Desktop Search

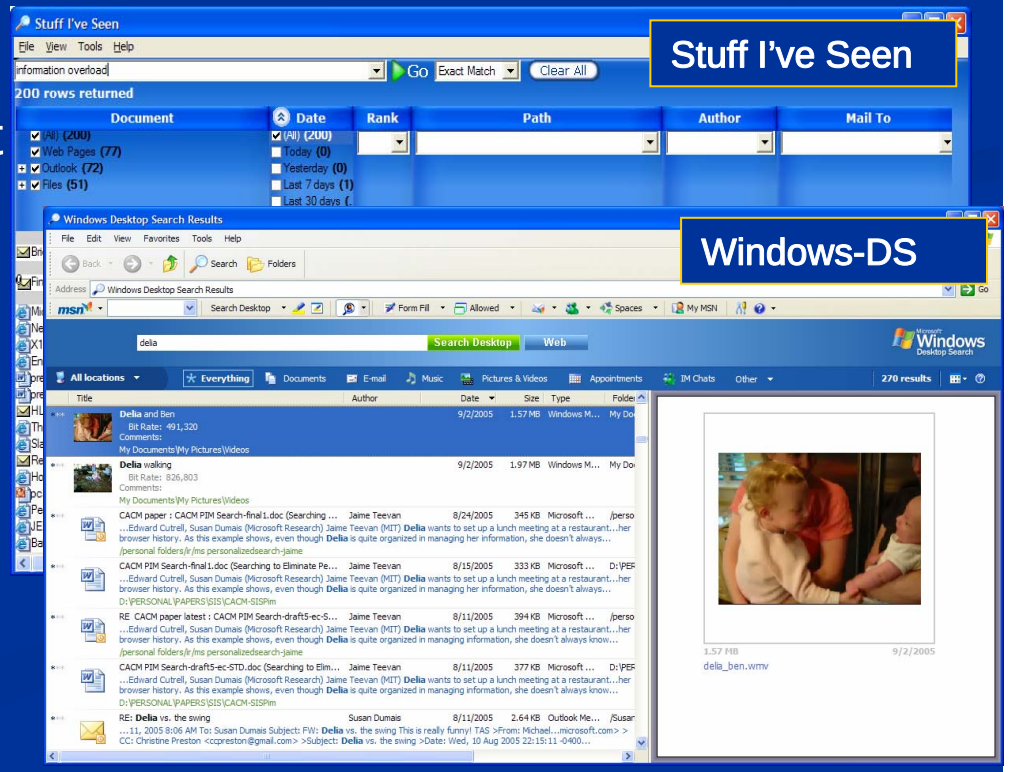
Susan Dumais

Microsoft Research

http://research.microsoft.com/~sdumais

Outline

- Search ≠ Web Search
  - E.g., Desktop Search, and many verticals
- Desktop Search ≈ *My Stuff*
- Stuff I've Seen (SIS)
  - *Case study:* Research prototype system, deployment experiences, usage data
  - → MSN Desktop Search; MS Vista Search (http://toolbar.msn.com)
- Future directions
  - Contextualized search
  - Personalized search

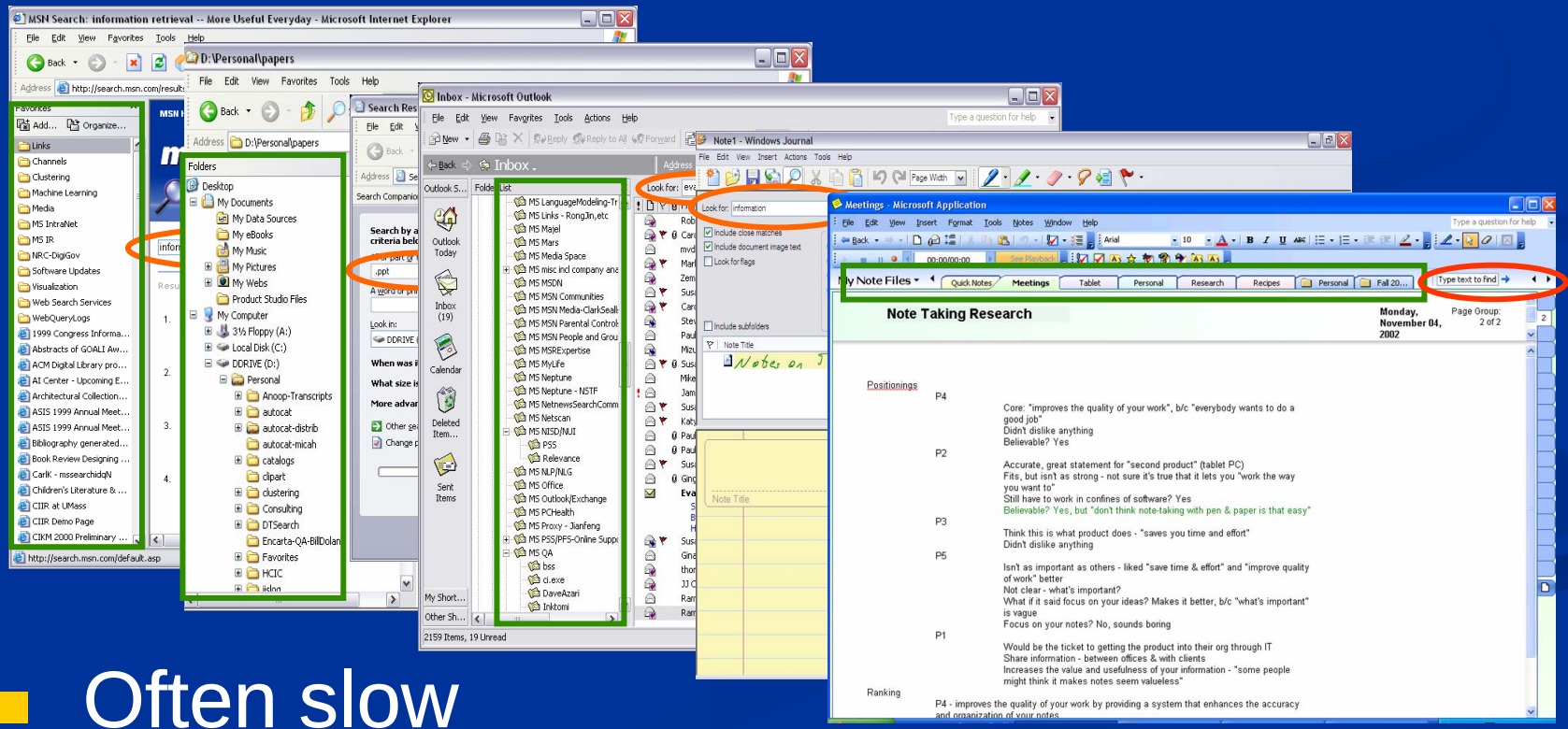


My Stuff

- Information acquisition vs. access
  - Easy to create or encounter lots of information
    - Types: email, docs, web pages, calendars, pictures, music, etc.
    - Amount: 100+ gig drives
  - Hard to organize
  - And, even harder to re-find
- Information discovery vs. recovery
  - Many tools for finding information (discovery)
  - Fewer tools for keeping information (recovery)
  - Yet, many tasks involve re-using information

Desktop Search Today

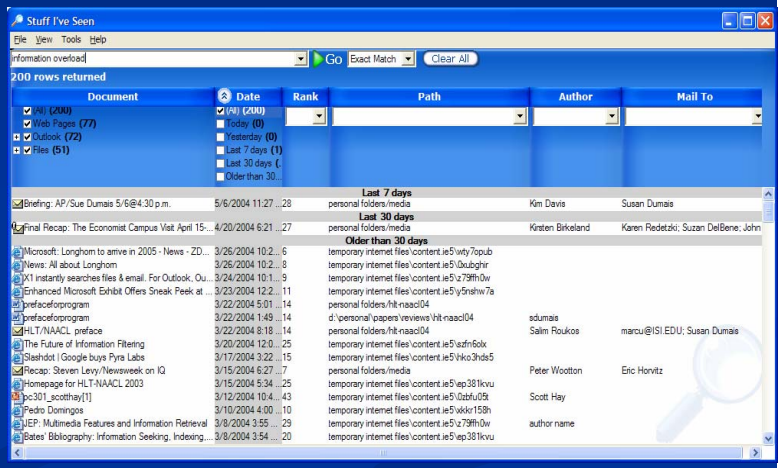
- Information silos
  - Many locations, interfaces for finding things (e.g., web, mail, contacts, docs, photos, notes)



