# Overlooked Implications of the Reconstruction Loss for VAE Disentanglement

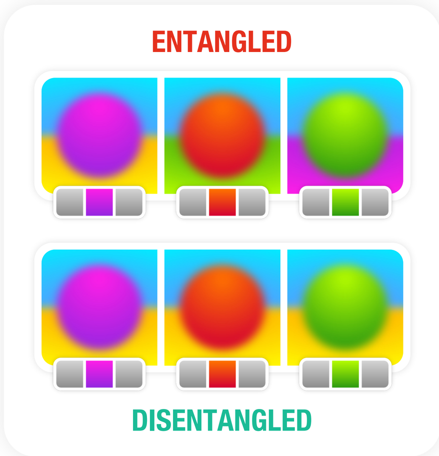
**Nathan Michlo[1,2], Richard Klein[1,2], Steven James[2]**

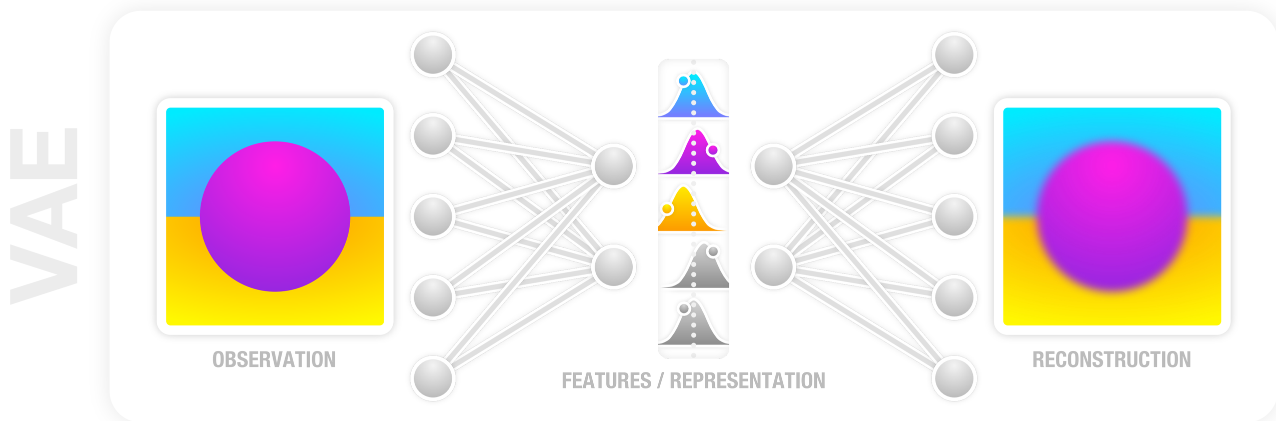
[all] University of the Witwatersrand, Johannesburg, South Africa [1] Prime Lab, Johannesburg, South Africa [2] Rail Lab, Johannesburg, South Africa

# We show how data & recon. loss lead to VAE disentanglement

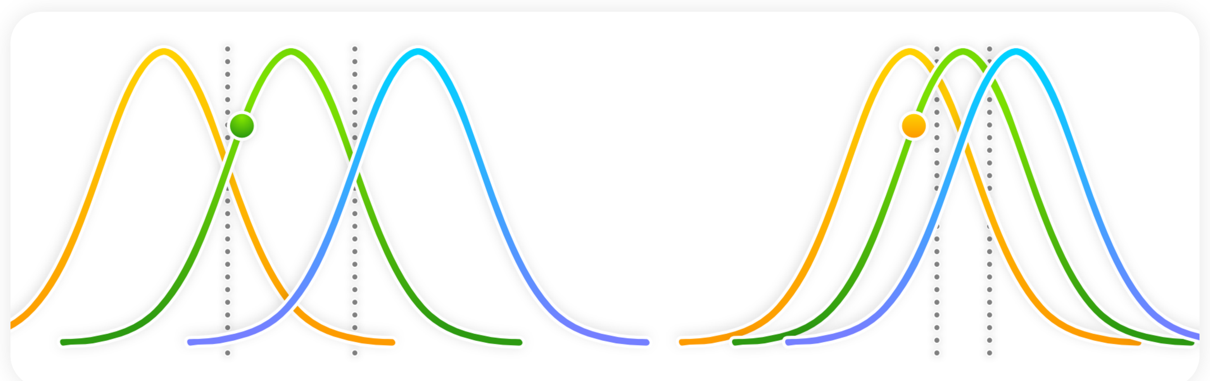
## Introduction

- Variational Auto-Encoders (VAEs) often learn **disentangled** representations of data.
- Disentanglement is evaluated using synthetic datasets with **subjective** ground-truth factors.




