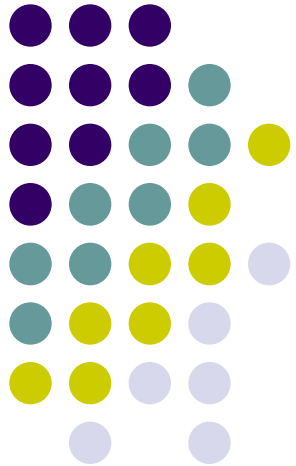
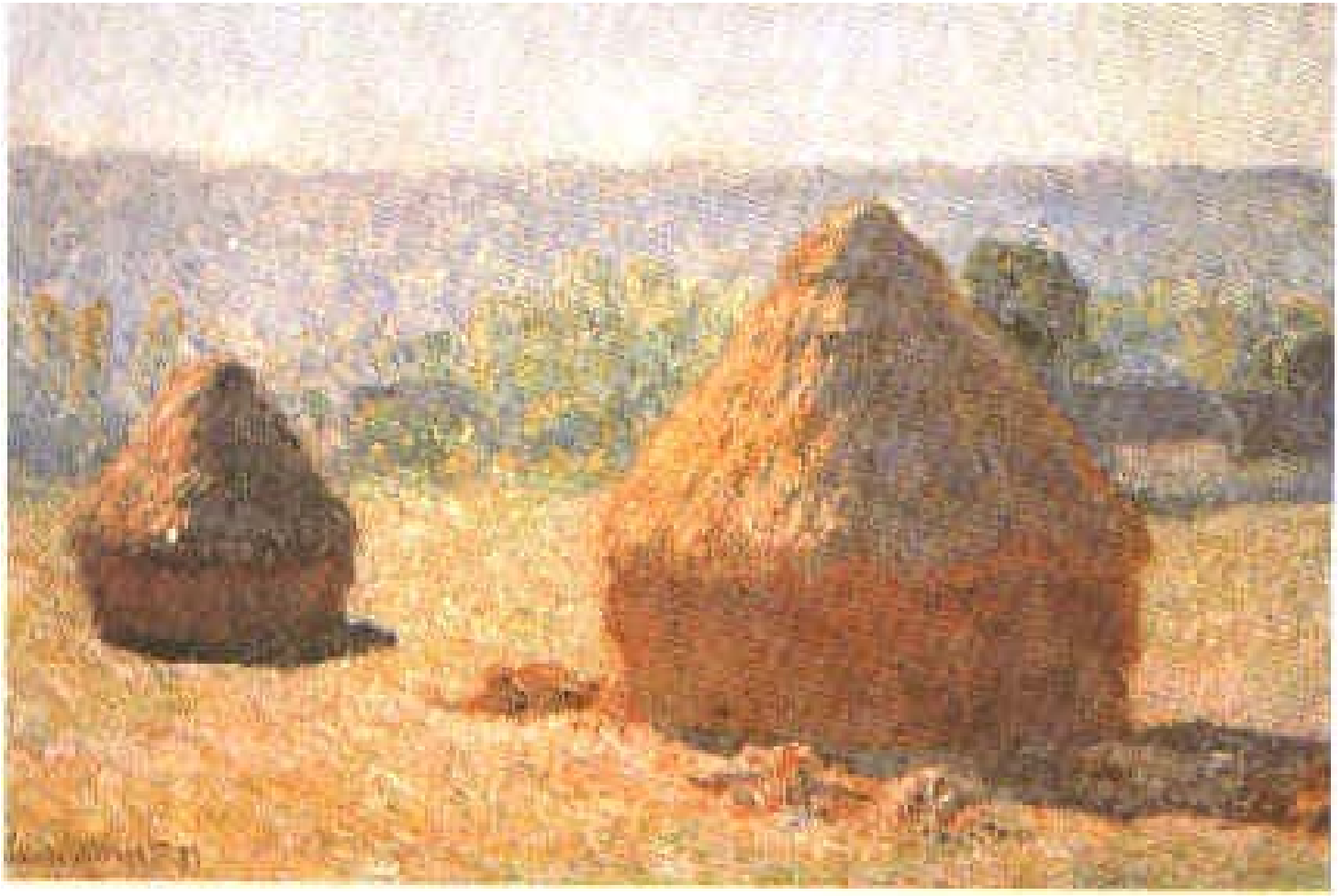
Needles, Haystacks and Smoking Guns – Searching for Legal Evidence in the Modern Email Archive

International Symposium:
Our Professional Identities in a World Gone Digital
The University of British Columbia
Vancouver, BC, Canada
February 13, 2009

Jason R. Baron
Director of Litigation
U.S. National Archives and Records Administration
College Park, Maryland



