# Machine learning in the context of large stellar surveys

**Friedrich Anders** for the GaiaUB group,
Tristan Cantat-Gaudin (now MPIA), Alfred Castro-Ginard (now Leiden),
Arman Khalatyan (AIP), Laia Casamiquela (Obs Paris-Meudon), Samir Nepal (AIP),
Anna Queiroz (now IAC), Cristina Chiappini (AIP), Rany Assaad (ex U Surrey),
Guillaume Guiglion (now MPIA)

Gaia

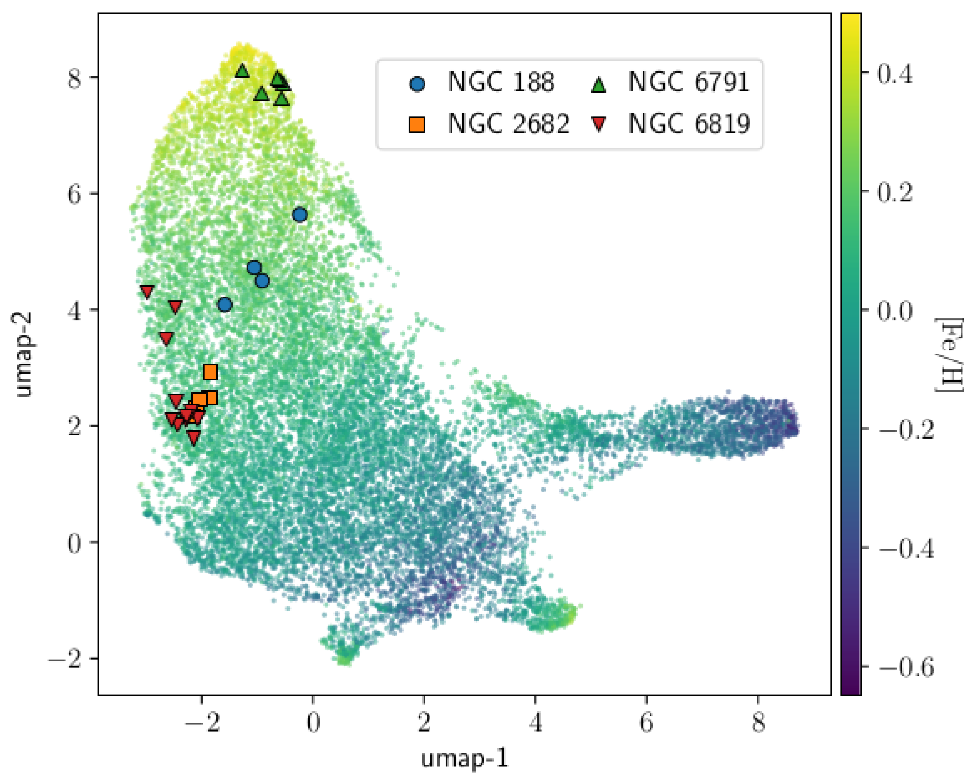
gaia

ICCUB
Institute of Cosmos Sciences

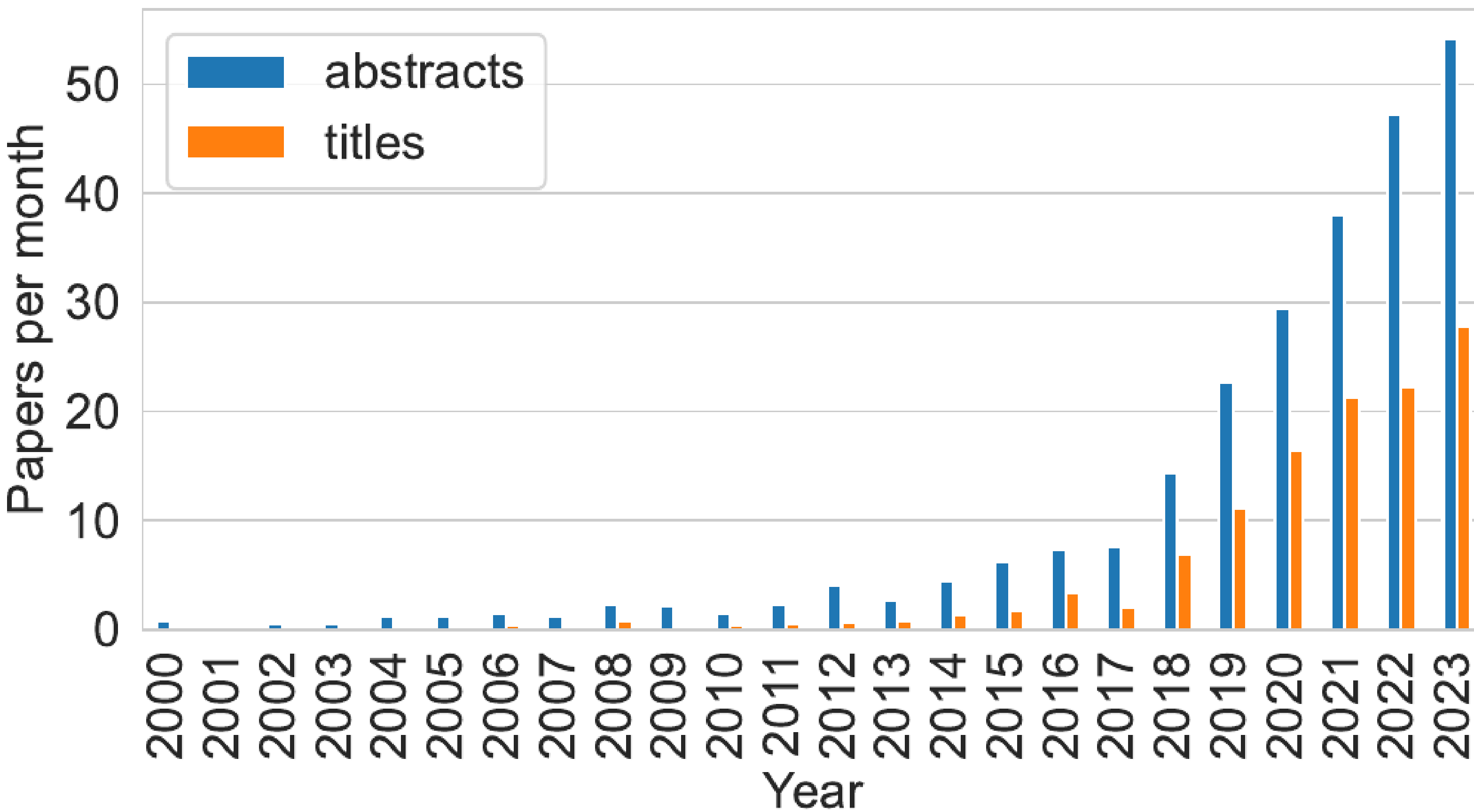
EXCELENCIA
MARÍA
DE MAEZTU

UNIVERSITAT DE BARCELONA

Marie Skłodowska-Curie Actions



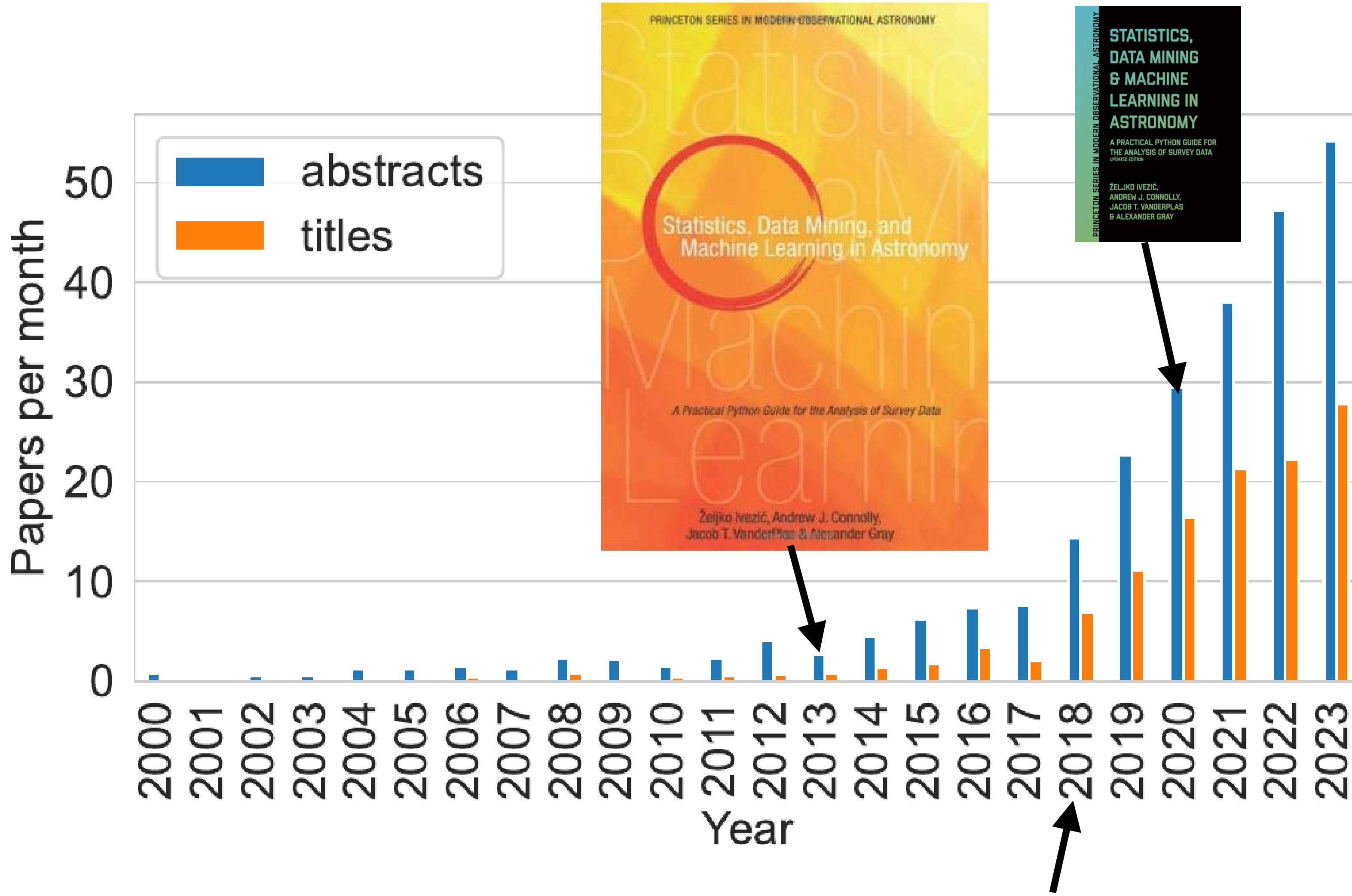
# Machine learning in astronomy




Number of refereed publications per month that include the terms "machine learning" or "artificial intelligence"

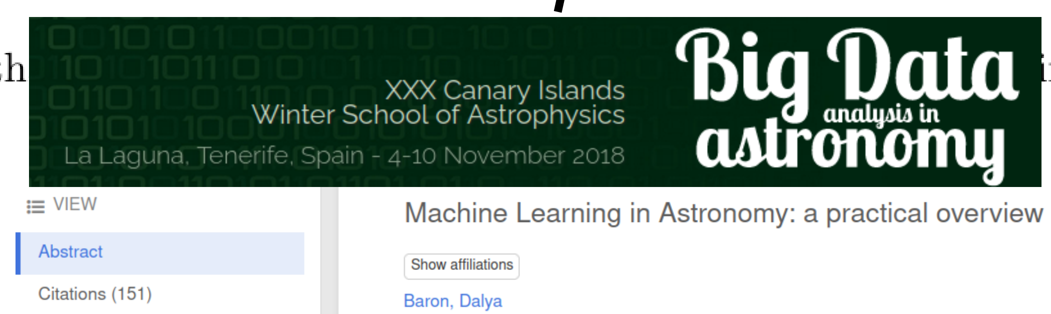
Huppenkothen+2023

# Machine learning in astronomy




Number of refereed publications per month th … intelligence”

Huppenkothen+2023



# „Machine learning in astronomy“ in reality

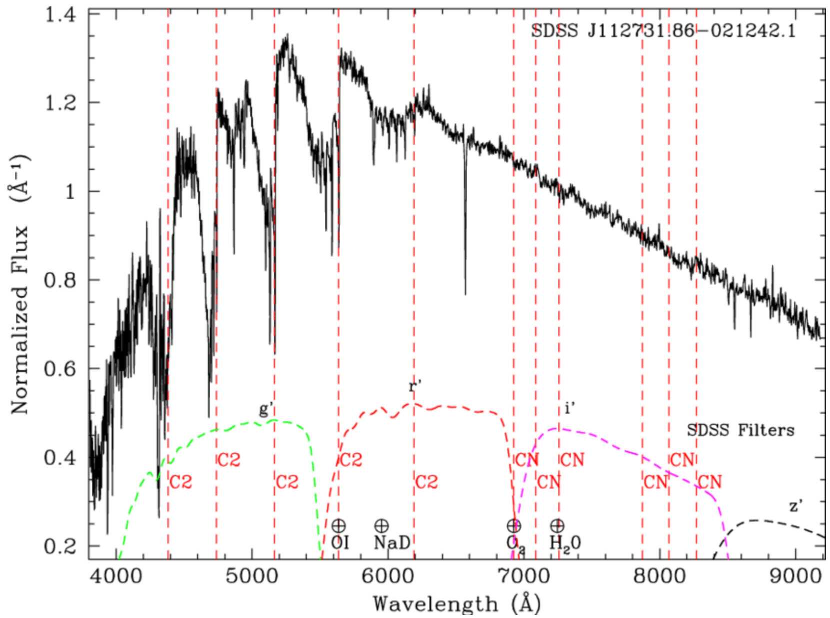
# Machine learning in astronomy: Useful references

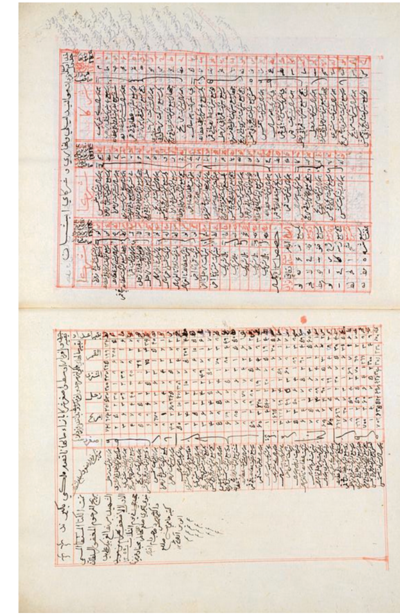
Books:
- Ivezic+2020: Statistics, Data Mining, and Machine Learning in Astronomy
- Aquaviva 2023: Machine Learning for Physics and Astronomy

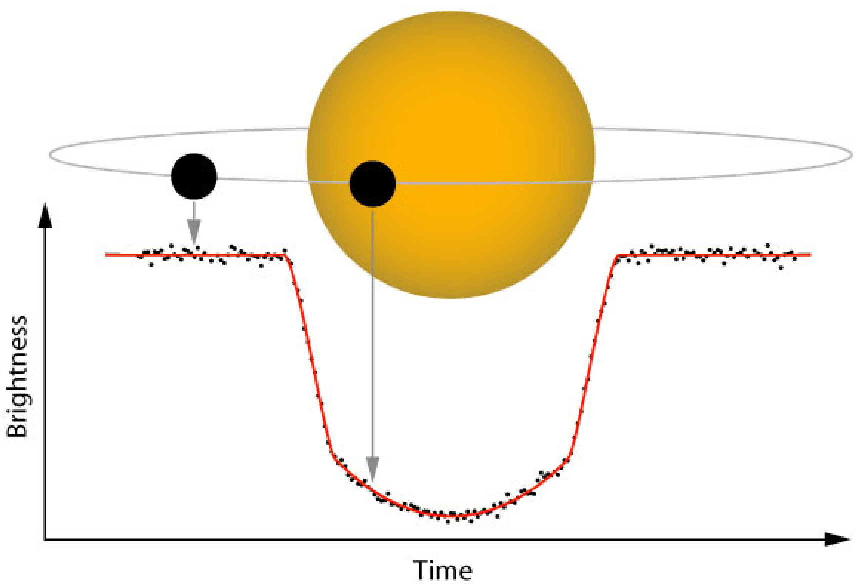
Recent reviews:
- Baron 2019: ML in astronomy – a practical overview
- Fluke & Jacobs 2020: Surveying the reach and maturity of machine learning and artificial intelligence in astronomy
- Reis+2021: Effectively using unsupervised machine learning in next generation astronomical surveys
- Sen+2022: Astronomical big data processing using machine learning
- Huppenkothen+2023: Impactful Machine Learning Research for Astronomy: Best Practices

# „7 types of astronomical data“

Fluke & Jacobs 2020





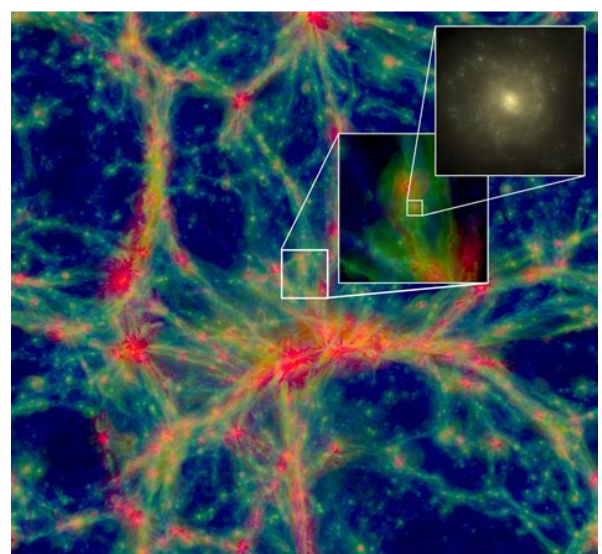
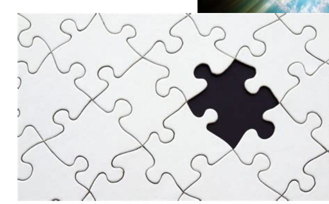



Image
Spectroscopy
Photometry
Light curve
Time Series
Catalogue
Simulation

# „7 types of astronomical problems“

Fluke & Jacobs 2020

| | |
|---|---|
| Classification | |
| Regression | |
| Clustering | |
| Forecasting | Learn from previous events, and predict or forecast that a similar event is going to occur |
| Generation | Missing information is created, expected to be consistent with the underlying truth |
| Discovery | New celestial objects, features, or relationships are identified |
| Insight | Insight is gained into the suitability of applying machine learning, choice of data set, hyperparameters, etc |



# „7 types of astronomical problems“

Fluke & Jacobs 2020

| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

# „7 types of computational problems“

Ivezic+2020

1) **Basic problems**: simple statistics: O(N) or O(N log N) at worst
2) **Generalized N-body problems**: any problem involving distances between tuples of points (nearest-neighbor searches, KDE): typically $O(N^2)$ or $O(N^3)$
3) **Linear algebraic problems**: linear systems, eigenvalue problems, and inverses: can be O(N) but in some cases the matrix of interest is $N \times N$, making the computation $O(N^3)$
4) **Optimization problems**: from unconstrained (O(N)) to constrained (e.g. nonlinear support vector machines: $O(N^3)$ convex and nonconvex.
5) **Integration problems**: e.g. estimation of Bayesian models: integration with high accuracy via quadrature has a computational complexity which is exponential in the dimensionality D → MCMC
6) **Graph-theoretic problems**: probabilistic graphical models or nearest-neighbor graphs for manifold learning
7) **Alignment problems**: “cross-matching” in astronomy: The worst-case cost is exponential in N...

