An Application of HEP Track Reconstruction Methods to Gaia EDR3

CTD 2023

Mine Gokcen[1], Maurice Garcia-Sciveres[2], Xiangyang Ju[2]

1. Istanbul Technical University
2. Lawrence Berkeley National Lab

Contents

Introduction

Galactic Archeology

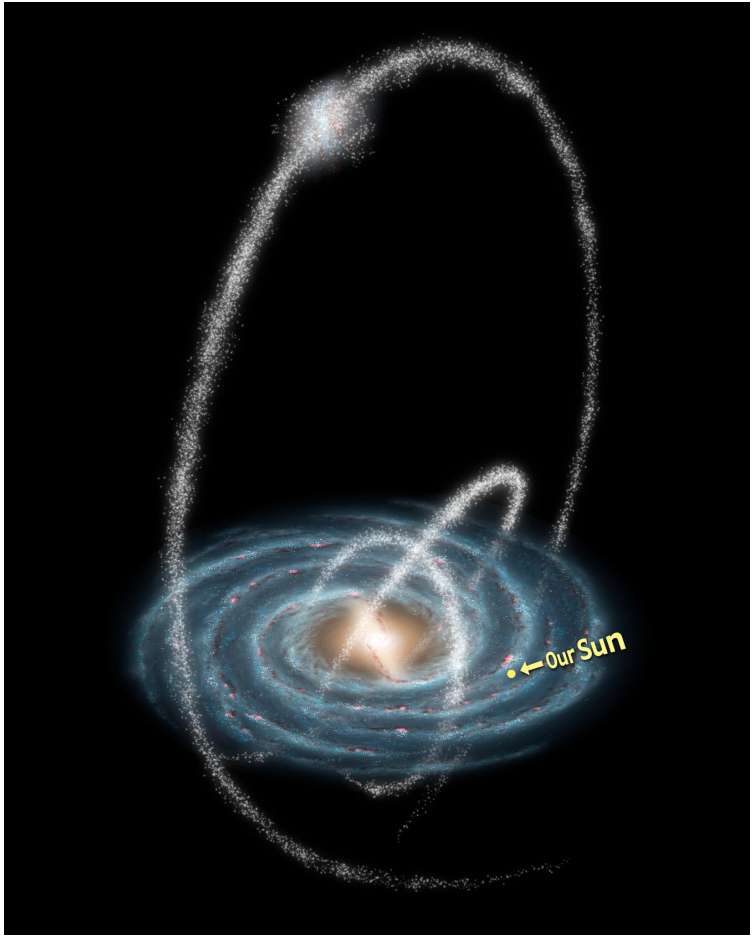
- Studies of galactic structure and objects in Milky Way (MW)
- To understand the origins of our galaxy

Galactic Structure and Stellar Streams

- Galaxies reside inside a dark matter (DM) halo
- DM halos play a major role in galaxy formation and evolution
- Different stellar populations provide insights into the DM distribution across our galaxy through their progenitors' merging histories.
- In particular, orbits of *stellar streams* show accretion patterns of new matter into our galaxy.

Stellar Streams (Cont.)

- Stellar streams form when stars from satellite star clusters get tidally disrupted
- Stream stars have different kinematics than that of background stars
- They also have distinct chemical compositions and ages



Our Research Question & Connection to Tracking

- Can we apply HEP track reconstruction methods to stars to find stellar streams and stars from other distinct stellar populations?
- To what extent can we automate our process to have as little human input/supervision as possible?
- We can establish the following analogy:
    - Detector hits → stars
    - Space points → kinematics of stars
    - Seeds → groups of stars with similar kinematics
    - Track fitting → finding seeds within vicinity of each other

Our Research Question & Connection to Tracking (Cont.)

- Today, we show our work in progress of an application of HEP track reconstruction methods to astronomical data.
- We will talk about;
  - ML-based Seeding
  - Seed aggregation -- equivalent to duplicate removal in HEP tracking

Related Works

- Many other groups that work on statistical and computational methods to find streams!

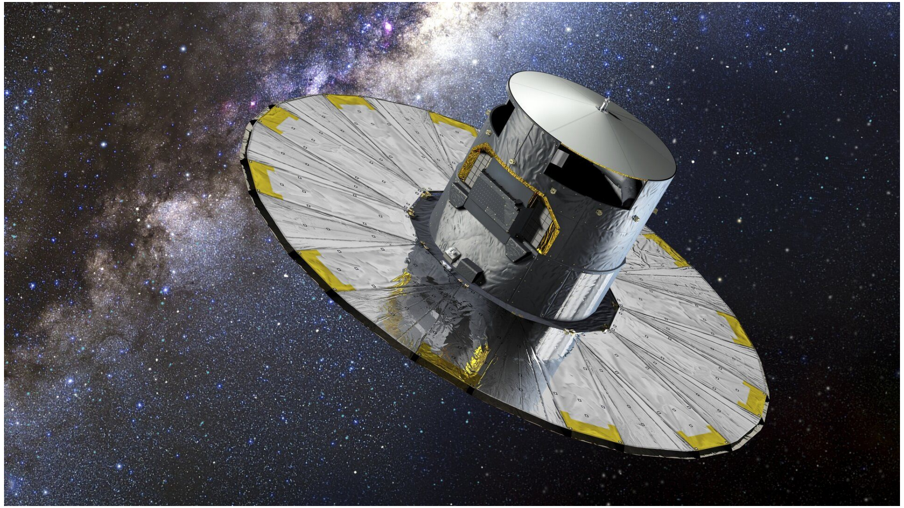
Some references:

  - M. Pettee et al. (2023) -- uses weakly-supervised anomaly detection
  - A. Bonaca et al. (2014) -- uses MCMC
  - N. Shipp et al. (2023) -- uses numerical simulations
  - V. Chandra et al. (2023) -- discovery of a gaseous stream
  - D. Shih et al. (2023) -- uses deep learning anomaly detection
  - …

Dataset

Gaia Mission

- Gaia is a space observatory and a mission of the European Space Agency (ESA)
- Goal: 3D map of the MW
- Since its launch in 2013, had three major data releases.
- Gaia Collaboration et al. (2016b)

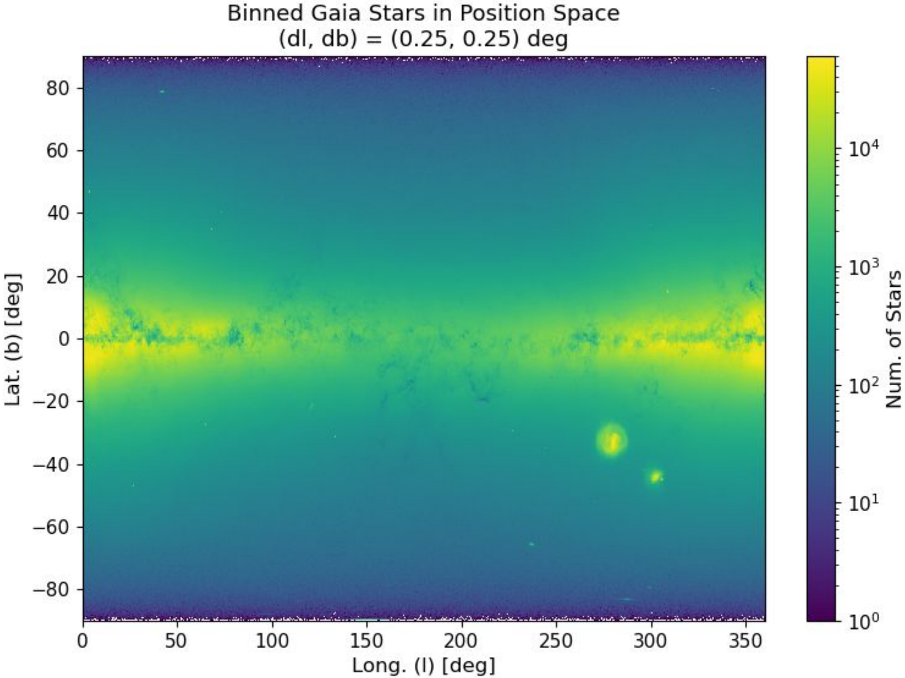


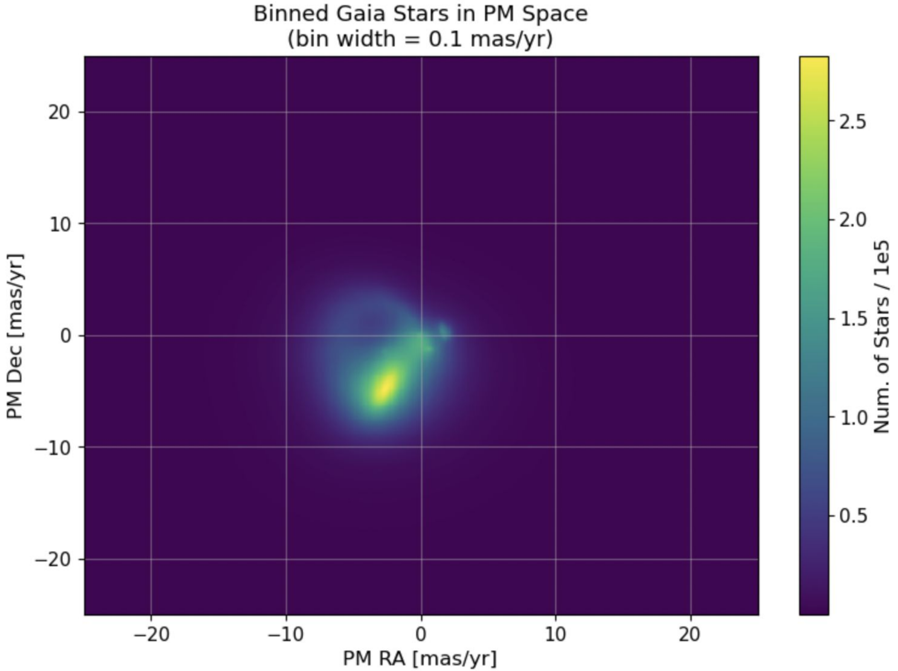
Gaia EDR3

- Catalog of astronomical objects (e.g.: stars, galaxies,...) and their properties.
  -- Gaia Collaboration et al. (2020a)
- Nearly 2 billion objects
- We look at their positional and kinematic properties
  - Longitude (l)
  - Latitude (b)
    } location
  - Proper motion in right ascension (PM RA)
  - Proper motion in declination (PM Dec)
    } angular velocity (kinematics)
  - Parallax ⟶ gives distance estimate

Catalog link: https://www.cosmos.esa.int/web/gaia/earlydr3

Gaia Stars in Position & Proper Motion Space

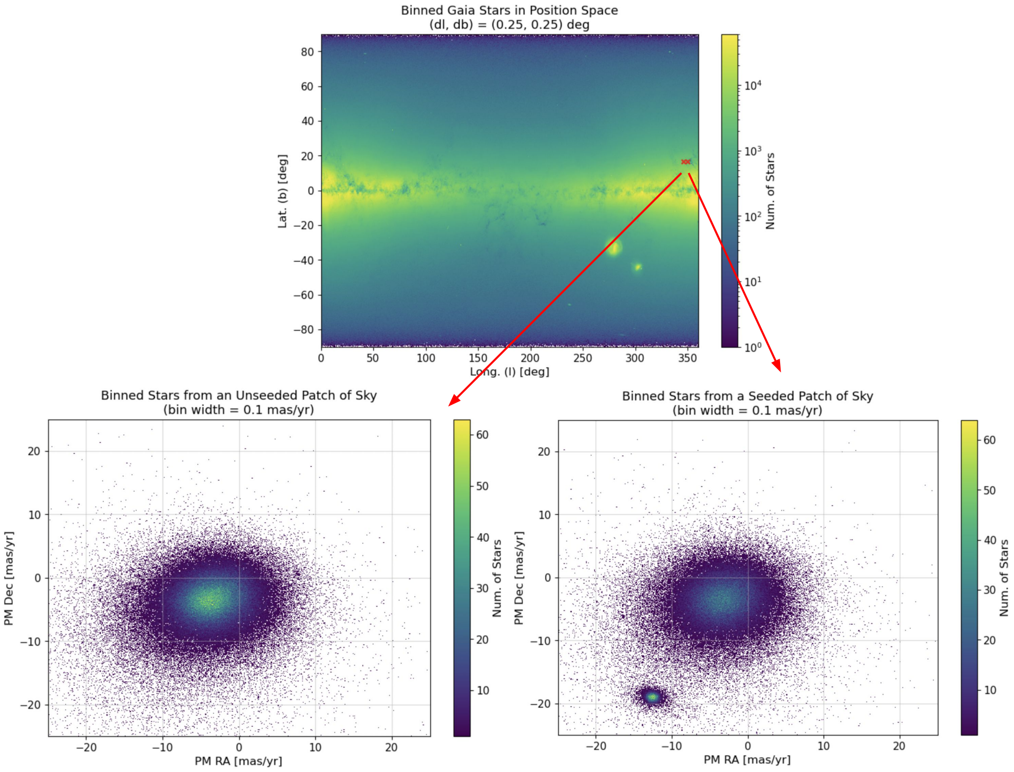




Seeding Stage

Proper Motion of Stars used for seeding

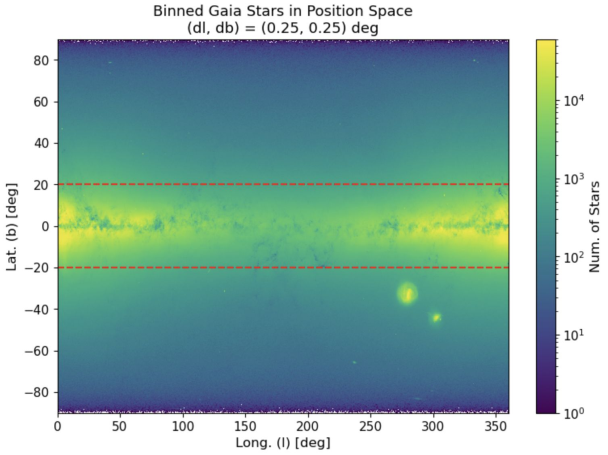
- Within a small patch of sky [(dl, db) ~ (1, 4) deg];
- stars from star clusters and streams have different kinematics than that of the background MW stars;
- meaning, they have a different statistical distribution in proper motion space.
- To form seeds, we will need to cluster stars based on their proper motions.

