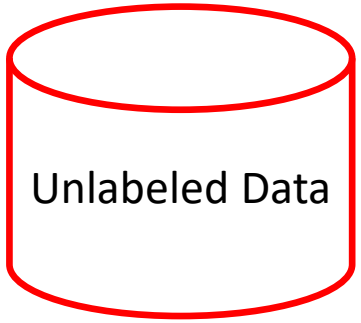
An Active Learning Approach for Botnet Detection

Mehrdad Hajizadeh

TU-Chemnitz

Problem Formulation



Supervised Learning
Deep Learning

For many more sophisticated supervised learning tasks, labeled instances are very difficult, time-consuming, or expensive to obtain.

Active Learning

- There are several scenarios in which active learners may pose queries, and there are also several different query strategies that have been used to decide which instances are most informative.
- Fig 1: A learner may begin with a small number of instances in the labeled training set L, request labels for one or more carefully selected instances, learn from the query results, and then leverage its new knowledge to choose which instances to query next. Once a query has been made, there are usually no additional assumptions on the part of the learning algorithm. The new labeled instance is simply added to the labeled set L, and the learner proceeds from there in a standard supervised way.

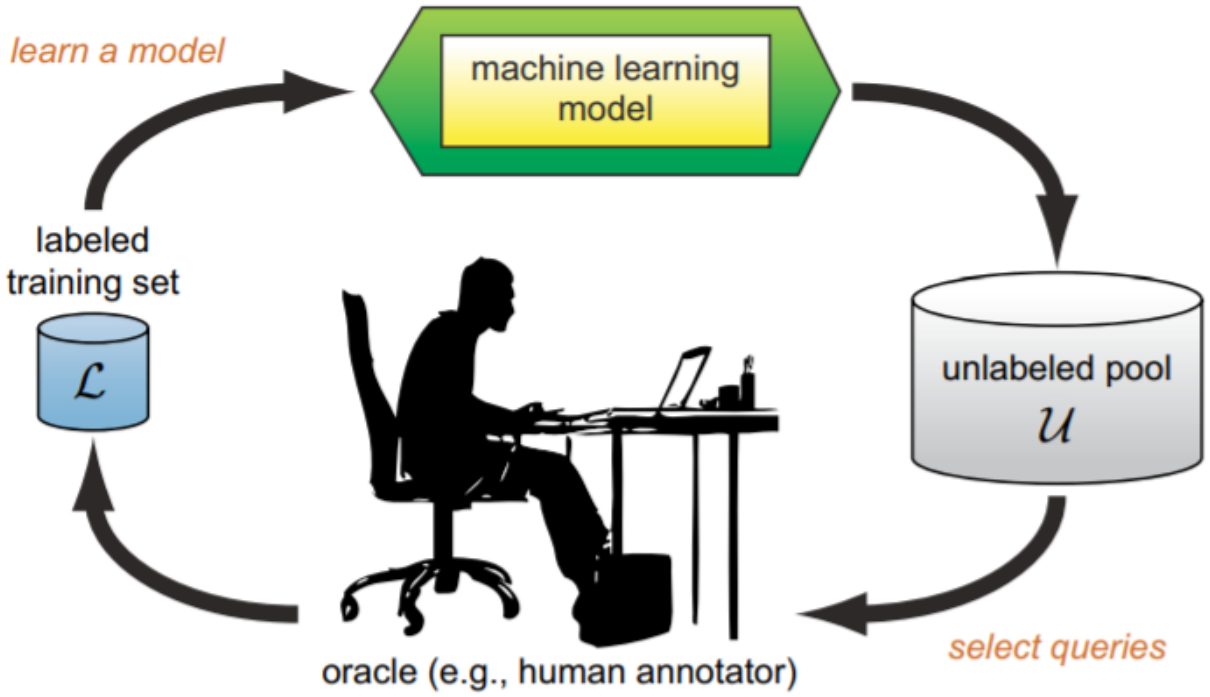


Figure 1: The pool-based active learning cycle.

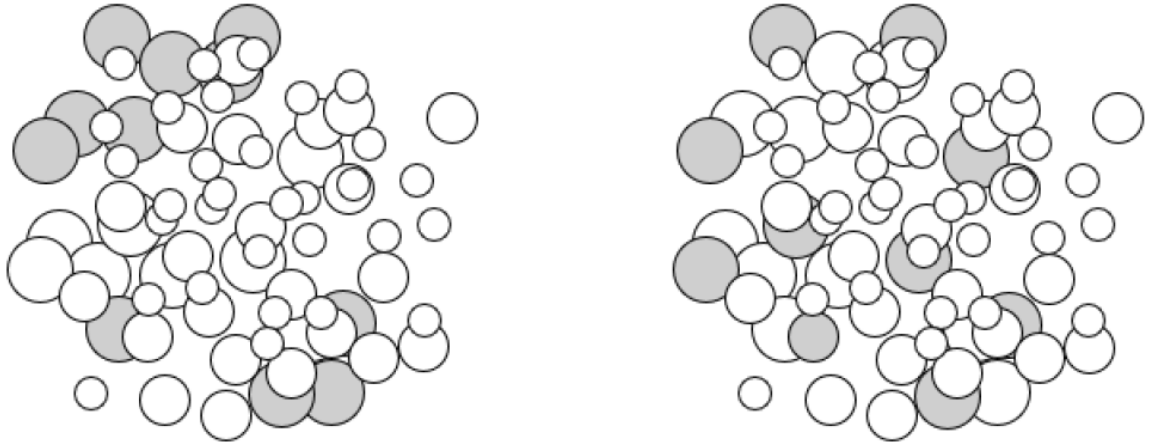
http://burrsettles.com/pub/settles.activelearning.pdf

