Learning more with less
Active Learning for Natural Language Processing

Shilpa Arora & Sachin Agarwal

Language Technologies Institute

School of Computer Science

Carnegie Mellon University

6th December 2007

Overview

- Introduction
- Evaluation Measures
- Selective Sampling
  - Uncertainty based
  - Query-by-committee
  - Other methods
- Conclusion

Active Learning

- *Reducing* the *number* of *labeled examples* required to learn a concept

Active Learning

- *Reducing* the *number* of *labeled examples* required to learn a concept

  Why …

  - Annotated data is expensive

Active Learning

- *Reducing* the *number* of *labeled examples* required to learn a concept

  Why …

  - Annotated data is expensive

  How ….

  - All examples are not *equally informative*

Active Learning

- *Not Equally Informative*

1. John ***lives in*** *New York*.
2. Tom ***lives in*** *California*.
3. Noah ***teaches in*** CMU.
4. Eric ***teaches in*** CMU.

1. John ***lives in*** *New York*.
2. Tom ***is settled in*** *California*.
3. Noah ***is a faculty at*** CMU.
4. Eric ***teaches in*** CMU.

Active Learning

Really what we want to do is…

- *Reduce* the *amount* of *user effort* required to learn a concept

Active Learning

Really what we want to do is…

- *Reduce* the *amount* of *user effort* required to learn a concept

And ….

- Number of examples ≠ user effort

Active Learning

Really what we want to do is…

- *Reduce* the *amount* of *user effort* required to learn a concept

And ….

- Number of examples ≠ user effort

Because …

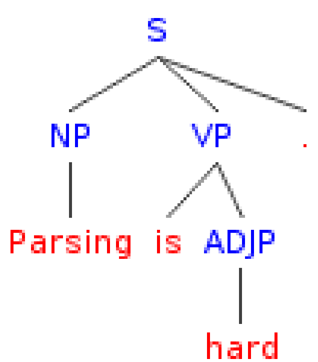
- All examples are not *equally easy to annotate*

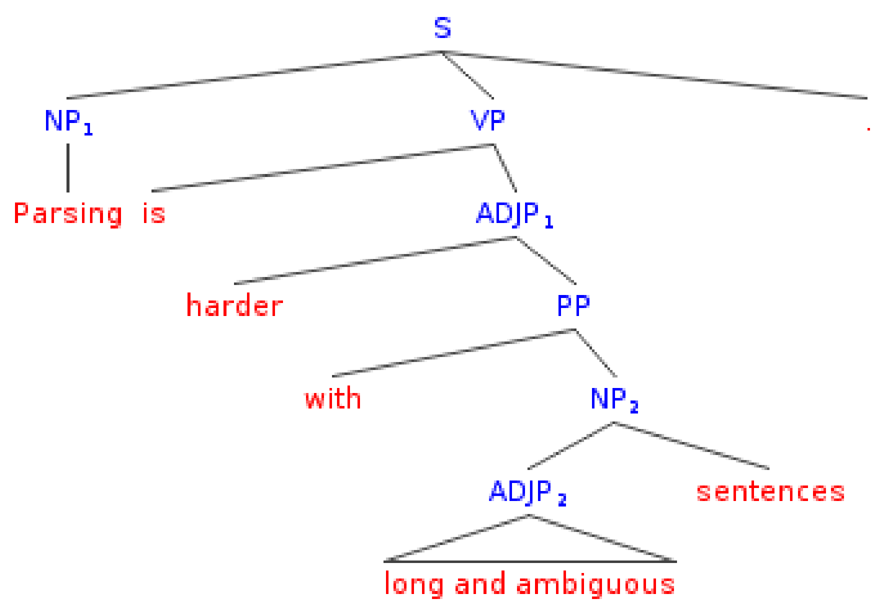
Active Learning

- *Not equally easy to annotate*

Parsing is hard.

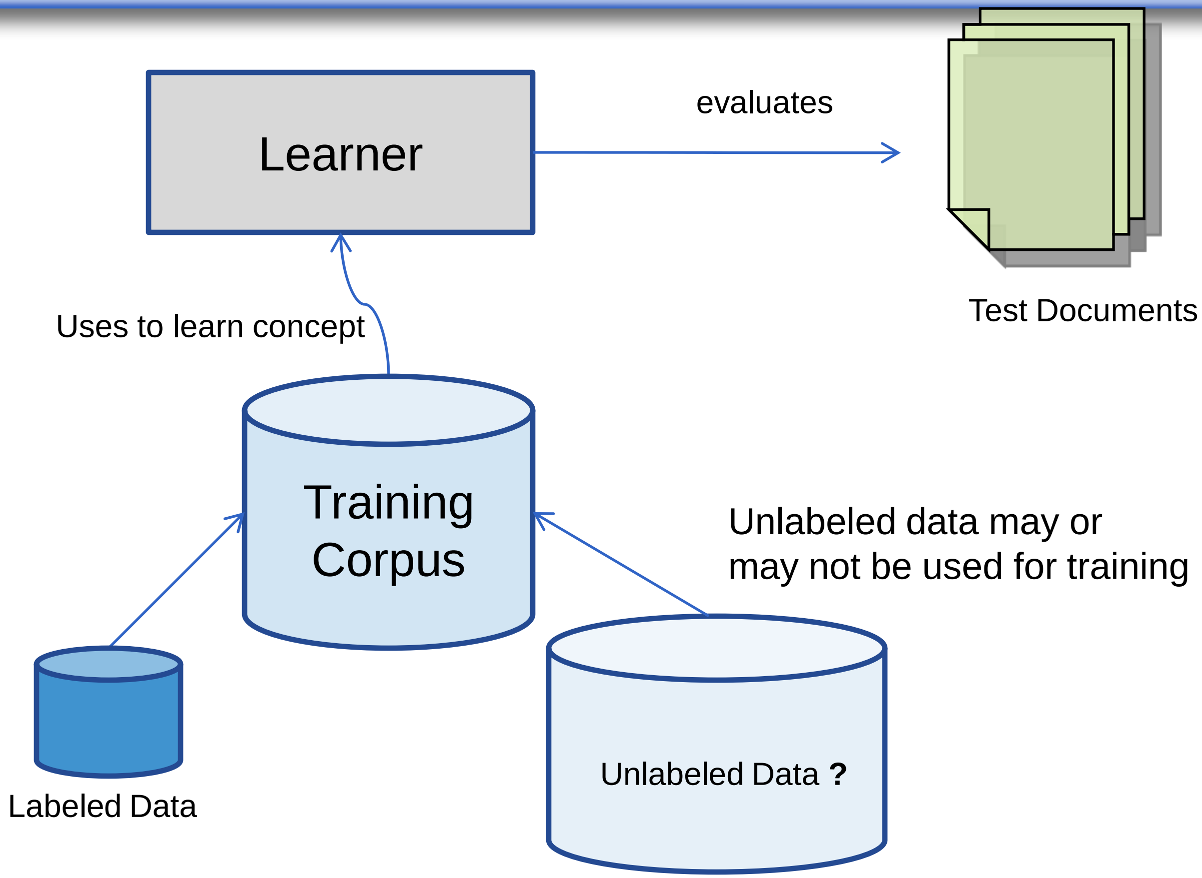
Parsing is harder with long and ambigous sentences .



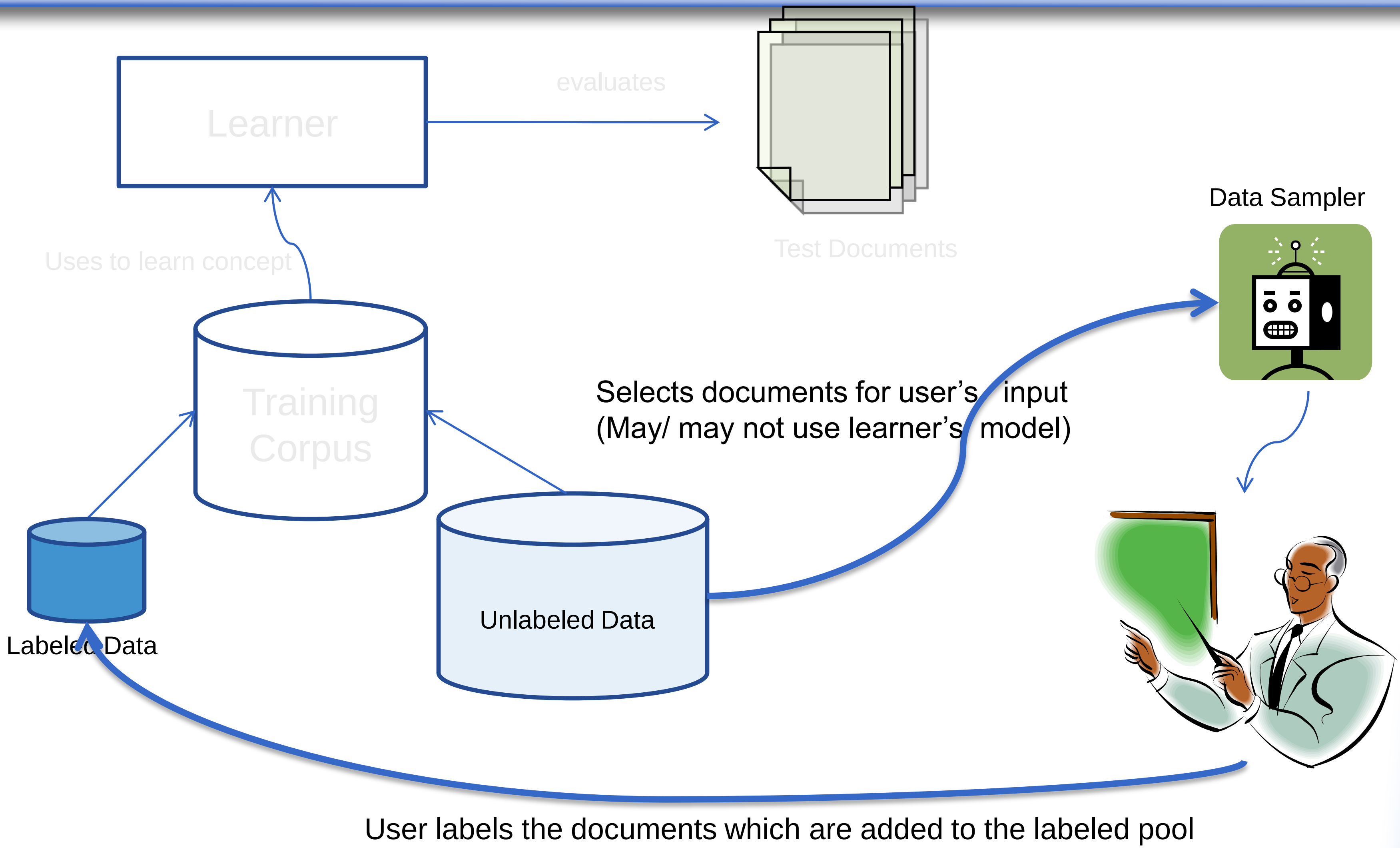


(Parses from: http://www.link.cs.cmu.edu/link/submit-sentence-4.html)

