Beyond Keywords: Emerging Best Practices in the Area of Search and Information Retrieval

New Mexico Digital Preservation Conference
(DIG IN)
June 5, 2008

Jason R. Baron
Director of Litigation
Office of General Counsel
National Archives and Records Administration

Overview

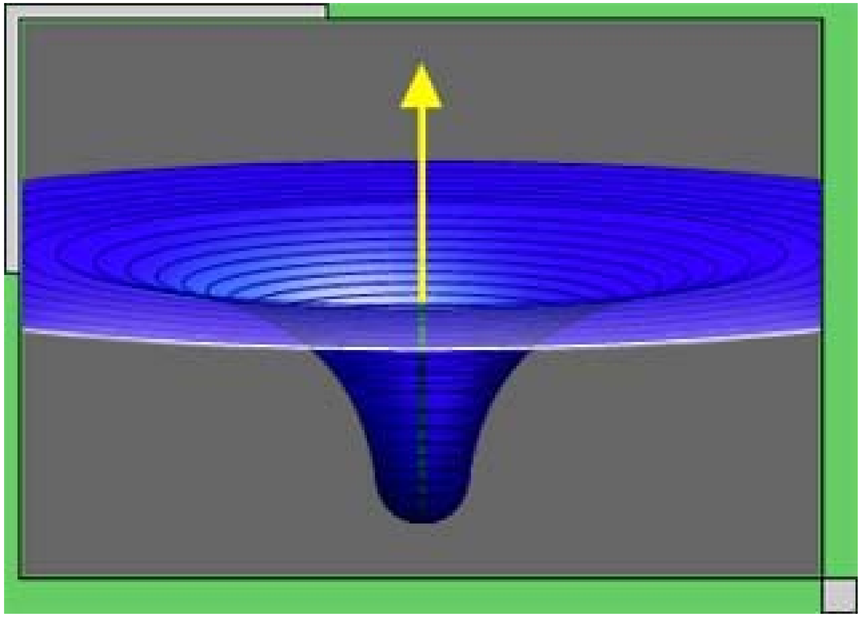
- Introduction: Myth, Hype, Reality – Information Retrieval and the Problem of Language
- Case Study: *U.S. v. Philip Morris*
- The TREC Legal Track: Initial Results
- Strategic Challenges & Litigation Risk: Thinking About Search Issues

References

Definition of “ESI”

- A new legal term of art: “electronically stored information” to supplement the older term “documents”:
    - *The wide variety of computer systems currently in use, and the rapidity of technological change,counsel against a limiting or precise definition of ESI…A common example [is] email … The rule … [is intended] to encompass future developments in computer technology. --Advisory Committee Notes to Rule 34(a), 2006 Amendments to the Federal Rules of Civil Procedure*

Information Inflation: The Expanding ESI Universe . . . .



The Myth of Search & Retrieval

When lawyers request production of “all” relevant documents (and now ESI), all or substantially all will in fact be retrieved by existing manual or automated methods of search.

Corollary: in conducting automated searches, the use of “keywords” alone will reliably produce all or substantially all documents from a large document collection.

The “Hype” on Search & Retrieval

Claims in the legal tech sector that a very high rate of “recall” *(i.e., finding all relevant documents) is easily obtainable provided one uses a particular software product or service.

The Reality of Search & Retrieval

+ Past research (Blair & Maron, 1985) has shown a gap or disconnect between lawyers' perceptions of their ability to ferret out relevant documents, and their actual ability to do so:

--in a 40,000 document case (350,000 pages), lawyers estimated that a manual search would find 75% of relevant documents, when in fact the research showed only 20% or so had been found.

More Reality: IR is Hard

+ Information retrieval (IR) is a hard problem: difficult even with English-language text, and even harder with non-textual forms of ESI (audio, video, etc.) caught up in litigation.

+ A vast field of IR research exists, including some fundamental concepts and terminology, that lawyers would benefit from having greater exposure with.

Why is IR hard (in general)?

+ Fundamental ambiguity of language
+ Human errors
+ OCR problems
+ Non-English language texts
+ Nontextual ESI (in .wav, .mpg, .jpg formats, etc.)
+ Lack of helpful metadata

Problems of language

Polysemy: ambiguous terms (e.g., “George Bush,” “strike,”)

Synonymy: variation in describing same person or thing in multiplicity of ways (e.g., “diplomat,” “consul,” “official,” ambassador,” etc.)

Pace of change: text messaging, computer gaming as latest examples (e.g., “POS,” “1337”)

Why is IR hard (for lawyers)?

- Lawyers not technically grounded
- Traditional lawyering doesn't emphasize front-end "process" issues that would help simplify or focus search problem in particular contexts
- The reality is that huge sources of heterogeneous ESI exist, presenting an array of technical issues
- Deadlines and resource constraints
- Failure to employ best strategic practices

Snapshot of 2008 ESI Heterogeneity

E-mail, integrated with voice mail & VOIP, word processing (including not in English), spreadsheets, dynamic databases, instant messaging, Web pages including intraweb sites, Blogs, wikis, and RSS feeds, backup tapes, hard drives, removable media, flash drives, new storage devices, remote PDAs, and audit logs and metadata of all types.

Sedona Guideline 11 (2007)

A responding party may satisfy its good faith obligation to preserve and produce potentially relevant electronically stored information by using electronic tools ad processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information.

www.thesedonaconference.org

Case Study: U.S. v. Philip Morris – Overall Discovery

- 1,726 Requests to Produce propounded by tobacco companies on U.S. (30 federal agencies, including NARA) for tobacco related records
- Along with paper records, email records were made subject to discovery
- 32 million Clinton era email records – government had burden of searching

Case Study: U.S. v. Philip Morris (con't) – Employing a limited feedback loop

- **Original set of 12 keywords searched unilaterally**
- **After informal negotiations, additional terms explored**
- **Sampling against database to find "noisy" terms generating too many false positives (Marlboro, PMI, TI, etc.)**
- **Report back and consensus on what additional terms would be in search protocol.**