Typical Active Learning Scenarios

- The classical scenario
  - specifes the number of labels provided by oracle.
- The cost-sensitive active learning scenario involves query cost and misclassifcation costs.
  - The learning task is to achieve a trade-of between them to minimize the total cost.
- The active learning from data streams scenario involves dynamic nature of the data.
  - The learner should query instances based on the data observed so far to maximize the classifcation accuracy on future instances

Batch Query Strategy

The information provided to the model by such similar samples is essentially the same, which not only wastes labeling resources, but also makes it difficult for the model to learn genuinely useful information. Therefore, it is important to query a set of samples that are both information-rich and diverse.



(a) Batch query strategy considering only the amount of information.

(b) Batch query strategy considering both information volume and diversity.

Fig. 2. A comparison diagram of two batch query strategies, one that only considers the amount of information and one that considers both the amount and diversity of information. The size of the dots indicates the amount of information in the samples, while the distance between the dots represents the similarity between the samples. The points shaded in gray indicate the sample points to be queried in a batch.

Paper:A Survey of Deep Active Learning

Active Learning Categories based on Query

- **Stream-based selective sampling**
  - In this scenario, the algorithm determines if it would be beneficial enough to query for the label of a specific unlabeled entry in the dataset. While the model is being trained, it is presented with a data instance and immediately decides if it wants to query the label. This approach has a natural disadvantage that comes from the lack of guarantee that the data scientist will stay within budget.
- **Pool-based sampling**
  - This is the most well known scenario for active learning. In this sampling method, the algorithm attempts to evaluate the entire dataset before it selects the best query or set of queries. The active learner algorithm is often initially trained on a fully labeled part of the data which is then used to determine which instances would be most beneficial to insert into the training set for the next active learning loop. The downside of this method is the amount of memory it can require.
- **Membership query synthesis**
  - This scenario is not applicable to all cases, because it involves the generation of synthetic data. The active learner in this method is allowed to create its own examples for labeling. This method is compatible with problems where it is easy to generate a data instance.

https://algorithmia.com/blog/active-learning-machine-learning

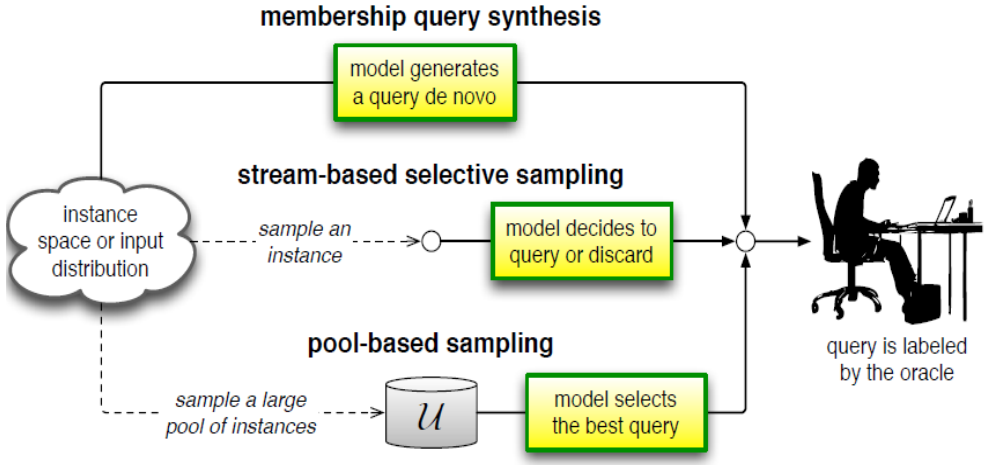
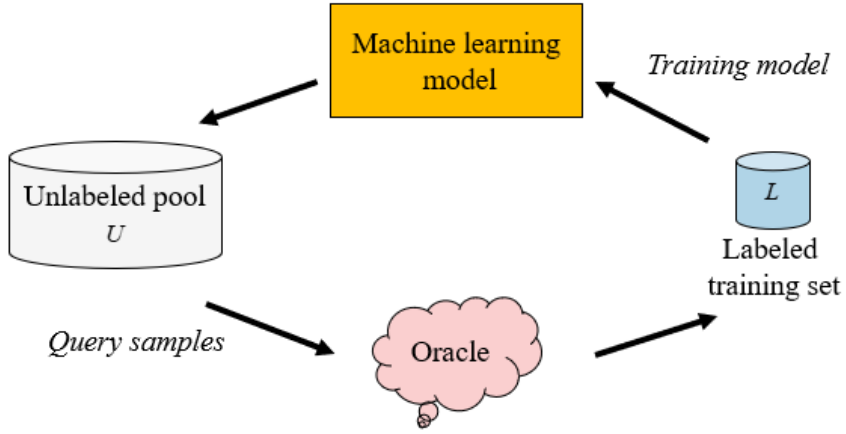


Diagram illustrating the three main active learning scenarios

http://burrsettles.com/pub/settles.activelearning.pdf



The pool-based active learning cycle

Ref: A Survey of Deep Active Learning

Pool Based vs Stream Based

- The main difference between stream-based and pool-based active learning is that the former scans through the data sequentially and makes query decisions individually, whereas the latter evaluates and ranks the entire collection before selecting the best query.
- While the pool-based scenario appears to be much more common among application papers, one can imagine settings where the stream based approach is more appropriate. For example, when memory or processing power may be limited, as with mobile and embedded devices.

