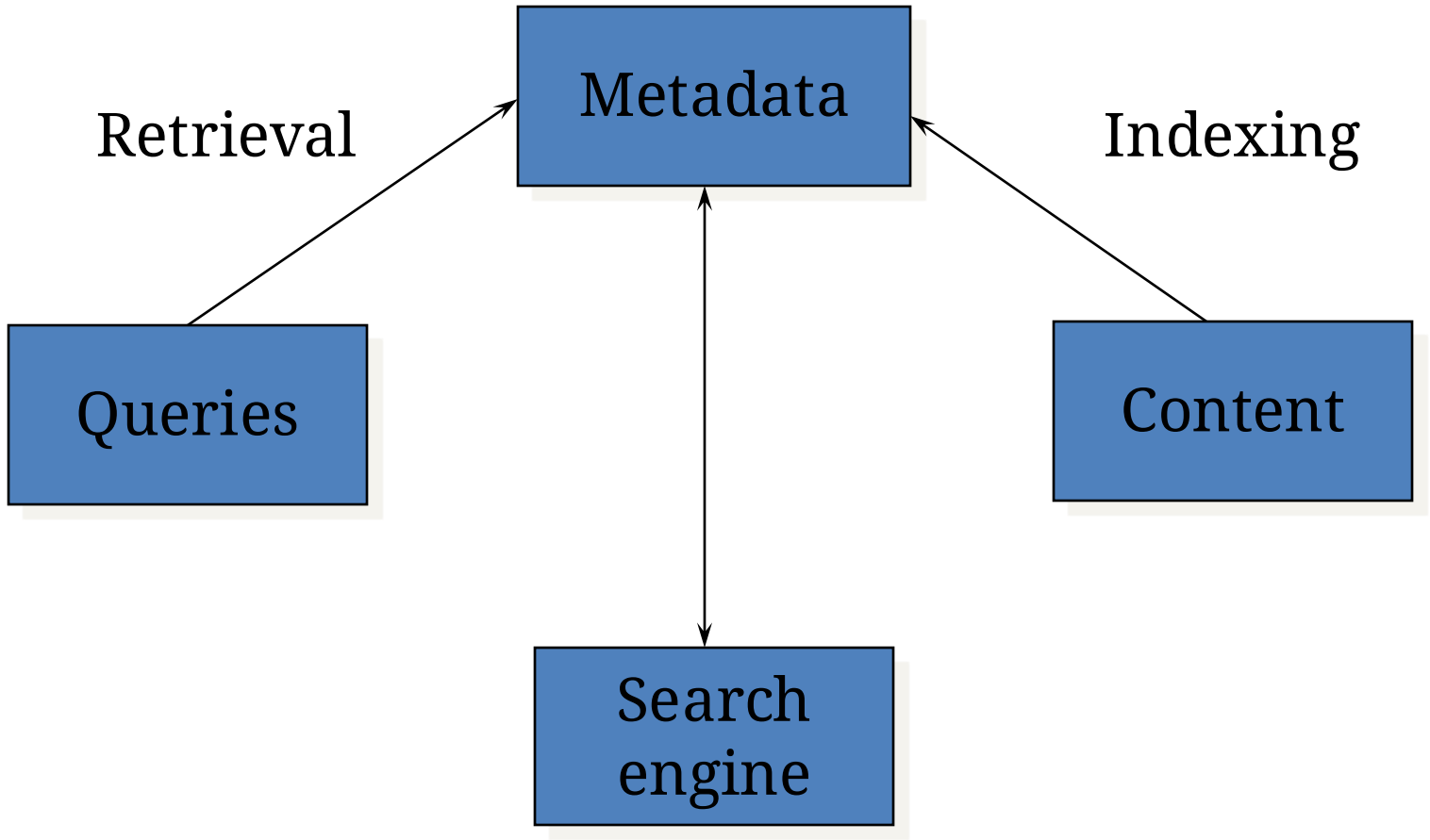
4.1 Information Retrieval Process, Information, and Challenges

What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh, MD, FACMI, FAMIA, FIAHSI

Information retrieval (IR)

