E-Discovery:
A Challenge for Search

ICIC 2009
Barcelona (Sitges), Spain

David A. Evans

JustSystems Evans Research, Inc.

October 20, 2009

JSE-PR-09-04

Abstract

Corporations increasingly use and retain information only in the form of electronically held data and documents. As a result, the production and sharing of information in legal proceedings will depend heavily on techniques for accessing, searching, organizing and analyzing electronic data – the principal focus of E-Discovery. Large corporations may have terabytes of e-mail and other files spanning many years that are potentially relevant to a case. In response to a court order, an E-Discovery team must identify, assemble, individuate and categorize an organization's files, segregate all "privileged" material (which may be withheld legally), and deliver a minimally comprehensive and exhaustive set of data to the opposing party – all in a relatively short amount of time. The techniques needed to accomplish such a task necessarily include search, clustering, classification, filtering, social network analysis, extraction, and more – and no one of these is sufficient. Such requirements challenge our traditional models for search. In particular, the appropriate user models do not reflect the standard "web" or "enterprise" conditions. This presentation explicates the requirements and types of solutions that dominate E-Discovery.

Background on E-Discovery

What is E-Discovery?

Briefly…

- **The requirement to provide to a party in an official investigation or lawsuit (court case) documents that exist in electronic form.**
- **Electronic documents include e-mail, text messages, electronic calendars, voicemail, audio files, graphics, photographs, drawings, spreadsheets, CAD files, metadata, animations, files on portable devices and storage media, digital data, etc.**
- **Note: E-Discovery currently supplements – but is gradually replacing – traditional discovery of "paper" materials. A great deal of E-Discovery practice is grounded in the experiences and expectations of people who are steeped in traditional paper-based (and manual) document discovery.**

Why Care about E-Discovery?

It's a Growing and Expensive Problem…

- **2009 Market size projected to be $4B…**
  [Source: Socha-Gelbmann Electronic Discovery Survey Public Report, 2007]
- **Expected 35% Annual Growth through 2011**
  [Source: Gartner MarketScope for E-Discovery and Litigation Support Vendors, 2007]
- **For the Enterprise…**
  - **High Cost of Compliance**
    (Far more expensive than the purchase of a large software system, typically)
  - **Very High Cost of Failure**
    (On the order of a small acquisition)

Cases Typically Involving E-Discovery

For U.S. and Foreign Enterprises Doing Business in the U.S.

- **Environmental Protection / Violations**
- **Pharmaceutical (Drug) & General Product Liability**
- **Infringement**
- **Antitrust**
- **Fraud**
- **Shareholder Actions**
- **Financial (Securities) Violations**
- **…**

Typical Challenges in E-Discovery

- Lots of data (order of terabytes)
- Little of actual value
- Short amounts of time for processing
- Importance of manual review
- Redundancy; near-redundancy; faux-redundancy
- Embeddings
- Encodings
- Co-mingling / Co-occurrence of data

Typical Goals of Processing / Analysis

- **Enumeration / Individuation of Items**
- **Determining & Defending Information Status – Privileged vs. non-Privileged**
- **Identifying Individuals (and Documents) that should be Involved in Depositions**
- **Establishing a Chain of Custody / Possession or Knowledge of Events at Points in Time**
- **…**

Some Framing Issues in E-Discovery

Distinctions from (Web) Search Practice

- Emphasis on Recall (avoiding false negatives; insuring exhaustive coverage)
- Human-in-the-Loop Processing – from initial formulation of the problem to evaluation (review) of the results
- Absence of Reference Data (uniqueness of circumstances in each case)
- No Standard Practice (including no standard evaluation metrics that translate into success in practical cases)

Some Framing Issues in E-Discovery

Distinctions from (Web) Search Practice, continued

- **Extreme Importance of Context – Social Network / Social Communications; Time; Replication (Protected vs. Public); Status of Agents; Status of Knowledge (before or after critical event); etc.**
- **Heterogeneous Information Typology, where one encounters text and non-text intimately intertwined and related; structured and non-structured**
- **Multi-Language Data (in every sense)**
- **Importance of Non-Textual Information**

Focus in E-Discovery

Distinctions from (Web / Enterprise) Search Practice

- **Compliance (exhaustive accountability)**
- **Argumentation (serving a forensic purpose; information that fits into a narrative)**
- **Evidence (information whose interpretation is determined by the circumstances of its discovery; contrast with alternative information)**
- **Explanation (not retrieval; not simple Q-A)**

What Do We Do When We Do E-Discovery?

The Work of (E-)Discovery

- We collect documents and other records
- We try to establish for a moment of time (statically) the information that will be potentially relevant, including
  - The subject matter / material of the investigation
  - The sources (locations, scope, etc.) of the material
  - The people (including certain roles) of interest
- We analyze the material
- We produce lists & items of interest
- (We support a legal team…)

The Work of (E-)Discovery, continued

- We explain (defend) what we have done to a court, including how we went about searching for and sorting out relevant documents.
- The court may rule on the (in)adequacy of our efforts…

- Traditionally, the safe bet is Boolean!

The Boolean “Standard”

- **Why Boolean?**
  - **It’s what everybody has used in the past**
  - **“Precision”**
  - **We can account for the results…**

- **And yet…**
  - **Performance in Precision ⇒ *Cost of Review***
  - **Performance in Recall ⇒ *Cost of Compliance***

The Challenge of Strict Compliance…

Judgment of Documents

- **Relevant**
  - **Hot**
    - ⇒ **"Smoking Gun"**
  - **Highly Relevant**
    - ⇒ **Deposition**
  - **Responsive**
    - ⇒ **Within Scope of Order**
- **Not Relevant**
    - ⇒ **Suggests Thoroughness**
- **Inaccurate**
- **Unknown / Unavailable**

The Court only wants to see Relevant Documents

Illustration by Charles Robinson from Carroll 1906, Front'spiece.

From: Evans, D.A. "E-Discovery: A Signature Problem for Search." Search Engine Meeting 2009.

New Directions from the Courts

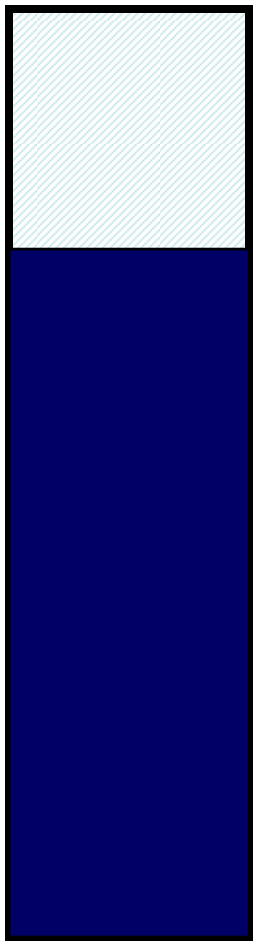
- **Magistrate Judge John Facciola**
  U.S. District Court, District of Columbia
  - **U.S. v. O'Keefe, 2008**
  - **Equity Analytics v. Lundin, March 7, 2008**
- **U.S. Chief Magistrate Judge Paul W. Grimm**
  U.S. District Court, District of Maryland
  - **Victor Stanley, Inc. v. Creative Pipe, Inc., 2008**
- **Search should be reasonable and appropriate to the task (and defensible!).**
- **Keyword-based search may not be sufficient.**
- **We know a lot about effective search (cf. TREC Legal Track) – consult experts!**

How Good / Bad is Search?
(In General? For E-Discovery?)

What We Knew in 1997

What Contributes to Accuracy?

Good Queries (10+ Terms) / 1M Documents



5 "False–Positives"

4 "Feedback" **Interactive**

2 Phrases ("Linguistics")

1 "Passages"

3 "Weighting" **"Advanced"**

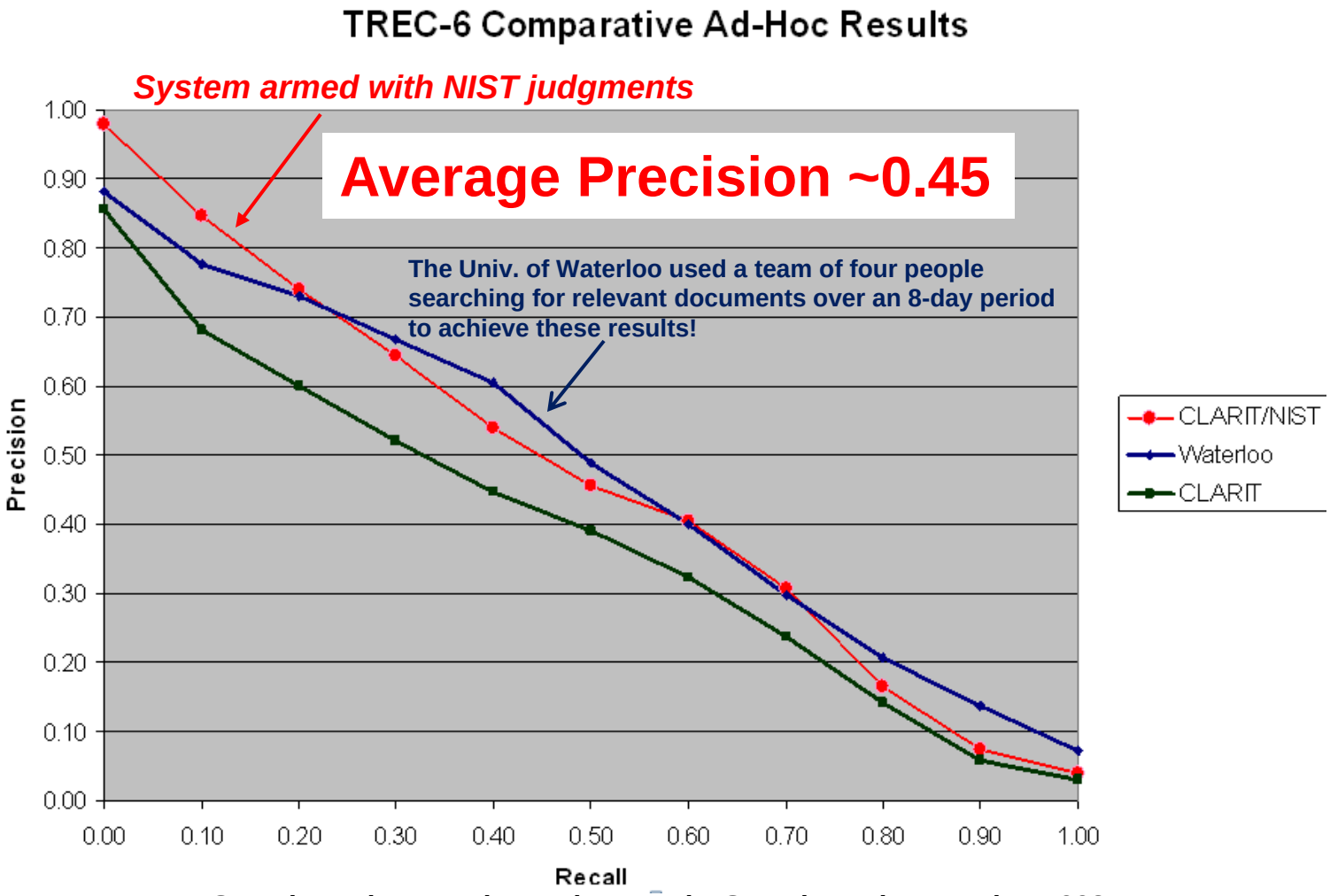
2 Combinations/Constraints

2 "Stemming"/Normalization

1 "Word" Match **Boolean**

Adapted from Evans, D.A. "Search Engines: Today and Beyond." Search Engine Meeting 1998.