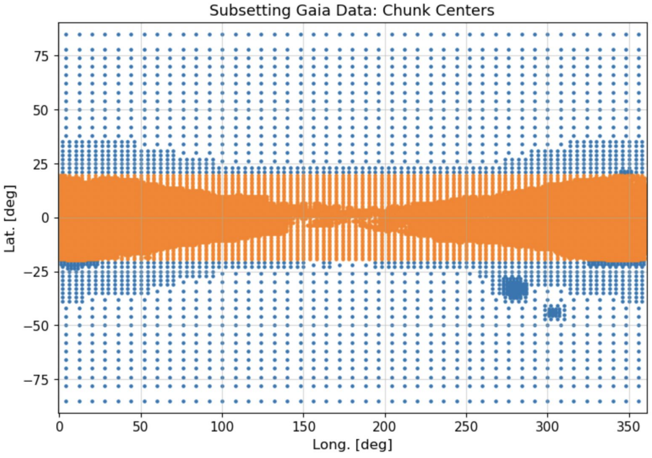


PM distribution in two neighboring small patches of sky

Subsetting Gaia Data: "Chunking"

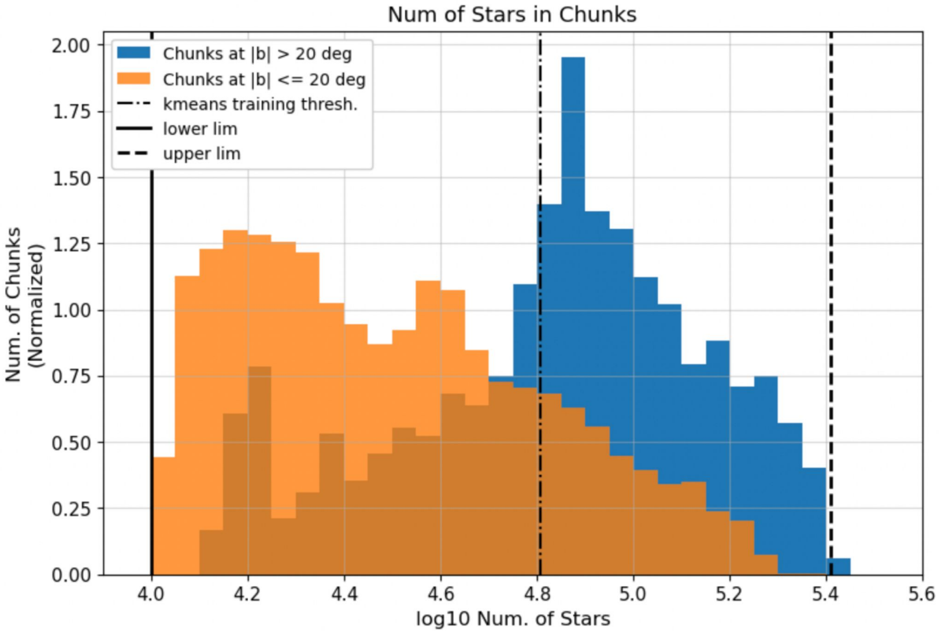
- Stream stars are locally in the same neighborhood → stars should be clustered locally for seeding
- Our *chunk* (2D patches of sky) sizes are picked based on the number of stars contained within, which is optimized for our algorithmic choices.
- We treat galactic disk (orange) separately.
- 2,500 + 30,500 ~ 33,000 chunks in total





Subsetting Gaia Data: "Chunking"

- Stream stars are locally in the same neighborhood → stars should be clustered locally for seeding
- Our *chunk* (2D patches of sky) sizes are picked based on the number of stars contained within, which is optimized for our algorithmic choices.
- We treat galactic disk separately.
- 2,500 + 30,500 ~ 33,000 chunks in total

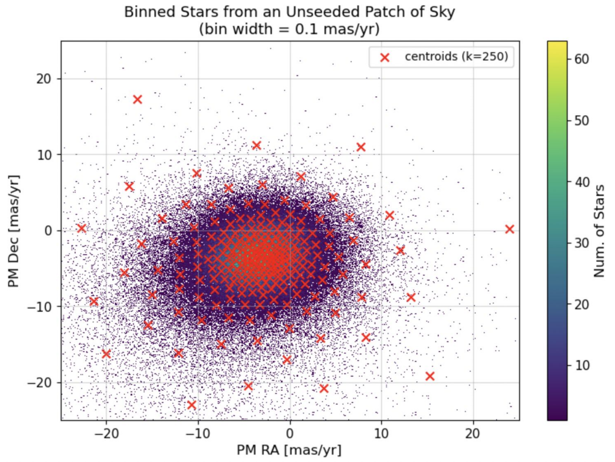


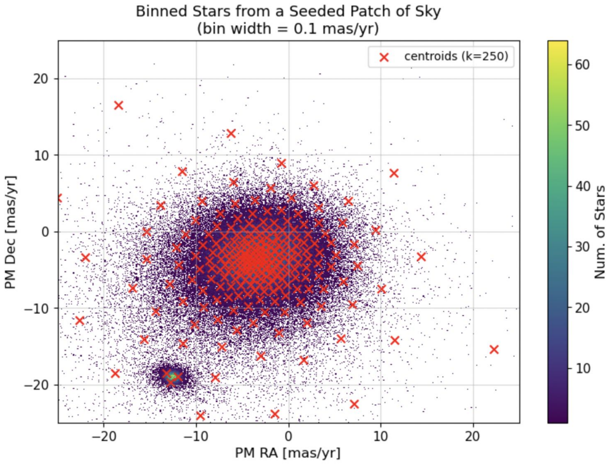
K-means and KNN for Clustering

- For the data clustering and classification, we use k-means and kNN algorithms from FAISS library.
- FAISS: Facebook AI Similarity Search
  - An open source library to conduct data mining over user data.
  - Extremely fast and powerful especially with GPU computing.
  - Clustering demo
- k-means: Suggests *k* centroids (cluster averages) that could represent the data based on how similar the parameter values are.
- kNN: Determines which points go into which centroids.

Clustering Proper Motion of Stars

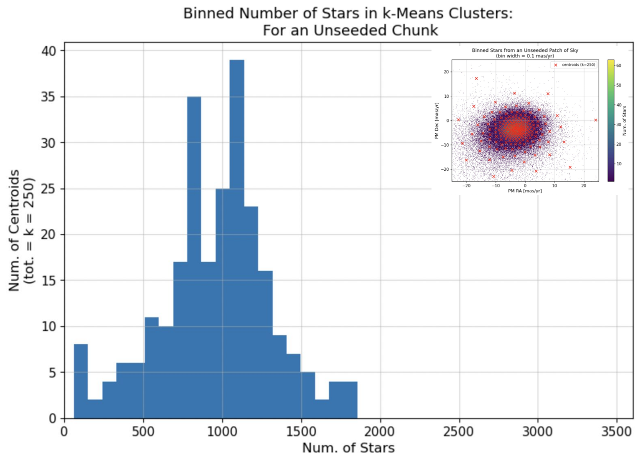
- We assume that stars in both the BG and the signal distributions follow a bivariate Gaussian distribution;
- and that signal-to-BG ratio is low enough so that;
- if we use a very high *k* for k-means, signal stars are clustered into very few centroids as opposed to the background stars.
- To test for seeds (explained in following slides), *k* needs to be high enough to create a statistical distribution.



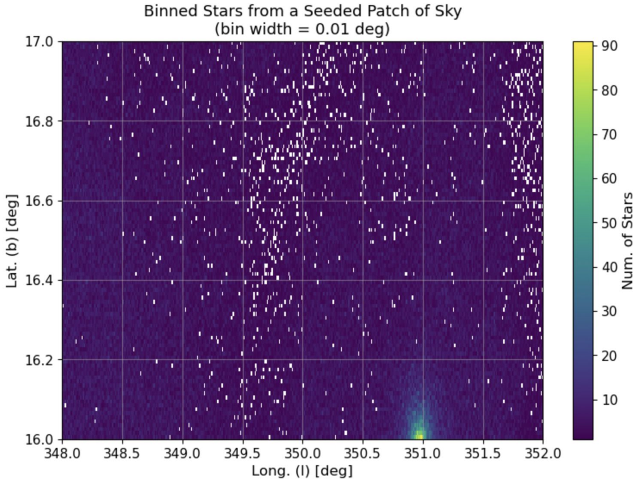


Determining If a Chunk Is Seeded

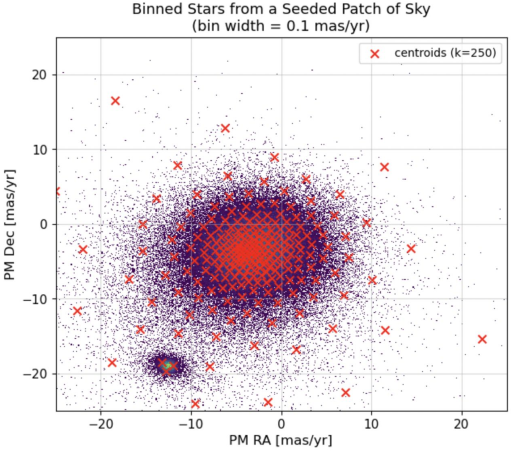
- For a given chunk, we bin the number of stars in each centroid.
- Chunks with signal stars should have an outlier in the distribution of number of stars in k-means centroids.
- To test for outliers, we look for gaps between the bins and the counts within final bins. (Exact test condition is optimized for 20 bins.)
- We have tried other methods to test for outliers, but this test also works for non-Gaussian distributions.

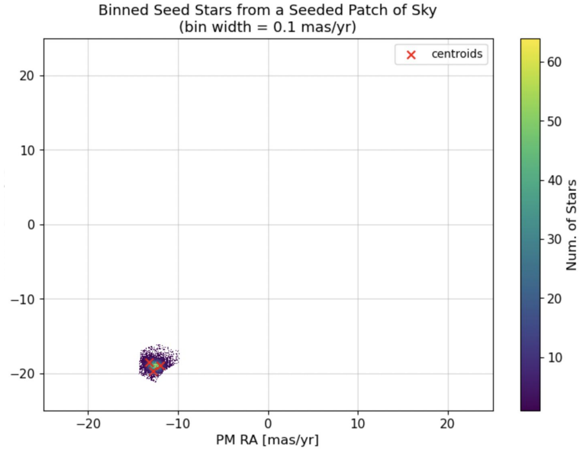


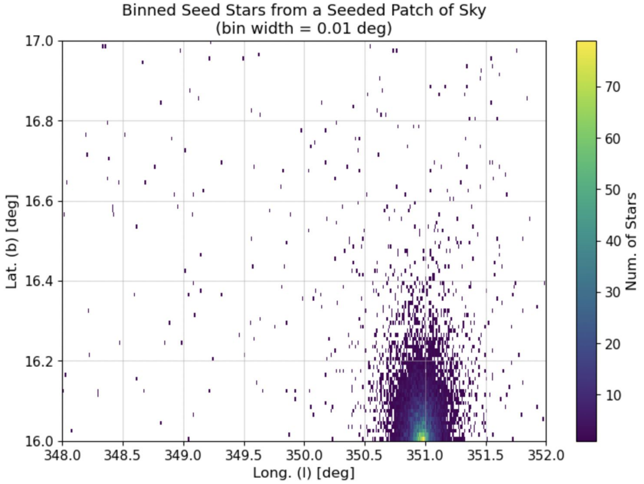
Seed Stars



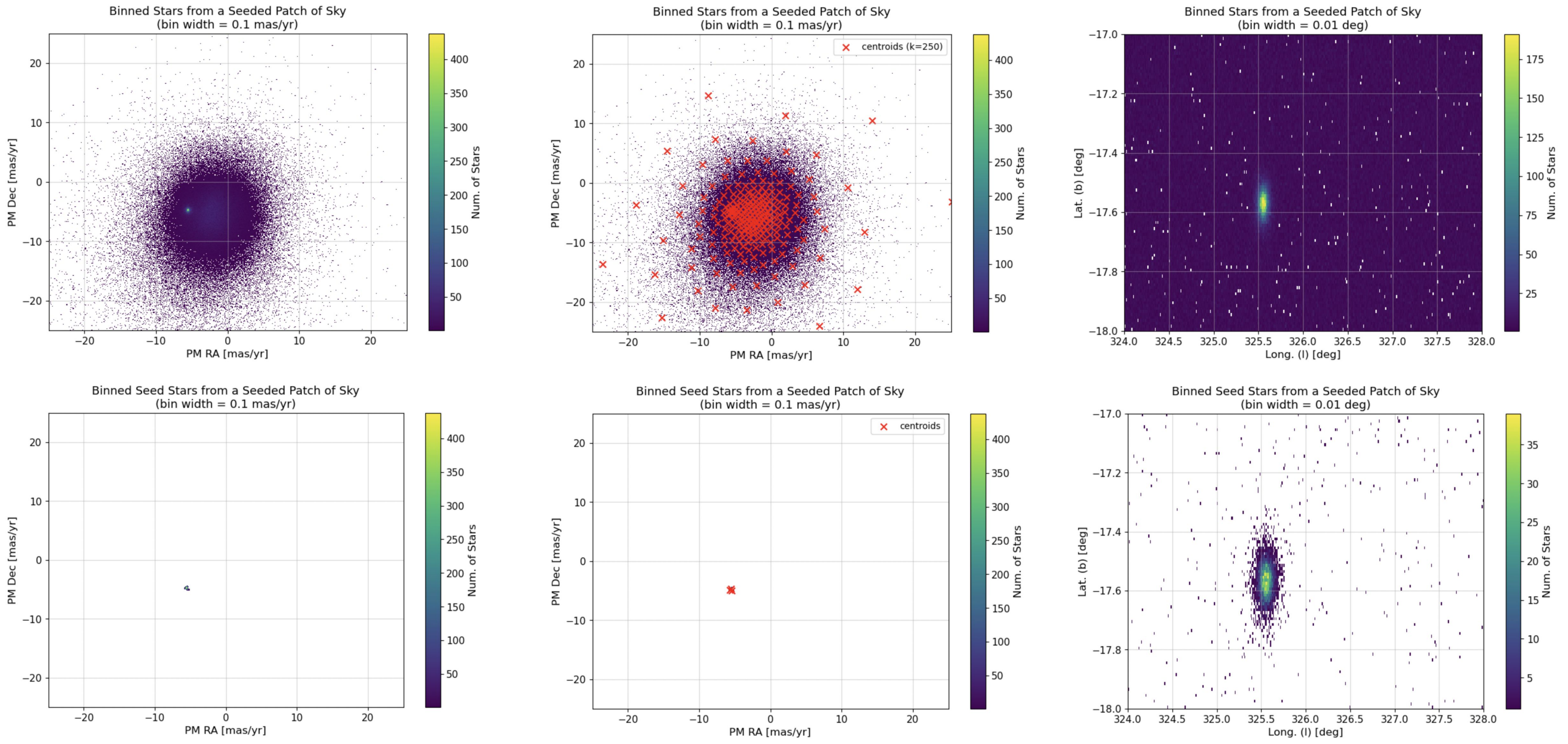
- If a chunk passes the seed test, we consider the stars within the top 3 k-means clusters with highest star counts to be seed stars.







