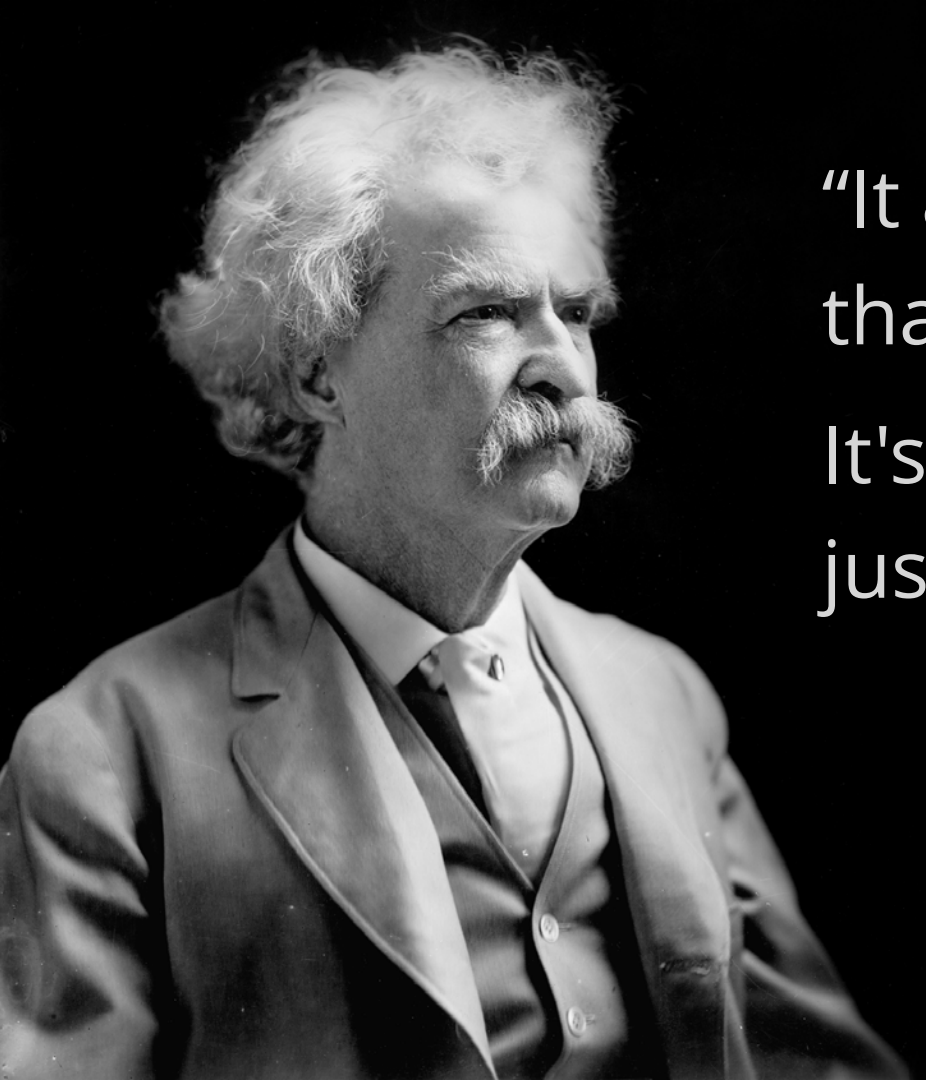
Catalyst

ASU Arkfeld Analytics

Five Years of CAL

The Case for Testing and What it Tells Us

John Tredennick
Founder Catalyst



“It ain't what you don't know that gets you into trouble.

It's what you know for sure th just ain't so.”

Mark Twain

Are Humans the Weak Link?

I've managed my share of TAR projects. I've used or seen used the various flavors of TAR and the outcomes these products produce.

LEGAL OPERATIONS

Are Humans The Weak Link In Technology-Assisted Review?

If there is any shortcoming of TAR technologies, the blame may fairly be placed at the feet (and in the minds) of humans.

By MIKE QUARTARARO

Oct 16, 2018 at 1:46 PM

There's been debate throughout the legal industry about which software product is the superior tool for conducting technology-assisted review (TAR). I've been involved in more discussions than I care to recount about the TAR process, the available tools, and the people using them. I'm not aware of any scientific study demonstrating that any particular TAR software or algorithm is dramatically better or, more importantly, significantly more accurate, than any other. In the end, it seems to me that the only real problem with TAR software — all of them — is the people who use it.

Mike Quartararo

Are Humans the Weak Link?

In the end, it seems to me that the only real problem with TAR software — all of them — is the people who use it.

That's not just the opinion of a somewhat cynical operations guy. It's true. And I would not write it if it weren't.

LEGAL OPERATIONS

Are Humans The Weak Link In Technology-Assisted Review?

If there is any shortcoming of TAR technologies, the blame may fairly be placed at the feet (and in the minds) of humans.

By MIKE QUARTARARO

Oct 16, 2018 at 1:46 PM

There's been debate throughout the legal industry about which software product is the superior tool for conducting technology-assisted review (TAR). I've been involved in more discussions than I care to recount about the TAR process, the available tools, and the people using them. I'm not aware of any scientific study demonstrating that any particular TAR software or algorithm is dramatically better or, more importantly, significantly more accurate, than any other. In the end, it seems to me that the only real problem with TAR software — all of them — is the people who use it.

Mike Quartararo

Are Humans the Weak Link?

truth·i·ness

/ˈtro͞oTHēnis/

noun INFORMAL

the quality of seeming or being felt to be true, even if not necessarily true.

Human Review

The Gold Standard?

"The idea that exhaustive manual review is the most effective – and therefore the most defensible – approach to document review is strongly refuted. Technology assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort."

Grossman and Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review, Richmond Journal of Law and Tech, Vol XVII, Issue 3 (2011).

Keyword Search

Attorneys worked with experienced paralegals to develop search terms. Upon finishing, they estimated that they had retrieved at least three quarters of all relevant documents.

What they ***actually*** retrieved:

Blair & Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System* (1985).

Lawyers Can be the Weak Link

truth·i·ness

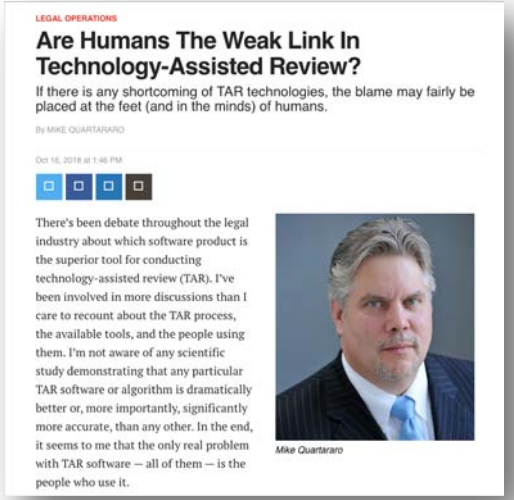
/ˈtro͞oTHēnis/

noun INFORMAL

the quality of seeming or being felt to be true, even if not necessarily true.

Another Gem

A few days ago, I began wondering what is known to be true about TAR that everyone in the eDiscovery space should be able to agree upon.



LEGAL OPERATIONS

Are Humans The Weak Link In Technology-Assisted Review?

If there is any shortcoming of TAR technologies, the blame may fairly be placed at the feet (and in the minds) of humans.

By MIKE QUARTARARO

Oct 16, 2018 at 1:46 PM

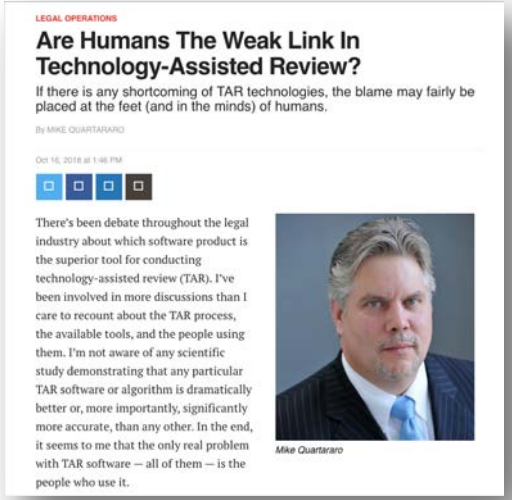
There's been debate throughout the legal industry about which software product is the superior tool for conducting technology-assisted review (TAR). I've been involved in more discussions than I care to recount about the TAR process, the available tools, and the people using them. I'm not aware of any scientific study demonstrating that any particular TAR software or algorithm is dramatically better or, more importantly, significantly more accurate, than any other. In the end, it seems to me that the only real problem with TAR software — all of them — is the people who use it.

Mike Quartararo

Another Gem

First, TAR is not artificial intelligence. . . When you cut through the chaff of the marketing hype, TAR is machine learning — nothing more, nothing less. . . There's nothing artificially intelligent about TAR. It does not think or reason on its own.

[Y]ou get out of a TAR project exactly what you put into it. Anyone who says otherwise is either not being honest or just doesn't know any better.



LEGAL OPERATIONS

Are Humans The Weak Link In Technology-Assisted Review?

If there is any shortcoming of TAR technologies, the blame may fairly be placed at the feet (and in the minds) of humans.

By MIKE QUARTARARO

Oct 16, 2018 at 1:46 PM

There's been debate throughout the legal industry about which software product is the superior tool for conducting technology-assisted review (TAR). I've been involved in more discussions than I care to recount about the TAR process, the available tools, and the people using them. I'm not aware of any scientific study demonstrating that any particular TAR software or algorithm is dramatically better or, more importantly, significantly more accurate, than any other. In the end, it seems to me that the only real problem with TAR software — all of them — is the people who use it.

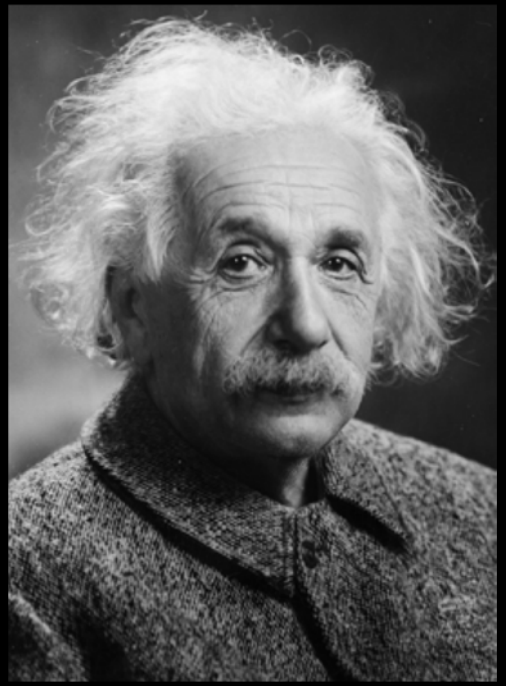
Mike Quartararo

Artificial Intelligence

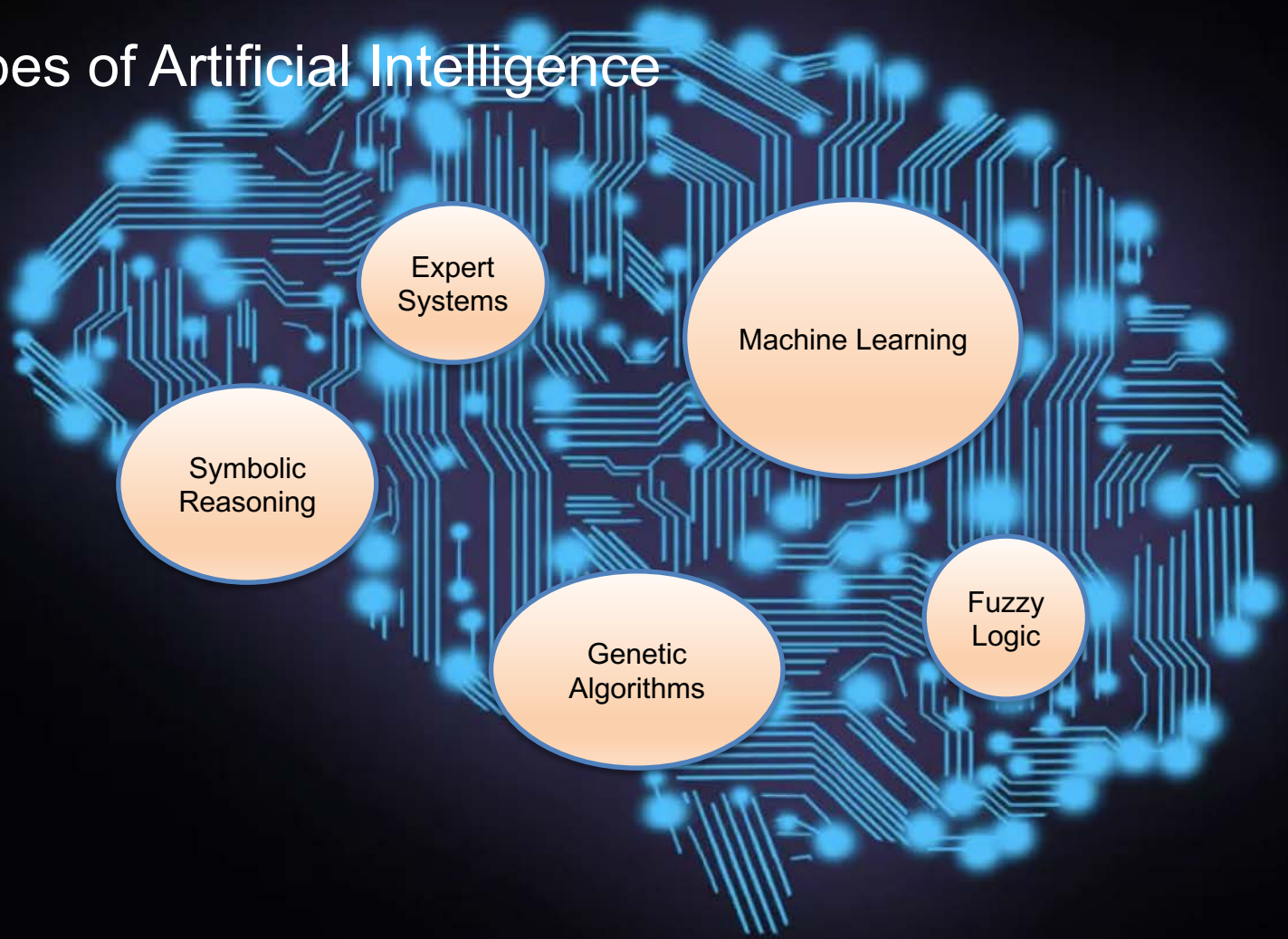
Is the ability of a computer to mimic certain operations of the human mind.

Is the term used when machines are able to learn, reason, discover meaning or generalize from large volumes of data

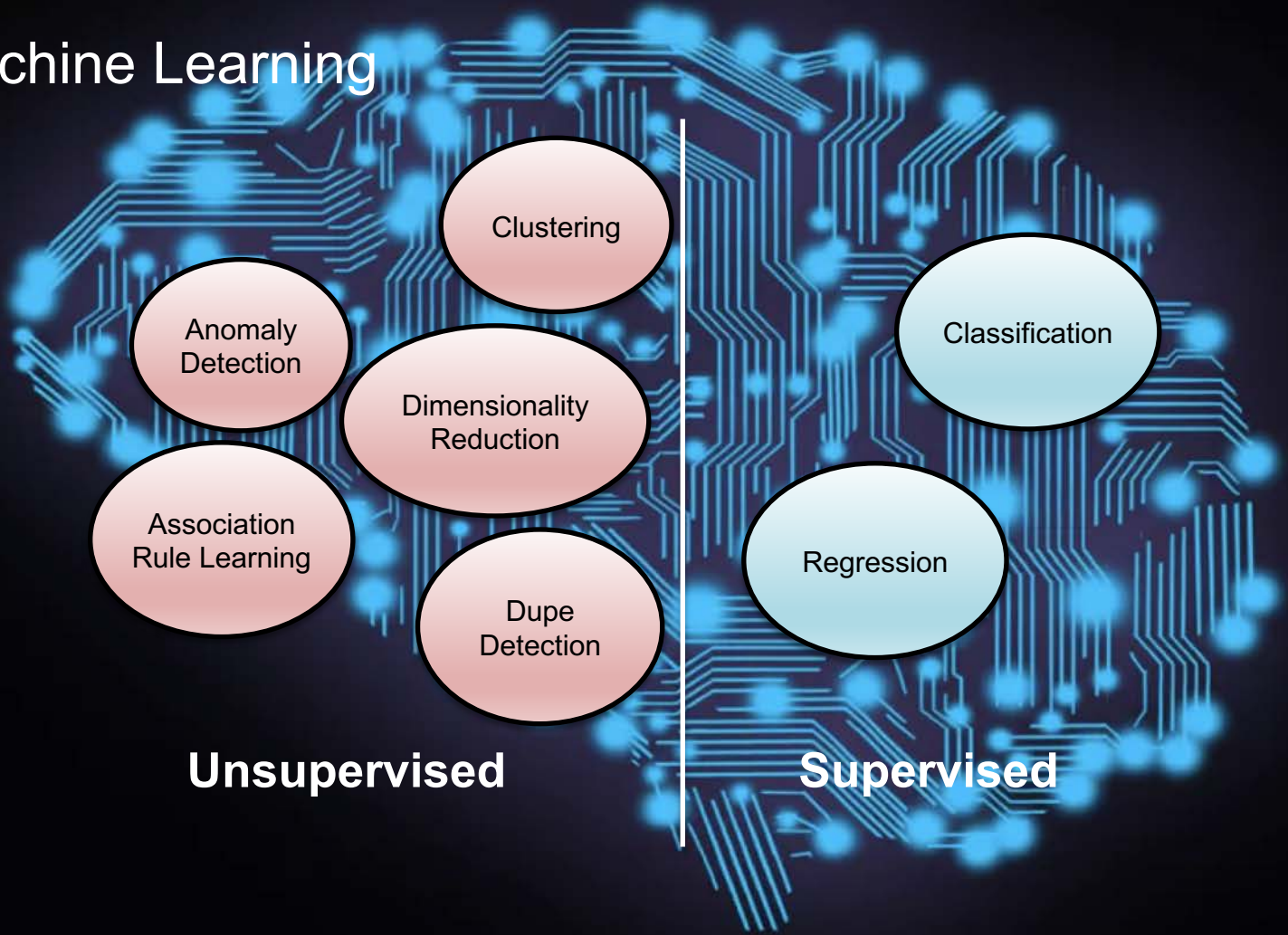
The goal is to arrive at a “reasoned” conclusion, simulating the human decision process, often with better decisions.



Types of Artificial Intelligence



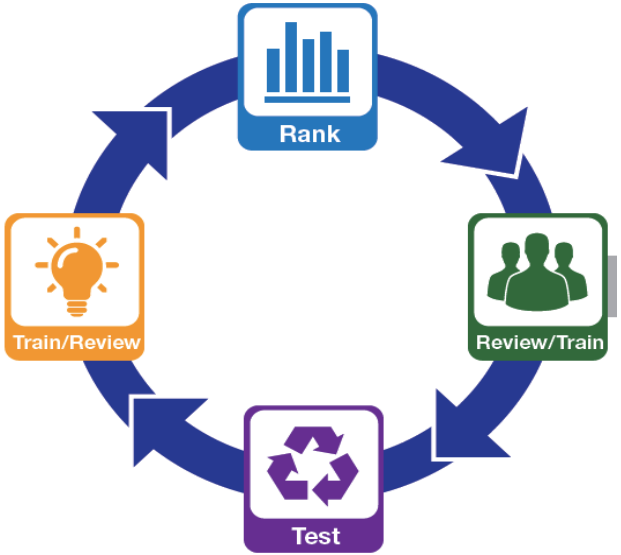
Machine Learning




Intelligent
Advanced
Predictive
Ranking
Adaptive
Automated
Responsive
Computer
Transparent
Text Retrieval
Relevancy
Analytics
Classification
Enhanced
Suggested
Presumptive
Coding
Machine Learning
Assisted
Technology
Concept
Information
AutoSuggest
Search
Document
Linguistic
Computer-Assisted
Semantic
Review
Artificial Intelligence
Meaning-based
Prioritized
Algorithms
Tagging
Data Meaning
Privilege

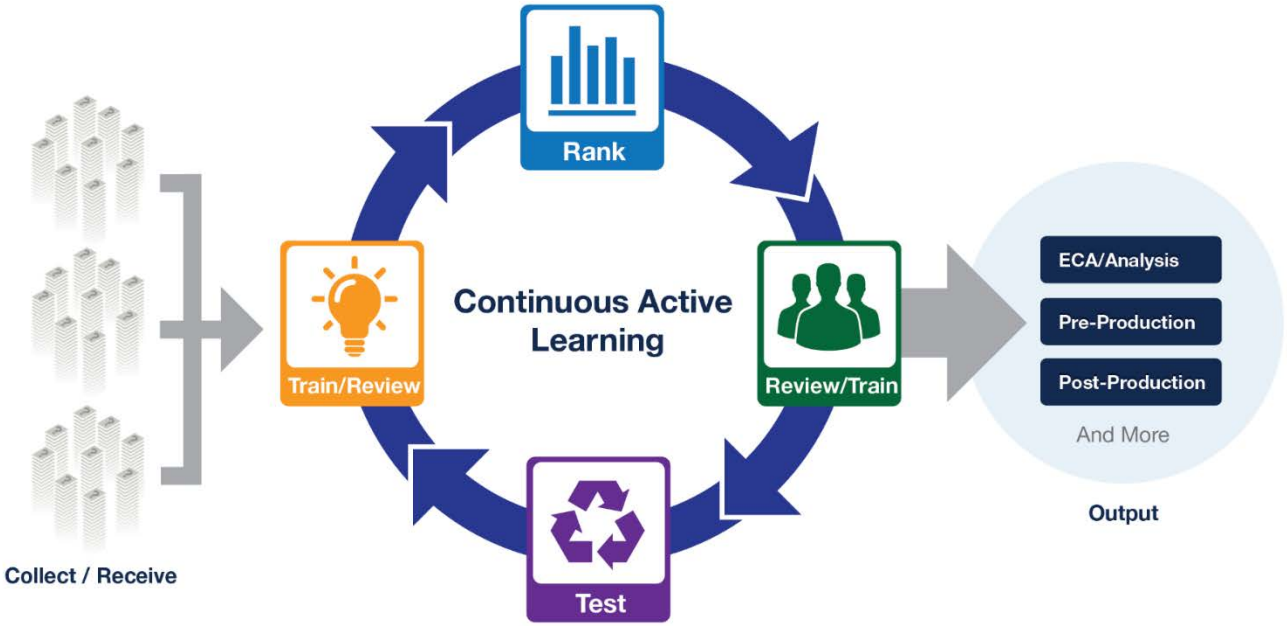
What is TAR?

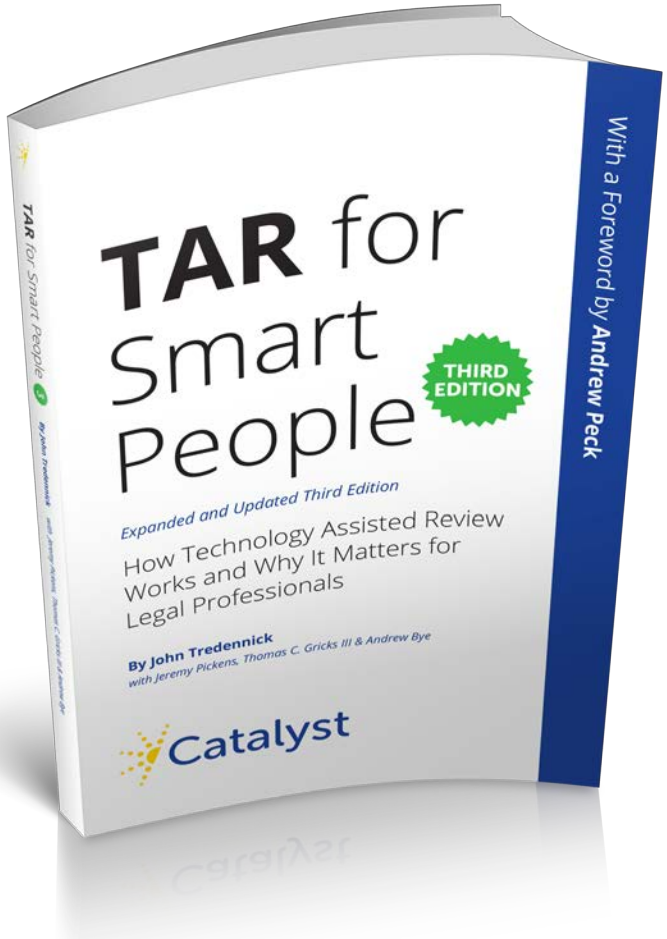
1. A process through which humans work with a computer to teach it to identify relevant documents.
2. Ordering documents by relevance for more efficient review.
3. Stopping the review after you have found a high percentage of relevant documents.



TAR 2.0: Continuous Active Learning

Review equals training

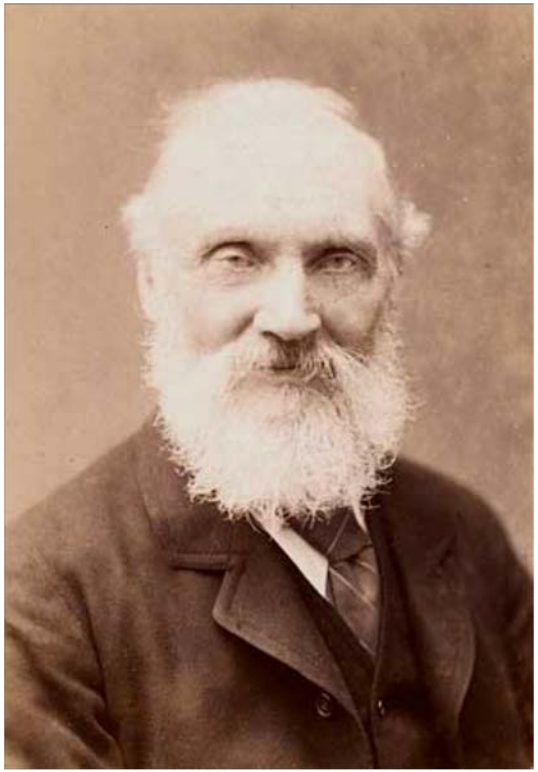



TAR for Smart People
THIRD EDITION
Expanded and Updated Third Edition
How Technology Assisted Review Works and Why It Matters for Legal Professionals
By John Tredennick
with Jeremy Pickens, Thomas C. Gricks III & Andrew Bye
Catalyst
With a Foreword by Andrew Peck

Lord Kelvin (1883)

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind."

If you cannot measure it you cannot improve it.



IR Testing: The Cranfield Model

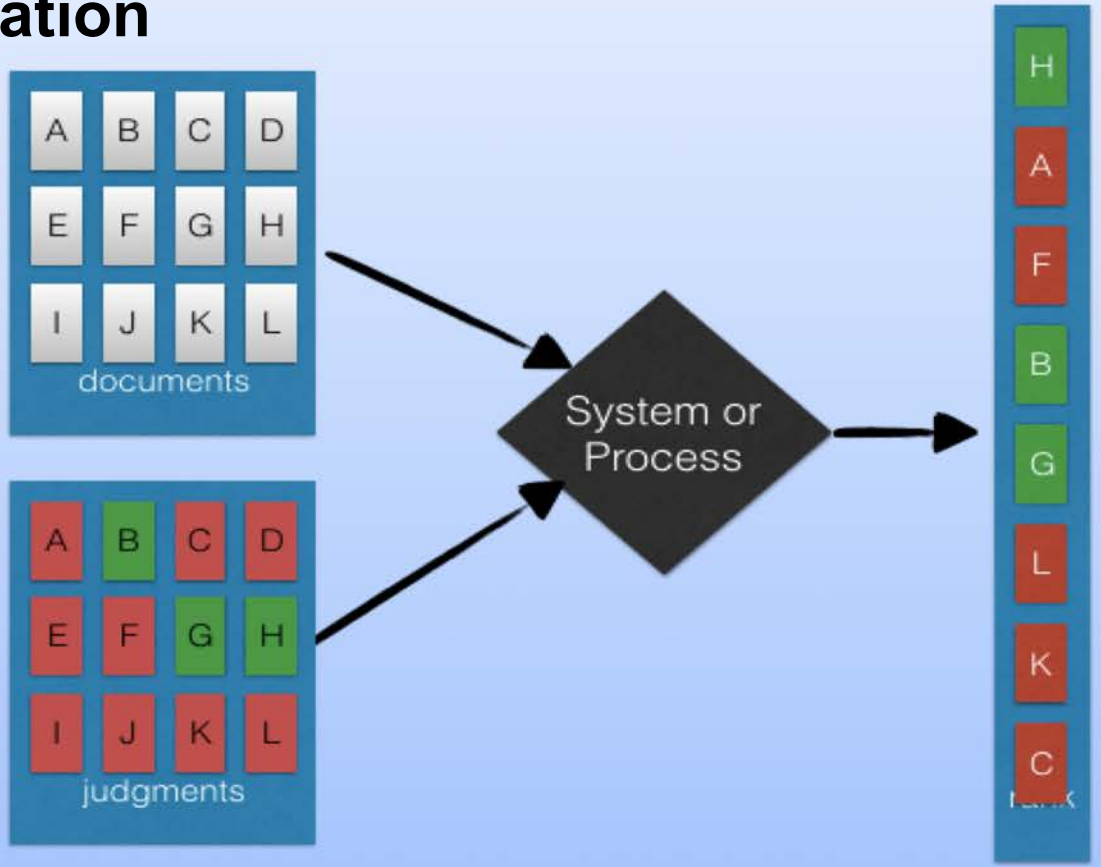
1. Assemble a test collection
   Document corpus
   Judgments
2. Choose an effectiveness metric
3. Vary some aspect of the TAR system (baseline and new idea)
4. Run (simulate) both
5. Compare using the effectiveness metric.

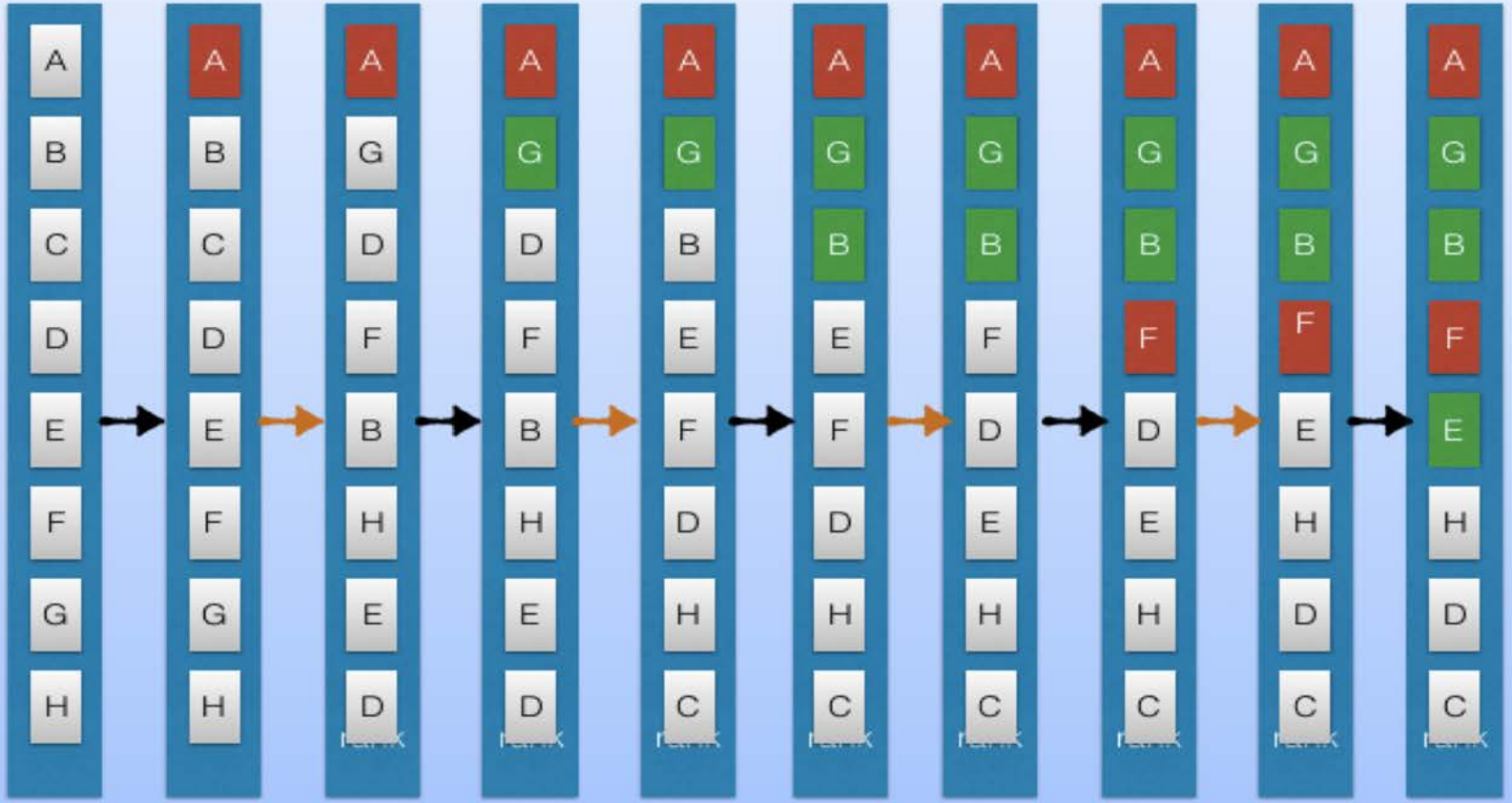
21

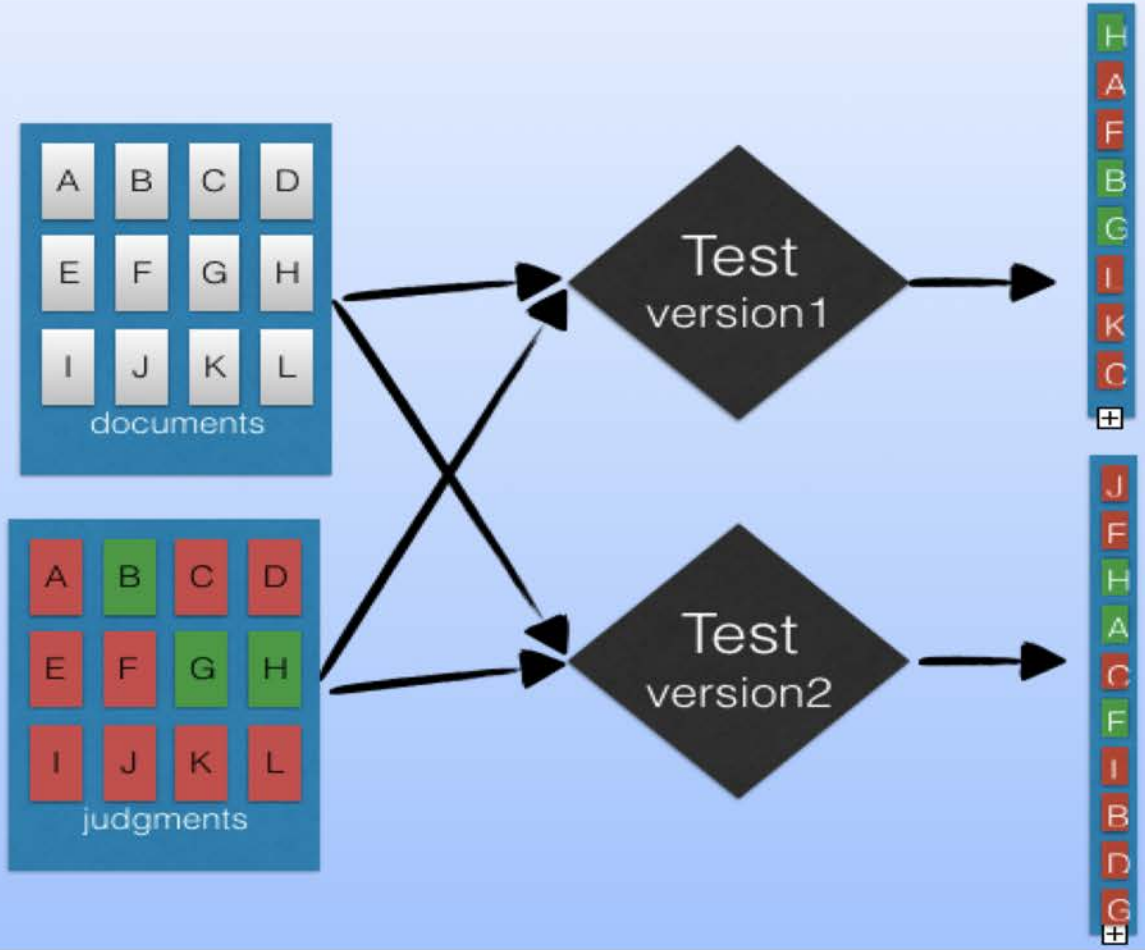
Simulations: How Could You Be So Sure?

Anyone who has watched the epic legal drama *My Cousin Vinny* realizes that the critical question in evaluating any claim is "How could you be so sure?" In our case, the answer is in large part: simulations.

