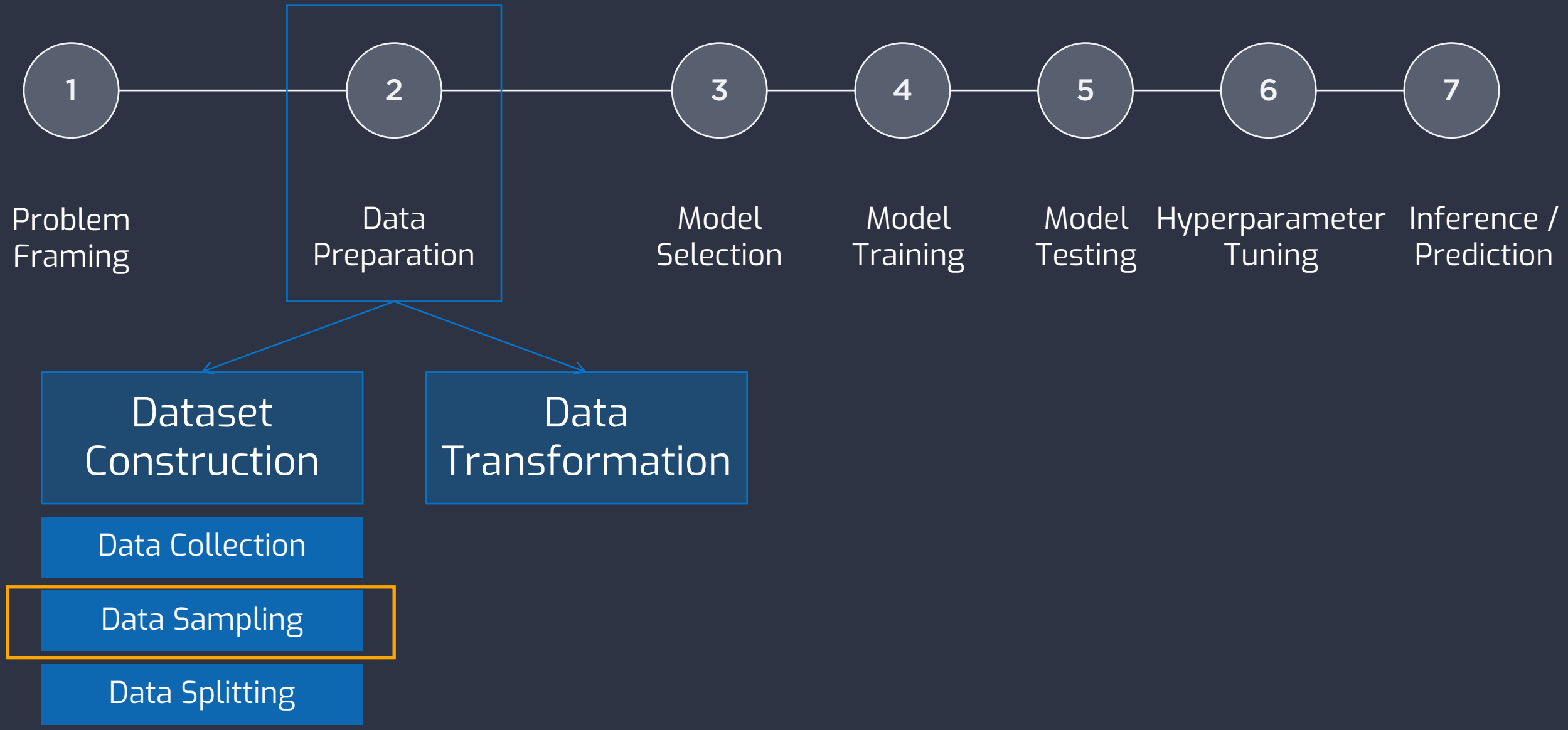
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Data Sampling

