CONSENSUS EXTENSION TO ACTIVE LEARNING: THEORY AND APPLICATIONS

Pattern Recognition in Bioinformatics 2010

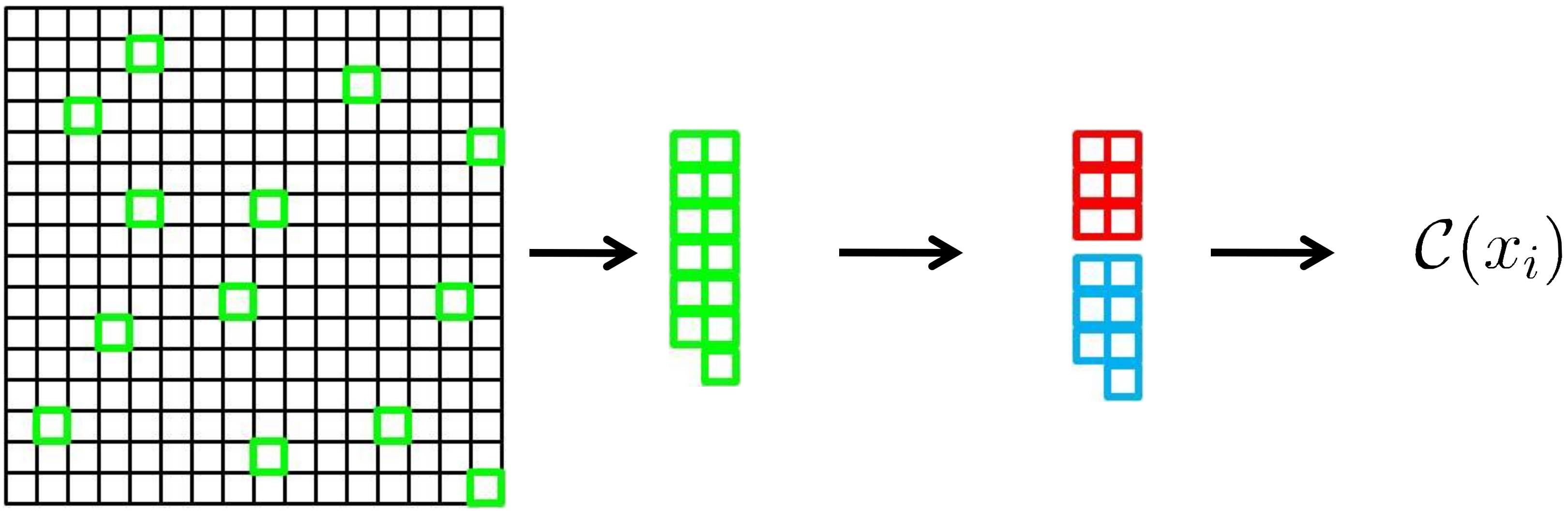
Scott Doyle and Anant Madabhushi

Rutgers University, Department of Biomedical Engineering, Piscataway, NJ

Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview
- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Consensus of Ambiguity: Combining AL methods
- Theory of CoA
- Experimental Results
- Concluding Remarks

Supervised Classification Paradigm



Dataset $x_i \in X$

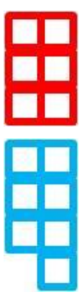
Randomly Sampled Training Points

Expert Annotated Training Points

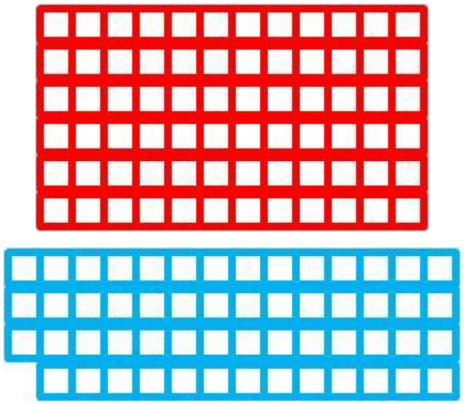
Supervised Classifier

Building Training with Random Learning

- Each sample is an observation of that sample's class
- With random sampling or learning (RL): **more samples are better** (more complete class model)
- Problem: Training samples are difficult to obtain!

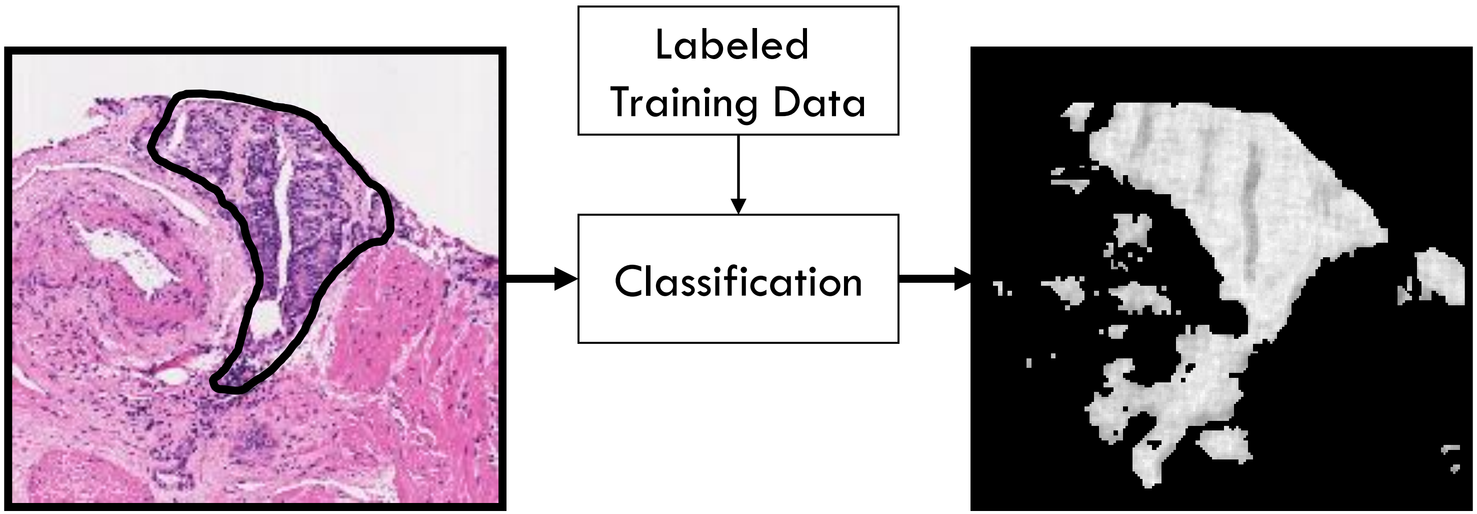


$\longrightarrow \mathcal{C}_1(x_i)$



$\longrightarrow \mathcal{C}_2(x_i)$

Building Training with Random Learning



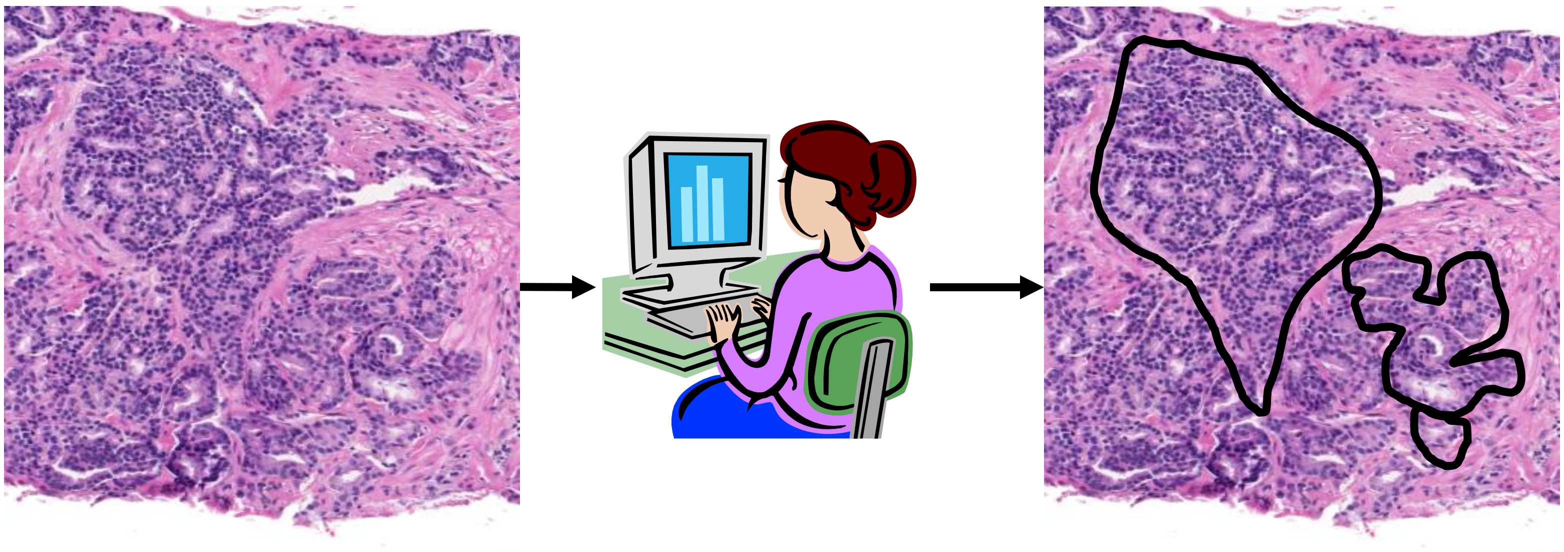
Cancer Region

Classification Result

- High accuracy requires training that is:
  - Accurate – Correctly labeled
  - Representative – Contains class information
  - Discriminative – Captures class differences

Doyle, et al. “Hierarchical Boosted Bayesian Ensemble for Prostate Cancer Detection from Digitized Histopathology", Biomedical Engineering, IEEE Transactions on. *(In Press)*