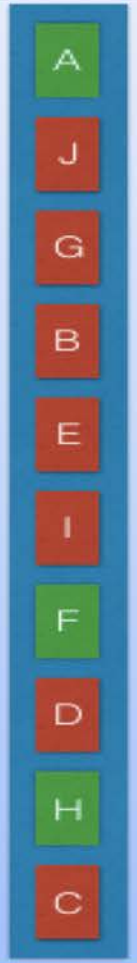
Simulation


A B C D
E F G H
I J K L
documents
A B C D
E F G H
I J K L
judgments
System or Process
H
A
F
B
G
L
K
C


A B C D E F G H
A B C D E F G H
A G D F B H E D
A G D F B H E D
A G B E F D H C
A G B E F D H C
A G B F D E H C
A G B F D E H C
A G B F E H D C
A G B F E H D C


A B C D
E F G H
I J K L
documents
A B C D
E F G H
I J K L
judgments
Test
version1
Test
version2
H A F B G I K C
J F H A C F I B D G

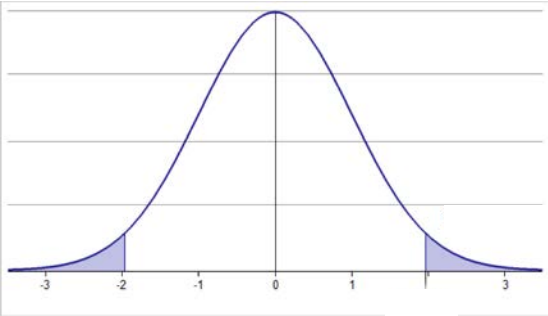
Test 1



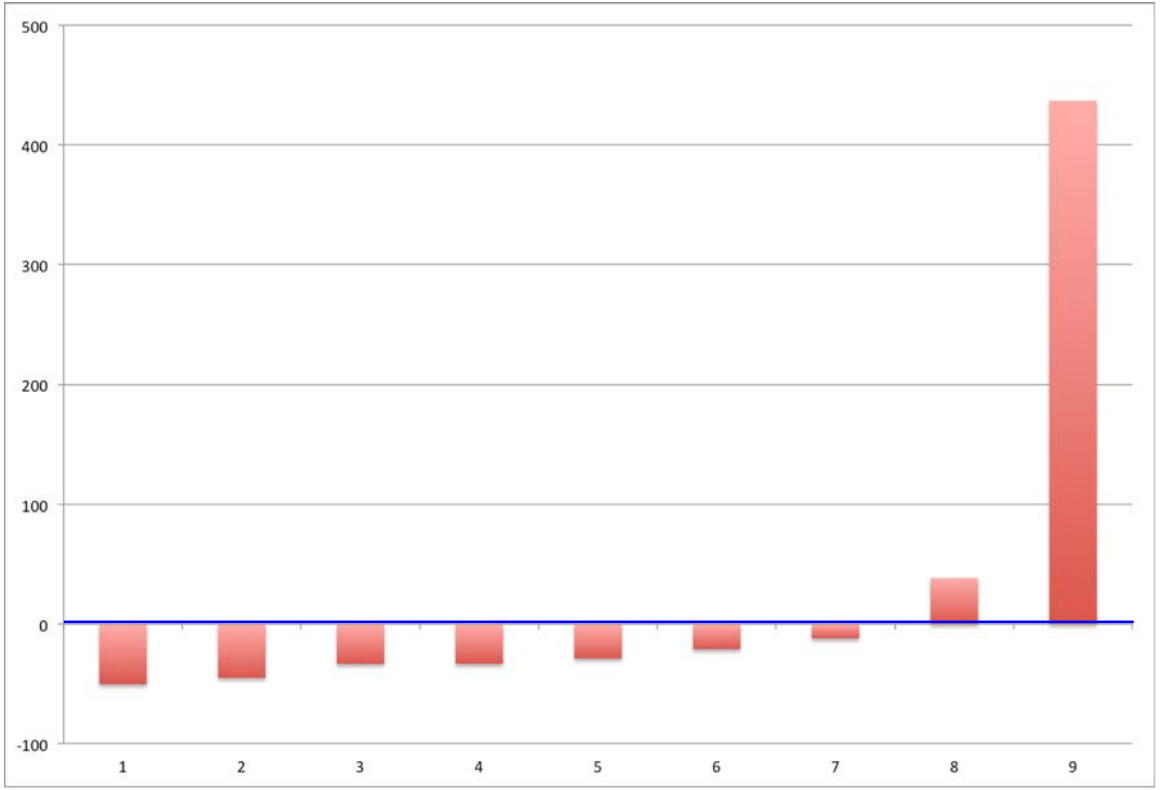
Test 2

Understanding Significance Tests

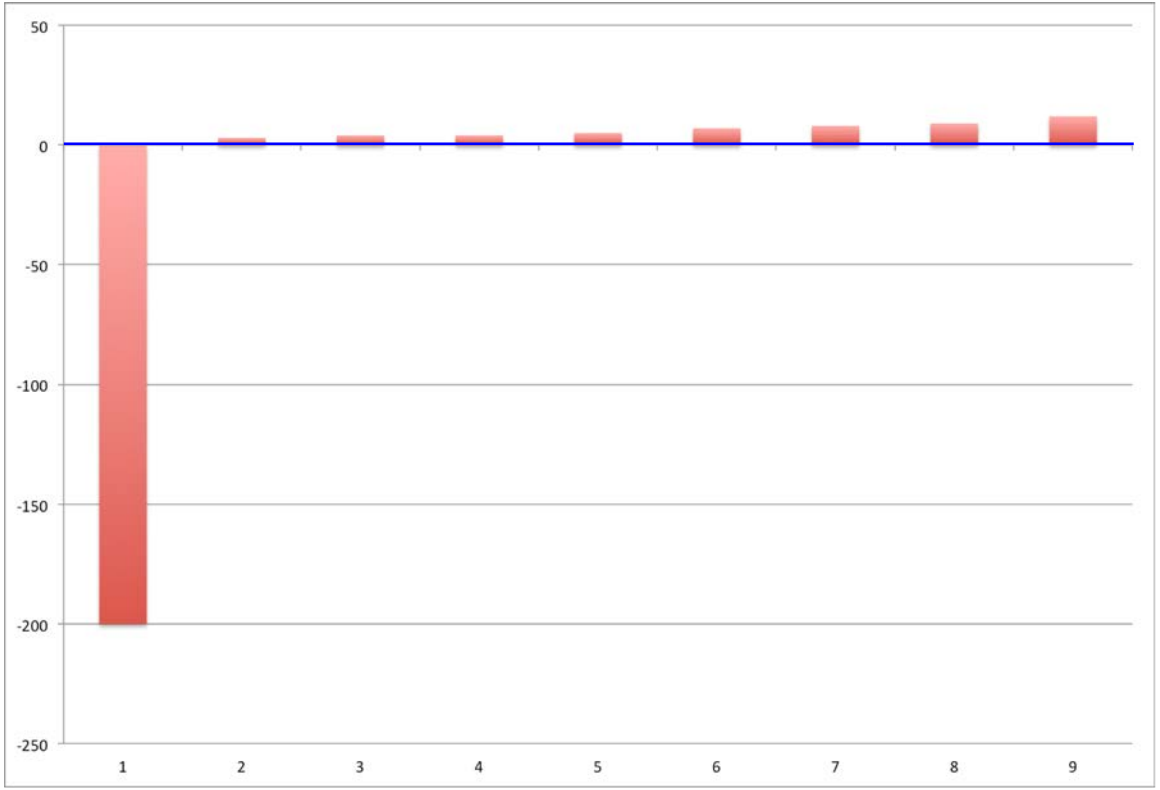
1. Null hypothesis: New system is no better than baseline
   - Compute effectiveness metric for each topic, for both systems (raw score)
   - Compare effectiveness metric for each topic, using test statistic (+/-, %improvement, etc.)
   - Compute p-value using test-statistic (probability that difference is due to chance)
   - Reject null hypothesis if $p \leq \alpha$ (typically 0.1 or 0.05)
2. More topics = more confidence
3. Common tests: t-test, Wilcoxon signed-rank test, sign test



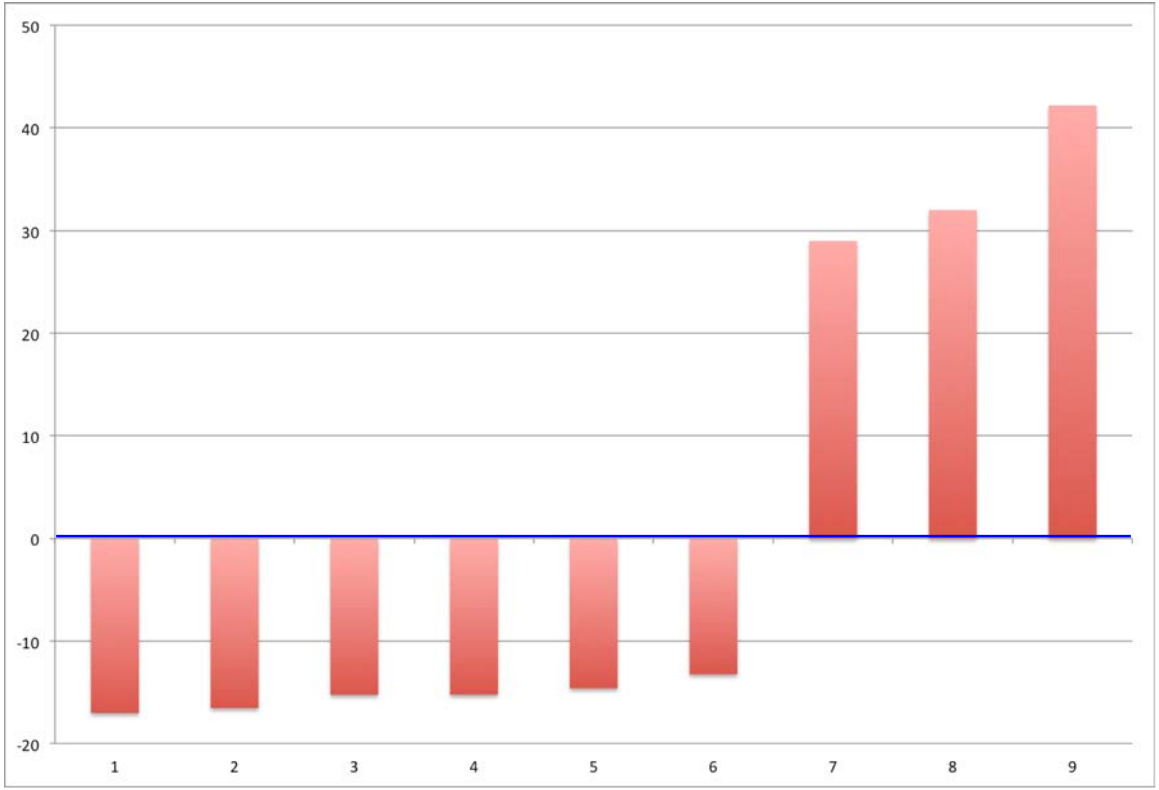
Alternative 1



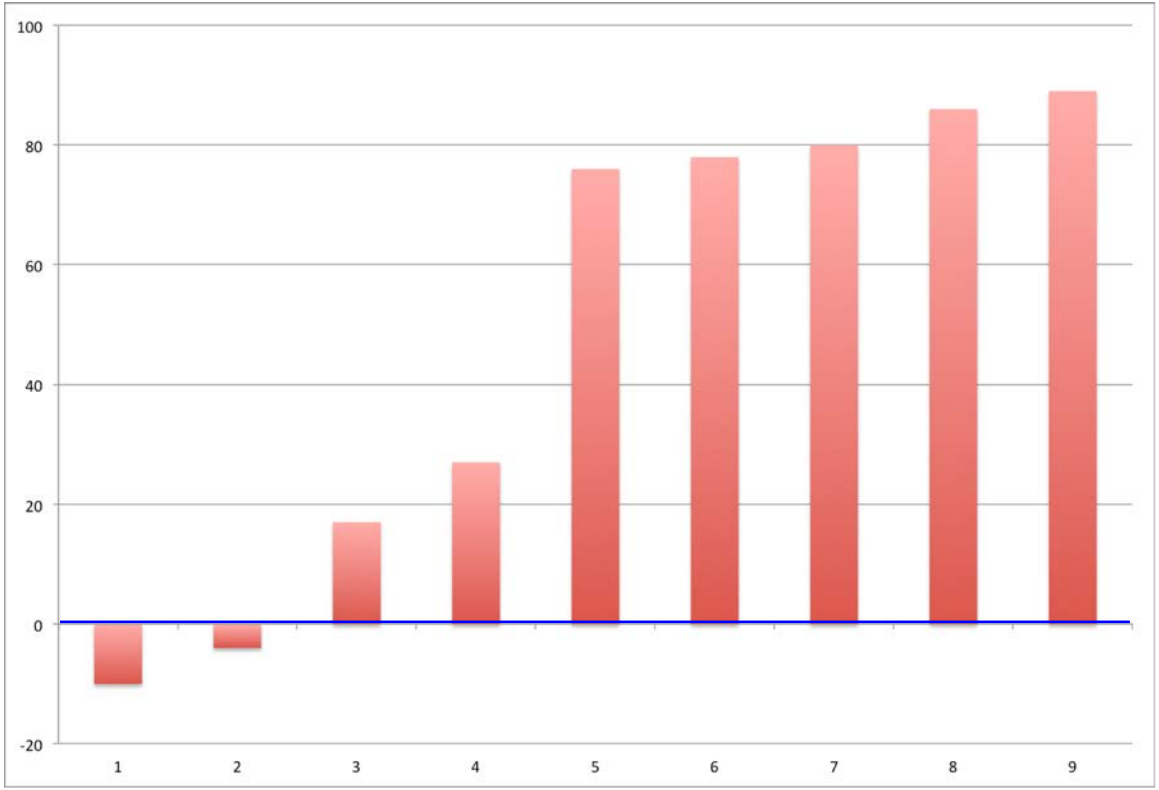
Alternative 2



Alternative 3

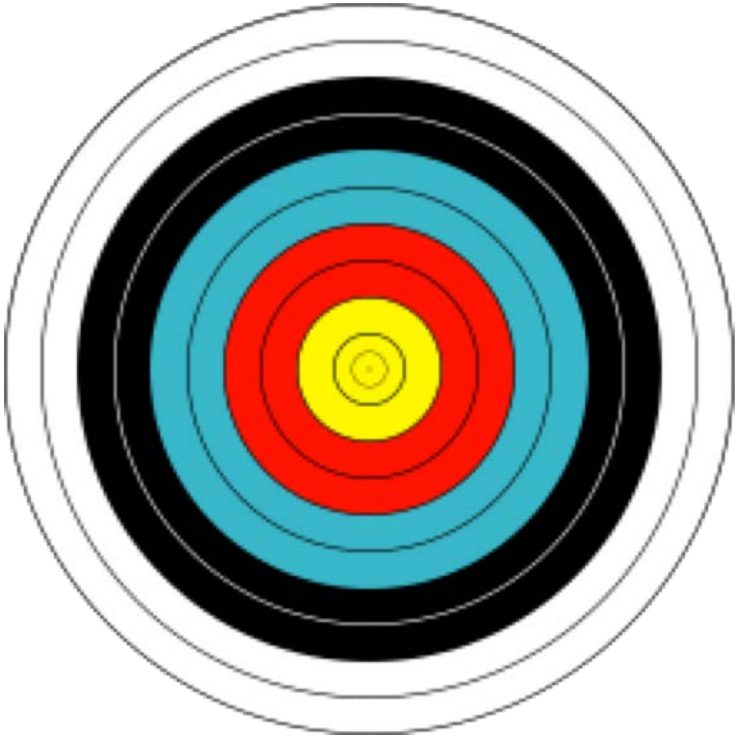


Alternative 4

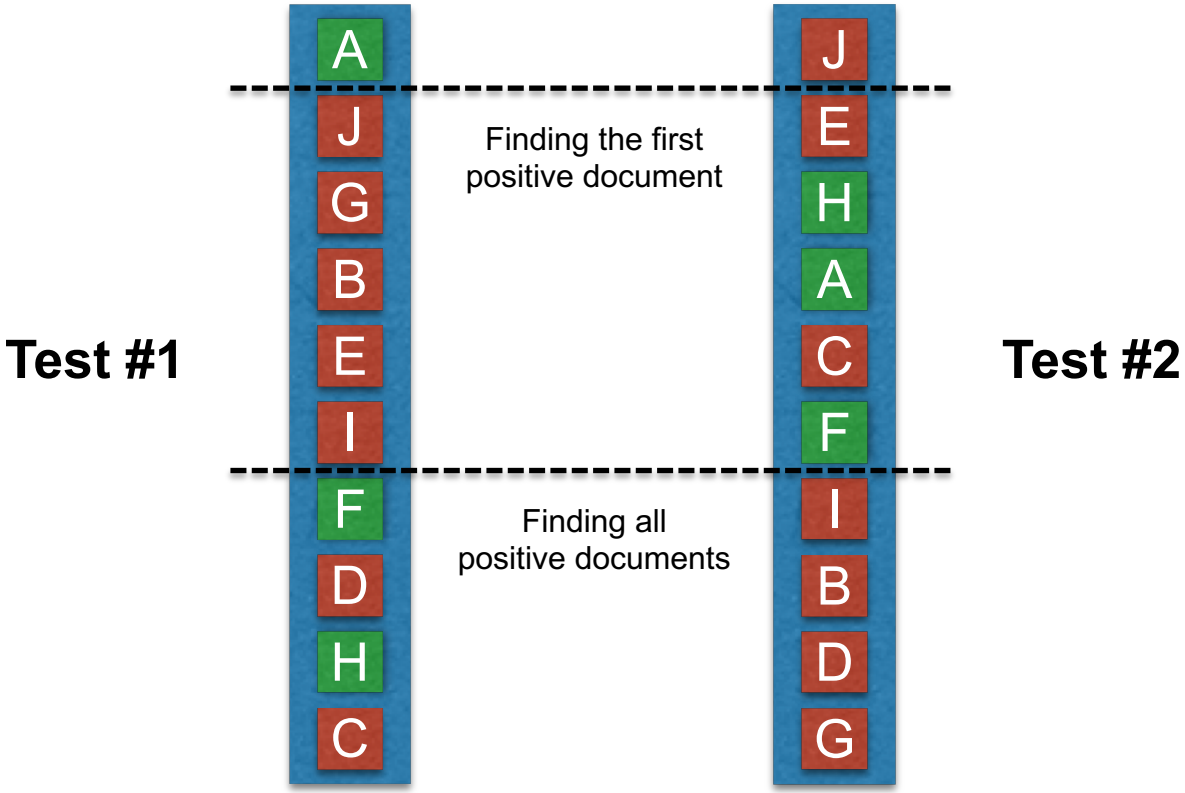


Effectiveness Metrics

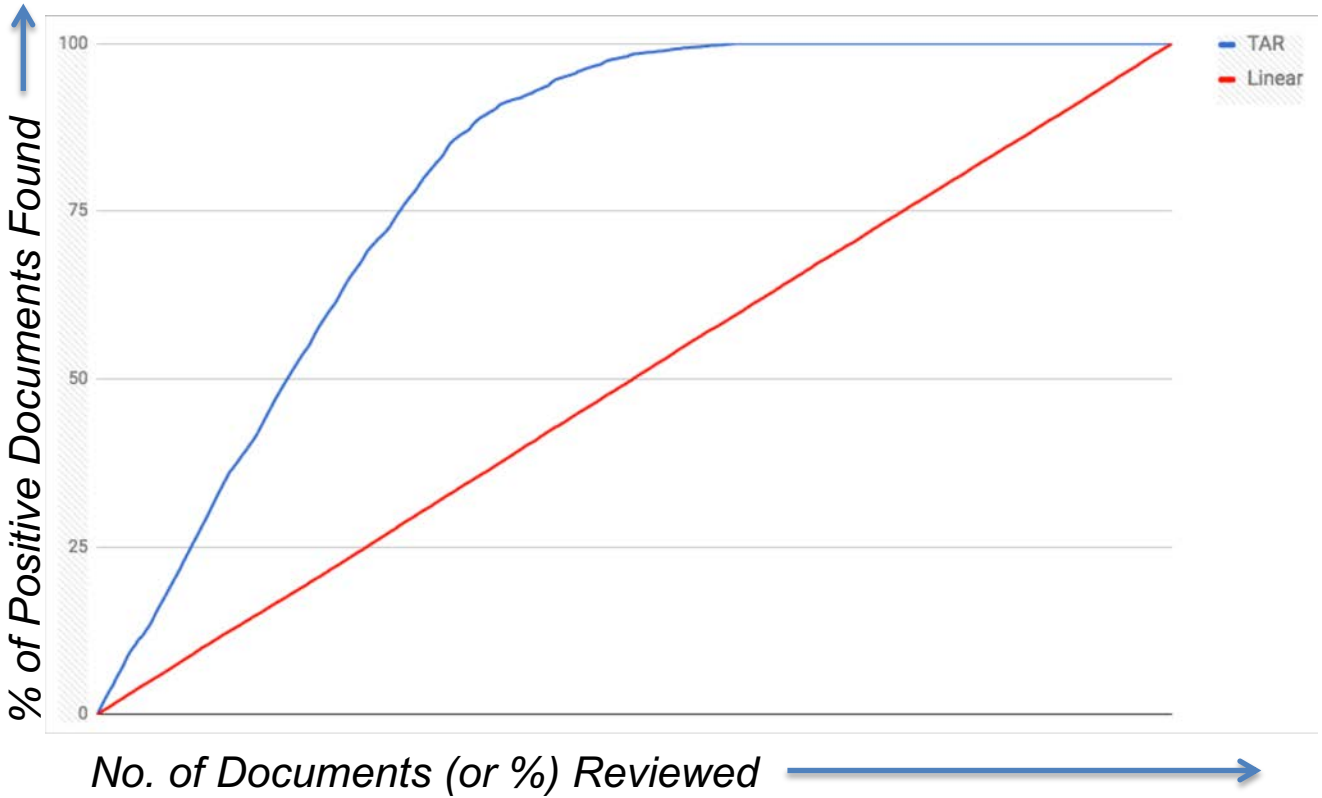
1. Recall
2. Precision
3. Some other goal?



