Rachit Sabharwal

10211 Camden Garden Lane, Katy, Texas, 77494 ✉ rachit-sabharwal@outlook.com

☎ (585) 281-1928 rachitest

Professional Summary

Quantitative researcher and data scientist with a PhD background in statistics and machine learning. Specializes in building predictive models, automating complex workflows, and quantifying model uncertainty and risk. Seeking to apply advanced analytical skills to drive data-informed strategies and generate value in the financial sector.

Education

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

- Advanced Certificate in Data Science

The University of California, Berkeley *Berkeley, CA*
Certificate in Software Development and Programming *June 2023 - Apr 2025*

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

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

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

- Minor in Psychology

Experience

Research & Development Intern, MiLOS (Machine Learning, Optimization, & Statistics), Engineering & Process Sciences, Core R&D *Lake Jackson, TX* *May 2025 - Aug 2025*
The Dow Chemical Company

- Developed and deployed an R application to automate 50% of the Life Cycle Assessment workflow, creating a projected $15M in annual operational savings and freeing up significant analyst time.
- Researched and compared frequentist and Bayesian uncertainty quantification methods for machine learning models, delivering a framework to assess model reliability and risk under noisy, real-world data conditions.
- Communicated complex quantitative findings on model performance and business impact to diverse audiences, including senior leadership, securing buy-in for project continuation.

Graduate Research Assistant (Doctoral) *Dallas, TX*
The University of Texas Health Science Center at Houston, School of Public Health - Dallas Campus *Feb 2025 - present*

- Engineered and implemented a full CI/CD and DevOps framework, which reduced testing time and streamlined development, increasing team productivity and code reliability.
- Processed and integrated complex, sensitive datasets using Python and R, establishing a clean data foundation for subsequent predictive modeling and analysis.
- Designed and executed a comprehensive testing suite for the entire data pipeline, ensuring data integrity and model accuracy from ingestion to final reporting.
- Developed and maintained dynamic dashboards to monitor key performance indicators for a large-scale clinical trial, providing stakeholders with real-time insights for decision-making.

Biostatistics and Data Science - Graduate Research Assistant (Doctoral) *Houston, TX*
The University of Texas Health Science Center at Houston, School of Public Health *Sept 2022 - Jan 2025*

- Designed and implemented robust ETL pipelines for datasets of varying scale, increasing data processing efficiency and reliability for downstream analysis.
- Applied advanced statistical and machine learning models to complex biomedical data, uncovering key insights into vaccine efficacy and disease comorbidity.
- Automated the generation of weekly research reports through a CI/CD pipeline, ensuring stakeholders received timely and accurate updates.
- Co-authored four peer-reviewed journal articles, successfully translating complex analytical findings into impactful scientific publications.

Biostatistics and Data Science - Graduate Research Assistant (Master's) *Houston, TX*
The University of Texas Health Science Center at Houston, School of Public Health *Feb 2020 - Aug 2022*

- Maintained public-facing COVID-19 dashboards using Python and Tableau, providing critical real-time data to health officials and the public.
- Developed and deployed web-based recommender systems on Streamlit and Heroku, enhancing user engagement and content discovery at academic conferences.
- Engineered and maintained ETL pipelines to power real-time dashboards and recommender systems, ensuring high data availability and performance.
- Conducted in-depth literature reviews on NLP and recommendation systems, informing model selection and development strategy for multiple projects.

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

- Developed a deep transfer learning model to predict adverse drug events, creating a novel framework for assessing product risk and safety.
- Engineered a Graph Neural Network to model complex relationships within biomedical data, enabling the generation of predictive signatures to identify high-potential drug candidates.
- Conducted a comparative analysis of ETL frameworks (Airflow, Prefect, Luigi), delivering a data-driven recommendation that was adopted to standardize the team's NLP pipelines.
- Presented complex research on GNNs and Transfer Learning to technical and business stakeholders, influencing the adoption of new modeling techniques.

