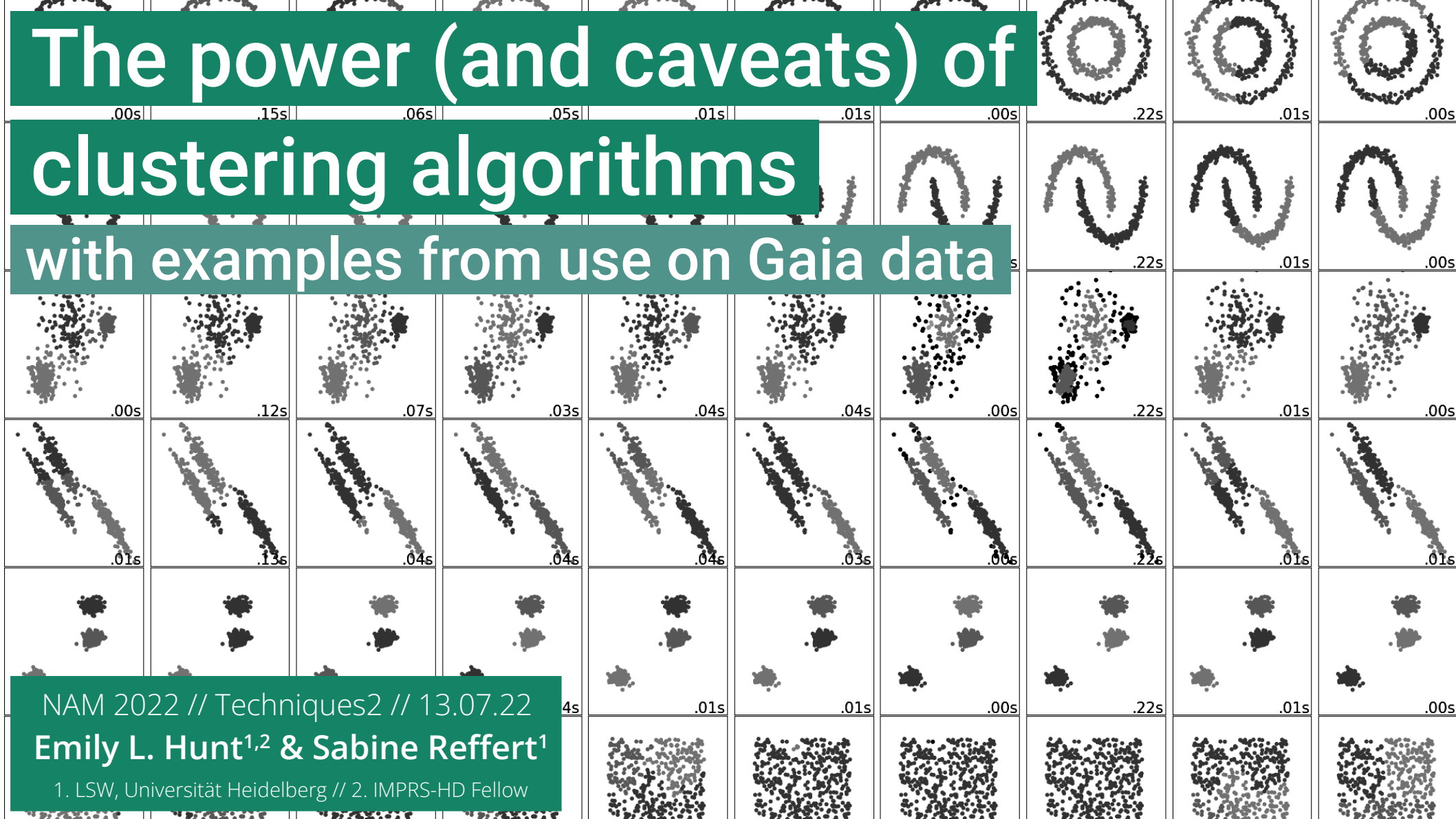
The power (and caveats) of clustering algorithms

with examples from use on Gaia data



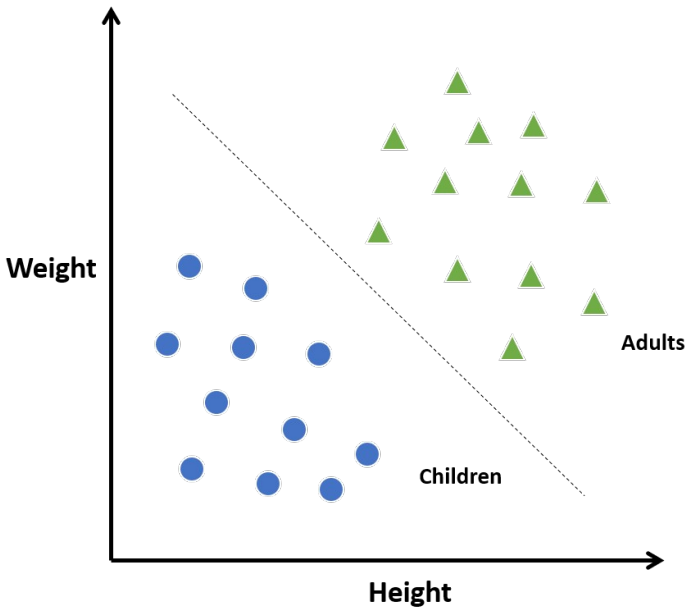
NAM 2022 // Techniques2 // 13.07.22

Emily L. Hunt[1,2] & Sabine Reffert[1]

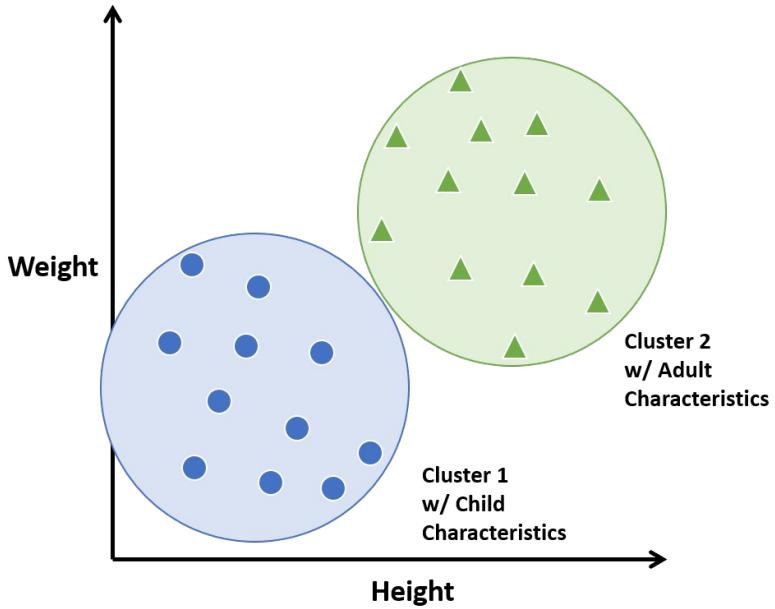
1. LSW, Universität Heidelberg // 2. IMPRS-HD Fellow

what is *unsupervised* ML?

Classification in supervised ML is about making a **decision boundary** between different regions



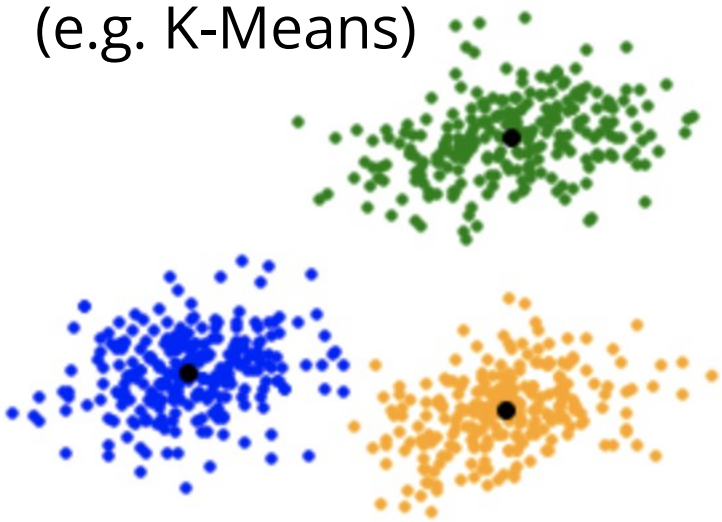
Clustering algorithms (a type of unsupervised ML) do this using **properties of the data itself**



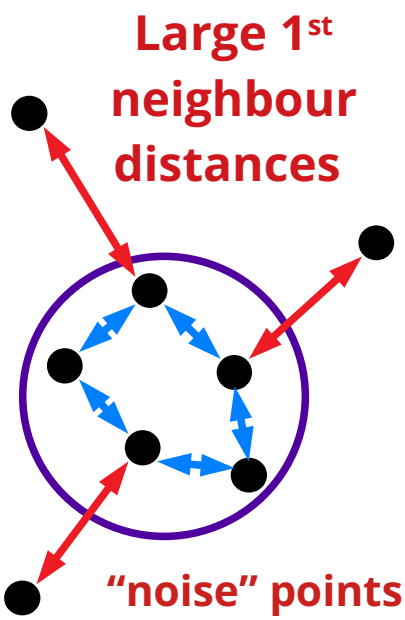
the many, many algorithms...

There are many different types, each suited to different problems

Partitioning
(e.g. K-Means)



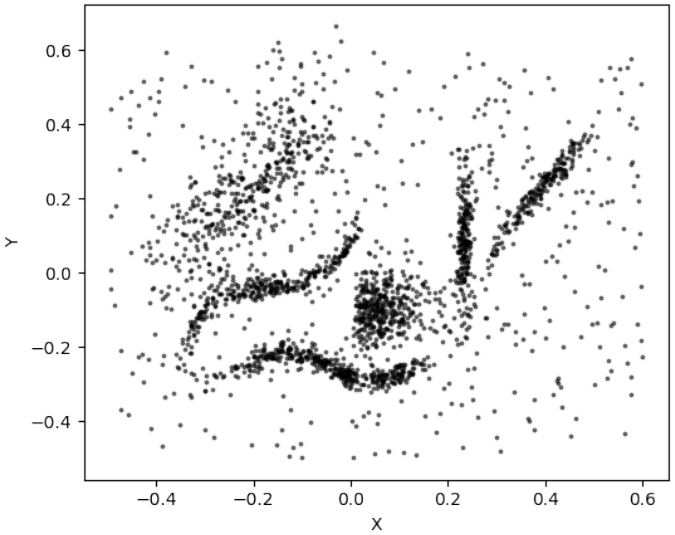
Density-based clustering
(e.g. DBSCAN)



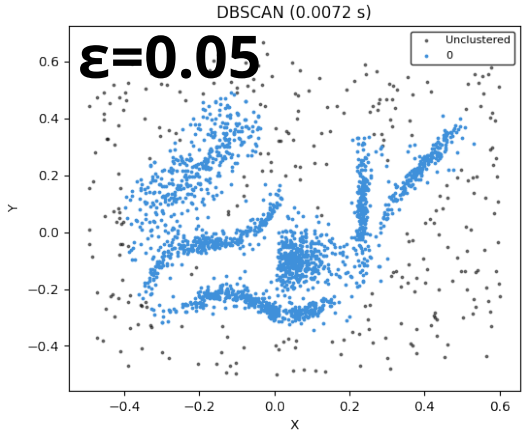
Link: Overview of various algorithms

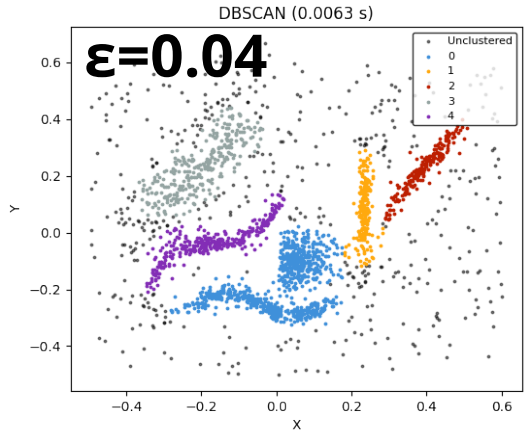
the hard part: setting parameters.

