Maximizing Science in Big Data Astronomy

Hayden Smotherman
University of Washington
Department of Astronomy

UNIVERSITY of WASHINGTON

DiRAC

Table of Contents:

Table of Contents:

The Five Vs

Volume: The magnitude of the data being processed

Velocity: The rate of data that must be processed

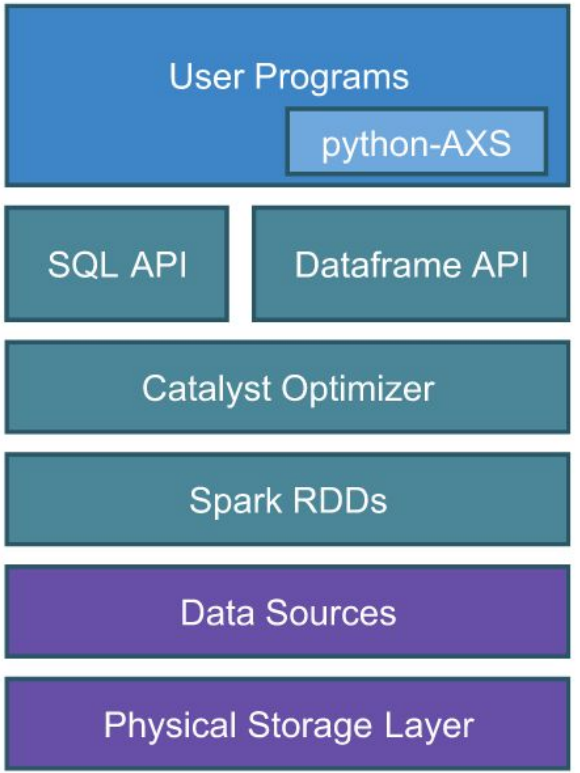
Variety: The complexity of the data: structured and unstructured

Veracity: The fidelity of the data and the associated noise

Value: The scientific value or information contained within data as well as the cost.

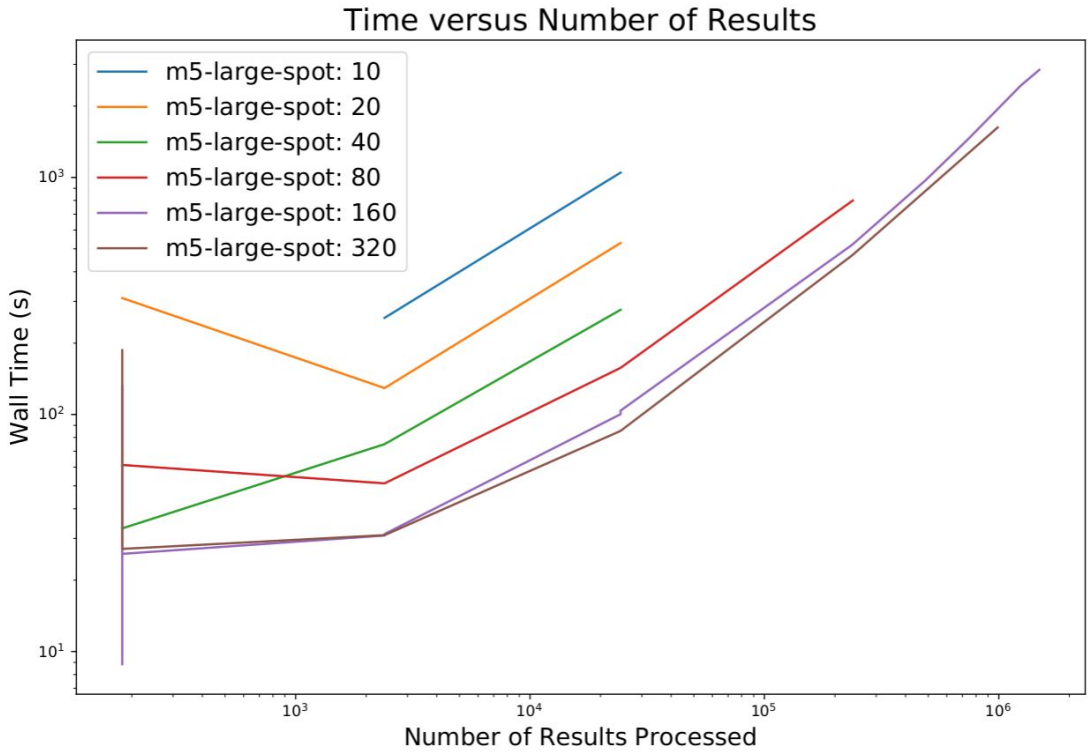
AXS: Astronomy eXtensions for Spark

- AXS enables big data science by implementing common operations using familiar python and astropy commands.
- AXS extends Spark with data partitioning scheme, sort-merge join optimization, and provides efficient parquet file access to light curves (e.g. 2.9 billion ZTF objects)



