Challenge Evaluations in Biomedical Information Retrieval

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

References

Amini, I, Martinez, D, et al. (2016). Improving patient record search: a meta-data based approach. *Information Processing & Management*. 52: 258-272.

Anonymous (2012). From Screen to Script: The Doctor's Digital Path to Treatment. New York, NY, Manhattan Research; Google. http://www.thinkwithgoogle.com/insights/library/studies/the-doctors-digital-path-to-treatment/

Baker, M (2016). 1,500 scientists lift the lid on reproducibility. *Nature*. 533: 452-454.

Bastian, H, Glasziou, P, et al. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine*. 7(9): e1000326. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000326

Blumenthal, D (2011). Implementation of the federal health information technology initiative. *New England Journal of Medicine*. 365: 2426-2431.

Blumenthal, D (2011). Wiring the health system--origins and provisions of a new federal program. *New England Journal of Medicine*. 365: 2323-2329.

Buckley, C and Voorhees, E (2000). Evaluating evaluation measure stability. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece. ACM Press. 33-40.

Buckley, C and Voorhees, EM (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, England. ACM Press. 25-32.

Demner-Fushman, D, Abhyankar, S, et al. (2012). NLM at TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute for Standards and Technology http://trec.nist.gov/pubs/trec21/papers/NLM.medical.final.pdf

Demner-Fushman, D, Abhyankar, S, et al. (2011). A knowledge-based approach to medical records retrieval. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology

Edinger, T, Cohen, AM, et al. (2012). Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. *AMIA 2012 Annual Symposium*, Chicago, IL. 180-188.

Egan, DE, Remde, JR, et al. (1989). Formative design-evaluation of Superbook. *ACM Transactions on Information Systems*. 7: 30-57.

Fidel, R and Soergel, D (1983). Factors affecting online bibliographic retrieval: a conceptual framework for research. *Journal of the American Society for Information Science*. 34: 163-180.

Fox, S (2011). Health Topics. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2011/HealthTopics.aspx

Harman, DK (2005). The TREC Ad Hoc Experiments. TREC: Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press: 79-98.

Hersh, W, Müller, H, et al. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*. 22: 648-655.

Hersh, W, Turpin, A, et al. (2001). Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management*. 37: 383-402.

Hersh, W and Voorhees, E (2009). TREC genomics special issue overview. *Information Retrieval*. 12: 1-15.

Hersh, WR (1994). Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*. 45: 201-206.

Hersh, WR (2001). Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management*. 37: 365-366.

Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.

Hersh, WR, Crabtree, MK, et al. (2002). Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*. 9: 283-293.

Hersh, WR and Greenes, RA (1990). SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. *Computers and Biomedical Research*. 23: 405-420.

Hersh, WR and Hickam, DH (1995). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science*. 46: 478-489.

Hersh, WR, Hickam, DH, et al. (1994). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*. 1: 51-60.

Hersh, WR, Müller, H, et al. (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association*. 13: 488-496.

Hersh, WR, Pentecost, J, et al. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*. 47: 50-56.

Ide, NC, Loane, RF, et al. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*. 14: 253-263.

Jarvelin, K and Kekalainen, J (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20: 422-446.

King, B, Wang, L, et al. (2011). Cengage Learning at TREC 2011 Medical Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology

Martinez, D, Otegi, A, et al. (2014). Improving search over electronic health records using UMLS-based query expansion through random walks. *Journal of Biomedical Informatics*. 51: 100-106.

Müller, H, Clough, P, et al., Eds. (2010). ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Heidelberg, Germany, Springer.

Mynatt, BT, Leventhal, LM, et al. (1992). Hypertext or book: which is better for answering questions? *Proceedings of Computer-Human Interface 92*. 19-25.

Roberts, K, Simpson, M, et al. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal*. 19: 113-148.

Safran, C, Bloomrosen, M, et al. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*. 14: 1-9.

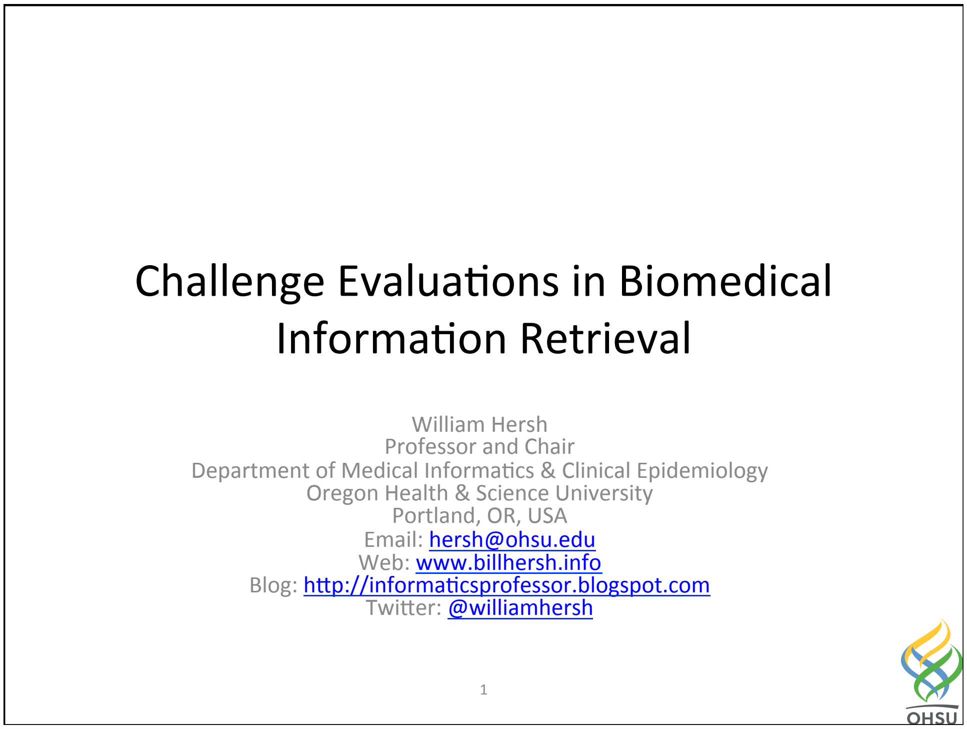
Stead, WW, Searle, JR, et al. (2011). Biomedical informatics: changing what physicians need to know and how they learn. *Academic Medicine*. 86: 429-434.

Tenenbaum, JD, Avillach, P, et al. (2016). An informatics research agenda to support precision medicine: seven key areas. *Journal of the American Medical Informatics Association*: Epub ahead of print.

