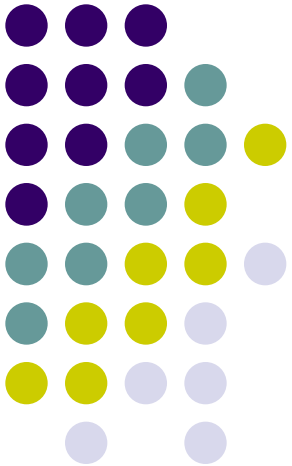
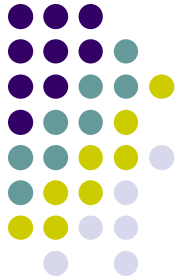
Session 8: Cooperation in E-Discovery Search

LBSC 708X/INFM 718X
Seminar on E-Discovery
Jason R. Baron
Adjunct Faculty
University of Maryland
March 15, 2012



Overarching Questions

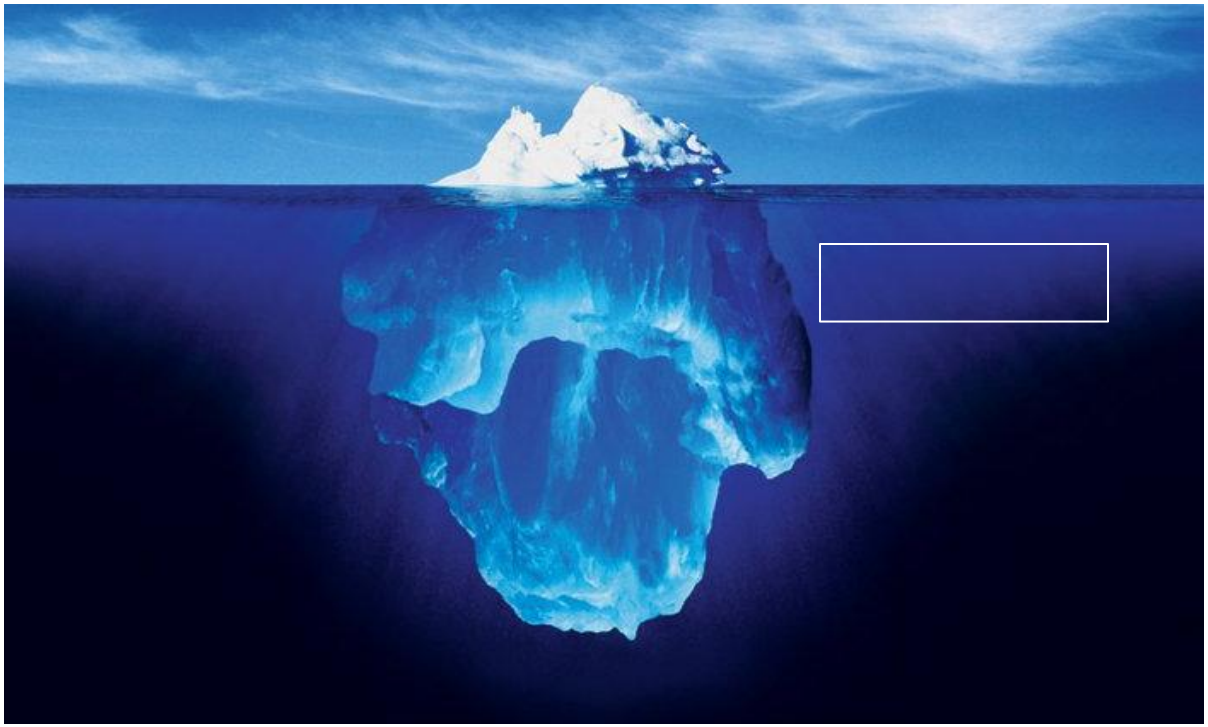


- How is the information retrieval task inside organizations like or unlike the problem of finding a good restaurant tonight after class in Baltimore?
- What constitutes “cooperation” when it comes to searching for relevant evidence?
- What are your ethical duties in light of the asymmetries inherent in e-discovery search?

Federated Searches Across The Enterprise & The Problem of Compartmentalization

Hradcany Castle and Charles Bridge, Prague

How much of one's client's ESI is known to in-house counsel at the onset of litigation?



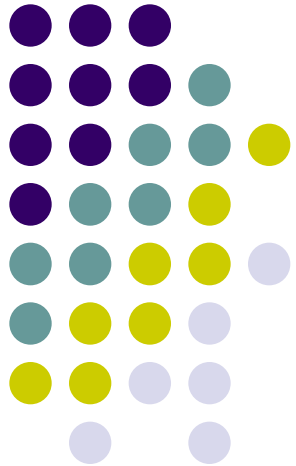
Information Governance Challenges

Data Silos

+ **Not Knowing What You Know, What Data You Have (In Which Silo)**
+ **Federated Search Challenges: Custodians, Legacy Systems, Complex Data Sets**

The Sedona Principles (2d ed. 2008), Principle 6

Responding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information.



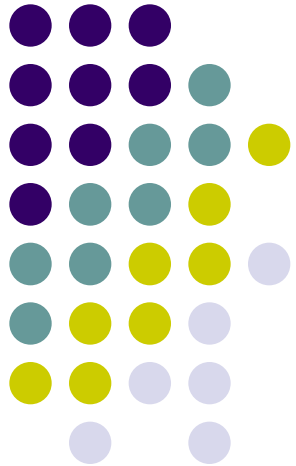
In Ford Motor Co. v. Edgewood Properties Inc., 257 F.R.D. 418, 427 (D. N.J. 2009), in the face of allegations of missing evidence, the Court upheld a manual collection process used by Ford Motor, acknowledging that "manual collection is sometimes even disfavored [citing to The Sedona Conference Commentary on Search and Retrieval], but going on to note that "absent an agreement or timely objection, the choice is clearly within the producing party's sound discretion."

Please your Honor, may I have some more (ESI, that is)?

"In the face of a protest of 'inexplicable deficiencies' in a party's production, vague and speculative notions that there, in essence, are insufficient to compel judicial action." Judge Facciola, writing in *U.S. v. O'Keefe,* 537 F. Supp. 2d 14, 22 (D.D.C. 2008)

But there are gale force winds in the case law…

Cf. Judge Scheindlin writing in *Pension Committee of the University of Montreal Pension Plan v. Banc of America,* found plaintiffs' litigation hold policy defective in part because:



"It does not direct employees to *preserve* all relevant records-both paper and electronic-nor does it create a mechanism for *collecting* the preserved records so that they can be searched by someone other than the employee. Rather, the directive places total reliance on the employee to search and select what that employee believed to be responsive records without any supervision from Counsel."