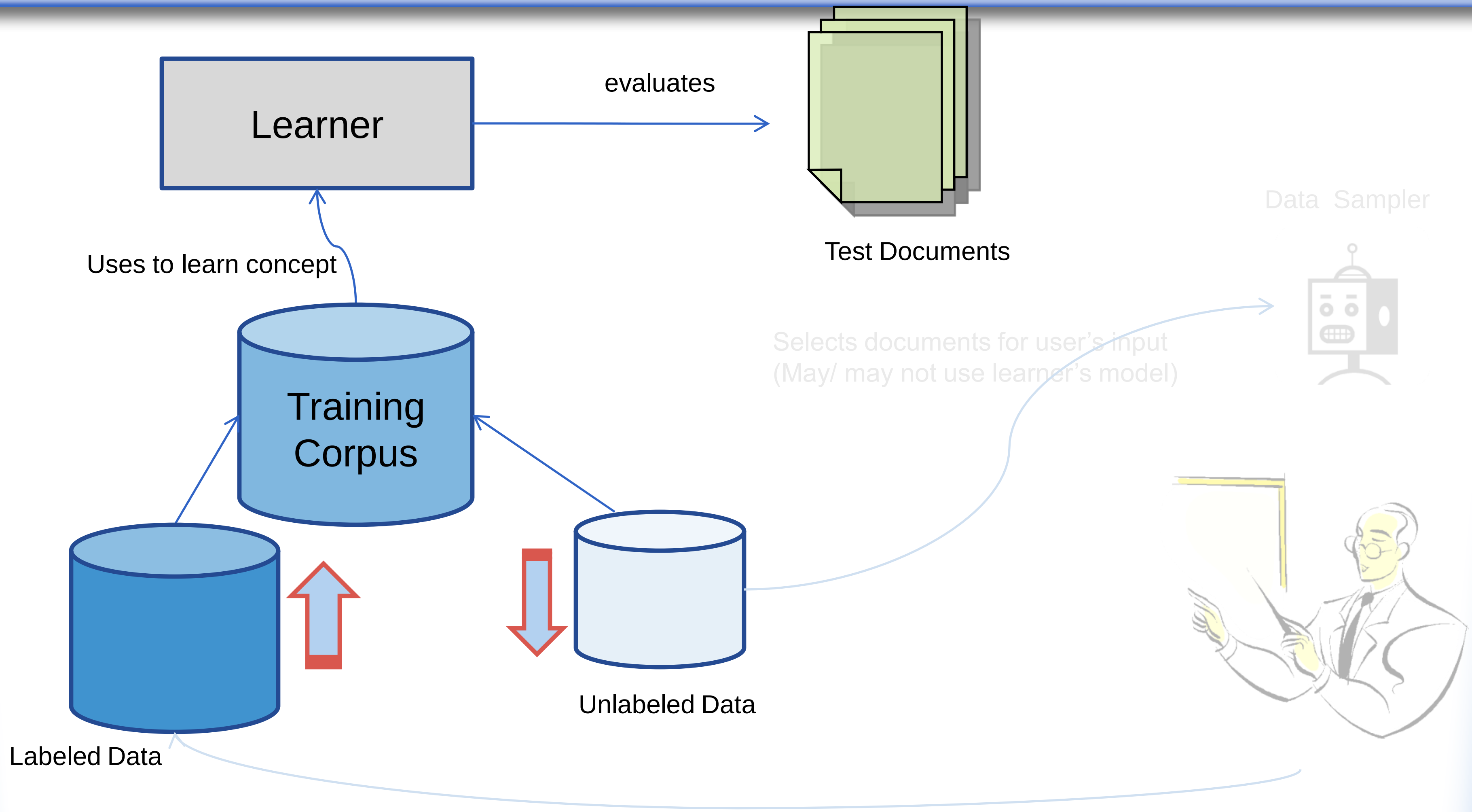
Active Learning Process



Active Learning Process



Active Learning Process



Evaluation Measures

- Accuracy Vs. Number of training examples

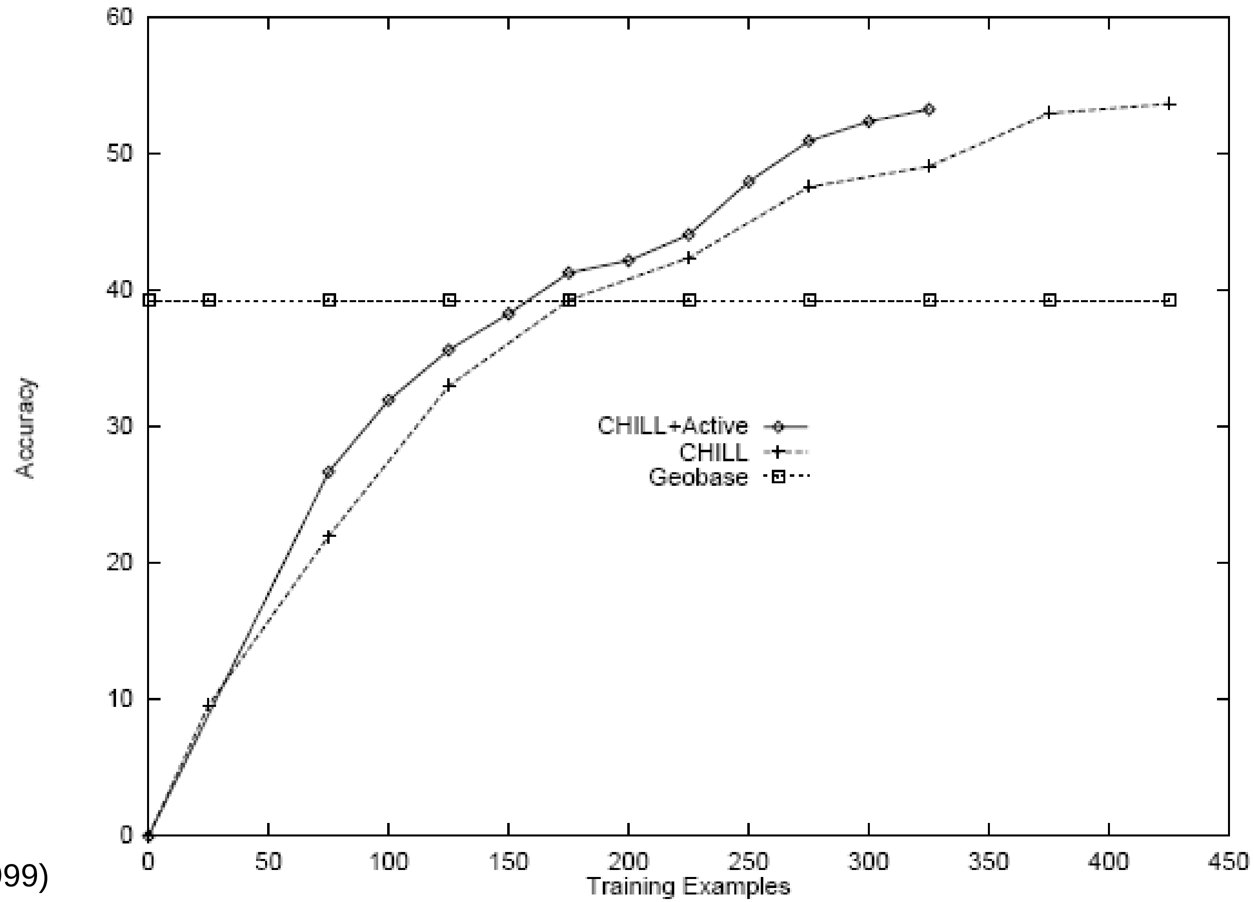


Figure from (Thompson et al., 1999)

Evaluation Measures

- Accuracy Vs. Number of training examples

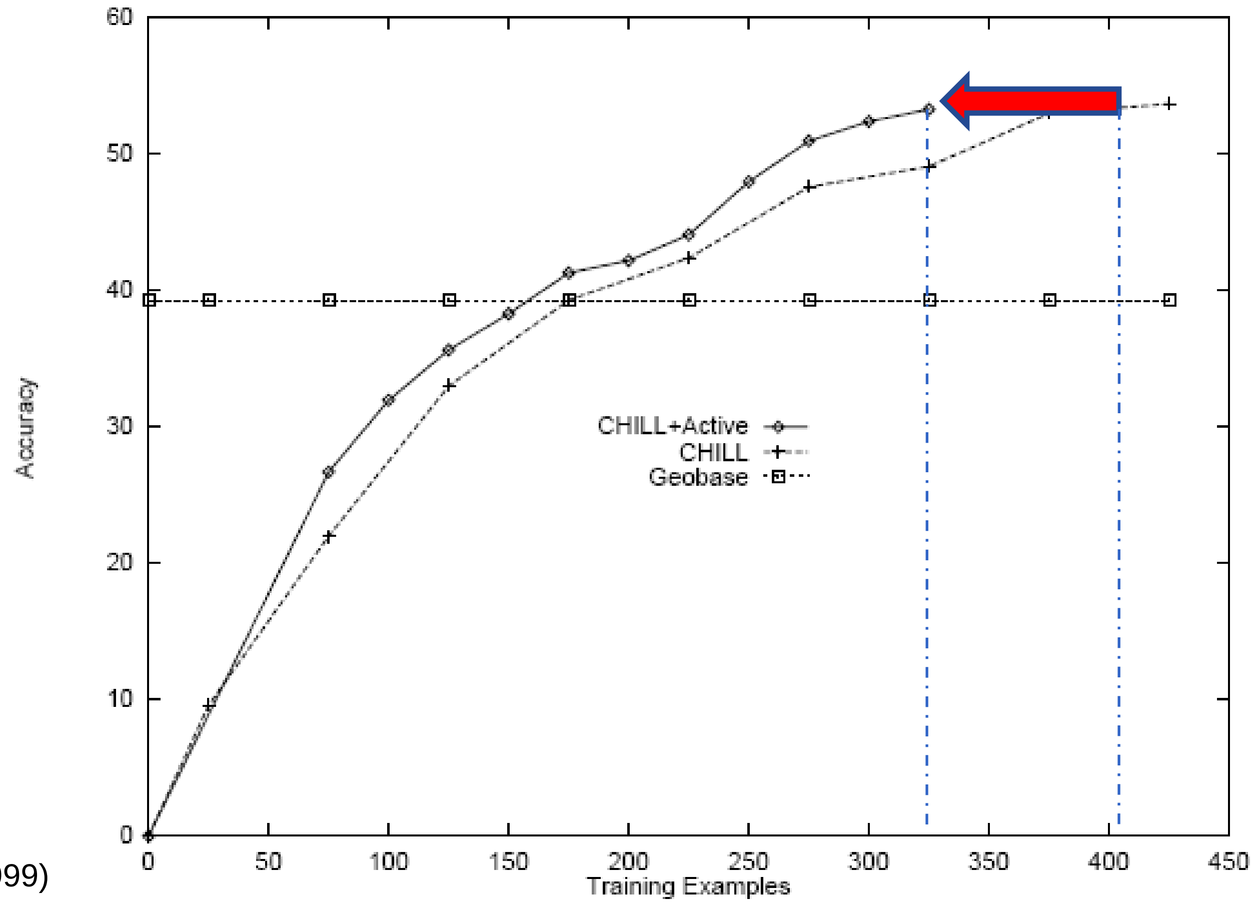


Figure from (Thompson et al., 1999)

Evaluation Measures

How do we measure user effort?

(Kristjannson et. al., 2004)

Evaluation Measures

How do we measure user effort?

Number of examples user has to correct?

(Kristjannson et. al., 2004)

Evaluation Measures

How do we measure user effort?

Number of examples user has to correct?

OR

Number of corrections user has to make?

(Kristjannson et. al., 2004)

Evaluation Measures

- Expected Number of User Actions (ENUA)
  - Number of User Actions, such as clicks, required to correctly label all the fields (Kristjannson et. al., 2004)
  - ENUA doesn't distinguish between *boundary detection* and *classification*
  - *Culotta and McCallum, (2005)* define 4 types of user actions: *Start, End, Type and Choose*

Evaluation Measures

- Expected Number of User Actions (ENUA)
  - Number of User Actions, such as clicks, required to correctly label all the fields (Kristjannson et. al., 2004)
  - ENUA doesn't distinguish between *boundary detection* and *classification*
  - *Culotta and McCallum, (2005)* define 4 types of user actions: *Start, End, Type and Choose*

What about effort in reading the text ?

Evaluation Measures

- Rebecca Hwa (2000), user effort in parsing:
  - *Number of brackets user adds* instead of number of sentences user has to annotate

Selective Sampling

- Active learning aims at reducing the number of labeled examples required to learn the target concept by *selectively sampling* from the unlabeled data for user's input

Selective Sampling

- Active learning aims at reducing the number of labeled examples required to learn the target concept by *selectively sampling* from the unlabeled data for user's input
- Strategies
  - Uncertainty-based
  - Query-by-committee

Selective Sampling

- Active learning aims at reducing the number of labeled examples required to learn the target concept by *selectively sampling* from the unlabeled data for user's input
- Strategies
  - Uncertainty-based
  - Query-by-committee

Uncertainty-based

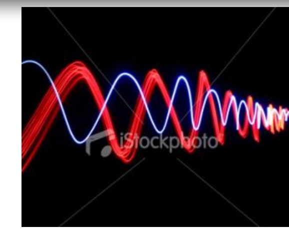
- Examples the learner is *least certain* about are presented to the user
  - *Interactive Information Extraction* (Kristjannson et al., 2004)
  - *Semantic Role Labeling* (Roth and Small, 2006)
  - *Grammar Learning* (Hwa, 2000)
  - *Online Learning for Spam Filtering* (Sculley, 2007)
  - *Parsing & Rule-based IE* (Thompson et al., 1999)

Interactive Information Extraction

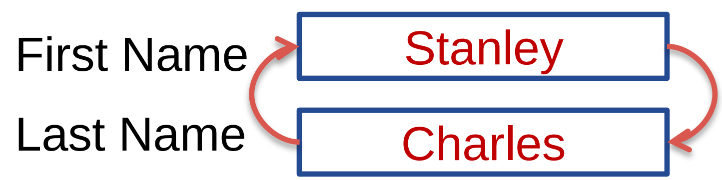
- Extracting contact addresses from web pages & emails
- Interface for users to make corrections
- CRFs with Viterbi algorithm for finding the most likely state sequence given the observation sequence

(Kristjannson et al., 2004)

Interactive Information Extraction

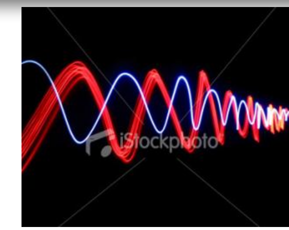


- Correction Propagation: *A correction propagates & corrects more fields*
  - *Constraints (Corrections)* can *affect* the optimal paths *before* and *after* the time steps specified in the constraint & this may *help* in correcting other fields
  - Constrained Viterbi



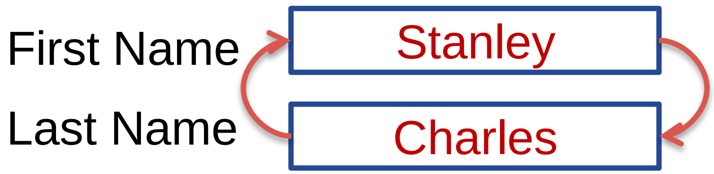
(Kristjannson et al., 2004)

Interactive Information Extraction



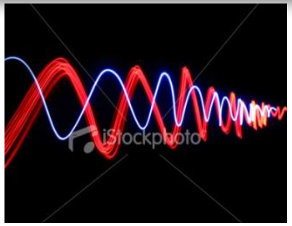
- Correction Propagation: *A correction propagates & corrects more fields*
  - *Constraints (Corrections)* can *affect* the optimal paths *before* and *after* the time steps specified in the constraint & this may *help* in correcting other fields

Correct the field that would result in most correction propagation ?



(Kristjannson et al., 2004)

Interactive Information Extraction



- Correction Propagation: *A correction propagates & corrects more fields*
  - *Constraints (Corrections)* can *affect* the optimal paths *before* and *after* the time steps specified in the constraint & this may *help* in correcting other fields

After how many corrections should we propagate ?

First Name | Stanley

Last Name | Charles

(Kristjannson et al., 2004)

Interactive Information Extraction

- Uncertainty-based Recommendation

How do we calculate uncertainty or confidence a learner has in its prediction?