Dr SHI Lei

Durham University

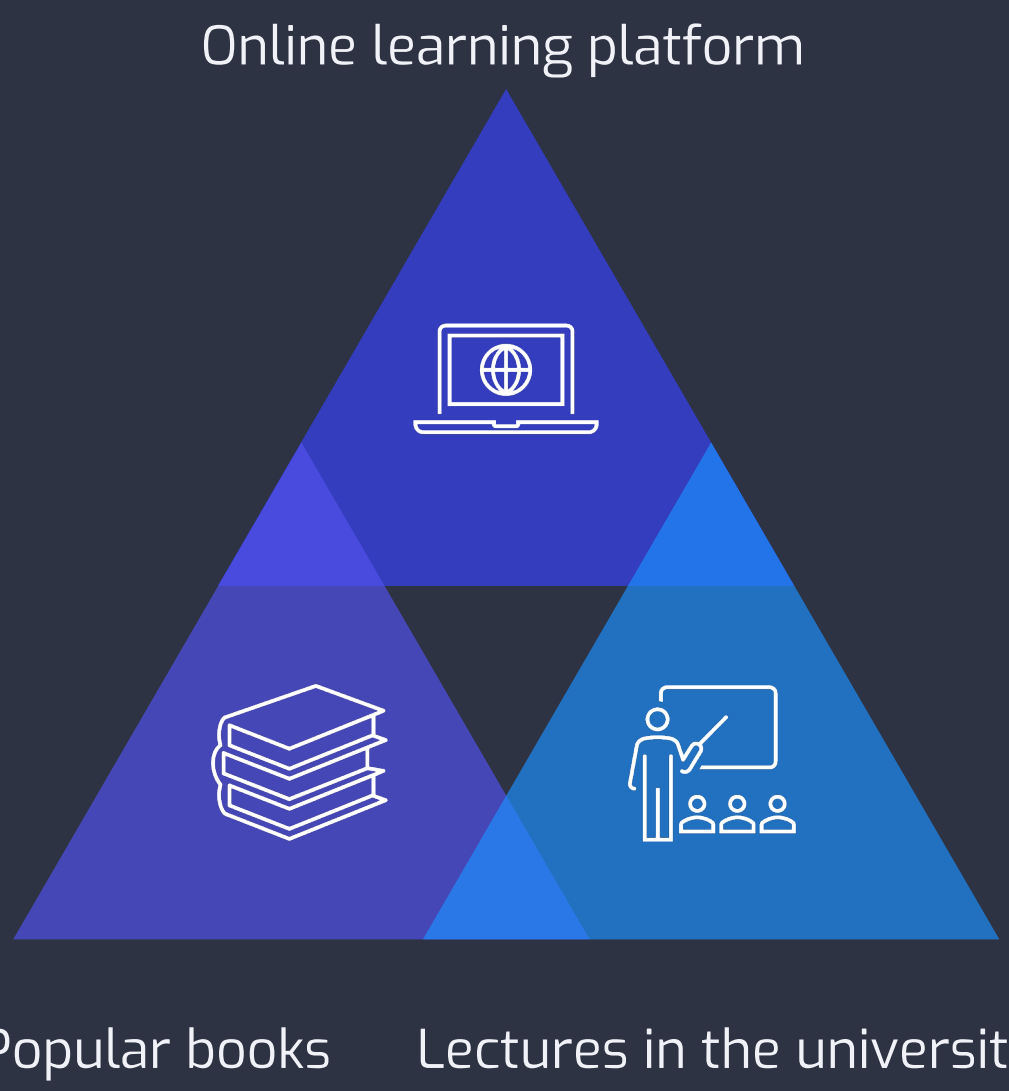
learning lab


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Transformation
Data Collection
Data Sampling
Data Splitting

Learning Objectives

- Understand what is data sampling and why we need it
- Learn techniques for data sampling

EXAMPLE. Learning machine learning



The best way is to pick up only one option, rather than being lost in the “information overload” or spending too much time just to find the “best option” which possibly doesn’t exists.

