Spam Track

Past, Present and Future

Gordon V. Cormack

15 November 2006

University of Waterloo

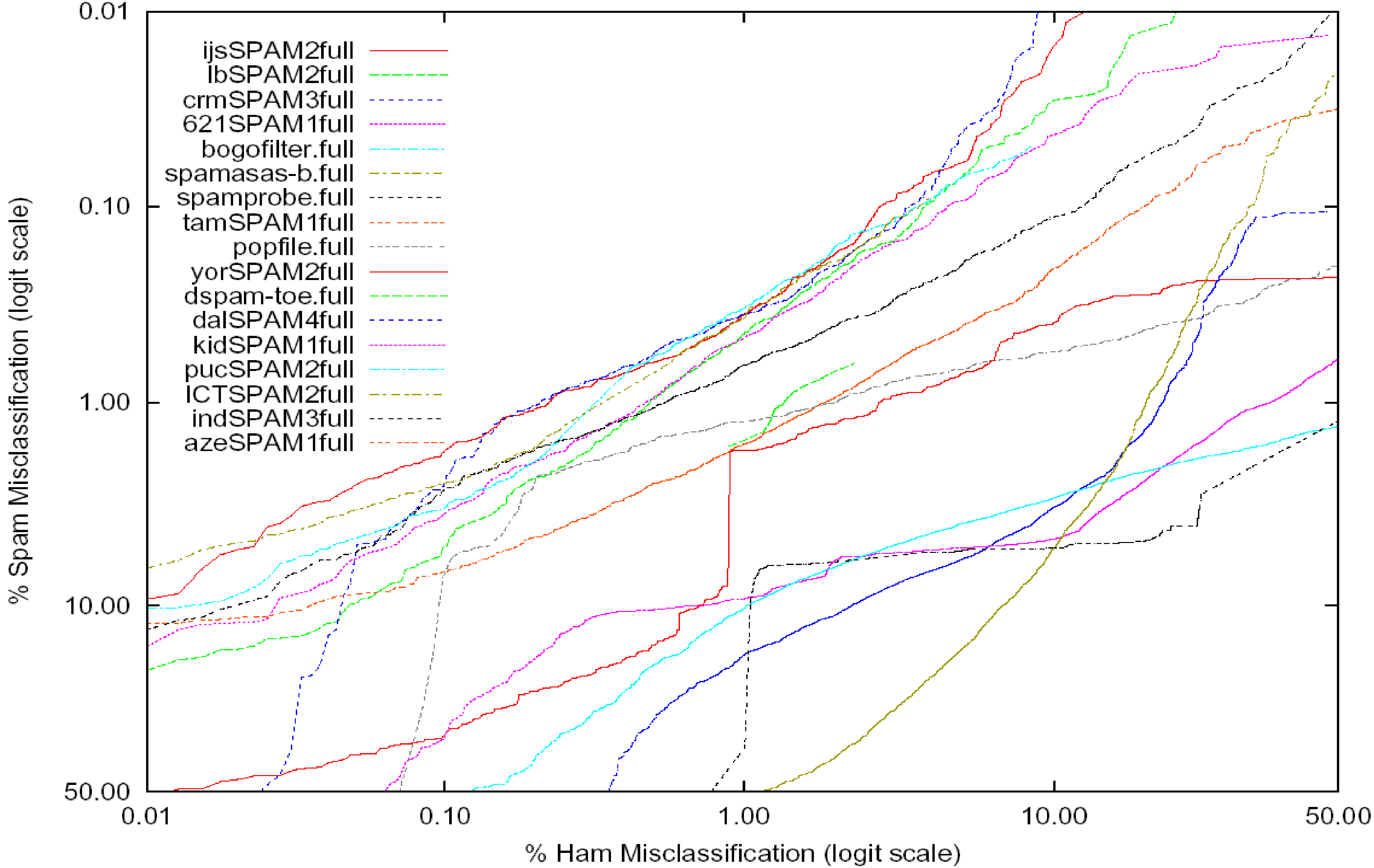
Pre-TREC History

- Academic evaluations
  - vector-space, batch test/training sets, machine learning methods, *accuracy* as evaluation measure
- In-house evaluations (& testimonials)
- MrX Corpus (Cormack & Lynam)
  - capture real user's email July '03 – Feb '04
  - careful construction of *gold standard*
  - on-line testing
  - open-source *'Bayesian'* and *rule-based* filters
  - ROC analysis

TREC 2005

- Tension between privacy and archival corpus
  - standardized filter interface and toolkit
- Private corpora (MrX, SB, TM)
  - MrX runs available on request
- Public corpus
  - 90,000 messages (Enron + seeded spam)
  - download: (google for TREC spam corpus)
  - amusement: spamorham.org (J. Graham-Cumming)
- Online classification task
  - *idealized* user gives immediate, accurate feedback

ROC Curves – TREC 2005



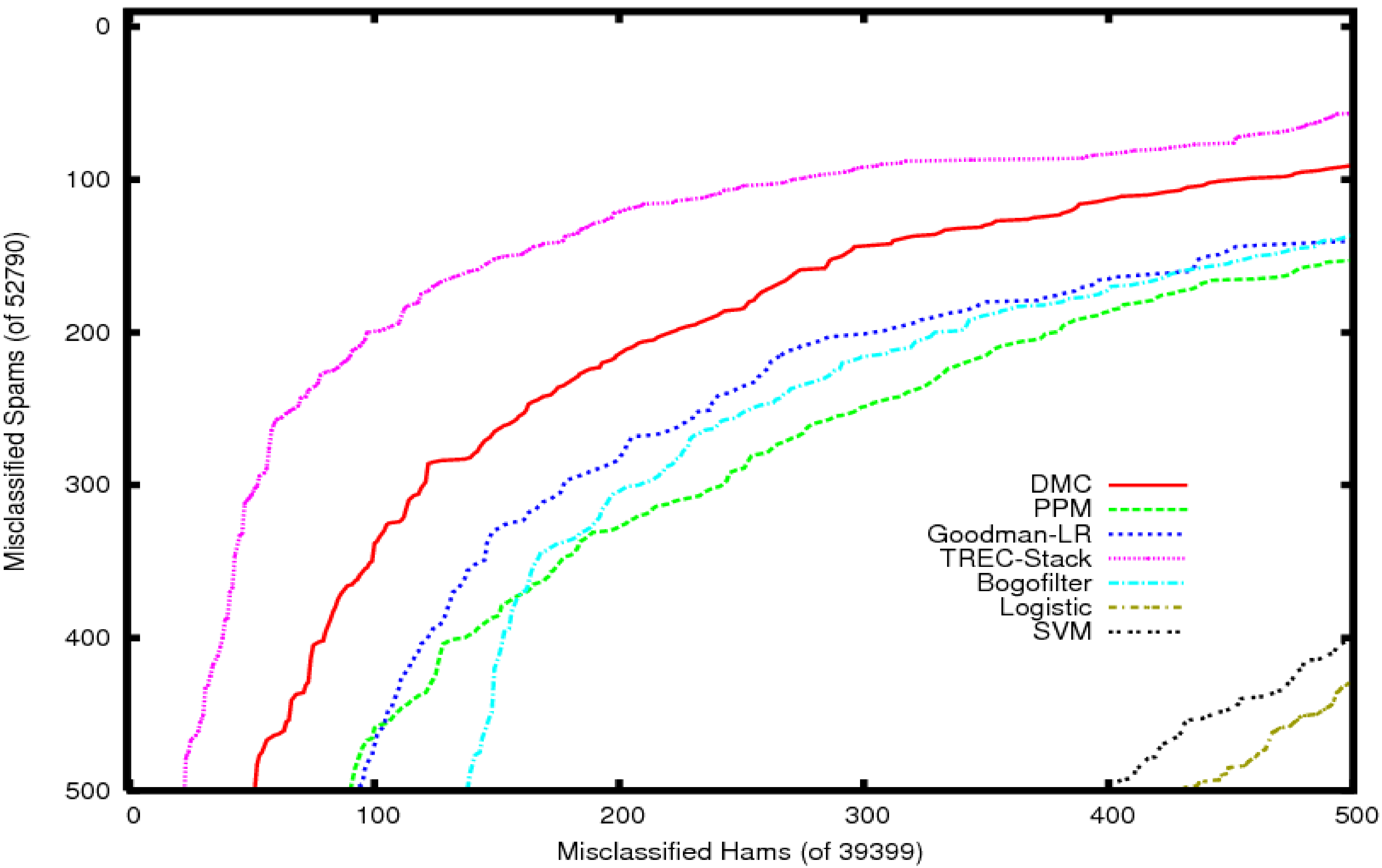
TREC 2005 Outcomes

- Logistics of preparing/submitting/evaluating
- Public & private corpora yield comparable results
- Compression models worked very well (Bratko)
- Why no (strong) machine learning methods?
- Is ideal user realistic? Effect of delay/error?
- Are spammers defeating these methods faster than we can evaluate them?
- What about other real-time aspects? Blacklists, greylists, spam warehouses?

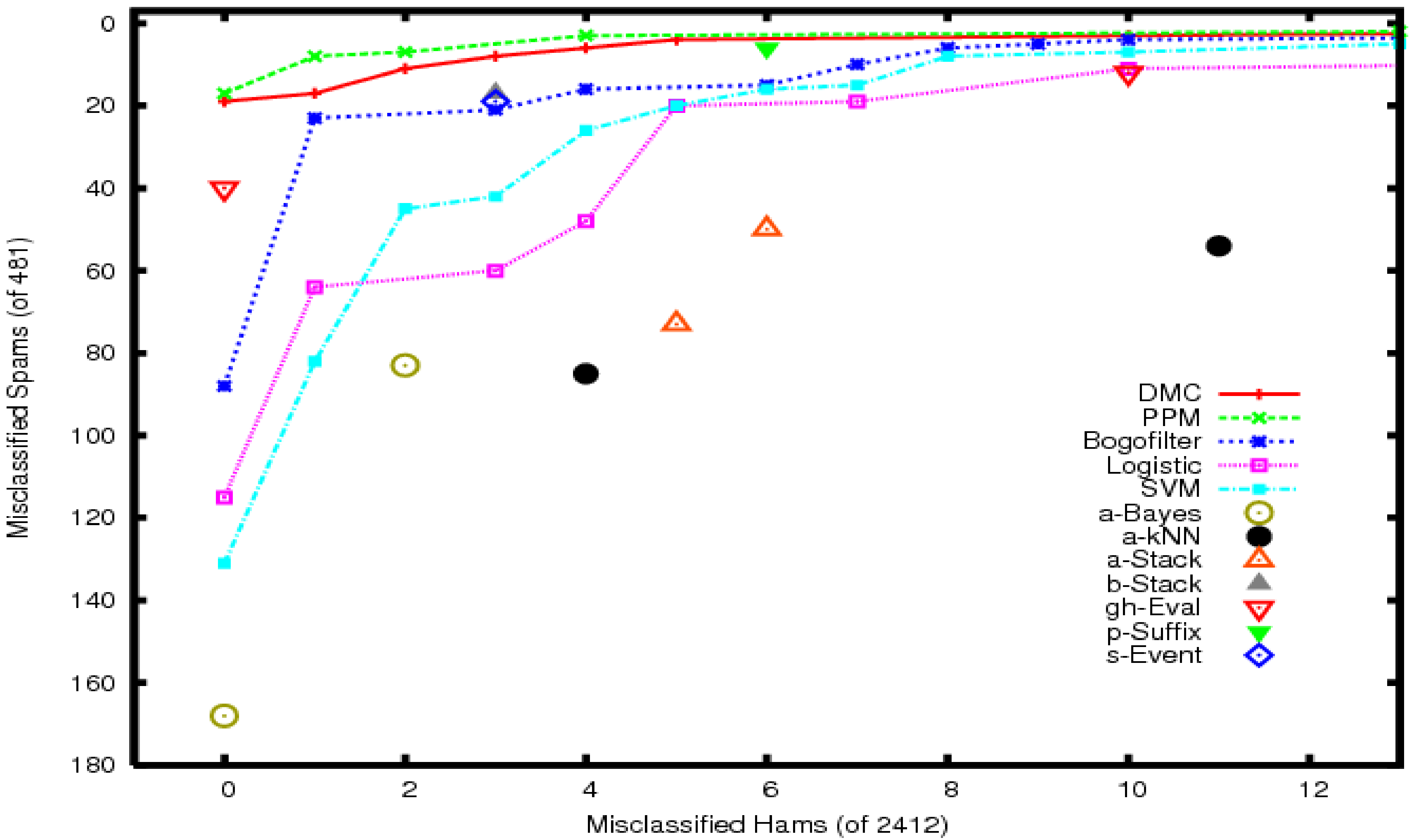
Since TREC 2005

- TREC vs ML-style evaluation
- DMC – bit-wise compression-based method
- Stacking (fusion) of the TREC 2005 filters
- ECML Discovery Challenge
- Design of TREC 2006
  - TREC 2005 + delayed feedback + active learning
- TREC 2006
- TREC 2007
- Other evaluations?

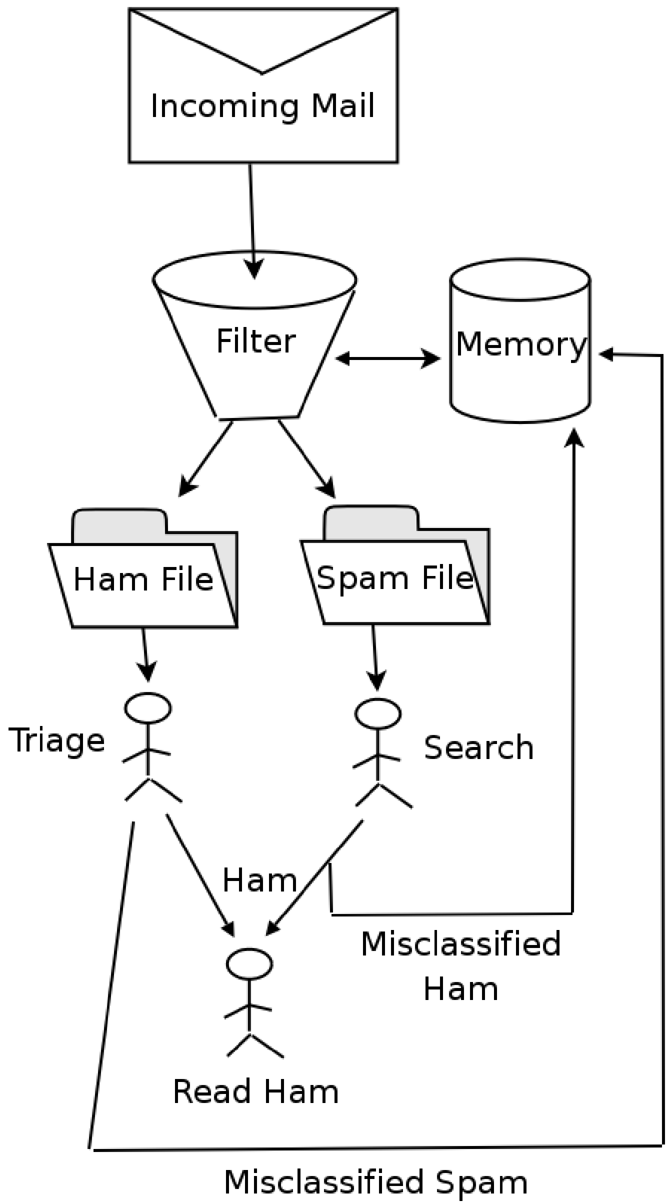
Since TREC 2005



Ling Spam Corpus



Spam Filter Usage



- Filter Classifies Email
- Human addressee
  - Triage on ham File
  - Reads ham
  - Occasionally searches for misclassified ham
  - Report misclassified email to filter

2006 Tasks

Immediate Feedback

- reprise TREC 2005 *idealized user*

Delayed Feedback

- lazy user reports classification later, in batches
- batch size random, avg 500 -- 1000 messages

Active Learning

- sequence of unclassified messages
- filter requests true classification for some
- predict future sequence of messages

Immediate and delayed feedback tasks

Filter is invoked through standard commands:

initialize

create necessary files & servers (cold start)

classify *filename*

read *filename* which contains exactly 1 email message

write one line of output:

classification score auxiliary_file

train *judgement filename classification*

take note of gold-standard *judgement*

finalize

clean up: kill servers, remove files

Active learning task

Filter implements a shell program:

for n = 100, 200, 400, ...

read training data (1st 90% of corpus)

for i from 1 to n

request classification for 1 message

for each message in test data (last 10% of corpus)

output classification

erase memory

2006 Data

- Newer versions of private corpora
  - MrX (2003-04) ==> MrX II (2005-06)
  - SB (2004-05) ==> SB II (2005-06)
- (Mostly) English public corpus
  - Web retrieval of mbox-format files (1993-2006)
  - Augmented by spam-trap spam (2006) spoofed to simulate delivery to (paired) web message
- Chinese public corpus (Courtesy CCERT)
  - Mailing list ham
  - Spam trap spam

Corpora

Private Corpora

	Ham	Spam	Total
MrX2	9039	40135	49174
SB2	9274	2695	11969
Total	18313	42830	61143

Public Corpora

	Ham	Spam	Total
trec06p	12910	24912	37822
trec06c	21766	42854	64620
Total	34677	67766	102442

Run Tag Suffixes

Corpus / Task	Filter Suffix
trec06p / immediate feedback	pei
trec06p / delayed feedback	ped
trec06c / immediate feedback	pci
trec06c / delayed feedback	pcd
MrX2 / immediate feedback	x2
MrX2 /delayed feedback	x2d
SB2 / immediate feedback	b2
SB2 / delayed feedback	b2d

Active Run Tag Suffixes

pei.*nnn* – Public English, *nnn* training examples

cei.*nnn* – Public Chinese, *nnn* training examples

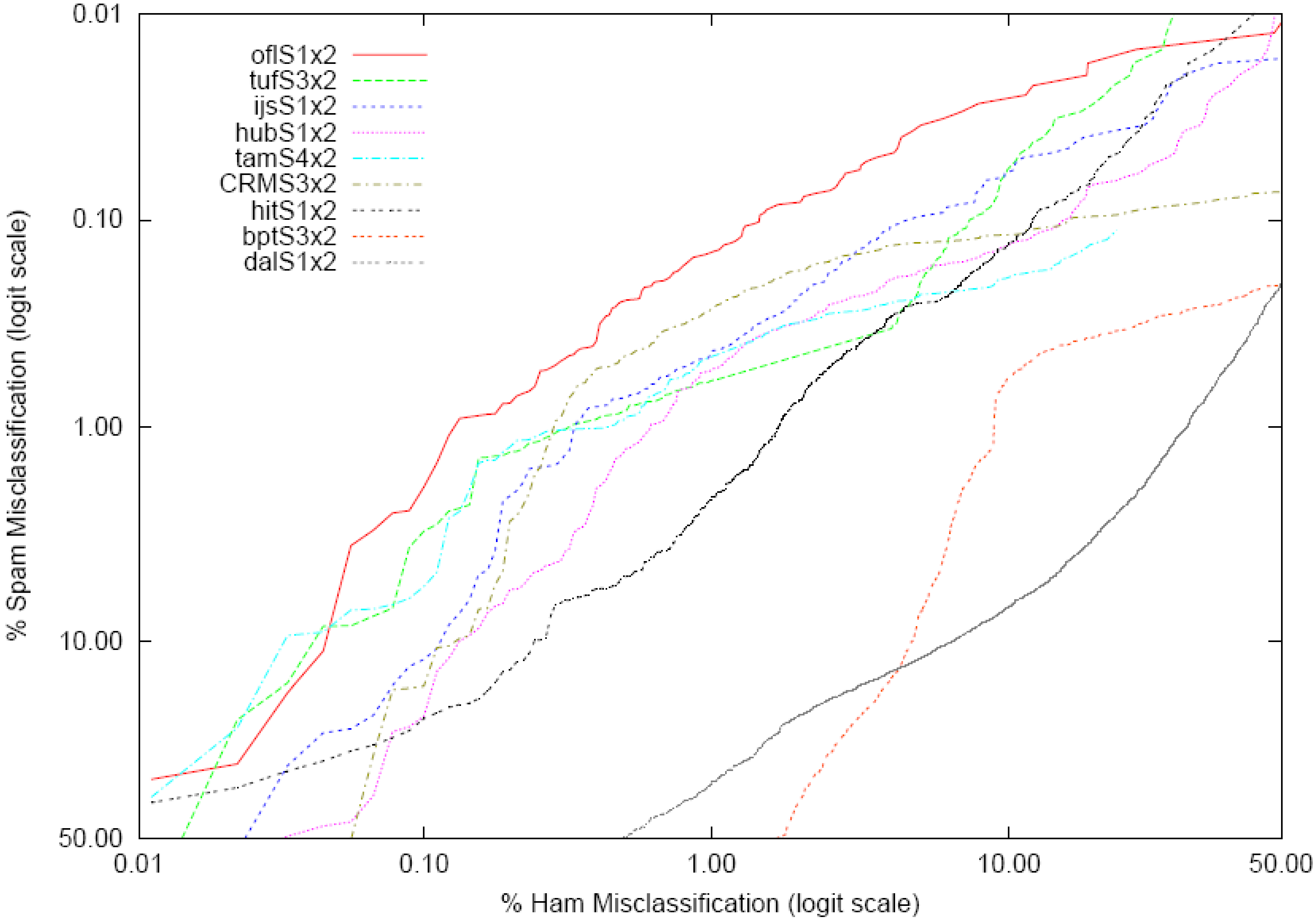
x2.*nnn* – MrX II, *nnn* training examples

b2.*nnn* – Mrx II, *nnn* training examples

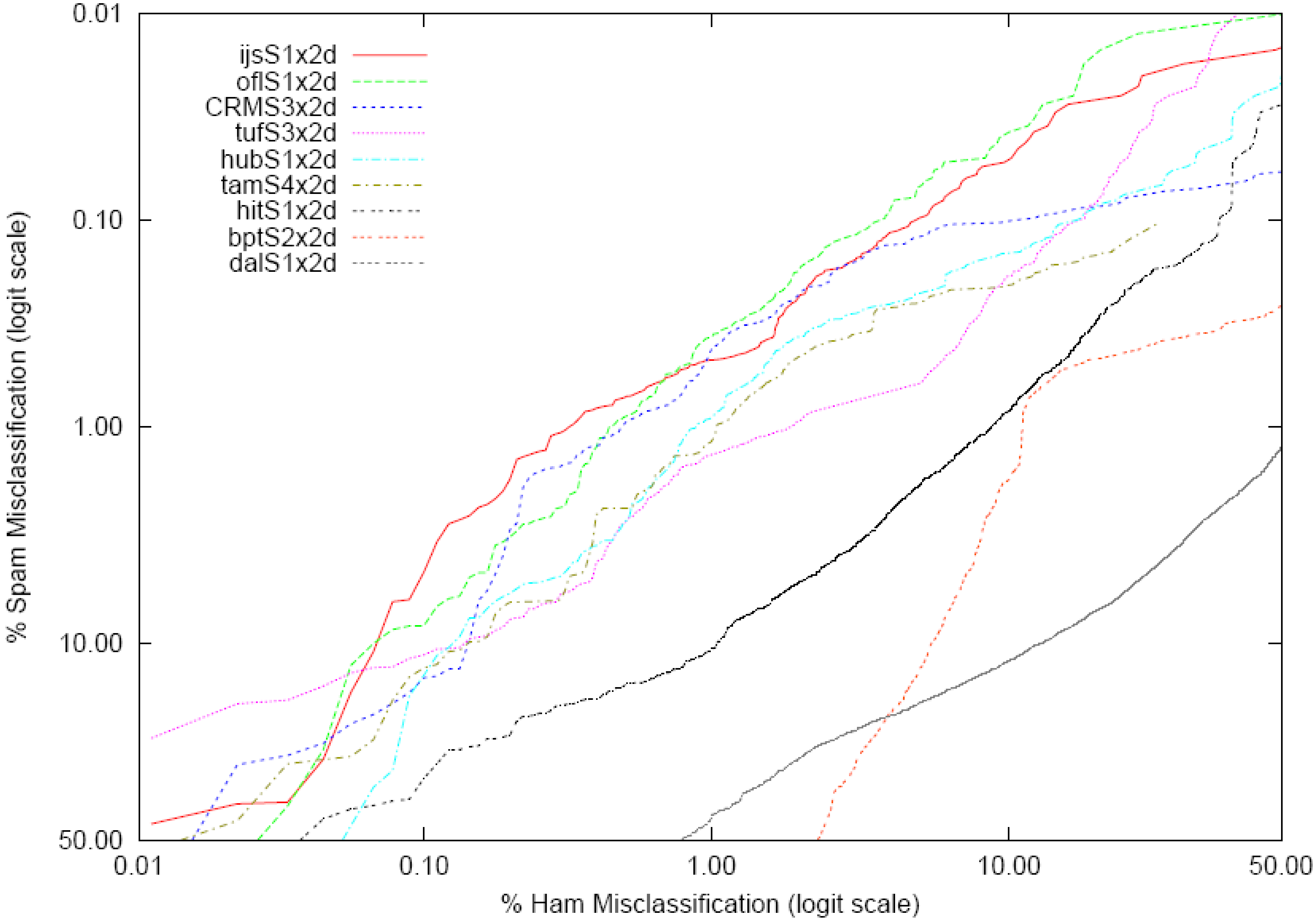
Participant Filters

Group	Filter Prefix
Beijing University of Posts and Telecommunications	bpt
Harbin Institute of Technology	hit
Humboldt University Berlin & Strato AG	hub
Tufts University	tuf
Dalhousie University	dal
Jozef Stefan Institute	ijs
Tony Meyer	tam
Mitsubishi Electric Research Labs (CRM114)	CRM
Fidelis Assis	ofl

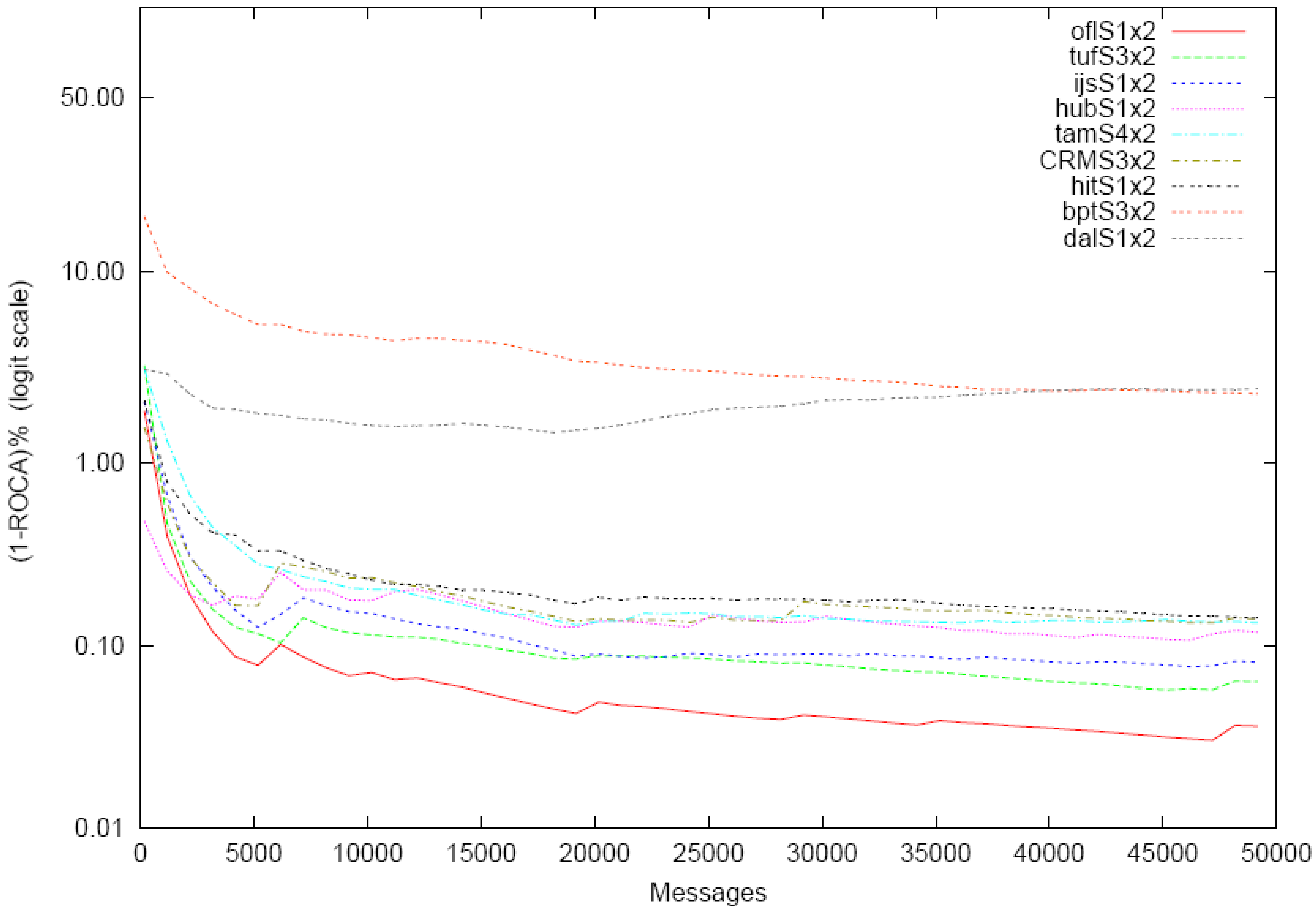
MrX II – Immediate Feedback



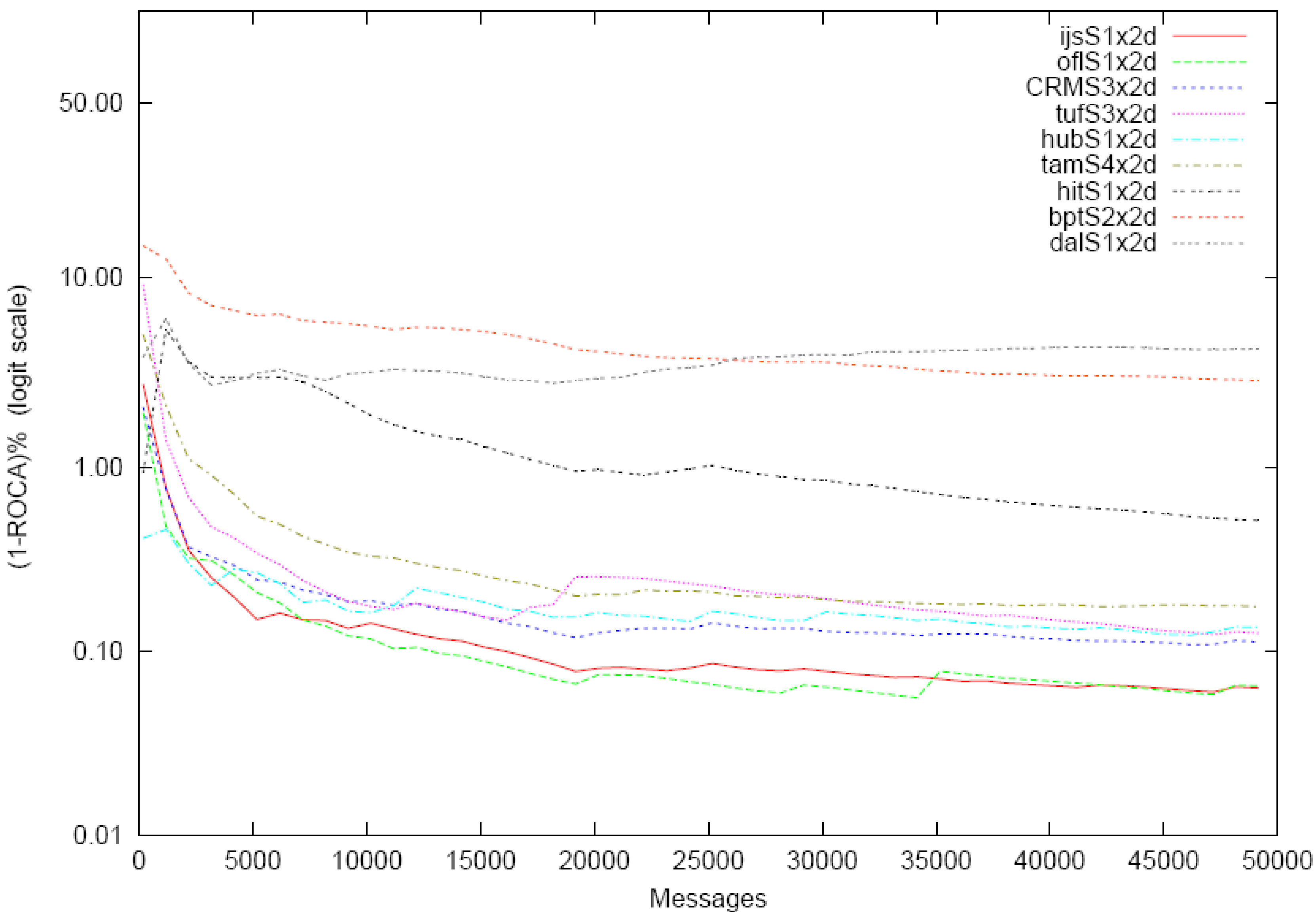
MrX II – Delayed Feedback



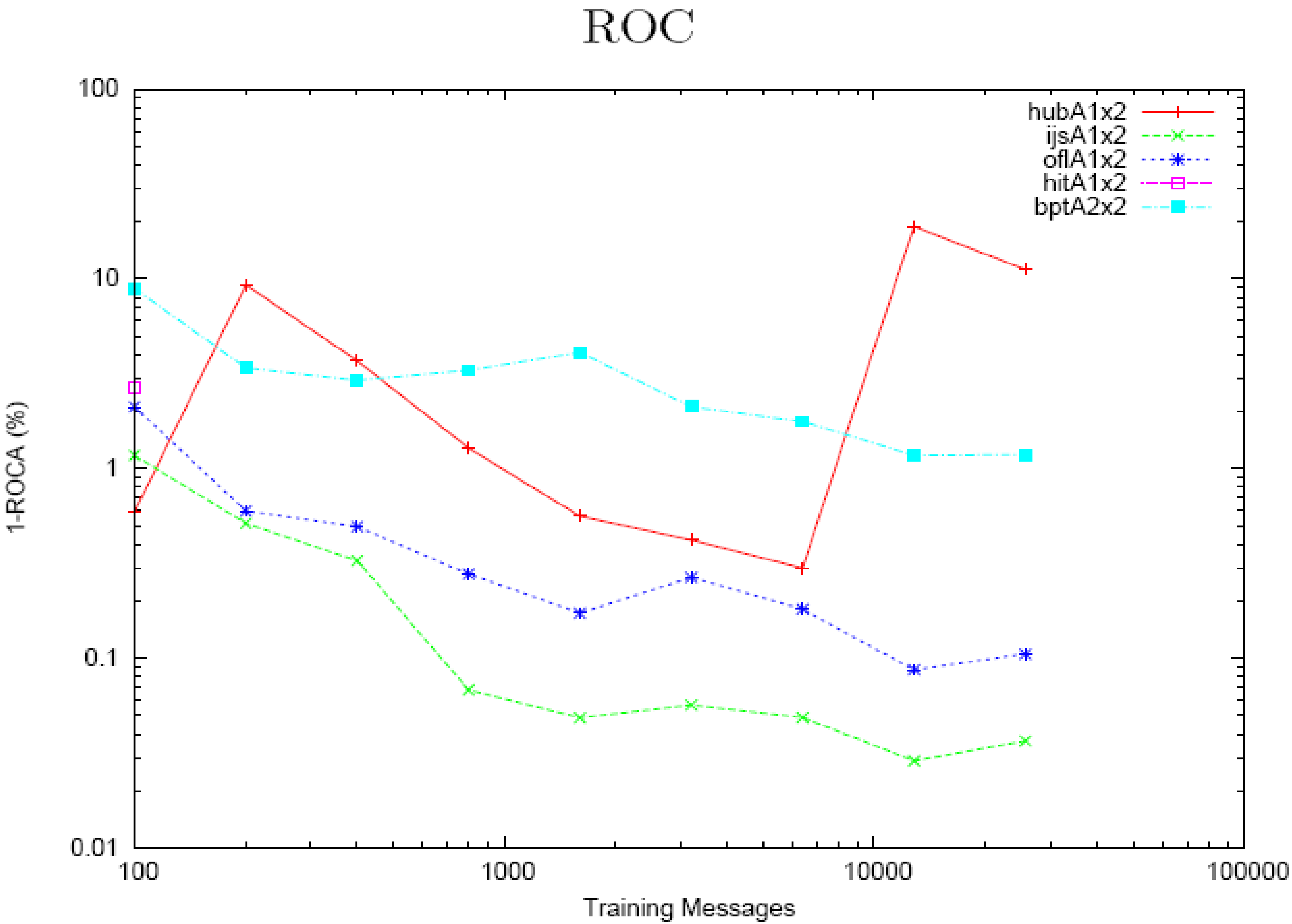
MrX II immediate learning curve



MrX II delay – learning curves



Mrx II – Active Learning



1-ROCA (%)