Example of Boolean search string from *U.S. v. Philip Morris*

- (((master settlement agreement OR msa) AND NOT (medical savings account OR metropolitan standard area)) OR s. 1415 OR (ets AND NOT educational testing service) OR (liggett AND NOT sharon a. liggett) OR atco OR lorillard OR (pmi AND NOT presidential management intern) OR pm usa OR rjr OR (b&w AND NOT photo*) OR phillip morris OR batco OR ftc test method OR star scientific OR vector group OR joe camel OR (marlboro AND NOT upper marlboro)) AND NOT (tobacco* OR cigarette* OR smoking OR tar OR nicotine OR smokeless OR synar amendment OR philip morris OR r.j. reynolds OR ("brown and williamson") OR ("brown & williamson") OR bat industries OR liggett group)

U.S. v. Philip Morris E-mail Winnowing Process

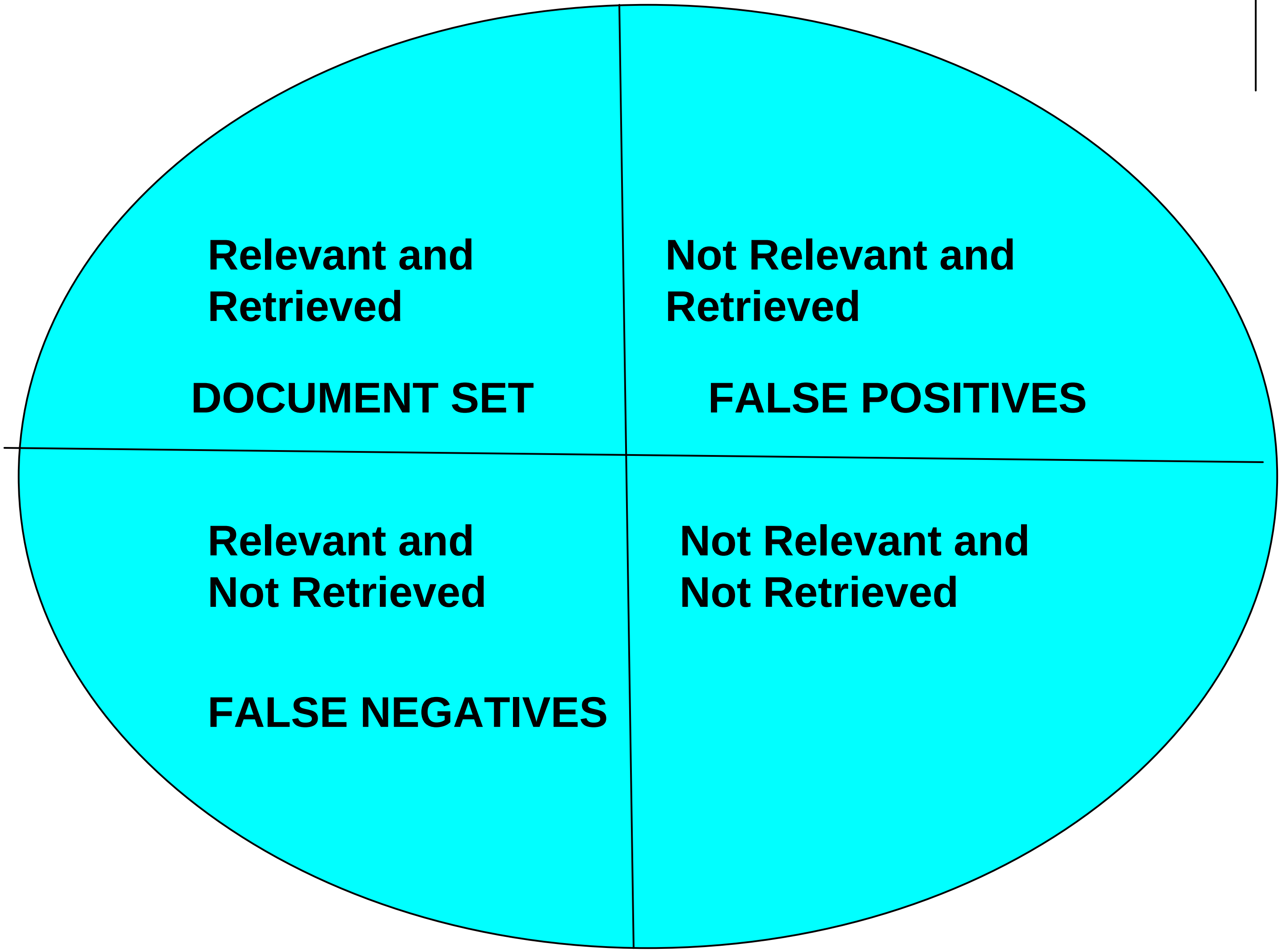
20 million	**→**	**200,000**	**→**	**100,000**	**→**	**80,000**	**→**	**20,000**
email		**hits based**		**relevant**		**produced**		**placed on**
records		**on keyword**		**emails**		**to opposing**		**privilege**
		terms used				**party**		**logs**
		(1%)						

- **→ A PROBLEM: only a handful entered as exhibits at trial**
- **→ A BIGGER PROGLEM: the 1% figure does not scale**

Litigation Targets

+ Defining “relevance”
+ Maximizing # responsive docs
+ Minimizing retrieval “noise” or false positives (non-responsive docs)

