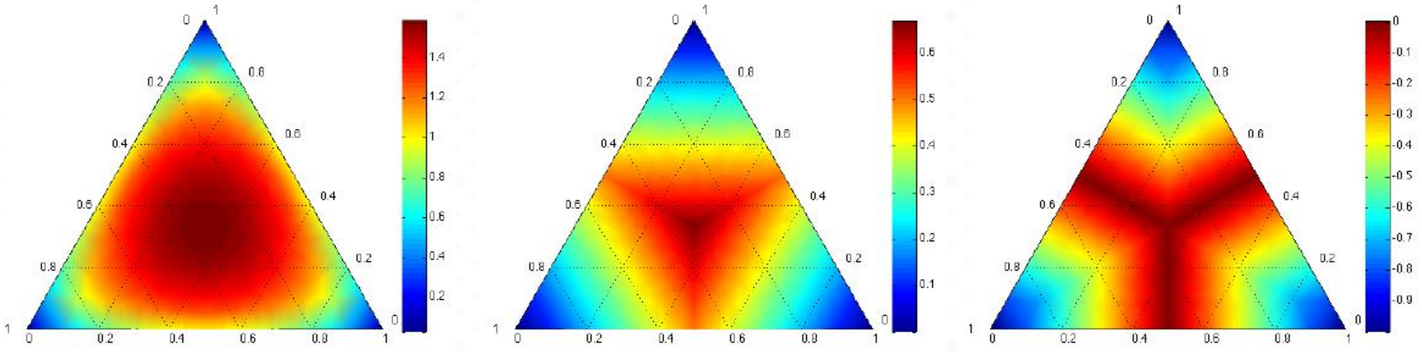
ECE 4252/8803: Fundamentals of Machine Learning (FunML)
Fall 2024

Lecture 25: Data and Label Efficient Learning – Active Learning



OLIVES @GeorgiaTech

Georgia Tech

Overview

In this Lecture..

Supervised Learning

- Labeling assumption:
  - **Labeling processes** are already **completed** beforehand.
- Goal:
  - To achieve good generalizability

Active Learning

- Labeling assumption:
  - Samples are **selectively** labeled with certain budget constraints
- Goal:
  - To achieve good generalizability with **fewer annotated samples**

Active Learning

Motivation

Unlabeled Data Points

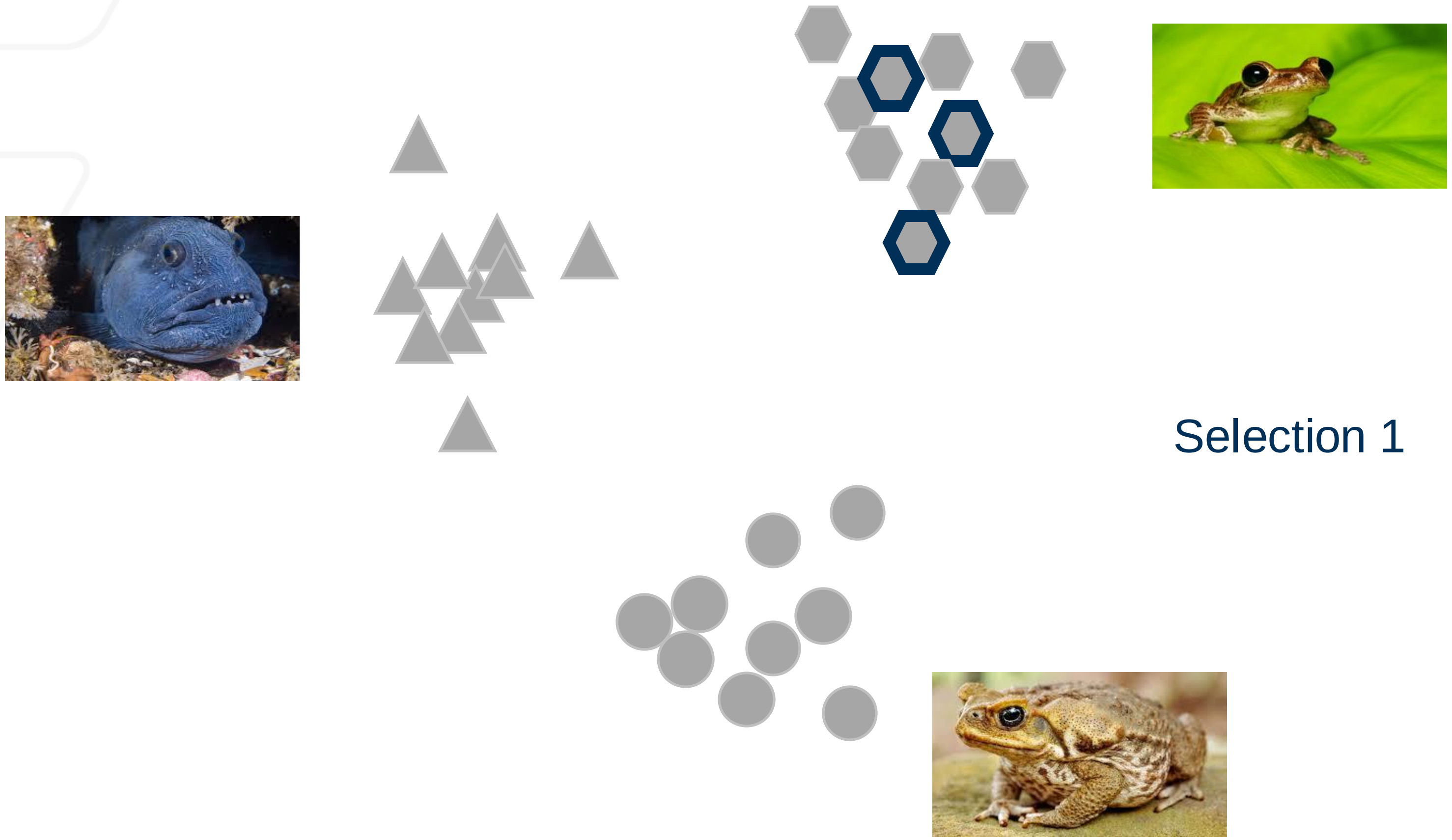
Labeling Constraints:

Cost per labeled sample: 3$
Available budget: 9$
Samples to label: 3

Active Learning

Motivation

Unlabeled Data Points

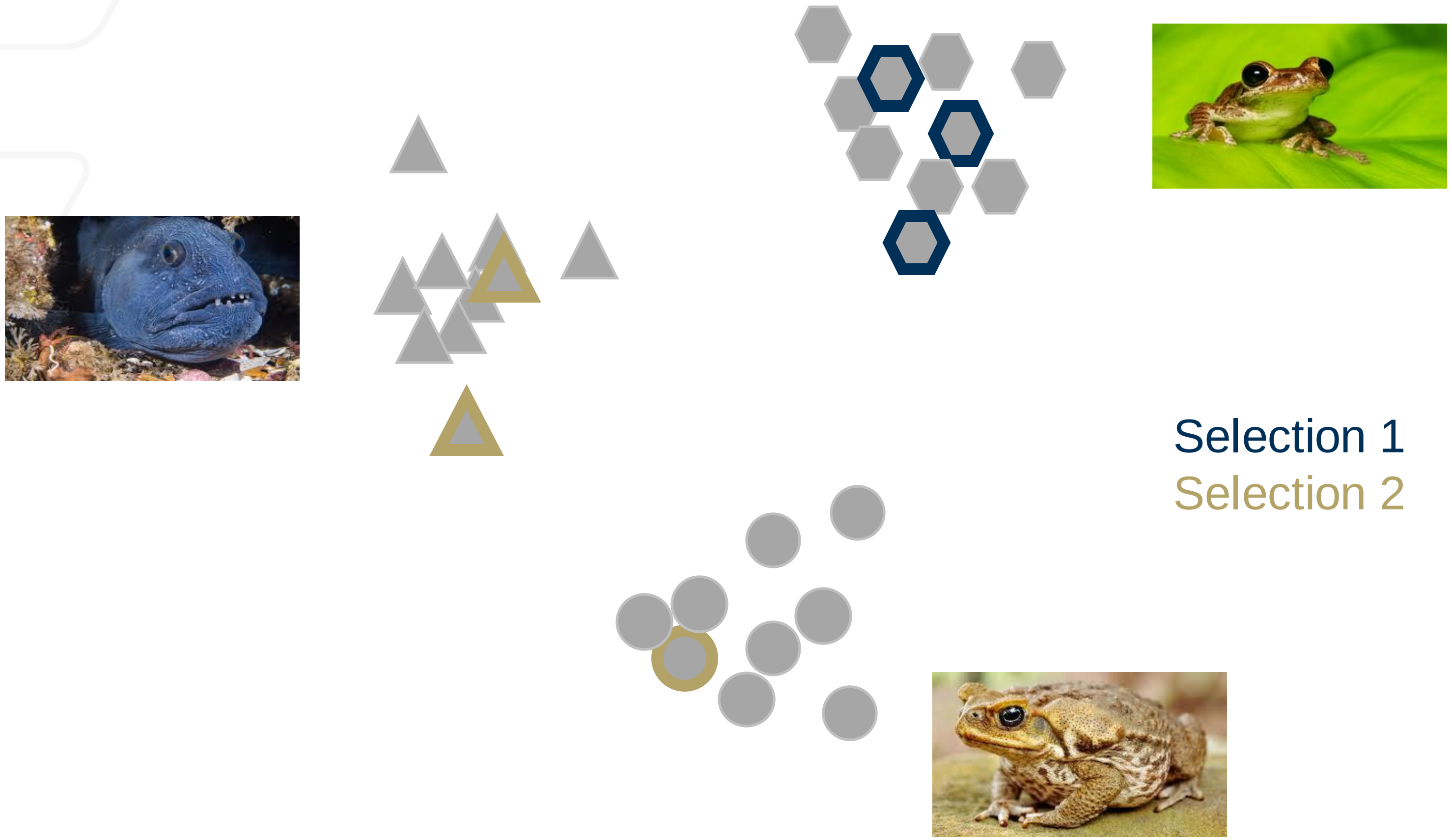


Labeling Constraints:

Cost per labeled sample: 3$
Available budget: 9$
Samples to label: 3

Active Learning
Motivation

Unlabeled Data Points

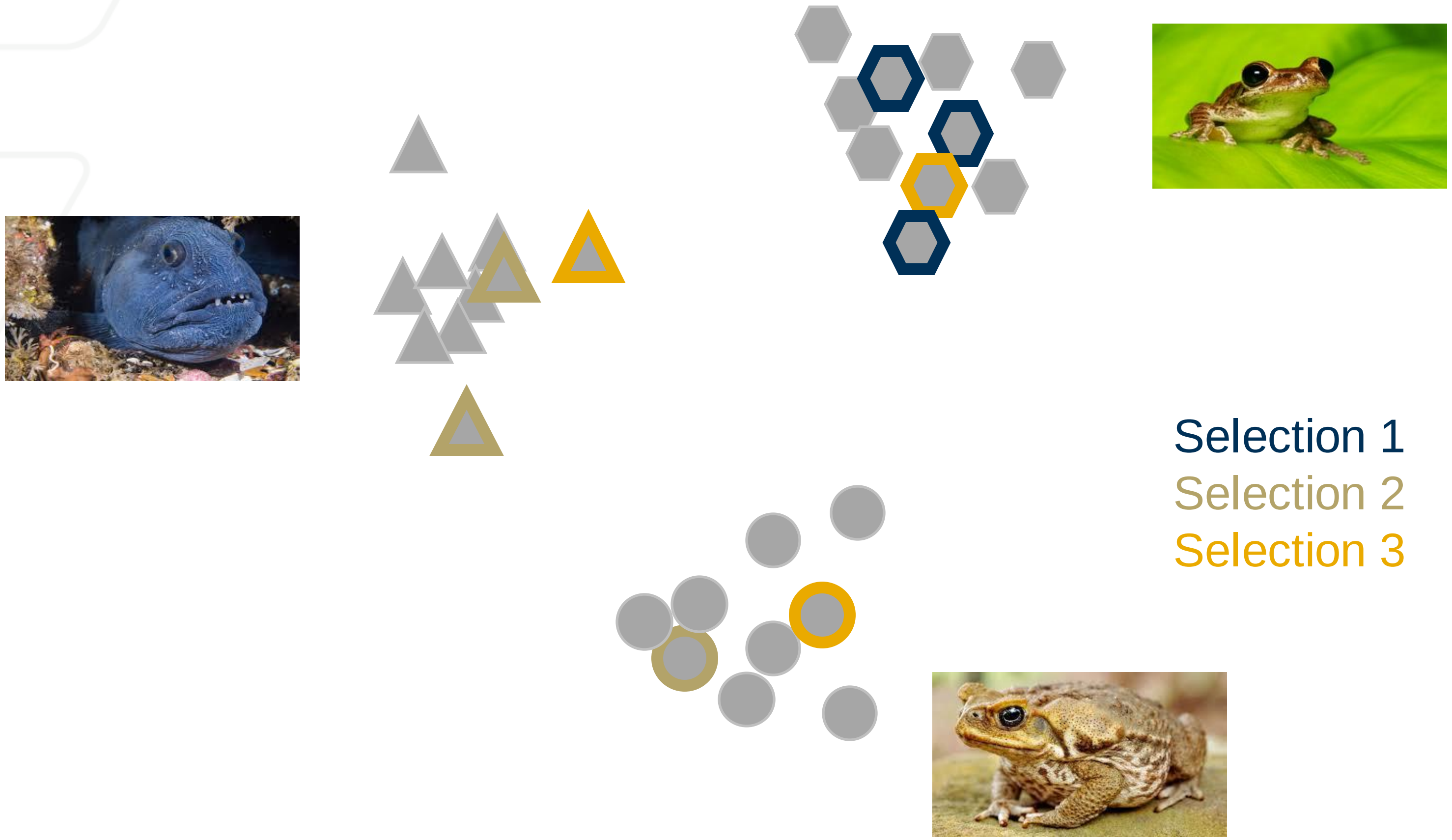


Labeling Constraints:

Cost per labeled sample: 3$
Available budget: 9$
Samples to label: 3

Active Learning
Motivation

Unlabeled Data Points



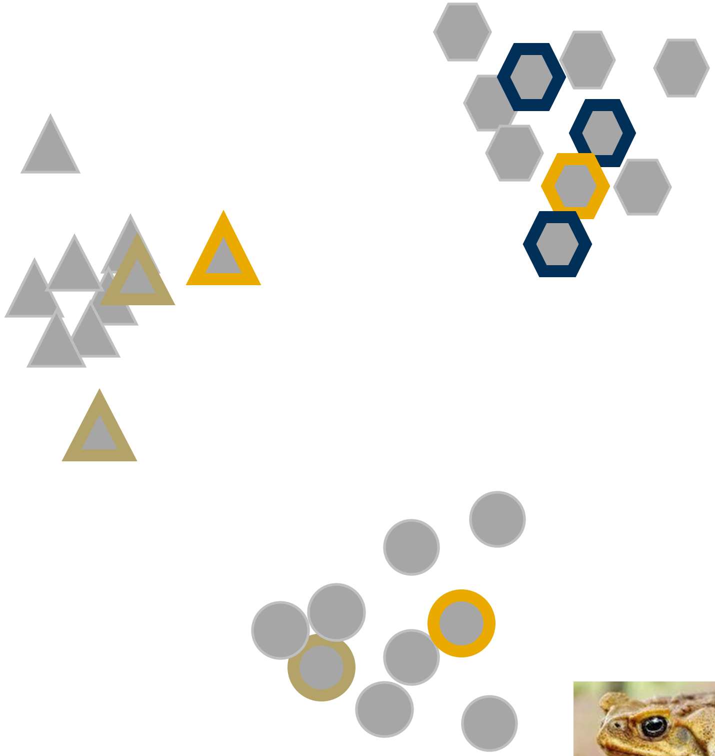
Labeling Constraints:

Cost per labeled sample: 3$
Available budget: 9$
Samples to label: 3

Active Learning

Motivation

Unlabeled Data Points



Selection 1
Selection 2
Selection 3

Labeling Constraints:

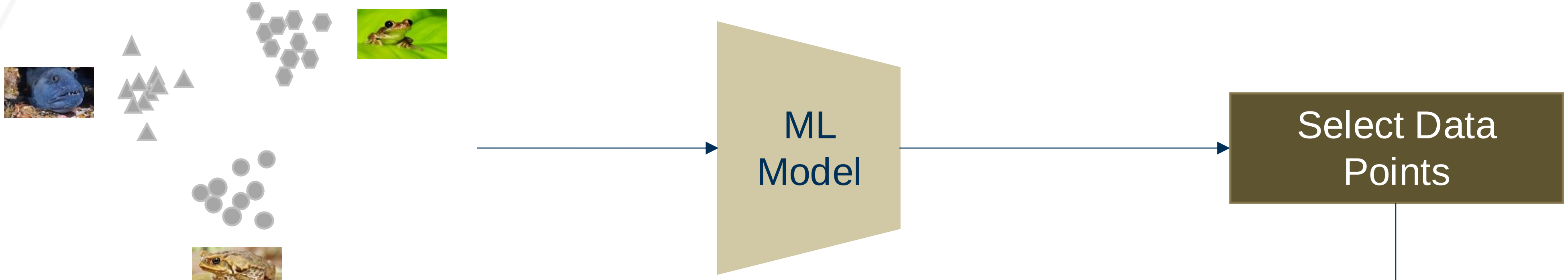
Cost per labeled sample: 3$
Available budget: 9$
Samples to label: 3

Which **selection** results in the **highest performance** of our machine learning model?

Overview

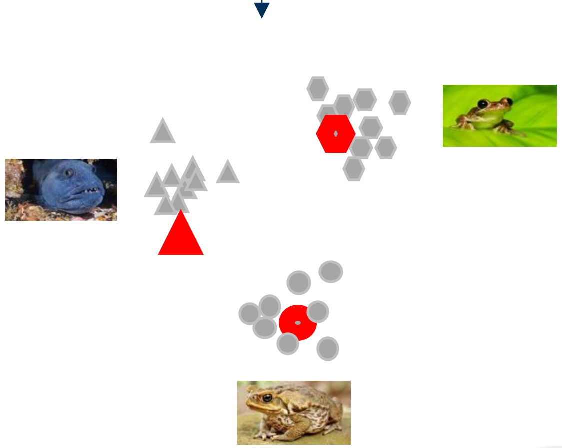
In this Lecture..

Active Learning
Basics

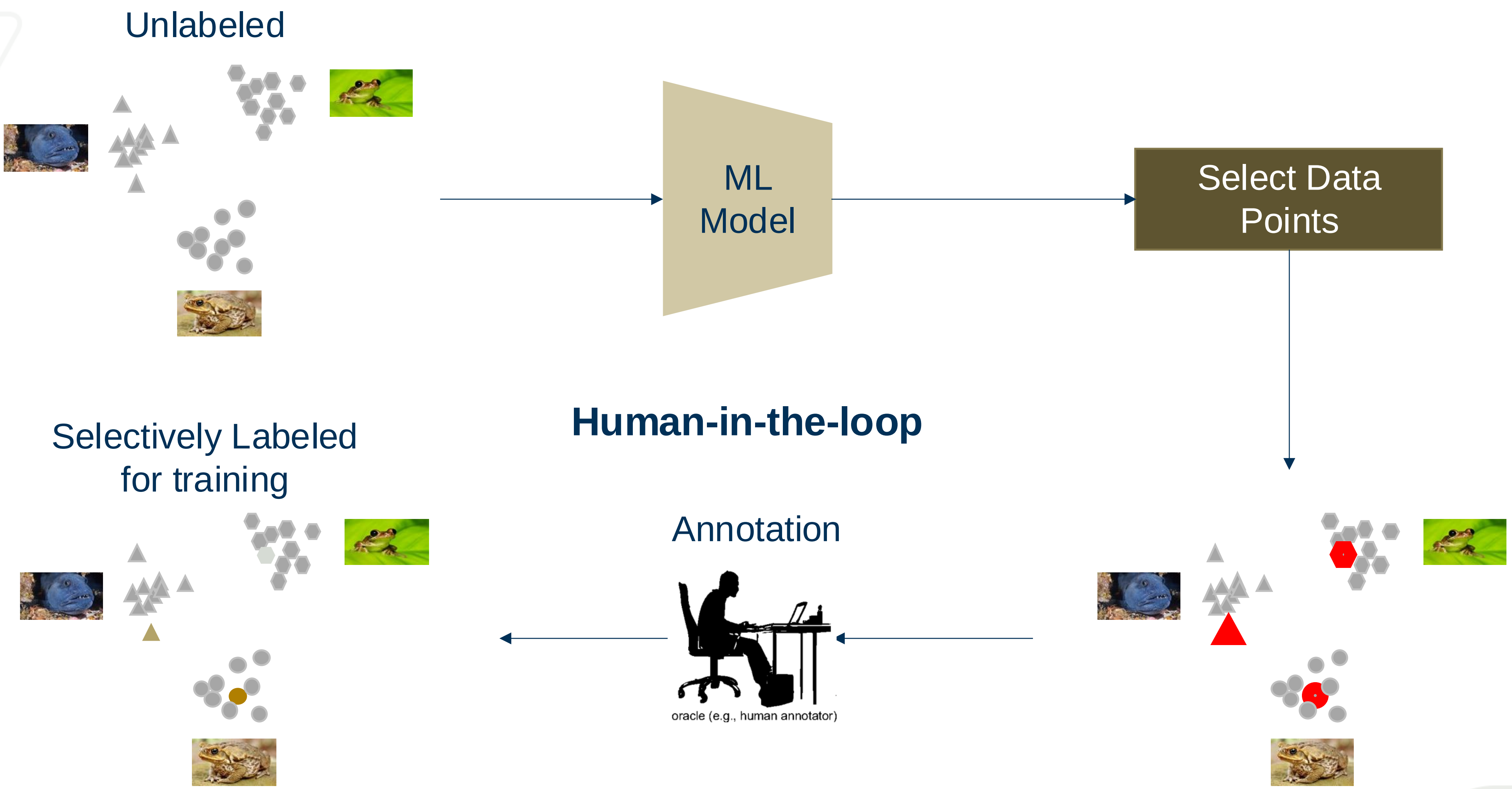


Unlabeled

We can leverage information extracted from the model to guide the selection.



Active Learning
Overview



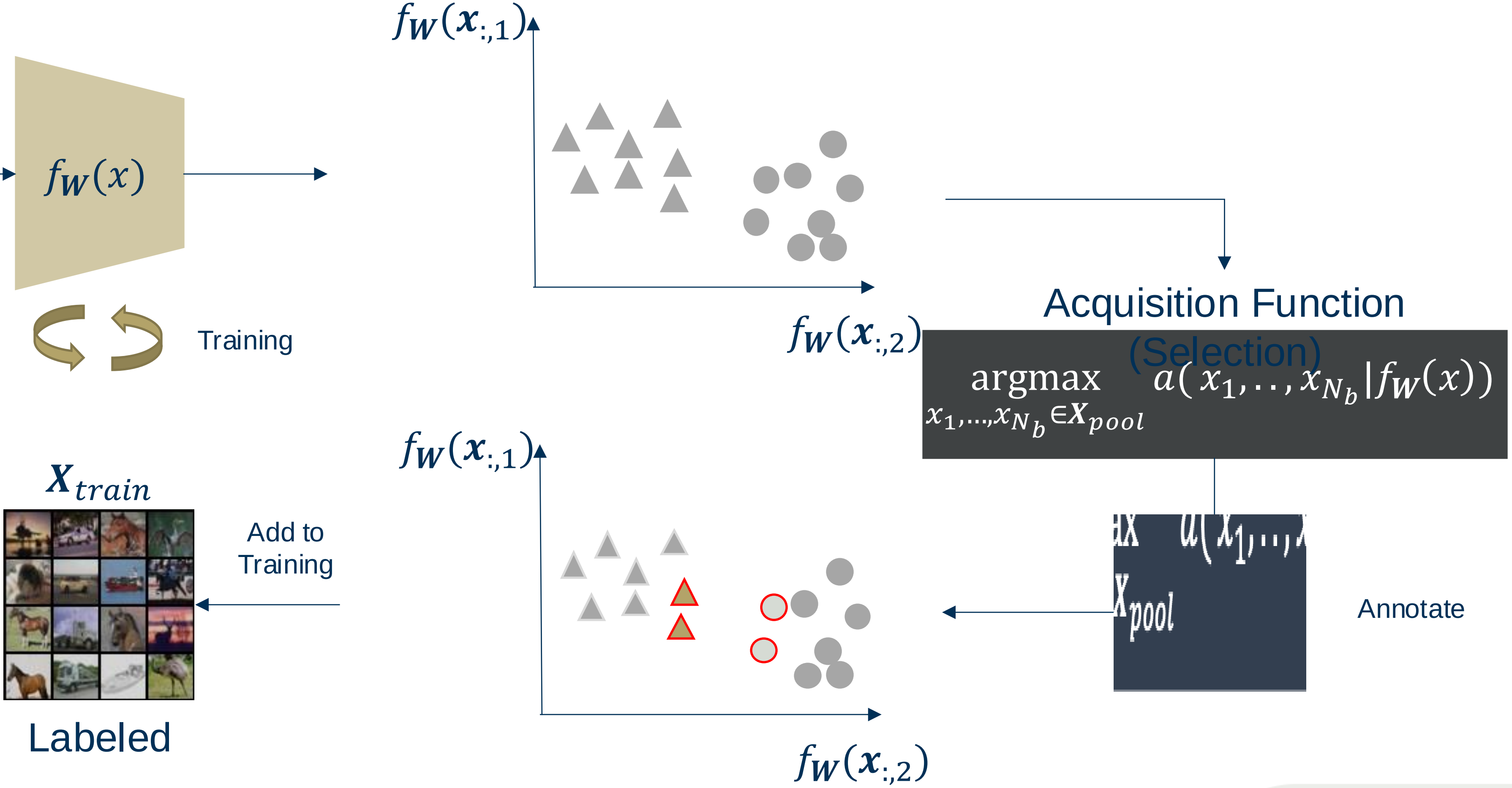
Terminologies

$\boldsymbol{X}_{pool}$

Unlabeled

$\boldsymbol{X}_{pool}$: Unlabeled Pool
$\boldsymbol{X}_{train}$: Labeled Training Set
a: Acquisition function
N_b: Acquisition size

Active Learning Workflow



Active Learning

Process

Input: Unlabeled pool, untrained network, acquisition function (also called query strategy), oracle (annotator)

- Round 1:
  - Choose a small and random subset of available data from the unlabeled training pool
  - Label the randomly chosen subset
  - Train the ML model based on the labeled subset
- Round 2:
  - Pass the remaining unlabeled data through the trained ML model
  - The acquisition function selects a subset of data from the unlabeled pool based on the outputs from the trained ML model
  - The oracle labels the selected subset
  - The ML model is retrained with all available labeled data (from round 1 and round 2)
- Round 3:
  - ...

The rounds continue until a stopping criteria is met

Overview

In this Lecture..

Introduction and Motivation

Fundamentals

Performance Metrics

- Performance Metrics
- Annotator Metrics

Difficulty-based Active Learning

Diversity-based Active Learning

Fusion-based Active Learning

Deployment Performance Metric: Regression

Active Learning
Stopping Criteria

Stopping criteria is constraints dependent:

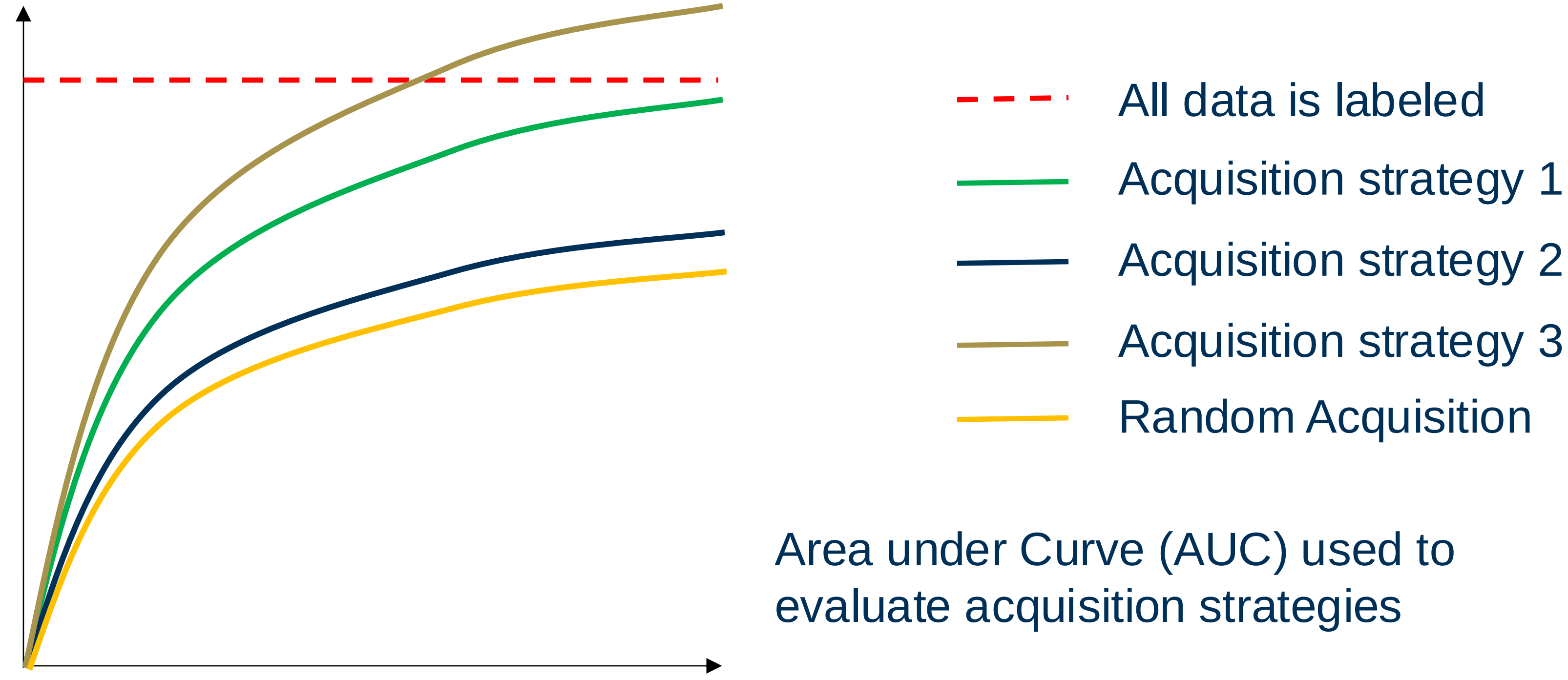
1. Performance constraints – application acceptable performance
   - If necessary and sufficient performance is achieved
2. Annotator constraints – Time/Money
   - If budgetary constraints are met

