# Outline

- noun phrase coreference resolution
- a (supervised) machine learning approach
  - evaluation
  - problems…some solutions
- weakly supervised approaches

Knowledge-based approaches are still common. E.g.

- Lappin & Leass [1994]
- CogNIAC [Baldwin, 1996]

# A Machine Learning Approach

- Classification
  - given a description of two noun phrases, $NP_i$ and $NP_j$, classify the pair as *coreferent* or *not coreferent*

*coref* ? *coref* ?

[Queen Elizabeth] set about transforming [her] [husband], ...

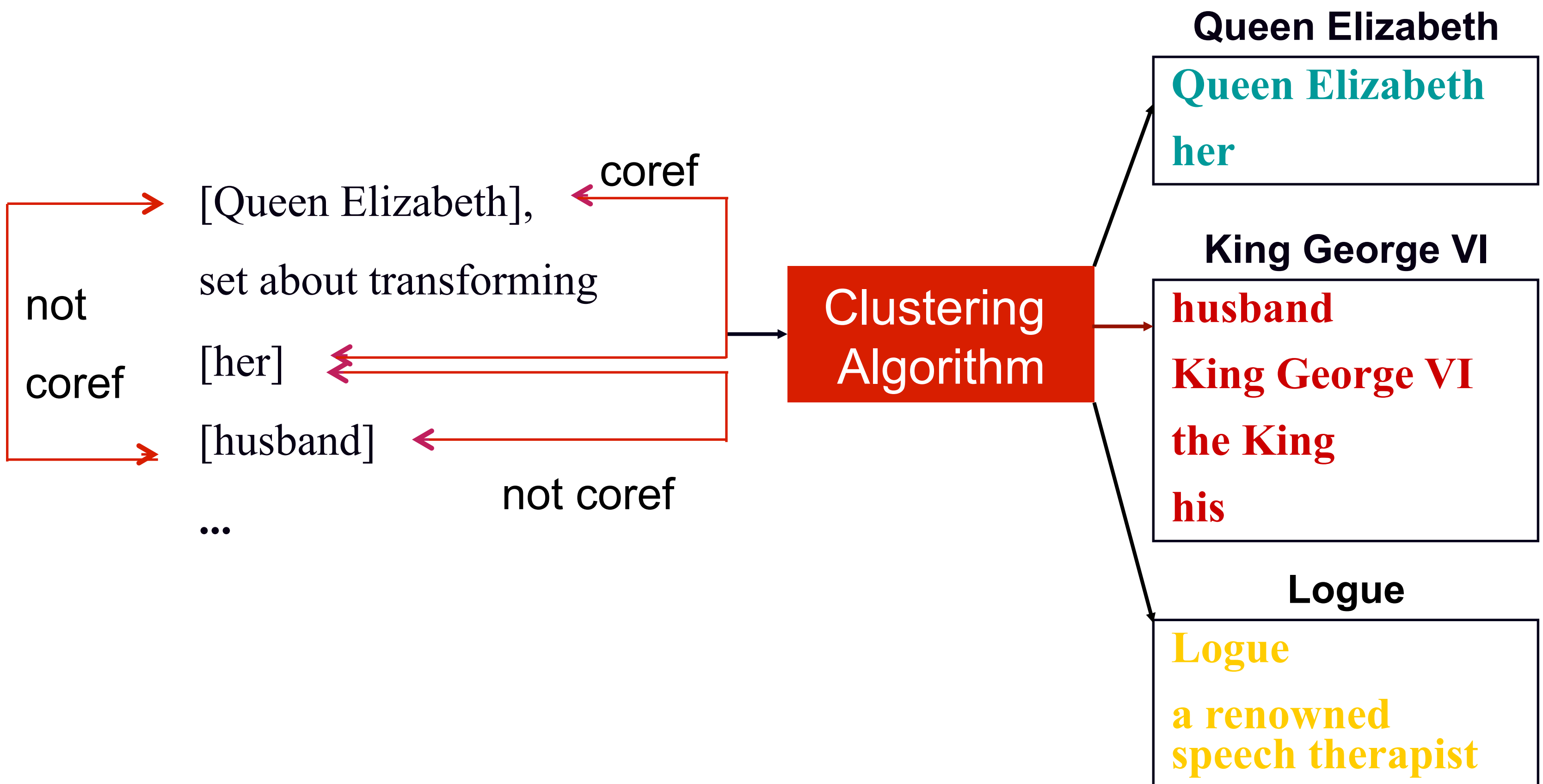
*coref* ?

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon et al. [2001]; Ng & Cardie [2002]; …

CORNELL

# A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



# Training Data Creation

- Creating training instances
  - texts annotated with coreference information

candidate antecedent anaphor

  - one instance $inst(NP_i, NP_j)$ for each *ordered* pair of NPs
    - » $NP_i$ precedes $NP_j$
    - » feature vector: describes the two NPs and context
    - » class value:

| | |
|---|---|
| *coref* | pairs on the same coreference chain |
| *not coref* | otherwise |

# Instance Representation

- 25 features per instance
  - lexical (3)
    - » string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - » pronoun_1, pronoun_2, demonstrative_2, indefinite_2, …
    - » number, gender, animacy
    - » appositive, predicate nominative
    - » binding constraints, simple contra-indexing constraints, …
    - » span, maximalnp, …
  - semantic (2)
    - » same WordNet class
    - » alias
  - positional (1)
    - » distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - » naïve pronoun resolution algorithm

# Learning Algorithm

- RIPPER (Cohen, 1995)
  C4.5 (Quinlan, 1994)
  - rule learners
    - » input: set of training instances
    - » output: coreference classifier

- Learned classifier
    - » input: test instance (represents pair of NPs)
    - » output: classification
      confidence of classification

# Clustering Algorithm

- Best-first single-link clustering
  - Mark each $NP_j$ as belonging to its own class: $NP_j \in c_j$
  - Proceed through the NPs in left-to-right order.
    - » For each NP, $NP_j$, create test instances, $inst(NP_i, NP_j)$, for all of its preceding NPs, $NP_i$.
    - » Select as the antecedent for $NP_j$ the highest-confidence coreferent NP, $NP_i$, according to the coreference classifier (or none if all have below .5 confidence);

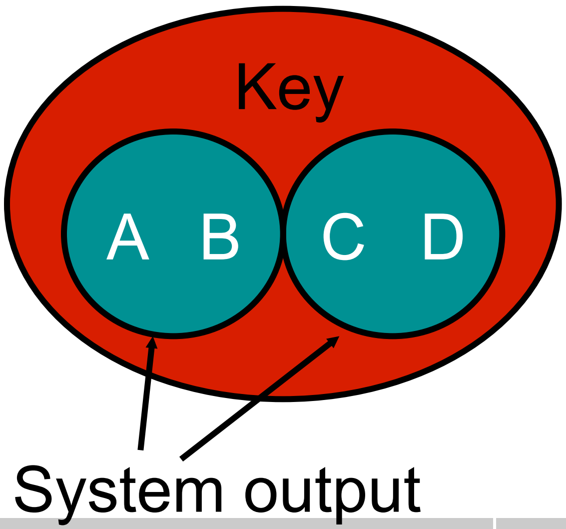
      Merge $c_j$ and $c_j$ .

# Outline

- noun phrase coreference resolution
- a (supervised) machine learning approach
  - evaluation
  - problems…some solutions
- weakly supervised approaches

# Evaluation

- MUC-6 and MUC-7 coreference data sets
- documents annotated w.r.t. coreference
- 30 + 30 training texts (dry run)
- 30 + 20 test texts (formal evaluation)
- scoring program
  - recall
  - precision
  - F-measure: 2PR/(P+R)



# Results

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Ng & Cardie** | 63.3 | 76.9 | **69.5** | 54.2 | 76.3 | **63.4** |
| **Best MUC System** | 59 | 72 | **65** | 56.1 | 68.8 | **61.8** |

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 40.7 | 73.5 | **52.4** | 27.2 | 86.3 | **41.3** |
| **Worst MUC System** | 36 | 44 | 40 | 52.5 | 21.4 | 30.4 |
| **Best MUC System** | 59 | 72 | 65 | 56.1 | 68.8 | 61.8 |

