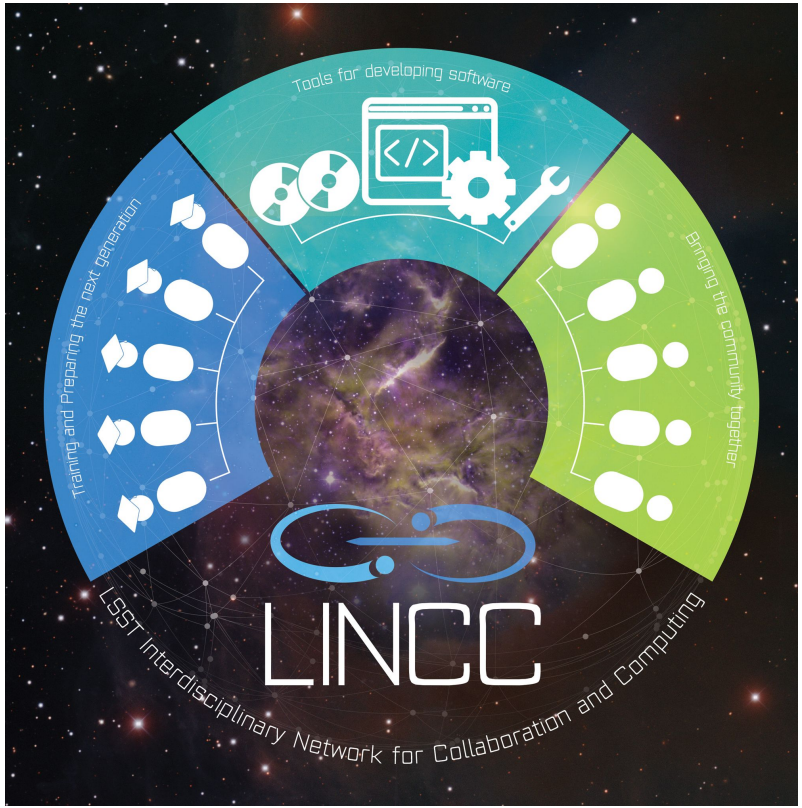
HiPSCat/LSDB Overview

ADASS Tutorial
Samuel Wyatt
11/05/2023

LINCC

- **L**SST **I**nterdisciplinary **N**etwork for **C**ollaboration and **C**omputing
- Science Frameworks:
  - Scalable Spatial Analysis (**LSDB**)
  - Time Domain (TAPE & **LSDB**)
  - Scalable Faint Object Detection (KBMOD)
  - Comprehensive Photo-Z infrastructure (RAIL)



The LINCC Frameworks Project

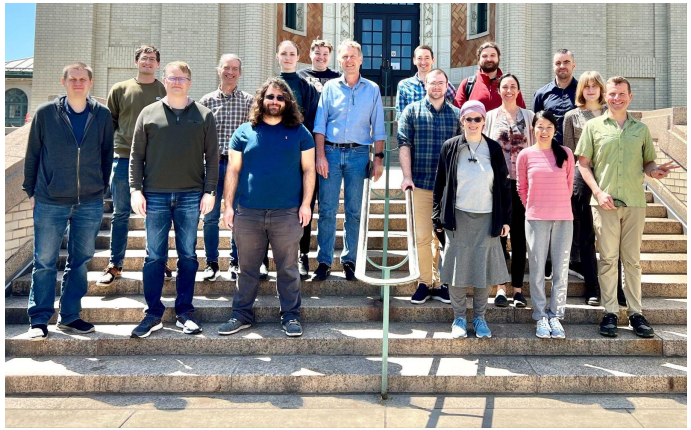
LSST Interdisciplinary Network For Collaboration And Computing

LSST Science Pipelines

A collaboration between UW, CMU, LSSTC, U Pitt, and NOIRLab to build software, frameworks, and systems for key LSST science.

PIs: Andy Connolly (UW), Rachel Mandelbaum (CMU)

Director of Engineering: Jeremy Kubica (CMU)



https://www.lsstcorporation.org/lincc/frameworks

LINCC Frameworks is supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Institute of Astrophysics (VIA)

LINCC Frameworks Mission

The LINCC Frameworks team's mission is to enable scientists by developing scalable and productionised software/algorithms in collaboration with broader community.

