Self-supervised representations for automated astronomical discoveries

Presenter: Koketso Mohale
Supervisor: Prof Michelle Lochner

UNIVERSITY of the WESTERN CAPE

Motives and challenges

Motives:

- Cosmology and galaxy evolution
- Discovering new phenomena

Problems:

- Data volume
- Data complexity

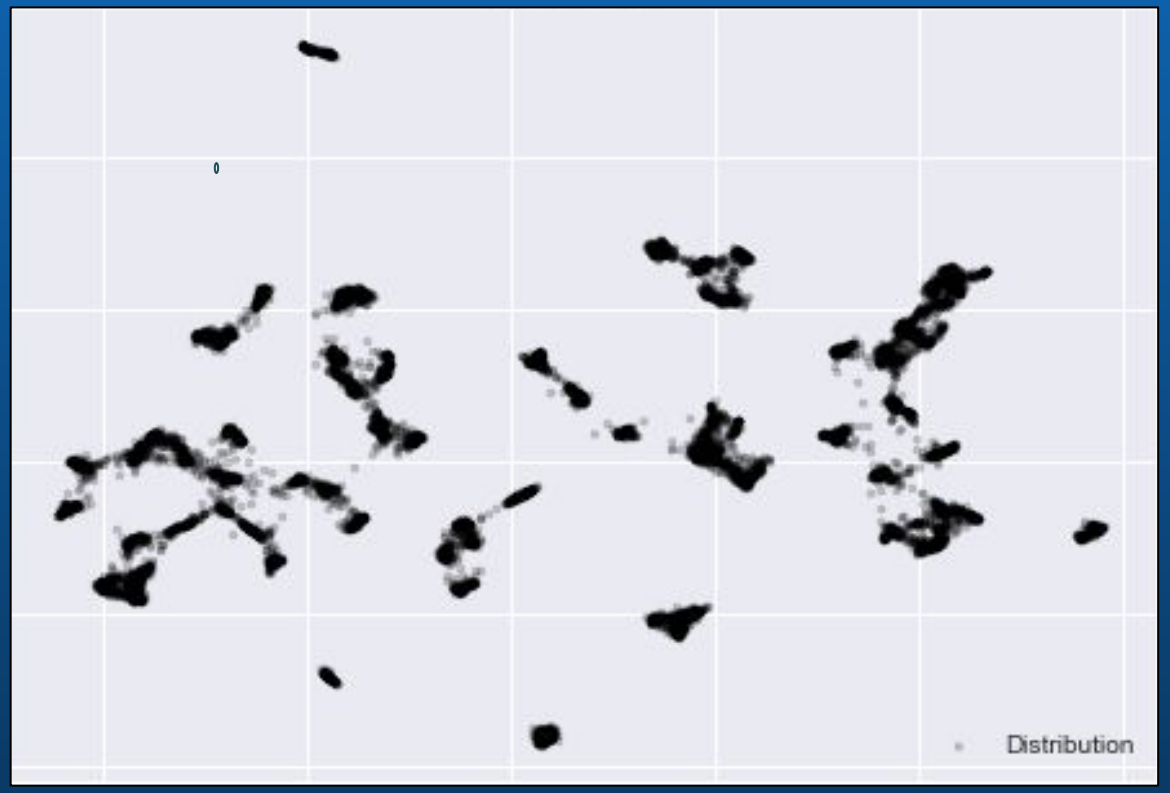
Credit: Square Kilometre Array Observatory

Data volume problem

- During 2007-2019 (**12 years**) *Galaxy Zoo had* ~ ***11 million classification tasks*** [Raddick MJ et al, 2019]
- Too **much data** for volunteers to inspect
- ML (Supervised Learning) published works **rely on quality and quantity of labels**
  - **Assume classes in the data**

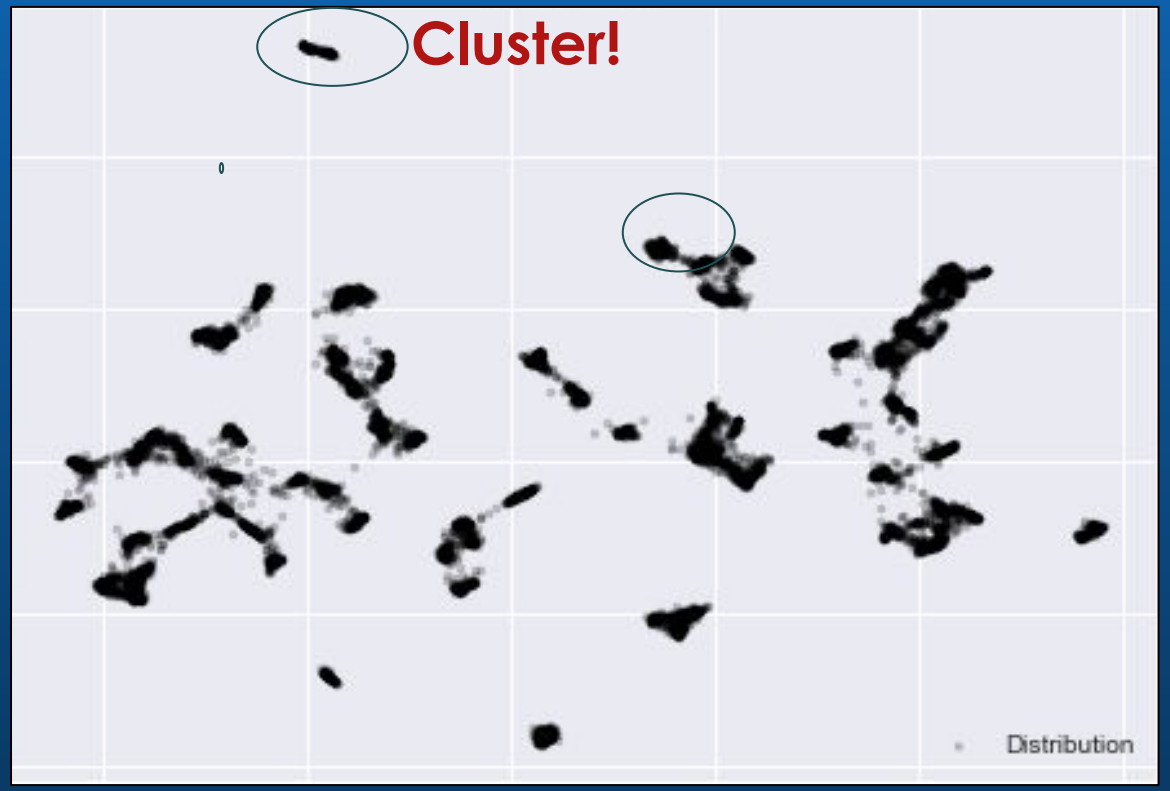
Unsupervised machine learning

- **Unsupervised Learning:**
  - **No need for labelled data**
  - Data exploration



