

# Rachit Sabharwal

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

☎ (585) 281-1928    🌐 www.rachitsabharwal.me    📺 rachitest

## Education

---

**The University of California, Berkeley**

*Certificate in Software Development and Programming*

*Berkeley, CA*

*June 2023 - Apr 2025*

**The University of Texas Health Science Center at Houston**

*Doctor of Philosophy in Biostatistics*

*Houston, TX*

*Aug 2022 - present*

- Advanced Certificate in Data Science

**The University of Texas Health Science Center at Houston**

*Master of Science in Biostatistics*

*Houston, TX*

*Jan 2020 - May 2022*

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

**University of Rochester**

*Bachelor of Science in Environmental Health*

*Rochester, NY*

*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*

- Authored manuscript \_"Trust and Uncertainty Quantification in Machine Learning Models Under Measurement Error"\_ comparing the difference between frequentist and bayesian approaches in creating "trustworthy" models for noisy (real-world like) simulated data
- Created R application allowing users to automate upto 50% of LCA ( Life Cycle Assesment) analysis, estimated to be worth ~15MM annually across Dow
- Successfully presented literature review and experimental results to both technical and non-technical stakeholders

**Graduate Research Assistant (Doctoral)**

*Dallas, TX*

*The University of Texas Health Science Center at Houston, School of Public Health*

*Feb 2025 - present*

*- Dallas Campus*

- Establish good DevOps practices, architect repository structures, author comprehensive documentation, and implement CI/CD pipelines to streamline development and testing processes
- Perform data cleaning, wrangling and integration on medium-sized datasets containing PII using Microsoft Excel, Python, and R
- Plan and enact a test suite for both code and data spanning the entire data engineering and data science life-cycle from data ingestion to report generation
- Author and edit manuscripts

**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 ETL pipelines for all sizes of datasets (small, medium, and large), ensuring efficient data munging and integration
- Established good DevOps practices, architected repository structures, authored comprehensive documentation, and implemented CI/CD pipelines to streamline development and testing processes

- Applied a mix of classical statistical models and advanced machine learning techniques for comprehensive data analysis on varied topics including — vaccine non-response, serostatus, pediatric comorbidities arising from Covid-19, etc.
- Developed a weekly report generation pipeline incorporating CI/CD, ensuring seamless integration of new data and automated report updates
- Planned and enacted a test suite for both code and data spanning the entire data engineering and data science life-cycle from data ingestion to report generation
- Authored and edited multiple manuscripts, contributing to the publication of at least four peer-reviewed journal articles
- Utilized R (tidyverse, tidymodels, data.table, gt), Python (polars, pandas, duckdb, statsmodels, scikit-learn, PyTorch, TensorFlow), Quarto, SQLite, Git, Github, and Gitlab for various projects

### **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*

- Performed data cleaning, wrangling and integration on medium-sized datasets containing PII using Microsoft Excel, Python, and R
- Maintained Covid-19 dashboards using Python and Tableau for the [Texas Covid-19 Dashboard Project](#)
- Created, deployed, and maintained accessible and responsive web apps on multiple platforms (Streamlit and Heroku) for academic conference recommender systems
- Built and serviced high content ETL pipelines using Python, R, and Cron to feed recommender systems and Covid-19 dashboards
- Created recommendation systems models using Python and Scikit-Surprise
- Conducted literature reviews on research concerning Recommendation Systems, and Natural Language Processing models such as word2vec and doc2vec

### **Research and Early Development, Development Sciences & Informatics**

*San Francisco, CA*

#### **- Informatics Intern**

*May 2021 - Jan 2022*

*Genentech*

- Used Deep Transfer Learning via PyTorch and Raytune to create a neural network to predict adverse events for drugs
- Created a Knowledge Graph with Neo4j and a Graph Neural Network using NetworkX and PyTorch to generate gene expression signature-likes for drugs
- Developed a framework for the tokenization of internal documents for ingestion into text-mining application
- Compared work flow management/ETL frameworks (Airflow vs. Prefect vs. Luigi) for use with all NLP pipelines and presented the results
- Conducted literature reviews on research concerning Graph Neural Networks, Transfer Learning, and Natural Language Generation
- Successfully presented literature review and experimental results to both technical and non-technical stakeholders

