Weakly Supervised Natural Language Learning Without Redundant Views

Vincent Ng and Claire Cardie
Department of Computer Science
Cornell University

Weakly Supervised Learning

- Supervised approaches
  - require a lot of annotated data that could be expensive or even impractical to obtain

- Weakly supervised approaches
  - address the need for cost-effective annotation methods
  - idea: bootstrap from a small set of labeled data

Multi-View Weakly Supervised Learning

- Multi-view weakly supervised learning algorithms
  - bootstrap from a small set of labeled data using separate, but redundant *views* (i.e. disjoint feature subsets) of the data
  - e.g. co-training (Blum and Mitchell, 1998)
    co-EM (Nigam and Ghani, 2000)
- 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)
- Conditional independence assumption can be relaxed (Nigam and Ghani, 2000; Abney, 2002)

Multi-View Weakly Supervised Learning

- Finding a pair of views that largely satisfies both conditions is non-trivial
  - in practice, users determine a natural feature split into views that are expected to satisfy the two conditions
  - precludes the use of multi-view weakly supervised algorithms on problems without *a natural feature split*

- Hypothesis: single-view weakly supervised algorithms can potentially better than their multi-view counterparts on these problems

Goals of the Study

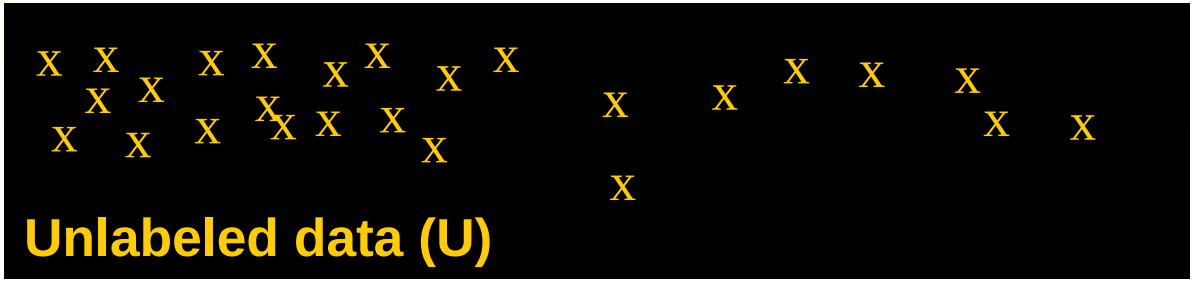
- Take one problem without a natural feature split and apply it to a multi-view weakly supervised learner and two single-view weakly supervised learners
  - Multi-view
    - co-training (Blum and Mitchell, 1998)
  - Single-view
    - self-training with bagging (Banko and Brill, 2001)
    - weakly supervised EM (Nigam *et al*., 2000)

Outline of the Talk

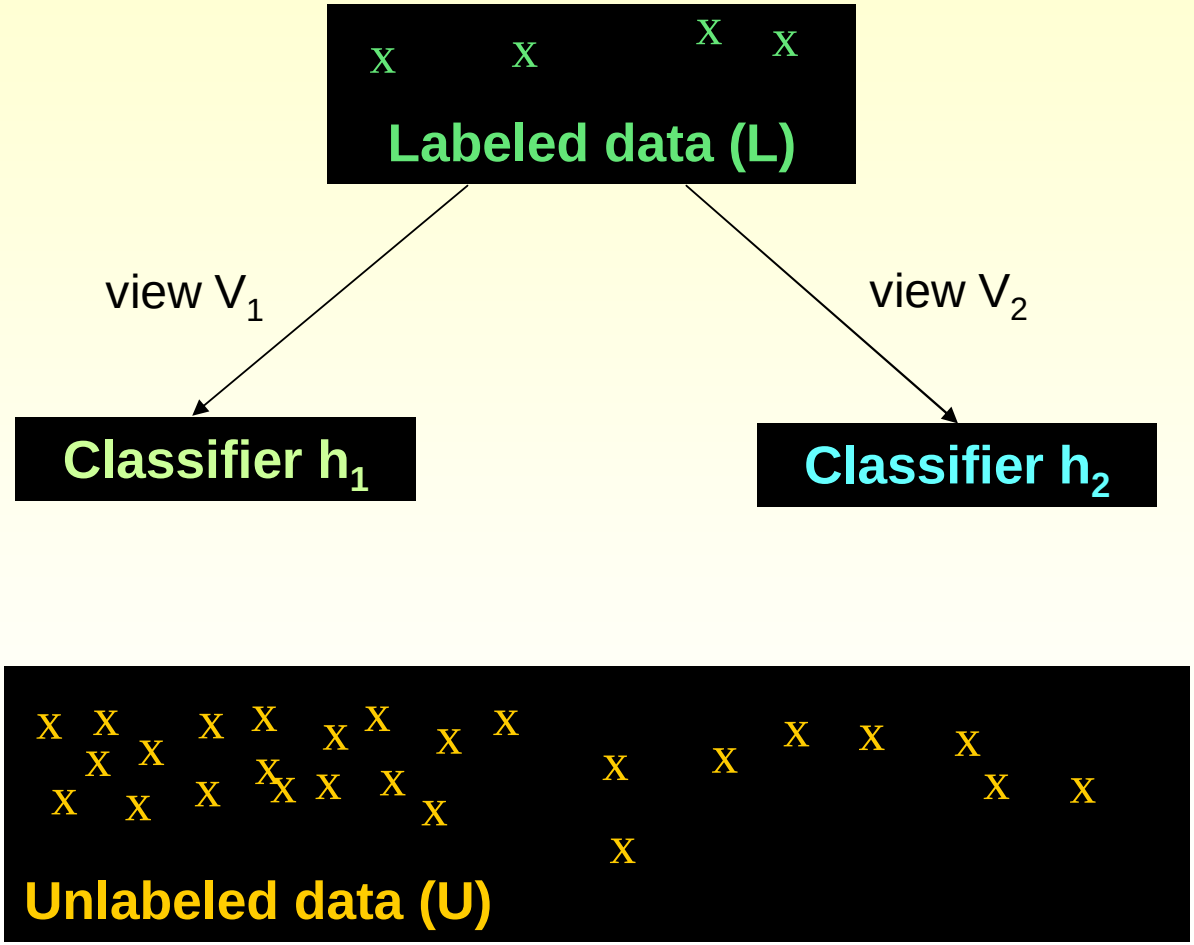
- Weakly supervised learning algorithms
  - co-training
  - self-training with bagging
  - weakly supervised EM
- A learning task without a natural feature split
- Evaluation
- An EM-based bootstrapping algorithm

Co-Training [Blum and Mitchell, 1998]

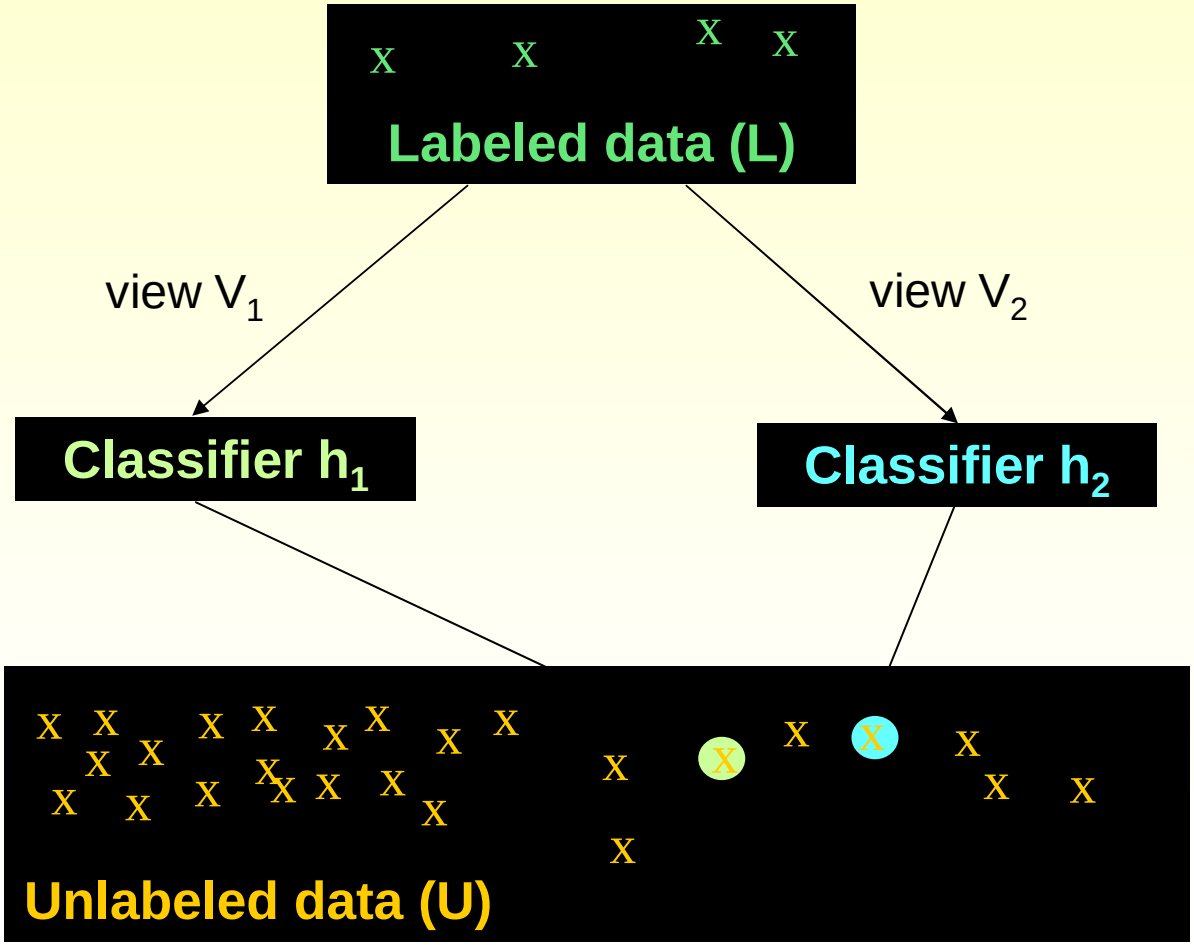
Co-Training [Blum and Mitchell, 1998]



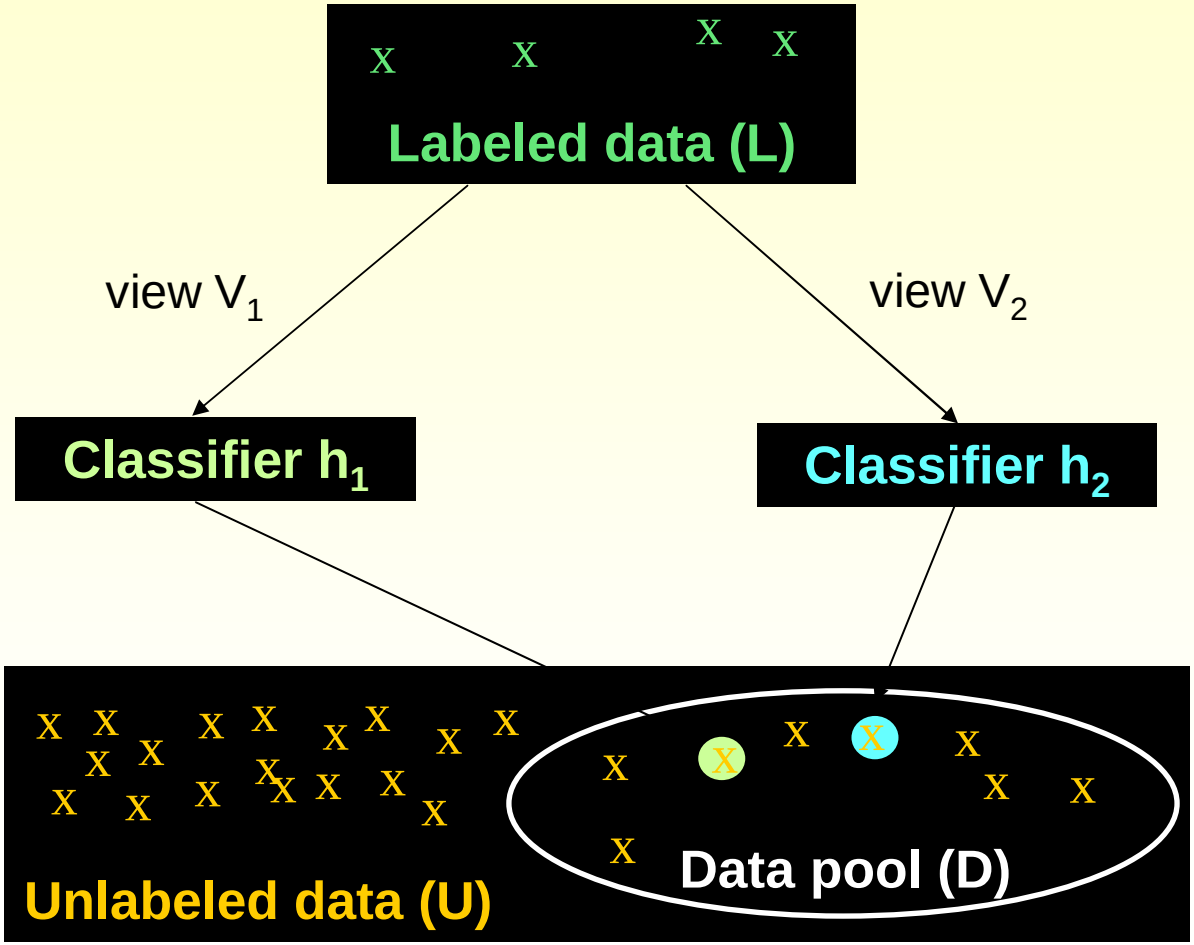
Co-Training [Blum and Mitchell, 1998]



