Vicky Suman

Sr. Data Science Engineer

April 6, 1992

House No.- 4018-3F, Block-H, Ansal Versalia, Sector 67, Gurugram, Haryana 122101

+91 9711404468

jatolia.vicky8@gmail.com

https://www.linkedin.com/in/vicky-suman-3a22a0123/

https://vsuman-ai.github.io/

https://medium.com/@vickysuman

Education

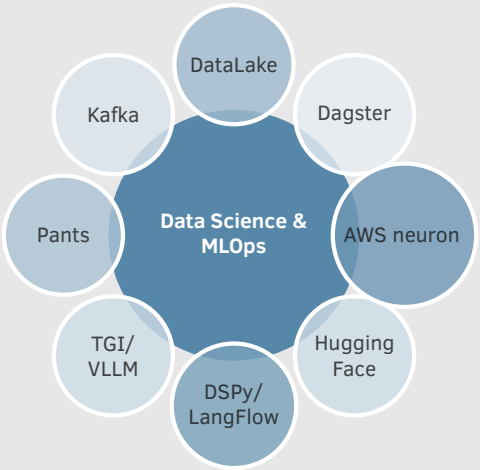
Indian Institute of Technology Bombay
M.Sc. Applied Statistics and Informatics
2015 – 2017

P.G. D.A.V. College, University of Delhi
B.Sc. Statistics (Hons)
2012 – 2015

About Me

AI and Data Engineering leader architecting scalable, production-grade AI platforms. I lead teams of 3 pax, define technical direction and deliver high-impact LLM and ML systems that translate complex data into measurable business value.

Platform & Tools



Working Experience

November, 2022 – present

Sr. Data Science Engineer — **Ontic Technologies, Noida**

Leading the architecture and delivery of enterprise scale AI platforms across LLM systems, NLP, computer vision, and distributed data infrastructure. Translating complex technical systems into reliable products with measurable business impact.

★ **Global Risk Event Intelligence Platform:** Architected and launched a revenue-generating AI-powered Risk Event Intelligence platform (~ 7.5K per client) and processing 10K+ messages per hour.

- Implemented **embedding-based de-duplication and trend clustering with summarization**, reducing alert noise and enabling event-level intelligence aggregation.
- Developed a real-time **embedding platform** (Kafka + Triton Inference Server + gRPC), storing vectors in Qdrant/MongoDB to power semantic search or clustering for trends and de-duplication.
- Engineered **LLM-based contextual location extraction** to identify incident-specific addresses and implemented reverse geocoding for precise geo-mapping.
- Built a **hypothesis-driven LLM severity classification** system with few-shot prompting to prioritize security events across diverse client scenarios.

★ **AI Systems & Infrastructure Engineering:** Developed AI systems across LLM applications, NLP pipelines, computer vision services, and inference infrastructure, with emphasis on scalability, reliability, and operational excellence.

- Built an **Agentic AI** framework using DSPy + tool-based reasoning to automate topic creation, web extraction, and topic query generation.
- Architected a **denoising face detection pipeline** (YOLOv11 ONNX + fine-tuned ViT) deployed via Triton with batched inference, significantly reducing false-positive image matches. Eliminated redundant third-party API calls, driving **$5K/month infrastructure cost savings**.
- Designed a **GenAI metrics** platform powered by **QLoRA** fine-tuned LLMs (served via TGI + DSPy), generating structured in backend API format.
- Built a **camera health monitoring** feature for customer support team to detect offline client cameras and trigger proactive alerts, reducing security blind spots.

★ **Data Platform & Architecture:** Build distributed data platforms and processing architectures for large-scale ingestion (millions of records in per hour), transformation, indexing, and retrieval to power enterprise AI and intelligence products.

- Designed and deployed a **production-grade LLM inference** using **(TGI)** on **k8s**, enabling scalable LLM serving across multiple modules.
- Developed a DSPy-based **prompt optimization** framework for LLM classification tasks using **GEPA, SIMBA, and Decision-Tree** optimizers.
- Established a **Monorepo** architecture using **Pants** enabling modular builds, dependency isolation, and faster multi-service development workflows.
- **Data lake pipeline** on AWS using **Lambda, Glue (PySpark), and Hudi** on S3, enabling downstream platform use cases in real-time event processing.