685 F.Supp.2d 456, 473 (S.D.N.Y. 2010) (as amended May 28, 2010)

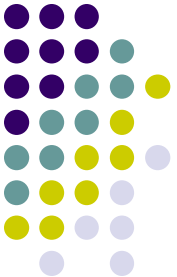
Judge Scheindlin's Opinion in Pension Committee goes on to say:

> I note that not every employee will require hands-on supervision from an attorney. However, attorney oversight of the process, including the ability to review, sample, or spot-check the collection efforts is important. The adequacy of each search must be evaluated on a case by case basis.

Citing to:

Adams v. Dell, 621 F.Supp.2d 1173, 1194 (D.Utah 2009) (holding that defendant had violated its duty to preserve information, in part because the defendant's preservation practices "place operations-level employees in the position of deciding what information is relevant")

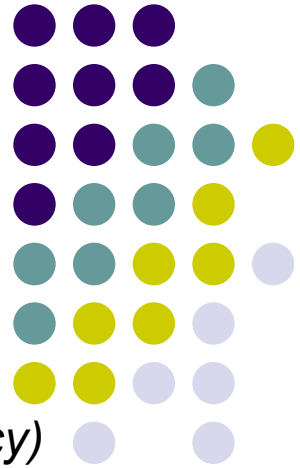
Jones v. Bremen High School District 228, 2010 WL 2106640 (N.D. Ill. May 25, 2010)



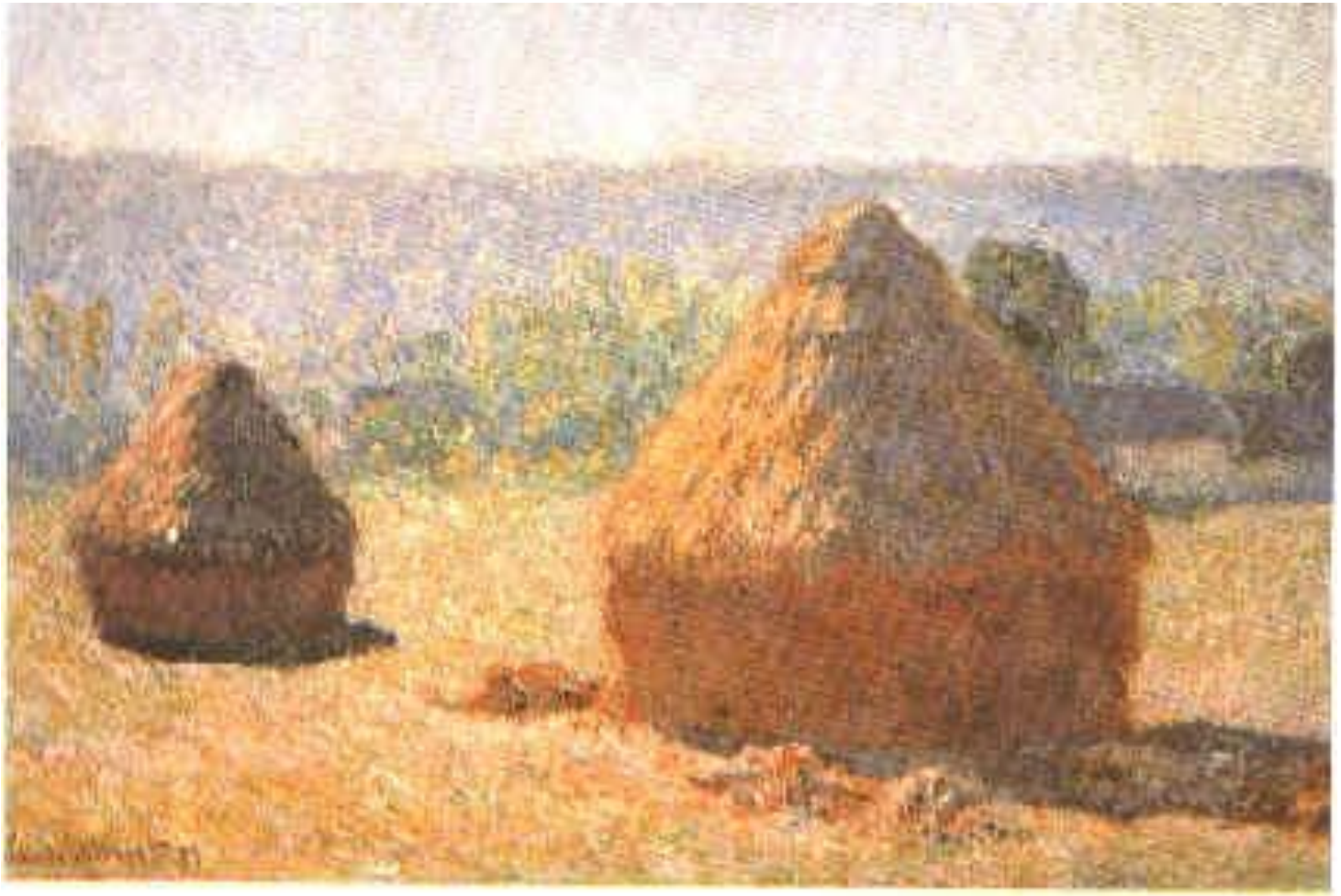
"The Court finds that defendant clearly breached its duty to preserve relevant documents in this litigation. * * * * Defendant also was aware that its employees were able to delete permanently emails. Despite these indisputable facts, defendant inexplicably did not request *all* employees who had dealings with plaintiff to preserve emails so that they could be searched further for possible relevance to plaintiff's case by counsel. Instead, defendant directed just three employees (one of whom was at the center of plaintiff's complaints) to search their own email without help from counsel and to cull from that email what would be relevant documents. It is unreasonable to allow a party's interested employees to make the decision about the relevance of such documents, especially when those same employees have the ability to permanently delete unfavorable email from a party's system. * * * * Most non-lawyer employees . . . . do not have enough knowledge of the applicable law to correctly recognize which documents are relevant to a lawsuit and which are not. Furthermore, employees are often reluctant to reveal their mistakes or misdeeds."

