Physics Without Frontiers: Chile

School on machine learning in physics

13-17 JANUARY 2025 | VALPARAÍSO, CHILE

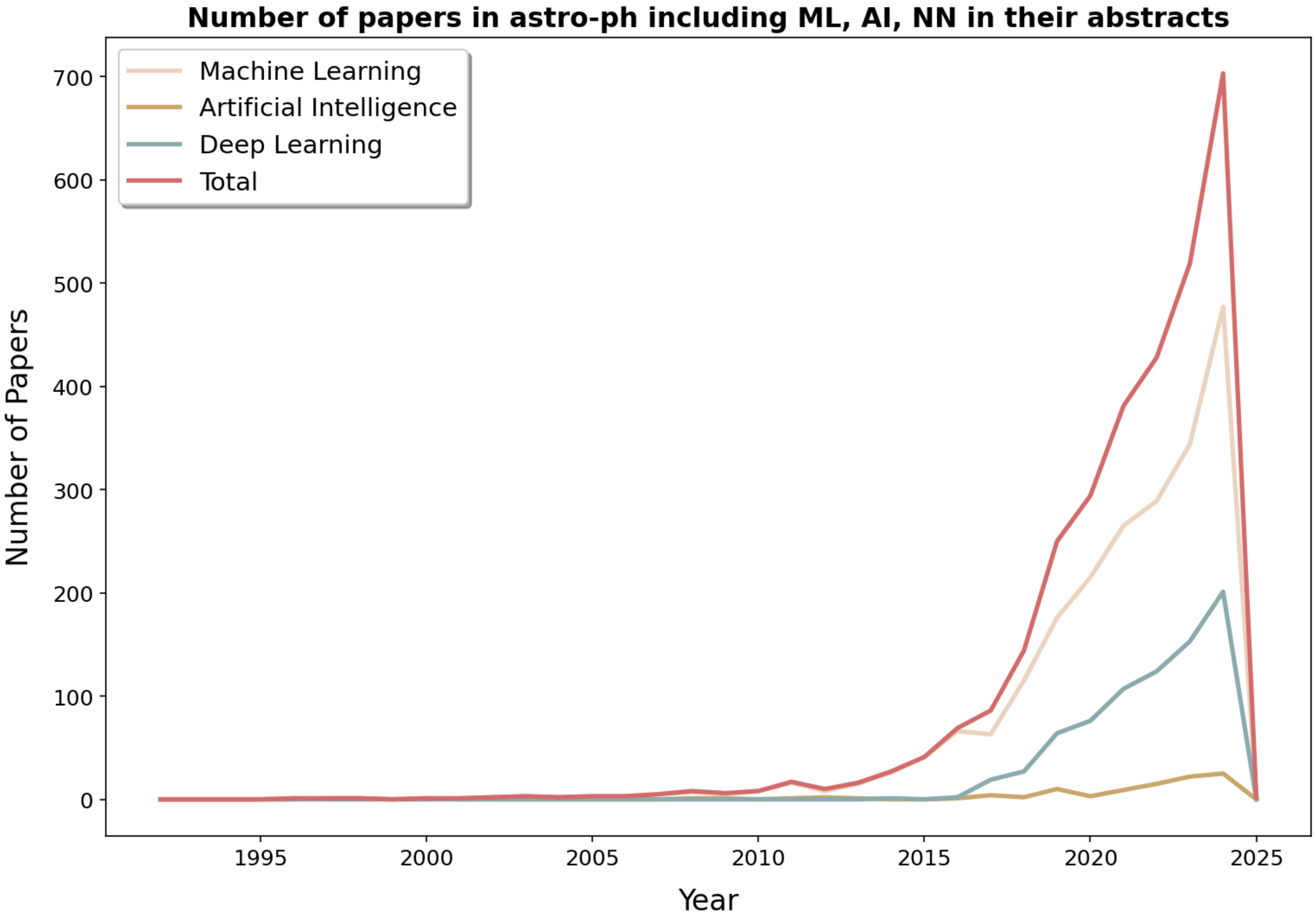
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

CCTVal

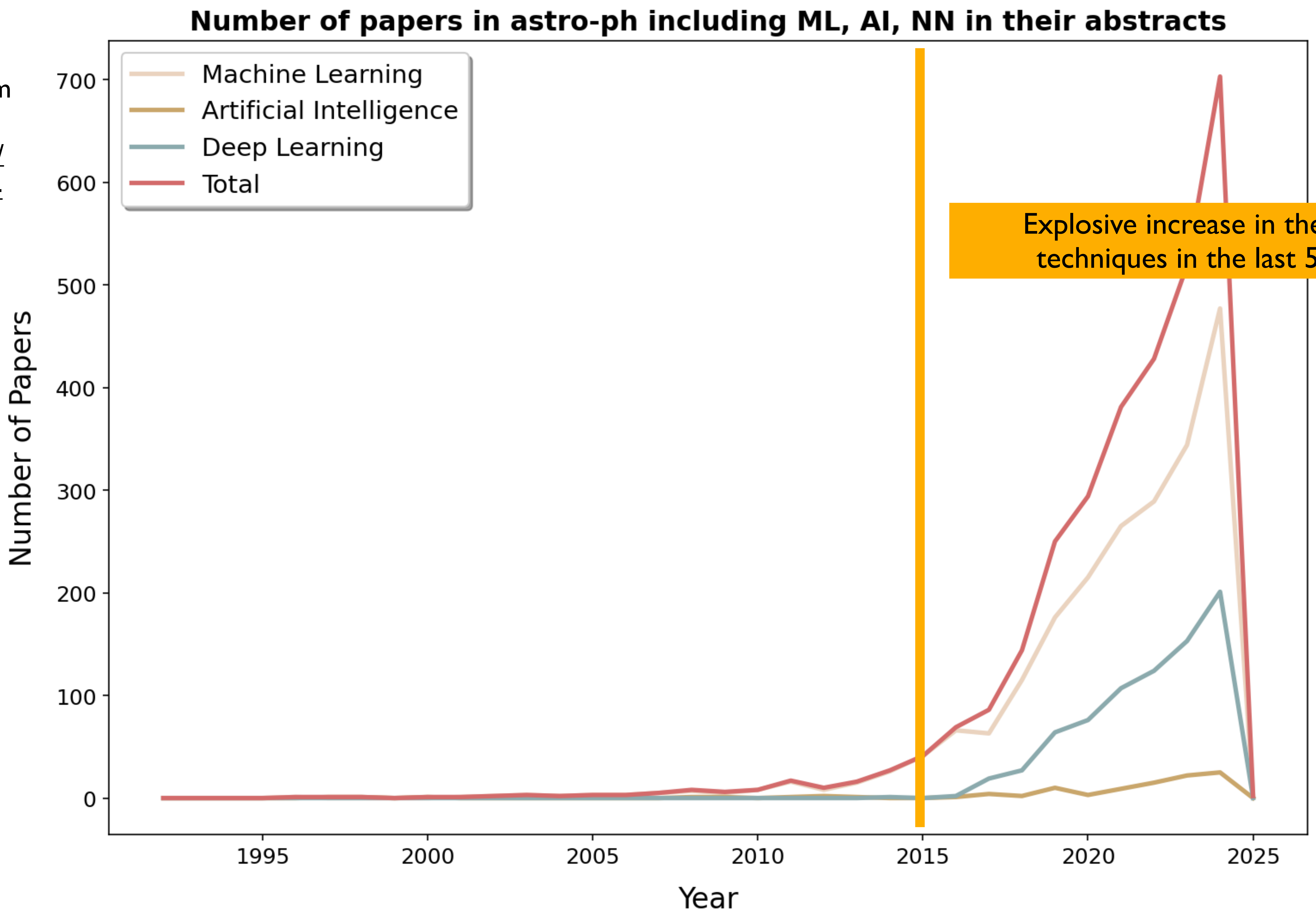
AI/ML Applications in Astrophysics

Pía Amigo
Departamento de Física, USM
pia.amigo@usm.cl

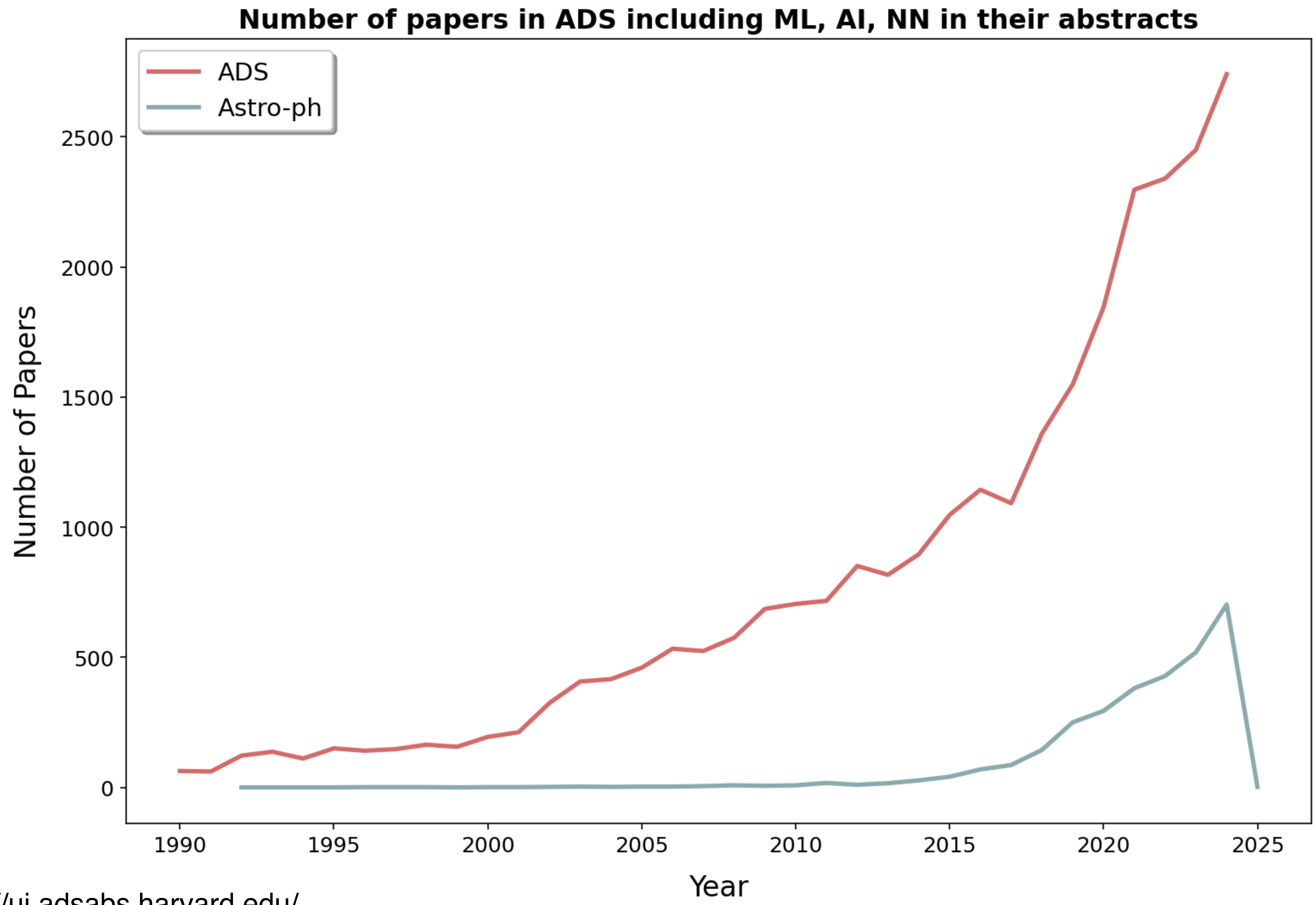
arXiv dataset from Kaggle, https://www.kaggle.com/datasets/Cornell-University/arxiv/code



arXiv dataset from Kaggle, https://www.kaggle.com/datasets/Cornell-University/arxiv/code



Explosive increase in the use of ML techniques in the last 5-10 years!

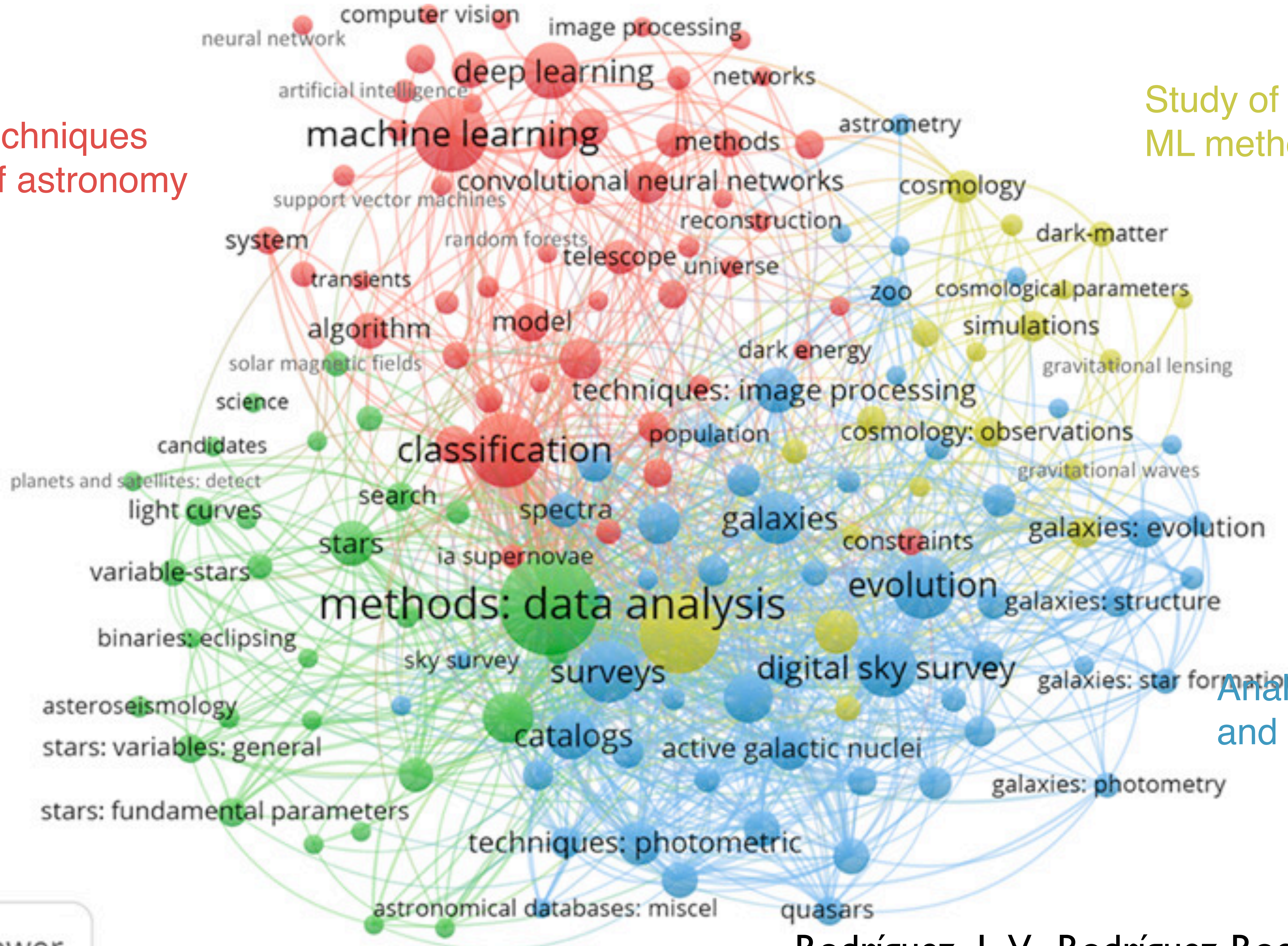


NASA ADS. https://ui.adsabs.harvard.edu/

~2700 papers

Includes AI and ML techniques applied to the fields of astronomy and astrophysics

Study of cosmology using AI and ML methods



VOSviewer

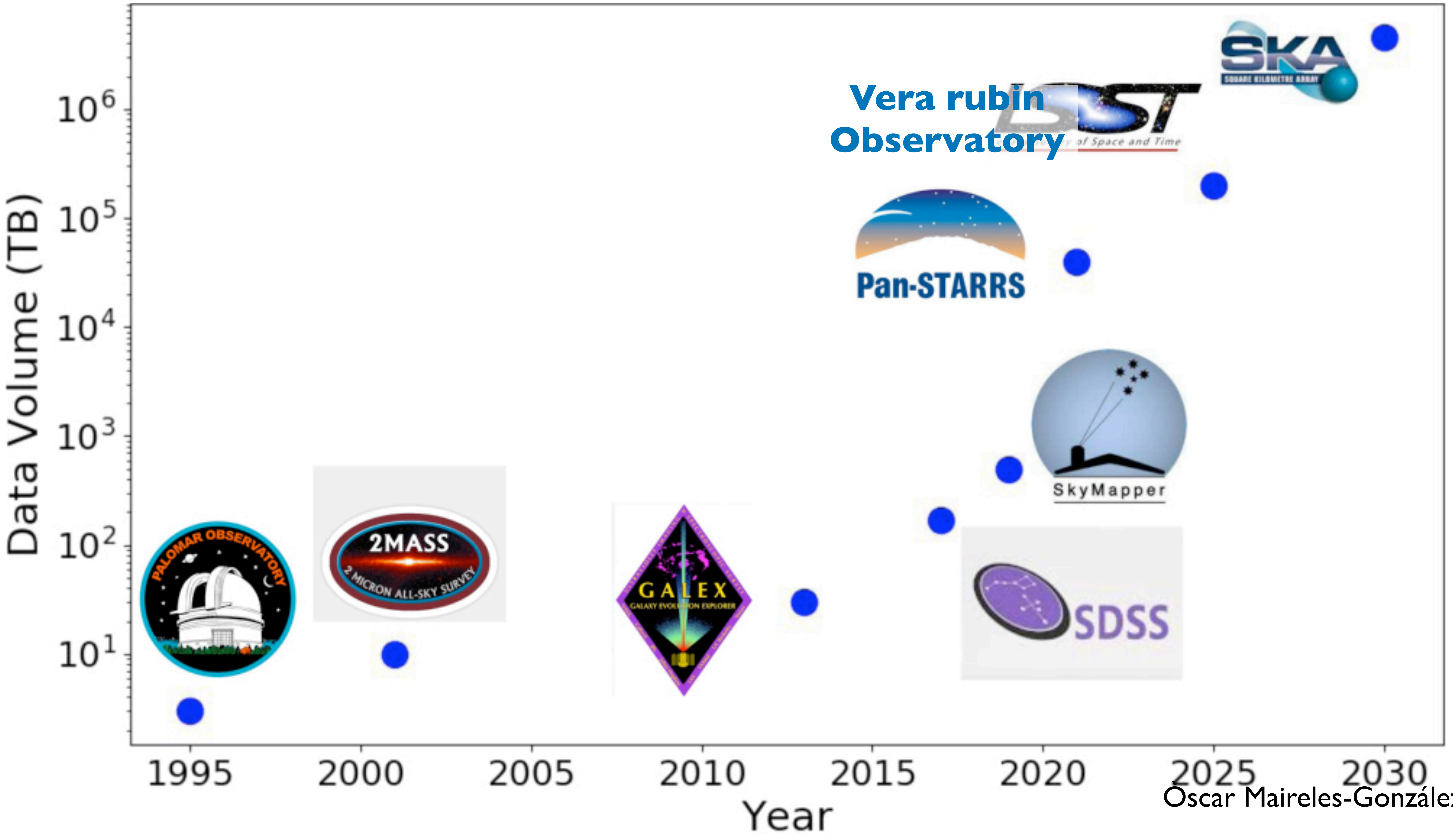
Application of AI and ML in the field of stellar analysis.

Analysis of galaxies by means of AI and ML techniques

Rodríguez, J.-V., Rodríguez-Rodríguez, I., & Woo, W. L. (2022). On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric analysis.

Why ML/AI in Astrophysics?

Evolution of Astronomical Surveys Data Volumes

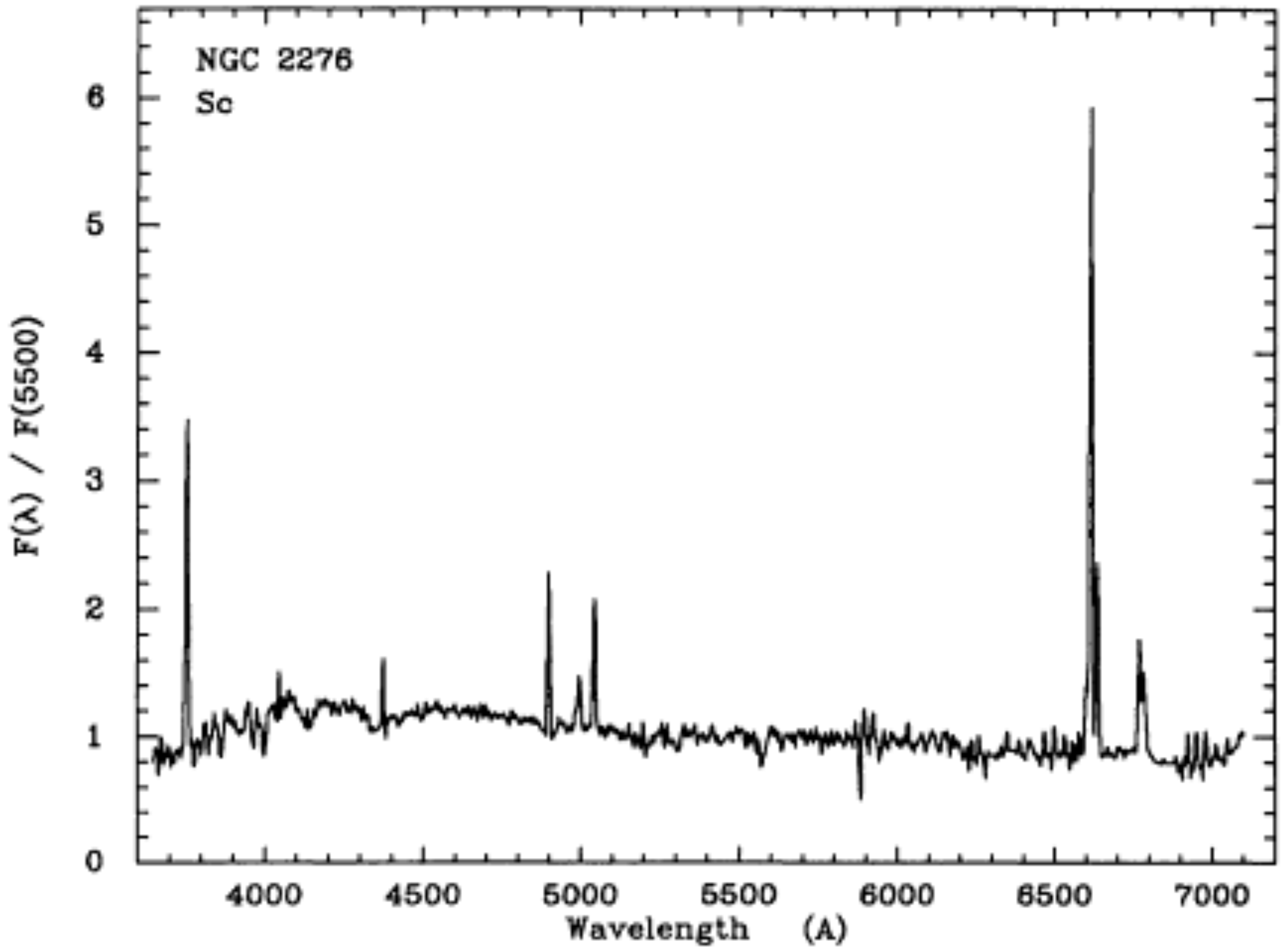


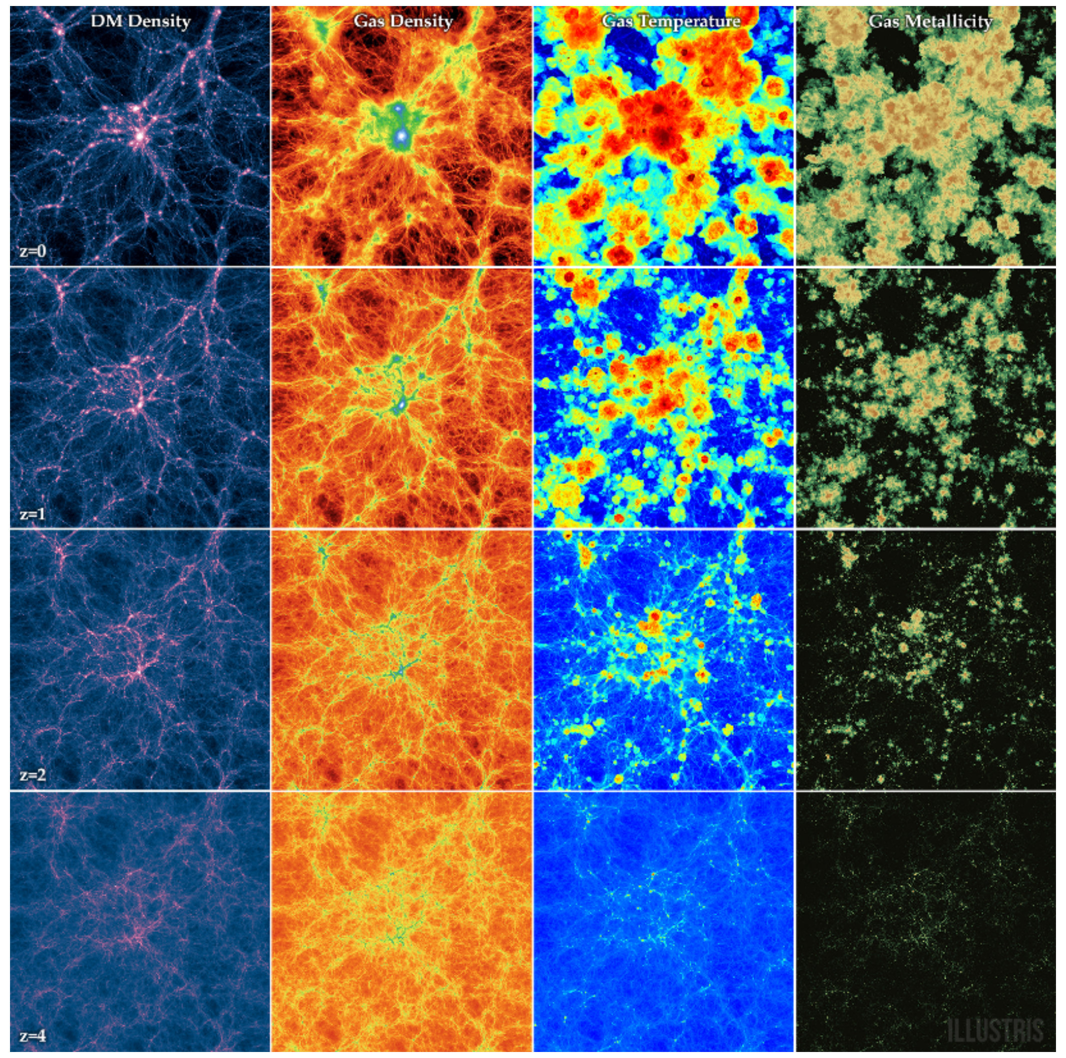
Óscar Maireles-González et al 2023

Why ML/AI in Astrophysics?

