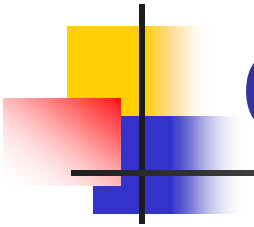
Information Retrieval: Grand Challenges in the 21st Century

William Hersh, M.D.
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
hersh@ohsu.edu
www.billhersh.info

Greetings from “blue” Portland

Portland, Canada



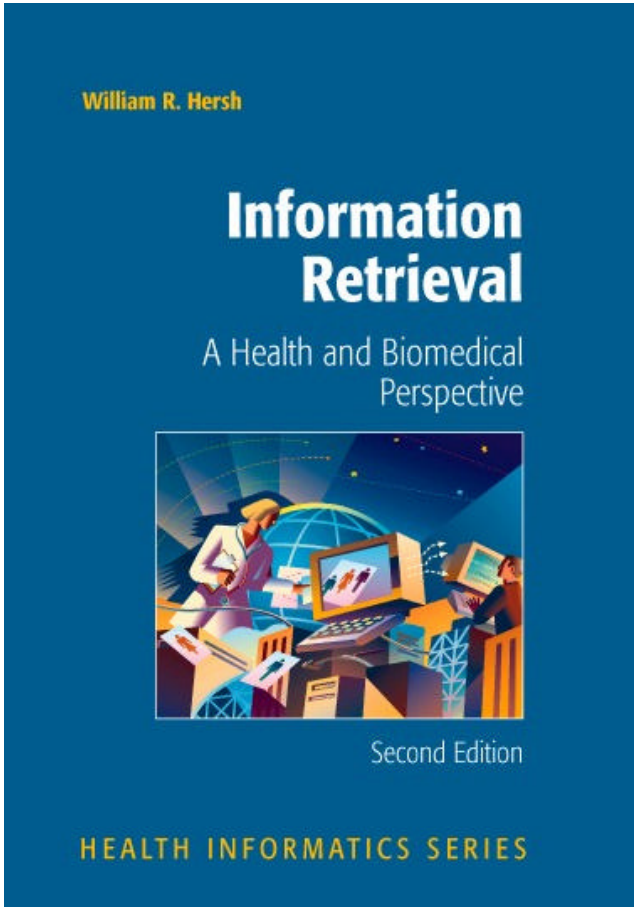
Overview

- Primer on information retrieval
- Non-grand challenges
- Grand challenges
- Final thoughts

Primer on information retrieval

Information retrieval – Hersh, 2003

- Focuses on indexing and retrieval of knowledge-based information
- Historically centered on text in documents, but increasingly associated with multimedia and even patient-specific information
- www.irbook.info

