

# Foundations of User-Oriented Information Retrieval: User Interfaces & User-Centered Evaluation

Diane Kelly, Professor and Director  
School of Information Sciences  
University of Tennessee, USA

*European Summer School in Information Retrieval*  
July 16, 2019

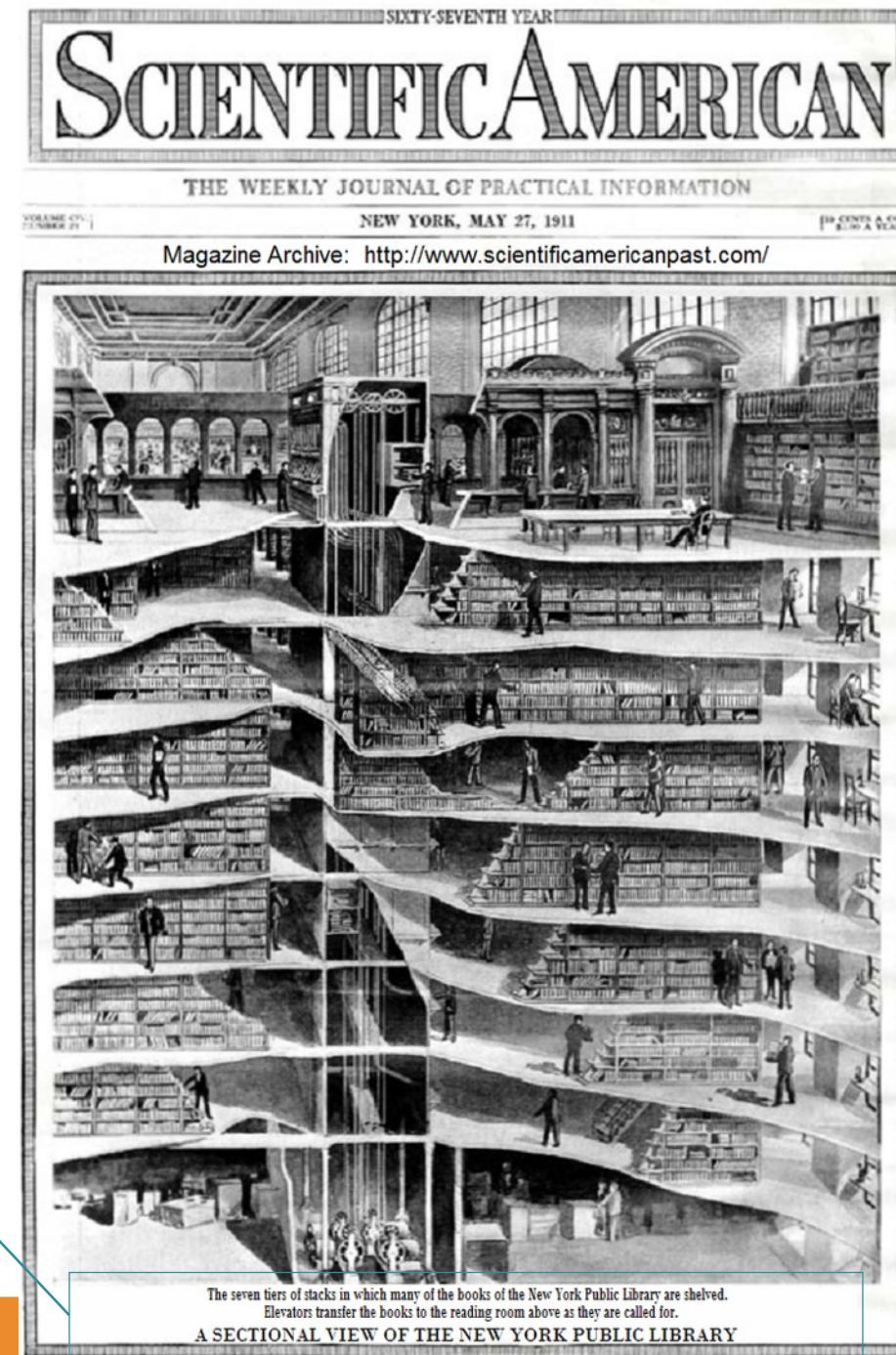
# Agenda

- Search User Interfaces (30 minutes)
- User-Centered Evaluation (1 hour)
  - Deep dive into some details

# Information Retrieval: Closed Stacks System

*Scientific American*  
*The Weekly Journal of Practical Information,*  
May 27, 1911

"The seven tiers of stacks in which many of the books of the New York Public Library are shelved. Elevators transfer the books to the reading room above as they are called for."



# Information Retrieval: Closed Stacks System?



# Interactive IR Evolution

Query (Croft)	Data & Collections	Producers	Information Needs	Users
1970 CE (Boolean search) 	Highly Regulated Highly Curated	Specialized	Homogenous Professional Serious	Specialized
1994 CE (web search) 	negligence navigation aids			
2005 CE (CQA) 	Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?	Highly Unregulated Curated and Uncurated	“Everybody”	“Everybody”

# Interaction Evolution



This is magnetic tape.

## Search Interface

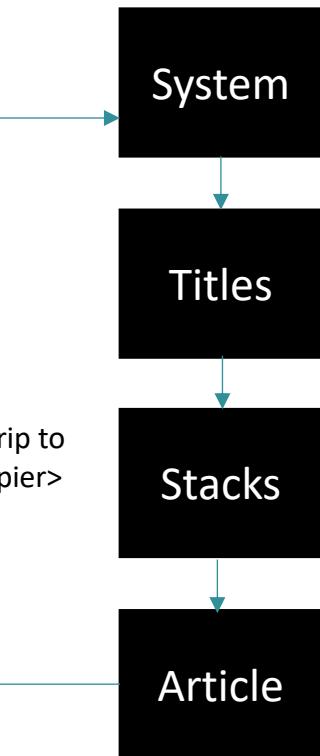
Choose the titles you would like to print.

Addressing the issue of cataloging and making chiropractic  
Healthnet: Connecticut Consumer Health Information Network.  
Leisure for creative thought: planned respites from classroom  
From Shamans to curators: bearers of tradition.  
Archives, data bases, and interactive computer programs: are  
Reviewing the research literature.  
Z\iterature search: a short description for nursing personnel]  
How to conduct a literature search.  
A reference shape library for computer aided socket design in

OK

Exit

<insert trip to  
photocopier>



TODAY: Search applications, information, information needs and searchers are diverse!



The New York Times

Westlaw



# Involving the User in Search

Since the user's original query is often inadequate, some sort of user interaction with the retrieval operation is desirable. The user of a manual retrieval system such as a library might at first ask a general and unclear question. The librarian, using his knowledge of the document collection, might then ask the user a few questions and show him a few books in an attempt to pinpoint his needs.

This study investigates relevance feedback, which is a procedure allowing user interaction with an automated information retrieval system. The user is given a small set of possibly relevant items, and is asked to judge each as relevant or non-relevant to his request. These user relevance judgments are then used for feedback to the information retrieval system, to produce a better subsequent set of retrieved items.

Ide, E. (1967, 1969). User interaction with an automated information retrieval system. In G. Salton (Ed.) *Information Storage and Retrieval: Scientific Report No. ISR-12 and ISR-15*.

# Involving the User in Search

- By the mid-1960s, several techniques had been introduced to assist users, including the:
  - Display of online thesauri to help with query formulation
  - Choice of novice or experienced searcher interface mode
  - Ability to save search queries to rerun at a later time or on a different database
  - Relevance feedback
  - System prompts for further information from user about his/her information need
- In 1971, the first workshop was held about interactive searching.
  - Walker, D.E. (1971). *Interactive bibliographic search: The user/computer interface*. Montvale, NJ: AFIPS Press.

A Positive End-Expiratory Pressure—Nasal-Assist Device (PEEP-NAD) for treatment of respiratory distress syndrome.; Tummons, *Anesthesiology*, 38, 592-5, June 73.

1. J L Tummons, 2. blood, 3. carbon dioxide, 4. human, 5. hydrogen-ion concentration, 6. infant, newborn, 7. masks, 8. methods, 9. nose, 10. oxygen, 11. oxygen inhalation therapy, 12. positive-pressure respiration, 13. respiration, 14. respiratory distress syndrome

► Yes, 13, not 6

We are not doing so well now. You may already have the important references.

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of Documentation* 33(1), 1-14.

SYSTEM MODE → \* FRED!  
\* PLEASE SELECT ONE OF THE FOLLOWING SYSTEMS

\* TR - TEXTUAL RETRIEVAL  
\* DB - DATA BASE MANAGEMENT SYSTEM

USER MODE → Δ TR ↴  
\* PLEASE SELECT ONE OF THE FOLLOWING FUNCTIONS

\* 1. SUBJECT MATTER INDEX  
\* 2. BOOLEAN REQUEST  
\* 3. THESAURUS  
\* 4. CITATOR  
\* 5. PRINT OPTION  
\* 6. HELP  
\* 7. STOP

Δ 2 ↴  
\* PLEASE SELECT ONE OF THE FOLLOWING BOOLEAN SUB-FUNCTIONS

\* 1. NEW REQUEST  
\* 2. ERROR  
\* 3. OLD REQUEST  
\* 4. HELP OF THESAURUS REQUEST  
\* 5. HELP  
\* 6. STOP

Δ 3 ↴  
\* PLEASE SELECT ONE OF THE FOLLOWING OLD REQUEST ROUTINES

\* 1. ADD A NEW SINGLE WORD  
\* 2. DELETE A SINGLE WORD  
\* 3. REPLACE A SINGLE WORD  
\* 4. ADD A NEW SEQUENCE OF WORDS  
\* 5. DELETE A SEQUENCE OF WORDS  
\* 6. REPLACE A SEQUENCE OF WORDS

Δ 2 ↴  
\* FRED!  
\* YOU SELECT TO USE THE ROUTINE FOR DELETING A SINGLE WORD FROM  
YOUR OLD BOOLEAN REQUEST. IF THIS SELECTION IS CORRECT INSERT  
THE WORD YOU LIKE TO DELETE ELSE PRESS THE ESCAPE KEY.

Δ PROGRAMMER ↴  
\* PLEASE WAIT. THANK YOU!

Slonim, J., Maryanski, F. J., & Fisher, P. S. (1978). Mediator: An integrated approach to information retrieval. *Proceedings of SIGIR*, 14-36.

# Information Intermediary Modeling

## *Information Need for Zentralblatt*

*Information need category:*

- Similar information need previously specified to Euromath
- Topic that you can describe PRECISELY
- Topic that you can only describe VAGUELY
- Specific document(s), e.g. Author known

*Number of documents you want to retrieve:*    From: **10**   To: **30**

*Display formats:*

- Title, Authors
- Title, Authors, Source
- Title, Index Terms
- All fields, including Abstract

*Experience in online retrieval:*    Little    Moderate    Extensive

**Find**

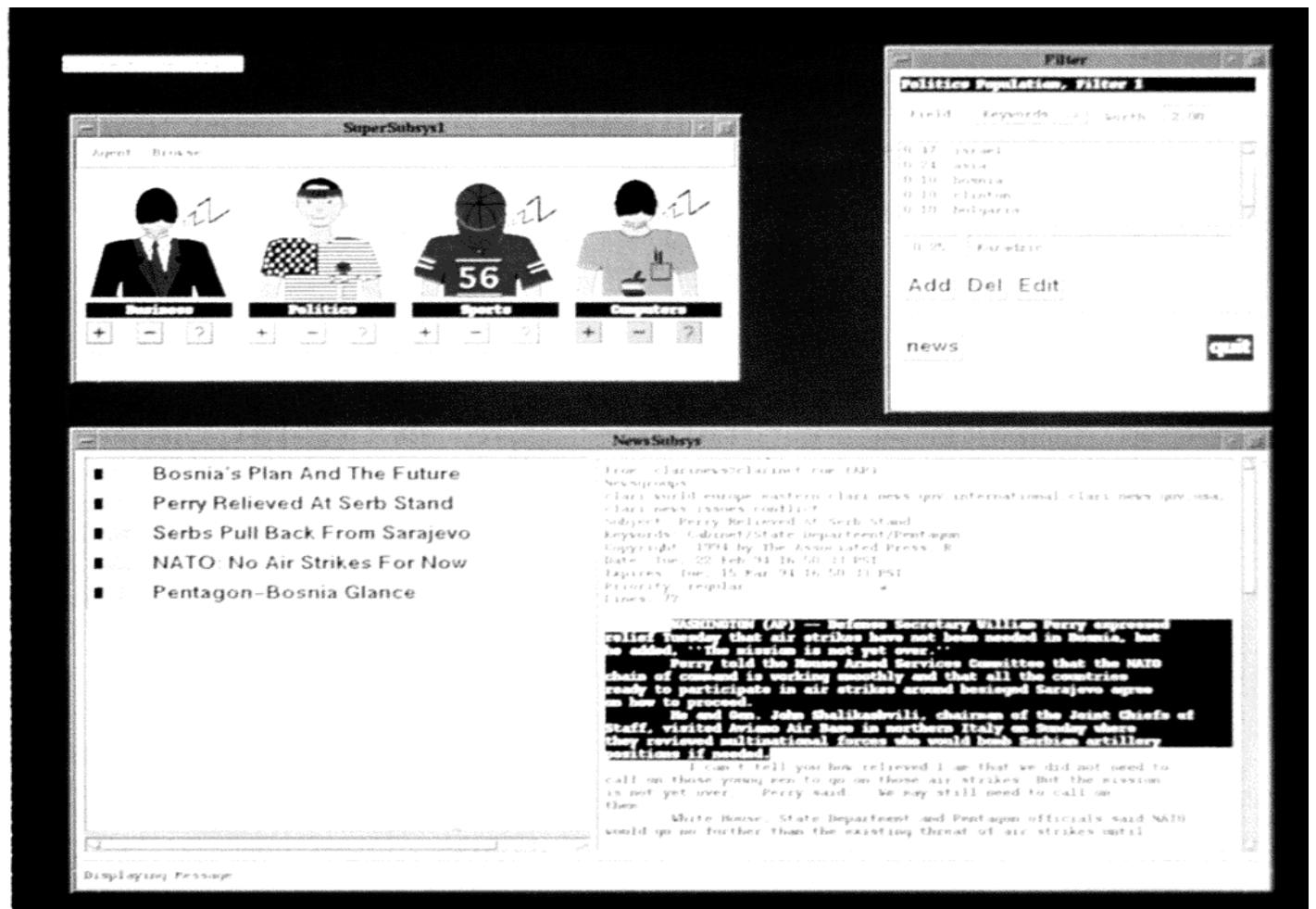
**Cancel**

McAlpine, G. & Ingwersen, P. (1989). Integrated information retrieval in a knowledge worker support system. *ACM SIGIR Forum*, 48-57.

# User Modeling and Profiles

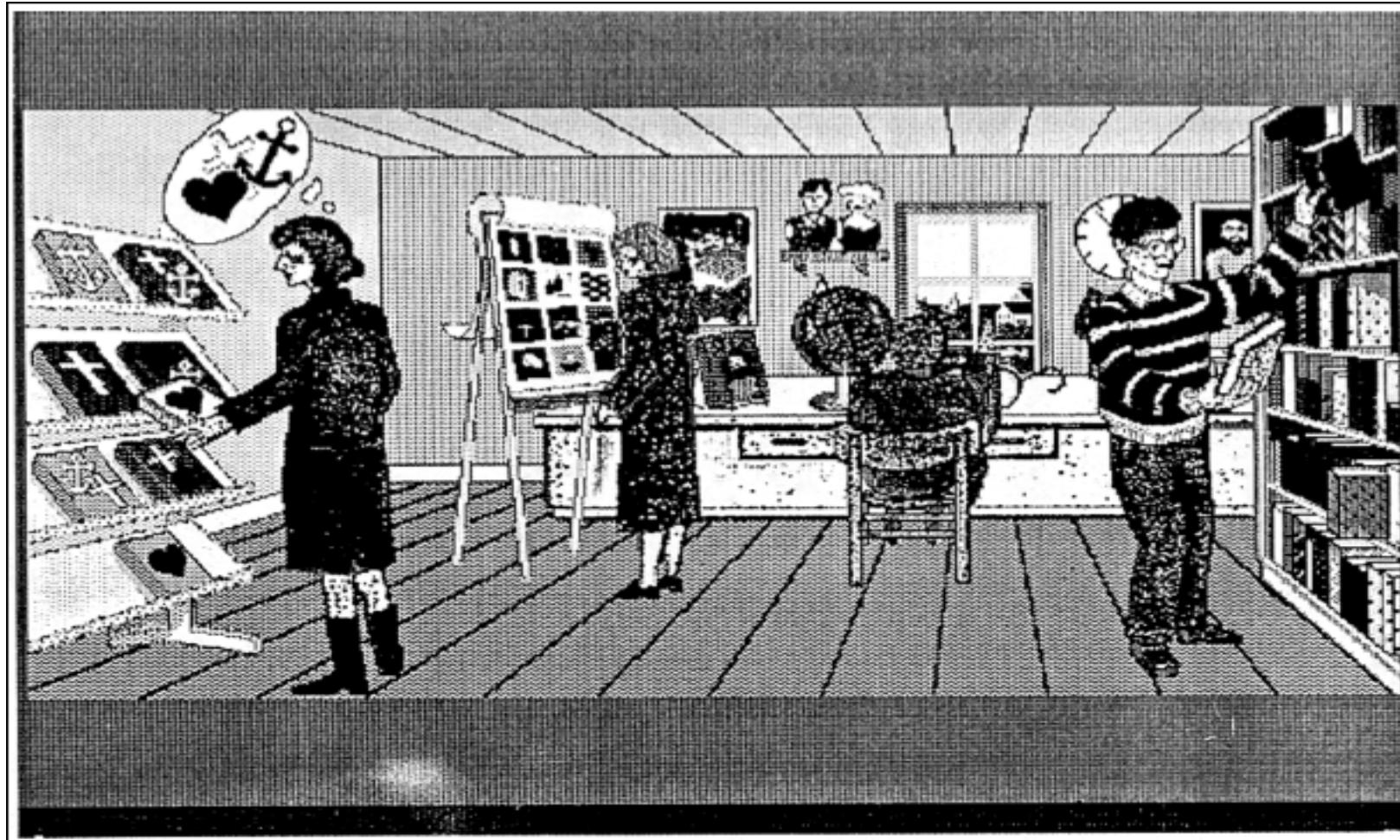
FACET	VALUE	RATING
Activated-by	Athletic-w-trig	
Genl	ANY-PERSON	
Movitivations		
Excite	800	600
Interests		
Sports	900	800
Thrill	5	700
Tolerate-violence	4	600
Romance	-5	500
Education	-2	500
Tolerate-suffering	4	600
Strengths		
Physical-strength	900	900
Perseverance	800	600
<u>SPORTS-PERSON</u>		

Rich, E. (1983). Users are individuals: Individualizing user models. *International Journal of Human-Computer Studies*, 51, 323-338.



