Scithon™
An evaluation framework for assessing research productivity tools

R. Wu, V. Stauber, V. Botev, J. Elosua, K. Marinov, A. Brede, M. Ritola
Iris.ai
Norway
{ronin, valentin, victor, jacobo, kaloyan.marinov, anita, maria}@iris.ai

Abstract

There is a current scarcity of tested methods to evaluate the performance of artificial intelligence-based science discovery tools. *Iris.ai*, an international start-up developing text understanding technology and products, has developed a novel framework for performing such evaluation tasks. The framework, organized around live events, involves a systematic and cross-disciplinary comparison that focuses on productivity gains and takes into account user engagement. Under this format, referred to as *Scithon™*, event participants are asked to address, in a compressed time frame, the early stages of a research challenge put forth by a third party. Submitted results are then evaluated externally by domain experts. The logged data, including user engagement with the system, is compared against the outcome of the *Scithon™*. In this paper, we present in detail the full mechanics of the *Scithon™* and the results obtained from a series of *Scithon™* competitions run since 2016, where the presented framework is used to evaluate the productivity gains of *Iris.ai*'s own intelligent research assistant. Initial findings show that, compared to conventional evaluation frameworks for search engines, *Scithon™* is a suitable platform for benchmarking intelligent research assistants and is able to identify advantages and disadvantages of such systems in deeper detail and complexity. Iris.ai provides the usage of the platform under an *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License*, which means we welcome the community to freely adopt its name and format with an appropriate acknowledgement to this paper and its authors.

Keywords: Artificial Intelligence, Intelligent Research Assistants, Performance Evaluation

1. Introduction

Mankind is stepping into a new era of artificial intelligence (AI). Everyday, we rely on AI-powered intelligent personal assistants (IPAs) for completing various tasks in our day-to-day lives. Since the 2010s, methods for evaluating IPA systems based on different metrics, such as the accuracy of voice recognition (Assefi et al., 2015) and the cognition workload on users (Strayer et al., 2017), have been discussed by various research groups. However, for scientific research, adequate evaluation methods for novel and AI-based intelligent research assistants (IRA), which help researchers navigate through a tremendous amount of literature, are still scarce. To benchmark the efficiency and accuracy of available IRA systems, *Iris.ai* have launched and hosted a series of *Scithon™* competitions[1,2] across Europe since 2016. Anyone interested in the format of the *Scithon™* is welcome to adopt it and develop it further following its licence regulations. In this paper, we present the *Scithon™* competition concept and methodology and the first statistical results as part of our ongoing efforts to objectively assess the productivity of an IRA system.

In Section 2., we highlight some literature works on the evaluation methods applied to existing research assistant tools and discuss our motivation behind the design of the *Scithon™*. In Section 3., we describe the general concept and process of *Scithon™* and present the collected statistical data in Section 4. A discussion on the outcome of *Scithon™* is presented in Section 5. We conclude this work and lay down our future perspectives in Section 6.

2. Background and Related Work

The development of IRA systems is still at its very early stage. There are lots of research assistant tools, such as *ArXiv*[3], *Semantic Scholar*[4], *Google Scholar*[5], etc., most of which, as the current date, provide basic functionality acting mainly as a search engine built on top of scientific publication databases. Lots of effort has been invested in the development of IRA systems, for example, the citation-based recommender system recently implemented in the COnnecting REpositories (CORE[6], Knoth and Zgrahal (2012)) system (Knoth et al., 2017)), and *Iris.ai*'s own IRA system for which the Word Importance-based Similarity of Documents Metric (WISDM) was developed to increase efficiency by preserving accuracy (Botev et al., 2017). As literature search is a crucial yet tedious process in scientific research, an efficient IRA system can help researchers gain a significant amount of time to focus on the innovative process more in depth. The question is: *how do we judge if the design of an IRA system fits well to the needs of researchers?*

Various groups of researchers have conducted experiments that compare different IRA tools based on the coverage of the database, recall rate, precision, and importance [7] as the main criteria. *Google Scholar* is perhaps one of the most widely used tools for searching scientific publications. In the domain of medical science, researchers have investigated its functionality and compared it with other available tools. A direct comparison of results returned from ten searches performed on both

[1] Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

[2] For more information, please refer to: https://iris.ai/scithons/

[3] https://arxiv.org

[4] https://www.semanticscholar.org

[5] https://scholar.google.com

[6] https://core.ac.uk/

[7] citation counts of search results

Google Scholar and *PubMed* (NCBI Resource Coordinators, 2017) shows that, although *Google Scholar* returns ~ 4 % more relevant results, it retrieves ~ 13 % of articles that are not published in scientific journals (Shultz, 2007). In another comparison between *Google Scholar* and *PubMed*, Anders and Evans (2010) show that, although *PubMed* and *Google Scholar* have similar recall rates, *PubMed* shows higher precision than *Google Scholar*. A similar conclusion on the recall rate and precision has also been drawn in Boeker et al. (2013).

