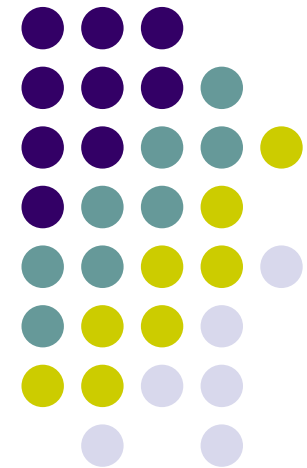


Through A Lawyer's Lens: Measuring Performance in Conducting Large Scale Searches Against Heterogeneous Data Sets in Satisfaction of Litigation Requirements

II Workshop
University of Pennsylvania
Philadelphia, PA

Jason R. Baron
Director of Litigation
Office of General Counsel
National Archives and Records Administration
October 26, 2006



Searching Through E-Haystacks: A Big Challenge





Overview

- ESI Overload
- ESI & The New Federal Rules
- Issues in Search and Info. Retrieval: A Case Example
- Challenges
- Additional Sources

The problem: heterogeneous “electronically stored information” (ESI)

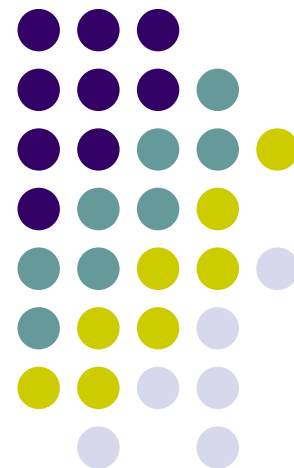


- E-mail
- Word processing, spreadsheets, all applications
 - Integrated with voice mail & VOIP
 - Instant messaging
 - Web pages including intraweb sites
 - Blogs, wikis, and RSS feeds
 - Backup tapes, hard drives
- Removable media, thumb drives & new storage devices
 - Remote PDAs, blackberrys
 - Metadata of all types

ESI and the New Federal Rules

“Electronically stored information” as a new category subject to Rule 34 requests for documents

Rule 34(b): request may specify form in which ESI is to be produced, subject to objection



ESI and the New Federal Rules



- Rule 26(f)

At the parties' planning meeting, issues expected to be discussed include:

“Any issues relating to disclosure or discovery of electronically stored information, including the form or forms in which it should be produced”

“Any issues relating to preserving discoverable information”

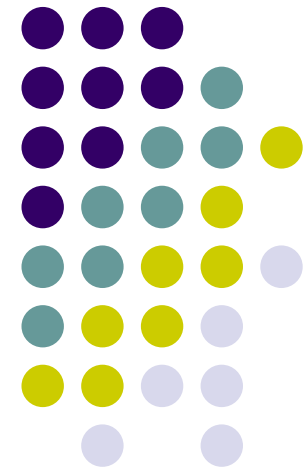
The background of the slide is a scenic photograph of Sedona, Arizona. It features prominent red sandstone rock formations, including several tall, pointed spires. The sky is a vibrant blue with scattered white clouds. In the upper right corner, the dark green needles of a pine tree are visible. The overall lighting suggests a bright, sunny day.

**The Sedona
Conference Working
Group on Best
Practices for Electronic
Document Retention
and Production**

www.thesedonaconference.org

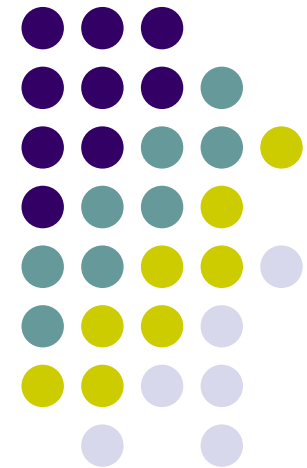
Sedona Principle 11

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



Case Precedent on Automated Searches

- *Current cases emphasize keyword searches under the rubric of “search protocols” (Treppel v. Biovail)*
- *Case law on employing concept searches and other alternative search methodologies is expected to emerge over time*



Real life example: *U.S. v. Philip Morris et al.*



- Civil lawsuit brought by Clinton Administration against tobacco companies in 1999
- RICO case – racketeering allegation that companies have conspired since 1953 to defraud the American public as to the true health effects of smoking
- Judge Kessler’s August 17, 2006 decision