Interactive Information Extraction

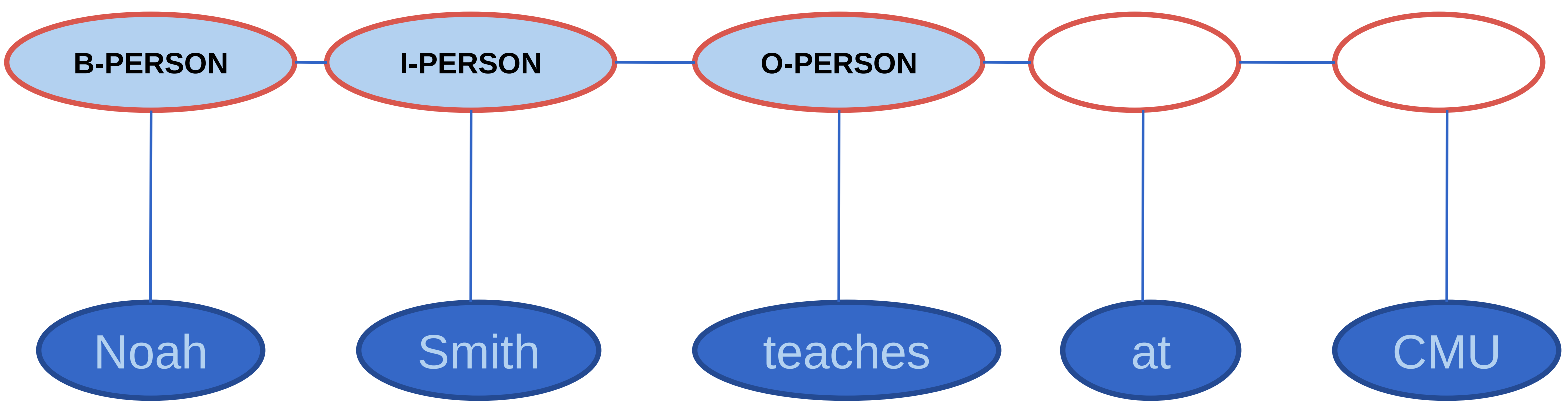
- Confidence estimation:
  - *How confident we are that Noah Smith is a person ?*

(Kristjannson et al., 2004)

Interactive Information Extraction

- Confidence estimation:
  - *How confident we are that Noah Smith is a person ?*

Constrained Forward Backward



(Kristjannson et al., 2004)

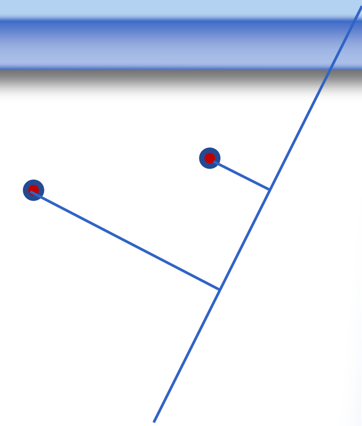
Savings from Active Learning

Interactive Information Extraction (Kristjannson et al., 2004):

- DataSet - 2187 web & email records, 25 classes
- Reduction in ENUA - 11.3%

(Kristjannson et al., 2004)

Margin-based classifiers



- Perceptron for Structured Output
- Certainty = Distance from hyperplane
- *Least* certainty = *Smallest* margin
- Multiclass
  - Margin between predicted label and 2nd highest activation value
- Global Vs Local Margin
  - Local margin - select examples with a small average local multi-class margin

(Roth and Small, 2006)

Quering Partial Labels

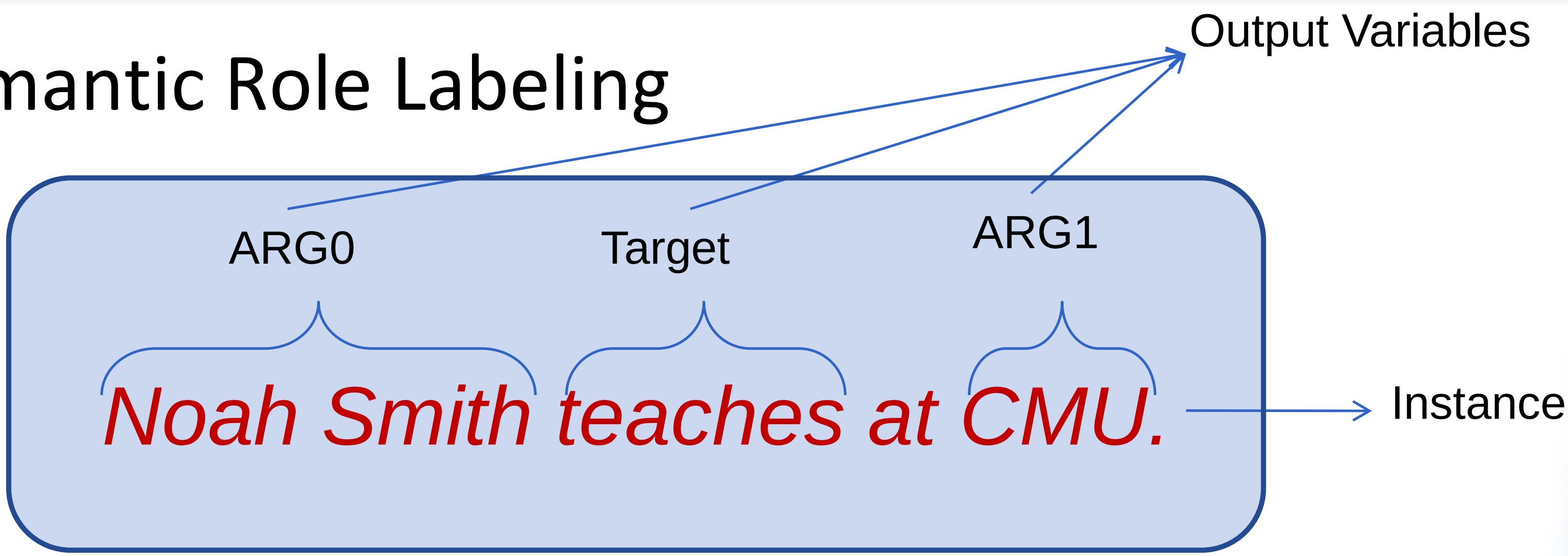
- Semantic Role Labeling

$$\overbrace{\textit{Noah Smith}}^{\text{ARG0}}\ \overbrace{\textit{teaches}}^{\text{Target}}\ \textit{at}\ \overbrace{\textit{CMU.}}^{\text{ARG1}}$$

(Roth and Small, 2006)

Quering Partial Labels

- Semantic Role Labeling



- All *output variables* in an *instance* are not equally informative
- Reduces output space for remaining local variables => similar to *Correction Propagation*

(Roth and Small, 2006)

Savings from Active Learning

Semantic Role Labeling (Roth and Small, 2006)

- DataSet - CoNLL-2004 shared task
- Complete label queries - 35% fewer examples
- Partial label queries - 50% fewer examples

Grammar Learning

- Inferring grammatical structure of a language from examples
- Variant of inside-outside algorithm to learn Probabilistic Lexicalized Tree Insertion Grammar (Hwa, 1998)
- Selective sampling to minimize the user annotation effort

(Rebecca Hwa, 2000)

Grammar Learning

- Select examples with high *Training Utility Value (TUV):*
  - Sentence length
    - Longer sentences -> complex & ambiguous
  - Tree entropy of the sentence
    - Classifier's distribution over all possible parse trees
    - Uniform distribution => higher entropy => higher uncertainty

(Rebecca Hwa, 2000)

Savings from Active Learning

Grammar Learning (Hwa, 2000)

- DataSet - WSJ Corpus: Penn Treebank
- Tree-entropy based – 36% fewer annotations (# of brackets added)
- Length based – 9% fewer annotations

Online Learning

- E.g., Spam filtering
- Online Active Learning
  - Messages come in a *stream*
  - Decision to recommend has to be made in *real time*
  - Pool-based Active Learning is *expensive*

Pool-based learning

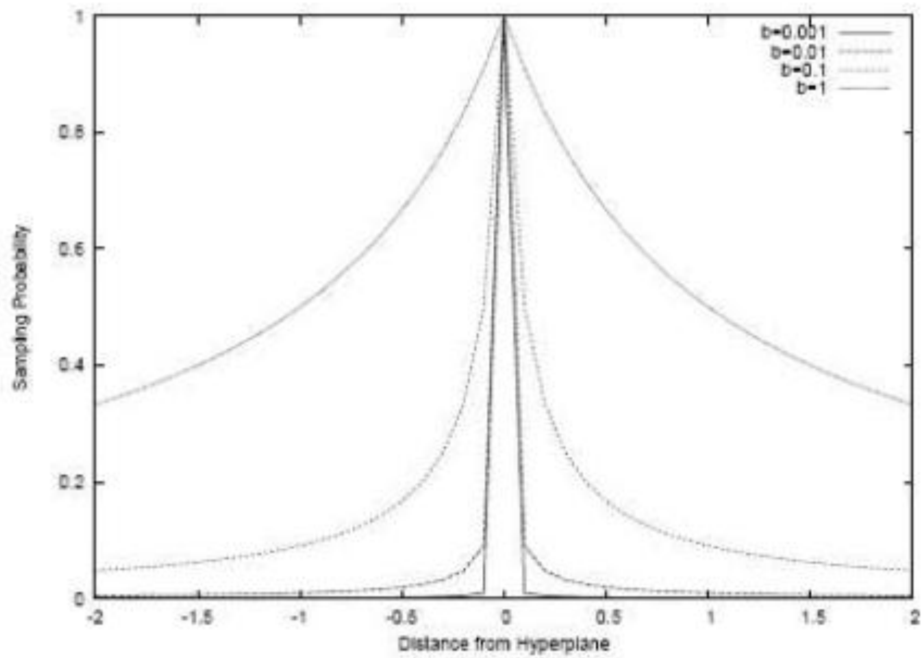
Online learning

(D. Sculley, 2007)

Online Learning

- Sampling probability:

$$P_i = \frac{b}{b + |p_i|}$$



b= Sampling parameter

$|p_i|$ = distance from hyperplane or classification confidence

(D. Sculley, 2007)

Savings from Active Learning

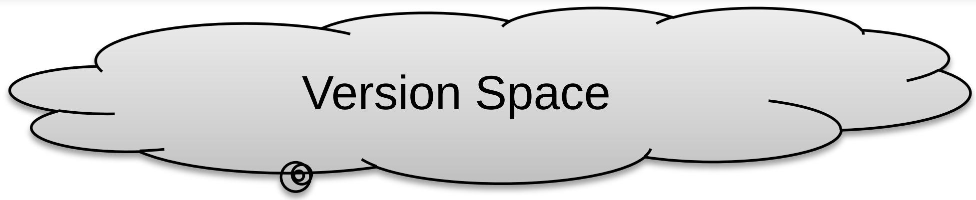
Online Learning for Spam Filtering (Sculley, 2007)

- DataSet – TREC 05 & 06
- Requires only 10% of examples required by uniform sampling

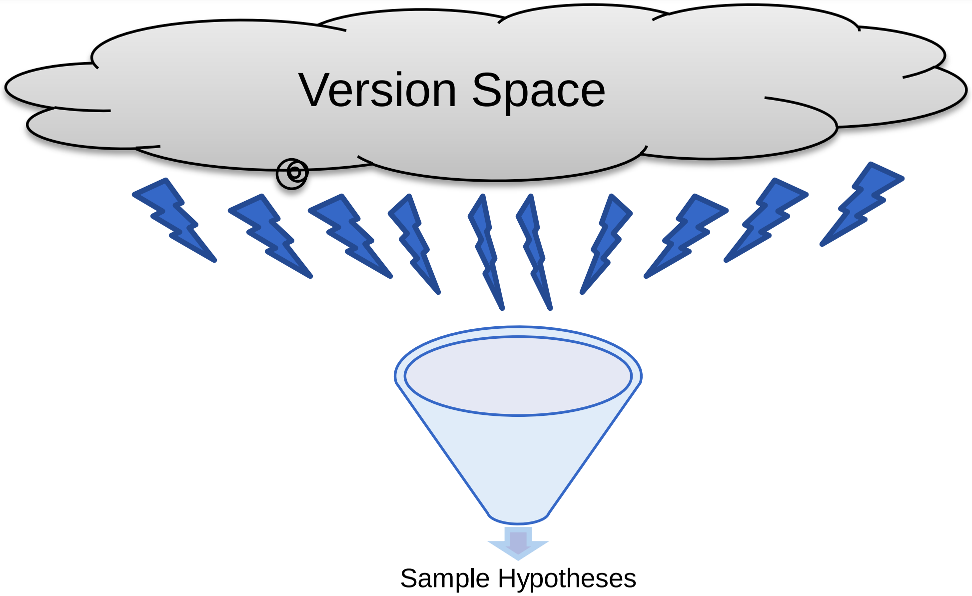
Query-by-committee

- Active learning aims at reducing the number of examples required to learn the target concept by *selectively sampling* from the unlabeled data
- Strategies
  - Uncertainty-based
  - Query-by-committee

