

# Approaches to Research in IR

Bruce Croft

Center for Intelligent Information Retrieval  
University of Massachusetts Amherst

Information Retrieval Research Group  
RMIT University

# What is this about?

- Broad overview of IR research
  - Historical context, unique characteristics
- How to do good IR research
  - Choose a topic, produce good papers, make an impact
  - Important for students and researchers, academics and industry (but focus here is on grad students)
- Why me and what is my bias?
  - Over 40 years in academic research, >40 Ph.D. students, many postdocs and visitors, lots of publications, many interactions with industry
  - System-oriented (vs. user-oriented) research

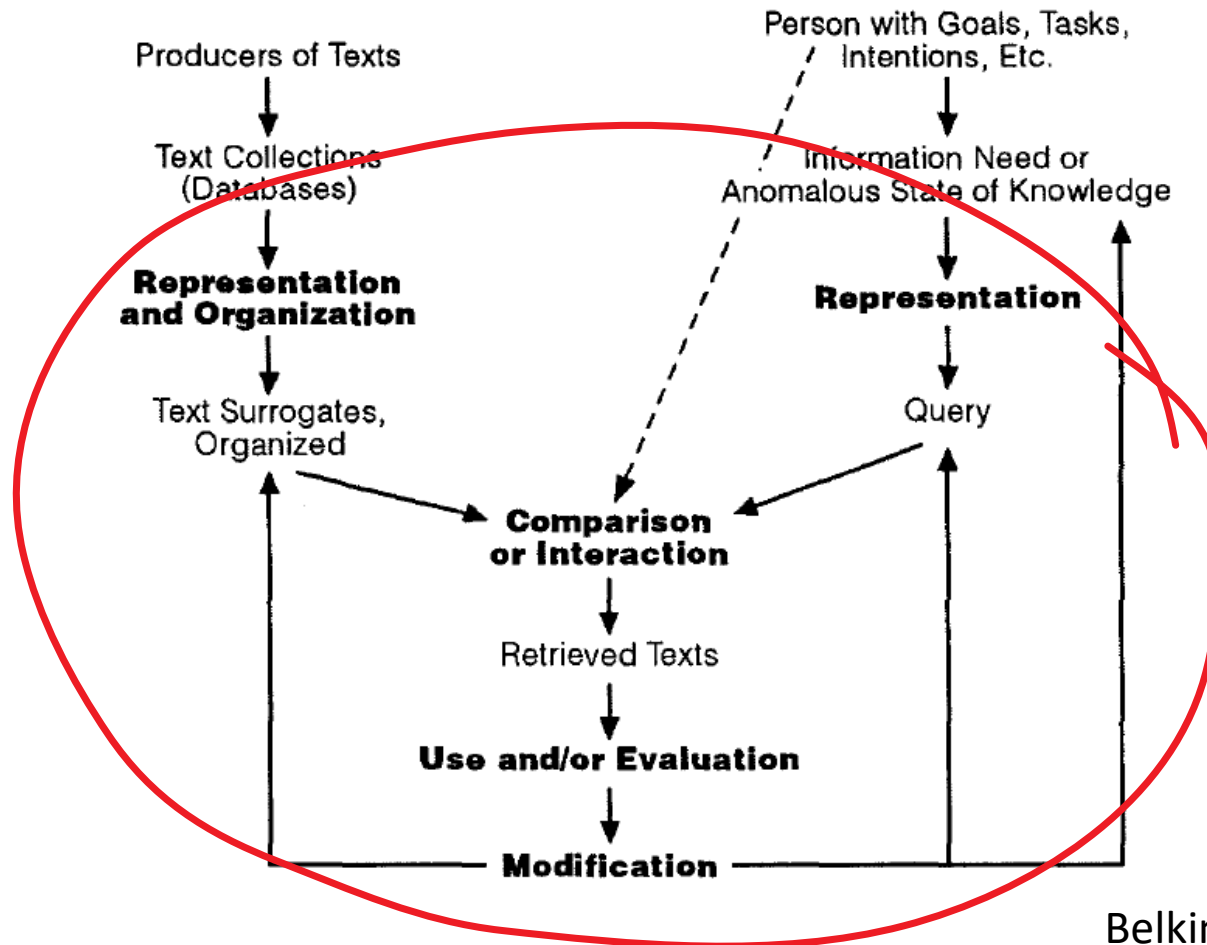
# Brief history of IR

- Information retrieval has deep roots in Information and Library Science
  - e.g., Cranfield experiments
  - Mostly user-oriented and evaluation research based on existing systems, but also design
    - e.g., Bush, Swanson, Maron, Cooper, Belkin
- More emphasis developed on how to build IR systems
  - e.g., Salton, Sparck Jones, Robertson
  - More computer science, although boundaries were fuzzy

