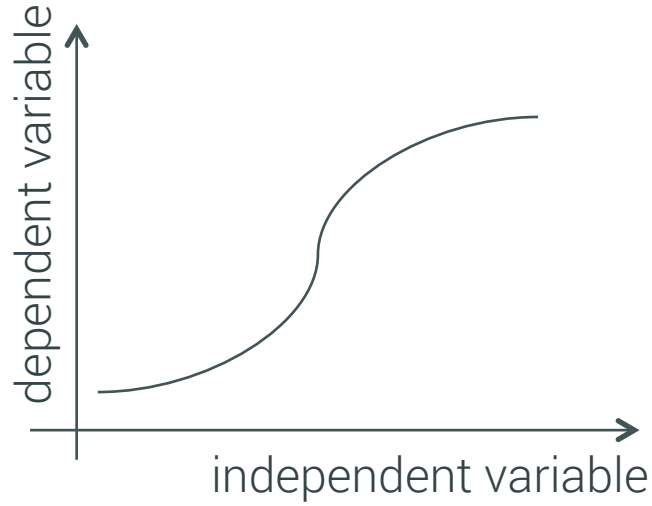
Source: IBM, 2014

taxonomy of **ANALYTICS**

- Descriptive – what HAS happened?
- Predictive – what COULD happen?
- Prescriptive – what SHOULD happen?

Source: IBM, 2013

know your **VARIABLES**



independent variable
explanatory variable
predictor variable
controlled variable
input variable
regressor
feature

dependent variable
response variable
response variable
experimental variable
output variable
regressand
target

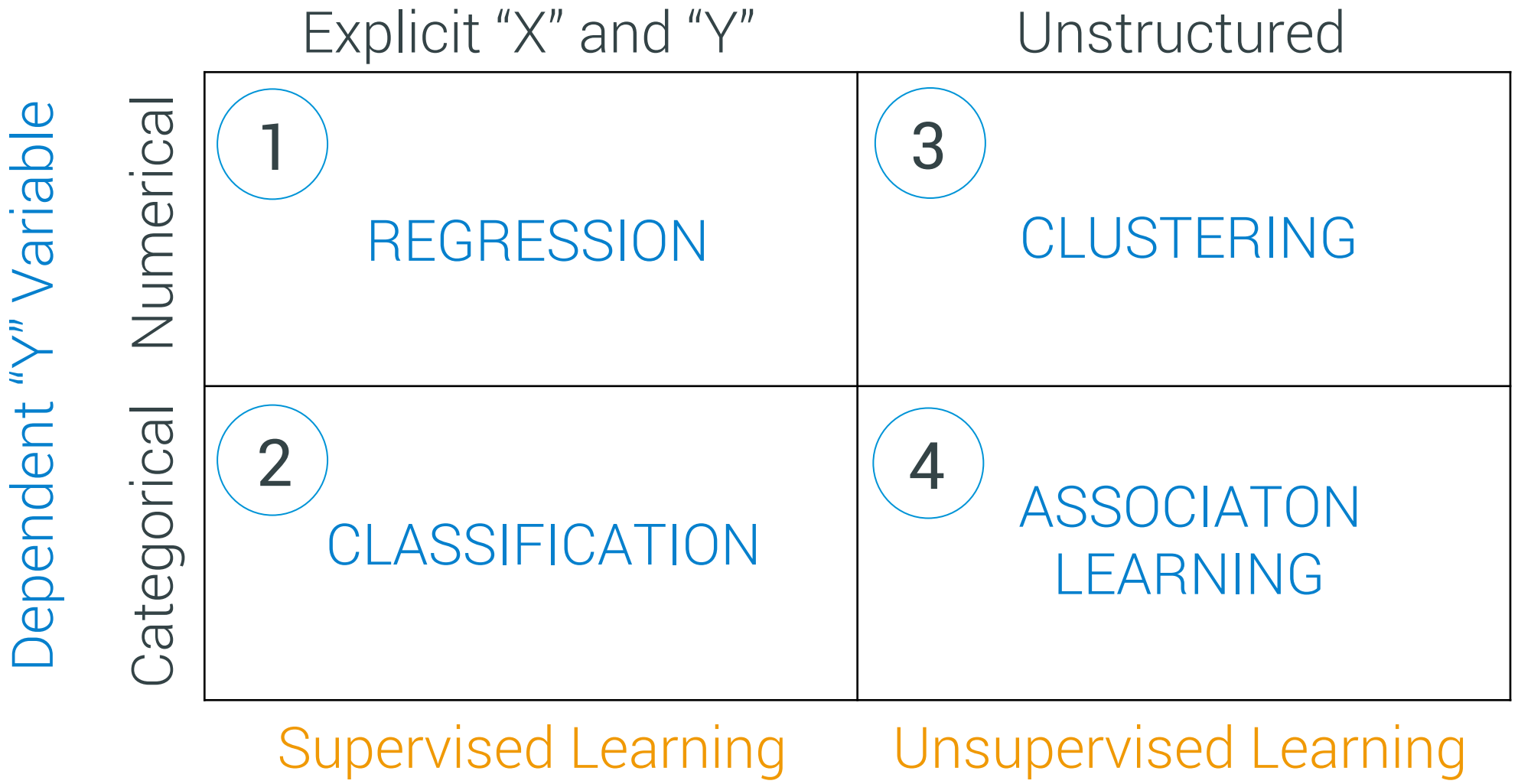
Each “sample” of a variable could be a scalar, vector, or matrix