U.S. v. Philip Morris – E-discovery



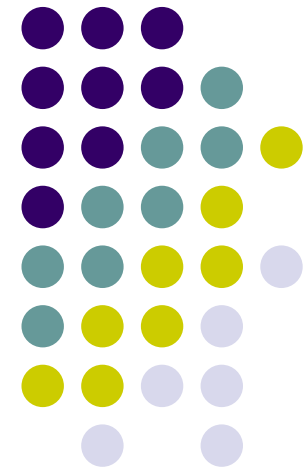
- 1,726 Requests to Produce propounded by tobacco companies on U.S. (30 federal agencies, including NARA) for tobacco related records
- Along with paper records, email records were made subject to discovery
- 32 million Clinton era email records – government had burden of searching

NARA keyword searches

Original Search Terms Used By NARA on email database:

- Tobacco
- Cigarette
- Smoking
- <Tar>
- Nicotine
- smokeless
- Synar Amendment

- Philip Morris
- R.J. Reynolds
- Brown and Williamson
- BAT Industries
- Liggett group

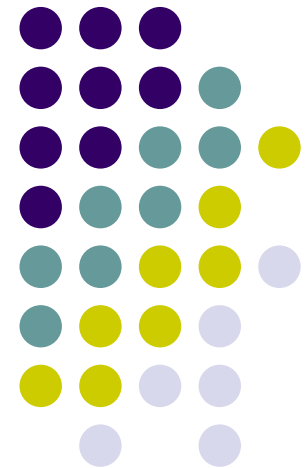


Problem with false positives

**Marlboro BUT NOT Upper Marlboro
(Maryland)**

**PMI (for Philip Morris Institute) BUT NOT
presidential management intern**

**TI (for Tobacco Institute) BUT NOT do re
me . . .**



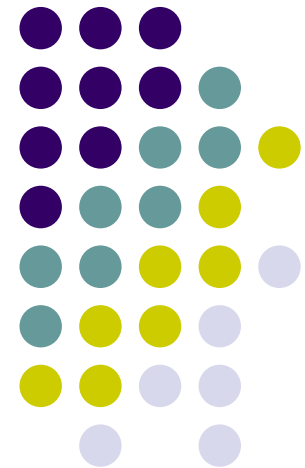


Example Search String

- **((((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)**

Litigation Targets

- + Defining “relevance”
- + Maximizing Recall
- + Maximizing Precision
- + Dealing with fuzzy language concerns
- + Recognizing Inherent ambiguity & variation in all texts



Beyond Boolean: Alternative Search Methods

Probabilistic models (Bayesian)

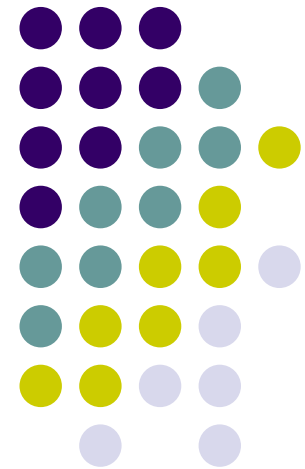
Fuzzy Search Models

Statistical methods (clustering)

Machine learning approaches to semantic representation

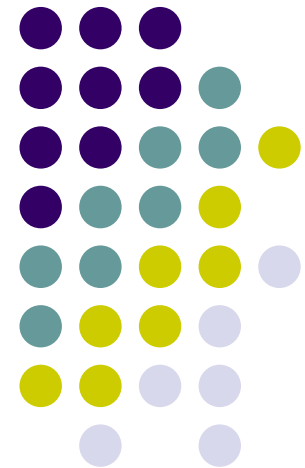
Categorization tools: taxonomies and ontologies

Visualization techniques & social network analysis



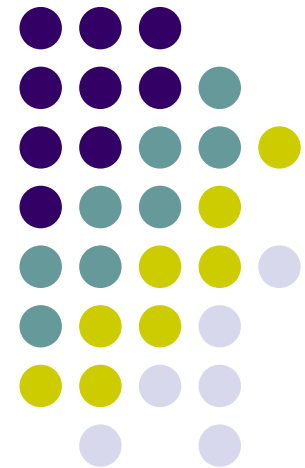
Information Integration Challenges

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



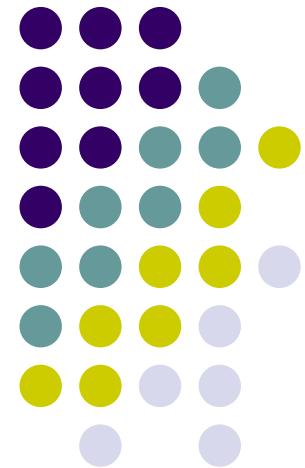
Challenges (con't)