# Brief history of IR

- *“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”*  
(Salton, 1968)
  - General definition that can be applied to many types of information and search applications
- Applications have changed radically over the past 60 years
  - Libraries, information services, legal and medical information, desktop search, government information, news, web search, recommendation, verticals (e.g., images), social media, question answering, mobile search, conversational assistants

# The Big Picture



Belkin and Croft, 1992

# Research Evolution

- Major research themes have stayed mostly the same, but research topics have evolved as systems and applications change
  - e.g., query evolution

1970 CE  
(Boolean search)



NEGLECT! FAIL! NEGLIG! /5 MAINT! REPAIR! /P  
NAVIGAT! /5 AID EQUIP! LIGHT BUOY  
"CHANNEL MARKER"

1994 CE  
(web search)



negligence navigation aids


2005 CE  
(CQA)



Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?

# Research Timeline

1960s



Evaluation of Boolean systems

Automatic vs manual indexing

Ranking vs. Boolean

Evaluation of ranking

Boolean queries vs “natural language”

Term weighting models

TF.IDF

Vector space model

Relevance feedback (including evaluation)

Clustering for retrieval

Query transformation (stopwords, stemming, expansion)

Ranking efficiency (inverted file optimization)

Probabilistic ranking models

Intermediaries (user studies and systems)

Full text vs. abstracts

Interfaces for effective search

Filtering (recommendation)

Word context models

Web search

1990s

Feature-based retrieval and fusion

Question answering (factoid)

Language models

Semi-structured documents

Multimedia and image search

Performance prediction

Query logs, click graphs, citation graphs, random walks

Adversarial information retrieval

Learning to rank

Long query and dependence models

Search aggregation and diversification

Personalized search

Exploratory search

Task-based search

Neural IR models

Conversational IR models

Fairness and bias models

Information privacy

# TREC

- Text Retrieval Conference started in 1992 as a workshop on evaluation
  - major influence on research in IR and related fields
  - TREC Tracks defined many major research areas
  - Spinoffs include CLEF, NTCIR, FIRE
  - Should be considered by any IR research program, but should not be considered as a prerequisite for a research program