1		2		3	4
“x”	“y”	“x”	“y”		
0.0	0.0	0.0	blue	0.0	butter
0.2	0.8	0.2	blue	0.2	eggs
0.4	0.9	0.4	blue	0.4	flour
0.5	1.0	0.5	yellow	0.5	mixer
0.7	-0.4	0.7	yellow	0.7	beef
1.1	-1.6	1.1	yellow	1.1	grill
2.8	6.5	2.8	green	2.8	foil
2.9	6.9	2.9	yellow	2.9	sugar
3.0	8.4	3.0	green	3.0	ketchup
5.3	-5.3	5.3	blue	5.3	rolls

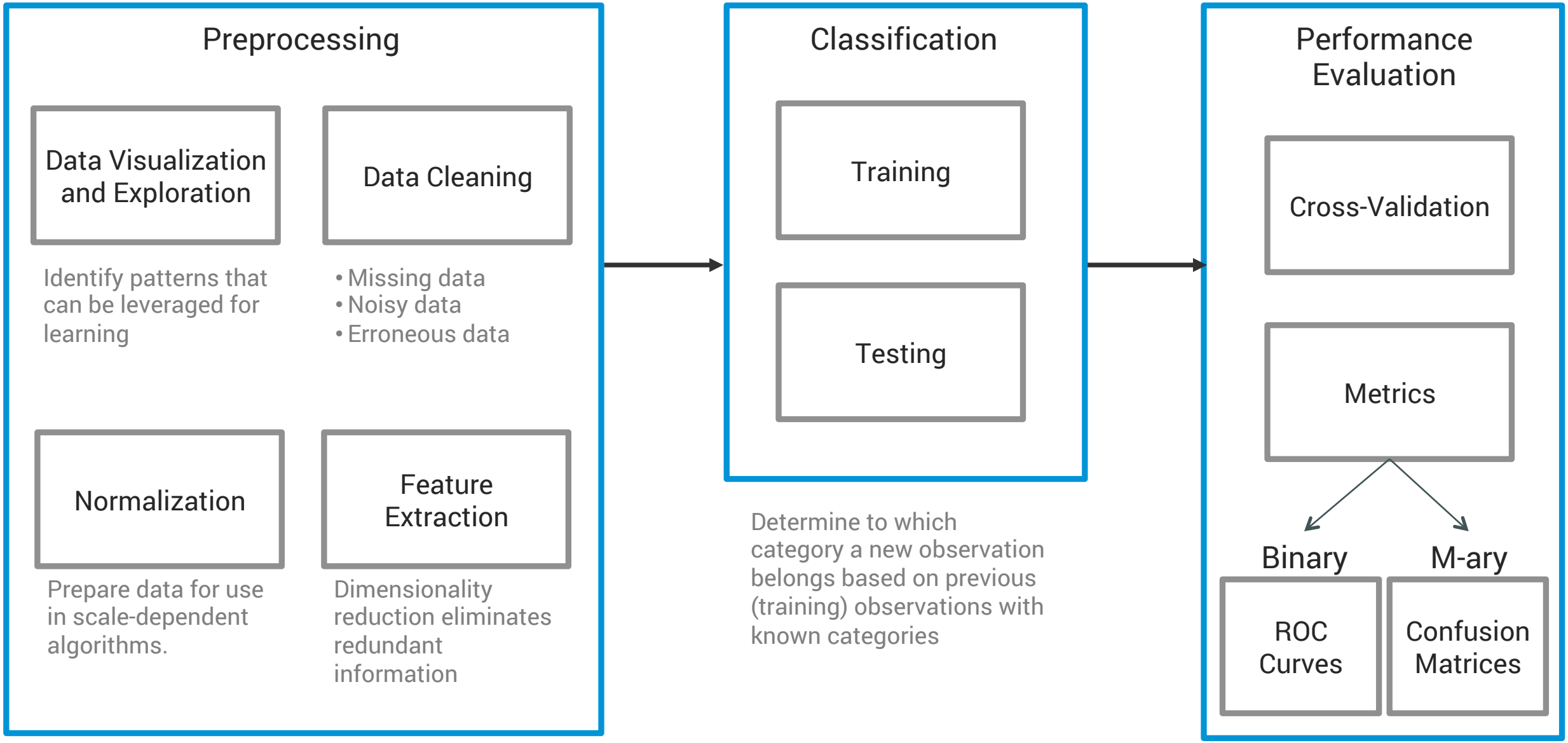
EXAMPLES

1. Forecast trends
   Estimate causal relationships
2. Face detection
   Spam detection
   Medical diagnosis
3. Market segmentation
   Estimate densities
4. Relationship discovery
   Point of sale product association