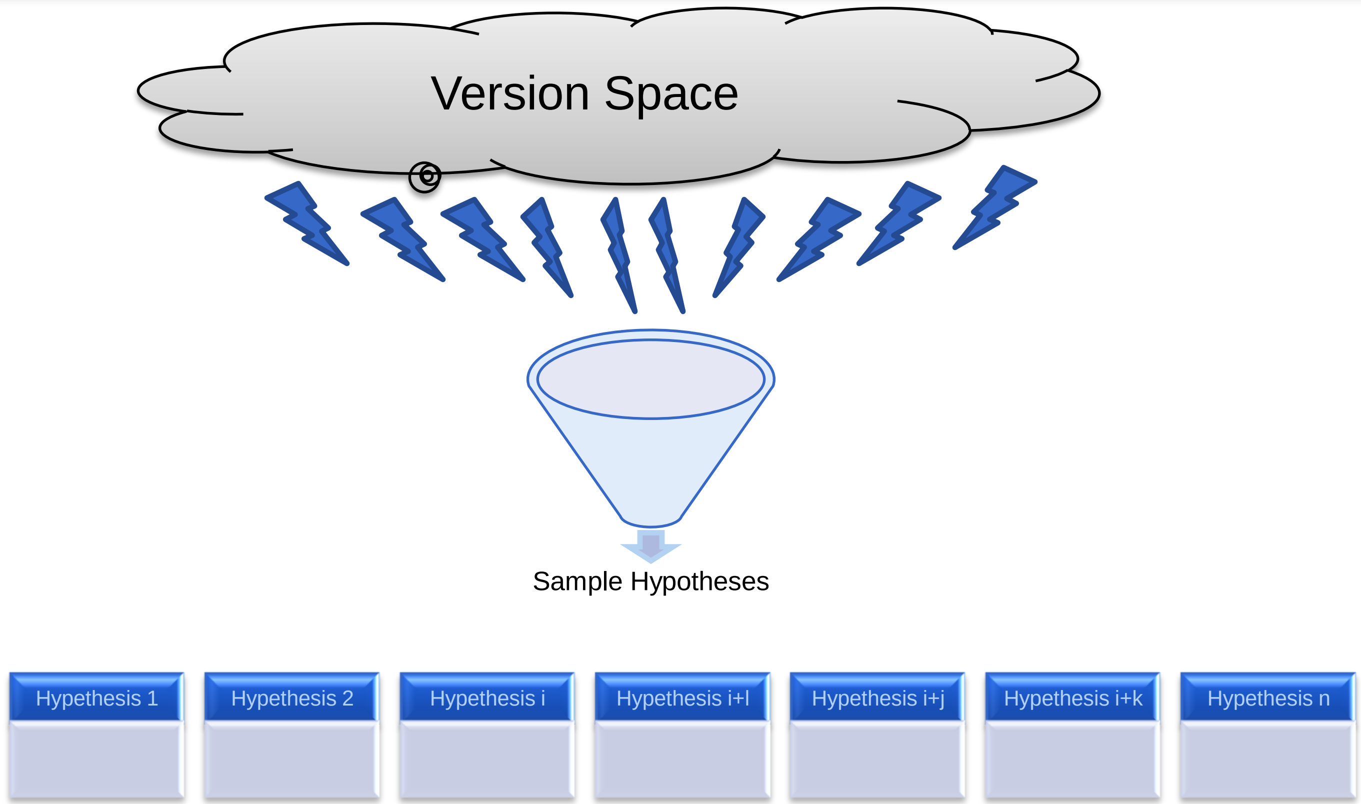
Query-by-Committee



Query-by-Committee

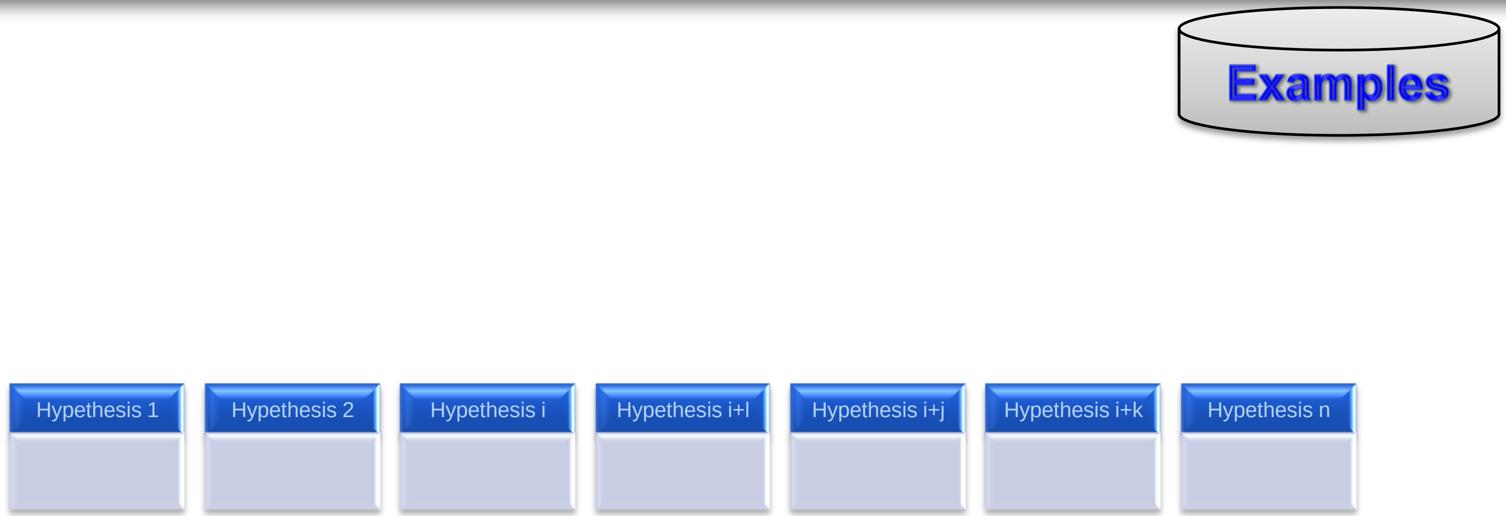


Query-by-Committee

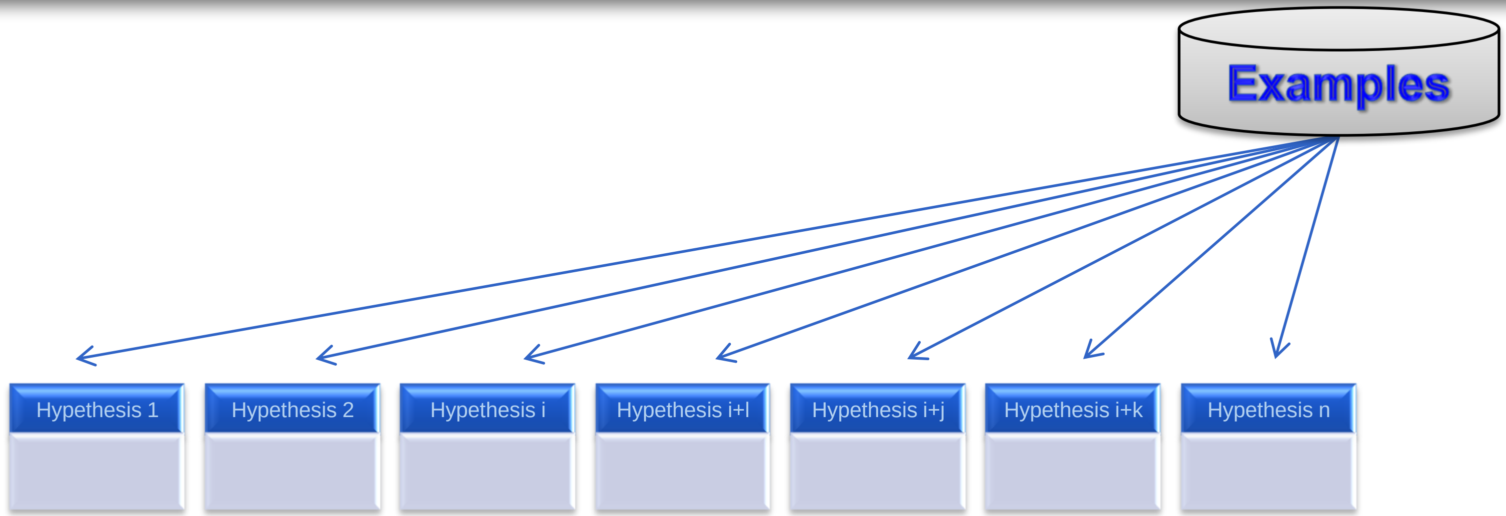


Committee of ‘n’ hypotheses

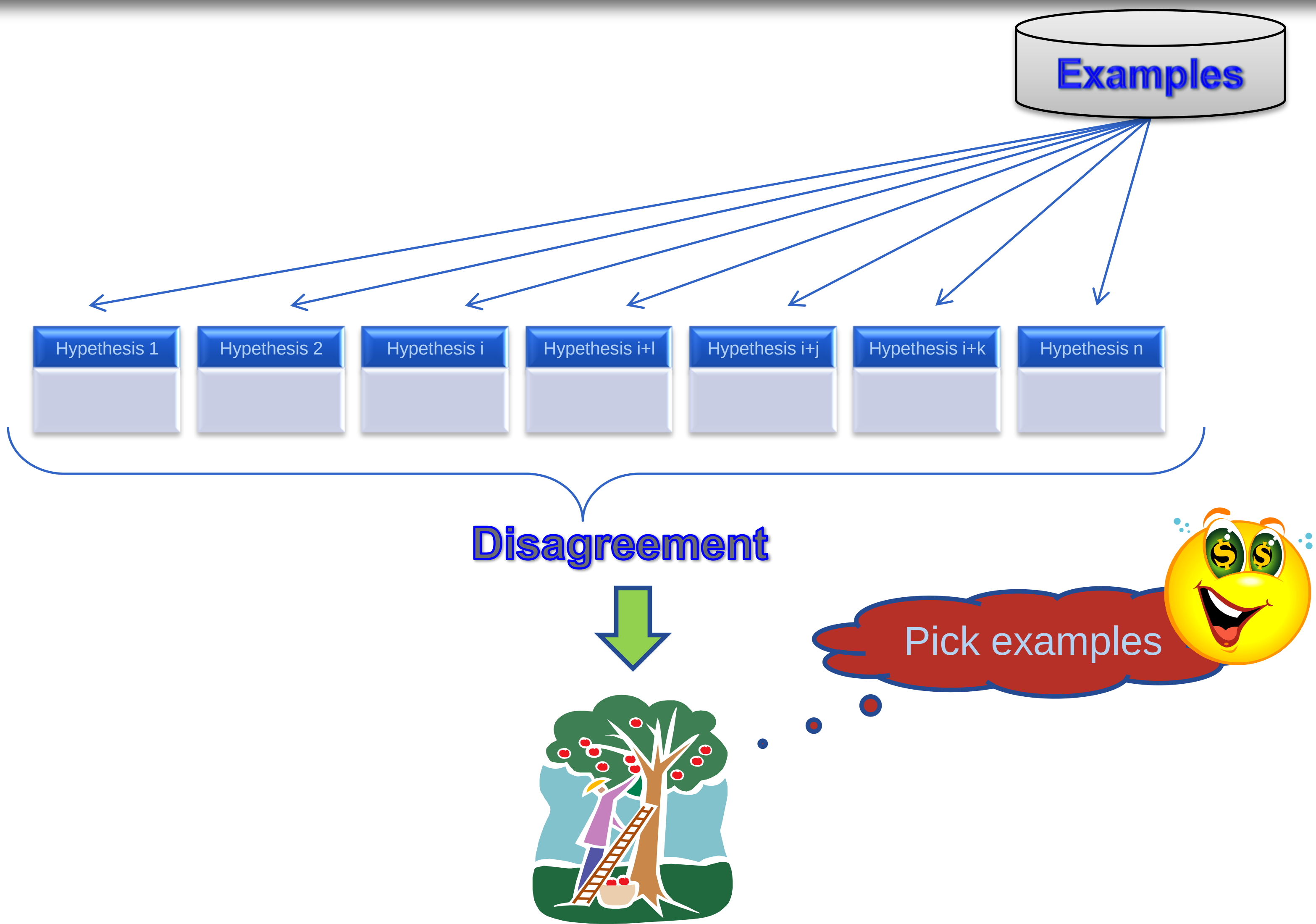
Query-by-Committee



Query-by-Committee



Query-by-Committee



Query-by-Committee

- Research covered in the literature review
  - *Semi-supervised learning using EM* (McCallum and Nigam, 1998)
  - *Multi-view active learning* (Muslea et al., 2006)
  - *Bootstrapping Statistical Parsers* (Steedman et al. 2003)

QBC Semi-supervised Learning using EM

- McCallum and Nigam, 1998
  - Combine QBC based active learning with EM
  - Use Naïve Bayes classifier for text classification
  - Committee of ‘k’ classfiers
    - Sample parameters using Gamma distribution ‘k’ times to create a committee of ‘k’ classifiers
    - Parameters of Gamma distribution depend upon the word and class counts in training data

QBC Semi-supervised Learning using EM

- Metrics for committee disagreement
  - Vote Entropy:
    - Each member votes for its winning class,
    - Vote Entropy = entropy of vote distribution
    - Does not consider confidence of classifier
  - KL divergence to the mean: Average of KL divergence between each member's class distribution and mean of all distributions $\frac{1}{k}\sum_{m=1}^{k} D(P_m(C \mid d_i) \| P_{avg}(C \mid d_i))$

  where $P_{avg}(C \mid d_i) = \frac{1}{k}\sum_{m} P_m(C \mid d_i)$

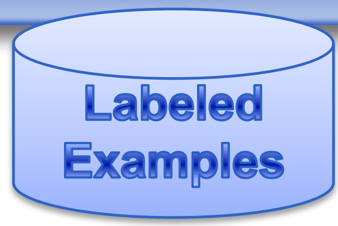
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM

- Document selection criteria
  - Stream-based
    - Decision to label is made on each document individually, irrespective of alternatives
  - Pool-based
    - Select from all documents in the pool which has largest disagreement
  - Density-weighted pool-based
    - Combine the similarity and disagreement measure

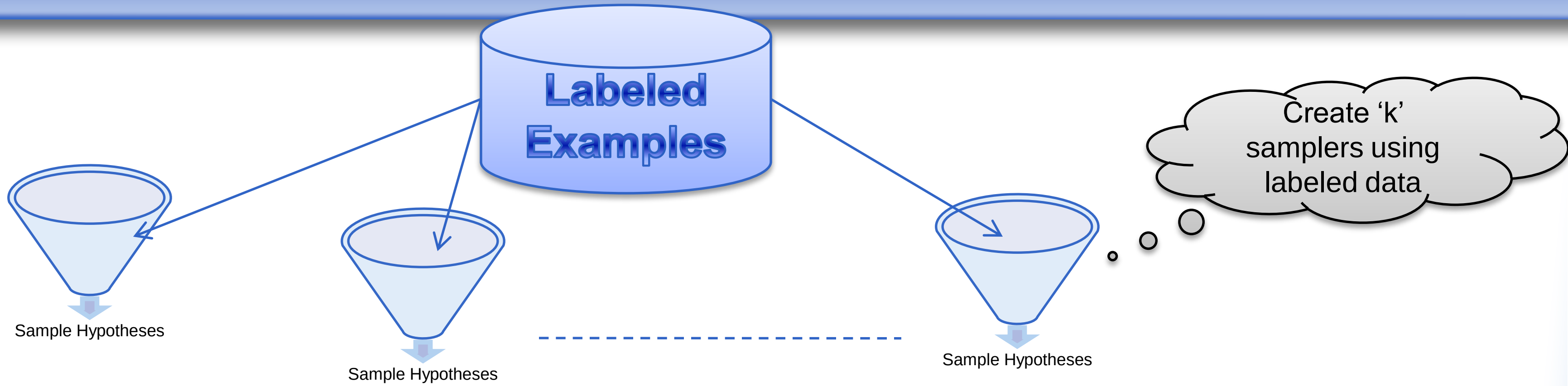
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



