Model Evaluation (P2)

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn

Content outline

- Machine learning: Basic concepts
- Evaluation strategies

Machine learning

An overview

Machine learning (ML) models

- ML models implement learning algorithms to identify patterns and make predictions or decisions based on **data**.
- It learns from training data and, once trained, can analyze and infer from new data.

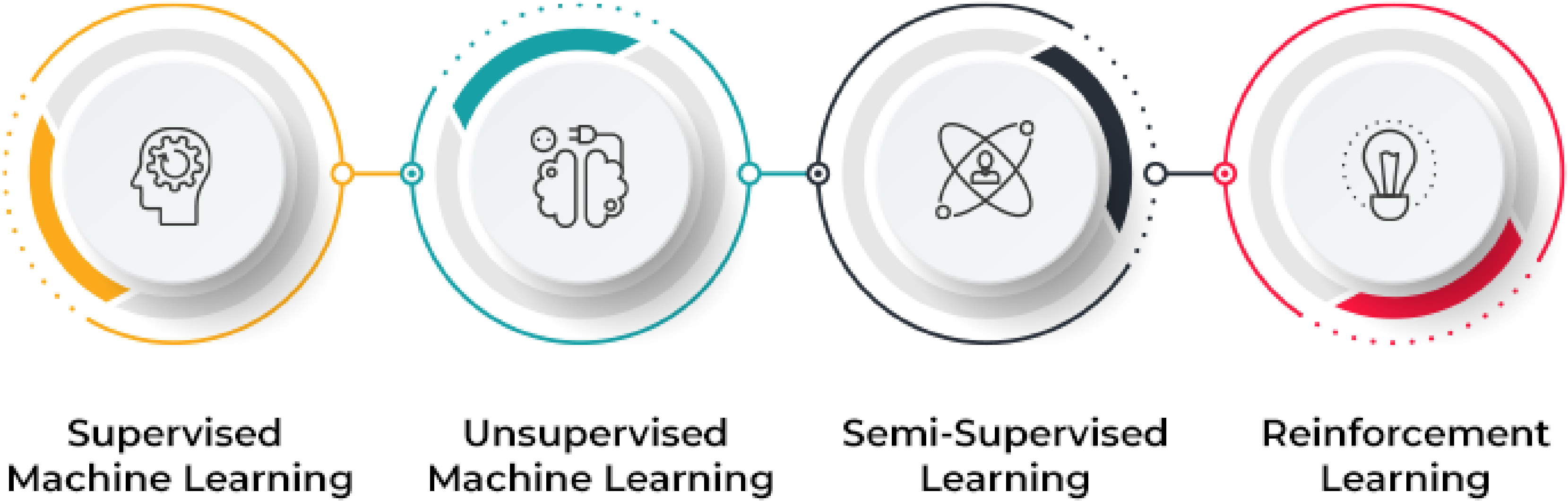
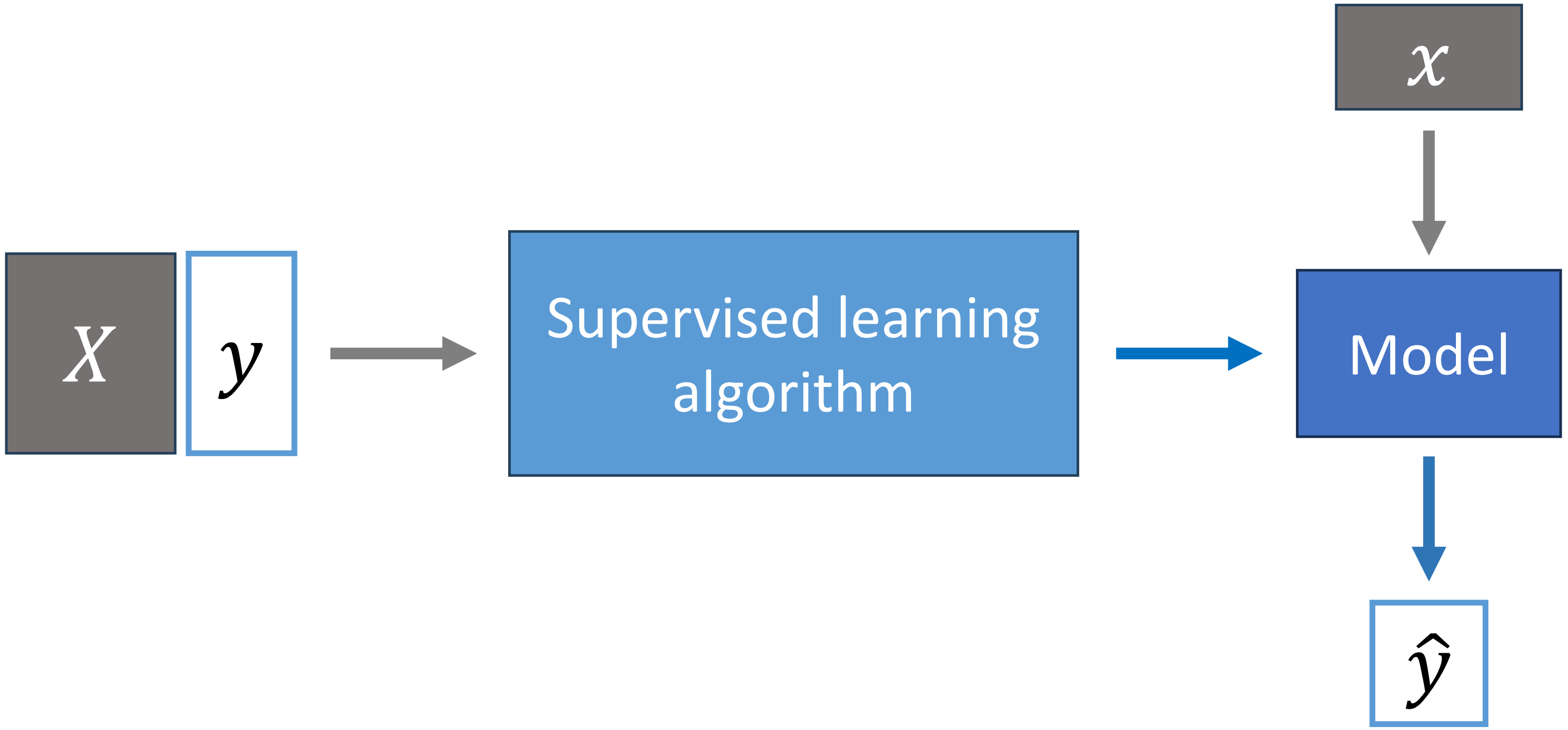


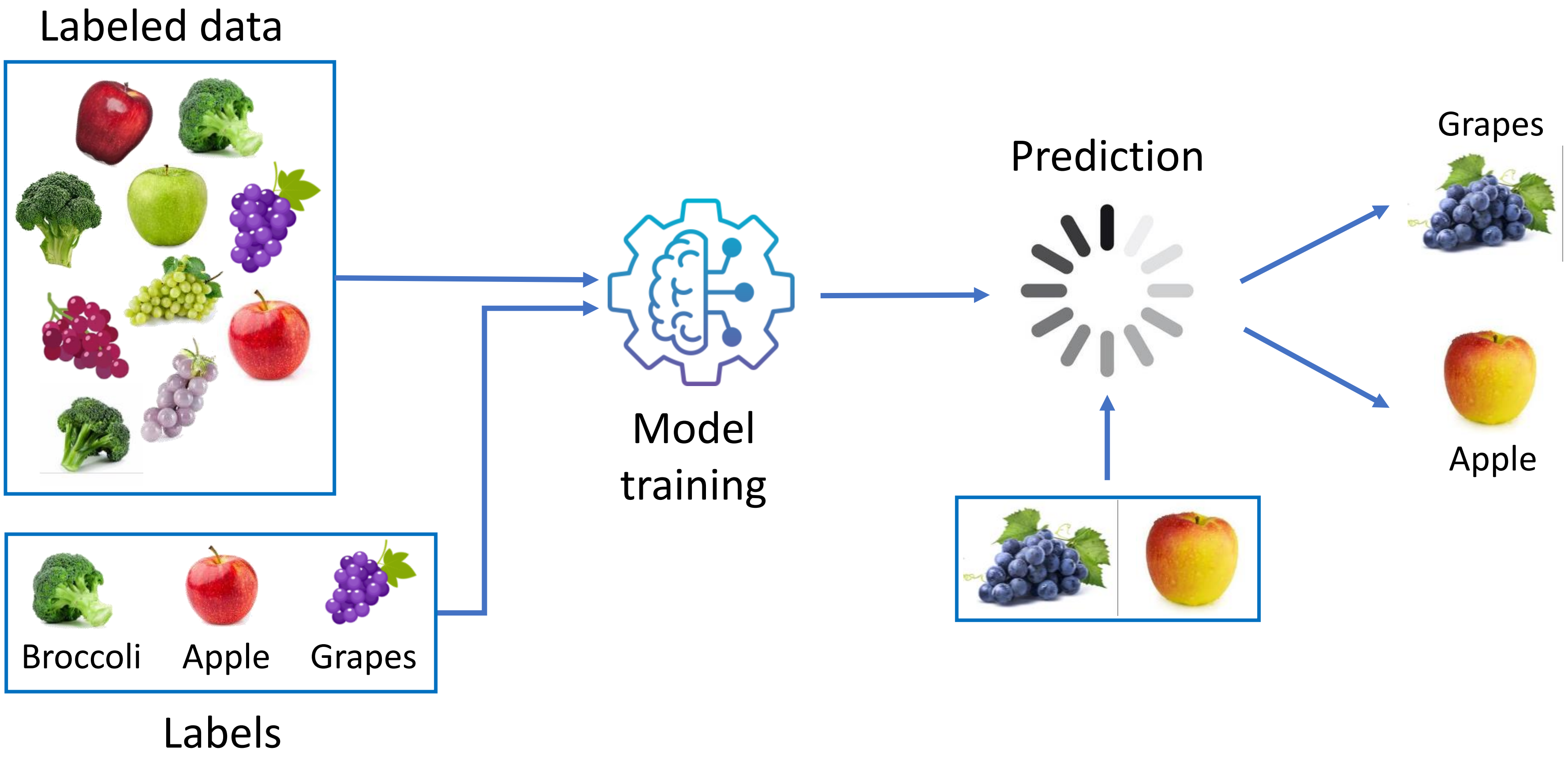
Image credit: Toolbox

Supervised learning (SL)

- A SL model learns from labeled examples to map from inputs to outputs.
- Each example is a pair of input and output values.