We want to:

- be engineering and algorithmically focused,
- collaborate with other software efforts (projects may be contributions to existing code bases),
- leverage existing tools (build on top of the Rubin Science Platform and standard community tools/libraries), and
- coordinate with community to avoid unnecessary duplication of effort.

Workshop: From Data to Software to Science with the Rubin Observatory LSST

Goal: Enable *interactive development* of exciting scientific use cases for early LSST data, and identifying the common computational/technical challenges and enabling technologies associated with them.

	Cross-matching	Photo-z	Selection functions	Time series	Image reprocessing	Image analysis
Cosmology	✓✓	✓✓	✓✓	✓✓	✓	✓
Extragalactic static	✓✓	✓✓	✓✓		✓✓	✓
Extragalactic transient	✓✓	✓✓	✓	✓✓	✓	✓
Extragalactic variable	✓✓	✓	✓	✓✓	✓	✓
Local Universe transient & variable	✓✓		✓	✓✓		
Local Universe static	✓✓		✓✓		✓	✓
Solar system	✓		✓✓	✓✓	✓	✓✓

White paper: https://arxiv.org/abs/2208.02781

Workshop: From Data to Software to Science with the Rubin Observatory LSST

Goal: Enable *interactive development* of exciting scientific use cases for early LSST data, and identifying the common computational/technical challenges and enabling technologies associated with them.

	Cross-matching	Photo-z	Selection functions	Time series	Image reprocessing	Image analysis
Cosmology	✓✓	✓✓	✓✓	✓✓	✓	✓
Extragalactic static	✓✓	✓✓	✓✓		✓✓	✓
Extragalactic transient	✓✓	✓✓	✓	✓✓	✓	✓
Extragalactic variable	✓✓	✓	✓	✓✓	✓	✓
Local Universe transient & variable	✓✓		✓	✓✓		
Local Universe static	✓✓		✓✓		✓	✓
Solar system	✓		✓✓	✓✓	✓	✓✓

White paper: https://arxiv.org/abs/2208.02781

HiPSCat & LSDB

New LINCC Frameworks development to enable

- 1. Scalable Cross-matching
- 4. Scalable job execution system

HiPSCat

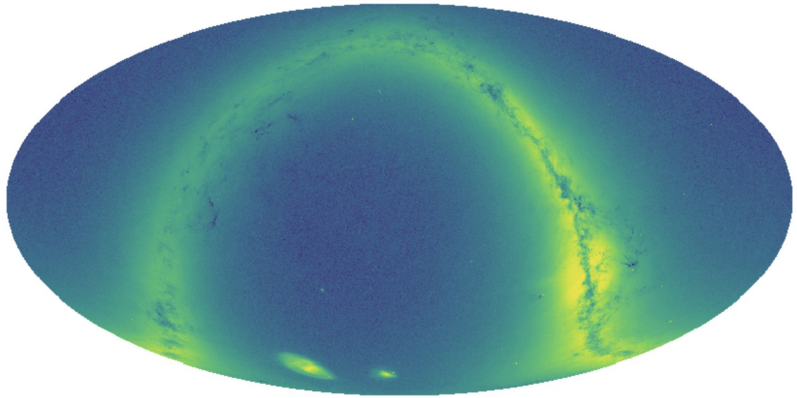
- Proposed file structure for spatially partitioning astronomical catalogs

LSDB

- Python library for HiPSCat analytics

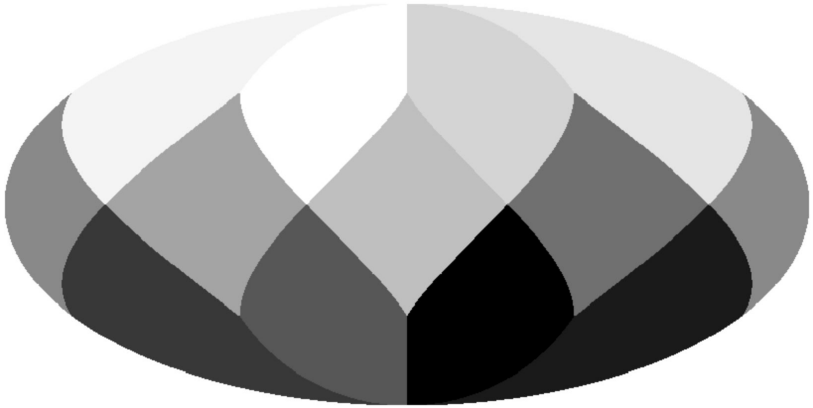
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?



