Topics in ~~User~~ Cognitive Modelling in Interactive Information Retrieval

Alan Medlar (alan.j.medlar@helsinki.fi)
Dorota Glowacka (glowacka@cs.helsinki.fi)

Organisation and assessment

- This seminar is worth 5 credits
- Duration: teaching periods 1 and 2
- Aim: introduce interactive information retrieval and cognitive modelling
- Assessment:
  - 8-10 page report (+ references) on using cognitive modelling in IIR
  - a presentation on the topic of the report
- Topics in IIR and cognitive modelling can be chosen freely, but we will provide suggestions

Schedule

- **Changes will appear on the course webpage**
- 04.09.19 Lecture 1: Introduction to IR and IIR
- 11.09.19 Lecture 2: Cognitive modelling
- 25.09.19 Deadline for topic selection (title + 3 papers min.)
- 09.10.19 Presentation of chosen topic (5 mins, 5 slides)
- 30.10.19 Feedback session
- 20.11.19 Final presentations (20 mins, 20 slides) - if necessary
- 27.11.19 Final presentations (20 mins, 20 slides)
- 11.12.19 Deadline for final paper submission

Essay structure

- Essay will have 3 sections:
  - **An IIR component (or search task)** (e.g. ranking, relevance feedback, implicit relevance feedback) - what does it do? how is it implemented? how is its effectiveness validated?
  - **A cognitive process** (e.g. categorisation, decision making, implicit learning) - what does it study? describe the model, what type of experiment is used to gather data?
  - **Cognitive modelling in IIR** (e.g. modelling relevance feedback as a categorisation process) - sketch an experimental design, what old results can be replicated? what new results will we get?

Information Retrieval (IR)

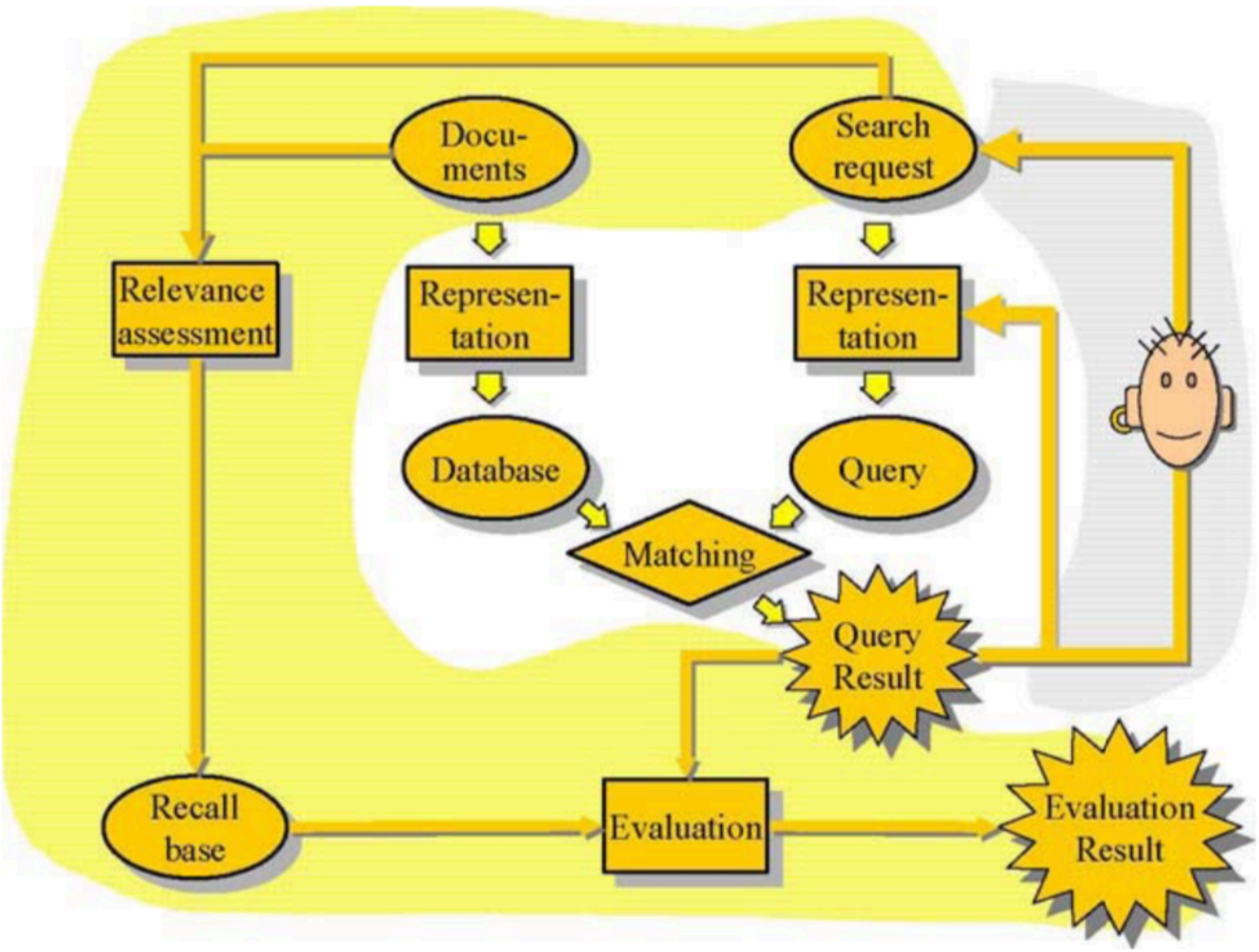
- Definition (Introduction to Information Retrieval, Manning, Raghavan and Schütze, 2010):