# TREC history

## Past tracks [\[ edit \]](#)

- **Chemical Track** - **Goal:** to develop and evaluate technology for large scale search in [chemistry](#)-related documents, including academic papers and patents, to better meet the needs of pr
  - **Clinical Decision Support Track** - **Goal:** to investigate techniques for linking medical cases to information relevant for patient care
  - **Contextual Suggestion Track** - **Goal:** to investigate search techniques for complex information needs that are highly dependent on context and user interests.
  - **Crowdsourcing Track** - **Goal:** to provide a collaborative venue for exploring [crowdsourcing](#) methods both for evaluating search and for performing search tasks.
  - **Genomics Track** - **Goal:** to study the retrieval of [genomic](#) data, not just gene sequences but also supporting documentation such as research papers, lab reports, etc. Last ran on TREC 2
  - **Dynamic Domain Track** - **Goal:** to investigate domain-specific search algorithms that adapt to the dynamic information needs of professional users as they explore in complex domains.
  - **Enterprise Track** - **Goal:** to study search over the data of an organization to complete some task. Last ran on TREC 2008.
  - **Entity Track** - **Goal:** to perform entity-related search on Web data. These search tasks (such as finding entities and properties of entities) address common information needs that are not
  - **Cross-Language Track** - **Goal:** to investigate the ability of retrieval systems to find documents topically regardless of source language. After 1999, this track spun off into [CLEF](#).
  - **FedWeb Track** - **Goal:** to select best resources to forward a query to, and merge the results so that most relevant are on the top.
  - **Federated Web Search Track** - **Goal:** to investigate techniques for the selection and combination of search results from a large number of real on-line web search services.
  - **Filtering Track** - **Goal:** to binarily decide retrieval of new incoming documents given a stable [information need](#).
  - **HARD Track** - **Goal:** to achieve High Accuracy Retrieval from Documents by leveraging additional information about the searcher and/or the search context.
  - **Interactive Track** - **Goal:** to study user [interaction](#) with text retrieval systems.
  - **Knowledge Base Acceleration Track** - **Goal:** to develop techniques to dramatically improve the efficiency of (human) knowledge base curators by having the system suggest modifications
  - **Legal Track** - **Goal:** to develop search technology that meets the needs of lawyers to engage in effective [discovery](#) in [digital document](#) collections.
  - **LiveQA Track** - **Goal:** to generate answers to real questions originating from real users via a live question stream, in real time.
  - **Medical Records Track** - **Goal:** to explore methods for searching unstructured information found in patient medical records.
  - **Microblog Track** - **Goal:** to examine the nature of real-time information needs and their satisfaction in the context of microblogging environments such as Twitter.
  - **Natural language processing Track** - **Goal:** to examine how specific tools developed by computational linguists might improve retrieval.
  - **Novelty Track** - **Goal:** to investigate systems' abilities to locate new (i.e., non-redundant) information.
  - **OpenSearch Track** - **Goal:** to explore an evaluation paradigm for IR that involves real users of operational search engines. For this first year of the track the task will be ad hoc Academ
  - **Question Answering Track** - **Goal:** to achieve more [information retrieval](#) than just [document retrieval](#) by answering factoid, list and definition-style questions.
  - **Real-Time Summarization Track** - **Goal:** to explore techniques for constructing real-time update summaries from social media streams in response to users' information needs.
  - **Robust Retrieval Track** - **Goal:** to focus on individual topic effectiveness.
  - **Relevance Feedback Track** - **Goal:** to further deep evaluation of relevance feedback processes.
  - **Session Track** - **Goal:** to develop methods for measuring multiple-query sessions where information needs drift or get more or less specific over the session.
  - **Spam Track** - **Goal:** to provide a standard evaluation of current and proposed [spam filtering](#) approaches.
  - **Tasks Track** - **Goal:** to test whether systems can induce the possible tasks users might be trying to accomplish given a query.
  - **Temporal Summarization Track** - **Goal:** to develop systems that allow users to efficiently monitor the information associated with an event over time.
  - **Terabyte Track** - **Goal:** to investigate whether/how the [IR](#) community can scale traditional IR test-collection-based evaluation to significantly large collections.
  - **Total Recall Track** - **Goal:** to evaluate methods to achieve very high recall, including methods that include a human assessor in the loop.
  - **Video Track** - **Goal:** to research in automatic segmentation, [indexing](#), and content-based retrieval of [digital video](#).
- In 2003, this track became its own independent evaluation named [TRECVID](#).
- **Web Track** - **Goal:** to explore information seeking behaviors common in general web search.

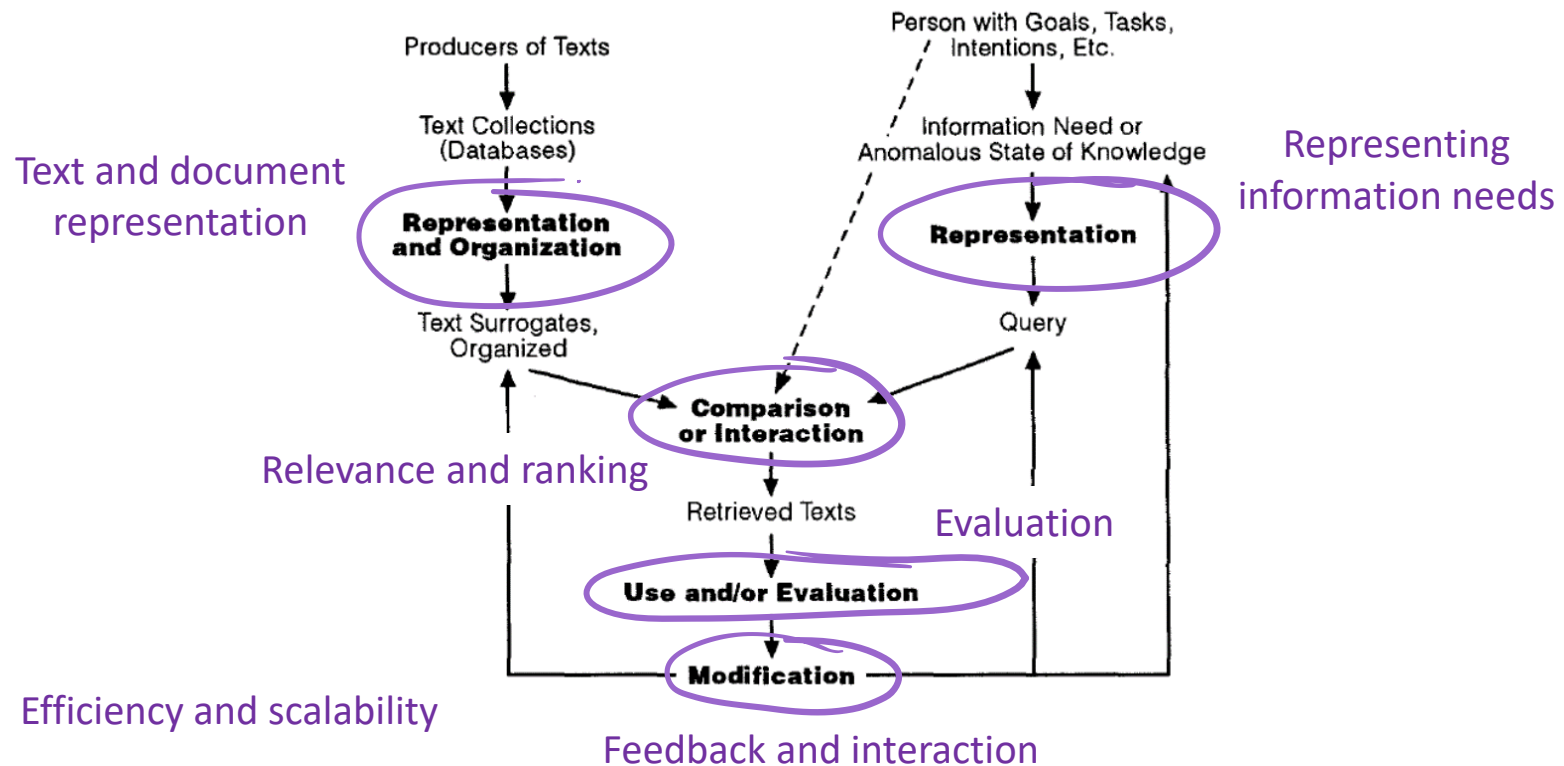
# Why choose IR?