# „7 types of astronomical problems“

Fluke & Jacobs 2020

| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

ML Methods used in our group: (XD)GMM, kNN, ANN (MLP), CNN, umap, t-SNE, XGBoost

# „7 types of astronomical problems“

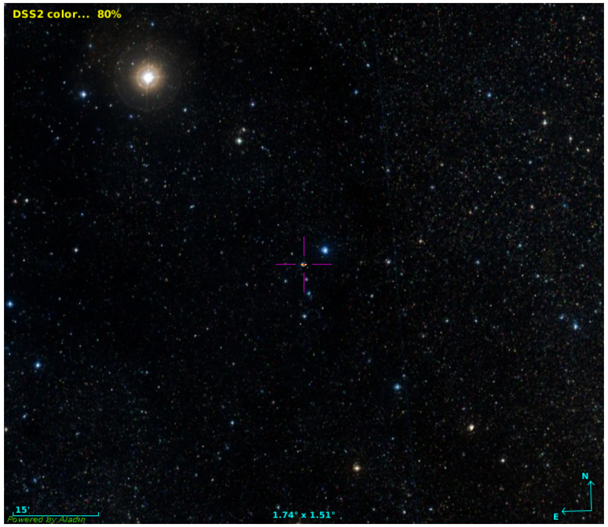
Fluke & Jacobs 2020

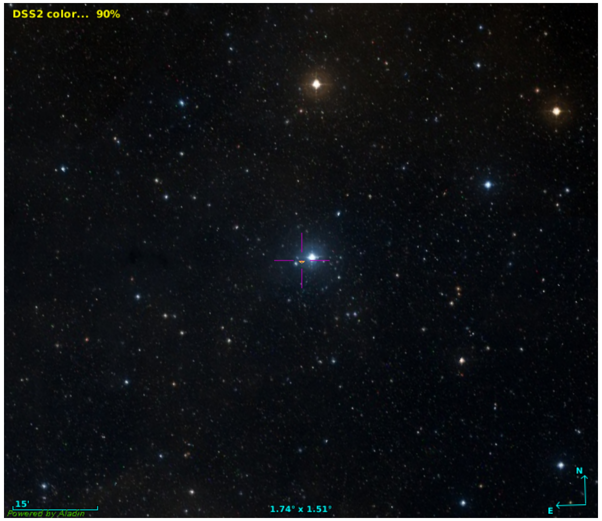
| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

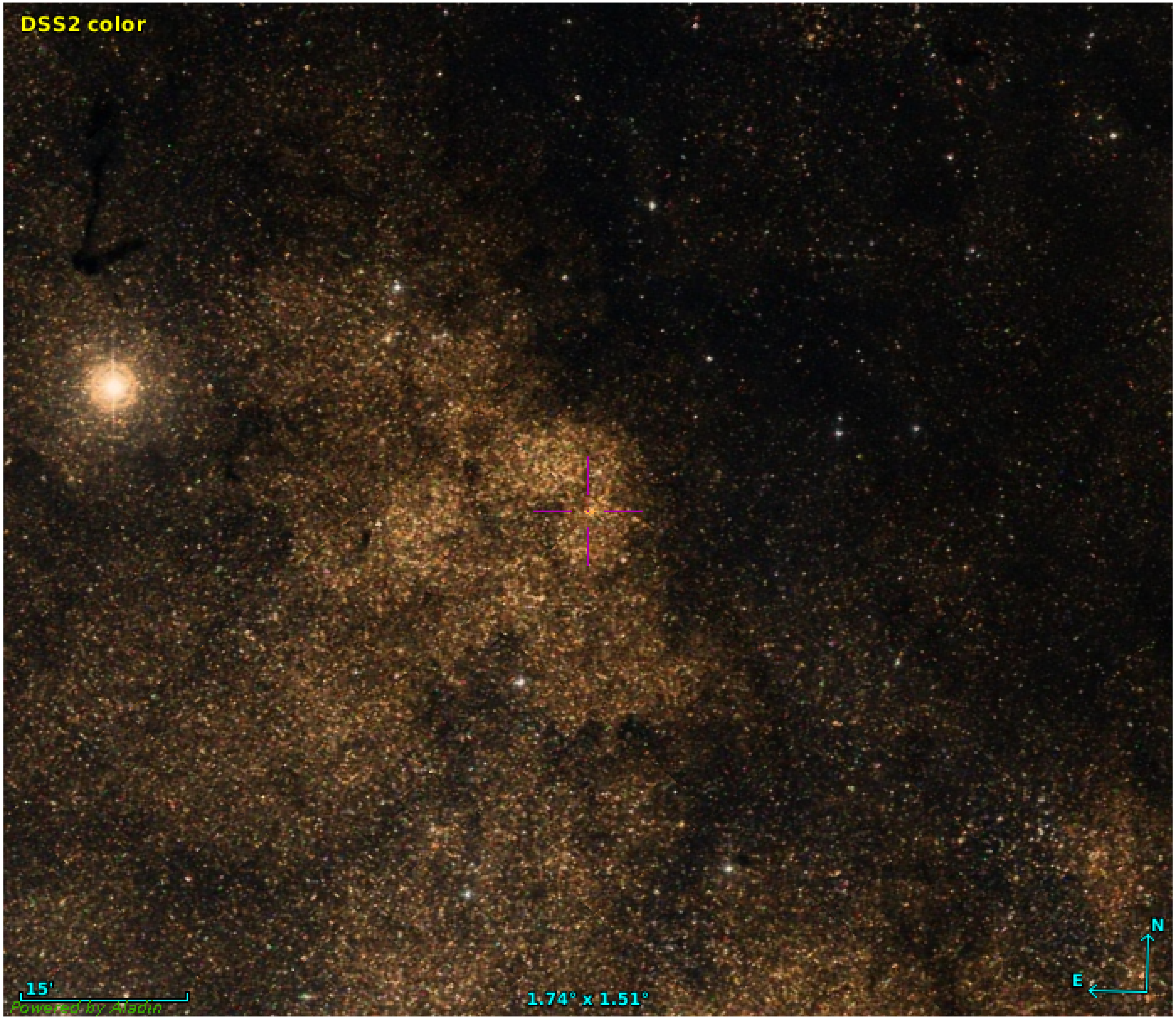
# Science case I:
# The Gaia open cluster census

| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

# What is an open star cluster?



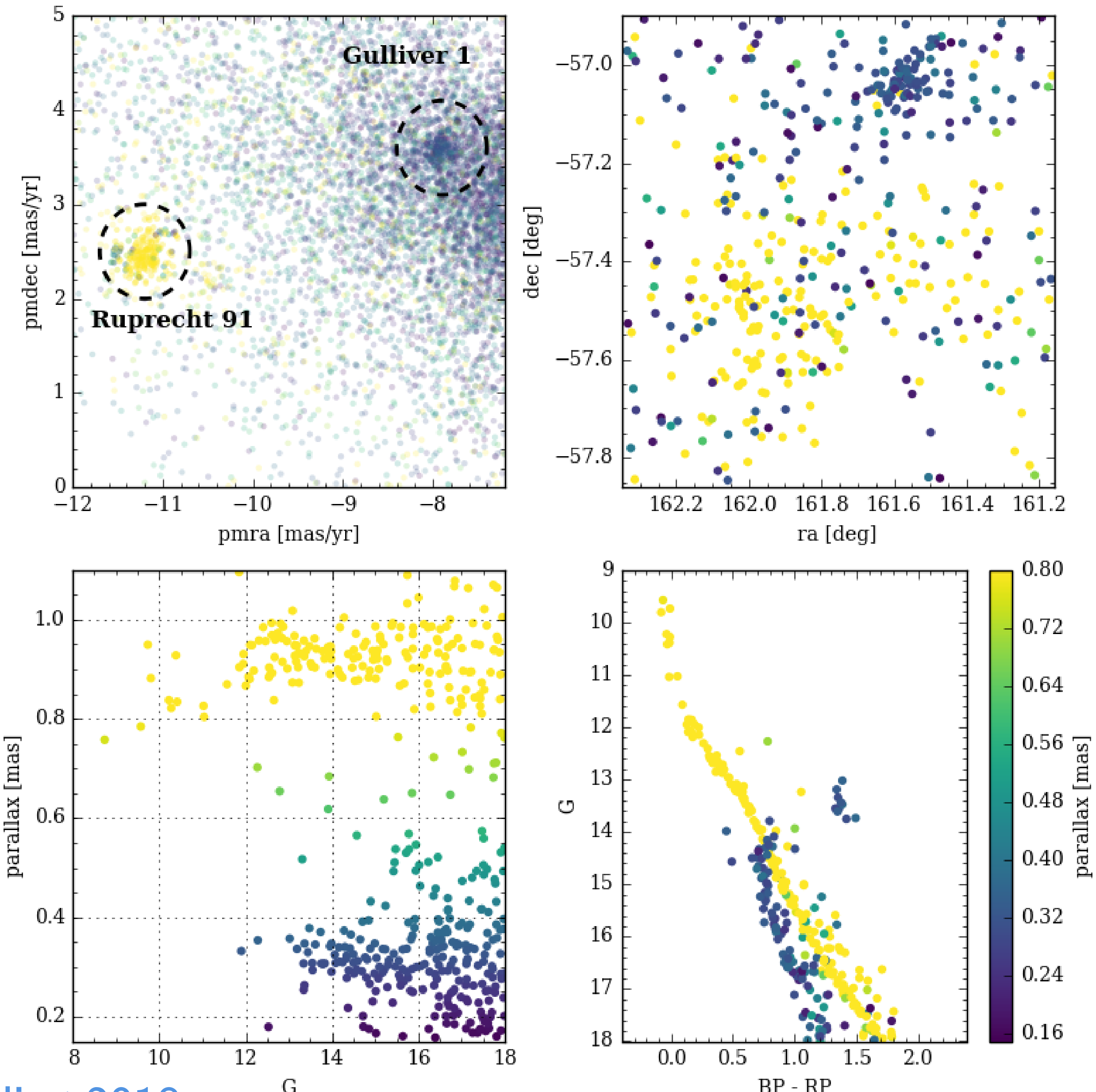
# Gaia DR2: cluster finding in phase space
# Serendipitous discoveries

Basically:

New discoveries could be made **by eye.**

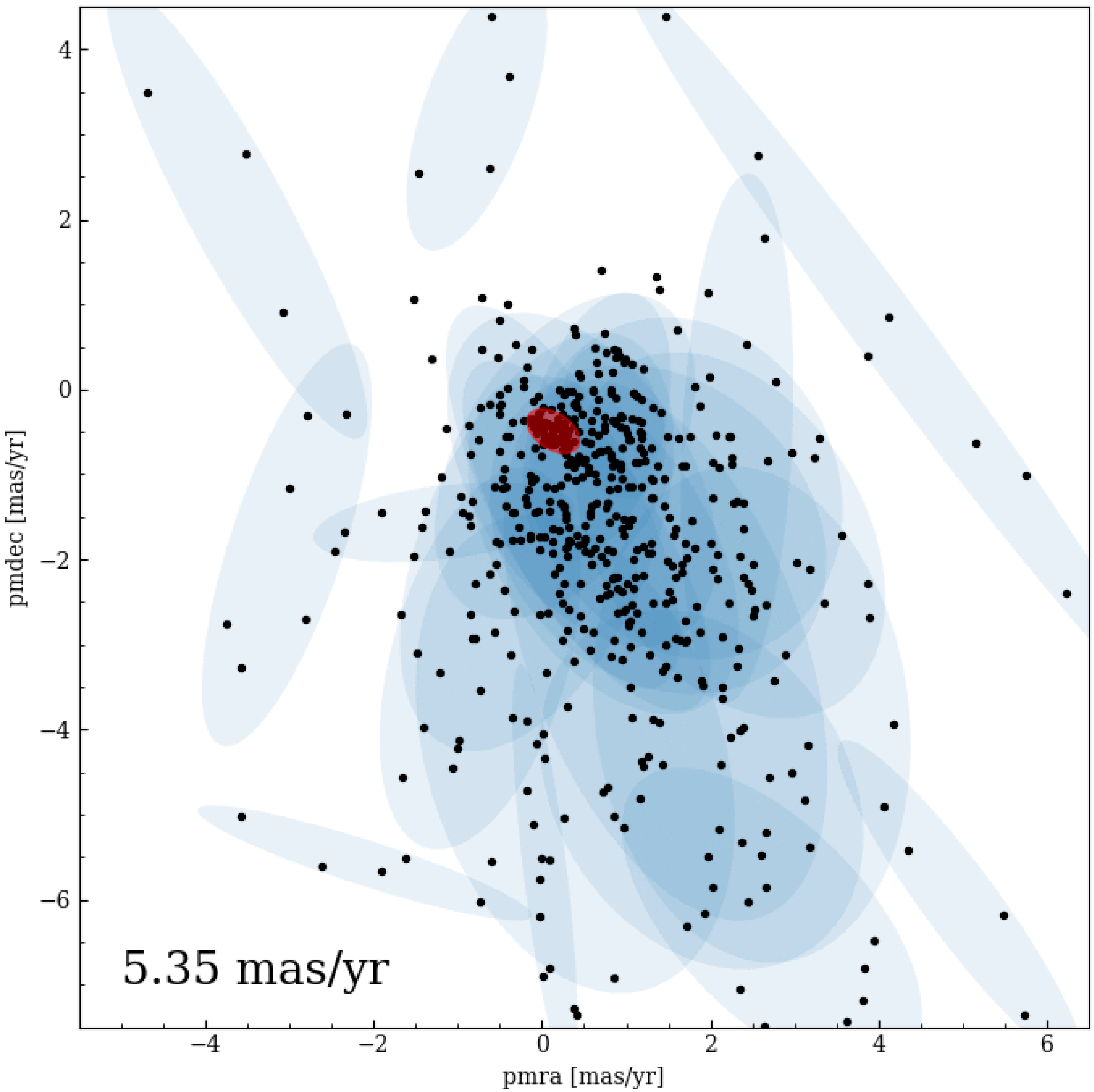
The Gaia DR2 data are often so good that they look like textbook examples for clustering algorithms.

It was clear that much more could be found...




Cantat-Gaudin+2018

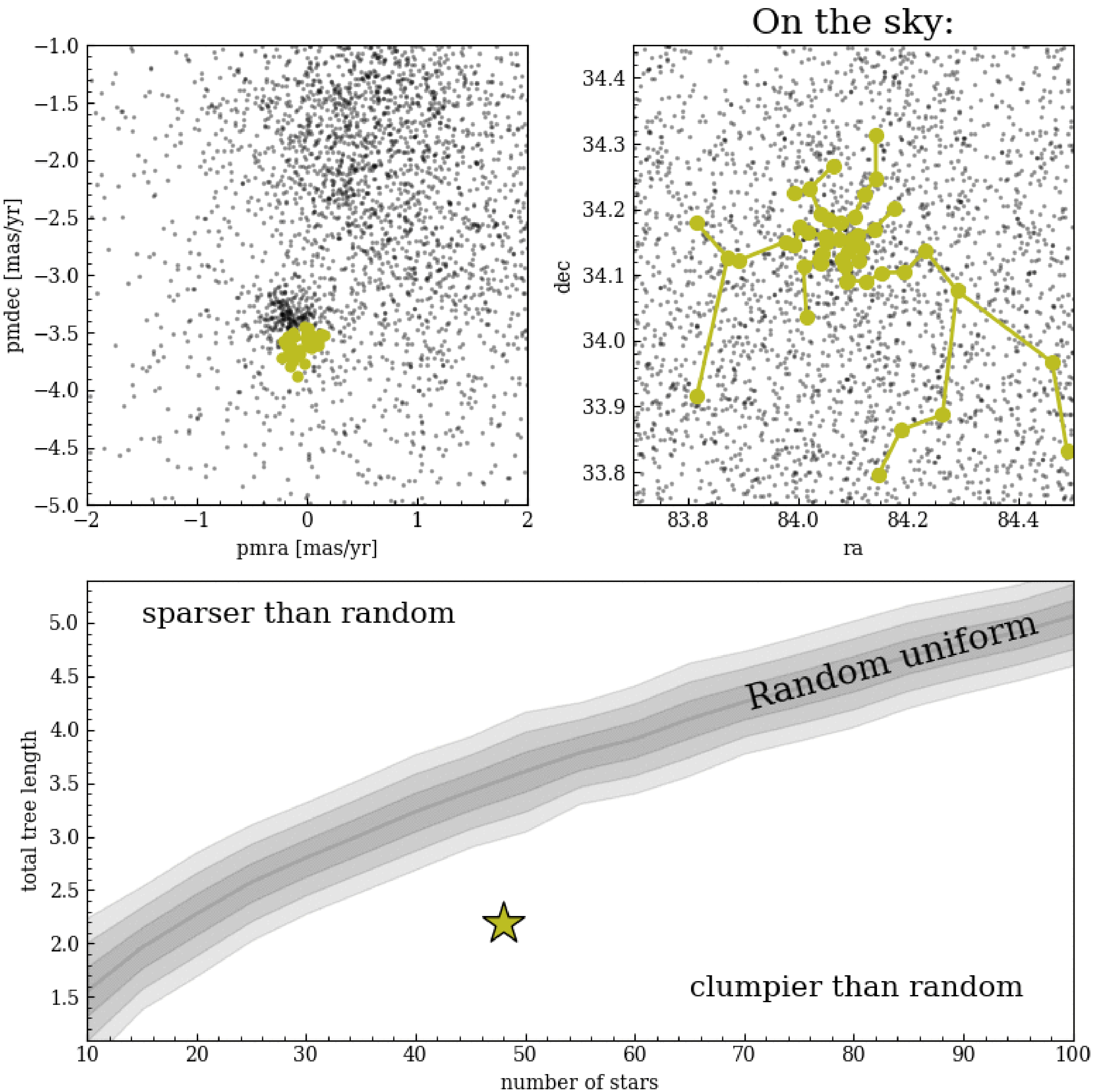
# Cantat-Gaudin+2019: GMM + K-Means




Step 1:
Use Gaussian mixture model to look for sharp peaks in the velocity distribution

