Searching for what no one is looking for

Sebastian Ratzenböck

Data Science @ Uni Vienna

AIA, Garching
July 25th, 2019

Searching for what no one is looking for

Blind searches in Gaia DR2

Sebastian Ratzenböck

Data Science @ Uni Vienna

universität wien

AIA, Garching

July 25th, 2019

Start PhD


Blind search
Save work
Start PhD


Blind search
Save work
Start PhD


Blind search
Save work
Start PhD
Abstract

Part I

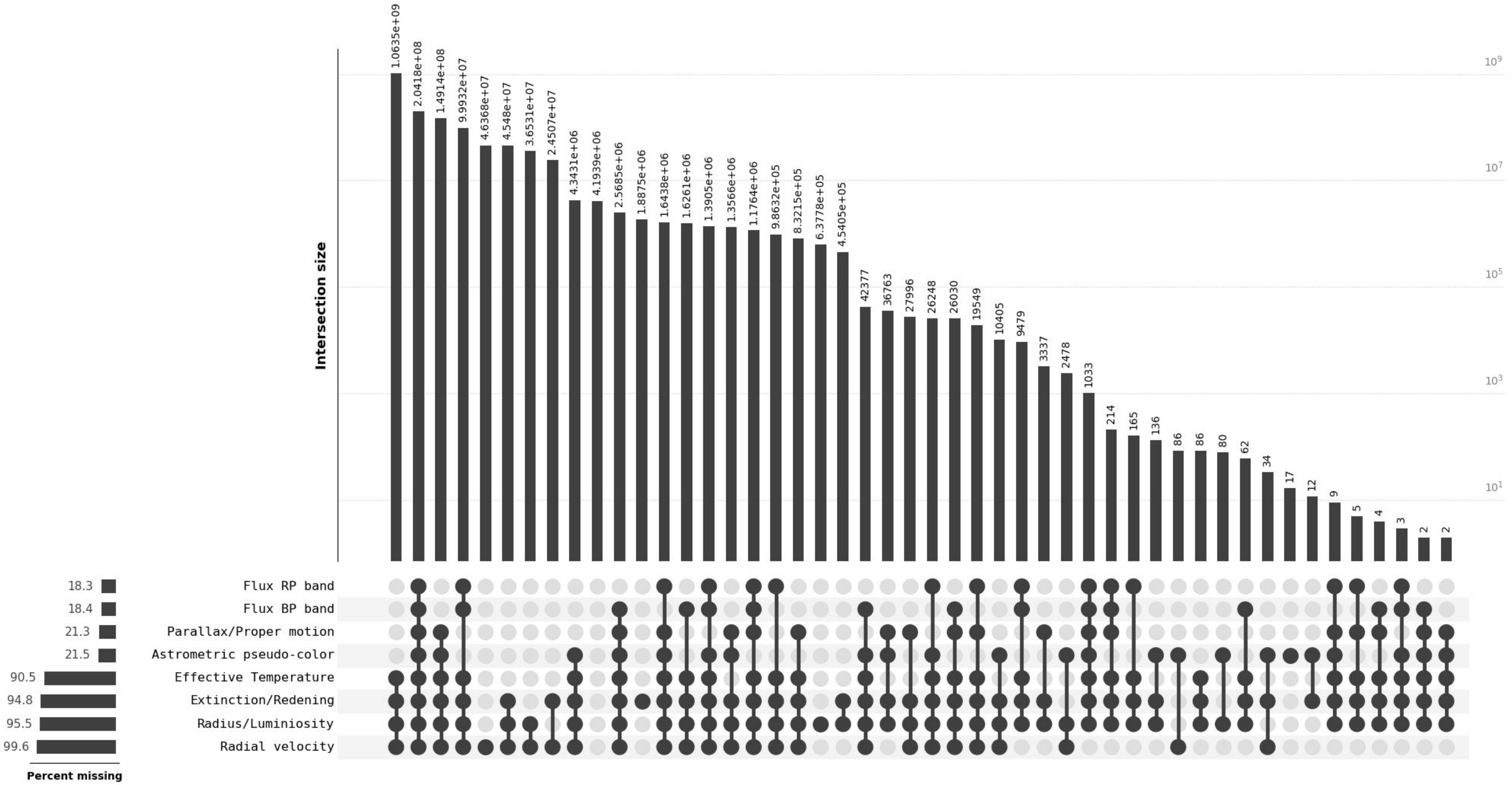
The blind search

Blind search in Gaia DR2

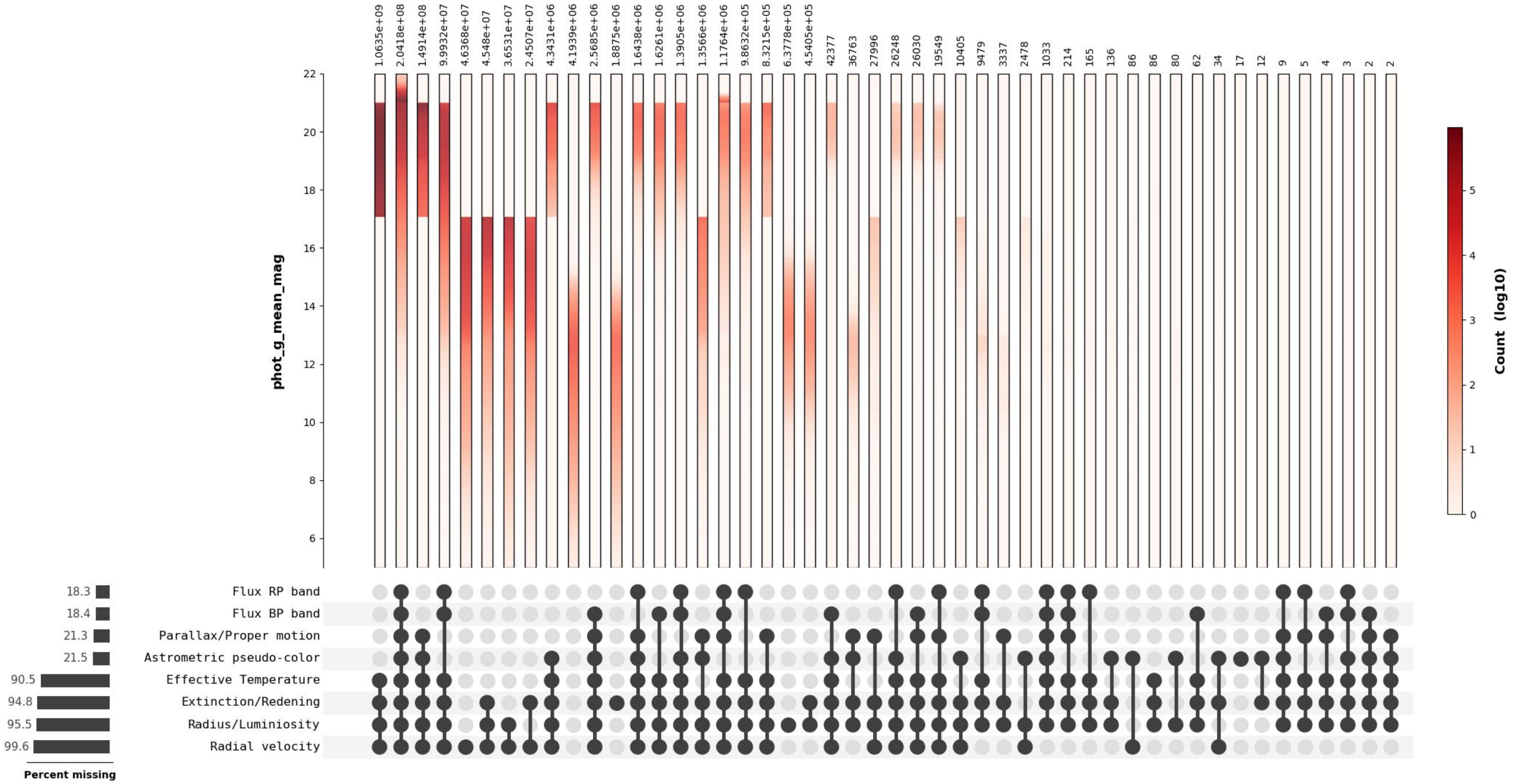
"Gaia does not exclusively observe stars: ***all*** objects brighter than G ≈ 20 mag are observed, [...]."

