Last updated in September 2024

Rachit Sabharwal

10211 Camden Garden Lane, Katy, Texas, 77494 ✉ rachit-sabharwal@outlook.com ☎ (585) 281-1928

in rachit-sabharwal-b11346102 rachitest

Education

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

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

- Certificate in Data Science

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

- Minor in Psychology

Experience

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 – Oct 2024*

- Design and implement ETL pipelines for all sizes of datasets (small, medium, and large), ensuring efficient data munging and integration
- Establish good DevOps practices, architect repository structures, author comprehensive documentation, and implement CI/CD pipelines to streamline development and testing processes
- Apply 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.
- Develop a weekly report generation pipeline incorporating CI/CD, ensuring seamless integration of new data and automated report updates
- 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 multiple manuscripts, contributing to the publication of at least four peer-reviewed journal articles
- Utilize 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 - Informatics Intern *San Francisco, CA*
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

Publications

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

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

Recommender System for Academic Conferences Manuscript in Progress

Sabharwal R, Yaseen A

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

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

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

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 (UTH),

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