FINDING RESPONSIVE DOCUMENTS IN A LARGE DATA SET: FOUR LOGICAL CATEGORIES



Relevant and Retrieved

DOCUMENT SET

Not Relevant and Retrieved

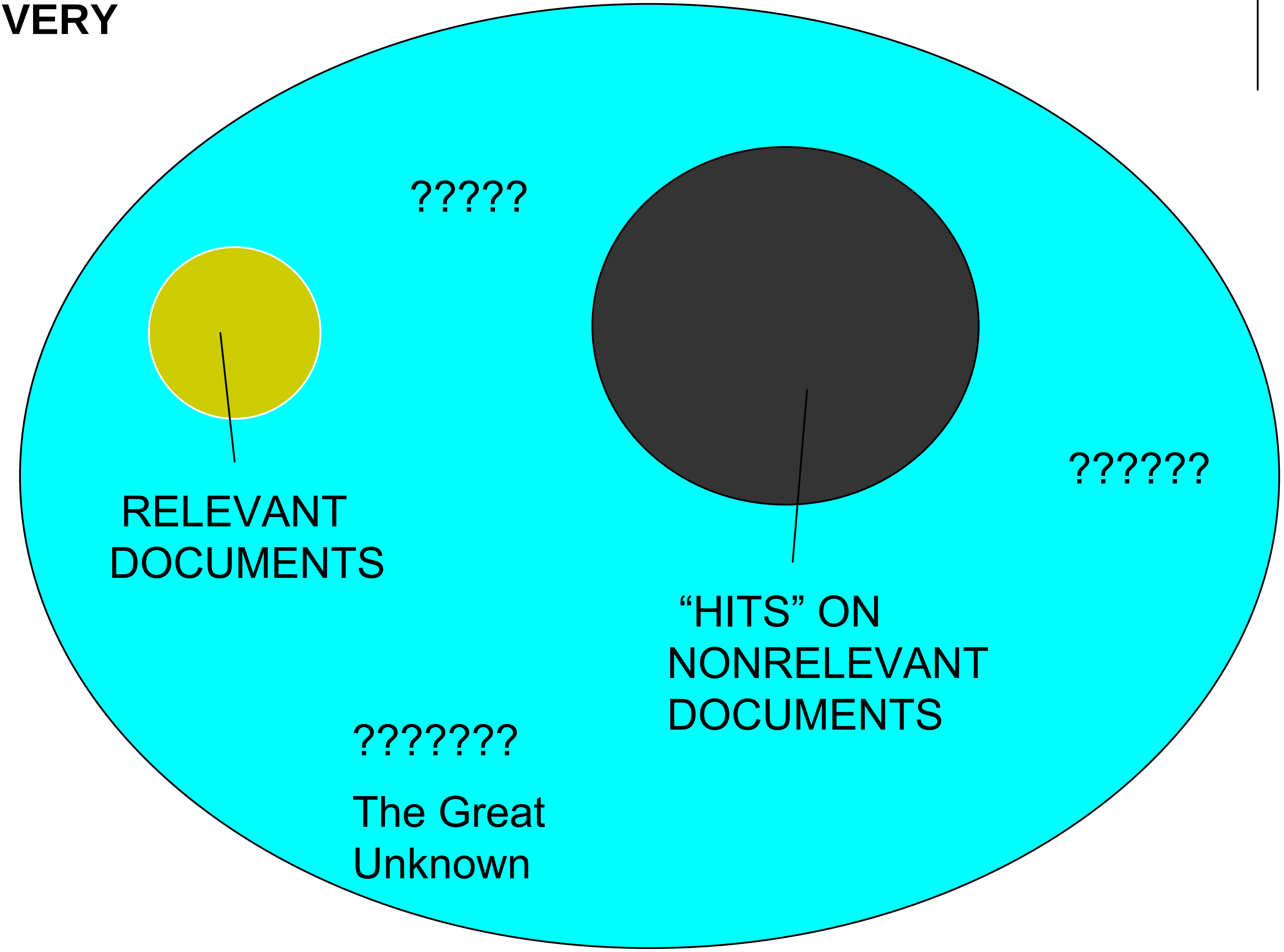
FALSE POSITIVES

Relevant and Not Retrieved

FALSE NEGATIVES

Not Relevant and Not Retrieved

FINDING RESPONSIVE DOCUMENTS IN A LARGE DATA SET: THE REALITY OF LARGE SCALE DISCOVERY



Measures of Information Retrieval

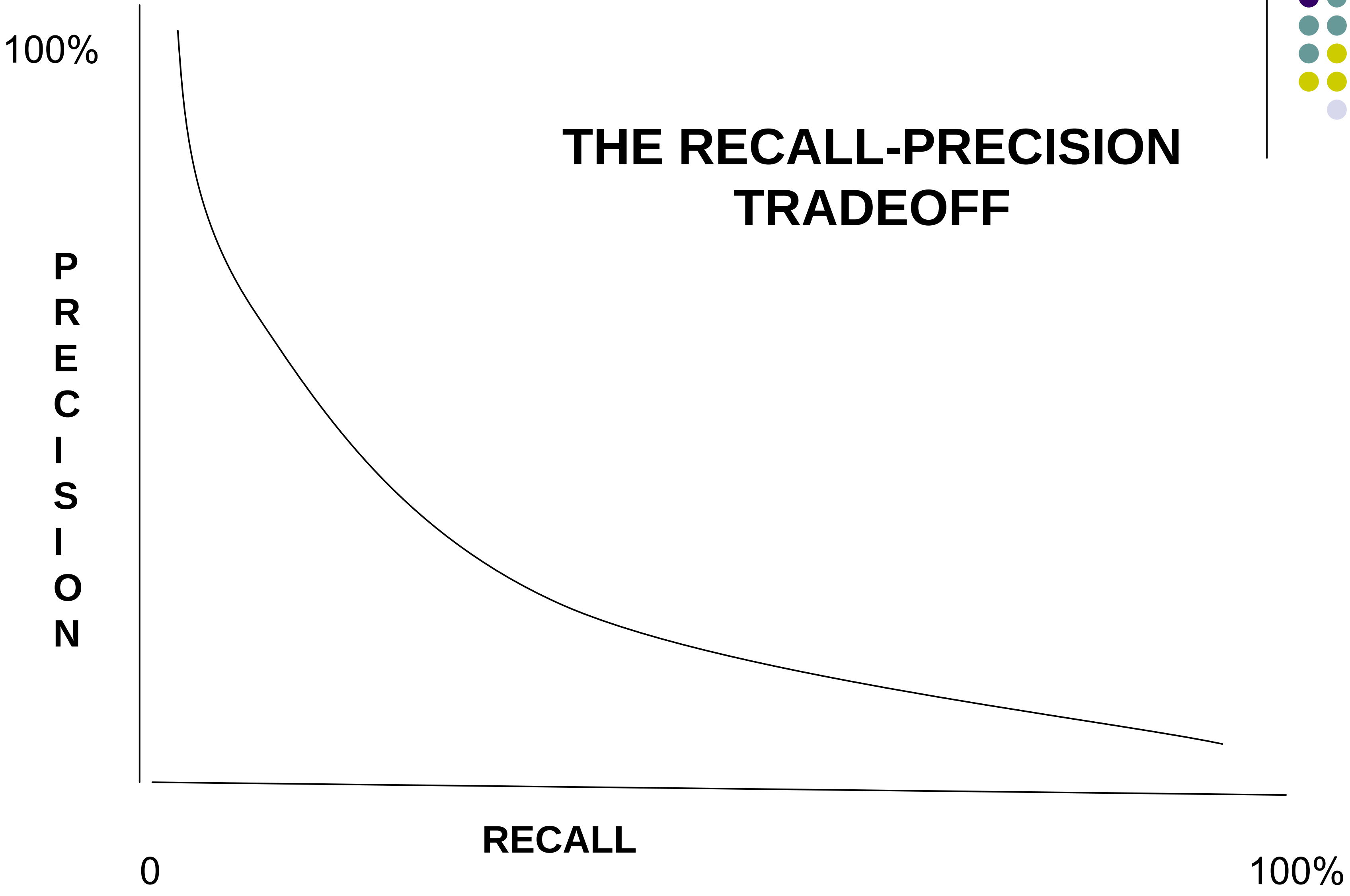
Recall =

$$\text{Recall} = \frac{\text{\# of responsive docs retrieved}}{\text{\# of responsive docs in collection}}$$

Measures of Information Retrieval

$$\text{Precision} = \frac{\text{\# of responsive docs retrieved}}{\text{\# of docs retrieved}}$$

THE RECALL-PRECISION TRADEOFF



Three Questions

(1) How can one go about improving rates of recall and precision (so as to find a greater number of relevant documents, while spending less overall time, cost, etc., sifting through noise?)

(2) What alternatives to keyword searching exist?

(3) Are there ways in which to benchmark alternative search methodologies so as to evaluate their efficacy?

Beyond Keywords: Alternative Search Methods

- *Greater Use Made of Boolean Strings*
- *Fuzzy Search Models*
- *Probabilistic models (Bayesian)*
- *Statistical methods (clustering)*
- *Machine learning approaches to semantic representation*
- *Categorization tools: taxonomies and ontologies*
- *Social network analysis*

Reference: *Appendix to The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (Aug. 2007 Public Comment Draft), available at http://www.thesedonaconference.org (link to publications)*

What is TREC?

- Conference series co-sponsored by the National Institute of Standards and Technology (NIST) and the Advanced Research and Development Activity (ARDA) of the Department of Defense
- Designed to promote research into the science of information retrieval
- First TREC conference was in 1992