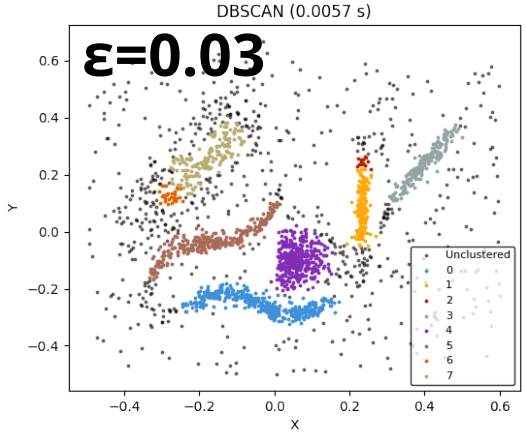
For some basic data...

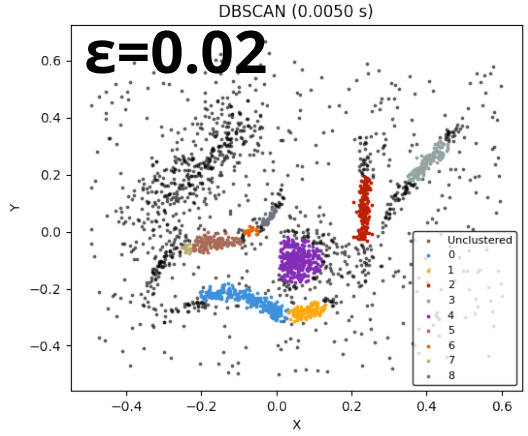


...what ε and m_{Pts} to use? (this is DBSCAN)









play around with this in a notebook!

github.com/emilyhunt/nam_2022_talk

(also includes these slides)

4. The clustering algorithms

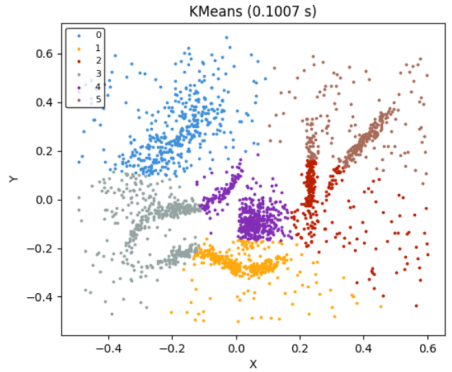
Let's try various algorithms! Feel free to play around with the parameters.

4.1. K-Means

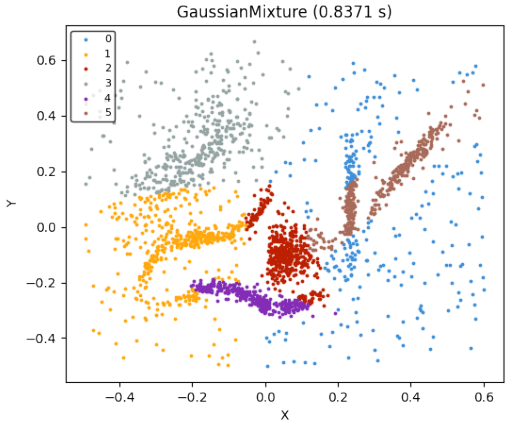
The archetypal (and most simple) clustering algorithm...

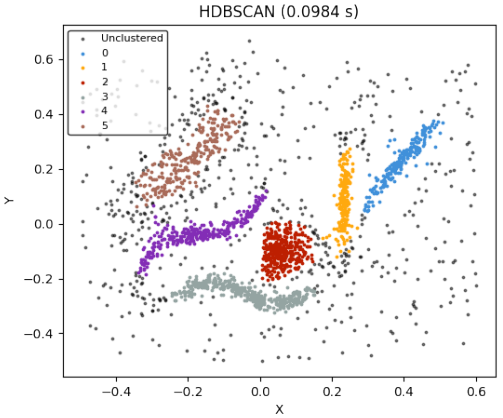
```
[118]: run_clustering_algorithm(data, sklearn.cluster.KMeans, n_clusters=6, savename='KMeans')
```

```
Running clustering algorithm!
  clustering took 0.1007 seconds
Plotting results!
```



Basic algorithm; some things are roughly separated, but not so well.

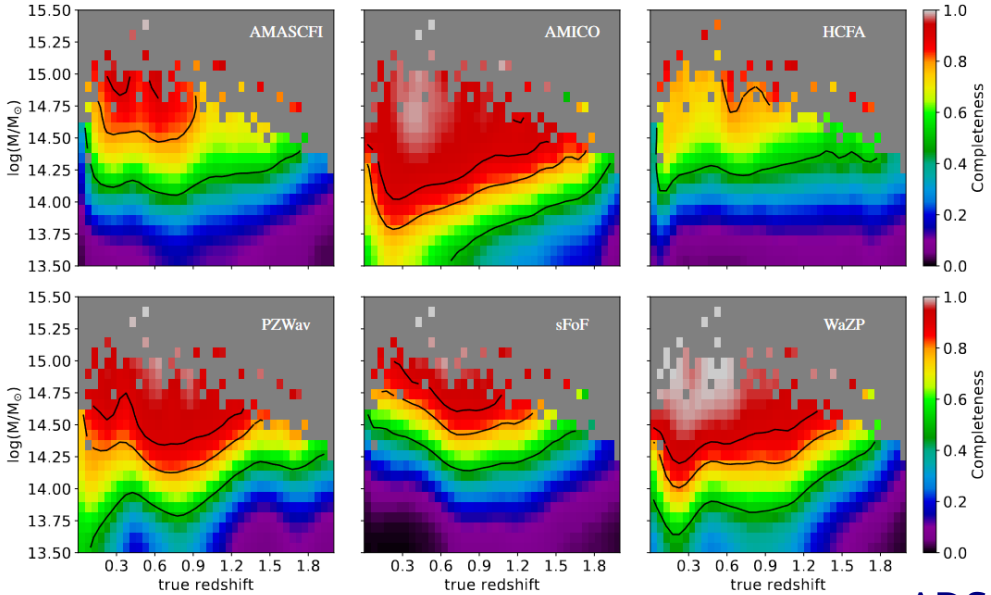




many astro problems come in clusters.

***Euclid* preparation III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection**

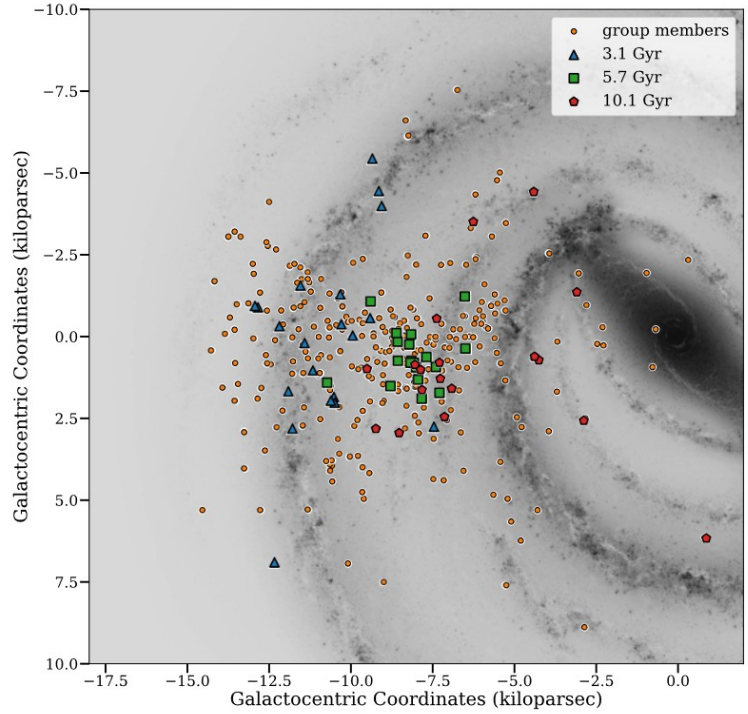
Euclid Collaboration, R. Adam[1,2,3⋆], M. Vannier[2], S. Maurogordato[2], A. Biviano[4], C. Adami[5], B. Ascaso[6],



ADS

Strong chemical tagging with APOGEE: 21 candidate star clusters that have dissolved across the Milky Way disc

Natalie Price-Jones[1,2⋆], Jo Bovy[1,2], Jeremy J. Webb[1], Carlos Allende Prieto[3,4],



ADS

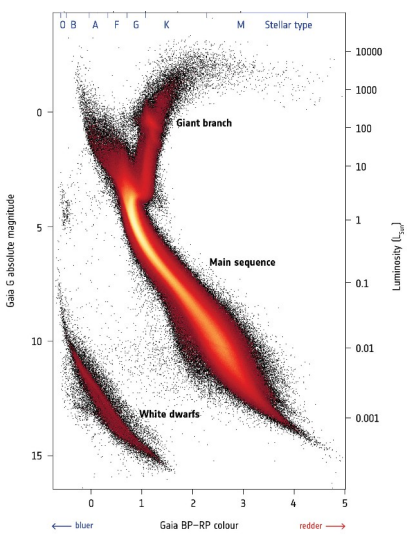
our problem: open clusters of stars.