Measuring Active Learning Performance

Application-specific performance on **test** set

For classification:
- Accuracy
- F1 score
- TPR/FPR
- …

For object detection:
- mAP (mean Average Precision)
- …



Number of annotated samples in $\boldsymbol{X}_{train}$

Area under Curve (AUC) used to evaluate acquisition strategies

$$AUC_3 > AUC_1 > AUC_2$$

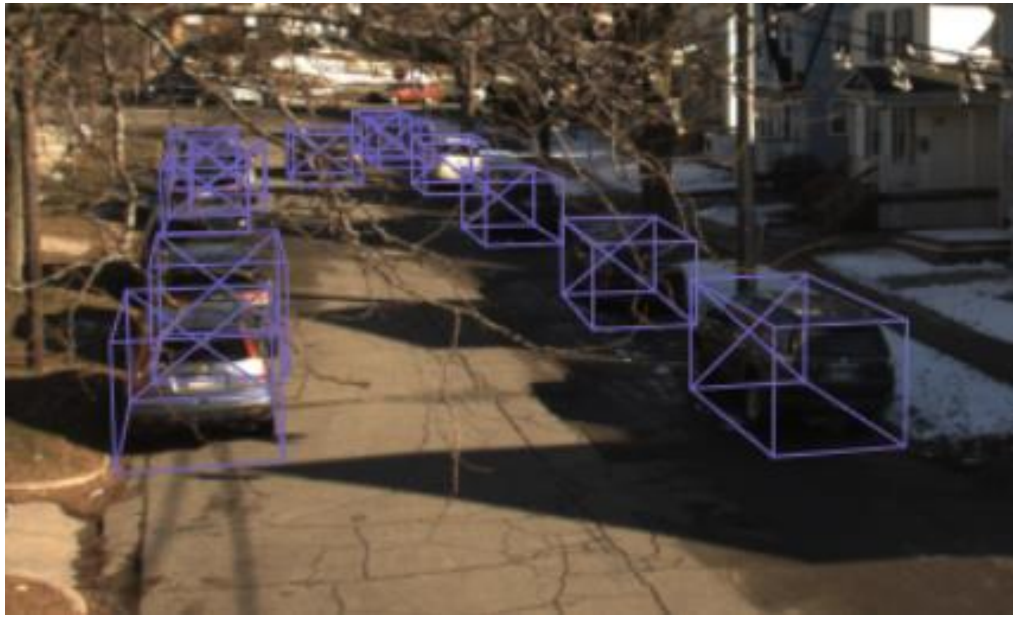
Active Learning

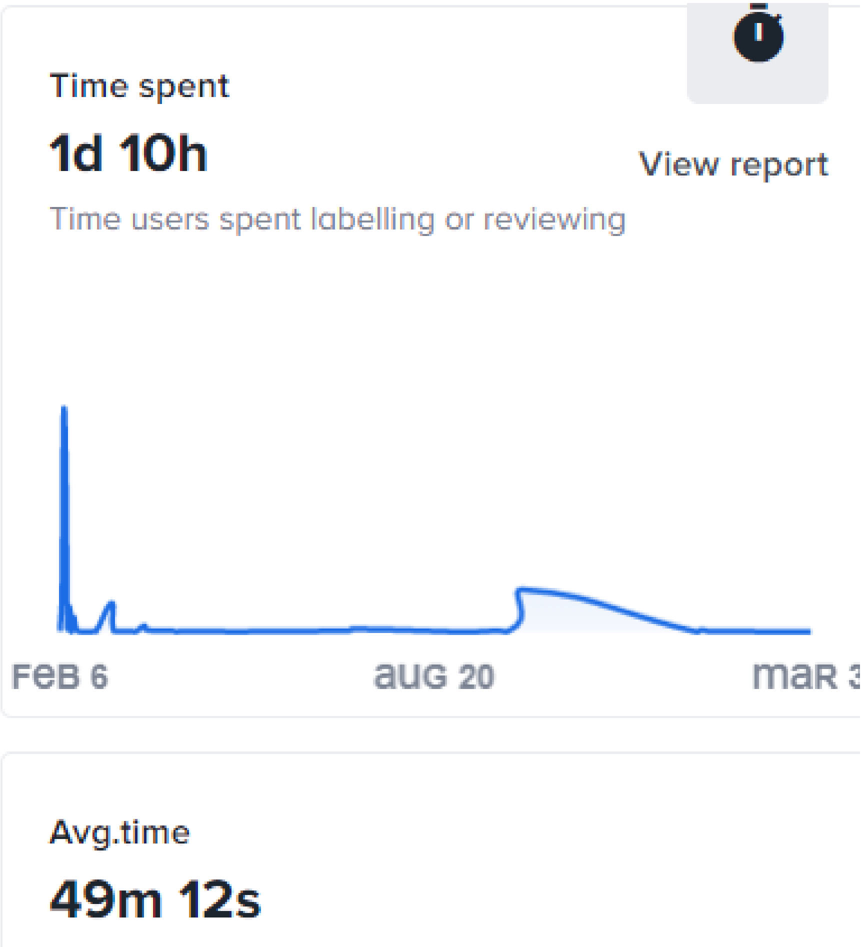
Annotator Constraints based-Metrics

1d 10h is a time constraint

Sequence

- 578 frames
- Labels for object detection, lidar, agent interactions





Kokilepersaud, Kiran, et al. "FOCAL: A Cost-Aware Video Dataset for Active Learning." *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023.

OLIVES @GeorgiaTech

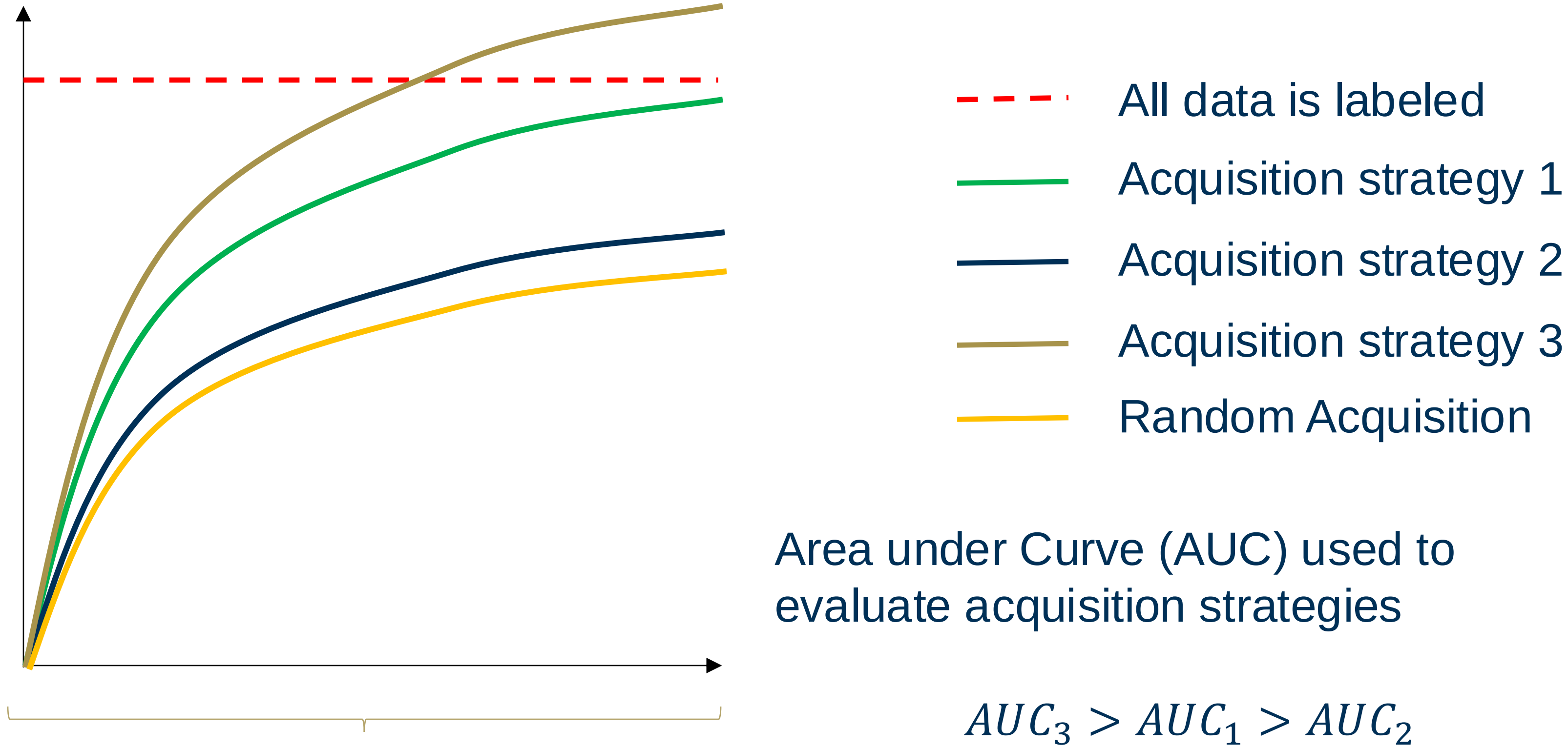
Georgia Tech

Active Learning

Annotator Constraints based-Metrics

Application-specific annotator cost on **train** set

For classification:
- Hours
- FLOPS
- $ Cost (Not preferred unless you are advertising your annotation company)
- …



Number of annotated samples in $\boldsymbol{X}_{train}$

Area under Curve (AUC) used to evaluate acquisition strategies

$$AUC_3 > AUC_1 > AUC_2$$

Kokilepersaud, Kiran, et al. "FOCAL: A Cost-Aware Video Dataset for Active Learning." *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023.

Overview

In this Lecture..

Introduction and Motivation

Fundamentals

Performance Metrics

Difficulty-based Active Learning

- Entropy
- Least Confidence
- Margin

Diversity-based Active Learning

Fusion-based Active Learning

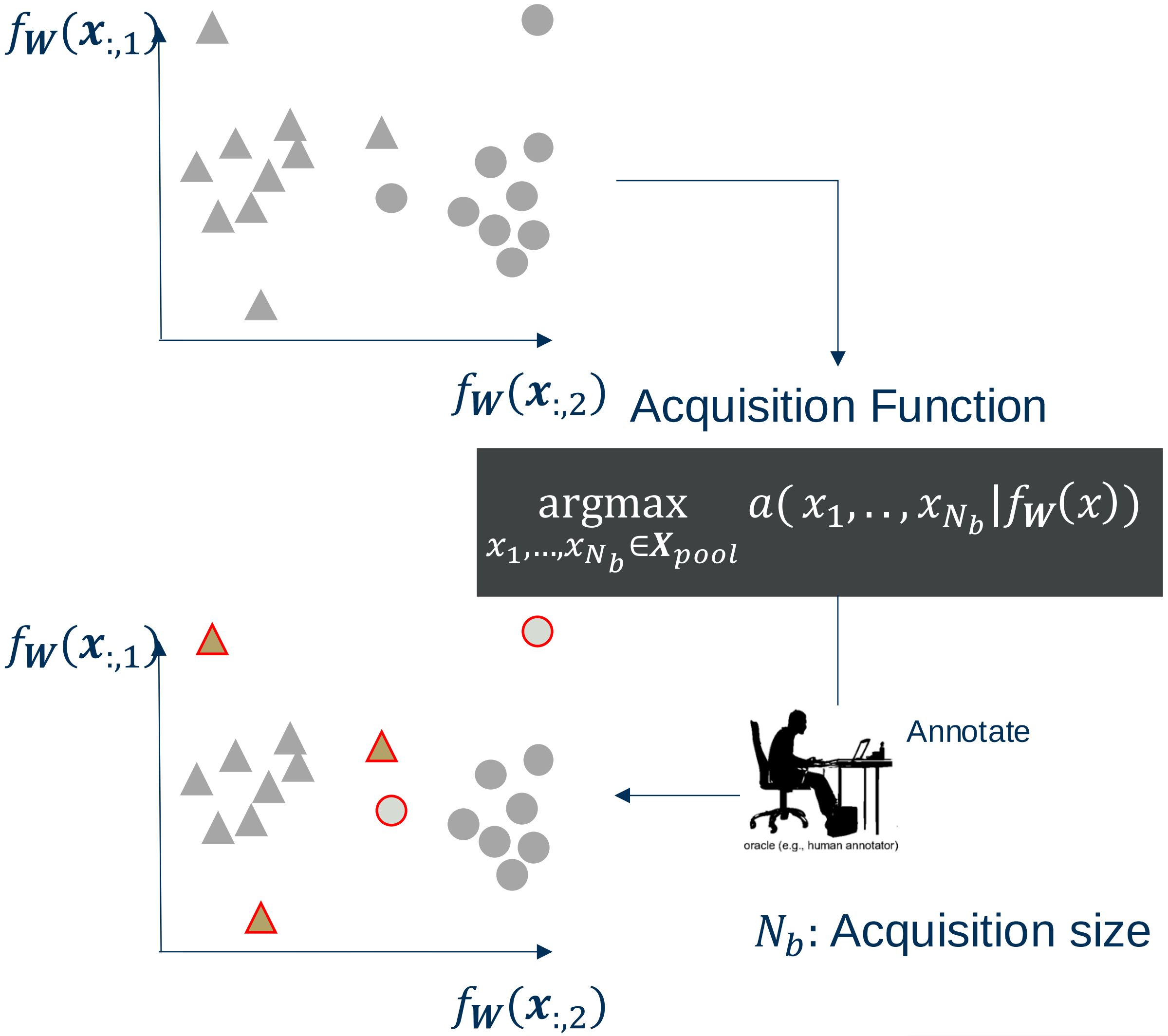
Deployment Performance Metric: Regression