### **Consumer & Market Knowledge - Advanced Analytics Co-Op**

*Cincinnati, OH*

*Procter & Gamble*

*Jan 2021 - May 2021*

- Created predictive models, analytics, and visualizations that facilitated a deep understanding of consumer and shopper behaviors
- Used parallel computing (Dask and Modin) to develop both predictive and explanatory models enabling insights into market trends and retailer behavior
- Created and serviced big data ETL pipelines utilizing the Google Cloud Platform, Python, and Apache Airflow
- Upgraded teams nascent DevOps by implementing unit-testing via Pytest and Great Expectations, containerization via Poetry and Docker/Kubernetes, and agile via Jira and Confluence
- Successfully presented experimental results and visualizations to both technical and non-technical stakeholders

## **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*

- Teaching assistant for PH 1998 — Introduction to Statistical and Data Science Programming
- Assisted in instructing a class of 20 students both individually and in groups
- Provided after-class instruction, individually and in groups, including hands-on technical demonstrations for both curricular and extracurricular topics
- Topics taught included (but were not limited to) — Data Types and Structures (R & Python), Loops (R & Python), Functional Programming (R & Python), NumPy (Python), Pandas (Python), Matplotlib (Python), Tidyverse (R), GGPlot2 (R)
- Created and assessed all assignments and exams

## **Data Engineering Intern**

*San Francisco, CA*

*Bristol Myers Squibb*

*June 2020 - Aug 2020*

- Utilized Python, HTML, CSS, and Javascript in creating a multifeatured patent recommendation app to significantly improve scientists' workflow
- Developed and serviced ETL pipelines using Python and Apache Airflow
- Performed data cleaning and data wrangling with R and Python on multiple datasets of varying sizes (small, medium, and large)
- Designed and maintained both relational and graph databases in PostgreSQL and Neo4j
- Conducted literature reviews on research concerning Recommendation Systems, Information Retrieval Systems, and BERT and BERT variations (BioBERT, SciBERT, etc.)
- Successfully presented literature review and experimental results to both technical and non-technical stakeholders

## **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)*

|                                 |   |
|---------------------------------|---|
| Google Cloud IAM and Networking | May 2020<br>Google Cloud (Coursera)<br>May 2020 |
| Machine Learning                | Stanford University<br>(Coursera)<br>May 2020   |
| Hadoop Foundations - Level 1    | IBM<br>May 2020                                 |
| Spark - Level 1                 | IBM<br>May 2020                                 |

## Publications

---

|   |           |
|---|-----------|
| <b>Factors associated with elevated SARS-CoV-2 immune response in children and adolescents</b>  | Aug 2024  |
| Messiah SE, Abbas R, Bergqvist E, Swartz MD, Talebi Y, <b>Sabharwal R</b> , 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<br><a href="https://doi.org/10.3389/fped.2024.1393321">10.3389/fped.2024.1393321</a> (Frontiers in Pediatrics)           |           |
| <b>Baseline characteristics of SARS-CoV-2 vaccine non-responders in a large population-based sample</b>   | May 2024  |
| Yaseen A, DeSantis SM, <b>Sabharwal R</b> , 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<br><a href="https://doi.org/10.1371/journal.pone.0303420">10.1371/journal.pone.0303420</a> (PLoS One)  |           |
| <b>An Interactive Online Dashboard with COVID-19 Trends and Data Analysis in Northeast and South Texas</b>  | Apr 2024  |
| Zhang Z, <b>Sabharwal R</b> , 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<br><a href="https://research.ebsco.com/linkprocessor/plink?id=894625e1-7146-30bf-aa2c-9f5637dac41e">research.ebsco.com/linkprocessor/plink?id=894625e1-7146-30bf-aa2c-9f5637dac41e</a> (Texas Public Health Journal) |           |
| <b>Long-term immune response to SARS-CoV-2 infection and vaccination in children and adolescents</b>  | Oct 2023  |
| Messiah SE, Talebi Y, Swartz MD, <b>Sabharwal R</b> , 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<br><a href="https://doi.org/10.1038/s41390-023-02857-y">10.1038/s41390-023-02857-y</a> (Pediatric Research)                                       |           |
| <b>Scholarly recommendation systems: a literature survey</b>  | June 2023 |
| Zhang Z, Patra BG, Yaseen A, Zhu J, <b>Sabharwal R</b> , Roberts K, Cao T, Wu H<br><a href="https://doi.org/10.1007/s10115-023-01901-x">10.1007/s10115-023-01901-x</a> (Knowledge and Information Systems)  |           |
| <b>Data Cleaning for eCommerce: Standardizing Data Handling Practices for eCommerce Datasets</b>  | May 2021  |
| <b>Sabharwal R</b><br>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, Raytune, Optuna, Pytorch, Tensorflow, Huggingface, JAX

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

**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:** 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