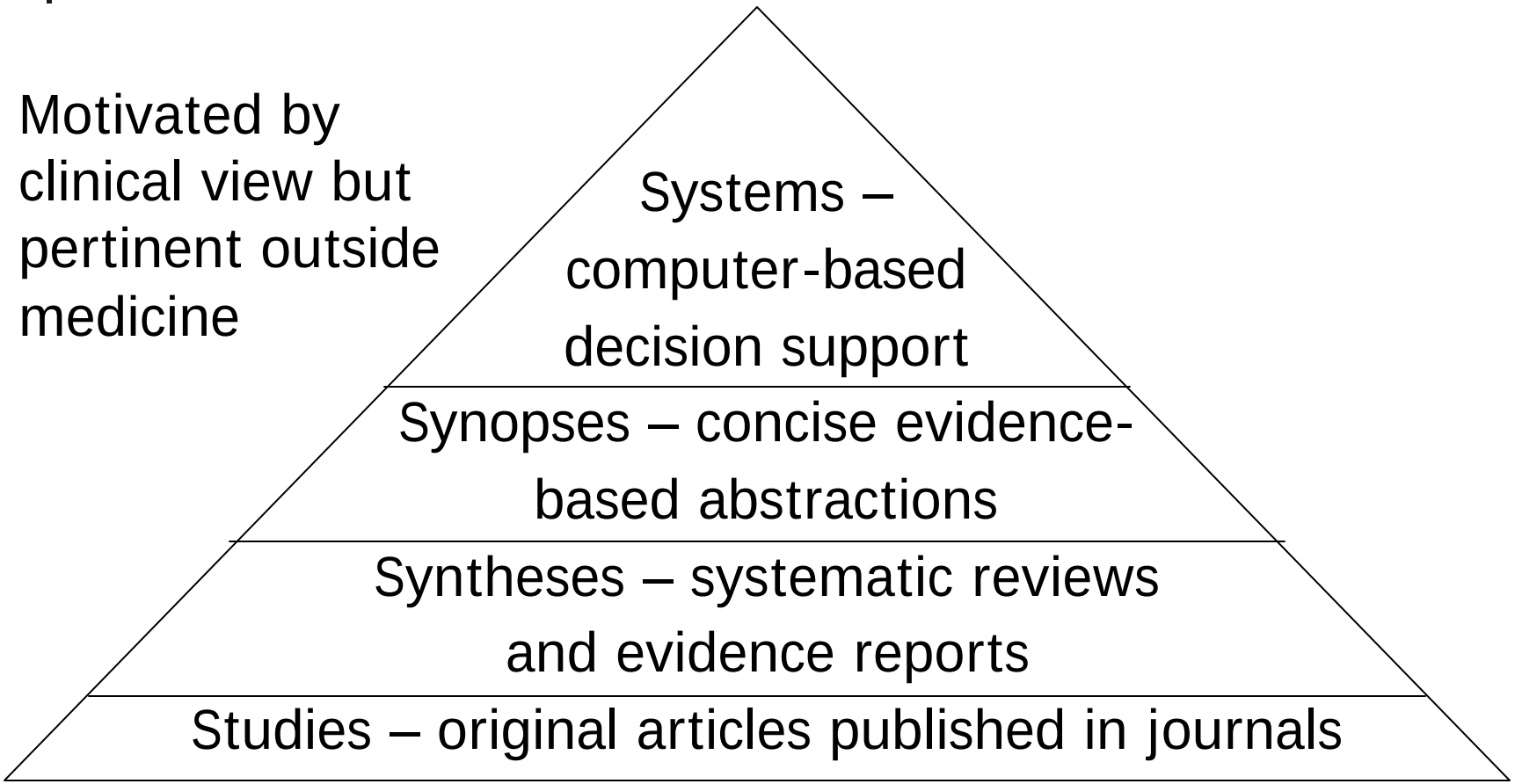


Overview of information from the clinical perspective

- Two basic types, with different uses and applications
  - *Patient-specific* information is generated in the care of patients
    - Applications: electronic health records, telemedicine, etc.
  - *Knowledge-based* information is the scientific literature of health care
    - Applications: information retrieval systems, evidence-based medicine

Hierarchy of information (Haynes, 2001)

Motivated by clinical view but pertinent outside medicine



Searching – everyone is doing it ...