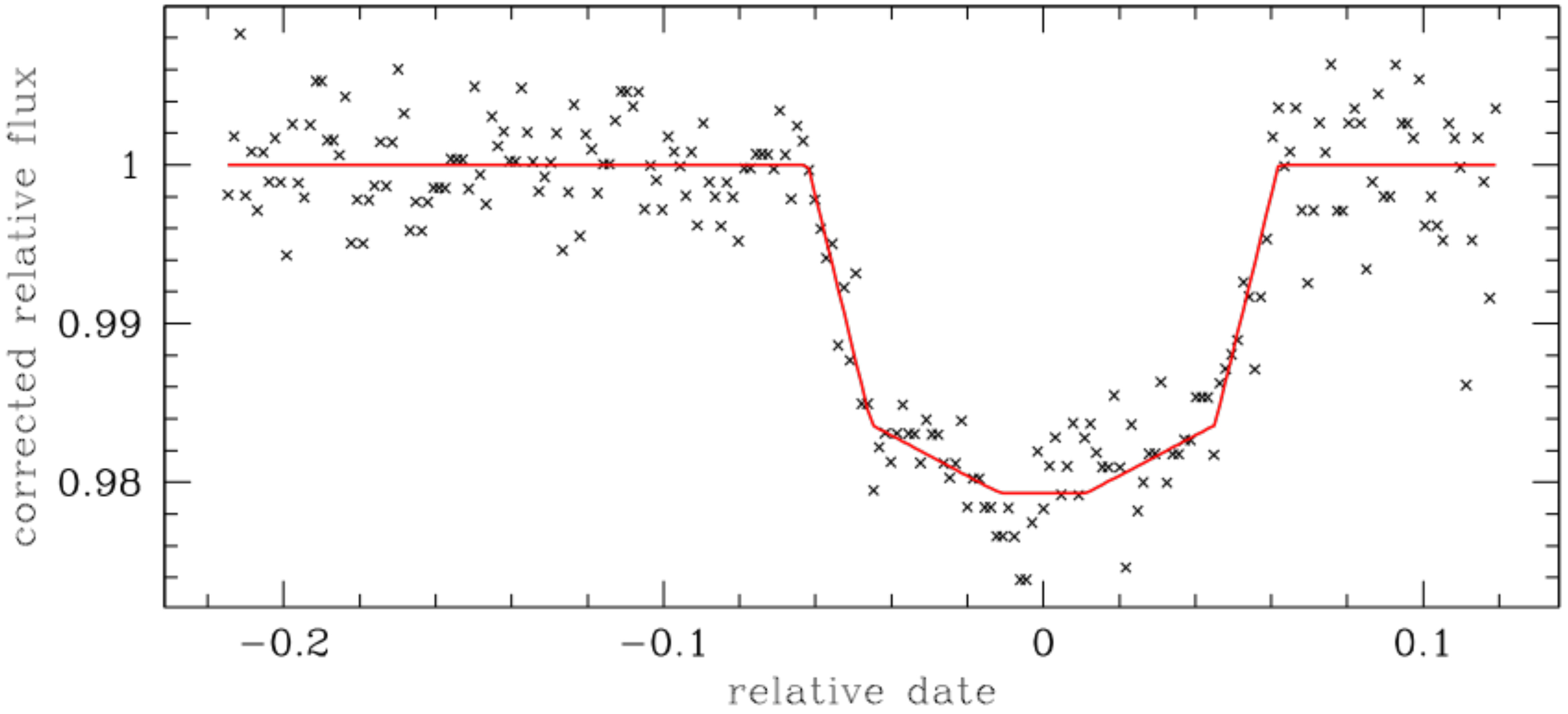
- Bigger and more complex datasets available (open source). ML is almost a necessity. And the future will become more demanding (Rubin, SKA).
- Techniques become more popular and better known.
- Availability of better computing infrastructures (GPUs, cloud services) and more funding for AI-based projects.
- No ethical issues like privacy concerns or biases that may affect other disciplines.
- Success stories involving citizen science projects.

Astronomical Data

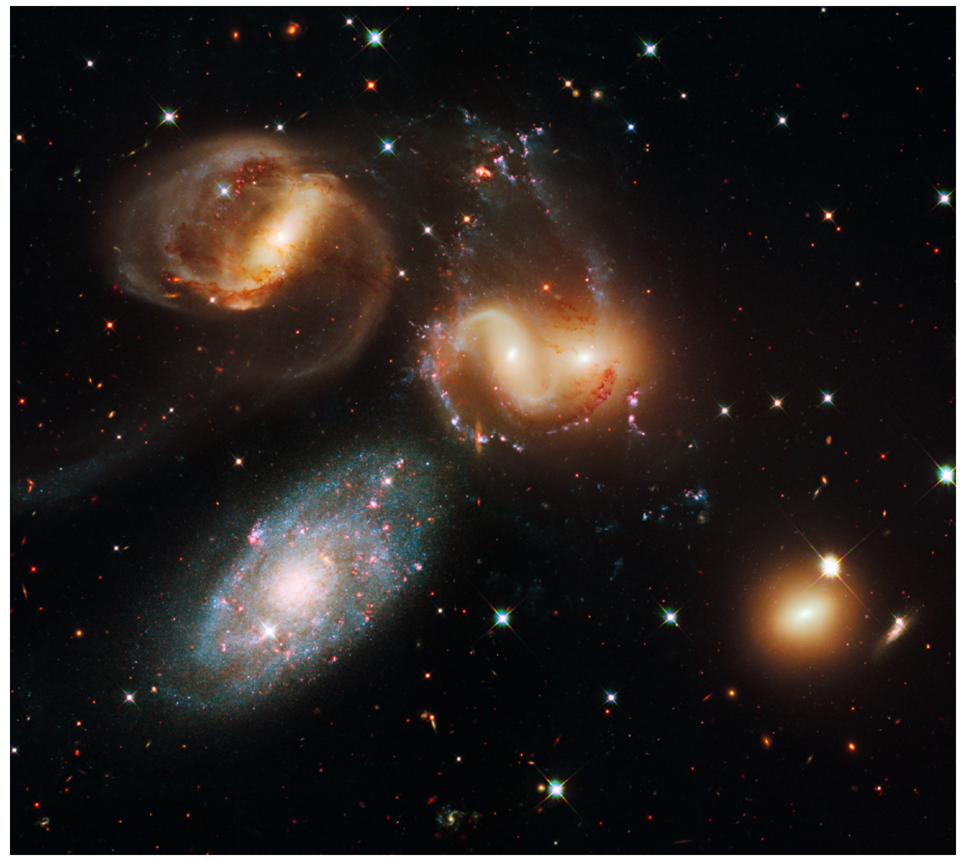


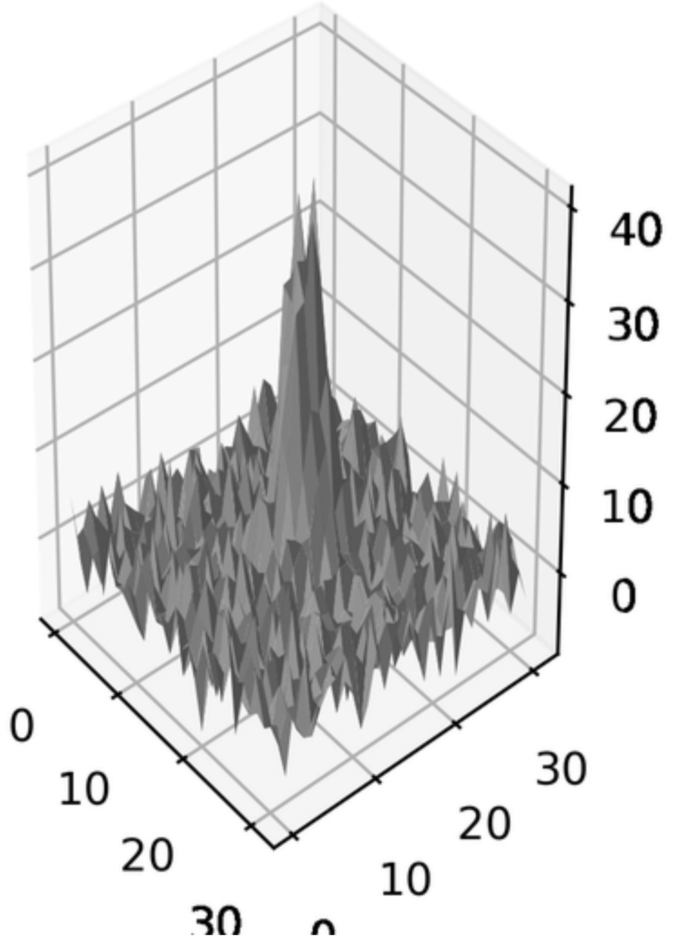


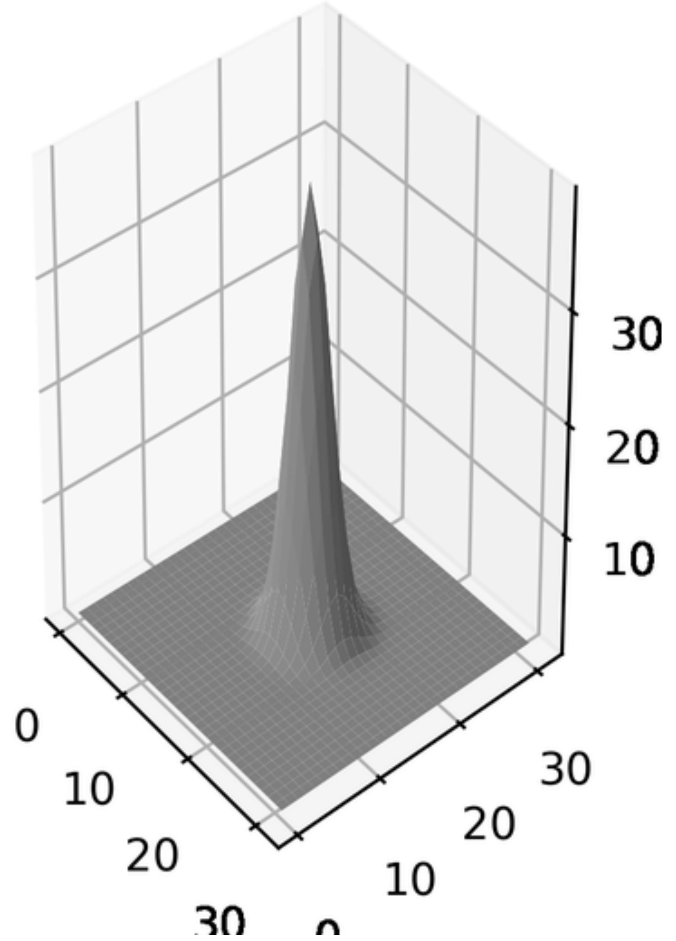
Name	RA(1950)	Dec(1950)	pm	angle	v_{rad}	Sp Type	m_V	B−V	U−B	R−I	π_{trig}	M_V
Sun						G2 V	-26.72	0.65	0.10			4.85
NN	00 00 06	-34 29.7	0.758	168.6		DC9	14.90	0.46	-0.44		75.2	14.28
GJ 1001	00 02 05	-40 57.8	1.618	154.5	-3	M3.5	12.84	1.63	1.30	1.23	104.2	12.93
NN	00 02 16	+34 22.8	0.776	83.0	+6.4 VAR	G2 V	6.11	0.62	0.09		29.8	4.56
NN	00 02 21	+22 59.5	0.380	91.5		G9 V	7.82	0.74	0.29	0.33		5.78
Gl 1	00 02 28	-37 36.2	6.097	112.5	22.9	M4 V	8.54	1.46	0.96	0.92	221.8	10.27
Gl 2	00 02 32	+45 30.6	0.894	100.5	0.1	dM2 e	9.93	1.49	1.18	0.85	87.0	9.63
NN	00 02 43	+48 12.0	0.009	305.5		G5	8.30					6.84
Gl 3	00 02 48	-68 06.2	0.582	190.7	41	K5 V	8.48	1.06	1.03	0.42	72.5	7.1
NN	00 02 54	-50 20.0	0.167	276.0		M5	11.95	1.50		+0.95t		10.31
Gl 4 A	00 03 02	+45 32.2	0.839	101.8	+0.0 SB	dK6 e	8.97	1.44	1.21	+0.71 J	87.0	8.67
Gl 4 B	00 03 02	+45 32.1	0.885	98.3	0.1	M0.5 V	9.02	1.45	1.20		87.0	8.72
Gl 4.1A	00 03 38	+58 09.5	0.260	76.7	-11.6	G5 V	6.43c	+0.64c	+0.11c		46.5	4.77c
Gl 4.1B	00 03 38	+58 09.5	0.260	76.7	-16	dG8	7.20c	+0.78c	+0.33c		46.5	5.54c
NN	00 03 40	-66 07.5	0.593	160.6		M4	12.16	1.55		1.04		10.86
Gl 4.2A	00 03 44	-49 21.2	0.592	93.9	2.6	G1 IV	5.71	0.52	0.03	0.17	48.3	4.13
Gl 4.2B	00 03 44	-49 21.2	0.592	93.9			11.50				48.3	9.9 *
Gl 5	00 04 01	+28 44.7	0.422	114.1	-5.5	K0 Ve	6.14	0.75	0.33		70.2	5.37
GJ 1002	00 04 13	-07 47.5	2.041	203.6	-42	M5-5.5	13.75	1.98	+1.60:	1.63	212.8	15.39
GJ 1003	00 04 46	+28 58.8	1.890	127.2		m	14.18	1.49	1.40	1.14	53.5	12.82

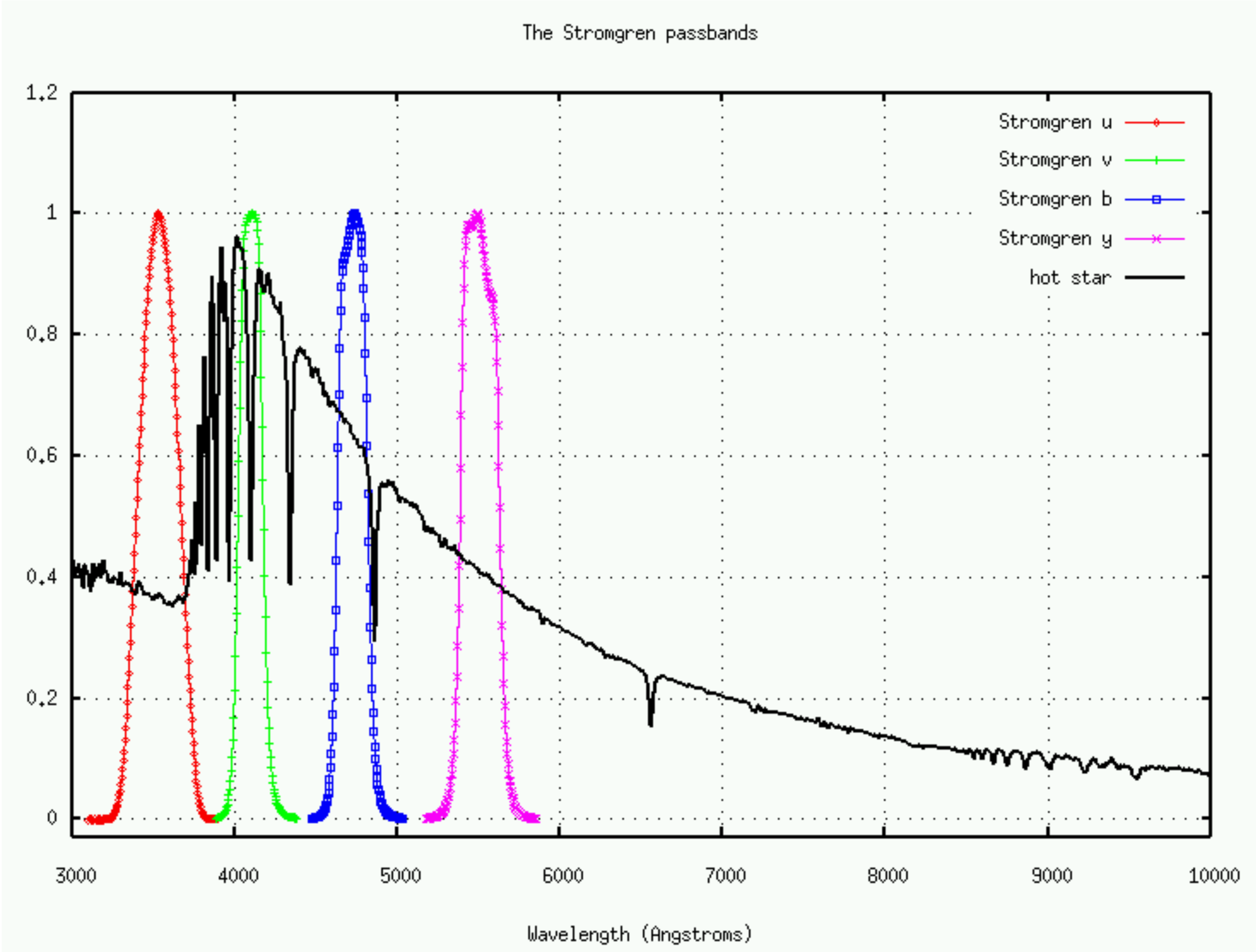


Astronomical Data: Images and Photometry





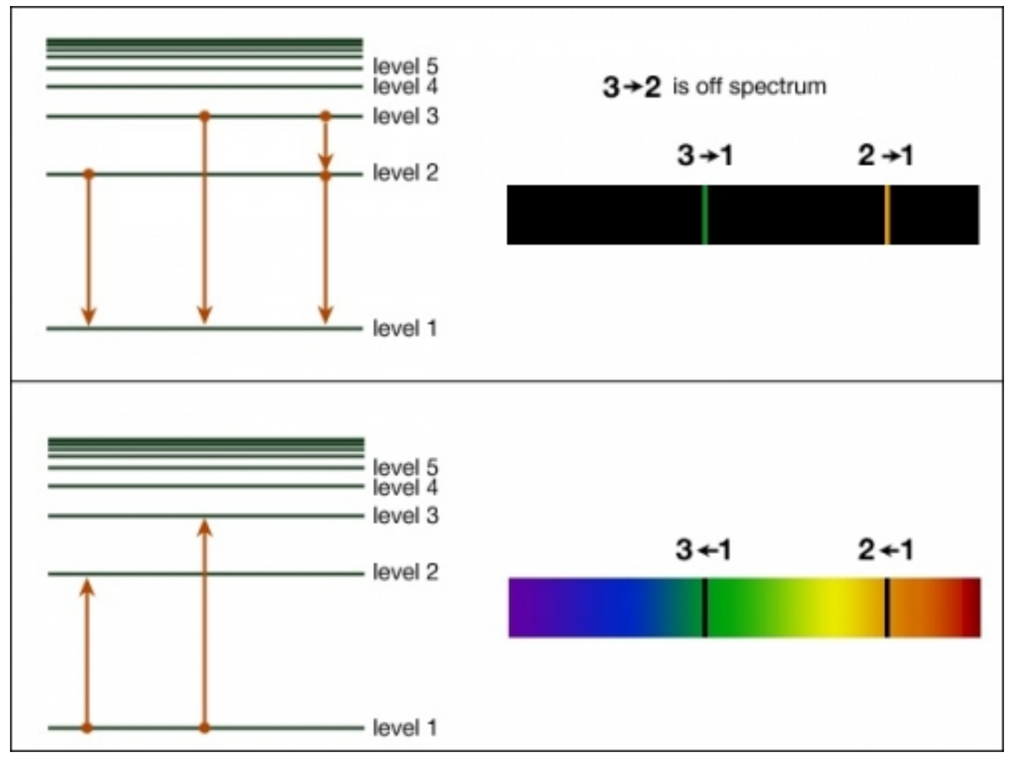


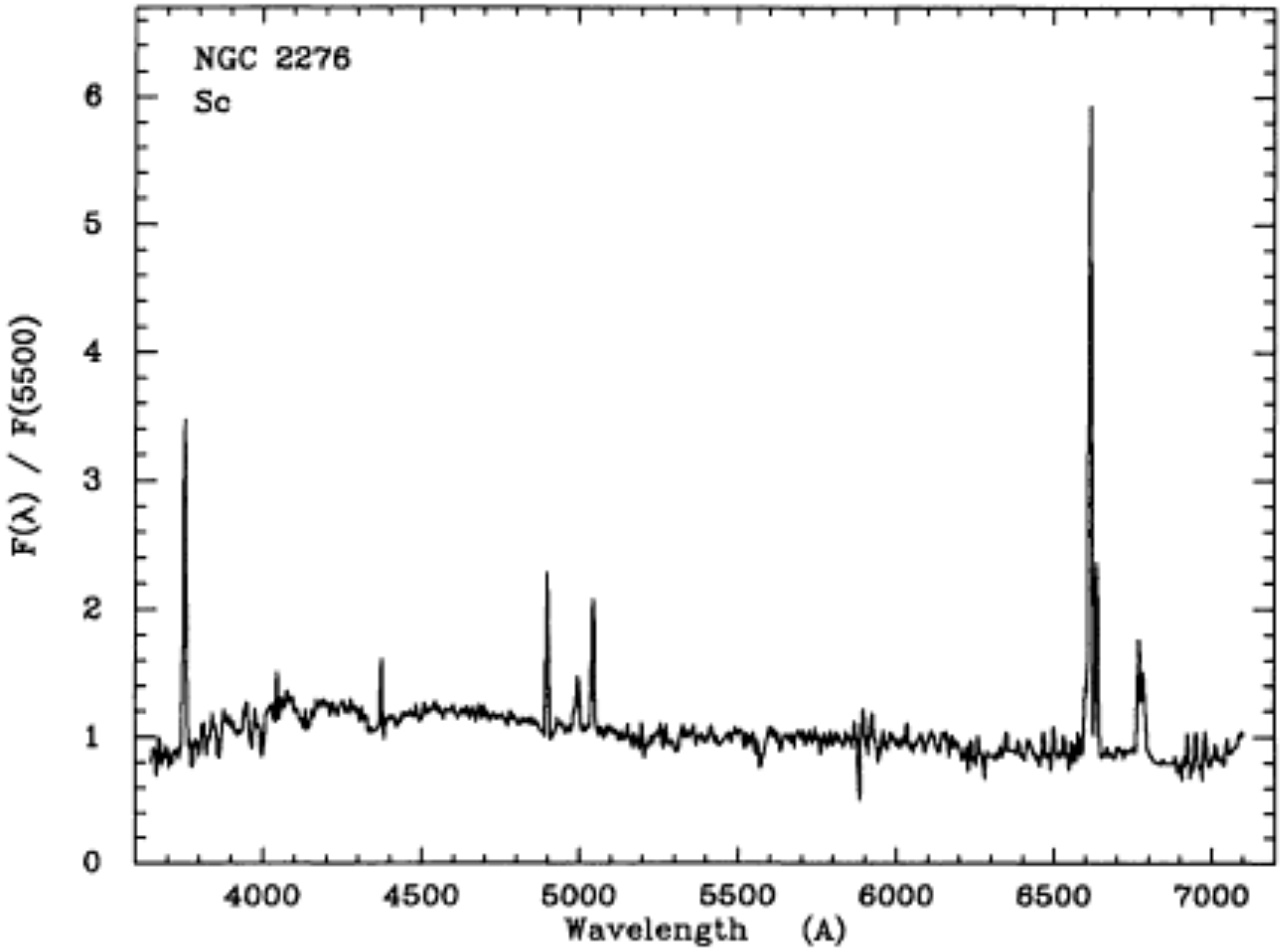


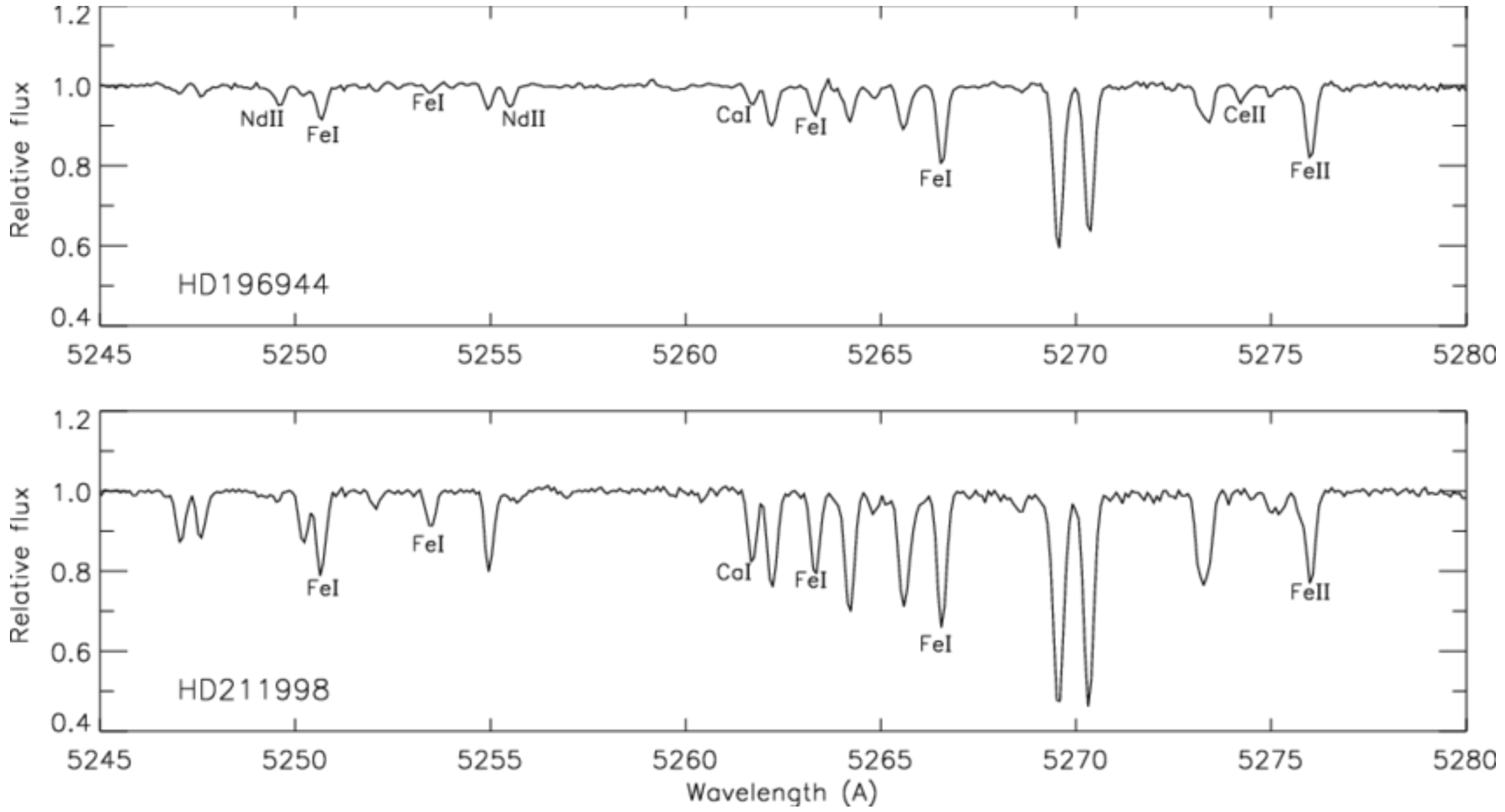
Astronomical Data: Images and photometry

