Subhabrata Majumdar

Senior Applied Scientist, Machine Learning Applied Research, Splunk
e-mail: zoom.subha@gmail.com; web: shubhobm.github.io

Expertise

Trustworthy machine learning: explainability, fairness, privacy, security and robustness.
Predictive models on big data: machine learning and statistical techniques, NLP.
Statistical machine learning: high-dimensional models, graphical models, feature selection, hypothesis testing.
Tools: Python (tensorflow, numpy, pandas, scikit-learn), R (data.table, ggplot, caret), C++, AWS, Hive/SQL, Hadoop, Spark.

Education

- PhD Statistics, University of Minnesota - Twin Cities, 2017. Advisor: Snigdhansu Chatterjee.
- MS Statistics, Indian Statistical Institute, 2012.
- BS Statistics, Indian Statistical Institute, 2010.

Experience

Splunk, Oct 2021-present

- As part of the Applied Research group, involved in R&D for innovative ML algorithms to power new Splunk products.
- Current projects include ML explainability for graph embeddings, and ML for cybersecurity.

AT&T Data Science and AI Research, Aug 2018-Sep 2021

- Led R&D in trustworthy ML methods in AT&T as part of the Data Science and AI Research group.
- As a founding member of AI governance core team implementing responsible ML practices across AT&T, proposed SIFT (System to Integrate Fairness Transparently) as an enterprise-level fairness monitoring and auditing framework.
- Technical lead for developing the internal SIFT platform, fairness analyses of business-critical use cases going through SIFT, and internal and external research collaborations.
- Major past projects in other areas involve engagement models, attribution, ad intelligence and content curation methods.
- Work led to 15+ research papers and 15+ patent filings in 3 years.

University of Florida Informatics Institute, July 2017-Aug 2018

- As a postdoctoral researcher Developed novel machine learning methods for integrative analysis of multimodal biological Omics data.
- Work led to publication in the Journal of Machine Learning Research.

IBM Research, May 2016-Aug 2016

- Research intern in the IBM Social Good fellowship program.
- Collaborated with scientists in the Data Science group, and Cary Institute of Ecosystem Studies to mine ecological data and devise cognitive algorithms that can determine which primates are carriers for the Zika virus in the wild.
- Work led to publication in the journal Epidemics.

Santander Consumer USA, May 2015-Aug 2015

- Worked with the Statistical Analysis team on implementing machine learning methods in Loss Forecast Score prediction.
- Used random forest, XGBoost and bagging to achieve performance improvements over current model in production.

University of Chicago Harris School of Public Policy, June 2014-Aug 2014

- As a Data Science for Social Good fellow, collaborated with the Chicago Department of Public Health and developed a ML model to predict childhood lead poisoning in Chicago.
- The model went on to be implemented for proactive home inspections, aided by a $3.5 million grant the CDPH received from the U.S. Department of Housing and Urban Development.
- Findings were published in 2015 KDD proceedings.

National Marrow Donor Program, June 2013-Aug 2013

- Developed a spatial algorithm for data-driven marrow donor recruitment for Leukemia patients with rare alleles.

Leadership

- Drove the development and adoption of trustworthy machine learning methods in AT&T through technical leadership.
- Co-founder and co-organizer of the Trustworthy ML Initiative (https://www.trustworthyml.org).
- Managed a team of 4 data scientists to collaborate with UNICEF officials on a volunteer project in developing a ML platform for air-quality prediction (https://www.solveforgood.org/proj/41). Work was presented in CHI-2021.