  "*Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*"

IR example


Google!
Help
google
10 results
Google Search
I'm feeling lucky
Showing results 1-10 of approximately 234,000 for google. Search took 0.06 seconds.
Take me back to the present
Google (Google RN)
...the web using Google Try our special searches: Uncle Sam Search millions...
...web's Linux resources ©1998 Google Inc....
www.google.com/ - Cached: 2k - GoogleScout
Google Search: <Unclesam>
...terms. Search the entire web from the Google home page! Copyright...
www.google.com/unclesam - Cached: 2k - GoogleScout
Google Search: <Linux>
...terms. Search the entire web from the Google home page! Copyright...
www.google.com/linux - Cached: 2k - GoogleScout
www.google.com/search
GoogleScout
Why Use Google?
...Why Use Google? Because Google delivers the most relevant search...
...search results--first and fast! Google uses sophisticated next-generation...
www.google.com/why_use.html - Cached: 8k -GoogleScout
www.google.com/netscape
GoogleScout
Google Help
...Basic Search To enter a query into Google, just type in a few descriptive...
...descriptive keywords and click on the Google Search button for your list...
www.google.com/help.html - Cached: 13k -GoogleScout
google.stanford.edu/
GoogleScout
Constructing and refining searches in Google: Detailed Searching Instructions
...refining searches in Google: Detailed Searching Instructions Teaching...
...University of California, Berkeley Google is a fairly new Web searching...
www.lib.berkeley.edu/TeachingLib/Guides/Internet/Google.html - Cached: 12k - GoogleScout
Gooooooooooogle
Result Page: 1 2 3 4 5 6 7 8 9 10 Next page
google
Google Search
Search within results?
Try your query on: AltaVista Excite HotBot Infoseek Lycos Deja Yahoo! Amazon Open Directory eGroups
Copyright ©1998 Google Inc. - About - Help