- 15th Conference held November 15-17, 2006 in U.S. in Gaithersburg, Maryland (NIST headquarters)

TREC 2006/2007 Legal Track

- The TREC Legal Track was designed to evaluate th effectiveness of search technologies in a real-world legal context
- First of a kind study using nonproprietary data since Blair/Maron research in 1985
- 9 hypothetical complaints and 80+ “requests to produce” drafted by Sedona Conference members
- “Boolean negotiations” conducted as a baseline for search efforts
- Documents to be searched were drawn from a publicly available 7 million document tobacco litigation Master Settlement Agreement database
- Participating teams of information scientists from around the world contributing computer runs

Legal Track Research Teams 2007

Carnegie Mellon U

Dartmouth College

Long Island U

Sabir Research, Inc.

U of Iowa

U of Massachusetts

U of Maryland

U of Missouri, Kansas City

U of Washington

Ursinus College

Fudan U (CH)

National U of Singapore (SG)

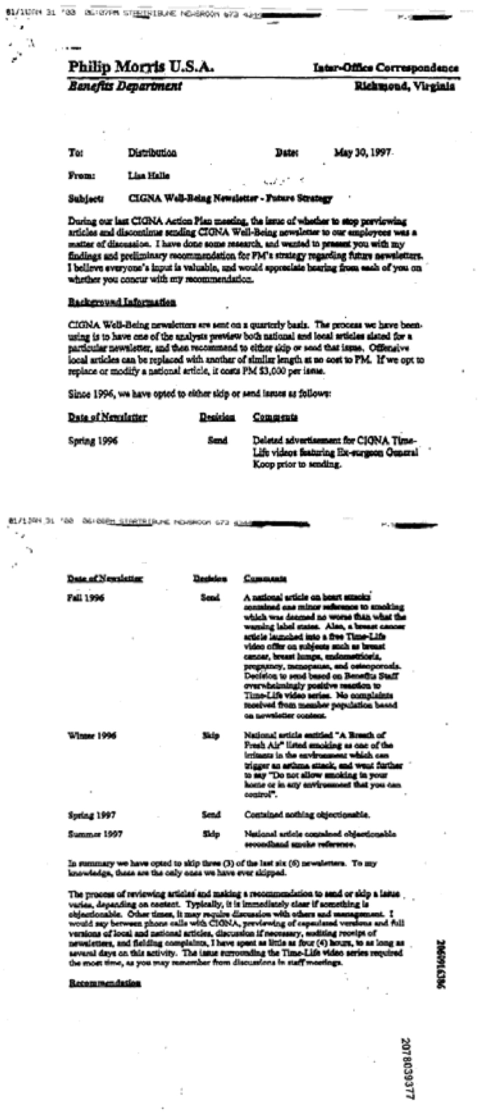
Open Text Corporation (CA)

U of Amsterdam (NL)

U of Waterloo (CA)

TREC Legal Track: Documents

Scanned



01/10AN 31 '00 06:07PM STARTRIBUNE NEWSROOM 673 4[redacted] P.[redacted]

Philip Morris U.S.A. **Inter-Office Correspondence**
Benefits Department **Richmond, Virginia**

To: Distribution Date: May 30, 1997

From: Lisa Halle

Subject: CIGNA Well-Being Newsletter - Future Strategy

During our last CIGNA Action Plan meeting, the issue of whether to stop previewing articles and discontinue sending CIGNA Well-Being newsletter to our employees was a matter of discussion. I have done some research, and wanted to present you with my findings and preliminary recommendation for PM's strategy regarding future newsletters. I believe everyone's input is valuable, and would appreciate hearing from each of you on whether you concur with my recommendation.

