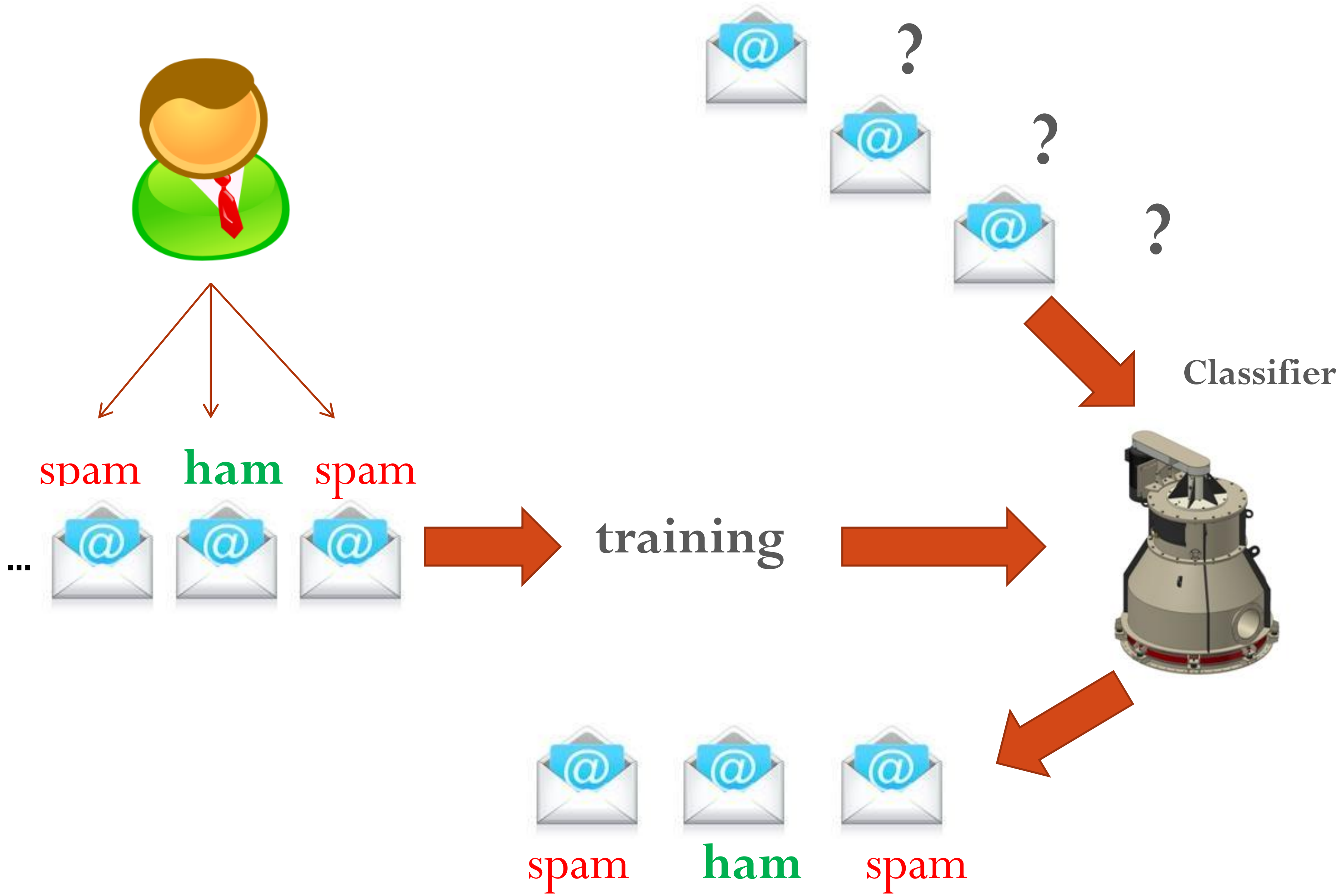
Spam Filtering: an Active Learning approach using Incremental Clustering

Kleanthi Georgala[1,2]
Aris Kosmopoulos[1,3]
Georgios Paliouras[1]

[1]National Center of Scientific Research "Demokritos", Agia Paraskevi, 153 10 Attiki, Greece
[2]LIACS, Leiden University, the Netherlands
[3]Athens University of Economics and Business, Athens, Greece

Classifiers as spam filters



Problem with Classifiers

- **User cannot provide labels for all messages**

Solution :

- Minimize manual labeling → **Active Learning**
  - Select a set of instances
  - Train classifier with this subset
    - Random selection, uncertainty sampling
      - Outliers, non-representative instances