Voorhees, E and Hersh, W (2012). Overview of the TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute of Standards and Technology http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf

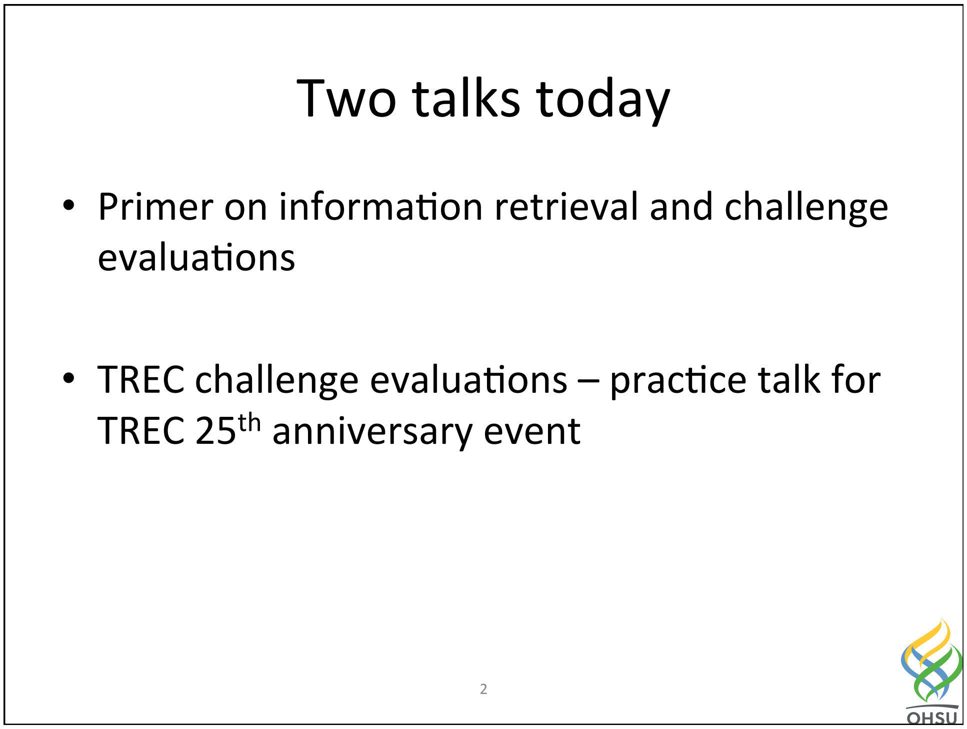
Voorhees, EM and Harman, DK, Eds. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA, MIT Press.

Yilmaz, E, Kanoulas, E, et al. (2008). A simple and efficient sampling method for estimating AP and NDCG. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore. 603-610.



Challenge Evaluations in Biomedical Information Retrieval

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh



Two talks today

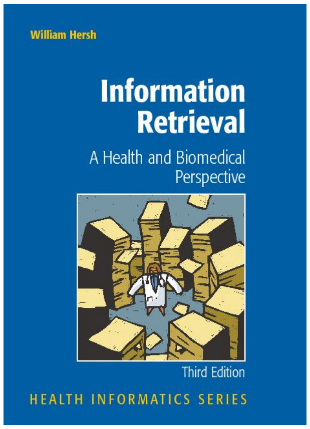
- Primer on information retrieval and challenge evaluations
- TREC challenge evaluations – practice talk for TREC 25th anniversary event

Learning objectives

- Define the evaluation measures used in information retrieval system evaluation and how they are used in challenge evaluations
- Describe the biomedical "tracks" in the Text Retrieval Conference (TREC) challenge evaluations
- Discuss the major results and findings of the TREC biomedical tracks

Information retrieval (IR, aka search)

- Focus on indexing and retrieval of (predominantly) knowledge-based information
- Historically centered on text in knowledge-based documents, but increasingly associated with many types of content
- www.irbook.info



William Hersh

Information Retrieval

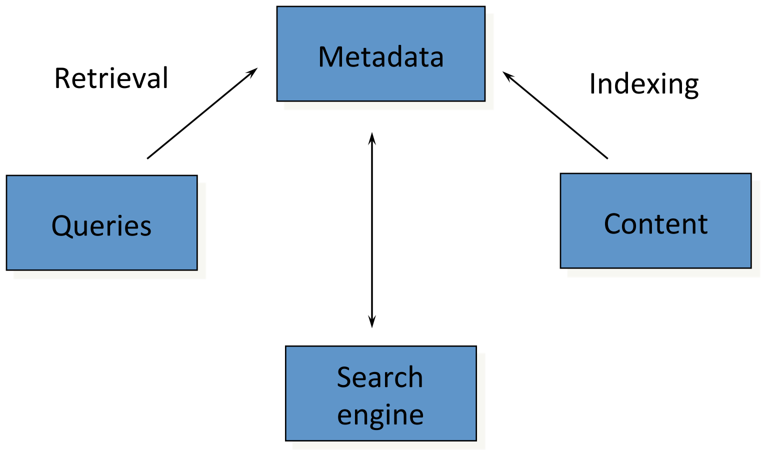
A Health and Biomedical Perspective

Third Edition

HEALTH INFORMATICS SERIES

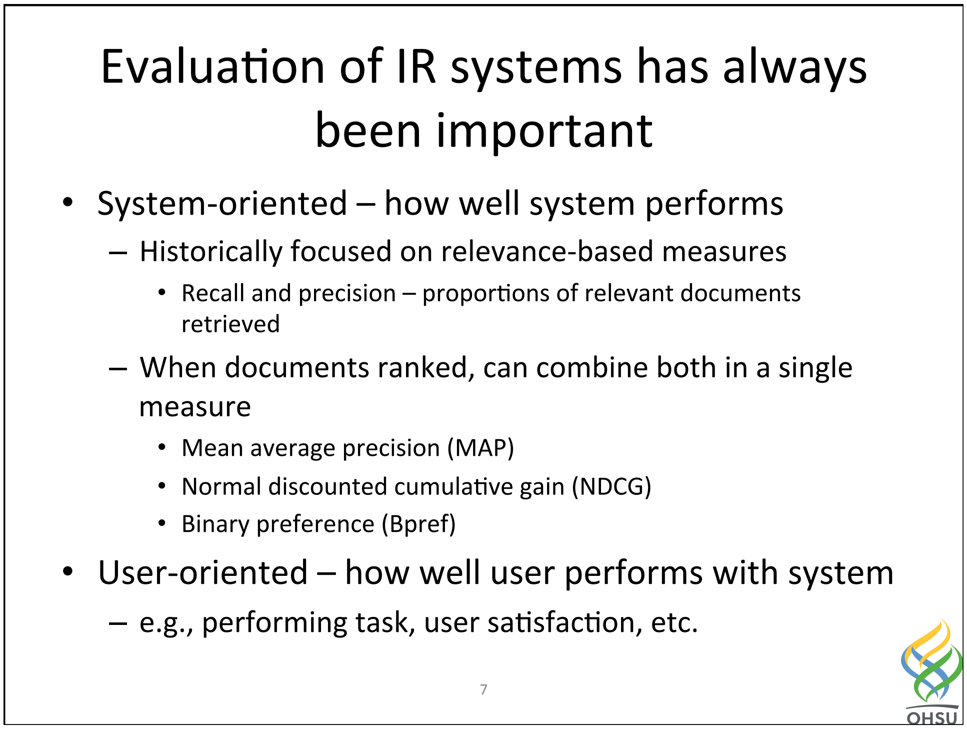
(Hersh, 2009)