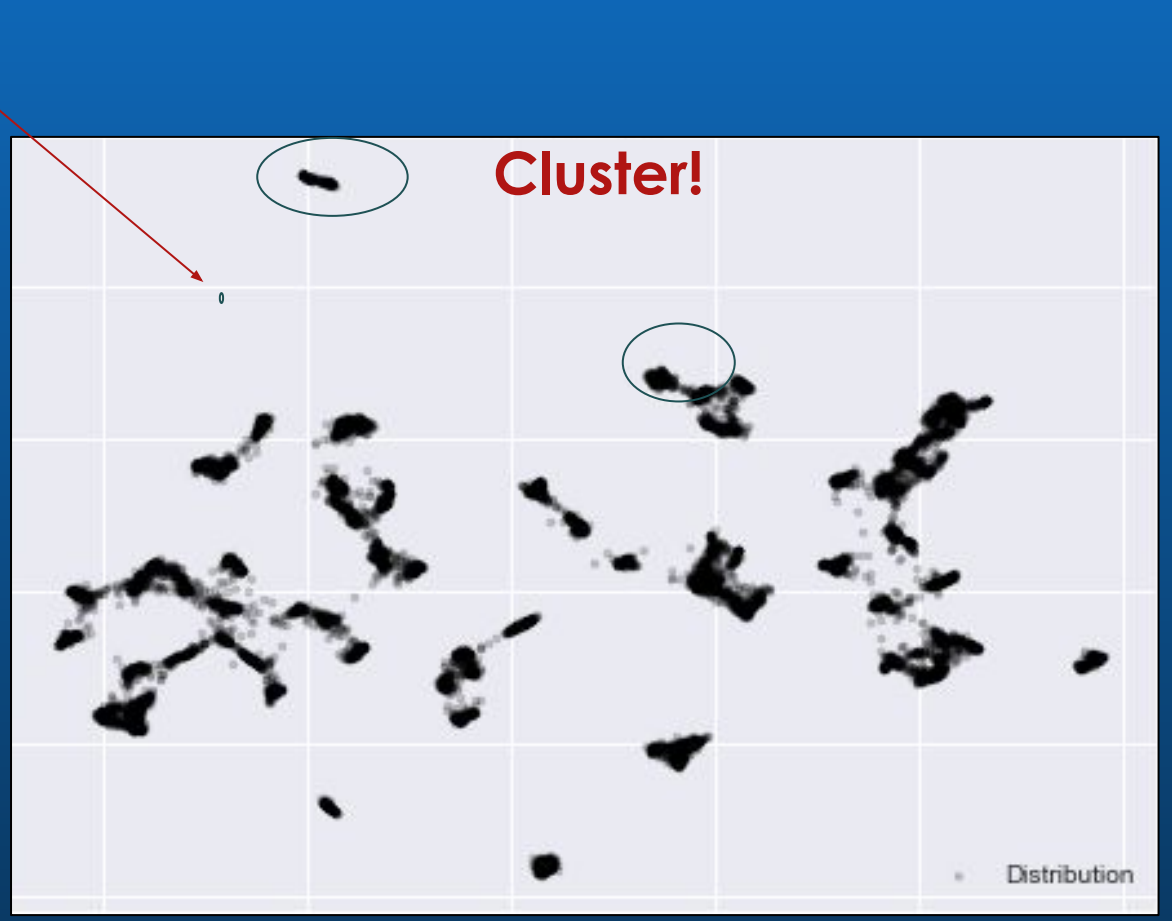
Unsupervised machine learning

- **Unsupervised Learning:**
  - **No need for labelled data**
  - Data exploration
- Useful tasks:
  - **Clustering**: Find the groups



Unsupervised machine learning

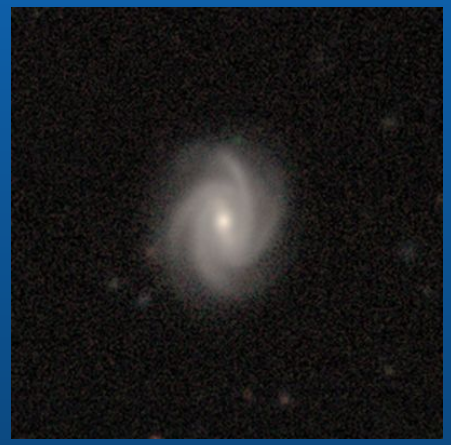
- **Unsupervised Learning:**
  - **No need for labelled data**
  - Data exploration
- Useful tasks:
  - **Clustering**: Find the groups
  - **Anomaly Detection:** Find the odd ones
  - Similarity queries



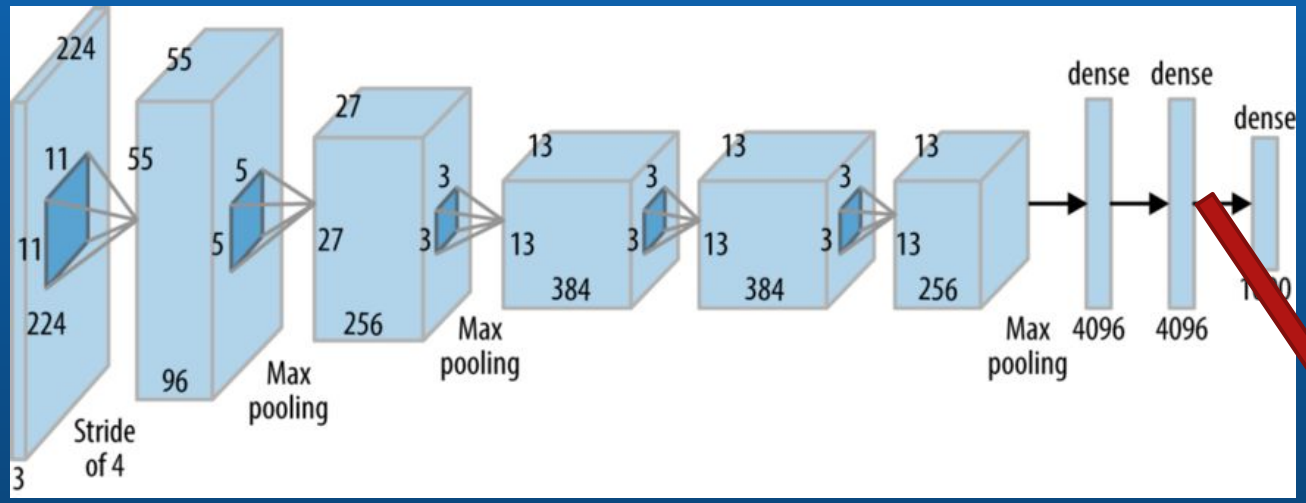
Data complexity challenge

- New telescopes => **high dimensional data**
  - Unsupervised learning algorithms fail:
    - Scale horribly with dimensions
    - **Curse of dimensionality**
    - Need to reduce dimensionality (representation learning)
- ***Unsupervised learning applied to representations (lower dimensional features)***

