Biomedical Information Retrieval

William Hersh, MD
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

References

Alsheikh-Ali, AA, Qureshi, W, et al. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*. 6(9): e24357. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0024357

Anonymous (2006). Fatally Flawed - Refuting the recent study on encyclopedic accuracy by the journal Nature. Chicago, IL, Encyclopedia Brittanica. http://corporate.britannica.com/britannica_nature_response.pdf

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

Anonymous (2015). The Beginner's Guide to SEO. Seattle, WA, Moz. http://moz.com/beginners-guide-to-seo

Anonymous (2016). Toward fairness in data sharing. *New England Journal of Medicine*. 375: 405-407.

Anonymous (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 45: D12-D17.

Bachrach, CA and Charen, T (1978). Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Medical Informatics*. 3: 237-254.

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

Brin, S and Page, L (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 30: 107-117. http://infolab.stanford.edu/pub/papers/google.pdf

Broder, A (2002). A taxonomy of Web search. *SIGIR Forum*. 36(2): 3-10. http://www.acm.org/sigir/forum/F2002/broder.pdf

Castillo, C and Davison, BD (2011). Adversarial Web Search. Delft, Netherlands, now Publishers.

Cerrato, P (2012). IBM Watson Finally Graduates Medical School. Information Week, October 23, 2012. http://www.informationweek.com/healthcare/clinical-systems/ibm-watson-finally-graduates-medical-sch/240009562

Coletti, MH and Bleich, HL (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*. 8: 317-323.

Davies, K (2006). Search and Deploy. Bio-IT World, October 16, 2006. http://www.bio-itworld.com/issues/2006/oct/biogen-idec/

DeAngelis, CD, Drazen, JM, et al. (2005). Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. *Journal of the American Medical Association*. 293: 2927-2929.

Ferrucci, D, Brown, E, et al. (2010). Building Watson: an overview of the DeepQA Project. *AI Magazine*. 31(3): 59-79. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303
Ferrucci, DA (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*. 56(3/4): 1. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6177724
Fox, S (2011). Health Topics. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2011/HealthTopics.aspx
Fox, S (2011). The Social Life of Health Information, 2011. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx
Fox, S and Duggan, M (2013). Health Online 2013. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2013/Health-online.aspx
Funk, ME and Reid, CA (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*. 71: 176-183.
Giles, J (2005). Internet encyclopaedias go head to head. *Nature*. 438: 900-901. http://www.nature.com/nature/journal/v438/n7070/full/438900a.html
Gorman, PN (1995). Information needs of physicians. *Journal of the American Society for Information Science*. 46: 729-736.
Hanbury, A, Müller, H, et al. (2015). Evaluation-as-a-Service: Overview and Outlook, arXiv. http://arxiv.org/pdf/1512.07454v1
Haynes, RB, McKibbon, KA, et al. (1990). Online access to MEDLINE in clinical settings. *Annals of Internal Medicine*. 112: 78-84.
Heilman, J (2013). Online encyclopedia provides free health info for all. *Bulletin of the World Health Organization*. 91: 8-9.
Hersh, W, Müller, H, et al. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*. 22: 648-655.
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 (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.
Hersh, WR, Bhupatiraju, RT, et al. (2006). Enhancing access to the bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*. 1: 3. http://www.j-biomed-discovery.com/content/1/1/3
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, Crabtree, MK, et al. (2000). Factors associated with successful answering of clinical questions using an information retrieval system. *Bulletin of the Medical Library Association*. 88: 323-331.
Hersh, WR and Hickam, DH (1998). How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature. *Journal of the American Medical Association*. 280: 1347-1352.
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.
Holan, AD (2016). 2016 Lie of the Year: Fake news. St. Petersburg, FL, Politifact. http://www.politifact.com/truth-o-meter/article/2016/dec/13/2016-lie-year-fake-news/
Huesch, MD (2013). Privacy threats when seeking online health information. *JAMA Internal Medicine*. 173: 1838-1839.

Insel, TR, Volkow, ND, et al. (2003). Neuroscience networks: data-sharing in an information age. *PLoS Biology*. 1: E17.
Kalpathy-Cramer, J, SecodeHerrera, AG, et al. (2015). Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at ImageCLEF 2004–2013. *Computerized Medical Imaging and Graphics*. 39: 55-61.
Laine, C, Horton, R, et al. (2007). Clinical trial registration: looking back and moving ahead. *Journal of the American Medical Association*. 298: 93-94.
Laurent, MR and Vickers, TJ (2009). Seeking health information online: does Wikipedia matter? *Journal of the American Medical Informatics Association*. 16: 471-479.
Lee, JS, Lorincz, C, et al. (2011). Should Healthcare Organizations Use Social Media? Falls Church, VA, Computer Sciences Corp. http://assets1.csc.com/health_services/downloads/CSC_Should_Healthcare_Organizations_Use_Social_Media.pdf
Libert, T (2015). Privacy implications of health information seeking on the Web. *Communications of the ACM*. 58(3): 68-77.
Lohr, S (2012). The Future of High-Tech Health Care — and the Challenge. New york, NY. New York Times. February 13, 2012. http://bits.blogs.nytimes.com/2012/02/13/the-future-of-high-tech-health-care-and-the-challenge/
Magrabi, F, Coiera, EW, et al. (2005). General practitioners' use of online evidence during consultations. *International Journal of Medical Informatics*. 74: 1-12.
Marcetich, J, Rappaport, M, et al. (2004). Indexing consistency in MEDLINE. *MLA 04 Abstracts*, Washington, DC. Medical Library Association. 10-11.
Markoff, J (2011). Computer Wins on 'Jeopardy!': Trivial, It's Not. New York, NY. New York Times. February 16, 2011. http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html
McHenry, R (2004). The Faith-Based Encyclopedia. Tech Central Station, November 15, 2004. http://www.techcentralstation.com/111504A.html
Mello, MM, Francer, JK, et al. (2013). Preparing for responsible sharing of clinical trial data. *New England Journal of Medicine*. 369: 1651-1658.
Metzger, J and Rhoads, J (2012). Summary of Key Provisions in Final Rule for Stage 2 HITECH Meaningful Use. Falls Church, VA, Computer Sciences Corp. http://skynetehr.com/PDFFiles/MeaningUse_Stage2.pdf
Nicholson, DT (2006). An evaluation of the quality of consumer health information on Wikipedia Capstone, Oregon Health & Science University.
Nielsen, J and Levy, J (1994). Measuring usability: preference vs. performance. *Communications of the ACM*. 37: 66-75.
Perrin, A (2015). One-fifth of Americans report going online 'almost constantly'. Washington, DC, Pew Research Center. http://www.pewresearch.org/fact-tank/2015/12/08/one-fifth-of-americans-report-going-online-almost-constantly/
Pluye, P and Grad, RM (2004). How information retrieval technology may impact on physician practice: an organizational case study in family medicine. *Journal of Evaluation in Clinical Practice*. 10: 413-430.
Pluye, P, Grad, RM, et al. (2005). Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies. *International Journal of Medical Informatics*. 74: 745-768.
Purcell, K, Brenner, J, et al. (2012). Search Engine Use 2012. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx
Rodwin, MA and Abramson, JD (2012). Clinical trial data as a public good. *Journal of the American Medical Association*. 308: 871-872.

