

Data Overlap: A Prerequisite for Disentanglement

And Metric Learning Solutions

Nathan Juraj Michlo

29 October 2021



&



My Research

Overview Of This Presentation

1. Disentanglement

- **Learning disentangled representations** with Variational Auto-Encoders (VAEs) is often attributed to the regularisation term.

2. Overlap In Data

- In this work we highlight the **interaction between data and the reconstruction loss** as a significant contributor to disentanglement in VAEs.

3. Adversarial Dataset

- We design an intuitive **adversarial dataset** that exploits this mechanism to break a large set of disentanglement frameworks.

4. Disentanglement As Metric Learning

- Finally we develop solutions to this problem in the form of **axis-aligned supervised and unsupervised metric learning** approaches.

1. Background: **Representations**

Representation Learning

Definition

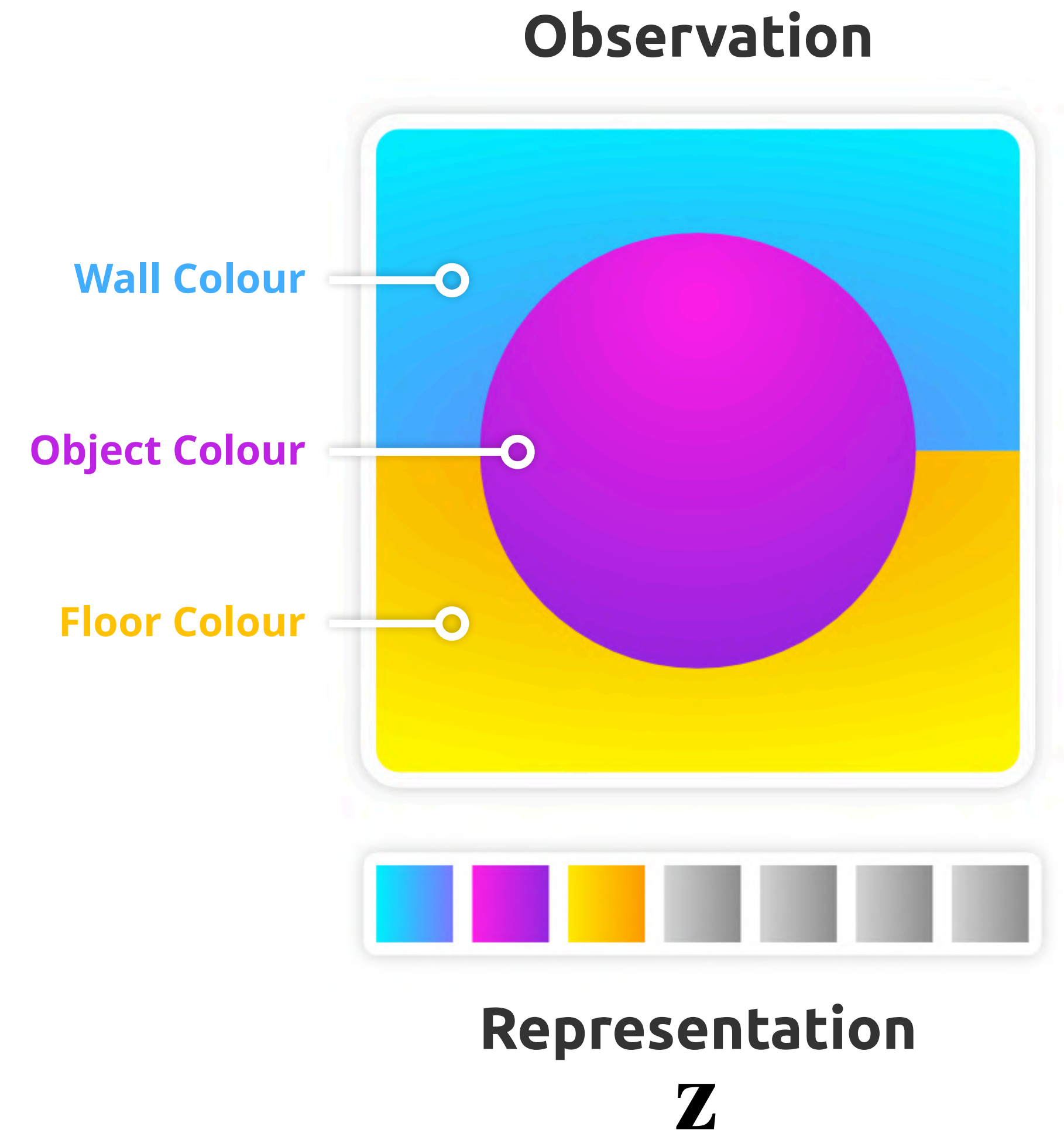
“Learning representations of data that make it easier to extract useful information when building classifiers or other predictors”^[1]

Example:

- Dataset has with **floor**, **object** and **wall** colours, that are the ground-truth factors.
- Learn **low dimensional vectors** that still capture these factors.

Q: How do we learn these representations?

- ~~Unsupervised~~^[2] or Weakly-Supervised.
- PCA, ICA, VAEs, GANs, Ranking Loss, etc...



[1] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

[2] Locatello, Francesco, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." *arXiv preprint arXiv:1811.12359* (2018).

Disentangled Representations

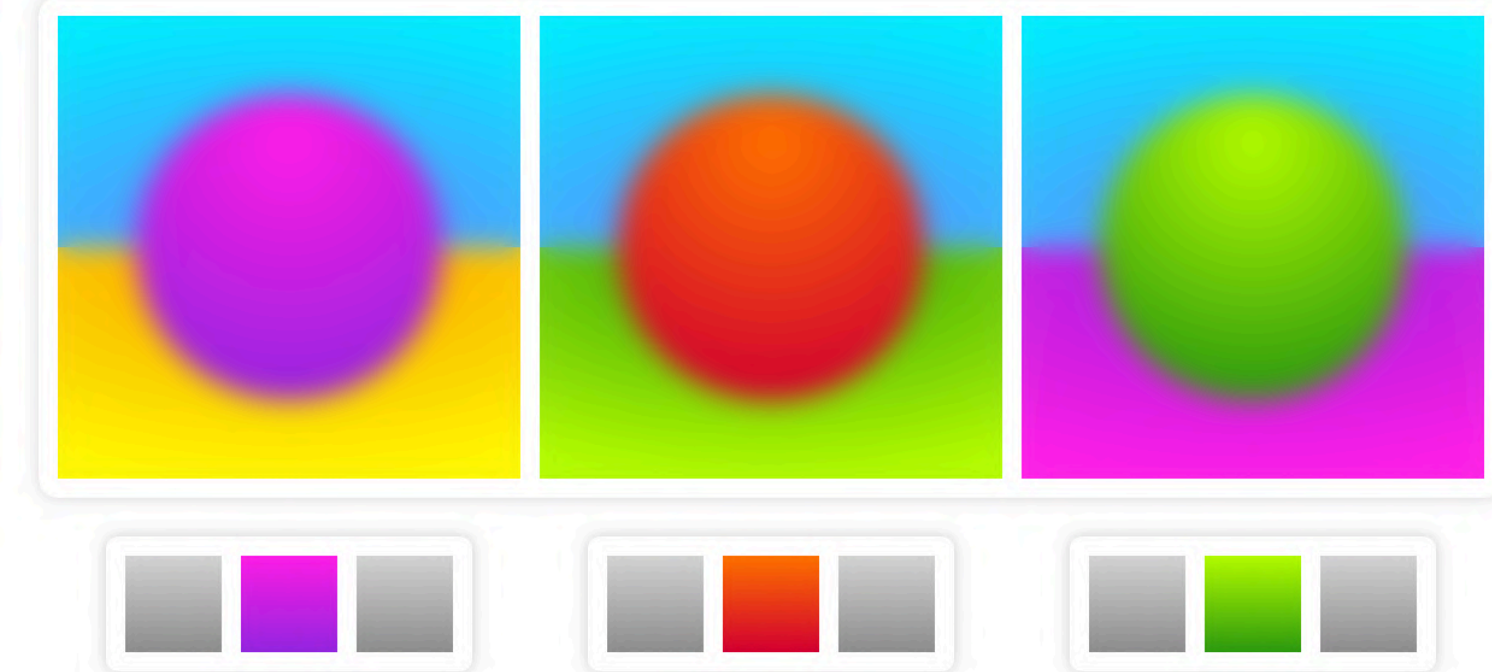
What Makes a Good Representation? ^[1]

- **Disentangled**

- Representations **align with ground-truth** factors
- **Sparse**: Only a subset of factors are relevant
- **Shared**: Differences attributed to a subset of factors
- Hierarchical components
- Interpretable & Fair

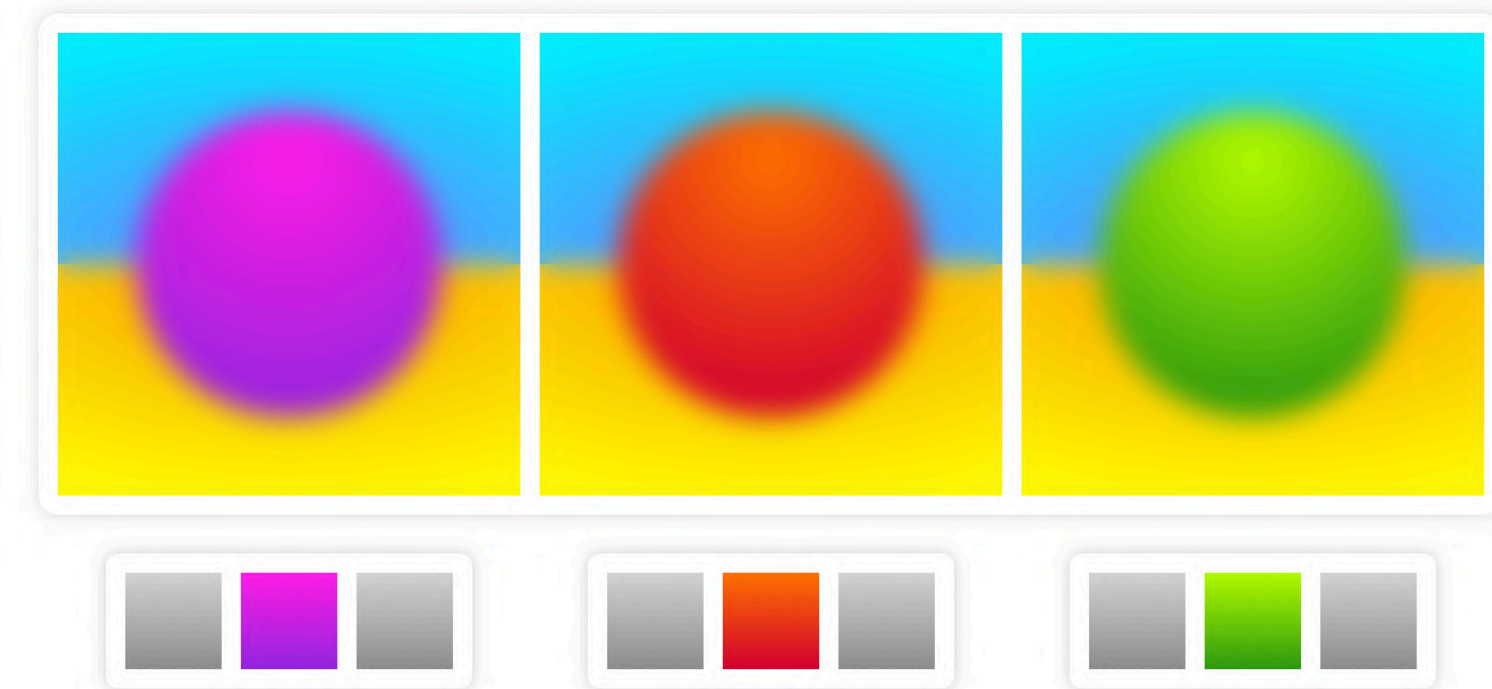
- **Other Properties**

- **Smooth**, without discontinuities
- Temporal and spatial coherence
- Natural clustering



Entangled Representations

Unintuitive - Latent variables affect multiple factors.



Disentangled Representations

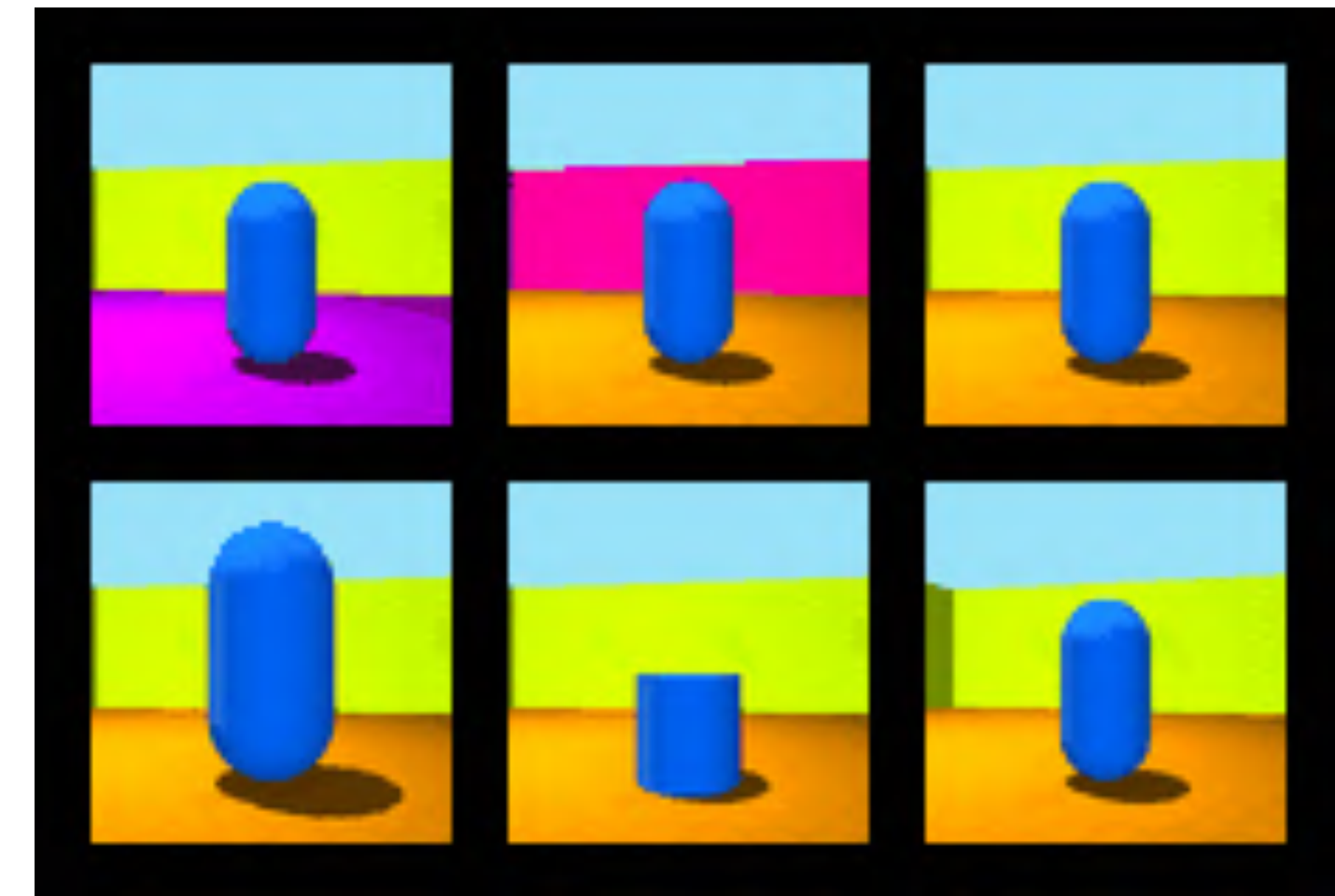
Axis-aligned - Latent variables affect distinct factors.

[1] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

How Do We Evaluate Representations?

Use Synthetic Datasets With Known Factors

- Use datasets with **known ground-truth factors**
 - 3DShapes, DSprites, MPI3D, Cars3D, Small NORB
- Using disentanglement metrics
 - **Compare learnt representations** and the **ground-truth factors** of each observation
 - β -VAE Score, FactorVAE Score, DCI, MIG, SAP Score, Modularity
 - Each score has assumptions & biases, use multiple



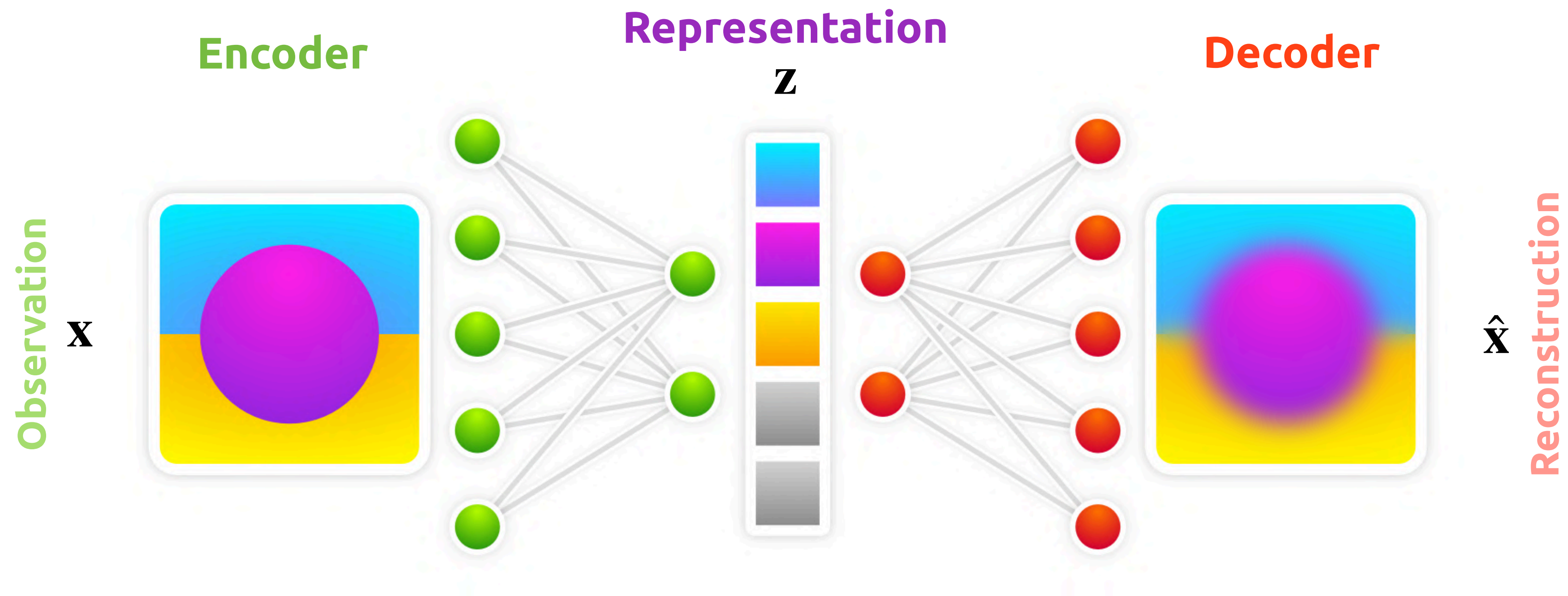
3DShapes
Ground-Truth Factors



DSprites
Ground-Truth Factors

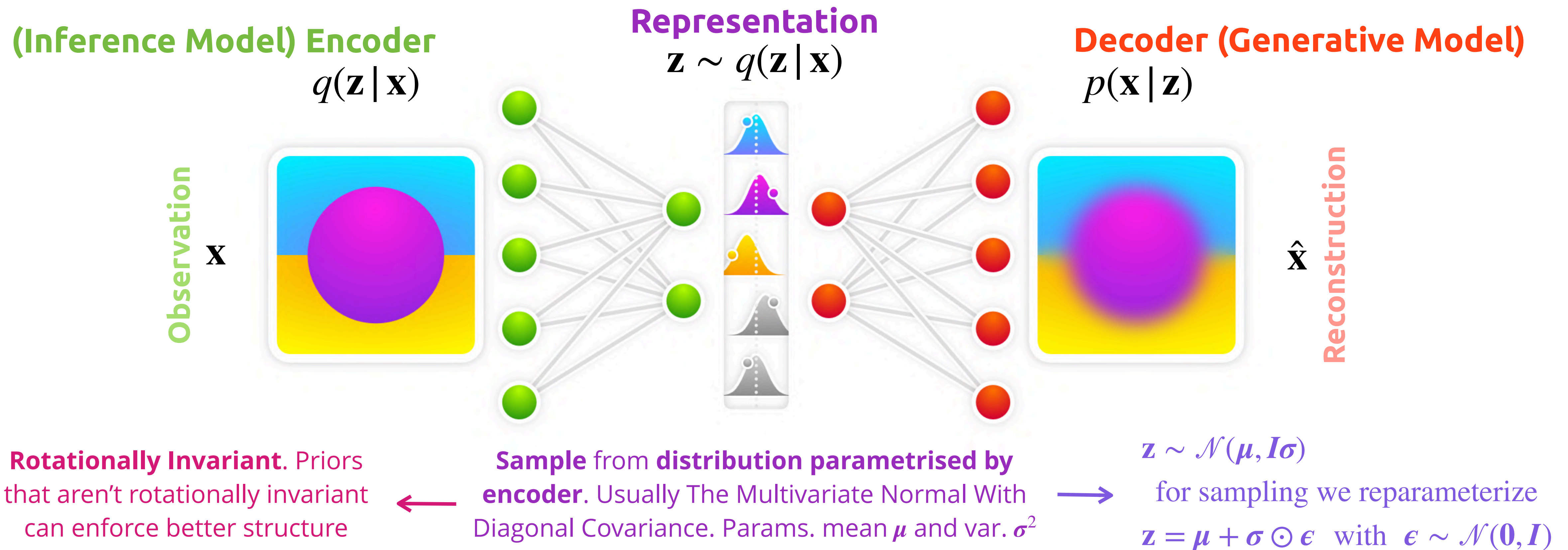
Auto-Encoders

Representation Learning With Auto-Encoders



Variational Auto-Encoders

Representation Learning With Variational Auto-Encoders (VAEs) ^[1]



