UNIVERSIDADE D
COIMBRA

IPOPORTO CISUC COGNITIVE & MEDIA SYSTEMS

DSPT
DATA SCIENCE PORTUGAL

Coimbra, February 13, 2019

Dealing with Imbalanced Data

the nuts and bolts

Miriam Seoane Santos
CISUC, DEI/FCTUC, University of Coimbra
IPO-Porto Research Centre (CI-IPOP),
Porto

IEEE Computational Intelligence Magazine

Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches

Research Frontier

Abstract

Although cross-validation is a standard procedure for performance evaluation, its joint application with oversampling remains an open question for researchers farther from the imbalanced data topic. A frequent experimental flaw is the application of oversampling algorithms to the entire dataset, resulting in biased models and overly-optimistic estimates. We emphasize and distinguish overoptimism from overfitting, showing that the former is associated with the cross-validation procedure, while the latter is influenced by the chosen oversampling algorithm. Furthermore, we perform a thorough empirical comparison of well-established oversampling algorithms, supported by a data complexity analysis. The best oversampling techniques seem to possess three key characteristics: use of cleaning procedures, cluster-based example synthetization and adaptive weighting of minority examples, where Synthetic Minority Oversampling Technique coupled with Tomek Links and Majority Weighted Minority Oversampling Technique stand out, being capable of increasing the discriminative power of data.

I. Introduction

Imbalanced Data (ID) occurs when there is a considerable difference between the

IMAGE LICENSED BY INGRAM PUBLISHING

class priors of a given problem. Considering a binary classification problem, a dataset is said to be imbalanced if there exists an under-represented concept (a minority class) when compared to the other (a majority class) [1]. Prediction models built from imbalanced datasets are most often biased towards the majority concept, which is especially critical when there is a higher cost of misclassifying the minority examples, such as diagnosing rare diseases [2].

Approaches to handle imbalanced scenarios can be mainly divided into data-level approaches, where the data is preprocessed in order to achieve a balanced dataset for classification, and algorithmic-level approaches, where the classifiers are adapted to deal with the characteristic issues of imbalanced data [3–6]. By far, data-level approaches are the most commonly used, as they have proven to be efficient, are simple to implement and completely classifier-independent [2], [7]. Data-level strategies fall into two main categories, undersampling and oversampling: the former consists in removing majority examples while the latter replicates the minority examples. Researchers often invest in oversampling procedures since they are capable

Digital Object Identifier 10.1109/MCI.2018.2866730
Date of publication: 15 October 2018

Corresponding Author: Pedro Abreu (pha@dei.uc.pt).

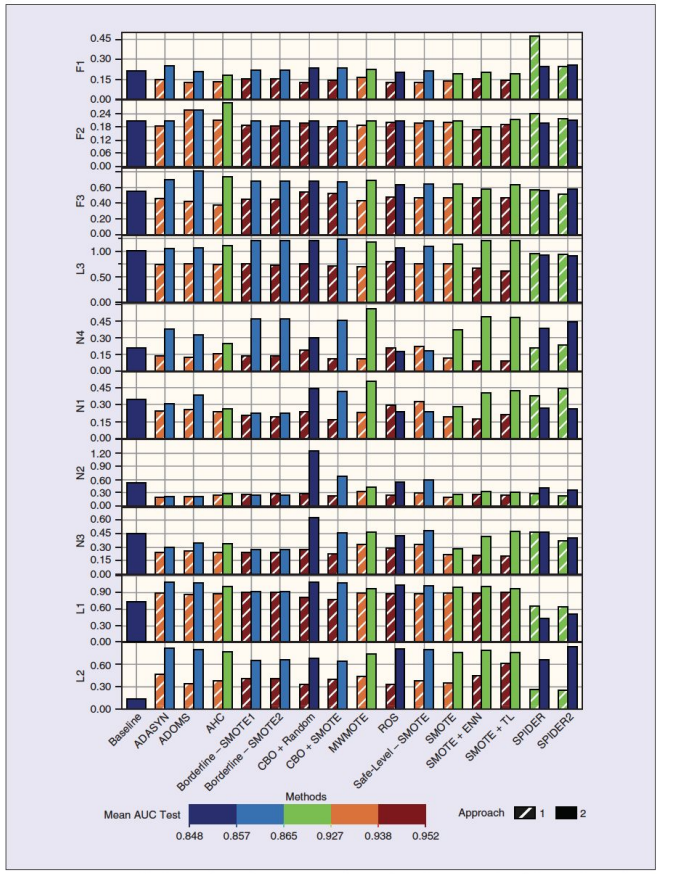


FIGURE 4 Differences (in module) between the complexity measures for all oversampling techniques, considering both Approaches 1 and 2.

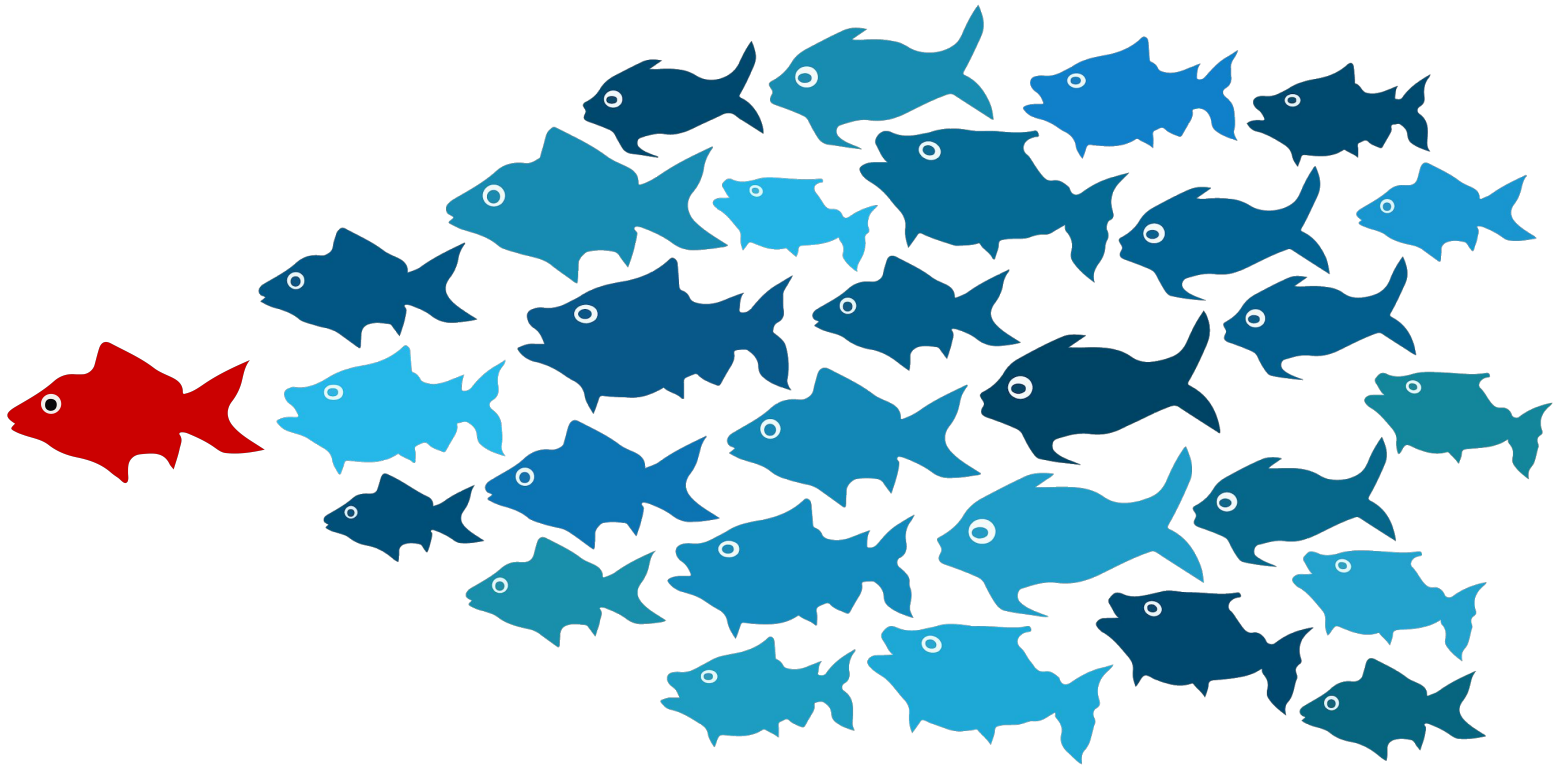
Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Computational Intelligence Magazine*, *13*(4), 59-76.

What

What

What is Imbalanced Data?

What is Imbalanced Data?

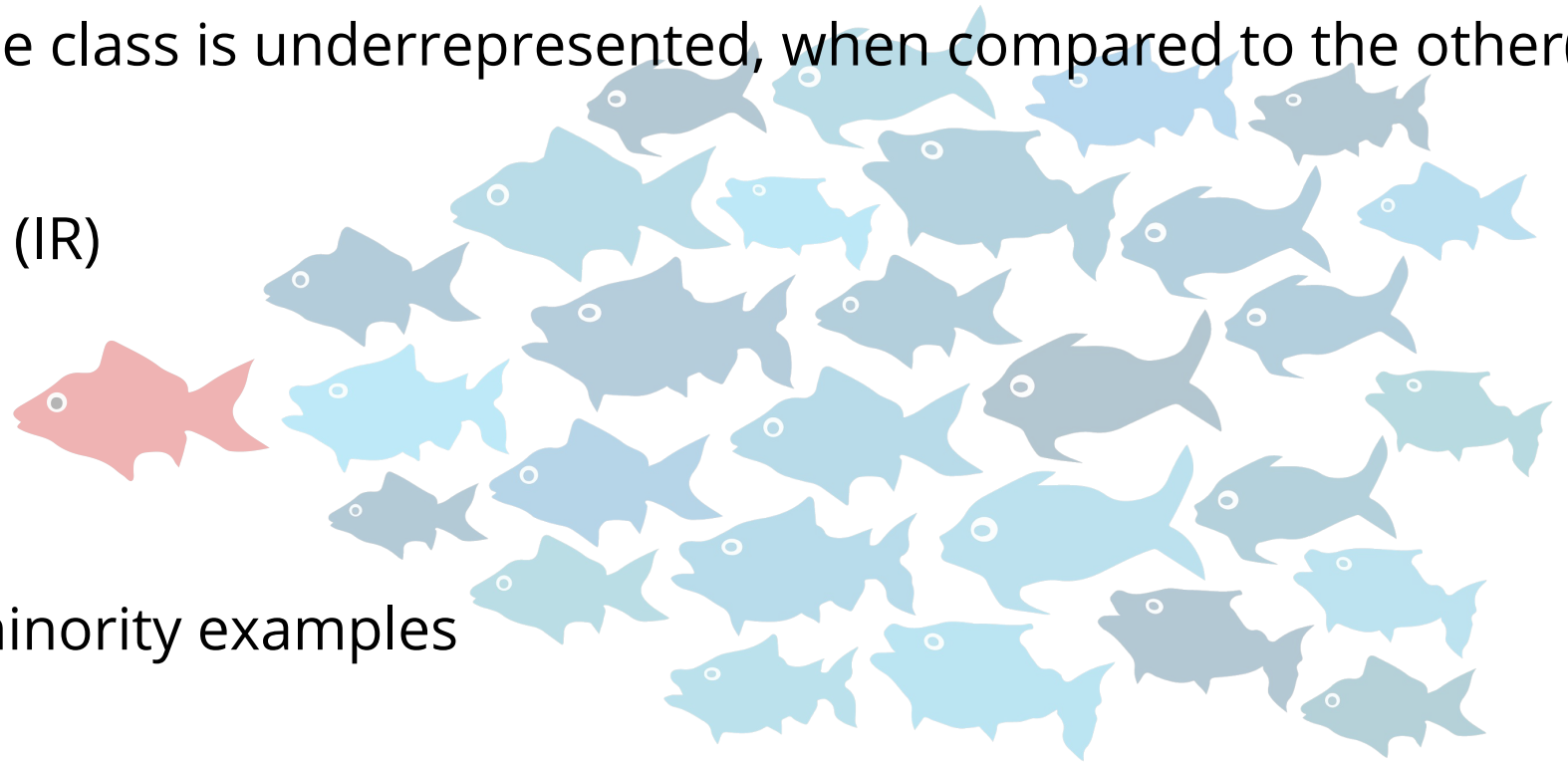


What is Imbalanced Data?

- Occurs when one class is underrepresented, when compared to the other(s)

- Imbalance Ratio (IR)

$$\frac{n_{majority}}{n_{minority}}$$



- Percentage of minority examples

Why

Why

Why should we handle Imbalanced Data?