Laboratory model of IR

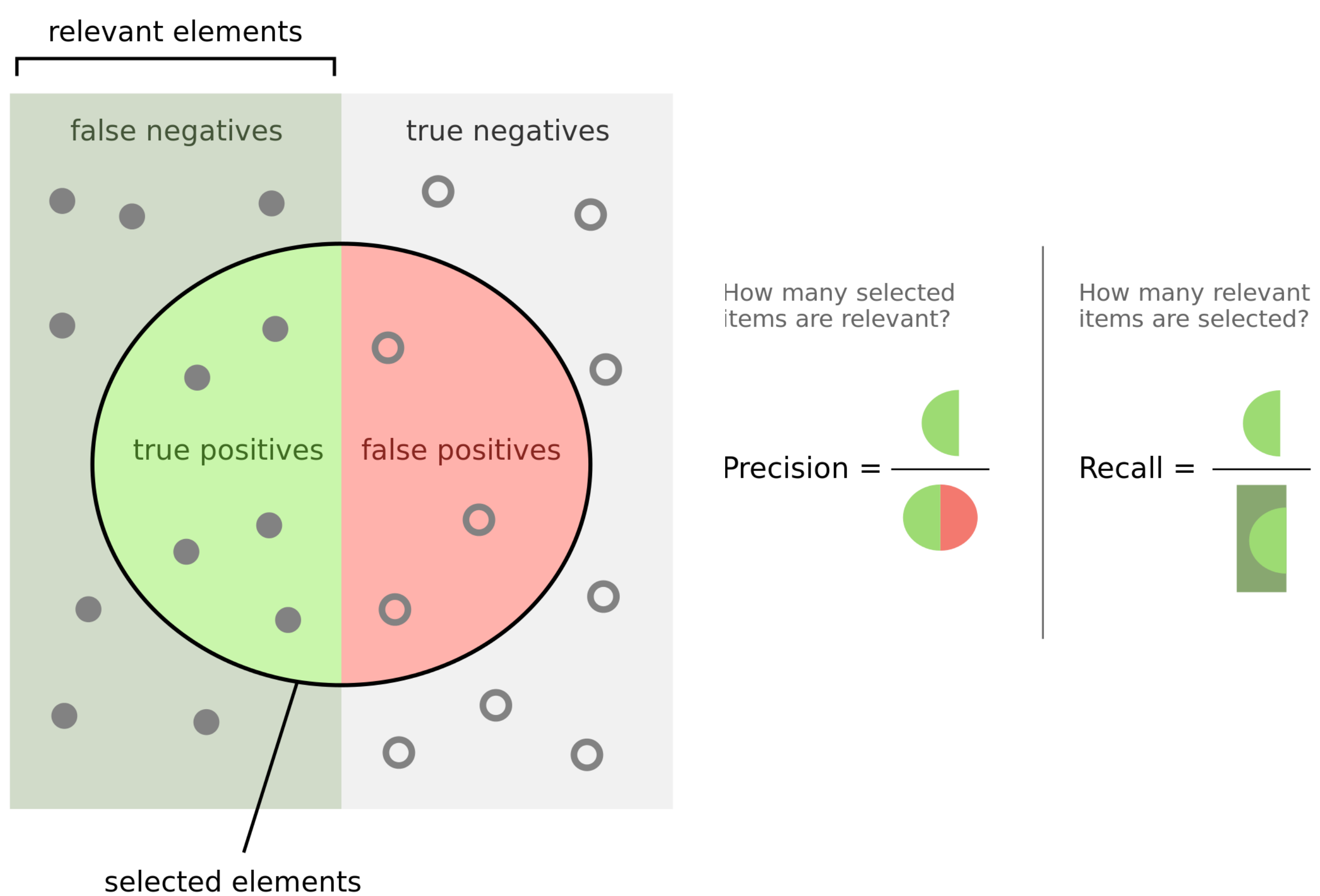


IR Evaluation (1)

- Based on the Cranfield experiments (1966)
- Assumes that the **relevance** of retrieved documents is a good proxy for whether the IR system satisfies users' **information needs**
- Requirements:
    - Document corpus
    - Information needs (queries)
    - Relevance judgments (binary assessment of relevant/not relevant for query-document pairs)

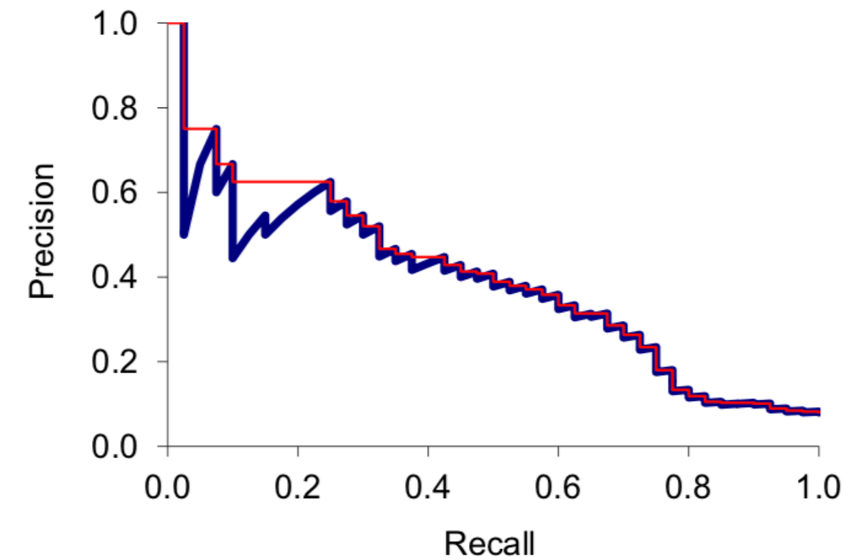
IR Evaluation (2)

- Precision and recall used to evaluate **unranked** search results
- **Tradeoff between precision and recall** - as the number of search results increases, precision decreases, but recall increases (on average)



IR Evaluation (3)

- Precision and recall does not reflect the efficacy of IR systems when search results are **ranked**
- Evaluation metrics for ranked search results:
  - Precision-recall curve
  - Precision@K (P@K)
  - Mean average precision (MAP)
  - Discounted cumulative gain (DCG)

