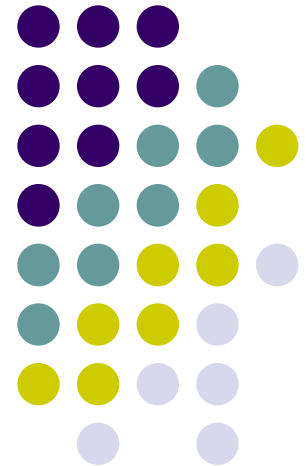


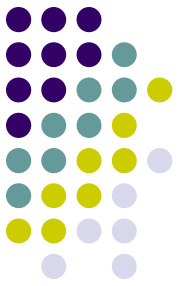
MER 2009

**Search & Information Retrieval Law 101: Hot
Topics and 7 Keys To Success**

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

May 19, 2009





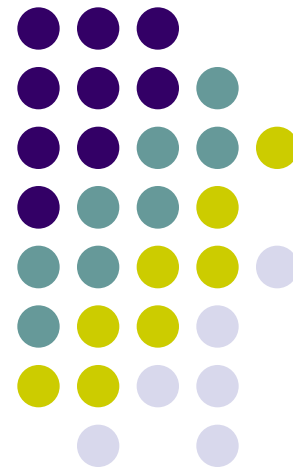
Overview

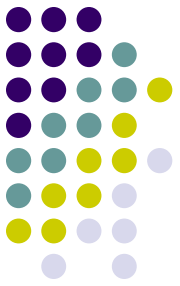
- **Introduction: Case Study: *U.S. v. Philip Morris***
- **Myth, Hype, Reality – Information Retrieval and the Problem of Language**
- **The TREC Legal Track**
- **Strategic Challenges & Seven Keys to Success**
- **Recent Case Law**
- **References**

Definition of “ESI”

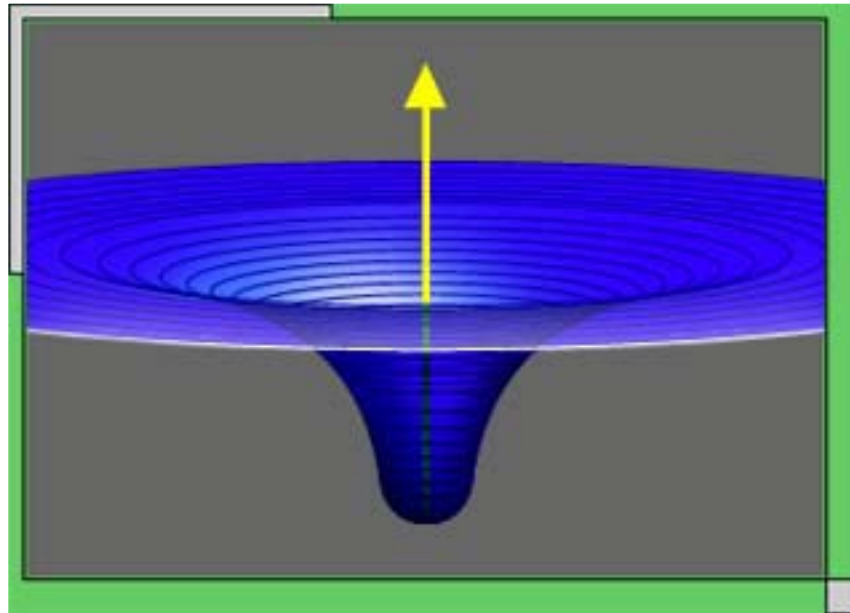
-A new legal term of art: “electronically stored information” to supplement the older term “documents”:

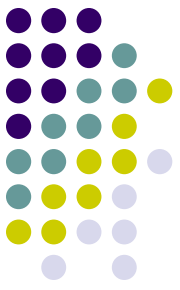
-The wide variety of computer systems currently in use, and the rapidity of technological change, counsel against a limiting or precise definition of ESI...A common example [is] email ... The rule ... [is intended] to encompass future developments in computer technology. --Advisory Committee Notes to Rule 34(a), 2006 Amendments to the Federal Rules of Civil Procedure





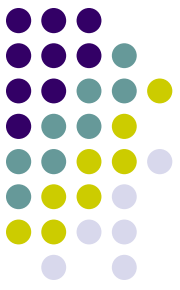
Information Inflation: The Expanding ESI Universe





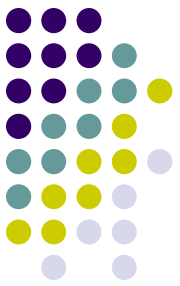
Case Study: U.S. v. Philip Morris – Overall Discovery

- 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



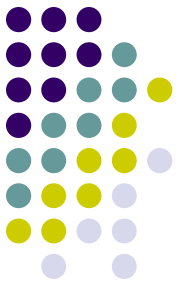
Case Study: U.S. v. Philip Morris (con't) – Employing a limited feedback loop

- **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.**



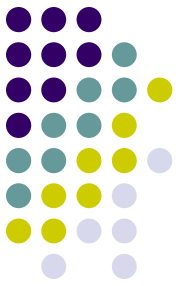
Example of Boolean search string from *U.S. v. Philip Morris*

- **((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 PROBLEM: the 1% figure does not scale



A Hypothetical

- 1 billion emails, 25% with attachments
- Reviewed at 50 hour
- Would take 100 people, 10 hrs per day, 7 days a week, 52 weeks a year

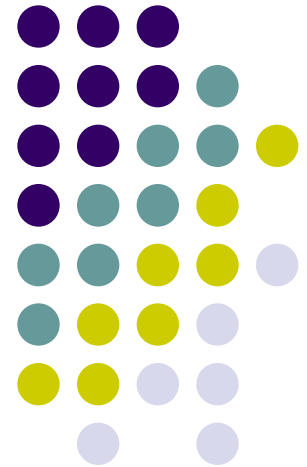
54 YEARS TO COMPLETE

- At \$100/hr, \$ 2 billion in cost
- Even 1% (10 million docs) ... 28 weeks and \$20 million in cost

The Myth of Search & Retrieval

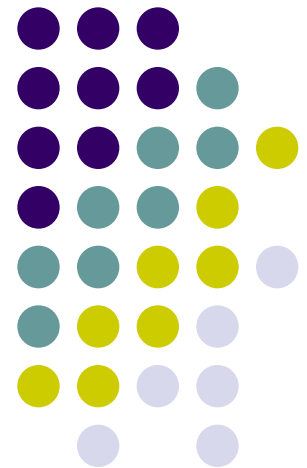
When lawyers request production of “all” relevant documents (and now ESI), all or substantially all will in fact be retrieved by existing manual or automated methods of search.