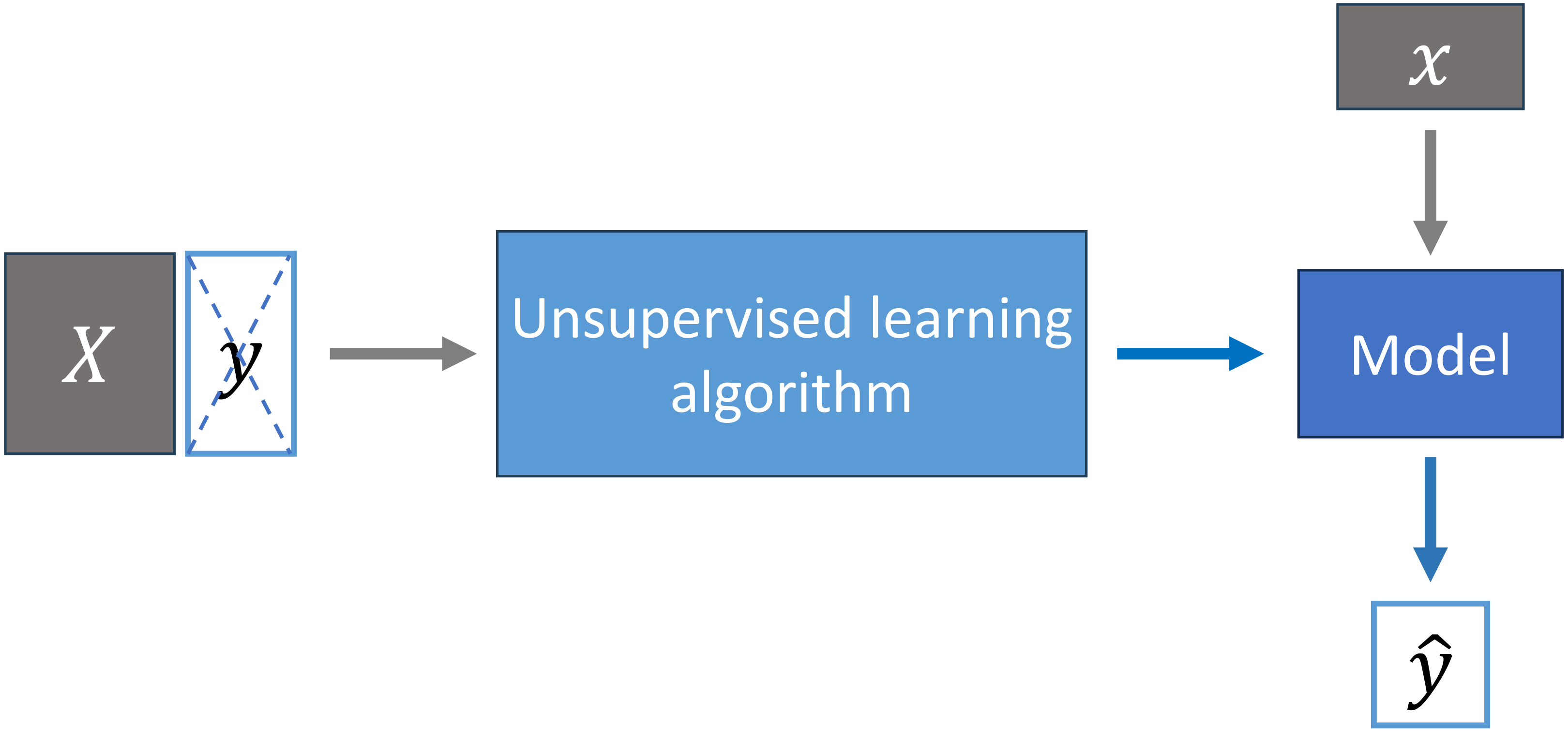


Supervised learning (SL)

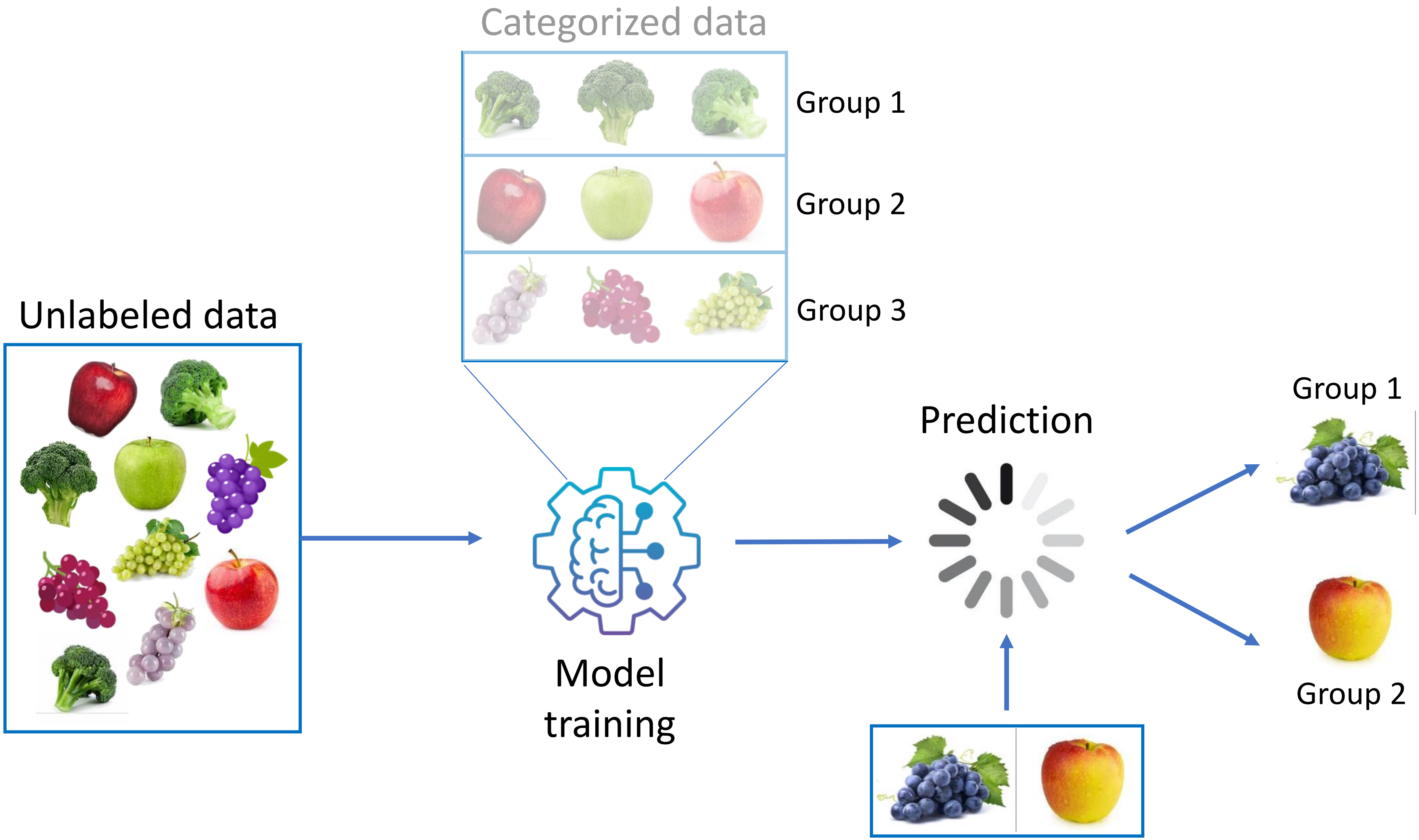


Unsupervised learning (USL)

- A USL model learns from unlabeled examples to describe patterns and insights in the data.



Unsupervised learning (USL)



Categorized data

Group 1

Group 2

Group 3

Unlabeled data

Model training

Prediction

Group 1

Group 2

Reinforcement learning (RL)

- The agent learns from the environment by interacting with it and receives rewards for performing actions.