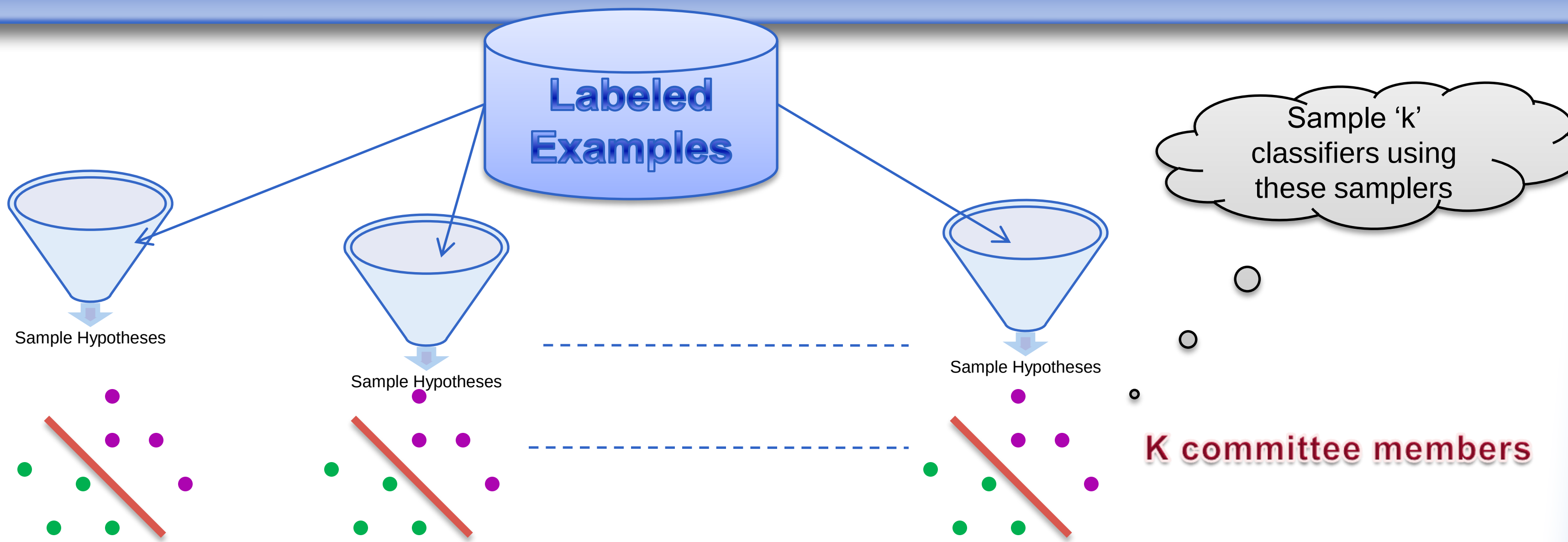
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



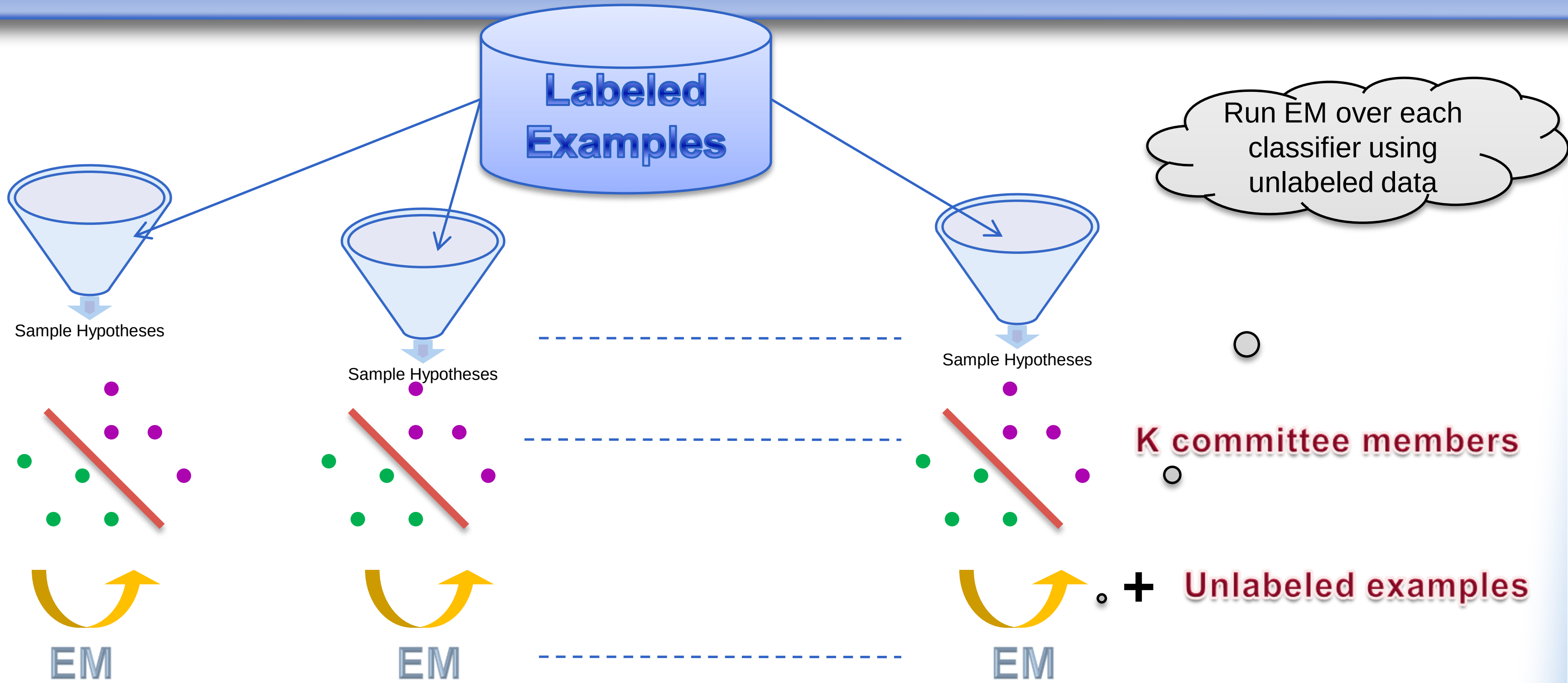
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



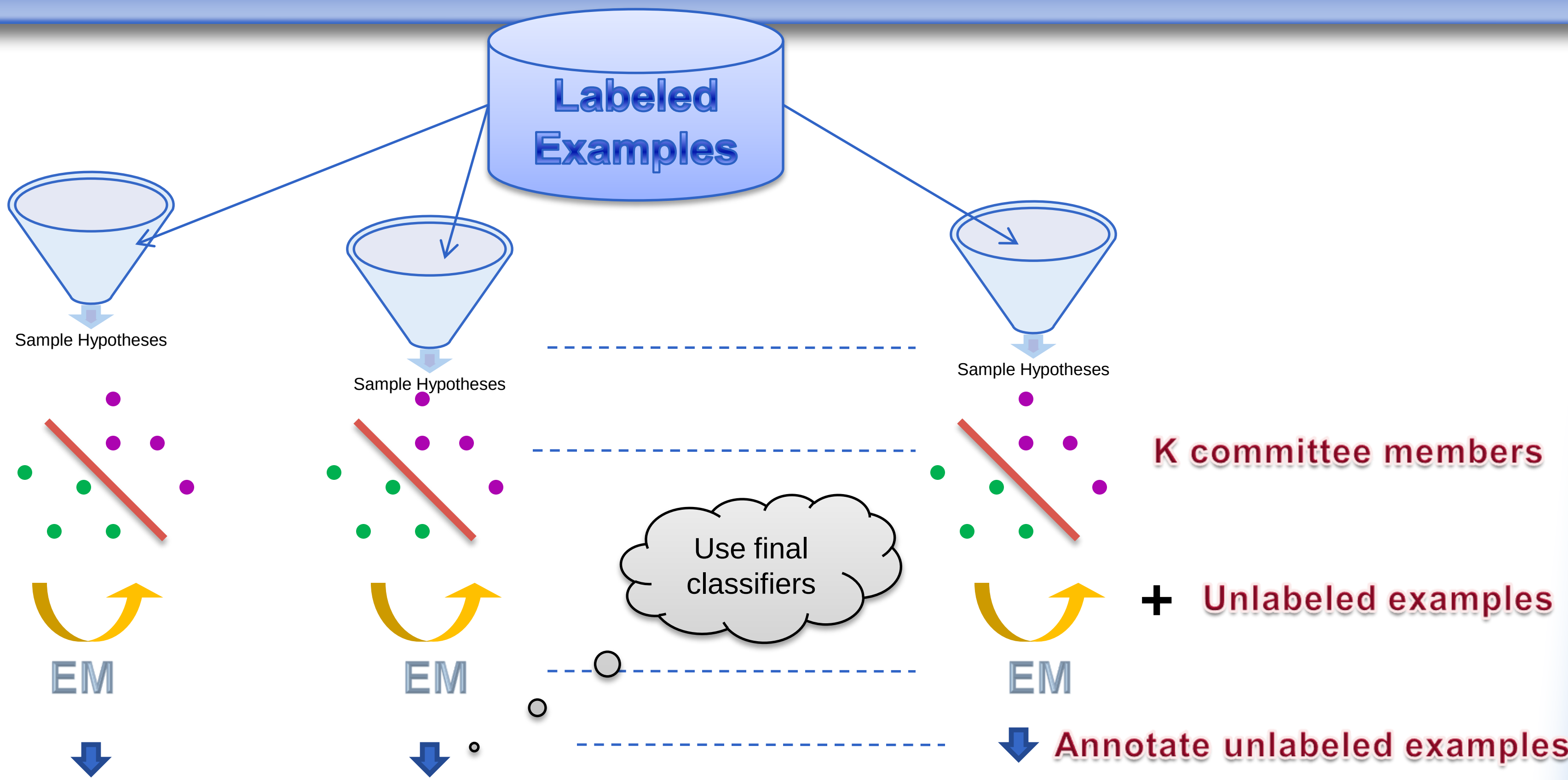
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



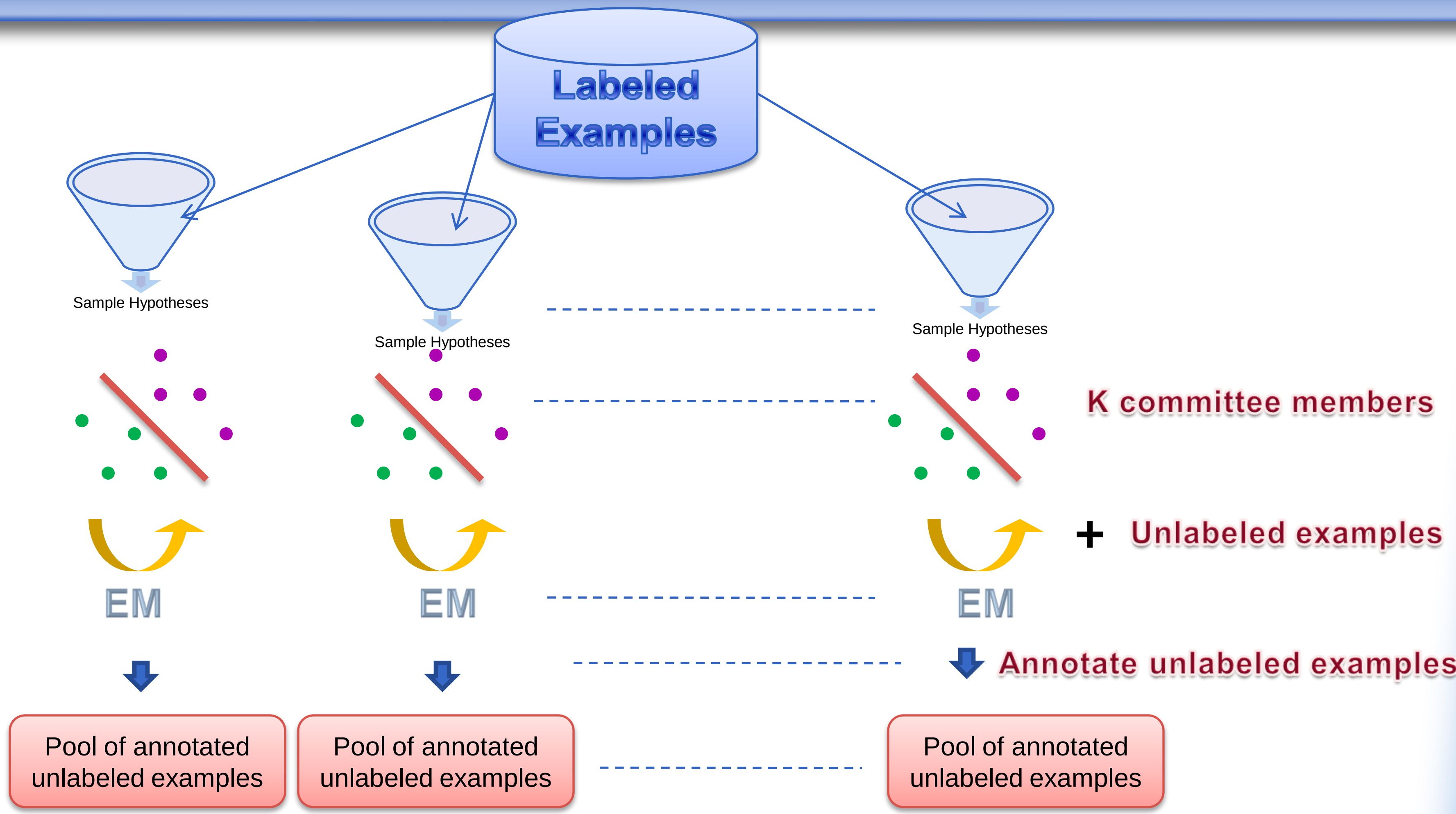
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