Zečević et al 2019

Cesium Scaling Test: Volume and Velocity



Selecting the right hardware configuration can reduce runtime by an order of magnitude.

Cesium Scaling Test: Volume and Value

Selecting the right hardware can reduce costs, but selecting the wrong hardware can cost more without saving time.

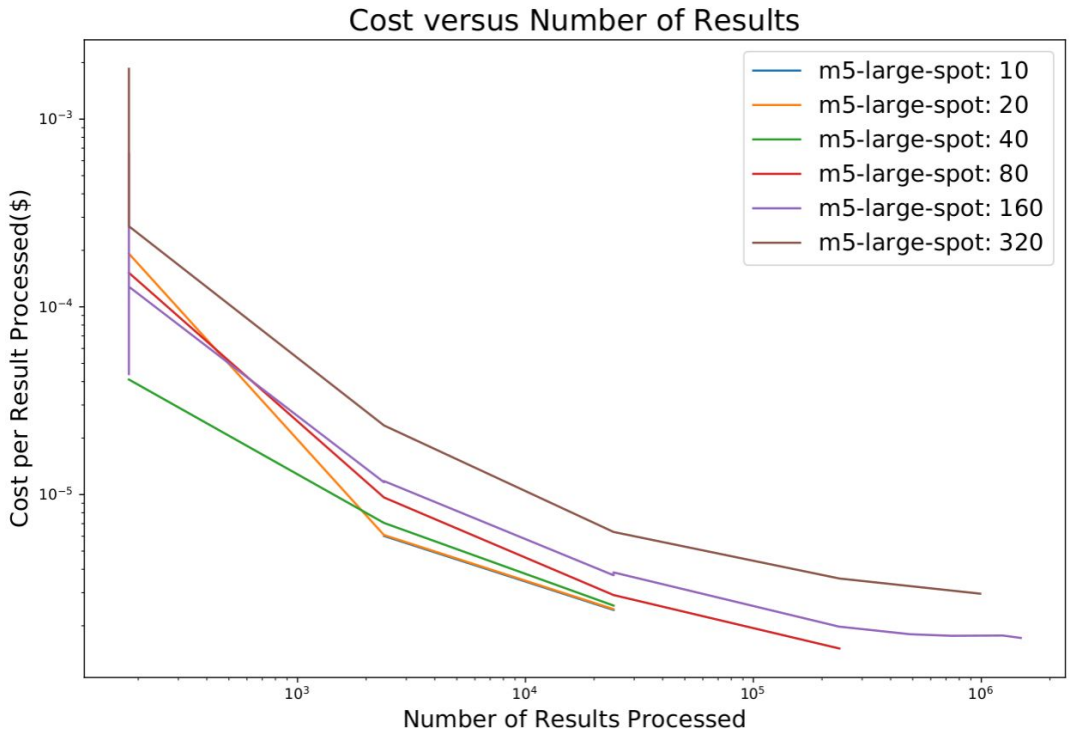
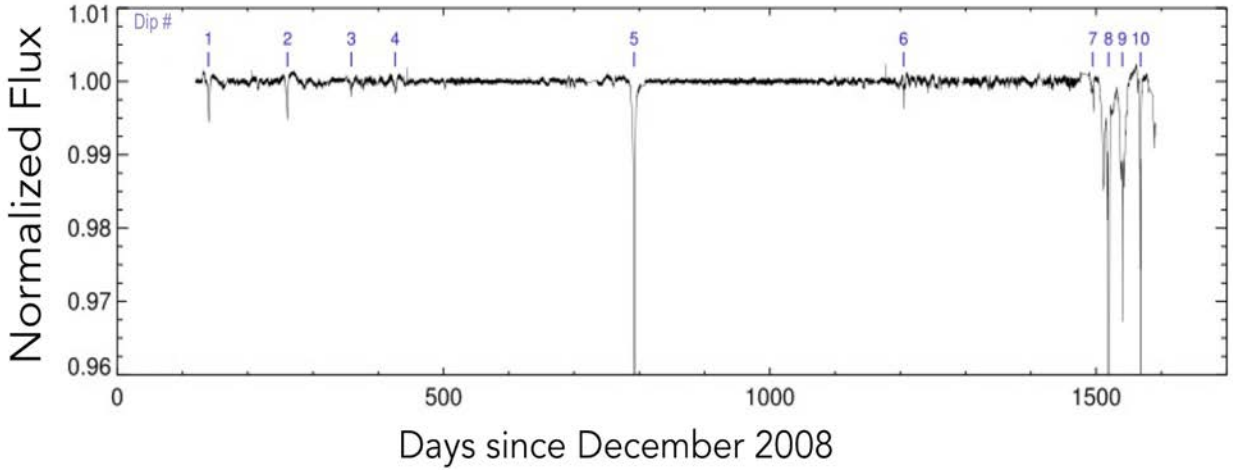