- Field concerned with organization and retrieval of predominantly text-based information
  - But multimedia (e.g., images, sounds, video, etc.) and more complex databases are increasingly a part
- When I began work in this area (circa 1989), few physicians or scientists and virtually no patients had done an on-line search
  - Now everyone is searching – from Web browser to mobile device

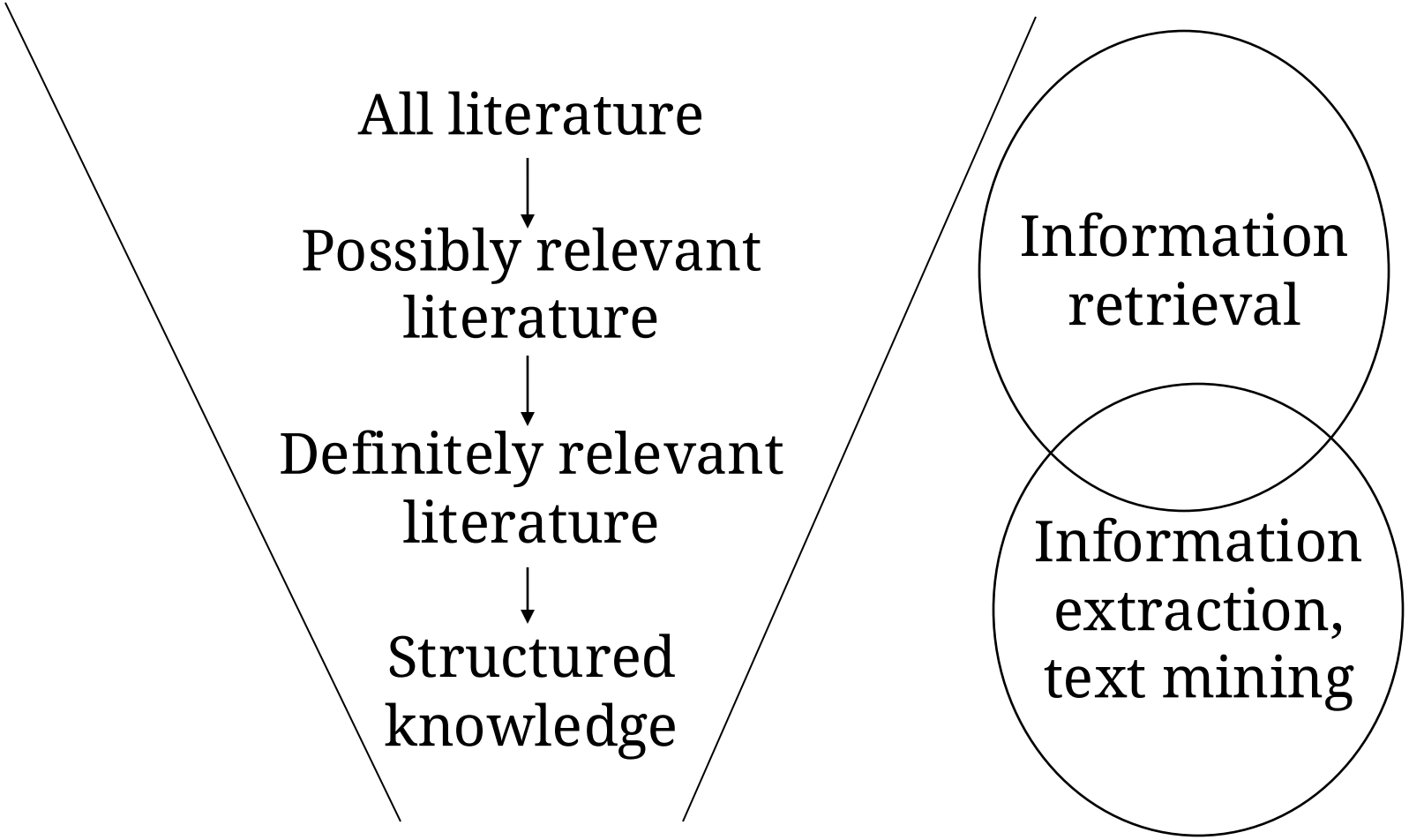
IR process


Retrieval
Metadata
Indexing
Queries
Content
Search
engine

Intellectual tasks of IR

- Indexing
  - Assigning metadata to content items
  - Can assign
    - Subjects (terms) – words, terms from controlled vocabulary
    - Attributes – e.g., author, source, publication type
- Retrieval
  - Most common approaches are
    - Boolean – use of AND, OR, NOT
    - Natural language – words common to query and content