Background Information

CIGNA Well-Being newsletters are sent on a quarterly basis. The process we have been using is to have one of the analysts preview both national and local articles slated for a particular newsletter, and then recommend to either skip or send that issue. Offensive local articles can be replaced with another of similar length at no cost to PM. If we opt to replace or modify a national article, it costs PM $3,000 per issue.

Since 1996, we have opted to either skip or send issues as follows:

Date of Newsletter	Decision	Comments
Spring 1996	Send	Deleted advertisement for CIGNA Time-Life videos featuring Ex-surgeon General Koop prior to sending.

01/12AN 31 '00 06:08PM STARTRIBUNE NEWSROOM 673 4[redacted] P.[redacted]

Date of Newsletter	Decision	Comments
Fall 1996	Send	A national article on heart attacks contained one minor reference to smoking which was deemed no worse than what the warning label states. Also, a breast cancer article launched into a free Time-Life video offer on subjects such as breast cancer, breast lumps, endometriosis, pregnancy, menopause, and osteoporosis. Decision to send based on Benefits Staff overwhelmingly positive reaction to Time-Life video series. No complaints received from member population based on newsletter content.
Winter 1996	Skip	National article entitled "A Breath of Fresh Air" listed smoking as one of the irritants in the environment which can trigger an asthma attack, and went further to say "Do not allow smoking in your home or in any environment that you can control".
Spring 1997	Send	Contained nothing objectionable.
Summer 1997	Skip	National article contained objectionable secondhand smoke reference.

In summary we have opted to skip three (3) of the last six (6) newsletters. To my knowledge, these are the only ones we have ever skipped.

The process of reviewing articles and making a recommendation to send or skip a issue varies, depending on content. Typically, it is immediately clear if something is objectionable. Other times, it may require discussion with others and management. I would say between phone calls with CIGNA, previewing of capsulated versions and full versions of local and national articles, discussion if necessary, auditing receipt of newsletters, and fielding complaints, I have spent as little as four (4) hours, to as long as several days on this activity. The issue surrounding the Time-Life video series required the most time, as you may remember from discussions in staff meetings.

Recommendation

2069916396

2078039377

OCR

Philip Moxx's. U.S.A. x.dr~am~c.
cvrrespoaa.aa
Benffrts Departmext Rieh>pwna, Yfe&ia
Ta: Dishlbutfon Data aday 90,1997.
From: Lisa Fislla
Sabj.csr CIGNA WeWedng Newsbttsr -
Yntsre StratsU
During our last CIGNA Aatfoa Plan
meadng, tlu iasuo of wLetSae to i0op
per'Irw+ng
artieles aod discontinue mndia6 CIGNA
Well-Being aawslener to om employees
was a
msiter of disanision . I Imvm done
somme reaearc>>, and wanted to
pruedt you with my
Sadings and pcdiminary
recwmmeadatioa for PM's atratezy
Ieprding I4aas aewelattee* .
I believe .vayone'a input is valusble,
and would epproolate hoarlng fmaa
aaeh of you on
whetlne you concur with my
reeommendatioa
...

Metadata

Title: *CIGNA WELL-BEING NEWSLETTER - FUTURE STRATEGY*

Organization Authors: *PMUSA, PHILIP MORRIS USA*

Person Authors: *HALLE, L*

Document Date: *19970530*

Document Type: *MEMO, MEMORANDUM*

Bates Number: *2078039376/9377*

Page Count: *2*

Collection: *Philip Morris*

TREC Legal Track: Topics

RequestNumber: **52**

RequestText: **Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.**

Proposal: **"high-phosphate fertilizer!" AND (boost! w/5 "crop yield") AND (commercial w/5 agricultur!)**

Rejoinder: **(phosphat! OR hpf OR phosphorus OR fertiliz!)
AND (yield! OR output OR produc! OR crop OR crops)**