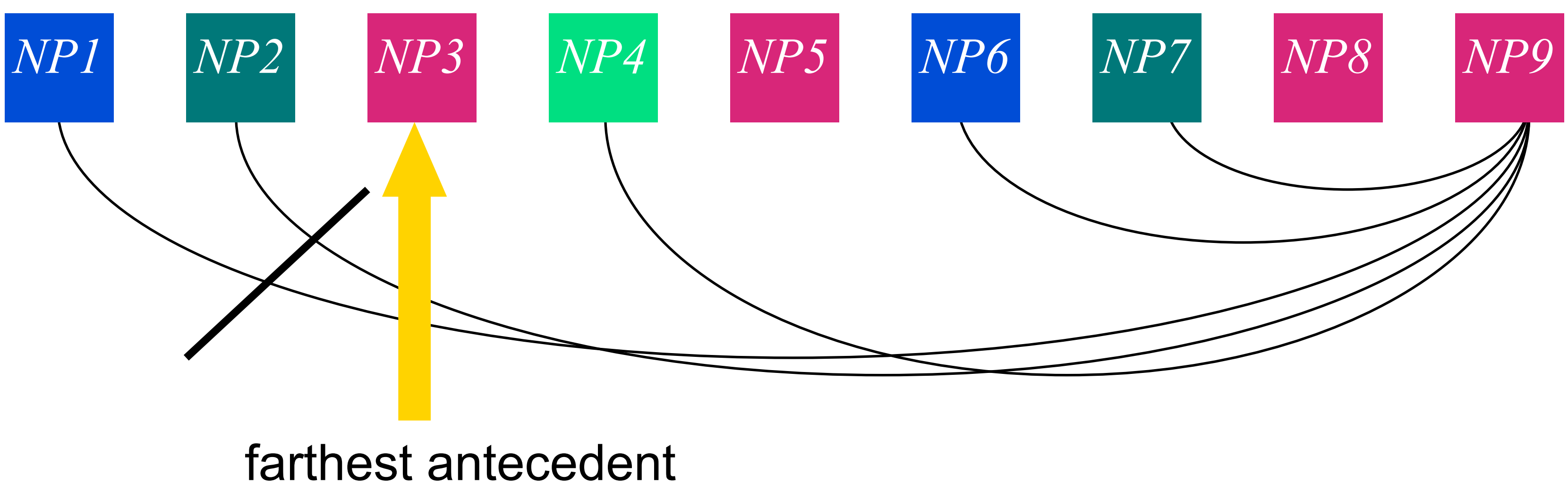
# Classifier for MUC-6 Data Set

```
ALIAS = C: +
ALIAS = I:
| SOON_STR_NONPRO = C:
| | ANIMACY = NA: -
| | ANIMACY = I: -
| | ANIMACY = C: +
| SOON_STR_NONPRO = I:
| | PRO_STR = C: +
| | PRO_STR = I:
| | | PRO_RESOLVE = C:
| | | | EMBEDDED_1 = Y: -
| | | | EMBEDDED_1 = N:
| | | | | PRONOUN_1 = Y:
| | | | | | ANIMACY = NA: -
| | | | | | ANIMACY = I: -
| | | | | | ANIMACY = C: +
| | | | | PRONOUN_1 = N:
| | | | | | MAXIMALNP = C: +
| | | | | | MAXIMALNP = I:
| | | | | | | WNCLASS = NA: -
| | | | | | | WNCLASS = I: +
| | | | | | | WNCLASS = C: +
| | | PRO_RESOLVE = I:
| | | | APPOSITIVE = I: -
| | | | APPOSITIVE = C:
| | | | | GENDER = NA: +
| | | | | GENDER = I: +
| | | | | GENDER = C: -
```

CORNELL

# Problem 1

- Coreference is a rare relation
  - skewed class distributions (2% positive instances)
  - *remove some negative instances*



# Problem 2

- Coreference is a discourse-level problem with different solutions for different types of NPs
    - » proper names: string matching and aliasing
  - – inclusion of “hard” positive training instances
  - – *positive example selection*: selects easy positive training instances (cf. Harabagiu *et al.* (2001))

> Queen Elizabeth set about transforming **her husband**,
> **King George VI**, into a viable monarch. Logue,
> the renowned speech therapist, was summoned to help
> **the King** overcome his speech impediment...

# Problem 3

- Coreference is an equivalence relation
  - loss of transitivity
  - need to tighten the connection between classification and clustering
  - *prune learned rules w.r.t. the clustering-level coreference scoring function*

*coref* ? *coref* ?

[Queen Elizabeth] set about transforming [her] [husband], ...

*not coref* ?

# Results

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 40.7 | 73.5 | 52.4 | 27.2 | 86.3 | 41.3 |
| **NEG-SELECT** | 46.5 | 67.8 | 55.2 | 37.4 | 59.7 | 46.0 |
| **POS-SELECT** | 53.1 | 80.8 | 64.1 | 41.1 | 78.0 | 53.8 |
| **NEG-SELECT + POS-SELECT** | 63.4 | 76.3 | 69.3 | 59.5 | 55.1 | 57.2 |
| **NEG-SELECT + POS-SELECT + RULE-SELECT** | 63.3 | 76.9 | **69.5** | 54.2 | 76.3 | **63.4** |

- Ultimately: large increase in F-measure, due to gains in recall

# Comparison with Best MUC Systems

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **NEG-SELECT + POS-SELECT + RULE-SELECT** | 63.3 | 76.9 | **69.5** | 54.2 | 76.3 | **63.4** |
| **Best MUC System** | 59 | 72 | **65** | 56.1 | 68.8 | **61.8** |

# Supervised ML for NP Coreference

- Good performance compared to other systems, but…**lots** of room for improvement
  - Common nouns < pronouns < proper nouns
  - Tighter connection between classification and clustering is possible
  - Need additional data sets
    - » ACE data from Penn's LDC
    - » General problem: reliance on manually annotated data…

# Outline

- noun phrase coreference resolution
- a (supervised) machine learning approach
- weakly supervised approaches
  - background
  - two techniques
  - evaluation

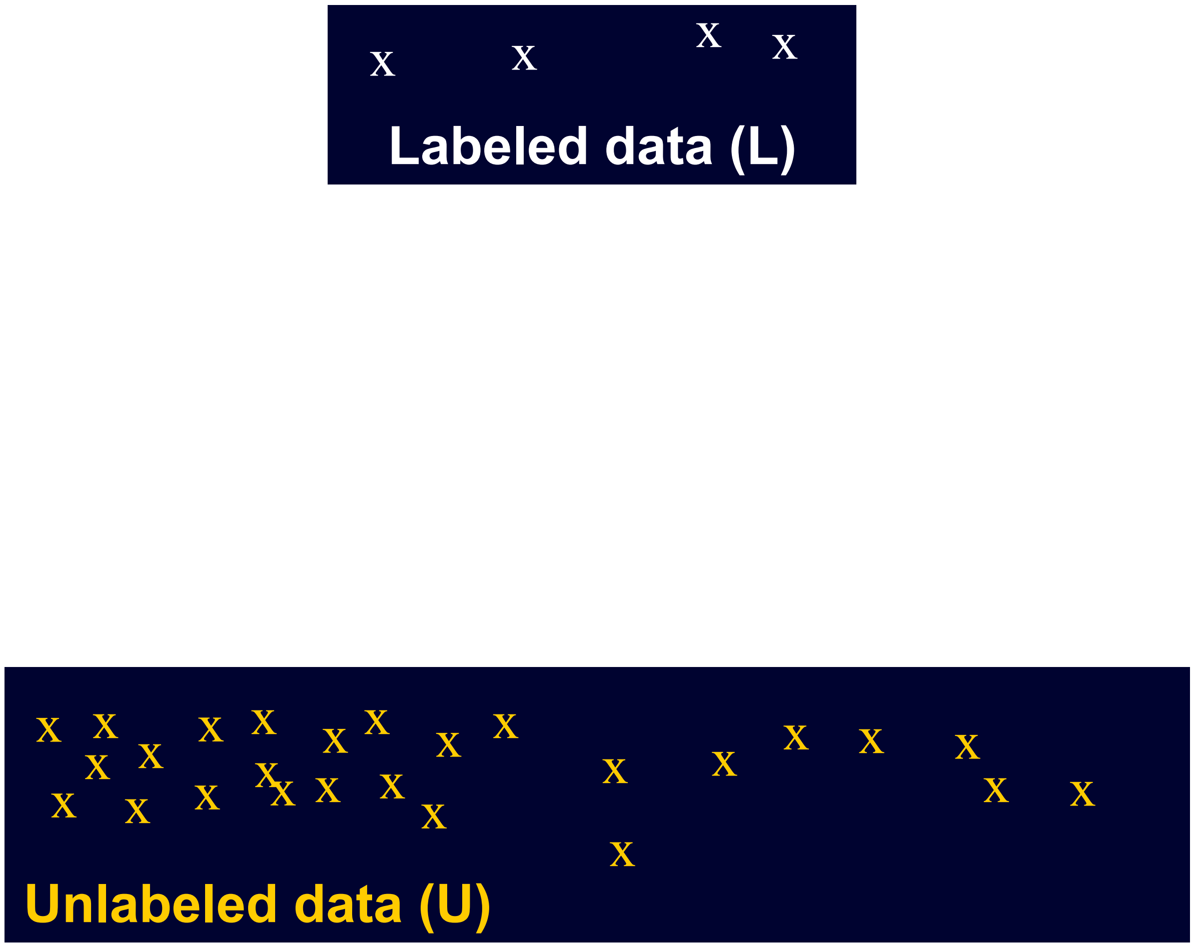
# Weakly Supervised Approaches

- Idea:

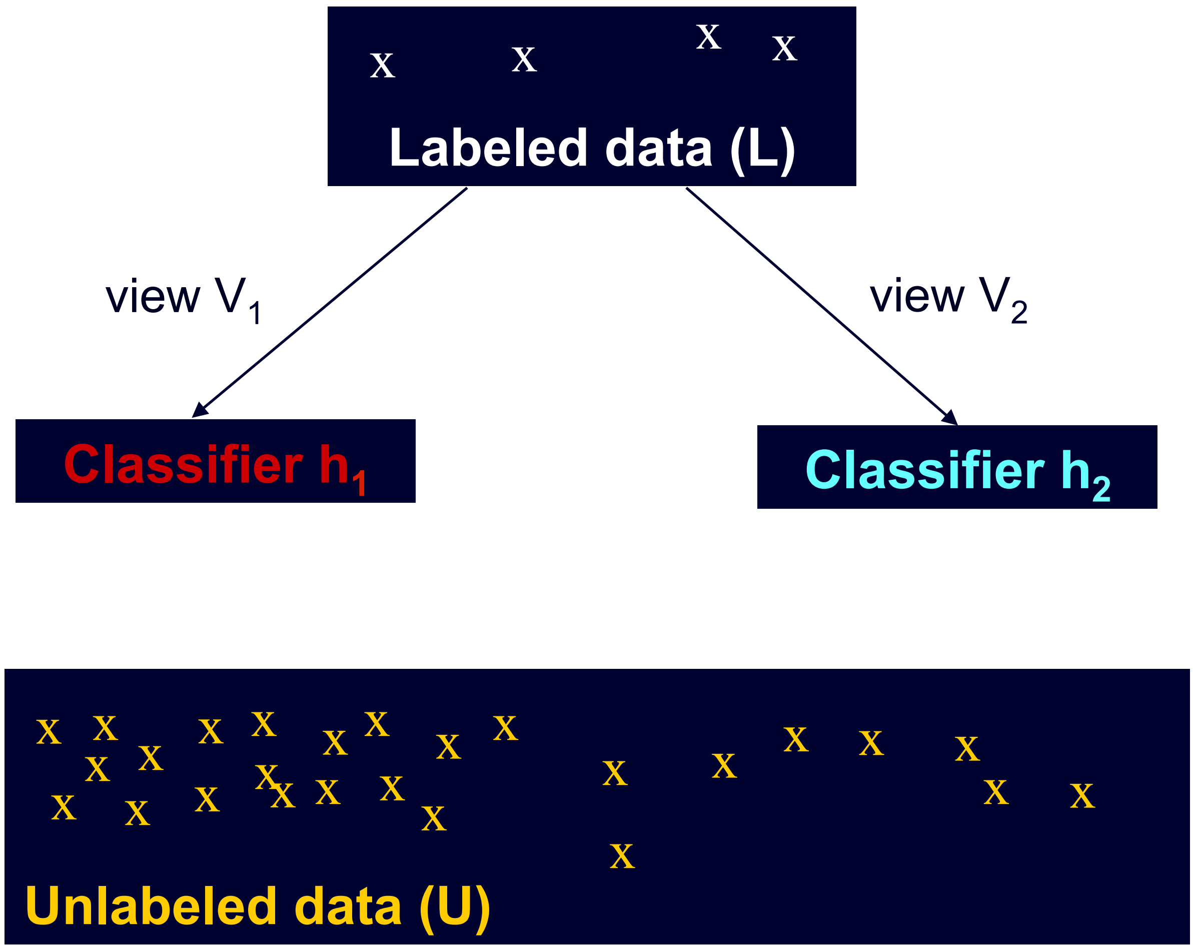
  bootstrap (NP coreference) classifiers using a *small amount of labeled data* (expensive) and a *large amount of unlabeled data* (cheap)

- Methods
  - Co-training
  - Self-training

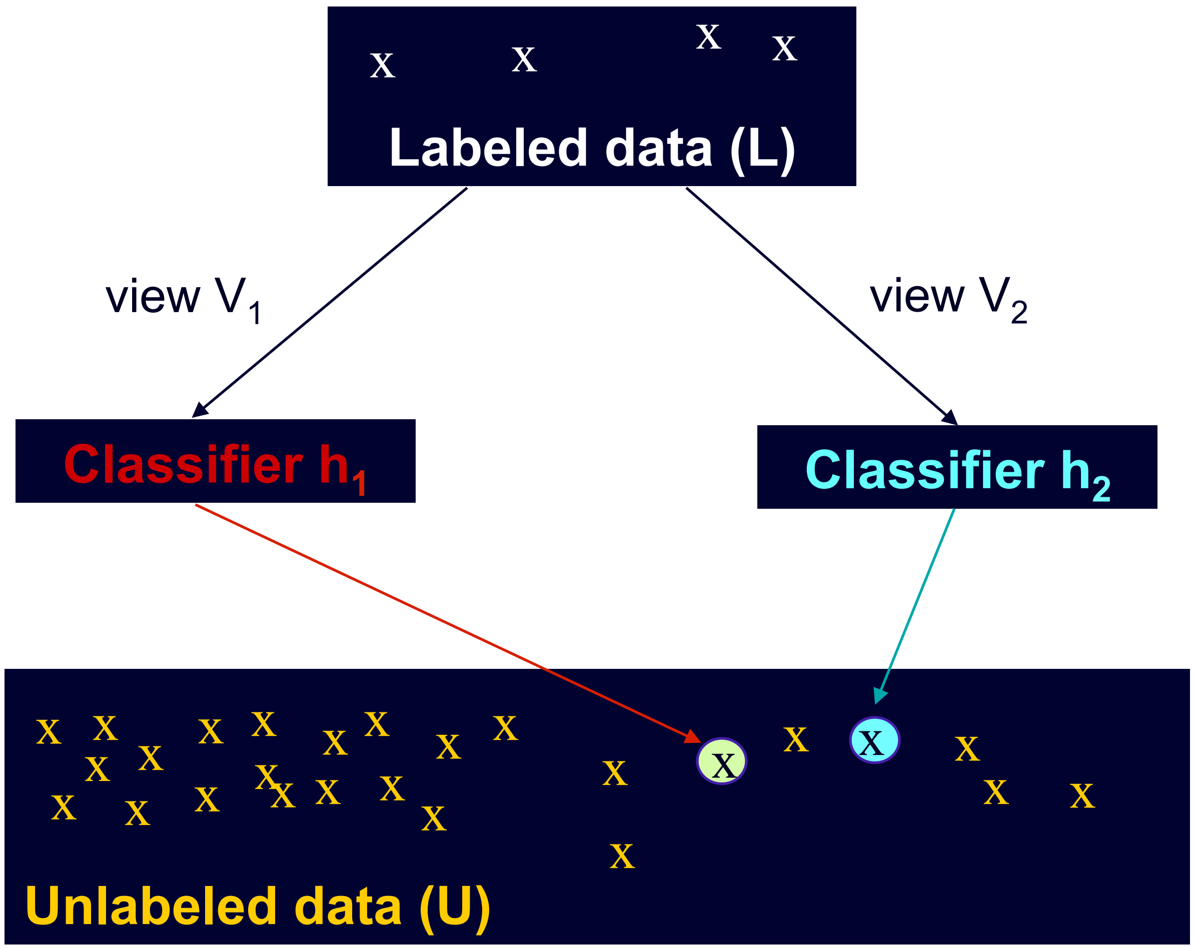
# Co-Training [Blum and Mitchell, 1998]