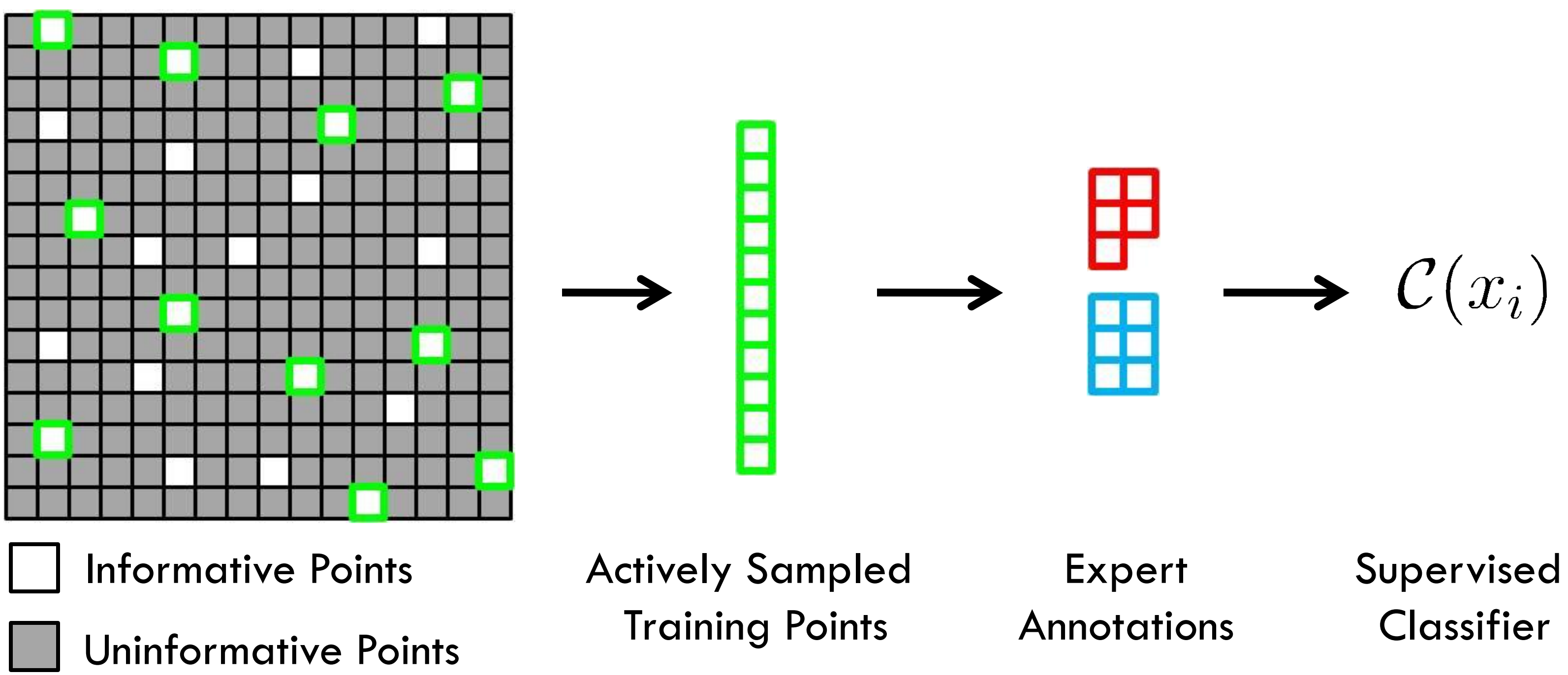
Building Training with Random Learning



- Expert medical knowledge is required
- Large images (1-2 GB): tedious, time-consuming to obtain detailed contours
- Each training image requires a great deal of effort

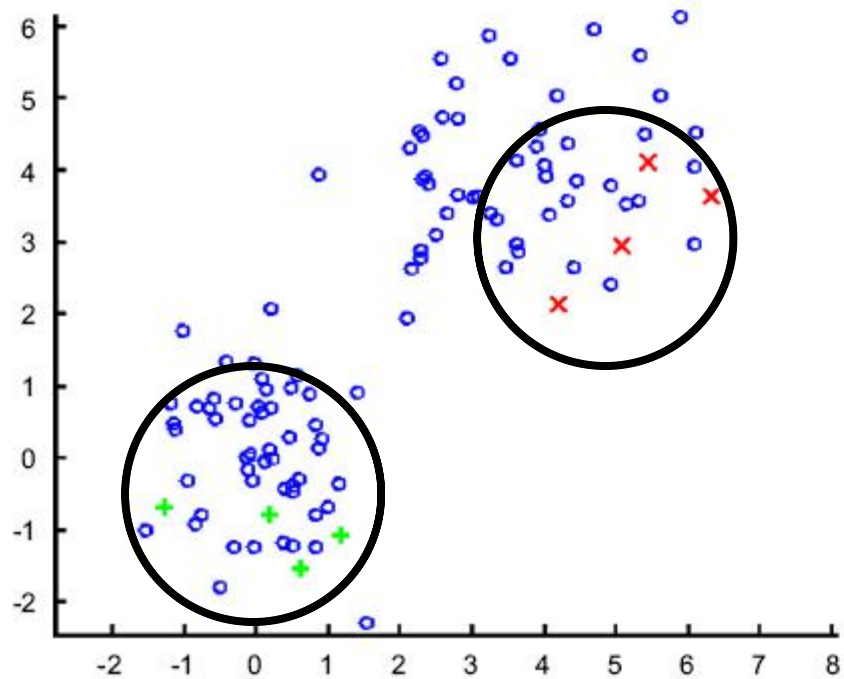
Active Learning (AL) Overview

- Active Learning (AL):
  - Selectively choose only informative samples for training
  - "Informative": samples that are likely to increase classifier performance



Active Learning (AL) Overview

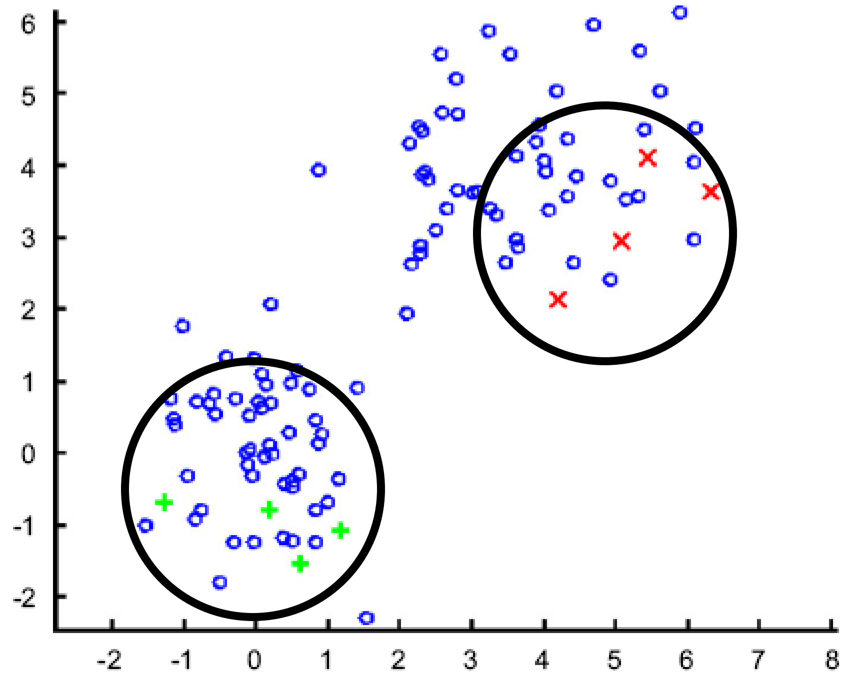
- How do we find "informative" annotation samples?
- Concept of "sample ambiguity"



Schohn and Cohn, "Less is More: Active Learning with Support Vector Machines." Proc. of the 17th Int. Conf. on Machine Learning, 2000, pp. 839-846.
Constantinopoulos, C. and Likas, A. "Semi-supervised and active learning with the probabilistic RBF classifier" Neurocomputing, Vol 71(13-15), pp. 2489-2498 (2006)

Active Learning (AL) Overview

- The more ambiguous a sample is, the more likely it is informative (should be selected for annotation)



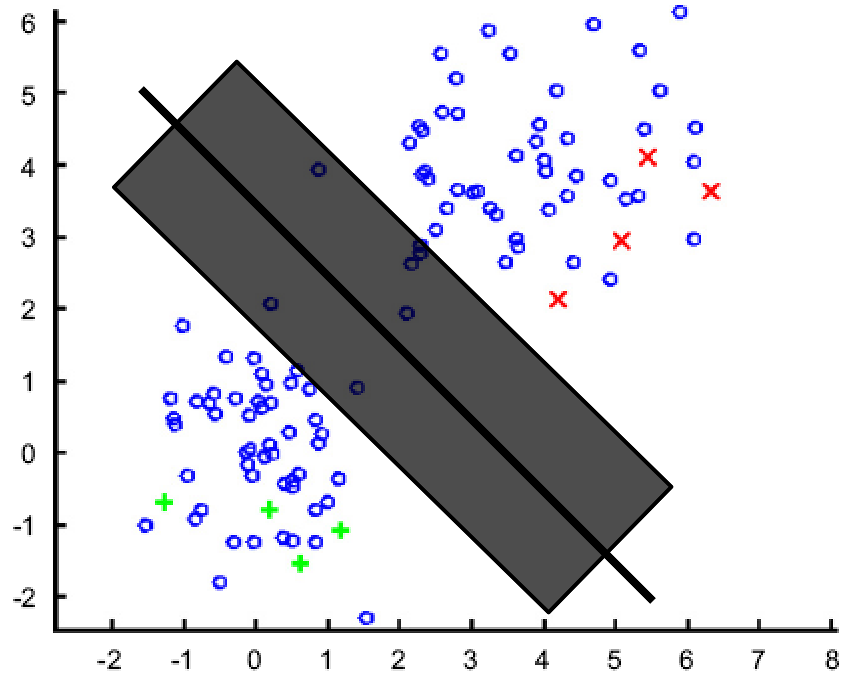
Schohn and Cohn, "Less is More: Active Learning with Support Vector Machines." Proc. of the 17th Int. Conf. on Machine Learning, 2000, pp. 839-846.
Constantinopoulos, C. and Likas, A. "Semi-supervised and active learning with the probabilistic RBF classifier" Neurocomputing, Vol 71(13-15), pp. 2489-2498 (2006)

Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview
- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Consensus of Ambiguity: Combining AL methods
- Theory of CoA
- Experimental Results
- Concluding Remarks

Measuring Sample Ambiguity

- Schohn (2000), Constantinopoulos (2006): SVMs
  - Distance to the decision hyperplane

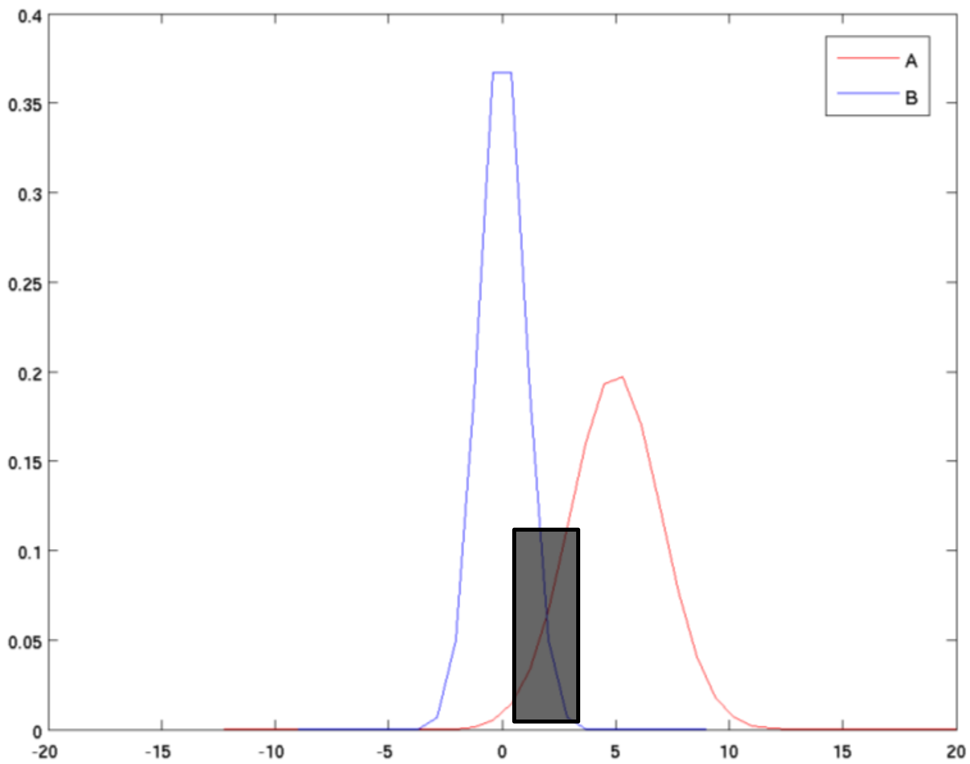


Schohn and Cohn, "Less is More: Active Learning with Support Vector Machines." Proc. of the 17th Int. Conf. on Machine Learning, 2000, pp. 839-846.
Constantinopoulos, C. and Likas, A. "Semi-supervised and active learning with the probabilistic RBF classifier" Neurocomputing, Vol 71(13-15), pp. 2489-2498 (2006)

Measuring Sample Ambiguity

- Bayes' Likelihood:
  - Based on likelihood of class membership

Posterior Probability of Class Membership



Feature Value

Measuring Sample Ambiguity

- Seung (1992), Freund (1997): Query-by-Committee
  - Based on disagreement among weak bagged classifiers

Weak Classifiers

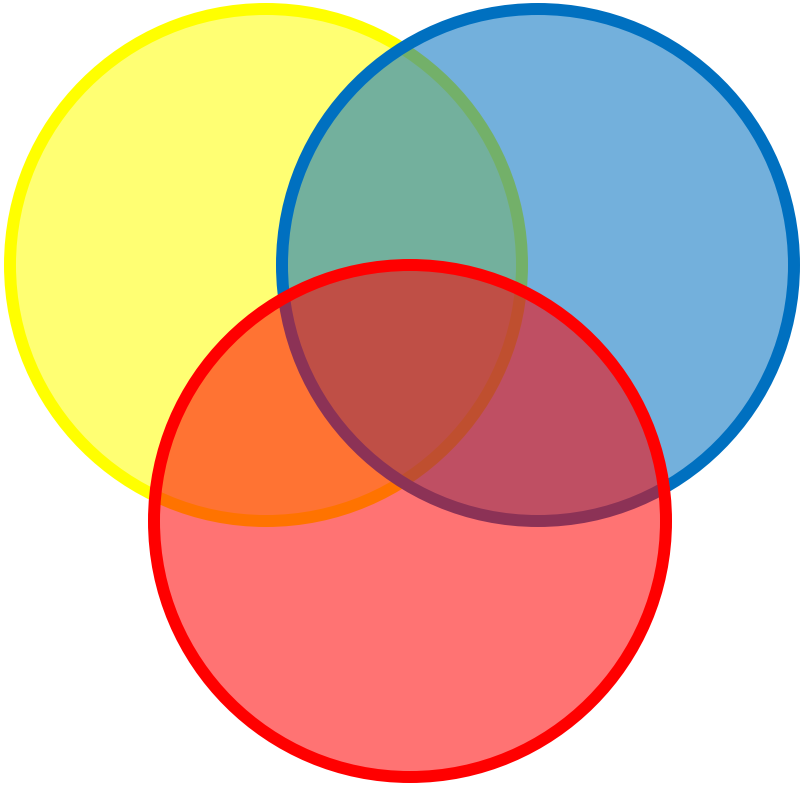
Average Classification

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	**TOTAL (0-1)**
1	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2/M = 0.2**
0	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**3/M = 0.3**
0	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**2/M = 0.2**
1	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**5/M = 0.5**

Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: 5th Annual ACM Workshop on Computational Learning Theory, pp. 287–294. (1992)
Freund, Y., et al. "Selective Sampling Using the Query by Committee Algorithm" Machine Learning, 28, 133–168 (1997)

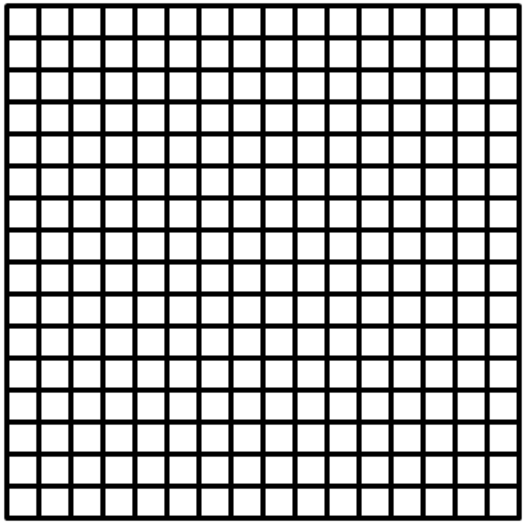
Combining AL Methods

- AL methods use one description of ambiguity each
- Ensemble methods combine multiple algorithms:
  - Variance is exploited to yield optimal results
  - Consensus classification: sample classification
  - Consensus Active Learning: sample ambiguity
- Consensus of Ambiguity (CoA)



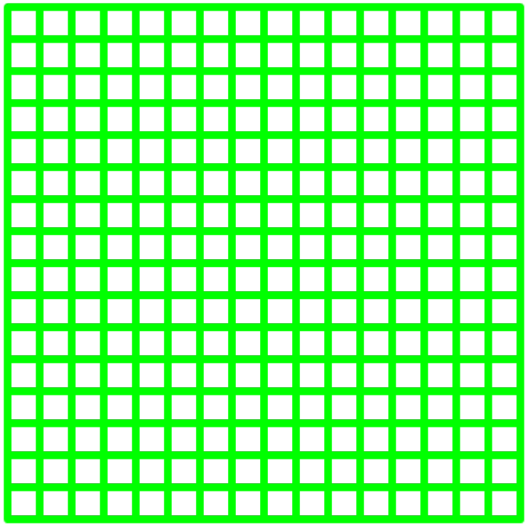
Advantage of CoA vs. AL

- RL: All samples to be annotated



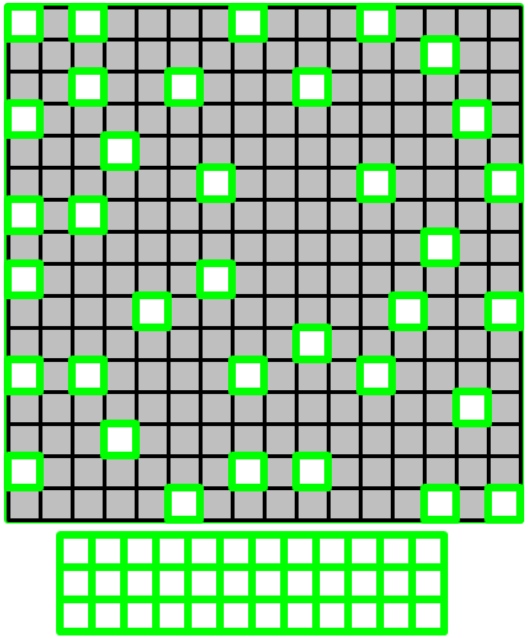
Advantage of CoA vs. AL

- RL: All samples to be annotated



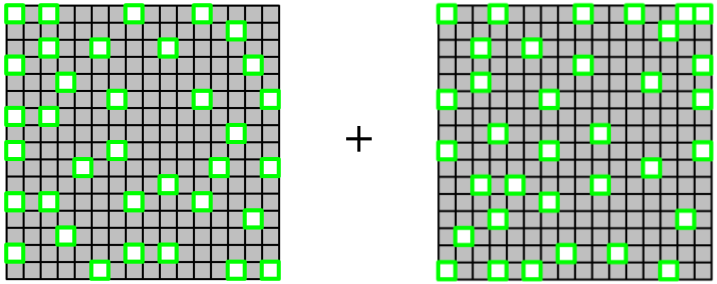
Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation



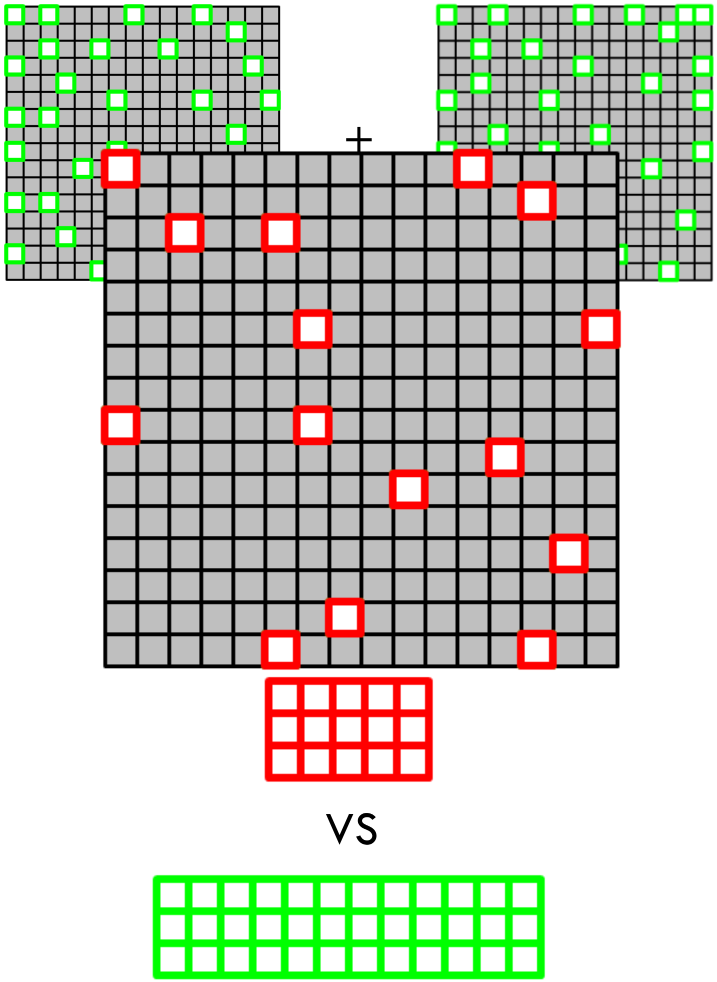
Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation
- CoA: FEWER eligible samples for annotation



Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation
- CoA: FEWER eligible samples for annotation



Overview

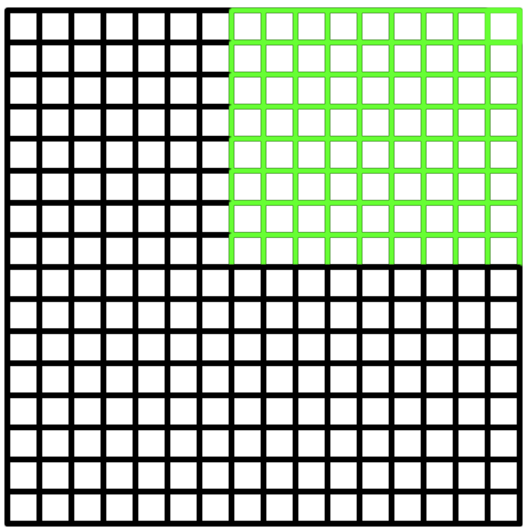
- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview
- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods
- Theory of CoA
- Experimental Results
- Concluding Remarks

CoA Theory: Specific Properties

- CoA Properties:
  - Multiple AL algorithms reduce ambiguous samples
  - Additional algorithms increase benefit of CoA
- Necessary Components:
  - General definition of ambiguous sample
  - Consensus among multiple algorithms (consensus ratio)
  - Identifying "strongly" ambiguous samples

CoA Theory: Basic Notation

Dataset X



Sample $x_i \in X$

Label $y_i \in \{w_1, w_2, \cdots, w_c\}$

Supervised Classifier $\mathcal{C}(x_i) \in \{w_1, w_2, \cdots, w_c\}$

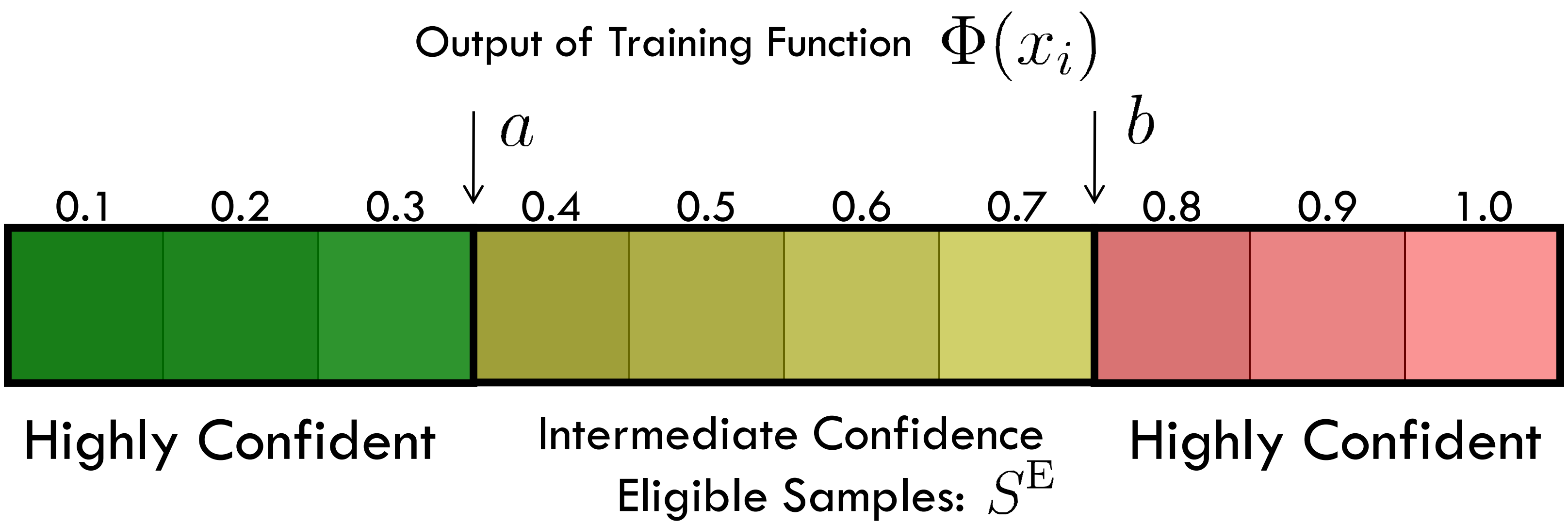
S^{tr} Training Set (Labeled)

Goal of the training algorithm: Build S^{tr} from unlabeled samples contained in X

Samples chosen according to a Training Function: $\Phi(x_i)$ which measures sample ambiguousness.

CoA Theory: Sample Ambiguity

- **Definition 1.** *A sample* $x_i \in X$ *is considered ambiguous if* $a < \Phi(x_i) < b$ *where a, b are lower and upper bounds for* $\Phi(x_i)$*, respectively.*



CoA Theory: Multiple Algorithms

- CoA employs multiple training algorithms:

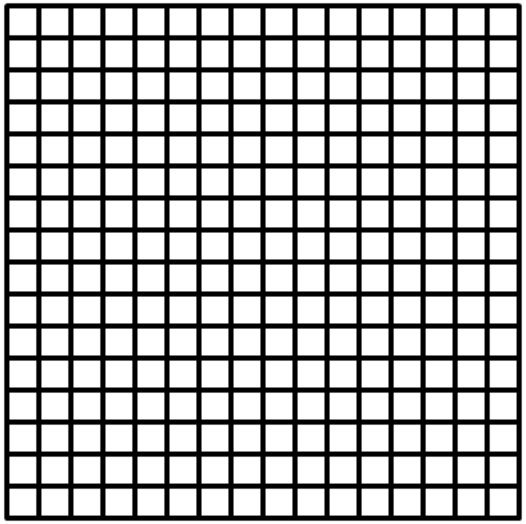
$$\Phi_j, j \in \{1, 2, \cdots, M\}$$

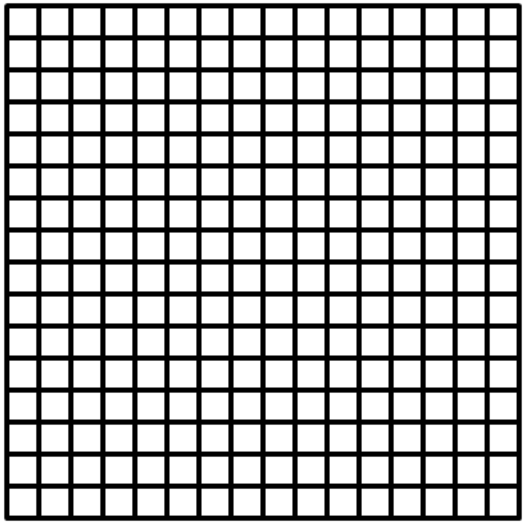
- Each algorithm returns a corresponding set of ambiguous (i.e. eligible-to-annotate) samples:

$$S_1^{\mathrm{E}}, S_2^{\mathrm{E}}, \cdots, S_M^{\mathrm{E}}$$

CoA Theory: Consensus Ratio

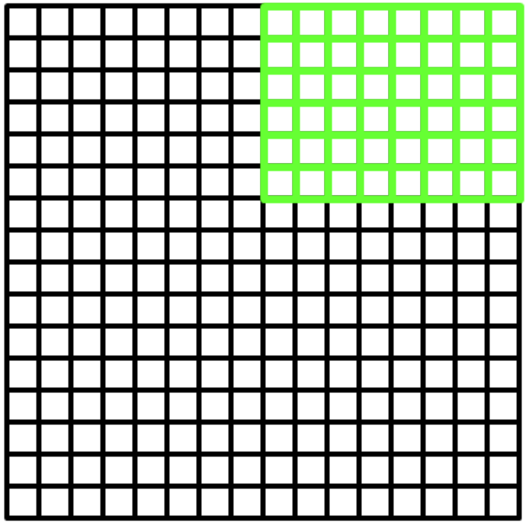
- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$ *.*