- Providing easy access to the world of information has always been a key motivation and still is
  - Good business and good for the world
  - More than social media?
- Huge range of research challenges and potential applications based on this simple motivation
  - Many of these are not solved or could be improved significantly, even given the commercial success

Exercise:

Why did you choose IR as a research area?

# Central Themes of IR Research



# Limitations of the “Big Picture”

- Some overviews of IR tend to emphasize a “one-shot” perspective on search
- Exercise: important research issues not covered in this simple view?
- history, context, sessions, external knowledge/data, provenance, multimedia, multiple representations and ranking models, aggregated search, diversification, relevance vs answers, generated answers, performance prediction...
- Conference descriptions can be useful but also confusing as a guide to research

# Beyond the Basics

- SIGIR 2012 CFP

- Document Representation and Content Analysis (e.g., text representation, document structure, linguistic analysis, non-English IR, cross-lingual IR, information extraction, sentiment analysis, clustering, classification, topic models, facets)
- Queries and Query Analysis (e.g., query representation, query intent, query log analysis, question answering, query suggestion, query reformulation)
- Users and Interactive IR (e.g., user models, user studies, user feedback, search interface, summarization, task models, query logs, personalized search)
- Retrieval Models and Ranking (e.g., IR theory, language models, probabilistic retrieval models, feature-based models, learning to rank, combining searches, diversity)
- Search Engine Architectures and Scalability ( e.g., indexing, compression, MapReduce, distributed IR, P2P IR, mobile devices)
- Filtering and Recommending (e.g., content-based filtering, collaborative filtering, recommender systems, profiles)
- Evaluation (e.g., test collections, effectiveness measures, experimental design)
- Web IR and Social Media Search (e.g., link analysis, social tagging, social network analysis, advertising and search, blog search, forum search, CQA, adversarial IR, vertical and local search)
- IR and Structured Data (e.g., XML search, ranking in databases, desktop search, entity search)
- Multimedia IR (e.g., Image search, video search, speech/audio search, music IR)
- Other Applications (e.g., digital libraries, enterprise search, genomics IR, legal IR, patent search, text reuse)

# Beyond the Basics

- SIGIR 2019 CFP

**Search and Ranking.** Research on core IR algorithmic topics, including IR at scale, covering topics such as:

- Queries and query analysis
- Web search, including link analysis, sponsored search, search advertising, adversarial search and spam, and vertical search
- Retrieval models and ranking, including diversity and aggregated search
- Efficiency and scalability
- Theoretical models and foundations of information retrieval and access