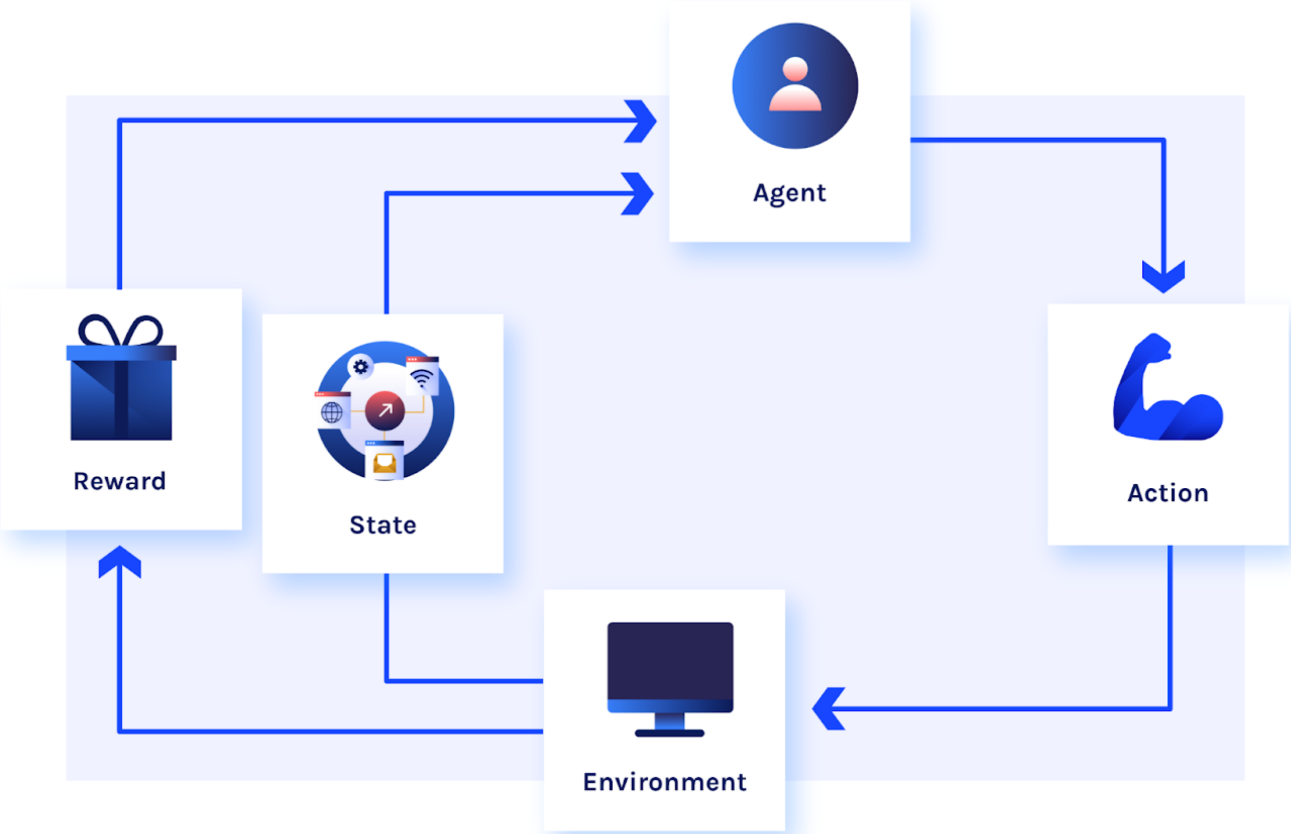


Image credit: Datatonic

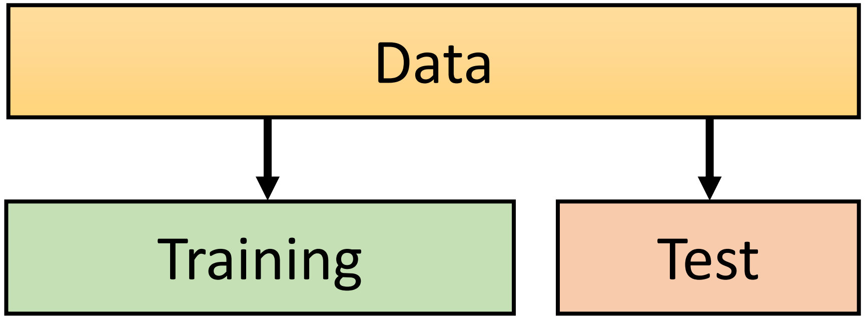
Reinforcement learning (RL)

Video credit: YouTube

Evaluation strategies

Holdout validation

- The data is randomly partitioned into two independent sets.



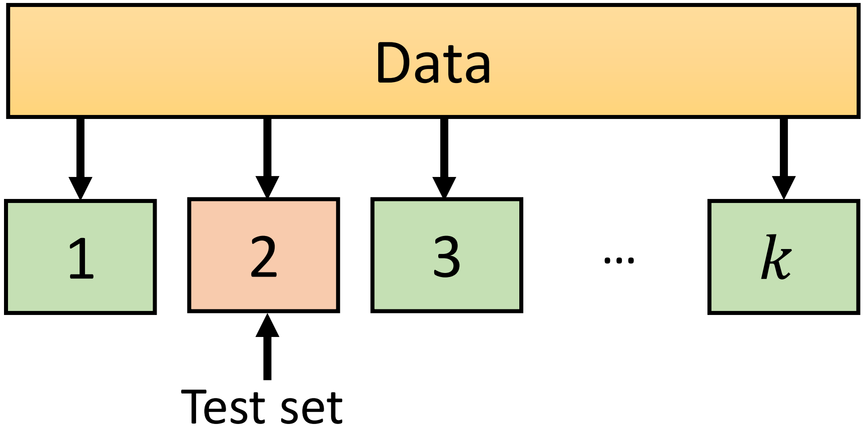
2/3 of data for model construction (training set) and 1/3 of data for accuracy estimation (test set)

- Useful for initial model assessment
- Unreliable if the split fails to reflect the entire data accurately.
  - **Workaround solution:** repeat holdout k times and take the average of the accuracies obtained.

k-fold cross validation

- The data is randomly split into k subsets (or folds, usually $k = 10$) that are mutually exclusive and roughly equal-sized.
- The model trains on all but one subset, which serves as the test set, and this process repeats for each subset.

At i^{th} iteration, use the i^{th} subset as validation set and others as training set.



- Ideal for scenarios with limited data, ensuring the model is rigorously tested under various conditions.

Leave-One-Out cross validation

- The model trains on all data points except one and tests on that excluded point, and this is repeated for each data point.
- Highly effective for small datasets
- Maybe computationally intensive for larger ones

- **Leave-$\boldsymbol{P}$-Out cross validation:** Every possible combination of P data points serves as a test set.

Stratified cross validation

- The class distribution in each fold mirrors that of the original data set.
- Crucial for handling imbalanced data, where some classes are underrepresented

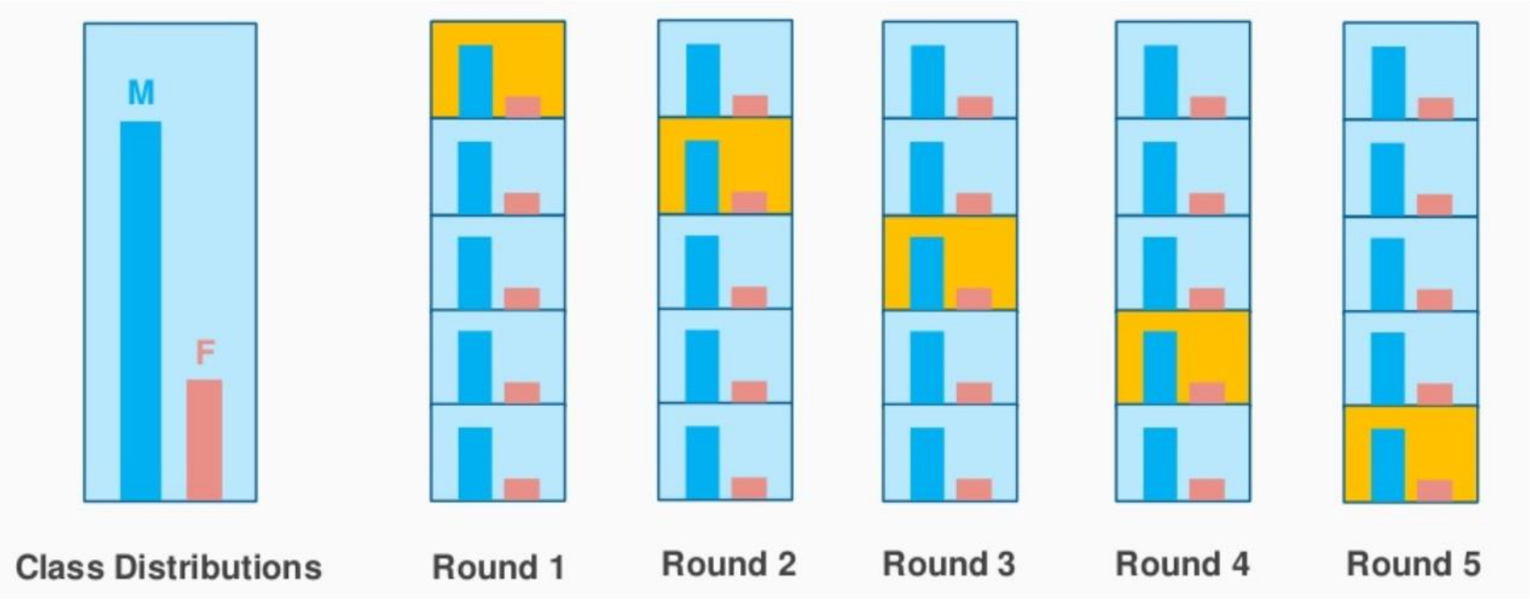
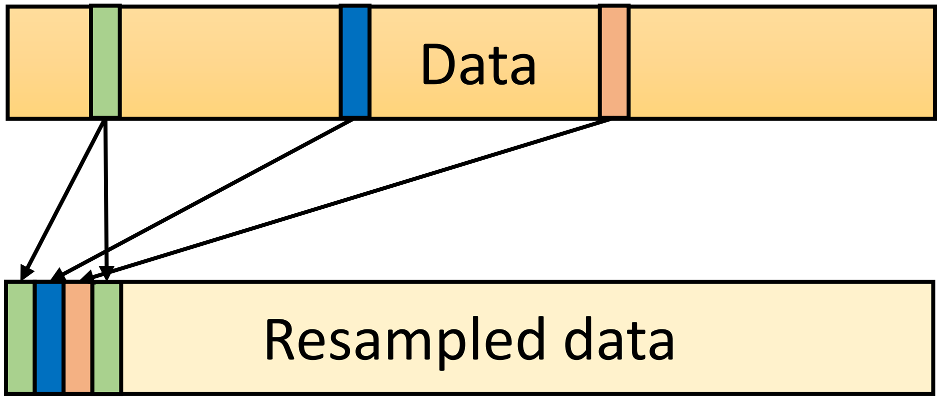


Image credit: Stack Exchange

Bootstrap

- Training tuples are sampled uniformly with replacement
- Work well with small data, yet tend to be overly optimistic



- .632 bootstrap is the **most common** method.