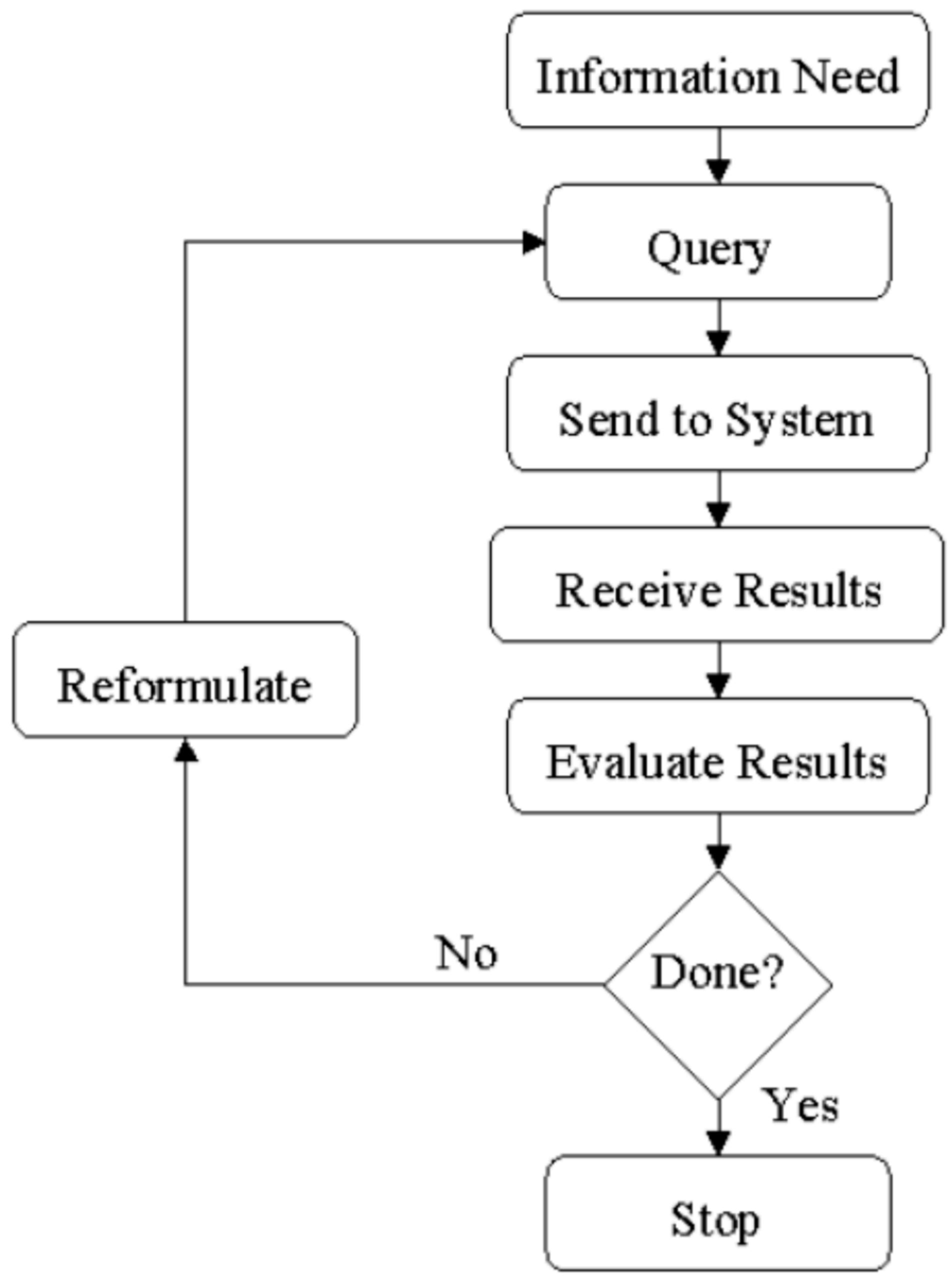


$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Models of search

- Classic IR
  - Content-related search in unstructured documents
  - System-oriented view
  - **Static** information needs