- Random sampling causes reconstruction mistakes. VAEs place similar observations close together in embedding space to minimise this error.

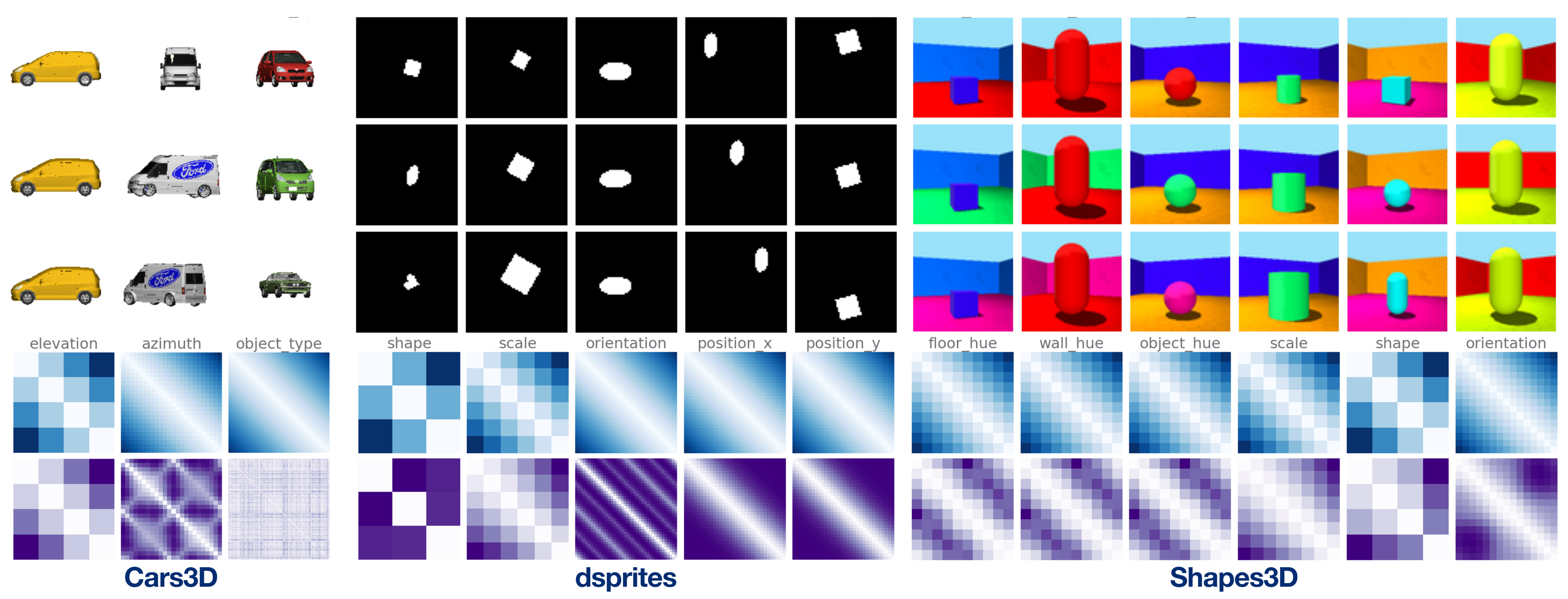


## Characterising Existing Datasets

- VAEs perceive distances between observations along ground-truth factors using their chosen reconstruction loss.