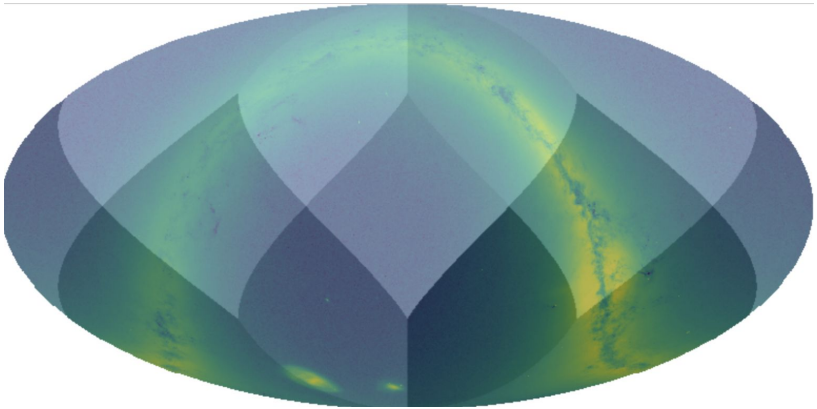
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Static Healpix?



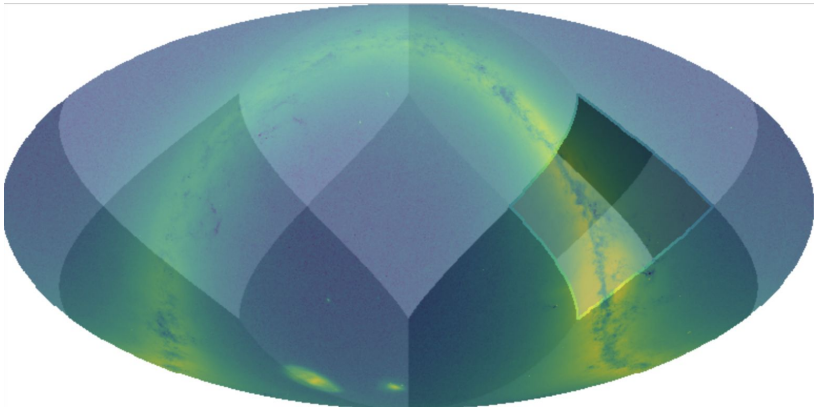
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Static Healpix?
    - Not the most efficient at dense areas



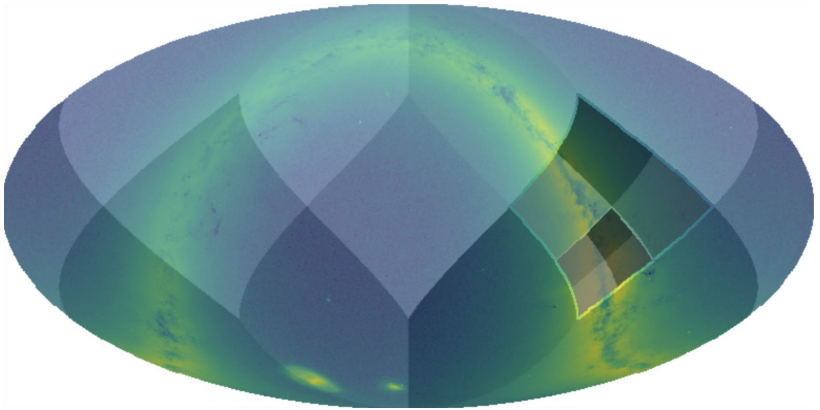
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Dynamic healpix?
    - recursive splitting based on source density at a threshold (row size, or source number)