- Often slow

Desktop Search With SIS

- Unified index of *stuff you've seen*
  - All types of info (e.g., files, email, calendar, contacts, web pages, rss, im)
  - Index content plus metadata (e.g., time, author, title, size, usage)
  - Automatic and immediate update of index
  - Rich UI possibilities, since it's *your* content (e.g., consider usage)



- Get back to information you've seen
  - Recovery vs. discovery

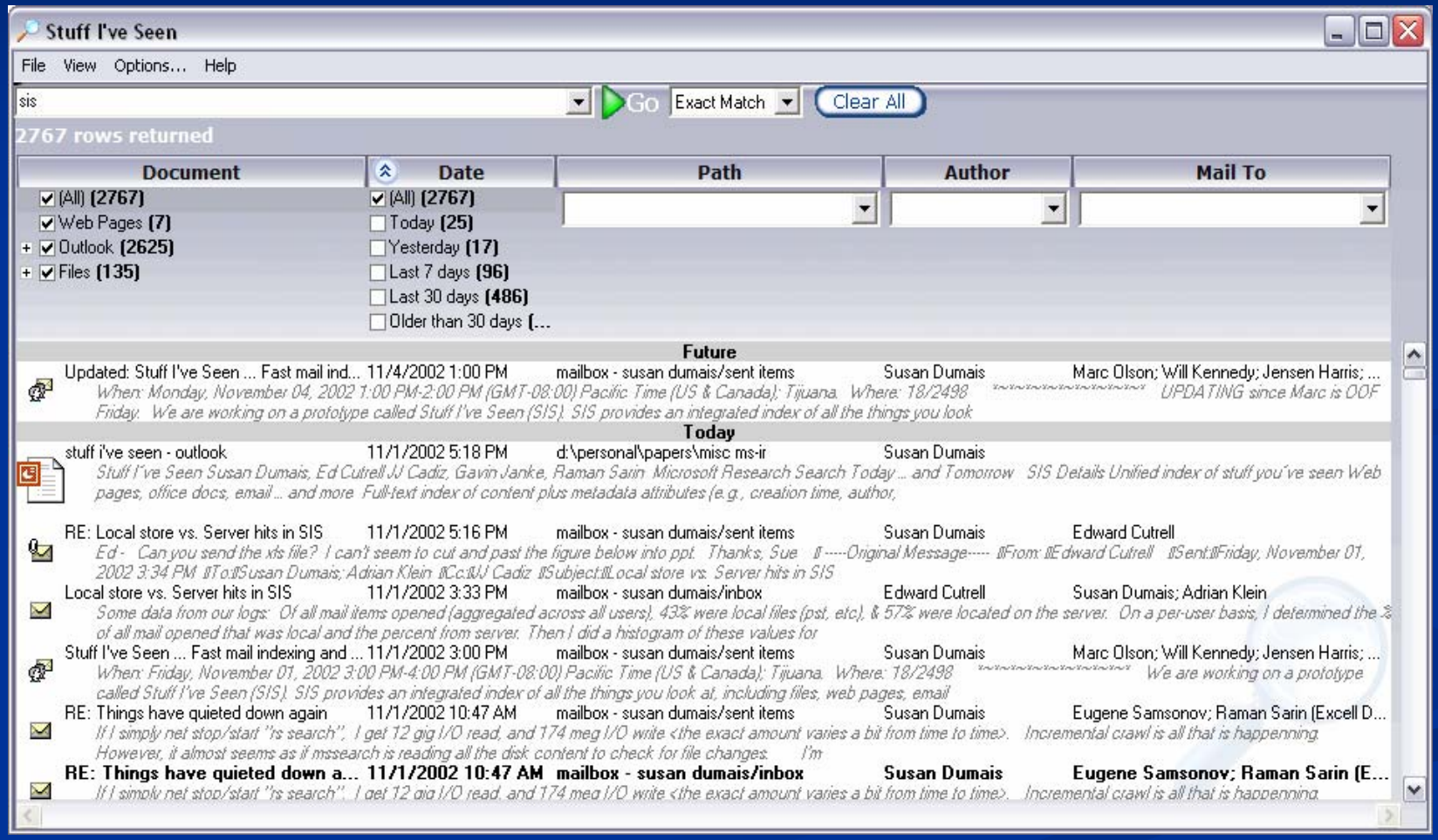
Related Work

- Research projects
  - Haystack [Adar et al., 1999; Huynh et al. 2002]
  - Keeping Found Things Found [Jones et al., 2001]
  - MyLife Bits [Gemmell, Bell et al., 2002]
  - Lifestreams/Scopeware [Fertig, Freeman, Gelernter, 1996]
- Commercial systems/software
  - OS: Mac OS X Spotlight, MS Vista Search
  - DS Apps: Enfish, dtSearch, Copernic, X1/Yahoo!, G-DS, MSN-DS, etc.
- What's new with SIS …
  - Full content and metadata for many different sources
  - Extensible architecture (gather, filter, word break)
  - Focus on user interface and user experience
  - Iterative design guided by usage and experimental data

SIS Design Principles

- Indexing experience …
  - No additional work is required
  - User sees something, and it gets indexed
- Retrieval experience …
  - Fast, flexible
  - Interactive refinement
    - Sort and filter on metadata
    - Note: Sort/filter automatically triggers query
  - UI innovations
    - Previews, Top/Side, Sort order
    - Richer visualizations

SIS Demo


Stuff I've Seen
File View Options... Help
sis
Go
Exact Match
Clear All
2767 rows returned
Document
Date
Path
Author
Mail To
(All) (2767)
Web Pages (7)
Outlook (2625)
Files (135)
(All) (2767)
Today (25)
Yesterday (17)
Last 7 days (96)
Last 30 days (486)
Older than 30 days (...
Future
Updated: Stuff I've Seen ... Fast mail ind... 11/4/2002 1:00 PM mailbox - susan dumais/sent items Susan Dumais Marc Olson; Will Kennedy; Jensen Harris; ...
When: Monday, November 04, 2002 1:00 PM-2:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana. Where: 18/2498 UPDATING since Marc is OOF Friday. We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look
Today
stuff i've seen - outlook 11/1/2002 5:18 PM d:\personal\papers\misc ms-ir Susan Dumais
Stuff I've Seen Susan Dumais, Ed Cutrell JJ Cadiz, Gavin Janke, Raman Sarin Microsoft Research Search Today ... and Tomorrow SIS Details Unified index of stuff you've seen Web pages, office docs, email ... and more Full-text index of content plus metadata attributes (e.g., creation time, author,
RE: Local store vs. Server hits in SIS 11/1/2002 5:16 PM mailbox - susan dumais/sent items Susan Dumais Edward Cutrell
Ed - Can you send the xls file? I can't seem to cut and past the figure below into ppt. Thanks, Sue -----Original Message----- From: Edward Cutrell Sent: Friday, November 01, 2002 3:34 PM To: Susan Dumais; Adrian Klein Cc: JJ Cadiz Subject: Local store vs. Server hits in SIS
Local store vs. Server hits in SIS 11/1/2002 3:33 PM mailbox - susan dumais/inbox Edward Cutrell Susan Dumais; Adrian Klein
Some data from our logs: Of all mail items opened (aggregated across all users), 43% were local files (pst, etc), & 57% were located on the server. On a per-user basis, I determined the % of all mail opened that was local and the percent from server. Then I did a histogram of these values for
Stuff I've Seen ... Fast mail indexing and ... 11/1/2002 3:00 PM mailbox - susan dumais/sent items Susan Dumais Marc Olson; Will Kennedy; Jensen Harris; ...
When: Friday, November 01, 2002 3:00 PM-4:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana. Where: 18/2498 We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at, including files, web pages, email
RE: Things have quieted down again 11/1/2002 10:47 AM mailbox - susan dumais/sent items Susan Dumais Eugene Samsonov; Raman Sarin (Excell D...
If I simply net stop/start "its search", I get 12 gig I/O read, and 174 meg I/O write <the exact amount varies a bit from time to time>. Incremental crawl is all that is happenning. However, it almost seems as if mssearch is reading all the disk content to check for file changes. I'm
RE: Things have quieted down a... 11/1/2002 10:47 AM mailbox - susan dumais/inbox Susan Dumais Eugene Samsonov; Raman Sarin (E...
If I simply net stop/start "its search", I get 12 gig I/O read, and 174 meg I/O write <the exact amount varies a bit from time to time>. Incremental crawl is all that is happenning.