NATIONAL GEOGRAPHIC

SPECIAL ISSUE

THE FUTURE OF MEDICINE

How personalized medicine is transforming your health care

Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

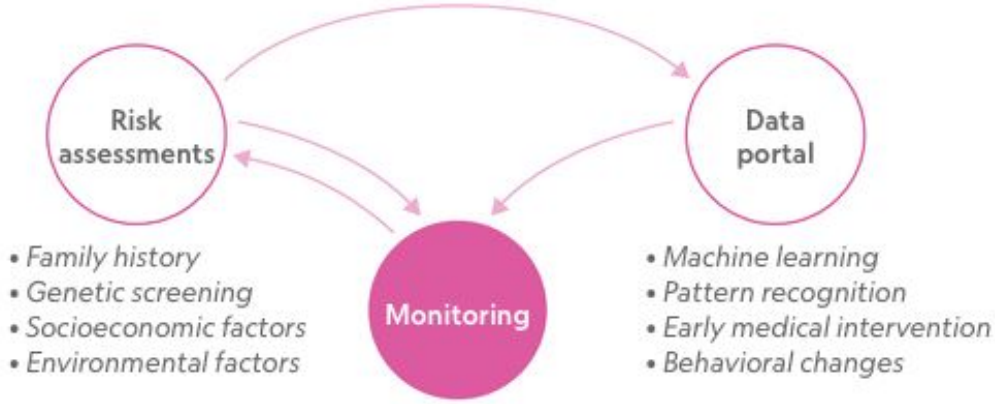
Ziad Obermeyer, M.D., and Ezekiel J. Emanuel, M.D., Ph.D.

By now, it's almost old news: big data will transform medicine. It's essential to remember, however, that data by themselves are useless. To be useful, data must be analyzed, interpreted, and acted on. Thus, it is algorithms — not data sets — that will prove transformative. We believe, therefore, that attention has to shift to new statistical tools from the field of machine learning that will be critical for anyone practicing medicine in the 21st century.

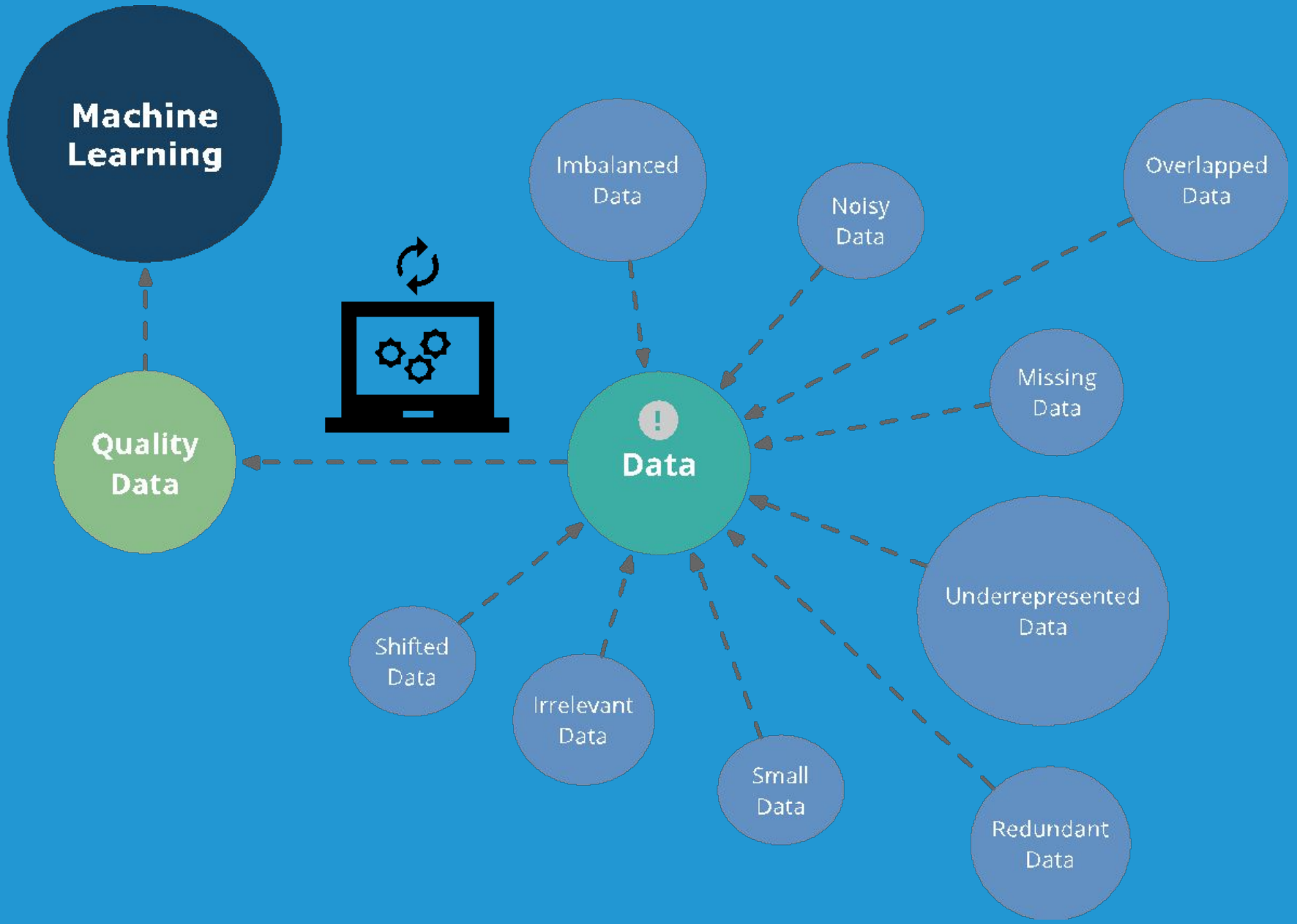
First, it's important to understand what machine learning is not. Most computer-based algorithms in medicine are "expert systems" — rule sets encoding knowledge on a given topic, which are applied to draw conclusions

