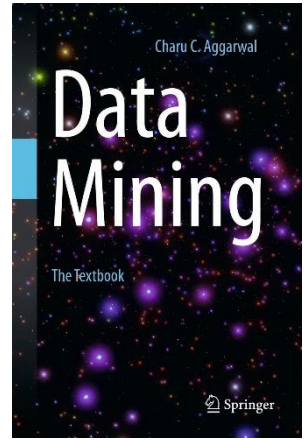
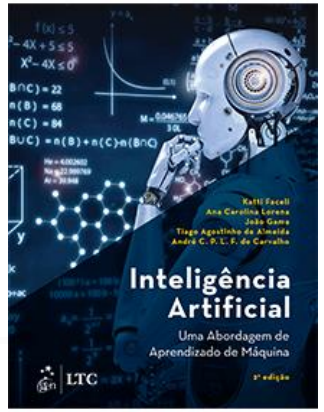
Active Learning Literature Survey
https://burrsettles.com/pub/settles.activelearning.pdf



Aprendizado Ativo

Introdução



The goal is to maximize the accuracy of classification at a specific cost of label acquisition

It integrates label acquisition and model construction

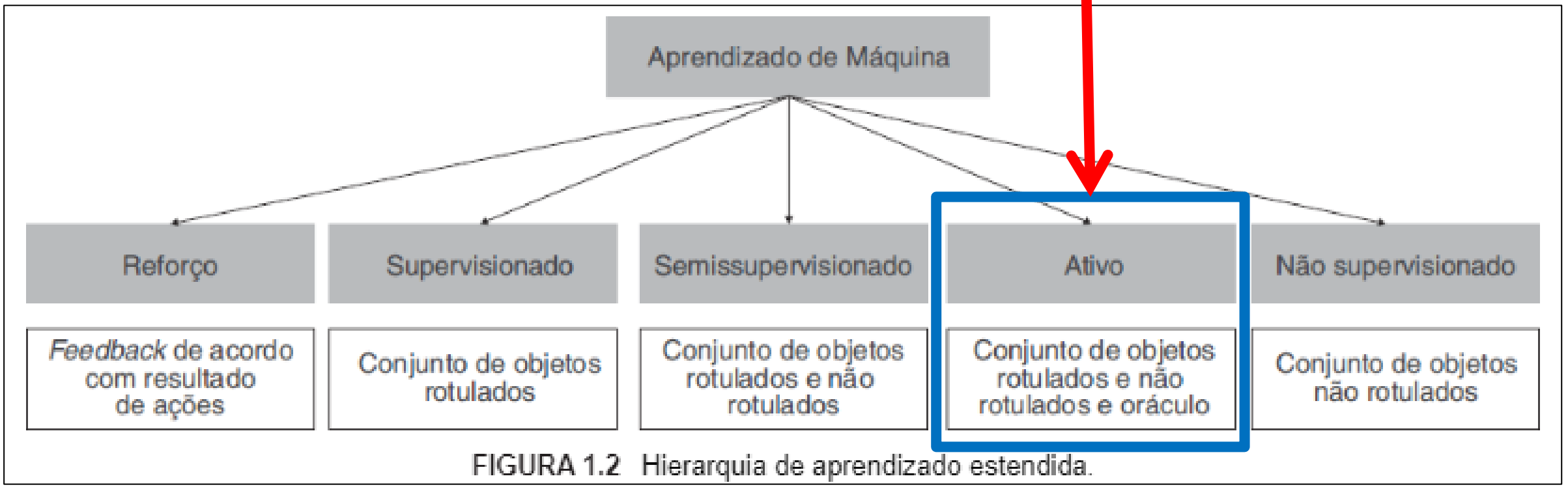
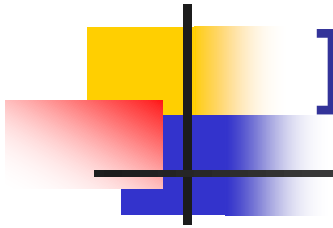


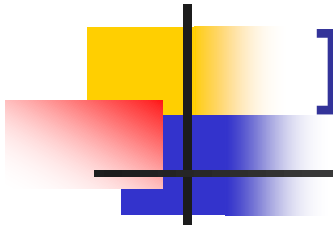
FIGURA 1.2 Hierarquia de aprendizado estendida.