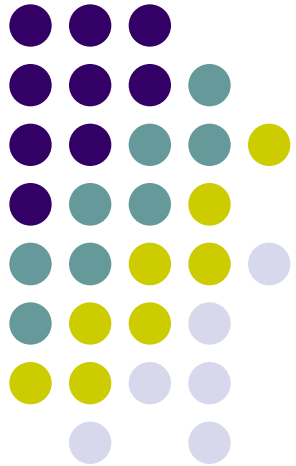
Searching Through E-Haystacks: A Big Challenge



A New Legal Term of Art Under the U.S. Federal Rules of Civil Procedure: *Electronically Stored Information* or "ESI"

"Electronically stored information":

- *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*



Common Forms of ESI

Email with attachments (all kinds)

Text files, powerpoint, spreadsheets

Voice mail, instant and text messaging

Databases, proprietary applications

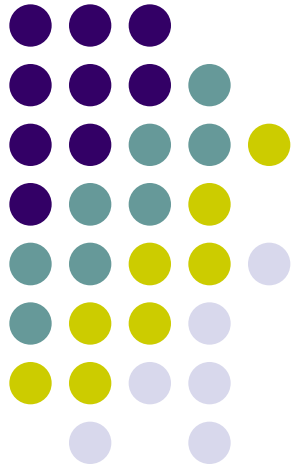
Internet, intranet, wikis, blogs, RSS feeds

(plus cache files, slack space data, cookies)

Data on PDAs, cellphones

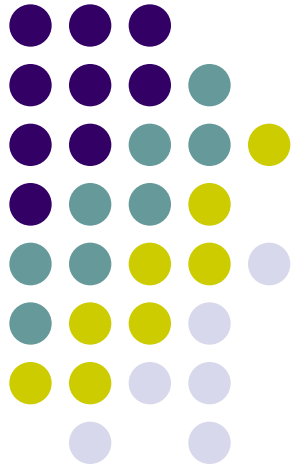
Videoconferencing & webcasting

Metadata

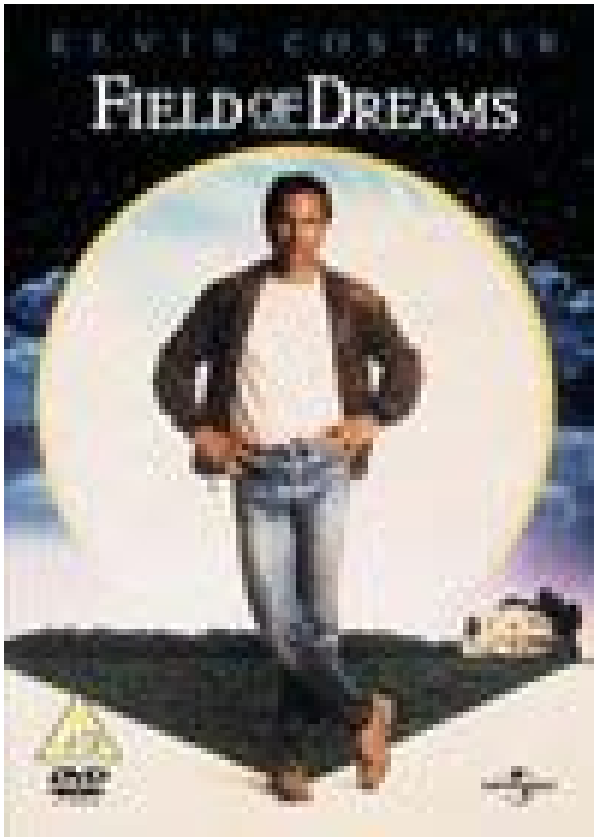


Common Sources of ESI

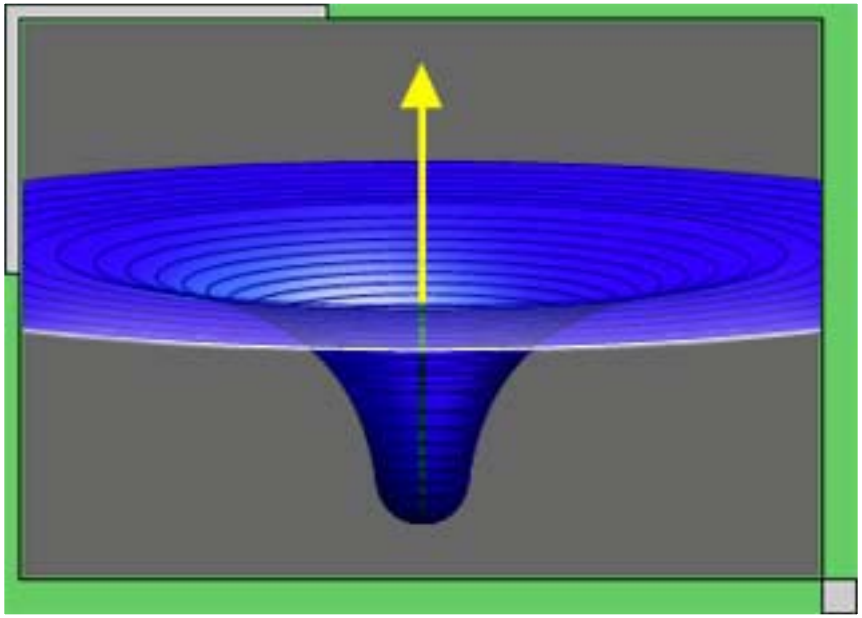
- Mainframes, network servers, local drives (including network activity logs)
- DVDs, CD ROMs, floppy disks
- Laptops
- Backup tapes
- External hard drives (e.g., flash, Zip, Jazz, ipods)
- Third party storage