.632 Bootstrap

- Consider a dataset with d tuples.
- Repeat the following sampling procedure k times
  - Sample d times to obtain a training set of d samples.
  - Tuples not in the training set form the validation set.
  - About 63.2% of the original data end up in the bootstrap, and the remaining 36.8% form the validation set (since $\left(1 - \frac{1}{d}\right)^d \approx e^{-1} = 0.368$ for large d)
- The overall accuracy of the model is

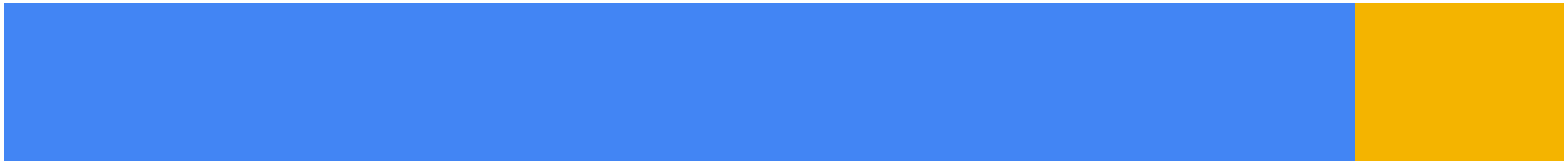
$$Acc(M) = \frac{1}{k}\sum_{i=1}^{k}(0.632 \times Acc(M_i)_{val\_set} + 0.368 \times Acc(M_i)_{train\_set})$$

  - where $Acc(M_i)_{train\_set}$ and $Acc(M_i)_{val\_set}$ are the accuracies obtained with bootstrap sample i when the model is applied to the training set and val set.

Splitting into data sets

Training set and Test set

- We commonly split the original data into two subsets: training set and test set.



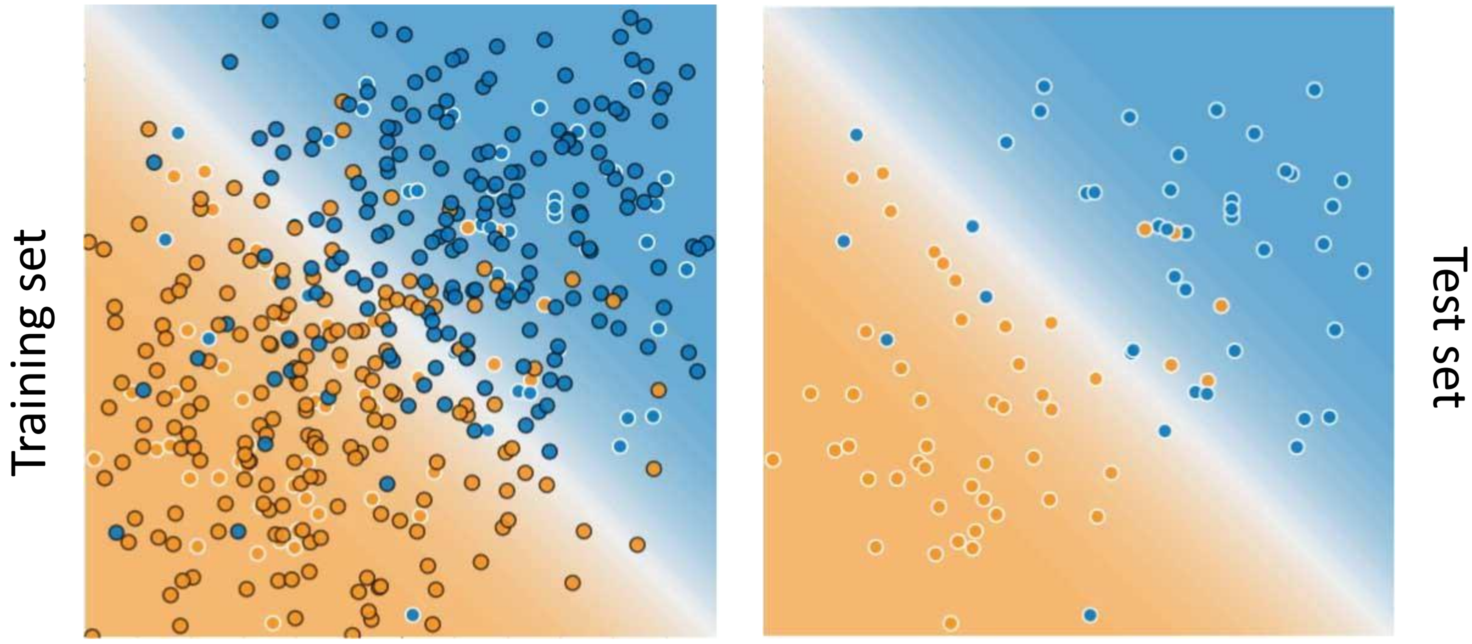
Training set:
a subset to train a model

Test set:
a subset to test the trained model

- The test set must meet the following conditions:
  - Large enough to yield statistically meaningful results
  - Be a representative of the entire data set

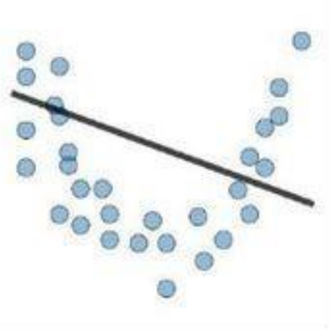
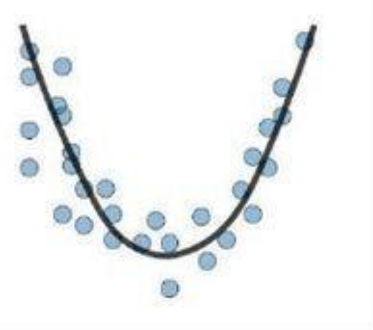
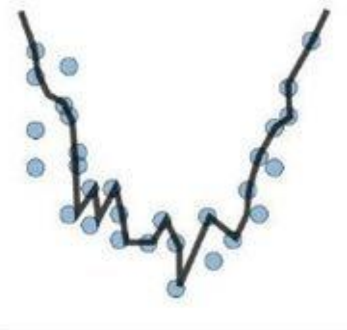
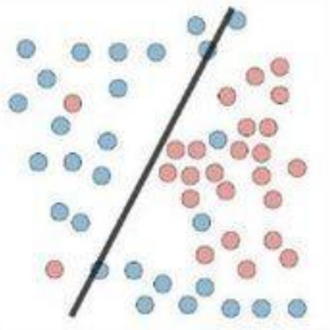
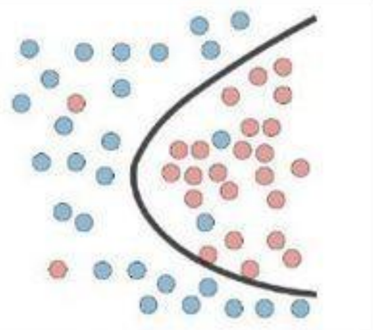
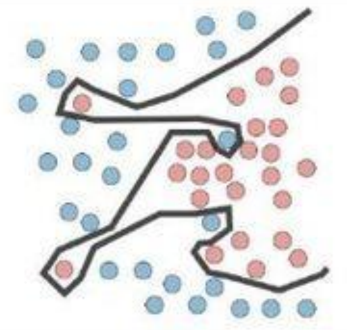
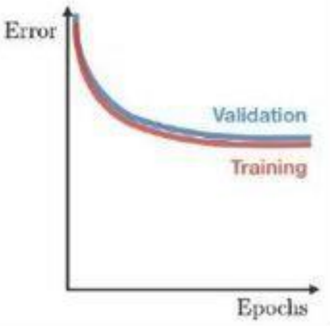
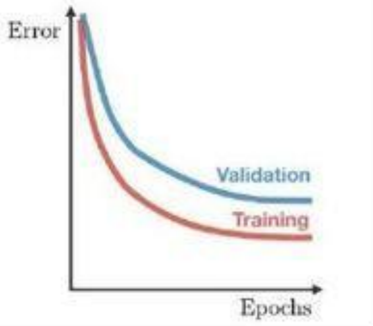
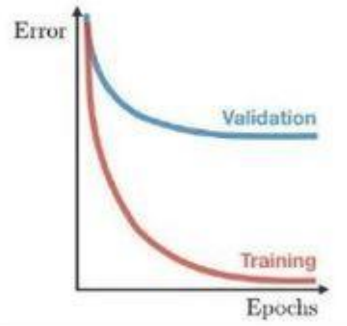
Validate the model against test set

- The test set serves as a proxy for new data, to which the model should generalize well.