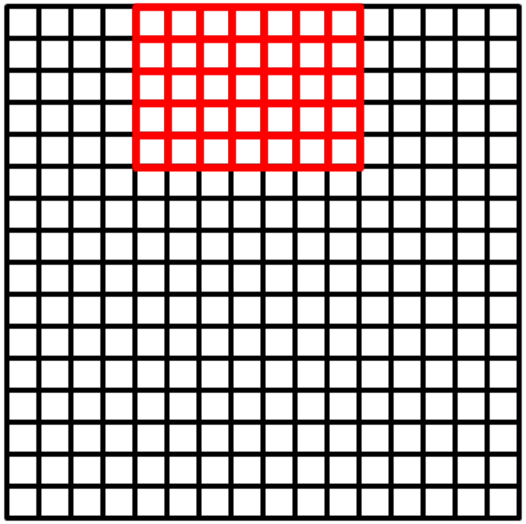




CoA Theory: Consensus Ratio

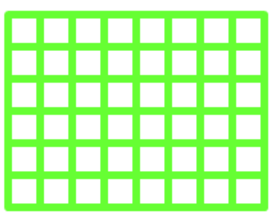
- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$ *.*



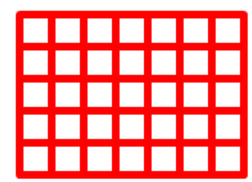


CoA Theory: Consensus Ratio

- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$ *.*



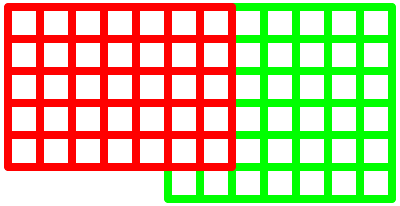
S_1^{E}



S_2^{E}

CoA Theory: Consensus Ratio

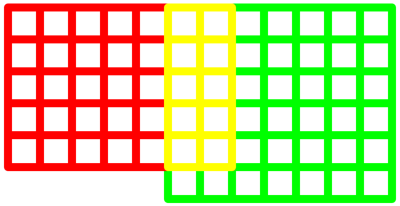
- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$.



CoA Theory: Consensus Ratio

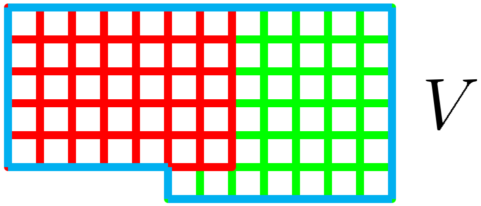
- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$.

U



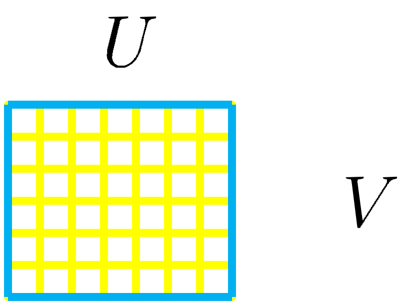
CoA Theory: Consensus Ratio

- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$.



CoA Theory: Consensus Ratio

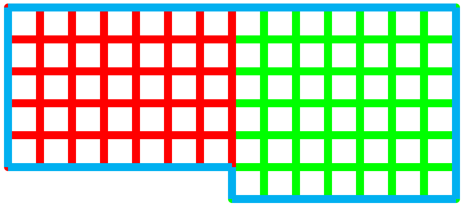
- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$.



If they overlap completely, then U = V and R = 1.

CoA Theory: Consensus Ratio

- **Definition 2.** *Given nonempty sets of ambiguous samples,* $S_j^{\mathrm{E}}, j \in \{1, 2, \cdots, M\}$ *, consensus ratio is defined as* $\mathcal{R} = U/V$ *where* $U = |\cap_{j=1}^{M} S_j^{\mathrm{E}}|$ *and* $V = |\cup_{j=1}^{M} S_j^{\mathrm{E}}|$ *.*

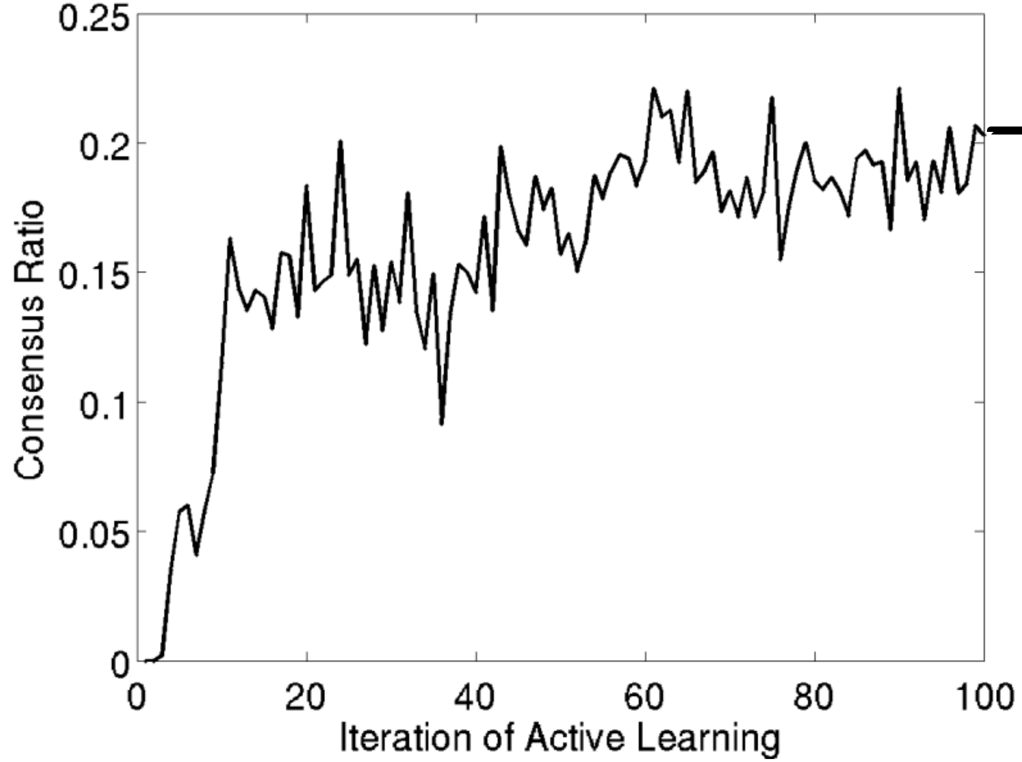


If they are independent, then U = 0, and R = 0.

Low ratios: greater benefit from the consensus scheme
High ratios: algorithms perform the same, so less benefit

CoA Theory: Consensus Ratio

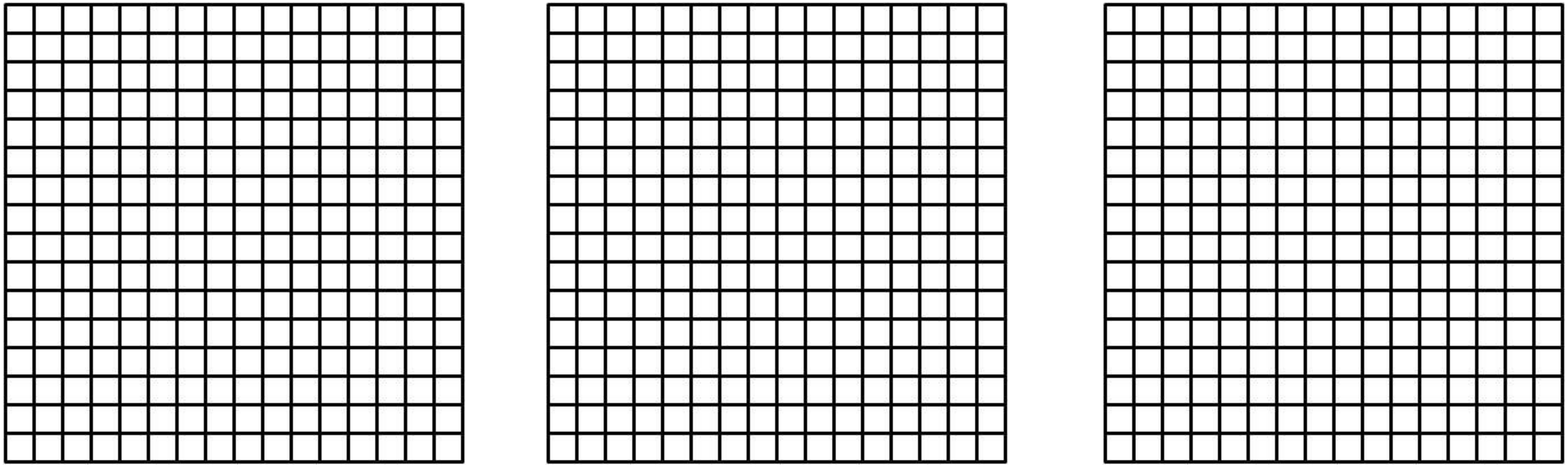
Consensus among three different AL algorithms



Plateau at 0.2: relatively little consensus
Motivates the use of ensemble approach