Basics of IR



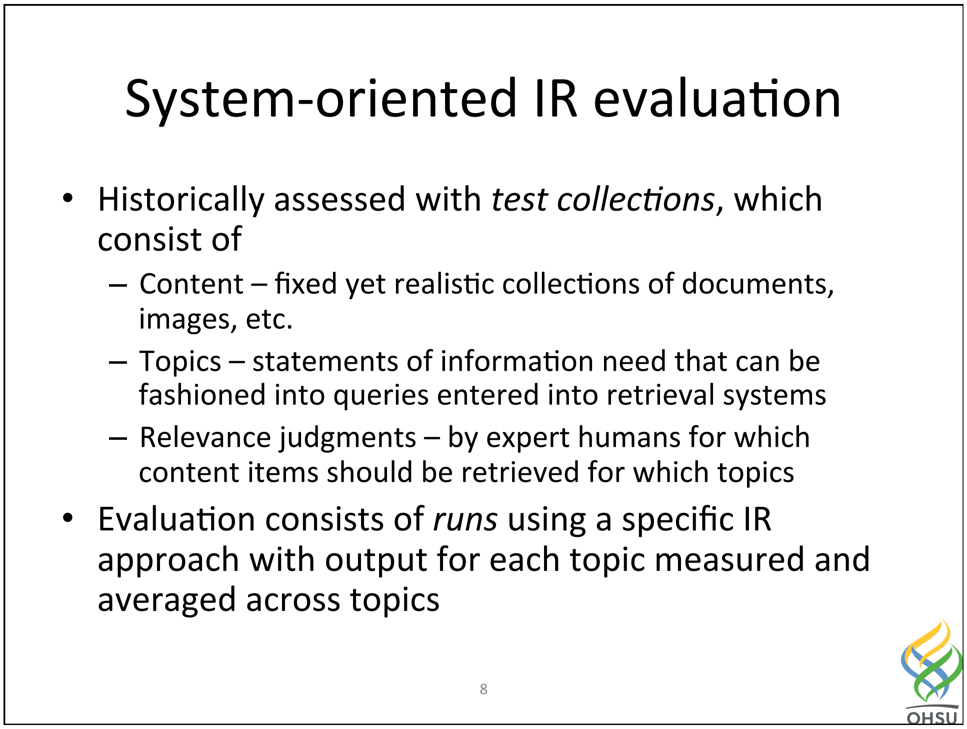
Use cases for IR

- Historically, retrieval of knowledge
  - Documents, especially journal articles (originally abstracts)
  - Multimedia – images, sounds, video, etc.
  - Hypermedia – Web-based content
- Newer foci
  - Clinical data – e.g., cohort discovery from electronic health records
  - Data – e.g., finding data sets



Evaluation of IR systems has always been important

- System-oriented – how well system performs
  - Historically focused on relevance-based measures
    - Recall and precision – proportions of relevant documents retrieved
  - When documents ranked, can combine both in a single measure
    - Mean average precision (MAP)
    - Normal discounted cumulative gain (NDCG)
    - Binary preference (Bpref)
- User-oriented – how well user performs with system
  - e.g., performing task, user satisfaction, etc.



System-oriented IR evaluation

- Historically assessed with *test collections*, which consist of
  - Content – fixed yet realistic collections of documents, images, etc.
  - Topics – statements of information need that can be fashioned into queries entered into retrieval systems
  - Relevance judgments – by expert humans for which content items should be retrieved for which topics
- Evaluation consists of *runs* using a specific IR approach with output for each topic measured and averaged across topics

Recall and precision

- Recall

$$R = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,relevant\ documents\ in\, collection}$$

  – Usually use *relative recall* when not all relevant documents known, where denominator is number of known relevant documents in collection

- Precision

$$P = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,retrieved\ documents}$$

Some measures can be combined into a single aggregated measure

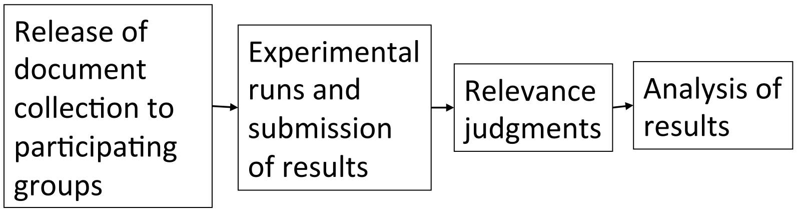
- *Mean average precision* (MAP) is mean of average precision for each topic (Harman, 2005)
  – Average precision is average of precision at each point of recall (relevant document retrieved)
  – Despite name, emphasizes recall
- *Bpref* accounts for when relevance information is significantly incomplete (Buckley, 2004)
- *Normal discounted cumulative gain* (NDCG) allows for graded relevance judgments (Jarvelin, 2002)
- MAP and NCDG can be "inferred" when there are incomplete judgments (Yilmaz, 2008)

Challenge evaluations

- A common approach in computer science, not limited to IR
- Develop a common task, data set, evaluation metrics, etc., ideally aiming for real-world size and representation for data, tasks, etc.
- In case of IR, this usually means
  - Test collection of content items
  - Topics of items to be retrieved – usually want 25-30 for "stability" (Buckley, 2000)
  - Runs from participating groups with retrieval for each topic
  - Relevance judgments of which content items are relevant to which topics – judged items derived from submitted runs

Challenge evaluations (cont.)