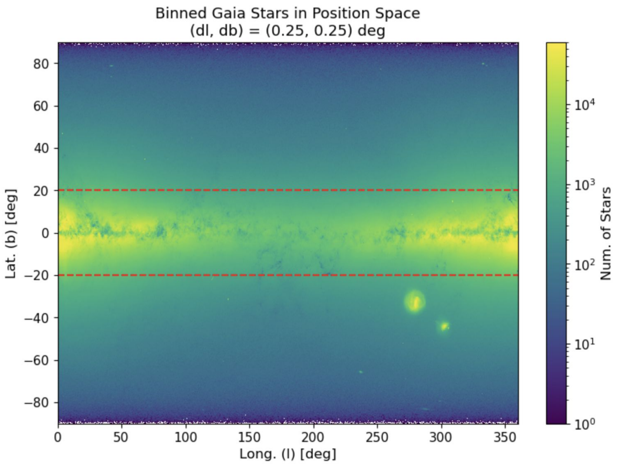
Another Example of a Seeded Chunk & Its Seed Stars



Results

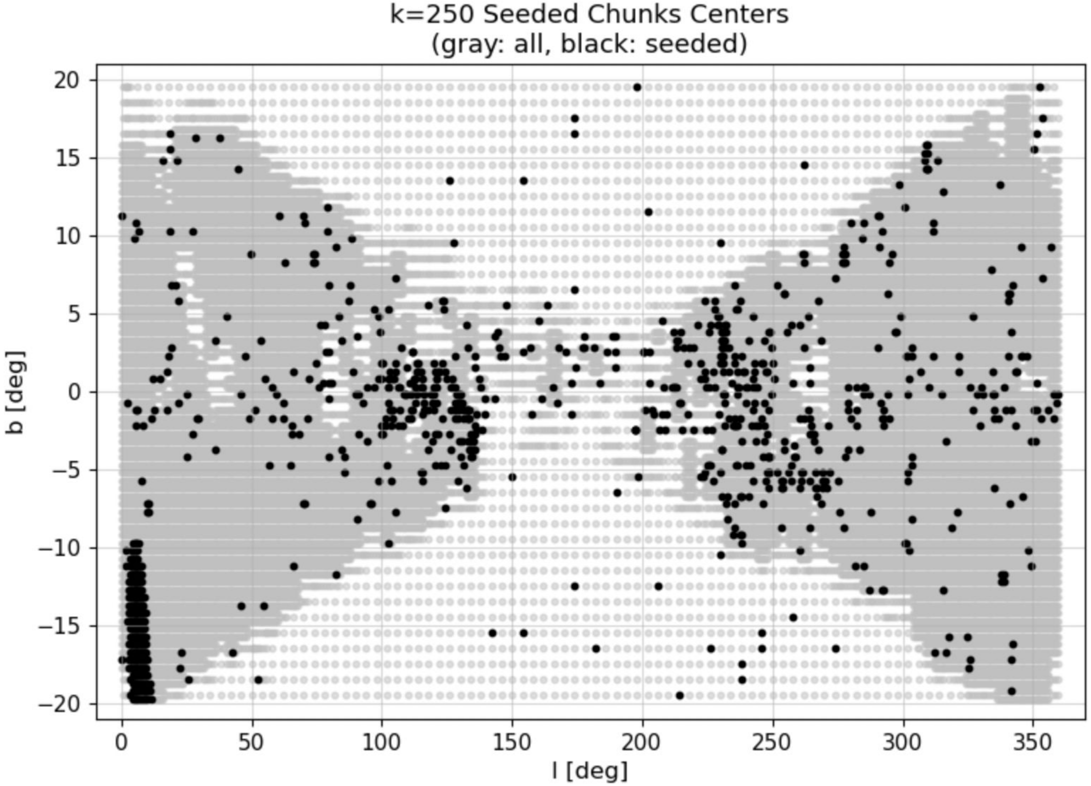
k=250 Seed Results: Summary Table

Today for the sake of time we will only go over the **galactic disk** (|b| < 20 deg) region results.

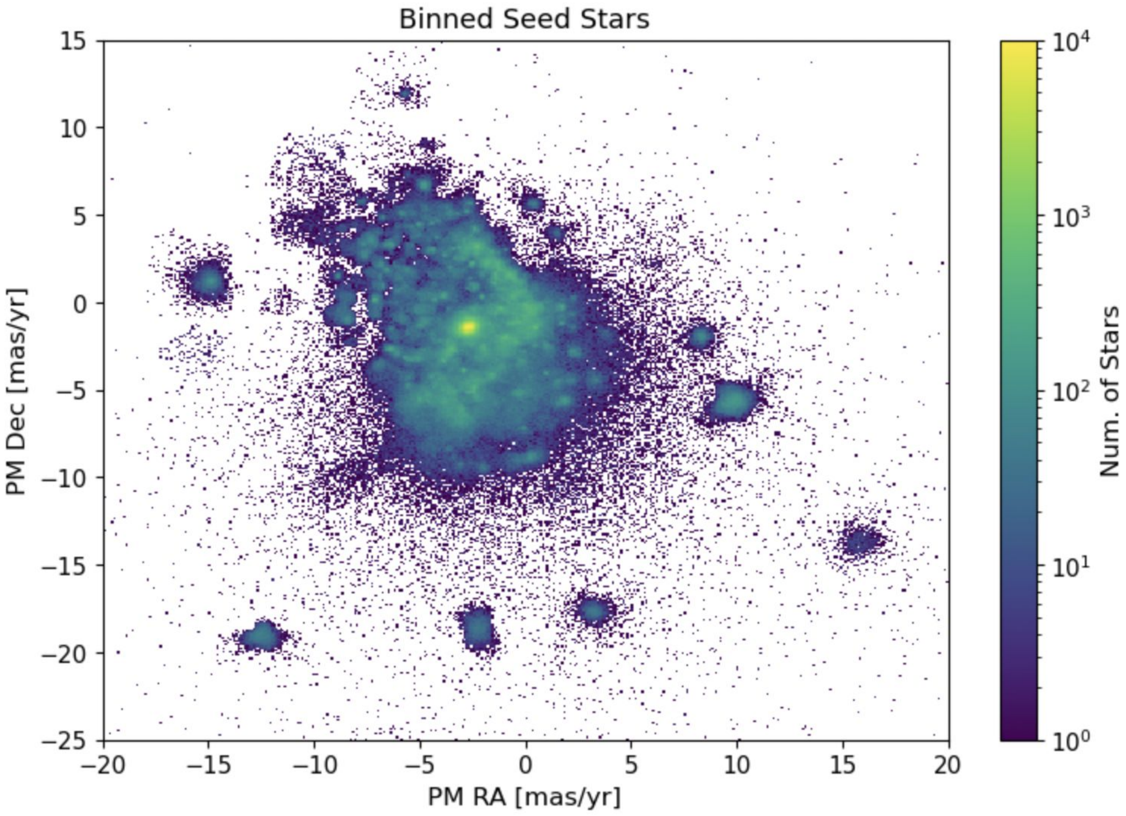


	# of chunks	# of seeded chunks	# of seeded chunks/ # of total chunks	# of seed stars
|b| > 20 deg	2,375	618	26.0 %	354,557
|b| < 20 deg	30,264	972	3.21 %	1,013,408

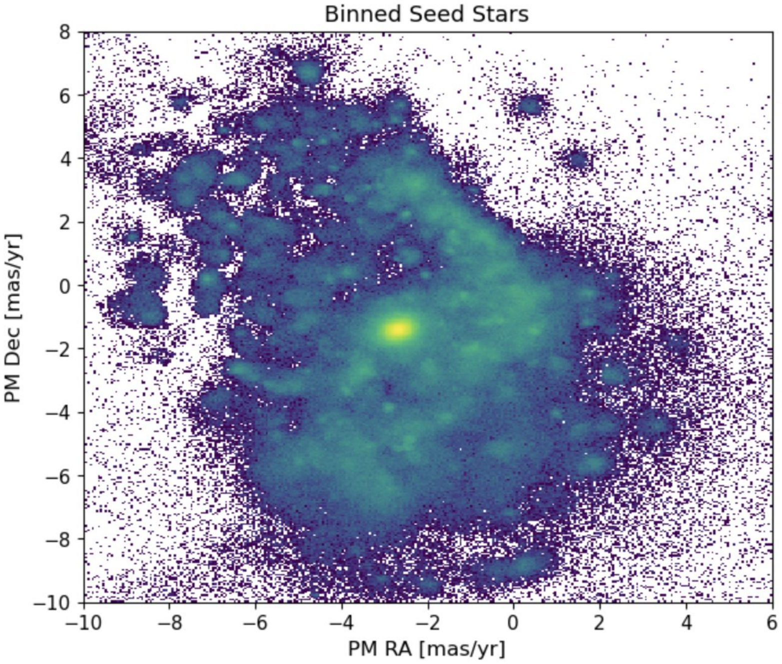
Seed Results in Position Space

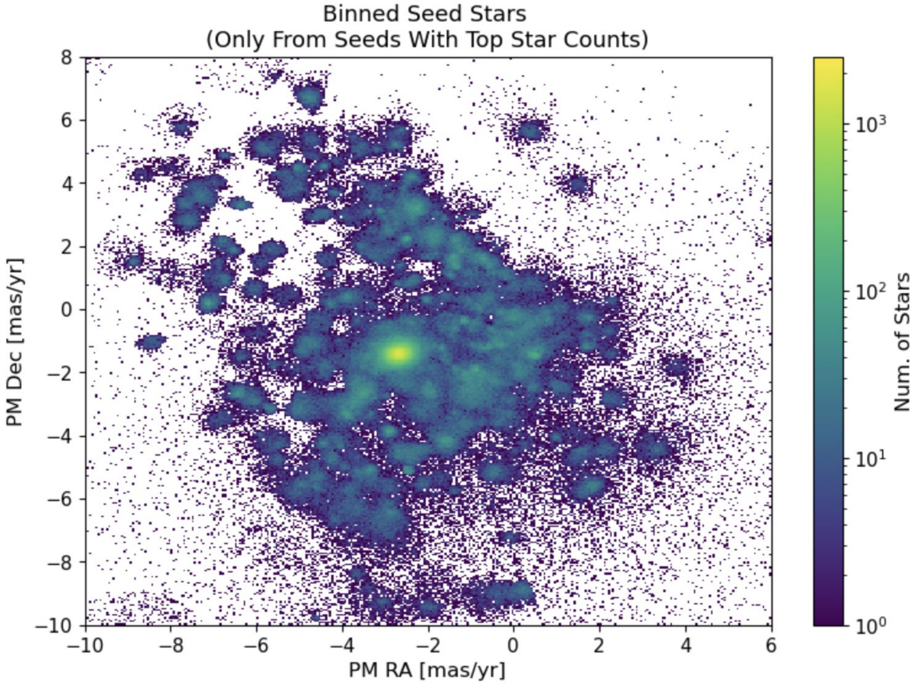


Seed Results in Proper Motion Space



Seed Results in Proper Motion Space (Cont.)