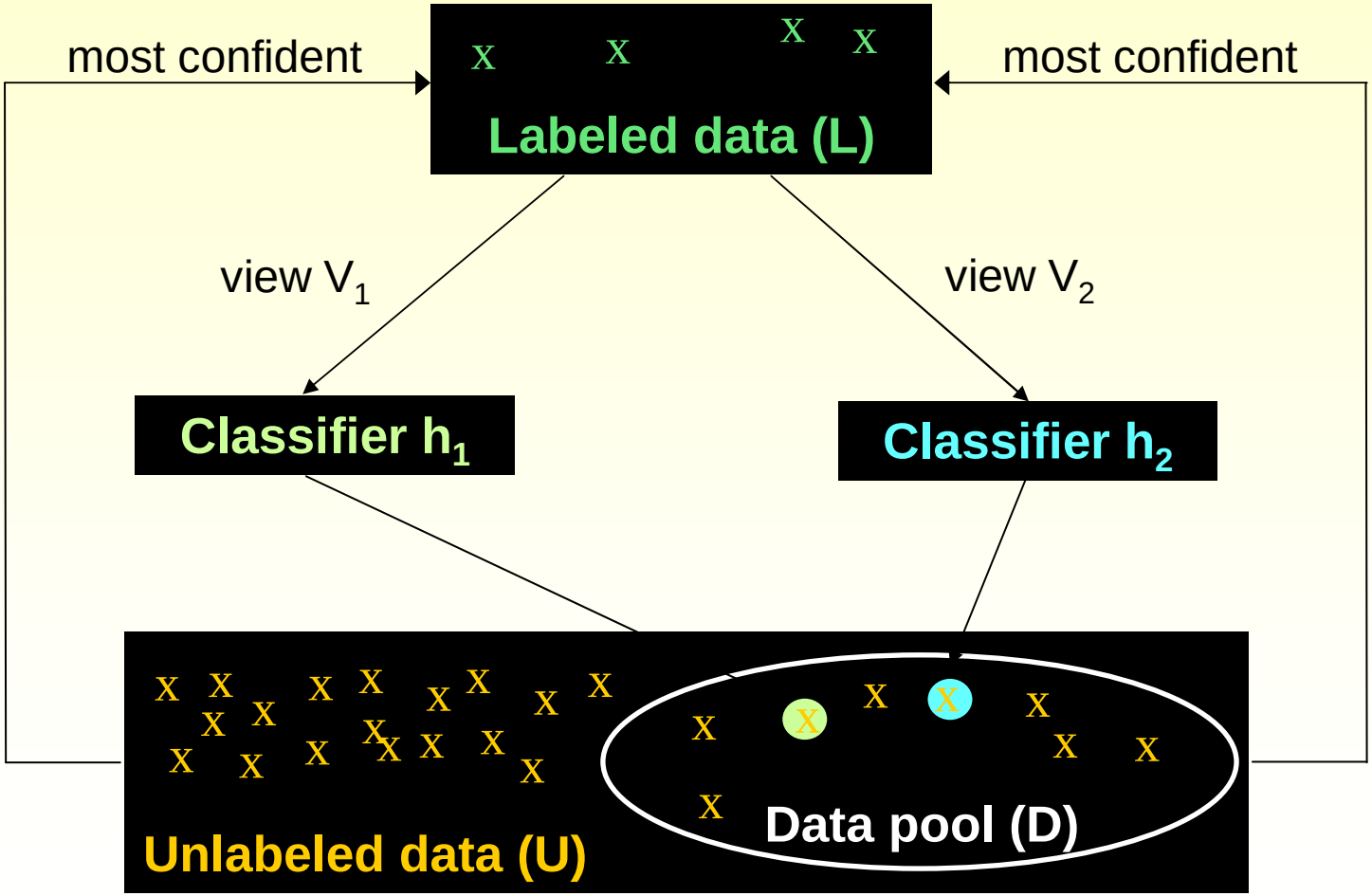
Co-Training [Blum and Mitchell, 1998]



Co-Training [Blum and Mitchell, 1998]



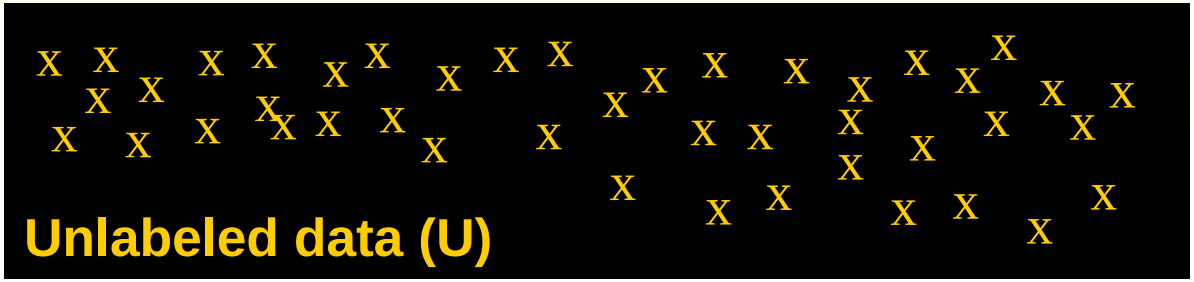
Co-Training [Blum and Mitchell, 1998]



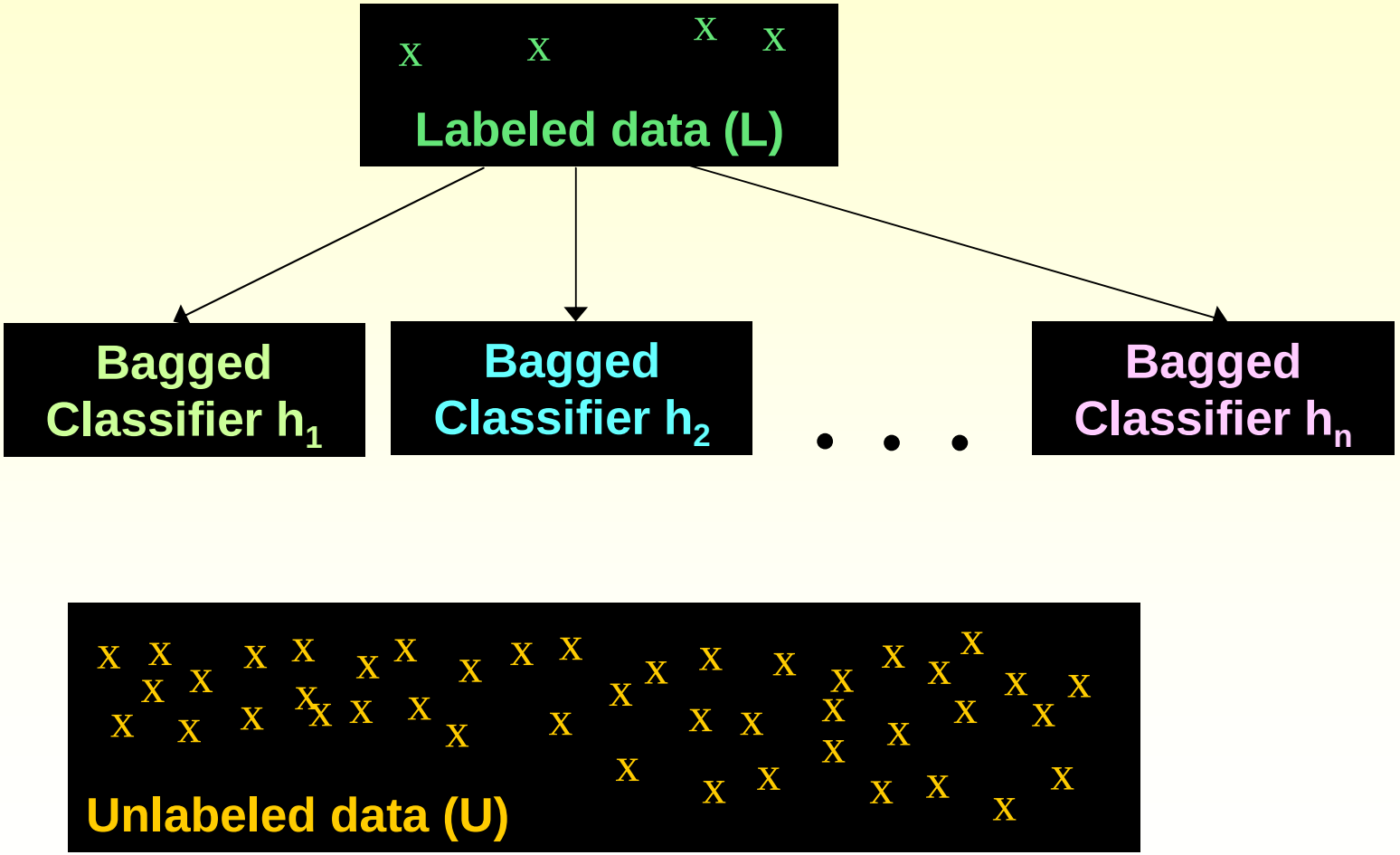
Co-Training [Blum and Mitchell, 1998]

- Multi-view algorithm
- A number of parameters need to be tuned
  - views, data pool size, growth size, number of iterations
- The algorithm is sensitive to its input parameters (Nigam and Ghani, 2000; Pierce and Cardie, 2001)

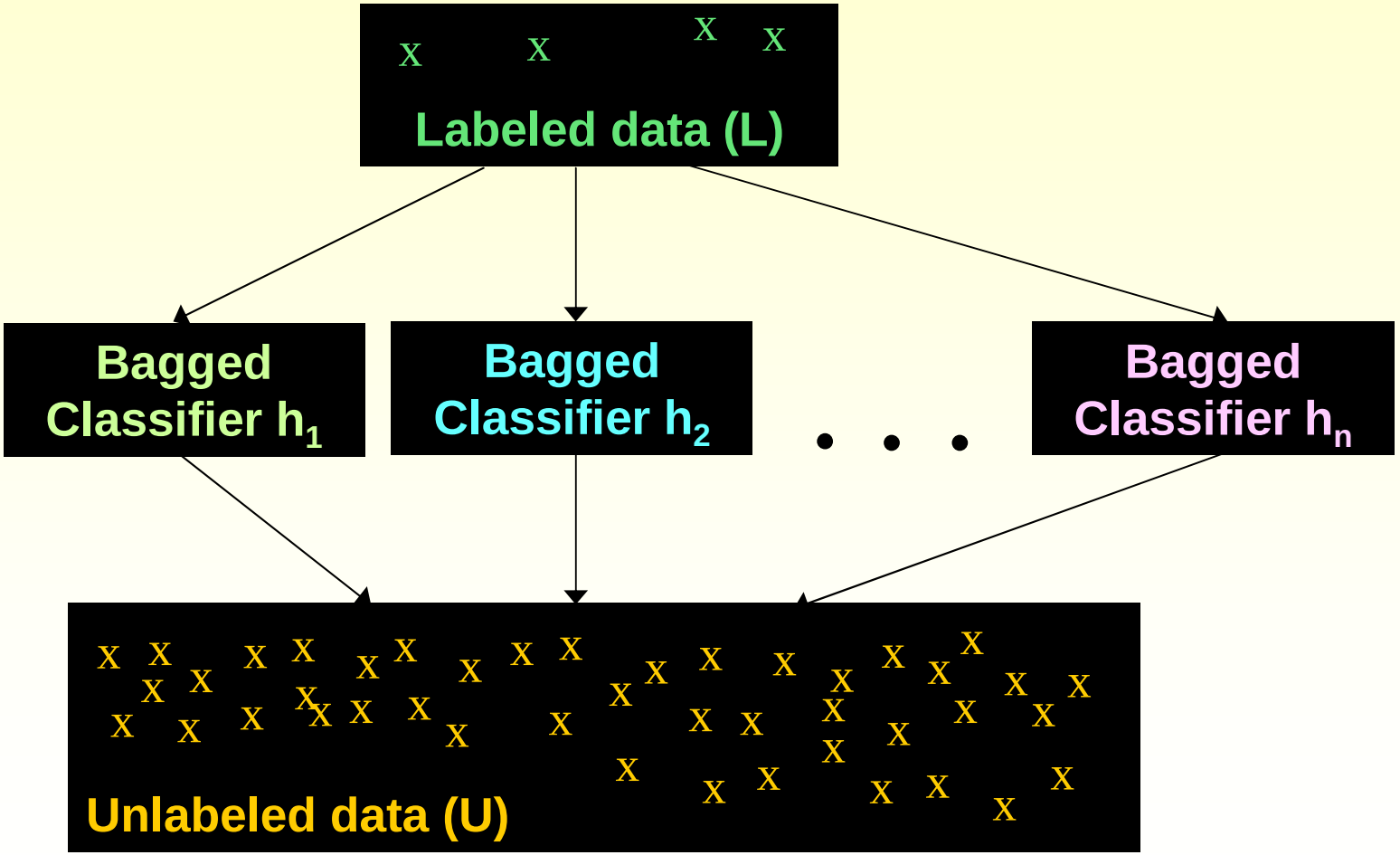
Self-Training with Bagging [Banko and Brill, 2001]



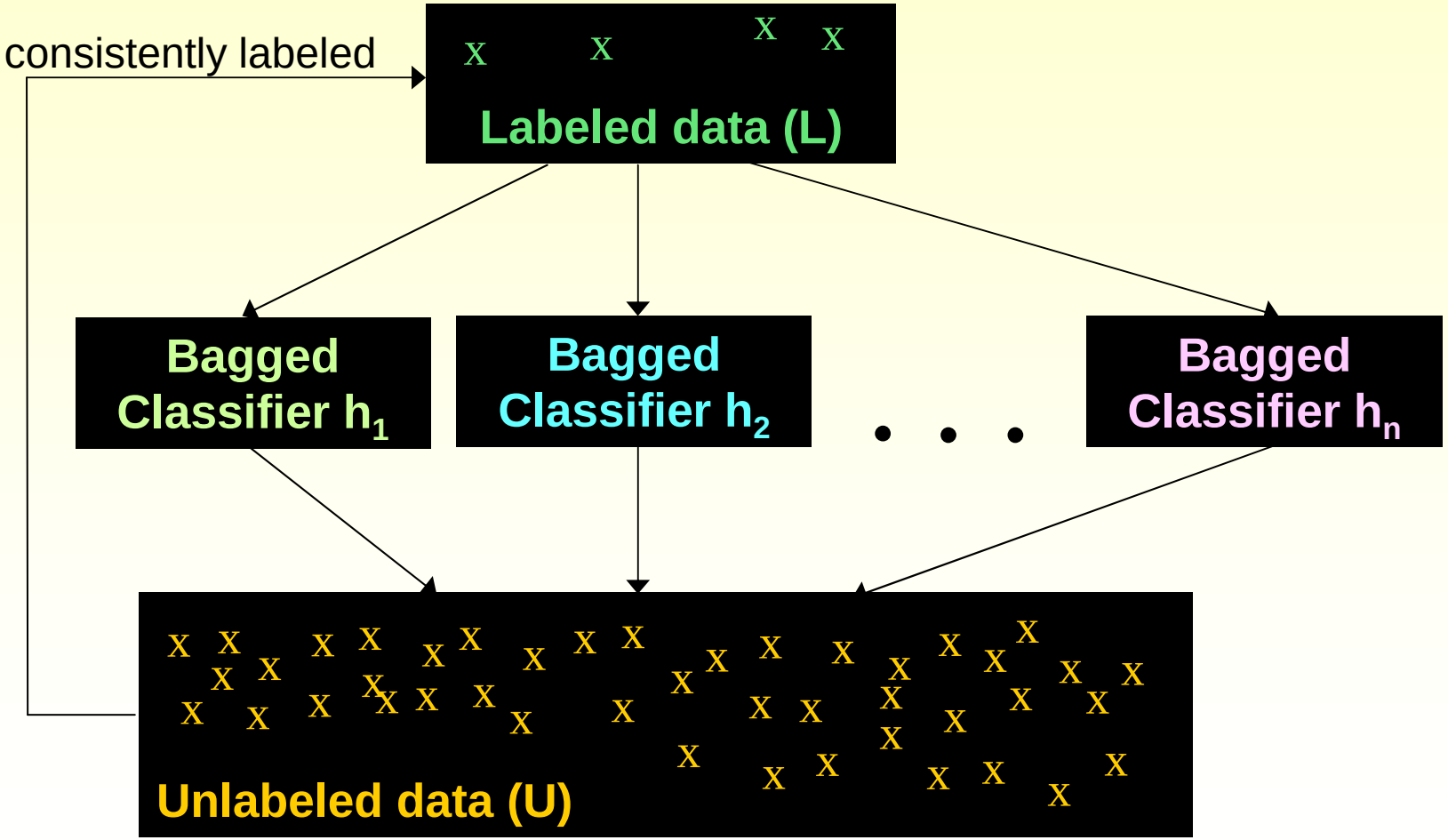
Self-Training with Bagging [Banko and Brill, 2001]