**Future Directions.** Research with theoretical or empirical contributions on new technical or social aspects of IR, especially in more speculative directions or with emerging technologies, covering topics such as:

- Novel approaches to IR
- Ethics, economics, and politics
- Applications of search to social good
- IR with new devices, including wearable computing, neuroinformatics, sensors, Internet-of-Things, vehicles

**Domain-Specific Applications.** Research focusing on domain-specific IR challenges, covering topics such as:

- Social search
- Search in structured data including email search and entity search
- Multimedia search
- Education
- Legal
- Health, including genomics and bioinformatics
- Other domains such as digital libraries, enterprise, news search, app search, archival search

**Content Analysis, Recommendation and Classification.** Research focusing on recommender systems, rich content representations and content analysis, covering topics such as:

- Filtering and recommender systems
- Document representation
- Content analysis and information extraction, including summarization, text representation, readability, sentiment analysis, and opinion mining
- Cross- and multilingual search
- Clustering, classification, and topic models

# Beyond the Basics

- SIGIR 2019 continued

**Artificial Intelligence, Semantics, and Dialog.** Research bridging AI and IR, especially toward deep semantics and dialog with intelligent agents, covering topics such as:

- Question answering
- Conversational systems and retrieval, including spoken language interfaces, dialog management systems, and intelligent chat systems
- Semantics and knowledge graphs
- Deep learning for IR, embeddings, and agents

**Human Factors and Interfaces.** Research into user-centric aspects of IR, including user interfaces, behavior modeling, privacy, and interactive systems, covering topics such as:

- Mining and modeling search activity, including user and task models, click models, log analysis, behavioral analysis, and attention modeling
- Interactive and personalized search
- Collaborative search, social tagging and crowdsourcing
- Information privacy and security

**Evaluation.** Research that focuses on the measurement and evaluation of IR systems, covering topics such as:

- User-centered evaluation methods, including measures of user experience and performance, user engagement and search task design
- Test collections and evaluation metrics, including the development of new test collections
- Eye-tracking and physiological approaches, such as fMRI
- Evaluation of novel information access tasks and systems such as multi-turn information access
- Statistical methods and reproducibility issues in information retrieval evaluation



# What is good IR research?

- Has *motivation*, clear research questions, novelty, theory, models, experiments, users, evaluation, impact
  - All related to IR research themes
  - What is the “elevator story”?
- Clearly defines prior research context
- Considers efficiency and implementation
  - In some cases as primary issues, otherwise secondary
- Reproducible
  - Code and data sharing, as much as possible

# What is good IR research?

- Produces good *science*, instead of leaderboard chasing or solely engineering work
  - Increases our understanding of IR

## Exercise:

Write down a research idea, focusing on the motivation, but including some description of the experiments, and the data you would need

- In other words, what is your elevator story?

# Industry vs Academic Research

- Google, Microsoft, Baidu, Amazon, Alibaba, Facebook etc. have hundreds of people working on R&D for search
- The same companies have most of the interesting data (i.e., query logs, purchase data)
- They can also pay for many people to do manual annotations of data (e.g., relevance judgments)
- Is it possible for academic researchers to do meaningful IR research in this environment?
  - or should we all just study evaluation metrics 😊

# Limitations of Academic IR research

- Small numbers of researchers at individual sites
- Small, very distributed community
- Very restricted access to user data
- Limited resources for computing
- Obtaining research funding can be difficult

# Advantages of Academic IR

- Large pool of motivated students
- Freedom to research anything, regardless of business case
  - Academic research agenda may not match “what’s the next hot thing in web search” and that’s ok
- Builds on, and acknowledges prior research
- Can freely share results and data with collaborators
- Easy to make the case that research will have an “impact”
  - IR has always thought about applications and defined research agendas in relation to them
- Innovative, passionate community including many industry partners

# IR vs ML and NLP

- Goal of IR research is to improve access to information - or understanding how people process, compare, analyze and retrieve information
  - Goal of ML is to develop more effective learning algorithms - the “application” is secondary
  - Goal of NLP is “developing systems for the understanding, analysis, manipulation, and/or generation of natural language” - although applications like translation, summarization, and sentiment analysis are an important part of the field
- IR research can make use of techniques and representations developed in ML and NLP and improve them for IR goals
  - Neural networks are an example

# IR vs ML and NLP

- IR focuses on retrieving existing text documents, passages, sentences or “factoids”
  - but also images, video, semi-structured data, even structured data
- Evaluations in IR are usually larger and more rigorous – this has had a major influence on other areas
- Emphasis on failure analysis, understanding models, and explainability
- Increasing overlap with NLP, particularly in QA and conversational search
  - More test collections available 😊
- Choose conferences for publications carefully since each area has somewhat different expectations