IR also growing part of knowledge discovery from scientific literature

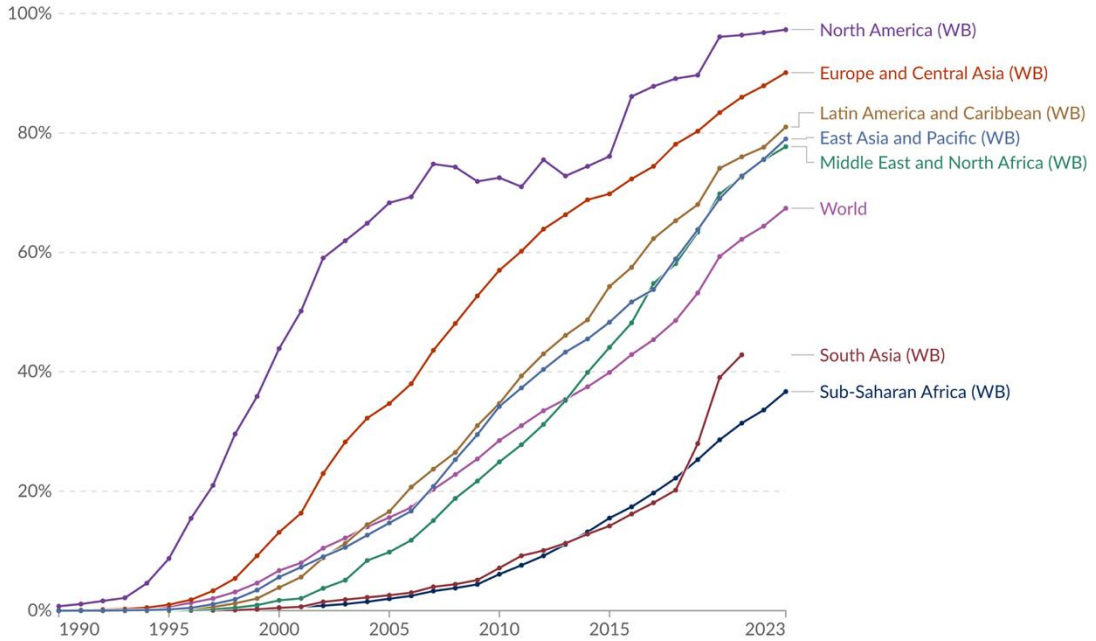


Major challenges in IR

- 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
- Determining quality and veracity of information

