# Dimensionality Reduction Methods and relevant applications to galactic images

Final Thesis Presentation

Dillon Loh Guan Hui (061801914)

# Flow of today’s presentation

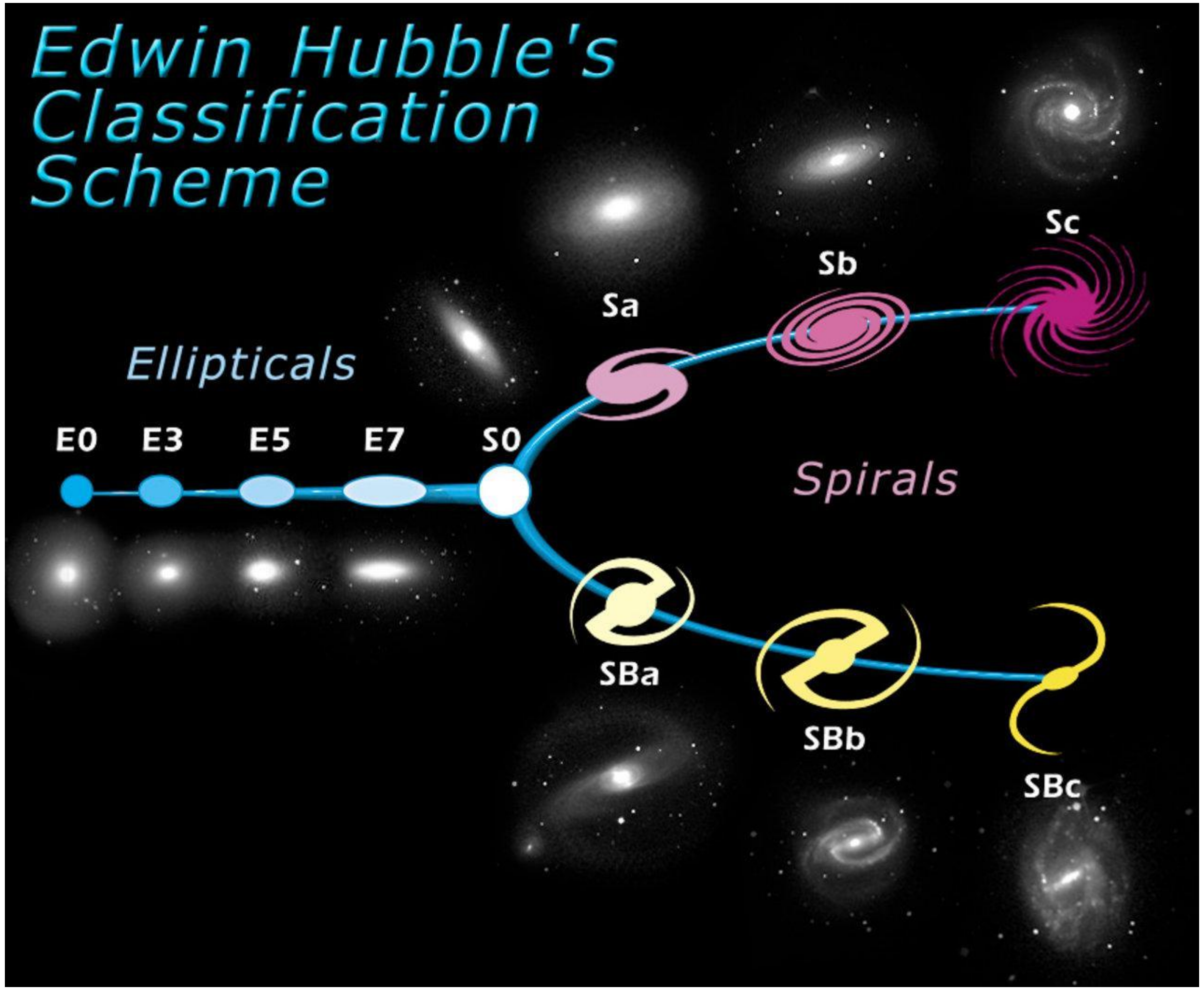
1. **Introduction**
   - Morphologies + Morphologies in Galactic Images
   - Dimensionality Reduction
   - Past Works
   - Morphology vs Physical Properties
2. **Data Preparation**
3. **Principal Component Analysis**
   - Methodology
   - Results/Discussion
4. **Autoencoders**
   - Methodology
   - Results/Discussion
5. **Conclusion**

# 1.	Introduction

この部分だけ日本語でやります。

# Galaxy Morphologies

- Galaxies are systems of stars and interstellar matter.
  - Come in all sorts of shapes and have defining features
- They are commonly grouped according to the ‘Hubble Tuning Fork’ classification system
- Galaxies that do not fall under Ellipticals or Spirals are called ‘Irregulars’

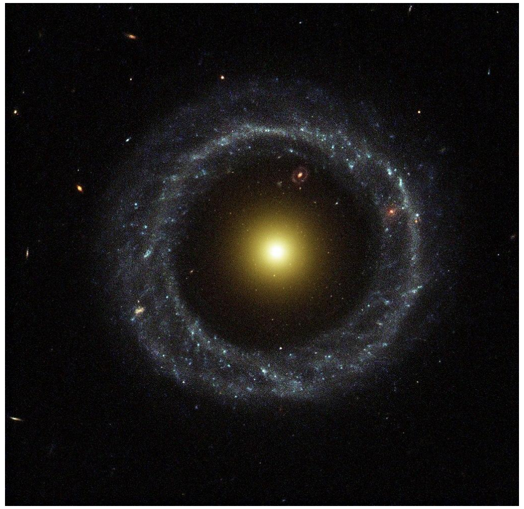


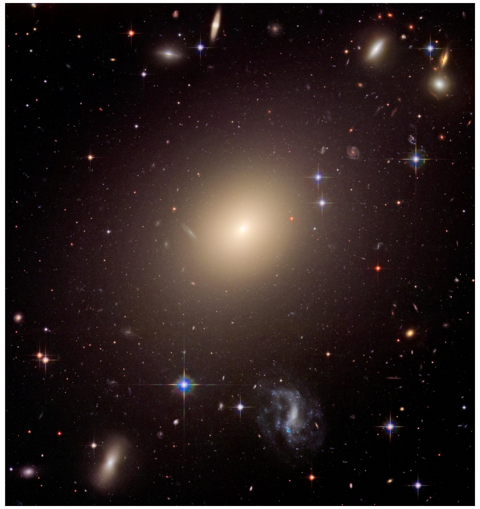
NASA

# Spatially-resolved Galactic Images

- Galactic images reveal the morphology of galaxies

ESA/Hubble





NASA and The Hubble Heritage Team

- These galaxies exhibit features that are similar.
  - The images themselves can be broken down into shared 'components' -> How?
- Each image is high dimensional (n x n pixels)
  - Not all dimensions are unique.
  - Could we also reduce redundant dimensions for better analysis?

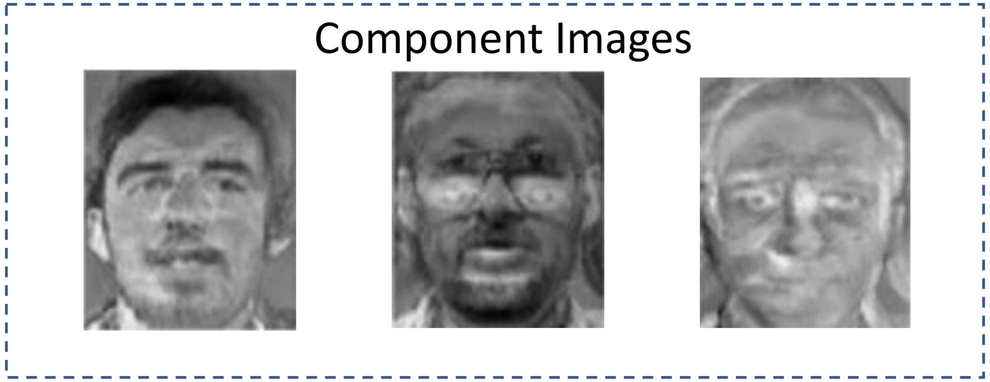
# Dimensionality Reduction?

- It can be <u>difficult to work with/extract information</u> from high-dimensional data.
  - ‘Curse of Dimensionality’

- Naturally, we want to find ways to ‘compress’ this information.
  - ‘Dimensional reduction’

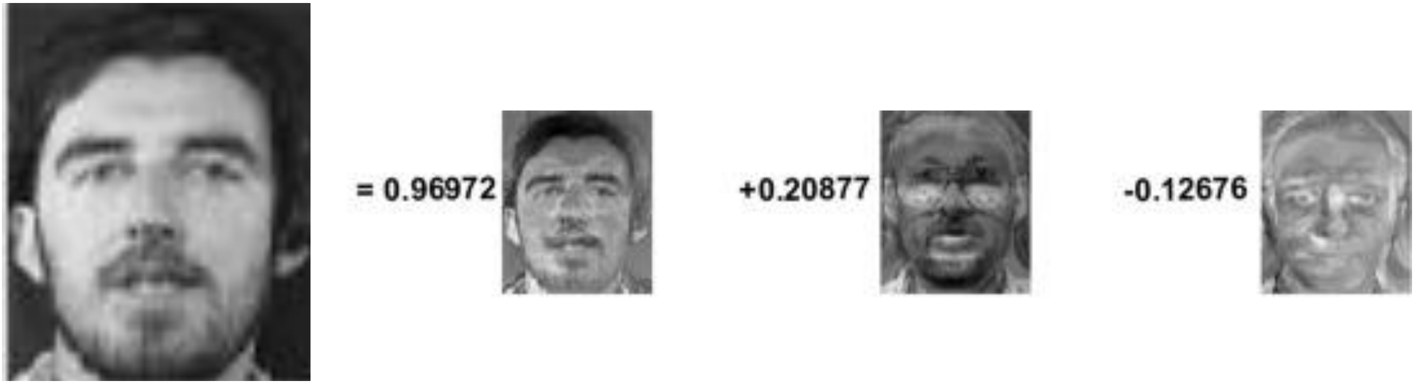
What if all faces in the world could be decomposed into a sum of these 3 component images?

⇩



Component Images

This image is linear combinations of the 3 component images above.

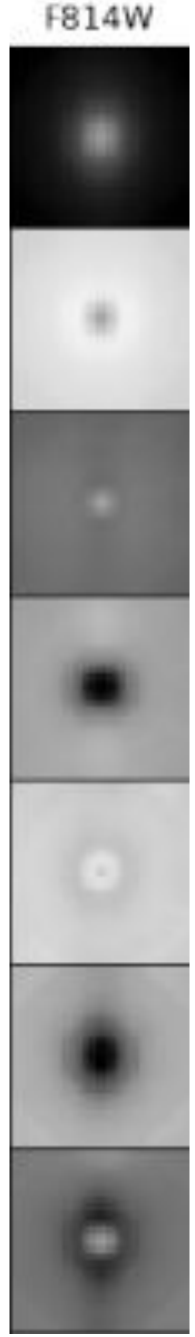


# Past works on Galactic Image Decomposition

*“Eigengalaxies: describing galaxy morphology using principal components in image space”* – 2020

- Demonstrated that galactic images could be decomposed into a linear sum of ‘eigen-galaxies’ via PCA.
  - Sample of 10243 galaxies represented by just **12 eigengalaxies**

- Massively **reduces dimensions** required to describe galactic structural properties



(Uzeirbegovic et al., 2020)

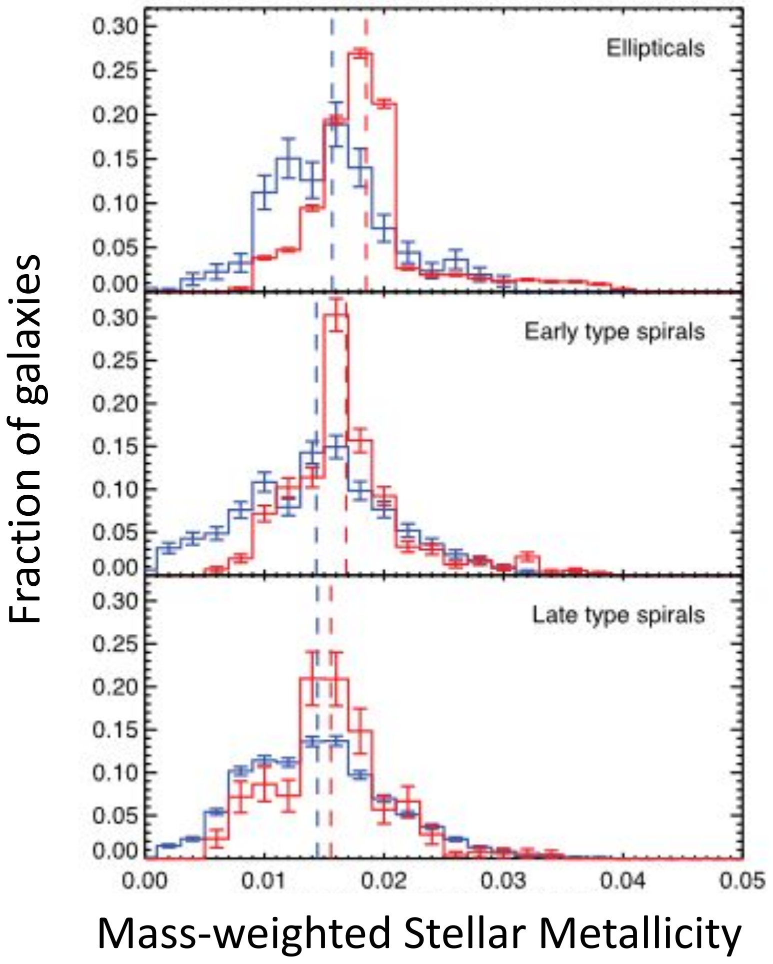
# Physical Properties of Galaxies

In general, there are well-defined connections between the morphological properties of a galaxy and its physical property. *

- Ellipticals tends to be older.*
- Ellipticals tend to have higher metallicity
- Ellipticals tend to have lower velocity distribution than spirals *
- Spirals tend to have higher SFR than Ellipticals**
- Spirals' bulges tend to be older than their disc components

**Can we then find relations between the low-dimensional representation of galactic images and physical properties?**

Fraction of galaxies with different metallicities for Ellipticals and Blue/Red Spirals



Tojeiro et al., 2013

# 2. Data Preparation

ここから英語で話します。

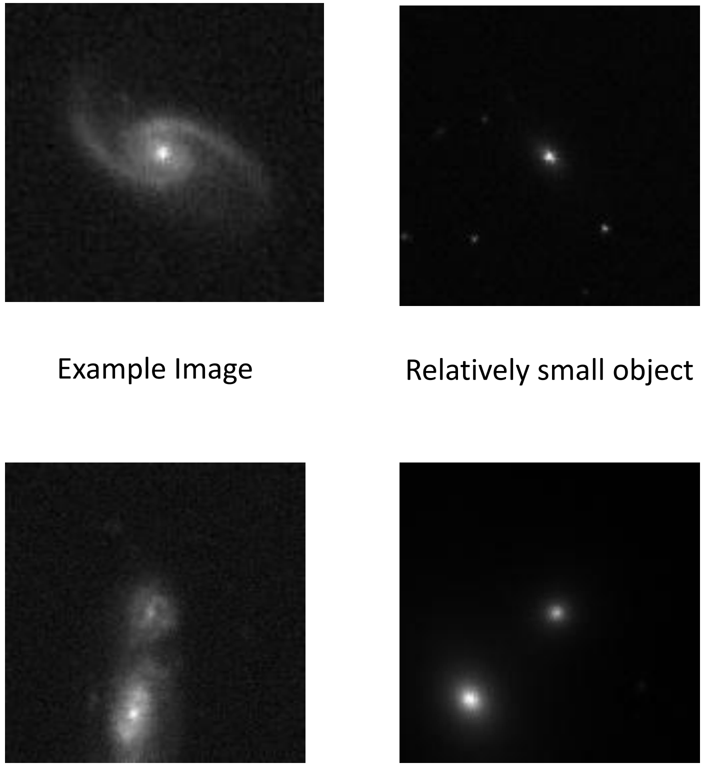
# Data used

- Sloan Digital Sky Survey (SDSS) is a multi-spectral imaging and redshift survey.
  - Data Release 17 optical images are downloaded via SkyView*.

- SDSS *i* Band Images Used
  - Near-Infrared wavelengths sensitive to long-living stars' emission
  - Representative of stellar mass distribution

- 150x150 pixel images

**Example Images**



Example Image

Relatively small object

Center object disturbed by extra objects

Presence of extra objects that have negligible disturbance

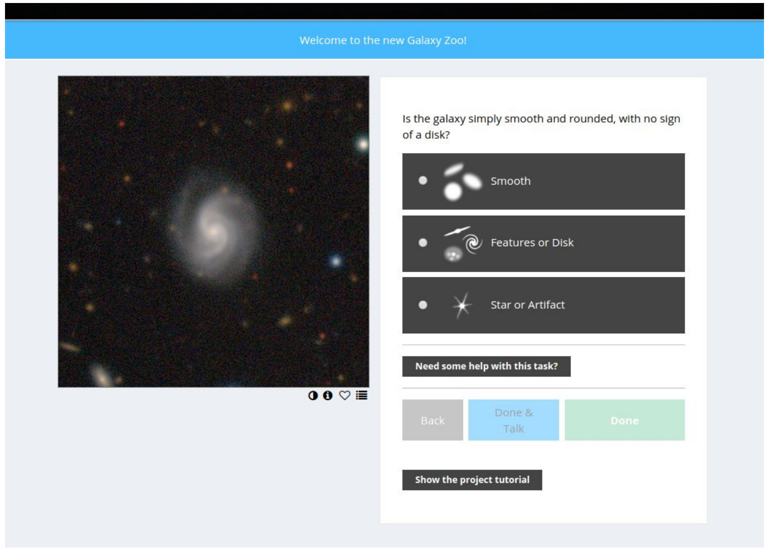
*https://skyview.gsfc.nasa.gov/current/cgi/titlepage.pl*

# Filtering by Galaxy Morphology with Galaxy Zoo 2

- Galaxy Zoo is a project **where humans assist in the manual morphological classification of galaxies**

- Classifications done include:
  - Spiral/Elliptical/Irregular
  - Presence of oddities (rings, merging, etc.)
  - Extent of morphological features (spiral arm tightness, bulge size, etc.)

- For each classification, a **clean flag** is raised if **>80%** of the responses to a question are the same.
  - We use these clean flags to decide if a galaxy belongs to a certain class.



Welcome to the new Galaxy Zoo!

Is the galaxy simply smooth and rounded, with no sign of a disk?

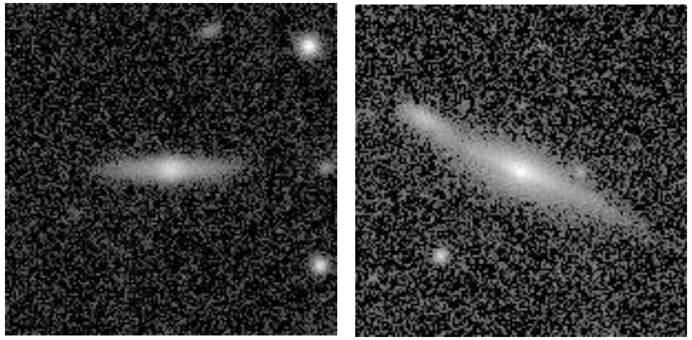
https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

# Data Filters

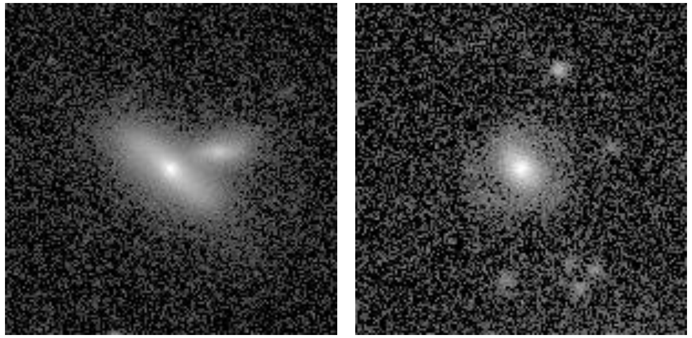
Images used were filtered based on several conditions:

- Edge-on galaxies rejected
  - Difficult to observe morphological features.
- Galaxies with ‘oddities’ rejected
  - Galactic merging, overlap, lensing, dust lane, etc.
- Of the filtered galaxies, 12539 galaxies were successfully downloaded via SkyView within a reasonable allocated time.