For a broader comparison, another search experiment has been conducted on four search engines, *PubMed*, *Science Direct*[8], *Google Scholar* and *Iranian National Medical Digital Library*[9]. From this experiment, Samadzadeh et al. (2013) show that *PubMed*, *Science Direct* and *Google Scholar* stand out in the recall rate, precision and importance respectively. In a detailed case study on two commonly used tools in medical science, *Web of Science (WOS)*[10], provided by Thomson Reuters, and *Scopus*[11], developed by Elsevier, Aghaei Chadegani et al. (2013) have taken into account various factors, such as recall rate, importance, license-cost and author-profile tracing, in the comparison. It has been demonstrated that *WOS* has stronger chronicle coverage in its database while *Scopus* covers a broader range of journals, although those sometimes can be of lower impact and limited to recent articles only.

One noteworthy fact is that all the above comparisons are domain-specific. Cross-disciplinary comparisons between available IRA systems are only a handful to this date, and most of which focus only on one aspect of functionality (de Winter et al., 2014; Harzing and Alakangas, 2016). As an ideal IRA system is designed to “assist” researchers, upon evaluating such a system, it is crucial to consider: (1) how the system engages its users, and (2) whether the system design is aligned the underlying research process. Moreover, in the new era of AI-powered IRA systems, searches of scientific publications will not be limited to simple textual inputs and outputs. Long list of sorted results may very well be replaced by more visually appealing systems with direct guidance by the AI. Therefore, the traditional methods introduced to compare simple IRA systems may be insufficient to benchmark all the required properties. In order to benchmark the functionality of different IRA systems with metrics more closely related to the researchers and their research processes, we propose the *Scithon™*, a systematic and cross-disciplinary comparison platform that focuses on the productivity gains and takes into account the user engagement.

3. Methodology

3.1. *Scithon™* Overview

Scithon™, or Science-Hackathon (Briscoe, 2014), is a platform designed for evaluating and comparing different IRA systems. It allows assessment of research assistant tools in a competition format where various key-factors, such as the required time for perceiving a problem, finding relevant literature, summarizing results, etc., are all taken into consideration. In contrast with classical hackathons, participants of *Scithon™* are challenged to “hack” a given scientific problem. As the main theme of the challenge is flexible, it provides an ideal platform for cross-disciplinary comparison of IRA systems.

The *Scithon™* competition process is inspired by the systematic literature mapping process (Petersen et al., 2008), which researchers adopt as the initial step to structurally approach a given scientific problem. During a *Scithon™*, interdisciplinary researchers of various professional backgrounds are given a pre-defined research problem and participate in teams to compete with each other within a limited amount of time (usually around eight hours). At the end of the competition, each team is required to summarize their discovery on the given research problem, based on the evidence found in scientific publications, and submit a report to an external jury panel. The winning team is thus rewarded for a report judged by the jury the best in its qualitative and quantitative values to the given research problem.

3.2. Teams

Typically, each *Scithon™* team consists of four to six members. To ensure a balanced team composition, the participants are pre-classified based on their professional backgrounds, experiences, and their degrees of higher education. During the team-assignment, it is ensured that participants of high qualifications are evenly distributed to teams.

3.3. Competition Format

In a typical *Scithon™* competition, participants follow the systematic process laid out in Figure 1. Before the competition, a research problem is prepared by an external party (the challenge provider), usually an industrial, academic and/or governmental entity (Step (1) in Figure 1). After being briefed during the competition, each team spends about eight hours highlighting important research questions for the given research problem and searching for relevant scientific publications. Each team is guided through the systematic mapping process with a provided template (see Table 1 in Appendix A) that involves identifying pertinent research questions, searching relevant scientific publications, screening and mapping scientific publications to the research questions, as indicated in Steps (2) to (4) in Figure 1, respectively.