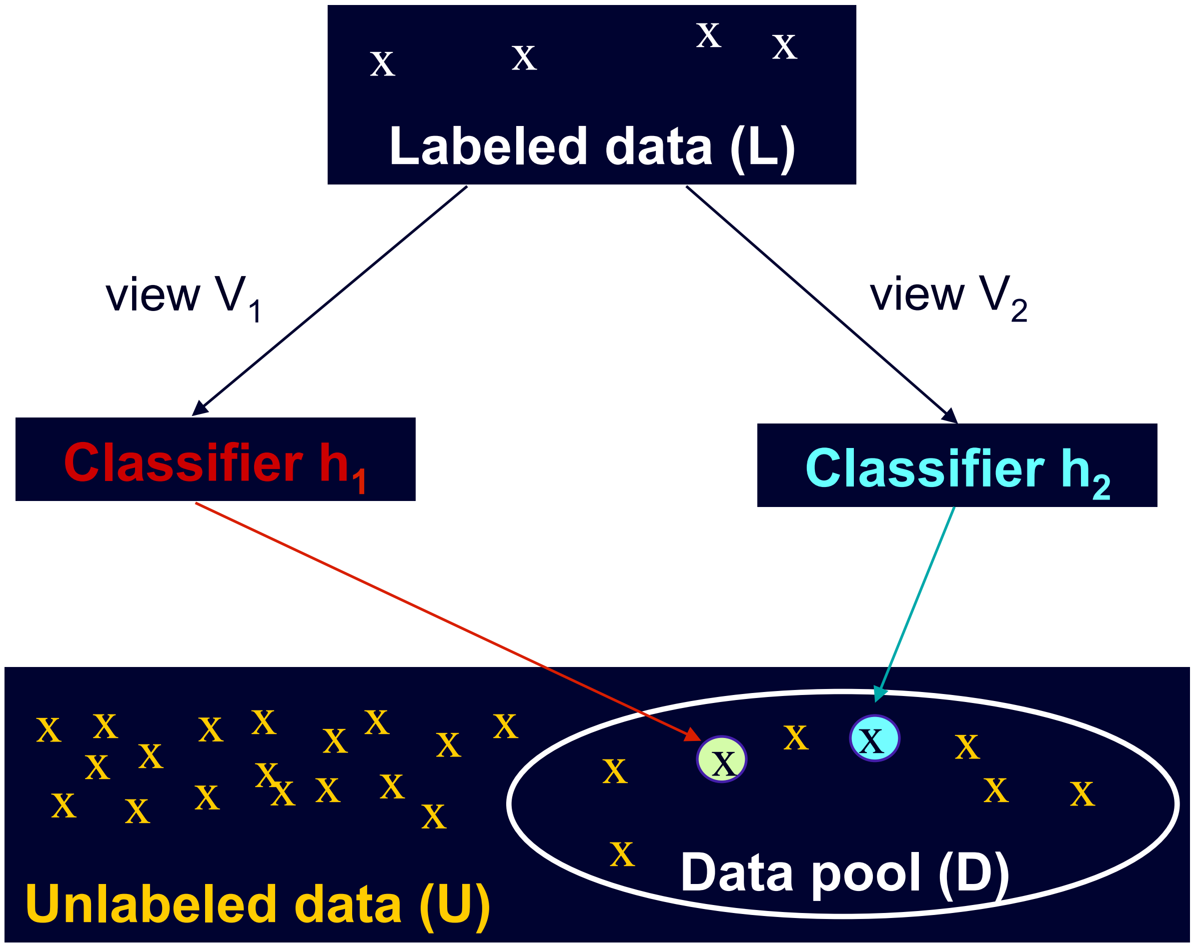
# Co-Training [Blum and Mitchell, 1998]



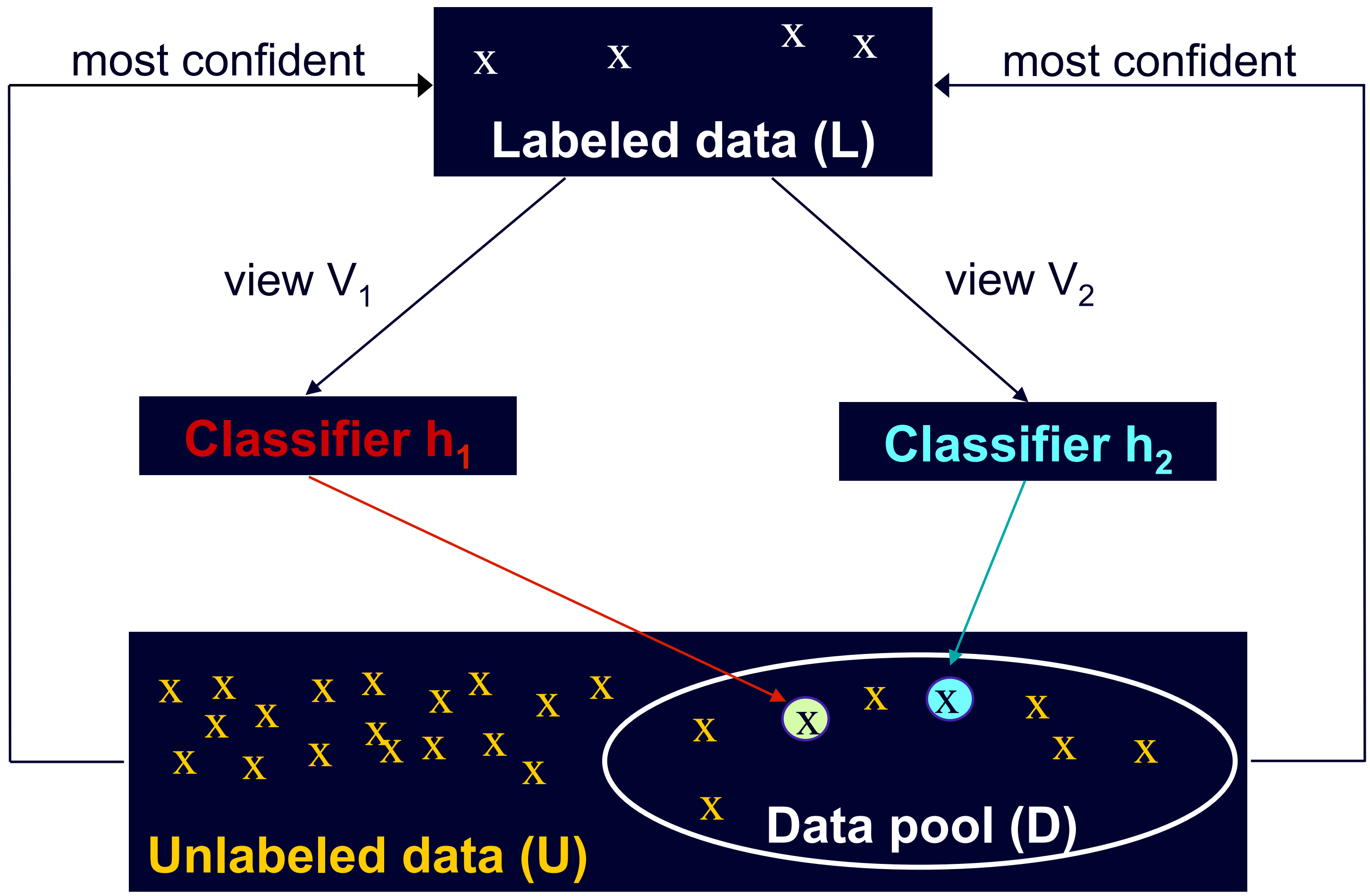
# Co-Training [Blum and Mitchell, 1998]



