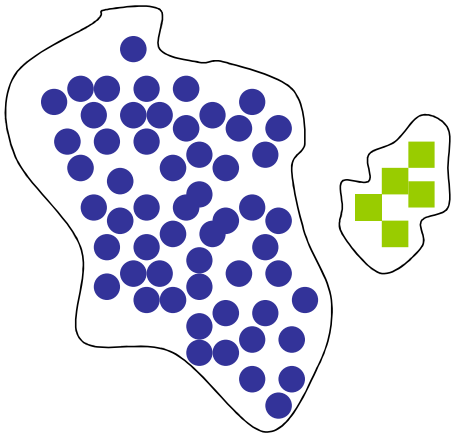
Class Imbalance Problem

A training dataset is called **imbalanced** if at least one of the classes are represented by significantly less number of instances than the others.

Why are some datasets imbalanced?

- **Natural reasons:** Normal examples are generally abundant and form the majority (negative) class. On the other hand, examples of interest are generally rare and form the minority (positive) class.
- **Limitations:** Cost, difficulty, privacy limits on collecting instances of some classes.
- **Multiclass classification:** One-against-rest schema



Medical diagnosis

Classification of surveillance events

identifying fraudulent activity in transactions

text categorization

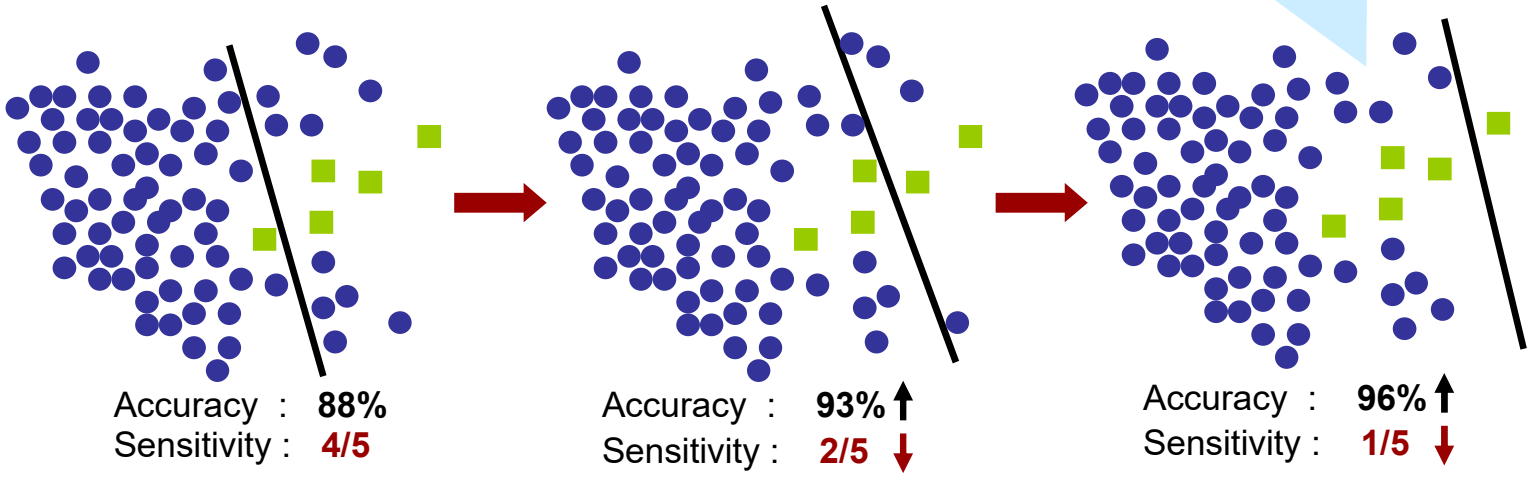
Classification Boundary & Sensitivity

The class boundary learned by standard machine learning algorithms can be severely skewed toward the positive class.

- Negative (majority) class (# of instances: 95)
- Positive (minority) class (# of instances: 5)

No classification error on the negative class.

BUT is it good for sensitivity?



Machine learning algorithms try to maximize the classification accuracy.
But **accuracy on positive class (sensitivity)** may decrease!

Methods to handle Imbalanced Data Classification

Oversampling

- Duplicating minority examples multiple times
- Introducing new synthetic examples for minority class

!!! Computational limitations

Undersampling

- Ignoring part of the majority class

!!! Possibility of discarding informative instances

Different misclassification penalty parameters

Active Learning