SIS Architecture

- Indexing infrastructure uses MS Search components (note: IR platform)
  - *Gatherer* – interface to content sources, e.g., files, http, MAPI
  - *Filters* – decode different file types, e.g., word, powerpoint, html, pdf, journal notes
  - *Tokenizer* – break into words, including date normalization, stemming, etc.
  - *Indexer* – standard inverted index
  - *Retriever* – Boolean, best match (Okapi), fielded
- User interface
- Client side indexing and storage

Evaluating SIS

- Internal deployment
  - ~3000 users
  - Users include: program management, test, sales, development, administrative, executives, etc.
- Research techniques
  - Free-form feedback
  - Questionnaires; Structured interviews
  - Usage patterns from log data
  - UI experiments (randomly deploy different versions)
  - Lab studies for richer UI (e.g., timeline, trends)

SIS Usage Data

Personal store characteristics

- 5k – 500k items

Sue's (Laptop) World		
Type	**N**	**Size**
Web	3k	0.2 Gb
Files	28k	23.0 GB
Mail	60k	2.2 Gb
Total	**91k items**	**25.4 Gb**
Index		**190 Mb**
		+1.5 Mb/week

Query characteristi

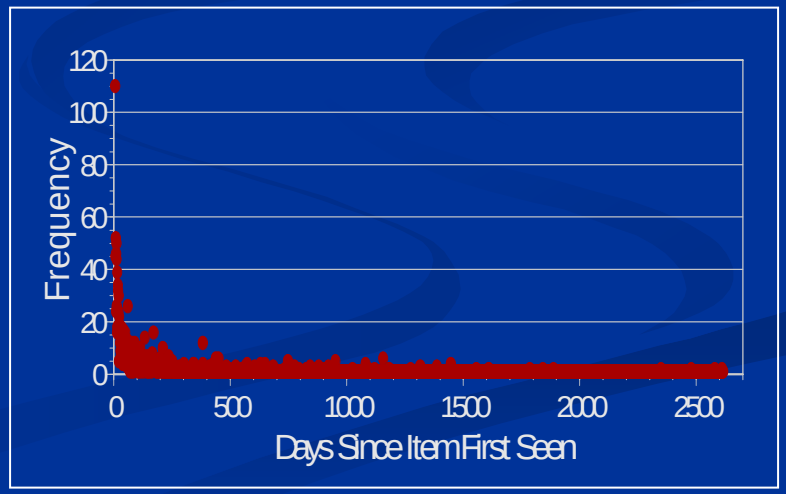
- Short queries (1.6 w
- Few advanced oper
- Many advanced operators and query iteration in UI (48%)
  - Filters (type, date); modify query; re-sort results

SIS Usage Data, cont'd

Characteristics of items opened

- File types opened
  - 76% Email
  - 14% Web pages
  - 10% Files
- Age of items opened
  - 5% today
  - 21% within the last week
  - 47% within the last month
  - 50% of the cases -> 36 days
    - Web: 11 days
    - Mail: 36 days
    - Files: 55 days

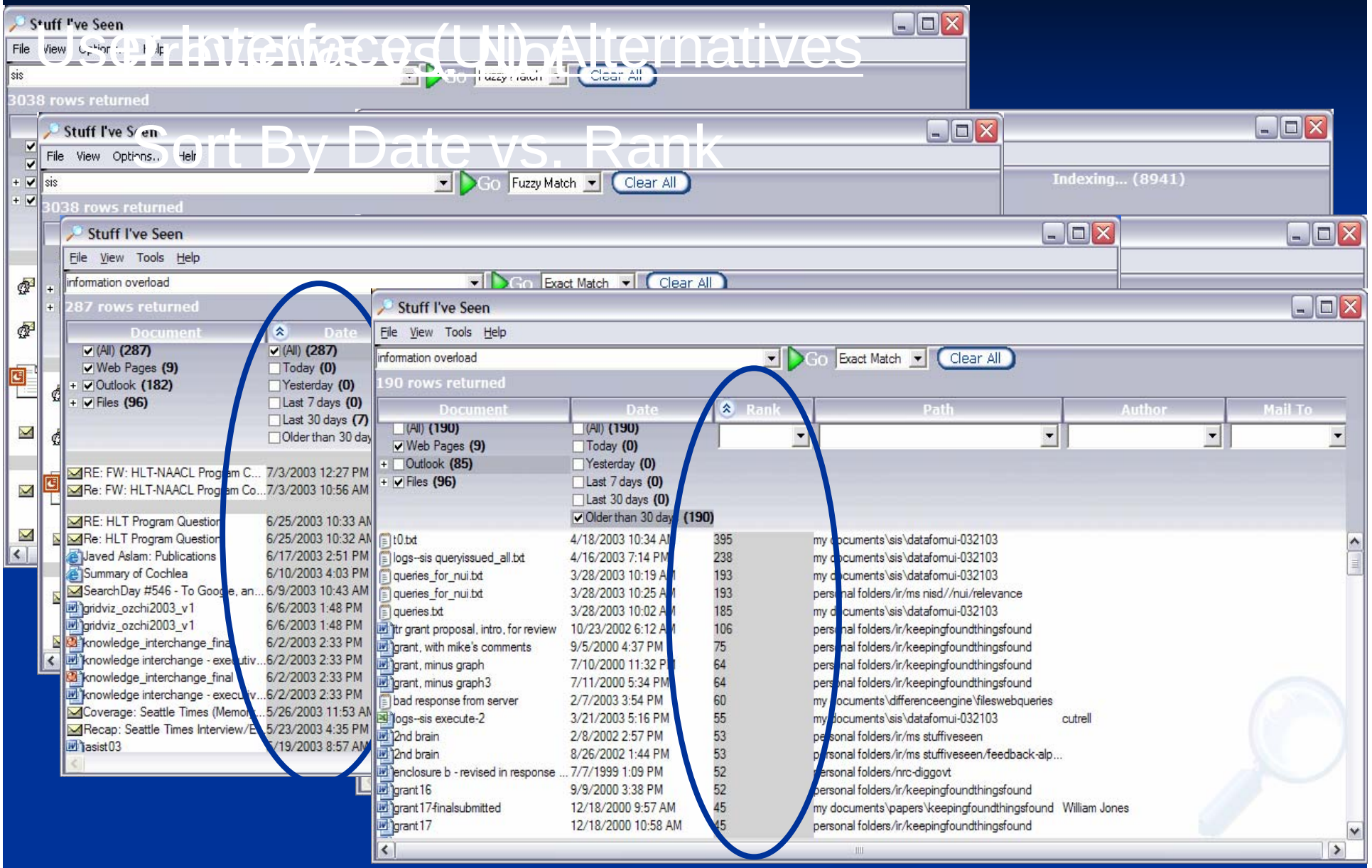
Log(Freq) = -0.68 * log(DaysSinceSeen) + 2.02