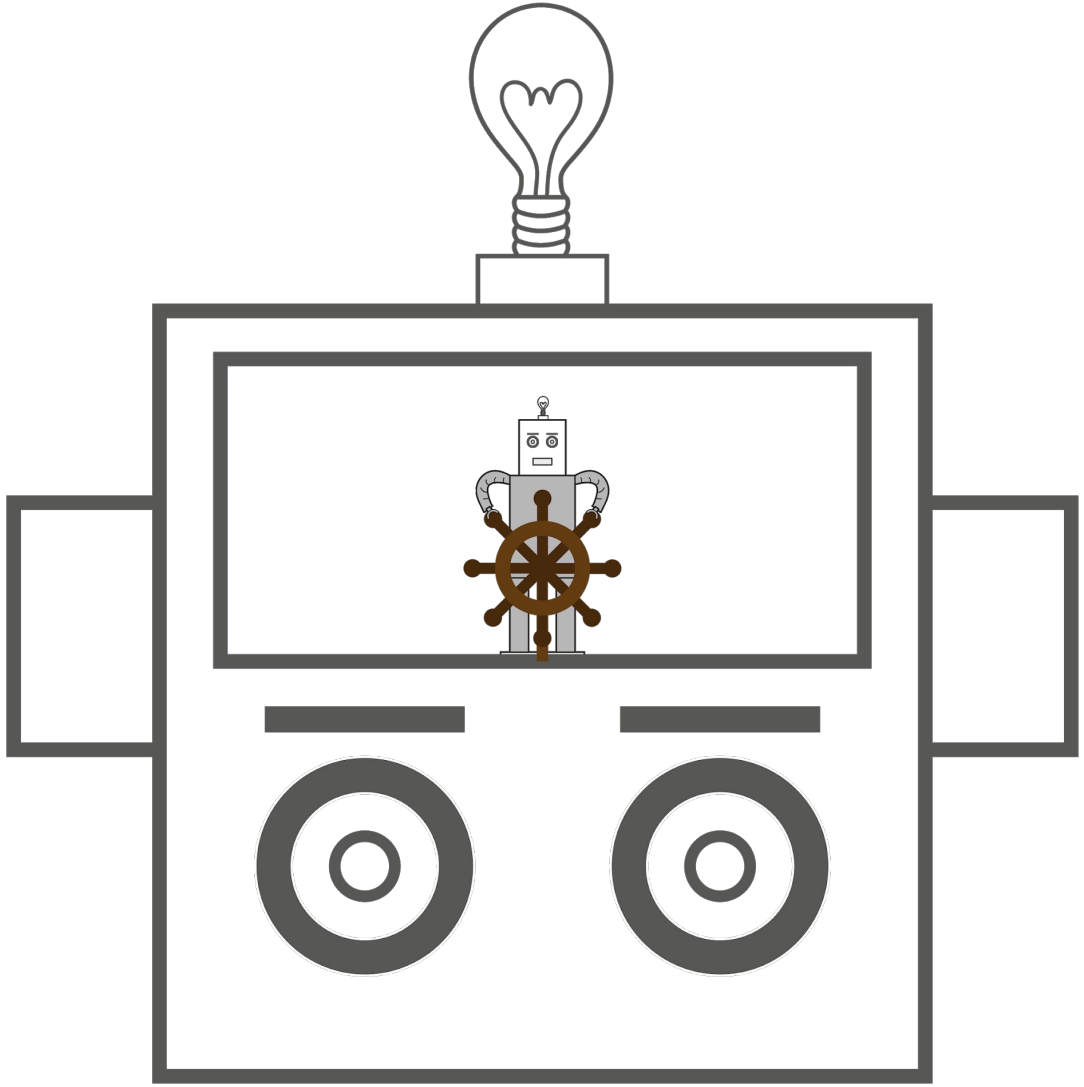
feedzai

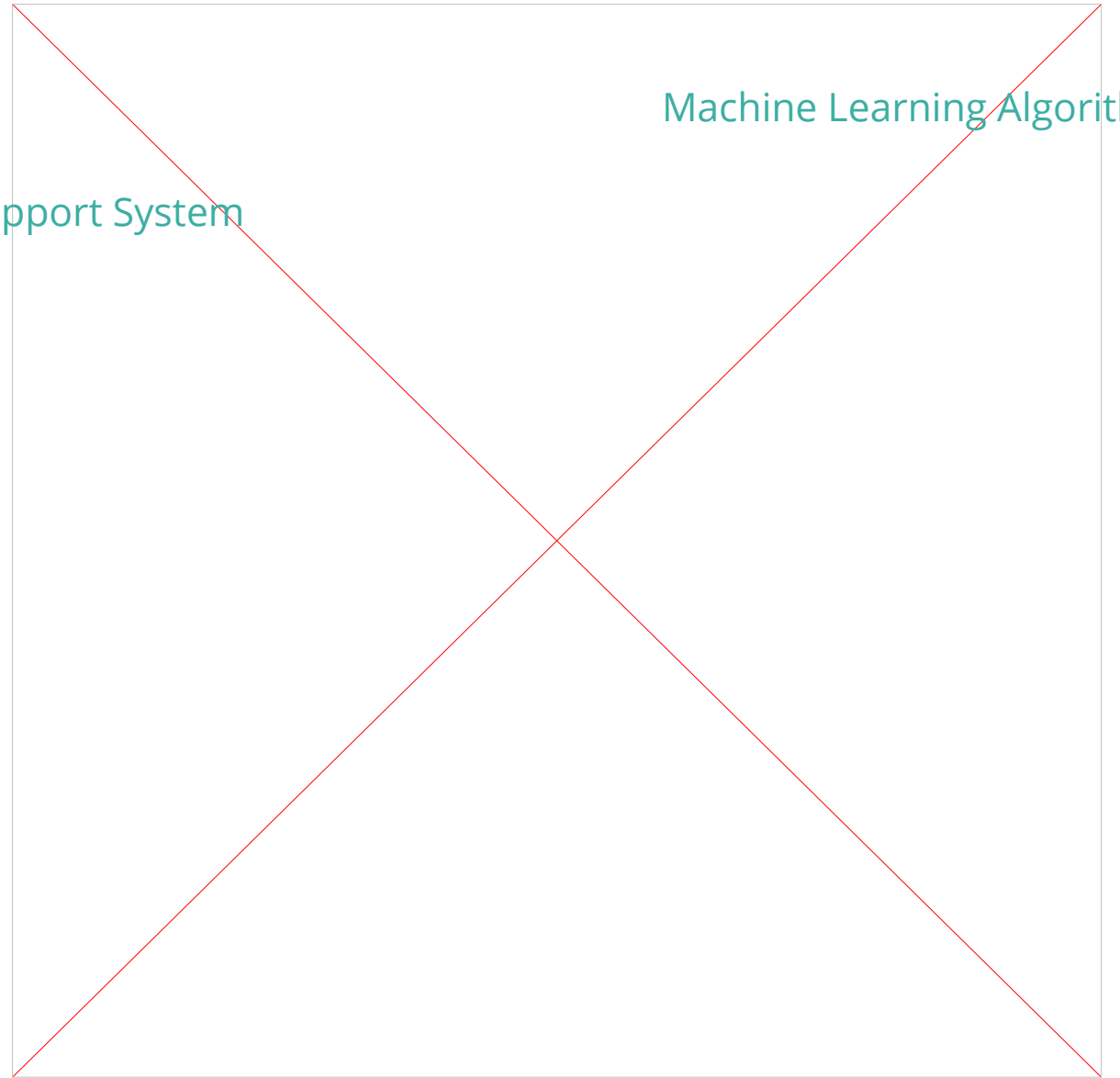
HOW MACHINE LEARNING WILL CHANGE THE WAY BANKS FIGHT FRAUD

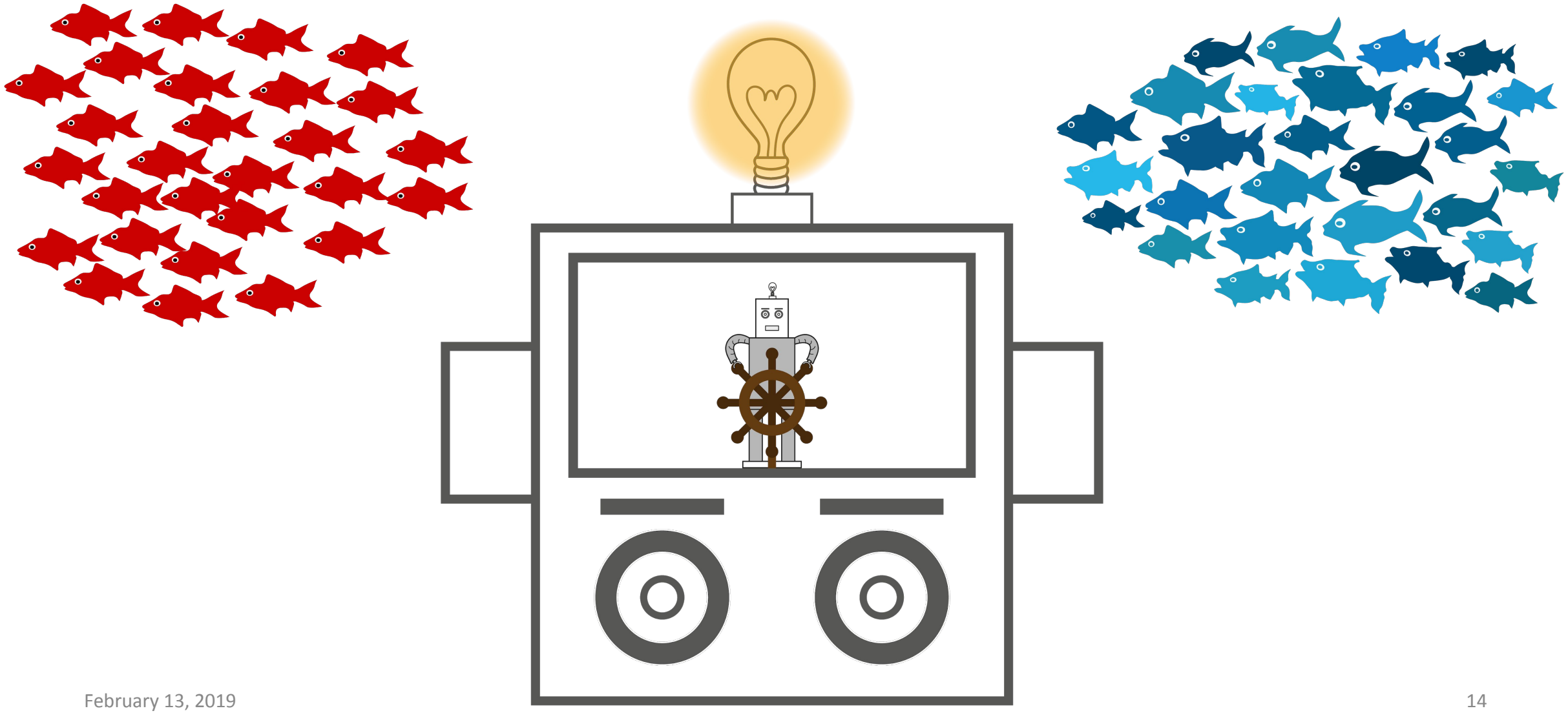


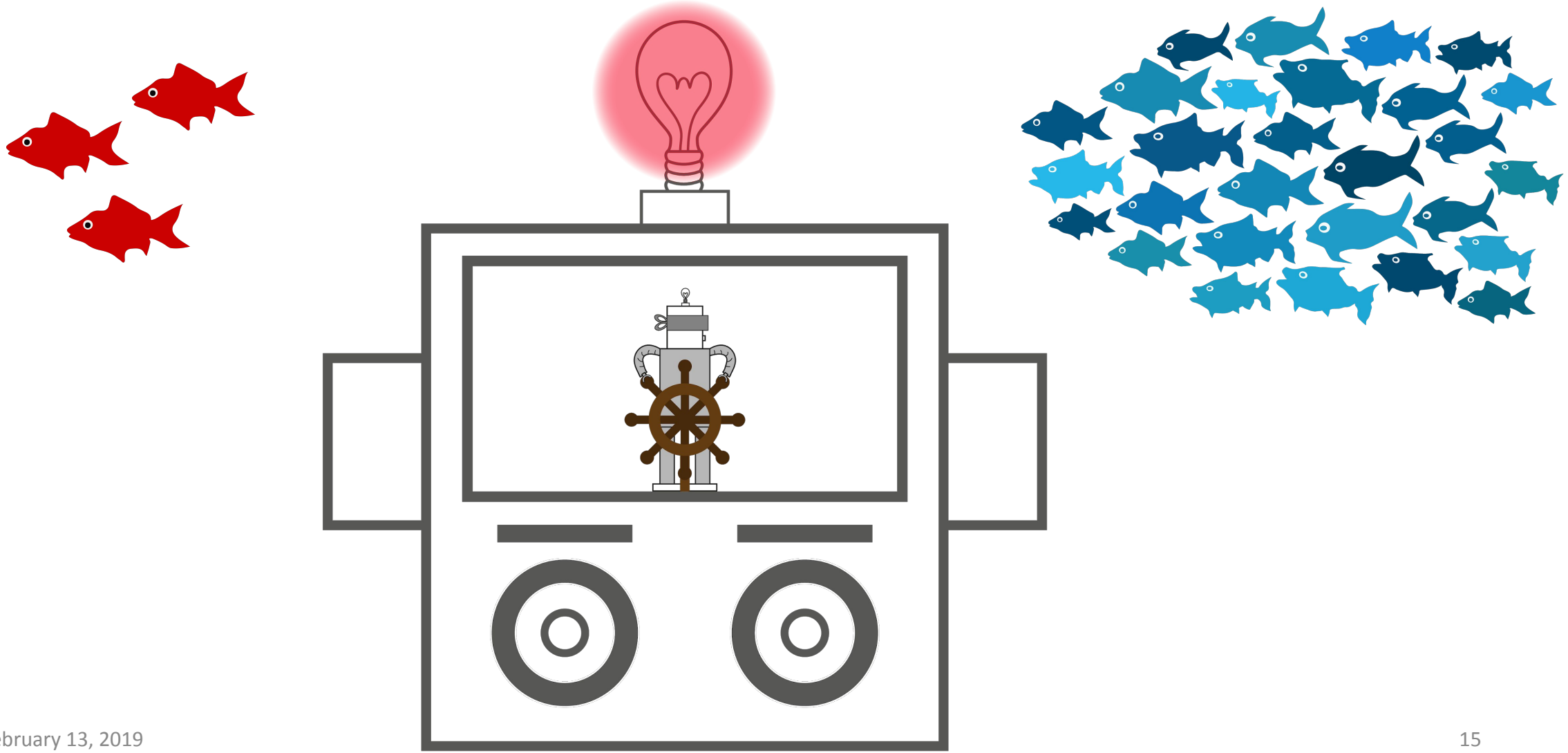

Machine Learning
Security
Automation
Software
Network
Information Technology
Web
Social
Financial Management
Biomedical Engineering
Monitoring
Disease diagnosis
Risk Scoring
Chemical Engineering
Software Defect prediction
Network intrusion
Fake website detection


Machine Learning
Quality Data
Data
Imbalanced Data
Noisy Data
Overlapped Data
Missing Data
Underrepresented Data
Redundant Data
Small Data
Irrelevant Data
Shifted Data




Machine Learning Algorithm
Decision Support System



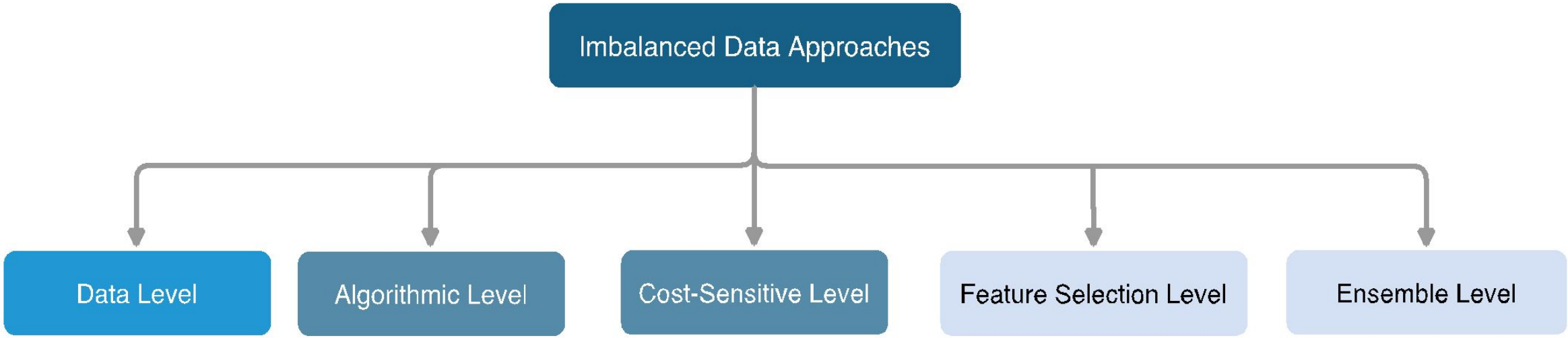


How

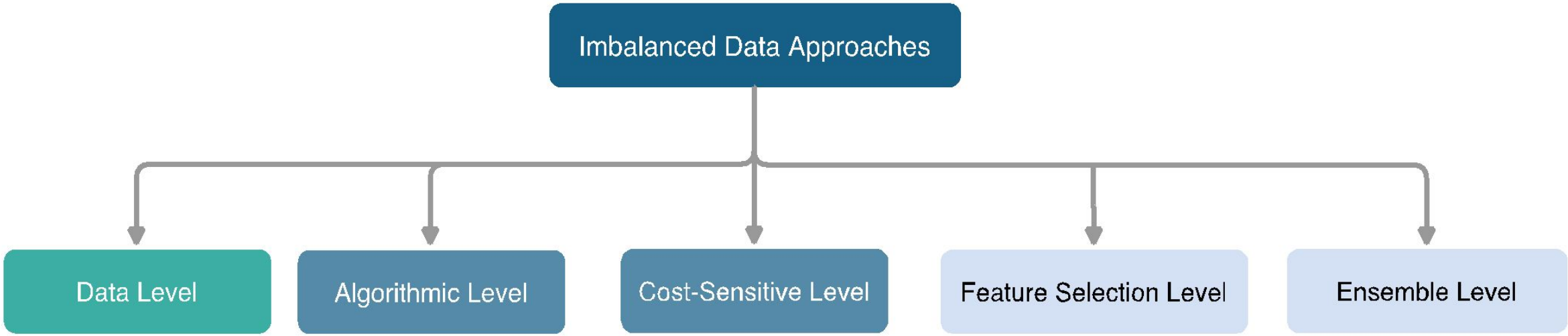
How

How can we handle Imbalanced Data?

Strategies to handle Imbalanced Data

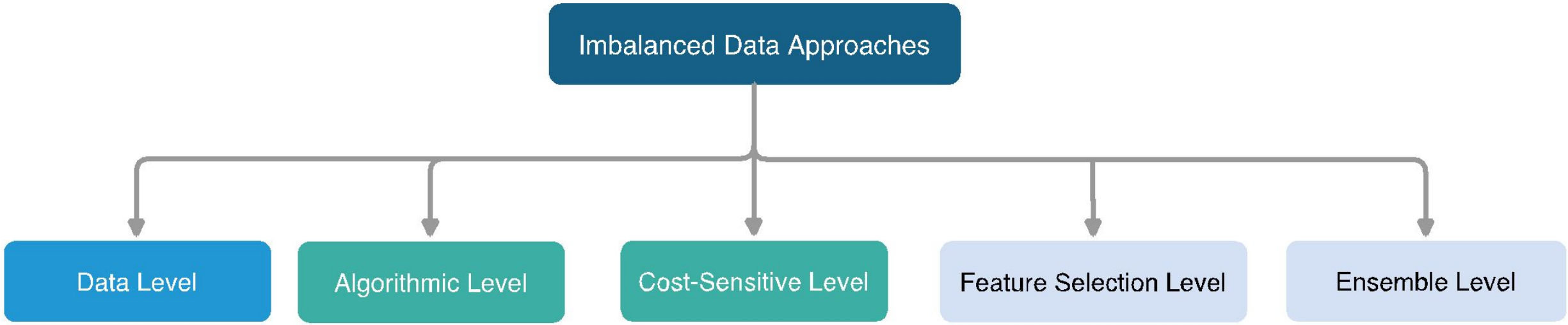


Strategies to handle Imbalanced Data



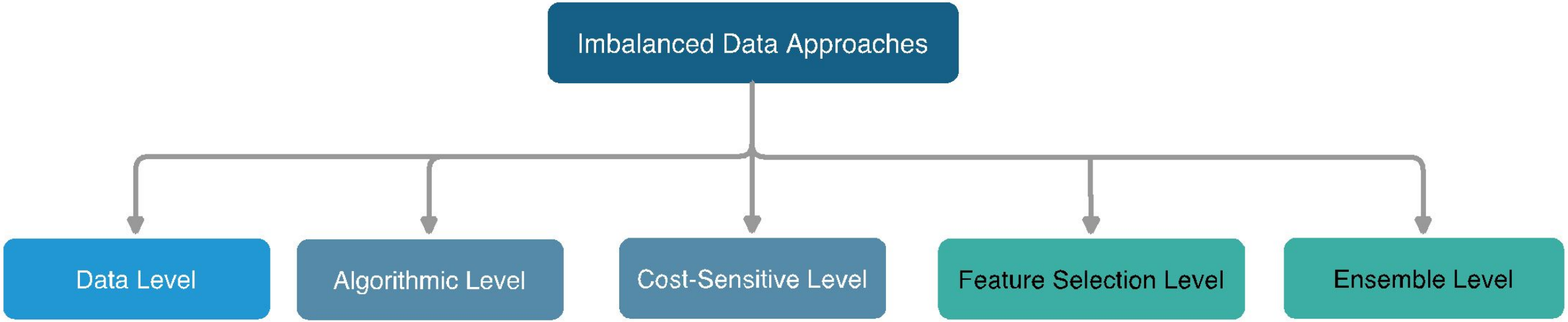
- (Re)Sampling Methods:
  - Modify the prior distribution of the majority and/or minority classes