Roegiest, A and Cormack, GV (2016). An architecture for privacy-preserving and replicable high-recall retrieval experiments. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy. 1085-1088.
Ross, JS and Krumholz, HM (2013). Ushering in a new era of open science through data sharing: the wall must come down. *Journal of the American Medical Association*. 309: 1355-1356.
Royle, JA, Blythe, J, et al. (1995). Literature search and retrieval in the workplace. *Computers in Nursing*. 13: 25-31.
Salton, G (1991). Developments in automatic text retrieval. *Science*. 253: 974-980.
Sánchez-Mendiola, M and Martínez-Franco, AI, Eds. (2014). Informática Biomédica, 2a Edición. Mexico City, MX, Elsevier.
Shortliffe, EH and Cimino, JJ, Eds. (2014). Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Fourth Edition). London, England, Springer.
Smith, M (2014). Targeted: How Technology Is Revolutionizing Advertising and the Way Companies Reach Consumers. Washington, DC, AMACOM.
Stanfill, MH, Williams, M, et al. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*. 17: 646-651.
Strzalkowski, T and Harabagiu, S, Eds. (2006). Advances in Open-Domain Question Answering. Dordrecht, Netherlands, Springer.
Taylor, H (2010). "Cyberchondriacs" on the Rise? Those who go online for healthcare information continues to increase. Rochester, NY, Harris Interactive. http://www.harrisinteractive.com/vault/HI-Harris-Poll-Cyberchondriacs-2010-08-04.pdf
Tuason, O, Chen, L, et al. (2004). Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pacific Symposium on Biocomputing*, Kona, Hawaii. World Scientific. 238-249.
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 (2005). Question Answering in TREC. TREC - Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press: 233-257.
Voorhees, EM and Harman, DK, Eds. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA, MIT Press.
Voorhees, EM and Tong, RM (2011). Overview of the TREC 2011 Medical Records Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute of Standards and Technology
Wanke, LA and Hewison, NS (1988). Comparative usefulness of MEDLINE searches performed by a drug information pharmacist and by medical librarians. *American Journal of Hospital Pharmacy*. 45: 2507-2510.
Westbrook, JI, Gosling, AS, et al. (2005). The impact of an online evidence system on confidence in decision making in a controlled setting. *Medical Decision Making*. 25: 178-185.
Wu, S, Liu, S, et al. (2017). Intra-institutional EHR collections for patient-level information retrieval. *Journal of the American Society for Information Science & Technology*: in press.
Yandell, MD and Majoros, WH (2002). Genomics and natural language processing. *Nature Reviews - Genetics*. 3: 601-610.
Zarin, DA and Tse, T (2013). Trust but verify: trial registration and determining fidelity to the protocol. *Annals of Internal Medicine*. 159: 65-67.
Zarin, DA, Tse, T, et al. (2015). The proposed rule for U.S. clinical trial registration and results submission. *New England Journal of Medicine*. 372: 174-180.
Zarin, DA, Tse, T, et al. (2011). The ClinicalTrials.gov results database--update and key issues. *New England Journal of Medicine*. 364: 852-860.

Biomedical Information Retrieval

William Hersh, MD
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

Topics to cover

- Content
- Indexing
- Evaluation

Content

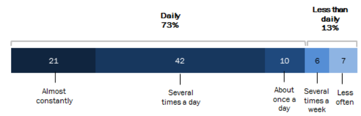
- Current status and challenges in biomedical information retrieval (IR)
- Classification and examples of knowledge-based information

Challenges in biomedical IR

- We have gone from information paucity to information overload
- Many topics we want to search on have multiple ways to be expressed
  - e.g., diseases, genes, symptoms, etc.
- The converse is a problem too: Many words and terms used to express topics have multiple meanings
- Balancing open access vs. providing for cost of production and maintenance

IR is now "mainstream"

- Internet (and likely search engine) use is now ubiquitous
  - Not only in developed countries (Perrin, 2015) but across world – http://www.internetworldstats.com/stats.htm
- 71% of Internet users (59% of US adults) have searched for health information, with 35% using it for self-diagnosis (Fox, 2013)
- "Search engine optimization" (SEO) is a key function used by many companies and organizations (Moz, 2015)
  - https://moz.com/beginners-guide-to-seo
  - Some are lucky, e.g., last name of "Hersh"



WORLD INTERNET USAGE AND POPULATION STATISTICS
MARCH 25, 2017 - Update

World Regions	Population (2017 Est.)	Population % of World	Internet Users 31 Mar 2017	Penetration Rate (% Pop.)	Growth 2000-2017	Users % Table
Africa	1,246,504,865	16.6 %	**345,676,501**	27.7 %	7,557.2%	9.3 %
Asia	4,148,177,672	55.2 %	**1,873,856,654**	45.2 %	1,539.4%	50.2 %
Europe	822,710,362	10.9 %	**636,971,824**	77.4 %	506.1%	17.1 %
Latin America / Caribbean	647,604,645	8.6 %	**385,919,382**	59.6 %	2,035.8%	10.3 %
Middle East	250,327,574	3.3 %	**141,931,765**	56.7 %	4,220.9%	3.8 %
North America	363,224,006	4.8 %	**320,068,243**	88.1 %	196.1%	8.6 %
Oceania / Australia	40,479,846	0.5 %	**27,549,054**	68.1 %	261.5%	0.7 %
WORLD TOTAL	**7,519,028,970**	**100.0 %**	**3,731,973,423**	**49.6 %**	**933.8%**	**100.0 %**