Issues and Challenges of Manual, Custodian Based Methods

1. *Under-collection*
2. *Inconsistent, idiosyncratic searching for purpose of collection*
3. *Late identification of key evidence*
4. *Metadata spoiliation*
5. *Self-interest, bias*
6. *End user's absence of legal knowledge (e.g., relevancy)*
7. *Failure of attorney supervision (being out of loop)*
8. *Burdens, costs, and the risk of a do-over*



Searching the Enterprise Haystack….



to find relevant needles…

ends up like searching in a maze…

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**

Reality:
Big data and litigation

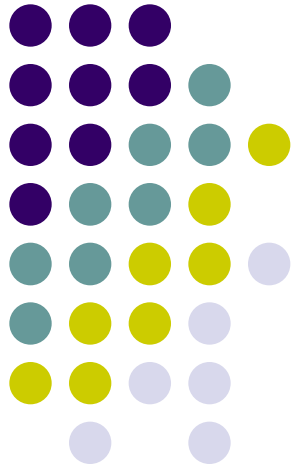
Lehman Brothers Investigation

-- 350 billion page universe (3 petabytes)

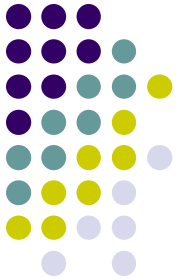
-- Examiner narrowed collection by selecting key custodians, using dozens of Boolean searches

-- Reviewed 5 million docs (40 million pages using 70 contract attorneys)

Source: Report of Anton R. Valukas, Examiner, In re Lehman Brothers Holdings Inc., et al., Chapter 11 Case No. 08-13555 (U.S. Bankruptcy Ct. S.D.N.Y. March 11, 2010), Vol. 7, Appx. 5, at http://lehmanreport.jenner.com/.

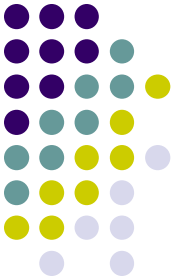


Judge Grimm writing for the U.S. District Court for the District of Maryland



"[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)

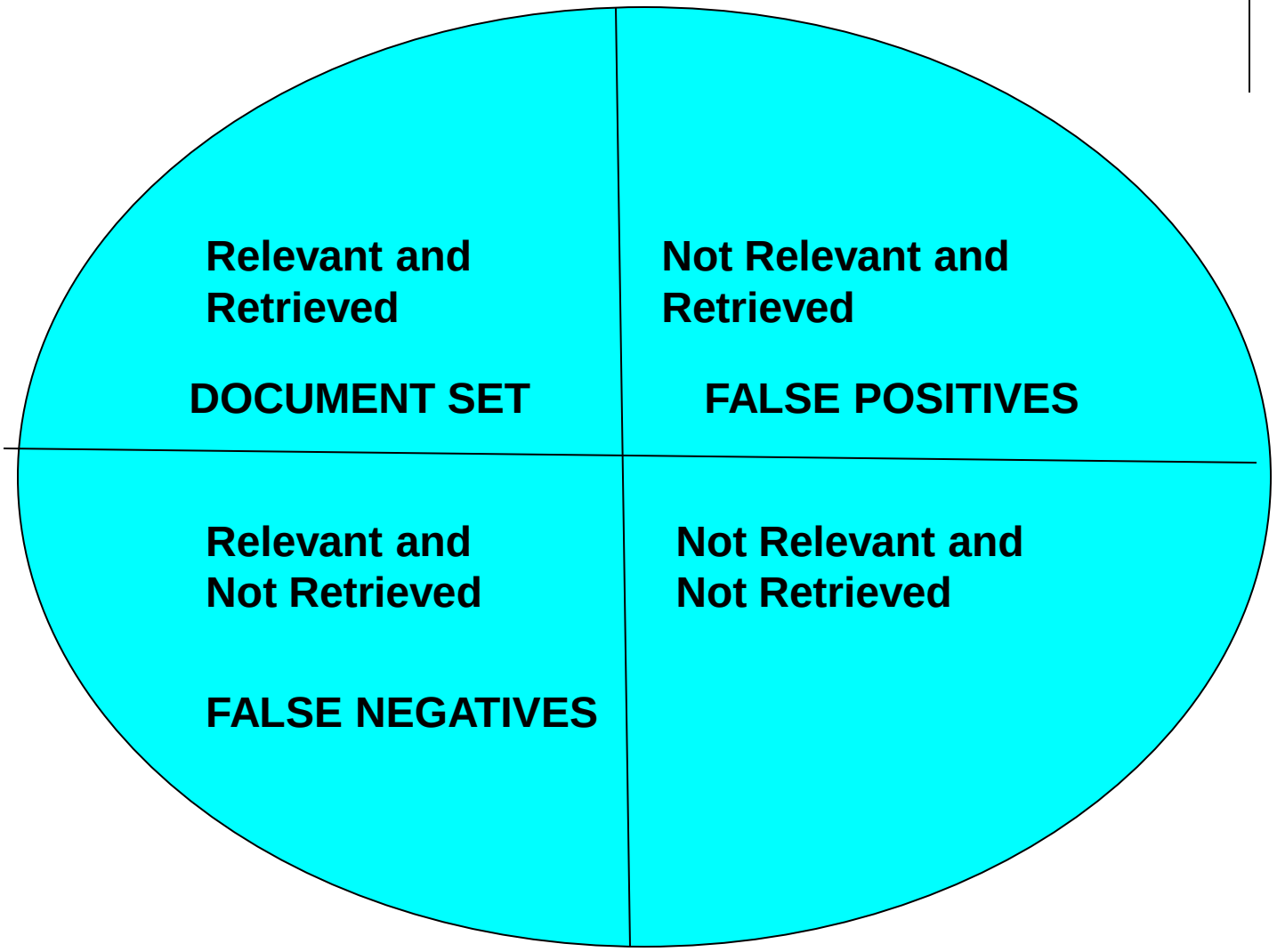
Judge Facciola writing for the U.S. District Court for the District of Columbia



"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).

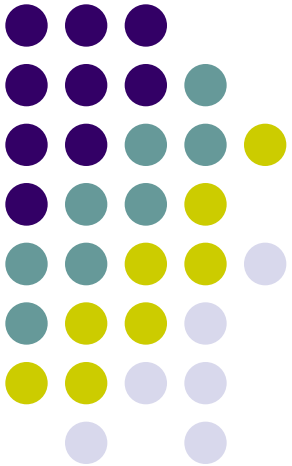
FINDING RESPONSIVE DOCUMENTS IN A LARGE DATA SET: FOUR LOGICAL CATEGORIES



The Myth of Search & Retrieval

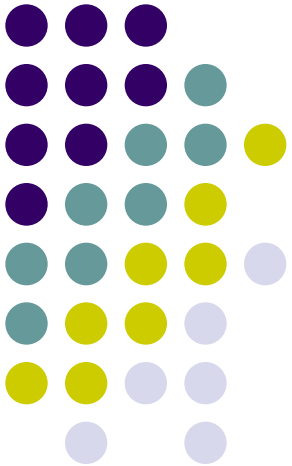
When lawyers request production of "all" relevant documents (and now ESI), all or substantially all will in fact be retrieved by existing manual or automated methods of search.