Table of Contents:

AXS: A Scientific Application



Boyajian's star

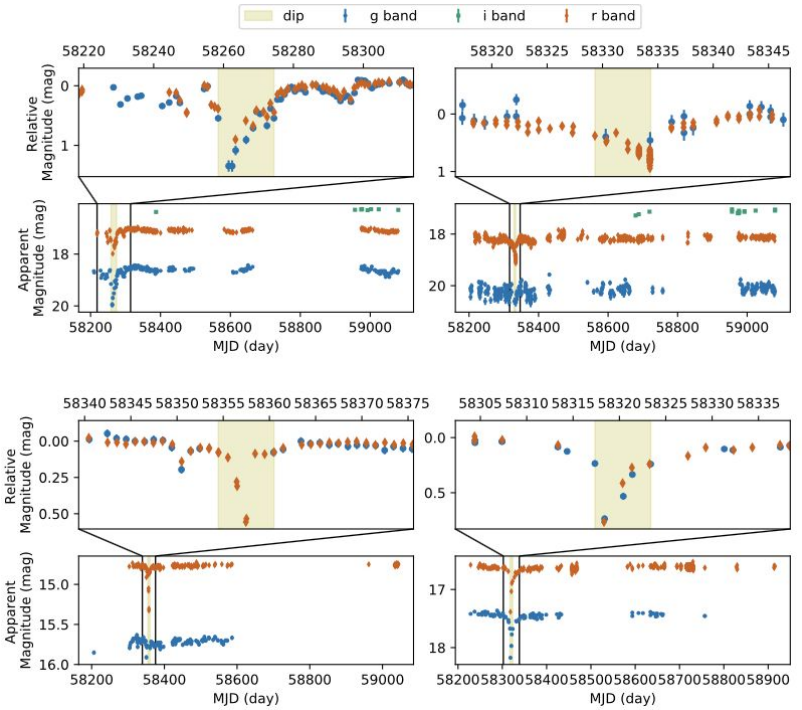
Science is about iteration and filtering

AXS helps create an intuitive approach for embedding light curve analyses within UDFs.

Iteratively process the data to filter and correct errors. Repeatedly pruning the data helps to manage the memory footprint.

Final computational resources are modest (450 core hours to process 1.4B light curves) because data are down selected for complex analyses.

There is a growing need for software engineering in astronomy.



Boone et al 2022

Table of Contents:

Sifting Through the Static: Moving Object Detection in Difference Images – Smotherman et al. (accepted AJ)

The Kuiper Belt

- A disk of objects beyond the orbit of Neptune extending from about 30 au to 45 au.

- The Kuiper Belt contains objects that have been only lightly-perturbed since shortly after the formation of the Solar System.