- Gaia: Science Performance

Data structure: Missing value sets in Gaia



Subset distributions: G-magnitude



Clustering pipeline

1. Give each data point a label

$$f(\vec{x}_i, \vec{\theta}) = s_i \qquad \vec{x}_i \in \mathbb{R}^n,\;\; s_i \in \mathbb{Z},\;\; \vec{\theta} = (\theta_1, \ldots, \theta_m)$$

2. Validate clustering

Clustering ***cycle***

1. Give each data point a label

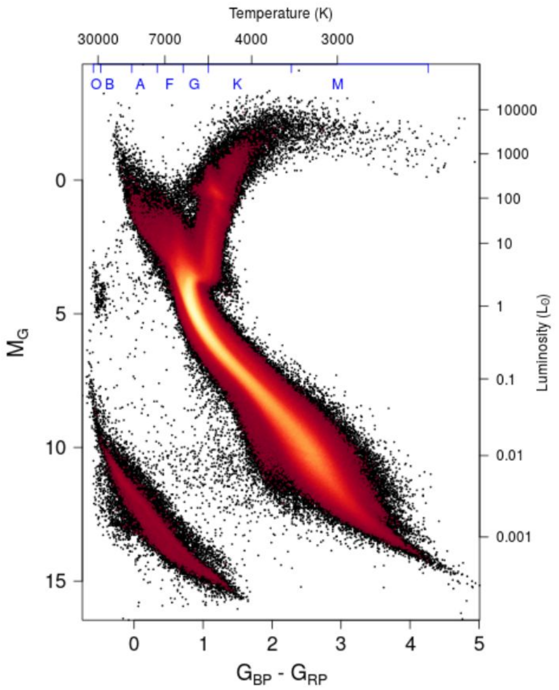
$$f(\vec{x}_i, \vec{\theta}) = s_i \qquad \vec{x}_i \in \mathbb{R}^n,\ \ s_i \in \mathbb{Z},\ \ \vec{\theta} = (\theta_1, \ldots, \theta_m)$$

2. Validate clustering (then go back to 1.)

Choosing the algorithm

Requirements

- Deal with non-linearities between variables
  - Flexible model
- Applicable to millions of data points



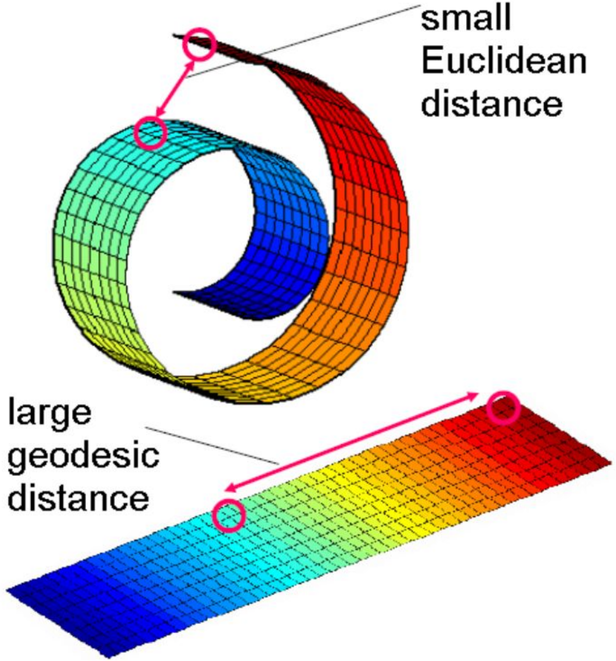
Source: Gaia Collaboration (2018)

Reducing dimensions - Manifold learning

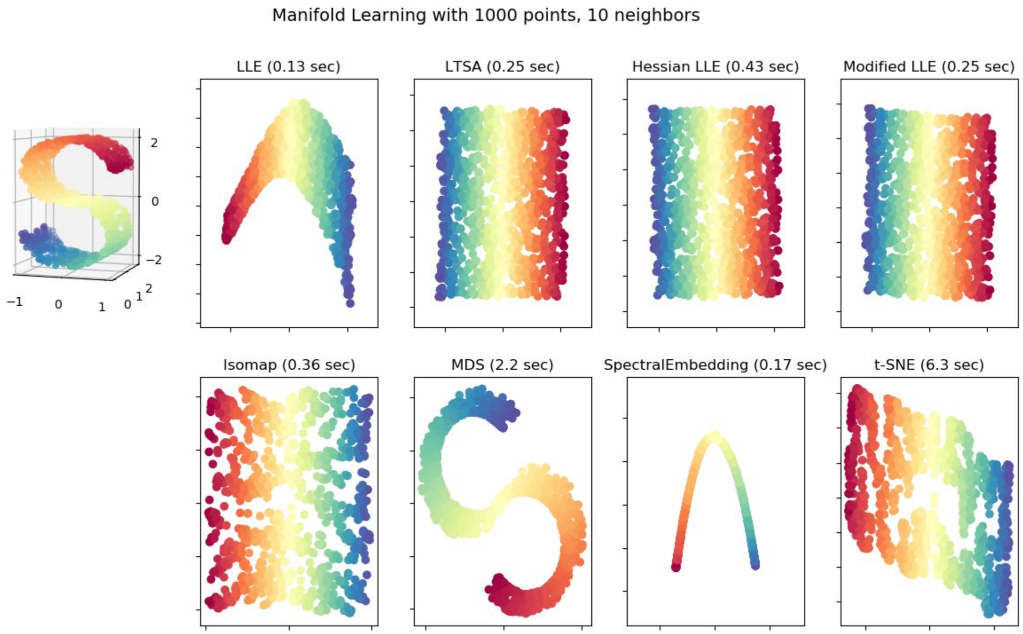
If variables **depend** on each other their joint distribution does not span the whole space

→ data lies on (around) the support of the joint distribution

Manifold: underlying support of the data distribution known only through finite sampling



Reducing dimensions - Manifold learning

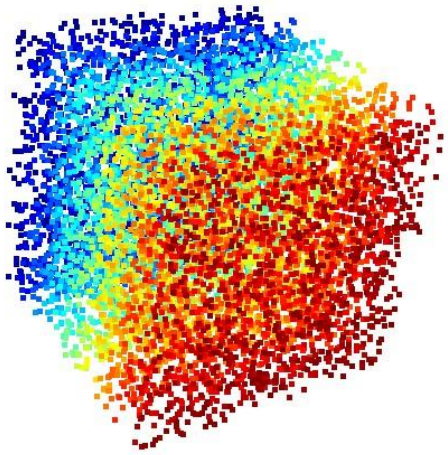


Source: Scikit Learn (scikit-learn.org/stable/modules/manifold.html)

Reducing dimensions - Manifold learning

BUT: We generally do **not** know the dimensionality of the intrinsic manifold

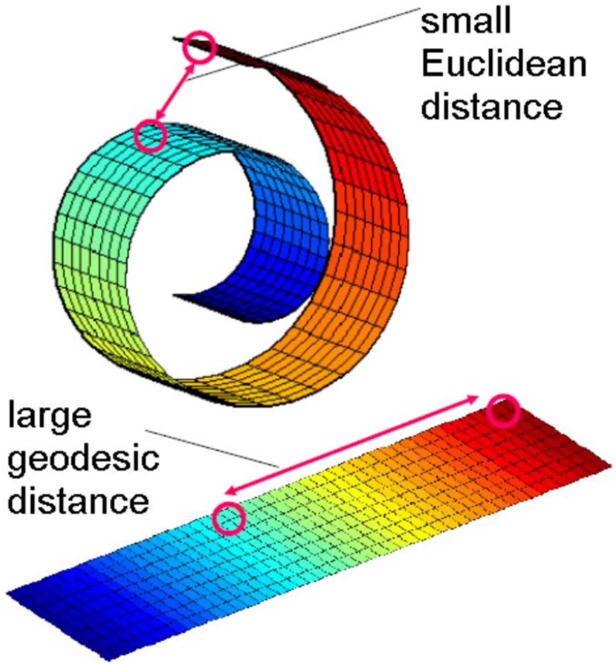
$\rightarrow$ Can lose valuable information



