Rubab Zahra Sarfraz

rubabzsarfraz@gmail.com · rubabzsarfraz.com · linkedin.com/in/rubabzsarfraz

Summary

I'm a data science professional with over 6 years of experience in building data science teams and engineering data-driven solutions that serve over half a million users. My industry experience in solving real-world data challenges has motivated me to pursue research in the field of databases to build efficient systems that are responsible for human-database interactions.

Education

Lahore University of Management Sciences (LUMS) Aug. 2016 – Jun. 2018
M.S. in Computer Science *Lahore, Pakistan*

- Thesis: Measuring the Impact of Fake News in Developing Regions
- Advisor: Dr. Ihsan Ayyub Qazi

University of Engineering & Technology (UET) Oct. 2012 – Jun. 2016
B.Sc. in Computer Engineering (with honors) *Lahore, Pakistan*

- Thesis: Monitoring Traffic on Virtual Routers of OpenStack
- Advisor: Dr. Irfan Ullah Chaudhary
- GPA: 3.74/4.00

Publications

Rubab Zahra Sarfraz, Samar Haider. "Vizard: Improving Visual Data Literacy with Large Language Models". In *7th International Workshop on Big Data Visual Exploration and Analytics (BigVis) at VLDB 2024*. [Slides][Code][Video] [Demo]

Rubab Zahra Sarfraz. "Towards Semi-Supervised Data Quality Detection in Graphs". In *13th International Workshop on Quality in Databases (QDB) at VLDB 2024*. [Slides][Code][Video]

Nida Munawar, **Rubab Zahra Sarfraz**, Maria Costello, David Robinson, Colm Bergin, Elaine Greene. "Risk Factors and Outcomes of Delirium in Hospitalized Older Adults with COVID-19: A Systematic Review and Meta-Analysis". In *Aging and Health Research (2023)*.

Research Experience

Lahore University of Management Sciences (LUMS) Oct. 2016 – Jun. 2018
Mentor: Dr. Ihsan Ayyub Qazi *Lahore, Pakistan*

- Investigated Facebook's content moderation policies and censorship practices during controversial political events by designing an algorithmic audit experiment using bot accounts. Analyzed the moderation feedback, identified flaws in the platform's approach, and proposed solutions to increase user transparency.
- Conducted a survey of fake news detection tools from an interdisciplinary lens and evaluated their applicability and limitations in Pakistan to understand their societal impact in the Global South.
- Designed and executed two large-scale online surveys during Pakistan's 2018 elections, collecting over 1200 user responses and analyzing 50 key pieces of fake news using a mixed-methods approach to understand people's news consumption habits, identify vulnerable demographics, and measure the impact of fake news in developing countries.

University of Engineering & Technology Jun. 2013 – Jun. 2016
Mentor: Dr. Irfan Ullah Chaudhary *Lahore, Pakistan*

- Conducted a comparative study of functional programming languages (Haskell, F#) to evaluate their educational applications for teenagers, proposing strategies to improve beginner programming pedagogy.
- Designed and implemented a monitoring plugin for OpenStack's Monasca framework to track virtual router traffic, integrating real-time traffic statistics with Grafana for actionable insights, addressing critical gaps in OpenStack's network monitoring capabilities.

Work Experience

BridgeLinx Oct. 2022 – Present

Data Lead *Lahore, Pakistan*

- Established the company's data infrastructure from the ground up, utilizing Snowflake for data lakes, Prefect for quality-driven data pipelines, and Tableau/Metabase for front-end, adopted by **40%** of the workforce.
- Built a temporal geospatial data pipeline, boosting customer acquisition by **25%** across **150+ transportation lanes** through optimized operational reach and resource allocation.
- Incorporated data quality checklist in product sprints. Created and deployed a data quality bot, improving data completeness from **45% to 98%**, resulting in 30% faster decision-making across **9 teams** and **200+ data points**.
- Engineered a real-time bidding system employing advanced analytics to optimize Return on Capital (ROC) through predictive client settlement behavior analysis upon order booking.

Finja Jan. 2020 – Oct. 2022

Data Lead *Lahore, Pakistan*

- Researched the low-resourced, undocumented SME market, engineered a machine learning pipeline identifying 40+ predictive variables, reducing loan approval time by **35%**, while keeping the **NPLs below 0.5%** with **$3M issuances/month.**
- Advocated for and expanded the ML model's business evaluation metric from overall to category-wise NPLs to protect vulnerable MSMEs from market-driven penalties, using a real-time feedback loop for classifying bad loans into categories.
- Led a multidisciplinary team to launch Pakistan's first data-driven SECP-approved P2P lending, investment and payment engine, resulting in **$1M+** lending issuance to **3000+ retail stores**.
- Led a 10-member data science team, achieving a **40%** reduction in data processing time and transitioning 8 departments to fully adopt data-driven decision-making, increasing operational efficiency by **25%**.
- Instrumental in securing a **$9M Series A** funding by managing the data rooms, drafting and presenting the impact of data-driven tech and insights to investors.