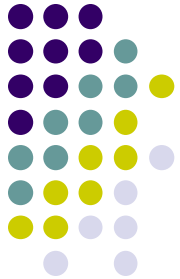
If you build it, the lawyers will come… (at least in the U.S.)



Information Inflation: The Expanding ESI Universe . . . .



Example: White House Email+



- Email system implemented in mid-1980s
- First lawsuit: 6,000 backups preserved
- Mid-1990s: introduction of a “total” electronic archiving scheme at White House
- Late 1990s: glitches
- 2001: 32 million emails transferred to NARA
- 2009: 200 million+ emails “ “ “
- The near-term future: a billion presidential emails at the end of the next eight years

The “Good” Archivist

- In the U.S., records are retained as part of “file plans” and “record retention schedules,” catalogued as discrete “record series” in accordance with retention and disposition instructions
- High degree of granularity present in records schedules as to specific retention periods for temporary records
- Appraisal by archivists is primarily a matter of justifying the segregation of “wheat” and “chaff,” i.e., the permanent from the temporary
- Email & ESI appraised under traditional methods: expected to be segregated by record series.

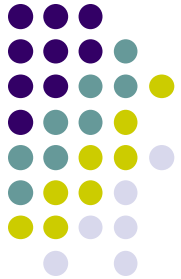
The “Bad” Lawyer

- **U.S. litigation increasingly demands the preservation of and access to all relevant documents, including in the form of “electronically stored information” or “ESI”**
- **Courts impose sanctions on parties for failing to preserve evidence under the “spoliation” doctrine**
- **Absent saving everything, often it is only with 20/20 hindsight that one can determine what *should* have been preserved in response to a lawsuit**
- **Recordkeeping solutions that rely on human judgment are prone to being second-guessed by litigants and judges.**

Four RecordkeepingPaths

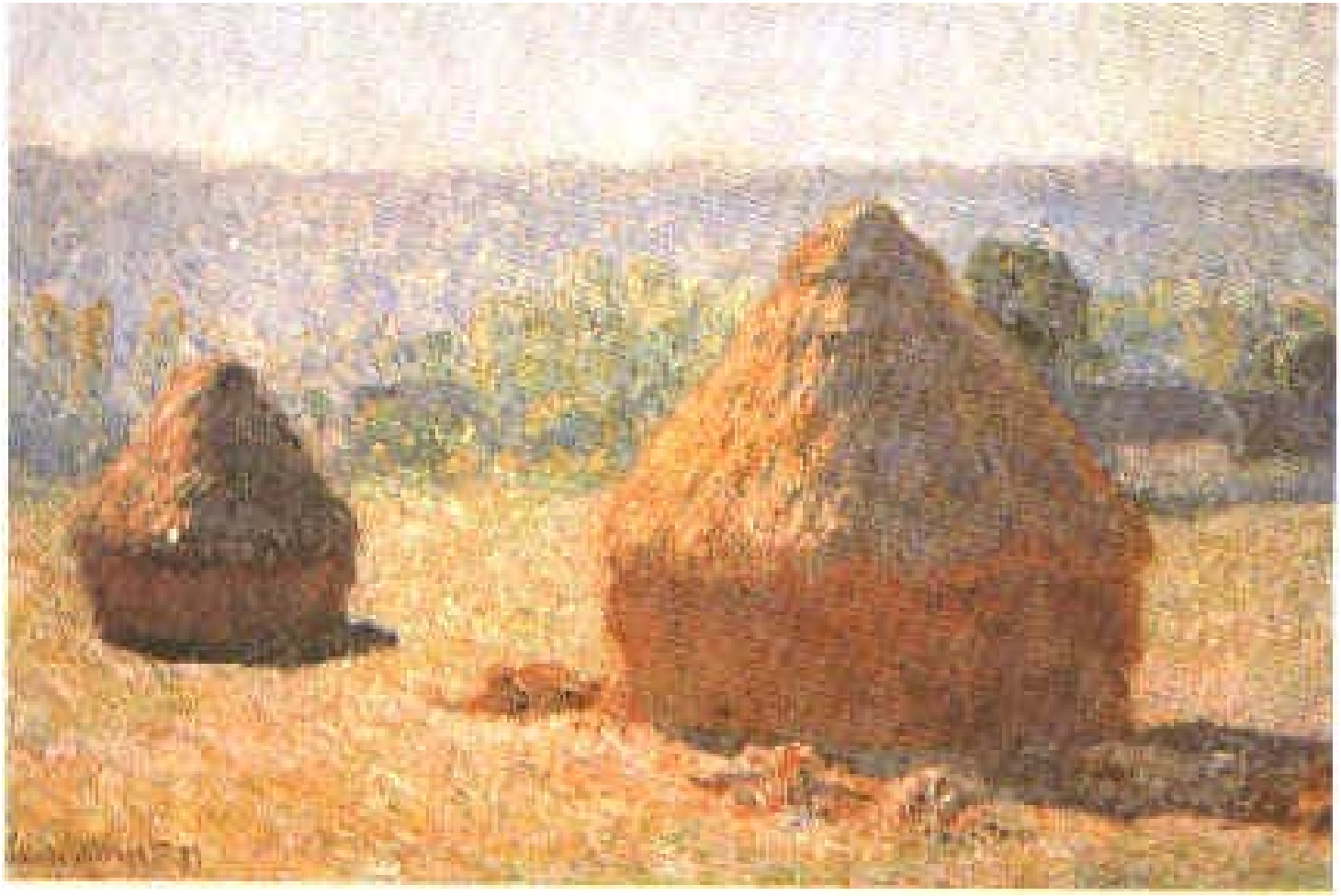
- **Print to paper: noncompliance and an increasingly failing paradigm**
- **Backup tapes: nonarchival and huge restoration costs to restore records**
- **Online user-based foldering in proprietary live systems: fractal recordkeeping devoid of KM**
- **Electronic recordkeeping under DoD 5015.2: reliance on human judgment & transactional costs**