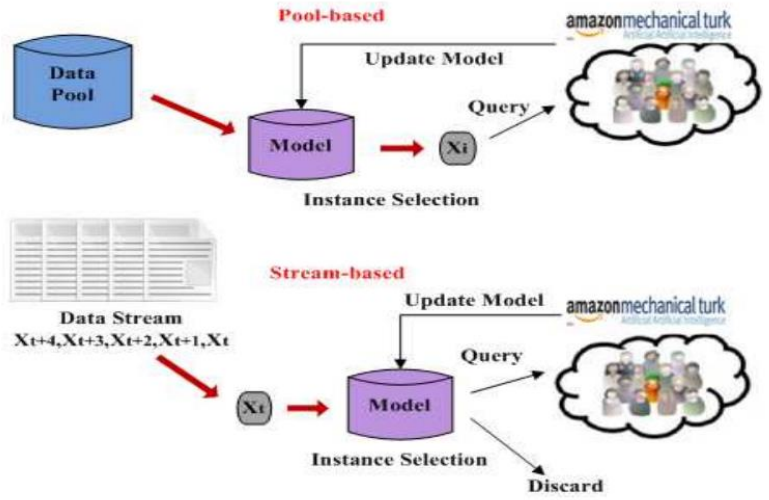


Figure 1: Stream-based active learning vs pool-based active learning.

Fig from Paper: Feedback-Driven Multiclass Active Learning for Data Streams

Query Strategy Frameworks

- Uncertainty Sampling
  - an active learner queries the instances about which it is least certain how to label. This approach is often straightforward for probabilistic learning models. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5
  - Uncertainty sampling strategies may also be employed with non-probabilistic classifiers.
    - decision tree classifier, nearest-neighbor, SVM (that involves querying the instance closest to the linear decision boundary)
- Query-By-Committee
  - The QBC approach involves maintaining a committee C = {clf1,clf2,…clf(n)} of models which are all trained on the current labeled set L, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.
- Expected Model Change
  - It uses a decision-theoretic approach, selecting the instance that would impart the greatest change to the current model if we knew its label.
  - E.g, “expected gradient length” (EGL) where the learner should query the instance x which, if labeled and added to L, would result in the new training gradient of the largest magnitude.
  - The intuition behind this framework is that it prefers instances that are likely to most influence the model (i.e., have greatest impact on its parameters), regardless of the resulting query label.
  - **Calculate gradient**? Read page 18, 19. Also see TensorFlow document

http://burrsettles.com/pub/settles.activelearning.pdf

Query Strategy Frameworks-Continue

- Expected Error Reduction
- Variance Reduction
- Density-Weighted Methods

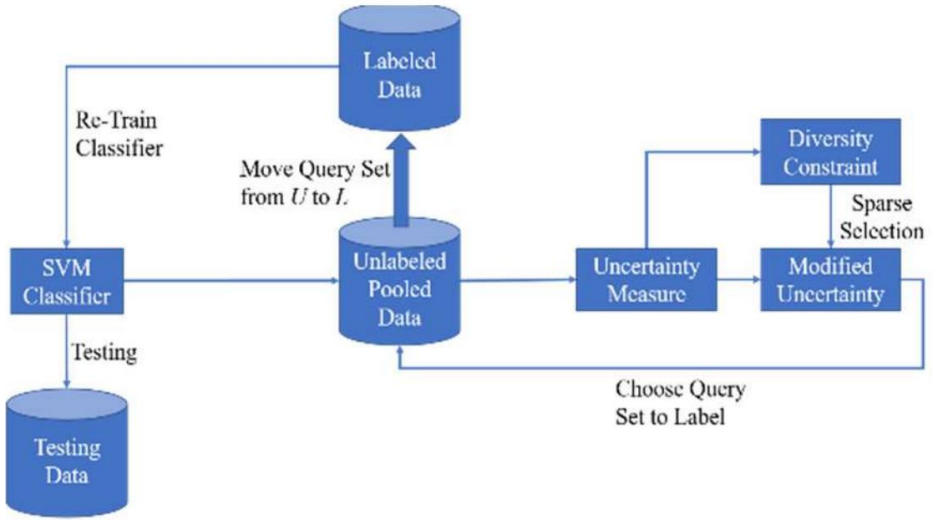
http://burrsettles.com/pub/settles.activelearning.pdf

Paper: Uncertainty Sampling Based Active Learning with Diversity Constraint by Sparse Selection

- Problem: Choose uncertainty samples to be queried. This paper considered:
- Uncertainty Measure Design
  - Choose the most uncertain samples from the unlabeled data pool to label. They used SVM to measure uncertainty score, see formula
- Active Learning via Sparse Selection Modeling
  - Given uncertainty scores generated from the classifiers, we would like to select the most informative samples for a query. The simplest selection strategy is that we always select the samples up to the batch size φ with the largest uncertainty scores.
  - However, this strategy ignores the relations among the pooled unlabeled samples. Sometimes the samples with top uncertainties are very similar to each other. Selecting all of such samples may obtain the same information as selecting just one sample. In other words, similar samples should not be selected at the same batch and diverse samples are needed for a query.
  - we want to make sure that only one sample among similar samples have large uncertainty and the uncertainties from other samples become smaller.