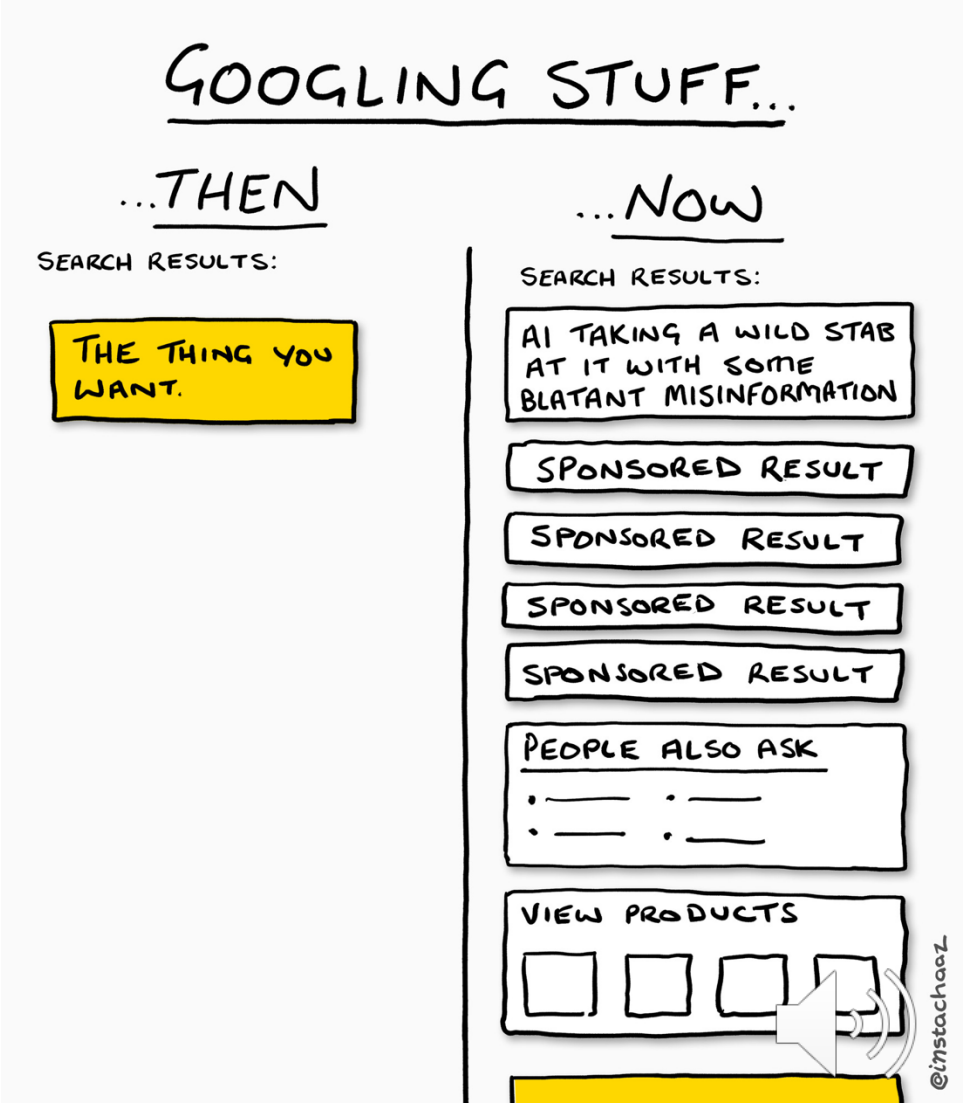
IR is now mainstream

- Internet (and likely search engine) use now ubiquitous
  - Not only in developed countries but across world
  - At least 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
  - Some are lucky, e.g., last name of "Hersh"



Web has changed 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 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)
- Hard to discern sponsored vs. organic results
- Privacy concerns about tracking our searching (Huesch, 2013; Libert, 2015)



