Report on the 1st ACM SIGIR/SIGKDD Africa School on Machine Learning for Data Mining and Search

Ben Carterette
Spotify / University of Delaware
carteret@udel.edu

Hussein Suleman
University of Cape Town
hussein@cs.uct.ac.za

Douglas W. Oard
University of Maryland
oard@umd.edu

Abstract

We report on the inception, organization, and activities of the 1st ACM SIGIR/SIGKDD Africa School on Machine Learning for Data Mining and Search, which took place at the University of Cape Town in South Africa January 14–18, 2019.

1 Introduction

In January of 2019, ACM SIGIR and ACM SIGKDD co-sponsored the 1st SIGIR/SIGKDD Africa School on Machine Learning for Data Mining and Search,[1] which took place at the University of Cape Town in South Africa. The event was conceived by SIGIR members with the aims of increasing opportunities in research from traditionally underserved communities, growing the IR and data mining communities in sub-Saharan Africa, and expanding the horizons of IR and data mining research.

This is a report on the inception, organization, and activities of the event. Section 2 gives context on how the event came to be. In Section 3 we describe the organizational decisions and activities including committees, funding, calls for participation, and scientific program. Section 4 details the five days of the event, including lectures, labs, and other sessions. In Section 5 we detail the next steps for the initiative.

2 From inception to inauguration

The inception of the event dates to the SIGIR conference in Tokyo in 2017. That year, the SIGIR Executive Committee organized the first SIGIR Diversity, Equity, and Inclusion (DEI) lunch for conference attendees. Several speakers were invited to share their thoughts on these topics. As SIGIR 2017 was also the 40th SIGIR, there was also an informal theme of thinking about the future of research in IR. To that end, the organizers had set up a whiteboard with a large map of the world, and asked attendees to write where they wanted SIGIR to go. As Doug Oard (Universtity of Maryland) noted in his DEI lunch talk, no one wrote anything specific on the entire continent of Africa. (See Fig. 1.) The question is, why? Are we as a community missing out on opportunities there? After all, over 1.2 billion people live in Africa, nearly as many as in all of Europe and North America combined,

As the SIG Governing Board at the time had been making a push for ACM-sponsored summer schools, a school seemed like a natural way to start. A push began from the SIGIR Executive Committee to start a summer school on IR in Africa. We identified SIGKDD as a possible partner and contacted Jian Pei, chair of that SIG, about the idea. This led to SIGKDD committing an equal amount to sponsoring the event.

Hussein Suleman at the University of Cape Town, South Africa, came on as an organizer around this time. Much of the early stages of planning revolved around establishing the school as an official ACM event, which required recruiting an organizing team, finding a location, and drafting a budget.

[1] http://www.sigir.org/afirm2019

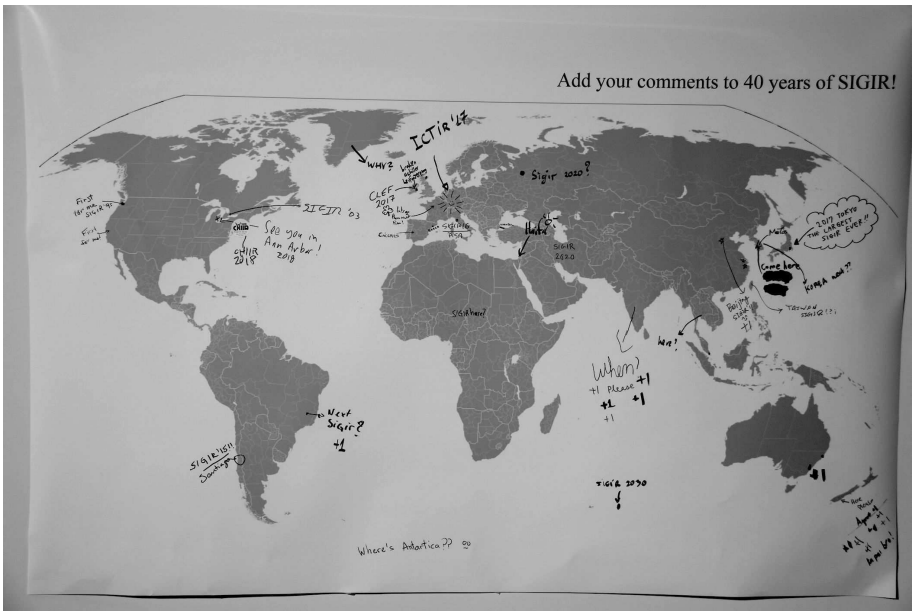

Figure 1: The world map at the 40th ACM SIGIR conference in Tokyo in 2017.