The Pleiades, age: ~100 Myr

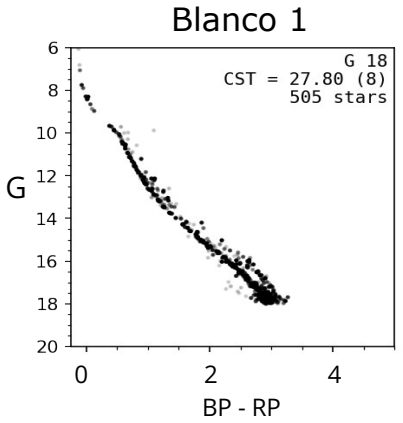
Credit: ESO

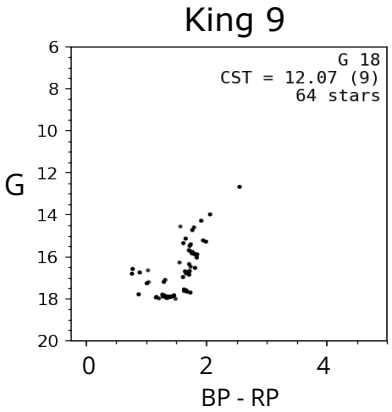
Open clusters are **homogenous** and **bound** groups of **coeval** stars

Highly useful to stellar & galactic science!



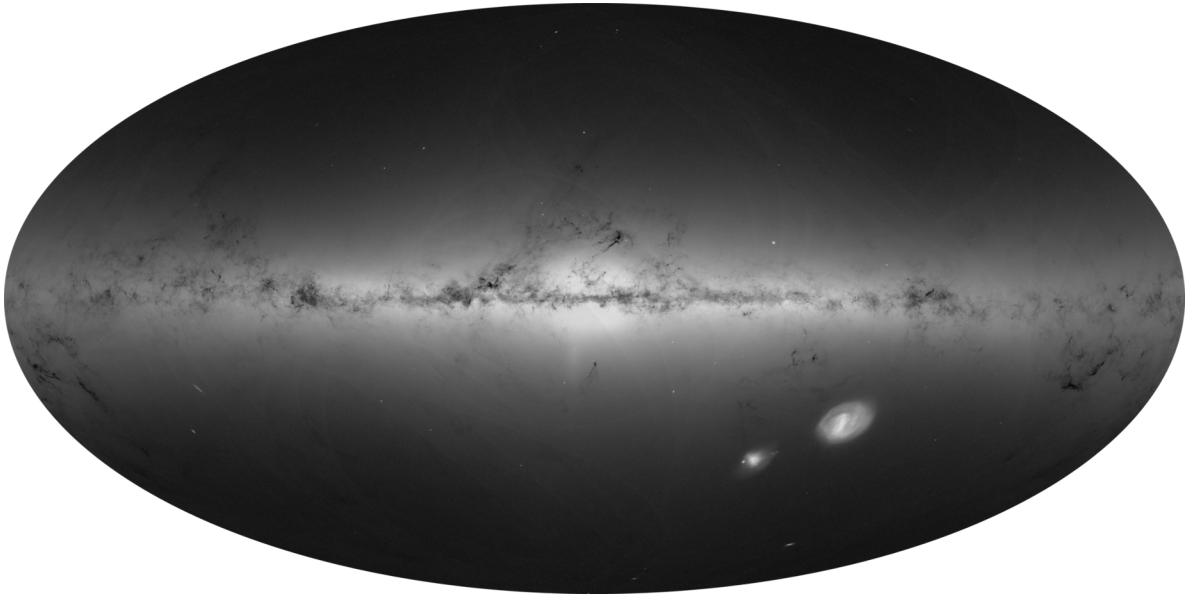
Credit: ESA





Gaia data is a challenge.

Gaia satellite has reliable astrometry for **~1 billion** stars
=> **~0.1%** of which in **open clusters**

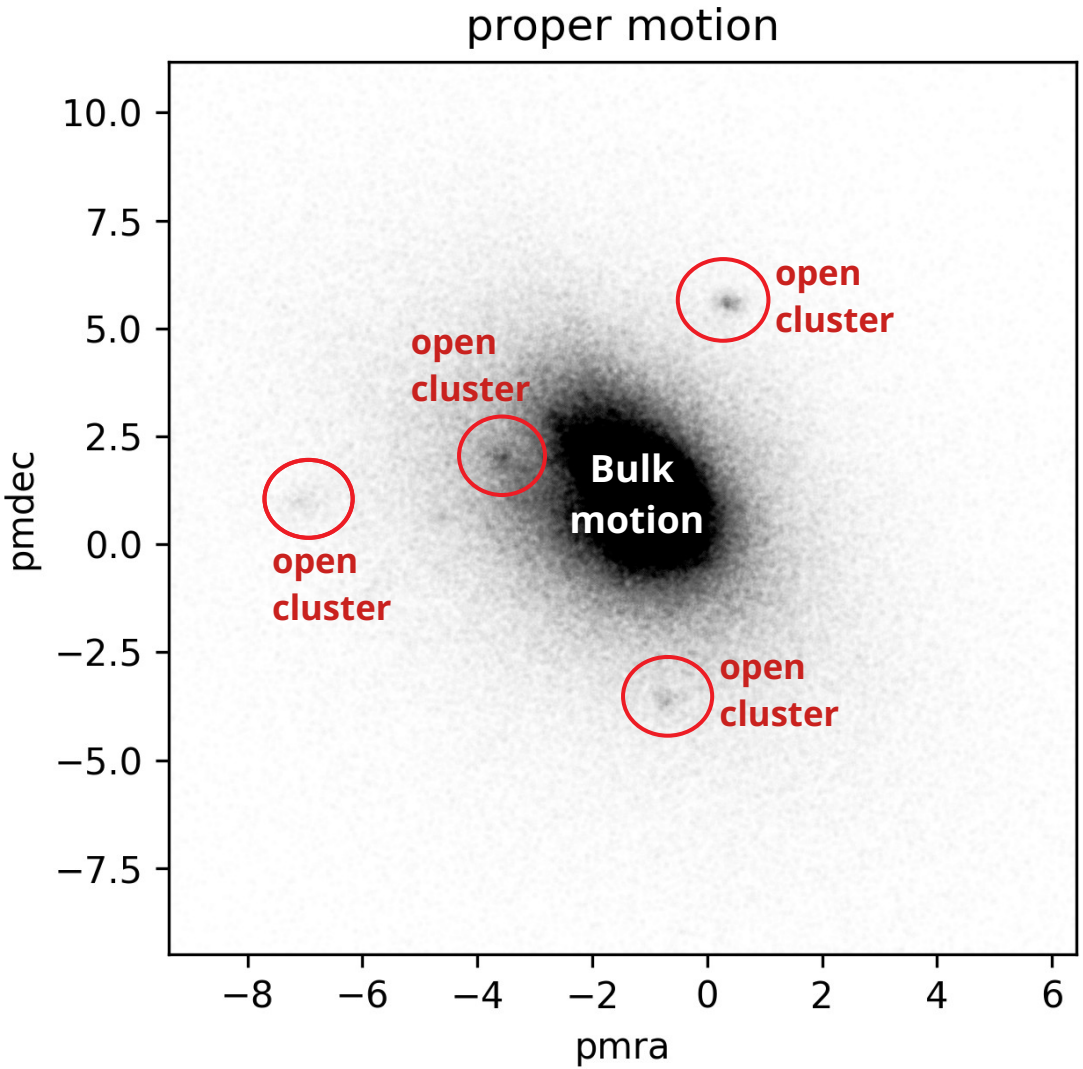


Gaia source density skymap. Credit: ESA

Some challenges:

- **So many sources**...
- **~99.9%** of which must be discarded!
- Open clusters vary from **~0.1°** to **~10°** in size

but... Gaia is still *perfect* for open clusters!

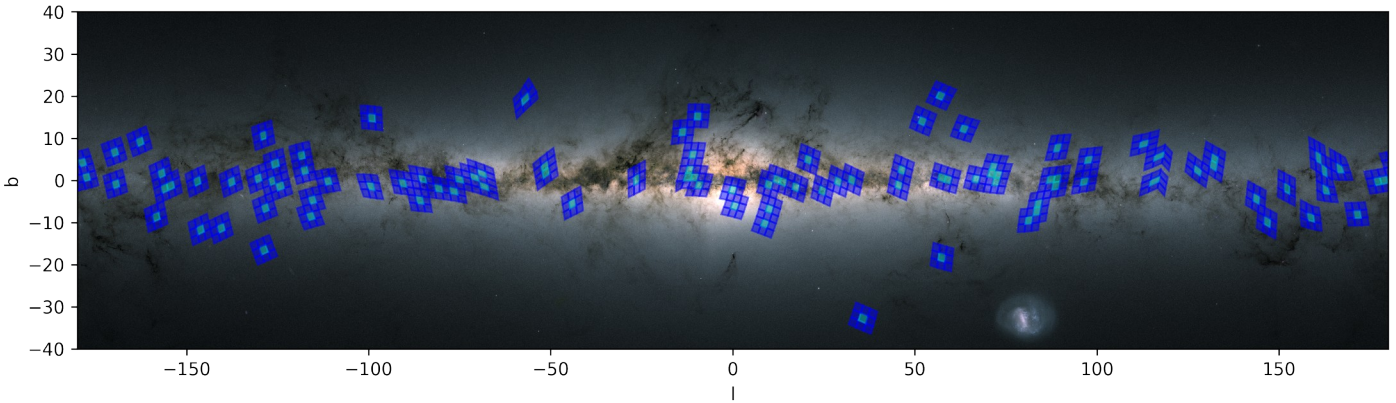


testing different algorithms.

Hunt & Reffert (2021): search for literature clusters in randomly selected fields

After initial trials of what worked best…

- Algorithms given **rescaled** positions, proper motions and parallaxes (5 dimensions) + corrected for spherical distortions
- Paper used **DBSCAN**, **HDBSCAN** and **Gaussian mixture models**



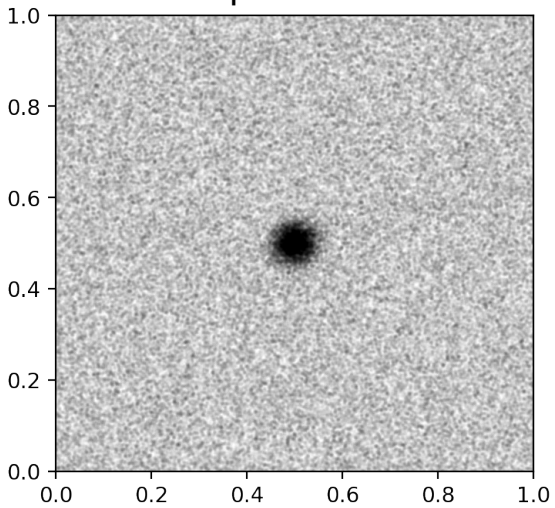
which one is best?

Algorithm	Speed	Sensitivity TP / (TP + FN)	Precision TP / (TP + FP)	
DBSCAN (Castro-Ginard+18 parameters)	**Fast**	0.53	**1.00**	← **Main literature approach**
DBSCAN (My parameters)	**Fast**	0.62	0.93	
HDBSCAN	Quite fast	**0.82**	**0.82**	← **My favourite**
Gaussian Mixtures	Slow	0.33	**1.00**	

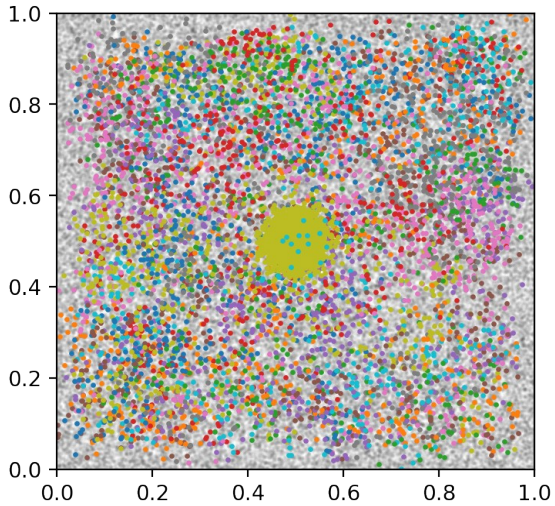
let's talk about false positives.