Run	X2	X2d	100	200	400	800	1600	3200	6400
Ofl	0.04	0.07	2.11	0.60	0.49	0.28	0.17	0.27	0.18
DMC	0.05	0.09	1.80	0.14	0.08	0.08	0.13	0.08	0.08
Tuf	0.06	0.13							
Ijs	0.08	0.06	1.17	0.51	0.33	0.07	0.05	0.06	0.05
Bogo	0.09								
Hub	0.12	0.14	0.59	0.60	0.37	0.50	0.36	0.42	0.28
Tam	0.13	0.18							
Crm	0.14	0.11							
Hit	0.14	0.52	2.66						
Bpt	2.35	3.08	9.10	3.40	2.90	3.27	3.91	2.12	1.77
Dal	2.50	4.34							

1-ROCA (%) Multi-corpus results

	Aggregate		trec06p		trec06c		MrX2		SB2	
Filter\Feedback	immediate	delay	immediate	delay	immediate	delay	immediate	delay	immediate	delay
oflS1	0.0295	0.1914	0.0540	0.1668	0.0035	0.0666	0.0363	0.0651	0.1300	0.3692
oflS3	0.0327	0.1908	0.0562	0.1702	0.0035	0.0601	0.0523	0.0824	0.1249	0.3174
oflS2	0.0365	0.2018	0.0597	0.2045	0.0104	0.1297	0.0525	0.0931	0.1479	0.3659
tufS2	0.0370	0.1079	0.0602	0.2038	0.0031	0.0104	0.0691	0.1449	0.3379	0.6923
oflS4	0.0381	0.1828	0.0583	0.1965	0.0077	0.0855	0.0718	0.1155	0.1407	0.2941
tufS1	0.0445	0.1262	0.0602	0.2110	0.0023	0.0081	0.0953	0.1991	0.3899	0.8361
ijsS1	0.0488	0.2119	0.0605	0.2457	0.0083	0.1117	0.0809	0.0633	0.1633	0.4276
tufS3	0.0705	0.1497	-	-	-	-	0.0633	0.1263	0.3350	0.6137
tufS4	0.0749	0.1452	-	-	-	-	0.0750	0.1314	0.3199	0.5696
CRMS3	0.0978	0.1743	0.1136	0.2762	0.0105	0.0888	0.1393	0.1129	0.2983	0.4584
CRMS2	0.1011	0.1667	0.1153	0.2325	0.0094	0.0975	0.1592	0.1143	0.4196	0.6006
CRMS1	0.1081	0.2165	0.1135	0.2447	0.0218	0.0784	0.1498	0.1341	0.3852	0.6346
hubS3	0.1674	0.2170	0.1564	0.1958	0.0353	0.0495	0.2102	0.2294	0.6225	0.8104
hubS4	0.1717	0.2400	0.1329	0.2006	0.0233	0.0330	0.1385	0.1763	0.5777	0.6784
hubS1	0.1731	0.2013	0.1310	0.1418	0.0238	0.0319	0.1180	0.1359	0.5295	0.5779
hubS2	0.1945	0.2716	0.1694	0.2952	0.0273	0.0369	0.1450	0.1827	0.4276	0.5306
hitS1	0.2112	0.8846	0.2884	0.5783	0.2054	1.3803	0.1412	0.5184	0.5806	1.2829
CRMS4	0.2375	1.5324	0.4675	2.1950	0.0579	1.7675	0.3056	0.4898	0.9653	2.0009
tamS4	0.2493	0.4480	0.2326	0.4129	0.1173	0.2705	0.1328	0.1755	0.4813	0.9653
tamS1	0.3008	1.0910	0.4103	0.8367	0.0473	0.1726	0.4011	0.6714	0.5912	4.5170
tamS2	0.9374	3.2366	1.2414	3.9352	0.4464	1.5370	-	-	6.5258	23.8125
tamS3	1.5309	2.2236	1.0602	1.8279	0.2899	1.0860	0.9514	1.5965	1.8462	6.0056

Results of Note

Orthogonal sparse bigrams with threshold training for headers (Assis, p 461 of notebook)

Perceptron with margin (Tufts) – incremental classical machine learning

Uncertainty sampling & pre-training (Humboltd U.)

Train on most recent examples (IJS)

Short message prefixes (Tufts, also DMC)

Are Spammers Winning?

	MrX	MrX II
Ijs	.08 (.04 - .10)	.08 (.05 - .12)
Ofl	.07 (.04 - .11)	.05 (.03 - .10)
Tuf	.04 (.03 - .05)	.06 (.04 - .09)
DMC	.04 (.03 - .05)	.05 (.03 - .09)
Bogofilter	.05 (.03 - .06)	.09 (.07 - .11)

What's Next?