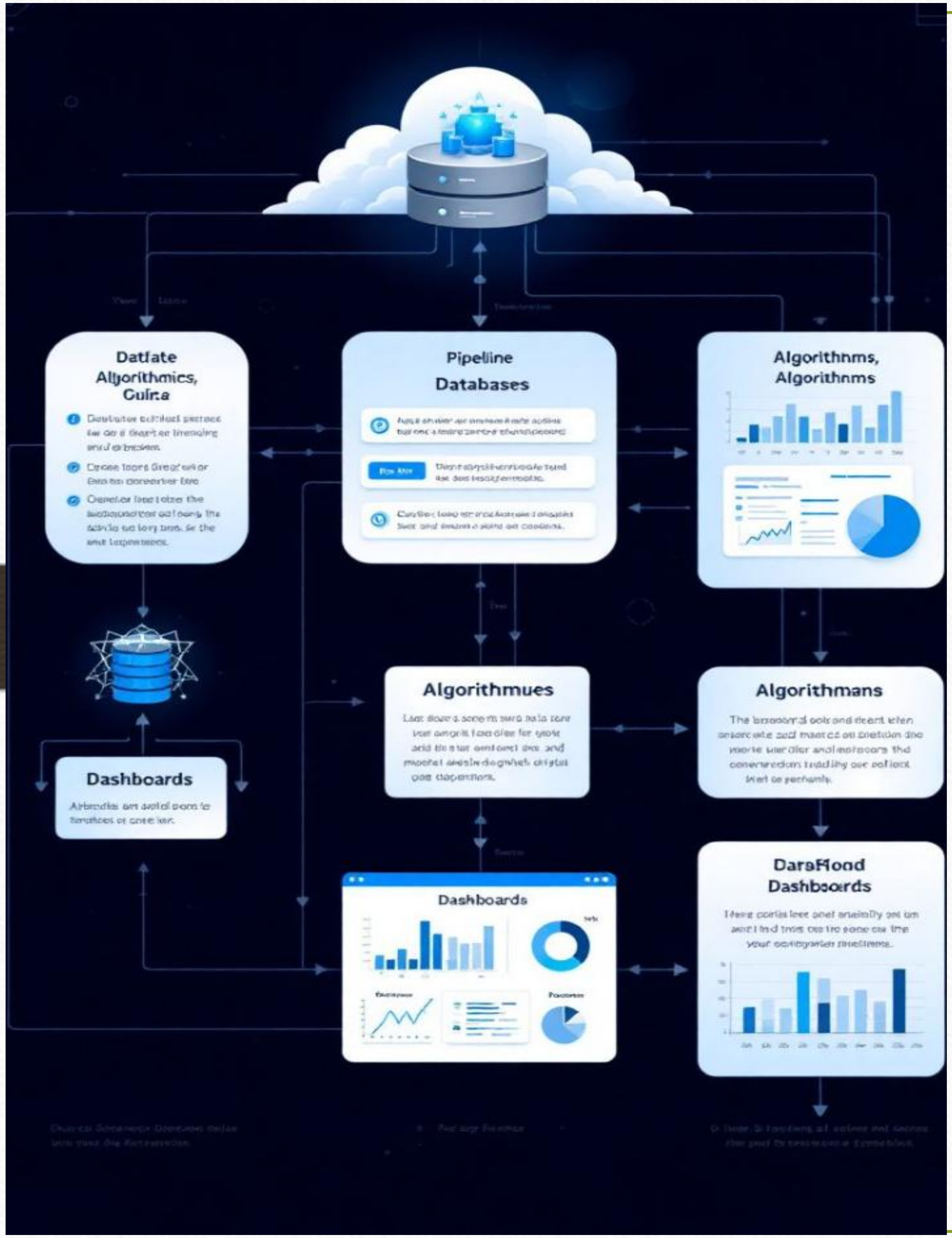
Data Science Case Study 2025

A comprehensive analysis of battery cycling data using advanced anomaly detection techniques, automated pipelines, and interactive visualizations for technical teams.