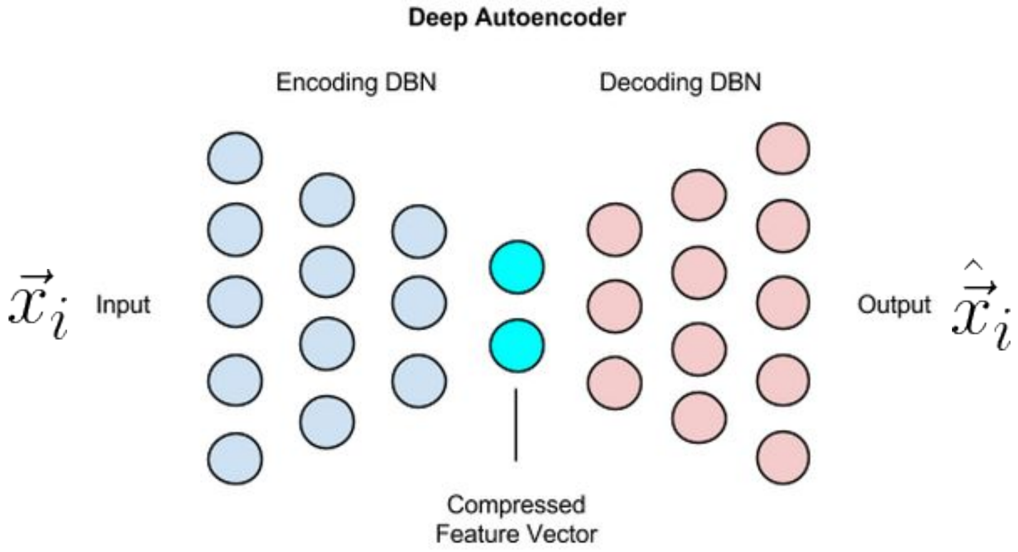
Reducing dimensions - Manifold learning

BUT: We generally do **not** know the dimensionality of the intrinsic manifold

$\rightarrow$ Can lose valuable information

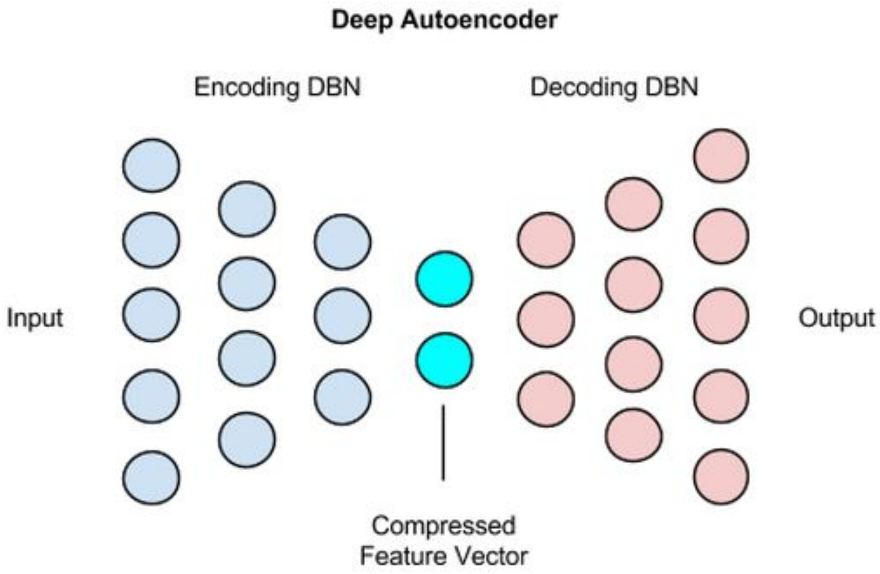


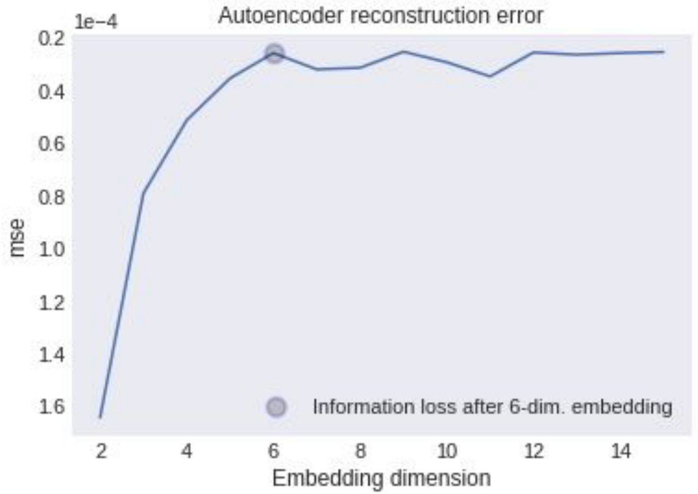
Reducing dimensions - Autoencoder



$$Loss = \sum_i ||\vec{x}_i - \hat{\vec{x}}_i||$$

Reducing dimensions - Autoencoder

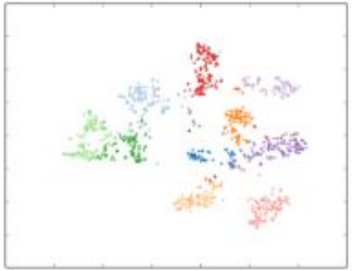




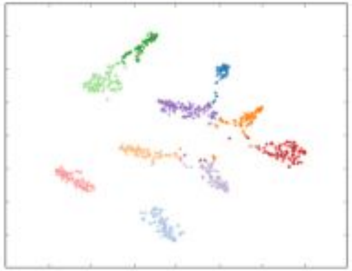
Deep embedded clustering (Xie et al. 2016)

- Use neural net as powerful feature extractor
- Introduce a second training phase where the representation in the mapping to the latent space is optimized for k-Means clustering
    - Set centroids (hyperparameter) in latent space and force points around these centroids to be t-distributed by minimizing KL divergence loss term

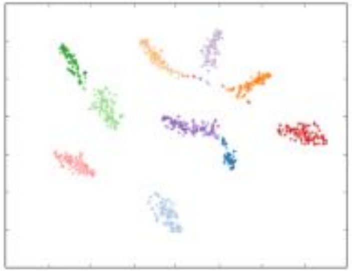
Deep embedded clustering (Xie et al. 2016)