# Co-Training [Blum and Mitchell, 1998]



# Co-Training [Blum and Mitchell, 1998]



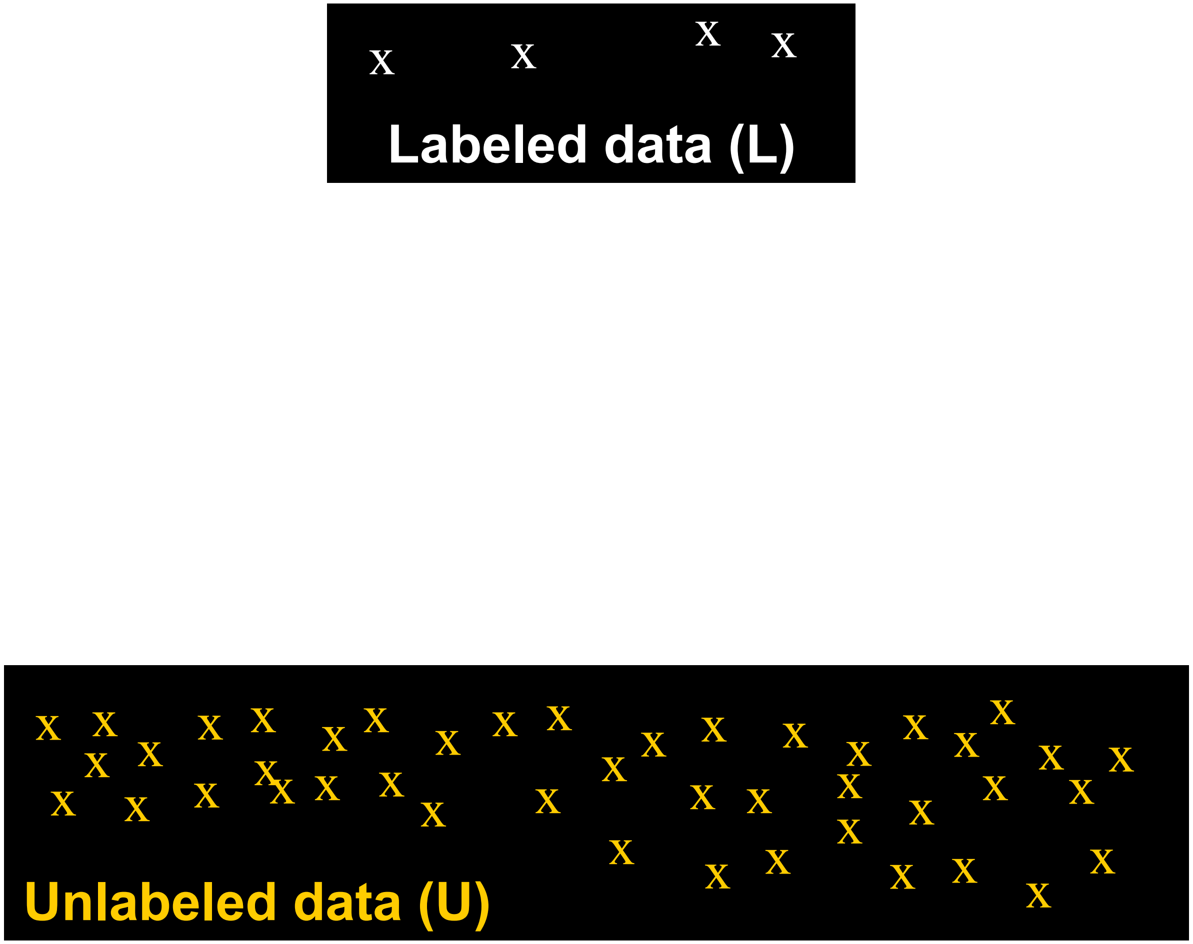
# Potential Problems with Co-Training

- Strong assumptions on the views (Blum and Mitchell, 1998)
  - each view must be sufficient for learning the target concept
  - the views must be conditionally independent given the class
  - empirically shown to be sensitive to these assumptions (Muslea *et al*., 2002)
- A number of parameters need to be tuned
  - views, data pool size, growth size, number of iterations, initial size of labeled data
  - algorithm is sensitive to its input parameters (Nigam and Ghani, 2000; Pierce and Cardie, 2001; Pierce 2003)

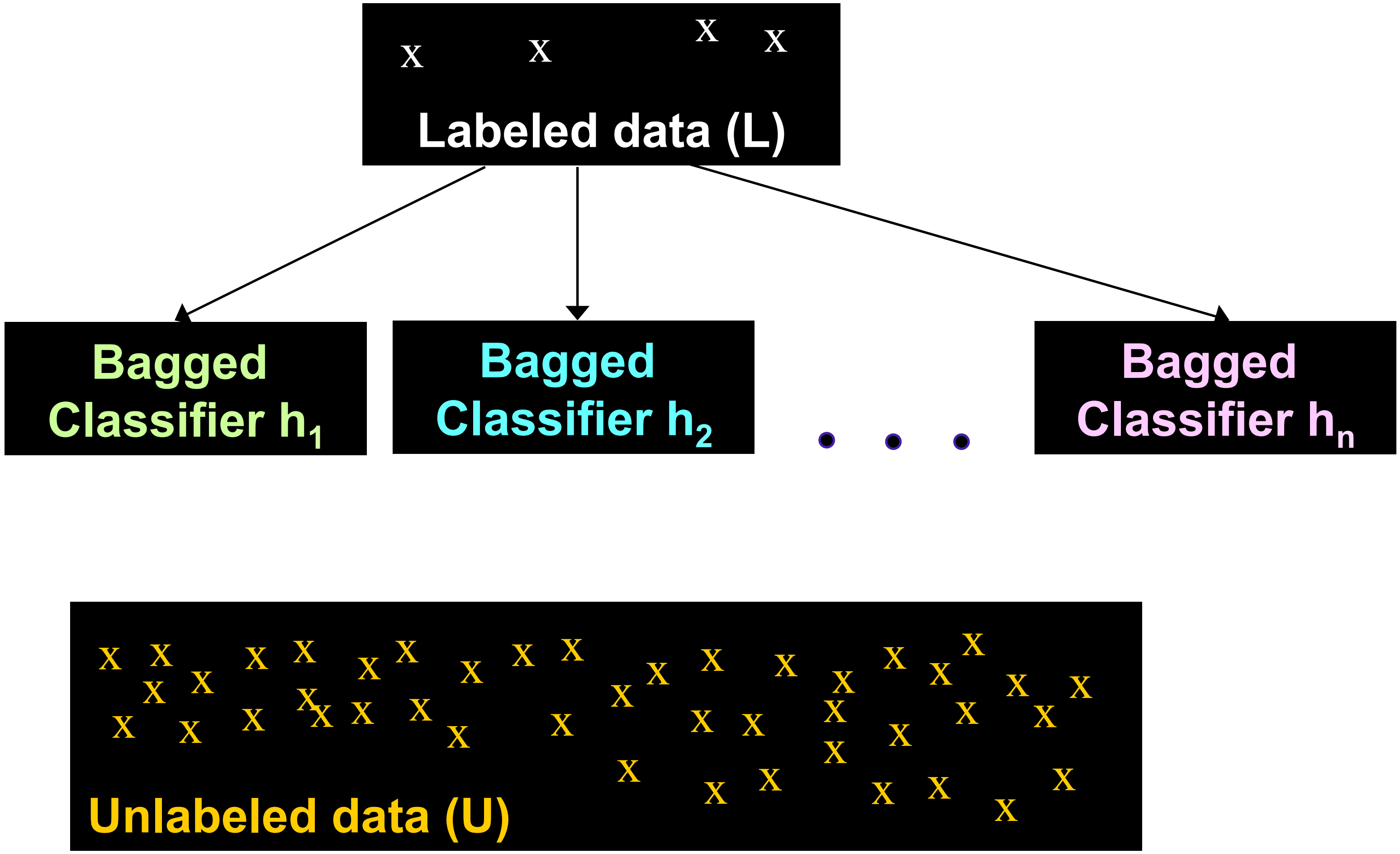
# Potential Problems with Co-Training

- Multi-view algorithm
  - Is there any natural feature split for NP coreference?
    - » view factorization is a non-trivial problem for coreference
      - ◆ Mueller *et al.*'s (2002) greedy method