Introdução

- Consider that, for any supervised learning system to perform well, it must often be trained on hundreds (even thousands) of labeled instances

- Sometimes these labels come at little or no cost, such as the the “spam” flag you mark on unwanted email messages
  - Many other supervised learning tasks, labeled instances are very difficult, time-consuming, or expensive to obtain
    - Ex.: documents, image, audio, video

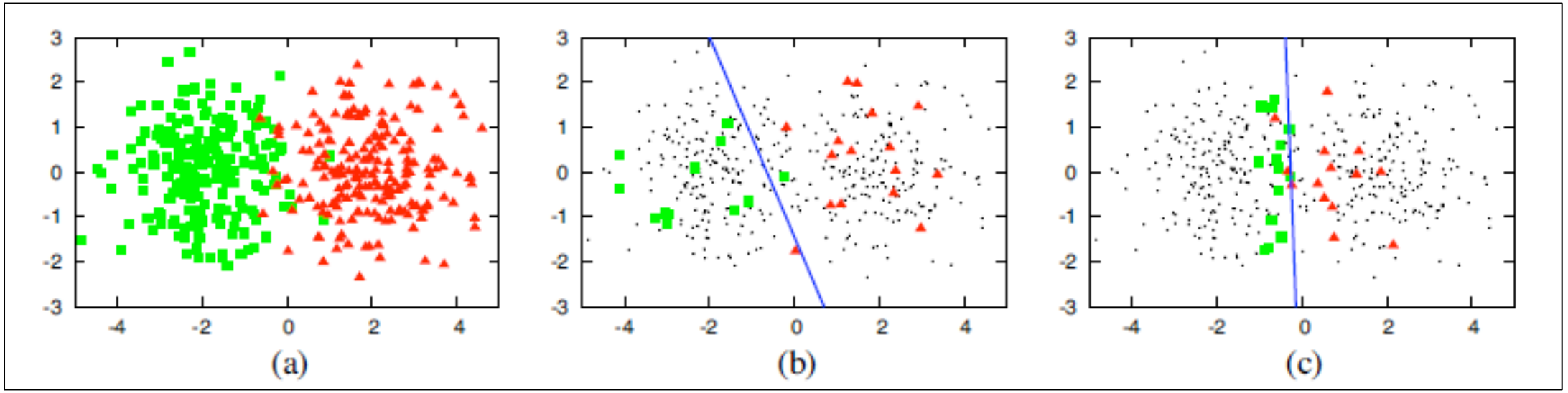


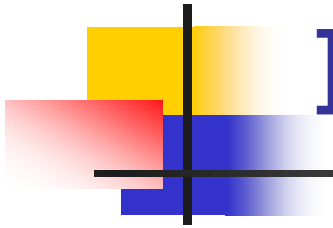
Introdução

- The key hypothesis is that if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training

Introdução

- Suppose that the acquisition of labels is so costly that one is only allowed to acquire the labels of few examples from the entire data set and use this set of few examples to train a model
- Clearly, the wrong choice of training examples may lead to significant overfitting
- The **goal** in active learning is to **integrate** the **labeling** and **classification process** in a **single framework** to create **robust models**

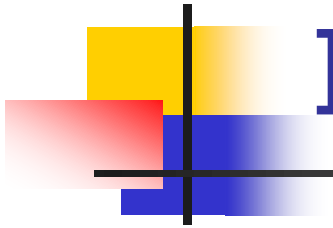




Introdução

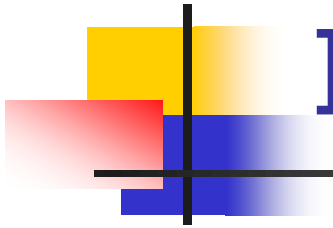
- Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle (e.g., a human annotator, Amazon Mechanical Turk)