IR and online access firmly planted in biomedicine and health

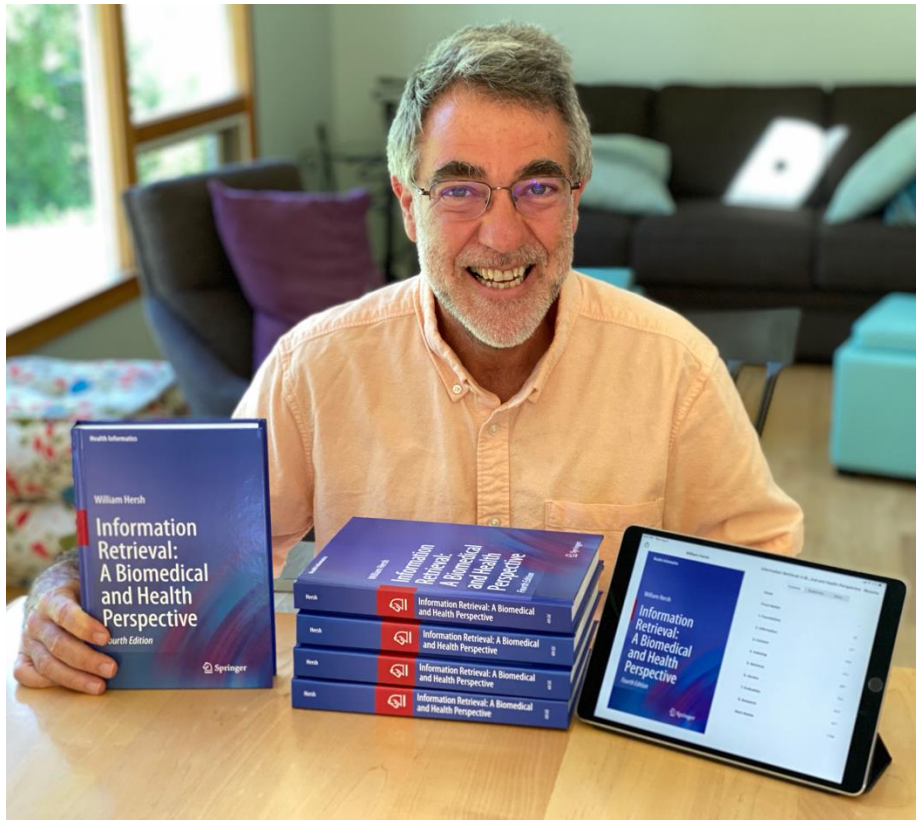
- Biology should be defined as an “information science” (Insel, 2003)
- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day (Bastian, 2010)
- Search for health information by clinicians, researchers, and patients/consumers is ubiquitous (Fox, 2011; Fox, 2013; Google/Manhattan Research, 2012)

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%

How to find more information about IR in biomedicine and health

- From me!
- Hersh WR, Information Retrieval: A Biomedical and Health Perspective, 4th Edition, 2020
- Chapters in other books
  - Informatics – Sanchez-Mendiola (2014), Shortliffe (2021), Hersh (2022)
  - IR – Alonso (2024), White (2025)
- Plenty of other books, journals, and other sources



Why is IR pertinent to biomedicine and health?

- 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 can be source of biomedical and health information (and misinformation)
- Growing but uncertain role for generative AI and large language models (LLMs) for managing and accessing medical knowledge