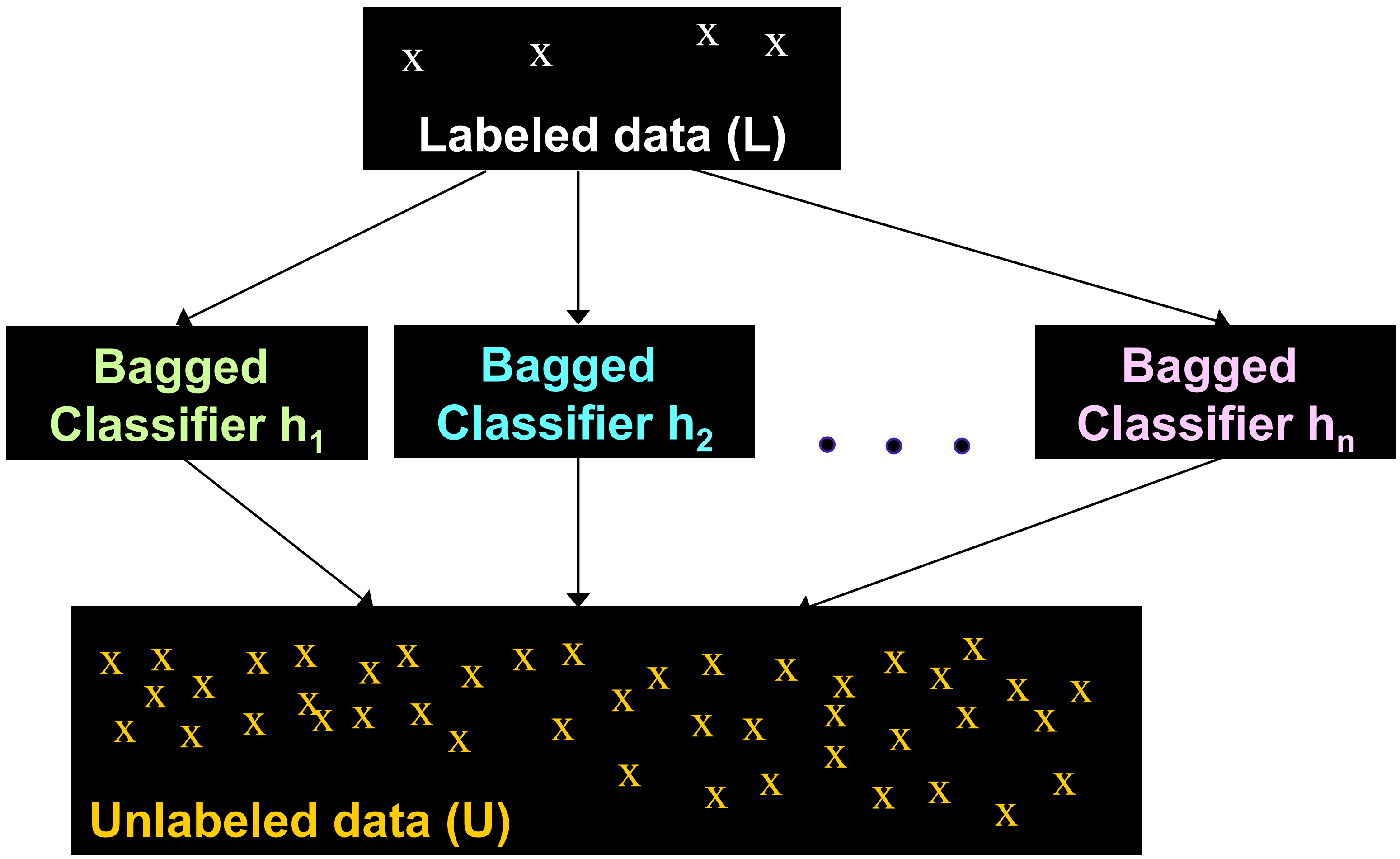
# Self-Training with Bagging
## [Banko and Brill, 2001]



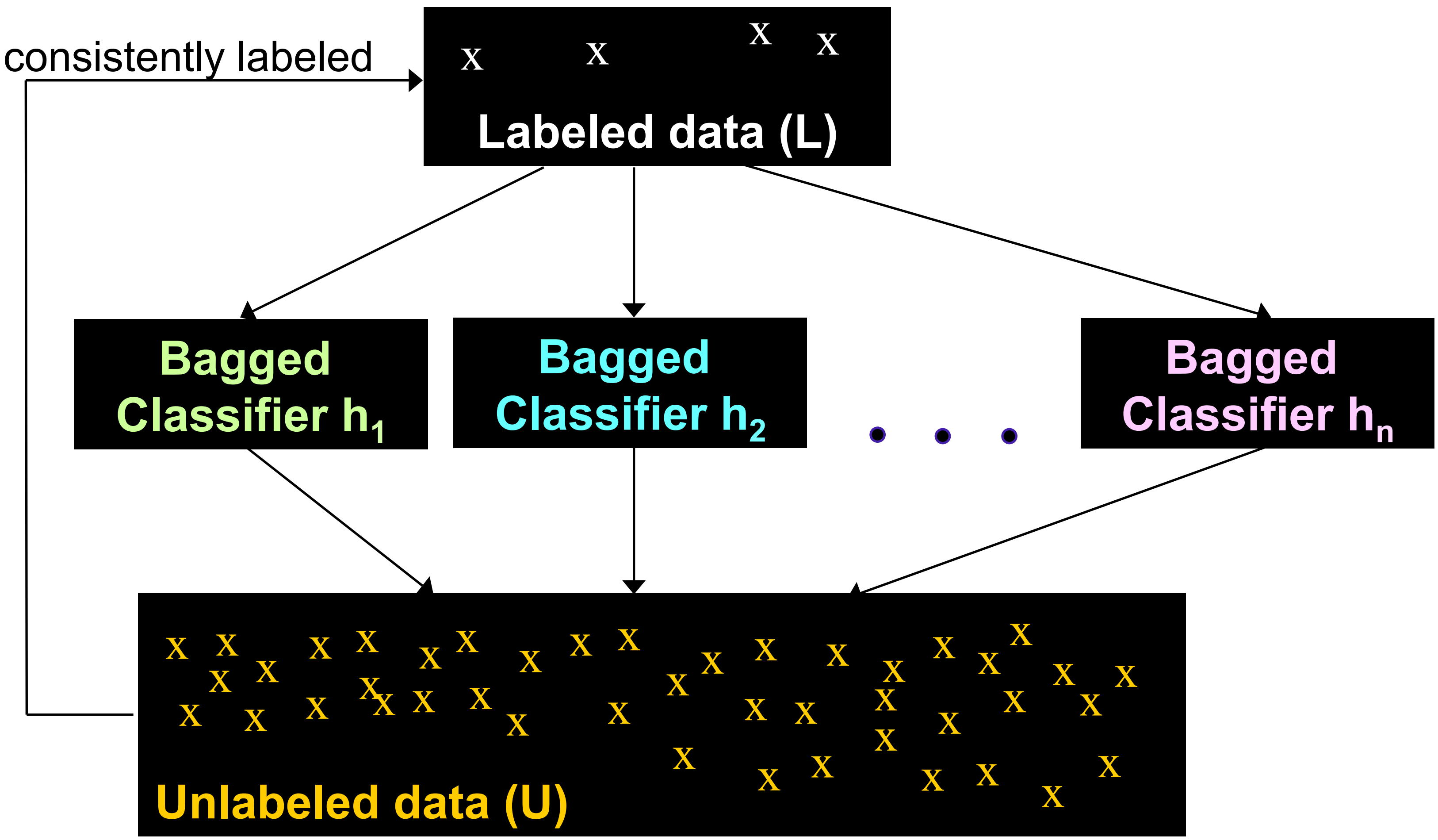
# Self-Training with Bagging [Banko and Brill, 2001]



# Self-Training with Bagging [Banko and Brill, 2001]



# Self-Training with Bagging
## [Banko and Brill, 2001]



# Plan for the Talk

- noun phrase coreference resolution
- a (supervised) machine learning approach
- weakly supervised approaches
  - background
  - two techniques
  - evaluation

# Evaluation

- MUC-6 and MUC-7 coreference data sets
- labeled data (L): one dryrun text
    - » 3500-3700 instances
- unlabeled data (U): remaining 29 dryrun texts
- vs. fully supervised ML
    - – ~500,000 instances (30 dryrun texts)

# Results (Baseline)

- train a naïve Bayes classifier on the single (labeled) text using all 25 features

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 58.3 | 52.9 | **55.5** | 52.8 | 37.4 | **43.8** |

# Evaluating the Weakly Supervised Algorithms

- Determine the best parameter setting of each algorithm (in terms of its effectiveness in improving performance)

# Co-Training Parameters

- Views (3 heuristic methods for view factorization)
  - Mueller *et al.*'s (2002) greedy method
  - random splitting
  - splitting according to the feature type
- Pool size
  - 500, 1000, 5000
- Growth size
  - 10, 50, 100, 200, 250
- Number of co-training iterations
  - run until performance stabilized

# Results (Co-Training)

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 58.3 | 52.9 | **55.5** | 52.8 | 37.4 | **43.8** |
| **Co-Training** | 47.5 | 81.9 | **60.1** | 40.6 | 77.6 | **53.3** |