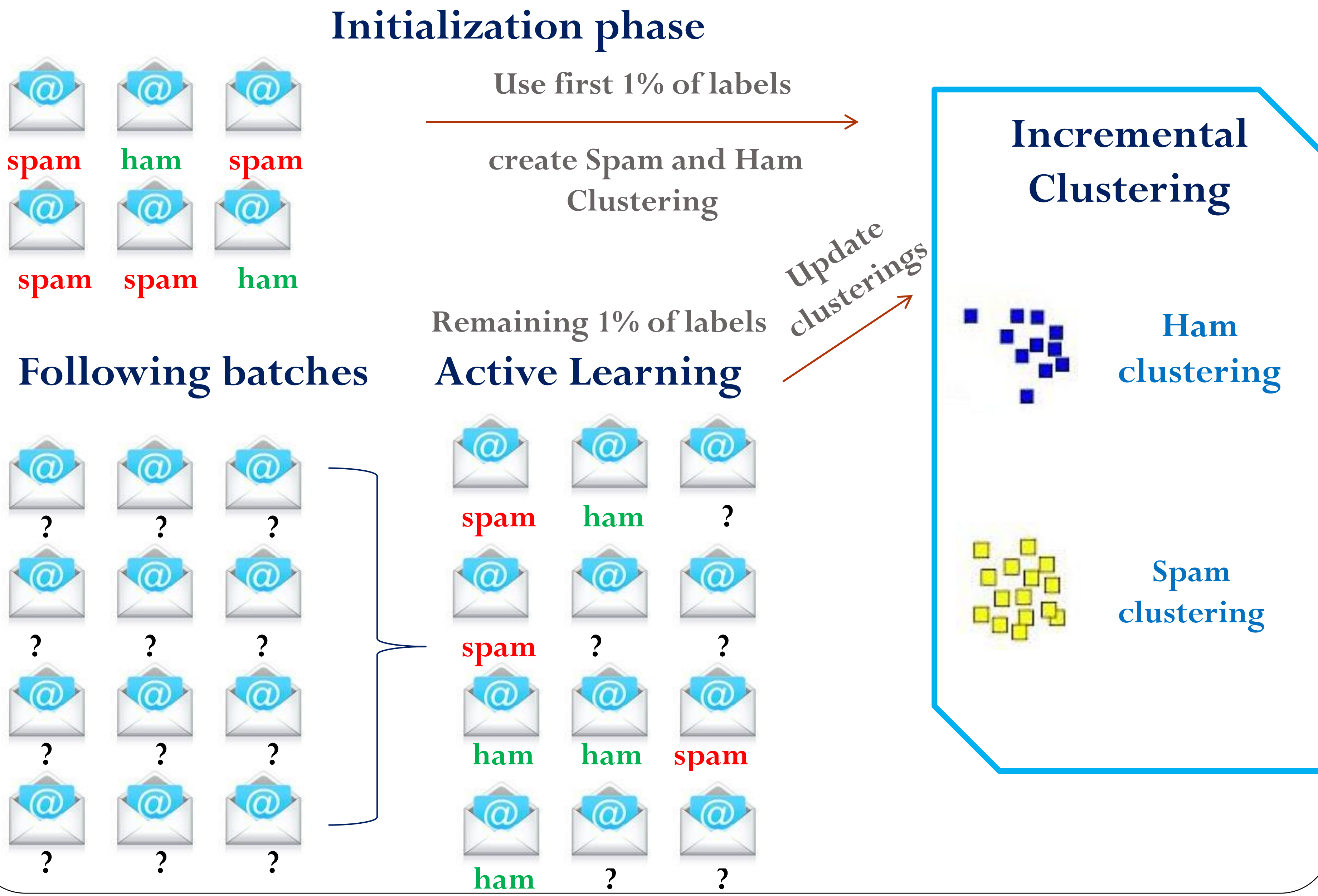
Problem with Classifiers

- Incorporate **Incremental Clustering**
  - Unsupervised learning, based on local structure
  - Create groups of highly correlated data-points
  - **No re-clustering of data**

- **Our contribution : Active Learning** combined with **Incremental Clustering**
  - Use only **2% of the overall message labels**
  - Consider **natural grouping** of data : select **representative instances** for training

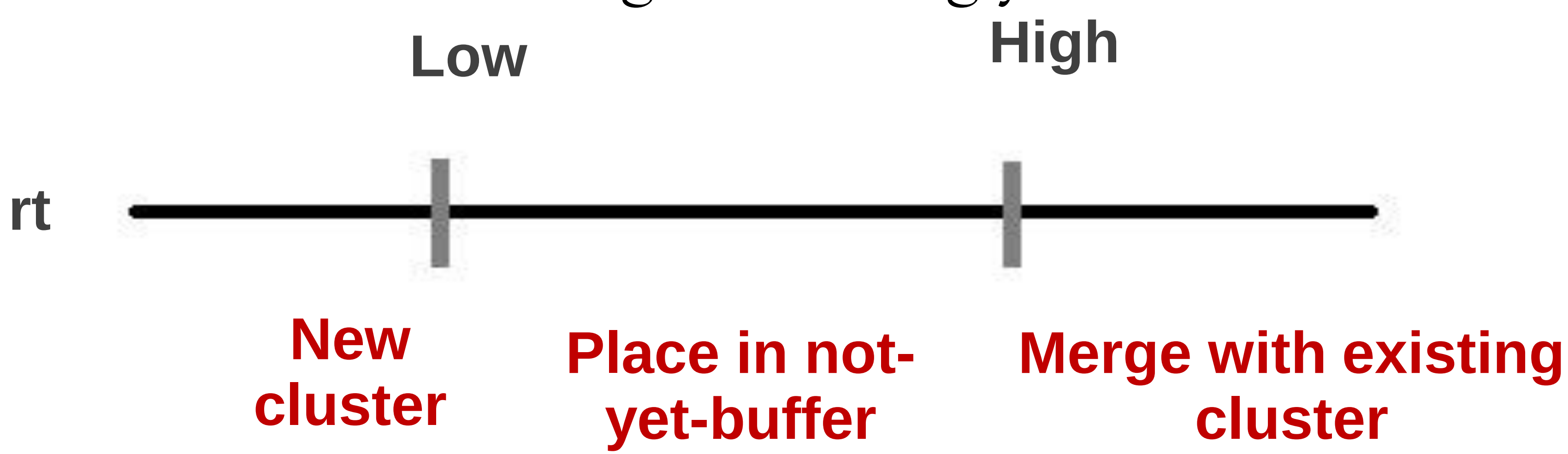
Active Learning combined with Incremental Clustering

Active Learning combined with Incremental Clustering



Initialization Phase

- Until the first 1% of labels is reached:
  - For each new message :
    - Request message's label
    - Compute ***rt*** based on label
    - Place the message accordingly:



Incremental Clustering

Given a message $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_n\}$ and a cluster $\boldsymbol{C}_{j,k}$ of a clustering $\boldsymbol{Cl}_j$ compute :

$$rt = \frac{\sum_{X_i} B_{X_i}}{\sum_{X_i} K_{X_i} + \sum_{X_i} K'_{X_i}}$$

- $\boldsymbol{X}_i$:word at position *i*
- $\boldsymbol{B}_{X_i}$: the number of already classified messages that contain X_i
- $\boldsymbol{K}_{X_i}$: the number of messages that belong to a cluster $\boldsymbol{C}_{j,k}$ and do not contain X_i
- $\boldsymbol{K`}_{X_i}$ **:** the number of messages in $\boldsymbol{Cl}_j$ that contain the word X_i but are not included in a cluster $\boldsymbol{C}_{j,k}$

Active Learning

- For the following batches :
  - For each new incoming message:
    - Compute ***rt*** for both clusterings → ***rtH / rtSp***
    - ***Until 1% of labels is reached :***
      - Select/place instances based on:

	rtH < low	low < rtH < high	rtH > high
rtSp < low	✓	✓	✗
low < rtSp < high	✓	♦	✗
rtSp > high	✗	✗	♦

Learning Algorithms

- **Limited training (*LT*)** : train classifier only with labelled messages
- **Semi-supervised training (*SST*)** : train classifier on all messages
  - true label for selected instances
  - classifier's predictions on unlabelled
- **Meta-classifier (*linear*):** weighted combination of LT and SST
  - Weights based on accuracy

Experiments

Experimental set-up

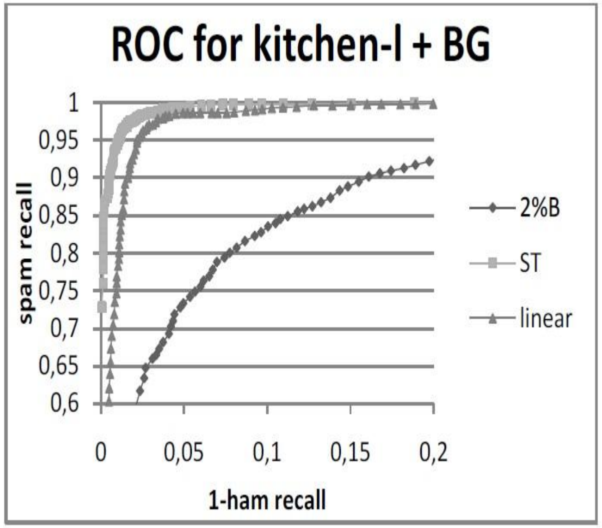
- ***Datasets :***
  - **Enron-Spam, NSCR "Demokritos"**
- ***Baseline :*** **2%B**
- ***Target Model :*** **Supervised Training (ST)**
- ***Thresholds tested:*** [0.3,0.5], [0.5,1.0]
- ***Evaluation***
  - *ROC curves*
    - **x-axis : 1-ham recall (**1-specifity**)**
    - **y-axis : spam recall (**sensitivity**)**
  - *Area Under Curve (AUC)*
  - *Statistical significance based on AUC*
- ***Classifier :*** *Naïve Bayes*