# Choosing a Research Topic

- Start by doing a variety of projects
  - Larger project themes usually develop from smaller pieces of work
- Focus on the project that interests you the most
  - and is interesting to your advisor!
- Don't expect everything to develop according to schedule
  - Thesis proposals are very helpful but changes can happen
  - Follow up on the most promising work, drop directions that are not working (after trying multiple approaches)

# Choosing a Research Topic

- Have clear motivation for the project, before anything else
- Be first to describe a research problem, if possible
- The problem should have interesting technical challenges
- Choose a problem that is feasible in the time allowed, with short-term milestones
- Always try to relate the topic to existing problems and solutions. Don't ignore prior work to justify the contribution

# Choosing a Research Topic

- Read papers - from different topics, even from other related fields, like ML, NLP, RecSys, and KDD
- Find and follow leaders in the field
- Having the data to do research is crucial, but don't wait for a TREC track
  - Also don't wait for industry partners to provide data!
- Don't do research in academic environment that would be better done in industry (generally)
  - Counterexamples – click model research from Joachims (Cornell) and Agichtein (Emory)

# Examples of academic IR areas

- “Long-tail” issues for web search
- Tasks other than standard “ad-hoc” search
  - e.g., cross-lingual, documents as queries, entity search
- Intersection of IR and NLP
  - e.g., QA and conversational IR
- Complex search and exploration
- Detailed user studies
  - e.g., interactive IR, eye-tracking
- Evaluation
  - e.g., metrics, significance, methodology

# Choosing a Research Topic

- Make sure that there is a sound evaluation plan, including appropriate baselines
  - Evaluation is key in the IR community
  - May have to implement other peoples' algorithms
- Importance of crowdsourcing and scraping
  - Spend sweat equity, don't wait for someone else to create the data
- Recycling data collections
  - Be smart about using existing data collections from TREC, CLEF, NLP projects, companies and repurposing them for your project

# Examples of Current Research Topics

- SWIRL 2018
  - Workshop focused on identifying important research directions in IR (also SWIRL 2012)
- Current TREC, CLEF and NTCIR tracks
- Major funding initiatives
- Examples of recent CIIR projects

# SWIRL 2018

- **Decision Support over Pathways:** Understanding and designing systems to help people in making decisions.
- **Generating New Information Objects:** *Ad hoc* generation, composition, and summarization of new text, and layouts in response to an information need.
- **Transparent/Explainable Information Retrieval:** Explaining ranking decisions. Providing reliable and responsible information access.
- **Cognitive-aware IR:** Tracking and modeling user behavior and perception. Modeling political-correctness of decisions. Identifying fake news and provenance.
- **Societal impact of information retrieval:** Understanding the long term impact of IR on society and the economy.
- **Personal information access:** Federated personal information search and management (e.g. knowledge graphs). Biometrics for affective state.
- **Next Generation Efficiency-Effectiveness Issues:** Efficient machine learning inference. Resource-constrained search.
- **Machine Learning and Search:** Developing effective machine-learned retrieval models (e.g. neural networks, reinforcement learning, meta-optimization).
- **Personalized interaction:** Diversified and personalized interactions.
- **Conversational information access:** Information-seeking conversations. Learning representations for conversations.
- **New approaches to evaluation:** Moving beyond the Cranfield paradigm, topical relevance, and queries. Controlling for variability. Counterfactual evaluation and off-policy evaluation.
- **New interaction modes with information, multi-device search:** Multi-device search.
- **Blending online and physical:** Search in the context of mobile, smart environments, and augmented/virtual reality.
- **Task-specific representation learning:** Adapting machine learned models for new search domains.
- **Pertinent Context:** Surfacing and using the relevant contextual information for search.
- **Success prediction:** Formal models and principles to inform retrieval system design (build the right bridge instead of build six bridges and see which survives).

# SWIRL 2018 Major Themes

- Conversational Information Seeking
- Fairness, Accountability, Confidentiality and Transparency in Information Retrieval (now FACTS)
- IR for Supporting Knowledge Goals and Decision-Making
- Evaluation
- Machine Learning in Information Retrieval (Learnable IR)
- Generated Information Objects
- Efficiency Challenges
- Personal Information Access



# 2018 TREC Tracks

## **CENTRE Track**

This is a new track for 2018, which will run in parallel (with somewhat different emphases) in CLEF 2018, NTCIR-14, and TREC 2018. The overall goal of the track is to develop and tune a reproducibility evaluation protocol for IR.