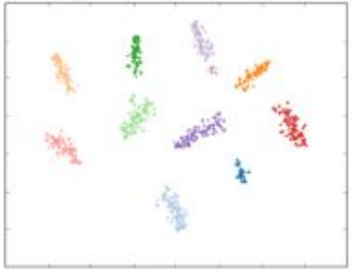
(a) Epoch 0



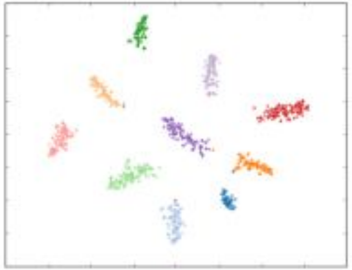
(b) Epoch 3



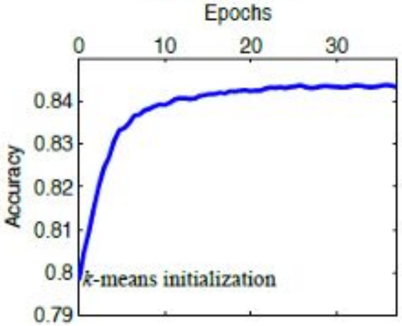
(c) Epoch 6



(d) Epoch 9

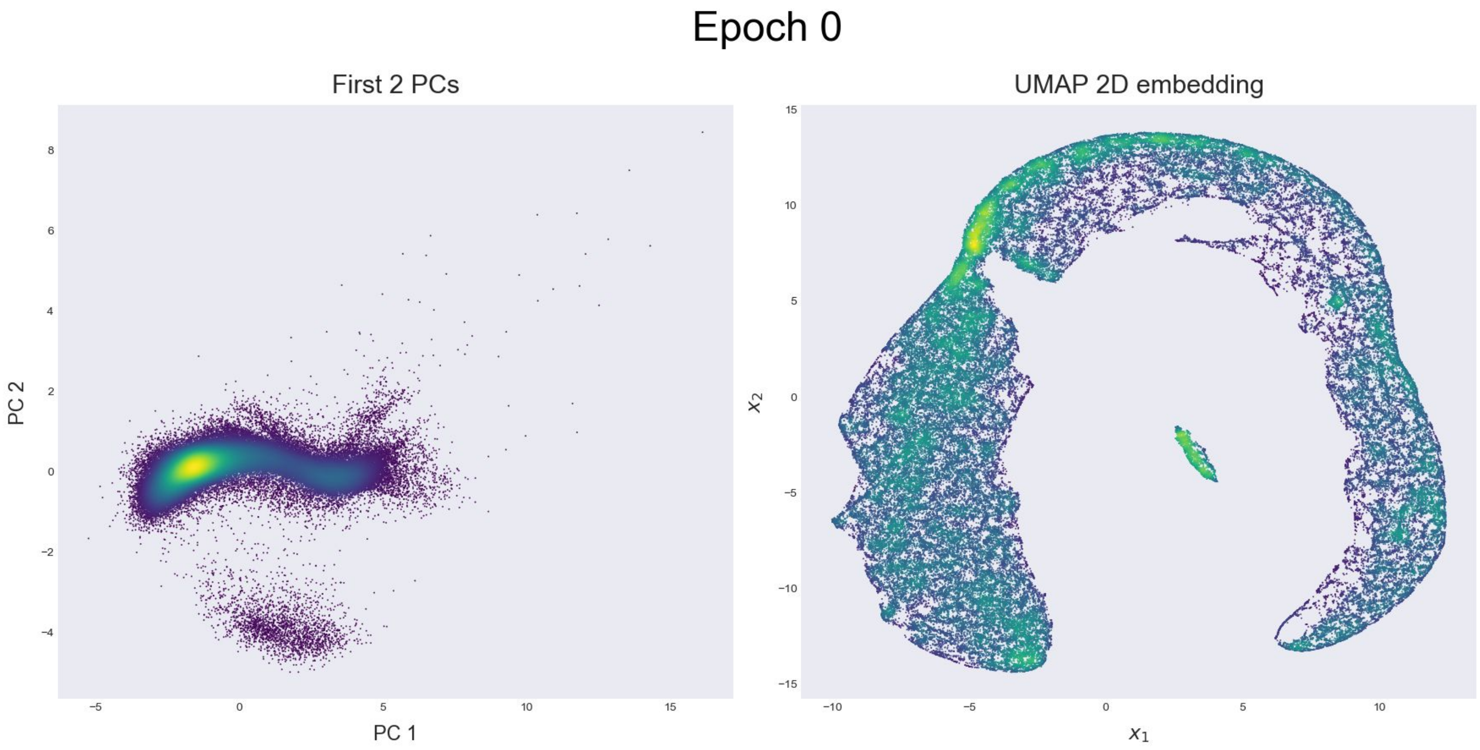


(e) Epoch 12

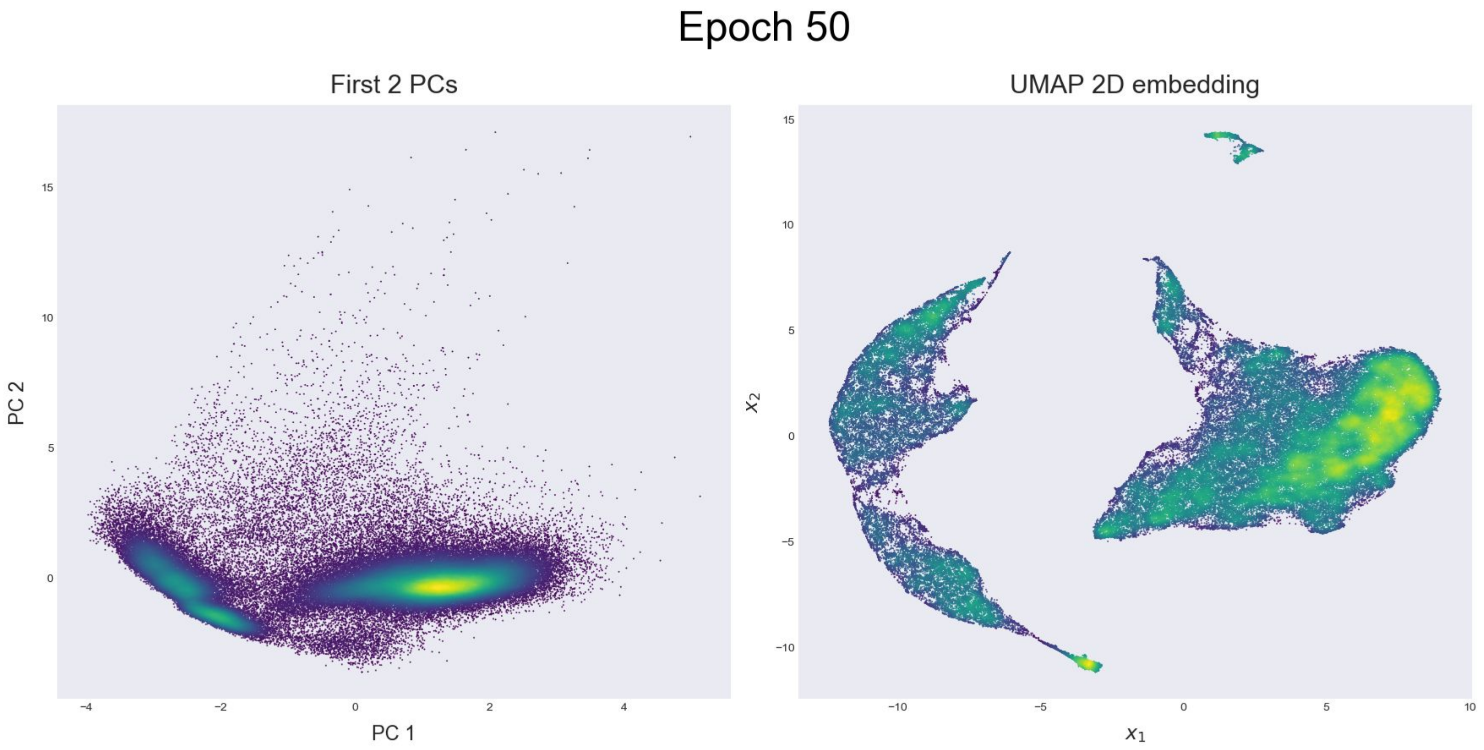


(f) Accuracy vs. epochs

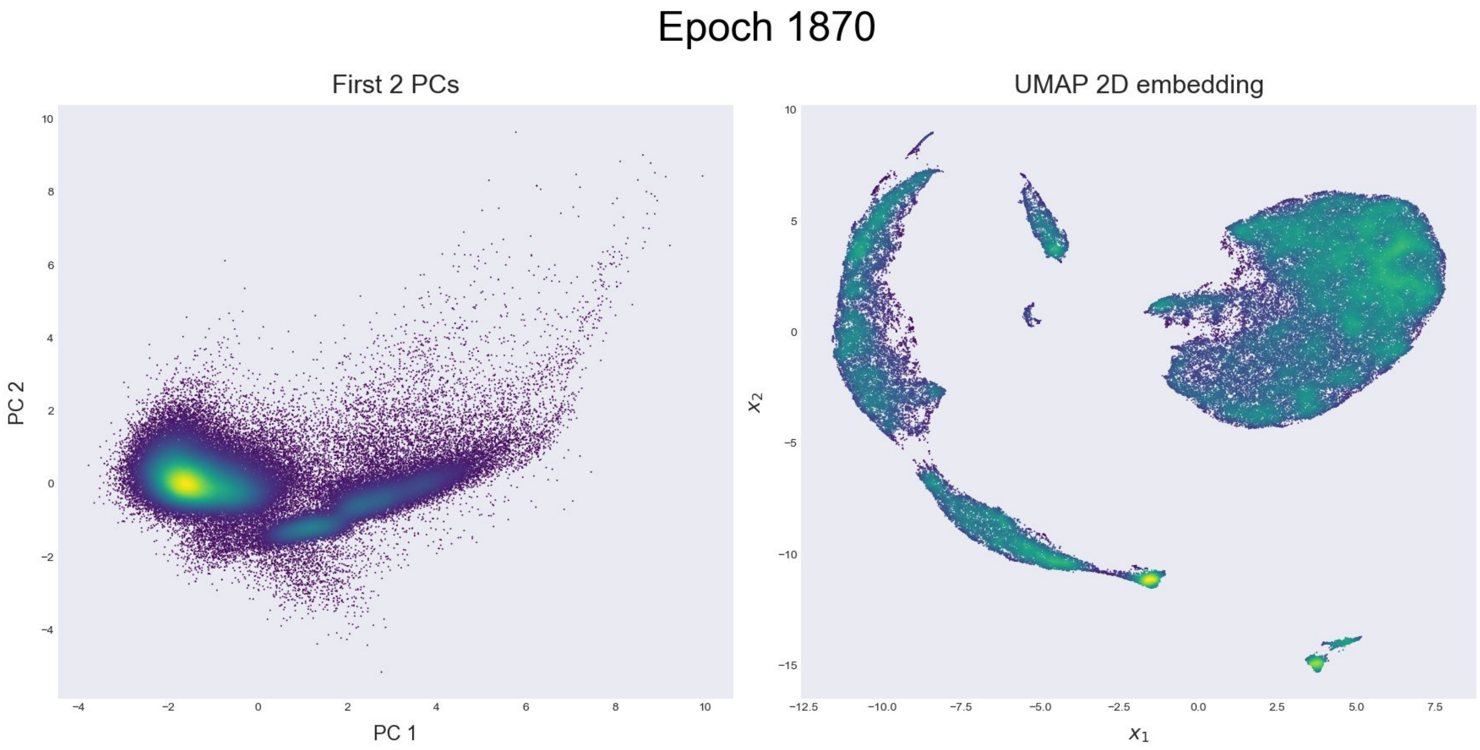
Deep embedded clustering on Gaia data



Deep embedded clustering on Gaia data

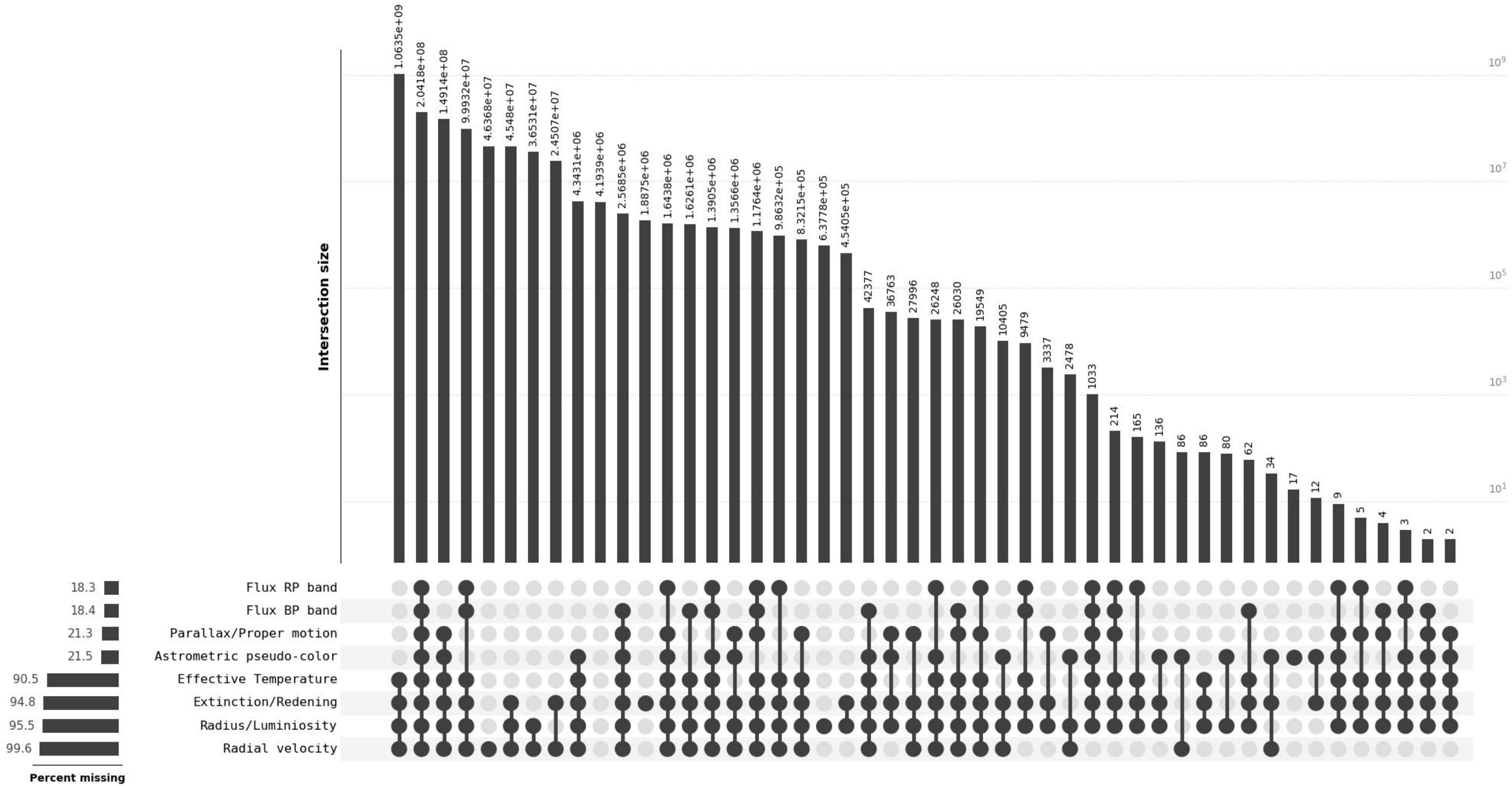


Deep embedded clustering on Gaia data



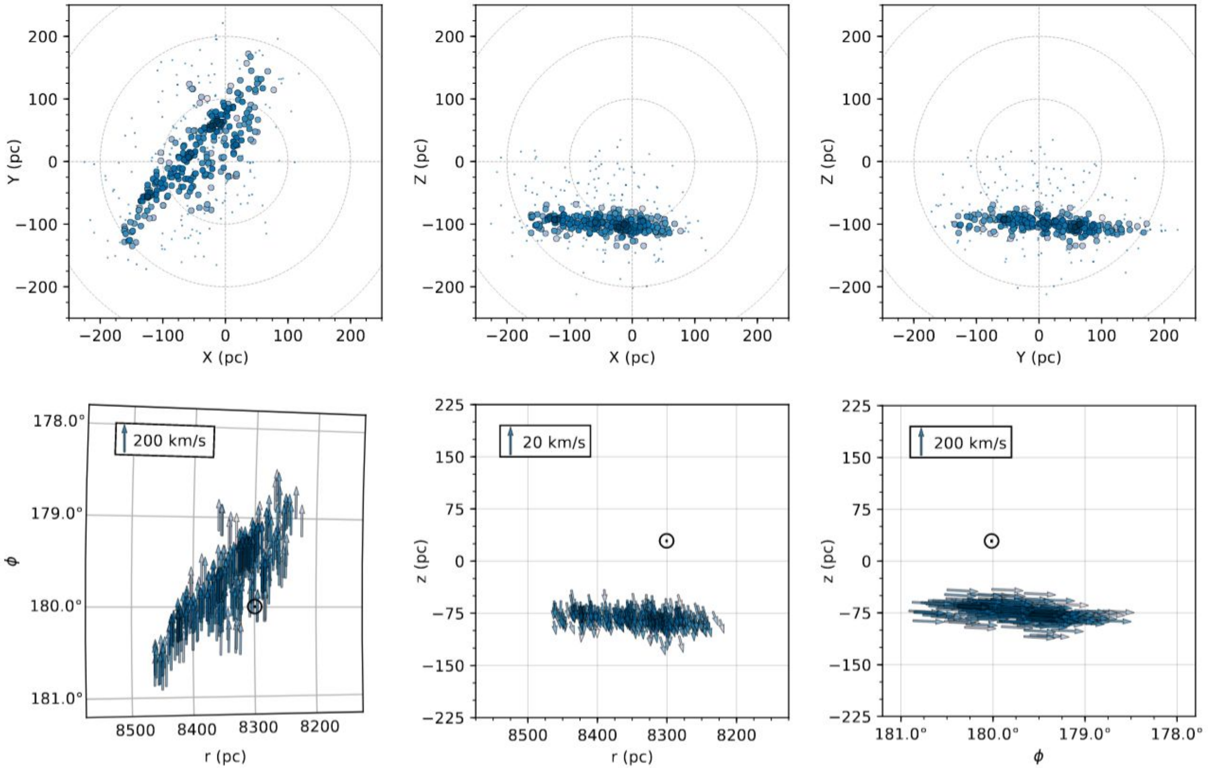
Part II

Feature engineering for robust OC extraction



Intersection size
1.0635e+09
2.0418e+08
1.4914e+08