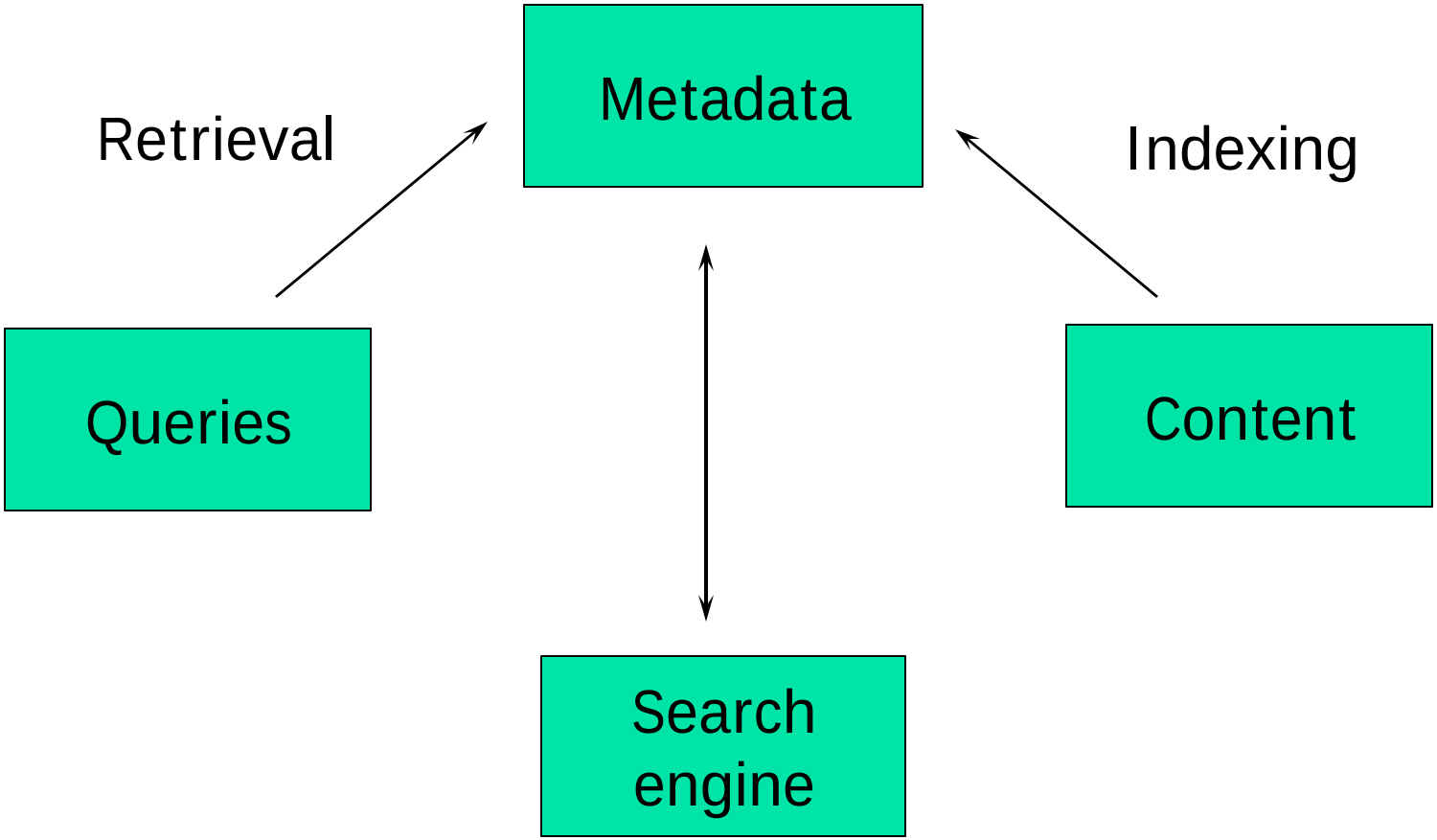
"First, they do an on-line search."

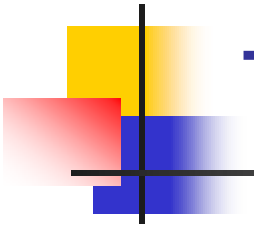
... everyone knows about it ...

(Am I a lucky father or what?)

... but new problems have emerged

IR system


Metadata
Retrieval
Indexing
Queries
Content
Search engine

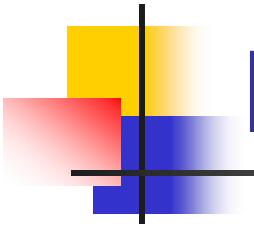


The intellectual tasks of IR

- Indexing
  - Assigning metadata to content items
  - Can assign
    - Terms – words, phrases 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, with output ranked by word or link frequency

Evaluation of information retrieval systems

- System-oriented
  - Historically focused on relevance-based measures
    - Recall – proportion of all relevant articles retrieved
    - Precision – proportion of retrieved articles that are relevant
  - When documents ranked, can combine both in mean average precision (MAP)
    - Average of precision at points of recall
- User-oriented
  - User satisfaction
  - Ability to complete tasks in laboratory setting
  - Outcomes of use in real-world setting



Non-grand challenges

- While occasionally challenging, these problems are, for the most part, not grand challenges
  - Using an IR system to “find some information” (e.g., simple search of local Web site, textbook, maybe even MEDLINE)
  - Finding a known item (Google Toolbar anyone?)
- These do not obviate need for users better learning how search systems work
  - Continued role for research by librarians, informaticians, computer scientists, etc.