→ cluster candidates

# Cantat-Gaudin+2019: GMM + K-Means

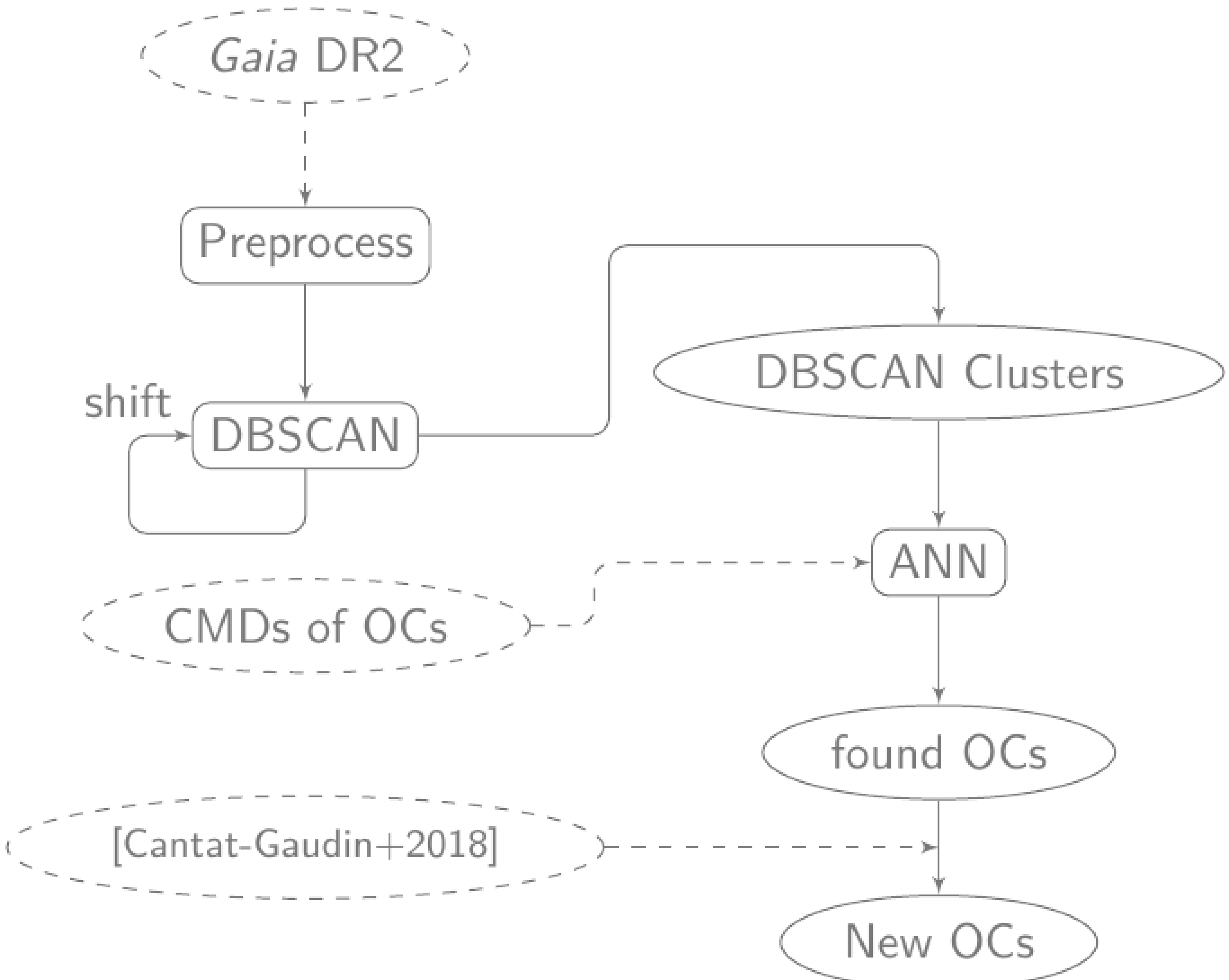



Step 2 (UPMASK):
K-means clustering in 5D astrometric space

Step 3 (UPMASK):
Compare minimum spanning tree distance with random distribution

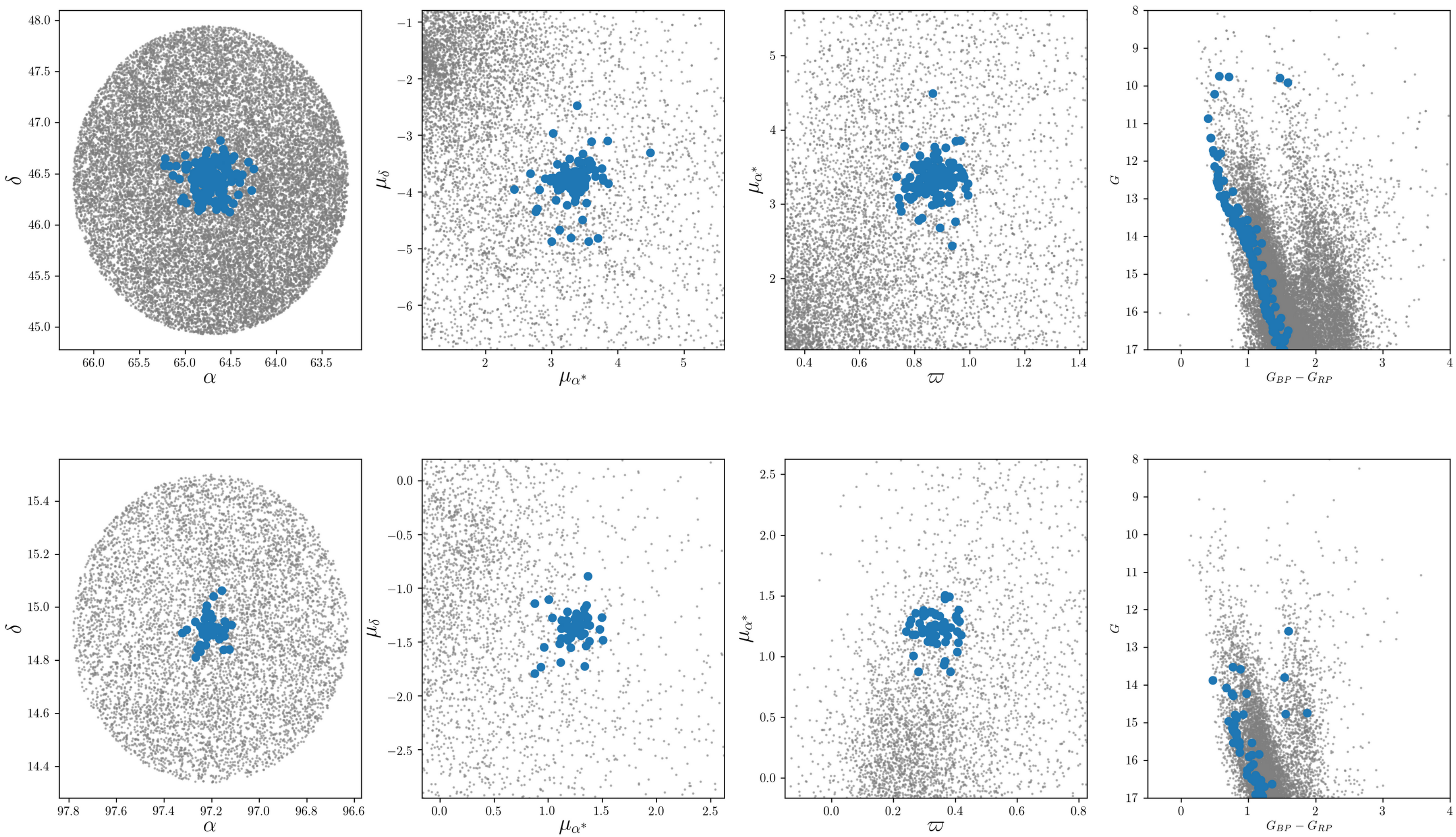
→ confirm cluster candidates

# Castro-Ginard+2019,2020,2022: Applying DBSCAN to Gaia data



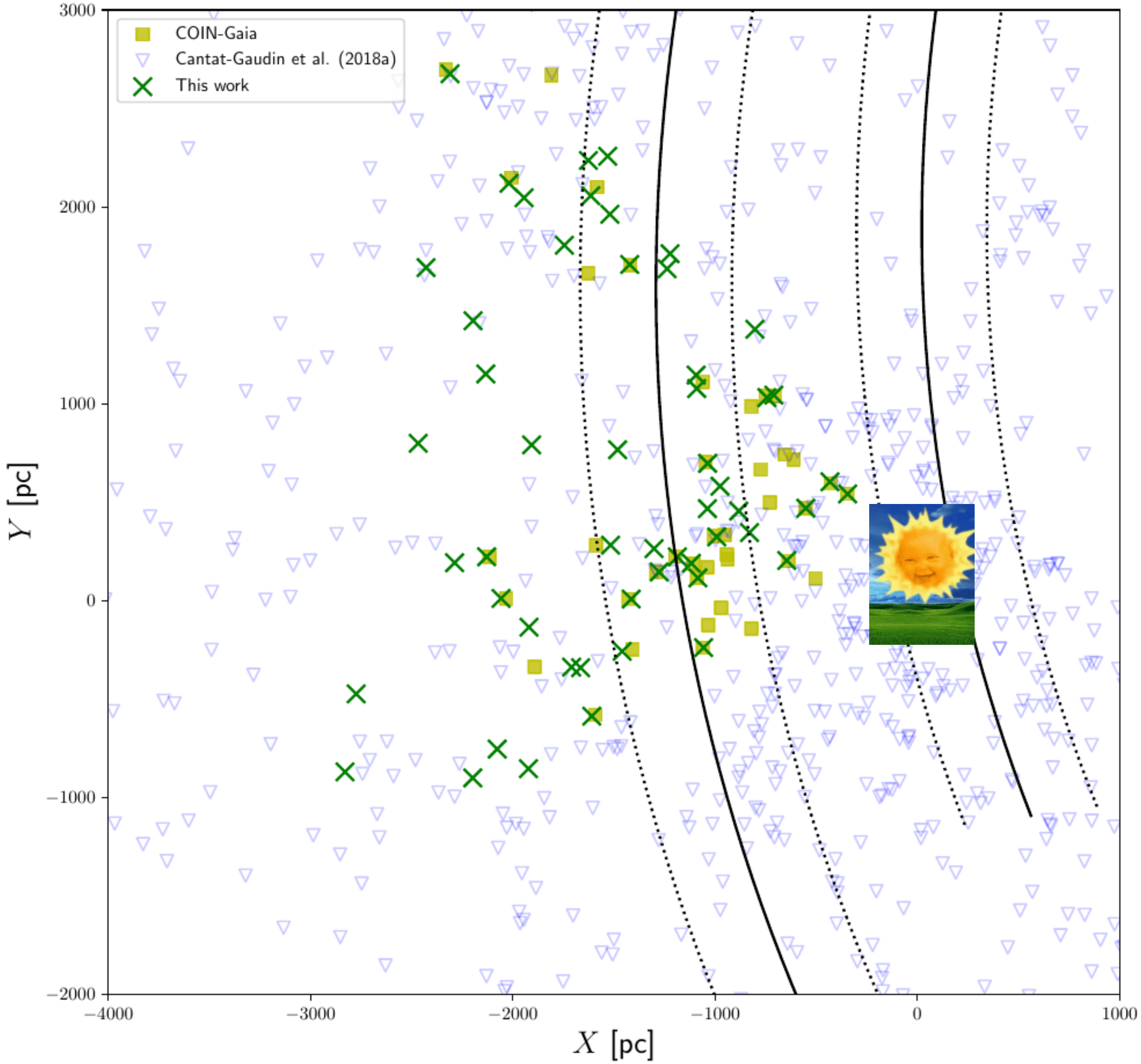
# Castro-Ginard+2019, 2020, 2022: Applying DBSCAN to Gaia data

5D **clustering + simple ANN classifier** (trained with expert humans)
to distinguish cluster and asterism colour-magnitude diagrams (90% accuracy)



# Castro-Ginard+2019: Applying DBSCAN in the Perseus arm

Some in common with Cantat-Gaudin+2019, but also ~40 new ones



# Castro-Ginard+2020:
# Applying DBSCAN to the full Galactic disc

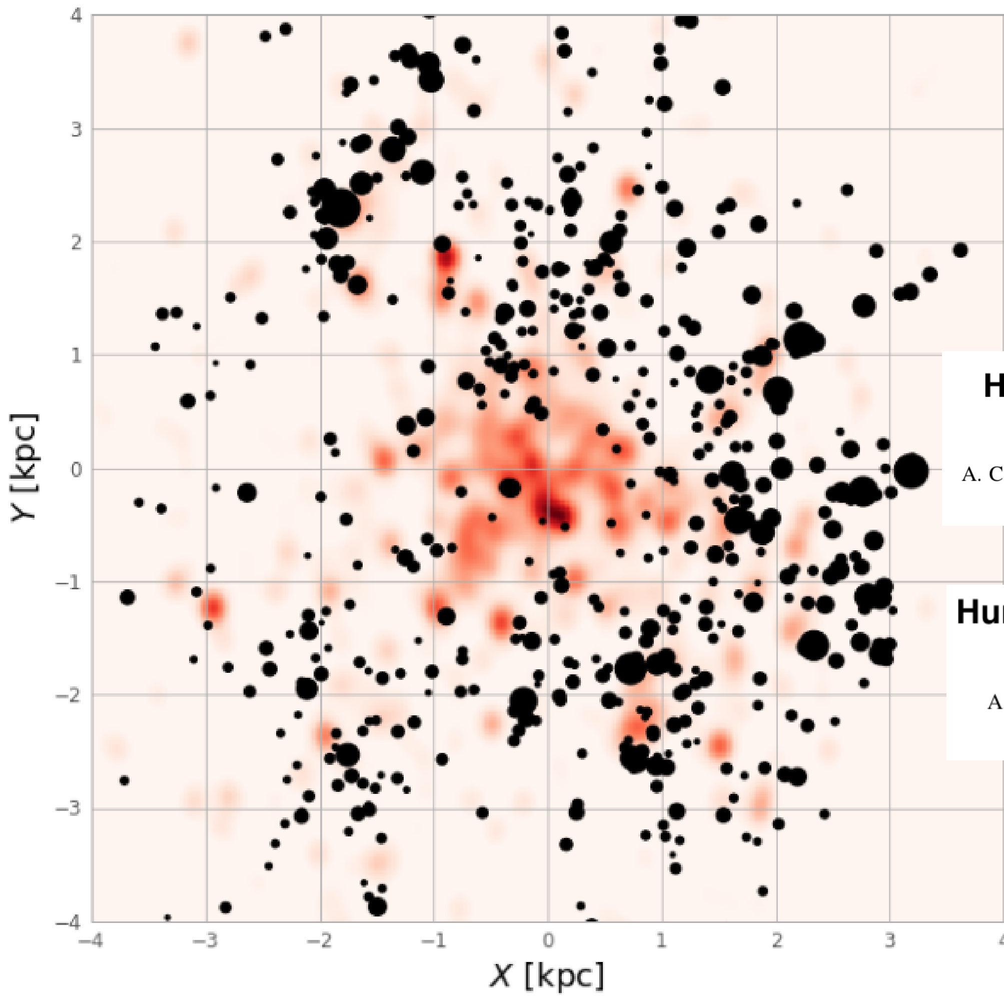



**Fig. 5.** Distribution of the OCs projected in the $X-Y$ plane. Previously known OCs (CG18,CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as a density map in red. Newly found OCs reported here are shown as black dots, where the size is proportional to the number of members of each cluster.

**Hunting for open clusters in *Gaia* DR2: 582 new open clusters in the Galactic disc★**

A. Castro-Ginard[1], C. Jordi[1], X. Luri[1], J. Álvarez Cid-Fuentes[2], L. Casamiquela[3], F. Anders[1], T. Cantat-Gaudin[1], M. Monguió[1], L. Balaguer-Núñez[1], S. Solà[2], and R. M. Badia[2]

**Hunting for open clusters in *Gaia* EDR3: 628 new open clusters found with OCfinder★**

A. Castro-Ginard[1,2], C. Jordi[2], X. Luri[2], T. Cantat-Gaudin[2,3], J. M. Carrasco[2], L. Casamiquela[4], F. Anders[2], L. Balaguer-Núñez[2], and R. M. Badia[5]

# Cantat-Gaudin+2020: Open cluster parameters with an ANN

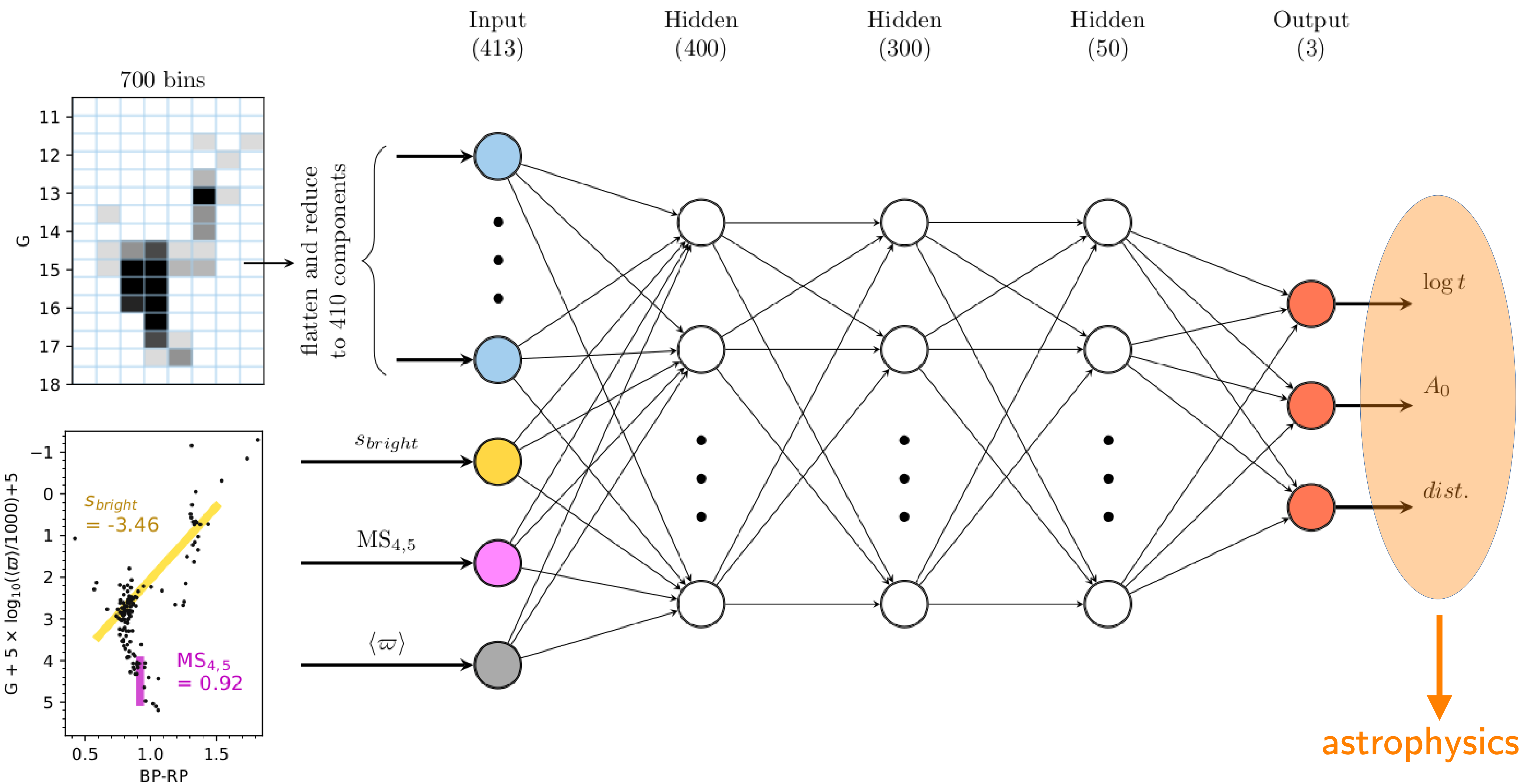



**Fig. 2.** Architecture of our artificial neural network, indicating the width (number of nodes) of each layer. The example cluster is Haffner 22. The input quantities are described in Sect. 3.1.

