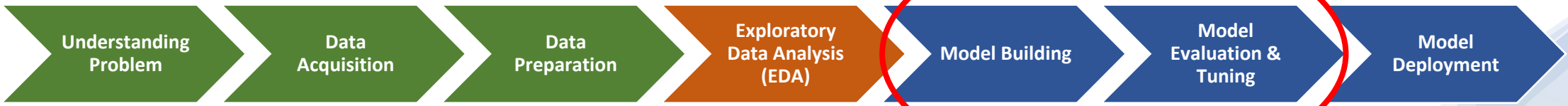
https://tinyurl.com/DSinMed

Core Processes in Data Science

- Statistical Analyses
- Data Visualization
- Data Summarization

- Cross-Validation
- Parameter Tuning
- Feature Selection
- Refining Algorithm



- Data Cleaning
- Data Organizing
- Data Transformation

- Algorithm Selection
- Identify Data Pattern
- Model Training

- Designing Pipelines
- Updating Model with new data

Introduction to Supervised Learning

Thammakorn Saethang

Machine Learning

Machine learning is a branch of artificial intelligence that enables computer programs to automatically learn and improve from experience. Machine learning algorithms learn from datasets, and then based on the patterns identified from the datasets, make predictions on unseen data.

Machine learning algorithms can be mainly categorized into two types:

- Supervised learning algorithms
- Unsupervised learning algorithms

Supervised & Unsupervised Learning

- **Supervised learning**
  - Outcome variable is available
  - Involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*.

- **Unsupervised learning**
  - No outcome variable (the true labels for the outputs are not known)
  - There are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

Regression Versus Classification Problems

Supervised learning algorithms are divided further into two types:

- Regression
  - Quantitative outcome variable
    - Predict a continuous value, for example, the price of a house, blood pressure of a person, a student's score in a particular exam, etc.
- Classification
  - Qualitative outcome variable
    - Predict a discrete value such as whether or not a tumor is malignant, whether a student is going to pass or fail an exam, etc.

Linear Regression (cont.)

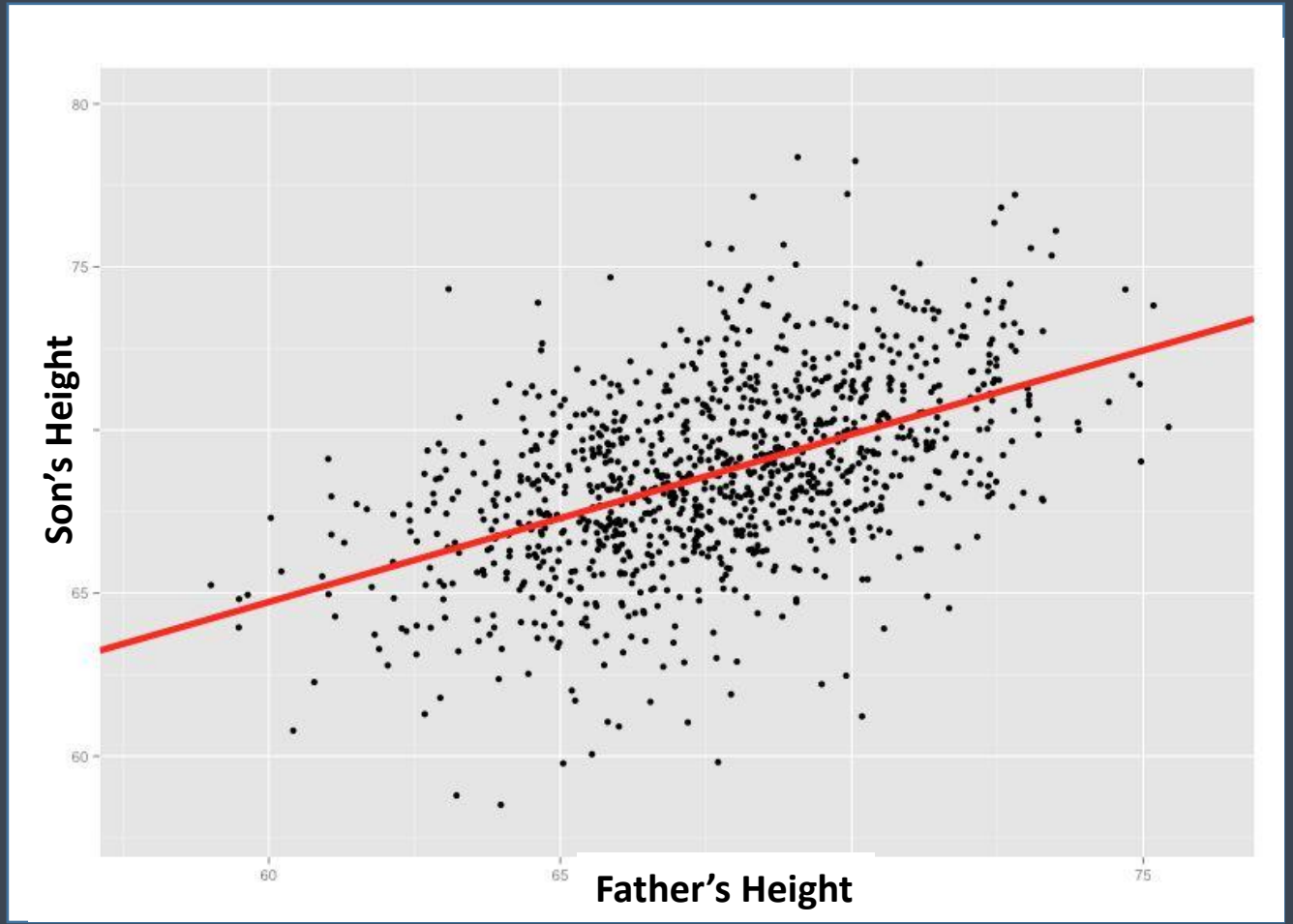
- Then we draw a line that minimize the distance between all the points and their distance to this line
- This line can be shown as:

intercept slope

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \quad ^*$$

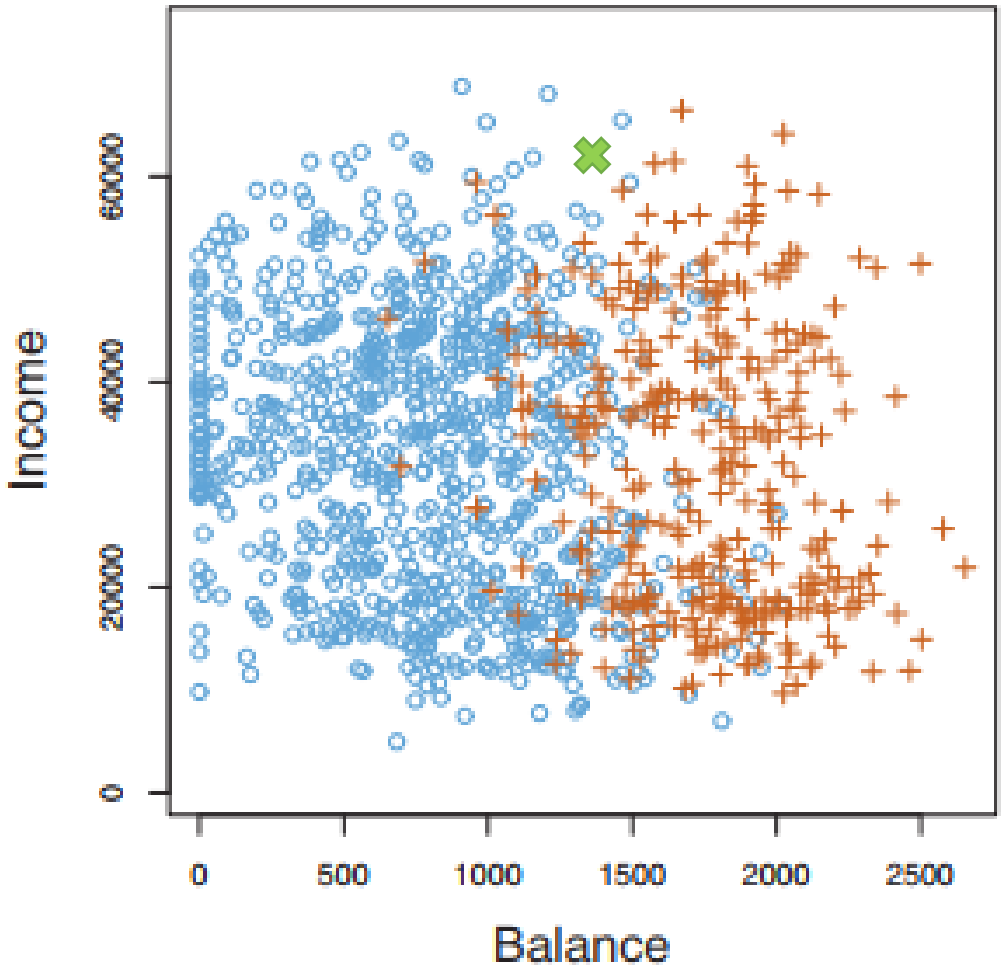
$$\widehat{Son's\ Height} = \widehat{\beta_0} + \widehat{\beta_1}(Father's\ Height)$$

