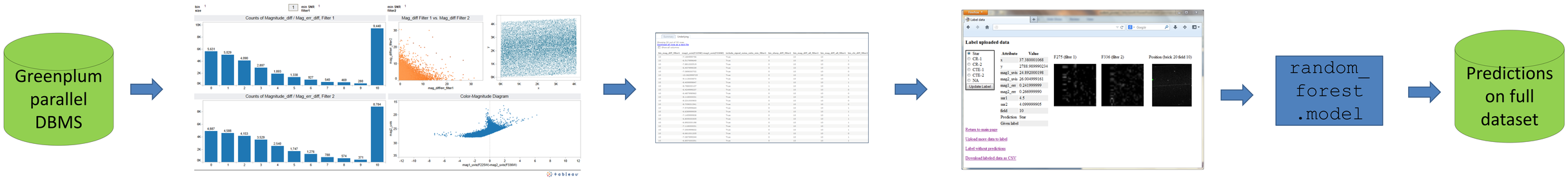
Applying Visualizations and Machine Learning to Detect Artifacts in Astronomy Catalogs

Martina Unutzer, Morgan Fouesneau, Ben Williams, Magdalena Balazinska, and Julianne Dalcanton
University of Washington
munutzer@cs.washington.edu, mfouesn@u.washington.edu, ben@astro.washington.edu, magda@cs.washington.edu, jd@astro.washington.edu

Problem: Catalogs of stars detected from telescope images often also contain cosmic rays. The challenge is to classify each record in the catalog as either a star or a cosmic ray.
Result: In this project, we developed (1) A set of online Tableau dashboards that facilitate collaborative, visual exploration of catalog data; (2) A web-based tool for easily labeling observations as either stars or cosmic rays, and (3) A Weka-based classifier for identifying cosmic rays in the catalog data. The classifier achieves 95% accuracy.

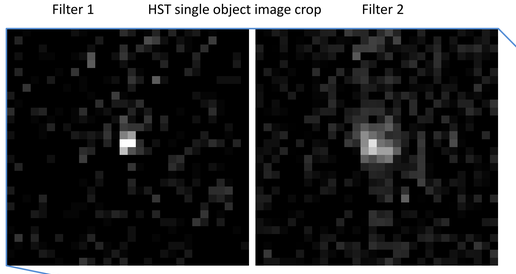


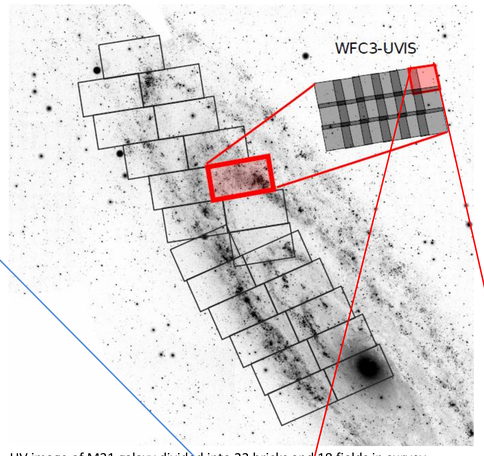
INTRODUCTION

A Hubble Space Telescope imaging survey of the Andromeda galaxy has been processed into a catalog of 40 million stars detected in the images. However, this catalog is largely affected by cosmic ray (CR) contamination even after preprocessing the images with common CR filters. We improved the CR detection by building a classifier which identifies cosmic rays at the catalog level, i.e, based on their measured properties from the images. The project made use of and developed Tableau-based visualization tools for collaborative data access between researchers and departments, making new and cross-disciplinary use of an existing data set to solve a common problem.

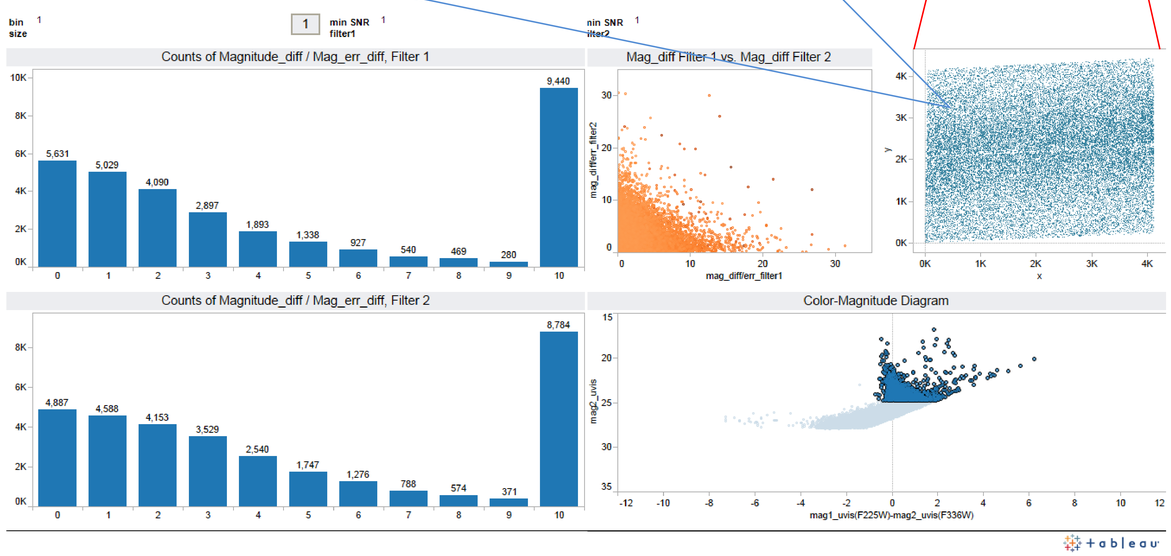
TABLEAU DASHBOARDS

- Goals: Manually explore and contrast the features of stars and cosmic rays. Select features for classifier.
- Greenplum parallel DBMS (173 attributes plus combinations) as input to Tableau.
- Dashboards with histogram for key feature, standard X-Y and color-mag diagrams.
- Linked filters: selected subset displayed on all plots.
- Graphical interface for quick data exploration.
- Collaborative, interactive analysis through Tableau Public: Astronomers save, share data cuts and plots.
- Prompted new discussions about existing data.
- No SQL queries or programming needed to access data.





UV image of M31 galaxy divided into 23 bricks and 18 fields in survey.



CLASSIFIER TRAINING

- Goal: Have scientists label a set of sources detected from images as either stars or cosmic rays. Use labeled data to train a classifier.
- Training set selected with Tableau dashboard filters.
- Labeling site: Python CGI scripts, SQLite.
- Objects displayed with telescope image and key attributes.
- Early classifiers provide predicted class on labeling site.
- Final training set: 600 points
- Increasing training set: no significant performance improvement after approx. 300 points.

CLASSIFIER

- RandomForest algorithm: 10 voting trees, each with 6 randomly selected attributes.
- Used subset of 36 features to classify.
- 10-fold cross-validation: approx. 95% accuracy.
- Final classifier applied to 40 million rows in Greenplum DB using Weka command line.
- Runtimes: training < 1 min, full labeling approx. 8h.

NEXT STEPS

- Image processing: mask out cosmic rays
- Choose certainty threshold to avoid losing mislabeled stars
- Reduce background noise from CRs in processed images
- Integrate model predictions into Tableau dashboards:
  - Spot unexpected trends , characteristics in predictions
  - Early assessment of model performance on full data set

ACKNOWLEDGMENTS: We thank Tableau Software for providing us with a license to Tableau. This work was funded in part by NSF CAREER award IIS-0845397.