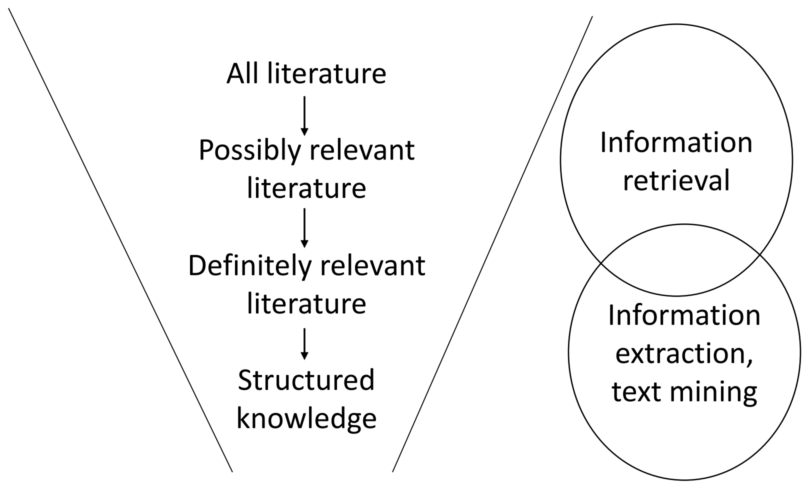
5

The Web has changed the nature of search

- Three major uses (Broder, 2002)
  - Informational – seeking information (39-48%)
  - Navigational – looking for a specific page, e.g., a home page (20-24%)
  - Transactional – perform transactions, e.g., on-line purchasing (30-36%)
- We are in the era of "adversarial" search – there is content we do not want to retrieve (Castillo, 2011; Smith, 2014)
  - Some of the content we might not want to retrieve is "fake news," which came to the fore in 2016 (Holan, 2016)
- Growing privacy concerns about tracking our searching (Huesch, 2013; Libert, 2015)

6

IR also a growing part of "knowledge discovery" from scientific literature



IR and online access firmly planted in health and biomedicine

- Biology is now defined as an "information science" (Insel, 2003)
- Pharmaceutical companies compete for informatics/library talent (Davies, 2006)
- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day (Bastian, 2010)
- Search for health information by clinicians, researchers, and patients/consumers is ubiquitous (Purcell, 2012; Google/Manhattan Research, 2012)
  - It's even part of "meaningful use" – text search over electronic health record notes (Metzger, 2012)

Use is ubiquitous among physicians (Google/Manhattan Research, 2012)

- Most have multiple devices – 99% with a desktop or laptop, 84% with a smartphone, and 54% with a tablet
- Spend twice as much time using online resources as print resources
- Even physicians aged 55+ heavy users – 80% own a smartphone, 84% use search engines daily, and 9 hours per week is spent online for professional purposes
- Search engine use a daily activity – 84%, with average of six searches done per day and 94% using Google
- When looking for clinical or treatment information, about a third click first on sponsored listings from a search
- About 93% say they take action based on searching – everything from pursuing more information to sharing with a patient or colleague to changing treatment decisions
- On smartphones, searching is preferred over mobile apps – 48% of use time with a search engine, 34% with mobile apps, and 18% going to specific Web sites in a browser or with a bookmark
- Spend about 6 hours per week watching online video, with about half of that time spent for professional purposes

OHSU

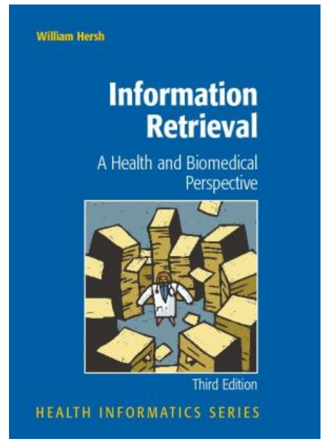
What kind of health information do consumers search for? (Fox, 2011)

Health topic	% searching
Specific disease or medical problem	66%
Certain medical treatment or procedure	56%
Doctors or other health professionals	44%
Hospitals or other medical facilities	36%
Health insurance – private or government	33%
Food safety or recalls	29%
Environmental health hazards	22%
Pregnancy and childbirth	19%
Medical test results	16%

OHSU

How to find more information about IR in health and biomedicine

- Hersh WR, *Information Retrieval: A Health and Biomedical Perspective*, Third Edition, 2009
  - Web site: www.irbook.info
- Chapters in other books, e.g., Shortliffe (2014), Sanchez-Mendiola (2014)
- Plenty of other books, journals, and other sources



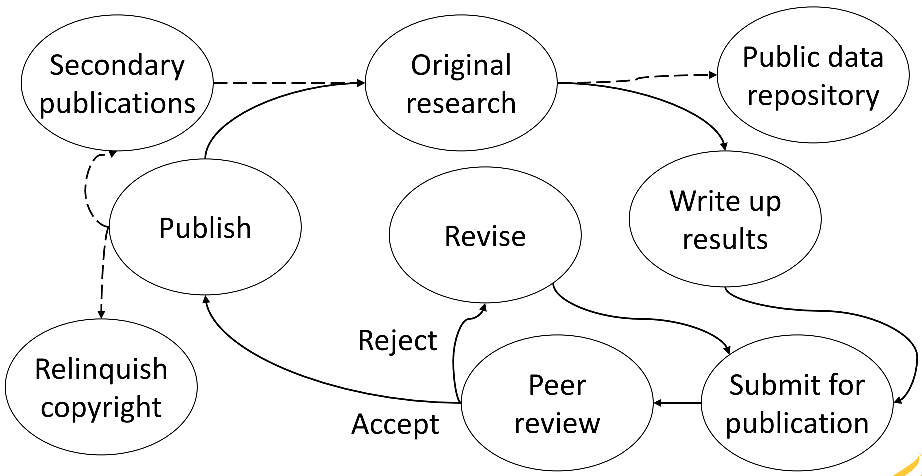
11

Why is IR pertinent to health and biomedicine?