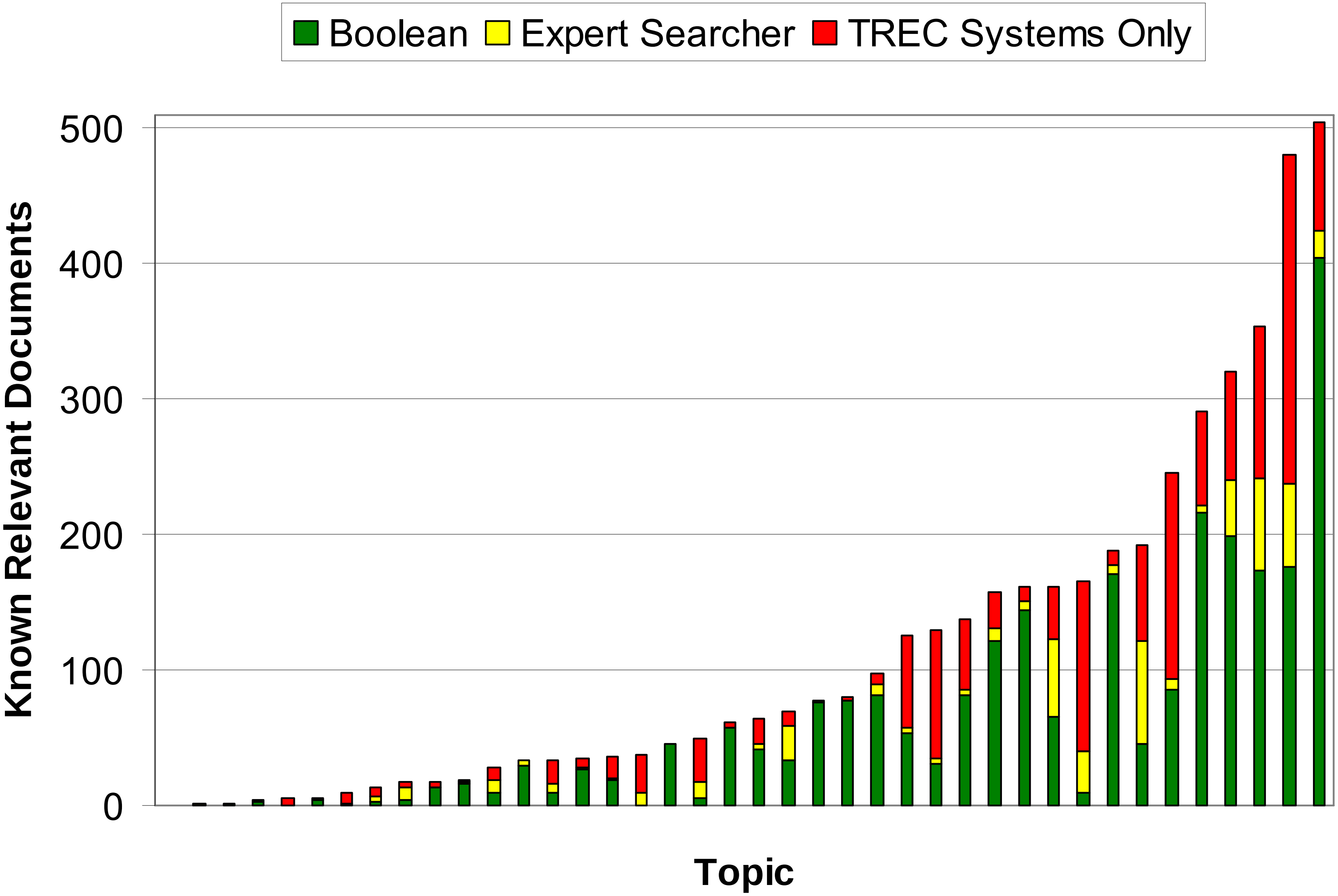
FinalQuery: **(("high-phosphat! fertiliz!" OR hpf) OR ((phosphat! OR phosphorus) w/15 (fertiliz! OR soil))) AND (boost! OR increas! OR rais! OR augment! OR affect! OR effect! OR multipl! OR doubl! OR tripl! OR high! OR greater) AND (yield! OR output OR produc! OR crop OR crops)**

B: **3078**

Beyond Boolean: getting at the "dark matter"

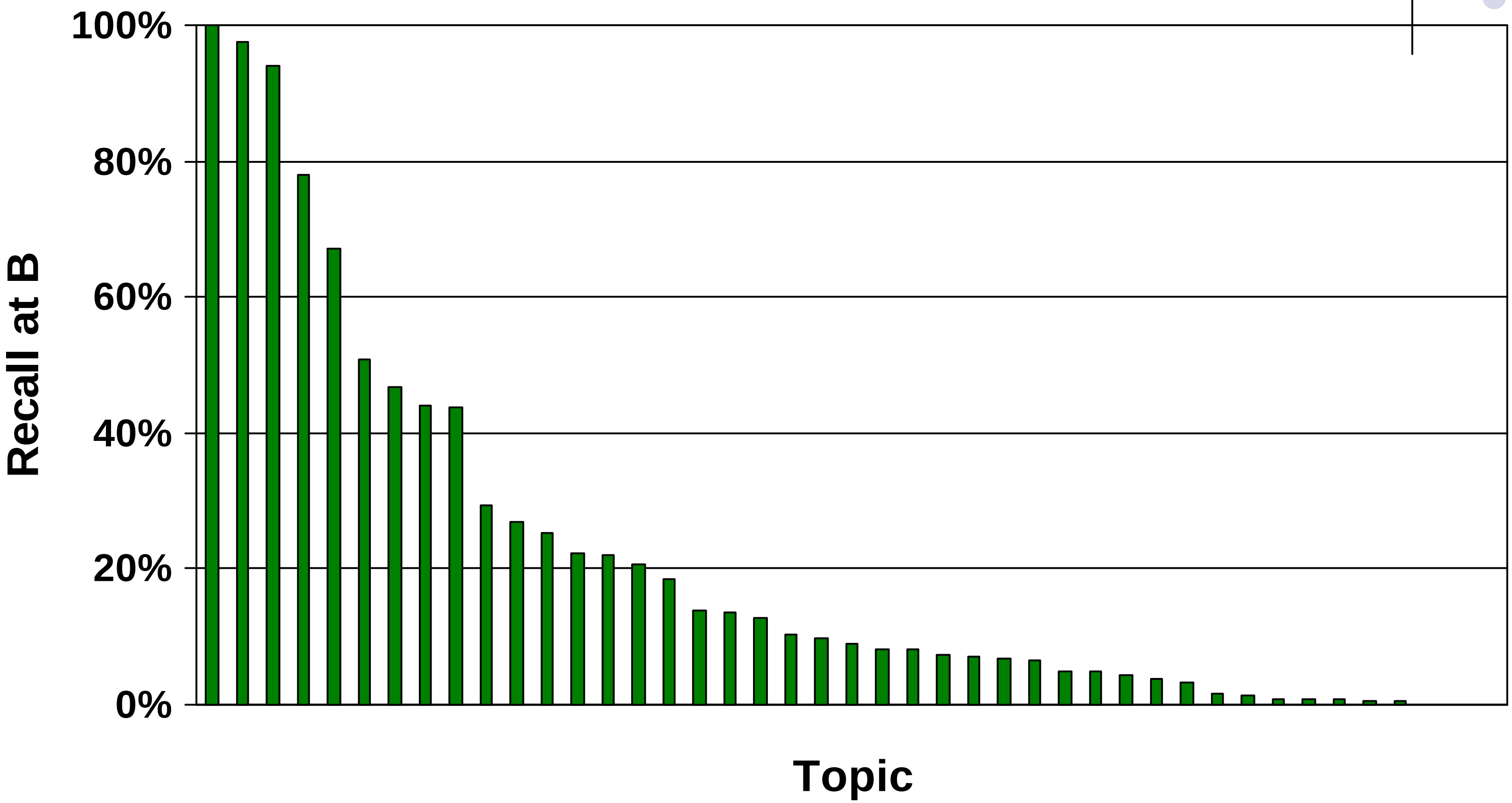
(i.e., relevant documents not found by keyword searches alone)