Deep learning



AlexNet doi:10.1145/3065386

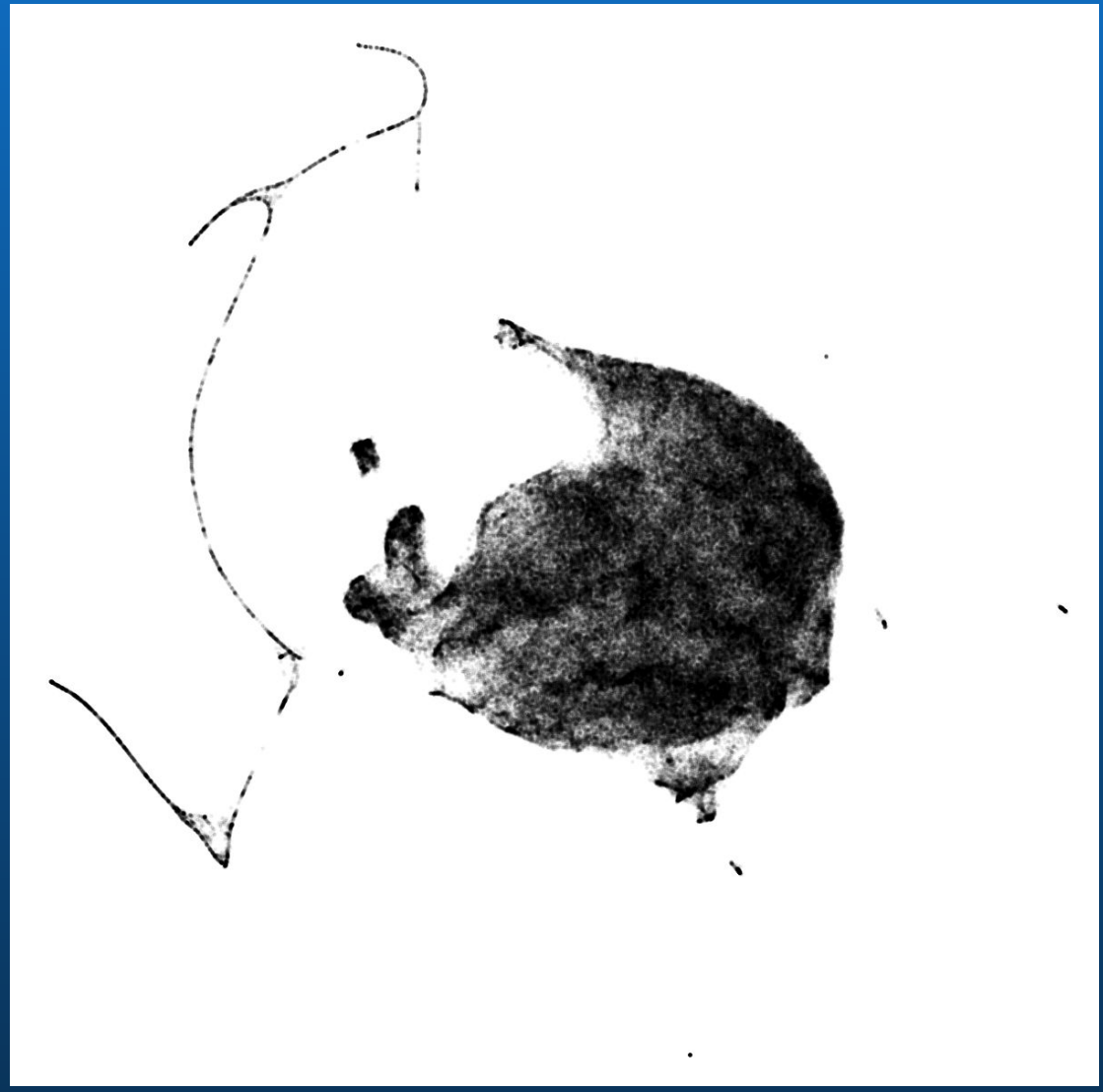


Spiral galaxy representation

- Take a **trained** DNN (Self-supervised learning)
- Use the outputs of a layer as features
  - *Experiments favour the second last layer (arXiv:2101.02767)*

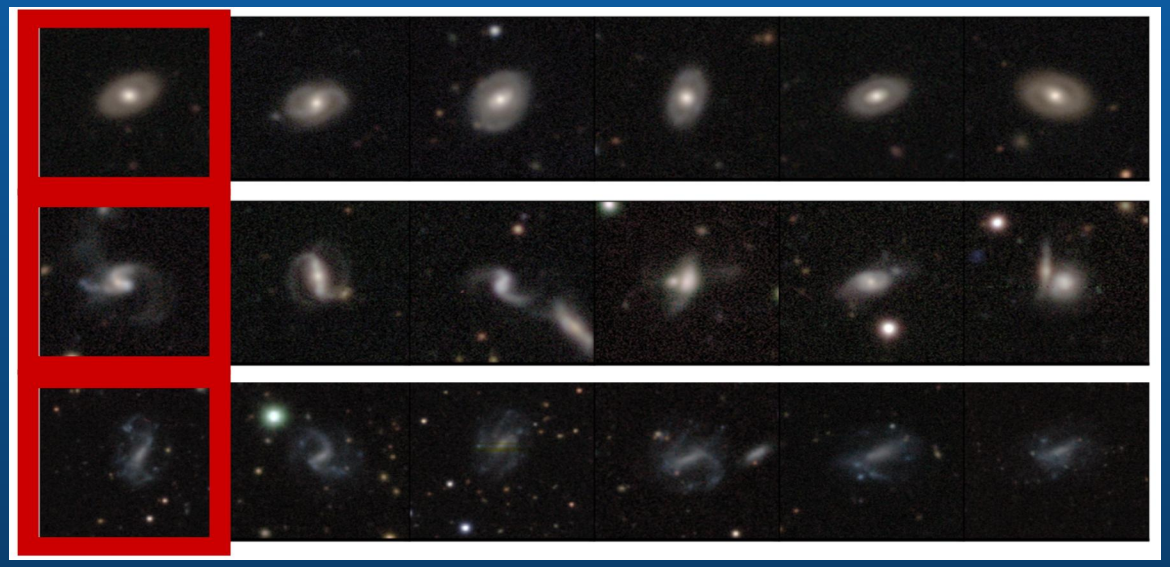
Good representations of Galaxy Zoo mapped to 2D

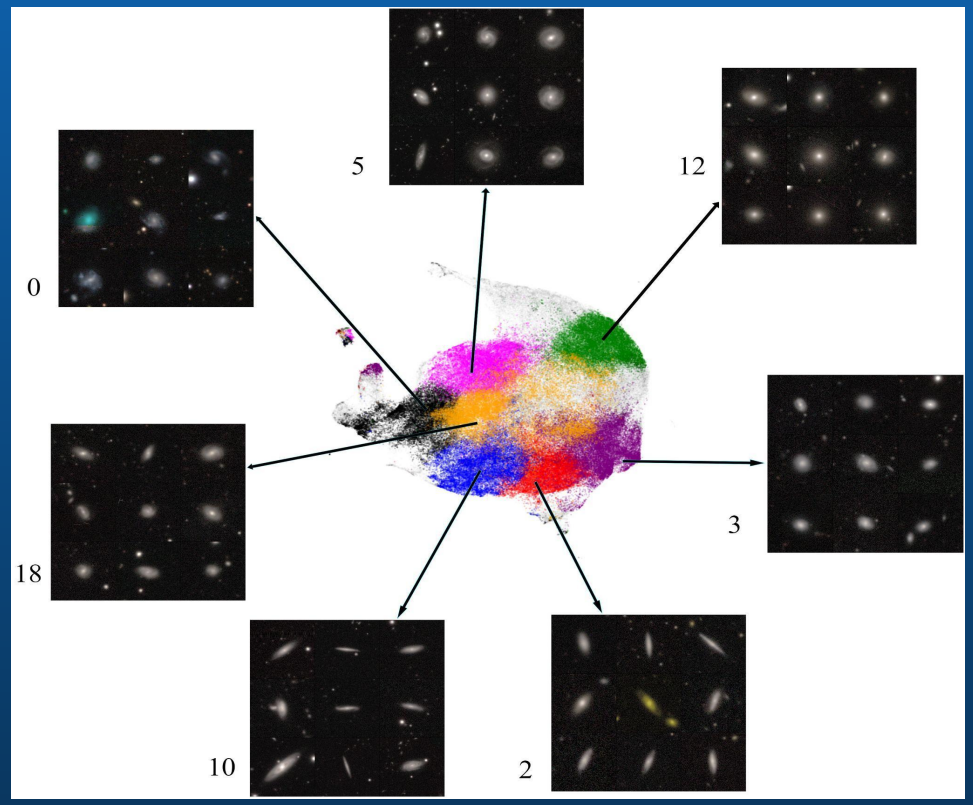
- ~300K Sources from the Dark Energy Camera Legacy Survey DR5[Dey et al. 2019]
- Trained a model using SSL: BYOL
- Representations (512D) mapped to 2D dimensions



What are useful representations?

Similar objects must have neighbouring representations;
Classes in clusters;
Anomalies in low density areas.





