

Introduction to SIAM Journal on Mathematics of Data Science (SIMODS)

Tamara G. Kolda, Editor-in-Chief*

1. Welcome. On behalf on the editorial board and SIAM, it is my sincere pleasure to introduce the inaugural issue of *SIAM Journal on Mathematics of Data Science* (SIMODS). The goal of SIMODS is to provide the reader with a curated collection of articles that highlight the key role of mathematics in advancing the field of data science. We strive to have articles that are relevant, timely, and accessible to a broad range of data scientists coming from different areas of mathematics, statistics, computer science, and engineering.

The ascent of data science is a boon for mathematics. Although mathematical innovation was initially overshadowed by advancements in the hardware and software architectures for efficiently managing petabytes and zetabytes of information, nowadays algorithms dominate, and advances in analysis of massive data require a combination of mathematical, computational, and inferential thinking [1].

To demonstrate how exciting this time is for mathematics, let me tell you about Sean McClure. Sean holds a Ph.D. in chemistry and manages a group working on artificial intelligence at Accenture. He's written a few blog posts on data science, including "What it Means to do Math in Data Science."¹ This post is remarkable because I can't think of a better descriptor of the power of mathematics in the field:

"...it isn't a candidate's mathematical dexterity we are after, it is their understanding of concepts. It is in grasping the appropriateness, or unsuitability, of an approach that matters... Abstraction is what gives academic discoveries its legs and this is where the data scientist operates. When we function above deep technical details we are able to use mathematically-driven tools to maneuver around hard challenges and solve interesting problems using software."

Did you get that? The author, who is not a mathematician, just summed up the entire purpose of advanced applied mathematics research: employ conceptual abstractions to overcome challenges and solve interesting real-world problems. Mathematical thinking has become a differentiating skill in the world of data science. The time has come for mathematics to have its own voice in the field, and SIMODS is that voice.

The genesis of SIMODS was a discussion with journal co-founder Joel Tropp² at the 2016 SIAM Annual Meeting in Boston, Massachusetts. We kvetched (there is no better word for it) about the unmet need for the SIAM journal program to serve the burgeoning generation of applied mathematicians working in data science. This notion that we were neglecting a

*Sandia National Laboratories, Livermore, CA.

¹<https://towardsdatascience.com/what-it-means-to-do-math-in-data-science-843f454fdddf6>.

²Steele Family Professor of Applied & Computational Mathematics, Department of Computing and Mathematical Sciences, California Institute of Technology.

significant contingent of up-and-coming mathematicians compelled us to action.³ Perhaps a less noble motivator was that we just wanted to create a journal where we and others like us could publish and find great papers to read. The problem was that it was hard to find a venue where there was appreciation for the *combination* of mathematics, statistics, and computational methods in the context of data science. In our roles as editors for other SIAM journals, we saw firsthand that these journals lacked sufficient expertise in data science and so were not always in a position to appreciate the context of data-science-oriented submissions. In publication venues outside of mathematics, such as machine learning conferences, mathematical developments were often relegated to appendices. Our colleagues in statistics have arguably always been data scientists, but the scale of today's data science problems requires a more interdisciplinary approach incorporating computational mathematics and computer science in new ways that don't easily fit into traditional statistical journals. So we set out to establish a journal for the best mathematical work in data science, where context was appreciated and the mathematical details were celebrated.

Ultimately, our intent is for SIMODS to serve as a home for papers at the intersection of five key mathematical fields: applied mathematics, computer science, statistics, signal processing, and network science. In the context of data science, it is rather difficult to separate these domains because major advances usually stem from a combination of techniques. Fields like machine learning and artificial intelligence are part of computer science, but they have deep roots in statistical methodology and probability theory. Network science is itself a subject that spans mathematics, computer science, and the physical sciences. Signal and image processing involve tools from approximation theory, optimization, and probability and statistics. Applied mathematics supplies other core ideas, including numerical linear algebra, numerical analysis, scientific computing, and mathematical modeling. To appreciate the interdisciplinary nature of data science, look no farther than the affiliations of our associate editors, which range across mathematics, statistics, computer science, and engineering departments.

As we developed the journal, Joel and I were incredibly fortunate to recruit Al Hero,⁴ Mike Jordan,⁵ and Rob Nowak⁶ to the leadership team. Their influence in shaping what you see before you has been considerable, to say the least. Together, we recruited an all-star editorial board, which we list in its entirety at the end of these opening remarks. I'd love to take credit for assembling this amazing team but, in truth, recruiting was easy because every member of the board appreciated the need for such a journal and was eager to help bring it to reality.