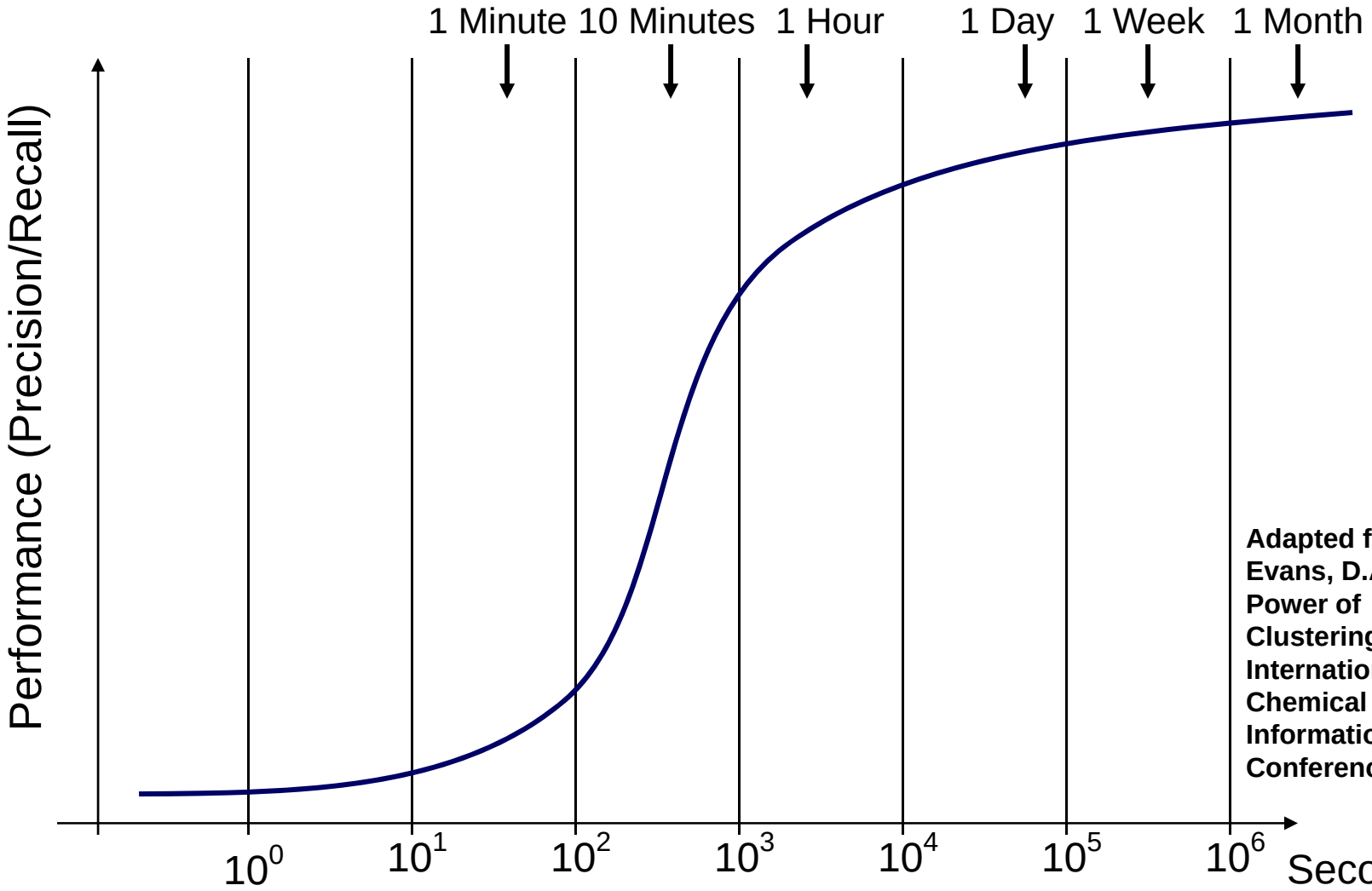
What is the Limit? (1997)



Adapted from Evans, D.A. "Search Engines: Today and Beyond." Search Engine Meeting 1998. (Cf. Milic-Frayling, et al. 1998; Cormack et al. 1998.)

How Much 'Quality' is Possible?

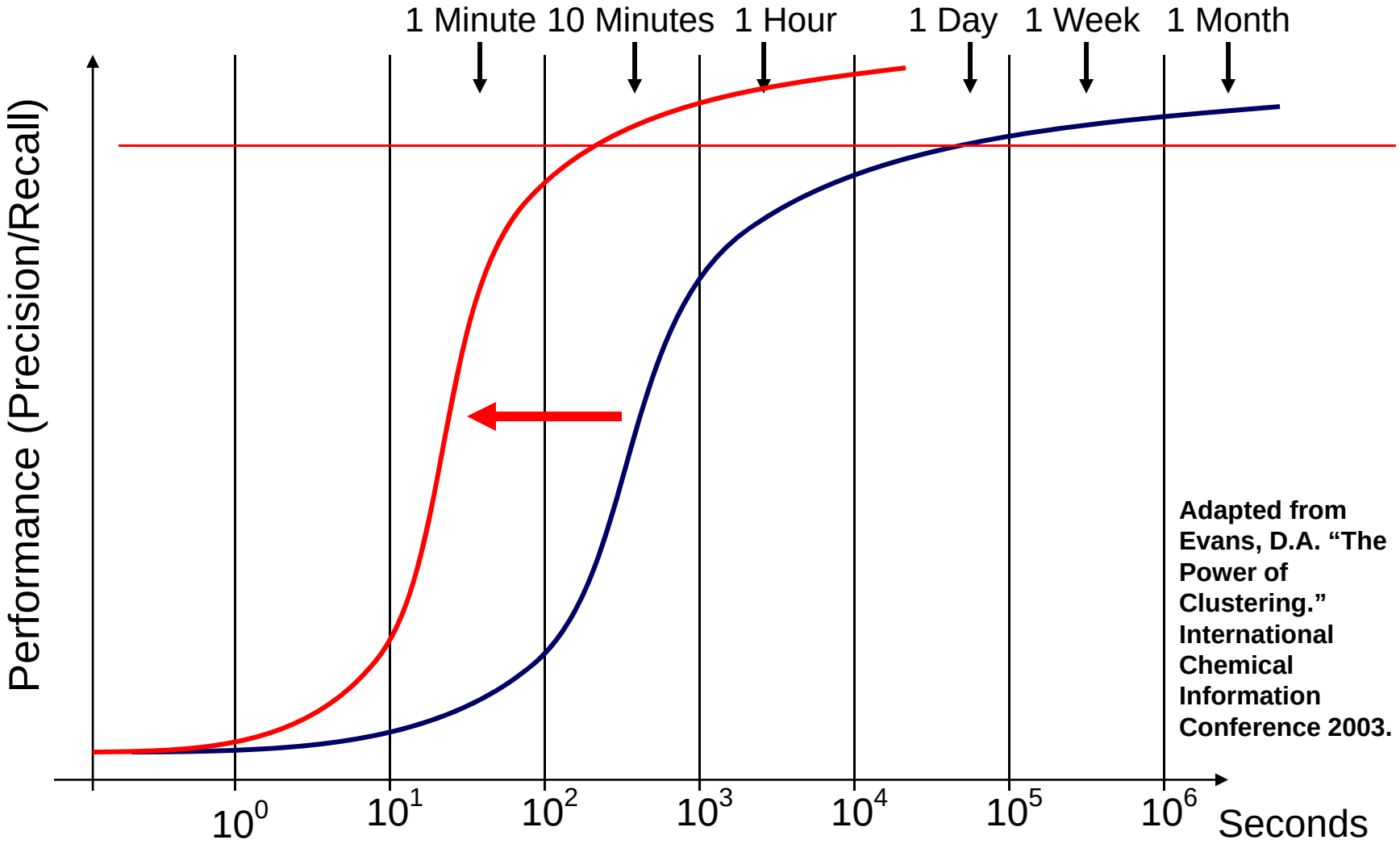
Human Effort vs. Retrieval Performance



Adapted from Evans, D.A. "The Power of Clustering." International Chemical Information Conference 2003.

How Much 'Quality' is Possible?

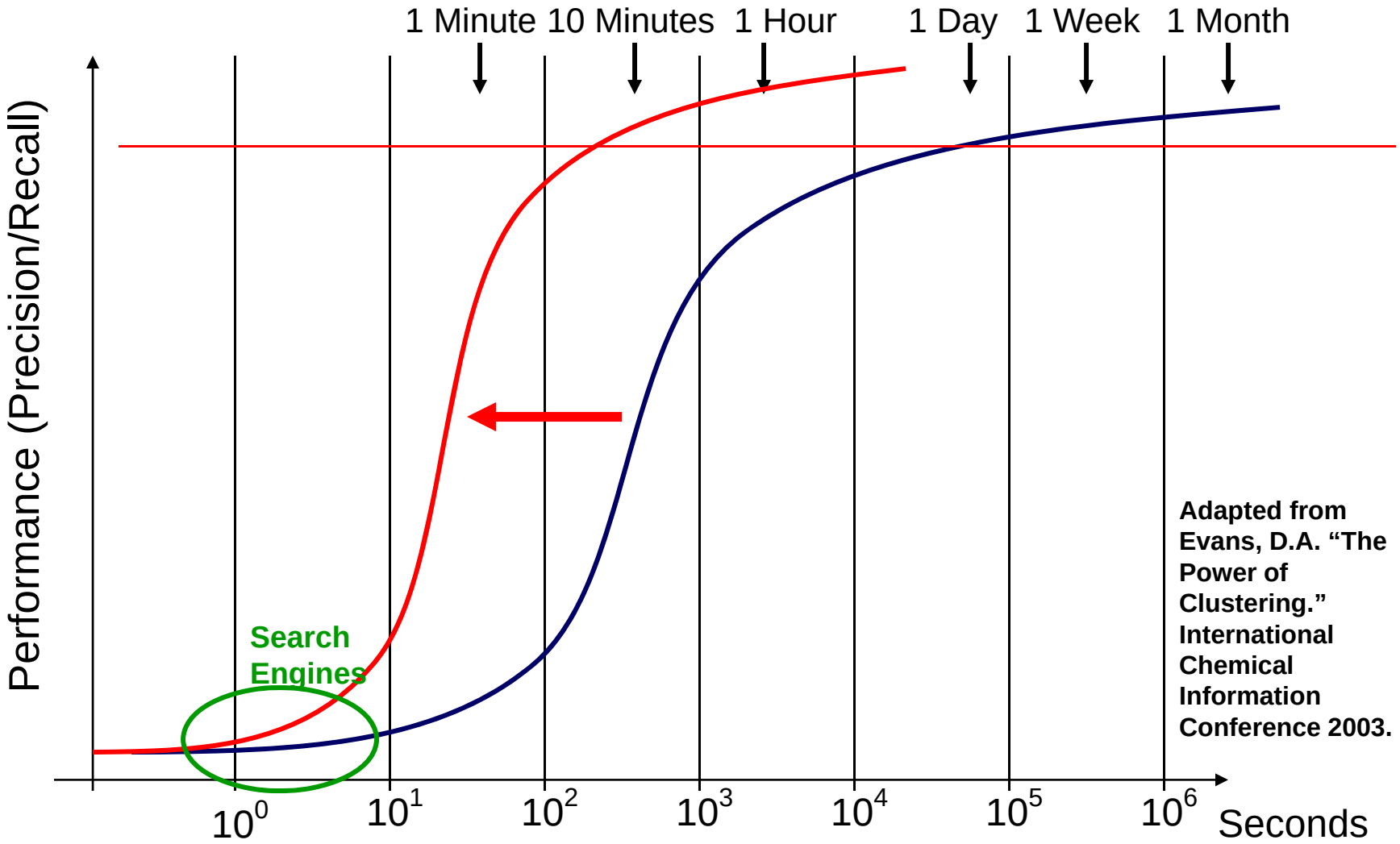
Human Effort vs. Retrieval Performance



Adapted from Evans, D.A. "The Power of Clustering." International Chemical Information Conference 2003.