Top vs. Side Views

User Interfaces (UI) Alternatives

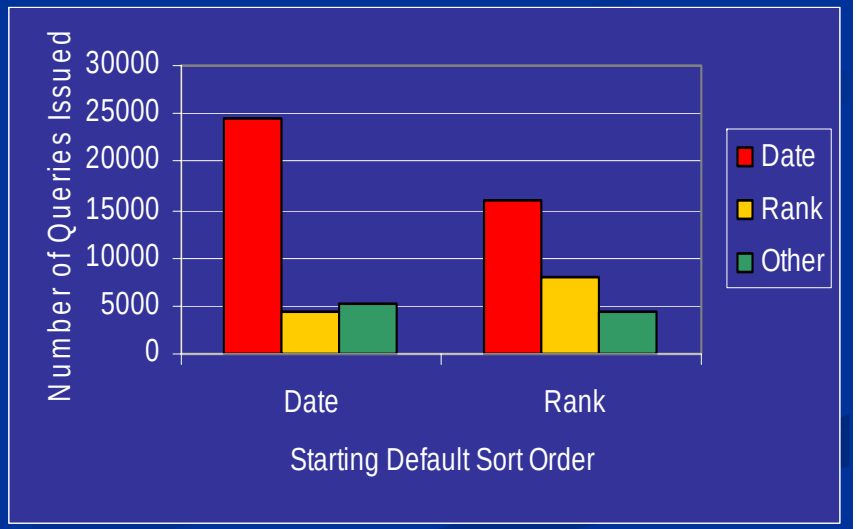
Sort By Date vs. Rank



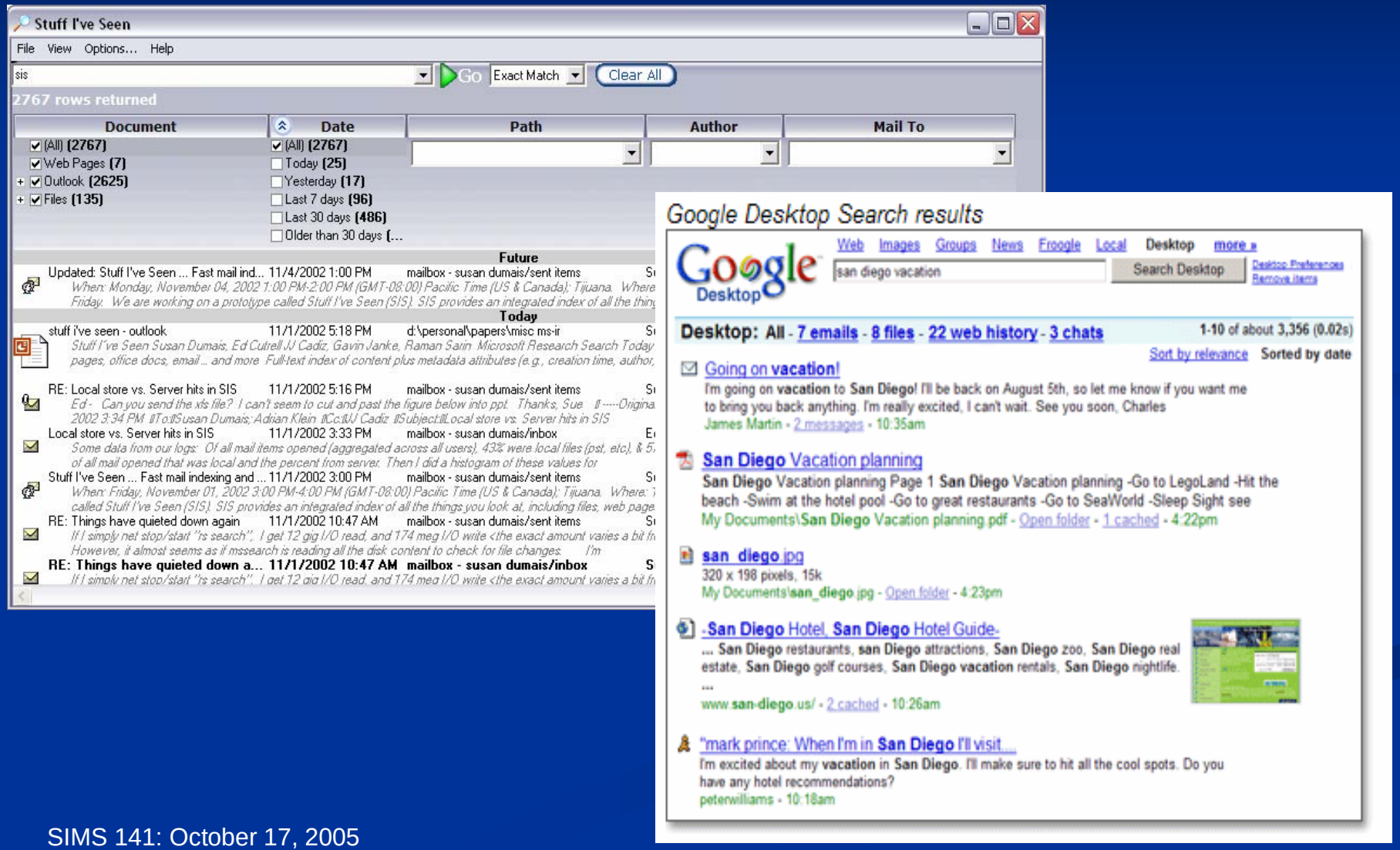
SIS Usage Data, cont'd

UI Usage

- Small effects of: Top/Side, Previews/NoPreviews
- Large effect of Sort Order:
  - Date by far the most common sort field, even for people who had Okapi Rank as default
  - Importance of time
  - Few searches for "best" match; many other criteria …



Metadata vs. Best-match list



SIS Usage Data, cont'd

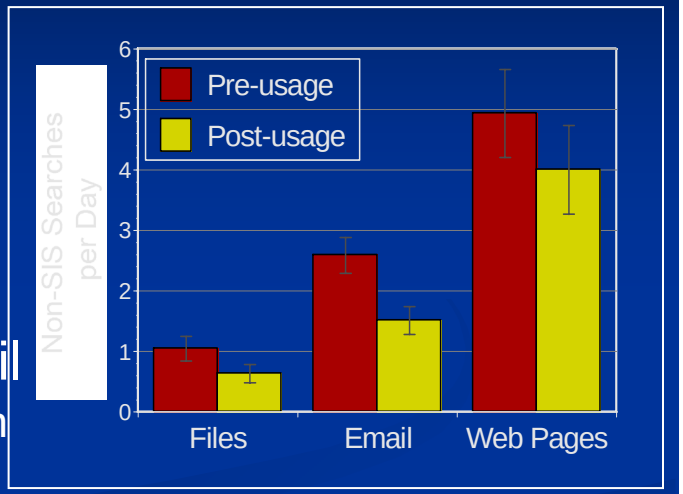
Observations about unified access

- Metadata quality is variable
  - Email: rich, pretty clean
  - Web: little, not very useful for retrieval
  - Files: some, but often wrong
- Need abstractions, e.g., "Useful date", "People", "Picture"
  - Initially, used 'date seen'
  - But …
    - Appointment, when it happens
    - File, when it is changed
    - Email and Web, when it is seen
  - "Useful date" abstraction

SIS Usage Data, cont'd

Ease of finding information

- Easier after SIS for web, email, files
- Non-SIS search decreases for web, email, files