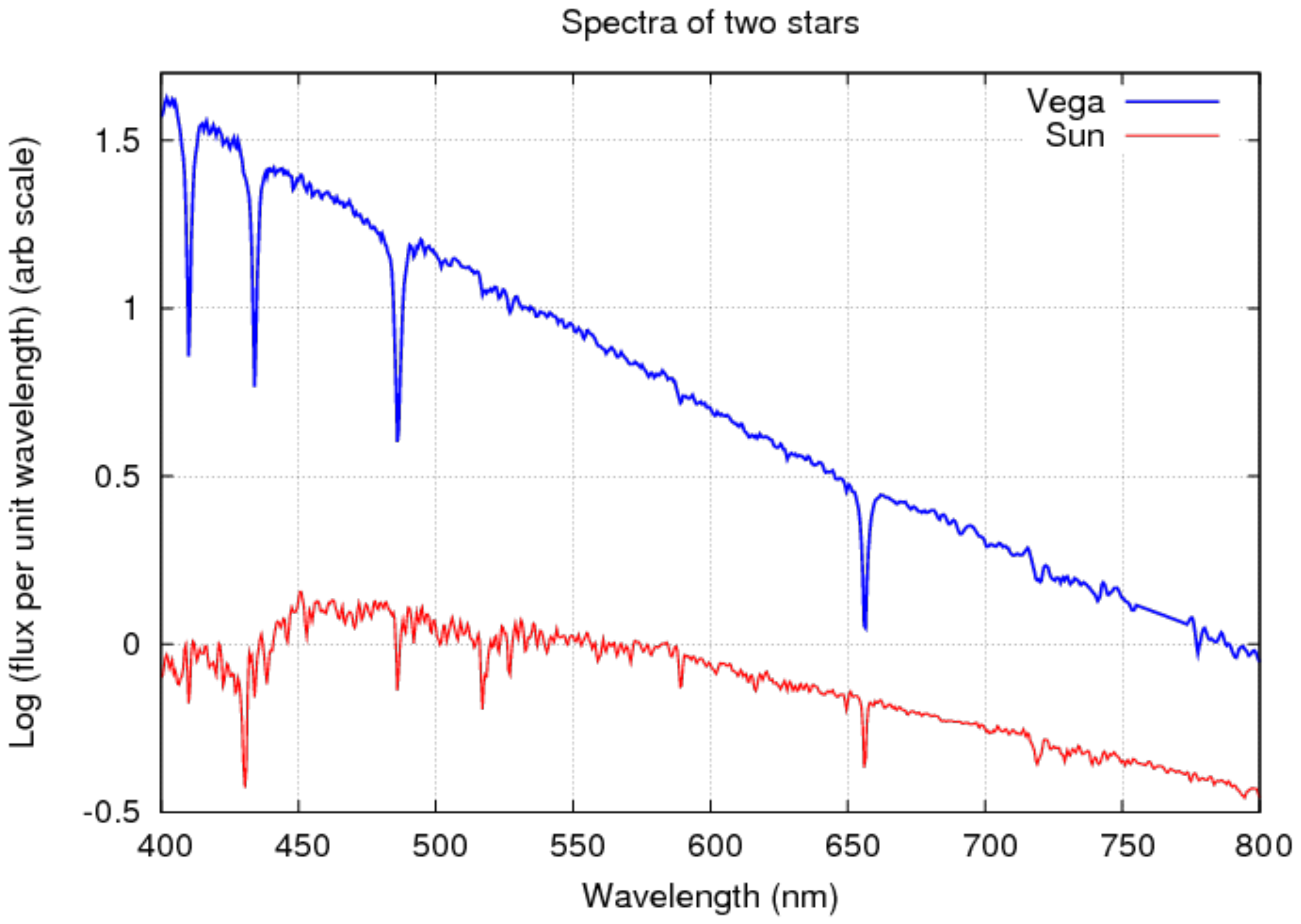
- They are not "photos" like those from regular cameras but maps of light intensity, where each pixel represents the amount of light coming from a point in the sky.
- Obtained using telescopes equipped with detectors like CCDs (Charge-Coupled Devices).
- **Photometry** is the precise measurement of the amount of light (flux) emitted by an astronomical object.
- Performed in specific wavelength bands (e.g., u,g,r,i,z from the SDSS photometric system).
- Helps study properties such as:
    - Apparent and absolute brightness.
    - Temperature and composition of stars.
    - Mass distribution in galaxies.

Astronomical Data: Spectra





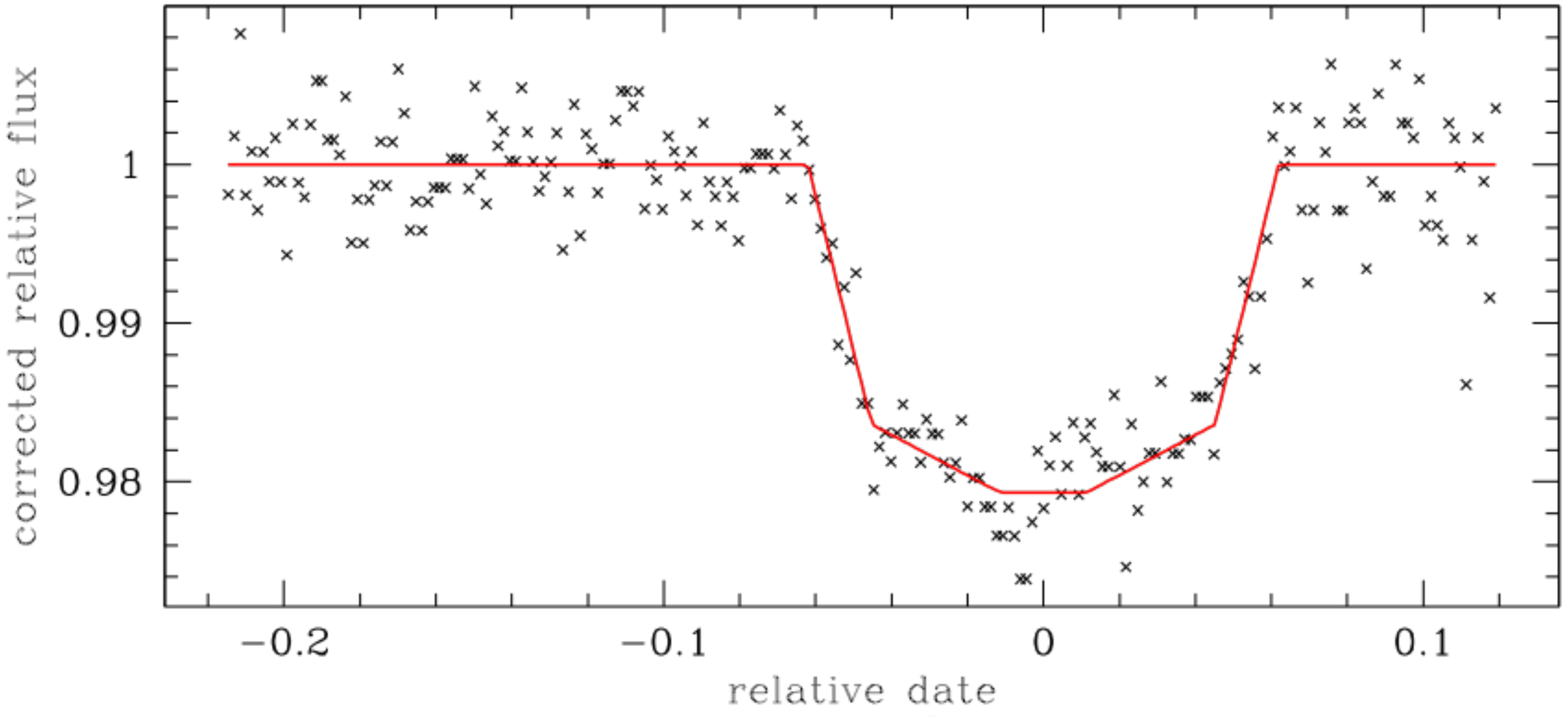


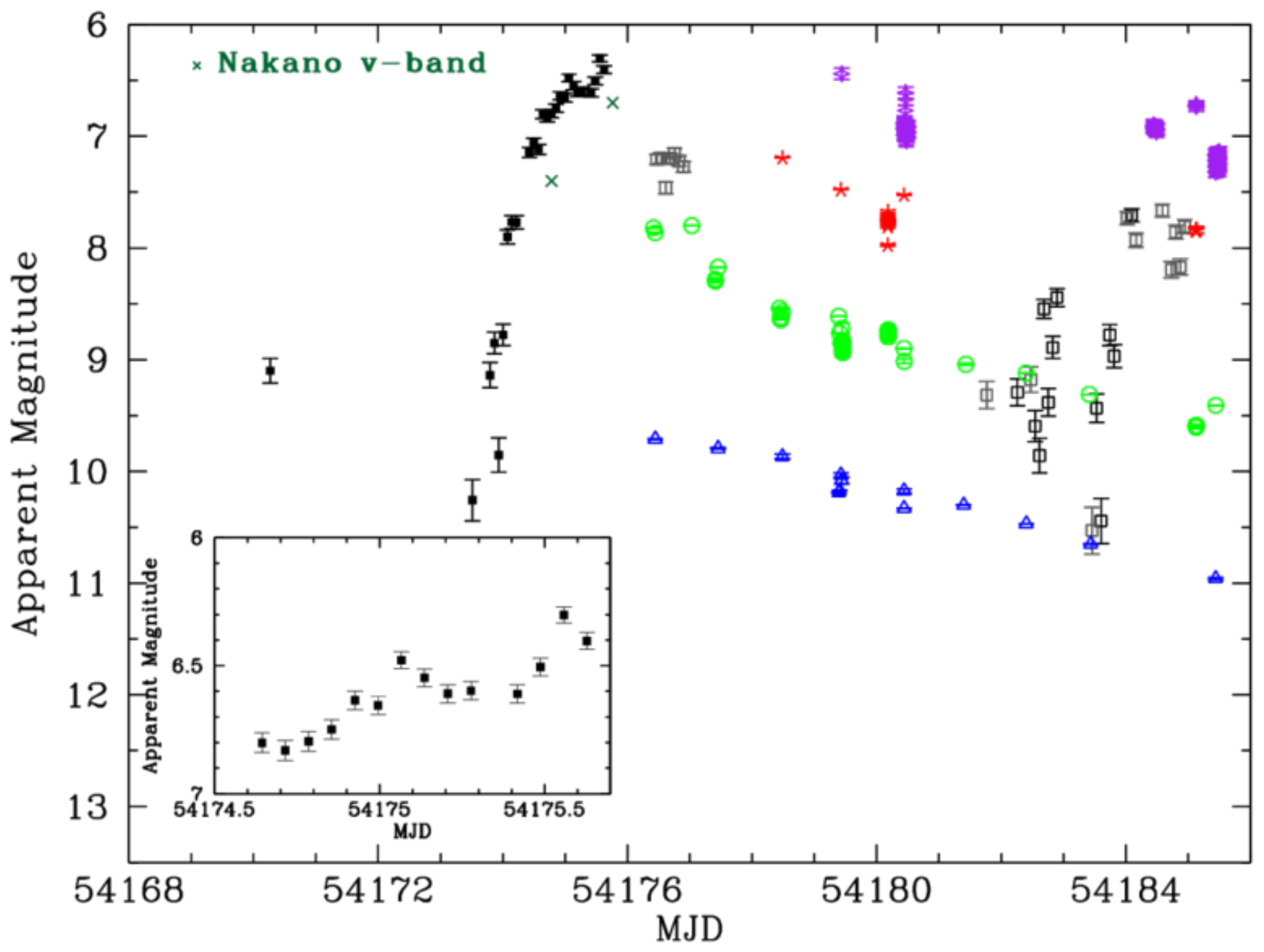


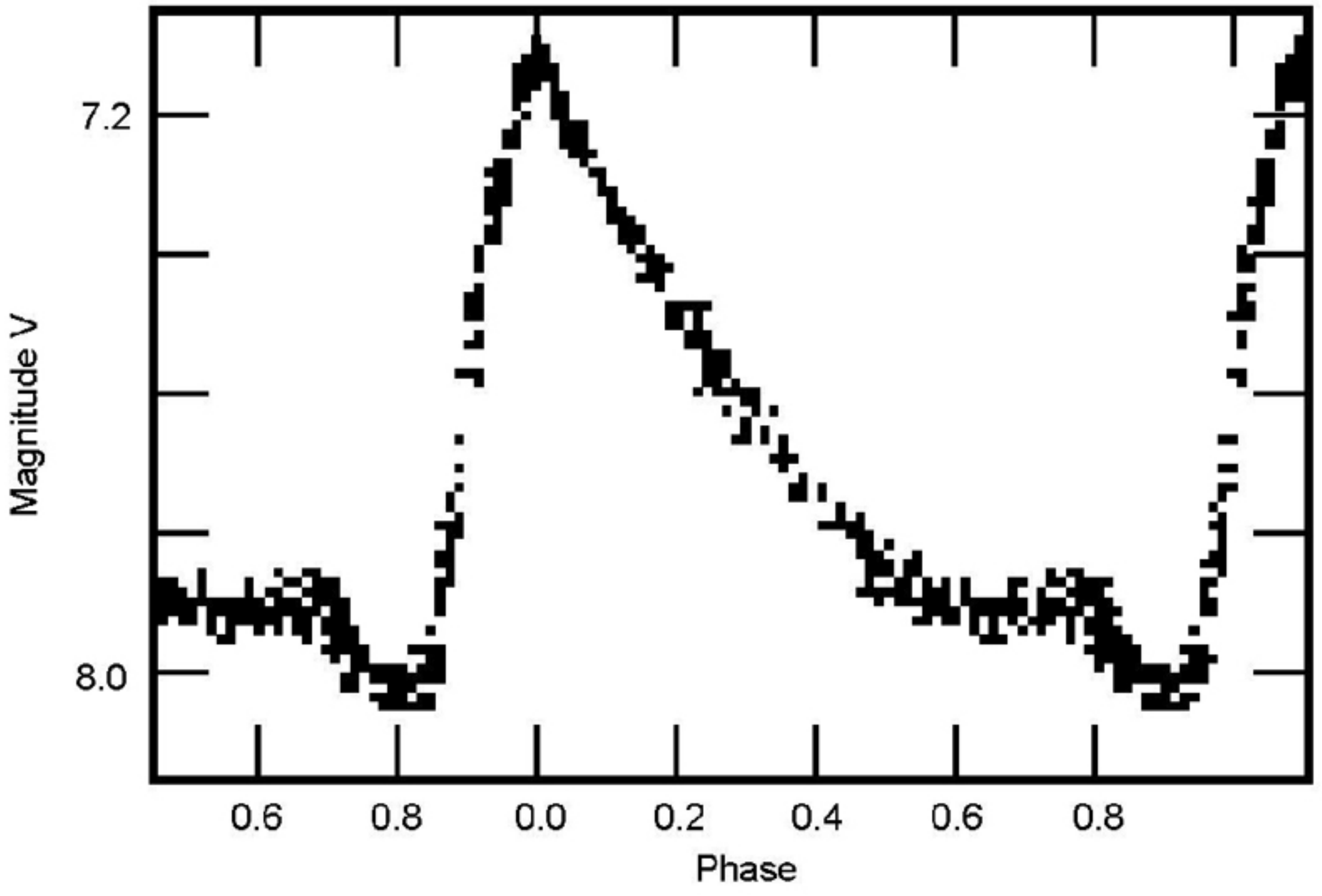
Astronomical Data: Spectra

- A spectrum is the distribution of light intensity from an astronomical object as a function of wavelength or frequency.
- Provides detailed information about the physical, chemical, and dynamical properties of celestial objects.
- Key Features in Spectra:
    - **Continuum:** Smooth emission from the object's surface or gas.
    - **Absorption Lines:** Dark lines where specific wavelengths are absorbed by elements.
    - **Emission Lines:** Bright lines where specific wavelengths are emitted by hot gas.

Astronomical Data: Transient Objects and Time Series



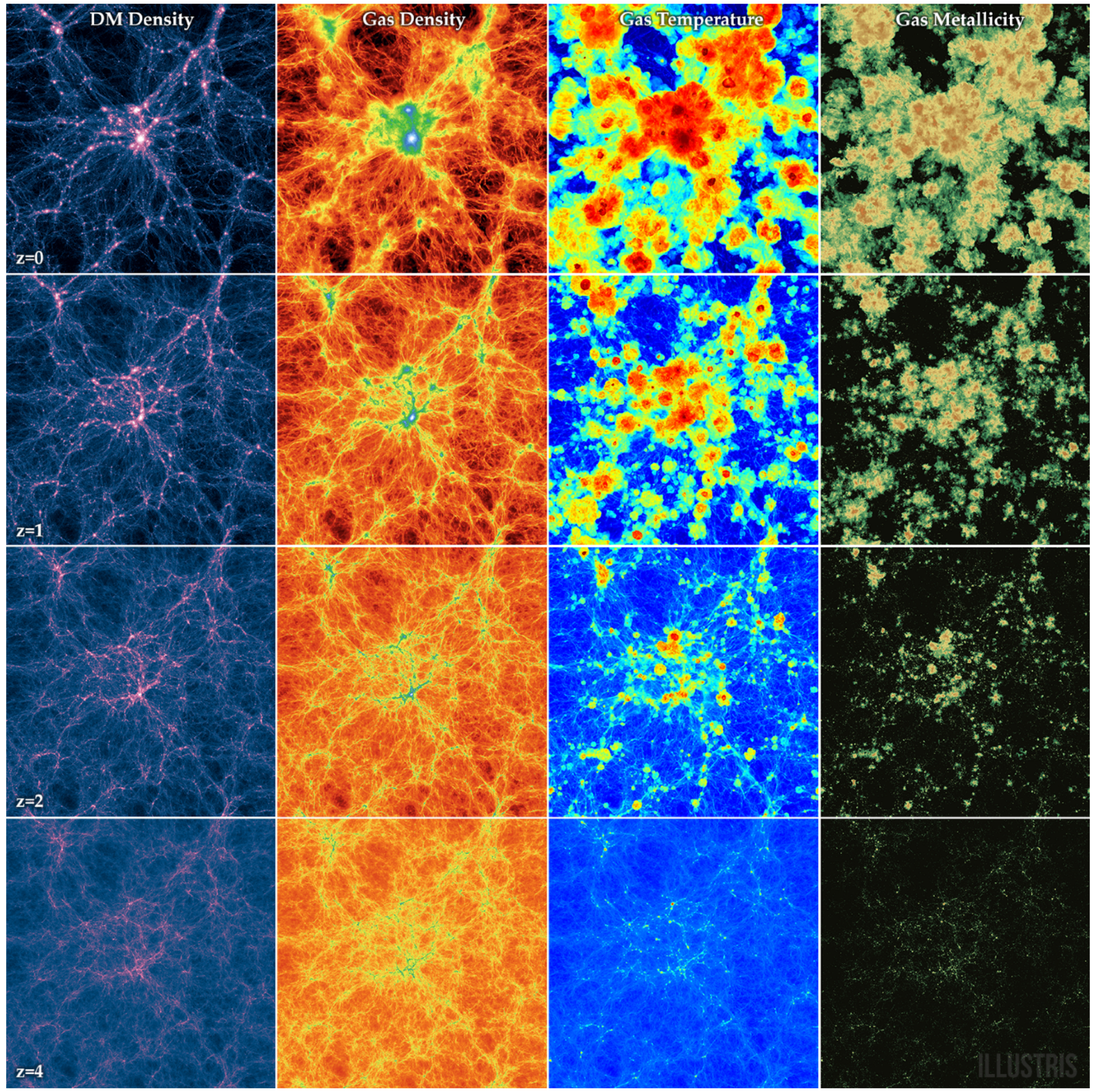




Astronomical Data: Transient Objects and Time Series

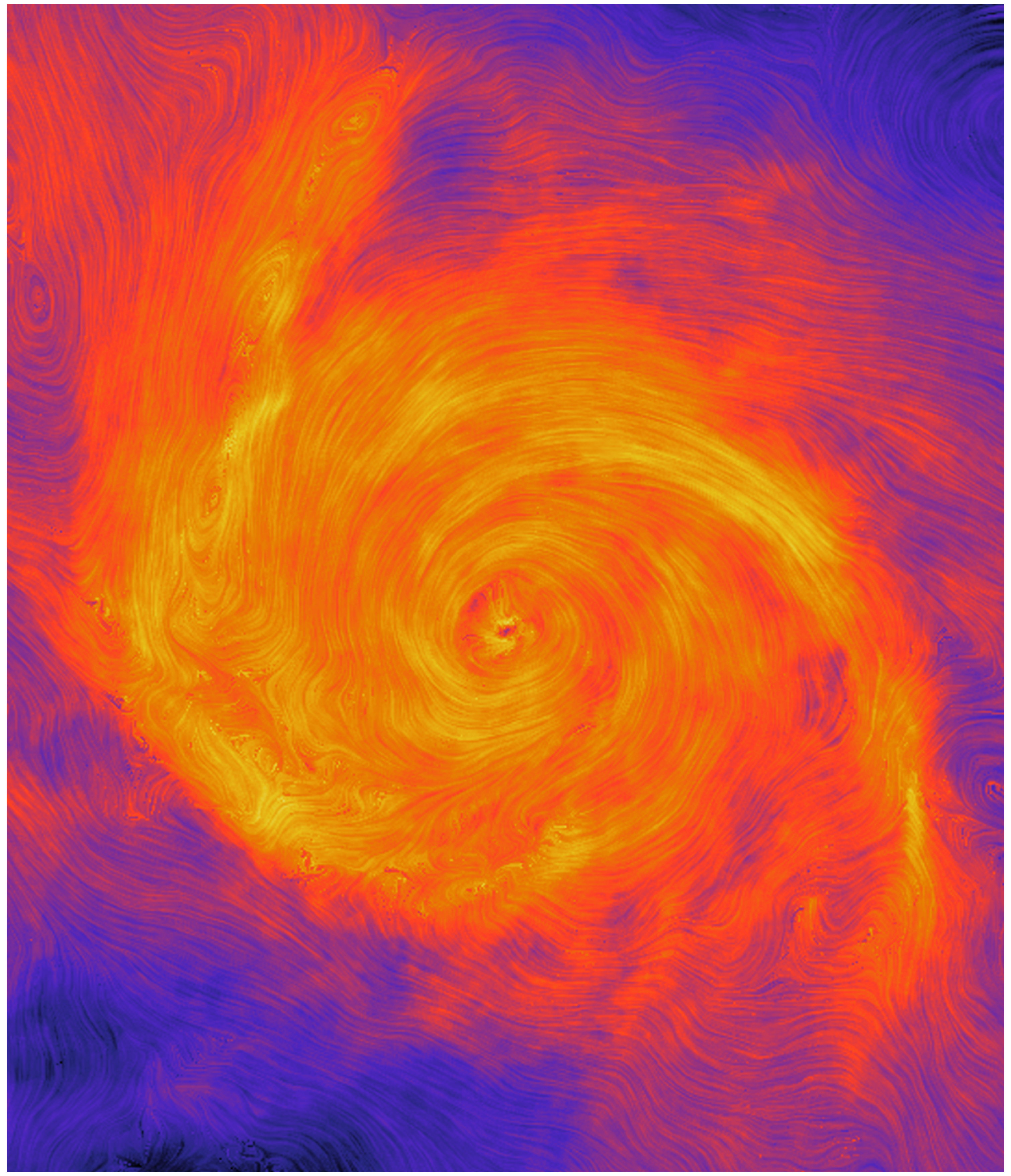
- Variations in the brightness, velocity, or other properties of astronomical objects.
- Periodicity: Repeated patterns in brightness or velocity (e.g., pulsating stars, eclipsing binaries).
- Transients: Sudden, non-repeating events (e.g., supernovae, microlensing).
- Types of transients:
  - Variable Stars: Study pulsating stars (e.g., Cepheids) to measure distances.
  - Exoplanet Detection: Detect transits as a planet passes in front of its host star.
  - Binary Systems: Measure radial velocity variations to determine orbital parameters.
  - Transient Events: Monitor supernovae or gamma-ray bursts.

