Whether it is searching patient records, helping medical professionals find best-practice evidence, or helping the public locate reliable and readable health information online, health search is a challenging area for IR research with an actively growing community and many open problems. This tutorial will provide attendees with a full stack of knowledge on health search, from understanding users and their problems to practical, hands-on sessions on current tools and techniques, current campaigns and evaluation resources, as well as important open questions and future directions.

CCS CONCEPTS
• Information systems → Expert search;

ABSTRACT
The HS2017 tutorial will cover topics from an area of information retrieval (IR) with significant societal impact — health search. Whether it is searching patient records, helping medical professionals find best-practice evidence, or helping the public locate reliable and readable health information online, health search is a challenging area for IR research with an actively growing community and many open problems. This tutorial will provide attendees with a full stack of knowledge on health search, from understanding users and their problems to practical, hands-on sessions on current tools and techniques, current campaigns and evaluation resources, as well as important open questions and future directions.

1 INTENDED AUDIENCE
Researchers of all levels seeking to understand the challenges, tasks and recent developments in information retrieval related to health (health search), be it consumer-oriented search, clinician search, or biomedical search. No prior knowledge in health search is required, making this tutorial ideal for those unfamiliar with this domain. The tutorial is also suitable for those familiar with health search as they will acquire insights from a hands-on session. The tutorial will also provide an analysis of successes and failures of current techniques, and an outline of the opportunities for IR research in the health domain.

2 MOTIVATION AND OVERVIEW
With modern medicine increasingly reliant on information technology, the demand for IR systems that search medical content has grown significantly. The increasing need to retrieve medical advice (by both consumers and clinicians), and the adoption of electronic medical records are two factors driving the demand for health search. IR research has much to offer here by developing new tools and techniques specific to this domain [17].

The range of health information available (primary research sources, secondary research sources, patient records, web pages and popular publications, etc.), the range of end users (health consumers, different clinicians — general practitioners, specialists, researchers, etc.), and the range of tasks (searching evidence-based-medicine literature [36], searching patient records and cohort selection [43], searching for medical advice on the Web [54], searching the literature for drug-drug interactions and co-morbidities [9], searching for clinical trials [23], etc.) leads to complex requirements that often require novel solutions to these different problems.

The key challenge in health search is how to bridge the semantic gap: the mismatch between the raw data and the way a human being interprets it. Although particularly prevalent in health search, the semantic gap problem is found in all domains [1]; advances in health search can thus advance the whole field. Key challenges include: how to leverage semantics and domain-knowledge resources for a better representation of documents and information need [26]; what characterises relevance, in particular how topicality is complemented by other dimensions of relevance [52] (understandability, authoritativeness, etc.), how bias and time pressure affects perception of relevance and decisions [33, 48] and how these influences the search process and evaluation.

Advances in health search will require familiarity with the tasks, users, successes, failures, and domain-specific resources. This tutorial will introduce researchers to the challenges and opportunities in health search, providing insights into current techniques and their results. It will also offer a hands-on overview of tools specific to the health domain made available by the clinical informatics and natural language processing communities.

This tutorial also draws on complementary efforts from other computer science fields in the health domain. Efforts from the clinical informatics and natural language processing communities have produced a wide array of tools that can dovetail with IR techniques. High quality domain knowledge resources (e.g., the UMLS¹ meta-thesaurus and SNOMED CT ontology²) have been developed to encode medical knowledge — these can be used for reasoning and inference within IR techniques [25, 53]. Natural language processing resources that identify medical concepts (from the aforementioned domain-knowledge resources) from free text have been developed (e.g., Metamap [5]). Similar tools have been created to extract other information from medical documents, such as negations, assertions and medications [15, 44]. The tutorial will provide

1https://www.nlm.nih.gov/research/umls/
2http://www.snomed.org/snomed-ct
hands-on demonstrations of how these tools and techniques can be exploited by IR systems.

