# Ryan R. Curtin

439 Calhoun St. NW
Atlanta, GA 30318
443.534.0378
ryan@ratml.org

*Understand algorithms. Make them faster.*

## Overview

Inefficiency permeates data science at every level. Have you ever wondered why so many gigabytes of dependencies are necessary for a typical data science project deployment? I certainly have. I wondered that while waiting on models to train. Then, I wondered if the models were really being trained using the right algorithms for the problem at hand. So I dug, and what I found was that the state of affairs is *not* great: data science as a whole is so much slower than it could be. Data science libraries with redundant computations; infrastructure tooling with too big a footprint; machine learning techniques that use asymptotically inefficient algorithms; software that does too much and in an effort to be 'easy' gives up efficiency altogether.

I'm an optimist. We can do better. I'm working to improve the situation. Over the past decade, I implemented large parts of the Armadillo, mlpack, and ensmallen libraries in C++: fast alternatives to standard data science tools. But clever implementation tricks aren't enough—we need better algorithms too. This led me to study *dual-tree algorithms*, where I formalized the class of algorithms, proved previously-unknown runtime bounds, and extended their applications to new problems. Even that's not the end of it: what's the use of a blazingly fast algorithm if just getting the data takes hours? We can do better there too: with others, I developed algorithms for in-database machine learning that provide orders-of-magnitude speedup by avoiding unnecessary joins and redundant computation. I know we can provide efficient, effective tools to data scientists by combining these three techniques:

→ optimizing low-level implementations,
→ selecting appropriate asymptotically efficient algorithms, and
→ eliminating redundant computations.

Those things are just the tip of the iceberg. There's so much to do! Let's get to it.

## Selected Relevant Publications

- "**mlpack 4: a fast, header-only C++ machine learning library**". R.R. Curtin, M. Edel, O. Shrit, *et al. Journal of Open Source Software*, vol. 8, issue 82, 2023.
- "**Rk-means: Fast Clustering for Relational Data**". R.R. Curtin, B. Moseley, H.Q. Ngo, X.L. Nguyen, D. Olteanu, M. Schleich. In *Proceedings of the 23rd Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, p. 2742–2752, 2020.
- "**On Coresets for Regularized Loss Minimization**". A. Samadian, K. Pruhs, B. Moseley, S. Im, R.R. Curtin. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, p. 482–492, 2020.
- "**Flexible numerical optimization with ensmallen**". R.R. Curtin, M. Edel, R.G. Prabhu, S. Basak, Z. Lou, C. Sanderson. *arXiv preprint arXiv:2003.04103*, 2020.
- "**Detecting adversarial samples from artifacts**". R. Feinman, R.R. Curtin, S. Shintre, A.B. Gardner. *arXiv preprint arXiv:1703.00410*, 2017.
- "**Armadillo: a template-based C++ library for linear algebra**". C. Sanderson, R.R. Curtin. *Journal of Open Source Software*, vol. 1, issue 26, pp. 1–2, 2016.
- "**Tree-independent dual-tree algorithms**". R.R. Curtin, W.B. March, P. Ram, D.V. Anderson, A.G. Gray, C.L. Isbell, Jr. In *Proceedings of the 30th International Conference on Machine Learning (ICML '13)*, pp. 1435–1443, 2013.

## Education

**Ph.D.** *in Electrical and Computer Engineering* August 2015
Georgia Institute of Technology, Atlanta, GA
Thesis: **"Improving Dual-Tree Algorithms"**

**Master of Science** *in Electrical and Computer Engineering* May 2009
Georgia Institute of Technology, Atlanta, GA

**Bachelor of Science** *with Highest Honors in Electrical Engineering* May 2008
Georgia Institute of Technology, Atlanta, GA

## Professional Experience

**Independent Consultant**, Atlanta, GA Spring 2023 - present

I build high-quality high-impact open source software for machine learning:

**mlpack** – **ensmallen** – **armadillo** – **bandicoot**

I spend my time making the vision I sketched out earlier a reality. I work with companies to eradicate inefficiencies in their existing data science pipelines, take a data science idea and turn it into a project, get a data science prototype into production, provide custom implementations of algorithms where standard tools don't suffice, and all kinds of things along these lines. In short, that's *data science consulting, but focused on efficiency at all levels*.