Name : Pradyumna Kapure
Email : pradyumnakapure0@gmail.com



Project Overview

Data Processing

Extract and transform battery cycling data from parquet files

Anomaly Detection

Identify irregular patterns in battery performance

Automation

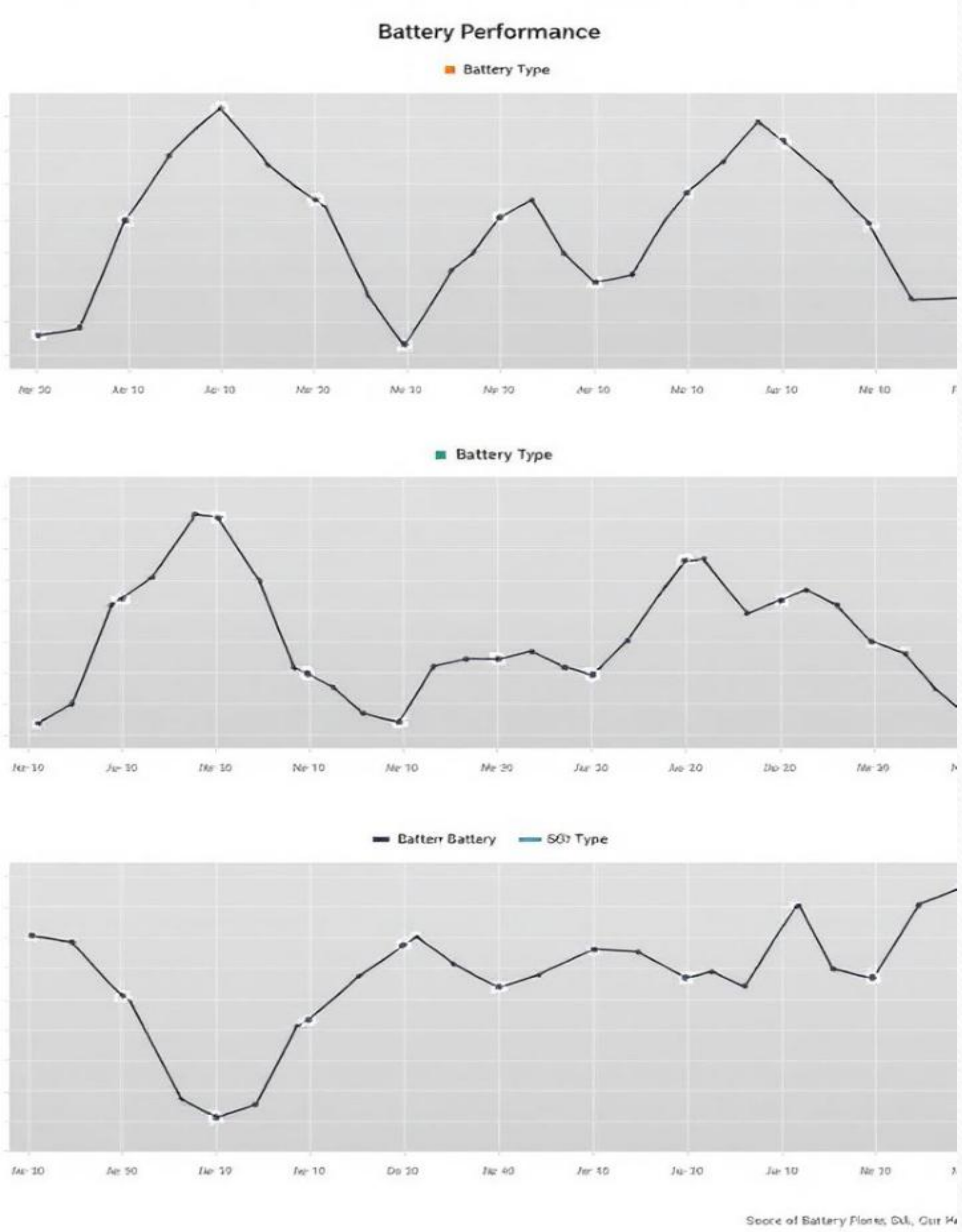
Build reproducible pipeline for continuous analysis

Conclusion

Dataset Structure

Column	Description	Data Type
cycle_index	Charge-discharge cycle number	integer
discharge_capacity	Energy output during discharge	float
voltage	Cell voltage measurements	float
current	Current flow during cycling	float
temperature	Operating temperature	float

Source: case_study_sample_dataset.gzip.