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

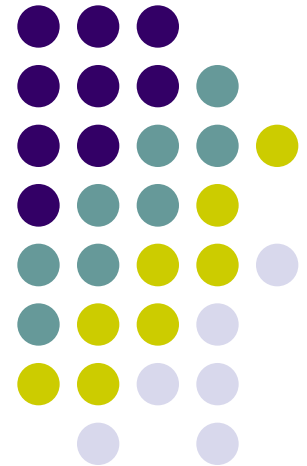
Claims in the legal tech sector that a very high rate of “recall” *(i.e., finding all relevant documents) is easily obtainable provided one uses a particular software product or service.



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:

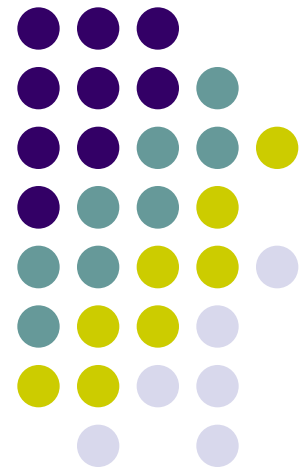
--in a 40,000 document case (350,000 pages), lawyers estimated that a manual search would find 75% of relevant documents, when in fact the research showed only 20% or so had been found.



More Reality: IR is Hard

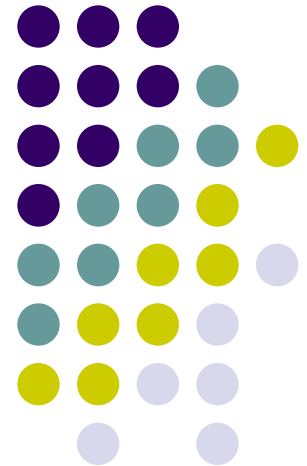
+ Information retrieval (IR) is a hard problem: difficult even with English-language text, and even harder with non-textual forms of ESI (audio, video, etc.) caught up in litigation.

+ 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

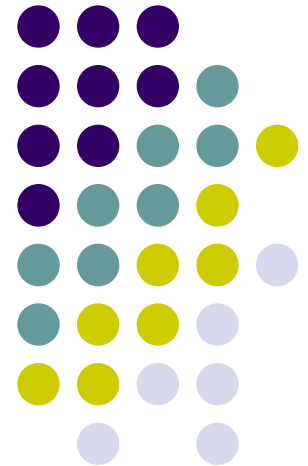


Problems of language

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

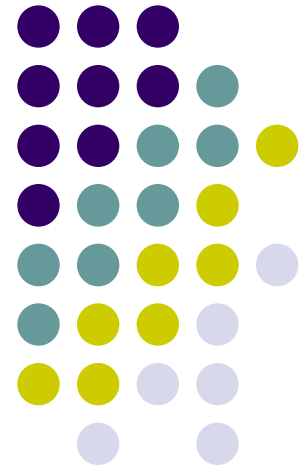
Synonymy: variation in describing same person or thing in multiplicity of ways (e.g., “diplomat,” “consul,” “official,” ambassador,” etc.)

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



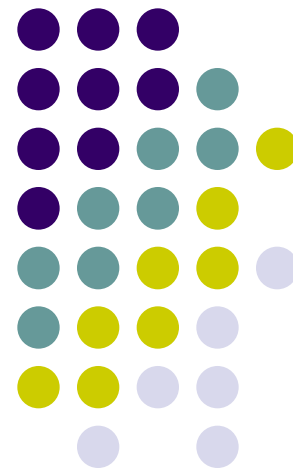
Why is IR hard (for lawyers)?

- + 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

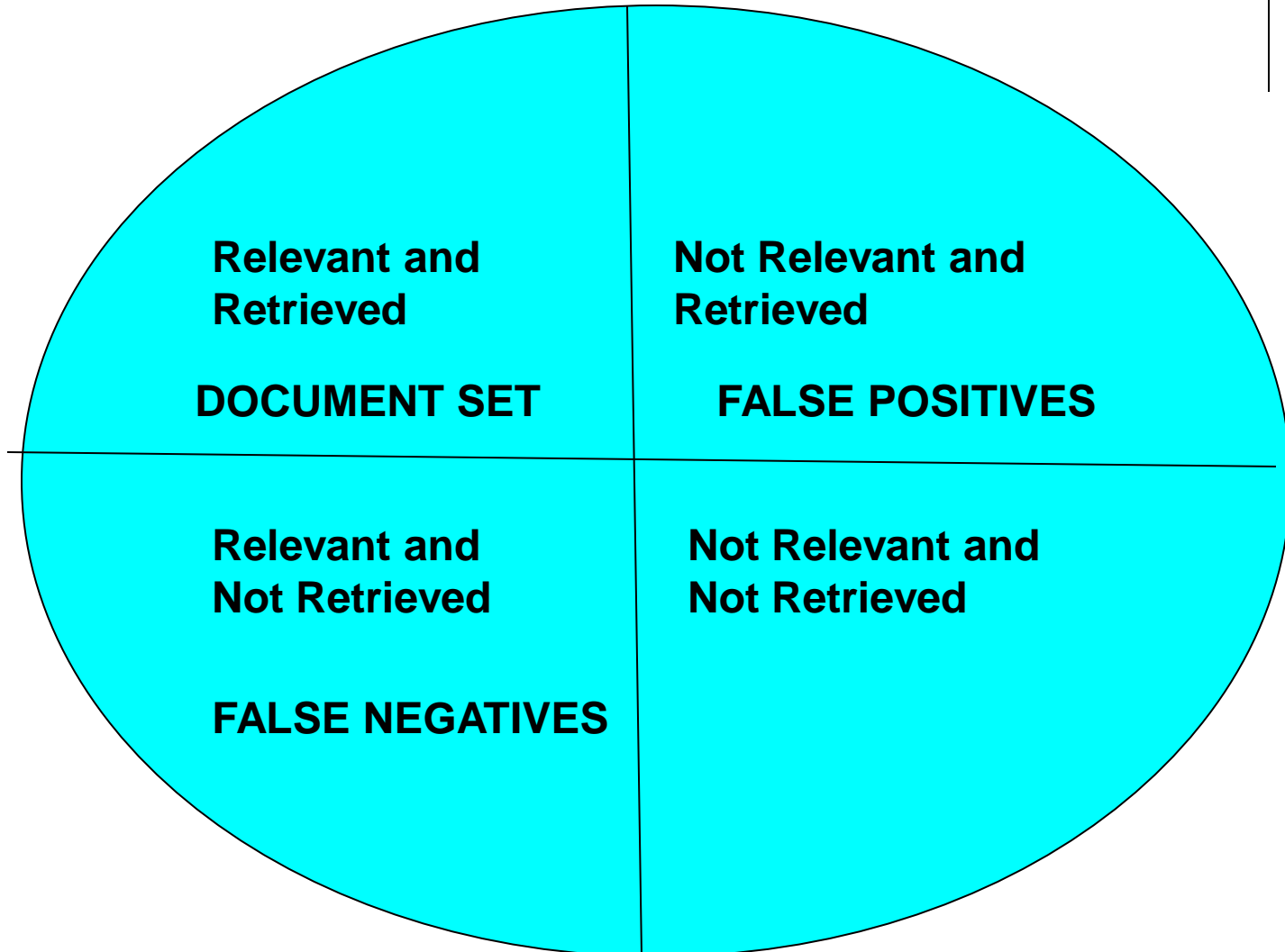
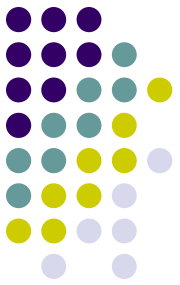


Snapshot of 2008 ESI Heterogeneity

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.