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



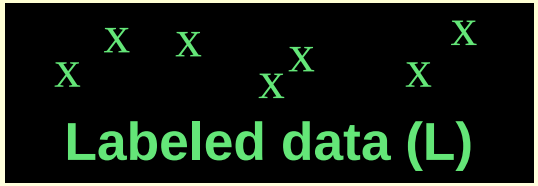
Self-Training with Bagging [Banko and Brill, 2001]

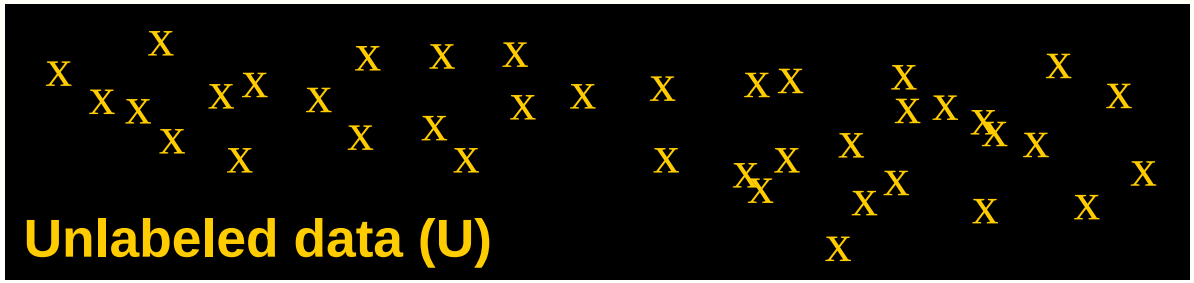


Self-Training with Bagging [Banko and Brill, 2001]

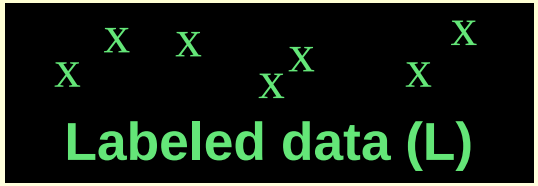
- Single-view algorithm
- Given the labeled and unlabeled data, only need to decide the number of bags to use

Weakly Supervised EM [Nigam *et al*., 2000]

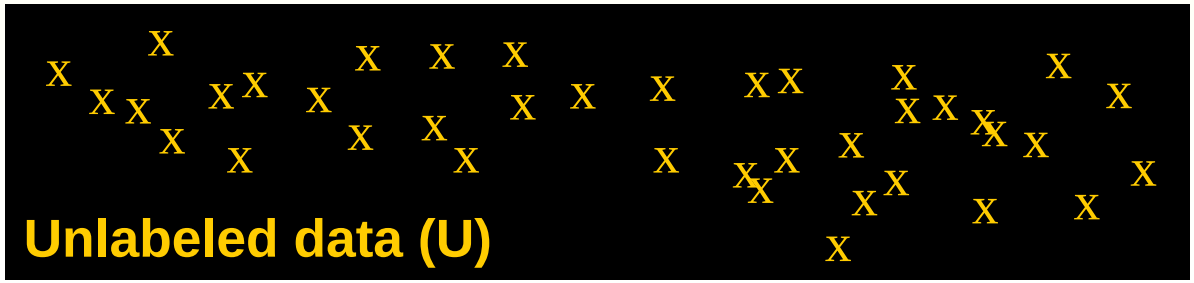




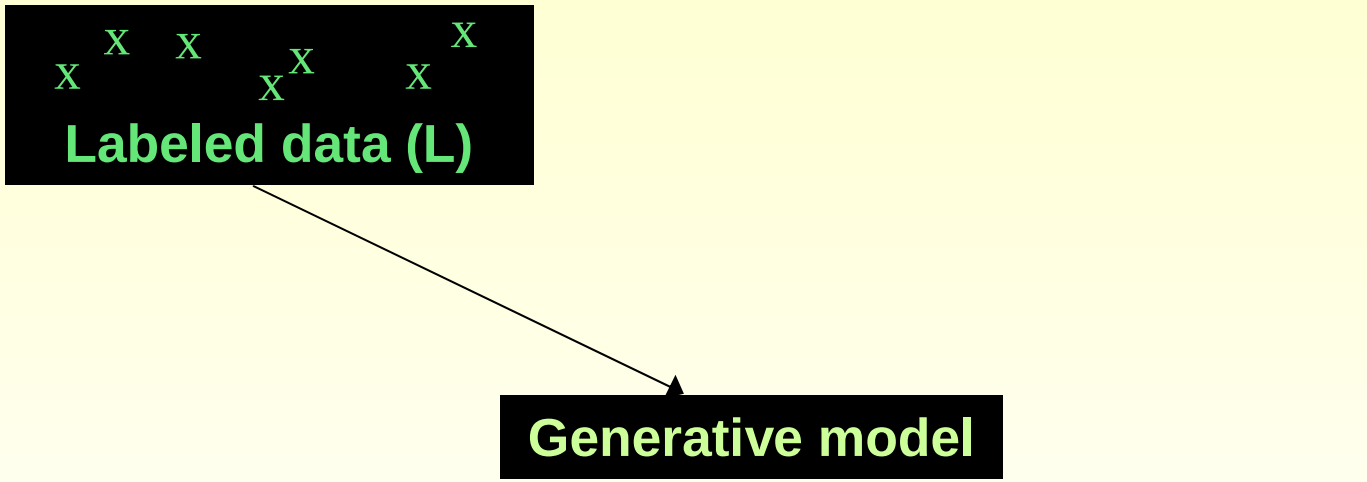
Weakly Supervised EM [Nigam *et al*., 2000]



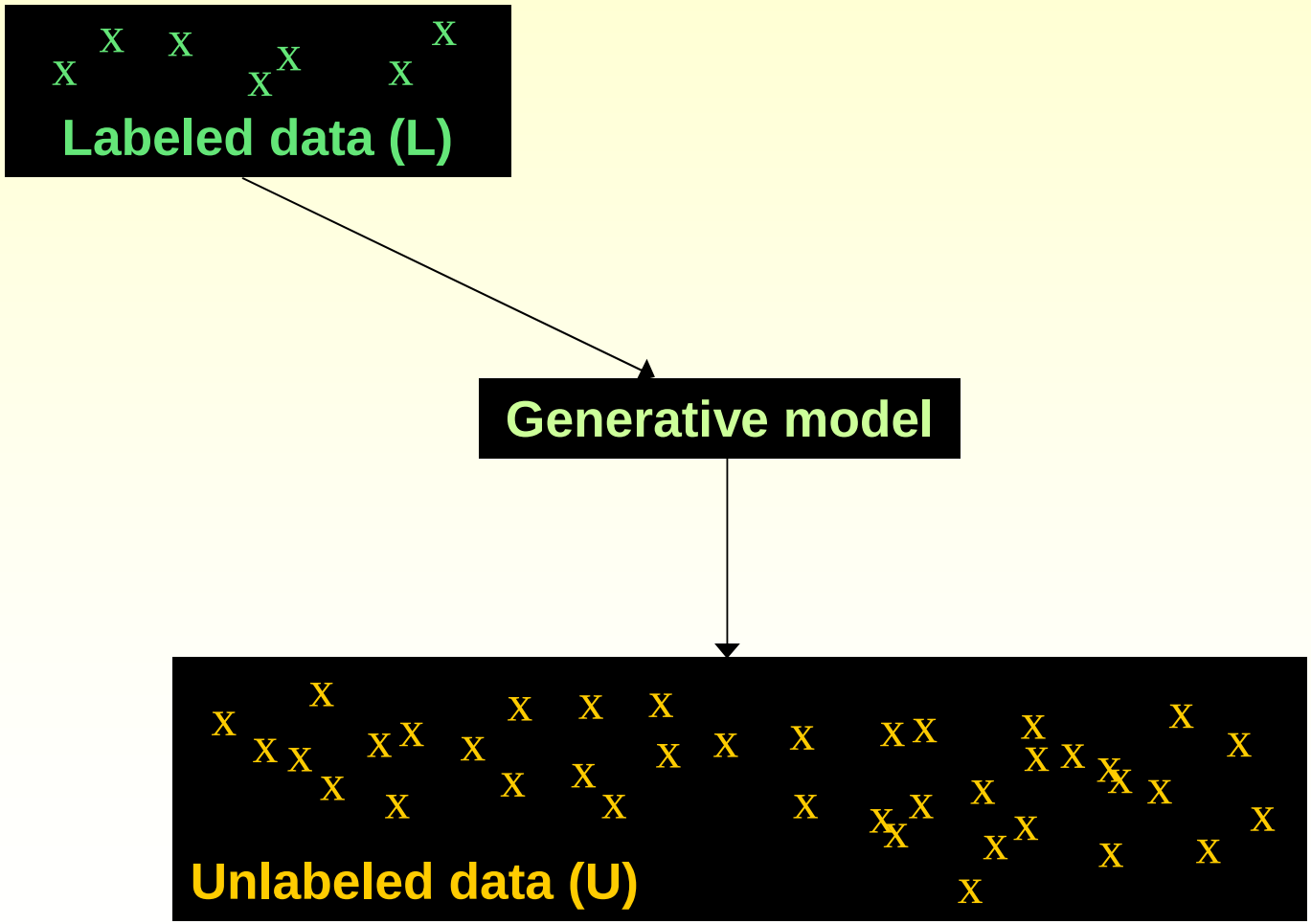
Generative model



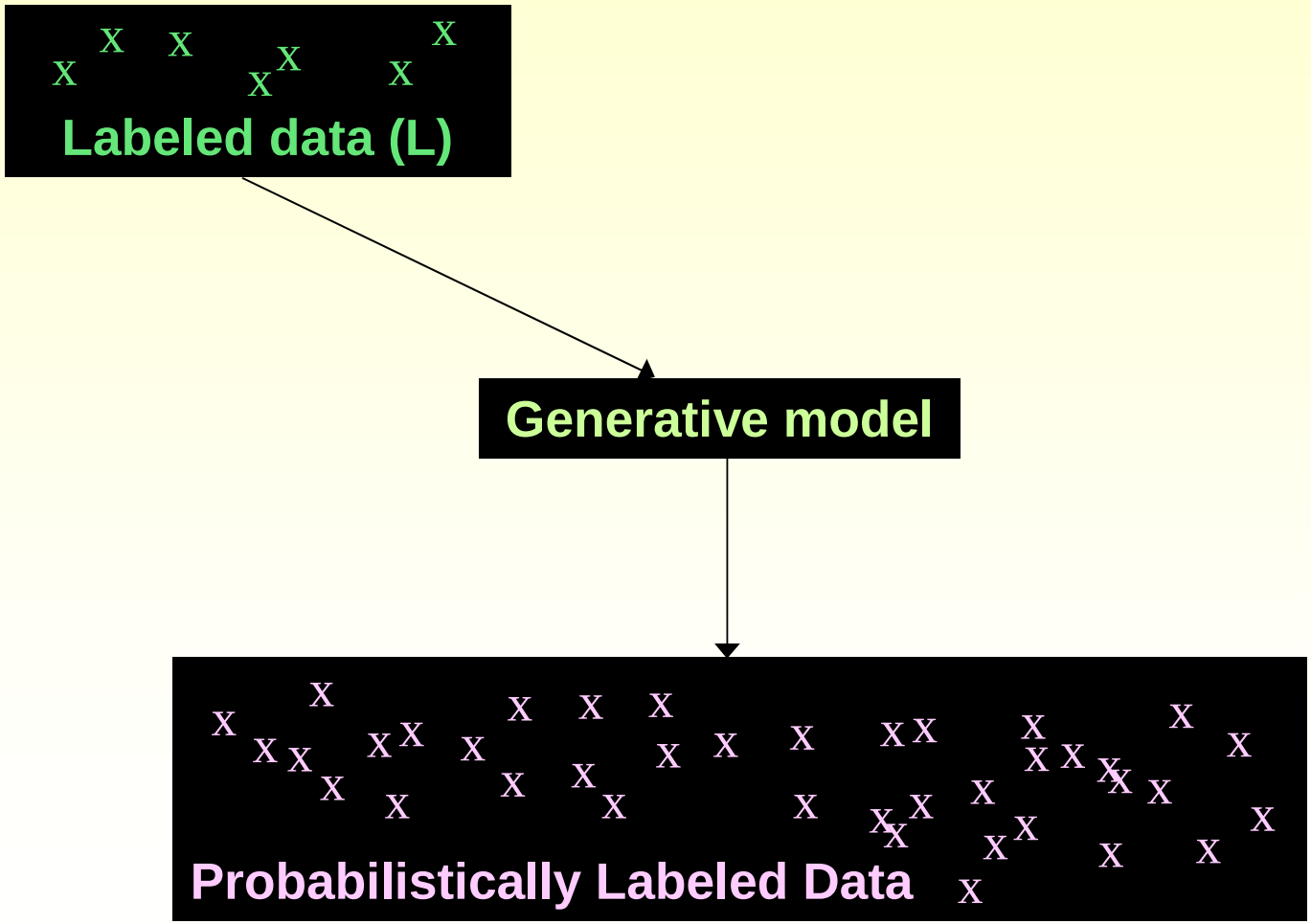
Weakly Supervised EM [Nigam *et al*., 2000]



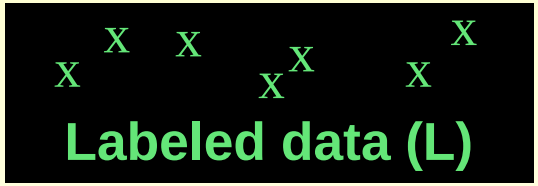
Weakly Supervised EM [Nigam *et al*., 2000]