2. A little history on the proposal and launch. We are pleased to be working with SIAM because they are known for having strong journals and for being highly interdisciplinary. To maintain the journal program's integrity, the process of creating a journal at SIAM is an arduous process. The initial proposal was submitted in early 2017 to Mike Miksis, the

³To give some data as motivation, the SIAM Activity Group (SIAG) on Data Mining and Analytics had the second highest number of student members, topped only by the SIAG on Computational Science and Engineering, signaling its importance to the next generation.

⁴John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and R. Jamison and Betty Williams Professor of Engineering, University of Michigan, Ann Arbor.

⁵Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics, University of California, Berkeley.

⁶Nosbusch Professor in Engineering, University of Wisconsin-Madison.

SIAM Vice President for Publications. At the time, the working title was *SIAM Journal on Mathematics of Information and Data (SIMID)*. The proposal was shared with the Journal Committee, the editors-in-chief for all the SIAM journals, and SIAM's officers. Their collective feedback led to many substantive changes in the proposal. The editors-in-chief were especially supportive, rising above any territorial claims.

The next step was approval of the SIAM Council at their July 2017 meeting in Pittsburgh, Pennsylvania. The working title had evolved to *SIAM Journal on Mathematics of Data (SIMOD)*. The proposal was formally submitted by Amr El-Bakry (ExxonMobil) and Bruce Hendrickson (LLNL) as well as Joel, Al, Rob, and me. By this point, we had reached out to potential editors and invited them to "indicate your willingness to seriously consider the offer if the proposal goes forward and you are indeed invited to be an associate editor." Because of our need to ensure representation across data science topics, not every one of those initial volunteers made it onto the inaugural editorial board, but we are nonetheless grateful for their support. I had the honor of presenting the proposal to the Council. Many a journal proposal has died in the hands of the SIAM Council, and none have made it through unscathed. We were no exception, so we had to make a few more changes before final approval.

The main request of the Council was to include even more statisticians in the journal leadership. Although SIAM has a joint journal, *Journal on Uncertainty Quantification (JUQ)*, with the American Statistical Association (ASA), fewer than 3% of SIAM members list statistics as their primary departmental affiliation. Getting more statisticians engaged was definitely going to be a challenge. Even though both Al and Rob had joint appointments in statistics, we needed more. Luckily for us, this is when we were able to convince Mike Jordan to join the leadership group and bring more statistical heft to our team. Thanks to this change and others, the result is an impressive engagement with the statistics community. I am especially indebted to those statisticians who were not previously engaged with SIAM but nonetheless willing to join our fledgling venture.

At the behest of Kivmars Bowling, SIAM's Publications Director, we also changed the journal title to its current form, *SIAM Journal on Mathematics of Data Science (SIMODS)*, which was a great improvement. Our revised proposal was approved by the SIAM Council via email vote in November 2017. Joel and I were able to give a revised and much improved presentation to the SIAM Board of Trustees at its December 2018 meeting in Philadelphia.⁷ The Board gave its unanimous approval, with one officer calling it a "slam dunk" and another saying it was "a big opportunity for SIAM." Shortly thereafter, I was appointed to serve as Editor-in-Chief.

3. Launch. As soon as we received final approval from SIAM's Board, we set about inviting members for our inaugural editorial board. Since the quality of the editorial board is what will ultimately ensure the success of the journal, we were delighted by the positive responses. Our editors are some of the most sought-after luminaries in the field, so to say they are busy is an understatement. Nevertheless, they agreed to be part of SIMODS because they shared our vision for how this journal would benefit the community. These editors and

⁷Full disclosure: I was a member of the SIAM Board of Trustees when the journal was approved. However, we followed conflict-of-interest policies, and I was absent for the deliberations and vote.

their successors will set and uphold the standards for the journal. The success of SIMODS is their success.

Although the editors are the outward face of the journal, the support of editorial staff is what makes the journal run smoothly. Quite a bit of work goes on behind the scenes. One of the major benefits of working with SIAM is its amazing publications department. SIMODS is especially indebted to Mitch Chernoff and Heather Blythe, who brought their considerable expertise in running journals to SIMODS. We were able to start taking submissions less than 6 months after we got the green light, in late April 2018. Six months later, we had received over 100 submissions, breaking all prior SIAM records for submissions to a new journal. I hope that this means we have filled a void in SIAM's journal program. In the end, we are publishing our first issue in less than 10 months from when we launched and only 2.5 years after the idea was conceived.

4. Contents of the inaugural issue. Our intention for SIMODS is to provide a curated collection of interdisciplinary articles that invite readers to delve into new topics. The papers in this issue are united by the theme of data science and the key role of mathematical thinking, but they are otherwise diverse. Some lean more toward theory while others toward applications, some are answering old questions while others pose new ones, and some quite serious while others are arguably fun.