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

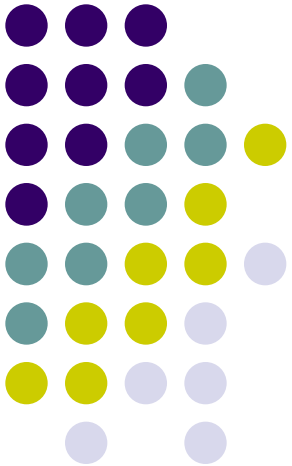
Claims in the legal tech sector that a very high rate of “recall” *(i.e., finding all relevant documents) is easily obtainable provided one uses a particular software product or service.



The Reality of Search & Retrieval

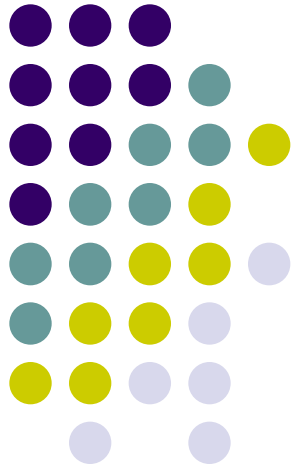
+ Past research (Blair & Maron, 1985) has shown a gap or disconnect between lawyers' perceptions of their ability to ferret out relevant documents, and their actual ability to do so:

--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.



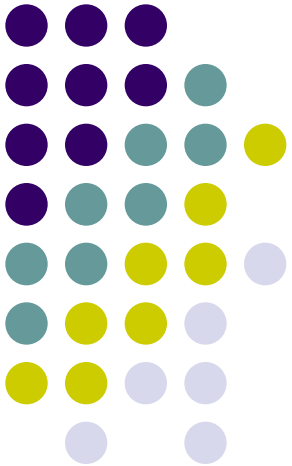
More Reality: IR is Hard

+ 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.



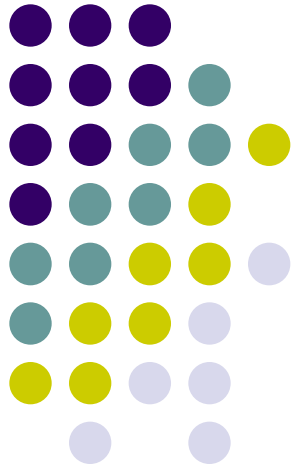
Why is IR hard (in general)?

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



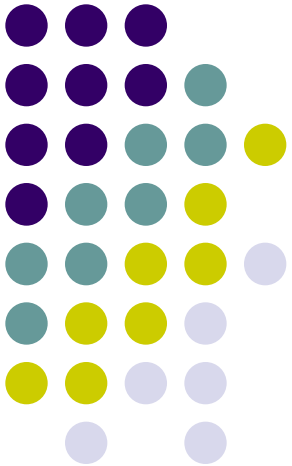
Problems of language

- 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



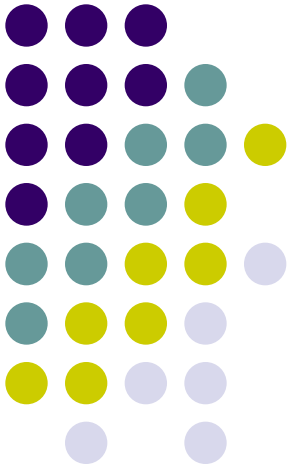
From *The Sedona Conference Best Practices in the Use of Search and Information Retrieval Methods in E-Discovery (2007)*

Practice Point 1. In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.

Practice Point 6. Parties should make a good faith attempt to collaborate on the use of particular search and information retrieval methods, tools, and protocols (including as to keywords, concepts, and other types of search parameters).

Practice Point 7. Parties should expect that their choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials).

Practice Point 8. Parties and the courts should be alert to new and evolving search and information retrieval methods.



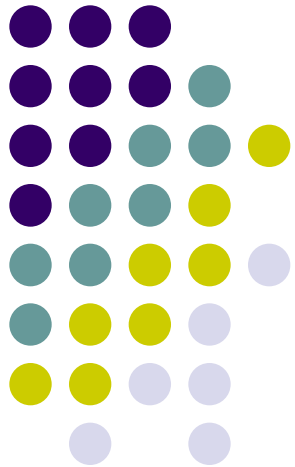
Step 1. The parties meet and confer on the nature of each others' computer hardware and software applications. Proposals are exchanged on the scope of search obligations, in terms of databases and applications to be searched, what active and possibly legacy media, key custodians, time periods. Additionally, keywords are proposed along with any other more sophisticated Boolean or concept search methods. A timetable for conducting searches after the propounding of discovery requests is agreed to.

Step 2. In the interval between meet and confers, parties conduct searches in accordance with prior representations and the actual wording of discovery requests. In doing so they may utilize sampling techniques, and estimates are gathered on the volume of data or "hits" made subject to search.

Step 3. The parties interact further in describing the result of initial searches and preliminary results. If the parties have agreed to a Rule 502 rubric, the parties may elect to share documents found to be potentially responsive. Search terms and protocols are adjusted and search methods are tuned or adjusted for the purpose of conducting more narrow, focused searches.

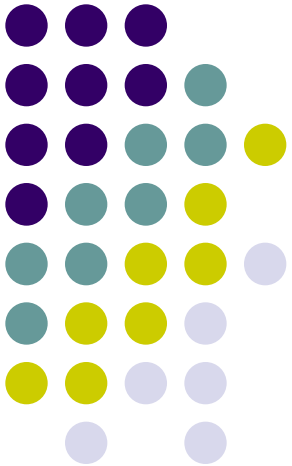
Step 4. The parties may elect to continue iteratively until a mutually agreed time or cap on numbers of responsive documents is reached.

From GPaul & JBaron, "Information Inflation: Can the Legal System Adapt" 13 Rich. J. Law & Tech 10 (2007)



Ethical Issues in Asymmetric Searches

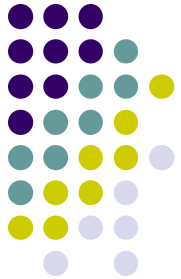
1) Requesting party role
2) Responding party role



ABA Model Rules

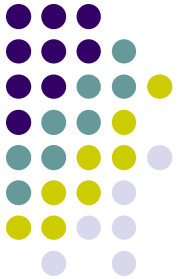
- American Bar Association Model Rule of Professional Conduct 3.4 calls for fairness to an opposing party and counsel. Under Rule 3.4(a), a lawyer shall not unlawfully obstruct another party's access to evidence or unlawfully conceal a document or other material having potential evidentiary value. Under Rule 3.4(d), a lawyer shall not, in pretrial procedure, make a frivolous discovery request or fail to make a reasonably diligent effort to comply with a legally proper discovery request by an opposing party.
- Cf. Model Rule 1.6: a lawyer shall not reveal information relating to the representation of a client unless the client gives informed consent.