How to find out closest samples to svm hyperplane?

This link helps



https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8122269&tag=1

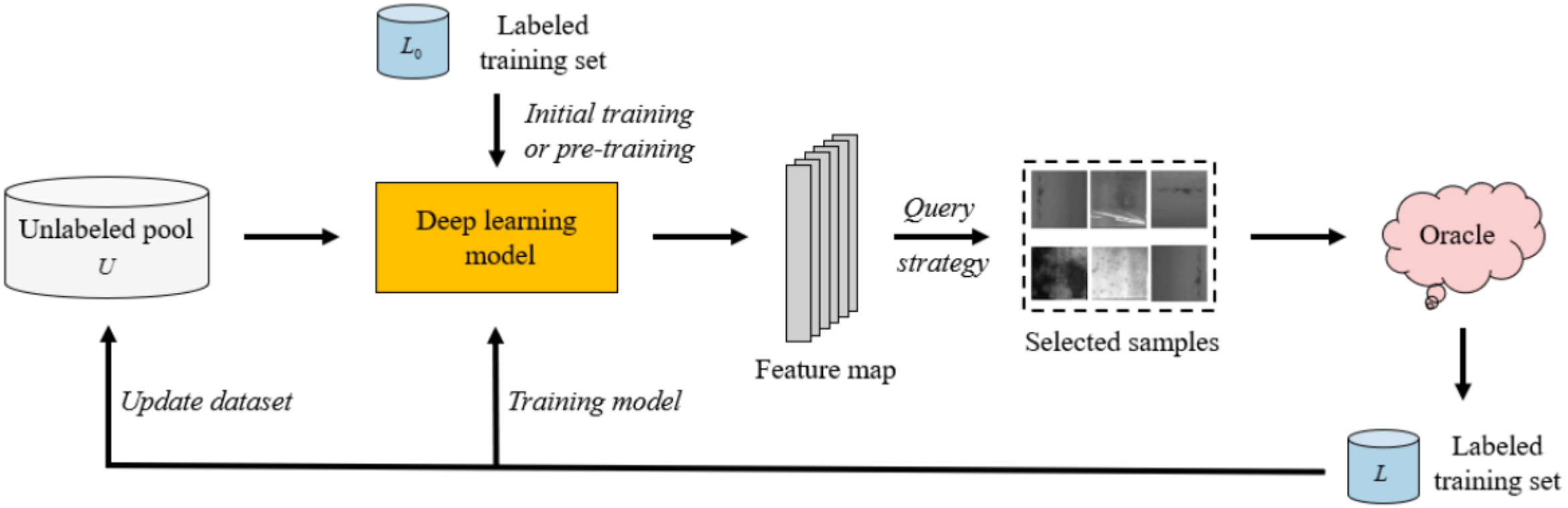
Emprical Analysis of Acrive Learning

- An important question is: “does active learning work?” Most of the empirical results in the published literature suggest that it does (e.g., the majority of papers in the bibliography of this survey).
- Furthermore, consider that software companies and large-scale research projects such as CiteSeer, Google, IBM, Microsoft, and Siemens are increasingly using active learning technologies in a variety of real world applications2

Real Challenges in practice

- As a result, the research question for active learning has shifted in recent years to "**can machines learn more economically if they ask questions**?"