In this era, no journal on data science would be complete without papers on deep neural networks. We are happy to feature several papers in this inaugural issue. The paper “Optimal Approximation with Sparsely Connected Deep Neural Networks,” by Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok (SIMODS editor), and Philipp Petersen, investigates to what extent a sparse deep neural network (DNN) can approximate general functions to within a prescribed accuracy. A fully connected DNN with hundreds of layers and hundreds of nodes per layer is difficult to train and use due to extensive memory and computational requirements; therefore, sparse DNNs are popular for use cases where memory and computational power is limited, such as smart phones. The beauty of this paper is that it enables the transfer of existing extensive mathematical results on functional approximation (such as results on wavelets from applied harmonic analysis) to the domain of neural networks. For instance, it shows that the bound for effective DNN approximations with M connections is identical to that for standard M -term approximations. The paper establishes novel bounds and includes numerical results that show that stochastic training methods provide close-to-optimal approximations. In “Multi-Layer Sparse Coding: The Holistic Way,” authors Aviad Aberdam, Jeremias Sulam, and Michael Elad (editor-in-chief of SIIMS, a sister publication) take a renewed look at sparse coding in signal processing, which has connections to deep neural networks. Using a new interpretation and different analysis techniques enables them to improve the theoretical state of the art. Ultimately, the mathematical ideas here are anticipated to impact how deep neural networks are trained. As one referee stated, “I consider the development of appropriate signal models as a key mission in the theoretical foundation of deep learning — and in fact, the presented results are one of the very few addressing this important issue.” In their paper “New Error Bounds for Deep ReLU Networks Using Sparse Grids,” Hadrien Montanelli and Qiang Du tie neural net approximation theory to existing mathematical results, in this case the theory of sparse grids. With these tools, the authors improves the best known theoretical

approximation bounds for ReLU DNNs for certain classes of functions.

We also include papers on network sciences. A nearest neighbor (NN) graphs creates a graph by connecting points (nodes) in n -dimensional space to their nearest neighbors in a point cloud. This can be done by taking the k nearest neighbors (no matter how far away) or all neighbors within a prescribed ϵ ball (no matter how many there are). The paper “Variational Limits of k -NN Graph-Based Functionals on Data Clouds,” by Nicolás García Trillo, considers the stability of variational problems on k -NN graphs, in contrast to prior works that have focused on ϵ -NN graphs. The analysis is motivated by contexts where the exact distances are not known but only relative relationships are (such as a is closer to c than b). Applications are abundant in machine learning, including clustering, classification, and semi-supervised learning. The paper “The Rankability of Data” by Paul Anderson, Timothy Chartier, and Amy Langville poses and solves a new problem. Ranking is a fundamental data science task. Its applications are numerous and include web search, data mining, cybersecurity, machine learning, and statistical learning theory. Yet little attention has been paid to the question of whether a graph dataset is suitable for ranking. The rankability problem asks: How can rankability be quantified? Can rankable subgraphs be identified? At what point is a dynamic, time-evolving graph rankable?

From recommender systems in marketing to image analysis in neuroscience, we often presume that data lives some to-be-discovered lower-dimensional subspace. What separates real-world matrices from indubitably full-rank random matrices? Madeleine Udell and Alex Townsend answer this question in their paper “Why Are Big Data Matrices Approximately Low Rank?” They prove that most matrices can be well-approximated entry-wise by a low-rank matrix. Their results hinge on a clever application of one of the most powerful results in data science: the Johnson–Lindenstrauss lemma, which proves that collections of points in high-dimensional space can be embedded into low-dimensional space with minimal distortion in the pairwise distances by using random orthogonal projections. Among other consequences, this arguably has implications for testing various methods for generating realistic artificial data.

Privacy is another major concern in data science. If we share only data summaries, is individual data still private? For instance, suppose a genetic screening company shares information about a particular genetic disease by giving counts of how many people have or don’t have it for a large subset of the full population? How many queries would it take to reveal individual results? The paper “Decoding from Pooled Data: Sharp Information-Theoretic Bounds” by Ahmed El Alaoui, Aaditya Ramdas, Florent Krzakala, Lenka Zdeborová, and Michael Jordan (section editor for SIMODS) develops bounds on the number of queries, assuming that the data is discrete and the histograms are exact.

Clever data sampling can revitalize even the most tried-and-true methods such as least squares. The paper “Sequential Sampling for Optimal Weighted Least Squares Approximations in Hierarchical Spaces” by Benjamin Arras, Markus Bachmayr, and Albert Cohen considers the problem of developing good approximations for continuous functions in the context of sequential least squares, which has applications to signal processing as well as numerical approximation of high-dimensional stochastic PDEs. In this case, the goal is to limit the number of samples but have freedom in specifying which samples. The authors propose a

method for reusing previous samples and provide a bound on samples that is linear (up to log factors) in the dimension of the approximation space.