3 Organization

Organization formally kicked off with a meeting of interested parties at the SIGIR 2018 conference in Ann Arbor, Michigan. During this meeting we decided on a name for the event, its timing, and its general structure. We also decided on an audience to target: junior faculty at institutions in sub-Saharan Africa. We discuss this more in Section 3.5.2 below.

3.1 Organizing team

The organizing team consisted of the following committees and individuals:

1. Steering committee
   - Ben Carterette, Chair (Spotify / University of Delaware)
   - Diane Kelly, SIGIR Representative (University of Tennessee Knoxville)
   - Jian Pei, SIGKDD Representative (Simon Fraser University, British Columbia)
   - Abdiganyi Diriye (IBM Kenya)
   - Ricardo Baeza-Yates (NTENT / Northeastern University)
2. General chairs and local support
   - Hussein Suleman, co-Chair (University of Cape Town)
   - Ben Carterette, co-Chair (Spotify / University of Delaware)
   - Catherine Chavula (University of Cape Town)
   - Zola Mahlaza (University of Cape Town)
   - Kelvin Meyer (University of Cape Town)
   - Jivashi Nagar (University of Cape Town)
   - Joseph Telemala (University of Cape Town)
3. Program committee
   - Doug Oard, Chair (University of Maryland)

- Ricardo Baeza-Yates (NTENT / Northeastern University)
- Praveen Chandar (Spotify)
- Charles L. A. Clarke (University of Waterloo, Canada)
- Rosie Jones (Spotify)
- Jaap Kamps (University of Amsterdam)
- Bhaskar Mitra (Microsoft)
- Jian Pei (Simon Fraser University)
- Hussein Suleman (University of Cape Town)
- Johanne Trippas (RMIT University)
- Emine Yilmaz (University College of London)

We note here the general paucity of African and absence of Asian representation on the organizing team. We address this in more detail below.

3.2 Event timing

As with many events, finding a good set of dates proved to be challenging. Because we envisioned this as a "summer school" (similar to other schools like RuSSIR, ESSIR, ASSIA, etc), we looked at dates in the southern hemisphere summer, which spans from December to early March. December and early January were inadvisable due to proximity to major holidays. February marks the start of the academic year in most South African universities. Thus mid-January seemed to be the best compromise, despite its proximity to the SIGIR deadline and some awkwardness with academic calendars in the US and Europe (which is where we thought the greatest number of instructors might come from).

3.3 Announcing the event

The event was first announced at the SIGIR 2018 conference in Ann Arbor, Michigan, during the Diversity, Equity, and Inclusion lunch on the second day of the main conference. A second announcement was made during the SIGIR business meeting on the last day. The steering committee chair announced the location and important dates for potential instructors to be aware of.

The event was announced again at the KDD 2018 conference in London, during the SIGKDD business meeting, by the steering committee chair.

3.4 Event funding

As mentioned above, ACM SIGIR and ACM SIGKDD funded the event with USD20,000 each. Additional support for local arrangements came from the University of Cape Town. This support meant that the majority of the ACM funds could be used to support travel for participants.

We also instituted a registration fee of USD200, though we obtained little revenue this way. Most participants were granted fee waivers.

Though we reached out to contacts in industry for financial support, we were not able to raise any additional funds this way.

3.5 Participation

We issued two open calls for participation, one for instructors and course proposals, and one for students.