IR has a long history of rigorous empirical evaluation; this is also the case for work related to health search. This tutorial will cover topics specific to health search evaluation: available test collection, evaluation resources, evaluation campaigns (TREC, CLEF, etc.), as well as insights on successes, failures and difficulties encountered. From an IR research perspective, novel aspects will include the consideration of multiple dimensions of relevance and new frameworks to evaluate systems based on task completion rather than relevance.

3 OBJECTIVES
The main aims of HS2017 will be to:

1. Summarise the basics of search in the health domain;
2. Present the different end user requirements for multiple user groups interested in health search, including tasks;
3. Provide an overview of the current use of IR techniques in the health domain;
4. Provide a hands-on introduction to domain-specific tools which can be exploited in health search;
5. Present resources and campaigns for evaluation in health search, including novel evaluation approaches;
6. Present challenges and opportunities for further research in the health domain and discuss how these could be met.

This knowledge will allow IR researchers to identify promising ways of applying their work to the health domain, allowing them to contribute to a domain of rapidly growing importance.

4 FORMAT AND SCHEDULE

4.1 Session 1: Background and theory
Introduction to the health domain and to the tutorial
The tutorial begins with an introduction to IR in health, giving an overview of the topics that will be covered in the tutorial and why they are important.
Duration: 15m

Types of health information
Health information comes in a myriad of forms. This section covers the characteristics of different types of health information sources important for health search. These range from patient-based information (e.g., electronic health records), knowledge-based information (e.g., scientific papers), through to consumer-based information (e.g., patient forums on the web). Also included are sources of domain knowledge such as medical ontologies, terminologies and classification systems, all of which are playing an increasingly important role in state-of-the-art IR systems.
Duration: 30m

End users and tasks
An analysis of the end user (from consumer [10, 49, 54] to clinician [34]) characteristics and tasks in health search is presented. Some groups of end users are addressed, and the information needs and search tasks they undertake are described. For example, the group of “physicians” can be divided into groups ranging from general practitioners operating from individual practices who require practice-oriented secondary literature, through specialists in a medical domain who wish to access more specialised works in this domain, to research physicians who have an interest in the primary literature.
Duration: 45m

4.2 Session 2: Techniques and methods
Methods in health search
This section covers the state-of-the-art in health search, summarising the most important research methods and results in this area with respect to the different tasks discussed in the previous session and highlighting common trends across tasks. This session will cover methods across different health search tasks, including: query expansion and reformulation [2, 32, 39, 42, 50, 51], use of domain knowledge and inference mechanisms [12, 24, 26, 29, 53], learning to rank and other learning methods [3, 6, 30, 31, 41], task-based information [21, 28], and specifically handling clinical text [7, 8, 16, 20, 22, 27]. We provide examples of health search systems already in use and lessons that can be drawn from their use.
Duration: 90m

4.3 Session 3: Practical
Use of clinical NLP tools for medical IR:
Hands-on session
The clinical informatics and clinical natural language processing community have developed a number of tools for extracting clinical information from free text. For example, statistical and dictionary-based named entity recognition systems have been developed to identify medical entities [5, 18, 40]; algorithms have been developed to extract associations, relationships and contexts and to enhance textual content with semantic information [11, 15, 47]. This part of the tutorial presents an overview of such techniques, providing a hands-on demonstration of how these tools work. In addition, the tutorial provides an outlook at how these tools have been used in the literature or can be integrated to enhance information representation and the whole information retrieval process.
Duration: 90m

4.4 Session 4: Evaluation and future directions
Evaluation
The tasks and challenges in evaluating health search are covered in this section. Evaluation that considers multiple dimensions of relevance (topicality, reliability, understandability [52], bias [48]) is presented. We touch on new frameworks to evaluate systems based on task completion rather than relevance. Evaluation campaigns and resources in this domain are presented, including TREC Medical Records Track [43, 45, 46], TREC Clinical Decision Support Track [36–38], CLEF eHealth (consumer health search [13, 14, 35, 55] and as of 2017 search systems for the compilation of systematic reviews), i2b2 Shared Task Challenges[4], ALTA Shared Task

(Query-based summarisation for evidence-based medicine) [4], clinical trial retrieval [23], and the use of ICD coded data as an automated relevance judgement mechanism [19].

