

**Working Group Report on Education and Publications  
In Preparation for the Seattle DB Meeting, October 9-10, 2018**

**Summary**

The working group on education and publications discussed how our community should react to challenges in conferences and peer reviewing, publications, education, careers, mentoring, and diversity.

*Conferences and Peer Reviewing.* One major challenge remains *review quality*. Reviews can lack appropriate feedback and sometimes contain factual errors. In addition, there is a perception that many reviewers look for reasons to reject a paper, and nitpick on details instead of focusing on the big picture, the novelty, and potential impact. Our current review process and common understanding among reviewers honors algorithmic research contributions as “syntactically correct papers”, with a clear definition of a baseline that an algorithm improves upon, whereas some members of our community do not sufficiently value other contributions, such as papers describing innovative systems, applications, experiments and analyses, or user studies. Novelty may be too ingrained in our community reviewer psyche to mean math equation, and theorems (with proof sketches!). It might be better to remove novelty as reviewing criteria, and better talk about about “contribution” and “potential impact”. We need to be more inclusive and embrace the diversity of contributions in our community. We also need to focus on practical impact and encourage work on open-source systems as well as industrial contributions through meaningful presence from industry beyond research labs. At the same time, the broadening of our field with a larger focus on applications presents tradeoffs we need to consider. From one side, there is a risk that our conferences could become a publication venue for tangential or B-rated papers from other fields, such as machine learning, due to a lack of expertise in non-database areas in our already large program committees. Some members of the community are concerned that conferences are becoming too broad, fragmenting into sub-communities with little overlap and joint interest. This becomes a problem in particular, if the program schedule does not account for the different sub-communities our broadened field and schedules all talks of a sub-community concurrently. On the other hand, with the convergence of different areas brought about by the explosion of data-intensive applications (in science and industry alike), and data management being at the core, without broadening we run the risk of missing many boats.

While there rightly is lots of focus on unfair reviewing from the point of view of authors (about which especially junior researchers are concerned), there is also the important yet less discussed concern that incorrect or irrelevant work keeps getting published, with potential long-term costs (e.g., 10-year best papers containing major errors). The push towards having reproducible publications that was pioneered in our community is an important first step towards enabling errors to be detected and corrected. The traditional conference peer review

model as gatekeeper for high-quality, novel research contributions may need to be reconsidered in view of the volume and variety of conference submissions.

*Actions and Recommendations.* Both pVLDB and SIGMOD are actively working towards improving the review process. For example, for 2019, SIGMOD has instituted a rebuttal phase when authors are given the opportunity to comment on factual errors in reviews. Throughout the year, pVLDB has constantly reminded AEs and PC members regarding review quality; AEs are tasked to closely control review quality. Clearly, we need to do more:

- We should educate the young generation on how to properly review papers. One idea would be to hold hands-on-workshops at our main conferences led by members of our community that have been recognized as outstanding reviewers
- We have started to recognize reviewers that do an outstanding job. We should continue to do this and also think of other creative ways of incentivizing reviewers. We should also more proactively track reviewers' performance and provide feedback to reviewers that are not doing a good job. For example, author rebuttals as well as AE feedback could be aggregated over multiple papers (and conferences).

We should continue to encourage a wider adoption of reproducibility in our community. Given the self-correcting nature of science, we should also consider having mechanisms for reporting mistakes and possible corrections, allowing for the publications (and contributions) in our community to be dynamic and evolve in a collaborative effort.

To address the diversity of publications, we should consider to clearly define the different types of research contributions (algorithms, systems, experiments, applications, user studies) and ensure that the program committees of our major conferences (SIGMOD, VLDB) will ensure that all of these contributions are valued. This could be achieved by asking authors to mark the respective contribution type of their submission and by assigning a special editor/co-chair for each contribution type, who monitors reviews and paper discussions and intervenes or pro-actively educates reviewers. Furthermore, in order to draw on broad expertise and be able to judge papers at the boundary of data management, PC chairs will dynamically have to draw external experts from related fields into a program committee as needed and for conferences like VLDB with short evaluation cycles have to postpone a paper decision. With respect to broadening conferences and the challenges of disjoint sub-communities, a task force of SIGMOD and VLDB should identify the major streams and encourage program chairs to schedule the program accordingly. Alternatively, voting by prospective attendees could be used to guide the generation of the program schedule.

*Publications.* With respect to publications, an issue that has been raised is that of the requirement of "too many LPU's" to be published at our prime venues VLDB and SIGMOD. There is a perception that our reviewers nitpick and conference papers take too long to get published -- after (many) successive rejections. This may be partly due to the fact that we consider our conference papers as hybrids between journals and conferences. Areas such ML,

in contrast, are more nimble and publish ideas at a faster pace. Requiring journal-quality submissions for conferences not only delays getting new ideas out, but it also favors incremental contributions which do not lend themselves to lively discussions at conferences. At the same time, we do not have enough of a journal publishing culture compared to other fields, even though our field has matured.