The Future “Promise” of Total E-record Archiving



- 100% archiving of email & ESI on the desktop
- Transport out of email store into generic format (e.g., XML)
- Use of smart filter technologies to segregate permanent from the temporary
- Culling for non-record material using certain agreed-upon rules and protocols
- Default temporary record status of remaining archived materials
- However: all “eggs” in the search basket

Searching the Haystack….



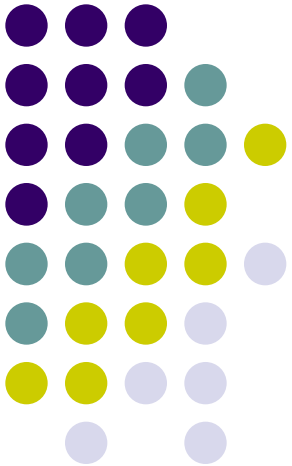
to find relevant needles…

ends up like searching in a maze…

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.

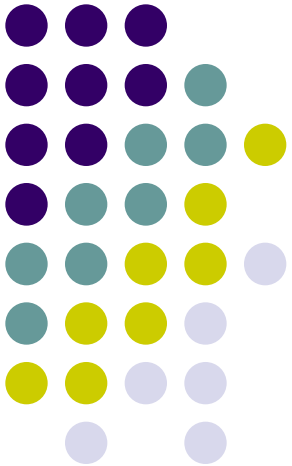
Corollary: in conducting automated searches, the use of "keywords" alone will reliably produce all or substantially all documents from a large document collection.



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.

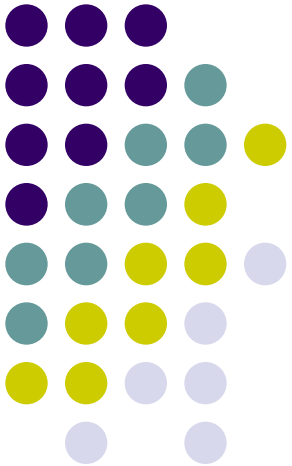
Corollary: claims that documents can be easily segregated by examination of content.



The Reality of Search & Retrieval

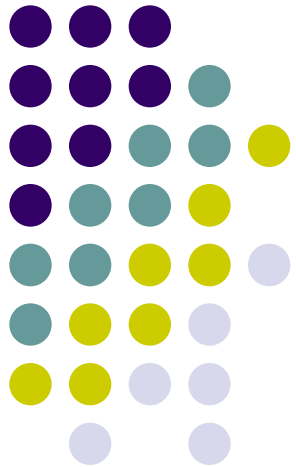
+ Past research (Blair & Maron, 1985) has shown a gap or disconnect between lawyers' perceptions of their ability to ferret out relevant documents, and their actual ability to do so:

--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.