In Machine Learning

We want the minimum amount of data containing sufficient information required to learn properly from the phenomenon without wasting time.

One way is to reduce information redundancy that doesn't give us useful information or doesn't contain any business value for a particular task.

How can we ensure our sample is not redundant or incomplete, i.e. our sample represents the real word phenomenon?

- High-level comparison between our sample and other sample or the whole population, to check if the chosen sample represents the whole population.
- Instead of learning from a large dataset, we could just make a sub-sampling of it yet keeping all the statistics intact.

Key to sampling, i.e. picking up a small, easy to handle sample...

Ensure the chosen sample does not lose statistical significance with respect to the whole population.

Sample size too small	not carry enough information to learn from.

We must compromise between the size & the information.

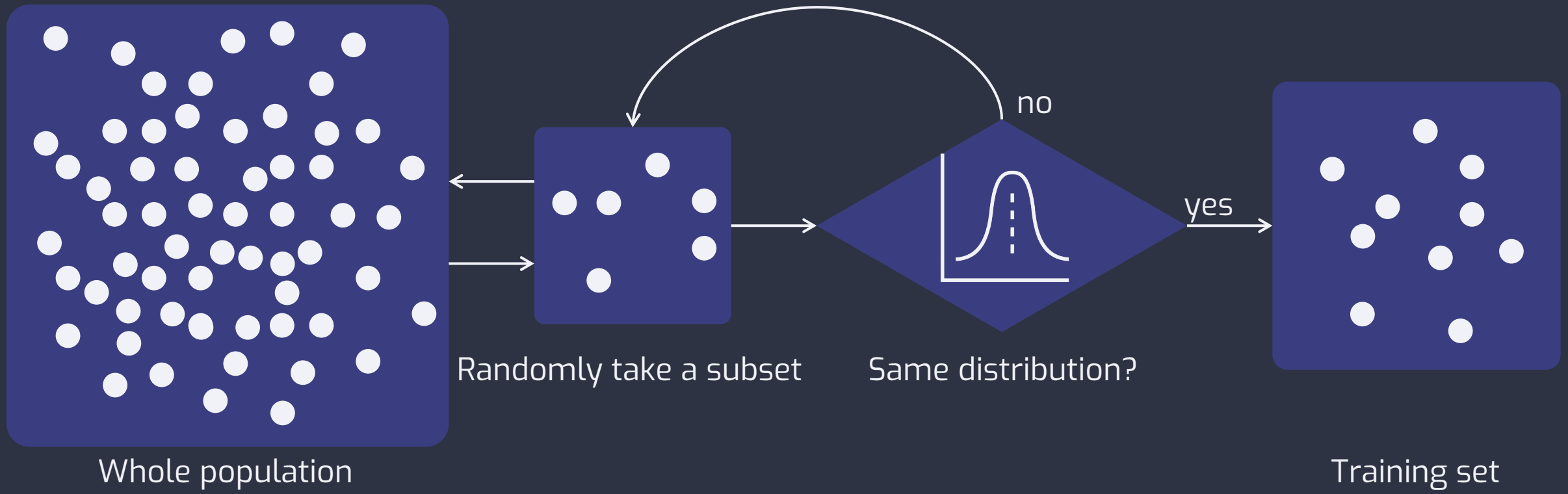
Sample size too large	time-consuming to proceed.

Sample to keep probability distribution of the population under reasonable significant level.