*Actions and Recommendations.* We need to make it more appealing to submit to journals. A good example is the approach adopted by the VLDBJ, where papers can be presented at VLDB. A stronger journal publication culture may also help address issues related to the diversity of publications as journals have more flexibility in finding appropriate reviewers for a specific paper. We also should establish a process to correct papers when errors are discovered. Conferences should also stop requiring others to shorten accepted papers, i.e., the short paper accept of ICDE for posters, as often the shortening will result in worse papers with less information. Either the paper is accepted as is (short or long), or it is rejected.

We should consider alternative paper types, including shorter papers (see Luna's proposal). We should also make sure to maintain an acceptance rate of at least 25%. This will contribute to a more inclusive program and help counter the nitpicking culture.

Rejection with inclusion. For papers that are rejected but look correct, there could be a special section of abstracts and links that are published in conference proceedings where the links point to arxiv papers. Authors of such papers could have one minute to present.

*Diversity.* With respect to diversity, our community should ensure that all activities (reviewing, conferences, etc.) are carried out in an inclusive and diverse environment with zero tolerance for discrimination, harassment, or any other form of misconduct. VLDB Endowment, IEEE ICDE and ACM SIGMOD should work together to enforce this principle, among conference attendees and members of our community. The establishment of DBCares and the creation of an inclusion and diversity officer by the VLDB Endowment illustrate that our community is already taking action with respect to this topic.

*Education.* Our field is broadening from data management to data science. In that respect, we must answer the question of what a data science curriculum should look like and position the data management community as an integral part of data science. While preserving concurrency, transaction management, query processing, query optimization as core concepts that every computer science student should know, we will need to broaden the generic curriculum to encompass the entire data analytics process from data discovery, data preparation/information extraction to information integration and data analysis / model building to information visualization.

*Jobs and Careers.* While data management is central to many different areas -- in science and industry alike, it is not properly recognized. With the increasing buzz around AI, there is a tendency to “call everything AI”. This is already having negative consequences for other computer science sub-areas, including data management. For example, the number of academic and industry jobs in DB seems to be decreasing (add pointer to CRA job report)-- most positions require expertise in Machine Learning and AI. We need to think and act strategically about how to position our field and ensure that its importance is well understood. With respect to careers, we can observe that due to the attractiveness of the field, it is a challenge to retain to researchers in academia.

### **Discussion points:**

- Education, Conferences and Peer Reviewing, Publications, Careers, Mentoring, Diversity: A recurring theme: our field vs. other fields
  - Our field is changing, data management has evolved and broadened beyond transactions and SQL. What constitutes data management today? What are the boundaries? We need a mission statement and a definition, compare recent desk reject discussions. Maybe “Application independent management and processing of data”.
  - Should we retire traditional DB work? how do we measure our success and reward/recognize impact? how does sigmod coexist with SysML? what do we want to do to fix the db reviewing process / why are we against open reviewing?
  - How should we position data management within data science? What is our unique perspective?
- Conferences
  - Some people consider the current conference system broken, with too many talks about low-impact work, overlapping talks on topics of interest in one session and hardly any interesting talk in other sessions. How do we ensure that our conferences remain interesting for sub-communities in a broadened field of data management / data science ?
  - There are different type of research contributions such as algorithms, experiments and analyses, innovative systems, user studies papers. What constitutes a research contribution for each of these types of papers?
  - Increase acceptance rates to at least 25%
  - Rejection with inclusion: For papers that are rejected but look correct, there could be a special section of abstracts and links that are published in conference proceedings where the links point to arxiv papers.
- Reviewing
  - Unethical behavior: How can we deal with reviewer cliques (i.e., paper IDs are being passed around by authors on a back channel, “friendly reviewers” are deliberately bidding for these papers, even if outside domain of expertise, and ensuring favorable reviews.). Possible solutions could be: manual assignment of reviewers by EICs and/or AEs, or EICs and AEs do a careful reassessment and fine-tuning of reviewer assignments post the bidding phase to ensure that cliques are not created. Another good practice would be to ensure, as far as possible,

that reviewers are from different geographical regions or ethnic backgrounds wrt authors.

- Can we get insights into the review process by analyzing the data? Bids, discussions, scores, reviewers? Can we obtain additional data, e.g., feedback from authors on reviews?
- Publications
  - Should we, and if so, how can we make papers live objects/liquid papers, with public reviews and a discussion forum as well as artefacts such as data and code?
  - How can we reward the creation and publication of open-source libraries, open-data, and make them citeable?
- Unethical behavior
  - We now have DBCARES. How should the community react to sexism, racism, or other improper misconduct? What kind of reprimands should be applied in case of violation?
- Education
  - Data management courses are not mandatory for CS or Data Science students. How to strengthen these in the ACM core? Has dealing with data changed relative to what we teach in a traditional database course. There should be more focus on the data analysis pipeline from information extraction through cleaning, integration, analysis, model building and visualization, with data management at its center.