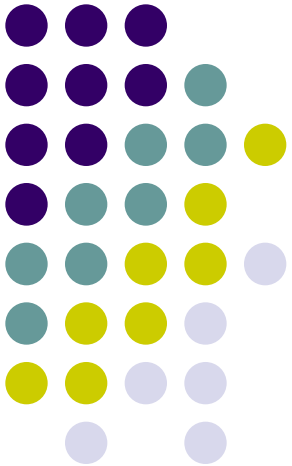
More Reality: IR is Hard

+ 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.

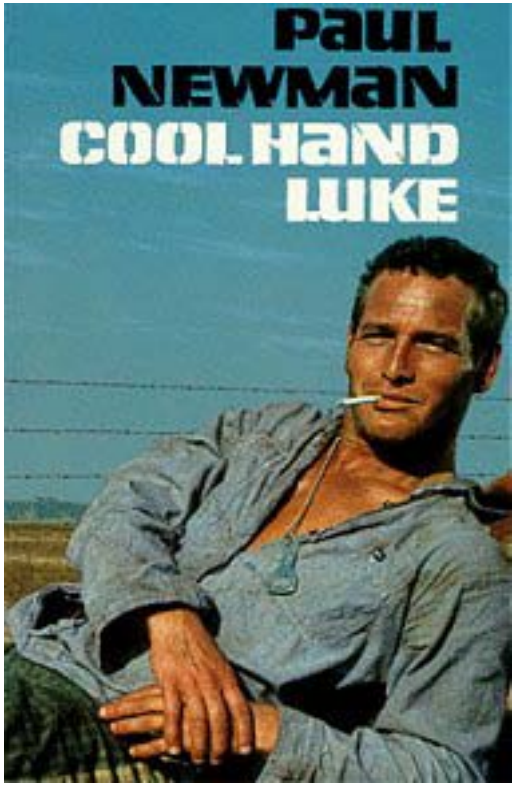


Why is IR hard (in general)?

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

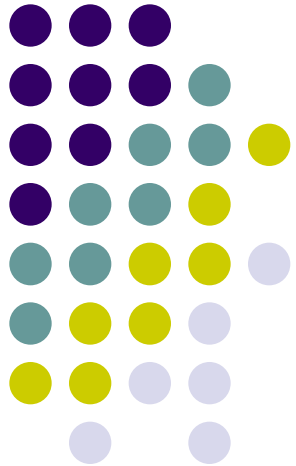


“What we’ve got here is a failure to communicate”

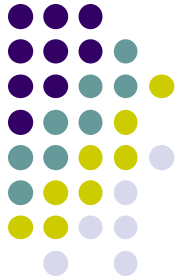


Problems of language

- 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”)



A Plug for Dealing With Search Issues In the Context of Real User Needs



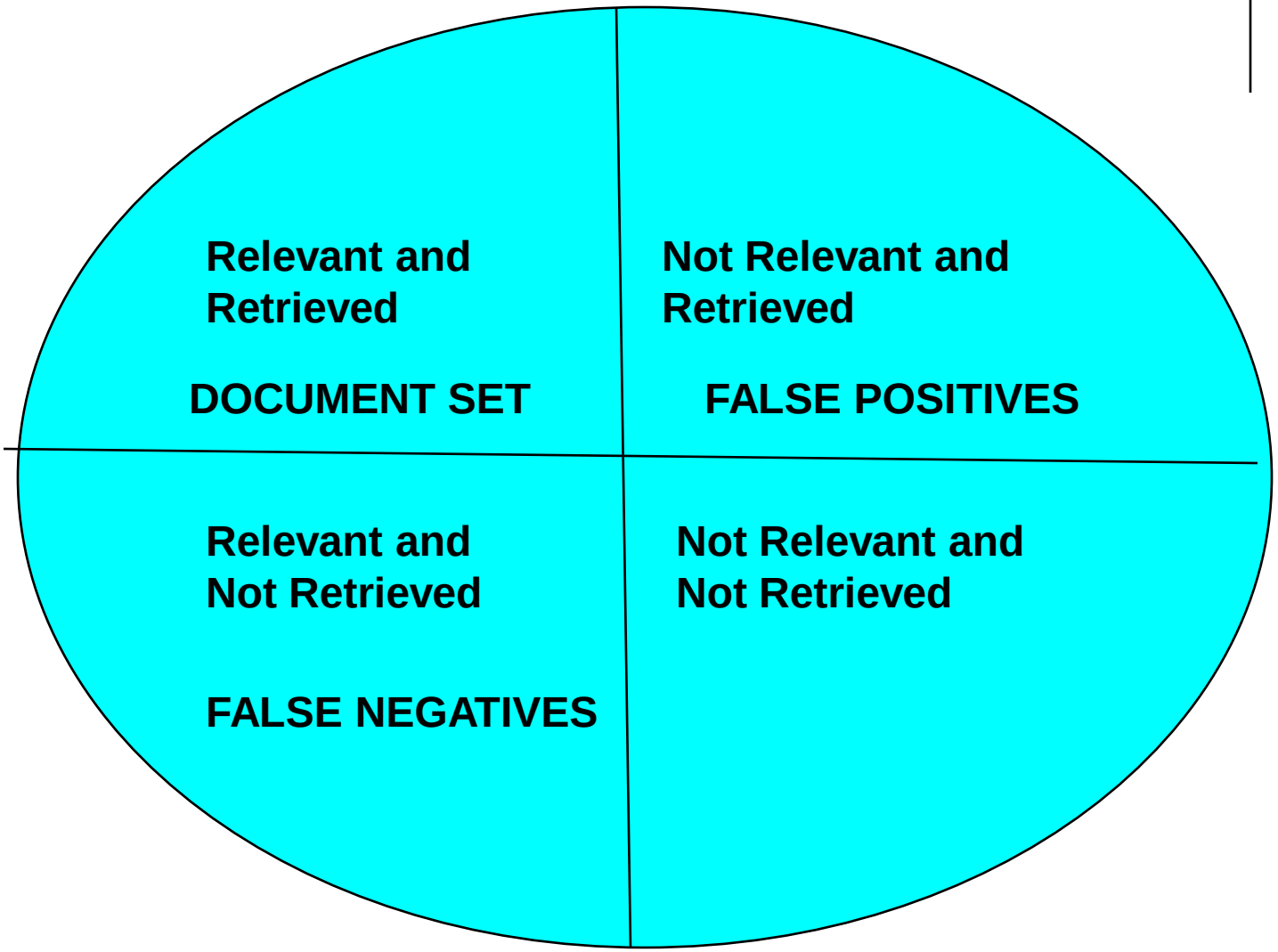
- Finding responsive needles in E-haystacks: the problems with keywords
- Maximizing recall of responsive docs
- Weeding out false positives
- Evaluating competing search products in the marketplace against some objective standard lawyers will embrace