The **most sensitive** clustering methods can also produce **the most false positives**. Care **must** be taken when using "off-the-shelf" methods

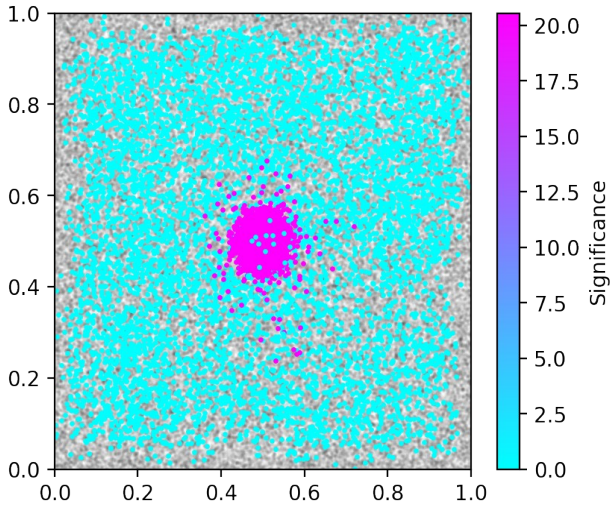
100,000 uniformly distributed 5D points + a 5000 point Gaussian



HDBSCAN, m_{clSize}=20…
151 clusters?!



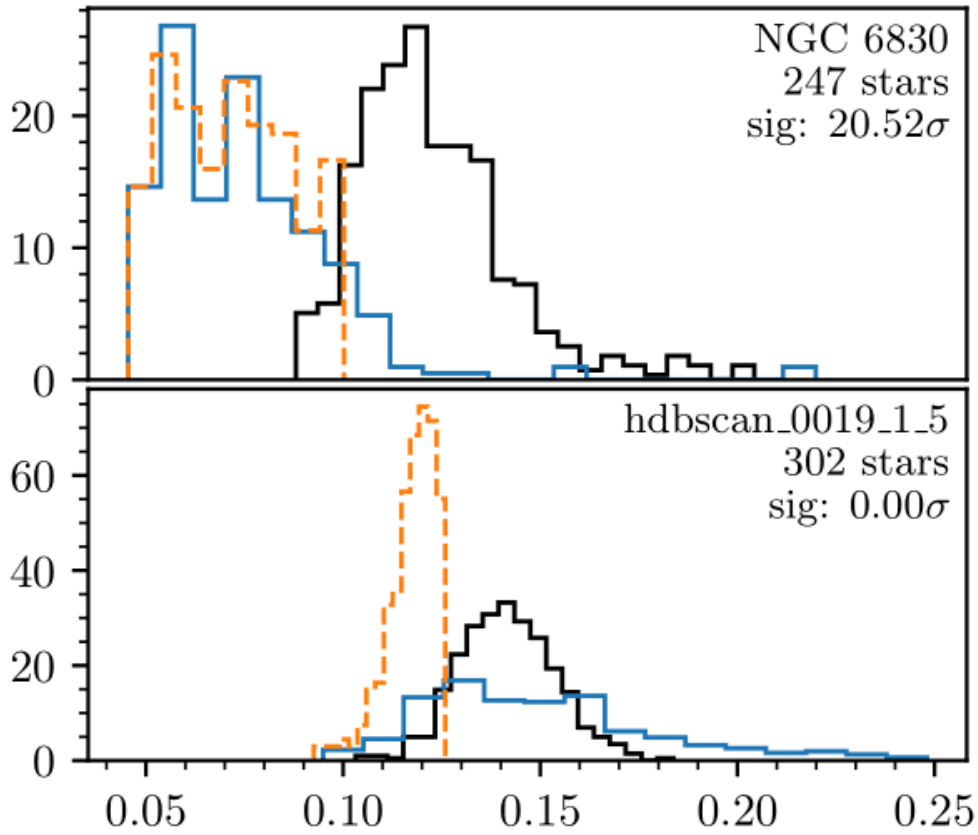
Density test → only the real cluster is significant



how the significance test works.

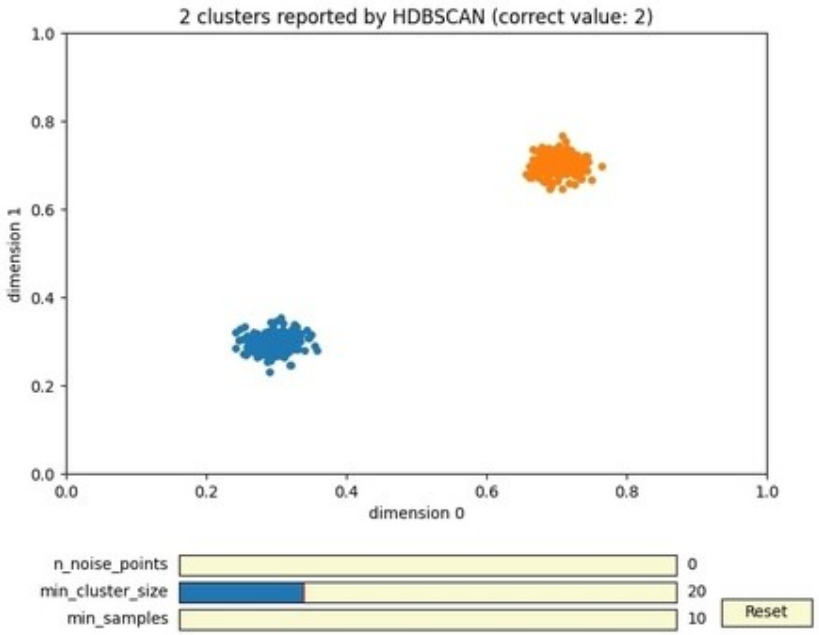
Simple density test compares **nearest neighbour distances** in cluster with field

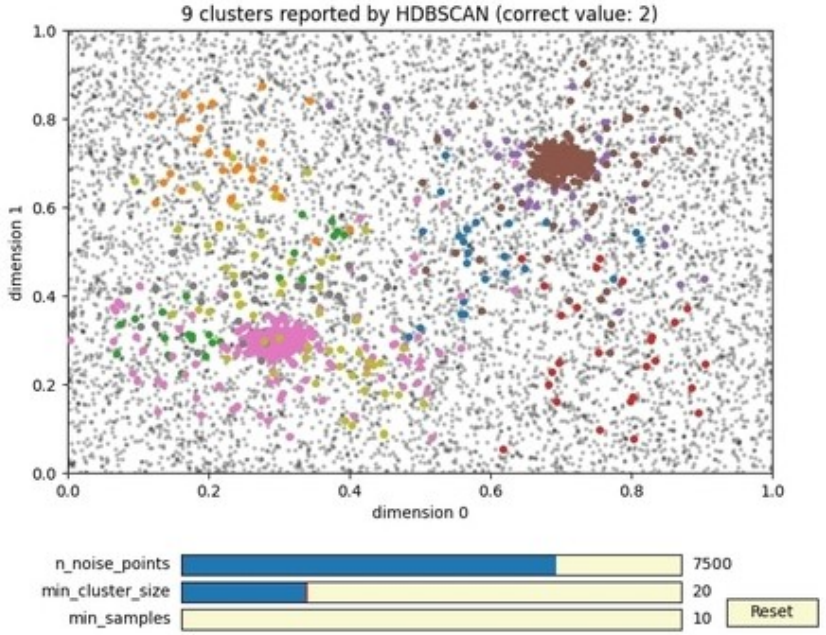
Can tell **random chance associations** and **real clusters** apart



a way to try this in real time!

I made a script that shows this in real time! (GitHub link)



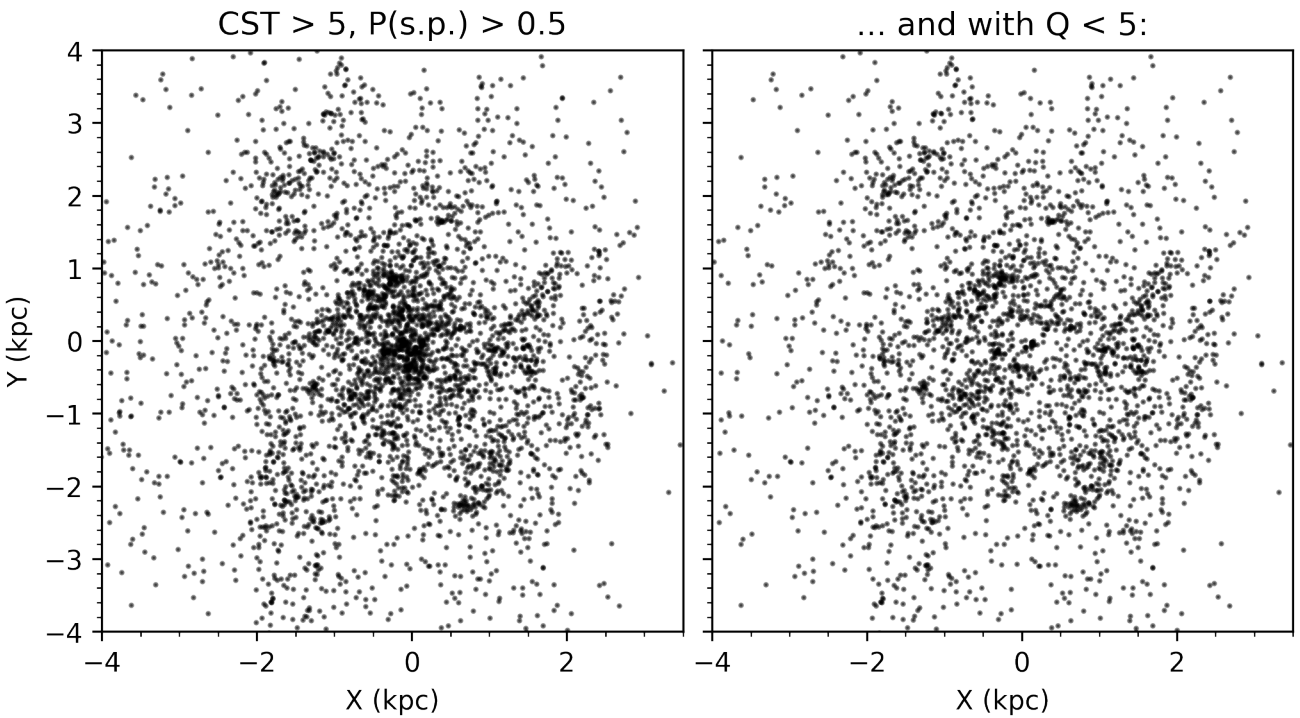


clustering 729 million stars in Gaia EDR3.