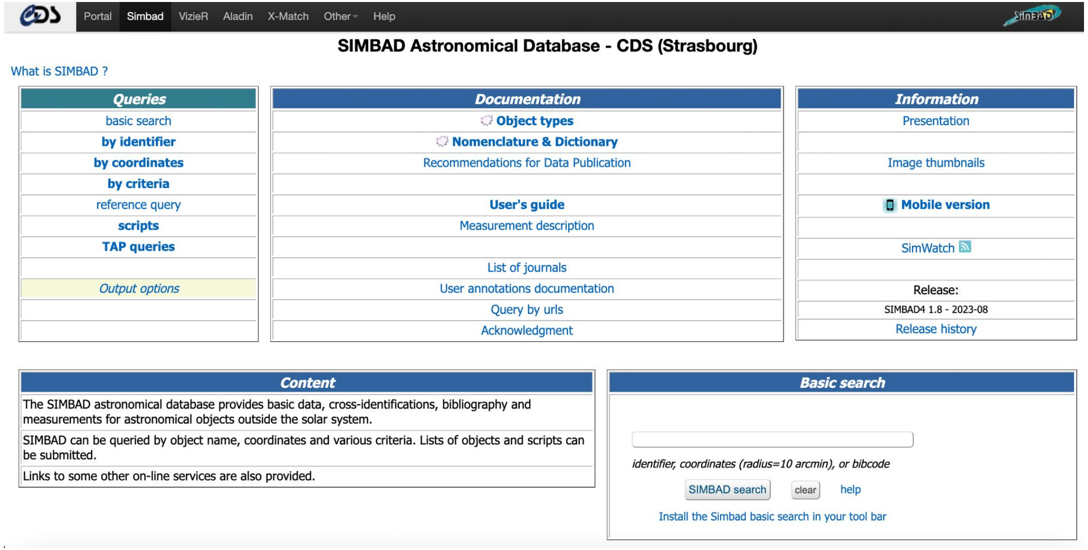


Re-Clustering Seed Stars & Seed Aggregation

- Once we obtain seed stars from various seed-search runs with different algorithmic choices,
- such as using a different *k* for k-means or phase-shifting our chunk centers,
- we combine seed stars and remove duplicates.
- We then re-cluster the seed stars into seeds again, this time using 4D k-means over position and proper motion variables.
- This process also can help us combine multiple seeds that actually represents the same physical object.
- (This is a work in progress.)

Comparing Seeds to Known Objects

- We use SIMBAD database to match our seeds to known objects.
- 2000,A&AS,143,9 , "The SIMBAD astronomical database", Wenger et al.



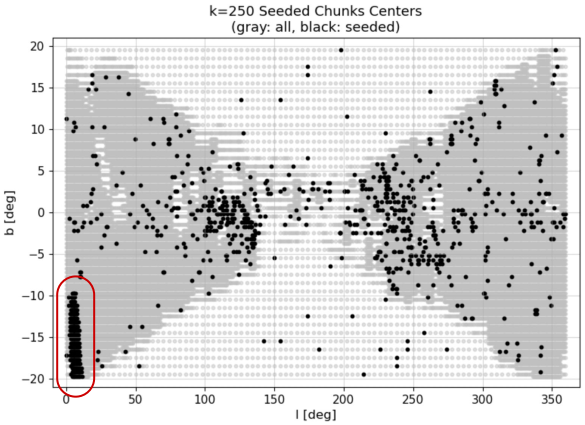
(Examples of) Potential Star Clusters Identified

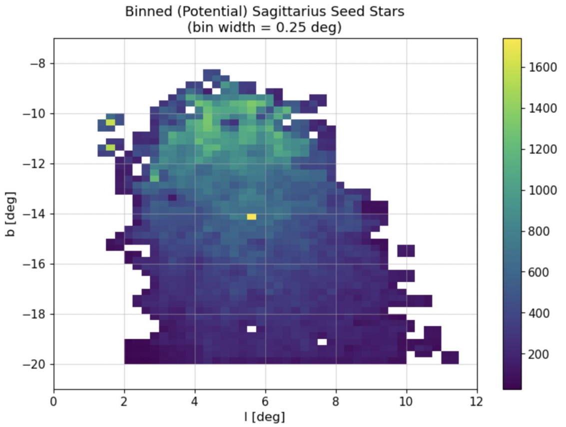
Sagittarius Dwarf Galaxy
and Stream

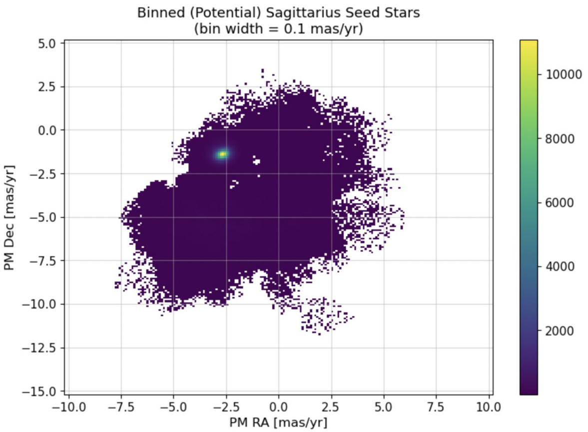
Basic data :

NAME SDG -- Galaxy

Other object types:	G (2014AJ)
ICRS coord. *(ep=J2000)* :	18 55 03.1 -30 28 42 [ ] D 2004AJ....127.2031K
FK4 coord. *(ep=B1950 eq=1950)* :	18 51 51.0 -30 32 34 [ ]
Gal coord. *(ep=J2000)* :	005.6081 -14.0858 [ ]
Proper motions *mas/yr* :	-2.650 -0.880 [0.080 0.080] D 1997AJ....113..634I







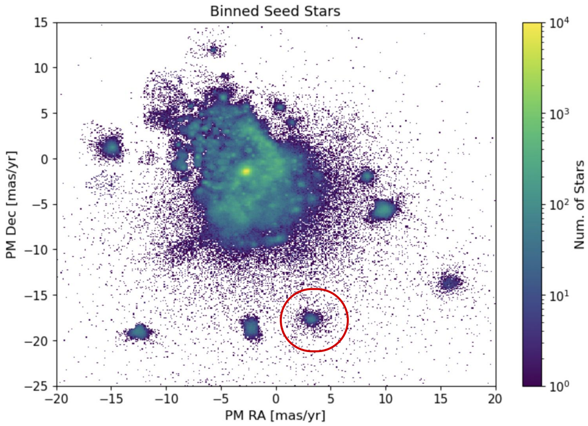
(Examples of) Potential Star Clusters Identified

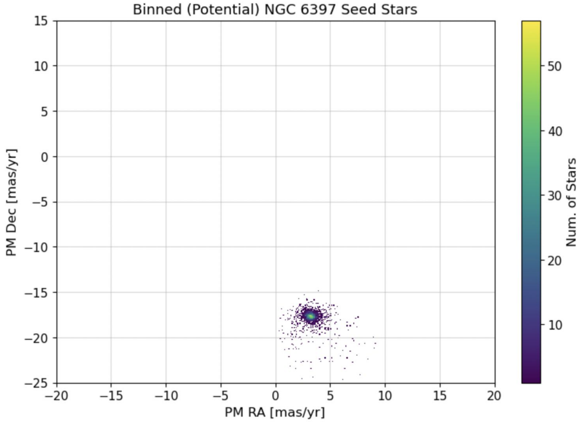
[NGC 6397]

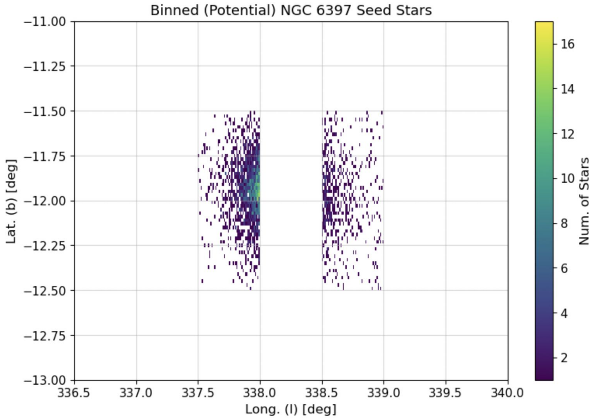
Basic data :

NGC 6397 -- Globular Cluster

Other object types:	**GlC** (2013A&A,GCl), **Cl*** (C,[KPS2012])
ICRS coord. *(ep=J2000) :*	**17 40 42.09 -53 40 27.6 (Optical) [ ] D 2010AJ....140.1830G**
FK4 coord. *(ep=B1950 eq=1950) :*	17 36 37.84 -53 38 53.5 [ ]
Gal coord. *(ep=J2000) :*	338.16501 -11.95952 [ ]
Proper motions *mas/yr* :	**3.30 -17.60 [0.01 0.01 90] C 2019MNRAS.482.5138B**
Radial velocity / Redshift / cz :	**V(km/s) 18.4 [0.1]** / z(spectroscopic) 0.000061 [0.000000] / cz 18.40 [0.10] **(Opt) A 2018MNRAS.478.1520B**
Parallaxes *(mas)*:	**0.416 [0.010] C 2021MNRAS.505.5978V**

