RAJAT GUPTA

DATA SCIENTIST

CONTACT

+33-783454133

rajatgupta116@gmail.com

Berlin, Germany
(Open to Remote)

https://www.linkedin.com/in/dr-rajat-gupta-a4a77391/

github.com/rajat116

www.rajat-gupta.dev

SKILLS

Programming & Tools

- Python, C++, Bash, SQL, Git, Jupyter, Docker, Conda, Advanced Excel, FastAPI

Machine Learning & Deep Learning

- Regression, Classification, Clustering, PCA, Anomaly Detection, Autoencoders (VAE), CNNs, Generative Models, XGBoost, Neural Networks, LLMs, Retrieval-Augmented Generation (RAG)

Frameworks & Libraries

- Scikit-learn, PyTorch, TensorFlow, Keras, Pandas, NumPy, Matplotlib, MLflow, Evidently, sentence-transformers, Qdrant

Data & Analysis

- Hypothesis Testing, feature engineering, semantic search over structured data

Platforms & Environments

- GitHub, CUDA (GPU), HTCondor, SLURM, Docker, Singularity, MLflow, Mage, Airflow, Terraform, Streamlit, AWS (S3)

Soft skills:

- Storytelling with data
- Solid interpersonal and collaborative skills

AWARDS & HONORS

- Breakthrough Prize in Fundamental Physics (2025) : – awarded as part of the ATLAS Collaboration at CERN.
- Ranked among top 3% of researchers globally (AD Scientific Index, 2023)
- Gold Medalist (1st rank) – M.Sc. Physics, Panjab University

LANGUAGES

- English (Fluent)
- French (Basics)

Data Scientist with a PhD in Physics and expertise in machine learning, data analysis, and scientific computing. Experienced in building robust models for complex, noisy datasets; applying model distillation to compress and accelerate deep learning workflows; and leveraging LLMs, semantic search, and retrieval-augmented generation (RAG) for data-centric applications. Skilled in Python, SQL, and modern ML frameworks, and in building end-to-end ML and LLM systems with automated workflows, monitoring, and model management for reliable production deployment.

PROJECT HIGHLIGHTS

GitHub Anomaly Detection (End-to-End MLOps Project)

- Built a real-time anomaly detection system for GitHub activity using Isolation Forest and engineered actor-level features.
- Orchestrated ETL, drift detection, and scheduled retraining with Airflow; tracked runs and models with MLflow and Terraform-provisioned infrastructure.
- Deployed a FastAPI inference service in Docker on AWS with CI/CD, and built monitoring dashboards using Streamlit, Evidently, and Slack/email alerts.

OverPrompt (LLM-Powered Cricket Analytics Assistant)

- Built an LLM assistant that answers free-form cricket questions with data-backed insights using ball-by-ball stats, semantic search, and retrieval-augmented generation (RAG).
- Designed a layered entity-resolution and vector-search stack (exact, fuzzy, embedding-based) with sentence-transformers, Qdrant, and Streamlit for interactive comparisons and leaderboards.

WORK EXPERIENCE

Data Scientist — 2022 - PRESENT

ATLAS@LHC (CERN and University of Pittsburgh)

- Leading an international team of 10 researchers on an ML-based rare event detection project analyzing over 1B samples, targeting signals occurring in <0.1% of cases.
- Applied decision tree–based ML models to suppress dominant background noise, achieving a 13% improvement in sensitivity for rare signal detection.
- Designing model distillation workflows and co-developing FPGA-friendly ML models with hardware engineers to enable sub-3 μs, low-latency, real-time applications.
- Building data-driven compression solutions for edge memory platforms, integrating ML techniques for efficient storage and retrieval.
- Mentoring one PhD and two undergraduate students in applied machine learning, data analysis and communication.

Data Analyst — 2016 - 2022

CMS@LHC (CERN and Panjab University)

- Analyzed experimental datasets with millions of records to extract structured patterns and validate data-driven hypotheses.
- Coordinated multi-year projects involving data quality checks, statistical evaluation, and reporting of key metrics.
- Led communication and documentation for collaborative research.

EDUCATION

Ph.D. in Particle Physics — April 2022

(Panjab University)