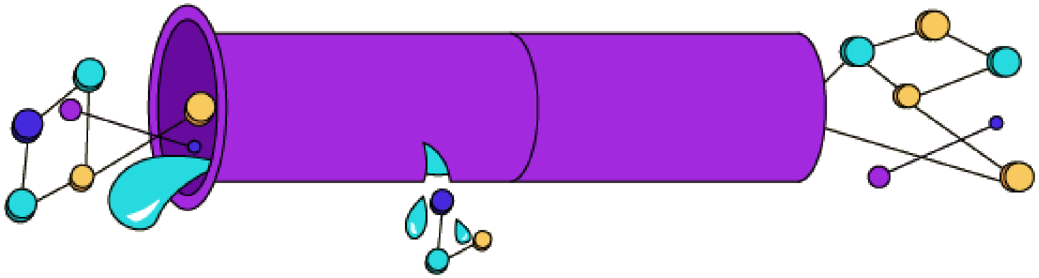
The model learned for the training data is very simple. This model makes some wrong predictions but performs similarly on both training and test data.

Overfitting vs. Underfitting

	Underfitting	**Just right**	**Overfitting**
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

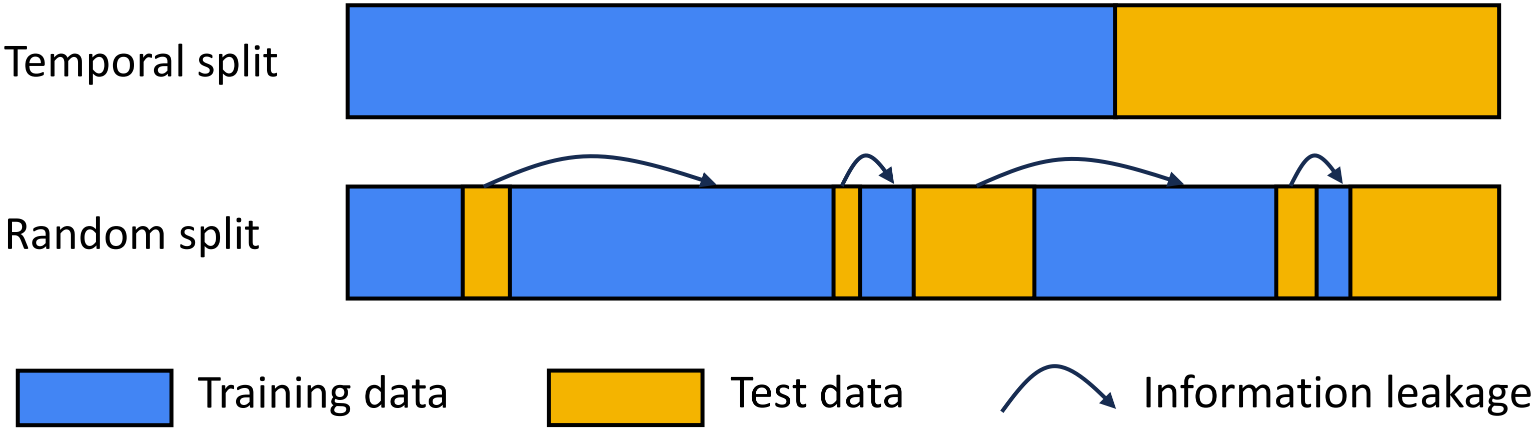
Data leakage: Never train on test data

- Data leakage occurs when external information not available during training is used to build the model.
- Artificially high performance during training and evaluation
- Poor generalization in practical use



Data leakage: Example causes

- Leaking target information: Include features that directly or indirectly reveal the target variable.
- Overlapping data: Share data between training and test sets
- Temporal leakage: Using future information to predict past or current events in time-series data.

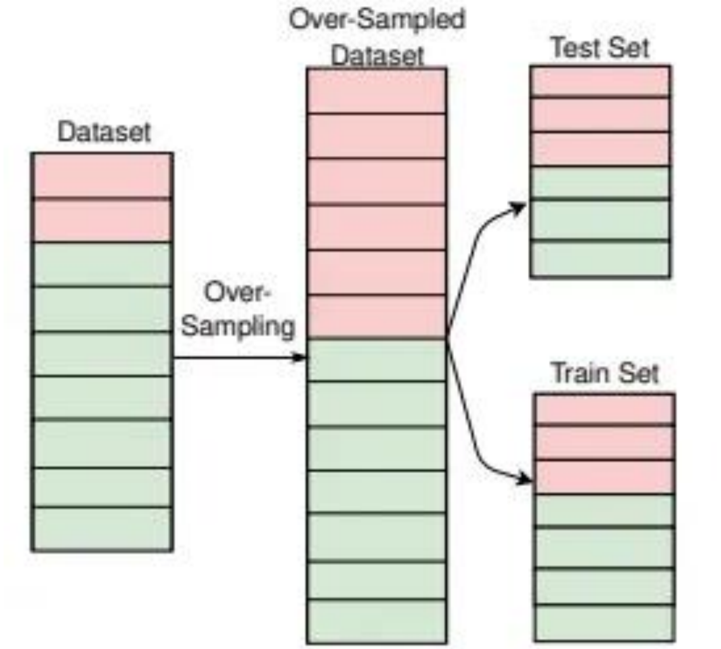


Information leakage in temporal data

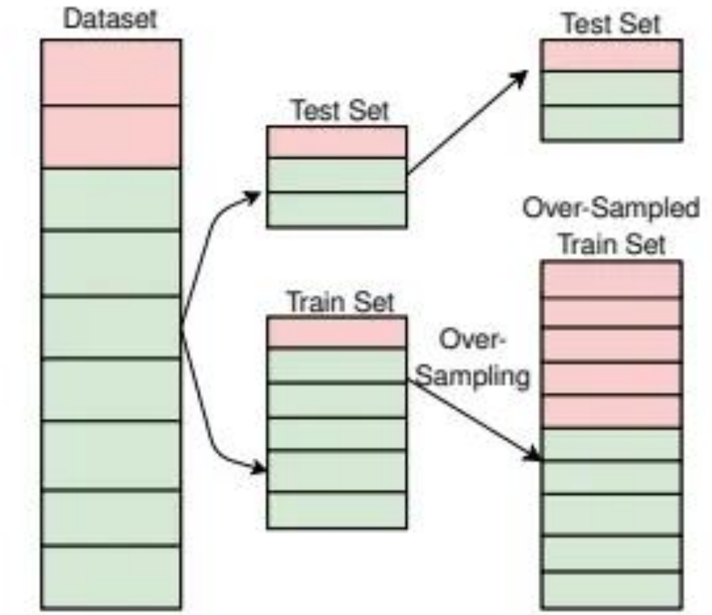
Quiz 01: Oversampling and Data leakage

Consider the following strategies, each of which perform train-test split and oversampling in different orders.

Which strategies is more likely exposed to data leakage?



Oversampling before train-test split



Oversampling only on train data after train-test split

Which split ratio is best?

- Less training data increases parameter variance, while less testing data raises performance variance.