Edge-on

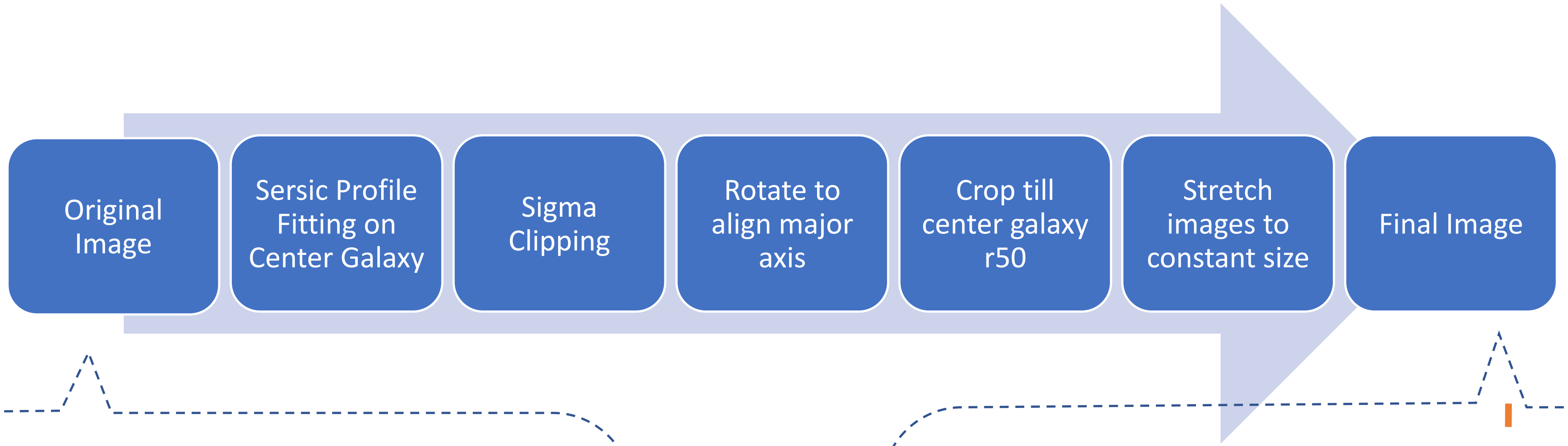


Oddities



*https://skyview.gsfc.nasa.gov/current/cgi/titlepage.pl*

# Preprocessing of Data



Original Image

Sersic Profile Fitting on Center Galaxy

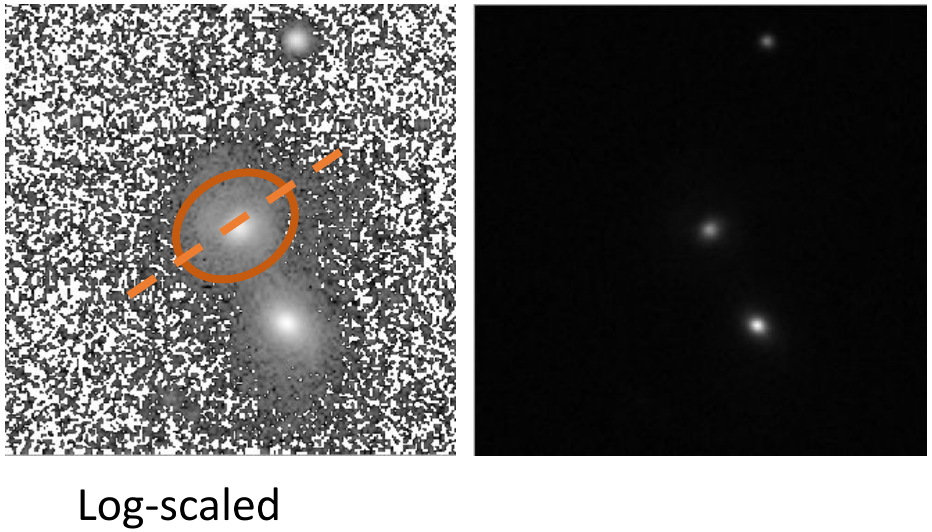
Sigma Clipping

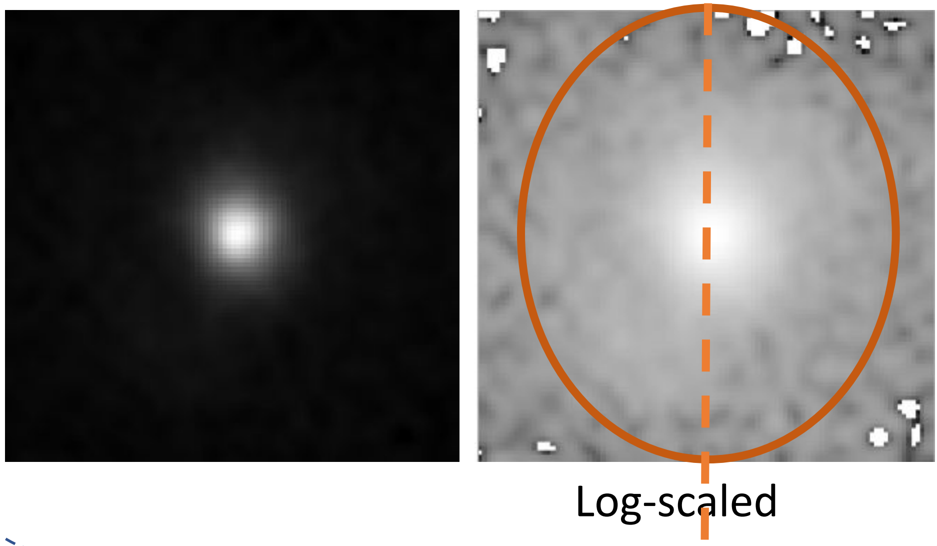
Rotate to align major axis

Crop till center galaxy r50

Stretch images to constant size

Final Image





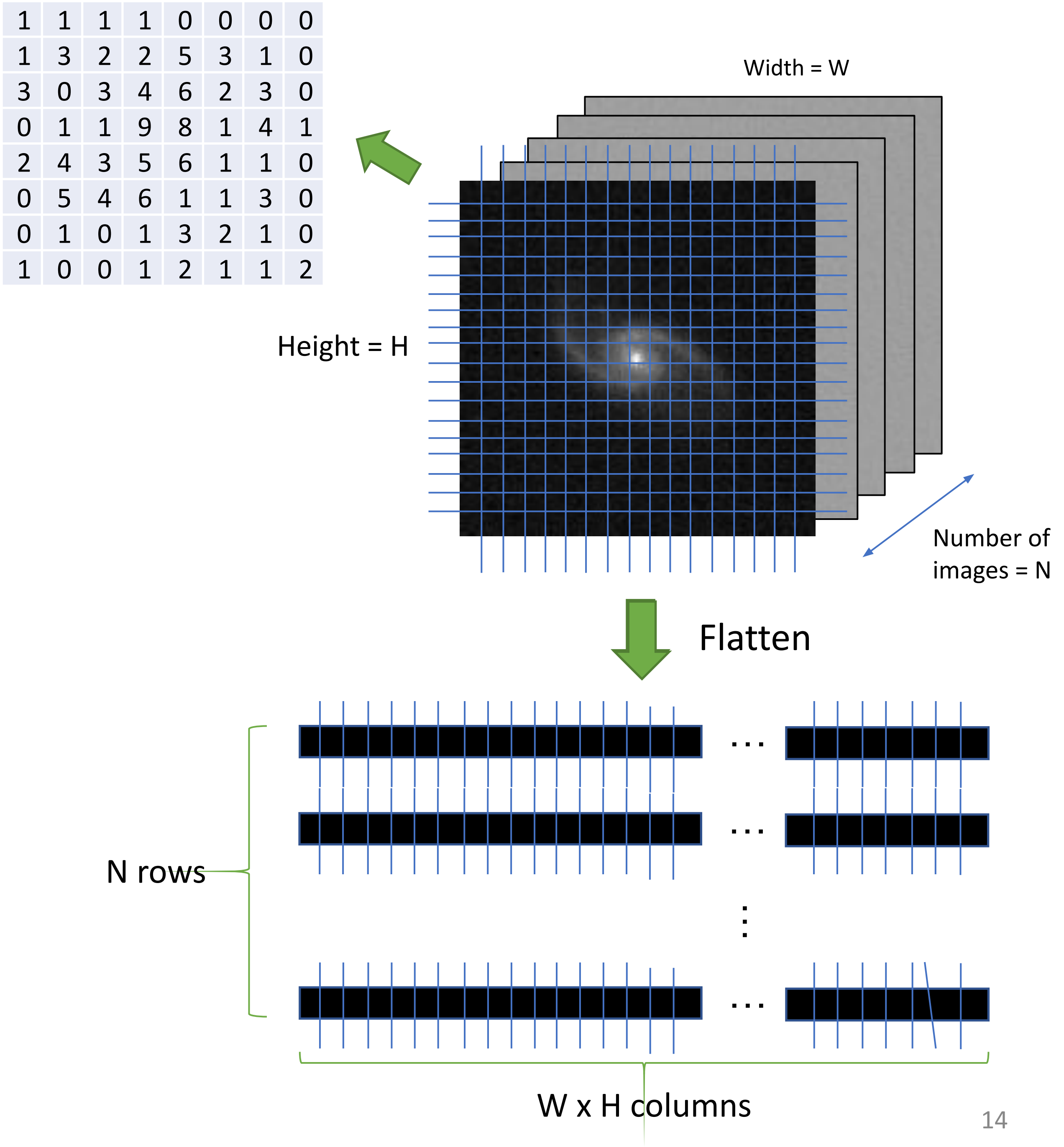
# Flattening

A galactic image is just arrays of pixel values that can be flattened into a vector.

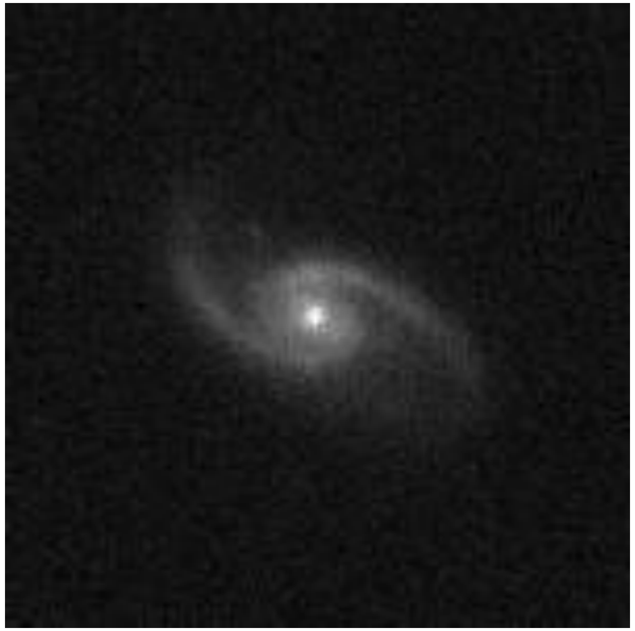
By considering sets of galactic images, one can construct a data set of images to apply PCA.

- Rows = One Galaxy
- Columns = Pixel Position in Image

We will then have a typical dataset where each row is a datapoint, and each column represents the ‘brightness at position i’



|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 3 | 2 | 2 | 5 | 3 | 1 | 0 |
| 3 | 0 | 3 | 4 | 6 | 2 | 3 | 0 |
| 0 | 1 | 1 | 9 | 8 | 1 | 4 | 1 |
| 2 | 4 | 3 | 5 | 6 | 1 | 1 | 0 |
| 0 | 5 | 4 | 6 | 1 | 1 | 3 | 0 |
| 0 | 1 | 0 | 1 | 3 | 2 | 1 | 0 |
| 1 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 3 | 2 | 2 | 5 | 3 | 1 | 0 |
| 3 | 0 | 3 | 4 | 6 | 2 | 3 | 0 |
| 0 | 1 | 1 | 9 | 8 | 1 | 4 | 1 |
| 2 | 4 | 3 | 5 | 6 | 1 | 1 | 0 |
| 0 | 5 | 4 | 6 | 1 | 1 | 3 | 0 |
| 0 | 1 | 0 | 1 | 3 | 2 | 1 | 0 |
| 1 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 3 | 2 | 2 | 5 | 3 | 1 | 0 | ··· | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |

# 3. Methods/Results

# Method 1: Principal Component Analysis

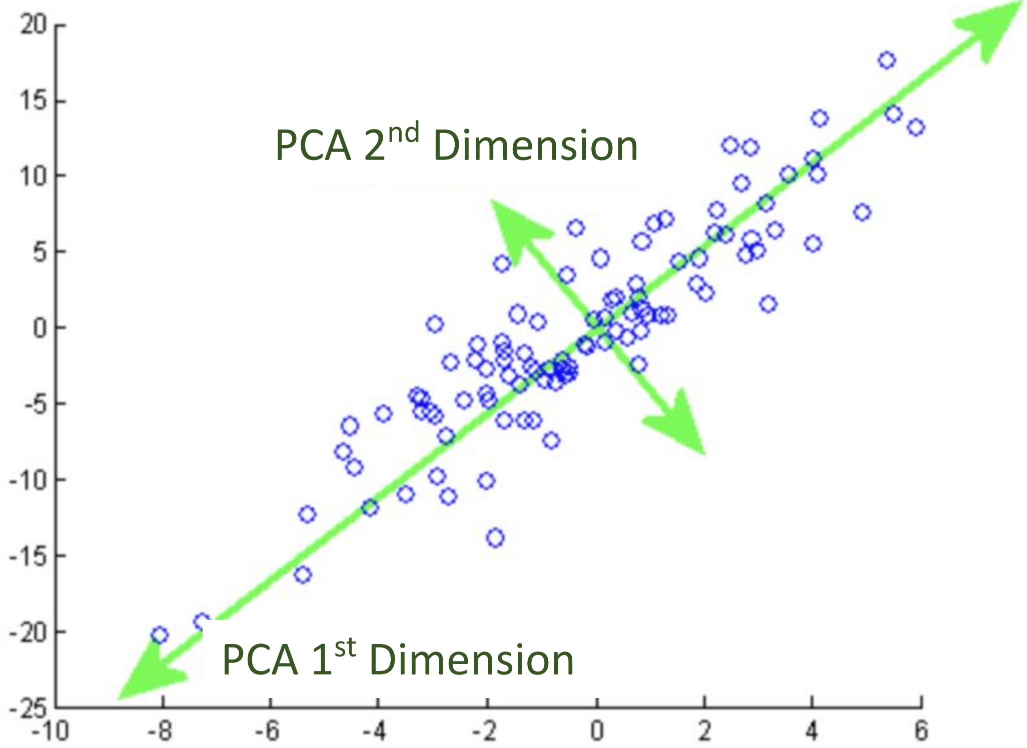
手法1：主成分分析

# PCA Basics

- The ‘Principal Components’ of a dataset are the set of orthogonal unit vectors where average squared distance from data points vector line is minimised.
    - Alternatively, the variance along PCs are maximised.

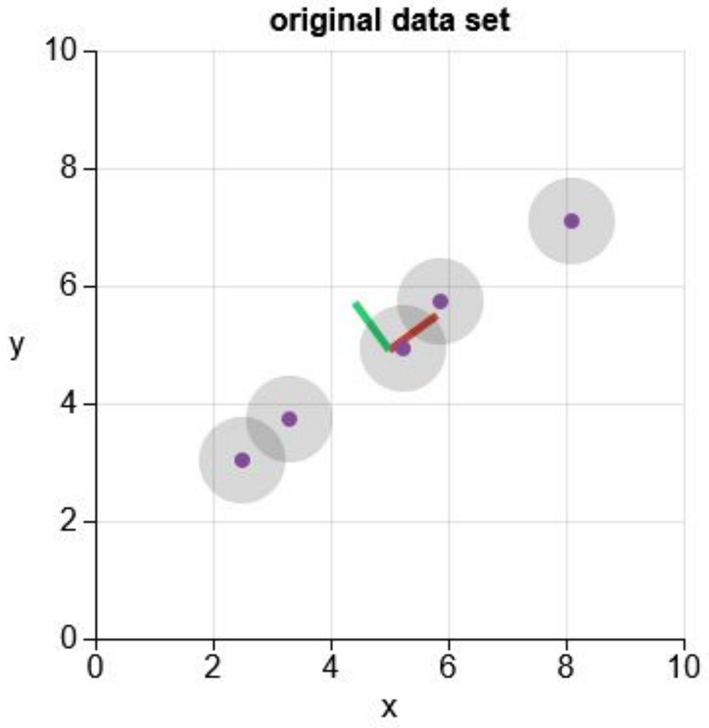
- We can select a subset of principal components, and use it as a new ‘eigenspace’ coordinate system.

- Our data points can now be represented in this lower dimensional space.



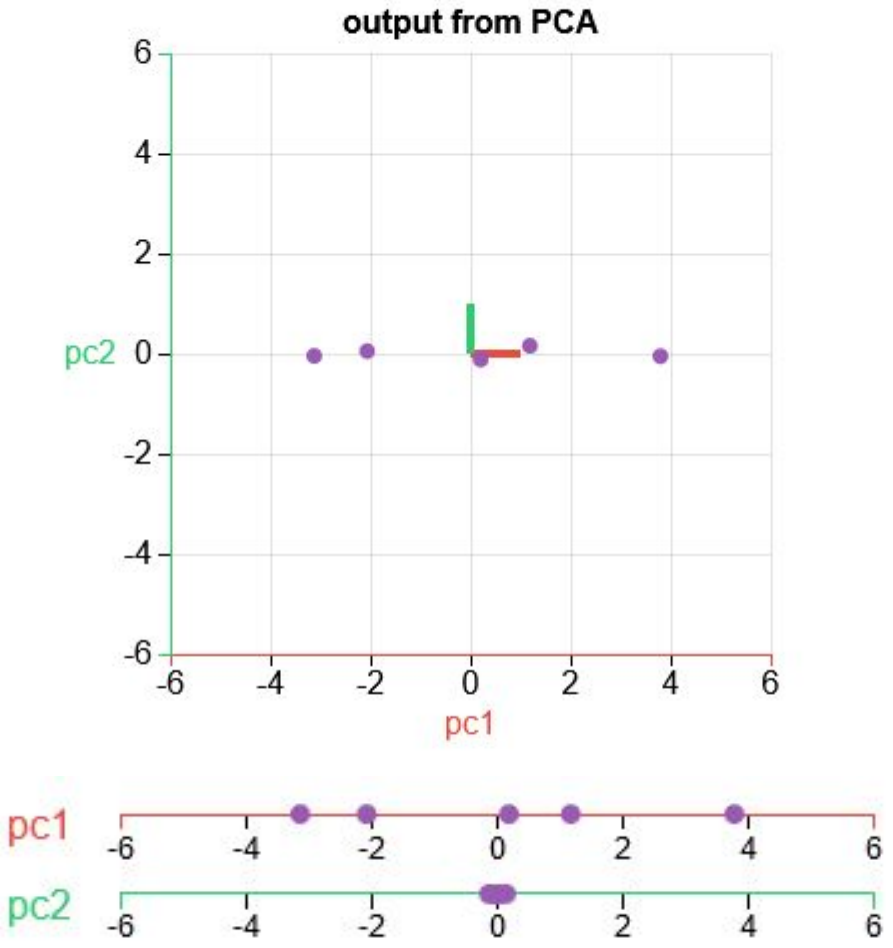
https://programmathically.com/principal-components-analysis-explained-for-dummies/

# PCA Intuition



This is a dataset of points in (x, y) space.
Notice that they all lie in close to a straight line.

Convert to PC coordinates



If we redefine our coordinate systems to be (PC1, PC2) space, we notice that since there is little variation along PC2, we could define each point just by their PC1 coordinate

# Finding the Principal Components

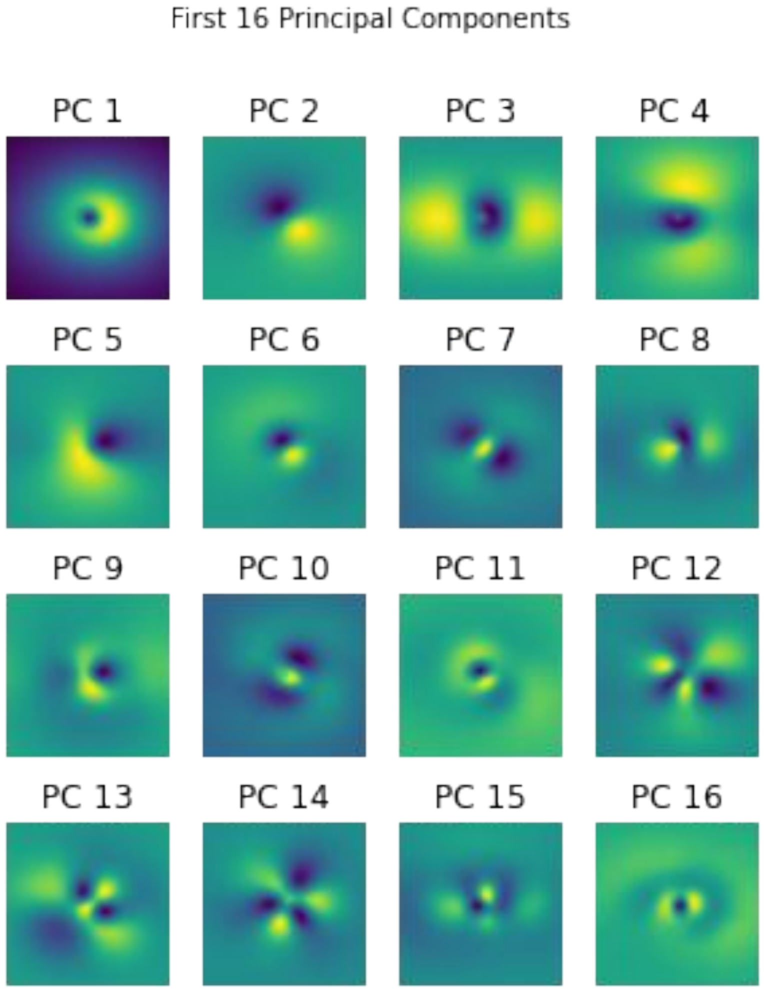
- Covariance matrix of a dataset is $Cov_{\boldsymbol{M}}(x, y) = \sum_{1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$
  - Covariance measures the deviation between the components of each data point.