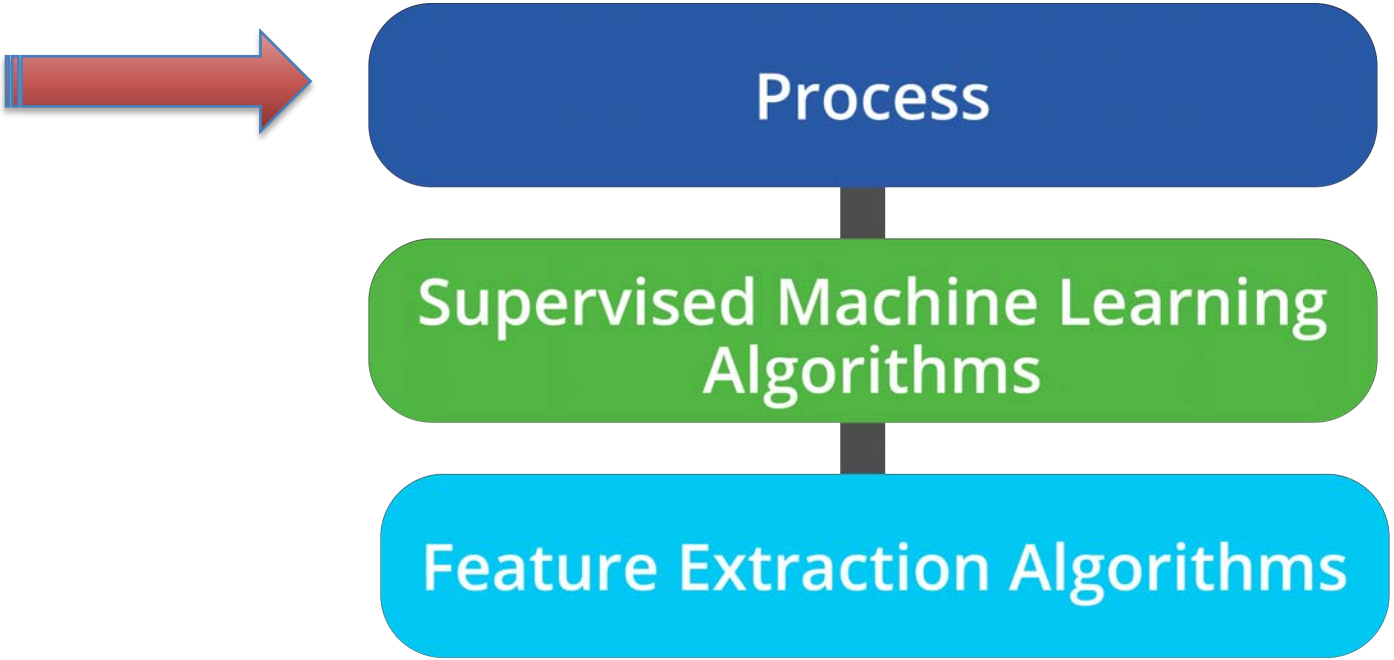
You Need to Know What's Important



Evaluating Results – The Yield Curve



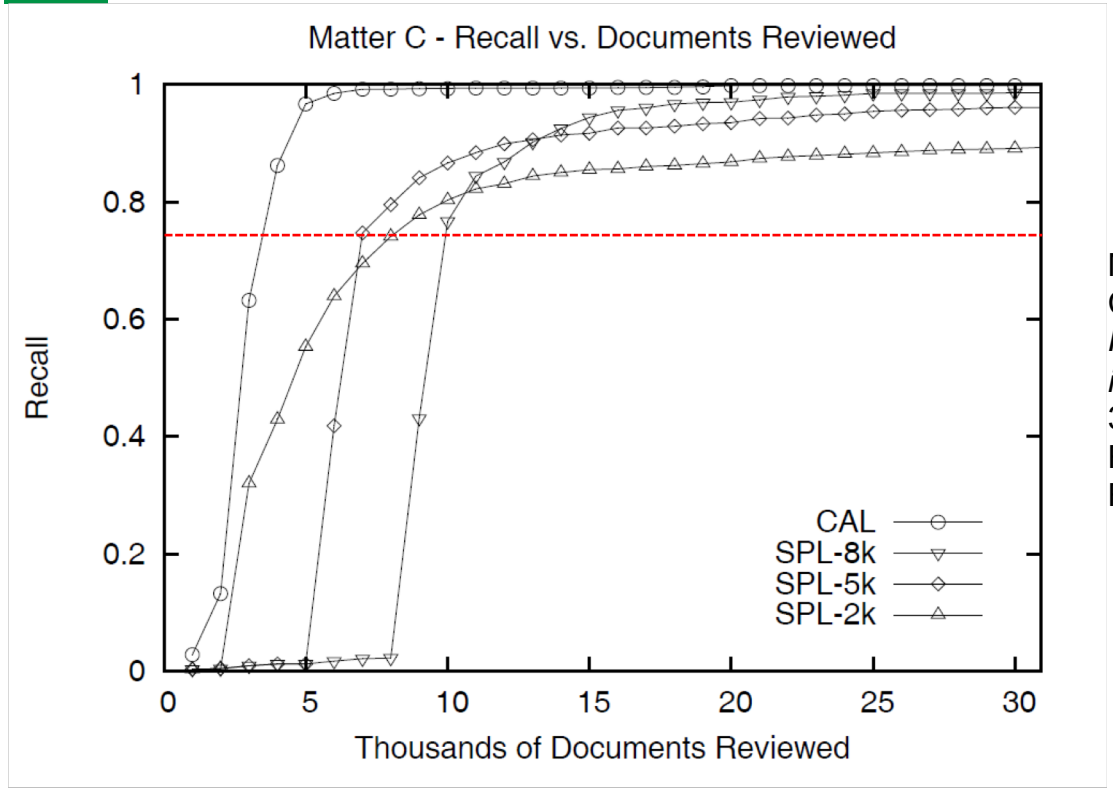
Three “Layers” of TAR



Simulation: Evaluate the Training/Review Protocol

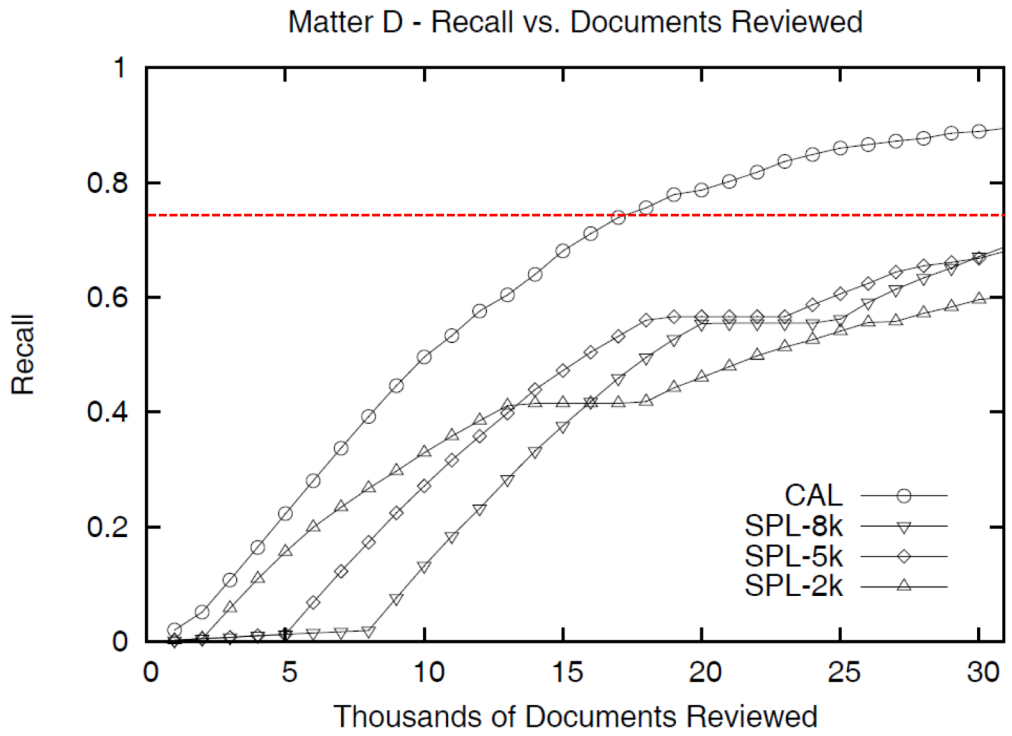
	Condition 1	Condition 2	Condition 3
Document Corpus	Corpus Z	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	[docid:7643 = true] [docid:225 = true]	[docid:7643 = true] [docid:225 = true]	[docid:7643 = true] [docid:225 = true]
Feature (Signal) Extraction	Character n-grams	Character n-grams	Character n-grams
Ranking Engine	Logistic Regresssion	Logistic Regresssion	Logistic Regresssion
Training/Review Protocol	**SPL**	**SAL**	**CAL**
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall	Precision@75% recall

Simulation Results (Metric: Precision at 75% Recall)



Maura R. Grossman and Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, Proceedings of The 37th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (2014)

Simulation Results (Metric: Precision at 75% Recall)

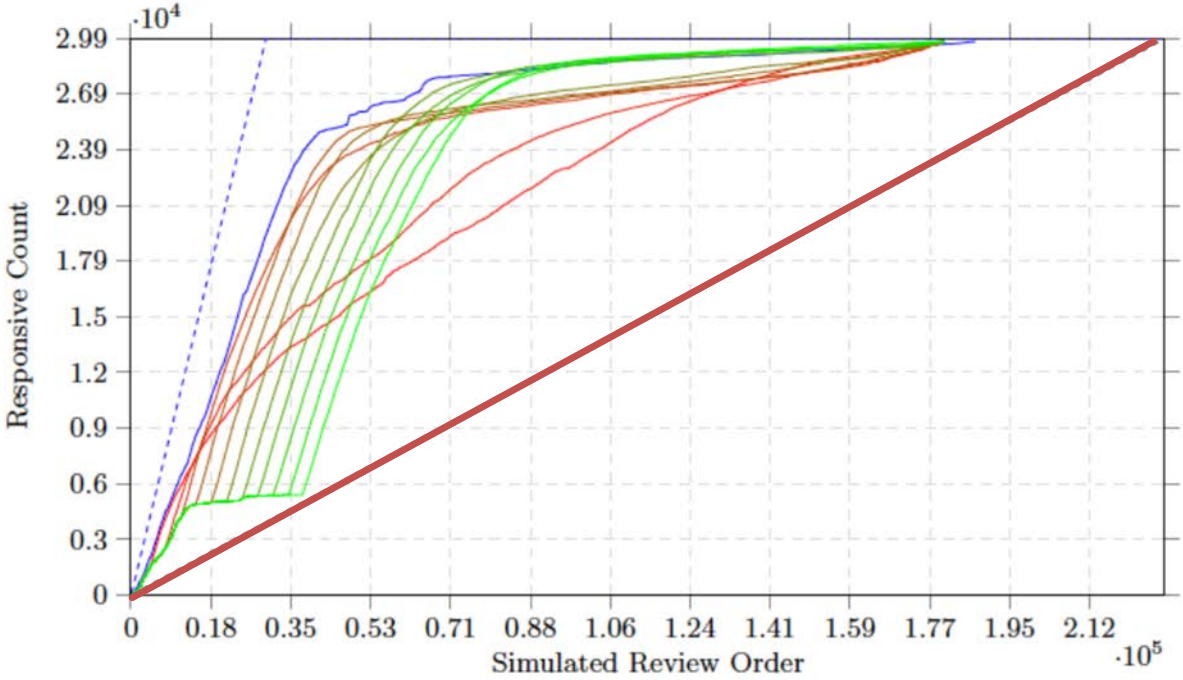


Maura R. Grossman and Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, Proceedings of The 37th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (2014)

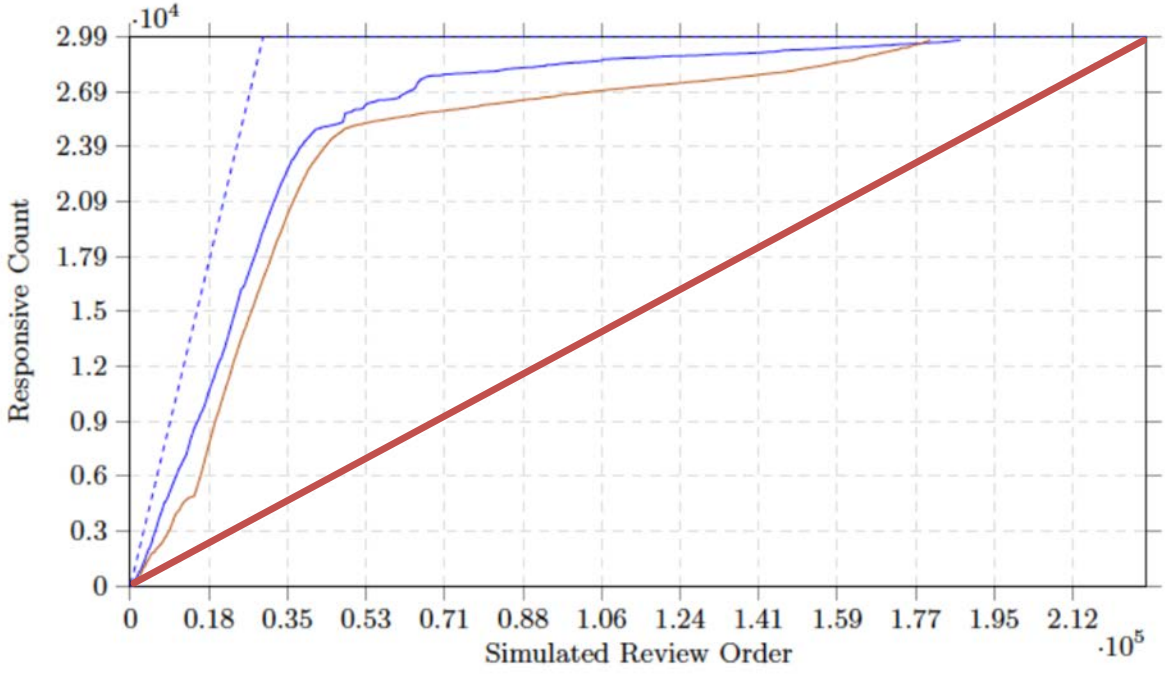
Simulation: Evaluate Expert TAR 1.0 – Non-expert TAR 2.0

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	**[docid:7643 = true] [docid:225 = false]**	**[docid:7643 = true] [docid:225 = true]**
Feature (Signal) Extraction	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]
Training/Review Protocol	**SAL**	**CAL**
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

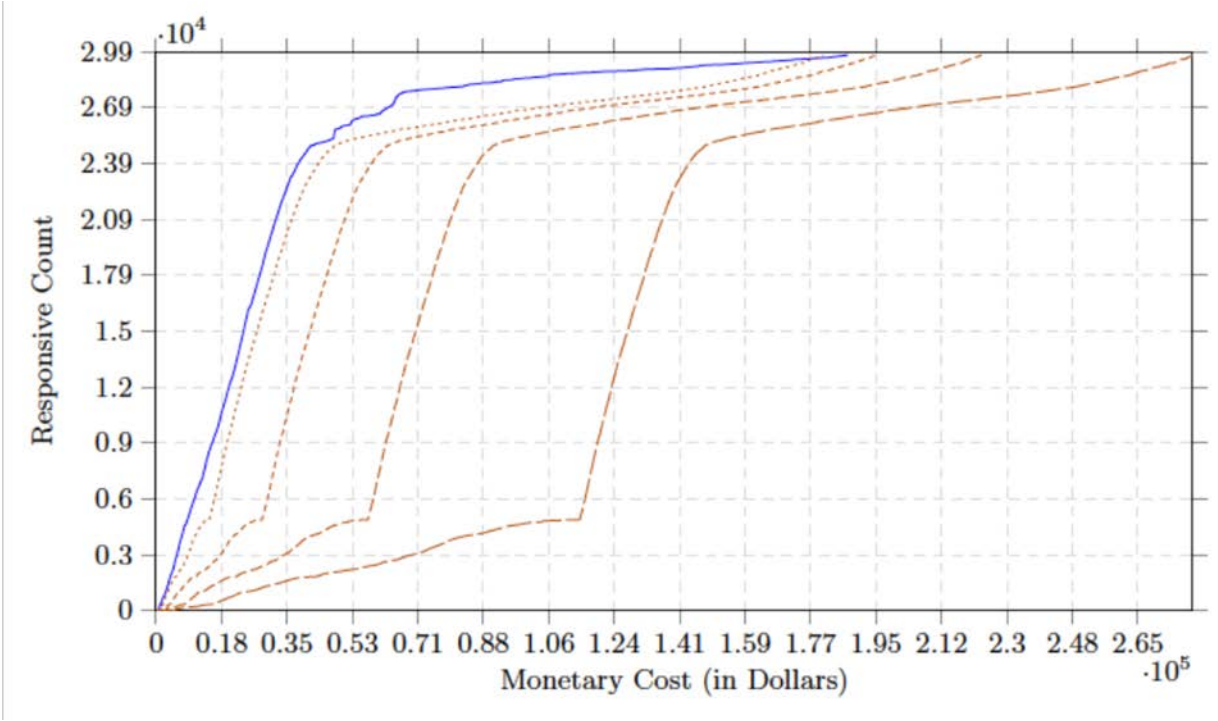
Review as a Function of Training (Metric: Precision at 75% Recall)



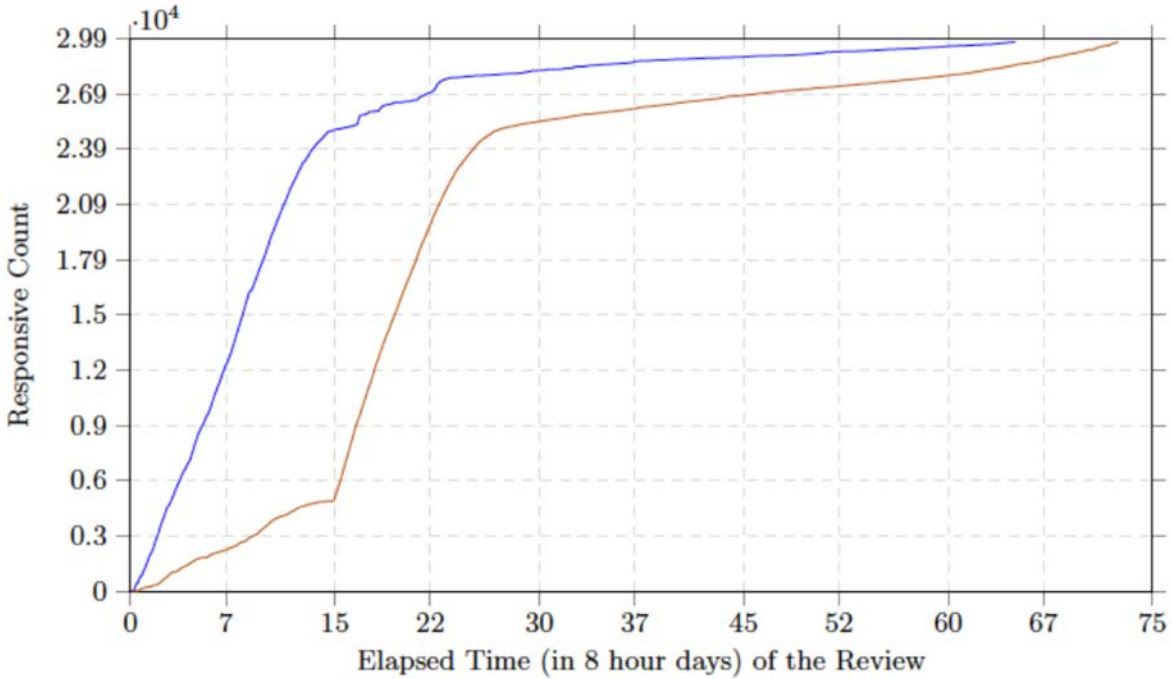
Review at Optimal Training (Metric: Precision at 75% Recall)