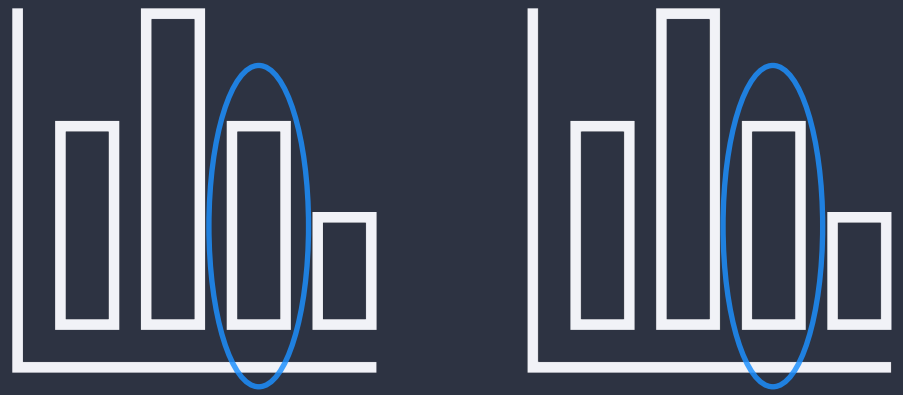
sample whole population

Have a look at the histogram of the sample and ensure it to be the same as the histogram of the whole population.

Statistical framework / Imbalanced data



1. What if our dataset is multivariate, i.e. having N variables (N > 1), especially that they could be a mix of numerical and categorical variables?



Consider each variable independently. If each one of single univariate histograms of the sample's columns is comparable with the correspondent histogram of the whole population's columns, we can assume the chosen sample is not biased.

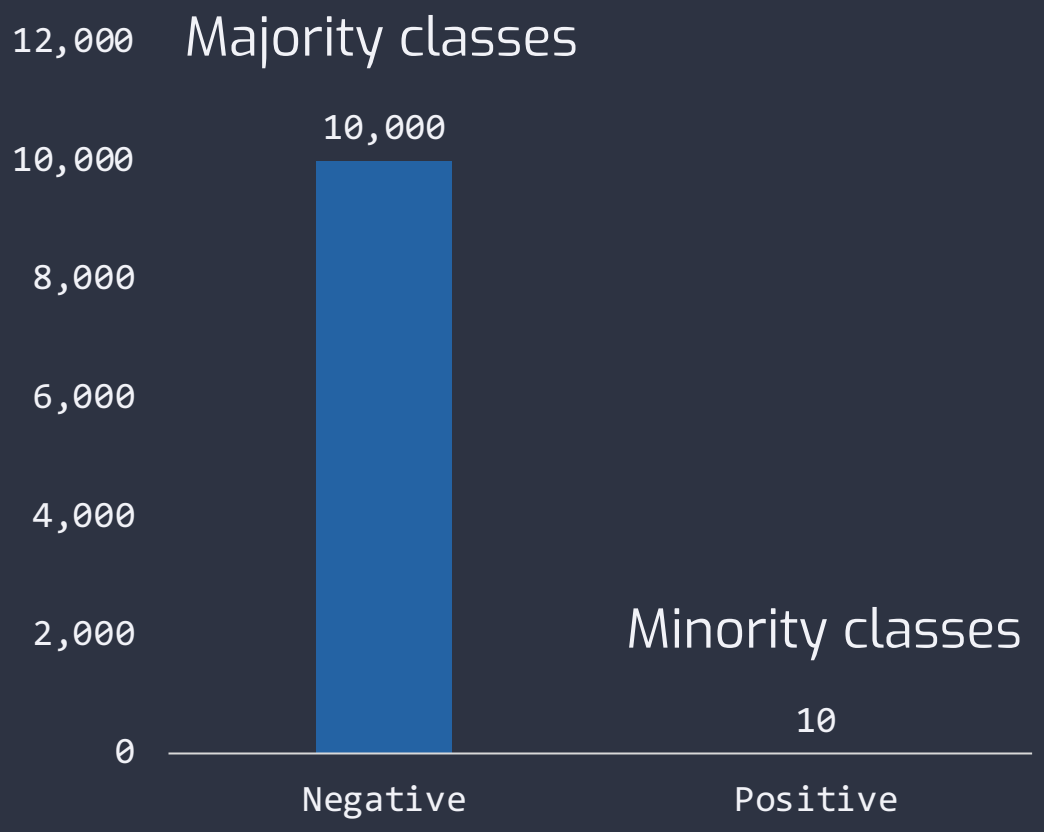
2. How to compare a sample with the whole population?