Grand challenges for Information Retrieval

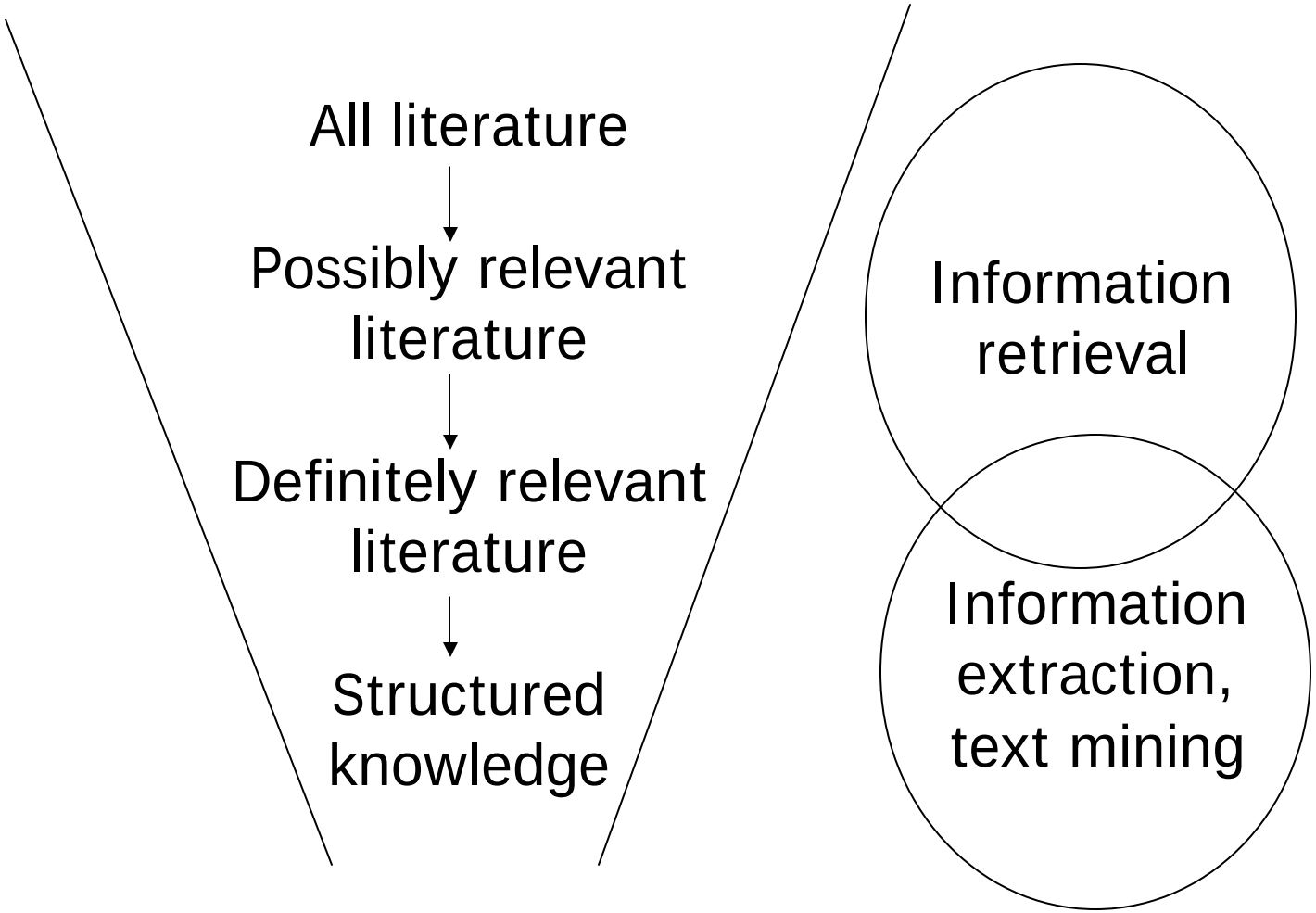
- Covered in this talk
  - Challenges for one important class of users, biomedical researchers
  - Open access to literature that is protective of intellectual property
- Others of great interest, for another day
  - Challenges for other classes of users, in particular clinicians and patients/consumers
  - Indexing – metadata for Web content
  - Evaluation – best measures, meaningful studies
  - Quality of health information on the Web