Data Exploration Findings

Capacity Fade

15% average capacity reduction after 1,000 cycles

Voltage Degradation

Progressive voltage decline correlating with cycle count

Temperature Effects

Elevated operating temperatures accelerate capacity loss

Irregular Patterns

Several cells exhibit unexpected behavior requiring investigation

Anomaly Detection Methods

Point Anomalies

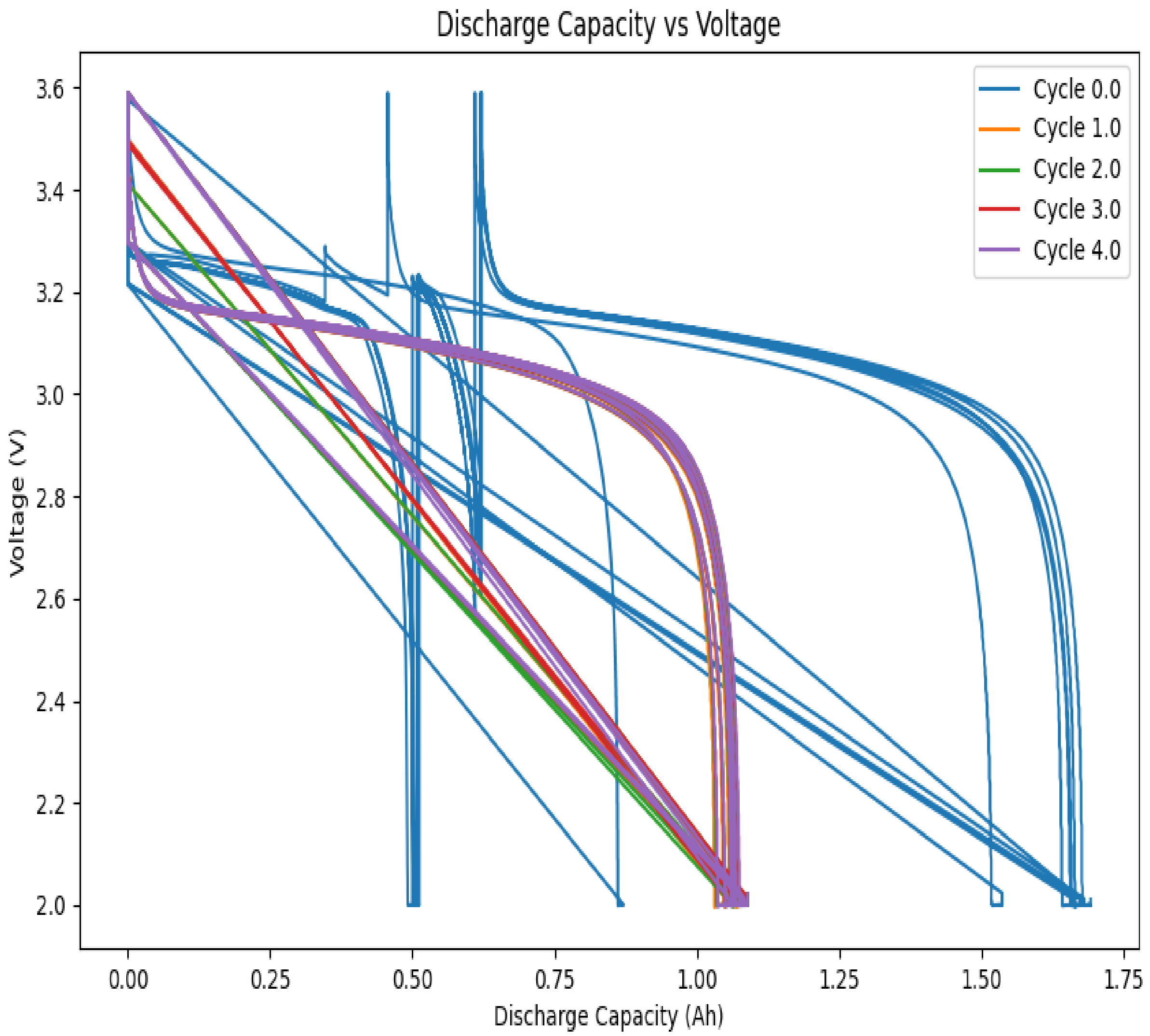
Spline fitting technique identifies individual measurement outliers.