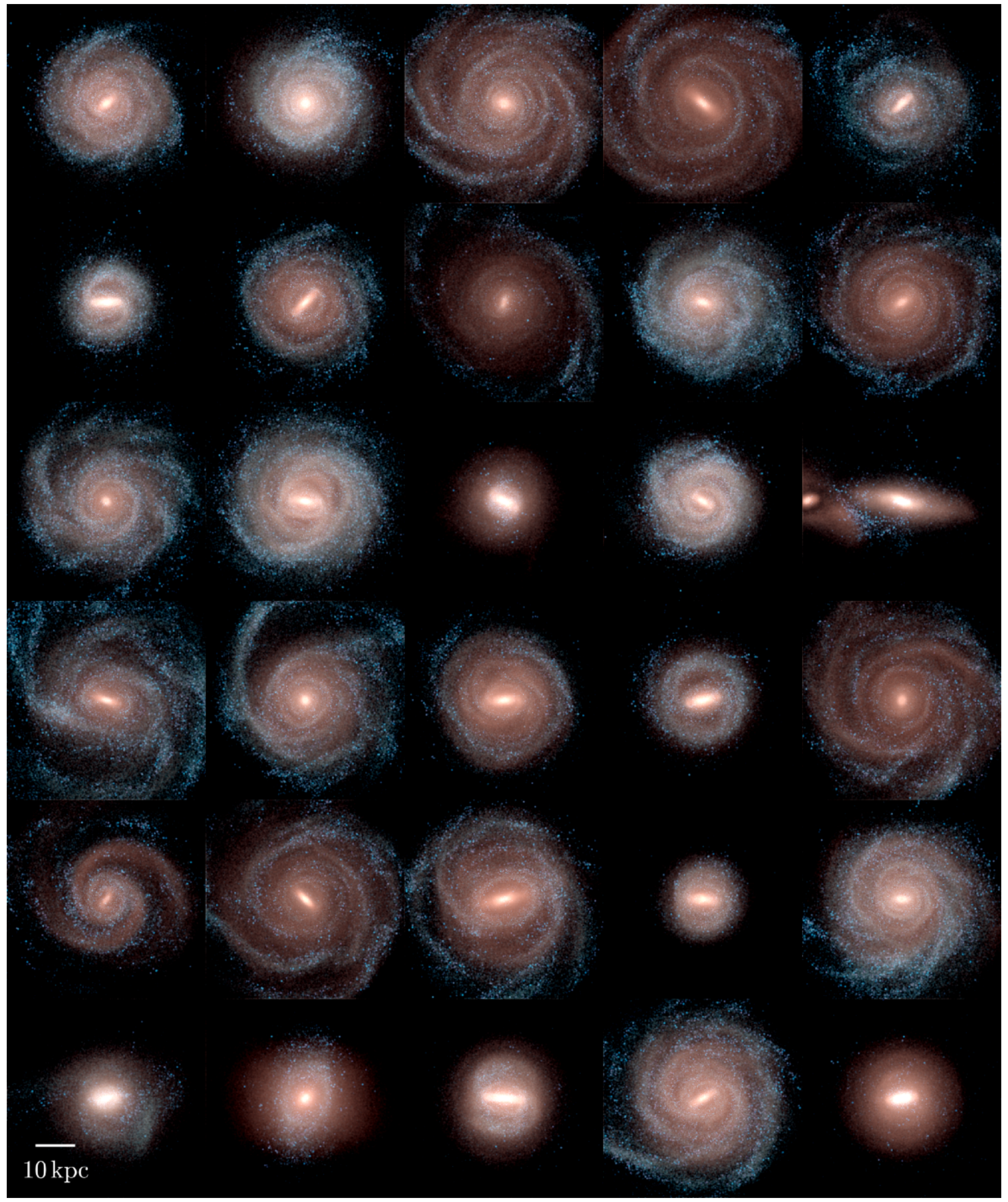
Astronomical Data: Simulations



Illustris Project

Astronomical Data: Simulations



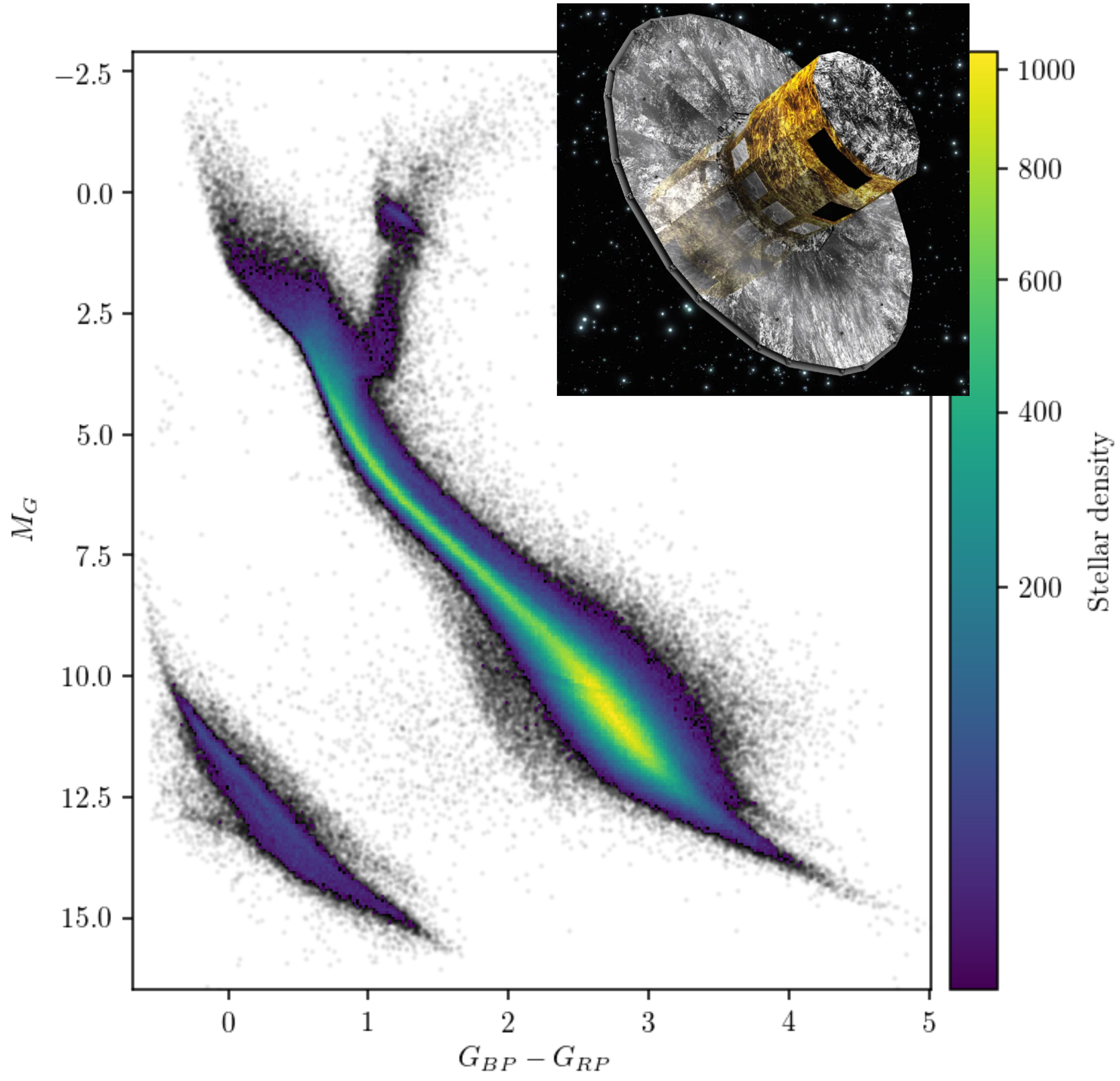


Auriga Simulation

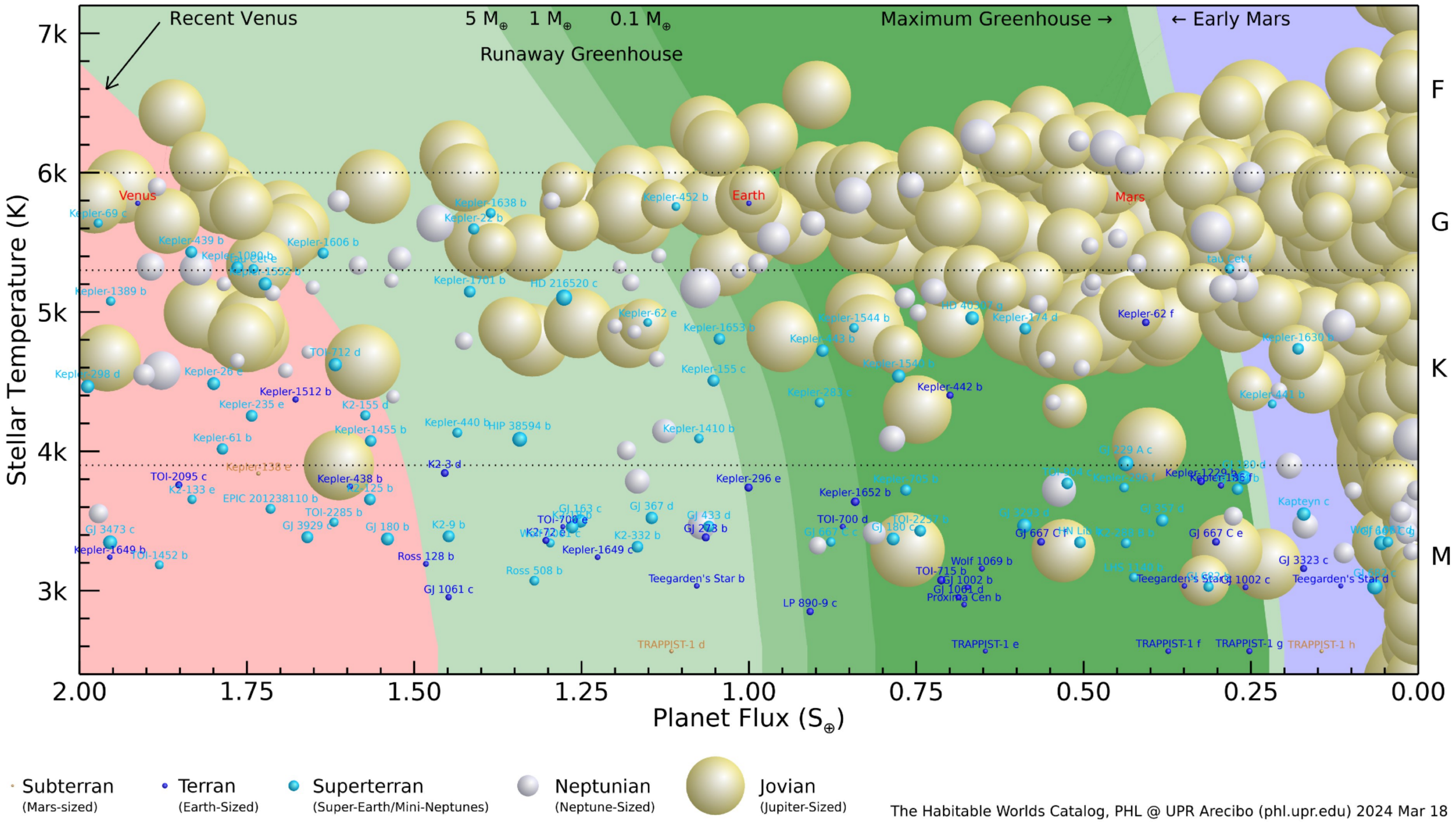
Astronomical Data: Simulations

- Simulations in astronomy model the evolution and behavior of astrophysical systems using computational methods and physical principles.
- The purpose is the study of processes that are difficult or impossible to observe directly, such as galaxy formation, large-scale structure evolution, or star formation.
- How Simulations Work
    - Input Physics: Include gravity, hydrodynamics, radiation, and feedback processes (e.g., from supernovae), magnetism, dark matter, and dark energy.
    - Numerical Methods: Use grids or particles to represent matter and solve equations governing astrophysical processes. Common methods include N-body simulations (gravity-dominated) and smoothed-particle hydrodynamics (SPH) for fluids.
    - Scale: Simulations can range from small (e.g., single star) to large-scale cosmological simulations.

Astronomical Data: Catalogs



Gaia



HWC


https://www.cosmos.esa.int/web/gaia/end-of-observations
Gaia
2013-2025

Astronomical Data: Catalogs

- Collections of data about astronomical objects, including their positions, brightness, motions, and other properties.
- Object Information: Include coordinates (e.g., right ascension and declination), magnitudes, distances, velocities, and classifications.
- Large Scale: Modern catalogs can contain millions or even billions of objects (e.g., Gaia DR3).
- Multidimensional Data: Combine photometric, astrometric, and spectroscopic data.
- Some examples: Gaia, SDSS, 2MASS, APOGEE, Hipparcos, NASA Exoplanets, ….

How ML/AI is been used in Astronomy

- **Classification:** Categories or labels are applied to objects or features. Based on a training set (labeled or unlabeled), the algorithm learns the characteristics that relate an instance to a category. When applied to a new instance, the algorithm assigns the most likely category label.

- **Regression:** Assignment of a numerical value (or values) based on the characteristics that are learnt or otherwise predicted by the machine learning algorithm. As with classification, a training set may be used or the characteristics may be inferred from the dataset.

- **Clustering:** Determine whether an object or a feature is part of (i.e., a member of) something. This might be a physical structure or association—as in the more familiar usage of the term in astronomy as applied to open, globular, or galactic clusters—or a region within an N-dimensional parameter space.

Fluke and Jacobs (2019), Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

How ML/AI is been used in Astronomy

- **Forecasting:** The purpose of the machine learning algorithm is to learn from previous events, and predict or forecast that a similar event is going to occur. There is an implicit time-dependence to the prediction

- **Generation and reconstruction:** Missing information is created, expected to be consistent with the underlying truth. The cause of the missing information might be due to the presence of noise, processing artifacts, or additional astronomical phenomena, all of which conspire to obscure the required signal.

- **Discovery:** New celestial objects, features, or relationships are identified as a consequence of the application of a ML or AI method.

- **Insight:** Moving beyond the discovery of celestial objects, new scientific knowledge is demonstrated as a consequence of applying machine learning or AI. This includes cases where insight is gained into the suitability of applying machine learning, choice of data set, hyperparameters, and comparisons with human-based classification.

Fluke and Jacobs (2019), Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

Techniques

Data/method	**ANN**	**CNN**	**GAN**	**SVM**	**DT**	**RF**	**DBSCAN**	***k*-NN**	***k*-M**
Image	•	•	•	•	•	•		•	
Spectroscopy	•	•		•		•			•
Photometry	•				•	•	•		•
Light curve		•				•			
Time series	•	•			•	•	•		
Catalogue	•			•	•	•	•	•	
Simulation	•	•	•	•		•			

Fluke and Jacobs (2019), Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

Techniques

Data/method	ANN	CNN	GAN	SVM	DT	RF	DBSCAN	*k*-NN	*k*-M
Image	•	•	•	•	•	•		•	
Spectroscopy	•	•		•		•			•
Photometry	•				•	•	•		•
Light curve		•				•			
Time series	•	•			•	•	•		
Catalogue	•			•	•	•	•	•	
Simulation	•	•	•	•		•			

Most used models!

Fluke and Jacobs (2019), Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

Well-Established Applications

- **Classification** and **forecasting** of solar flares. Segmentation for identification of umbra/penumbra/ photosphere in the solar surface.
- **Identification** of candidates to extrasolar planets from stellar lightcurves (Kepler)
- Stellar and photometric **classification** of stars, leading to finding new objects of specific types of stars (WR, hot sub-dwarfs, etc).
- **Classification** of galaxies from optical and radio imaging surveys. **Prediction** of physical properties from emission-line spectra. **Identification** of galaxies undergoing a special evolutionary phase as predicted by simulations.
- **Identification** and **classification** of transient objects.
- Accurate **estimation** of distance to extragalactic objects from photometric information (photometric redshift).
- **Identification** of systems affected by gravitational lenses in wide-area surveys
- **Discriminating** noise from signal in the detection of gravitational waves.

Progressing Applications

- **Reduction of false detections** from the moving objects detection pipelines. **Detection** and **classification** of asteroids.
- **Assigning** morphological types to radio-detected AGNs.
- **Identification** of blazar candidates in catalogues of high-energy sources (Fermi-LAT)
- **Detecting** high-redshift extremely luminous quasars
- **Discriminating** populations of BAL QSOs from non-BAL QSOs
- **Examination** of the output of cosmological simulations to connect physical properties of galaxies, dark matter halos and the cosmic environment.
- **Classification** of DM sub-halos. **Assignment** of galaxies to halos in simulations.

Emerging Applications

- **Classification** of atmospheric features on the Surface of Mars aiming at predicting dust storms.

- **Discovery** of previously unknown impact craters.

- Study of the ISM in our Galaxy. Spatial or chemical **clustering** of components in atomic and molecular clouds.

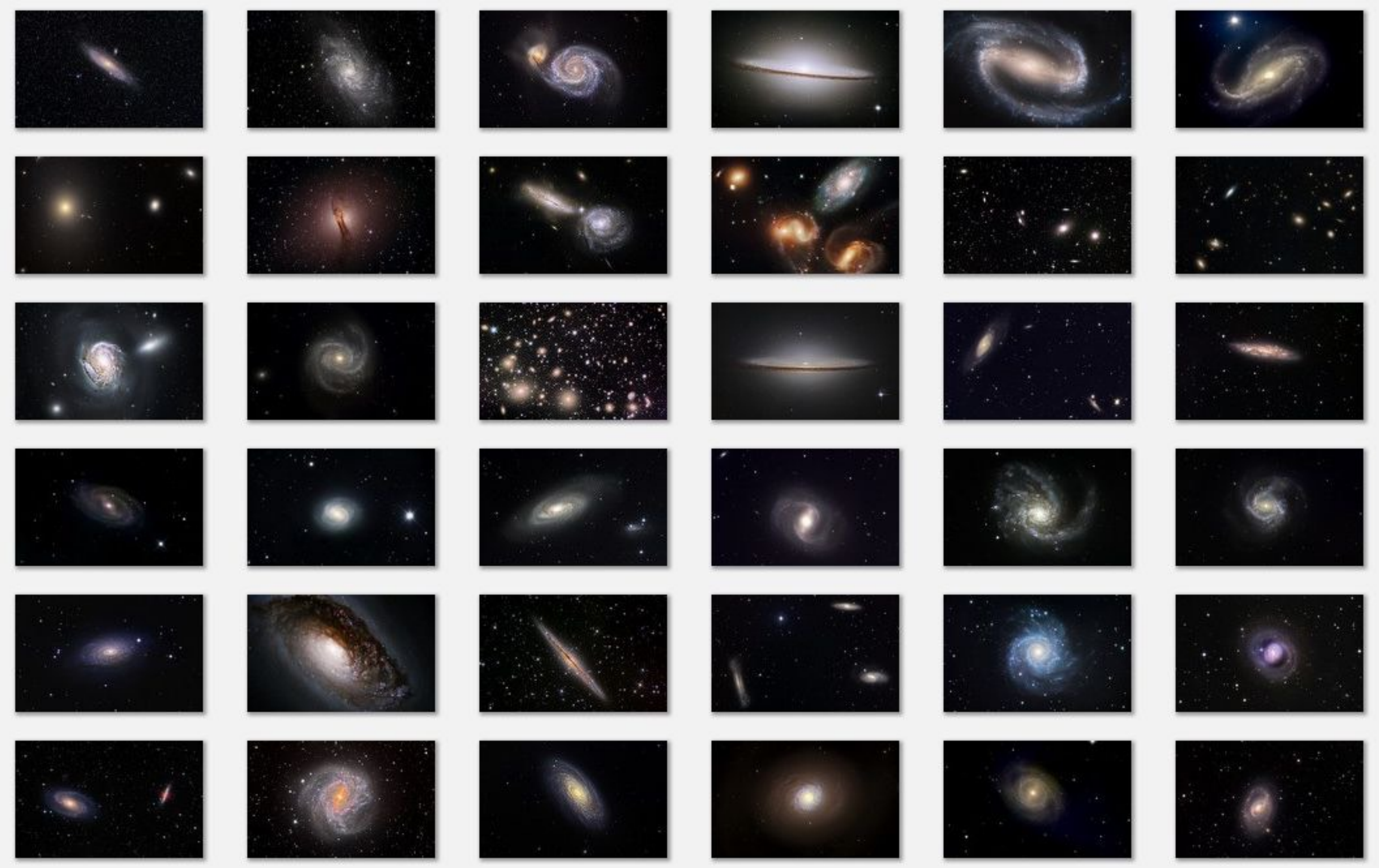
- **Determination** of dust reddenning in millions of stars, with application to GAIA data.

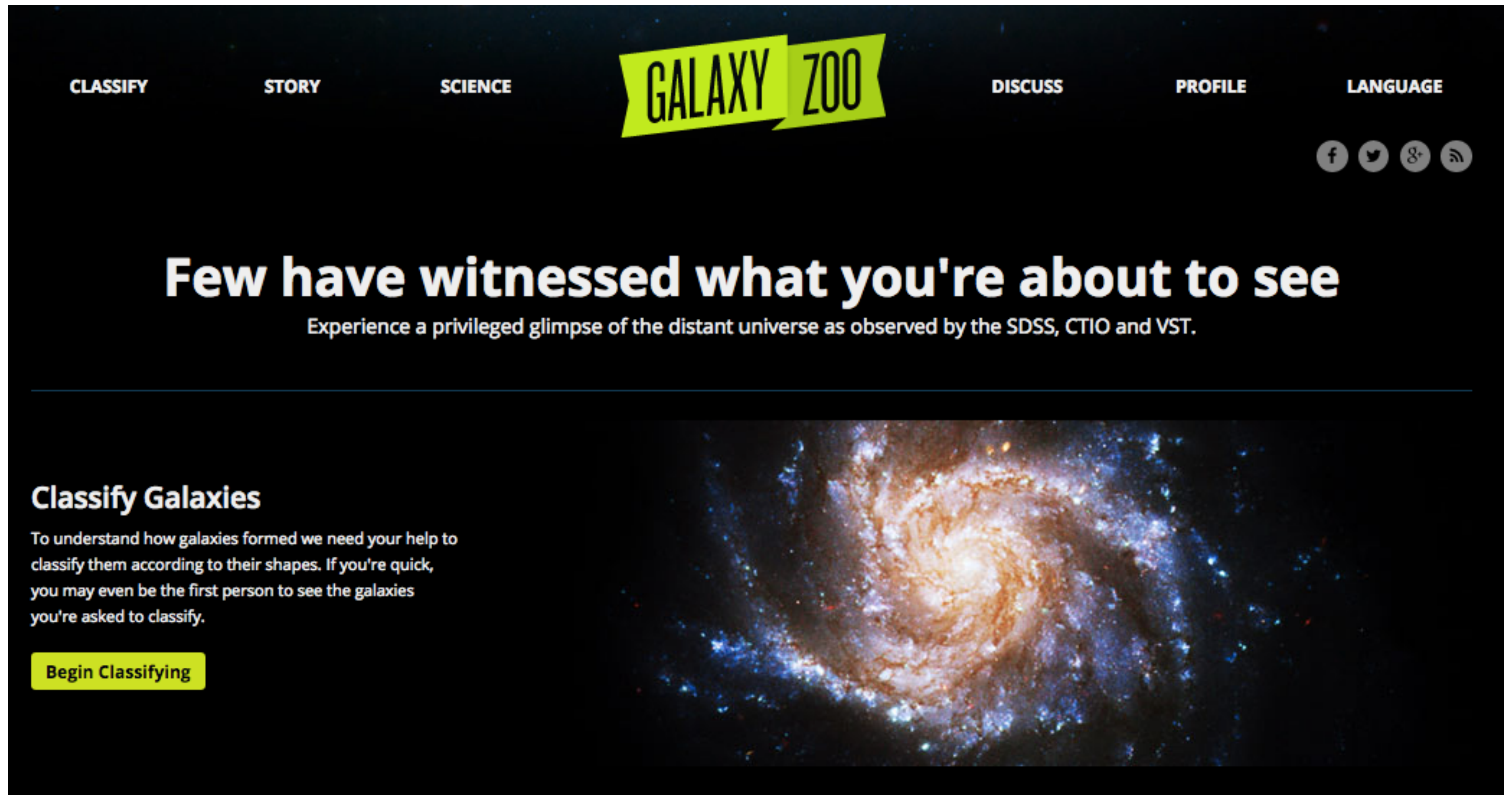
- **Discovery** of new open clusters from overdensities in GAIA DR2 data

- **Identify** faults in telescope drive systems that can be tackled in real time with automated expert systems

Some examples…

Galaxy Zoo





Goal: Train a model to classify the morphology of galaxies based on their images

Datasets (> 300.000 galaxies):

- Sloan Digital Sky Survey (APO, US, 2.5m)
- Dark Energy Camera Legacy Survey (DECaLS) (V. Blanco Telescope, Chile, 4m)
- Hawaii H2O Survey (Subaru Telescope, US, 8.2m)
- Cosmic Evolution Early Research Science (CEERS) with JWST (Space, 6.5m)

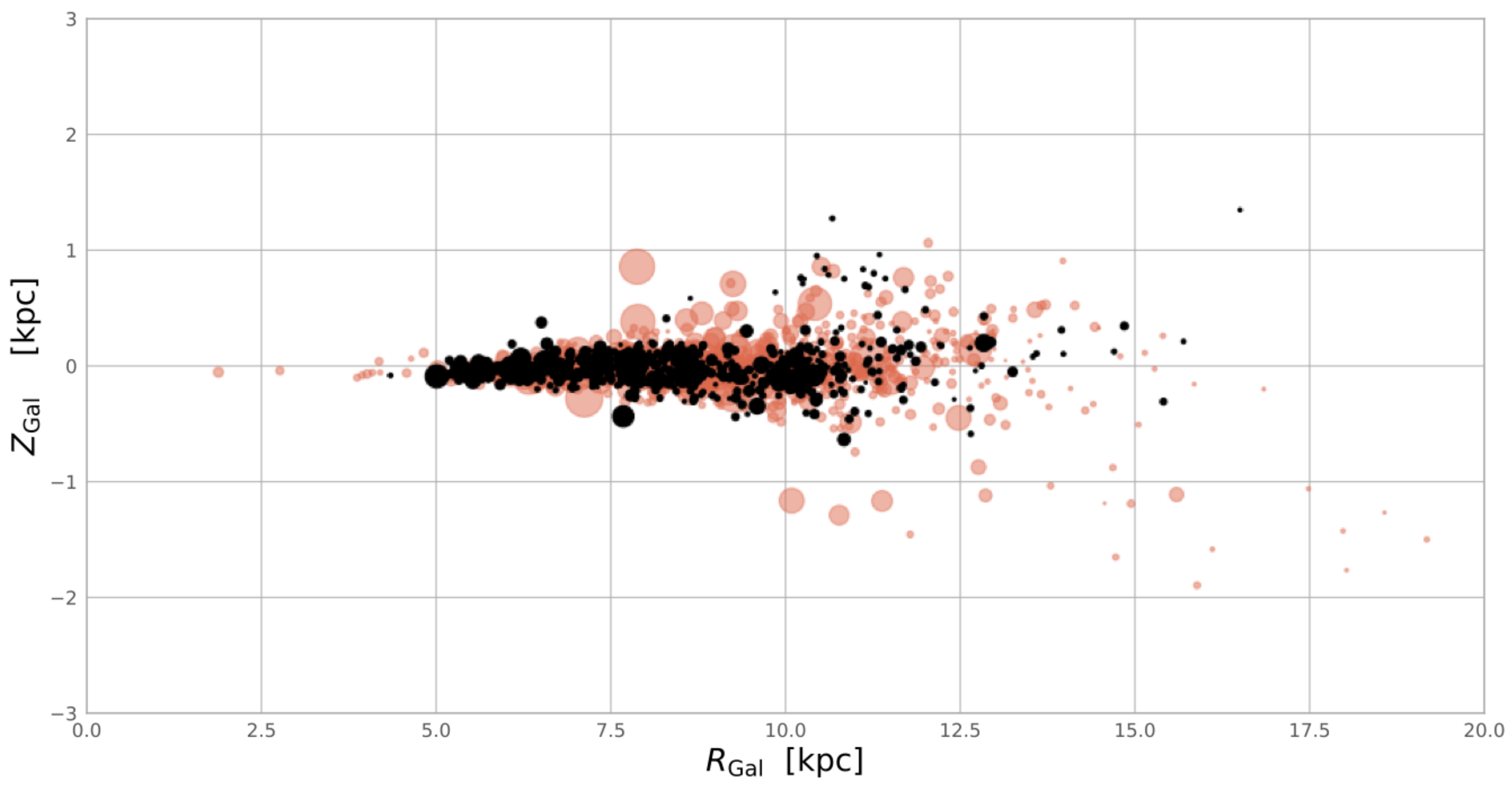
Methodology:

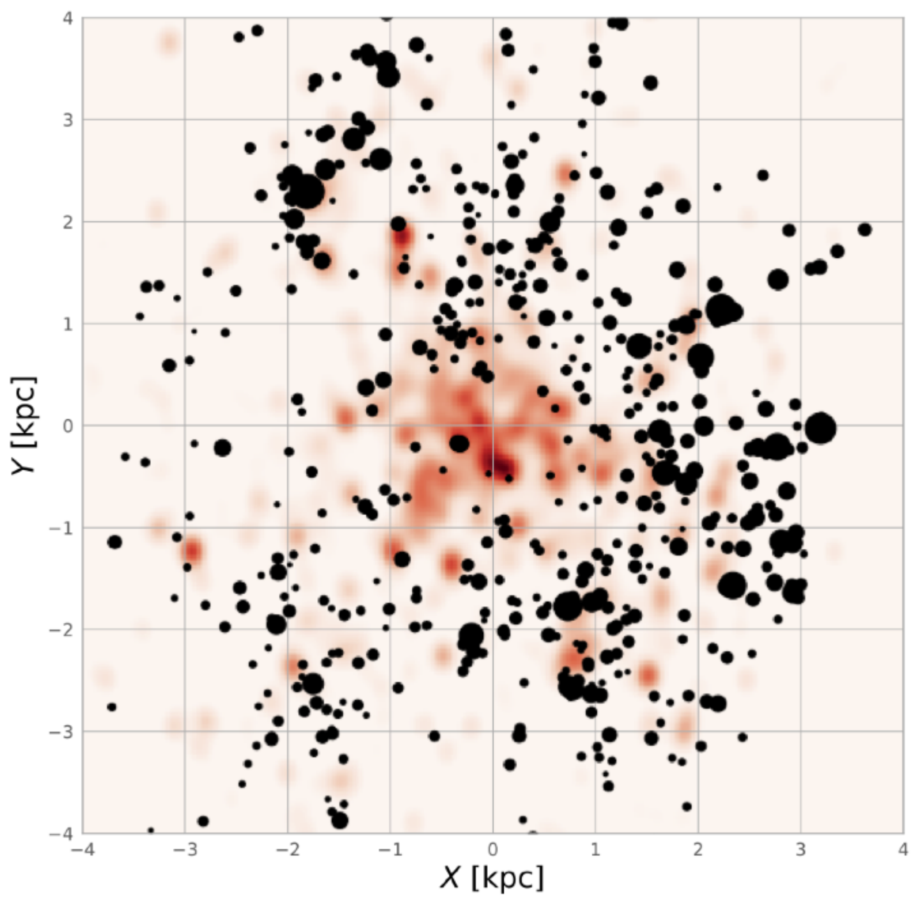
- Train a classifier with labelled data
- Labels are put by thousands of volunteers with no specific field knowledge

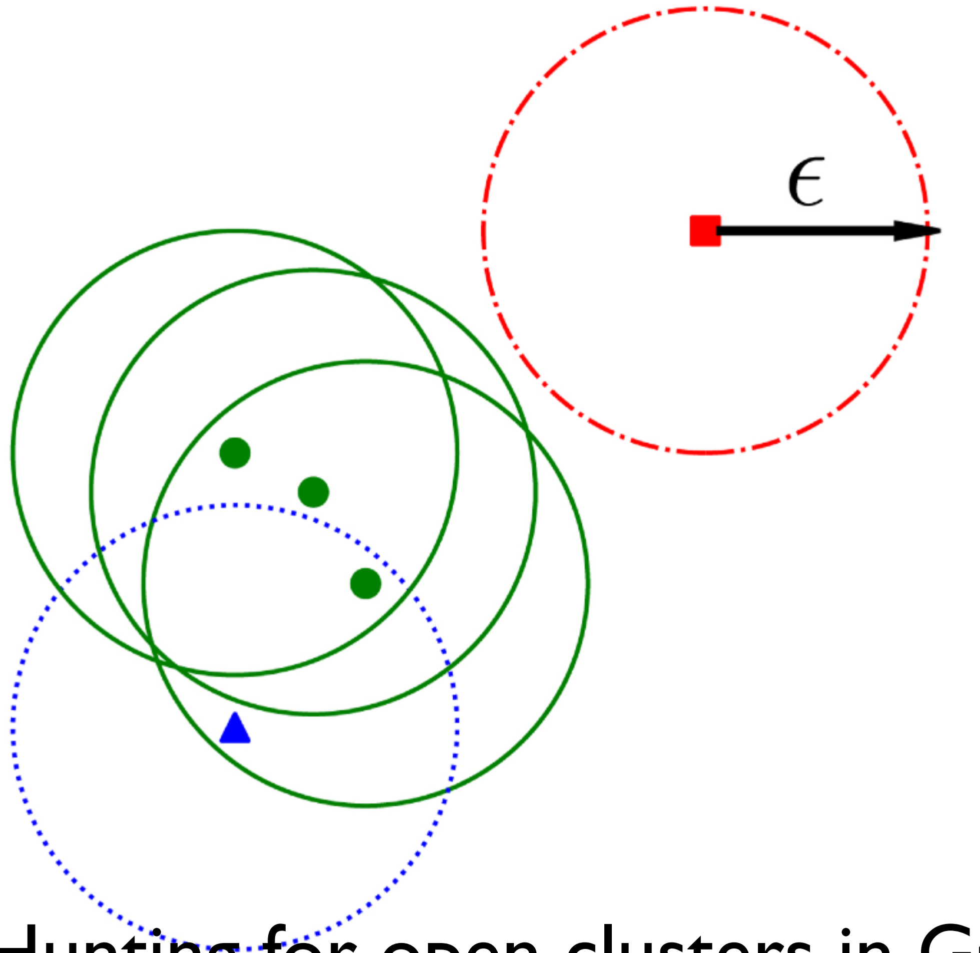
Results:

- 7.5M classifications(!) from which 140.000 get > 30 classifications
- Robust classifications
- Some work to be done to flag misclassifications
- This enabled producing >75 publications 2008-2023

Discovery of new clusters in the Galaxy







Goal: Identify new open clusters within the Galactic disc using Gaia DR2 data.

Methodology:

- Applied a clustering algorithm (DBSCAN) to Gaia DR2 astrometric data.
- Validated findings with color-magnitude diagrams and proper motion analysis.

Results:

- Discovered 582 new nearby open clusters
- Confirmed the existence of these clusters through independent methods.
- Enhanced understanding of the Galactic disk's structure.

Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc (Castro-Ginard et al., 2020)

Planet Hunters TESS

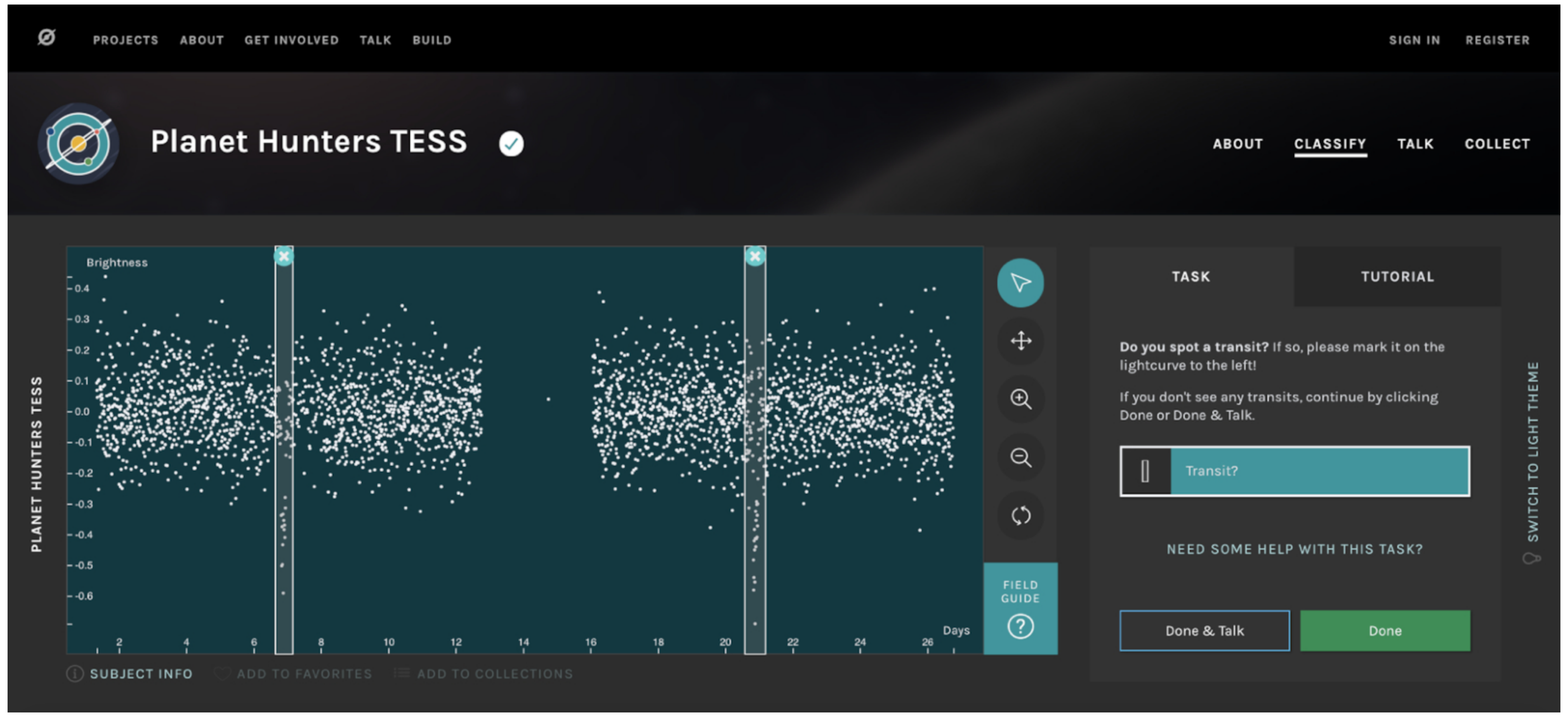
Goal: Identify exoplanets by using the transit method.

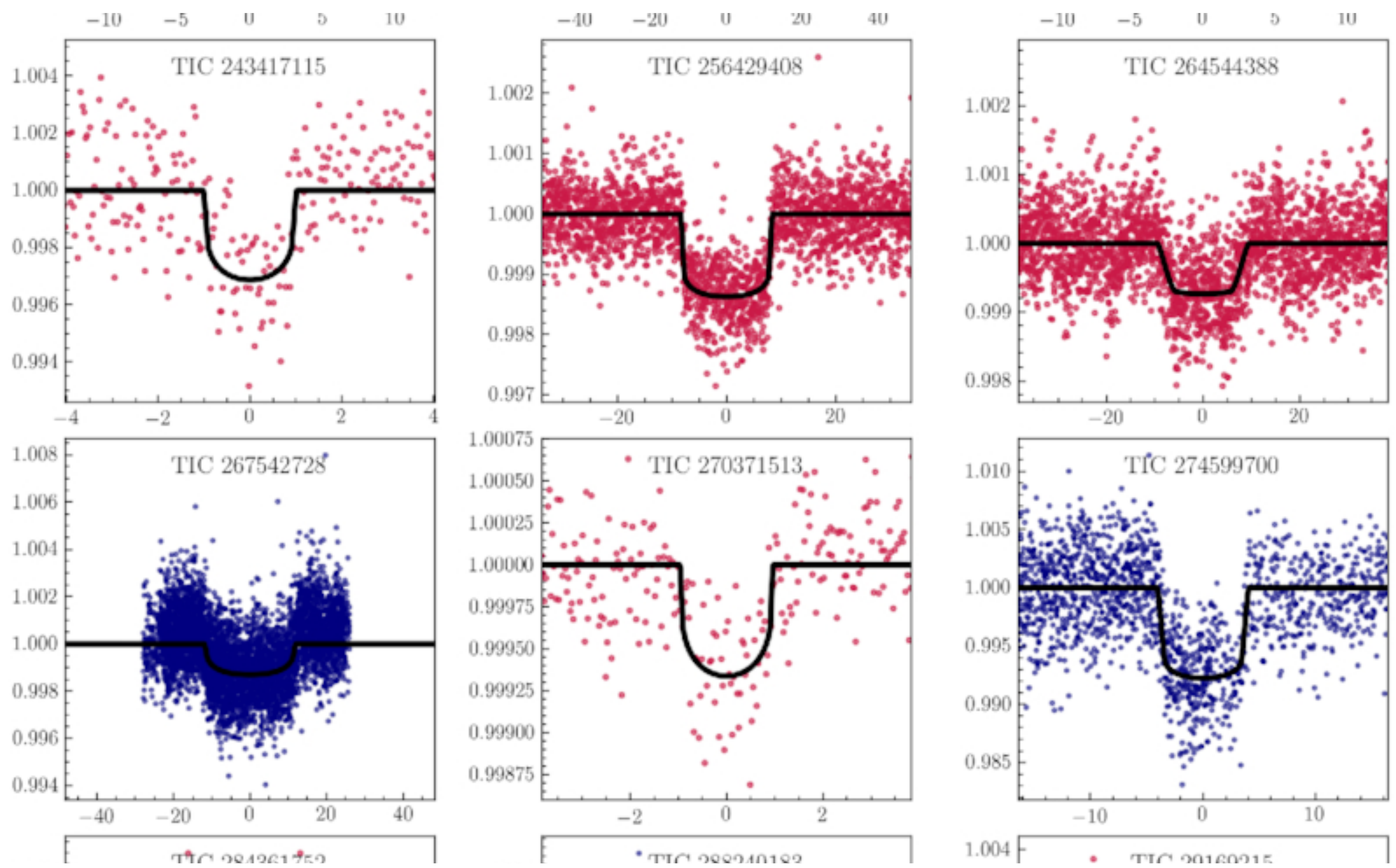
Data: Lightcurves observed with dedicated space missions: Kepler or TESS

Method:

- Build a classifier that automatically identifies potential candidates
- Volunteers help to identify tricky or borderline patterns, suggesting a classification (variable star, data glitch, potential planet)

Results: More than 100 new planetary systems identified in Kepler data





What's next? **Tons of data!!!**

Vera Rubin: Massive data processing in (near) real life!

- Cover all the visible sky every 2-3 nights (~20TB per night)
- Exhaustive study of the transient sky. About 10 million of alerts per night (20.000 alerts per minute)
- Latency of alert: 60 seconds

VERA C. RUBIN
OBSERVATORY

What's next? **Tons of data!!!**

SKAO: World's largest radio observatory

- The SKA will detect hundreds of millions of astrophysical systems
- Expected to generate 600 PB/year

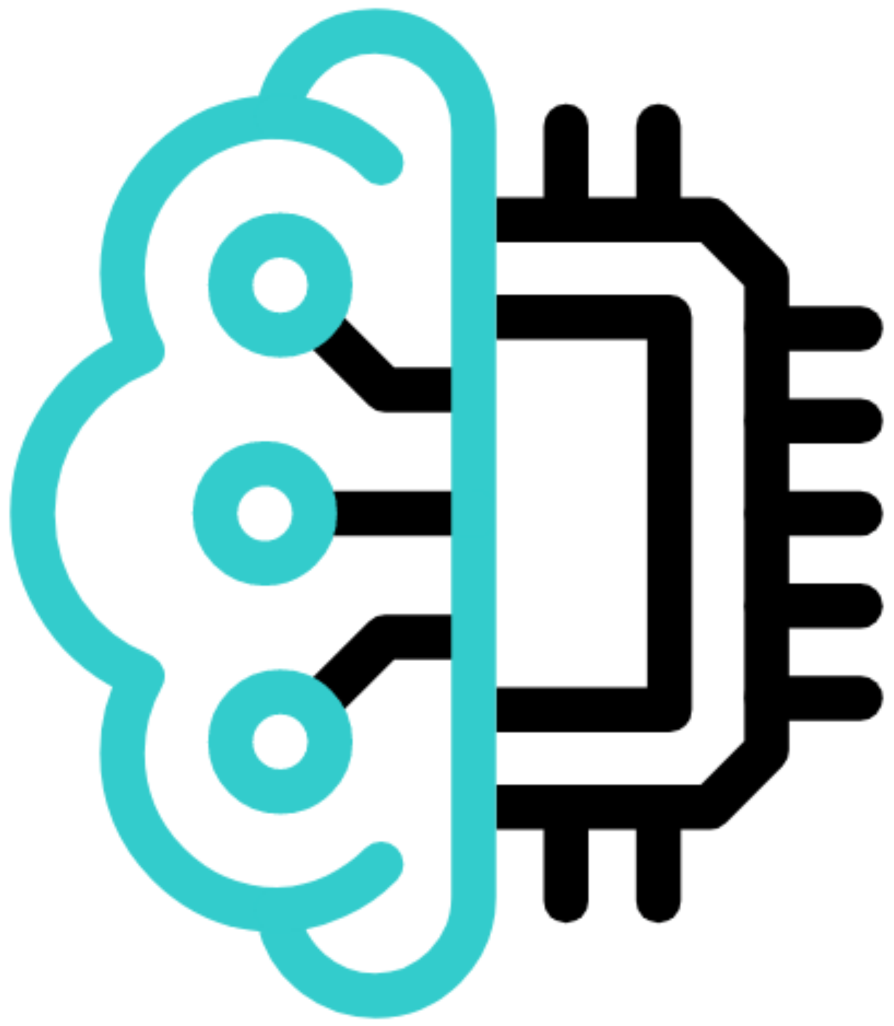
Challenges

- Scalability: Ensuring AI methods can handle the exponential growth in data volume.
- Data Quality: Dealing with noisy, incomplete, or biased astronomical datasets.
- Computational Resources: Making AI accessible for institutions with varying computational capabilities.

Hands-on Sessions

- Wednesday: **Membership determination in open clusters using DBSCAN**

- Thursday: **Photometric redshift using Decision Trees, Random Forest and Neural Networks**



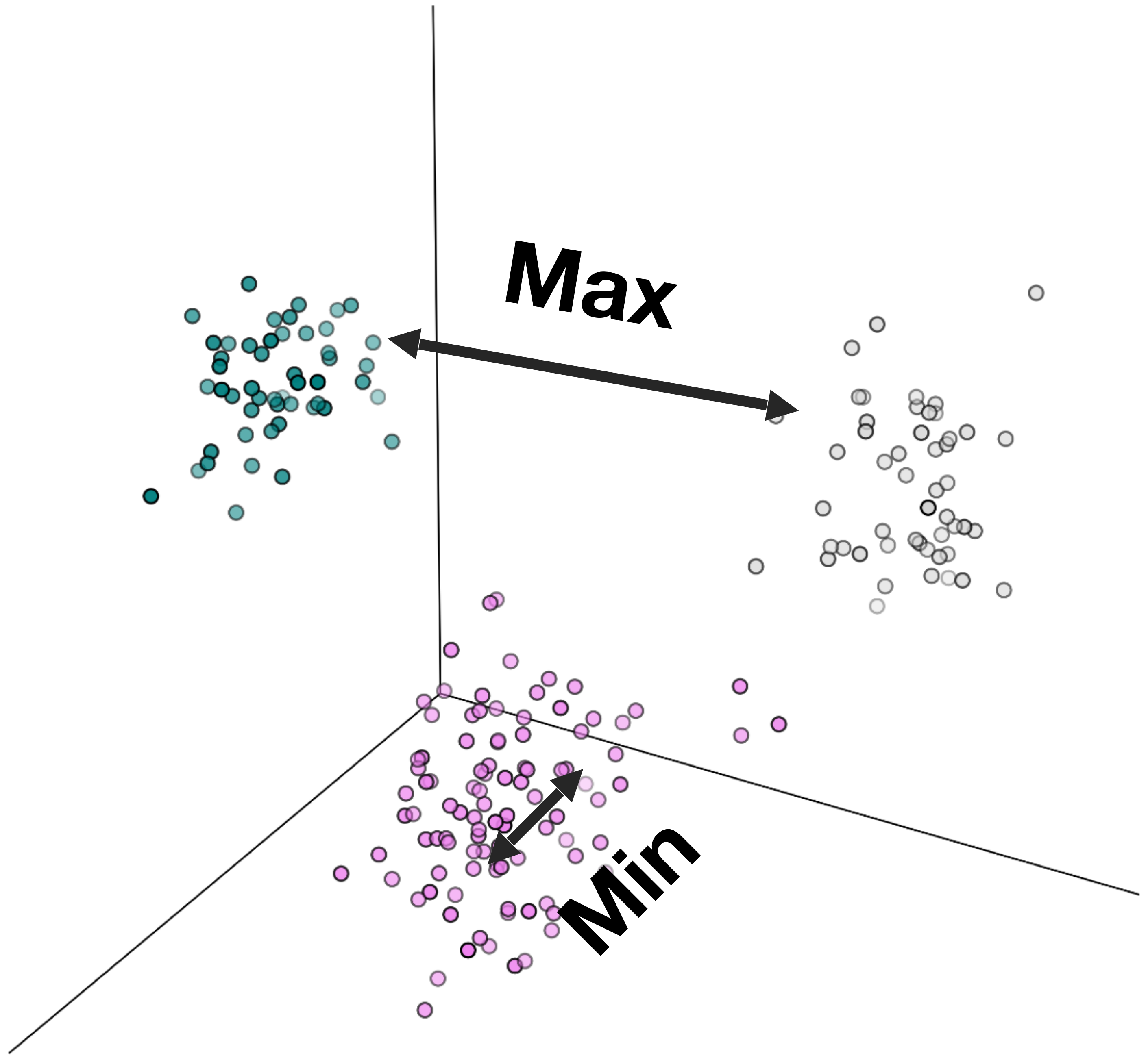
Membership determination in open clusters using DBSCAN

Basic concepts

Clustering

Clustering is the grouping of objects into a "cluster" such that they are similar (or related) to each other and different (or unrelated) from objects in other clusters.

A successful clustering scheme is one where the distances between clusters are large, and the distances within a cluster are small.



Density-based clustering

Density-based clustering algorithms, such as **DBSCAN**, identify clusters by finding **areas of higher density** in the data. This allows them to work with **arbitrarily shaped clusters** and automatically determine the number of clusters.

The operation of DBSCAN is controlled by hyperparameters: the proximity threshold that defines cluster density (`eps`) and the minimum number of samples in a cluster (`min_samples`). Finding the optimal values for these hyperparameters is challenging (similar to finding the optimal k in K-means) because tuning hyperparameters in unsupervised algorithms is not straightforward.

This method is particularly useful for identifying **outliers**.

Nice visualization in https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

DBSCAN pseudocode

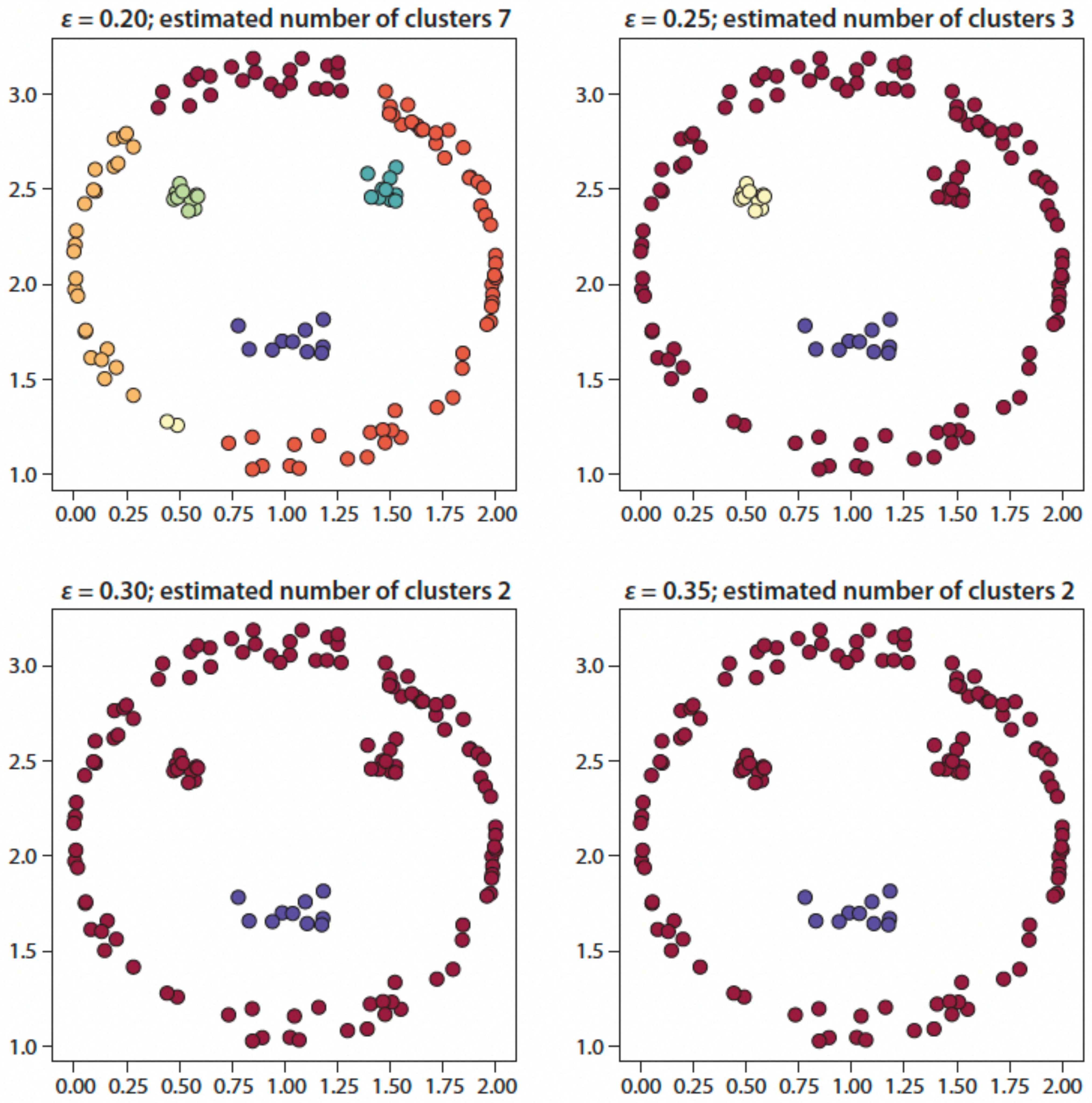
For each *unassigned* example x_i:

- Check whether there are at least $n_{\min}$ points within a distance of ϵ (that is, whether the sample is a core sample);
- If yes, implement the "expand the cluster" sequence.

