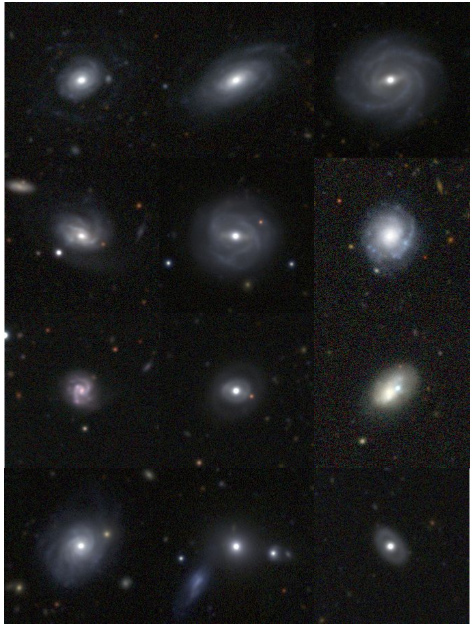
Searching for similarities and anomalies in a pool of galaxy images

Matias Carrasco Kind
Senior Research Scientist, National Center for Supercomputing Applications
Assistant Research Professor, Astronomy
University of Illinois at Urbana-Champaign

Artificial Intelligence for Data Discovery and Reuse, May 13-15 2019
Carnegie Mellon University, Pittsburgh, PA

Motivation



Astronomy is just one example where image exploration needs to be automated.

Large catalogs, Large number of images, many unexpected objects/problems → <u>Anomaly detection</u>

Large Synoptic Survey Telescope

- In operations 2021
- Every night for 10 years
  - 15 TB per night
- 18 billions objects (first year), ~40 billions by the end of survey
- ~1500 images per night
- Stream and static data
- Target to capture new physics (moving and variable objects)

DARK ENERGY SURVEY

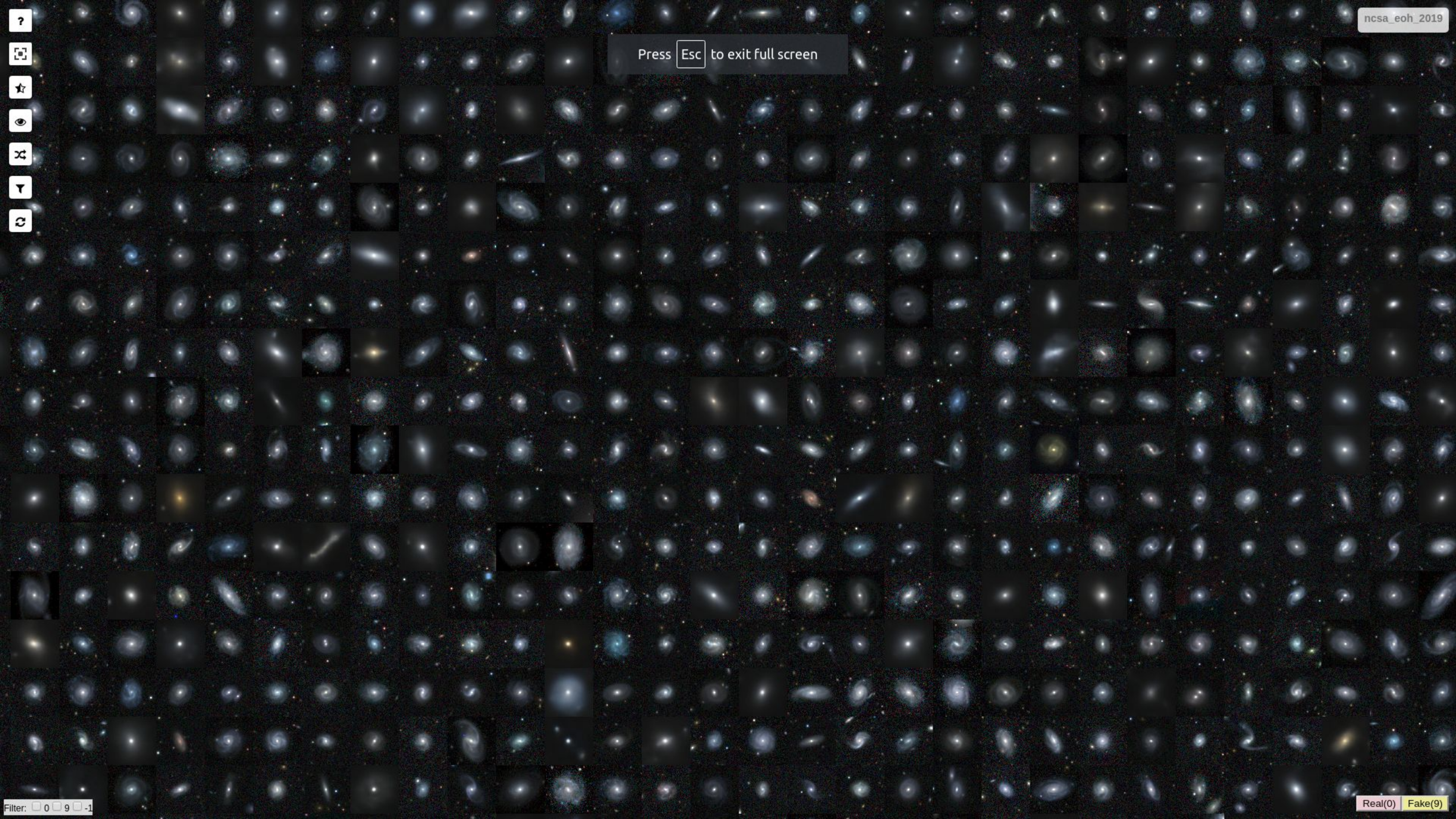
- More than 500 nights of observation over 5 years, 2TB per night
- 500 millions cataloged galaxies and 100 millions stars
- Many open problems: Systematics, new objects, new physics, etc.
- Almost completed

Current (personal) data discovery challenges

- Visualize large set of galaxy (or other) images
- Quickly classify images for AI using multiple experts
- Compress the important information in a efficient way
- Quickly search images by (dis)similarity (several science cases)
- Find anomalies in a image data set (new phenomena, errors, unrepresentative samples)

Not covered here

- Generate and sample realistic fake images based on a training for modeling and Monte-Carlo Sampling
- Generate and sample realistic fake images based on a training in a controlled manner (with a prior)

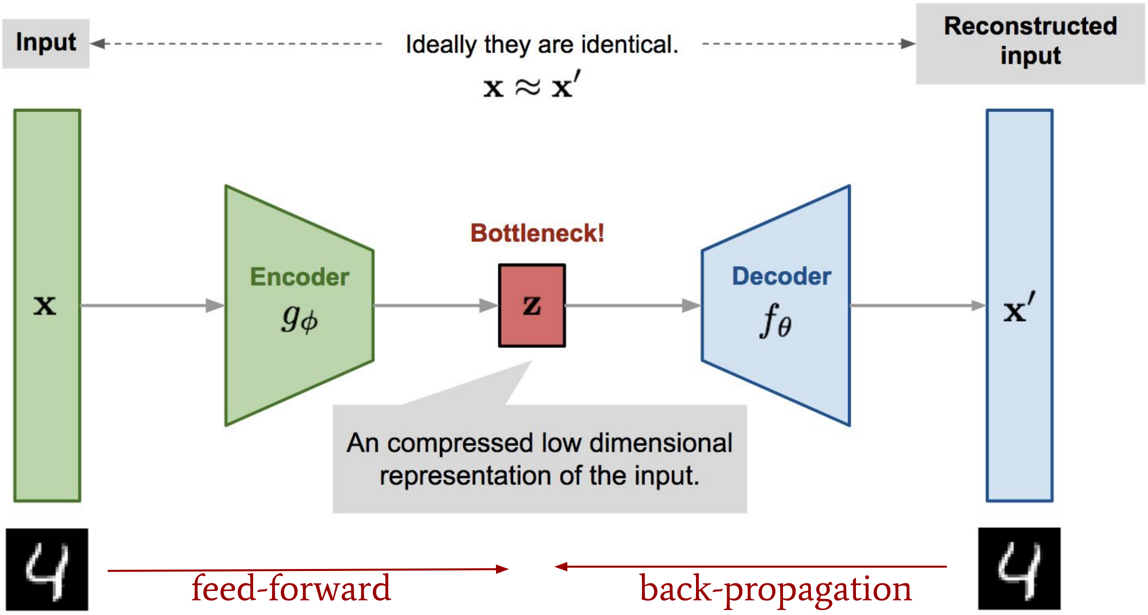

ncsa_eoh_2019
Press Esc to exit full screen
Filter: 0 9 -1
Real(0)
Fake(9)

Galaxy Image Exploration and Classification

- Image Exploration
- Resize is done dynamically
- Quick Classification/Label
- Works fine with 10,000 images
- Individual classifications are saved and aggregated
- Keyboard control

https://github.com/mgckind/cutouts-explorer

Autoencoders review



$$z = g_\phi(x)$$

$$x' = f_\theta(g_\phi(x))$$

$$\mathcal{L}(x, x') + reg.$$

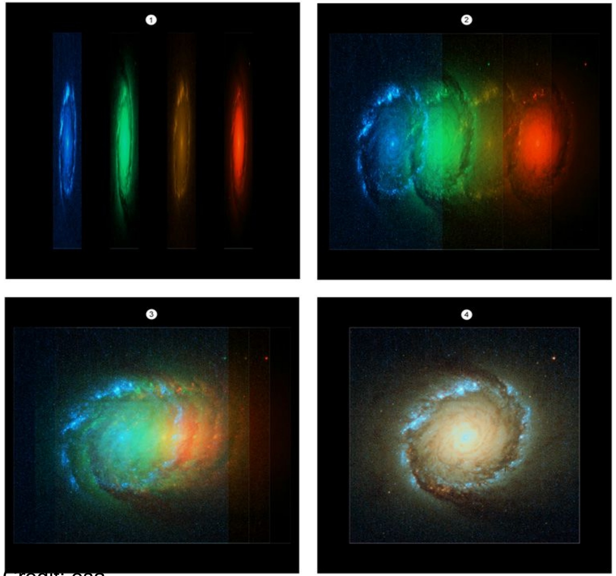
$$\ell_1 = \lambda \sum_i |a_i^{(h)}|$$

$$\mathcal{L}(x, x') = ||x - x'||^2$$

- Around since the 80's
- Data compression
- Anomaly detection
- Denoising
- Regular Machine Learning
- PCA

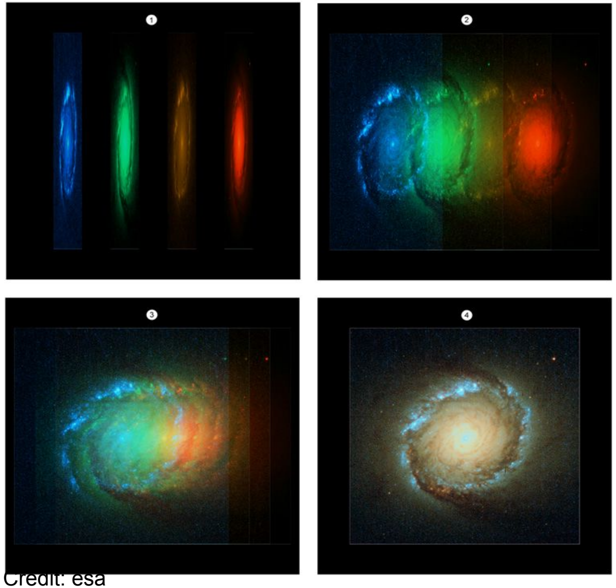
Many variants: Sparse AE, Contractive AE, Stacked AE, etc...