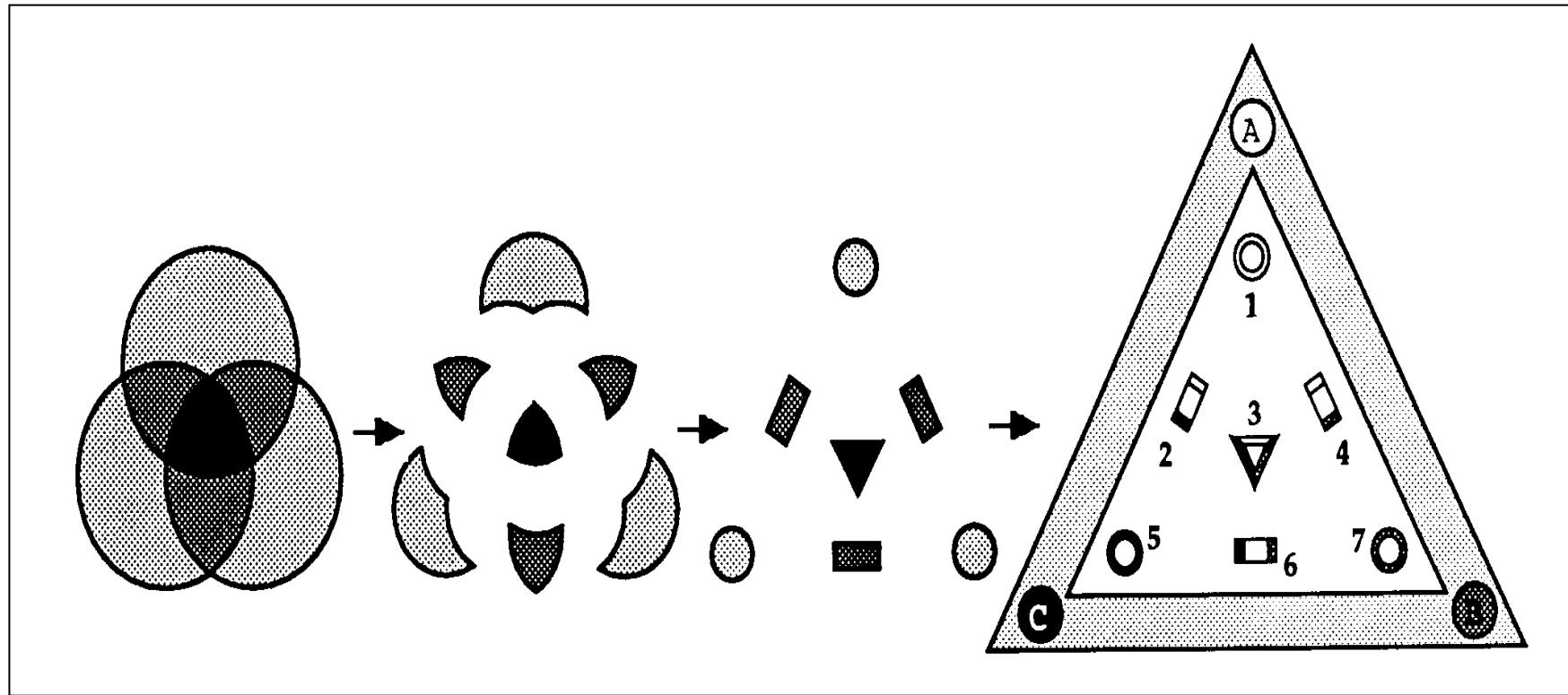
Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 30-40.

# Interactive Spaces



Pejtersen, A. M. (1989). *The BOOK House: Modeling user needs and search strategies as a basis for system design*. Roskilde, Risø National Laboratory. (Risø report M-2794).

# Help with Complex Query Languages



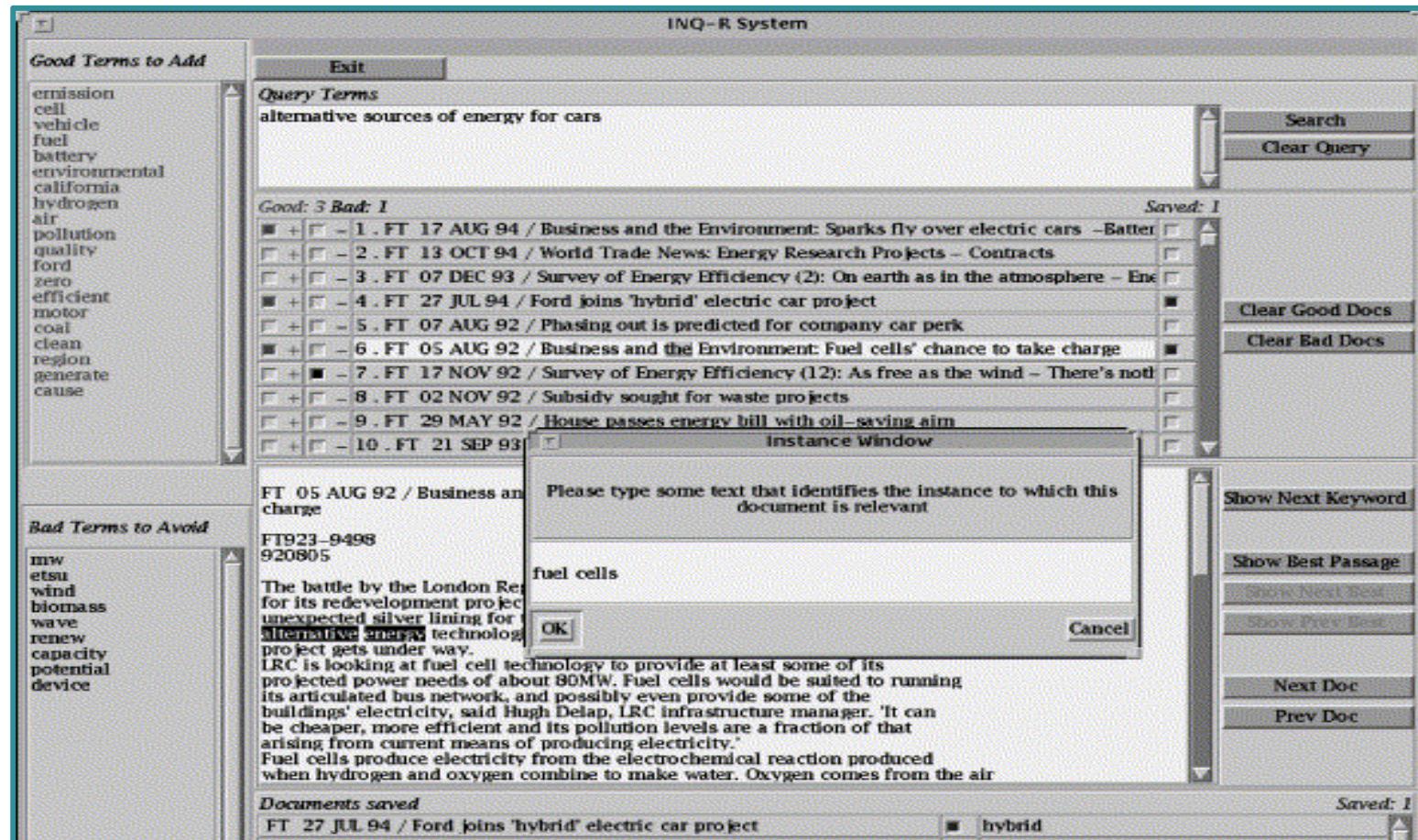
Spoerri, A. (1993). InfoCrystal: A visual tool for information retrieval. *Proceedings of the IEEE Visualization Conference*, 150-157.

# TREC Interactive Track

- Ran from TREC 3 to TREC 12
- Explored a variety of tasks including filtering (query writing), ad-hoc, aspectual recall, fact-finding and topic-distillation
- Most noted for establishing the ‘model’ user study and some guidelines for reporting experiments
- Finding: Difficult to do interactive IR studies in the context of TREC

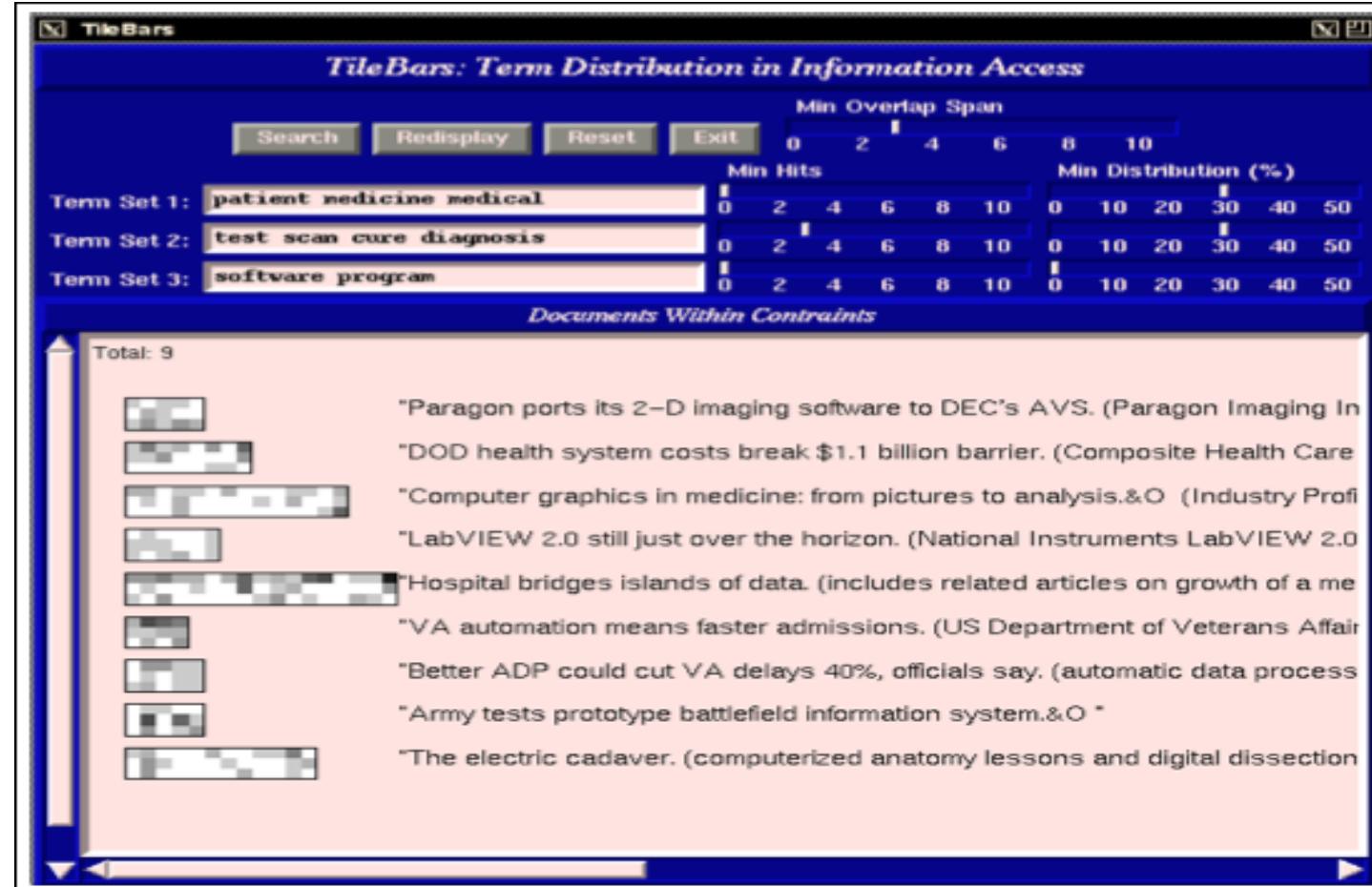
Dumais, S. T. & Belkin, N. J. (2005). The TREC interactive tracks: Putting the user into search. In *TREC: Experiment and Evaluation in Information Retrieval*, (E. M. Voorhees and D. K. Harman, eds.), pp. 123–153, Cambridge, MA: MIT Press, 2005.

# Relevance Feedback



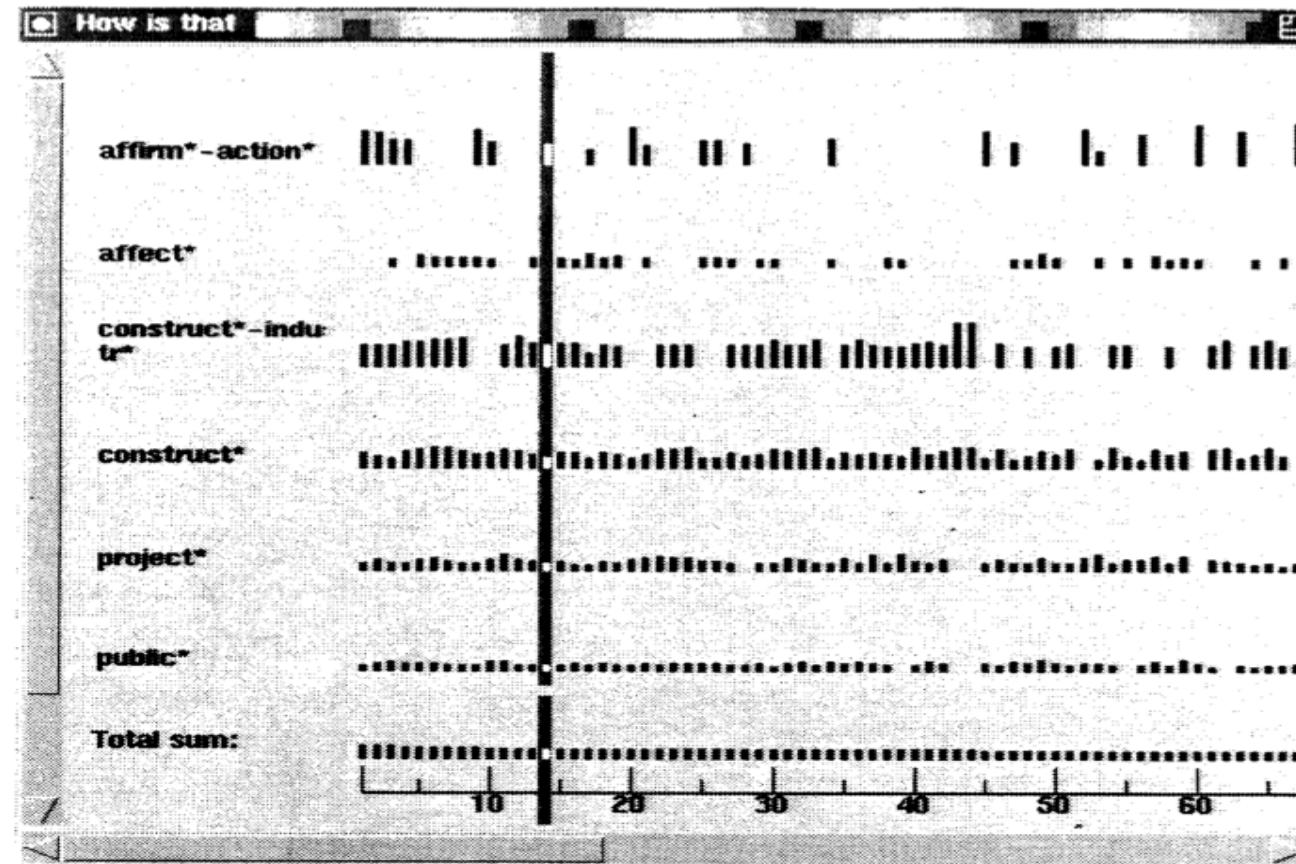
Belkin, N. J., et al. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management* 37(3), 404-434.

# Evaluating Results



Hearst, M. A. (1995). TileBars: Visualization of term distribution in full text information access. *Proceedings of CHI '95*, 59-66.

# Navigating and Comparing Results



Veerasamy, A. & Belkin, N. J. (1996). Evaluation of a tool for visualization of information retrieval results. *Proceedings of SIGIR '96*, 85-92.

# Single View Comparison

**The Wall Street Journal.Hypertext**

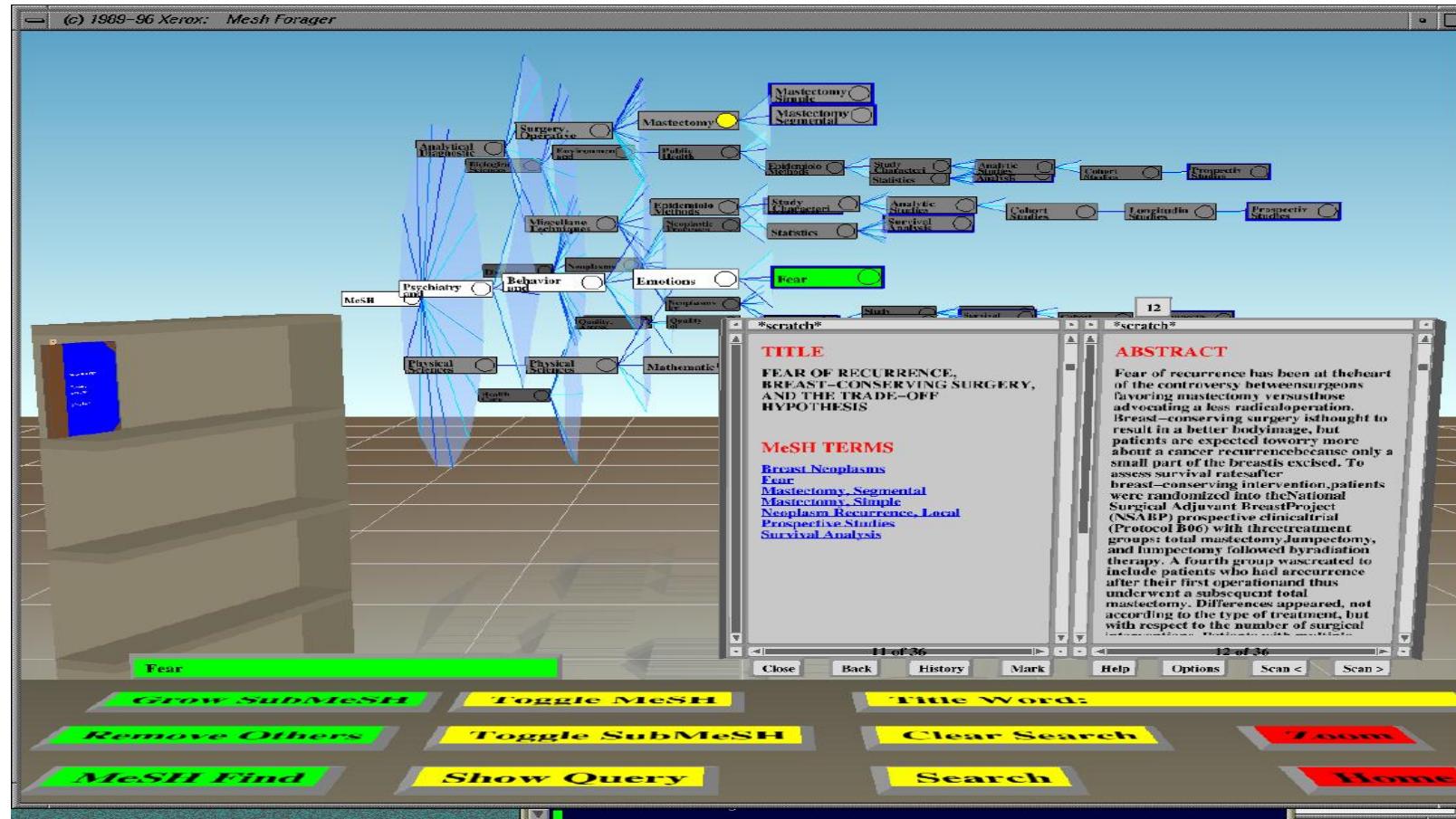
--> "Quebec seeks special status"