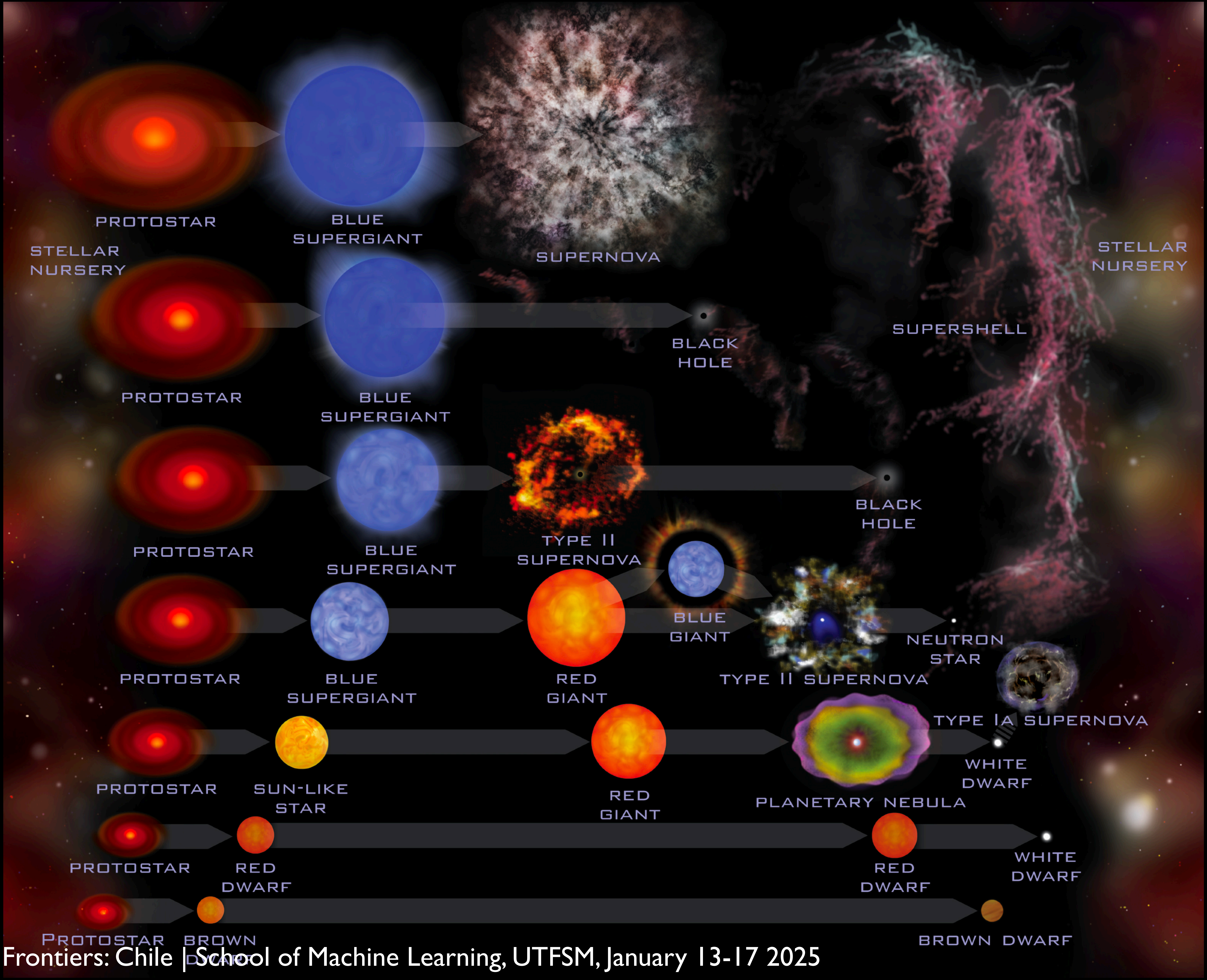
"Expand the cluster" sequence:

- Assign all samples within distance ϵ of the current core sample to cluster;
- For each newly assigned neighbor x_j that is a core point, implement the "expand the cluster" sequence around x_j.

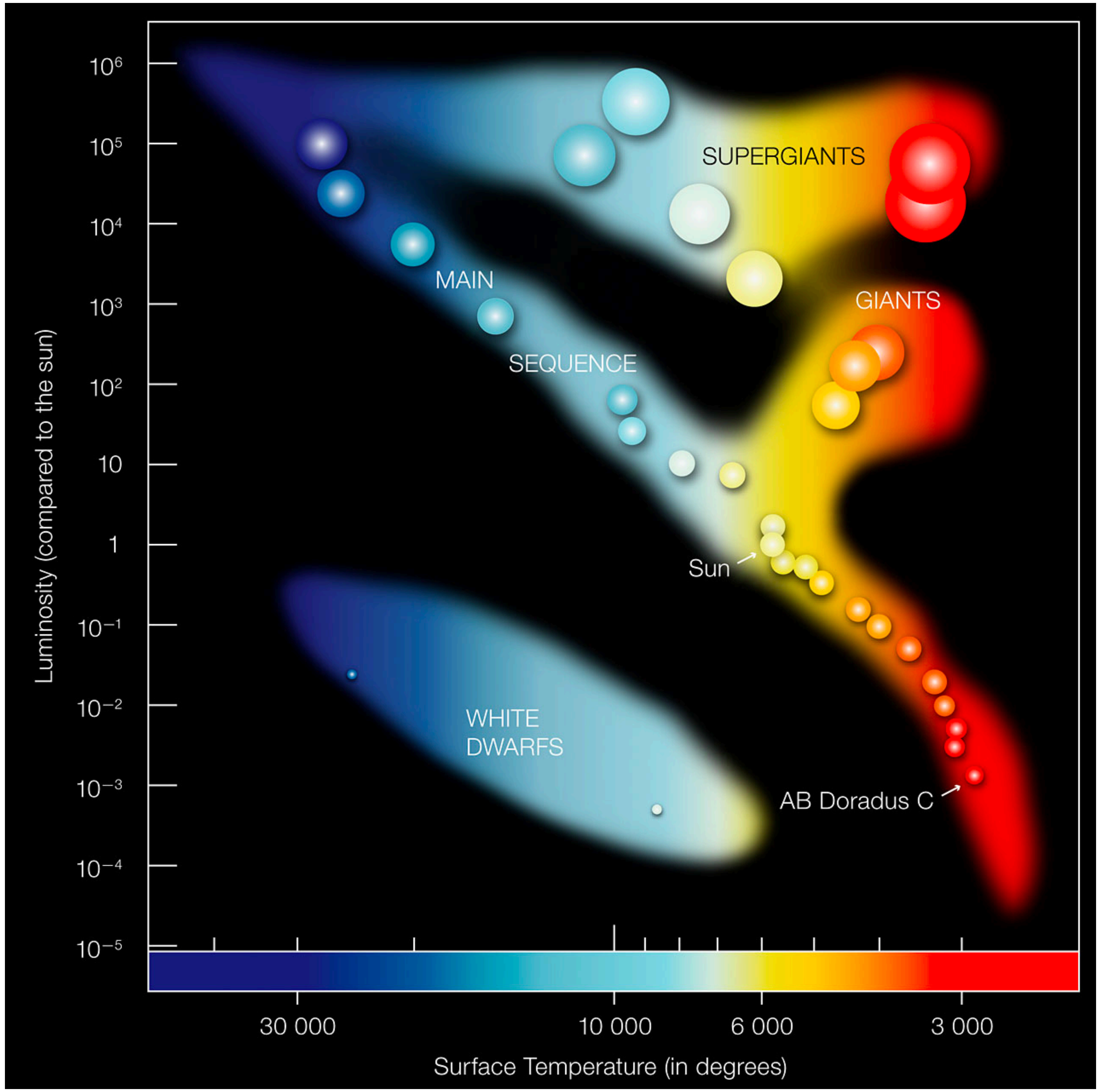
DBSCAN example

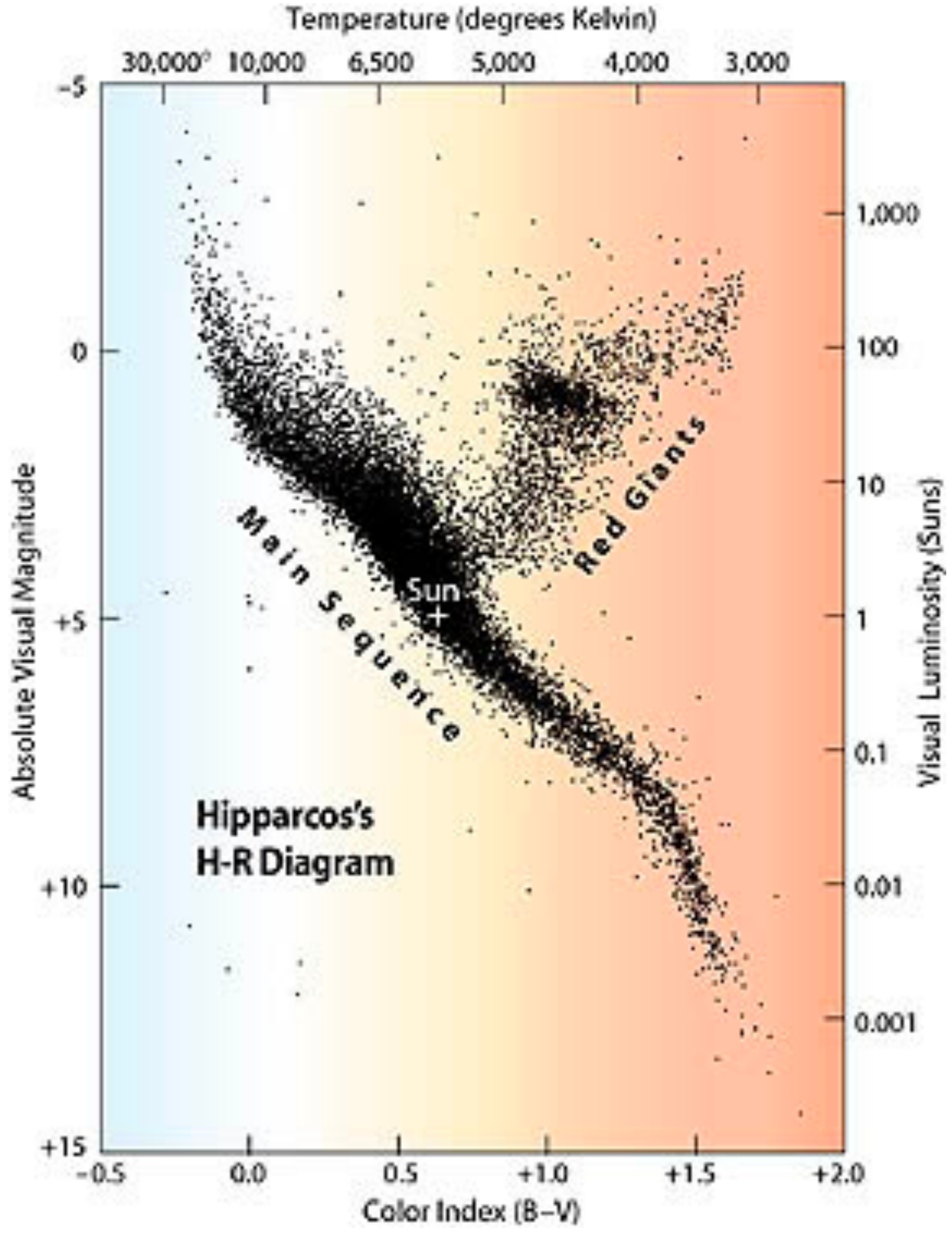


Stellar evolution in one slide!

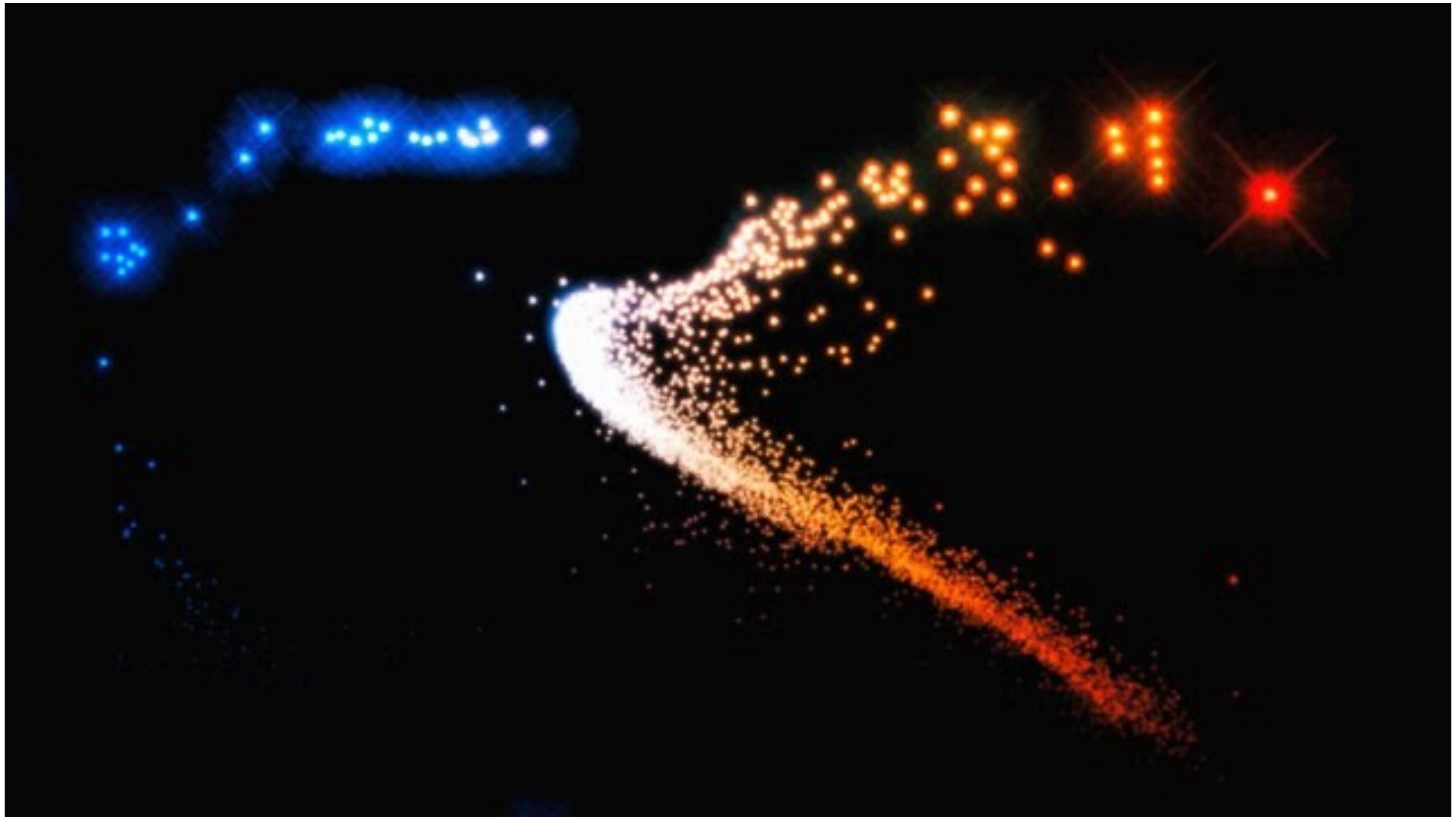


Herztprung-Russell and Color-Magnitude diagram



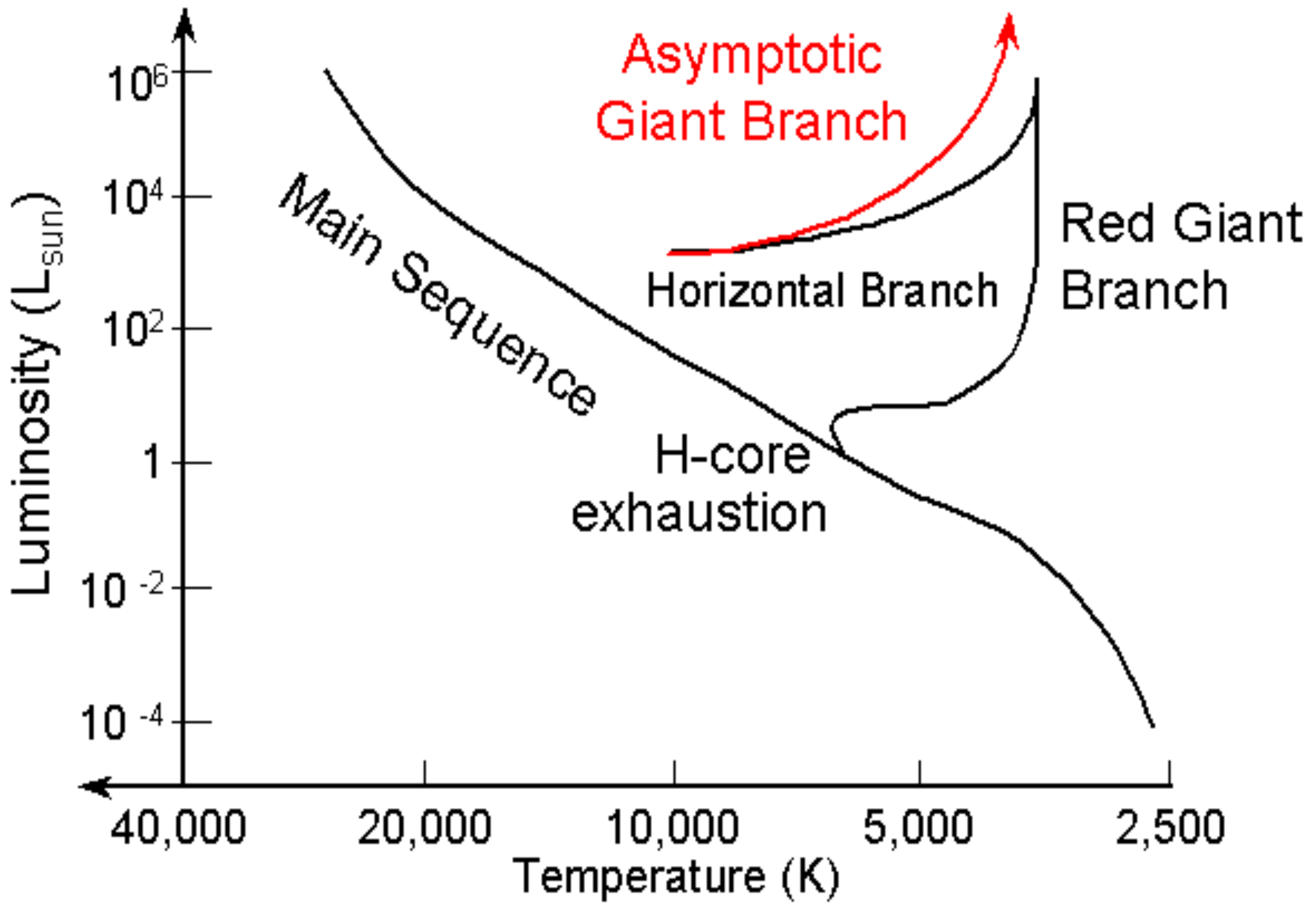


Herztprung-Russell and Color-Magnitude diagram



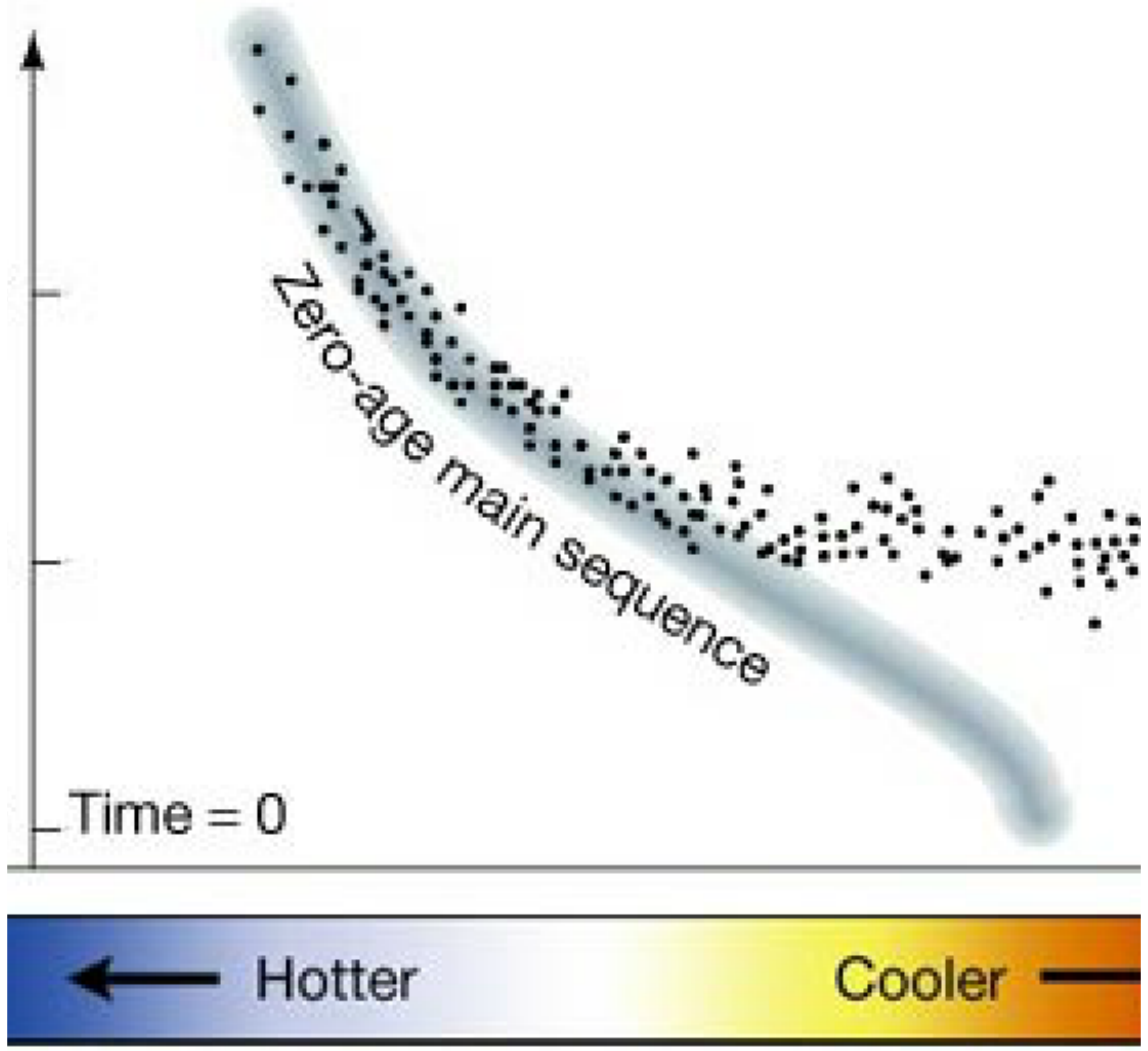
Stellar evolution and the HR diagram

- Main sequence (MS): Core hydrogen burning phase. Longest phase of evolution
- Turn-Off: Hydrogen exhausted in core.
- Red Giant Branch (RGB): Hydrogen Burning in shell around inert helium core.
- RGB tip: end of the RGB
- HB (RC): Helium burning in the core (details depends on the mass loss)
- Asymptotic Giant Branch (AGB): He burning in shell around an inert C/O core. Complicated mass dependent evolution from now on.