The latent space

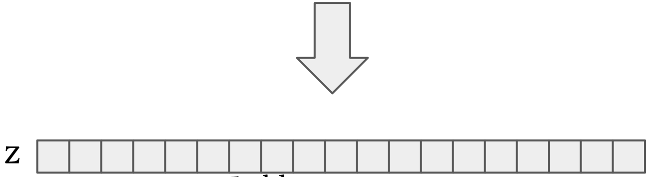


Credit: esa

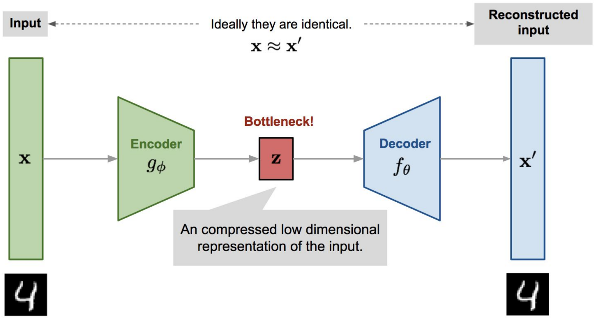
The latent space



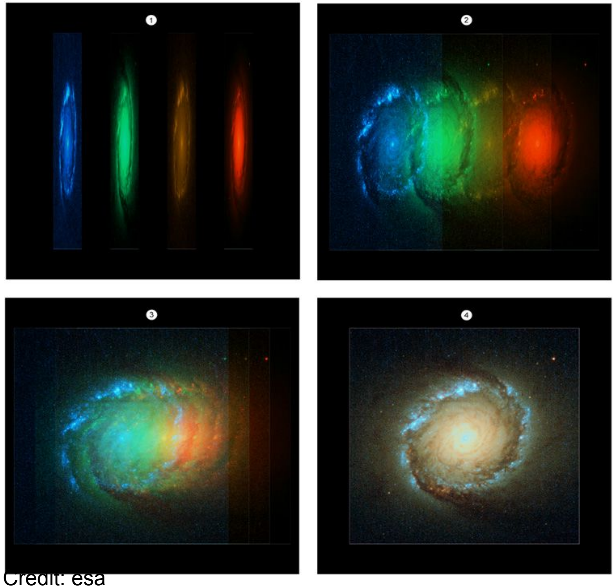
Credit: esa



1-d latent space

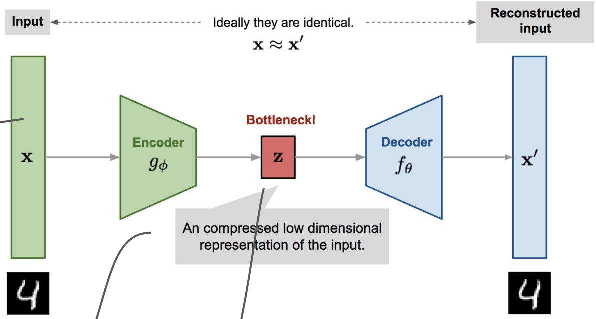


The latent space

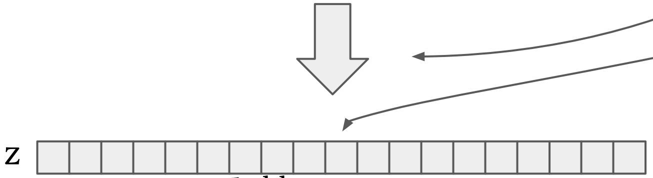


Credit: esa

flatten

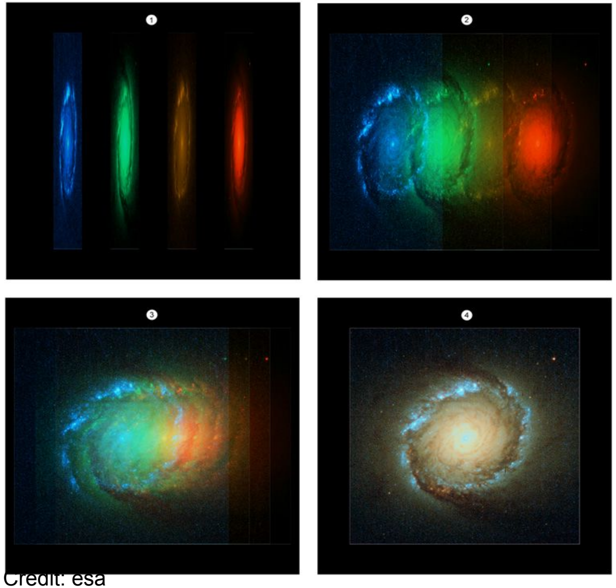


encoding



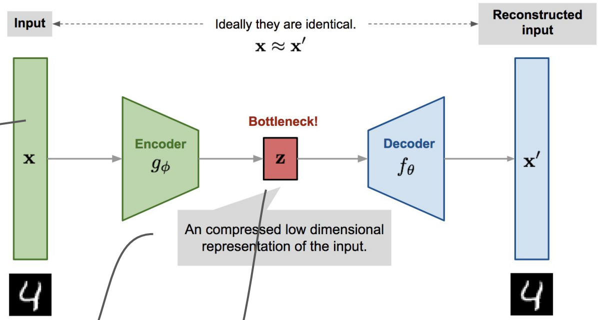
1-d latent space

The latent space

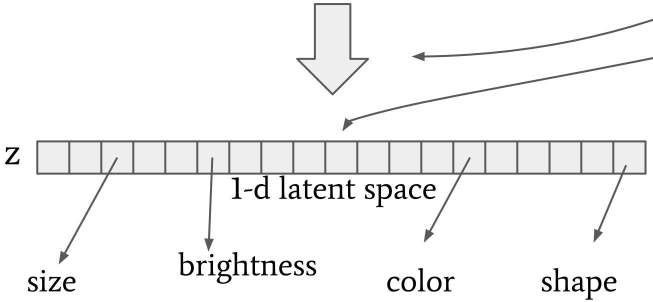


Credit: esa

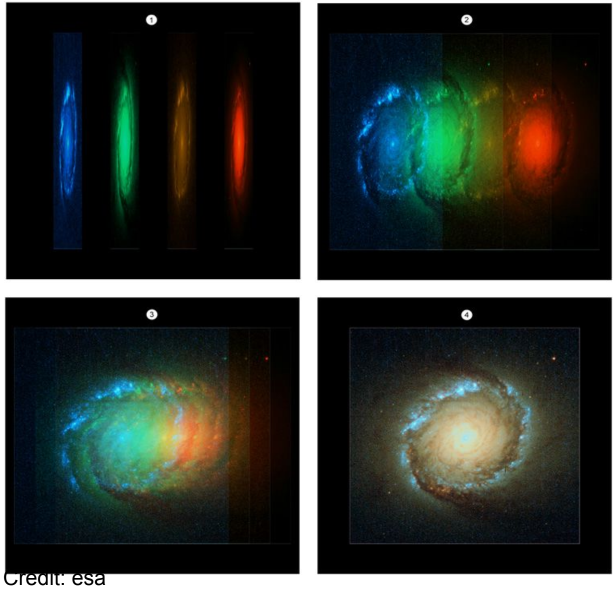
flatten



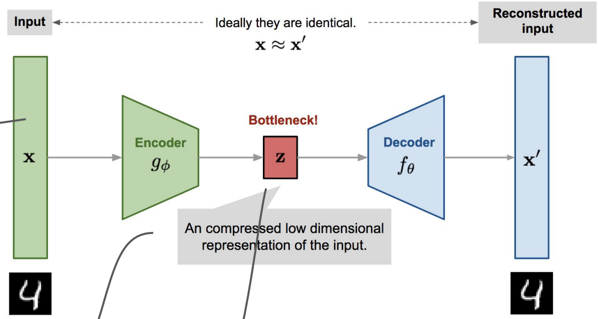
encoding



The latent space

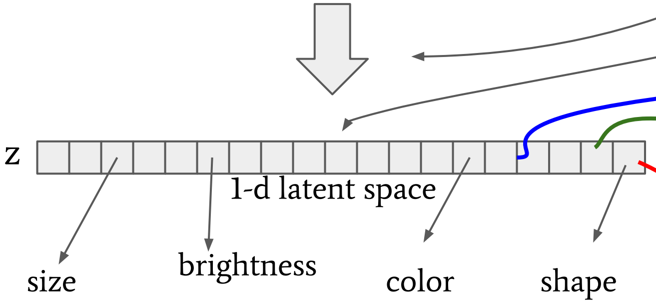


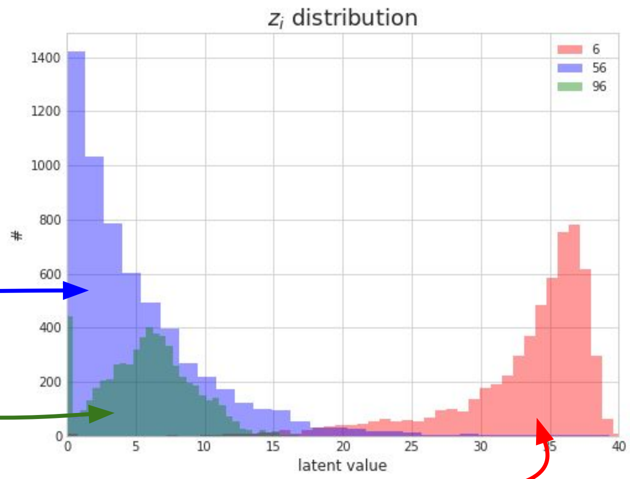
flatten



Each galaxy image is encoded to a unique 1-d vector

encoding

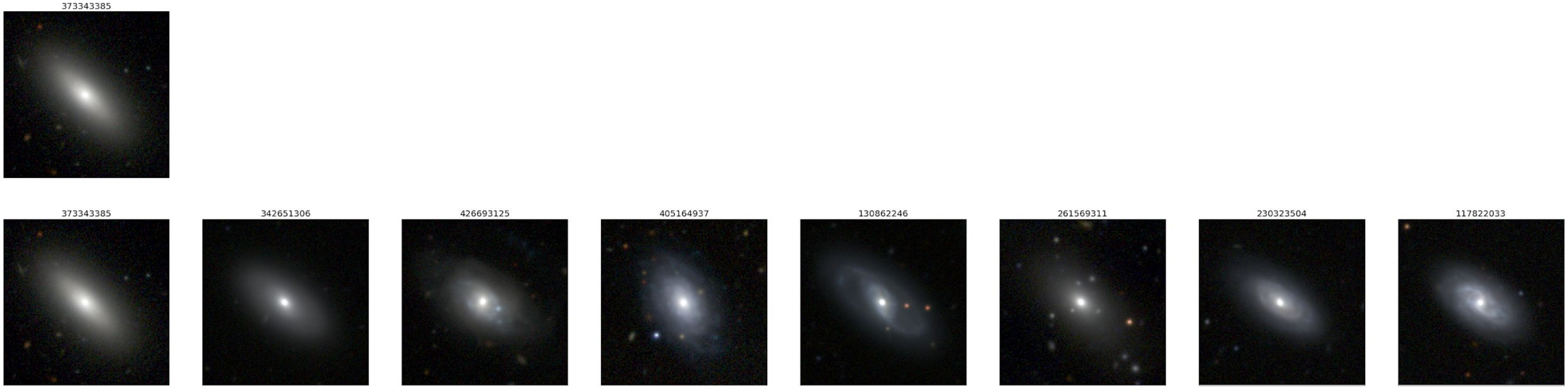




Similarity ranking

Using standard ML in latent space to look for neighbors, outliers, etc.

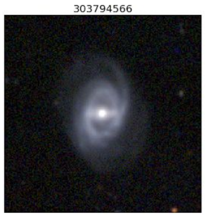
Compress images from 220x220x3 pixels to 100-vector (2000x), for fast similarity search, anomaly detection, etc...
No need decoder (only for Loss)

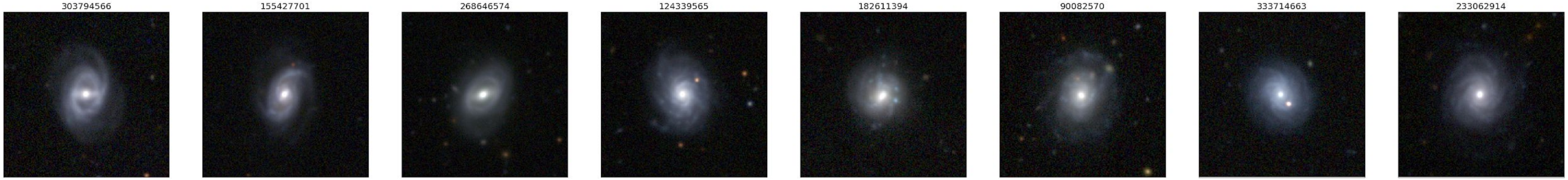


Similarity ranking

Using standard ML in latent space to look for neighbors, outliers, etc.