See "Information Inflation: Can The Legal System Adapt?," George L. Paul and J.R. Baron, in 13 Richmond Journal of Law & Technology 10 (2007), http://law.richmond.edu/jolt/v13i3/article10.pdf, and The Sedona Conference ® Commentary on The Use of Search and Information Retrieval Methods in E-Discovery (2007 draft), http://www.thesedonaconference.org

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)**

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



What is TREC?

- Conference series co-sponsored by the U.S. National Institute of Standards and Technology (NIST) and the Advanced Research and Development Activity (ARDA) of the U.S. 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 Gaithersburg, Maryland (NIST headquarters)

TREC 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
- Hypothetical complaints and “requests to produce” drafted by members of The Sedona Conference®
- “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 from around the world including in academia and legal service providers in commercial space

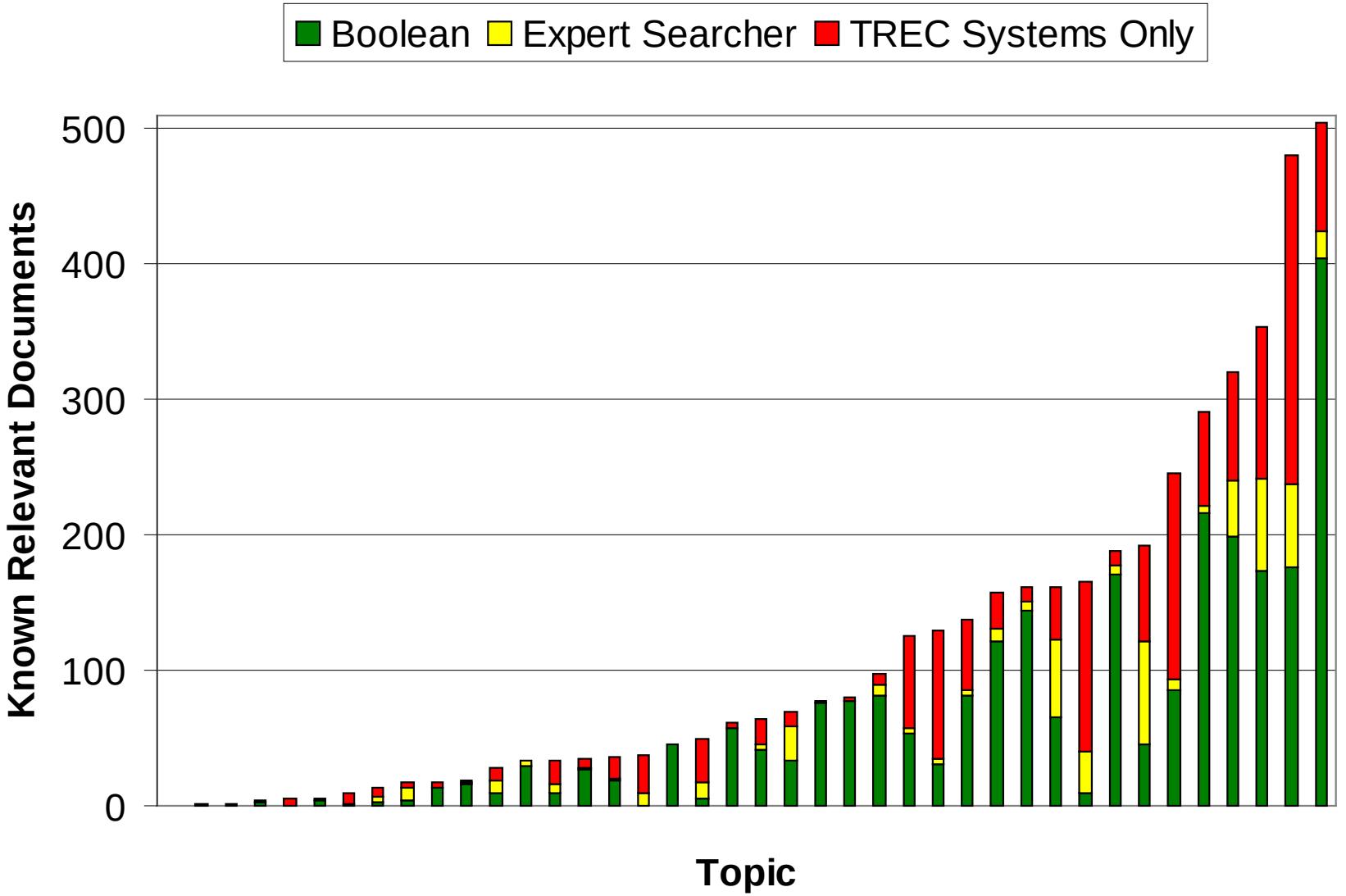
TREC 2006 LEGAL TRACK XML ENCODED TOPICS WITH NEGOTIATION HISTORY – ONE EXAMPLE

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <TrecLegalProductionRequest>
- <ProductionRequest>
<RequestNumber>**6**</RequestNumber>
<RequestText>**All documents discussing, referencing, or relating to company guidelines or internal approval for placement of tobacco products, logos, or signage, in television programs (network or cable), where the documents expressly refer to the programs being watched by children.**</RequestText>
- <BooleanQuery>
<FinalQuery>**(guide! OR strateg! OR approval) AND (place! OR promot! OR logos OR sign! OR merchandise) AND (TV OR "T.V." OR televis! OR cable OR network) AND ((watch! OR view!) W/5 (child! OR teen! OR juvenile OR kid! OR adolescent!))**</FinalQuery>
- <NegotiationHistory>

Beyond Boolean: getting at the “dark matter”

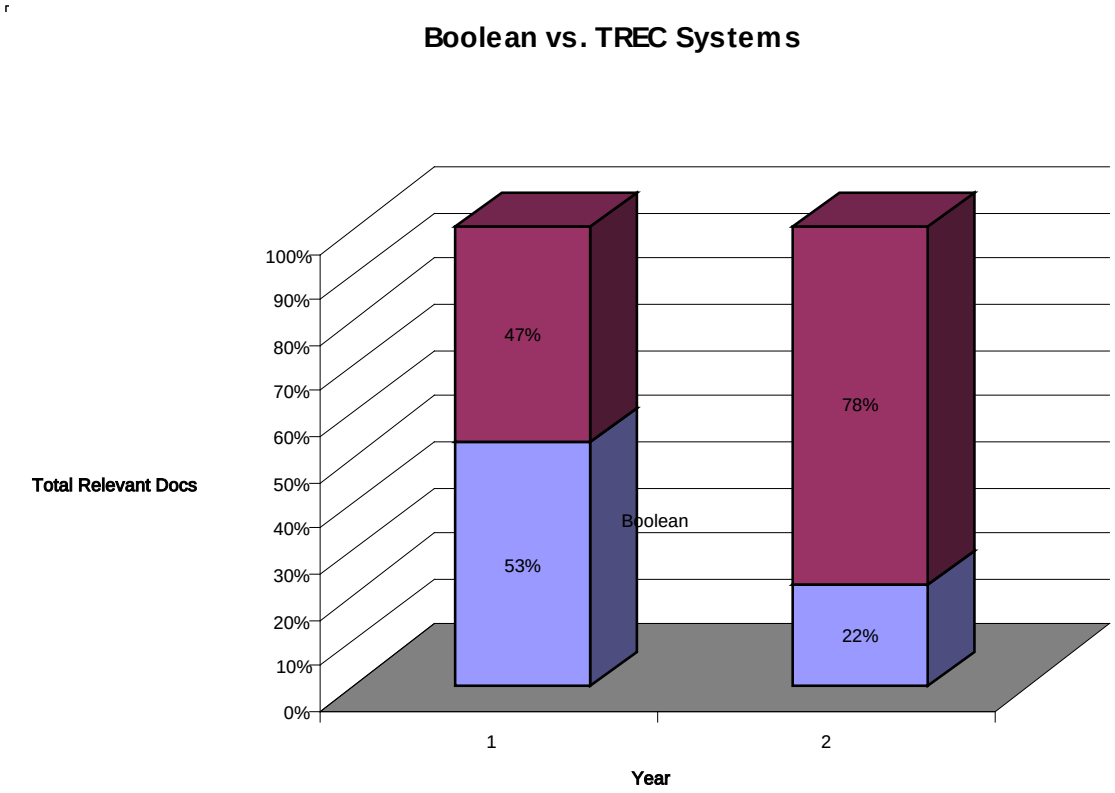
(*i.e., relevant documents not found by keyword searches alone)*