Nobody Finds Everything



Source: TREC 2006 Legal Track

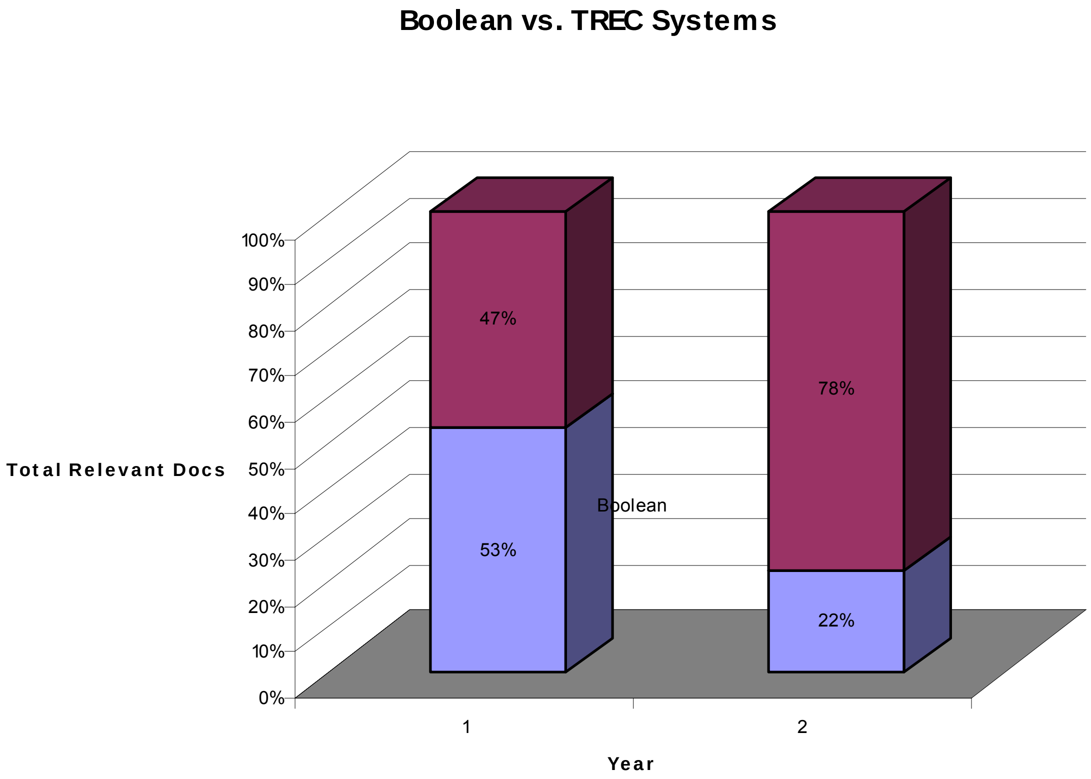
“Boolean” Searches May Miss A Large Percentage of Relevant Documents



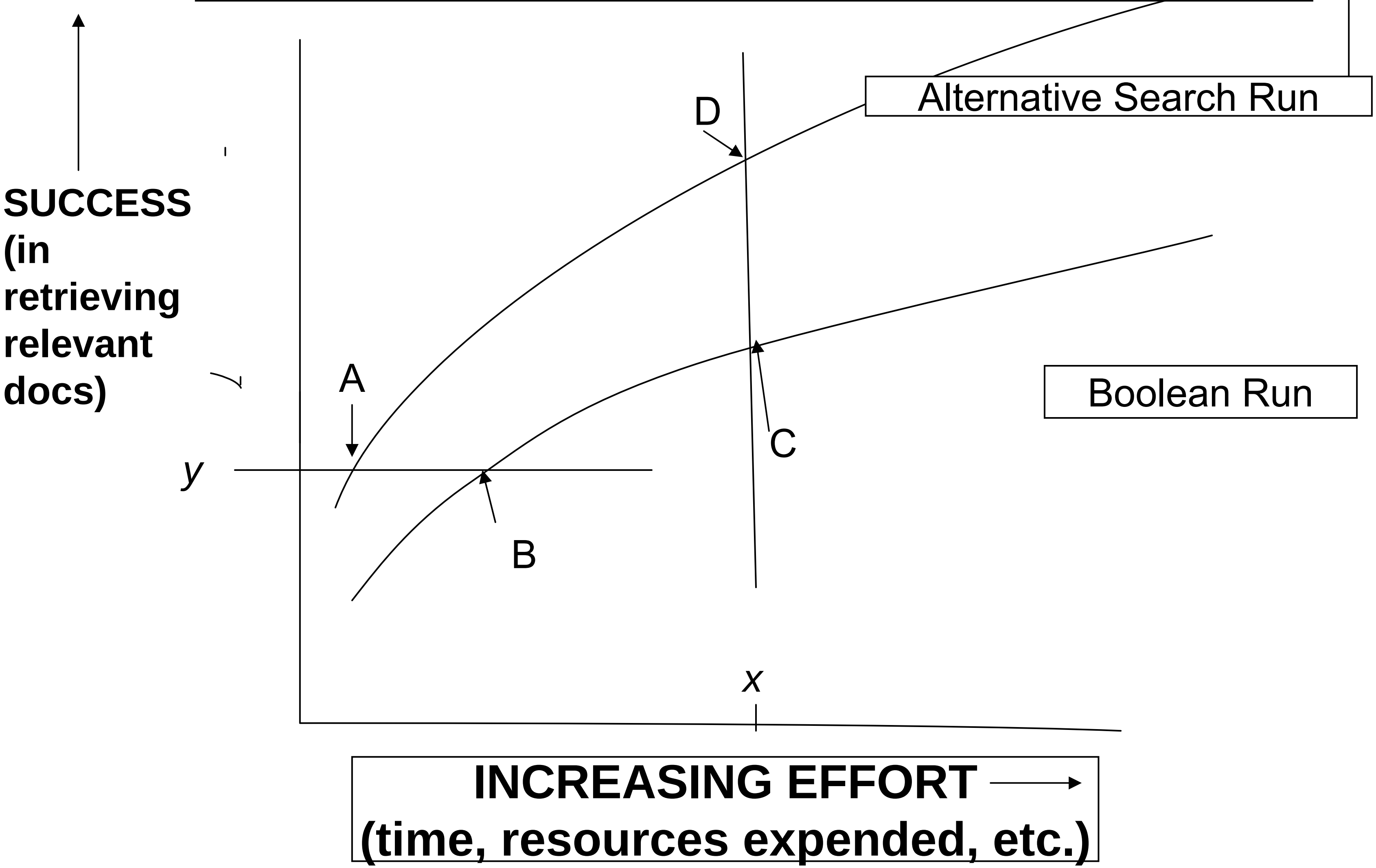
78% of relevant documents were only found by some other technique