[1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

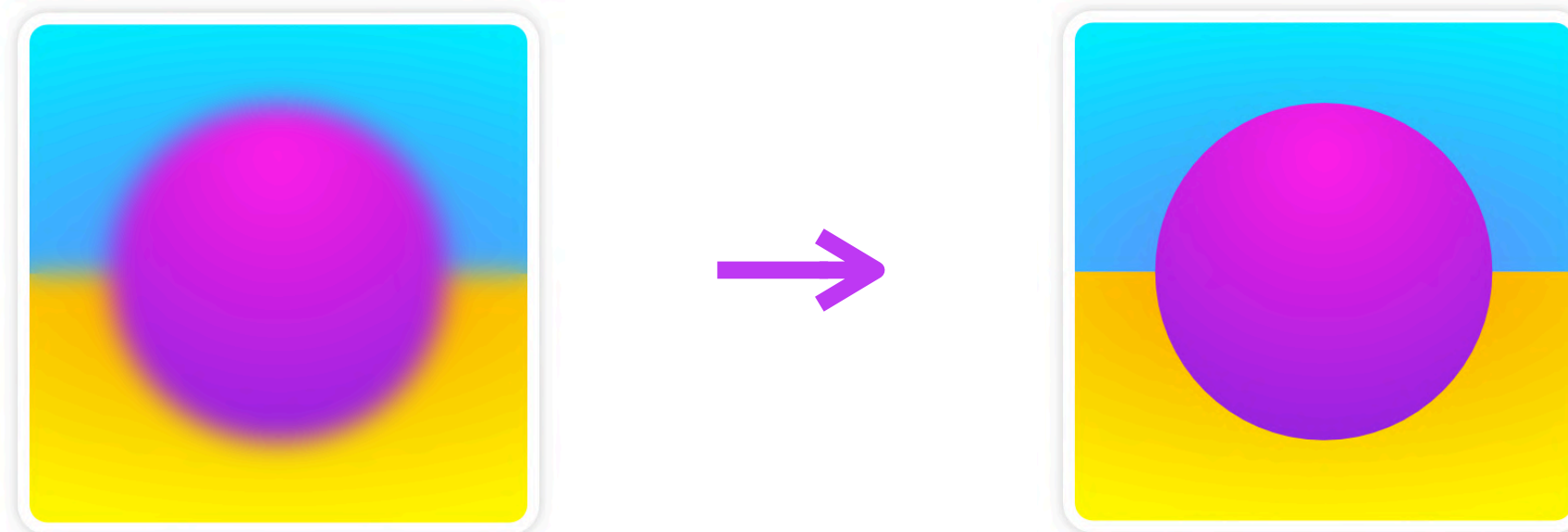
Variational Auto-Encoders

β -VAE Objective ^[1]

$$L_{\beta\text{VAE}} = L_{\text{rec}} + \beta L_{\text{reg}}$$

Coefficient adjusts strength of regularisation. ^[1] Higher values encourage better disentanglement but worse reconstruction quality. ^[2]

Reconstruction Term (Decoder)



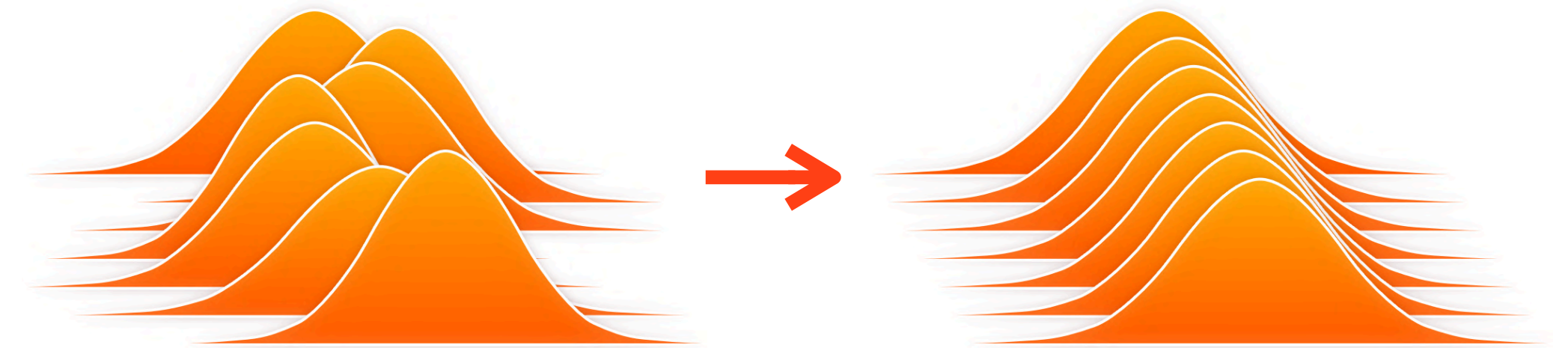
Reconstructions

Improves the **decoder** by encouraging reconstructions to match observations.

Observations

Stronger regularisation, sends less information to the decoder, resulting in **more blurred** reconstructions.

Regularising Term (Encoder)



Approx. Posterior

Improves the **encoder** by encouraging the approximate posterior to match the prior by minimising the KL Divergence.

Assumed Prior

The prior is usually assumed to be **standard normal distributions**, with a mean of 0 and variance of 1.

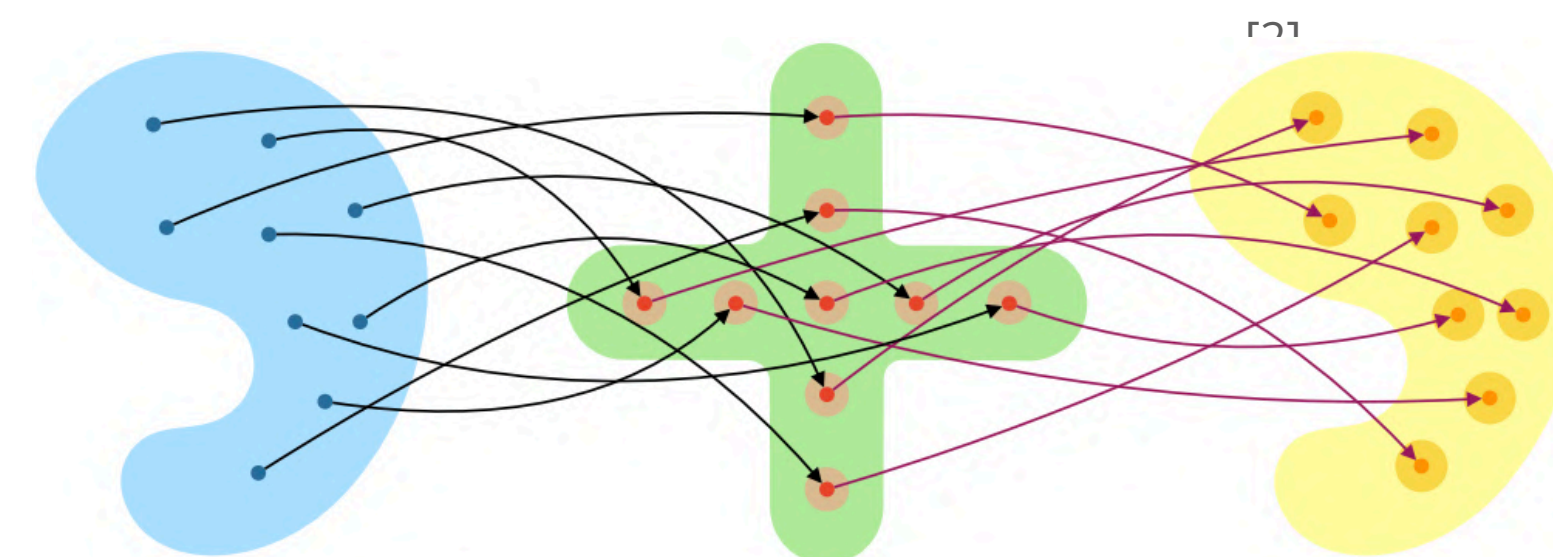
[1] Higgins, Irina, et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." ICLR 2.5 (2017): 6.

[2] Burgess, Christopher P., et al. "Understanding disentangling in β -VAE." arXiv preprint arXiv:1804.03599 (2018).

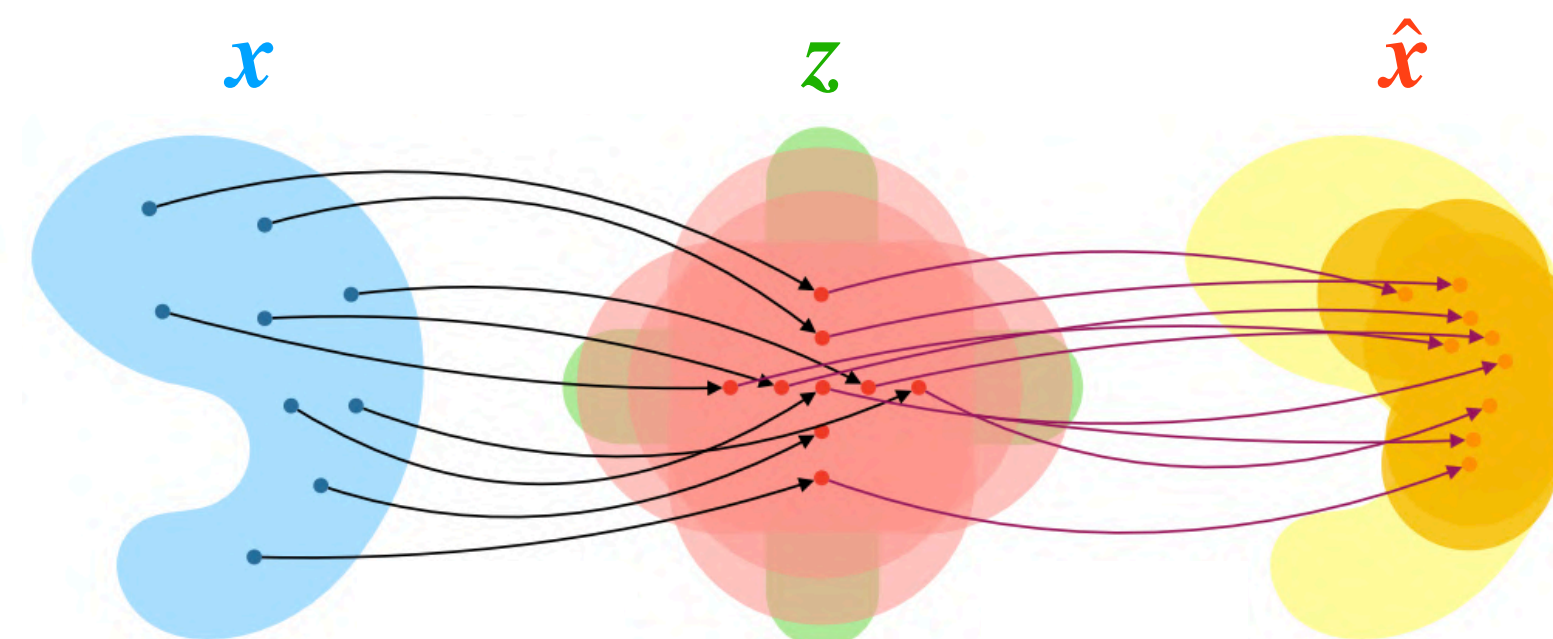
Disentanglement In Beta-VAEs

β -VAEs Control Latent Overlap^[2] And Sampling Errors^[1]

- **Regulariser encourages** the latent distributions to **overlap**
 - Good for **smoothness & disentanglement**
 - Decoder can confuse sampled representations
- **Disentanglement?**
 - **Reorders latent space** to minimise reconstruction error because of sampling^[1]
 - **Similar observations** in terms of loss **are placed together** in the latent space.



Weak Beta: $\beta \rightarrow 0$
Acts like lookup table + Less sampling errors + Worse disentanglement



Strong Beta: $\beta \rightarrow \infty$
Carries little information + More sampling errors + Better Disentanglement

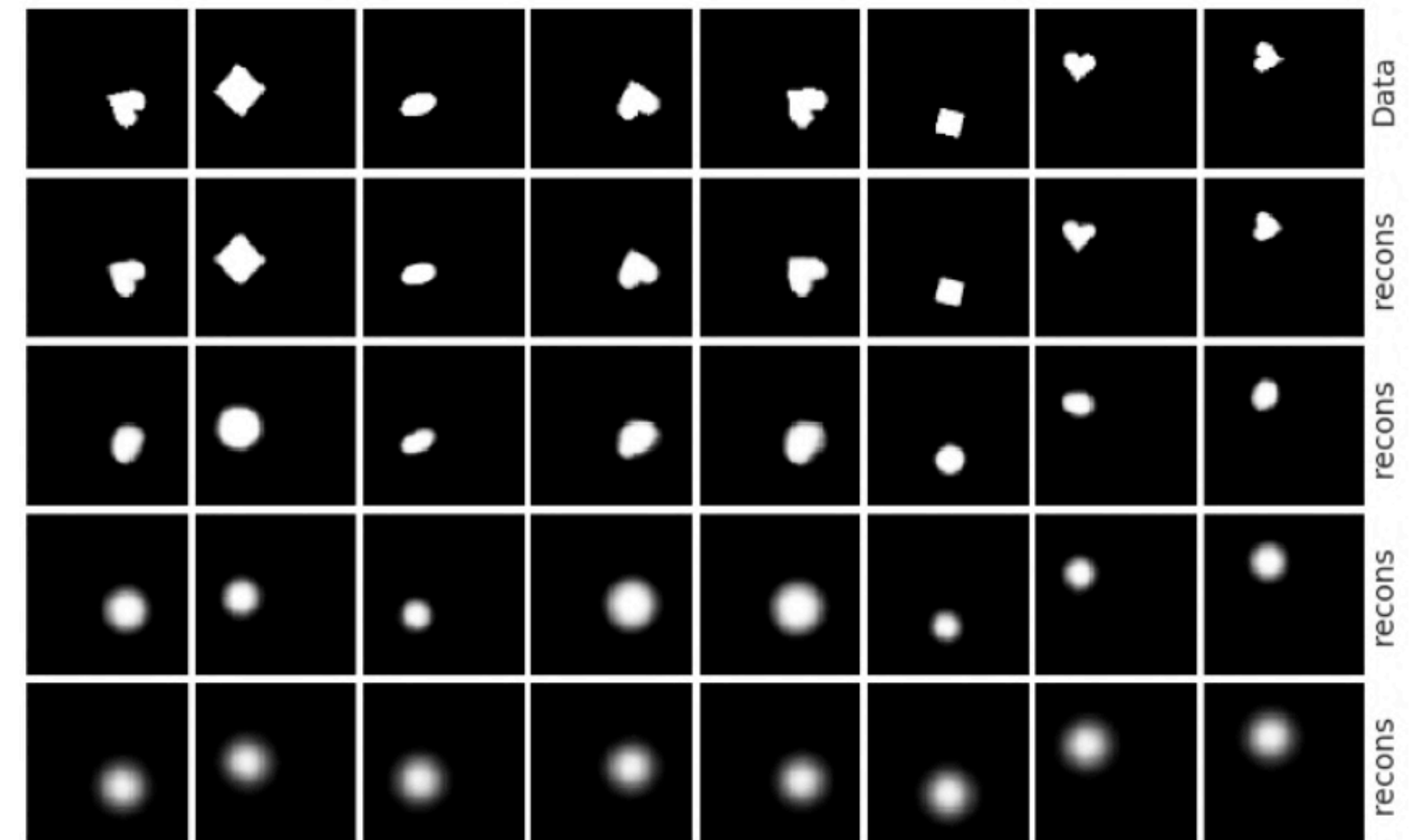
[1] Burgess, Christopher P., et al. "Understanding disentangling in β -VAE." arXiv preprint arXiv:1804.03599 (2018).

[2] Mathieu, Emile, Tom Rainforth, N. Siddharth, and Yee Whye Teh. 2018. "Disentangling Disentanglement In Variational Autoencoders". Arxiv.Org. <https://arxiv.org/abs/1812.02833>.

Information Bottleneck

β -VAEs From The Information Bottleneck Perspective ^[1]

- Information Bottleneck
 - **Tradeoff** between accuracy and compression.
 - As β is increased, **less information can be encoded** by the representation
- Since this bottleneck exists
 - The **factors** that minimise the reconstruction loss are prioritised.
 - Sampling **errors reorder latent space**



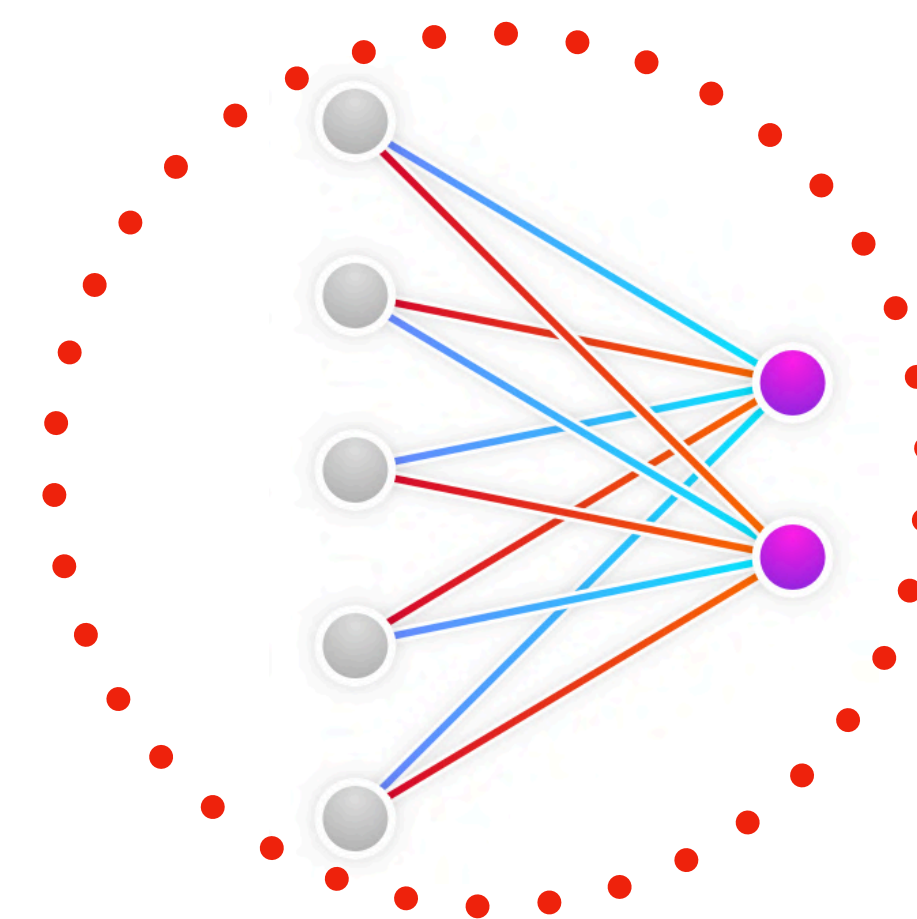
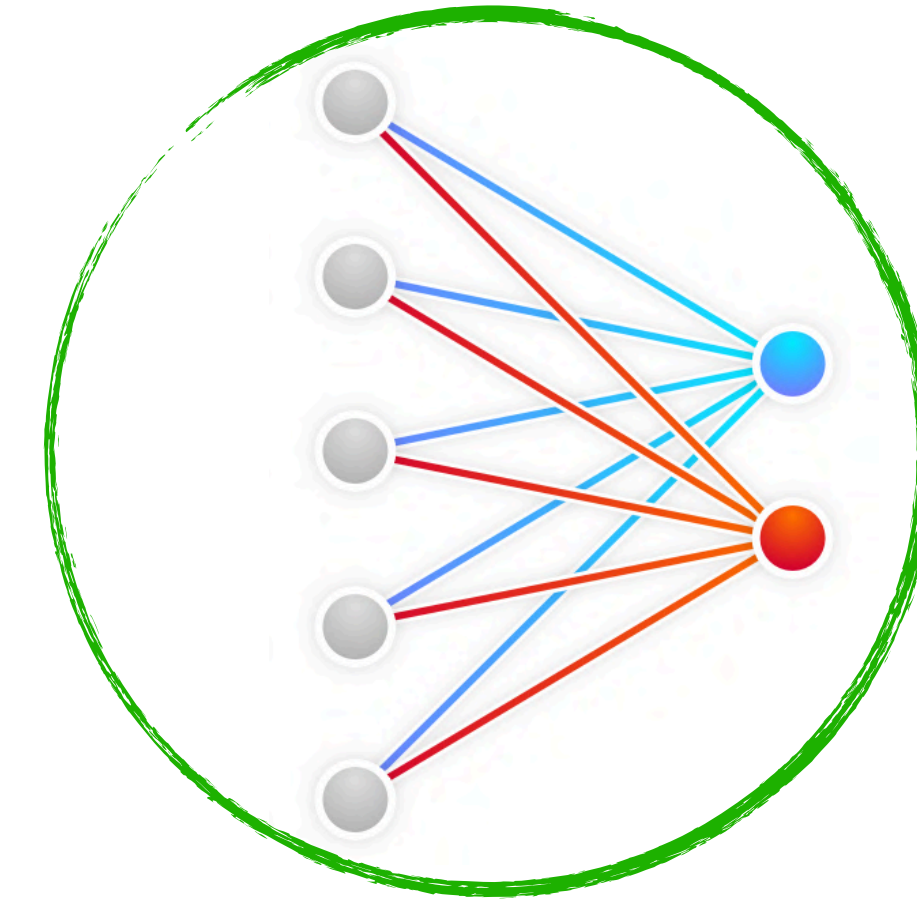
Q: Where does the axis-alignment come from?
Bottleneck doesn't explain this!

[1] Burgess, Christopher P., Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. "Understanding Disentangling In β -VAE". Arxiv.Org. <https://arxiv.org/abs/1804.03599>.

VAEs In Practice - Impossibility Result

Non-Identifiability of Unsupervised Models ^[1]

- Fundamentally **impossible** to reliably learn disentangled representations **without inductive biases**.
 - Performance of **unsupervised** methods largely **depends on** the initial **random seeds**.
 - We can use **model selection**, but this is impractical and still **requires** knowledge of the domain and **supervision**.