- The active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data



Introdução

- Every active learning system has two primary components:
  - Oracle
    - Provides the responses to the underlying query in the form of labels of specified test instances
    - It is viewed as a black-box that is part of the input to the process
  - Query System
    - The job of the query system is to pose queries to the oracle for labels of specific records
    - The design of the query system may depend on the application



Introdução

- There are several **scenarios** in which active learners may pose queries
- There are also several different **query strategies** that have been used to decide which instances are most informative

Scenarios x Query Strategy Frameworks

- **Scenarios**
  - *Pool-Based Sampling*
  - Stream-Based Selective Sampling

- **Query Strategy Frameworks**

Heterogeneity-based models

- *Uncertainty Sampling*
- Query-By-Committee
- Expected Model Change

Performance-based models

- Expected Error Reduction
- Variance Reduction

Representativeness-based models

- Density-Weighted Methods

Workflow

Como escolher os primeiros exemplos? (seed): random, clustering

Bias?

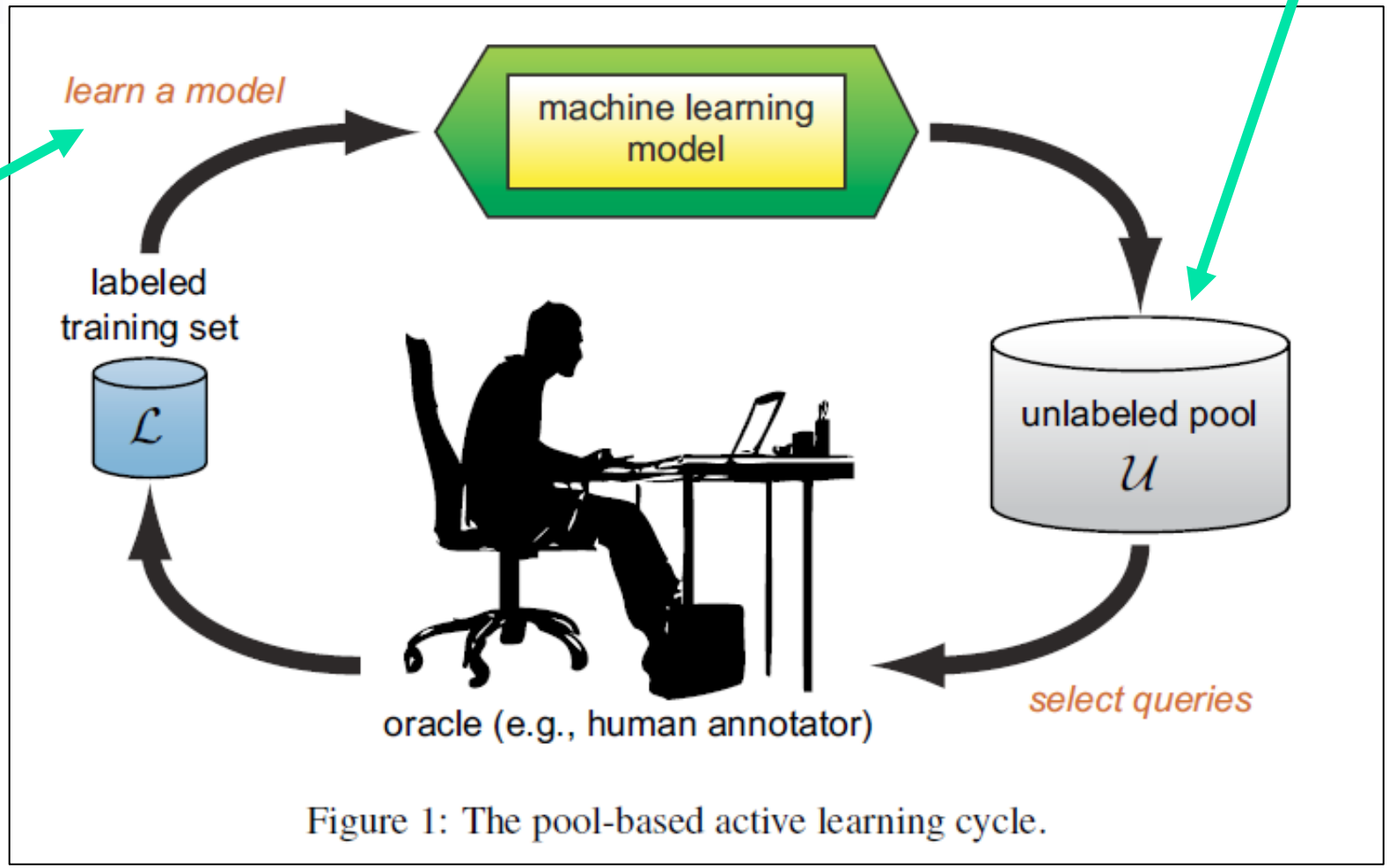


Figure 1: The pool-based active learning cycle.

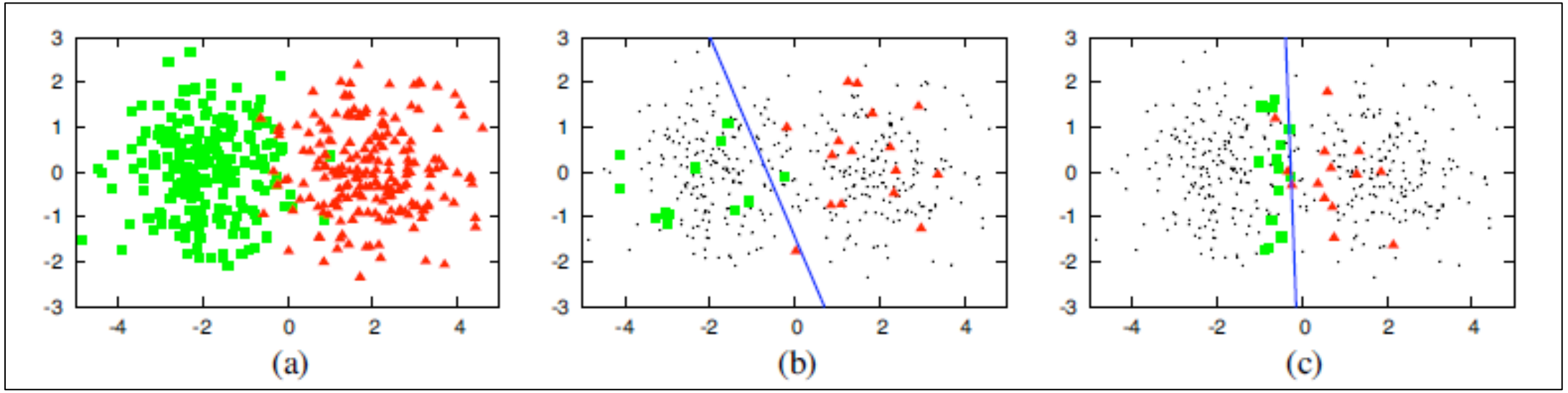
Pool-based active learning setting

- Queries are selected from a large pool of unlabeled instances

Uncertainty sampling query strategy

- Selects the instance in the pool about which the model is least certain how to label

Exemplo



A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain – The line represents the decision boundary of the classifier (70% accuracy)

A logistic regression model trained with 30 actively queried instances using uncertainty sampling (90%)