- Typical flow of events in an IR challenge evaluation

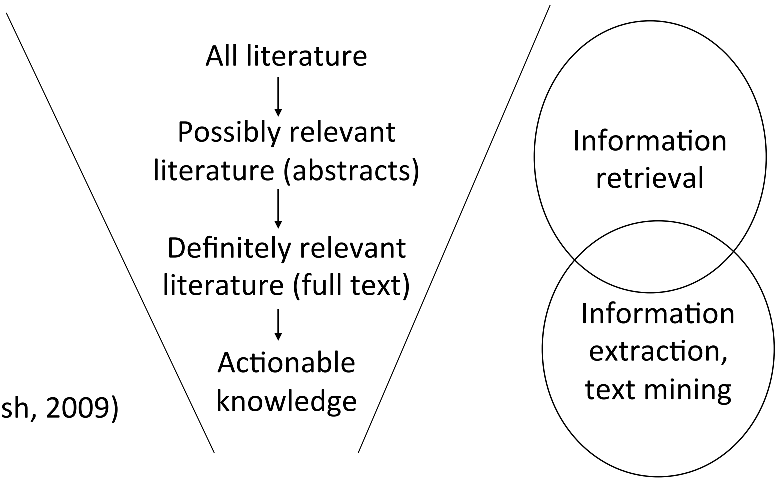

Release of document collection to participating groups
Experimental runs and submission of results
Relevance judgments
Analysis of results

- In IR, challenge evaluation results usually show wide variation between topics and between systems
  - Should be viewed as relative, not absolute performance
  - Averages can obscure variations

Some well-known challenge evaluations in IR

- Text Retrieval Conference (TREC, http://trec.nist.gov; Voorhees, 2005) – sponsored by National Institute for Standards and Technology (NIST), started in 1992
  - Many "tracks" of interest, such as routing/filtering, Web searching, question-answering, etc.
  - Mostly non-biomedical, but some tracks focused on genomics, EHRs, etc.
- Conferences and Labs of the Evaluation Forum (CLEF, www.clef-initiative.eu)
  - Started as track in TREC in 1996, spun off in 2000 to Cross-Language Evaluation Forum
  - Focus on retrieval across languages, European-based
  - Additional focus on image retrieval, which includes medical image retrieval tasks – www.imageclef.org (Hersh, 2009; Müller, 2010)
- TREC has inspired other challenge evaluations, e.g.,
  - i2b2 NLP Shared Task, https://www.i2b2.org/NLP/
  - bioCADDIE Dataset Retrieval Challenge – https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration

IR and text mining in context of biomedical knowledge management

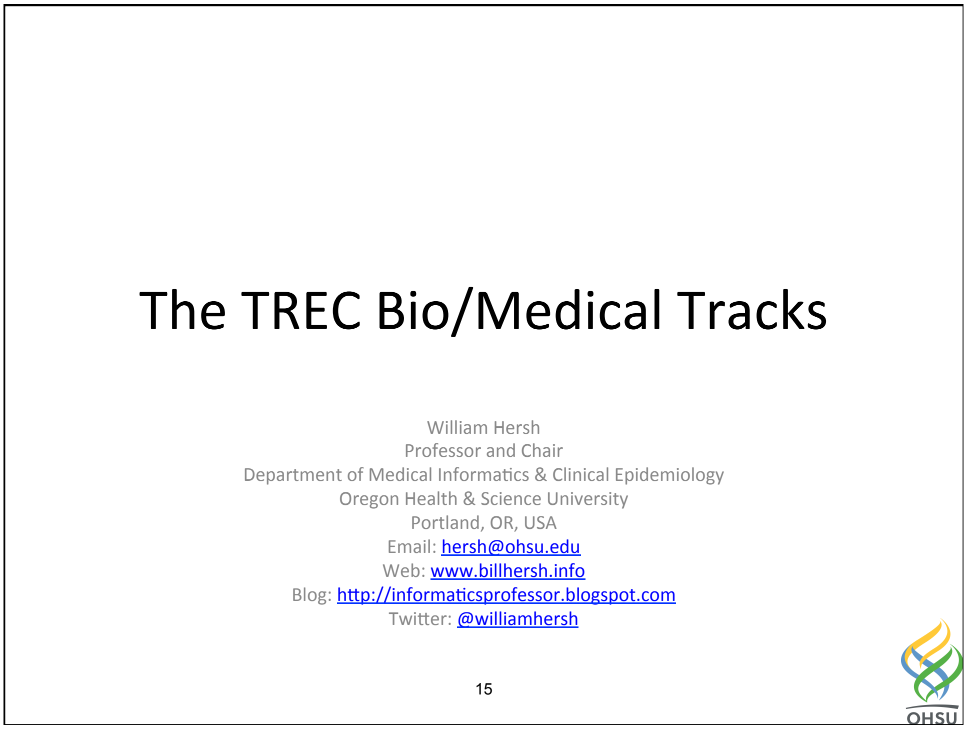


All literature → Possibly relevant literature (abstracts) → Definitely relevant literature (full text) → Actionable knowledge

Information retrieval

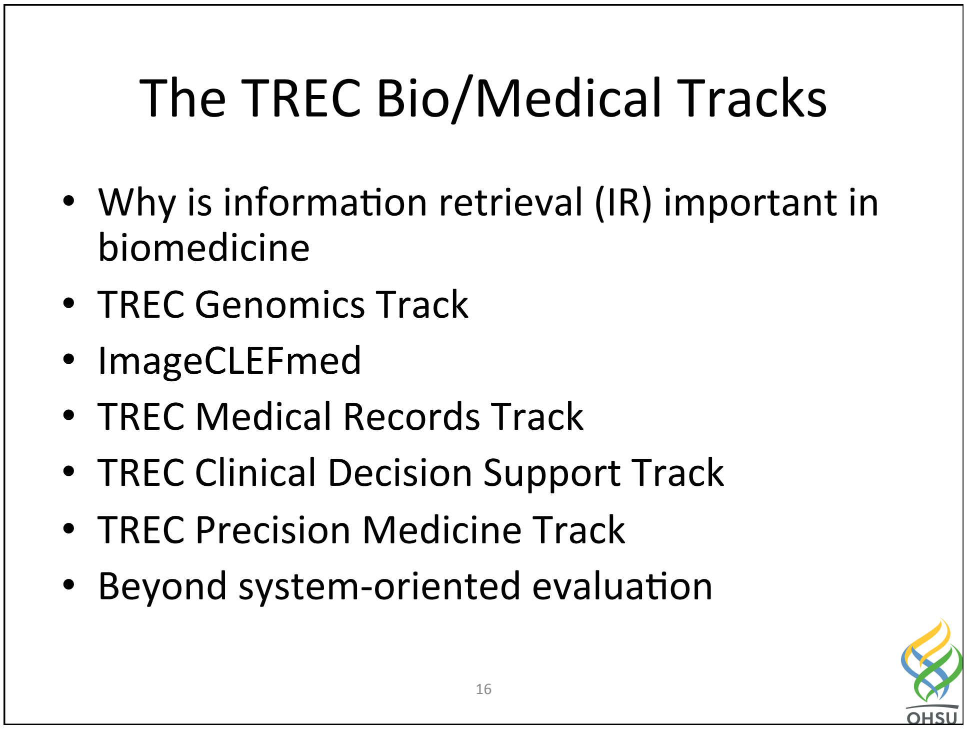
Information extraction, text mining

(Hersh, 2009)



The TREC Bio/Medical Tracks

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

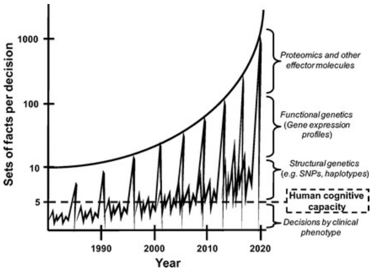


The TREC Bio/Medical Tracks

- Why is information retrieval (IR) important in biomedicine
- TREC Genomics Track
- ImageCLEFmed
- TREC Medical Records Track
- TREC Clinical Decision Support Track
- TREC Precision Medicine Track
- Beyond system-oriented evaluation

Why is IR important in health and biomedical domain?

- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day (Bastian, 2010)
- Data points per clinical decision increasing (Stead, 2011)
- Search for health information by clinicians, researchers, and patients/ consumers is ubiquitous (Fox, 2011; Google/Manhattan Research, 2012)
- Concerns about reproducibility of science (Baker, 2012)
- “Precision medicine” will increase quantity and complexity of data (Tenenbaum, 2016)



TREC Genomics Track (Hersh, 2009)

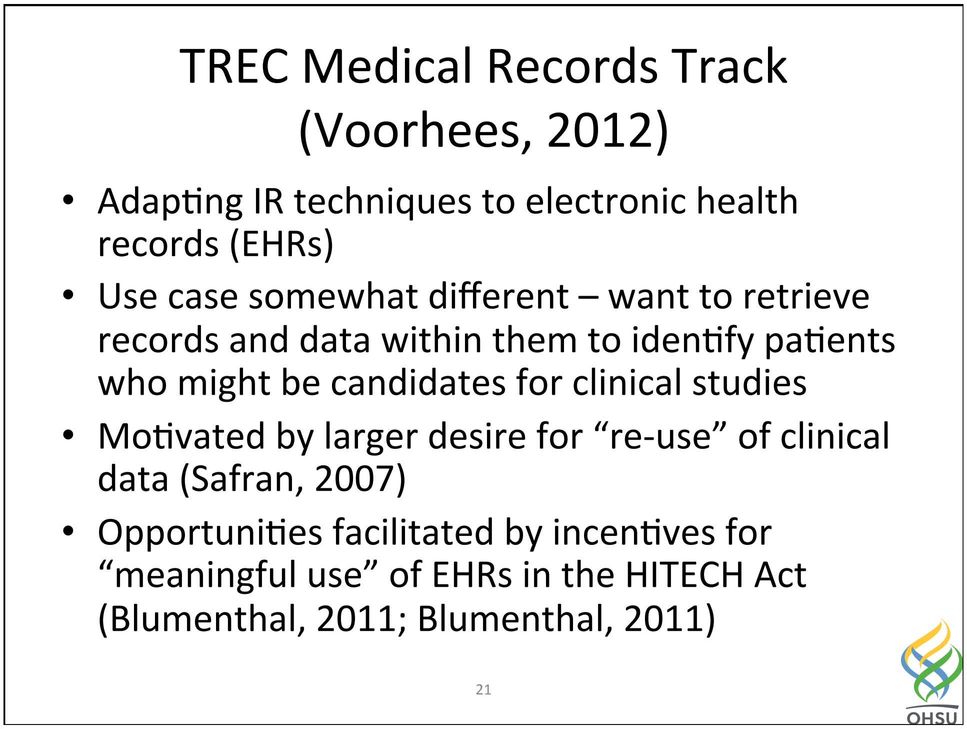
- Motivated by exploding research in genomics and inability to biologists to know all that might impact work
- First TREC track devoted to “domain-specific” retrieval, with focus on IR systems for genomics researchers
  - Supported by NSF Information Technology Research (ITR) grant
- History
  - 2004-2005 – focus on ad hoc retrieval and document categorization
  - 2006-2007 – focus on passage retrieval and question-answering as means to improve document retrieval

Lessons learned (Hersh, 2009)

- Ad hoc retrieval
  - Modest benefit for techniques known to work well in general IR, e.g., stop word removal, stemming, weighting
  - Query term expansion, especially domain-specific and/or done by humans, helped most
- QA
  - Most consistent benefit from query expansion and paragraph-length passage retrieval
- For all experiments (and papers describing them), major problems were
  - Lack of detailed description of systems
  - Use of low-performing baselines

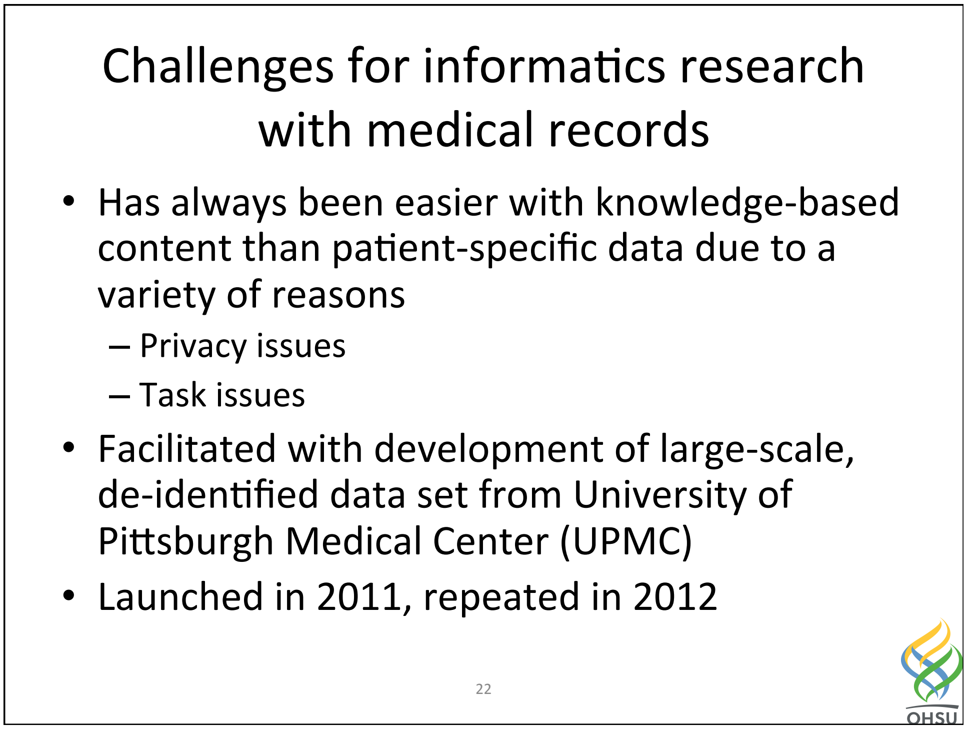
Image retrieval – ImageCLEF medical image retrieval task