categories of **MACHINE LEARNING**



machine learning: **CLASSIFICATION**



classification **ALGORITHMS**

Fisher's Linear Discriminant (FLD)
Logistic Regression (Logit)
Naïve Bayes
Perceptron

Linear

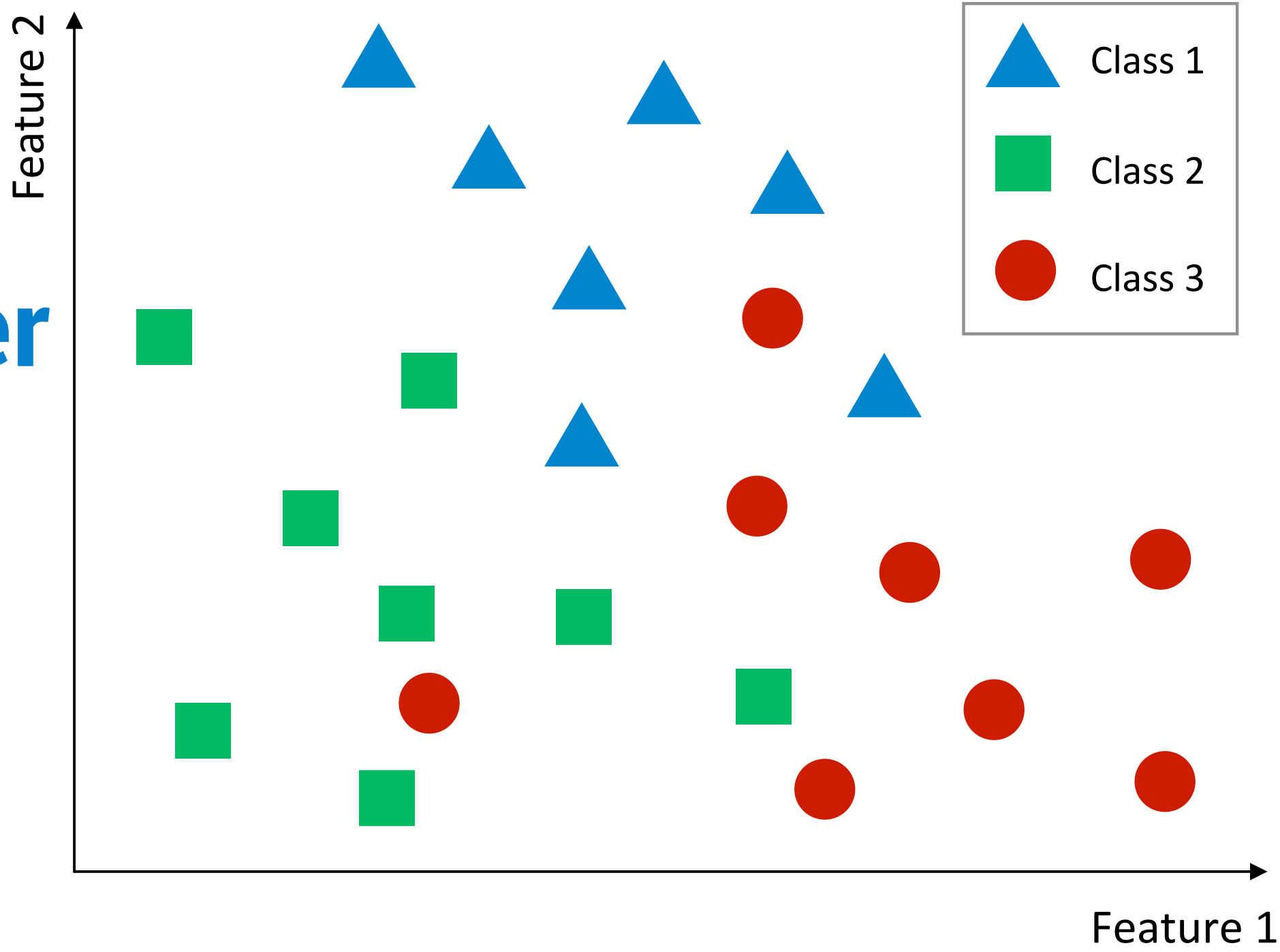
Support Vector Machine (SVM)
Kernel Estimation (e.g. k-Nearest Neighbor)
Decision Trees
Neural Networks

How to choose?