Stellar evolution and the HR diagram

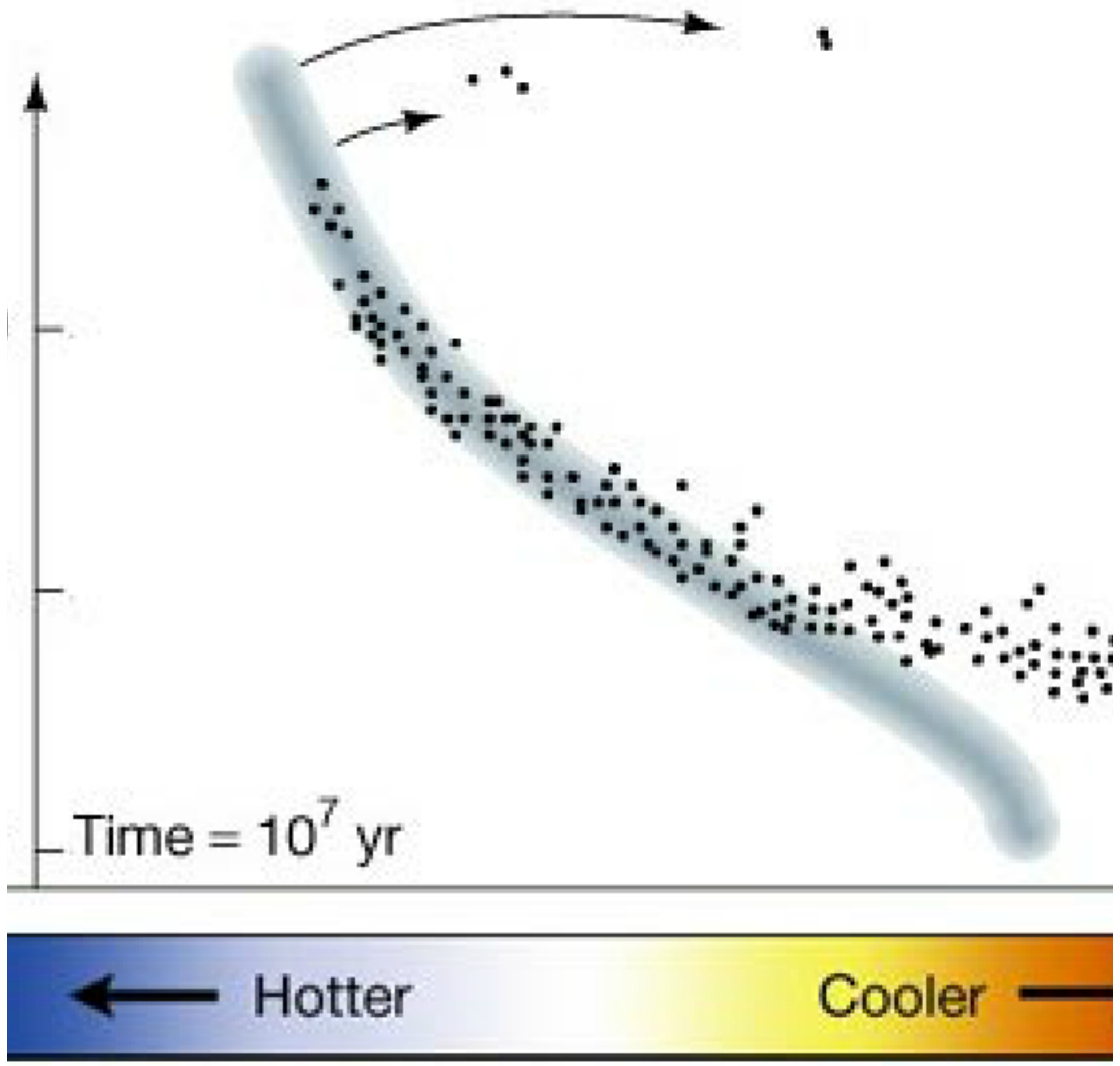
$$T = 0$$



Most of the massive stars are in the MS, while low-mass stars are in the T-Tauri stage

Stellar evolution and the HR diagram

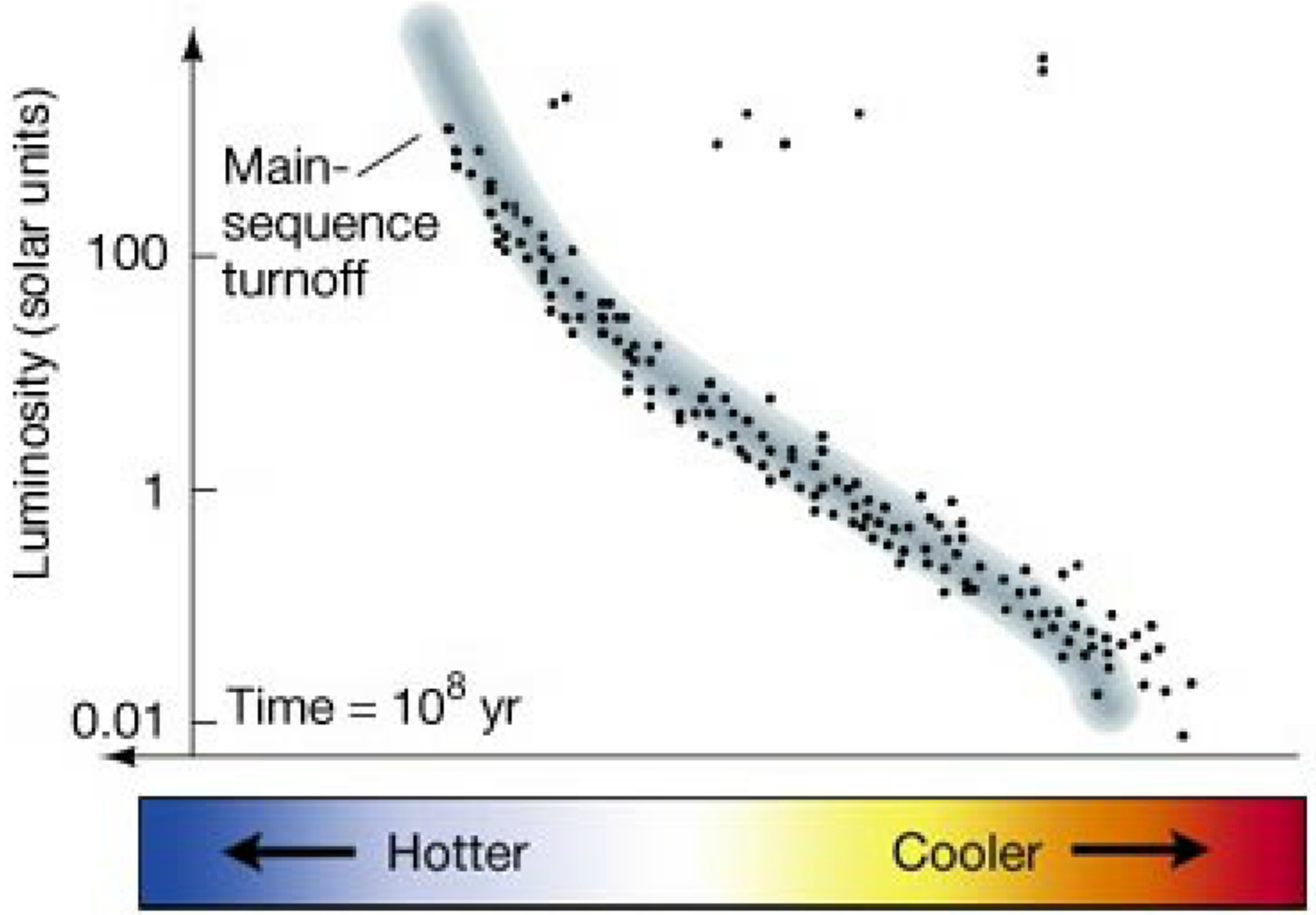
$T = 10^7$ years



O-type stars have exhausted all their hydrogen and evolve off the MS

Stellar evolution and the HR diagram

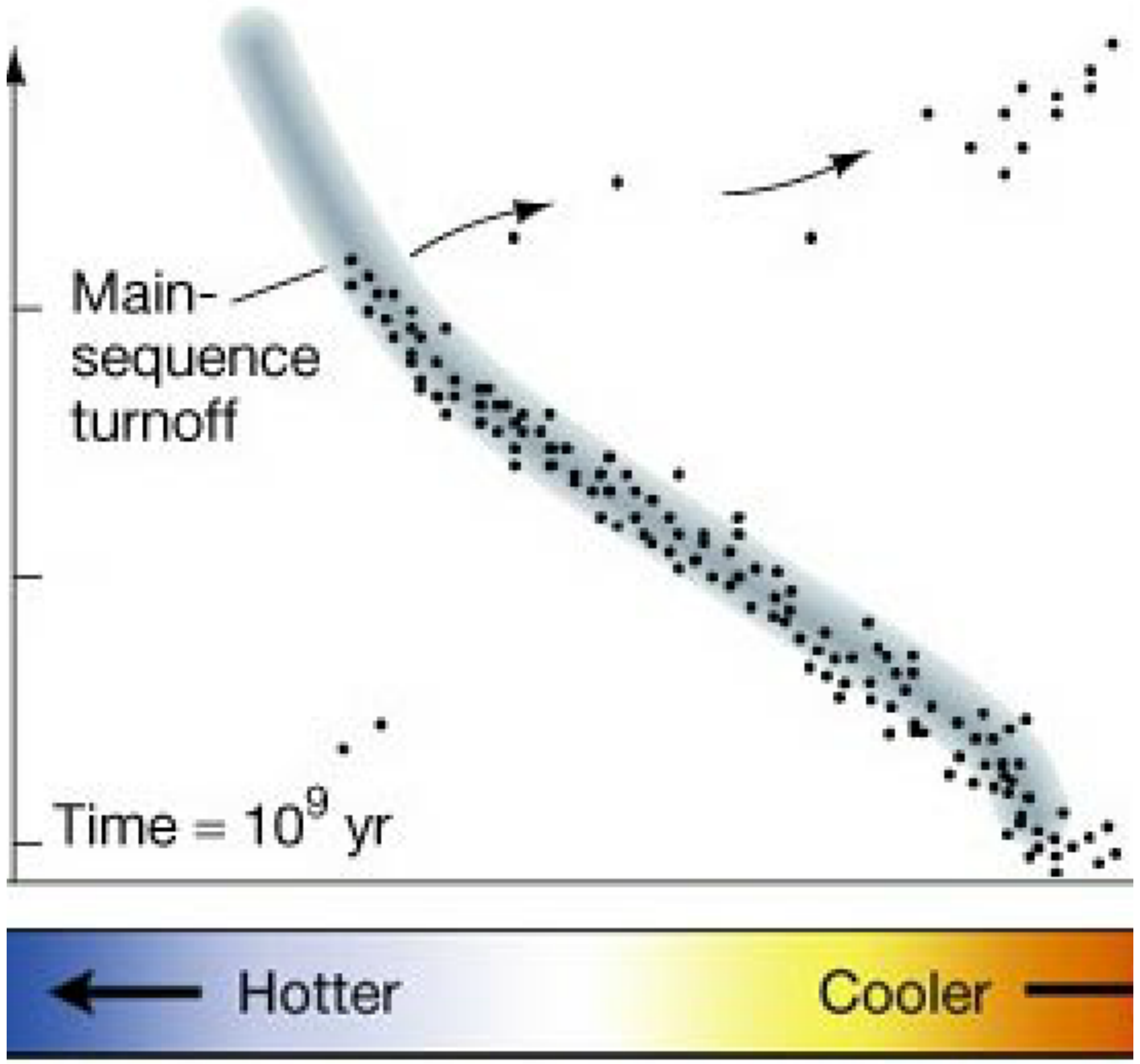
$T = 10^8$ years



O-type stars exploded as supernovae, while B-type stars evolve off the MS

Stellar evolution and the HR diagram

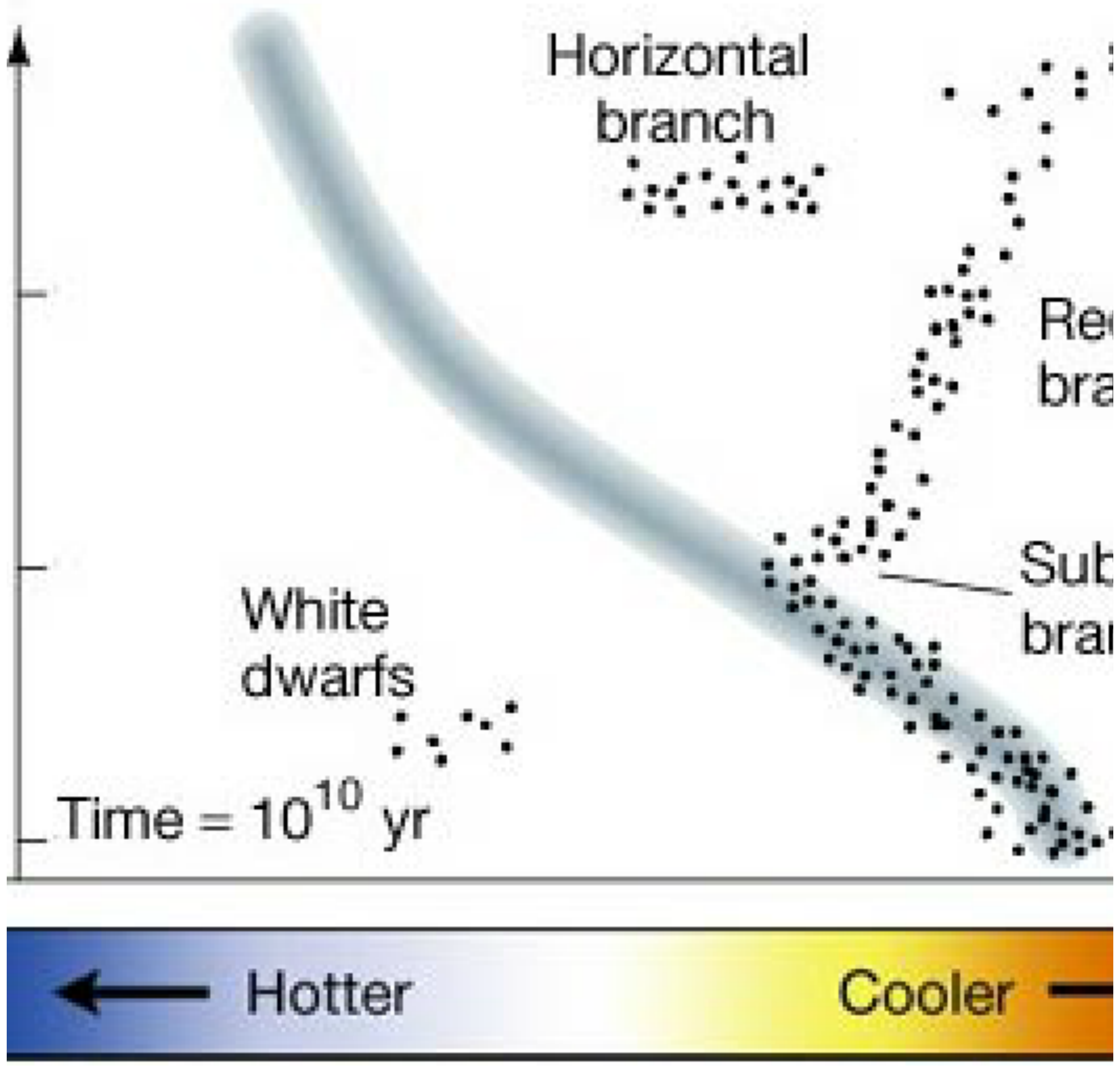
$T = 10^9$ years



B-type stars that are sufficiently massive explode as supernovae, while the rest evolve into red giants. A-type stars begin to leave the MS

Stellar evolution and the HR diagram

$T = 10^{10}$ years



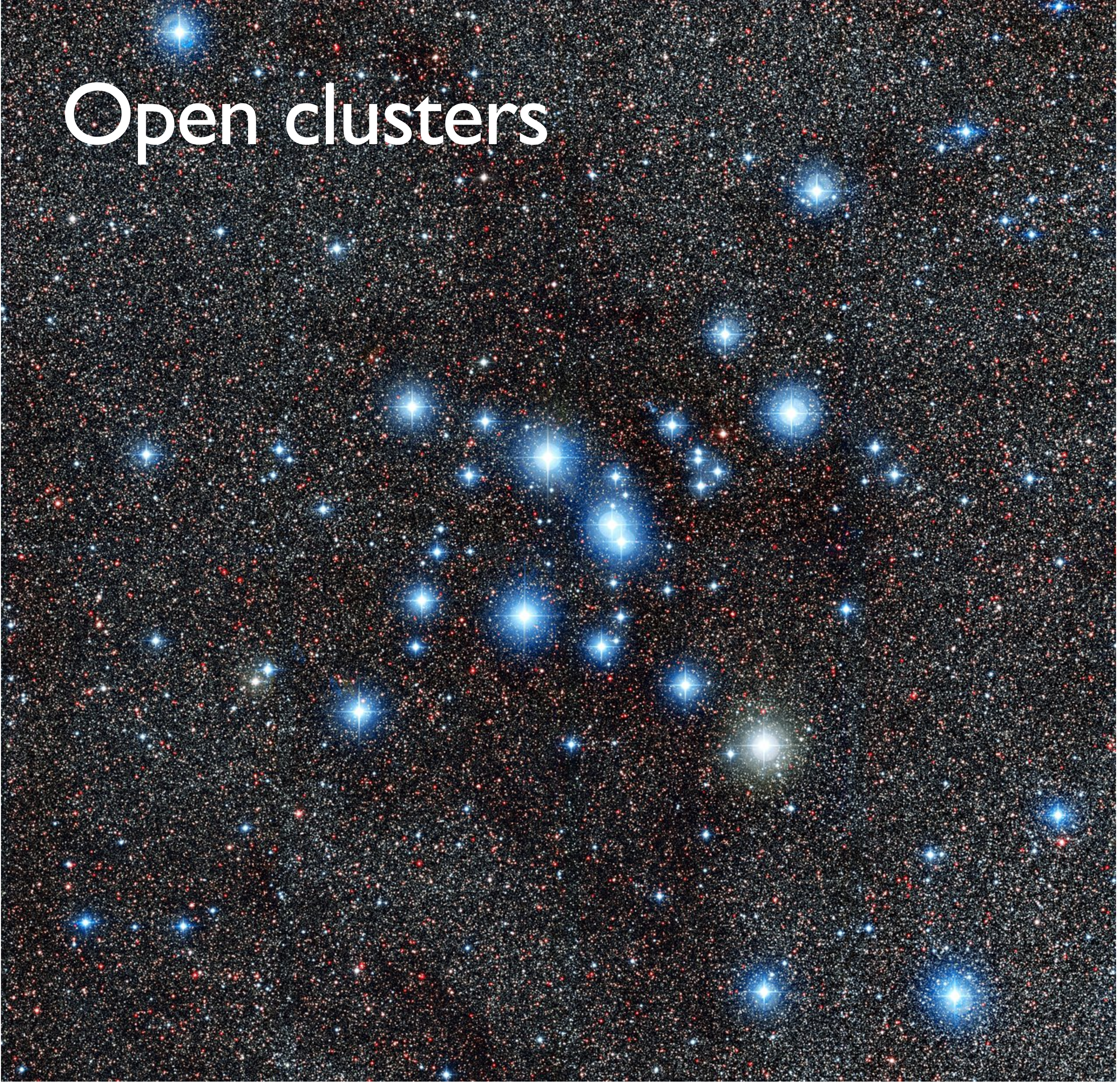
OBAFG-type stars have evolved off the MS, the giant branch is heavily populated, and there are already several white dwarfs. The MS is primarily composed of K and M-type stars

CMD of star clusters

A star cluster is crucial for understanding stellar evolution because historically, they are considered **simple stellar populations (SSPs):**

- All stars form at the same time **(same age)**.
- All stars have the **same composition.**
- All stars are at the **same distance.**