<p><b>Related</b></p> <p><b>① Canadian Parties Clear Plan to Rewrite Constitution to Meet Provincial Demands</b></p> <p>By John Uquhart and G. Pierre Goad Staff Reporters of The Wall Street Journal 03/02/92</p> <p>OTTAWA -- Canada's major political parties approved a plan to rewrite the constitution to deal with the longstanding demands of Quebec, other provinces and native Canadians.</p> <p>The plan would give Quebec some or all powers it wants, but whether it can win the necessary broad support in the rest of Canada was uncertain. Most of the constitutional changes proposed in the plan require the approval of the federal government and at least seven of Canada's provinces; some require the consent of all 10 provinces.</p> <p>Quebec Premier Robert Bourassa plans to comment in detail on the federal blueprint tomorrow, a spokeswoman said.</p> <p>At first glance the federal plan falls short of the demands by the ruling Quebec Liberal Party. Opposition leader Jacques Parizeau told reporters in Quebec City that the federal plan is</p>	<p><b>Related</b></p> <p><b>③ International: Canada to Unveil Plan for Changes in Its Constitution</b></p> <p>By John Uquhart Staff Reporter of The Wall Street Journal 09/24/91</p> <p>OTTAWA -- The Canadian government intends to announce today its proposals to rewrite the Canadian constitution, which is being attacked by social activists, provincial governments and native groups.</p> <p>The initiative is expected to provoke fights among political parties and interest groups seeking constitutional reforms. It was drafted following the collapse last year of a constitutional accord that would have recognized Quebec as a "distinct society" within Canada.</p> <p>The government, in a bid to win support for its new plan, has opened up the process to include the constitutional demands of western Canadians, women, natives and other groups, as well as those of Quebec. It also has invited proposals that would broaden its own plan.</p> <p>For example, Ontario's left-wing New Democratic Party government will push for a</p>	<p><b>Related</b></p> <p><b>International: Canada's Leaders Clear a Hurdle In Impasse Involving Constitution</b></p> <p>Tentative Accord Is Reached On Status for Quebec, But Opposition Remains</p> <p>By G. Pierre Goad and John Uquhart Staff Reporters of The Wall Street Journal 06/11/90</p> <p>OTTAWA -- Canada's leaders patched together a tentative agreement that would give Quebec special status and end a bitter constitutional impasse that has revived talk of independence in the French-speaking province.</p> <p>But continuing opposition in English Canada could still kill the deal, setting off a full-blown political crisis. Even if the agreement is ratified, Canada's French-English squabbling won't end soon. The bitter constitutional debate "opened wounds in the national psyche," said Prime Minister Brian Mulroney. After seven days of tough bargaining, Mr. Mulroney and the nation's ten provincial premiers signed the accord Saturday night. "This is a happy day for Canada," Mr. Mulroney said.</p>	<p><b>Related</b></p> <p><b>Mulroney Detects Progress in Talks Over Constitution</b></p> <p>By John Uquhart and G. Pierre Goad Staff Reporters of The Wall Street Journal 06/05/90</p> <p>OTTAWA -- Canadian Prime Minister Brian Mulroney reported a "small degree of progress" on the third day of negotiations on Canada's constitutional impasse.</p> <p>Mr. Mulroney declined to disclose details of the negotiations, which involve himself and 10 provincial government leaders. But government officials said that much of yesterday's negotiations involved differences over a proposed constitutional amendment that would set new rules for changing the federal Senate.</p> <p>The proposed amendment is part of the constitutional accord that was negotiated three</p>
<p><b>Related</b></p> <p><b>② The Americas: English Canadians Get Ready to Say Goodbye to Quebec</b></p> <p>By David Frum 04/05/91</p> <p>What if they gave a constitutional crisis and nobody came?</p> <p>A special commission convened by the Quebec government urged on March 26 that Quebec hold a referendum by 1992 on independence. Even before the commission presented its report, Prime Minister Robert Bourassa -- theoretically the leader of the anti-separatist forces in Quebec -- had issued a constitutional proposal of his own, which called for a massive</p>	<p><b>Related</b></p> <p><b>④ International: Quebec Issues Ultimatum On Political Independence</b></p> <p>Province Proposes Altering Canada's Constitution, Diminishing Its Powers</p> <p>By G. Pierre Goad Staff Reporter of The Wall Street Journal 01/30/91</p> <p>OTTAWA -- Quebec's ruling party said the rest of Canada must agree to a radical constitutional makeover and a broad</p>	<p><b>Related</b></p> <p><b>⑤ Canada Talks Stall After Fifth Day Over Quebec's Bid for Special Status</b></p> <p>By John Uquhart and G. Pierre Goad Staff Reporters of The Wall Street Journal 05/08/90</p> <p>OTTAWA -- Negotiations to resolve Canada's constitutional impasse stalled last night after five days of talks, and several provincial premiers warned that the situation is critical.</p> <p>Canadian Prime Minister Brian</p>	<p><b>Related</b></p> <p><b>⑥ Foreign Exchange: Dollar Is Mixed on Technical Factors; Traders Are Mulling Politics, Economy</b></p> <p>By John Bader Special to The Wall Street Journal 06/06/90</p> <p>NEW YORK -- The dollar was mixed as the market concerned itself primarily with the potential impact of future political and economic developments.</p> <p>The dollar was slightly higher</p>

History  
① "Quebec seeks special status"  
② The Americas: English Canadians Get Ready to Say Goodbye to Quebec  
③ International: Canada to Unveil Plan for Changes in Its Constitution  
④ International: Quebec Issues Ultimatum On Political Independence  
⑤ Canada Talks Stall After Fifth Day Over Quebec's Bid for Special Status  
⑥ Foreign Exchange: Dollar Is Mixed on Technical Factors; Traders Are Mulling Politics, Economy  
⑦ Another Filter  
Topic  
30 matching articles found

Golovchinsky, G. & Chignell, M. H. (1997). The newspaper as an information exploration metaphor. *Information Processing & Management*, 33(5), 663-683.

# Interacting with System Components



Hearst, M. A. & Karadi, C. (1997). Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. *Proceedings of SIGIR '97*.

# Saving and Sorting

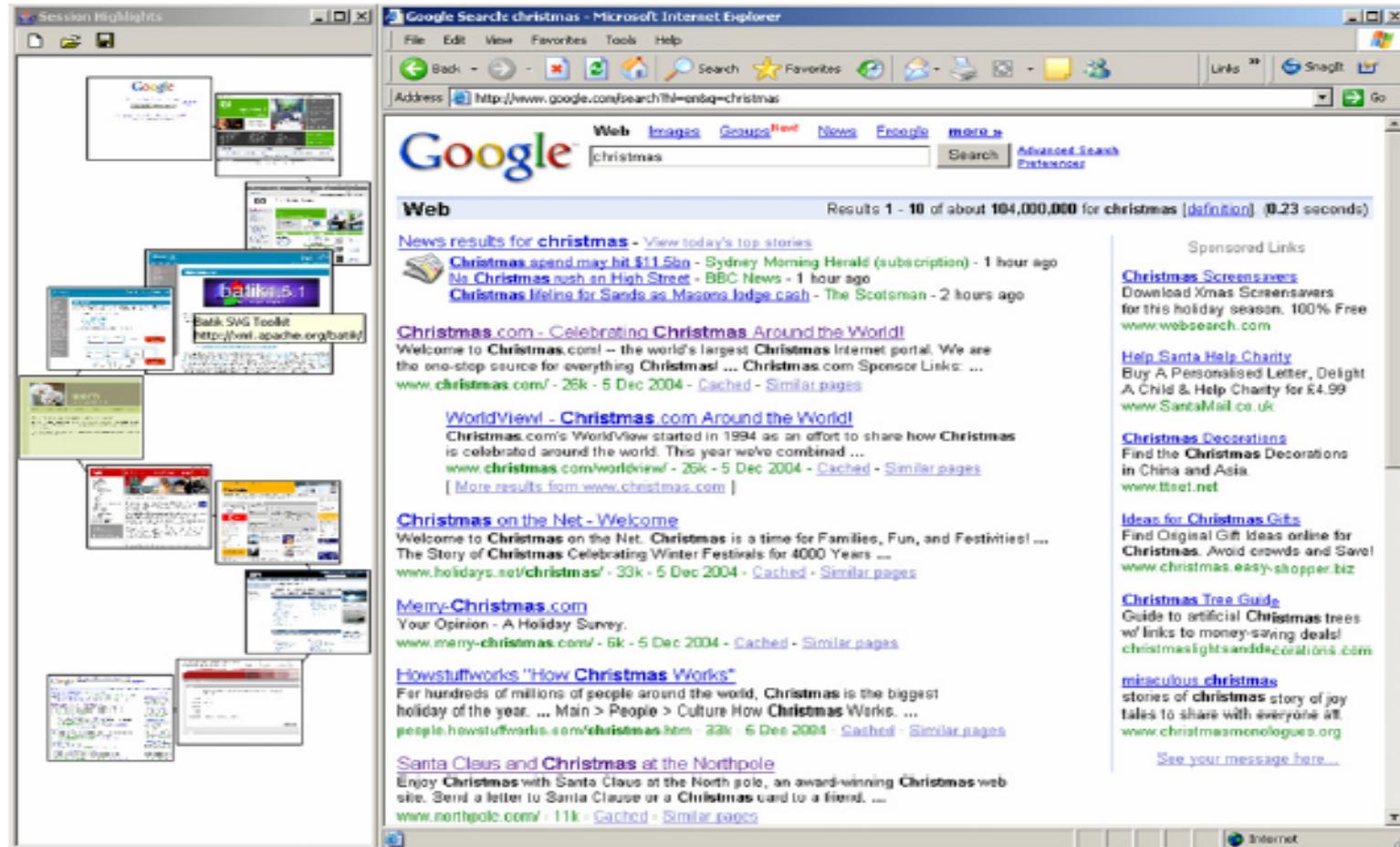


Robertson, et al. (1998). Data Mountain: Using spatial memory for document management. *Proceedings of UIST '98*, 153-162.

# Google Beta

The screenshot shows the Google Beta homepage. At the top is the iconic multi-colored 'Google!' logo with the word 'BETA' underneath it. Below the logo is a search bar with the placeholder text 'Search the web using Google!'. Underneath the search bar are two buttons: 'Google Search' and 'I'm feeling lucky'. The main content area has a teal background. On the left, under 'Special Searches', are links to 'Stanford Search' and 'Linux Search'. In the center, there are links to 'Help!', 'About Google!', 'Company Info', and 'Google! Logos'. On the right, there's a section titled 'Get Google! updates monthly:' with a field for 'your e-mail', a 'Subscribe' button, and an 'Archive' link. At the bottom of the page, the copyright notice 'Copyright ©1998 Google Inc.' is visible.

# Saving and Sorting



Jhaveri, N. & Raiha, K.-J. (2005). The advantages of a cross-session web workspace. *Proceedings of CHI*.

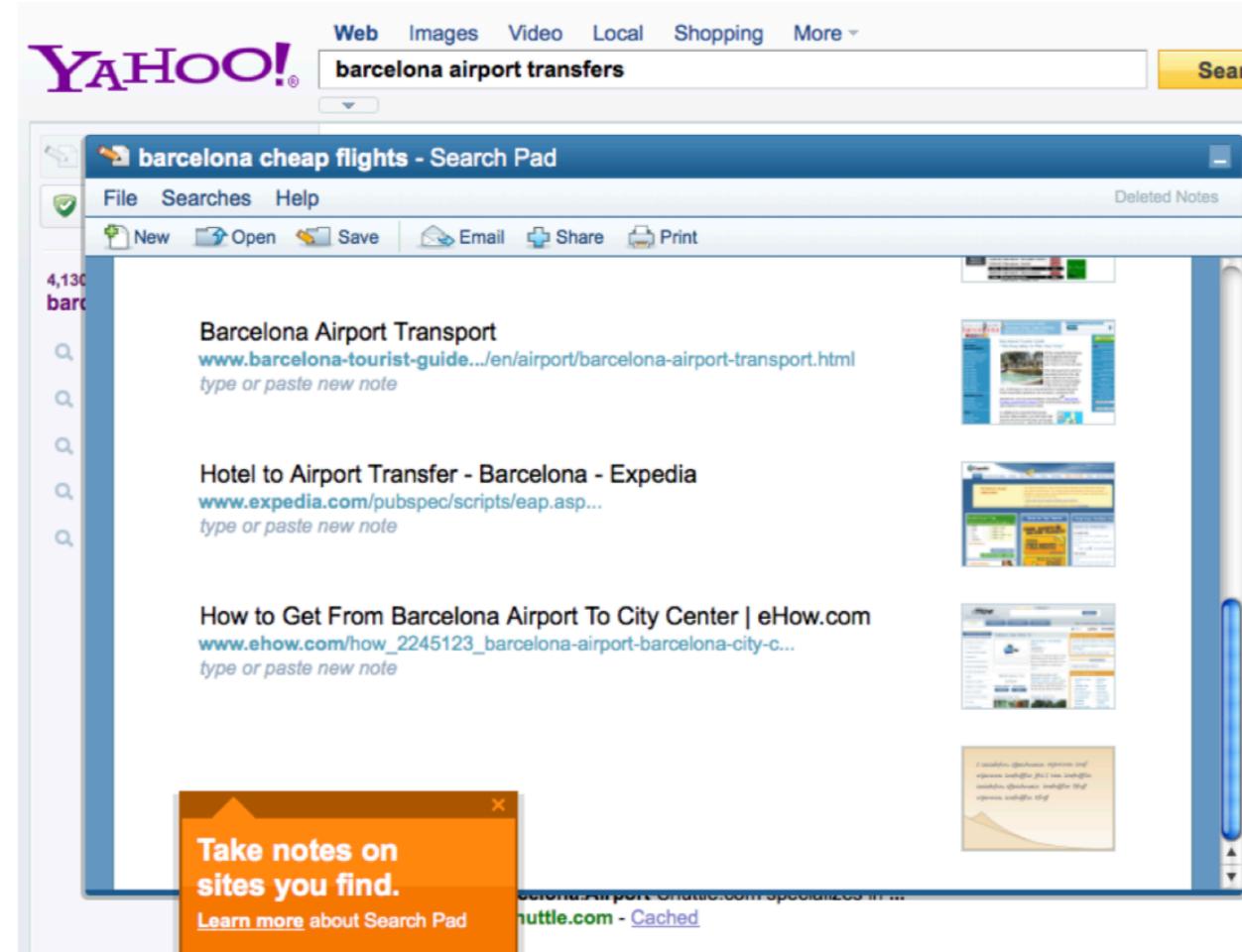
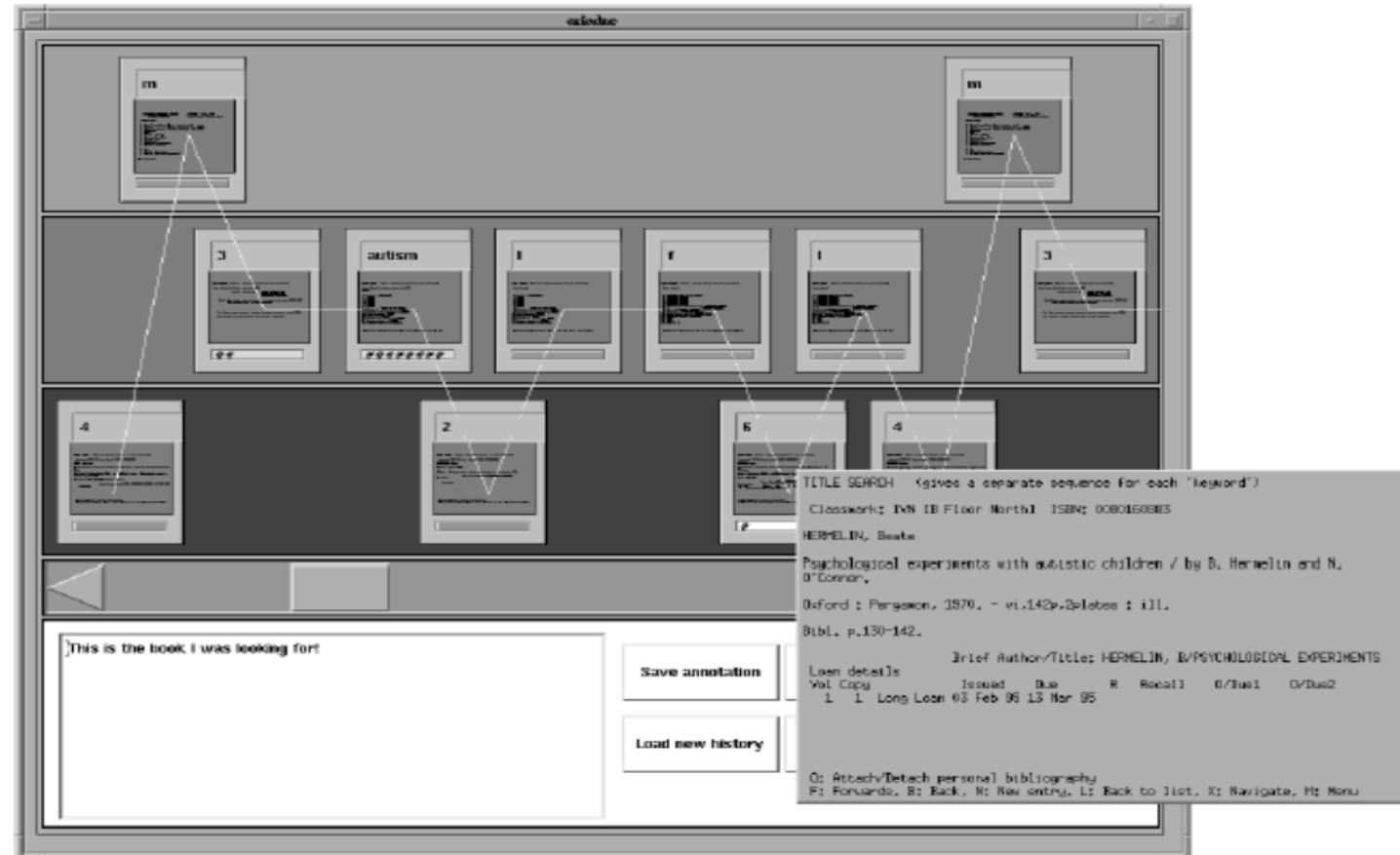


Figure 2: An opened pad object

Donato, D., Bonchi, F., Chi, T., & Maarek, Y. (2010). Do you want to take notes? Identifying research missions in Yahoo! Search Pad. *Proc. WWW Conference*, 321-330.

