


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 <i>Certificate in Software Development and Programming</i>	<i>Berkeley, CA</i> <i>June 2023 – present</i>
The University of Texas Health Science Center at Houston <i>Doctor of Philosophy in Biostatistics</i> <ul style="list-style-type: none">◦ Advanced Certificate in Data Science◦ Expected Completion Date - 08/2026	<i>Houston, TX</i> <i>Aug 2022 – present</i>
The University of Texas Health Science Center at Houston <i>Master of Science in Biostatistics</i> <ul style="list-style-type: none">◦ Certificate in Data Science	<i>Houston, TX</i> <i>Jan 2020 – May 2022</i>
University of Rochester <i>Bachelor of Science in Environmental Health</i> <ul style="list-style-type: none">◦ Minor in Psychology	<i>Rochester, NY</i> <i>Sept 2014 – May 2018</i>

Experience

Biostatistics and Data Science - Graduate Research Assistant (Doctoral) <i>The University of Texas Health Science Center at Houston, School of Public Health</i> <ul style="list-style-type: none">◦ 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	<i>Houston, TX</i> <i>Sept 2022 – Oct 2024</i>
Biostatistics and Data Science - Graduate Research Assistant (Master's) <i>The University of Texas Health Science Center at Houston, School of Public Health</i> <ul style="list-style-type: none">◦ 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	<i>Houston, TX</i> <i>Feb 2020 – Aug 2022</i>

Processing models such as word2vec and doc2vec

**Research and Early Development, Development Sciences & Informatics
- Informatics Intern**
Genentech

*San Francisco, CA
May 2021 – Jan 2022*

- 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
Procter & Gamble

*Cincinnati, OH
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

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

*Houston, TX
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), Tidiverse (R), GGPlot2 (R)
- Created and assessed all assignments and exams

Data Engineering Intern
Bristol Myers Squibb

*San Francisco, CA
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

Group 1 Biomedical Researcher and Key Personnel	<i>CITI Program Mar 2023</i>
Group 2 Social and Behavioral Researchers and Key Personnel	<i>CITI Program Mar 2023</i>
Data Acquisition and Management	<i>CITI Program Oct 2020</i>
Big Data Foundations - Level 1	<i>IBM May 2020</i>
Big Data Foundations - Level 2	<i>IBM May 2020</i>
Data Science Math Skills	<i>Duke University (Coursera) May 2020</i>
AWS Machine Learning	<i>AWS (Coursera) May 2020</i>
Google Cloud IAM and Networking	<i>Google Cloud (Coursera) May 2020</i>
Machine Learning	<i>Stanford University (Coursera) May 2020</i>
Hadoop Foundations - Level 1	<i>IBM May 2020</i>
Spark - Level 1	<i>IBM May 2020</i>

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

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](https://doi.org/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](https://doi.org/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, 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), Big-Query

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