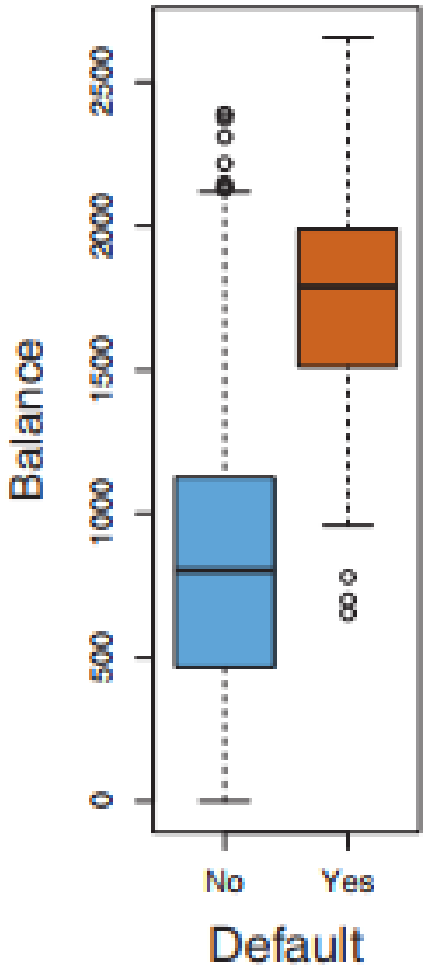
$$\hat{f}(X) = \widehat{\beta_0} + \widehat{\beta_1} x$$

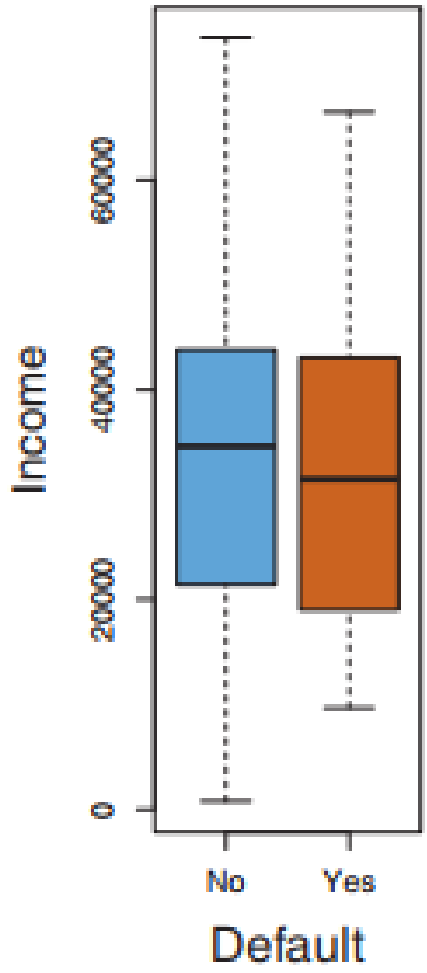


* Also, y = a+bx, y = ax+b , y = mx+c

Classification Problem





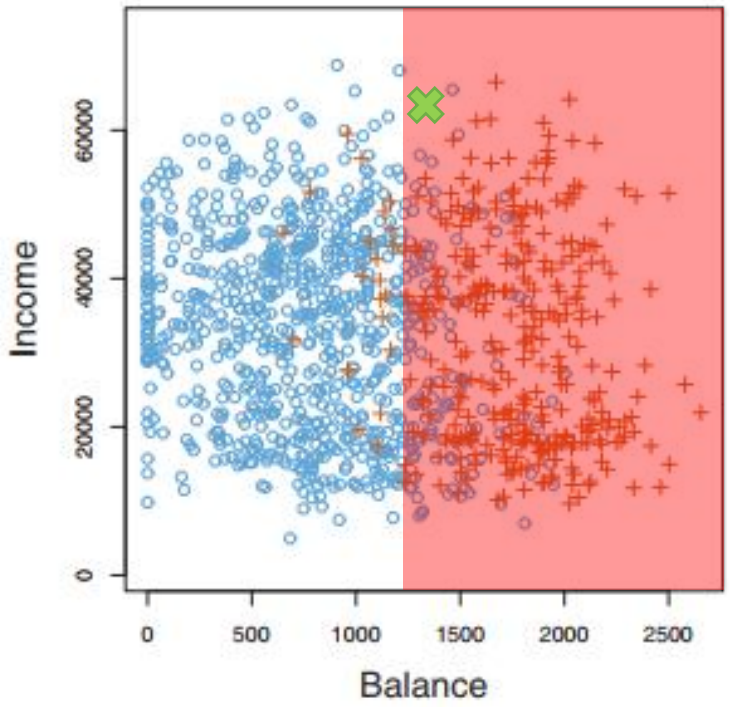


1. **What will be our model to explain this data?**
2. **If we have a new data, can we make a prediction for that data?**