Content

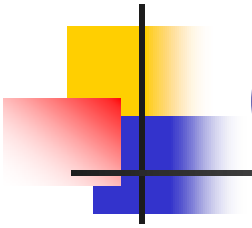
The information challenges for biomedical researchers

- We are in an era of “high throughput,” data-intensive science
- Biology and medicine provide many information challenges for information retrieval, extraction, mining, etc.
- Many reasons to structure knowledge with development of annotation, model organism databases, cross-data linkages, etc.
- Growing array of publicly accessible data resources and tools that may aid these tasks

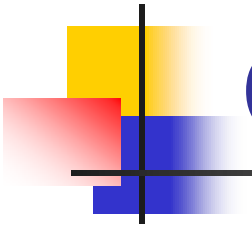
Emerging approach to biological knowledge management



Text Retrieval Conference (TREC, trec.nist.gov)



- Forum for comparative evaluation of IR systems
  - Competition minimized; collegiality maximized
- Organized by NIST
- Annual cycle consisting of
  - Distribution of test collections and queries to participants
  - Determination of relevance judgments and results
  - Annual conference for participants at NIST
- Began in 1992 and has continued annually



Organization of TREC

- Began with two major tasks, both of which have been discontinued
  - Ad hoc retrieval – standard searching
  - Routing – identify new documents with queries developed for known relevant ones
- Has evolved to a number of tracks devoted to specific interests
  - 7 tracks per year – each usually runs for 2-4 years
  - Past and current tracks have included question-answering, interactive, cross-language, Web, etc.
  - And now, retrieval in a domain (genomics)