Categorical Variables	Pearson's chi-square test
Numerical Variables	Kolmogorov-Smirnov test

Test a *null* hypothesis – the frequency distribution of certain instances in a sample is consistent with the whole population of the dataset.

In multivariate case, all the variables need to be significative, and we reject the *null* hypothesis if the p-value of at least one of the tests is low than the usual 5% confidence level, i.e. to accept the sample, we must have all the variables pass the significance test.

EXAMPLE. to classify positive and negative COVID-19 cases



Much more negative instances than the positive ones

The classification dataset is with skewed class proportions – imbalanced.

Imbalanced dataset could be very problematic!

Imbalanced dataset could be very problematic!

Not learnt enough from the minority class

- low efficiency in detecting them
- but, often the minority class is of the most interests

Misleading high accuracy

- excellent accuracy does not necessarily mean the model performs very well
- it may be just reflecting the underlying class distribution

Imbalanced dataset is common!

Most machine learning algorithms for classification were designed under the assumption that there were equal numbers of instances for each class.

However, it is common that our dataset does not have exactly equal number of instances in each class.

e.g. fraud-detection, spam detection, outlier detection, anomaly detection…
claim prediction, default prediction, conversation prediction, churn prediction…

There isn't a clear definition or division, we could roughly have the following categories:

Degree of imbalance	Minority %
mild	20% ~ 40%
moderate	1% ~ 20%
extreme	< 1%

Techniques to combat imbalanced dataset:

As a start: try training on the true distribution, e.g. using the whole dataset

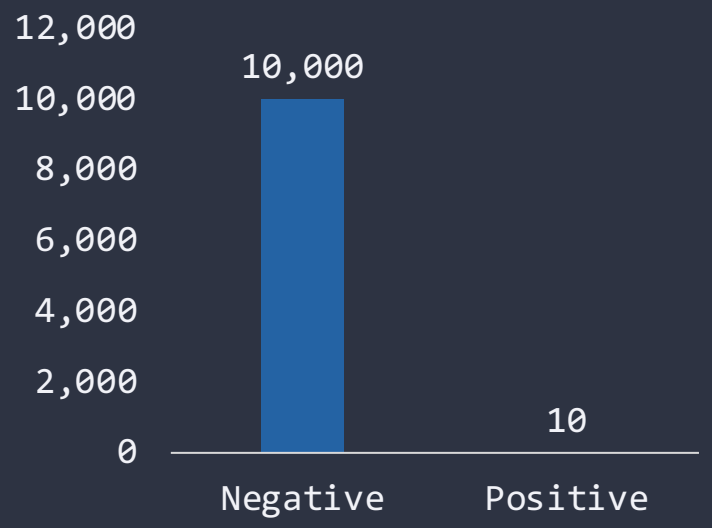
If the trained model works well and generalises well – well done!

If not…

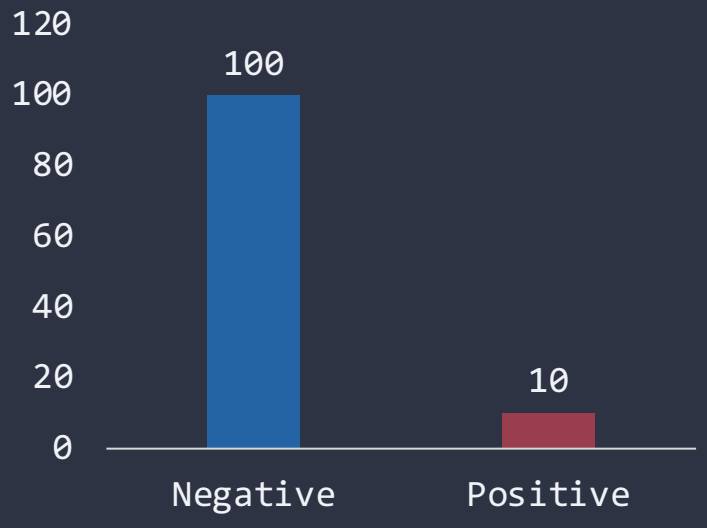
Down-sampling	training model on a disproportionately low subset of the majority class instances.
Up-weighting	adding an instance weight to the down-sampled class equal to the factor by which we down-sampled.