**RelationalAI, Inc.**, Atlanta, GA Fall 2018 - Fall 2022
*Computer Scientist*

At RelationalAI my work consisted of developing novel, accelerated in-database algorithms for machine learning problems, as well as helping design and implement the database system on which those algorithms ran. My works at RelationalAI had a few main themes:

- Design the machine learning API for RelationalAI's database query language (Rel), and bind in support for external machine learning libraries such as mlpack.
- Work as part of a research team to develop and implement an automatic differentiation system for use inside of Rel (a fully declarative language).
- Implement query optimization strategies to allow efficient training of machine learning models on relational data, without materializing joins unnecessarily.

**Symantec Corporation**, Atlanta, GA Fall 2015 - Summer 2018
*Principal Research Scientist*

My responsibilities at Symantec were to pursue a research programme loosely focused on Symantec-relevant applications such as malware classification and related tasks, while continuing work as the lead developer of mlpack.

**Georgia Institute of Technology**, Atlanta, GA Fall 2009 - present
*Graduate Research Assistant*

During my graduate school career, I worked for four different labs on different research projects, and simultaneously led development of **mlpack**, an open-source C++ machine learning library currently with over 4k stars on Github and over 250 contributors from around the world.

**Google, Inc.**, Mountain View, CA Summer 2010
*Software Engineering Intern*
Worked with the Similar Pages team to provide improved search results.

**Georgia Tech Research Institute**, Atlanta, GA Spring 2009 - Spring 2010
Food Processing Technology Division
ELSYS Lab
*Graduate Research Assistant*
Applied machine learning techniques for stress detection in broiler chickens.
Investigated techniques for the analog-to-digital frontend of a radar warning receiver.

**Nexidia, Inc.**, Buckhead, GA Summer 2007
*Research Intern*
Created voice synthesizers that can generate missing samples, and still be comprehensible.

## Skills and Miscellany

- Extensive knowledge of Linux and related UNIX-like systems (as well as Windows)
- Good understanding of and experience with 1930s automotive technology
- Extremely comfortable with C and C++ as well as a plethora of other languages and design paradigms
- Basic machining knowledge: lathes, mills, drill presses, routers, saws, etc.
- Knowledgeable with state-of-the-art machine learning techniques for classification, regression, density estimation, and other similar tasks
- Experienced with hand-optimizing programs for substantial runtime improvement
- Amateur metallurgist
- Nationally-known indoor kart racer (ranked 12th in 2019 Kart World Championship)
- Trained and active pollworker
- Mentor for *Google Summer of Code* for 10 years.
- Coffee snob (I use gradient descent for pour-over optimization; yes, you can judge me for that)

## Full Publication List

Journal publications.

1. "**mlpack 4: a fast, header-only C++ machine learning library**". R.R. Curtin, M. Edel, O. Shrit, S. Agrawal, S. Basak, J.J. Balamuta, R. Birmingham, K. Dutt, D. Eddelbuettel, R. Garg, S. Jaiswal, A. Kaushik, S. Kim, A. Mukherjee, N.G. Sai, N. Sharma, Y.S. Parihar, R. Swain, C. Sanderson. *Journal of Open Source Software*, vol. 8, issue 82, 2023.

2. "**The ensmallen library for flexible numerical optimization**", R.R. Curtin, M. Edel, R. Prabhu, S. Basak, Z. Lou, C. Sanderson. *The Journal of Machine Learning Research (JMLR)*, vol. 22, p. 1–6, 2021.

3. "**Functional Aggregate Queries with Additive Inequalities**", M.A. Khamis, R.R. Curtin, B. Moseley, H.Q. Ngo, X. Nguyen, D. Olteanu, M. Schleich. *ACM Transactions on Database Systems (TODS)* 45.4, pp. 1–41, 2020.

4. "**Practical Sparse Matrices in C++ with Hybrid Storage and Templated-Based Expression Optimisation**", C. Sanderson, R.R. Curtin. *Mathematical and Computational Applications*, vol. 24, no. 3, article 70, 2019.

5. "**mlpack 3: a fast, flexible machine learning library**", R.R. Curtin, M. Edel, M. Lozhnikov, Y. Mentekidis, S. Ghaisas, S. Zhang. *The Journal of Open Source Software*, vol. 3, issue 26, pp. 726, 2018.

6. "**Exploiting the structure of furthest neighbor search for fast approximate results**". R.R. Curtin, J. Echauz, A.B. Gardner. *Information Systems*, vol. 80, pp. 124–135, 2018.

7. "**`gmm_diag` and `gmm_full`: C++ classes for multi-threaded Gaussian mixture models and Expectation-Maximisation**", C. Sanderson, R.R. Curtin. *The Journal of Open Source Software*, vol. 2, 2017.