- Biomedical professionals increasingly use images for research, clinical care, and education, yet we know very little about how to best retrieve them
- Developed test collection and exploration of information needs motivating use of image retrieval systems (Hersh, 2006; Hersh, 2009; Müller, 2010)
- Started with ad hoc retrieval and added tasks
  - Modality detection
  - Case finding
- Overall conclusions: text yielded most consistent results with image features providing variable value
- Continues on with highly defined tasks



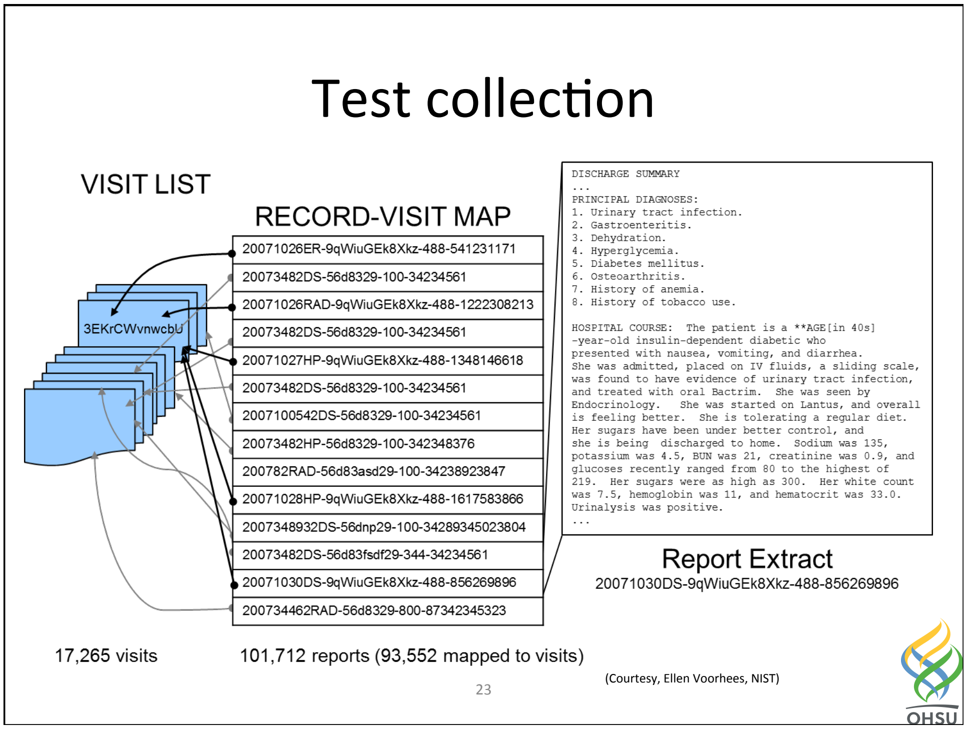
TREC Medical Records Track (Voorhees, 2012)

- Adapting IR techniques to electronic health records (EHRs)
- Use case somewhat different – want to retrieve records and data within them to identify patients who might be candidates for clinical studies
- Motivated by larger desire for "re-use" of clinical data (Safran, 2007)
- Opportunities facilitated by incentives for "meaningful use" of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011)



Challenges for informatics research with medical records

- Has always been easier with knowledge-based content than patient-specific data due to a variety of reasons
  - Privacy issues
  - Task issues
- Facilitated with development of large-scale, de-identified data set from University of Pittsburgh Medical Center (UPMC)
- Launched in 2011, repeated in 2012



Test collection

VISIT LIST

RECORD-VISIT MAP

20071026ER-9qWiuGEk8Xkz-488-541231171
20073482DS-56d8329-100-34234561
20071026RAD-9qWiuGEk8Xkz-488-1222308213
20073482DS-56d8329-100-34234561
20071027HP-9qWiuGEk8Xkz-488-1348146618
20073482DS-56d8329-100-34234561
2007100542DS-56d8329-100-34234561
20073482HP-56d8329-100-342348376
200782RAD-56d83asd29-100-34238923847
20071028HP-9qWiuGEk8Xkz-488-1617583866
2007348932DS-56dnp29-100-34289345023804
20073482DS-56d83fsdf29-344-34234561
20071030DS-9qWiuGEk8Xkz-488-856269896
200734462RAD-56d8329-800-87342345323

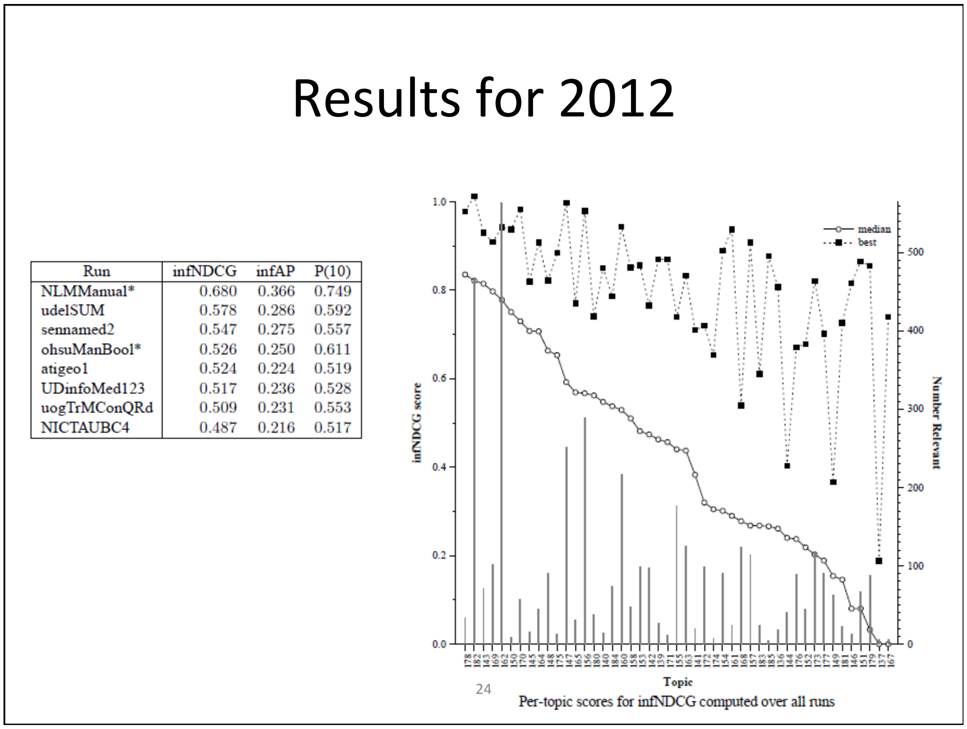
```
DISCHARGE SUMMARY
...
PRINCIPAL DIAGNOSES:
1. Urinary tract infection.
2. Gastroenteritis.
3. Dehydration.
4. Hyperglycemia.
5. Diabetes mellitus.
6. Osteoarthritis.
7. History of anemia.
8. History of tobacco use.

HOSPITAL COURSE:  The patient is a **AGE[in 40s]
-year-old insulin-dependent diabetic who
presented with nausea, vomiting, and diarrhea.
She was admitted, placed on IV fluids, a sliding scale,
was found to have evidence of urinary tract infection,
and treated with oral Bactrim.  She was seen by
Endocrinology.   She was started on Lantus, and overall
is feeling better.  She is tolerating a regular diet.
Her sugars have been under better control, and
she is being  discharged to home.  Sodium was 135,
potassium was 4.5, BUN was 21, creatinine was 0.9, and
glucoses recently ranged from 80 to the highest of
219.  Her sugars were as high as 300.  Her white count
was 7.5, hemoglobin was 11, and hematocrit was 33.0.
Urinalysis was positive.
...
```

