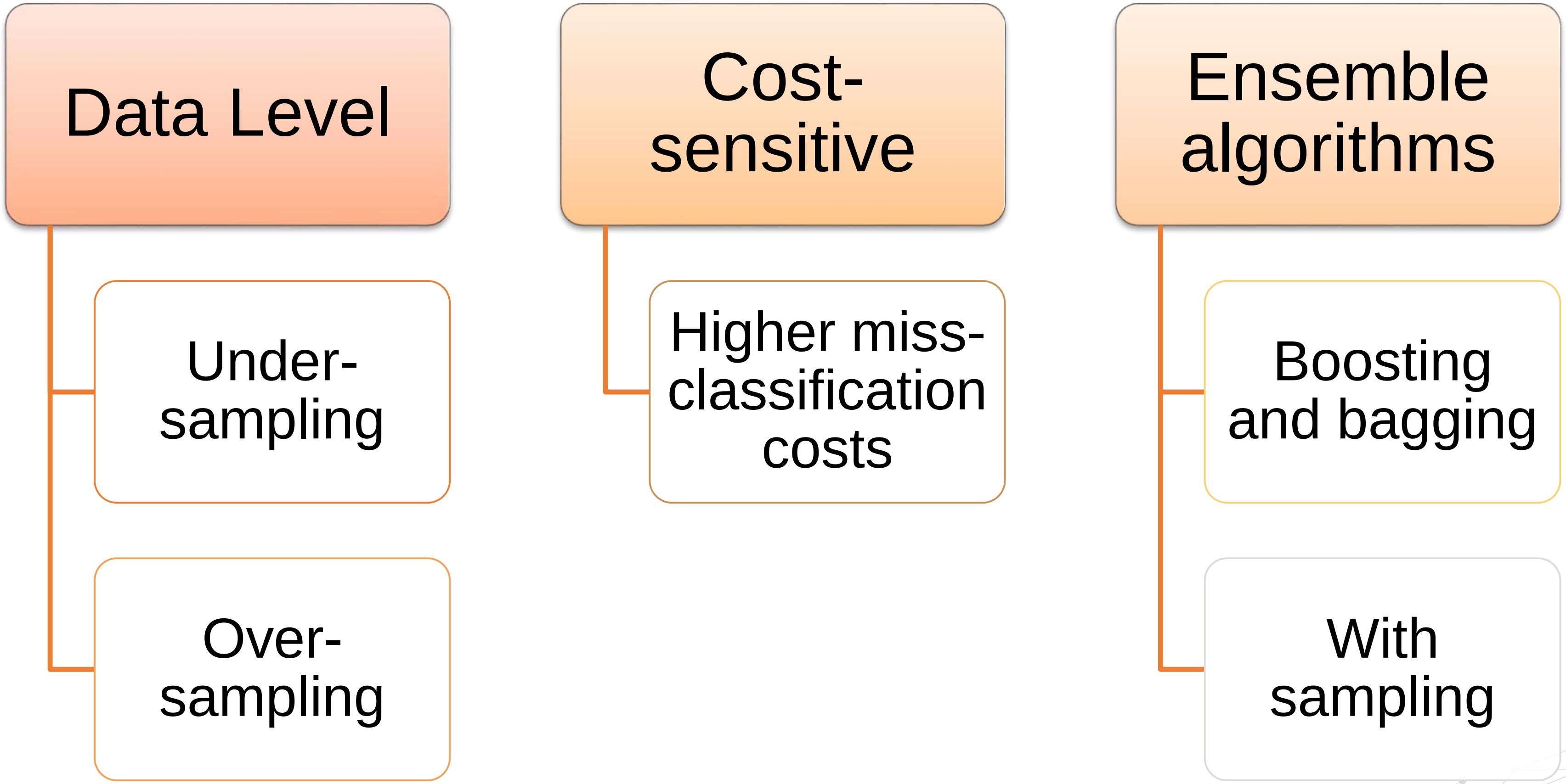
Train In Data

Solutions for Imbalanced Datasets

Solutions for Imbalanced Datasets



- Data Level
  - Under-sampling
  - Over-sampling
- Cost-sensitive
  - Higher miss-classification costs
- Ensemble algorithms
  - Boosting and bagging
  - With sampling

Data-level approaches

Changing the distribution of the data.

- Random Over- or Under-sampling
- Creating new synthetic data
- Removing noise or alternatively, removing easy observations to classify

Cost-sensitive approaches

Different cost to different errors.

The cost of misclassifying an instance of the minority class outweighs the cost of misclassifying an instance from the majority.

The cost-sensitive learning process seeks to minimize the cost error.

Ensemble approaches

Combine weak learners

Construct multiple classifiers from the original data and then aggregate their predictions.

Combining classifiers generally improves their generalization ability