**Introduction:**

When it comes to applying computer vision in the medical field, most tasks involve either 1) image classification for diagnosis or 2) segmentation to identify and separate lesions. However, in pathology cancer detection, this is not always possible. Obtaining labels is time-consuming and labor-intensive. Furthermore, pathology slides can be up to 200k x 100k pixels resolution, and they will not fit in memory for classification since for example, the ImageNet CNN models only use 224 x 224 pixels for training. Downsampling normally is not an option because we are trying to detect a tiny area, such as a cancerous area varying from 300 x 300 pixels area.  
Suppose we have pathology slides and the label for each slide. Since we cannot train the classifier on the whole slide, we divide each slide into tiles or patches and only process a few of them at a time on GPU. However, we do not know labels for each patch, so we need Multiple Instance Learning. In the MIL framework, slides are the “bag” and patches are “instances”. By using it, we are able to save the labeling effort and leverage weakly labeled data. When we have pathology slides from patients and we want to predict if a large slide contains cancer cells or, zooming out if the patient has malignant cells, Multiple Instance Learning is a good option because doctors don’t need to segment individual cells or label each patch. Only the whole slide needs a label.

In general, Multiple Instance Learning can deal with classification problems, regression problems, ranking problems, and clustering problems, but we will mainly focus on classification problems.

**Technical Test:**

1. **MIL-MNIST Dataset Construction:**
   * Utilize the MNIST dataset, a standard benchmark for classification tasks.
   * Transform it into a Multiple Instance Learning (MIL) problem by grouping several digits into a bag.
   * Assign bag labels: If a bag contains at least one instance labeled as "1", the bag label is "1"; otherwise, if all instance labels are "0-9" except "1", the bag label is "0".

As the graph shown below, the bag filled with red has a bag label “1” and the bag filled with blue has a bag label “0”.

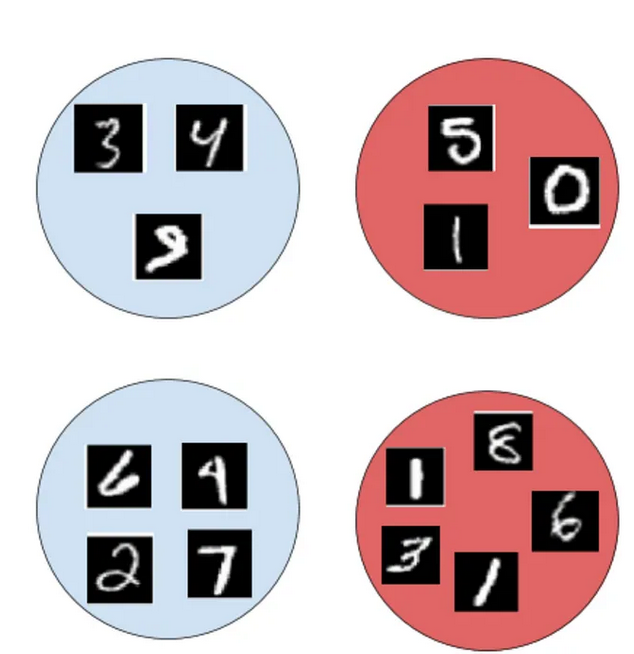


Figure 1: bags and instance labels

# 

1. **Dataset Generation and evaluation:**
   * Generate a dataset with the following parameters:
     + Target number: 1
     + Bag length: 64
     + Number of bags for training: 100
     + Number of bags for testing: 50
   * Implement the classical weakly supervised pipeline described in the attached paper to evaluate the generated dataset.
   * For feature generation, simply convert each image into a vector.
   * Utilize a simple classification model as follows:

