

Selective Digital Amnesia

Maura R. Grossman

maura.grossman@uwaterloo.ca

Gordon V. Cormack

gvcormac@uwaterloo.ca

DESI VII Workshop – ICAIL 2017

June 12, 2017

University of
Waterloo



Commingled in Large Datasets:

- Information subject to privacy rights
- Information subject to access rights

Multiple Aspects of Privacy/Access:

- Rights holder – individual • class • entity • public
- Information type – subject-specific • generic
- Risk tolerance for false positives and false negatives
- Time and resources available to accomplish the task

Aspirational Goal:

- Specify the criteria for privacy/access rights
- Identify the records that meet the criteria
- Remove/retrieve/segregate the identified records

Really Just the Flip Side of Information Retrieval (“IR”)

- Criterion is called “relevance” (or “non-relevance”)
- “High-Recall IR” / “High-Stakes IR”

HRIR / HSIR is a Classification Task:

- Inductive (*i.e.*, build the best possible classifier)
- Transductive (*i.e.*, categorize every record in the dataset as effectively and efficiently as possible)

Imprecision in the definition of privacy/access rights and specification of the criteria

- Experts disagree surprisingly often
- When experts disagree, what is the right answer?

As a practical matter, the assessment of a single expert is often considered “good enough”

Even so, expert assessment is too costly and time consuming for large datasets, and does not scale well

Expert assessment, by definition, violates privacy rights

Non-expert assessment exacerbates all of the above

The Ultimate Question

Given these challenges the \$64,000£ Question is:

Can technology-assisted review (“TAR”) improve on currently accepted practice?

An Historical Overview

TREC: The Text REtrieval Conference

Spam Track 2005 – 2007

Legal Track 2006 – 2011

Total Recall Track 2015 – 2016



The TREC 2007 Spam Track

Used both public and private email datasets

Introduced an “evaluation toolkit” (or “test jig”) for simulating a user-in-the-loop and sandboxed evaluation of participating spam filters

TREC 2007 Negotiated Boolean Query

Request #53: “Please produce any and all documents concerning the effect of Maleic hydrazide (MH) on the tumorigenicity in hamsters”

(“maleic hydrazide” OR (MH AND (pesticide! OR “weed killer” OR herbicide! OR (growth OR sprout!) w/3 (inhibitor! OR retardant))) OR “potassium salt” OR De-cut OR “Drexel MH” OR Gro-taro OR C₄N₂H₄O₂) AND (tumor! OR oncogenic OR oncology! OR pathology! OR pathogen!) AND (hamster! OR mice OR mouse OR rat OR rats OR rodent!)

TREC 2008

Team	Topic	Reviewed	Produced	Recall	Precision	F_1
H5	103	n/a	608,807	62.4%	81.0%	70.5%

TREC 2009

Team	Topic	Reviewed	Produced	Recall	Precision	F_1
Waterloo	201	6,145	2,154	77.8%	91.2%	84.0%
Waterloo	202	12,646	8,746	67.3%	88.4%	76.4%
Waterloo	203	4,369	2,719	86.5%	69.2%	76.9%
H5	204	20,000	2,994	76.2%	84.4%	80.1%
Waterloo	207	34,446	23,252	76.1%	90.7%	82.8%
	Average:	15,521	7,973	76.7%	84.7%	80.0%

Humans vs. (Two Kinds of) TAR

TECHNOLOGY-ASSISTED REVIEW IN E-DISCOVERY CAN BE MORE EFFECTIVE AND MORE EFFICIENT THAN EXHAUSTIVE MANUAL REVIEW

By Maura R. Grossman^{*} & Gordon V. Cormack^{†**}

Cite as: Maura R. Grossman & Gordon V. Cormack,
*Technology-Assisted Review in E-Discovery Can Be More
Effective and More Efficient Than Exhaustive Manual
Review*, XVII RICH. J.L. & TECH. 11 (2011),
<http://jolt.richmond.edu/v17i3/article11.pdf>.

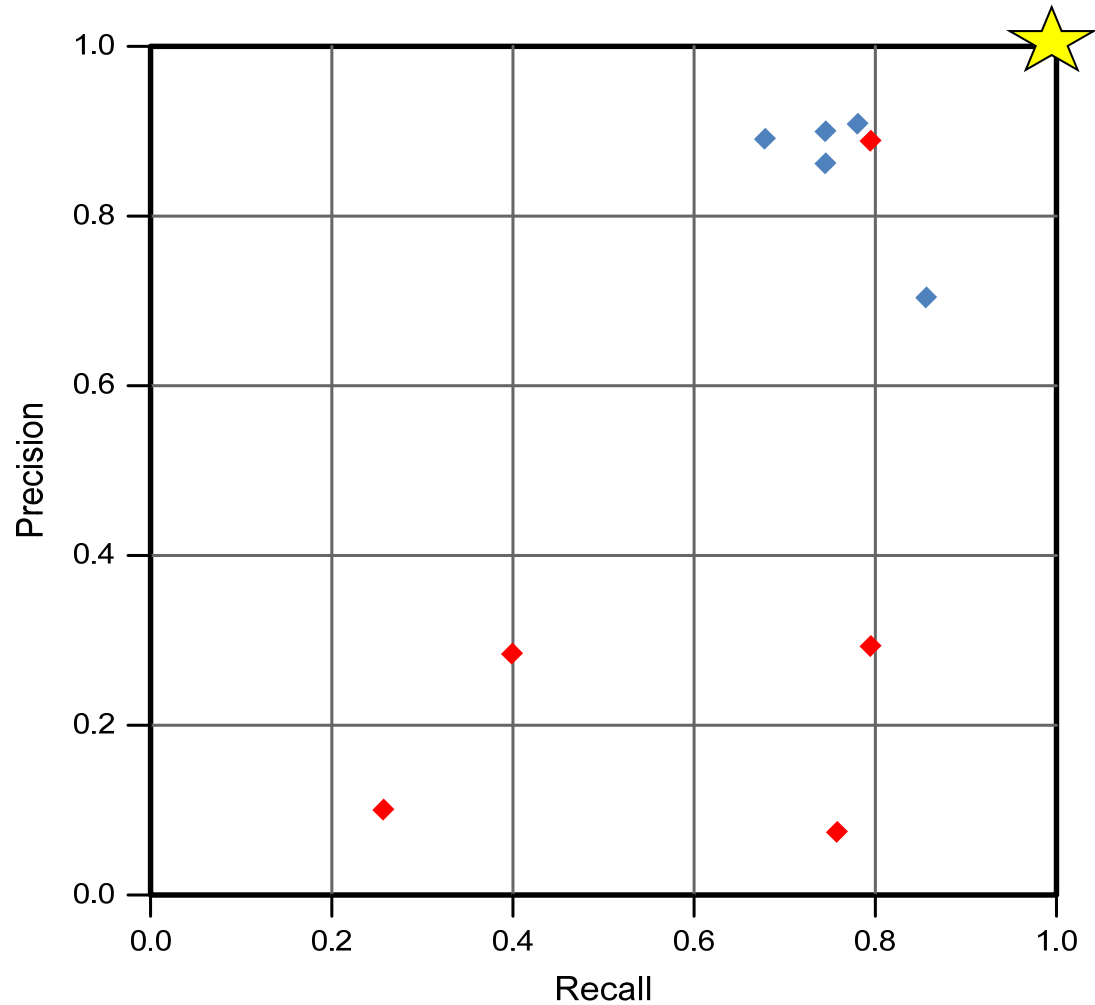
Humans vs. (Two Kinds of) TAR (Cont'd)

Topic	Team	Recall	Precision	F_1
201	Waterloo	(†) 77.8%	(*) 91.2%	(*) 84.0%
	TREC (Law Students)	75.6%	5.0%	9.5%
202	Waterloo	67.3%	(*) 88.4%	(*) 76.4%
	TREC (Law Students)	(†) 79.9%	26.7%	40.0%
203	Waterloo	(*) 86.5%	(*) 69.2%	(*) 76.9%
	TREC (Professionals)	25.2%	12.5%	16.7%
204	H5	(*) 76.2%	(*) 84.4%	(*) 80.1%
	TREC (Professionals)	36.9%	25.5%	30.2%
207	Waterloo	76.1%	(†) 90.7%	82.8%
	TREC (Professionals)	(†) 79.0%	89.0%	(†) 83.7%
Avg.	H5/Waterloo	(†) 76.7%	(*) 84.7%	(*) 80.0%
	TREC	59.3%	31.7%	36.0%

Results marked (*) are superior and overwhelmingly significant ($P < 0.0001$)

Results marked (†) are superior but not statistically significant ($P > 0.1$)

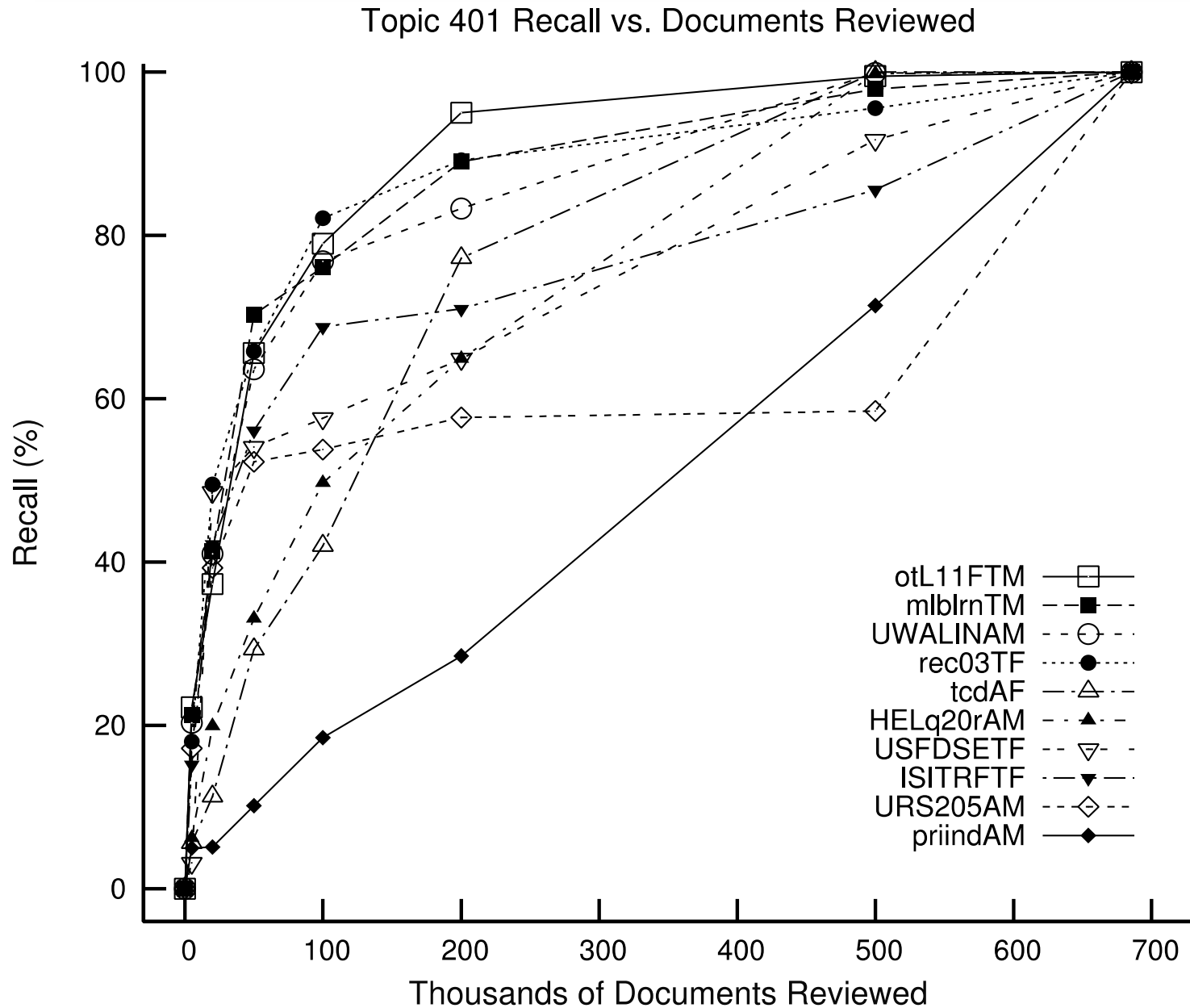
Humans vs. (Two Kinds of) TAR (Cont'd)



◆ TREC 2009 Technology-Assisted Reviews

◆ TREC 2009 Manual Reviews

2011 Legal Track Learning Task





TAR vs. TAR (SPL, SAL, and CAL)

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07.

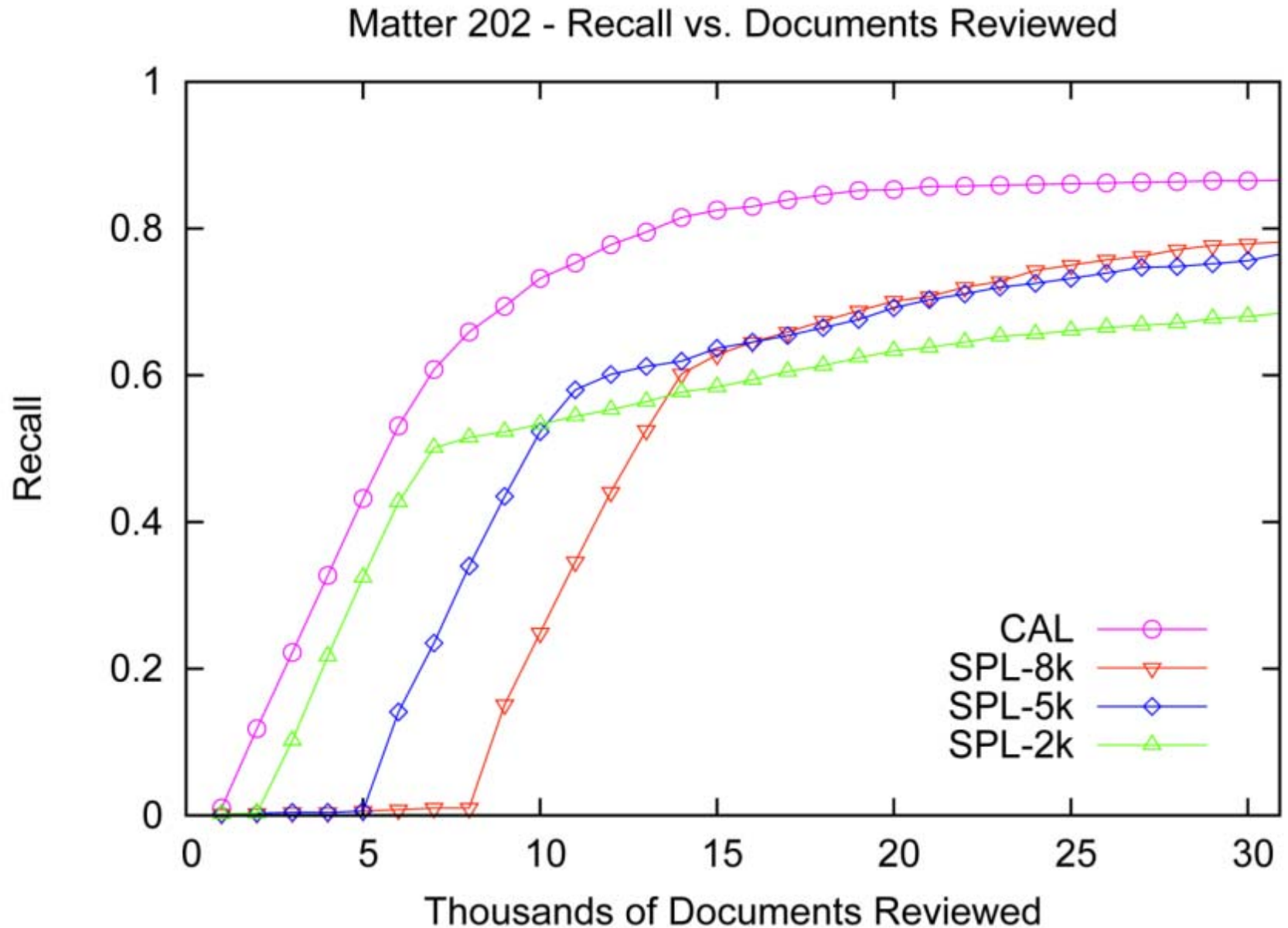
<http://dx.doi.org/10.1145/2600428.2609601>.

Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

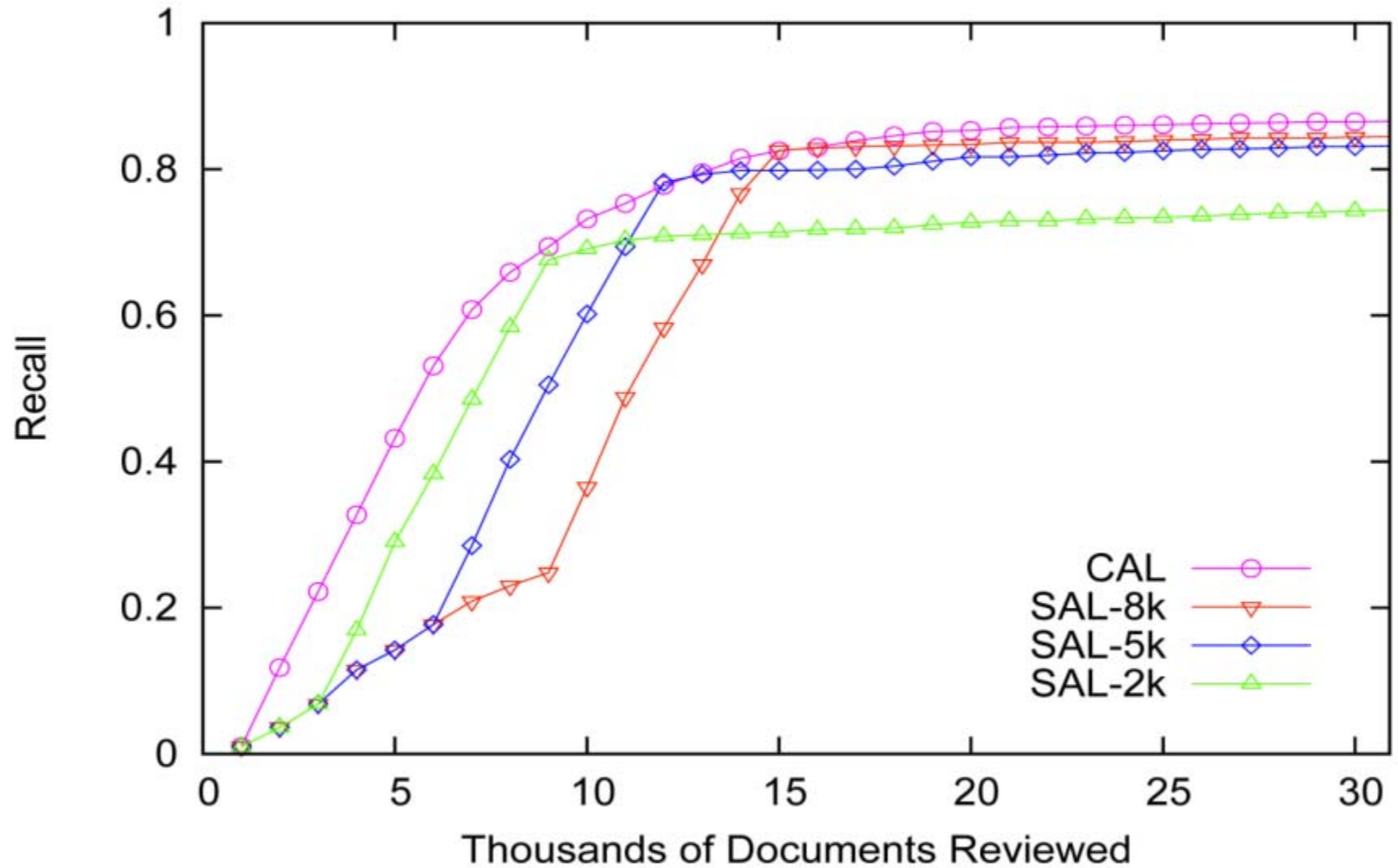
Maura R. Grossman^{*}
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

CAL vs. SPL (SIGIR 2014)



CAL vs. SAL (SIGIR 2014)

Matter 202 - Recall vs. Documents Reviewed



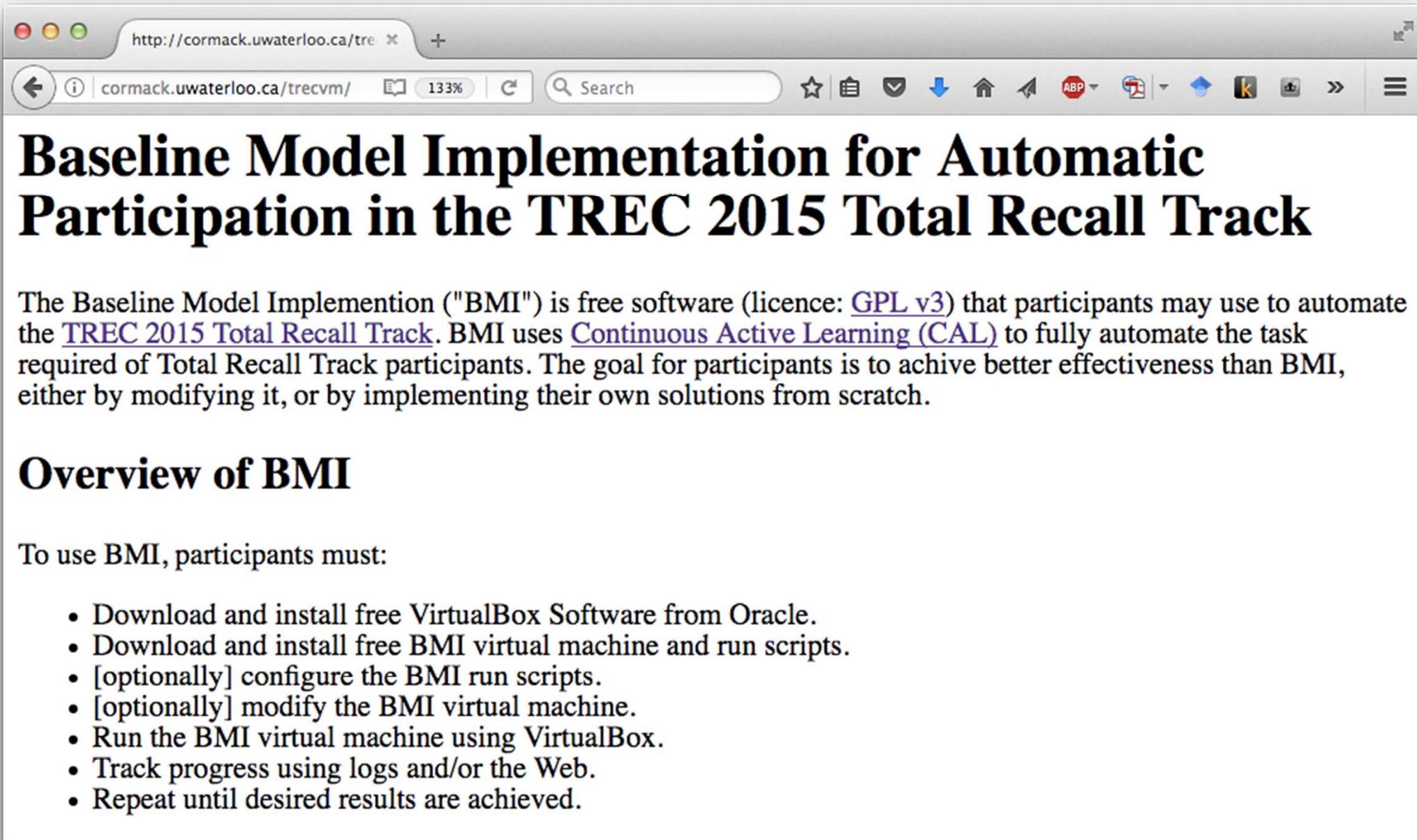
arXiv:1504.06868v1 [cs.IR] 26 Apr 2015

CAL

Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review

GORDON V. CORMACK, University of Waterloo
MAURA R. GROSSMAN, Wachtell, Lipton, Rosen & Katz*

We enhance the autonomy of the continuous active learning method shown by Cormack and Grossman (SIGIR 2014) to be effective for technology-assisted review, in which documents from a collection are retrieved and reviewed, using relevance feedback, until substantially all of the relevant documents have been reviewed. Autonomy is enhanced through the elimination of topic-specific and dataset-specific tuning parameters, so that the sole input required by the user is, at the outset, a short query, topic description, or single relevant document; and, throughout the review, ongoing relevance assessments of the retrieved documents. We show that our enhancements consistently yield superior results to Cormack and Grossman's version of continuous active learning, and other methods, not only on average, but on the vast majority of topics from four separate sets of tasks: the legal datasets examined by Cormack and Grossman, the Reuters RCV1-v2 subject categories, the TREC 6 AdHoc task, and the construction of the TREC 2002 filtering test collection.



The screenshot shows a web browser window with the address bar containing `http://cormack.uwaterloo.ca/tre`. The page title is "Baseline Model Implementation for Automatic Participation in the TREC 2015 Total Recall Track". The main content includes a paragraph about the BMI software and a list of instructions for using it.

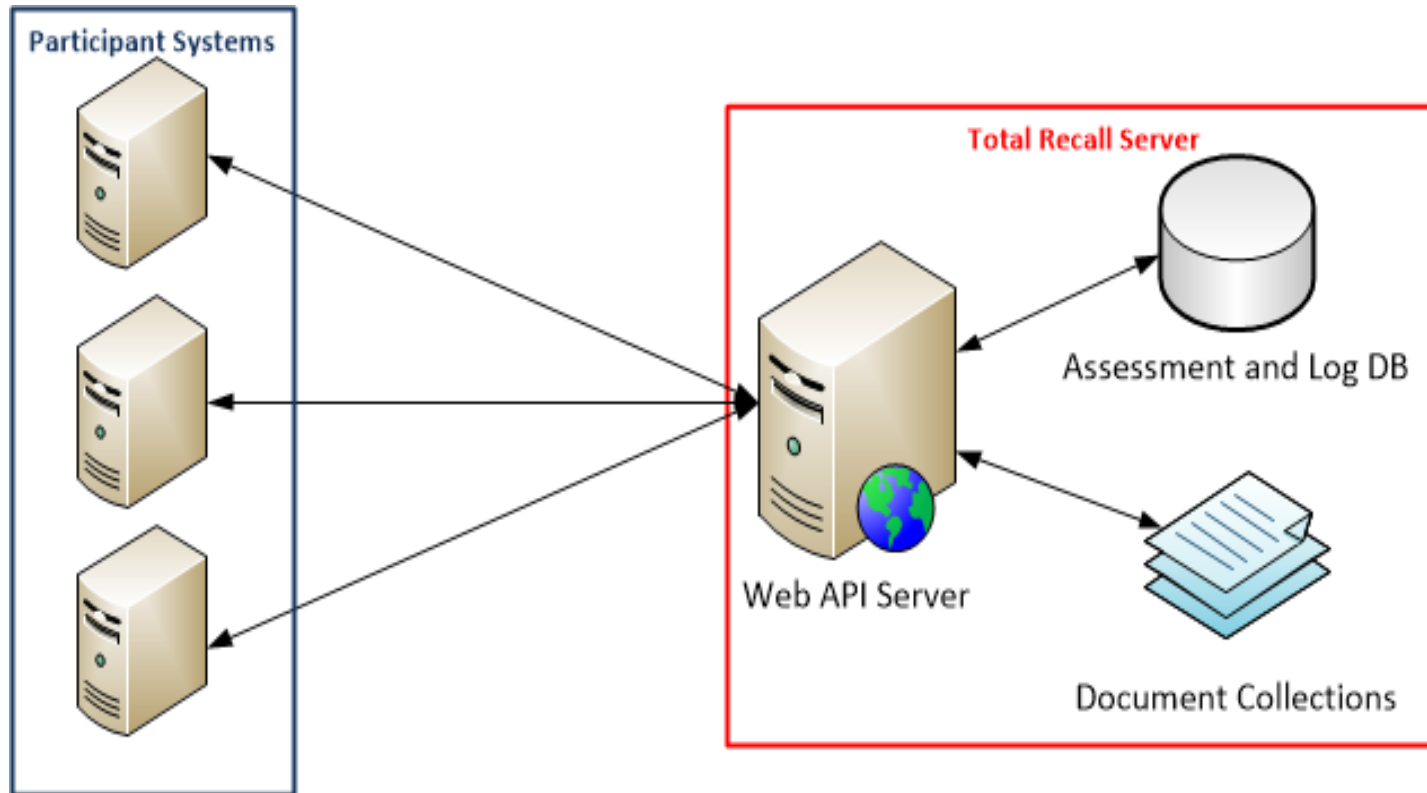
Baseline Model Implementation for Automatic Participation in the TREC 2015 Total Recall Track

The Baseline Model Implementation ("BMI") is free software (licence: [GPL v3](#)) that participants may use to automate the [TREC 2015 Total Recall Track](#). BMI uses [Continuous Active Learning \(CAL\)](#) to fully automate the task required of Total Recall Track participants. The goal for participants is to achieve better effectiveness than BMI, either by modifying it, or by implementing their own solutions from scratch.

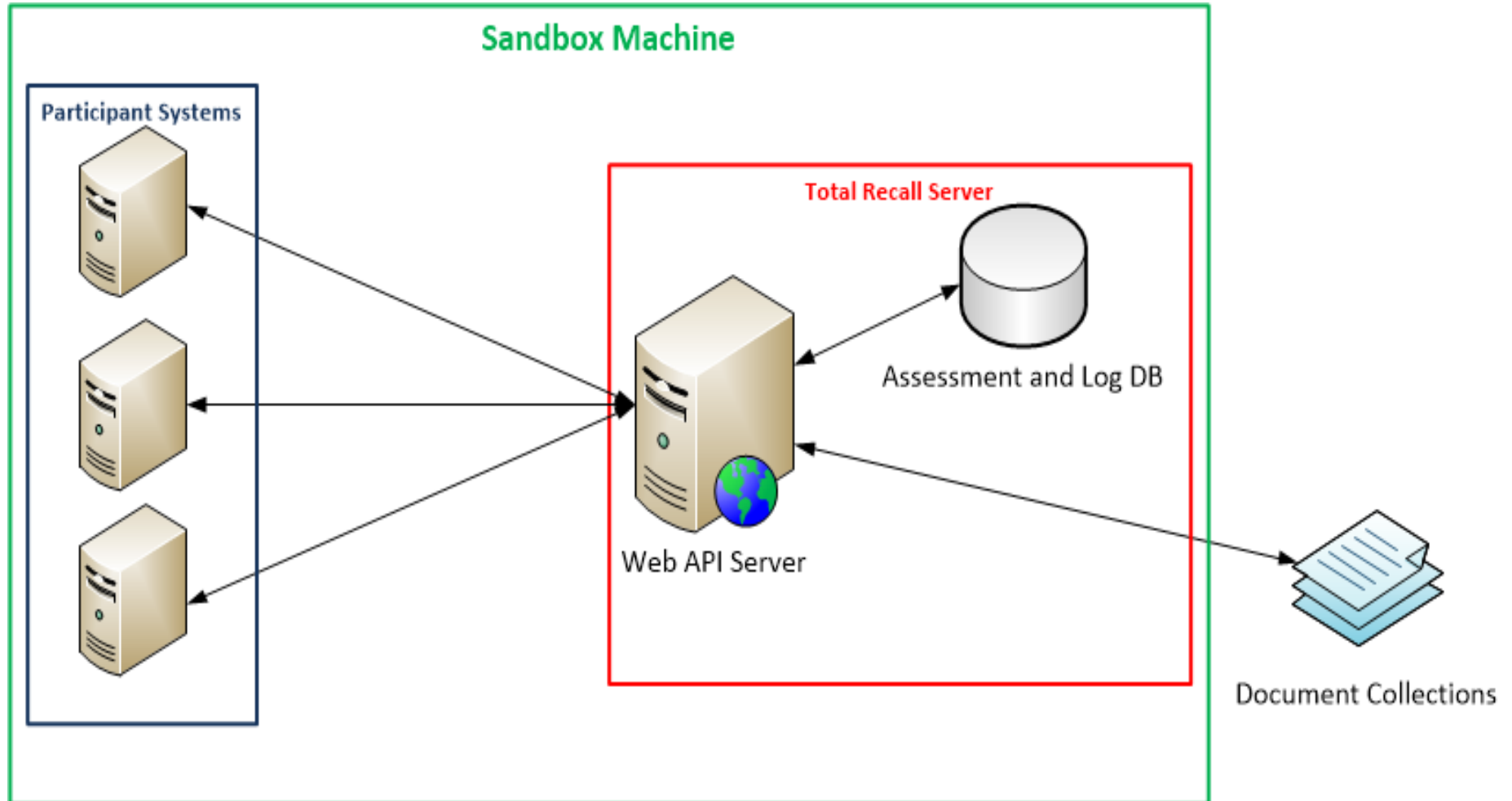
Overview of BMI

To use BMI, participants must:

- Download and install free VirtualBox Software from Oracle.
- Download and install free BMI virtual machine and run scripts.
- [optionally] configure the BMI run scripts.
- [optionally] modify the BMI virtual machine.
- Run the BMI virtual machine using VirtualBox.
- Track progress using logs and/or the Web.
- Repeat until desired results are achieved.



Sandbox Architecture: TREC Total Recall Track



Open Web Datasets (10 topics per dataset):

- Jeb Bush email
- Web crawls of *Blackhat World* and *Hacker Forum*
- Web crawls of Pacific Northwest news sites
- 10 participating teams (2 manual; 8 automatic)

Sandbox Datasets:

- **Virginia Governor Tim Kaine email (at Library of Virginia)**
- MIMIC II Clinical Dataset (at University of Waterloo)
- 6 participating teams (necessarily all automatic)

Virginia Governor Tim Kaine



Virginia Memory: Kaine Email

www.virginiamemory.com/collections/kaine/

virginiamemory
LIBRARY OF VIRGINIA

DIGITAL COLLECTIONS READING ROOM EXHIBITIONS ONLINE CLASSROOM ABOUT VIRGINIA MEMORY

HOME » COLLECTIONS » KAINE

KAINE EMAIL PROJECT @ LVA

Welcome to the Library of Virginia's Kaine Email Project, where we make accessible the email records from the administration of Governor Timothy M. Kaine, Virginia's 70th governor (2006–2010). Users can search and view email records from the Governor's Office and his Cabinet Secretaries; learn about other public records from the Kaine Administration; go behind the scenes to see how the Library of Virginia made the email records available; and read what others are saying about the collection. The Library of Virginia received [approximately 1.3 million email messages](#) from the Kaine Administration. We are processing and releasing these records in batches, so please check back often for new content.



[Search the Collection](#)



[Related Content](#)



[Look Under the Hood](#)



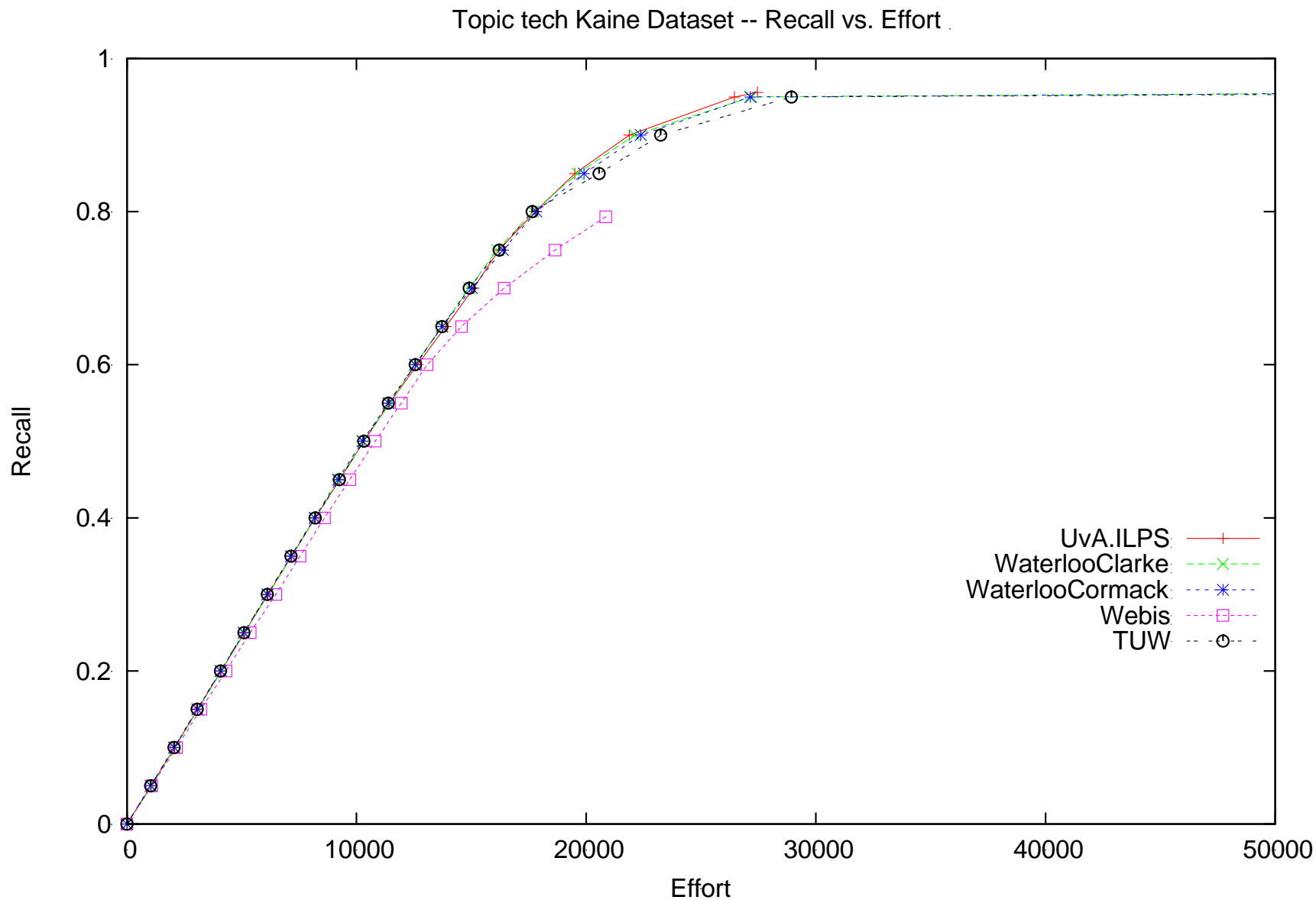
[What's the Buzz](#)



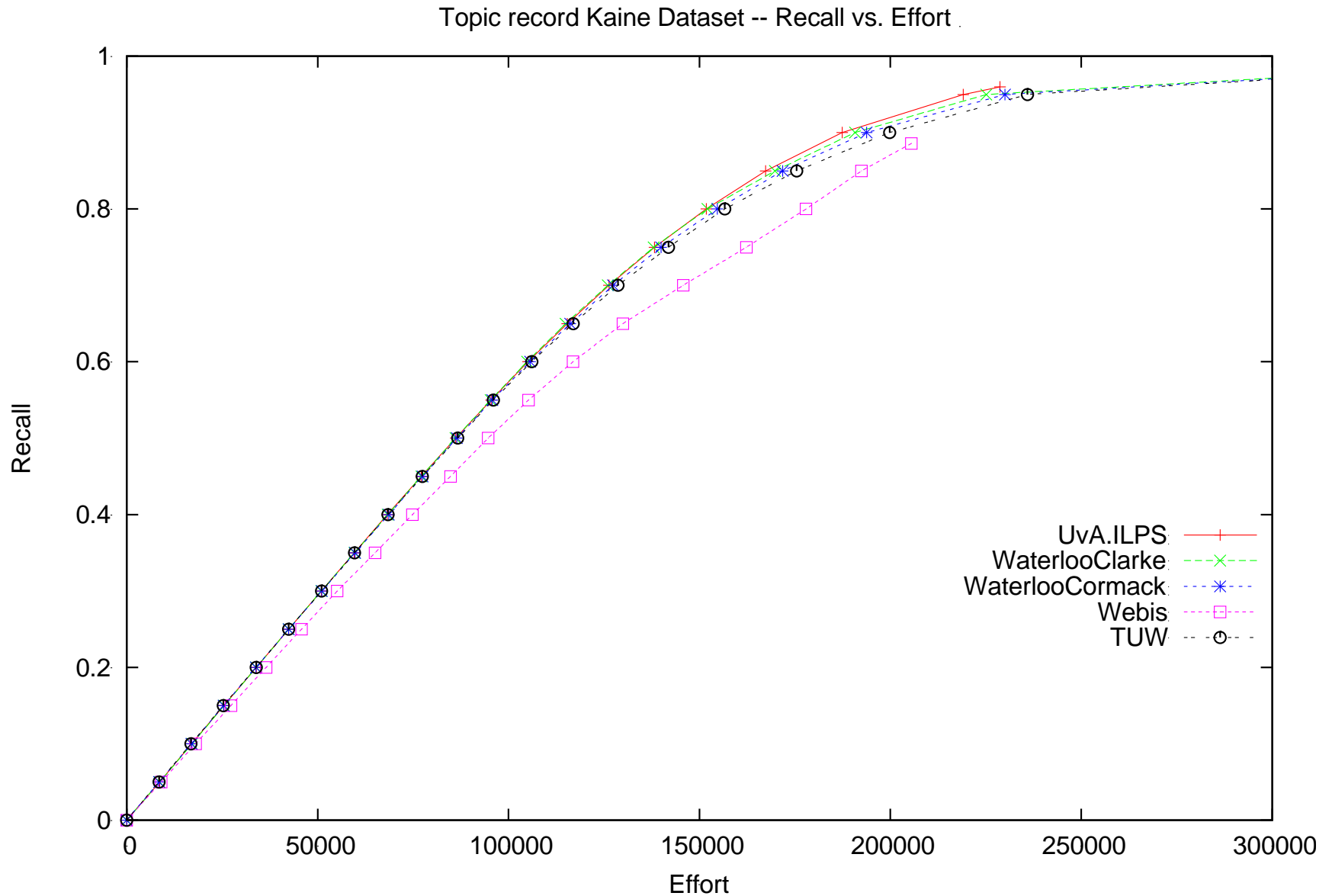
INSTITUTE of
Museum and Library
SERVICES

This project is made possible by federal funding provided through the Library Services and Technology Act program administered by the Institute of Museum and Library Services.

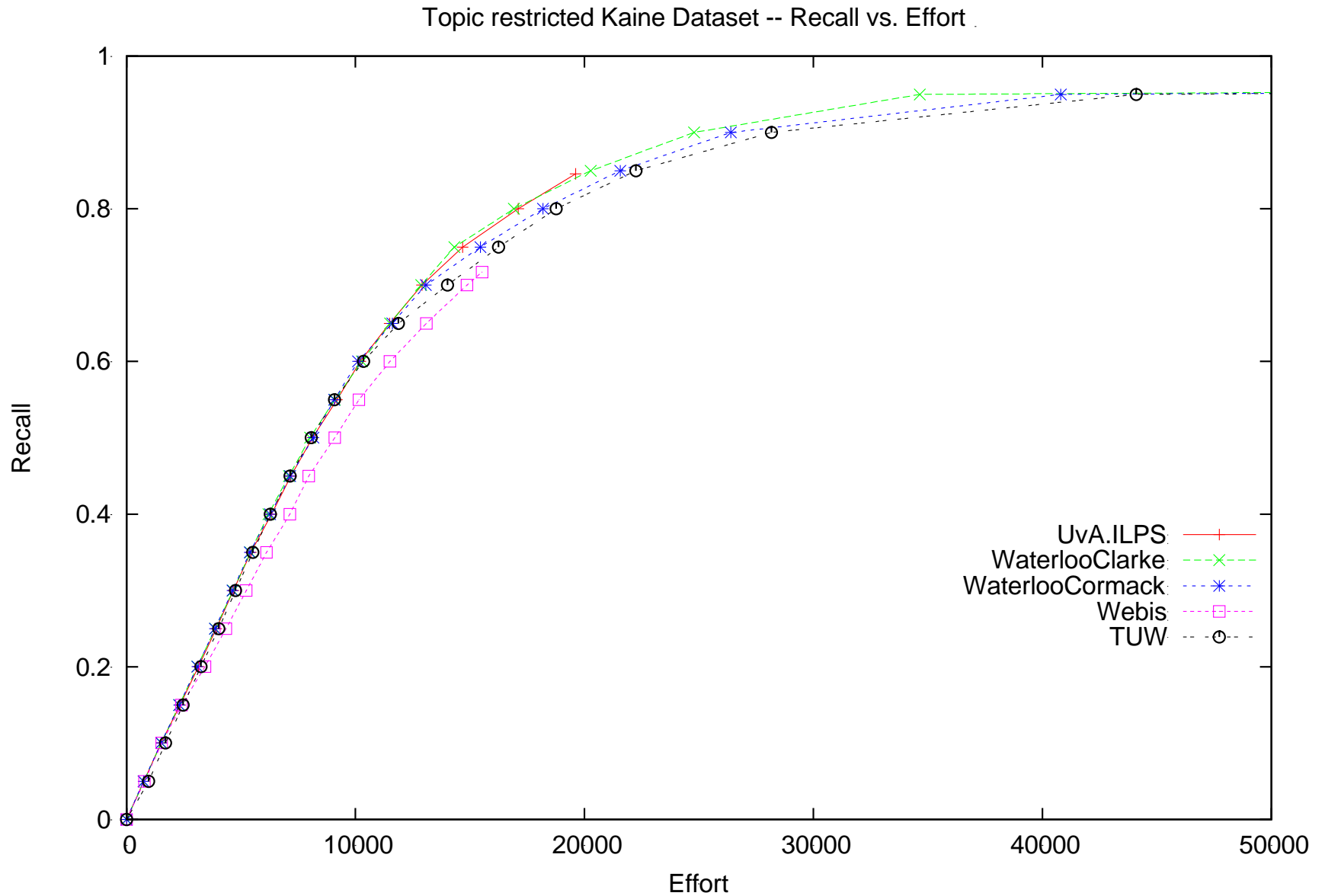
TREC 2015 Kaine VA Tech Shooting



TREC 2015 Kaine Record vs. Non-Record



TREC 2015 Kaine Restricted Records



Determining When to Stop

Gain Curves tell us what might be *if we knew when to stop*

We investigated two stopping procedures (SIGIR 2016):

- Statistical “target method”
- Non-statistical “knee method”

Results* on the Kaine 2015 Total Recall dataset

Topic	Target Method		Knee Method	
	Recall	Effort	Recall	Effort
Legal hold	87.6%	24,232	95.3%	44,085
Record vs. Non-record	92.3%	197,600	98.8%	301,464
Restricted vs. Open Record	94.3%	26,252	98.8%	44,085

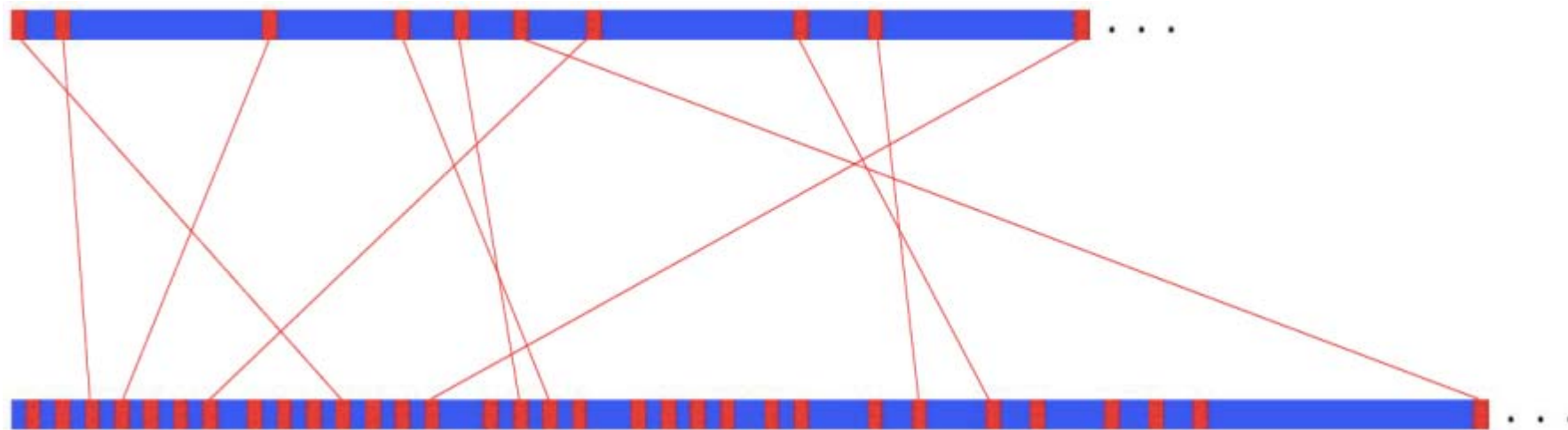
* Predicated on the fiction that human assessment (for user feedback and for evaluation) is infallible.

The Target Method

1. Assess documents selected at random, until ***ten relevant documents*** are identified.
2. Apply any generic TAR method, selecting documents for assessment until every one of the ***same ten relevant documents*** has been assessed.
3. Provided the documents for assessment are **independently selected**, steps 1 and 2 may be transposed or interleaved.
4. With **95% probability**, step 2 will select **at least 70% of the relevant documents**.

The Target Method

Randomly selected order

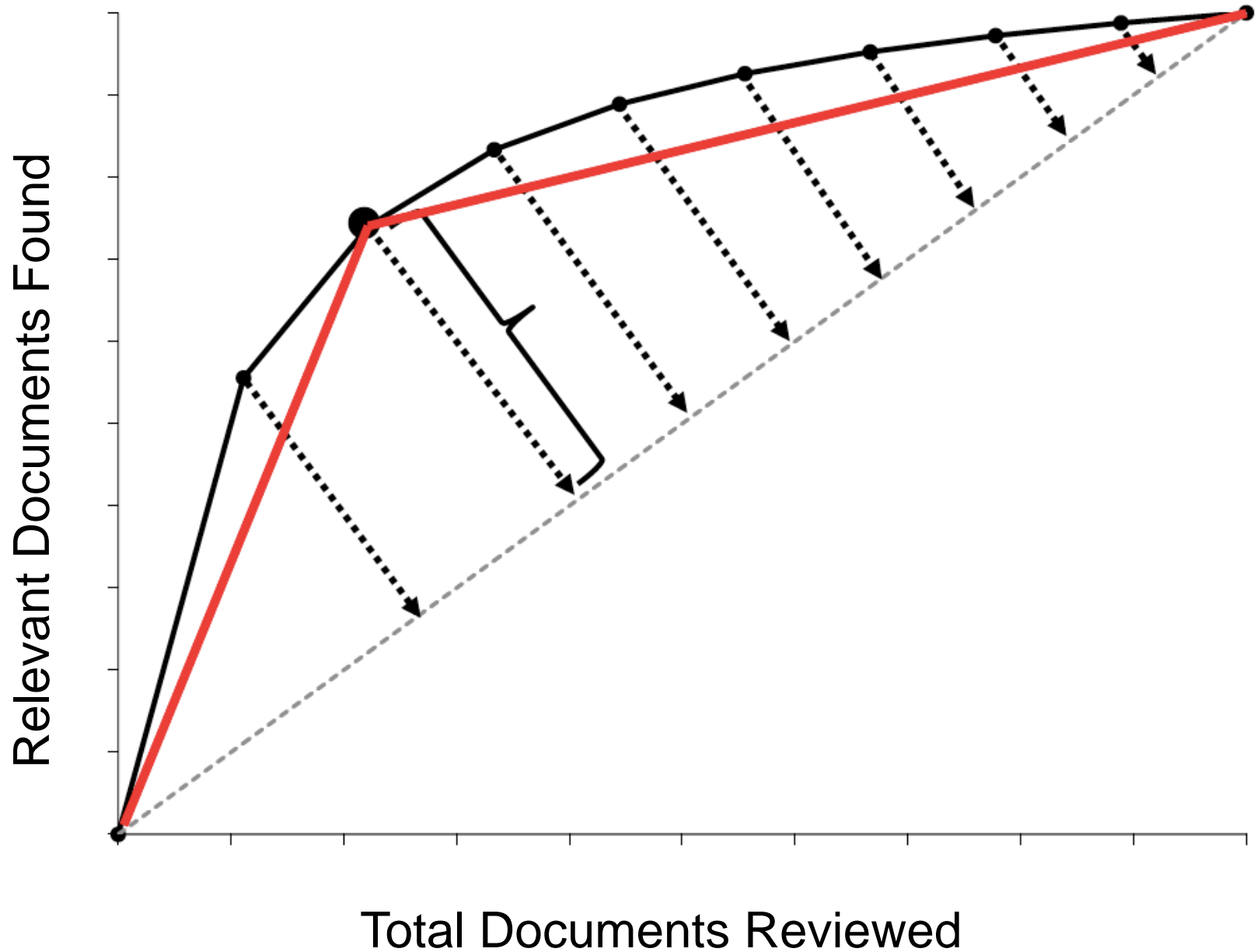


Generic method ranking

The Knee Method

1. Use AutoTAR to select exponentially larger batches of documents for assessment.
2. After assessing each batch, calculate a **gain curve**, plotting the number of relevant documents found as a function of the number of documents reviewed.
3. The **knee** is the point of maximal distance from the diagonal to the gain curve.
4. Stop when the **slope ratio** (*i.e.*, the slope from the origin to the knee divided by the slope from the knee to the extremum) exceeds a pre-determined threshold: $156 - \min(150, \textit{assessed relevant docs})$.

Gain Curve – Knee – Slope Ratio



Three methods were evaluated:

1. **System** – Documents retrieved by the TAR system
2. **User** – Documents retrieved by the TAR system *and* judged relevant by a fallible user
3. **Adjudicated** –
 - Documents retrieved by the TAR system and judged relevant by the user; or
 - Documents retrieved by the TAR system, judged non-relevant by the user, “appealed” by the TAR system, and judged relevant by a second assessor; or
 - Documents retrieved by the TAR system, judged relevant by the user, “appealed” by the TAR system, and judged relevant by a second assessor.

Accounting for Fallible Assessments

To adequately test HRIR methods, we require:

- Realistic (*i.e.*, fallible) simulated feedback
- An independent (and hopefully better) gold standard

We used three sets of Roger's assessments:

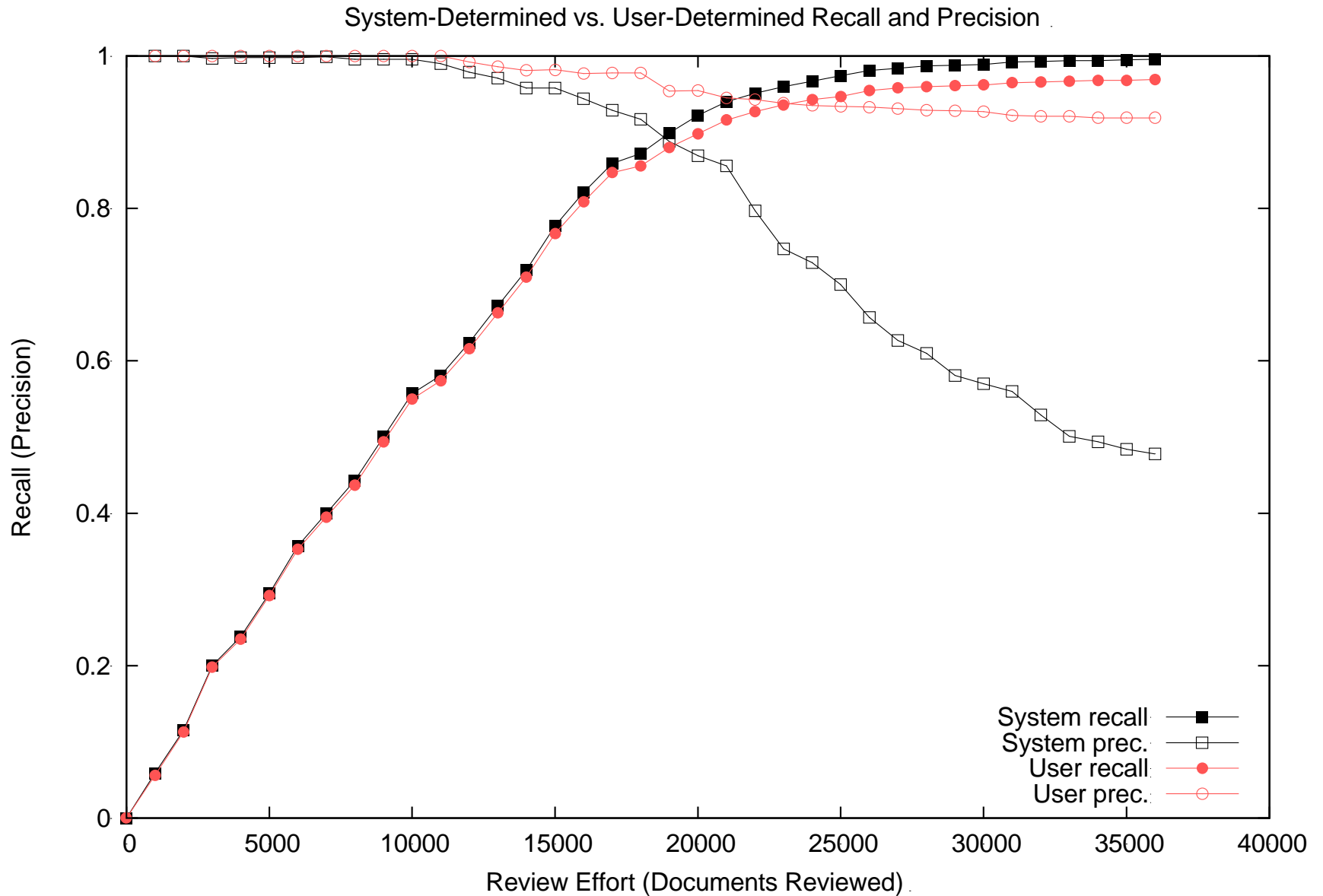
- Roger I = Roger's original assessments
- Roger II = Roger's assessment of a stratified sample, more than 2 years after Roger I
- Roger III = Roger's assessment of all cases of disagreement between Roger I and Roger II

Roger I was used to simulate user feedback

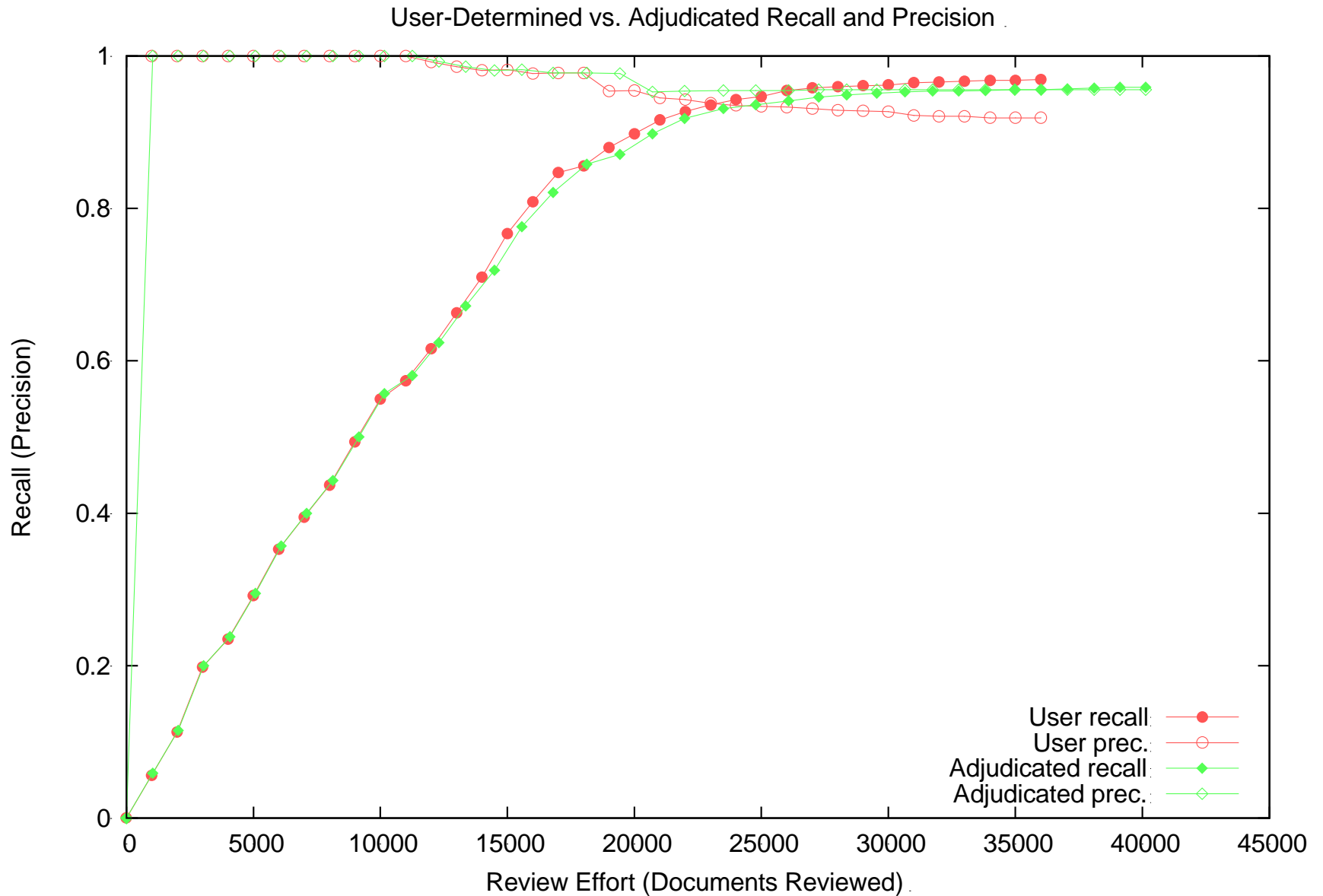
Roger II was used to simulate assessment for quality control

Roger III was used as the gold standard for evaluation

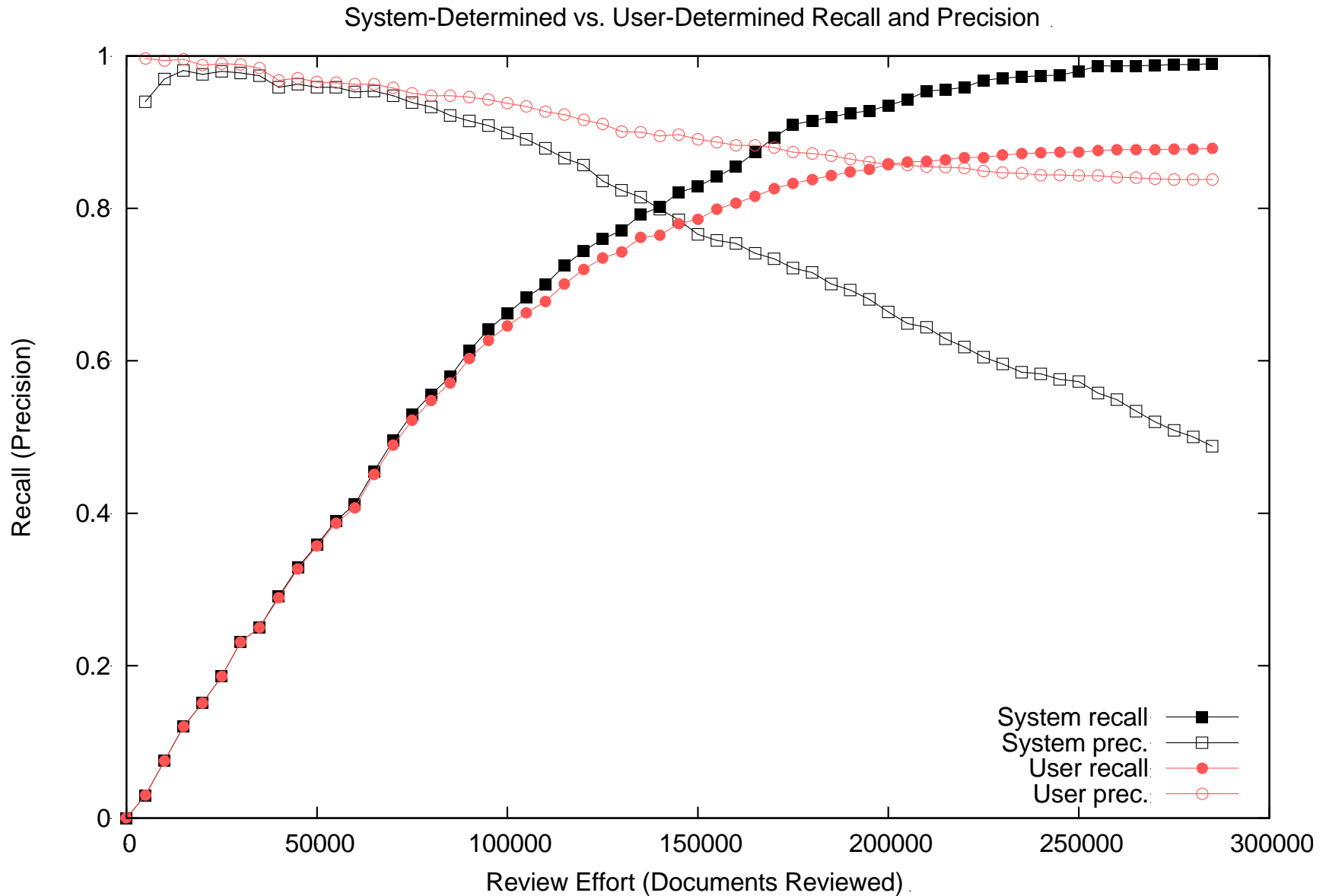
VA Tech Shooting



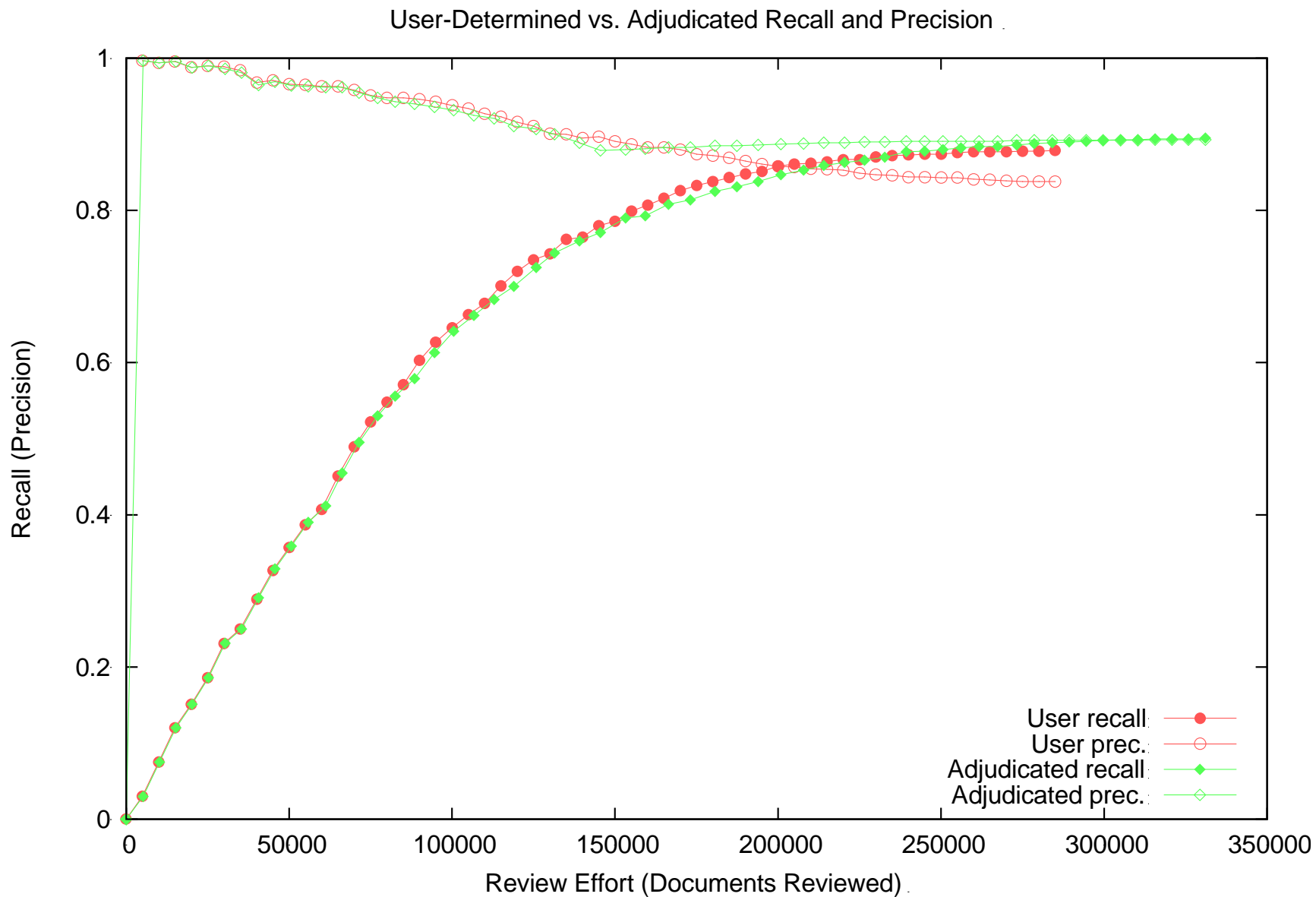
VA Tech Shooting (Adjudicated)



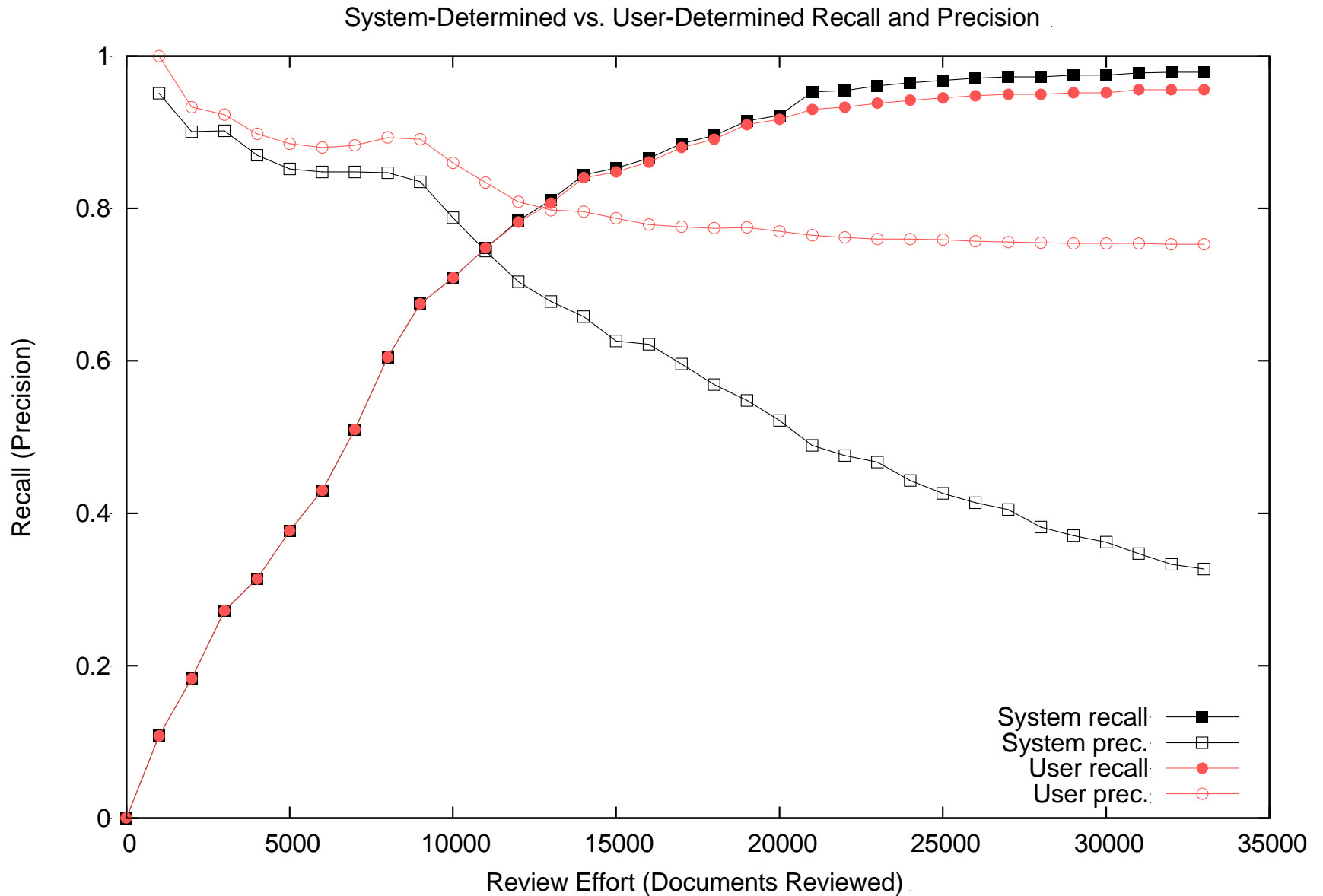
Record vs. Non-Record



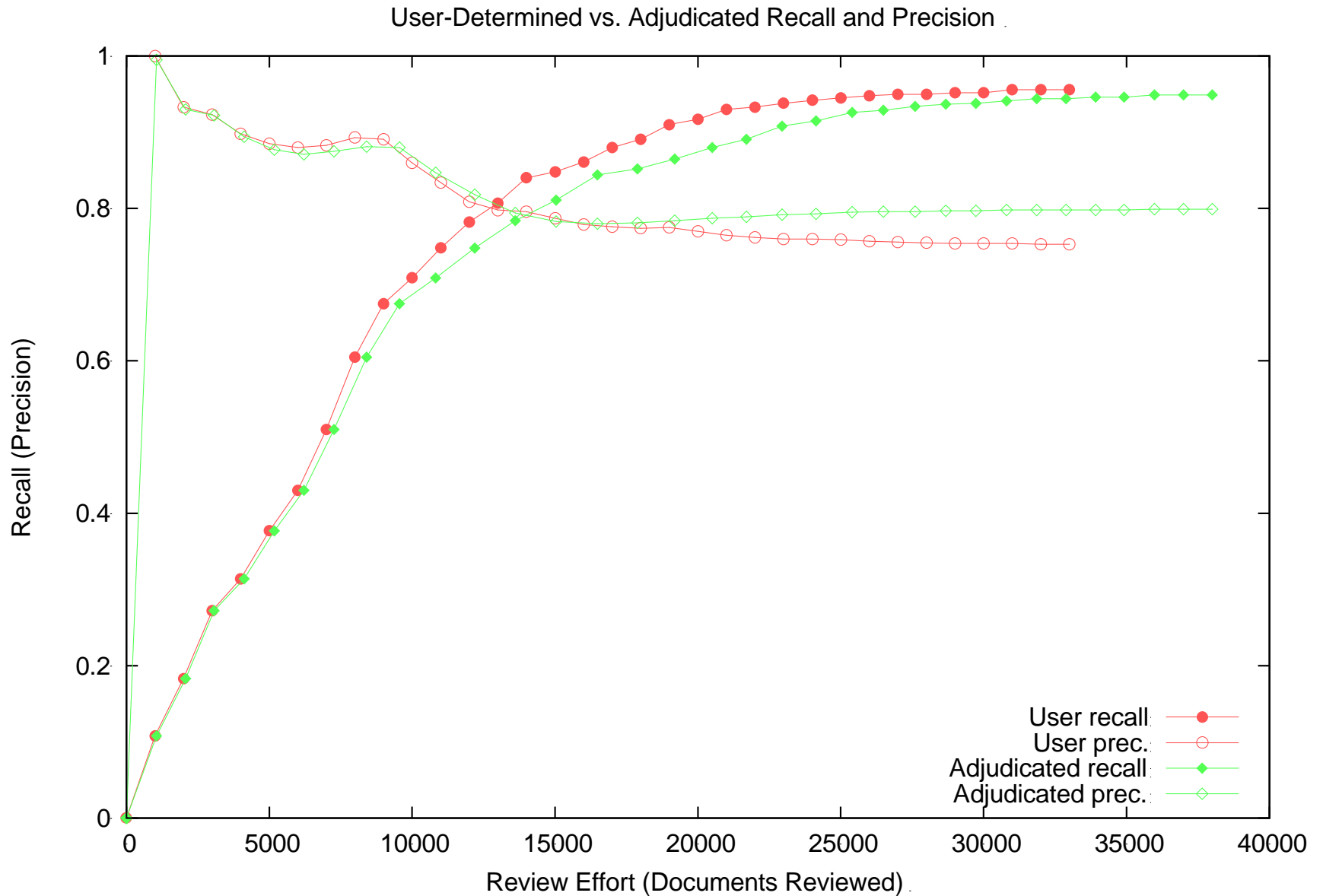
Record vs. Non-Record (Adjudicated)



Restricted Record



Restricted vs. Non-Restricted (Adjudicated)





KAINE EMAIL PROJECT @ LVA

Topic	Manual Review				Adjudicated			
	Recall	Precision	F_1	Effort	Recall	Precision	F_1	Effort
Legal Hold	0.97	0.91	0.94	401,960	0.96	0.96	0.96	40,522
Archival	0.89	0.84	0.86	381,819	0.90	0.89	0.89	332,410
Restricted	0.98	0.75	0.84	146,594	0.95	0.80	0.87	38,048

Replication on the TREC 4 Dataset

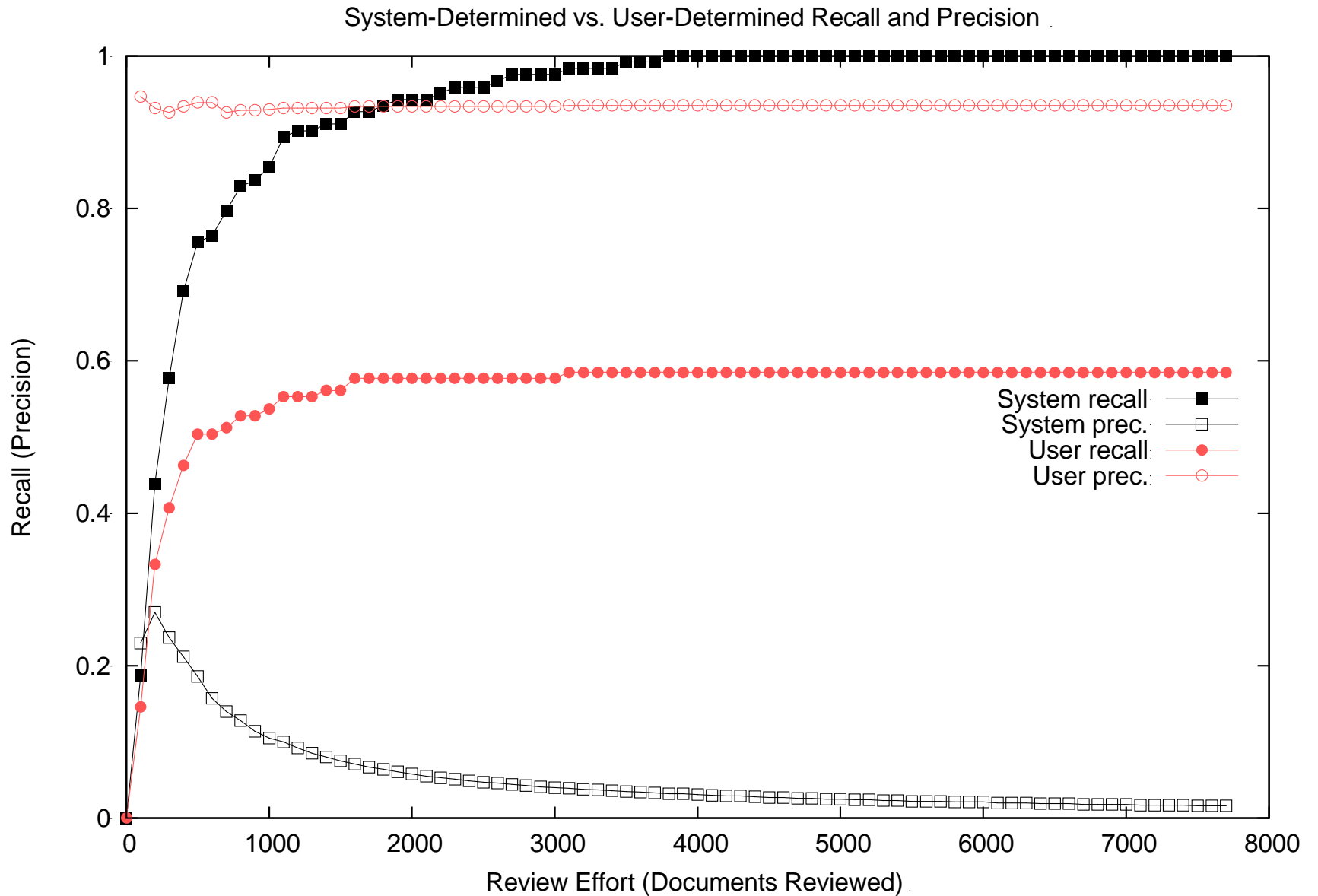
From Voorhees 2000 (Info. Proc. & Mgmt.)

Three NIST assessments for each document:

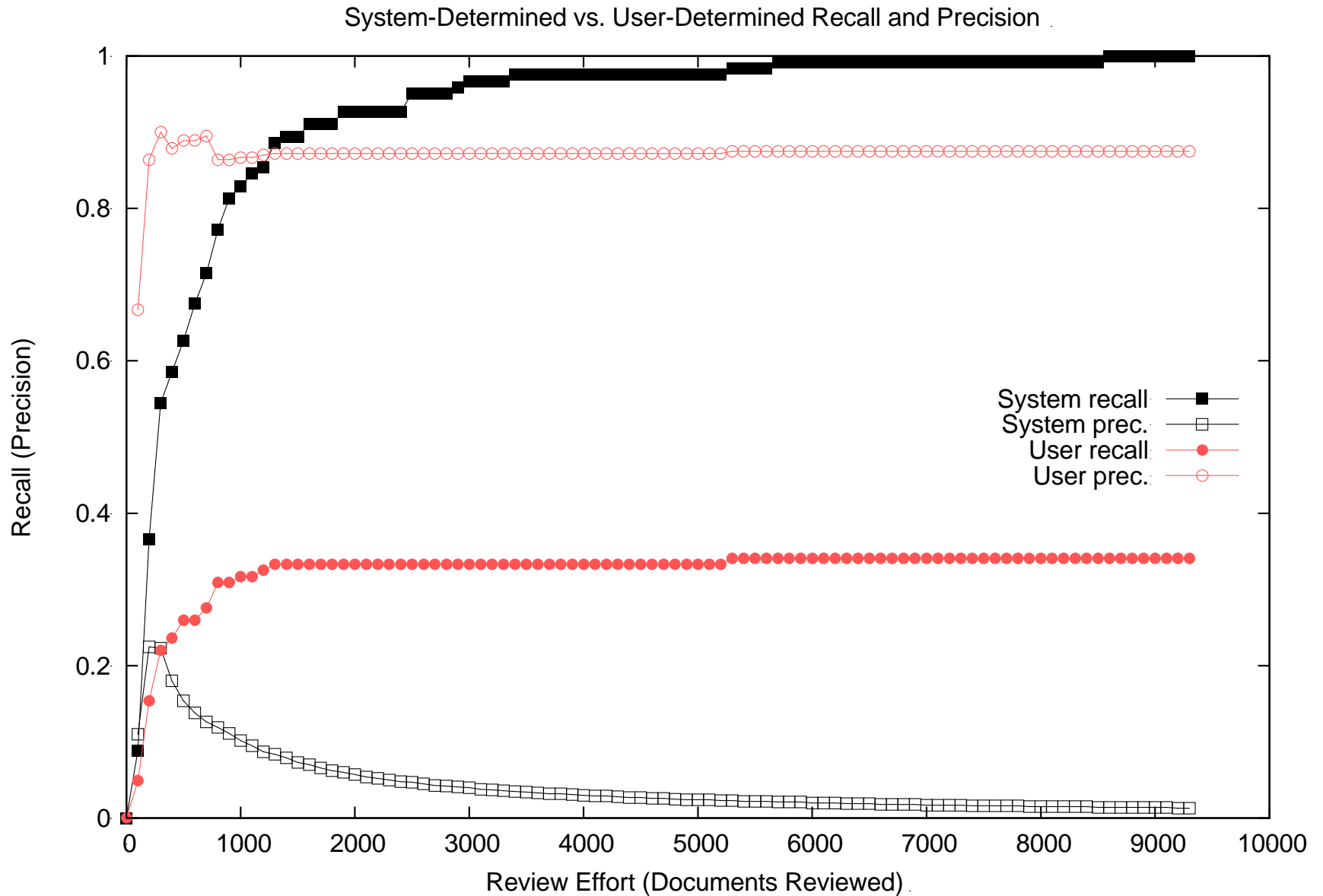
- Primary assessor (used as the gold standard)
- Two secondary assessors (used as the user and the adjudicator)
- 48 topics, two runs for each topic, swapping user and alternate assessor roles

Strategy	Recall	Precision	F_1	Effort
Manual	0.57	0.69	0.63	567,528
System	0.94	0.06	0.10	22,911
User	0.55	0.81	0.62	22,911
Adjudicated	0.64	0.82	0.69	23,662

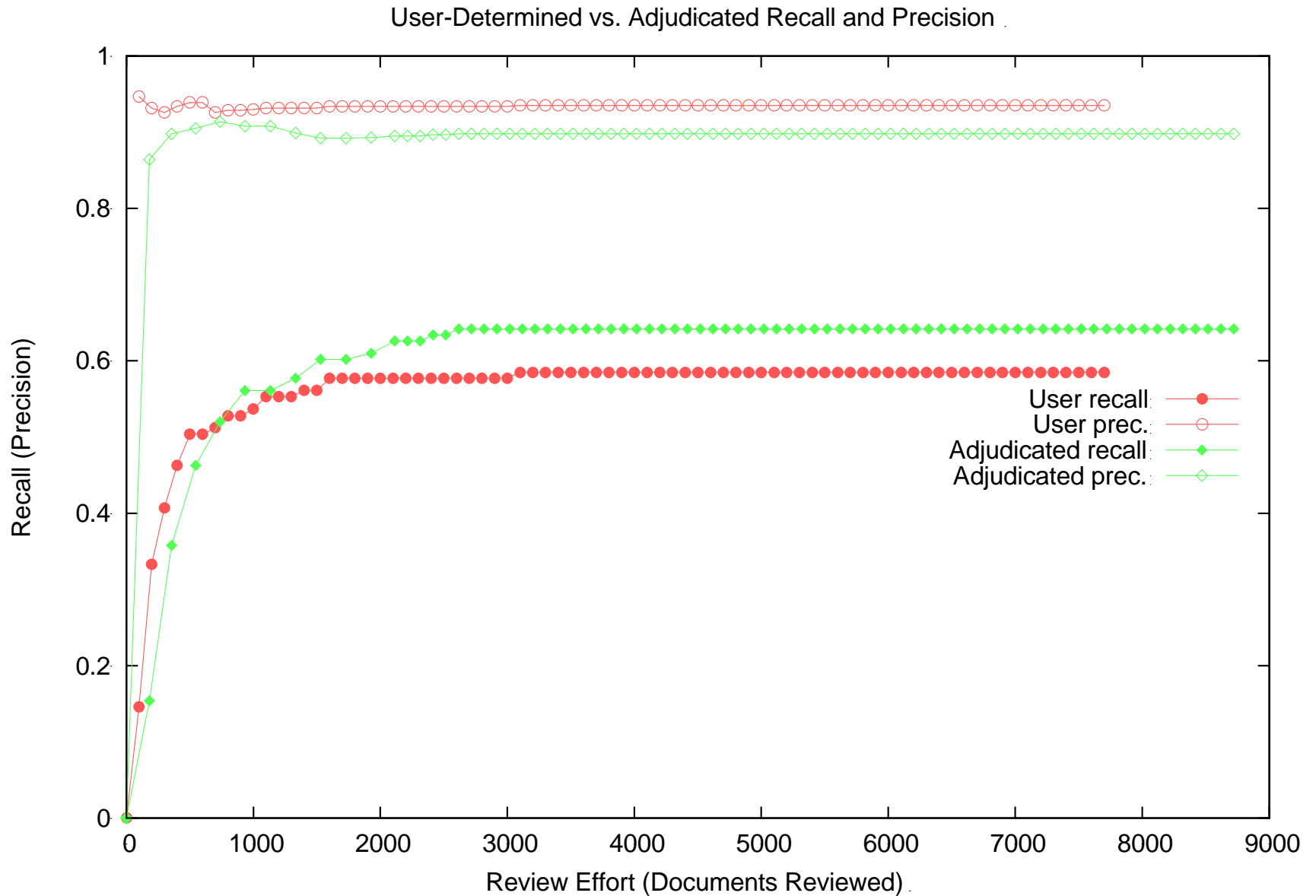
TREC 4 (Alternate Assessor 1)



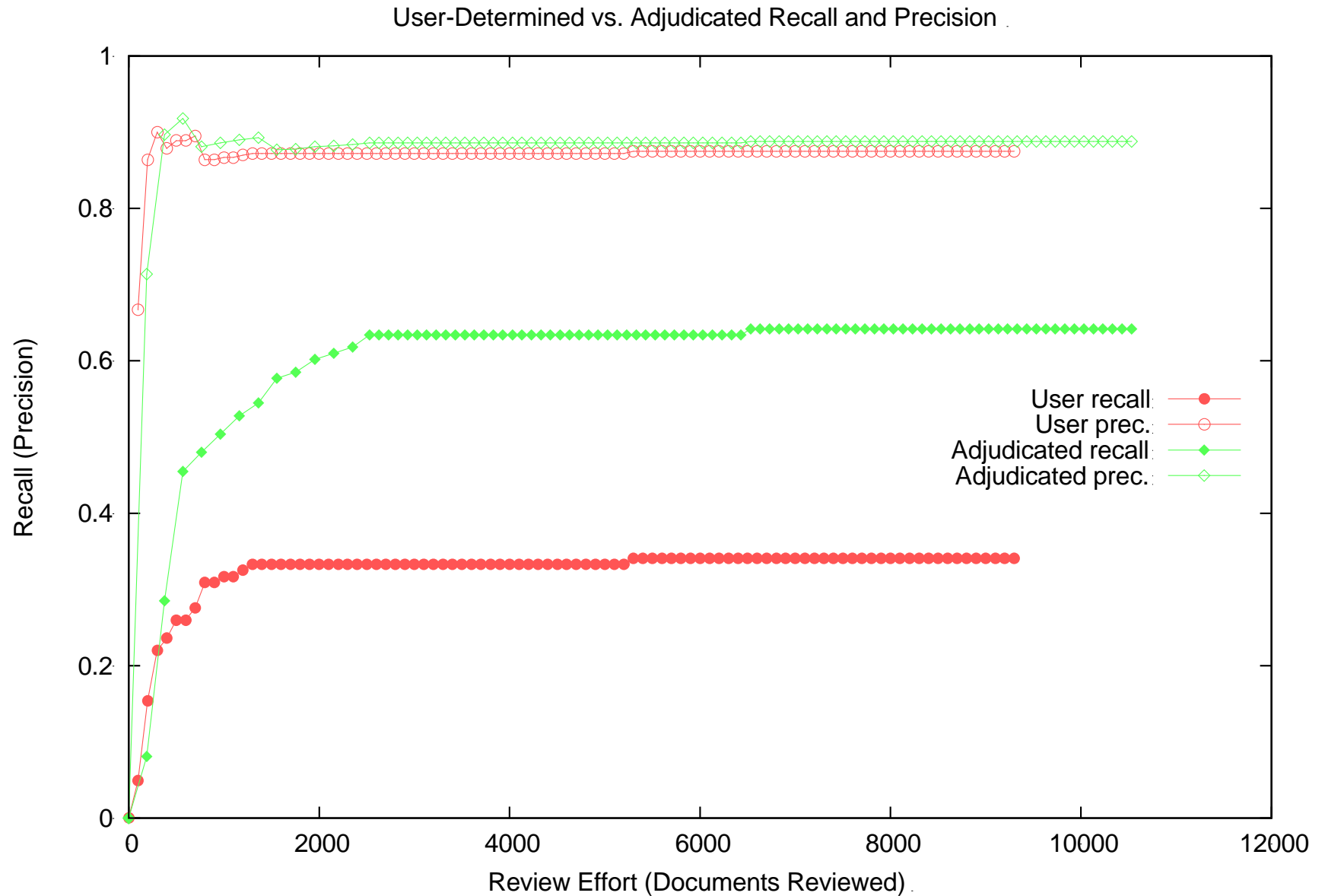
TREC 4 (Alternate Assessor 2)



TREC 4 (Alt. 1 Adjudicated by Alt. 2)



TREC 4 (Alt. 2 Adjudicated by Alt. 1)



Similarities (*accord* Roitblat et al. 2010)

- Recall: Too close to call
- Precision: TAR superior
- Efficiency: TAR superior

Differences

- Uncontrolled human input vs. AutoTAR
- Volunteer vs. officially rendered expert coding
- Open vs. blind relevance adjudication



© 2014 American Psychological Association

0096-3445/14/\$12.00 <http://dx.doi.org/10.1037/xge0000033>

Journal of Experimental Psychology: General

Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

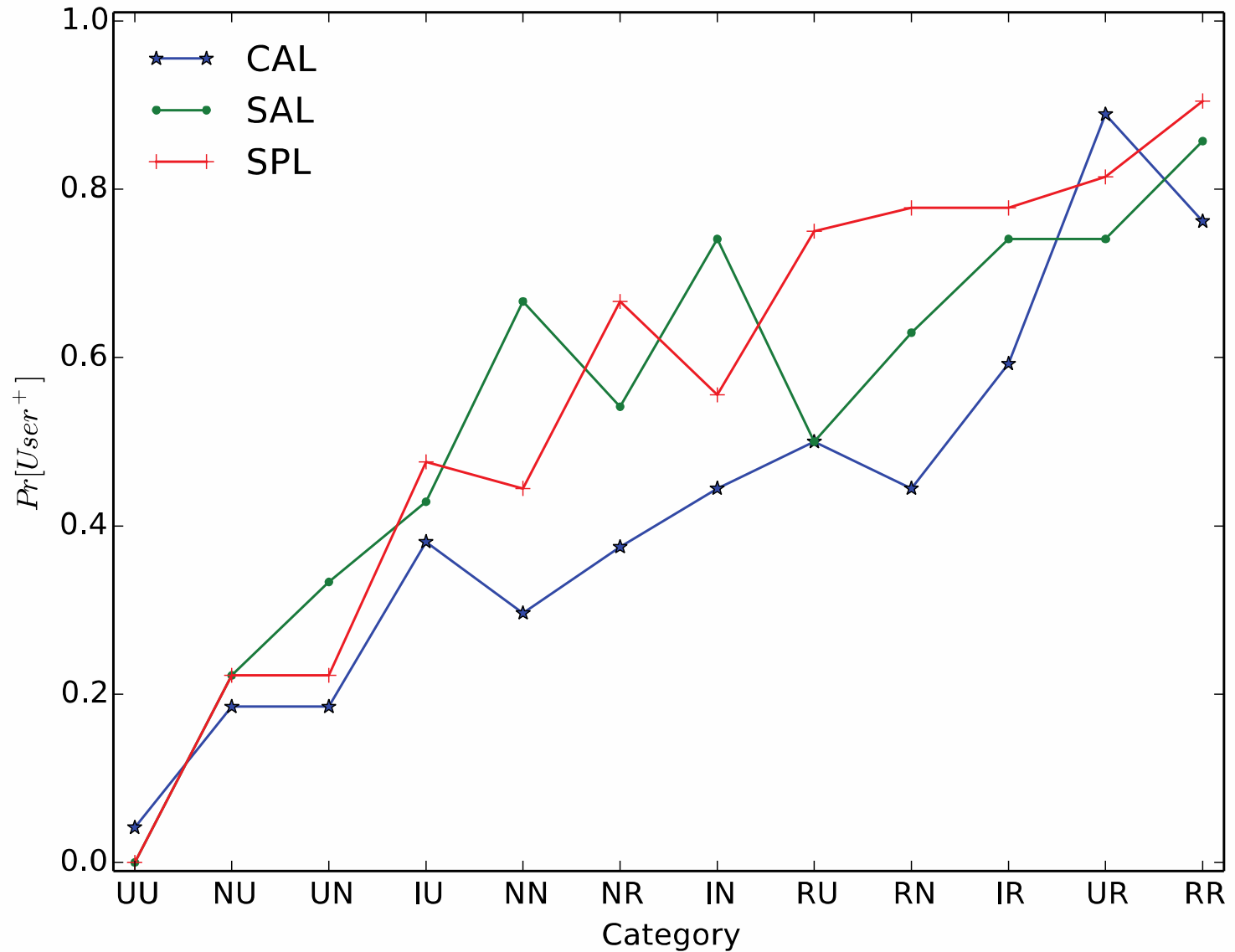
Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion*, is costly, and it is important to understand its causes. We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In 5 studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human. Participants who saw the algorithm perform were less confident in it, and less likely to choose it over an inferior human forecaster. This was true even among those who saw the algorithm outperform the human.

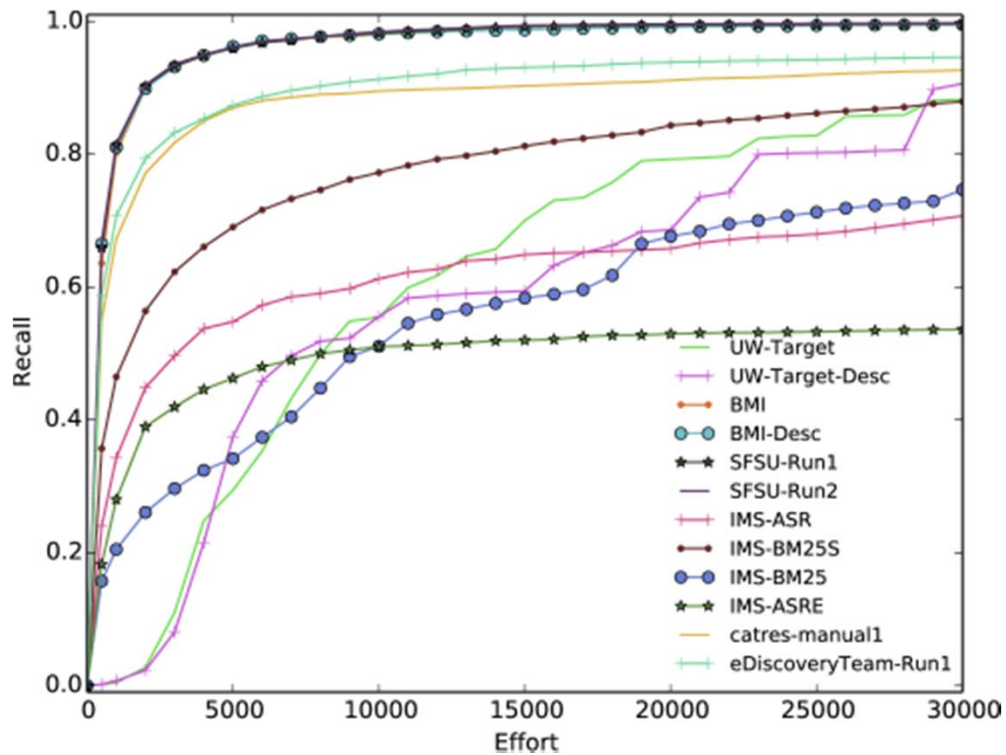
Conclusions and Moving Ahead

1. We have to stop kidding ourselves that the only problem with human review is scalability and we have to stop fearing that we will sacrifice effectiveness for efficiency if we use TAR.
2. TAR has shown promise for privacy-preserving HRIR applications.
3. We have to move away from evaluation paradigms that assume an infallible user and use the same assessments for training and for evaluation purposes.
4. We need more research into how we can do privacy-preserving HRIR tasks better, how we can demonstrate their superiority or improvement, and how we can overcome the human tendency towards algorithm aversion.

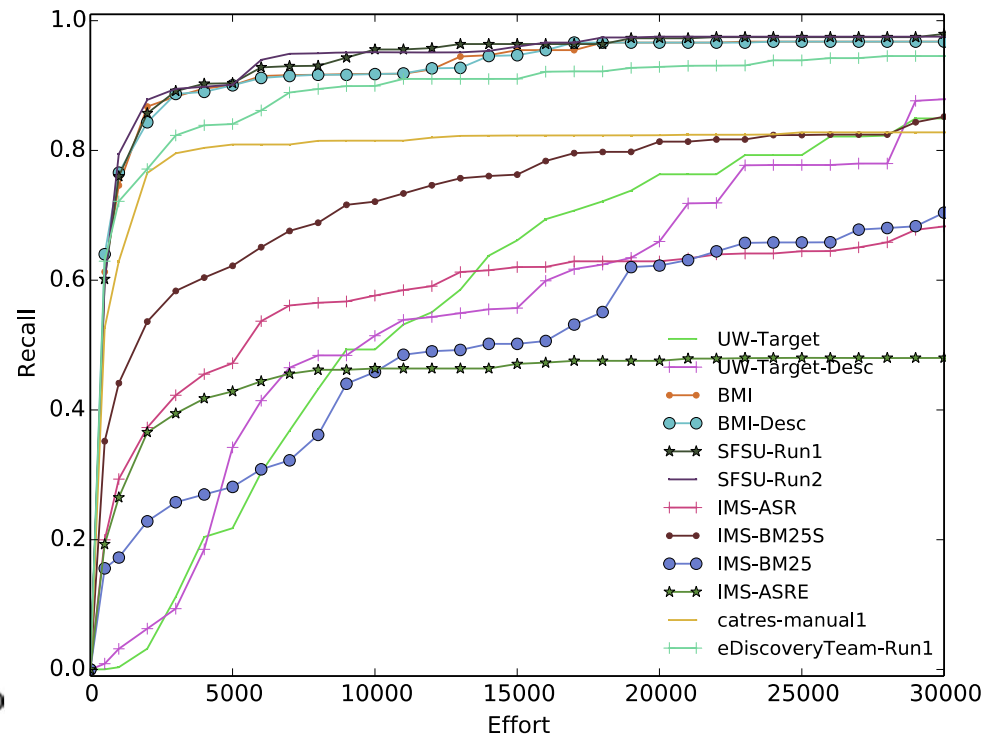
Relevance is Relative!



TREC 2016 Primary vs. 3 Alternates

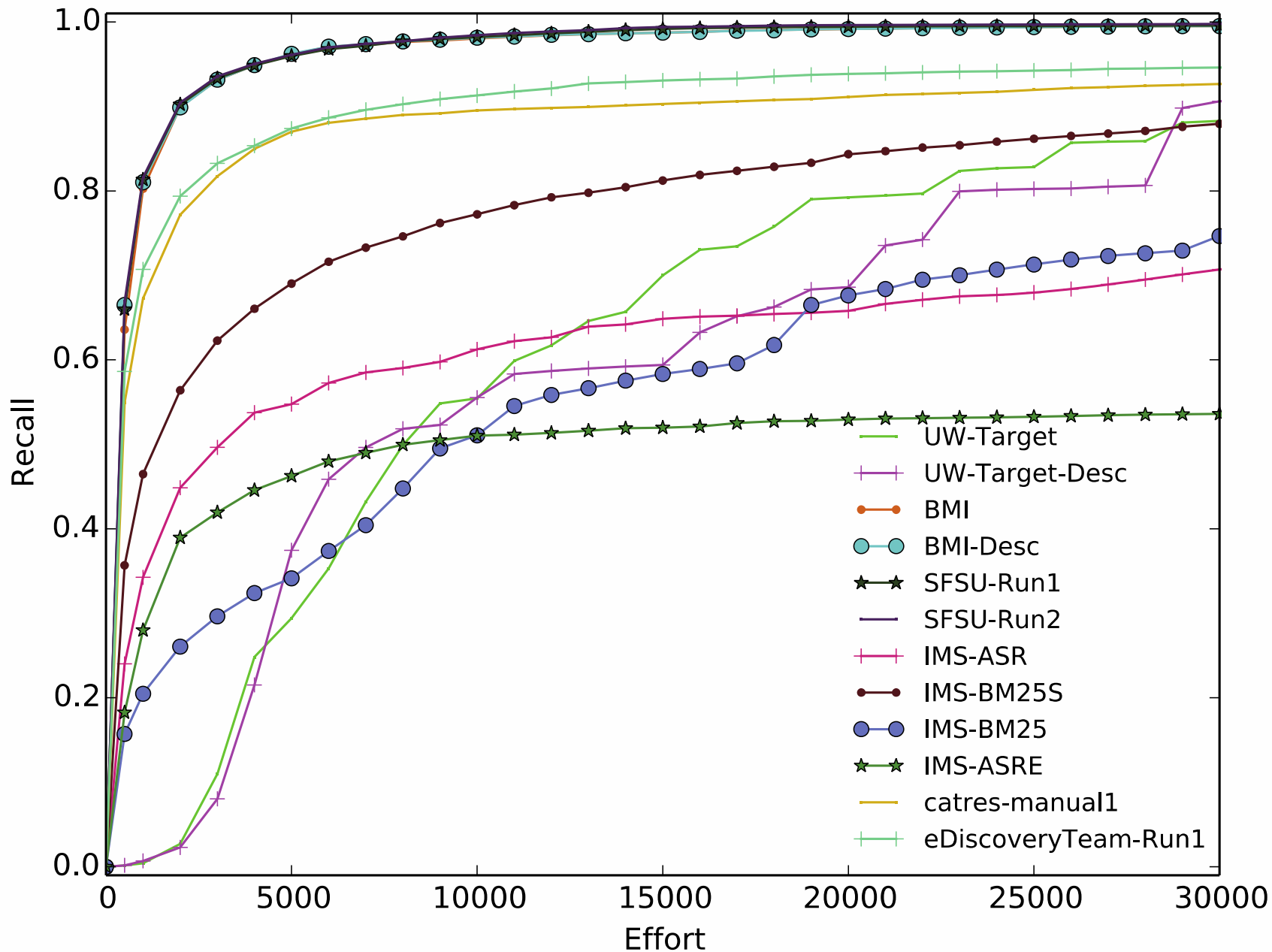


Primary Assessments

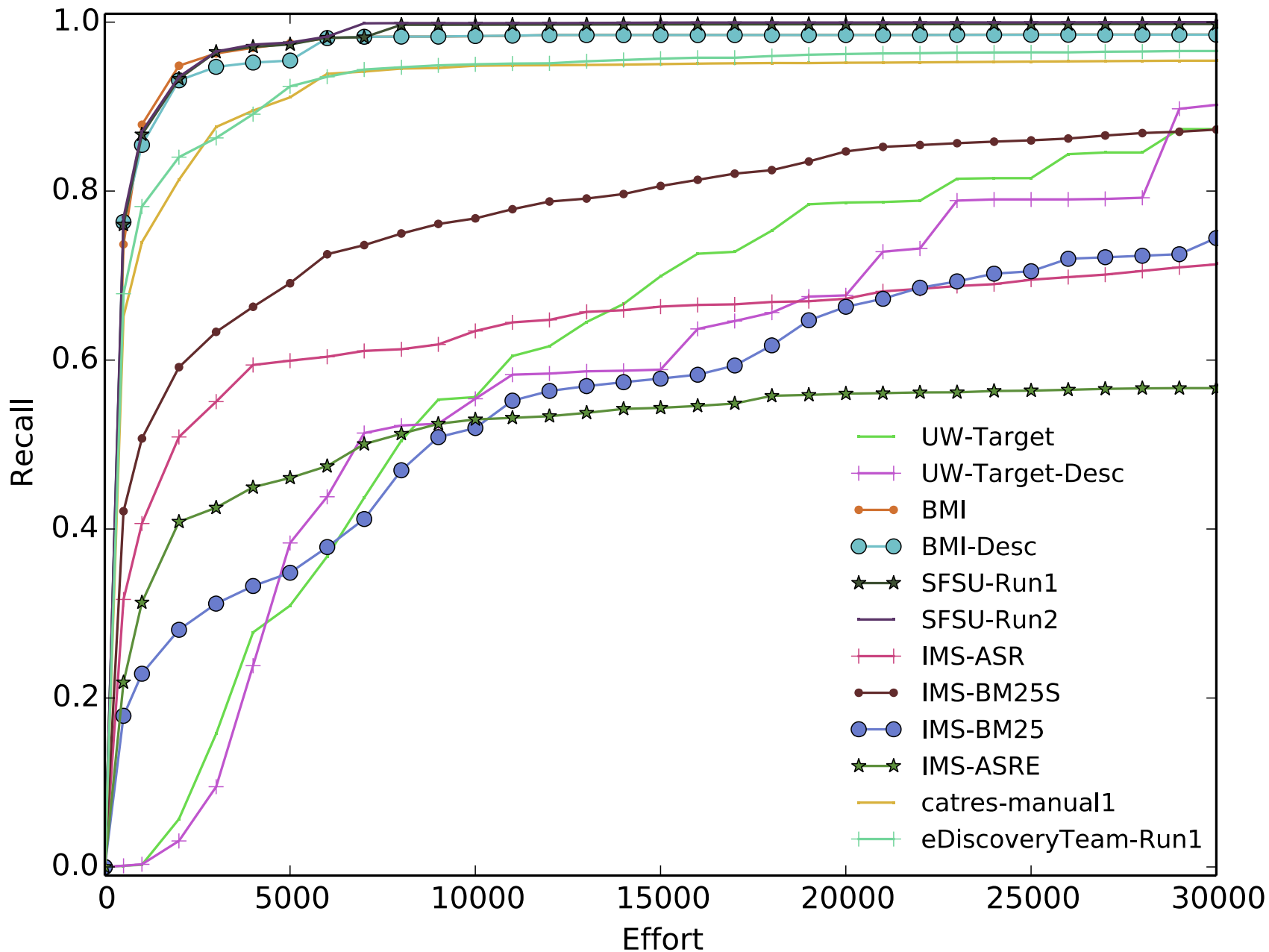


Alternate Majority Vote

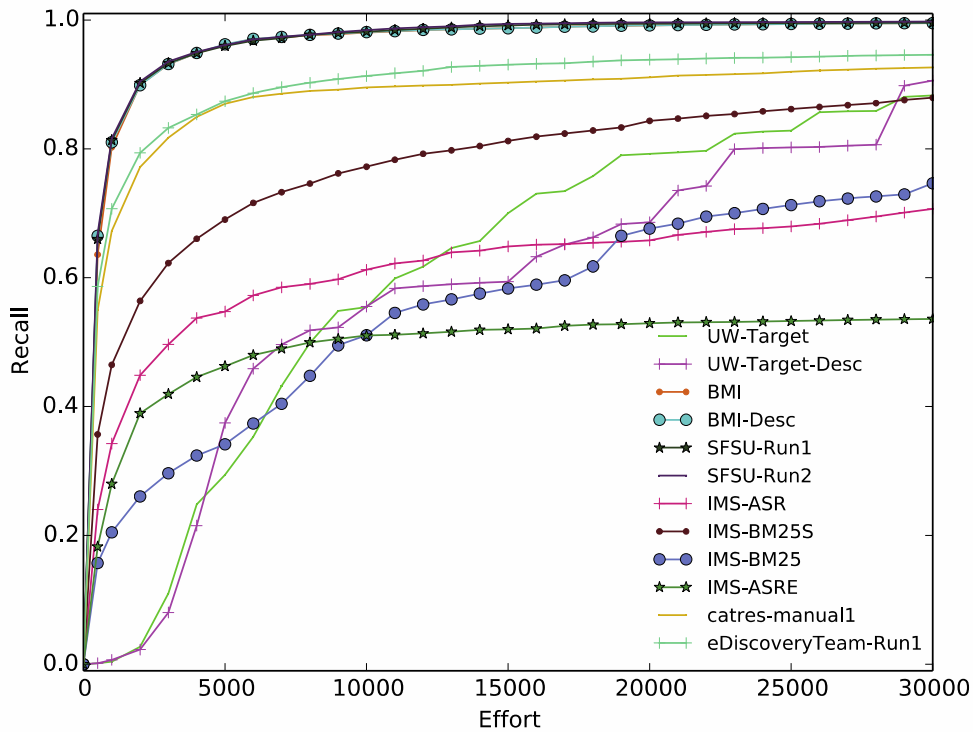
TREC 2016 Total Recall Track: Overall Recall



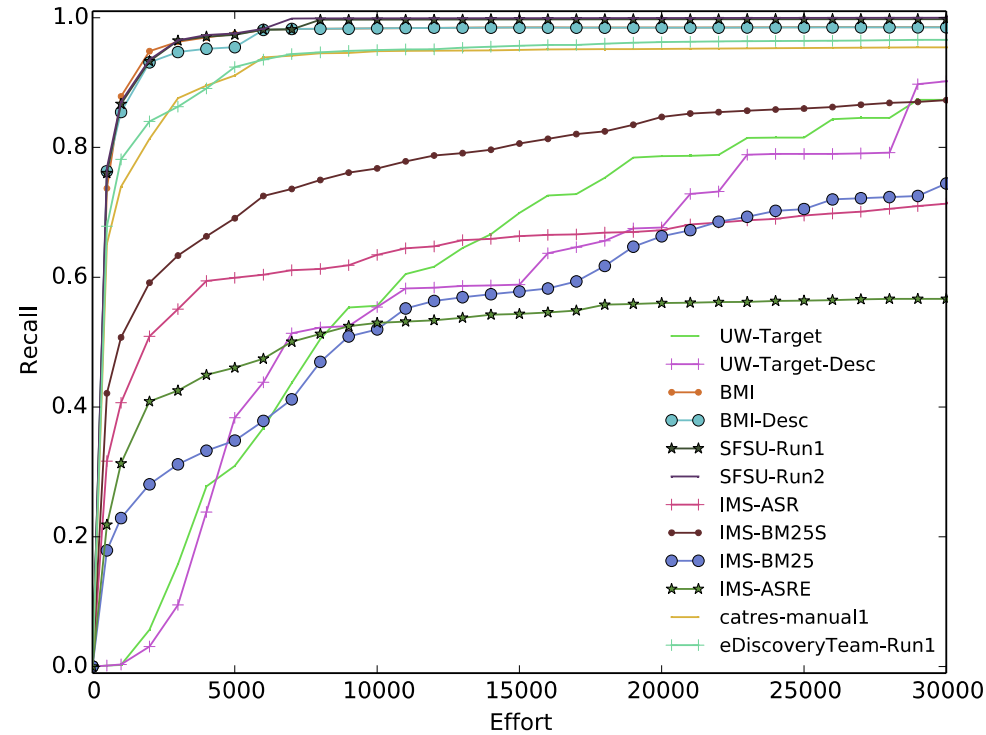
Recall for Important Documents Only



Overall Recall vs. Important Documents Only

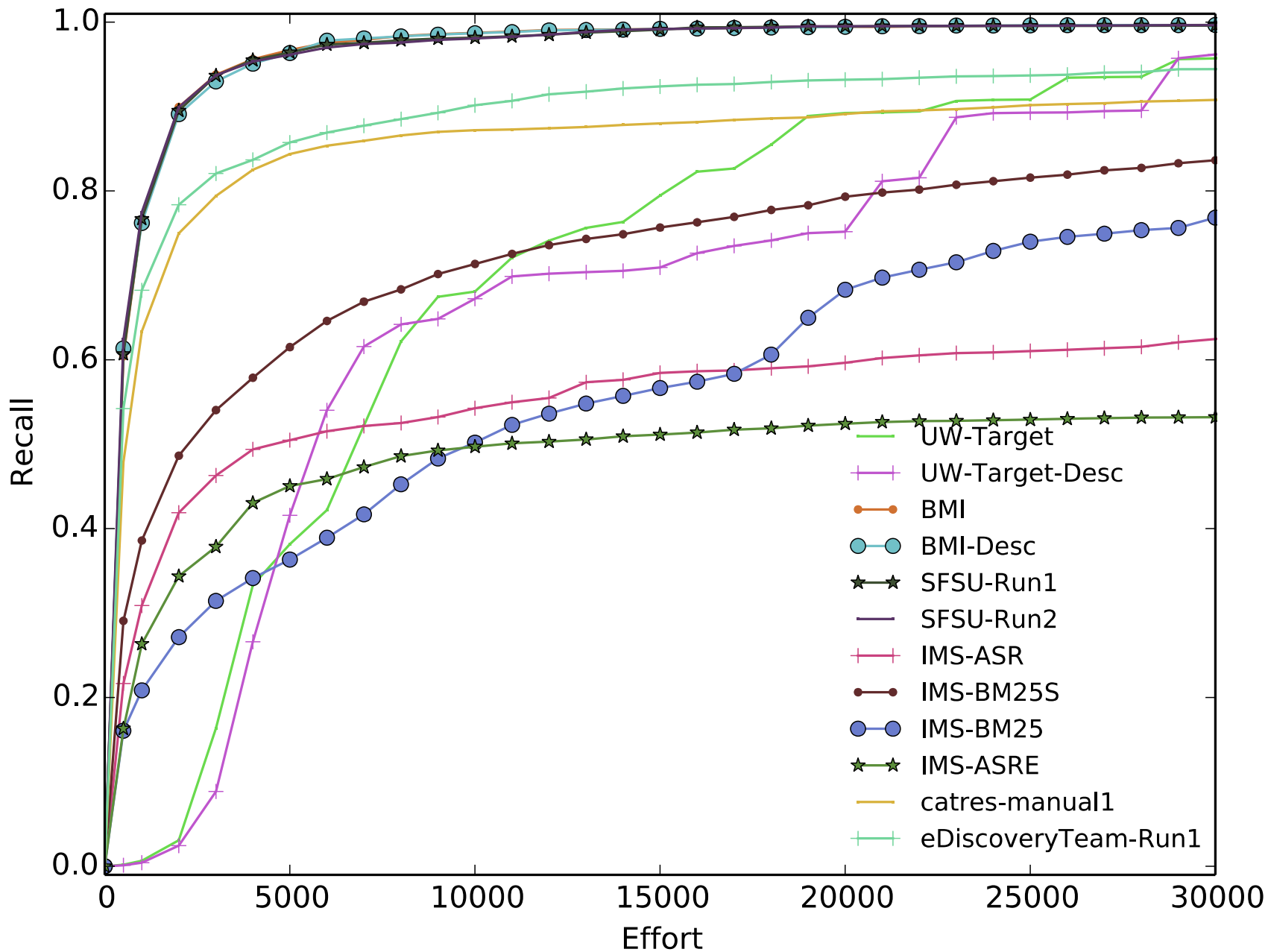


Overall Recall

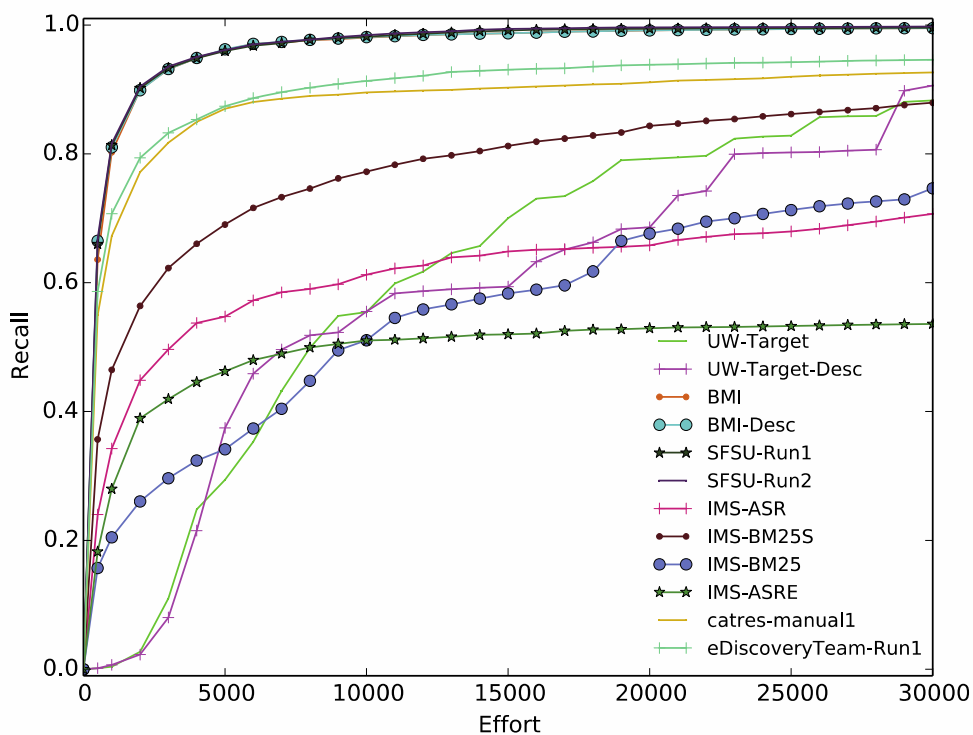


Important Recall

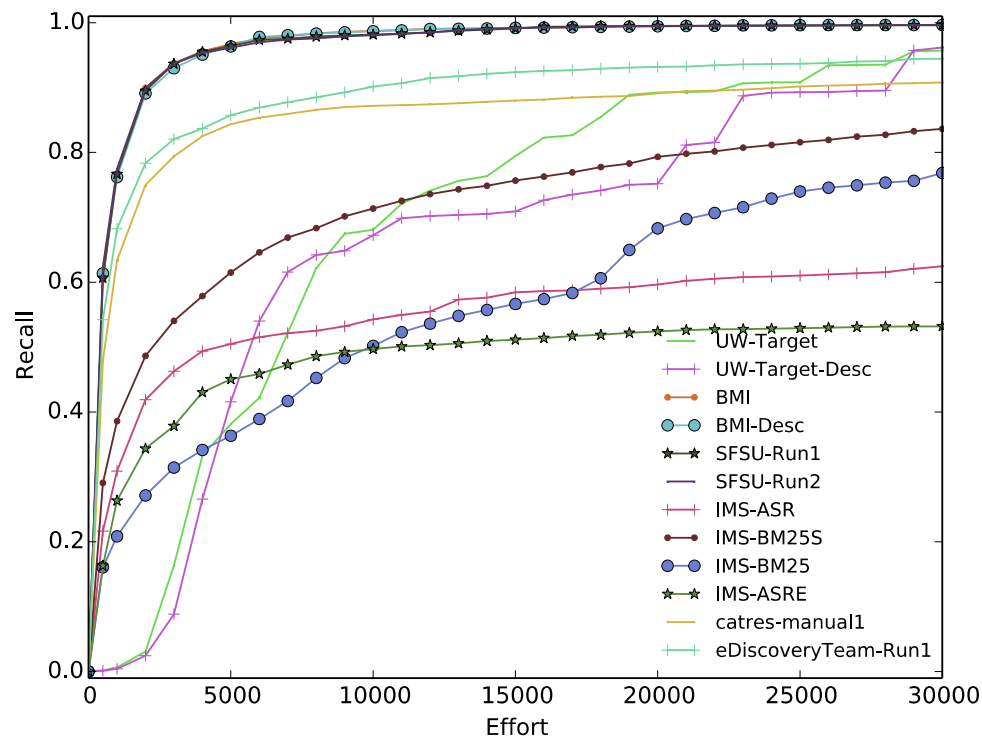
Average Subtopic Recall



Overall Recall vs. Subtopic Recall



Overall Recall



Subtopic Recall