9.9932e+07
4.6368e+07
4.548e+07
3.6531e+07
2.4507e+07
4.3431e+06
4.1939e+06
2.5685e+06
1.8875e+06
1.6438e+06
1.6261e+06
1.3905e+06
1.3566e+06
1.1764e+06
9.8632e+05
8.3215e+05
6.3778e+05
4.5405e+05
42377
36763
27996
26248
26030
19549
10405
9479
3337
2478
1033
214
165
136
86
86
80
62
34
17
12
9
5
4
3
2
2
10^9
10^7
10^5
10^3
10^1
Flux RP band
Flux BP band
Parallax/Proper motion
Astrometric pseudo-color
Effective Temperature
Extinction/Redening
Radius/Luminiosity
Radial velocity
18.3
18.4
21.3
21.5
90.5
94.8
95.5
99.6
Percent missing

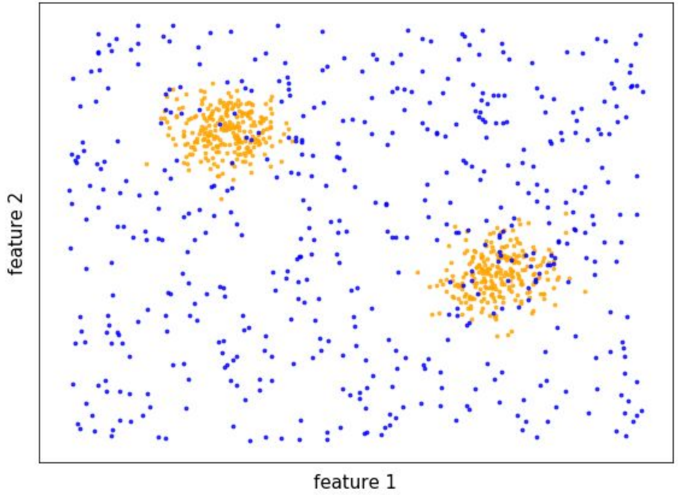
Stellar clusters



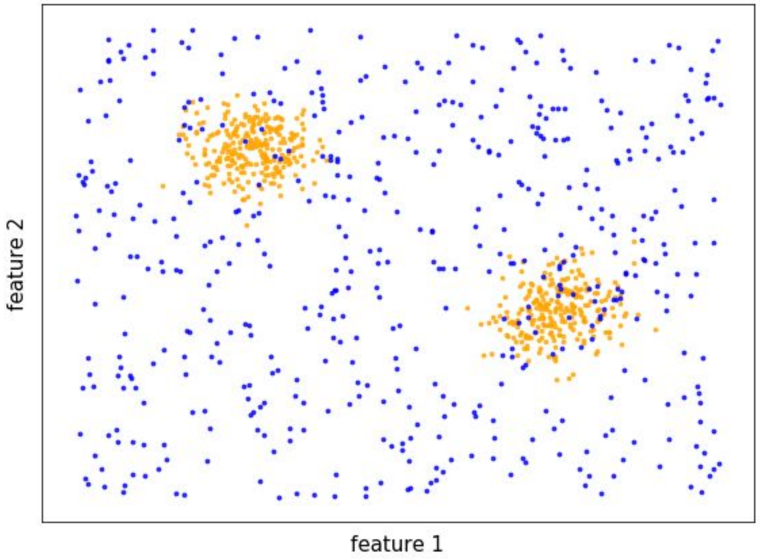
Source: Meingast et al. (2019)