# Collaborative Search



Twidale, M. B. & Nichols, D. M. (1998). Designing interfaces to support collaboration in information retrieval. *Interacting with Computers*, 10(2), 177-193.



Figure 1. The SearchTogether client. (a) integrating messaging, (b) query awareness, (c) current results, (d) recommendation queue, (e)(f)(g) search buttons, (h) page-specific metadata, (i) toolbar, (j) browser

Morris, M. R., & Horvitz, E. (2007). SearchTogether: An interface for collaborative web search. *Proc. of UIST*, 3-12.

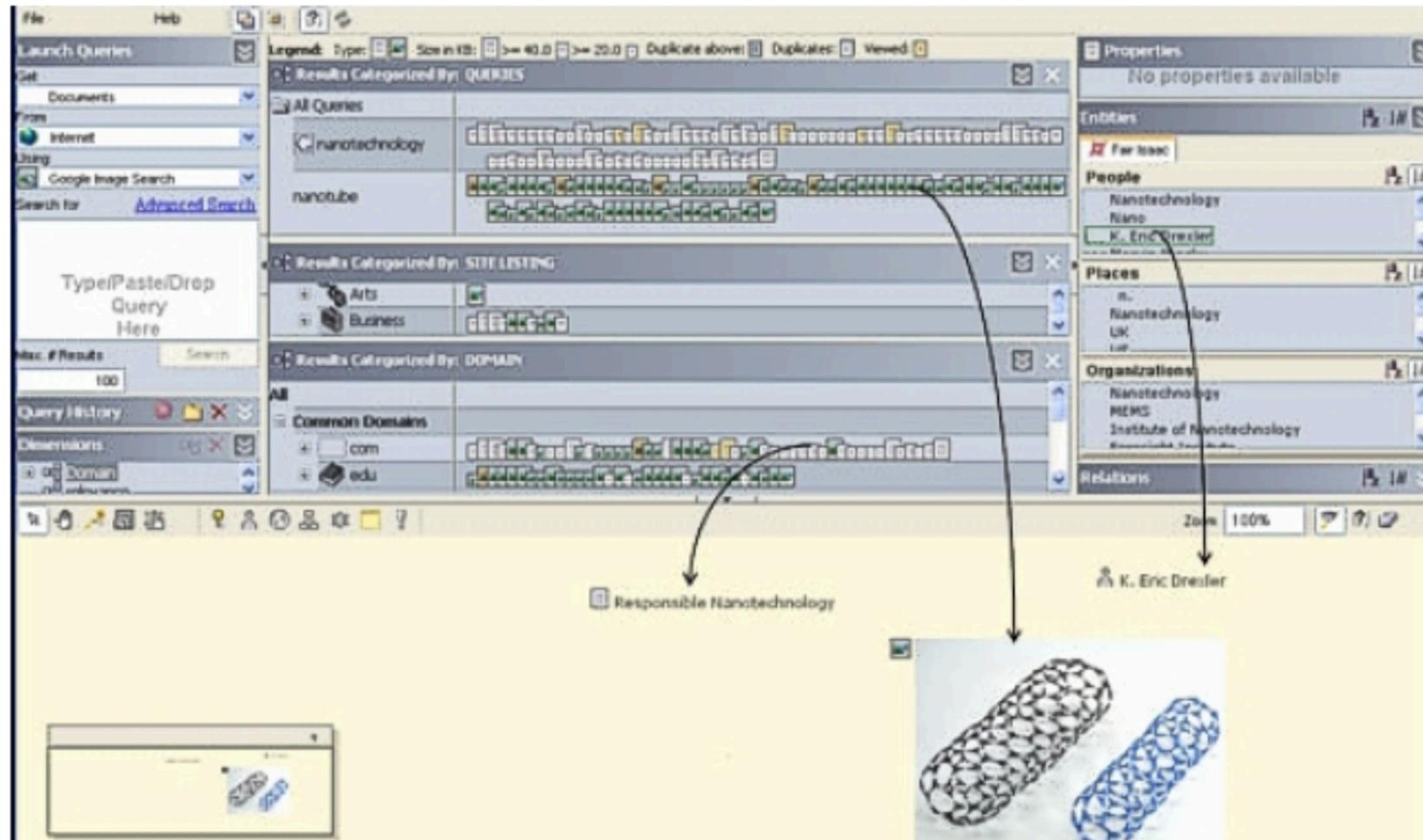
# Integrated Environments

The left screenshot shows the Querium interface, a search environment. It features a sidebar with tasks (e.g., gene's research on Mar 24), views (Search, Summary, Queries, Documents), and filters (Normal [96], All years [100]). The main area displays a 'Current query' search bar ('interactive session-based search query history') and a results list. The results are sorted by rank and include various academic papers and reports, such as 'The Use of Relevance Feedback on the Web: Implications for Web IR System Design' and 'Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web'.

The right screenshot shows a web browser window titled 'Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web'. The page content discusses user queries on the Web, mentioning Bernard J. Jansen, Armand Spink, and Tekla Saracevic. It includes sections on abstracts, introductions, and detailed analyses of user query logs from Excite. The URL is <http://jimjansen.tripod.com/academic/pubs/webnet99.pdf>.

Golovchinsky, G., Biriye, A., & Dunnigan, T. (2012). The future is in the past: Designing for exploratory search. *Proceedings of IliX '12*.

# Specialized Audiences



Wright, et al. (2006). The Sandbox for analysis-concepts and methods. *Proceedings of SIGCHI Conference*.

# Persuading & Nudging People to Change



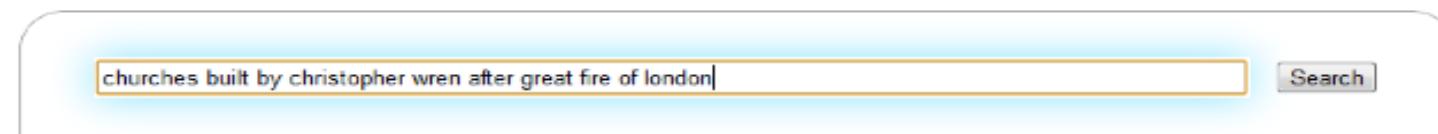
**Figure 1. Empty query box.**



**Figure 2. As the person starts to type, the halo changes.**



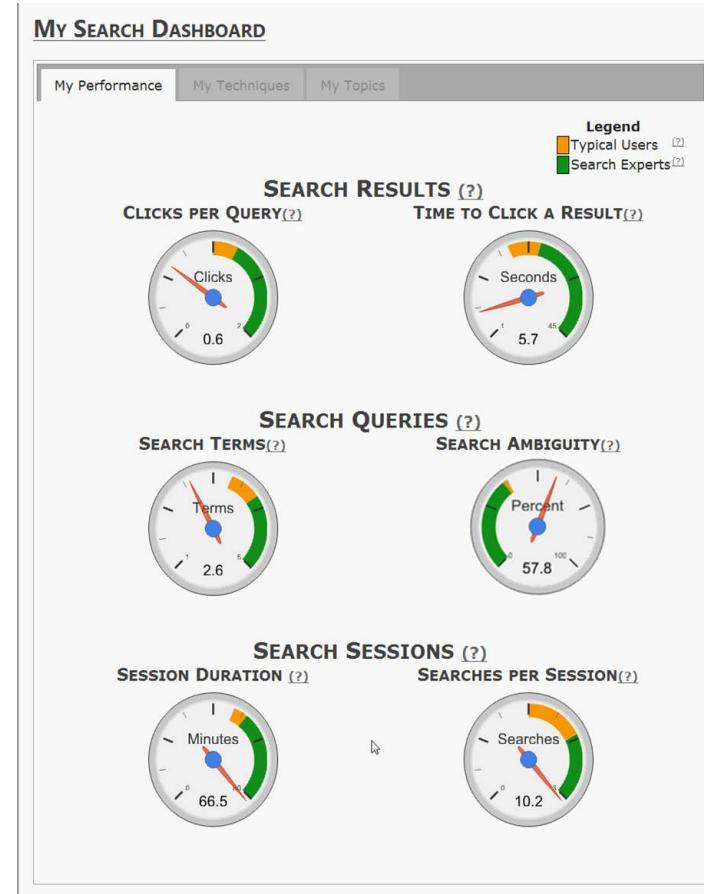
**Figure 3. A longer query with a bluer halo.**



**Figure 4. A long query with a bluish halo.**

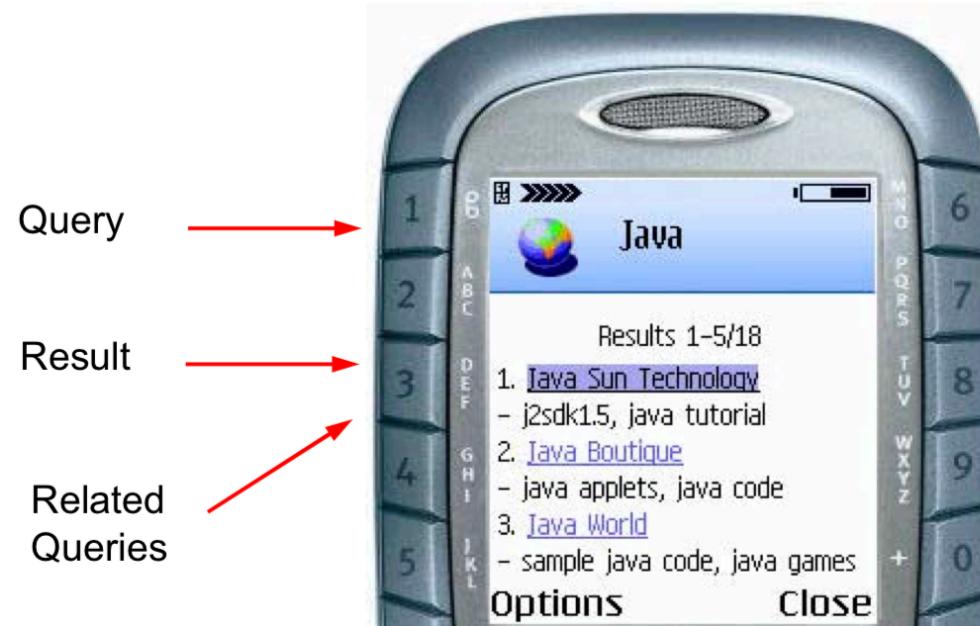
Agapie, E., Golovchinsky, G. & Qvardordt, P. (2012). Encouraging behavior: A foray into persuasive computing. *Proc. of HCIR*.

# Reflective Practice & Learning



Bateman, S., Teevan, J., & White, R. W. (2012). The search dashboard: How reflection and comparison impact search behavior. *Proceedings of CHI '12*, Austin, TX, 1785-1794.

# Mobile Search



**Figure 2: Illustration of the Results and Related Queries Generated for the Query ‘Java’ on a Mobile Phone**

Church, K., Smyth, B., & Keane, M. T. (2006). Evaluating interfaces for intelligent mobile search. *Proc. of the 2006 International Cross-disciplinary workshop on Web accessibility (W4A): Building the mobile web: rediscovering accessibility?* 69-78.

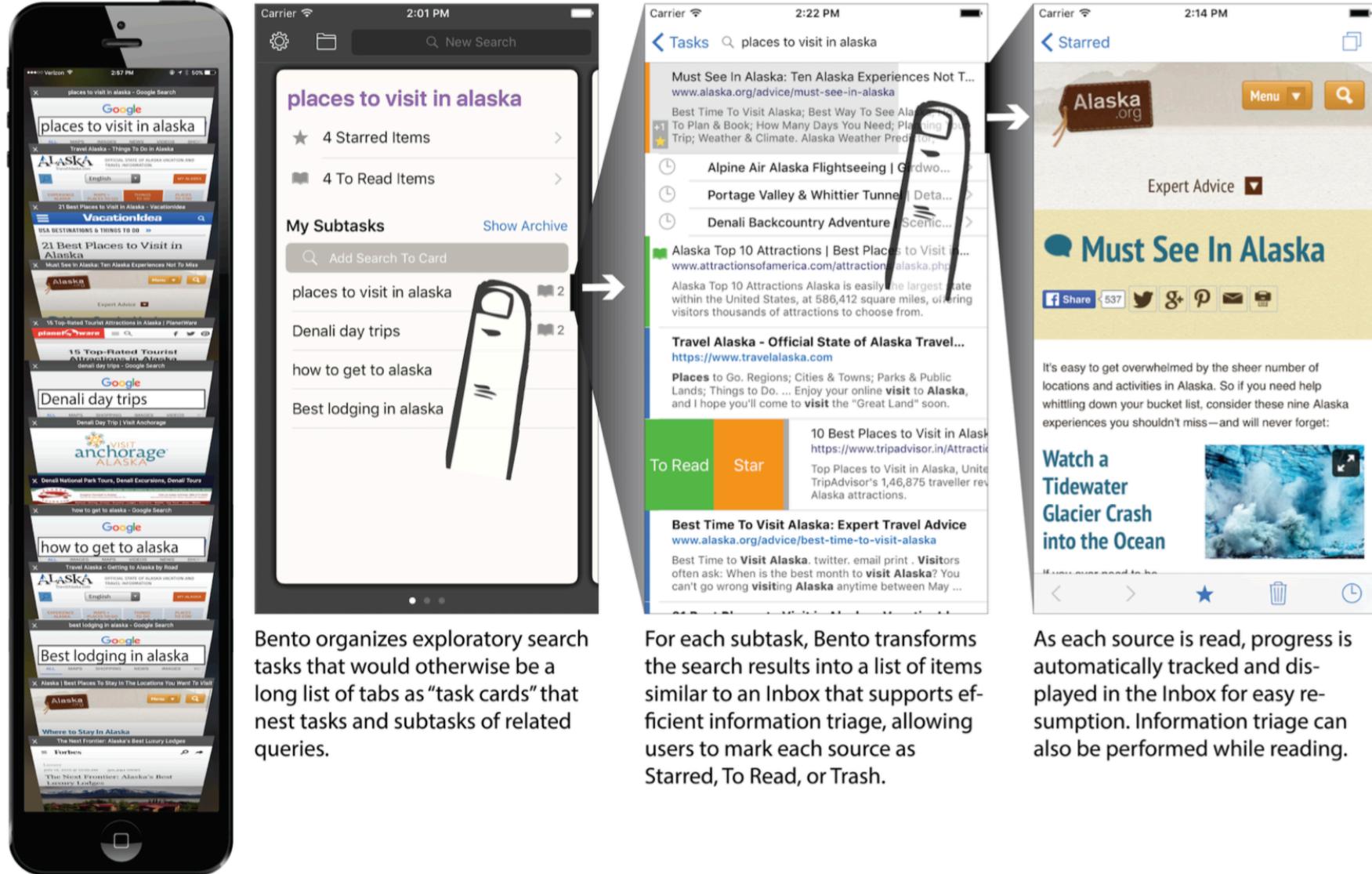


Figure 1. Comparing a typical list of tabs (left) with Bento's search centered navigation from the same exploratory search task.

Hahn, N., Chang, J. C., & Kittur, A. (2018). Bento Browser: Complex mobile search without tabs. *Proc. of ACM CHI*, paper 251.

# Proactive Search during Conversations

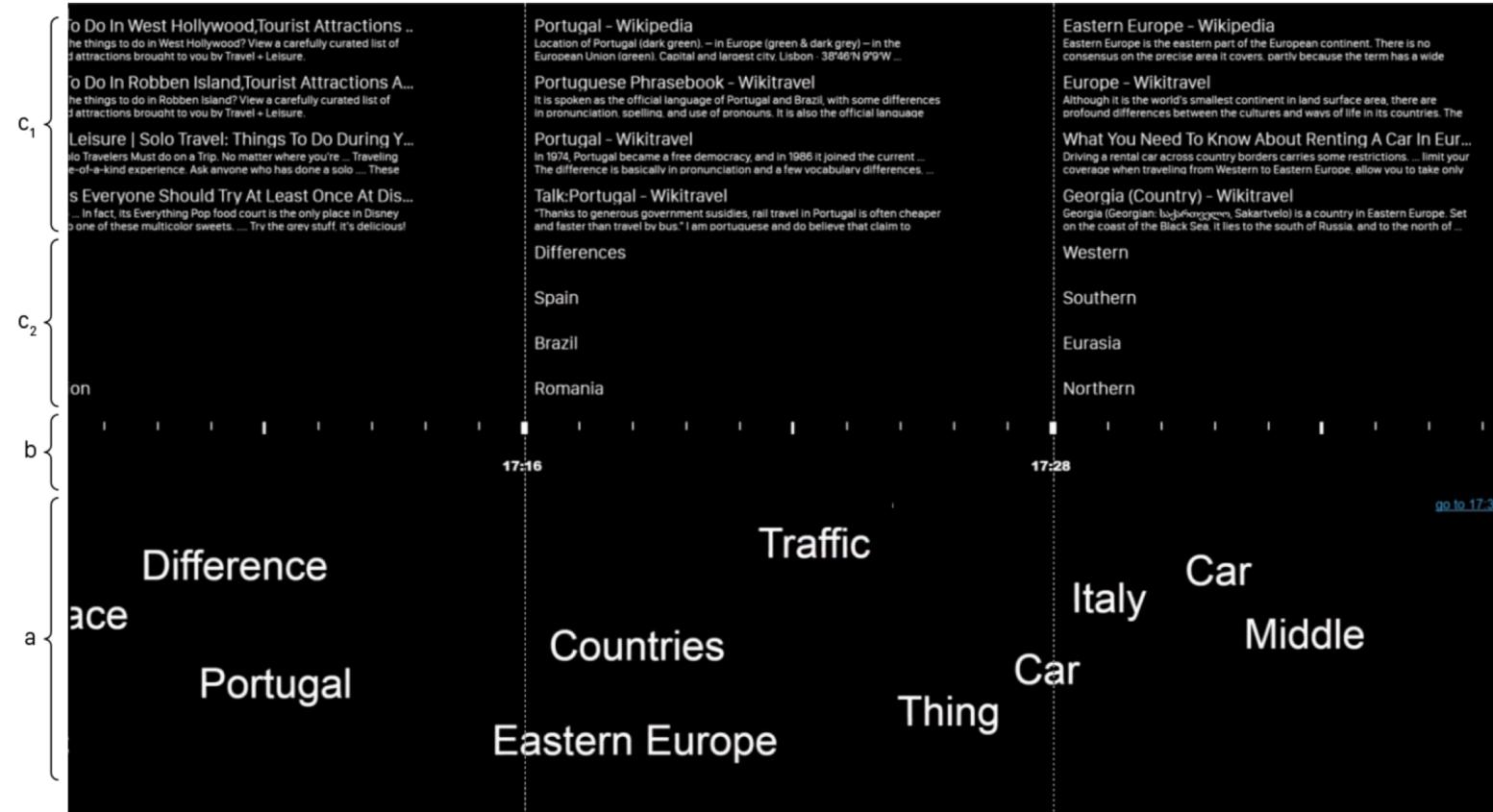
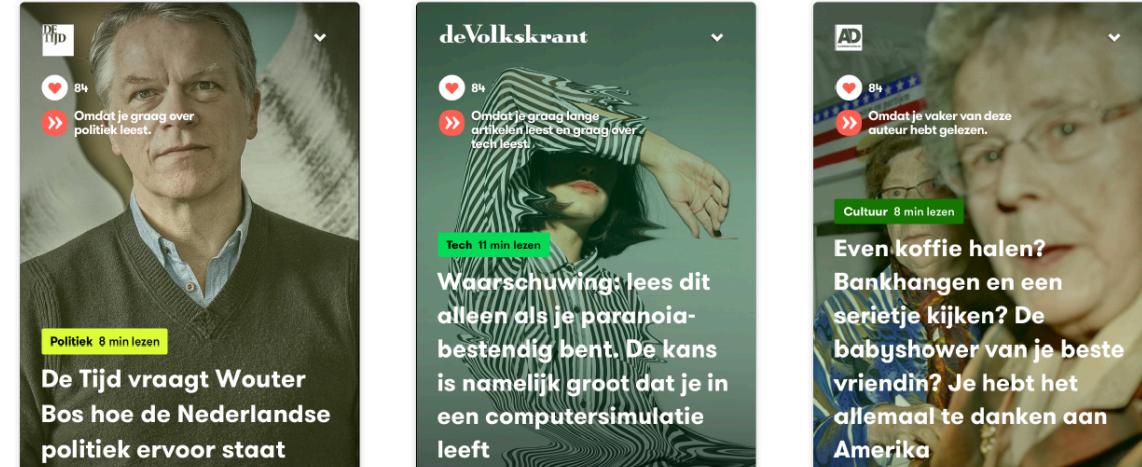


Figure 2. The user interface of the SearchBot system. The system monitors a conversation and provides continuous recommendations of related documents and entities in a non-intrusive way. a) Stream of recognized entities; b) timescale with timecodes; c<sub>1</sub>) recommended documents; c<sub>2</sub>) recommended entities.