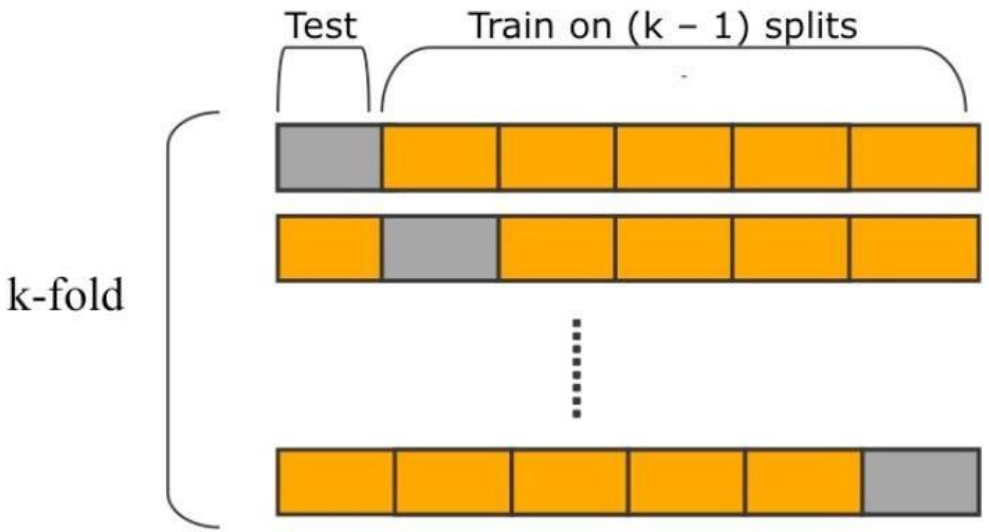
Deep Active Learning



A typical example of deep active learning

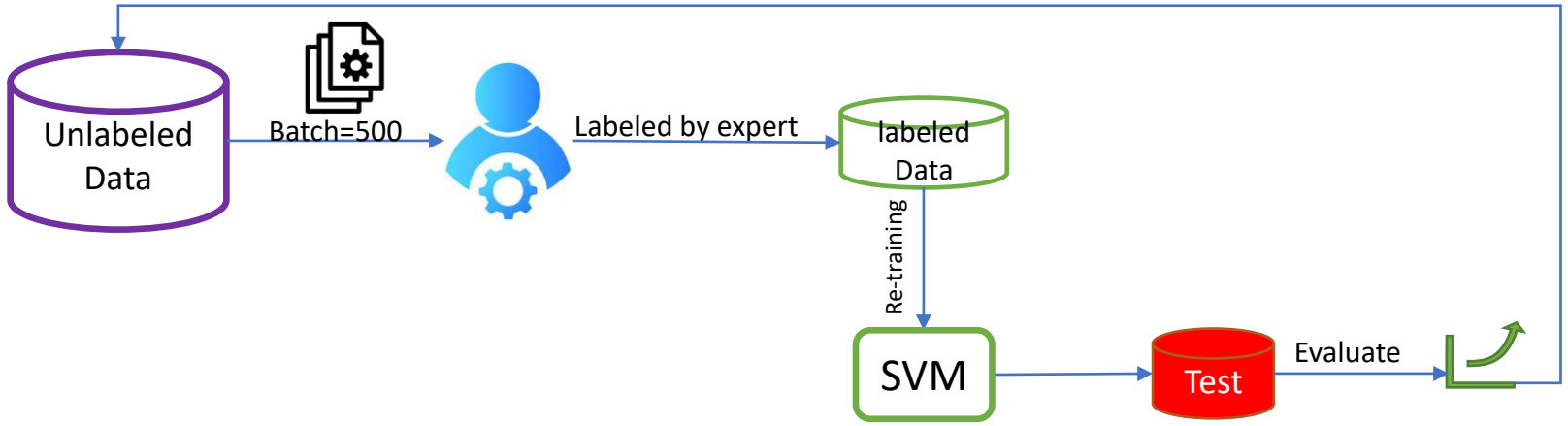
Paper:A Survey of Deep Active Learning

Our Scenario

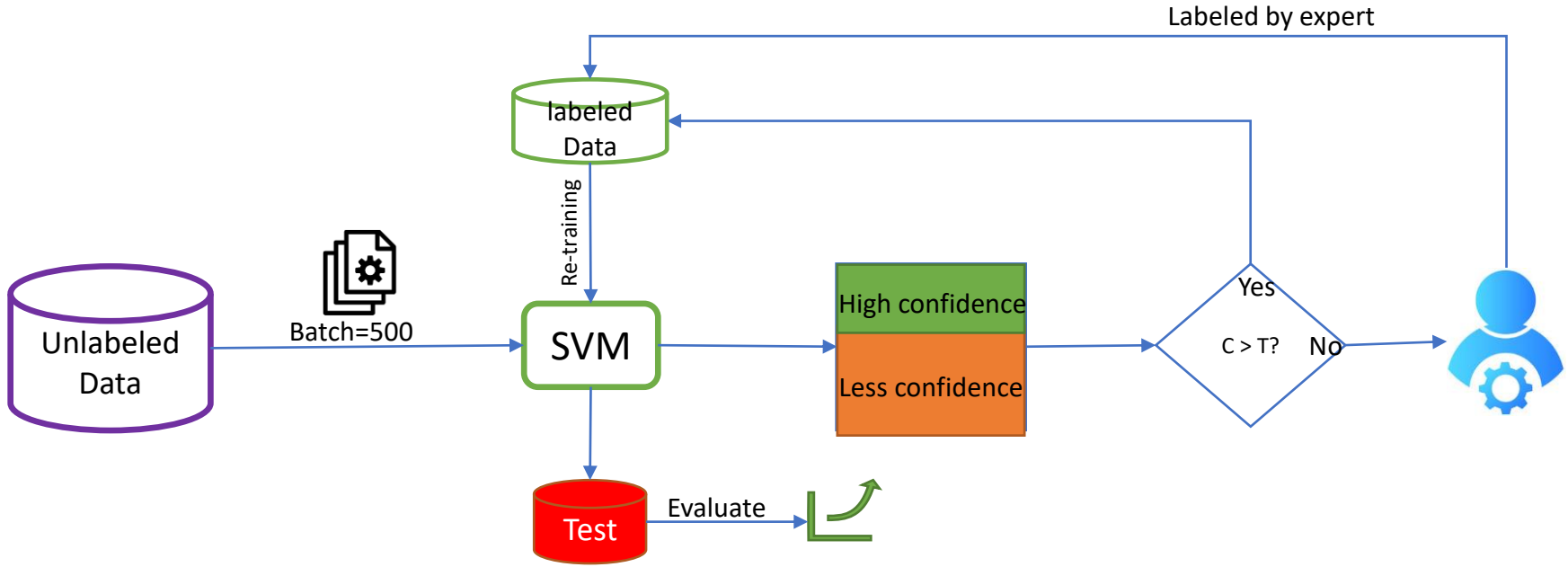


Active Learning with SVM

Passive Learning (Random Sampling)