Acquisition Functions

Difficulty-based Acquisition Functions

Difficulty-based Acquisition Functions

- Idea: select the most “difficult” samples in the dataset
- Examples of difficult samples:
    - Rare classes
    - Outliers
    - High data noise
    - Class imbalance
    - …
- Also referred to as uncertainty-based acquisition functions



Difficulty-based Acquisition Functions

- Select the most **difficult** samples for **generalization**
- Popular examples:
  - Entropy

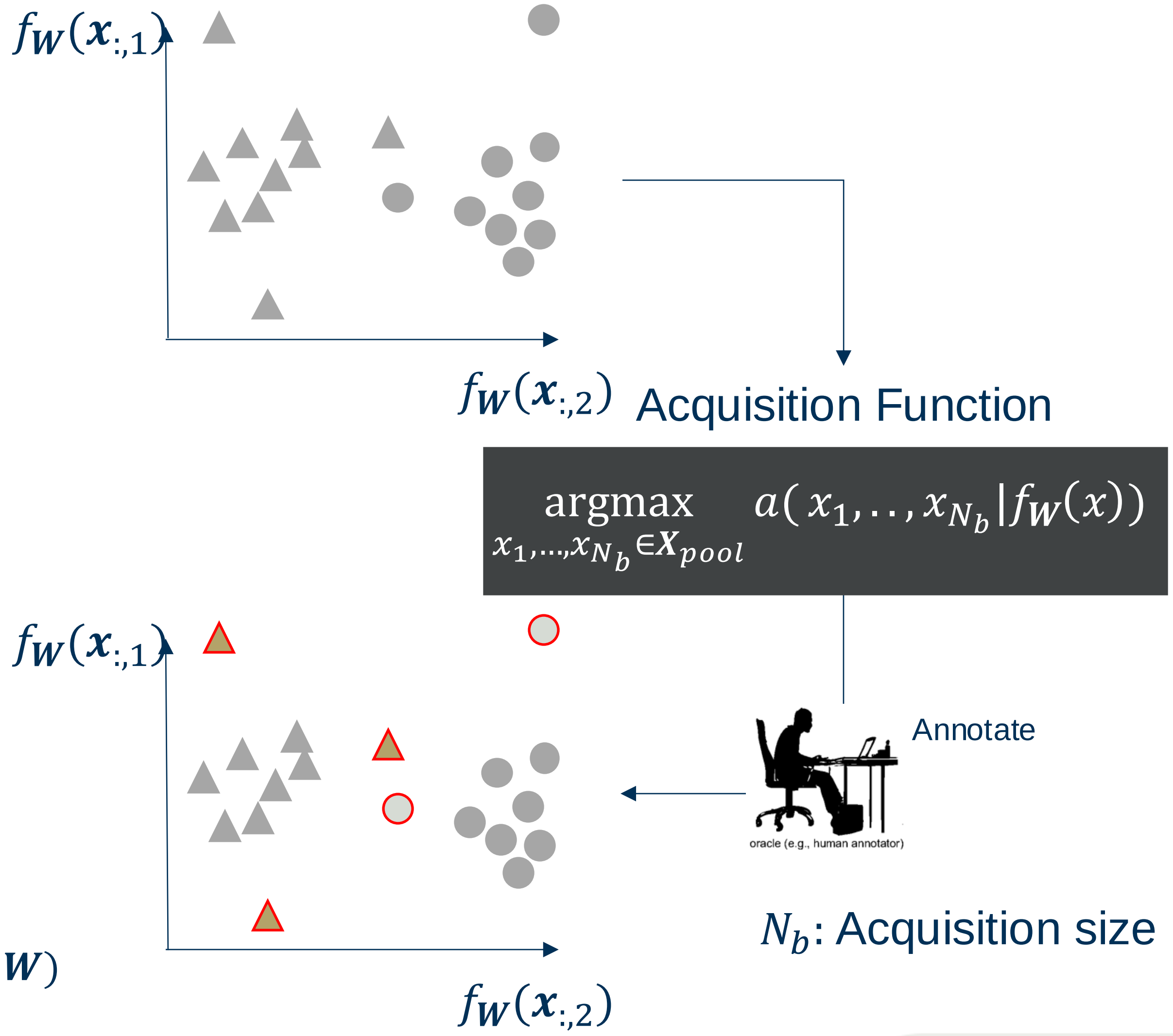
$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} H(p(y|\boldsymbol{x}_i, \boldsymbol{W}))$$

  - Least Confidence

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = -\sum_{i=1}^{N_b} \max_{y_c} p(y = y_c | \boldsymbol{x}_i, \boldsymbol{W})$$

  - Margin

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} p(y = c_s | \boldsymbol{x}_i, \boldsymbol{W}) - p(y = c_h | \boldsymbol{x}_i, \boldsymbol{W})$$



Entropy

- Select the most **difficult** samples for **generalization**
- Popular examples:
  - **Entropy**

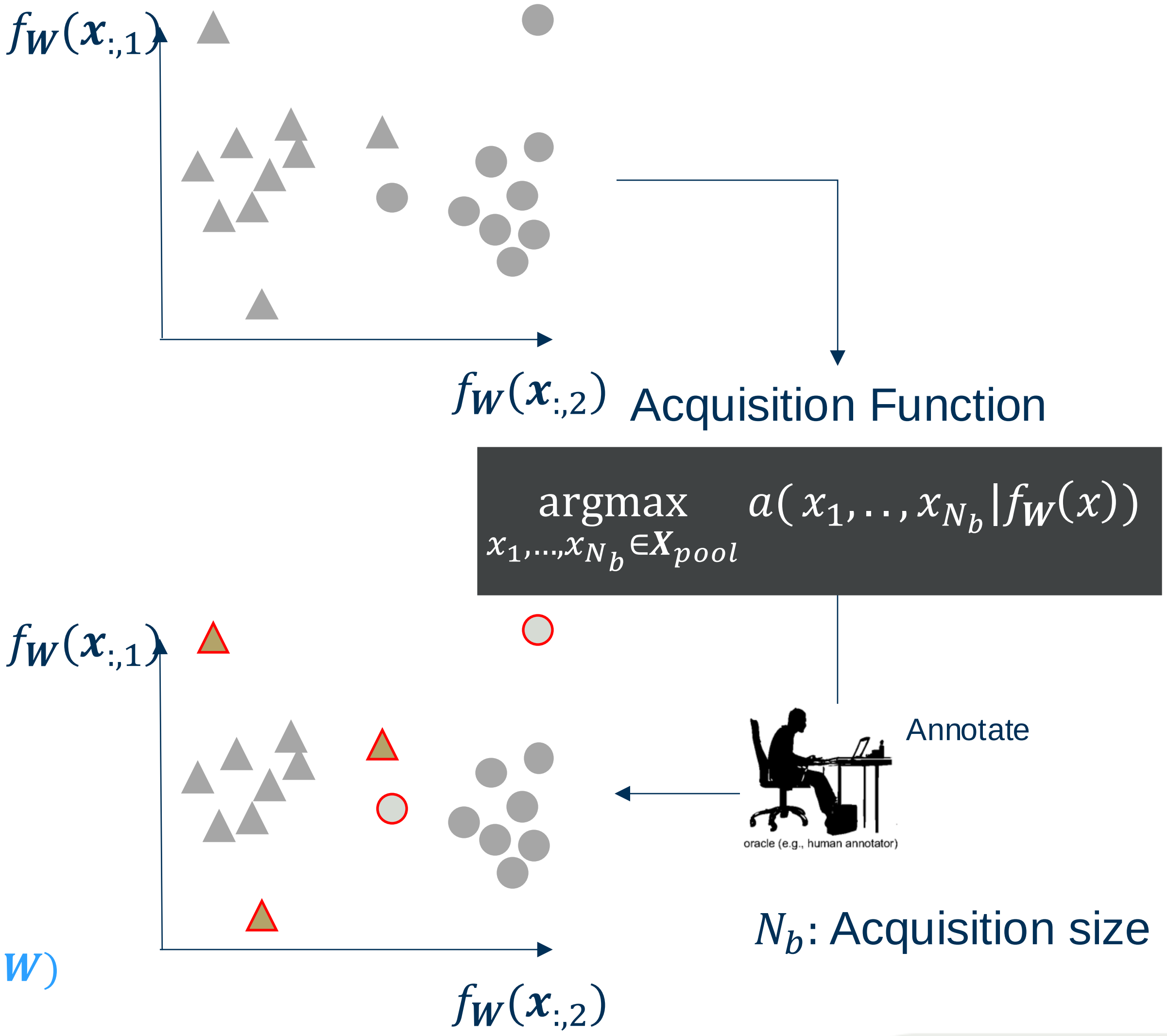
    $$a\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N_b}\middle|f_W(\boldsymbol{x})\right)=\sum_{i=1}^{N_b}H(p(y|\boldsymbol{x}_i,\boldsymbol{W}))$$

  - Least Confidence

    $$a\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N_b}\middle|f_W(\boldsymbol{x})\right)=-\sum_{i=1}^{N_b}\max_{y_c}p(y=y_c|\boldsymbol{x}_i,\boldsymbol{W})$$

  - Margin

    $$a\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N_b}\middle|f_W(\boldsymbol{x})\right)=\sum_{i=1}^{N_b}p(y=c_s|\boldsymbol{x}_i,\boldsymbol{W})-p(y=c_h|\boldsymbol{x}_i,\boldsymbol{W})$$



N_b: Acquisition size

Entropy Sampling

- Select the most **difficult** samples of the dataset
- Entropy represents the information content of a distribution
- Idea: annotate samples with the **highest entropy**

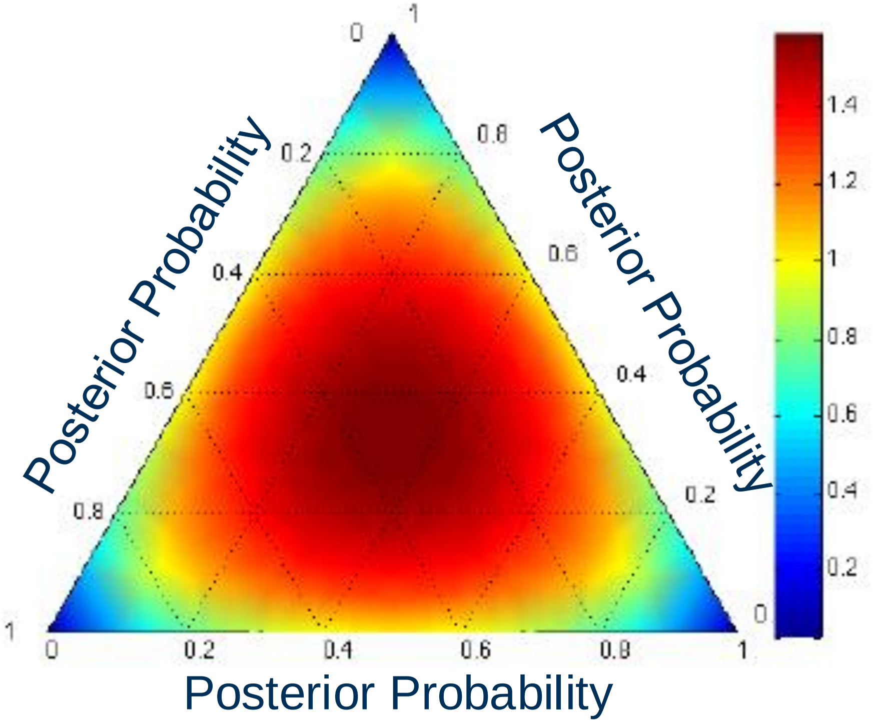
- Highest Entropy Sampling

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_{\boldsymbol{W}}(\boldsymbol{x})\right) = \sum_{i=1}^{N_b} H(p(y|\boldsymbol{x}_i, \boldsymbol{W}))$$

where

$$H(p(y|\boldsymbol{x}_i, \boldsymbol{W})) = -\sum_{c=1}^{C} p(y = c|\boldsymbol{x}_i, \boldsymbol{W}) \log(p(y = c|\boldsymbol{x}_i, \boldsymbol{W}))$$

SoftMax Entropy of 3-Classes



Three Label Classification:
- Red: Requires annotation
- Blue: Does not require annotation

Acquisition Functions

Difficulty-based Acquisition Functions

Least Confidence

- Select the most **difficult** samples for **generalization**

- Popular examples:
  - Entropy

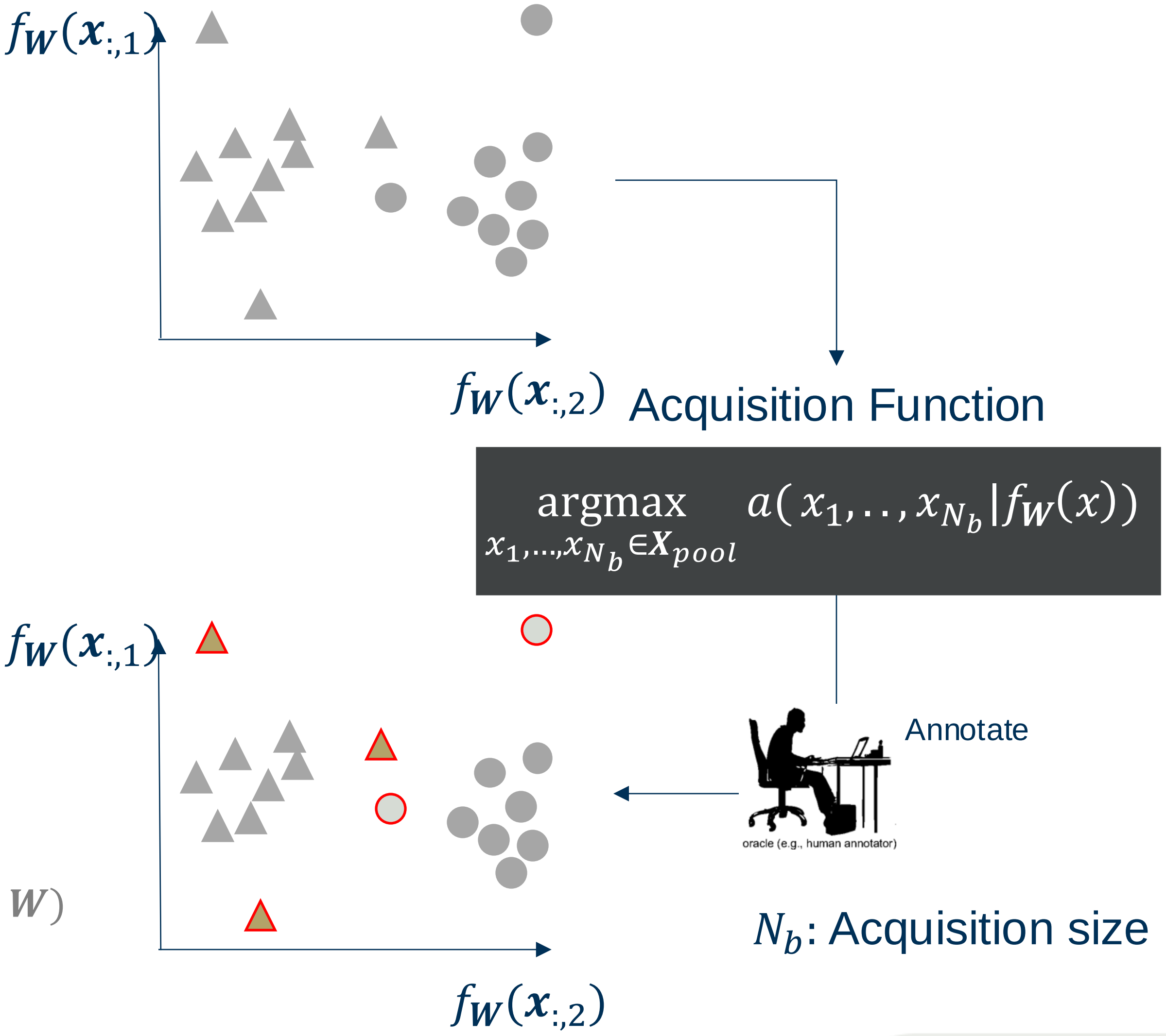
$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} H(p(y|\boldsymbol{x}_i, \boldsymbol{W}))$$