8. "**Armadillo: a template-based C++ library for linear algebra**", C. Sanderson, R.R. Curtin. *The Journal of Open Source Software*, vol. 1, issue 26, pp. 1–2, 2016.

9. "**Plug-and-play runtime analysis for dual-tree algorithms**", R.R. Curtin, D. Lee, W.B. March, P. Ram. *The Journal of Machine Learning Research (JMLR)*, vol. 16, pp. 3269–3297, 2015.

10. "**Dual-tree fast exact max-kernel search**", R.R. Curtin, P. Ram. *Statistical Analysis and Data Mining*, vol. 7, issue 4, p. 229–253, 2014.

11. "**mlpack: a scalable C++ machine learning library**". R.R. Curtin, J.R. Cline, N.P. Slagle, W.B. March, P. Ram, N.A. Mehta, A.G. Gray. *The Journal of Machine Learning Research*, vol. 14, p. 801–805, 2013.

Conference publications.

12. "**Intermediate N-Gramming: Deterministic and Fast N-Grams For Large N and Large Datasets**", R.R. Curtin, F. Lu, E. Raff, P. Ranade. Submitted to *The 40th Annual AAAI Conference on Artificial Intelligence (AAAI 2026)*.

13. "**Zipf-Gramming: Scaling Byte N-Grams Up To Production Sized Malware Corpora**", E. Raff, R.R. Curtin, D. Everett, R.J. Joyce, J. Holt. Accepted to *The 34th ACM International Conference on Information and Knowledge Management (CIKM 2025)*, 2025.

14. "**Bandicoot: A Templated C++ Library for GPU Linear Algebra**, R.R. Curtin, M. Edel, C. Sanderson. Accepted to *The 26th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2025)*, 2025.

15. "**Optimizing the Optimal Weighted Average: Efficient Distributed Sparse Classification**", F. Lu, R.R. Curtin, E. Raff, F. Ferraro, J. Holt. Accepted to *2025 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2025)*, 2025.

16. "**The Deconstructed Warehouse: An Ephemeral Query Engine Design for Apache Iceberg**", R.R. Curtin, J. Tagliabue. Accepted to *51st International Conference on Very Large Data Bases (VLDB 2025), Composable Data Management Systems Workshop (CDMS)*, 2025.

17. "**Armadillo: An efficient framework for numerical linear algebra**", C. Sanderson, R.R. Curtin. In *Proceedings of the 17th International Conference on Computer and Automation Engineering (ICCAE 2025)*, 2025.

18. "**FaaS and Furious: abstractions and differential caching for efficient data pre-processing**", J. Tagliabue, R.R. Curtin, C. Greco. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, p. 3562–3567, 2024.

19. "**High-Dimensional Distributed Sparse Classification with Scalable Communication-Efficient Global Updates**", F. Lu, R.R. Curtin, E. Raff, F. Ferraro, J. Holt. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*, p. 2037–2047, 2024.

20. "**An Approximation Algorithm for the Matrix Tree Multiplication Problem**", M. Abo Khamis, R.R. Curtin, S. Im, B. Moseley, H. Ngo, K. Pruhs, A. Samadian. *The 46th International Symposium on Mathematical Foundations of Computer Science (MFCS 2021)*, vol. 202, p. 6:1–6:14, 2021.

21. "**An Adaptive Solver for Systems of Linear Equations**", C. Sanderson, R.R. Curtin. In *The 14th International Conference on Signal Processing and Communication Systems (ICSPCS '20)*, pp. 1–6, 2020.

22. "**Rk-means: Fast Clustering for Relational Data**", R.R. Curtin, B. Moseley, H.Q. Ngo, X.L. Nguyen, D. Olteanu, M. Schleich. In *Proceedings of the 23rd Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, pp. 2742–2752, 2020.

23. "**On Coresets for Regularized Loss Minimization**", A. Samadian, K. Pruhs, B. Moseley, S. Im, R.R. Curtin. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, pp. 482–492, 2020.

24. "**On functional aggregate queries with additive inequalities**", M.A. Khamis, R.R. Curtin, B. Moseley, H.Q. Ngo, X.L. Nguyen, D. Olteanu, M. Schleich. In *Proceedings of the 2019 ACM SIGMOD/PODS International Conference on Management of Data*, pp. 414–431, 2019.

25. "**Detecting DGA domains with recurrent neural networks and side information**", R.R. Curtin, A.B. Gardner, S. Grzonkowski, A. Kleymenov, A. Mosquera. In *Proceedings of The 14th International Conference on Availability, Reliability, and Security*, pp. 1–10, 2019.