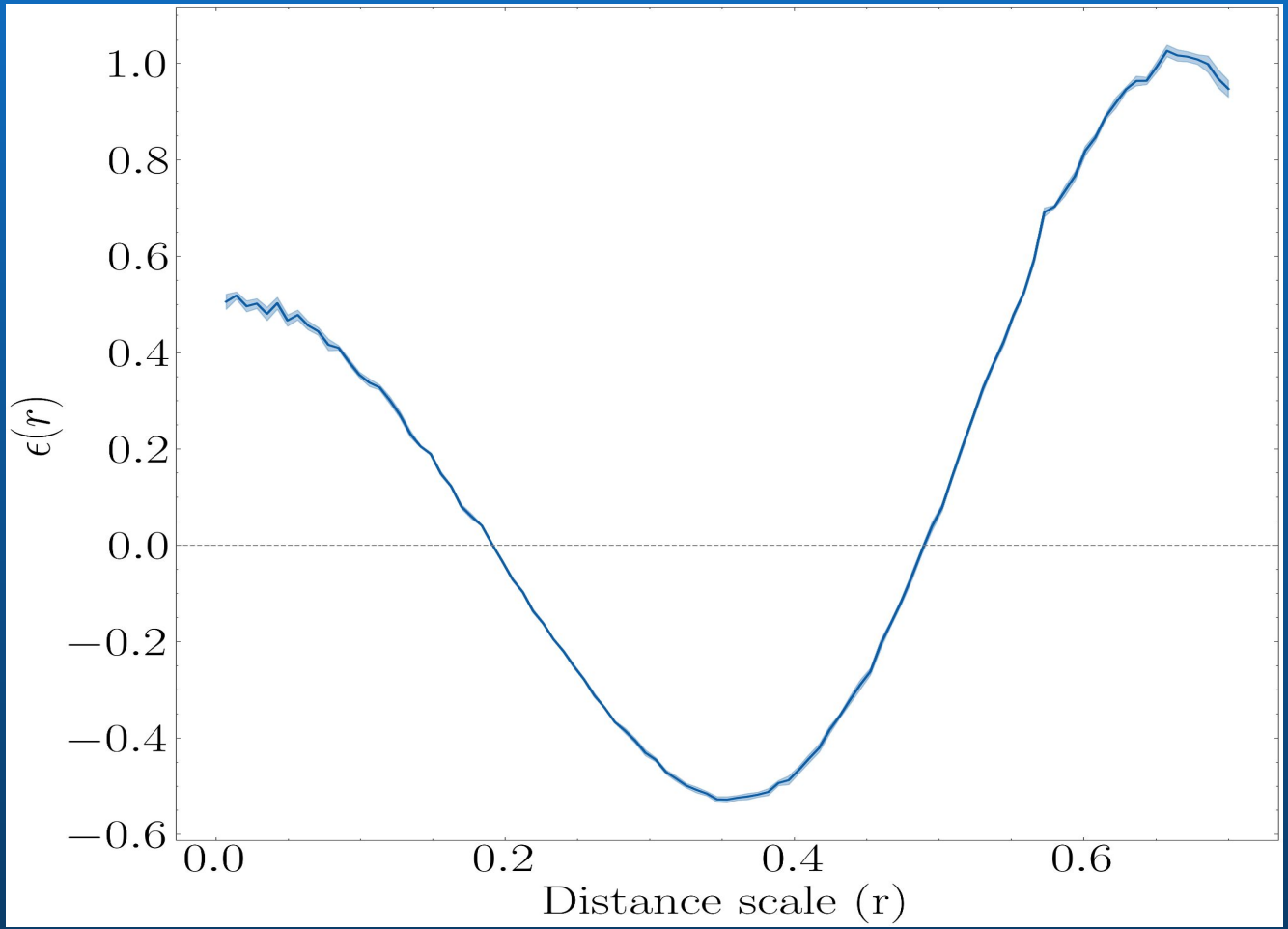
Mohale and Lochner, 2024

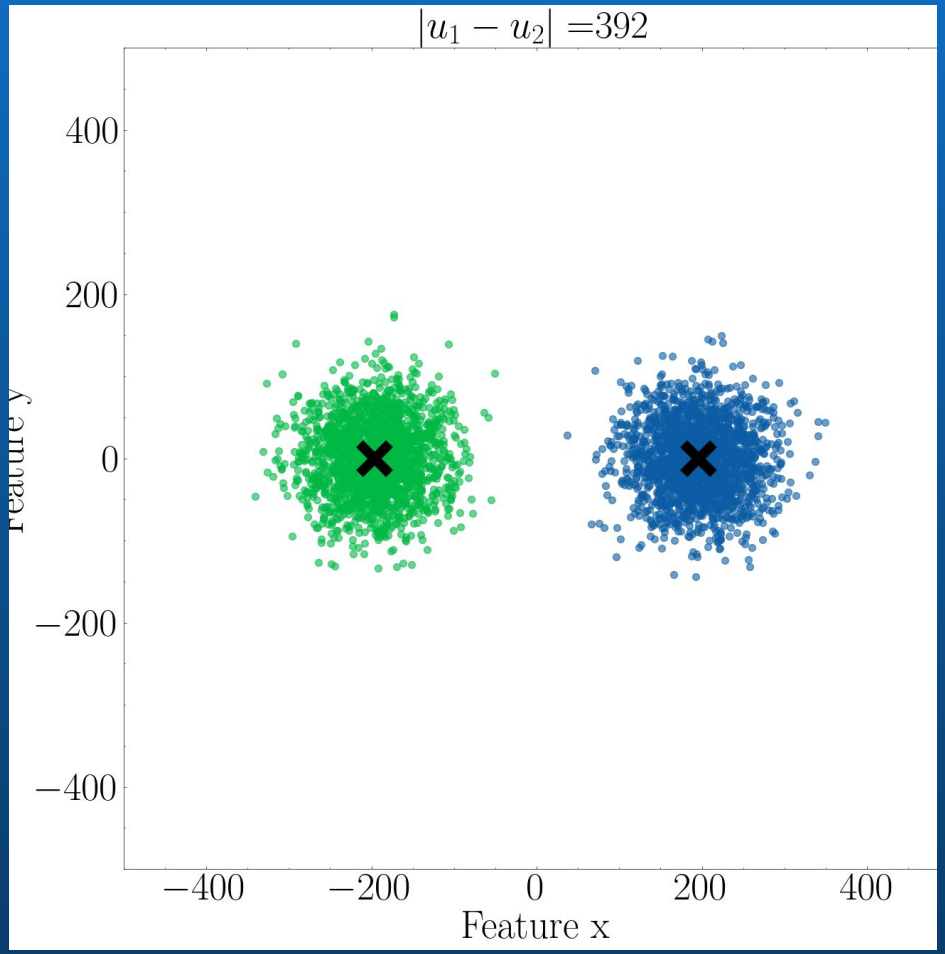
The 2-point correlation function:
A measure of representation utility

- The **2-point correlation function (2PCF)** for deep representations
- Excess **2-point distances as a function of scale**
- Our formulation:
  - How hard is it to describe a representation space as one **generated from a Gaussian**