Strategies to handle Imbalanced Data



- Algorithmic modification:
  - Learning methods are adapted to me more attuned to the class imbalance issues
- Cost-Sensitive Learning:
  - Considers different misclassification costs for different classes

Strategies to handle Imbalanced Data



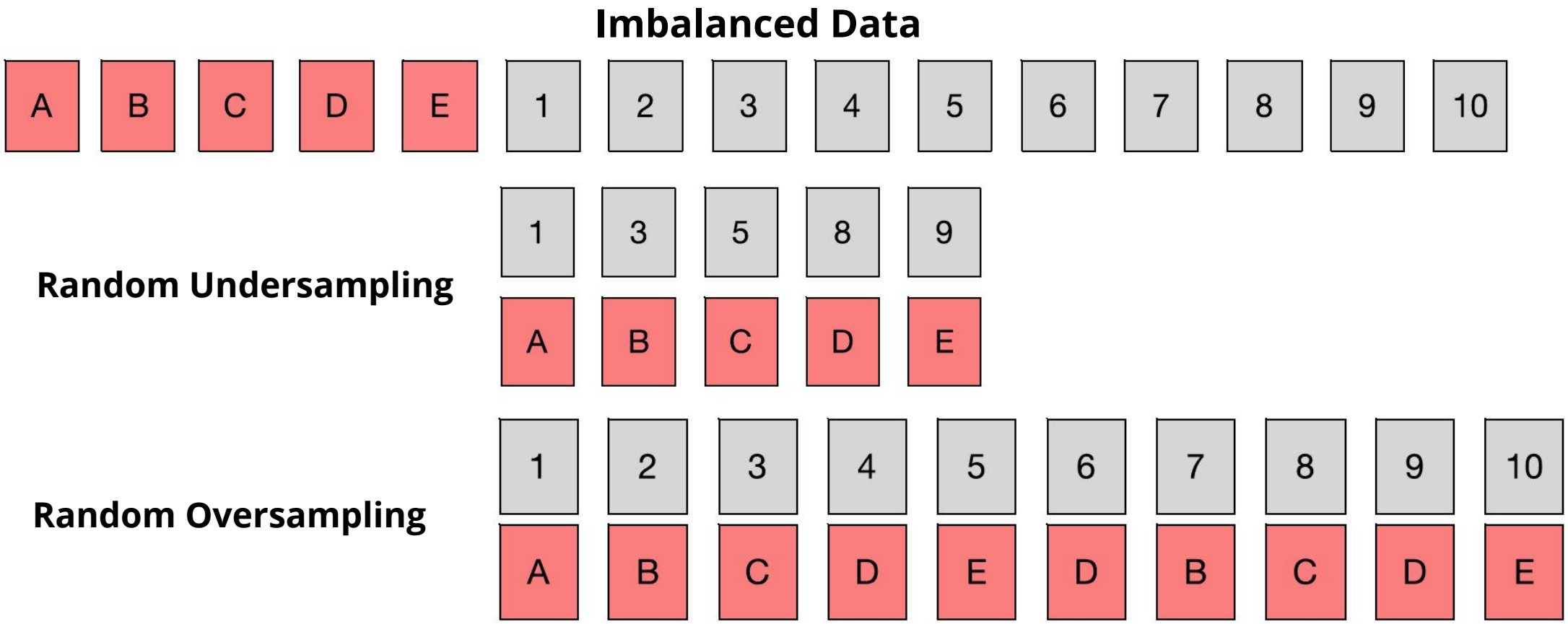
- Feature Selection:
  - Select an informative subset of features
- Ensembles:
  - Aggregate the predictions of several classifiers

Data-Level Approaches

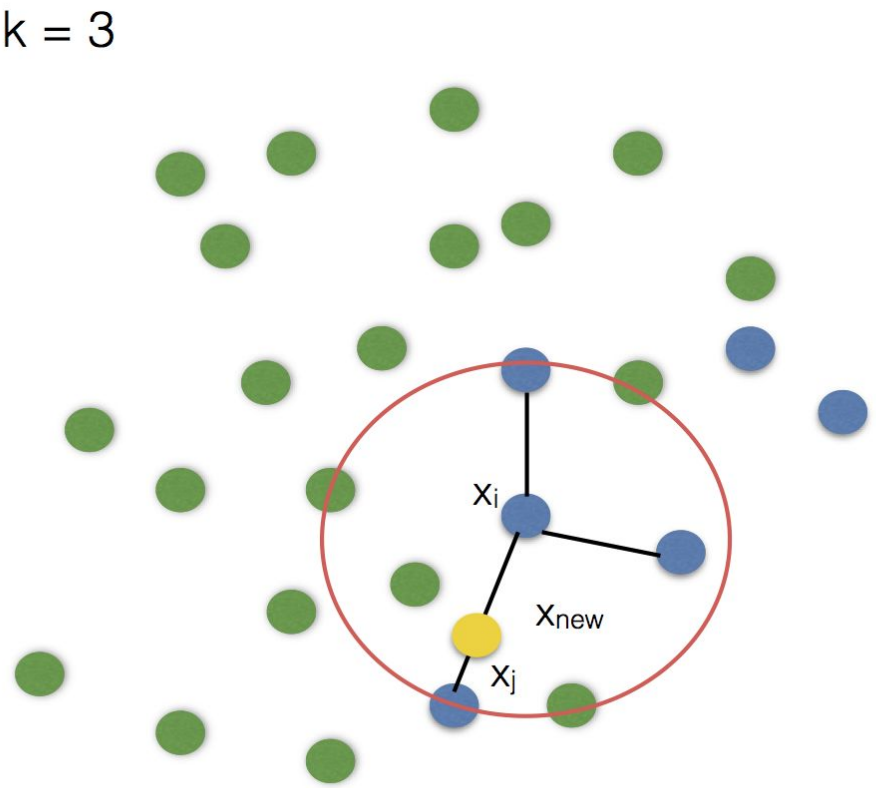
Data Level

- Data-Level approaches are the most commonly used:
  - Have proven to be efficient
  - Are rather intuitive and simple to implement
  - Classifier-independent
- Two main categories:
  - Undersampling: Removing majority examples
  - Oversampling: Adding minority examples

Undersampling and Oversampling

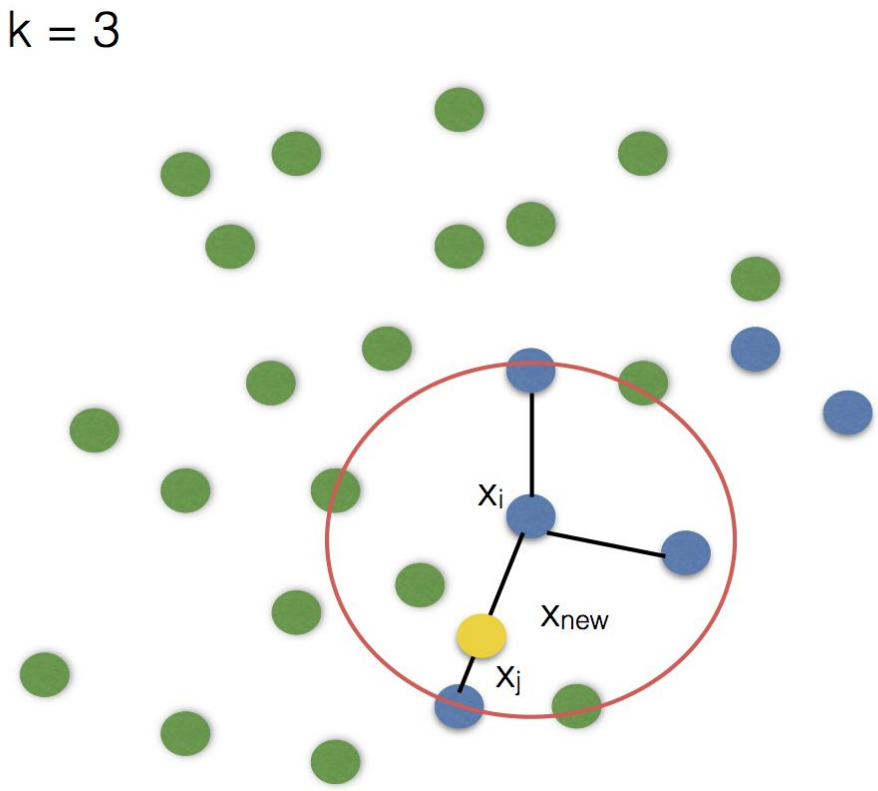


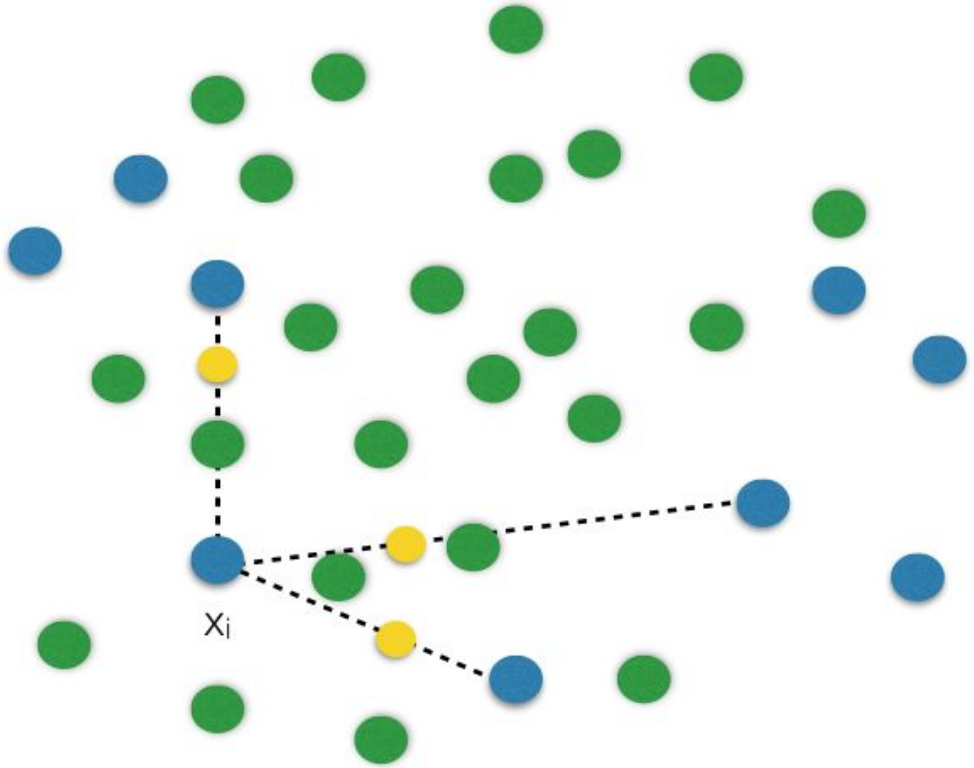
SMOTE: **S**ynthetic **M**inority **O**versampling **T**echnique