How Much 'Quality' is Possible?

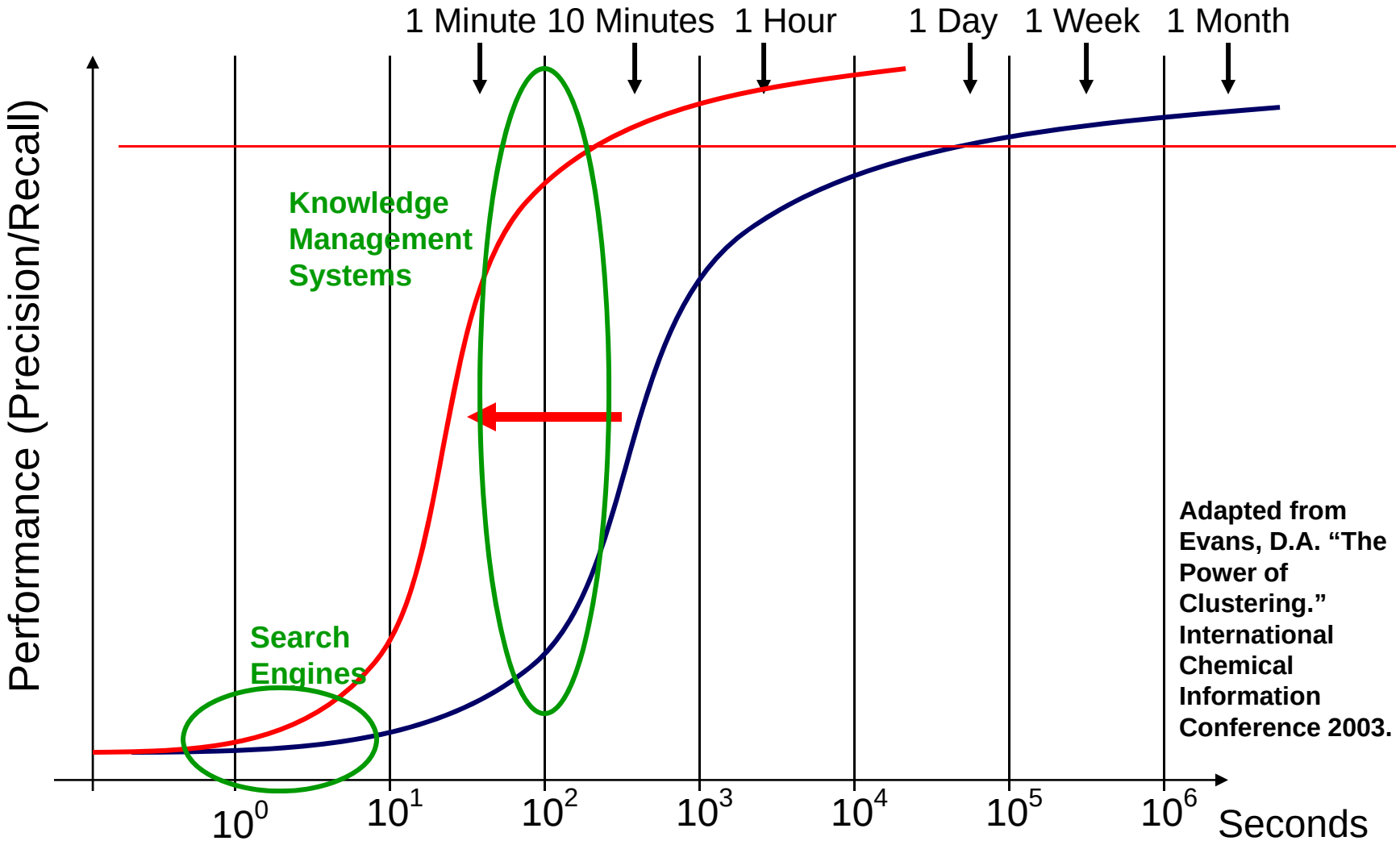
Human Effort vs. Retrieval Performance



Adapted from Evans, D.A. "The Power of Clustering." International Chemical Information Conference 2003.

How Much 'Quality' is Possible?

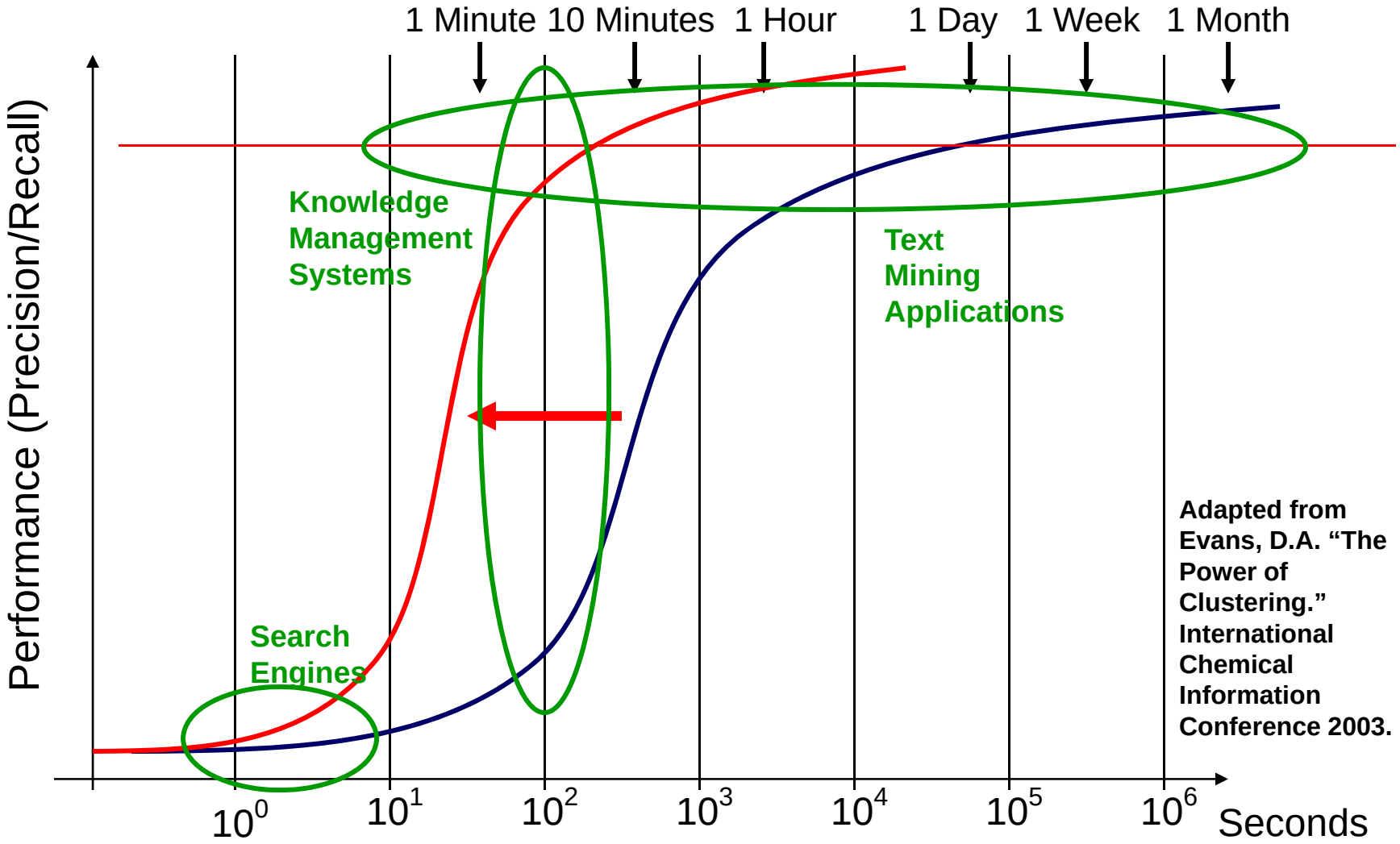
Human Effort vs. Retrieval Performance



Adapted from Evans, D.A. "The Power of Clustering." International Chemical Information Conference 2003.

How Much 'Quality' is Possible?

Human Effort vs. Retrieval Performance



What We Know about Search in E-Discovery Today (TREC Legal Track)

Sample “Production Request” (TREC)

<ProductionRequest>

<RequestNumber>82</RequestNumber>

<RequestText>All documents discussing the color of the paper used to make cigarettes in connection with increasing sales.</RequestText>

<BooleanQuery>

<FinalQuery>((color! OR shade! OR pastel! OR tint!) w/5 paper) AND (increas! w/15 (sale! OR sell!))</FinalQuery>

<NegotiationHistory>

<ProposalByDefendant>(color! w/2 paper) AND (increas! w/3 sales)</ProposalByDefendant>

<RejoinderByPlaintiff>(color! OR shade! OR pastel! OR tint!) AND paper AND (sale! OR sell!)</RejoinderByPlaintiff>