Accuracy
Efficiency
Interpretability

k-nearest neighbor **Classifier**

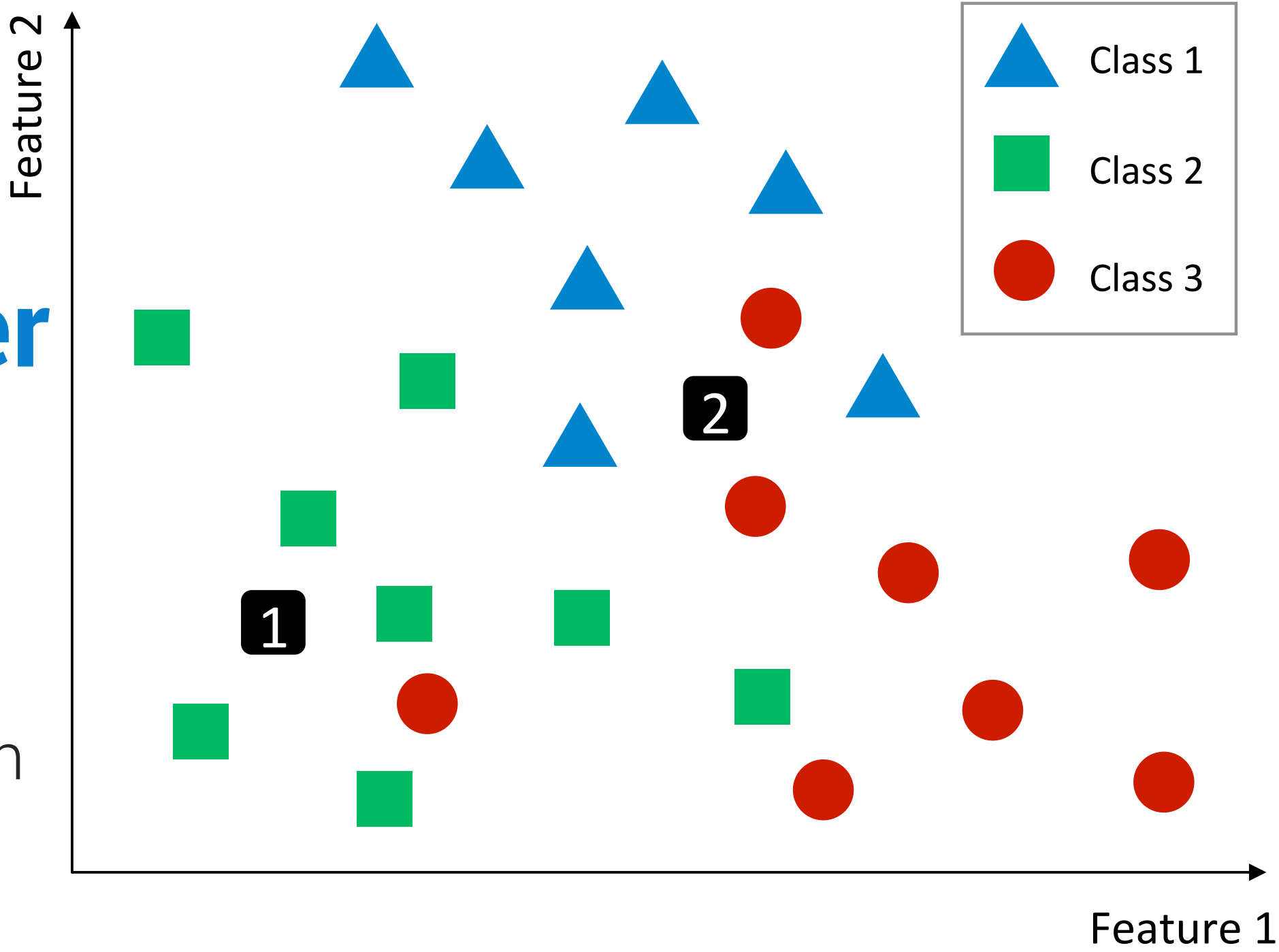
Step 1:
Training



k-nearest neighbor **Classifier**

Step 2:

Place new observations in feature space

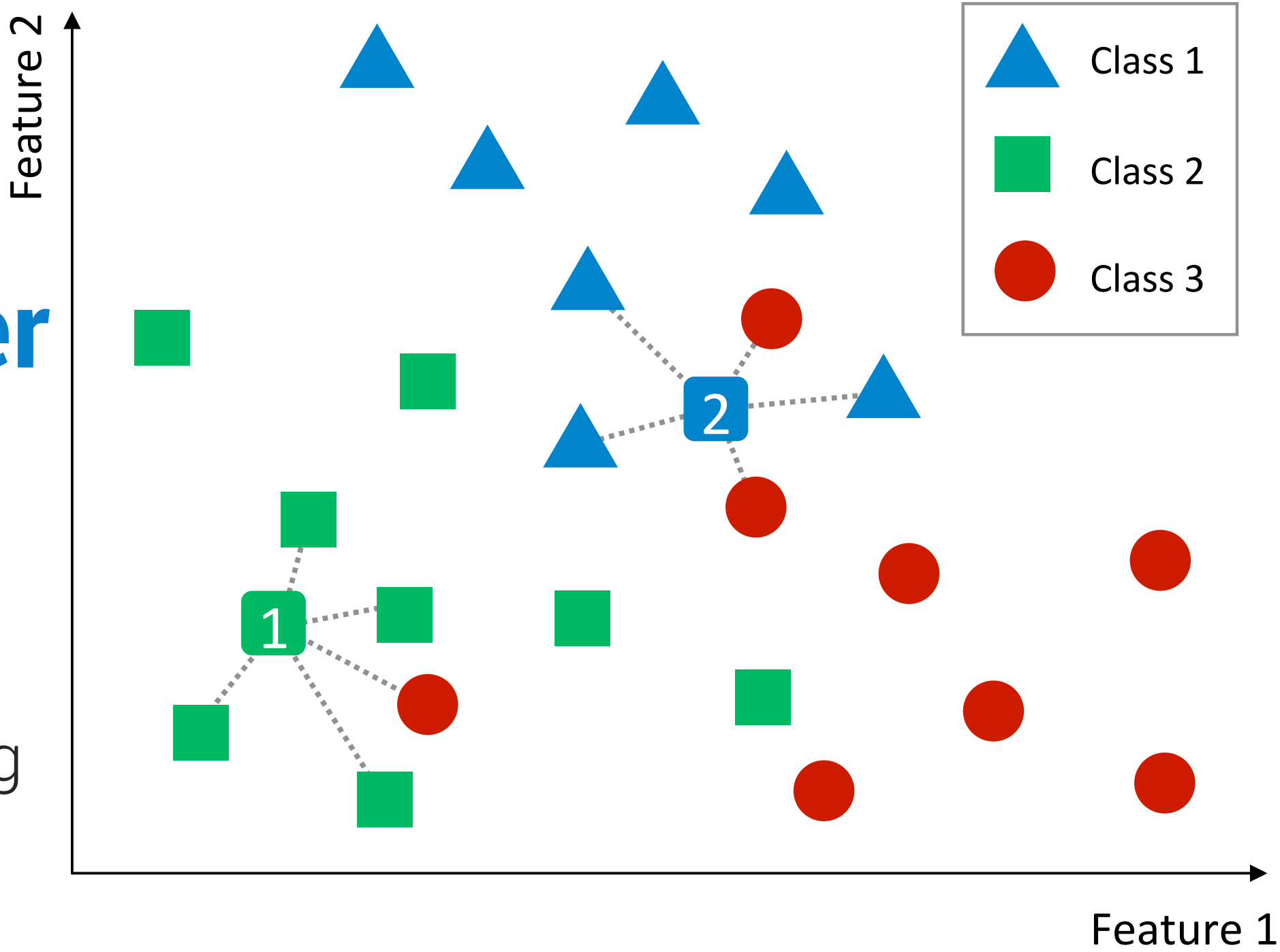


k-nearest neighbor

Classifier

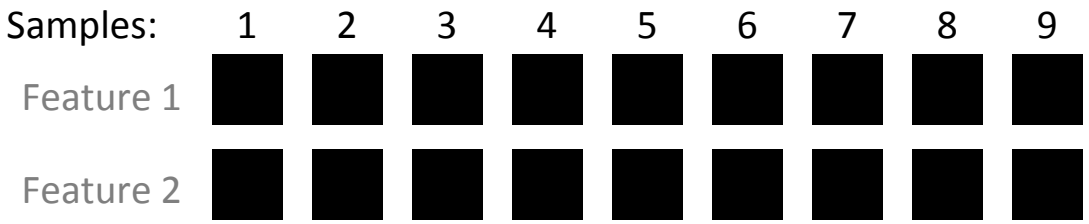
Step 3:

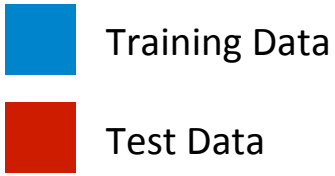
Classify the new data using the trained classifier