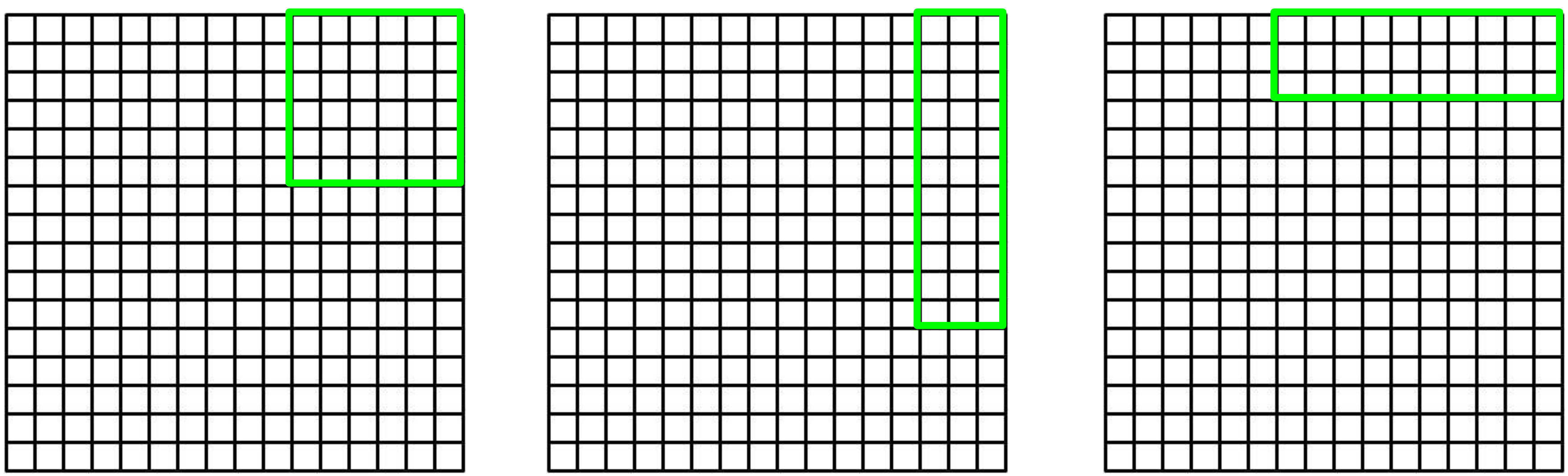
CoA Theory: Strong Ambiguity

- **Definition 3.** *A sample* $x_i \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^{\mathrm{E}} = \cap_{j=1}^{M} S_j^{\mathrm{E}}$ *; that is, if the sample is designated as ambiguous by* Φ_j *, for all algorithms* $j \in \{1, 2, \cdots, M\}$*.*



CoA Theory: Strong Ambiguity

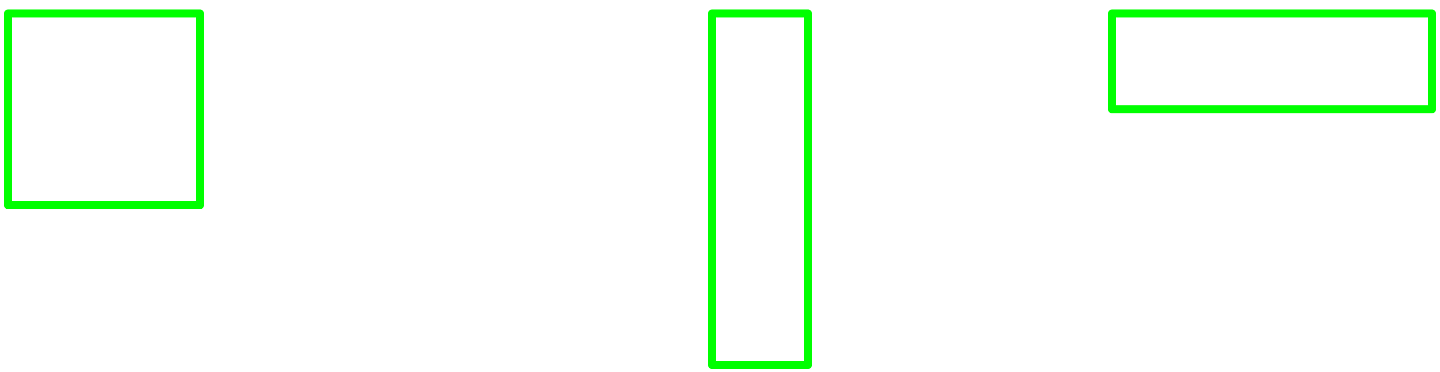
- **Definition 3.** *A sample* $x_i \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^{\mathrm{E}} = \cap_{j=1}^{M} S_j^{\mathrm{E}}$ *; that is, if the sample is designated as ambiguous by* Φ_j *, for all algorithms* $j \in \{1, 2, \cdots, M\}$.



S_1^{E} S_2^{E} S_3^{E}

CoA Theory: Strong Ambiguity

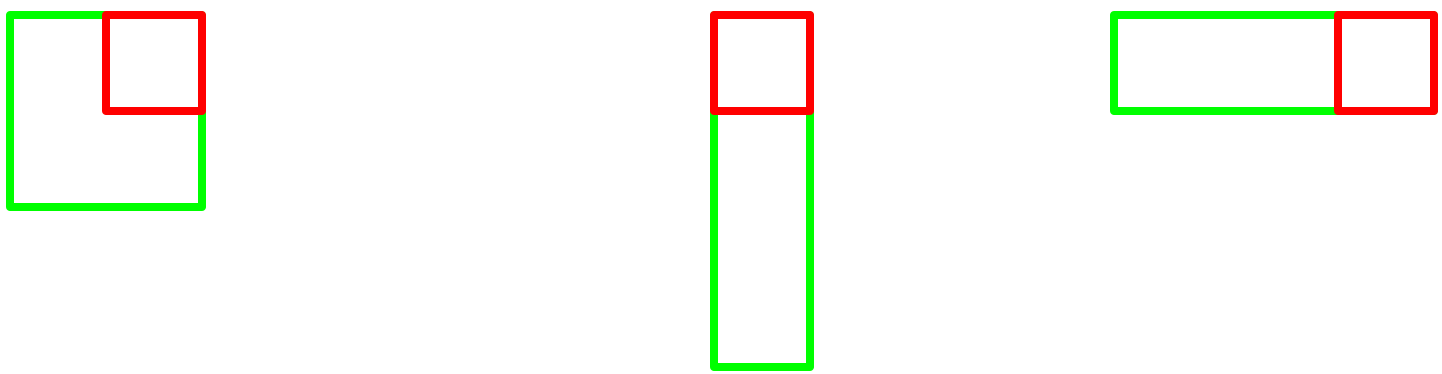
- **Definition 3.** *A sample* $x_i \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^{\mathrm{E}} = \cap_{j=1}^{M} S_j^{\mathrm{E}}$ *; that is, if the sample is designated as ambiguous by* Φ_j *, for all algorithms* $j \in \{1, 2, \cdots, M\}$.



S_1^{E} S_2^{E} S_3^{E}

CoA Theory: Strong Ambiguity

- **Definition 3.** *A sample* $x_i \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^{\mathrm{E}} = \cap_{j=1}^{M} S_j^{\mathrm{E}}$ *; that is, if the sample is designated as ambiguous by* Φ_j *, for all algorithms* $j \in \{1, 2, \cdots, M\}$.



S_1^{E} S_2^{E} S_3^{E}

CoA Theory: Strong Ambiguity

- **Definition 3.** *A sample* $x_i \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^{\mathrm{E}} = \cap_{j=1}^{M} S_j^{\mathrm{E}}$ *; that is, if the sample is designated as ambiguous by* Φ_j *, for all algorithms* $j \in \{1, 2, \cdots, M\}$.

S_1^{E} S_2^{E} S_3^{E}

CoA Theory: Addition of Algorithms

- Proposition. *As the number of algorithms being combined* M *increases, the consensus ratio* $\mathcal{R}$ *will either remain the same or will decrease.*

- Analogy to a sieve: as you add more layers of filtering, fewer samples will "get through".
- Remember: Small consensus ratio means better motivation for using consensus algorithm

Overview

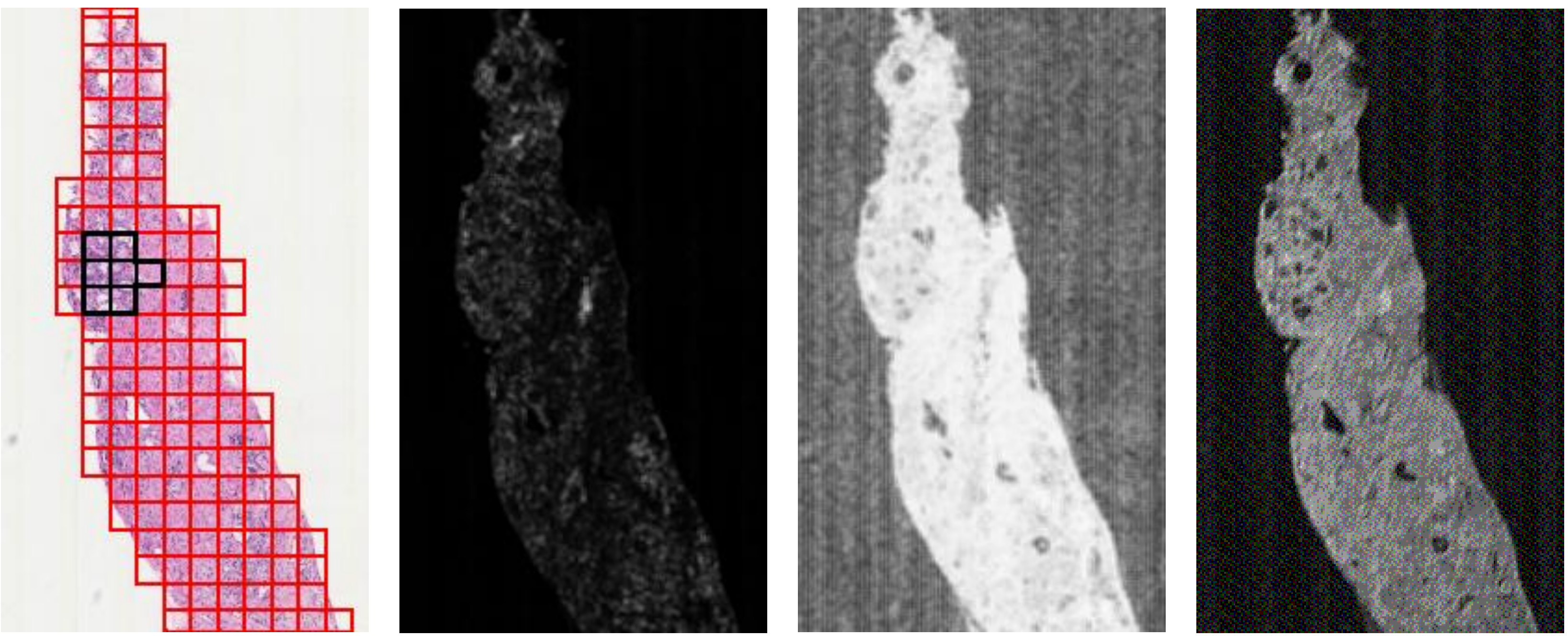
- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview
- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods
- Theory of CoA
- Experimental Results
- Concluding Remarks