Oct. 2018 – Dec. 2019

Data Engineer *Lahore, Pakistan*

- Architected a scalable data architecture for financial transactions (lending, payments, investments), optimizing for performance, security, and compliance across a multi-service platform serving **300K customers**.
- Built and managed a high-availability data lake and hybrid data pipelines, integrating real-time streaming and batch processing to prioritize workloads efficiently across **50+ dashboards**, **8 departments** and **3 products**.
- Developed a high-accuracy sanctions screening system in-house for KYC, enabling real-time fraud detection on incoming wallets and saving **$150K**. Successfully identified **two high-risk matches**, mitigating compliance risks and preventing financial losses.
- Enhanced search engine relevance for a B2C marketplace (from **40%** to **80%** relevance ratio) by improving vendor product data quality and using machine learning models in an adaptive pipeline with human-in-the-loop validation.

CERN Jul. 2018 – Sept. 2018

Intern, Software Engineer *Geneva, Switzerland*

- Spearheaded the test deployment Ceph clusters with Rook on Kubernetes, leveraging both OpenStack VMs and Ironic hosts.
- Defined and implemented robust evaluation metrics, demonstrating the superior latency and user-friendliness of the new cluster deployment approach, reducing setup time significantly.
- Enhanced the Orchestrator CLI with RGW support, enabling rapid S3 service provisioning in just **1-2 minutes**.
- Published a blog post with the findings in collaboration with CERN and Ceph team.

RedHat (via Outreachy) Dec. 2017 – Mar. 2018

Intern, Software Engineer *Remote*

- Improved distributed cluster management by adding performance dashboards. Pull requests: [1] [2] [3]
- Converted the decentralized architecture of Ceph Manager to a centralized one for improved coherence with the codebase.

Meta Feb. 2017 – Mar. 2017

Mentee, Software Engineer *Remote*

- Extended the Osquery open source project by implementing a virtual table in C++ for listing Python packages installed on a server.
- This functionality was actively used for securing Meta's data center servers against vulnerabilities introduced by PyPI in 2017.

Teaching Experience

Instructor, Introduction to Data Science — Summer 2023
Lahore School of Economics

Teaching Assistant, Advanced Operating Systems — Spring 2017
Lahore University of Management Sciences
Instructor: *Dr. Muhammad Hamad Alizai*

Teaching Assistant, Programming Fundamentals — Spring 2016
University of Engineering & Technology
Instructor: *Dr. Irfan Ullah Chaudhary*

Awards & Honors

Invited Participant, International Visitor Leadership Program (IVLP), U.S. Department of State — 2024

Finalist, USAID FDI Grant of $100K (on behalf of BridgeLinx) — 2024

Finalist, U.K. Climate Finance Accelerator (on behalf of BridgeLinx) — 2023

Runner Up, CERN Openlab Lightning Talks — 2018

Diversity Scholar, KubeCon by the Linux Foundation — 2018

Dean's Honor List, University of Engineering & Technology — 2012 – 2016

Community Involvement

Reviewer, IEEE VIS Workshop on Visualization for AI Explainability (VISxAI) — 2024

Board Member, Computer Science Alumni Network, LUMS — 2024

Advisor, Securities and Exchange Commission of Pakistan (for drafting regulations for P2P lending) — 2022

Open Source Contributor, Ceph Blog: Evaluating Ceph Deployments with Rook — 2018

Open Source Contributor, Docker Docs Hackathon. Pull requests: [1] [2] [3] [4] — 2017

Talks

"Elevating Trust in Your Data with Python", PyCon Pakistan — 2024

"Reshaping the Financial Sector with Artificial Intelligence", Information Technology University — 2022

"Debugging the Startup Space: Tech Careers", LUMS Women in Computing — 2022

"Best Practices in Operationalizing Machine Learning", AWS Startup CTO Forum — 2022

"Breaking the Bias", Systems Limited — 2021

"How Technology can be Your Best Friend", Meta Developers Circle — 2020

"Evaluating Ceph Deployment with Rook", CERN — 2018

Technical Skills

Expertise: Data Engineering, Data Architecture, Data Modeling, Data Product Management

Languages & Scripts: Python, C++, SQL, HTML, LaTeX

Machine Learning: Scikit-learn, PyTorch, Transformers

Data Lake & Quality: BigQuery, Snowflake, S3, Great Expectations, Pandera

Data Visualization: Tableau, Looker Studio, Streamlit

Data Infrastructure: GCP (Pub/Sub, Cloud Functions, Storage, Dataflow, SQL), AWS (SageMaker, Lambda), Prefect, FastAPI, Flask, Docker