Classical search process model



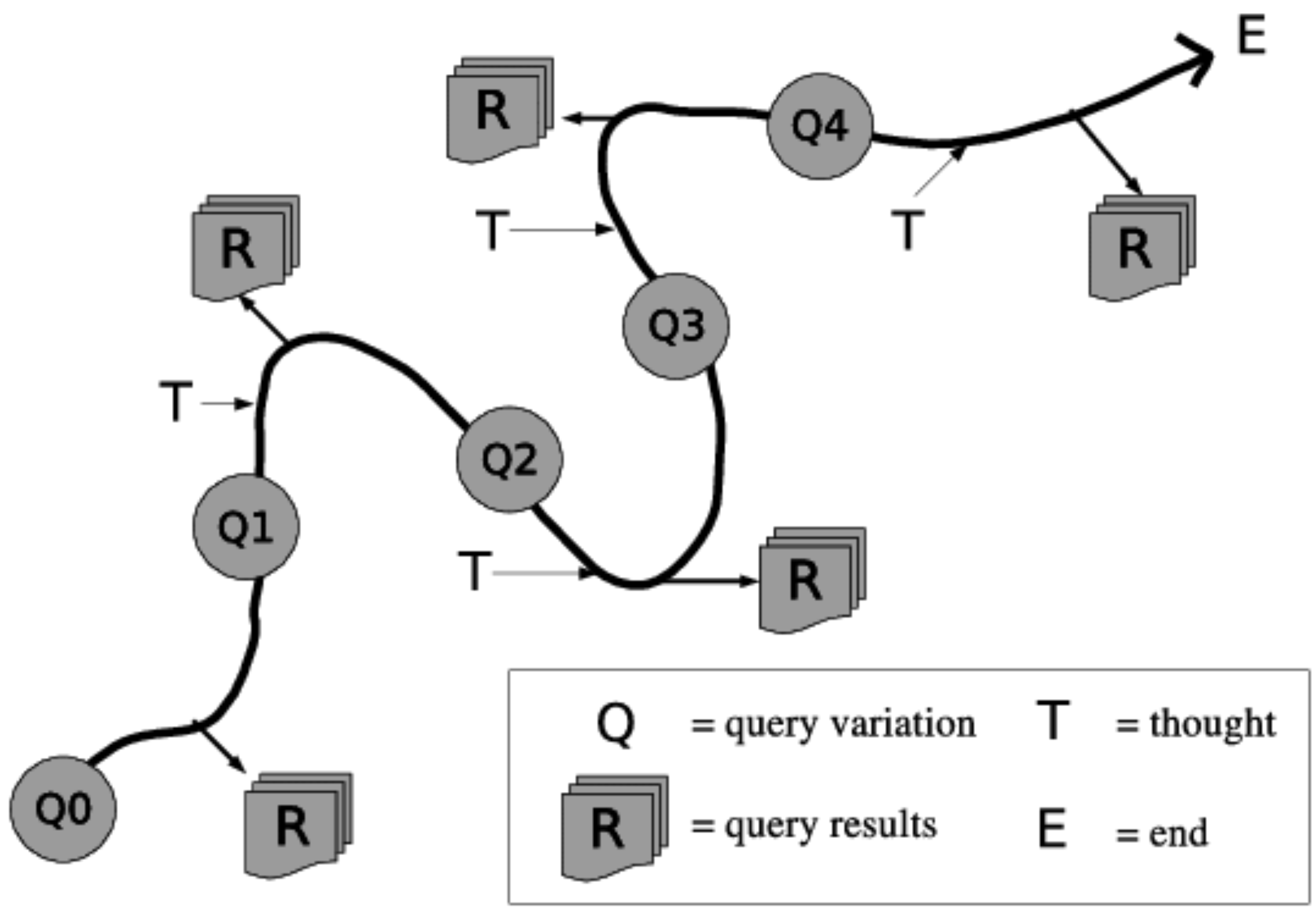
Empirical studies

- Search consists of a **sequence** of connected, but different searches
- Search results trigger new searches, the task context remains constant
- Goal of search is to **accumulate information** and learn about the search topic