</NegotiationHistory>

</BooleanQuery>

<FinalB>888</FinalB>

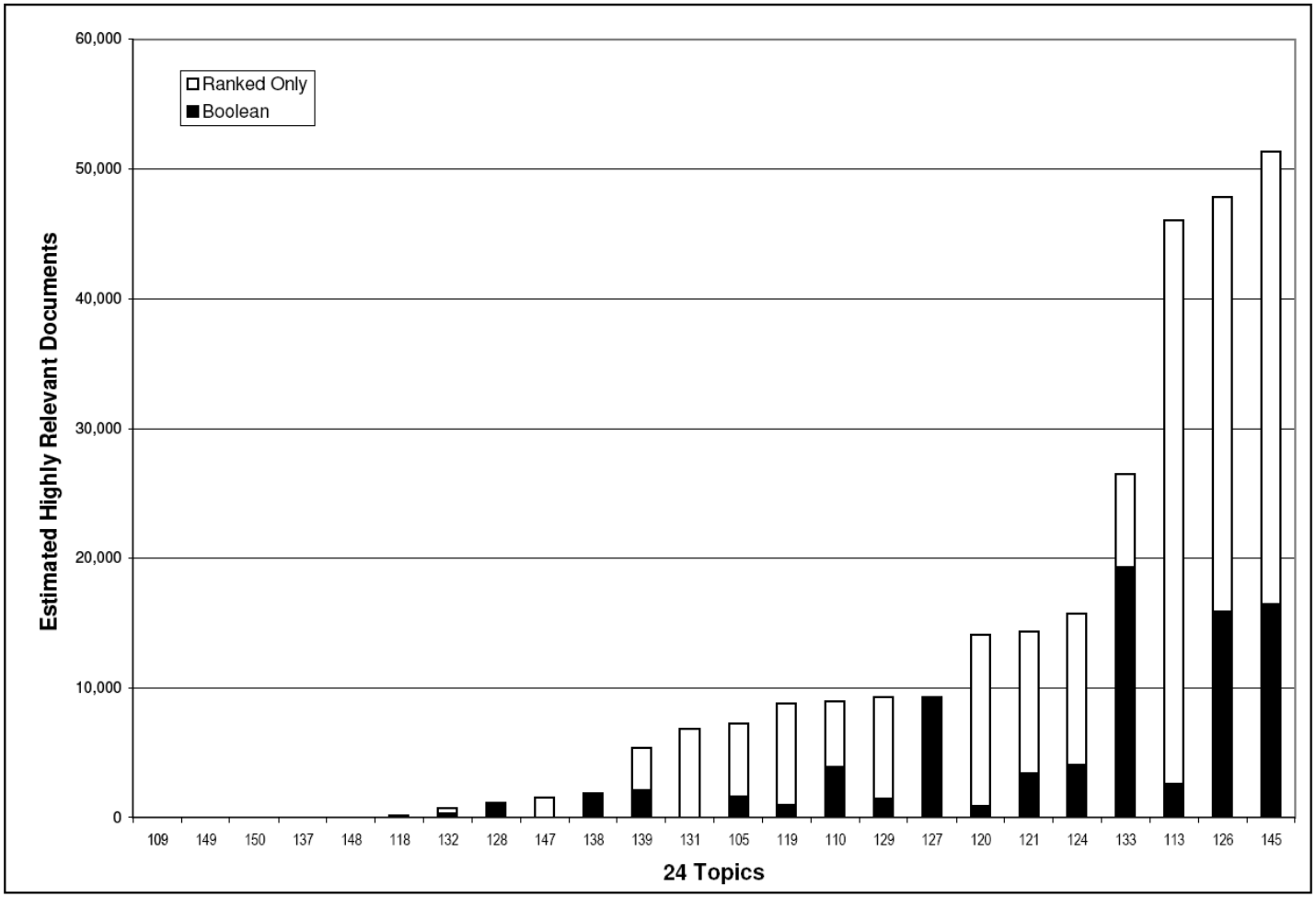
<RequestSource>2007-C-4</RequestSource>

<Instruction>

<P>1. These requests require the production of all responsive documents within the sole or joint possession, custody or control of the Defendants, including their agents, departments, attorneys, directors, officers, employees, consultants, investigators, insurance companies, or other persons subject to Defendants' custody or control.</P>

<P>2. All documents that respond, in whole or in part, to any portion of these Requests must be produced in their entirety, including all attachments and enclosures.</P> …

The Boolean Shortfall…



Highly Relevant Documents not Found by the Consensus Boolean Run
From: Oard et al. 2009

Retrieval Results

All Relevant (26 topics)	Retrieved	Precision	Recall	F_1
Defendant	3,180	0.41	0.04	0.05
Plaintiff	219,606	0.23	0.43	0.19
Consensus1	93,190	0.24	0.33	0.20
Final	40,402	0.28	0.24	0.16
	Avg. K			
Median (23 request runs)	14,363	0.26	0.12	0.10
Median (41 other runs)	40,402	0.28	0.25	0.16
Highly Relevant only	(24 topics)			
Defendant	3,445	0.14	0.06	0.06
Plaintiff	234,016	0.08	0.57	0.09
Consensus1	97,259	0.07	0.42	0.09
Final	39,930	0.08	0.33	0.09
	Avg. K_h			
Median (23 request runs)	5,838	0.10	0.22	0.05
Median (41 other runs)	19,965	0.09	0.34	0.08

$F_1 \approx 2\ R\ P\ /\ (R+P)$

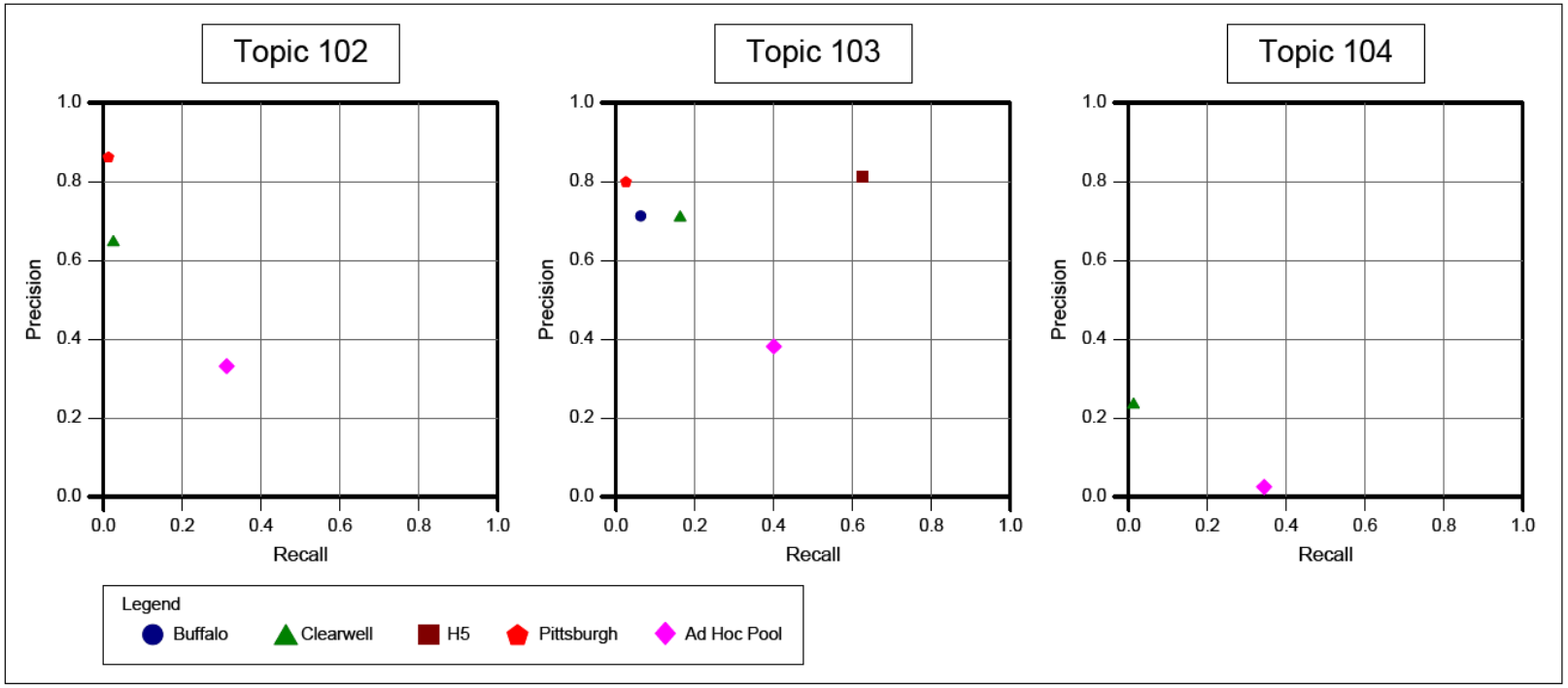
TREC 2008

Mean Scores of the Negotiated Boolean Queries and Median Mean Scores of the Participant Runs

From: Oard et al. 2009

Interactive Task

TREC 2008

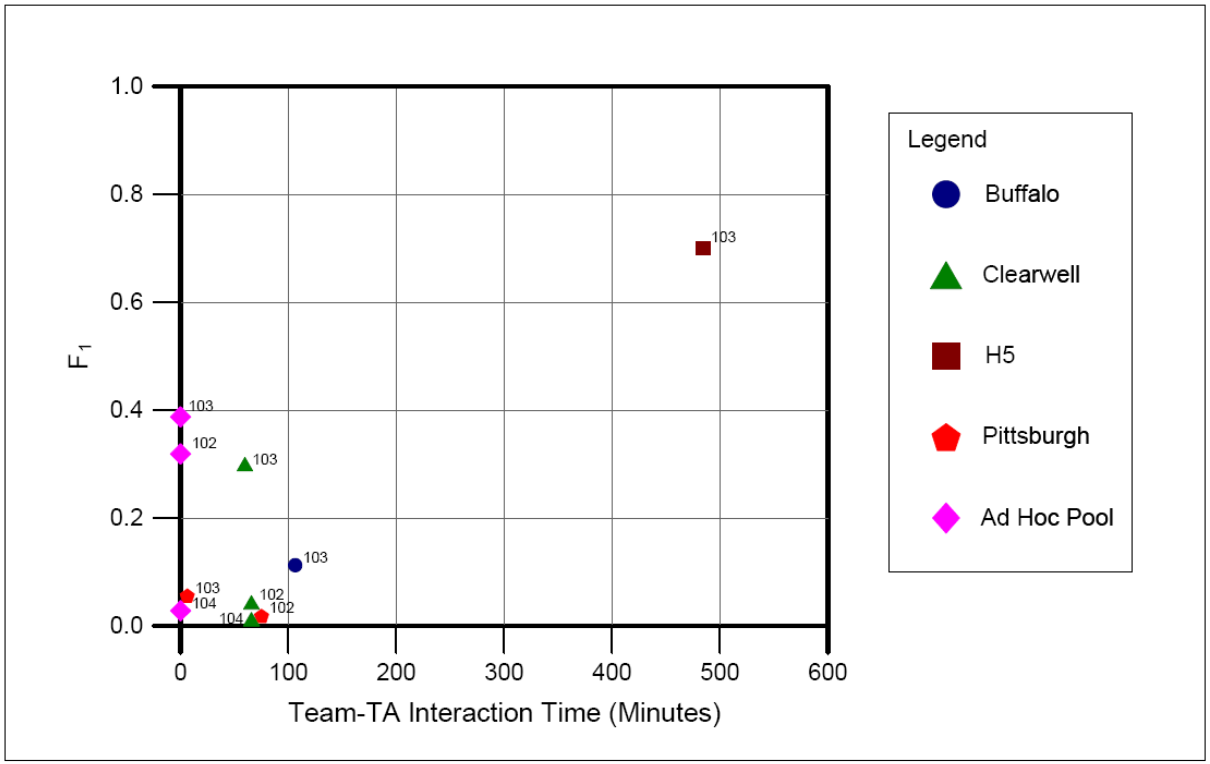


Recall & Precision – Post Adjudication

From: Oard et al. 2009

Interactive Task

TREC 2008



F_1 vs. Team-TA Interaction Time
From: Oard et al. 2009

What if Better Search is *not* the Answer?