During the systematic mapping process, all teams are required to conduct search activities on designated team computers equipped with key-logger software, which tracks their discovery activities. At the end of the systematic mapping process, each team is required to submit a report that summarizes the findings following the provided template given in Table 1 in Appendix A. To ensure fairness and transparency of the evaluation process of *Scithon™*, the challenge provider is asked to form an external expert jury panel that assesses the results and se-

[8] https://www.sciencedirect.com/

[9] https://librarytechnology.org/library/64586

[10] https://webofknowledge.com/

[11] https://www.scopus.com/

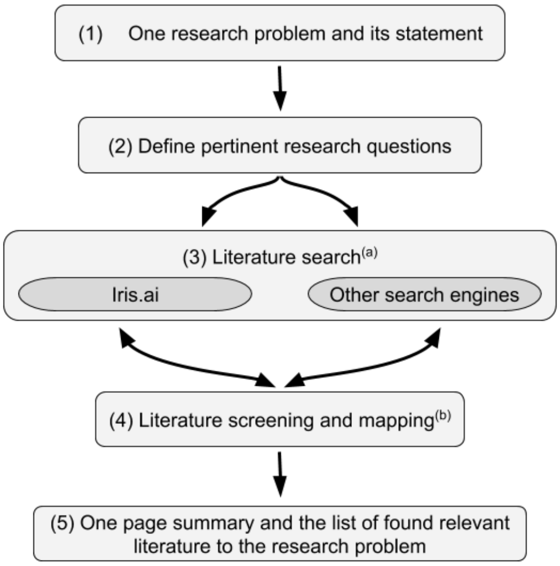
(1) One research problem and its statement
(2) Define pertinent research questions
(3) Literature search(a)
Iris.ai
Other search engines
(4) Literature screening and mapping(b)
(5) One page summary and the list of found relevant literature to the research problem

Figure 1: This flowchart describes the process during a typical *Scithon™* competition. Participants would follow the indicated steps to systematically map scientific publications into research questions, defined by each team in Step (2). The final results produced in Step (5) are submitted to the jury by each team on the provided *Scithon™* template (see Table 1 in Appendix A). [*a*] All steps are monitored by a key-logger that records user-engagement with the provided tools. [*b*] Steps (2) and (4) are guided by the same provided template (Table 1 in Appendix A).

lects the winning team, based on various qualitative and quantitative evaluation criteria (see provided template in Table 3 of Appendix A).

3.4. *Scithon™* Results Evaluation

The result evaluation process consists of two phases: (1) the jury-evaluation phase and (2) the log-analysis phase. The jury-evaluation phase starts with a briefing to the expert jury panel on the metric template (see Table 3 in Appendix A). The purpose is for the jury to know the intention behind each evaluation criterion and adapt it to the current research problem. The briefing usually happens at the beginning of the *Scithon™*. After the competition, participants submit their final reports and the jury panel starts the evaluation. The submitted reports are anonymized to ensure fairness and transparency. Every jury member first individually judges each anonymous report. Final scores are then determined based on the consensus of all jury members. This process typically takes around one to two hours. After scores are assigned, the winner is announced and that concludes the *Scithon™* competition.

To objectively assess the productivity gains from an IRA system, a post-competition log-analysis phase is needed. During this phase the logged data gathered from designated computers is processed. The raw data gathered consist of key-logs, which contain teams' exact search terms and problem statements, screenshots taken every five seconds, and browser histories. This phase is offline, *i.e.* outside of the *Scithon™* competition, and usually takes a few days. The gathered information provides quantified properties of user engagement that may influence productivity gains.

We believe that the proposed system constitutes a controlled and efficient environment for assessing IRA systems, and that it produces objective and transparent evaluation reports (see Table 4 in Appendix A for *Scithon™* results). In the following section, we present a case study on applying the *Scithon™* framework to *Iris.ai*'s own IRA system followed by a results discussion and conclusion.