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

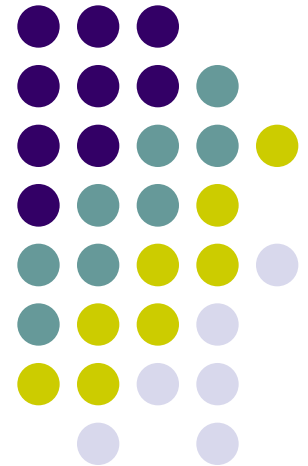


Measures of Information Retrieval

Recall =

of responsive docs retrieved

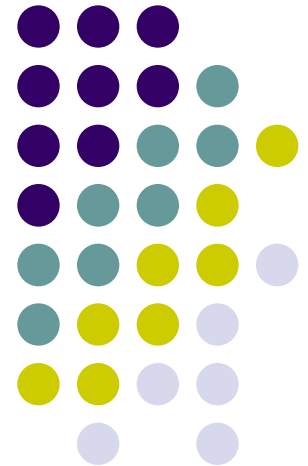
of responsive docs in collection

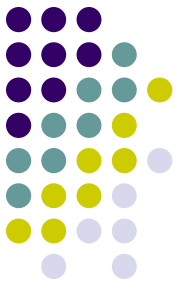


Measures of Information Retrieval

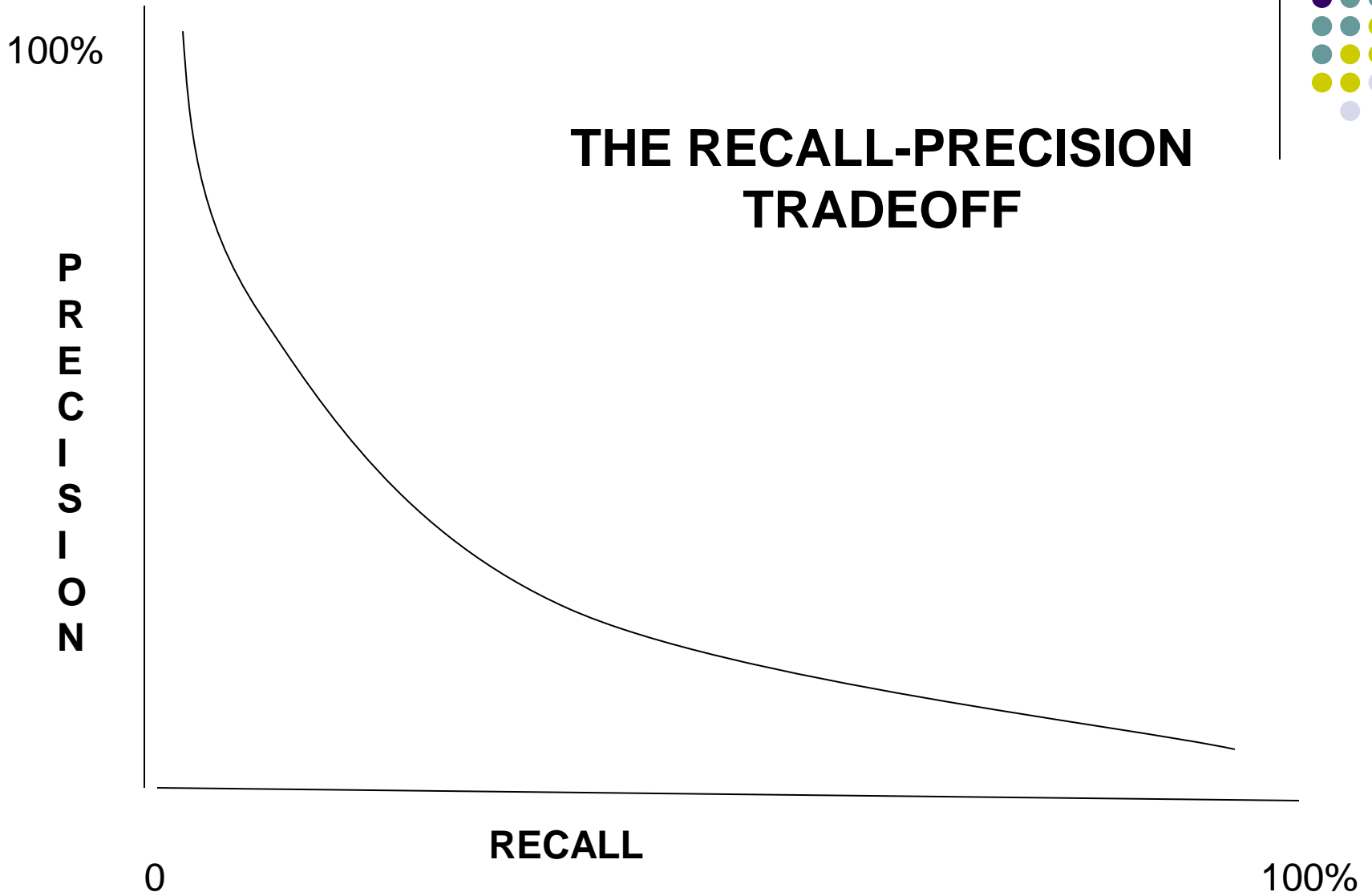
Precision =

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



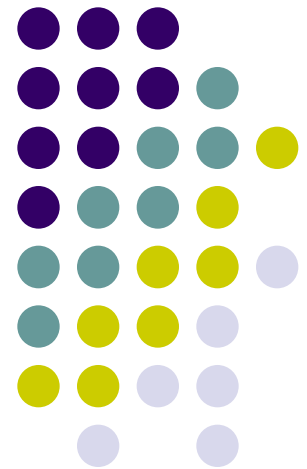


THE RECALL-PRECISION TRADEOFF

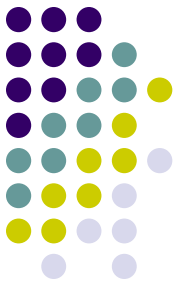


Three Questions

- (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?**



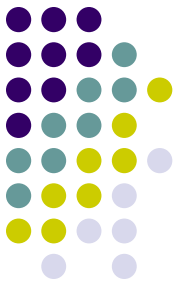
Beyond Keywords: Alternative Search Methods



- *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*

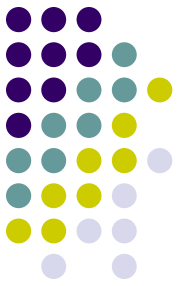
Reference: *Appendix to The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (Aug. 2007 Public Comment Draft), available at <http://www.thesedonaconference.org> (link to publications)*

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 Legal Track



- **The TREC Legal Track was designed to evaluate the effectiveness of search technologies in a real-world legal context**
- **First of a kind study using nonproprietary data since Blair/Maron research in 1985**
- **Hypothetical complaints and 100+ “requests to produce” drafted by Sedona Conference members**
- **“Boolean negotiations” conducted as a baseline for search efforts**
- **New Interactive Task added in 2008 using Topic Authorities and a post-adjudication round**
- **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, plus in 2008 from legal service providers**

TREC Legal Track: Documents



Scanned

OCR

Metadata

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 newsletters to our employees was a matter of discussion. I have later 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 \$5,000 per issue.