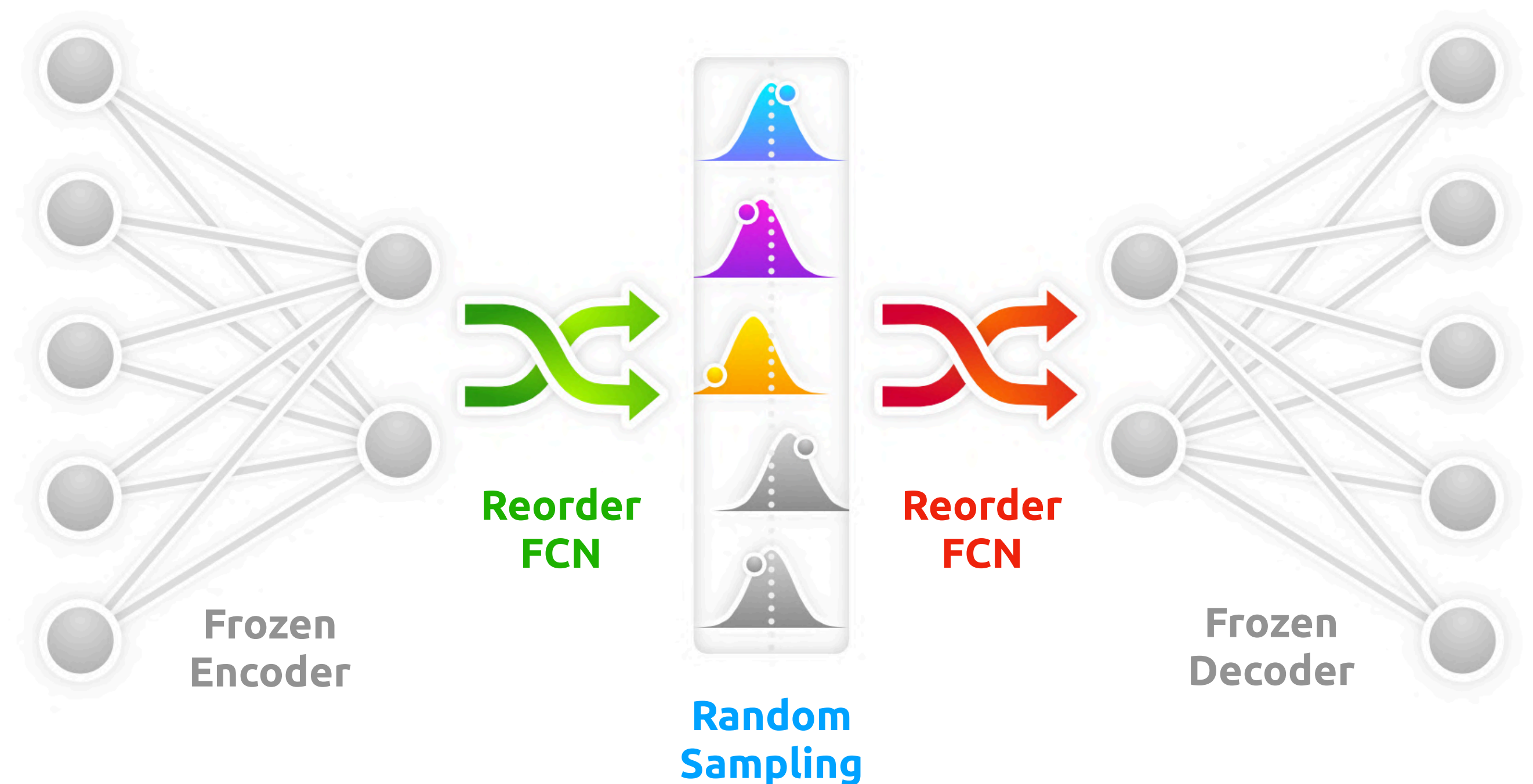
[1] Locatello, Francesco, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." arXiv preprint arXiv:1811.12359 (2018).

2. Overlap In Existing Datasets

Minimise Sampling Error To Disentangle

A Simplified Example Of How VAEs Disentangle

1. **Consider** pre-trained **auto-encoder** with **perfect reconstructions**
2. **Freeze** auto-encoder **weights**
3. **Convert into VAE** by adding fully connected layers
4. **Train as VAE**
 - Training **reorders** the **latent space** to **minimise** reconstruction loss **error** due to **random sampling**

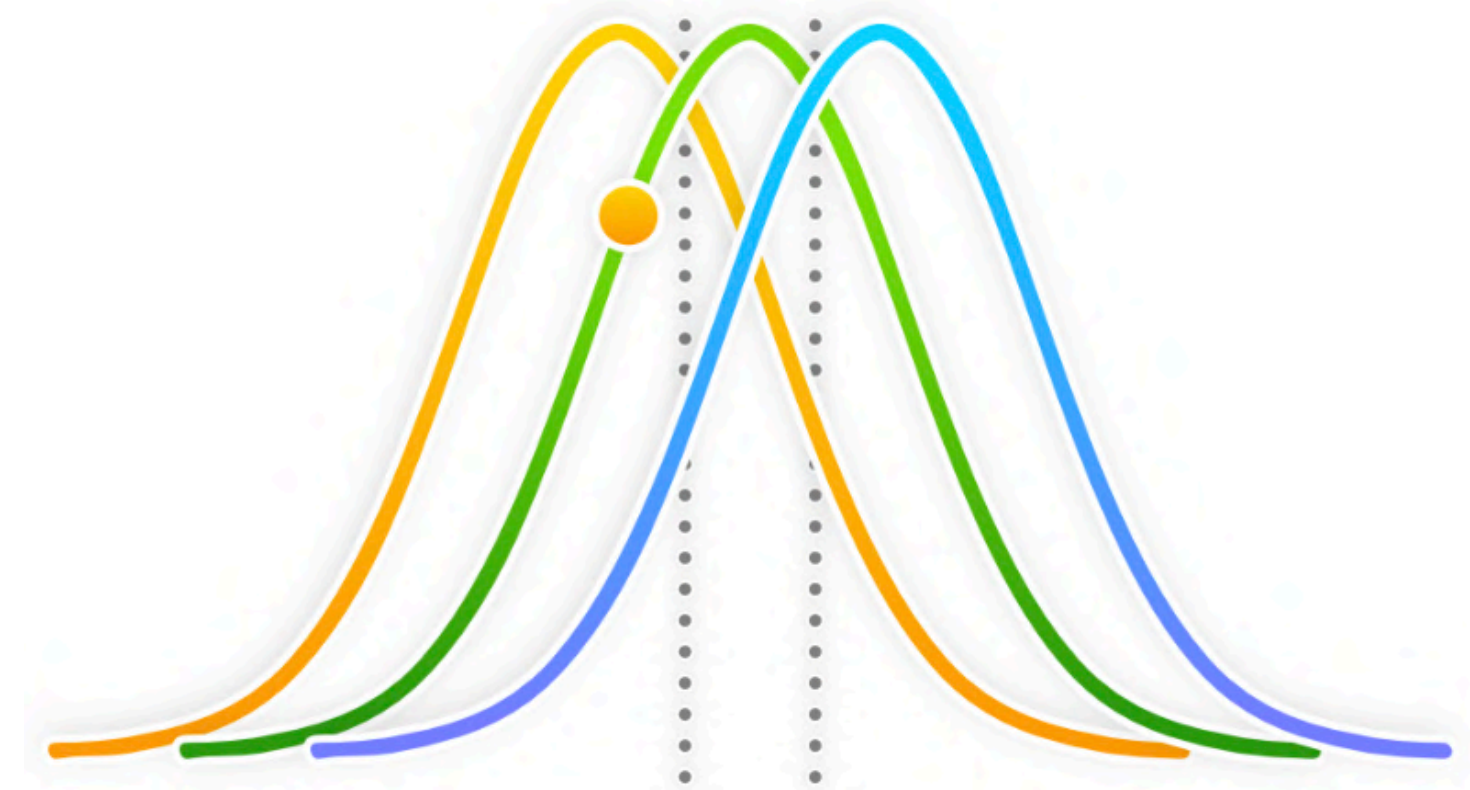


Visual Distance (Visual Overlap)

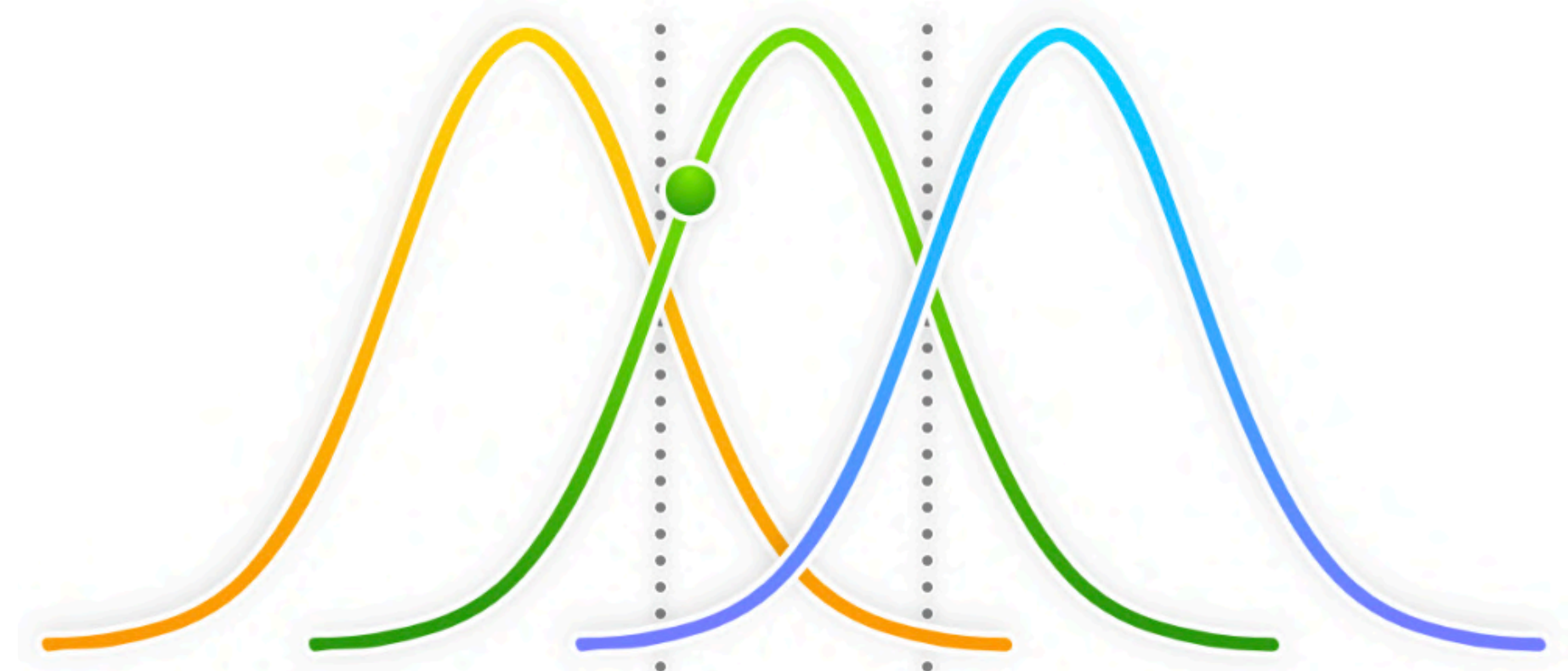
A Way To Examine Existing Datasets

- Use reconstruction loss as a distance metric
 - Compare error between true reconstruction and sampled reconstruction
 - Assume perfect reconstructions we are just comparing elements from the dataset itself
 - MSE is just square euclidean distance

$$\begin{aligned}d_{\text{vis}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) &= \lim_{\hat{\mathbf{x}} \rightarrow \mathbf{x}} \mathcal{L}_{\text{rec}}(\hat{\mathbf{x}}^{(a)}, \hat{\mathbf{x}}^{(b)}) \\ &= \mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})\end{aligned}$$