Hypotheticals



- Case 1: The misspelled material term
- Case 2: Corporate slang, acronyms, jargon, abbreviations
- Case 3: Synonyms (an incomplete bag of proposed keywords)

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*
- *Hybrid approaches*

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

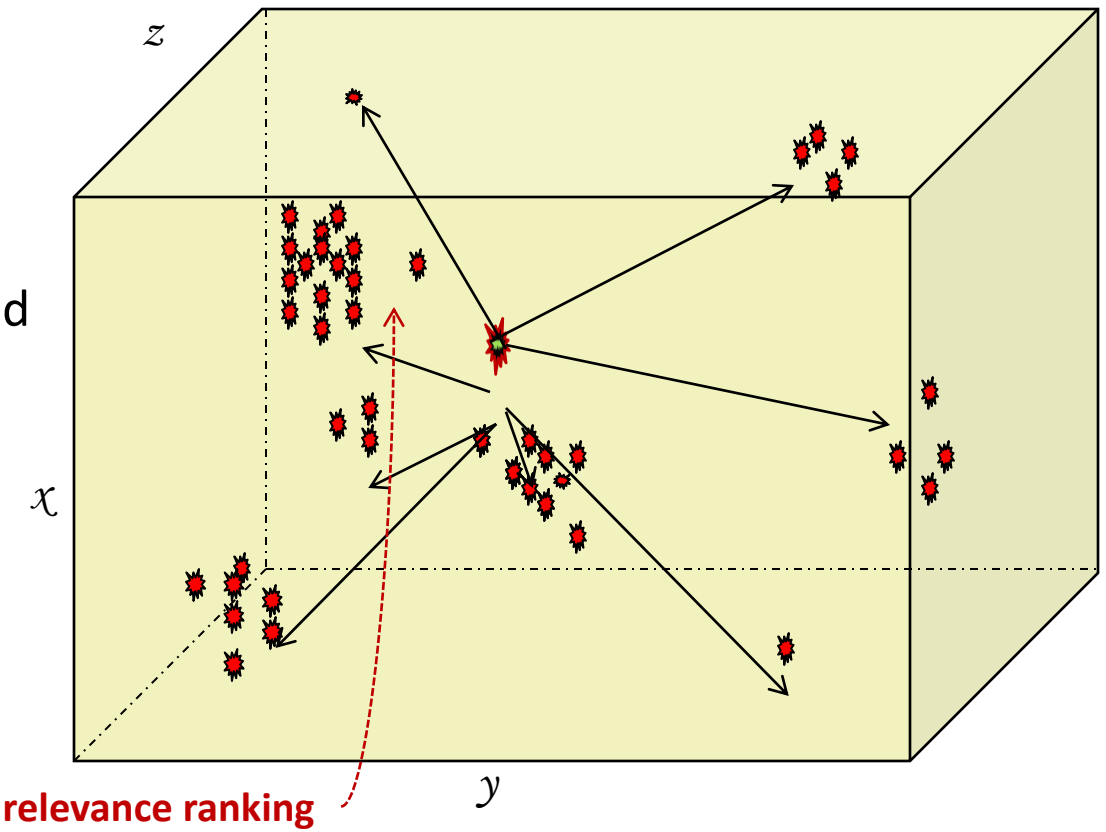
Bayesian Statistical Models

Based on mathematical models of Statistical Probability to recognize documents of similar content.

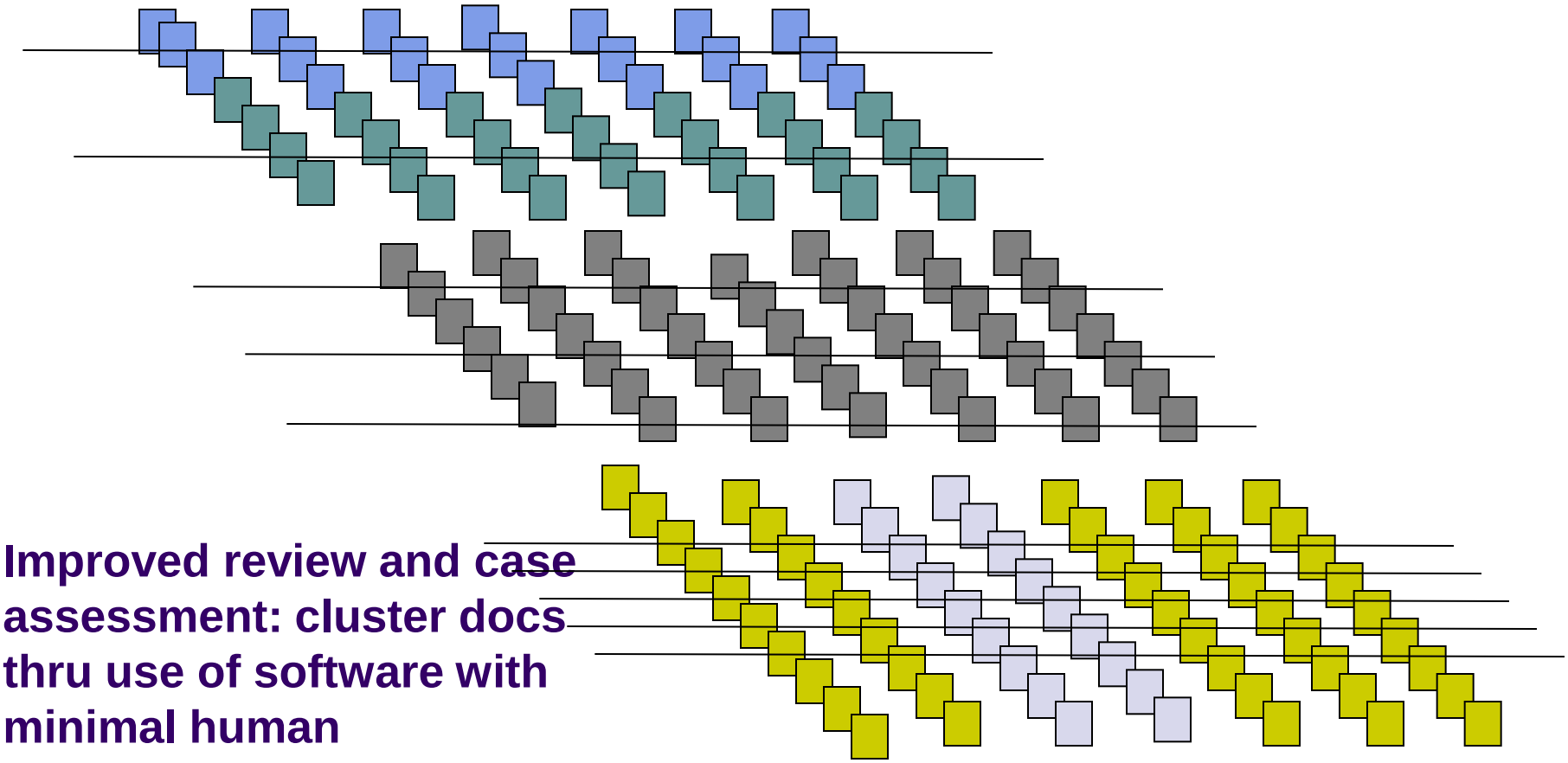
- Learns passively from the document content
- Position, frequency and proximity of terms (language independent) combine to create a mathematical “thumbprint” of concepts contained in documents.
- Useful to “cluster” documents by content
- Can “learn” to build clusters from exemplar sets
- Requires re-indexing and assessment can change

