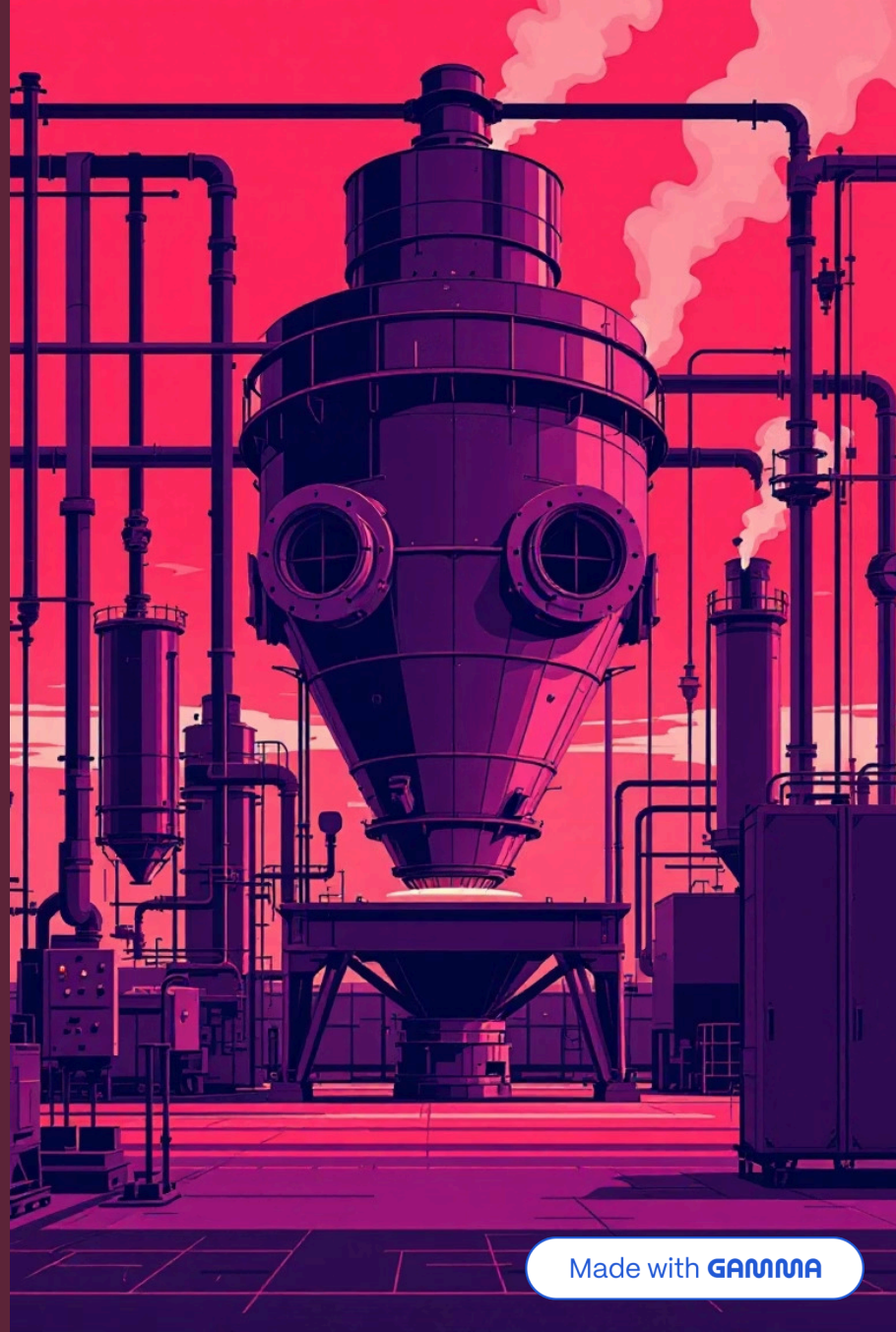