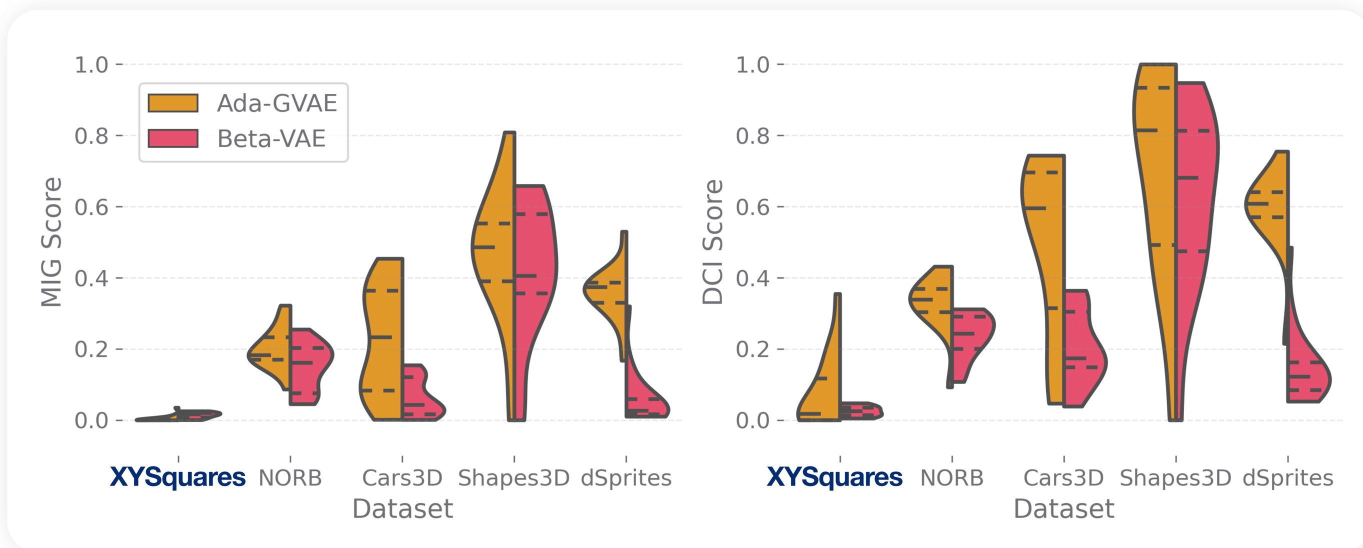
$$d_{pcv}(\bullet, \bullet) = \lim_{\hat{\boldsymbol{x}}\to\boldsymbol{x}} \mathcal{L}_{rec}(\bullet, \bullet)$$

- Factors which disentangle easily, usually have high correlation with this.