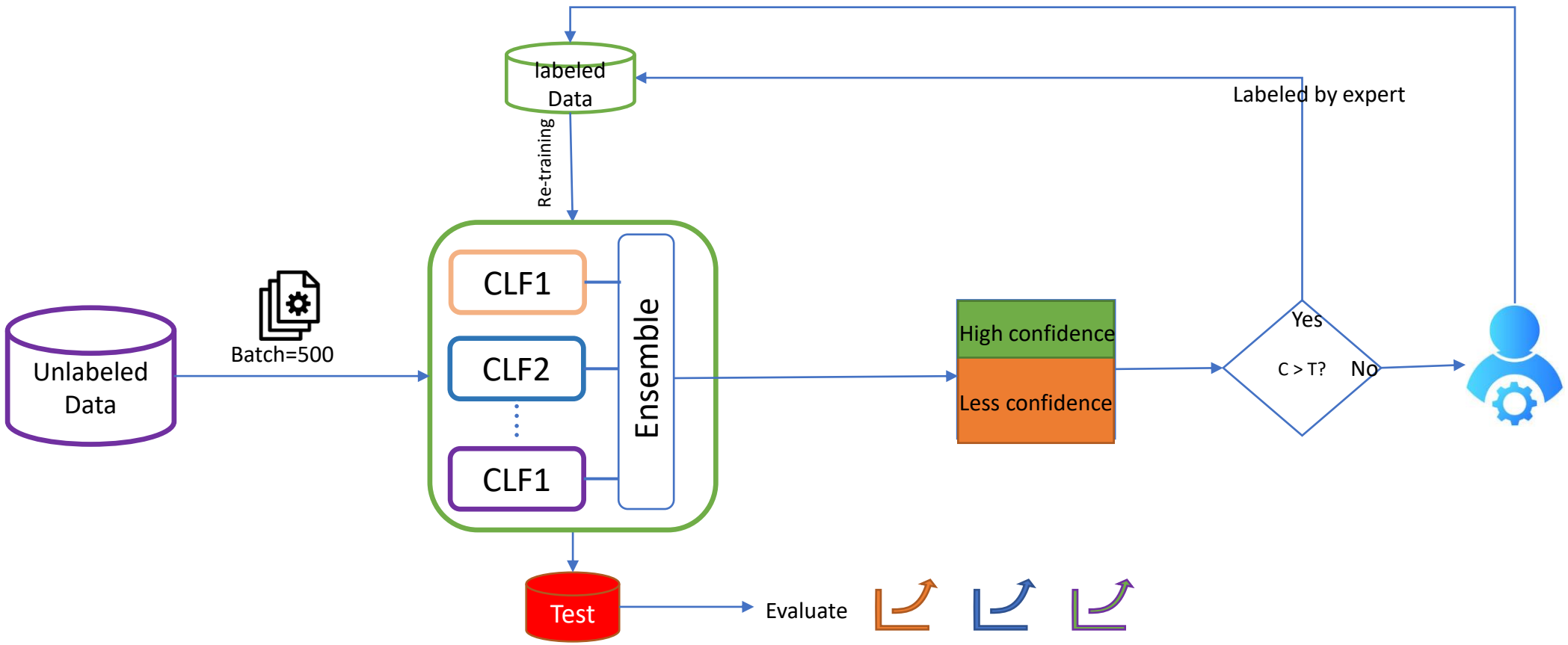
Active Learning (Uncertainty Sampling)



Active Learning via Ensmeble Learning

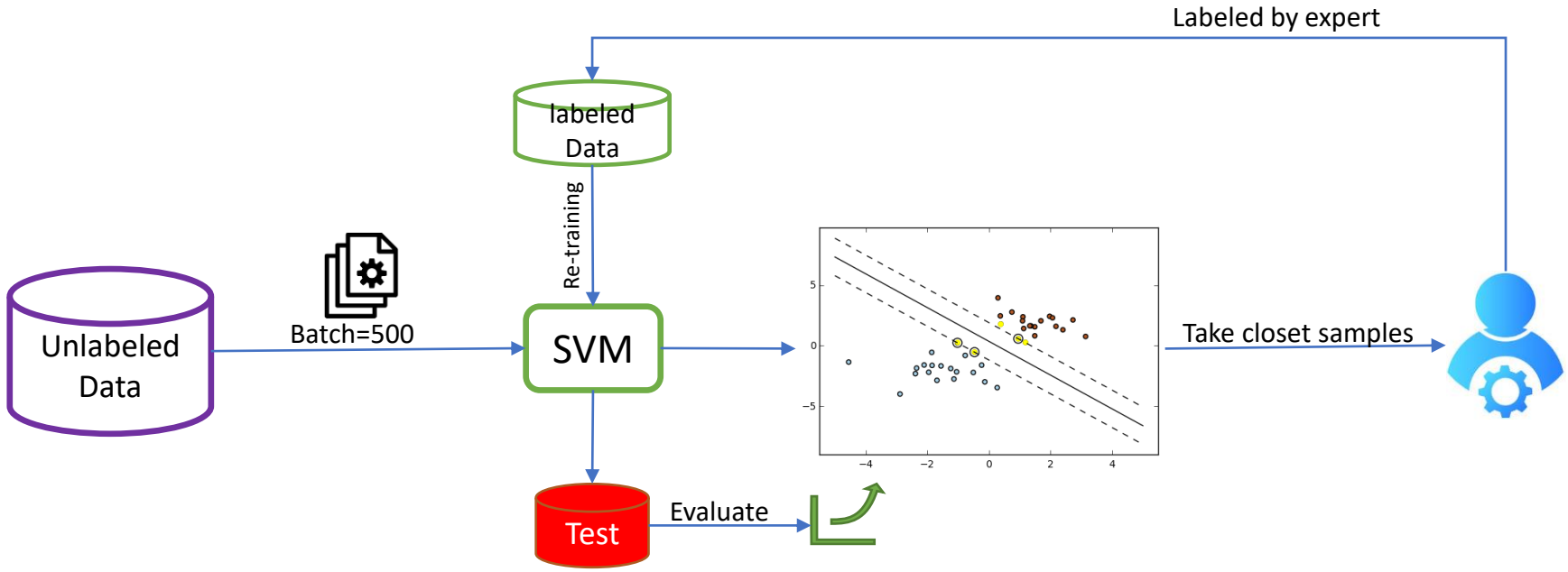
Active Learning (Uncertainty Sampling)

- Query-By-Committee
  - The QBC approach involves maintaining a committee C = {clf1,clf2,…clf(n)} of models which are all trained on the current labeled set L, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.