We conclude with a pair of related papers on automatically selecting informative landmarks for three-dimensional objects, with application to geometric morphometrics. Ideally, the landmarks capture morphologically distinct shape variables so that objects can be analyzed independently of their size, position, and orientation. In part one, “Gaussian Process Landmarking on Manifolds,” Tingran Gao, Shahar Z. Kovalsky, and Ingrid Daubechies develop an optimally convergent procedure based on Gaussian process active learning. In part two, “Gaussian Process Landmarking for Three-Dimensional Geometric Morphometrics,” Tingran Gao, Shahar Z. Kovalsky, Doug M. Boyer, and Ingrid Daubechies demonstrate the real-world effectiveness of their technique for developing registration maps between biological structures such as fossilized bones and teeth. This approach has the potential to replace very tedious manual determination of landmarks.

5. Closing thoughts. As we move forward, SIMODS will establish mathematics’ importance in the fast-growing domain of data science and serve as a home for those that work at this crossroads of mathematics, statistics, computer science, network science, signal processing, and other fields. We started taking submissions in late April 2018 and received extremely strong submissions from a broad spectrum of authors. Here we are, in February 2019, with the debut of this publication, starting with ten of those early submissions. We appreciate the many excellent submissions that we have received and hope that they continue. Our intent is to have a balance of theoretical developments that have practical implications and algorithmic and methodological innovations that directly advance the practice. We foresee fresh new ideas alongside innovative improvements to existing methods. We aim to have articles that challenge traditional thinking next to ones that explain the successes (and failures) of existing methods. Papers published in SIMODS will develop useful theories, propose new algorithms, describe clever implementations, and share novel methodologies across disciplines. We anticipate that these papers will not only be useful to data sciences but also have ramifications for traditional areas of applied mathematics research as they incorporate methods that have advanced in the data science regime. We are indebted to SIAM for its welcoming atmosphere for multidisciplinary research. It is hard to imagine a better home for SIMODS, and we look forward to seeing many new SIAM members coming from data sciences as well as the launch of a new conference in the same theme coming in 2020. Stay tuned!

Tamara G. Kolda
Editor-in-Chief
SIAM Journal on Mathematics of Data Science

Appendix A. Inaugural SIMODS Editorial Board. **Editor-in-Chief:** Tamara G. Kolda, Sandia National Laboratories. **Section Editors:** Alfred Hero, University of Michigan; Michael Jordan, University of California, Berkeley; Robert D. Nowak, University of Wisconsin, Madison; Joel A. Tropp, California Institute of Technology. **Associate Editors:** Maria-Florina Balcan, Carnegie Mellon University; Rina Foygel Barber, University of Chicago; Robert Calderbank, Duke University; Venkat Chandrasekaran, California Institute of Tech-

nology; Jennifer Chayes, Microsoft Corporation; Alexandre d'Aspremont, CNRS; Ioana Dumitriu, University of Washington; Maryam Fazel, University of Washington; Emily Fox, University of Washington; Mark Girolami, Imperial College London; David F. Gleich, Purdue University; Ashish Goel, Stanford University; Ilse Ipsen, North Carolina State University; Eric Kolaczyk, Boston University; Gitta Kutyniok, Technische Universität Berlin; Monique Laurent, Centrum Wiskunde & Informatica; Elizaveta Levina, University of Michigan; Yi Ma, University of California, Berkeley; Michael Mahoney, University of California, Berkeley; Long Nguyen, University of Michigan; Ivan Oseledets, Skolkovo Institute of Science and Technology; Natesh Pillai, Harvard University; Ali Pinar, Sandia National Laboratories; Mason A. Porter, University of California, Los Angeles; Bala Rajaratnam, University of California, Davis; Philippe Rigollet, Massachusetts Institute of Technology; Justin Romberg, Georgia Institute of Technology; Ronitt Rubinfeld, Massachusetts Institute of Technology; Richard Samworth, University of Cambridge; Katya Scheinberg, Lehigh University; Amit Singer, Princeton University; Marc Teboulle, Tel Aviv University; Ramon van Handel, Princeton University; Weichung Wang, National Taiwan University; Rachel Ward, University of Texas, Austin; Rebecca Willett, University of Chicago.

REFERENCES

- [1] NATIONAL RESEARCH COUNCIL, *Frontiers in Massive Data Analysis*, The National Academies Press, Washington, DC, 2013, <https://doi.org/10.17226/18374>.