**Duration:** 50m

**Open challenges for health search and conclusions (including discussion)**

There are many open problems in health search which are fertile ground for information retrieval (IR) research. Examples include: (i) searching for “similar” anonymised patient records or “similar” medical images within a hospital to assist in diagnosis or treatment; (ii) linking treatment guidelines to patient records based on their content; (iii) searching within a patient record to obtain an overview of the medical history. This part of the tutorial briefly presents some of the areas of medicine and consumer-health experience that have a potential to be improved through the use of IR techniques, leading to a discussion with the participants on meeting these challenges.

**Duration:** 40m

### 5 RELEVANCE TO THE IR COMMUNITY

Health related topics have become a common theme within IR. A number of venues, including SIGIR, have dedicated workshops, tutorials or tracks dedicated to health search. In recent times, SIGIR has had both tutorials (2012 and workshops (2014 and 2016). Similarly, other venues such as WSDM, WWW, KDD, and ACL have all hosted health related tutorials or workshops. This shows the interest from the community in health search. Even so, a number of important aspects of health search were never covered and are thus the focus of this tutorial:

- This tutorial will clearly provide an understanding of the users, their information needs, tasks and challenges that exist in this domain. This is critical as many of these differ from other domains.
- Practical sessions will provide attendees with hands-on experience with health search techniques, tools and problems.
- The tutorial will conclude with an analysis of open questions in the domain.

It is our intention to provide attendees with a full stack of knowledge in health search, from understanding the problems, practical solutions and fruitful areas of future work. It is our hope that this will equip them to contribute to an area of growing interest in the field with significant societal impact.

### A PRESENTERS

**Guido Zuccon**

Guido is a lecturer within the School of Electrical Engineering and Computing Science at the Queensland University of Technology. His research interests include formal models of search and evaluation methods, in particular applied to health search. Guido has actively contributed to the areas of document ranking, search result diversification, formal models, and evaluation. Guido is the IR Task leader for the CLEF eHealth Evaluation Lab, a shared-task that aims to evaluate systems for consumer health search. He has already presented tutorials at SIGIR 2015, CIKM 2016 and ICTIR 2016 on formal models of search.

Guido received a Ph.D. in Computing Science from the University of Glasgow in 2012. Before joining the Queensland University of Technology as a lecturer in 2014, he was a postdoctoral research fellow at the CSIRO, Australia, working on health search technology at intersection of academic research and applied research with real customers.

**Bevan Koopman**

Bevan is a Research Scientist at the Australian e-Health Research Centre, part of the Commonwealth Science and Industrial Research Organisation (CSIRO), based in Brisbane, Australia. He leads projects dedicated to health search, that is the research and development of novel search engine technology to improve access, retrieval and analysis of different health data. He is working on search engines for electronic patient records and search engines for evidence-based medicine. This work is done at intersection of academic research and applied research with real customers.

He received a PhD from the Queensland University of Technology in 2014 with a thesis titled Semantic search as inference: applications in health informatics and holds a visiting researcher position at that university. He also spent a number of years in industry and will therefore provide both an academic and commercial expertise to this tutorial.

### B TYPE OF SUPPORT MATERIALS TO BE SUPPLIED TO ATTENDEES

A selected list of references is presented at the end of this proposal. In addition to those, the materials will consist of (1) slides and the corresponding hand-outs, (2) an annotated bibliography of works in health search, (3) a virtual machine distribution and/or Git repository containing software and example programs that will be demonstrated in the hands-on session.

Participants will require WiFi access and power for the practical sessions.

### REFERENCES