Since 1994, we have opted to either skip or send issues as follows:

Date of Newsletter	Decision	Comments
Spring 1996	Skip	Deleted advertisement for CIGNA Time-Life videos featuring Dr.urgem General. Keep prior to sending.

Date of Newsletter	Decision	Comments
Fall 1996	Skip	A national article on heart attacks' contained one minor reference to smoking which was deemed so vague that what the warning label meant. Also, a breast cancer article included into a Free Time-Life video offer on adjuvant such as breast cancer, breast lumps, endometriosis, pregnancy, menopause, and osteoporosis. Decision to send based on benefits their 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 causes in the environment which can trigger an asthma attack, and went further to say "Do not allow smoking in your house or in any environment that you can control".
Spring 1997	Skip	Combined 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 expensed systems and full reviews of local and national articles, discussion if necessary, editing notes of newsletters, and holding conclusions, 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 past meetings.

Recommendation

Philip Moxx's. U.S.A. x.dr~am~c.
cvrrespoaa.aa
Benffrts Department Rieh>pwna, Yfe&ia
Ta: Dishlbutfon Data aday 90,1997.
From: Lisa Fislla
Sbj.csr CIGNA WeWedng Newsbttsr -
Yntsre StratsU
During our last CIGNA Aatfoa Plan
meadng, tlu iasuo of wLetSae to iOop
per'Irw+ng
artieles aod discontinue mndia6 CIGNA
Well-Being aawslener to om employees
was a
msiter of disanision . I Imvm done
somme reearc>>, and wanted to
pruedt you with my
Sadings and pcdiminary
recwmmeadatioa for PM's atratezy
leprding l4aas aewelattee* .
I believe .vayone'a input is valusble, and
would epproolate hoarlng fmaa aeah of
you on
whetlne you concur with my
reecomendatioa

Title: CIGNA WELL-BEING
NEWSLETTER - FUTURE
STRATEGY

Organization Authors:
PMUSA, PHILIP MORRIS
USA

Person Authors: HALLE, L

Document Date: 19970530

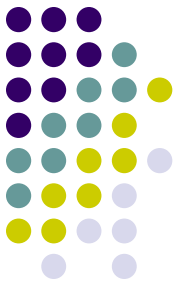
Document Type: MEMO,
MEMORANDUM

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!)**

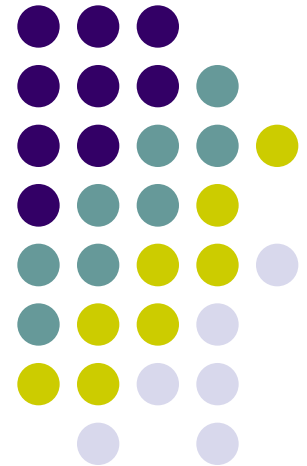
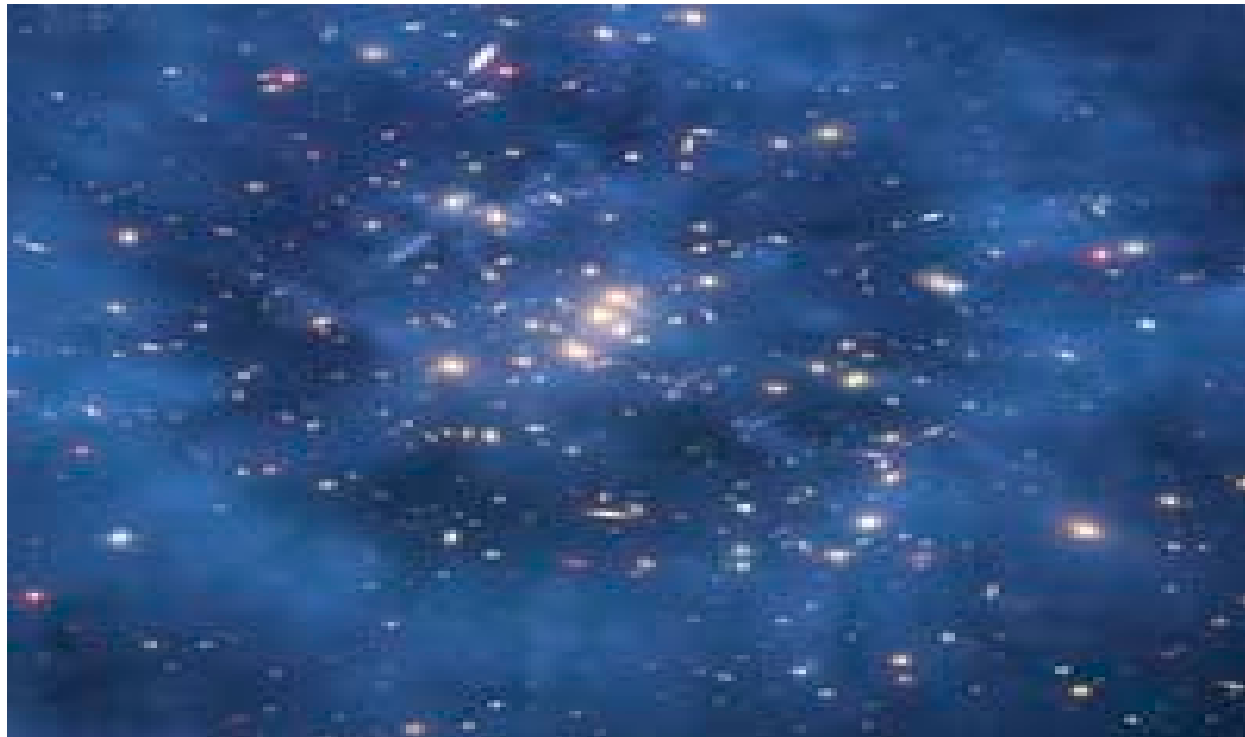
Rejoinder: **(phosphat! OR hpf OR phosphorus OR fertiliz!) AND (yield! OR output OR produc! OR crop OR crops)**