Consumer & Market Knowledge - Advanced Analytics Co-Op *Cincinnati, OH*
Procter & Gamble *Jan 2021 - May 2021*

- Built predictive models to forecast consumer behavior, delivering key insights into market dynamics that informed marketing strategy and resource allocation.
- Leveraged parallel computing frameworks (Dask, Modin) to analyze massive datasets, identifying key market trends and drivers of retailer performance.
- Architected and maintained scalable ETL pipelines on Google Cloud Platform, ensuring a timely and reliable data flow for all downstream analytics and modeling efforts.
- Led the adoption of modern DevOps practices, implementing unit testing, containerization (Docker), and agile methodologies (Jira) to improve team velocity and code quality

Biostatistics and Data Science - Teaching Assistant *Houston, TX*
The University of Texas Health Science Center at Houston, School of Public Health *Sept 2020 - Dec 2020*

- Instructed a class of 20 graduate students on foundational data science programming concepts in R and Python, improving overall class comprehension and skill acquisition.
- Developed and delivered curriculum modules on key data science libraries and paradigms, including Tidyverse, Pandas, and functional programming.
- Designed and graded all course assignments and exams, providing constructive feedback to foster student development.

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

- Developed a full-stack patent recommendation application that significantly improved the research workflow efficiency for research scientists.
- Engineered and maintained automated ETL pipelines using Python and Airflow, ensuring reliable and timely data for the recommendation engine.
- Designed and administered PostgreSQL and Neo4j databases to efficiently store and query complex patent and scientific data.
- Researched state-of-the-art Information Retrieval and NLP models (e.g., BERT variants), informing the technical direction of the patent recommendation system.

Honors and Awards

Delta Omega Honors Society: Alpha Iota Chapter

Tau Sigma Honors Society: Beta Rho Chapter

Rochester Innovation Grant: University of Rochester

Innovation and Creativity Award: Rochester Institute of Technology

Certifications

Good Clinical Practice (GCP) *CITI Program*
Jan 2025

Group 1 Biomedical Researcher and Key Personnel *CITI Program*
Mar 2023

Group 2 Social and Behavioral Researchers and Key Personnel *CITI Program*
Mar 2023

Data Acquisition and Management *CITI Program*
Oct 2020

Big Data Foundations - Level 1 *IBM*
May 2020

Big Data Foundations - Level 2 *IBM*
May 2020

Data Science Math Skills *Duke University (Coursera)*
May 2020

AWS Machine Learning *AWS (Coursera)*
May 2020

Google Cloud IAM and Networking *Google Cloud (Coursera)*
May 2020

Machine Learning *Stanford University (Coursera)*
May 2020

Hadoop Foundations - Level 1 *IBM*
May 2020

Spark - Level 1 *IBM*
May 2020

Publications

Trust and Uncertainty Quantification in Machine Learning Models Under Measurement Error Aug 2025
Sabharwal R
The Dow Chemical Company, Internal White Paper

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

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

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

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

Scholarly recommendation systems: a literature survey June 2023
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)

Data Cleaning for eCommerce: Standardizing Data Handling Practices for eCommerce Datasets May 2021
Sabharwal R
Procter & Gamble, Internal White Paper

Skills

Languages: English (Native/Bilingual), Hindi (Native/Bilingual), French (Intermediate)

Work Authorization: US Citizenship, Canadian Citizenship

Technical Skills

Machine Learning: Scikit-learn, TidyModels, Pytorch, Tensorflow, Raytune, Optuna, Huggingface, JAX

Programming Languages: Python, R, Javascript, C, Java, HTML, CSS, SAS, MATLAB

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

Cloud and Distributed Computing: AWS (AWS HPC), GCP, Azure, Spark, Hadoop, Slurm, On-Prem HPC

DevOps: Git, GitHub, GitLab, Docker, GitHub/GitLab CI/CD, Jenkins, Kubernetes, Jira, Confluence

Workflow Orchestration: Airflow, Prefect, Cron, Luigi

Frameworks and Platforms: Shiny, Streamlit, FastAPI, Django, Flask, Heroku, Replit, Great Expectations, PyTest

Tooling: VSCode, RStudio, Quarto, Jupyter, PyCharm, CLion, IntelliJ IDEA, Confluence, Slack, Tableau, Power BI, Stata, DBeaver

Operating Systems: Windows, Linux (Ubuntu, and Mint), MacOS

General Computing: Microsoft Office, Google Workspace