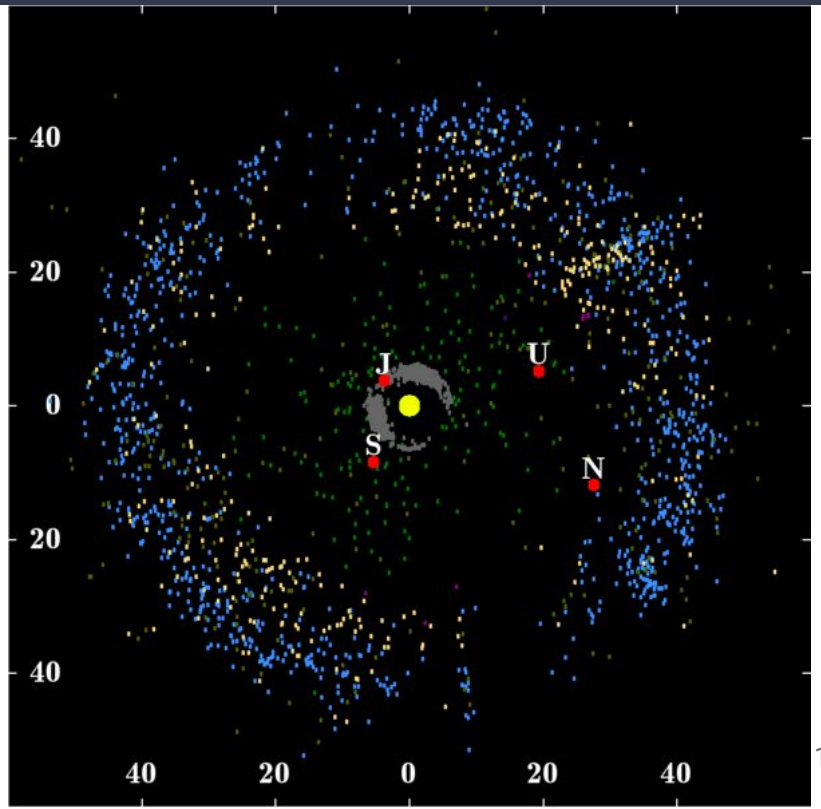
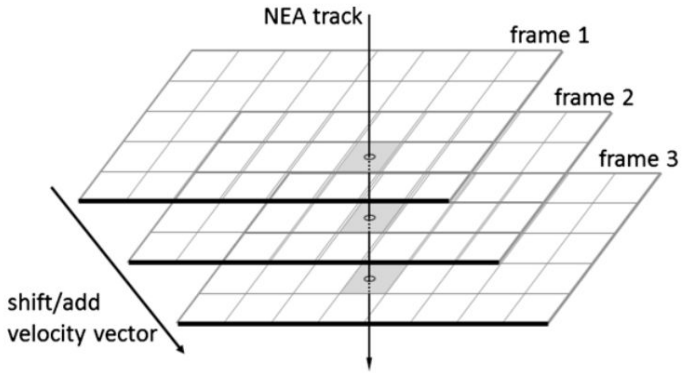


Image Credit: https://commons.wikimedia.org/wiki/File:Kuiper_belt_plot_objects_of_outer_solar_system.png

Current Methods: Digital Tracking

- Classical searches use “shift and stack”
  - If examining stacks by eye, limited to small grid of velocity/angle shifts
- Update of this method is “digital” or “synthetic” tracking
  - Algorithmically search for objects along many trajectories and then add the likelihood to find detections
  - Computationally expensive