- Growth of knowledge has long surpassed human memory capabilities
- Clinicians have frequent and unmet information needs
- Researchers must frequently update their knowledge in new areas quickly
- Primary literature on a given topic can be scattered and hard to synthesize
- Non-primary literature sources are often neither comprehensive nor systematic
- Web is increasingly used as source of health and biomedical information

12

Life-cycle of knowledge-based information



Classification of knowledge-based scientific information

- Primary – original research
  - Published mainly in journals but also in conference proceedings, technical reports, books, etc.
  - Can include re-analysis, e.g., meta-analysis and systematic reviews
- Secondary – reviews, condensations, and/or synopses of primary literature
  - Textbooks and handbooks are staples of clinical practitioners, researchers, and others
  - Guidelines are important for normalizing care and measuring quality

Classification of knowledge-based content

- Bibliographic
  - By definition rich in metadata
- Full-text
  - Everything on-line
- Annotated
  - Non-text or structured text annotated with text
- Aggregations
  - Bringing together all of the above
- These categories are admittedly fuzzy, and increasing numbers of resources have more than one type

15

Bibliographic content

- Bibliographic databases
  - The old (e.g., MEDLINE) have been revitalized with new features
  - New ones (e.g., National Guidelines Clearinghouse) have emerged
- Web catalogs
  - Share many characteristics of traditional bibliographic databases
- Real simple syndication/Rich site summary (RSS)
  - "Feeds" provide information about new content

16

Bibliographic databases

- Contain metadata about (mostly) journal articles and other resources typically found in libraries
- Produced by
  - U.S. government – most produced by National Library of Medicine (NLM, www.nlm.nih.gov)
    - e.g., MEDLINE, genomics information, etc.
  - Commercial publishers, e.g.,
    - EMBASE – part of larger SciVal
    - CINAHL – Cumulative Index to Nursing and Allied Health Literature
    - ACM Guide to Computing Literature – computer science and related areas

MEDLINE

- References to biomedical journal literature
  - Original medical IR application – system for searching MEDLINE launched in 1971 with literature maintained in MEDLARS system dating back to 1966
    - Name derives from MEDLARS On-Line – MEDLINE
  - Free to world since 1997 via PubMed – http://pubmed.gov
    - Now with links to full text of articles and other resources
- Statistics
  - http://www.nlm.nih.gov/bsd/bsd_key.html
  - Over 23M references to peer-reviewed literature
  - Over 5600 journals, mostly English language
  - Nearly 900,000 new references added yearly

National Guidelines Clearinghouse

- Produced by Agency for Healthcare Research and Quality (AHRQ)
  - www.guideline.gov
- Contains detailed information about guidelines
  - Including degree they are evidence-based
  - Interface allows comparison of elements in database for multiple guidelines
- Has links to those that are free on Web and links to producers when proprietary

19

Web catalogs

- Generally aim to provide quality-filtered Web sites aimed at specific audiences
  - Distinction between catalogs and sites blurry
- Some are aimed towards clinicians
  - HON Select – http://www.hon.ch/HONselect/
  - Translating Research into Practice – www.tripdatabase.com
- Others are aimed towards patients/consumers
  - Healthfinder – www.healthfinder.gov

20

RSS

- RSS “feeds” provide short summaries, typically of news, journal articles, or other recent postings on Web sites
- Users receive RSS feeds by an RSS aggregator that can typically be configured for the site(s) desired and to filter based on content
  - Work as standalone, in Web browsers, in email clients, etc.
- Two versions (1.0, 2.0) but basically provide
  - Title – name of item
  - Link – URL of full page
  - Description – brief description of page

Full-text content

- Contains complete text as well as tables, figures, images, etc.
- If there is corresponding print version, both are usually identical
- Includes
  - Periodicals
  - Books
  - Web sites – may include either of above

Full-text primary literature

- Almost all biomedical journals available electronically
  - Many published by Highwire Press (www.highwire.org), which adds value to content of original publisher, including *British Medical Journal*, *Journal of the American Medical Association*, *New England Journal of Medicine*, etc.
  - Also published by leading commercial scientific publishers, e.g., Elsevier, Kluwer, Springer, etc.
  - Growing number available via open-access model, e.g., Biomed Central (BMC), Public Library of Science (PLoS)
  - Another source of full-text papers is PubMed Central (PMC; http://pubmedcentral.gov)

Books

- Textbooks
  - Most well-known clinical textbooks are now available electronically
    - e.g., *Harrison's Principles of Internal Medicine*
  - Most are bundled into large collections by publishers
    - e.g., Access Medicine (McGraw-Hill), Elsevier, Kluwer
  - NLM has developed books site as part of Entrez
    - http://www.ncbi.nlm.nih.gov/books
- Compendia of drugs, diseases, evidence, etc.
- Handbooks – very popular with clinicians
- Increasingly published on mobile devices

Value added for electronic books

- Multimedia, e.g., skin lesions, shuffling gait of Parkinson's Disease, etc.
- Bundling of multiple books
- Can be updated in between "editions"
- Linkage to other information, e.g., to references, self-assessments, updates, other resources, etc.

25

Web sites

- Defined more narrowly here to refer to coherent collections of information on Web
- Usually take advantage of Web features, such as linking, multimedia
- Increasingly integrated with other resources and available on different platforms (e.g., integrated into electronic health records [EHRs], on smartphones, etc.)

26

Some notable full-text content on Web sites

- Government agencies
  - National Cancer Institute
    - www.cancer.gov
  - Centers for Disease Control – travel and infection information
    - http://www.cdc.gov/DiseasesConditions
    - http://www.cdc.gov/travel/
  - Other NIH institutes, e.g., National Heart, Lung, and Blood Institute (NHLBI)
    - www.nhlbi.nih.gov

Full-text Web sites (cont.)