4. Results - case study on *Iris.ai*'s IRA

For a case study, we tested *Iris.ai*'s own IRA system in a series of *Scithon™* competitions since September 2016. The goal of these competitions was for each team to digest and map relevant scientific publications, as accurately as possible, to the research questions defined around the given research problem. Within the given time limit, participating teams compete by using the IRA system designed and developed at *Iris.ai* and/or any other market-ready research assistant tools *e.g.* *ArXiv*, *Google Scholar*, *Science Direct*, *OnePetro*[12], etc. (Step (5) in Figure 1). To this date, there have been six public and three private *Scithon™* competitions. As the data from the private competitions is confidential, we focus our discussion on the results from the public *Scithon™* competitions.

To gauge user engagement, the parameters extracted during the log-analysis phase include: the percentage of papers found using *Iris.ai*'s IRA (PP), the time spent with *Iris.ai*'s IRA (T), and the number of maps created by *Iris.ai*'s IRA (MC), which corresponds to number of search queries for conventional search engines. These parameters are then correlated with the total jury percentile score (JS) given to each team. The relationships between total jury percentile scores and the three extracted parameters, PP, T, and MC, are presented from left to right panels in Figure 2. The different colors represent results from different *Scithon™* competitions and full data from those events is available in Table 4 in Appendix A. The Pearson's r values estimated from the data are 0.63, 0.42, and 0.5, respectively with $p < 5\,\%$.

5. Discussion

As Figure 2 shows, in general, total jury percentile scores are positively related to the extracted parameters, PP, T, and MC. This implies that participants gain fuller scopes of the given research problems, both quantitatively and qualitatively, from deeper engagement with *Iris.ai*'s IRA system. These results also indicate that *Iris.ai*'s IRA is capable of boosting the efficiency and refining the outcome of systematic mapping processes for researchers.

The positive relationships between total jury scores and T and MC (middle and right panels of Figure 2) indicate that the more participants explore with *Iris.ai*'s

[12] https://www.onepetro.org/

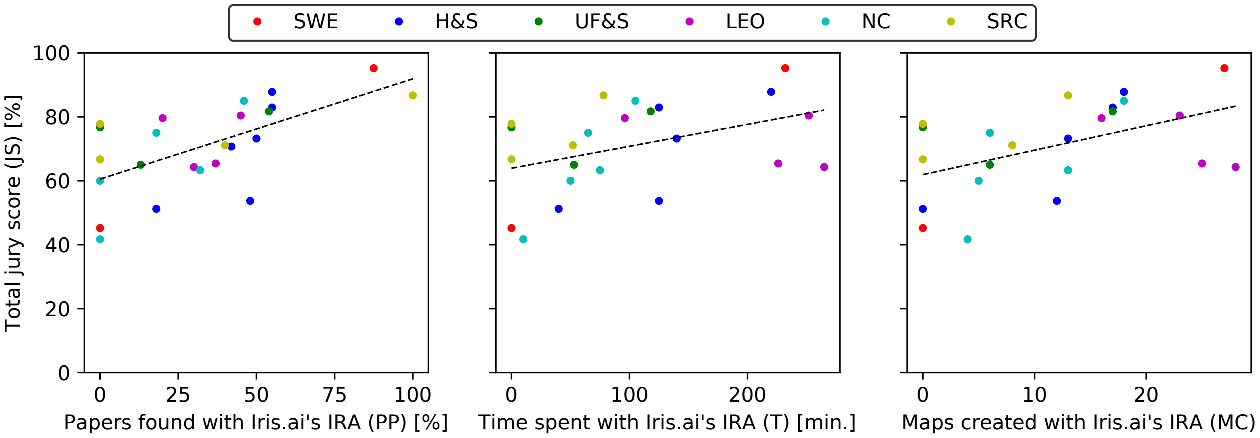

Figure 2: Results from our case study of *Iris.ai*'s own IRA (see Section 4. and 5. for further details). For raw data and legend abbreviations see Table 4.

IRA (measured by both T, MC), the better they can formulate the scope of the given research problem, and thus the better their findings and summary reports are. Although there is clearly a positive relationship between total jury percentile score and the percentage of papers found with *Iris.ai*'s IRA (indicated by PP in the left panel of Figure 2), from the Pearson's r value (0.63), it may be questioned whether there is sufficient statistical evidence to undoubtedly and objectively conclude and quantify the productivity gains by the tool. Gathering more data points could strengthen the conclusion, but even the current results could be used as a base hypothesis for the potential of the IRA system.