3.5.1 Instructor participation

The call for courses went out shortly after SIGIR 2018, to SIG-IRlist and KDNuggets, as well as Twitter and other open distribution channels. The call requested proposals that included lectures and

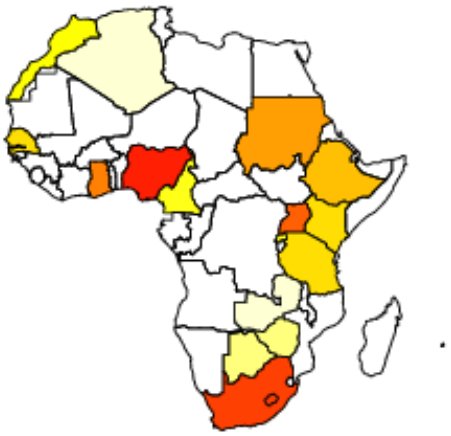

Figure 2: Map of Africa indicating origin of applications for travel support. The two countries that provided the largest number of applications were South Africa and Nigeria.

lab activities from experts in the fields of search and data mining, and indicated that instructor teams' travel would be supported up to USD2,000. We used an EasyChair site to accept proposals

In total we received 15 course proposals, which we considered a success. These were reviewed by the program committee consisting of the individuals listed above. Reviews were taken into consideration in forming the tentative program from accepted proposals; the final program was decided on by the steering committee and the program committee chair. We invited some instructors to take on topics that we viewed as necessary but that had not been proposed.

The final list of courses accepted through proposal and invitation were:

1. Introduction to Information Retrieval, Grace Yang (Georgetown University, USA)
2. Multilingual Information Retrieval, Doug Oard (University of Maryland, USA)
3. Information Retrieval Systems, Guido Zuccon (University of Queensland, Australia)
4. IR for Development, Hussein Suleman (University of Cape Town, South Africa)
5. Evaluation, Nicola Ferro (University of Padua, Italy), Maria Maistro (University of Copenhagen, Denmark), and Ian Soboroff (NIST, USA)
6. Text classification, Fabrizio Sebastiana and Alejandro Moreo (CNR-ISTI, Italy)
7. Algorithmic Bias, Josh Kroll (University of California, Berkeley, USA)
8. Deep Learning for IR, Bhaskar Mitra, Nick Craswell, Daniel Campos, and Emine Yilmaz (Microsoft, University College London)
9. Learning from User Interactions, Rishabh Mehrotra (Spotify, UK)
10. Health Search, Guido Zuccon (University of Queensland, Australia)
11. Music Recommendation at Spotify, Ben Carterette (Spotify / University of Delaware, USA)

Unfortunately Ian Soboroff and Rishabh Mehrotra were unable to attend the event. The latter two lectures by Guido and Ben were last-minute additions made during the event itself.

3.5.2 Student participation

As mentioned above, we targeted junior faculty at institutions in sub-Saharan Africa as our primary audience. We had several reasons for this: first, other schools taking place in sub-Saharan Africa have

been targeting undergraduates (the Deep Learning Indaba) or Ph.D. students (the Machine Learning Summer School series); an event for faculty would be distinct and unique. Second, faculty members as a group are the most well-positioned to have impact on the growth of a community, given their influence on trainees. Third, faculty members tend to be excellent students, not just in terms of understanding material, but also in the sense that they will know best how to evaluate the event and help improve it in the future. Finally, given that many universities are at an early stage in establishing research and graduate programs, the junior faculty are often seen as trailblazers in those processes and can significantly influence emerging research areas.

Calls for participation were issued in two phases after the program had been finalized (so that the call could include information about the courses and instructors). The first call was for applications for travel support. We sent emails to SIG-IRlist as well as to communication channels for African research and development. In response we received over 275 applications. Figure 2 shows the countries in Africa from which we received applications, with deeper shades indicating higher volume of applicants.

Based on the remaining budget after estimating instructor travel, we allocated USD25,000 for supporting participant travel. As we wanted to ensure this would be a truly multi-national event, we tried to allocate awards such that the majority of people receiving support would receive nearly all of what they needed for their travel—it is not common for African institutions to support international travel, so many of them would not be able to attend at all unless their travel was fully supported. We also wanted a diverse event, in terms of nationality, gender, background, and other considerations, and we strove as much as possible to achieve this. It must be noted that we did not accomplish gender diversity to the degree we wished; this is something to work on next time.