Nobody Finds Everything

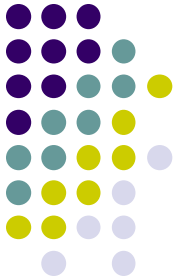


Source: TREC 2006 Legal Track

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

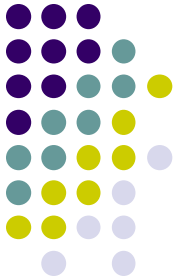


Judge Grimm writing for the U.S. District Court for the District of Maryland



"[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)

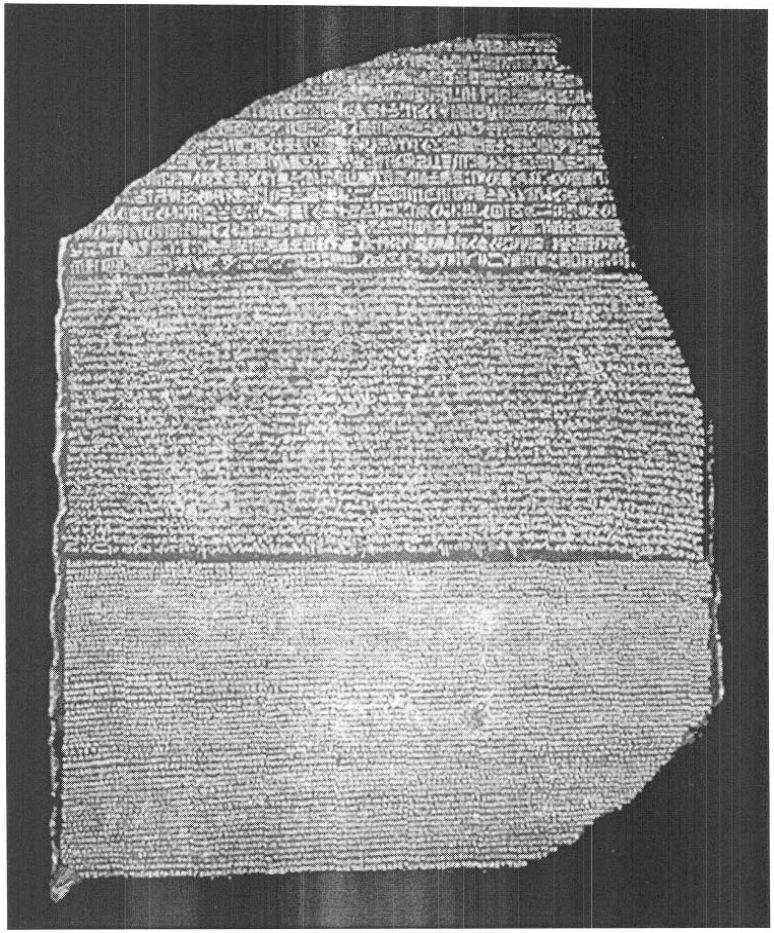
Judge Facciola writing for the U.S. District Court for the District of Columbia



"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).

Rosetta Stone Approach: The Archivist's Need To Master 3 Languages: Legal, RM, IT



Whither Appraisal?

- **The IT savvy archivist will be able to crosswalk between technological solutions for filtering exponentially increasing volumes of information, and traditional notions of appraisal**
- **The "information retrieval" savvy archivist will understand that a range of alternative search methods – including AI methods yet to be developed -- hold the *key* to the efficient future accessing of information**
- **The neutral archivist may serve as an unbiased resource for the filtering of information in an increasingly partisan (untrustworthy) world.**

Jason R. Baron

Adjunct Professor

College of Information Studies

University of Maryland, College Park, MD

(301) 837-1499

Email: jason.baron@nara.gov