Unfortunately, as with all experiments conducted with direct user engagement, several subjective defects should be taken into account on an individual *Scithon™* basis. Such defects include the integrated team capacity, attitude and motivation. The chief motivations for the *Scithon™* participants are the reward and their interest in the given scientific challenge. In some cases, a team may suffer from a significant knowledge gap to the other teams that inevitably leads to inferior results regardless of the time spent with the IRA. In other cases, during some private *Scithon™* competitions, we observe that the motivation can sometimes be hindered by pre-established group-dynamics among colleagues which can thus influence the final results. To be able to mitigate the impact of such defects or to pinpoint their compromised nature, the *Scithon™* setup should benefit from logging team engagement and ensuring the neutrality of external jury panel. In future *Scithon™* competitions, additional markers, such as for example the Rhythm Badge technology developed at the Massachusetts Institute of Technology Lederman2018, will be added to detect and flag such cases.

Despite the discussed threats to validity, our current results show that, in order to acquire the most accurate and objective results, it is recommended that participants focus on problem solving with full dedication. Our results also demonstrate that *Scithon™* competitions allow us to benchmark *Iris.ai*'s IRA against other market-ready IRA systems to understand better its advantages and disadvantages. A clear insight drawn using *Scithon™* for *Iris.ai* was that teams not familiar with the field experience much higher productivity gains compared to other teams more familiar with the problem area. Moreover, a certain amount of time spent with the tool is necessary for productivity gains to start to appear. Such qualities cannot be easily measured by conventional metrics but can give insightful information about the strengths of the IRA system.

As our team aims to introduce novel features into *Iris.ai*'s IRA for its perfection, we are certainly interested in the one-on-one benchmarks of its performance against other market-ready IRA systems. Such work, however, is under development and is beyond the scope of this paper. Here we aim to investigate the efficiency of *Scithon™*for getting insights and measuring performance focusing on the usage metrics of one specific tool only. Thus we discuss whether teams can perform better with *Iris.ai*'s IRA and compare the results with the efficiency of their research process before being introduced to the tool. Further more, we want to avoid restraining them from using tools that are familiar to their work process outside the usage of *Iris.ai*'s IRA, because in many cases the introduction of a new tool could be supplementary and not necessarily a replacement of an existing tool in the teams' research process. In a future work, we will discuss the performance of *Iris.ai*'s IRA alongside other specific IRA systems in an one-on-one comparison.

6. Conclusion and future perspectives

In this paper, we introduce the *Scithon™*, a systematic and cross-disciplinary comparison platform for intelligent research assistant (IRA) systems, that evaluates productivity gains and takes into account user engagement. In *Scithon™*, participants are challenged to "hack" a given scientific problem. As the main theme of the challenge is flexible, it provides an ideal platform for cross-disciplinary comparison of intelligent research assistant (IRA) systems.

Based on the case study on *Iris.ai*'s own IRA system, we conclude:

1. *Scithon™* is a suitable platform for benchmarking IRA systems and it can be used to measure end-to-end user productivity gains compared to more conventional metrics for search engine evaluation. As comes with an *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License*, we welcome the community to freely adopt its name and format with an appropriate acknowledgement to this paper and its authors.

2. With the help of *Scithon™*, we identified the strengths of *Iris.ai*'s IRA and possible target customer groups

3. The framework was able to unearth complex properties of *Iris.ai*'s IRA. These include that, compared with users who are familiar with the given research problem, users that are not familiar with the field experience much higher productivity gains.

4. Additional experiments with a refined keylogger software need to be conducted with the *Scithon™* framework to understand how to prevent subjective threats to validity.

In the future *Iris.ai* will continue developing its IRA systems, and this will require searching for evaluation frameworks that include measuring user engagement and provide a controlled environment with objective evaluation criteria. We believe assisting tools are meant to assist humans and should be measured against how well they manage to achieve their ultimate goal.

7. Bibliographical References

Aghaei Chadegani, A., Salehi, H., Md Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., and Ale Ebrahim, N. (2013). A comparison between two main academic literature collections: Web of science and scopus databases. *Asian Social Science*, 9(5):18–26.

Anders, M. E. and Evans, D. P. (2010). Comparison of PubMed and Google Scholar literature searches. *Respiratory care*, 55(5):578–83, may.

Assefi, M., Liu, G., Wittie, M. P., and Izurieta, C. (2015). An experimental evaluation of apple siri and google speech recognition. In *Proccedings of the 2015 ISCA SEDE*.