- Physician-oriented medical news and overviews, e.g.,
  - Medscape – www.medscape.com
  - Many professional societies provide to members, e.g., http://www.acponline.org/clinical_information/
- Patient/consumer-oriented, e.g.,
  - NetWellness – www.netwellness.com
  - WebMD – www.webmd.com
- Many mobile apps provide health information, e.g.,
  - iTriage – www.itriagehealth.com
  - Epocrates – www.epocrates.com

Other interesting types of Web content

- Wikipedia – www.wikipedia.org
  - Encyclopedia with free access and distributed authorship
  - Some concerns about manipulation (McHenry, 2004) but
    - Comparable to *Encyclopedia Britannica*? (Giles, 2005 – rebuttal: Anonymous, 2006)
    - Health information quality is reasonably good (Nicholson, 2006)
    - Content retrieved prominently in most Web searches (Laurent, 2009)
    - Making attempt to improve quality of medical content (Heilman, 2013)
- Body of knowledge
  - Software Engineering Body of Knowledge (SWEBOK, www.swebok.org) organizes knowledge of field
- Social media/Web 2.0 and beyond (Lee, 2011)

29

OHSU

Annotated

- Non-text or structured text annotated with text
- Includes
  - Image collections
  - Citation databases
  - Evidence-based medicine databases
  - Clinical decision support
  - Genomics databases
  - Other databases

30

OHSU

Image collections

- Most prominent in the "visual" medical specialties, such as radiology, pathology, and dermatology
- Well-known collections include
  - Visible Human – http://www.nlm.nih.gov/research/visible/visible_human.html
  - Lieberman's eRadiology – http://eradiology.bidmc.harvard.edu
  - WebPath – http://library.med.utah.edu/WebPath/webpath.html
  - More pathology – PEIR, www.peir.net
  - DermIS – www.dermis.net
  - More dermatology, also a decision-support system – www.visualdx.com
- Many have associated text, which assists with indexing and retrieval

Citation databases

- *Science Citation Index* and *Social Science Citation Index*
  - Database of journal articles that have been cited by other journal articles
  - Now part of a package called *Web of Science*, which itself is part of a larger product, *Web of Knowledge* (Clarivate)
    - http://clarivate.com/scientific-and-academic-research/research-discovery/web-of-science/
- SCOPUS – http://www.elsevier.com/online-tools/scopus
- Google Scholar – http://scholar.google.com

Evidence-based medicine databases

- Cochrane Database of Systematic Reviews – http://www.cochrane.org
  - Collection of systematic reviews, kept updated
- Evidence "formularies"
  - Clinical Evidence (BMJ) – http://clinicalevidence.bmj.com/x/index.html
  - JAMAevidence – http://jamaevidence.com
- PubMed Health – https://www.ncbi.nlm.nih.gov/pubmedhealth/
  - Systematic reviews and summaries of systematic reviews
- Many resources part of aggregations

Clinical decision support (CDS)

- Content used in CDS systems, usually part of EHRs
  - Order sets (usually "evidence-based")
  - CDS rules
  - Health/disease management templates
- Growing and evolving commercial market for such tools, especially as EHR adoption increases; leaders include
  - Zynx – www.zynxhealth.com
  - Thomson Reuters Cortellis – http://cortellis.thomsonreuters.com
  - EHR vendors themselves and partners

Genomics databases

- National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov; NCBI, 2017) collection links
  - Literature references – MEDLINE
  - Textbook of genetic diseases – On-Line Mendelian Inheritance in Man
  - Sequence databases – Genbank
  - Structure databases – Molecular Modeling Database
  - Genomes – Catalog of genes
  - Maps – Locations of genes on chromosomes

OHSU

Other databases

- ClinicalTrials.gov
  - www.clinicaltrials.gov
  - Originally database of clinical trials funded by NIH
  - Now used as register for clinical trials, with results reporting for some (DeAngelis, 2005; Laine, 2007; Zarin, 2013; Zarin, 2015)
- NIH RePORTER
  - http://projectreporter.nih.gov/reporter.cfm
  - Database of all research grants funded by NIH
  - Replaced the CRISP database

OHSU

Data publishing

- Internet makes it technologically feasible
- Many fields have long tradition of requiring depositing of data in public repository as a condition to publish, e.g., genomics, although availability incomplete (Alsheikh-Ali, 2011)
- Growing advocacy for clinical trials data
  - A "public good" (Rodwin, 2012) for new era of "open science" (Ross, 2013)
  - Calls for doing so by journal editors (Taichman, 2016) and others (Ross, 2013; Mello, 2013)
  - Pushback from trialists who want time-limited protection of those who generate data for rewards of their work and from those who aim to discredit or undermine original research (Anonymous, 2016)
- biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE)
  - Database of metadata about available biomedical data sets
  - https://datamed.org/

Aggregations – integrating many resources

- Clinical – growing tendency of publishers to aggregate resources into comprehensive products
  - Merck Medicus – www.merckmedicus.com
    - Collection of many resources available to any licensed US physician
  - Up to Date – www.uptodate.com
    - Very popular among clinicians
  - Essential Evidence Plus (includes InfoPOEMS, "Patient-oriented evidence that matters") – www.essentialevidenceplus.com
  - Dynamed – www.dynamed.com

Other aggregations

- Biomedical research: Model organism databases, e.g., Mouse Genome Informatics
  - www.informatics.jax.org
  - Combines genomics and related data, bibliographic database, gene references, etc.
- Consumer: MEDLINEplus
  - http://medlineplus.gov
  - Integrates a variety of licensed resources and public Web sites

Indexing

- Assignment of metadata to content to facilitate retrieval
- Two major types
  - Human indexing with controlled vocabulary
  - Automated indexing of all words
- Also address
  - Indexing other "objects"
  - UMLS Metathesaurus
  - Web indexing