## **Common Core Track**

The Common Core track uses an ad hoc search task over news documents. As such, it serves as a common task for a wide spectrum of IR researchers to attract a diverse run set that can be used to investigate new methodologies for test collection construction.

## **Complex Answer Retrieval Track**

The focus of the Complex Answer Retrieval track is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus.

## **Incident Streams Track**

This is a new track for TREC 2018. The Incident Streams track is designed to bring together academia and industry to research technologies to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators.

## **2018 TREC Tracks (cont'd)**

### **News Track**

The News Track is a new track for 2018. It will feature modern search tasks in the news domain. In partnership with The Washington Post, we will develop test collections that support the search needs of news readers and news writers in the current news environment.

### **Precision Medicine Track**

This track is a specialization of the Clinical Decision Support track of previous TRECs. It will focus on building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments as well as evidence-based literature to identify the most effective existing treatments.

### **Real-Time Summarization Track**

The Real-Time Summarization (RTS) track explores techniques for constructing real-time update summaries from social media streams in response to users' information needs.

# CIIR project examples

- H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, and J. Kamps. “From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing”. CIKM 2018.
- K. Bi, Q. Ai, and W B. Croft. “Iterative Relevance Feedback for Answer Passage Retrieval with Passage-Level Semantic Match”. ECIR 2019.
- L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. “Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems”. SIGIR 2018.
- C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu. “User Intent Prediction in Information-seeking Conversations”. CHIIR 2019.

Exercise:

Discuss selection of audience research topics

# Publications

- How, when, where, why and what to publish
- Conference papers, journal articles, proposals....
- Publish early and often
  - Use deadlines as motivation to advance your research
  - Should have solid results but don't wait too long
  - Place your stake in the ground!
  - Don't forget arXiv
- Conferences and workshops are the place to establish your reputation
- Journal articles if your career requires them
- Best venues – SIGIR, ECIR then CIKM, ICTIR, WSDM, CHIIR, TheWebConf
  - Smaller venues like DESIRES are also very good

# Conference Papers

- “Bread and butter” of most CS areas
- Used as metric for progress for grad students, important for Ph.D. portfolio, major component of CV for hiring, promotion
- Most important dissemination tool (look at Scholar, for example)
- Short papers are hardest to write

# Conference Papers

- Typically 8-12 pages, two-column
  - Short papers range from 2-6 pages
- Very low acceptance rates at good conferences
- Conference receives hundreds of papers
- Reviewing can be somewhat random
  - In particular, can get reviewers who are not that familiar with the topic area of your paper
- Rejection is part of academic life and it is possible to improve papers based on reviews (sometimes)

# Conference Papers

- Most important parts of a conference paper?
  - Title and abstract
  - Introduction
  - References
  - Try to justify all of your decisions, by referring to the literature, convincing reasoning, or doing experiments.
  - The contributions should be clear. The experiments should support the contributions.
  - Have clear conclusions. What do people learn from the paper?



# General Points

- Spelling and typos!!!
- Grammar!!
- Casual language, slang, abbreviations
  - “a bunch of sources”, “lots of data”, “ur account”, “IR”
- Some of my personal favorites
  - “it’s”
    - it is
  - “utilize”
    - use
  - “traditionally”
    - typically, often

# Editing

- Advisors and managers are not editors!
  - Although advisors are at least supposed to teach you how to write
  - They want to focus on content and maybe style
- Use colleagues or even friends to do copy editing/first pass
  - Maybe a professional for a serious document (e.g., Ph.D. thesis)

# Conference Papers

- Rule 1: Avoid “cute” titles
  - e.g., To sing, you must close your eyes and draw
  - Freshness Matters: In Flowers, Food, and Web Authority
  - To Translate or Not to Translate?
  - Web Search Solved? All Result Rankings the Same?
- Rule 2: Keep the abstract short
  - Motivation, main goal, what you did, overview of results
- Rule 3: Live in the present
  - Mostly stick to present tense, except for prior work, but *be consistent*
- Rule 4: Don't be too passive
  - Avoiding personal pronouns is generally good, but sometimes active voice can sound better

# Conference Papers

- Rule 5\*: Tell a story
  - Introduction needs to lay out the “plot”, give convincing motivation, overview of approach, why this approach is different to previous work, and clearly state the main contributions.
  - The rest of the paper has to stick to this story so reader always has a “map” in their mind to relate what they are reading to your overall theme