# Cantat-Gaudin & Anders 2020: Heaps of asterisms in the OC literature

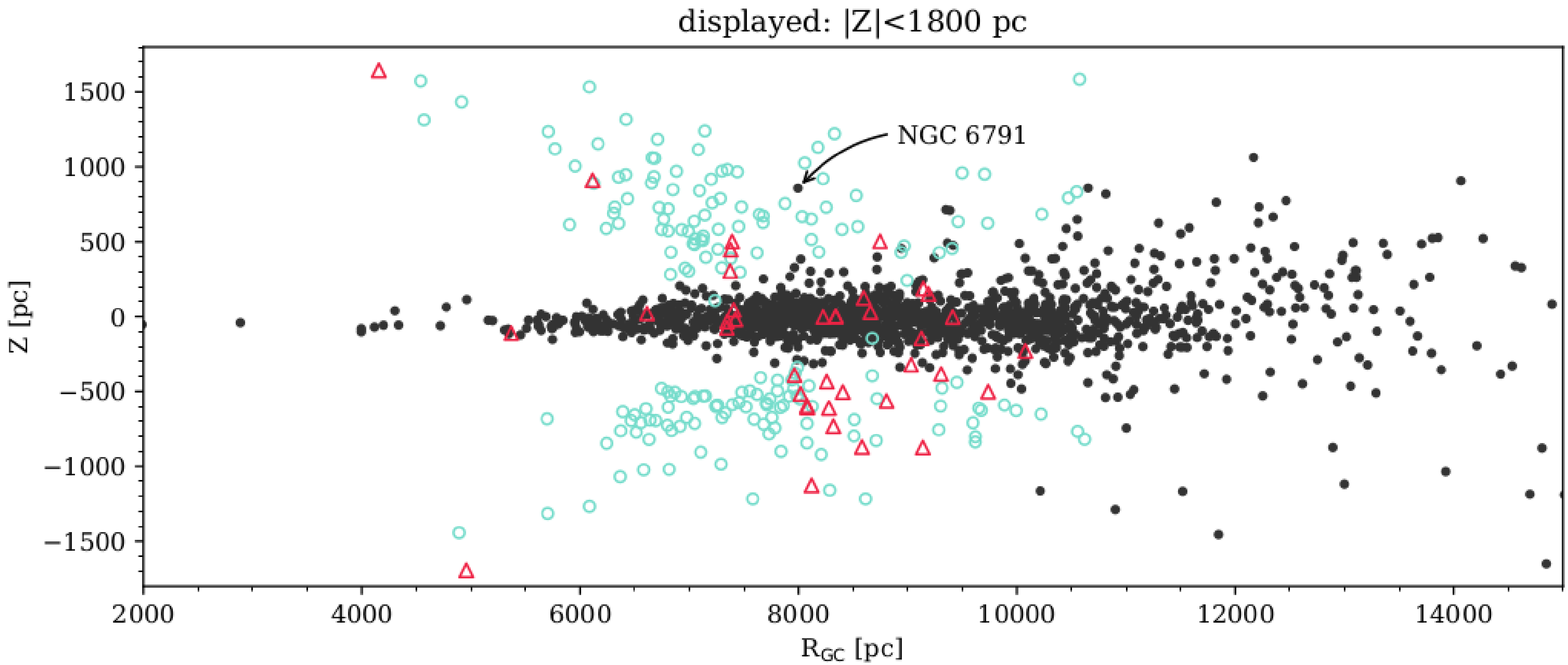



**Fig. 3.** Black dots: clusters whose existence has been confirmed with *Gaia* DR2 data. Open circles: expected location of the candidate clusters of Schmeja et al. (2014) and Scholz et al. (2015). Open triangles: expected location of the other groupings that this study argues are asterisms.

→ Also an **un-discovery** leads to astrophysics

# Asterisms:
# Why we get fooled


In the discovery process:

- because **we want to discover *something***
- because 1-2 bright stars dominate the light profile and **fool the eye**
- because of holes in the **dust** distribution
- because of zonal **distortions in** ancillary Schmidt plate **data**

In literature work:

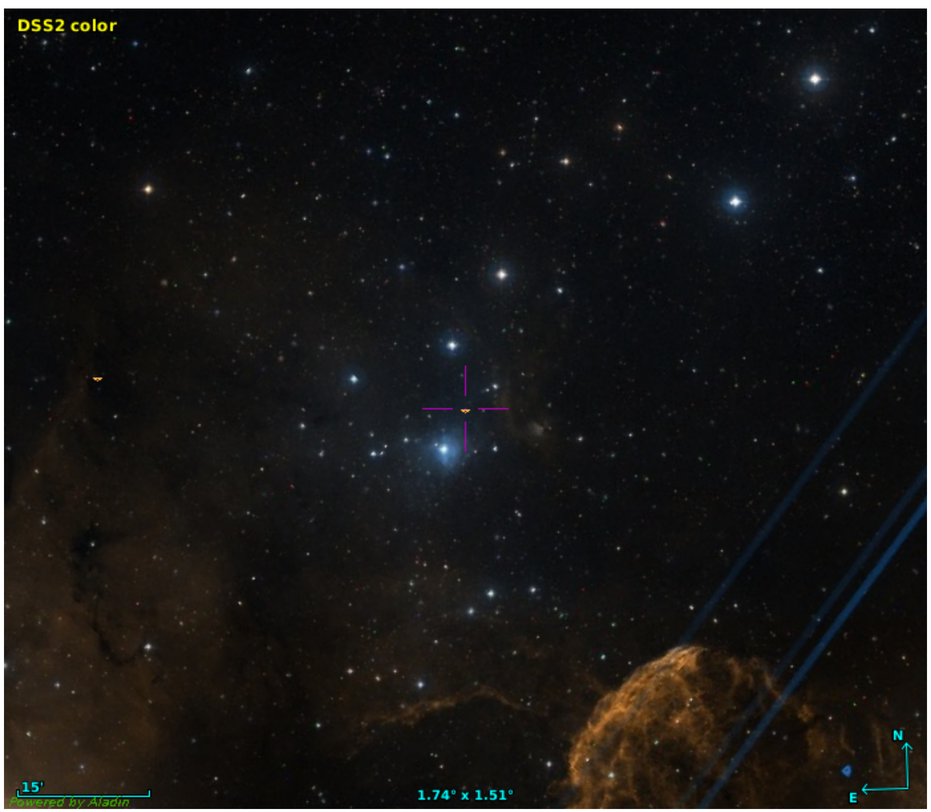
- because of **tradition**
- **to avoid conflicts** with referees and established researchers

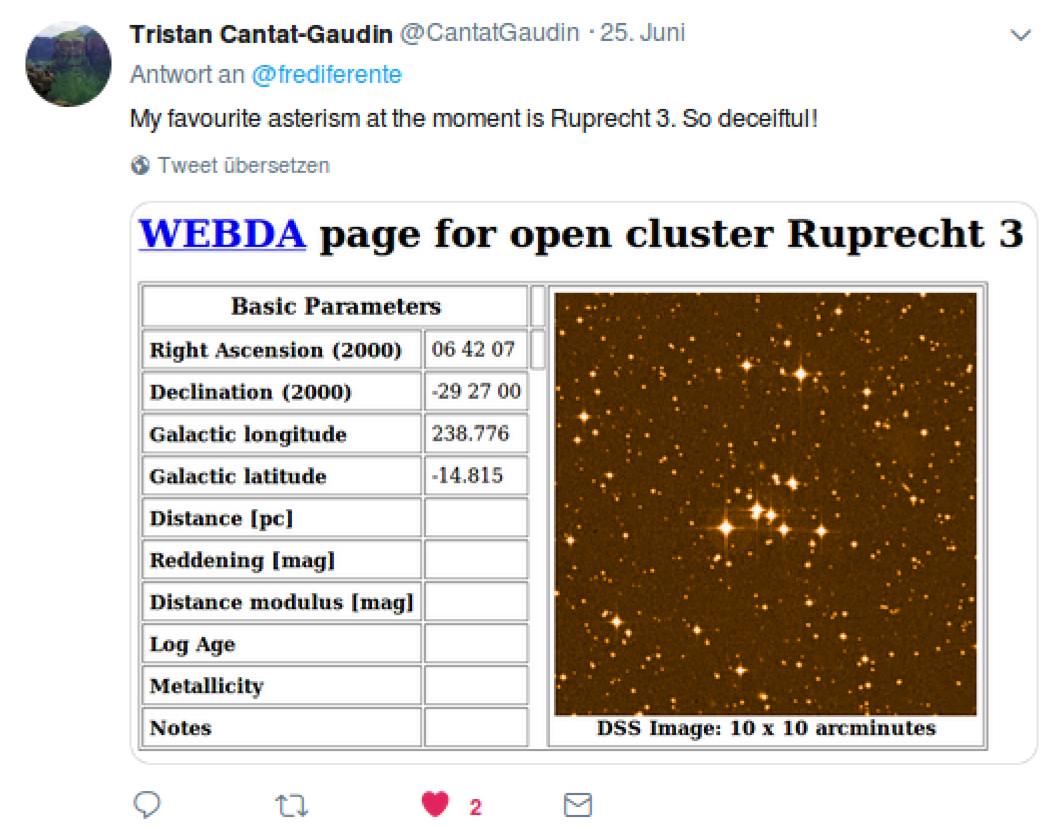
# Conclusions after Gaia DR2

1) Gaia is an optimal playground for clustering! We are not even doing the most fancy things. Almost every time we try a new algorithm, we find something new.

2) **The hard work is** not in *finding* candidates - but in vetting and **interpreting** them. What is a good metric in 5D astrometric space?

3) Don‘t use catalogues blindly.
Nobody will ever write a paper about the non-existence of Basel 5.
That doesn‘t mean it actually exists.





**Tristan Cantat-Gaudin** @CantatGaudin · 25. Juni
Antwort an @frediferente
My favourite asterism at the moment is Ruprecht 3. So deceitful!
Tweet übersetzen

**WEBDA page for open cluster Ruprecht 3**

| Basic Parameters | |
|---|---|
| Right Ascension (2000) | 06 42 07 |
| Declination (2000) | -29 27 00 |
| Galactic longitude | 238.776 |
| Galactic latitude | -14.815 |
| Distance [pc] | |
| Reddening [mag] | |
| Distance modulus [mag] | |
| Log Age | |
| Metallicity | |
| Notes | |

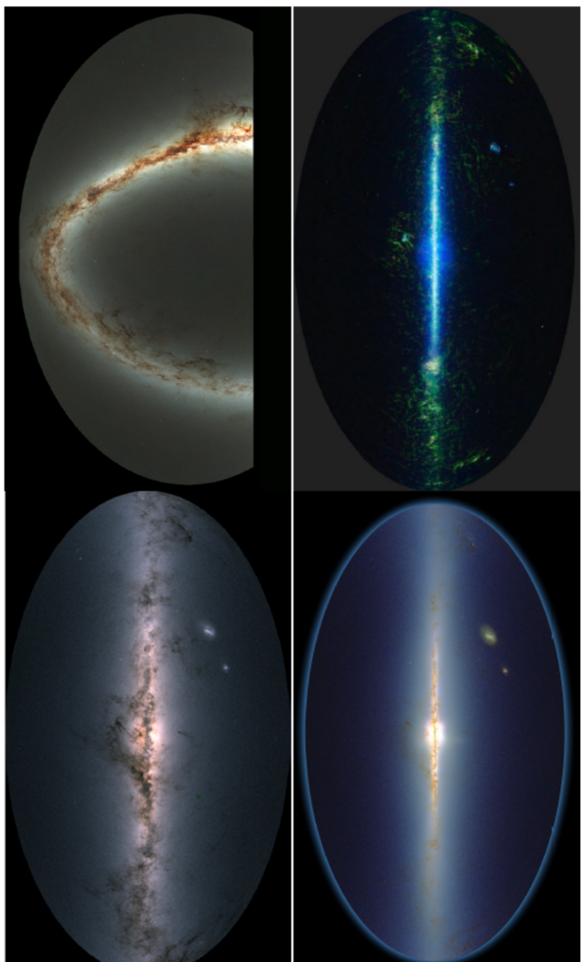
DSS Image: 10 x 10 arcminutes

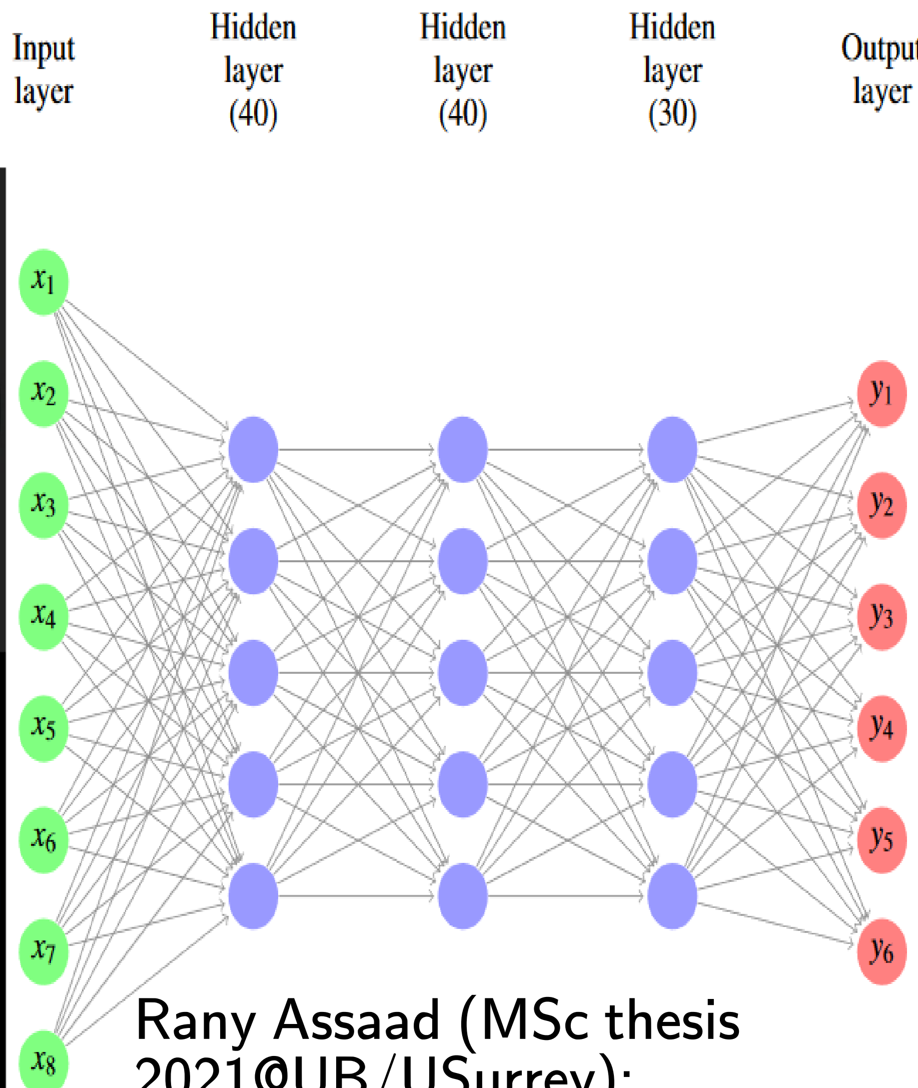
2

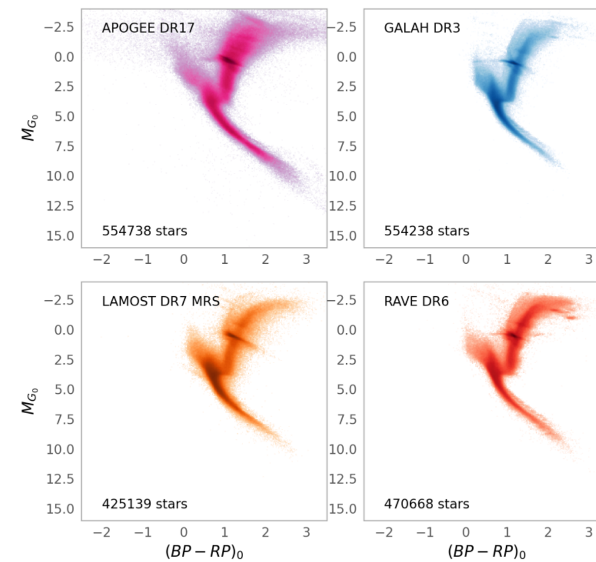
# Science case II:
# Stellar parameter estimation