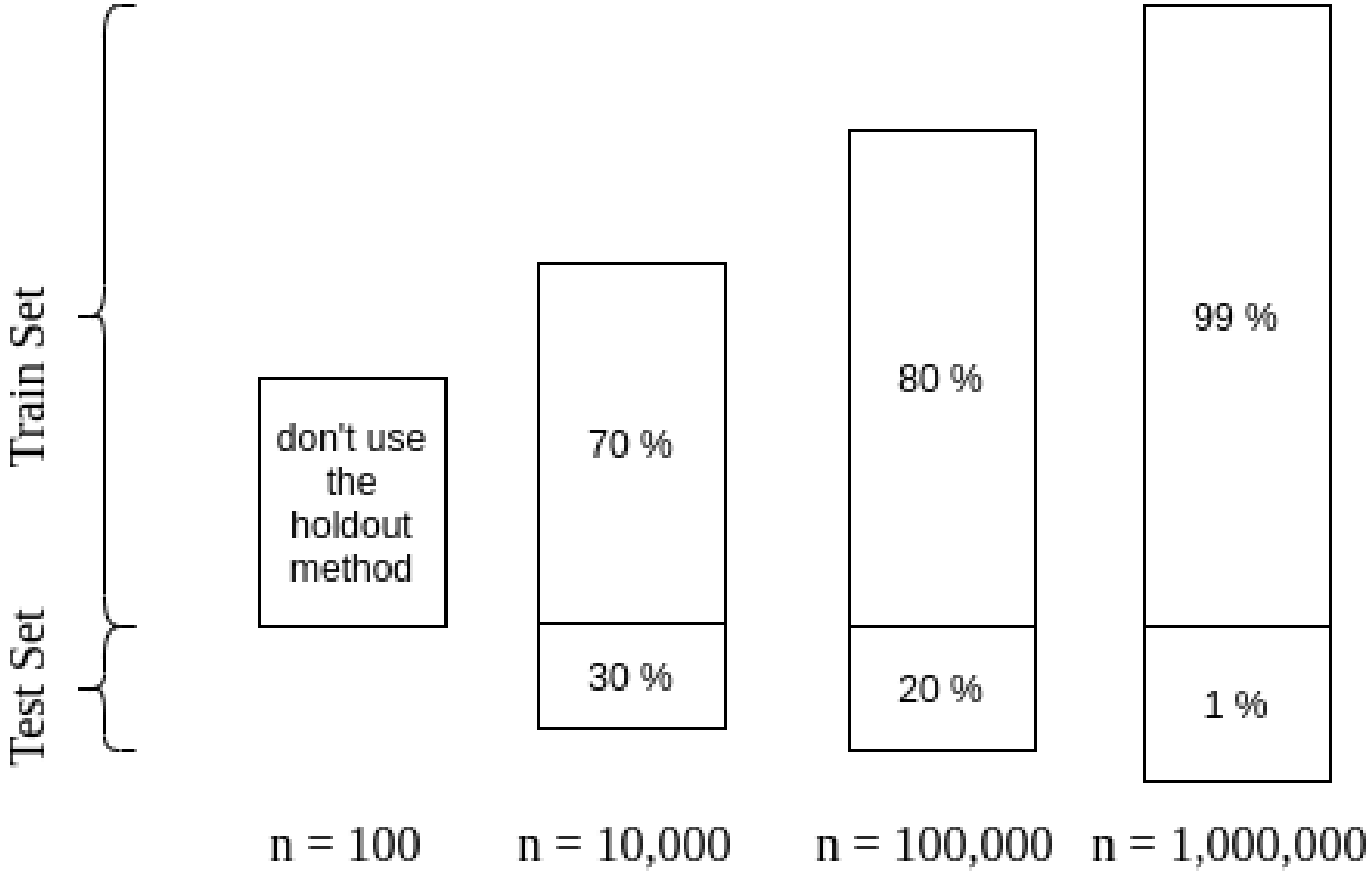
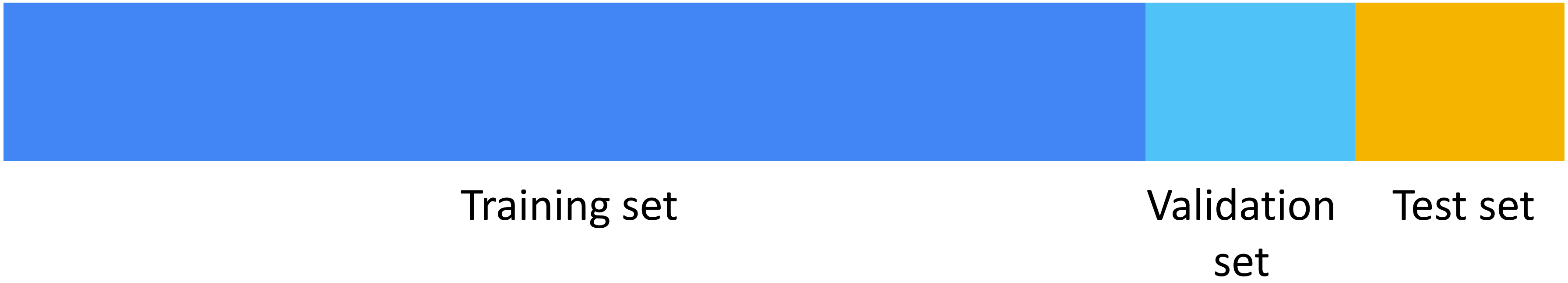


Image credit: Baeldung

Validation set

- Partitioning the original data into three subsets, training set, validation set, and test set can reduce the risk of overfitting.



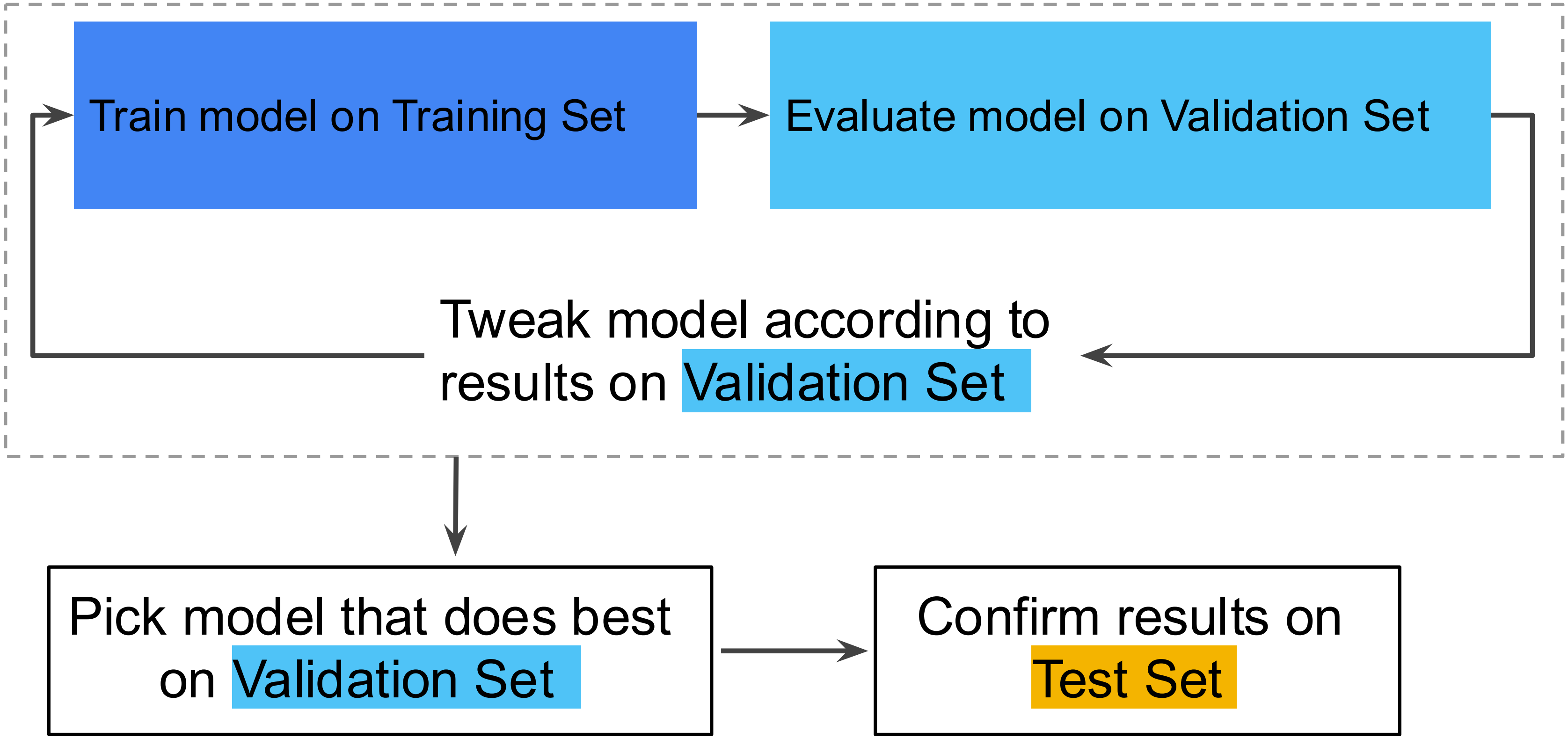
Training set

Validation set

Test set

- The validation set evaluates the trained model and helps fine-tuning parameters.
- The test set is now for double-checking after the model has "passed" the validation set.

Model evaluation workflow

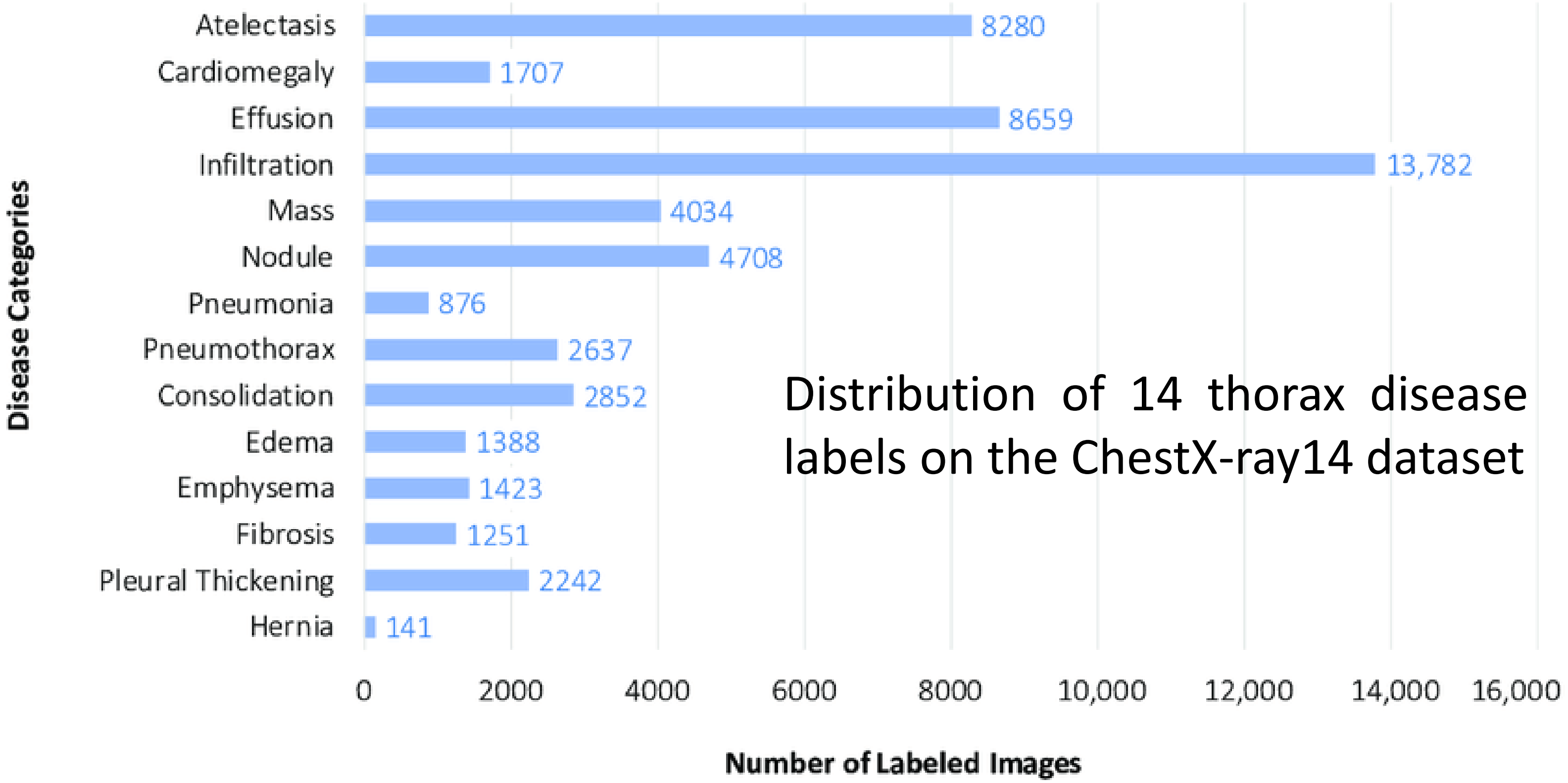

Train model on Training Set
Evaluate model on Validation Set
Tweak model according to results on Validation Set
Pick model that does best on Validation Set
Confirm results on Test Set