Additional benefits

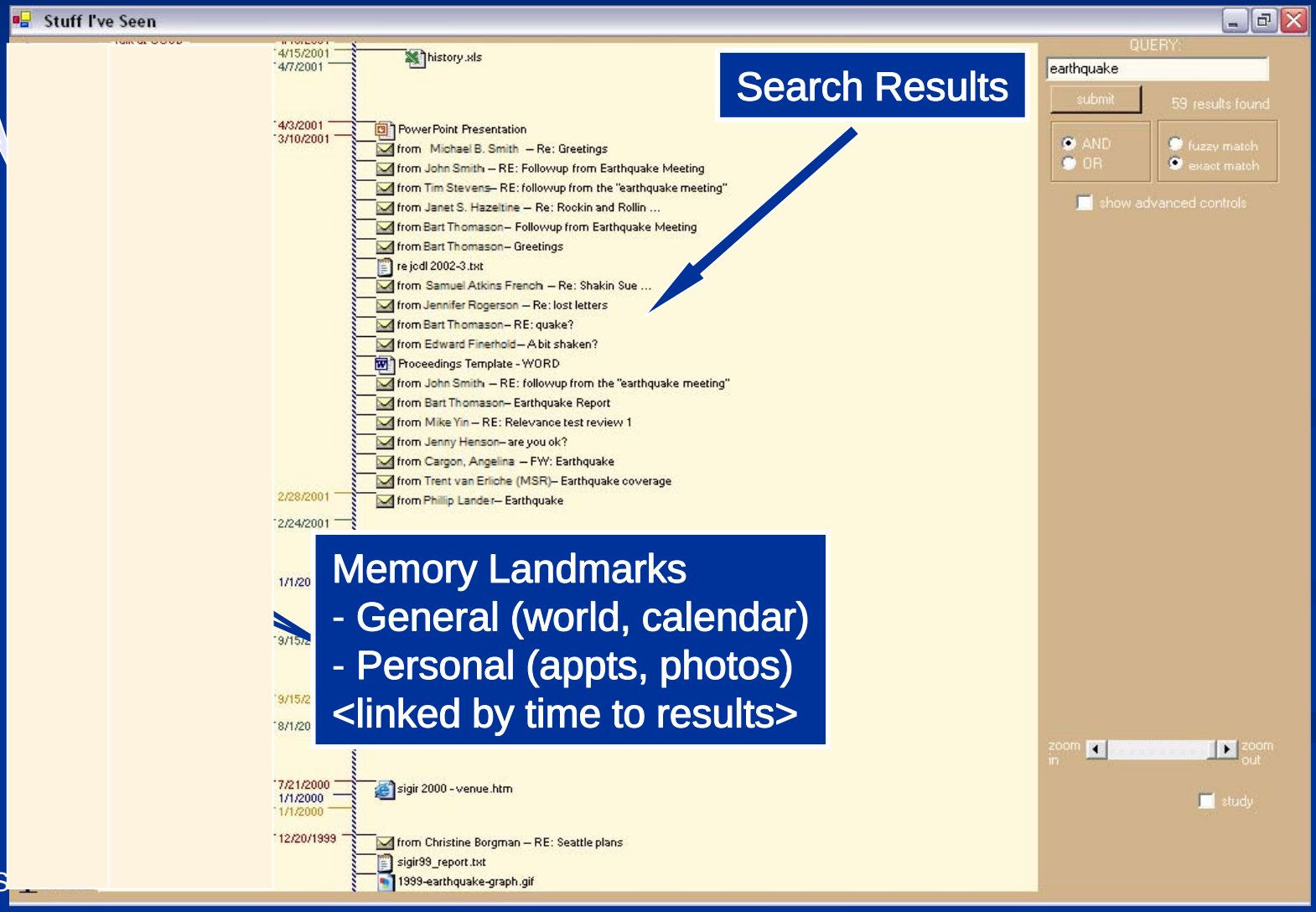
- "The ability to *find misfiled documents* and email has been extremely helpful." -- A sales executive in Washington D.C.
- "Thanks again for the MARVELOUS tool! I find myself unable to live without it! It saves me at least 10-15 minutes a day looking for information; saves even more time *not having to file things*. It makes me more effective, as more time goes to thinking and deciding, and less to overhead." -- An executive in Redmond.

SIS, Timeline w/ Landmarks

- SIS: time as important access cue
- Importance of “landmarks” in human memory
- Identify and use landmarks to facilitate information management and search
- Timeline interface, augmented with landmarks
  - General landmarks: holidays, world events
  - Personal landmarks: important photos, appointments
  - Heuristics or Bayesian models to identify memorable events