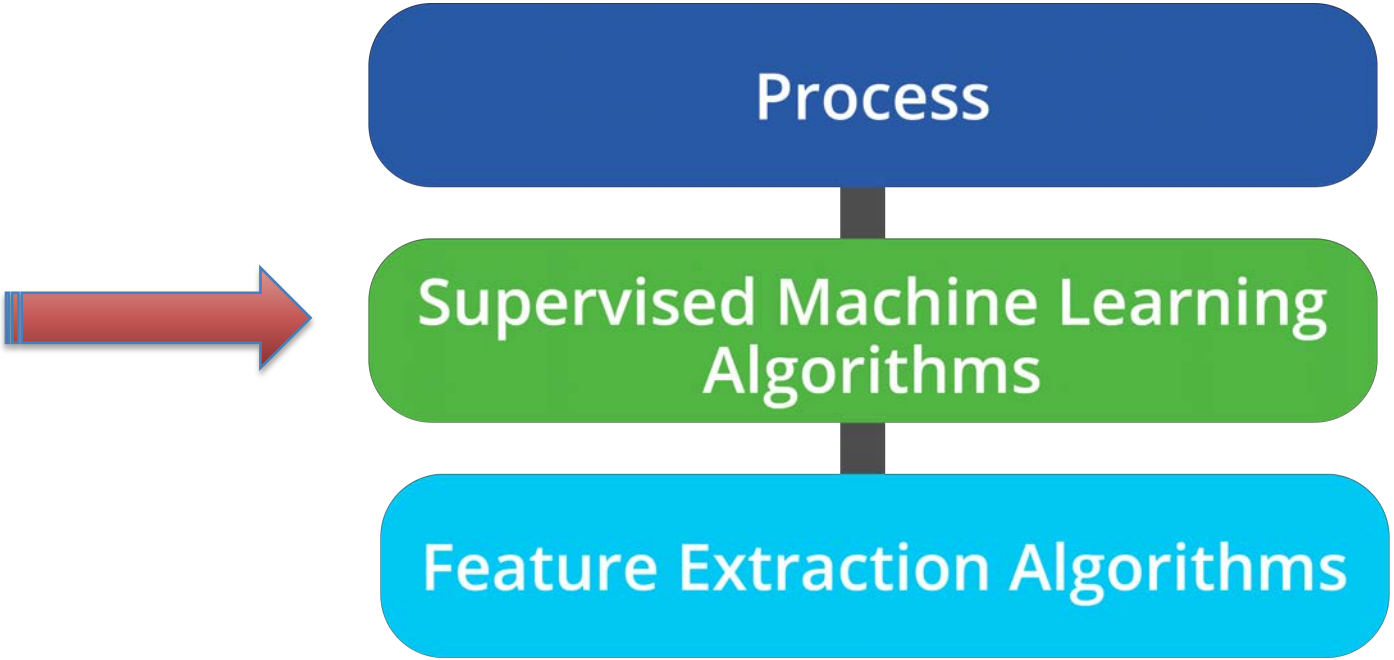
Total Cost of Review (Metric: Total Review Cost at 75% Recall)



Total Time of Review (Metric: Total Review Time at 75% Recall)



Three “Layers” of TAR



Simulation: Evaluate Core Algorithms

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	[docid:7643 = true] [docid:225 = true]	[docid:7643 = true] [docid:225 = true]
Feature (Signal) Extraction	n-grams	n-grams
Ranking Engine	**Logistic Regresssion**	**Support Vector Machine**
Training/Review Protocol	One-shot	One-shot
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Recall at 20k reviewed	Recall at 20k reviewed

Core Algorithm Results (Metric: Recall at 20% Reviewed)

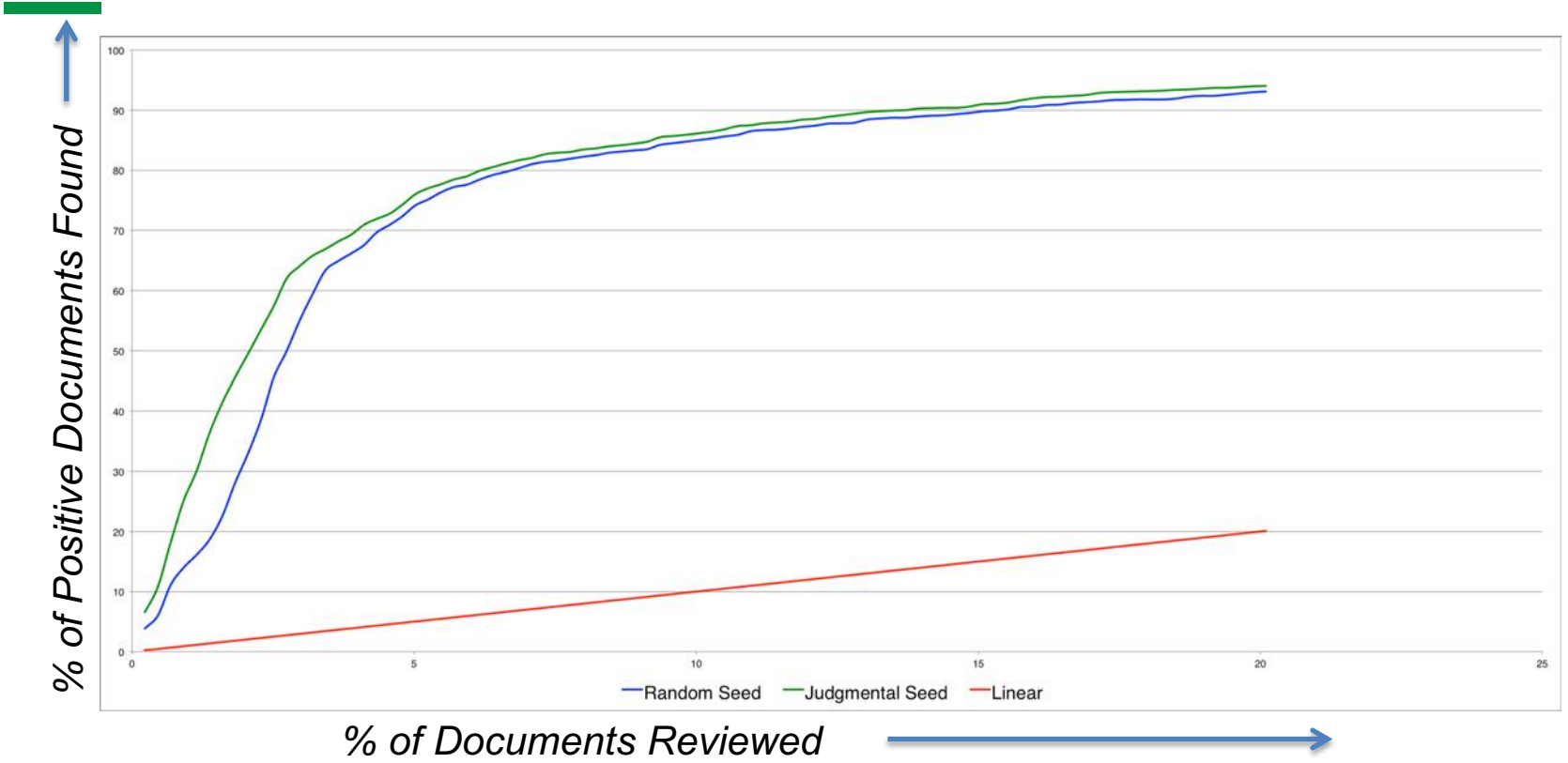
Algorithm	Topic 201	Topic 202	Topic 203	Topic 207
Logistic Regression	92%	96%	90%	90%
Linear SVM	95%	97%	98%	92%
XGBoost	93%	96%	87%	85%
Deep Learning	***74%***	***87%***	***65%***	***86%***
1-NN	89%	92%	92%	84%

Yang et al., *Effectiveness Results for Popular e-Discovery Algorithms*, International Conference on AI and Law, June 2017

Simulation: Random vs. Judgmental Seeds

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	**docids 5738, 83, 29973 (RANDOM)**	**docids 8282, 1209, 36 (JUDGMENTAL)**
Feature (Signal) Extraction	1-grams	1-grams
Ranking Engine	Logistic Regresssion	Logistic Regresssion
Training/Review Protocol	CAL	CAL
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

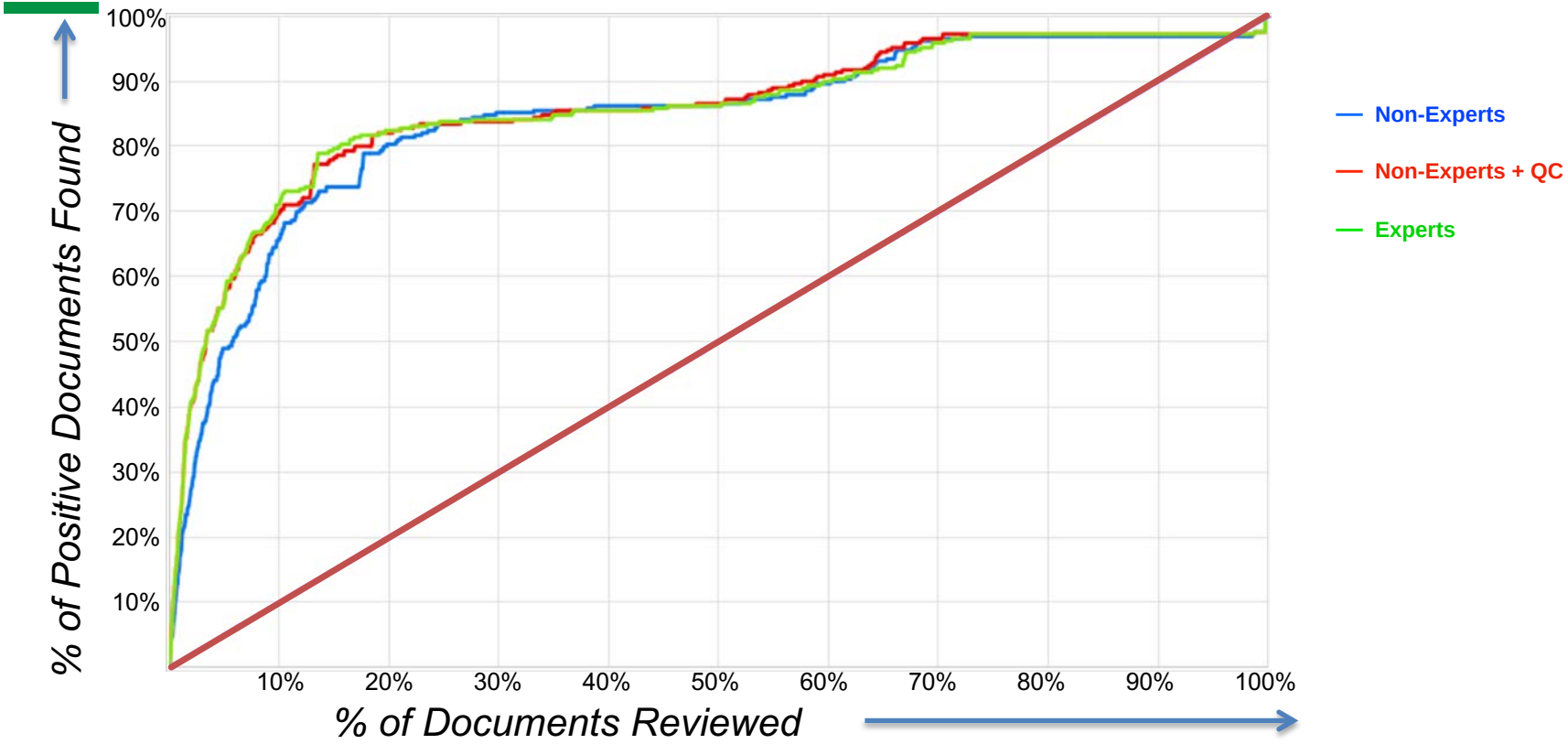
Random – Judgmental Results (Metric: Precision at 75% Recall)



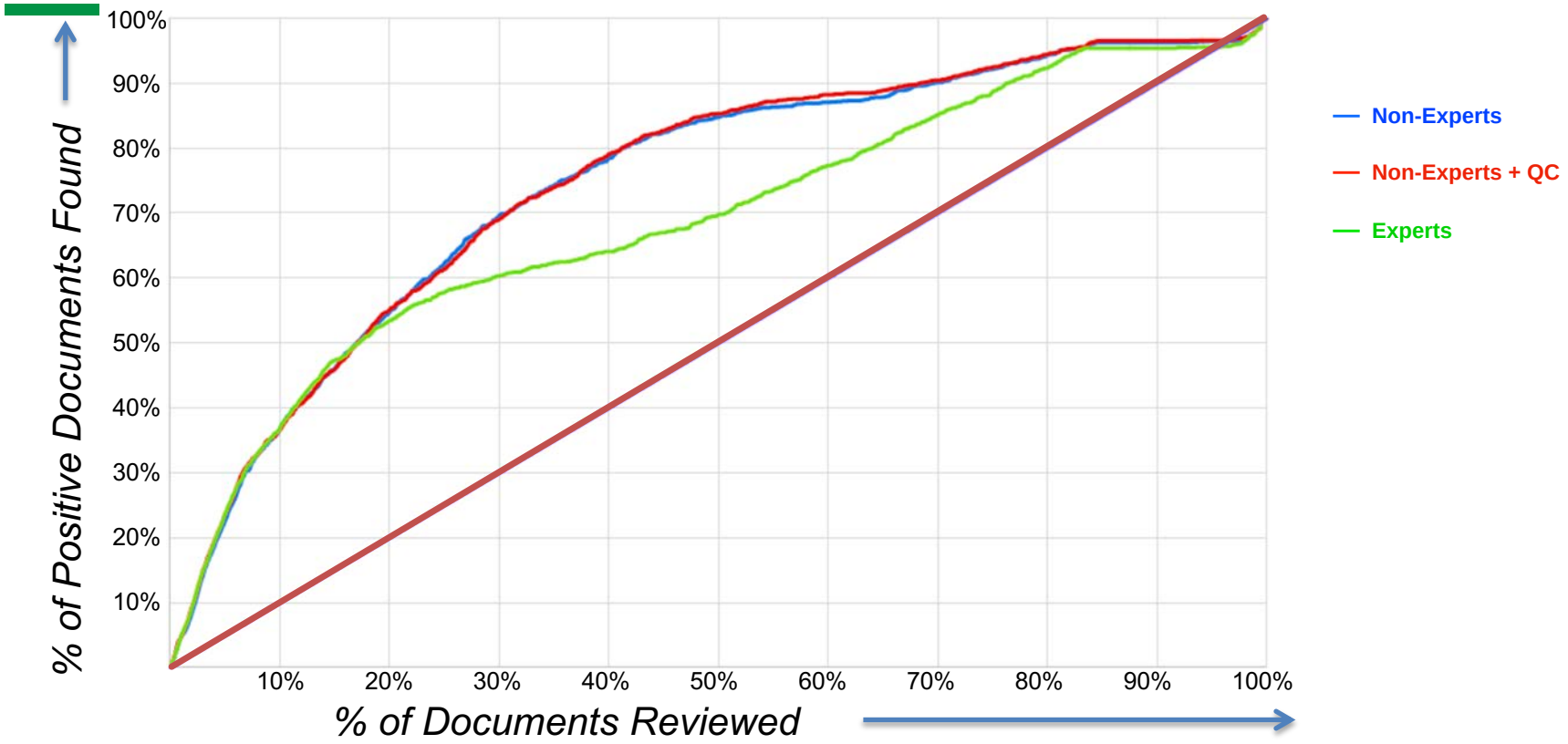
Simulation: Expert vs. Non-expert Training

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	**[docid:7643 = true]** **[docid:225 = false]**	**[docid:7643 = true]** **[docid:225 = true]**
Feature (Signal) Extraction	1-grams	1-grams
Ranking Engine	Logistic Regresssion	Logistic Regresssion
Training/Review Protocol	CAL	CAL
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

Expert – Non-Expert Results (Metric: Precision at 75% Recall)



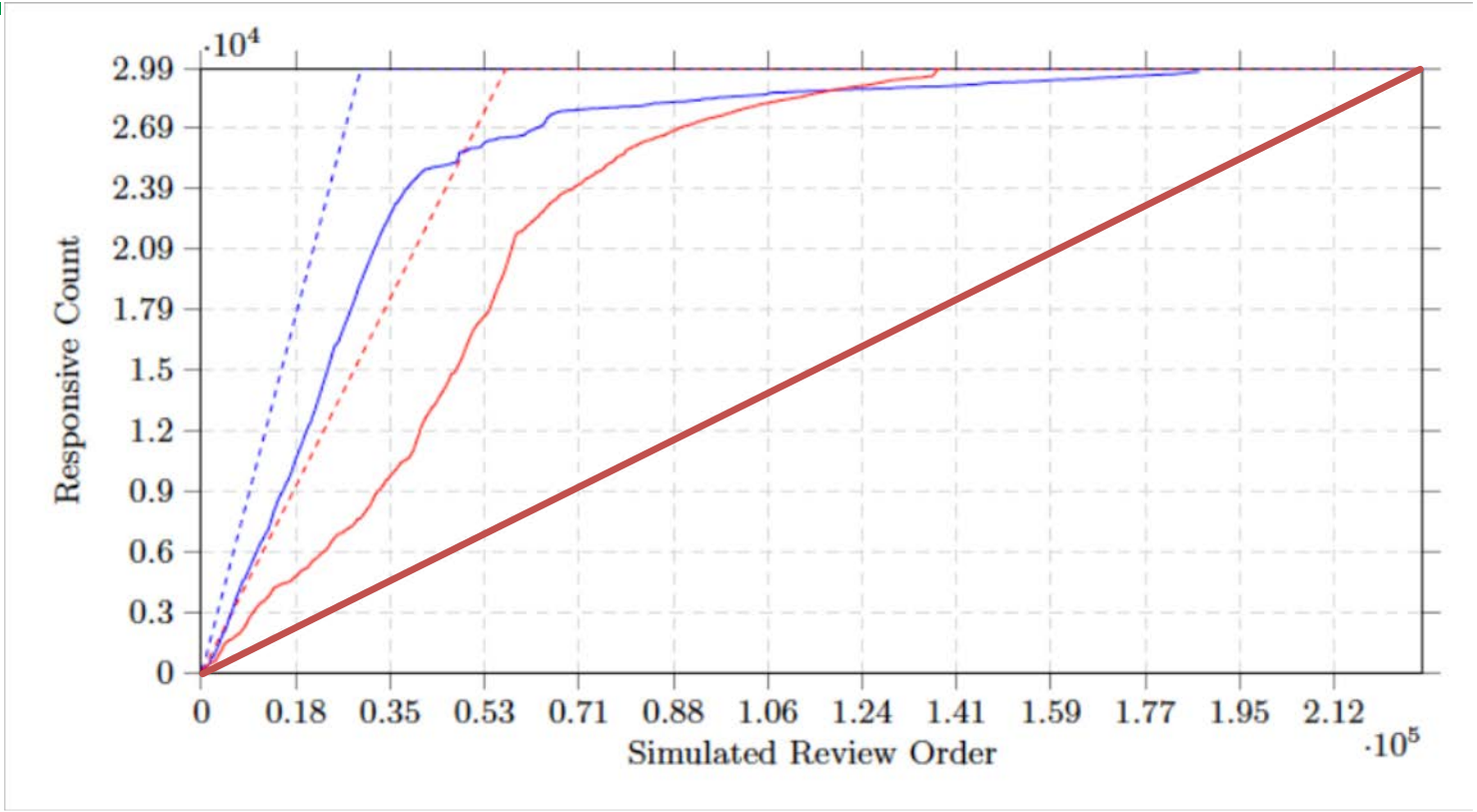
Expert – Non-Expert Results (Metric: Precision at 75% Recall)



