SOUMIK PURKAYASTHA

E-mail: soumikp@umich.edu **Website:** soumikp.github.io **Phone:** +1-734-881-5075

Education

University of Michigan, Dept. of Biostatistics *Sep. 2019 - Apr. 2024 (expected)*

PhD in Biostatistics, Advisor: Peter X. K. Song

MS in Biostatistics *(Sep. 2019 - Apr. 2021)* **GPA 4.0+**

Awarded 2021 Richard G. Cornell Fellowship.

Indian Statistical Institute *Jul. 2017 - Jun. 2019*

MS in Statistics, with specialization in Biostatistics. **GPA 4.0**

Awarded 2017-19 Government of India-sponsored scholarship.

Awarded 2019 Sabyasachi Roy Gold Medal.

St. Xavier's College, Kolkata *July 2014 - June 2017.*

BS, Major: Statistics. Minors: Math and computer science. **GPA 4.0**

Standardized test scores

GRE	Score: 332/340 (V: 163, Q: 169, AWA: 5.0)	*Oct. 2018*
TOEFL	Score: 120/120 (R: 120, L: 120, S: 120, W: 120)	*Oct. 2018*
ISI-MS degree qualification	All-India rank: 11	*May 2017*
IIT-MS degree qualification	All-India rank: 10	*Feb. 2017*

Professional skills

Language skills: Bengali and English (native), Hindi (proficient at speaking, reading and writing).

Programming Languages: Python, R, C++, SQL, SAS.

Frameworks: Pyspark (for Spark), Snakebite (for Hadoop), Sklearn, Scikit, Pandas, NumPy.

Summary of statistical skills:

- Handle large tracts of data (cleaning, processing and quality control) using **Hadoop** and **SQL**.
- Provide insights about **experimental design** and perform **statistical analyses** (using a range of supervised and unsupervised learning methods for regression and classification) in **R**, **Python** or **SAS**.
- Develop interactive visualization and tabulation tools using **RShiny**, **Plotly** and **Tableau**.

Professional experience

Apple Inc., Cupertino, USA. **AI-ML intern for Siri Data** *May - 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. Helped build Siri Search products by leveraging human annotation data, implemented **semi-supervised language models** on unlabelled user data in **Python**.

Walmart Labs, Bangalore, IND. **Statistical analyst intern** *May - 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 interactive apps using **RShiny**, with special emphasis on data visualisation using **Plotly**. Built real-time spike detection models using **state space models** and **ensemble classification models** to find unusual demand patterns in stores in **R**.

Professional and volunteer service

Manuscript review *May 2021 +*

- Annals of Applied Statistics, New England Journal of Statistics in Data Science and PLOS One.

Memberships *May 2021 +*

- International Biometric Society, Institute of Mathematical Statistics and American Statistical Association.

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

STATCOM is a community outreach program provided by graduate students in data organization, analysis, and interpretation. STATCOM is involved with multiple community partners in the Southeast Michigan area 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.
- 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.

Selected publications

h-index: 9 (Google scholar); † denotes equal contribution. Citation counts accessed on 02/23/2023.

- **Purkayastha, S.** and Song, P.X.K. *fastMI: a fast and consistent copula-based estimator of mutual information.* 2022. **Under peer-review.**
- **Purkayastha, S.** and Song, P.X.K. *Asymmetric predictability in causal discovery: an information theoretic approach.* 2022. **Under peer-review.**
- Salvatore, M.†, **Purkayastha, S.**†, [12 authors] *Lessons from SARS-CoV-2 in India: A data-driven framework for pandemic resilience.* **Science Advances**, 8(24), 2022. ***Cited by 42 independent sources.***
- **Purkayastha, S.**, [7 authors] *Estimating the wave 1 and wave 2 infection fatality rates from SARS-CoV-2 in India.* **BMC Research Notes** 14(262), 2021. ***Cited by 25 independent sources.***
- **Purkayastha, S**, [9 authors] *A comparison of five epidemiological models for transmission of SARS-CoV-2 in India.* **BMC Infectious Diseases**, 533, 2021. ***Cited by 28 independent sources.***
- Salvatore, M., Basu, D., Ray, D., Kleinsasser, M., **Purkayastha, S** [7 authors] *Comprehensive public health evaluation of lockdown as a non-pharmaceutical intervention on COVID-19 spread in India: national trends masking state-level variations.* **BMJ Open**, 10(12), 2021. ***Cited by 42 independent sources.***
- Tang, L., Zhou, Y., Wang, L., **Purkayastha, S.**, . . . [8 authors] *A Review of Multi-Compartment Infectious Disease Models.* **International Statistical Review** 88(2), 2020. ***Cited by 75 independent sources. Top Cited Article for 2020-21 in International Statistical Review.***
- **Purkayastha, S.**, Salvatore, M. and Mukherjee, B. *Are women leaders significantly better at controlling the contagion during the COVID-19 pandemic?* **Journal of Health and Social Sciences** 5(2), 2020. ***Cited by 29 independent sources.***
- Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., **Purkayastha, S.**, [18 authors] *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, 2020. ***Cited by 148 independent sources.***