Compress images from 220x220x3 pixels to 100-vector (2000x), for fast similarity search, anomaly detection, etc...
No need decoder (only for Loss)

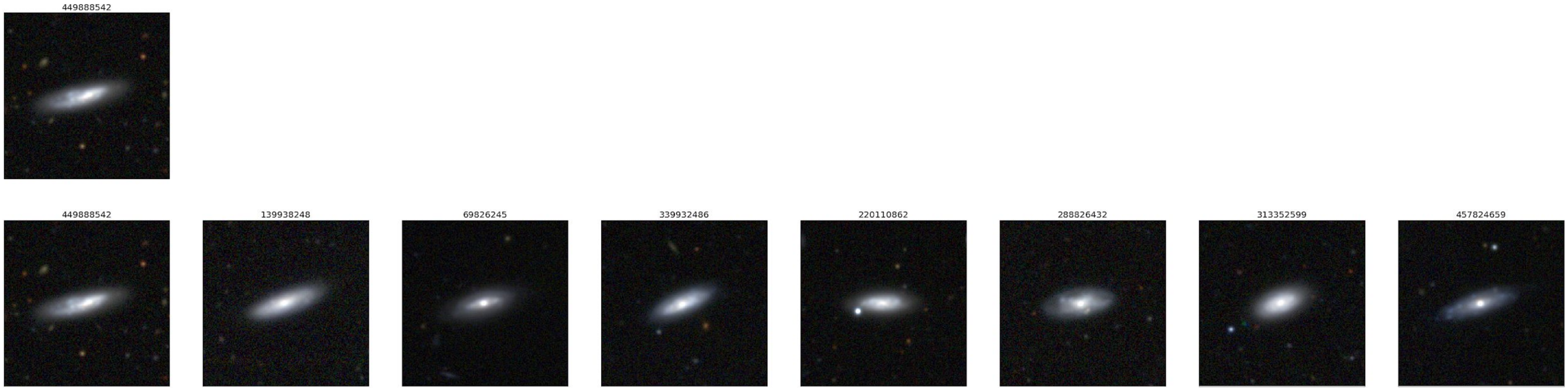




Similarity ranking

Using standard ML in latent space to look for neighbors, outliers, etc.

Compress images from 220x220x3 pixels to 100-vector (2000x), for fast similarity search, anomaly detection, etc...
No need decoder (only for Loss)



Anomaly detection with Extended Isolation Forest

Isolation Forest:

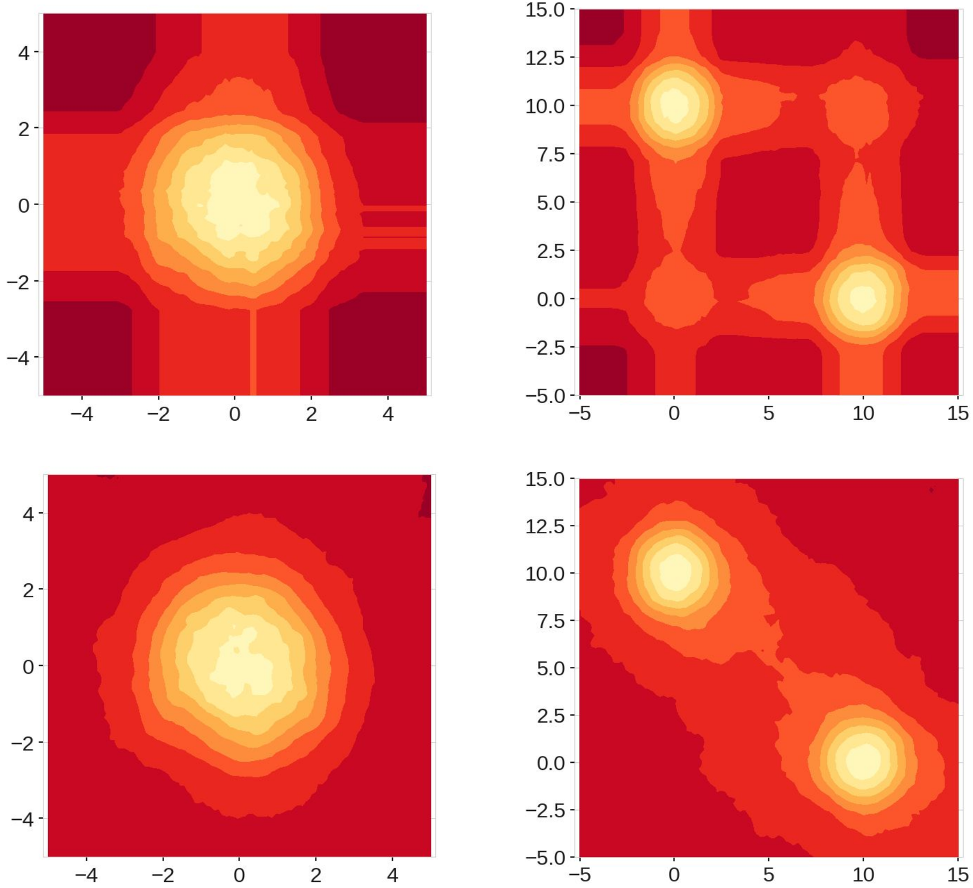
- ✔ Model free
- ✔ Computationally efficient
- ✔ Readily application to high dimensional data
- ❌ Inconsistent scoring seen in score maps

Extended Isolation Forest:

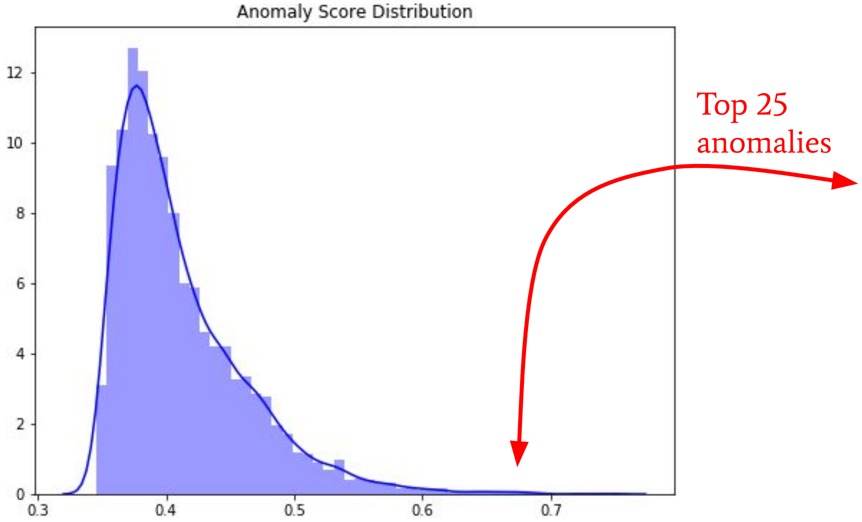
- ✔ Model free
- ✔ Computationally efficient
- ✔ Readily application to high dimensional data
- ✔ Consistent scoring

Hariri, Carrasco-Kind, Brunner, 2019 , arXiv: 1811.02141

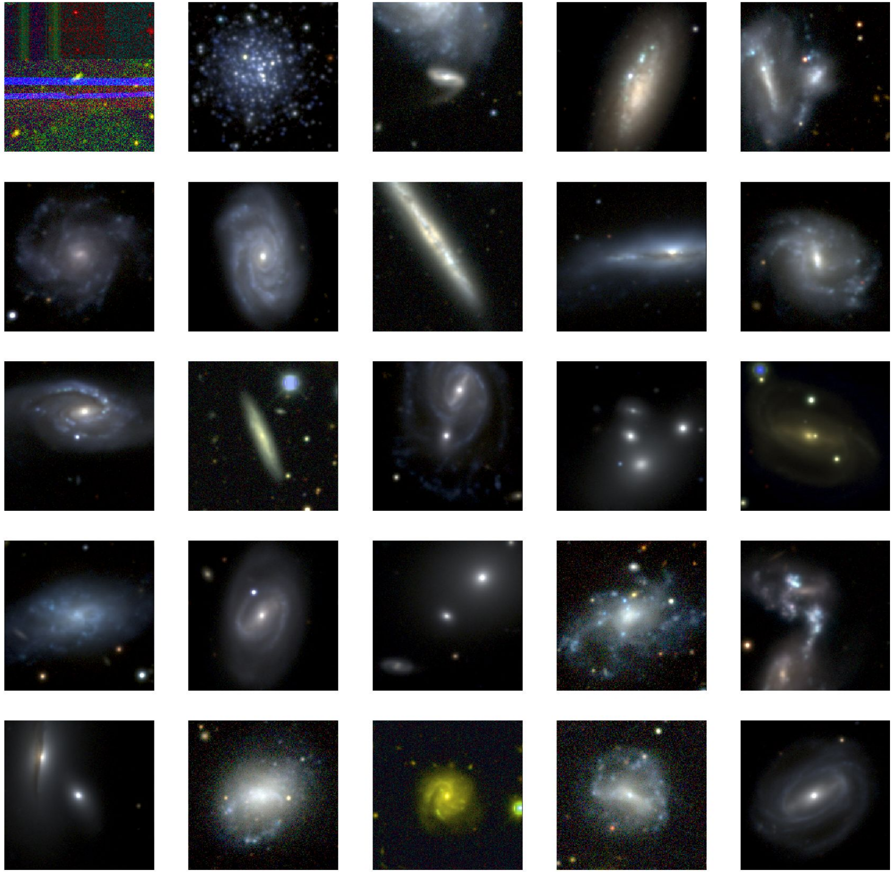
https://github.com/sahandha/eif



Anomaly detection with EIF in high dimensional latent space



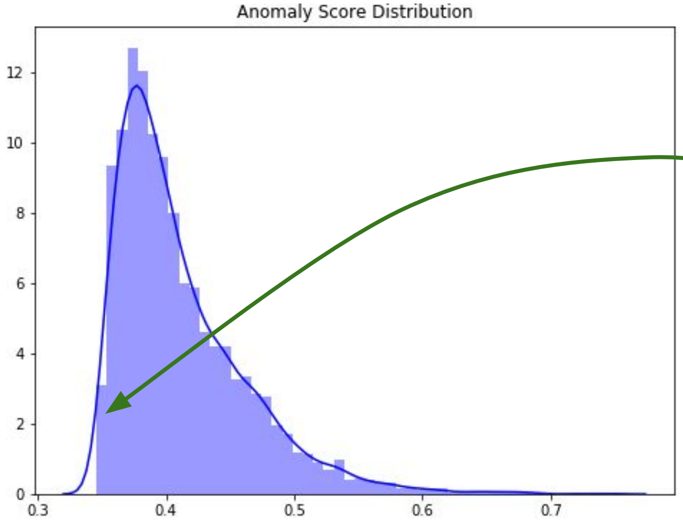
Anomalies →



The EIF algorithm produces a anomaly score that can be used to select outliers galaxies.

Errors, special cases, unrepresented galaxies

Anomaly detection with EIF in high dimensional latent space



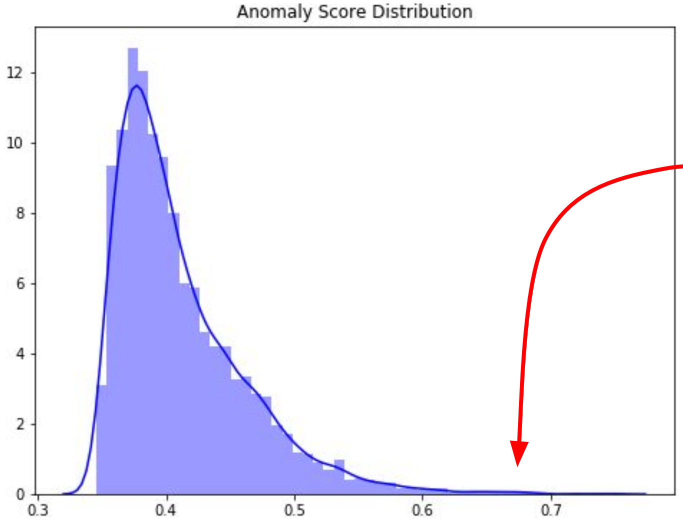
Top 25 nominals

Anomalies

But also can tell us about the repeated and common cases as shown here

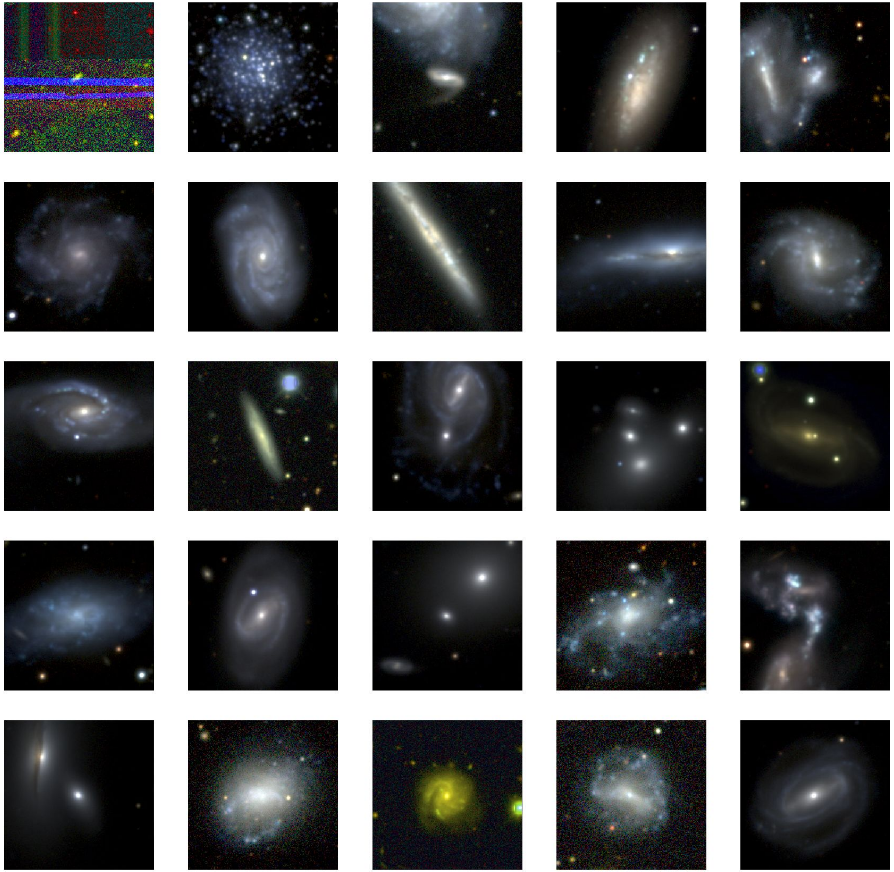
These are all different sources!