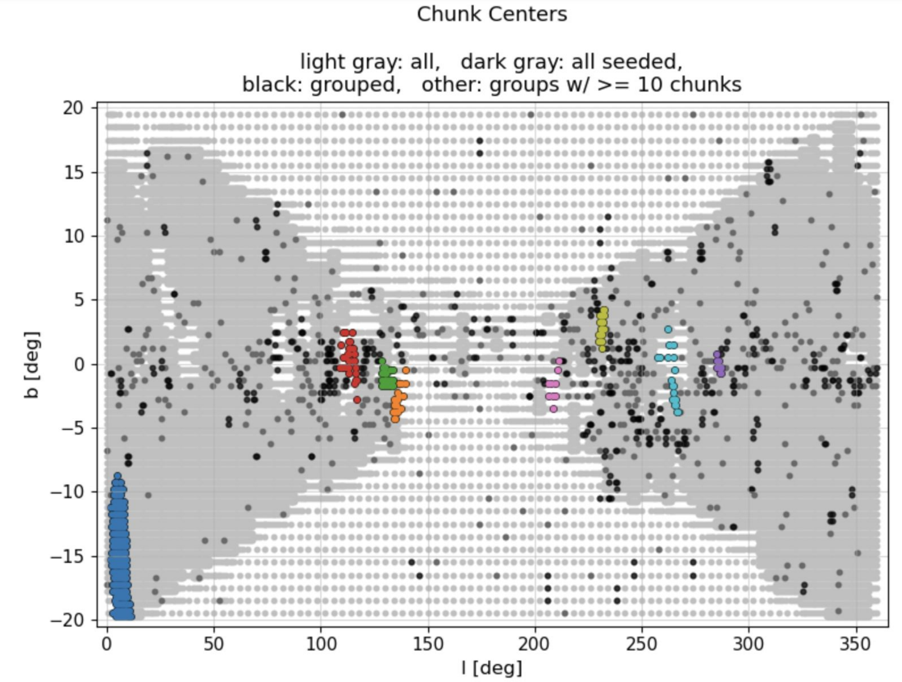




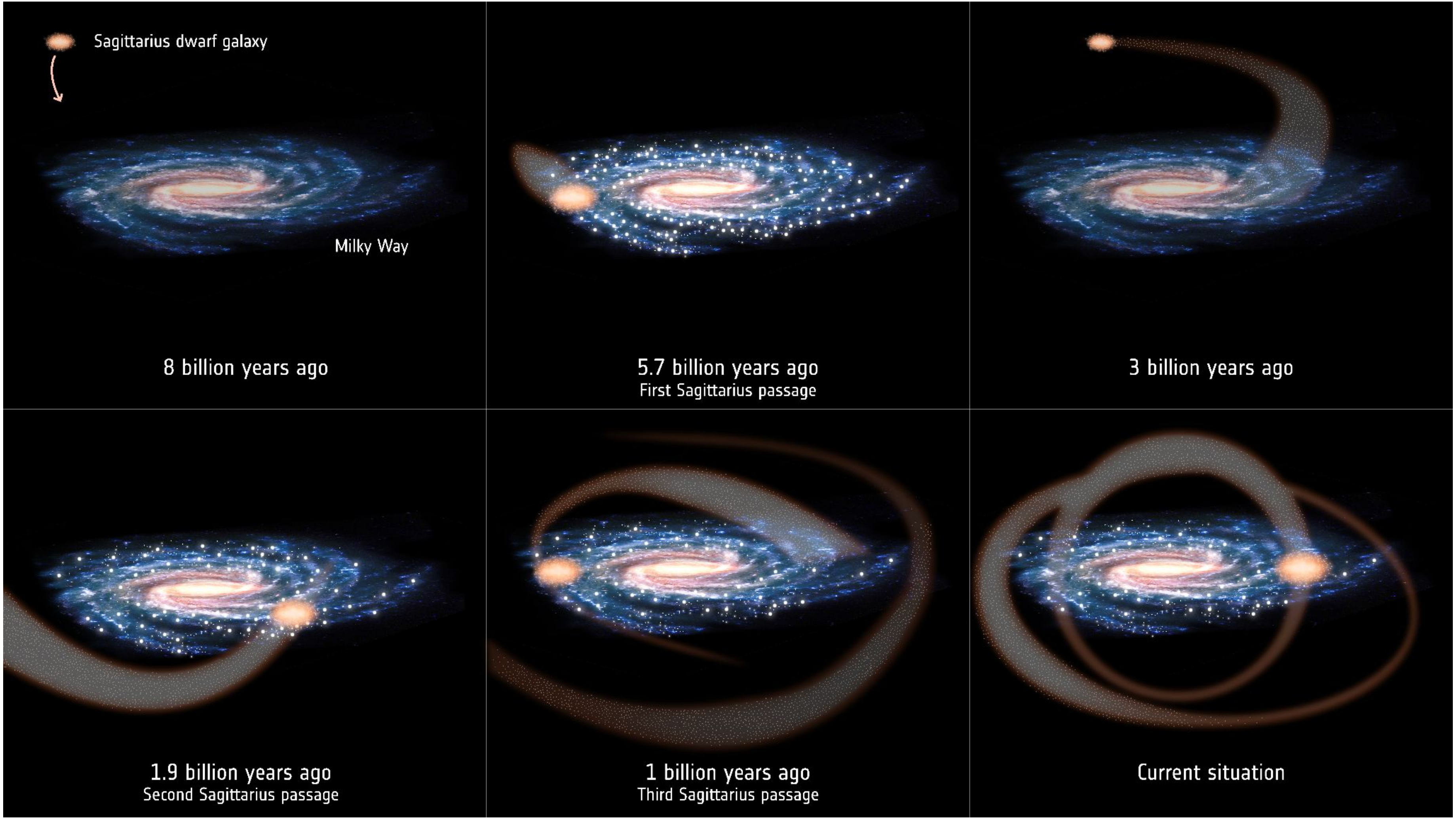


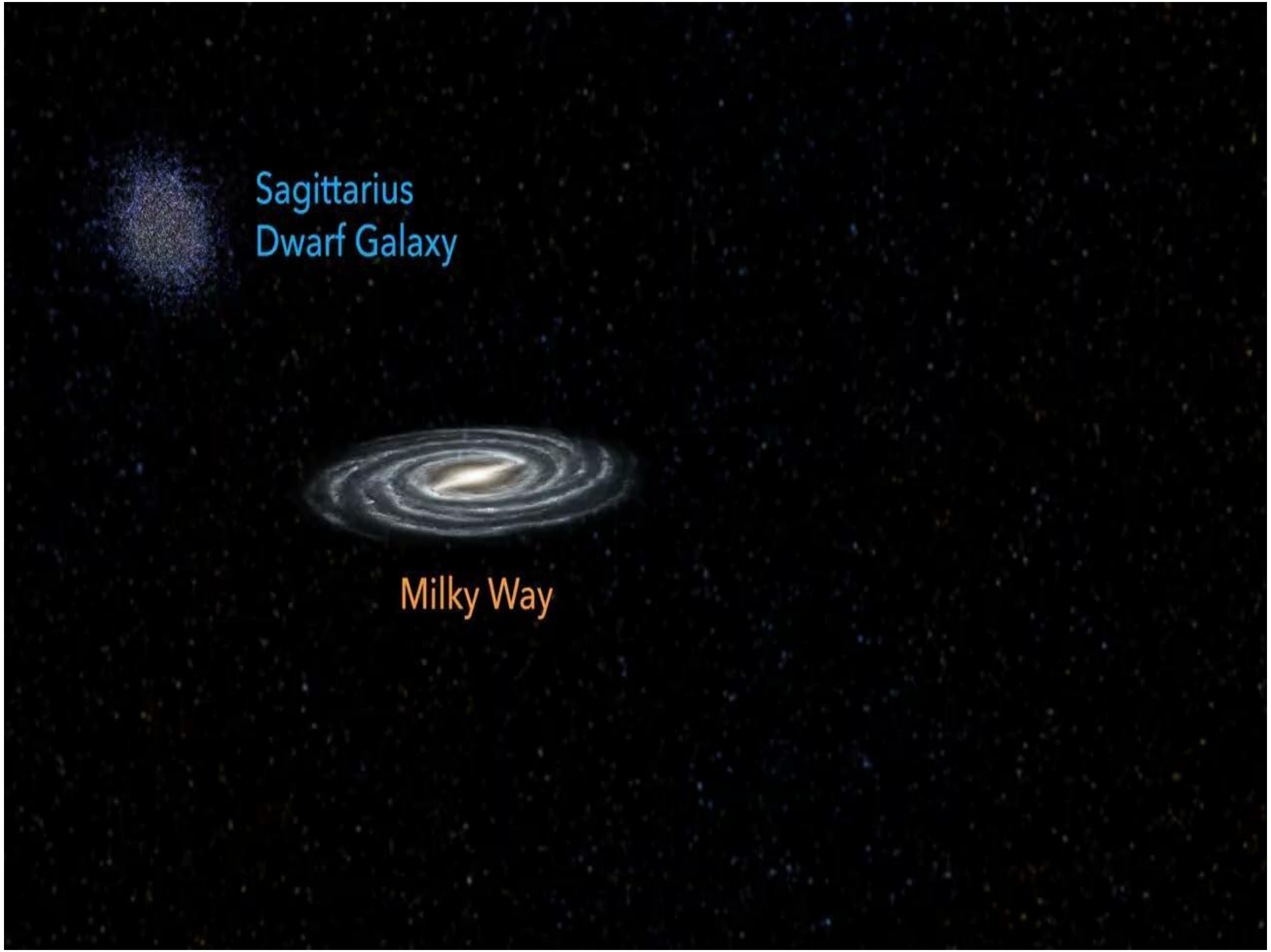
Summary & Future Directions

- We established an ML-based seeding method to find distinct stellar populations.
- We have identified examples of seeds belonging to known objects and continuing the process of matching seeds to known objects.
- We have begun to explore tracking methods for combining our seeds.

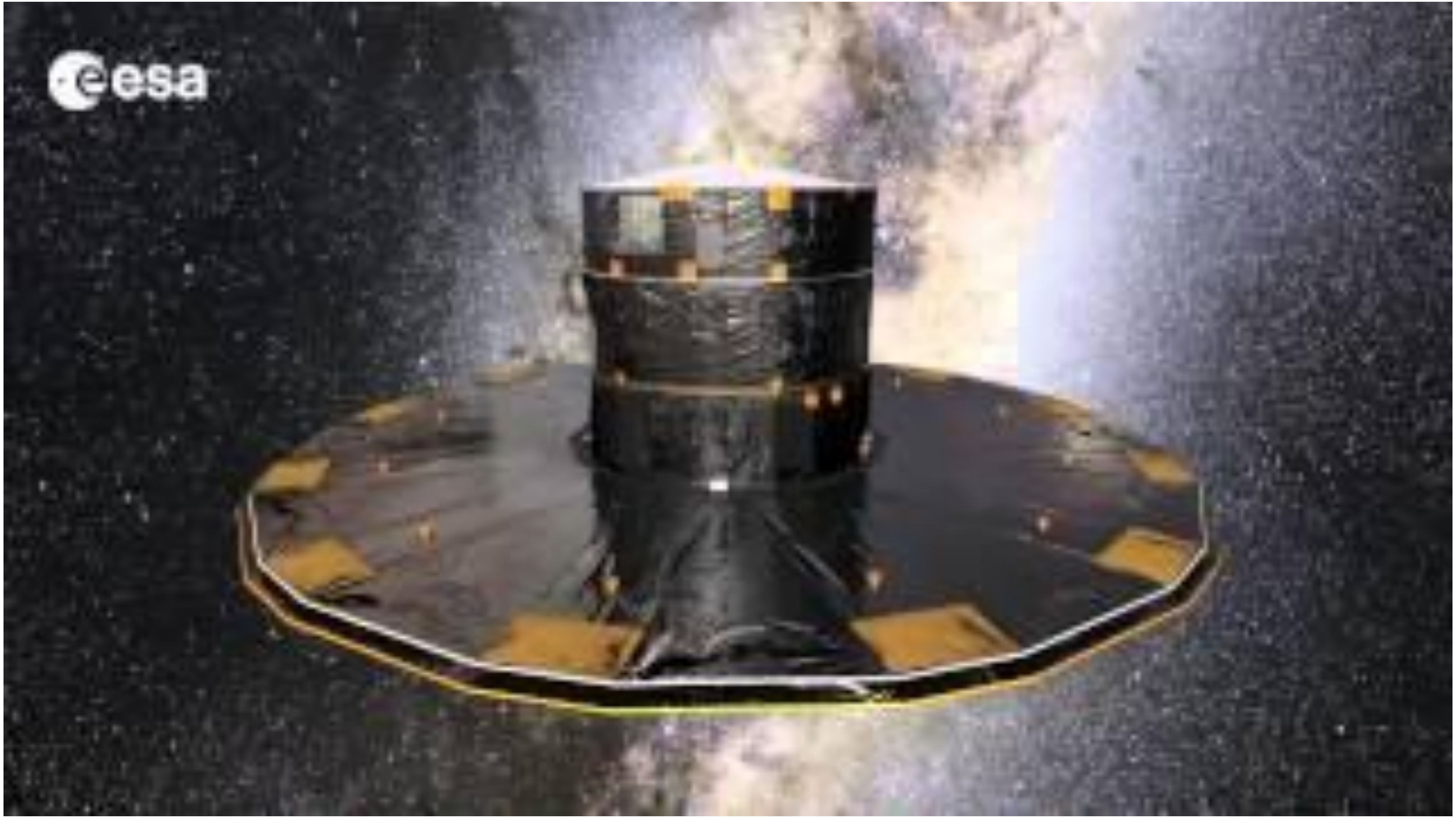


Backup Slides

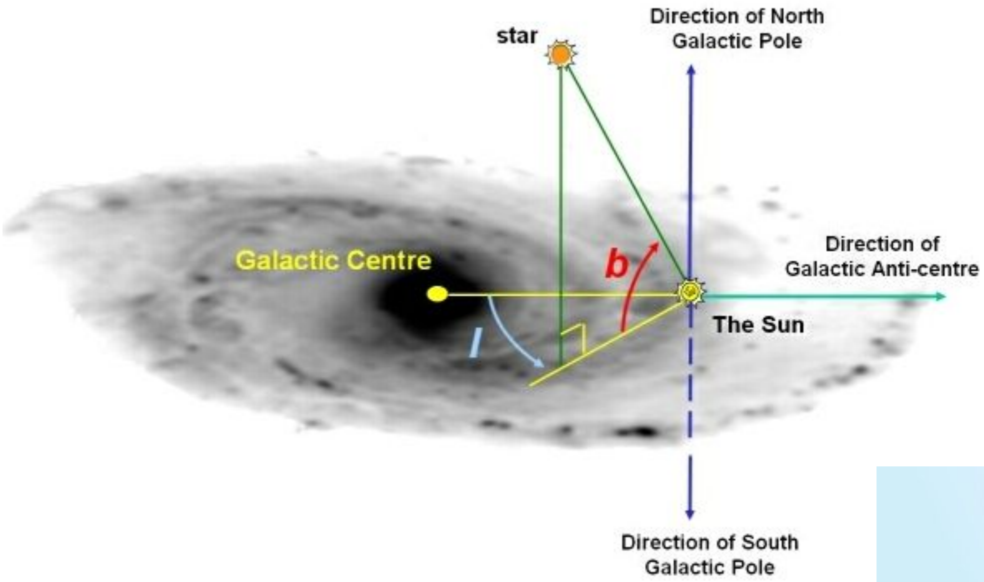

Sagittarius dwarf galaxy
Milky Way
8 billion years ago
5.7 billion years ago
First Sagittarius passage
3 billion years ago
1.9 billion years ago
Second Sagittarius passage
1 billion years ago
Third Sagittarius passage
Current situation

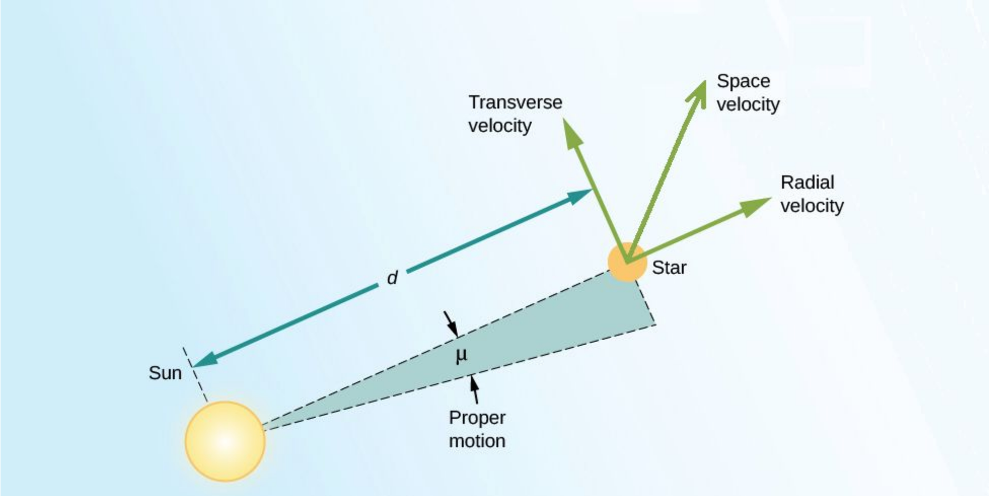


https://sci.esa.int/web/gaia/-/galactic-crash-may-have-triggered-solar-system-formation