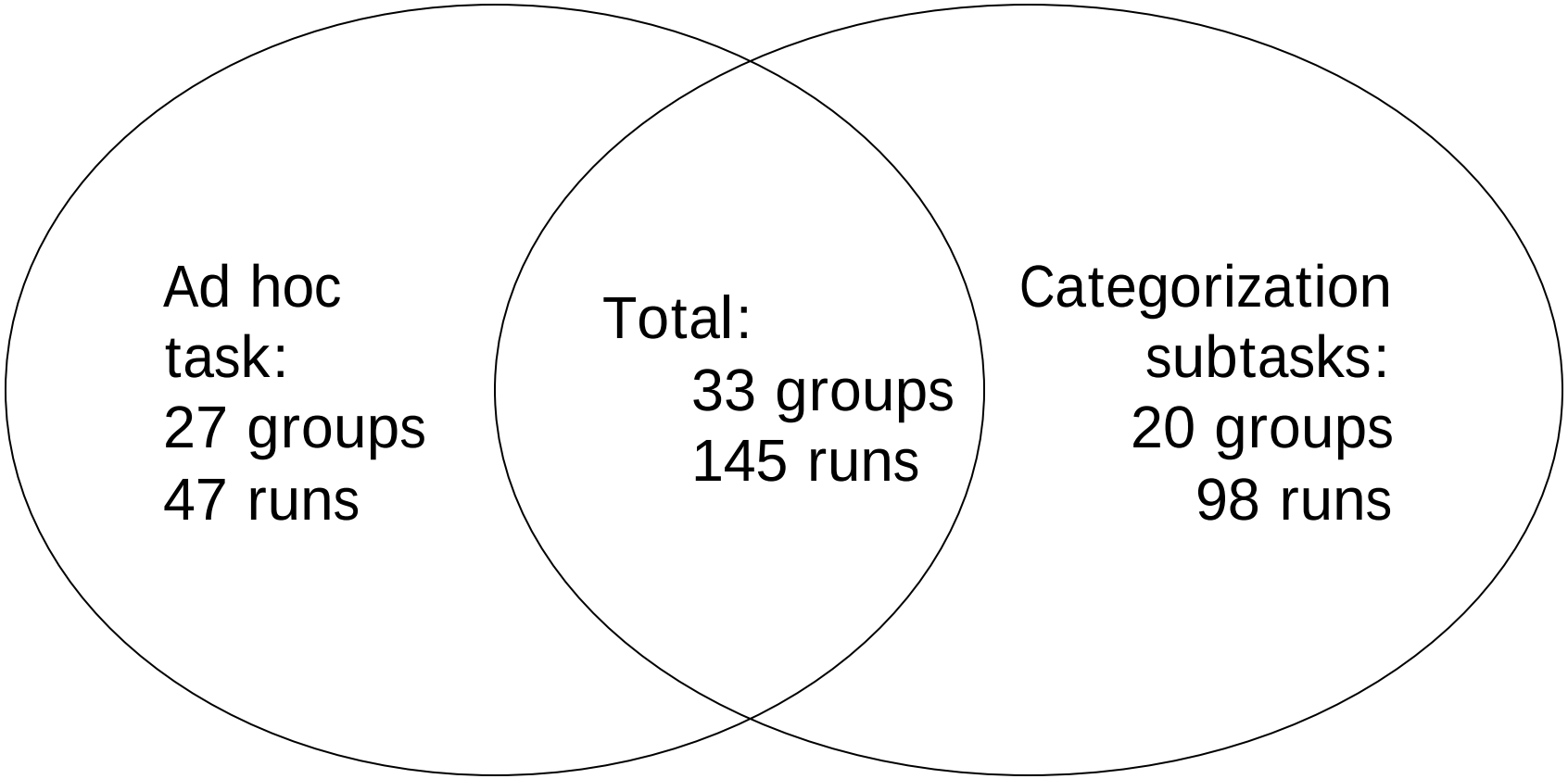
TREC 2004 Genomics Track

- Second year of track (Hersh, 2004), first year fully funded
  - http://medir.ohsu.edu/~genomics
- Two tasks
  - Ad hoc retrieval
    - Modeled after biologist with acute information needs
    - Used MEDLINE bibliographic database – despite proliferation of full-text journals, still entry point into literature for most searchers
  - Categorization
    - Motivated by real-world problems faced by Mouse Genome Informatics (MGI) curators, e.g., choosing articles and applying Gene Ontology (GO) terms for gene function
    - Divided into subtasks of article triage and annotation

A personal change of direction

- My previous work
  - Focus on clinical informatics
  - Emphasis on user-oriented evaluation based on limitations of system-oriented approaches
- But it will come full circle
  - Genomics is the future of medicine
  - Will add user studies and clinical pertinence
- This work also demonstrates
  - Bioinformatics has more in common with medical informatics than we may think!

Track participation – largest number of participants



Ad hoc retrieval task

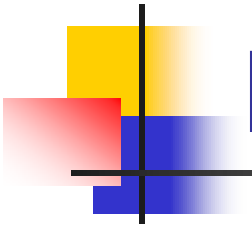
- Documents
  - MEDLINE subset
    - 10 years from 1994 to 2003
    - ~4.5M documents
      - About one-third of entire database, which goes back to 1966
    - ~9 GB text (MEDLINE format)
- Topics
  - Based on real biologist information needs
  - 50 topics (and 5 samples) based on
    - 74 real information needs
    - Collected from 43 biologists by 11 interviewers
    - Each reviewed by 1-2 others who turned into “searchable” topic

Example topic

<TOPIC>
<ID>**51**</ID>
<TITLE>**pBR322 used as a gene vector**</TITLE>
<NEED>**Find information about base sequences and restriction maps in plasmids that are used as gene vectors.**</NEED>
<CONTEXT>**The researcher would like to manipulate the plasmid by removing a particular gene and needs the original base sequence or restriction map information of the plasmid.**</CONTEXT>
</TOPIC>