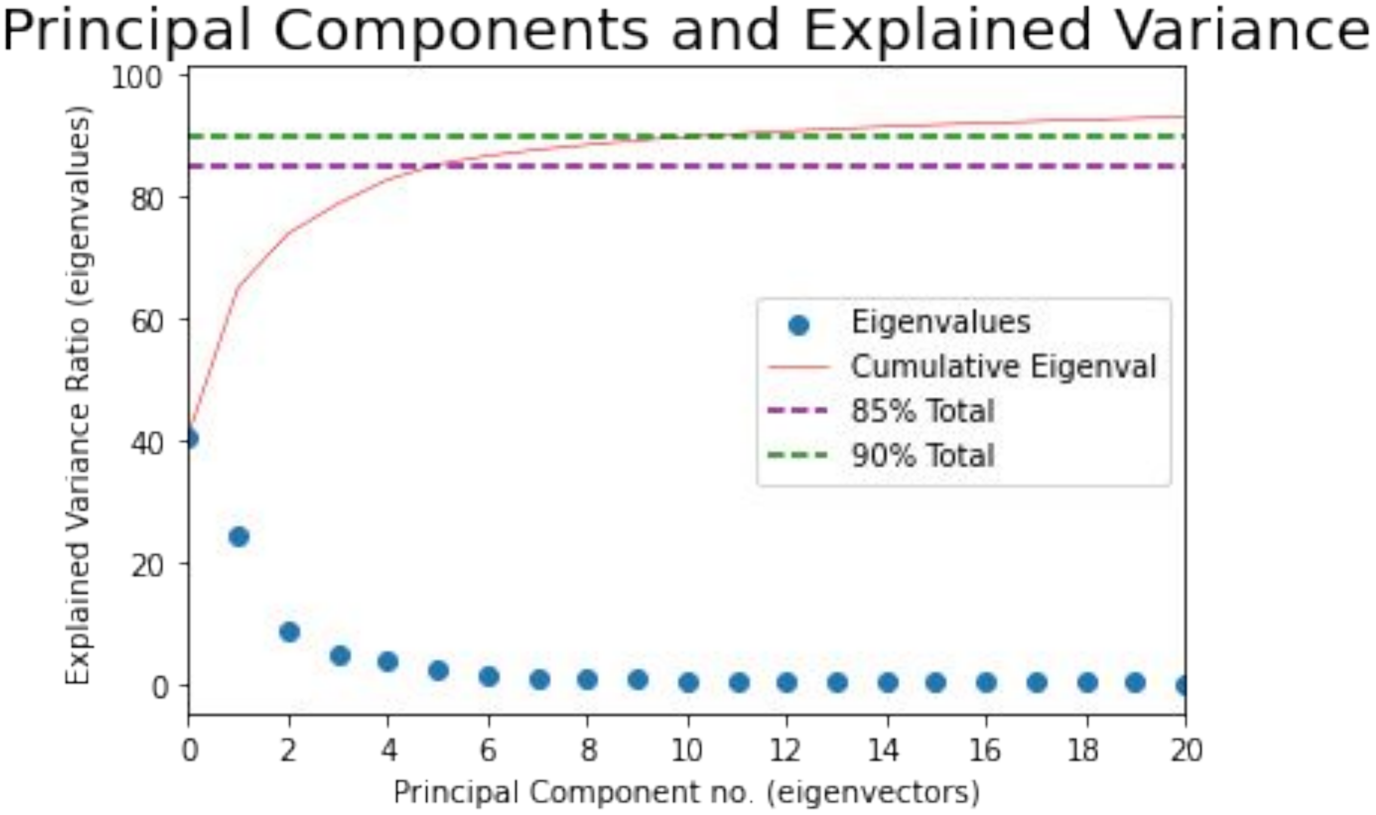
$$\begin{bmatrix} cov(x_1, x_1) & \dots & cov(x_1, x_n) \\ \dots & & \\ cov(x_n, x_1) & \dots & cov(x_n, x_n) \end{bmatrix}$$

- Eigenvectors of $Cov_{\boldsymbol{M}}$ give the Principal Components
- Corresponding Eigenvalues give the ‘Explained Variance’ of that PC.

# PCA Results

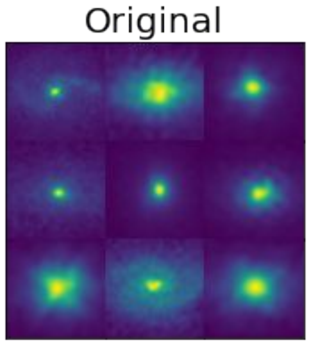
- 12 components sufficient to explain 90% variance
- PC 1 with 40% variance looks to correspond to nucleus of galaxies.
- PC 7 shows possible bar structure
- PC 11, PC 16, possibly showing spiral features

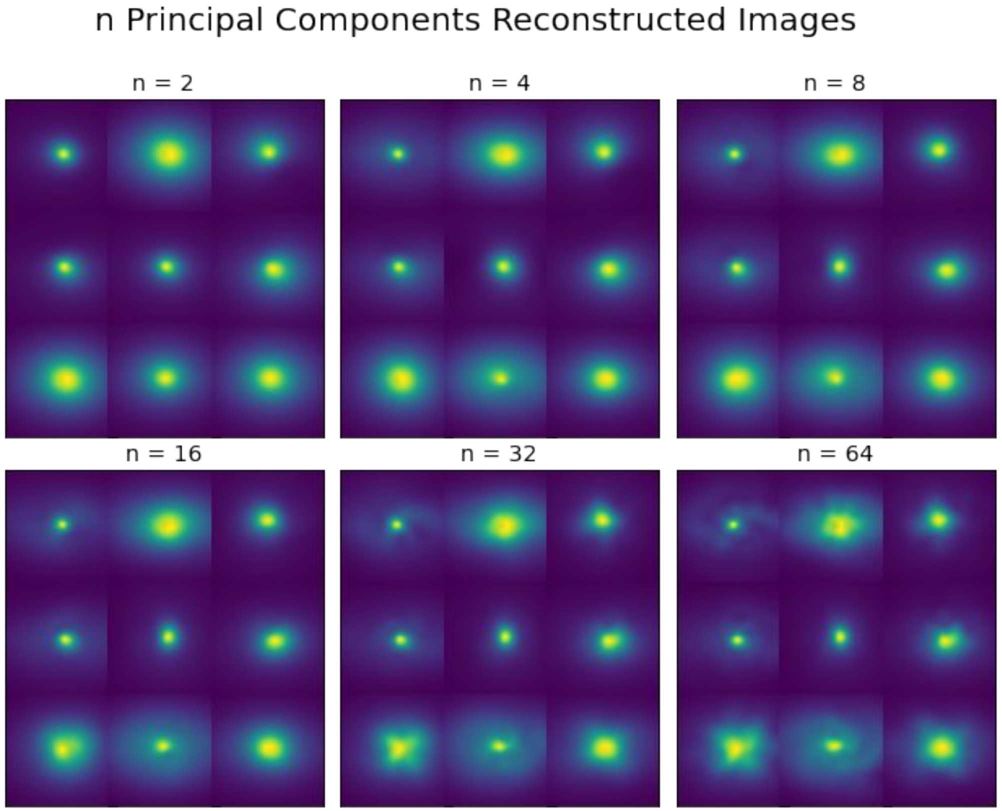


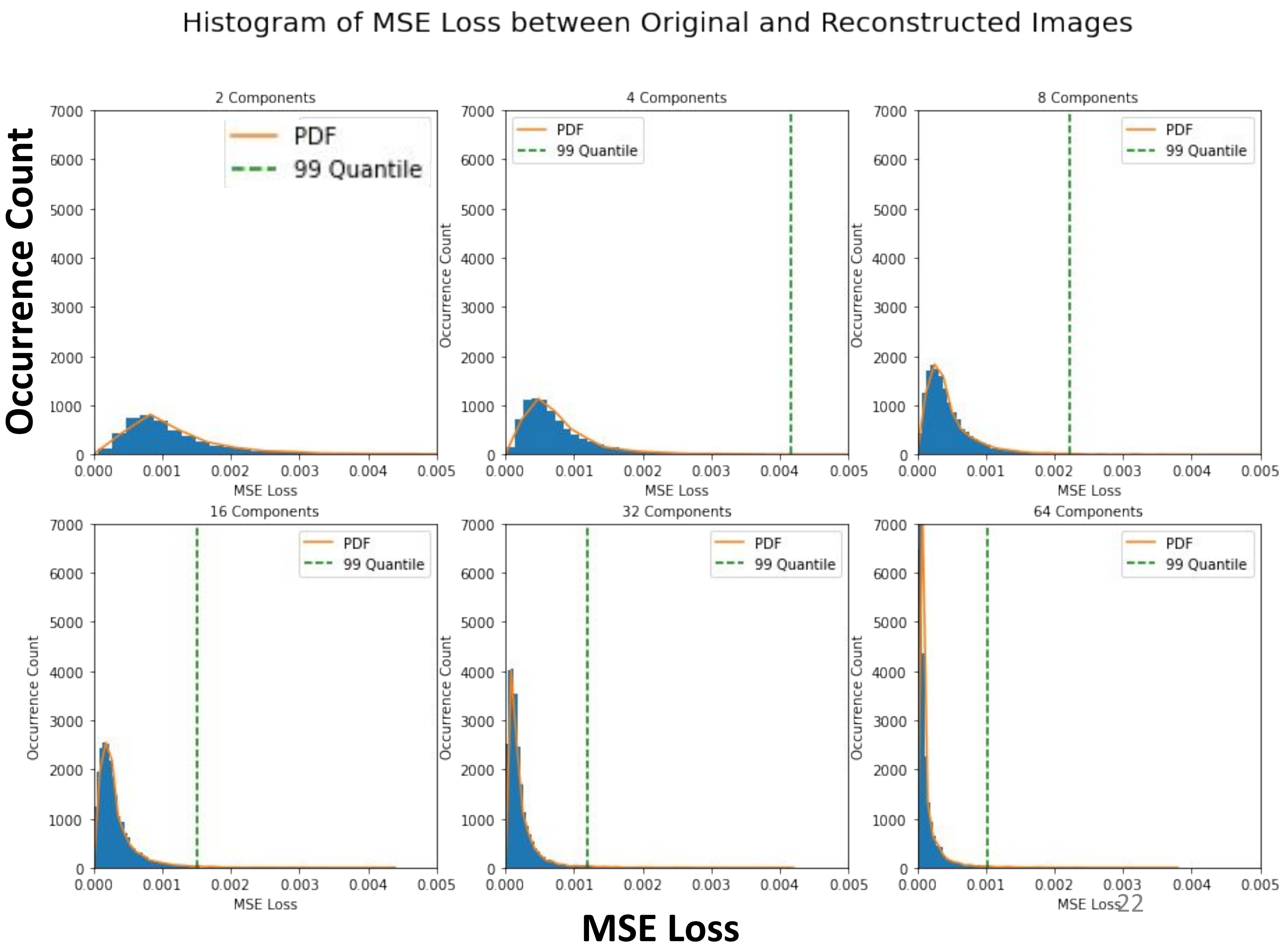


# *PCA – Results/Discussion*

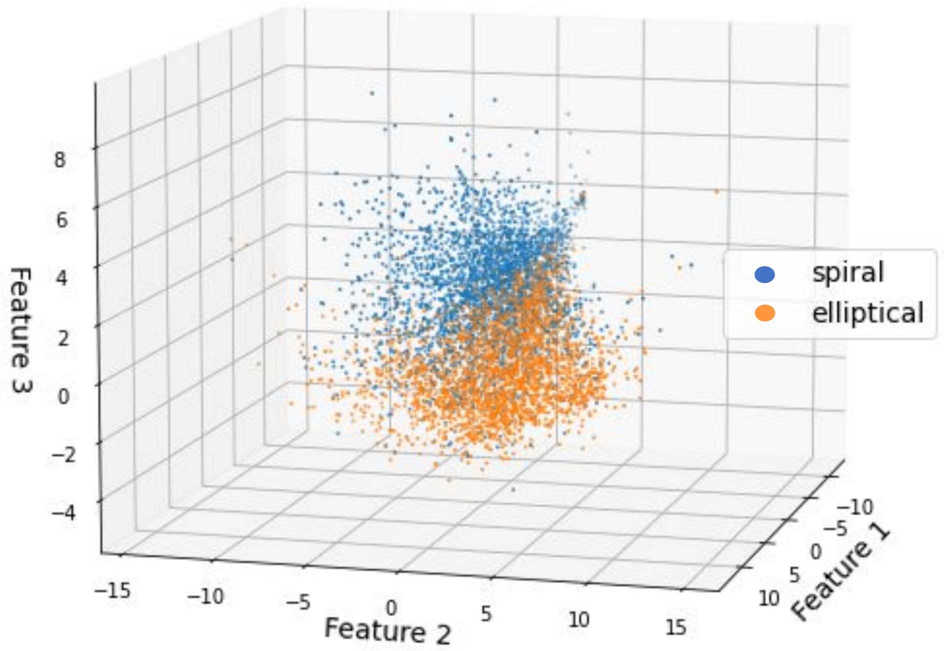
- At 2 components, 99% of reconstructed images fall within **0.0070 Mean Squared Error Loss**.
- At 32 components, 99% of reconstructed images fall within **0.0015 Mean Squared Error Loss**.
  - Spiral Features are visible in reconstructed images

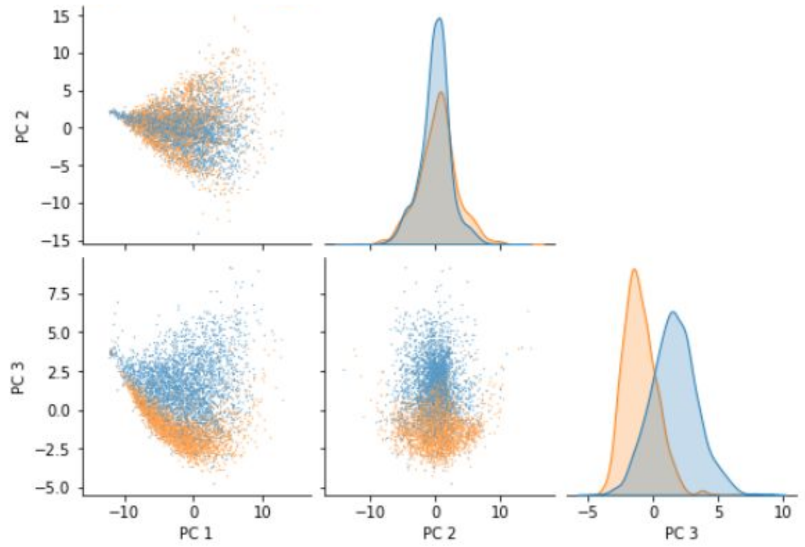






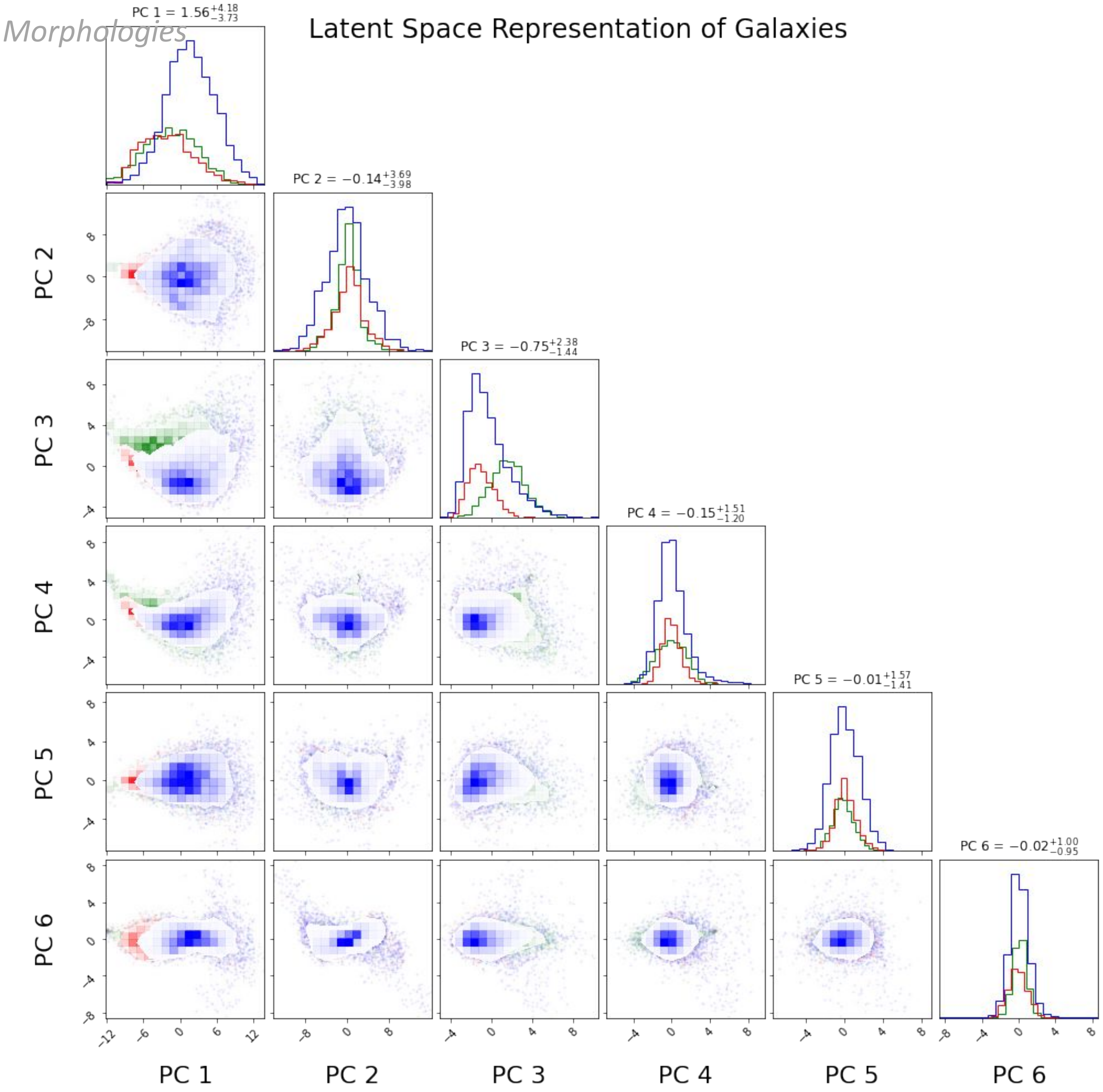
# Latent Space Representation of Galaxies





- In the latent space, we see that elliptical and spiral galaxies are found predominantly in clusters.

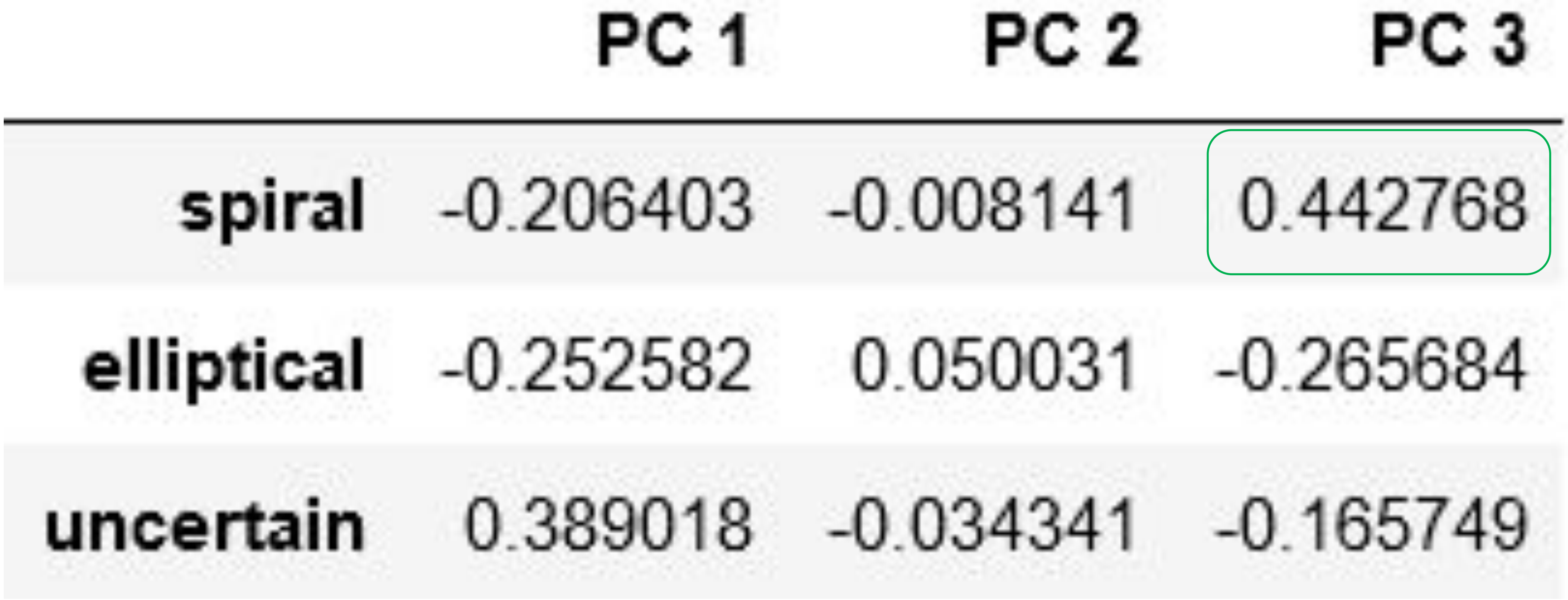
- Spiral galaxies tend to have bigger PC3 components than Ellipticals



- The ‘unclear’ galaxies tend to have bigger PC1 components than Spirals and Ellipticals.

