Rachit Sabharwal

10211 Camden Garden Lane, Katy, Texas, 77494 rachit-sabharwal@outlook.com (585) 281-1928
www.rachitsabharwal.me rachitest

Technical Skills

Machine Learning: Scikit-learn, Pytorch, Tensorflow, Transfer Learning, LLMs (Gemini, Haiku, o1), Graph NNs (PyG), Reinforcement Learning (OpenAI Gymnasium)

Programming Languages: Python, R, Javascript

Databases: RDBMS (PostgreSQL, SQLite, MySQL), NoSQL DBMS (MongoDB, Elasticsearch, Neo4J), BigQuery

Cloud and Distributed Computing: AWS, GCP, Azure, On-Prem HPC

DevOps: Git, GitHub, GitLab, Docker, GitHub/GitLab CI/CD, Jenkins

Workflow Orchestration: Airflow, Prefect, Cron

Relevant Experience

Sept 2022 - Oct 2024 **Biostatistics and Data Science - Graduate Research Assistant (Doctoral)**, The University of Texas Health Science Center at Houston, School of Public Health – Houston, TX

- Established DevOps practices, architected repository structures, and implemented CI/CD pipelines. Designed and implemented ETL pipelines for datasets of various sizes, achieving a **24x speed increase over prior solution leading to faster model iteration and deliverable generation.**
- Applied statistical models and machine learning techniques for data analysis on vaccine non-response and Covid-19 pediatric comorbidities. **Contributed to the publishing of three peer-reviewed articles, with two more under review.**
- Developed a weekly report generation pipeline with CI/CD, ensuring seamless integration of new data. Planned and enacted a comprehensive test suite, **increasing team lead confidence and reducing publishing turnaround times.**

May 2021 - Jan 2022 **Research and Early Development, Development Sciences & Informatics - Informatics Intern**, Genentech – San Francisco, CA

- Contributed to the development of frameworks for an in-house data annotation tool, utilizing advanced deep learning NLP models and custom processing pipelines. Created a efficient document tokenization framework **allowing for fast and high-quality data annotation for NLP model building** and downstream informatics systems.
- Engineered a neural network using deep transfer learning, to predict adverse drug events. Focusing on Drug-Induced Liver Injury created a **low parameter model that matched the prediction accuracy of the SOTA model.**
- Created a Knowledge Graph with Neo4j and a Graph Neural Network using NetworkX and PyTorch to generate gene expression signature-likes for drugs. Began the creation of a knowledge repository — an accessible data source which **reduced the need to manually sift through dense primary sources.**

June 2020 - Aug 2020 **Data Engineering Intern**, Bristol Myers Squibb – San Francisco, CA

- Developed a multifeatured patent recommendation web application. The application implemented a mix of classical NLP algorithms (TF-IDF, BM 25) and deep learning algorithms (BERT). **Improved bench scientists' productivity by streamlining patent search and analysis workflow.**
- Engineered and maintained performant ETL pipelines with Python and Apache Airflow. Undertook extensive data cleaning and data wrangling on datasets of varying sizes (small, medium, and large) **ensuring data quality and readiness for analysis at a high velocity.**
- Designed and maintained robust relational and graph databases in PostgreSQL and Neo4j, **optimizing data storage and retrieval to support various internal projects and analyses.**

Education

June 2023 - present **The University of California, Berkeley**, Certificate in Software Development and Programming – Berkeley, CA

Aug 2022 - present **The University of Texas Health Science Center at Houston**, Doctor of Philosophy in Biostatistics – Houston, TX

- Advanced Certificate in Data Science

Jan 2020 - May 2022 **The University of Texas Health Science Center at Houston**, Master of Science in Biostatistics – Houston, TX

- Thesis — BioRec: A Biomedical Recommendation System for Academic Conferences and Journals
- Certificate in Data Science

Sept 2014 - May 2018 **University of Rochester**, Bachelor of Science in Environmental Health – Rochester, NY

- Minor in Psychology

Publications

Aug 2024 **Factors associated with elevated SARS-CoV-2 immune response in children and adolescents**

Messiah SE, Abbas R, Bergqvist E, Swartz MD, Talebi Y, **Sabharwal R**, Han H, Valerio-Shewmaker MA, DeSantis SM, Yaseen A, Gandhi HA, Amavisca XF, Ross JA, Padilla LN, Gonzalez MO, Wu L, Silberman MA, Lakey D, Shuford JA, Pont SJ, Boerwinkle E

10.3389/fped.2024.1393321 (Frontiers in Pediatrics)

May 2024 **Baseline characteristics of SARS-CoV-2 vaccine non-responders in a large population-based sample**

Yaseen A, DeSantis SM, **Sabharwal R**, Talebi Y, Swartz MD, Zhang S, Leon Novelo L, Pinzon-Gomez CL, Messiah SE, Valerio-Shewmaker M, Kohl HW 3rd, Ross J, Lakey D, Shuford JA, Pont SJ, Boerwinkle E

10.1371/journal.pone.0303420 (PLoS One)

Apr 2024 **An Interactive Online Dashboard with COVID-19 Trends and Data Analysis in Northeast and South Texas**

Zhang Z, **Sabharwal R**, Lee M, Zhang K, McGaha P, Crum M, Bauer C, Fisher-Hoch SP, McCormick JB, Reininger BM, Thomas S, Guajardo E, Pinon D, Yaseen A

research.ebsco.com/linkprocessor/plink?id=894625e1-7146-30bf-aa2c-9f5637dac41e (Texas Public Health Journal)

Oct 2023 **Long-term immune response to SARS-CoV-2 infection and vaccination in children and adolescents**

Messiah SE, Talebi Y, Swartz MD, **Sabharwal R**, Han H, Bergqvist E, Kohl HW 3rd, Valerio-Shewmaker M, DeSantis SM, Yaseen A, Kelder SH, Ross J, Padilla LN, Gonzalez MO, Wu L, Lakey D, Shuford JA, Pont SJ, Boerwinkle E

10.1038/s41390-023-02857-y (Pediatric Research)

June 2023 **Scholarly recommendation systems: a literature survey**

Zhang Z, Patra BG, Yaseen A, Zhu J, **Sabharwal R**, Roberts K, Cao T, Wu H

10.1007/s10115-023-01901-x (Knowledge and Information Systems)

May 2021 **Data Cleaning for eCommerce: Standardizing Data Handling Practices for eCommerce Datasets**

Sabharwal R

Procter & Gamble, Internal White Paper