Report Extract
20071030DS-9qWiuGEk8Xkz-488-856269896

17,265 visits

101,712 reports (93,552 mapped to visits)

(Courtesy, Ellen Voorhees, NIST)



Results for 2012

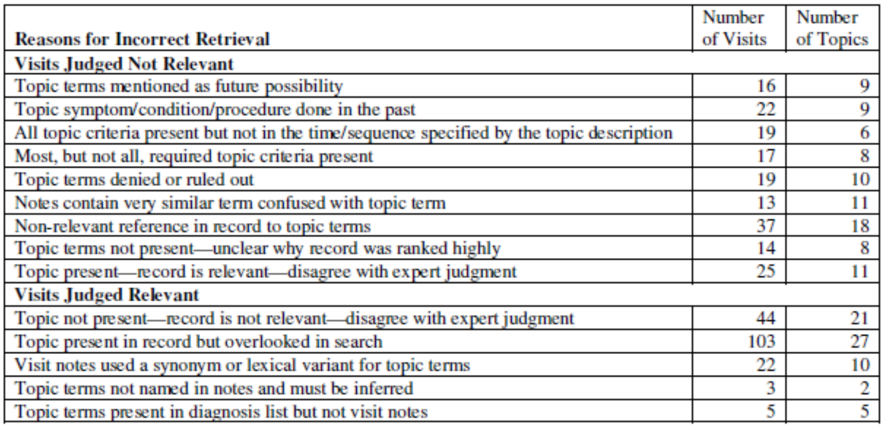
Run	infNDCG	infAP	P(10)
NLMManual*	0.680	0.366	0.749
udelSUM	0.578	0.286	0.592
sennamed2	0.547	0.275	0.557
ohsuManBool*	0.526	0.250	0.611
atigeo1	0.524	0.224	0.519
UDinfoMed123	0.517	0.236	0.528
uogTrMConQRd	0.509	0.231	0.553
NICTAUBC4	0.487	0.216	0.517

Per-topic scores for infNDCG computed over all runs

Which approaches did (and did not) work?

- Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011; Demner-Fushman, 2012)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
- Many approaches known to work in general IR fared less well, e.g., term expansion, document focusing, etc.
  - Other domain-specific approaches also did not show benefit, e.g., creation of PICO frames, negation
- Some success with
  - Results filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus (King, 2011)
  - Expansion by concepts and relationships in UMLS Metathesaurus (Martinez, 2014)
  - Pseudorelevance feedback using ICD-9 codes (Amini, 2016)

Failure analysis for 2011 topics (Edinger, 2012)



Reasons for Incorrect Retrieval	Number of Visits	Number of Topics
Visits Judged Not Relevant		
Topic terms mentioned as future possibility	16	9
Topic symptom/condition/procedure done in the past	22	9
All topic criteria present but not in the time/sequence specified by the topic description	19	6
Most, but not all, required topic criteria present	17	8
Topic terms denied or ruled out	19	10
Notes contain very similar term confused with topic term	13	11
Non-relevant reference in record to topic terms	37	18
Topic terms not present—unclear why record was ranked highly	14	8
Topic present—record is relevant—disagree with expert judgment	25	11
Visits Judged Relevant		
Topic not present—record is not relevant—disagree with expert judgment	44	21
Topic present in record but overlooked in search	103	27
Visit notes used a synonym or lexical variant for topic terms	22	10
Topic terms not named in notes and must be inferred	3	2
Topic terms present in diagnosis list but not visit notes	5	5

TREC Clinical Decision Support Track (Roberts, 2016)

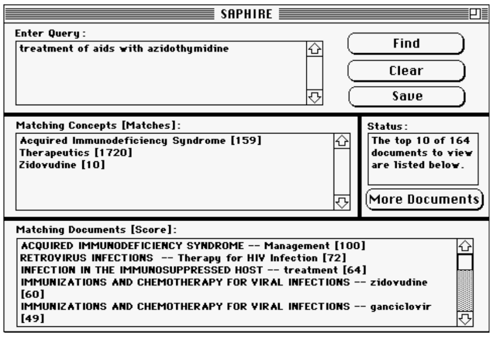
- www.trec-cds.org
- Ad hoc search of biomedical literature (PubMed Central Open Access Subset – 1.25M articles)
- Topics are patient descriptions in three information need categories
  - Diagnosis
  - Test
  - Treatment
- Currently in third year of operation
- Transitioning to Precision Medicine Track

TREC has inspired and guided other challenge evaluations in biomedicine

- i2b2
  - https://www.i2b2.org/NLP
  - Various NLP-related tasks, including extraction and de-identification
- CLEF eHealth
  - https://sites.google.com/site/clefehealth/home
  - Information extraction and patient-centered IR
- bioCADDIE
  - https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration
  - Data set retrieval

System-oriented retrieval is not enough

- My initial focused on concept-based searching (Hersh, 1990)
  - Did not impart value over word indexing and searching (Hersh, JAMIA, 1994)
- Experience of several evaluations led to concern with evaluation focus on recall/precision (Hersh, JASIS, 1994)
  - How much difference is meaningful?
  - How valid is batch evaluation for understand how well user will search?