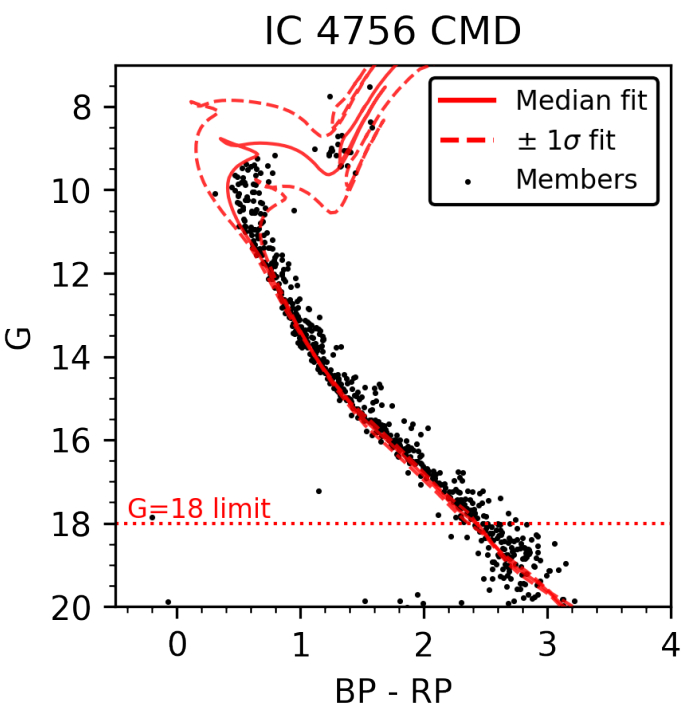
- Ran **HDBSCAN** on largest-ever sample of stars
- Used **HEALPix** tesselation scheme for regions
- About **8 days of wall time** on a powerful machine (actually not bad)

after much work: results.

Distribution, |z| < 500 pc:



Example CMD:



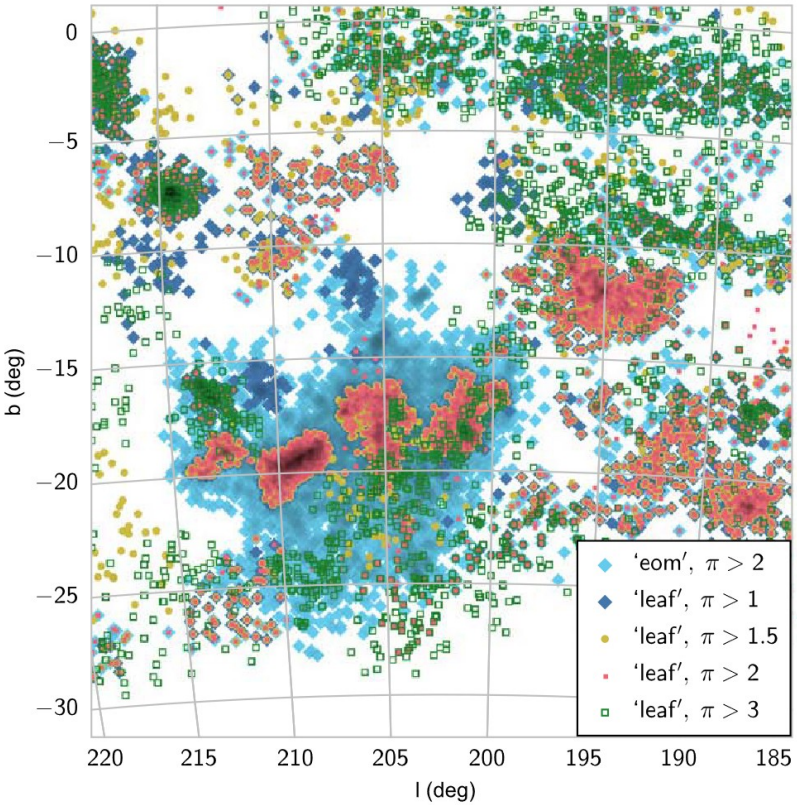
and a cautionary tale on false positives...

Kounkel+19,20: reported **thousands** of groups and 'strings' of stars using **HDBSCAN** on Gaia DR2 data

(+ without using a significance test to clean HDBSCAN results)

In EDR3: we re-detect just **18.1%** of their groups (with better data!)

Also echoes **Zucker+22:** "many Kounkel+ group members inconsistent with having common origin"



where next for clustering algorithms?

- Clustering algorithms are **fantastic** tools for astronomy
- Many off-the-shelf algorithms available (e.g. scikit-learn)
- BUT: they aren't designed for astronomy problems, and can run into weird issues

In the future: astro must collaborate with computer scientists, mathematicians etc. to develop algorithms

datasets to come...