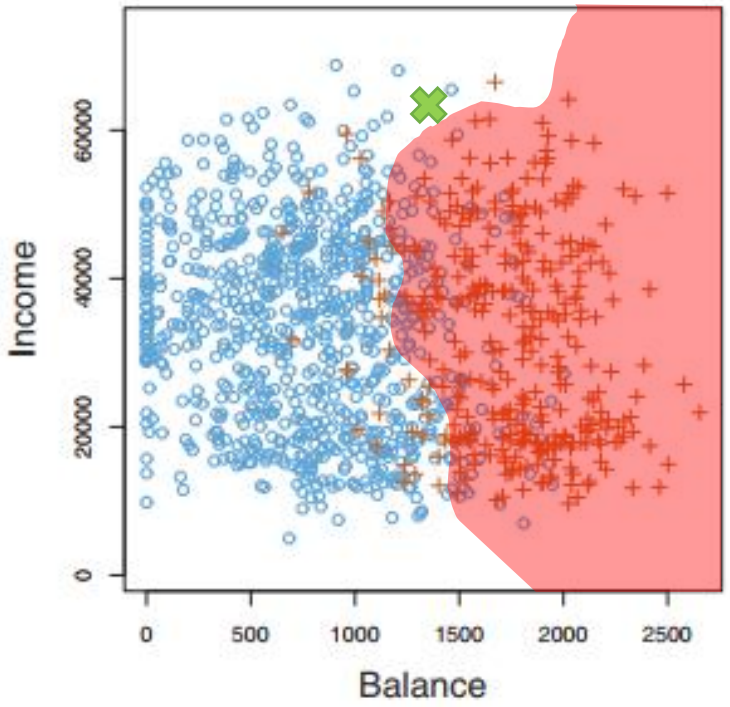
(✖)

Classification Problem

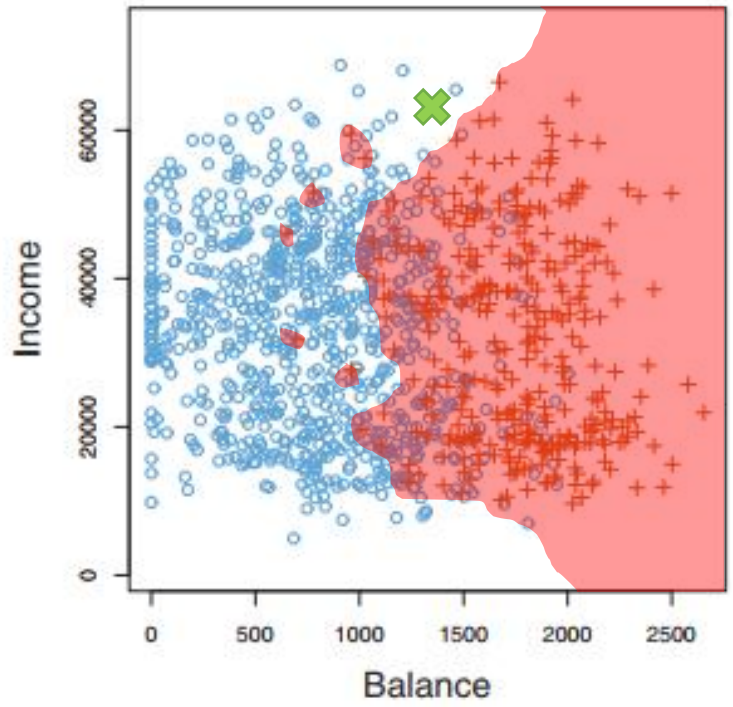
Model 1



Model 2



Model 3



Data Preparation for Machine Learning (ML)

Data preparation process:

- Cleaning
- Organizing
- Removing outliers
- Denoising
- Validation
- Standardization
- Transformation*
- etc.

Importance of data preparation:

- ML algorithms require specific data format*
- Garbage in, garbage out (GIGO)
- Modern ML models requires data normalization
- etc.