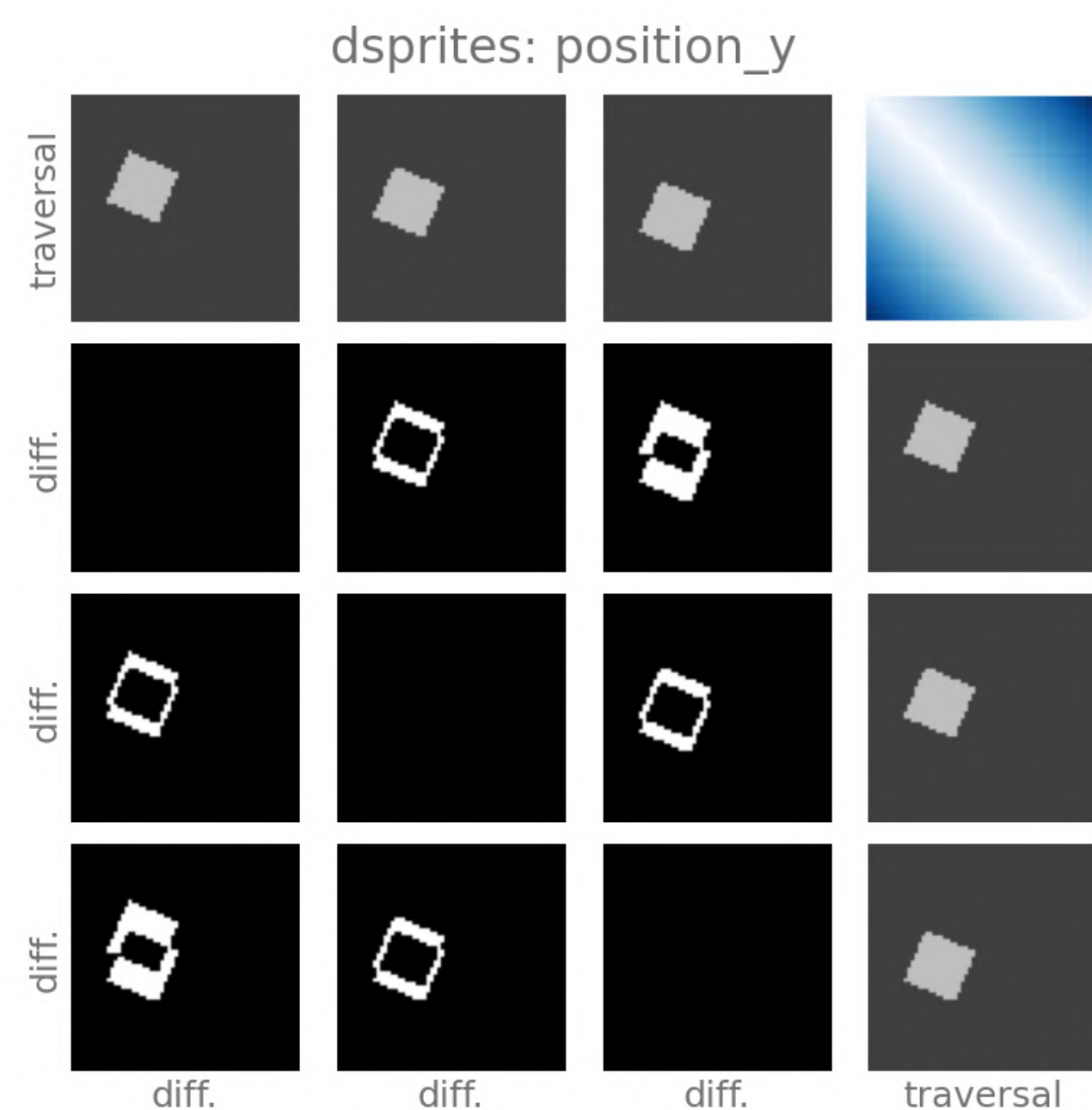
b: Confused Distribution



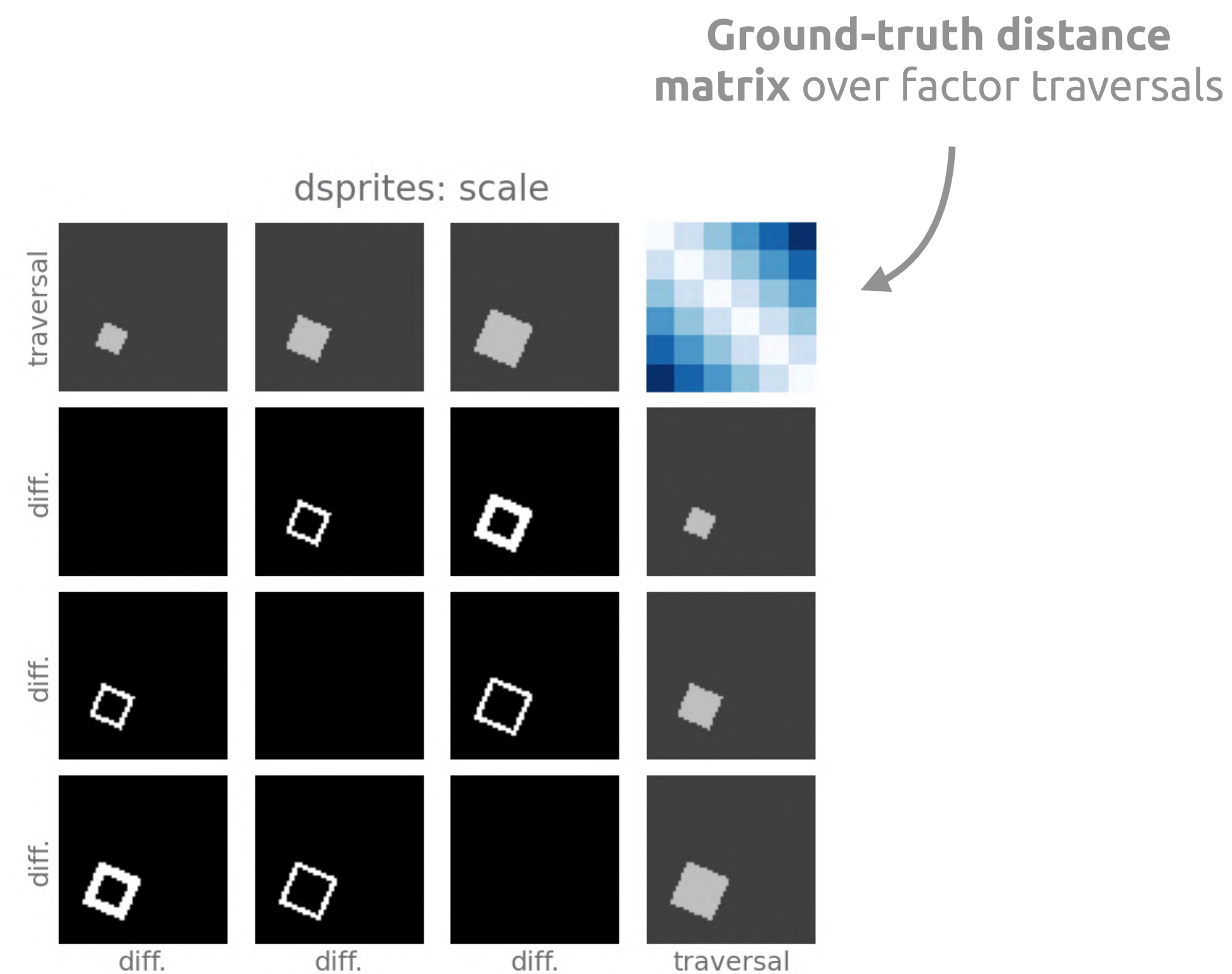
a: Correct Distribution

Distance Matrices

A Way To Examine Existing Datasets

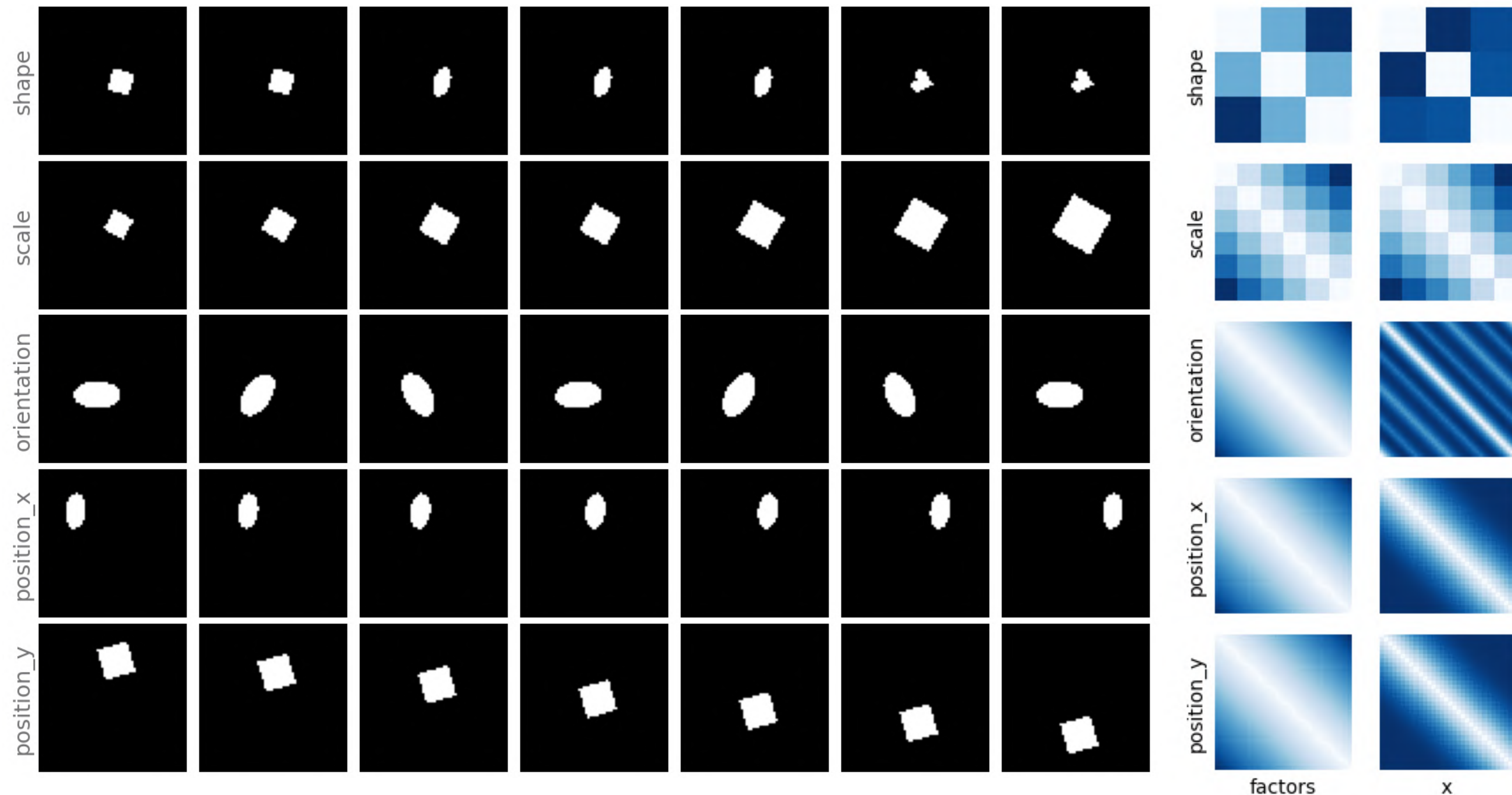


Visual distance matrix
between observations
over factor traversals



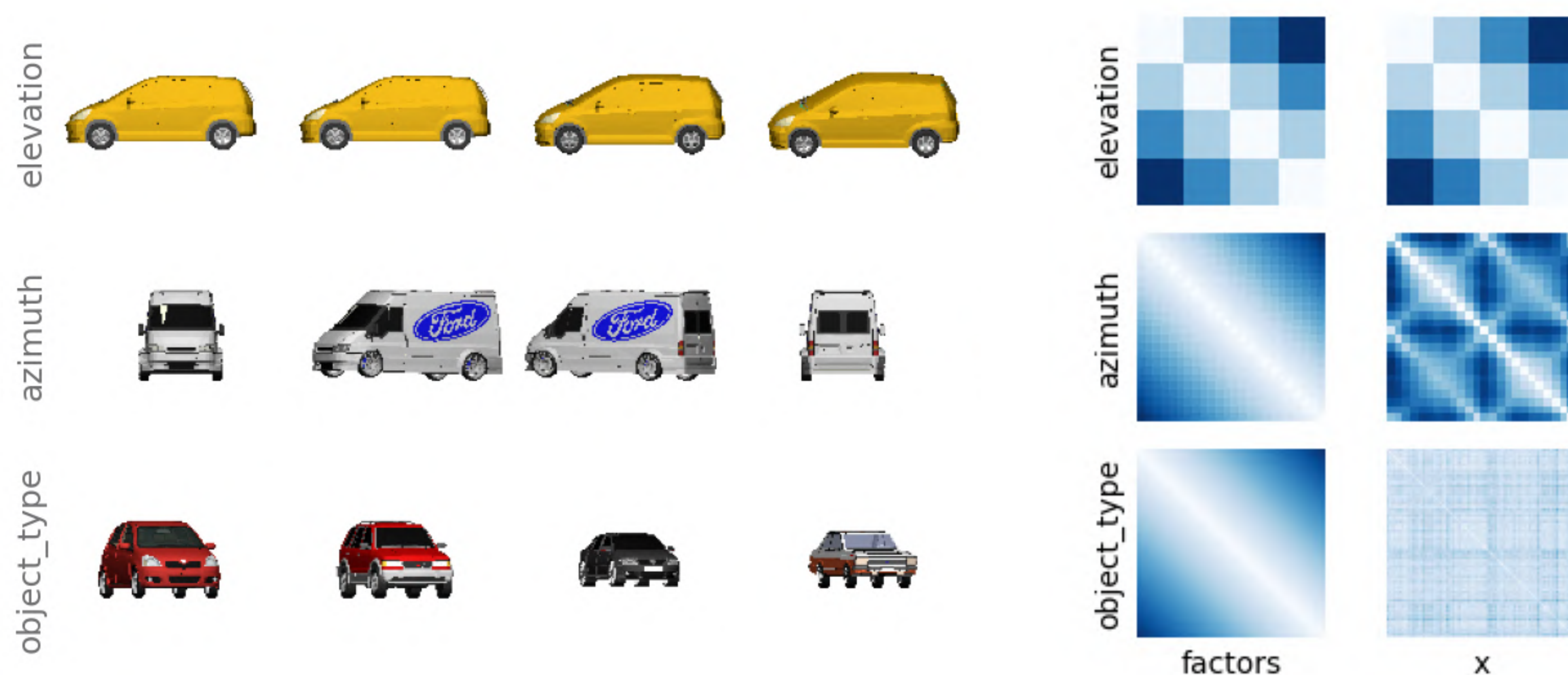
Existing Ground-Truth Datasets

DSprites: Ground-Truth Factors Correspond To Observation Distances



Existing Ground-Truth Datasets

Cars3D: Ground-Truth Factors Sometimes Correspond To Observations



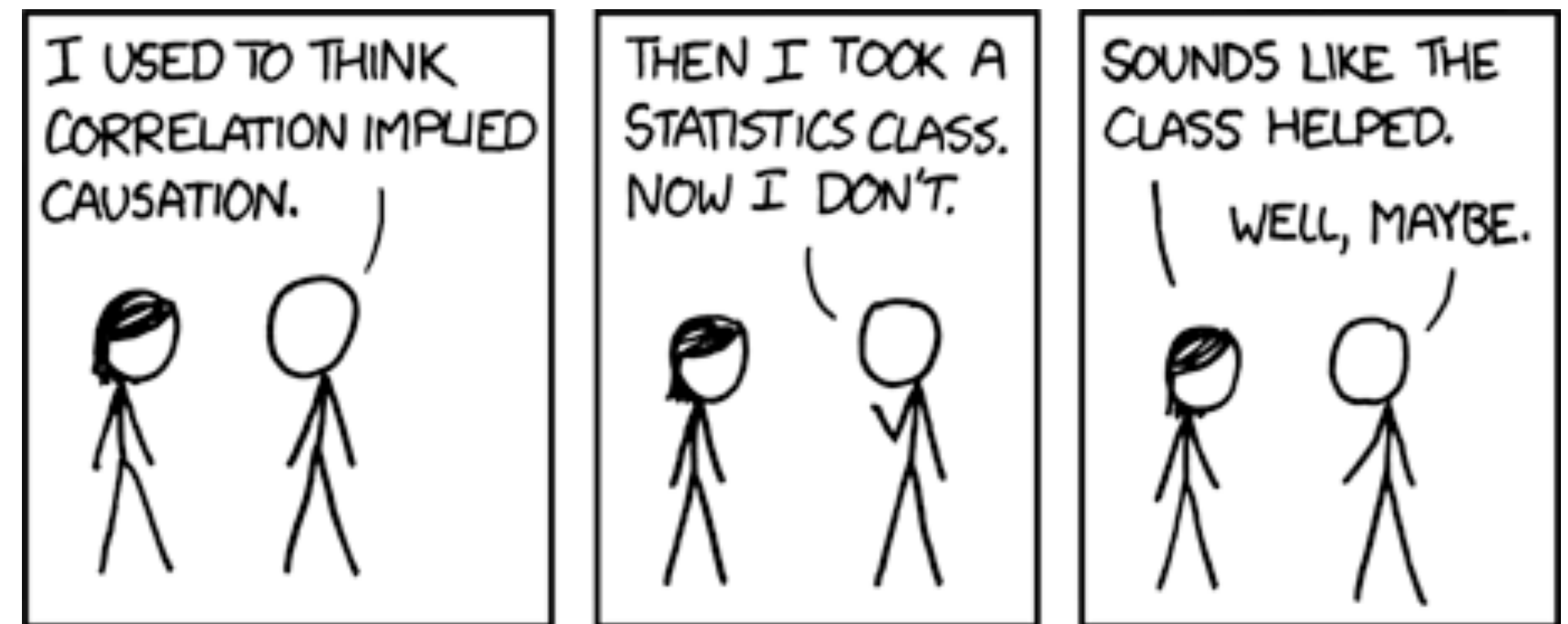
Same factors are easier to decompose into sub-factors

Something Seems Wrong

Are We Disentangling By Chance?

- Distances between ground-truth **factors naturally align with** distances between **observations** in the datasets based on the loss
 - Does this mean everything is happening by accident?
 - Can we purposely break VAEs by carefully designing the input data?
- Is this why these methods struggle on real world data?

Then... Again...



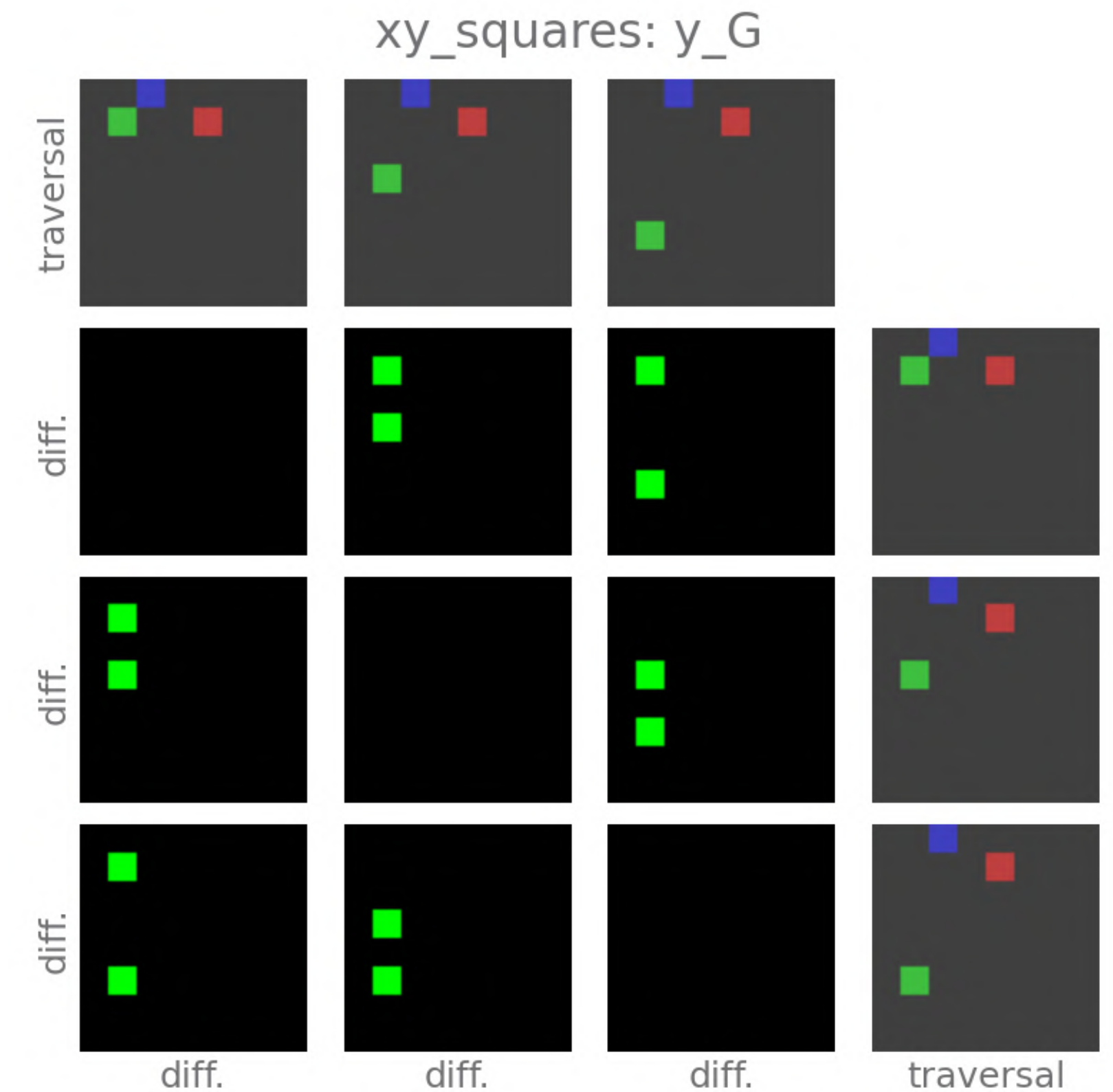
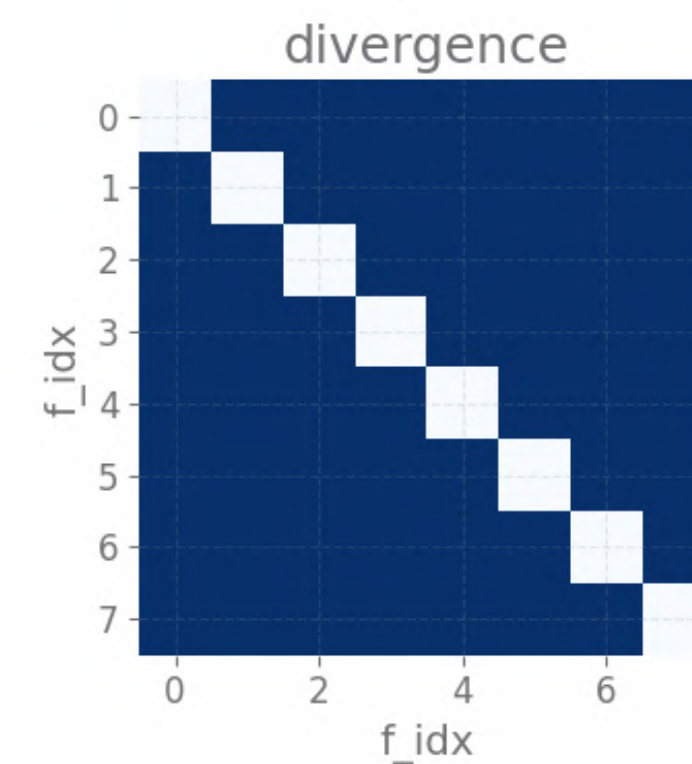
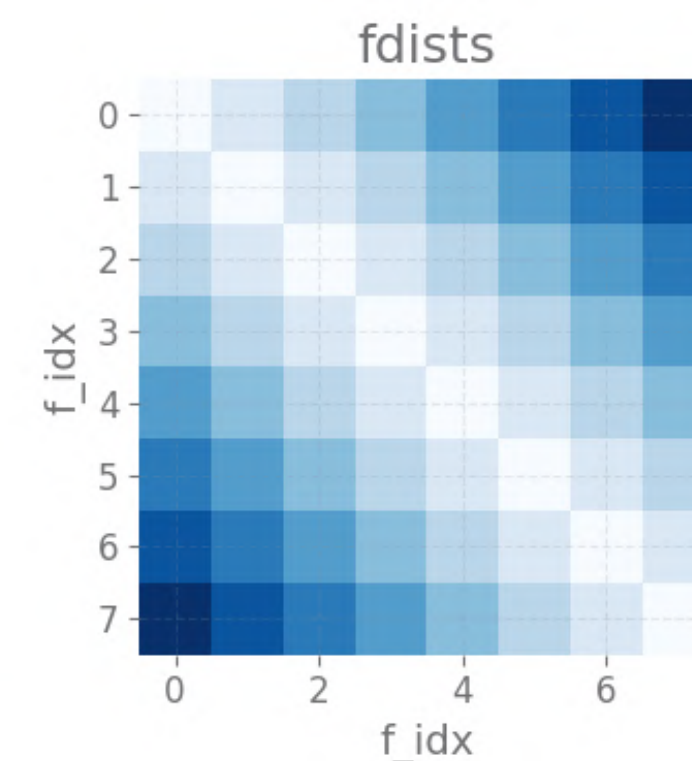
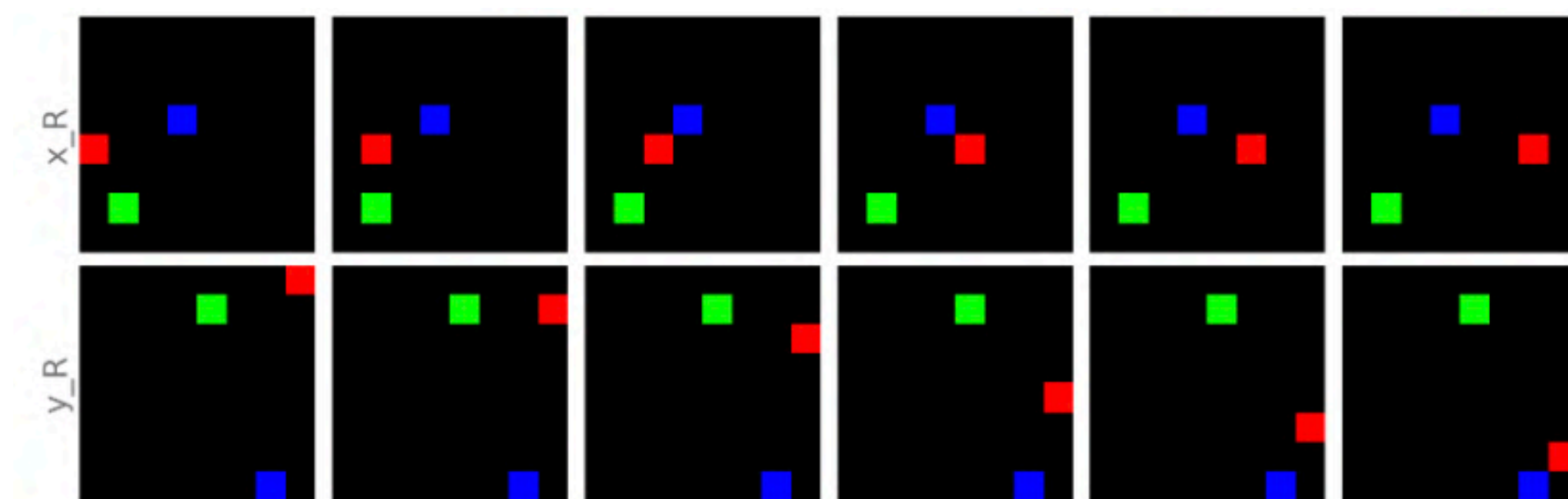
XKCD: Correlation from <https://xkcd.com/552>

3. Adversarial Dataset

Constant Overlap

How Can We Fool The Frameworks?

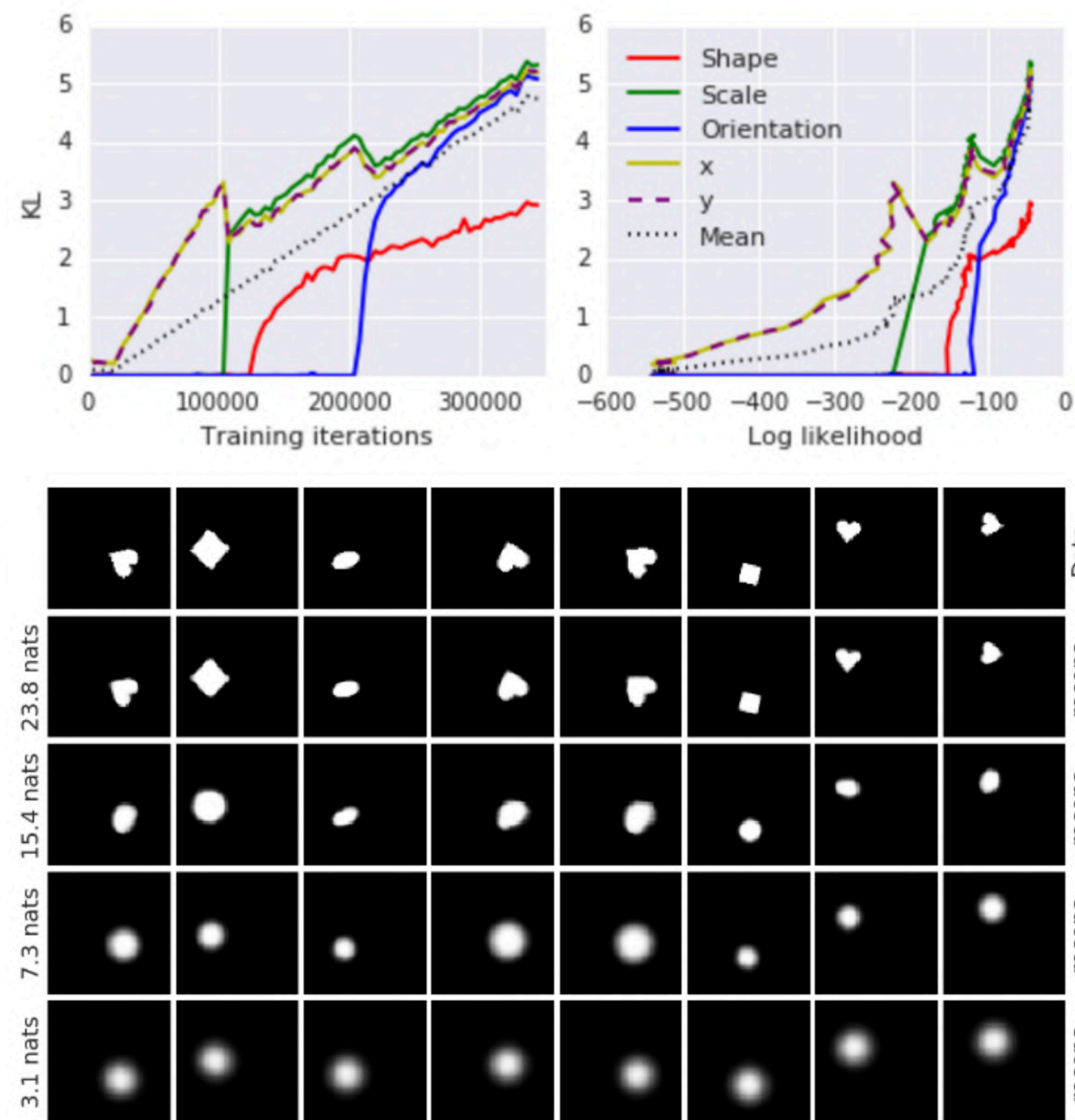
- Design dataset with **constant overlap along factor traversals**
 - Minimising reconstruction error due to sampling **cannot induce ordering**
 - “no overlap” is technically a wrong description for this dataset



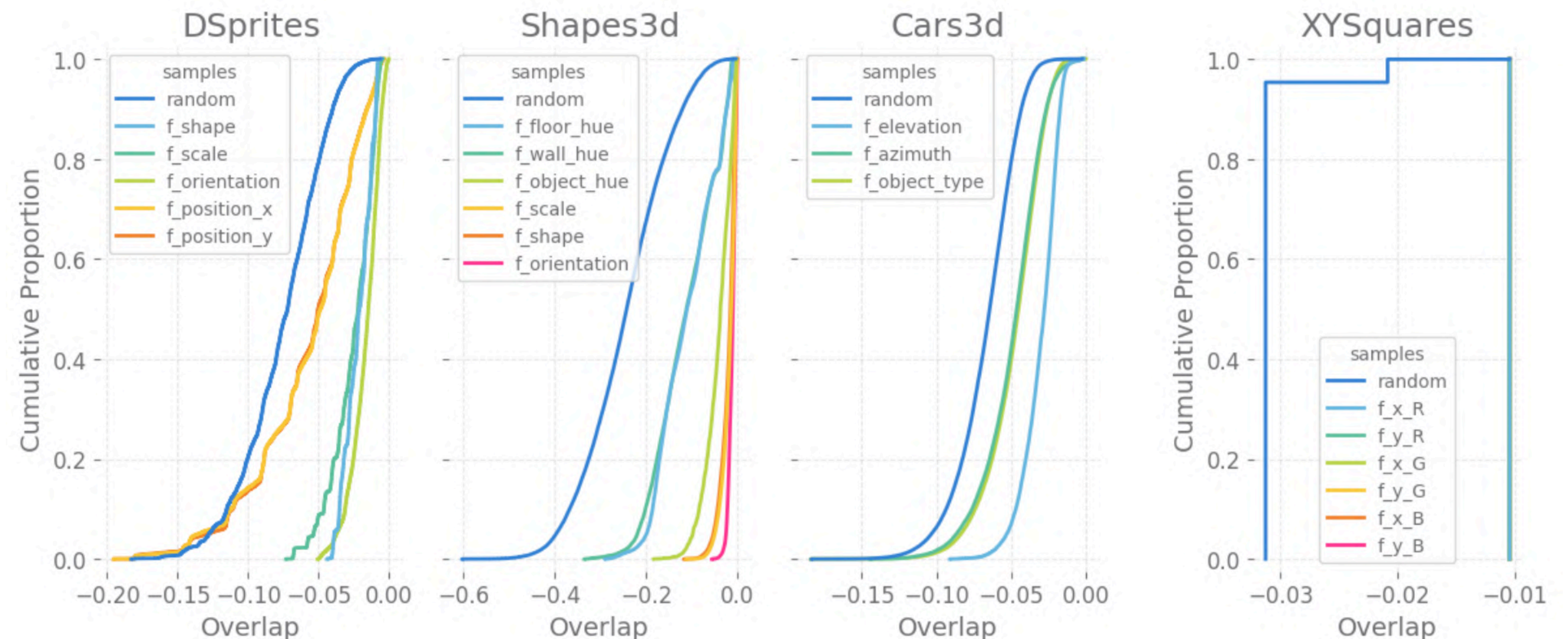
Same amount of overlap
between each obs. in a
factor traversal