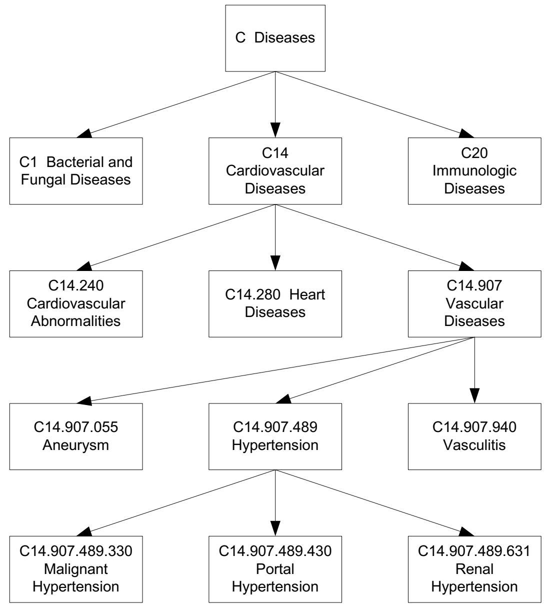
Human indexing

- Usually performed by professional indexer with some background in biomedicine
- Follows protocol to scan resource and select terms from a controlled vocabulary
- Most vocabularies are hierarchical and have specific definitions for when term is to be assigned

Medical Subject Headings (MeSH) vocabulary (Colletti, 2001)

- Over 26,000 terms, with many synonyms for those terms
  - Over 230,000 Supplementary Concept Records, formerly mostly chemicals and drugs, now rare diseases and genes
- Hierarchical, based on 16 trees, e.g., *Anatomy*, *Diseases*, *Chemicals and Drugs*
- Contains 83 subheadings, which can be used to make a heading more specific, such as *Diagnosis* or *Therapy*
- MeSH browser allows exploration
  - http://www.nlm.nih.gov/mesh/MBrowser.html

A slice of MeSH



43

MEDLINE indexing

- Indexing done by professionals who follow protocol first devised by Bachrach (1978)
  - Read title, introduction, and conclusion and then scan methods, results, figures, tables, and, lastly, abstract
  - Ignore "key words" of publisher
  - Assign 2-4 headings (with or without subheadings) as central concepts (or major headings) and another 5-10 as minor headings
  - Use most specific headings in hierarchy assigned
- Important additional tag is Publication Types
  - e.g., Randomized Controlled Trial, Meta-Analysis, Practice Guideline, Review
- Many modern tools have been developed to assist indexing, such as term suggestion and look-up

44

Other bibliographic indexing

- Other NLM databases use MeSH
- Some non-NLM resources use MeSH
  - MeSH freely available from NLM at http://www.nlm.nih.gov/mesh/filelist.html
- Other non-NLM databases have their own subject headings, e.g.,
  - CINAHL subject headings
  - EMTREE

45

Other metadata

- Indexing covers more than content
- Other attributes of documents to index can include
  - Author(s)
  - Source: journal name, issue, pages
  - Publication or resource type
  - Relationship to other information
    - e.g., gene identifier, grant number, etc.

46

Automated indexing

- Indexing of all words that occur in content items
  - In bibliographic databases, will usually include title, abstract, and often other fields, e.g., author or subject heading
  - In full-text documents, will usually include all text, including title
- Often use a stop word list to remove common words (e.g., *the*, *and*, *which*)
- Some systems "stem" words to root form (e.g., *coughs* or *coughing* to *cough*)

Weighted indexing (Salton, 1991)

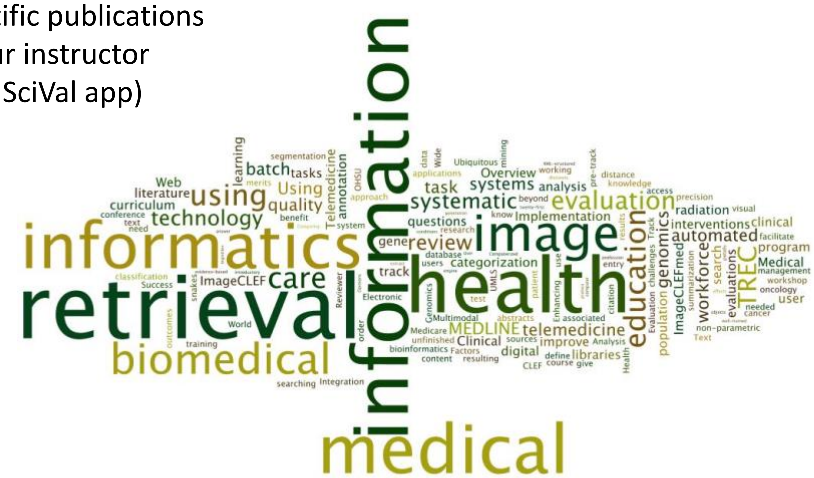
- Usually used with automated indexing
- Gives weight to words that are frequent but discriminating
- Most common approach is for weight to equal product TF*IDF
  - Inverse document frequency of word i
    - IDF_i = log(# documents/# documents with word)+1
  - Term frequency of word i in document j
    - TF_{ij} = frequency of word in document

Weighted indexing examples

- From a database on AIDS
  - The word AIDS will likely occur in almost every document, while *retinopathy* will be much more "discriminating"
- In a general medical database
  - AIDS will occur much less frequently, so is better indexing term

"Visual" indexing – e.g., Wordle, www.wordle.net

Scientific publications of your instructor (from SciVal app)



Citation indexing

- Other content items that "cite" this one, e.g., references, links, etc.
- Indexing is at content item level
- Goal is to designate related or important content items
- Citation databases list all other articles that cite a specific article in journals
  - e.g., *Science Citation Index*, SCOPUS, and *Google Scholar*
- Novel feature of Google search engine (Brin, 1998) was giving higher weight to Web pages that have more links to them

Limitations of human indexing

- Inconsistency
  - When MEDLINE records indexed in duplicate, consistency varies from 63% for central concept headings to 36% for heading-subheading combination (Funk, 1983)
  - Results verified even with modern indexing tools and methods (Marcetich, 2004)
- Inadequate indexing vocabulary
  - Up to 25% of all concepts not represented in MeSH (Hersh, 1994)
  - Ambiguities and other naming problems with genes, proteins, etc. (Yandell, 2002; Tuason, 2004)

Limitations of word indexing

- Synonymy – e.g., *cancer/carcinoma*
- Polysemy – e.g., *lead*
- Context – e.g., *high blood pressure*
- Focus – e.g., central vs. incidental concepts
- Granularity – e.g., antibiotics vs. specific ones

53

Research

- Evaluation
  – How valuable are systems to users?
  – How well do systems and users perform?
- Future directions
  – Applying IR techniques to electronic health records
  – Beyond retrieval – question-answering

54

Evaluation

- Questions often asked
  - Is system used?
  - Are users satisfied?
  - Do they find relevant information?
  - Do they complete their desired task?
- Most studied group is physicians, with systematic reviews of results (Hersh, 1998, Pluye, 2005)
- Most IR evaluation research has focused on retrieval of relevant documents, which may not capture full spectrum of usage
  - Often consists of challenge evaluations that develop "test collections" – best known is (non-medical) Text Retrieval Conference (TREC, http://trec.nist.gov) (Voorhees, 2005)

55

OHSU

Is system used?