# Conference Papers

- Rule 6\*: Be careful with citations
  - Reviewers are always annoyed if you leave out their favorite citation (especially one of theirs)
  - Too many citations better than too few
  - Previous research does exist pre-Google (i.e., 2000)
  - Try to provide a framework for your literature review in “related work” section
  - Related work section usually is better at start
  - Don’t just cite – comment and relate to your work
  - Be careful of self-plagiarism

# Conference Papers

- Rule 7: Don't make definitive statements
  - e.g., “it is well-known that approach x is the most effective”
- Rule 8: Avoid explosions
  - Don't say anything about the Internet, Web, social media, conversational assistants” or anything else “exploding” or “rapidly increasing” in the first paragraph
- Rule 9: Give examples
  - Running example is best

# Conference Papers

- Rule 10\*: Don't leave out details in experiments
  - Describe test collections, user annotations, training, parameter settings, significance tests, effectiveness metrics, efficiency metrics in detail (doesn't mean repeating well-known formulas)
  - Use multiple metrics
  - Describe baselines and why they are the best comparisons
  - Don't make claims that aren't justified by results
  - Analyze your successes and failures

# Journal Papers

- Similar rules to conference papers
  - Keeping track of theme/story can be harder
- Even less tolerance for missing detail, citations
- Must have comprehensive experiments and analysis of results
- If you believe in some aspect of the paper that is criticized, argue with reviewers
- Specific rules for allowed overlap with conference papers



# ACM Policy

- The technical contributions appearing in ACM journals are normally original papers which have not been published elsewhere. Widely disseminated conference proceedings and newsletters are a form of publication. **A submission based on a paper appearing elsewhere must have major value-added extensions to the version that appears elsewhere.** For conference papers, there is little scientific merit in simply sending the submitted version to a journal once the paper has been accepted for the conference. The authors learn little from this, and the scientific community gains little.
- The submitted manuscript should thoroughly consolidate the material, should extend it to be broader, and should more carefully cover related research. **It should have at least 30% new material.** The new material should be content material, not just the addition of proofs or a few more performance figures. This affords an opportunity to describe the novel approach in more depth, to consider the alternatives more comprehensively, and to delve into some of the issues listed in the other paper as future work.

# ACM Policy

- Widely disseminated refereed conference proceedings, in addition to journal papers, are considered publications, but technical reports and CORR articles (neither of which are peer reviewed) are not. All overlapping papers appearing in workshop proceedings and newsletters should be brought to the editor's attention; they may be considered publications if they are peer reviewed and widely disseminated.
- Regardless of policy, if the introduction is nearly identical, you will almost certainly get rejected

# The Thesis

- Many local variations, depends heavily on institution and advisor
- How long should it take?
  - 3-6 years
- How good does it have to be?
  - Good enough to be accepted without major revisions
- What is in it?
  - Contributions equivalent to at least three conference papers IMHO
- Emphasize contributions, relationship to prior research, evaluation, details for reproducibility

# Proposals

- Critical skill for your survival as an academic
  - Also important in companies
- Many proposals done collaboratively
  - More difficult than writing it yourself
  - Have to get consistent theme, tone
  - Have to allow for other people being slackers
  - Be prepared to take charge, give assignments, set deadlines, merge inputs

# Presentations

- Typically 20-25 minutes at conferences
  - 1 hour for job talks
- Don't have too many slides
  - About 1 min/slide
  - Leave time for questions
- Don't have too much on a slide
  - Avoid repeating detail that is in the paper
  - Summarize main results, no huge tables, graphs ok
  - Have 1-3 “hard” slides to show there is depth

# Presentations

- Don't assume your audience knows everything you're talking about
  - Explain the basics clearly
- Tell the story!
  - Focus on take-away message

# Internships

- Many opportunities for internships at search companies
- Provide valuable experience and potential job opportunities
  - but can delay research progress
- Best situation is when internship research is related to thesis research, and company agrees that your work can be published
- Do 2-3 internships during a thesis, probably not your first year

# Collaboration and Independent Research

- Collaboration is crucial for success
- Sharing ideas and results is important
- We are scientists and researchers, not start-up CEOs
- Work on multiple projects during thesis
- Discuss ideas with different people with different backgrounds
- Work with students from other research groups
- Make sure your name is prominent in the research for your thesis, author list ordering policies vary but first author is important



# Summary

- IR is an exciting research area
- Choose a good research topic
- Do good science
- Publish often
- Remember that life is not all about research 😊