Stellar clusters - feature space

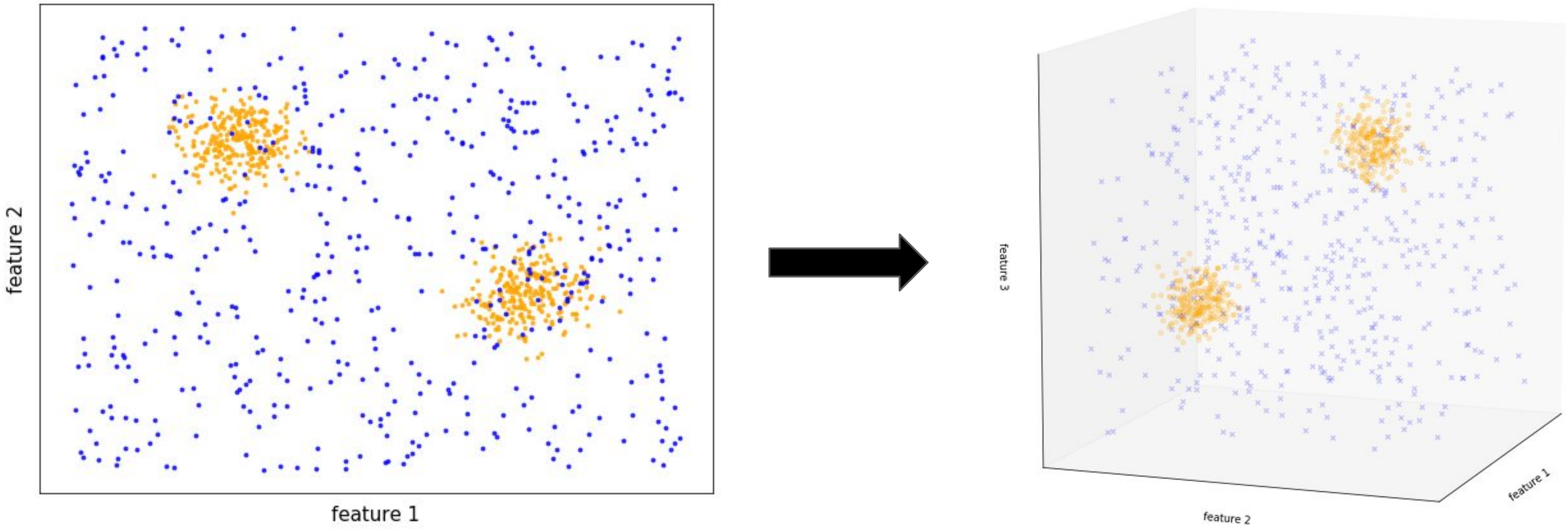
- 5D feature space: XYZ + PMs
- The feature space is dominated by noise
- OC have different densities



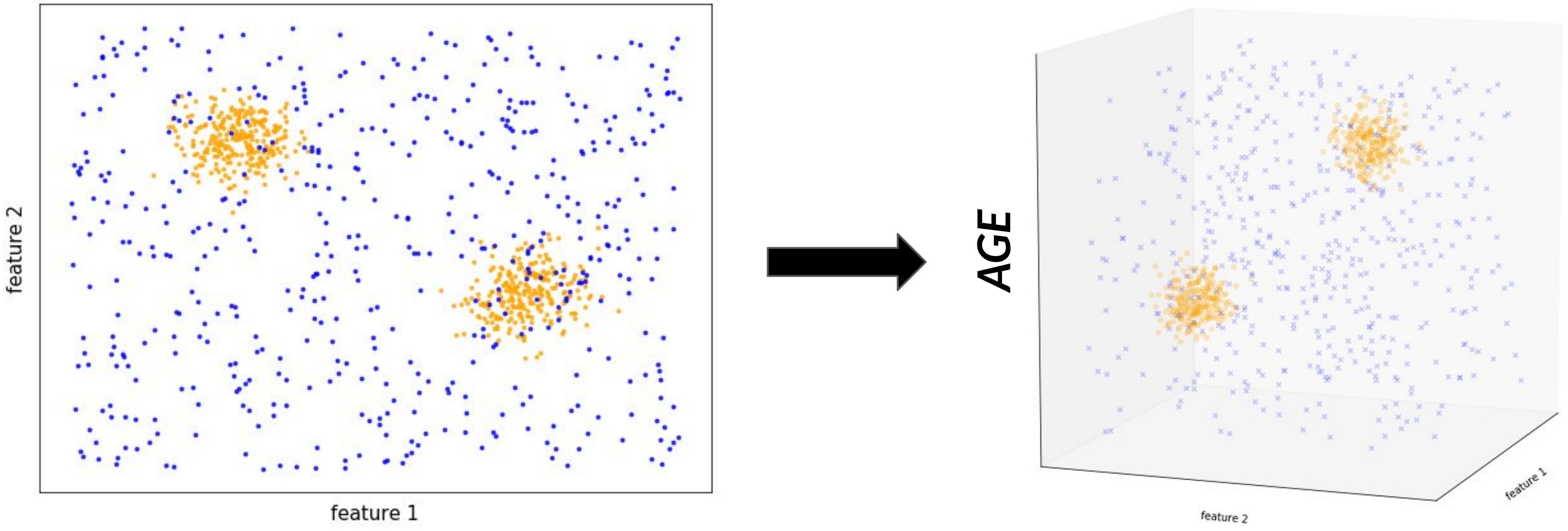
Feature engineering



Feature *adding*

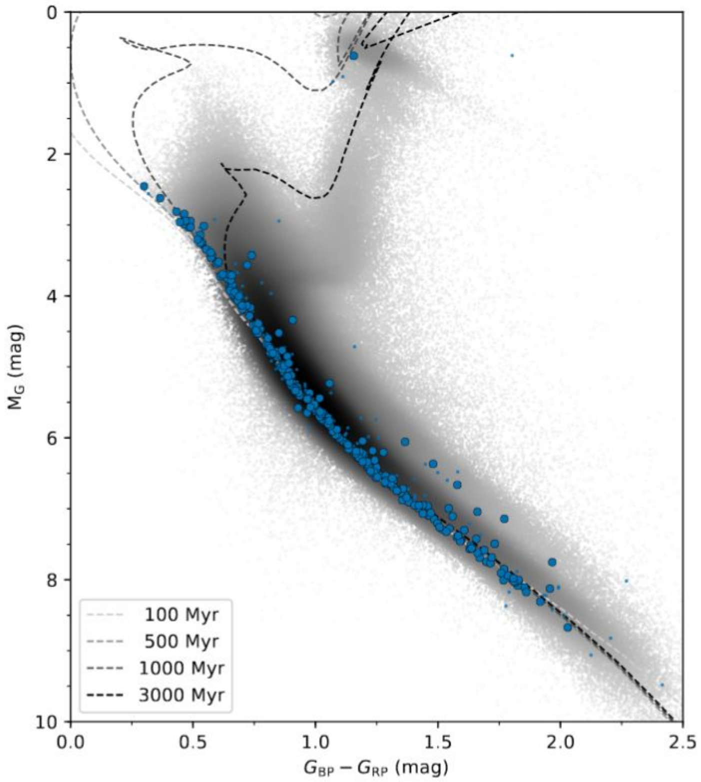


Feature *adding*



Age engineering

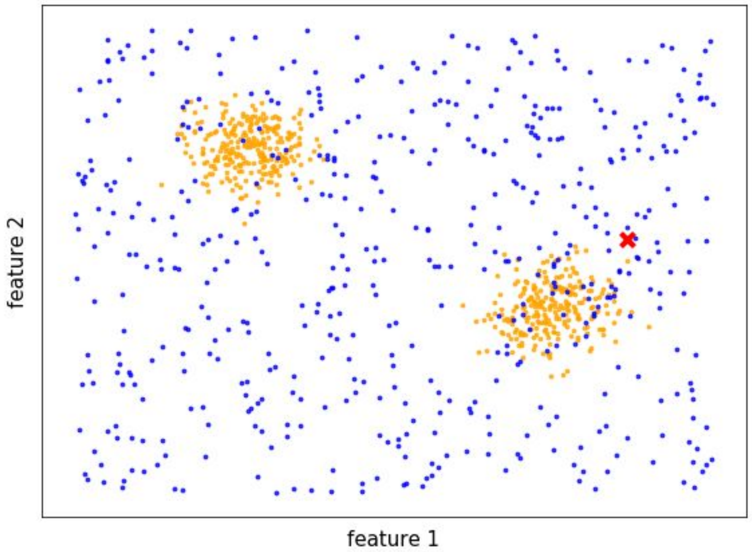
- Fitting a curve to the data which you believe are members of your cluster
- Usually quite messy, isochrone models are not perfect