$$x_{new} = x_i + (x_j - x_i) \times \delta, \text{where } \delta \in [0, 1]$$

SMOTE: **S**ynthetic **M**inority **O**versampling **T**echnique



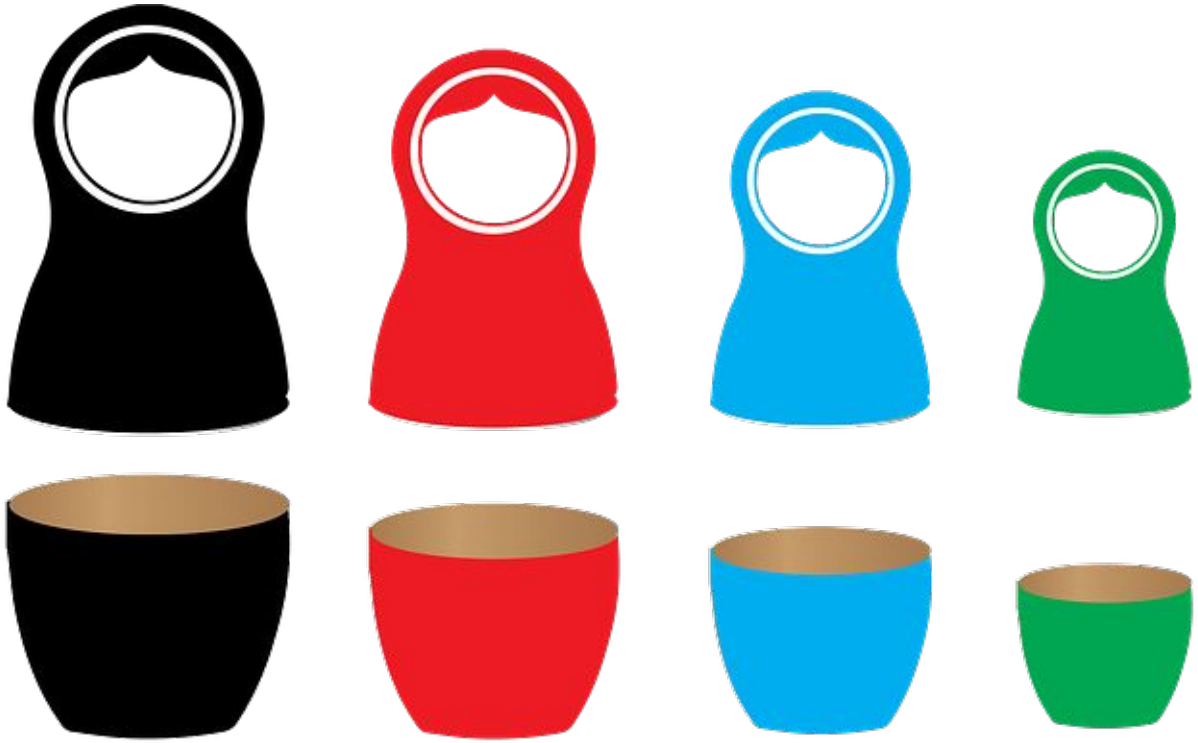


SMOTE variants

Safe-Level-SMOTE

ADOMS

SMOTE



SMOTE-ENN

SMOTE-TL

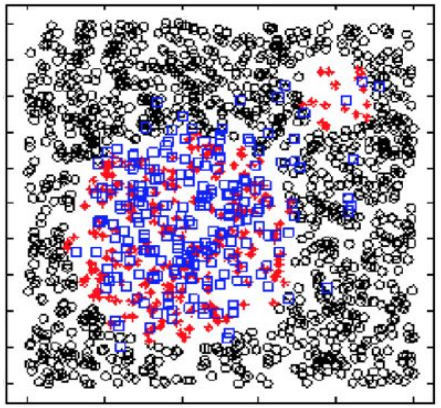
ADOMS

Borderline-SMOTE

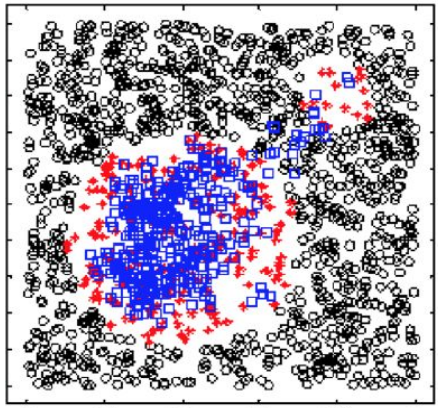
SMOTE-CBO

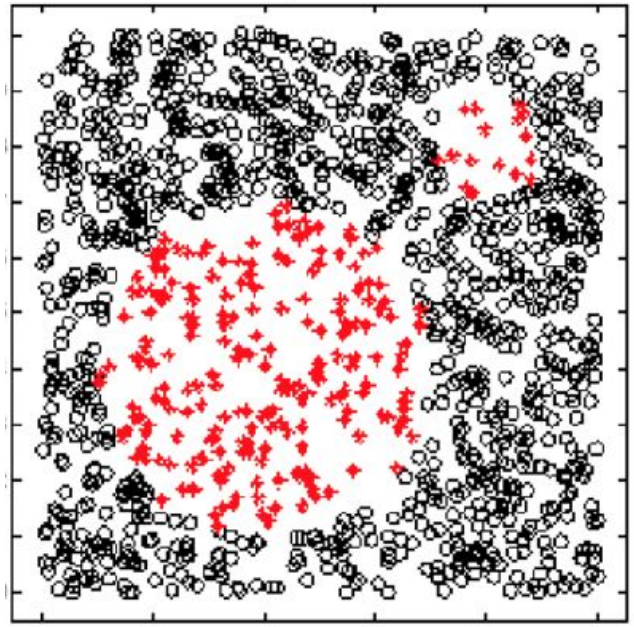
ADASYN

Safe-Level-SMOTE

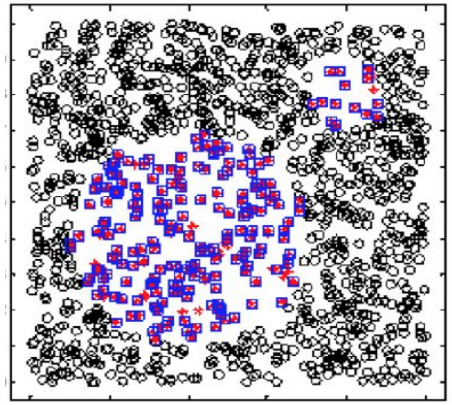


SMOTE

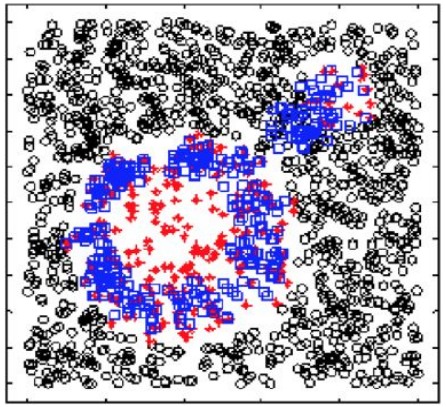




ROS

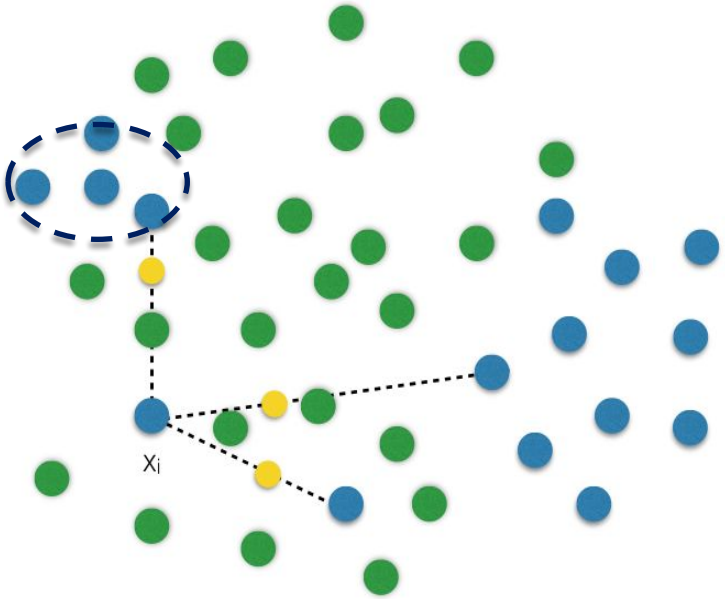


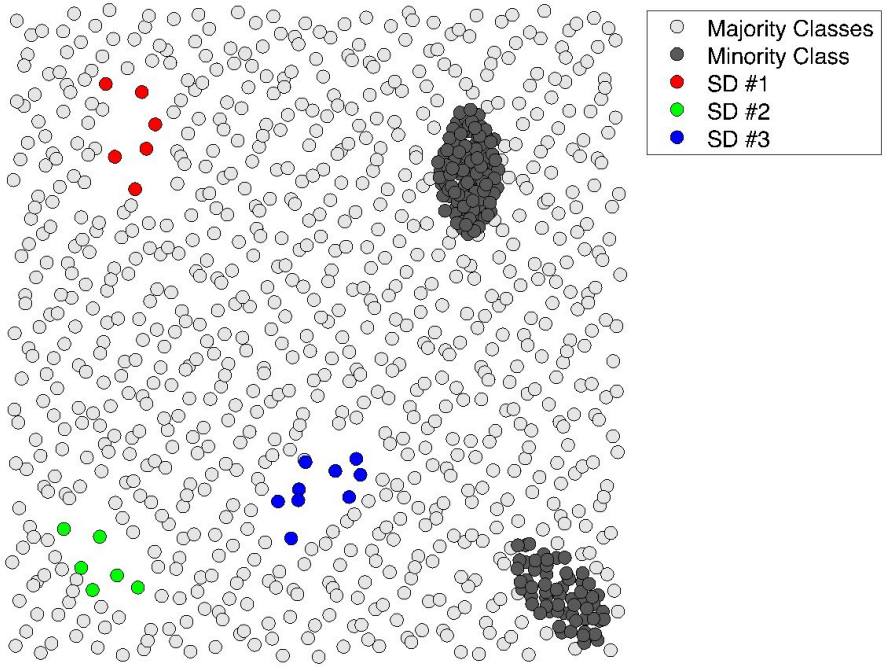
Borderline-SMOTE



Class Imbalance and other Difficulty Factors

- Between and within class imbalance
  - Classifiers are typically biased towards classifying larger disjuncts





CBO: **C**luster-**b**ased **O**versampling

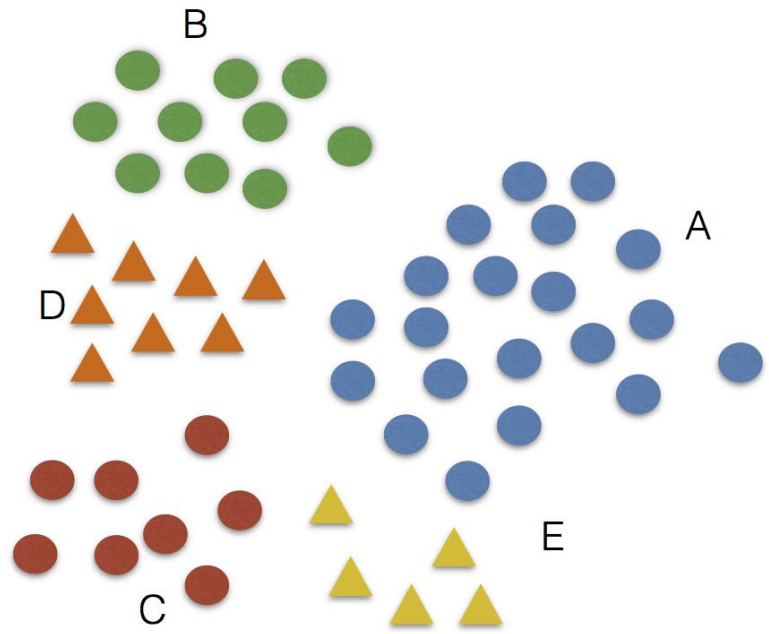
Number of examples in each cluster:

Majority class: A: 20; B: 10; C: 8

$c_{maj} = 3$

Minority class: D: 8; E: 5

$c_{min} = 2$



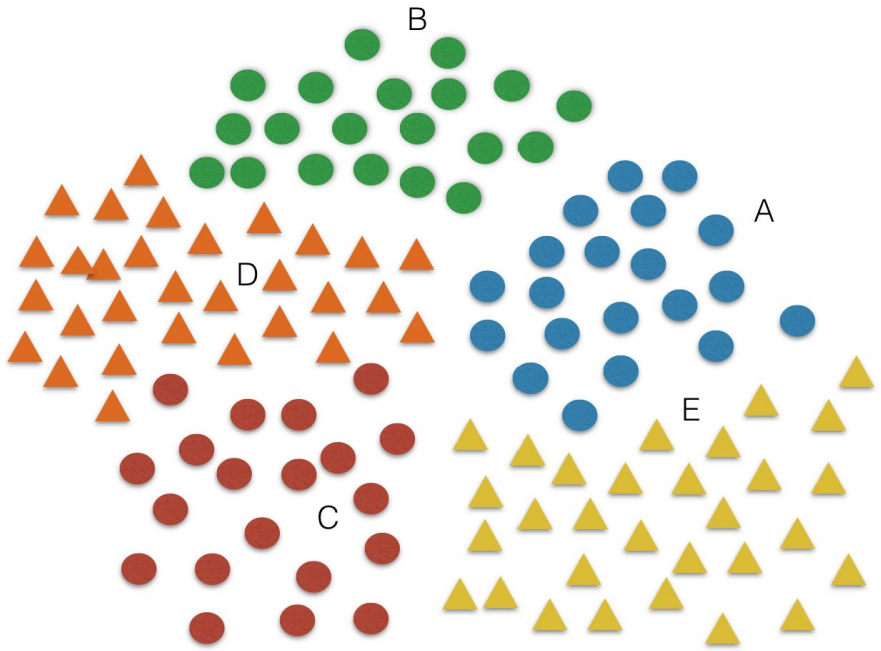
Number of examples in each cluster:

Majority class: A: 20; B: 20; C: 20

$c_{maj} = 3$

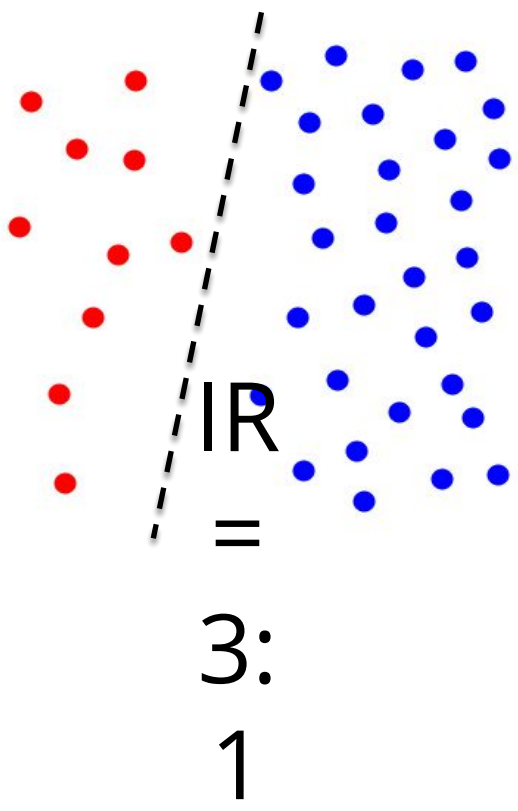
Minority class: D: 30; E: 30

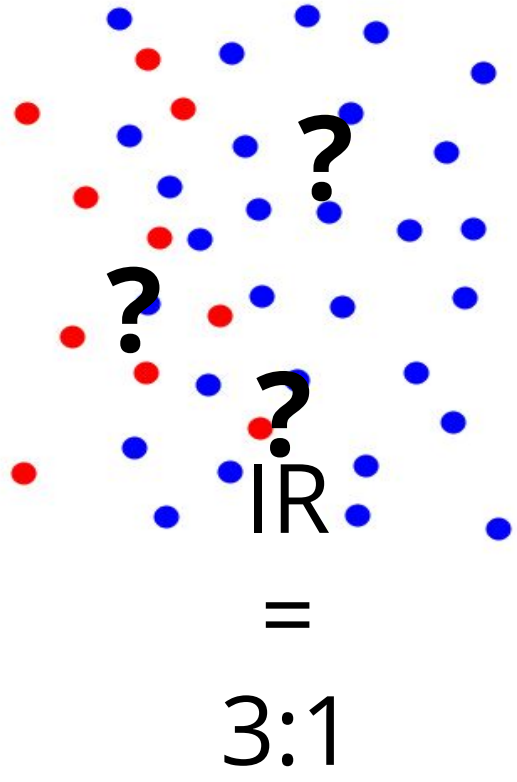
$c_{min} = 2$



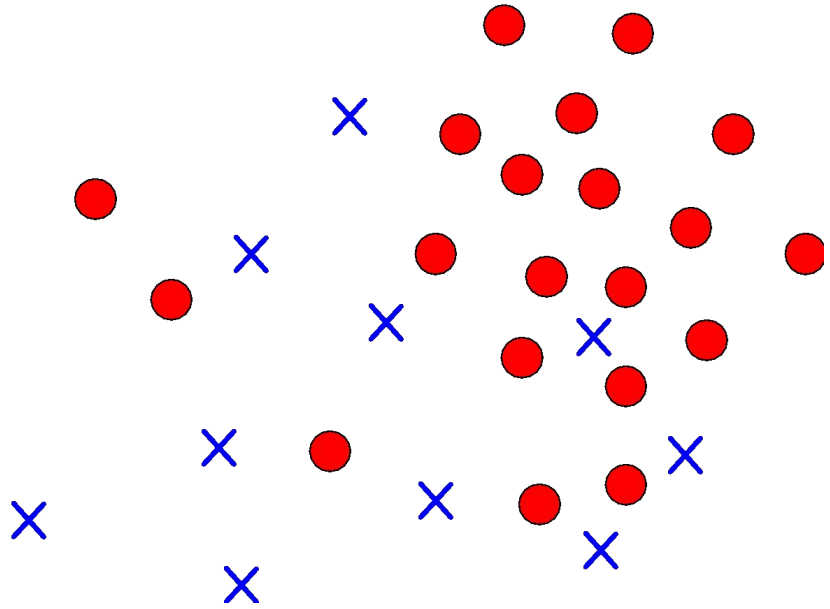
Class Imbalance and other Difficulty Factors

- Class Overlap

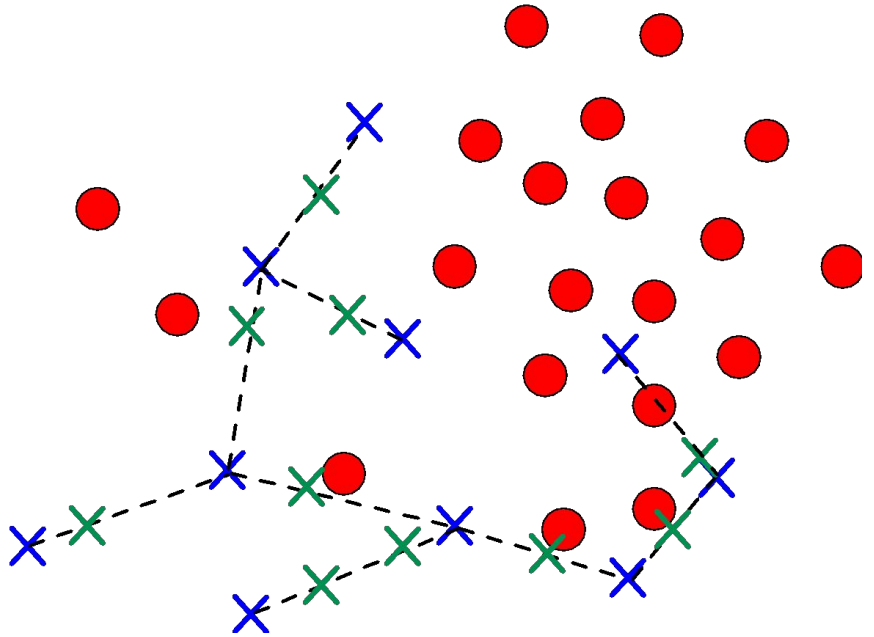




SMOTE-TL and SMOTE-ENN

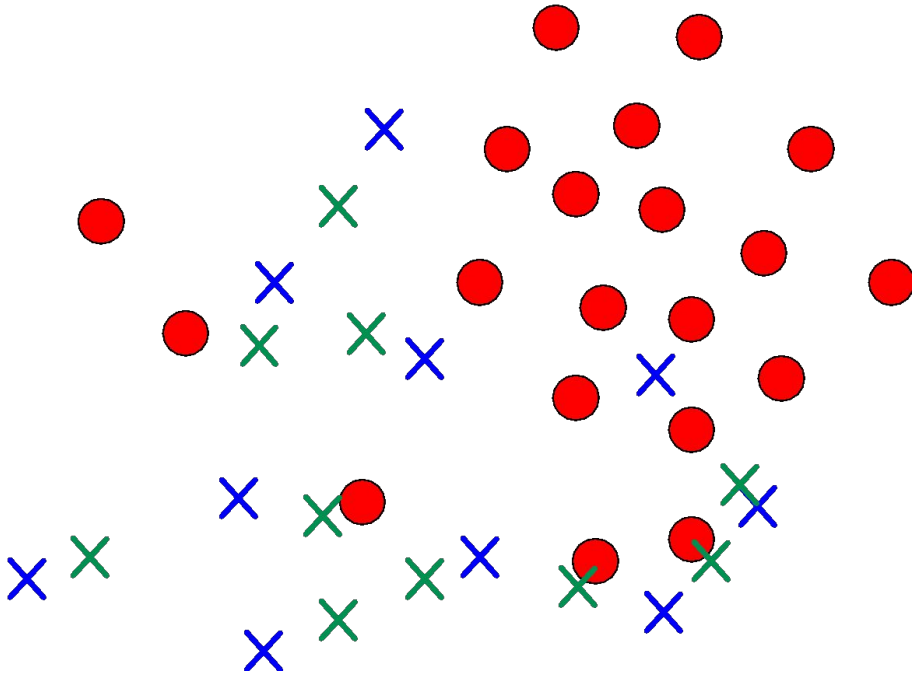


Imbalanced Data

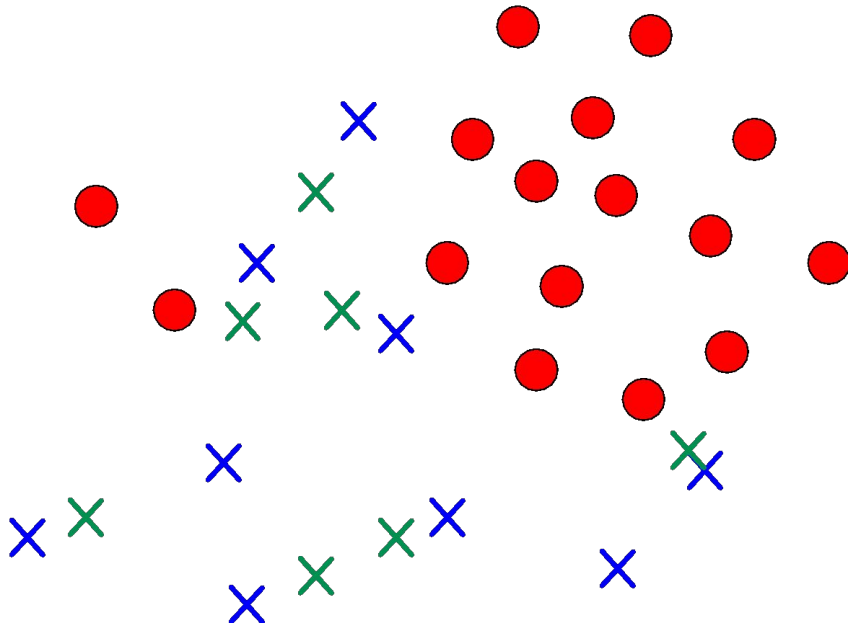


SMOTE

SMOTE-TL

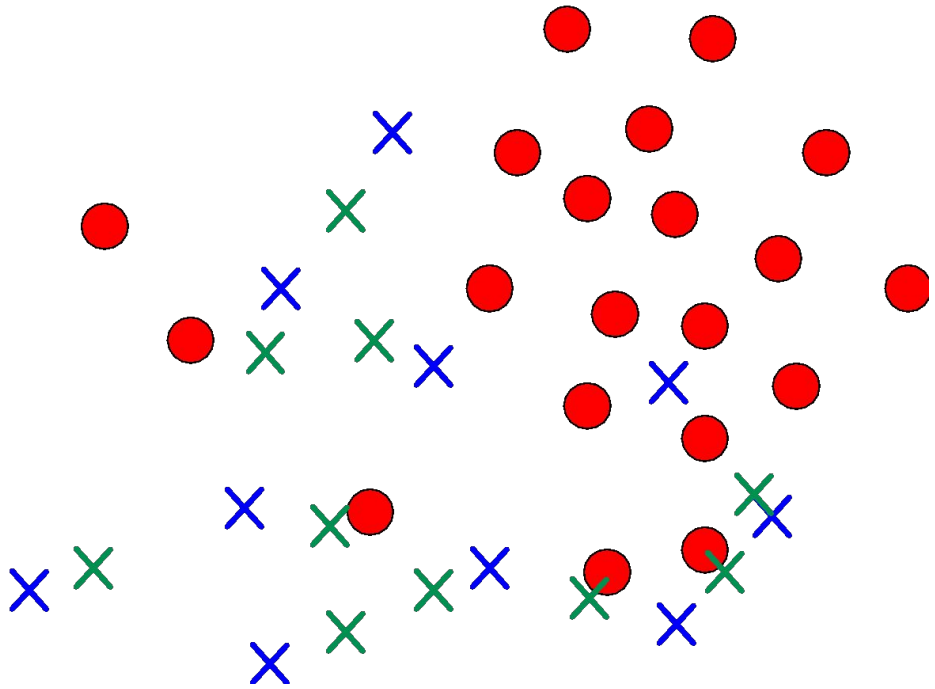


Tomek Links

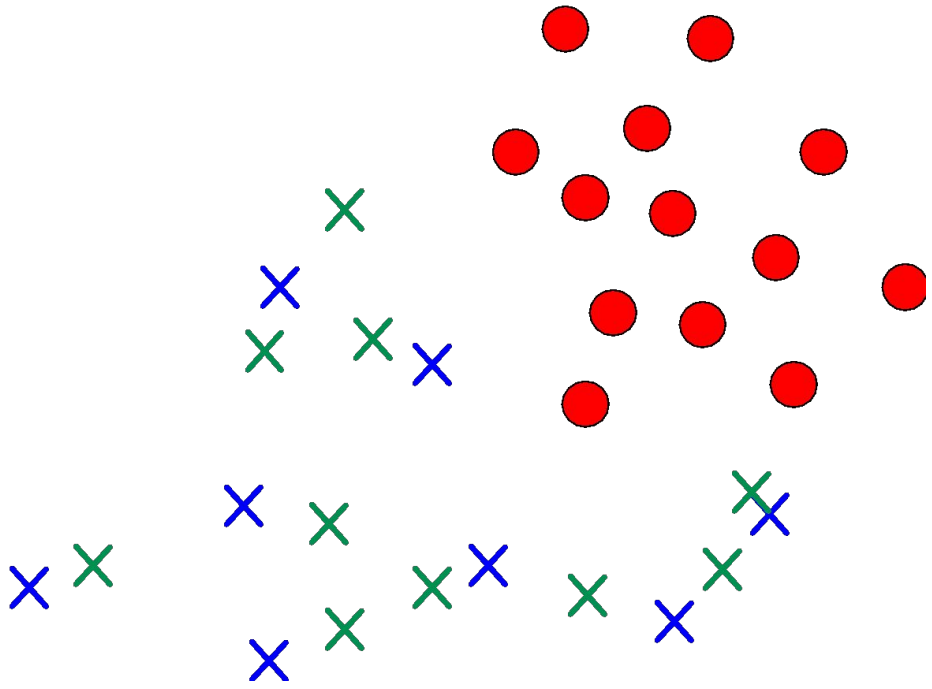


SMOTE-TL

SMOTE-ENN



ENN



SMOTE-ENN

Software for Imbalanced Learning

- KEEL (keel.es)
- WEKA (cs.waikato.ac.nz/ml/weka)
- Rapid Miner (rapidminer.com)
- KNIME (knime.com)