# Explanations

Maartje Ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, and Maarten de Rijke. Do News Consumers Want Explanations for Personalized News Rankings?. In *FATREC Workshop on Responsible Recommendation*, August 2017.



(a) Single reason, visible – "Because you like reading about politics."  
 (b) Single reason, invisible – "Because you like long reads and tech."  
 (c) Multiple reasons, visible – "Because you often read from this author."

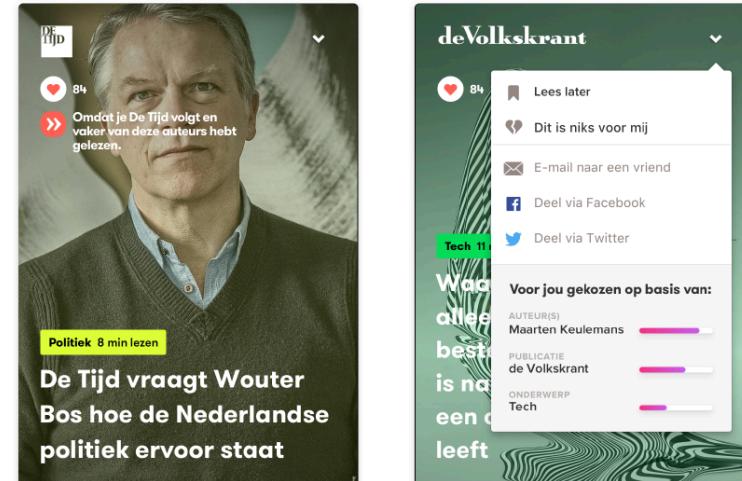


Figure 1: Examples of reason types as shown to users in our user study. Textual reasons are in the lines that start with "Omdat" (because). For the bar chart layout the reasons starts with "Voor jou gekozen" (selected for you). Translations are given below each article.

# User-Centered Evaluation

# Different Types of Methods

	Observational	Experimental
<b>Lab Studies</b> <i>Controlled interpretation of behavior with detailed instrumentation</i>	In-lab behavior observations	In-lab controlled tasks, comparison of systems
<b>Field Studies</b> <i>In the wild, ability to probe for detail</i>	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
<b>Log Studies</b> <i>In the wild, little explicit feedback but lots of implicit signals</i>	Logs from a single system	A/B testing of alternative systems or algorithms

Table 1. Different types of user data in HCI research.

Dumais, S., Jeffries, R., Russell, D. M., Tang, D. & Teevan, J. (2014). Understanding user behavior through log data and analysis. J.S. Olson and W. Kellogg (Eds.), *Human Computer Interaction Ways of Knowing*. New York: Springer.

# User Studies in Controlled Settings

- Allows the researcher to interact with users to better understand their experiences.
- Allows for the use of a wider-variety of data collection techniques, including:
  - Logs, scales, questionnaires
  - Interviews, think-aloud
  - Eye-tracking, physiological data, EEG, fMRI
- Enables manipulation, control and isolation of effects

# What is an experiment?

“By experiment we refer to that portion of research in which variables are **manipulated** and their effects upon other variables are **observed**.”

(Campbell & Stanley, 1963, p. 1)

- Manipulated variables: independent variables.
- Observed variables: dependent variables.

# What is an experiment?

“By experiment we refer to that portion of research in which variables are manipulated and their **effects** upon other variables are observed.”

(Campbell & Stanley, 1963, p. 1)

- Experiments are conducted to evaluate **causal** relationships.

# Simplest Case: Interface Evaluation

Non-blended Interface

**Web**

[Portland, Oregon Tourist & Vacation Information ? Travel Portland](#)  
Travel Portland is your resource for Portland vacations, meeting planning and more. Discover Portland events, restaurants, hotel deals and more.  
<http://www.travelportland.com/>

**Images**

[Things to do in Portland, Oregon ? Travel Portland](#)  
Explore Portland Oregon. Find things to do, places to see, our top attractions and activities. Travel Portland offers a comprehensive visitor guide to ...  
<http://www.travelportland.com/things-to-see-and-do/things-to-see-and-do-home>

**Videos**

[Visiting - City of Portland, Oregon](#)  
Learn about events, attractions, hotels, and more at Travel Portland. See what Downtown ... © 2012 City of Portland, Oregon Privacy Policy - Accessibility: ...  
<http://www.portlandonline.com/index.cfm?c=25782>

**News**

[Portland Tourism and Vacations: 403 Things to Do in Portland, OR ...](#)  
Portland Tourism: TripAdvisor has 42,178 reviews of Portland Hotels, Attractions, and Restaurants making it your best Portland Vacation resource ...  
[http://www.tripadvisor.com/Tourism-g52024-Portland\\_Oregon-Vacations.html](http://www.tripadvisor.com/Tourism-g52024-Portland_Oregon-Vacations.html)

**Blogs**

[Portland, Oregon - Wikipedia, the free encyclopedia](#)  
The Portland metropolitan area in the U.S. states of Oregon and Washington has a variety of tourist attractions. 24 Hour Church of Elvis, exhibit and museum ...  
[http://en.wikipedia.org/wiki/Tourism\\_in\\_Portland,\\_Oregon](http://en.wikipedia.org/wiki/Tourism_in_Portland,_Oregon)

**Q & A**

[Portland, Oregon Vacations, Tourism, Guides, Hotels, Things to Do ...](#)  
Portland is an eclectic city, where sophisticated and alternative styles coexist peacefully. It is known for its friendliness, rich culture and variety of outdoor ...  
[http://travel.yahoo.com/p/travelguide-191501995-portland\\_vacations-](http://travel.yahoo.com/p/travelguide-191501995-portland_vacations-)

**Shopping**

[Greater Portland - Travel Oregon](#)  
Portland has been described as America's most European city. If that means a great walking city with tons of public transportation, a progressive atm ...  
<http://traveloregon.com/cities-regions/greater-portland/>

Vs.

Blended Interface

**Everything**

**Web**

[Portland, Oregon Tourism & Vacation Information ? Travel Portland](#)  
Travel Portland is your resource for Portland vacations, meeting planning and more. Discover Portland events, restaurants, hotel deals and more.  
<http://www.travelportland.com>

**Images**

[Things to do in Portland, Oregon ? Travel Portland](#)  
Explore Portland Oregon. Find things to do, places to see, our top attractions and activities. Travel Portland offers a comprehensive visitor guide to ...  
<http://www.travelportland.com/things-to-see-and-do/things-to-see-and-do-home>

**Videos**

[Visit - City of Portland, Oregon](#)  
Learn about events, attractions, hotels, and more at Travel Portland. See what Downtown ... © 2012 City of Portland, Oregon Privacy Policy - Accessibility: ...  
<http://www.portlandonline.com/index.cfm?c=25782>

**News**

**Shopping**

**Image Results**

**Video Results**

**Web**

[Portland, Oregon Tourism & Vacation Information ? Travel Portland](#)  
Portland Tourism: TripAdvisor has 42,178 reviews of Portland Hotels, Attractions, and Restaurants making it your best Portland Vacation resource ...  
[http://www.tripadvisor.com/Tourism-g52024-Portland\\_Oregon-Vacations.html](http://www.tripadvisor.com/Tourism-g52024-Portland_Oregon-Vacations.html)

[Portland, Oregon - Wikipedia, the free encyclopedia](#)  
The Portland metropolitan area in the U.S. states of Oregon and Washington has a variety of tourist attractions. 24 Hour Church of Elvis, exhibit and museum ...  
[http://en.wikipedia.org/wiki/Tourism\\_in\\_Portland,\\_Oregon](http://en.wikipedia.org/wiki/Tourism_in_Portland,_Oregon)

[Portland, Oregon Vacations, Tourism, Guides, Hotels, Things to Do ...](#)  
Portland is an eclectic city, where sophisticated and alternative styles coexist peacefully. It is known for its friendliness, rich culture and variety of outdoor ...  
[http://travel.yahoo.com/p/travelguide-191501995-portland\\_vacations-](http://travel.yahoo.com/p/travelguide-191501995-portland_vacations-)

**News Results**

[Travel Oregon launches Chinese language site](#)  
Travel Oregon has launched a Chinese language website catering to one of the state's fastest growing visitor markets. The new site offers potential visitors a raft ...  
[http://www.visitoregon.org/News/Travel\\_Oregon\\_launches\\_Chinese\\_language\\_site](http://www.visitoregon.org/News/Travel_Oregon_launches_Chinese_language_site)

**Blog Results**

[Film & Tourism: A Dynamic Duo for Oregon's Economy - blog\\*spot](#)  
This post provide highlights Oregon's tourism industry, and discusses growth throughout the state. Lodging revenue picked up in 2011, increasing 5.4 percent over ...  
<http://katherinewilliams30.tumblr.com/post/1000000000000000000/film-tourism-a-dynamic-duo-for-oregons-economy-blog-spot>

[Portland economy subspacess rest of state](#)  
The Business Journal 21 hours ago  
[http://www.bizjournals.com/oregon/article/1000000000000000000/portland\\_economy\\_subspacess\\_rest\\_of\\_state.html](http://www.bizjournals.com/oregon/article/1000000000000000000/portland_economy_subspacess_rest_of_state.html)

[Appeals court allows lawsuit regarding no-fly list to proceed in Oregon](#)  
<http://www.usatoday.com/story/travel/oregon/2012/07/10/oregon-no-fly-list-suit/5604111>

[Washington Post 4 days ago](#)

[Portland economy subspacess rest of state](#)  
The Business Journal 21 hours ago  
[http://www.bizjournals.com/oregon/article/1000000000000000000/portland\\_economy\\_subspacess\\_rest\\_of\\_state.html](http://www.bizjournals.com/oregon/article/1000000000000000000/portland_economy_subspacess_rest_of_state.html)

[Appeals court allows lawsuit regarding no-fly list to proceed in Oregon](#)  
<http://www.usatoday.com/story/travel/oregon/2012/07/10/oregon-no-fly-list-suit/5604111>

[Washington Post 4 days ago](#)

**Blog**

[Portland, Oregon: A Dynamic Duo for Oregon's Economy - blog\\*spot](#)  
This post provide highlights Oregon's tourism industry, and discusses growth throughout the state. Lodging revenue picked up in 2011, increasing 5.4 percent over ...  
<http://katherinewilliams30.tumblr.com/post/1000000000000000000/film-tourism-a-dynamic-duo-for-oregons-economy-blog-spot>

[Portland economy subspacess rest of state](#)  
The Business Journal 21 hours ago  
[http://www.bizjournals.com/oregon/article/1000000000000000000/portland\\_economy\\_subspacess\\_rest\\_of\\_state.html](http://www.bizjournals.com/oregon/article/1000000000000000000/portland_economy_subspacess_rest_of_state.html)

[Appeals court allows lawsuit regarding no-fly list to proceed in Oregon](#)  
<http://www.usatoday.com/story/travel/oregon/2012/07/10/oregon-no-fly-list-suit/5604111>

[Washington Post 4 days ago](#)

**Blog**

[Portland, Oregon: A Dynamic Duo for Oregon's Economy - blog\\*spot](#)  
This post provide highlights Oregon's tourism industry, and discusses growth throughout the state. Lodging revenue picked up in 2011, increasing 5.4 percent over ...  
<http://katherinewilliams30.tumblr.com/post/1000000000000000000/film-tourism-a-dynamic-duo-for-oregons-economy-blog-spot>

[Portland economy subspacess rest of state](#)  
The Business Journal 21 hours ago  
[http://www.bizjournals.com/oregon/article/1000000000000000000/portland\\_economy\\_subspacess\\_rest\\_of\\_state.html](http://www.bizjournals.com/oregon/article/1000000000000000000/portland_economy_subspacess_rest_of_state.html)

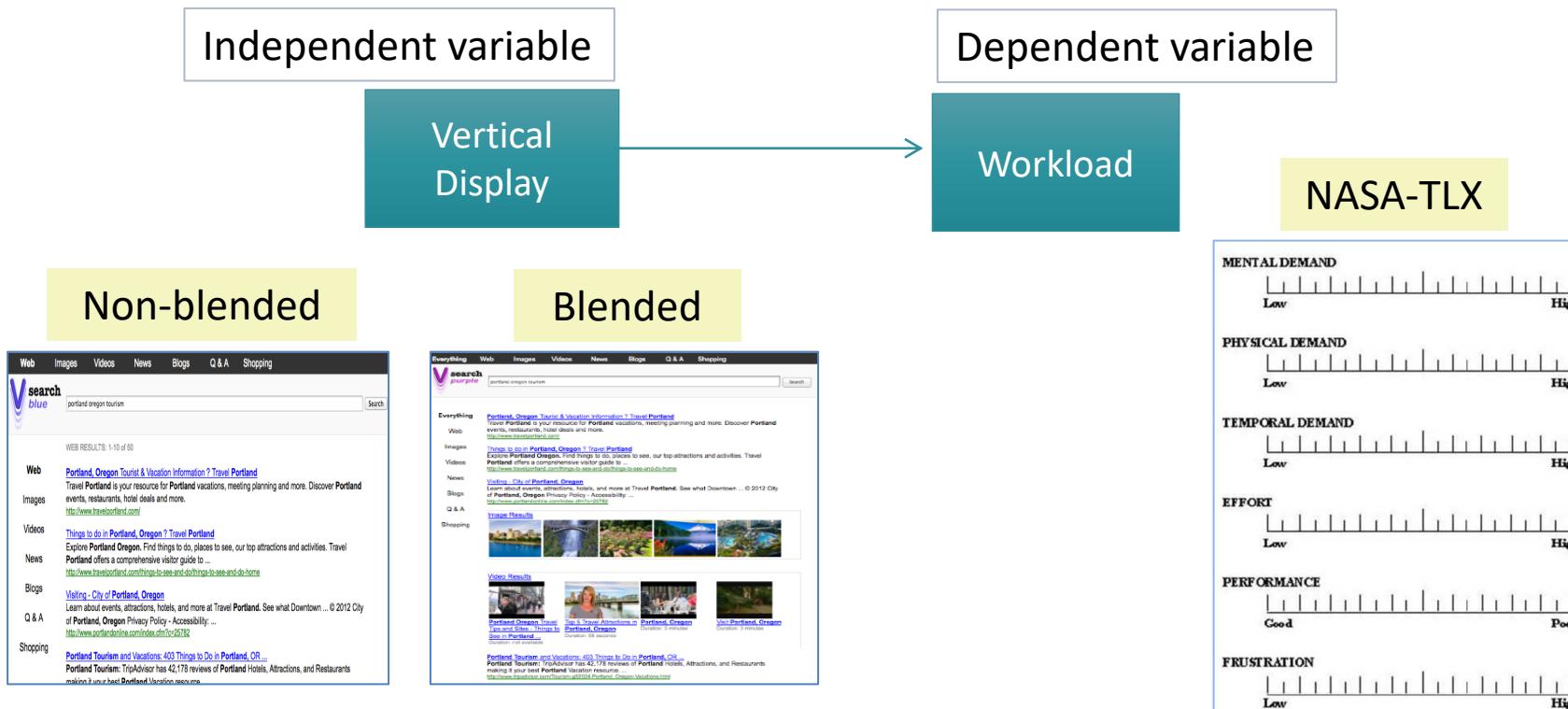
[Appeals court allows lawsuit regarding no-fly list to proceed in Oregon](#)  
<http://www.usatoday.com/story/travel/oregon/2012/07/10/oregon-no-fly-list-suit/5604111>

[Washington Post 4 days ago](#)

Arguello, J., Wu, W.C., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search. *Proceedings SIGIR '12*, 435-444.

# Simplest Case: Interface Evaluation

How does **vertical display** impact the **workload** experienced by people during search?



# Field and Lab Experiments

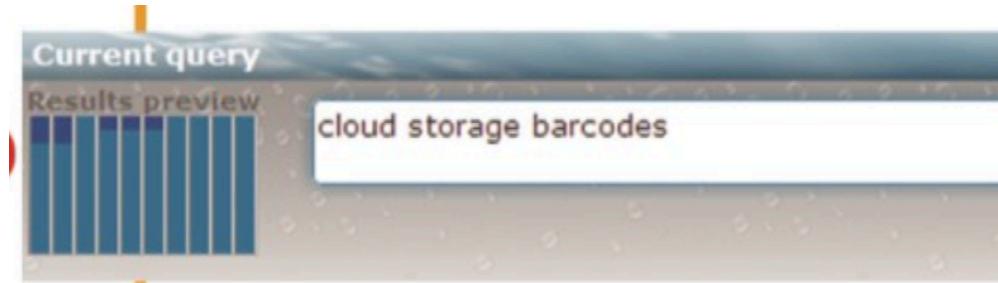
- In a **field experiment**, researchers conduct a live experiment ‘in-the-wild’ and participants often do not know they are participating.
  - Example: AB Test
  - Note: Using Mechanical Turk is not a field experiment
- Field experiments often have better **ecological validity** than laboratory experiments, but the researcher has far **less control**, which makes it harder to isolate cause and effect.