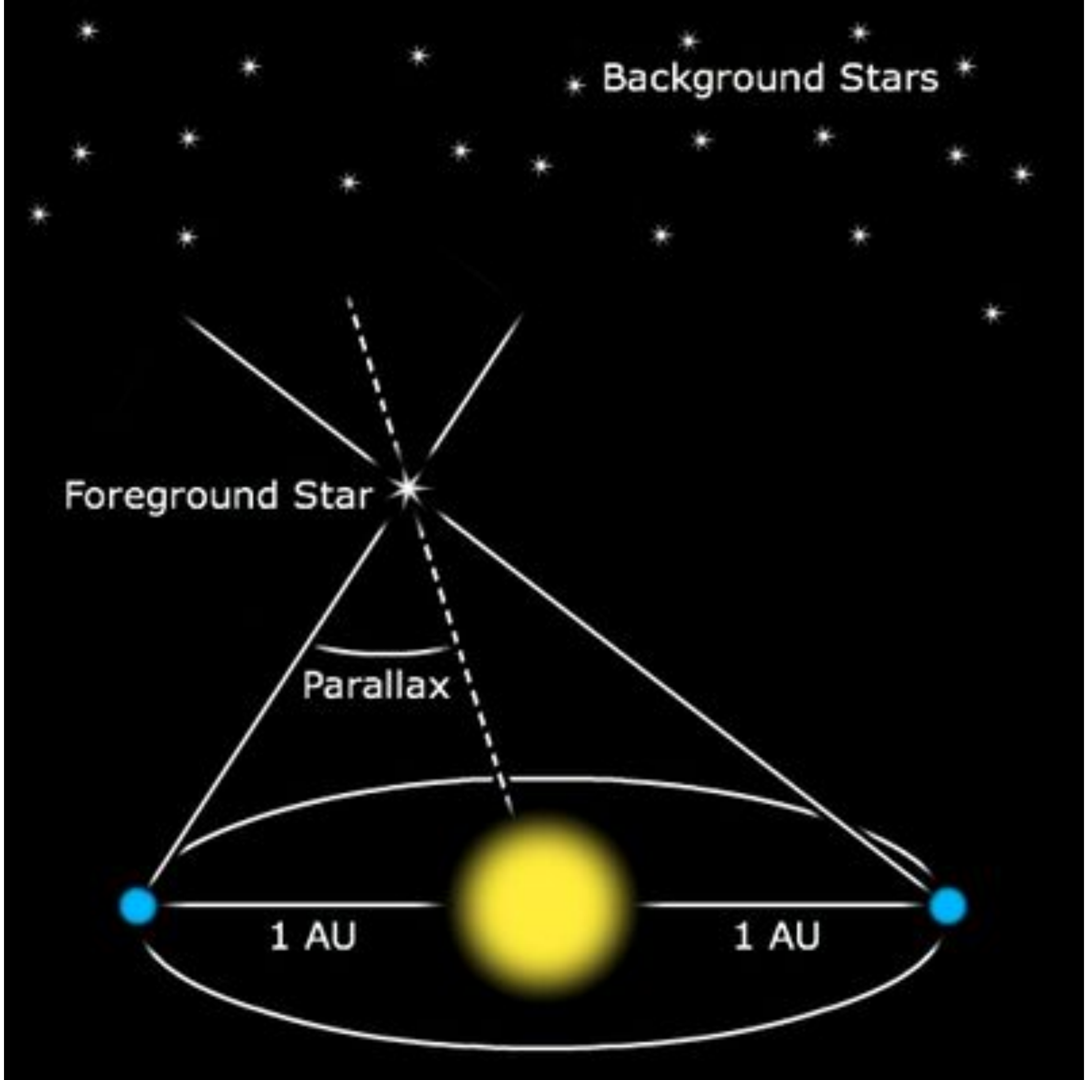


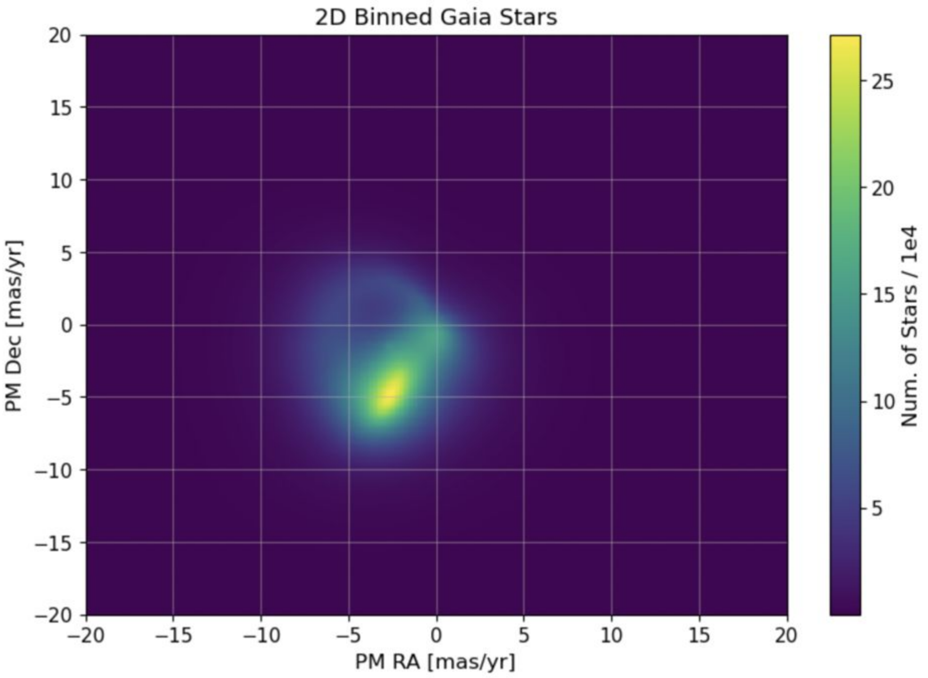
https://www.esa.int/ESA_Multimedia/Videos/2013/06/Gaia_scanning_the_sky





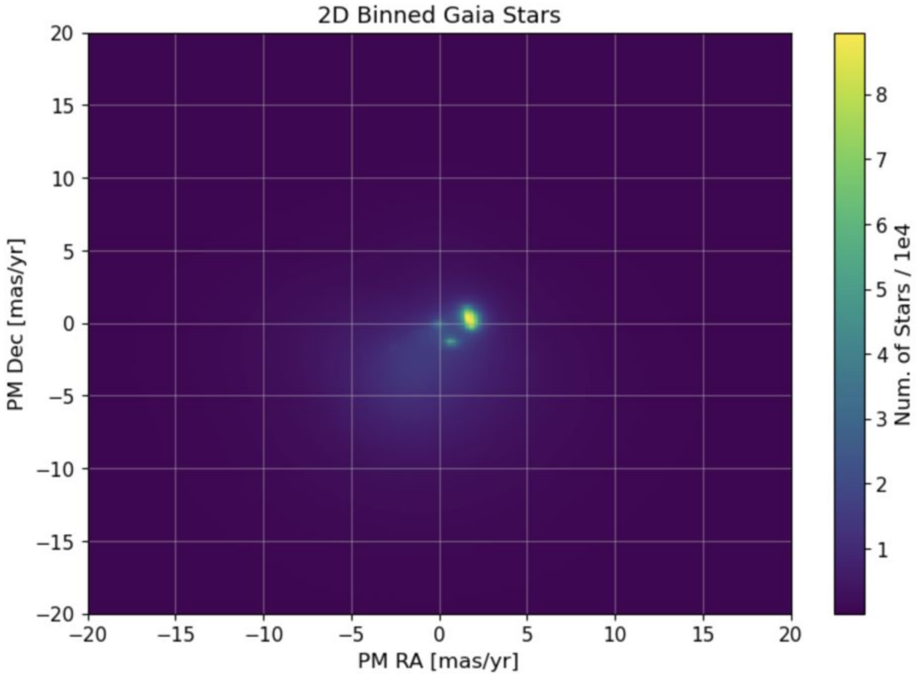
Proper Motion and Velocity of a Star

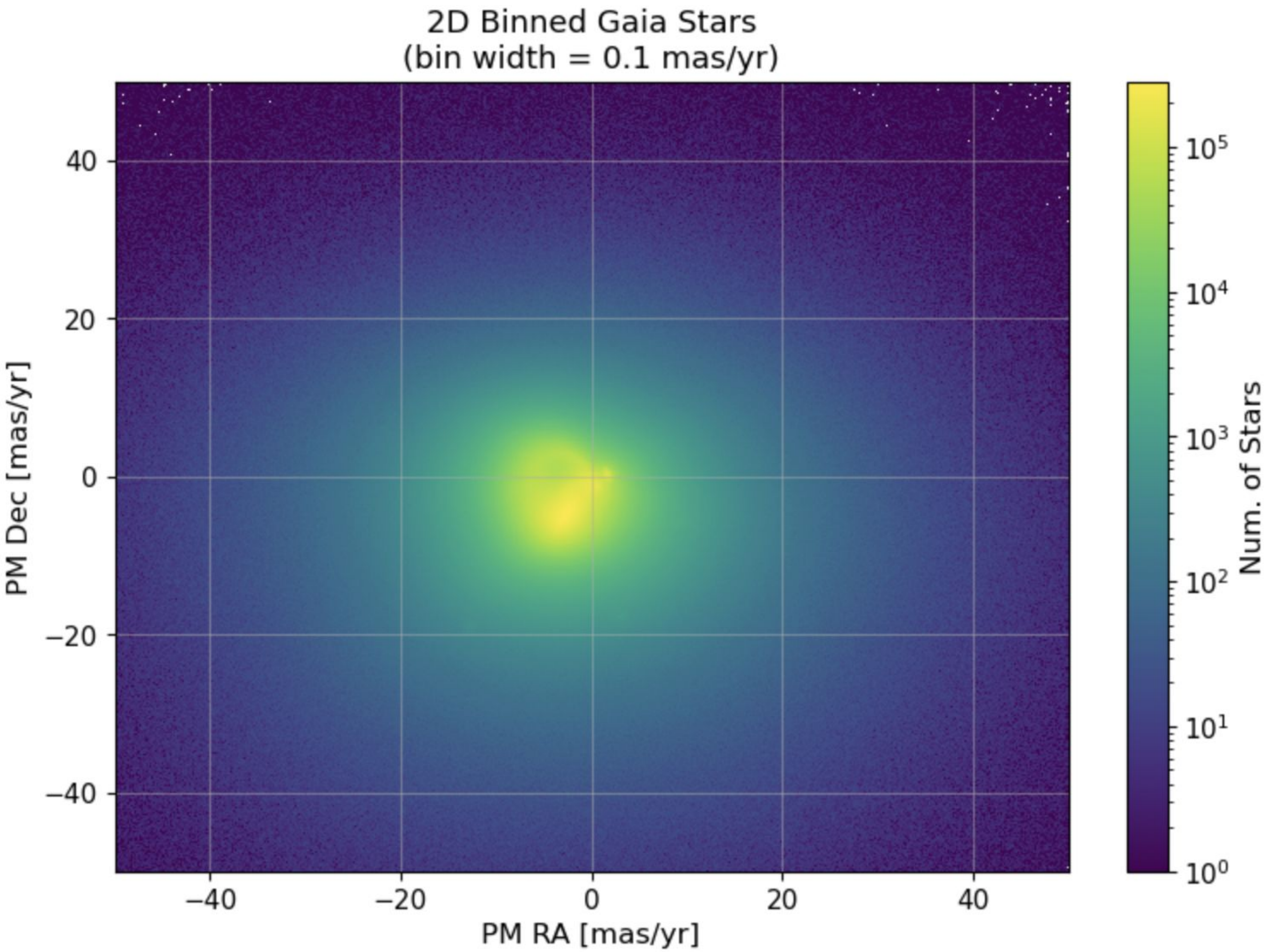

Background Stars
Foreground Star
Parallax
1 AU
1 AU



PM RA vs PM Dec
of Disk (|b| <= 20 deg) Stars

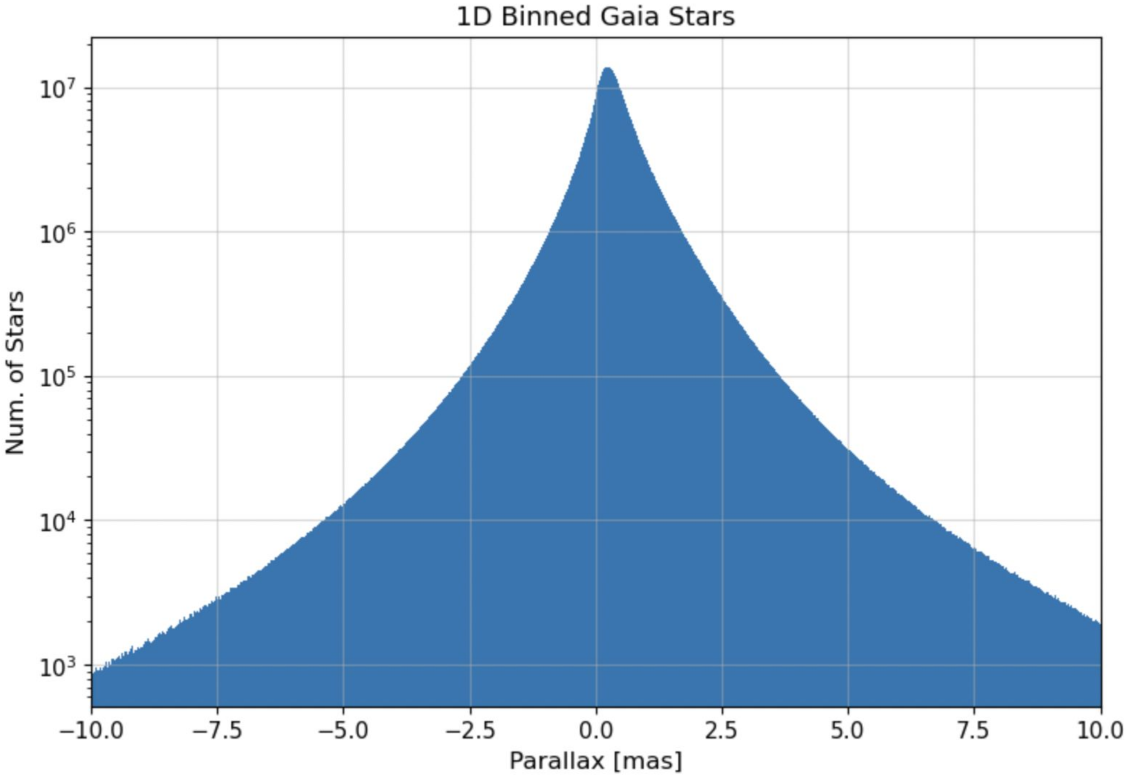
PM RA vs PM Dec
of Halo (|b| > 20 deg) Stars




2D Binned Gaia Stars
(bin width = 0.1 mas/yr)
PM Dec [mas/yr]
PM RA [mas/yr]
Num. of Stars
40
20
0
−20
−40
−40
−20
0
20
40
10^0
10^1
10^2
10^3
10^4
10^5