What Do We Do When We Do E-Discovery? (Again)

Example Case – Mostly, Not Search

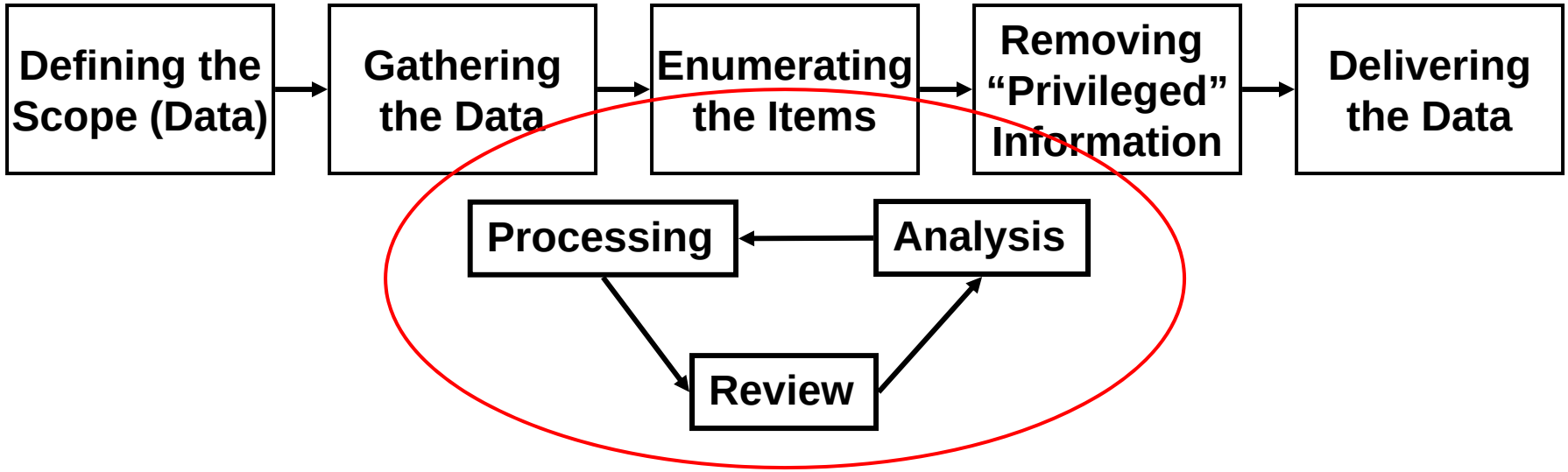
Example Case

Sorry, No Subject-Matter Details!

- **Multinational Corporation (Defendant)**
- **Source Data: Approximately 300 GB of E-Mail**
- **Number of Directories / “People”: ~5,000**
- **Number of files: ~1,000,000**
- **Many Attachments, Compressed Archives**
- **Text, Data, Multiple Human Languages**
- **Rate of Redundancy / Duplication: ~50%**
- **Rate of Errors in Individuation: ~10%**

Ingredients of a Solution

Process Flow and Techniques…



- code normalization
- unzipping compressed data
- language ID
- lexical-atom discovery
- NLP (multi-lingual)
- term EQ-class discovery
- person identification

- indexing (term/feature selection)
- duplicate/near-duplicate ID
- enumeration/individuation
- cross-linking related items
- social network analysis
- clustering (for topic threads)

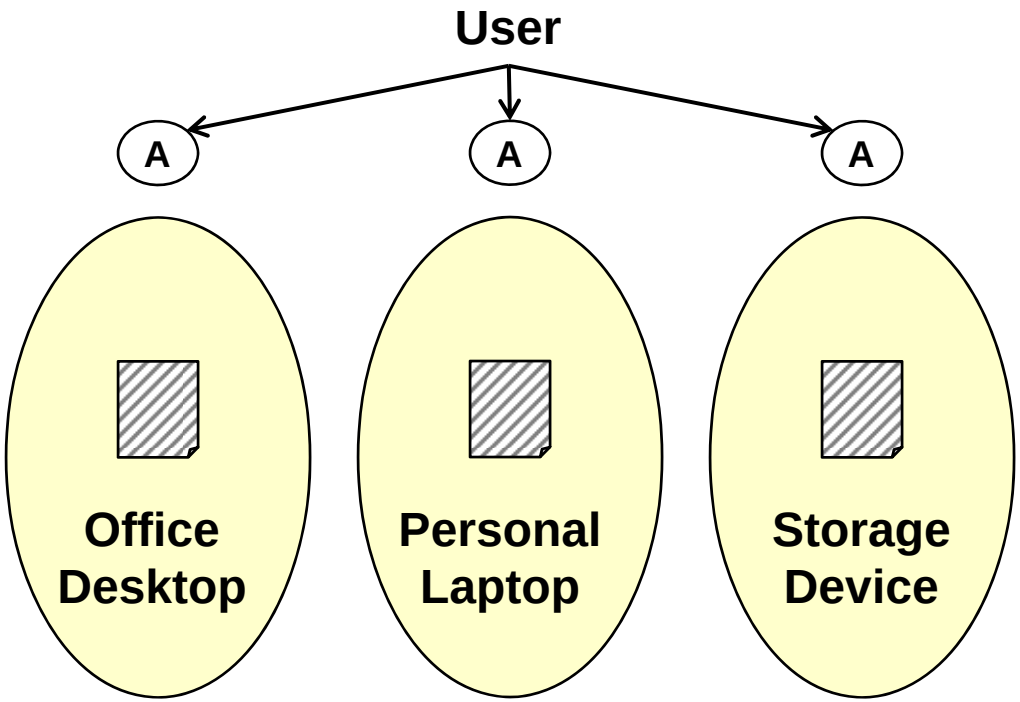
- filtering
- classification (P/~P)
- topic mapping
- time series analysis
- pseudo-causal modeling

Adapted from: Evans, D.A. "Why E-Discovery is a CIKM-Hard Problem." ACM CIKM 2008.

Illustrations of Typical Problems

Counting Instances of a Document

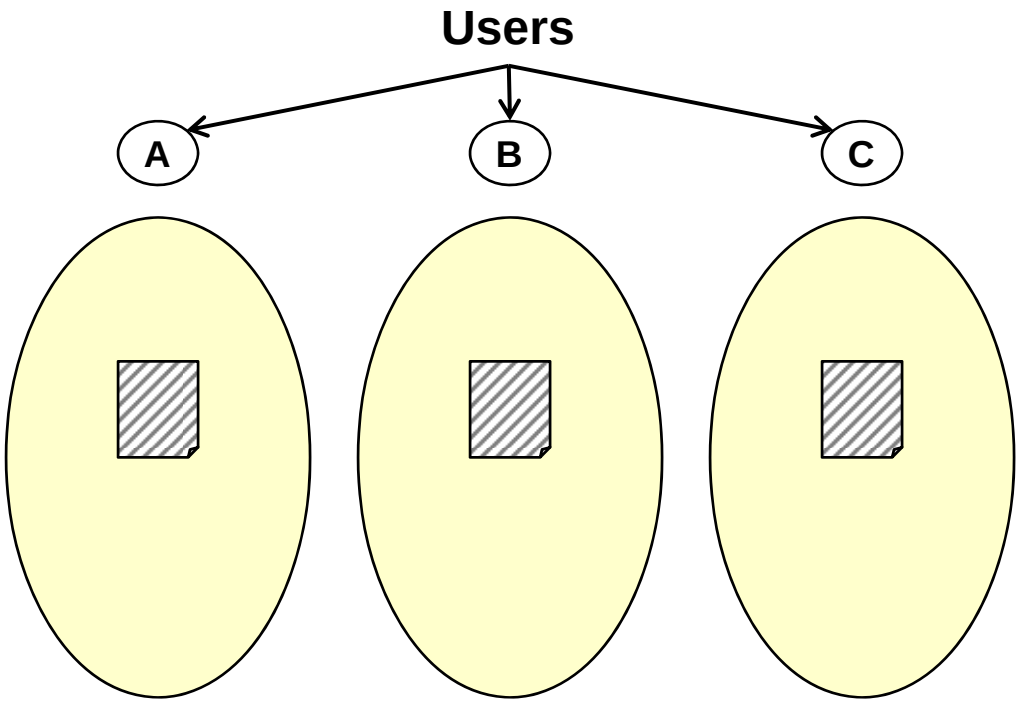
"Same" Document, Different Locations, Same User, But Co-Mingled With Different Other Documents



One Instance? Or Three?

Counting Instances of a Document

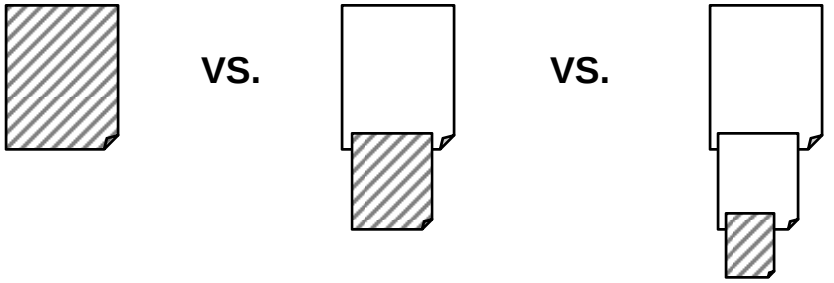
"Same" Document, Different Locations, Different Users, Co-Mingled With Different Other Documents



One Instance? Or Three?

Counting Instances of a Document

"Same" Document, Different Contexts



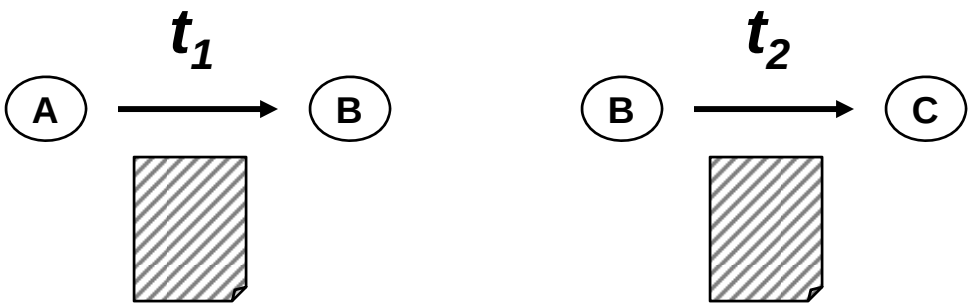
Embedded Attachment

Embedded Attachment

One Instance? Or Three?

Context Matters

Privileged Communication at time t_1

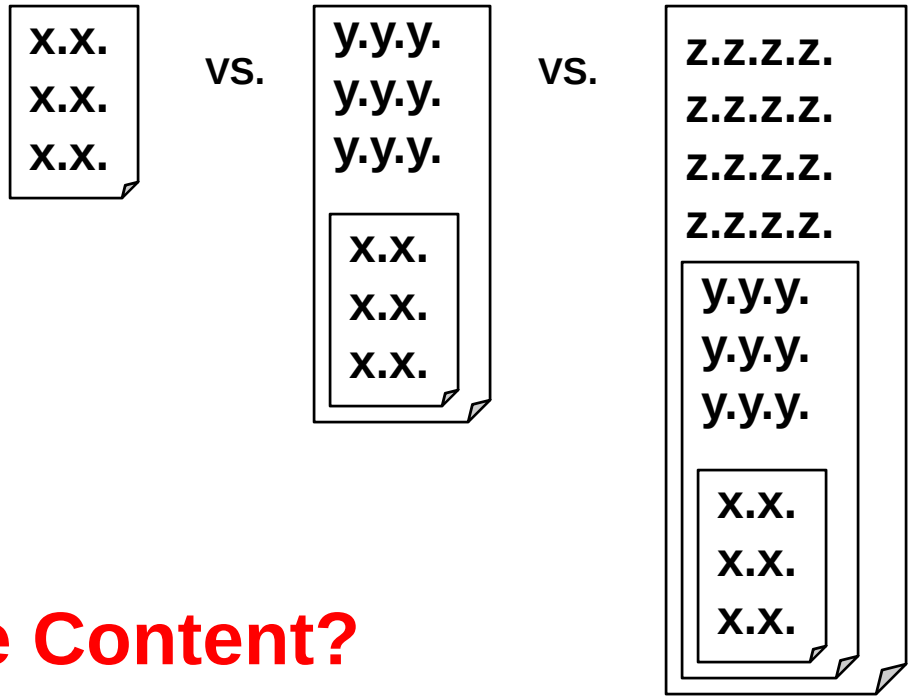


Still Privileged Communication at time t_2 ?