Latent Semantic Indexing (LSI)

1. SVD (Singular Value Decomposition) assigns each record to a place creating "clusters"
2. "Query" documents are SVD analyzed and placed in the matrix
3. "Hits" and rankings are determined by the distance from clusters

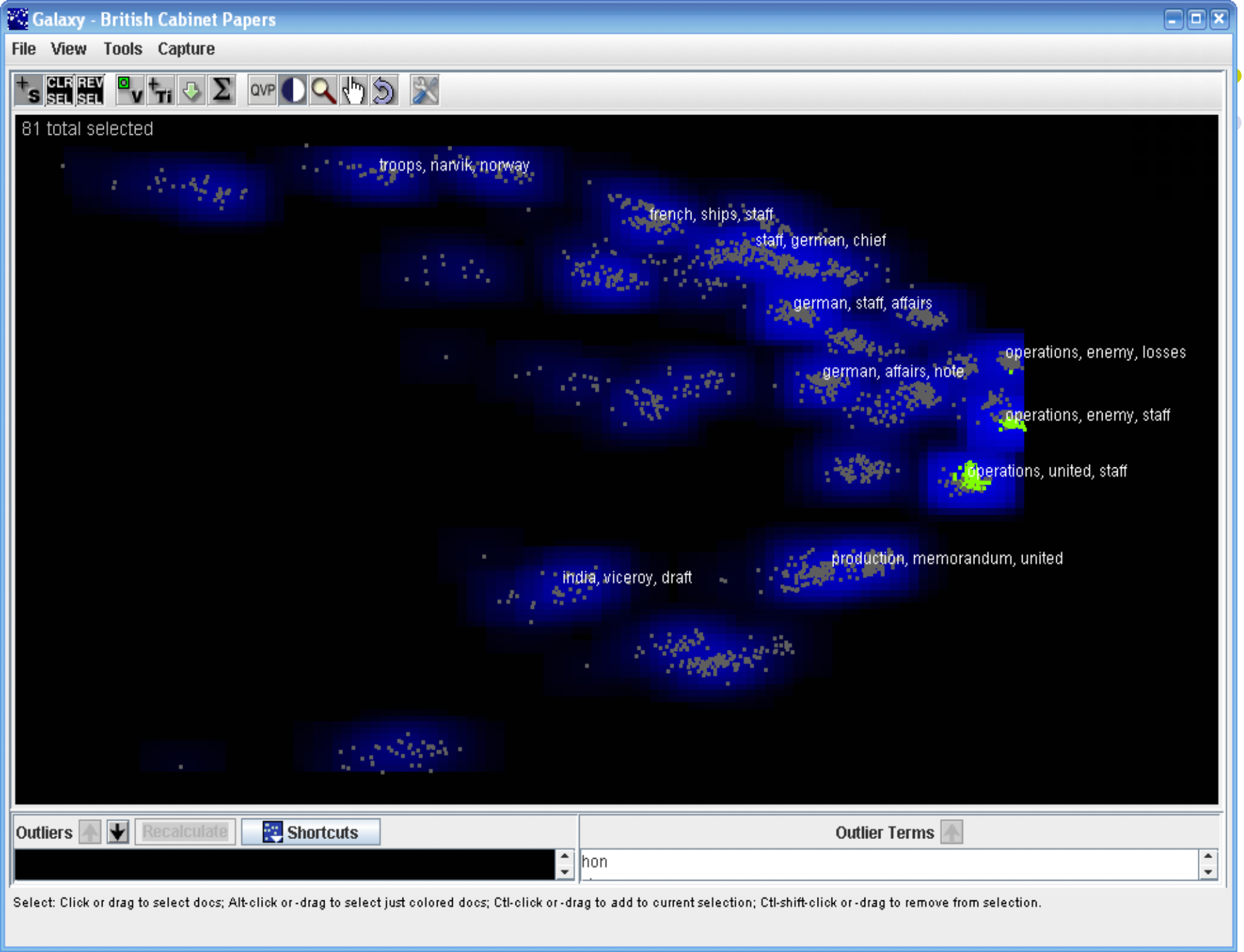


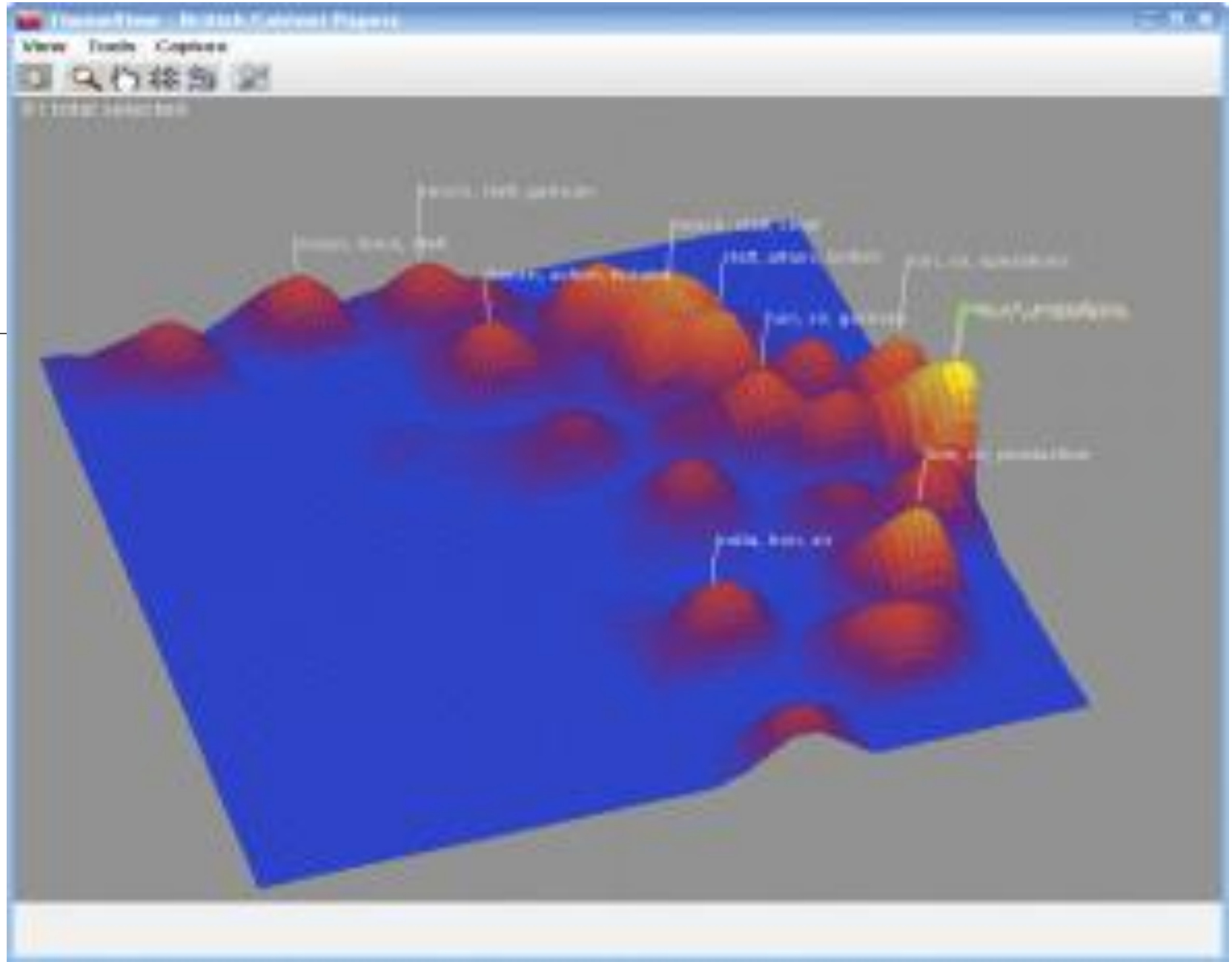
Emerging New Strategies: "Predictive Analytics"