We offered support to 36 applicants. Nearly all accepted the support by registering for the event, though in the end many did not attend. Nigerian applicants in particular experienced a high degree of difficulty obtaining visas to travel to South Africa.

A second call, without travel support, was issued closer to the dates of the event. Several more people, mainly from South Africa, registered via this call.

In the end there were 27 participants representing 13 different African countries: Eswatini (formerly Swaziland), Ethiopia, Ghana, Kenya, Malawi, Nigeria, Senegal, South Africa, Sudan, Tanzania, Uganda, Zambia, and Zimbabwe. This turned out to be an ideal number.

4 The event

The majority of the event took place in two locations: the Hlanganani Junction lecture room in the Chancellor Oppenheimer Library at the University of Cape Town Upper Campus, where participants attended lectures, and the Computer Science Senior Lab in the Computer Science building (a short walk from the library), where participants worked on hands-on activities for learning about search and data mining technologies.

Each day was split between lectures and labs, with lectures starting at 9 in the morning and going through about 2 o'clock in the afternoon, with a break for catered lunch in the same room. Lab sessions were two to three hours in the afternoon, adjourning at 5:30.

The full schedule is shown in Table 1.

4.1 Day one

The event kicked off with a keynote entitled "Retrieval as Interaction" by Maarten de Rijke (University of Amsterdam, Netherlands). Maarten spoke about *safety* and *explicability* in the development of online search and recommendation systems. Systems deployed online must be safe in the sense that they perform well, they learn from interactions, they don't privilege any group over others, and they provide diversity or breadth rather than depth; Maarten connected this to his and his colleagues' work on online learning with bandits and other research. Systems deployed online must be explicable in the sense that they can explain how the technology works and how a particular decision was arrived at; Maarten discussed this in the context of his and his colleagues' work on contrastive explanations.

	Mon 14 Jan	Tue 15 Jan	Wed 16 Jan	Thu 17 Jan	Fri 18 Jan
0900	Keynote	IR for Development	Classification 1	Deep learning 1	Health search
1030	IR intro	Evaluation 1	Classification 2	Deep learning 2	Recommendation
1200	Multilingual IR	Lunch	Lunch	Lunch	Lunch
1300	Lunch	Evaluation 2	Algorithmic bias 1	Algorithmic bias 2	Choose-your-lab
1430	IR systems lab	Evaluation lab	Classification lab	Deep learning lab	
1600					Panel
1700					Graduation
1730	End of working day				
1815	Reception				
1900				Gala dinner	

Table 1: Full schedule of lectures, labs, and other sessions over the five days of the event. Materials from lectures and labs are available for download at `http://www.sigir.org/afirm2019`.

Following Maarten, Grace Yang (Georgetown University, USA) presented an introduction to information retrieval and research in the field. She presented an overview of work needed to build an end-to-end text retrieval system: preprocessing and indexing text, retrieval and ranking functions, evaluating effectiveness, and incorporating user feedback. Grace also showed how IR can be understood in relation to other disciplines of CS.

After a short break, Doug Oard (University of Maryland, USA) spoke about multilingual information retrieval. Included in his talk were some of the challenges of retrieval in other languages, in particular encoding characters, identifying terms, stemming, representation of queries and documents across languages, difficulties of retrieval in languages lacking resources, and more. Doug gave some examples from some of his current work on two African languages, Kaswahili and the Somali language.

After lunch, we reconvened in the lab for a brief lecture on tools for IR and labs on Apache Elasticsearch by Guido Zuccon (University of Queensland, Australia). Guido first gave a lecture on "IR in Practice", showing architectures of IR systems and the querying process and giving an overview of tools used to do work on IR in both research and industry. He focused mainly on Elasticsearch and Lucene, as Elasticsearch was the topic of the hands-on lab activities. These activities were structured into seven parts that built on one another (though few particiapnts completed all seven—a recurring theme of the labs): from installing Elasticsearch, to indexing and searching a small TREC collection, to producing runs in TREC format, to accessing term vectors, and so on.

We concluded the day with a welcome reception for all attendees that featured a small poster session in which some participants described their current research.