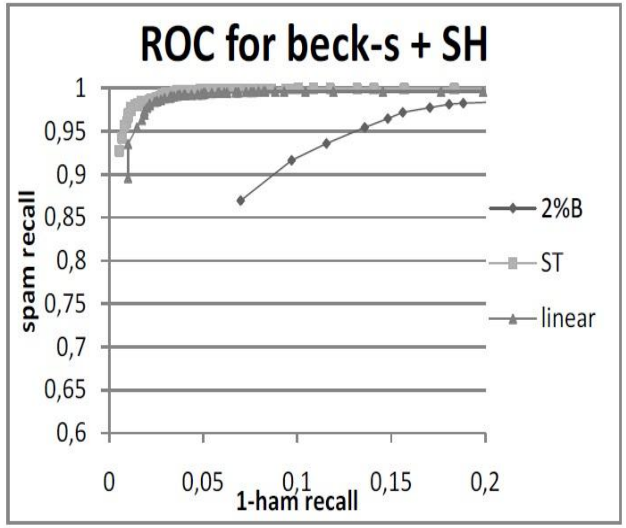
Experimental results

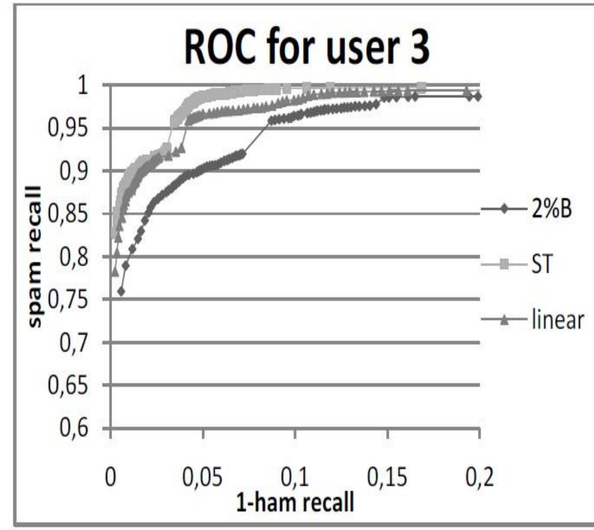
Proportion of spam/ham requested labels

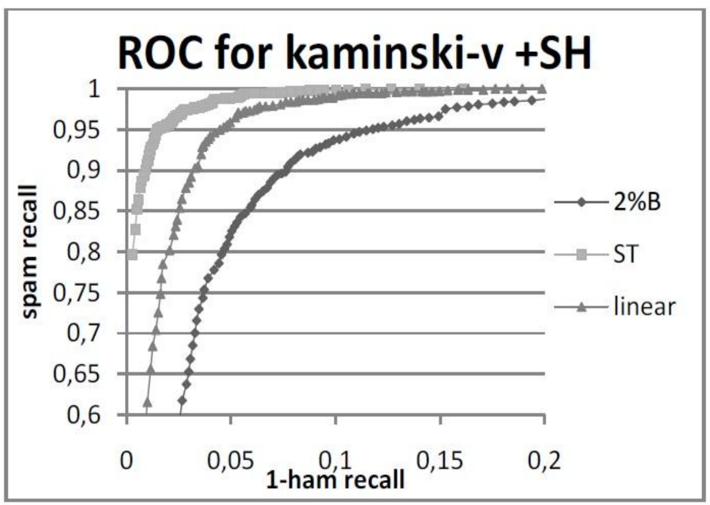
Datasets	First 1%		Extra 1% [0.3-0.5]		Extra 1% [0.5-1]		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
farmer-d +GP	42	9	27	24	33	18	51
kaminski-v	46	12	44	14	46	12	58
kitchen-l + BG	40	15	40	15	41	14	55
williams-w3 + GP	17	43	9	51	13	47	60
beck-s + SH	16	35	8	43	13	38	51
lokay-m + BG	16	44	2	58	13	47	60
User 1	67	31	80	18	69	29	108
User 2	36	98	44	90	55	79	134
User 3	68	108	98	76	68	93	161
User 4	59	23	62	20	58	24	86