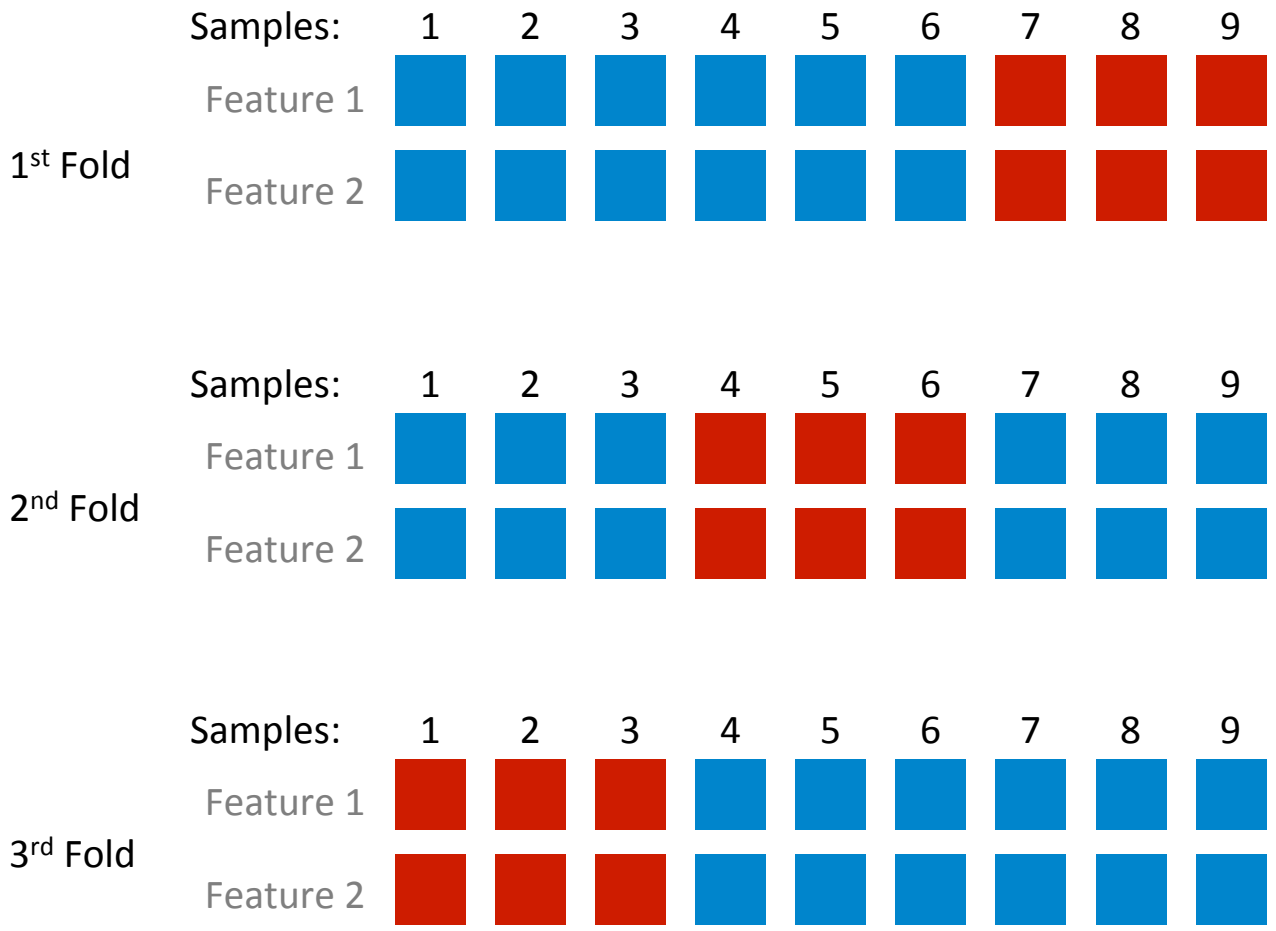
k-folds **CROSS-VALIDATION**

Original feature set with two features and 9 samples