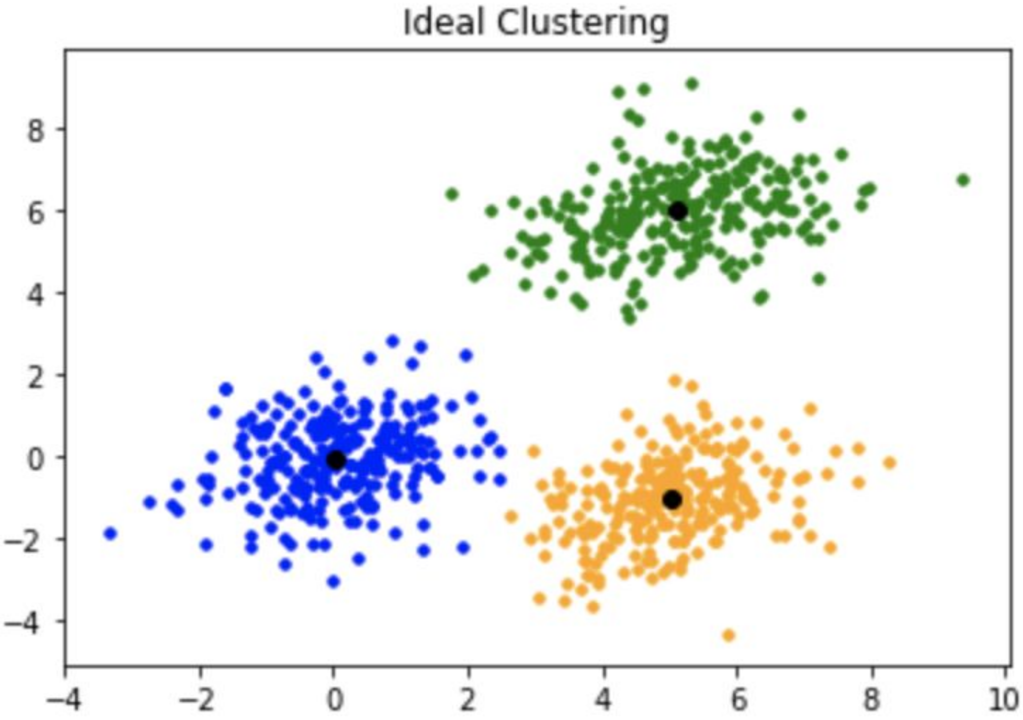
Gaia Stars in Parallax Space

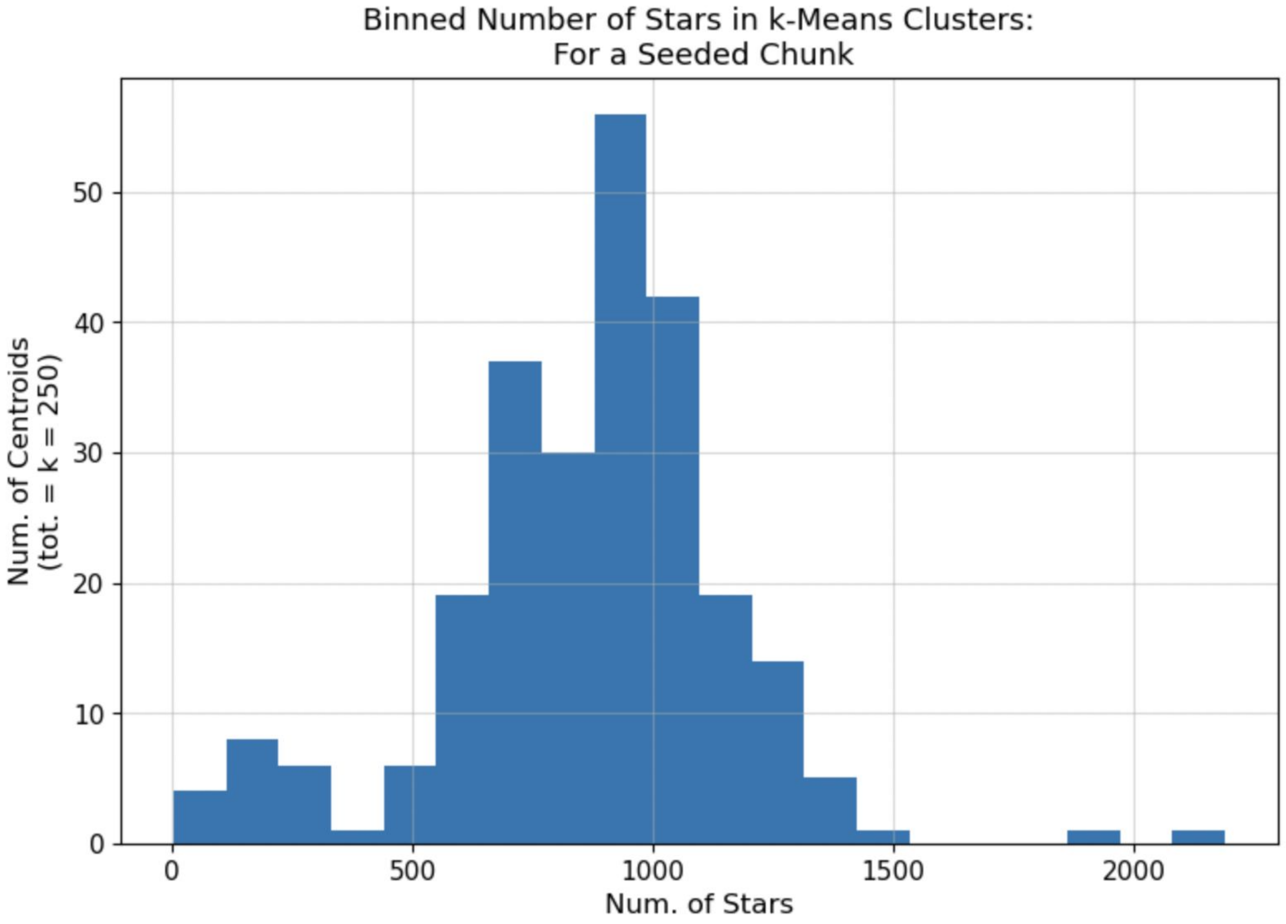
(Parallax variable is not very reliable.)

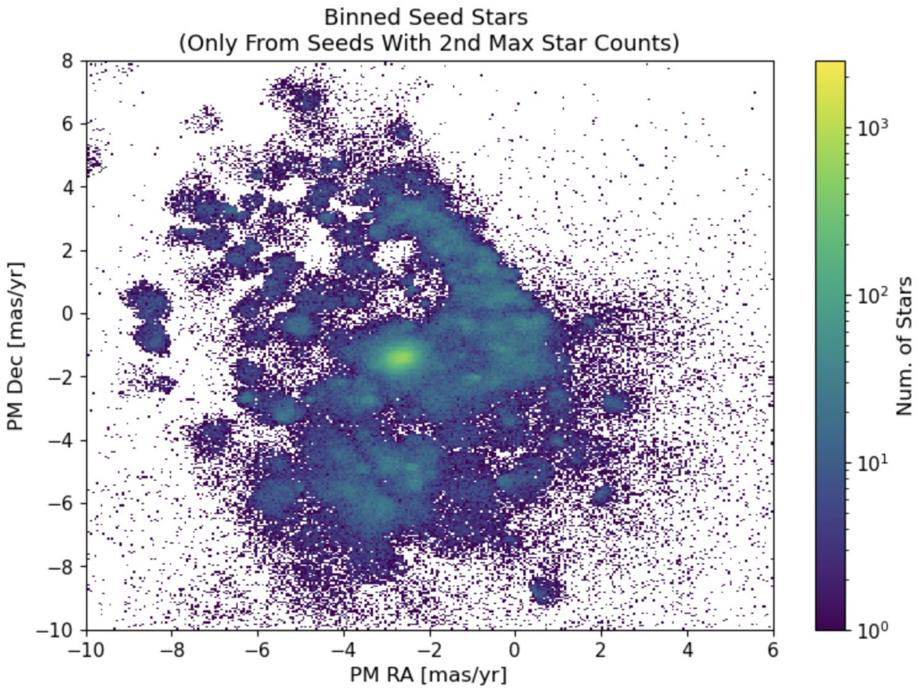


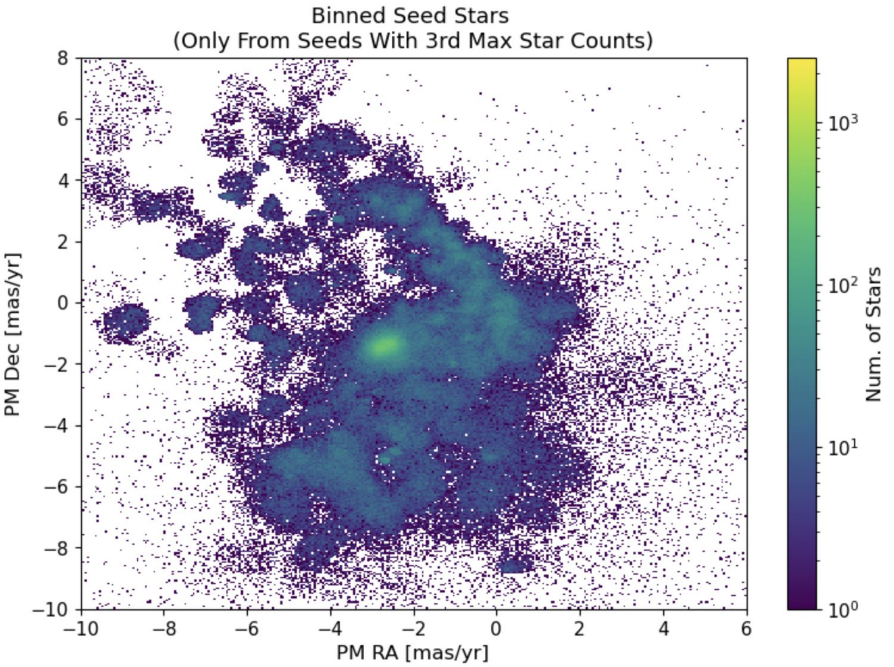
K-means and KNN Example

k=3 clusters in 2 dimensions

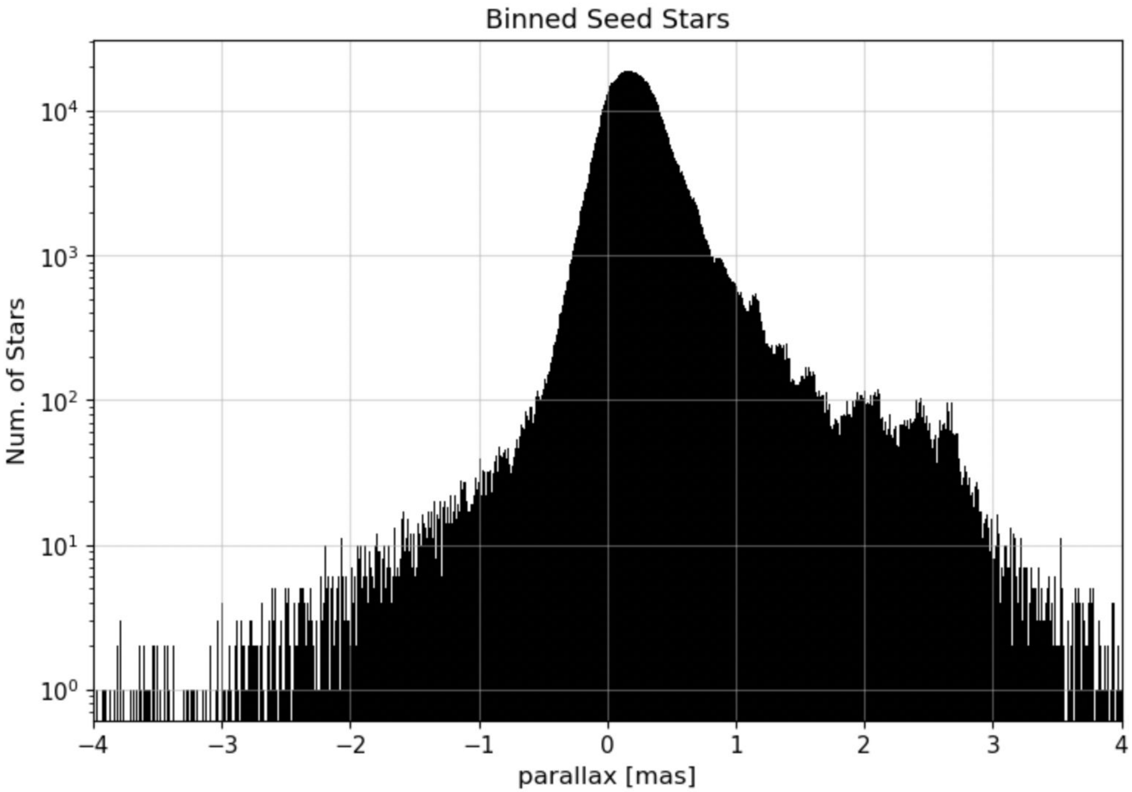



Binned Number of Stars in k-Means Clusters:
For a Seeded Chunk
Num. of Centroids
(tot. = k = 250)
Num. of Stars
0
10
20
30
40
50
0
500
1000
1500
2000