  - **Least Confidence**

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = -\sum_{i=1}^{N_b} \max_{y_c} p(y = y_c|\boldsymbol{x}_i, \boldsymbol{W})$$

  - Margin

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} p(y = c_s|\boldsymbol{x}_i, \boldsymbol{W}) - p(y = c_h|\boldsymbol{x}_i, \boldsymbol{W})$$



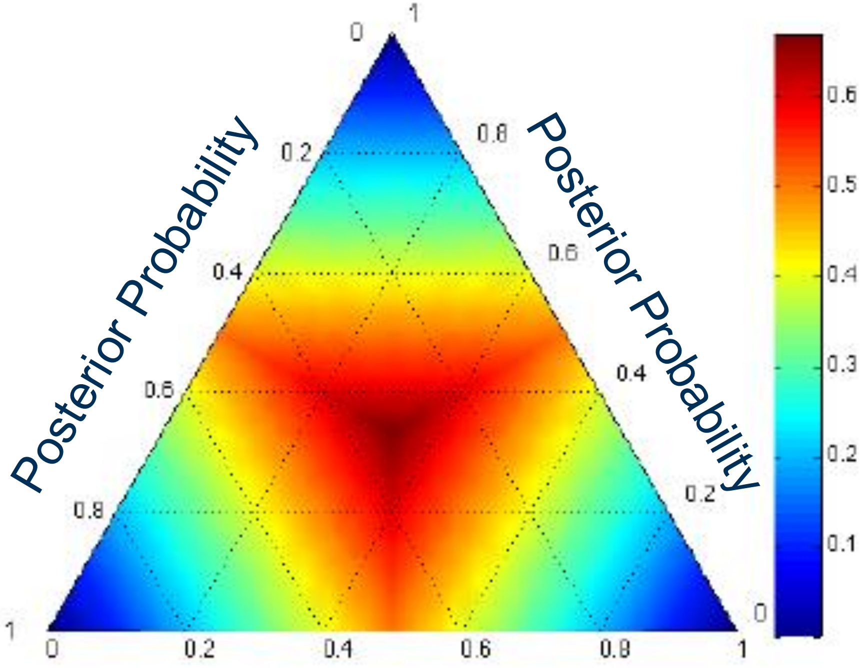
Least Confidence Sampling

- Select the most **difficult** samples of the dataset
- Samples with the lowest prediction confidence are the most difficult
- Idea: sample data points with **lowest output probability**

- Least Confidence

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_{\boldsymbol{W}}(\boldsymbol{x})\right) = -\sum_{i=1}^{N_b} \max_{y_c} p(y = y_c | \boldsymbol{x}_i, \boldsymbol{W})$$

SoftMax Least Confidence of 3-Classes



Three Label Classification:
- Red: Requires annotation
- Blue: Does not require annotation

Acquisition Functions

Difficulty-based Acquisition Functions

Margin

- Select the most **difficult** samples for **generalization**

- Popular examples:
  - Entropy

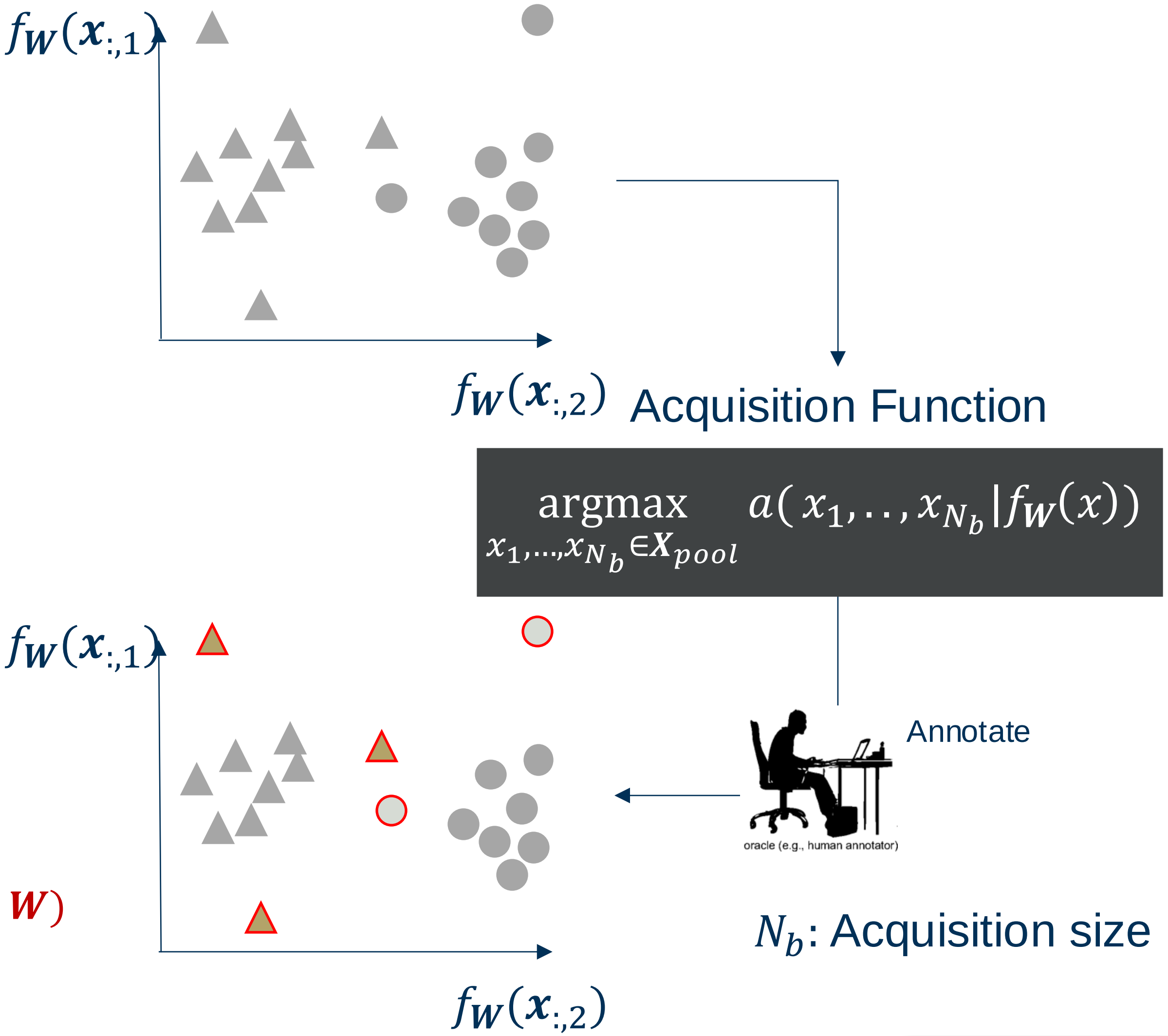
    $$a(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} H(p(y|\boldsymbol{x}_i, \boldsymbol{W}))$$

  - Least Confidence

    $$a(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = -\sum_{i=1}^{N_b} \max_{y_c} p(y = y_c|\boldsymbol{x}_i, \boldsymbol{W})$$

  - **Margin**

    $$a(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} p(y = c_s|\boldsymbol{x}_i, \boldsymbol{W}) - p(y = c_h|\boldsymbol{x}_i, \boldsymbol{W})$$



Margin Sampling

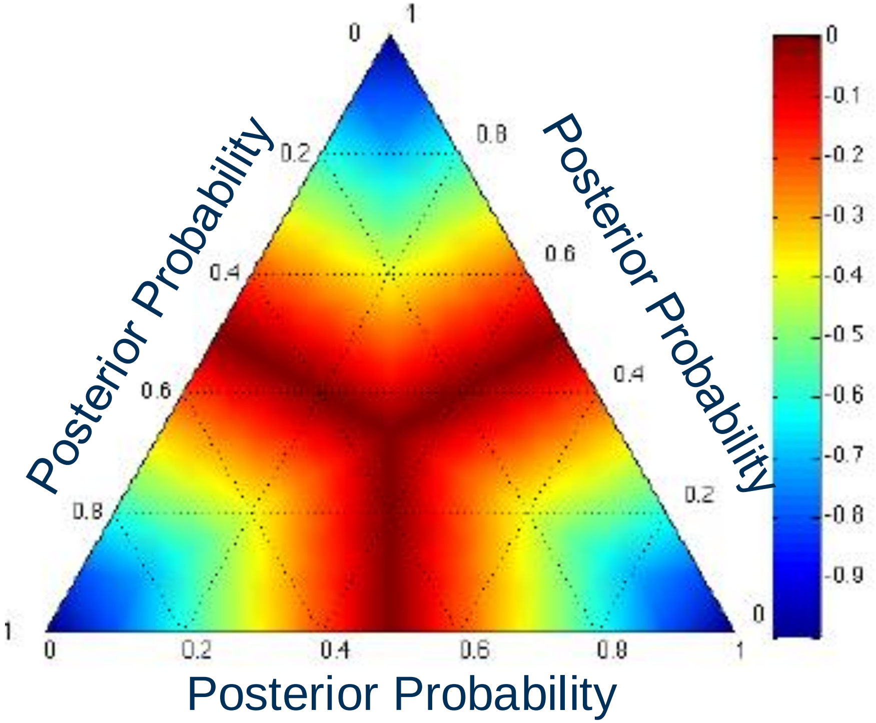
- Select the most **difficult** samples of the dataset
- Margin between highest probability class and second highest probability class measures prediction insecurity
- Idea: annotate samples with the **smallest class margin**

- Margin

$$a(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} | f_{\boldsymbol{W}}(\boldsymbol{x})) = \sum_{i=1}^{N_b} p(y = c_s | \boldsymbol{x}_i, \boldsymbol{W}) - p(y = c_h | \boldsymbol{x}_i, \boldsymbol{W})$$

c_s: Class with second highest probability
c_h: Class with highest probability

SoftMax Margin of 3-Classes



Three Label Classification:
- Red: Requires annotation
- Blue: Does not require annotation

Overview

In this Lecture..

Introduction and Motivation

Fundamentals

Performance Metrics

Difficulty-based Active Learning

Diversity-based Active Learning

- Coreset
- Discriminative Samping

Fusion-based Active Learning

Deployment Performance Metric: Regression

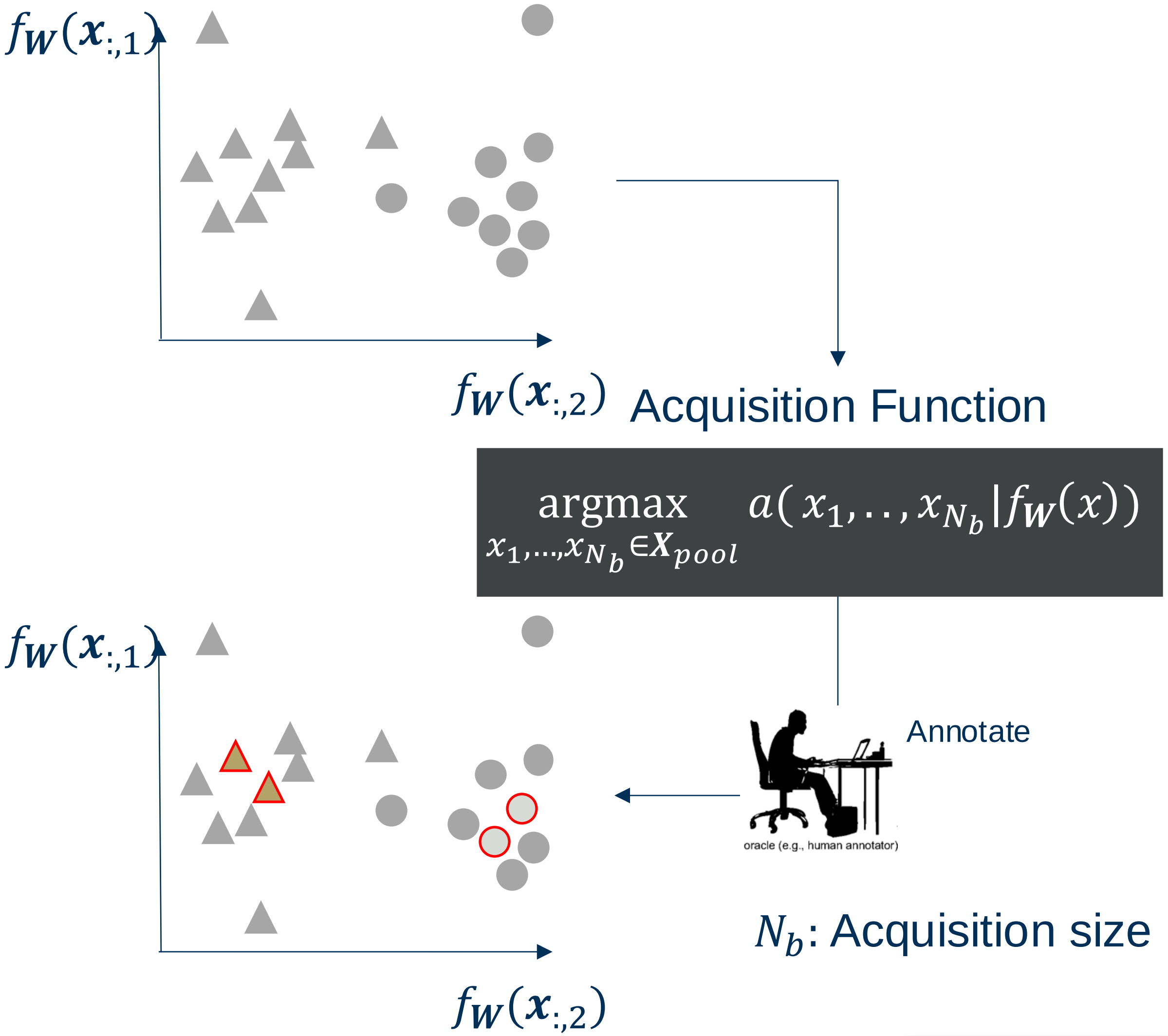
Acquisition Functions

Diversity-based Acquisition Functions

Diversity-based Acquisition Functions

- <u>Idea</u>: select samples that best "represent" the dataset

- Examples of representative samples:
  - Class centroids
  - Coreset samples
  - …