Overlap In Datasets

The Most Important Factors Can Be Found By Looking At Their Overlap



From: Understanding Disentanglement in Beta-VAEs



(a) Common Disentanglement Datasets

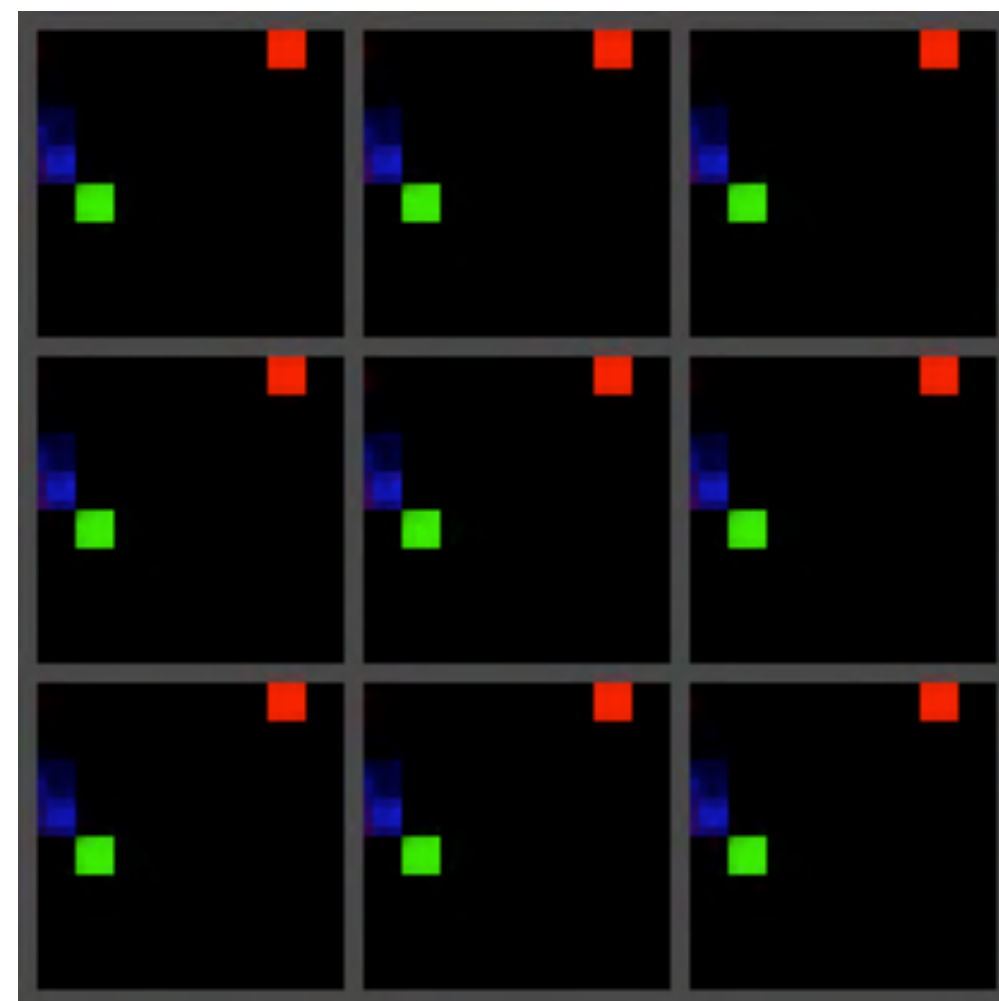
(b) Adversarial Dataset

$$\mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})$$

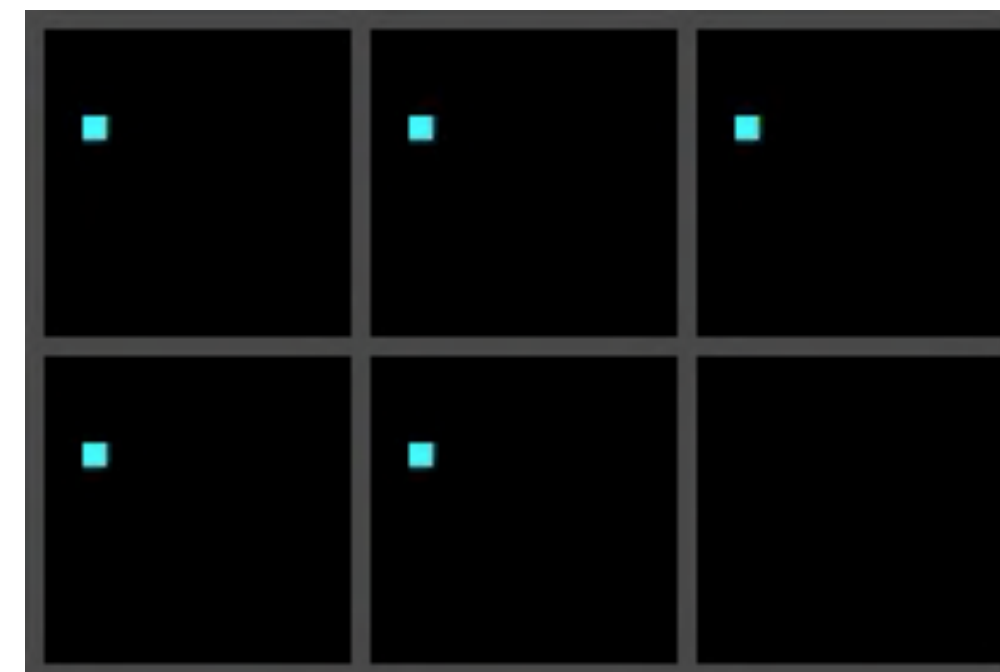
Our Results: Direct Pairwise Loss Between Dataset Samples

Adversarial Results

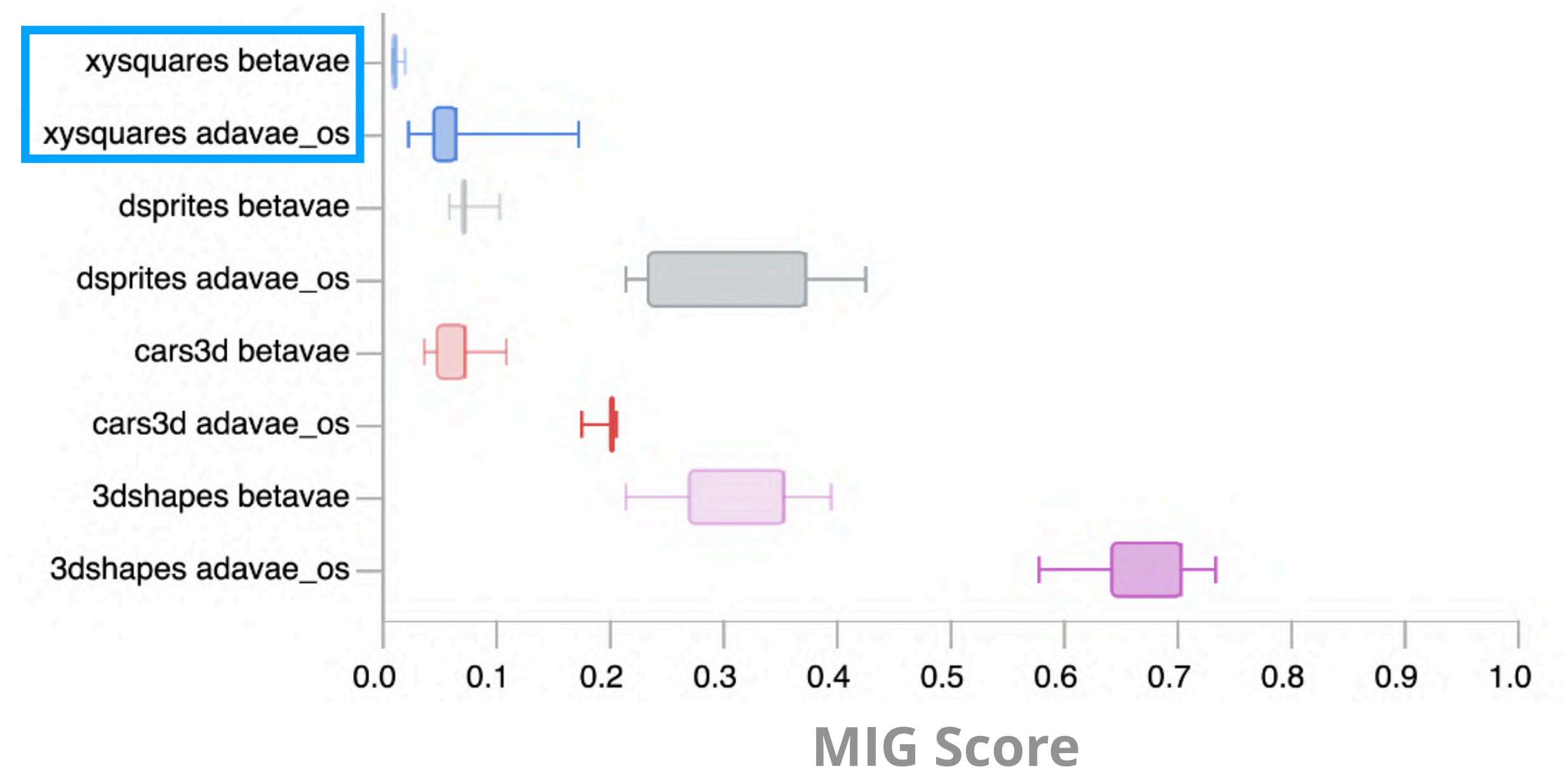
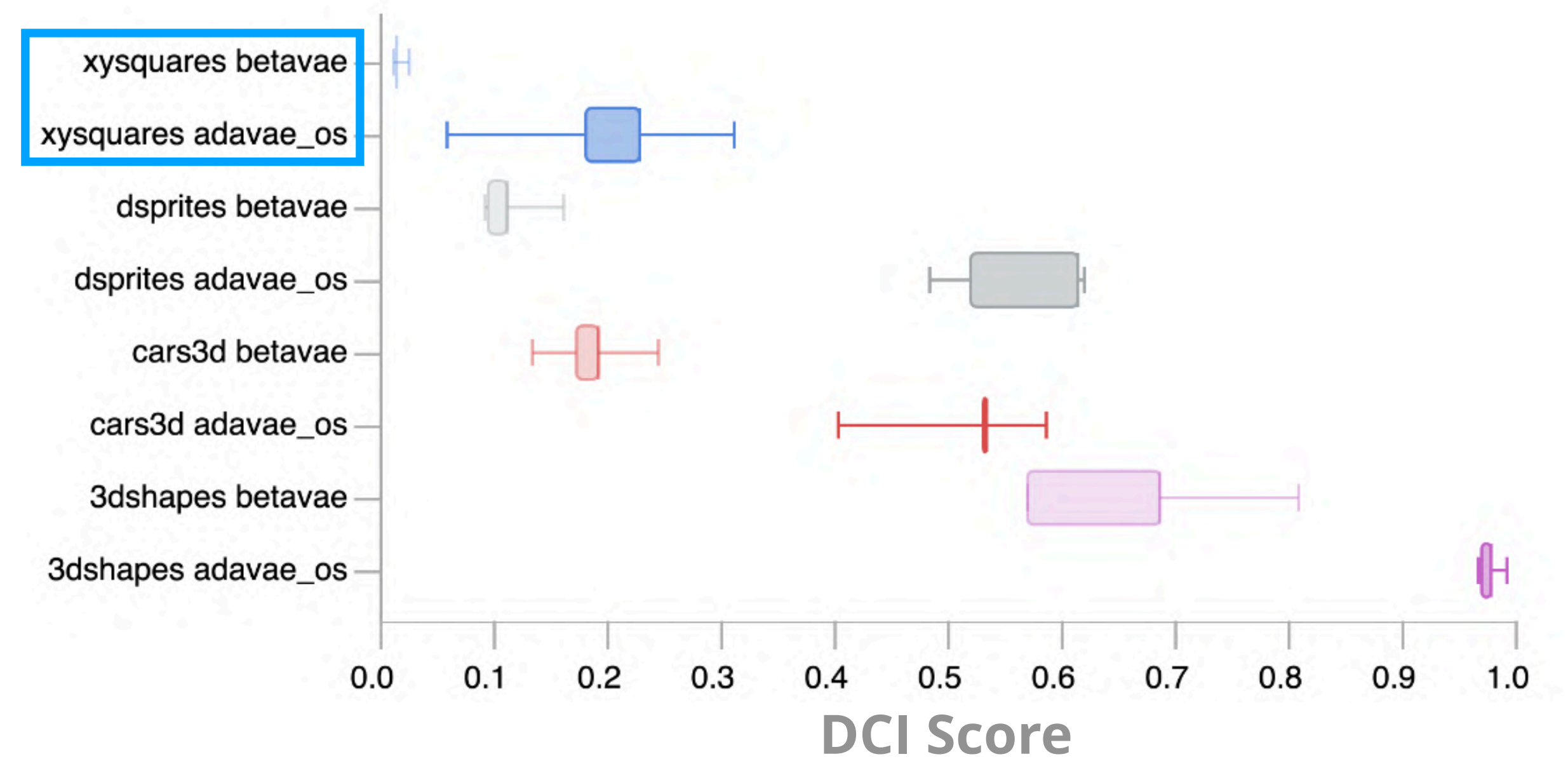
Everything Breaks!



Non-Overlapping Dataset
Ada-GVAE Latent Traversal

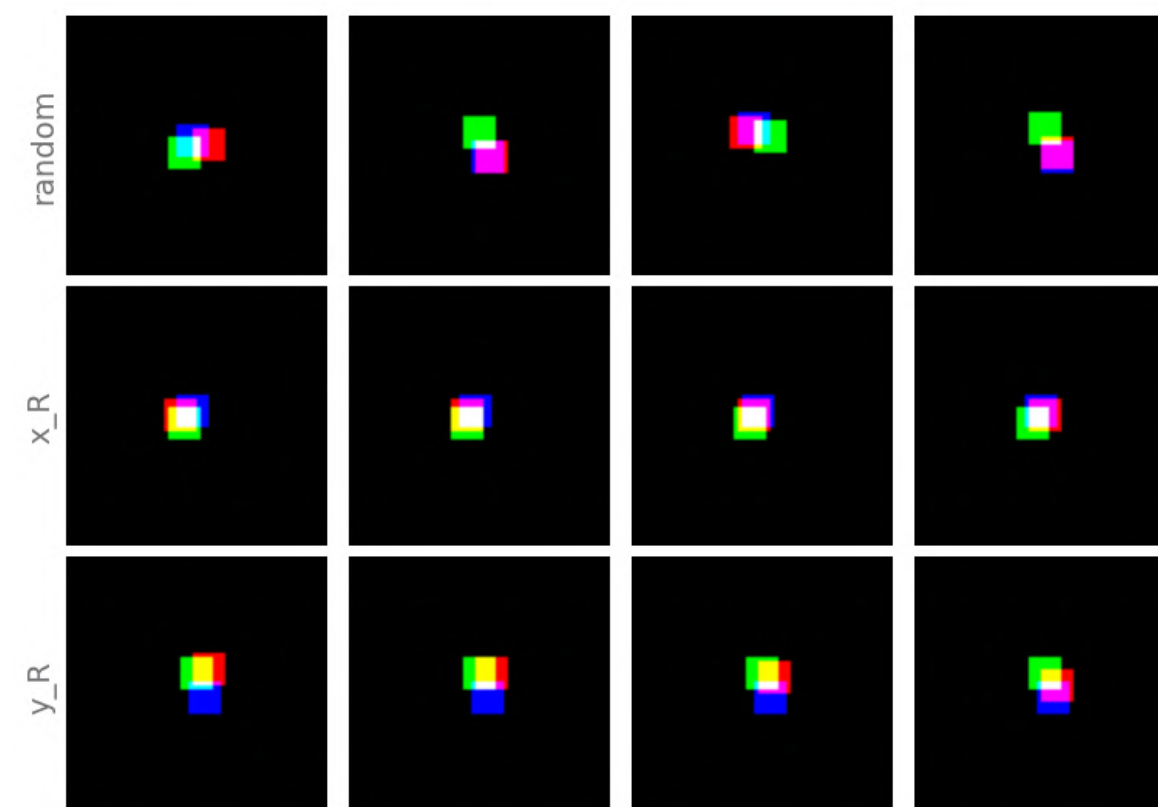


Overlapping Dataset
Ada-GVAE Latent Traversal

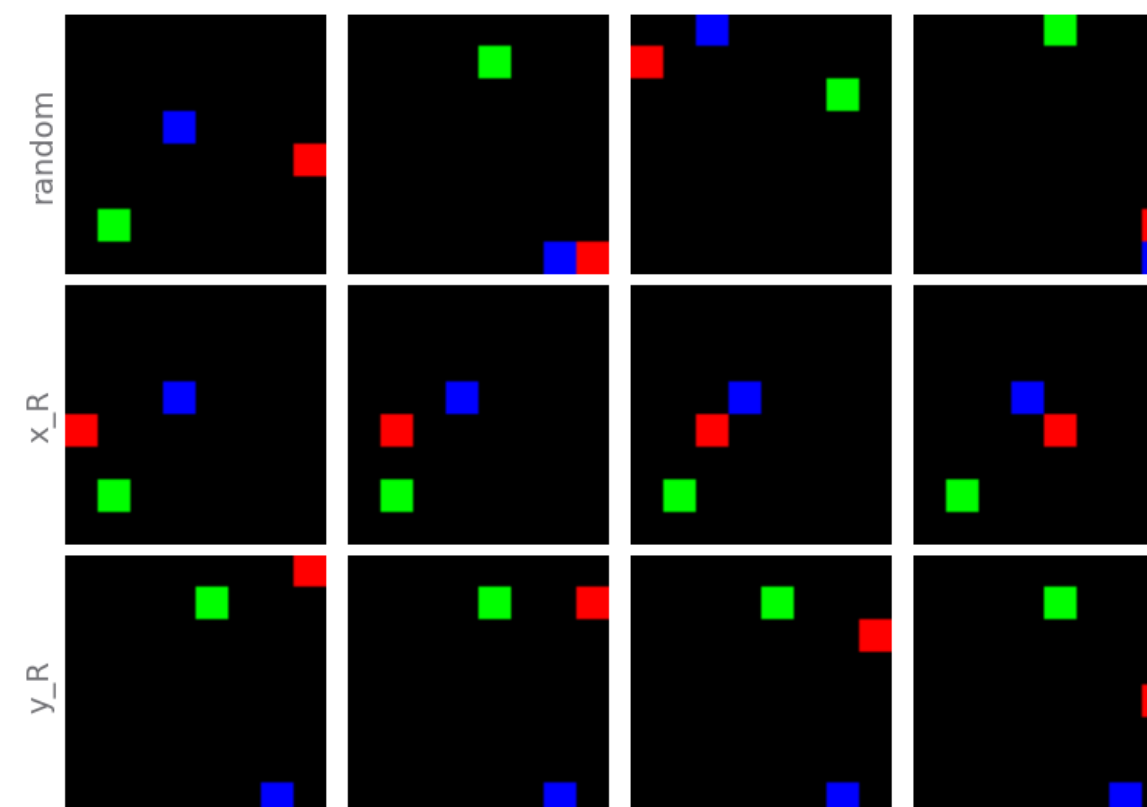


Increasing Overlap

Changing The Data

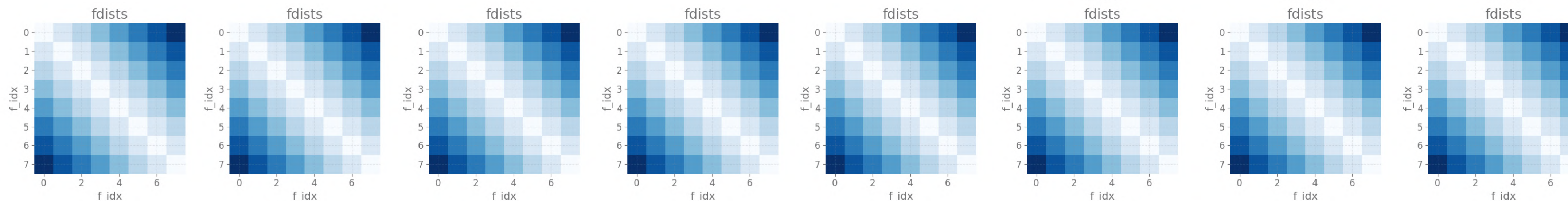


Overlapping Dataset
7 Pixels Overlap

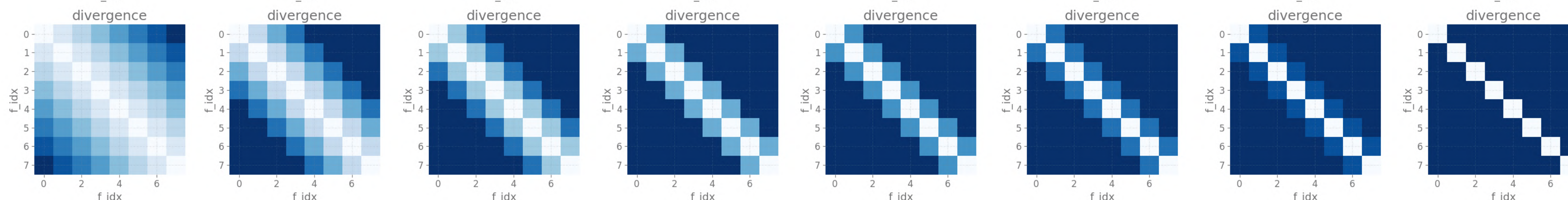


Original Adversarial Dataset
0 Pixels Overlap

Factor Dist.
Mat.



Overlap Dist.
Mat.