Good Cases



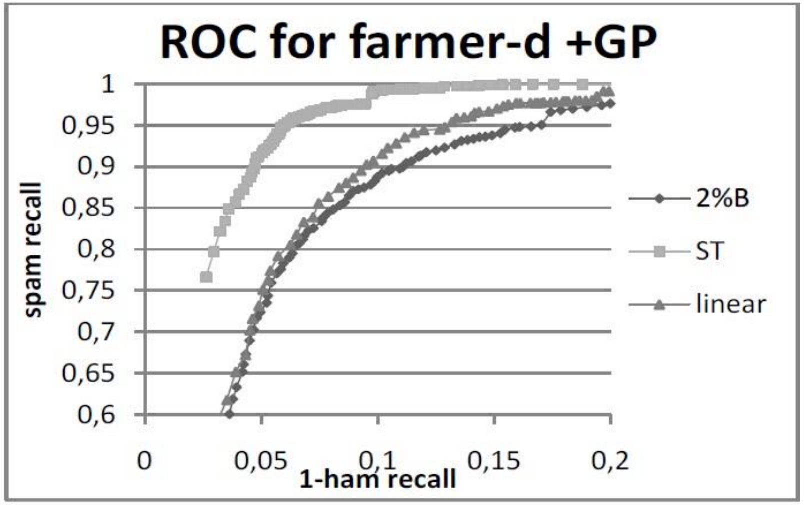




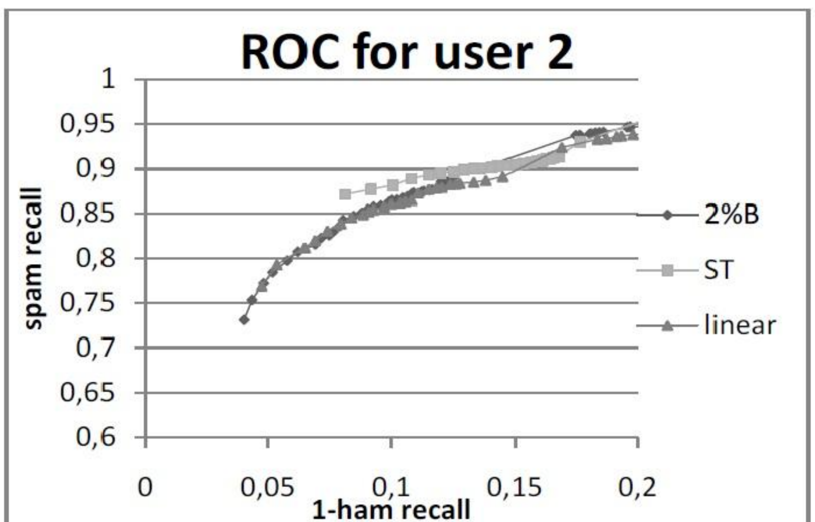


- Limit x-axis between 0.0-0.2
- Statistical significant differences between ***2%B*** and ***linear***
- Similar performance between ***ST*** and ***linear***

Problematic Cases



• *2%B* similar to ***linear***



• ***ST*** similar to ***2%B***

• ***Low overall performance of methods***

Experimental results

Datasets	Descending order of the methods based on their AUC				
farmer-d +GP	ST	Linear	LT*	2%B*	SST
kaminski-v	ST	Linear	LT*	SST	2%B
kitchen-l + BG	ST	SST*	Linear*	2%B	LT*
williams-w3 + GP	ST	SST	Linear	2%B	LT
beck-s + SH	ST	SST*	Linear	LT	2%B*
lokay-m + BG	ST	SST*	Linear	2%B	LT*
User 1	ST	Linear	LT*	2%B*	SST
User 2	ST	Linear*	LT*	2%B*	SST
User 3	ST	SST*	Linear	2%B	LT*
User 4	ST	Linear	LT	2%B	SST

- ** :The difference between this method and the method on the left is not statistically significant*
- ***In half datasets : LT > SST. In other half : SST > LT***
- ***linear > 2%B : statistical significant differences***
- ***linear*** achieves similar results with the ***ST***

Recap and Conclusions

- Spam filtering incorporating **Active Learning and Incremental Clustering**
- Selectively request labels for messages
- Request **only 2%** of overall labels

- Good performance with limited data
- Best learning method outperforms baseline
- Similar results with fully supervised approach

Thank you!

Questions?