Scenarios

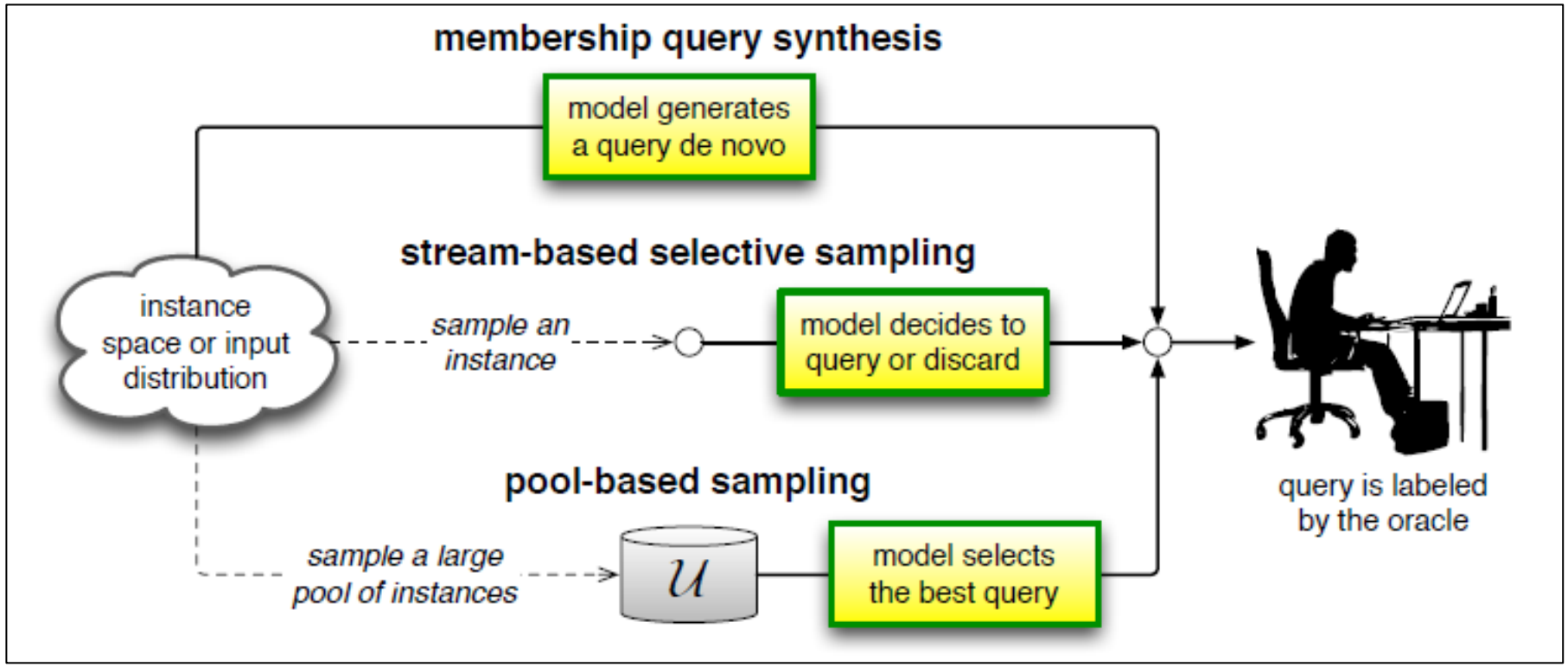
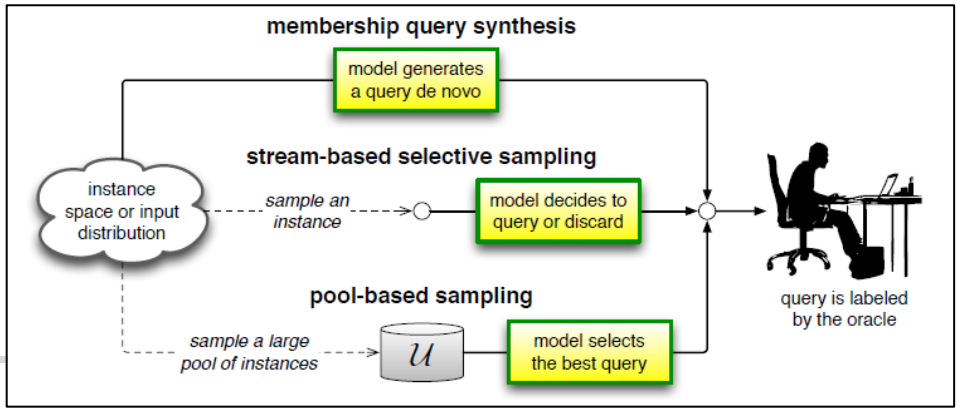