Relevance judgments

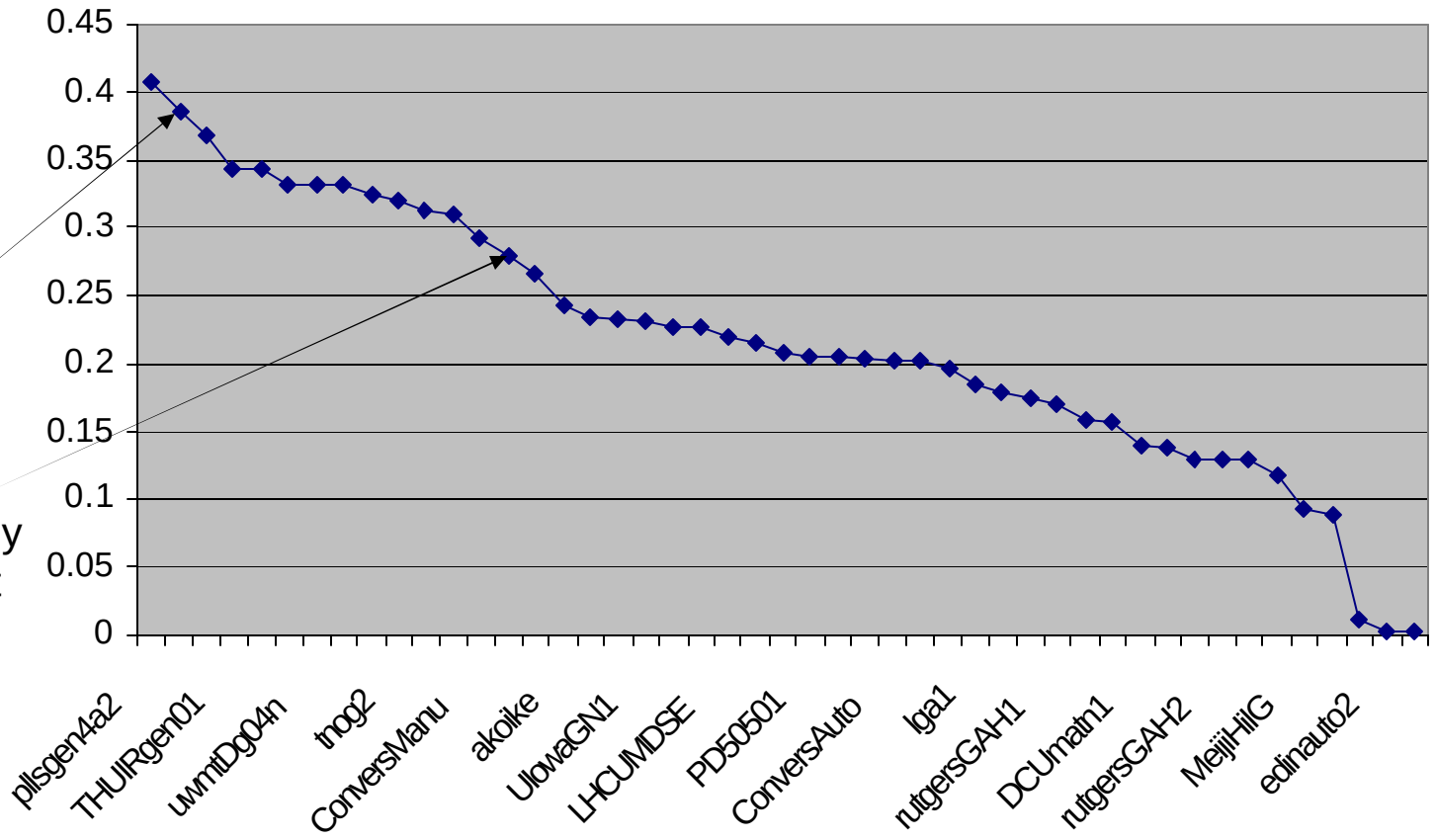
- Using standard TREC pooling method
  - Assessed top designated runs of the 27 groups who submitted results
- Performed by two judges – a PhD biologist and undergraduate biologist
  - Kappa = 0.51 – agreement “fair”; typical for IR
- Averages per topic
  - Documents assessed: 975
  - Definitely relevant: 93 (9%; range 1-506)
  - Possibly relevant: 73 (7%; range 0-485)
  - Definitely + possibly relevant (relevance for runs): 166 (16%; range 1-697)
    - Three topics had no definitely relevant documents