4.2 Day two

The second day kicked off with a lecture by Hussein Suleman (University of Cape Town) on the topic of Information Retrieval for Development (IR4D) (see Figure 3). The talk started with the question of how we can use IR, data mining, and other technologies to support human and socio-economic development within Africa. Some of the challenges include tools that do not offer good support for local languages, a dearth of quality data that can be used to build tools, the sheer number of languages and the fact that most people know more than one, the difficulty of even defining separate languages in Africa, and wildly varying understanding and ability to use these technologies. Hussein spoke about some of the research work going on in Africa, in his own group and in others', to address these challenges.

The rest of the day centered on evaluation. Nicola Ferro (University of Padua, Italy) and Maria Maistro (University of Copenhagen, Denmark) presented two lectures on many aspects of evaluation, from test collections, to relevance judging, to measurement and metrics, to statistical testing, to online feedback.

In the first lecture, Nicola focused on offline evaluation: test collections, acquiring relevance judgments, evaluation metrics, and statistical hypothesis testing. He gave special attention to IR evaluation campaigns, asking whether there might soon be one located in Africa and centered on problems of

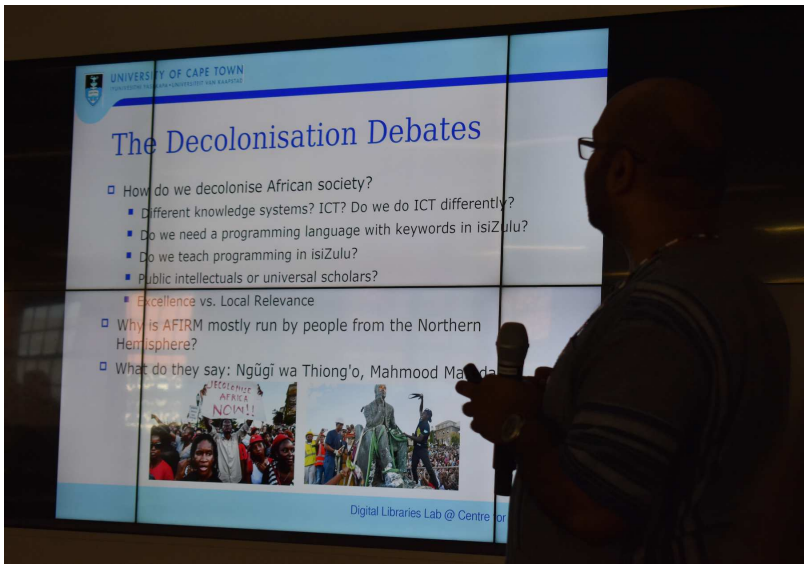

Figure 3: Prof. Hussein Suleman of the University of Cape Town gave a thought-provoking talk on using IR, data mining, and other technologies for human and socio-economic development in Africa.

interest to African researchers and practitioners. In the second lecture, Maria focused on online evaluation: designing controlled experiments with live users, logging implicit and explicit feedback from user interaction, calibrating offline measures with logged data, modeling users based on their logged interactions, and more. She made the important point that systems are becoming increasingly contextual and personal, which makes offline experimentation and evaluation much more difficult.

For the afternoon lab activities, participants learned about using `trec_eval` to evaluate retrieval system outputs in the TREC style. They used Jupyter Notebooks to import evaluation tools and compute standard evaluation metrics, as well as see how to implement new metrics. Most participants again did not progress as far in the exercises as planned, but materials are still available for those who want to learn offline.

4.3 Day three

Day three was mainly about classification. In a two-part lecture, Fabrizio Sebastiani (CNR-ISTI, Italy) presented on text classification and sentiment analysis. The first part focused on defining text classification tasks, representing text in documents for the purposes of learning classification models, and evaluating text classification. In this part Fabrizio provided some links to collections and tools for experimentation. The second part was about the specific tasks of sentiment analysis and opinion mining, noting applications from predicting election results, to sensitivity with advertising, to online reputation management. Sentiment analysis is a particularly interesting task because it is difficult but has many important applications.