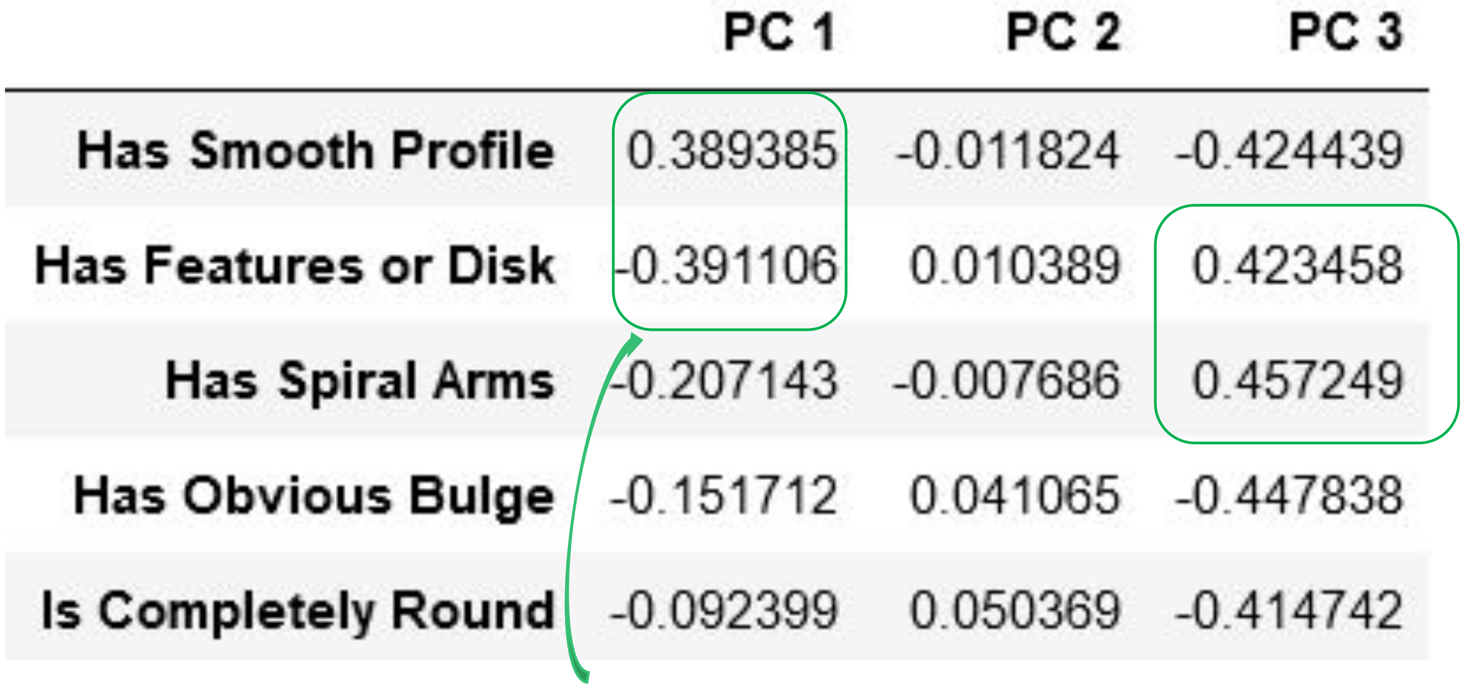
# Correlation with Classification

- Moderate correlations between PC3 and Spiral classification



| | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| **spiral** | -0.206403 | -0.008141 | 0.442768 |
| **elliptical** | -0.252582 | 0.050031 | -0.265684 |
| **uncertain** | 0.389018 | -0.034341 | -0.165749 |

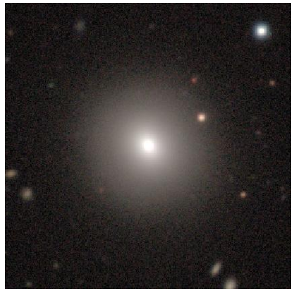
# Correlation with Galaxy Zoo 2 Weighted Votes on Presence of Various Features



| | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| Has Smooth Profile | 0.389385 | -0.011824 | -0.424439 |
| Has Features or Disk | -0.391106 | 0.010389 | 0.423458 |
| Has Spiral Arms | -0.207143 | -0.007686 | 0.457249 |
| Has Obvious Bulge | -0.151712 | 0.041065 | -0.447838 |
| Is Completely Round | -0.092399 | 0.050369 | -0.414742 |

- Higher PC 1 indicates higher likelihood of smooth profile + lower likelihood of features/discs.
    - More likely to be elliptical?

- More PC2 -> Higher likelihood of features/discs/spirals
    - Characteristics of Spirals?



-> Smooth Profile
-> Obvious Bulge

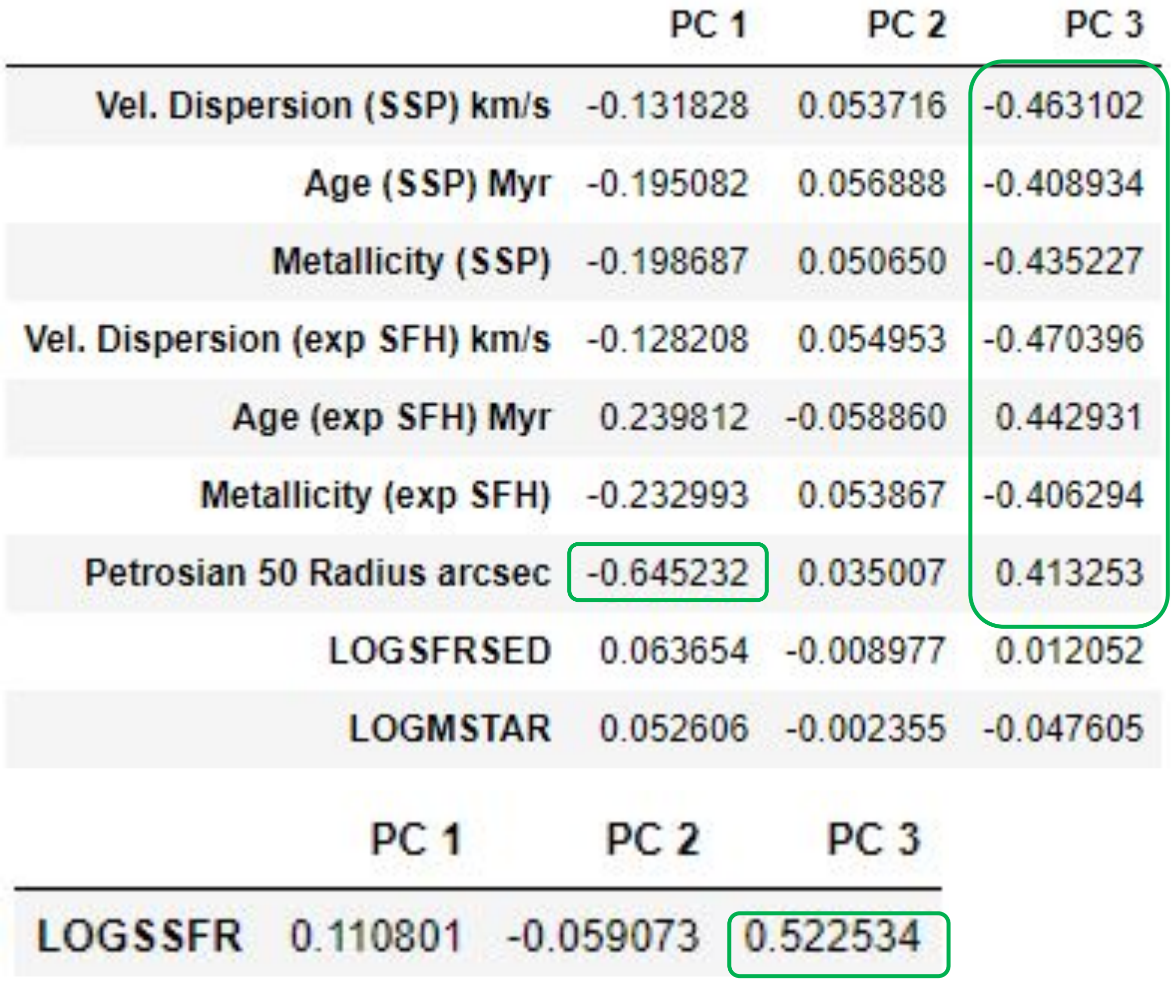
-> Spiral Arms

-> Has Features/Disc

# Correlations with Physical Properties

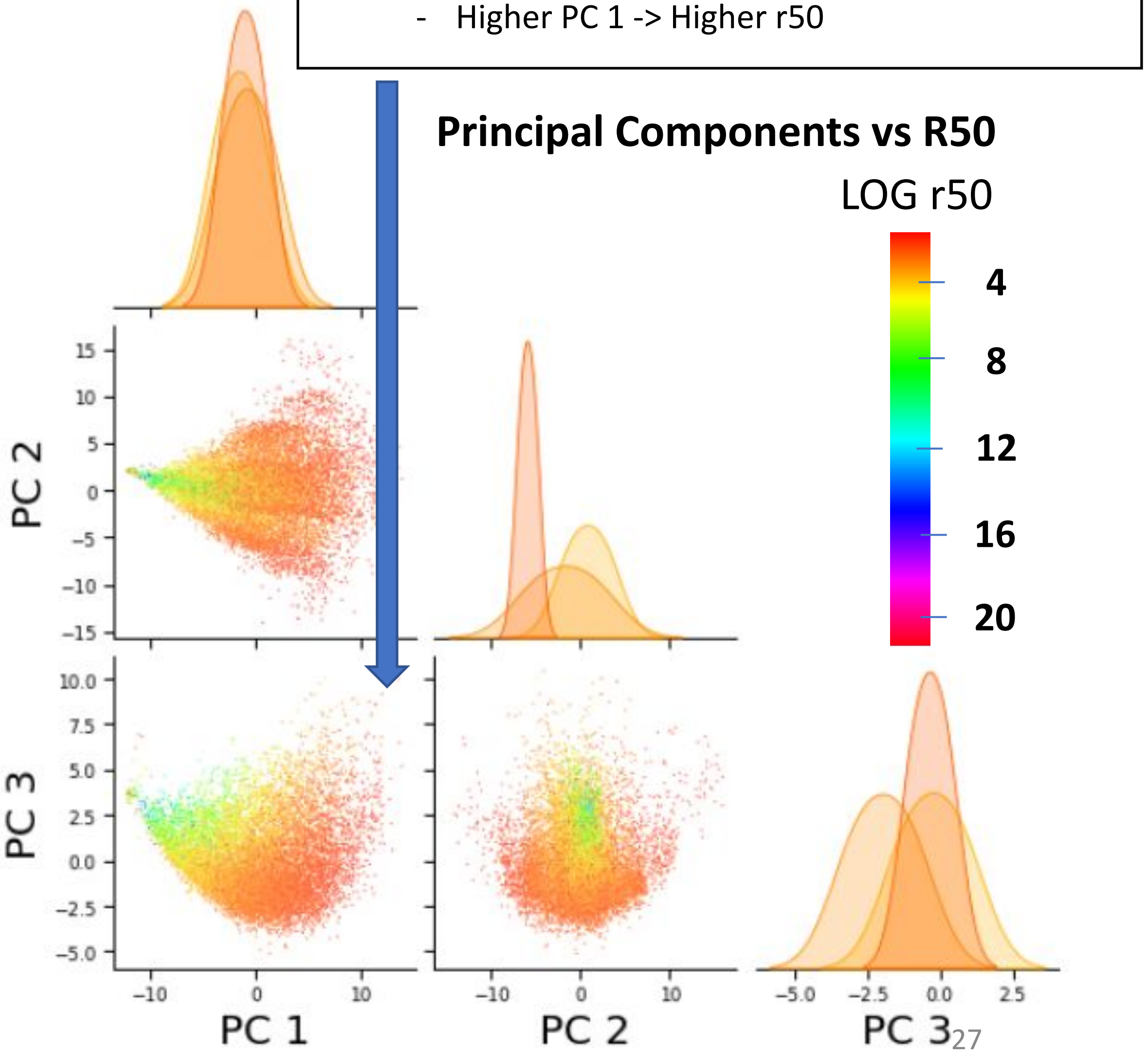
- We see clear variation of r50 along latent space
  - Higher PC 1 -> Higher r50

**Correlation Matrix of PC vs Physical Properties**



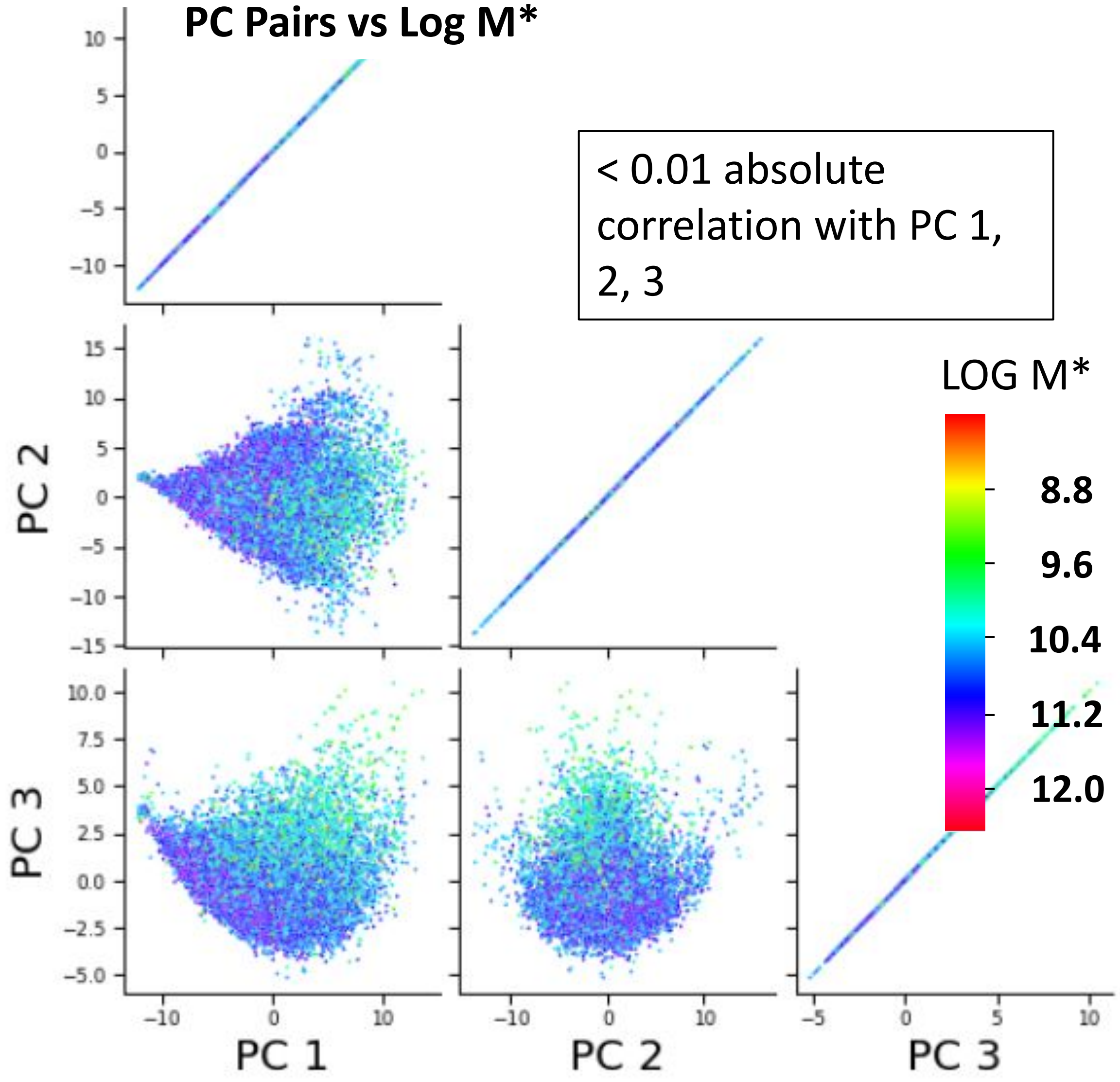
| | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| Vel. Dispersion (SSP) km/s | -0.131828 | 0.053716 | -0.463102 |
| Age (SSP) Myr | -0.195082 | 0.056888 | -0.408934 |
| Metallicity (SSP) | -0.198687 | 0.050650 | -0.435227 |
| Vel. Dispersion (exp SFH) km/s | -0.128208 | 0.054953 | -0.470396 |
| Age (exp SFH) Myr | 0.239812 | -0.058860 | 0.442931 |
| Metallicity (exp SFH) | -0.232993 | 0.053867 | -0.406294 |
| Petrosian 50 Radius arcsec | -0.645232 | 0.035007 | 0.413253 |
| LOGSFRSED | 0.063654 | -0.008977 | 0.012052 |
| LOGMSTAR | 0.052606 | -0.002355 | -0.047605 |

| | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| LOGSSFR | 0.110801 | -0.059073 | 0.522534 |



