VLDB Scalable Data Science Category: The Inaugural Year

Arun Kumar
University of California, San Diego
arunkk@eng.ucsd.edu

Alon Halevy
Meta Reality Labs Research
ayh@fb.com

Nesime Tatbul
Intel Labs and MIT
tatbul@csail.mit.edu

1. INTRODUCTION

As part of the International Conference on Very Large Data Bases (VLDB) 2021 / Proceedings of the VLDB Endowment Volume 14, a new Research Track category named Scalable Data Science (SDS) was launched [2, 6]. The goal of SDS is to attract cutting-edge and impactful real-world work in the scalable data science arena to enhance the impact and visibility of the VLDB community on data science practice, spur new technical connections, and inspire new follow-on research. The inaugural year proved to be successful, with numerous interesting papers from a wide cross section of both industry and academia, spanning several data science topics, and originating from several countries around the world.

In this report, we reflect on the inaugural year of SDS with some statistics on both submissions and accepted papers, SDS invited talks, and our observations, lessons, and tips as inaugural Associate Editors for SDS. We hope this article is helpful to future authors, reviewers, and organizers of SDS, as well as other interested members of the wider database / data management community and beyond.

2. SOME SALIENT STATISTICS

Out of 882 Research Track submissions to Volume 14, 112 (13%) were under SDS, 692 (78%) under Regular, 55 (6%) under Experiments, Analysis & Benchmarks (EA&B), and 23 (3%) under Vision. PVLDB does not impose acceptance rates per se. 31 SDS submissions underwent revision and 26 were accepted in the end, yielding an acceptance of 23% for SDS, which was roughly the same as the Regular research category.

2.1 Affiliation Types of Authors

As Figure 1 shows, SDS proved attractive to authors from both academia and industry. There were several submissions from top software companies and research universities from around the world. About half the submissions had at least one industry co-author, although the majority of first authors were from academia, mostly students. Interestingly, papers with a mix of industry and academic co-authors saw a significantly higher acceptance rate than papers with only academic or only industry authors.

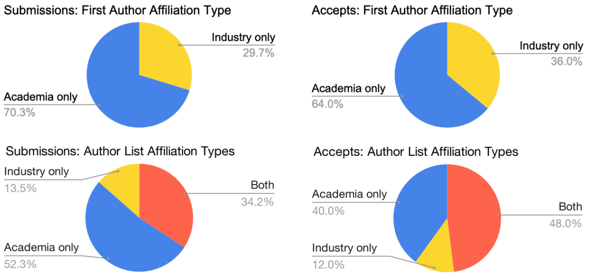

Figure 1: Statistics on author affiliation types.

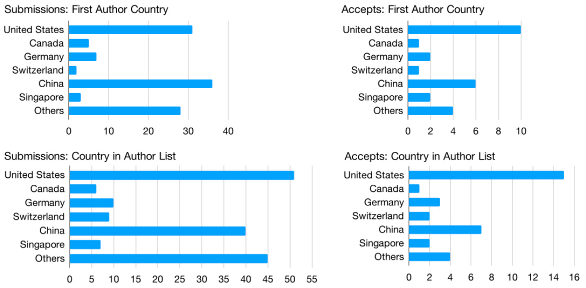

Figure 2: Statistics on countries of authors.

2.2 Countries of Authors

There was significant geographic diversity among both submissions and accepted papers, with the US and China accounting for the largest numbers. Several papers were also cross-national collaborations. Figure 2 shows the top 6 countries based on the nation of affiliation of the lead author and co-author mixtures. The countries in the "Others" category in the plots include Australia, Austria, Belgium, Brazil, Denmark, France, Greece, Hong Kong SAR, India, Israel, Italy, Morocco, New Zealand, Portugal, Qatar, Russia, Saudi Arabia, South Korea, Sweden, Turkey, and the UK.

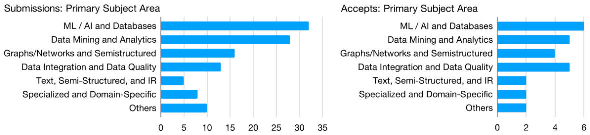

Figure 3: Statistics on primary sub-areas of papers.

2.3 Primary Sub-Areas of Papers

Given the nature of SDS, sub-areas that overlap with Data Science topics naturally saw the highest representation among the primary subject areas. The recently growing sub-area of "ML/AI and Databases" turned out to be the largest, followed by popular classical topics at PVLDB such as data mining/analytics, data integration/quality, and semistructured/graph data. The 2 accepted papers in the "Others" category in the plot had a primary area of provenance/workflows and distributed data systems. Figure 4 shows the "word clouds" produced from the titles of all SDS submissions and accepted papers.

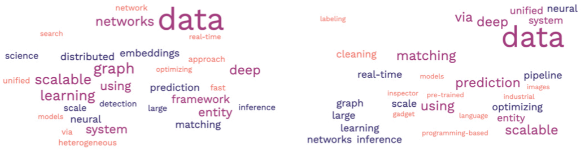

Figure 4: Word clouds of paper titles: submissions on the left and accepted papers on the right.

2.4 Code Release and Reproducibility

In line with the other Research Track categories of PVLDB, SDS encourages authors to open source their research code and data (or at least make them available privately to reviewers) but this is not mandatory. Out of 26 accepted papers, 16 had provided open-sourced code by submission time; 2 others shared their code privately to the reviewers; the remaining 8 did not provide code, with most of their authors being from industry.

3. SDS INVITED TALKS

At the VLDB 2021 conference itself, we had an exciting lineup of 6 eminent invited speakers on SDS-related themes. The talks covered a variety of topics from both industry and academia spanning algorithmics, systems, and applications of Data Science, including Internet data analytics, graph mining, biomedical informatics, and infrastructure for ML and Data Science. The talk titles and speakers are listed below. The talk abstracts and speaker biographies are available online [4] and as part of the conference proceedings [1].