1 Down-sampling

down-sample the majority class, the positive cases in our covid-19 dataset. Since we have 10 positive to 10,000 negatives, we can down-sample by a factor of 100, thus taking one out of a hundred negatives. So now, we have 1:10 instead of 1:1,000, and our sample is less biased, i.e. about 10/(100+10), 9%, instead of 0.1% of our data is positive.



Extremely biased



Less biased

2 Up-weighting

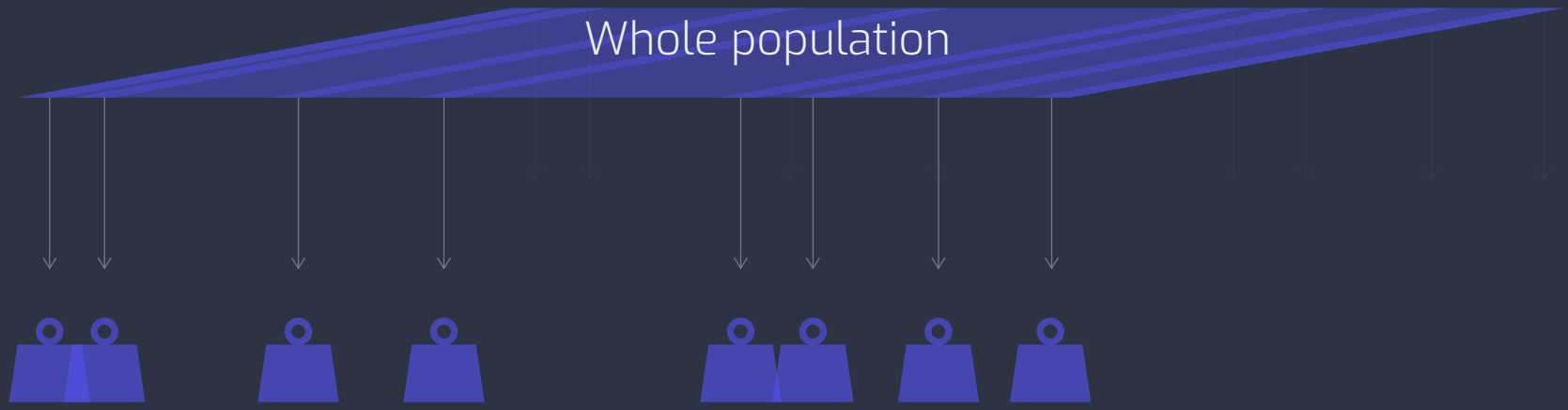
up-weight the down-sampled majority class. As our down-sample factor is 100, the instance weight should also be 100. This way we count an individual negative instance more importantly during training. A weight of 100 means the model treats those negative instances 100 times as important as it would a positive instance with a weight of 1.

Down-sampling

Randomly pick instances from the whole population

Up-weighting

Add weight to the down-sampled instances



Weight should be equal to the factor used for down-sampling:

```
{instance weight} = {original instance weight} x {down-sampling factor}
```

Takeaway Points

- Need to ensure sample is not redundant or incomplete.
- Need to ensure sample does not lose statistical significance with respect to the whole population.
- Statistical framework (categorical / numerical)
- Imbalanced data (down-sampling & up-weighting)