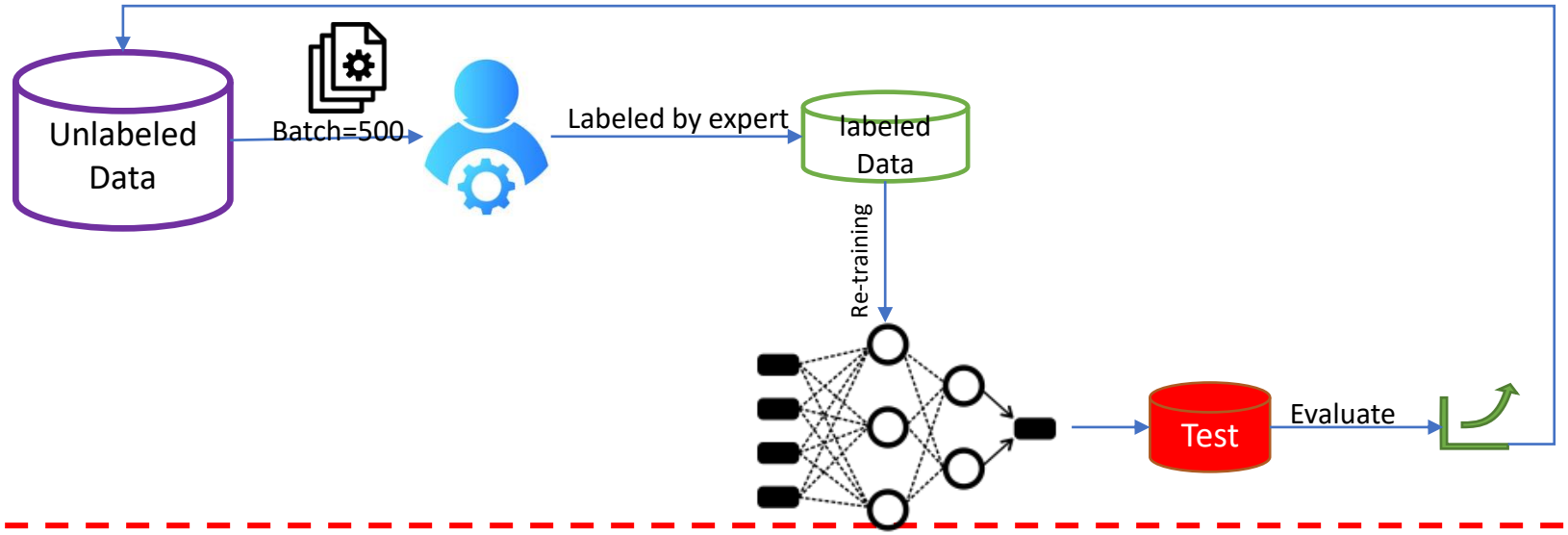
Active Learning with SVM

Active Learning (Sampling of top datapoints closest to the hyperplane) : Use decision Function