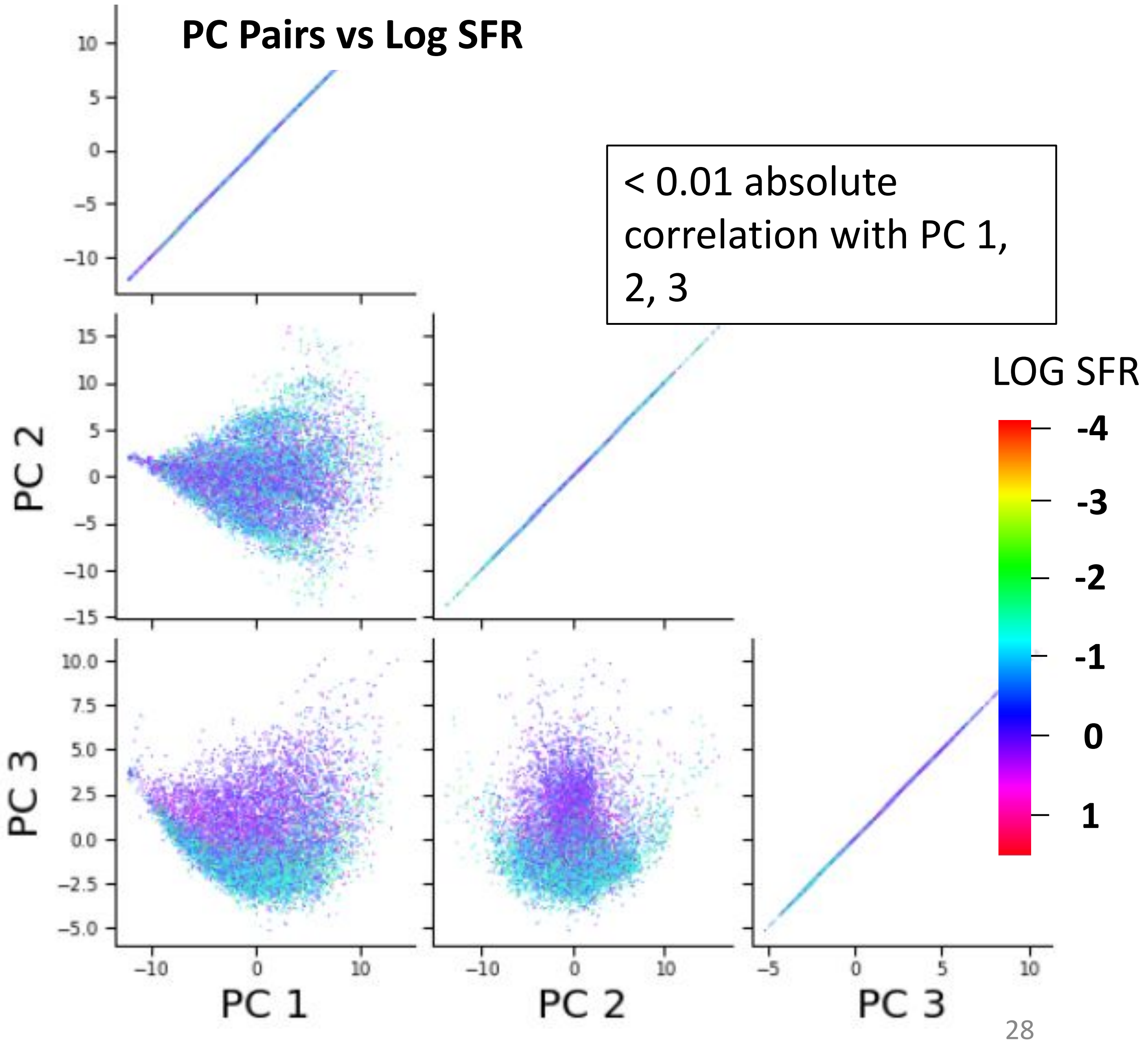
**Principal Components vs R50**

# PC vs Stellar Mass (M*)/Star Formation Rate (SFR)



**PC Pairs vs Log M***

< 0.01 absolute correlation with PC 1, 2, 3

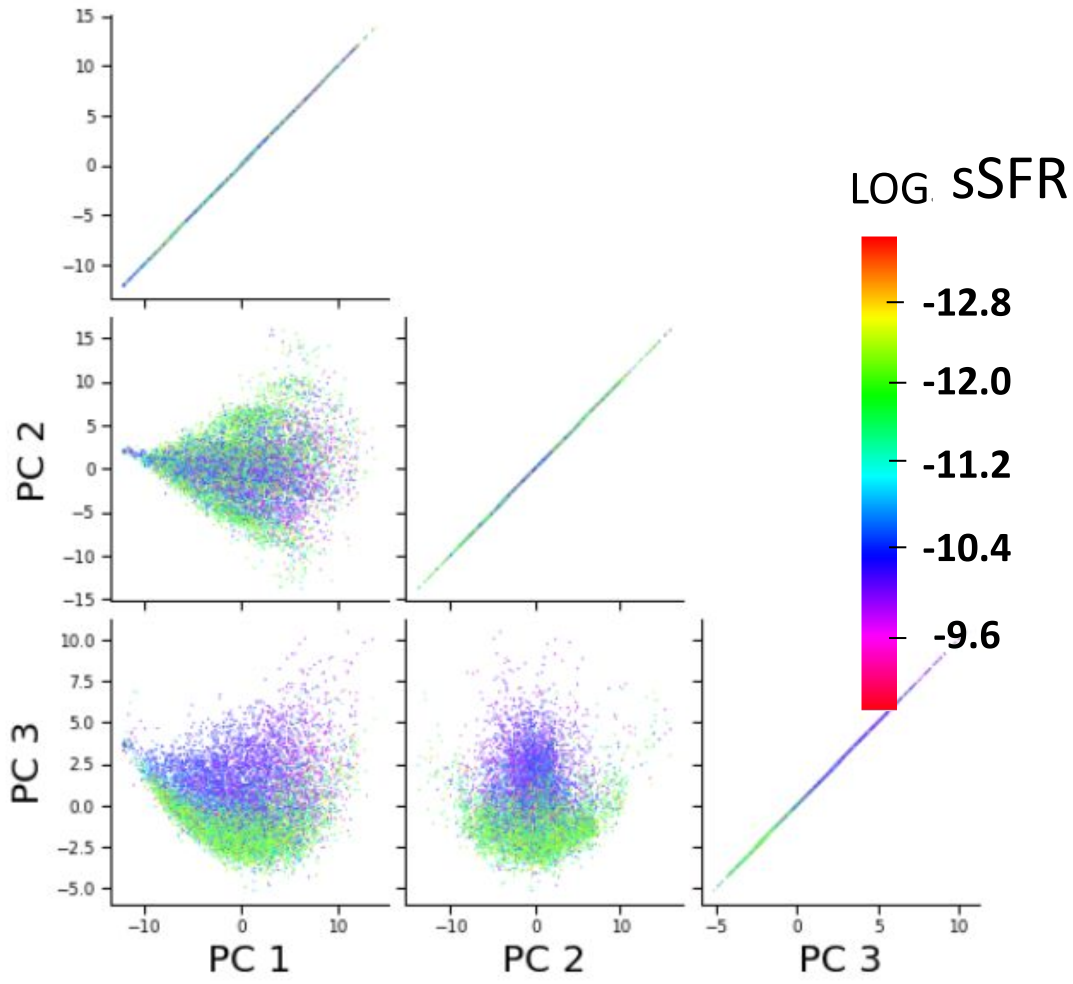


**PC Pairs vs Log SFR**

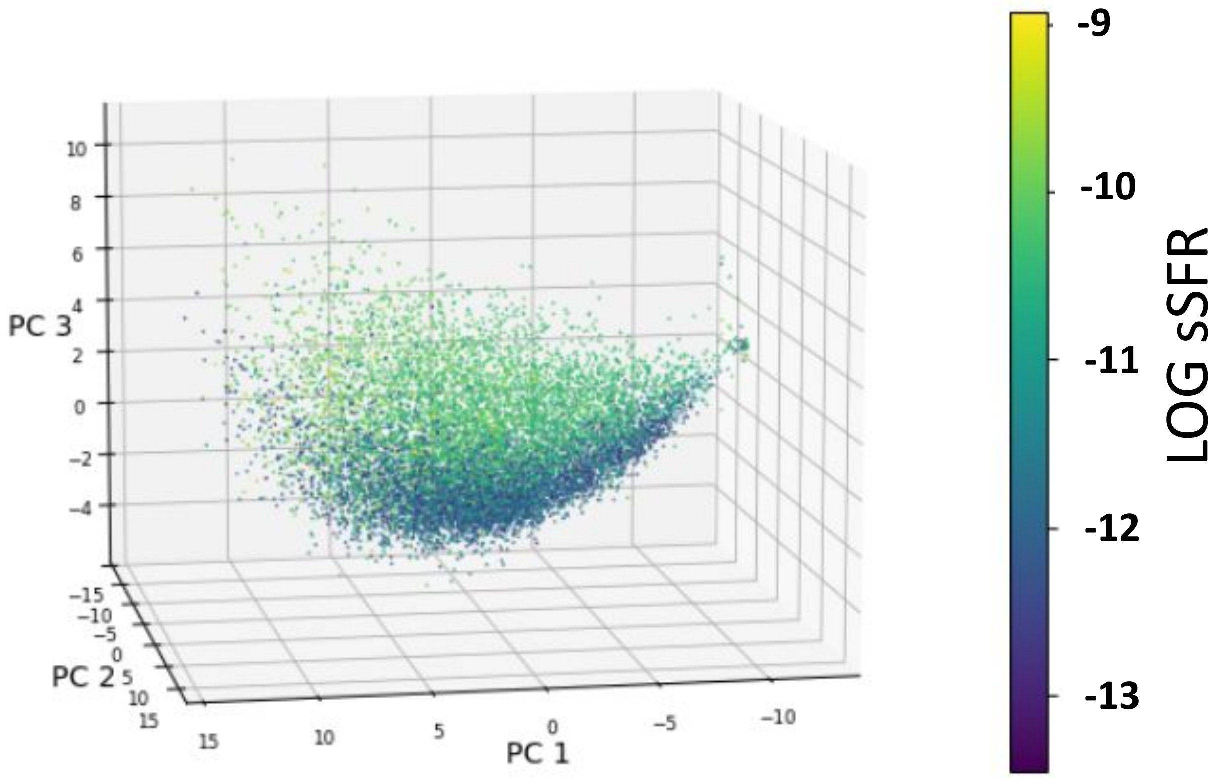
< 0.01 absolute correlation with PC 1, 2, 3

# Principal Components vs Specific SFR (sSFR)

PC Pairs vs LOG sSFR



3D Latent Representation LOG sSFR



- Moderately strong correlation between PC3 and SSFR
  - Pearson Correlation Ratio = 0.52

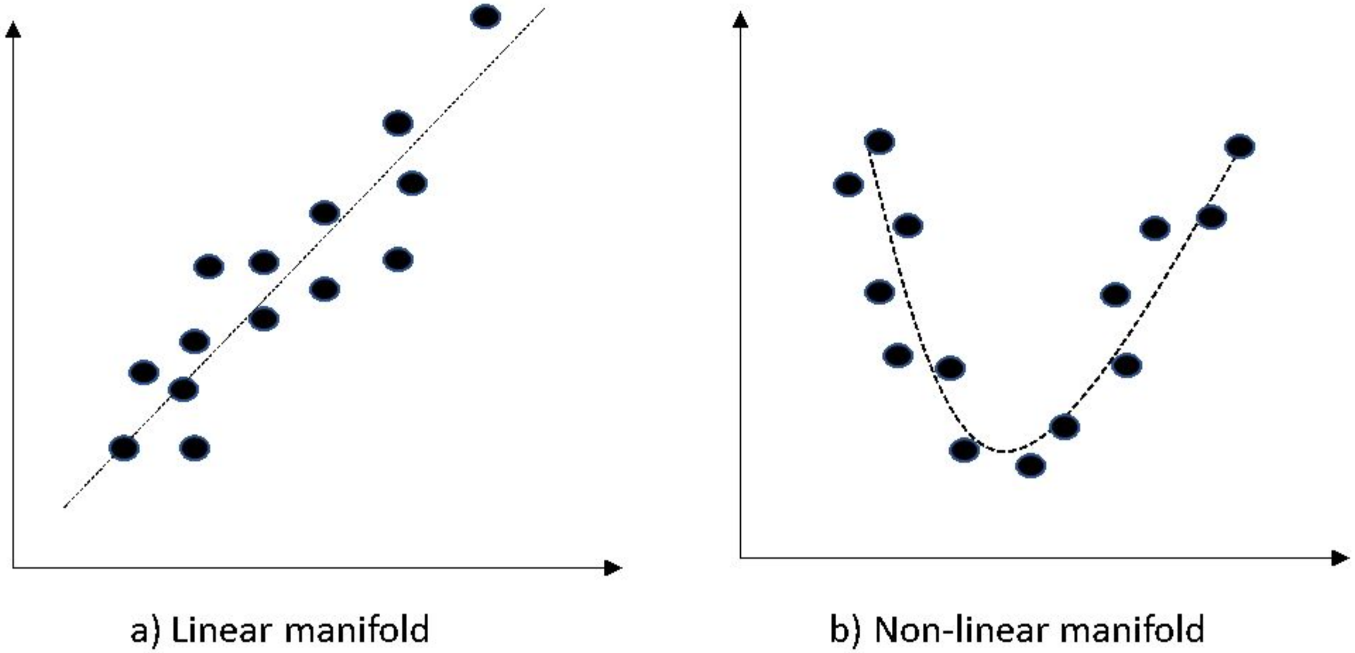
# Conclusion

- PCA applied to galactic images helped us find a lower dimensional representation of our data.
  - We can convert our data from a 10000 feature representation to just 64 latent feature representation
  - The latent features are linearly independent from each other -> no redundant information.
- The principal components can be directly interpreted as a galaxy's individual components.
- There are observable relationships between the latent space and a galaxy's physical properties.

# Method 2: Autoencoders

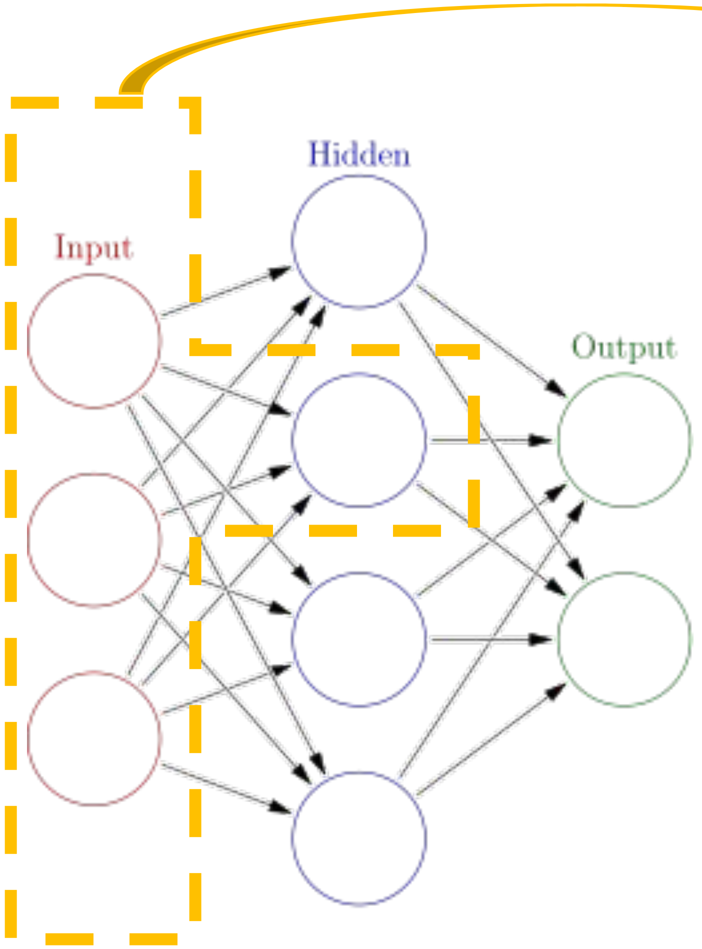
手法２：オートエンコーダ

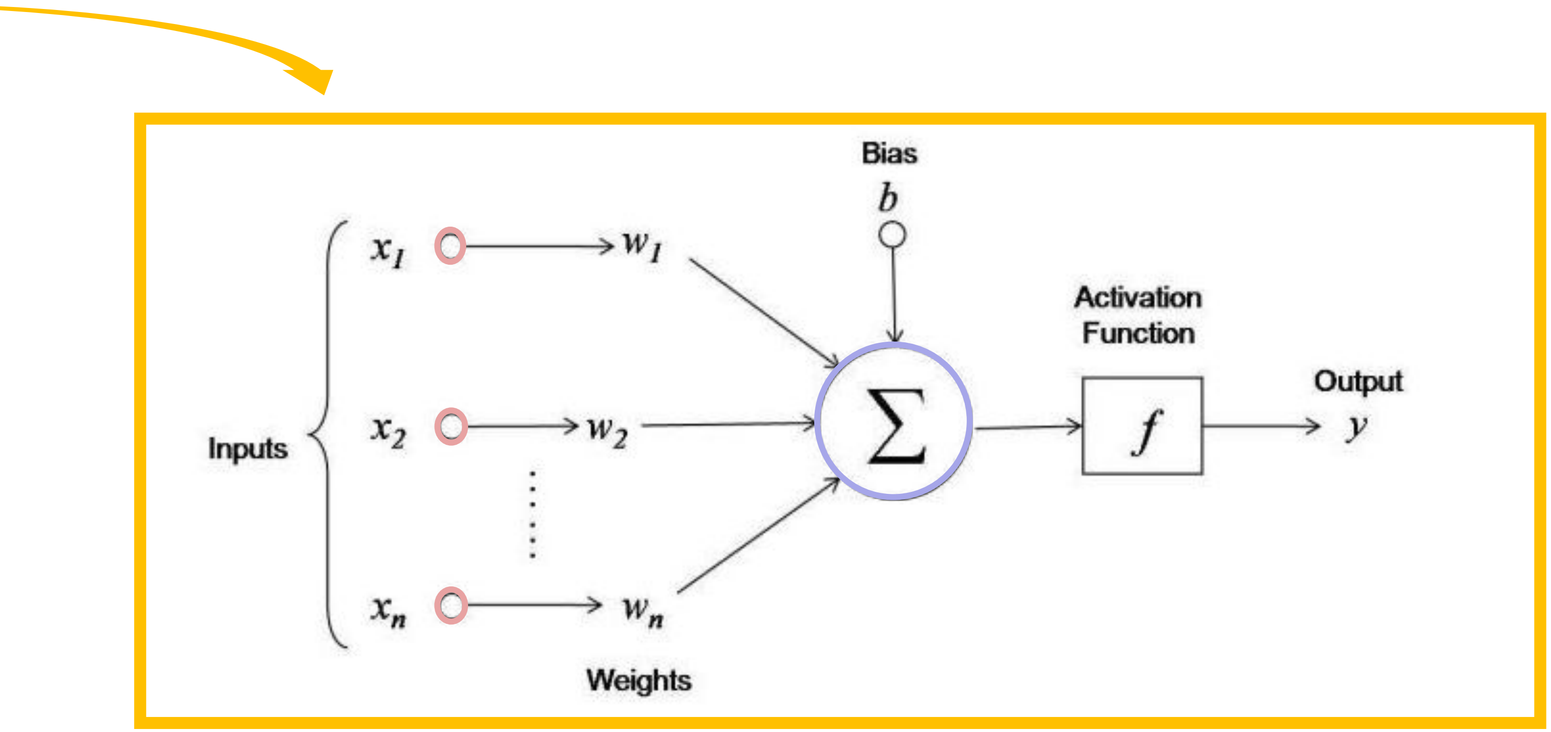
# Why use neural networks?



- PCA is linear method (sum of eigenvalues * eigenvectors)
- Data might not lie on a linear manifold/hyperplane
- Neural Networks can learn ‘non-linear’ manifolds

# Neural Networks

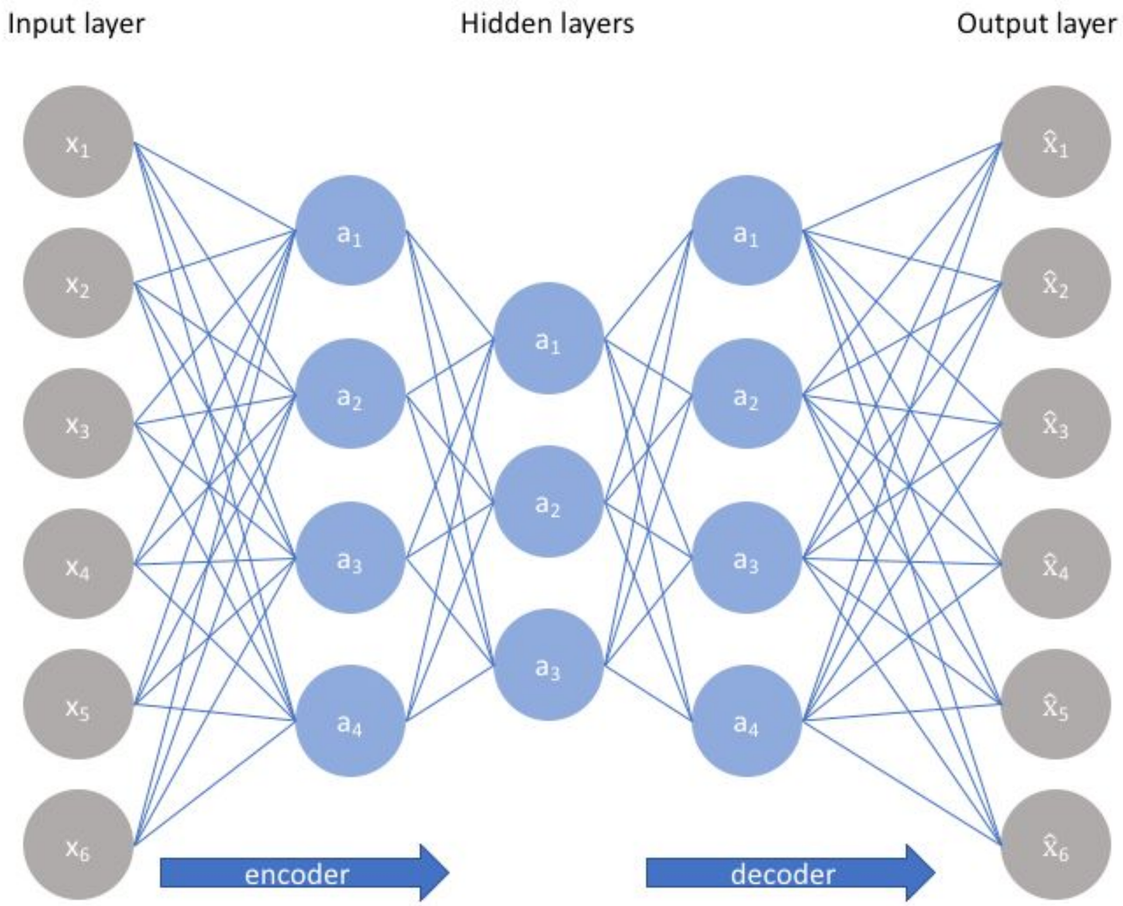


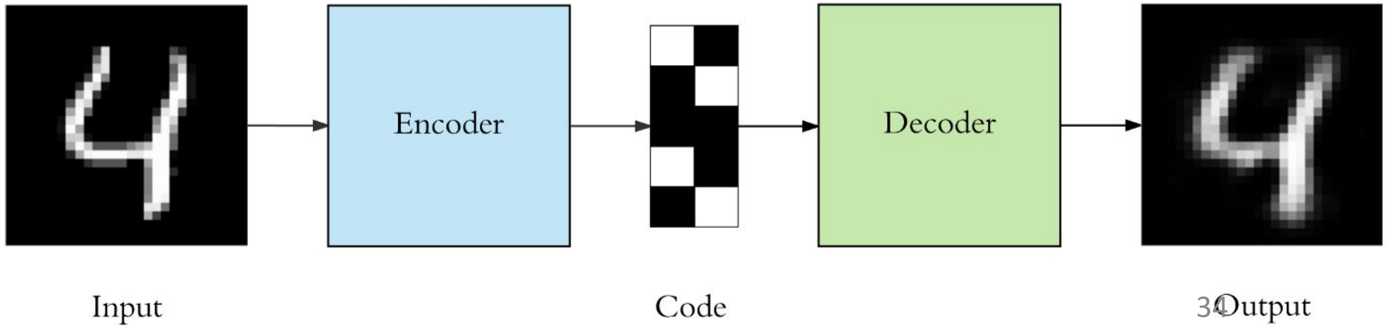


- Each **node** sums the (**inputs*** **weight**) from ALL the nodes of the previous layer.
- The sum is passed through an activation function.
- Activation function output is passed as inputs to next layer's nodes.
- "Training" the NN involves adjusting the **weight** until it gives "good" outputs from the final layer.