SIS, Timeline w/ Landmarks

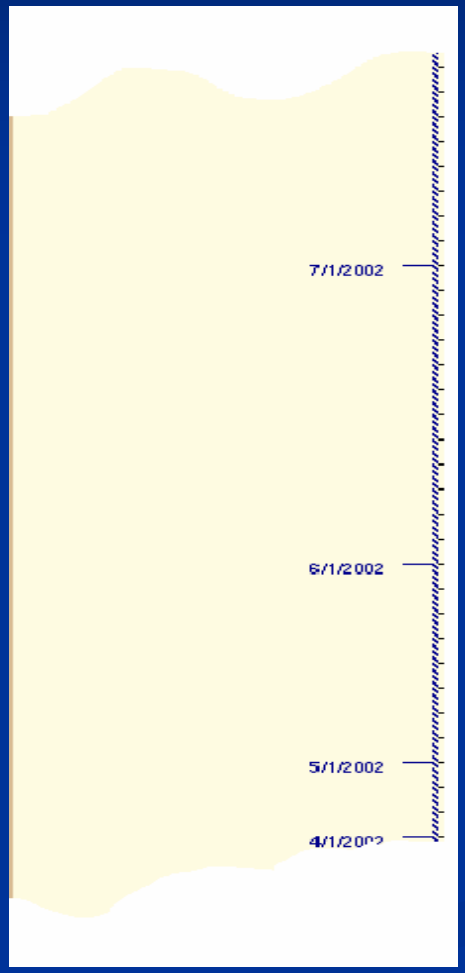
Distribution of Results Over Time

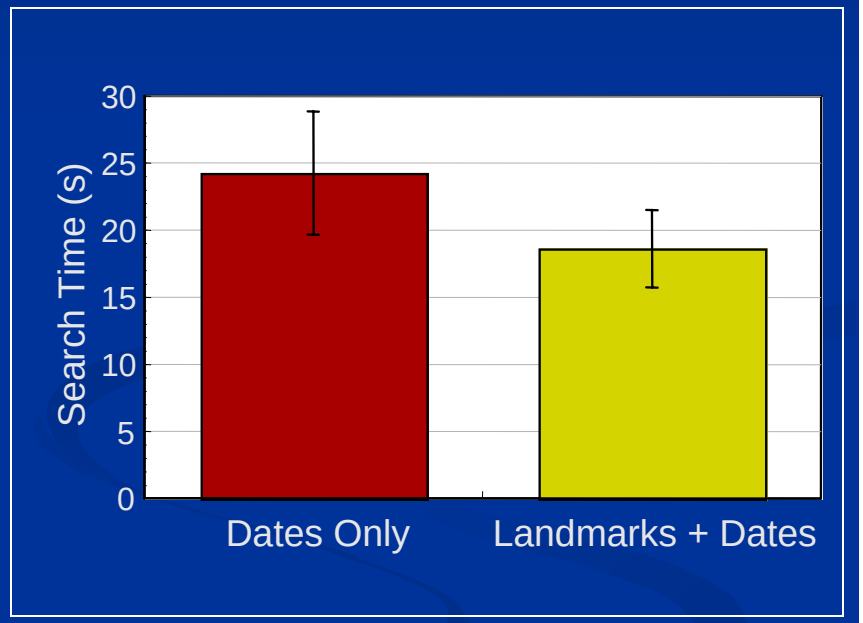


SIS, Timeline Experiment

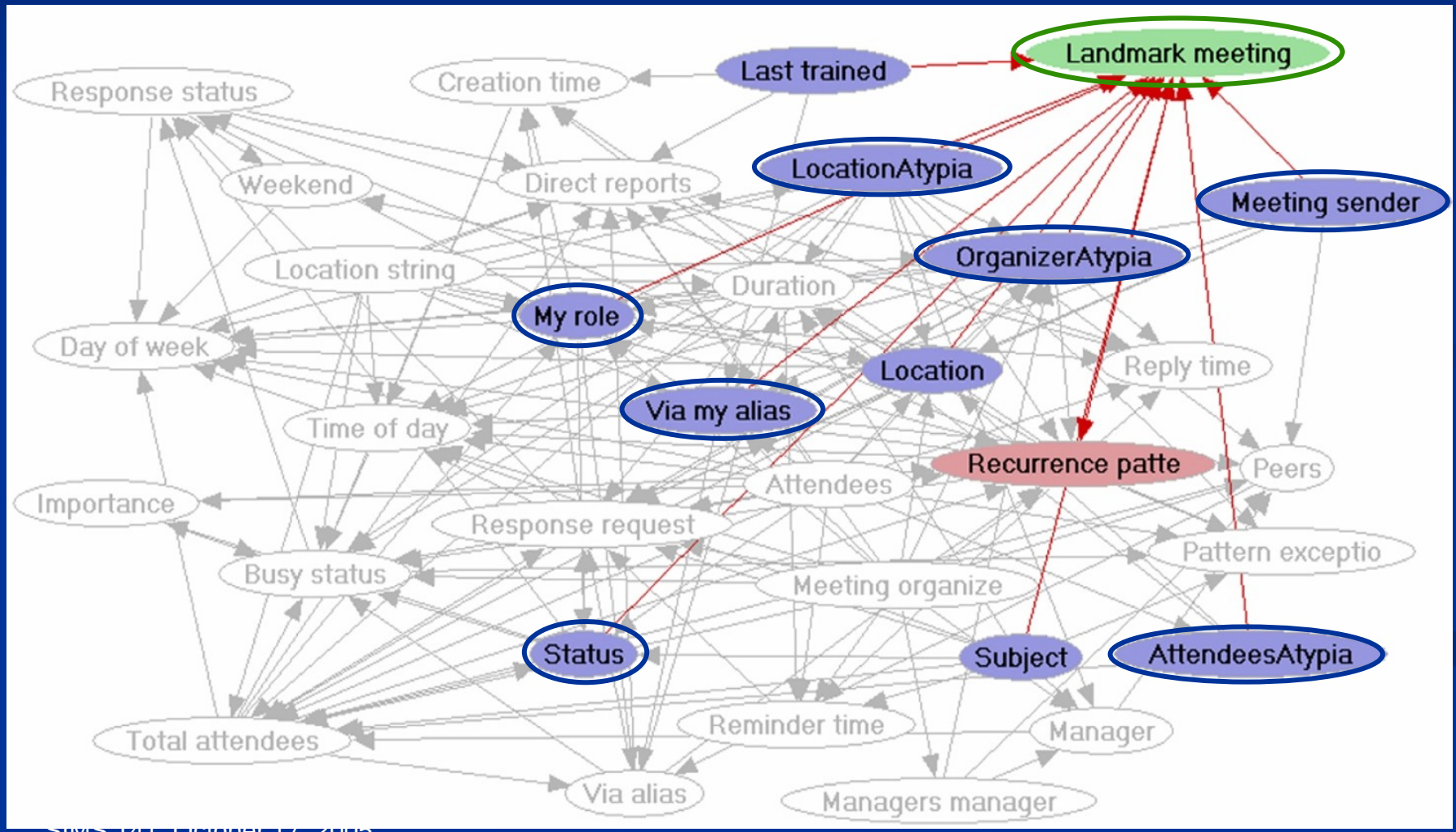
With Landmarks

Without Landmarks





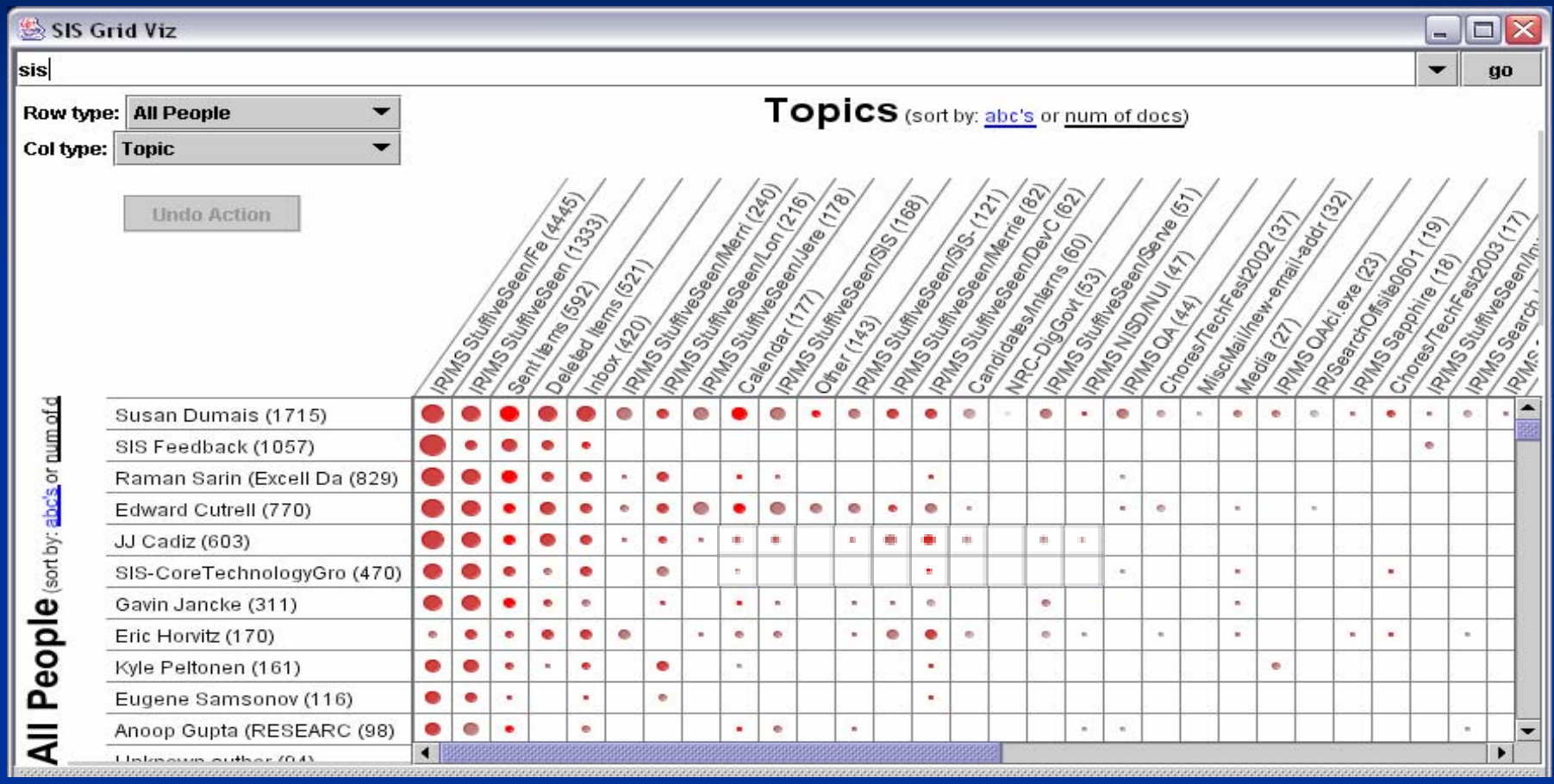
Landmarks, key dependencies
(from learned graphical model)



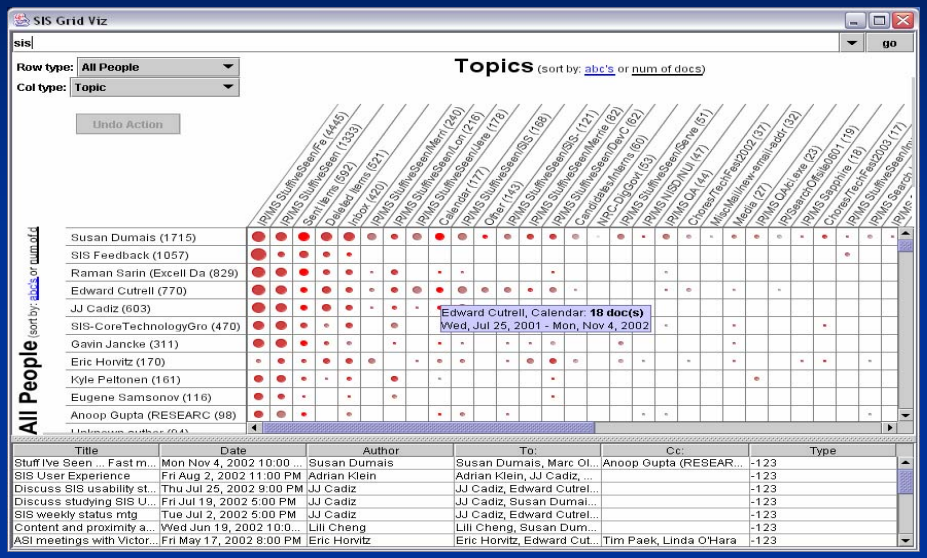
SIS, Visualizing Patterns

- Summarize the results of a search
  - Abstraction beyond individual results
- Grid-based design
  - Axes represent topic, time, people, etc.
  - Cells encode frequency, recency
- Supports activities like:
  - What newsgroups are active (on topic x)?
  - What people are active, authoritative (on topic x)?
  - When did I last interact w/ people?