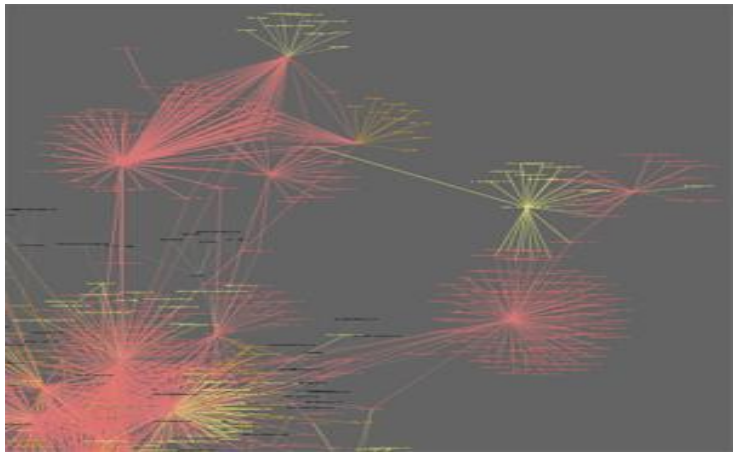
Improved review and case assessment: cluster docs thru use of software with minimal human intervention at front end to code "seeded" data set

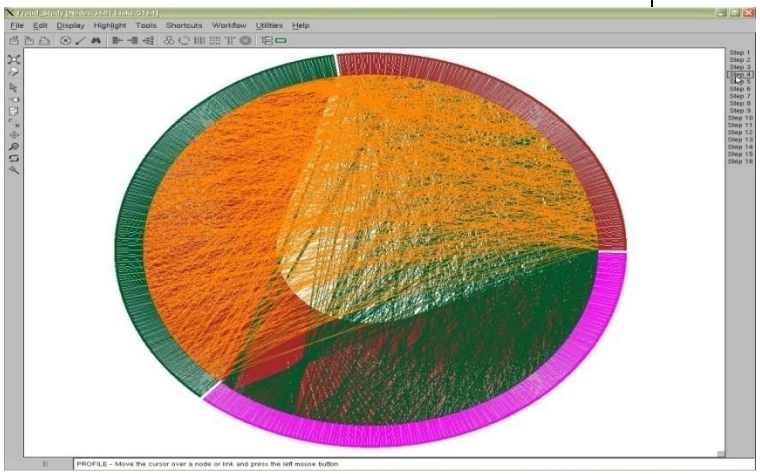
Slide adapted from Gartner Conference
June 23, 2010 Washington, D.C.

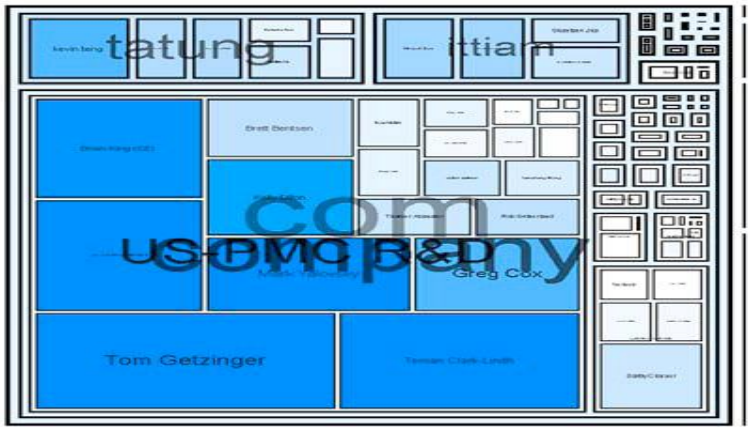

Galaxy - British Cabinet Papers
File View Tools Capture
81 total selected
troops, narvik, norway
french, ships, staff
staff, german, chief
german, staff, affairs
operations, enemy, losses
german, affairs, note
operations, enemy, staff
operations, united, staff
production, memorandum, united
india, viceroy, draft
Outliers
Recalculate
Shortcuts
Outlier Terms
hon
Select: Click or drag to select docs; Alt-click or -drag to select just colored docs; Ctl-click or -drag to add to current selection; Ctl-shift-click or -drag to remove from selection.


View
Tools
Capture

Visual Analysis Examples
(Presentation by Dr. Victoria Lemieux, Univ. British Columbia, at Society of American Archivist Annual Mtg. 2010, Washington, D.C.)