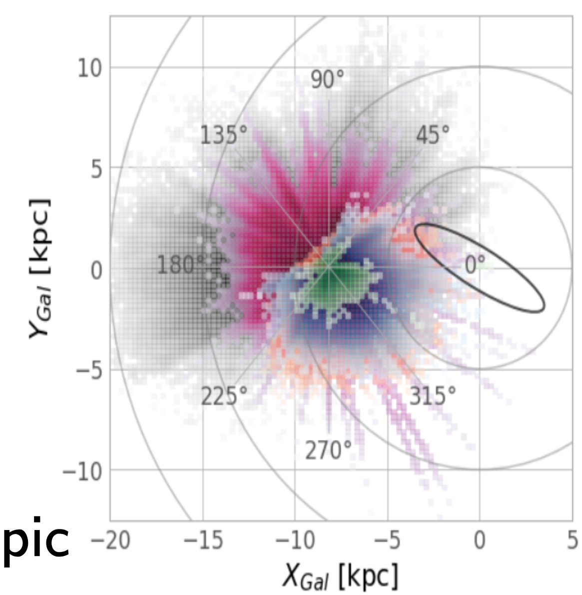
| **Nature/Type** | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

# Label transfer: Spectroscopy → Gaia + photometry (>300M stars)





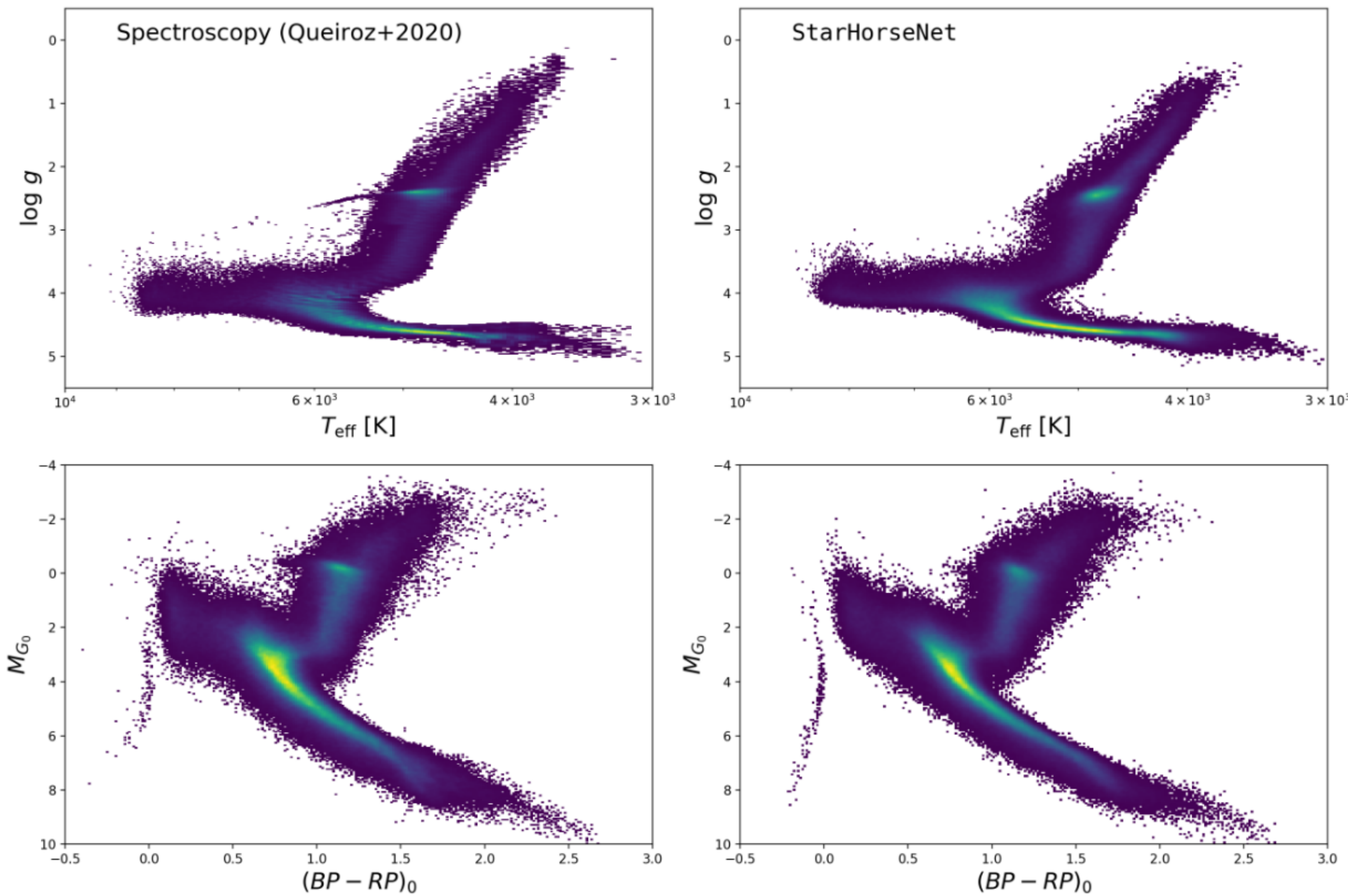




Rany Assaad (MSc thesis 2021@UB/USurrey):

- Tested **ANN regression** of Gaia DR2 + photometry to spectroscopic labels (Queiroz+2020)

# Label transfer: Spectroscopy → Gaia DR2 + photometry (>300M stars)

Test dataset:



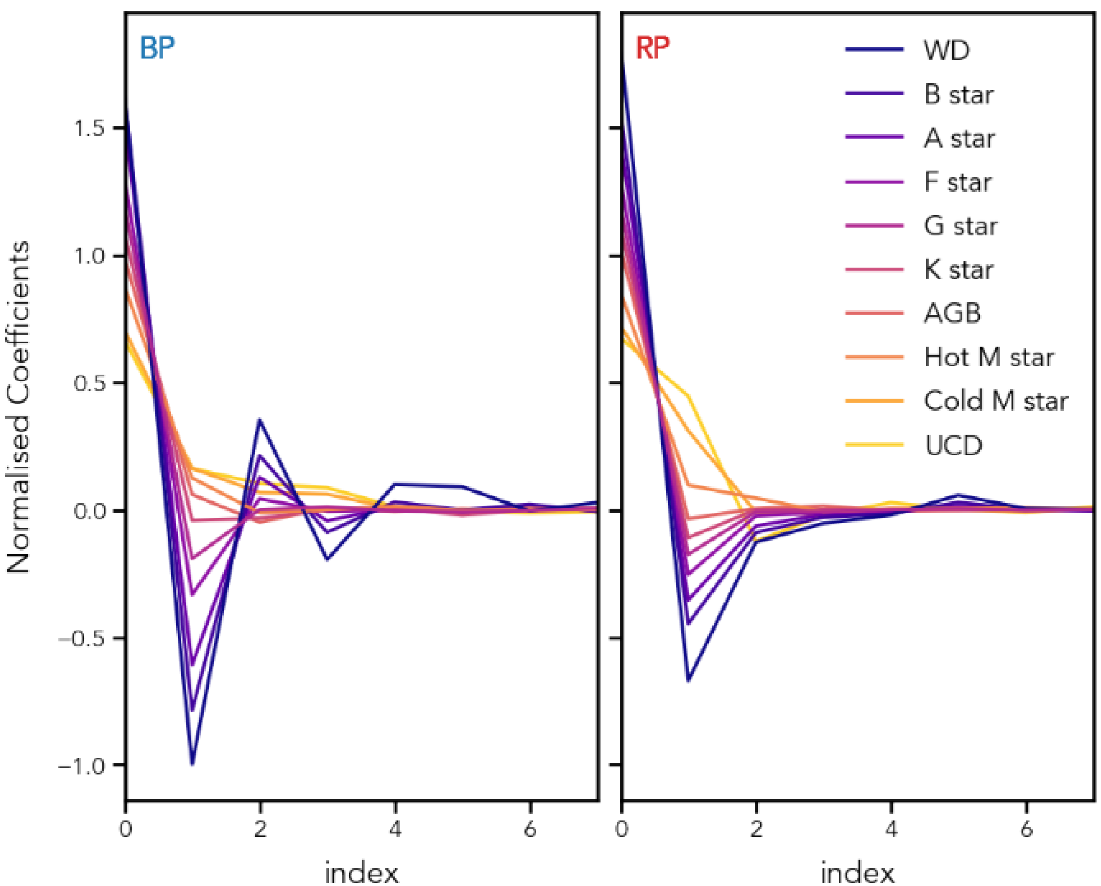

R. Assaad MSc thesis 2021, Anders+2023 proceeding

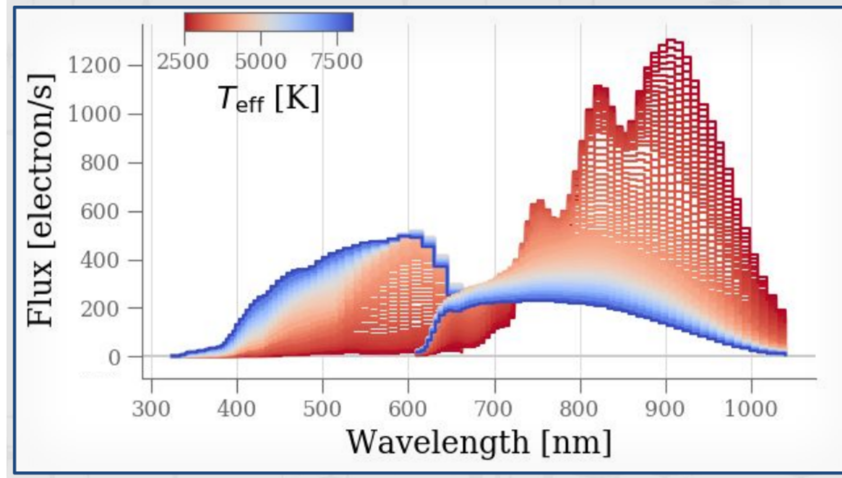
# gaia DR3: 220M BP/RP spectra!

CU8 presentation
O. Creevey EAS 2021

eas 2021 Leiden Virtual

Use of mean Bp and Rp Spectra



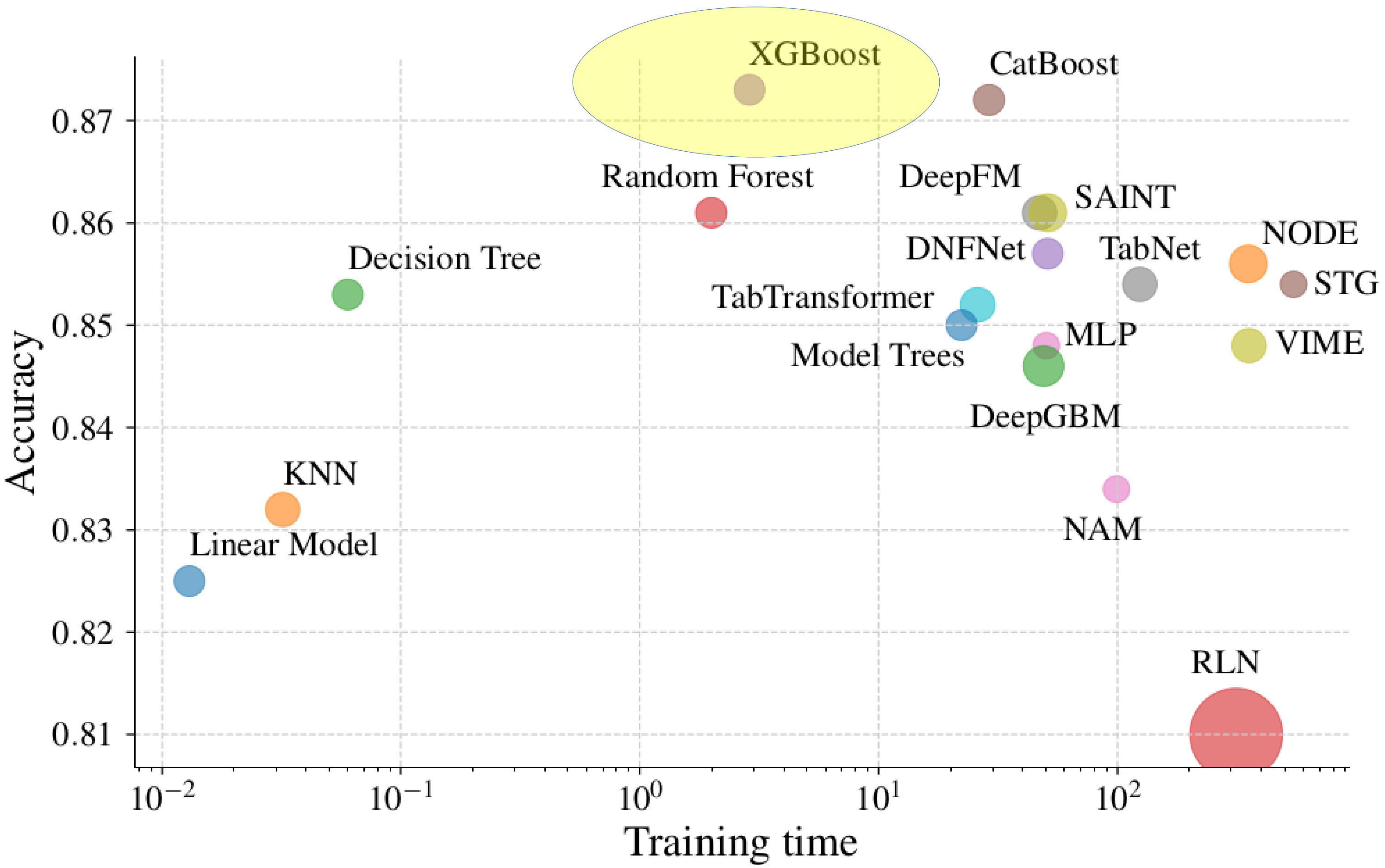

Bp and Rp spectra produced by Gaia-DPAC-CU5/DPCI
Astrophysical Parameters (APs) produced by Gaia-DPAC-CU8/DPCC

De Angeli+2023:

- First eight XP coefficients (see Carrasco+2021) of the continuous representation in BP and RP

# gaia DR3 StarHorseNet → SHBoost

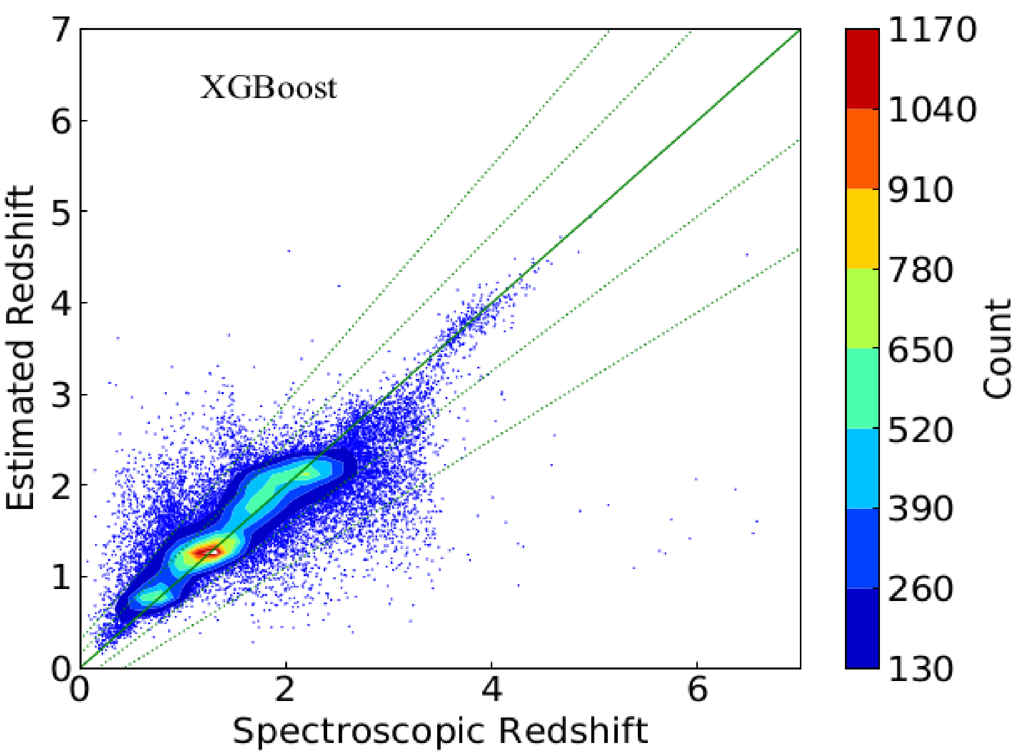
- DR3 includes 220M XP spectra → they can be easily fed to a StarHorseNet-like code
  → project to derive higher-precision stellar labels (with/without XP spectra)