Models of search

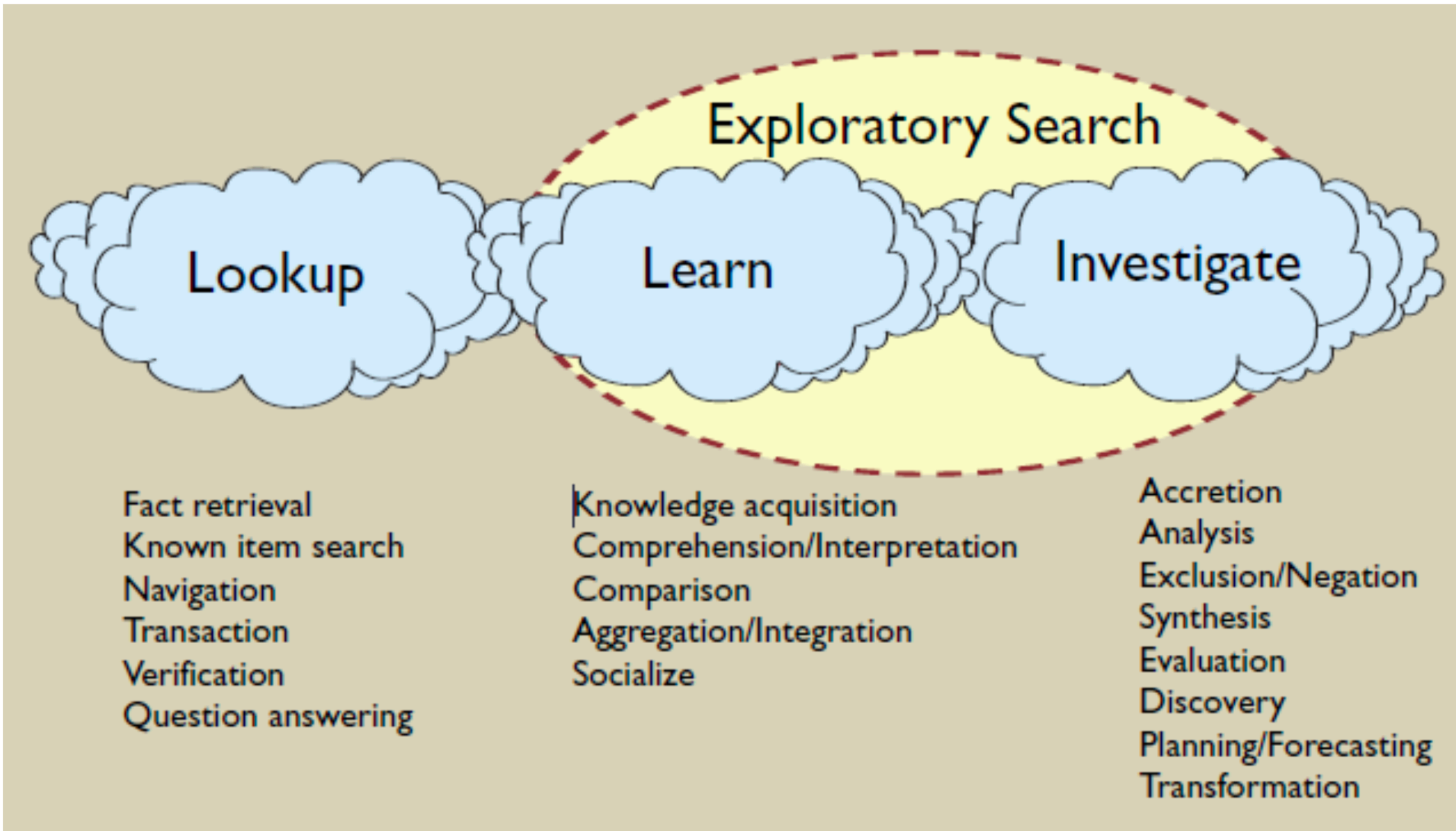
- Classic IR
  - Content-related search in unstructured documents
  - System-oriented view
  - **Static** information needs
- Interactive IR
  - Focus on user interaction with information system
  - **Dynamic** information needs

Berry-picking model of information seeking



M.Bates (1989). "**The design of browsing and berrypicking techniques for the online search interface**".
Online review, 13.5, pp. 407–424.

Marchionini's search activities



G.Marchionini (2006), **"Exploratory search: from finding to understanding",** Communications of the ACM, 49(4), 41-46.

Interactive Information Retrieval (IIR)