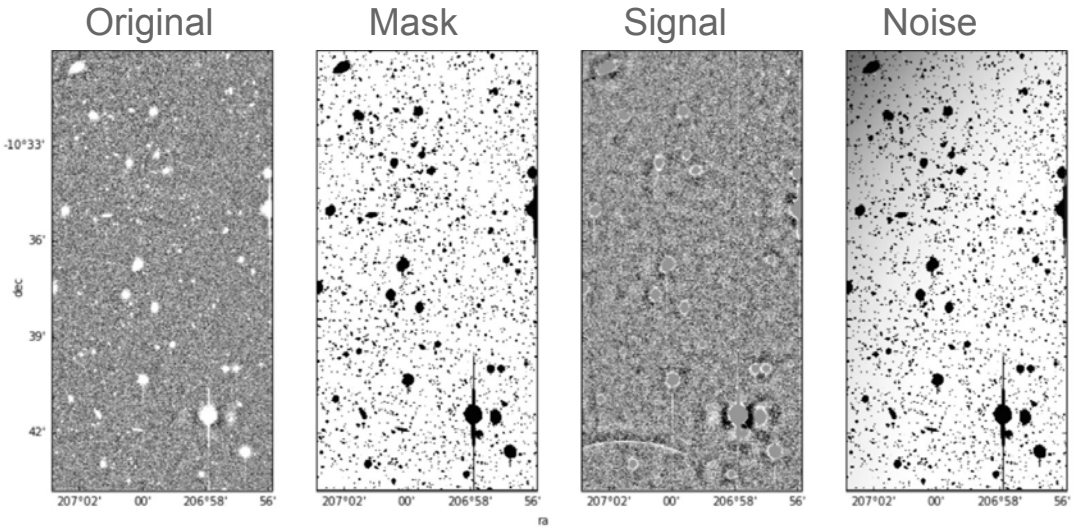


Credit: Zhai et al. (2014)

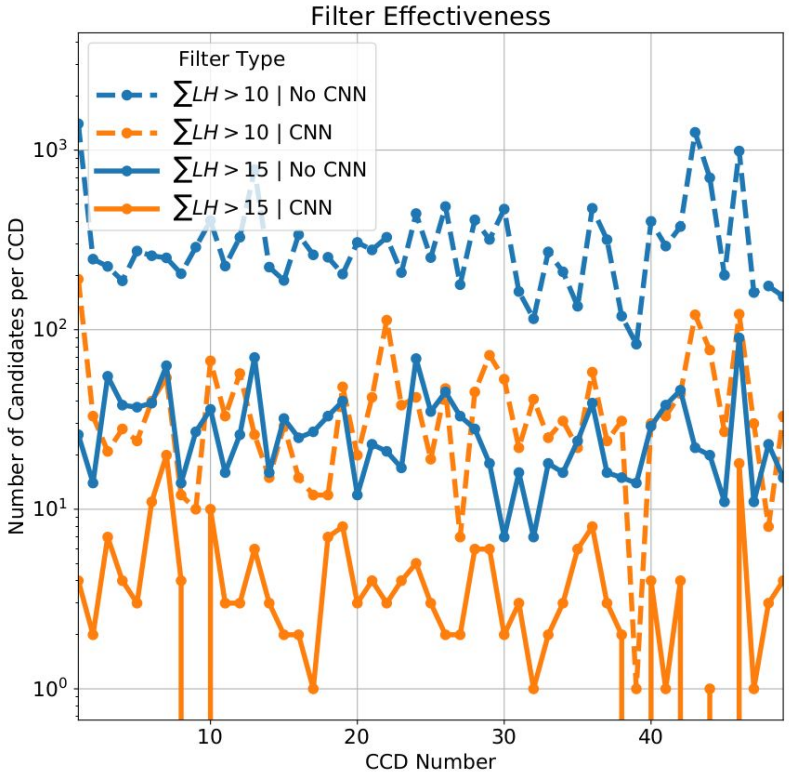
https://github.com/dirac-institute/kbmod