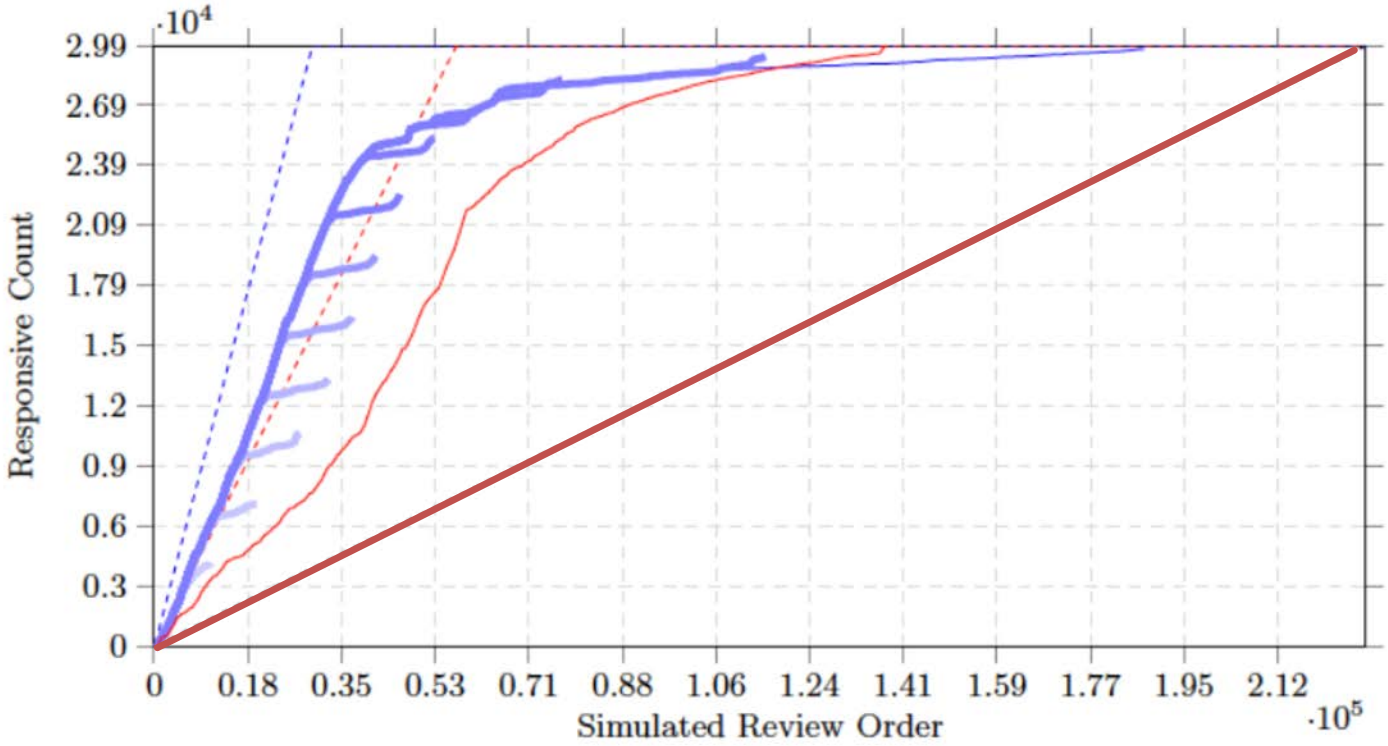
Simulation: Family vs. Document Batching

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	docids 5738, 83, 29973	docids 5738, 83, 29973
Feature (Signal) Extraction	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]
Training/Review Protocol	**CAL with Family Batching**	**CAL with Individual Doc**
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

Family Batching Results (Metric: Precision at 75% Recall)



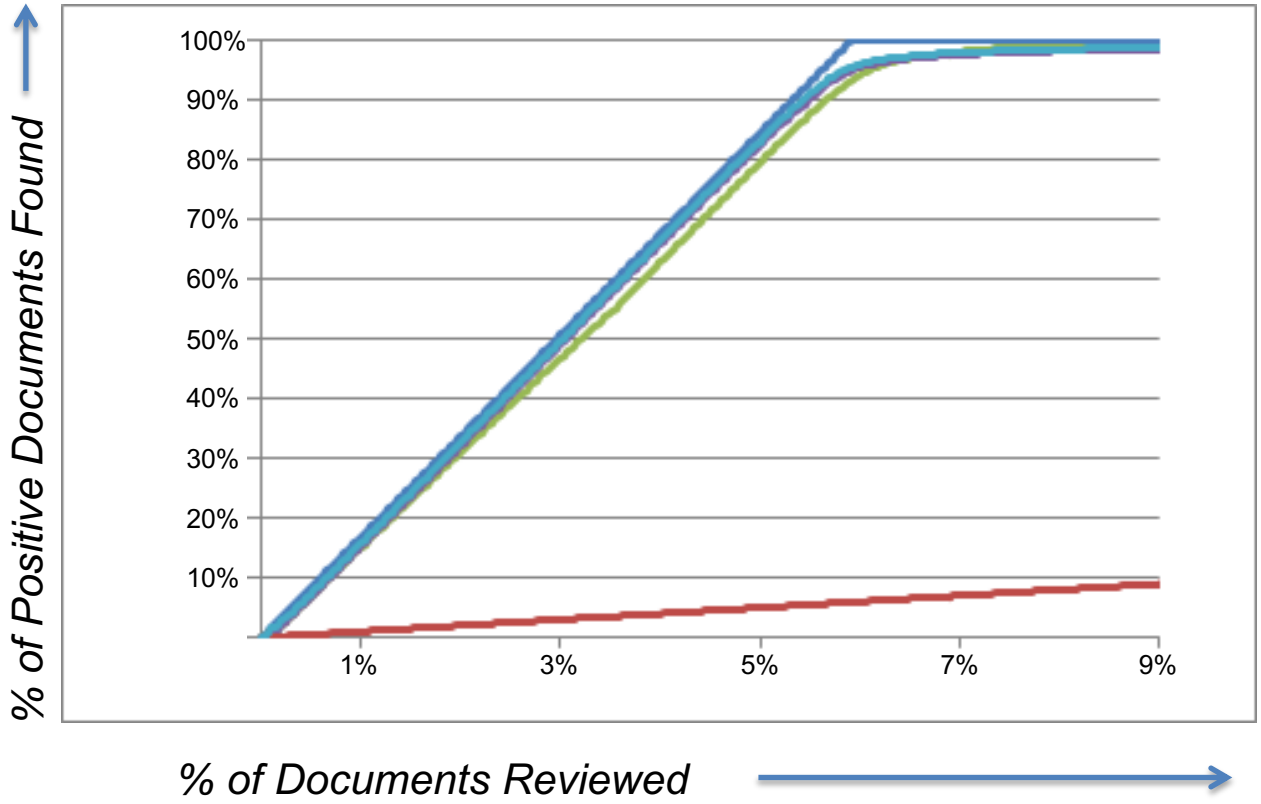
Family Batching Results (Metric: Precision at 75% Recall)



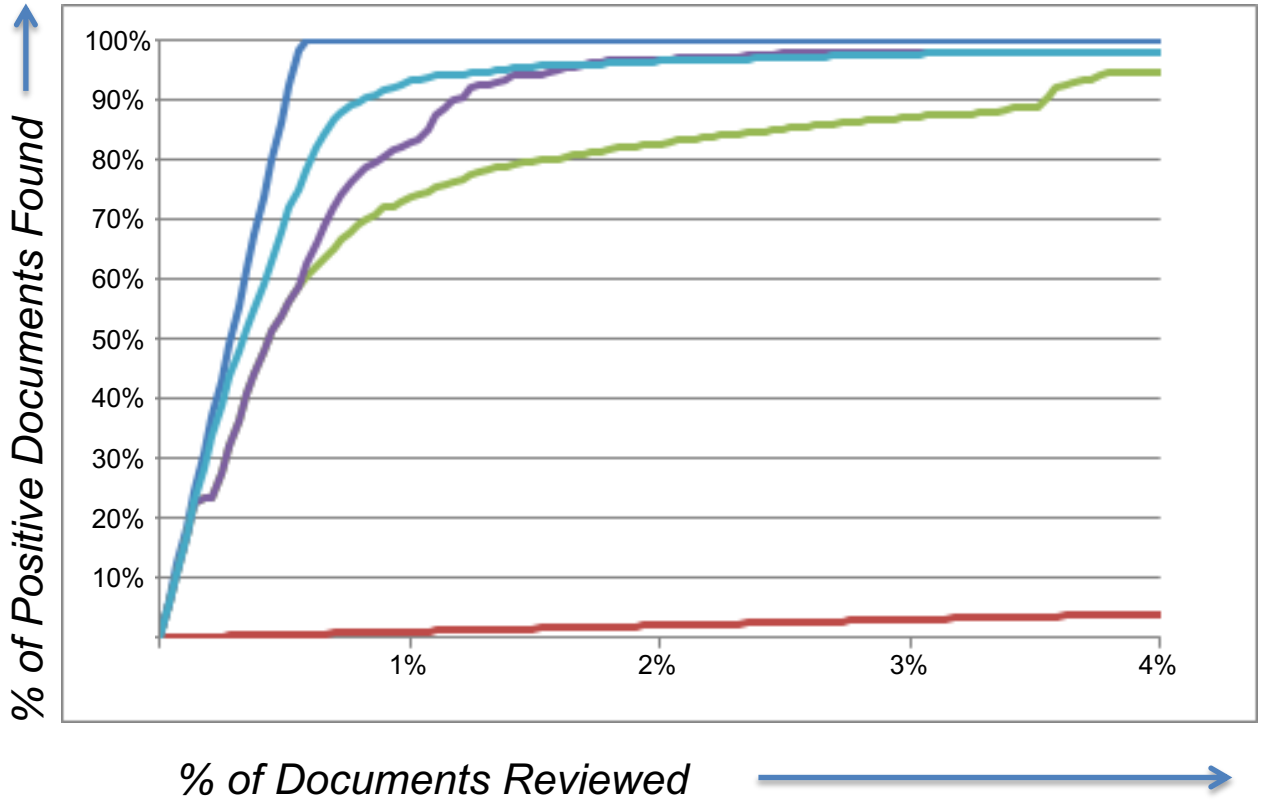
Simulation: Evaluate CAL Update Rate

	Condition 1	Condition 2	Condition 3
Document Corpus	Corpus Z	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	docids 5738, 83, 29973	docids 5738, 83, 29973	docids 5738, 83, 29973
Feature (Signal) Extraction	n-grams	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]	[Catalyst]
Training/Review Protocol	**CAL updated weekly**	**CAL updated daily**	**CAL updated 10 minutely**
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall	Precision@75% recall

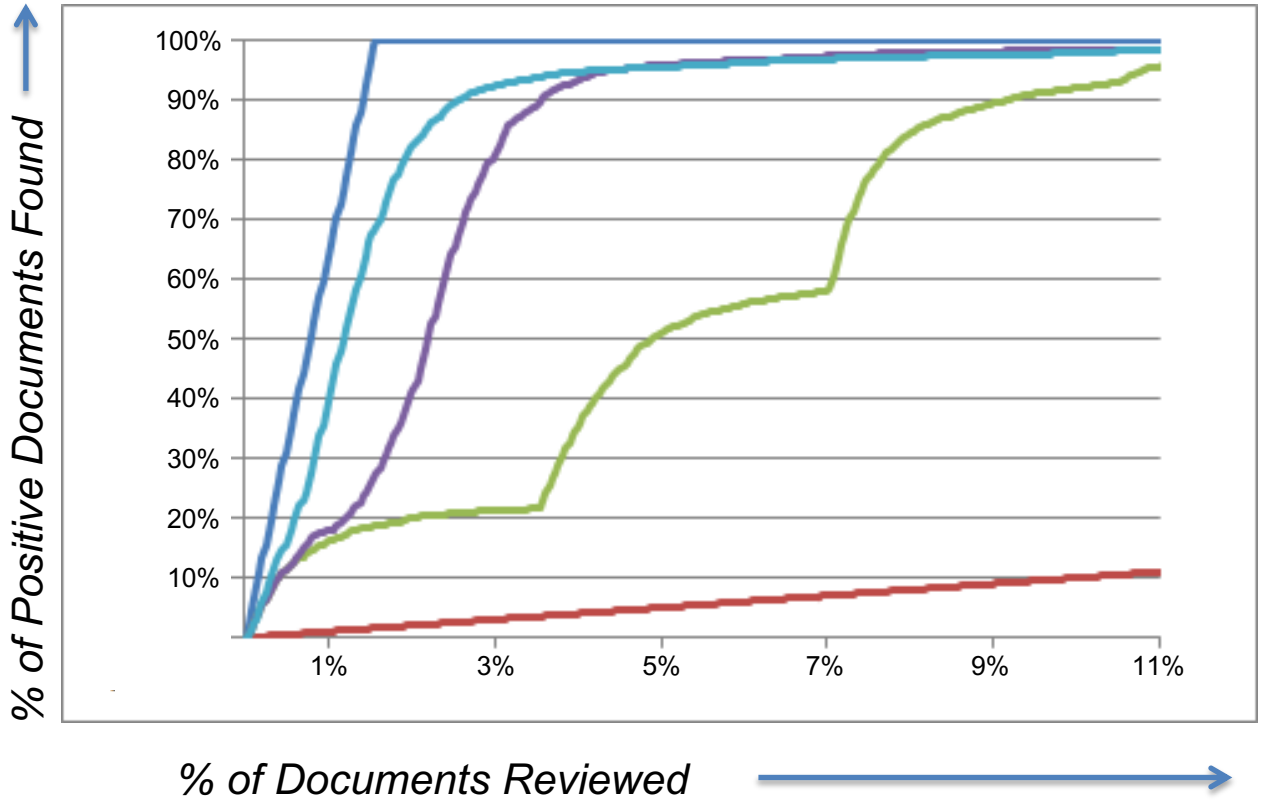
Update Rate Results (Metric: Precision at 75% Recall)



Update Rate Results (Metric: Precision at 75% Recall)



Update Rate Results (Metric: Precision at 75% Recall)

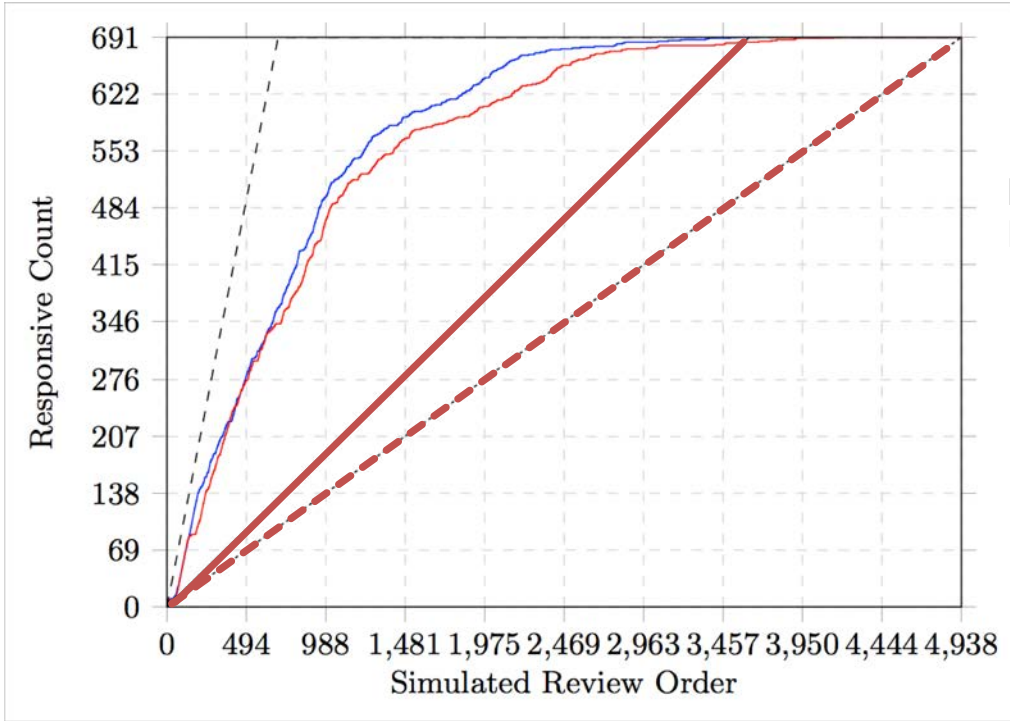


Simulation: Evaluate the Need for Culling

	Condition 1	Condition 2
Document Corpus	**Unculled Corpus X**	**Culled Corpus X'**
Starting Condition (e.g. seed documents, ad hoc query, etc.)	docids 5738, 83, 29973	docids 5738, 83, 29973
Feature (Signal) Extraction	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]
Training/Review Protocol	CAL	CAL
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

The Impact of Culling (Metric: Precision at 75% Recall)

Is it worth fighting over keyword culling?

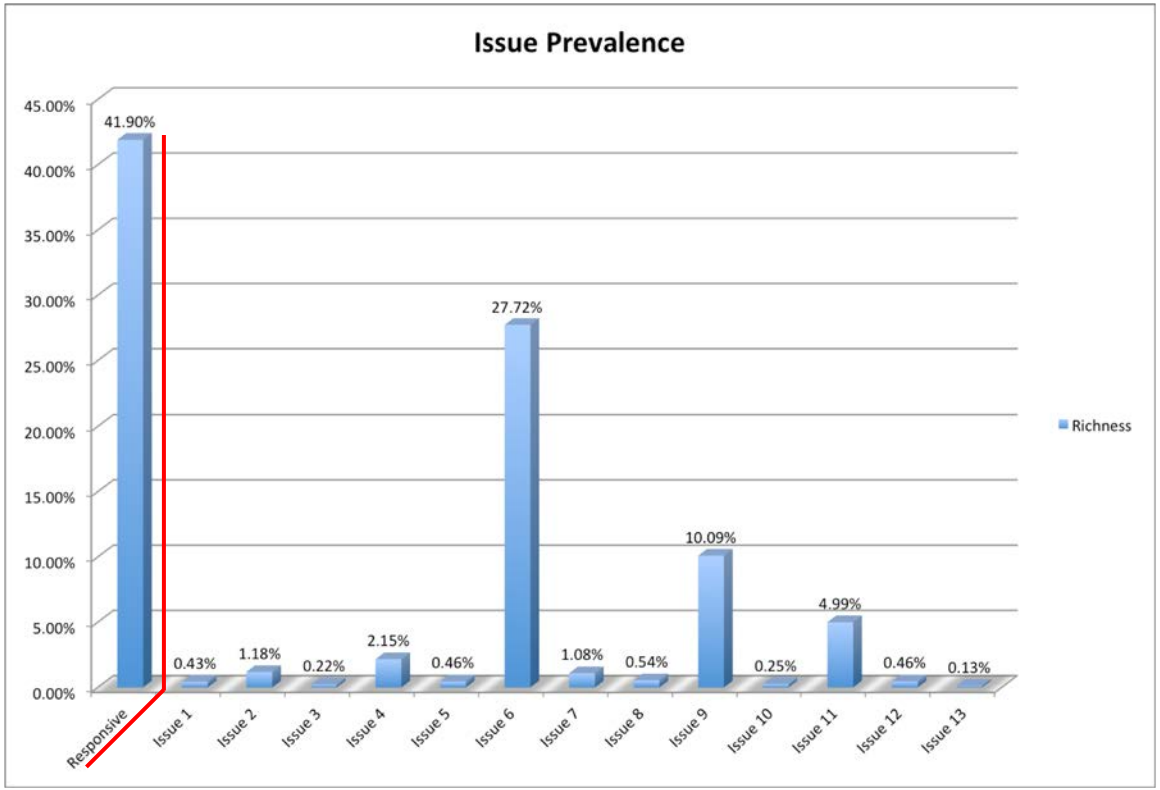


Blue: Culled Collection
Red: Not Culled

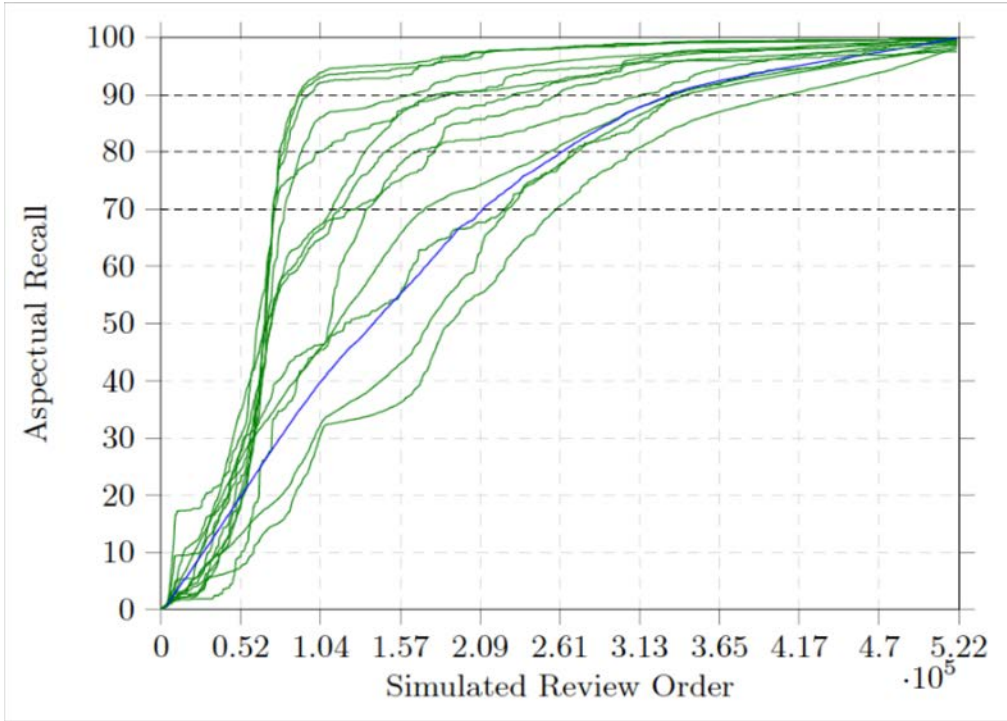
Simulation: Issue/Facet Effectiveness

	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	docids 5738, 83, 29973	docids 5738, 83, 29973
Feature (Signal) Extraction	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]
Training/Review Protocol	**CAL**	**Linear**
Ground Truth	true/false for responsive true/false for each facet	true/false for responsive true/false for each facet
Evaluation Metric	Precision@70%, 80%, 90% recall	Precision@70%, 80%, 90% recall