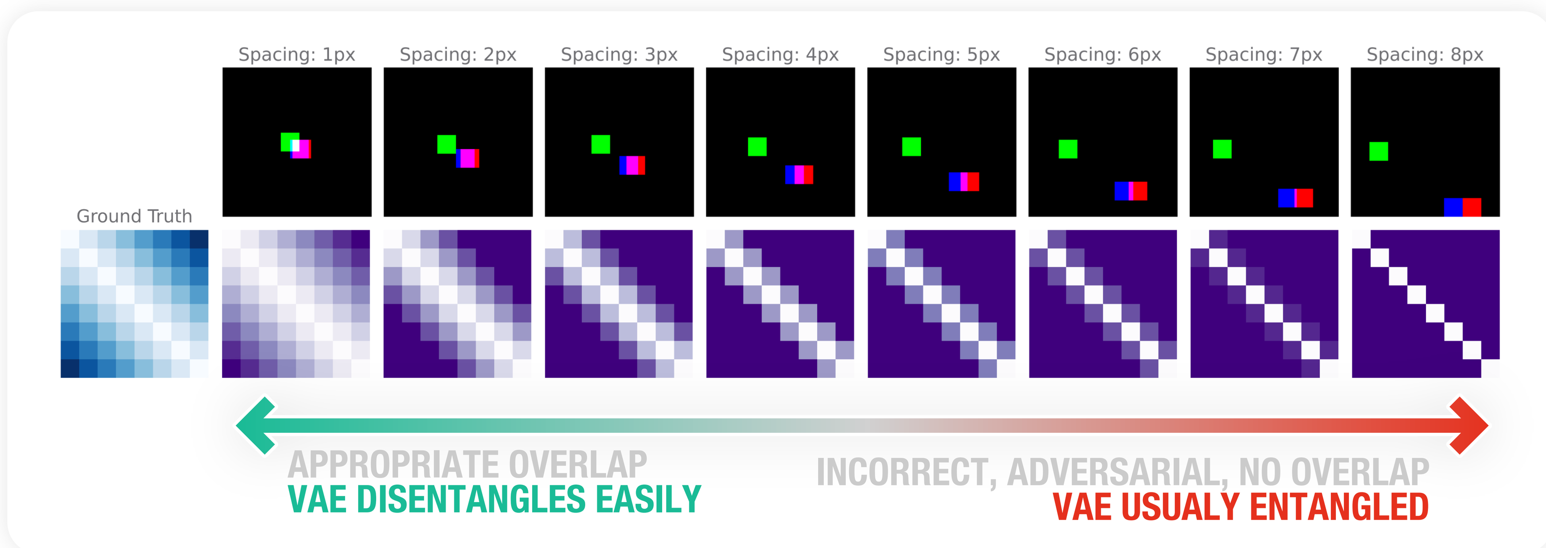


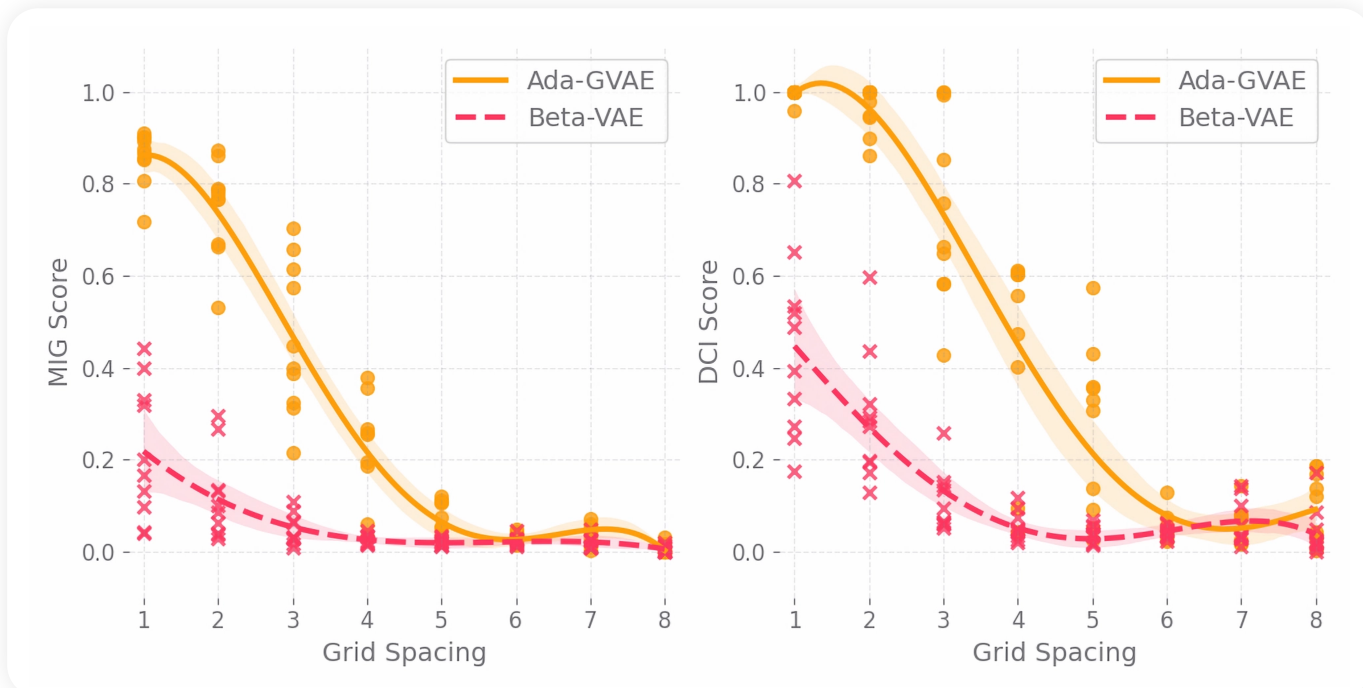

## Adversarial Data Example

- Datasets are considered **adversarial** when distances between datapoints along ground-truth factors are constant. No order can be found, and no latent space re-organisation takes place.
- We design a simple 8x8 gridworld domain called **XYSquares** with adjustable spacing of x and y ground-truth factors to test this.
- Disentanglement performance in the adversarial case is low.