# Lab Experiments are Useful For ...

- Examining search behavior and user differences
  - How does task complexity impact search behavior?
  - How does time pressure impact search behavior?
- Examining the relevance judgment process
  - How does the presentation order of documents impact a person's relevance assessments?
- Evaluating interfaces and systems
  - Is my new interface any good?
  - Is my query suggestion technique any good?
- Testing theory
  - To what extent does Information Foraging Theory explain people's stopping behaviors during search?

# Example: Interface Evaluation

Experimental Treatment

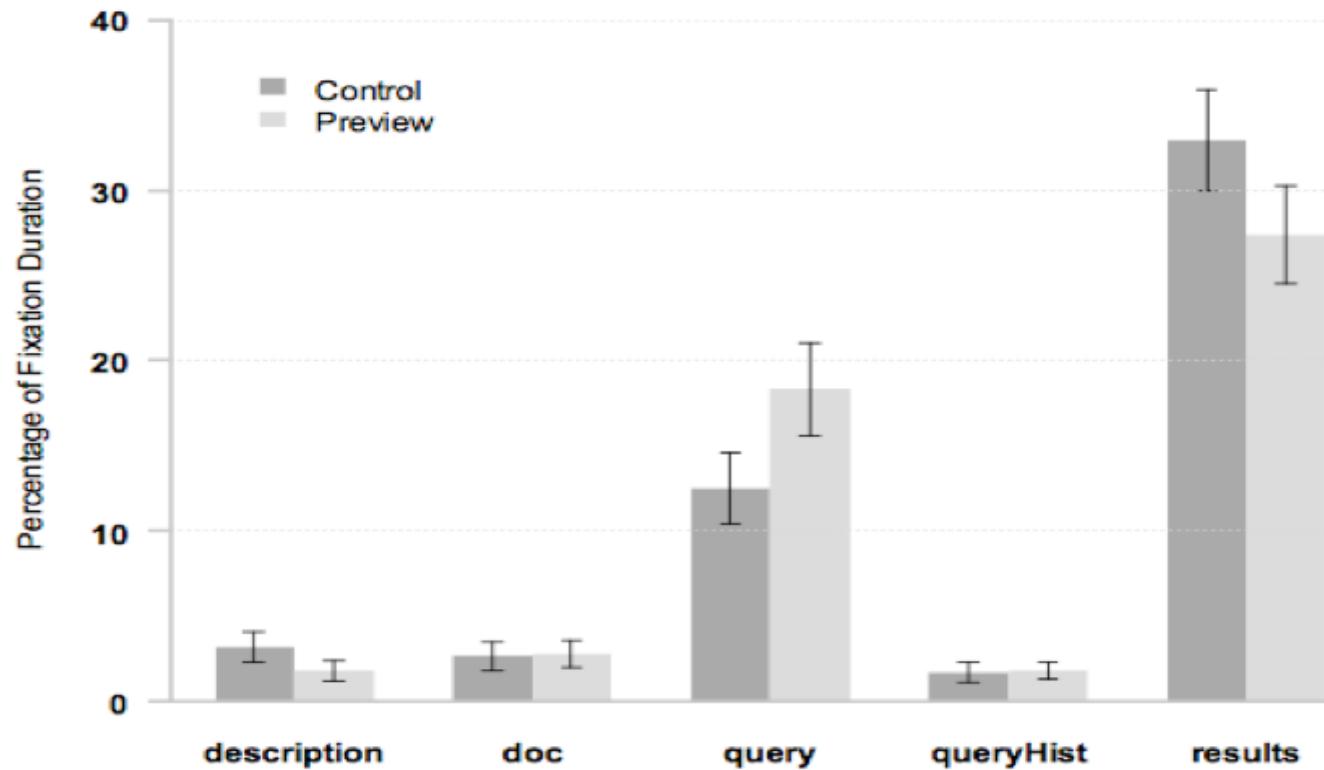


Baseline; Control Condition



Qvarfordt, P., Golovchinsky, G., Dunnigan, T. & Agapie, E. (2013). *Looking ahead: Query preview in exploratory search*. Proc. of ACM SIGIR Conference, 243-252.

# Example: Interface Evaluation



**Figure 6. Percentage of attention on UI elements *during* query formulation (total fixation duration on UI element).**

# Example: Individual Differences Research

- The basic premise is that people differ along many characteristics; researchers in this area seek to explain variations in behaviors by variations in characteristics.
- Some of the earliest work in interactive IR research focused on investigating how search success was related to individual differences.
- Example differences that have been studied include: age, biological sex, undergraduate major, learning styles, and **cognitive abilities**.
  - Cognitive Abilities
    - **Perceptual speed**
    - Associative memory
    - Visualization ability

# Example: Cognitive Ability

Do people with different levels of perceptual speed exhibit different search behaviors when solving the same information search tasks?

	Low Perceptual Speed	High Perceptual Speed
Session length (min)	6.85 (3.71)	6.43 (3.73)
Query length	4.62 (2.24)	5.85 (1.98)
SERP clicks	1.91 (1.78)	3.52 (2.68)
URLs viewed	2.38 (1.96)	4.63 (3.81)
URLs per query	1.26 (0.88)	1.84 (1.58)
All differences significant at $p<0.01$		

Brennan, K., Kelly, D., & Arguello, J. (2014). The effect of cognitive abilities on information search for tasks of varying levels of complexity. *Proceedings of the Information Interaction in Context Conference (IICX)*, Regensburg, Germany, 165-174.

# What is an experiment?

- In an **experiment**, we **deliberately manipulate** the environment to **isolate** the effect of one variable on another.
- This can be contrasted with **correlational research**, where we take a **snapshot** of several variables at a time and investigate how they co-relate (often in natural settings).
  - Example: retrospective log analysis
  - Correlation is a necessary condition for establishing causality, but is not sufficient.

# Components of a User-Centered Study

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2).

# Components of a User-Centered Study

- Participants
- Experimental “Conditions”
  - Systems/Algorithms
  - Interfaces
  - Instructions
  - ...
- Search Tasks (sometimes called topics; can be used as independent variables)
- Collection/Corpus of Information Objects
- Measures
- Data Collection Techniques
- Data Analysis Procedures

# Participants

- Participants should be *target* users.
  - E.g., Asking undergraduate students to evaluate an enterprise search system is unlikely to be meaningful.
- Studying computers scientists isn't always bad.
- Important to describe recruitment methods and characteristics of participants.
- Sample sizing is not magic or ad-hoc, although it often appears this way.

# Digression: Sample Sizing

- There are statistical methods to help you understand *risks* associated with sample sizes
  - The goal of *statistical power analysis* is to identify a sufficient number of participants to keep the risk of *Type I errors* and risk of *Type II errors* at acceptably low levels given a particular *effect size* without making the study unnecessarily expensive or difficult.
- Bigger ≠ Better
  - i.e., don't confuse sample size with sample representativeness

# Type I and Type II Errors

- A Type I error ( $\alpha$ ) occurs when the researcher rejects a null hypothesis that is actually ‘true.’ (False positive)
- A Type II error ( $\beta$ ) occurs when the researcher fails to reject the null hypothesis that is actually ‘false.’ (False negative)

# Type I and Type II Errors

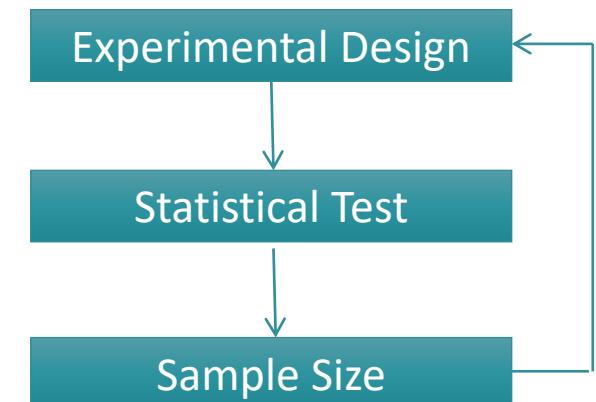
- The amount of risk a researcher is willing to tolerate with respect to making Type I or Type II errors can be quantified.
- In fact, researchers already state their risk for Type I errors when they set  $\alpha= 0.05$  as the threshold for statistical significance ( $\alpha$  is the *p-value* before the statistic is computed).
  - When  $\alpha= 0.05$ , you are risking a 5% chance of a Type I error.
- Side note: The more tests you do, the greater the likelihood of Type I errors.
- Tip: probability theory says to try something 20 times and then you'll find at least one significant result!

# Type I and Type II Errors

- Most people are less familiar with when and how to declare tolerance for **Type II errors ( $\beta$ )**.
- Works the same way in that risks are expressed as decimals (i.e., 0.05, or 5% risk).
- $\beta$  is not used as an input to most statistical tests, but it is used during power analysis.
- **Statistical power** is expressed as  $1 - \beta$

# What is Statistical Power?

- Statistical significance testing is computed after a study is completed.
- Statistical power is computed before a study is conducted. It involves two major steps:
  - Hypothesizing an **effect size**
  - Estimating risks associated with Type I and Type II errors
- It is also a function of the type of statistical test one wants to perform (which is a function of the experimental design).



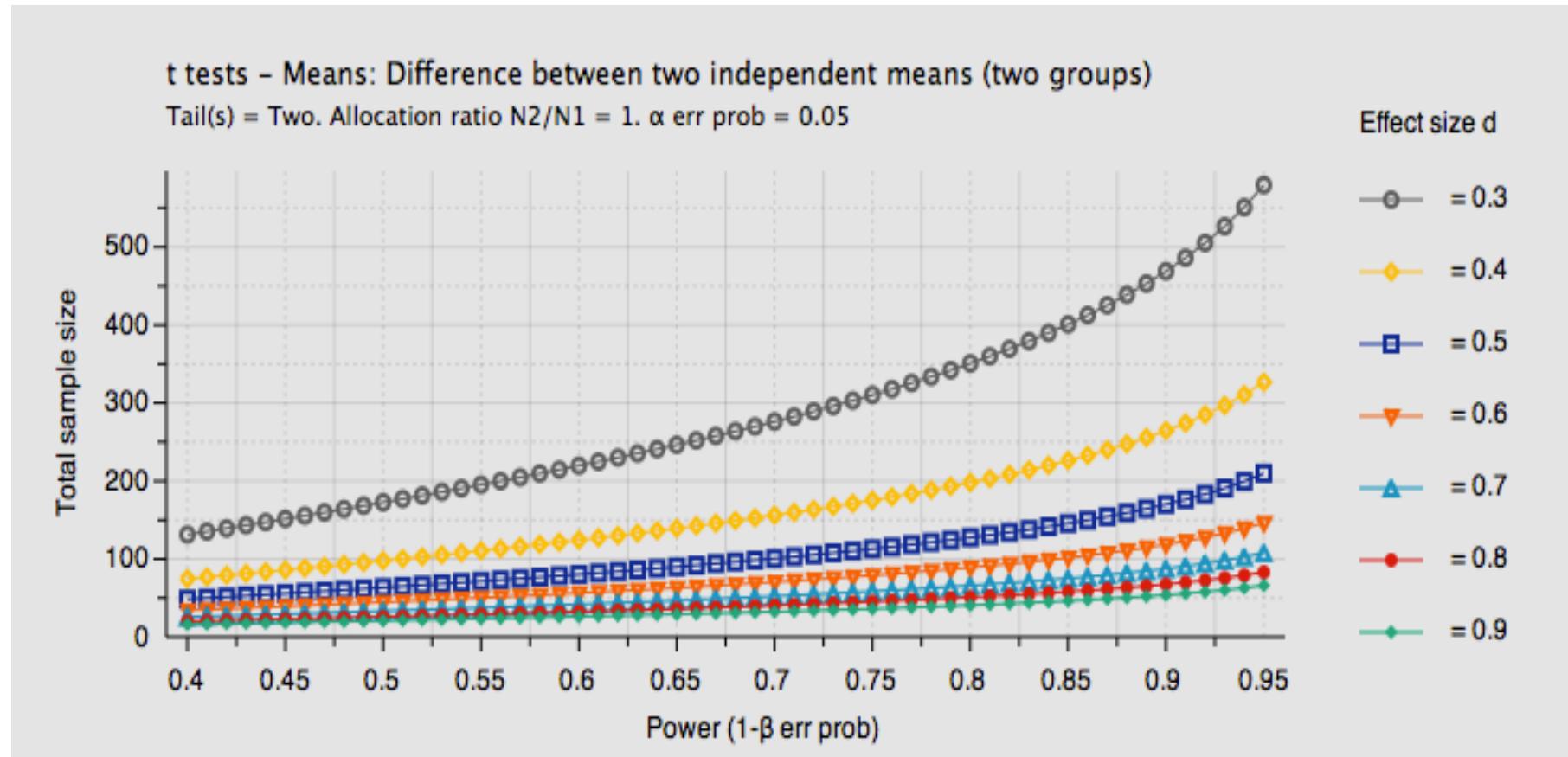
# What is Effect Size?

- Effect Size is the strength of a statistically significant finding. The *actual value* is computed after one runs a statistical test, but its *estimate* is used in power analysis.
- Effect size is a **standardized measure of the magnitude** of the association between the treatment and outcome variables.
  - Does task complexity lead to a (10%? 20%? 70%) increase in number of queries entered?
  - Does Interface A lead to a (10%? 20% 70%) increase in the performance over Interface B?
- Most uncertain aspect of sample size planning
  - Requires a hypothesis of the expected effect size

# Repeat

- Statistical power analysis allows a researcher to estimate sample size given:
  - Specific type of statistical test (e.g., independent samples t-test) (which is a function of the research design)
  - Anticipated effect size
  - Alpha value (risk of Type I errors)
  - Desired power (risk of Type II errors)
    - Reminder: Power =  $1 - \beta$
- Demo of G\*Power

# Power Analysis of Independent Sample T-Test



# Summary

- Statistical power analysis and G\*Power are analytical tools to help you better understand the risks (mostly Type II) associated with your experiment given a particular sample size.
- You can pick any sample size you want (there are no correct answers!), but you should understand the risks associated with Type II errors.
- An underpowered study mostly affects you ... and potentially what we know as a community.

# Experimental Designs

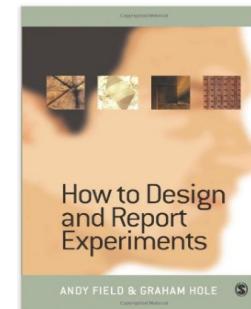
Note to ESSIR Students: We weren't able to cover the next set of slides, but I leave them in here for your reference if they are useful.

# Types of Experimental Designs

- Between subjects factor (or variable)
  - Participants are only exposed to one level of the factor.
  - Example: You are testing two interfaces and each participant only uses one interface.
  - Also called between groups design.
- Within subjects factor (or variable)
  - Participants are exposed to more than one level of the factor.
  - Example: You are testing two interfaces and participants use both interfaces.
  - Also called *repeated measure(s)* design.

# Between Subjects Designs

- Advantages
  - Simplicity
  - Less chance of practice and fatigue effects.
  - Useful when it is impossible for an individual to participate in all experimental conditions.
- Disadvantages
  - More expensive in terms of time, effort and number of participants.
  - Insensitive to experimental manipulations.



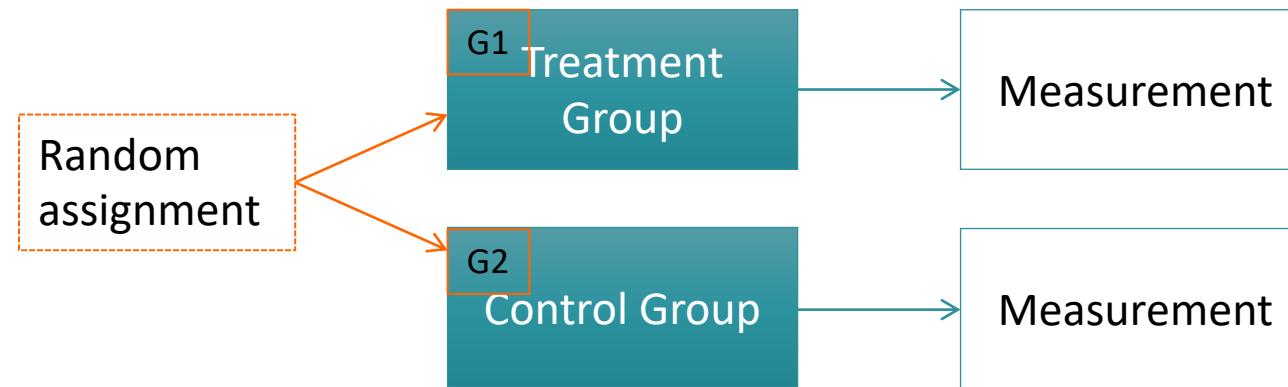
# (Not) an Experimental Design

Classic Usability Test



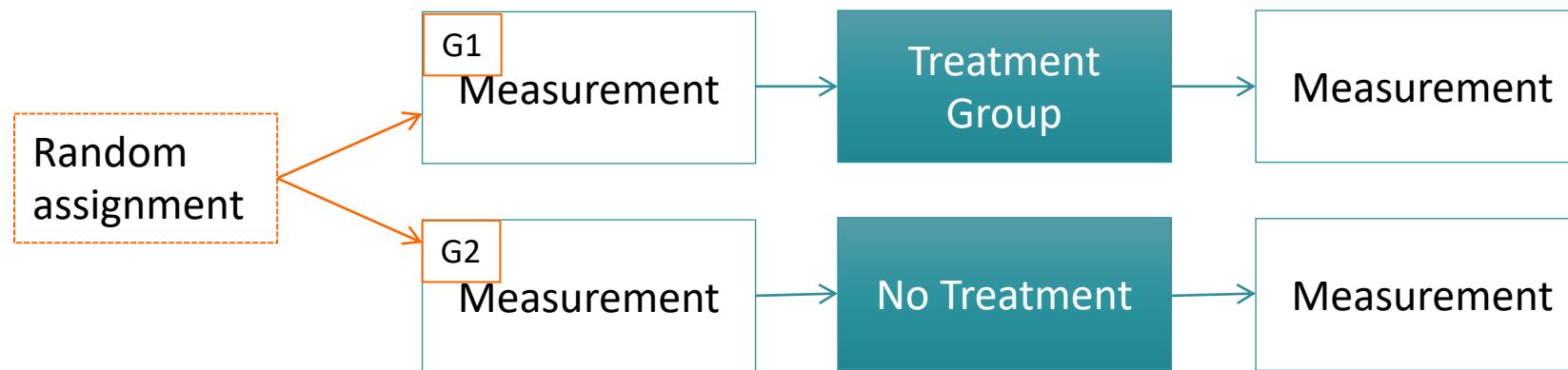
# Between Subject Designs

Post-test only/control group design



# Between Subject Designs

Pre-test/Post-test control group design



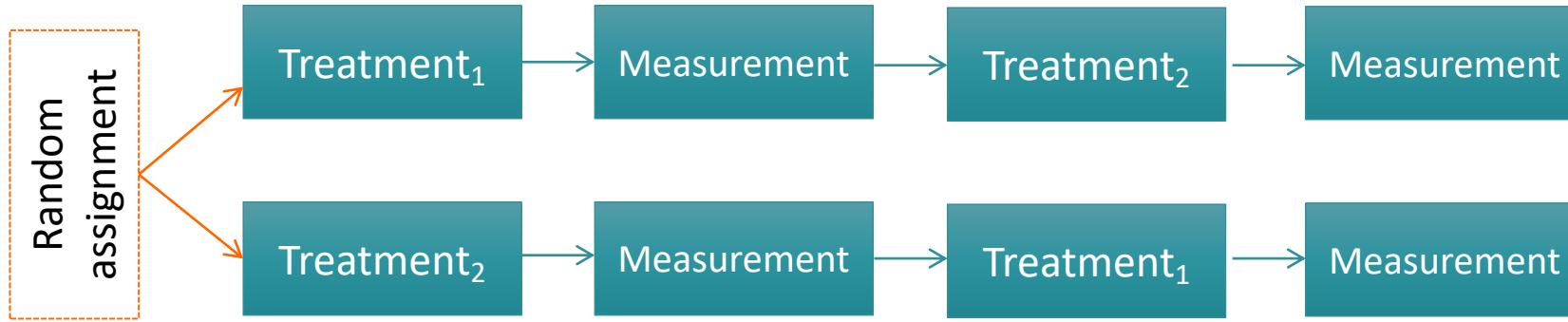
# Within Subject Design

- All things being equal, a repeated-measure (within subject design) is more sensitive than a between subjects design because there are fewer sources of random variation in our outcome measures.
  - Participants act as their own controls.
- More economical with respect to time and effort, but ...