Diversity-based Acquisition Functions

- Select the most **representative** samples of the dataset
- Acquisition function based on the representation **directly**
- Popular examples:

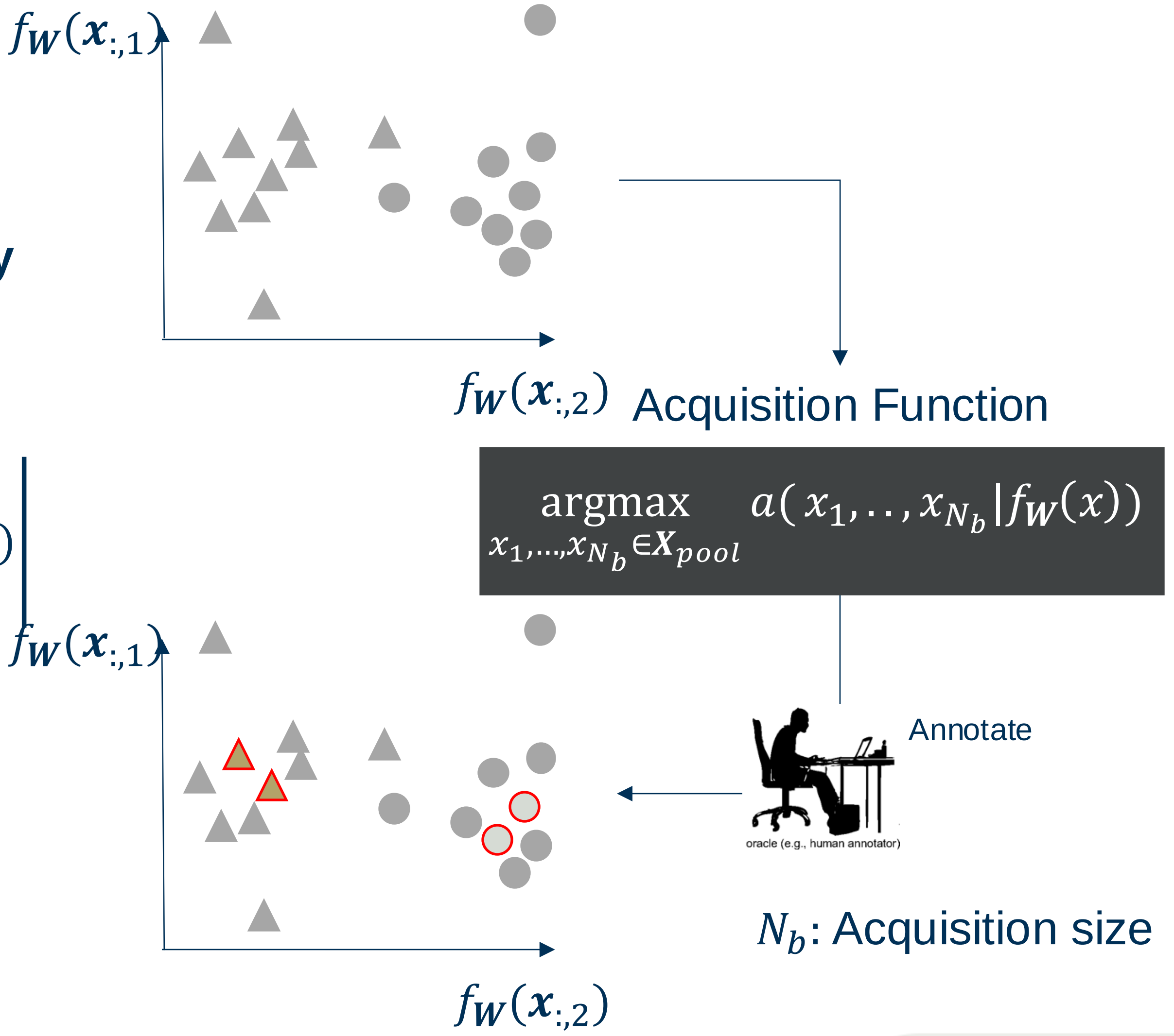
  - Core Set

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_W(\boldsymbol{x})\right) = -\left| \frac{1}{n} l_{pool} - \frac{1}{N_b} \sum_{i=1}^{N_b} l(\boldsymbol{x}_i, y_i; f_W) \right|$$

  - Discriminative Active Learning

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_W(\boldsymbol{x})\right) = -\sum_{i=1}^{N_b} p(y = u | \varphi(\boldsymbol{x}_i))$$

$\varphi(\boldsymbol{x}_i)$: Binary classifier for labeled/unlabeled



Coreset

- Select the most **representative** samples of the dataset
- Acquisition function based on the representation **directly**
- Popular examples:

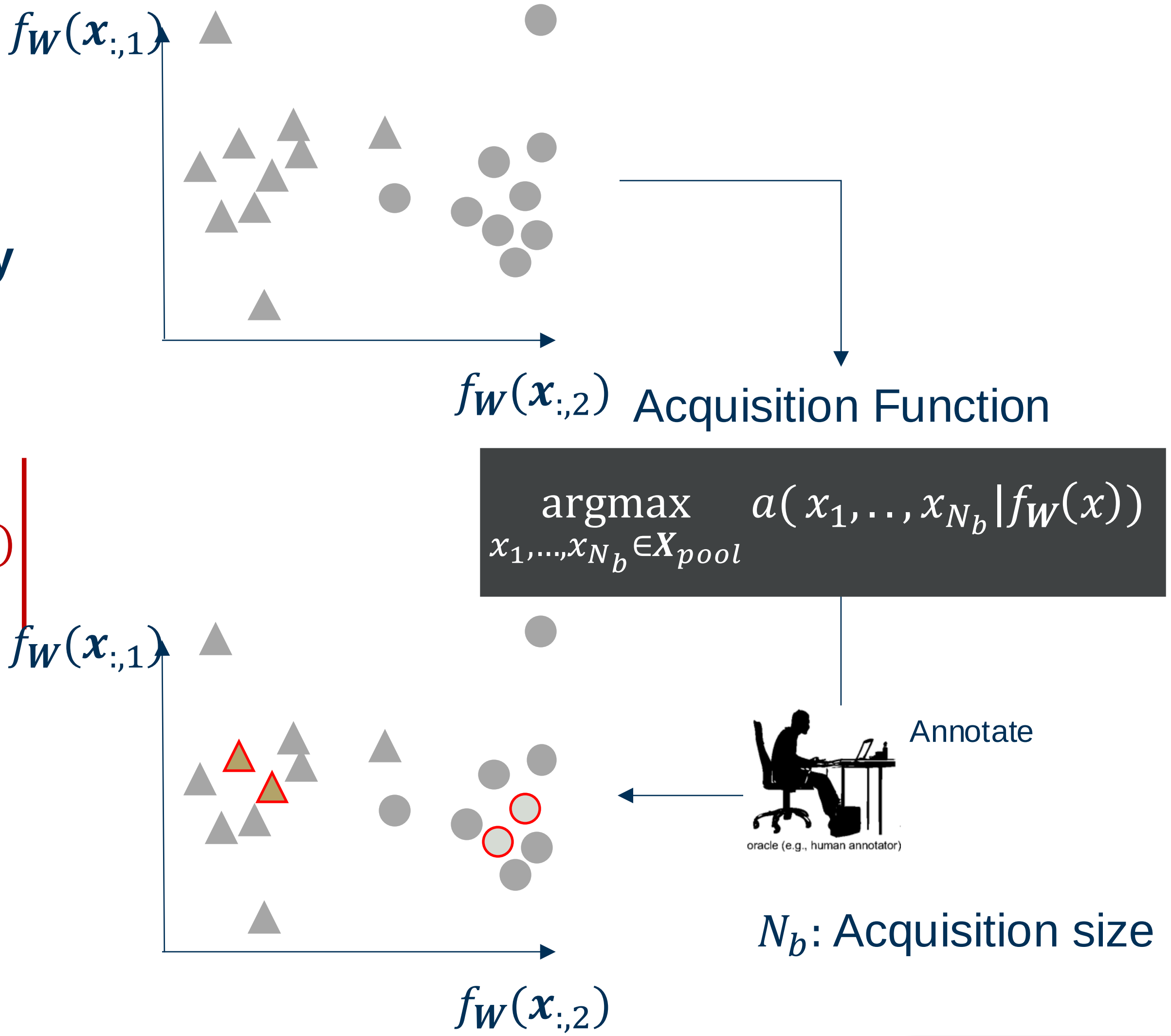
  - **Core Set**

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_W(\boldsymbol{x})) = -\left| \frac{1}{n} l_{pool} - \frac{1}{N_b} \sum_{i=1}^{N_b} l(\boldsymbol{x}_i, y_i; f_W) \right|$$

  - Discriminative Active Learning

$$a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} | f_W(\boldsymbol{x})) = -\sum_{i=1}^{N_b} p(y = u | \varphi(\boldsymbol{x}_i))$$

$\varphi(\boldsymbol{x}_i)$: Binary classifier for labeled/unlabeled



Coreset Sampling

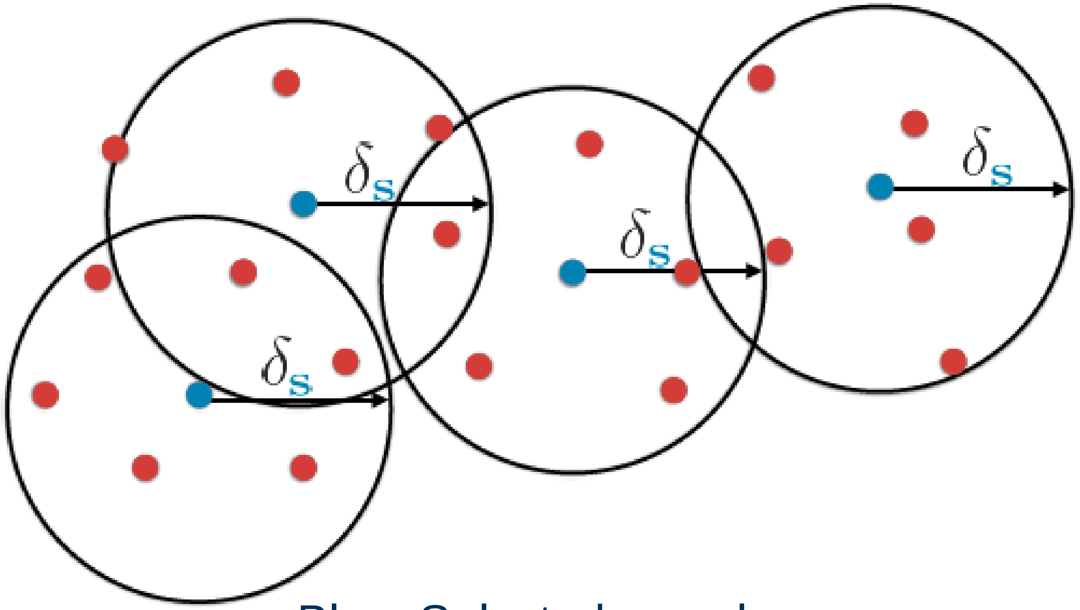
- Select the most **representative** samples of the dataset
- Greedy k-centers algorithm
  - Initialize small labeled subset
  - Calculate the utility function of unlabeled samples
  - Add the sample with the highest utility to the subset
  - Loop until convergence

- Core Set

$$a\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_b} \middle| f_W(\boldsymbol{x})\right) = -\left|\frac{1}{n} l_{pool} - \frac{1}{N_b}\sum_{i=1}^{N_b} l(\boldsymbol{x}_i, y_i; f_W)\right|$$

l_{pool}: Accumulated Loss of Unlabeled Pool
$l(\boldsymbol{x}_i, y_i; f_W)$: Loss of individual sample x_i

Coreset Example of a Dataset



Blue: Selected samples
Red: Unlabeled samples

Selected samples cover the dataset!