$$\tilde{\epsilon}(r) = \frac{N_{PR}}{N_{DP}} \frac{DD(r)}{RR(r)} - 1$$

A cluster signal

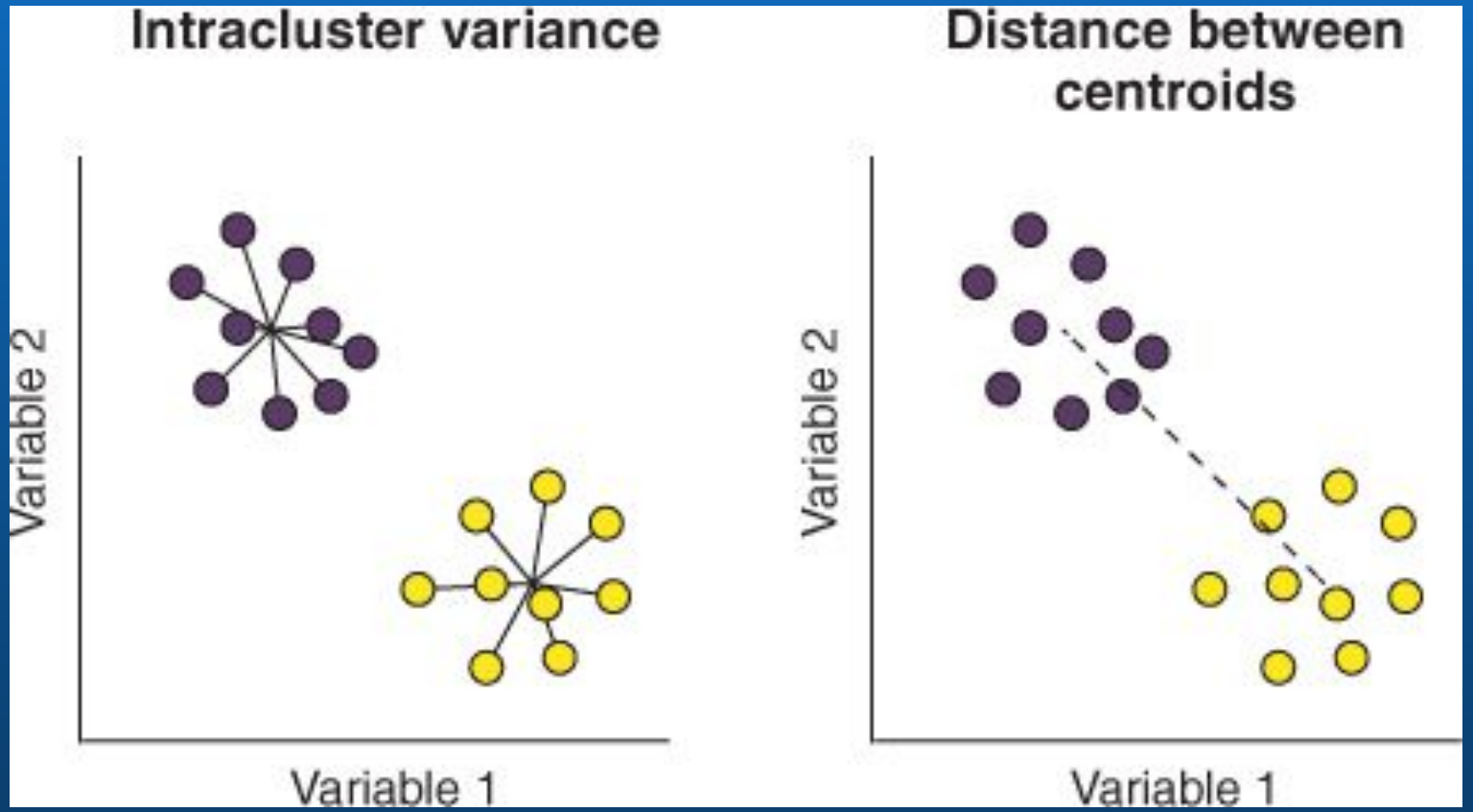




- Fatemi-Ghomi et al. (1999): A cluster appears as a bump on the 2PCF

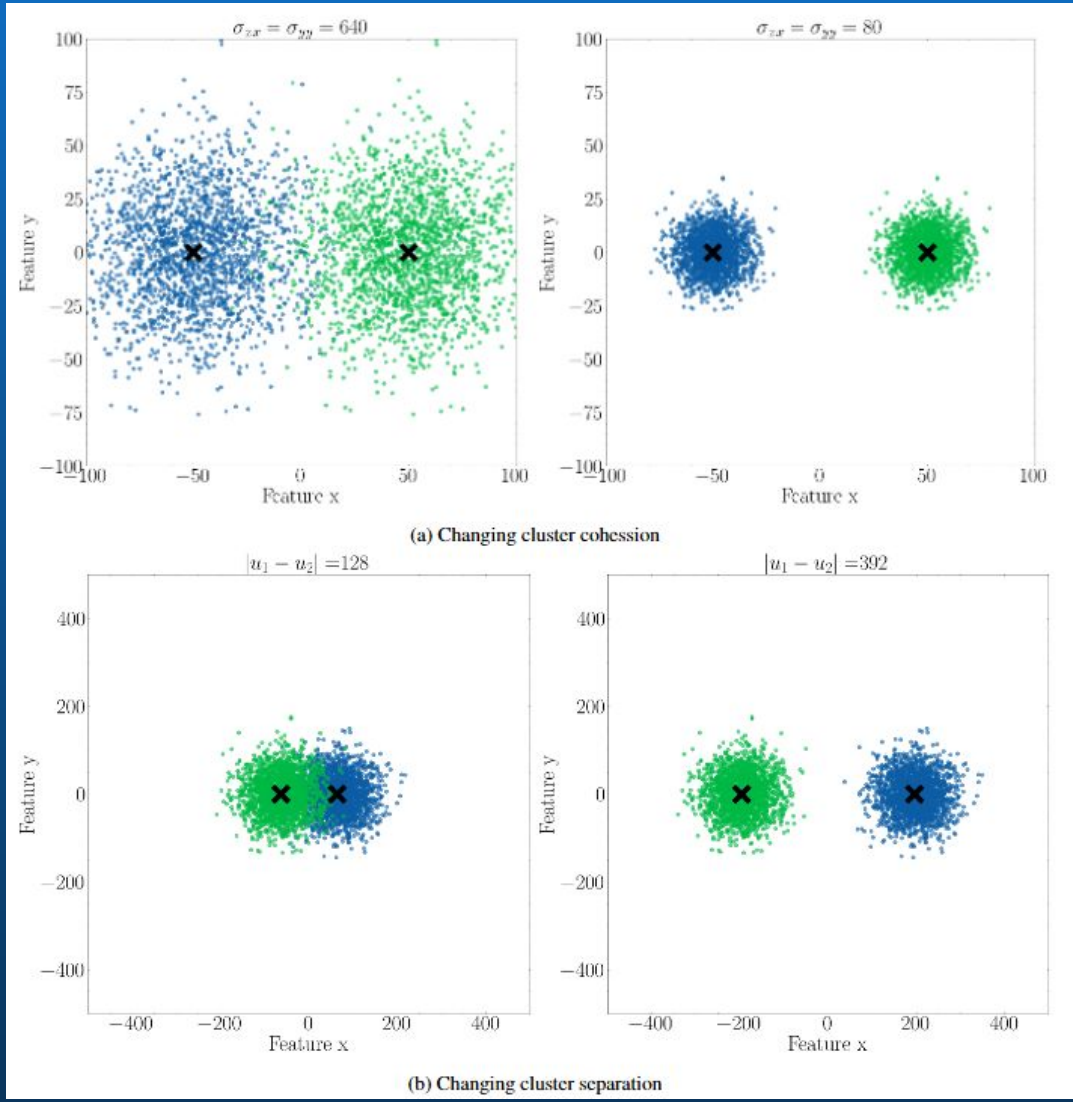
What is structure?

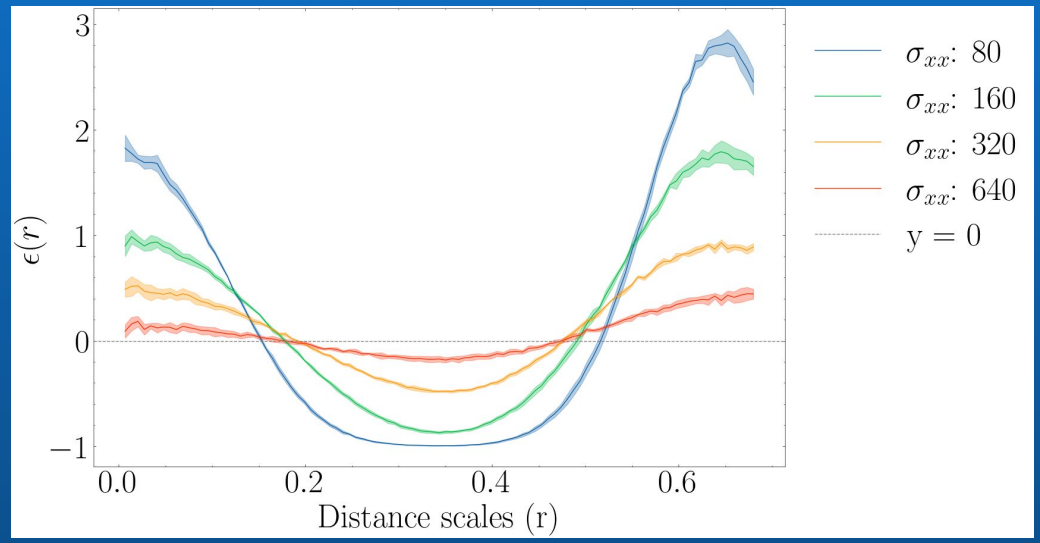
- Existing clustering metrics measure:
  - Intra-cluster cohesion.
  - Intercluster distances.
- Cluster metrics measure both
  - **Davies-Bouldin score**, Silhouette score, etc.