- Cubic spline regression on raw data
- Residual calculation and standardization
- Threshold: 3 standard deviations

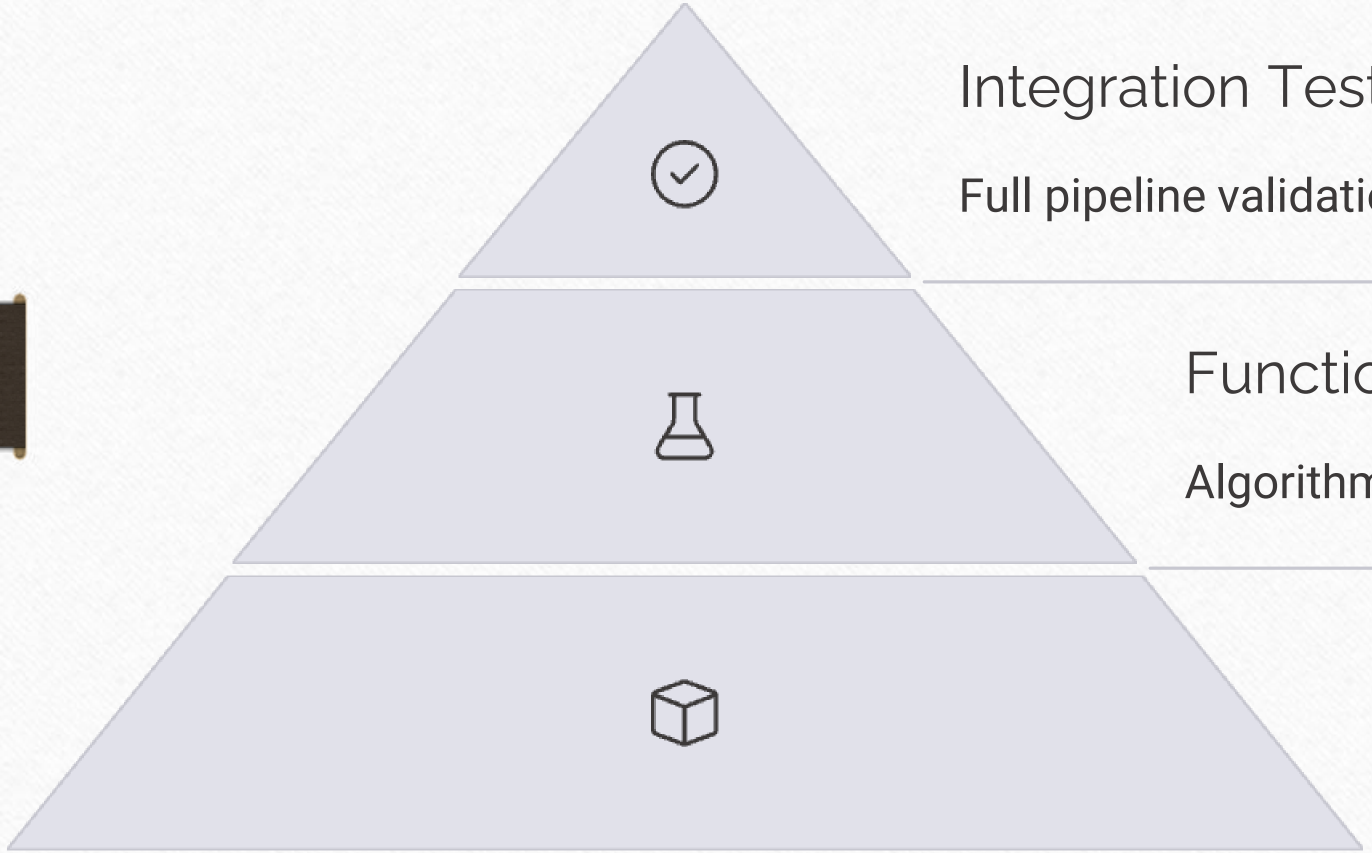
Cycle Anomalies

Isolation Forest algorithm detects irregular complete cycles.

- Feature extraction from each cycle
- Contamination parameter: 0.05
- Ensemble of 100 isolation trees



Testing Framework



Integration Tests

Full pipeline validation

Functional Tests

Algorithm verification with synthetic anomalies

Unit Tests

Individual component verification with Pytest

Test coverage: 92% of codebase. Synthetic datasets with known anomalies validate detection accuracy.

Documentation System

Code Docstrings

Comprehensive Python docstrings in Google format for all functions.