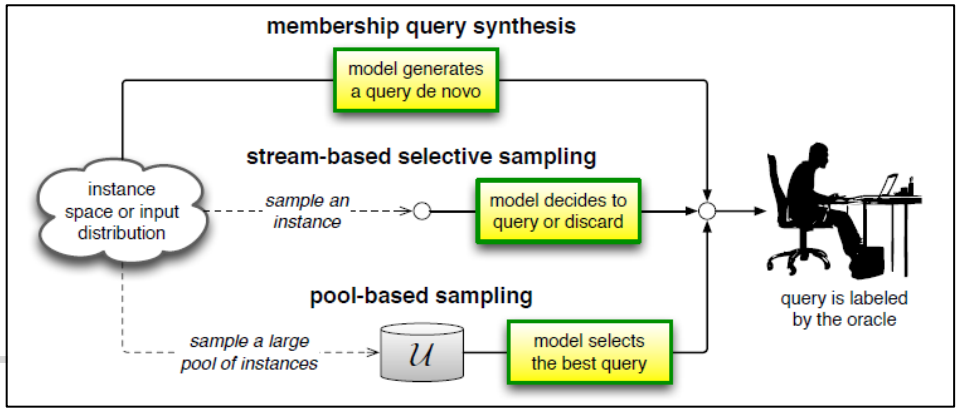
Diagram illustrating the three main active learning scenarios

Scenarios
Pool-Based Sampling



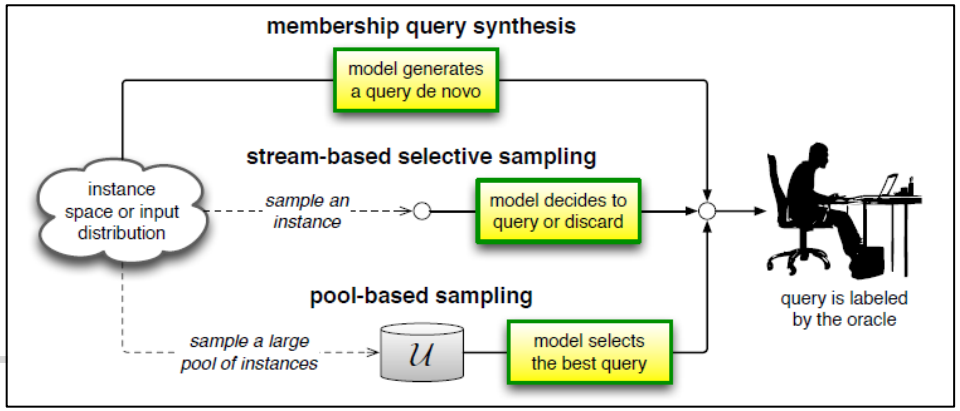
- It assumes that there is a small set of labeled data L and a large pool of unlabeled data U available
- Queries are selectively drawn from the pool, which is usually assumed to be closed (i.e., static or non-changing)
- Typically, instances are queried in a greedy fashion, according to an informativeness measure used to evaluate all instances in the pool

Scenarios
Stream-Based **Selective Sampling**



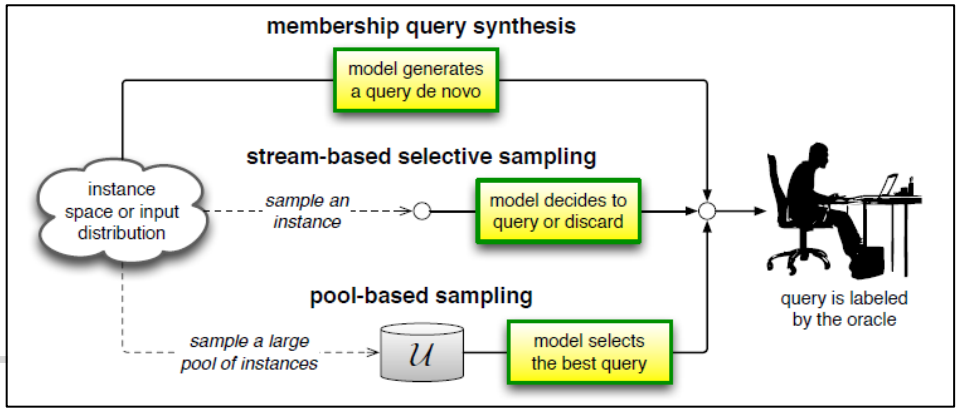
- The key assumption is that obtaining an unlabeled instance is free (or inexpensive), so it can first be sampled from the actual distribution, and then the learner can decide whether or not to request its label
- Also known as stream-based or sequential active learning, as each unlabeled instance is typically drawn one at a time from the data source, and the learner must decide whether to query or discard it