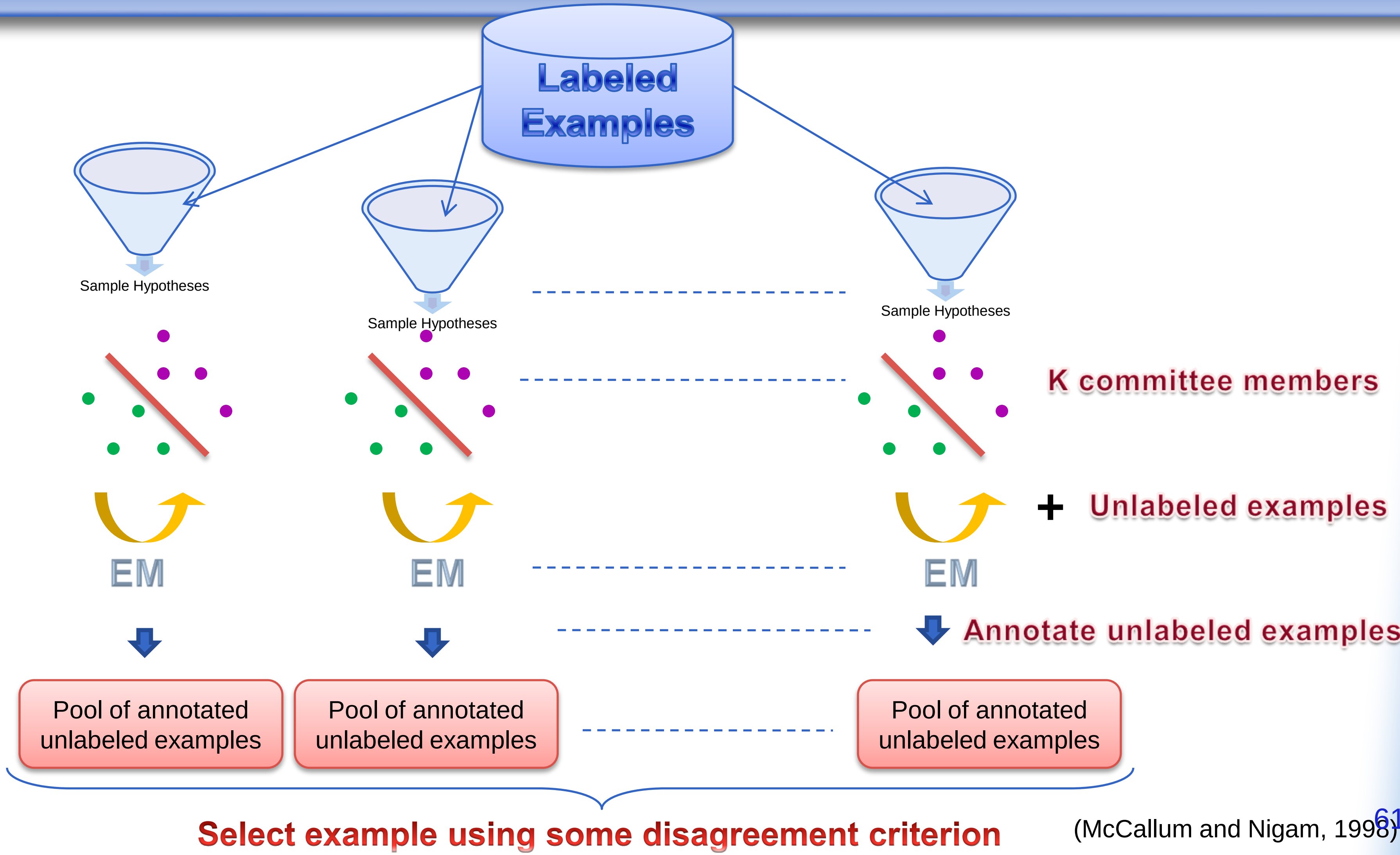
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



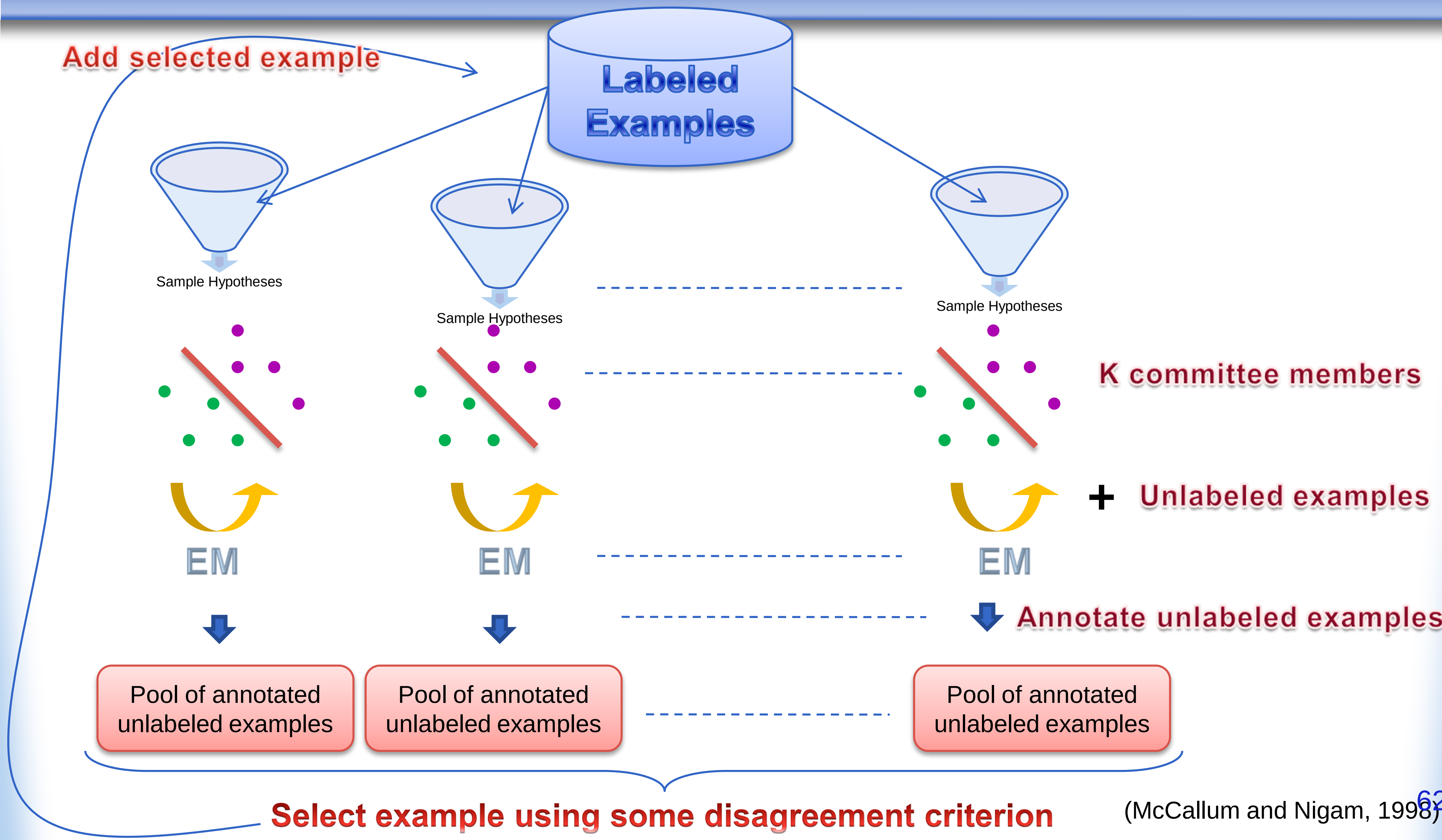
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



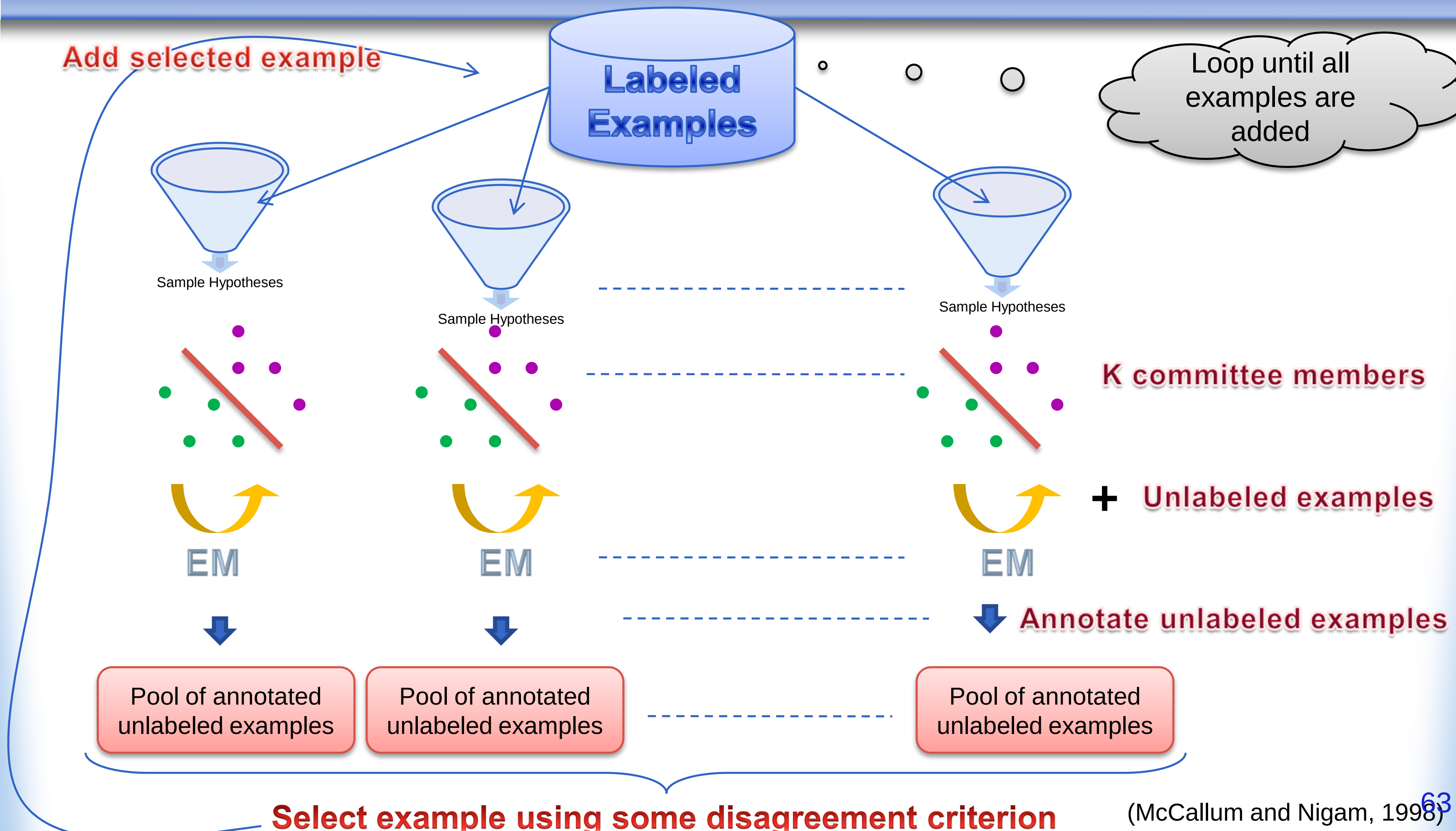
(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



(McCallum and Nigam, 1998)

QBC Semi-supervised Learning using EM



(McCallum and Nigam, 1998)

Savings from Active Learning

- Results
  - Usenet and Reuters data for experiments
  - Algorithm requires 32 labeled documents for achieving an accuracy of 64% as compared to 59 labeled documents for random sampling.

(McCallum and Nigam, 1998)

Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept

Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept

Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept

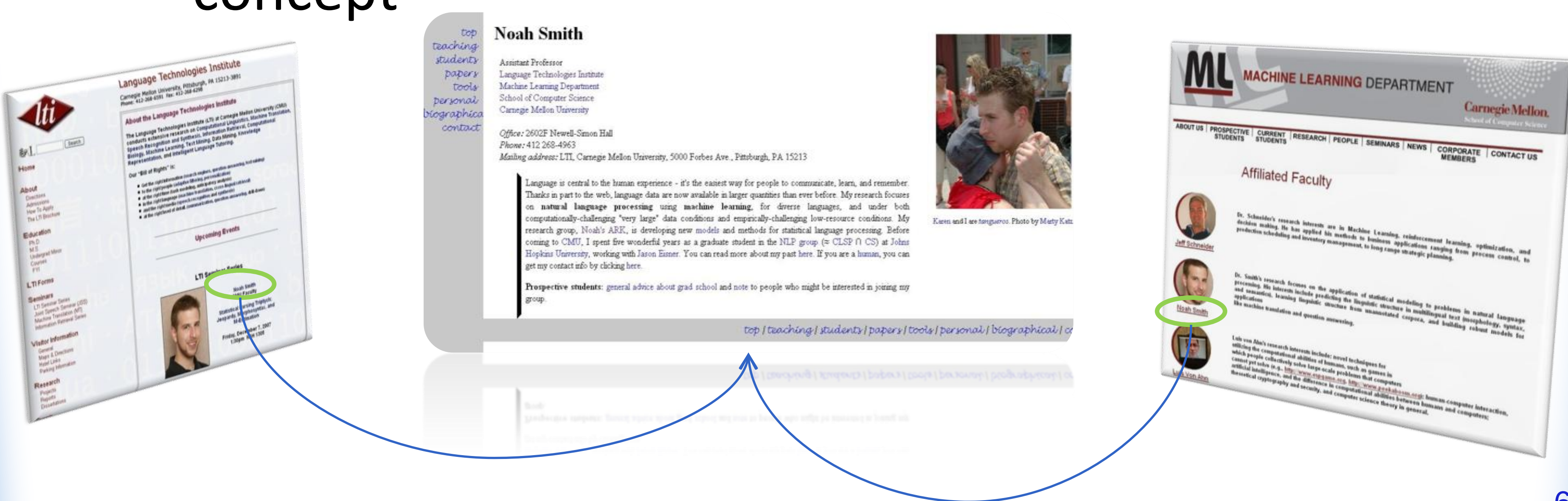
Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept

Words in document as features

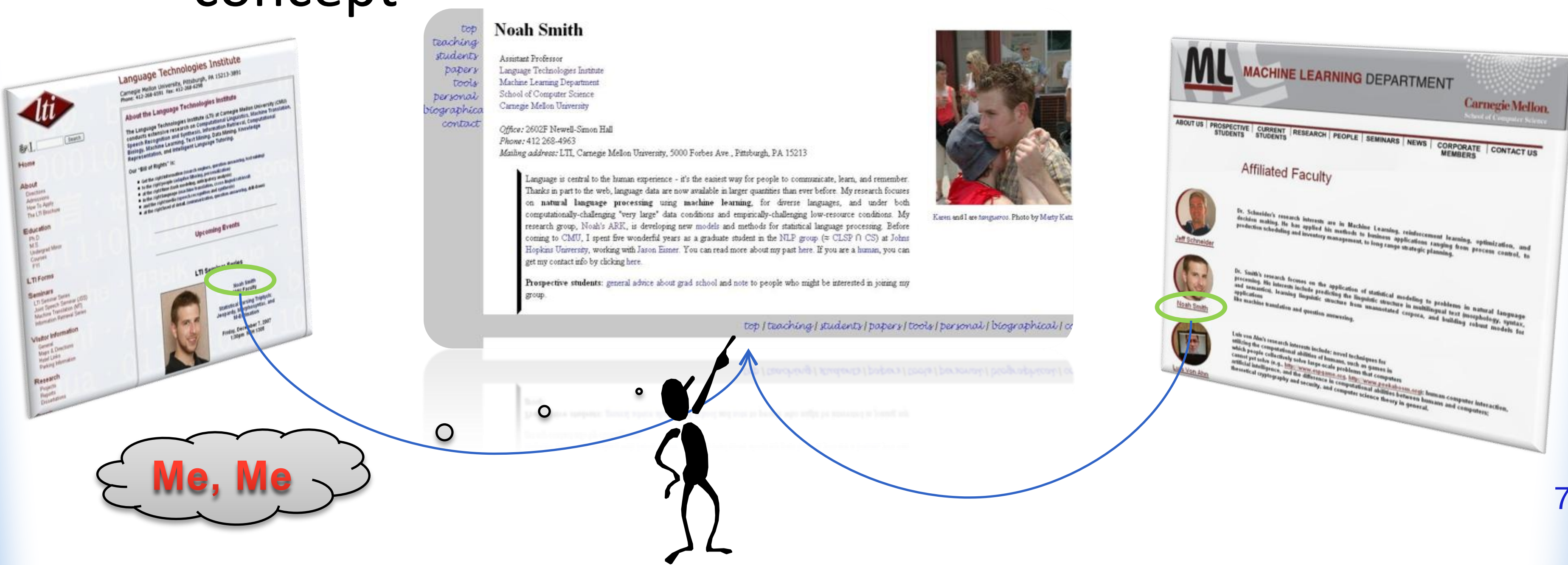
Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept



Multi-view Active Learning

- Multiple views
  - Disjoint sets of features
  - Each of the sets sufficient to learn the target concept



Multi-view Active Learning

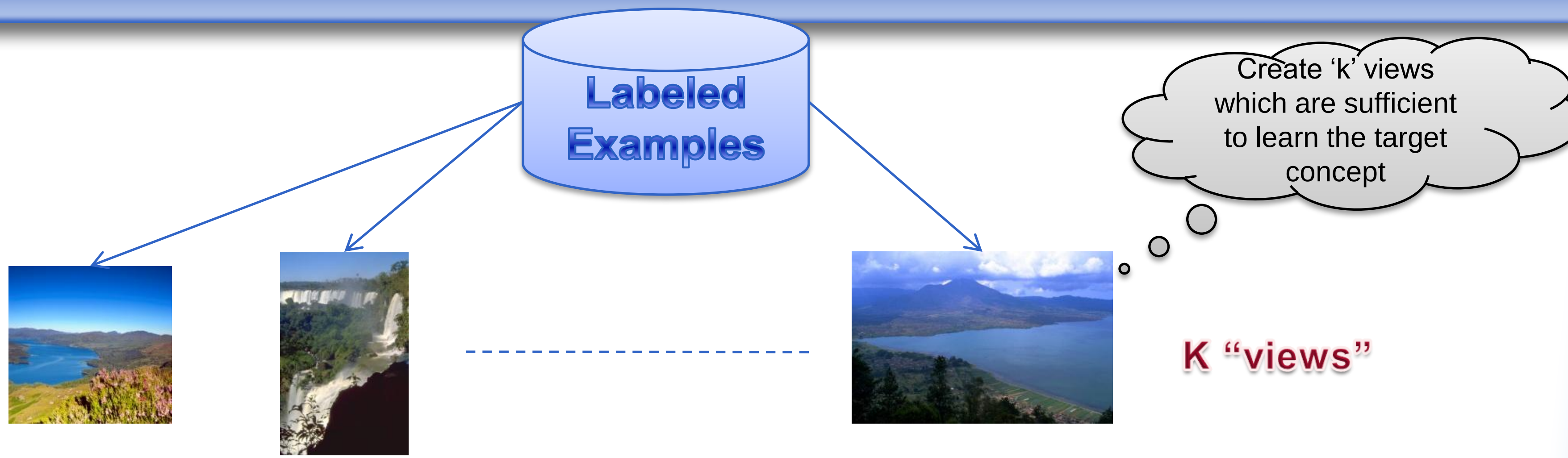
- Co-Testing
  - A family of active learners for multi-view learning tasks.
  - Two step iterative algorithm
  - Requires as input a few labeled and many unlabeled examples.

(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing

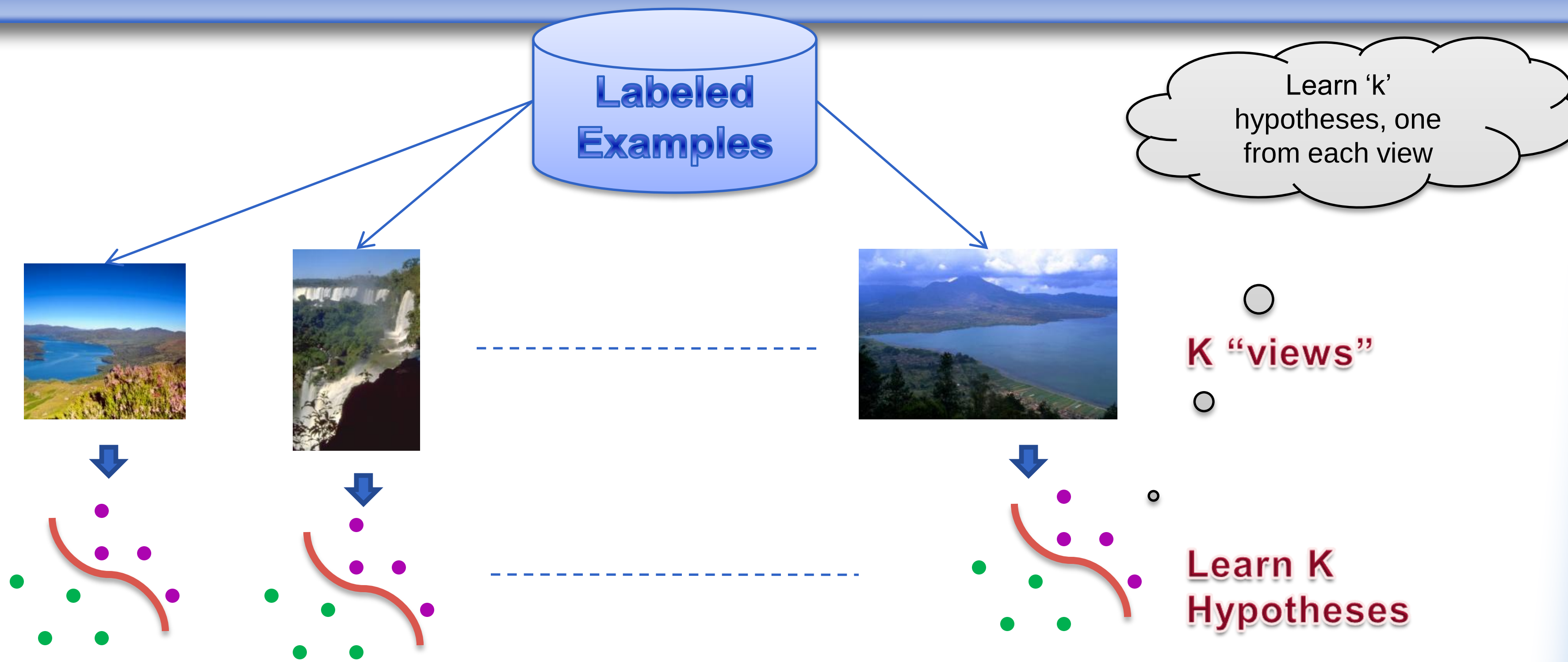
(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing



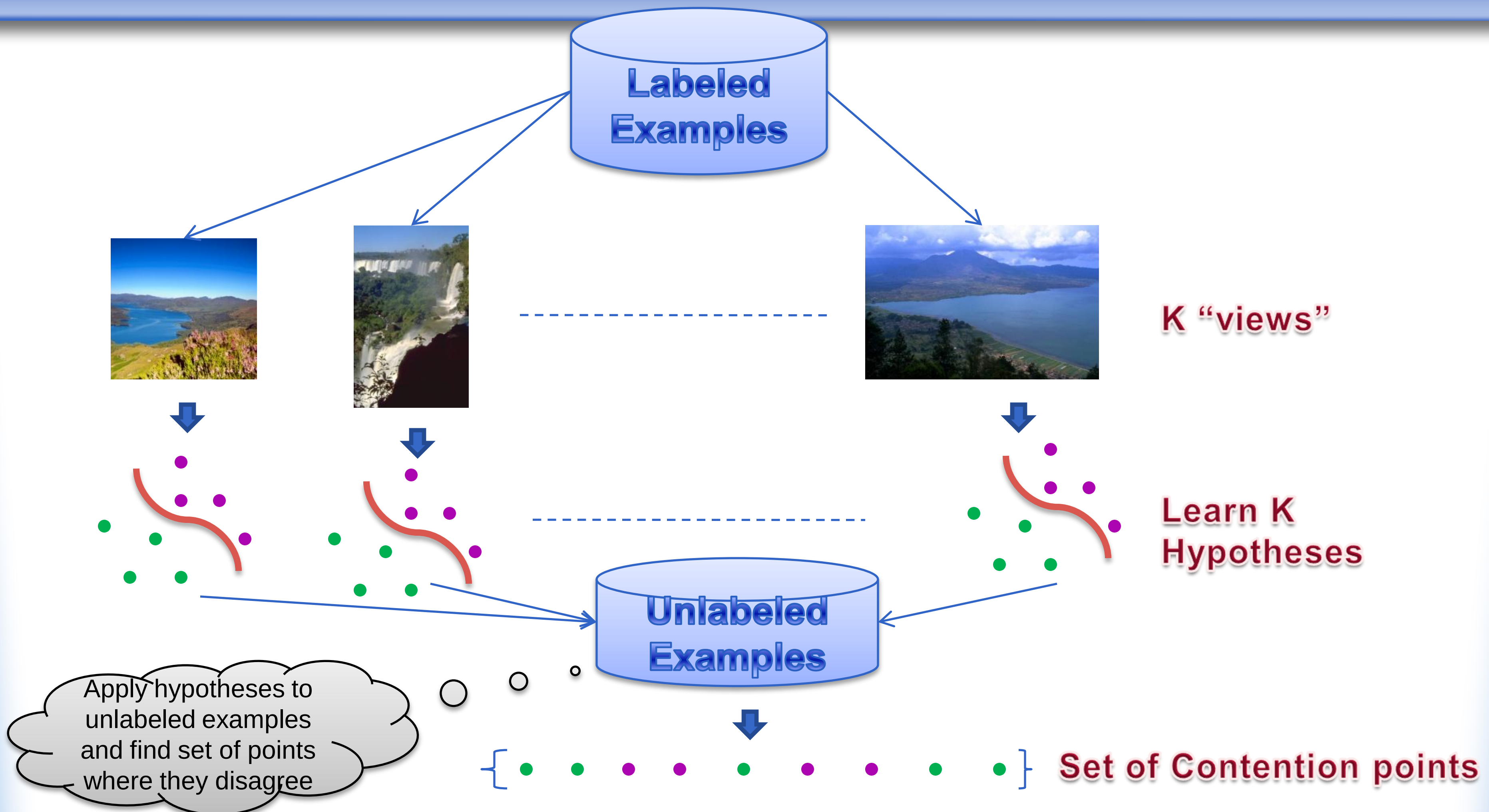
(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing



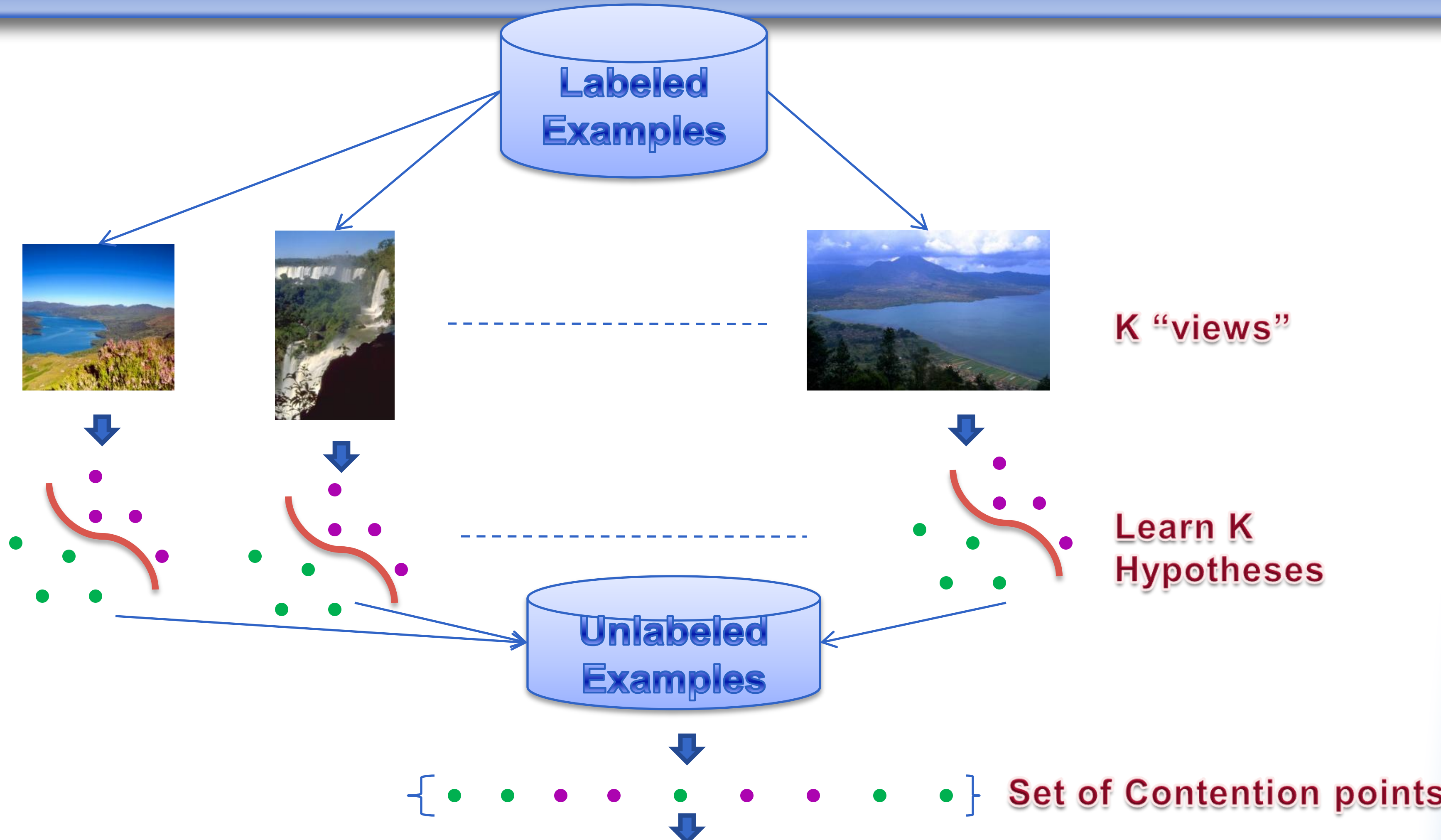
(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing



(Muslea et al., 2006)

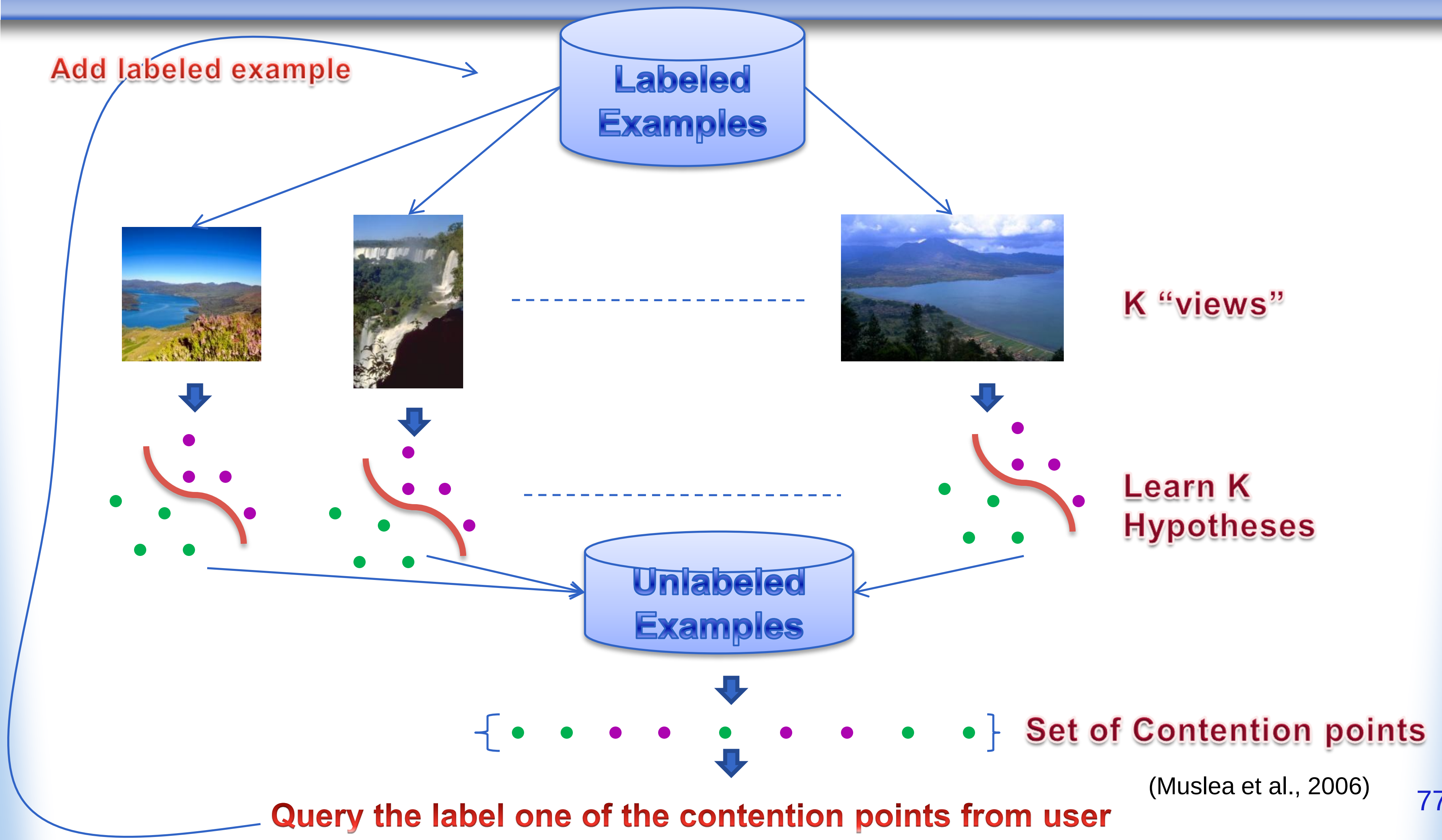
Multi-view Active Learning
Co-Testing



(Muslea et al., 2006)

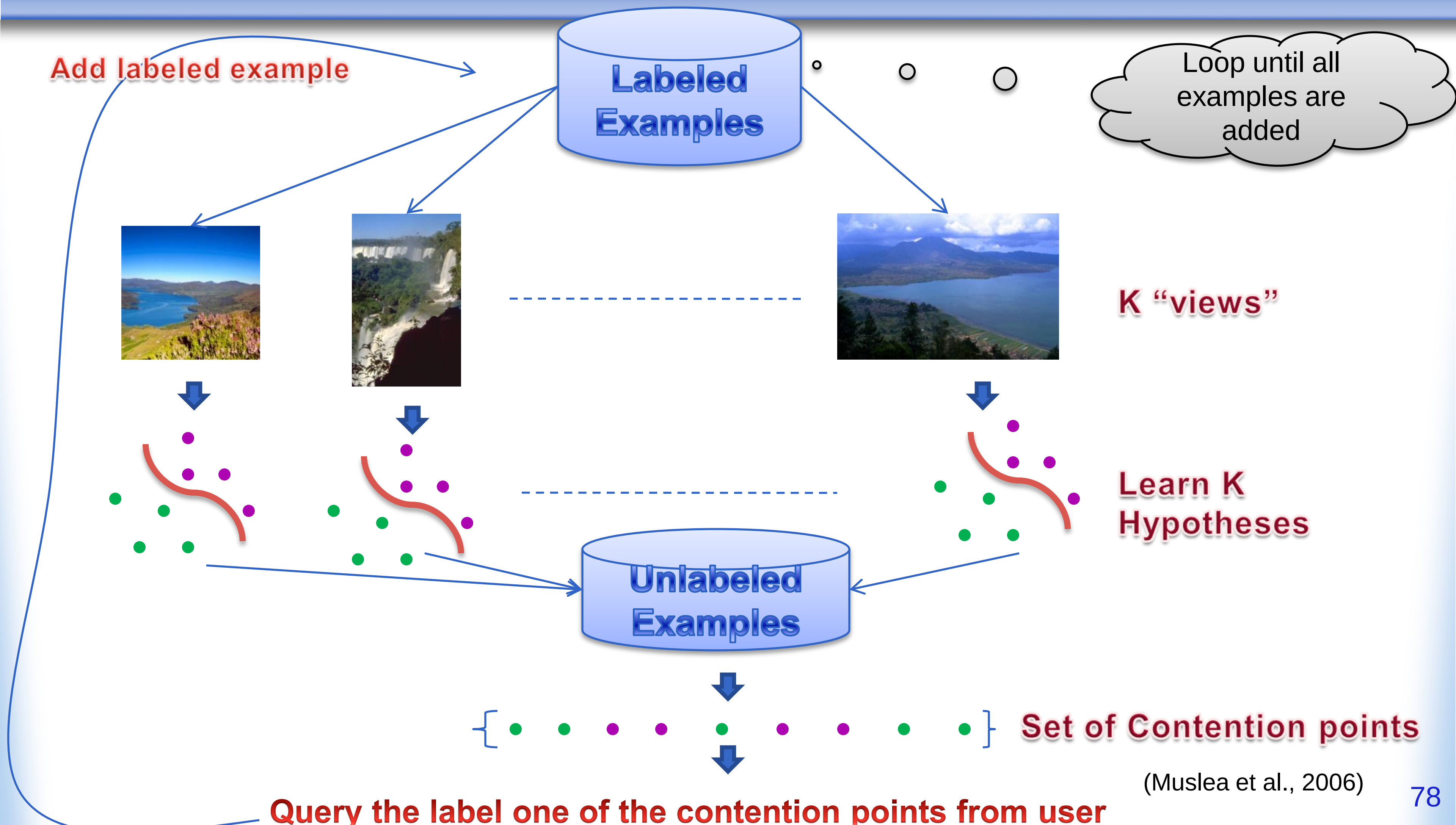
Query the label one of the contention points from user

Multi-view Active Learning
Co-Testing



(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing



(Muslea et al., 2006)

Multi-view Active Learning Co-Testing

- The above algorithm refers to a family of Co-Testing algorithms
- Each algorithm is defined by the choice of
  - Selection of contention point to be queried
  - Creation of final output hypotheses

(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing

- Selection of contention point to be queried
  - *Naïve*: random selection
  - *Aggressive*: choose contention point where least confident hypotheses make most confident prediction

$$Q = \arg\max_{x \in Contention\ Points} \left\{ \min_{i \in \{1,2,\ldots,k\}} Confidence(h_i(x)) \right\}$$

  - *Conservative*: choose contention point where confidence of prediction of hypotheses is as close as possible

$$Q = \arg\min_{x \in Contention\ Points} \left\{ \max_{f \in \{h_1,\ldots,h_k\}} Confidence(f(x)) - \min_{g \in \{g_1,\ldots,g_k\}} Confidence(g(x)) \right\}$$

(Muslea et al., 2006)

Multi-view Active Learning
Co-Testing

- Creation of final output hypotheses
  - *Weighted vote*: combines the vote of each hypothesis, weighted by the confidence of their respective predictions.
  - *Majority vote*: chooses the label that was predicted by most of the hypotheses
  - *Winner-takes-all*: the output hypothesis is the one learned in the view that makes the smallest number of mistakes over the N queries

(Muslea et al., 2006)

Savings from Active Learning

- Results
  - Results presented over 3 domains: web-page classification, discourse tree parsing and advertisement removal
  - Results show that Co-Testing outperforms all the tested single-view algorithms statistically significantly (t-test confidence of atleast 95%)

(Muslea et al., 2006)

Other strategies

- Diversity Sampling: To maximize the training utility of batch
  - *Global:* Cluster based on similarity & select examples from different clusters
  - *Local:* Select examples that are most different from the examples already selected from the pool

- Representativeness
  - Number of examples similar to it
  - Choose centroids of the clusters
  - Less likely to be outliers and most informative

(Shen et al., 2004)

Other strategies

- Diversity Sampling: To maximize the training utility of batch
  - *Global:* Cluster based on similarity & select examples from different clusters
  - *Local:* Select examples that are most different from the examples already selected from the pool

- Representativeness
  - Number of examples similar to it
  - Choose centroids of the clusters
  - Less likely to be outliers and most informative

(Shen et al., 2004)

Conclusion & Discussion

- Selective sampling methods
  - Uncertainty-based
  - Query-by-committee
- *Interesting ideas…*
  - Querying partial labels
  - Combination with semi-supervised and multi-view techniques
  - Appropriate measures for user-effort

Questions

Please send your feedback to:
shilpaa@cs.cmu.edu & sachina@cs.cmu.edu

References

- McCallum, A. and Nigam, K. (1998). Employing EM and pool-based active learning for text classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning.*
- Muslea, I., Minton, S., and Knoblock, C. A. (2006). Active learning with multiple views, *Journal of Artificial Intelligence Research (JAIR), 27:203-233.*
- Steedman, M., Hwa, R., Clark, S., Osborne, M., Sarkar, A., Hockenmaier, J., Ruhlen, P., Baker, S., and Crim, J. (2003). Example selection for bootstrapping statistical parsers. In *NAACL '03, pages* 157-164, Morristown, NJ, USA.

References

- Shen, D., Zhang, J., Su, J., Zhou, G., and Tan, C.-L. (2004). Multi-criteria-based active learning for named entity recognition. In *ACL '04: page 589, Morristown, NJ, USA.*
- Thompson, C. A., Cali, M. E., and Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. In *Proceedings of 16th ICML-1999, pages 406-414. Morgan Kaufmann, San* Francisco, CA.
- Sculley, D. (2007). Online active learning methods for fast label-efficient spam filtering.In *CEAS 2007: Proceedings of the Fourth Conference on Email and Anti-Spam.*
- Roth, D. and Small, K. (2006). Active learning with perceptron for structured output. In *ICML 06: Workshop on Learning in Structured Output Spaces.*

References

- Kristjannson, T., Culotta, A., Viola, P., and Callum, A. M. (2004). Interactive information extraction with constrained conditional random fields. In *AAAI 2004, San Jose, CA.*
- Hwa, R. (2000). Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing andvery large corpora, pages 45-52, Morristown, NJ, USA.*