Metrics and analysis

- Primary performance metric – mean average precision (MAP)
- Also measured precision@10 and precision@100 documents
- Groups had additional measurements from trec_eval
- Statistical analysis – repeated measures ANOVA with posthoc Tukey pairwise comparisons

Ad hoc task results



Univ. of Waterloo!

Statistically significant from top run

n = 47, Max = .4075, Median = .2074, Min = .0012

Ad hoc task analysis

- Best runs used a variety of techniques, including
  - Domain-specific query expansion
  - Language modeling techniques, e.g., smoothing
- Of note, simple OHSU runs using Lucene “out of the box” (TF*IDF weighting) scored above mean/median
  - OHSUNeeds = .2343, OHSUAll = .2272
  - In other words, many groups did detrimental things!

Categorization task

- Motivation
  - Apply text categorization to full-text documents for tasks that assist work of MGI
- Sub-tasks
  - Triage – determine if articles have experimental evidence warranting GO assignment
    - A pertinent task beyond gene function annotation
  - Annotation – determine if article warrants assignment of GO category, with or without evidence code(s)
- Why not annotate actual GO terms?
  - Avoid exact overlap with Biocreative
  - A hard task, as learned from Biocreative
- More details and results in overview paper on Web site

Access

Access to the archive of science

- Problem more acute in biomedicine than most other scientific fields, e.g., computer science
- Impediments to wider dissemination are economic and political, not technical
  - Journals have monopolies due to promotion and tenure concerns
- There is growing concern over
  - Cost of journals in era of constrained library budgets
  - Shift from paper to electronic access – you no longer get to keep your back issues

Call for “open access” to scientific research results

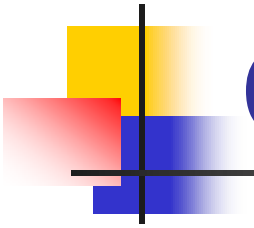
- Rationale: Most research publicly funded, yet reports of results copyrighted by publishers
  - If such information may be life-saving, it should be freely available
- Challenges: Production of information is not free and where do you draw the line with secondary publications
- Perspectives: Weiss, 2003; DeAngelis, 2004
- Proposed legislation: Sabo bill would prohibit copyrighting of all US government-funded research (McLellan, 2003)

Better-known open access publishing initiatives

- PubMed Central – pubmedcentral.gov
- BioMed Central (Hersh, 2001) – www.biomedcentral.com
- Public Library of Science (Butler, 2003) – www.plos.org
- Latter two bring publishing model full circle back to electronic equivalent of page charges in exchange for open access
  - Assumption that cost should be built into research budgets, with provisions for those unable to pay

Current status of open access publishing

- NIH issued request for information (RFI) on proposed regulations
  - Over 6,000 replies, most in favor of open access policies
- Pushback from major journals and publishers, expressing concern about
  - Financial viability for both for-profit and non-profit publishers
  - Access to publishing for unfunded and developing world researchers
  - Government control of publishing
- Compromise policy was to be unveiled this month, calling for free release of research papers after 6-12 months
  - Withdrawn due to political pressure, in part because of new incoming Secretary of HHS
- Future direction unclear, but most journals opening up archives after 6-12 months already and some adopting open access



Conclusions

- IR systems have become “mainstream”
- Searching is an essential skill for knowledge workers and perhaps the rest of the world as well
- Basic searching is simple and easy to do
- Challenges remain in providing access to the appropriate and/or best information while preserving the incentive to produce it