Document Status is not Transitive…

Context Matters

Threads, Embedded Text in E-Mail



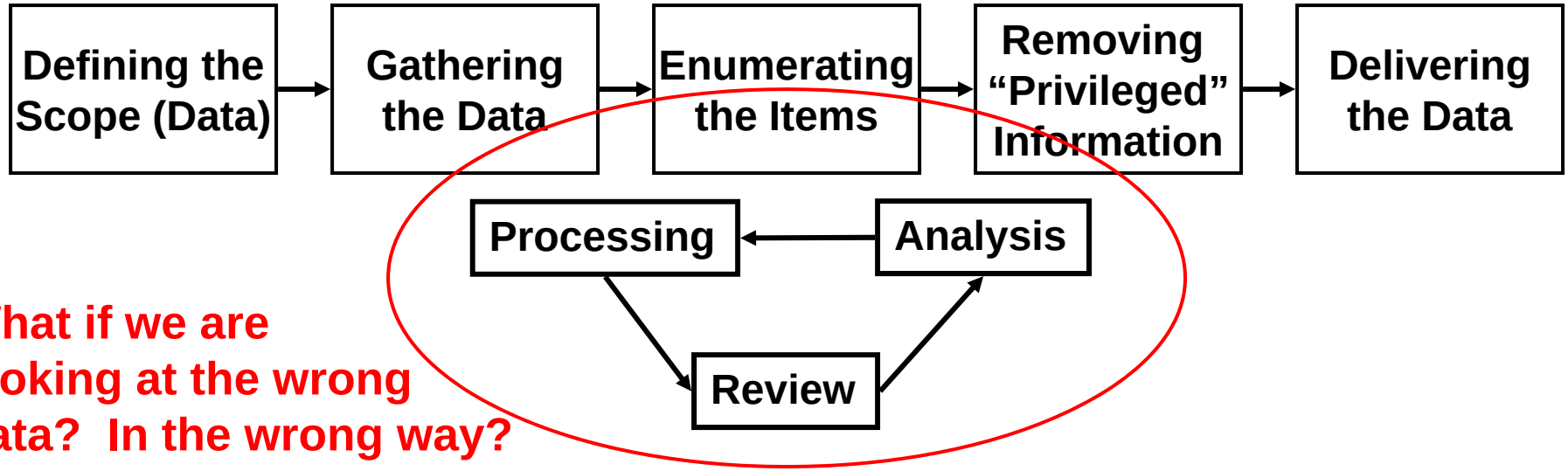
What is the Content?
What is the Comment?
What is the Role of Sender? Receiver? CC? BCC?

In Practice, Typically…

- **Questionable Recall**
- **Poor Enumeration of Items (and only that – no other context)**
- **Incomplete Classification of "Privileged"**
- **…**
- **No Attempt to Analyze Foreign-Language Material, except quite superficially**
- **No (or very little) Attempt to Accommodate Cultural Differences in Work and Work Products**

Ingredients of a Solution

Process Flow and Techniques…



What if we are looking at the wrong data? In the wrong way?

- code normalization
- unzipping compressed data
- language ID
- lexical-atom discovery
- NLP (multi-lingual)
- term EQ-class discovery
- person identification
- indexing (term/feature selection)
- duplicate/near-duplicate ID
- enumeration/individuation
- cross-linking related items
- social network analysis
- clustering (for topic threads)
- filtering
- classification (P/~P)
- topic mapping
- time series analysis
- pseudo-causal modeling

Adapted from: Evans, D.A. "Why E-Discovery is a CIKM-Hard Problem." ACM CIKM 2008.

What Do We Do When We Work?

“Work” & Its Artifacts

Meetings (Face-to-Face)
Videoconferences
Teleconferences
Presentations / Lectures
Conferences
Visits & “Occasions”
Supervision
Collaboration
Consultation
Conversation

Reports
Spreadsheets
Graphics
Design Documents
Forms
E-Mail
Images
Timesheets/Logs
Text Messages

Work Products in Multinational Orgs

Problems Legal Teams Have with E-Discovery Material

- **Non-English Texts**
  - **“We asked for the <<foreign-language>> stuff, but we could never get anything out of it.”**
- **Non-Useful / Unusual Document Types**
  - **“All we got was some e-mail and a huge amount of repetitious forms.”**
- **Different Boundaries for Trusted Communication**
  - **“We tried to keep all the privileged stuff out, but there were copies all over the place.”**

Cultural & Behavioral Dimensions

Culture & Behavior in Interactions

- **In general…**
  - **We internalize culture**
  - **We externalize behavior**
- **Culture is the "software" that encodes our sensibilities, values, expectations, default orientation and attitudes towards people in degrees of relationship to us, …**
- **Behavior is the "application" running in social contexts, through which we manifest degrees of trust, sharing, following, leading, accepting responsibility, …**

Characterization of Cultural Dimensions

Geert Hofstede, *Cultures and Organizations*, 1991

- **Power Distance**
  - How a society handles inequalities – "the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally."
- **Individualism / Collectivism**
  - Behavior towards the group – "Individualism pertains to societies in which the ties between individuals are loose: everyone is expected to look after himself or herself and his or her immediate family. Collectivism as its opposite pertains to societies in which people from birth onwards are integrated into strong, cohesive in groups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty."
- **Masculinity / Femininity**
  - Behavior according to gender – "Masculinity pertains to societies in which social gender roles are clearly distinct; femininity pertains to societies in which social gender roles overlap."
- **Uncertainty Avoidance**
  - The need for Structure – "the extent to which the members of a culture feel threatened by uncertain or unknown situations."
- **Long-Term Orientation**
  - "Long-term orientation" includes values such as thrift and perseverance. "Short-term orientation" includes respect for precedence and tradition, fulfilling social obligations, saving "face."

Example Differences by Country

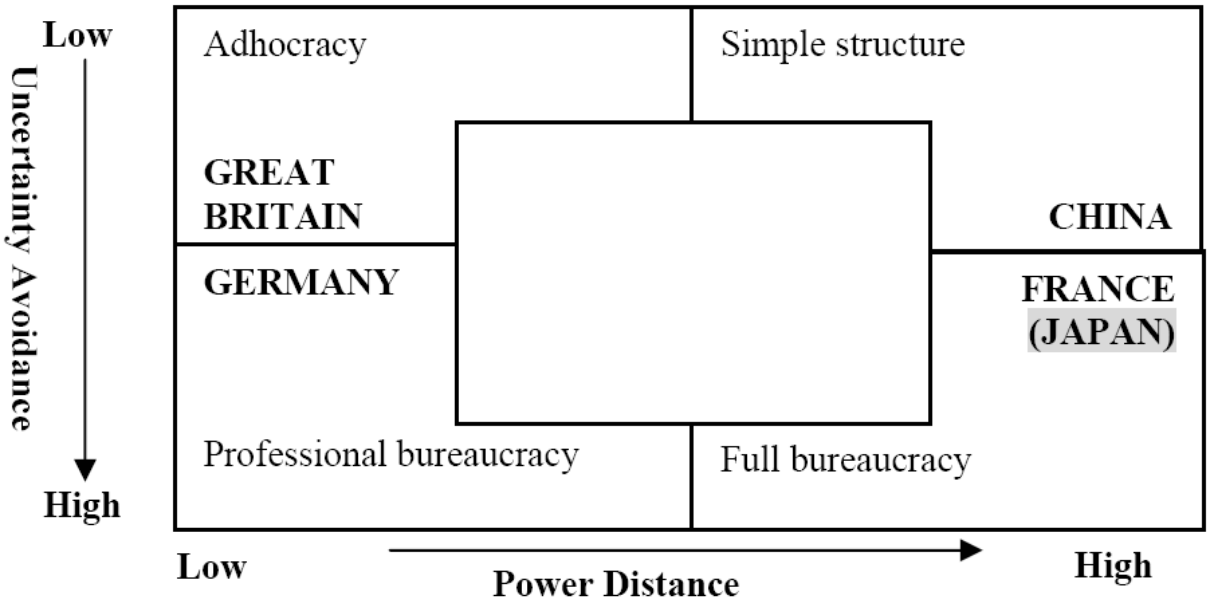


Figure 10: Preferred Coordination Mechanism (Adapted from Hofstede, 1991, p.152)

Note: This figure illustrates the typical organization structure predicted by "power distance index" and "uncertainty avoidance index." "Power distance index" refers to the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally. "Uncertainty avoidance index" indicates the extent to which a culture programs its members to feel comfortable in unstructured situations such as unknown, surprising, and different from the usual. Uncertainty-avoiding cultures try to minimize the possibility of such situations by using strict laws and rules, and safety and security measures

From Horii, Jin & Levitt 2005b

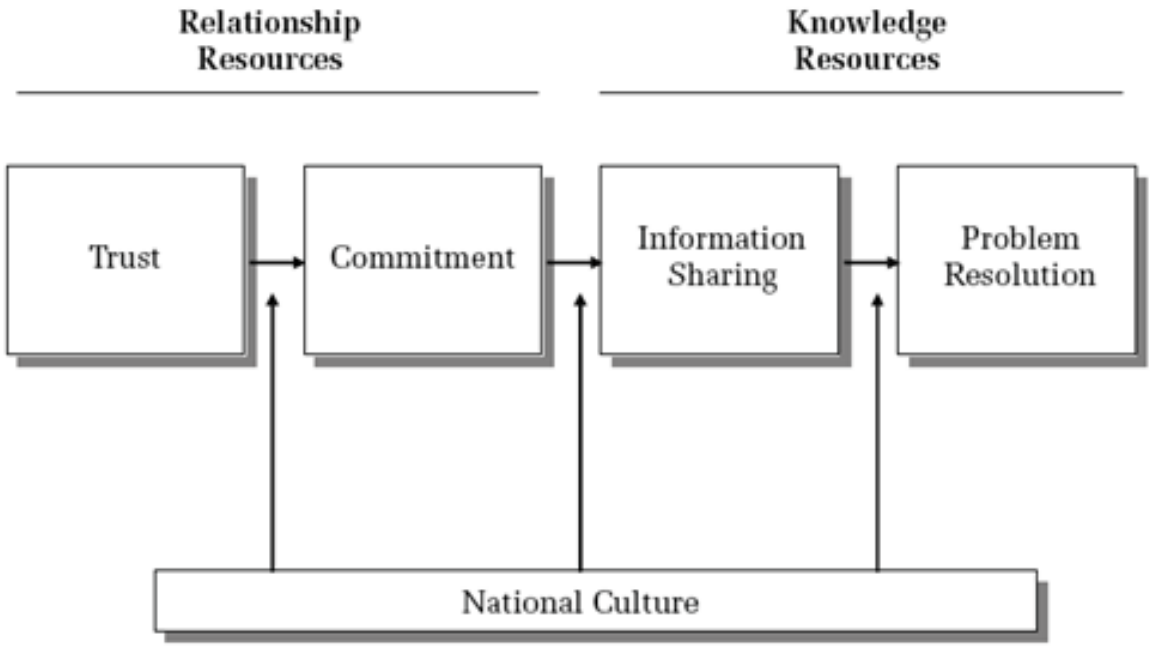
Parameterizing Work-Group Cultures

		Culture A (American)	Culture J (Japanese)
Practices	Centralization	Decentralized authority	Centralized authority
	Formalization	Medium level of formalization	High level of formalization
	Org. hierarchy	Flatter hierarchy	Multiple Hierarchy
Values	Decision Making	Individual decision making	Consensual decision making
	Communication	Individually-based	Group-based

Figure 1: Summary of Cultural Differences

From Horii, Jin & Levitt 2004

Other Cultural Influences on Behavior



Associated factors:

- **Societal Monitoring**
- **Social Exchange**
- **Individual Responsibility**

From: Griffith, Myers & Harvey 2006

Observed Differences in Behavior

H_1: Japanese firms perceive a stronger positive association between trust and commitment in intracultural, interorganizational relationships than U.S. firms.