1. "Towards Scalable Online Machine Learning Collaborations with OpenML" by Joaquin Vanschoren of the Eindhoven University of Technology.

2. "Internet Traffic Analysis at Scale" by Anja Feldmann of The Max Planck Institute for Informatics.

3. "The Power of Summarization in Graph Mining and Learning: Smaller Data, Faster Methods, More Interpretability" by Danai Koutra of the University of Michigan.

4. "Designing Production-Friendly Machine Learning" by Matei Zaharia of Databricks and Stanford University.

5. "From ML models to Intelligent Applications: The Rise of MLOps" by Manasi Vartak of Verta.

6. "Summarizing Patients Like Mine via an On-demand Consultation Service" by Nigam Shah of Stanford University.

4. OBSERVATIONS, LESSONS, AND TIPS

Finally, we now share some interesting aspects of serving as inaugural Associate Editors (AEs) of SDS. We also share some lessons for future AEs and other organizers, as well as tips for both authors and reviewers.

4.1 Salient Observations and Lessons

Data Science is a fast-growing field with high interest in both industry and academia around the world. We were happy to see a high volume of high-quality submissions from both universities and companies, including from outside North America and Europe. For instance, the inaugural Best SDS Paper Award went to a paper from Singapore with authors from NUS and Grab [5].

As the CFP explains, SDS welcomes submissions of two types: deployed and evaluated solutions [2]. We were happy to see a healthy mix of both. Many papers described interesting research methods or tools that were at the cusp of practical impact or had just been adopted in practice, helping advance key industrial or domain science applications.

The SDS-specific guidance provided to Research Track reviewers proved sufficient for the most part. That said, in several cases we had to call attention to the special focus on potential for impact and scalability, as well as not overemphasizing novelty of techniques vs. other valuable forms of novelty, such as a new important problem, new key application impact, etc. We hope such multifaceted research evaluation continues.

We oversaw just under a dozen papers being moved across categories of the Research Track as part of a Revision, e.g., SDS to Regular, SDS to EA&B, or Regular to SDS. As can be expected, some reviewers found these moves hard to judge due the differing evaluation criteria. We took this feedback to improve the guidelines for reviewers and also passed it along as an action item to the Editors in Chief and SDS AEs of PVLDB Vol 15. SDS-specific guidelines on technical contributions and scalability were also refined further [3].

4.2 Common Reasons for Rejections

We now list some common reasons we saw why SDS papers got rejected. Some reasons are clearly similar to Regular but their interpretation is often SDS-specific.

1. The paper did not effectively articulate the practical importance of the problem or the potential for practical impact of the work. This criterion is more important for SDS than Regular, since the latter also focuses on speculative or long-term basic research.

2. The empirical evaluation was too weak. While this issue is similar to Regular, SDS-specific criteria that were often underappreciated by authors included showing meaningful scalability metrics and use of real-world datasets.

3. The methods were considered too straightforward or simplistic and there were not enough other forms of novelty to rebalance. Such papers were especially tricky to handle, since this difference in emphasis for SDS vs. Regular is new. It is possible that even a simplistic solution could lead to interesting insights at scale and in deployed scenarios.

4. The paper was entirely ML algorithmics-oriented, with little relevance for the database / data management audience. Data Science is a broad new interdisciplinary field, and PVLDB will not be a suitable venue for many kinds of Data Science papers. Most of the 9 desk-rejected SDS submissions fell under this rationale.

4.3 SDS-Specific Tips

We conclude with some SDS-specific tips based on our experiences as SDS AEs:

- **To SDS AEs:** Recall the SDS-specific evaluation criteria to reviewers when needed during discussions. Keep an open mind on moving a submission across the different categories under the Research Track if the paper's merits warrant that and after discussing with the Editors in Chief.

- **To SDS reviewers:** Read the SDS-specific evaluation criteria again before reviewing such papers. Keep an open mind on the various forms of novelty a paper can bring to the table and also its potential for practical impact, especially if it is of the evidential type.

- **To SDS authors:** Read the CFP on SDS-specific criteria carefully before submitting [3]. Also check out the above list of common reasons for rejections to avoid those pitfalls.

Overall, it was an honor for us to serve PVLDB as AEs to help shape and launch SDS. We are delighted by the enthusiastic response to this new publication category from both academia and industry. We hope this momentum continues and grows alongside the impact and visibility of the VLDB community in the exciting interdisciplinary arena of Data Science.

Acknowledgments. We thank Xin Luna Dong and Felix Naumann for their feedback on this article.

5. REFERENCES

[1] Proceedings of the VLDB Endowment, Volume 14, 2020-2021. Online at `https://vldb.org/pvldb/vol14-volume-info/`.

[2] PVLDB Volume 14 CFP. Online at `https://vldb.org/pvldb/vol14-contributions/`.

[3] PVLDB Volume 15 CFP. Online at `http://vldb.org/pvldb/vol15-contributions/`.

[4] VLDB 2021 Invited SDS Talks. Online at `https://vldb.org/2021/?program-schedule-sds-invited`.

[5] ABEYWICKRAMA, T., LIANG, V., AND TAN, K.-L. Optimizing Bipartite Matching in Real-World Applications by Incremental Cost Computation. *Proc. VLDB Endow. 14*, 7 (mar 2021), 1150–1158.

[6] HALEVY, A., KUMAR, A., AND TATBUL, N. ACM SIGMOD Blog post: "Scalable Data Science: A New Research Track Category at PVLDB Vol 14 / VLDB 2021". Online at `https://wp.sigmod.org/?p=3033`, 2020.