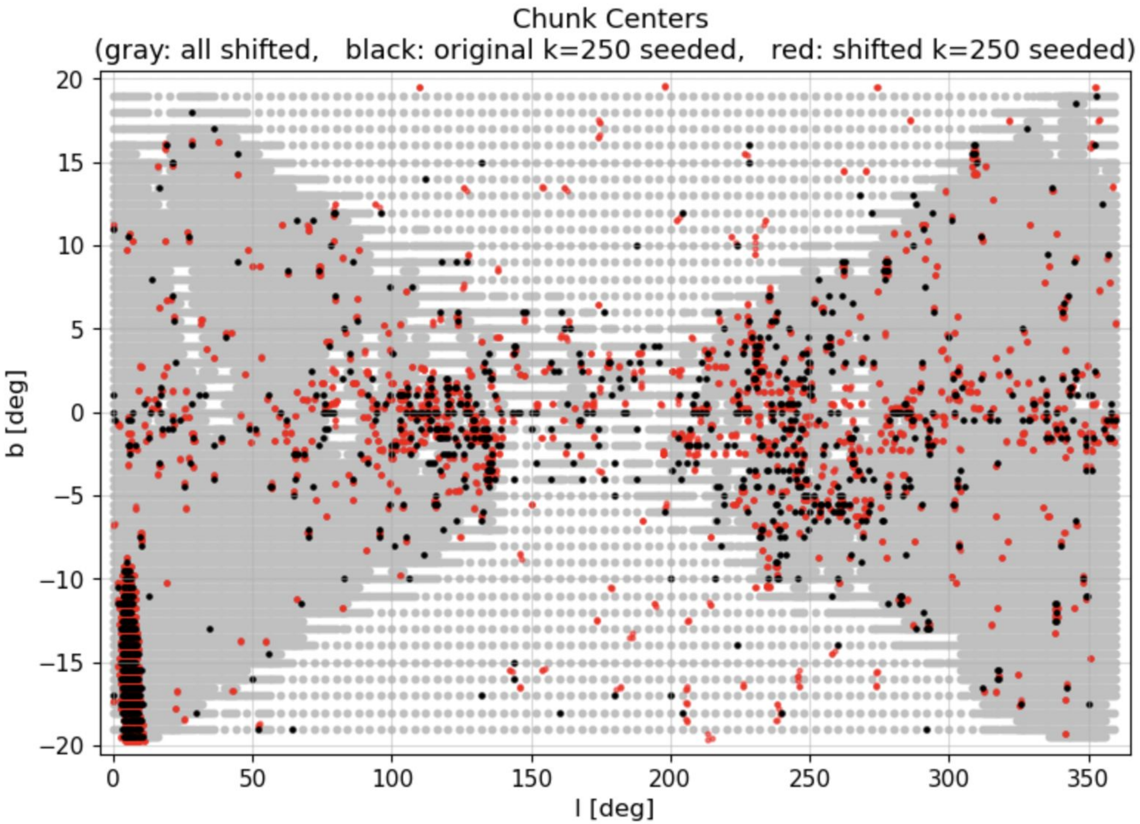

Binned Seed Stars
(Only From Seeds With 2nd Max Star Counts)
PM Dec [mas/yr]
PM RA [mas/yr]
Num. of Stars


Binned Seed Stars
(Only From Seeds With 3rd Max Star Counts)
PM Dec [mas/yr]
PM RA [mas/yr]
Num. of Stars

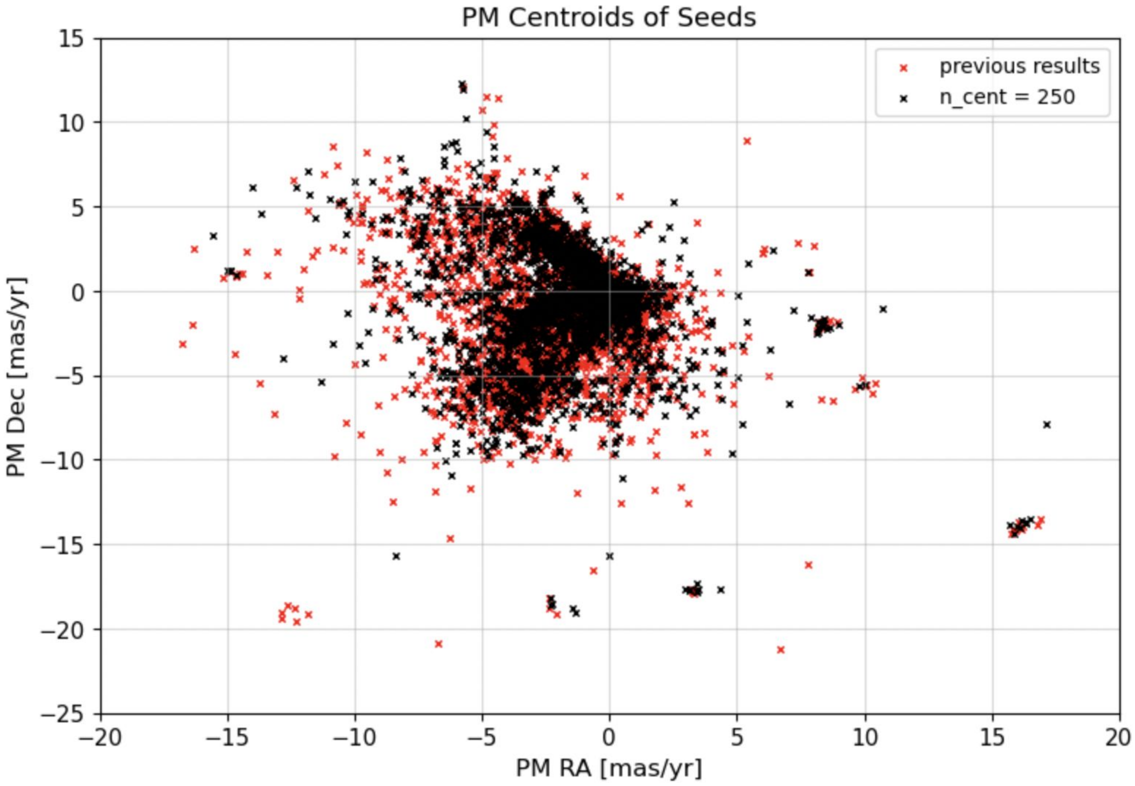
Seed Results in Parallax Space



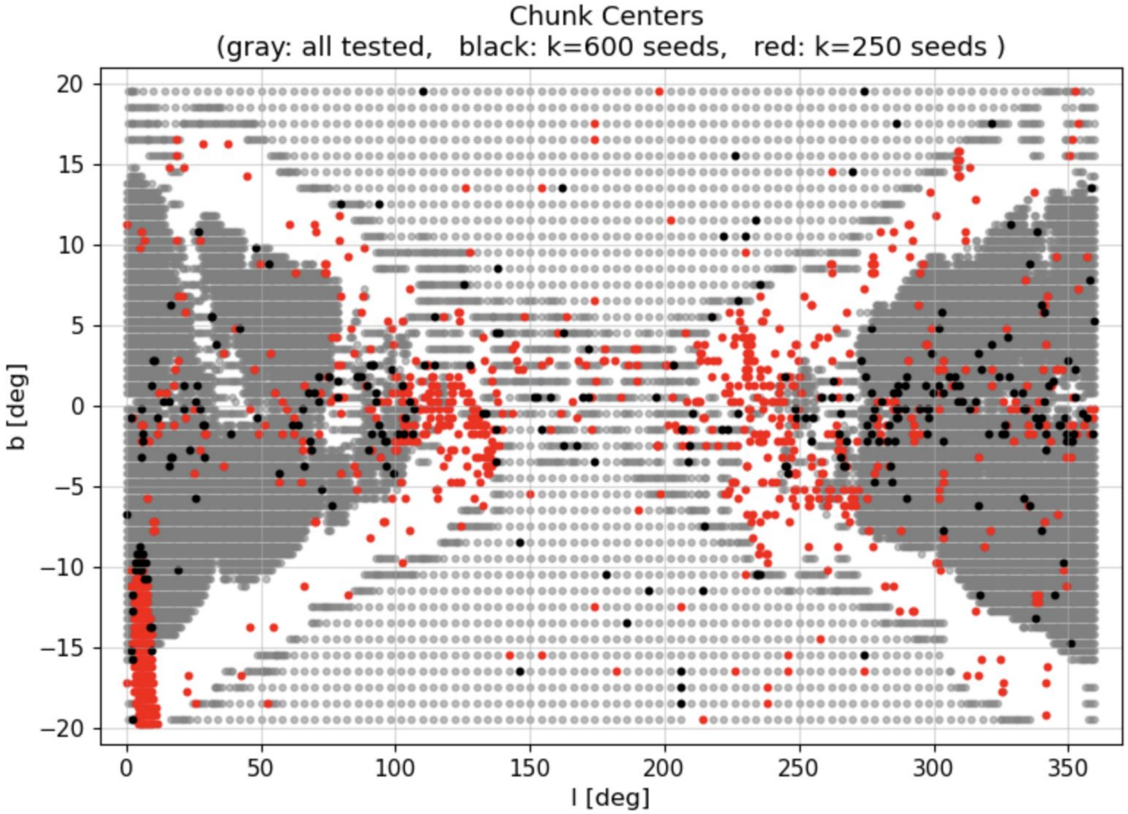
Window-Slided Chunk Centers



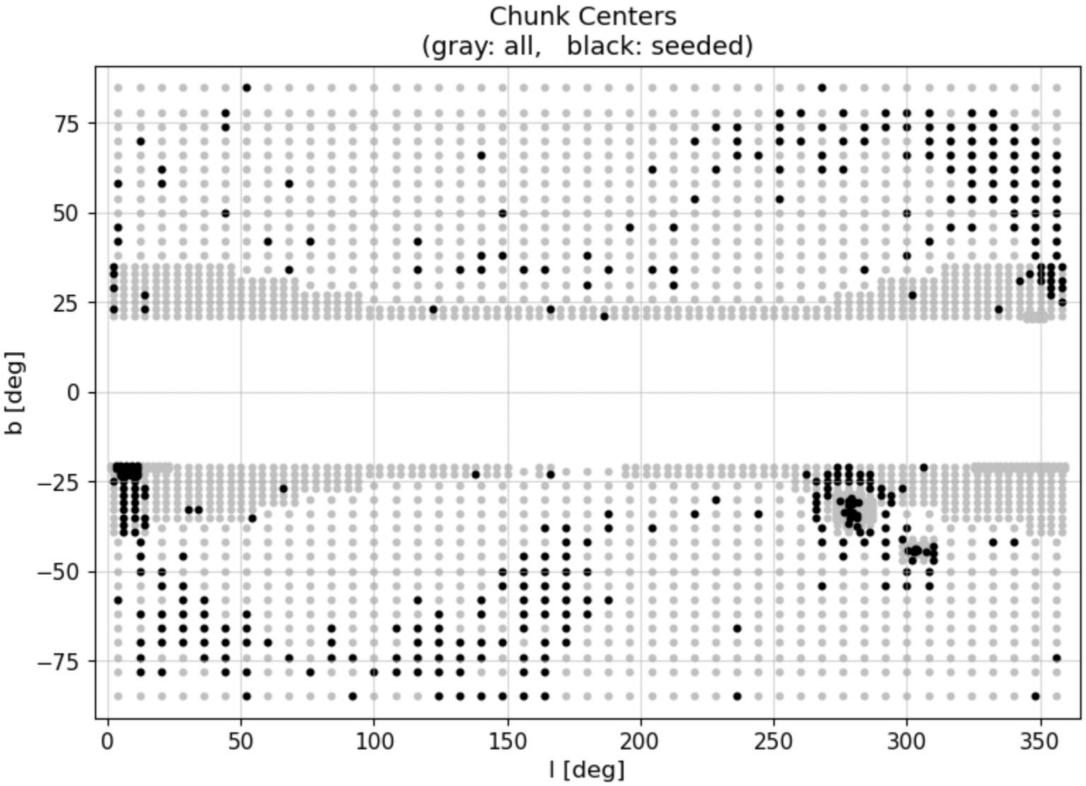
Window-Slided Chunk Seeds



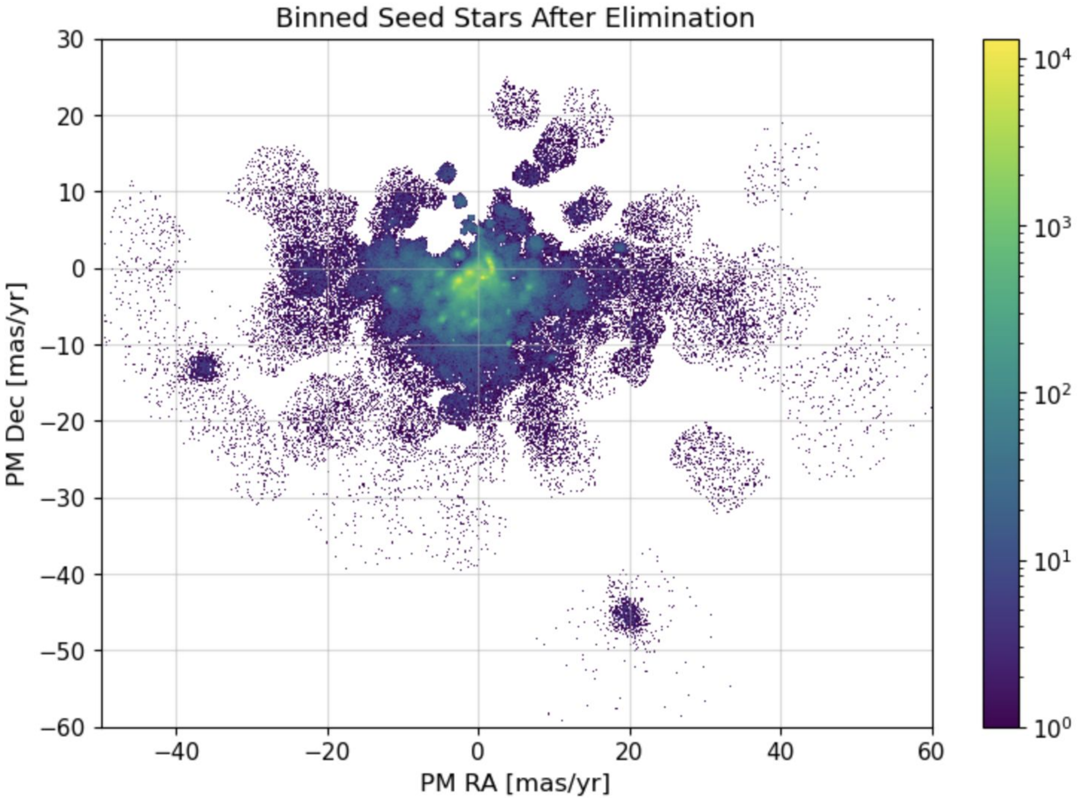
Varying *k* for K-Means



Galactic Halo (|b| > 20 deg Region) Seeds



Galactic Halo (|b| > 20 deg Region) Seeds



Galactic Halo (|b| > 20 deg Region) Seeds