Boeker, M., Vach, W., and Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Medical Research Methodology*, 13(1):131, dec.

Botev, V., Marinov, K., and Schäfer, F. (2017). Word importance-based similarity of documents metric (WISDM). In *Proceedings of the 6th International Workshop on Mining Scientific Publications - WOSP 2017*, pages 17–23, New York, New York, USA. ACM Press.

Briscoe, G. (2014). Digital innovation: The hackathon phenomenon. *Creativeworks London*.

de Winter, J. C. F., Zadpoor, A. A., and Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, 98(2):1547–1565, feb.

Harzing, A.-W. and Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, feb.

Knoth, P. and Zgrahal, Z. (2012). Core: Three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), nov.

Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N., and Bayer, V. (2017). Towards effective research recommender systems for repositories. pages 1–6, may.

NCBI Resource Coordinators. (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 45(D1):D12–D17, jan.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK. BCS Learning & Development Ltd.

Samadzadeh, G., Rigi, T., and Ganjali, A. (2013). Comparison of Four Search Engines and their efficacy With Emphasis on Literature Research in Addiction (Prevention and Treatment). *Int J High Risk Behav Addict*, 1(4):166–171.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association : JMLA*, 95(4):442–445, oct.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., and Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 71(2):93–110.

Appendices

A The *Scithon™* templates and data

An example of a *Scithon™* user template is given in Table 1, with a glossary in Table 2. The evaluation criteria for the external judges are given in Table 3. Finally, data collected from various events organized by *Iris.ai*, which is plotted in Figure 2, is given in Tables 4.

Scithon™evaluation matrix

MISSION
Science hackathons, or "Scithons™", are focused on addressing scientific research challenges.

Teams of researchers are asked to:

Map relevant scientific articles to solve a particular, pre-determined problem.
By mapping we mean finding relevant scientific papers, articles, reports, etc. and categorizing them using various approaches. I.e. providing an overview.

Summarize the key findings by skimming through the categories and papers.
Describing these findings and how well supported by research they are.
Drawing preliminary conclusions based on relevant research trends, including the latest ones.
If possible, suggest well-motivated future research activities.

PROCESS
Each research team identifies relevant research papers classifying them in a structured manner (see below).

As a starting point teams save the template to their computers.
At the end of the Scithon™ all teams will send the filled out templates to the organizers.

Teams will be provided with computers to optimize process monitoring ex post.
Activities throughout the Scithon™will be recorded for analysis after the event.

TEAM TEMPLATE

TEAM NAME:

RESEARCH QUESTION:

PROBLEM STATEMENT:

TOPIC AREA **topic area title #1**

Research type facet	Contribution type facet	Paper title	Paper conclusions	Paper ID / URL
What aspect of research are papers about?	**What specific element do papers relate to?**	**Please provide the full paper title**	**Please provide the paper's key conclusion**	**Please provide a unique paper ID or URL**

SUMMARY OF FINDINGS

Table 1: The *Scithon™* user template

APPENDIX
GLOSSARY

Category	**Description**
Systematic review	A systematic review summarizes the results of available carefully designed health-care studies and provides a high level of evidence on a transparent, a-priori driven approach.
Literature review	Narrative reviews tend to be mainly descriptive, do not involve a systematic search of the literature, and thereby often focus on a subset of studies in an area chosen based on availability or author selection.
Validation research	Techniques investigated are novel and have not yet been implemented in practice. Techniques used are for example experiments, i.e., work done in the lab.
Evaluation research	Techniques are implemented in practice and an evaluation of the technique is conducted. That means, it is shown how the technique is implemented in practice (solution implementation) and what are the consequences of the implementation in terms of benefits and drawbacks (implementation evaluation). This also includes to identify problems in industry.
Evaluation research	A solution for a problem is proposed, the solution can be either novel or a significant extension of an existing technique. The potential benefits and the applicability of the solution is shown by a small example or a good line of argumentation.
Philosophical papers	These papers sketch a new way of looking at existing things by structuring the field in form of taxonomy or conceptual framework.
Opinion papers	These papers express the personal opinion of somebody whether a certain technique is good or bad, or how things should been done. They do not rely on related work and research methodologies.
Experience papers	Experience papers explain on what and how something has been done in practice. It has to be the personal experience of the author.