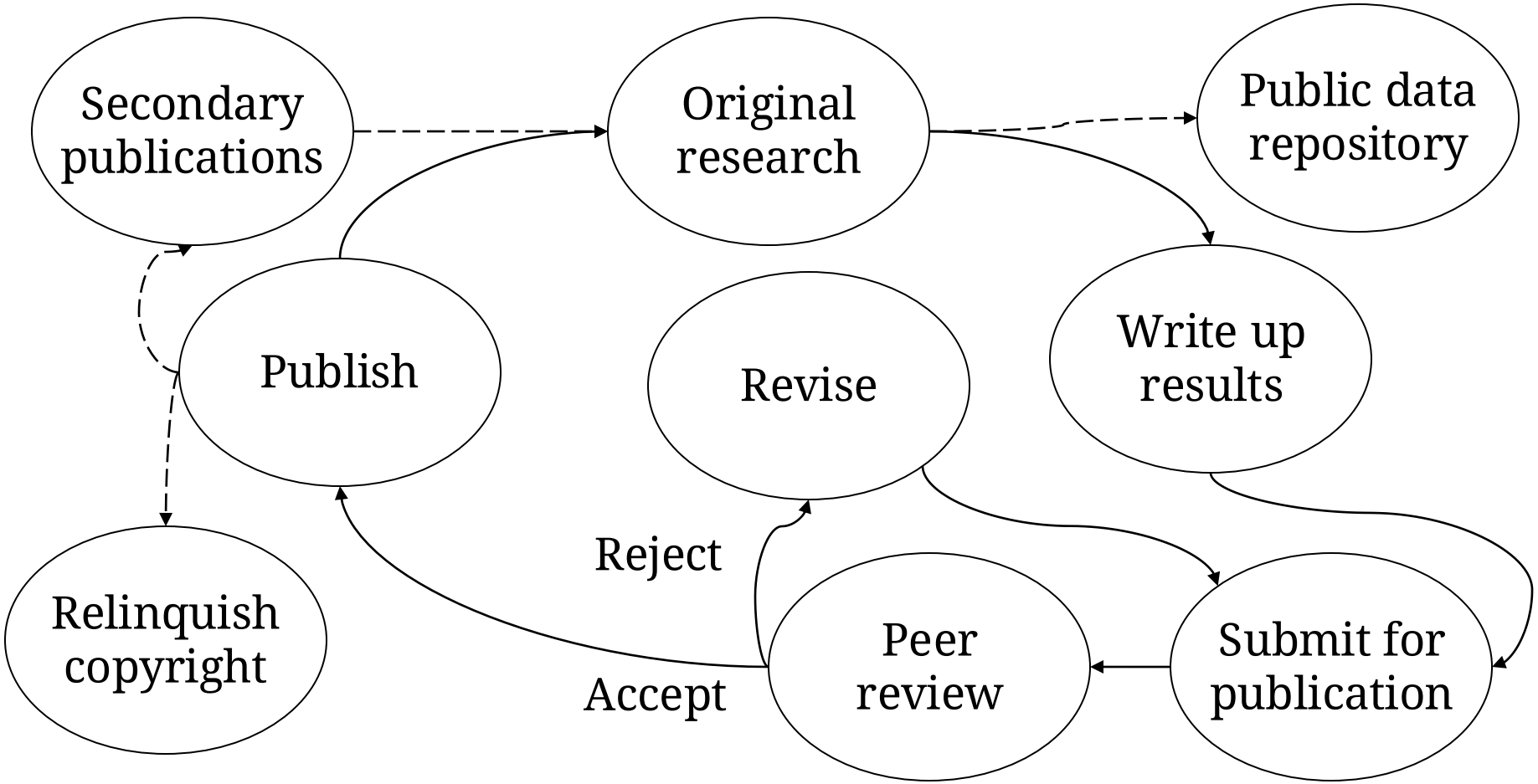
Information accessed in IR

- Knowledge-based information
- Life cycle of knowledge-based information
- Electronic publishing

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., systematic reviews and meta-analysis
- 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

Historical life-cycle of knowledge-based information



Changes to life-cycle: data publishing

- Internet makes feasible – long history in some fields, e.g., genomics
- Growing advocacy for clinical data
  - A “public good” (Rodwin, 2012) for new era of “open science” (Ross, 2013; National Academies, 2018)
  - One-third of re-analyses led to different conclusions (Ebrahim, 2014)
  - Calls by journal editors (Taichman, 2017)
  - All research funded by National Institutes of Health (NIH) must have data-sharing policy (Kozlov, 2022; Ross, 2023)
- Concerns
  - Allowing primary publication (Merson, 2016)
  - Worth the cost (Strom, 2016)?
  - In prominent medical journals, JAMA, Lancet, and New England Journal of Medicine, gap between papers that purport availability and have actual availability (Danchev, 2020)
  - NIH policy difficult for resource-strapped researchers (Bauchner, 2023)?

Changes to life-cycle: preprints

- Long-standing use in physics, computer science, and other fields (Chiarelli, 2019; Puebla, 2021)
  - arXiv (Garisto, 2022)
- Biomedicine mostly historically bound by “Inglefinger Rule” (1969) of no prior publication
- But now growing use in
  - Biology
  - Medicine
- Concerns about subverting scientific process (Maslove, 2018)
  - Most final publication results congruent with preprint findings (Janda, 2022)
  - Most COVID-related submissions published, although with higher rate of retraction (Kodvanj, 2022)
  - Some retraction designations do not propagate back to preprints (Avissar-Whiting, 2022)
- Alternative: registered reports with peer review of methods before results (Chambers, 2019; Brock, 2020)?

Changes to life-cycle: electronic publishing

- Virtually all scientific publishing now electronic
- Challenges to access mostly economic and not technical (Hersh 2000; Sox, 2009)
  - Prestigious journals have monopolies due to academic promotion and tenure concerns
  - Unlike paper-based content, digital content usually owned by publisher and licensed to user
  - Who will pay cost of value journals add via peer review, editing, etc.?
- Is scientific paper obsolete?
  - Use computational notebooks with access to data instead (Ritchie, 2022)?