7px overlap

6px overlap

5px overlap

4px overlap

3px overlap

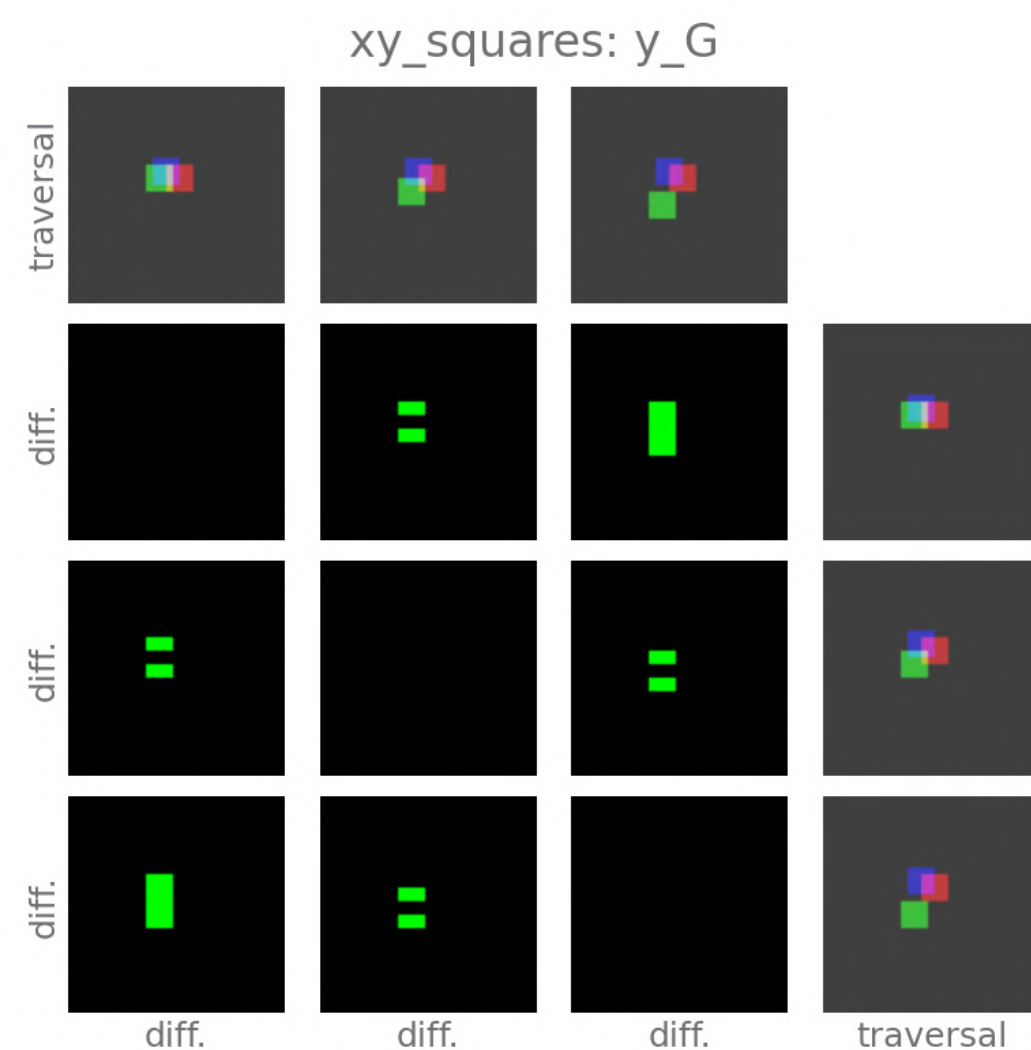
2px overlap

1px overlap

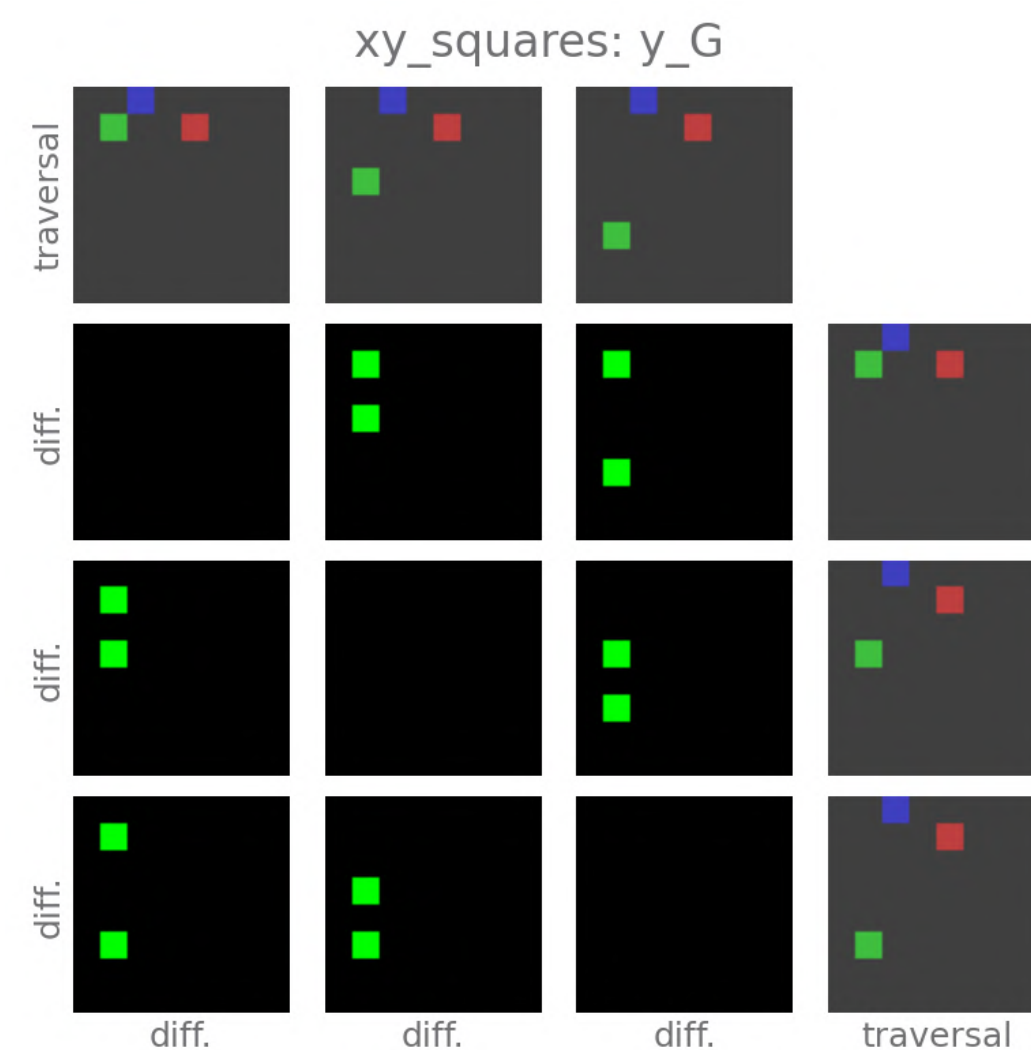
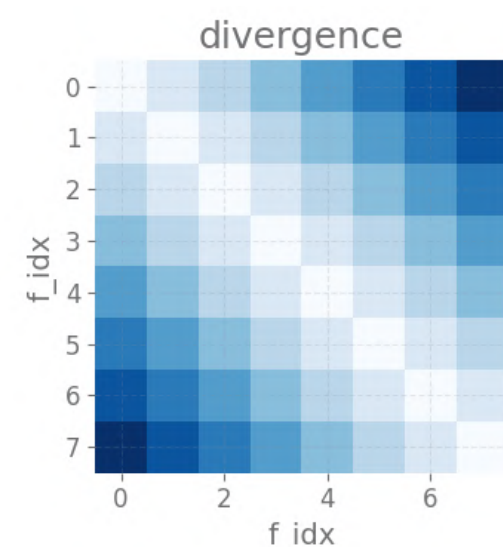
0px overlap

Increasing Overlap

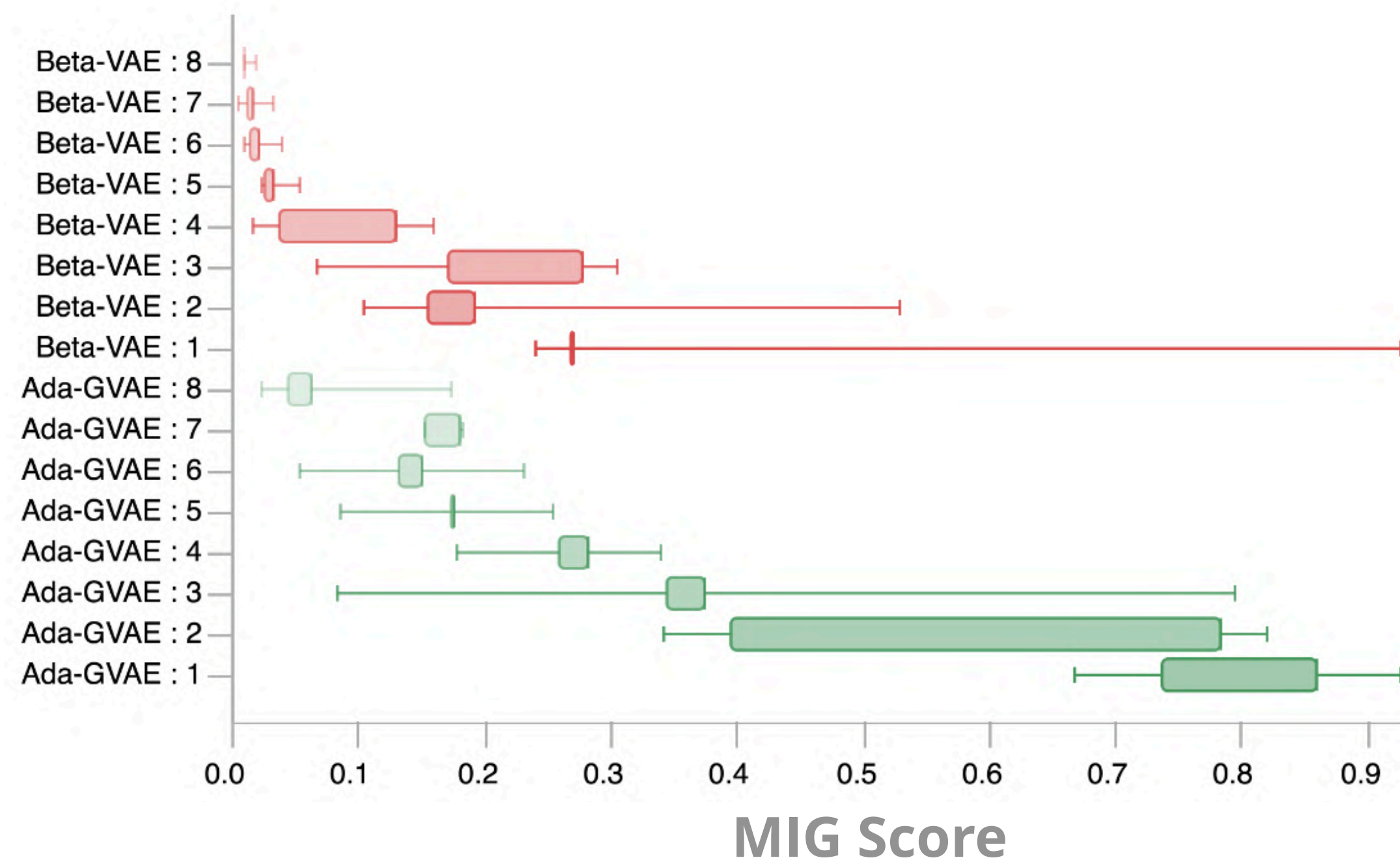
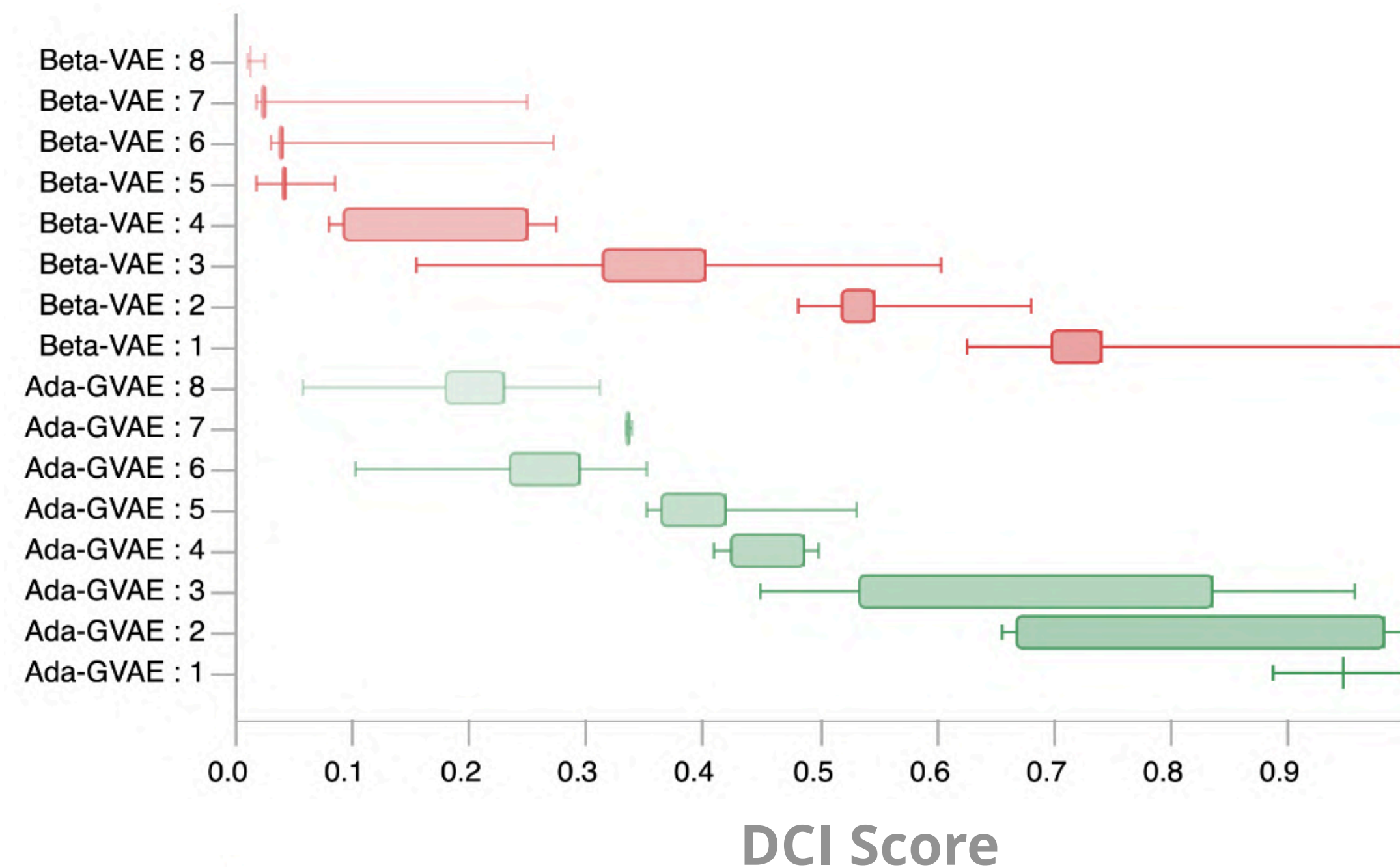
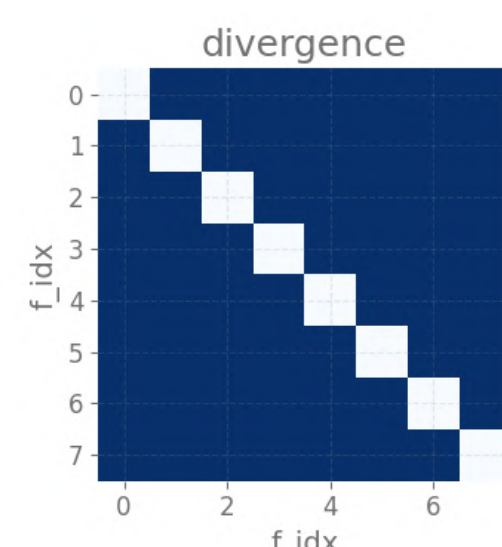
Constant Overlap Breaks SOTA



Overlapping Dataset
7 Pixels Overlap



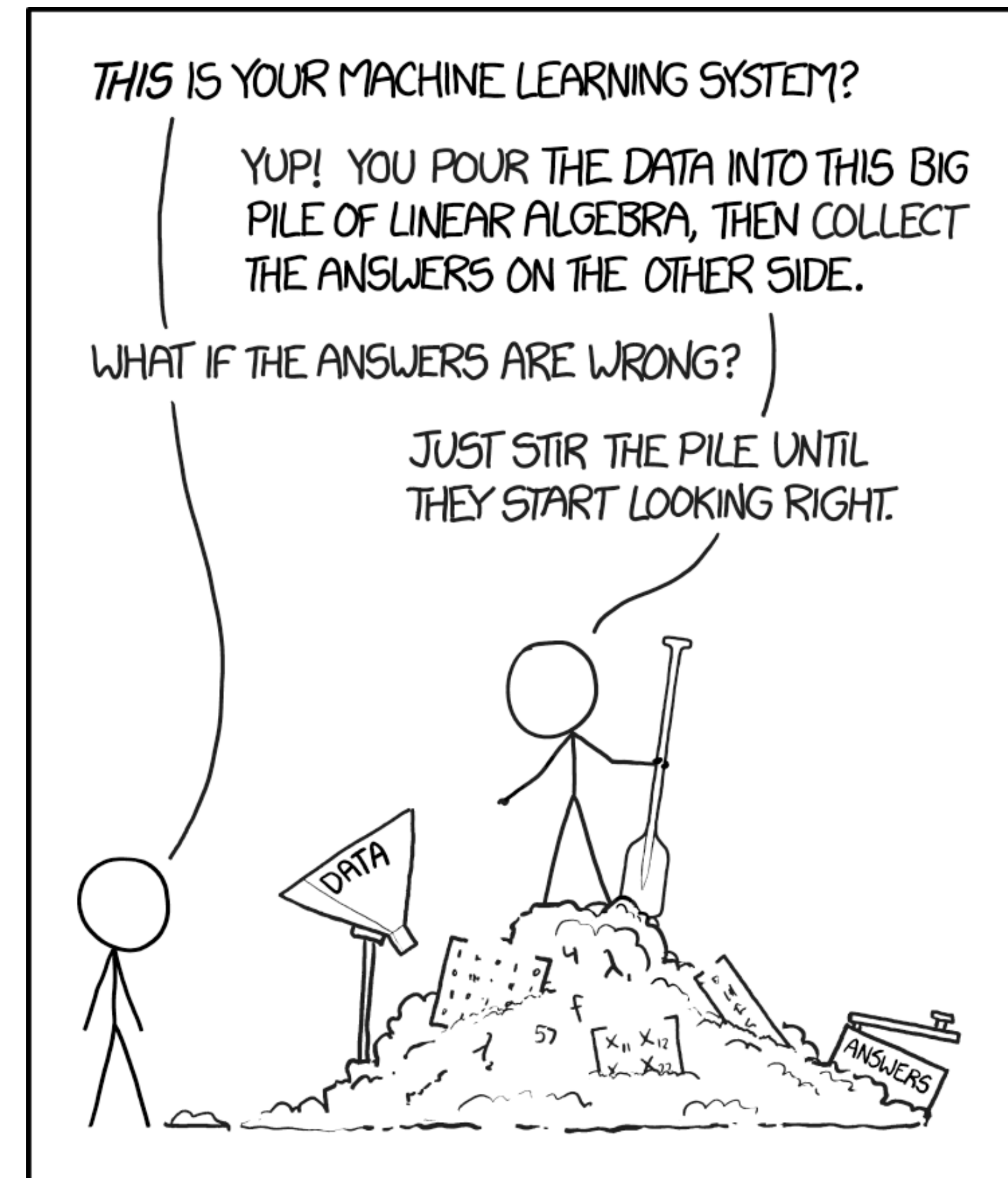
Original Adversarial Dataset
0 Pixels Overlap



Can We Disentangle The Adversarial Data?

We Could Use Metric Learning?

- Are VAEs Local Metric Learners?
 - Linear auto-encoders perform **PCA**
 - Observation **distances correspond** to ground-truth factors
 - Ordering is induced by **minimising sampling error** of reconstruction loss
- Can we use **metric learning** directly for disentanglement? (No decoder)
 - Start with **supervised** approach **first**. Then we can work on unsupervised.