Vicky Suman

Sr. Data Science Engineer

</> Skills

- Python
- Docker
- Cloud Services: AWS
- MongoDB
- Kubernetes
- SQL

Awards

- **Yearly Rockstar Performance Dedication** and **Commitment** towards work, **Ontic 2023**.
- **Exceptional Performance** and **Commitment** towards work, **RMS 2019**.
- **All India Rank 60** in **IIT** JAM 2015 MS exam.

Leadership Position

2016 – 2017 **Head of Logistic Team**, 36th Mathematics Olympiad, **Mathematics Association, IITB**

2016 – 2015 **Key Member of Logistic Team and Registration Team** of 35th Mathematics Olympiad, **Mathematics Association, IITB**

Soft Skills

- Team Player
- Adaptability
- Communication
- Business Acumen
- Storytelling
- Critical Thinking

December, 2020 – October, 2022 **Modeler, Data Science & Analytics (India)** — **RMS A Moody's Analytics Company, Noida**

Working on **Casual Models** which are mathematical model represent of casual relationship of a **Cyber Risk Ecosystem** .

* Build a model that explains the likelihood of a company given that size-sector and jurisdiction is impacted by **Business Email Compromise** attack of cyber risk, with add-on **DVC** as model architecture.
* Built a **Dagster** pipeline consists of three assets group at high level, all centered around extract news from domains with different languages, apply multiple NLP models and aggregate activity analysis. Run are **Schedule** on daily basis with partition and send a **Slack message** on failure and **Materialized KPI** asset for each run.
* Leverage **M2M-100 and mBART-50** multilingual (seq-to-seq) model primarily intended for **language translation tasks**.
* Generates a **synthetic catalog of enterprise software** with associated **vulnerabilities** that threat actors may exploit, To have a comprehensive view of the **worldwide software market**, Which determines the software product **"footprint"** of firms attacked when **threat actors exploit vulnerabilities** in that software.
* Assisted in development **contagious malware worm simulations model**, analysis and tune the hyper-parameter by generating the experiment's by leveraging **MLflow**.
* Developed **Cyber Risk IT Model's life cycle** (experimentation, reproducibility and central model registry) for model calibration and validation by using **MLflow**.
* Developed **Linguistic rule by using dependence parser**, to find the type of incident and then relation between the performs the action and receives the action from title of the given article.

May, 2018 – December, 2020 **Senior Modeling Analyst, Data Science & Analytics (India)** — **RMS A Moody's Analytics Company, Noida**

Responsible for build a framework to automate **CID**.

* Designed and Developed integrated **Cloud Incident and Exposure** Databases for top CSPs, unifying outage events, product-service-region mappings, revenue, market share, and cloud adoption metrics into a unified schema for **analytical reporting**.
* **Clustering DDOS attacks** into similar groups, where the attacks have similar characteristics.
* Designed and deployed a multilingual cyber threat **NLP pipeline** with REST APIs, enabling automated text extraction from diverse open source (e.g., media blogs, RSS feeds) and **advanced analysis** through a **sentence classifier** (SpaCy Vectors + SVM) for identifying cyber attack relevance. Integrated a custom **entity extraction** module using CRF and SpaCy NER to detect critical details such as affected organizations, attack types, financial impact, and timelines **enhancing threat intelligence** accessibility across teams.

June, 2017 – April, 2018 **Data Scientist Trainee** — **Phenom People, Hyderabad**

Concern project were **chat-bot and job search engine**.

* Built and deployed a personalized **AI chatbot using RASA Stack and Dialogflow** to handle job search queries, FAQs, and chit-chat, integrating NLU to enhance conversational relevance and UX by storing of session-based conservation, browser history and chatbot interaction to drive personalized recommendations.
* Developed and exposed REST APIs for ML models, including an SVM-based sentence classifier for **job intent detection and a recommendation engine** to match users with relevant job profiles.