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))**

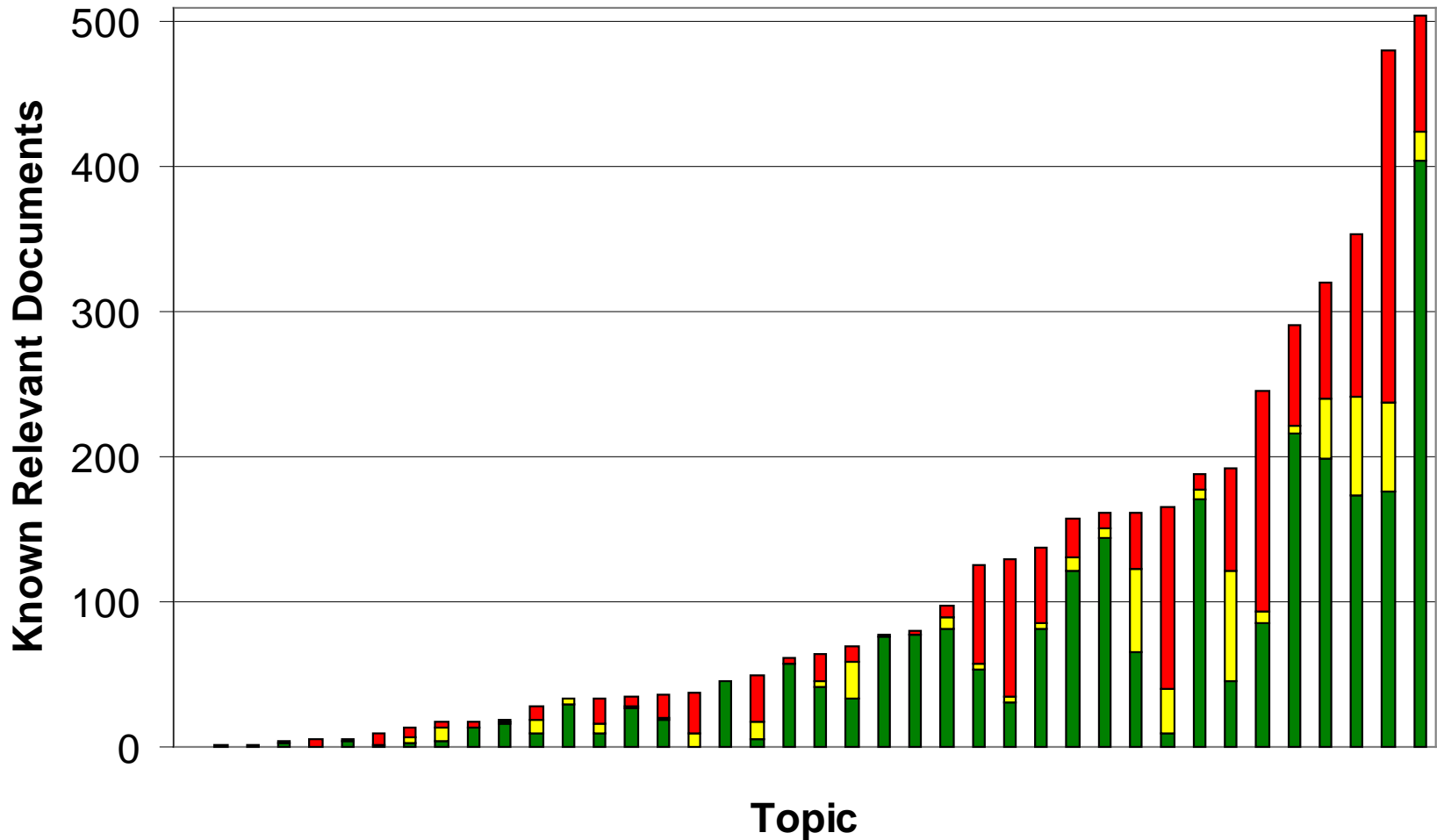
B: **3078**

Beyond Boolean: getting at the “dark matter”

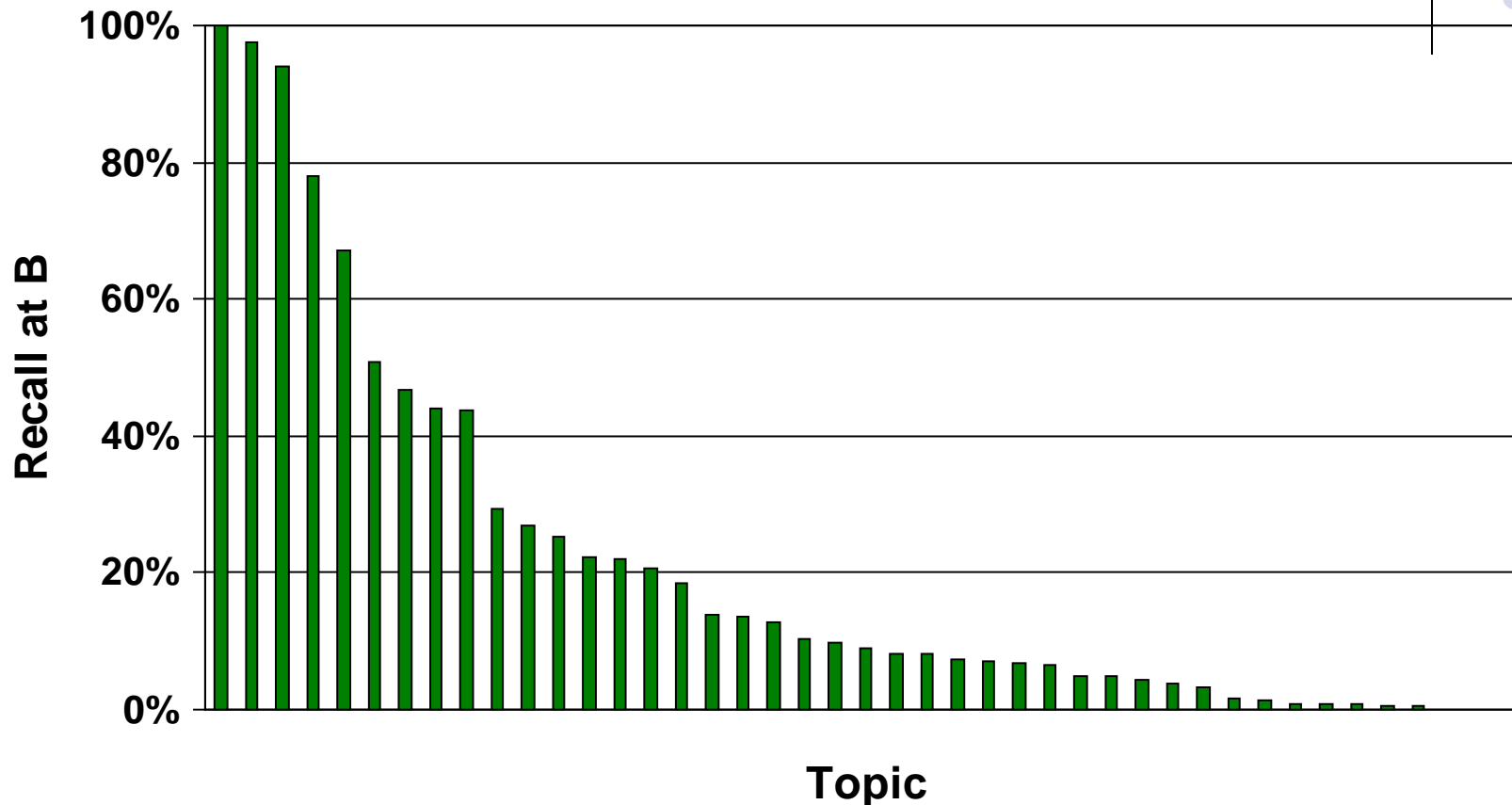
*(i.e., relevant documents not found by keyword searches
alone)*