Anomaly detection with EIF in high dimensional latent space



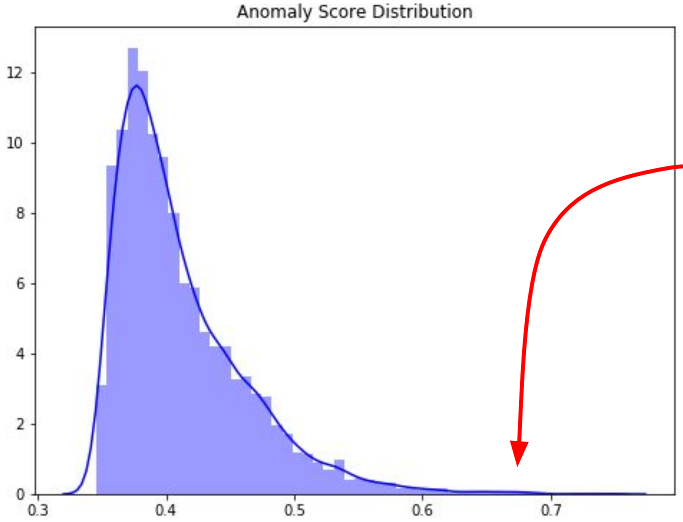
Anomalies →

Top 25 anomalies



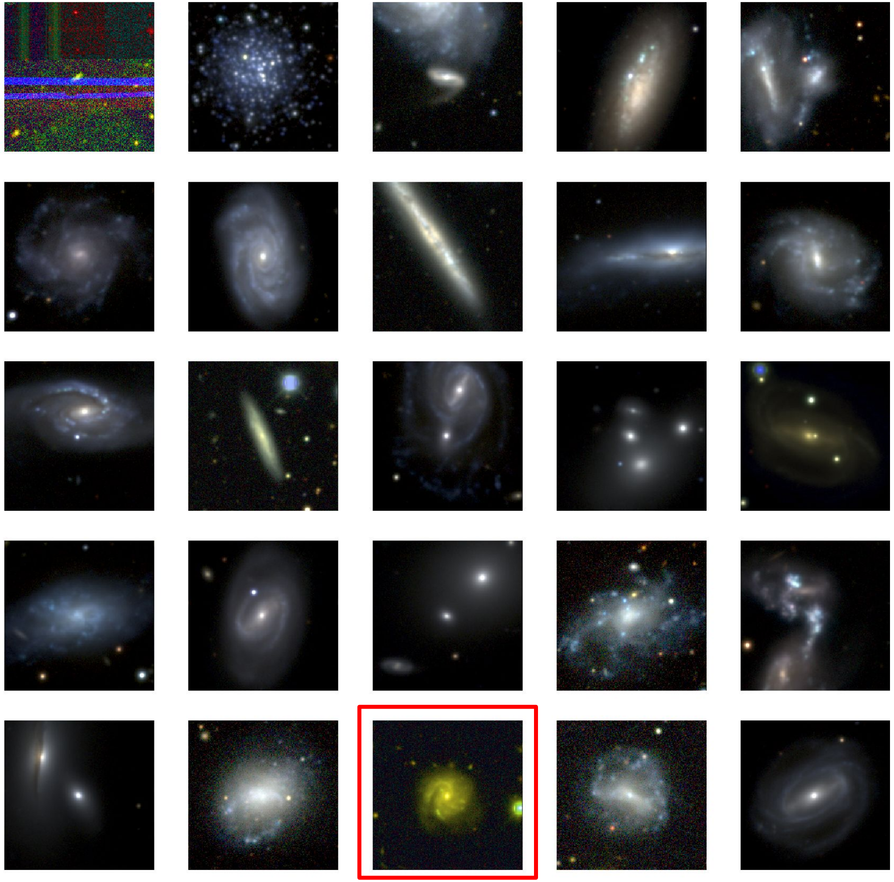
We can find other errors by looking at similarities

Anomaly detection with EIF in high dimensional latent space



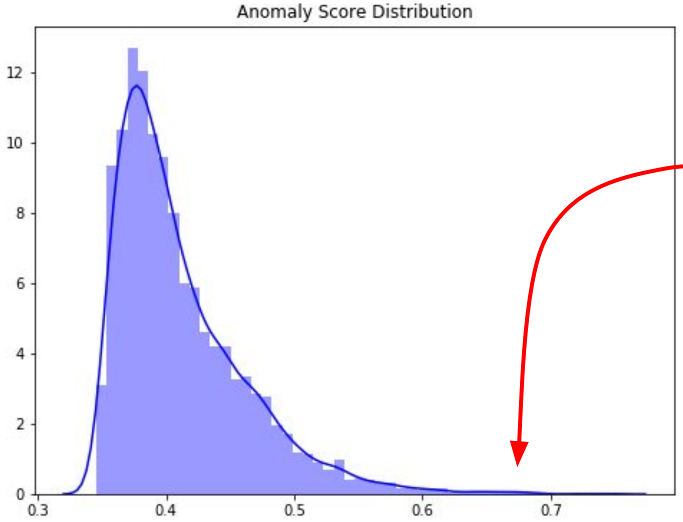
Anomalies →

Top 25 anomalies

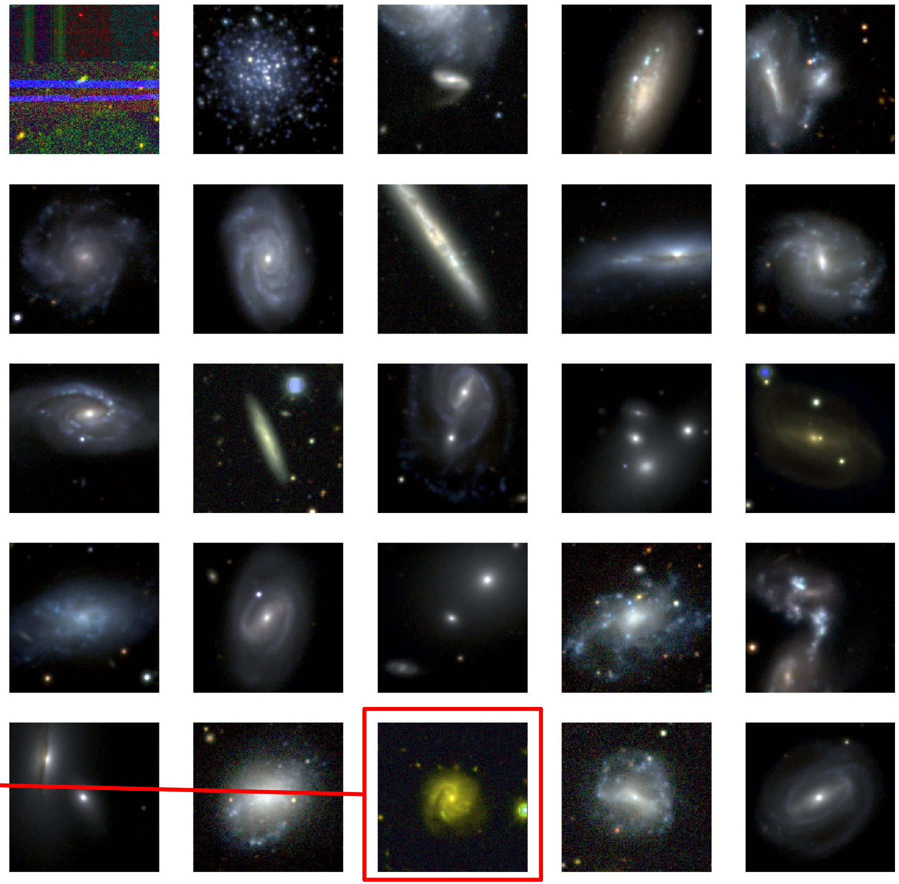


We can find other errors by looking at similarities

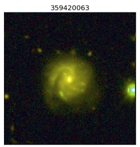
Anomaly detection with EIF in high dimensional latent space



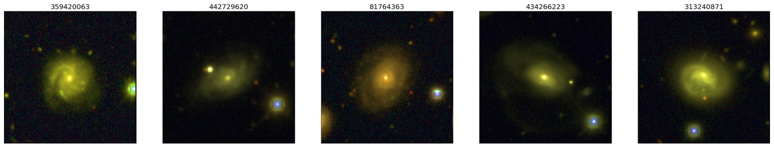
Top 25 anomalies



Anomalies

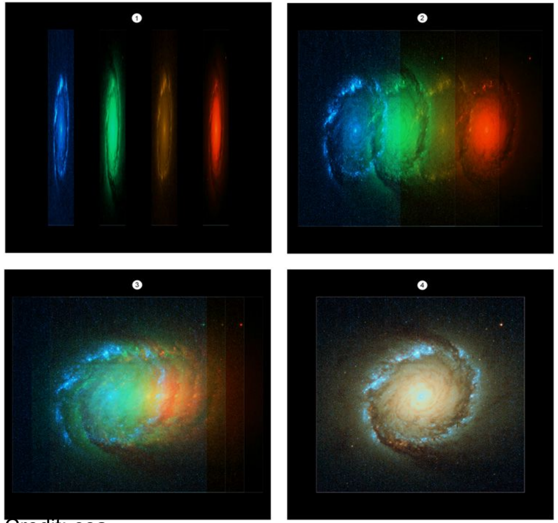


We can find other errors by looking at similarities

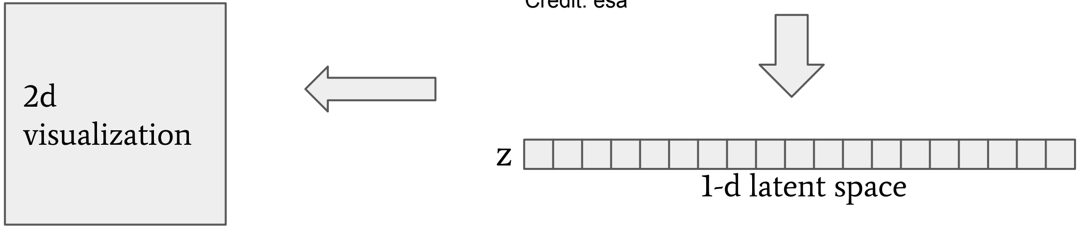


Can we keep reducing dimensions?