Selected publications *(see Google scholar for full list)*

- (Book) Pruksachatkun, Y., McAteer, M., **Majumdar, S.** Developing Trustworthy Machine Learning, in press, 2022, published by O'Reilly Media.
- (Book chapter) **Majumdar, S.** Fairness, Explainability, Privacy, and Robustness for Trustworthy Algorithmic Decision Making. In: Big Data Analytics in Chemoinformatics and Bioinformatics, in press, 2022, published by Elsevier.
- **Majumdar, S.**, Flynn, C., and Mitra, R. Evaluating Fairness in the Presence of Spatial Autocorrelation, NeurIPS 2021 AFCR workshop, PMLR 171, 6-18, 2022.
- **Majumdar, S.** and Michailidis, G. Joint Estimation and Inference for Data Integration Problems based on Multiple Multi-layered Gaussian Graphical Models, Journal of Machine Learning Research, 23(1), 1-53, 2022.
- *Derzsy, N., **Majumdar, S.**, Malik, R. An Interpretable Graph-based Mapping of Trustworthy Machine Learning Research, CompleNet-2021. **Alphabetical authors*.
- Farias, V., Timbo, F., Flynn, C., Machado, J., **Majumdar, S.,** Srivastava, D. Local Dampening: Differential Privacy for Non-numeric Queries via Local Sensitivity, PVLDB, 14(4), 521-533, 2020.
- Rustamov, R. and **Majumdar, S.** Intrinsic Sliced Wasserstein Distances for Comparing Collections of Probability Distributions on Manifolds and Graphs, arXiv preprint arXiv:2010.15285. 2020.
- *Dodwell, E., Flynn, C., Krishnamurthy, B., **Majumdar, S.** and Mitra, R. Towards Integrating Fairness Transparently in Industrial Applications, arXiv preprint arXiv:2006.06082, 2020. **Alphabetical authors*.
- Ghosh, A. and **Majumdar, S.** Ultrahigh-dimensional Robust and Efficient Sparse Regression using Non-Concave Penalized Density Power Divergence, IEEE Transactions on Information Theory, 66(12), 7812-7827, 2020.
- 7 authors. Computer-Assisted and Data Driven Approaches for Surveillance, Drug Discovery, and Vaccine Design for the Zika Virus, Pharmaceuticals, 12(4), 157, 2019.
- 11 authors. Confronting data sparsity to identify potential sources of Zika virus spillover infection among primates, Epidemics, 27, 59-65, 2019.
- **Majumdar, S.** and Basak, S. C. Beware of external validation!-A Comparative Study of Several Validation Techniques used in QSAR Modelling, Current Computer-Aided Drug Design, 14(4), 284-291, 2018.
- 10 authors. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning, KDD-2015, 2039-2047, 2015.

Major invited/refereed talks *(see my website for full list)*

09/2022 Keynote at 8th Indo-US Workshop on Mathematical Chemistry.
07/2022 NAACL 2022 Workshop on Trustworthy Natural Language Processing.
04/2022 University of Washington Responsibility in AI Systems and Experiences (RAISE) Lab.
12/2021 NeurIPS 2021 Workshop on Algorithmic Fairness through the Lens of Causality and Robustness.
06/2021 ASA 2021 Symposium on Data Science and Statistics.
05/2021 International Indian Statistical Association Conference-2021.
01/2021 Plenary talk in All India Council for Technical Education Faculty Development Programme.
12/2020 Data Science Salon Virtual.
11/2020 Indian Institute of Technology, Kanpur Data Science Seminar Series (3 invited talks).
02/2020 Invited talk on Data Science Research in AT&T in 3rd NISS Virtual Industry Career Fair.
11/2019 3rd AT&T Graduate Student Symposium.
03/2019 NYC Women in Machine Learning & Data Science meetup.
05/2018 Savvysherpa, Inc., Minneapolis, MN.
05/2018 International Indian Statistical Association Conference-2018, Gainesville, FL.

Major awards

- University of Minnesota (UMN) Martin-Buehler Award in Statistics 2016-2017, awarded by School of Statistics.
- UMN Interdisciplinary Doctoral Fellowship 2016-2017, awarded by the Graduate School.
- Best Student Paper in theory and methods, 2016 International Indian Statistical Association conference, Corvallis, OR.
- UMN Social Community Building Grant 2015-2016, awadred by Council of Graduate Students.
- UMN School of Statistics Travel Awards, 2014-2016.
- 5th International Workshop on Climate Informatics travel award (funded by National Science Foundation (NSF)), 2015.