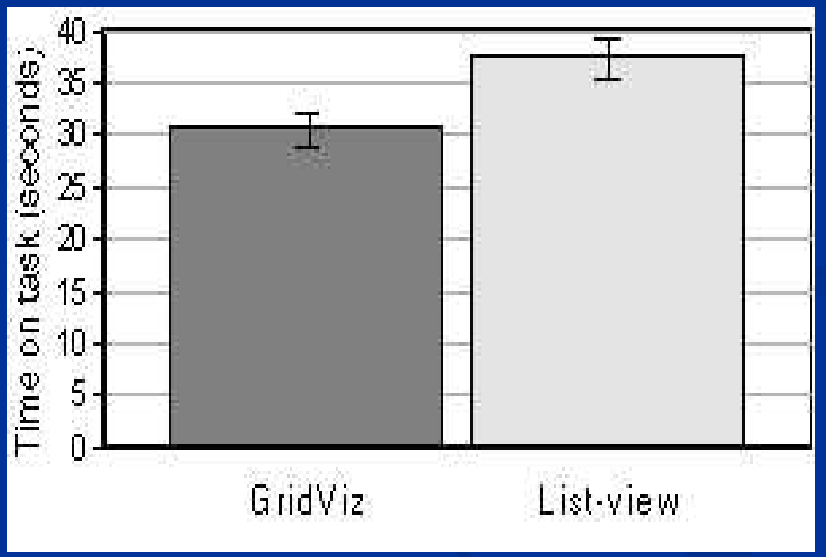
SIS, Visualizing Patterns

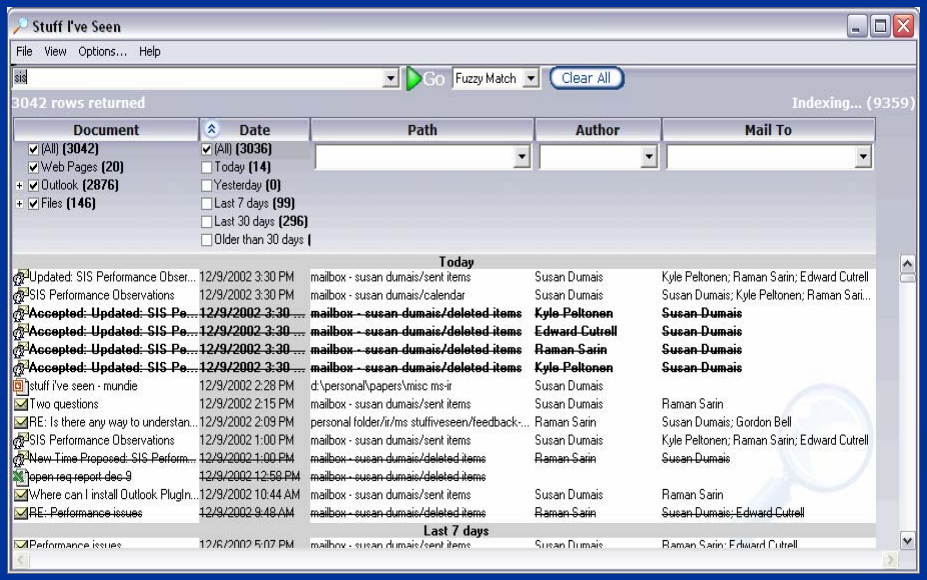


SIS, Grid vs. List Experiment



Grid View

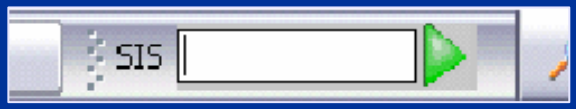




List View

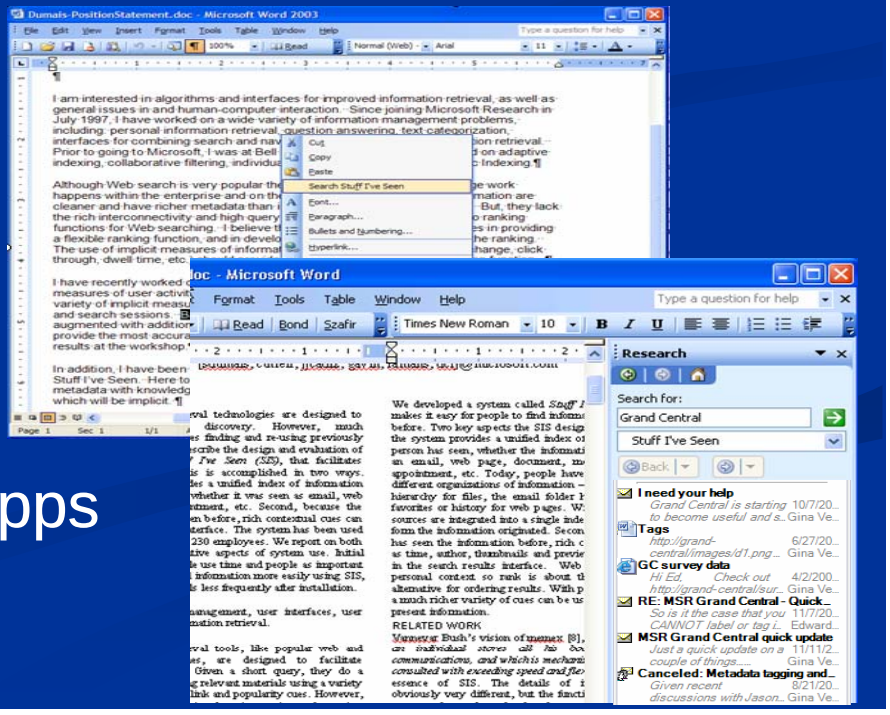
Contextualized Search

- Search is not the end goal …
- Need to support information management in the context of ongoing work activities
  - Search always available