# Autoencoders

- Autoencoders are a type of feed forward neural network.
- Consists of a ‘encoder’ and ‘decoder’.
- Can be used to learn non-linear relationships.
  - Manifolds
- Loss function helps check training progress





# Architecture

**Data Input**

- Images were Log-scaled -> normalized to (0, 1)

**Encoder**

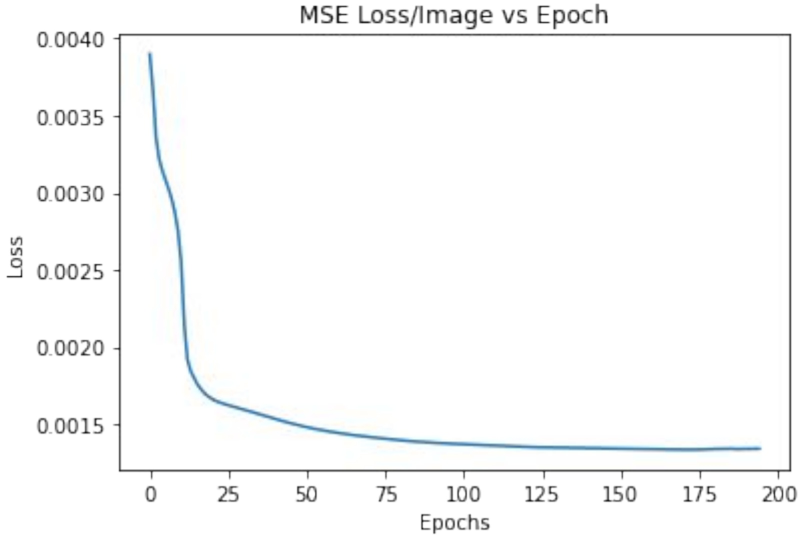
- 4 Hidden Layers (100000 -> 1000 -> 100 -> 64 -> 32)
- Bottleneck Layer (3 Dimensions)

**Decoder**

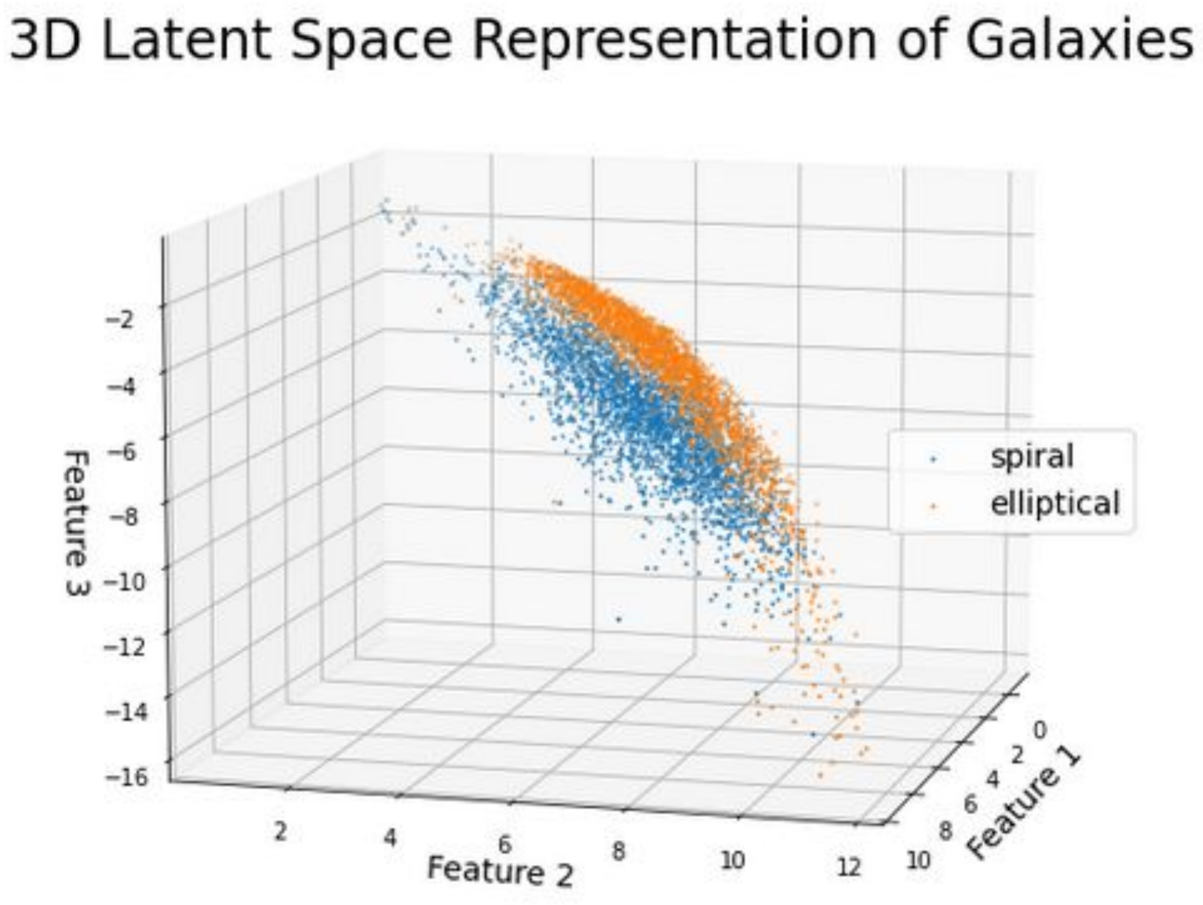
- Mirrored the Encoder Layers

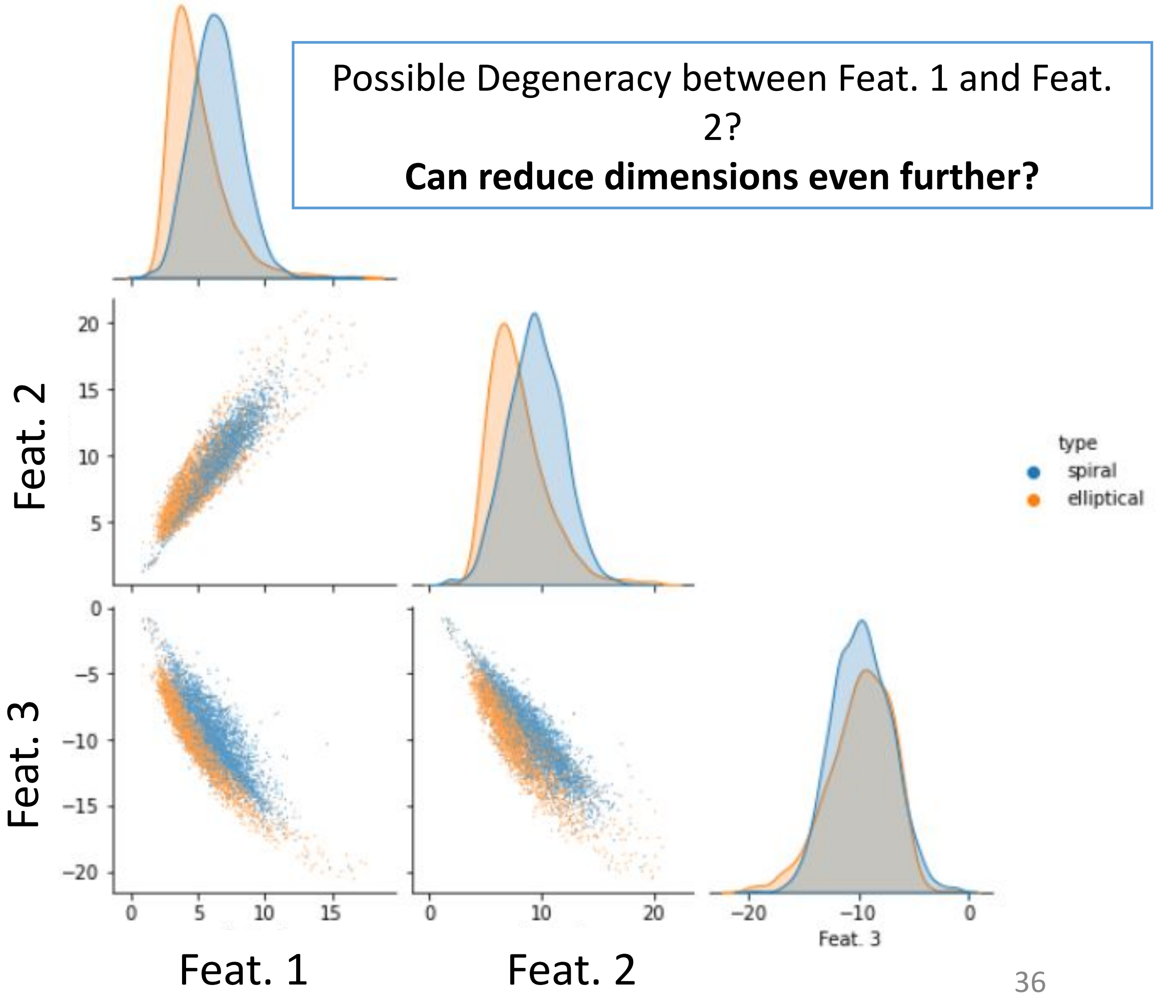
**Loss Function**

- MSE Loss



# Results: 3D Latent Space Representation

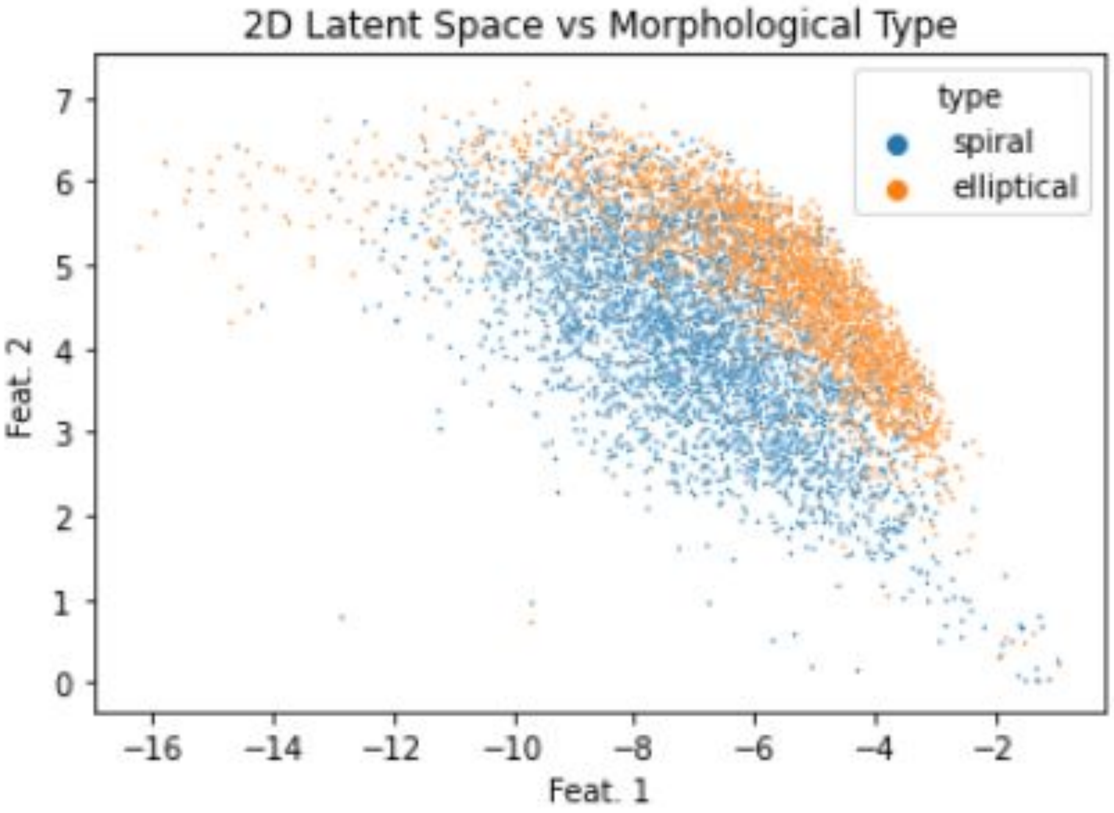


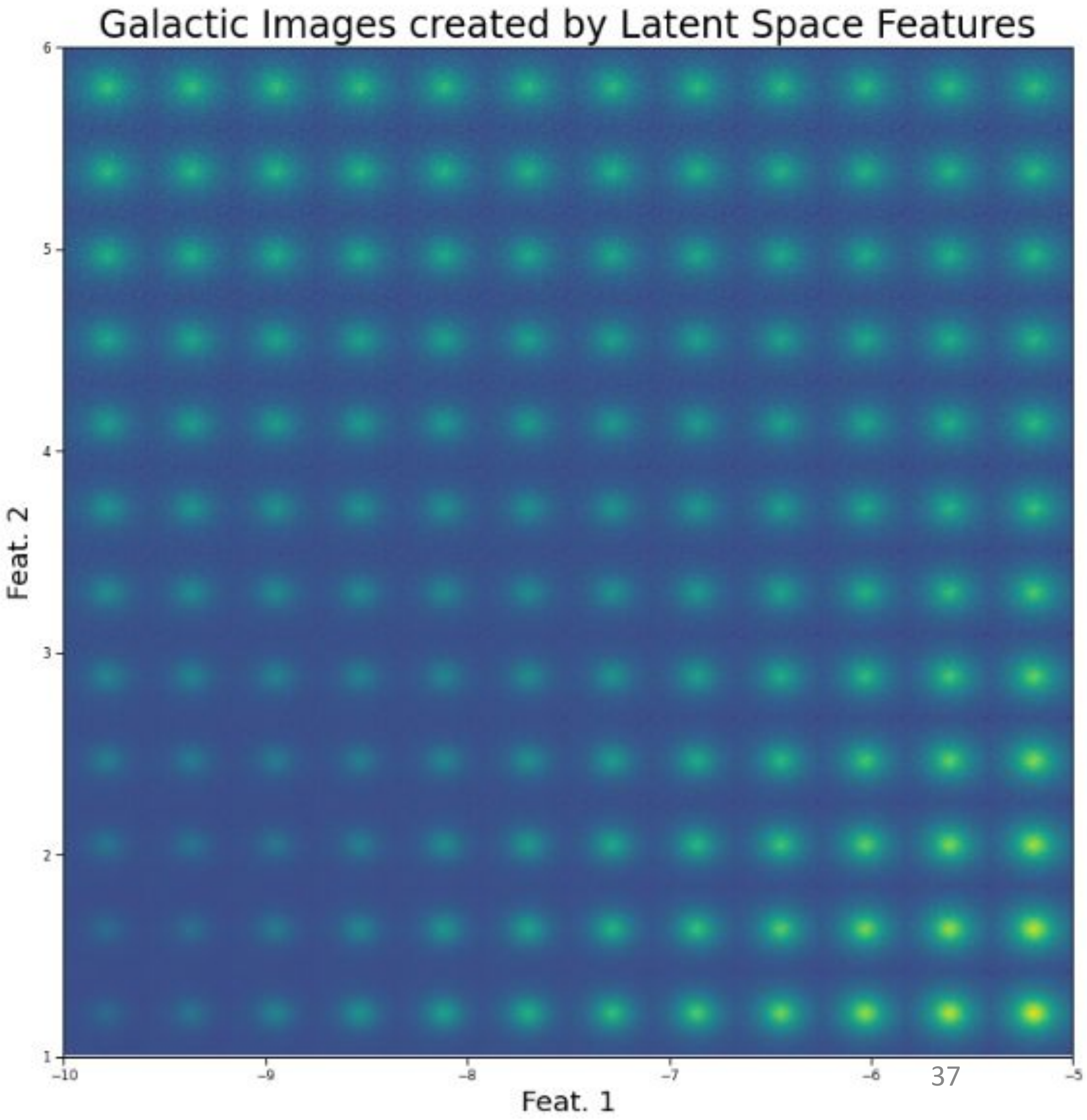


Possible Degeneracy between Feat. 1 and Feat. 2?

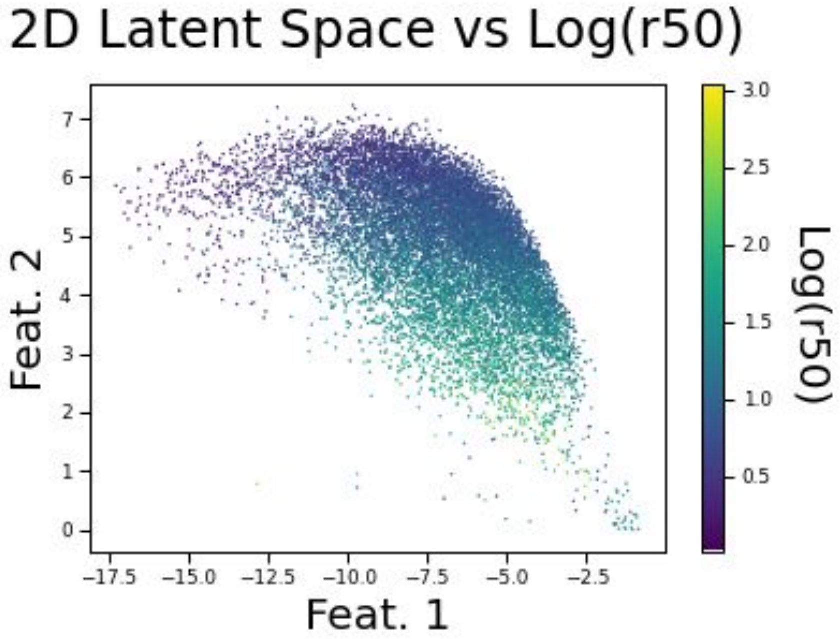
**Can reduce dimensions even further?**

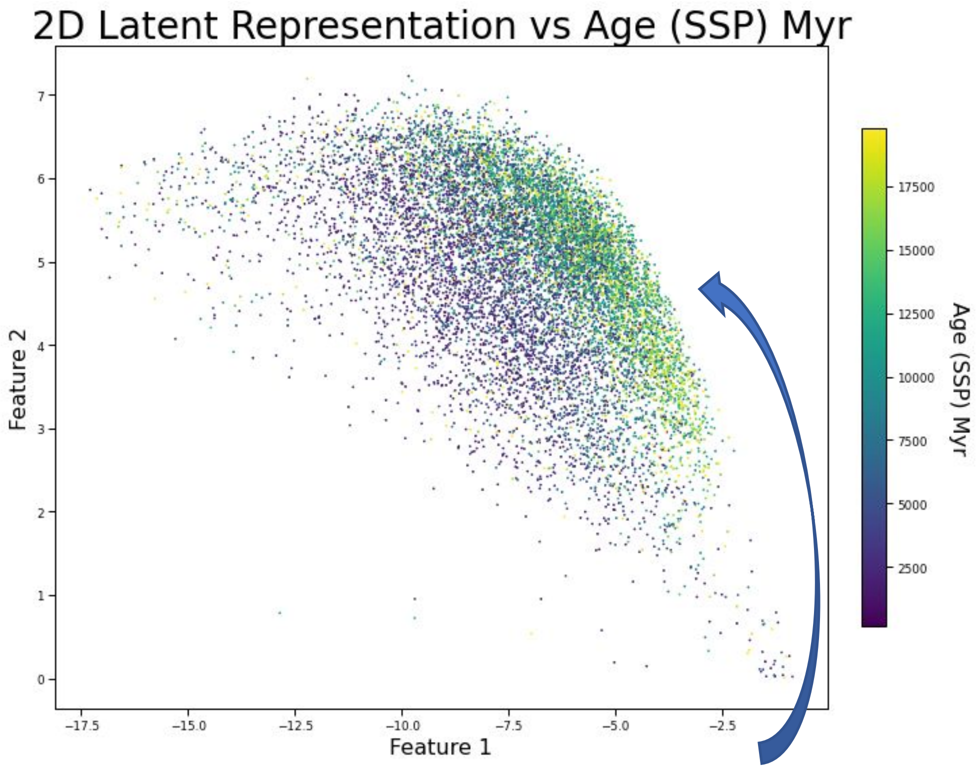
# Results: 2D Latent Space Representation