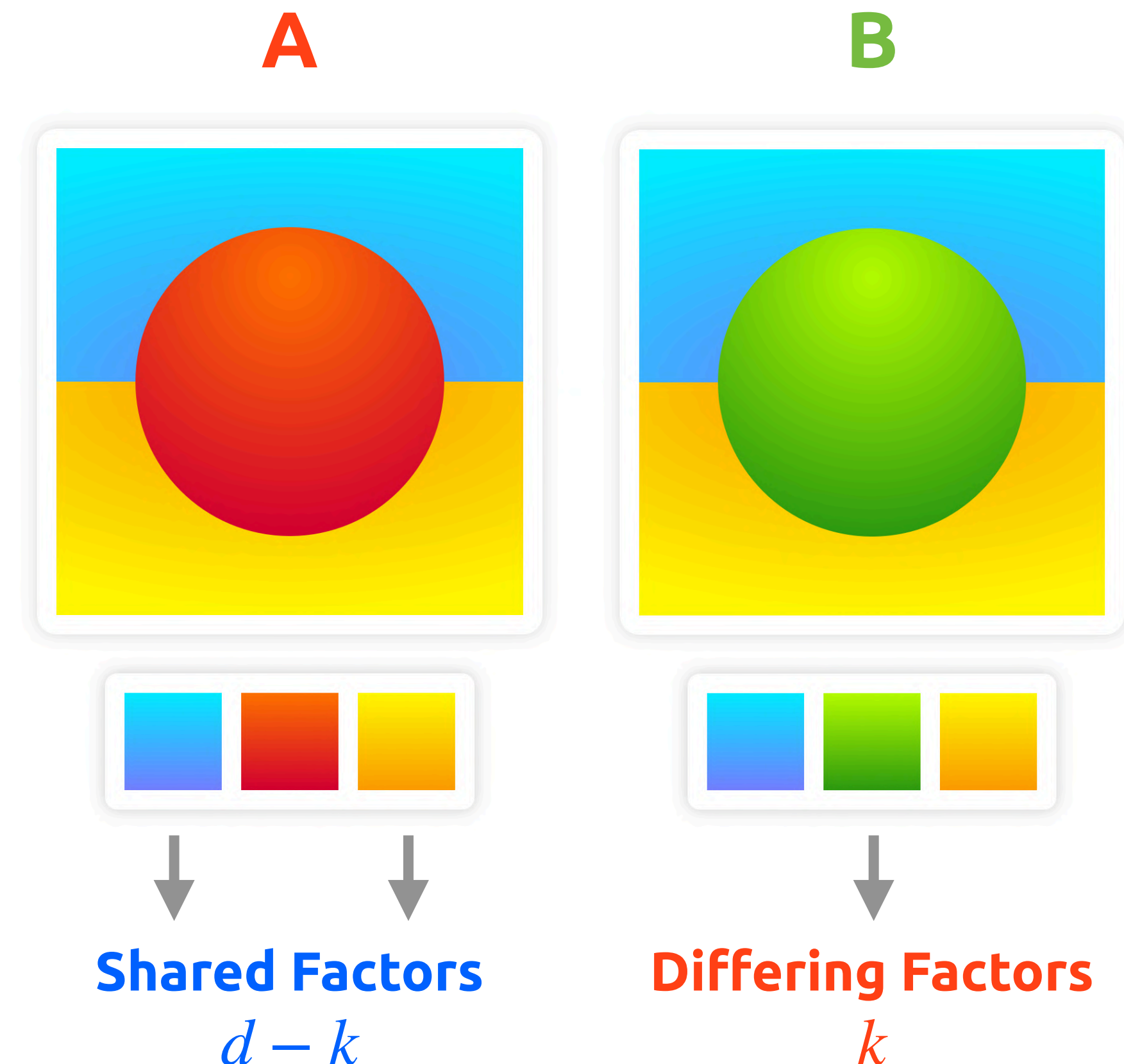


4. Background: **Adaptive VAEs**

Adaptive VAEs

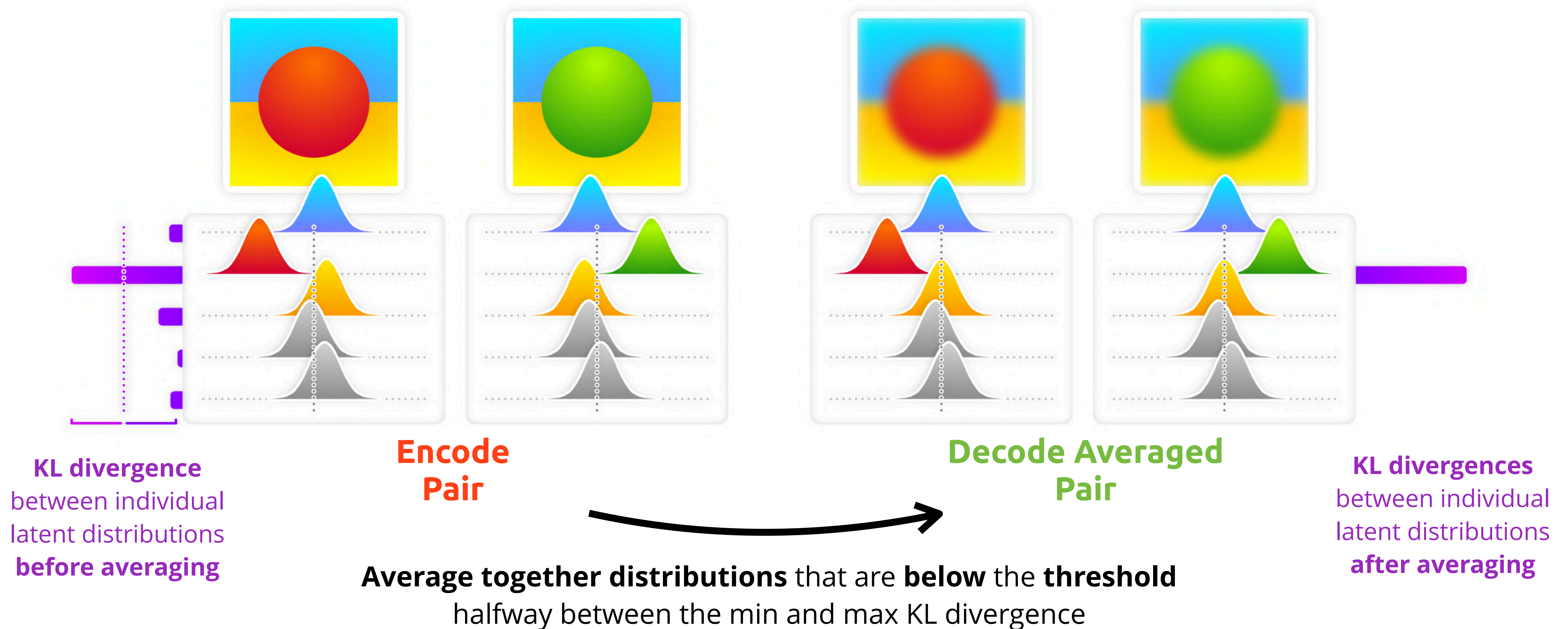
Weakly-Supervised Disentanglement Without Compromises ^[1]

- **Estimates which factors have changed** between observation pairs.
 - **Encourages** the desirable property of partially **shared representations**
 - Variants are: Ada-GVAE & Ada-ML-VAE changing how latents are averaged
- **Weakly-Supervised** by sampling pairs
 - Assumes **at least one factor** in a pair is **shared**



Adaptive VAEs

Algorithm Overview - Averaging



5. Supervised Adaptive Triplet

Triplet Loss

Background: Standard Triplet Loss

Goal

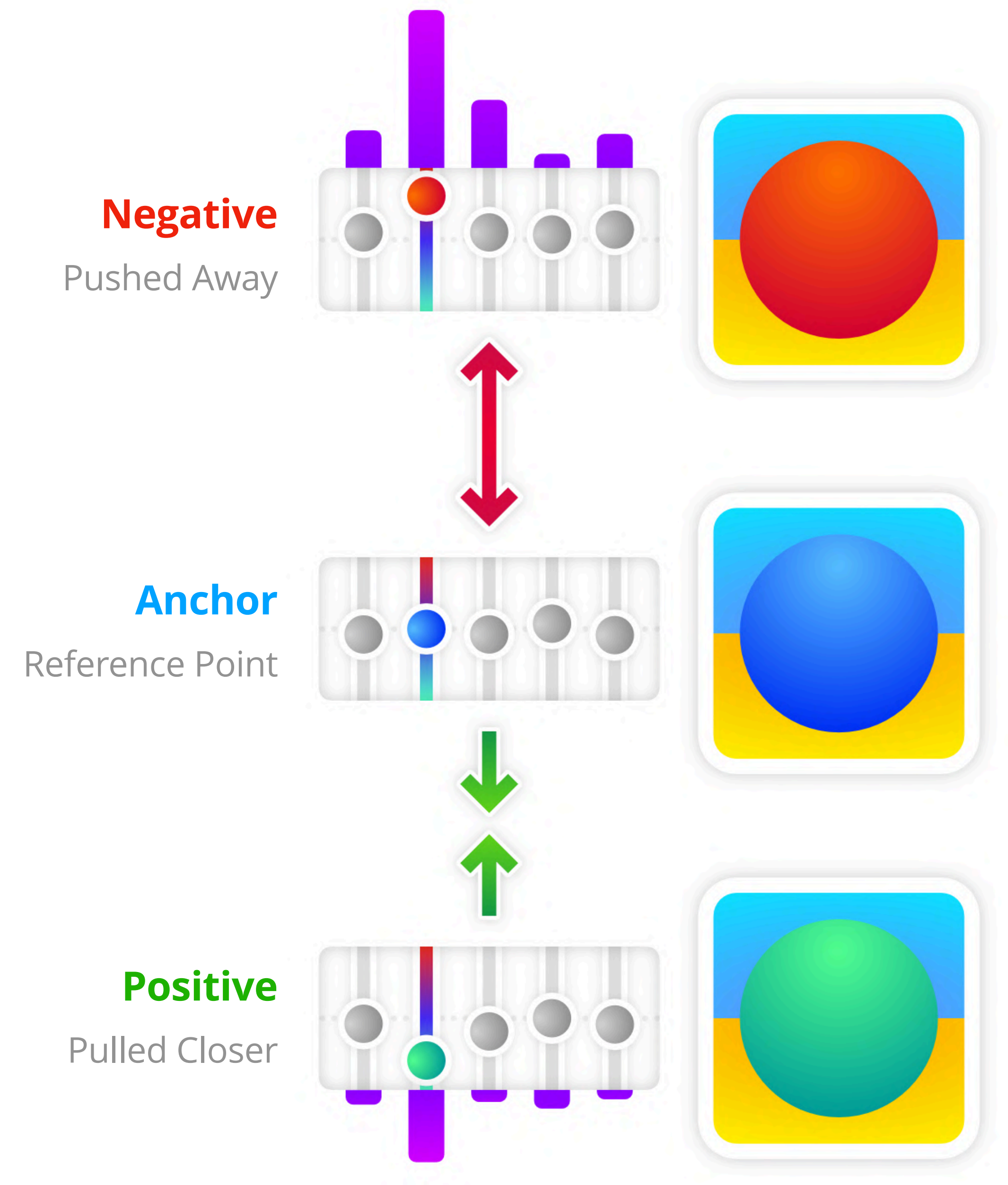
- Learn **embeddings** that correspond to some notion of similarity between **labeled observations**

How

- Push **negative** representation further away
- Pull **positive** representation closer together
- Negative has no effect beyond **margin**
- Applied to mean of distributions: $\mathbf{z} = \mu$

$$\mathcal{L}_{triplet} = \max(0, \psi + d(\mu^{(a)}, \mu^{(p)}) - d(\mu^{(a)}, \mu^{(n)}))$$

NB we no longer need a decoder, we just use it to visualise everything



Adaptive Triplet Loss

Weight Shared Elements Less

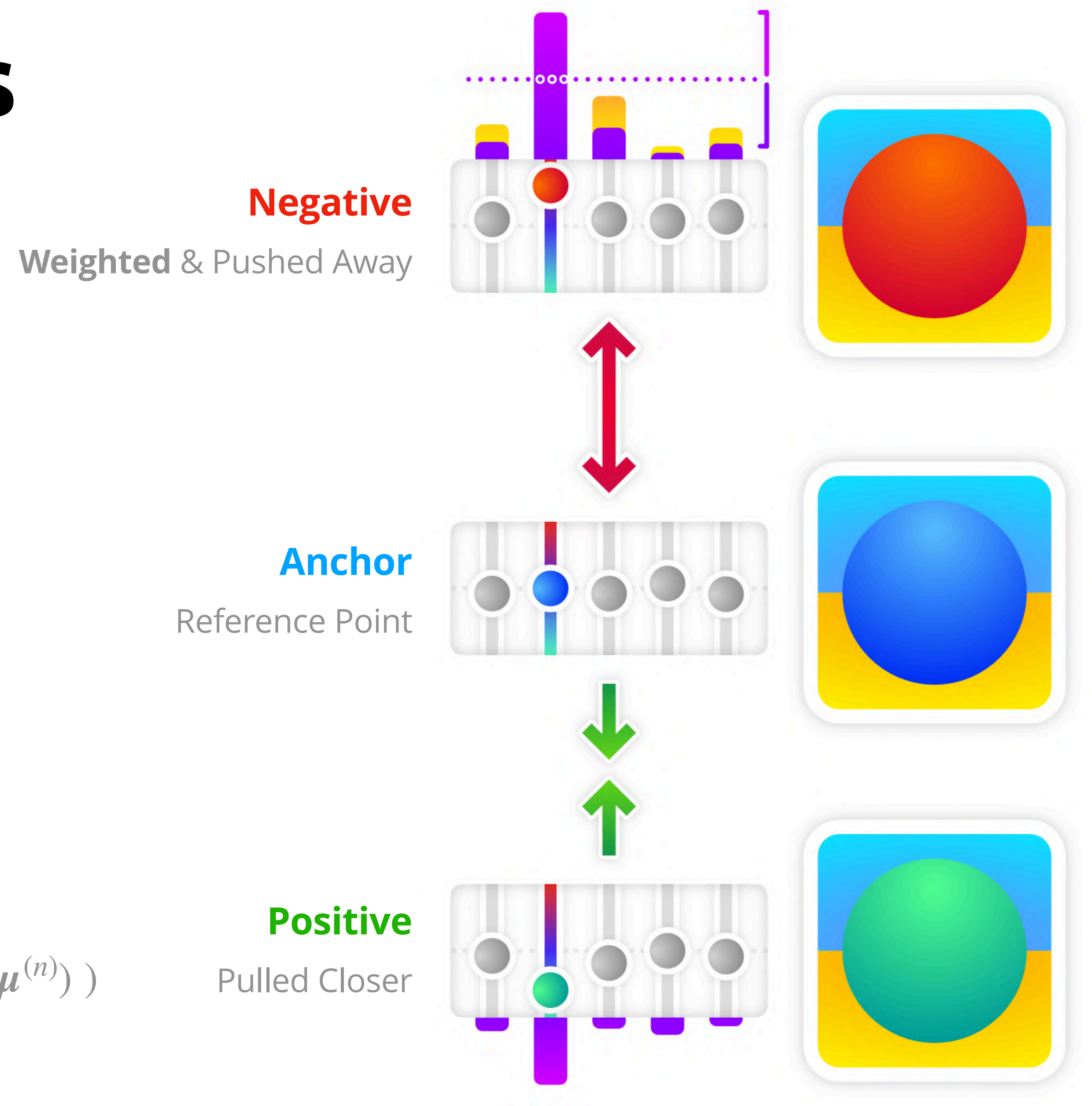
Goal

- Introduce axis alignment into triplet loss

How

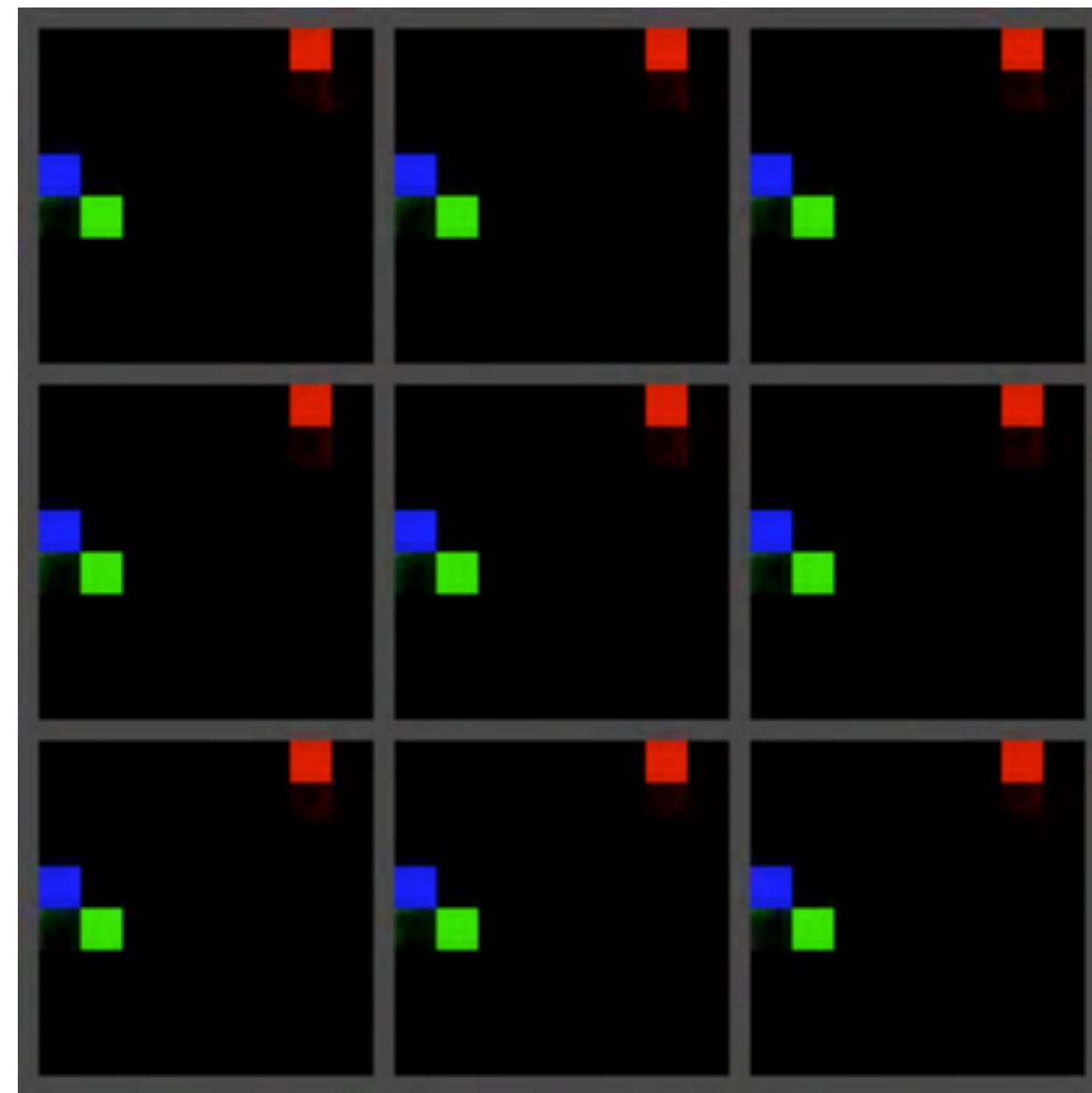
- **weight** ω shared elements of **anchor-negative** distance less
- Unchanged **anchor-positive** “fixes mistakes”
- Replace **L2** distance with **L1** distance. Better in high dimensions and lets latent variables die

$$\mathcal{L}_{ada-triplet} = \max(0, \psi + d(\mu^{(a)}, \mu^{(p)}) - d(\omega \odot \mu^{(a)}, \omega \odot \mu^{(n)}))$$

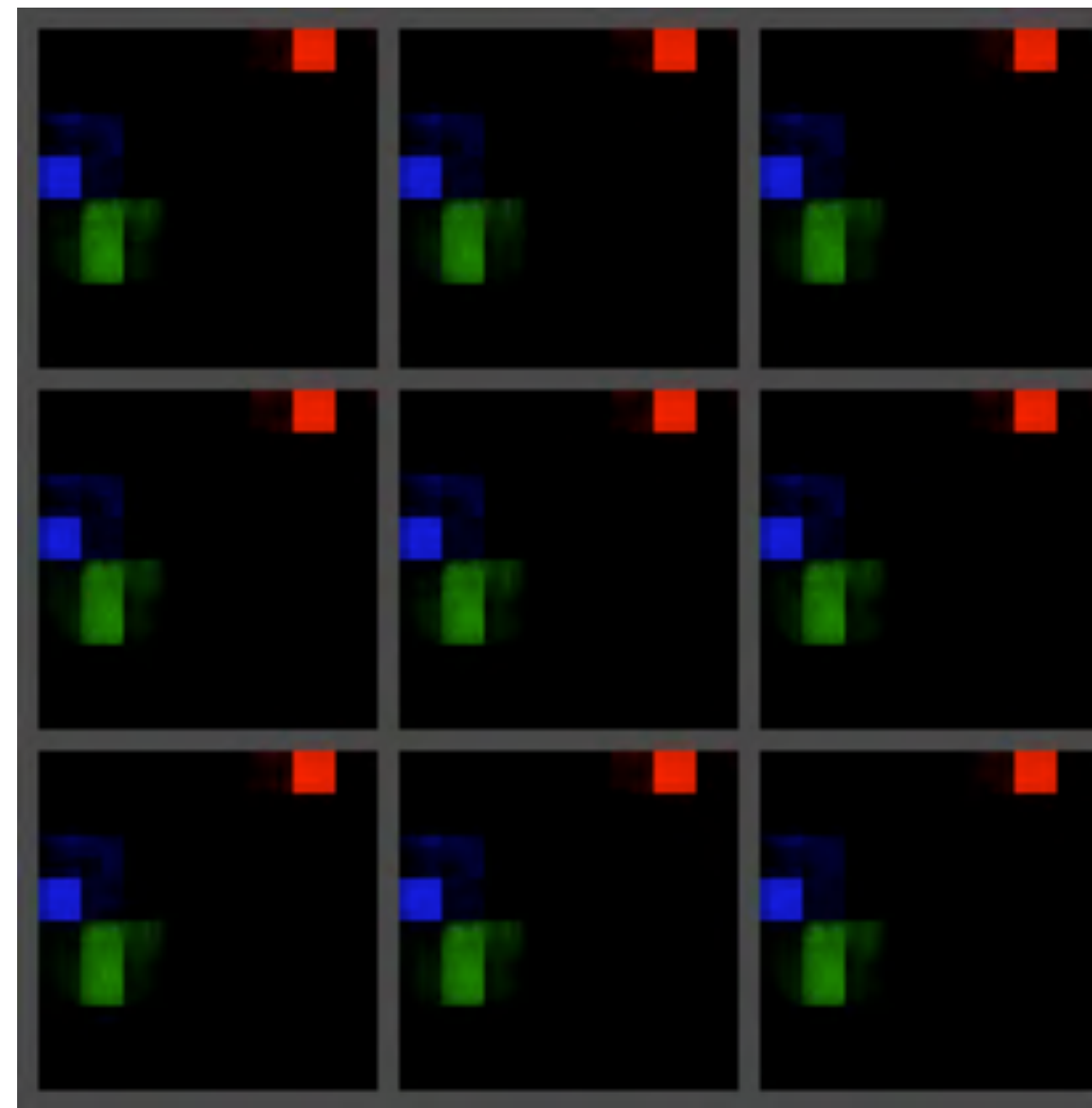


Supervised Triplet Results

Detached Decoder & Slowly Increasing Averaging Threshold



Supervised **Triplet**

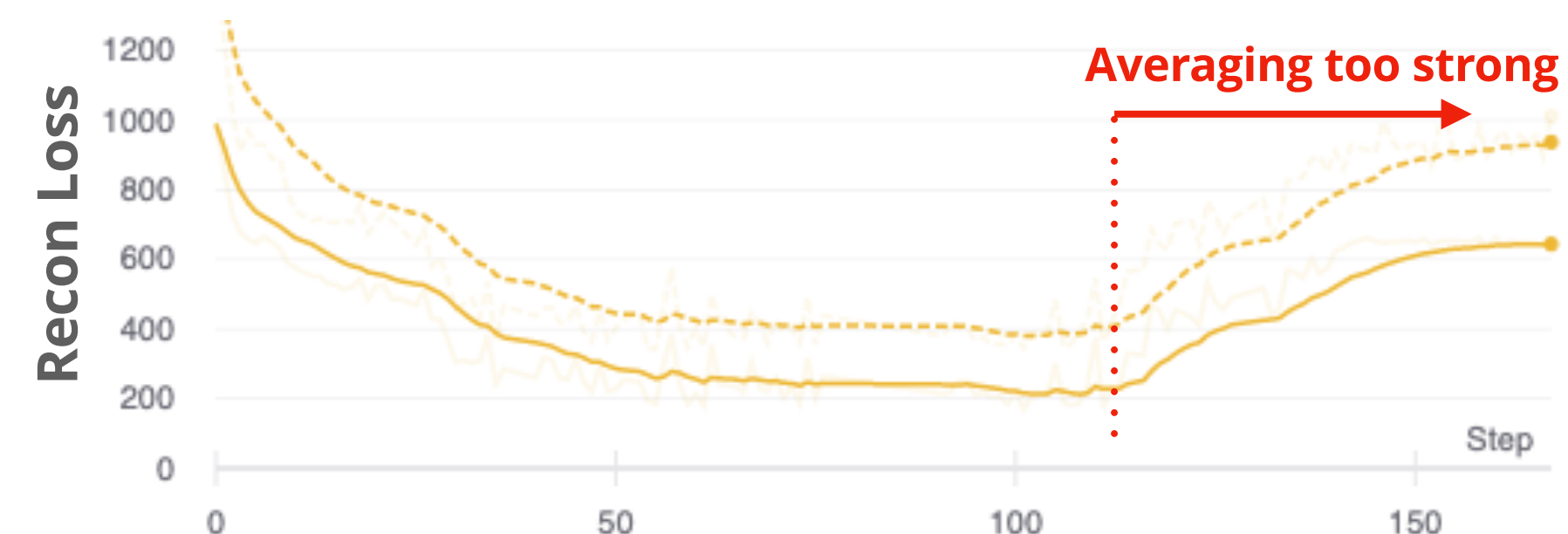


Supervised **Adaptive Triplet**



Supervised **Adaptive Triplet**

Squares move in the same pattern in different blocks - **not axis aligned**, but aligned along some arbitrary axis