- Most studies done prior to ubiquitous Internet, electronic health records, mobile devices, etc.
- Studies in various clinical settings (Hersh, 2009; Magrabi, 2005) showed average use varied from 0.3 to 8.7 accesses per person-month
- Whatever the actual number, this paled in comparison to known physician information needs (Gorman, 1995) of two questions per every three patients

56

OHSU

Are users satisfied?

- Most studies report good user satisfaction, but some interesting studies to note
  - Nielsen (1994) meta-analysis found association (though imperfect) between user satisfaction and ability to use computer systems
  - Most Internet users believe they mostly find information they are seeking (Taylor, 2010; Fox, 2011)

Do they find relevant information?

- Most common approach to evaluation
- Usually measured by relevance-based measures of recall and precision
  - Recall (R)

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

  - Precision (P)

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

- And various aggregations, e.g., F, MAP, NDCG, etc.

Comments about recall and precision

- There tends to be a trade-off between the two
- "Relevance" can be an ambiguous notion (Hersh, 1994)
- It is unclear whether they correlate with a user's success in using an IR system
- The proliferation of standard test collections leads to a great deal of research that excludes real users

59

How well do clinicians search? Early results from Haynes (1990)

Searcher Type	Recall	Precision
Novice clinicians	27%	38%
Expert clinicians	48%	48%
Librarians	49%	57%

Other findings

- Little overlap among retrieval sets
  - Searchers tended to find similar quantities of disparate relevant documents
- Novice searchers satisfied with results
  - Adequate information or ignorant bliss?

60

Extending evaluation beyond physicians and documents

- Other clinicians
  - Nurses – Rolye, 1995
  - Pharmacists – Wanke, 1988
  - Nurse practitioners – Hersh, 2000; Hersh, 2002
- Biomedical researchers
  - Very little study of their use of IR systems
  - Investigated by TREC Genomics Track (Hersh, 2006; Hersh, 2009) – http://ir.ohsu.edu/genomics/
- Image retrieval – ImageCLEFmed (Hersh, 2006; Hersh, 2009; Kalpathy-Cramer, 2015)
  - Retrieval performance related to query type, measure selection
  - http://ir.ohsu.edu/image/

OHSU

Recall and precision studies yield useful results, but

- Are searchers able to solve their information problems by using system?
  - Some results research have used "task-oriented approach" to measure question-answering
  - Hersh (2002) – use of MEDLINE to answer clinical questions
    - Medical students answered 34% of questions before system, 51% afterwards
    - Nurse practitioner students answered 34% of questions before system but did not change with system
    - Time to answer a question was ~30 minutes
    - No association of recall or precision with correct answering

OHSU

Another task-oriented study

- Westbrook (2005) – use of online evidence system
  - Physicians answered 37% of questions before system, 50% afterwards
  - Nurse specialists answered 18% of questions before system, 50% afterwards
  - Those who had correct answers had higher confidence in their answers, but those not knowing answer initially had no difference in confidence whether answer right or wrong

How do IR systems impact physician practice? (Pluye, 2004)

- Qualitative study found four themes mentioned by physicians
  - Recall – of forgotten knowledge
  - Learning – new knowledge
  - Confirmation – of existing knowledge
  - Frustration – that system use not successful
- Researchers also noted two additional themes
  - Reassurance – that system is available
  - Practice improvement – of patient-physician relationship

Challenges for IR evaluation moving forward

- Must understand tasks of user and focus evaluation accordingly
- Ultimate measure, like any other informatics application, might be health outcome
  - This may be difficult with IR systems since usage may not directly impact outcomes of patient care or research activity

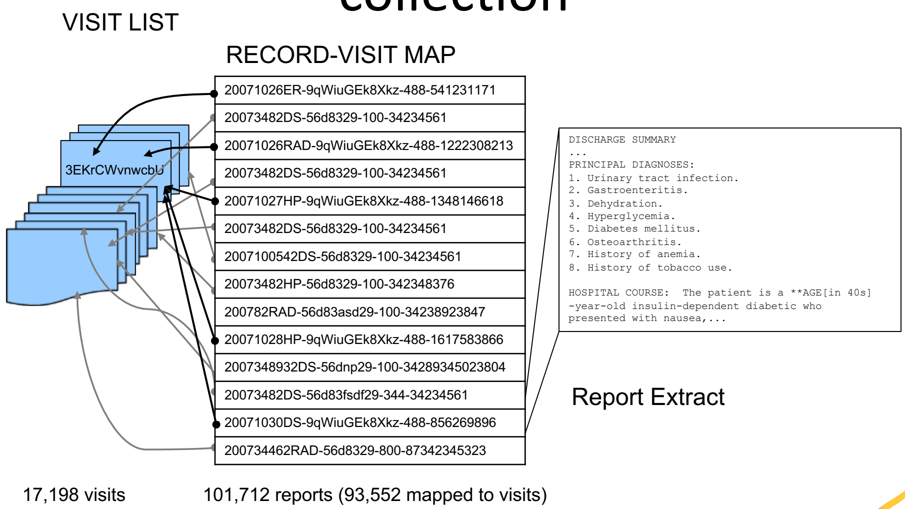
65

Research directions – applying IR to medical records

- Most medical records still in narrative documents, where natural language processing (NLP) techniques are improving but still imperfect (Stanfill, 2010)
- For some tasks, can we take an IR approach?
  - TREC Medical Records Track used de-identified corpus of medical records in initial task of identifying patients as candidates for clinical research studies (Voorhees, 2011; Voorhees, 2012)