Source: Meingast et al. (2019)

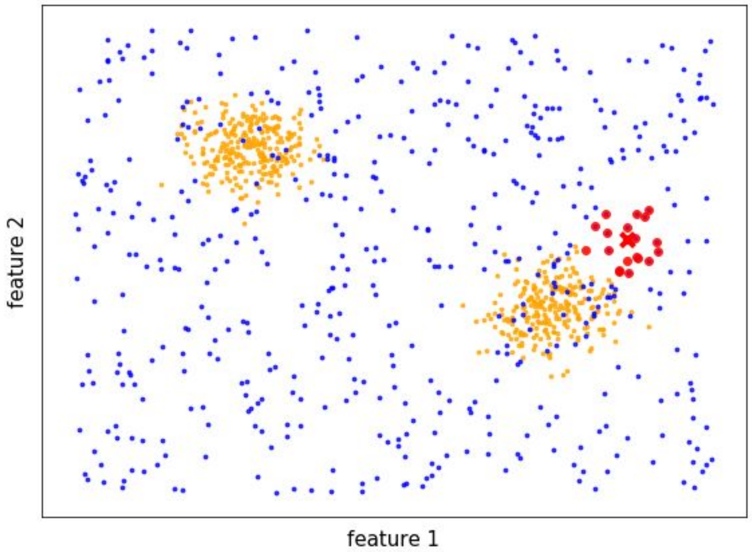
Age engineering

1. Take a point from the sample



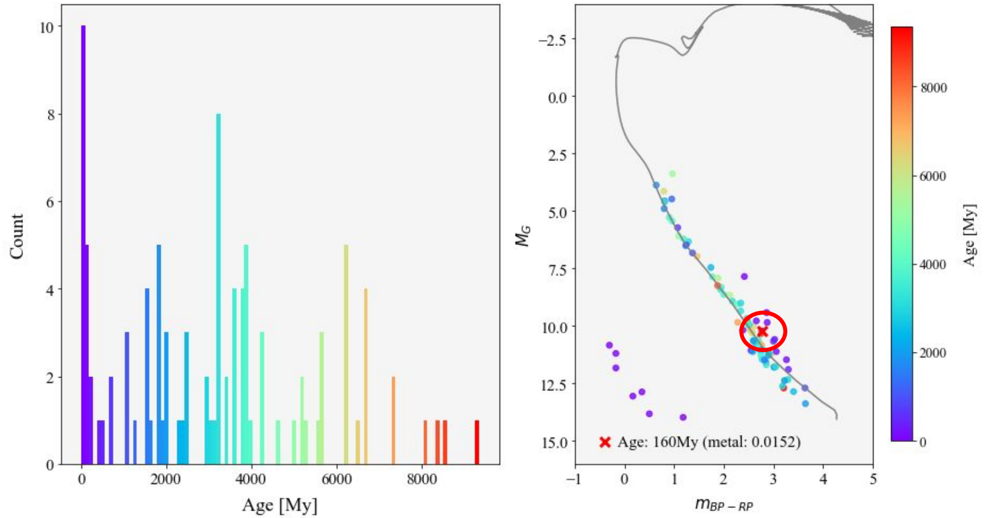
Age engineering

1. Take a point from the sample
2. Get its neighborhood



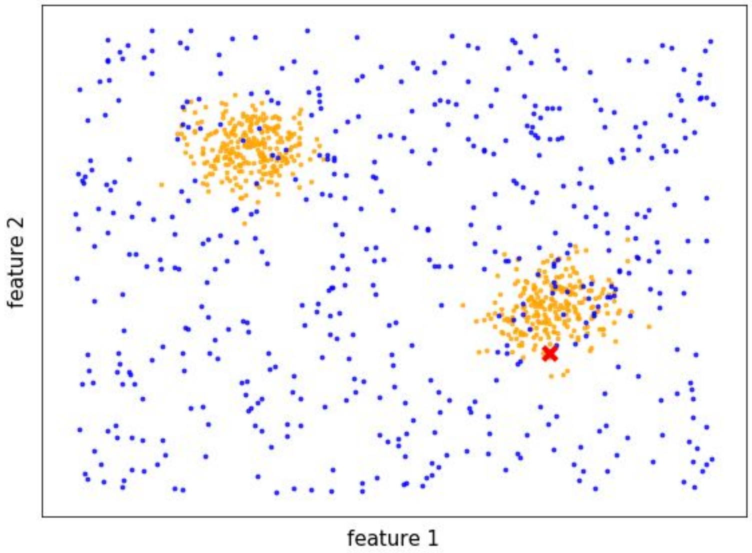
Age engineering

1. Take a point from the sample
2. Get its neighborhood
3. Plot neighborhood points in CMD & fit age to points



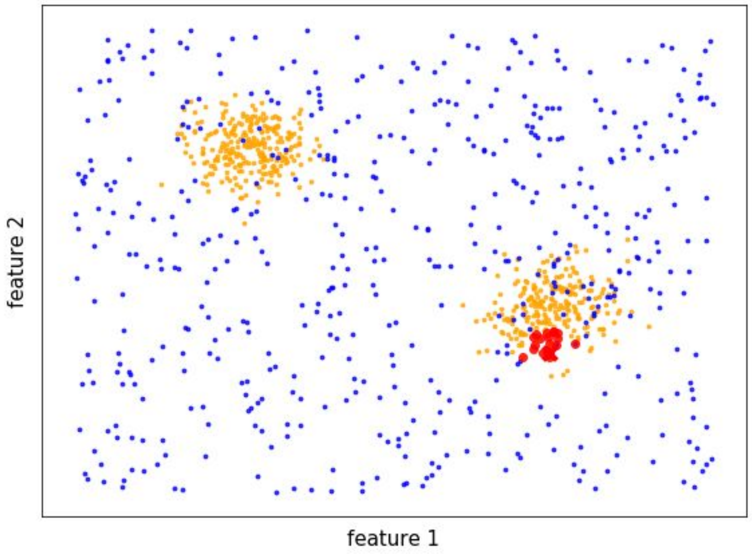
Age engineering

1. Take a point from the sample



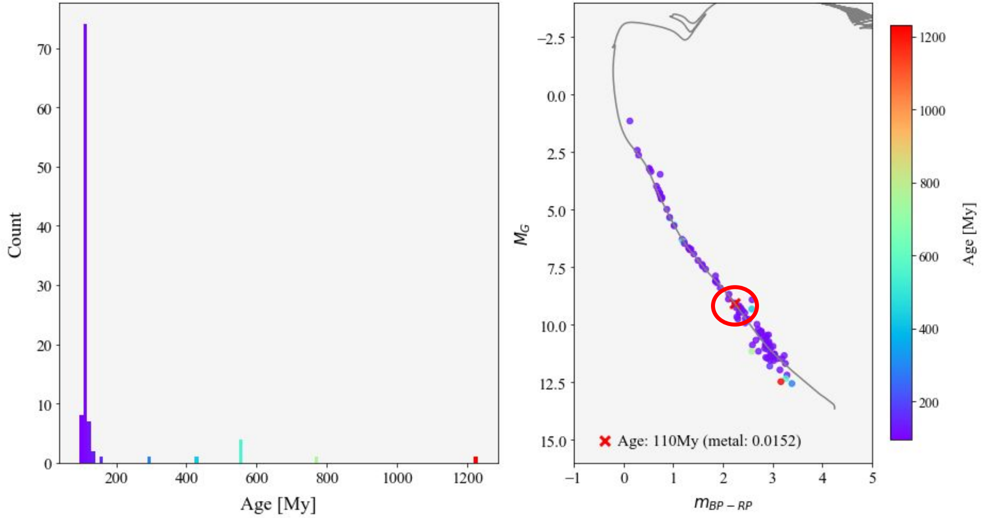
Age engineering

1. Take a point from the sample
2. Get its neighborhood



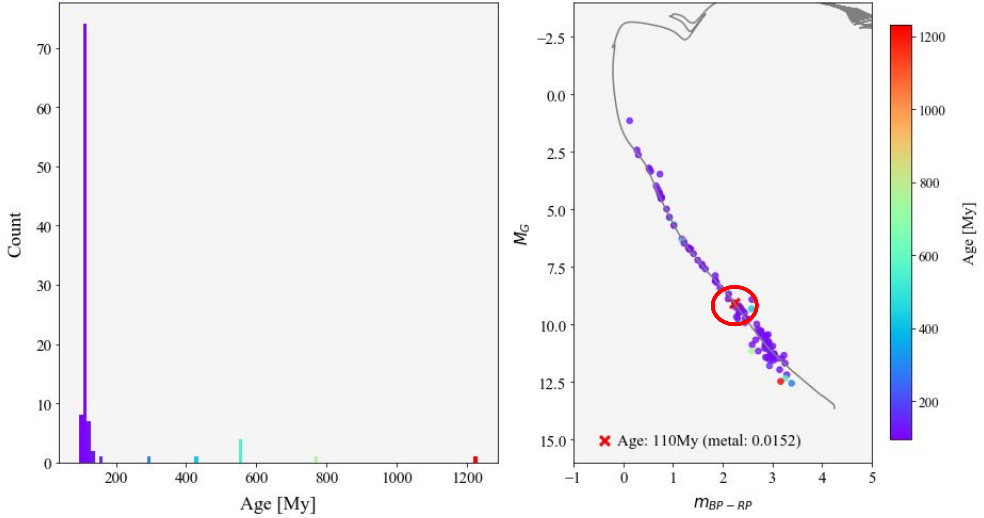
Age engineering

1. Take a point from the sample
2. Get its neighborhood
3. Plot neighborhood points in CMD & fit age to points



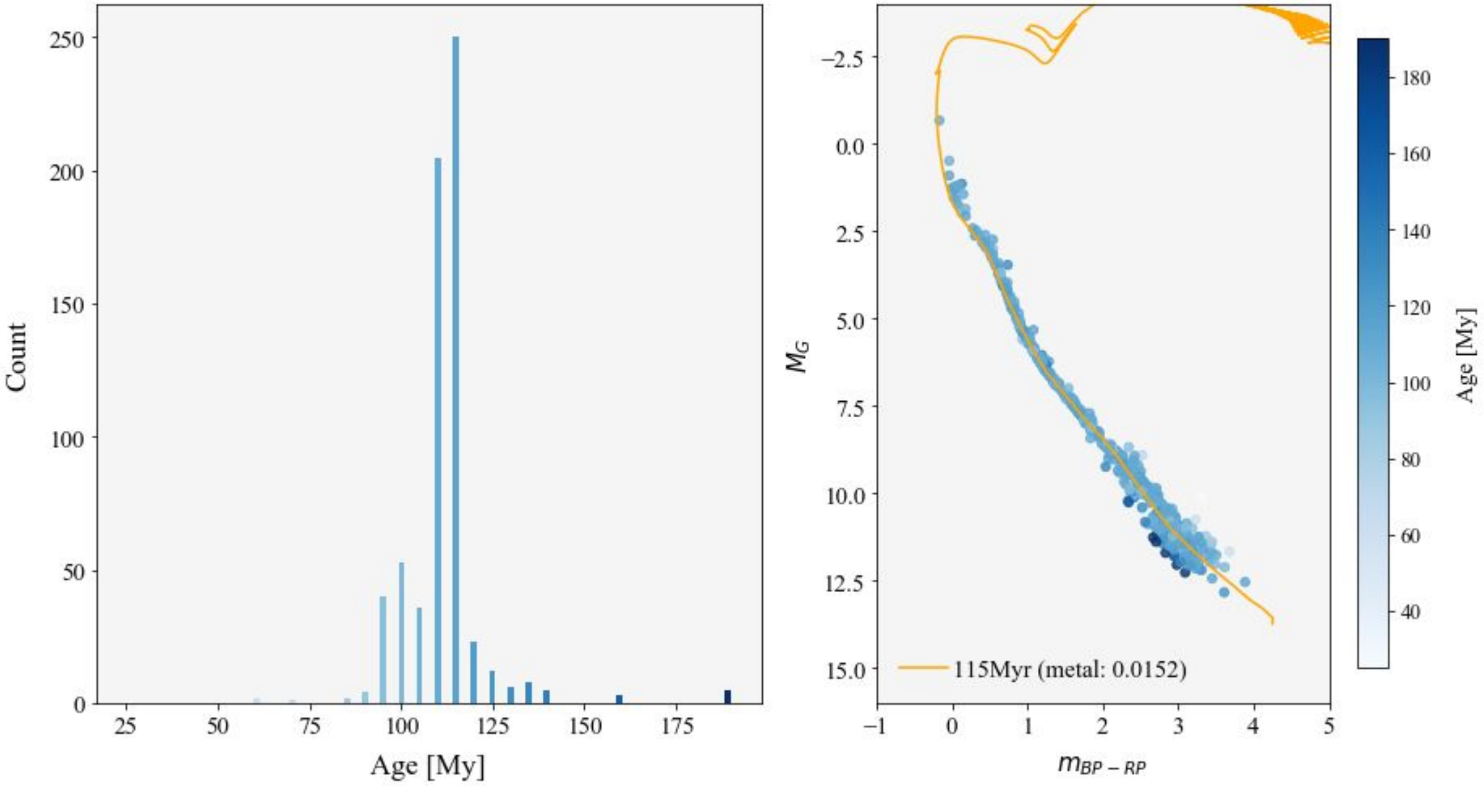
