Week 1 Advisor Meeting - AstrID Project Overview

Project Summary

AstrID (Astronomical Identification) is a comprehensive system for temporal dataset preparation and anomaly detection in astronomical observations. The project focuses on detecting transient events (supernovae, unusual celestial phenomena) using machine learning approaches on time-series astronomical imaging data.

Core Anomaly Detection Strategies

1. Multi-Modal ML Approach

- **Primary**: U-Net deep learning architecture for image segmentation and anomaly detection
- **Secondary**: Traditional ML methods (Isolation Forest, One-Class SVM) for feature-based detection
- **Ensemble**: Combined scoring system leveraging multiple detection methods

2. Image Differencing Pipeline

- **ZOGY Algorithm**: Optimal image subtraction for transient detection
- **Classic Differencing**: Simple subtraction with scaling/offset correction
- **Source Detection**: SEP/photutils for candidate extraction from difference images

3. Training Data Strategy

- **Synthetic Anomaly Generation**: Creating artificial transients (bright spots, dark spots, streaks)
- **Real-Bogus Classification**: Filtering artifacts from genuine astrophysical events
- **Historical Survey Data**: Leveraging SDSS, Pan-STARRS, ZTF datasets

Tool Integration & Usage

MLflow Implementation

- **Experiment Tracking**: Model training runs, hyperparameters, metrics logging
- **Model Registry**: Version control and deployment management for U-Net models
- **Artifact Storage**: Cloudflare R2 integration for model weights and training data
- **Energy Monitoring**: GPU power consumption tracking during training/inference

Prefect Orchestration

- **Processing Flows**: Automated pipelines (ingestion → preprocessing → differencing → detection)
- **Model Training Workflows**: Scheduled retraining based on new data and performance metrics
- **System Monitoring**: Health checks, performance monitoring, alerting system
- **Worker Management**: Dramatiq workers for parallel processing across pipeline stages

Technical Architecture

Data Pipeline

1. **Observation Ingestion** → FITS file processing, metadata extraction
2. **Preprocessing** → Calibration, WCS alignment, quality assessment
3. **Image Differencing** → ZOGY implementation, candidate detection
4. **ML Inference** → U-Net anomaly detection, confidence scoring
5. **Human Validation** → Curation interface, expert review workflow

Infrastructure Stack

- **Backend**: FastAPI, PostgreSQL (Supabase), Redis
- **ML**: PyTorch, scikit-learn, TensorFlow/Keras
- **Storage**: Cloudflare R2, local filesystem
- **Containerization**: Docker, docker-compose for development

Research Questions to Discuss

1. Model Performance & Validation

- How do we establish ground truth for astronomical anomaly detection?
- What metrics best evaluate performance on rare transient events?
- How do we handle class imbalance in astronomical datasets?

2. Data Quality & Preprocessing

- What preprocessing steps are most critical for anomaly detection accuracy?
- How do we handle varying image quality across different surveys/instruments?
- Should we focus on single-epoch or multi-epoch detection strategies?

3. Scalability & Production Deployment

- How do we scale to handle millions of alerts per night (LSST-scale)?
- What's the optimal balance between automated detection and human validation?
- How do we ensure model robustness across different survey conditions?

4. Scientific Impact & Applications

- Which types of astronomical transients should we prioritize?
- How can this system contribute to time-domain astronomy research?
- What collaboration opportunities exist with observatories/survey teams?

5. ML/MLOps Methodology

- How do we implement continuous learning from expert feedback?
- What's the strategy for model retraining as new data becomes available?
- How do we detect and handle model drift in production?

Current Development Status

- **Phase 1**: Core infrastructure and domain models
- **Phase 2**: ML pipeline integration and workflow orchestration
- **Phase 3**: Production deployment and monitoring (planned)

This plan implements a practical application of modern MLOps tools (MLflow, Prefect) to real-world astronomical research challenges, with potential for significant scientific impact in transient astronomy.