The KBMOD algorithm

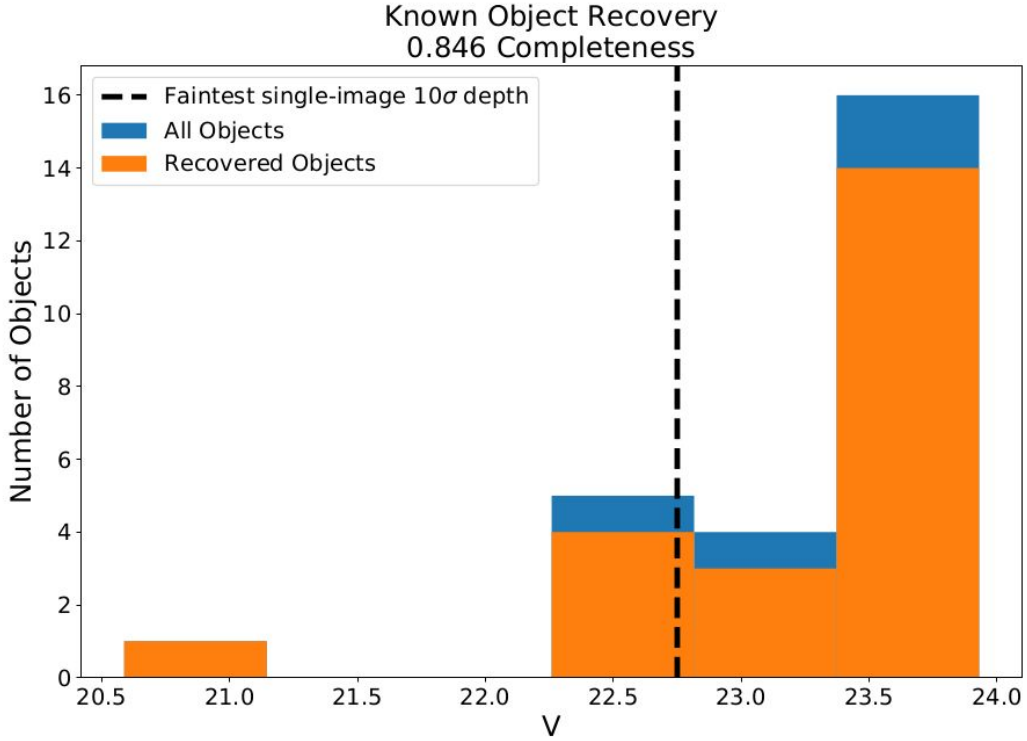
- Add masks to images, or use difference imaging
- Convolve images with PSF to create Maximum Likelihood Images
- Sum the likelihoods of trajectory locations of the individual images
- KBMOD can search >10^10 trajectories in about a minute on 10-15 4Kx4K images using a 1080 GPU