Evaluating the Training Set

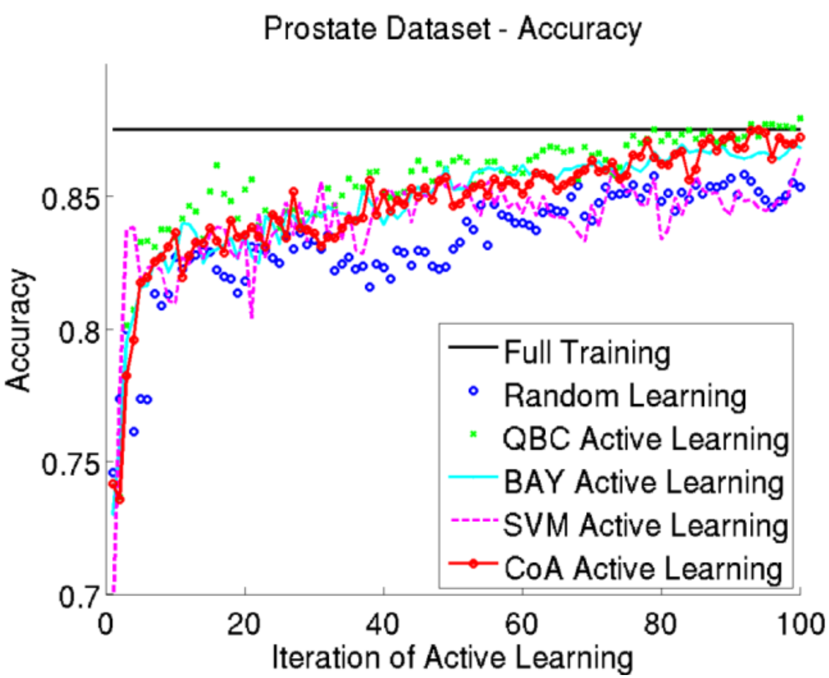
- Training set evaluation: Probabilistic Boosting Tree
- Two medical image analysis databases:
  - Prostate cancer detection from histopathology
  - Breast cancer grading from histopathology
- Three training algorithms:
  - Query-By-Committee (QBC)
  - Bayes Likelihood (BAY)
  - Support Vector Machine Distance (SVM)

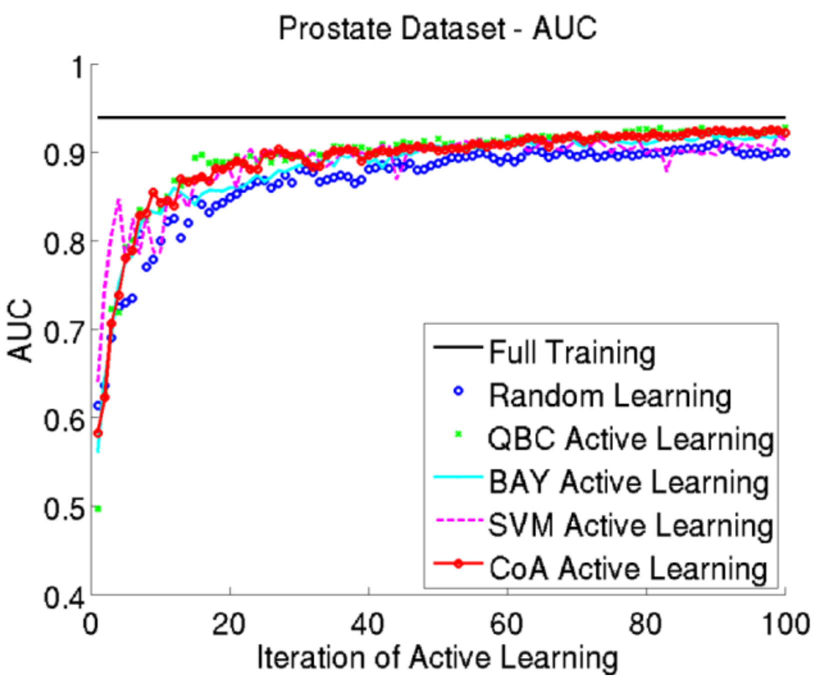
Experiment 1: Prostate Dataset

- Experiment 1 – Prostate Histopathology
  - 30 x 30 pixel grid on prostate biopsy samples
  - 14 texture features extracted from each ROI
  - 12, 000 ROIs classified

