RDAP ASSOCIATION

RDAP Happenings

Hi RDAP members,

This month, our newsletter focuses on activities which bring our community together — our forthcoming conference, as well as broader topics that affect our work. Recently, I was reminded of the role data and data visualizations play in everyday experience.

Generally, my passion for sports statistics is roused in the spring, with the beginning of baseball season, but during the last few weeks it was football data that came to the fore. (I have been a long-time fantasy football player, well before Yahoo.) The other day, I tuned in to see a Sunday afternoon NFL playoff game, and instead of being able to focus on the action, my attention was drawn to the minute-by-minute statistical feed, from yards to attempts to percentages. The game itself was barely visible behind all the graphics. I finally found a full-screen broadcast on another channel. That said, I realize the first type of broadcast is popular in sports parlors and on live betting sites, supporting activities besides purely watching the game.

During studio banter between plays, Terry Bradshaw (the quarterback of the Pittsburgh Steelers during their dominant run in the 70s) mused that NFL records were being broken left and right primarily due to longer seasons and more playing time, suggesting that game statistics should be normalized. (Of course, if you have watched Mr. Bradshaw for years, you will know he did not actually use the term “normalized”.) He was both right and wrong about the merits of such a change, as neither normalized nor “raw” stats tell the whole story of the game — or the players.

I was reminded on that winter Sunday that the data we work with needs to reflect the “game” we play, but it is not the “game” in itself. The stories we tell are more complex than they first appear. As data practitioners, we should continue to tackle hard topics within our society while opening a running lane to new perspectives.

[Editor’s note: The “I” above refers to sports numbers enthusiast Lora Lennertz, principal author of this newsletter’s introduction, but Ali Krzton takes responsibility for all editorializing about the Steelers’ greatness and any sports metaphors stretched beyond their utility.]

Updates from the Board

With the start of the new year, the RDAP Board strives to continue supporting our committees and members as we all navigate shifting circumstances while struggling to balance capacity and output. We hope everyone was able to take time at year's end to rest and recuperate, and we are very excited for our upcoming work in 2022.

In particular, we want to welcome the members of the new Diversity, Equity, Inclusion, and Accessibility (DEIA) Action Committee and extend thanks to Erin Carillo and Monica Ihli for volunteering to co-chair the group through its formation and beginning. This work will be based on the committee's charge to "support RDAP's efforts to advocate for and actively support DEIA within our organization and within the fields of data management and curation", guided by the 95 RDAP member responses to the DEIA Task Force survey we received in the fall. We are also working to hire an Accessibility Fellow to assist in developing a strategic plan focused on standardizing accessibility for the RDAP annual summit and other RDAP-run virtual educational events.

The RDAP Board,

- Jonathan Petters (President)
- Amy Koshoffer (President-elect)
- Rachel Woodbrook (Secretary)
- Patti Condon (Treasurer)
- Jen Darragh (Past President)

RDAP Feature Article

"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI

Summary and discussion by Ali Krzton, Research Data Management Librarian, Auburn University

Auburn University

Those who work with data have learned the importance of provenance, documentation, standardization, context, and metadata in maintaining the quality of datasets. This was historically done to preserve their utility for human reuse and re-examination, but in recent years the emphasis on machine-readability of datasets has increased, in part to allow for their use in AI (artificial intelligence) applications. Just as those involved in creating and maintaining datasets benefit from an improved understanding of how they might be used with AI, the developers of AI systems should pay attention to issues that affect the data upon which their models rely. Several Google researchers present this perspective in "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI", a conference paper based on their qualitative study of AI practitioners (Sambasivan et al., 2021).

Sambasivan et al. (2021) interviewed 53 practitioners in East and West Africa, India, and the US who work on high-stakes AI, or AI applied in critical domains where failures have profound negative impacts on people. Through this work they identified a persistent problem, *data cascades*, originating from quality issues in the datasets used to build AI models rather than features of the models themselves. The report defines a data cascade as "compounding events causing negative, downstream effects from data issues, that result in technical debt over time" (Sambasivan et al., 2021: 5). That technical debt incurs human costs including harms to intended beneficiaries, abandonment of projects, alienation of project partners, and wasted time and effort (Sambasivan et al., 2021: 8).

While data cascades had multiple causes, the authors point to devaluation of data work with respect to model work as a central theme. This extended even to the domain expertise needed to understand and interpret the data in the first place, leaving the programmers to make classification and cutoff decisions they admitted they were not qualified to make (Sambasivan et al., 2021: 7). AI practitioners were aware that people involved with data collection and organization were not rewarded for their work, or if they were, they might be rewarded in ways that worked against data quality (Sambasivan et al., 2021: 9). Shortcomings in the education and training background of AI practitioners also contribute to data cascades. Most of them learned AI methodologies on extremely clean "toy" datasets or a selection of commonly used open datasets that were nothing like the real-world data on which they were required to build and train their models (Sambasivan et al., 2021: 11). Consequently, they were not prepared to deal with issues such as inaccurate, incomplete, non-representative, or poorly-documented data, leading to data cascades.