Discriminative Active Learning

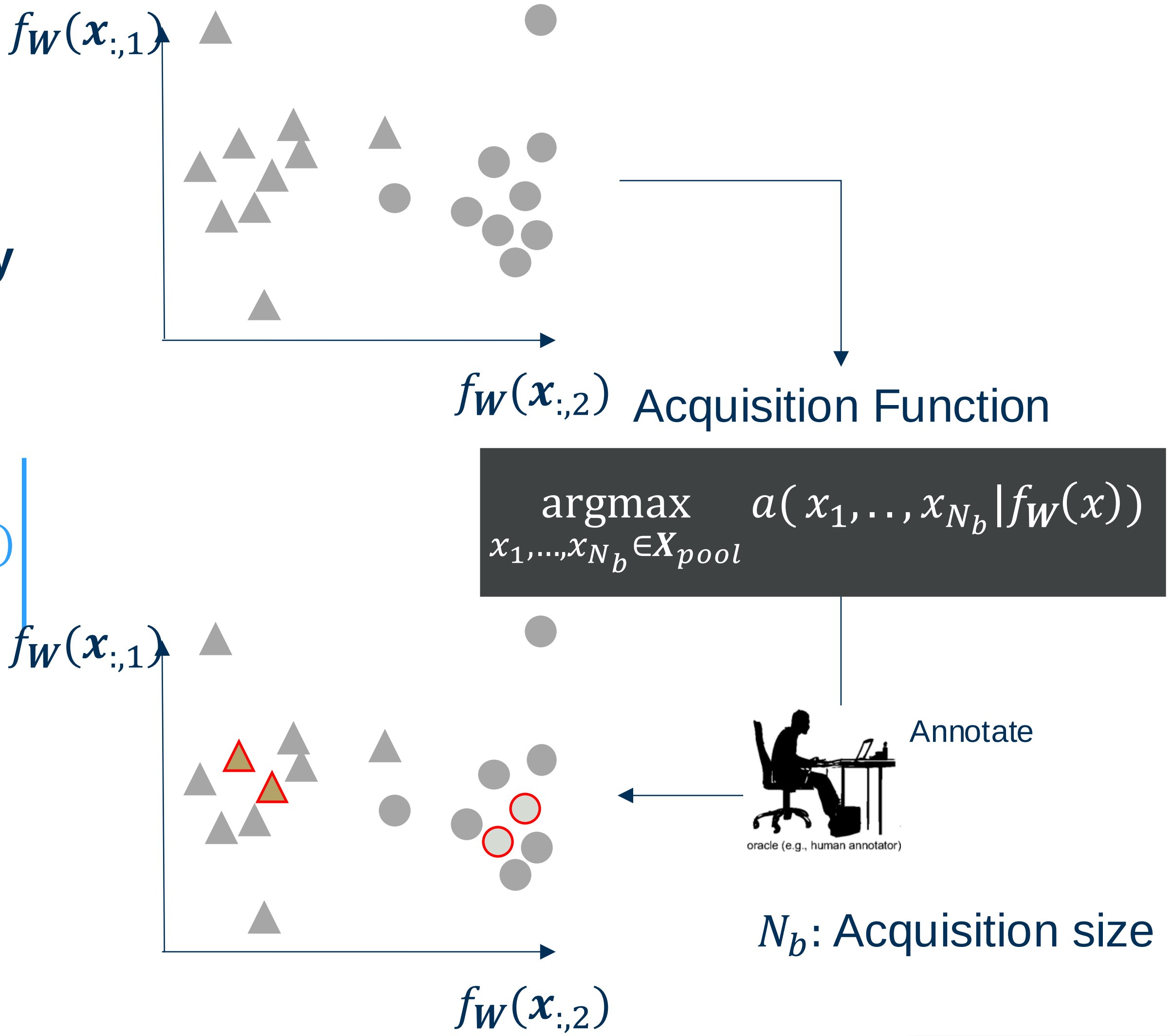
- Select the most **representative** samples of the dataset
- Acquisition function based on the representation **directly**
- Popular examples:
  - Core Set

$$a(\boldsymbol{x}_1,\dots,\boldsymbol{x}_{N_b}|f_W(\boldsymbol{x})) = -\left|\frac{1}{n} l_{pool} - \frac{1}{N_b}\sum_{i=1}^{N_b} l(\boldsymbol{x}_i, y_i; f_W)\right|$$

  - **Discriminative Active Learning**

$$a(\boldsymbol{x}_1,\dots,\boldsymbol{x}_{N_b}|f_W(\boldsymbol{x})) = -\sum_{i=1}^{N_b} p(y=u|\varphi(\boldsymbol{x}_i))$$

$\varphi(\boldsymbol{x}_i)$ **: Binary classifier for labeled/unlabeled**



Discriminative Sampling

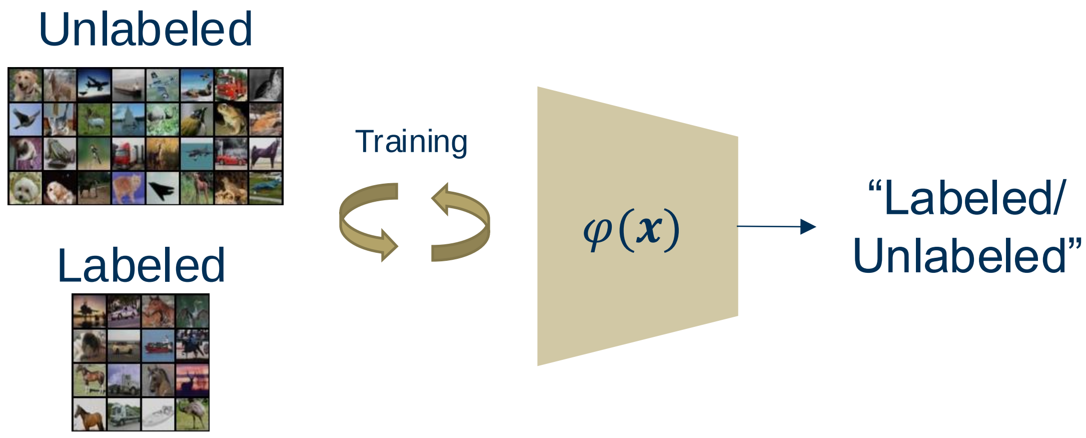
- Select the most **representative** samples of the dataset
- Selection workflow
  - Train binary classifier to distinguish labeled from unlabeled data
  - Evaluate binary classifier on unlabeled training pool
  - Annotate samples with highest probability of being unlabeled

- Discriminative active learning

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_W(\boldsymbol{x})\right) = -\sum_{i=1}^{N_b} p(y = u | \varphi(\boldsymbol{x}_i))$$

$\varphi(\boldsymbol{x}_i)$: Binary classifier for labeled/unlabeled

Step 1: Train Binary Classifier



Step 2: Select Points with Highest Unlabeled Probability

$$a\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N_b} \middle| f_W(\boldsymbol{x})\right) = -\sum_{i=1}^{N_b} p(y = u | \varphi(\boldsymbol{x}_i))$$

Overview

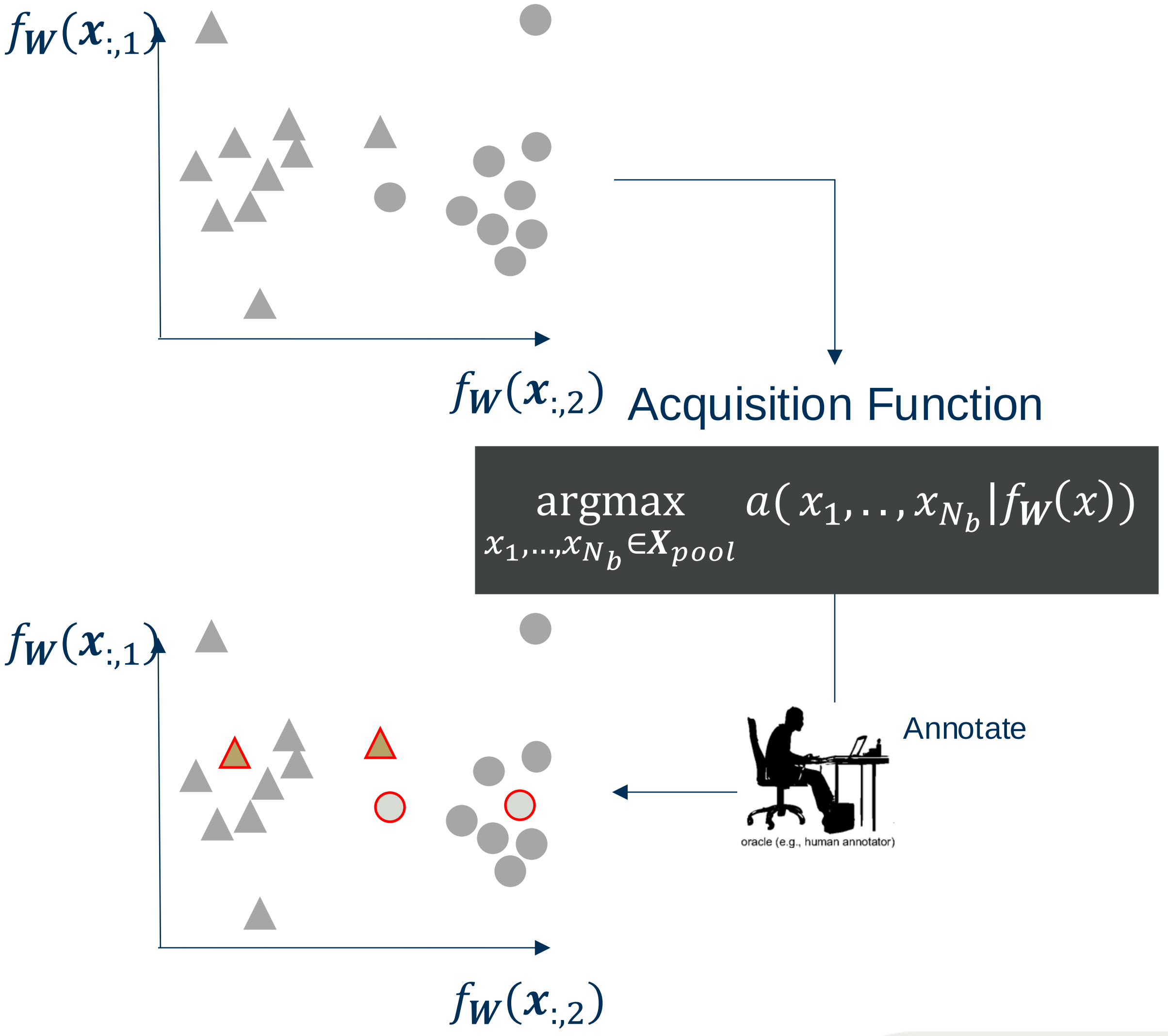
In this Lecture..

- Introduction and Motivation
- Fundamentals
- Performance Metrics
- Difficulty-based Active Learning
- Diversity-based Active Learning
- Fusion-based Active Learning
  - Motivation
- Deployment Performance Metric: Regression

Fusion-based Active Learning

- <u>Idea</u>: select the both **diverse and difficult** samples interactively

- Both difficult and diverse samples are important for performance:
  - e.g., diverse samples may be important in early rounds while difficult samples are more relevant at later rounds
  - Difficulty is typically integrated in sample ranking
  - Diversity frequently involves probabilistic sampling algorithms



Overview

In this Lecture..