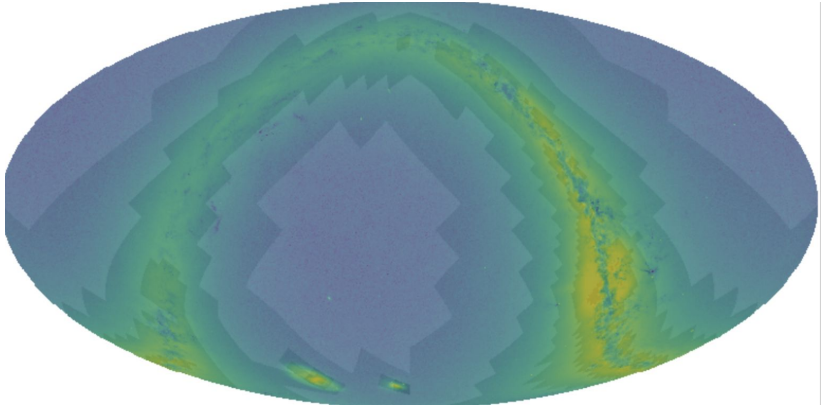
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Dynamic healpix?
    - recursive splitting based on source density at a threshold (row size, or source number)



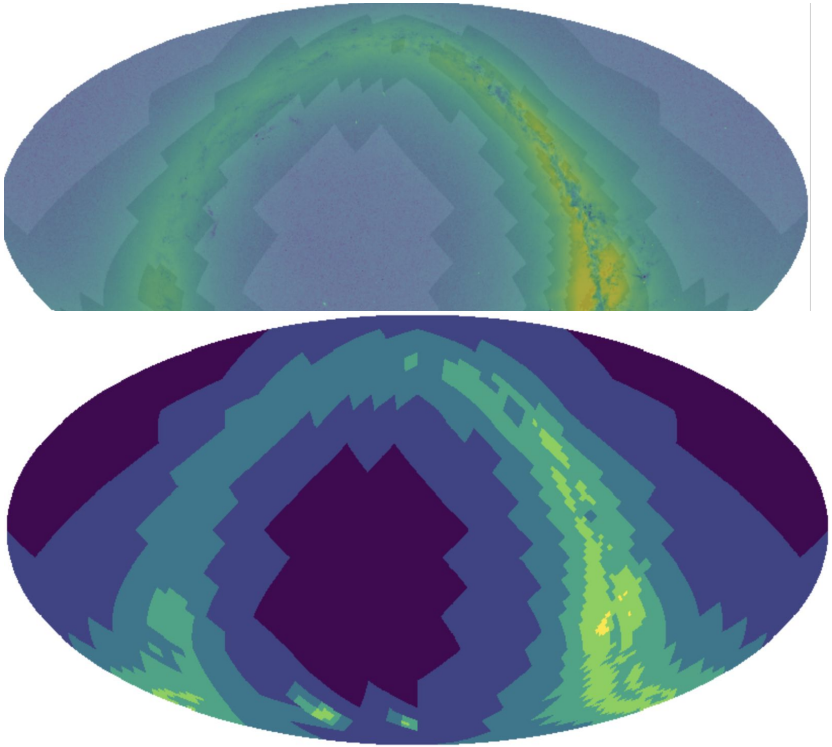
HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Dynamic healpix?
    - recursive splitting based on source density at a threshold (row size, or source number)



HiPSCat

- How do we plan to store/access LSST sized object and source catalogs?
- **Spatial Partitioning**
  - Dynamic healpix?
    - recursive splitting based on source density at a threshold (row size, or source number)
    - for gaia_dr3: ~4000 partitions at a threshold of 1 million