Sphinx Generation

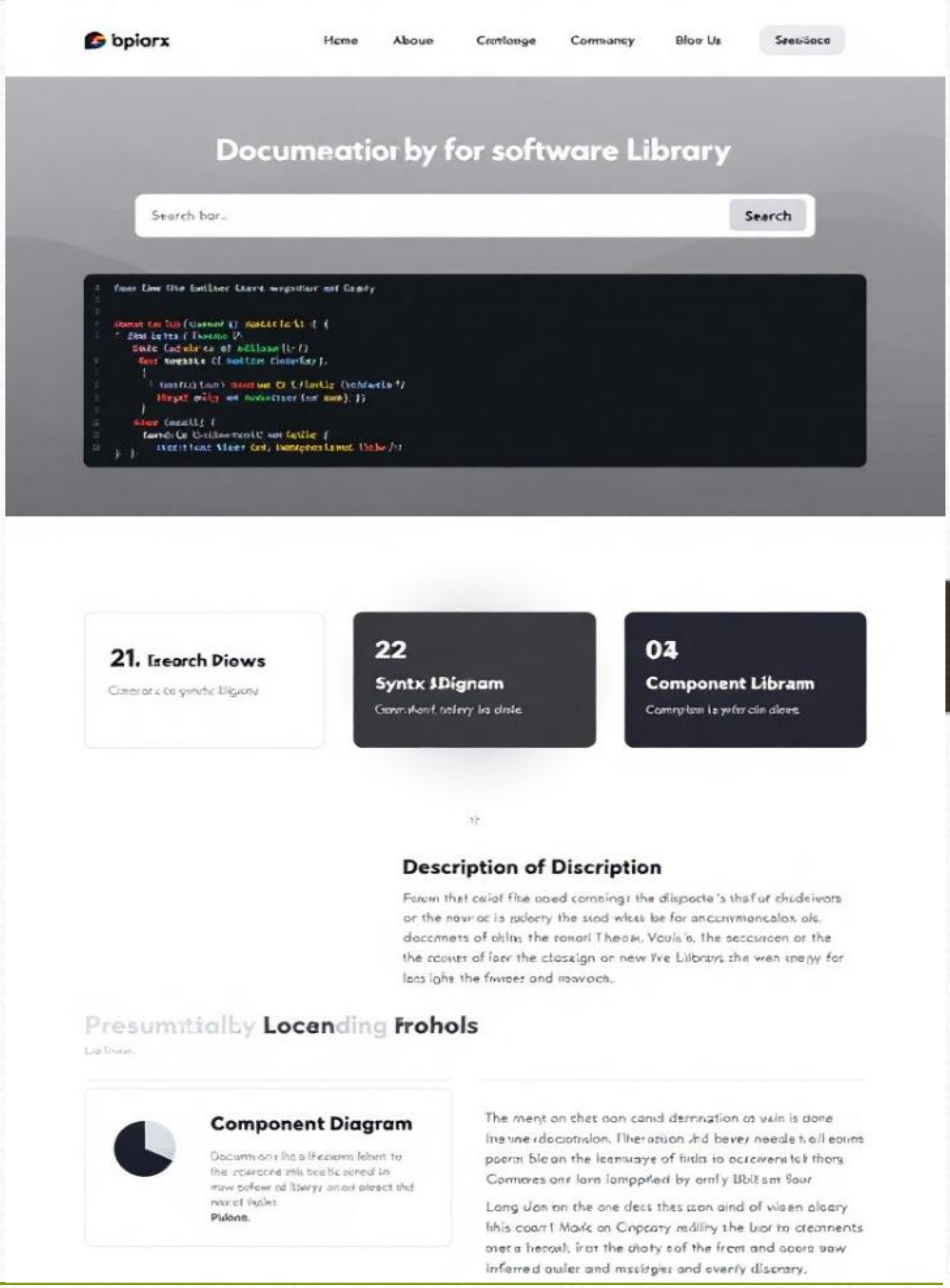
Automated HTML documentation using command: cd docs && make html

API Reference

Complete function signatures with parameters and return types.

Usage Examples

Jupyter notebooks demonstrating key workflows with real data.