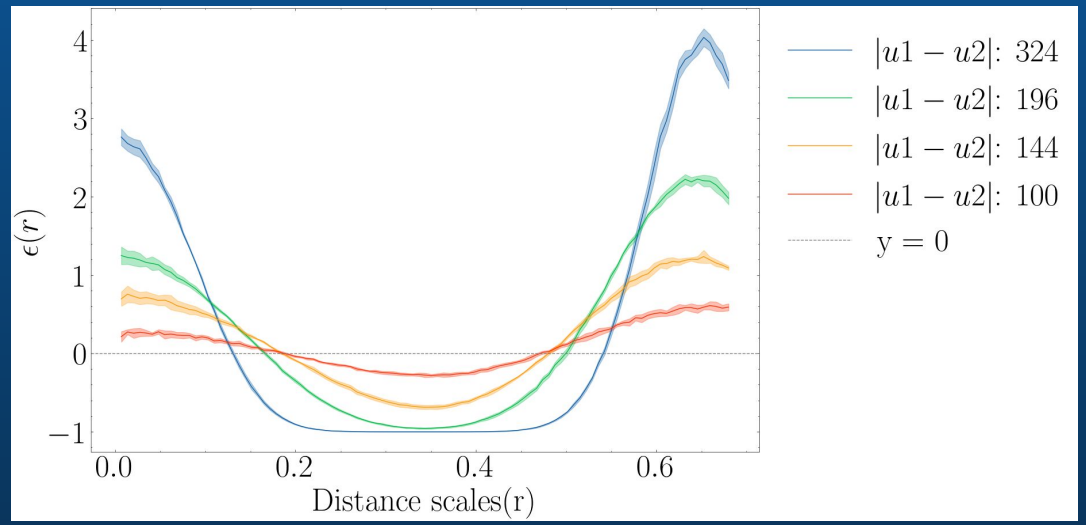


Davies and Bouldin,1979; Rousseeuw,1987

2PCF detects changes in structure





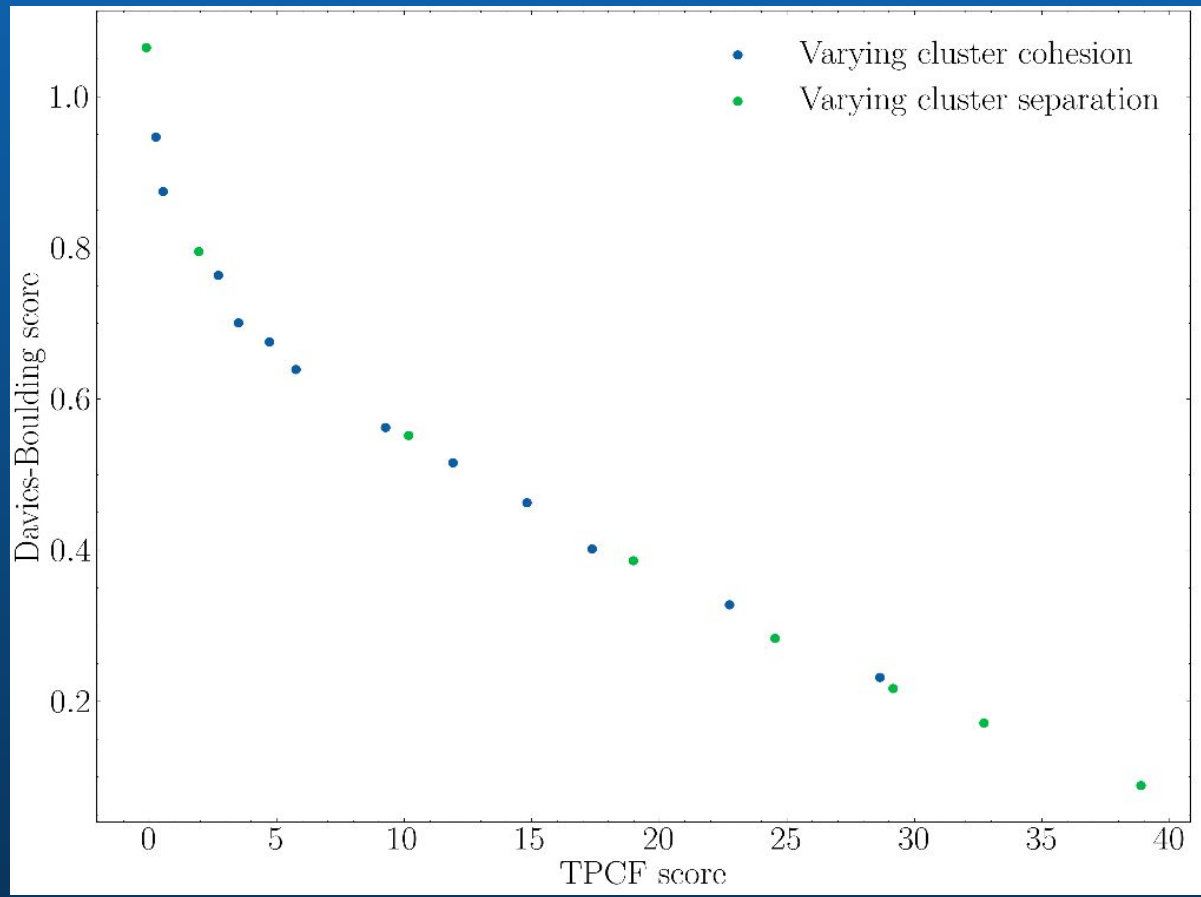


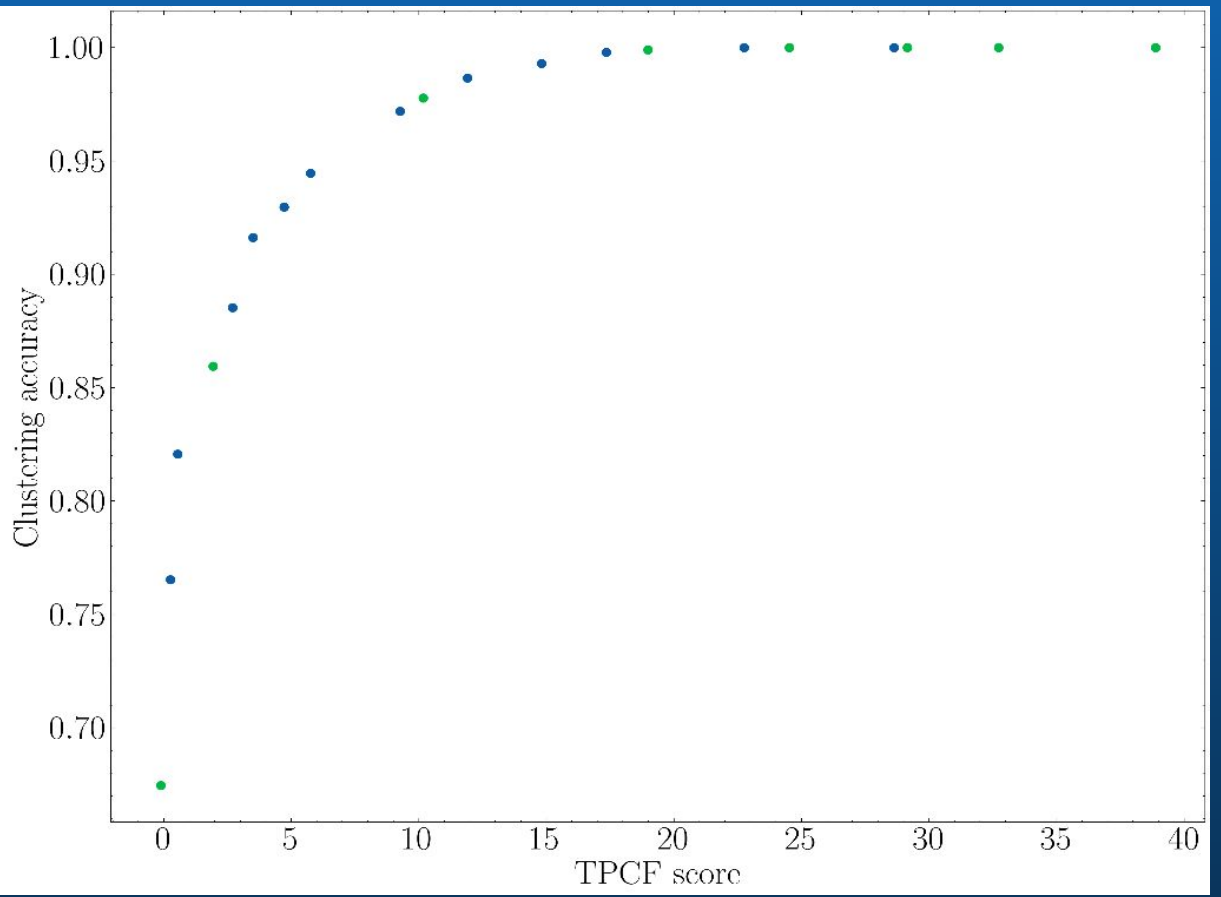
Structure summary

- We define the 2PCF score:
- Compare against:
  - Davies-bouldin on class labels
  - Clustering accuracy on K-Means assignments

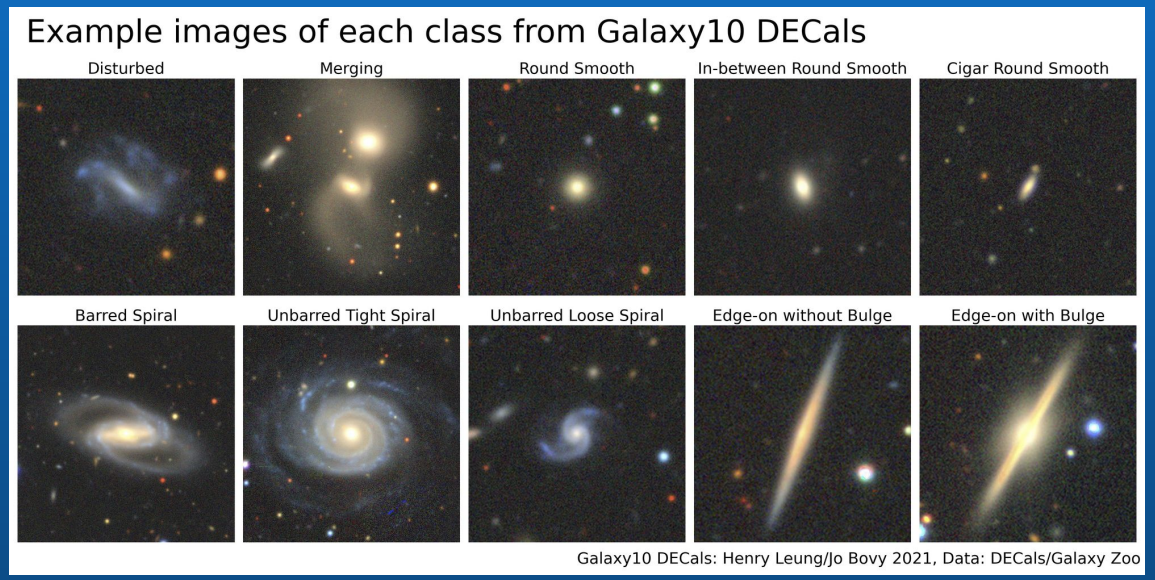
$$TPCF\_score := \sum_{i}^{N} \tilde{\epsilon}(r_i) \pm \sum_{i}^{N} \sigma(r_i)$$

