RAJAT KAPGATE

Open to Relocation | +1 (812) 553-2822 | rkapgate@iu.edu | linkedin.com/in/rajat-kapgate | github.com/rajatk9962

PROFESSIONAL EXPERIENCE

Research Data Scientist | Indiana University School of Optometry, Bloomington, USA   **May 2024 – Present**

- Analyzed 500+ GB of infant data, performing **statistical analysis** to correlate **head and eye movement** for identifying potential early markers of **infant eye disorders**, as part of an NIH-funded study (**Grant EY032897**).
- Preprocessed time series using **IQR filtering, rolling sum, low-pass filtering,** and **non-max suppression**; integrated RigidMask deep learning model for motion segmentation, collectively enhancing head motion extraction accuracy by 40%.
- Performed **ANOVA** across five age groups, identifying statistically significant motor control variations with p-value less than 0.05.
- Engineered **advanced data visualizations**, including head movement reconstruction with Unity and Open3D, Autocorrelation, KDE, polar plots to analyze infant head dynamics.

Data Science Co-op | Boehringer Ingelheim Pharmaceuticals, Ridgefield, USA   **May 2024 – Nov 2024**

- Implemented a **FinOps** cost analytics dashboard with **Streamlit**, integrating cloud cost data for better visibility. Monitored Jenkins CI/CD pipelines and performed **root cause analysis**, uncovering inefficiencies and cutting compute costs by **$200K+.**
- Accomplished a **70%** reduction in **reporting** turnaround for the drug Jardiance by leveraging **Large Language Models, LangChain** CSV **agents**, **Azure** Chat API, and Python-pptx**,** leading to an increase in decision making efficiency.
- Orchestrated an ETL pipeline to process **2M+** drug price data points from the Nuro API, optimizing SQL workflows with window functions on AWS RedShift, automating with cron jobs, and storing results in AWS S3.
- Devised a **Retrieval-Augmented Generation (RAG)** system on proprietary organizational data using **Azure GPT-4o** and **FAISS** for vector-based similarity search, cutting research effort by 40%.

Data Scientist| TCS Research, Mumbai, India   **Jun 2021 – Aug 2023**

- Developed a **deep learning** model for Indian Sign Language to text conversion, **optimizing** architecture and **reducing** parameters from **200K+ to 478** for efficient CPU deployment.
- Orchestrated a data preprocessing and augmentation pipeline using **Generative Gradient Origin Networks**, enhancing data diversity and increasing dataset volume by 40% to improve model performance.
- Implemented a **MultiStream CNN-LSTM** model with **CTCLoss** and Visual Alignment Constraint, achieving a **24.4% WER** on the RWTH Phoenix benchmark, surpassing conventional models while ensuring computational efficiency.
- Served as a key resource for **AI and machine learning**, mentoring six associates and demystifying model workflows for non-technical stakeholders, resulting in a **20% boost** in team efficiency and informed decision-making.
- Utilized **Google BigQuery** for crafting intricate database queries and harnessing **Power BI** to craft impactful dashboards. Managed a high-performing team of IT professionals, resulting in a **40%** increase in project efficiency.

EDUCATION

Indiana University Bloomington, USA   **Aug 2023 – May 2025**
Master of Science in Data Science   **GPA: 3.94/4.00**
Coursework: Data Mining, Machine Learning, Advanced Database Concepts, Statistics, Data Visualization, Algorithms

University of Mumbai, India   **Aug 2017 – Jun 2021**
Bachelor of Engineering in Computer Engineering   **GPA: 3.50/4.00**
Coursework: Big Data Analytics, Elements of Artificial Intelligence, Advanced DB, Data Structures, Exploratory Data Analysis

TECHNICAL SKILLS

- **Programming Languages:** Python, R, SQL, C++, Java, JavaScript, ReactJS, HTML, CSS, XML
- **Business Intelligence:** Tableau, Power BI, Excel, Looker Studio, GeoPandas, Seaborn, NumPy, Pandas
- **Machine Learning:** Tensorflow, PyTorch, Keras, Scipy, XGBoost, RAG, LDA, NLTK, NLP, LLM, NumPy, Pandas, Scikit-learn, OpenCV
- **Databases:** MySQL, BigQuery, PostgreSQL, MongoDB, Neo4j, SSIS, SSRS, Parquet
- **Cloud Tools:** Azure (Databricks, Data Factory), AWS (S3, Redshift, Glue, Lambda, Sagemaker), GCP (BigQuery), Kubernetes
- **Project Management:** Confluence, Jira, Agile, Waterfall, SDLC, Kanban, CRM, ERP

ACADEMIC PROJECTS

Photo Realistic Face Generation using Generative Adversarial Networks

- Devised a controllable **GAN** model to generate high-resolution human images with options to select from over 10 facial features. Incorporated the Wasserstein Loss function and an auxiliary 2D CNN to unravel the latent space and estimate feature distributions.

Big Data & Climate Change Prediction Pipeline (PySpark)

- Engineered a scalable data pipeline using MongoDB and PySpark to ingest and process global temperature, performing distributed ETL and forecasting trends via **Spark mLib**; deployed on cloud compute with visualization for high-volume datasets.

A/B Testing & Marketing Campaign Optimization

- Conducted A/B testing and regression analysis on 365-day Facebook and Google AdWords data to assess conversions and cost efficiency. Used hypothesis testing and cointegration analysis to optimize ad spend allocation and improve ROI.

Vio-Later: Predictive Prevention of Traffic Violations and Road Accidents

- Developed a traffic violation prevention system using New York traffic data, leveraging **GeoPandas** for city map visualization using **ensemble** of **XGBoost, ANN, Decision Trees** to predict high-risk violations and reduce accidents.