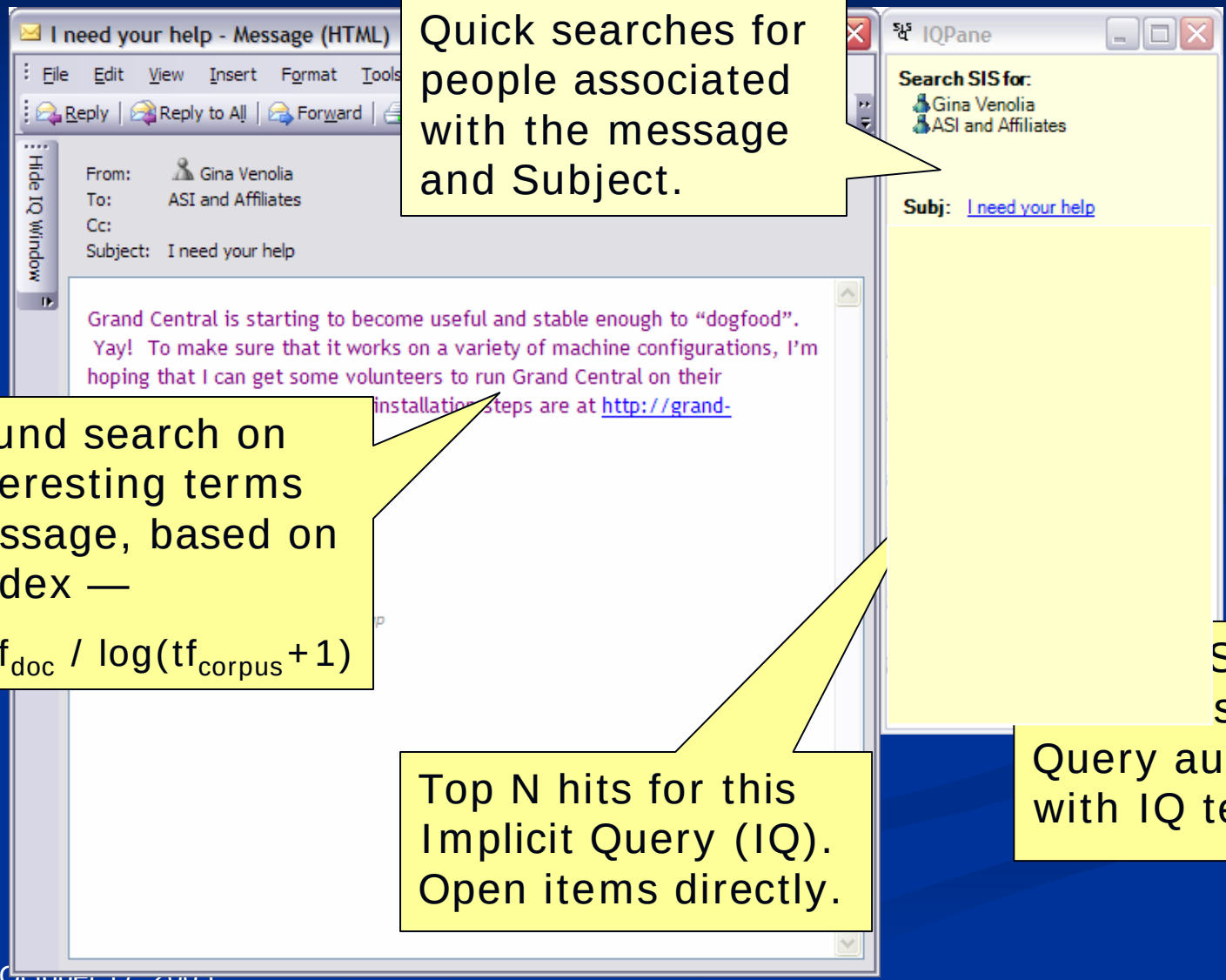
  - Search from within apps

    (select keywords/regions, full docs)



  - Show results in context of apps

Context, Implicit Queries



Quick searches for people associated with the message and Subject.

Background search on top *k* interesting terms from message, based on user's index —

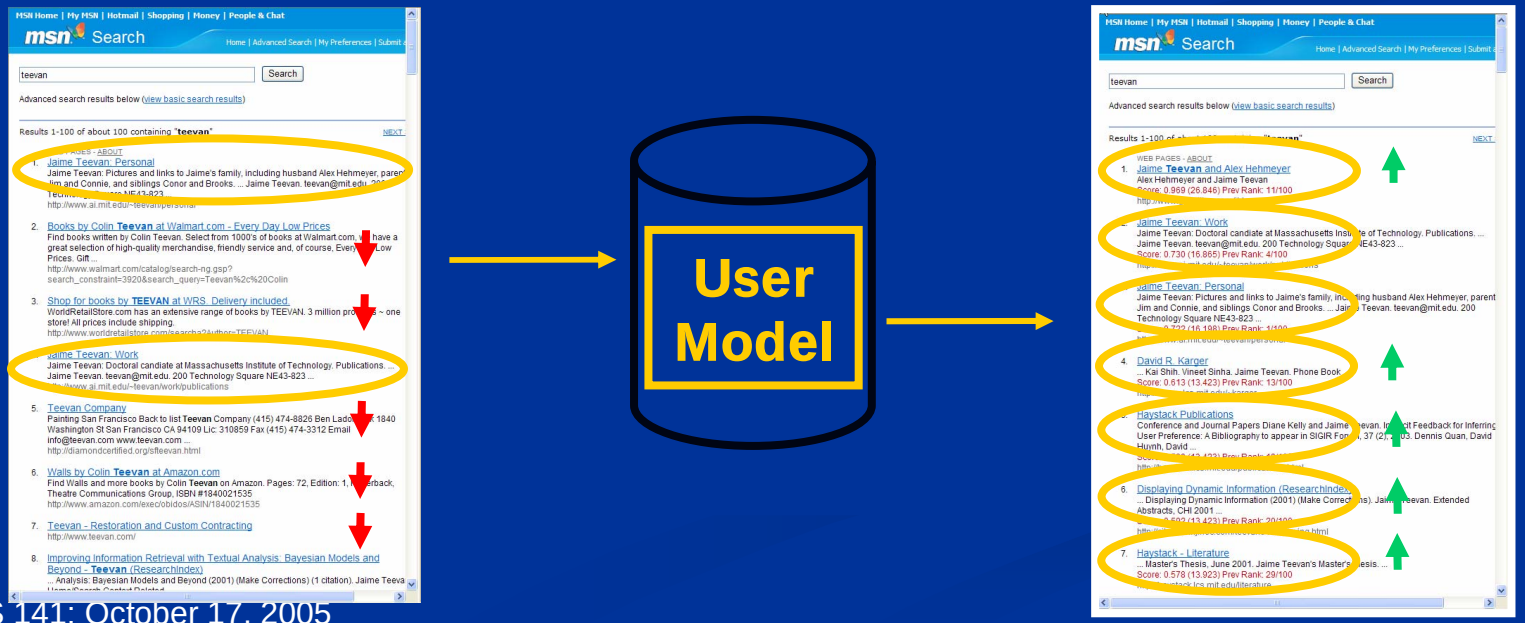
Score = tf_{doc} / log(tf_{corpus}+1)

Top N hits for this Implicit Query (IQ). Open items directly.

S for search. Query autofills with IQ terms.

Personalized Search

- Today:
  - All users get the same results, independent of previous search history, current context, etc.
- Personalized Search Prototype:



Desktop Search Summary

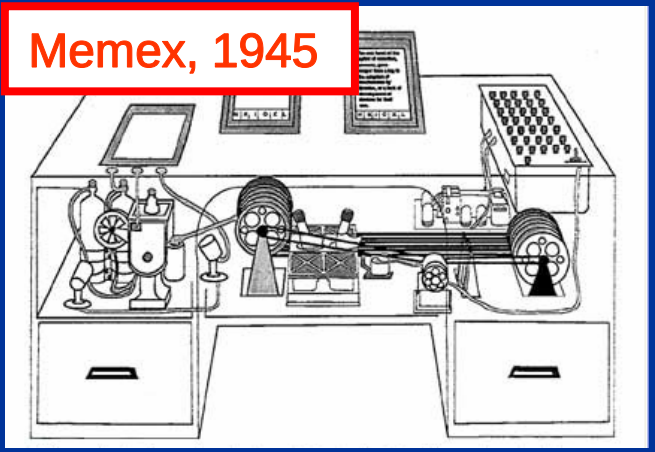
- Desktop search with SIS
  - Unified index to *stuff you've seen*
  - Heterogeneous content: files, email, web, etc.
  - Fast and flexible access
- Supports new capabilities for personal information management
  - Rich metadata vs. single hierarchy or ranked list
  - Landmarks, patterns, implicit queries, etc.
- New directions
  - Contextualized search
  - Personalized search

Vannevar Bush's Vision

V. Bush (1945). As we may think. *Atlantic Monthly, 176*, July 1945, 101-108.

- *Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.*

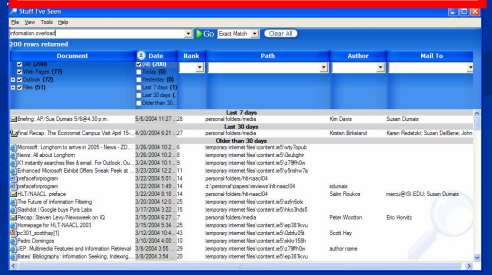
Memex, 1945



NLS/Augment, 1968

The oN-Line System display, keyboard and mouse

DesktopSearch, 2003



Thank You

- Questions/Comments …
- More info, http://research.microsoft.com/~sdumais
- MSN Toolbar and Desktop Search, http://toolbar.msn.com