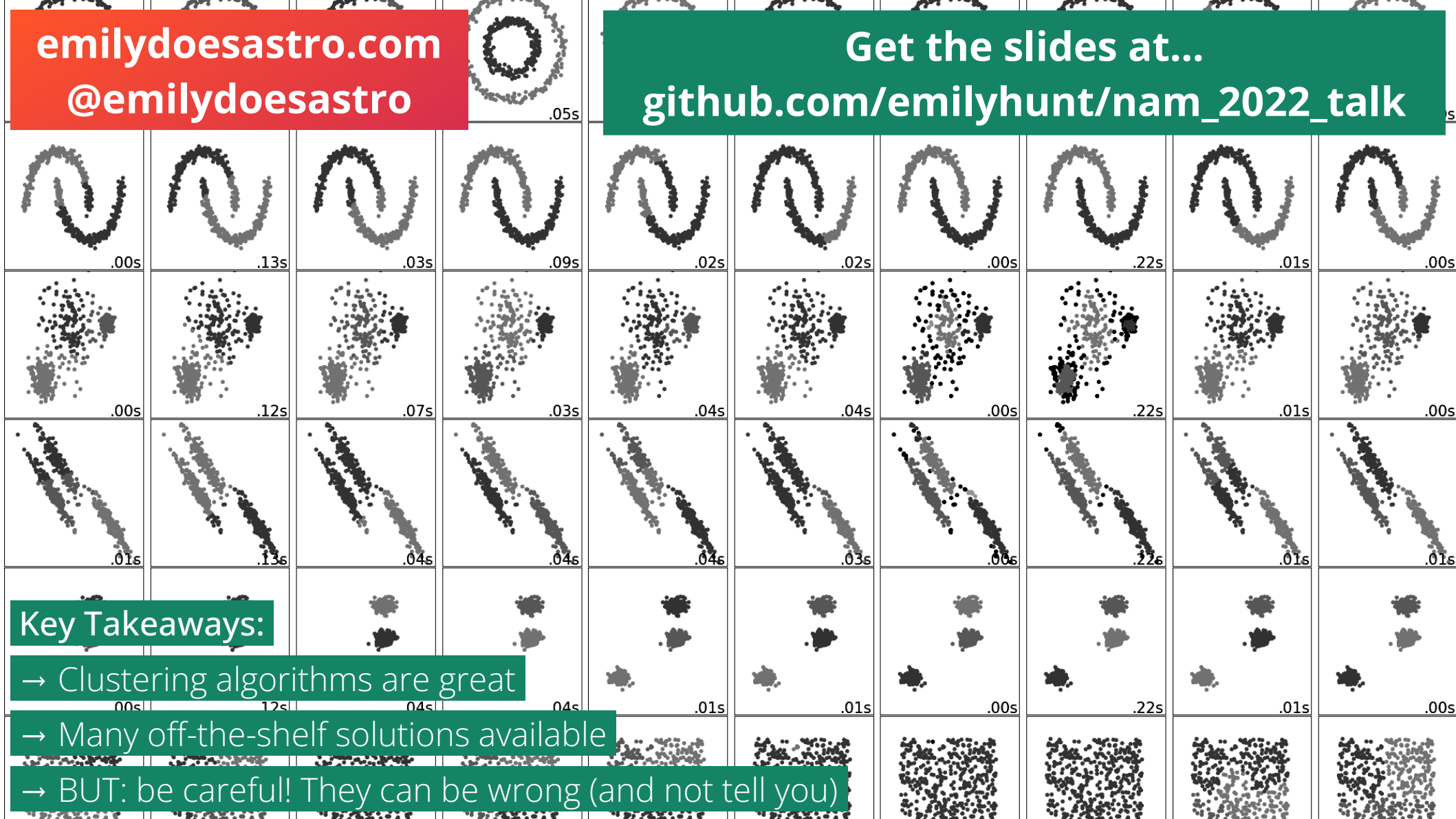
Gaia DR5 (~2030): 2 billion stars

Euclid (late 2020s): ~1 billion galaxies with redshifts

LSST (final data ~2035): 17 billion stars, 20 billion galaxies

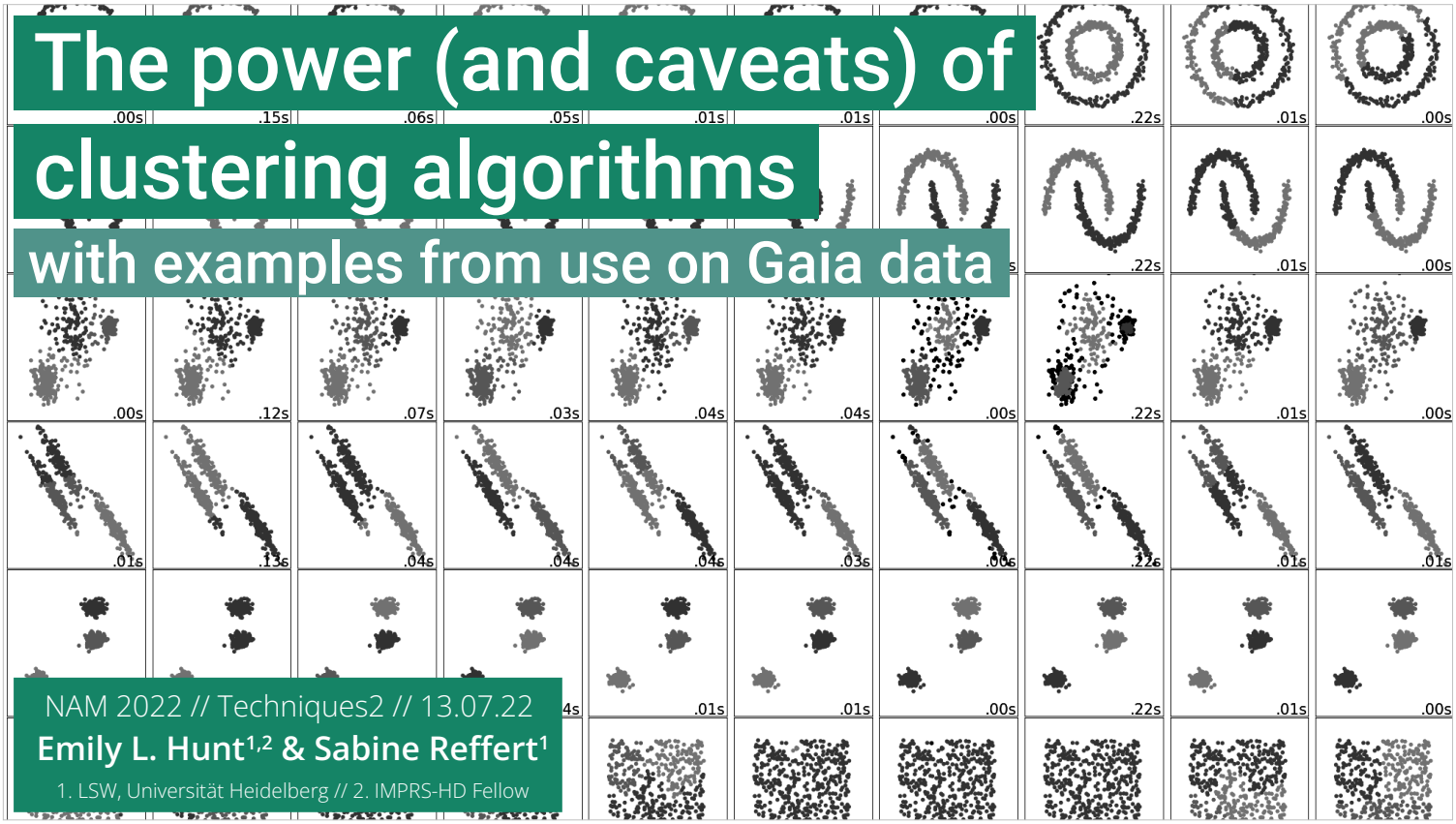
GaiaNIR (~2050): 12 billion stars

+ many, many other uses across astronomy for clustering algorithms


emilydoesastro.com
@emilydoesastro
Get the slides at...
github.com/emilyhunt/nam_2022_talk
Key Takeaways:
→ Clustering algorithms are great
→ Many off-the-shelf solutions available
→ BUT: be careful! They can be wrong (and not tell you)

The power (and caveats) of clustering algorithms

with examples from use on Gaia data



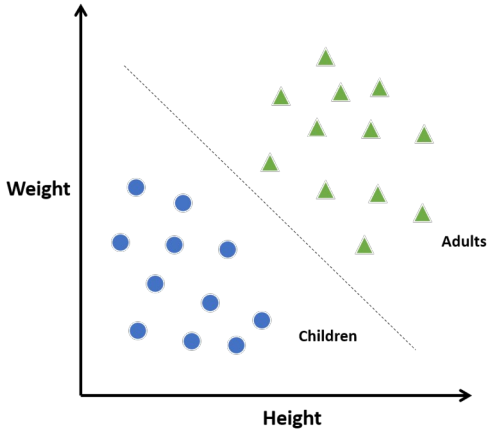
NAM 2022 // Techniques2 // 13.07.22

Emily L. Hunt[1,2] & Sabine Reffert[1]

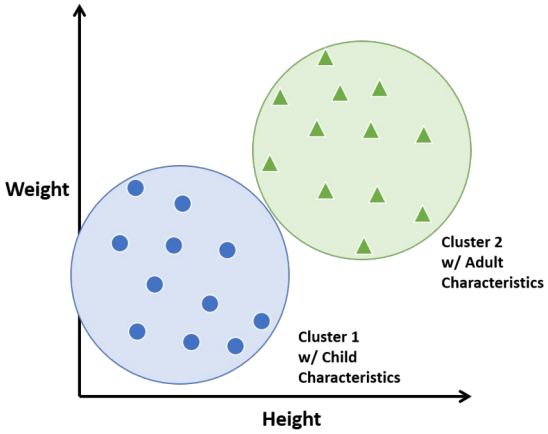
1. LSW, Universität Heidelberg // 2. IMPRS-HD Fellow

what is *unsupervised* ML?

Classification in supervised ML is about making a **decision boundary** between different regions



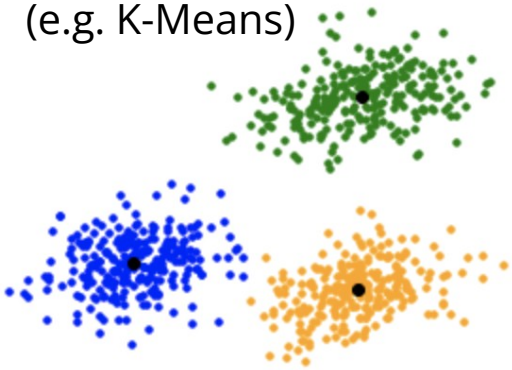
Clustering algorithms (a type of unsupervised ML) do this using **properties of the data itself**



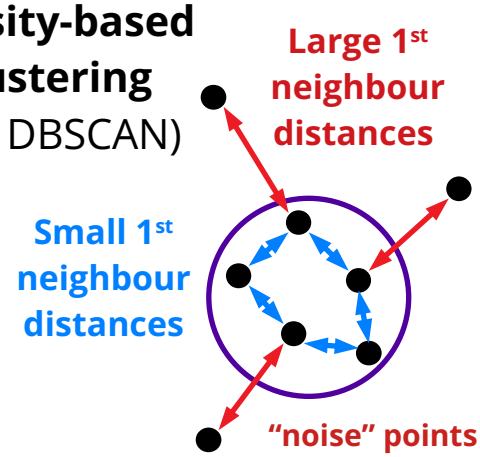
the many, many algorithms...

There are many different types, each suited to different problems

Partitioning
(e.g. K-Means)



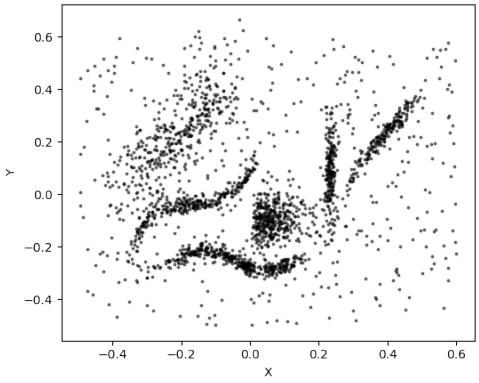
Density-based clustering
(e.g. DBSCAN)



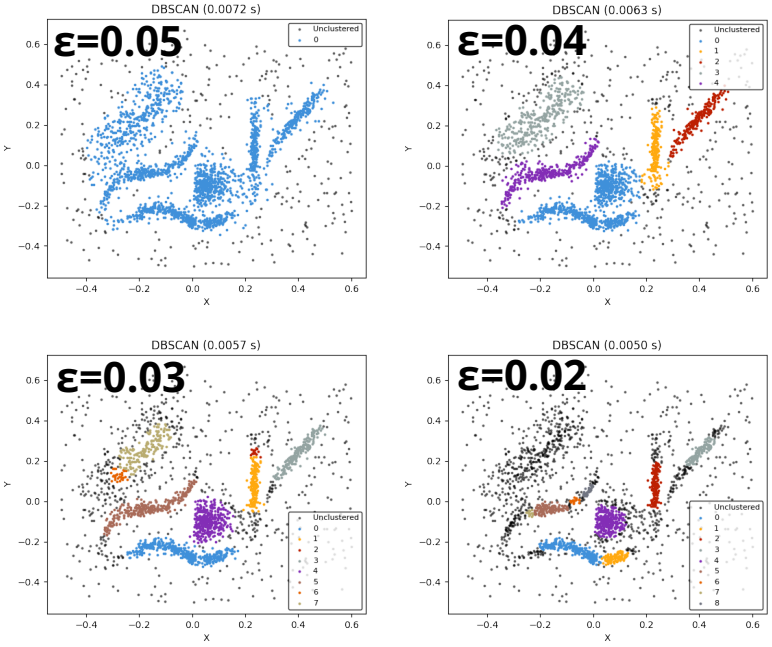
Link: Overview of various algorithms

the hard part: setting parameters.

For some basic data...



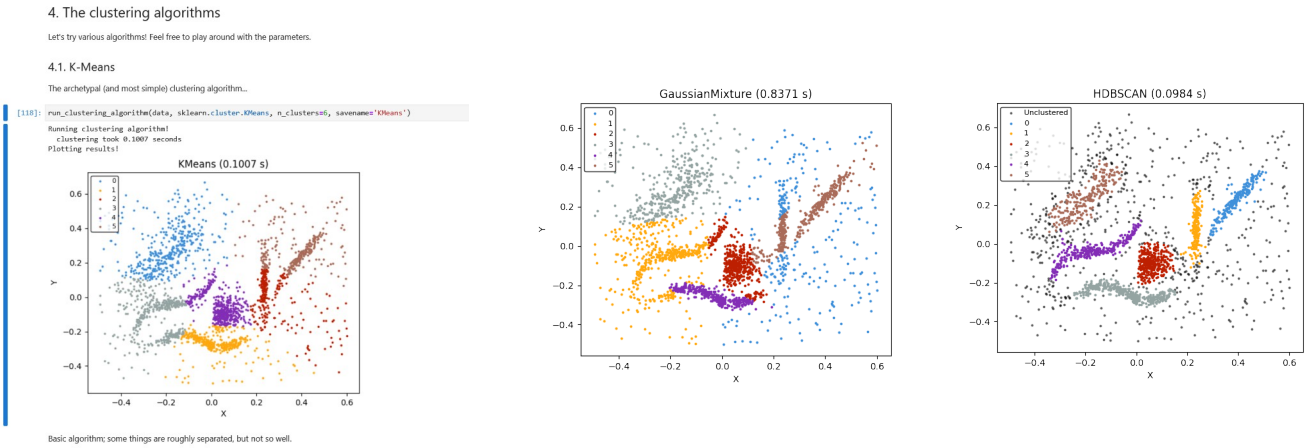
...what ε and m_{Pts} to use? (this is DBSCAN)



play around with this in a notebook!

github.com/emilyhunt/nam_2022_talk

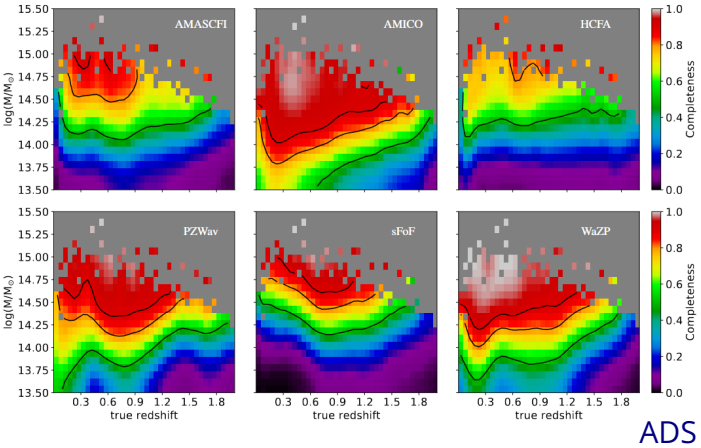
(also includes these slides)



many astro problems come in clusters.