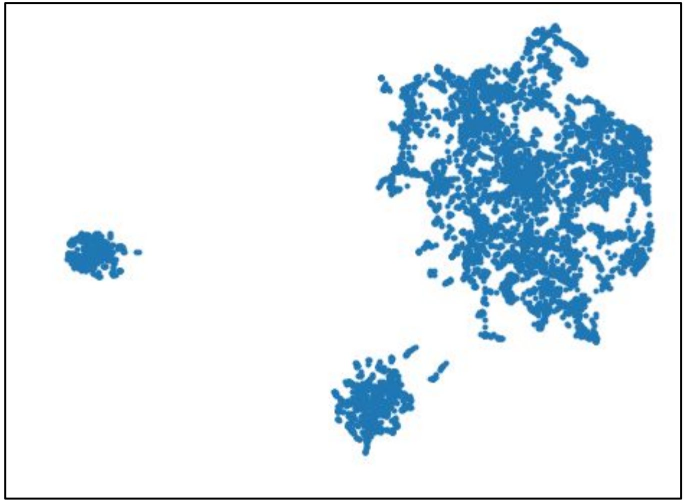
- Apply clustering and unsupervised techniques to latent space to find patterns
- Self Organizing Maps and T-SNE are perfect candidates
- We have used Uniform Manifold Approximation and Projection (UMAP)



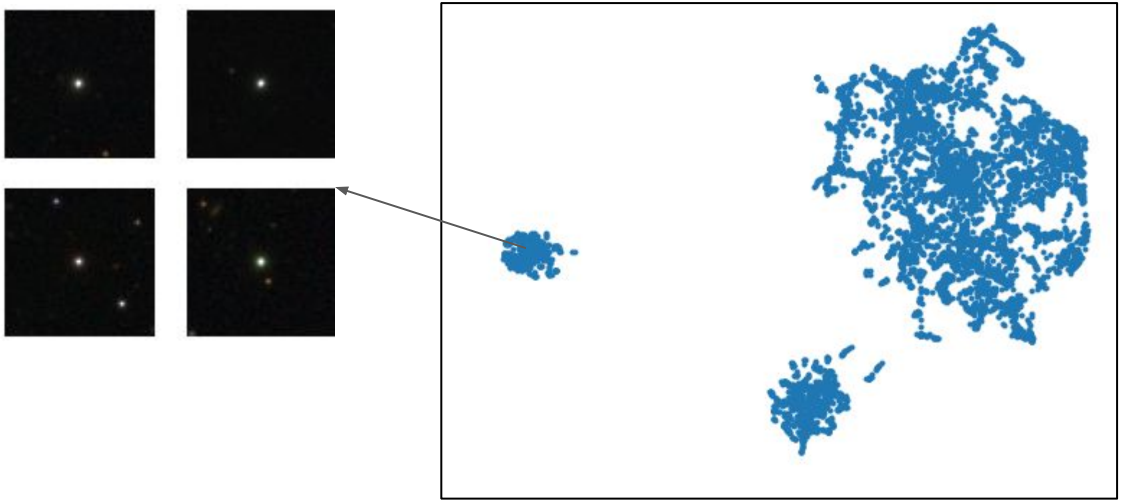
Credit: esa



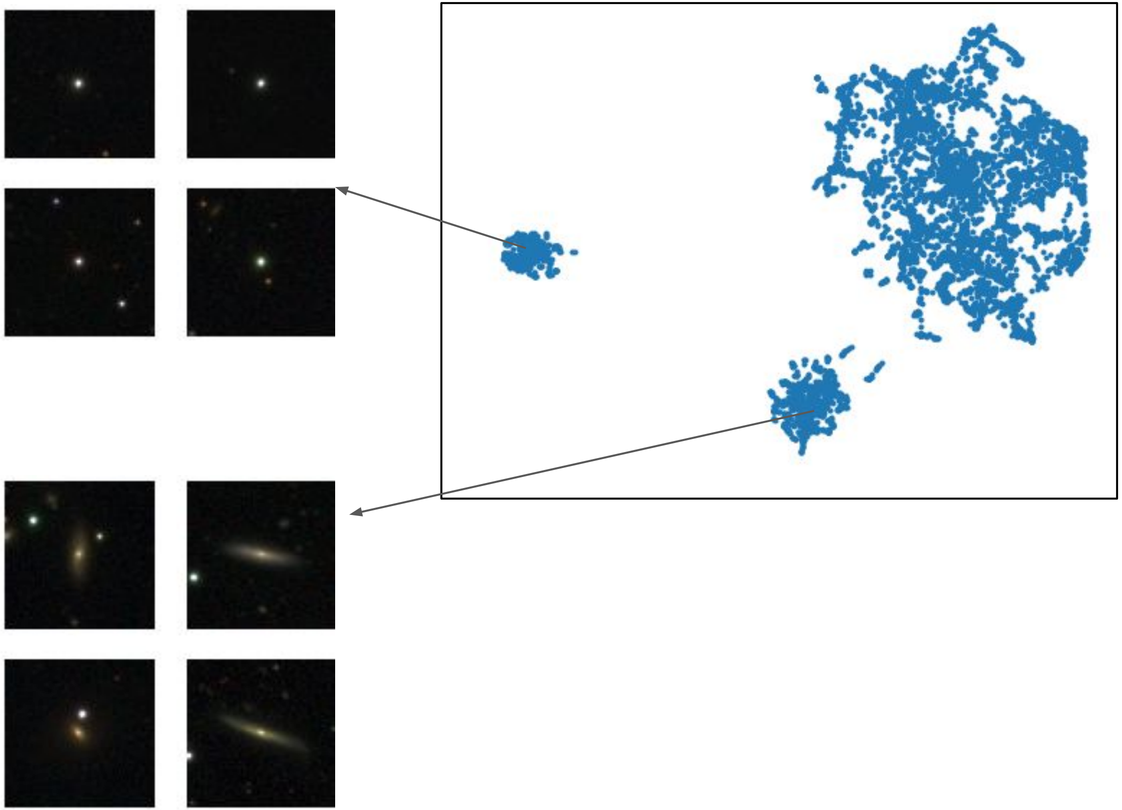
UMAP Representation (6000 galaxies)



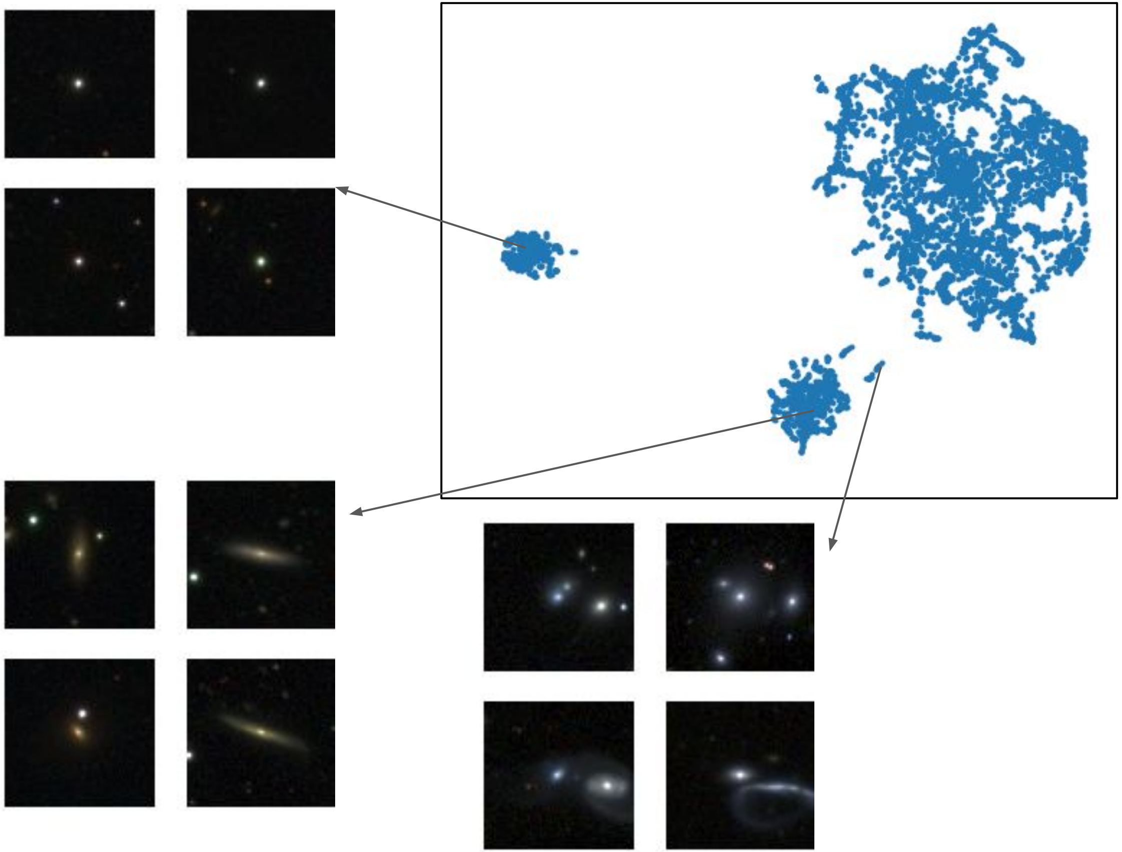
UMAP Representation (6000 galaxies)



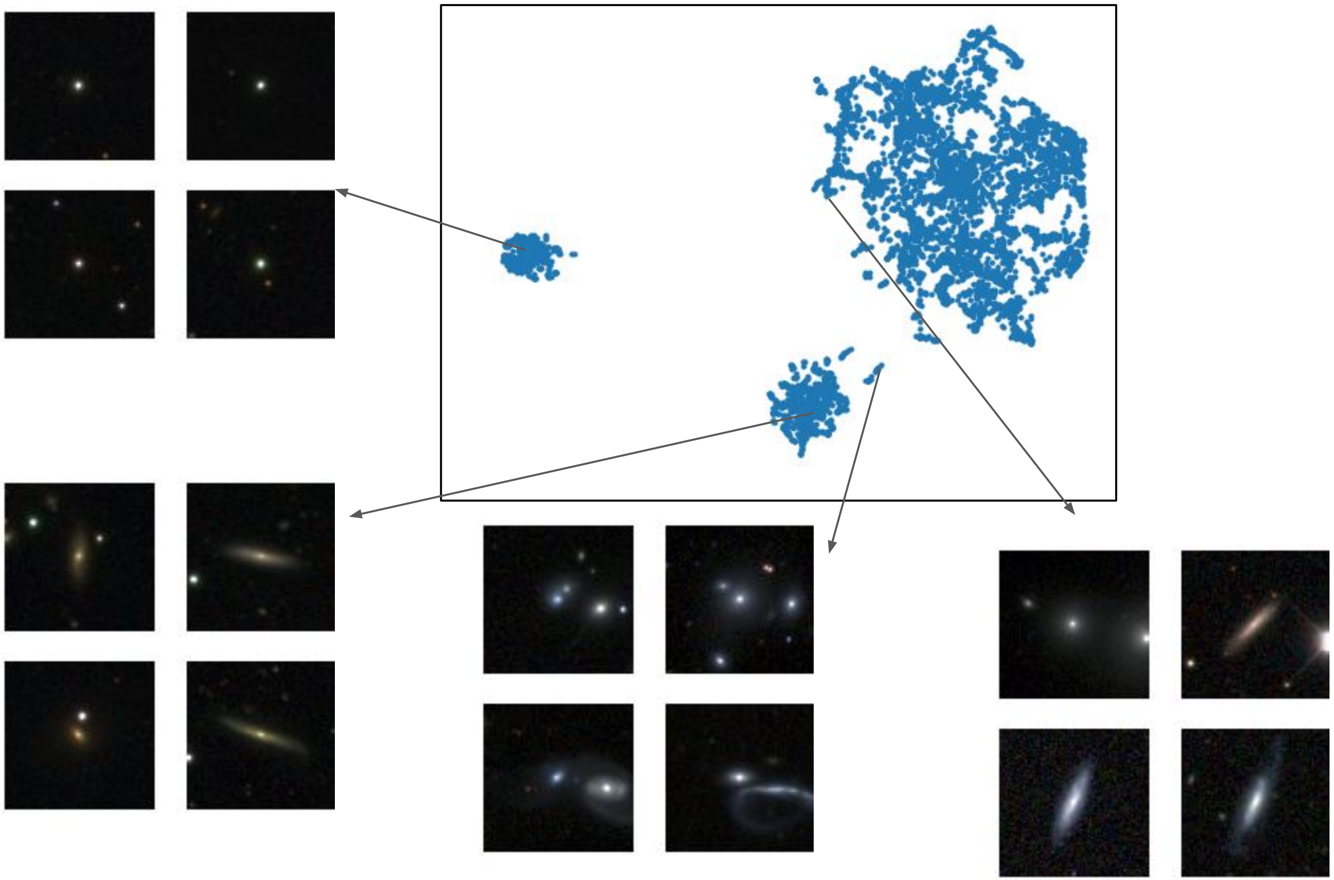
UMAP Representation (6000 galaxies)