A Closer Look at the Facets



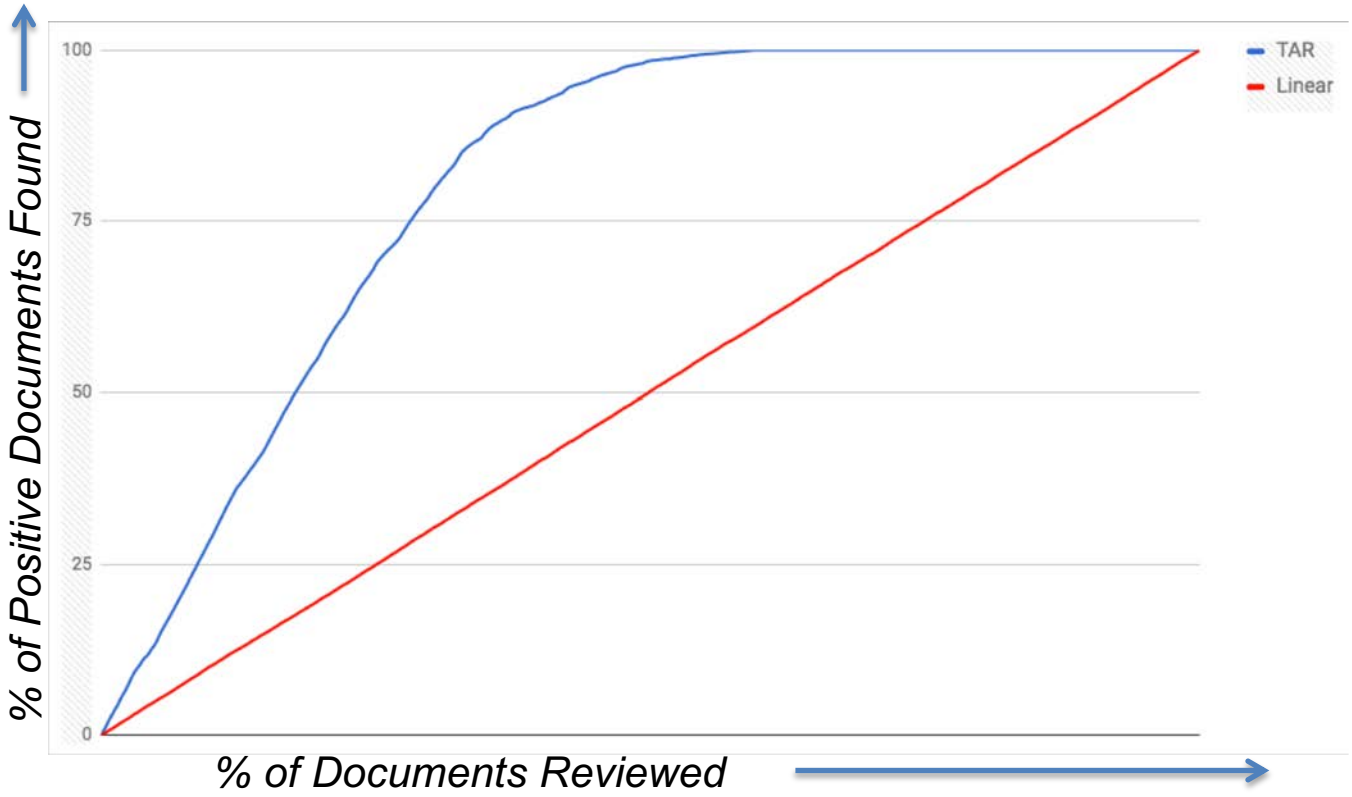
Facet Effectiveness (Metric: Precision at 75% Recall)



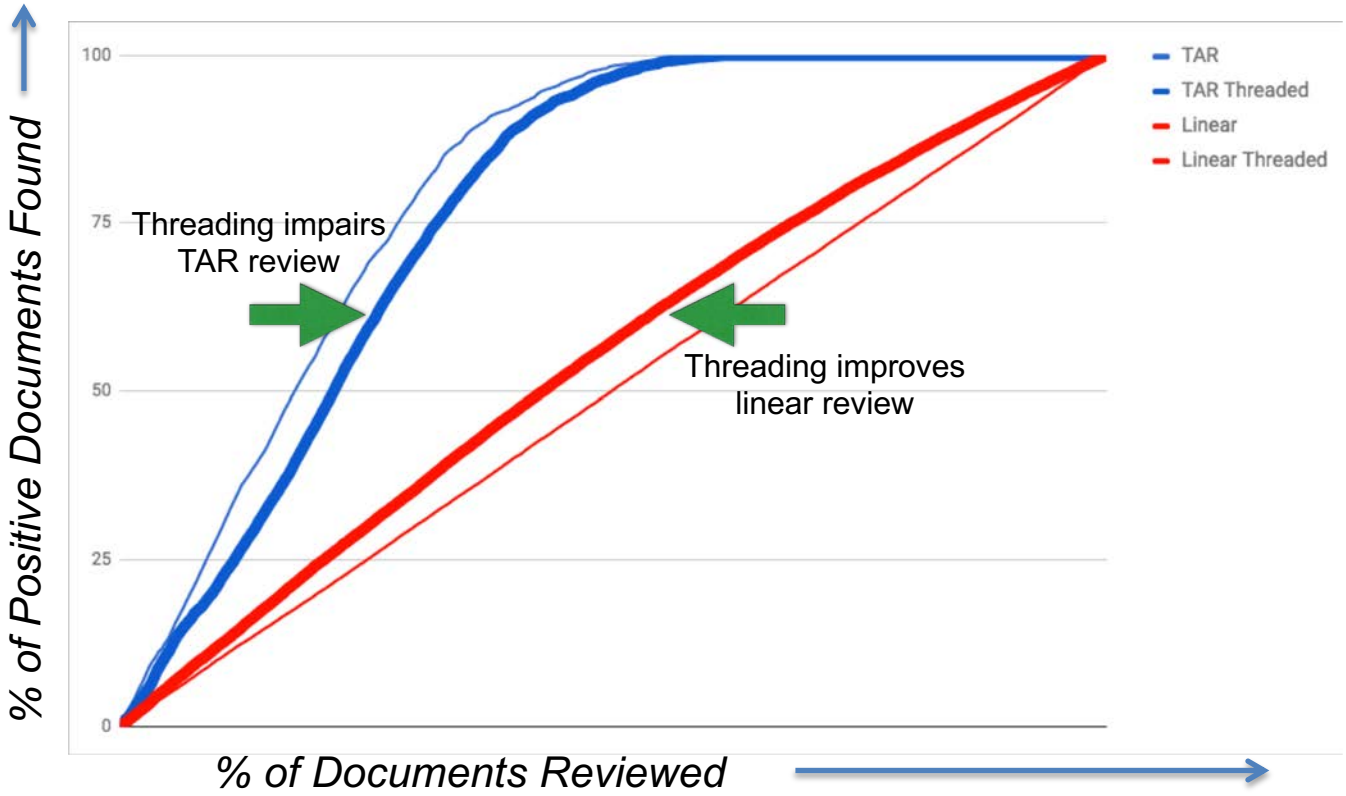
Simulation: Evaluate Threading Impact on Review Protocol

	Condition 1	Condition 2	Condition 3	Condition 4
Document Corpus	Corpus Z	Corpus Z	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	docids 5738, 83, 29973	docids 5738, 83, 29973	docids 5738, 83, 29973	docids 5738, 83, 29973
Feature (Signal) Extraction	n-grams	n-grams	n-grams	n-grams
Ranking Engine	[Catalyst]	[Catalyst]	[Catalyst]	[Catalyst]
Training/Review Protocol	**CAL without threading**	**CAL with threading**	**Linear without threading**	**Linear with threading**
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall	Precision@75% recall	Precision@75% recall

Review Without Threading (Metric: Precision at 75% Recall)



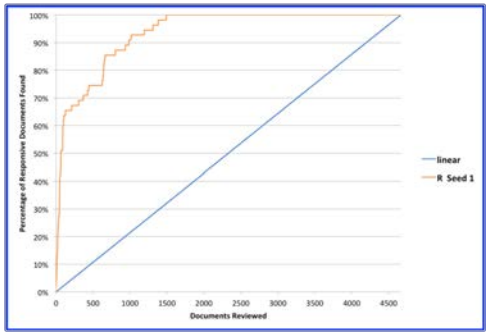
The Impact of Threading (Metric: Precision at 75% Recall)

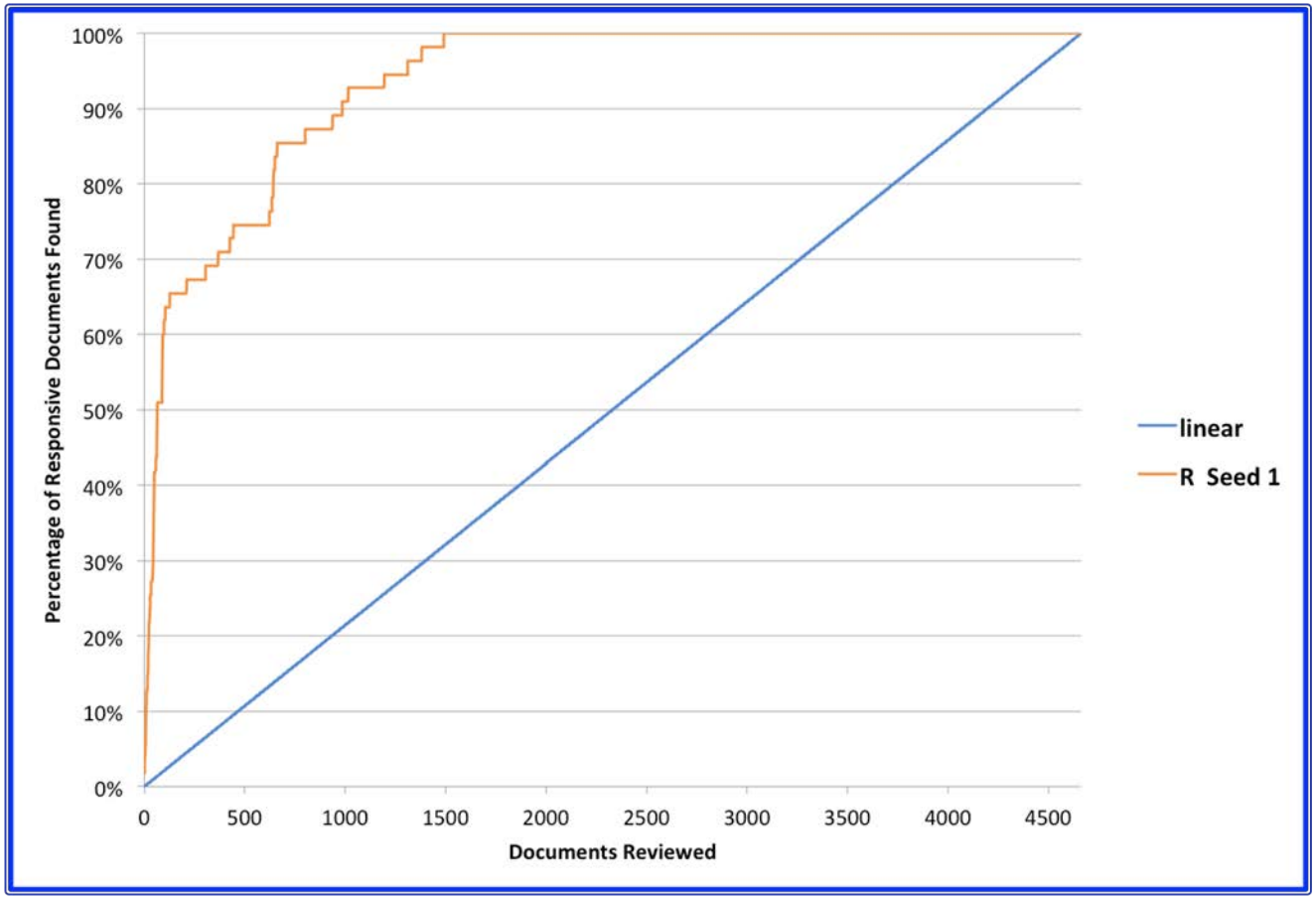


Simulation: Starting Seeds

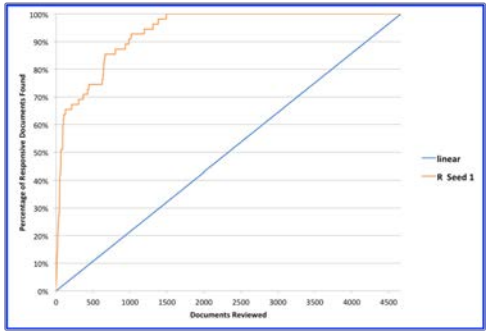
	Condition 1	Condition 2
Document Corpus	Corpus Z	Corpus Z
Starting Condition (e.g. seed documents, ad hoc query, etc.)	**Seed One**	**Seed 2-57**
Feature (Signal) Extraction	1-grams	1-grams
Ranking Engine	Logistic Regresssion	Logistic Regresssion
Training/Review Protocol	CAL	CAL
Ground Truth	[docid:7643 = true] [docid:225 = true] [docid:42 = false]	[docid:7643 = true] [docid:225 = true] [docid:42 = false]
Evaluation Metric	Precision@75% recall	Precision@75% recall

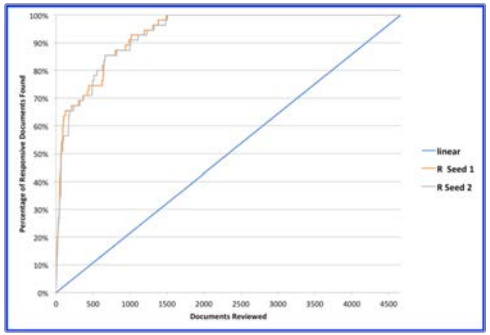
Single Seed

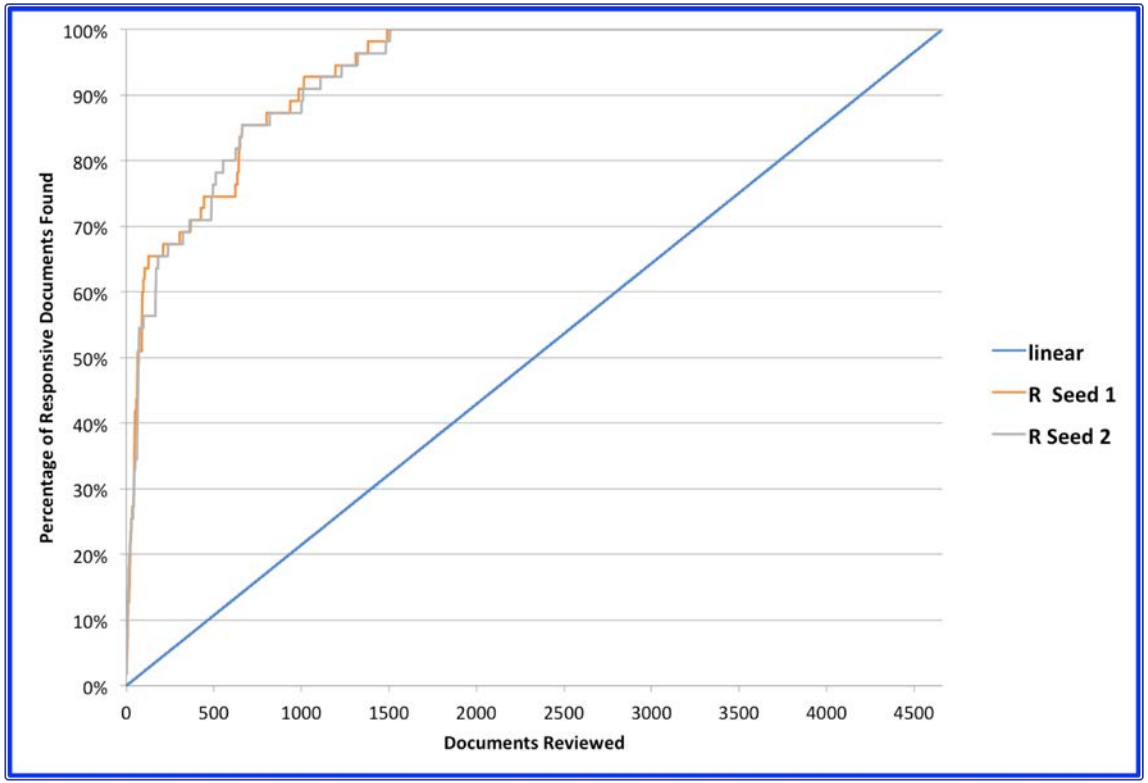

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Responsive Documents Found
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Documents Reviewed
linear
R Seed 1


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Responsive Documents Found
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Documents Reviewed
linear
R Seed 1