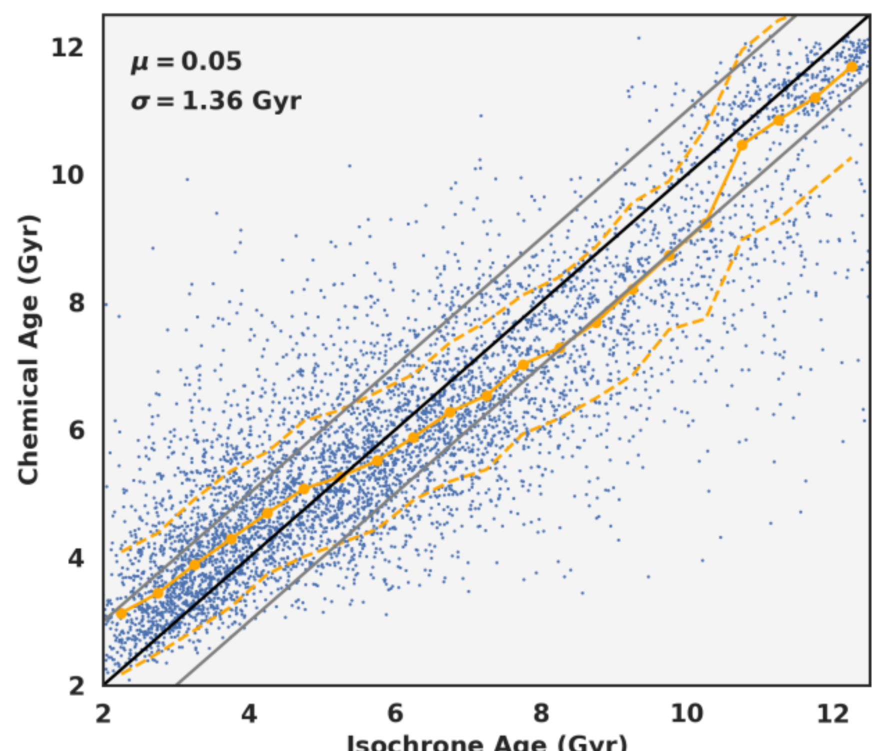



Borisov+2021 benchmark-tested regression algorithms for tabular data

# XGBoost regression in astronomy

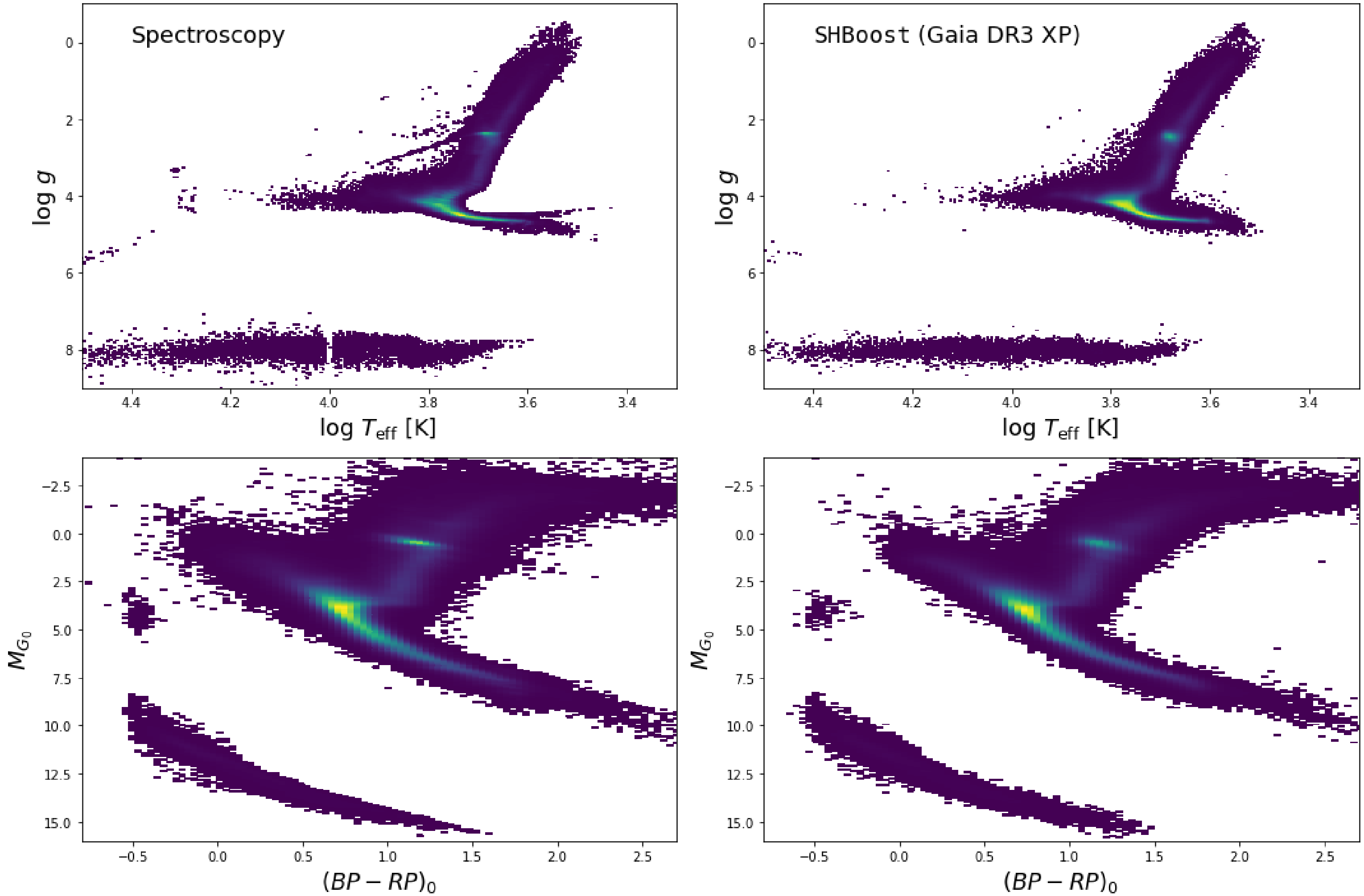
- Extreme Gradient-Boosted Trees *python* package *xgboost*: „scalable, portable and accurate" implementation of boosted trees (Friedman 2001, Chen & Guestrin 2016)
- Widely used for classification in astro: e.g. Bethapudi & Desai 2018; Yi et al. 2019; Li et al. 2019; Cunha & Humphrey 2022
- Examples for regression:
  - **Photometric redshifts** (Chong & Yang 2019; Li+2022; Humphrey+2023)
  - **Number of sunspots** (Dang+2022)
  - **Spectroscopic stellar ages** (Hayden+2022; Anders+2023)



# gaia DR3 StarHorse → SHBoost

- XGBoost works very nicely (even without parameter tuning) to predict StarHorse-like output parameters (d, Av, $T_{eff}$, log g, [M/H], mass)
- Much better metallicities (~0.15 dex) thanks to Gaia XP spectra
- Also works well for white dwarfs and hot stars (larger training set)

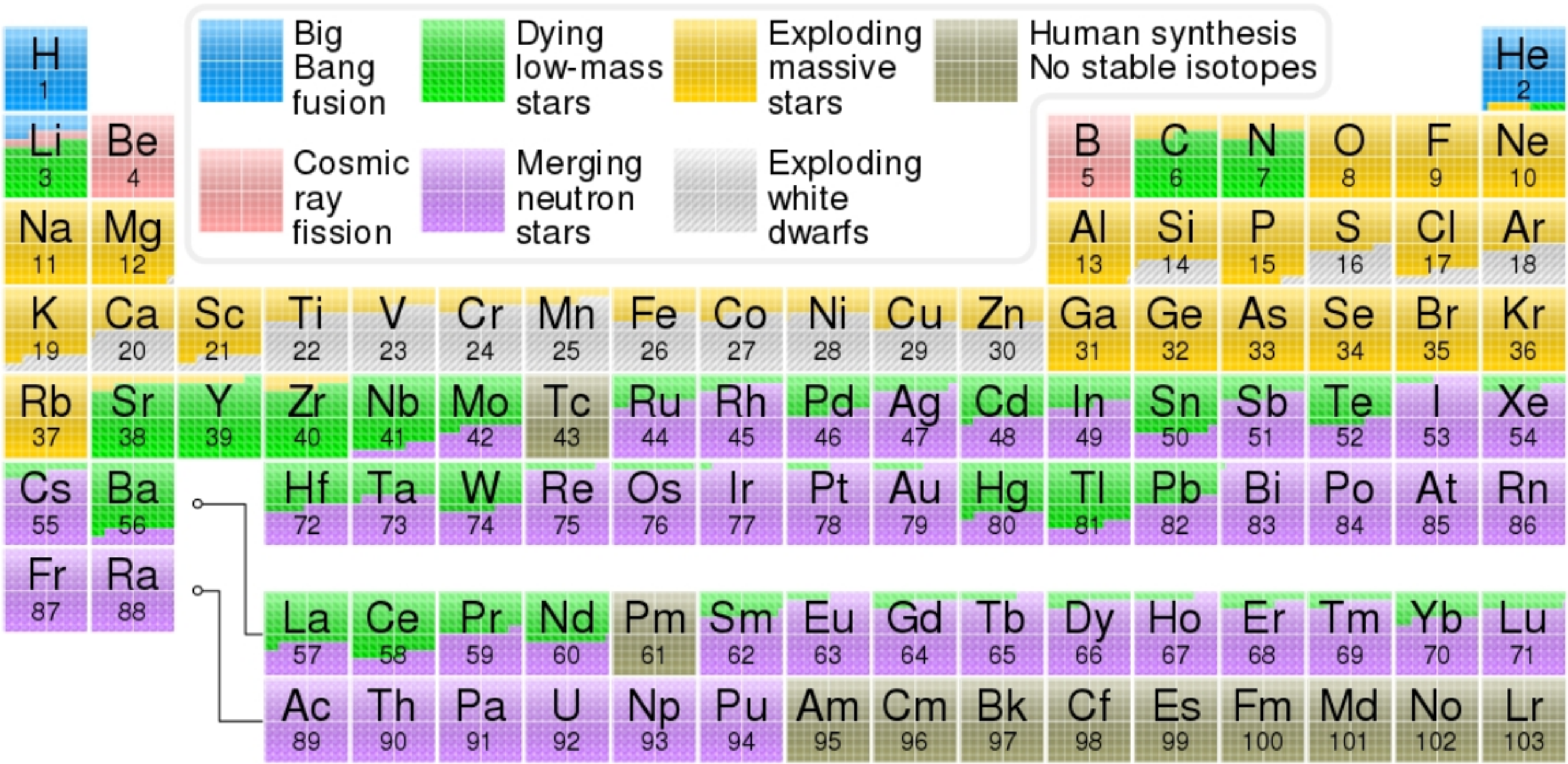


# Science case III: Chemical tagging

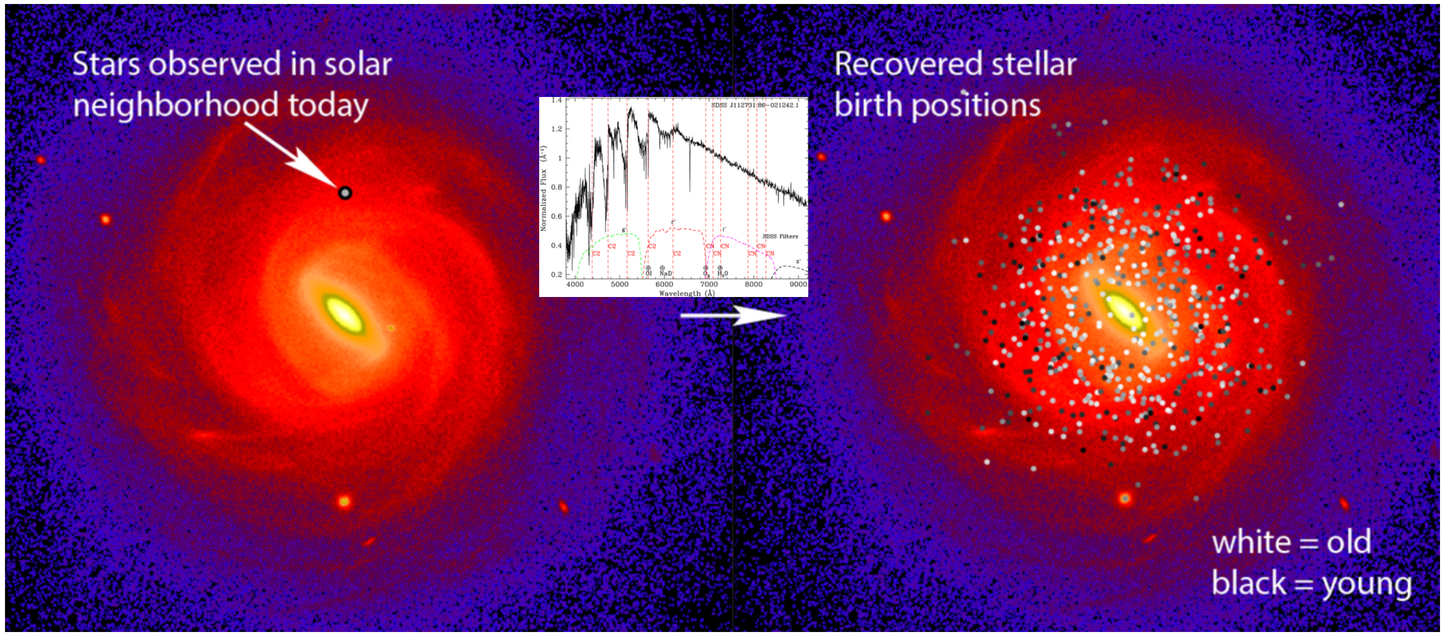
| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

# Science case III:
# Chemical tagging

## The origin of the elements



# Science case III: Chemical tagging



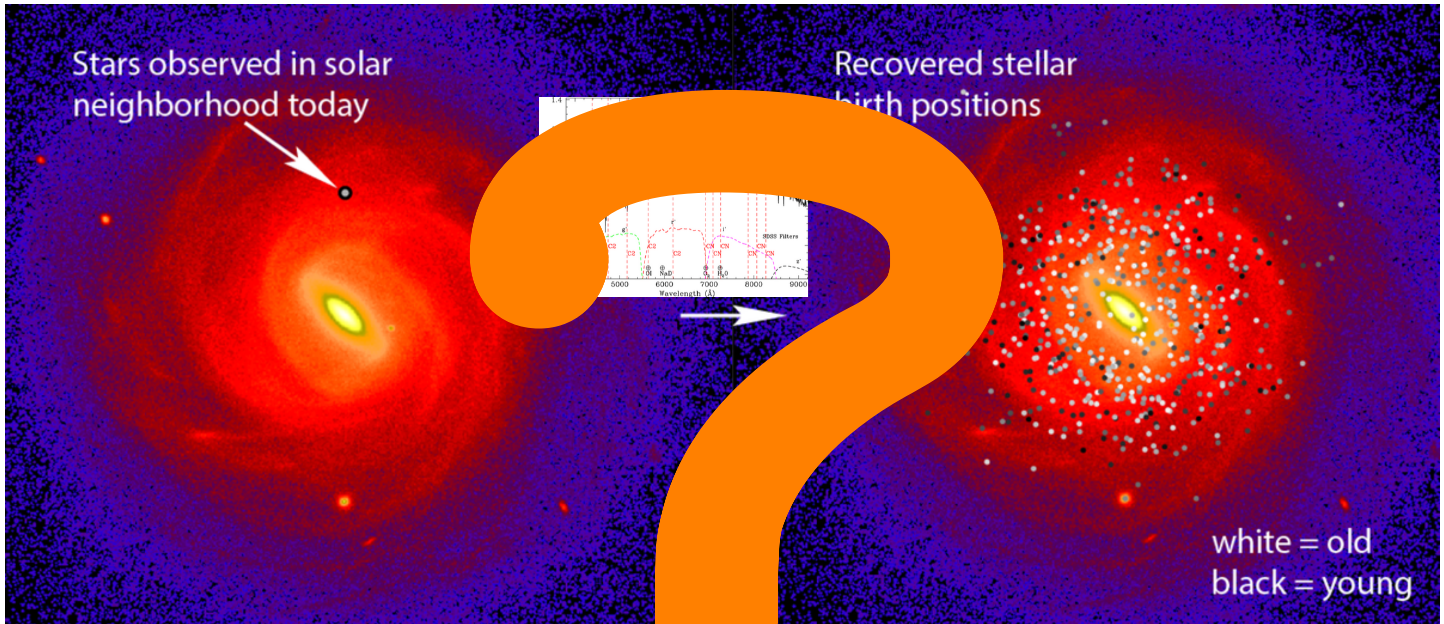

Minchev+2018 PR

## Chemical Signatures

A major goal of near-field cosmology is to tag or to associate individual stars with elements of the protocloud. For many halo stars, and some outer bulge stars, this may be possible with phase space information provided by Gaia. But for much of the bulge and the disk, secular processes cause the populations to become relaxed (i.e., the integrals of motion are partially randomized). In order to have any chance of unravelling disk formation, we must explore chemical signatures in the stellar spectrum. Ideally, we would like to tag a large sample of representative stars with a precise time and a precise site of formation.

Freeman & Bland-Hawthorn 2002

# Science case III: Chemical tagging




Minchev+2018 PR

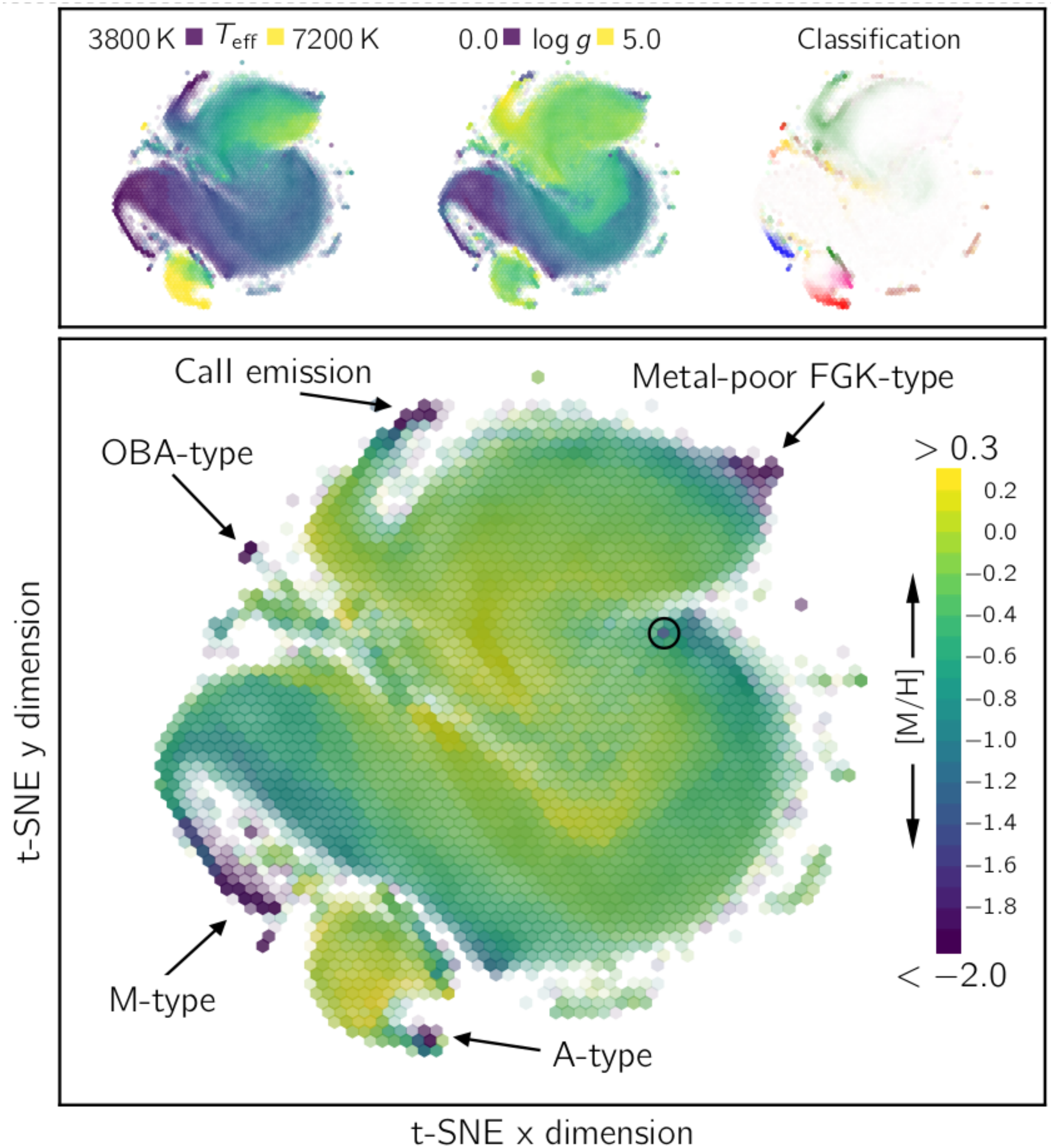
## Chemical Signatures

A major goal of near-field cosmology [illegible] tag or to associate individual stars with elements of the protocloud. For ma[illegible] stars, and some outer bulge stars, this may be possible with phase space inf[illegible]ation provided by Gaia. But for much of the bulge and the disk, secular processes cause the populations to become relaxed (i.e., the integrals of motion are partially randomized). In order to have any chance of unravelling disk formation, we must explore chemical signatures in the stellar spectrum. Ideally, we would like to tag a large sample of representative stars with a precise time and a precise site of formation.