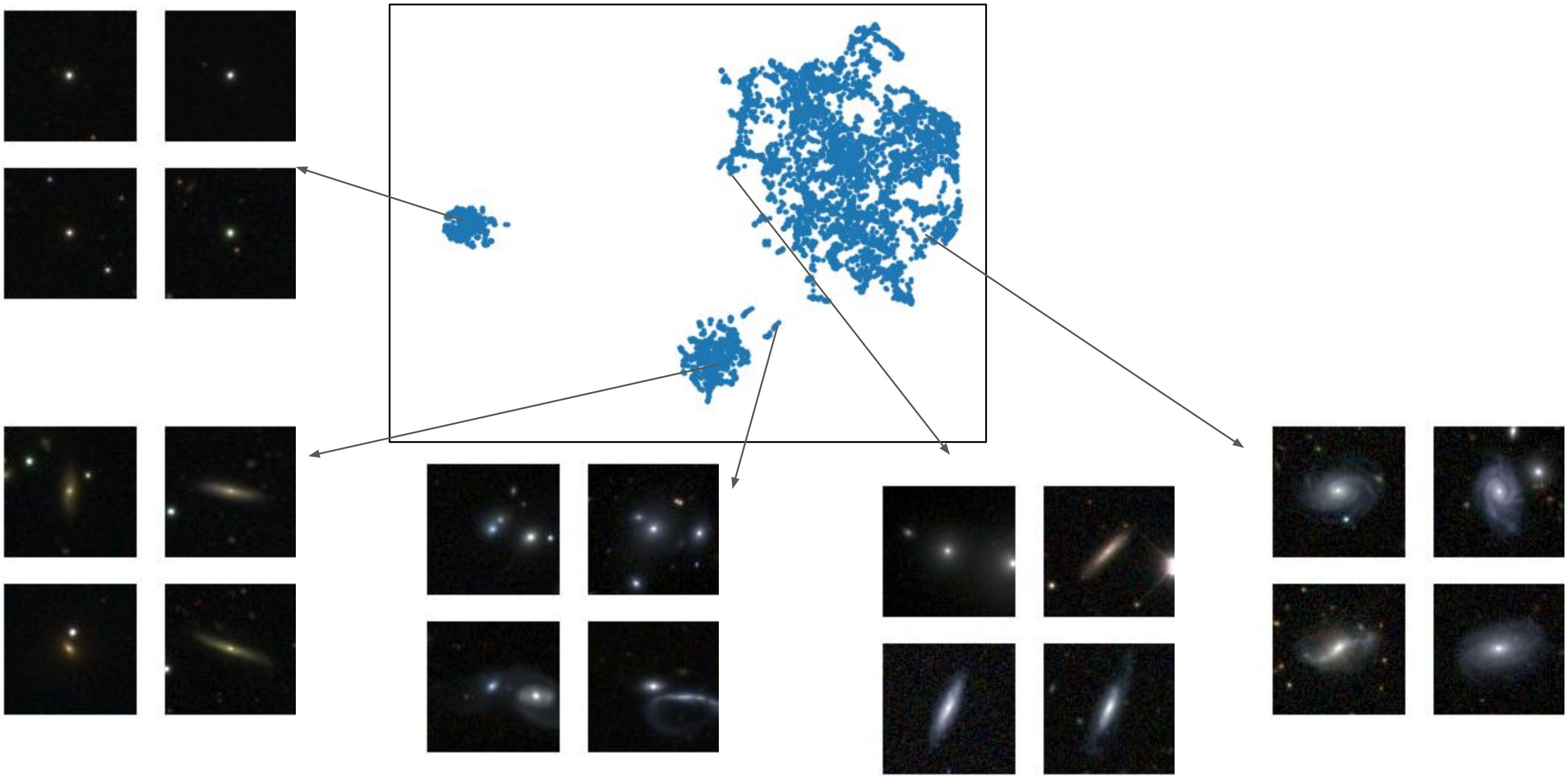
UMAP Representation (6000 galaxies)



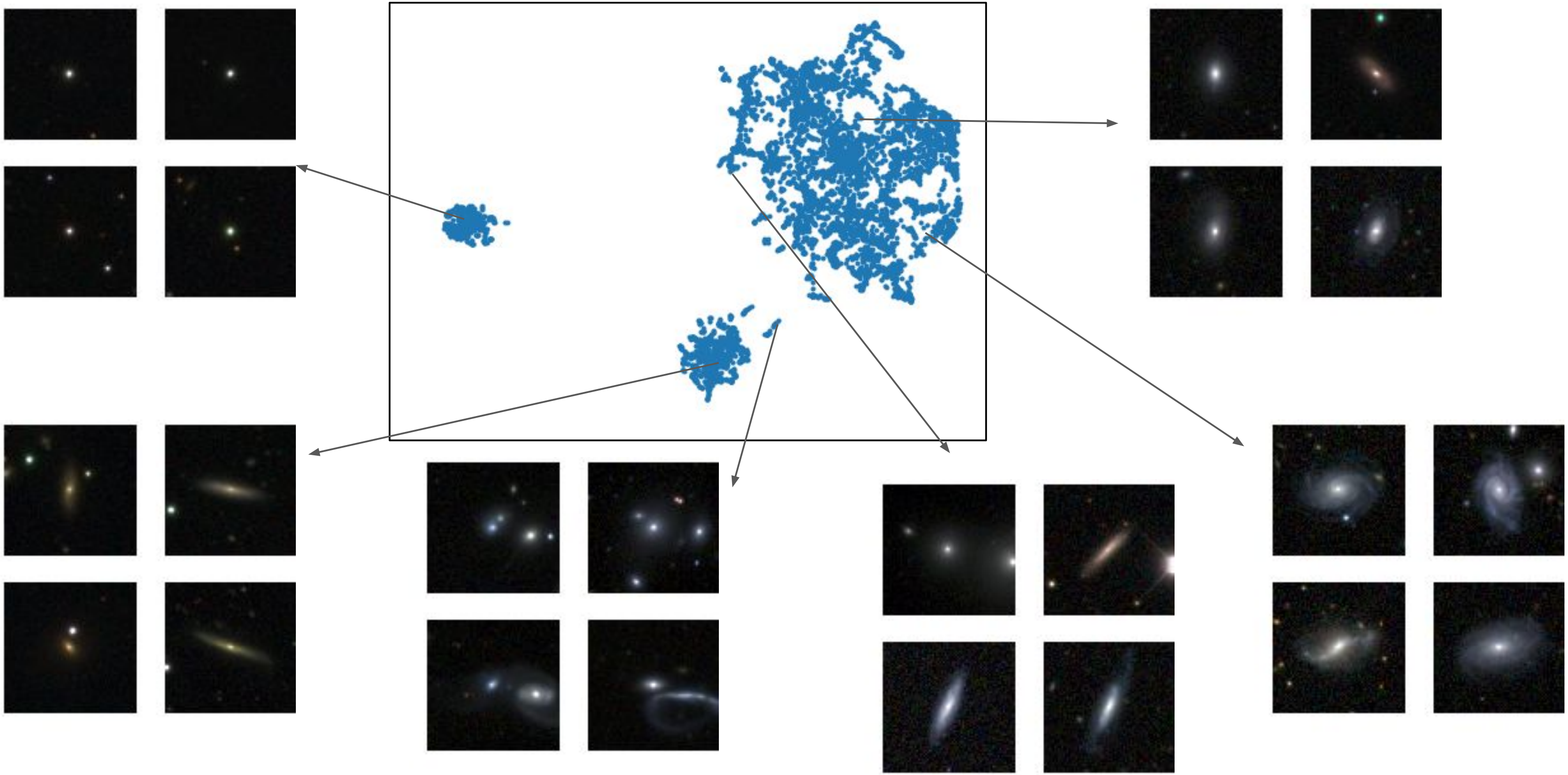
UMAP Representation (6000 galaxies)



UMAP Representation (6000 galaxies)



UMAP Representation (6000 galaxies)

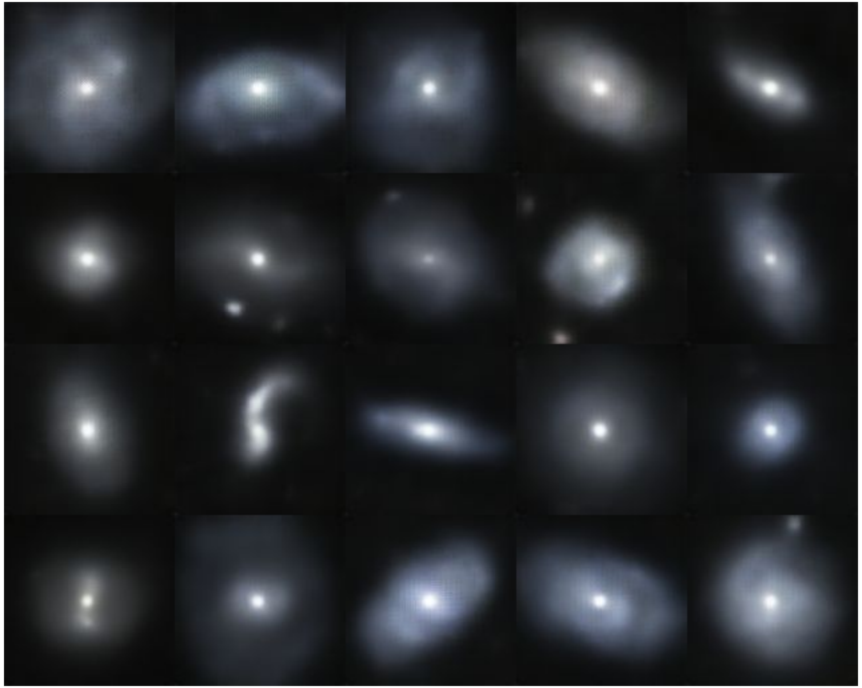


Conclusions

- We developed a visualization and classification tool for multiple images
- Using Autoencoders we can compress images to small (but high-n) latent space
- Look for similarities and anomalies in that space
- Represent even more in a 2d graph using t-SNE, SOM or UMAP
- Scientific driven cases
- State-of-the-art models allows a bayesian manipulation of the latent space

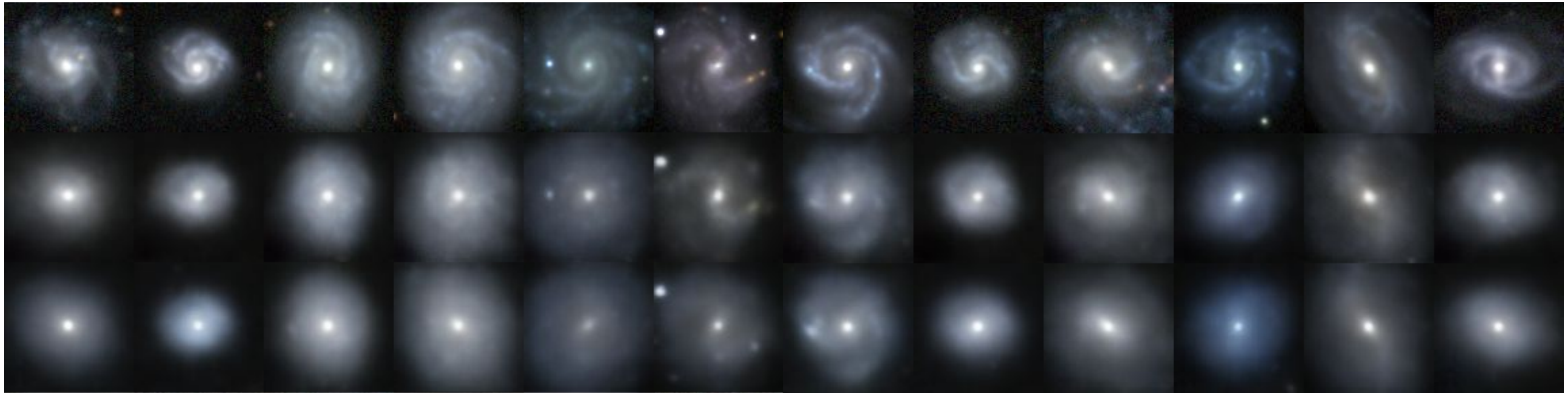
Thank you!

Questions?