HiPSCat

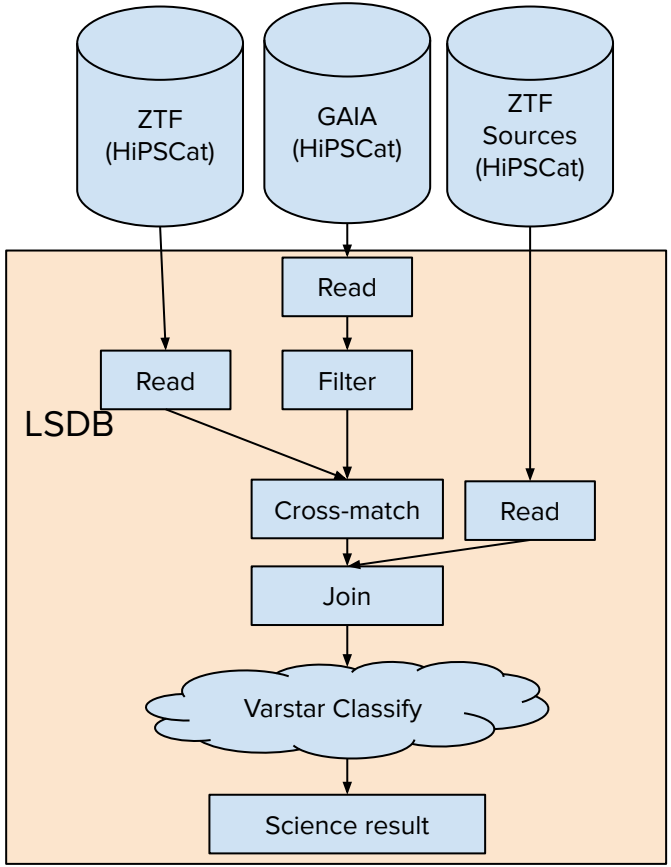
- How do we plan to store/access LSST sized object and source catalogs?
- **Storage**:
  - Apache Parquet files seem like the best bet (compression vs. speed)
  - Each pixel corresponds to a partition file
    - 4000 pixels = 4000 files
  - On-disk organization: HiPS-like (parquet hive partition schema)
    - `/Norder=0/Dir=0/Npix=0.parquet`
    - `/Norder=5/Dir=10000/Npix=10000.parquet`
- **NOTE: Hi**erarchical ~~**P**rogressive~~ **S**urvey -> Data is only stored in leaf nodes = No lower resolution data at lower orders. (new name?)

HiPSCat

- What can we do with this?
  - Anything that can read parquet can interact with HiPSCats
    - Dask, Ray, Hive, Hadoop …
  - Download subsets of large catalogs
    - given an arbitrary region, it's trivial to find which partitions that cover the region and download those files
    - parquet allows for easy/efficient selection of specific columns
  - Scalable analysis of single HiPSCats (parallel per-file analysis)
    - Complex searches (spatial and columnar comparison)
    - Feature computation (creating new columns, or mapping ufuncs on partitions)
  - Scalable analysis of many HiPSCats
    - Partitioning enables scalable **Cross Matching** and **Joins** inherently

LSDB

- **L**arge **S**urvey **D**ata**b**ase: Astronomy aware layer for HiPSCat
  - Python analysis library built on dask (parallelized pandas) with catalog cross-matching as fundamental use
  - The framework handles parallelization and should unlock individual researchers to perform robust analysis on full LSST data
  - Goal: Perform parallelized spatial join of O(10B) object catalogs and enable downstream analysis of the join result



LSDB

- **L**arge **S**urvey **D**ata**b**ase: Astronomy aware layer for HiPSCat
  - Python analysis library built on dask (parallelized pandas) with catalog cross-matching as fundamental use
  - The framework handles parallelization and should unlock individual researchers to perform robust analysis on full LSST data
  - Goal: Perform parallelized spatial join of O(10B) object catalogs and enable downstream analysis of the join result

```
img = gaia
      .query("pm > 10")
      .crossmatch(ztf)
      .join(ztf_sources)
      .for_each(varstar_classify)
      .query("pRRLy > 0.95")
      .skymap()

hp.mollview(img)
```

LSDB

- Tutorial Notebooks
  - ADASS Tutorial

Immediate plans

- HiPSCat
  - Functional catalogs we're testing with
  - Working with archive partners to test format on variety of catalogs
  - Incorporating IVOA standards into metadata and API
- LSDB
  - Scalable cross-matching with dask
  - Propagation of metadata through pipelines
  - Integration with TAPE (LINCC Frameworks **T**imeseries **A**nalysis & **P**rocessing **E**ngine)
  - Focus on usability and maintainability
- Find more info: https://github.com/lincc-frameworks/docs/wiki/LSDB
- Follow along by joining: https://groups.google.com/g/hipscat-wg

LSDB

- **L**arge **S**urvey **D**ata**B**ase
- Supporting LSST science questions requires key functionality in an analysis framework with the ability to:
  - Store and manipulate catalog data at scale
  - Perform distributed computation over this data
  - Use spatial structure within searches and statistical computation
  - Interoperate with data from other surveys
  - Access these catalogs without having to directly download them.