Open Challenges

- Focus on the structure and nature of minority examples to gain a better insight into the source of learning difficulties

- Study of certain data irregularities (difficulty factors)
  - Overlapping classes
  - Within-class imbalance
  - …

- Multi-class imbalanced problems

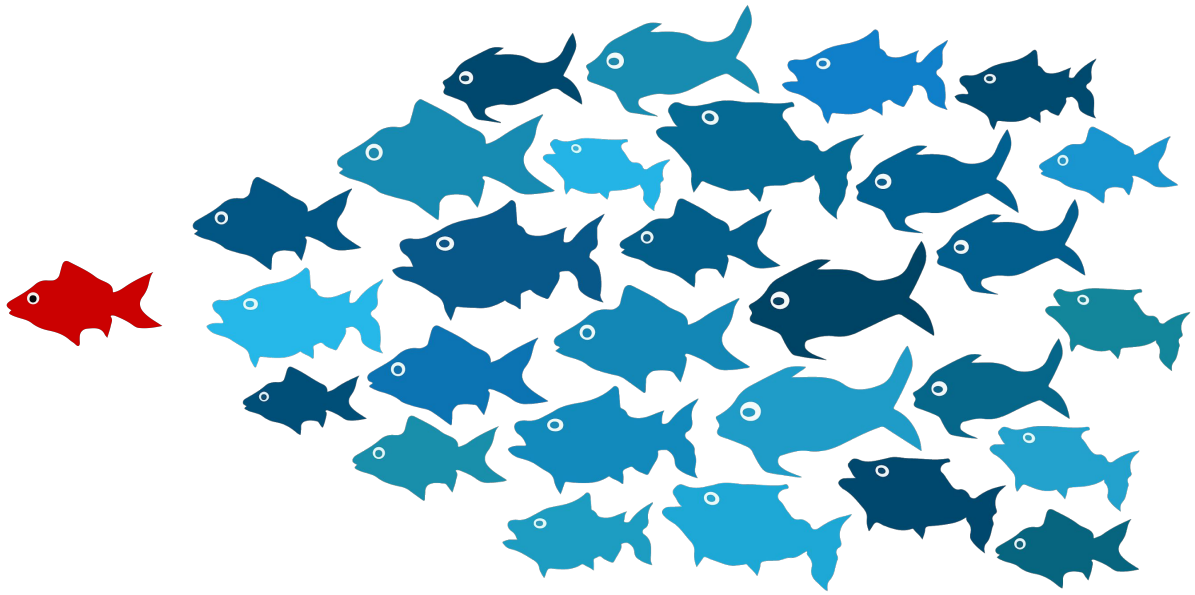
Experimental Design Pitfalls

Rookie mistakes in Imbalanced Data

Experimental Design Pitfalls: Rookie mistake #1

- Inappropriate performance measures
  - 95 blue fish and 5 red fish

		Blue fish	Red fish
Predicted Class	Blue fish	95	5
	Red fish	0	0



  - Accuracy is 95%
  - Blue fish recognition:
    - Specificity: 100%
  - Red fish recognition:
    - Sensitivity: 0%

Experimental Design Pitfalls: Rookie mistake #1

- Common performance measure in Imbalanced Learning literature
  - Sensitivity
  - F-measure
  - G-mean
  - AUC

$$Sensitivity = \frac{TP}{Total\ Positive} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$G_{mean} = \sqrt{Sensitivity \times Specificity}$$

Experimental Design Pitfalls: Rookie mistake #2

3.4. Data balancing

Among the huge number of features extracted (39,600×8) from 300 EHG signals, majority are from EHG recordings of normal pregnancies (term EHG signal). This is due to the imbalance in the term and preterm EHG data available from the database as the database included 262 term birth signals and only 38 preterm birth signals. To avoid this bias introduced by the data imbalance and to maintain the balance of features contributed by both the term and preterm EHG

Experimental Design Pitfalls: Rookie mistake #2

3.4. Data balancing

Among the huge number of features extracted (39,600×8) from 300 EHG signals, majority are from EHG recordings of normal pregnancies (term EHG signal). This is due to the imbalance in the term and preterm EHG data available from the database as the database included 262 term birth signals and only 38 preterm birth signals. To avoid this bias introduced by the data imbalance and to maintain the balance of features contributed by both the term and preterm EHG

Experimental Design Pitfalls: Rookie mistake #2

signals, we have employed data balancing using adaptive synthetic sampling approach (ADASYN) [36]. The algorithm initially calculates the degree of class imbalance and for each data example belonging to minority class; it finds the K nearest neighbors and calculates the density distribution ratio. Finally using the density distribution ratio, algorithm automatically computes the number of synthetic data examples that required to be generated for each minority class. Thus, ADASYN method is used to up sample or increase the features of minority class EHG signal and to provide a balanced representation of the data distribution in the resulting dataset. We have applied this ADASYN method to increase the number of signals of the minority class from 38 to 244.

Experimental Design Pitfalls: Rookie mistake #2

signals, we have employed data balancing using adaptive synthetic sampling approach (ADASYN) [36]. The algorithm initially calculates the degree of class imbalance and for each data example belonging to minority class; it finds the K nearest neighbors and calculates the density distribution ratio. Finally using the density distribution ratio, algorithm automatically computes the number of synthetic data examples that required to be generated for each minority class. Thus, ADASYN method is used to up sample or increase the features of minority class EHG signal and to provide a balanced representation of the data distribution in the resulting dataset. We have applied this ADASYN method to increase the number of signals of the minority class from 38 to 244.

Experimental Design Pitfalls: Rookie mistake #2

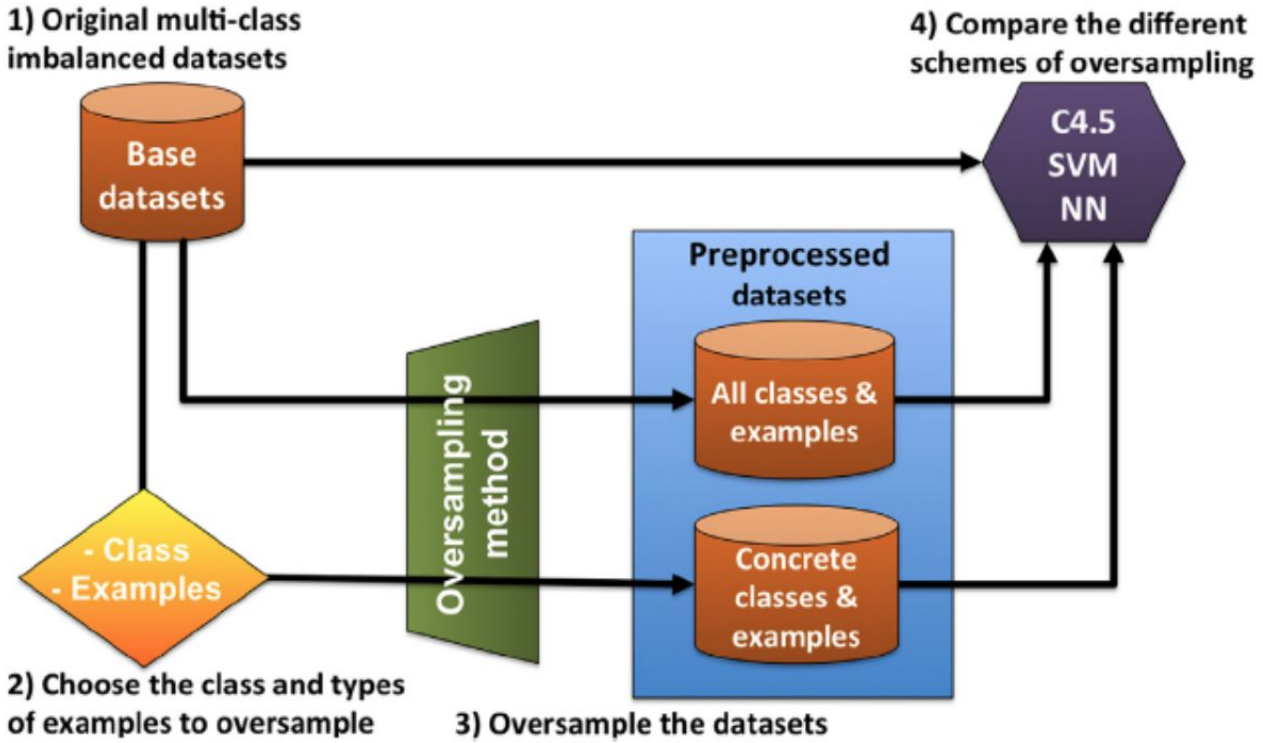


Fig. 3. Methodology to compare the preprocessing of different classes and types of examples.

Experimental Design Pitfalls: Rookie mistake #2

SMOTE oversampling and cross-validation

I am working on a binary classification problem in Weka with a highly imbalanced data set (90% in one category and 10% in the other). I first applied SMOTE (http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/node6.html) to the entire data set to even out the categories and then performed 10-fold cross-validation over the newly obtained data. I found (overly?) optimistic results with F1 around 90%.

Is this due to oversampling? Is it bad practice to perform cross-validation on data on which SMOTE is applied? Are there any ways to solve this problem?

machine-learning weka text-classification

e mistake #2

What is a possible solution for cross validation of an imbalanced data set problem?

What is a possible solution for cross validation of an imbalanced data set problem? The question is in three sections. 1. 1- Oversample the minority class examples using (SMOTE, ADASYN etc), then split it into 10 folds, train the classifier on first nine folds and test on 10th fold and repeat this process 10 times and take the average of metric measure then what about overfitting problem? 2. what about if we divide the data set into 10 folds, oversample the minority class examples in first ninth folds and train the classifier and test the trained classifier on the original (Not oversampled) 10th fold repeat this process 10 times and take the average .. question is what about distribution because basic assumption is training and test set follow the same distribution. 3. If we oversample the minority class examples same as number of majority class examples, then it is necessary to measure F-Measure, G-mean and AUC or accuracy measure is sufficient.

Data Analysis | Data Mining and Knowledge Discovery | Cross-Validation

SM ition

I am with a highly imbalanced data set (90% in one c E (http: chawla02a-html/node6.html) to the entire data 0-fold cross-validation over the newly obtai around 90%.

Is thi cross-validation on data on which SMOTE is applied? Are there any ways to solve this problem?

machine-learning | weka | text-classification

What is a possible solution for cross vali an imbalanced data set problem?

What is a possible solution for cross validation of an imbalance problem? The question is in three sections. 1. 1- Oversample the examples using (SMOTE, ADASYN etc), then split it into 10 folds classifier on first nine folds and test on 10th fold and repeat this times and take the average of metric measure then what about problem? 2. what about if we divide the data set into 10 folds, ov minority class examples in first ninth folds and train the classifie trained classifier on the original (Not oversampled) 10th fold rep 10 times and take the average .. question is what about distribut basic assumption is training and test set follow the same distrib oversample the minority class examples same as number of ma examples, then it is necessary to measure F-Measure, G-mean a accuracy measure is sufficient.

Data Analysis | Data Mining and Knowledge Discovery | Cross-Validation

SM

I a

on

(ht

da

ob

Is

is applied? Are there any ways to solve this

machine-learning | weka | text-classification

CrossValidated take #2

I am working on severely imbalanced data. In literature, several methods are used to re-balance the data using re-sampling (over- or under-sampling). Two good approaches are:

- SMOTE: Synthetic Minority Over-sampling TEchnique (SMOTE)
- ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN)

I have implemented ADASYN because its adaptive nature and ease to extension to multi-class problems.

My question is how to test the oversampled data produced by ADASYN (or any other oversampling methods). It is not clear in the mentioned two paper how they performed their experiments. There are two scenarios:

1- Oversample the whole dataset, then split it to training and testing sets (or cross validation).

2- After splitting the original dataset, perform oversampling on the training set only and test on the original data test set (could be performed with cross validation).

In the first case the results are much better than without oversampling, but I am concerned if there is overfitting. While in the second case the results are slightly better than without oversampling and much worse than the first case. But the concern with the second case is if all minority class samples goes to the testing set, then no benefit will be achieved with oversampling.

I am not sure if there are any other settings to test such data. Waiting for your inputs.

classification | dataset | resampling | unbalanced-classes | oversampling

What is a possible solution for cross vali… an imbalanced data set problem?

What is a possible solution for cross validation of an imbalance… problem? The question is in three sections. 1. 1- Oversample the… examples using (SMOTE, ADASYN etc), then split it into 10 folds… classifier on first nine folds and test on 10th fold and repeat this… times and take the average of metric measure then what about … problem? 2. what about if we divide the data set into 10 folds, ov… minority class examples in first ninth folds and train the classifie… trained classifier on the original (Not oversampled) 10th fold rep…

CrossValidated take #2

I am working on severely imbalanced data. In literature, several methods are used to re-balance the data using re-sampling (over- or under-sampling). Two good approaches are:

- SMOTE: Synthetic Minority Over-sampling TEchnique (SMOTE)
- ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN)

I have implemented ADASYN because its adaptive nature and ease to extension to multi-class problems.