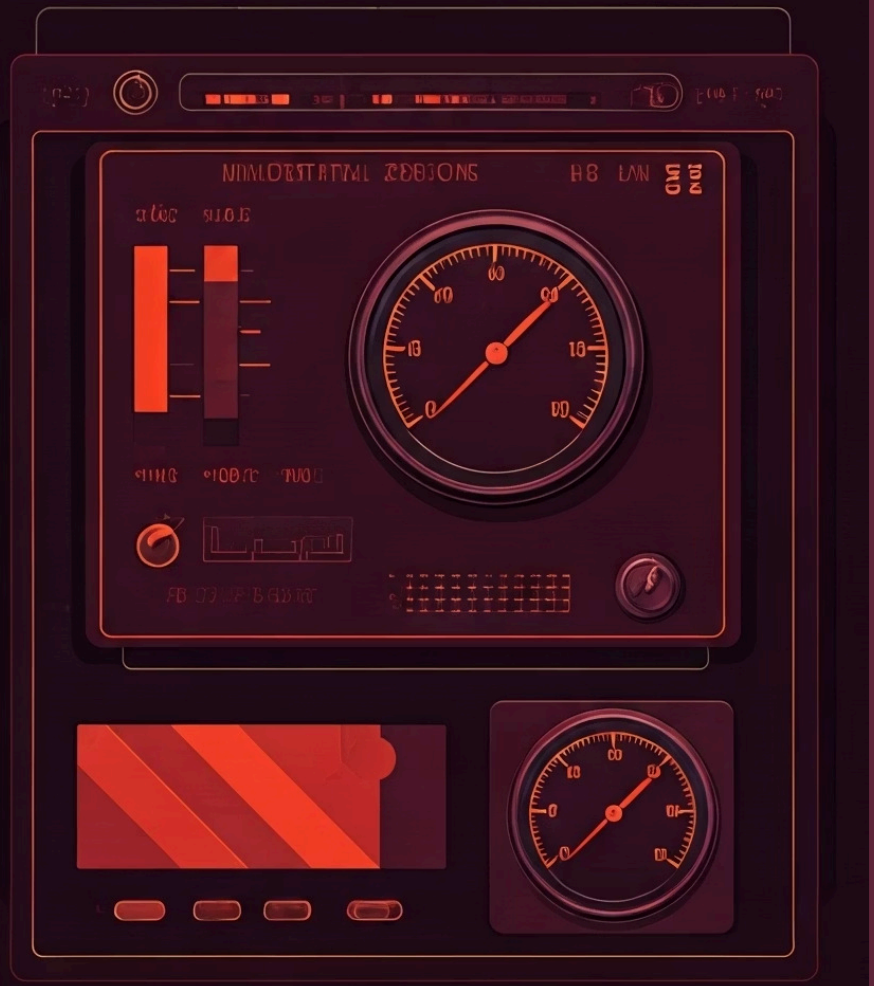
Cyclone Sensor Machine Analysis