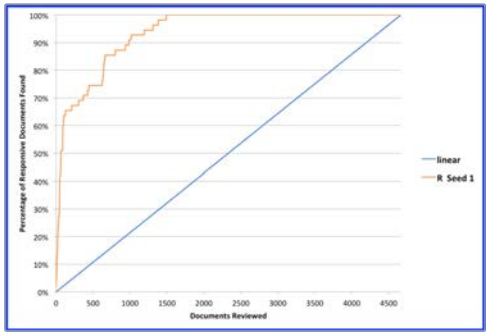
Single Seed

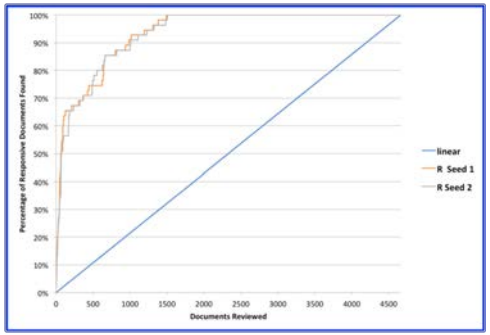

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1

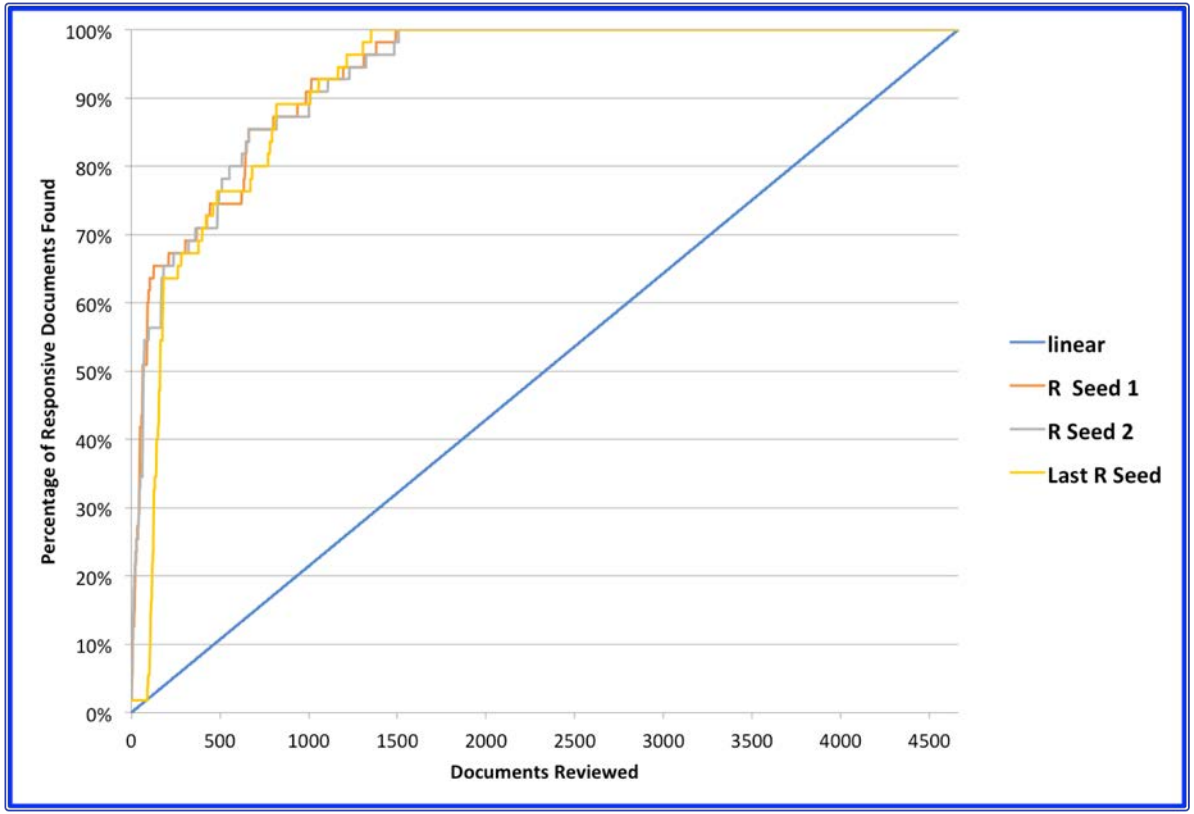

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Responsive Documents Found
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Documents Reviewed
linear
R Seed 1
R Seed 2

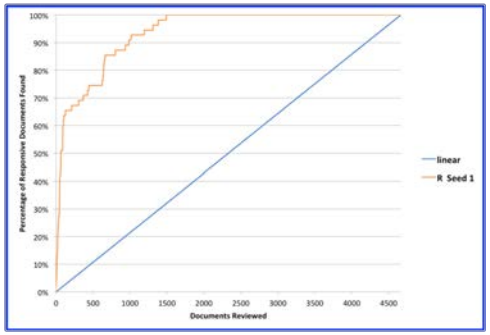
Single Seed

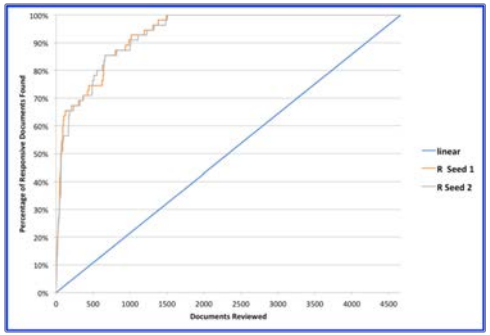

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1

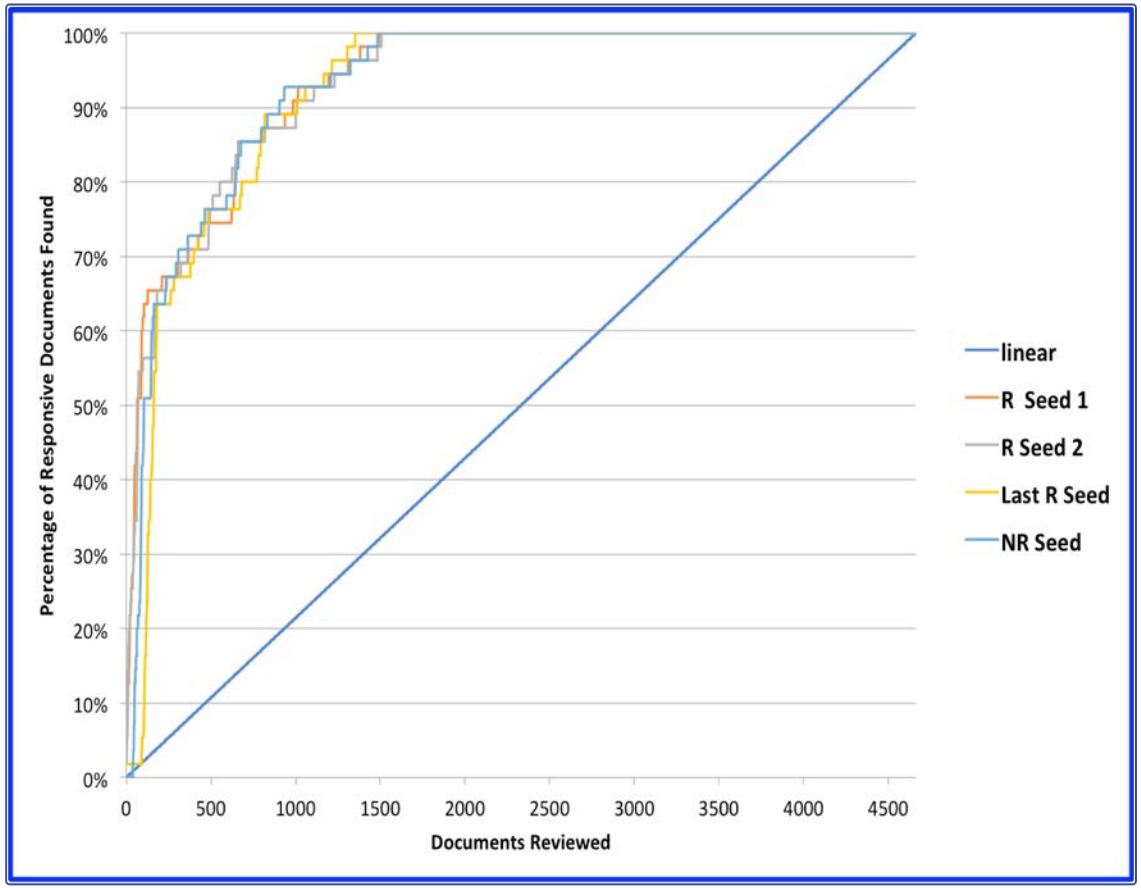

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Responsive Documents Found
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Documents Reviewed
linear
R Seed 1
R Seed 2
Last R Seed

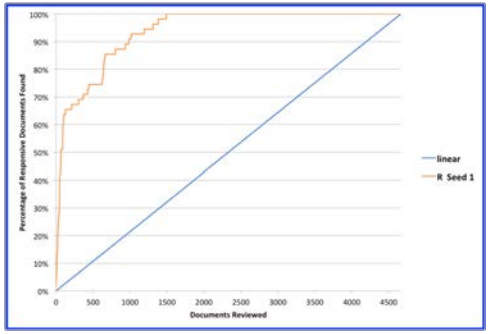
Single Seed

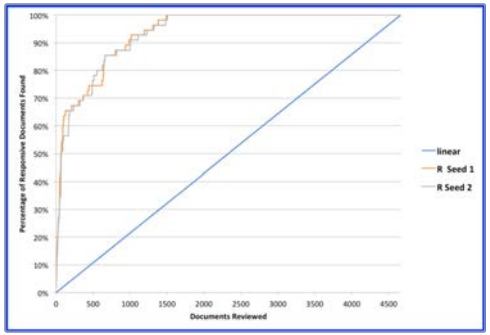

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1

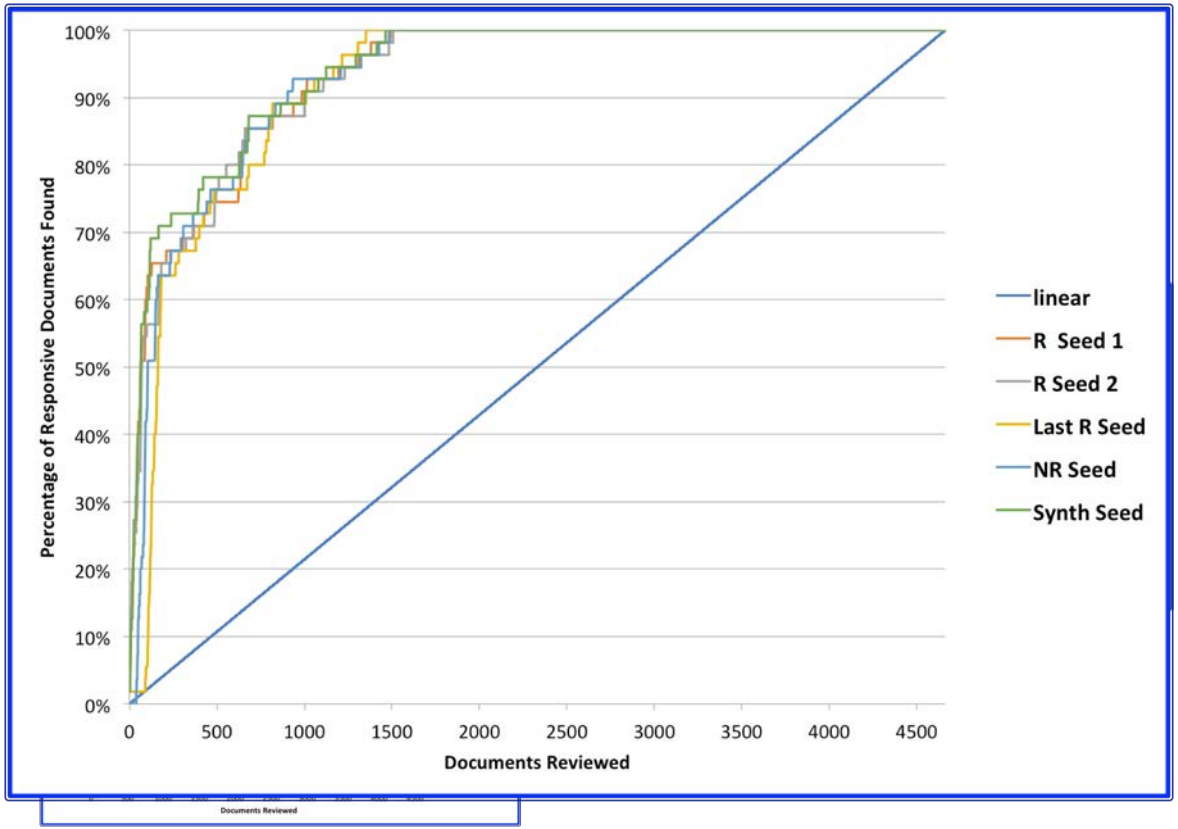

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2


Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2
Last R Seed
NR Seed

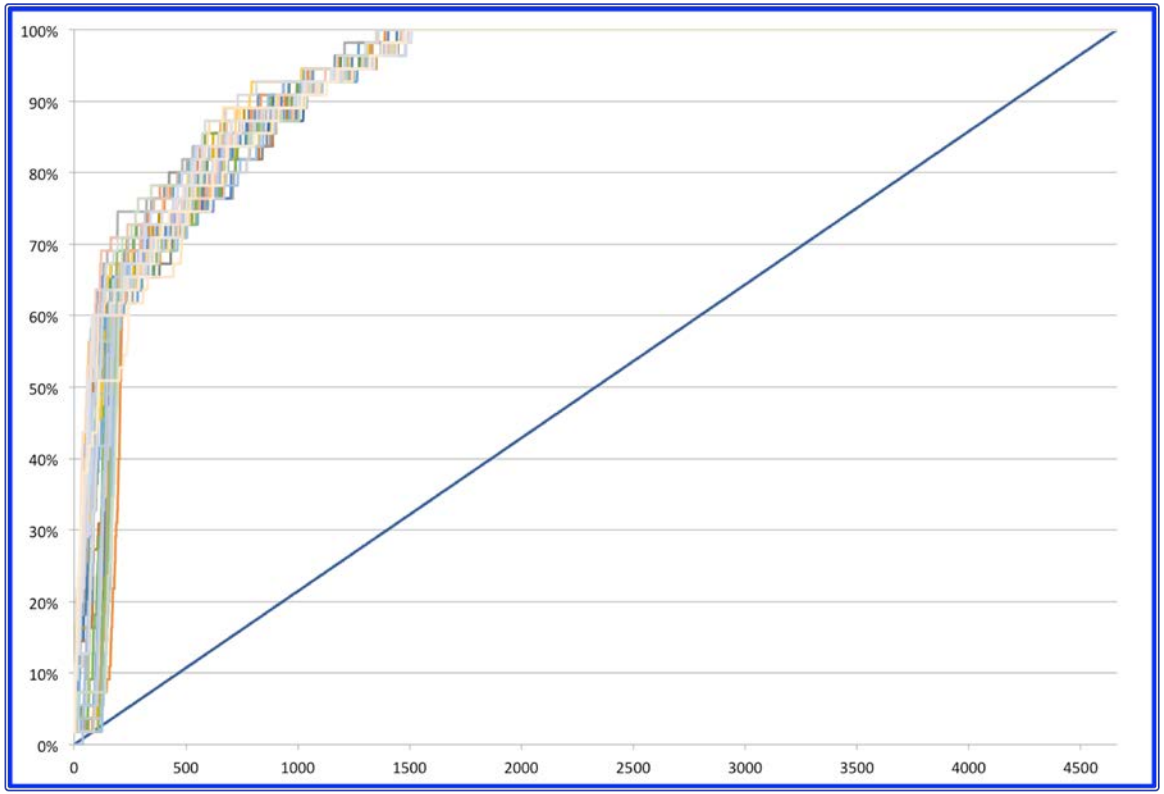
Single Seed


Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1


Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2

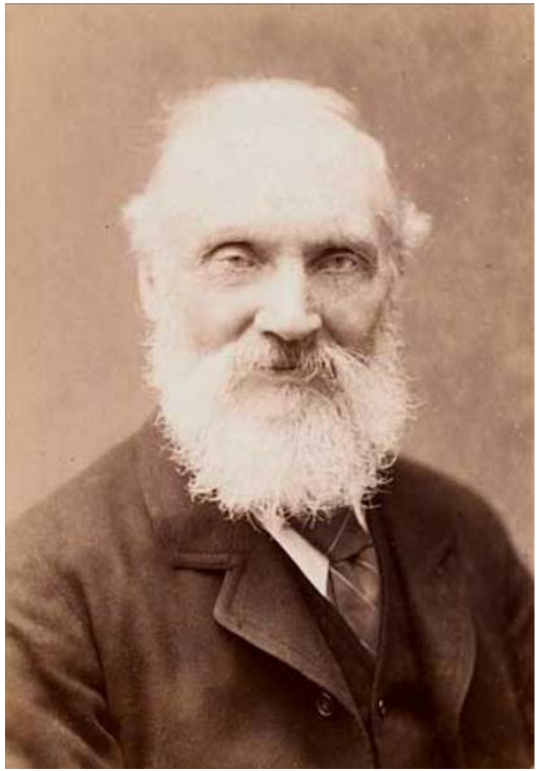

Percentage of Responsive Documents Found
Documents Reviewed
linear
R Seed 1
R Seed 2
Last R Seed
NR Seed
Synth Seed

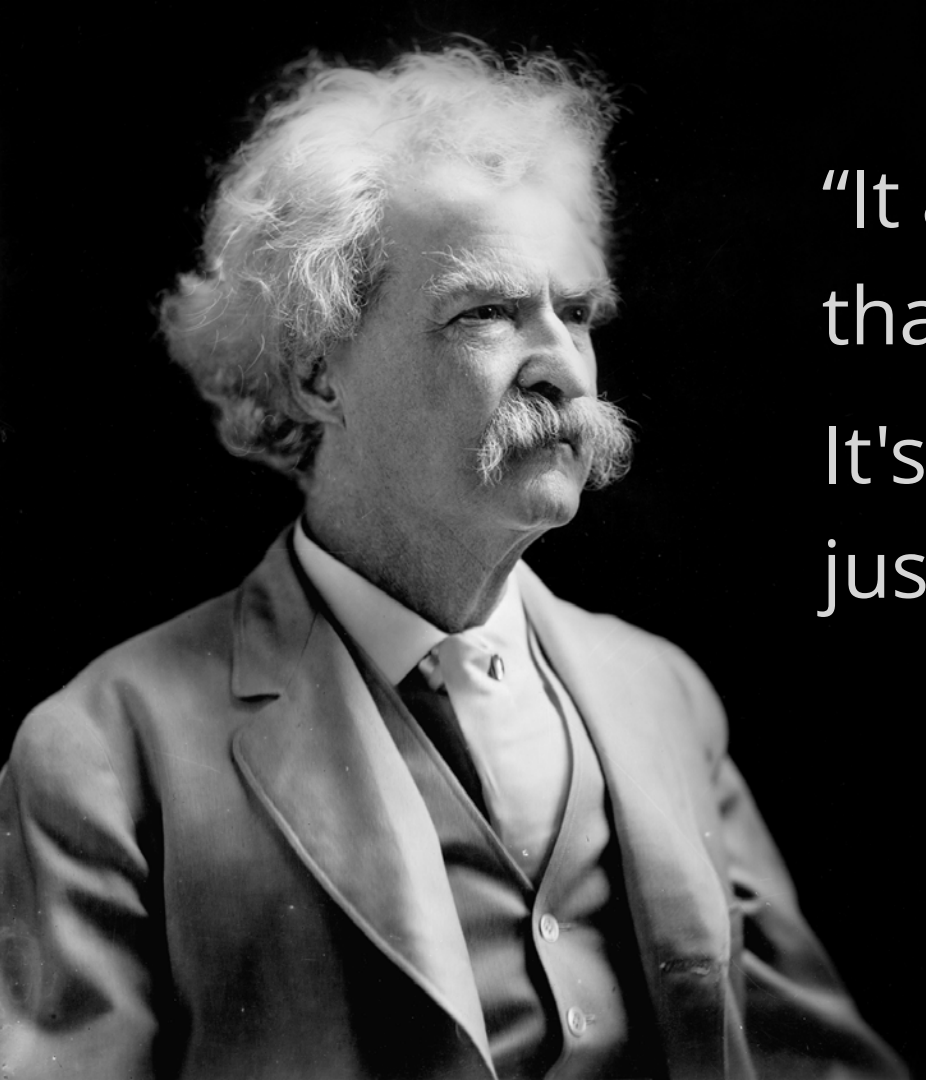
Single Seed – All Runs



What You Cannot Measure, You Cannot Improve

1. TAR is *not* a checklist of techniques or features
2. Combining techniques is not necessarily additive
   - If X is good and Y is good, then X + Y must be great! – WRONG!
3. Must consider all aspects of system *and* human performance as a holistic package





“It ain't what you don't know that gets you into trouble.

It's what you know for sure th just ain't so.”

Mark Twain

Catalyst

ASU Arkfeld Analytics

Five Years of CAL

The Case for Testing and What it Tells Us

John Tredennick
Founder Catalyst