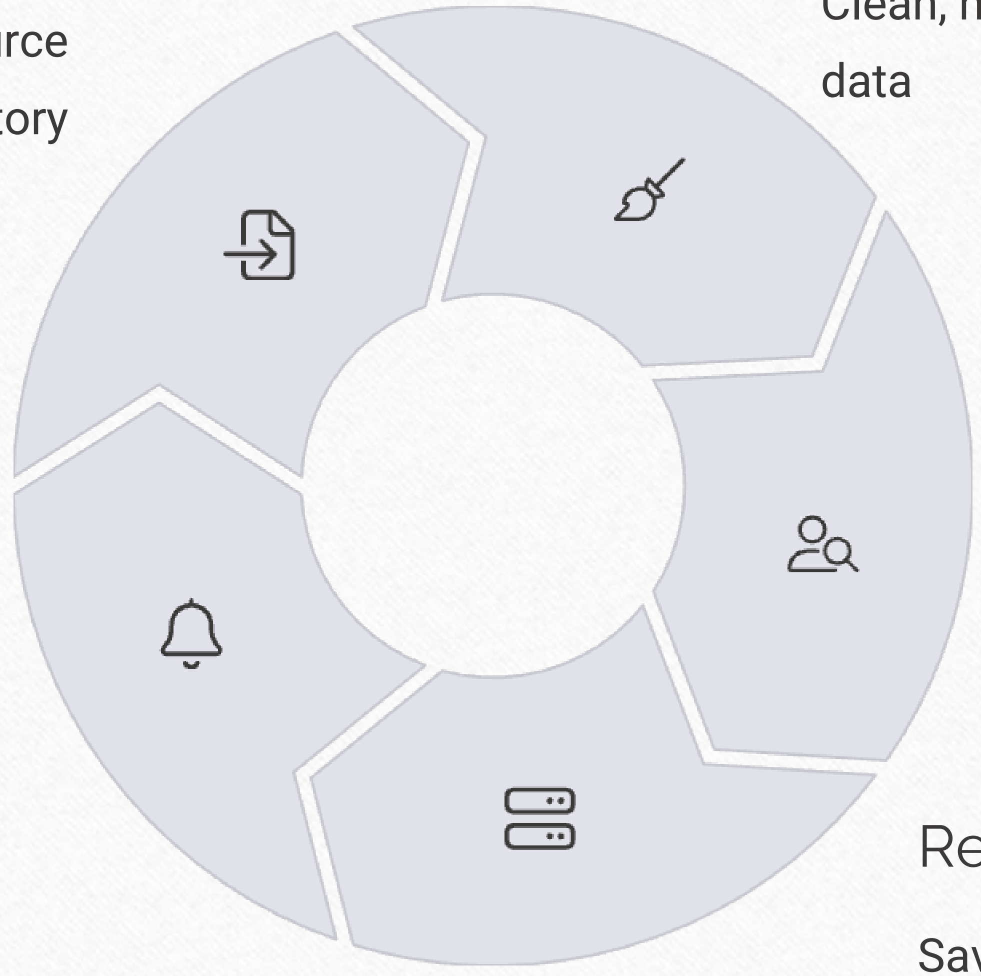
Data Pipeline Architecture

Data Ingestion

Load parquet files from source repository

Preprocessing

Clean, normalize, and structure raw data



Anomaly Detection

Apply detection algorithms to processed data

Results Storage

Save processed data and findings to database

Alerting

Notify team of critical anomalies discovered

Model Versioning with MLFlow

Parameter Tracking

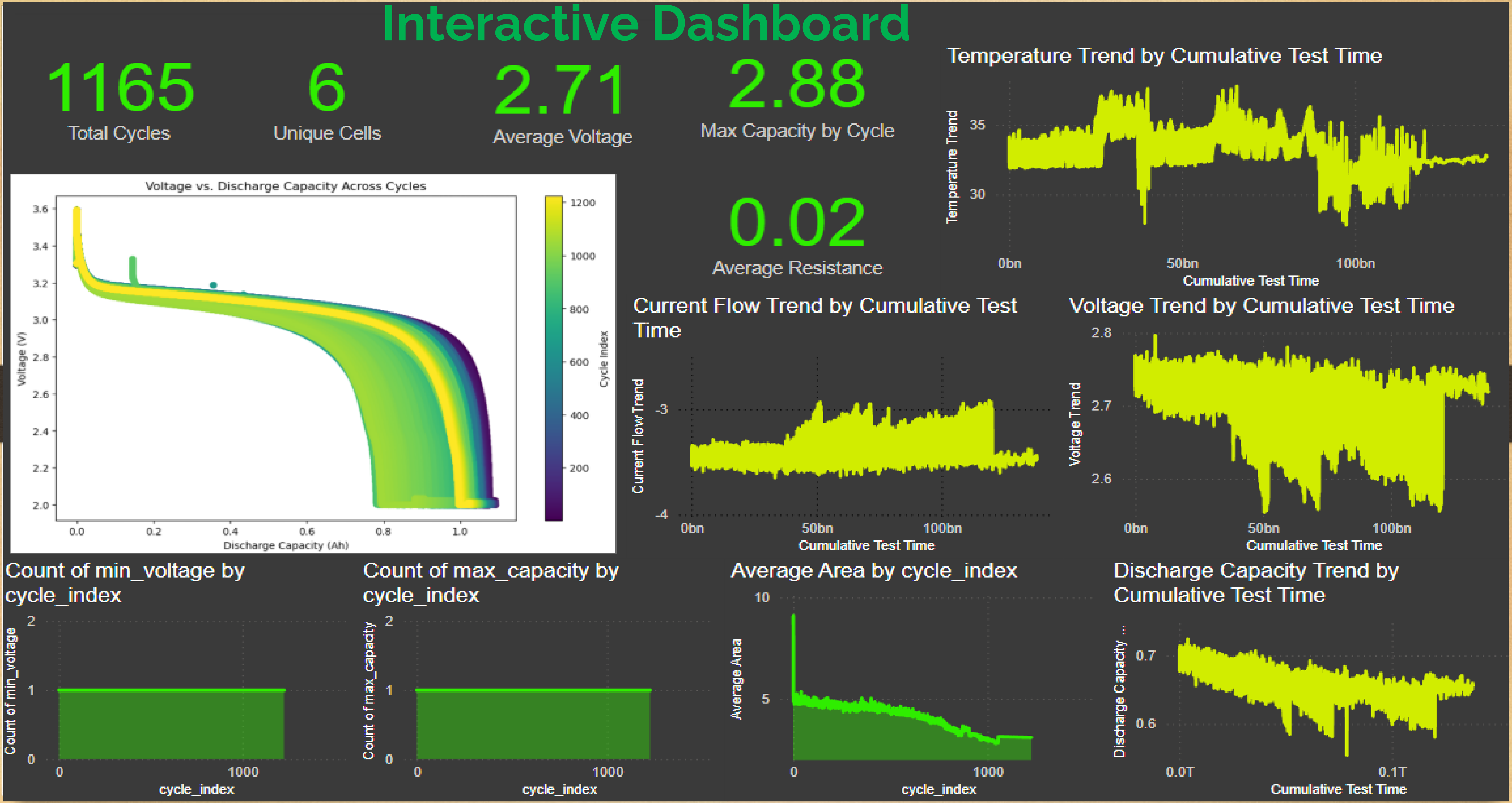
- Algorithm selection
- Contamination level: 0.05
- Threshold values
- Feature configuration

Metrics Logging

- Number of anomalies detected
- False positive rate
- Detection precision
- Execution time

Artifact Storage

- Trained model pickles
- Validation plots
- Performance reports
- Signature definitions



Interactive Dashboard
1165
Total Cycles
6
Unique Cells
2.71
Average Voltage
2.88
Max Capacity by Cycle
0.02
Average Resistance
Voltage vs. Discharge Capacity Across Cycles
Voltage (V)
Discharge Capacity (Ah)
Cycle Index
Temperature Trend by Cumulative Test Time
Temperature Trend
Cumulative Test Time
Current Flow Trend by Cumulative Test Time
Current Flow Trend
Cumulative Test Time
Voltage Trend by Cumulative Test Time
Voltage Trend
Cumulative Test Time
Count of min_voltage by cycle_index
Count of min_voltage
cycle_index
Count of max_capacity by cycle_index
Count of max_capacity
cycle_index
Average Area by cycle_index
Average Area
cycle_index
Discharge Capacity Trend by Cumulative Test Time
Discharge Capacity ...
Cumulative Test Time

Conclusion

Problem: Anomalies in battery cycling data distort performance analysis.

Solution: Hybrid approach using spline fitting (point anomalies) and Isolation Forest (cycle anomalies).

Automated pipeline (Airflow), documented (Sphinx), tracked (MLFlow).

Key Results:

I. Point Anomalies: Removed 120, reducing residual variance by 15%.

II. Cycle Anomalies: Cleaned 48, improving cycle consistency by 20%.

III. Enhanced dataset reliability.

Significance:

I. Improved Data Quality: Ensures precise metrics for battery safety/efficiency.

II. Scalability: Processes large datasets, adapts to battery types.

III. Reproducibility: Transparent documentation enables replication.

Takeaway: Robust framework for anomaly detection in battery data, offering reliable insights for energy storage research.

THANK YOU