Comprehensive data science project demonstrating end-to-end machine learning workflows

- Task 1: Machine Data Analysis & Time Series Forecasting
- Task 2: RAG + LLM System Design for Industrial Documentation

- By Anish Joshi





Task 1: Problem Definition & Dataset Overview

Dataset Characteristics

- 3 years of continuous sensor data
- ~370,000 records at 5-minute intervals
- Multiple temperature and pressure sensors
- Material outlet temperature monitoring

Business Objectives

- Detect and predict machine shutdowns
- Segment operational states automatically
- Identify anomalous behavior patterns
- Forecast inlet gas temperature trends

Data Preparation & Preprocessing

01

Data Quality Assessment

Identified missing values, timestamp gaps, and statistical outliers across all sensor channels

02

Time Series Standardization

Enforced strict 5-minute indexing and handled irregular sampling intervals

03

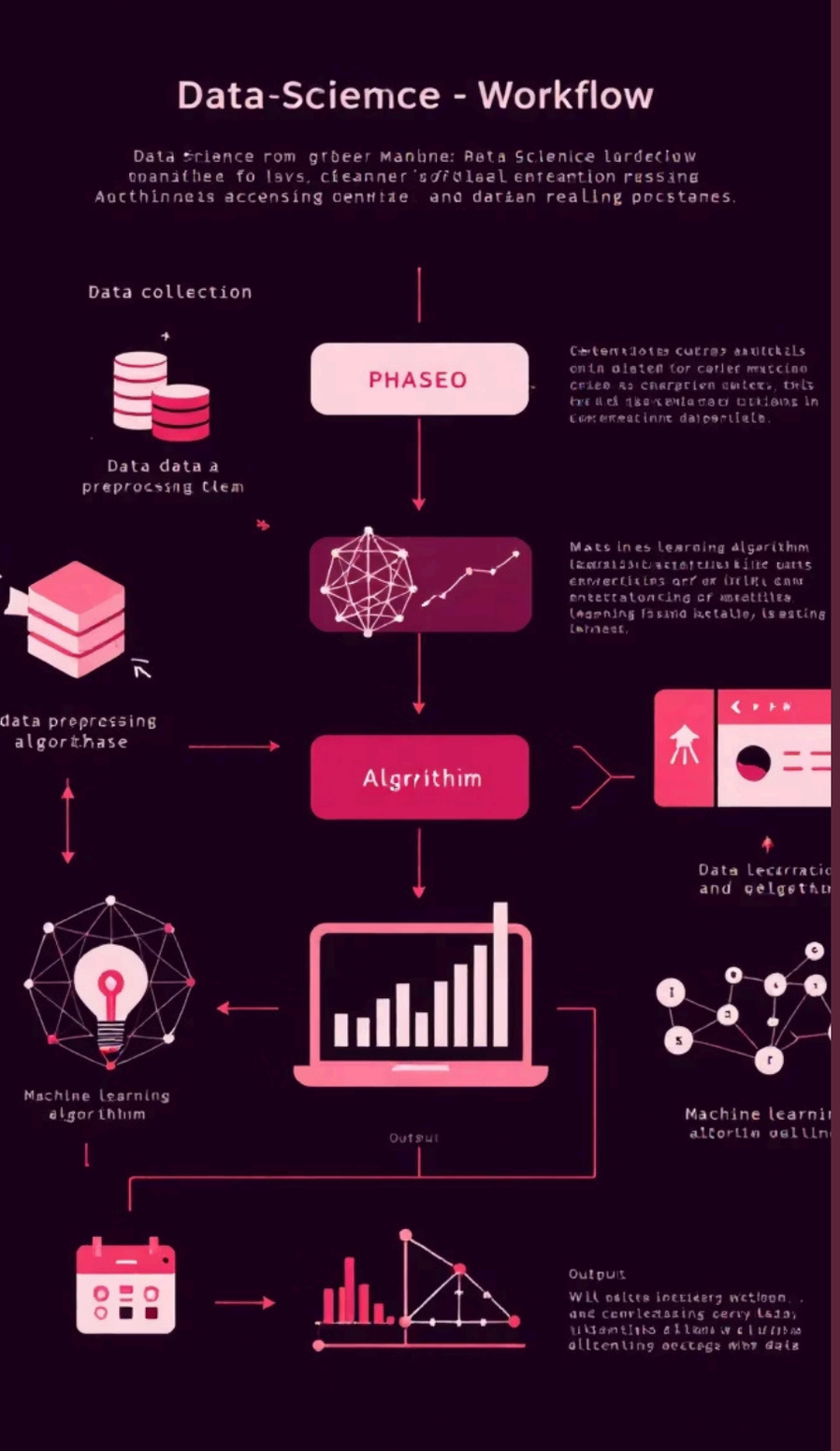
Exploratory Analysis

Generated summary statistics, correlation matrices, and temporal behavior patterns

04

Behavior Visualization

Analyzed weekly and yearly patterns to distinguish normal vs. variant operational states



Analysis Methods & Algorithms

Shutdown Detection

Developed automated flagging system for idle periods and calculated total operational downtime metrics

State Clustering (K-Means)

Identified interpretable machine states: Normal, High Load, Degraded, and Transitional operations

Anomaly Detection

Implemented contextual, state-aware detection using Isolation Forest and rolling MAD thresholds

Time Series Forecasting

Deployed ARIMA and Prophet models, benchmarked against persistence baseline performance