Nobody Finds Everything

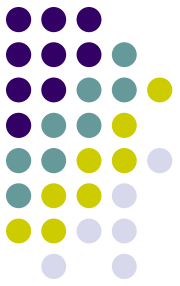


“Boolean” Searches May Miss A Large Percentage of Relevant Documents

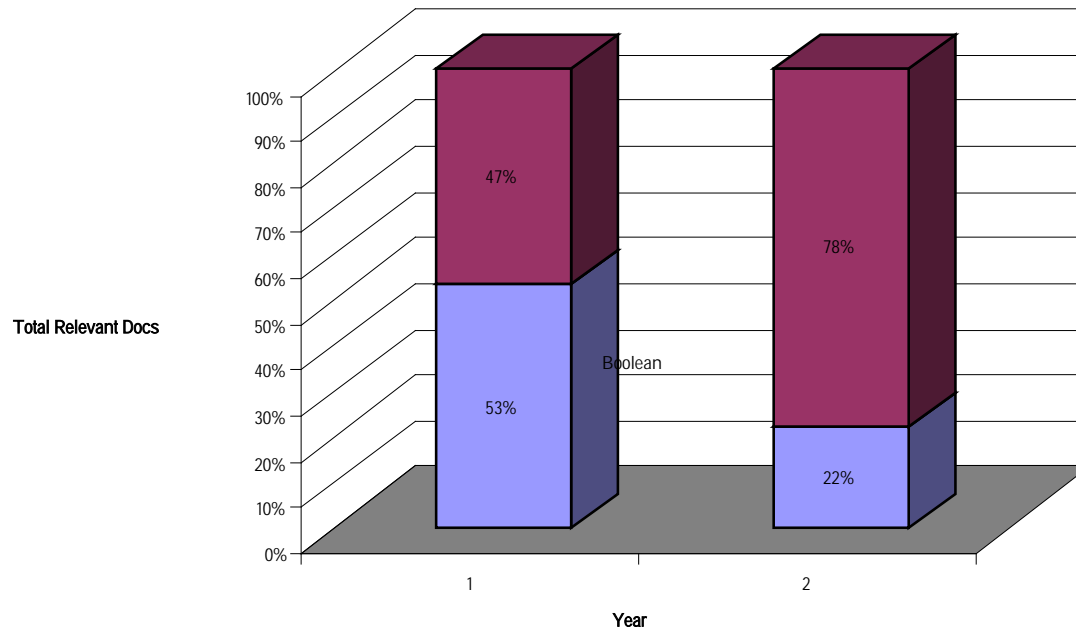


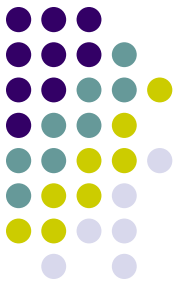
78% of relevant documents were only found by some other technique