A canonical example of a high-stakes AI domain is healthcare. As AI tools are increasingly brought to bear on healthcare decisions, there is a growing risk that data cascades will lead to model problems discovered long after they have harmed substantial numbers of people. A separate study of the performance of a proprietary algorithm designed to provide early warning of sepsis by University of Michigan personnel found it flagged too many false positives while missing real cases of sepsis; the study's lead author, Karandeep Singh, surmised that the model might be flawed because it was validated with billing codes rather than clinical data (Simonite, 2021). This is a data cascade, in this case arising from the use of data that was not a reliable indicator of the phenomenon of interest.

As AI researchers and practitioners alike discover the value of data work, data practitioners are presented with an opportunity to start new conversations and draw attention to the need for data expertise in AI-driven projects.

References

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L.M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 8–13, 2021, Yokohama, Japan, 1–15. doi: 10.1145/3411764.3445518 Open access copy available from https://research.google/pubs/pub49953/

Simonite, T. (2021-06-21). An algorithm that predicts deadly infections is often flawed. *Wired*. https://www.wired.com/story/algorithm-predicts-deadly-infections-often-flawed/

RDAP Summit Updates

Happy New Year from the Conference Committee! The RDAP Summit, March 15th–17th, is only two short months away. Registration for the Summit is now open and costs $25 USD for RDAP members, students, and those with financial hardships.

The schedule of events is posted on the Summit's website, and a more detailed schedule will be posted closer to March 15th. We've also introduced a new program for 2022: the Summit Buddy Program. This program partners new attendees with experienced attendees as buddies for the Summit. You can sign up for the program using the Registration Form.

Last but by no means least, we're very much looking forward to hearing from our keynote speaker, Jordan Harrod, on March 15th at 12pm ET/9am PT. Jordan is a PhD candidate in Medical Engineering and Medical Physics in the Harvard-MIT Health Sciences and Technology program. She works at the intersection of non-invasive brain-machine interfaces and machine learning for pain and anesthesia. You can check out Jordan before her keynote via her YouTube channel.

Action Committee Updates

DEIA

The new Diversity, Equity, Inclusion, and Anti-Racism (DEIA) Action Committee was formed in November 2021 to implement the recommendations developed by the DEIA Task Force. The DEIA Action Committee is responsible for supporting RDAP efforts to advocate for and actively support DEIA within our organization and within the fields of data management and curation. The members of the committee are: Monica Ihli (co-chair), Erin Carrillo (co-chair), Michele Matz Hayslett, Josh Sadvari,Amy C. Schuler, and John Watts.

Education and Resources

Last year, RDAP Education & Resources held three town hall-style webinars: “Final NIH Policy for Data Management and Sharing”, “Writing More Equitable Job Postings”, and “Using Social Media Research Data Responsibly: Considerations for Librarians and Researchers”.

This year, we plan to once again host three or four webinars, and we are seeking community input on topics for Spring 2022. These opportunities can take the form of lecture or workshop-based webinars, town hall Q&A, or other innovative ideas. Our events will take place on the RDAP Sponsored Zoom account with live transcription, and we hope to be able to provide recordings to RDAP members.

Do you have a topic you would like to present to the RDAP community? Propose it! Want a webinar on a specific topic? Suggest it along with potential speakers we might invite. We particularly welcome topics with a DEIA focus.

We are interested in suggestions from RDAP members to either attend or present.

Please complete this proposal form to submit your suggestion.

If you have questions or would like to discuss an idea, please reach out to one of the co-chairs, Courtney Kearney (ckearney@tulane.edu) or Shannon Sheridan (Shannon.sheridan@pnnl.gov). The RDAP Education and Resources Committee will review proposals on a rolling basis.

Marketing

No updates at this time.

Membership

Through a joint effort from the Sponsorship and Membership Action Committees and supported by sponsorship funds from the University of Wisconsin-Madison, scholarships are available to cover registration for the Research Data Access & Preservation Summit 2022 and an RDAP membership for one year. Please submit your application by January 28 through theonline application form. Contact membership@rdapassociation.org if you have any questions. Winners will be notified by February 11.

Publishing

Publishing

The Publishing Committee chair met with the Editor-in-Chief of the *Journal of eScience Librarianship*. We will once again contribute guest editors from this committee to help produce an RDAP Special Issue of *JeSLIB* featuring scholarship from the Summit. This is the fifth year of this partnership, and we are thankful for the opportunity to continue the critical discussions that originate from the RDAP Summit. Additionally, other members of the Publishing Committee will serve as mentors to authors who want additional feedback and support as they navigate the peer review process.

Sponsorship

No updates at this time.

Website

No updates at this time.

STAY CONNECTED!

Follow us on Twitter, join our listserv, or send us an email.

Visit us!

Unsubscribe