AstrID – Week 3 Plan - *Meeting 3*

Project Summary

Focus this week: consolidate real-data ingestion wins (MAST, SkyView, API flows), stand up the initial training dataset pipeline, and run the first U-Net baseline with simple comparisons.

Ingestion Status (Highlights)

- MAST real queries (astroquery): functional with validation and filtering
  - Example: M31 – 5,355 raw obs; 2,821 HST after filtering; time-filtered subsets behave as expected
  - Multi-region sweep (M31, M42, M51, M81, NGC 5128): 24,472 observations total across HST/JWST/TESS
- SkyView / CDS HiPS2FITS: DSS2 cutouts retrieved; normalized and displayed; WCS verified
- Reference datasets: 512×512 FITS at ~0.25° FOV created locally and uploaded to R2 (signed URLs)
- Service endpoints: batch ingest and reference-dataset endpoints exercised successfully with API key

Near-Term Milestones

1. Preprocessing defaults (alignment, background subtraction, pixel-scale normalization)
2. Pilot dataset assembly (reference + difference images; synthetic injections for labels)
3. Train first U-Net baseline and log to MLflow (metrics + artifacts)
4. Classical baselines (Isolation Forest, One-Class SVM) for sanity-check comparisons
5. Prefect flow extension to include training and evaluation steps

Risks & Open Questions

- Limited ground truth; dependence on synthetic injections and a small set of real events
- External API/rate limits; benefit of caching and staged downloads