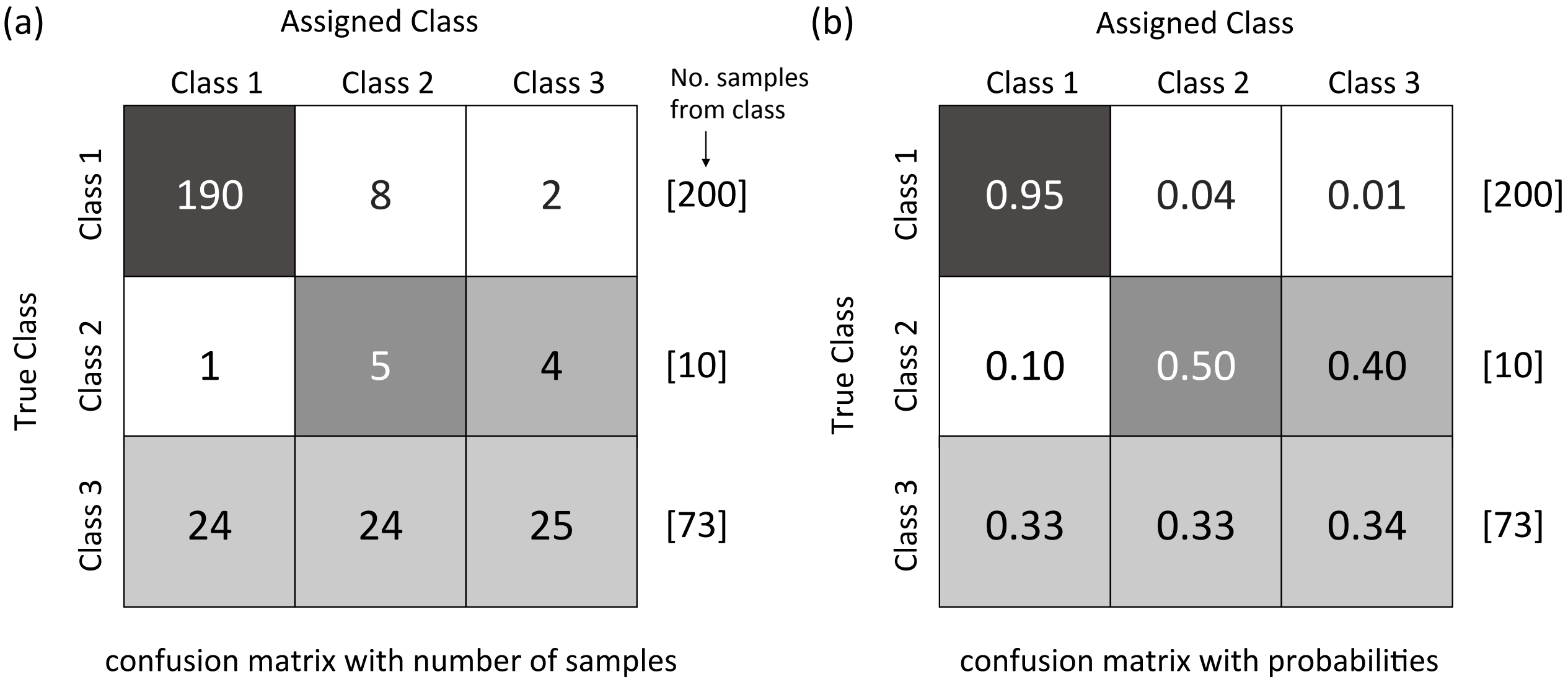




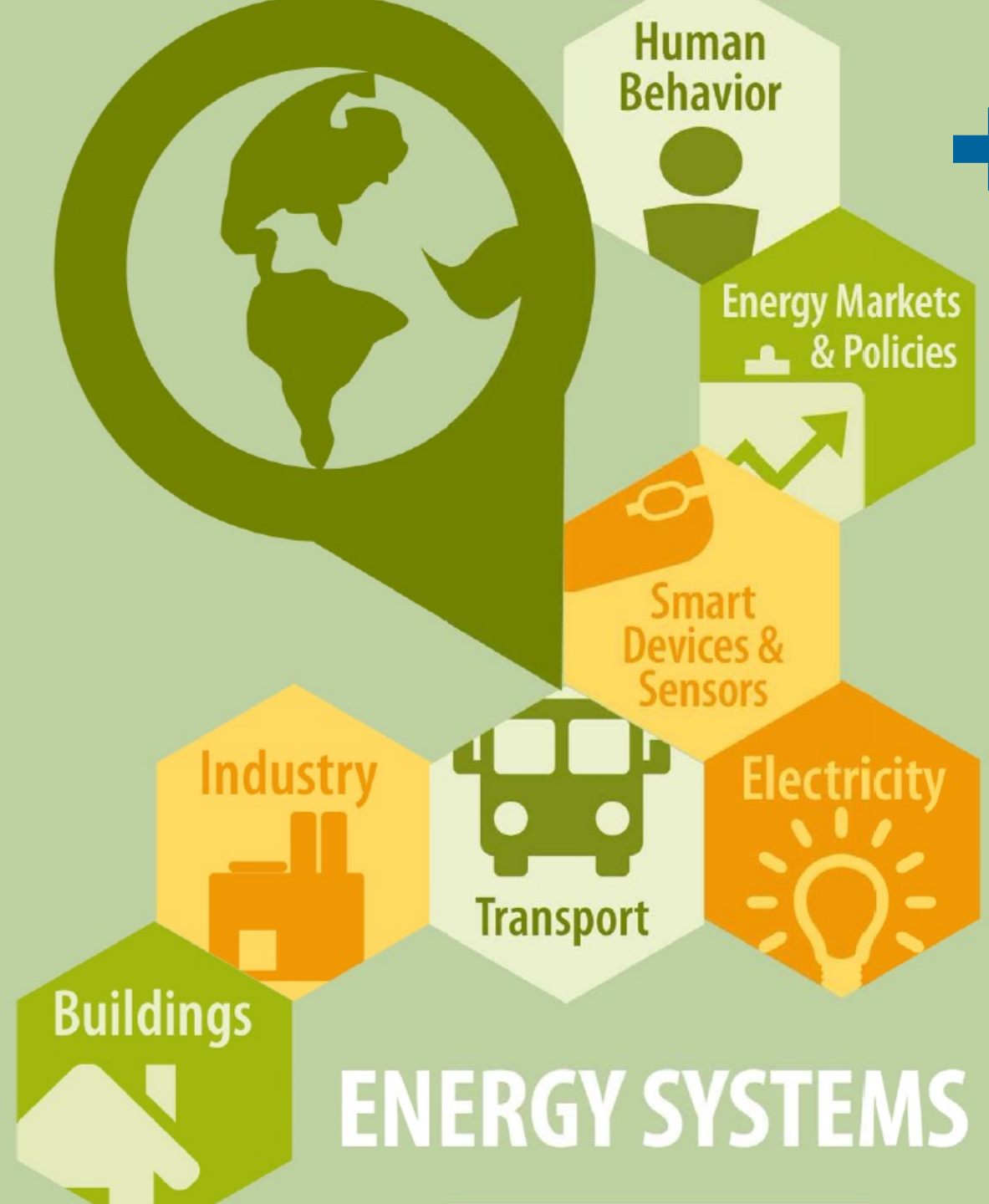
Each of the 3 folds in 3-fold cross validation