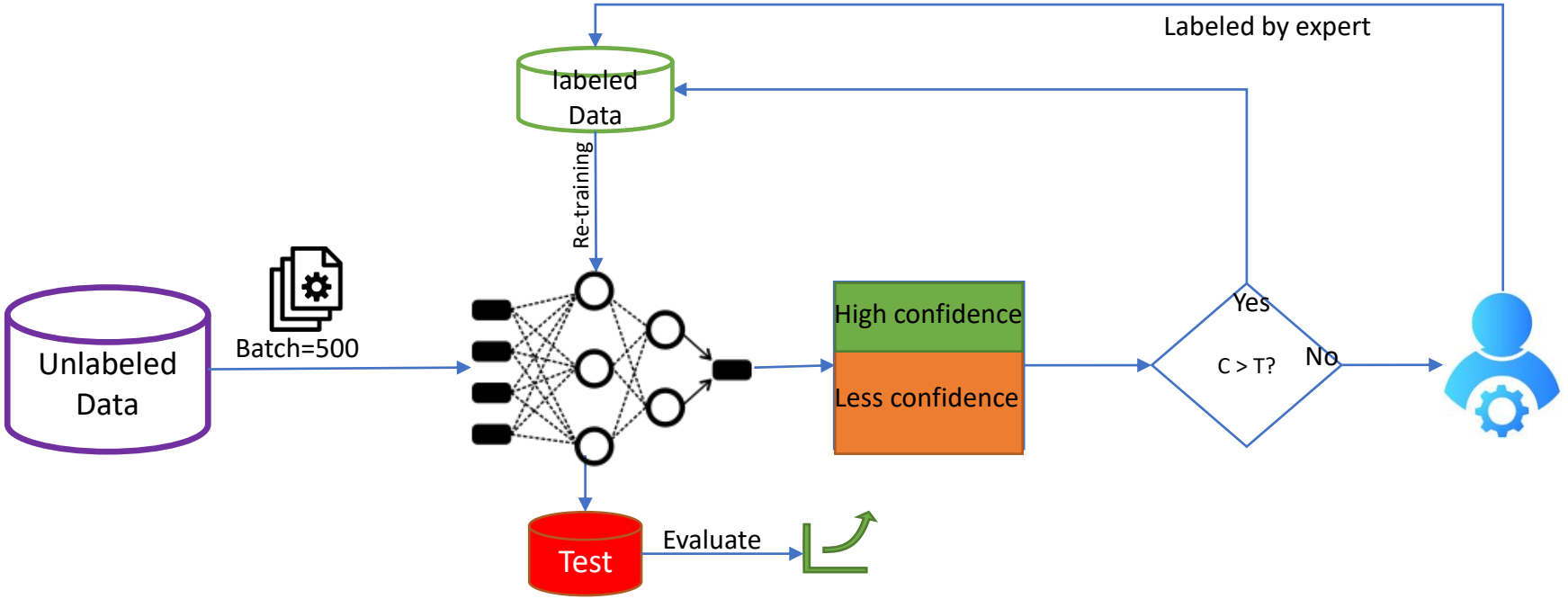


Deep Active Learning with CNN

Passive Learning (Random Sampling)

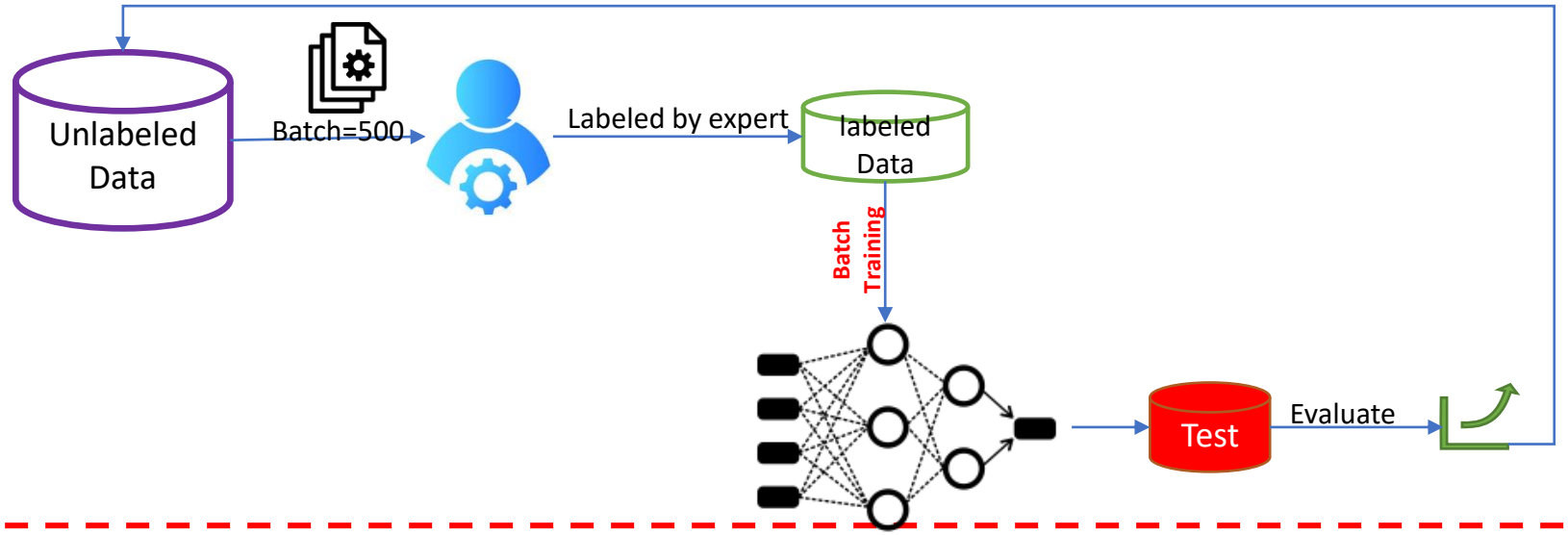


Active Learning (Uncertainty Sampling)



Deep Active Learning with CNN: Batch Training

Passive Learning (Random Sampling)



Active Learning (Uncertainty Sampling)

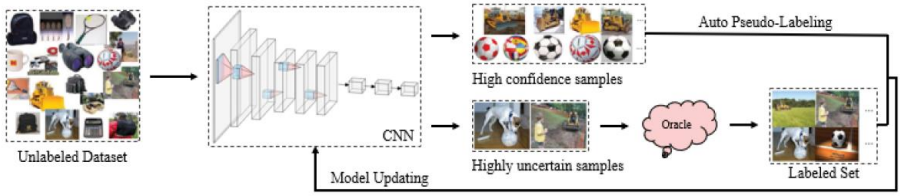


Fig. 4. In CEAL [166], the overall framework of DAL is utilized. CEAL [166] gradually feeds the samples from the unlabeled dataset to the initialized CNN, after which the CNN classifier outputs two types of samples: a small number of uncertain samples and a large number of samples with high prediction confidence. A small number of uncertain samples are labeled through the oracle, and the CNN classifier is used to automatically assign pseudo-labels to a large number of high-prediction confidence samples. These two types of samples are then used to fine-tune the CNN, and the updated process is repeated.

Paper: A Survey of Deep Active Learning