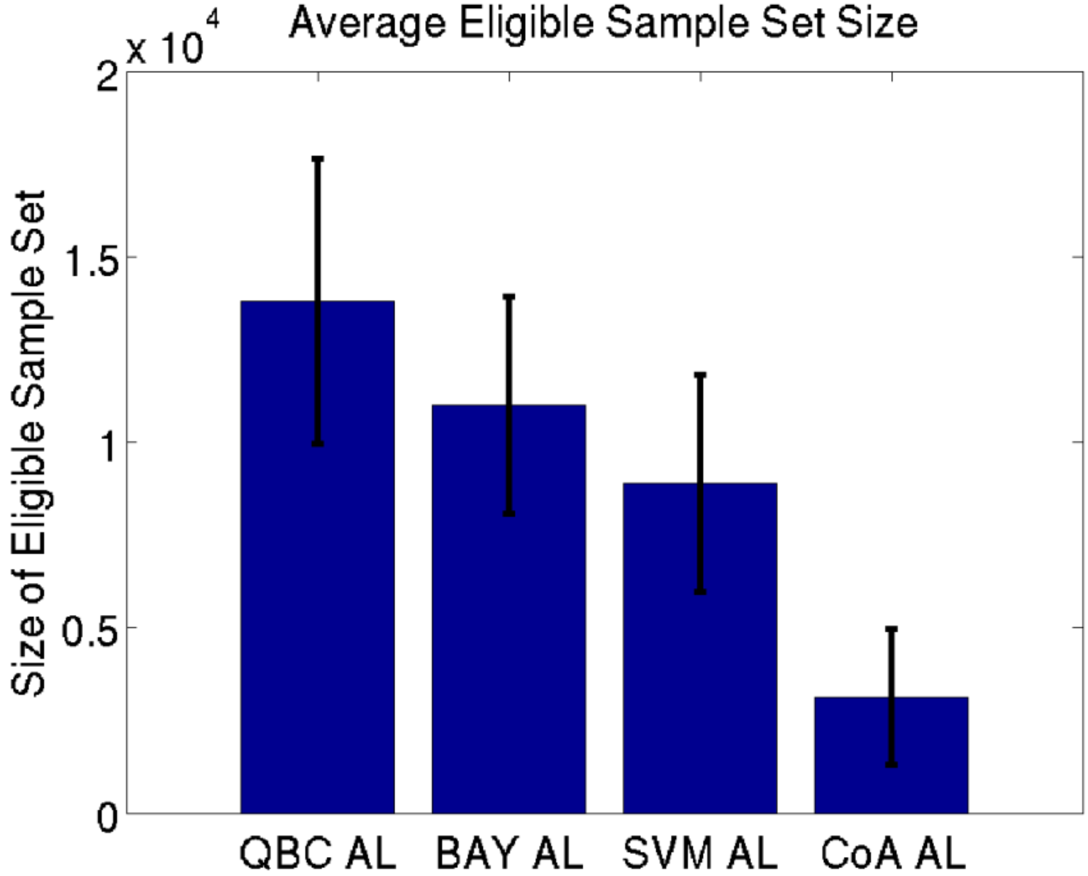


Experiment 1: Prostate Dataset



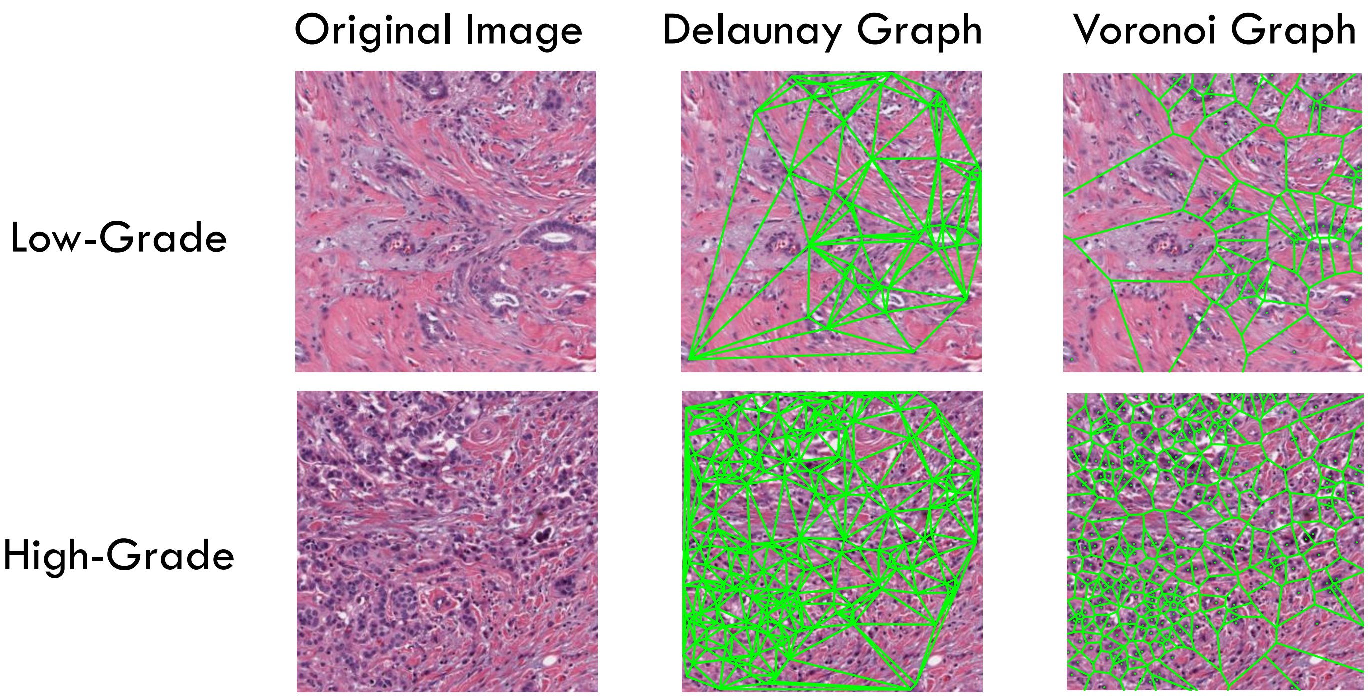


Experiment 1: Eligible Sample Size

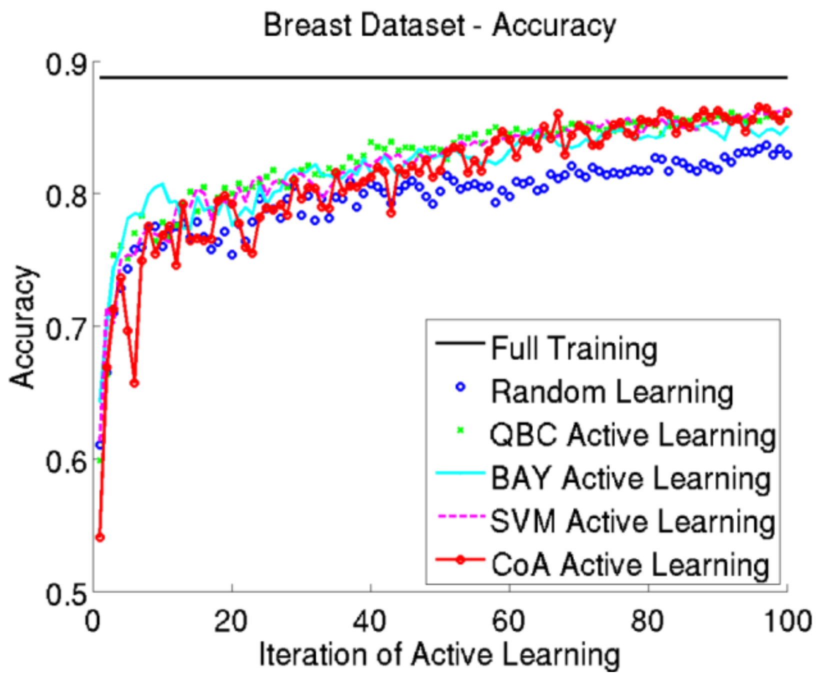


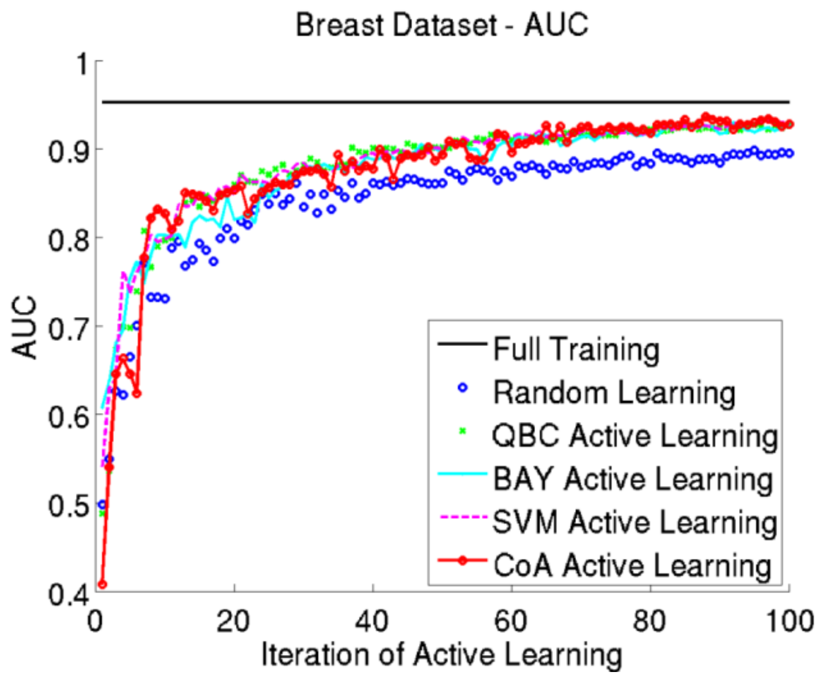
Experiment 2: Breast Grading

- Experiment 2 – Breast Histopathology Grading
  - 9,000 ROIs of homogeneous tissue (500 x 500 pixels)
  - Graph-based features to describe nuclear arrangement

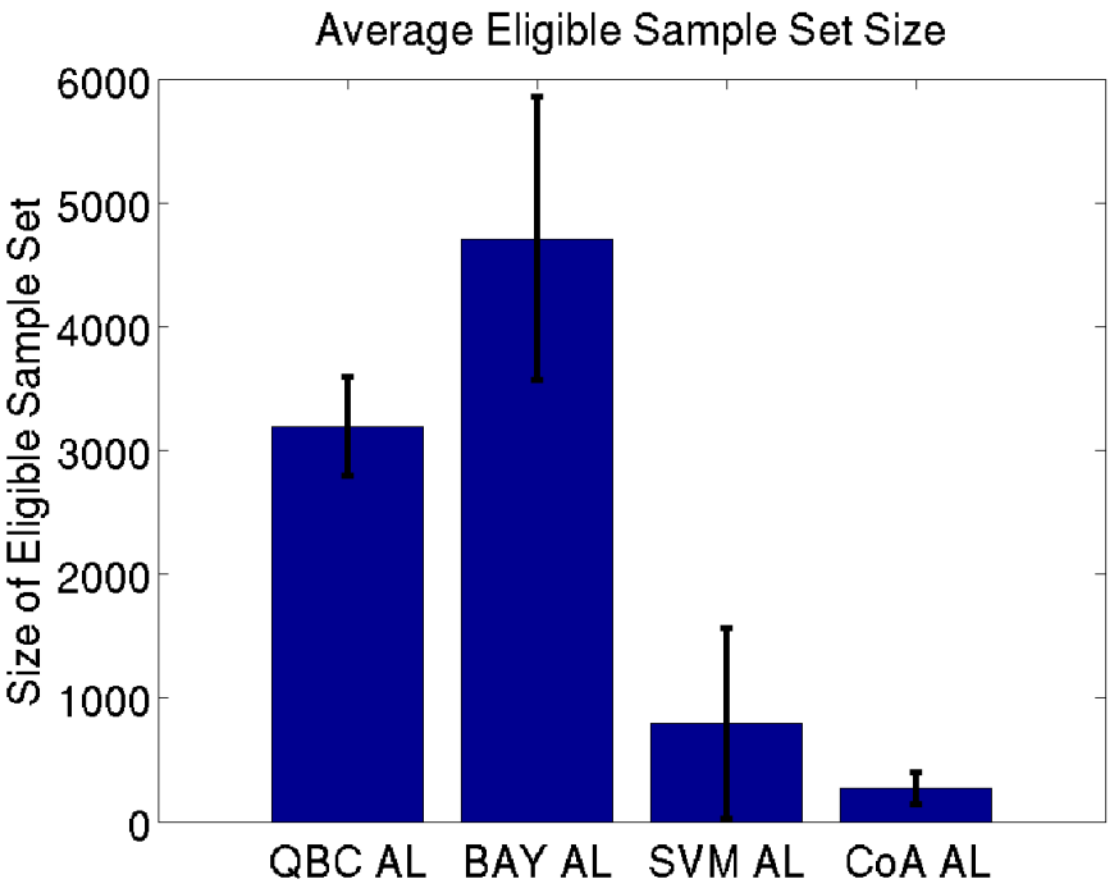


Experiment 2: Breast Grading





Experiment 2: Breast Grading



Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Overview of Active Learning (AL)
- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods
- Theory of CoA
- Experimental Results
- Concluding Remarks

Concluding Remarks

- CoA: Using multiple AL algorithms reduces set of informative samples, making annotation easier
- Additional methods increase CoA benefit if the consensus ratio decreases (but is still >0)
- Generalizable to any supervised classification problem where:
  - Data are costly, difficult to annotate
  - Target class is complex, RL requires many samples
  - Multiple AL algorithms can be leveraged simultaneously

Acknowledgments

- Wallace H. Coulter Foundation
- New Jersey Commission on Cancer Research
- National Cancer Institute
  - R01CA140772-01
  - R01CA136535-01
  - R21CA127186-01
  - R03CA143991-01
- Cancer Institute of New Jersey
- US Department of Defense (W81XWH-08-1-0145)
- Bioimagene, Inc.