Source:
`http://www.bcs.org/upload/pdf/ewic_ea08_paper8.pdf`
Lockwood, Craig; Sfetcu, Raluca; Oh, Eui Geum. Synthesizing Quantitative Evidence.
Joanna Briggs Institute (JBI), 2011.
`http://consumers.cochrane.org/what-systematic-review`
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024725/`

Table 2: Glossary for the *Scithon™* user template (Table 1)

Qualitative
Overview - How well did the team manage to explore the overview of the problem? Are there essential parts that are missing? Are there essentials parts that they got wrong? Do they cover all main approaches? Do they address all main challenges? (Max Score 10)
Findings - Are there interesting findings in the results? If yes what is the quality of the findings - are they well described, are they well supported by research, etc. (Max Score 10)
Conclusion - Is the conclusion following latest trends in research? Is it well supported by research? Does suggest future research activities? To what extend the proposed future activities follow the trends and are well-motivated? (Max Score 10)
Quantitative
Number of related papers to the field (10 are required) (Max Score 10)
Number of identified relevant approaches (Max Score 10)
Number of “spot on” papers (surprisingly interesting) (Max Score 10)

Table 3: The *Scithon™* evaluation criteria template for the jury.

	Team composition	Jury total score	papers from Iris.ai IRA	Time using Iris.ai IRA	maps generated
Swerea SICOMP (SWE)	*Can a reusable rocket be built exclusively with composite materials?*				
Team B	Researchers and university students, without direct field specialists	95%	88%	3:52:00	27
Team A	Researchers and university students, without direct field specialists	45%	0%	0:00:00	0
Hotus & Skhole (H&S)	*What health care interventions can affect life-style factors to prevent a risk for noncommunicable diseases?*				
Team E	Masters (political science, health science, health & wellbeing)	88%	55%	3:40:00	18
Team F	Masters and higher (information management, public health and rehabilitation)	83%	55%	2:05:00	17
Team C	Bachelor (sports and physiotherapy)	73%	50%	2:20:00	13
Team A	PhD (health sciences, physiotherapist, nurse, student majoring in automation)	71%	42%	N/A	N/A[13]
Team G	Masters (economics), Bachelor (computer science, education, nurse)	54%	48%	2:05:00	12
Team B	Bachelor (physiotherapy)	51%	18%	0:40:00	0
Team D	Bachelor (health & wellbeing, physiotherapist)	49%	44%	5:00:00	14
Uniklinik Freiburg & Stryker (UF&S)	*How can Augmented Reality be deployed to improve surgical education?*				
Team B	Cross-disciplinary researchers and doctors, including direct field specialists	82%	54%	1:58:00	17
Team C	Cross-disciplinary researchers and doctors, including direct field specialists	77%	0%	0:00:00	0
Team A	Cross-disciplinary researchers and doctors, including direct field specialists	65%	13%	0:53:00	6
Leo Pharma (LEO)	*Identifying new pathways and targets to treat eczema*				
Team Y	Bio-informatics; Masters and Bachelors	80%	45%	4:12:00	23
Team Z	Bio-informatics, computer-science, chemistry; Masters	79%	20%	1:36:00	16
Team X	Computer science and medicine, Masters, Bachelors, PhD	70%	37%	3:46:00	25
Team T	Bio informatics and biology, Master students final year	64%	30%	4:25:00	28
Norwegian Customs (NC)	*Automatic support for selection of objects viable for control for the customs*				
Team 1	Software professionals, Masters and Bachelors	85%	46%	1:45:00	18
Team 3	Masters final year	75%	18%	1:05:00	6
Team 5	Software professionals, PhD and Masters	63%	32%	1:15:00	13
Team 4	PhDs	60%	0%	0:50:00	5
Team 2	Business people, Professionals in logistics, Almost no software background	42%	0%	0:10:00	4
Stockholm Resilience Centre (SRC)	*Q1: In which time frame can global urbanization become biosphere positive?*				
Team 2	Researchers	71%	40%	0:52:00	8
Team 1	PhD and business people	67%	0%	0:00:00	0
	Q2: Can biosphere positive fibres and textiles clothe the world?				
Team 4	PhD, researchers, business people	87%	100%	1:18:00	13
Team 3	PhD, researchers, business people	78%	0%	0:00:00	0

Table 4: Data collected from various *Scithon™* events.