Types of Data & Transformation Process

Numerical data

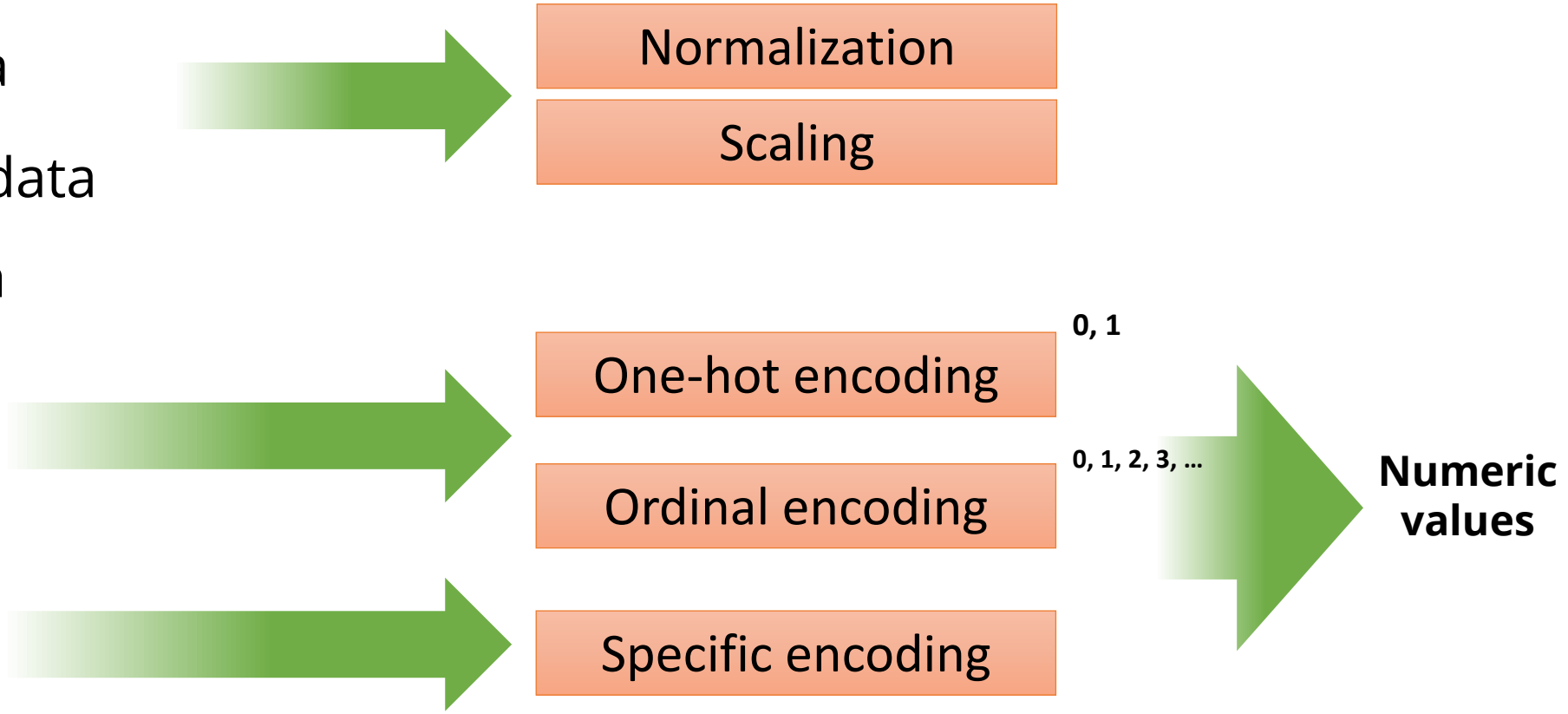
- Discrete data
- Continuous data

Categorical data

- Nominal
- Ordinal

Textual data

Image/sound/ other data



One-Hot Encoding Example

ID	gender
1	M
2	F
3	M

ID	gender_F	gender_M
1	0	1
2	1	0
3	0	1

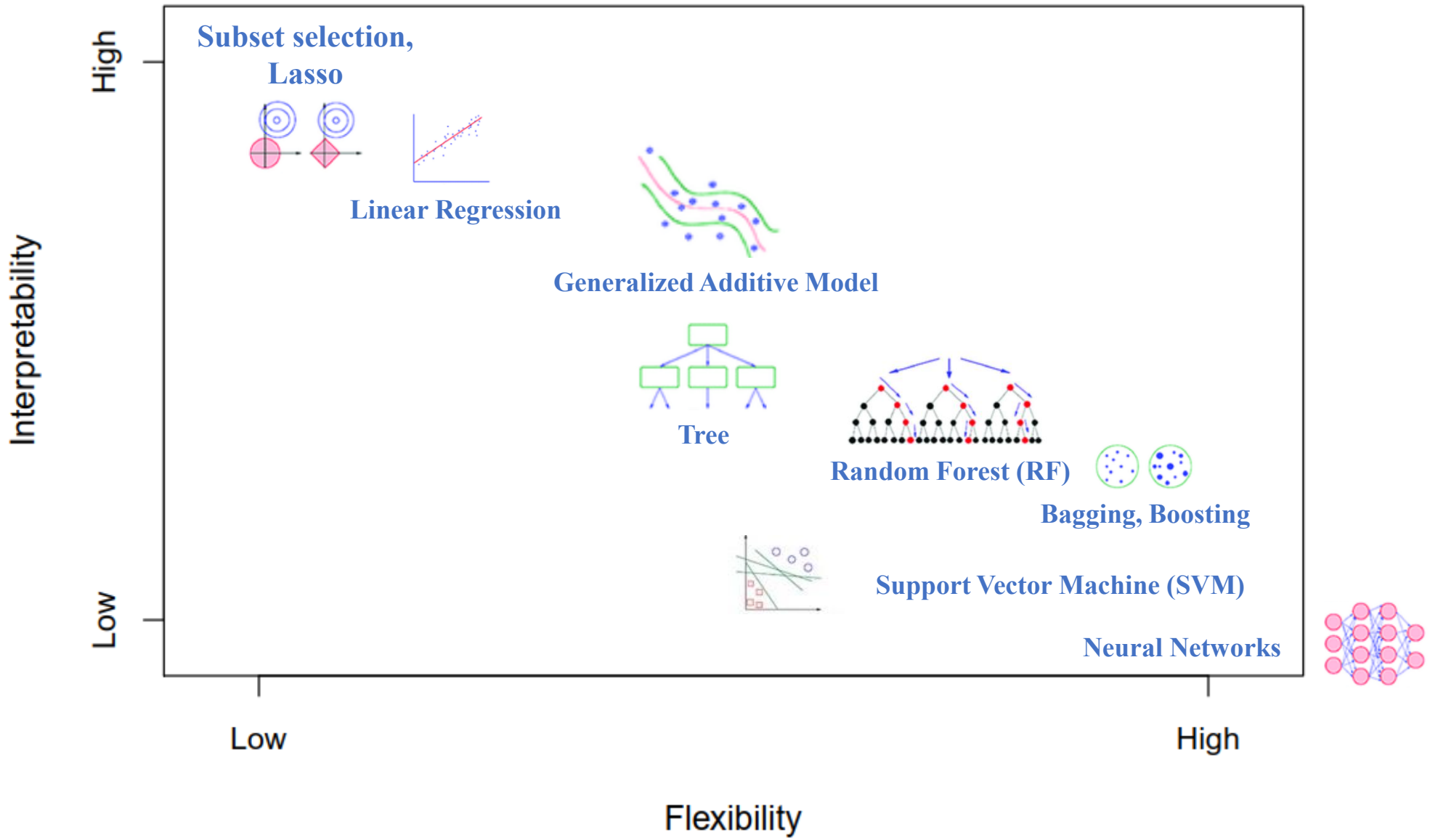
One-hot encoding

- get_dummies() in Pandas

ID	blood_gr
1	A
2	B
3	AB
4	0

ID	bg_A	bg_B	bg_AB	bg_O
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

Interpretability vs. Flexibility of Predictive Models



https://www.ncbi.nlm.nih.gov/pubmed/29354080

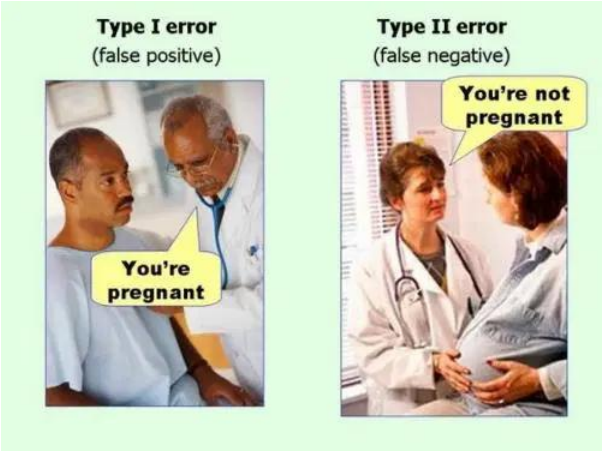
Evaluating Model Performance

- In regression problems
  - Mean squared error
  - Root mean squared error
  - R-squared

- In classification problems
  - Accuracy
  - Sensitivity
  - Specificity
  - Area under receiver operating characteristic curve

Confusion matrix

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN) — Type 2 error
	N	False Positives (FP) — Type 1 error	True Negatives (TN)



Accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

Balanced Accuracy (BACC)

$$BACC = (TP/P + TN/N)/2$$

F1 Score

is the harmonic mean of Precision and Sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews Correlation Coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Sensitivity or True Positive Rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

Specificity (SPC) or True Negative Rate (TNR)

$$SPC = TN/N = TN/(FP + TN)$$

Precision or Positive Predictive Value (PPV)

$$PPV = TP/(TP + FP)$$

Negative Predictive Value (NPV)

$$NPV = TN/(TN + FN)$$

Fall-out or False Positive Rate (FPR)