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 (i.e., Recall \neq 1.000)



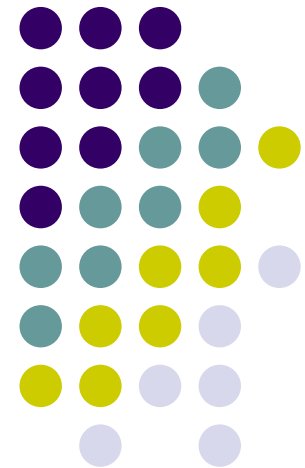
Challenges (con't)

Parties making a good faith attempt to collaborate on the use of particular search methods



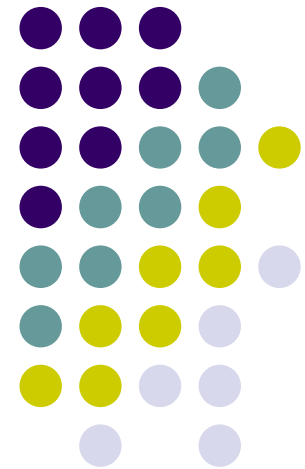
Challenges (con't)

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



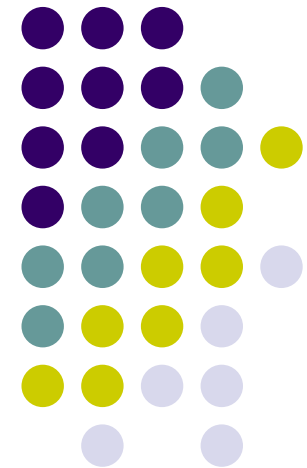
Challenges (con't)

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)



TREC Legal Track 2006

- + National Institute of Standards and Technology (NIST) Text Retrieval Conference
- + New legal track project for 2006: evaluating efficacy of search methods in a legal discovery context
- + Results to be reported in Nov. 2006



Case law re: keywords & search protocols



Recent case law where judges have intervened to adjudicate or order search protocols:

- *Balboa Threadworks v. Stucky*, 2006 WL 763668 (D. Kan. Mar. 24, 2006)
- *Medtronic v. Sofamor Danek, Inc. v. Michelson*, 229 F.R.D. 550 (W.D. Tenn 2003)
- *Treppel v. Biovail Corp.*, 233 F.R.D. 363 (S.D.N.Y. 2006)
- See also *Zubulake v UBS Warburg LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004) (discussing keyword searches generally)



References

- Baron, “Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery,” 6 *Sedona Conference Journal* 237 (2005) (available on Westlaw and LEXIS)
- K. Withers, “Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure,” 4 *Nw. J. of Tech. & Intell. Prop.* 171 (2006), available at <http://www.law.northwestern.edu/journals/njtip/v4/n2/3>

Additional References

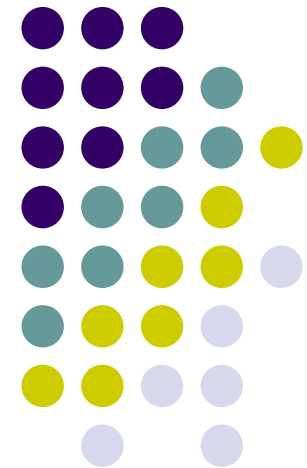
+ Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivities*, held Nov. 8, 2005, see

http://colab.cim3.net/cgi-bin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

+ NIST/TREC Legal Track 2006 home page, see <http://trec-legal.umiacs.umd.edu>

+ Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), Principle 11, see http://www.thesedonaconference.org/content/miscFiles/publications_html.

+ Sedona Conference, *Commentary on Search & Retrieval Issues* (forthcoming 2007)





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

8601 Adelphi Road Suite 3110

College Park, MD 20740

(301) 837-1499

Email: jason.baron@nara.gov