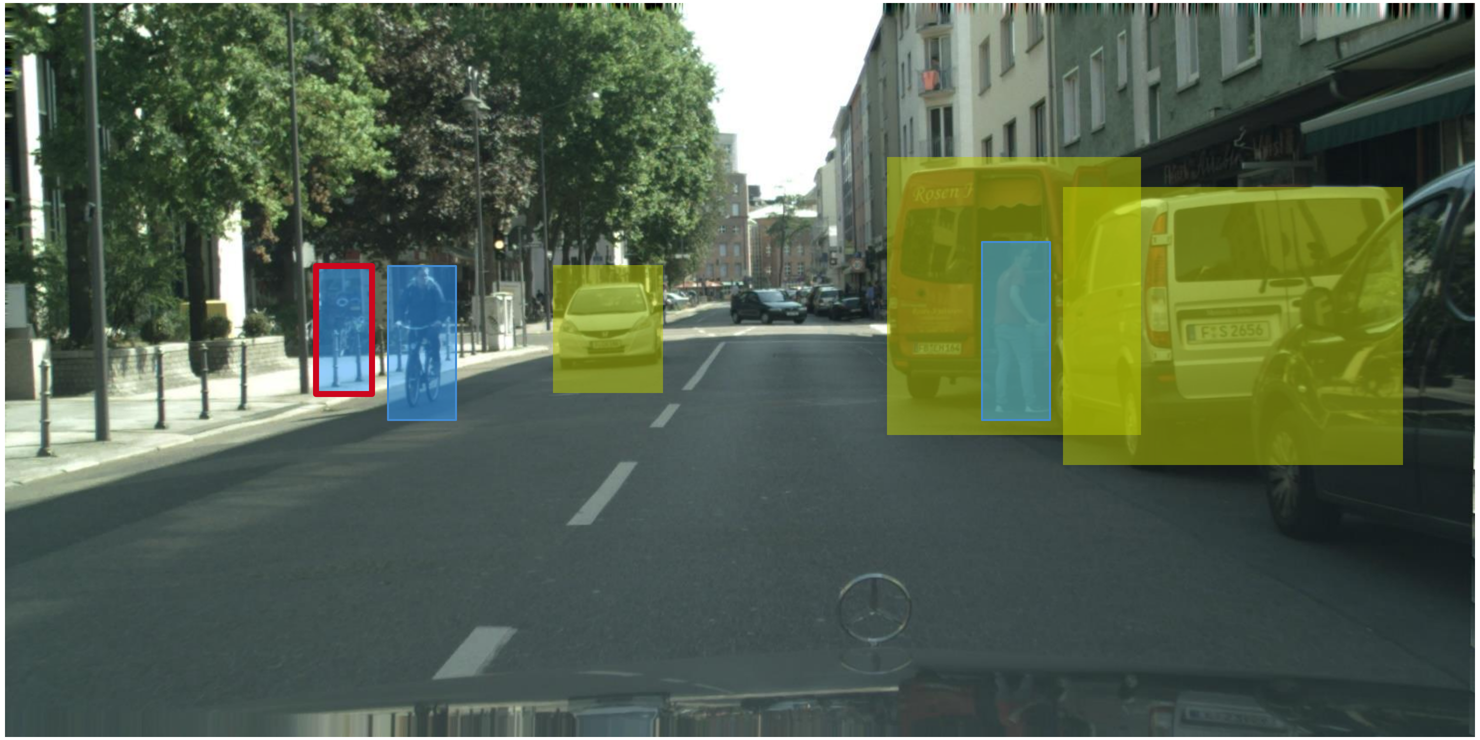
Performance Regression
Example

Regression Example in Autonomous Driving

Performance Regression

Example

Regression Example in Autonomous Driving



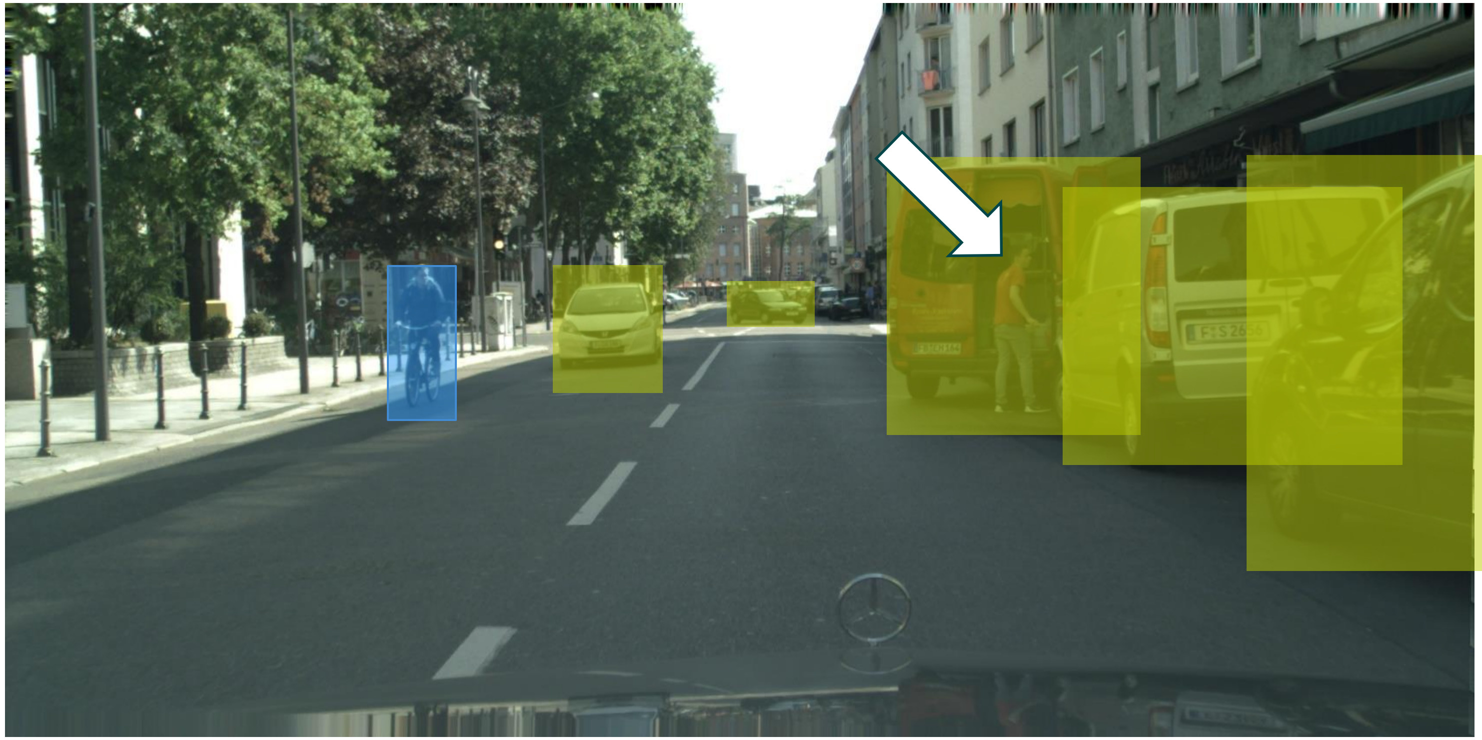
Active Learning Round N-1

Accuracy: **~87%**

Performance Regression

Example

Regression Example in Autonomous Driving



Active Learning Round **N**

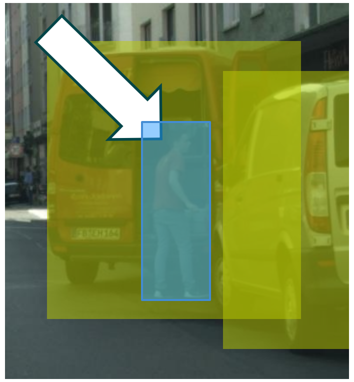
Accuracy: **~89%**

Performance Regression

Definition

Regression Example in Autonomous Driving

Active Learning Round N-1



Accuracy: ~87%

Regression

Active Learning Round N

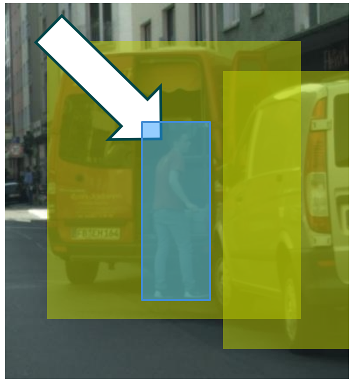
Accuracy: ~89%

Regression refers to the deterioration of performance after updates and puts model predictability at risk

Performance Regression
Definition

Regression Example in Autonomous Driving

Active Learning Round N-1



Accuracy: ~87%

Regression

Active Learning Round N

Accuracy: ~89%

Important for Active Learning:
- Stopping criteria
- Deployment decisions for online active learning
- Protocol quality

Performance Regression

Measuring Regression

Regression Manifests in Negative Flips

Both refers to across round N-1 and round N

Regression is **measured by the number of negative flips** in between active learning rounds

Performance Regression

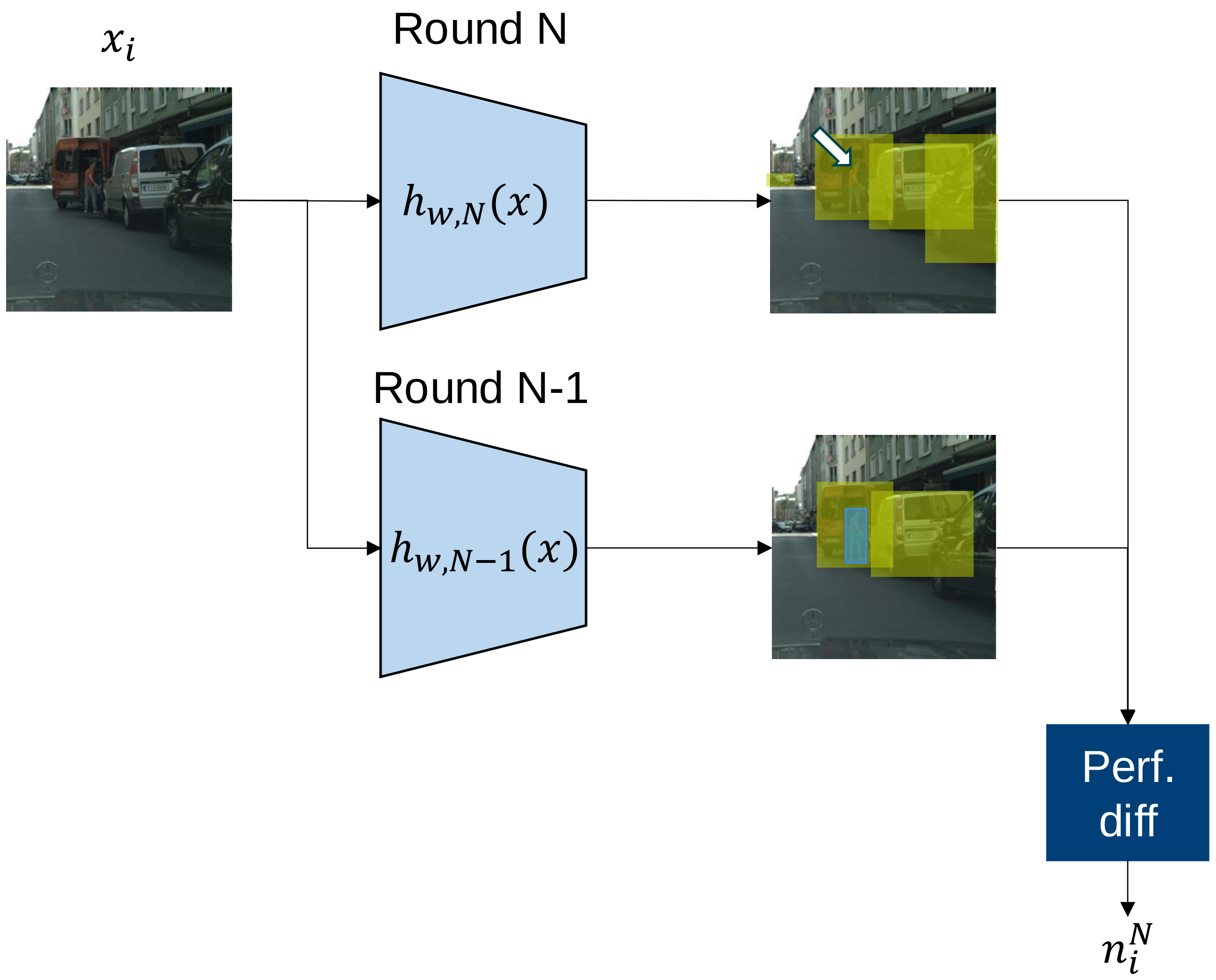
Measuring Regression

Negative Flip Rate (NFR):

Captures **performance deterioration** after updating the model

$$m = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{y_i'^{\,old}=y_i,\, y_i'^{\,new}!=y_i}$$

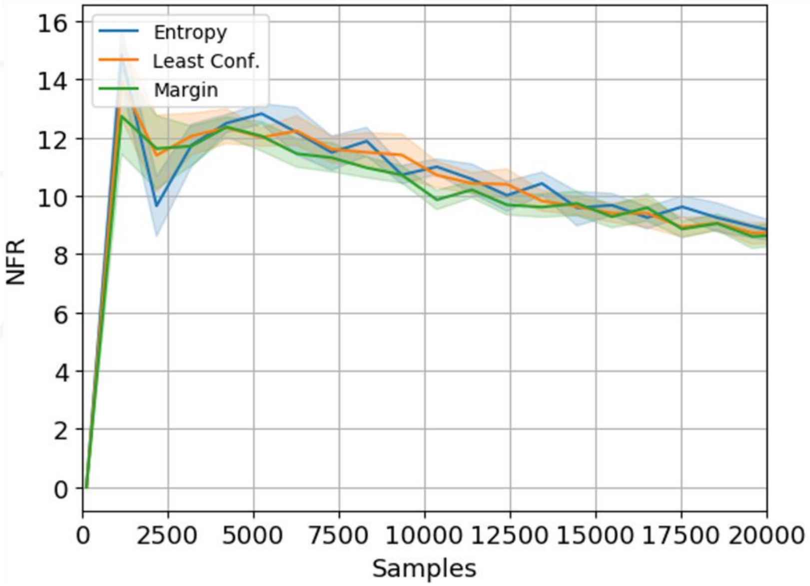
$y_i'^{\,old}$: Prediction of model at N-1
$y_i'^{\,new}$: Prediction of model at N
y_i: True label



Performance Regression

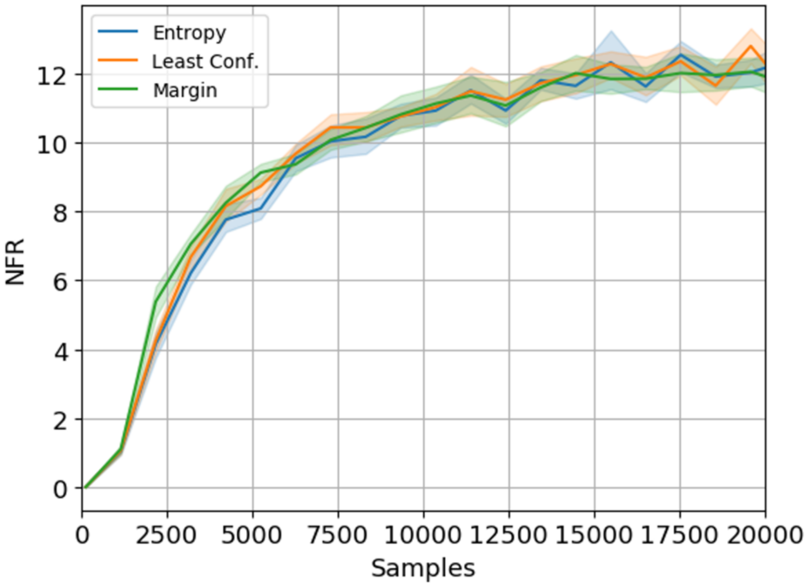
NFR in CIFAR10 vs CIFAR100

CIFAR10



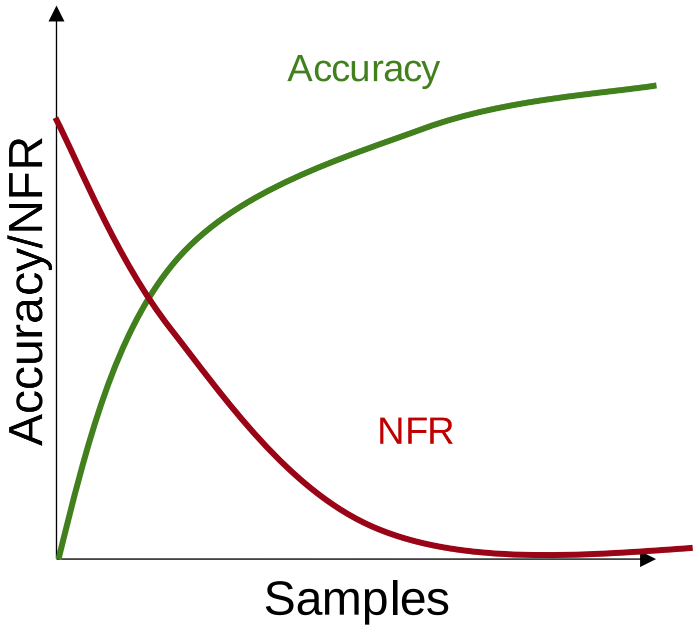
Expected NFR

CIFAR100



Unexpected NFR

Expectation



Appendix

References

Active Learning:
Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan. "Active learning with statistical models." *Journal of artificial intelligence research* 4 (1996): 129-145.

Difficulty-based Approaches:
Wang, Dan, and Yi Shang. "A new active labeling method for deep learning." *2014 International joint conference on neural networks (IJCNN)*. IEEE, 2014.

Diversity-based Approaches
Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." *arXiv preprint arXiv:1708.00489* (2017).

Gissin, Daniel, and Shai Shalev-Shwartz. "Discriminative active learning." *arXiv preprint arXiv:1907.06347* (2019).

Ash, Jordan T., et al. "Deep batch active learning by diverse, uncertain gradient lower bounds." *International Conference of Learning Representations* (2020)

Appendix
References

Fusion-based Approaches:
Ash, Jordan T., et al. "Deep batch active learning by diverse, uncertain gradient lower bounds." *International Conference of Learning Representations* (2020)

Kirsch, Andreas, Joost Van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning." *Advances in neural information processing systems* 32 (2019)

Active Learning under Limited Data:
Benkert, Ryan, Mohit Prabushankar, Ghassan AlRegib, Armin Pacharmi, and Enrique Corona. "Gaussian Switch Sampling: A Second Order Approach to Active Learning." *IEEE Transactions on Artificial Intelligence* (2023).

R. Benkert, M. Prabhushankar, and G. AlRegib, "Forgetful Active Learning With Switch Events: Efficient Sampling for Out-of-Distribution Data," in *IEEE International Conference on Image Processing (ICIP),* Bordeaux, France, 2022

Appendix
Notations

- x_i: a single feature
- $\boldsymbol{x}_i$: feature vector (a data sample)
- $\boldsymbol{x}_{:,i}$: feature vector of all data samples
- $\boldsymbol{X}$: matrix of feature vectors (dataset)
- $\boldsymbol{W}$: weight matrix
- $\boldsymbol{Z}$: latent representation
- E_θ: encoding function
- G_ϕ: decoding function
- $\widehat{\boldsymbol{X}}$: reconstruction of data
- N: number of data samples
- P: number of features in a feature vector
- $P^{(k)}$: the number of neurons in layer k
- α: learning rate
- Bold letter/symbol: vector
- Bold capital letters/symbol: matrix
- h_i and h_j representation space vectors
- z_i and z_j embedding space vectors