# Within Subject Designs

- Disadvantages
  - ‘Carry-over’ effects from one condition to another.
    - These designs take more time for the participant and so ‘systematic’ variations like fatigue, boredom, practice, learning can become more of an issue.
  - The need for conditions to be reversible.
    - Example: Time pressure

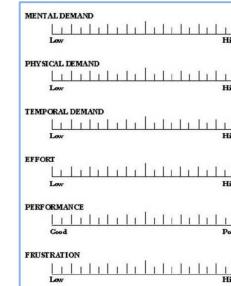
# Within Subject Designs



Treatment 1



Treatment 2



Measurement

[Treatments are *Counterbalanced*]

# Mixed Between/Within Design: Example

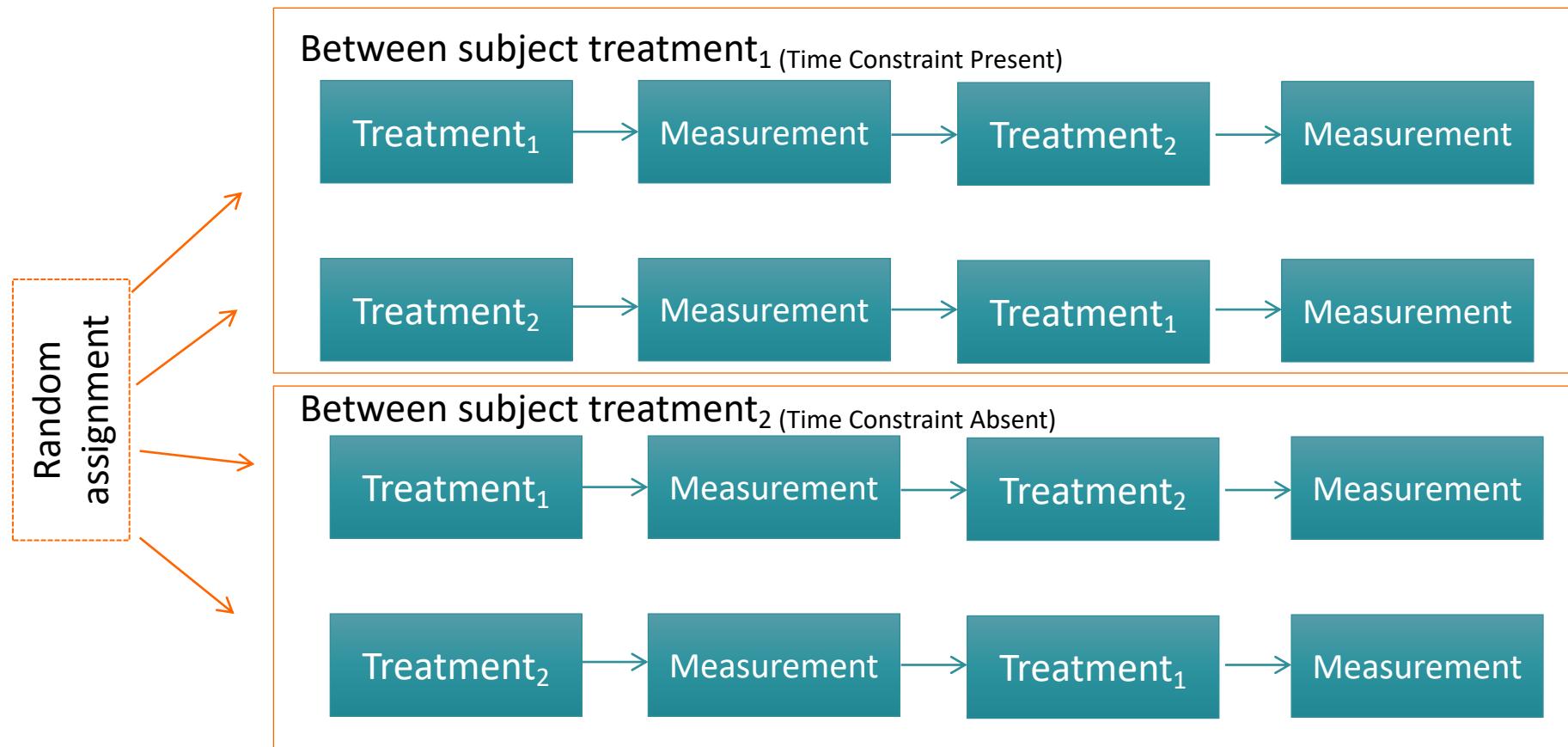
- **Time constraint** (between-subjects):
  - none vs. 5 min to complete each task
- **Delay** (within-subjects):
  - none vs. 5 sec. added to each query and SERP click
  - Every other task (delay order counterbalanced)
- 4 search tasks

Crescenzi, A., Kelly, D. and Azzopardi, L. 2016. Impacts of Time Constraints and System Delays on User Experience. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, CHIIR 2016*, ACM. New York. 141-150.

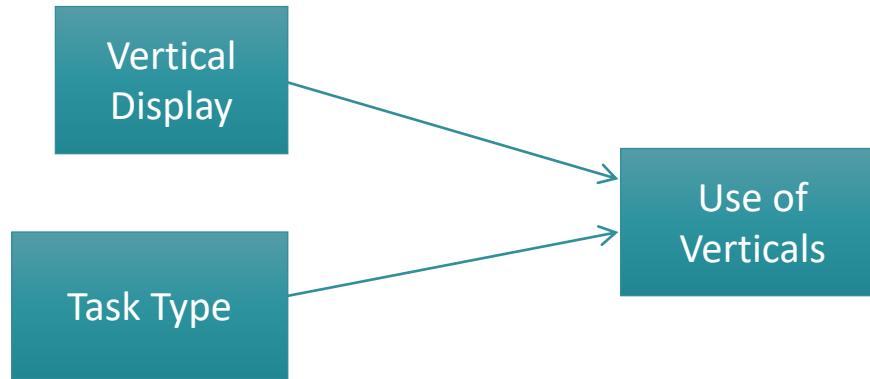
# Mixed Between/Within Design

Between: Time constraint (present or absent)

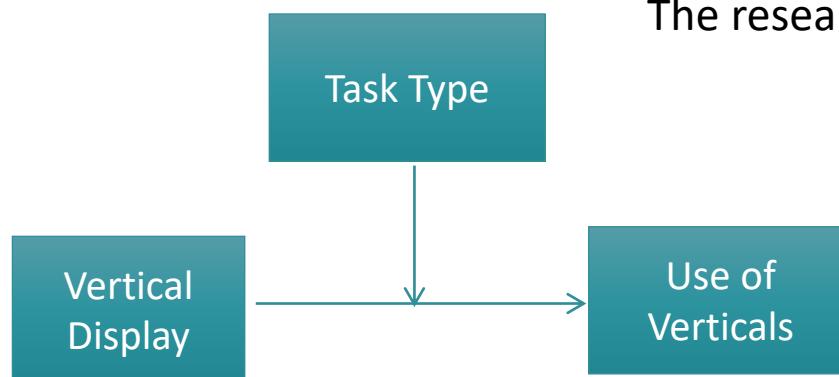
Within: Time delay (present or absent)



# Other Ways to View Research Designs



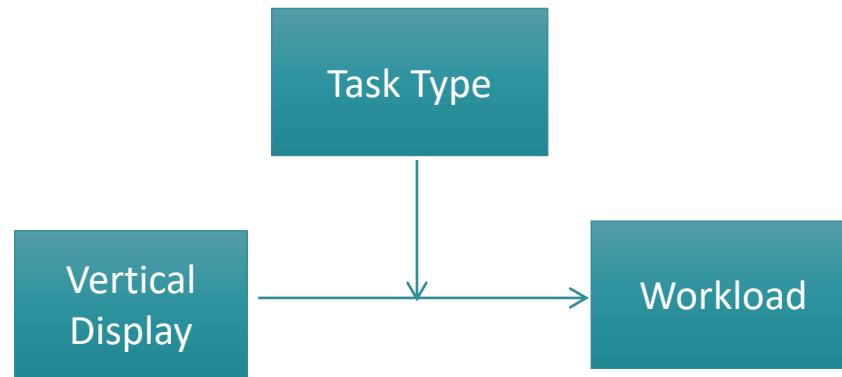
- Two independent variables
- One dependent variable



The research MODEL matters.

- One independent variable
- One **moderating** variable
- One dependent variable

# Research Designs



- One independent variable
- One **moderating** variable
- One dependent variable
- **Factorial Design (new concept)**
- 2X2 Factorial Design
- Main Effects
- Interaction Effects

		Vertical Display		
Task Type		Blended	Non-Blended	Total
	Fact Finding	18.1	6.3	11.7
	Exploratory	5.2	16.8	11.6
Total		12.25	11.05	11.65

Workload:  
1 ... 20

# What is a Task?

“activity t  
and a  
(V)

“search tas  
carried o  
(Wildemu

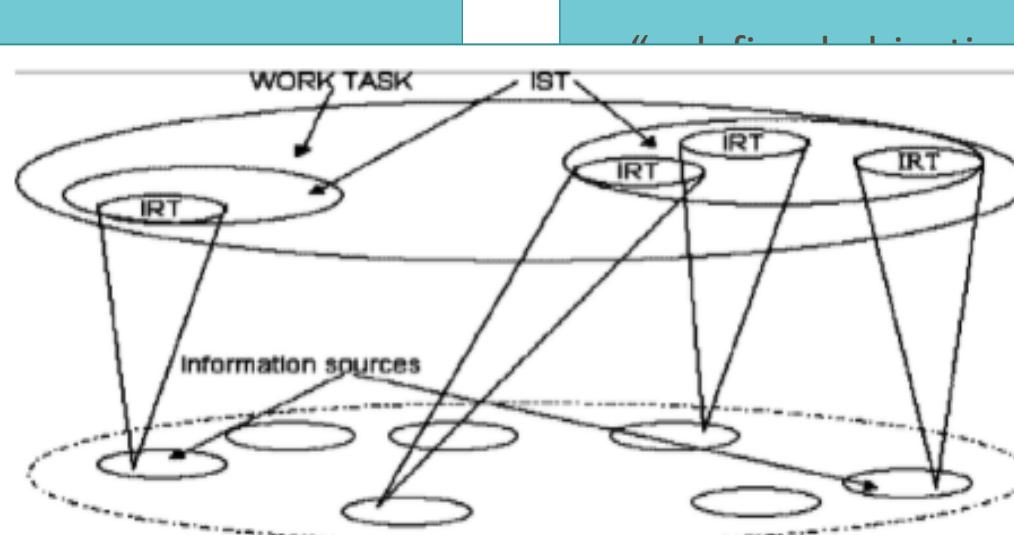


FIG. 1. Information seeking and searching (and retrieval) embedded in a work task (Byström & Hansen, 2002).

with an  
known  
ve known  
onal  
, p. 45)

omplish  
n with

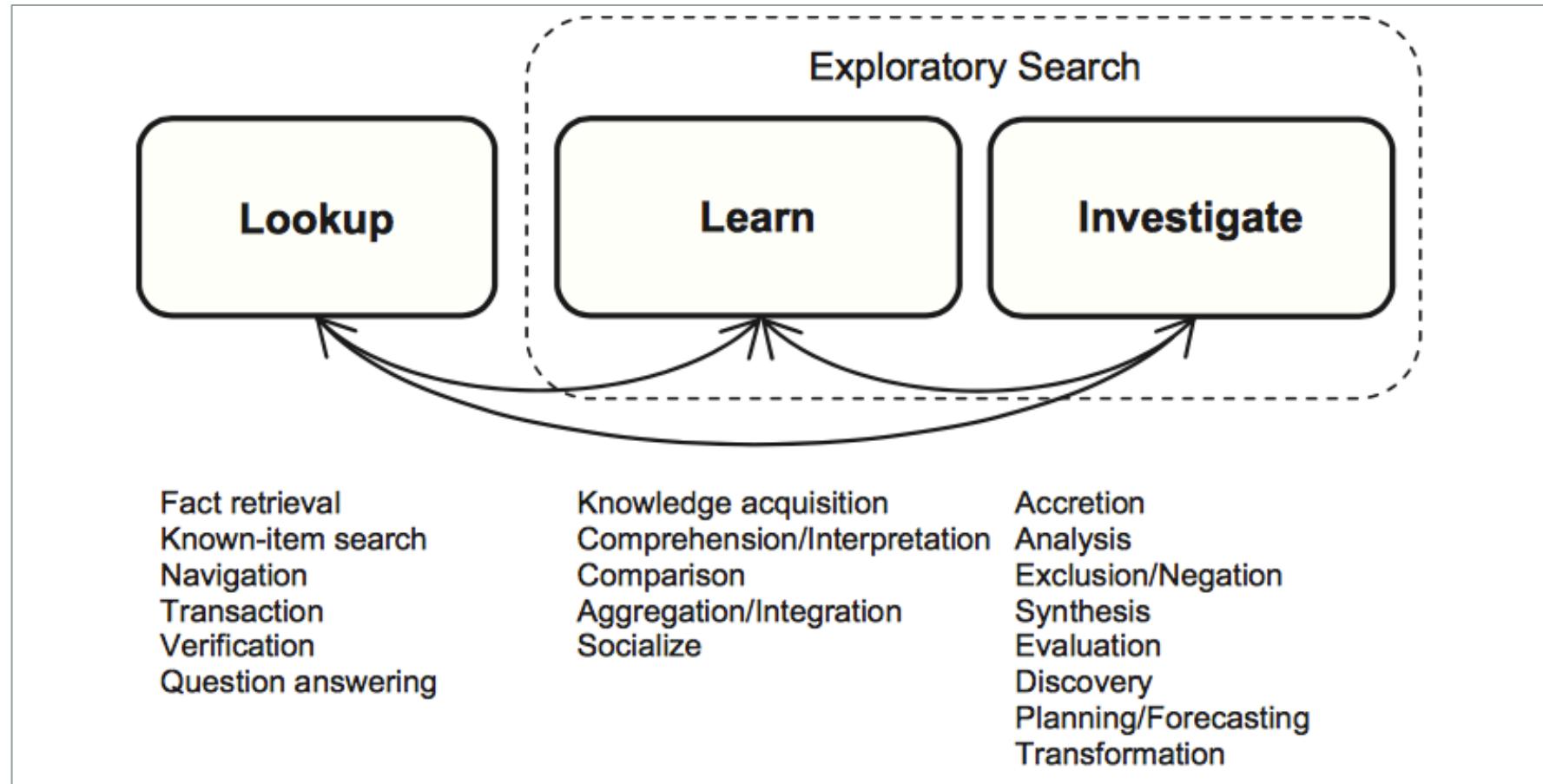
23)

# What is a Task?

- soft surroundings
- belly dancing music
- christian dior large bag
- best western airport sea tac
- [www.bajawedding.com](http://www.bajawedding.com)
- marie selby botanical gardens
- big chill down coats

Planning to attend a wedding?

# Task Types



White, R. W. & Roth, R.A. (2009). *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool. modified from Marchionini (1995)

# Example Search Task

You recently heard a story on National Public Radio about the use of biomass as fuel. Biomass refers to material created from living organisms. What are different types of biomasses that are used as fuels and how are they created? How do biomass fuels compare with fossil fuels when it comes to environmental impact? Which do you think is better? Why?

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIR*.

# Task Have Lots of Attributes ...

- Type (Fact-finding, decision-making, etc.)
- Complexity (objective)
- Difficulty (subjective)
- Saliency
- Urgency
- Importance
- ...

Li, Y. & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *IP&M, 44*, 1822-1837.

# ... Which Makes Task Design Difficult

- Task type, task complexity, and task difficulty are all variables which can quickly pile-up into a complicated, experimental mess.
- Complicated experiments are often difficult to understand and analyze.
- Some task attributes are experiential, which make them hard to predict, model and control.

# Tasks as Experimental Infrastructure

- The IR community has a strong culture of creating experimental infrastructures.
- Tasks are experimental infrastructure for IIR experiments.
- The development of search tasks can be difficult and time consuming, and often requires specialized knowledge and expertise.
- Clear guidelines for creating tasks do not exist.
- No reference models and few sets of shared search tasks.

# Why Task Design Matters

- Poor task design can:
  - Confound results
  - Generate undesirable search behaviors
  - Result in wasted time and money

# Genuine vs. Assigned Tasks

- Previous work on genuine and assigned tasks has shown that participants generally indicate greater interest in genuine tasks than assigned tasks.
  - Borlund, Dreier and Byström's (2012) participants rated their genuine tasks more interesting, spent more time searching, and generally found genuine tasks more difficult than simulated work tasks.
  - Poddar and Ruthven (2010) found participants had greater positive emotions, made more use of various search strategies, and had more confidence in their ability to succeed on their own tasks versus assigned tasks.
- So, why would we want to assign tasks?