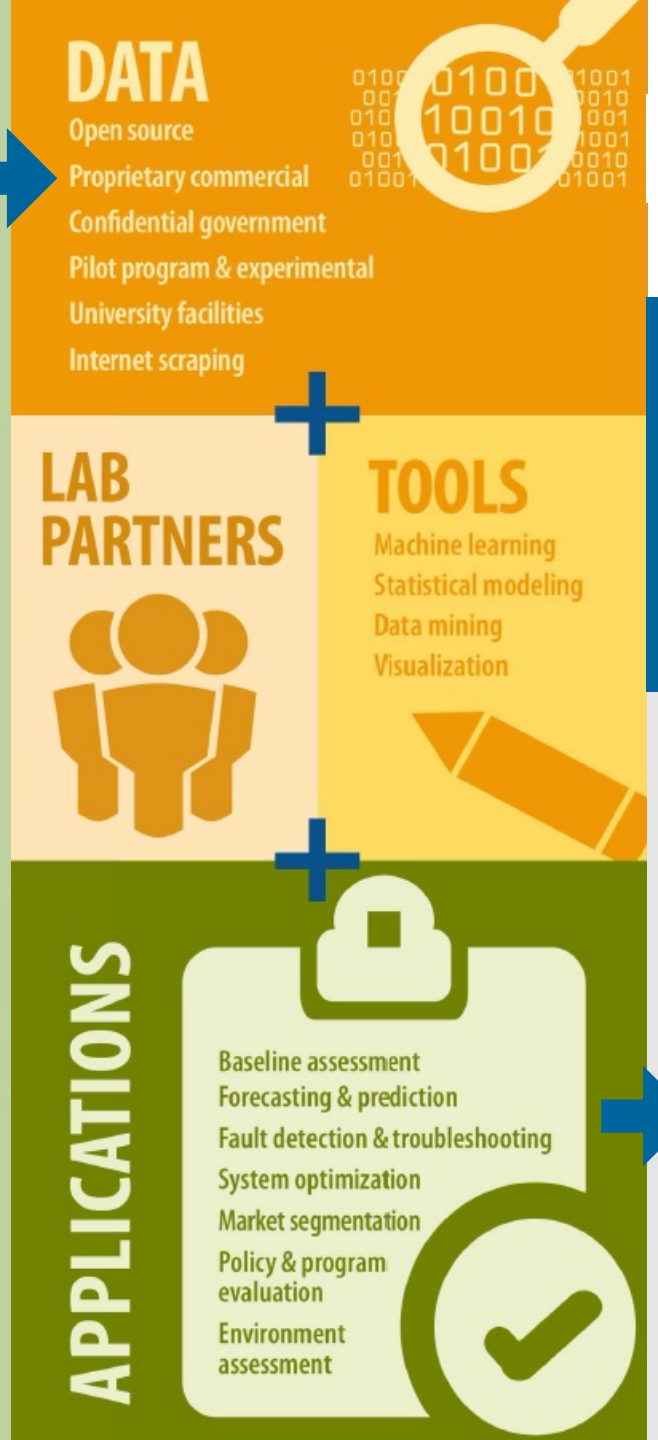


CONFUSION MATRIX



demo





Duke UNIVERSITY | ENERGY INITIATIVE

ENERGY DATA ANALYTICS LAB

email: kyle.bradbury@duke.edu