- **TPCF score correlates with clustering metrics**
  - Case ideal for clustering metric
- Independent of clustering algorithm
- Independent on labels

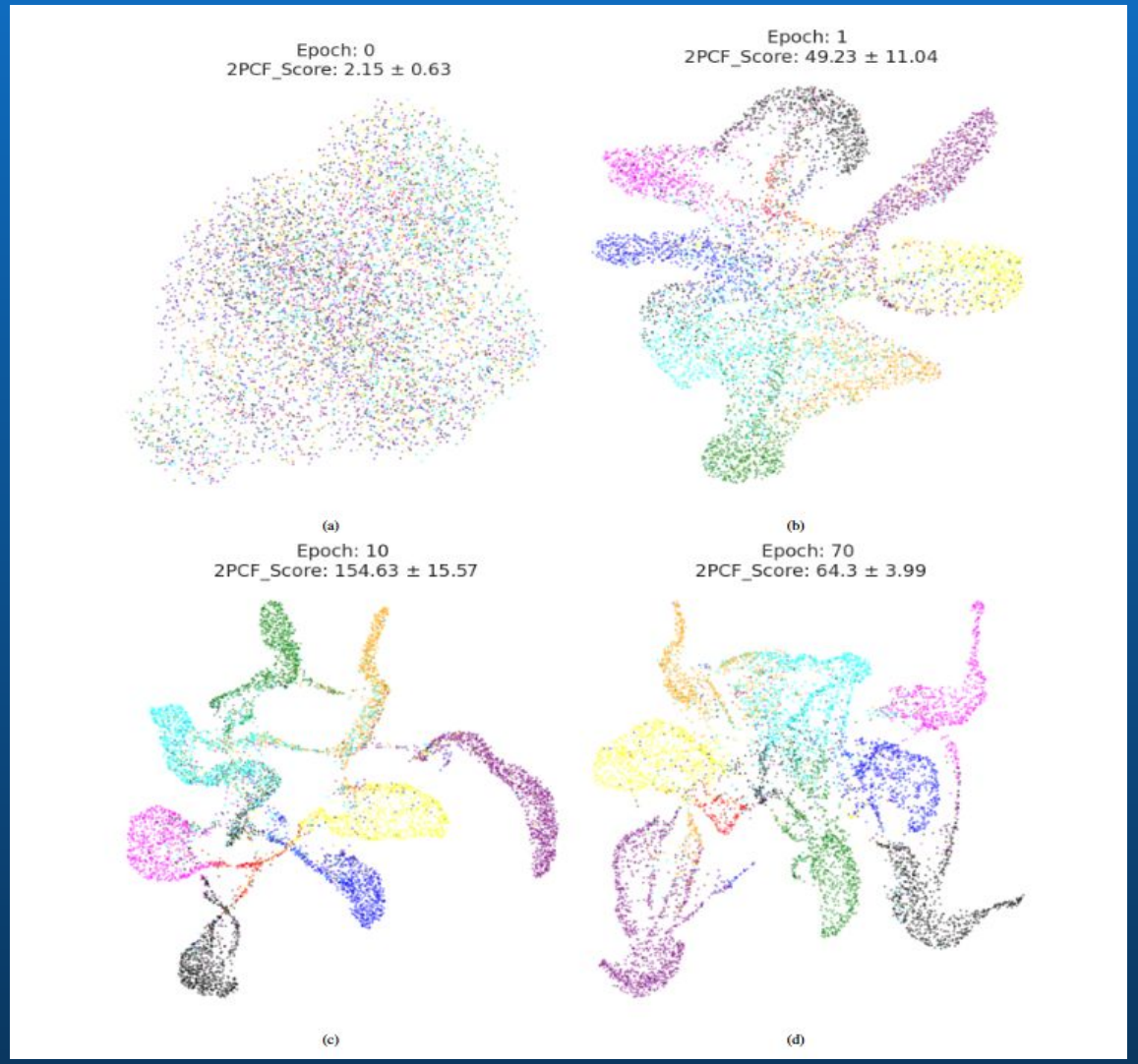




On deep representations of Galaxy10 DECaLS



- Representations Galaxy10 *(Leung and Bovy 2019)* from a model at different training stages.
  - Here mapped to 2-D (UMAP)
- Better deep representations correspond to higher 2PCF scores (on PCA components)