Scenarios
Stream-Based **Selective Sampling**



- The decision whether or not to query an instance can be framed several ways
- One approach is to evaluate samples using some "informativeness measure" or "query strategy" and make a decision such that more informative instances are more likely to be queried

Scenarios
Stream-based x Pool-based



- The main difference is that SB scans through the data sequentially and makes query decisions individually, whereas the PB evaluates and ranks the entire collection before selecting the best query

Query Strategy Frameworks

- **Heterogeneity-based models**: these models are based on the assumption that regions near the decision boundary are more likely to be heterogeneous and instances in these regions are more valuable for learning the decision boundary

- **Query Strategy Frameworks**
  - *Uncertainty Sampling*
  - Query-By-Committee
  - Expected Model Change
  - Expected Error Reduction
  - Variance Reduction
  - Density-Weighted Methods

Query Strategy Frameworks

- **Performance-based models**: these models quantify the impact of adding the queried instance to the classifier performance on remaining unlabeled instances

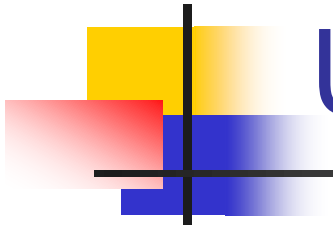
- **Query Strategy Frameworks**
  - *Uncertainty Sampling*
  - Query-By-Committee
  - Expected Model Change
  - Expected Error Reduction
  - Variance Reduction
  - Density-Weighted Methods

Query Strategy Frameworks

- **Representativeness-based models**: these models are based on the idea that informative instances should not only be those which are uncertain, but also those which are "representative" of the underlying distribution

- **Query Strategy Frameworks**
  - *Uncertainty Sampling*
  - Query-By-Committee
  - Expected Model Change
  - Expected Error Reduction
  - Variance Reduction
  - Density-Weighted Methods

Query Strategy Frameworks Uncertainty Sampling



- The simplest and most commonly used query framework
- In this framework, 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

Query Strategy Frameworks Uncertainty Sampling

- For problems with three or more class labels, a more general uncertainty sampling variant might query the instance whose prediction is the least confident

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

$\hat{y}$ = the class label with the highest posterior probability under the model θ

$$\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$$

Query Strategy Frameworks Uncertainty Sampling

- However, the criterion for the least confident strategy only considers information about the most probable label
- To correct this, some researchers use a variant called margin sampling

$$x_M^* = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

$\hat{y}_1, \hat{y}_2$ = the first and the second most probable class labels under the model

- Instances with large margins are easy, since the classifier has little doubt in differentiating between the two most likely class labels
- Instances with small margins are more ambiguous, thus knowing the true label would help the model discriminate more effectively between them