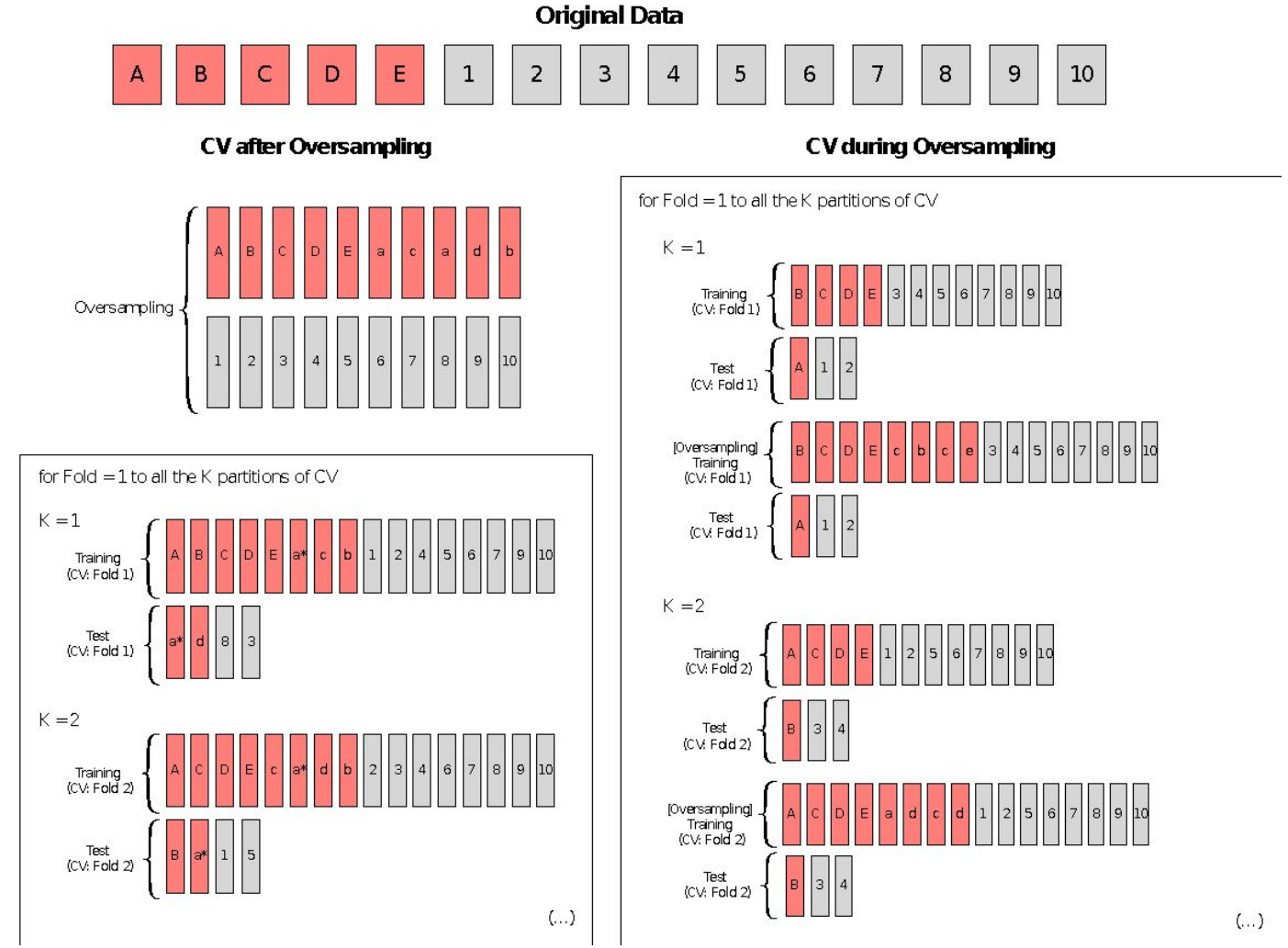
Will SMOTE make you more prone to overfit? (self.datascience)

submitted 1 year ago by Cjh411

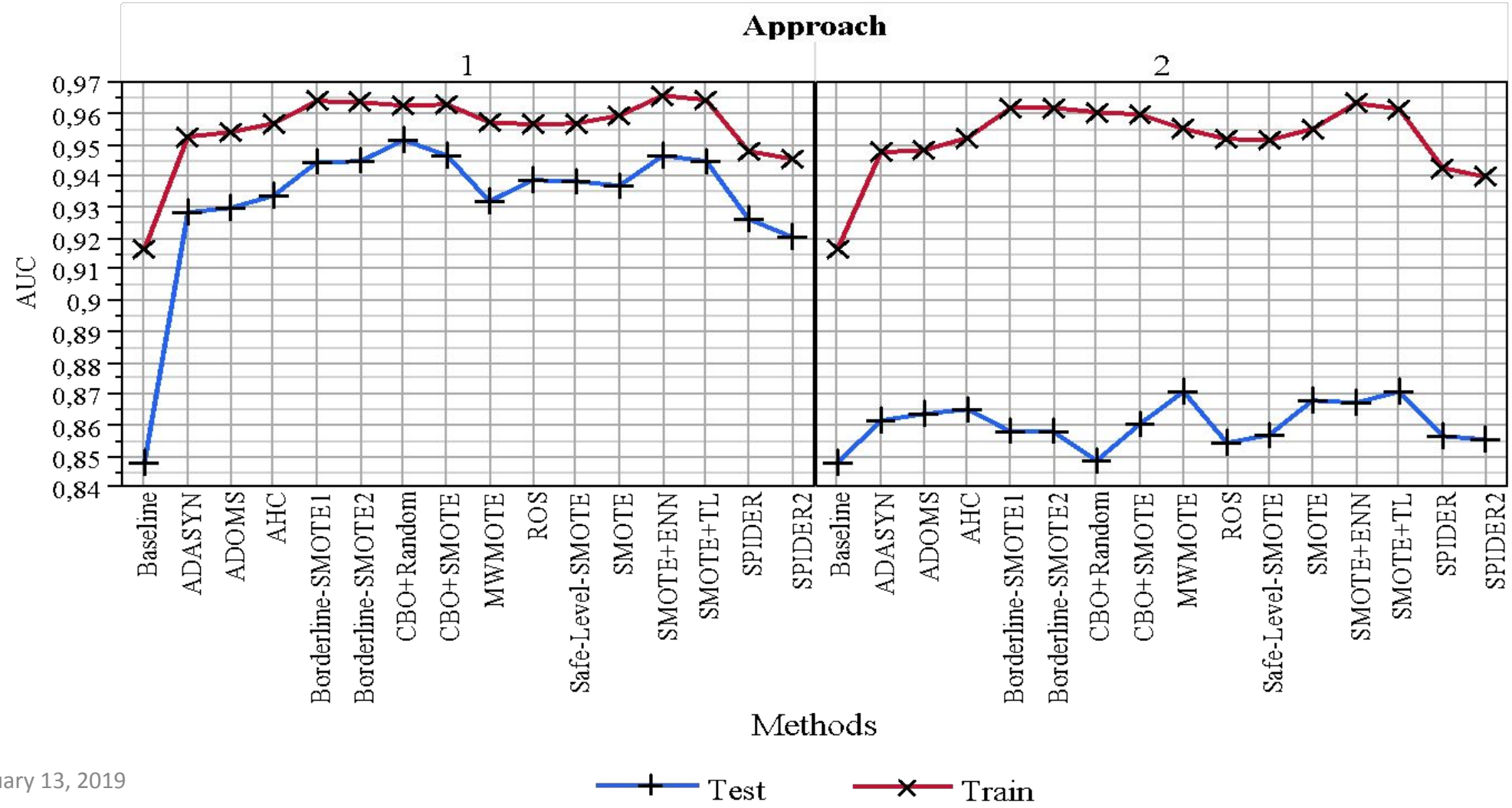
Someone brought up using SMOTE on a project where there is not only a large data imbalance but also very few minority records in absolute value (50 out of 1000). The data is very noisy and there is a lot of overlap between the minority and majority cases.

I understand the benefit of SMOTE in correcting for bias in your model. But In this situation do you run the risk of overfitting your model and losing generalizability on future data? I've tried searching the web and can't find much on the relationship between SMOTE and overfitting on small samples. Intuitively it feels dangerous to strengthen a signal in your data that you're skeptical of to begin with. We're using 10 fold cv for evaluation.

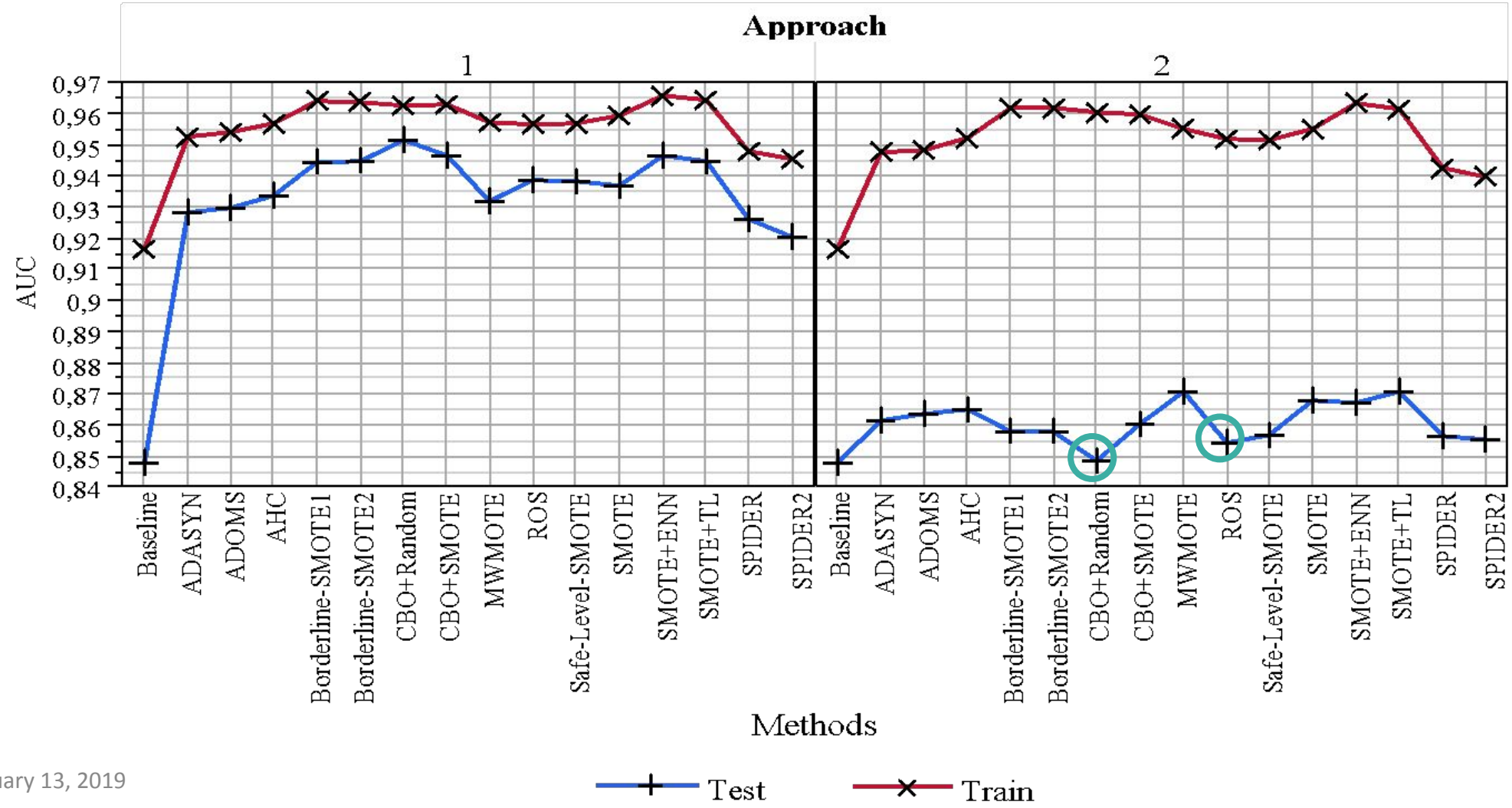
Experimental Design Pitfalls: Rookie mistake #2



Experimental Design Pitfalls: Rookie mistakes #2 and #3

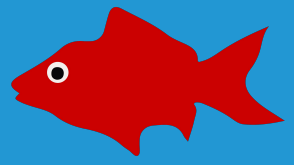


Experimental Design Pitfalls: Rookie mistakes #2 and #3



Take-home message!

- Imbalanced Data requires more informative measures
  - Accuracy/Error Rate are not appropriate
- Overoptimism is most often associated with inappropriate validation setups
  - Accuracy/Error Rate are not appropriate
- Overfitting is mostly related to the oversampling algorithm
  - Creating exact replicas of existing patterns is the most prejudicial technique



Thank you

UNIVERSIDADE D
COIMBRA

IPOPORTO CISUC COGNITIVE & MEDIA SYSTEMS

DSPT
DATA SCIENCE PORTUGAL

Coimbra, February 13, 2019

Dealing with Imbalanced Data

the nuts and bolts

Miriam Seoane Santos
CISUC, DEI/FCTUC, University of Coimbra
IPO-Porto Research Centre (CI-IPOP),
Porto