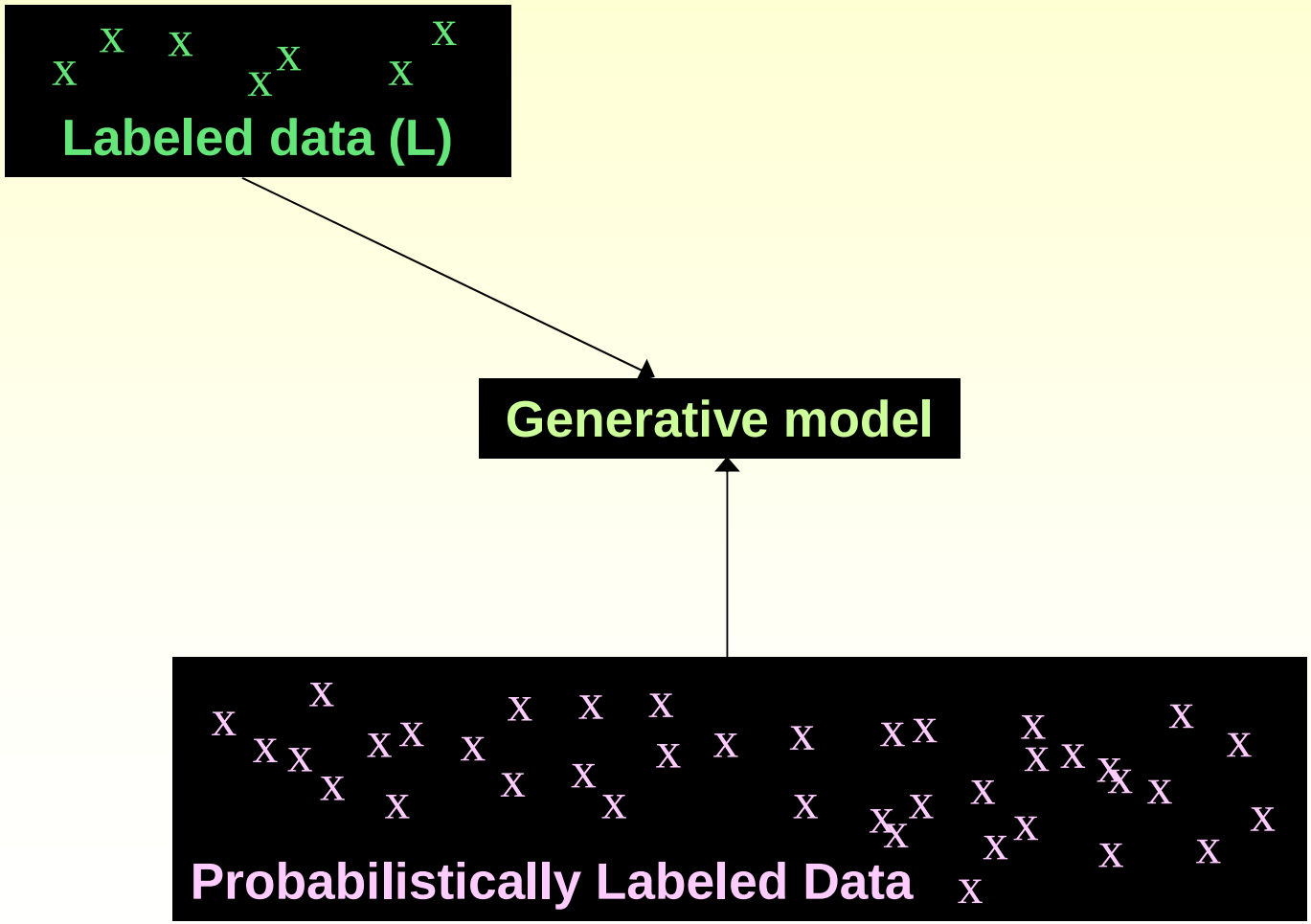
Weakly Supervised EM [Nigam *et al*., 2000]



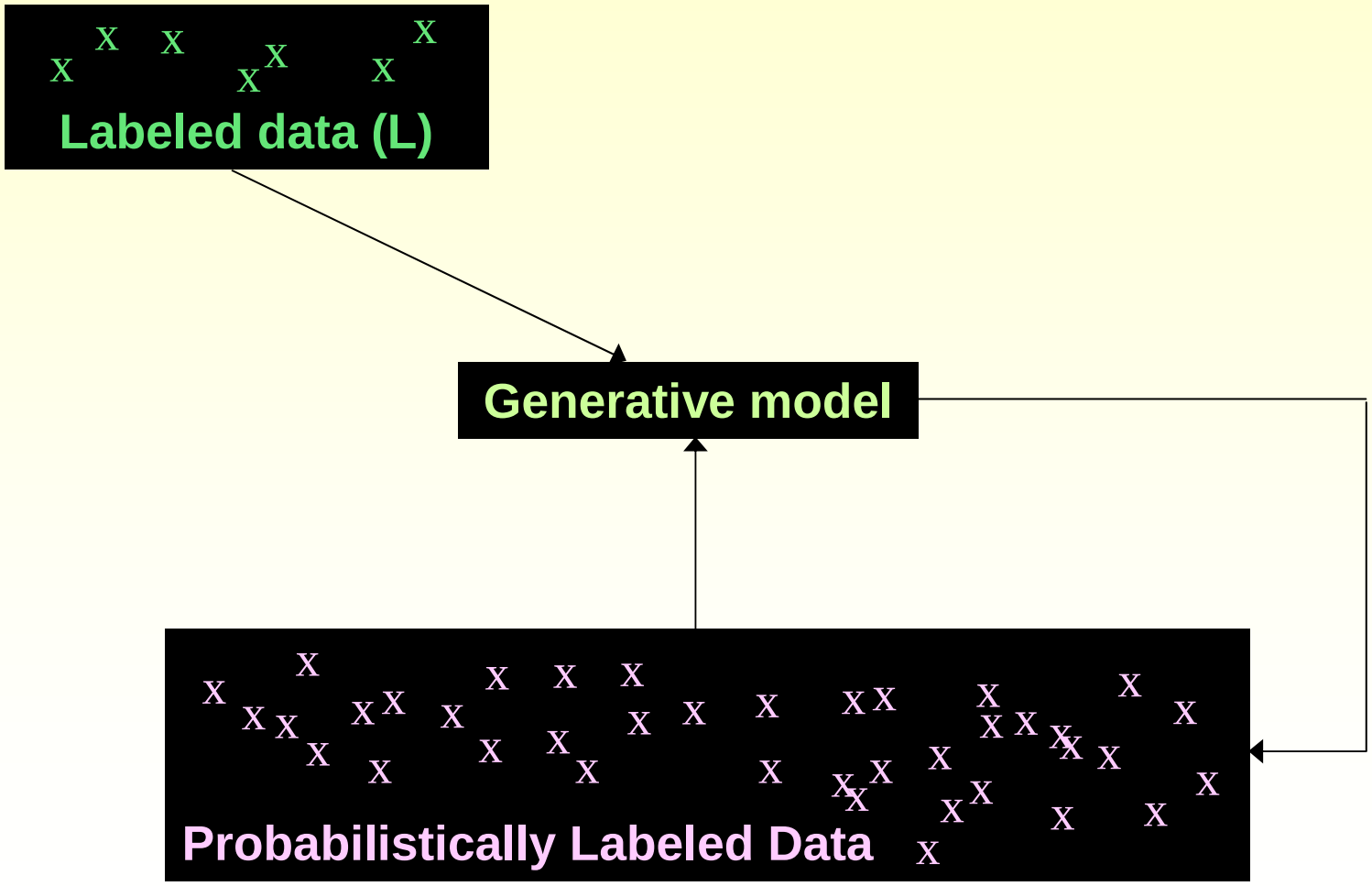
Weakly Supervised EM [Nigam *et al*., 2000]



Weakly Supervised EM [Nigam *et al*., 2000]



Weakly Supervised EM [Nigam *et al*., 2000]



Weakly Supervised EM [Nigam *et al*., 2000]

- A single-view algorithm
- Given the labeled and unlabeled data, only need to decide the number of iterations to run EM

Outline of the Talk

- Weakly supervised learning algorithms
  - co-training
  - self-training with bagging
  - weakly supervised EM
- A learning task without a natural feature split
  - supervised approaches
  - weakly supervised approaches
- Evaluation
- An EM-based bootstrapping algorithm

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Condoleezza Rice is a tenured professor in Stanford's political science department. Her interest in political science was stimulated by Josef Korbel, who is former Secretary of State Madeleine Albright's father...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Condoleezza Rice is a tenured professor in Stanford's political science department. Her interest in political science was stimulated by Josef Korbel, who is former Secretary of State Madeleine Albright's father...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Condoleezza Rice is a tenured professor in Stanford's political science department. Her interest in political science was stimulated by Josef Korbel, who is former Secretary of State Madeleine Albright's father...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Condoleezza Rice is a tenured professor in Stanford's political science department. Her interest in political science was stimulated by Josef Korbel, who is former Secretary of State Madeleine Albright's father...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Condoleezza Rice is a tenured professor in Stanford's political science department. Her interest in political science was stimulated by Josef Korbel, who is former Secretary of State Madeleine Albright's father...

Outline of the Talk

- Weakly supervised learning algorithms
  - co-training
  - self-training with bagging
  - weakly supervised EM
- A learning task without a natural feature split
  - supervised and weakly supervised approaches
- Evaluation
- An EM-based bootstrapping algorithm

The Supervised Learning Approach

- A two-step approach: classification + clustering
- Classification
  - classifies a pair of NPs as coreferent or not based on constraints learned from annotated data
- Clustering
  - coordinates the possibly contradictory pairwise classifications and constructs a partition on the set of NPs

The Weakly Supervised Learning Approach

- Use a weakly supervised algorithm to bootstrap the coreference classifier from a small set of labeled data

- The clustering mechanism is not manipulated by the bootstrapping procedure

The Coreference Resolution System

- Learning algorithm
  - naïve Bayes
- Clustering algorithm
  - best-first clustering
- Instance representation
  - 25 features per instance (created from each pair of NPs)
    - lexical
    - grammatical
    - semantic
    - positional

Any Natural Feature Split for Coreference?

- Views cannot be drawn from the left-hand and right-hand context
- Views cannot be drawn from features inside and outside the phrase under consideration
- View factorization is a non-trivial problem for coreference
  - Mueller *et al*.'s (2002) greedy method

Outline of the Talk

- Weakly supervised learning algorithms
  - co-training
  - self-training with bagging
  - weakly supervised EM
- A learning task without a natural feature split
  - supervised and weakly supervised approaches
- Evaluation
- An EM-based bootstrapping algorithm

Data Sets

- MUC-6 and MUC-7 coreference data sets
  - Documents annotated with coreference information
  - MUC-6: 30 dryrun texts + 30 evaluation texts
  - MUC-7: 30 dryrun texts + 20 evaluation texts

- Evaluation texts
  - reserved for testing

- Dryrun texts
  - one used as labeled data (L)
  - remaining 29 as unlabeled data (U)

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

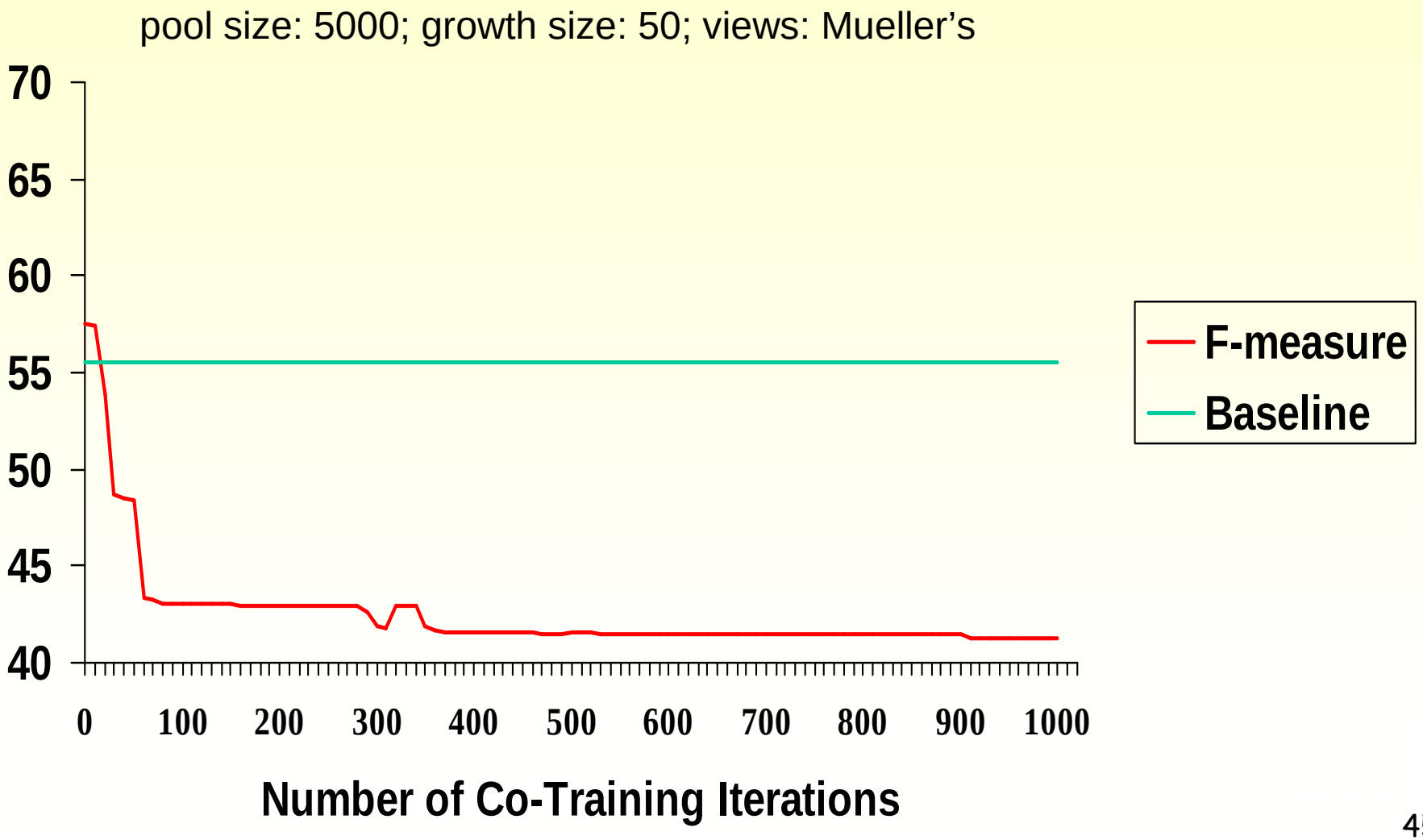
- Views (3 heuristic methods for view factorization)
  - Mueller *et al*.'s (2002) greedy method
  - random splitting
  - splitting according to the feature type
- Pool size
  - 500, 1000, 5000
- Growth size
  - 10, 50, 100, 200, 250
- Number of co-training iterations
  - run until performance stabilized

Results (Co-Training)