Matias Carrasco Kind -- NCSA
mcarras2@illinois.edu
github.com/mgkind
matias-ck.com

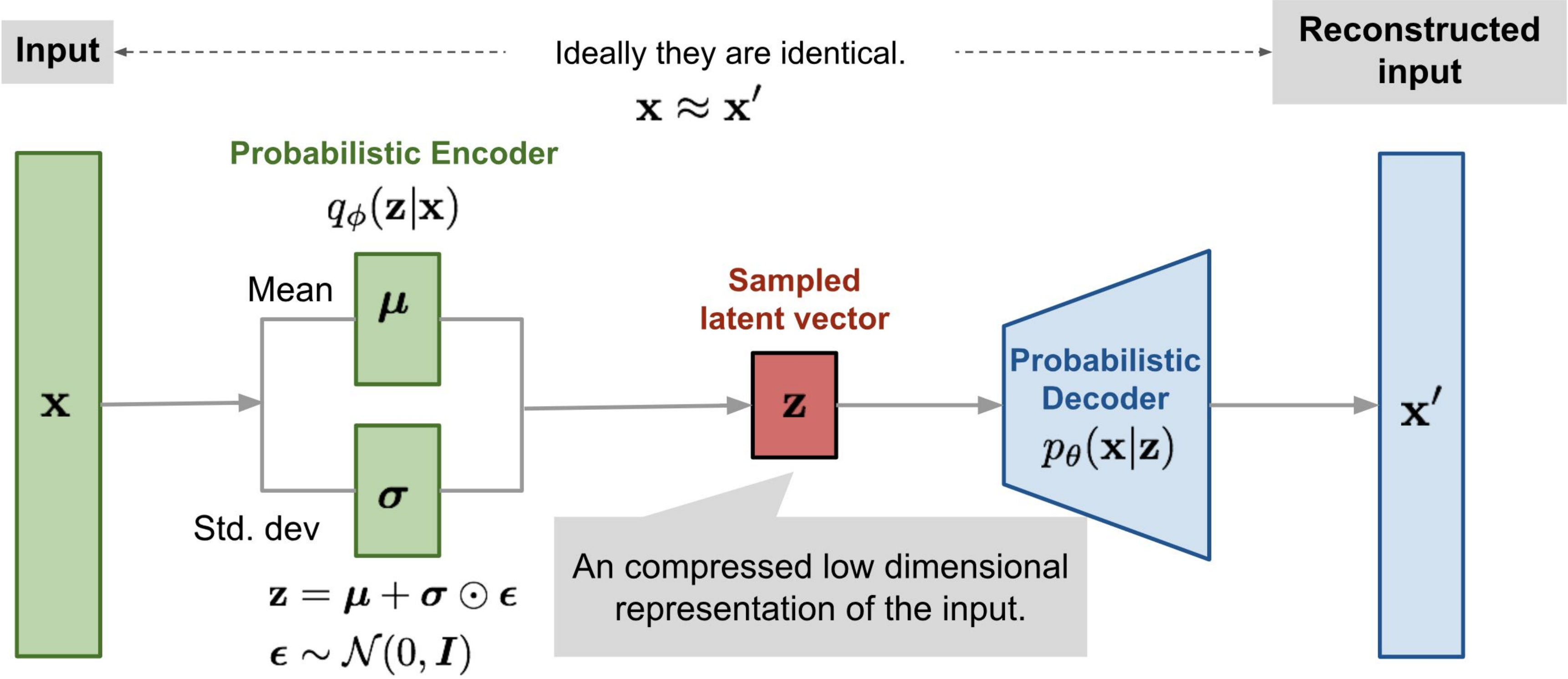
Autoencoders Applications: Reconstruction



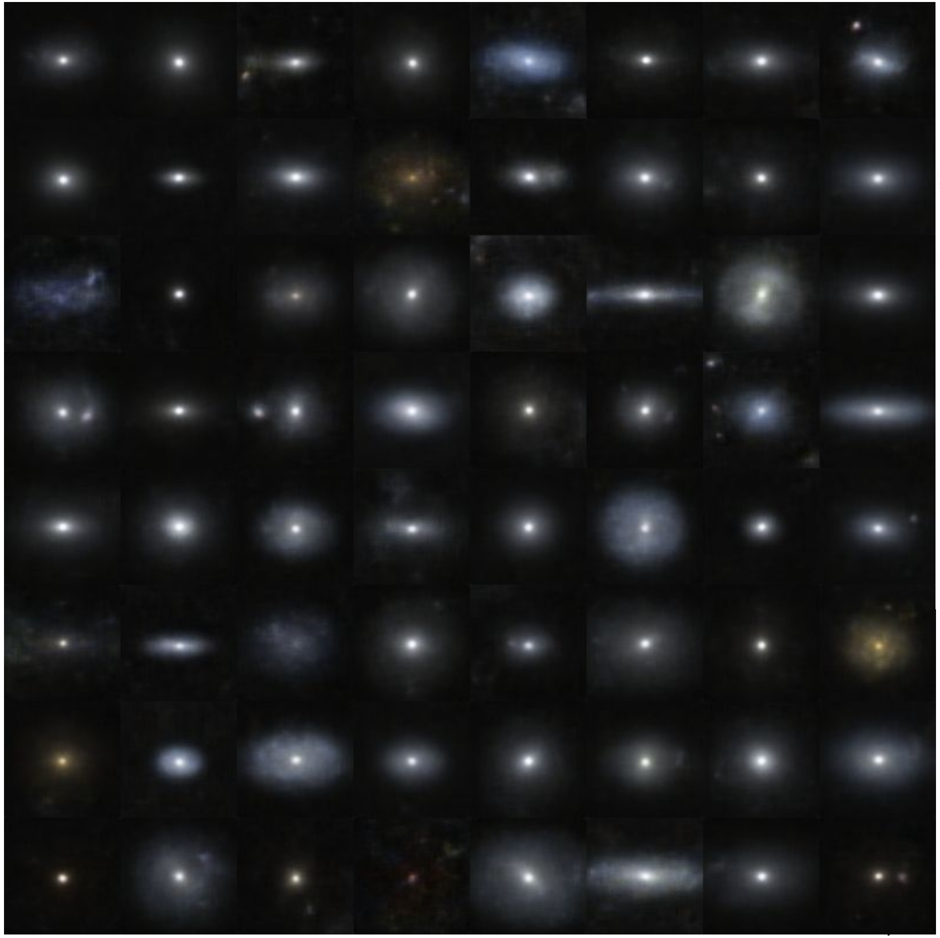
Blurry images and structure is lost, but angular sizes, radial profiles and brightness are a match.

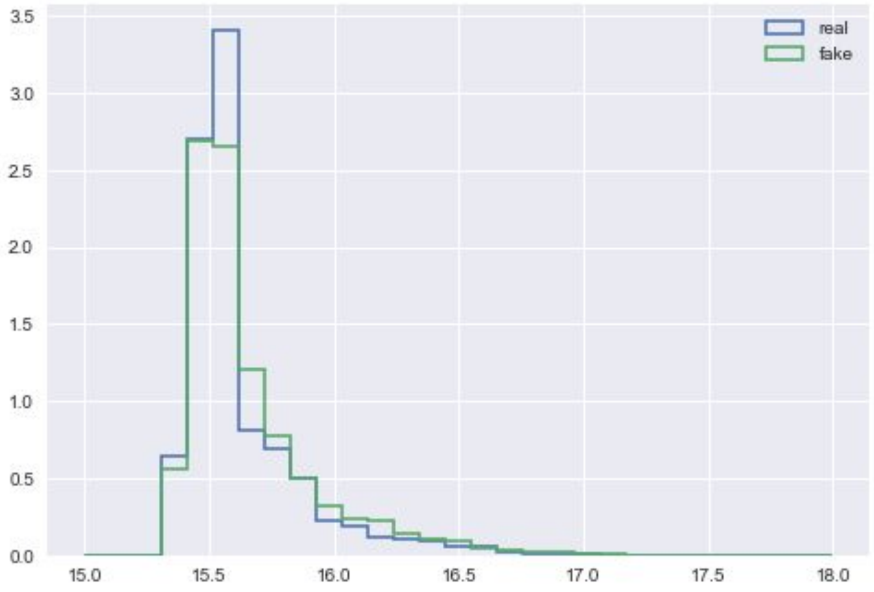
What if we can make the model learn properties at the same time as images.? What if can sample from the latent space?

Summarizing VAE basics, more complex models built on top



VAE Sampling example (no reconstruction)





We can generate samples from z, next step is can we constrain what's being sampled?

CMU

Multimodal VAE: Training modalities

Multimodal Generative Models for Scalable Weakly-Supervised Learning

Mike Wu
Department of Computer Science
Stanford University
Stanford, CA 94025
wumike@stanford.edu

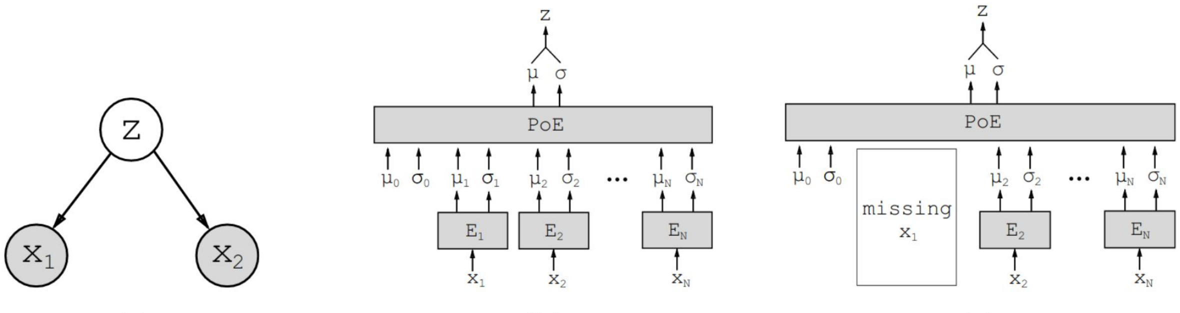
Noah Goodman
Departments of Computer Science and Psychology
Stanford University
Stanford, CA 94025
ngoodman@stanford.edu

[multimodal-generative-models-for-scalable-weakly-supervised-learning]()

Learning joint representation of conditionally independent modalities using product of experts.

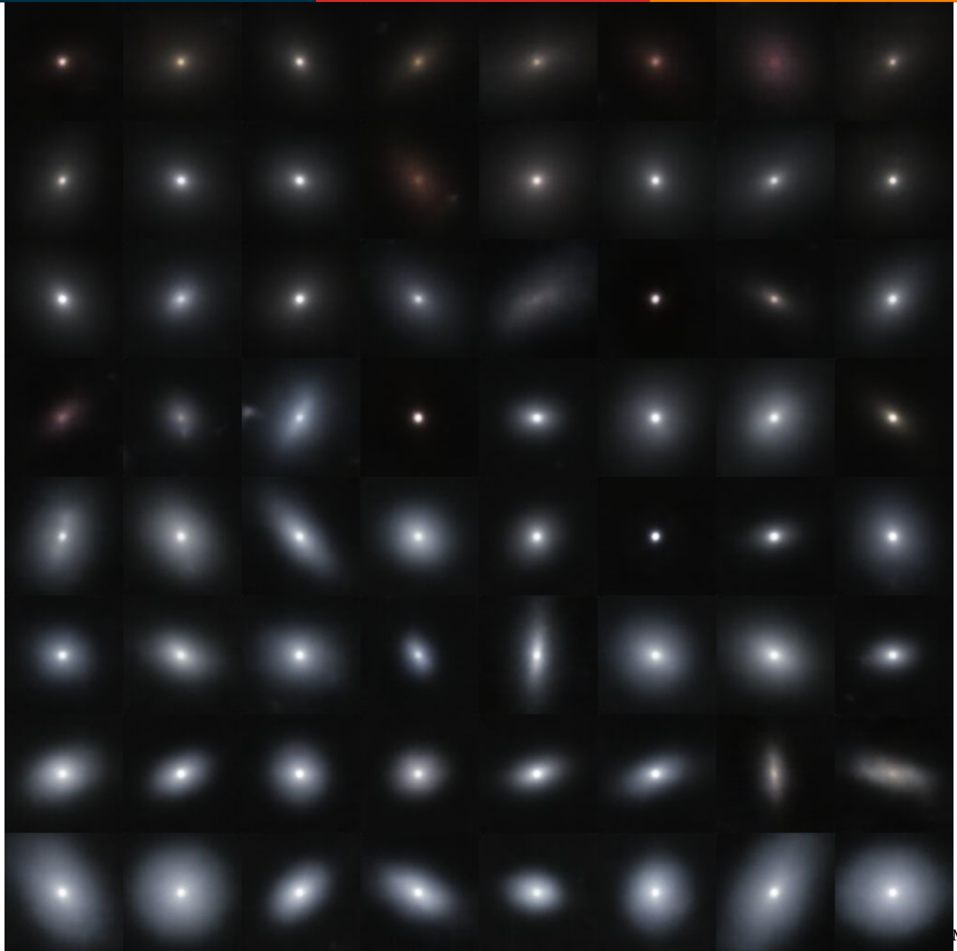
We can:

- Conditional sample with certain attributes
- Sample without any limitations
- Change the attribute of an existing input data
- Similarity search and anomaly detection
- Predict one modality from the others
- Sample and train with missing modalities