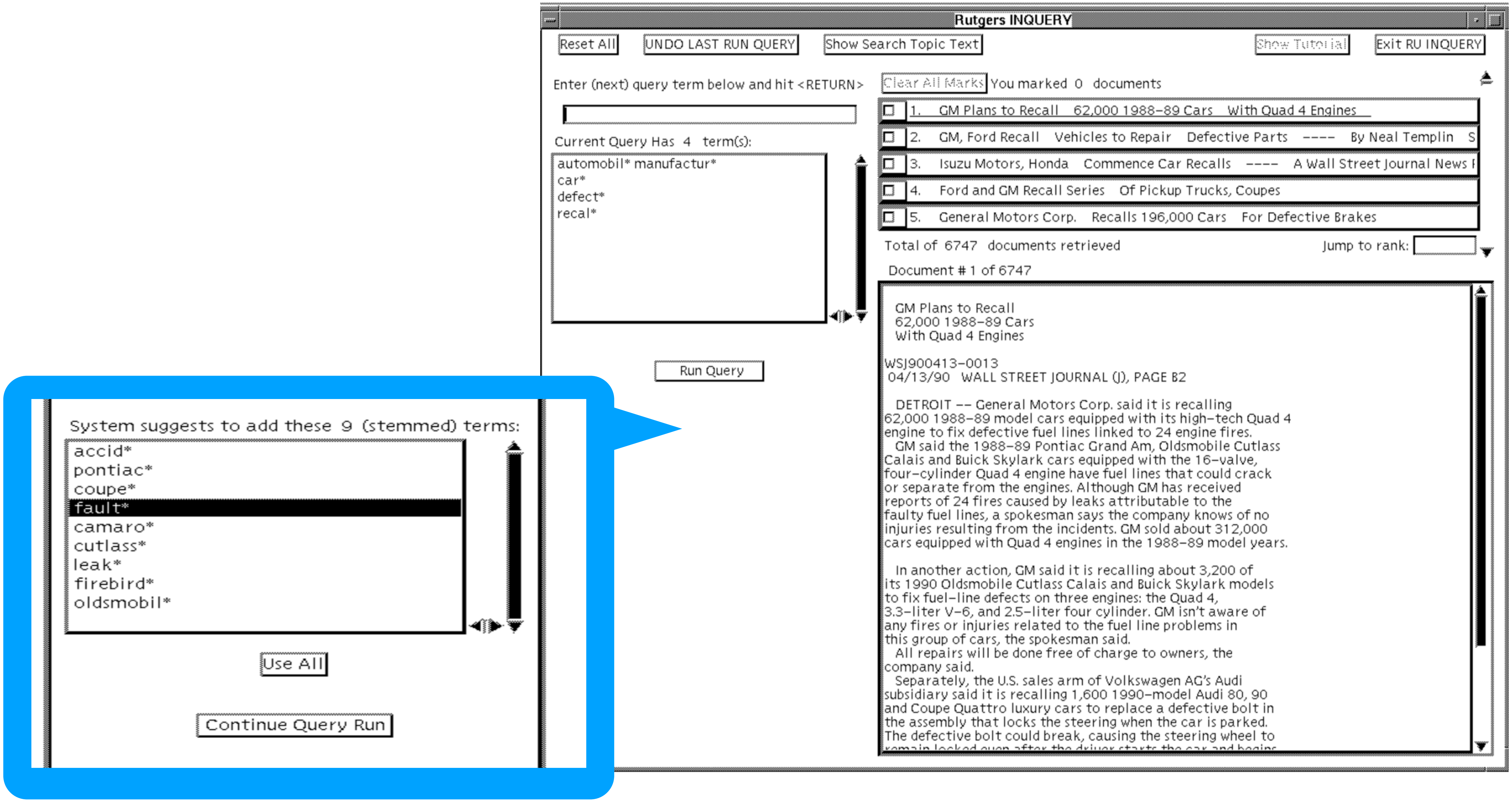
- Definition (Methods for evaluating interactive information retrieval systems with users, Kelly, 2009):

  "In interactive information retrieval (IIR), users are typically studied along with their interactions with systems and information. While classic IR studies abstract humans out of the evaluation model, ***IIR focuses on users' behaviors and experiences****—including physical, cognitive and affective—and the interactions that occur between users and systems, and users and information."*

Relevance in IIR

- **CANNOT** assume that the **relevance** of retrieved documents is a good proxy for whether the IIR system satisfies users' **information needs**
- Relevance assumed to be **subjective**: related to the user's knowledge, interests, etc.
- Cannot use relevance judgments, no substitute for real users!

IIR example (mid90s)



J. Koenemann, N. Belkin (1996). "**A case for interaction: A study of interactive information retrieval behavior and effectiveness**". ACM SIGCHI.

IIR example (2016)

- PULP video

IIR Evaluation

- "... there is no strong evaluation or experimental framework for IIR evaluations as there is for IR studies."
- Study design (search task, within-subjects vs. between-subjects)
- Measurements (think-aloud, observation, logging, questionnaires, semi-structured interviews)
- Data analysis (statistical tests, non-parametric tests, repeated-measures ANOVA, regression)

Search behavior measures

- Search behavior measures are logged by the interface or the backend
- Example from Kelly et al. 2015

Table 5. Search Behavior Measures

Measure	Definition
Queries	Total number of unique queries submitted by a participant when completing a task.
Query length	Average number of query terms in all unique queries issued for a task.
Unique query terms	Total number of unique query terms used by a participant when completing a task.
SERP clicks	Total number of clicks participants made on SERPs.
URLs visited	Total number of unique URLs visited by participants (includes URLs accessed directly and indirectly via SERP)
Queries w/o SERP clicks	Total number of unique queries where participants did not click on the search engine results page (SERP).
Time to completion	The amount of time (in seconds) participants spent completing search tasks.
SERP dwell time	Average time spent between issuing a new query and clicking on the first search result (in seconds).
Query diversity	Number of queries issued that were not issued by another participant completing the exact same task.
Query term diversity	Number of query terms used that were not used by another participant completing the exact same task.
URL diversity	Number of URLs visited that were not visited by another participant completing the exact same task.

D.Kelly, J.Arguello, A.Edwards, and W.Wu (2015). "**Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework**". ACM ICTIR, pp. 101–110.

