AstrID – *Meeting 3*

Objective

Review ingestion progress with real survey data and align on the initial model-training framework and research plan.

Recap (Since Meeting 2)

- Built and validated end-to-end ingestion for real astronomical data.
- Generated reference FITS cutouts and verified external API integrations.
- Stood up the scaffolding for training data preparation and experiment tracking.

Ingestion Progress (Concrete Results)

- MAST (astroquery) position queries: working with validation and filtering
  - M31 (RA 10.6847°, Dec 41.269°): 5,355 raw observations; 2,821 HST observations after filtering; time-filtered queries return expected subsets (e.g., 1 in last 30 days)
  - Multi-region test (M31, M42, M51, M81, NGC 5128): 24,472 observations total across HST/JWST/TESS
  - Metadata & products retrieval: successful (e.g., `26370626_drz.fits`)
- SkyView / CDS HiPS2FITS: working
  - Retrieved DSS2 cutouts; produced normalized images and displayed successfully
  - Created reference FITS (e.g., 512×512 at 0.25° FOV); verified WCS and data ranges
- Internal DataIngestionService: working
  - Ingested 2,821 observations for M31 (HST/JWST filter) into domain objects
  - Full pipeline test (radius 0.2°) ingested 4,764 observations and produced reference FITS
- API-based ingestion (FastAPI)
  - `POST /observations/ingest/batch-random` (API-key) tested; endpoint healthy
  - `POST /observations/ingest/reference-dataset` produced FITS and stored to Cloudflare R2 with signed URLs
  - Verified survey IDs for HST/JWST/DSS2/TESS and modular survey testing
- Robustness
  - Coordinate validation (RA/Dec bounds) enforced; error handling and timeouts verified; graceful fallbacks present

Data & Artifacts

- Real observations and reference datasets saved locally (e.g., `/tmp/astrid_reference_*/reference_10.6847_41.2690.fits`)
- Cloud artifacts stored in R2 (e.g., `reference-datasets/DSS/83.6330_22.0145_0.250deg_512px.fits`)

Training Framework (Initial Setup)

- Dataset preparation:

  - Confirmed FITS I/O and WCS extraction for building tensors from reference and observation cutouts
  - Plan to assemble pilot training/validation splits from DSS2 references and HST stamps
- Experiment tracking:
  - MLflow planned for runs/metrics/artifacts; model registry to version U-Net baselines
- Orchestration:
  - Prefect flows to automate ingest → preprocess → difference → detect; hooks ready to add training steps

Next: Model Training Research Plan

- Baselines to evaluate first:
  - U-Net/UNet++ segmentation for transient highlighting
  - Classical baselines for comparison: Isolation Forest, One-Class SVM
- Techniques to explore:
  - Self-supervised/contrastive pretraining on sky cutouts; synthetic transient injection for supervised fine-tuning
  - Real–bogus post-classifier on U-Net candidates
  - Image differencing with ZOGY as input channel to models
- Data curation focus:
  - Consistent pixel scale and FOV; band selection strategy; balance across surveys

Near-Term Milestones

1. Finalize preprocessing defaults (alignment, background subtraction, scaling)
2. Produce a small labeled pilot set (reference + difference images; synthetic injections)
3. Train first U-Net baseline; log to MLflow; save example masks/candidates
4. Run classical baselines; compare precision/recall and false alarms per image
5. Wire training into Prefect flow; schedule a repeatable pilot run

Evaluation (Plain Language Metrics)

- Precision, Recall, Localization match (IoU), False alarms per image, Processing time per image

Risks / Open Questions

- Ground truth scarcity; reliance on synthetic injections and limited known events
- Cross-survey variability (filters/seeing) impacting generalization
- API/rate limits; need caching and staged downloads

Requests for Advisor

- Feedback on baseline priorities (U-Net variants vs. classical) and target bands
- Guidance on preferred sky regions/events for the pilot ground truth
- Input on acceptable precision/recall and false alarm targets for first baseline