Source: TREC 2007 Legal Track

Boolean v. TREC Systems: Results of Legal Track Years 1 and 2



Boolean vs. Hypothetical Alternative Search Method



Managing Litigation Risk

↑ Success per amount of effort

= ↓ Litigation Risk

Takeaway Messages

- (1) Success in using any automated method of technology will be enhanced by a well thought out process with substantial human input on the front end.
- (2) The choice of specific search & retrieval methods is necessarily dependent on the specific legal context in which it is to be employed.
- (3) The use of search & retrieval tools does not guarantee that all responsive documents will be identified in large data sets. Moreover, different search methods may produce differing results.
- (4) There are alternatives to keyword searching that should be explored, and what is expected as a matter of due diligence is awareness of what alternatives are available in the marketplace.
- (5) This is a field where new and evolving methods are bursting on the scene continuously.

Successful Searching For Purposes of E-Discovery Also Must Involve….

(1) Utilizing a ‘fusion’ of various search methods

(2) A more structured, iterative process, involving sampling & relevance feedback methods (to facilitate both human-in-the-loop and machine learning)

(3) Necessitating greater transparency and collaboration among lawyers otherwise engaged in an adversary proceeding.

Judge Grimm writing for the U.S. District Court for the District of Maryland

“[W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying on such searches for privilege review.” ***Victor Stanley, Inc. v. Creative Pipe, Inc.,*** --- F.Supp.2d ----, 2008 WL 2221841, * 3 (D. Md. May 29, 2008); *see id., text accompanying nn. 9 & 10* (citing to Sedona Search Commentary & TREC Legal Track research project)

Judge Facciola writing for the U.S. District Court for the District of Columbia

"Whether search terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?'*, 13 RICH. J.L. & TECH.. 10 (2007) * * * Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread."

-- ***U.S. v. O'Keefe*****,** 537 F.Supp.2d 14, 24 D.D.C. 2008).

Future Research

TREC 2008 Legal Track

http://trec-legal.umiacs.umd.edu/

(Including Open Letter to Legal Community)

DESI II Workshop June 25, 2008 London

http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/index.html

The Sedona Conference Search & Retrieval Commentary

Additional US Case Law on Search Protocols

- *Ameriwood Industries, Inc. v. Liberman,* 2007 WL 685623 (E.D. Mo.) (court orders expert report with number of "hits" based on negotiated search terms, with expectation that parties will continue to meet and confer to refine search based on false positives)
- *Disability Rights Council of Greater Washington, et al. v. Washington Metropolitan Transit Authority,* 242 F.R.D. 139 (D.D.C. 2007) (Facciola, J.) (proposes use of concept searching as possible supplement to keyword searches)
- *Qualcomm Inc. v. Broadcom Corp.*, 2007 WL 2296441, at *33 (S.D. Cal.) (sanctions opinion involving underlying failure to disclose 200,000 emails prior to trial, where court found "incredible that Qualcomm never conducted such an obvious search" using certain keywords).
- *Williams v. Taser Intern, Inc.*, 2007 WL 1630875 (N.D. Ga.) (court adjudicates search protocol with keywords plus use of simple Boolean operators)
- *Treppel v. Biovail,* 233 F.R.D. 363 (S.D.N.Y. 2006) (court urged parties to consider negotiating keywords under rubric of coming up with a search protocol)

Additional References

- **Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year,* 8 Sedona Conference Journal 251 (2007) (available on WESTLAW and LEXIS)**
- **Jason R. Baron, Douglas W. Oard, David D. Lewis, *TREC-2006 Legal Track Overview,* http://trec.nist.gov/pubs/trec15/t15_proceedings.html (item 4)**
- **Mia Mazza, Emmalena K. Quesada, & Ashley L. Stenberg, *In Pursuit of FRCP 1: Creative Approaches to Cutting and Shifting Costs of Discovery of Electronically Stored Information,*, 13 RICH. J.L. & TECH. 11 (2007), http://law.richmond.edu/jolt/v13i3/article11.pdf. (concept searching)**

Additional References (con't)

- **George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007), http://law.richmond.edu/jolt/v13i3/article10.pdf. (concept searching)**
- ***Sedona Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* (August 2007 public draft), http://www.thesedonaconference.org/ content/miscFiles/publications_html**
- **Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, Paul Thompson, *Overview of TREC 2007 Legal Track,* at http://trec-legal.umiacs.umd.edu/**

Additional References (con't)

- ICAIL 2007 (International Conference on Artificial Intelligence and the Law), Workshop on Supporting Search and Sensemaking for ESI in Discovery Proceedings, see http://www.umiacs.umd.edu/~oard/desi-ws/
  - see also J. Baron and P. Thompson, "The Search Problem Posed By Large Heterogeneous Data Sets in Litigation: Possible Future Approaches to Research," ICAIL 2007 Conference Paper, June 4-8, 2007, available at http://www.umiacs.umd.edu/~oard/desi-ws/ (click link to conference paper).

+ TREC LEGAL TRACK HOMEPAGE:

http://trec-legal.umiacs.umd.edu/

Jason R. Baron

Director of Litigation

Office of General Counsel

National Archives and Records Administration

8601 Adelphi Road # 3110
College Park, MD 20740
(301) 837-1499
Email: jason.baron@nara.gov