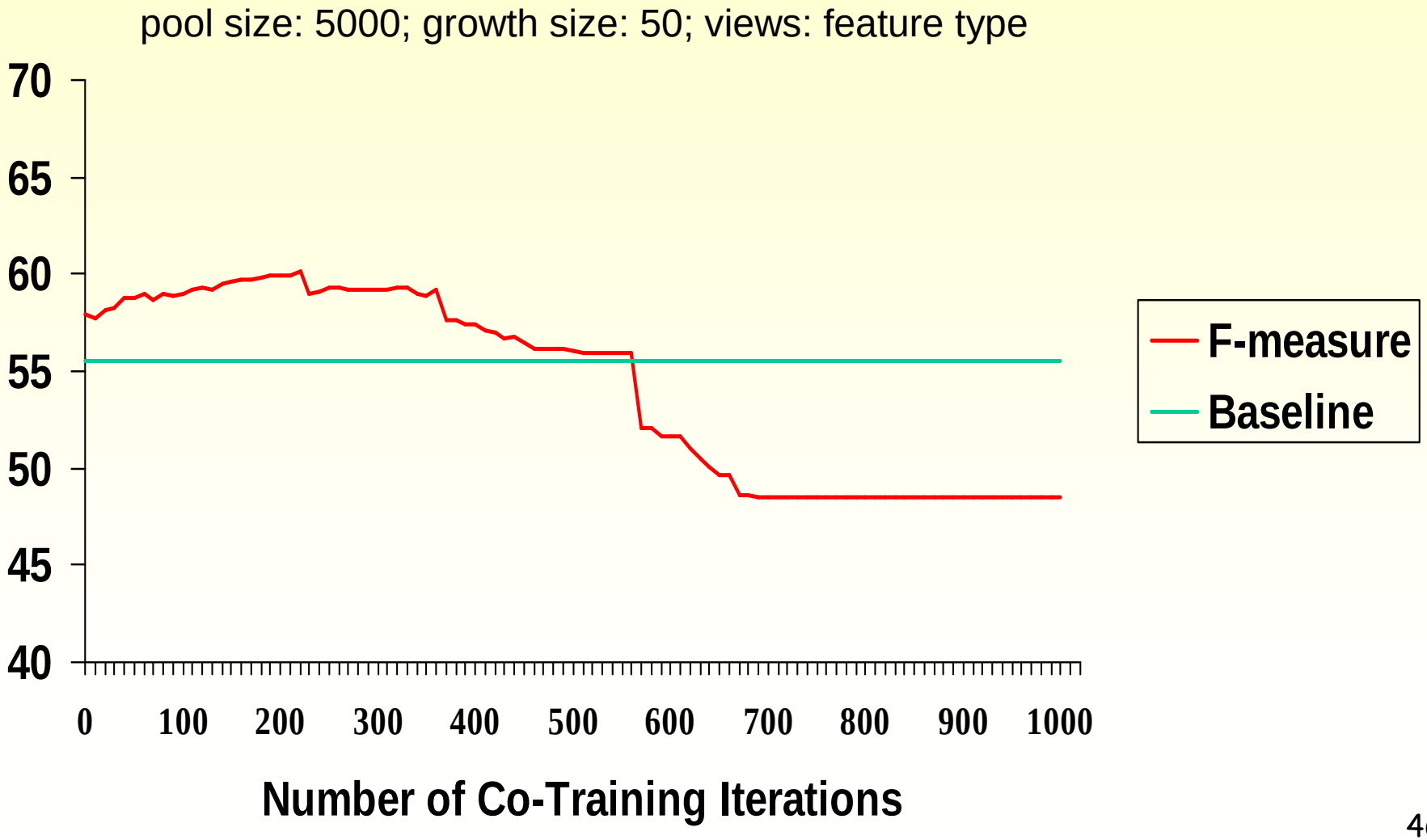
	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	**55.5**	52.8	37.4	**43.8**
Co-Training	47.5	81.9	**60.1**	40.6	77.6	**53.3**

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

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

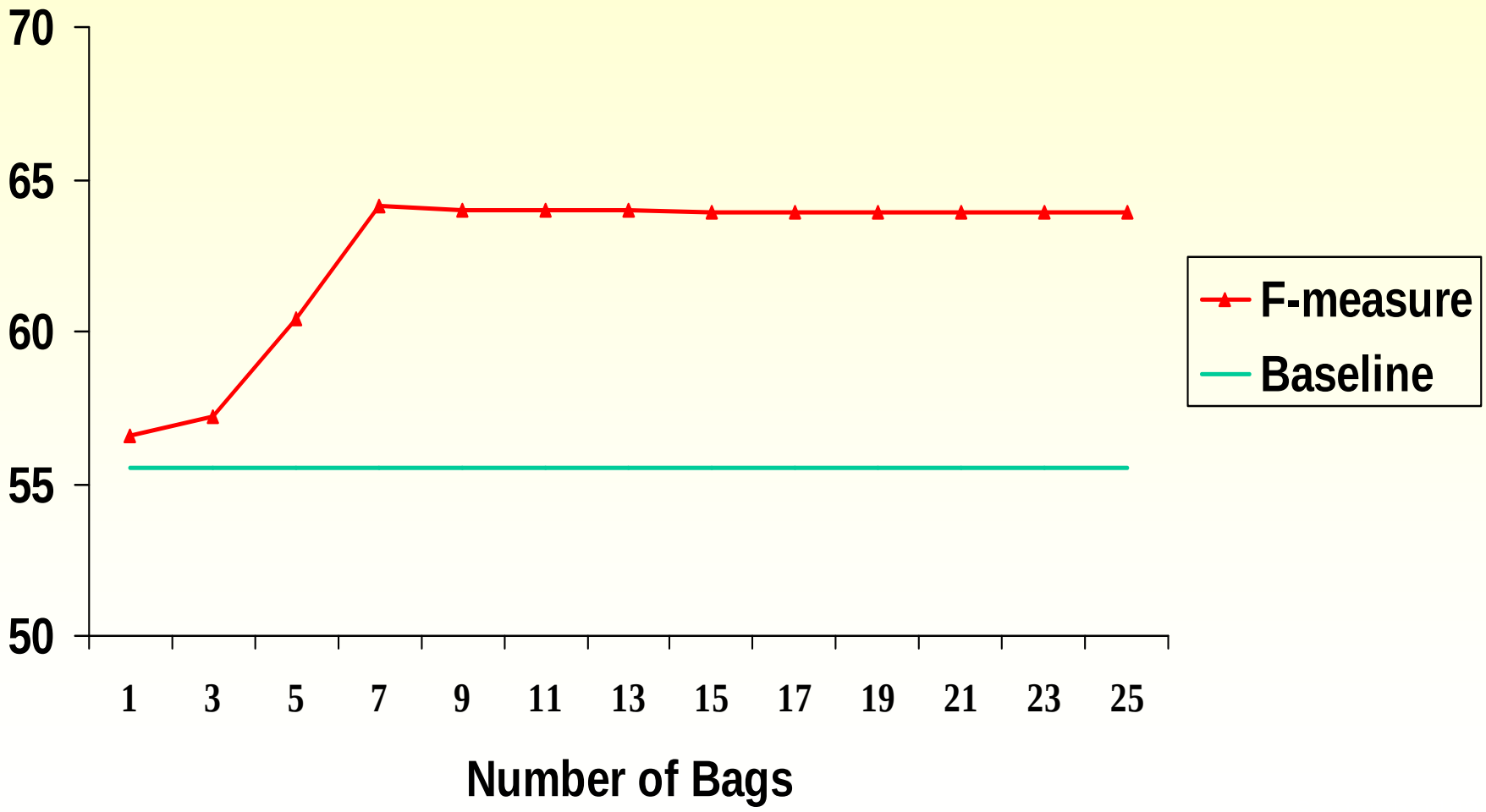
- 25 bags are sufficient for most learning tasks (Breiman, 1996)

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)



EM Parameters

- Number of iterations
  - run until convergence

Results (EM)

	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**
EM	64.8	51.8	**57.6**	54.1	40.7	**46.4**

- EM only gives rise to modest performance gains over the baseline

Results (EM)

	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**
EM	64.8	51.8	**57.6**	54.1	40.7	**46.4**

- EM does not perform as well as co-training

Summary of Results

- Applied one multi-view weakly supervised algorithm and two single-view algorithms to coreference resolution
  - Co-training outperforms the baseline at its best parameter setting
  - Self-training with bagging significantly outperforms co-training
  - EM only performs slightly better than the baseline

Why EM Doesn't Work Well

- Hypothesis: generative model is not correct
- Plausible solution: improve the model via feature selection
- The feature selection algorithm
  - imposes a total ordering on the features
  - selects the first *n* features

The Feature Selection Algorithm

The Feature Selection Algorithm

Training data

The Feature Selection Algorithm

Training data

Validation data

The Feature Selection Algorithm

Training data

Validation data

Scoring function

The Feature Selection Algorithm

Features selected thus far (S)

Training data

Validation data

Scoring function

The Feature Selection Algorithm

Features selected thus far (S)