H_2: Japanese firms perceive a stronger positive association between trust and commitment in intercultural, interorganizational relationships with U.S. firms than U.S. firms perceive with Japanese firms.

H_3: Japanese firms exhibit a stronger positive association between commitment and information sharing in intracultural, interorganizational relationships than U.S. firms.

H_4: Japanese firms exhibit a stronger positive association between commitment and information sharing in intercultural, interorganizational relationships with U.S. firms than U.S. firms exhibit with Japanese firms.

H_5: Japanese firms exhibit a weaker positive association between information sharing and problem resolution in intracultural, interorganizational relationships than U.S. firms.

H_6: Japanese firms exhibit a weaker positive association between information sharing and problem resolution in intercultural, interorganizational relationships with U.S. firms than U.S. firms exhibit with Japanese firms.

Hypotheses	Japan	United States	Assessment ($p < .05$)
H_1: Trust → commitment (intracultural, interorganizational)	.339 t = 3.39**	.042 t = 2.36*	Supported
H_2: Trust → commitment (intercultural, interorganizational)	.404 t = 3.54**	.018 t = .90	Supported
H_3: Commitment → information sharing (intracultural, interorganizational)	.592 t = 4.78**	.184 t = 2.90*	Supported
H_4: Commitment → information sharing (intercultural, interorganizational)	.611 t = 5.33**	.102 t = 2.14*	Supported
H_5: Information sharing → problem resolution (intracultural, interorganizational)	.855 t = 4.92**	.851 t = 5.43**	n.s.
H_6: Information sharing → problem resolution (intercultural, interorganizational)	.743 t = 4.71**	.980 t = 5.00**	n.s.

Stacked Model Assessment

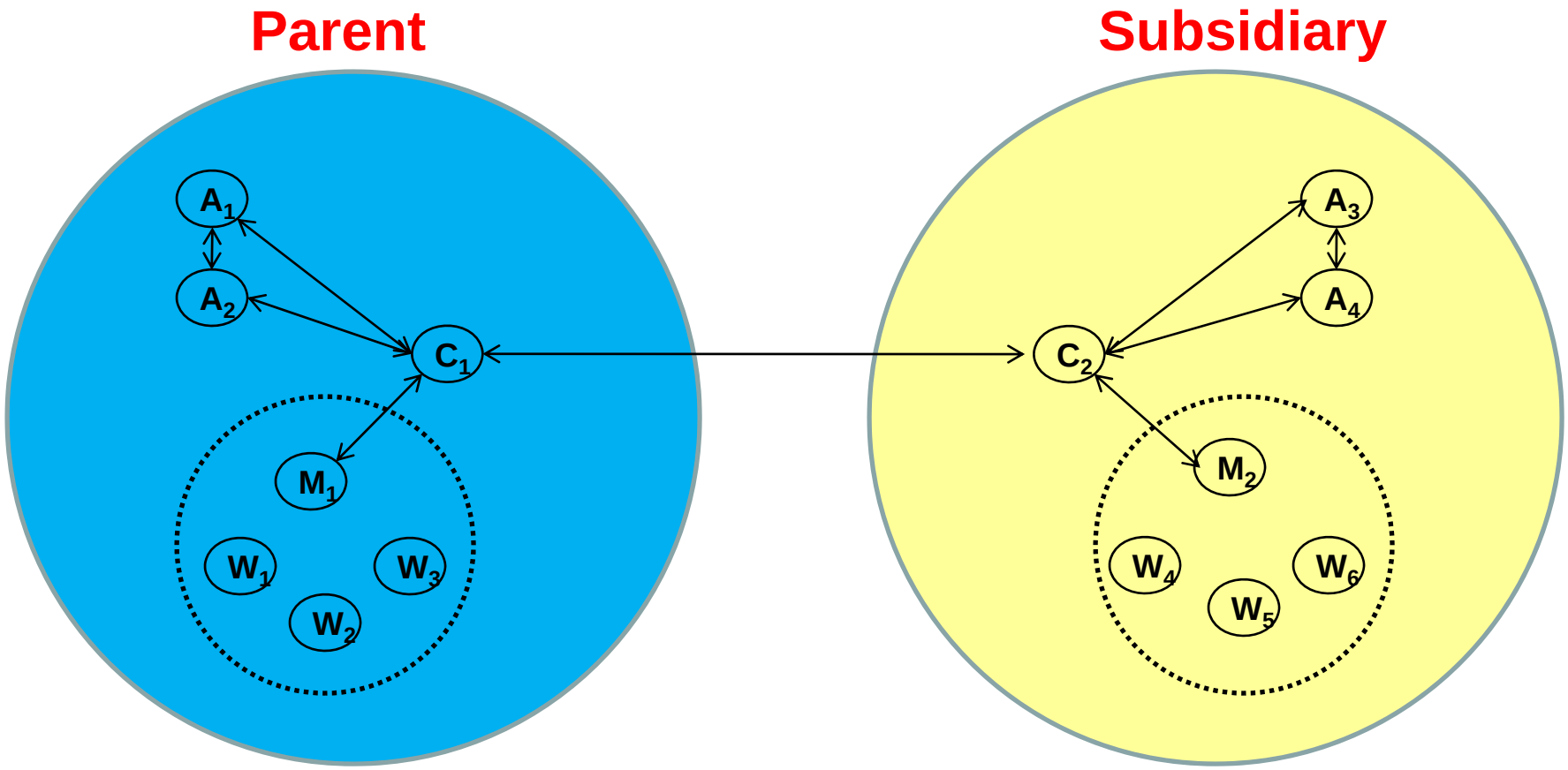
Intracultural: $\chi^2 = 515.74$, d.f. = 174; NFI = .95; CFI = .97.

Intercultural: $\chi^2 = 418.39$, d.f. = 174; NFI = .96; CFI = .97.

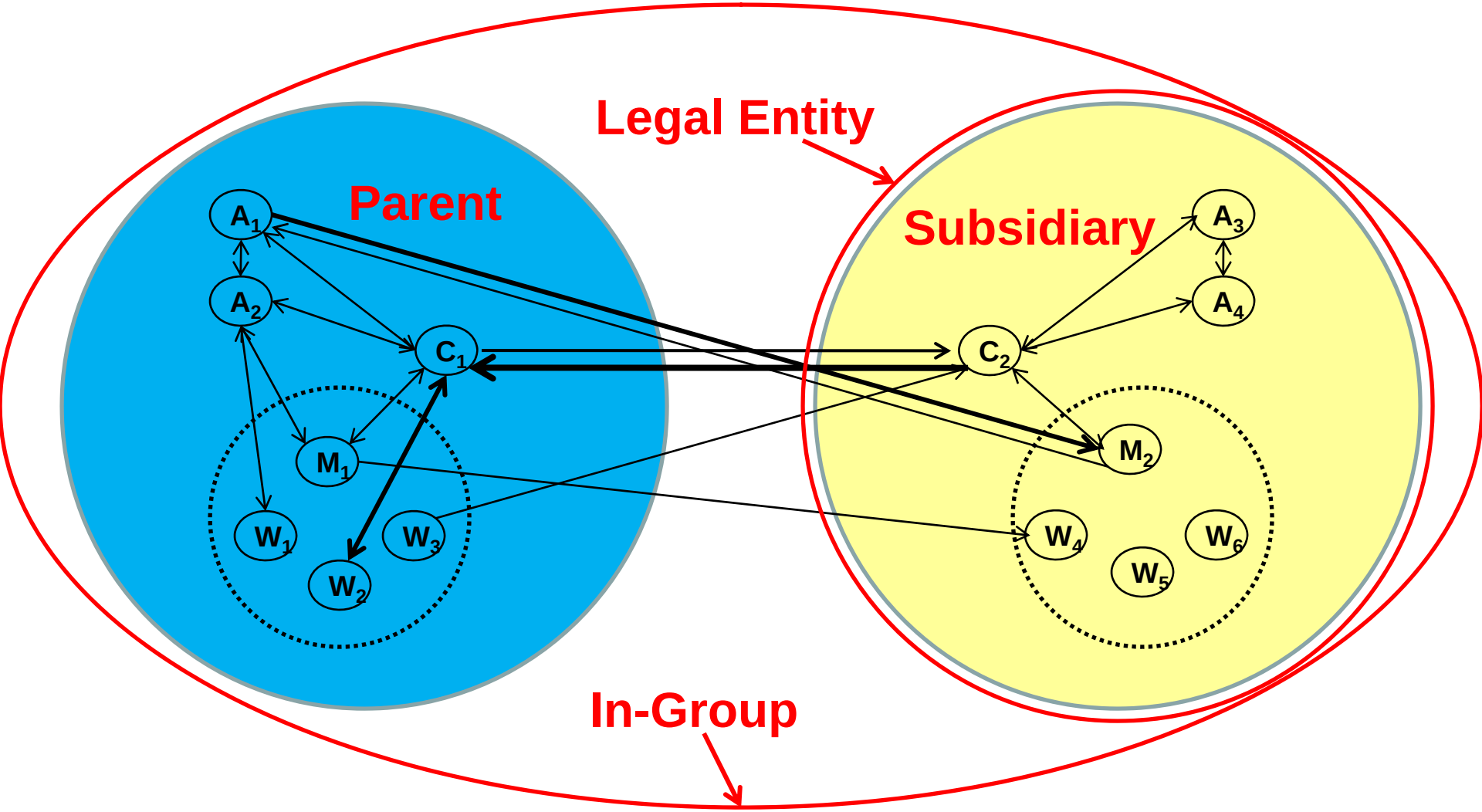
*$p < .05$.
**$p < .01$.
Notes: We report path coefficients and t-values. n.s. = not significant.