***Euclid* preparation III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection**

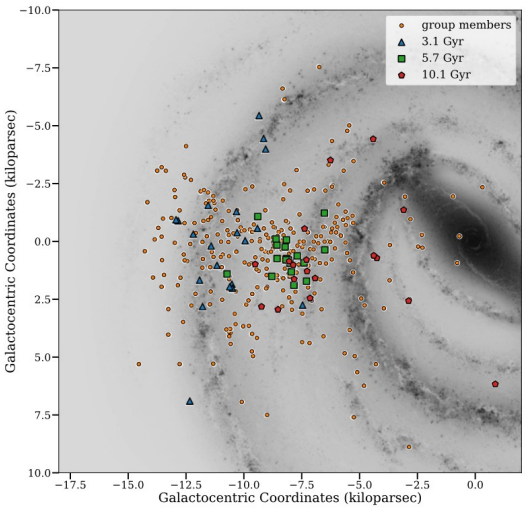
Euclid Collaboration, R. Adam[1,2,3]★, M. Vannier[2], S. Maurogordato[2], A. Biviano[4], C. Adami[5], B. Ascaso[6],



ADS

Strong chemical tagging with APOGEE: 21 candidate star clusters that have dissolved across the Milky Way disc

Natalie Price-Jones[1,2]★, Jo Bovy[1,2], Jeremy J. Webb[1], Carlos Allende Prieto[3,4],



ADS

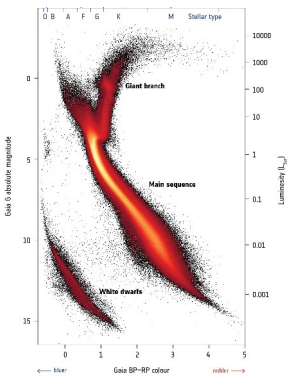
our problem: open clusters of stars.

The Pleiades, age: ~100 Myr

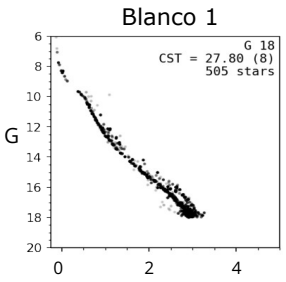
Credit: ESO

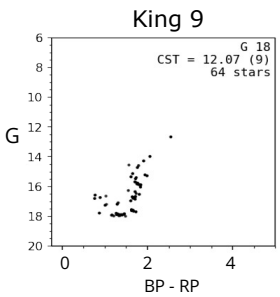
Open clusters are **homogenous** and **bound** groups of **coeval** stars

Highly useful to stellar & galactic science!



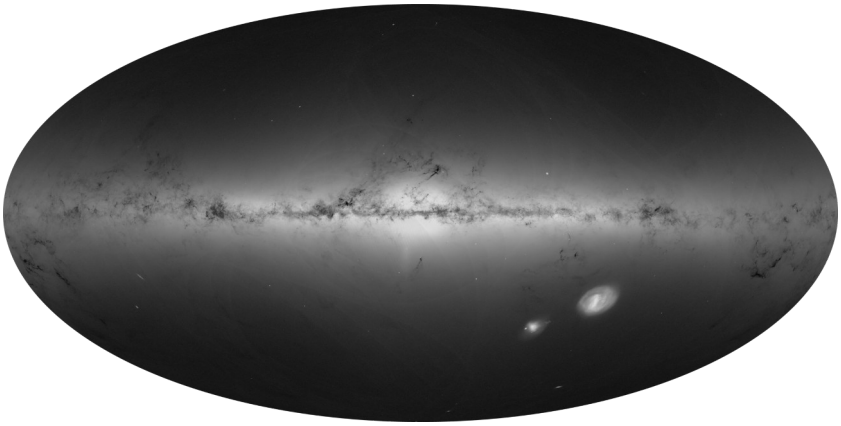
Credit: ESA





Gaia data is a challenge.

Gaia satellite has reliable astrometry for **~1 billion** stars
=> **~0.1%** of which in **open clusters**

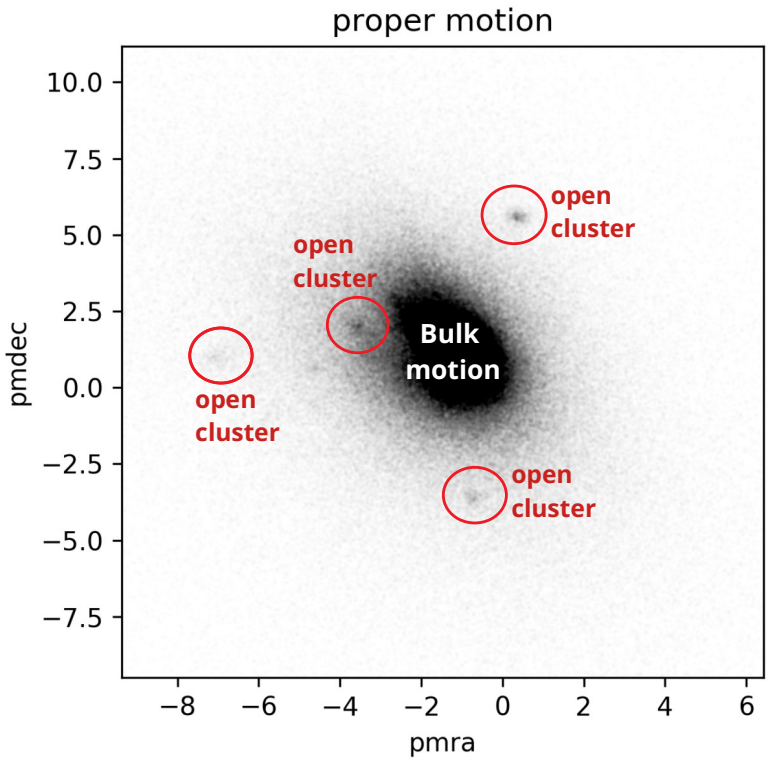


Gaia source density skymap. Credit: ESA

Some challenges:

- **So many sources**...
- **~99.9%** of which must be discarded!
- Open clusters vary from **~0.1°** to **~10°** in size

but... Gaia is still *perfect* for open clusters!

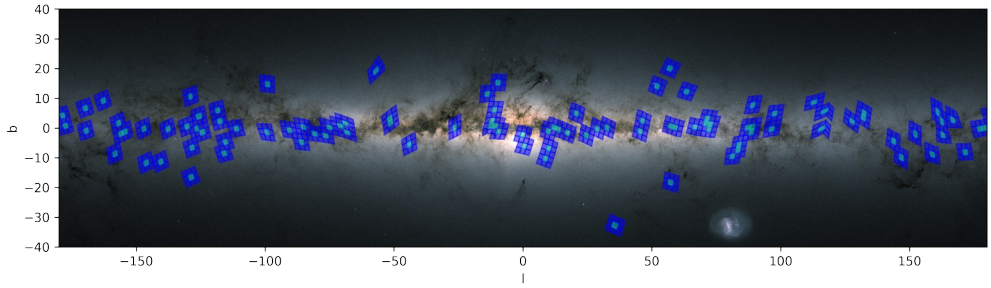


testing different algorithms.

Hunt & Reffert (2021): search for literature clusters in randomly selected fields

After initial trials of what worked best…

- Algorithms given **rescaled** positions, proper motions and parallaxes (5 dimensions) + corrected for spherical distortions
- Paper used **DBSCAN**, **HDBSCAN** and **Gaussian mixture models**



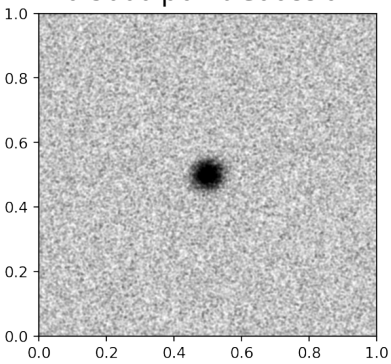
which one is best?

Algorithm	Speed	Sensitivity TP / (TP + FN)	Precision TP / (TP + FP)	
DBSCAN (Castro-Ginard+18 parameters)	**Fast**	0.53	**1.00**	← **Main literature approach**
DBSCAN (My parameters)	**Fast**	0.62	0.93	
HDBSCAN	Quite fast	**0.82**	**0.82**	← **My favourite**
Gaussian Mixtures	Slow	0.33	**1.00**	

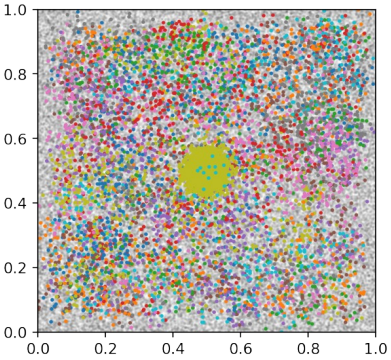
let's talk about false positives.

The **most sensitive** clustering methods can also produce **the most false positives**. Care **must** be taken when using "off-the-shelf" methods

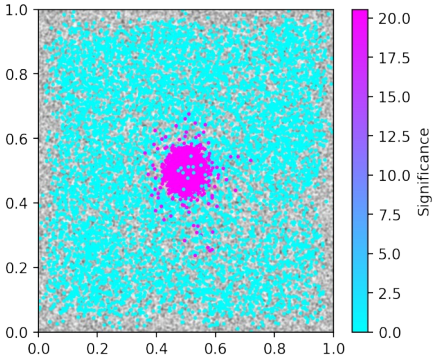
100,000 uniformly distributed 5D points + a 5000 point Gaussian



HDBSCAN, m_{clSize}=20… 151 clusters?!



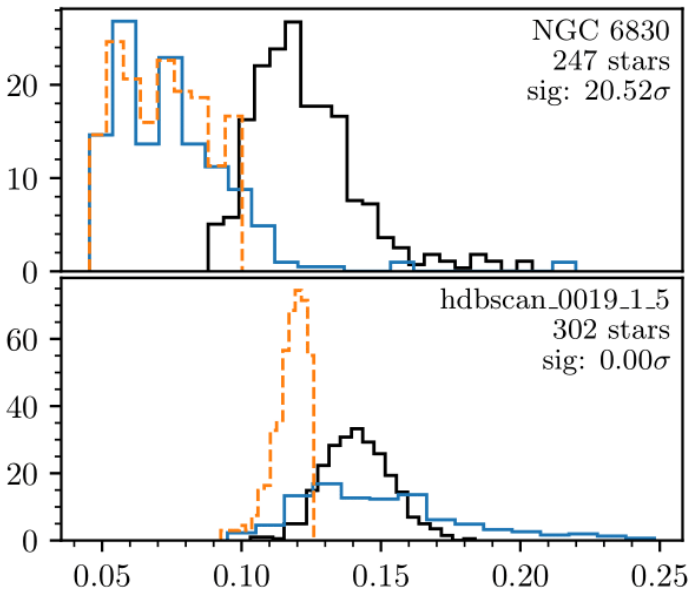
Density test → only the real cluster is significant



how the significance test works.