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

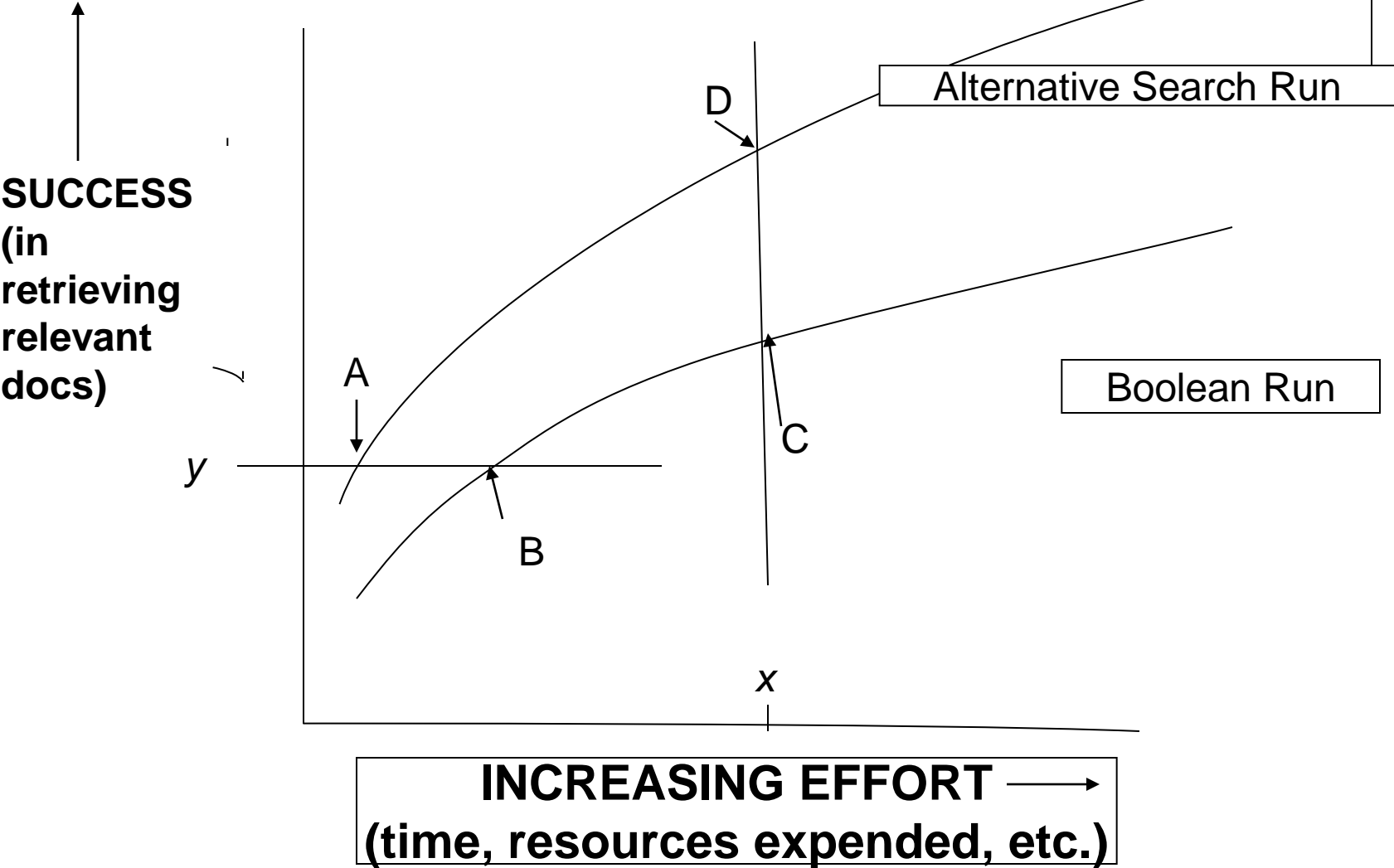


Boolean vs. TREC Systems

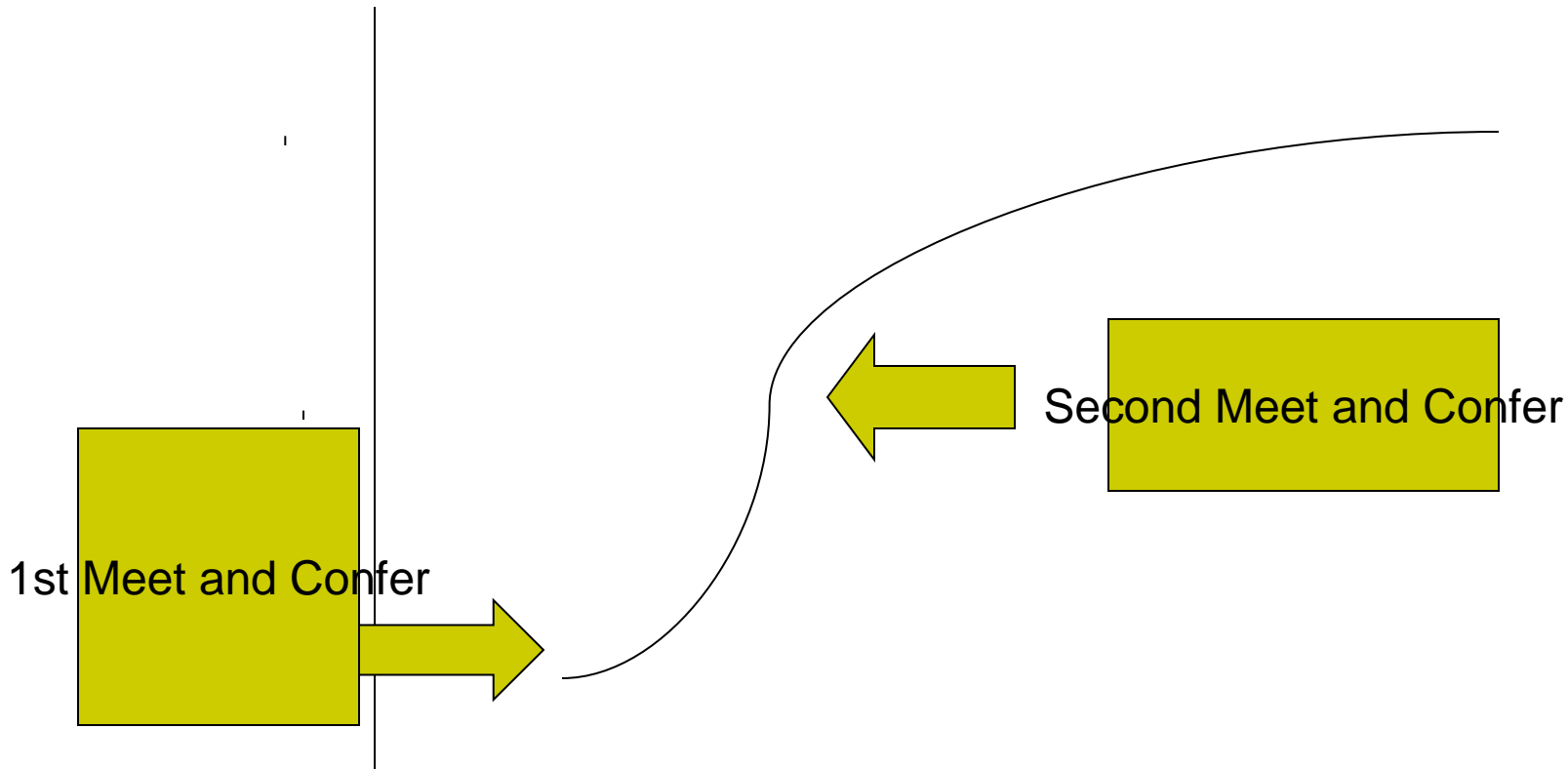
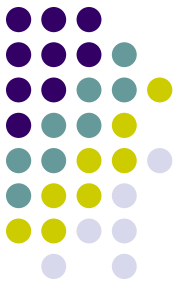




Boolean vs. Hypothetical Alternative Search Method



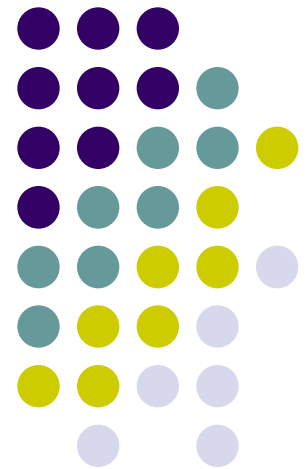
Improving Search Effectiveness Through Relevance Feedback and Multiple Meet and Confers



Source: F.C. Zhao, D. W. Oard, and J.R. Baron, "Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback" (forthcoming 2009)

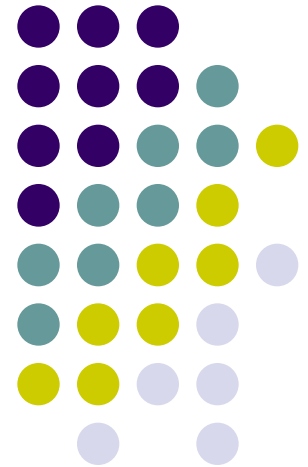
Strategic Challenges

Convincing lawyers and judges that automated searches are not just desirable but necessary in response to large e-discovery demands.



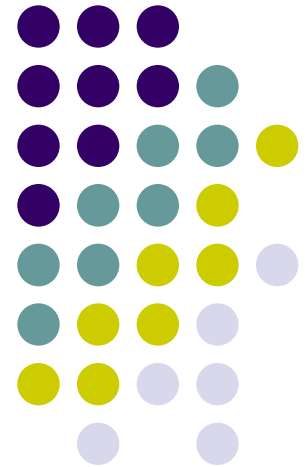
Challenges (con't)

Having all parties and adjudicators understand that the use of automated methods does not guarantee all responsive documents will be identified in a large data collection.