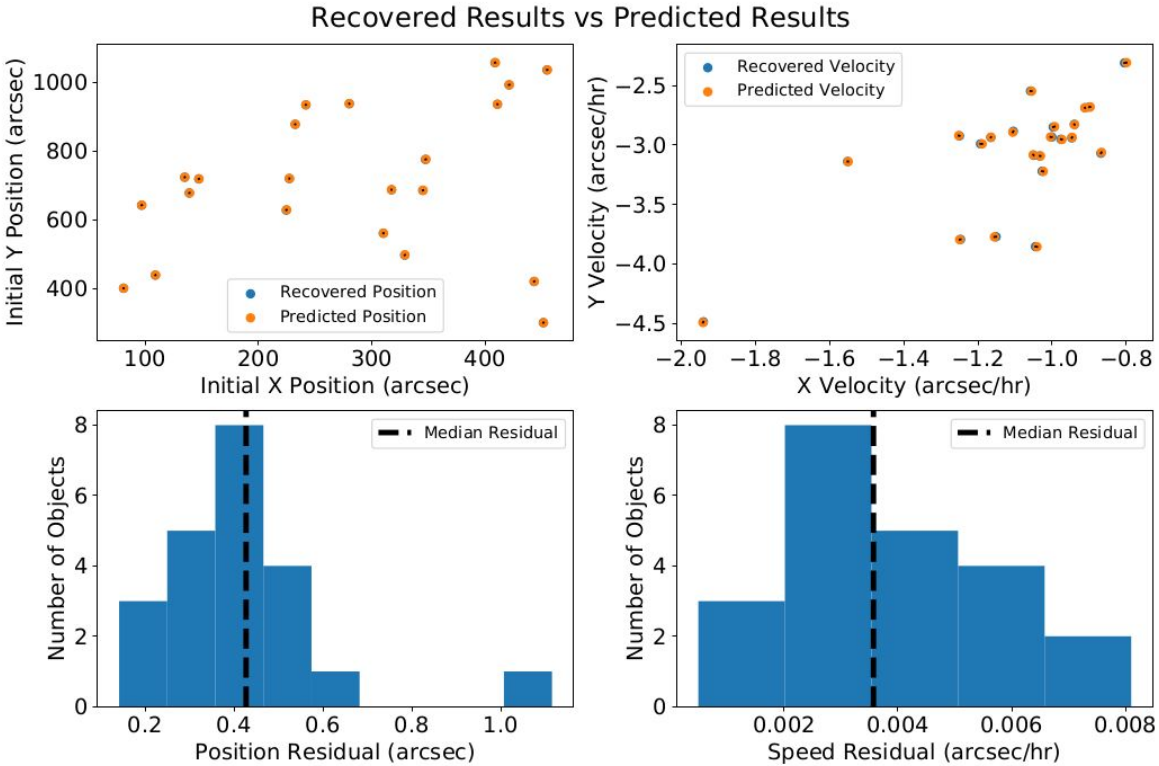
CNN Filter of False Positives



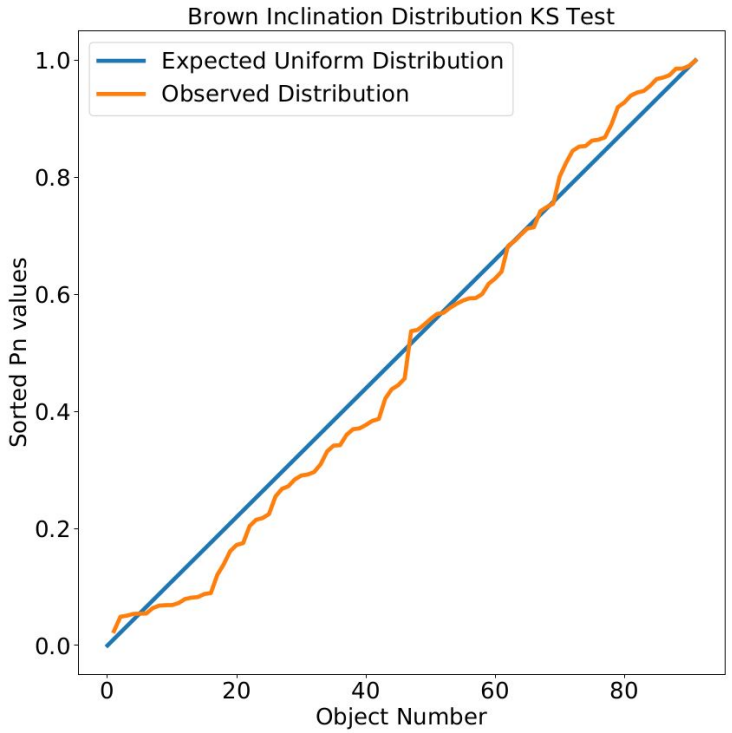
Recovery as a Function of Known V Mag

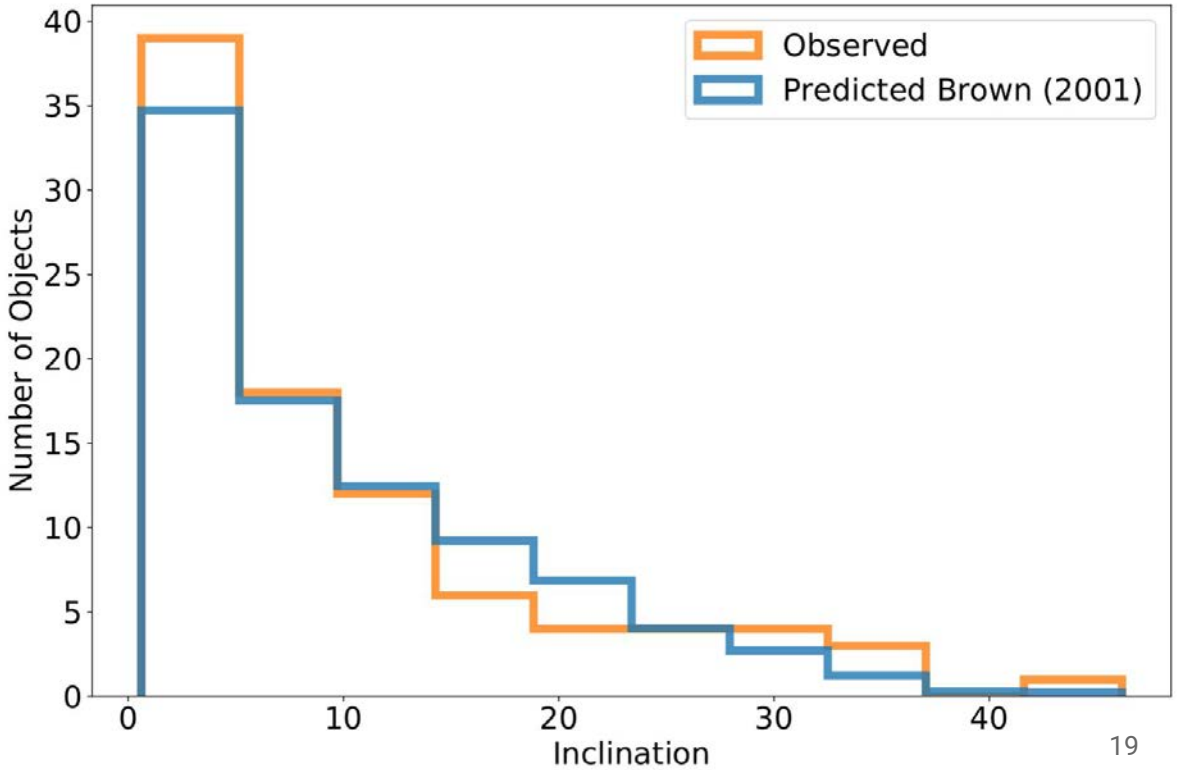


Comparison to Known Data



Inclination Distribution





Comparison to Known Magnitude Distribut

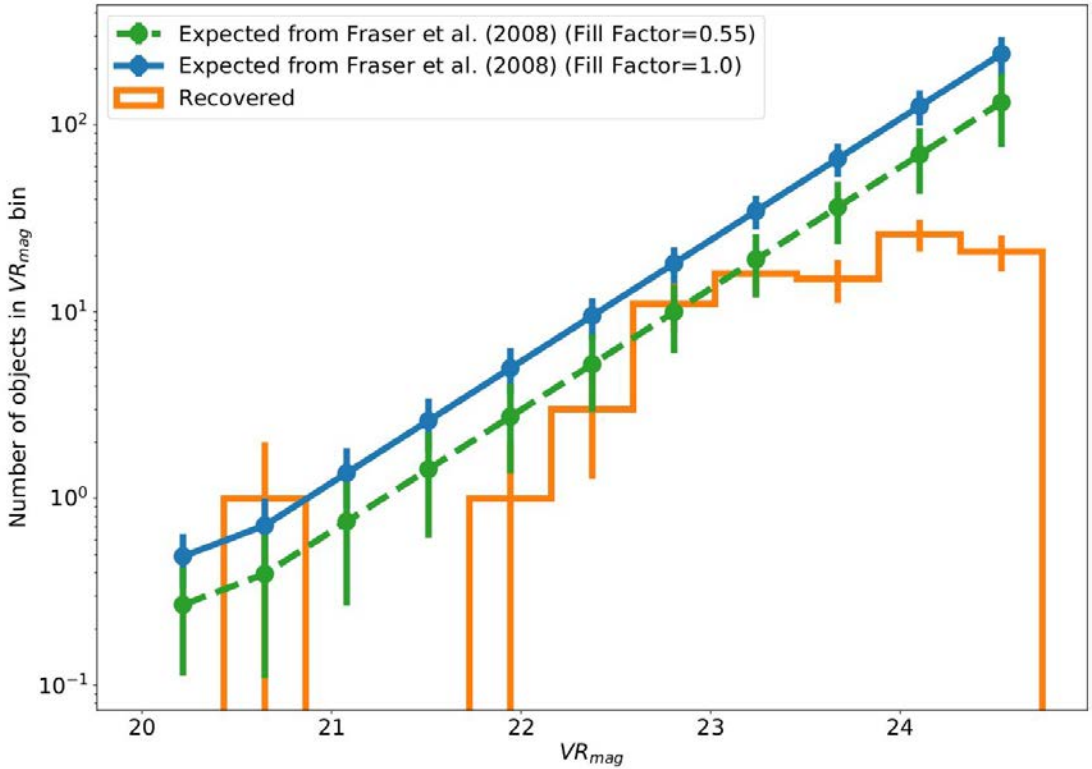


Table of Contents:

Summary and Challenges

- **It's not just the volume of data; it's the complexity of the science**
  - We should focus on how the physics of scientific questions might simplify the computational models.
  - Data quality can place important yet hard-to-characterize constraints on large-scale analysis.

- **Scaling science is more than just more hardware**
  - Tools don't always exist that are robust enough to scale for a typical science analysis.
  - Many of the scalable analysis frameworks do not scale easily with memory constraints.

Questions?