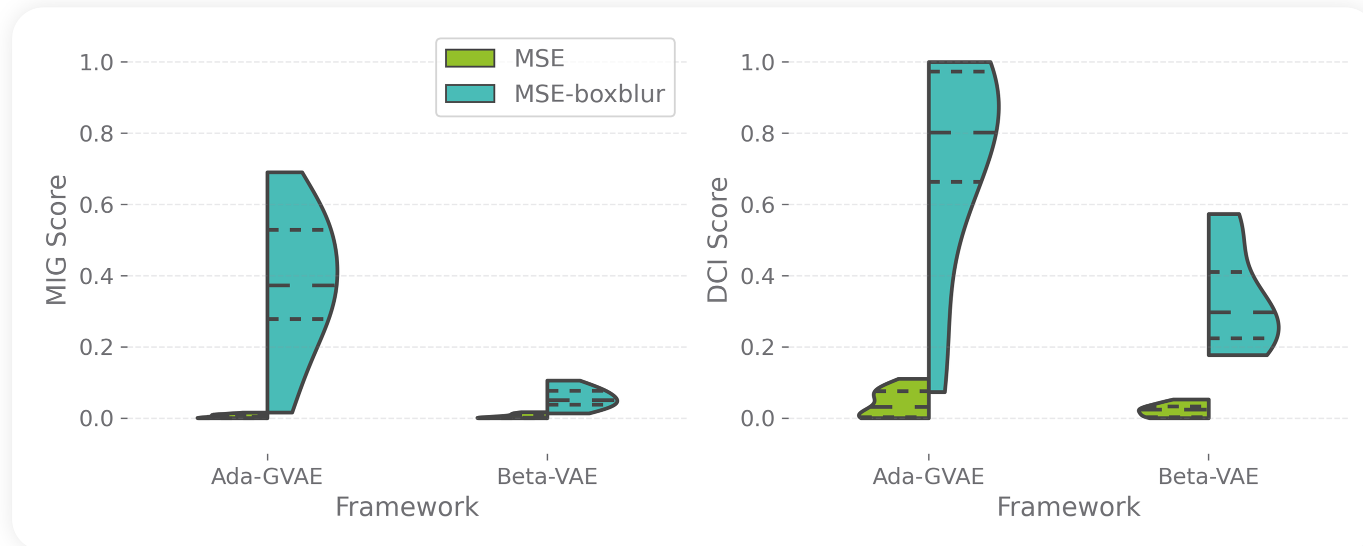

- We train VAEs over XYSquares datasets with varying spacing. More overlap in the data gives better disentanglement.




## Re-enable Disentanglement Example

- If perceived distances once again correspond to ground-truth distances, disentanglement takes place.
- We adjust the reconstruction loss to re-enable disentanglement. An example that is appropriate for XYSquares is a **box blur** augment.




## Considerations For Disentanglement

- In practise there are infinitely many datasets with **infinitely** many **choices** of what constitutes their ground-truth factors.
- Disentanglement choices are **subjective**, e.g. RGB, HSV or categorical representations for colours, binary or continuous encodings for positions, split or combined factors.
- Benchmark datasets, metrics and **literature largely ignore this**.
- Disentanglement is ultimately not from special algorithmic choices.

## Conclusion

- Fundamental characteristics of existing datasets encourage VAEs to learn disentangled representations.
- The focus on regularisation for disentanglement is misplaced, rather, **disentanglement is largely accidental**, and careful choice of the reconstruction loss or data is needed.

**SCAN FOR PAPER & RESOURCES**

or visit github.com/nmichlo