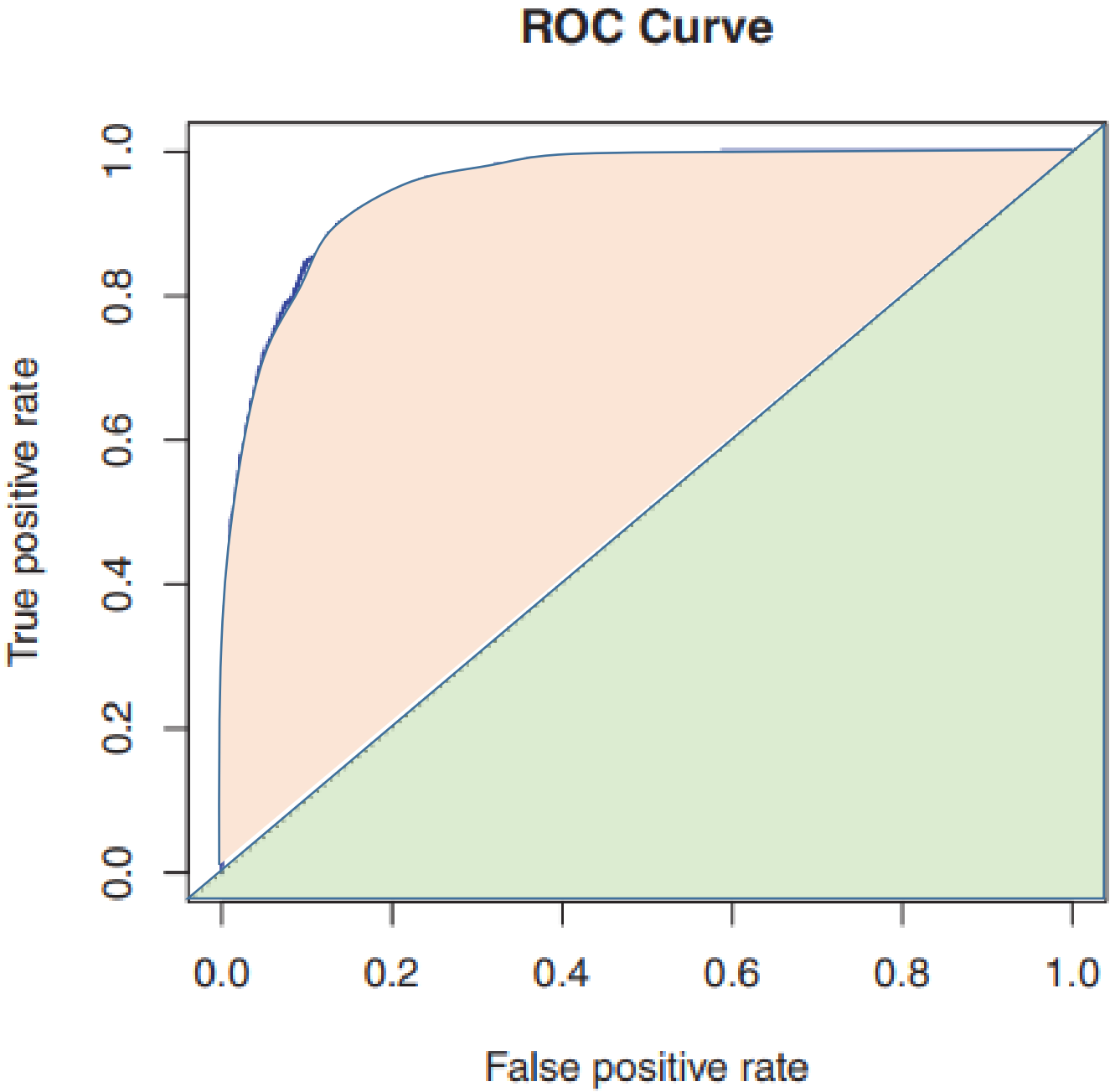
$$FPR = FP/N = FP/(FP + TN) = 1 - TNR$$

False Discovery Rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Miss Rate or False Negative Rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