After lunch, Josh Kroll (University of California, Berkeley, USA) presented the first part of his two lectures on algorithmic bias. Algorithmic bias is an increasingly important topic, as technologies such as speech recognition, facial recognition, recommendation, etc, gain more and more ground in the public sphere. Josh spoke about different types of bias and how they affect the work of ML and data scientists and engineers.

Participants then went to the lab for activities on text classification led by Alejandro Moreo (CNR-ISTI, Italy). Alejandro walked participants through classification exercises in a Jupyter Notebook using `scikit-learn`. They were shown how to load the 20 Newsgroups data, parse it, compute features,

Figure 4: The event closed with a panel of four participants. From left: Lighton Phiri (University of Zambia), Brian Mwandau (Strathmore University, Kenya), Justine Nakirijja (Makerere University, Uganda), and Reuben Dlamini (University of the Witwatersrand, South Africa).

and train and test an SVM classifier.

4.4 Day four

The fourth day centered around deep learning, led by a team from Microsoft: Bhaskar Mitra, Nick Craswell, Emine Yilmaz (also University College London, UK), and Daniel Campos. The two lectures, given by the first three participants, started with discussion of features used for IR and a discussion of neural networks. This was followed by discussion of the use of vector representations in IR, leading to a brief introduction to term embeddings with word2vec. The next part of the lecture featured an extended presentation of the use of term embeddings in IR to aid in query and document representation. This then transitioned into a discussion of approaches to supervised learning to rank, contrasting with the unsupervised nature of using pre-trained embeddings for representation. Finally, the team brought everything together to discuss supervised training of deep neural networks to perform IR tasks.

After lunch, Josh continued with the second part of his lecture on algorithmic bias. Josh noted that the technologies mentioned above often show biases towards certain population subgroups, leading with an example on how "targeted policing" using machine learning and data science over-targets African Americans. He talked about bias vs. variance, confounding effects, and theoretical results on fairness, as well as aspects of experimental validity that one should consider when trying to design fair ML systems. He concluded with some advice on how to mitigate bias.

Afterwards, Daniel Campos of Microsoft, with Bhaskar Mitra, led lab activities on deep learning, including word2vec and learning-to-rank for IR. Thanks to NOW Publishers, during the lab attendees were given a complementary printed copy of Mitra & Craswell's Foundations and Trends in Information Retrieval book on Deep Learning for IR.

Day four concluded with a dinner for all attendees at Gold Restaurant, which features foods from a variety of African cuisines. Everyone enjoyed performances including singing and dancing by local talent.

4.5 Day five

The fifth day was originally planned to feature lectures on learning from user interactions by Rishabh Mehrotra of Spotify London. Unfortunately, due to visa processing issues, Rishabh was unable to attend. Instead, Guido Zuccon and Ben Carterette filled in with lectures on health search and music recommendation, respectively.

Guido began with an overview of the types of health information that professional providers use and the challenges in using them: for example, noisy data, missing context, synonyms of health-specific terms, and varying quality of information. He then discussed the kinds of people that use health information technology and, in particular for search, how they query. He concluded with a discussion of resources for health IR.

Ben spoke about how Spotify uses machine learning, specifically contextual bandits and counterfactual analysis, to optimize recommendations on Spotify's home screen. He used this as a way to present various aspects of recommendation, from representation to optimization to evaluation and A/B testing.

Instead of the planned labs on interaction, the lab time was used to continue labs from previous days—as noted above, participants were typically not able to finish lab activities in the allotted time. Participants gathered in groups depending on which lab they wanted to continue.

4.5.1 Panel and graduation

We concluded the event with a discussion panel followed by a "graduation" ceremony. The panel, moderated by Hussein Suleman, featured four of our participants: Lighton Phiri, a lecturer at the University of Zambia; Brian Mwandau, a lecturer at Strathmore University in Kenya; Justine Nakirijja, a Ph.D. student at Makerere University in Uganda; and Reuben Dlamini, a senior lecturer at the University of the Witwatersrand in South Africa (Fig. 4).