Freeman & Bland-Hawthorn 2002

# An „easy“ example: Finding very metal-poor stars in surveys

Matijevic+2017:
First t-SNE analysis of stellar spectra



# An „easy“ example:
# Finding very metal-poor stars in surveys

Matijevic+2017

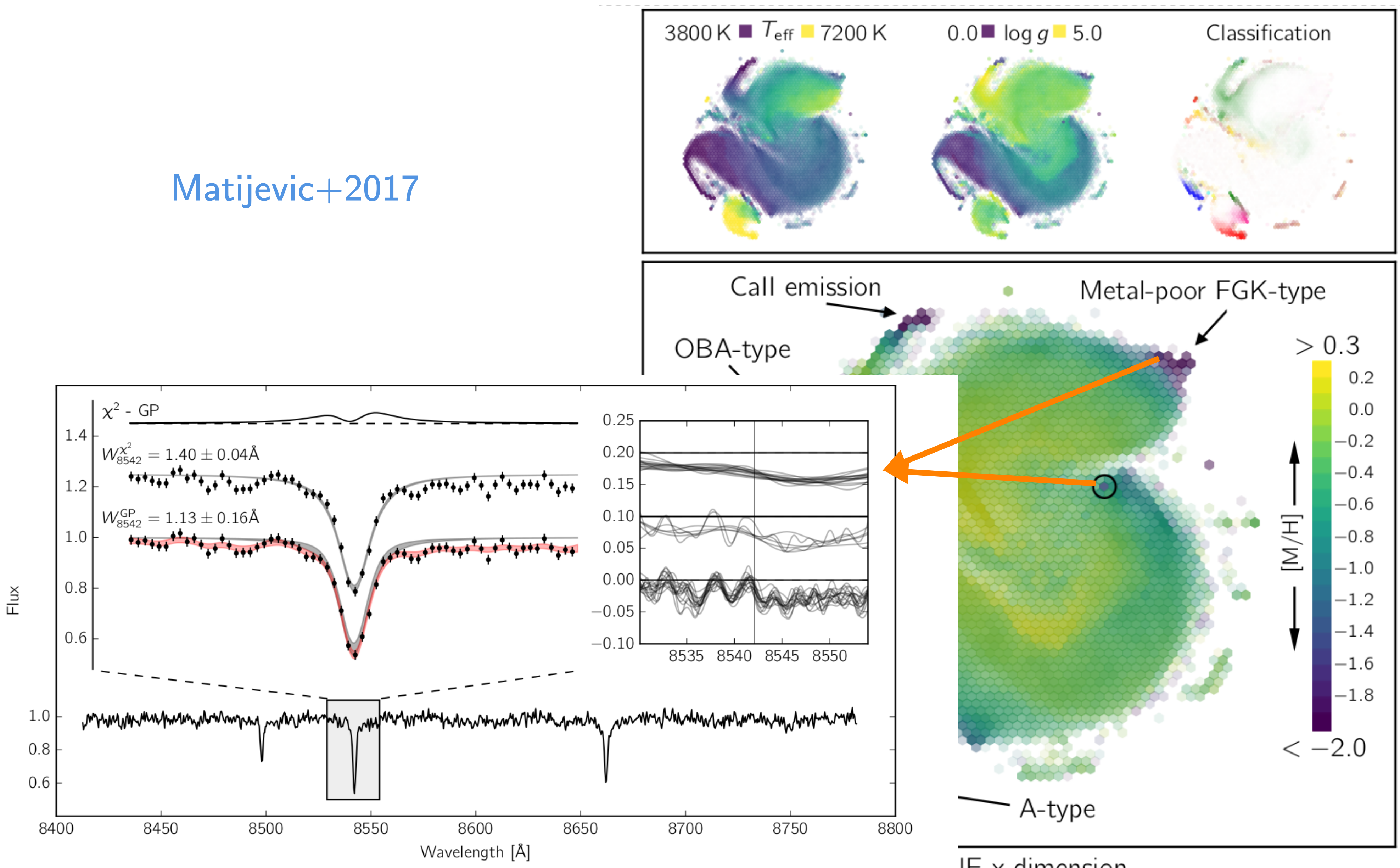
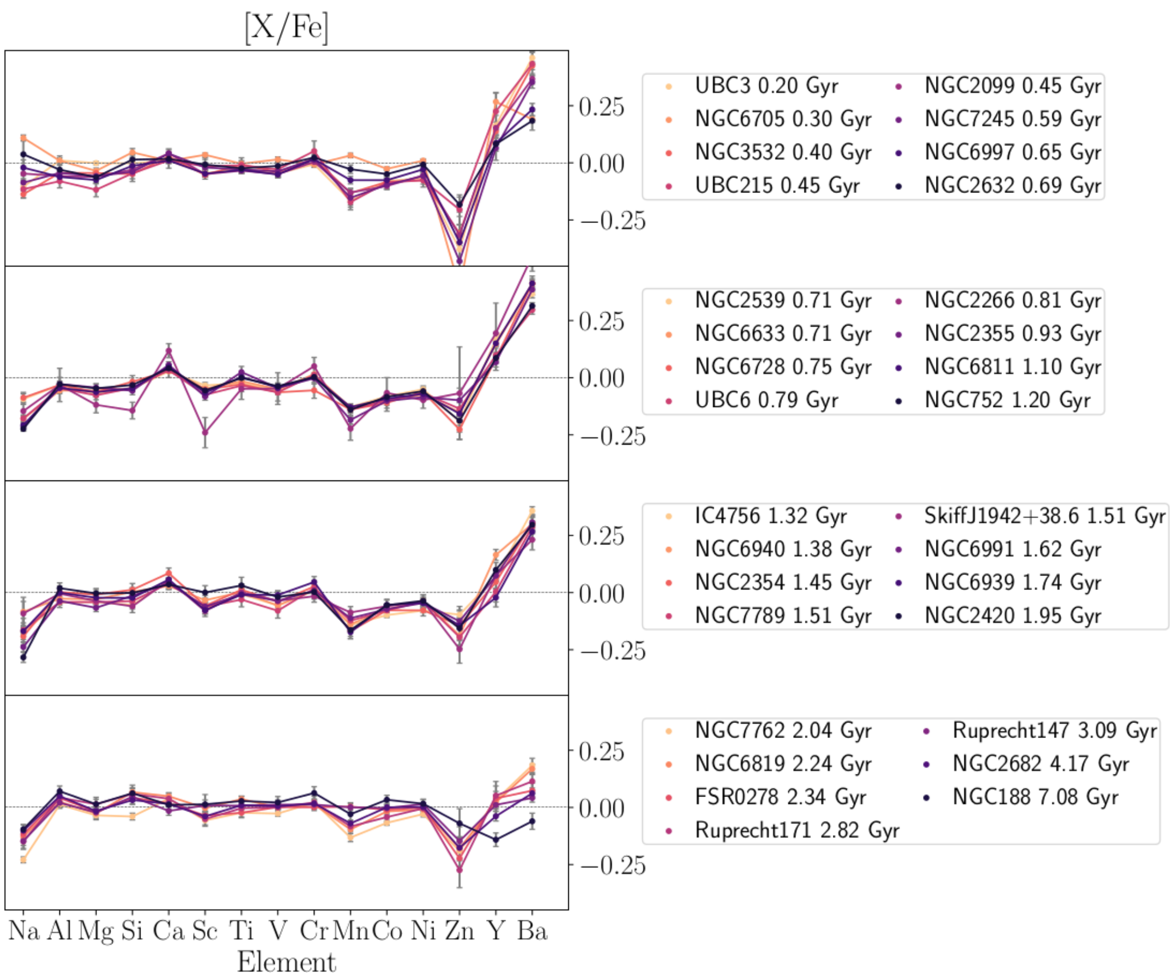



**Fig. 4.** Bottom part: one extremely metal-poor star spectra of RAVE. The metallicity of this star was estimated by our study to be [Fe/H] = −3.04 ± 0.22 dex and the measured S/N per pixel is ~52. The gray box marks the region around the central Ca II line for which the two solutions

# Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 **Idealised scenario**:
- All sample stars are open cluster members
- High-resolution (R>45,000), high SNR (>70) spectra
- Only red-clump stars → no abundance trends with temperature or gravity




Casamiquela+2016

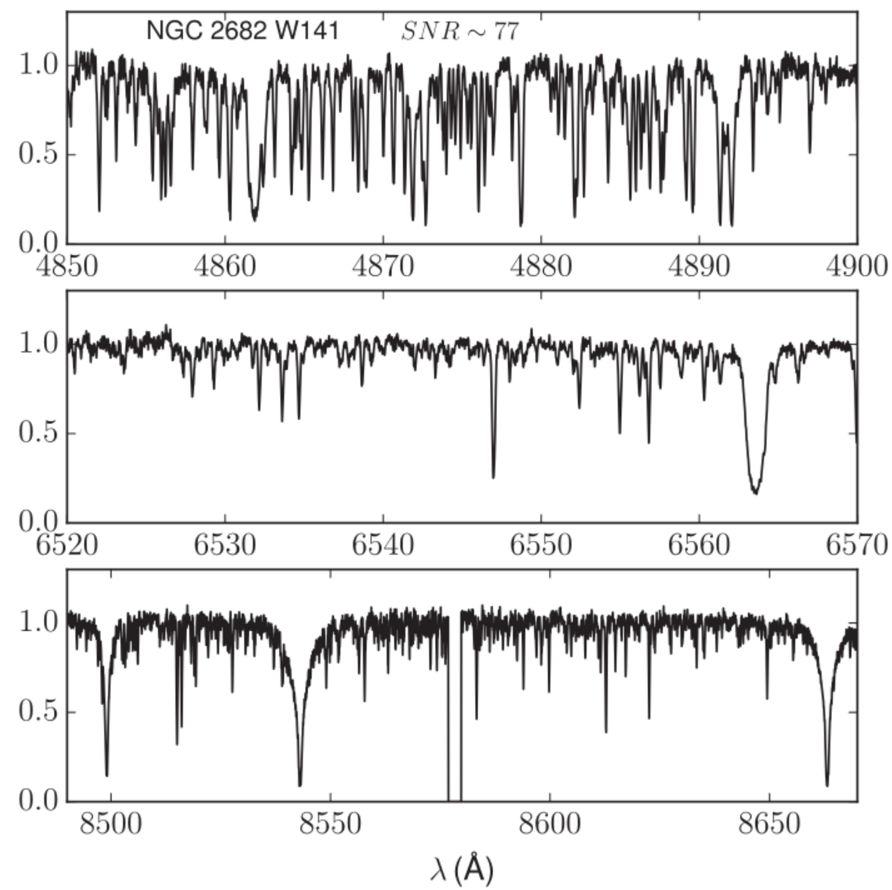
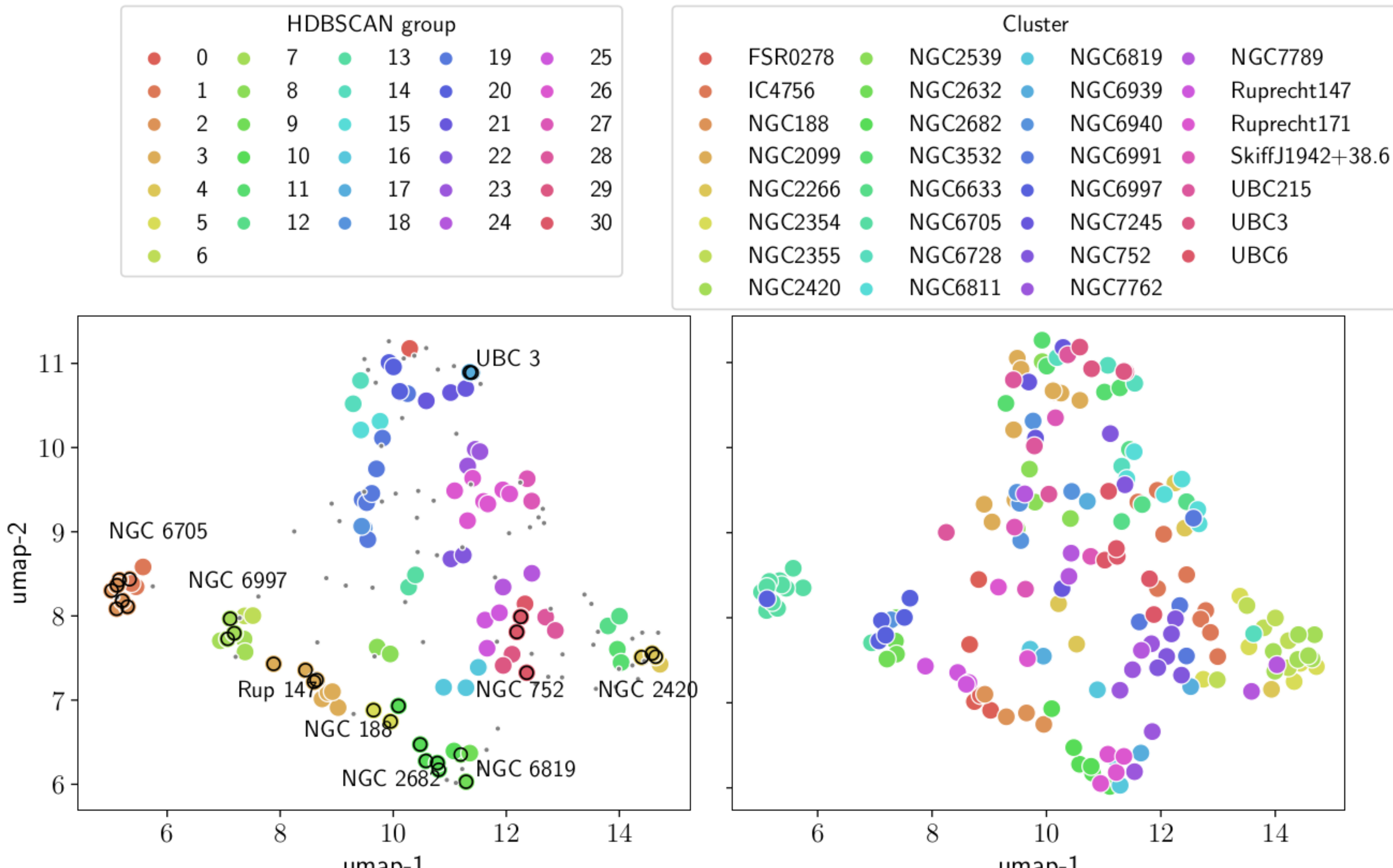



**Figure 3.** The Ca triplet (bottom), Hα (middle) and Hβ (top) regions of the final combined and normalized spectrum of the star NGC 2682 W141 observed with HERMES (SNR ∼ 77). A small gap from the order merging

# Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 **Idealised scenario**:

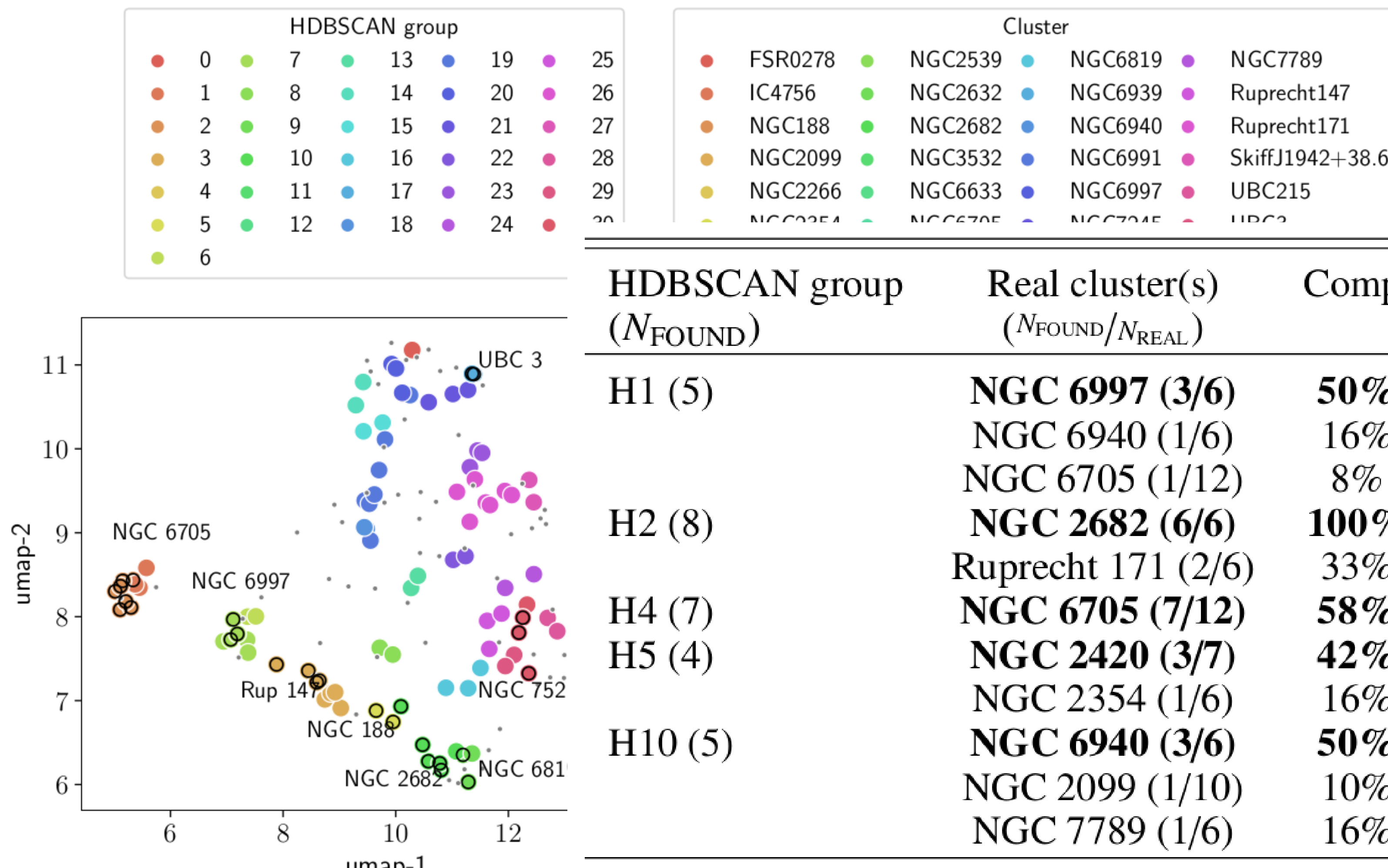
- HDBCSAN finds groups in abundance space → **no correspondence** with physical clusters
- umap (similar to t-SNE) only used for visualisation



# Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 **Idealised scenario**:

- HDBCSAN finds groups in abundance space → **never 1:1 correspondence** with physical clusters

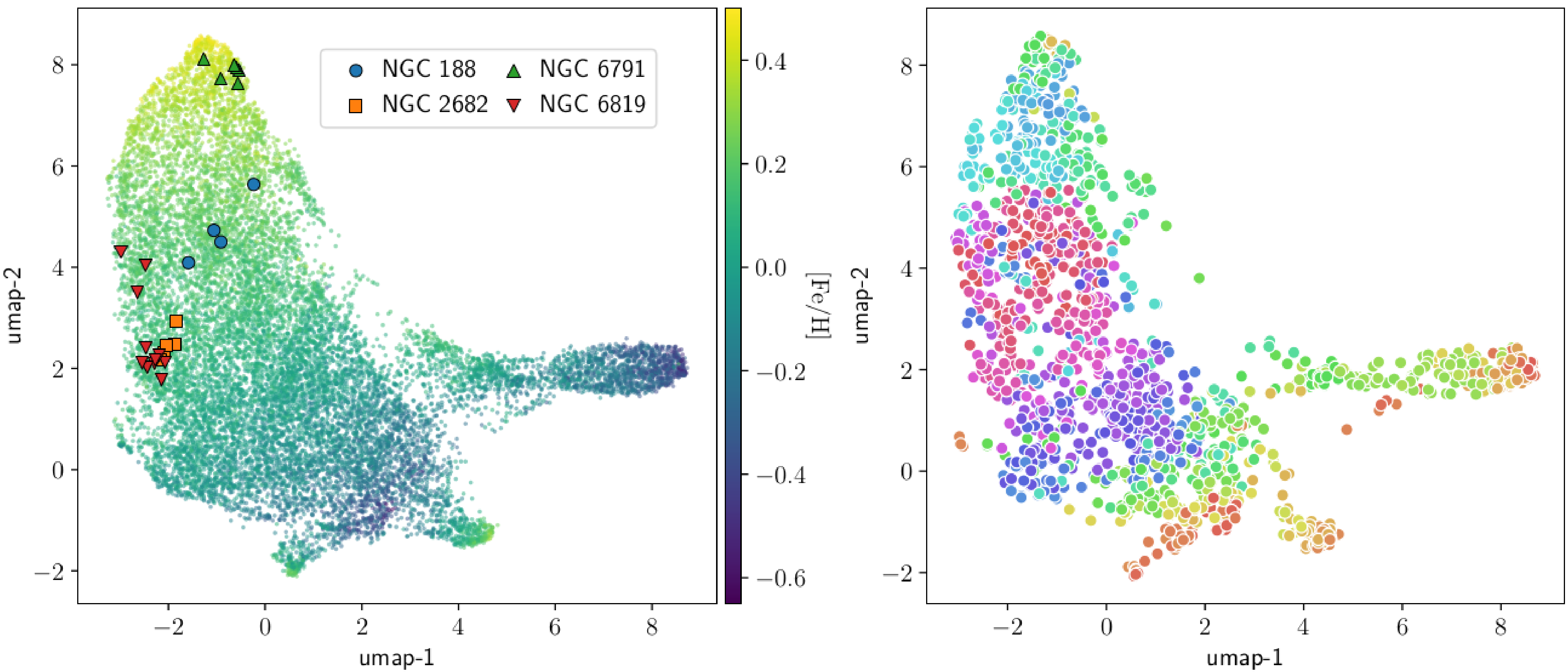



| HDBSCAN group ($N_{FOUND}$) | Real cluster(s) ($N_{FOUND}/N_{REAL}$) | Comp. | Hom. |
|---|---|---|---|
| H1 (5) | **NGC 6997 (3/6)** | **50%** | **60%** |
| | NGC 6940 (1/6) | 16% | 20% |
| | NGC 6705 (1/12) | 8% | 20% |
| H2 (8) | **NGC 2682 (6/6)** | **100%** | **75%** |
| | Ruprecht 171 (2/6) | 33% | 25% |
| H4 (7) | **NGC 6705 (7/12)** | **58%** | **100%** |
| H5 (4) | **NGC 2420 (3/7)** | **42%** | **75%** |
| | NGC 2354 (1/6) | 16% | 25% |
| H10 (5) | **NGC 6940 (3/6)** | **50%** | **60%** |
| | NGC 2099 (1/10) | 10% | 20% |
| | NGC 7789 (1/6) | 16% | 20% |

# Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 More realistic scenario (APOGEE survey):

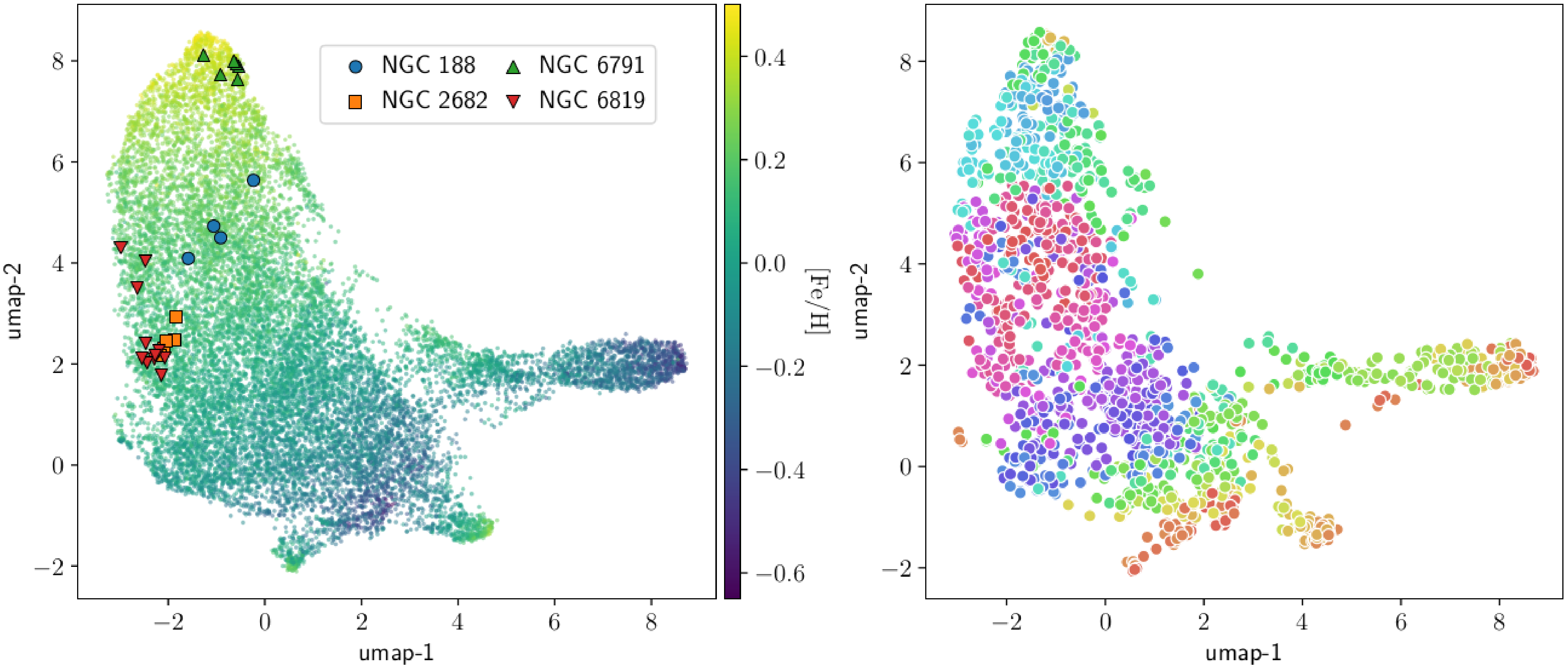
- As expected, HDBSCAN finds groups, but the physical correspondence is even worse



# Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 More realistic scenario (APOGEE survey):

- As expected, HDBSCAN finds groups, but the physical correspondence is even worse



# Summary

| Nature/Type | Classification | Regression | Clustering | Forecasting | Generation | Discovery | Insight |
|---|---|---|---|---|---|---|---|
| Image | ● | ● | ● | ● | ● | ● | ● |
| Spectroscopy | ● | ● | ● | | | ● | ● |
| Photometry | ● | ● | ● | ● | | ● | ● |
| Light curve | ● | ● | | | | ● | ● |
| Time Series | ● | ● | ● | | | ● | ● |
| Catalogue | ● | ● | ● | ● | | ● | ● |
| Simulation | ● | ● | | | ● | ● | |

# Summary

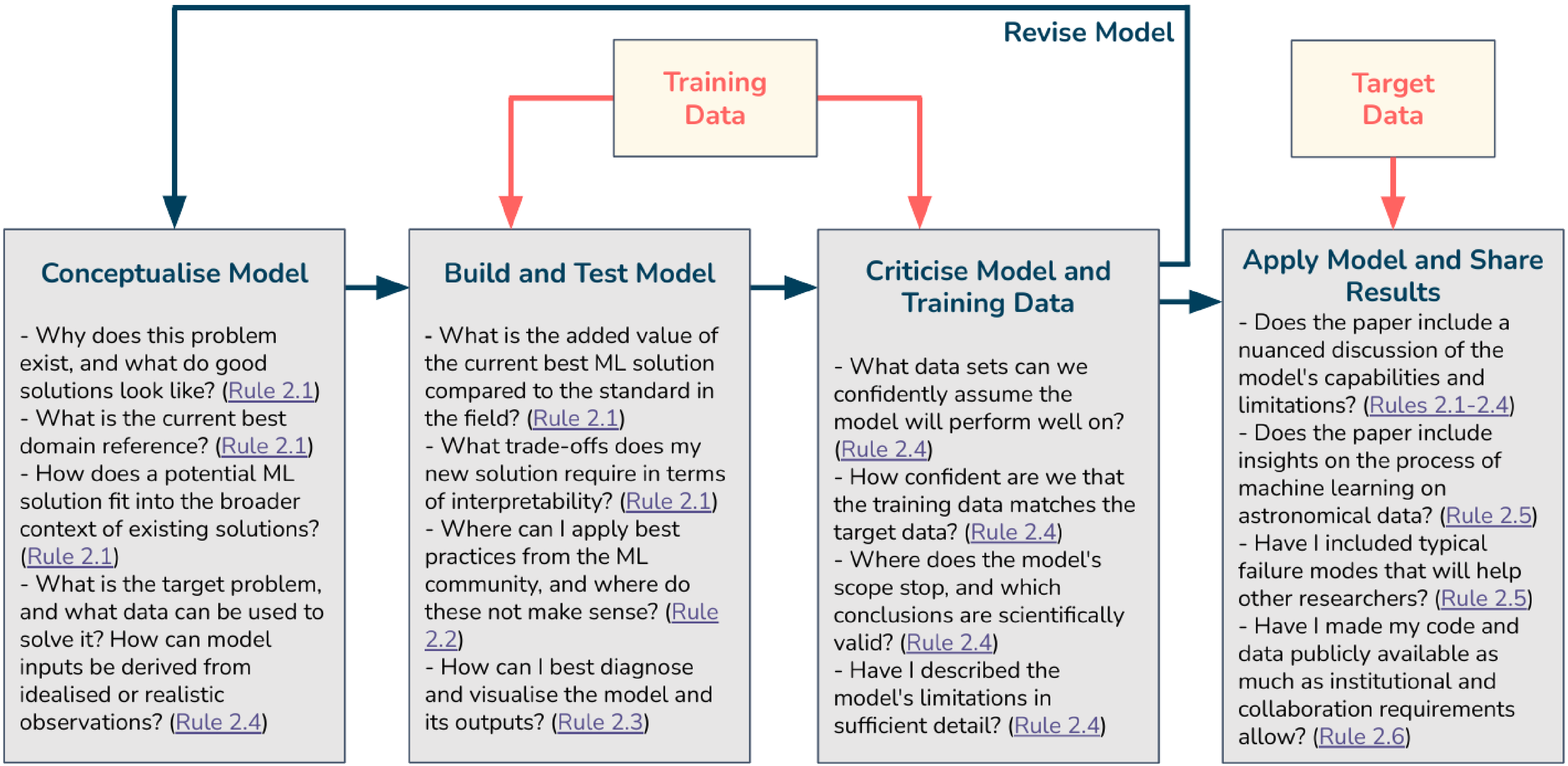



**Figure 2.** Box's loop for ML in astronomy.

Huppenkothen+2023

# Summary

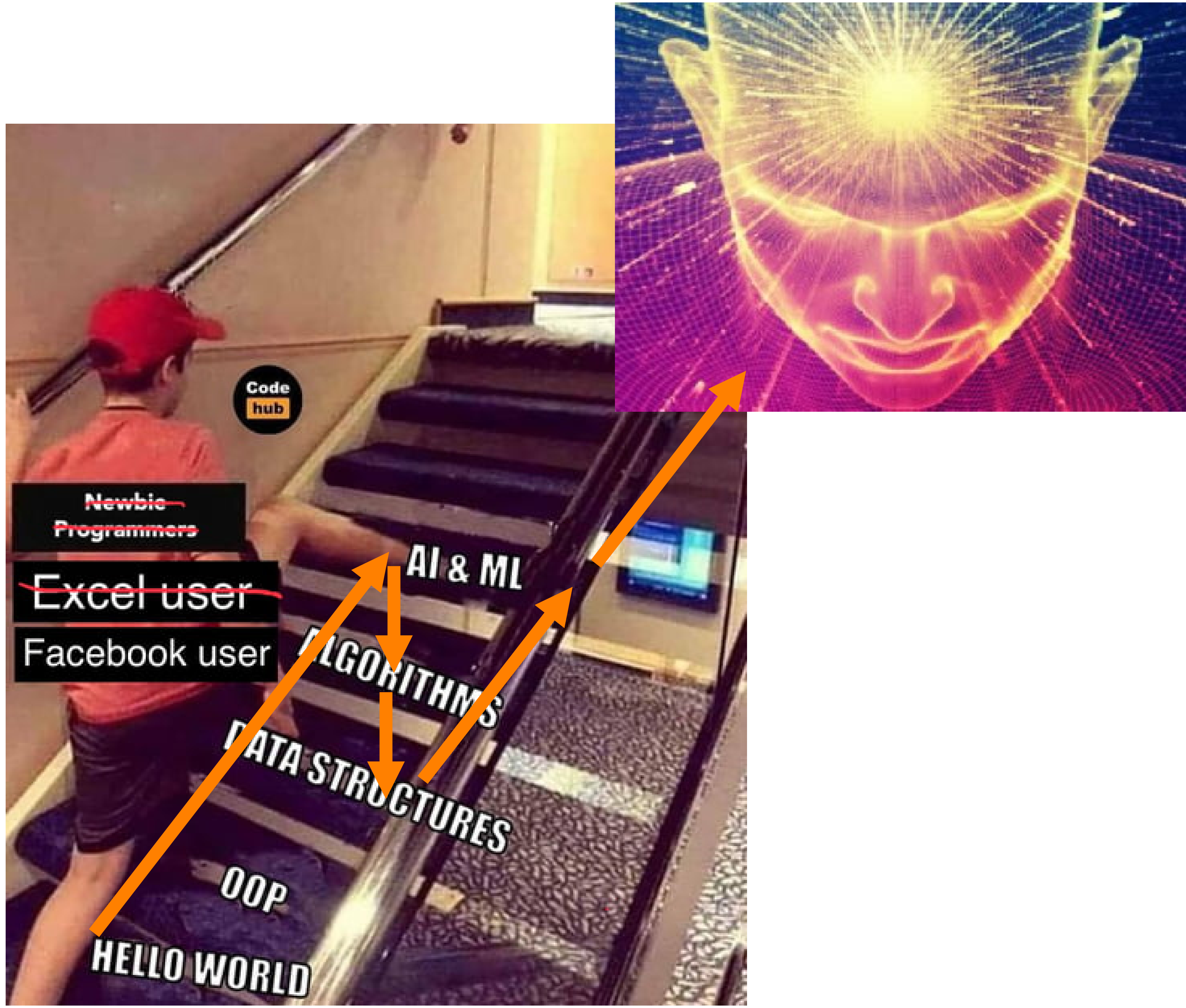
## A. A QUICK-START GUIDE FOR ASSESSING ML ASTRONOMY RESEARCH

Interdisciplinary results can be difficult to assess because they require a deep understanding not only of the scientific domain, but also of the methodology. This quick-start guide is intended as a starting point for readers and referees to assess new research for which they have a domain understanding but may lack methodological context. For referees, we caution that this guide is not absolutely prescriptive, nor exhaustive. Referees should consult journal expectations (e.g. American Astronomical Society 2022; MNRAS 2022; Nature 2022) and more general refereeing references (Wager et al. 2002; Nicholas & Gordon 2011; Raff 2013; Ntampaka et al. 2022) for guidelines on best practices for providing evaluations of manuscripts. The following considerations do not replace refereeing best practices; instead, they are additions to best practices that are specific to evaluating ML astronomy research.

1. Compare against a domain reference and put results in the broader context.
   - (a) Are the results put in the appropriate context? For example, if the method replaces an existing "traditional" technique, are results (accuracy, compute cost, robustness, etc.) from the traditional technique used as a comparison?
   - (b) In some cases, the new ML method enables an analysis that was not possible before; no traditional benchmark exists. In this case, it may be more difficult to put the results in context.
   - (c) If the outputs of the model are likely to be used downstream (e.g. for population-level analyses), do the authors consider how biases in their model might propagate into these analyses?
2. Adopt best practices from the ML community.
   - (a) Have the authors included citations in their literature review to summarize particular best practices applied in their work?

Huppenkothen+2023

# „Machine learning in astronomy“


Code hub
~~Newbie Programmers~~
~~Excel user~~
Facebook user
AI & ML
ALGORITHMS
DATA STRUCTURES
OOP
HELLO WORLD