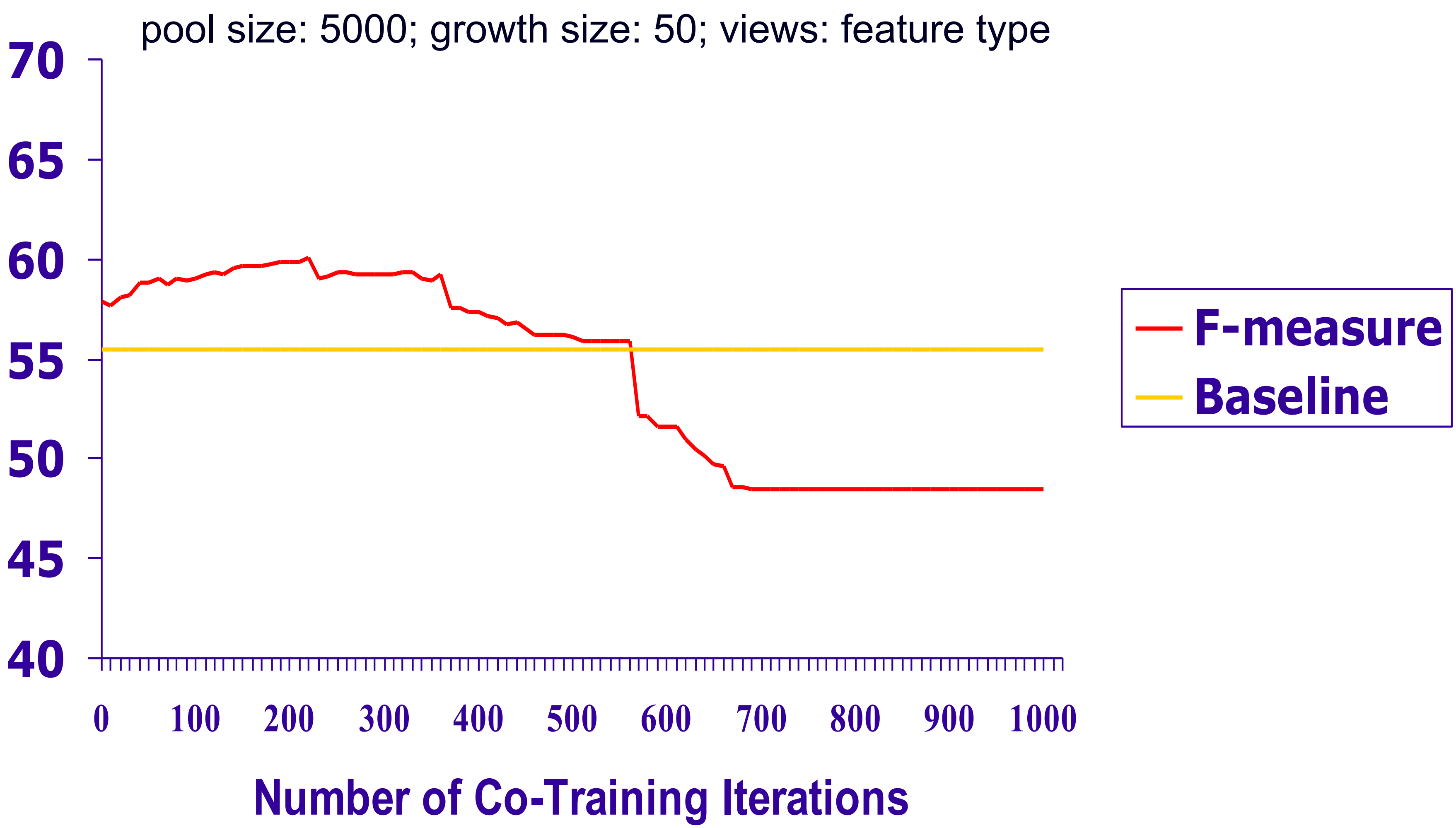
- co-training produces improvements over the baseline at its best parameter settings

# Results (Co-Training)

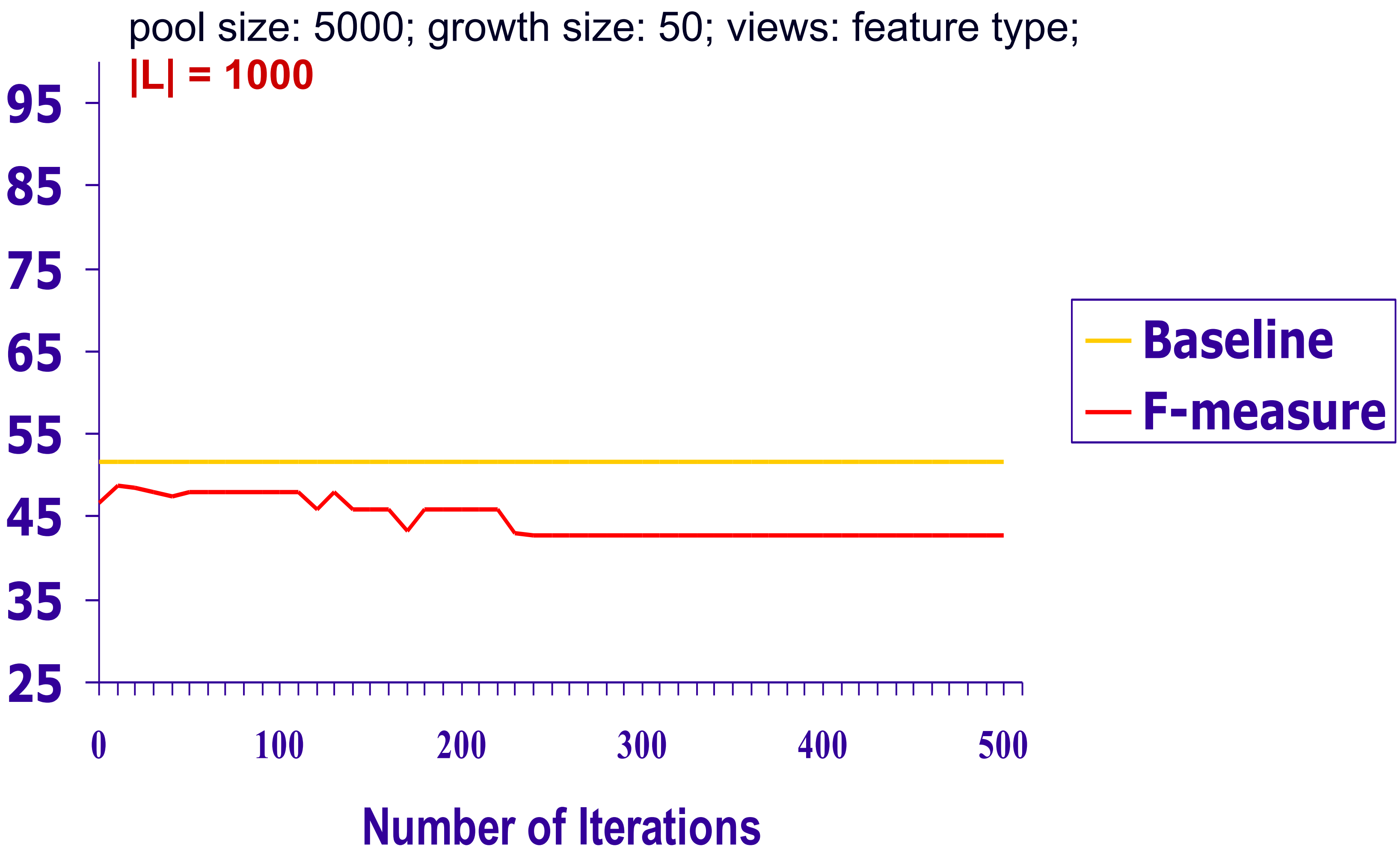
| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 58.3 | 52.9 | **55.5** | 52.8 | 37.4 | **43.8** |
| **Co-Training** | 47.5 | 81.9 | **60.1** | 40.6 | 77.6 | **53.3** |
| **Supervised ML*** (~500,000 insts) | 63.3 | 76.9 | 69.5 | 54.2 | 76.3 | 63.4 |

- co-training produces improvements over the baseline at its best parameter settings