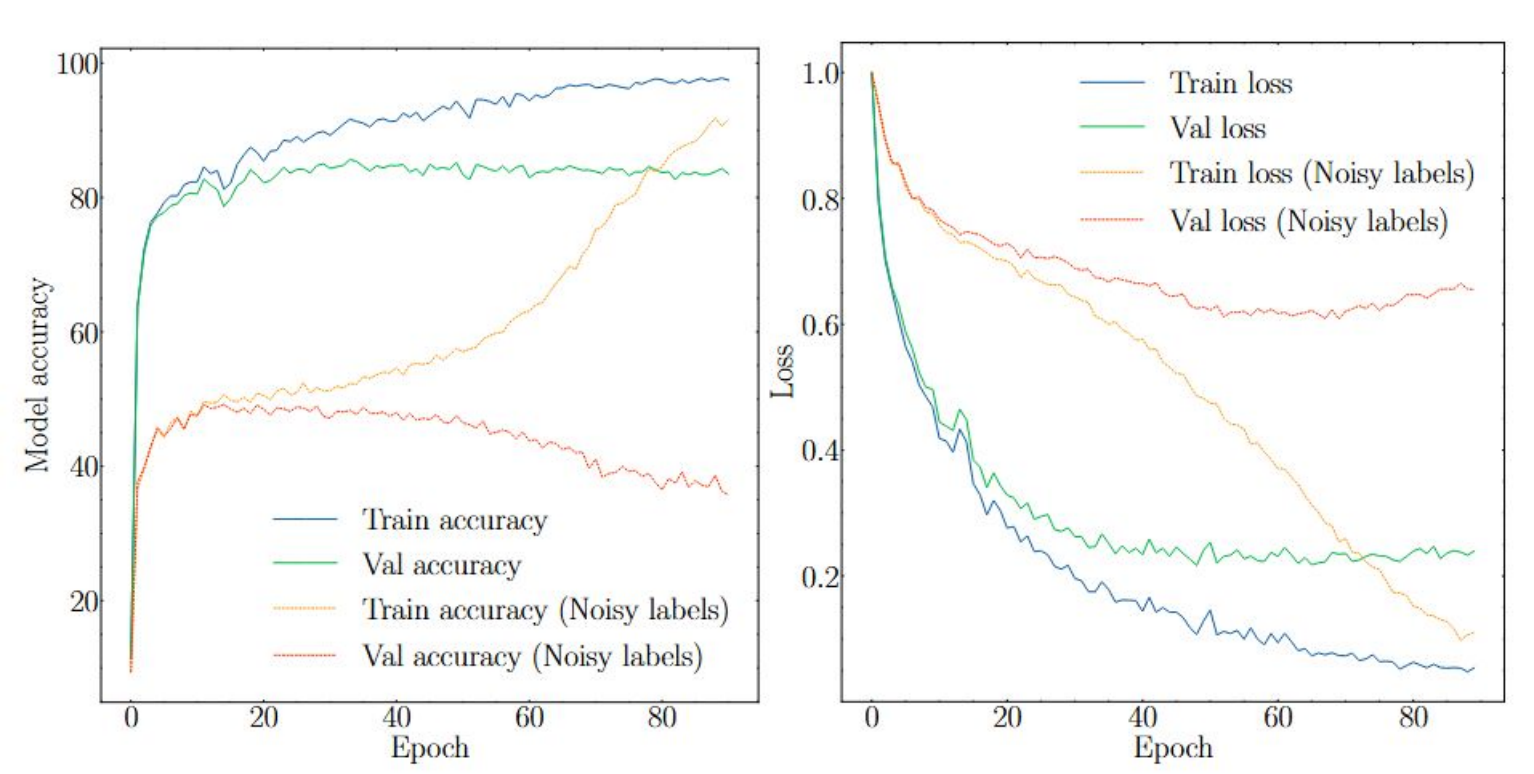
Email: Koketso.kjay@gmail.com

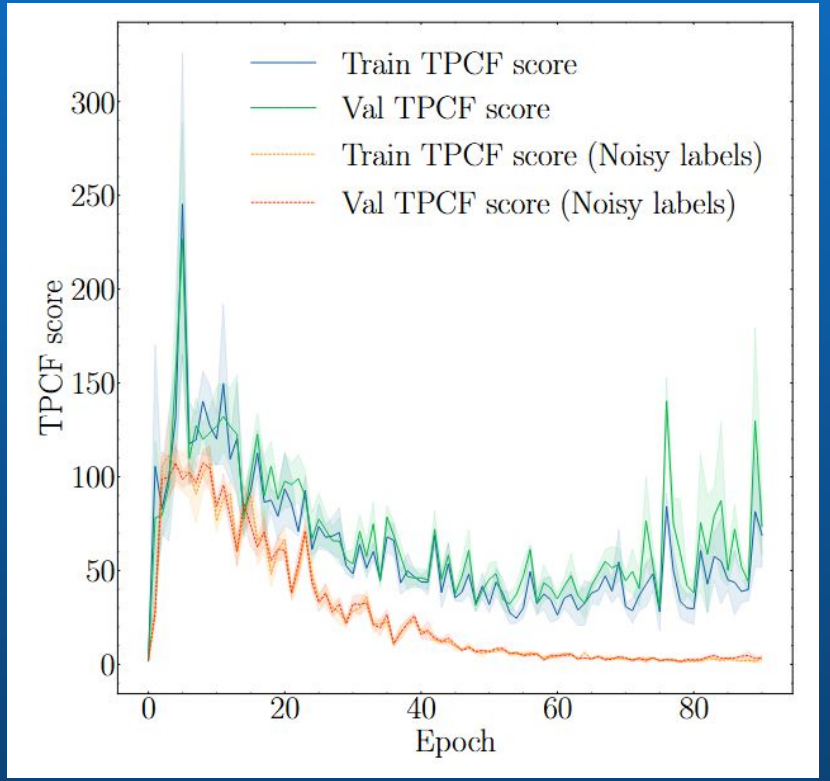
Conclusion

- Unsupervised learning (UL) is a promising pathway for mining large astronomical datasets
    - Unsupervised learning requires that low dimensional representations with structure.
- Good representations allow for automated clustering and anomaly detection
- We designed a methodology based on the 2-point correlation function for determining the utility of representations for UL
- We showed the ability to detect structure in simulated feature spaces

We aim to:

- Develop methods for gauging the quality of deep representations that
    - do not depend on models used, or rely on labels
    - specific for unsupervised learning downstream tasks
- Demonstrate utility for improving representation learning methods


Model accuracy
Epoch
Train accuracy
Val accuracy
Train accuracy (Noisy labels)
Val accuracy (Noisy labels)
Loss
Epoch
Train loss
Val loss
Train loss (Noisy labels)
Val loss (Noisy labels)


TPCF score
Epoch
Train TPCF score
Val TPCF score
Train TPCF score (Noisy labels)
Val TPCF score (Noisy labels)

- Intra-Cluster Distance

$$S_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k)$$

- Inter-Cluster Distance

$$M_{kl} = d(\mathbf{c}_k, \mathbf{c}_l)$$

- Define

$$R_{kl} = \frac{S_k + S_l}{M_{kl}}$$

- Let

$$R_k = \max_{\substack{1 \le l \le K \\ l \ne k}} R_{kl}$$

- Davies-Bouldin Index

$$DB = \frac{1}{K} \sum_{k=1}^{K} R_k$$

- Small Davies-Bouldin Index indicates the clusters are compact and cluster centers are far away.