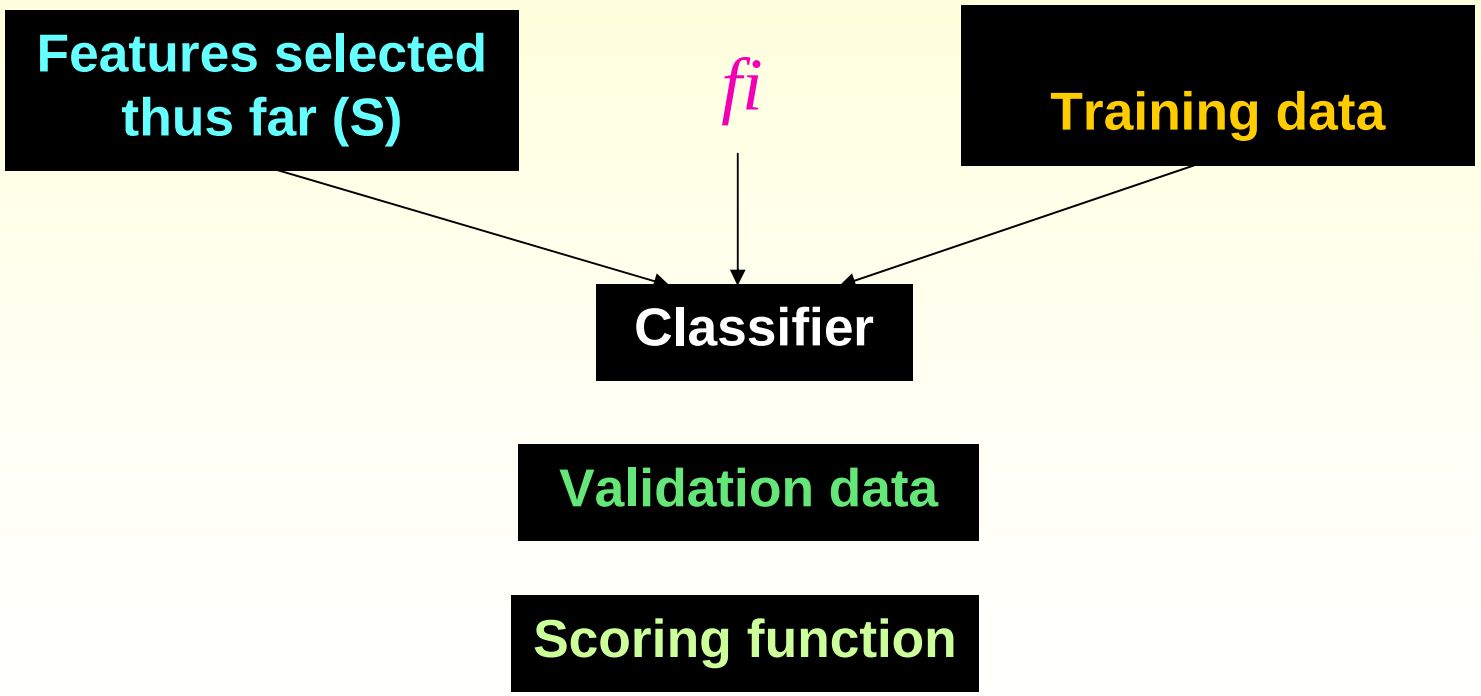
fi

Training data

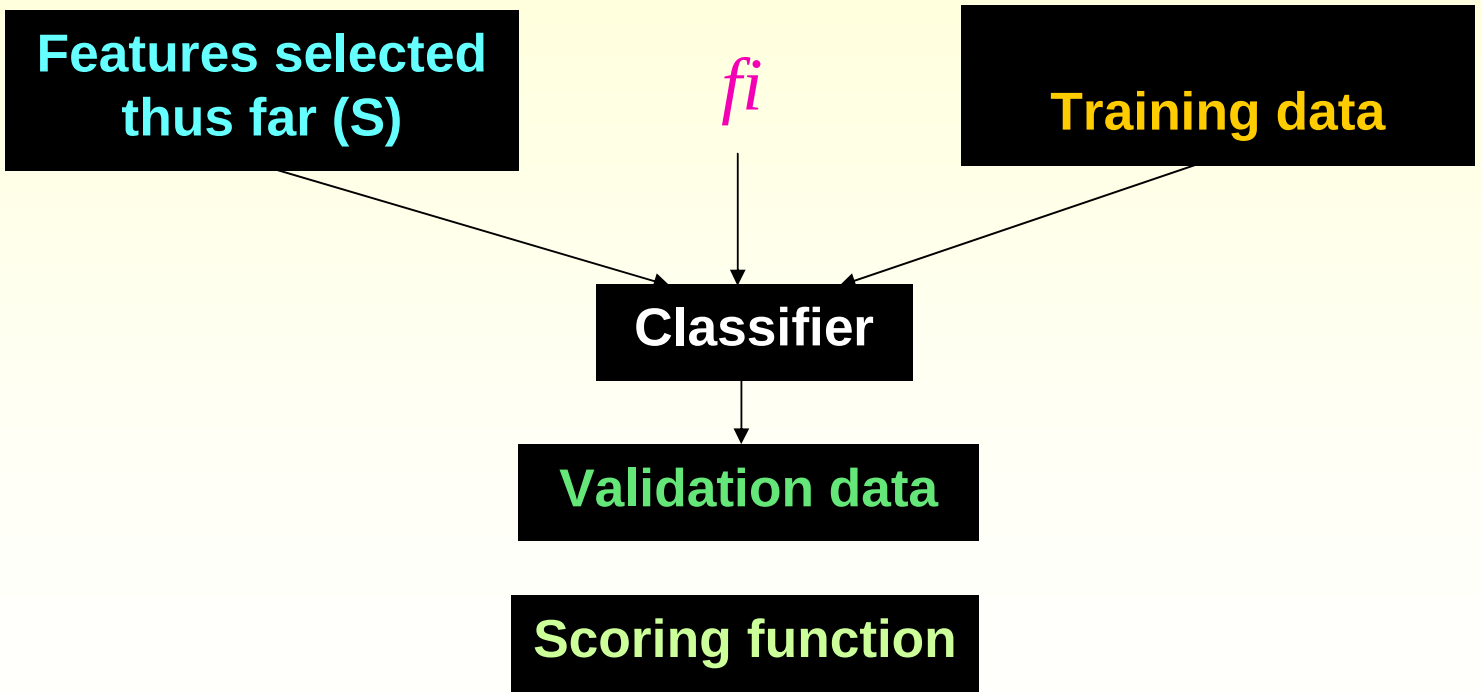
Validation data

Scoring function

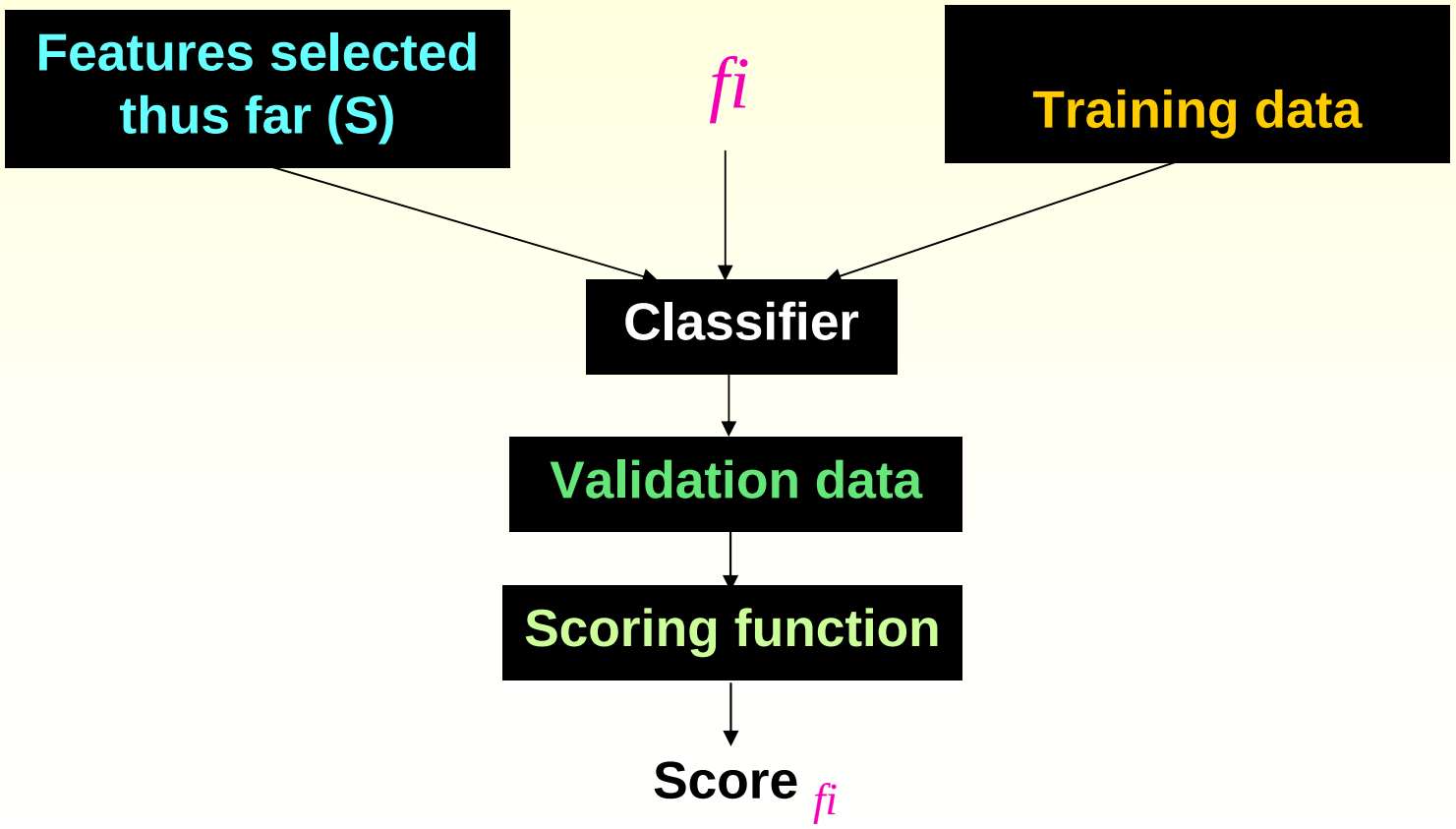
The Feature Selection Algorithm



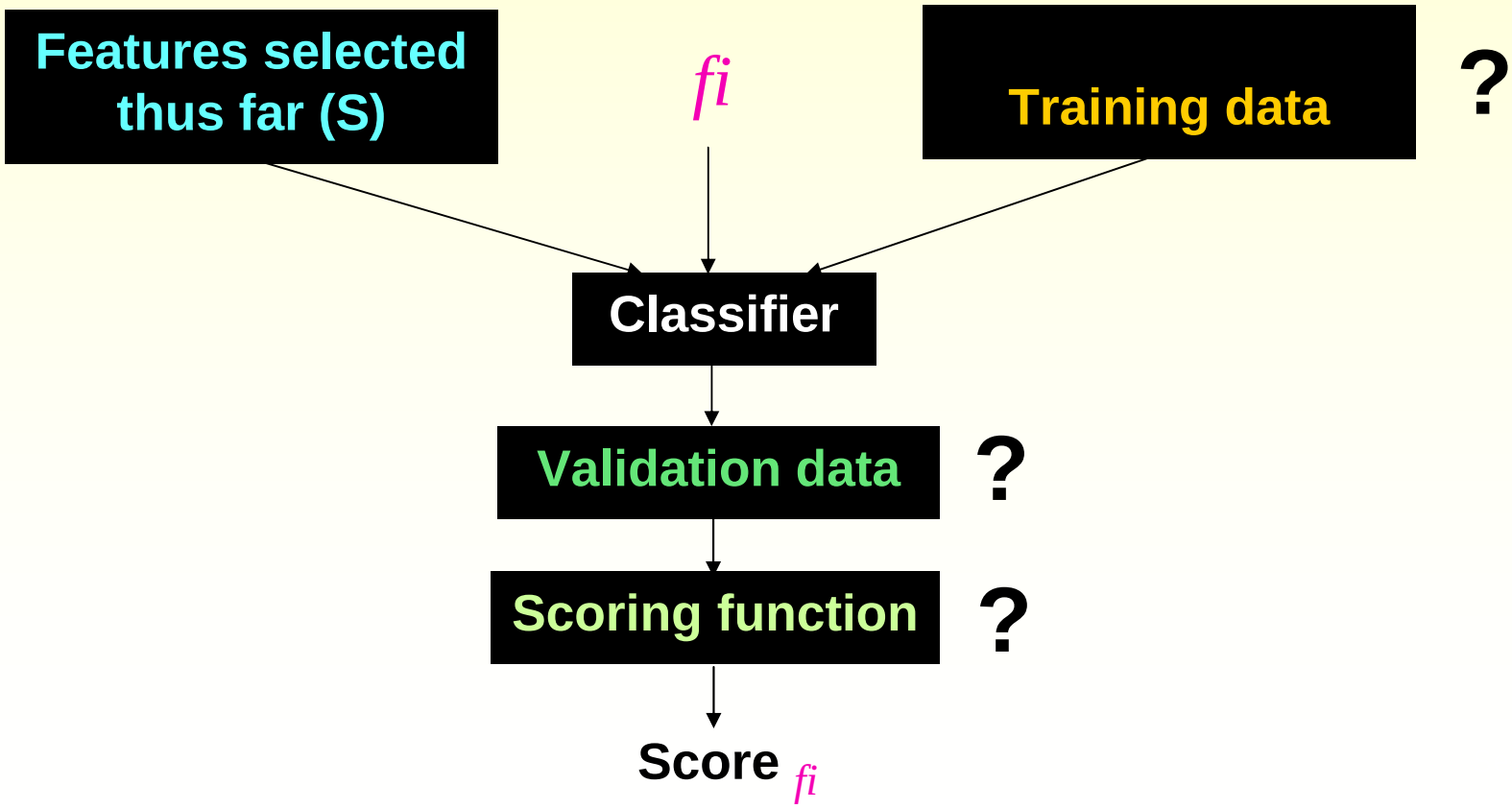
The Feature Selection Algorithm



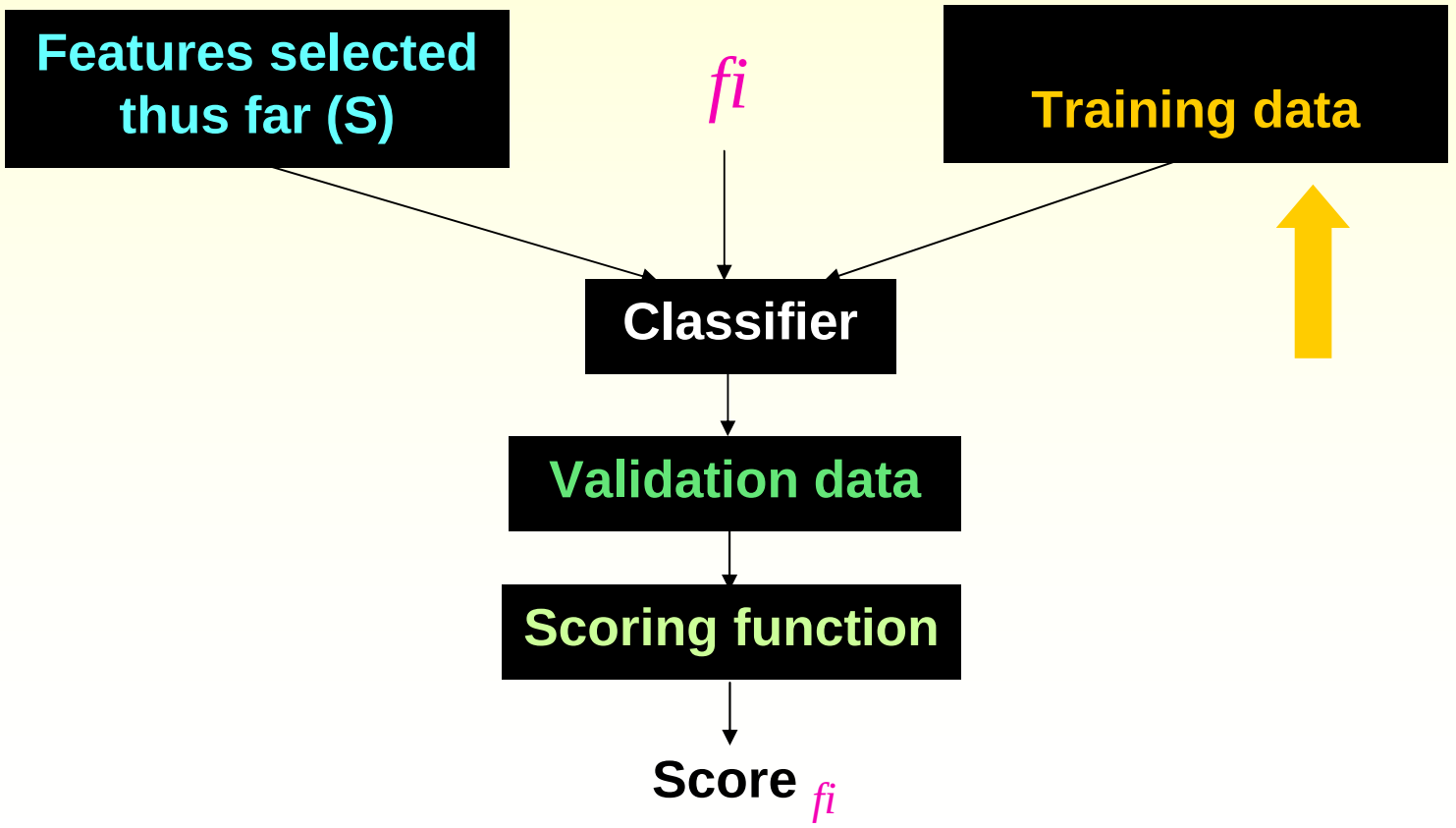
The Feature Selection Algorithm



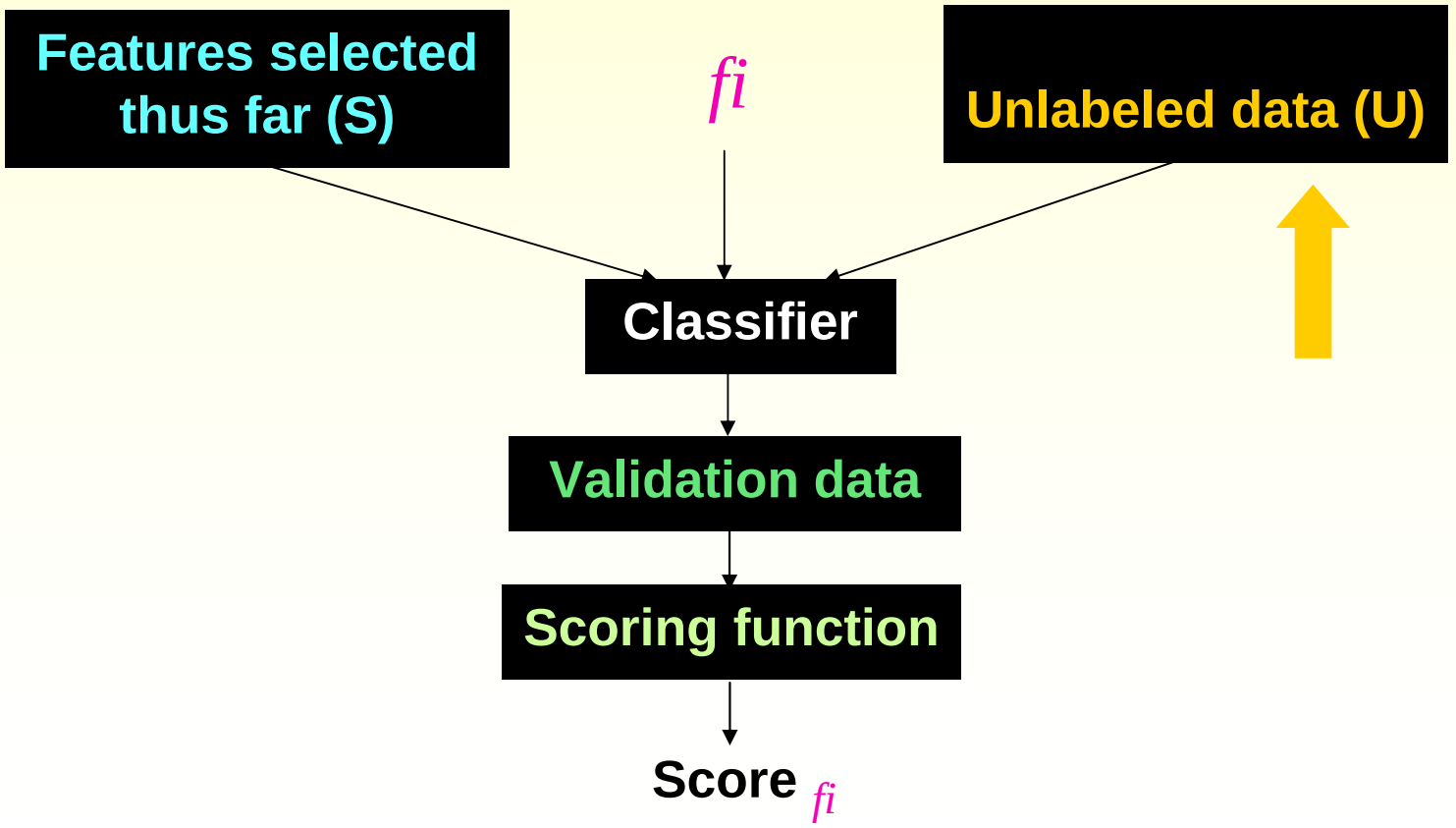
The Feature Selection Algorithm



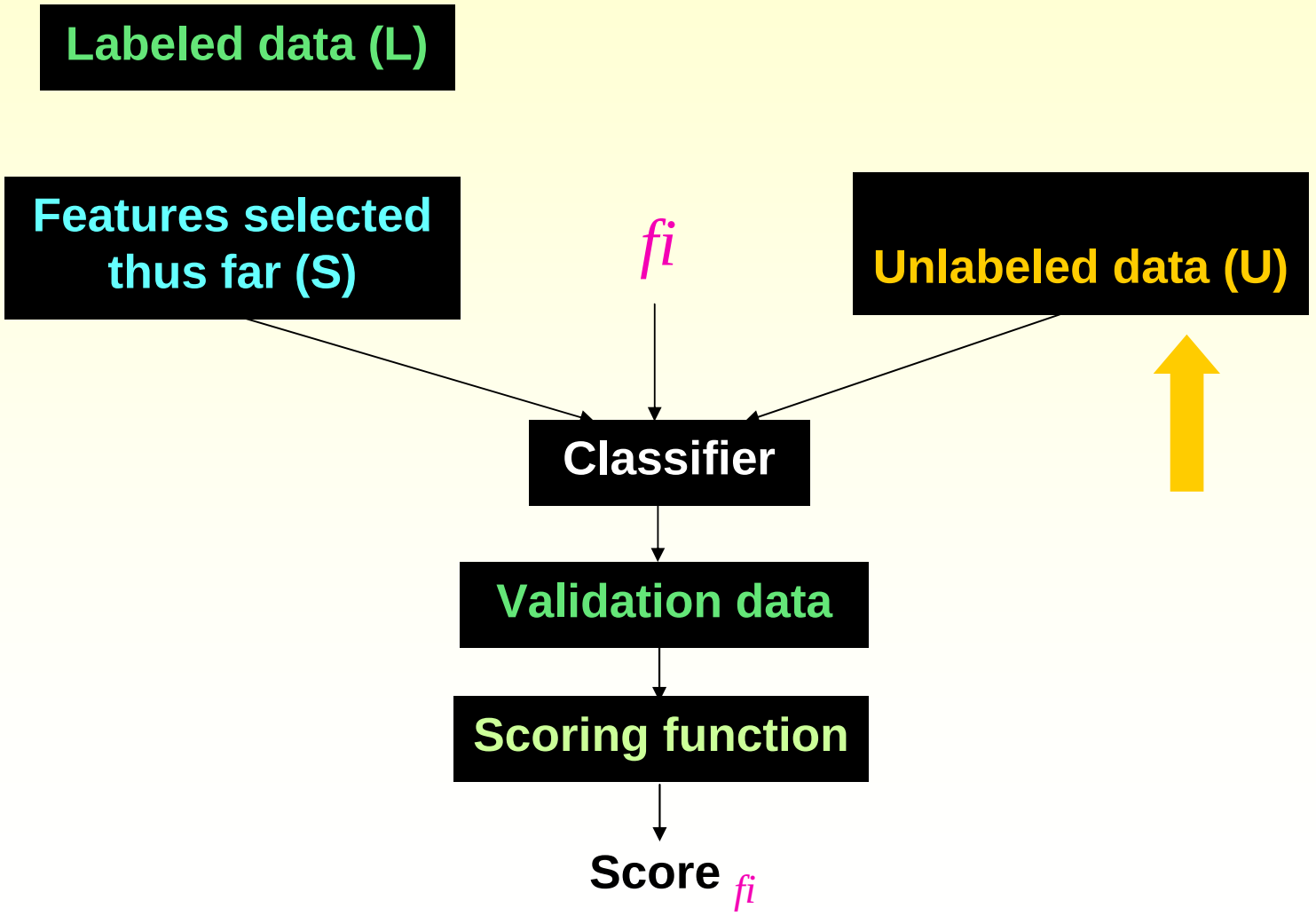
The Feature Selection Algorithm



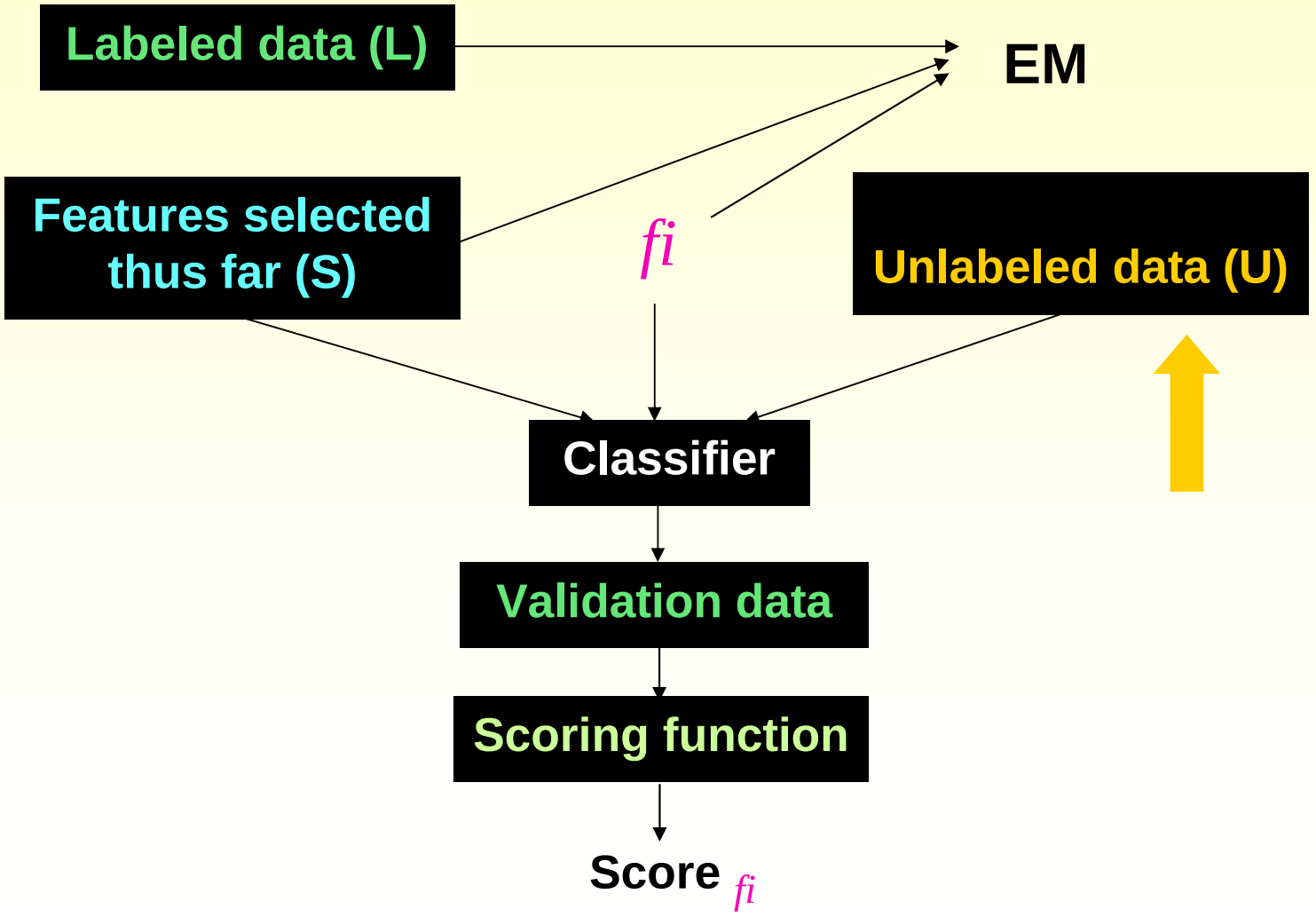
The Feature Selection Algorithm



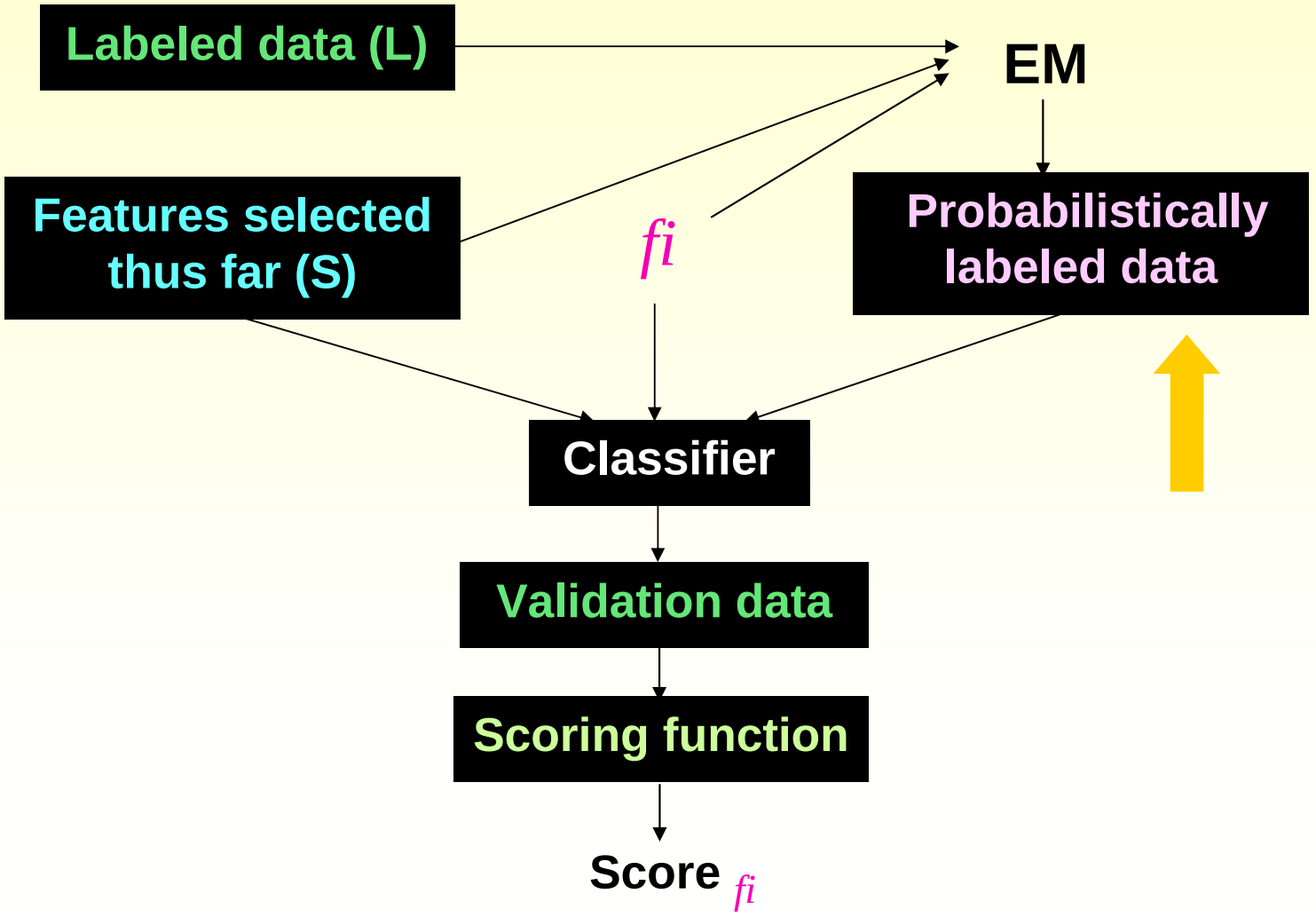