# Creating Interesting Tasks

1. How can we assign tasks to participants that will interest them?
2. To what extent with search behaviors and experiences differ for participants when searching for tasks that interest them versus tasks that do not interest them?

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIR*.

# Example Task and Procedure

You recently heard a story on National Public Radio about the use of biomass as fuel. Biomass refers to material created from living organisms. What are different types of biomasses that are used as fuels and how are they created? How do biomass fuels compare with fossil fuels when it comes to environmental impact? Which do you think is better? Why?

- Eight “evaluate” tasks were created using four domains.
- Day before experimental session, participants were asked to rank tasks from 1-8, with 1=most interesting and 8=least interesting.
- Task order randomized for each participant.
- Participants assigned two most interesting and two least interesting tasks to complete during the study.

# Participants' Task Rankings

Task	Interesting (1 or 2)	Uninteresting (7 or 8)
Online Communication	25	3
Energy Sources	11	6
Lupus	9	10
Endurance Sports	10	8
Vehicle Purchase	8	15
Tattoo Removal	7	13
Video Game Violence	4	12
Biomass Fuel	6	13
<b>Total</b>	<b>80</b>	<b>80</b>

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIIR*.

# Manipulation Check

**Table 2. Means, standard deviations, and t-test results of pre-search questionnaire, df=79, \*\*\* $p<.001$**

	Interesting Tasks	Uninteresting Tasks	t-values
Interest	4.01 (0.96)	2.77 (1.15)	7.39***
Prior Knowledge	3.00 (1.13)	2.20 (1.80)	5.27***
Relevance	3.61 (1.23)	2.24 (1.23)	7.06***
Search Frequency	2.44 (1.20)	1.64 (0.94)	4.69***
Difficulty	3.35 (1.15)	3.81 (0.93)	2.72***

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIIR*.

# Were people's search experiences different?

**Table 3. Means, standard deviations and *t*-test results for post-search questionnaire (all non-sig), df = 79**

	Interesting Tasks	Uninteresting Tasks	<i>t</i> -values
Difficulty	3.61 (0.83)	3.38 (0.89)	1.35
Skill	3.85 (0.70)	3.62 (0.85)	1.57
System Ability	3.40 (0.86)	3.35 (1.04)	0.27
Success	3.77 (0.69)	3.80 (0.77)	-0.17
Frustration	2.45 (0.56)	2.52 (0.84)	-0.65

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIR*.

# Were people's search experiences different?

**Table 4. Means, standard deviations and t-test results for User Engagement Scale, df = 79, \*\* $p < .01$ , \*\*\* $p < .001$**

	Interesting Tasks	Uninteresting Tasks	t-values
Perceived Usability	3.63 (0.60)	3.56 (0.79)	0.50
Focused Attention	2.87 (0.79)	2.64 (0.66)	2.70**
Felt Involvement	3.61 (0.50)	3.27 (0.72)	3.00***
Endurability	3.41 (0.58)	3.20 (0.76)	1.60
Novelty	3.72 (0.67)	3.09 (0.90)	4.43***
Total Engagement	3.45 (0.42)	3.15 (0.64)	3.04***

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIIR*.

# Did people behave differently?

**Table 5. Means, standard deviations, and t-test results of search behaviors, df = 79, \*p<.05**

	Interesting Tasks	Uninteresting Tasks	t-values
Total Time (in minutes)	6.86 (3.84)	5.88 (3.06)	2.06*
Time on SERP	2.04 (1.80)	1.74 (1.30)	1.40
Time on Docs	2.72 (2.25)	2.33 (1.79)	1.45
Queries	5.65 (4.45)	5.06 (4.13)	1.01
Query Length (in words)	2.56 (1.77)	2.55 (1.78)	0.94
Clicks Per Query	7.77 (4.77)	6.79 (3.97)	1.66
SERPs viewed	13.80 (9.16)	12.29 (7.88)	1.42
Bookmarks	3.60 (2.06)	3.42 (1.83)	0.63

Edwards, A. & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? *Proceedings of CHIIR*.

# Simulated Work Task

- Borlund proposed the use of **simulated work tasks**, which provide a short “cover story that describes an IR requiring situation.”
- These tasks are tailored to target participants.
- Allows participants to direct their own searches, while providing a common scenario to anchor experimental comparisons and relevance judgments.
- Better motivates people to search by providing them with tasks that interest them.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8.

# Simulated Work Task

## **Simulated situation:**

**Simulated work task situation:** After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

**Indicative request:** Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

**Figure 1. Example of a simulated situation/simulated work task situation ([Borlund, 2000a](#); [2000b](#)).**

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8.

# Data Collection Techniques & Measurement

# Data Collection Techniques

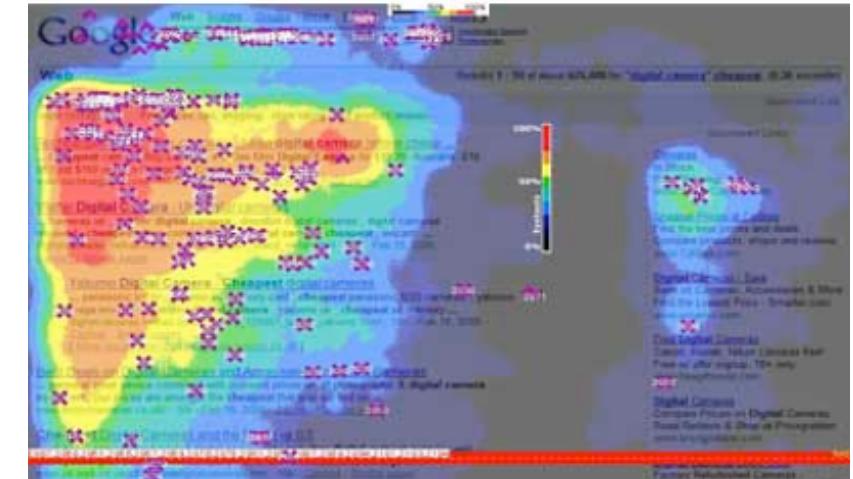
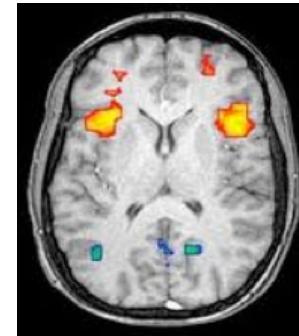
- Logging & Observation
- Self-report
  - Questionnaires (many types)
  - **Scales (standardized)**
  - Relevance measures
- Think-aloud & Stimulated Recall
- Interviews & Open-ended Questions
- Evaluation of End Products
- Learning



Morae

# Data Collection Techniques

- Eye-tracking
- Physiological Signals
- EEG
- Brain Scans (fMRI)



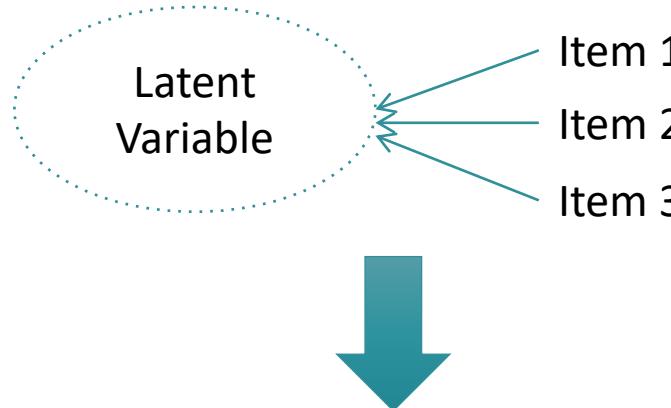
But ... What does this tell us?

# Measures

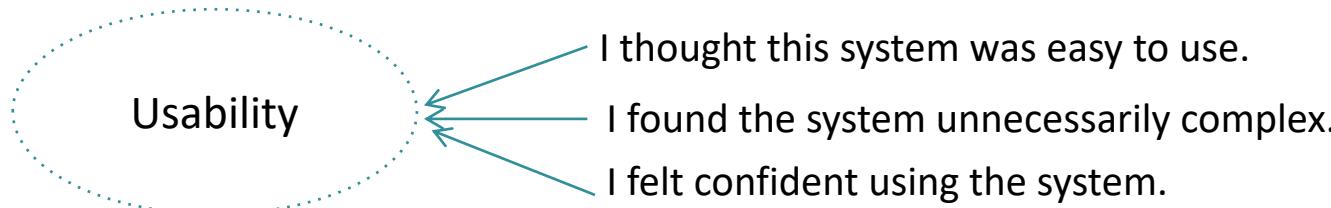
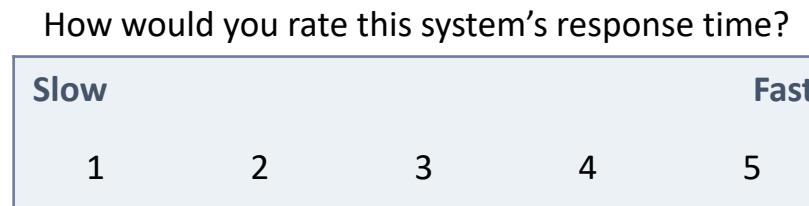
- Contextual
  - Individual Differences
    - Cognitive Ability
    - Need for Cognition
    - ...
  - Tasks
    - Type
    - Difficulty
    - ...
- Interaction
  - Queries issued
  - SERP clicks
  - Time spent
  - ...
- Performance
  - Number saved
  - Query diversity
  - ...
- User Experience
  - Usability
  - Preferences
  - Mental Effort, Workload
  - Flow and Engagement
  - Emotion
  - ...

# Digression: Self-Report Measures and Scales

Are you assuming some latent construct  
(psychometric theory)?



Or just want to ask a question?



# Example: Workload

Operationalization links a conceptual definition to a specific set of measurement techniques or procedures.

An operational definition could be:

- A scale (NASA TLX)
  - One or more survey questions
  - A method of observing events in a field setting

A conceptual definition of workload is:

- “Workload is a term that represents the cost of accomplishing mission requirements for the human operator.”

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, 904-908. Santa Monica: HFES.

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date
------	------	------

Mental Demand How mentally demanding was the task?



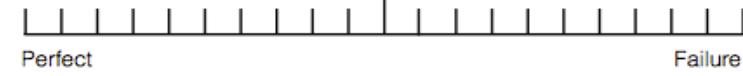
**Physical Demand** How physically demanding was the task?



**Temporal Demand** How hurried or rushed was the pace of the task?



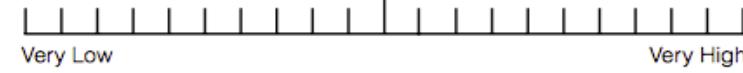
Performance How successful were you in accomplishing what you were asked to do?



**Effort** How hard did you have to work to accomplish your level of performance?



Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?



# Psychometrics

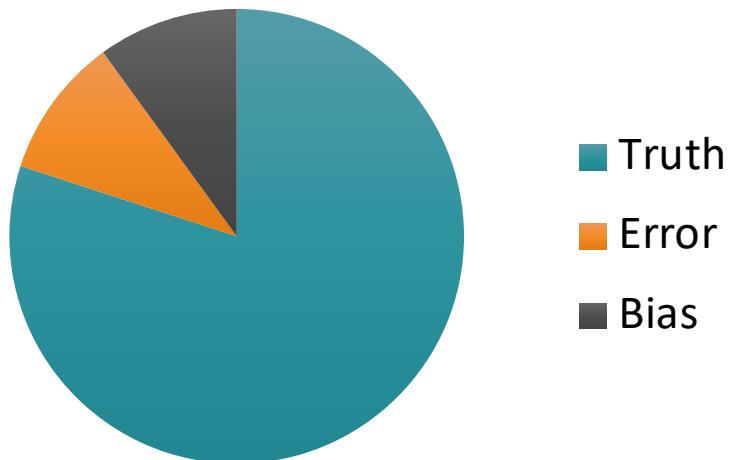
- Psychometrics is a field of study concerned with the theory and technique of psychological measurement.
- The field is concerned with the objective measurement of skills and knowledge, abilities, attitudes, personality traits and educational achievement.
- Two major tasks:
  - Construction of instruments
  - Development of procedures for measurement

# Psychometrics

- Measurement is, “the assignment of numerals to objects or events according to some rule.”
  - Stevens (1946)
- “Measurement in psychology and physics are in no sense different. Physicists can measure when they can find the operations by which they use to measure meet the necessary criteria; psychologists have but to do the same. They need not worry about the mysterious differences in meaning of measurement in the two sciences.”
  - Reese (1943)

# Psychometrics Model

$$O = T + E + B$$



# Example: Error

- Which option do you prefer?
  - Option A
  - Option B
  - Option C
  - Option D

It's not just how you ask questions; the format of the question can also introduce error.

# Example: Error

The system provided results quickly.

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

# Example: Bias

- How much time did you spend preparing for ESSIR?
  - 0 hours
  - 1 hour
  - 2 hours
  - 3 hours
  - 4 hours
  - 5 hours
  - More than 5 hours

# Note: All measurement is manmade.

## Stanford-Binet Fifth Edition (SB5) classification

IQ Range ("deviation IQ")	IQ Classification
---------------------------	-------------------

<b>145–160</b>	Very gifted or highly advanced
<b>130–144</b>	Gifted or very advanced
<b>120–129</b>	Superior
<b>110–119</b>	High average
<b>90–109</b>	Average
<b>80–89</b>	Low average
<b>70–79</b>	Borderline impaired or delayed
<b>55–69</b>	Mildly impaired or delayed
<b>40–54</b>	Moderately impaired or delayed

## Levine and Marks 1928 IQ classification IQ Range ("ratio IQ") IQ Classification

<b>175 or above</b>	Precocious
<b>150–174</b>	Very superior
<b>125–149</b>	Superior
<b>115–124</b>	Very bright
<b>105–114</b>	Bright
<b>95–104</b>	Average
<b>85–94</b>	Dull
<b>75–84</b>	Borderline
<b>50–74</b>	Morons
<b>25–49</b>	Imbeciles
<b>0–24</b>	Idiots

# Be Aware of “Objective” Instruments

[Home](#) > Current Issue > vol. 113 no. 28 > Anders Eklund, 7900–7905, doi: 10.1073/pnas.1602413113

 CrossMark [click for updates](#)

## Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund<sup>a,b,c,1</sup>, Thomas E. Nichols<sup>d,e</sup>, and Hans Knutsson<sup>a,c</sup>

Author Affiliations [✉](#)

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

[Abstract](#) [Full Text](#) [Authors & Info](#) [Figures](#) [SI](#) [Metrics](#) [Related Content](#) [PDF](#) [PDF + SI](#)

### Significance

Functional MRI (fMRI) is 25 years old, yet surprisingly its most common statistical methods have not been validated using real data. Here, we used resting-state fMRI data from 499 healthy controls to conduct 3 million task group analyses. Using this null data with different experimental designs, we estimate the incidence of significant results. In theory, we should find 5% false positives (for a significance threshold of 5%), but instead we found that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in false-positive rates of up to 70%. These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results.

# Measurement Variability

- No standardized measures for most concepts.
  - Everyone creates their own measures.
- Different studies measure the same concept in different ways.
  - For example, task complexity.
- The same measure (or observation) is used to represent different concepts, which are oftentimes completely opposite of one another.
  - For example,
    - # clicks = engagement
    - # clicks = frustration
    - Long display time = interest, engagement
    - Long display time = difficulty, failure
- Observations often not mapped to concepts.

# Method Variance & Measurement Error

- Method variance refers to variance that is attributable to the measurement method rather than to the construct of interest.
  - For example,
    - A protocol that primes participants to behave in a particular way.
    - Variation in how the study is delivered (e.g., researcher effects).
    - An instrument that introduces error (a poorly designed Likert scale).
    - A researcher who doesn't randomly assign participants to conditions.
- Crowdsourced and naturalistic studies necessarily contain a lot of method variance, which usually cannot be measured. Moreover, the sources of variance are not well-understood, or even known in most cases.

# Method Variance & Measurement Error

- Use of outdated test collections that are mismatched to target participants
- Use of unrealistic tasks that are mismatched to target participants

# Data Analysis

# Data Analysis

- Analytical methods are closely tied to experimental design.
- Techniques that model relationships, such as structural equation modeling, have not been used very much, but are starting to get more attention.
- Qualitative data analysis is often used:
  - Superficially
  - Rigorously

Kelly, D. & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967-2006. *Journal of the American Society for Information Science and Technology*, 64(4), 745-770.

# Statistical (Un)Reliability

- Several practices are of concern:
  - Undirected search for patterns (“fishing expedition”) (“p-hacking”)
  - Multiple-tests with no correction (increases risks of Type I error)
    - Limited use of data reduction techniques
    - Tests conducted using highly correlated measures
  - Inadequate number of study participants (under-powered studies) (increases risk of Type II error)
    - Lack of understanding of how to conduct formal, a priori power analysis
    - Don’t like results of power analysis 😞
    - Resource constraints (e.g., time, money)

# Statistical (Un)Reliability

- More concerning practices:
  - Use of statistical methods designed for "little" data on "big" data produce many significant (and often meaningless) differences
  - Lack of evidence about the strength of findings (i.e., not computing, or publishing effect size measures)

# Close: Contemporary Problems

# Contemporary Problems

- User Interfaces
  - Conversational
  - Proactive
  - Explanations
  - Decision-making

# On the reception and detection of pseudo-profound bullshit

Gordon Pennycook\*    James Allan Cheyne†    Nathaniel Barr‡    Derek J. Koehler†  
Jonathan A. Fugelsang†

## Abstract

Although bullshit is common in everyday life and has attracted attention from philosophers, its reception (critical or ingenuous) has not, to our knowledge, been subject to empirical investigation. Here we focus on pseudo-profound bullshit, which consists of seemingly impressive assertions that are presented as true and meaningful but are actually vacuous. We presented participants with bullshit statements consisting of buzzwords randomly organized into statements with syntactic structure but no discernible meaning (e.g., “Wholeness quiets infinite phenomena”). Across multiple studies, the propensity to judge bullshit statements as profound was associated with a variety of conceptually relevant variables (e.g., intuitive cognitive style, supernatural belief). Parallel associations were less evident among profundity judgments for more conventionally profound (e.g., “A wet person does not fear the rain”) or mundane (e.g., “Newborn babies require constant attention”) statements. These results support the idea that some people are more receptive to this type of bullshit and that detecting it is not merely a matter of indiscriminate skepticism but rather a discernment of deceptive vagueness in otherwise impressive sounding claims. Our results also suggest that a bias toward accepting statements as true may be an important component of pseudo-profound bullshit receptivity.

Keywords: bullshit, bullshit detection, dual-process theories, analytic thinking, supernatural beliefs, religiosity, conspiratorial ideation, complementary and alternative medicine.

# Contemporary Problems

- Many problems and challenges to tackle in user-centered evaluation including:
  - Development of research infrastructure including standardized measures and tasks
  - Analysis of issues related to statistical power, validity, reliability and reproducibility.
- Development and evaluation of theoretical explanations for information interactions with information systems.