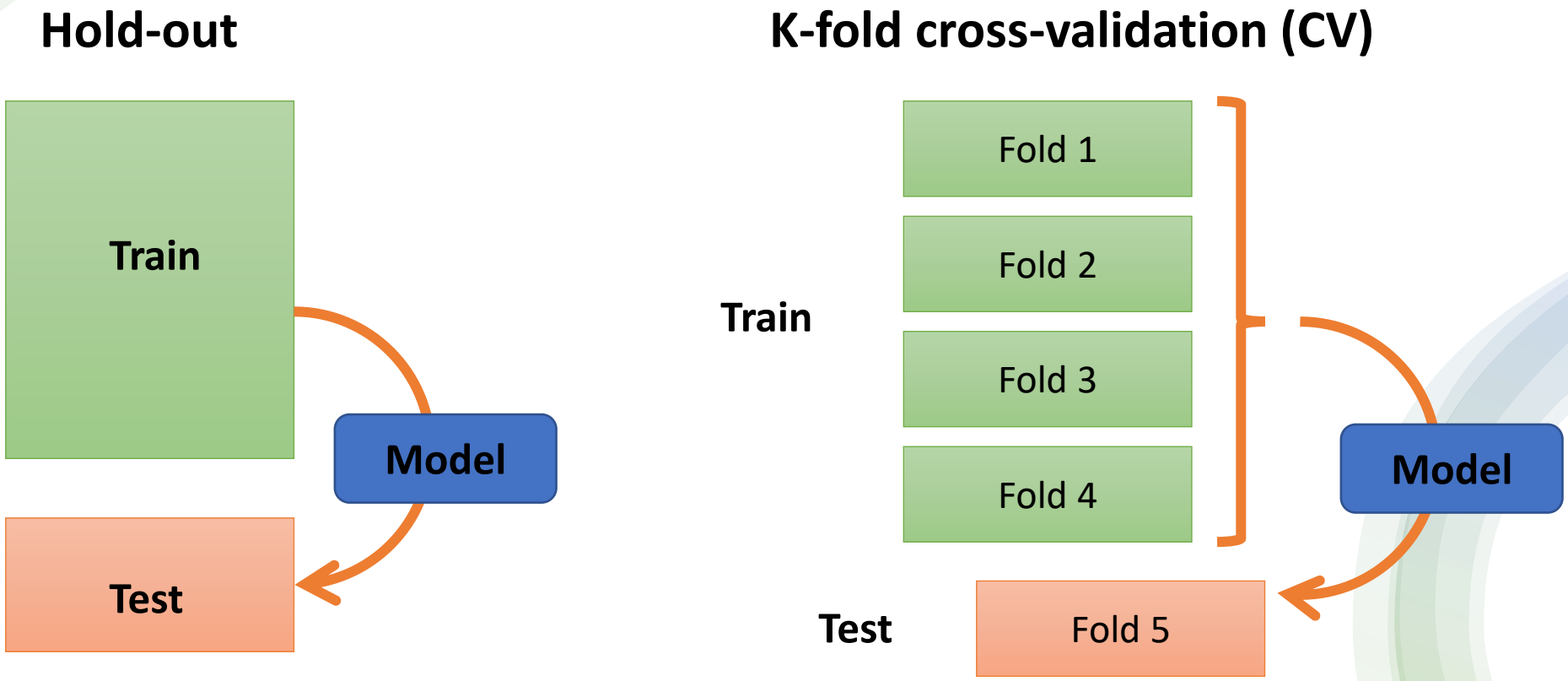


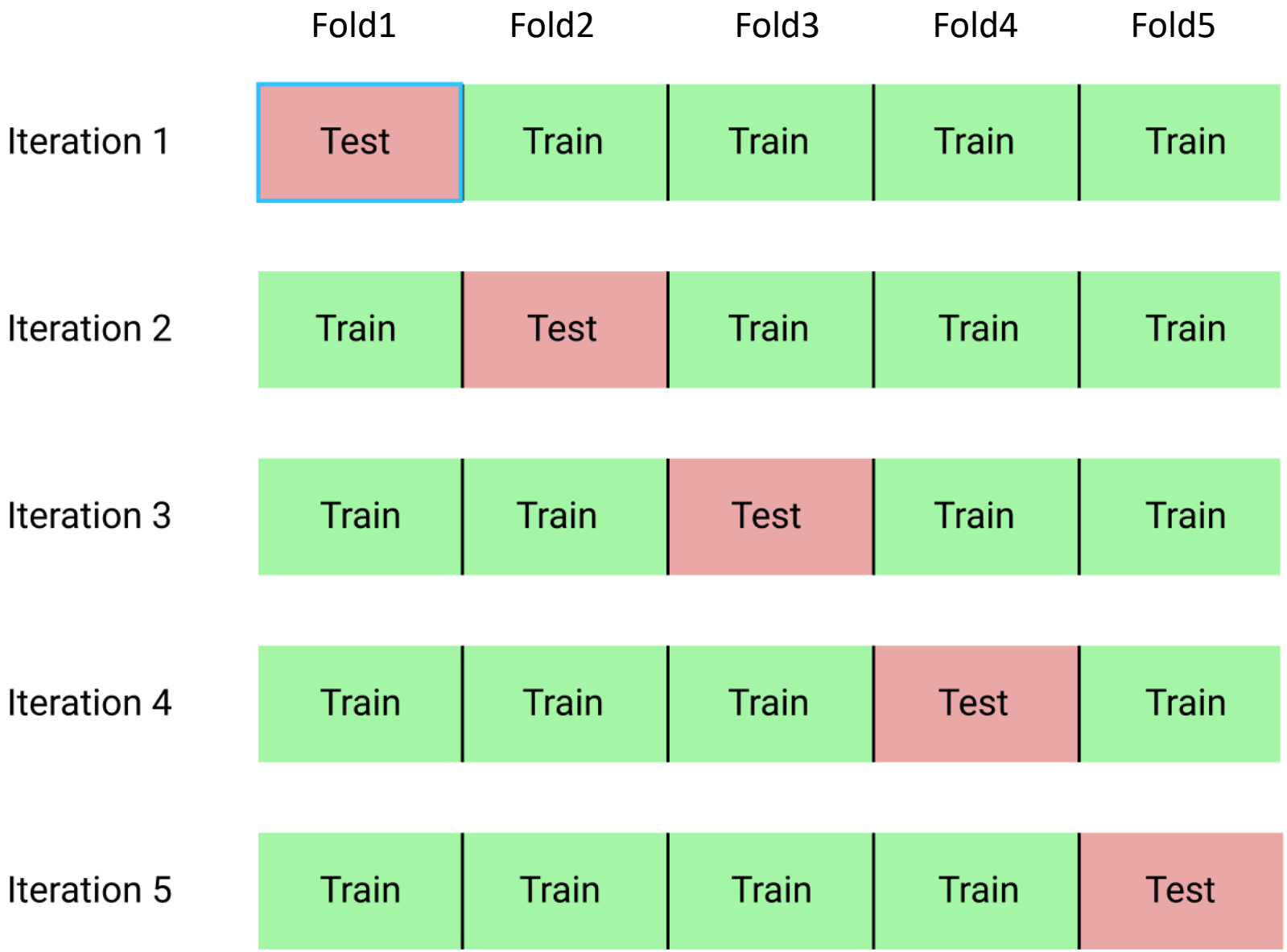
Area Under the roc Curve (AUC)

Random guess

Better than random guess

Model Evaluation Technique Using Cross-Validation




Fold1
Fold2
Fold3
Fold4
Fold5
Iteration 1
Test
Train
Train
Train
Train
Iteration 2
Train
Test
Train
Train
Train
Iteration 3
Train
Train
Test
Train
Train
Iteration 4
Train
Train
Train
Test
Train
Iteration 5
Train
Train
Train
Train
Test