Enter Query:
treatment of aids with azidothymidine

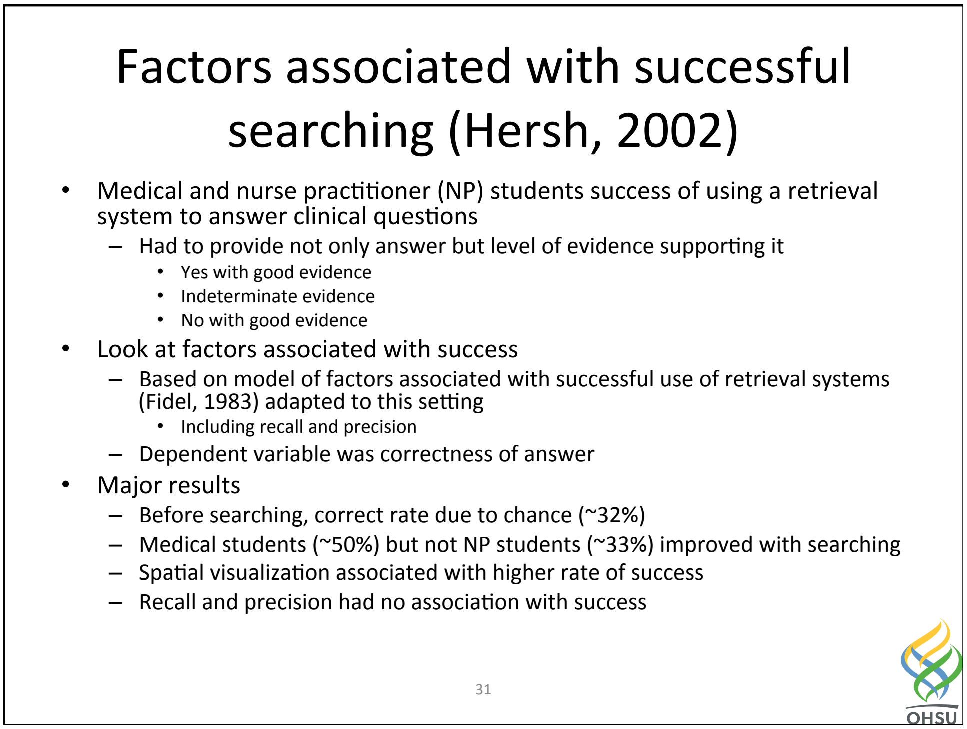
Matching Concepts [Matches]:
Acquired Immunodeficiency Syndrome [159]
Therapeutics [1720]
Zidovudine [10]

Status:
The top 10 of 164 documents to view are listed below.

Matching Documents [Score]:
ACQUIRED IMMUNODEFICIENCY SYNDROME -- Management [100]
RETROVIRUS INFECTIONS -- Therapy for HIV Infection [72]
INFECTION IN THE IMMUNOSUPPRESSED HOST -- treatment [64]
IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- zidovudine [60]
IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- ganciclovir [49]

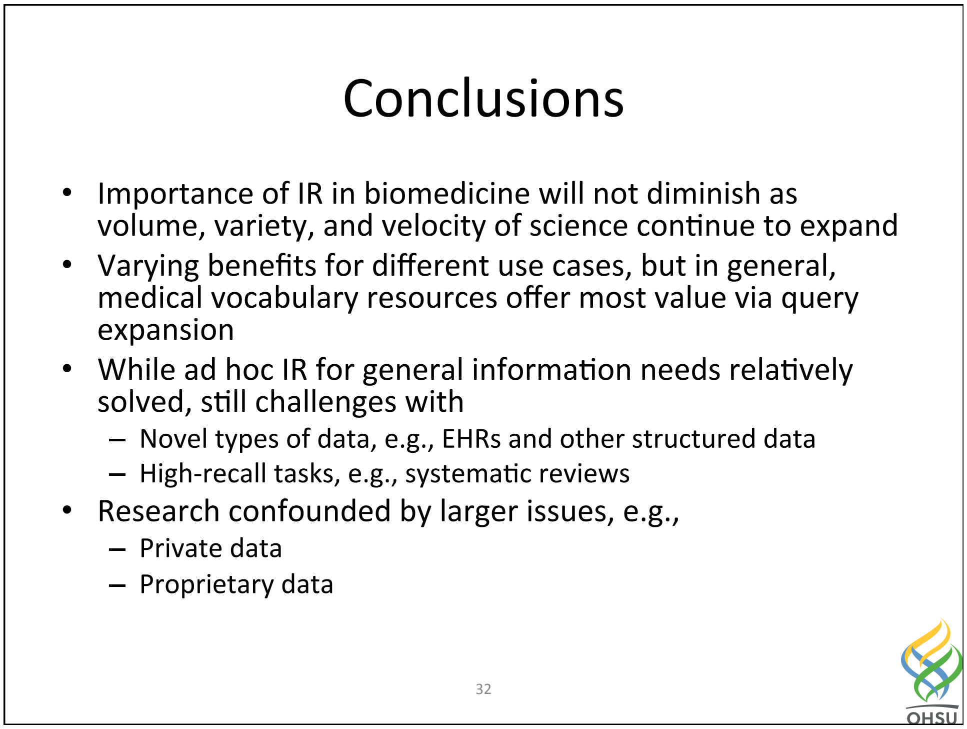
Led to "task-oriented" evaluation approaches

- Motivated by Egan (1989) and Mynatt (1992)
- Major task in medicine: answering questions
- How can we evaluate systems in interactive use for answering questions?
- Undertook parallel approaches in
  - Medicine – using
    - Electronic textbook – Scientific American Medicine (Hersh, 1995)
    - Bibliographic database – MEDLINE (Hersh, 1996)
  - General news – TREC Interactive Track (Hersh, 2001)



Factors associated with successful searching (Hersh, 2002)

- Medical and nurse practitioner (NP) students success of using a retrieval system to answer clinical questions
  - Had to provide not only answer but level of evidence supporting it
    - Yes with good evidence
    - Indeterminate evidence
    - No with good evidence
- Look at factors associated with success
  - Based on model of factors associated with successful use of retrieval systems (Fidel, 1983) adapted to this setting
    - Including recall and precision
  - Dependent variable was correctness of answer
- Major results
  - Before searching, correct rate due to chance (~32%)
  - Medical students (~50%) but not NP students (~33%) improved with searching
  - Spatial visualization associated with higher rate of success
  - Recall and precision had no association with success



Conclusions

- Importance of IR in biomedicine will not diminish as volume, variety, and velocity of science continue to expand
- Varying benefits for different use cases, but in general, medical vocabulary resources offer most value via query expansion
- While ad hoc IR for general information needs relatively solved, still challenges with
  - Novel types of data, e.g., EHRs and other structured data
  - High-recall tasks, e.g., systematic reviews
- Research confounded by larger issues, e.g.,
  - Private data
  - Proprietary data