- Pick the model that does best on the validation set.
- Double-check that model against the test set.
- This workflow creates fewer exposures to the test set.

Stretching your data

Class imbalance problem

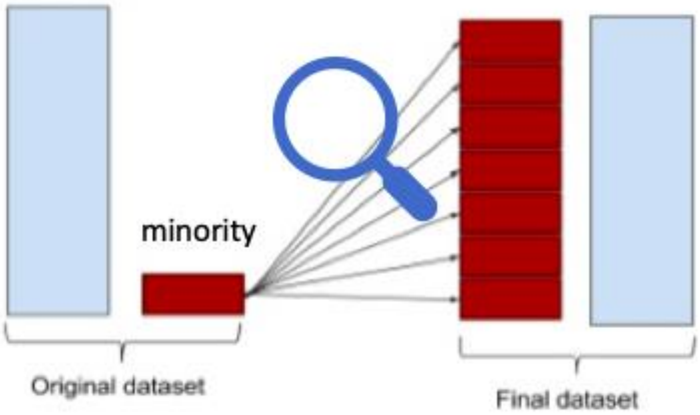
- One or more classes have very low proportions in the data as compared to the other classes.
    - E.g., medical diagnosis, anomaly detection, fraud detection, etc.

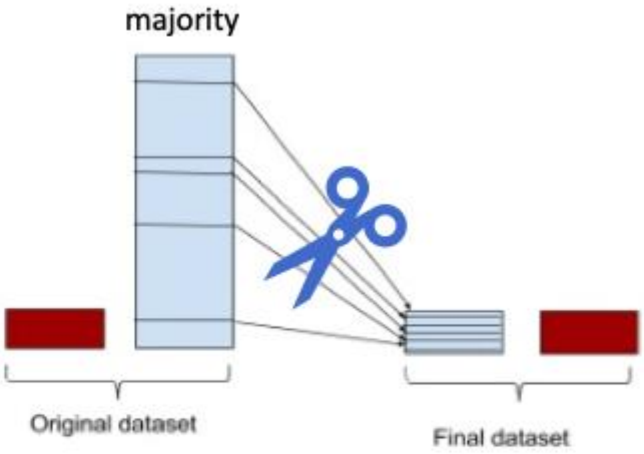


Distribution of 14 thorax disease labels on the ChestX-ray14 dataset

Class imbalance problem: Solutions

- Oversampling: re-sample the data from the minority classes
- Undersampling: randomly eliminate tuples from the majority classes

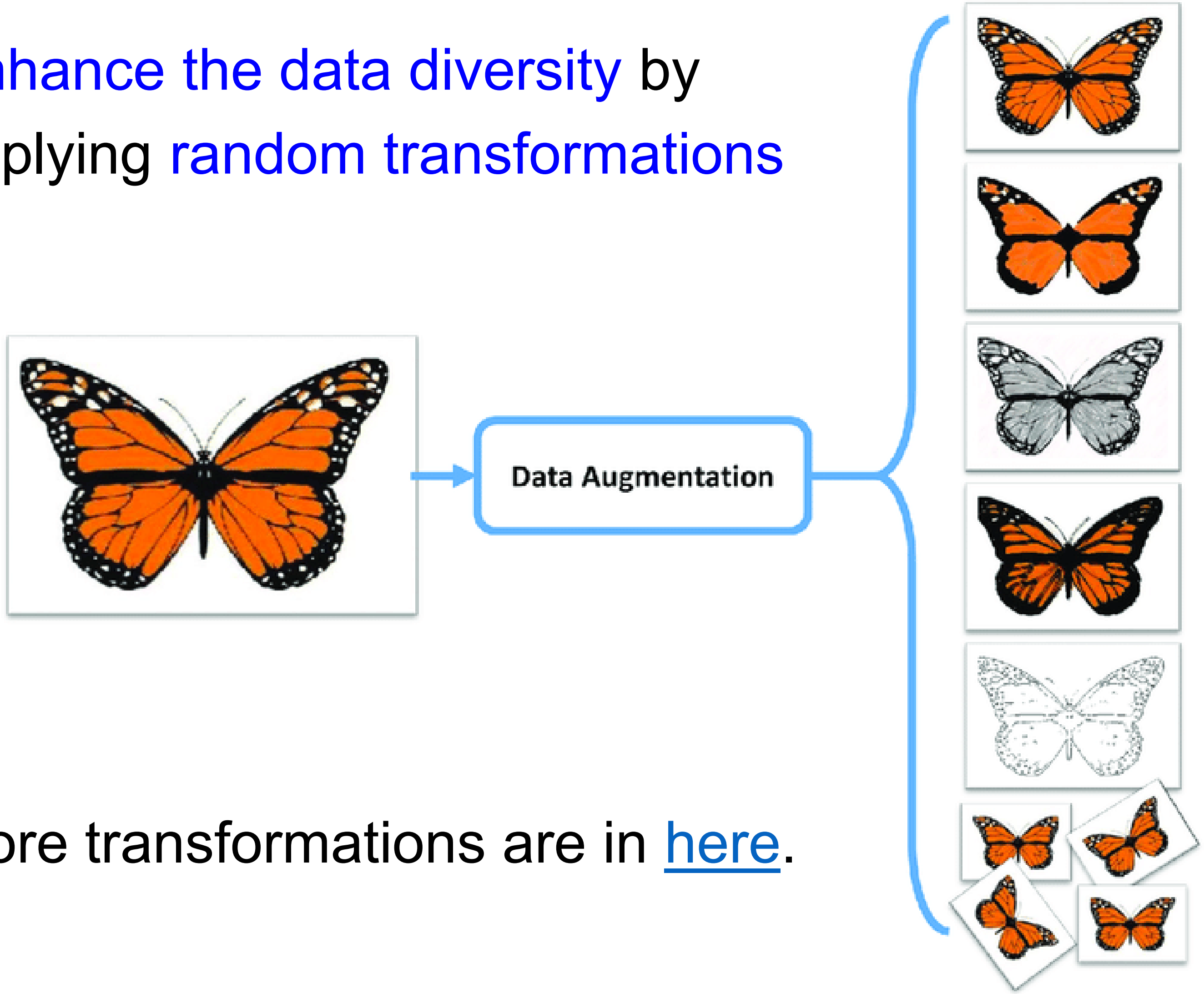




- Threshold-moving: move the decision threshold so that the rare class tuples are easier to classify → less chance of costly false negative errors
- Ensemble learning: multiple classifiers works together to vote