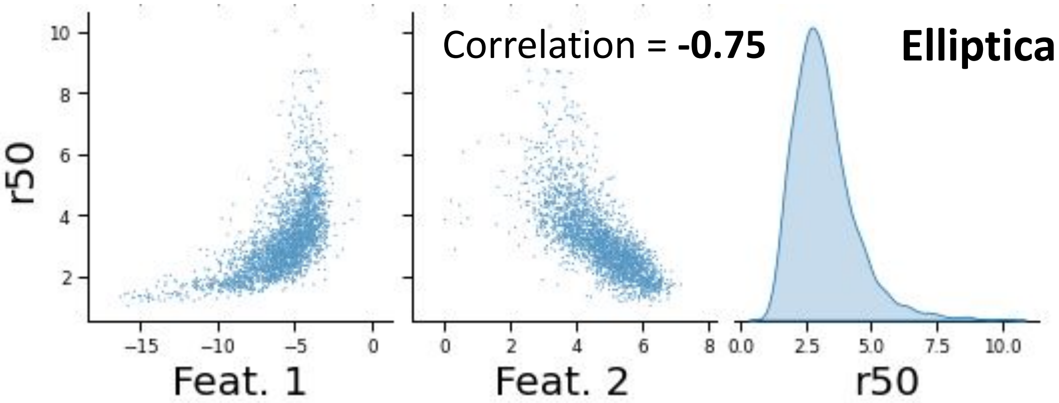

2D Latent Space vs Morphological Type


Galactic Images created by Latent Space Features

# 2D Latent Space vs Physical Properties







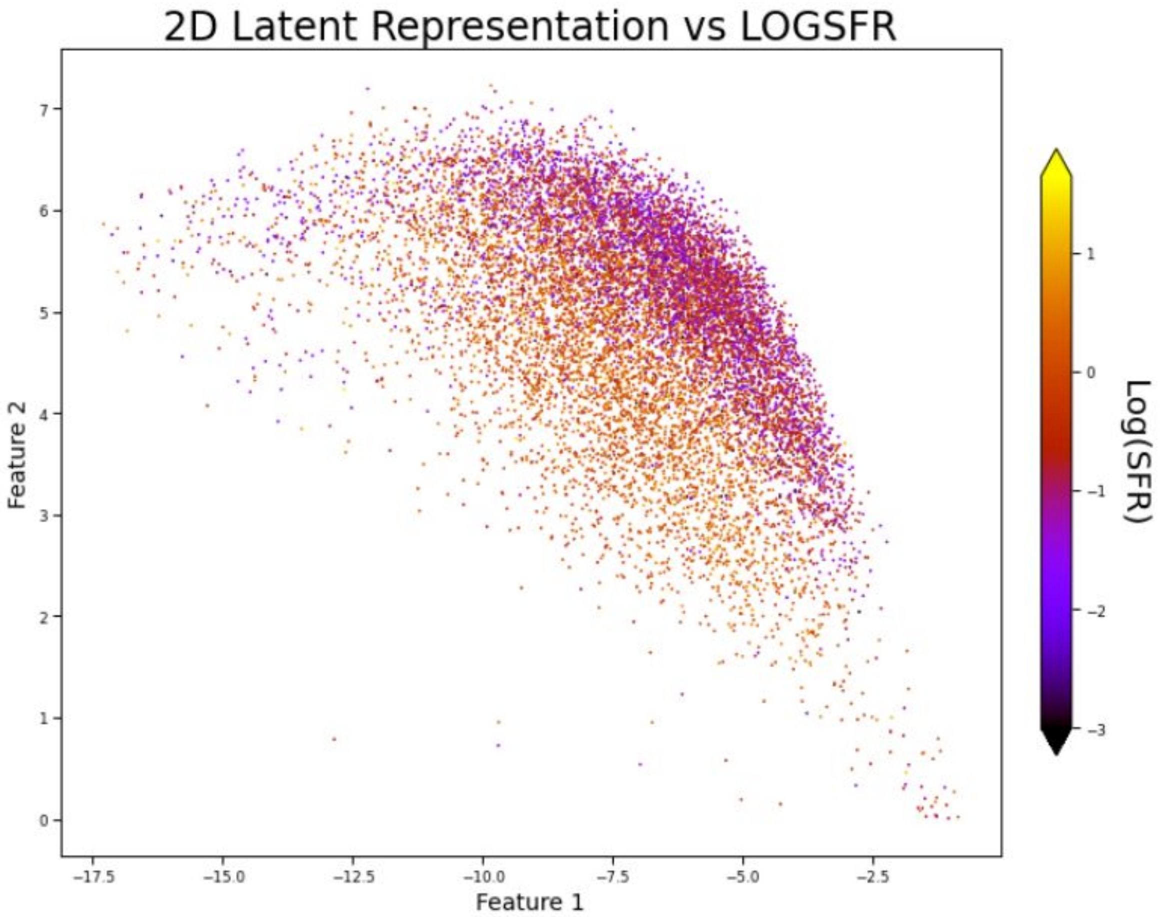
Correlation = **-0.75**

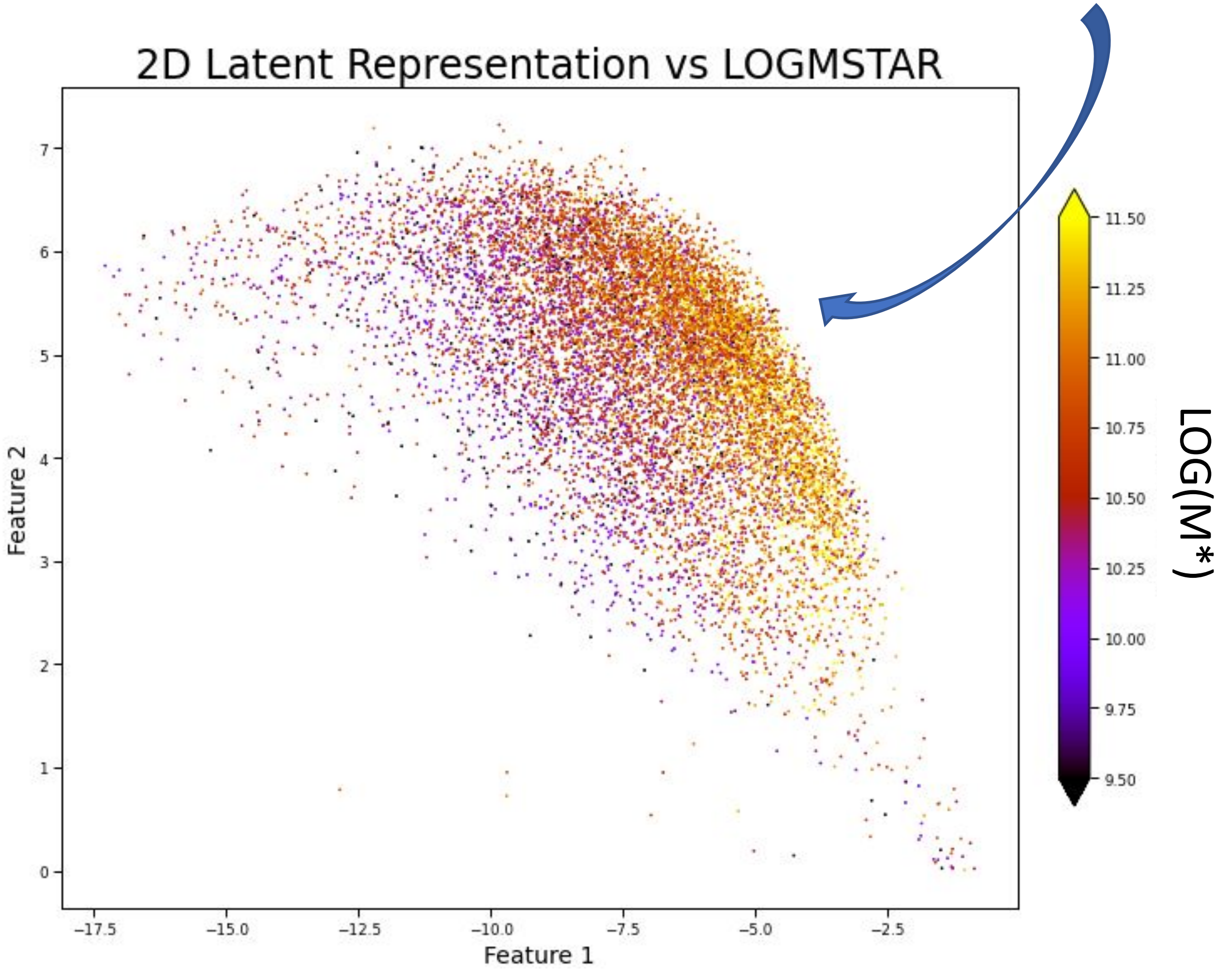
**Elliptical Galaxies Only**

- Lower Feat.1 -> Age tends to be lower.
  - Moderate-weak correlation = **<u>0.32</u>**
  - If Feat.1 represents bulge size in spirals, makes intuitive sense since center bulges tend to have **older** stars.

# 2D Latent Space vs Stellar Mass (M*)/ Star Formation Rate (SFR)

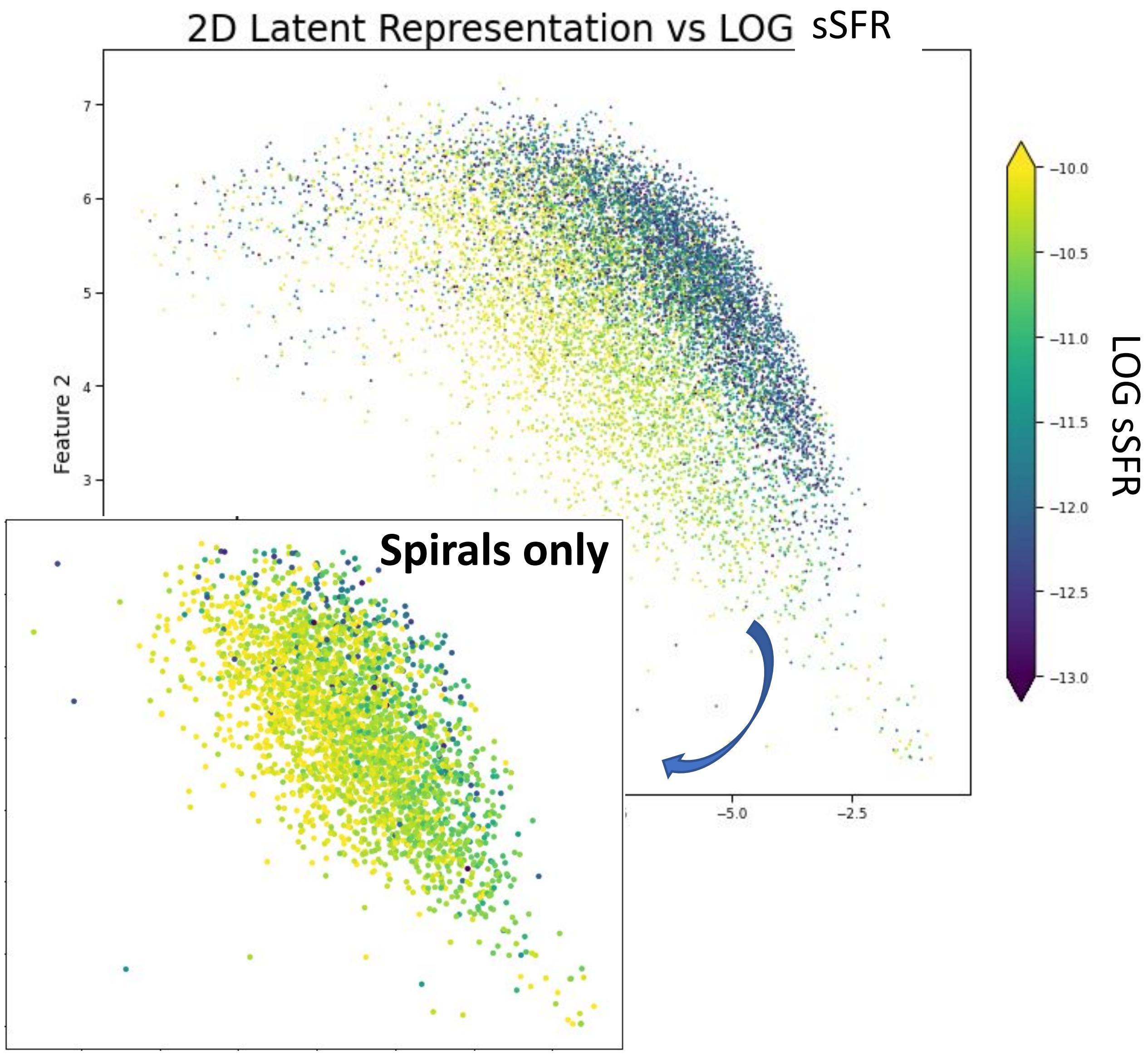
As Feat.1 increase and Feat.2 decrease, Log(MSTAR) increases





Clustering of SFR present, but no linear correlation between Latent Features and SFR

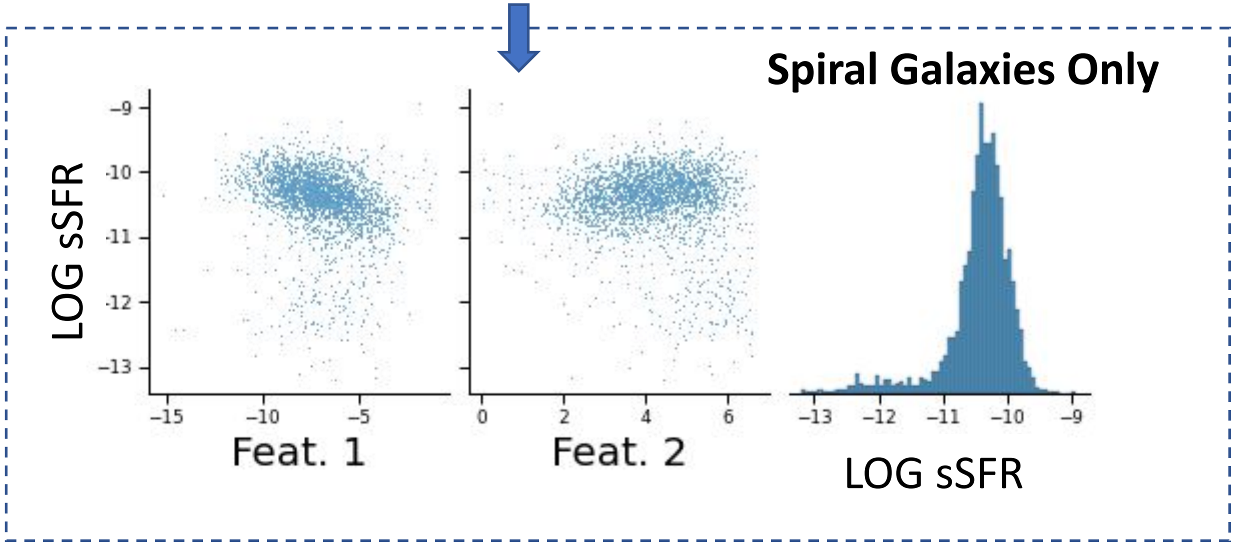
# 2D Latent Space vs sSFR



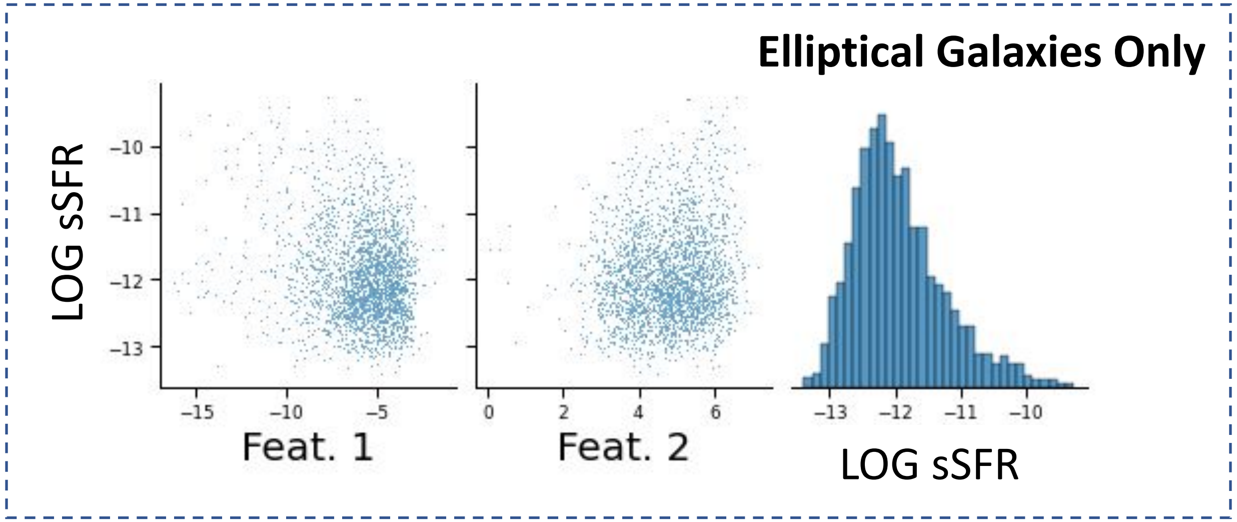
2D Latent Representation vs LOG sSFR

Smaller 'bulge'/Bigger 'disc' -> Higher sSFR?

- However, weak linear correlations
- Feat.1: -0.2, Feat.2: 0.2



**Spiral Galaxies Only**



**Elliptical Galaxies Only**

# Conclusion

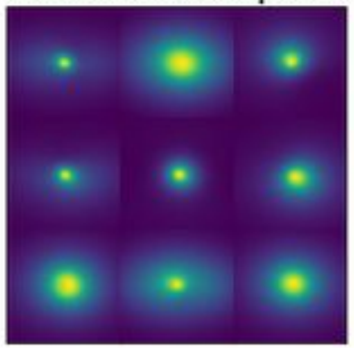
- Our simple autoencoder was able to find a low dimensional latent space whose features had correlations with certain physical properties.
    - 10000 -> 2 dimensions

- Autoencoders were able to find a non-linear manifold.
- However, the low dimensional features are not constrained to be linearly independent (unlike PCA).

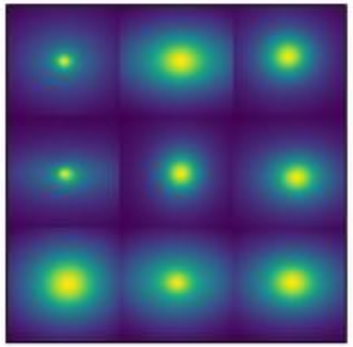
# Discussion/Conclusion

# Comparison of Reconstruction quality vs PCA

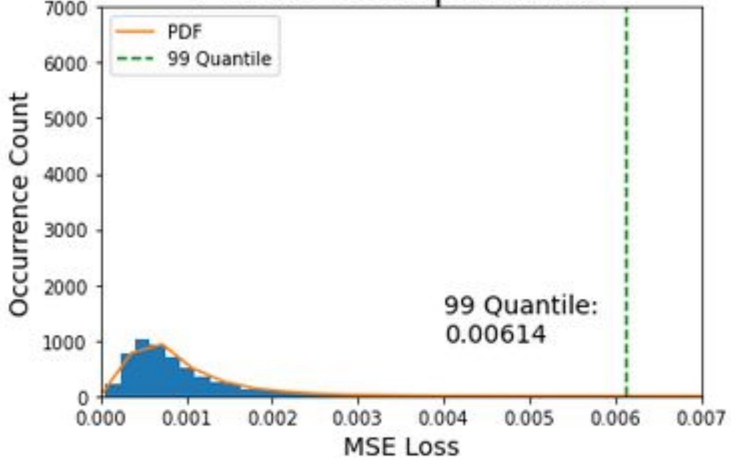
PCA with 3 Components



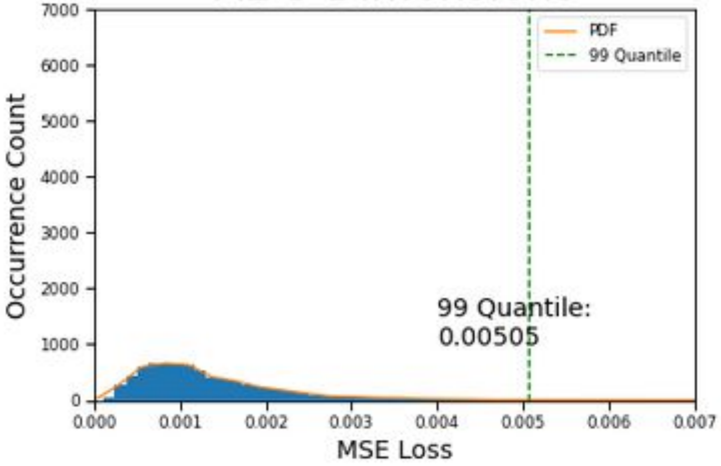
Autoencoder with 3 Dimensional Bottleneck



PCA 3 Components



AE 3 Dimensions



# Discussion

- The outcome of dimensionality reduction methods is highly dependent on preprocessing methods.

