# Enhanced Supervised Learning with Unlabeled Data

**Sally Goldman and Yan Zhou**
**Department of Computer Science**
**Washington University**
**St. Louis**

# A.K.A. Co-Training

But no need for either

- Redundantly sufficient features *or*
- Respecting the class distributions explictly when adding labeled examples

# So Where's the Magic?

The magic comes from

- Using two learners with different biases
  - Each hopefully finds patterns in the data hidden to the other
  - Each *must* produce a hypothesis that partitions the instance space
- Using statistics to
  - Analyse the expected performance of both hypotheses at each stage
  - Analyse the expected performance of each partition within each hypothesis at each stage
  - Decide when to label examples for the other learner
- Keeping two sets of automatically labeled examples

# Eh?

"To estimate the accuracy of (the hypotheses) we use a 95%-confidence interval for a binomial parameter. For each hypothesis and for each equivalence class within the two hypotheses we use *10-fold cross validation* to compute how many correct predictions are made."

# When to Label?

Learner A should label an example *x* for learner B only when:

- A's confidence in its label for *x* is better than B's
  - Use the confidence intervals for this
- The amount of data labeled for B will decrease the estimated error rate
  - Use the lower bound of the confidence interval to estimate # of mislabeled examples
  - Use a funky equation to estimate existing and potential error rates.

# Show Me The Money

Experiment:

- 8 UCI problems
- Unlabeled data: Test data plus some of the labeled data
- 3 runs each
- ID3 and HOODG are the two learners

Results

| Algorithm | Average Error Rate |
|---|---|
| ID3 | 16.4% |
| HOODG | 18.9% |
| Better of ID3 & HOODG | 14.1% |
| Co-Training | 11.9% |