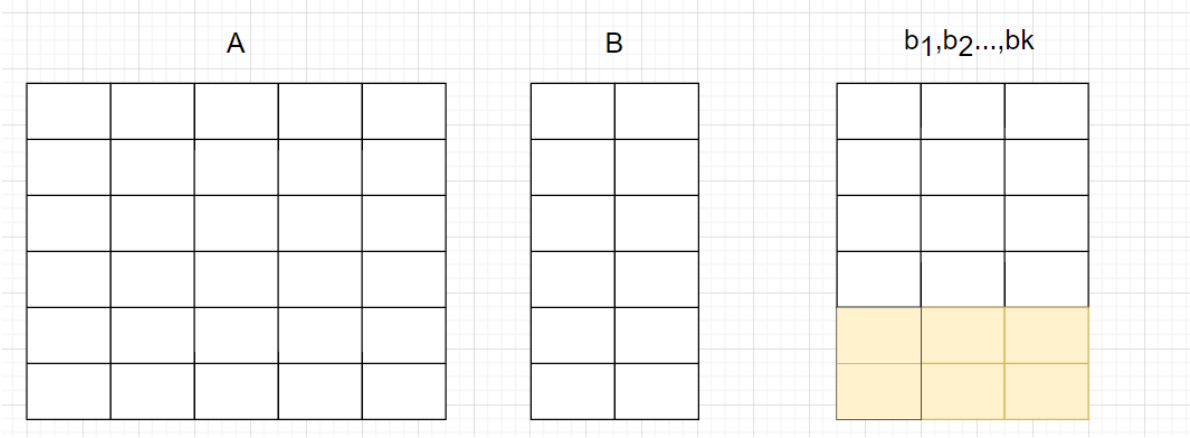
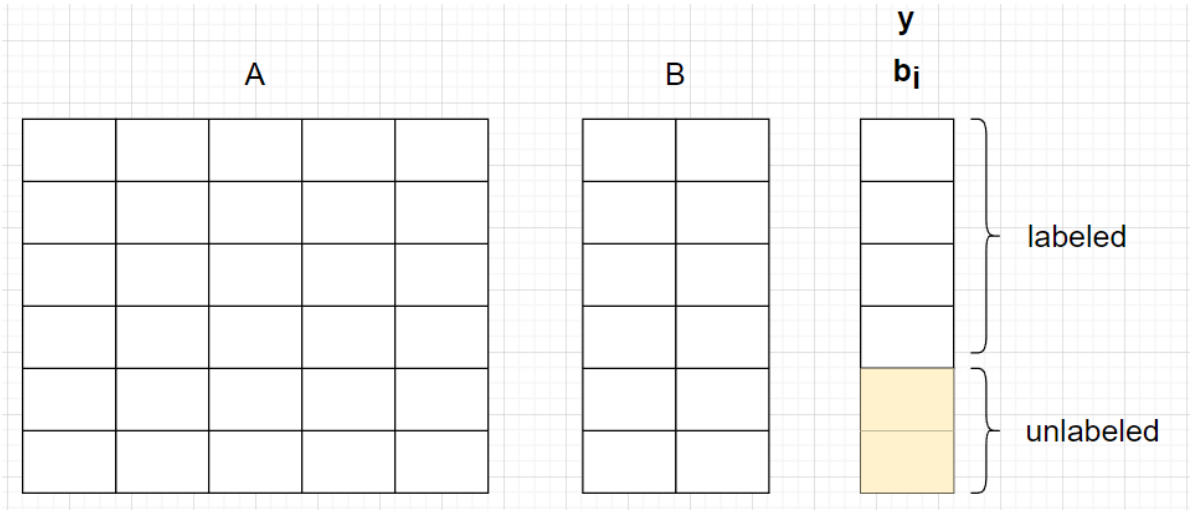
# 问题描述




对于B方特征缺失的第$i$列，$i=1,2,3...,k$，将其作为标签列:

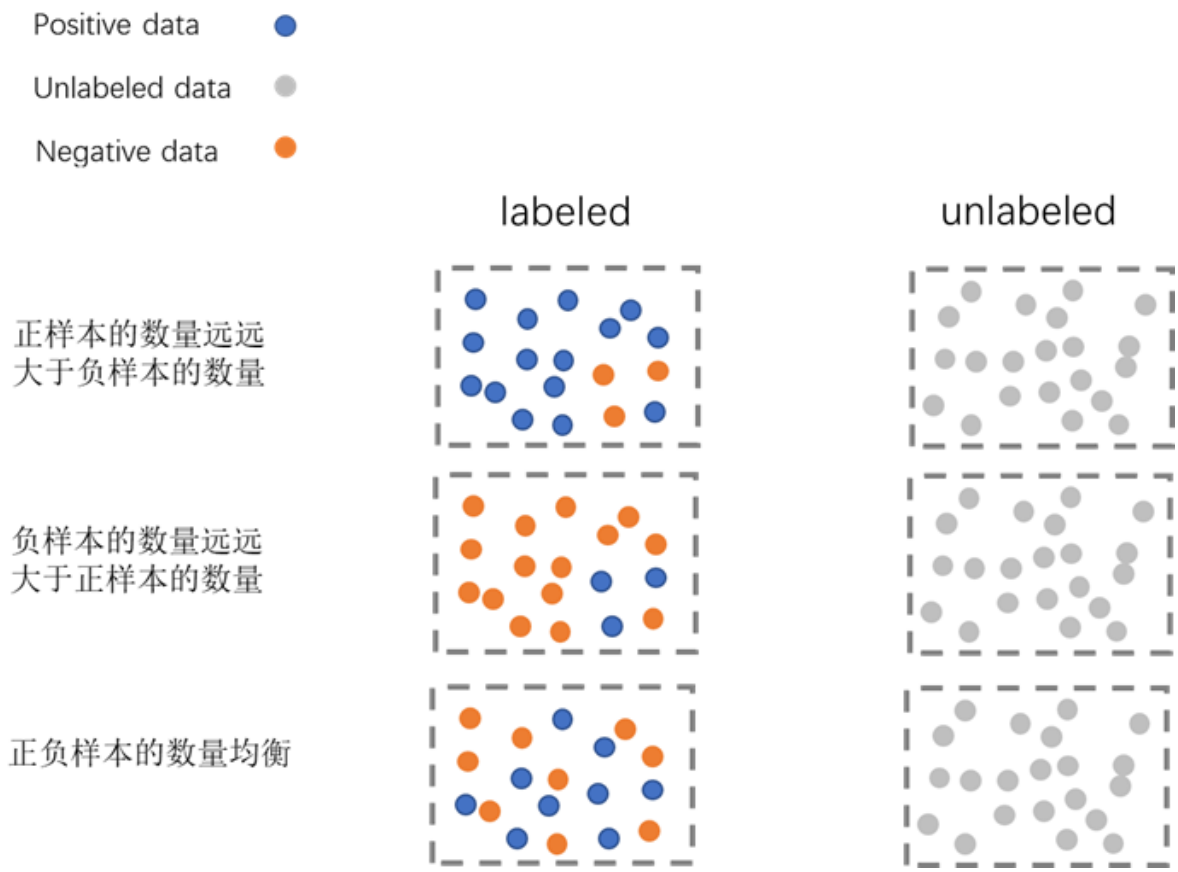



# bi列为只有两个类别的分类列

由于$b_i$列只有两个类别：正样本和负样本，分别用0和1表示，对于未标记样本用-1表示。未标记样本的情况不必多说，始终都是-1，主要看有标签的部分，正样本和负样本的分布比例，一共有三种情况:

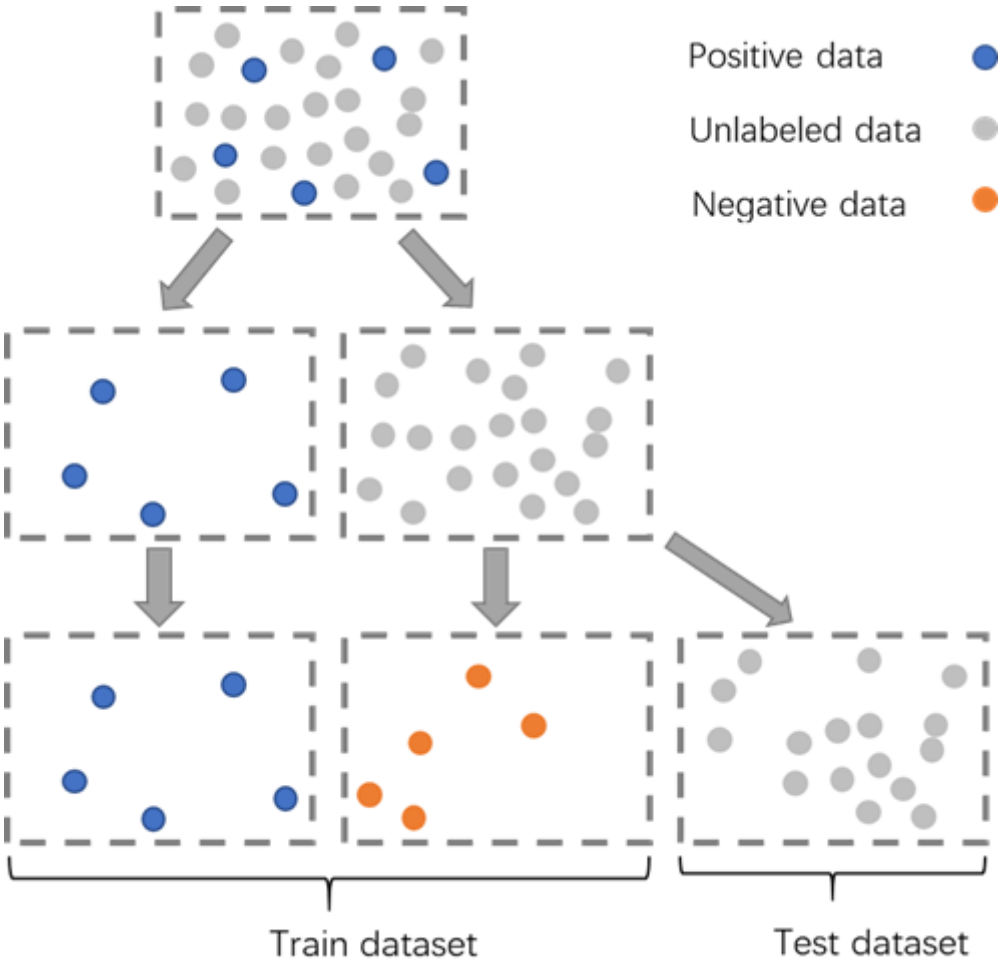
1. 正样本的数量远远大于负样本的数量
2. 负样本的数量远远大于正样本的数量
3. 正负样本的数量均衡

这三种情况对应的图示如下:




对于第一种情况，正样本数量远远大于负样本数量，符号VFPU算法的数据形式，由于正样本数量远远大于负样本数量，导致无法形成以一个正负样本均衡的数据集，所以采用随机抽样的方式，从未标记样本中抽取部分样本，作为负样本，形成一个正负样本均衡的数据集，让后交给纵向的分类器进行训练，未标记样本中没有被采样到的样本作为测试集，，并预测每个样本为正样本的概率，重复执行多次，最会对每个未标记样本为正样本的概率求平均作为最终的分数，取排名靠前的样本，将其标签作为正样本标签。

随机抽样，形成正负样本均衡的数据集方式如下:




假设有正样本的数量为$|P|$，负样本的数量为$|N|$。在原始的VFPU算法中，初始数据集$|N| = 0$，因此每次会从未标记样本中随机抽取与正样本数量相同的样本作为负样本，形成一个正负样本均衡的二分类数据集，作为训练集。未被随机抽样到的样本则作为测试集。训练数据集被送到纵向联邦分类器进行训练，并在测试集上进行测试。

现在，面临的问题是$|P|$和$|N|$之间的大小关系可能存在以下三种情况：

1. $|P| >> |N|$
2. $|P| << |N|$
3. $|P| == |N|$

目标是优化VFPU的随机采样步骤。在原始的VFPU中，是从未标记样本中随机抽取与正样本数量相同的样本，并将这些样本作为负样本。而在优化后的步骤中，打算这样做：

1. 仍然从未标记样本中进行采样，但是采样的目的是构造一个正负样本均衡的二分类数据集。
2. 采样的数量应为$||P| - |N||$，即正样本和负样本数量的差的绝对值。
3. 应该将采集到的未标记样本的标签设置为样本数量较少的一方，以此作为补偿。

这样的优化可以使得在训练模型时，更好地处理样本不均衡的问题，提高模型的性能。

sampling rate of reliable positive samples.

**Output:** $R$, a set of reliable positive samples.

1: **procedure** PARTY C EXECUTES
2: **for** $m = 1, 2, \ldots, M$ **do** 第一层循环
3: $P_m = \{i | \mathcal{Y}_i^C = 1,\ i \in \mathcal{I}_C\}$
4: $U_m = \{i | \mathcal{Y}_i^C = -1,\ i \in \mathcal{I}_C\}$
5: **for** $t = 1, 2, \ldots, T$ **do** 第二层循环，随机采样，优化
6: $N_m^t = \{Randomly\ select\ |P_m|\ elements\ from\ U_m\}$
7: $O_m^t = U_m - N_m^t$
8: Encrypt and send $N_m^t$, $P_m$, and $O_m^t$ to other parties.
9: Notify parties to set training data and testing data.
10: $S_m^t$ = Base_Estimator_Learning()