26. "**ensmallen: a flexible C++ library for efficient function optimization**", S. Bhardwaj, R.R. Curtin, M. Edel, Y. Mentekidis, C. Sanderson. In *Proceedings of the Systems for ML Workshop at NeurIPS 2018*, 2018.

27. "**A User-Friendly Hybrid Sparse Matrix Class in C++**", C. Sanderson, R.R. Curtin. In *Proceedings of The 2018 International Congress on Mathematical Software (ICMS 2018)*, pp. 422–430, 2018.

28. "**An open source C++ implementation of multi-threaded Gaussian Mixture Models, k-means and expectation maximisation**", C. Sanderson, R.R. Curtin. In *Proceedings of the 11th International Conference on Signal Processing and Communication Systems (ICSPCS 2017)*, pp. 1–8, 2017.

29. "**pfsuper: simulation-based prognostics to monitor and predict sparse time series**", J. Echauz, A.B. Gardner, R.R. Curtin, N. Vasiloglou, G.J. Vachtsevanos. In *Annual Conference of the Prognostics and Health Management Society 2017 (PHM '17)*, pp. 1–9, 2017.

30. "**A dual-tree algorithm for fast k-means clustering with large k**", R.R. Curtin. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 300–308, 2017.

31. "**Fast approximate furthest neighbors with data-dependent candidate selection**", R.R. Curtin, A.B. Gardner. In *Similarity Search and Applications 2016 (SISAP 2016)*, pp. 221–235, 2016.

32. "**Faster dual-tree traversal for nearest neighbor search**", R.R. Curtin. In *Similarity Search and Applications 2015 (SISAP 2015)*, pp. 77–89, 2015.

33. "**Collaborative filtering via matrix decomposition in mlpack**", S. Agrawal, R.R. Curtin, S. Ghaisas, M.R. Gupta. In *ICML 2015 Workshop on Machine Learning Open Source Software*, 2015.

34. "**An automatic benchmarking system**", M. Edel, A. Soni, R.R. Curtin. In *NIPS 2014 Workshop on Software Engineering for Machine Learning*, 2014.

35. "**Classifying broiler chicken condition using audio data**", R.R. Curtin, W. Daley, D.V. Anderson. In *GlobalSIP 2014 Symposium on Signal Processing Applications Related to Animal Environments*, 2014.

36. "**Tree-independent dual-tree algorithms**", R.R. Curtin, W.B. March, P. Ram, D.V. Anderson, A.G. Gray, C.L. Isbell, Jr. In *Proceedings of The 30th International Conference on Machine Learning (ICML '13)*, pp. 1435–1443, 2013.

37. "**Fast exact max-kernel search**", R.R. Curtin, P. Ram, A.G. Gray. In *SIAM International Conference on Data Mining (SDM '13)*, pp. 1–9, 2013. *Nominated for Best Paper Award.*

38. "**mlpack: a scalable C++ machine learning library**", R.R. Curtin, J.R. Cline, N.P. Slagle, M.L. Amidon, A.G. Gray. In *NIPS 2011 Workshop on Big Learning*, 2011.

39. "**Learning distances to improve phoneme classification**", R.R. Curtin, N. Vasiloglou, D.V. Anderson. In *Proceedings of the 2011 IEEE International Workshop on Machine Learning in Signal Processing (MLSP 2011)*, pp. 1–6, 2011.

Technical reports and other.

40. "**Flexible numerical optimization with ensmallen**", R.R. Curtin, M. Edel, R.G. Prabhu, S. Basak, Z. Lou, C. Sanderson. *arXiv preprint arXiv:2003.04103*, 2020.

41. "**A generic and fast C++ optimization framework**", R.R. Curtin, S. Bhardwaj, M. Edel, Y. Mentekidis. *arXiv preprint arXiv:1711.06581*, 2017.

42. "**Designing and building the mlpack open-source machine learning library**", R.R. Curtin, M. Edel. Submitted to *The Fourth International Conference of PUST (ICOPUST 2017)*—conference cancelled, 2017.

43. "**Detecting adversarial samples from artifacts**", R. Feinman, R.R. Curtin, S. Shintre, A.B. Gardner. *arXiv preprint arXiv:1703.00410*, 2017.

44. "**Improving dual-tree algorithms**", Ph.D. thesis, Georgia Institute of Technology, 2015.

45. "**Single-tree GMM training**", R.R. Curtin. *Technical report GT-CSE-2015-01*, Georgia Institute of Technology, School of Computational Science and Engineering, 2015.

**References available upon request.**