SOUMIK PURKAYASTHA

E-mail: soumikp@umich.edu **Cell phone:** +1-734-881-5075
Website: soumikp.github.io **ORCID:** 0000-0002-3619-2804

Education

University of Michigan, Dept. of Biostatistics *Sep. 2019 - Apr. 2024 (expected)*
PhD in Biostatistics, Advisor: Peter X. K. Song
Rackham Predoctoral Fellowship awardee. *2023-*
MS in Biostatistics *(Sep. 2019 - Apr. 2021)* **GPA 4.0+/4.0**
Richard G. Cornell Fellowship awardee. *2020-21*

Indian Statistical Institute *Jul. 2017 - Jun. 2019*
MS in Statistics, with specialization in Biostatistics. **GPA 4.0/4.0**
Government of India-funded scholarship awardee. *2017-19*
Sabyasachi Roy Gold Medal awardee. *2019*

St. Xavier's College, Kolkata *Jul. 2014 - Jun. 2017*
BS, Major: Statistics. Minors: Math and computer science. **GPA 4.0/4.0**

Professional experience

Michigan Medicine, Ann Arbor, USA. **Research Assistant** *May 2020 -*
Perform **statistical analyses** in **SAS** and **R** for the NIH-funded Diabetes Foot Consortium. Built and presently maintain an **automated data-pooling and analysis pipeline** and an **RShiny**-based dashboard for faster dissemination of interactive **Plotly visualization** and **model-based** findings that is **accessible to clinicians**.

Apple Inc., Cupertino, USA. **AI-ML intern for Siri Data** *May 2021 - Aug. 2021*
Developed **Pytorch**-based natural language models to analyze **user speech patterns**. Built multi-level predictors of **user search intent** in **Python** to improve data quality for algorithm training and evaluation. Built Siri Search products by implementing **semi-supervised language models** on partially labelled user data in **Python**.

Walmart Labs, Bangalore, IND. **Statistical analyst intern** *May 2018 - Jul. 2018*
Worked on data query and analysis of very large data sets and improved existing online grocery **forecasting models** in **R** and **C++**. Built real-time spike detection models using **state space models** and **ensemble classification models** to find unusual demand patterns in stores in **R**.

Language, programming and statistical skills

Language skills: Bengali and English (native), Hindi (proficient at speaking, reading and writing).
Programming languages and frameworks: Python, R, C++, SQL, SAS and Snakebite (for Hadoop).
Summary of statistical skills:

- Handle large tracts of data (cleaning, processing, and quality control) using **Hadoop** and **SQL**.
- Provide insights on **experimental design** and perform **statistical analyses** in **R**, **Python**, **C++**, **SAS**.
- Develop interactive visualization and tabulation tools using **RShiny**, **Plotly** and **Tableau**.

Professional and volunteer service

Journal peer review: Annals of Applied Statistics (2022), New England Journal of Statistics in Data Science (2022), and PLOS One (2021).

Professional affiliations: International Biometric Society, Western North American Region (WNAR) (2022+), American Statistical Association (2021+), Institute of Mathematical Statistics (2021+), International Biometric Society, Eastern North American Region (ENAR) (2021+)

Statistics in the Community **Co-president** (*May 2022 -*), **Member** (*Sep. 2021 -*)

STATCOM is a community outreach consultancy program provided by graduate students in data organization, analysis, and interpretation. STATCOM provides free consulting services for multiple community partners such as:

– The Michigan Center for Youth Justice to understand the patterns of special investigations and violations occurring in juvenile justice facilities throughout the state of Michigan.

– Poverty Solutions and the Detroit Housing Commission to reduce the number of evictions among families with children in Detroit by connecting people with financial assistance and case managers.

For my work with STATCOM, I was awarded the 2023 Rising Star Award by the University of Michigan.

Selected publications *h-index: 10 (Google scholar); † denotes equal contribution.*

– **Purkayastha, S.** & Song, P. X. K. (2023). fastMI: A fast and consistent copula-based nonparametric estimator of mutual information. *The Journal of Multivariate Analysis (105270)*. doi: 10.1016/j.jmva.2023.105270.

– Salvatore, M.[†], **Purkayastha, S.**[†], Ganapathi, L., Bhattacharyya, R., Kundu, R., Zimmermann, L., Ray, D., Hazra, A., Kleinsasser, M., Solomon, S., Subbaraman, R. & Mukherjee, B. (2022). Lessons from SARS-CoV-2 in India: A data-driven framework for pandemic resilience. *Science Advances (Vol. 8, Issue 24)*. American Association for the Advancement of Science (AAAS). doi: 10.1126/sciadv.abp8621.

– **Purkayastha, S.**, Kundu, R., Bhaduri, R., Barker, D., Kleinsasser, M., Ray, D. & Mukherjee, B. (2021). Estimating the wave 1 and wave 2 infection fatality rates from SARS-CoV-2 in India. *BMC Research Notes (Vol. 14, Issue 1)*. Springer Science and Business Media LLC. doi: 10.1186/s13104-021-05652-2.

– **Purkayastha, S.**, Bhattacharyya, R., Bhaduri, R., Kundu, R., Gu, X., Salvatore, M., Ray, D., Mishra, S. & Mukherjee, B. (2021). A comparison of five epidemiological models for transmission of SARS-CoV-2 in India. *BMC Infectious Diseases (Vol. 21, Issue 1)*. Springer Science and Business Media LLC. doi: 10.1186/s12879-021-06077-9.

– Tang, L., Zhou, Y., Wang, L., **Purkayastha, S.**, Zhang, L., He, J., Wang, F. & Song, P. X. K. (2020). A Review of MultiCompartment Infectious Disease Models. *International Statistical Review (Vol. 88, Issue 2, pp. 462513)*. Wiley. doi: 10.1111/insr.12402.

– Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., **Purkayastha, S.**, Halder, A., Rix, A., Barker, D., Kleinsasser, M., Zhou, Y., Bose, D., Song, P. X. K., Banerjee, M., Baladandayuthapani, V., Ghosh, P. & Mukherjee, B. (2020). Predictions, Role of Interventions, and Effects of a Historic National Lockdown in Indias Response to the COVID-19 Pandemic: Data Science Call to Arms. *Harvard Data Science Review, (Special Issue 1)*. doi: 10.1162/99608f92.60e08ed5.