The Feature Selection Algorithm



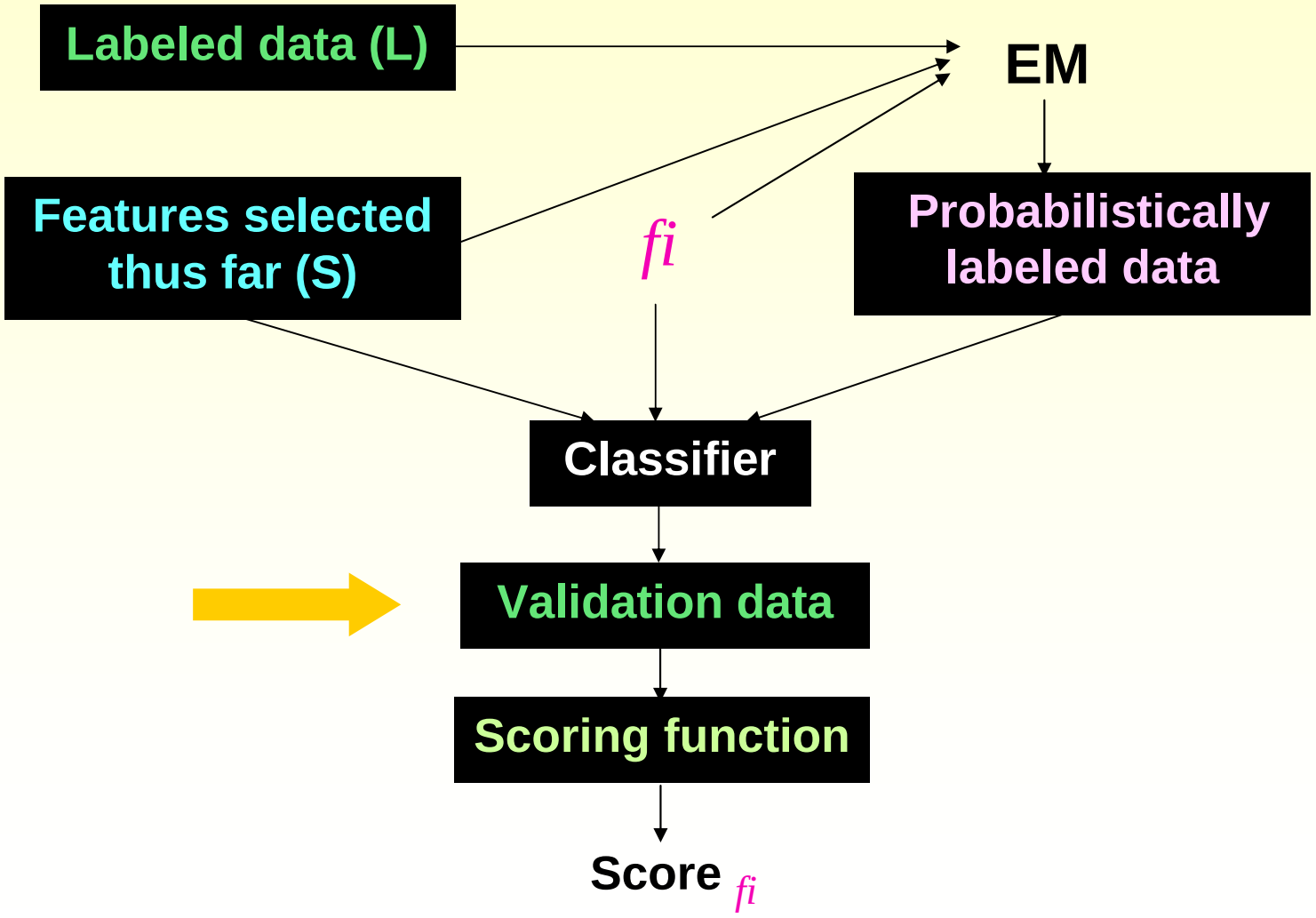
The Feature Selection Algorithm



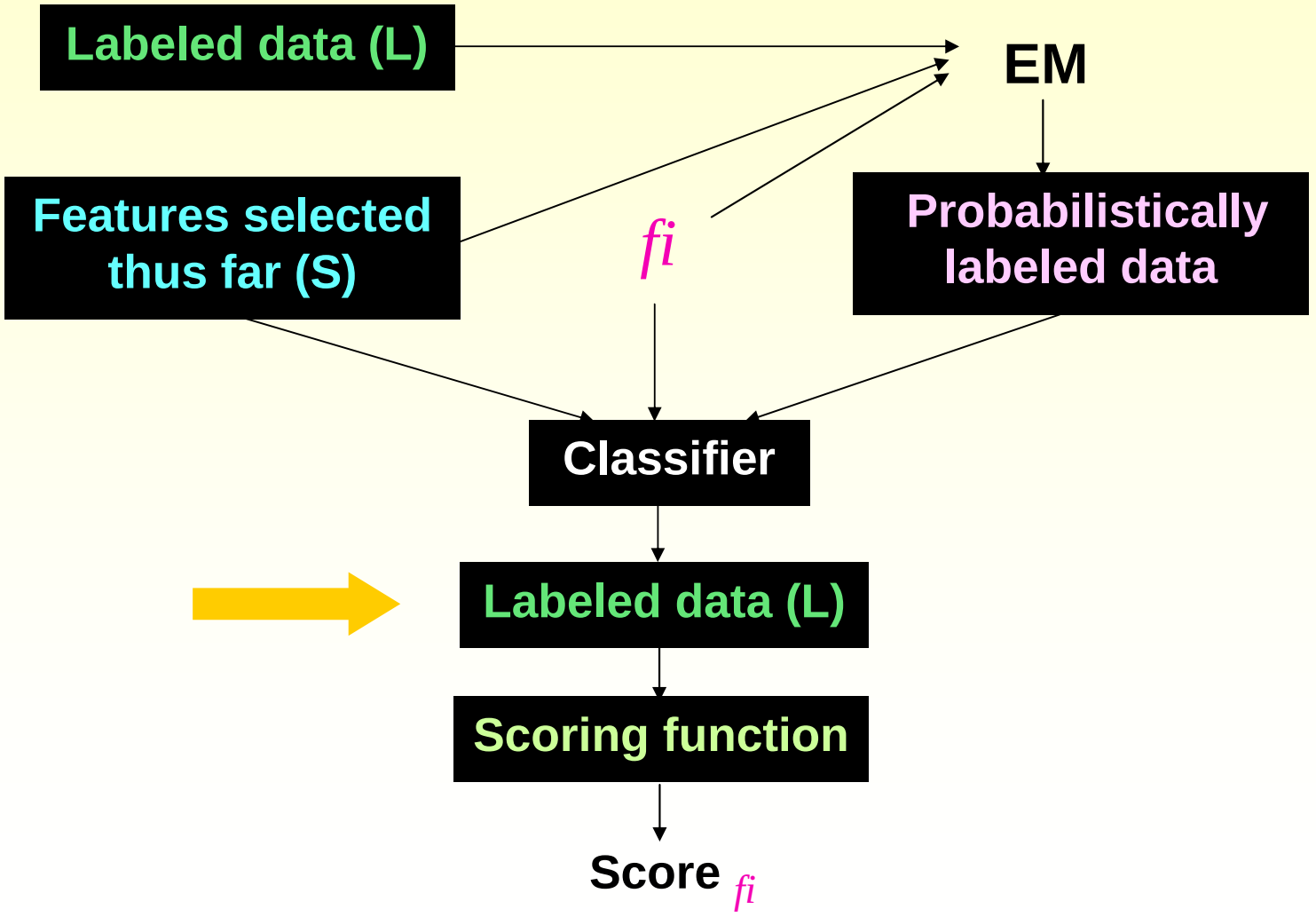
The Feature Selection Algorithm



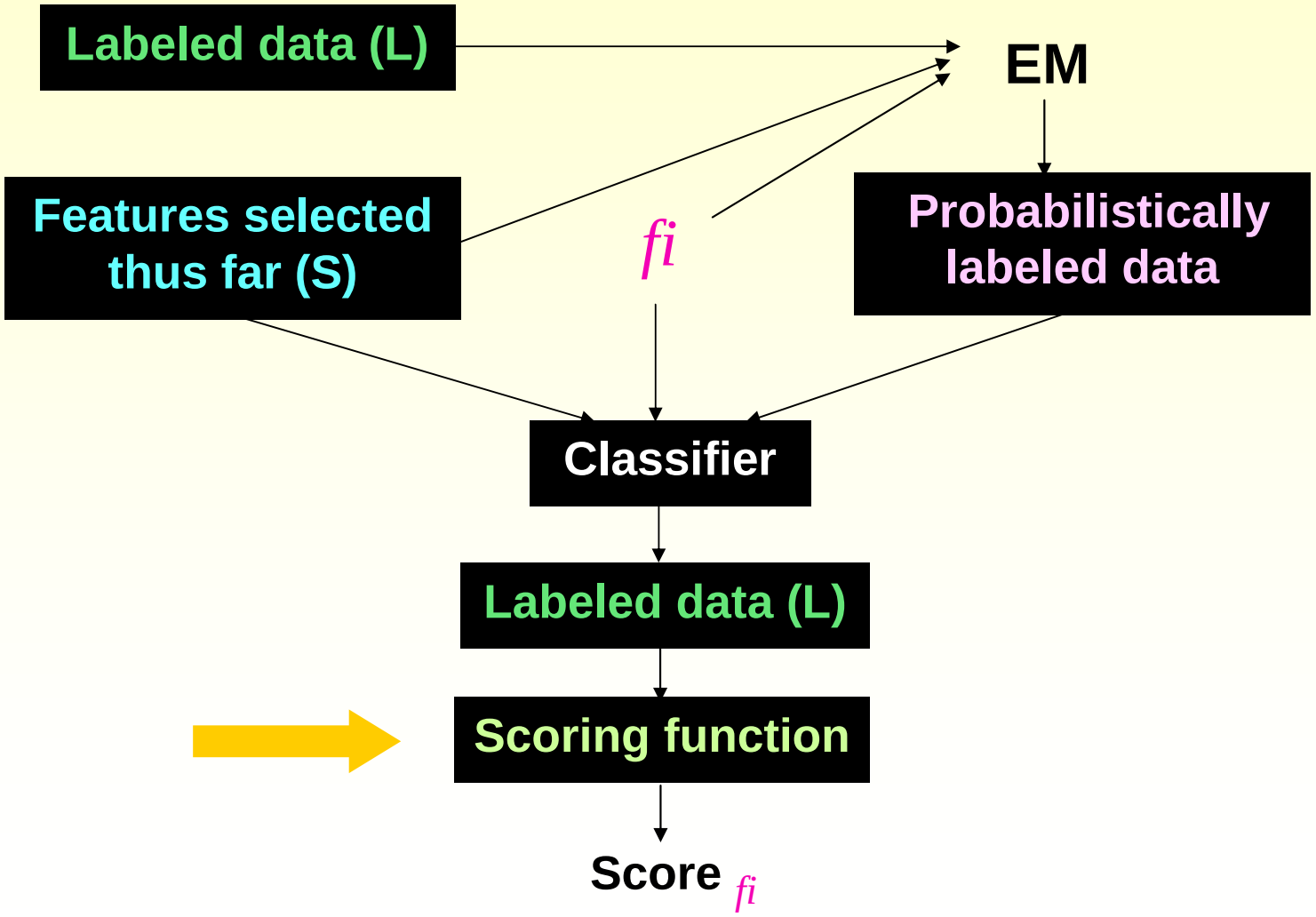
The Feature Selection Algorithm



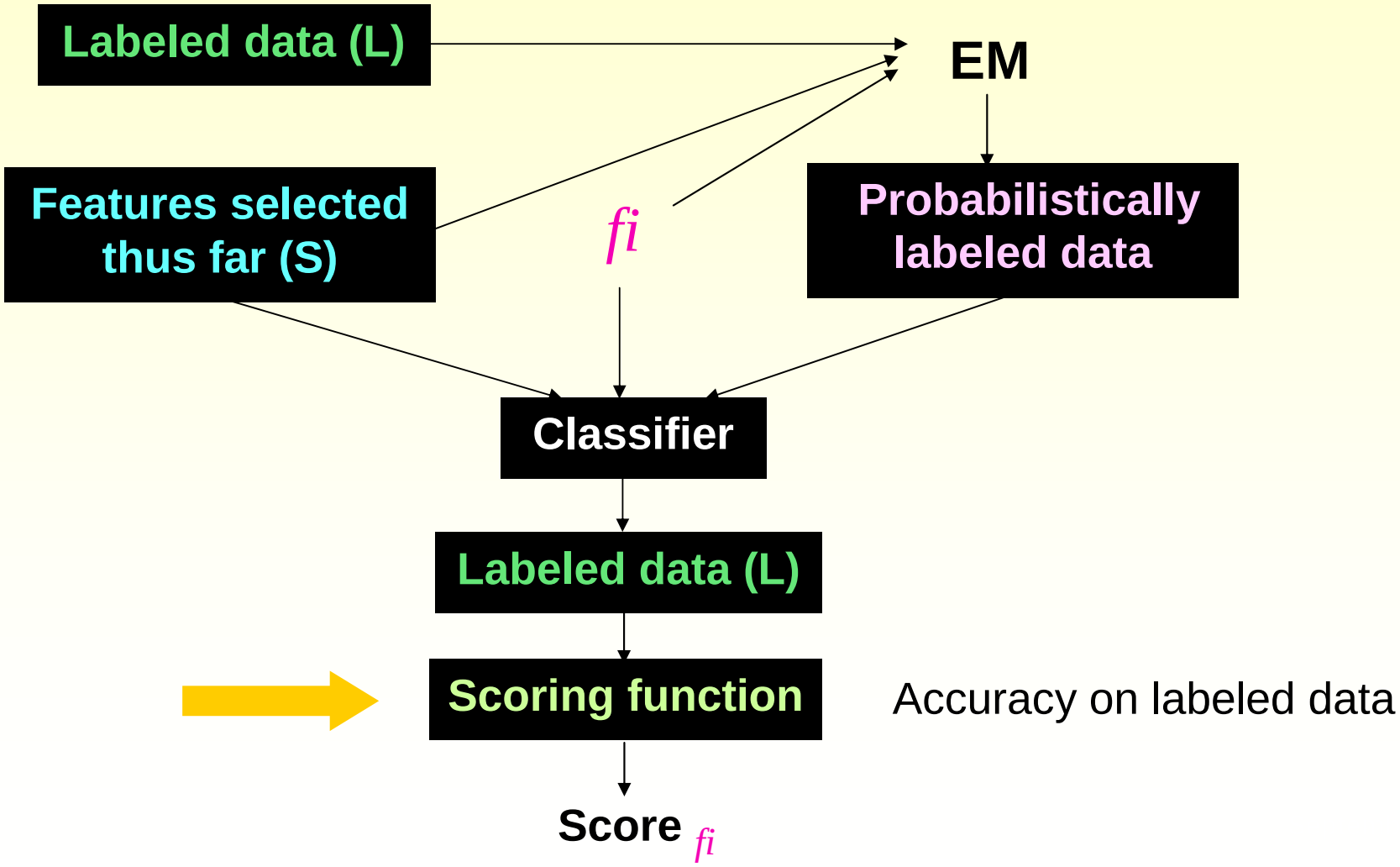
The Feature Selection Algorithm



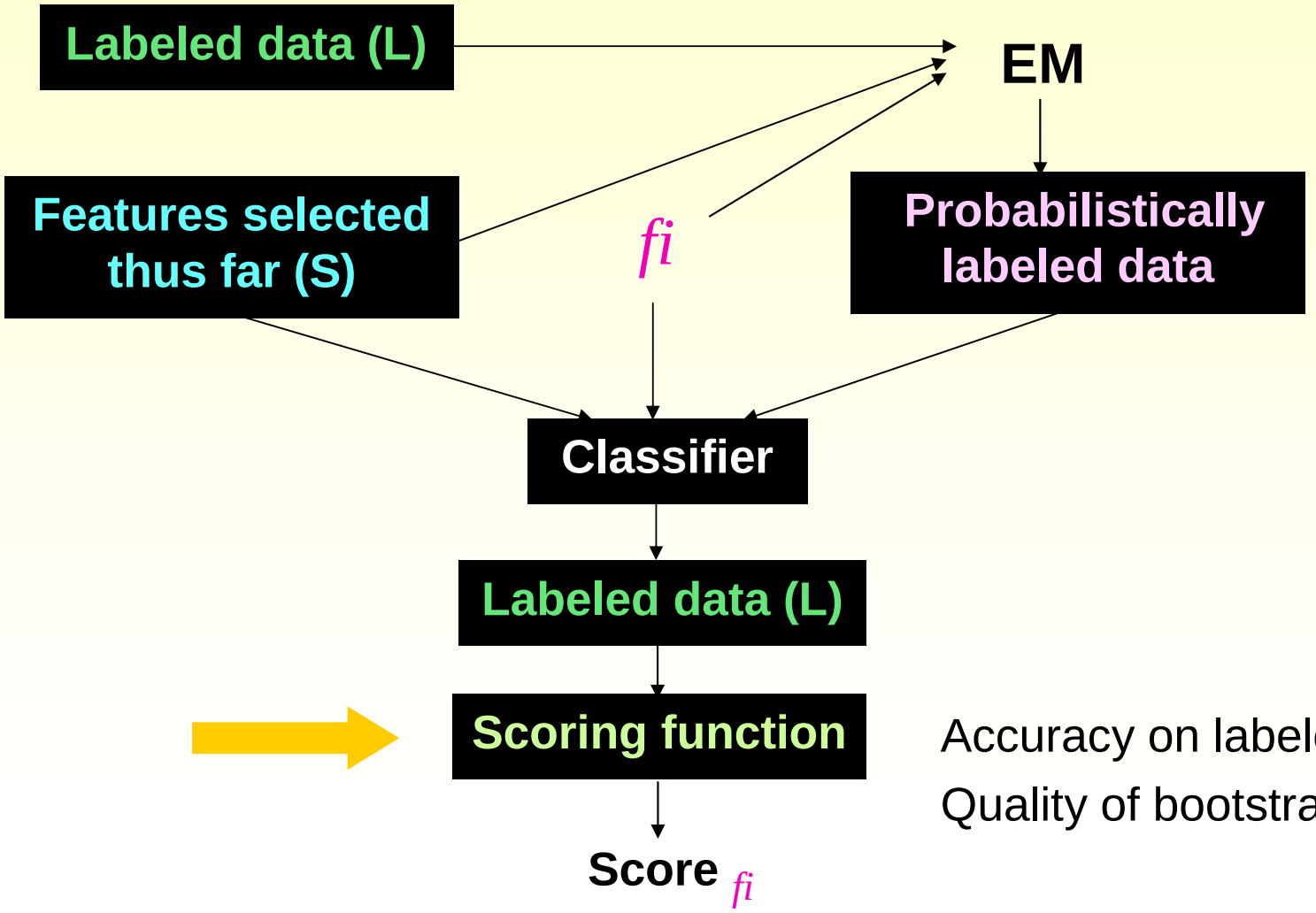
The Feature Selection Algorithm



The Feature Selection Algorithm



The Feature Selection Algorithm



Accuracy on labeled data

Quality of bootstrapped data

Results (FS-EM)

	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**
EM	64.8	51.8	**57.6**	54.1	40.7	**46.4**
FS-EM	64.2	66.6	**65.4**	53.3	70.3	**60.5**

- FS-EM performs better than co-training

Results (FS-EM)

	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**
EM	64.8	51.8	**57.6**	54.1	40.7	**46.4**
FS-EM	64.2	66.6	**65.4**	53.3	70.3	**60.5**

- FS-EM performs slightly better than self-training with bagging

Summary

- Investigated single-view weakly supervised algorithms as an alternative to multi-view algorithms for coreference resolution
  - Self-training with bagging outperforms co-training under various parameter settings
  - EM does not outperform co-training, but FS-EM does