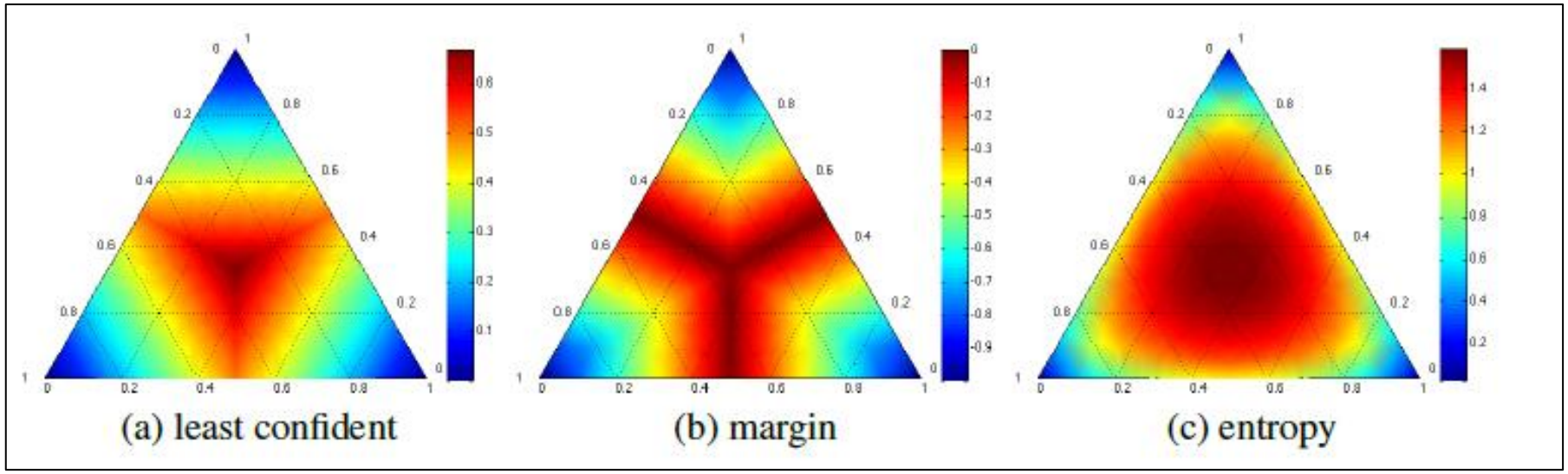
Query Strategy Frameworks Uncertainty Sampling

- For problems with very large label sets, the margin approach still ignores much of the output distribution for the remaining classes
- A more general uncertainty sampling strategy (and possibly the most popular) uses entropy as an uncertainty measure

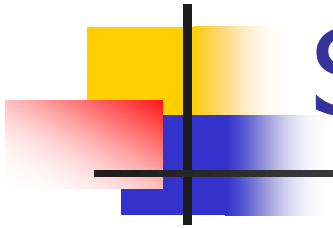
$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

y_i ranges over all possible labelings

Query Strategy Frameworks
Uncertainty Sampling



Heatmaps illustrating the query behavior of common uncertainty measures in a three-label classification problem



Stopping Criteria

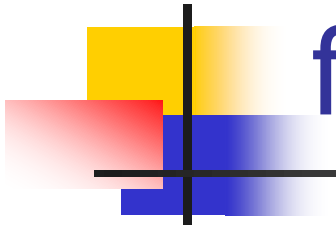
- When to stop learning?
  - One option is when the accuracy of a learner has reached a plateau, and acquiring more data is likely a waste of resources
- Several criteria have been proposed
  - These methods are generally based on the notion that there is an intrinsic measure of stability within the learner, and active learning ceases to be useful once that measure begins to level-off or degrade

Related Research Areas
Semi-Supervised Learning

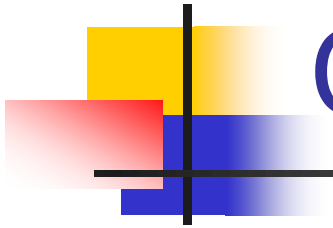
- Active learning and semi-supervised learning attack the same problem from opposite directions
  - While semi-supervised methods exploit what the learner thinks it knows about the unlabeled data, active methods attempt to explore the unknown aspects
    - For example, self-training uses the most confident unlabeled instances, while active learning with uncertainty sampling the instances about which the model is least confident

Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance [Paper]

modAL: A modular active learning framework for Python



- https://modal-python.readthedocs.io/en/latest/
- https://github.com/modAL-python/modal



Outros

- https://prodi.gy/