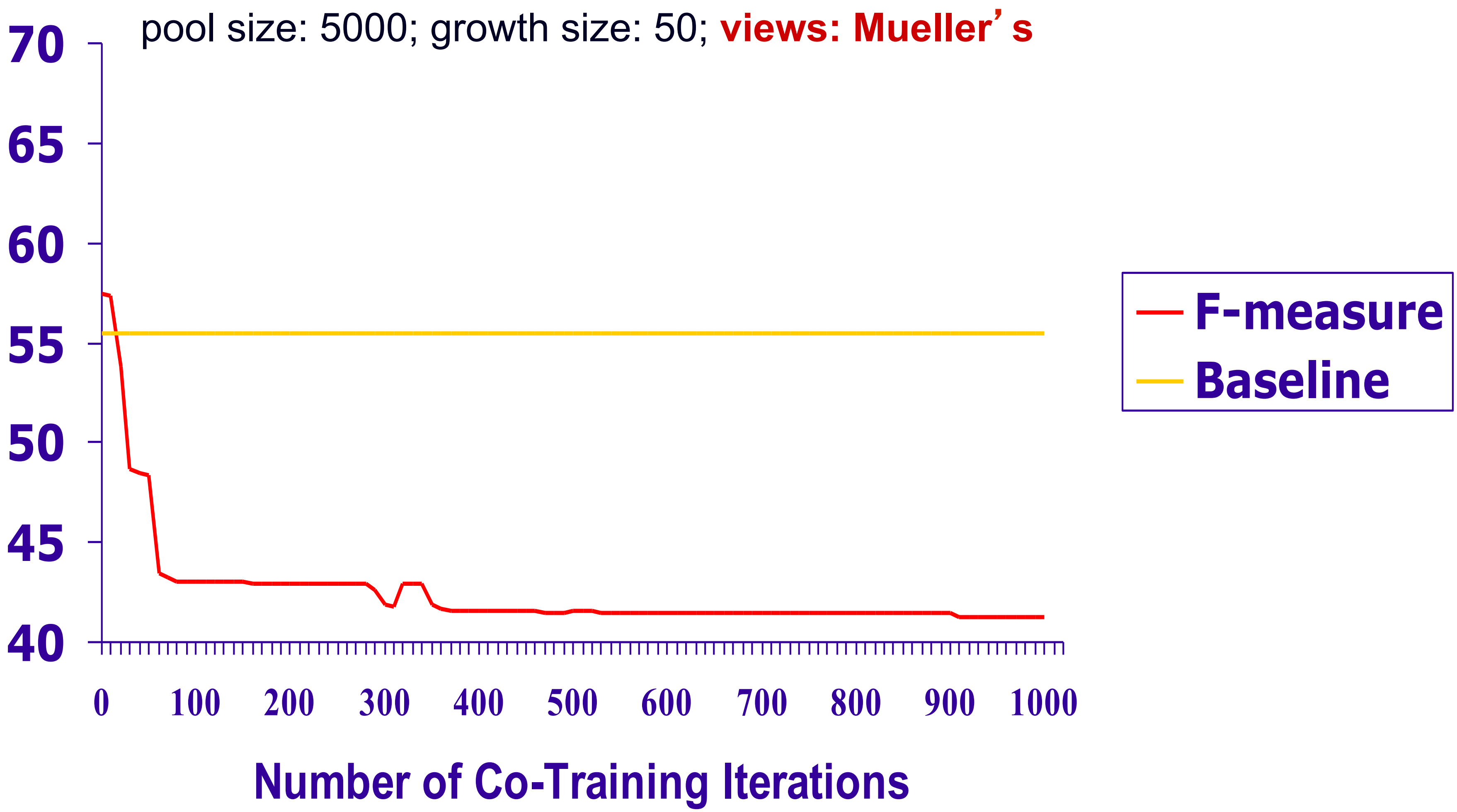
# Learning Curve for Co-Training (MUC-6)



# Learning Curve for Co-Training (MUC-6)



# Learning Curve for Co-Training (MUC-6)



# Self-Training Parameters

- Number of bags
  - tested all odd number of bags between 1 and 25

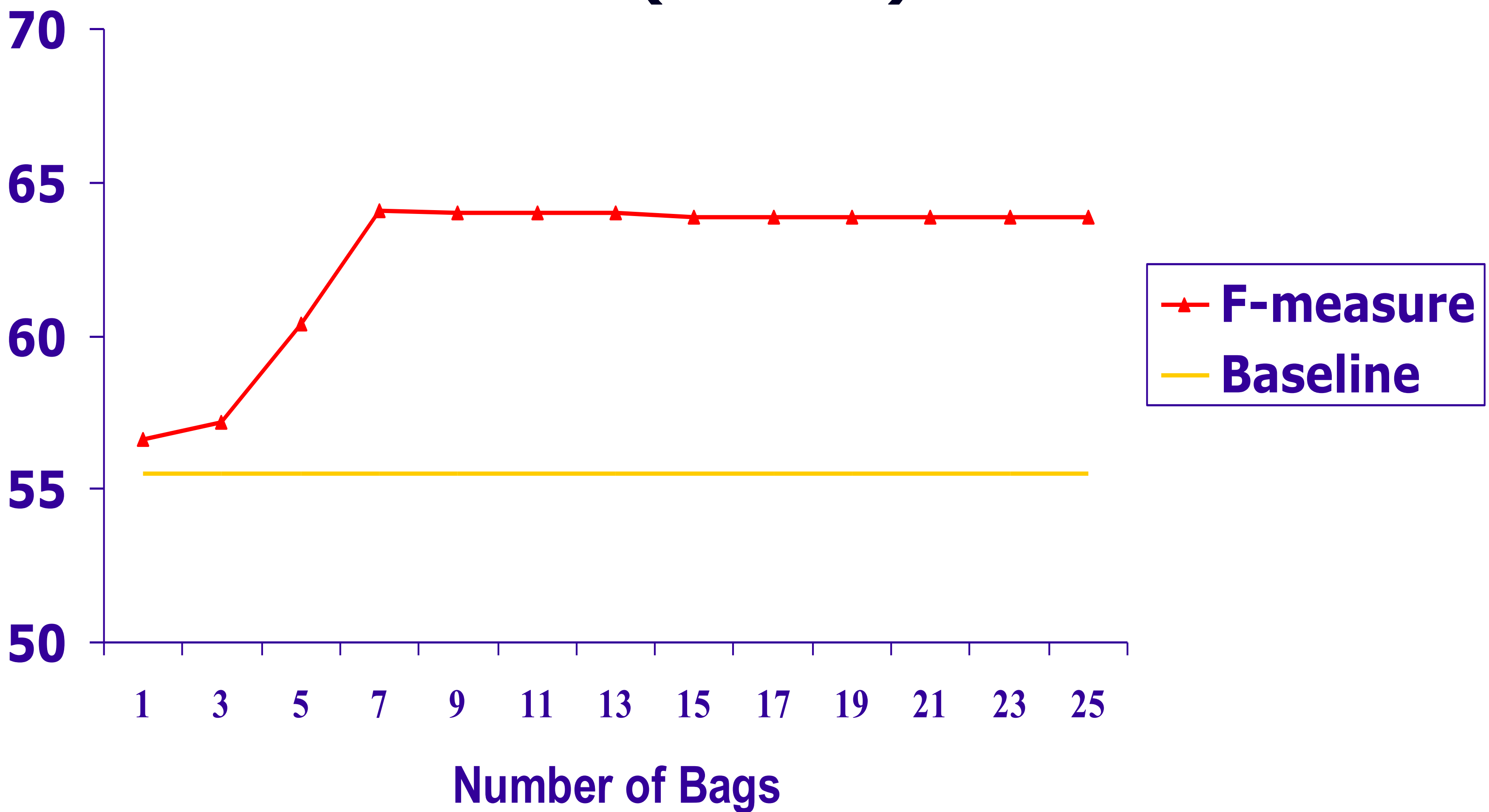
- 25 bags are sufficient for most learning tasks (Breiman, 1996)

# Results (Self-Training with Bagging)

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 58.3 | 52.9 | **55.5** | 52.8 | 37.4 | **43.8** |
| **Co-Training** | 47.5 | 81.9 | **60.1** | 40.6 | 77.6 | **53.3** |
| **Self-Training with Bagging** | 54.1 | 78.6 | **64.1** | 54.6 | 62.6 | **58.3** |

- Self-training performs better than co-training

# Self-Training: Effect of the Number of Bags (MUC-6)



# Results

| | MUC-6 | | | MUC-7 | | |
|---|---|---|---|---|---|---|
| | R | P | F | R | P | F |
| **Baseline** | 58.3 | 52.9 | **55.5** | 52.8 | 37.4 | **43.8** |
| **Co-Training** | 47.5 | 81.9 | **60.1** | 40.6 | 77.6 | **53.3** |
| **Self-Training with Bagging** | 54.1 | 78.6 | **64.1** | 54.6 | 62.6 | **58.3** |
| **Supervised ML*** (~500,000 insts) | 63.3 | 76.9 | 69.5 | 54.2 | 76.3 | 63.4 |

# Summary

- Supervised ML approach to NP coreference resolution
  - Good performance relative to other approaches
  - Still lots of room for improvement
- Weakly supervised approaches are promising
  - Not as good performance as fully supervised, but use much less manually annotated training data
- For problems where no natural view factorization exists…
  - Single-view weakly supervised algorithms
    - » Self-training with bagging

# …and also

1. Illustrate how much you've learned
2. Realities of doing work in NLP+ML
3. Introduce some cool weakly supervised learning methods