- Intricate features like spiral arms, bars, etc. were not learnt for both methods, hence the physical properties that are related to them showed low correlation with the latent space.
  - We can increase our latent space dimensions to better find features with correlations to such properties.

# Possible Developments

- Convolutional/Variational Autoencoders could be used instead of our basic MLP architecture.
  - Convolutional networks have been shown to work well with images.
  - Variational networks would allow us to get more continuous distributions in latent space.
- Disentangled Beta-VAE imposes independence constraints on lower dimensional features and could be used.
- Multi-channel images could be used instead of our monochrome images.

# Thank you

## ご清聴いただきありがとうございました。

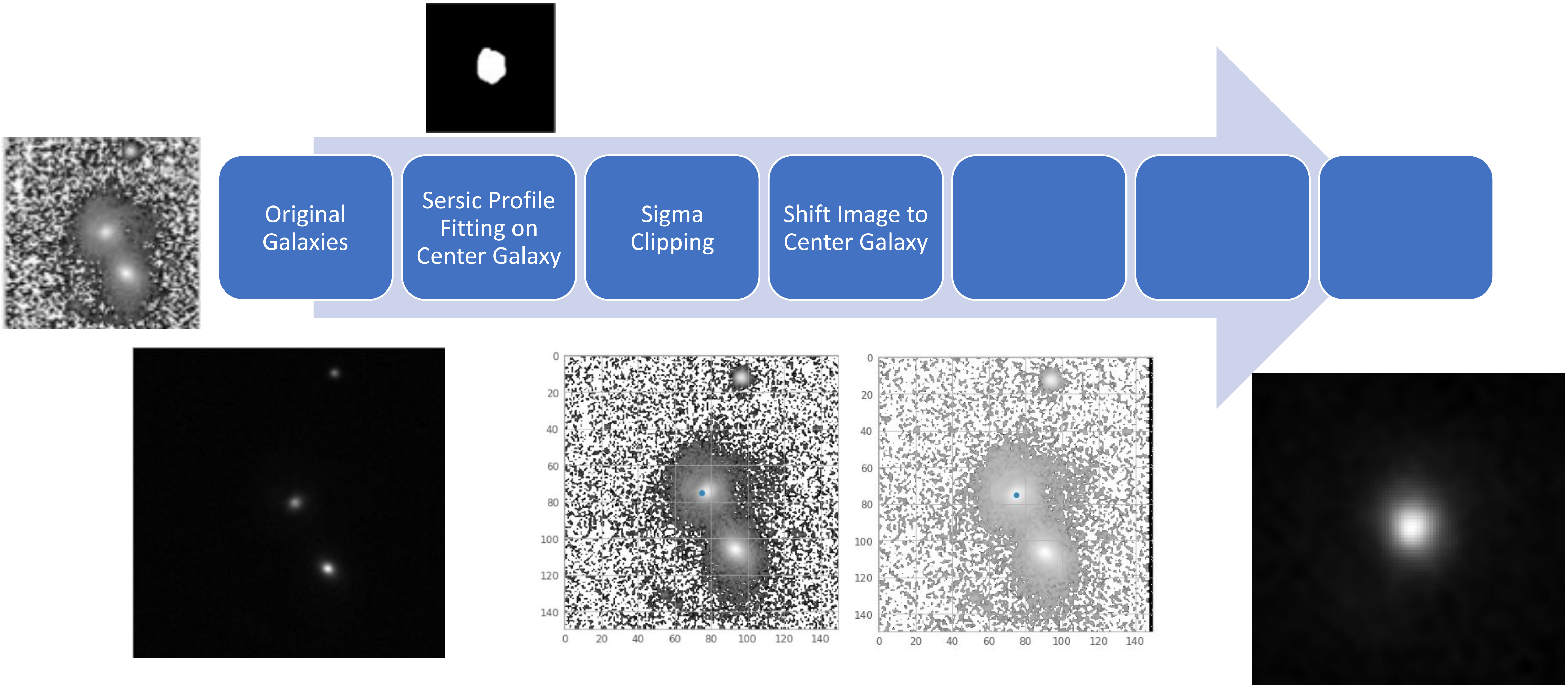
I would like to give special thanks to my mentor, Suchetha Cooray, for guiding me along this research from start to finish.

I would also like to thank my Takeuchi-san and my Omega Lab mates for all the help they have given me over the past 1.5 years.

Finally, thank you everyone here for taking the time to listen to my presentation.

# Appendix Slides

# Preprocessing of Data



Original Galaxies

Sersic Profile Fitting on Center Galaxy

Sigma Clipping

Shift Image to Center Galaxy

# Singular Value Decomposition in 2D

- Consider some 2 x 2 Transform Matrix $\boldsymbol{M}$**.**
- Let $\boldsymbol{v_1}, \boldsymbol{v_2}$ be two orthogonal vectors in our 2D space.
- If we apply our transform to these vectors, we get $\boldsymbol{Mv_1}, \boldsymbol{Mv_2}$.
- If we define the 'singular values' to be $\sigma_i = \frac{1}{|Mv_i|}$ , then we can write:

$$M\boldsymbol{v_1} = \boldsymbol{u_1}\sigma_1$$

$$M\boldsymbol{v_2} = \boldsymbol{u_2}\sigma_2$$

where $\boldsymbol{u_i} = \frac{1}{|\boldsymbol{Mv_i}|}\boldsymbol{Mv_i}$

- One can then express $M$ in the form

$$\boldsymbol{M} = [\boldsymbol{u_1}\ \boldsymbol{u_2}]\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}[\boldsymbol{v_1^T}\ \boldsymbol{v_2^T}] = \mathbf{U\Sigma V}^{\boldsymbol{T}}\ (eigendecomposed)$$

# Singular Values

- Consider an real *m x n* matrix **A**, of rank r. We note that

$$B = A^T A$$

will be an *n x n* REAL SYMMETRIC matrix since

- One can prove* that this REAL SYMMETRIC matrix will have *n* linearly independent eigenvectors $\lambda_n$

- We define the ‘singular values’ of matrix **A** to be

$$\sigma_n = \sqrt{\lambda_n}$$

# Finding PCs via Singular Value Decomposition

- While one can directly eigen-decompose the covariance matrix and get the PCs, it is not the most numerically stable**
- One can instead apply SVD on the dataset directly.

# Singular Value Decomposition (General)

- ANY m x n matrix can be eigen-decomposed into the form

$$\boldsymbol{M} = [\boldsymbol{u_1} \ldots \boldsymbol{u_m}] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} [\boldsymbol{v_1} \; \ldots \boldsymbol{v_n}] = \boldsymbol{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$$

*If n != m, this is padded with 0s*

- Since the singular values $\sigma_i$ can range from big to small, we can ignore terms below a certain cutoff, and use the remaining terms to estimate M.

$$\boldsymbol{M} \approx \boldsymbol{M_k} = \boldsymbol{U_k}\boldsymbol{\Sigma}_{\mathbf{k}}\mathbf{V}_{\mathbf{k}}^{\mathbf{T}}$$

# From Truncated SVD to PCA

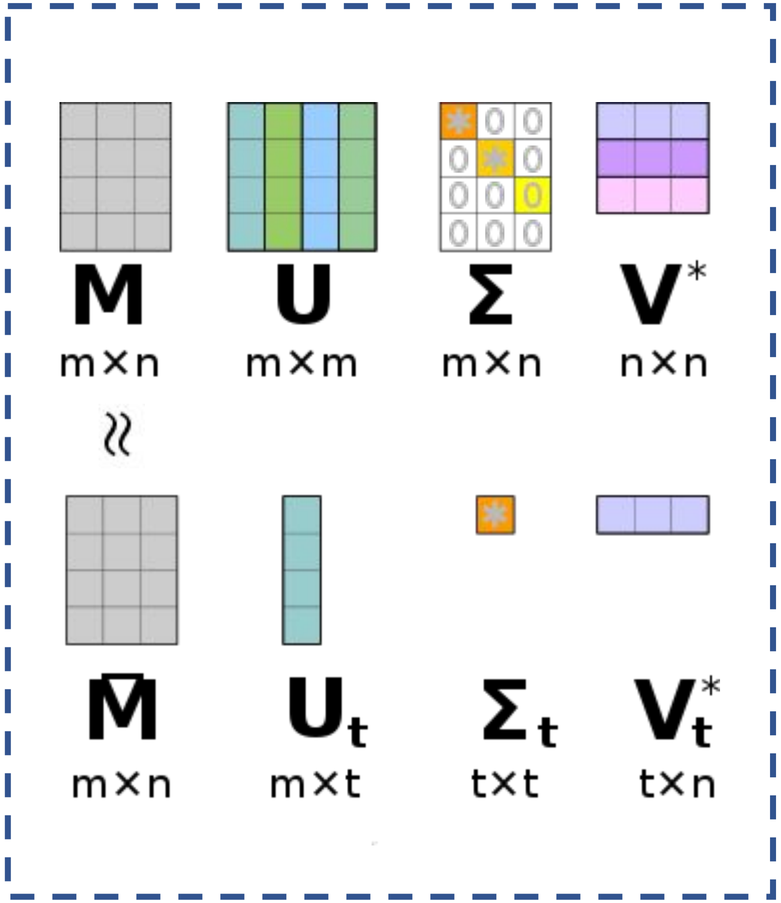
Assuming that the data has been centered (mean = 0):

- One can show that the columns of $\mathbf{V}$ in $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$ are precisely the principal components.
    - The eigenvalues $\lambda_i$ of each PC can be gotten from $\boldsymbol{\Sigma}$, where

$$\lambda_i = \frac{\sigma_i^2}{n-1}$$

    - Each row of $\boldsymbol{U\Sigma}$ is a datapoint transformed into the eigen/PC space.

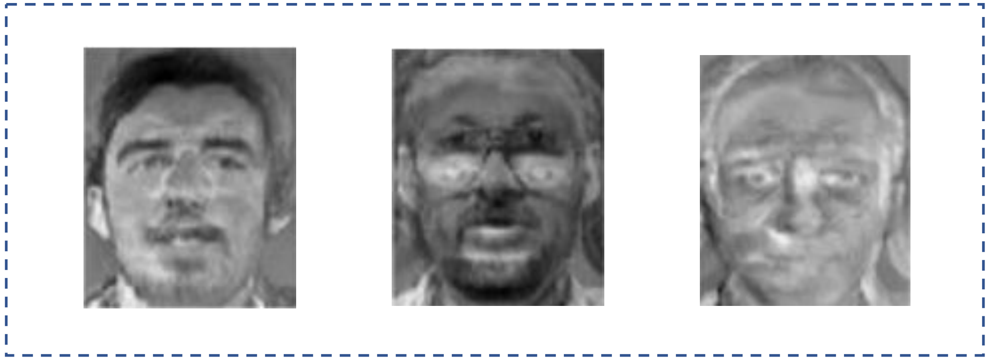
- We can choose how many PC based on the eigenvalues/explained variance.



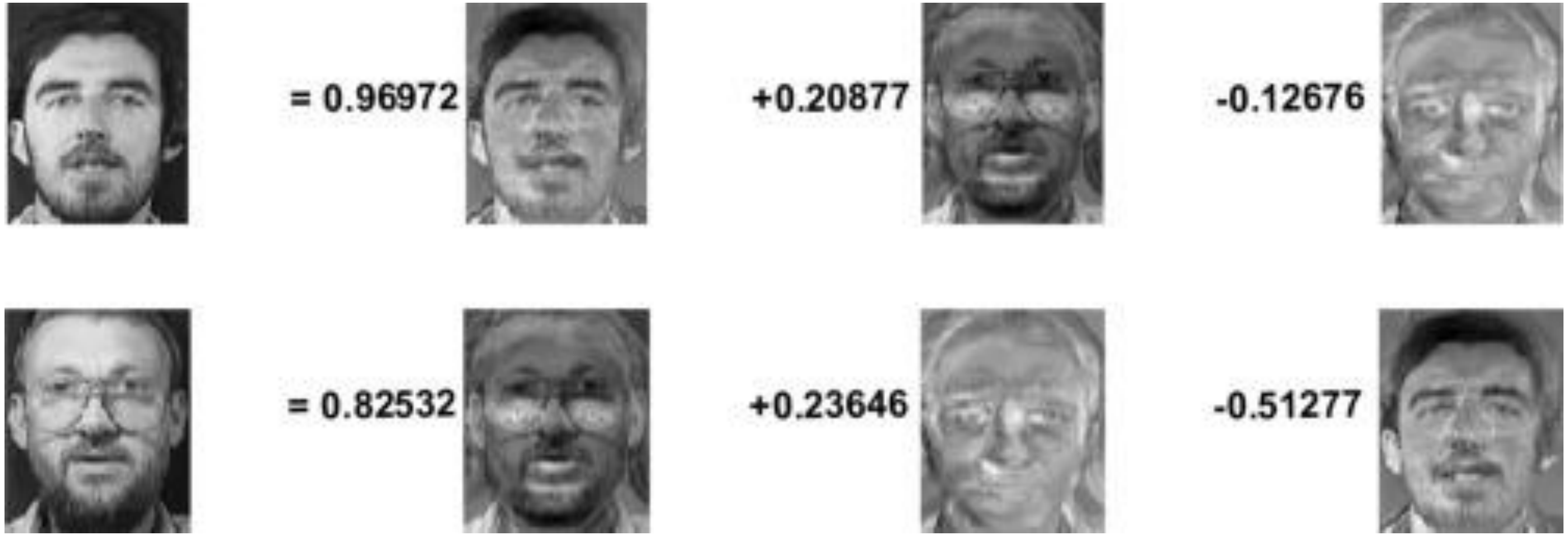
# Introduction/Rationale

- In recent years, amount of data available has increased explosively.
  - ‘Big Data’
- It can be difficult to extract information from high-dimensional data.
  - ‘Curse of Dimensionality’
- It can be difficult to work with high-dimensional data.
  - ‘Computationally Intractable’
- Naturally, we want to find ways to ‘compress’ this information.
  - ‘Dimensional reduction’

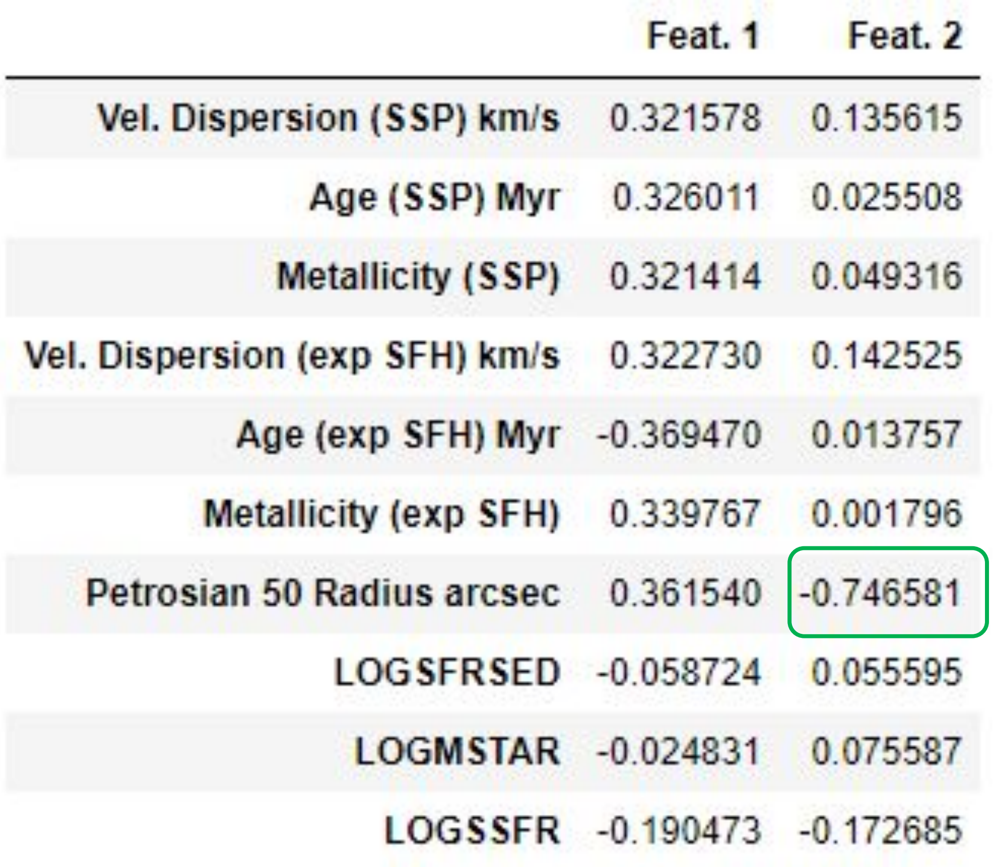
# 直感



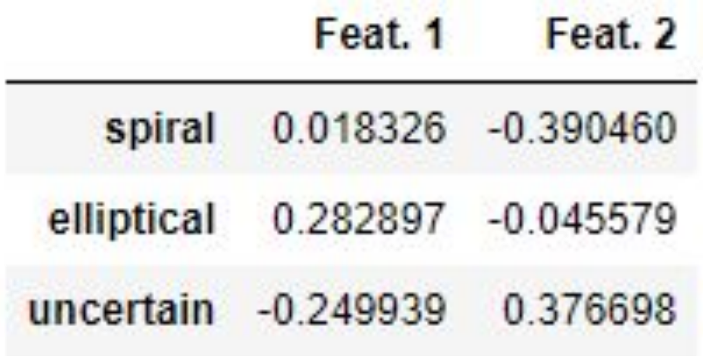
世の中のすべての顔イメージを、この3つのイメージの総和に分解できたらいいと思いませんか？



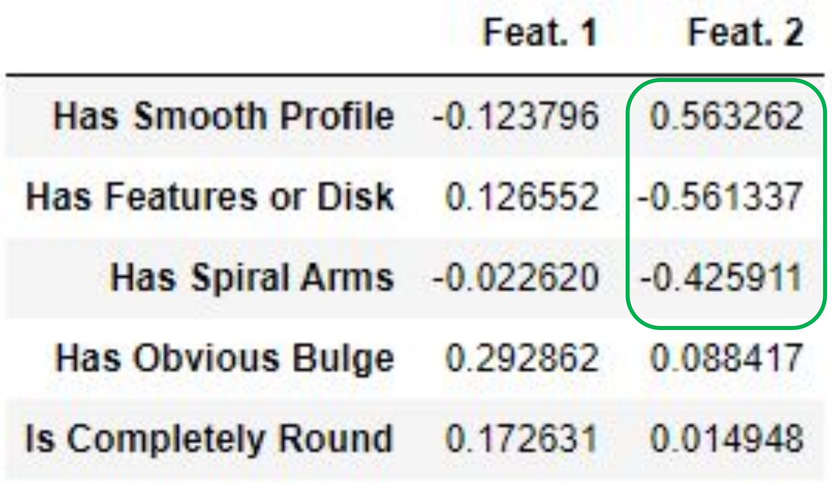
# Correlations between Feats. and Properties



| | Feat. 1 | Feat. 2 |
|---|---|---|
| Vel. Dispersion (SSP) km/s | 0.321578 | 0.135615 |
| Age (SSP) Myr | 0.326011 | 0.025508 |
| Metallicity (SSP) | 0.321414 | 0.049316 |
| Vel. Dispersion (exp SFH) km/s | 0.322730 | 0.142525 |
| Age (exp SFH) Myr | -0.369470 | 0.013757 |
| Metallicity (exp SFH) | 0.339767 | 0.001796 |
| Petrosian 50 Radius arcsec | 0.361540 | -0.746581 |
| LOGSFRSED | -0.058724 | 0.055595 |
| LOGMSTAR | -0.024831 | 0.075587 |
| LOGSSFR | -0.190473 | -0.172685 |



| | Feat. 1 | Feat. 2 |
|---|---|---|
| spiral | 0.018326 | -0.390460 |
| elliptical | 0.282897 | -0.045579 |
| uncertain | -0.249939 | 0.376698 |



| | Feat. 1 | Feat. 2 |
|---|---|---|
| Has Smooth Profile | -0.123796 | 0.563262 |
| Has Features or Disk | 0.126552 | -0.561337 |
| Has Spiral Arms | -0.022620 | -0.425911 |
| Has Obvious Bulge | 0.292862 | 0.088417 |
| Is Completely Round | 0.172631 | 0.014948 |