Adapted from: Griffith, Myers & Harvey 2006

Multinational Organization (1)



Multinational Organization (2)



Examples of Observed Differences

Work & Artifacts

• Face-to-Face Meetings	Japan > U.S.
• Video/Teleconferences	Japan > U.S.
• Tape Recordings	Japan > U.S.
• Reports	Japan < U.S.
• Slides/Presentations	Japan > U.S.
• E-Mail	Japan < U.S.
• Hand-Written Notes	Japan > U.S.
• Forms	Japan >> U.S.
• Text Messages	Japan < U.S.
• Spreadsheets	Japan > U.S.

Examples of Observed Differences

Sources & Information

- **Sharing of Information*** **Japan > U.S.**
- **Cross-Group Knowledge** **Japan > U.S.**
- **Legal Boundaries** **Japan << U.S.**
- **Repetition / Redundancy** **Japan > U.S.**

Observations

- **Culture Affects Work and Work Products**
- **Search Strategies May Need to Vary According to Record Type – High- vs. Low-Densities of Information**
- **Organizational Titles May Not Reflect Roles; Roles May Not Reflect Expertise or Knowledge**
- **Boundaries May Not Be Visible or Effective**

Lessons Learned

Culture & Behavior

- **We should change search strategies to mirror the cultural types and patterns of document generation.**
- **Weighting of “terms” should reflect the expected information-density of the target documents.**
- **We should anticipate “unusual” modes and patterns of information sharing, including the crossing of “legal” boundaries, when looking for sources.**

Conclusions

Conclusion 1

E-Discovery Problem *not* Solved by…

- Faster Search Response / Query Processing
- Bigger DBs
- "2.7 Terms" / Optimizing on User Queries
- "Better" Results above the Fold
- Generalized / Personalized Models of Users
- Freshness of Data
- Emergent Semantics / The "Crowd"
- …

Conclusion 2

- **Multinational corporations present special challenges (!)**
- **Culture affects behavior. In work groups, this may be reflected in different patterns of communication and different types of "document" artifacts.**
- **E-Discovery practices should take account of the linguistic and cultural-behavioral contexts of companies and individual workers.**

The End

Thanks!

References

Carroll, Lewis. Alice's Adventures in Wonderland. New York: Cassell and Company, Ltd. 1906.

Cormack, Gordon V.; Clarke, Charles L.A.; Palmer, Christopher R.; To, Samuel L. Passage Based Refinement. (MultiText Experiments for TREC-6.) In E.M. Voorhees and D.K. Harman (Editors), *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office, 1998.

Griffith, David A.; Myers, Matthew B.; Harvey, Michael G. An Investigation of National Culture's Influence on Relationship and Knowledge Resources in Interorganizational Relationships Between Japan and the United States. *Journal of International Marketing*, Vol. 14, No. 2, 2006. pp. 1-32.

Hofstede, G. (1991). *Culture and Organizations: Software of the Mind, Intercultural Cooperation and its Importance for Survival. McGraw-Hill, New York.*

Horii, T.; Jin, Y.; Levitt, R.E. Modeling and Analyzing Cultural Influences on Team Performance through Virtual Experiments. *Proceeding of the NAACSOS Conference 2004,* Pittsburgh, PA.

Horii, T.; Jin, Y.; Levitt, R.E. Modeling and Analyzing Cultural Influences on Project Team Performance. *Comput. Math. Organ. Theory* 10, 4 (Jan. 2005), 305-321. [2005a]
DOI= http://dx.doi.org/10.1007/s10588-005-6283-1

Horii, T.; Jin, Y.; Levitt, R.E. Impact of Multiple Normative Systems on Organization Performance of International Joint Ventures. Symposium on Normative Multi-Agent Systems, NORMAS 2005, University of Hertfordshire, Hatfield, UK, 12-15 April 2005, 54-64. [2005b] URL: http://dblp.uni-trier.de/db/conf/normas/normas2005.html#HoriiJL05

Milic-Frayling, Nataša; Zhai, ChengXiang; Tong, Xiang; Jansen, Peter; Evans, David A. Experiments in Query Optimization. The CLARIT System TREC-6 Report. In E.M. Voorhees and D.K. Harman (Editors), *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office, 1998, 415–454.

Najork, Marc; Zaragosa, Hugo; Taylor, Michael. HITS on the Web: How does it Compare? *SIGIR 2007 Proceedings*. ACM Press. 2007. 471–478.

Oard, Douglas W.; Hedin, Bruce; Tomlinson, Stephen; Baron, Jason R. Overview of the TREC 2008 Legal Track. In E.M. Voorhees and L. Bucklin (Editors), *The Seventeenth Text Retrieval Conference Proceedings (TREC 2008*). NIST Special Publication 500-277. 2009. (Cf. http://trec.nist.gov/pubs/trec17/t17_proceedings.html)

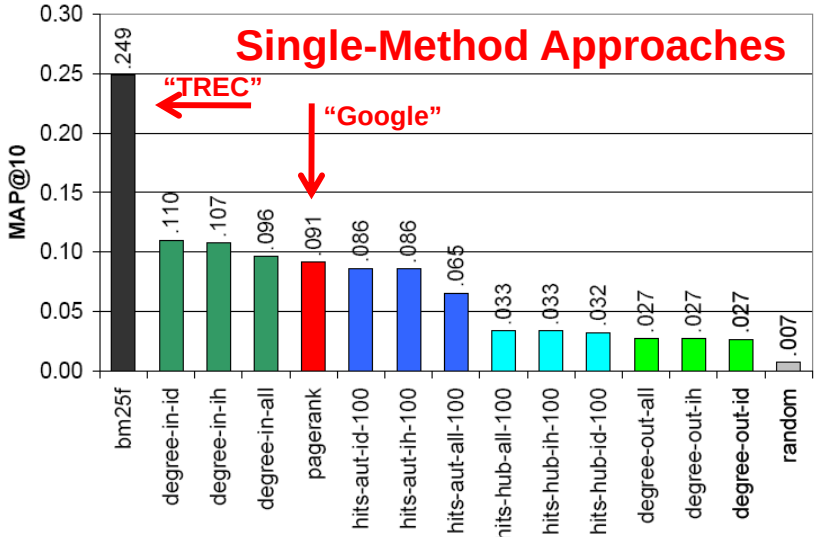
Appendix
Notes on Najork et al. 2007

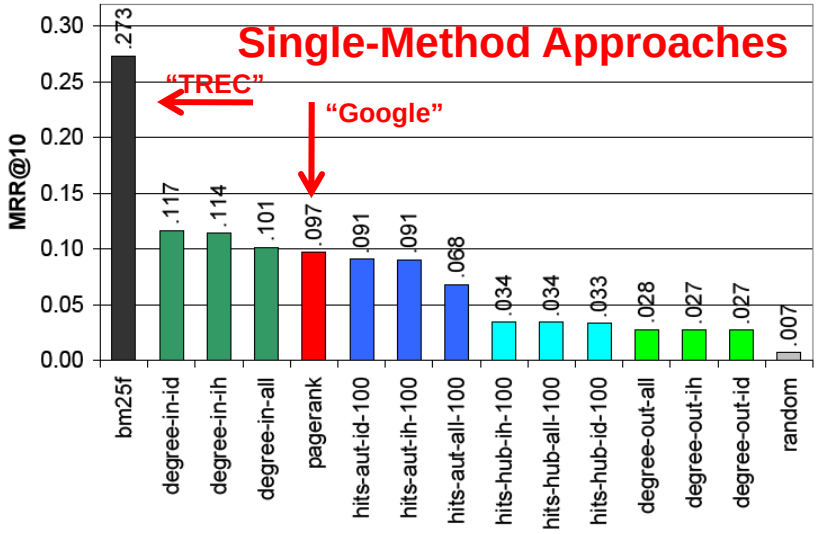
A Modern Evaluation of Web Search

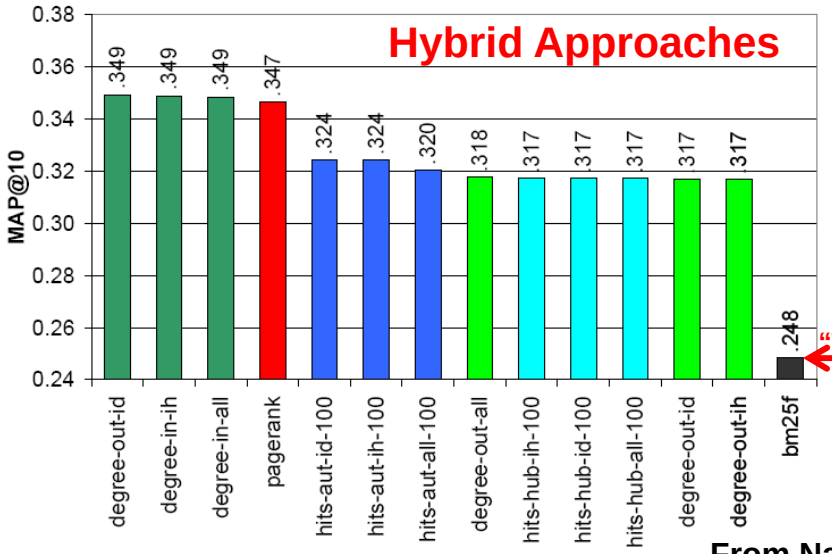
Najork et al. 2007

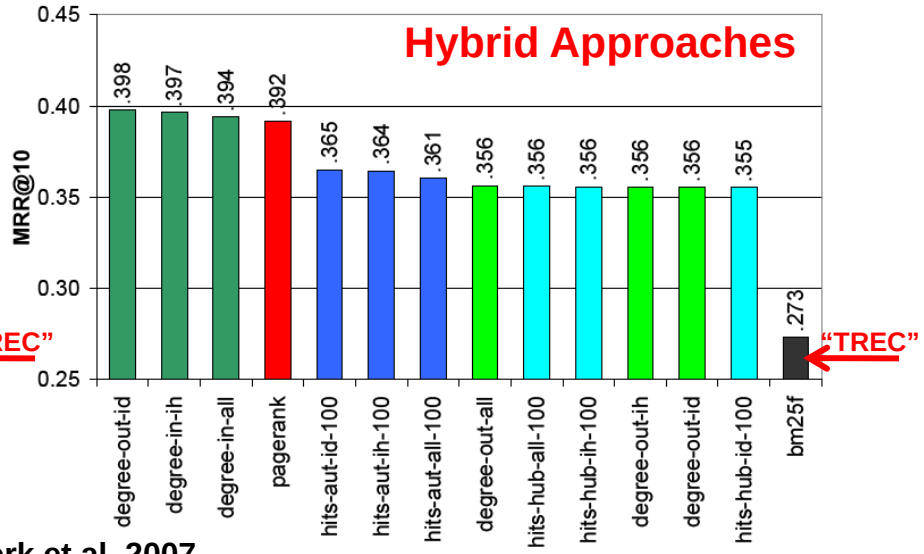
- **463,685,607 HTML Pages**
- **17,672,011,890 Non-Duplicate Hyperlinks**
- **2,897,671,002 URLs – 2,433,985,395 in Frontier**
- **28,043 Queries (Sampled from User Logs)**
- **66,846,214 Result URLs for Queries (2,838/Q)**
- **485,656 Results Evaluated for Relevance by Humans (on a six-point scale)**

State-of-the-Art Search Performance?









From Najork et al. 2007.