Key Insights & Findings

Predictive Shutdown Patterns

Shutdown events concentrated during specific operational states, enabling early warning systems for maintenance scheduling

State-Based Anomaly Distribution

Degraded operational states showed significantly higher anomaly density, indicating equipment stress patterns

Forecasting Performance Insights

Effective predictions achieved only under stable operational regimes due to inherent non-stationarity challenges

Actionable Recommendation: Monitor transitions into "risky states" to trigger proactive maintenance alerts and prevent unexpected downtime

Task 2: RAG System Problem & Requirements

Business Challenge

Operations team manages 50+ PDF documents including manuals, SOPs, and troubleshooting guides

Need for intelligent natural-language Q&A system: *"What does a sudden draft drop indicate?"*

Technical Requirements

- Open-source technology stack only
- Reliable responses with source citations
- Scalable architecture for future growth
- Robust against AI hallucination risks

RAG System Architecture

1

Document Ingestion & Preprocessing

Automated PDF parsing and text extraction with metadata preservation

2

Smart Chunking Strategy

Optimized 200-400 token segments with strategic overlap for context preservation

3

Embeddings & Vector Storage

all-MiniLM-L6-v2 embeddings stored in FAISS index for efficient similarity search

4

Hybrid Retrieval System

Dense semantic search combined with lexical reranking for optimal relevance

5

LLM Generation with Citations

flan-t5-small model with enforced source attribution and response validation

System Guardrails & Evaluation Framework

Quality Guardrails

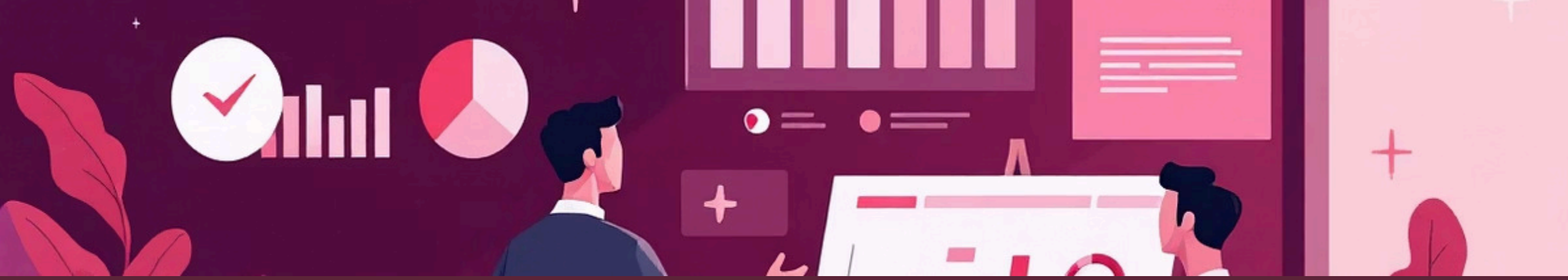
- Graceful fallback for low-relevance queries
- Mandatory source citations in all responses
- Sensitive query filtering and blocking

Performance Evaluation

- Precision@k and recall@k metrics
- Faithfulness scoring for accuracy
- Results stored in evaluation.csv

Scalability Planning

- Sharded FAISS or Chroma for 10x documents
- Microservices + autoscaling for 100+ users
- CPU embeddings and caching for cost optimization



Project Outcomes & Business Impact

1

Task 1: Machine Analytics Framework

Delivered comprehensive end-to-end analysis pipeline covering shutdown detection, state clustering, anomaly identification, and forecasting capabilities

2

Task 2: RAG System Prototype

Designed and implemented production-ready RAG system with robust guardrails, comprehensive evaluation metrics, and detailed scalability roadmap

Next Steps & Future Enhancements

Task 1 Improvements

- Integrate domain expert labels for supervised anomaly classification
- Implement real-time monitoring dashboard
- Develop predictive maintenance scheduling

Task 2 Enhancements

- Upgrade to Llama-2 7B model for higher fidelity responses
- Enhanced reranking algorithms
- Multi-language document support

Key Demonstration

Both projects showcase complete **end-to-end data science workflows** from raw sensor data analysis to applied NLP system development