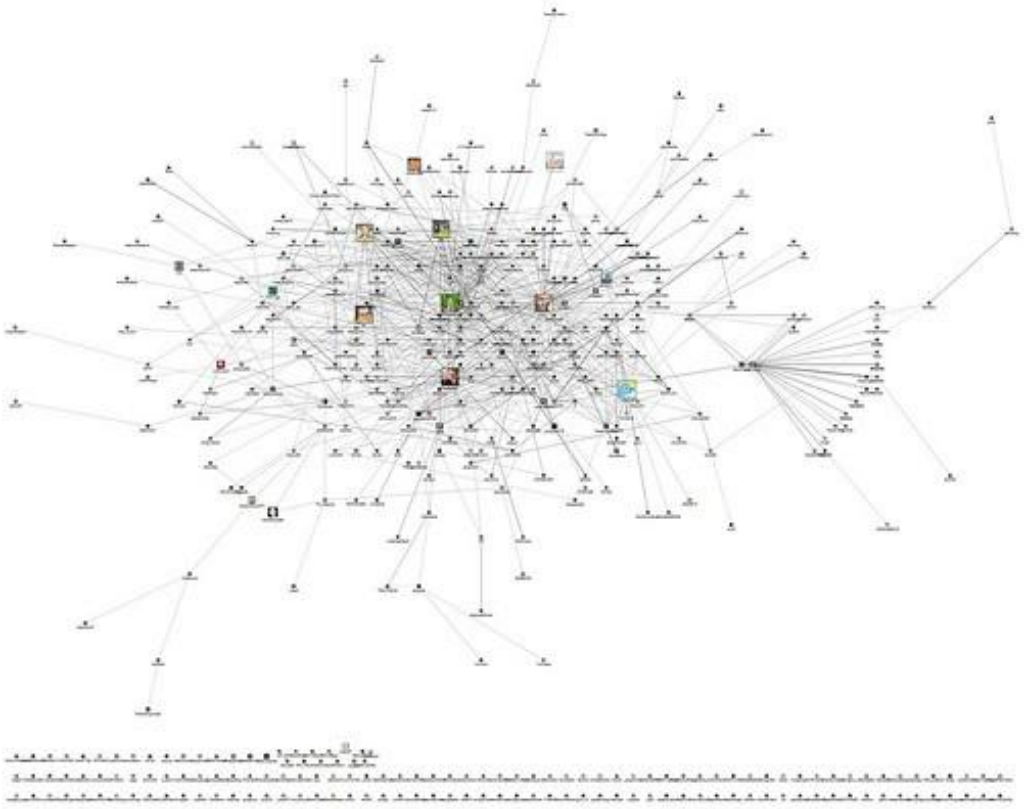






With acknowledgments to Jeffrey Heer, *Exploring Enron*, http://hci.stanford.edu/jheer/projects/enron/, Adam Perer, *Contrasting Portraits*, http://hcil.cs.umd.edu/trs/2006-08/2006-08.pdf, and Fernanda Viegas, *Email Conversations*, http://fernandaviegas.com/email.html

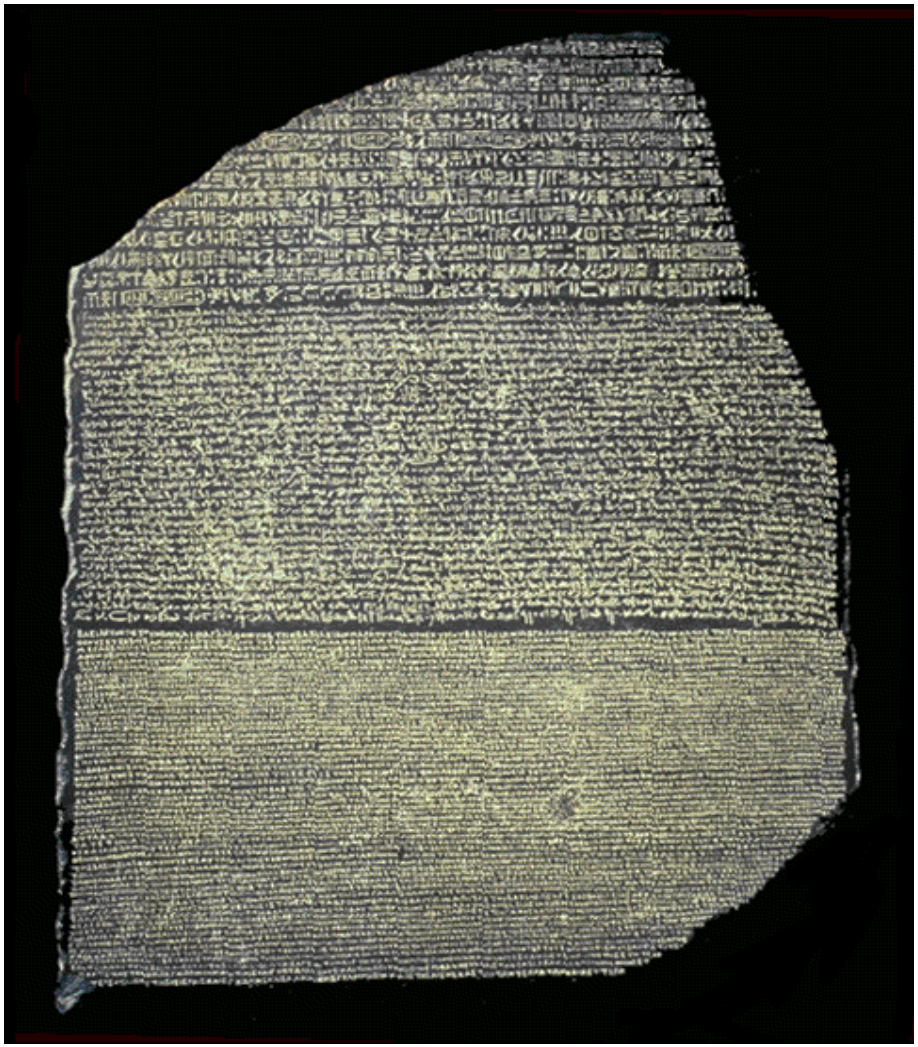
Social Networking/Links Analysis Example



From Marc Smith
Posted on Flickr
Under Creative Commons License

Interdisciplinary Approaches--

Three Languages: Legal, RM, and IT



References

- The Sedona Conference, ***Best Practices Commentary on The Use of Search & Retrieval Methods in E-Discovery*** **(2007)**
  www.thesedonaconference.org
- The Sedona Conference, ***Achieving Quality in E-Discovery*** **(2009)**
  www.thesedonaconference.org
- **TREC 2011 Legal Track Home Page**
  http://trec-legal.umiacs.umd.edu/

References

- J. Baron, “Law in the Age of Exabytes: Some Further Thoughts on ‘Information Inflation’ and Current Issues in E-Discovery Search, 17 *Richmond J. Law & Technology* (2011), see http://law.richmond.edu
- J. Baron, “Information Inflation: Can The Legal System Adapt?” (with co-author George L. Paul), 13 *Richmond J. Law & Technology* (2007), vol. 3, article 10, available at http://law.richmond.edu/jolt/index.asp
- D. Oard, J. Baron, B. Hedin, D. Lewis, S. Tomlinson, “Evaluation of Information Retrieval for E-Discovery,” *Artificial Intelligence and Law* (2010)