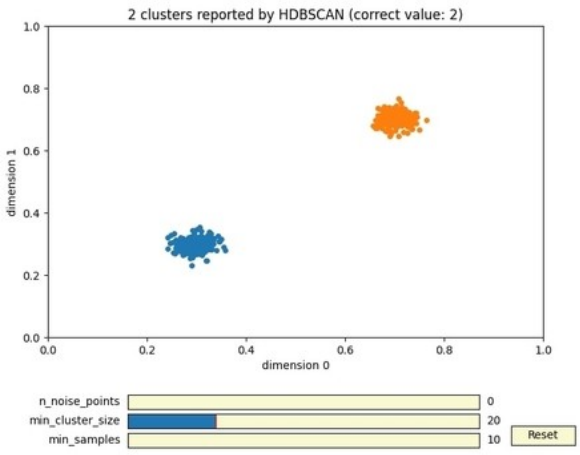
Simple density test compares **nearest neighbour distances** in cluster with field

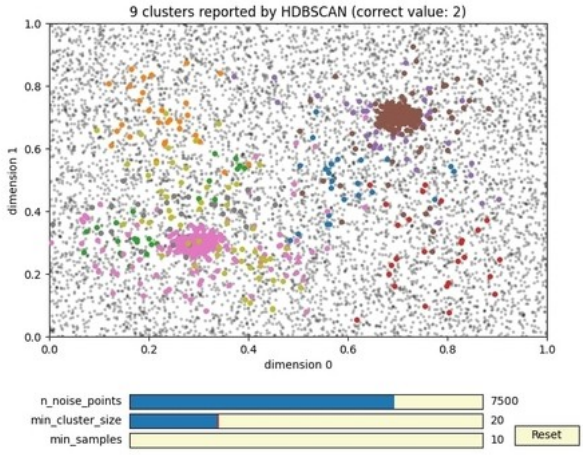
Can tell **random chance associations** and **real clusters** apart



a way to try this in real time!

I made a script that shows this in real time! (GitHub link)



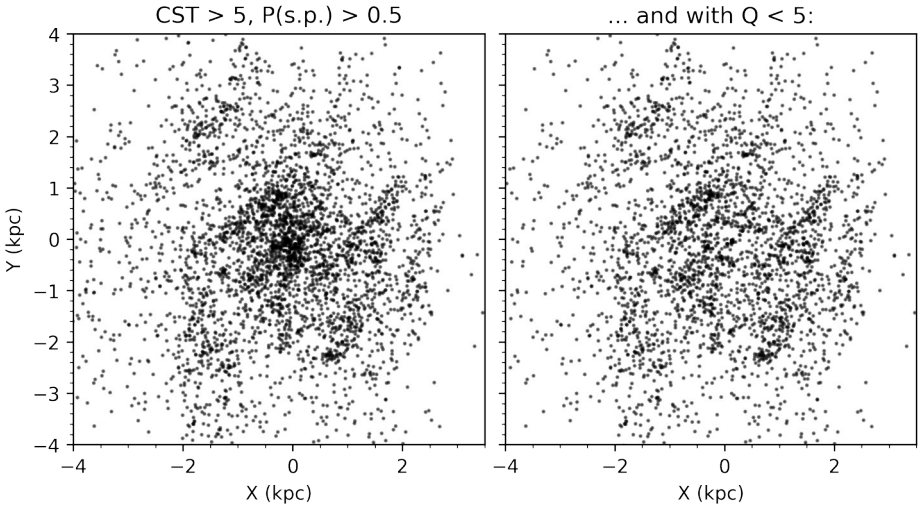


clustering 729 million stars in Gaia EDR3.

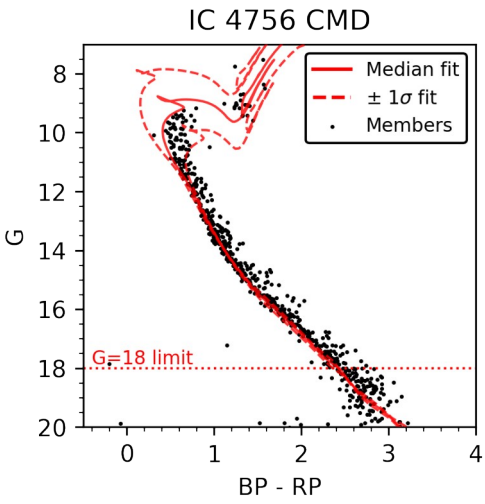
- Ran **HDBSCAN** on largest-ever sample of stars
- Used **HEALPix** tesselation scheme for regions
- About **8 days of wall time** on a powerful machine (actually not bad)

after much work: results.

Distribution, |z| < 500 pc:



Example CMD:



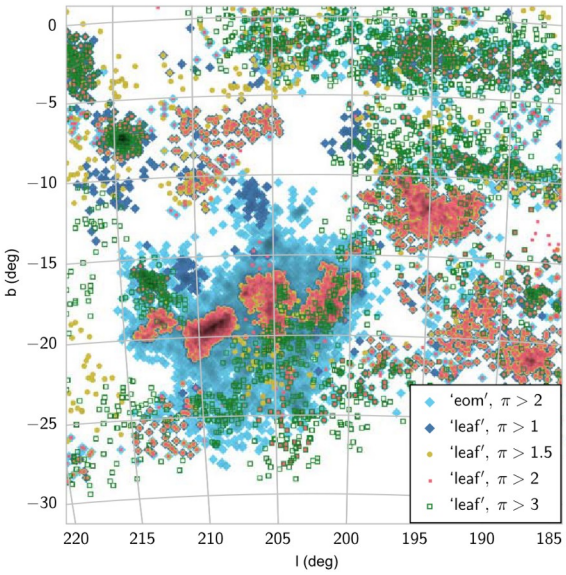
and a cautionary tale on false positives...

Kounkel+19,20: reported **thousands** of groups and 'strings' of stars using **HDBSCAN** on Gaia DR2 data

(+ without using a significance test to clean HDBSCAN results)

In EDR3: we re-detect just **18.1%** of their groups (with better data!)

Also echoes **Zucker+22:** "many Kounkel+ group members inconsistent with having common origin"



where next for clustering algorithms?

- Clustering algorithms are **fantastic** tools for astronomy
- Many off-the-shelf algorithms available (e.g. scikit-learn)
- BUT: they aren't designed for astronomy problems, and can run into weird issues

In the future: astro must collaborate with computer scientists, mathematicians etc. to develop algorithms

datasets to come...

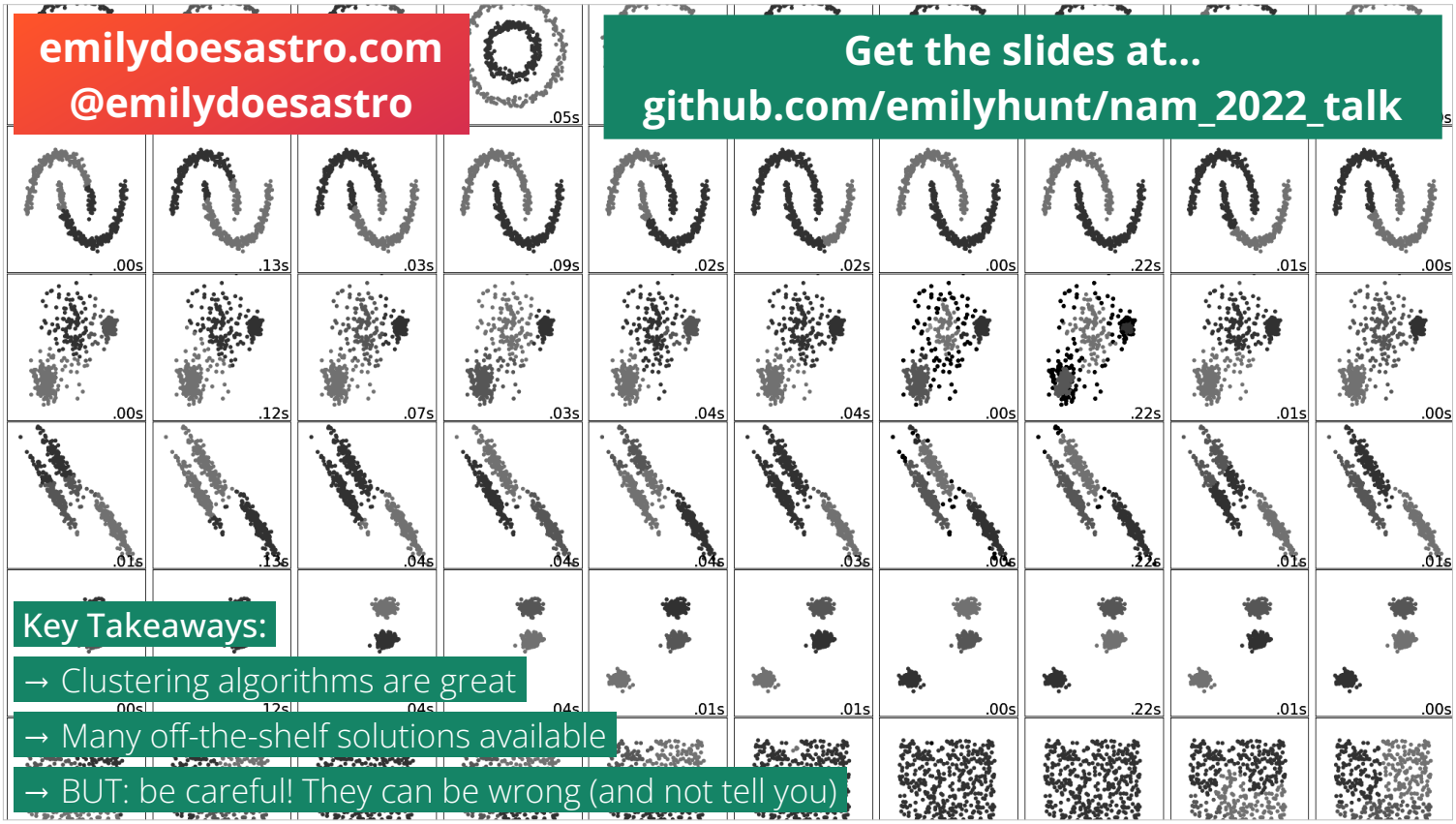
Gaia DR5 (~2030): 2 billion stars

Euclid (late 2020s): ~1 billion galaxies with redshifts

LSST (final data ~2035): 17 billion stars, 20 billion galaxies

GaiaNIR (~2050): 12 billion stars

+ many, many other uses across astronomy for clustering algorithms


emilydoesastro.com
@emilydoesastro
Get the slides at...
github.com/emilyhunt/nam_2022_talk
Key Takeaways:
→ Clustering algorithms are great
→ Many off-the-shelf solutions available
→ BUT: be careful! They can be wrong (and not tell you)