Age engineering

1. Take a point from the sample
2. Get its neighborhood
3. Plot neighborhood points in CMD & fit age to points



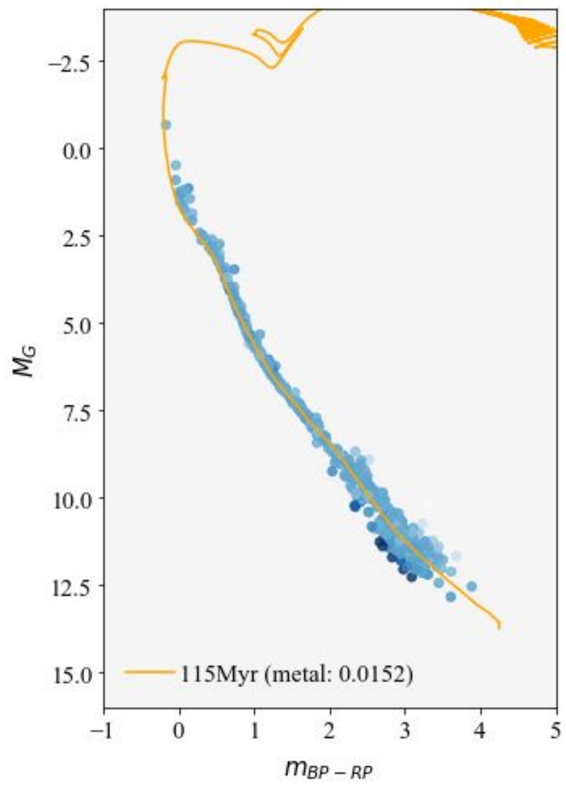
$$d_{ij} = c_x \times \sqrt{(\vec{x}_i - \vec{x}_j)^2} + c_v \times \sqrt{(\vec{v}_i - \vec{v}_j)^2} + c_{age} \times |age_i - age_j| + c_m \times |m_i - m_j|$$

Results

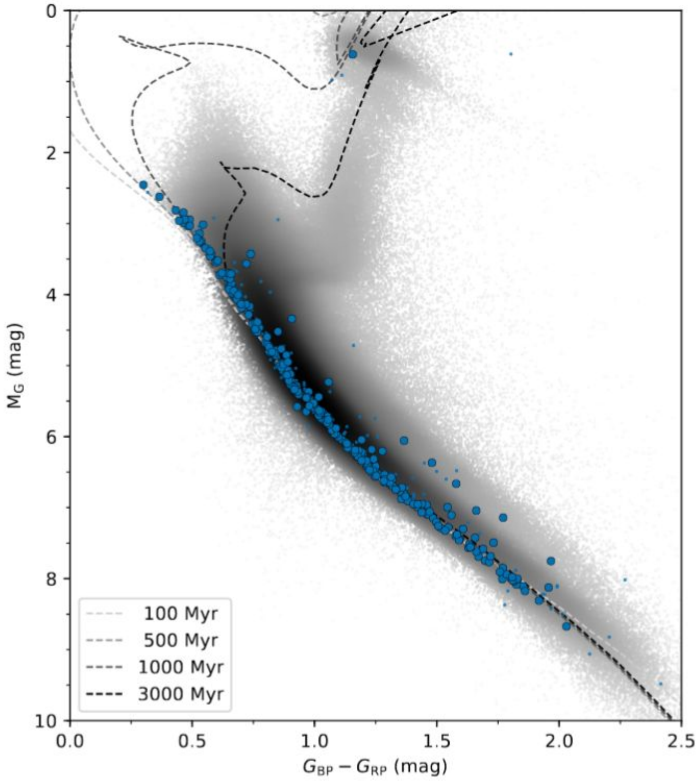


Count
250
200
150
100
50
0
25
50
75
100
125
150
175
Age [My]
M_G
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
115Myr (metal: 0.0152)
−1
0
1
2
3
4
5
m_{BP-RP}
180
160
140
120
100
80
60
40
Age [My]

Results



−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
M_G
115Myr (metal: 0.0152)
−1
0
1
2
3
4
5
m_{BP-RP}



0
2
4
6
8
10
M_G (mag)
100 Myr
500 Myr
1000 Myr
3000 Myr
0.0
0.5
1.0
1.5
2.0
2.5
$G_{BP} - G_{RP}$ (mag)