66

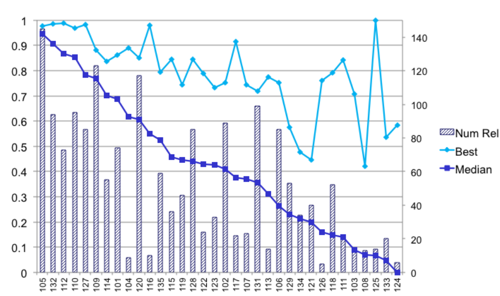
TREC Medical Records Track test collection



(Courtesy, Ellen Voorhees, NIST)

TREC Medical Records Track results

- Highly variable across different topics
  - Easiest – consistently best results
    - 105: Patients with dementia
  - Hardest – consistently worst results
    - 108: Patients treated for vascular claudication surgically
  - Large differences between best and worst results
    - 125: Patients co-infected with Hepatitis C and HIV
- Overall results show substantial room for improvement
  - Best results involve manual modification of queries



(Voorhees, 2011; Voorhees, 2012)

Subsequent work in medical records search

- Public test collections of medical records stymied by privacy concerns
- Funded for project using parallel corpora with common topics at OHSU and Mayo Clinic (Wu, 2017)
- Exploring options for Evaluation as a Service (EaaS) to allow others to use data without seeing it (Hanbury, 2015)
  - Similar situation to TREC Total Recall Track searching over email and corporate repositories (Roegeist, 2016)

69

More recent TREC tracks

- Clinical Decision Support, 2014-2016 (Roberts, 2016)
  - Given patient case, find relevant full-text articles (from PMC snapshot) about diagnosis, tests, or treatments
- Precision Medicine (2017)

70

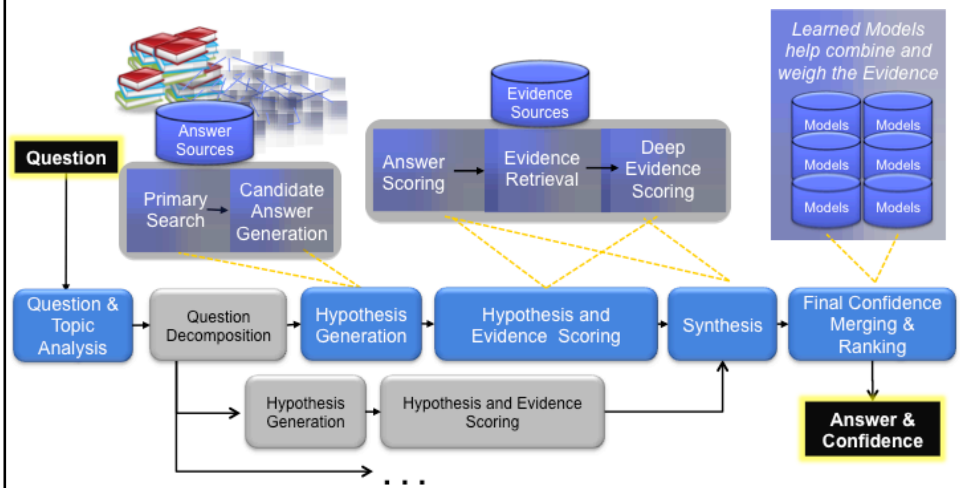
Research directions – question-answering

- Users may retrieve documents, but usually want answers to questions
- Subarea of IR research has focused on question-answering systems (Strzalkowski, 2006)
- Highest-profile system is IBM Watson
  - Developed out of TREC Question-Answering Track (Voorhees, 2005; Ferrucci, 2010)
  - Additional (exhaustive) details in special issue of *IBM Journal of Research and Development* (Ferrucci, 2012)
  - Beat humans at Jeopardy! (Markoff, 2011)
  - Now being applied to healthcare (Lohr, 2012); has "graduated" medical school (Cerrato, 2012)

How does Watson work (Ferrucci, 2010)?

- Built around a system called DeepQA, which uses massively parallel computing to acquire knowledge from resources of a given domain
- Learning process builds around sample questions from the domain
  - A key step is to identify lexical answer types (LATs) in the domain
  - Among general questions, some common LATs include `he`, `country`, `city`, `man`, `film`, `state`, `she`, `author`, `group`, `here`, `company`, etc.
  - NLP then applied to text and knowledge representation and reasoning (KRR) applied to structured knowledge
  - Machine learning then applied to questions and their answers

Watson architecture (Ferrucci, 2010)



73

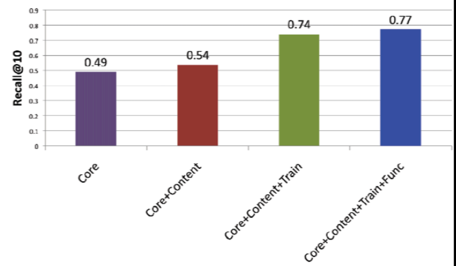
Applying Watson to medicine (Ferrucci, 2012)

- Trained using several resources from internal medicine: *ACP Medicine*, *PIER*, *Merck Manual*, and *MKSAP*
- Concept adaptation process required
  - Named entity detection – e.g., disambiguation of terms and their senses
  - Measure recognition and interpretation – e.g., age or blood test value
  - Recognition of unary relations – e.g., elevated <test result>
- Trained with 5000 questions from *Doctor's Dilemma*, a competition like Jeopardy!, in which medical trainees participate and is run by the ACP each year
  - Sample question is, `Familial adenomatous polyposis is caused by mutations of this gene`, with the answer being, `APC Gene`
    - Googling the question gives the correct answer at the top of its ranking to this and two other sample questions listed

74

Evaluation of Watson on internal medicine questions (Ferrucci, 2012)

- Evaluated on an additional 188 unseen questions
- Primary outcome measure was recall at 10 answers
  - How would Watson compare against other systems, such as Google or Pubmed, or using other measures, such as MRR?
- Future use case for Watson is applying system to data in EHR, ultimately aiming to serve as a clinical decision support system (Cerrato, 2012)
- Not much peer reviewed literature since then…



75