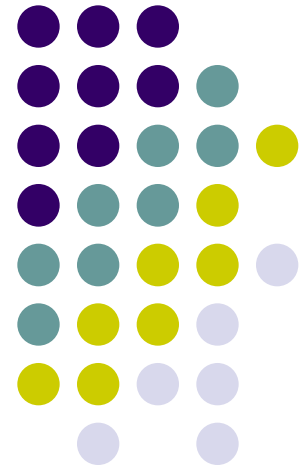
Challenges (con't)

Designing an overall review process which maximizes the potential to find responsive documents in a large data collection (no matter which search tool is used), and using sampling and other analytic techniques to test hypotheses early on.

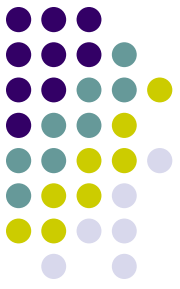


Challenges (con't)

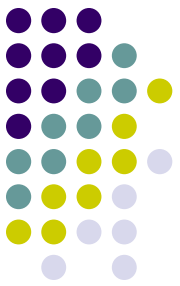
Being open to using new and evolving search and information retrieval methods and tools.



Seven Keys To Success in Search & Retrieval



- **(1) Project Management:** 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 (not the lone attorney thinking up keywords at her desk)
- **(2) Know Your Case:** the choice of specific search & retrieval methods is necessarily dependent on the specific legal context in which it is to be employed
- **(3) Practice Fuzzy Thinking:** think outside the box by considering misspellings and variations, using the power of Boolean
- **(4) Think About Alternatives:** 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.



Seven Keys To Success (continued)

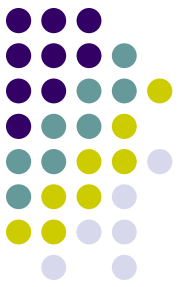
- (5) Structure the Process, with Sampling:** a more structured, iterative process, involving sampling & relevance feedback methods will improve results (to facilitate both human-in-the-loop and machine learning)
- (6) Practice Transparency:** greater transparency and cooperation among lawyers otherwise engaged in an adversary proceeding will facilitate the goals of Fed R. Civ. P. 1
- (7) Defensibility Starts With Documentation:** be able to tell the story of your search protocol by documenting your efforts at every stage of the process.

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.***, 250 F.R.D. 251 (D. Md. 2008); *see id.*, *text accompanying nn. 9 & 10* (citing to Sedona Search Commentary & TREC Legal Track research project)

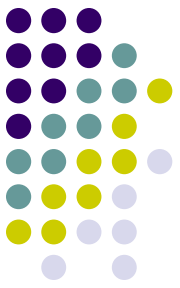
Judge Facciola writing for the U.S. District Court for the District of Columbia



“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.”

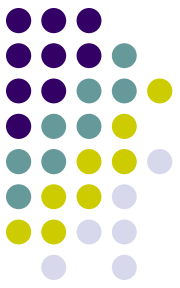
-- ***U.S. v. O'Keefe***, 537 F.Supp.2d 14, 24 D.D.C. 2008).

Judge Scheindlin writing for the U.S. District Court for the Southern District of New York

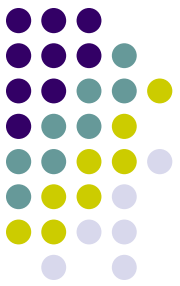


SEC v. Collins & Aikman Corp., 2009 WL 94311 (S.D.N.Y. Jan. 23, 2009) (court rejects SEC position that defendant has adequate opportunity to search through 10 million pages to find substantially the same documents as identified by the SEC,” where court notes significant expense and delay of doing so as well as the fact that “the inaccuracy of [keyword] searches is by now relatively well known”; court further rejects what it terms “SEC’s blanket refusal to negotiate a workable search protocol responsive to defendants’ requests,” citing The Sedona Conference proclamation)

Judge Peck writing for the U.S. District Court for the Southern District of New York



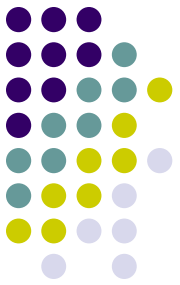
William A. Gross Construction Associates Inc. v. American Manufacturers Mutual Ins. Co., 2009 WL 724954 (S.D.N.Y. March 19, 2009) (in multi-million dollar dispute, where issue involved production of 3rd party emails where none of the three parties could agree on keyword search terms, court fashioned a compromise; court stated at the outset that “this Opinion should serve as a wake-up call to the Bar ... about the need for careful thought, quality control, testing and cooperation with opposing counsel in designing search terms or ‘keywords’ to be used.”)



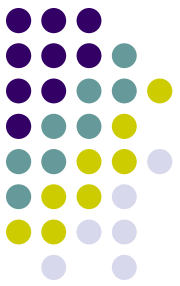
Additional case law on “search”

In re Fannie Mae Litigation, 552 F.3d 814 (D.C. Cir. 2009) (appellate court upholds contempt citation against government agency, where it had failed to meet court-imposed deadlines after agreeing to produce nonprivileged documents found responsive to 400 keyword search terms, where the production set consisted of 660,000 recovered documents that needed to be, but could not be, reviewed in time)

Additional case law on “search”



Spieker v. Quest Cherokee, 2008 WL 4758604 (D. Kan. Oct. 30, 2008) (where keyword search provisionally agreed to by parties produced 32 gigabytes of data, comprising 31,000 documents and 1,400,000 pages, court rejected defendants’ request for more specificity from plaintiffs on matter of keywords, holding that “defendants cannot escape its burden of production by now arguing that plaintiffs’ suggested search terms are ‘not specific enough,’” and that defendants should modify the terms accordingly)



Additional case law on “search” (con’t)

- + **Clearone Communications Inc. v. Chiang**, 2008 WL 920336 (D. Utah April 1, 2008) (court adjudicates dispute over Boolean conjunctive vs. disjunctive operators (“and” vs. “or”) between search terms)
- + **Qualcomm v. Broadcom Corp.**, 539 F.Supp. 2d 1214 (S.D. Cal. 2007) (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)

Future Research

TREC 2009 Legal Track

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

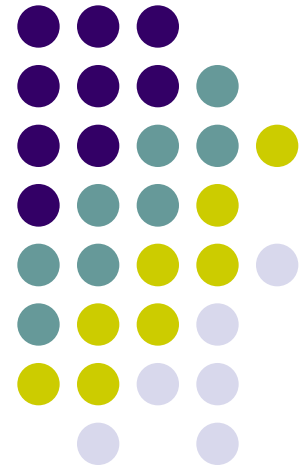
(Including Open Letter to Legal Community)

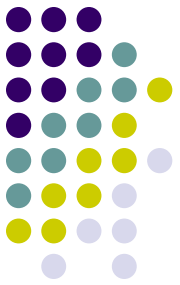
ICAIL 2009 Barcelona

DESI III Workshop -- June 8, 2009

http://www.law.pitt.edu/DESI3_Workshop/

The Sedona Conference Search & Retrieval Commentary





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