Open clusters



Density: $0.1 - 10^2$ stars pc^{-3}

Core radii: ~ 2 pc

Mass: $10^2 - 10^3\ M_\odot$

Age: $0.01 - 10$ Gyr

Median age: 0.3 Gyr

Gravitationally bound

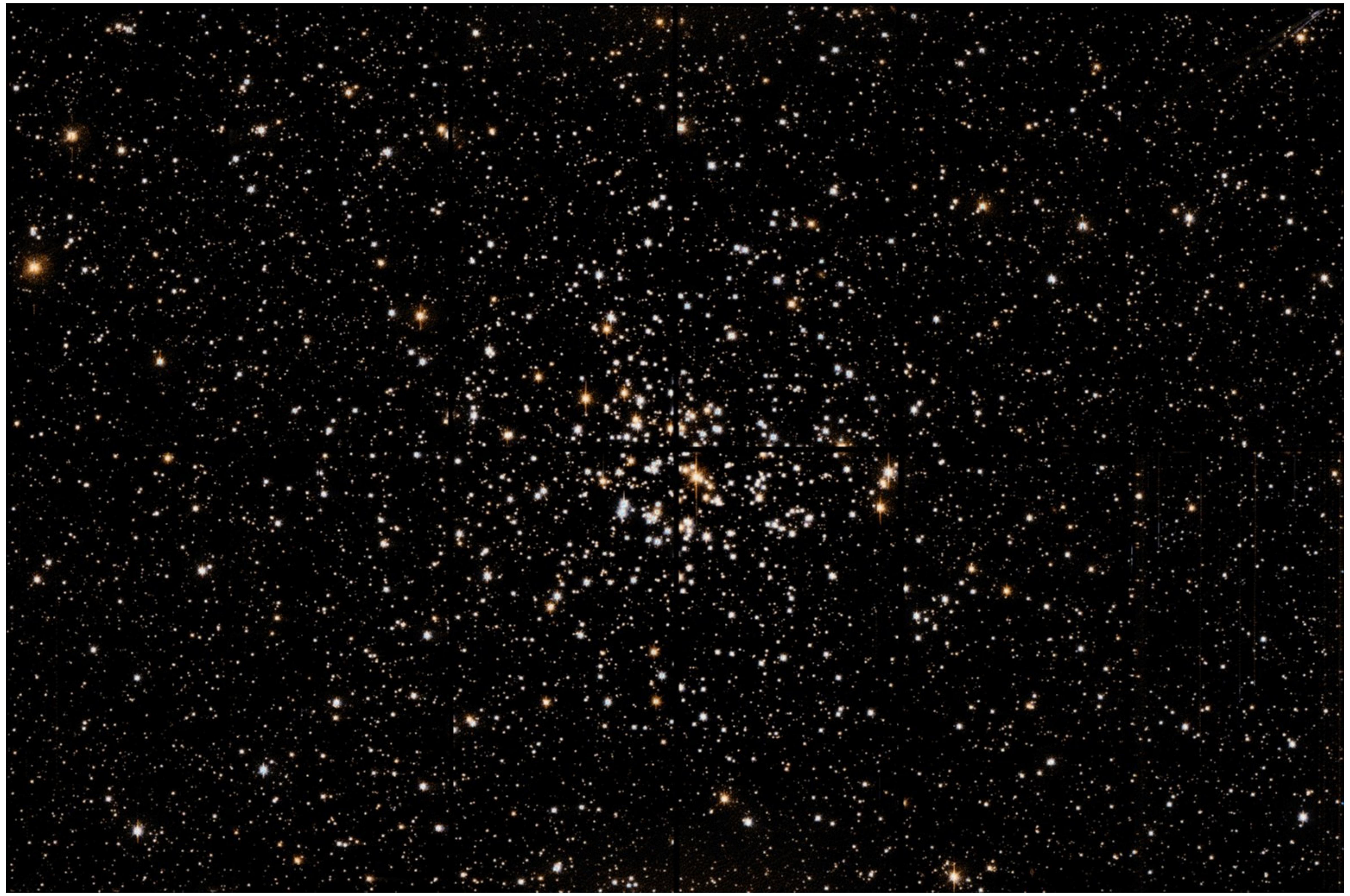
Chemically homogeneous

No gas left

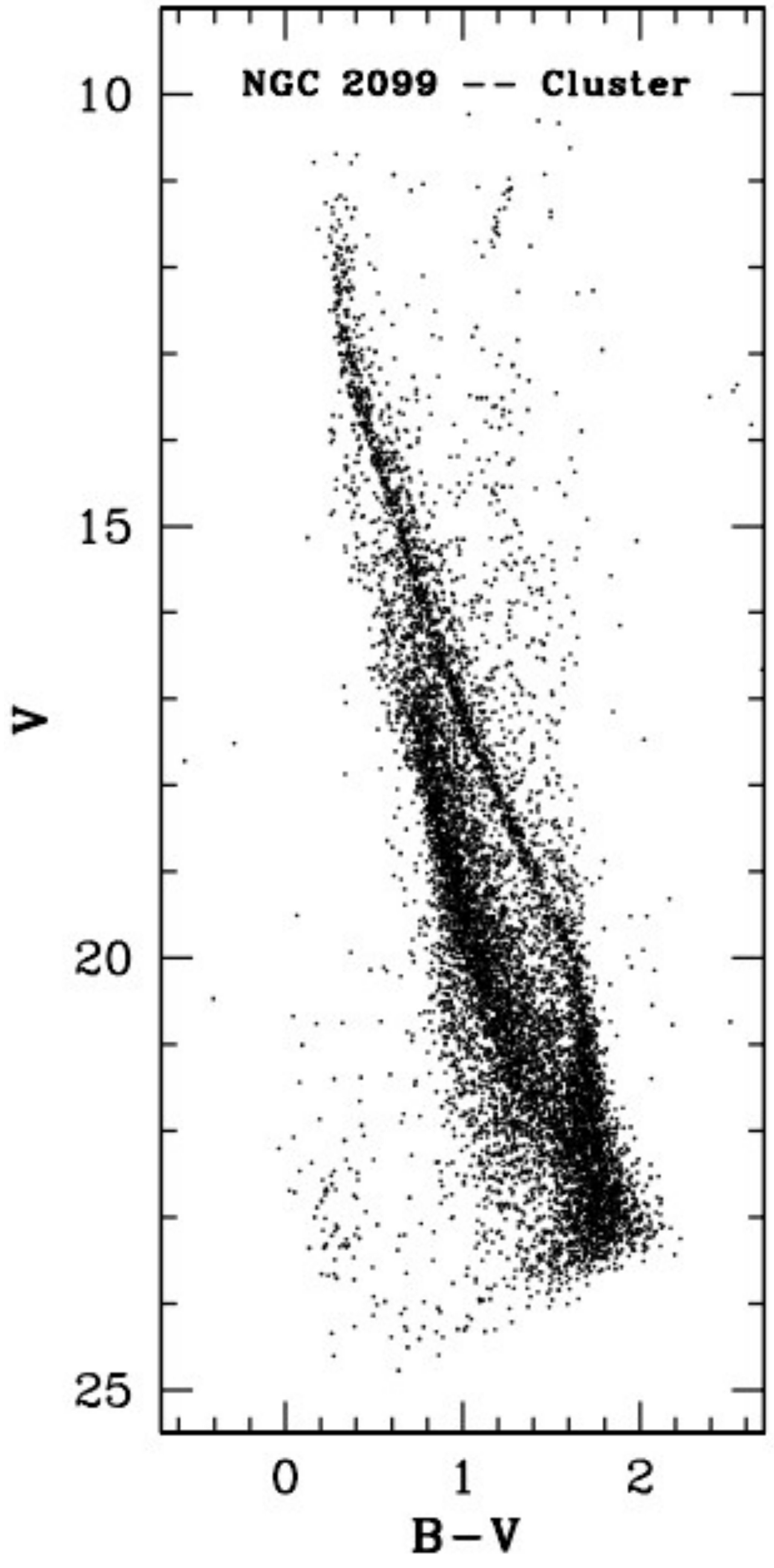
Almost coeval

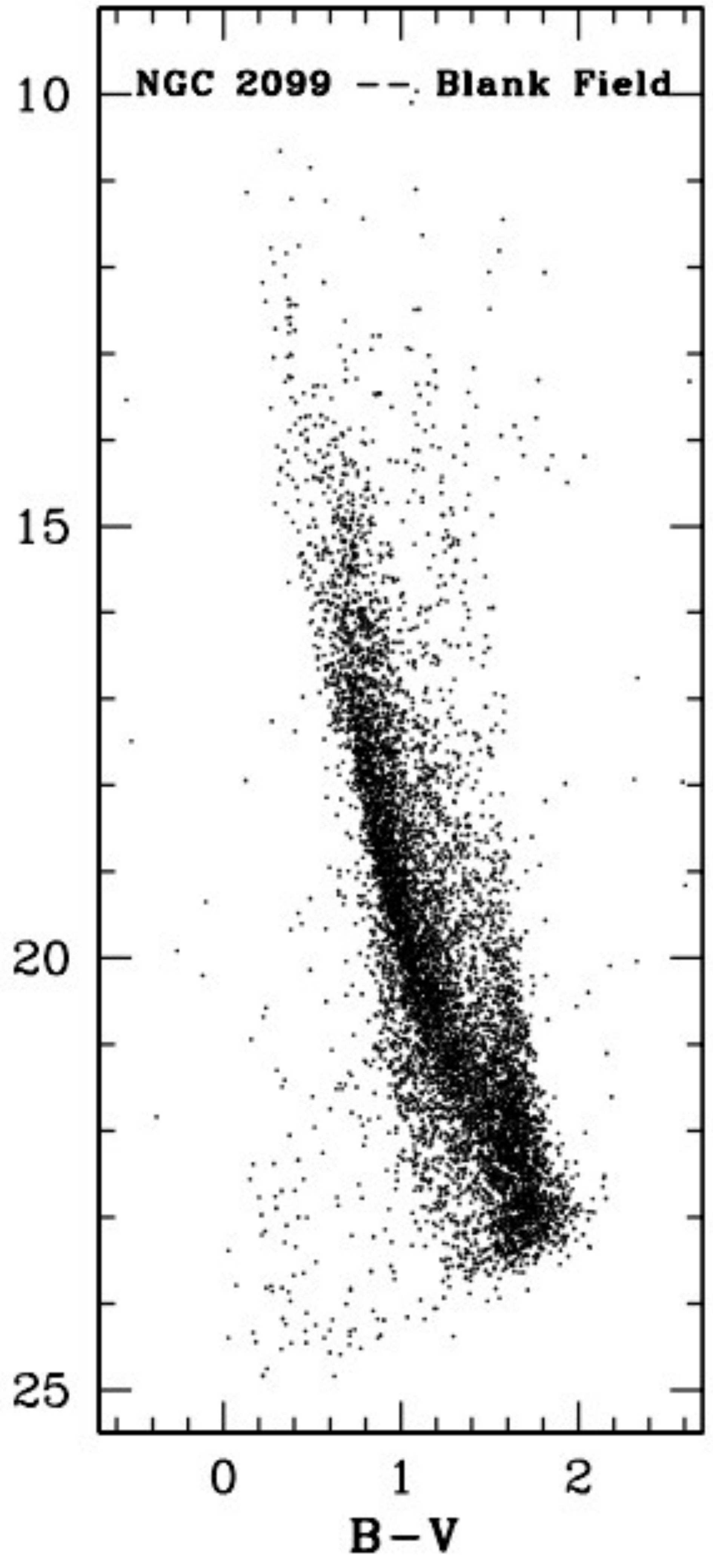
Location: Galactic disk

Open clusters



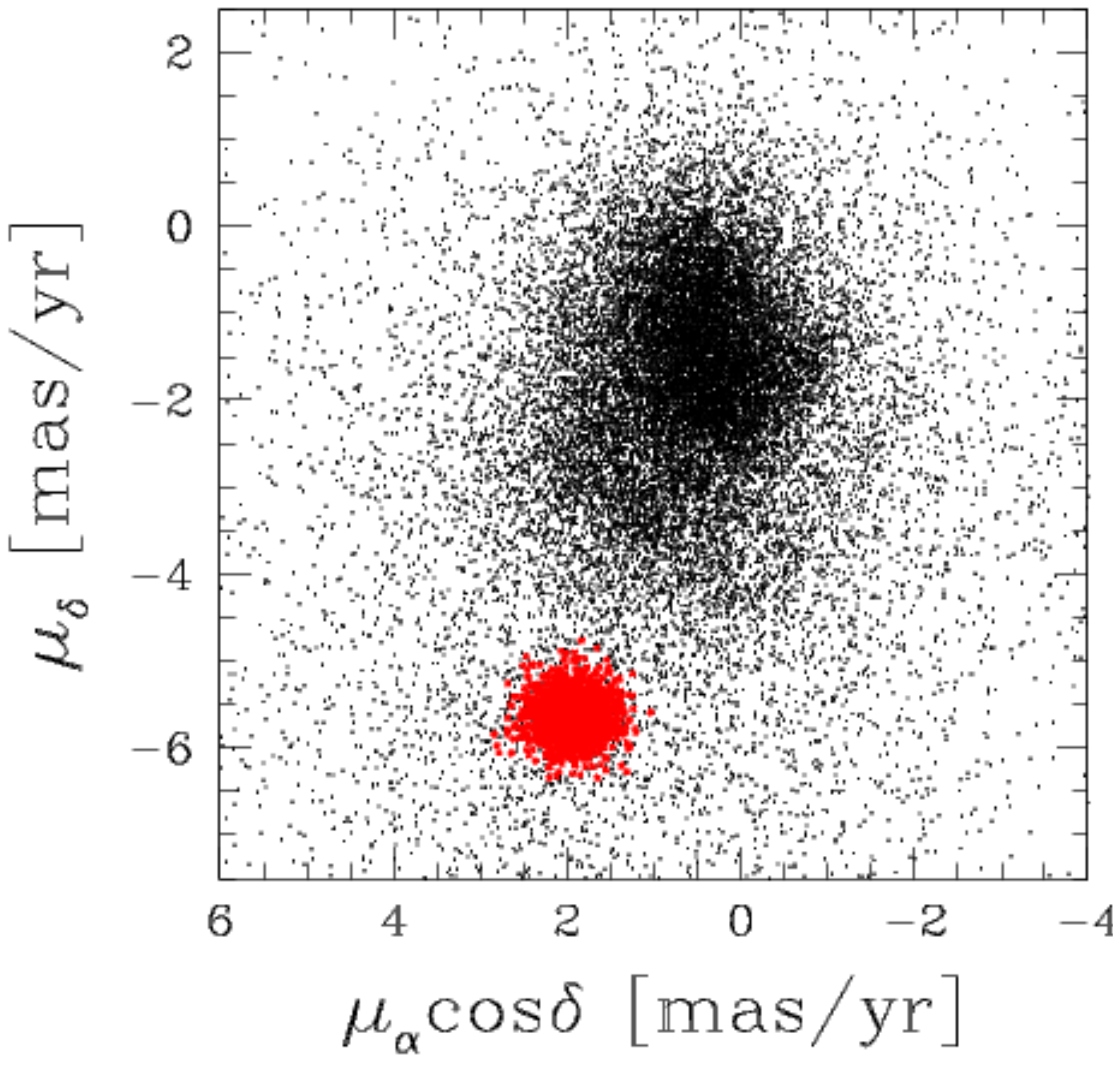
NGC2099
Age ~ 500 Myr

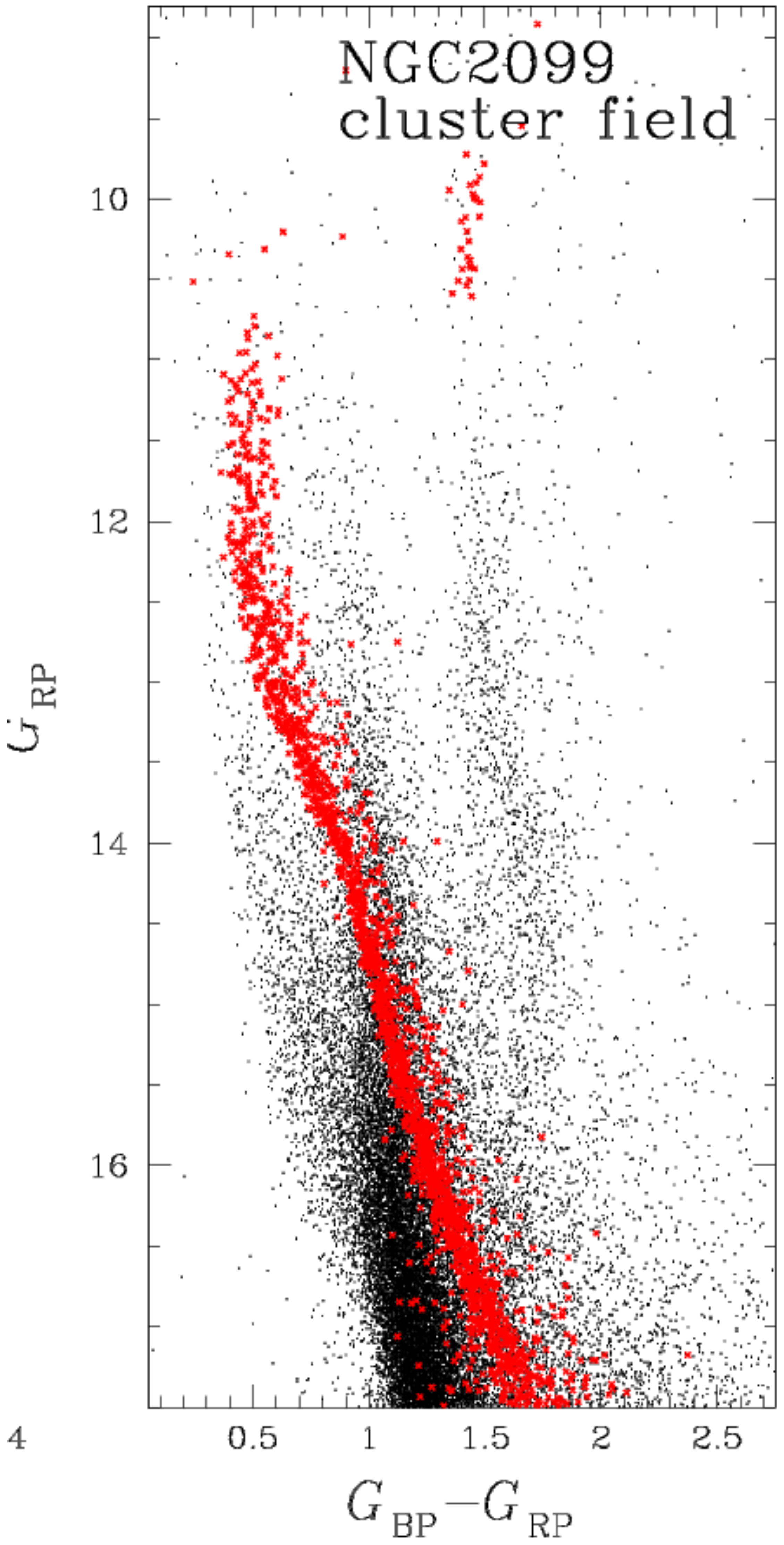


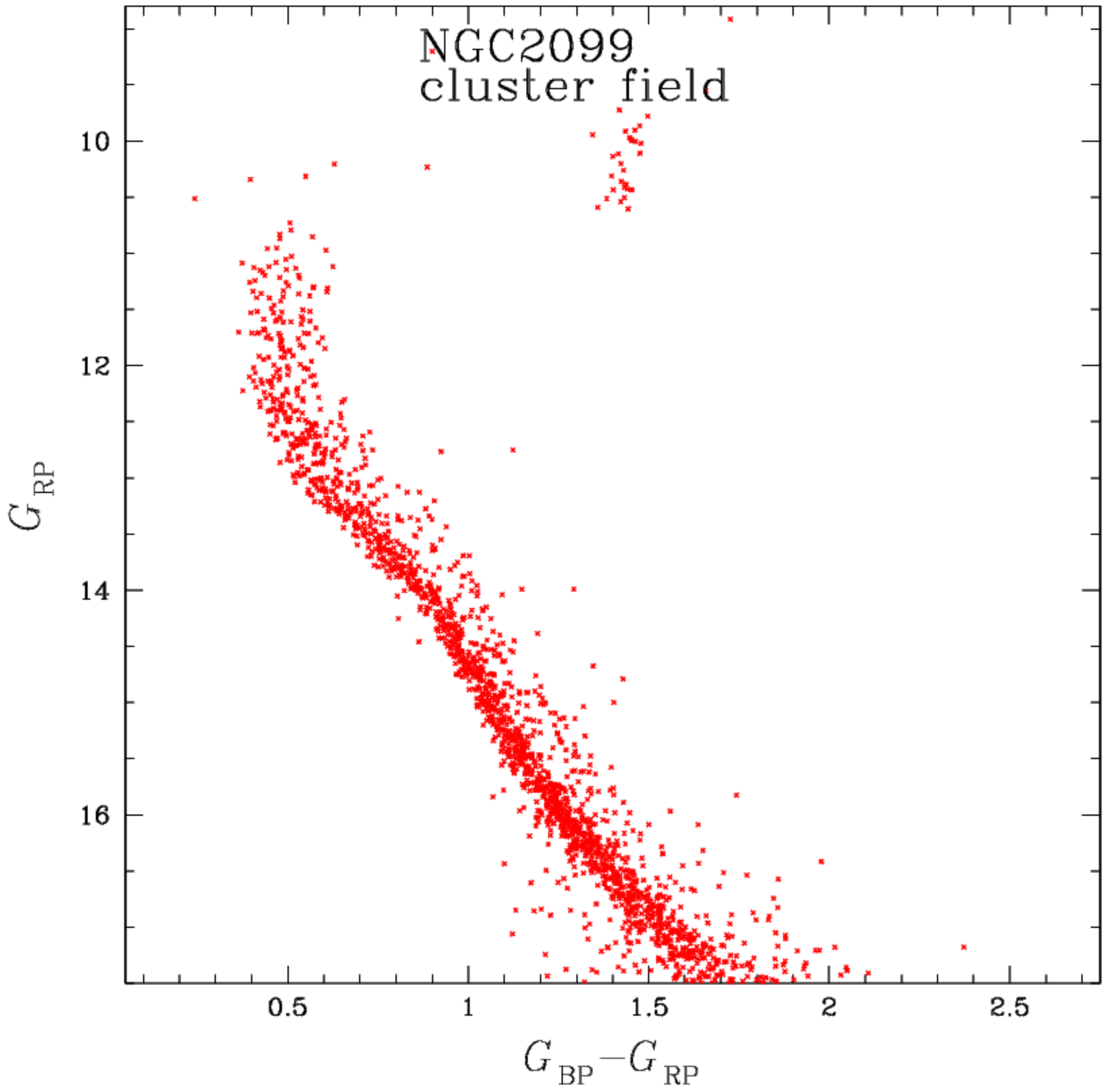


Kalirai et. al. (2001)

Open cluster membership



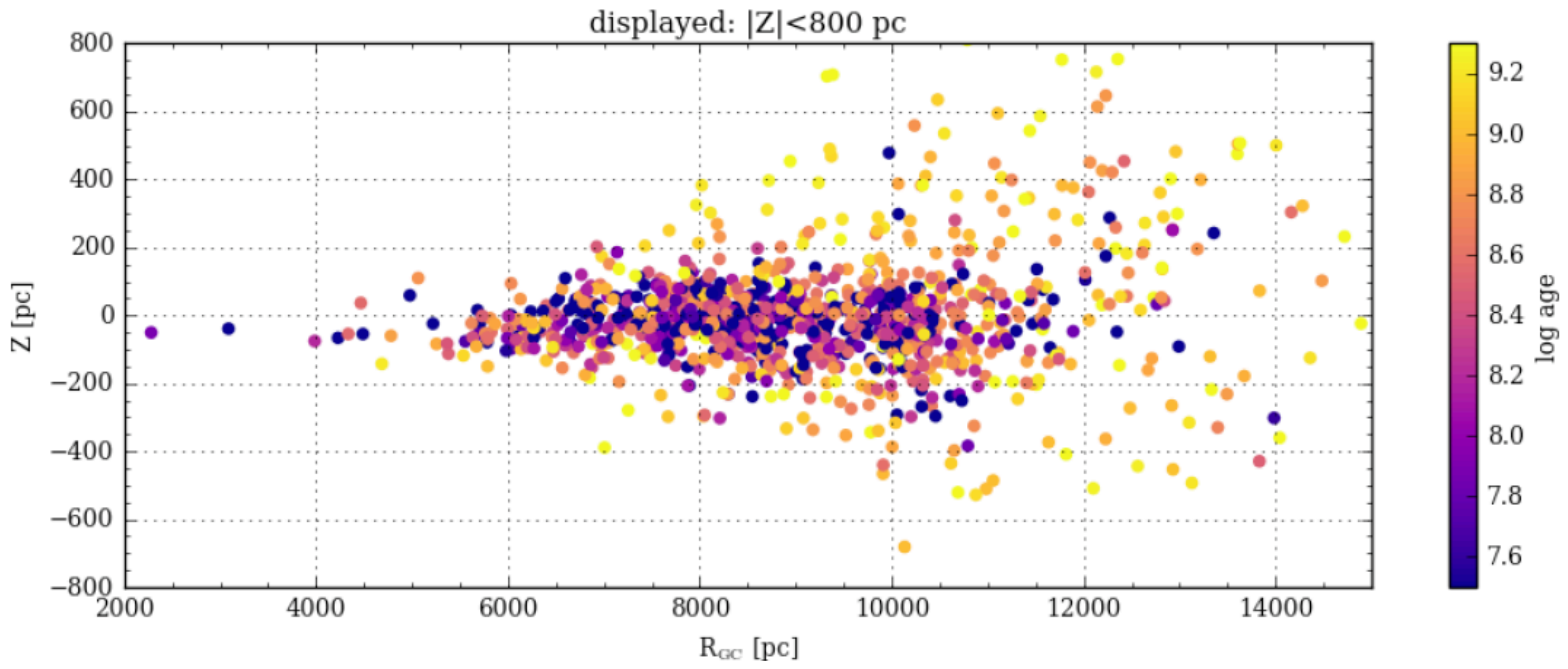


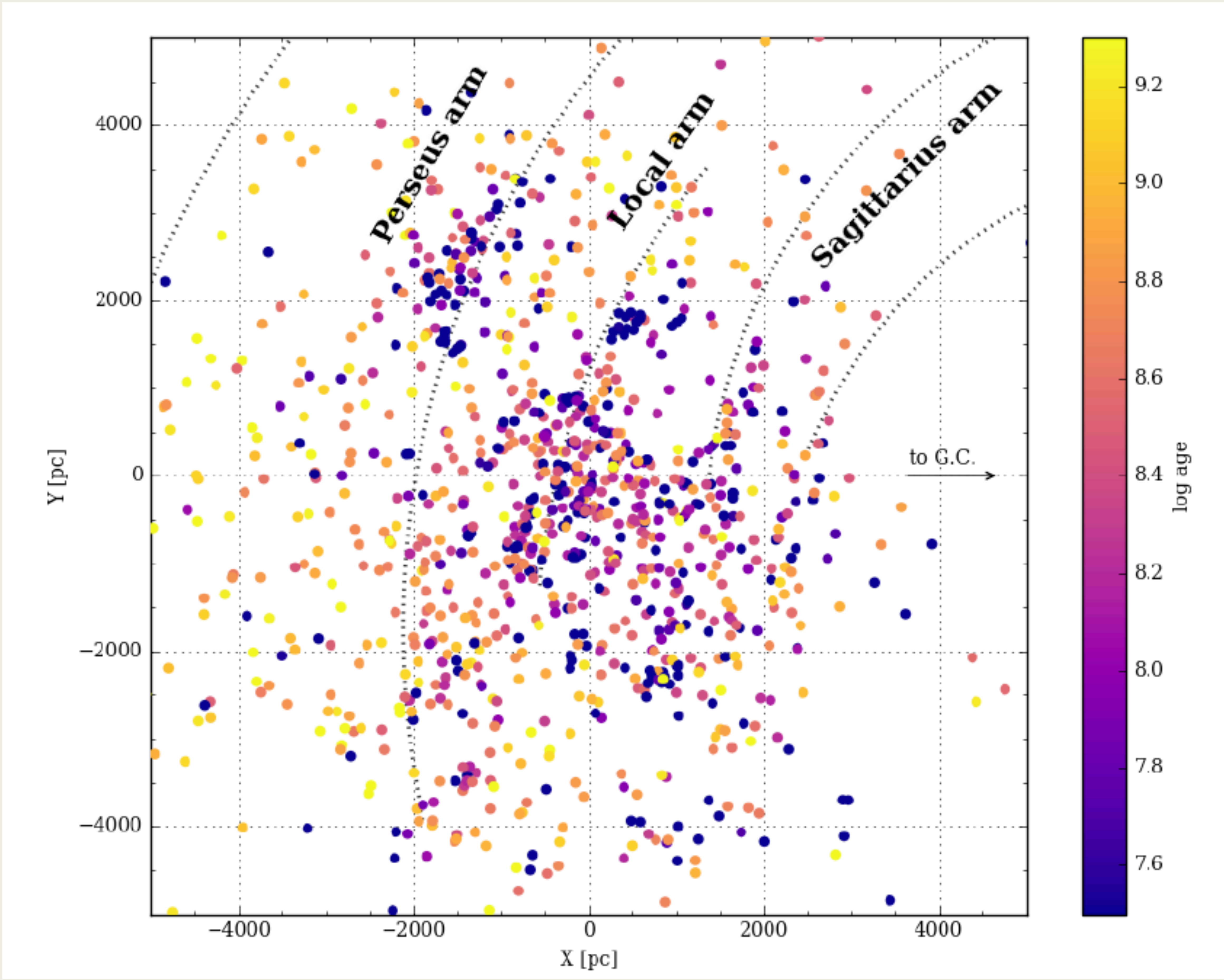


Cordoni et. al. (2018)

Open clusters

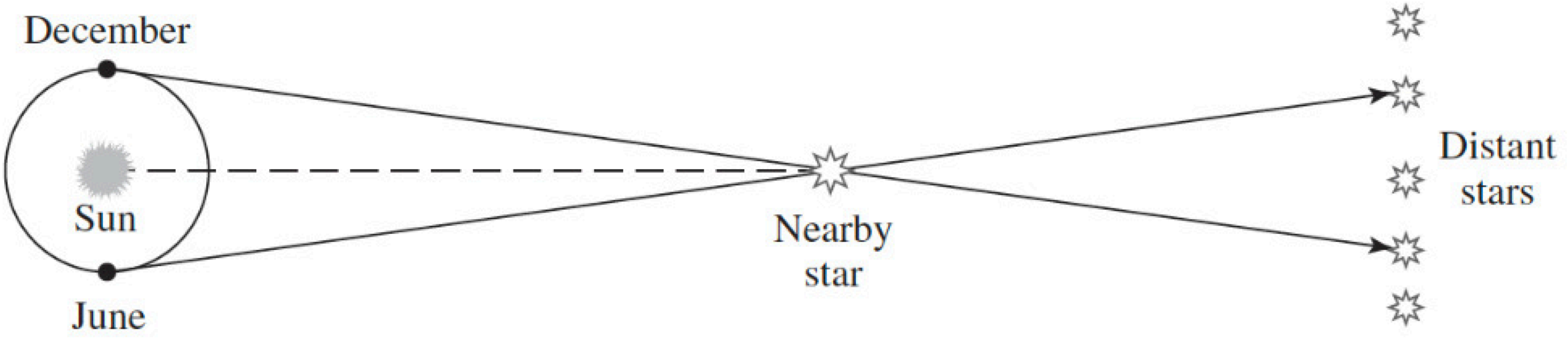
~2000 identified





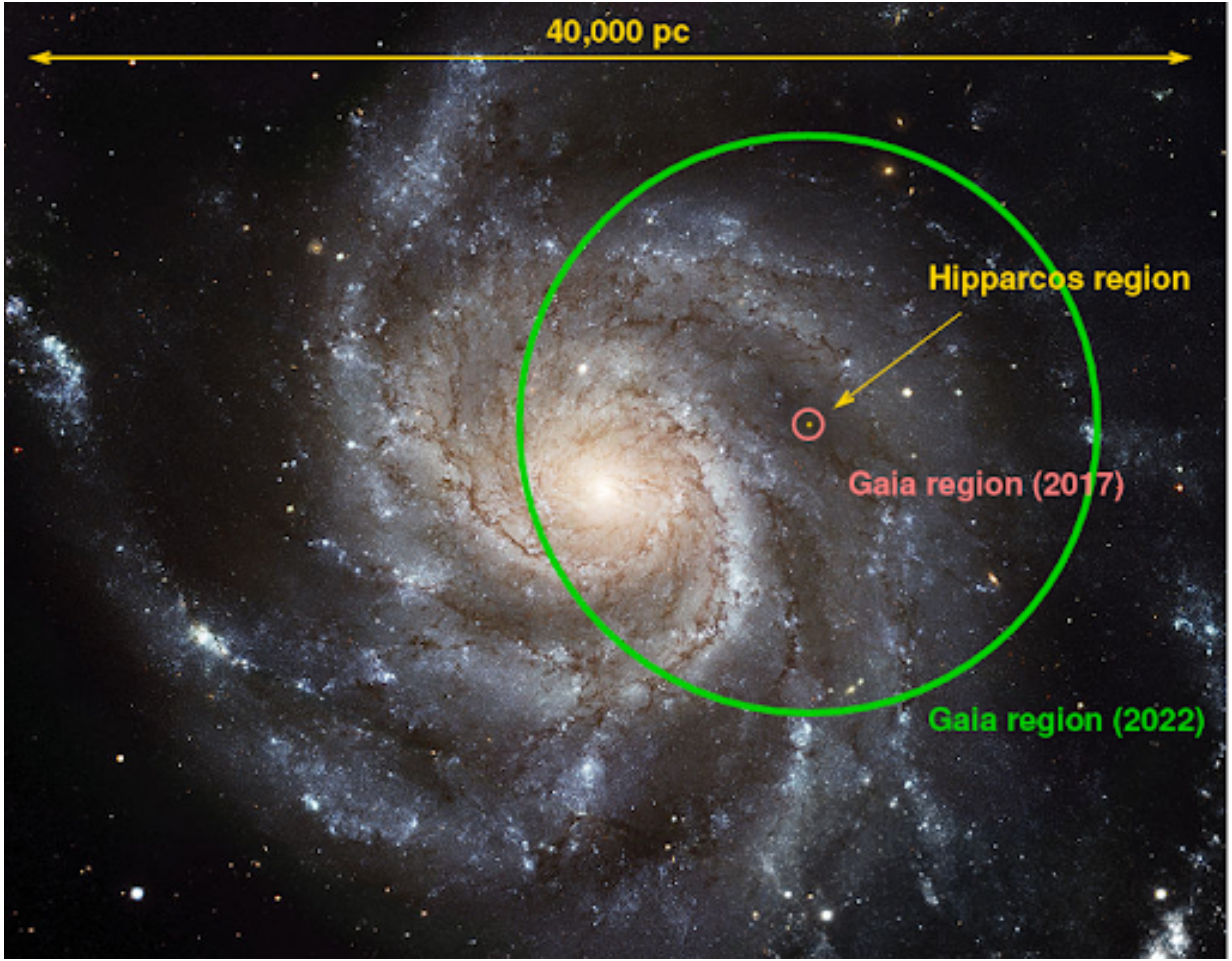
Cantat-Gaudin et al. (2018)

Parallax



$$d_*[\text{parsecs}] = \frac{1}{p('')}$$

Parallax


40,000 pc
Hipparcos region
Gaia region (2017)
Gaia region (2022)

For the hands-on sessions

https://github.com/pcamigo/ML_HEP_school_2025/

Physics Without Frontiers: Chile

School on machine learning in physics

13-17 JANUARY 2025 | VALPARAÍSO, CHILE

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

CCTVal

Thanks!