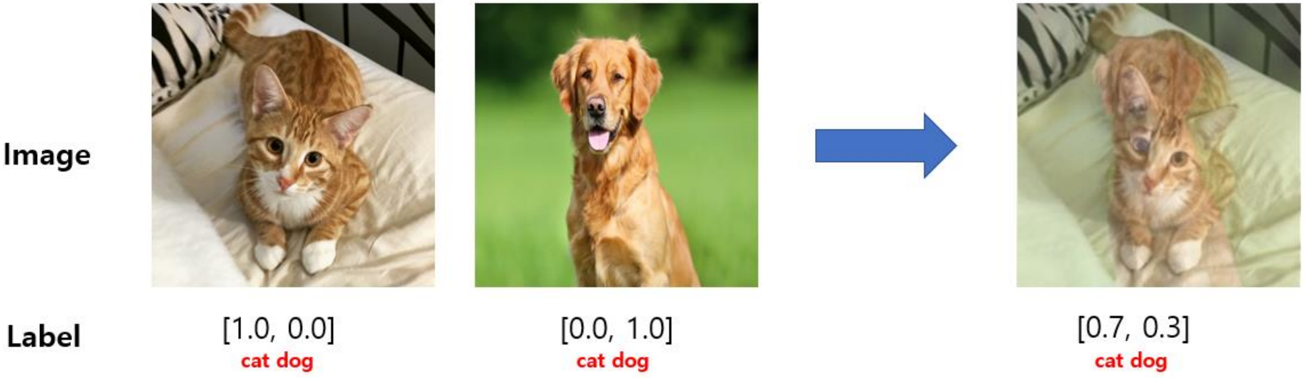
Data augmentation

- Enhance the data diversity by applying random transformations



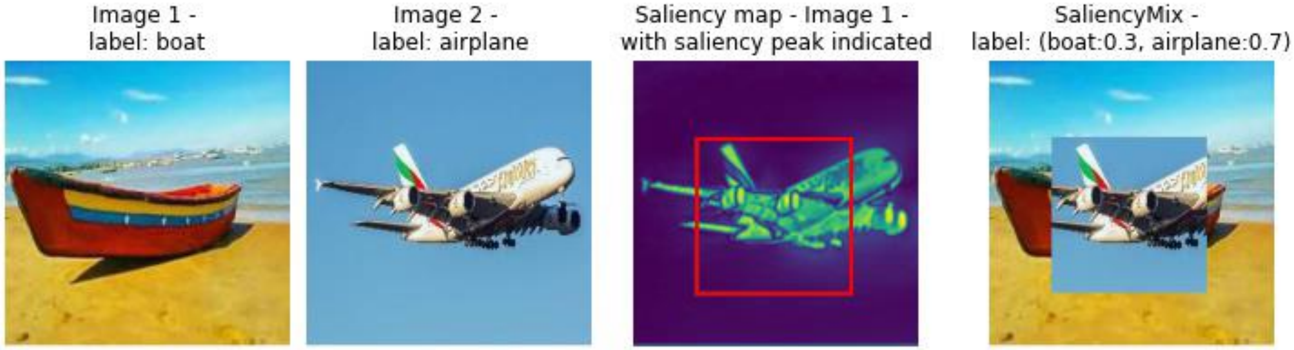
- More transformations are in here.

Data augmentation: Images



Mixup

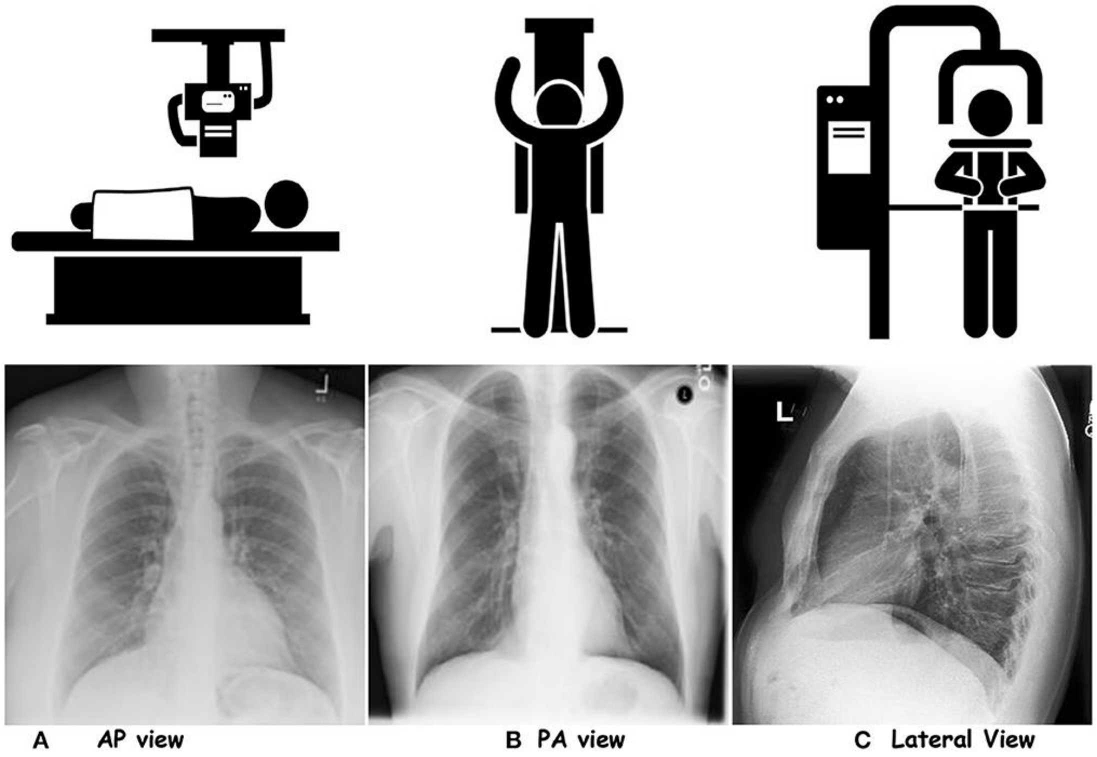
SaliencyMix



Shahab Uddin et al. “SaliencyMix: A saliency guided data augmentation strategy for better regularization,” in *ICLR*, 2021.
H. Zhang et al., “mixup: Beyondempirical risk minimization,” in *ICLR*, 2018.

Data augmentation: Medical images

- The augmented data must preserve the contextual validity.



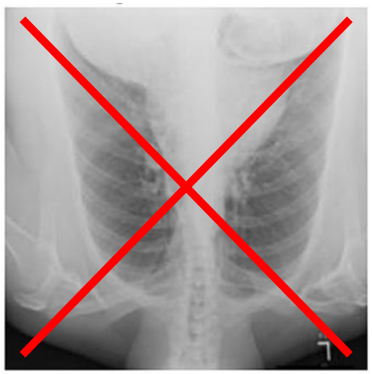


Image credit: frontiers

Data augmentation: Text data

- Synonym replacement

"The cat sat on the mat."
→ "The feline rested on the mat."

- Random deletion

"The quick brown fox jumps over the lazy dog.“
→ "The fox jumps over the dog."

- Back-translation

"The weather is nice today.“
→ (to French: "Le temps est beau aujourd'hui")
→ "Today's weather is pleasant."

- Contextual word embedding replacements

"The car is fast." → "The vehicle is speedy."

Data augmentation: Graph data

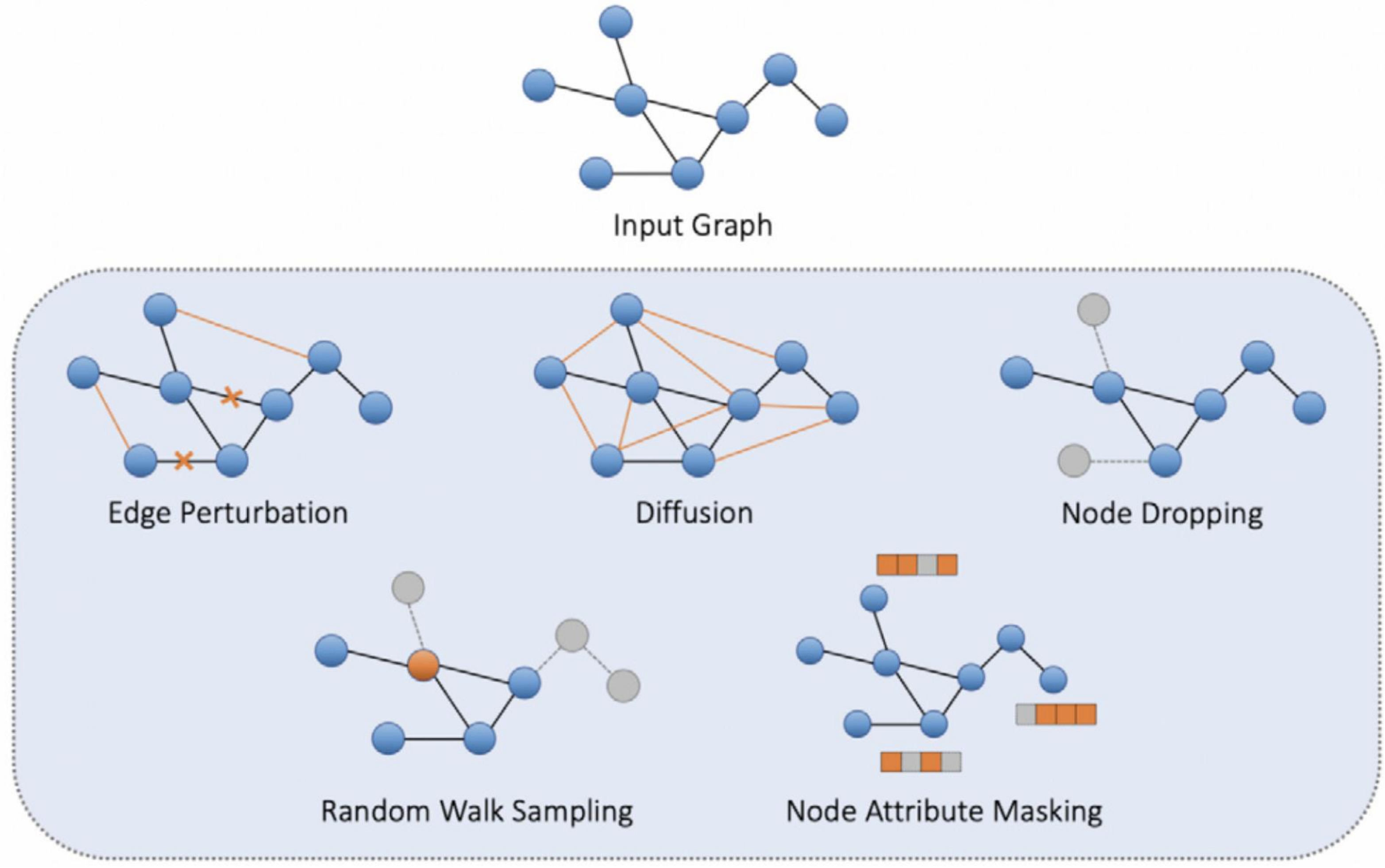


Image credit: ScienceDirect

Quiz 02: Data augmentation

- Present a data augmentation method used in computer vision or natural language processing.

...the end.