Satadru Bag

Data Scientist | Machine Learning | Computer Vision & Time Series | PhD

satadru.iucaa@gmail.com | github.com/deltasata | linkedin.com/in/satadru-bag

Munich, Germany · EU Blue Card holder (Germany) · Available at short notice

deltasata.github.io

Data scientist with 12+ years of analysing large, noisy datasets, including 7+ years of hands-on machine learning on images and time series. Trained first in theoretical physics and mathematics; invents statistical methods where standard ones fail. Developed and maintains a real-time Google Cloud pipeline built for **~10M events (~1 TB) per night**; designed deep-learning models that flag rare events in image streams with **70%+ recall at only 0.01% false positives**. Now applying the same methods to Earth observation (EO-Agent: flood and change questions answered from satellite images). Builds agentic AI workflows that run large model searches under pre-registered quality checks. PhD in astrophysics; leads a rare-event search in a 10,000-member international collaboration.

10M events per night | **110M** images screened | **70%+** recall at 0.01% FPR | **12+** yrs data analysis, **7+** ML

CORE SKILLS

Machine learning	PyTorch · TensorFlow / Keras · scikit-learn · Vision Transformers · CNN / ResNet / ConvNeXt / U-Net · ConvLSTM / LSTM · Graph neural networks · Multi-task learning · Ensembles & gradient boosting (LightGBM) · Clustering & k-NN · Explainable AI (XAI) for false-positive control
Generative AI	Generative models (diffusion, normalizing flows, VAEs) · Self-supervised pretraining (MAE, masked-image modelling) · Foundation-model fine-tuning
Vision & time series	Image classification & segmentation · Object detection & tracking · Change detection (image differencing) · Multi-sensor image fusion · Satellite imagery (Sentinel-1/2, STAC) · Image time series · Online (per-frame) prediction · Time-series forecasting · Rare-event & anomaly detection · Synthetic data generation
Statistics	Bayesian inference · MCMC / Hamiltonian Monte Carlo (Stan, emcee) · Simulation-based inference · Uncertainty quantification · Probabilistic modelling · Hypothesis testing · Time-series statistics (autocorrelation, spectral analysis) · Feature engineering & EDA
LLMs & agents	Large language models · RAG · Agentic AI (LangGraph, MCP, Claude Agent SDK, OpenAI Agents SDK)
Data & MLOps	Google Cloud (GCP) · AWS · BigQuery / SQL · DuckDB · REST APIs · Kafka (Avro) · Parquet / pandas · Docker / Apptainer · ONNX · Kubernetes · CI/CD · DVC · FastAPI · Git · Linux · HPC (SLURM, CUDA/ROCm, NVIDIA A100)
Programming	Python · SQL · C / C++ (MPI) · R / Stan · Bash · Testing (pytest, ruff, mypy)
Currently learning	MLflow · Apache Spark · Databricks · Apache Airflow · Google Earth Engine
Languages	English fluent · German basic · Bengali native · Hindi fluent

EXPERIENCE

Data Scientist / ML Researcher (Postdoctoral Researcher)

Jun 2025 – present

Technical University of Munich (TUM) & Max Planck Institute for Astrophysics (MPA) · Munich, Germany

- Lead a rare-event search in a petabyte-scale imaging survey (working-group lead in a **10,000-member** collaboration since 2025); developed and maintain its official real-time Google Cloud pipeline, built for **~10M events (~1 TB) per night** from six APIs and Kafka streams, with BigQuery cost guards, tests and CI/CD.
- Designed **OASIS**, transformers (PyTorch) that classify irregular multi-band image sequences frame by frame, chosen from **227 GPU runs** of 100 configurations in 19 architecture families (the 7 published on 10+ seeds each). The best model cut the baseline's loss by **30%** and hit 70% recall at 0.01% false positives after **6 images, not 9**.
- Rebuilt the survey's production image-processing chain (change detection, object detection, cross-frame tracking, two CNN quality scorers) as a lean container that runs on small image tiles, making training-data generation at scale feasible on HPC.

ML Research Scientist (Alexander von Humboldt Postdoctoral Fellow)

Mar 2023 – May 2025

Technical University of Munich (TUM) & Max Planck Institute for Astrophysics (MPA) · Munich, Germany

- Built a ConvLSTM (TensorFlow) that flags rare events in multi-band image time series, rescoring each frame: **70%+ recall at a 0.01% false-positive rate** by the 9th image; trained on 128,000 simulated sequences (~200 GB).
- Co-developed an ensemble of five CNNs that screened **110M images**: 75% recall at a 0.01% false-positive rate; post-processing removed 29% of false candidates with no loss in recall, surfacing 92 previously unknown objects.

- Co-authored a multi-sensor fusion study: a two-branch ResNet on high- and low-resolution imagery raised recall from 41% to **55%** at a 0.01% false-positive rate.