Questionnaires

Table 3. Pre-Task Questionnaire Items

Interest & Knowledge	How interested are you to learn more about the topic of this task? How many times have you searched for information about this task? How much do you know about the topic of the task?
Task Complexity	How defined is this task in terms of the types of information needed to complete it? How defined is this task in terms of the steps required to complete it? How defined is this task in terms of its expected solution?
Expected Task Difficulty	How difficult do you think it will be to *search* for information for this task using a search engine? How difficult do you think it will be to *understand* the information the search engine finds? How difficult do you think it will be to *decide* if the information the search engine finds is *useful* for completing the task? How difficult do you think it will be to *integrate* the information the search engine finds? How difficult do you think it will be to determine *when you have enough* information to finish the task?

Table 4. Post-Task Questionnaire Items

Engagement	How enjoyable was it to do this task? How engaging did you find this task? How difficult was it to concentrate while you were doing this task?
Interest	How much did your interest in the task increase as you searched? How much did your knowledge of the task increase as you searched?
Experienced Task Difficulty	Same five items from Table 3 except items started with, "How difficult was it to …"
Overall Difficulty	Overall, how difficult was this task?
Overall Satisfaction	Overall, how satisfied are you with your solution to this task? Overall, how satisfied are you with the search strategy you took to solve this task?

D.Kelly, J.Arguello, A.Edwards, and W.Wu (2015). "**Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework**". ACM ICTIR, pp. 101–110.

Research questions and hypotheses

Research Questions, Hypotheses, and Theory. Explicit research questions were found in 19.3% of the studies ($n = 29$), explicit hypothesis were found in 10.7% ($n = 16$) of the studies, and both a research question and a hypothesis were found in 4.7% of the studies ($n = 7$). In 65.3% ($n = 98$) of the studies, there was neither an explicitly stated research question nor hypothesis.

[objectives are to] **"compare two search systems,"** which suggests an implicit research question focused on basic evaluation.

D. Kelly and C. Sugimoto (2013). "**A systematic review of interactive information retrieval evaluation studies, 1967–2006**". JASIST 64.4, pp. 745–770.

In the majority of studies ($n = 57$, 45%), ANOVA was used as the method of analysis. This was followed by *t* test ($n = 33$, 26%), Mann-Whitey ($n = 11$, 9%), chi-square ($n = 8$, 6%), and Wilcoxon signed-rank test ($n = 6$, 5%). Correlation, Kruskal-Wallis and factor analysis were observed in fewer than 5% of the articles. Fifteen percent ($n = 19$) of the articles presented only descriptive statistics, while 9% ($n = 11$) did not provide any indication of which type of analysis was used, despite claiming statistically significant results or presenting probability values. Almost all the analyses were performed variable-by-variable and were conducted to compare the systems. Only a small percentage of articles described statistical analyses that attempted to model performance using multiple input variables ($n = 6$, 5%).

statistics, while 9% ($n = 11$) did not provide any indication of which type of analysis was used, despite claiming statistically significant results or presenting probability values.

Researcher degrees-of-freedom

General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Coined the term "researcher degrees-of-freedom"

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments ... demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. ... and straightforwardly effective disclosure-based solution to this problem. The solution in... authors and four guidelines for reviewers, all of which impose a minimal burden on the pu...

Choice between two dependent variables nearly doubles false positive rate

Keywords

methodology, motivated reasoning, publication, disclosure

Received 3/17/11; Revision accepted 5/23/11

Our job as scientists is to discover truths about the world. We generate hypotheses, collect data, and examine whether or not the data are consistent with those hypotheses. Although we aspire to always be accurate, errors are inevitable.

Perhaps the most costly error is a *false positive*, the incorrect rejection of a null hypothesis. First, once they appear in the literature, false positives are particularly persistent. Because null results have many possible causes, failures to replicate previous findings are never conclusive. Furthermore, because it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive

Which control variables should be considered? Should specific measures be combined or transformed or both?

It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

This exploratory behavior is not the by-product of mali-

Reading

- M.Bates (1989). “The design of browsing and berrypicking techniques for the online search interface”. Online review, 13.5, pp. 407–424.
- **D.Kelly (2009). “Methods for evaluating interactive information retrieval systems with users”. Foundations and Trends in Information Retrieval, 3.1–2, pp. 1–224.**
- **C.Manning, P.Raghavan, and H.Schütze. (2010) "Introduction to information retrieval.". Chapter 8.**
- D.Kelly and C.Sugimoto (2013). “A systematic review of interactive information retrieval evaluation studies, 1967–2006”. JASIST 64.4, pp. 745–770.
- D.Kelly, J.Arguello, A.Edwards, and W.Wu (2015). “Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework”. ACM ICTIR, pp. 101–110.

Next lecture: Cognitive models and applications