The panel, titled "Developing Research in IR and DM in Africa", was mainly about AFIRM, eliciting feedback from participants and ideas for the future. Discussion was structured around four questions:

1. *What was your favorite aspect of the event?* Panelists commented positively on the organization and professionalism, the split between lectures and labs, the instructors, the diversity of content, and the fact that materials would be available to reuse. Doug interjected that the students were the favorite aspect of many of the instructors!
2. *What would you do differently?* Panelists and participants had many good suggestions. Several noted the diversity of backgrounds and experience among participants meant that not everyone started at the same level; they offered suggestions for trying to get everyone ramped up to a similar level before the event: making lab material available beforehand, pre-readings, brush-up material (e.g. on the linux command line, which many participants struggled with).

   Panelists also mentioned finding the lab time especially valuable, and suggested having more of that, and in particular, having some lab time devoted to problems specifically of interest to African researchers and practitioners.

   Finally, panelists raised the issue of having more presenters from Africa. This is clearly an important area to address for future events.

   When the question was opened to the floor, participants mentioned several other items, most notably:

   - open travel support applications earlier to allow time for visa processing;
   - connect participants in advance to arrange travel and accommodations;
   - find other ways to support attendee travel.
3. *What's next? How should we follow up?* There was wide agreement that the AFIRM school should happen a second time. In addition, panelists raised the possibility of a dedicated IR/DM conference in Africa and an evaluation campaign like TREC dedicated to problems important

Figure 5: Group photo.

to Africans—possibly co-located with AFIRM so that there would be common datasets between labs.

Panelists also wanted to ensure that there would be continued communication among participants, in particular some ongoing mentorship. As well, AFIRM could be a platform for forming collaborations between African universities.

4. *Where should future schools be held?* Although panelists did mention the importance of moving around Africa, they also voiced concern that it will not be easy. Challenges raised included the lack of infrastructure for labs, the lack of institutional support for such an event, and security risks in some countries.

   We also discussed the focus on sub-Saharan Africa as opposed to the entire continent. Many attendees were in favor of participation from North Africa. Some noted that the divide between North and sub-Saharan Africa is quite wide, and if the event were held in North Africa it would likely have a negative effect on participation from sub-Saharan Africa.

   Over the course of the discussion, the room reached broad consensus that it was probably best to return to South Africa for the second year, applying what we learned from the first year, and to think about locations elsewhere in Africa the third year. Not all agreed—Reuben for example emphasized the importance of gaining the context of the rest of the continent, and suggested inviting senior people from institutions outside of South Africa so they could see the benefit themselves and sell the idea at their home institution.

Finally, to end the event, participants were awarded graduation certificates in an informal graduation "ceremony". Hussein Suleman was awarded the Best Lecture award for his talk on IR for Development.

5 Conclusions

The 1st ACM SIGIR/SIGKDD Africa School on Machine Learning for Data Mining and Search ran from January 14–18, 2019, in Cape Town, South Africa. By all indications it was a great success. We did not achieve everything we set out to do, and in particular we have much to do in terms of diversity of the event: diversity of presenters in terms of subject matter (more data mining), origin (more organizers and instructors from Africa and Asia), and gender (more women among organizers and all participants). Nevertheless, the response was overwhelmingly positive, and we are highly encouraged to move forward with the initiative.

A photo album is available at `https://photos.app.goo.gl/1156Ut6FanD7sEEH6`.

5.1 Next steps

We have set up an open-subscription Google Group at `afirm-africa@googlegroups.com` for continued communication among those who were there in person as well as any one else who is interested in joining.

The event will be held again in 2020, very likely in Cape Town around the same time of year. Towards active involvement by the sub-Saharan African community, we have recruited two of our participants to join the steering committee: Lighton Phiri and Reuben Dlamini. We have also identified a local PC co-Chair, Maria Keet of the University of Cape Town, to help ensure that the program is relevant to the needs of researchers and practitioners in sub-Saharan Africa. Finally, a number of our participants will be invited to the PC for 2020, and may even return as lecturers.

In addition, we are inviting a PC co-Chair from the SIGIR community and a PC co-Chair from the SIGKDD community to ensure more balance in topics. We will also review the timing of our attendee funding process to ensure greater diversity and participation from countries like Nigeria.