Machine Learning Researcher (Postdoctoral Researcher) — Sep 2019 – Nov 2022

Korea Astronomy and Space Science Institute (KASI) · Daejeon, South Korea

- Invented **new statistical methods** that detect a hidden, time-delayed copy of a signal in one blended time series and measure the delay: transients at under 5% false positives; random signals with no signal model, held impossible since 1996, at **92% precision** vs 53% for autocorrelation, with a proof and real-data confirmation.
- Built LSTM and 1D-CNN time-series classifiers and Bayesian models fitted with Hamiltonian Monte Carlo (Stan).

Visiting Researcher – Data Analysis (Guest Postdoctoral Researcher) — Apr 2019 – Jul 2019

Tartu Observatory, University of Tartu · Tartu, Estonia

- Measured the 3D shape of 956 large-scale clusters in a point cloud of 584,449 galaxies; open data.

Researcher – Statistics & Data Analysis (PhD Research Scholar in Astrophysics) — Aug 2012 – Mar 2019

Inter-University Centre for Astronomy and Astrophysics (IUCAA) · Pune, India

- Wrote **SURFGEN2**, a C/C++ and MPI package for 3D surface modelling and shape statistics, used by several research groups.
- Strong foundation in mathematics and statistics (differential equations, topology, Bayesian inference); 8 papers.

PORTFOLIO PROJECTS

EO-Agent, Earth-observation AI analyst — PyTorch, TerraMind, Prithvi-EO, Claude Agent SDK, MCP, LangGraph, vLLM

Answers plain-language questions about floods, fires, deforestation and construction anywhere on Earth from free Sentinel-1/2 data, shows the evidence and a calibrated confidence, or answers "cannot tell". Each learned model (cloud mask, U-Net / TerraMind flood maps, Siamese change network, burn-scar model, transformer over gapped time series) replaced a rule only after a pre-registered held-out test: water IoU **0.833 vs 0.690**; flood questions answered correctly **0.889 vs 0.722**.

github.com/deltasata/satellite-change-agent

Self-supervised ViTs for rare-object detection — PyTorch, ViT, ConvNeXtV2, MAE, AMD MI300A

Finds rare objects (1 in thousands) in 5-band survey images. Automated model search (~60 architectures, ~300 GPU runs) with pre-registered comparisons; in-domain masked-autoencoder pretraining on top of ImageNet weights halved the loss; a ViT + ConvNeXtV2 ensemble reached **AUC 0.9998** on a locked 30,000-object holdout, finding 90.6% at a 0.1% false-positive rate.

github.com/deltasata/self-supervised-vit-lsst-lenses

Real-time event pipeline — Python, REST APIs, Kafka (Avro), SQL (BigQuery, DuckDB), Google Cloud Storage, Parquet, pytest

Stream and archive pipelines over six external data providers (REST APIs and Kafka, retries with back-off), a space-time SQL join on BigQuery and DuckDB, Parquet storage, a cost-guarded BigQuery layer and 500+ automated tests (ruff, mypy) in CI.

github.com/deltasata/LSST_alert_pipeline_LSN_search

OASIS, transformers for asynchronous image streams — PyTorch, Transformers, GPU

Seven transformer designs that update a classification with every new frame of an irregular multi-band image stream: factorized vs joint space-time attention, windowed and pyramid variants, Perceiver-style readout, causal masking; arXiv 2610.17168.

github.com/deltasata/OASIS--On-the-fly_ASynchronous_Image-Stream-classification

ConvLSTM rare-event detector — TensorFlow / Keras

Spatio-temporal classifier for multi-band image time series that rescores every new frame, trained on realistic synthetic sequences built from single real images (per-frame blur, noise-map corrections, Poisson noise, hard look-alike negatives). A&A 707, A123 (2026); Part II in press.

github.com/deltasata/ConvLSTM-LSNe-Detection

EDUCATION

PhD in Astrophysics Inter-University Centre for Astronomy and Astrophysics (IUCAA), Pune, India — 2019

M.Sc. in Physics Indian Institute of Technology (IIT) Kanpur, Kanpur, India — 2012

B.Sc. in Physics (Hons.) St. Paul's Cathedral Mission College, University of Calcutta, Kolkata, India — 2010

LEADERSHIP & RECOGNITION

- **Publications:** 35+ research papers, 30+ peer-reviewed in Q1 journals.
- **Fellowships, grants & awards:** Alexander von Humboldt Fellowship (2023 – 2025); Wenner-Gren Fellowship (2023, Sweden); co-secured a DFG collaborative research centre grant (2024 – 2029); Best Postdoc Award, KASI (2022).
- **Talks, reviewing & mentoring:** 50+ talks worldwide (Stanford, Harvard, Imperial College, Tokyo); referee for several journals, incl. Nature Portfolio; mentored five students (bachelor to PhD level), three of whom lead-authored papers.