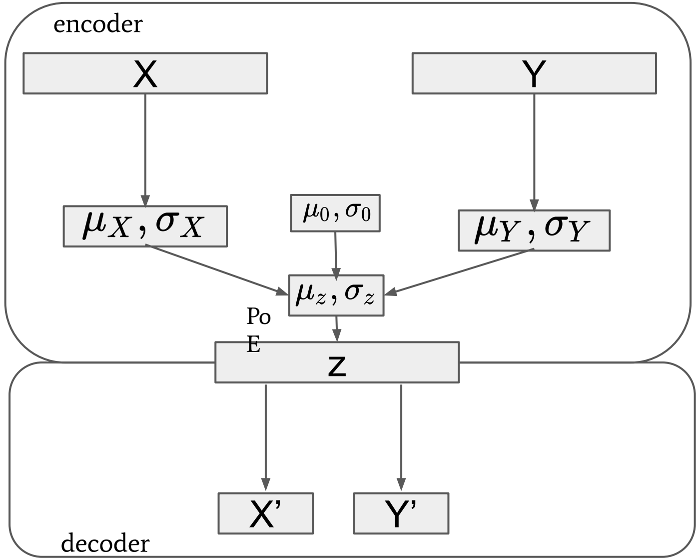


(a) (b) (c)

MVAE: Examples



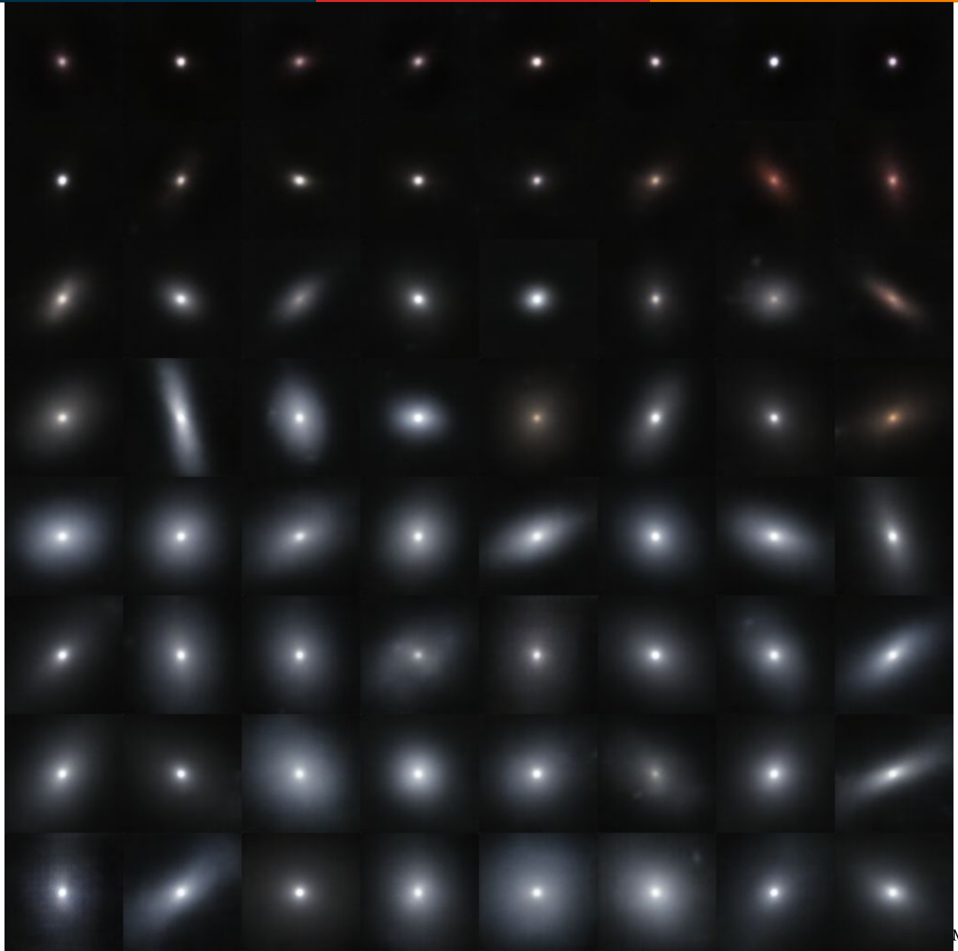
Samples with changing brightness (increasing downwards)



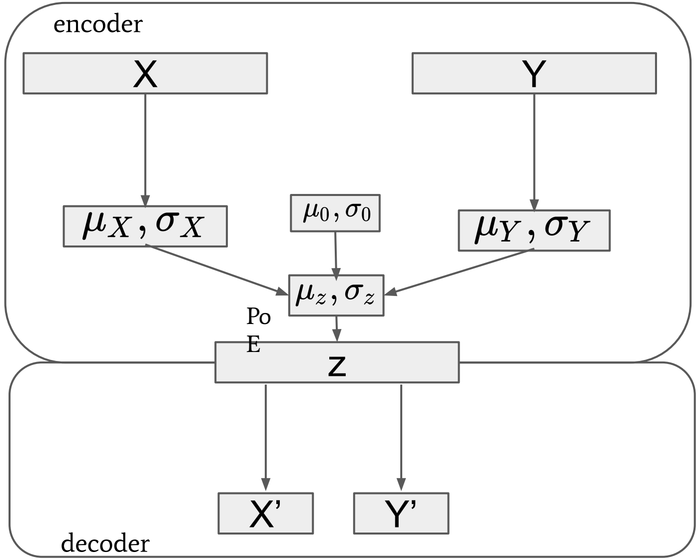
Bowen, Carrasco-Kind, et al. in prep.

MU

MVAE: Examples



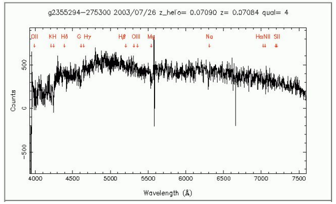
Samples with changing area
(increasing downwards)

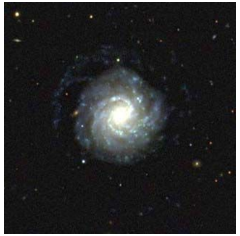


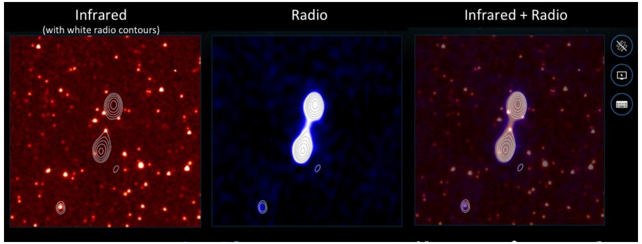
MU

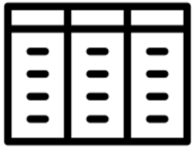
Bowen, Carrasco-Kind, et al. in prep.

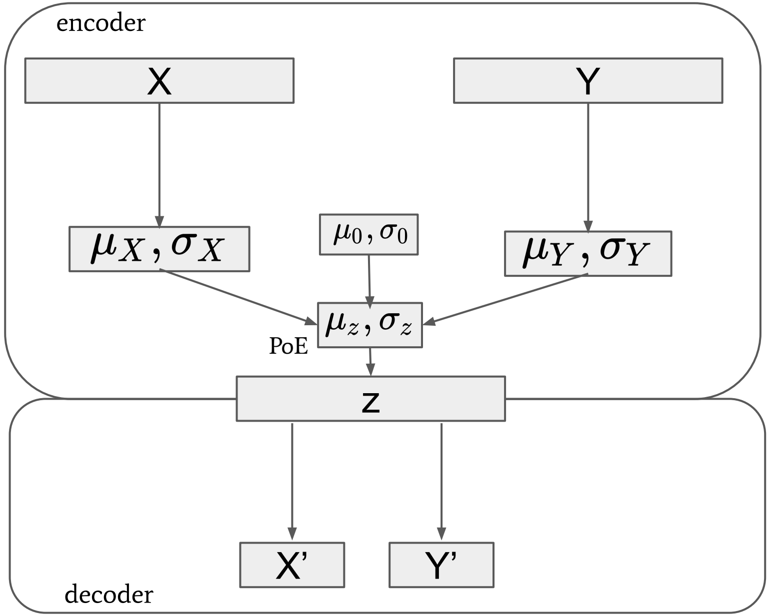
MVAE: Opens very interesting options





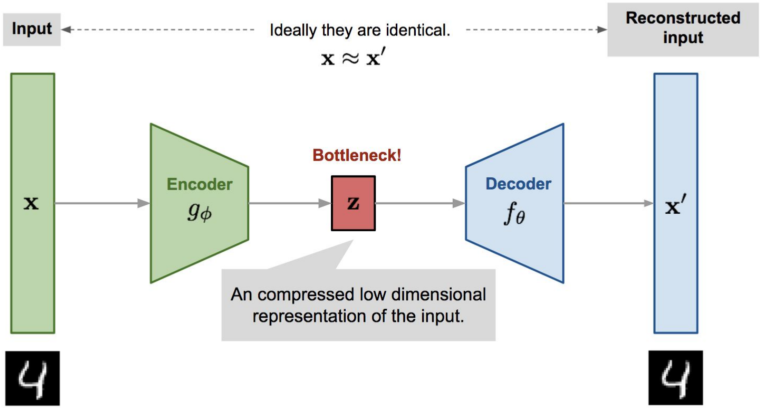






Bowen, Carrasco-Kind, et al. in prep.

Can we sample z to generate fake data?



Exist θ for max the likelihood

$$p(x) = \int p(x|z,\theta)p(z)dz$$

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

Too expensive

We need an approximate posterior (prob. encoder)

$$q_\lambda(z|x) \approx p(z|x)$$

And we can use q to be Gaussian (there are other alternatives)

$$q_\lambda(z|x) = \mathcal{N}(z; \mu_\lambda(x), \sigma_\lambda(x))$$

$$p(z) = \mathcal{N}(0, I)$$

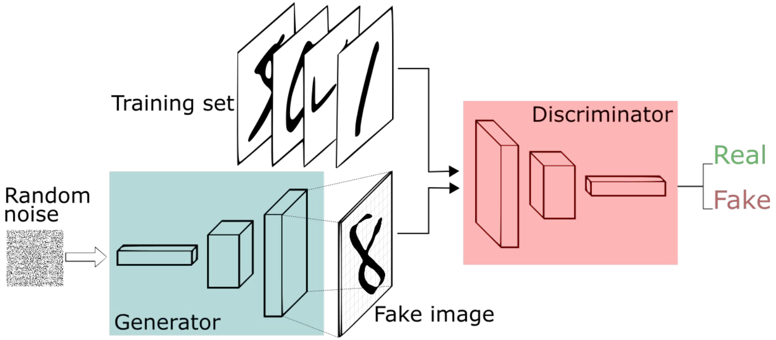
- Map x to a distribution $p(z|x)$
- Sample from distribution $z_i \sim p(z)$
- Generate fake data $x'_i \sim p_\theta(x'|z)$
- Probabilistic approach

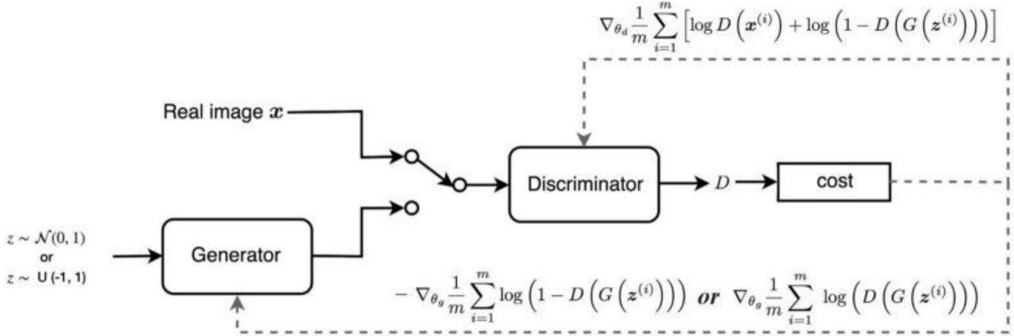
$$p(x,z) = p(x|z)p(z)$$

Variational "Autoencoder"

Generative Adversarial Networks (GAN)

Gan can be VERY good to specific image generation and create realistic images. Very powerful discriminator





But:

- Very hard to train (unstable)
- Not really sampling methods
- Hard to evaluate likelihood of data p(x)
- Tend to underfit data distribution
- Main goal is to fool the discriminator

Very powerful if combined with VAE