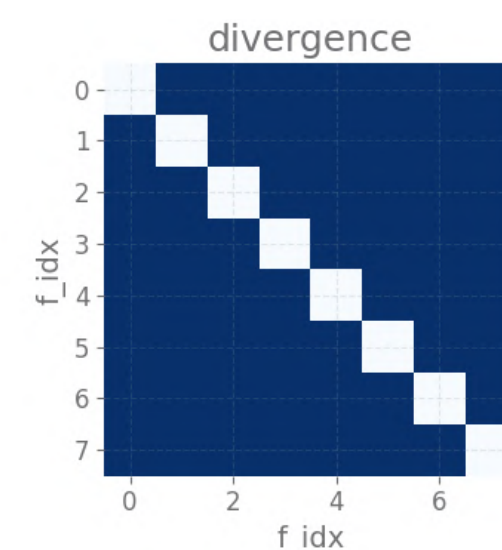
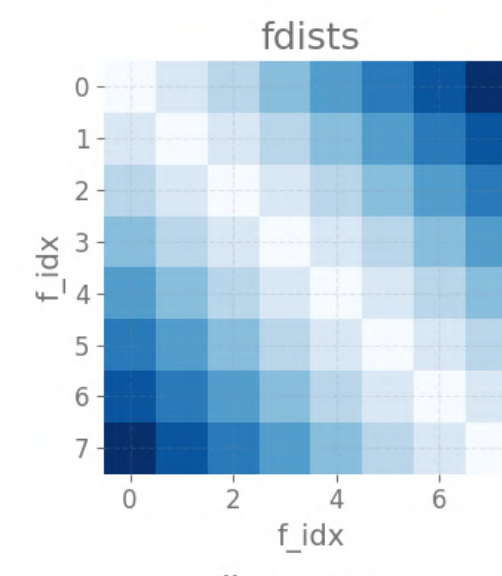
Increasing Ave. Threshold & Decreasing Ave. Weights

6. Unsupervised Adaptive Triplet

Unsupervised Adaptive Triplet

Order Triplets Using Visual Distance

- Use visual distance to sample triplets
 - **Measure visual overlap**/distance between anchor-positive and anchor-negative pairs
 - **Swap positive and negative** if distance is wrong
- Existing datasets are suited to this approach
 - **Adversarial dataset will still fail**
 - **Solution:** Move away from pixel-wise distance function!



Supervised:

$$S_{\text{fdist}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \|\mathbf{y}^{(a)} - \mathbf{y}^{(b)}\|_1$$

“Order triplets according to **Manhattan distance along factors**”

Unsupervised:

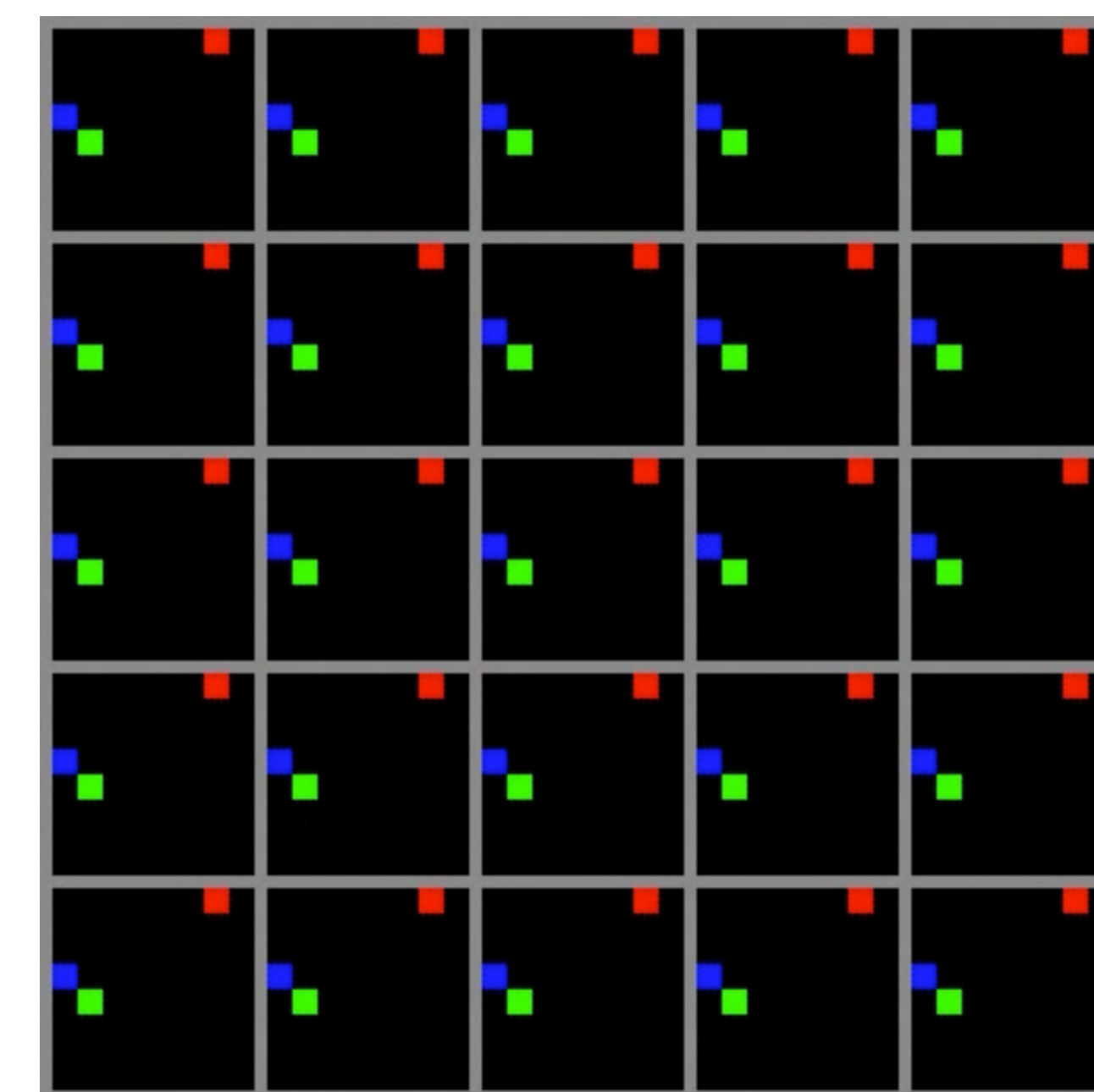
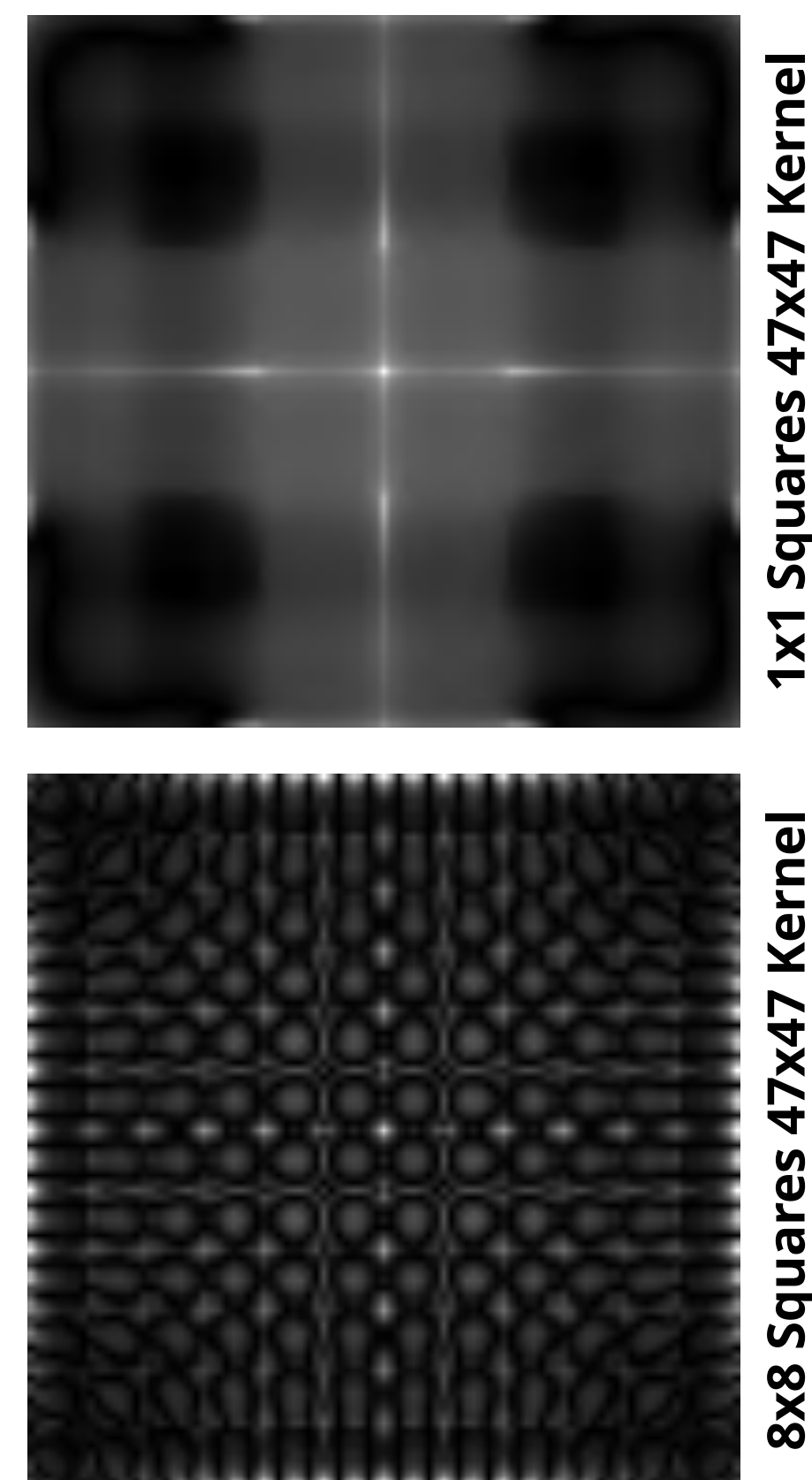
$$S_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})$$

“Order triplets according to reconstruction loss / **visual distance**”

Changing The Distance Function

Learn Distance Functions That Will Disentangle Certain Types Of Data

- **To disentangle** the adversarial dataset we need to **change** away from a pixel-wise **distance function** / loss
 - **Augment** the image **inside** the **distance function**.
- Optimise **spearman rank correlation** between factor distances and data visual distances
 - How well the latent ordering corresponds to the ground truth ordering

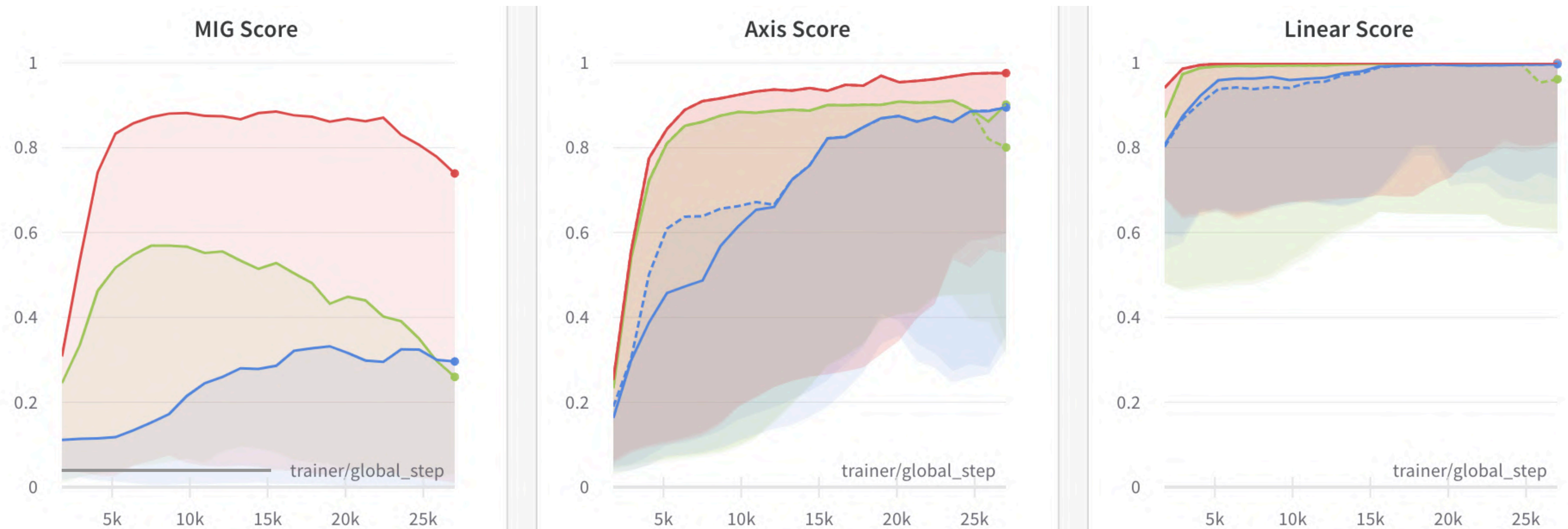


Unsupervised Adaptive Triplet
1x1 Squares 47x47 Kernel

NB decoder just trained for visualisation

Introducing Overlap

Results With Different Augmentations Before Distance Function



XY1 Kernel
Box Kernel
Gaussian Kernel
~ No Kernel

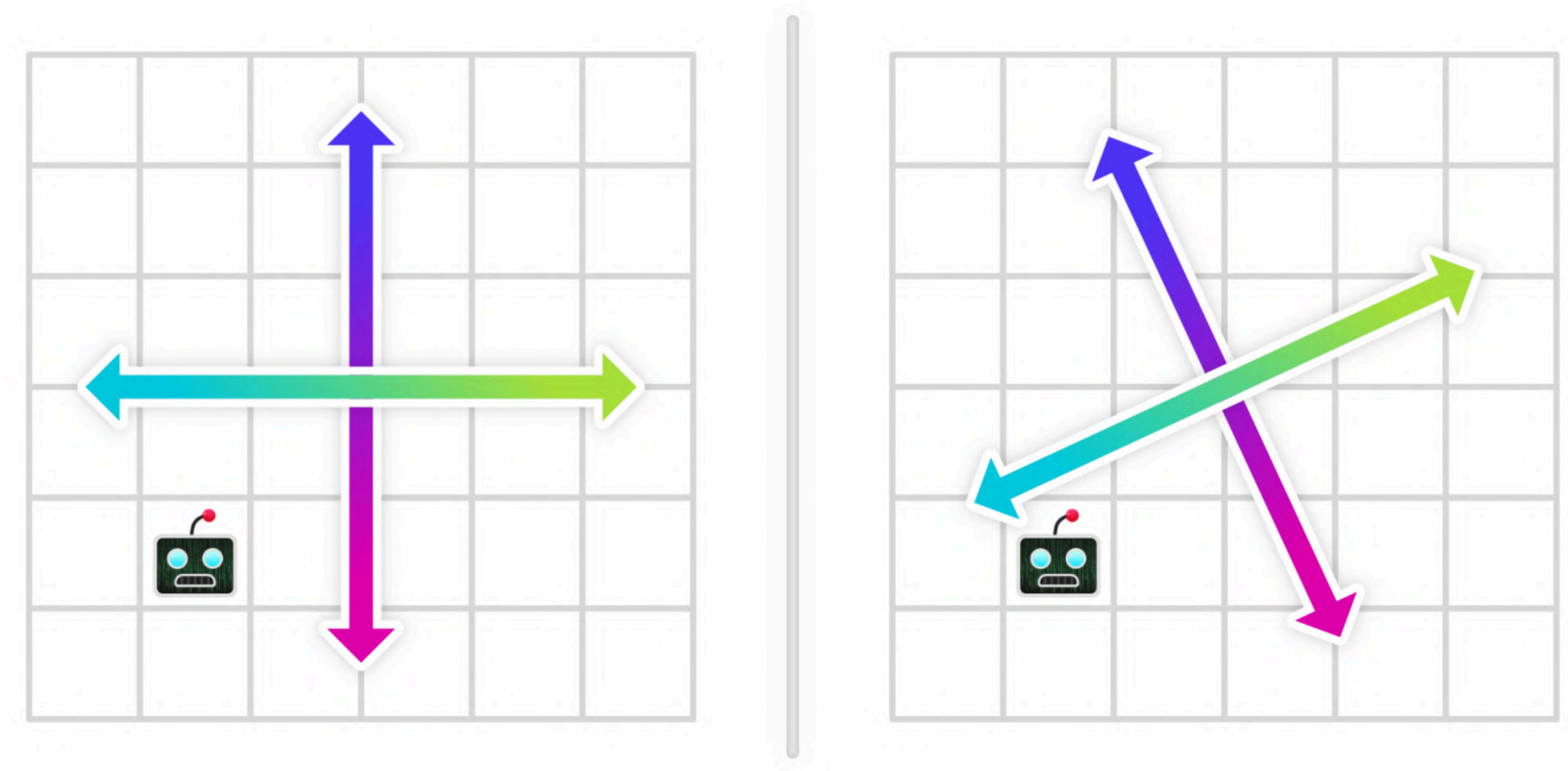
7. Take Aways

Take Aways

How Can We Apply All Of This?

Disentanglement With VAEs is not always what it seems.

- You need to **introduce biases** to disentangle
- VAEs find **axes in data**, but these **may not correspond to ground truth factors**. This is by chance with synthetic datasets
- VAEs often **fail on real world data**

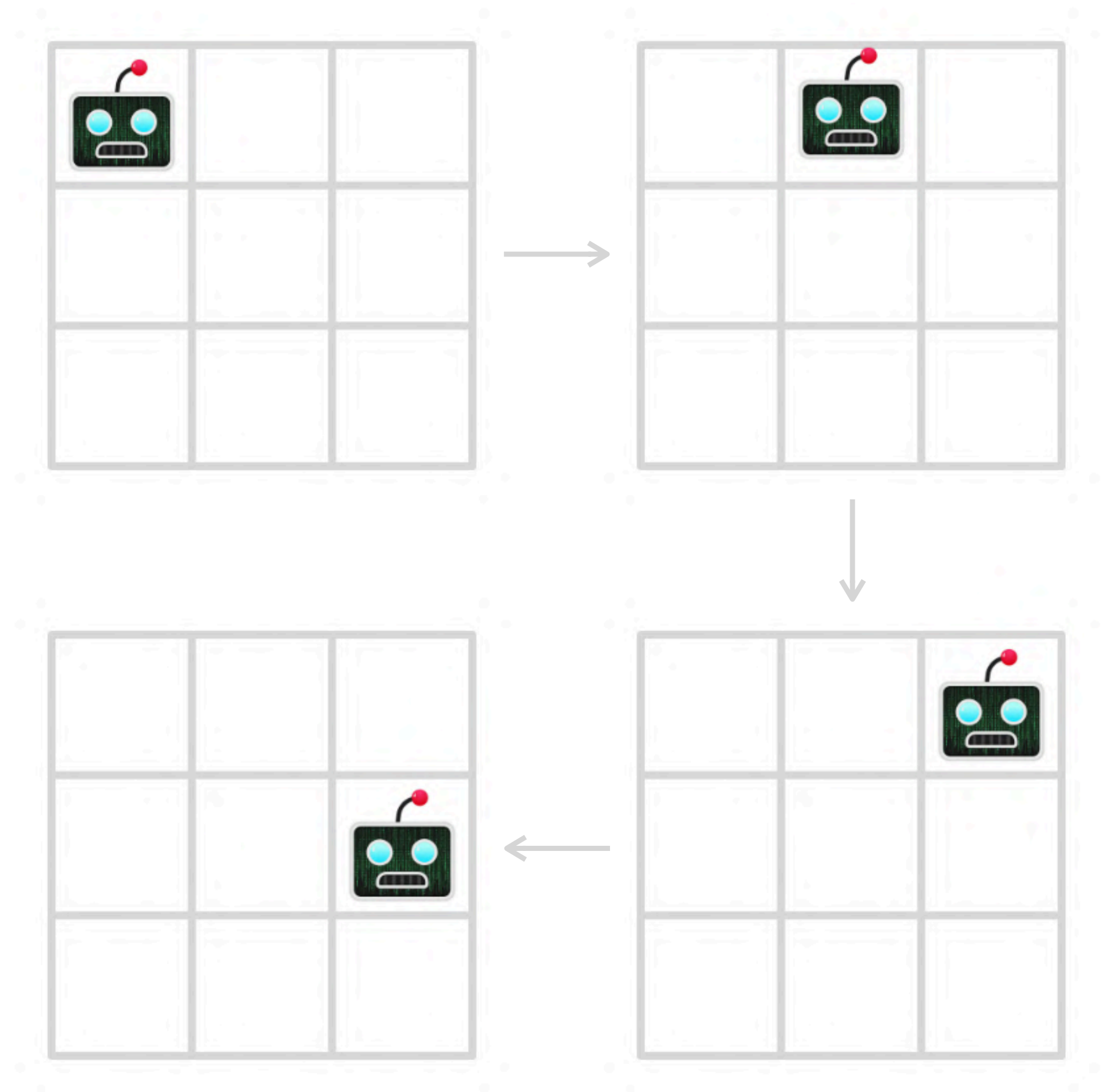


Take Aways

How Can We Apply All Of This?

Be careful of the data you feed your models

- E.g. Reinforcement learning: consider **not skipping frames** if we need to learn a representation off of them. Missed **overlap may be important**.
- Can't expect to disentangle if the ground truth factors you want don't align with the distances in the data



Take Aways

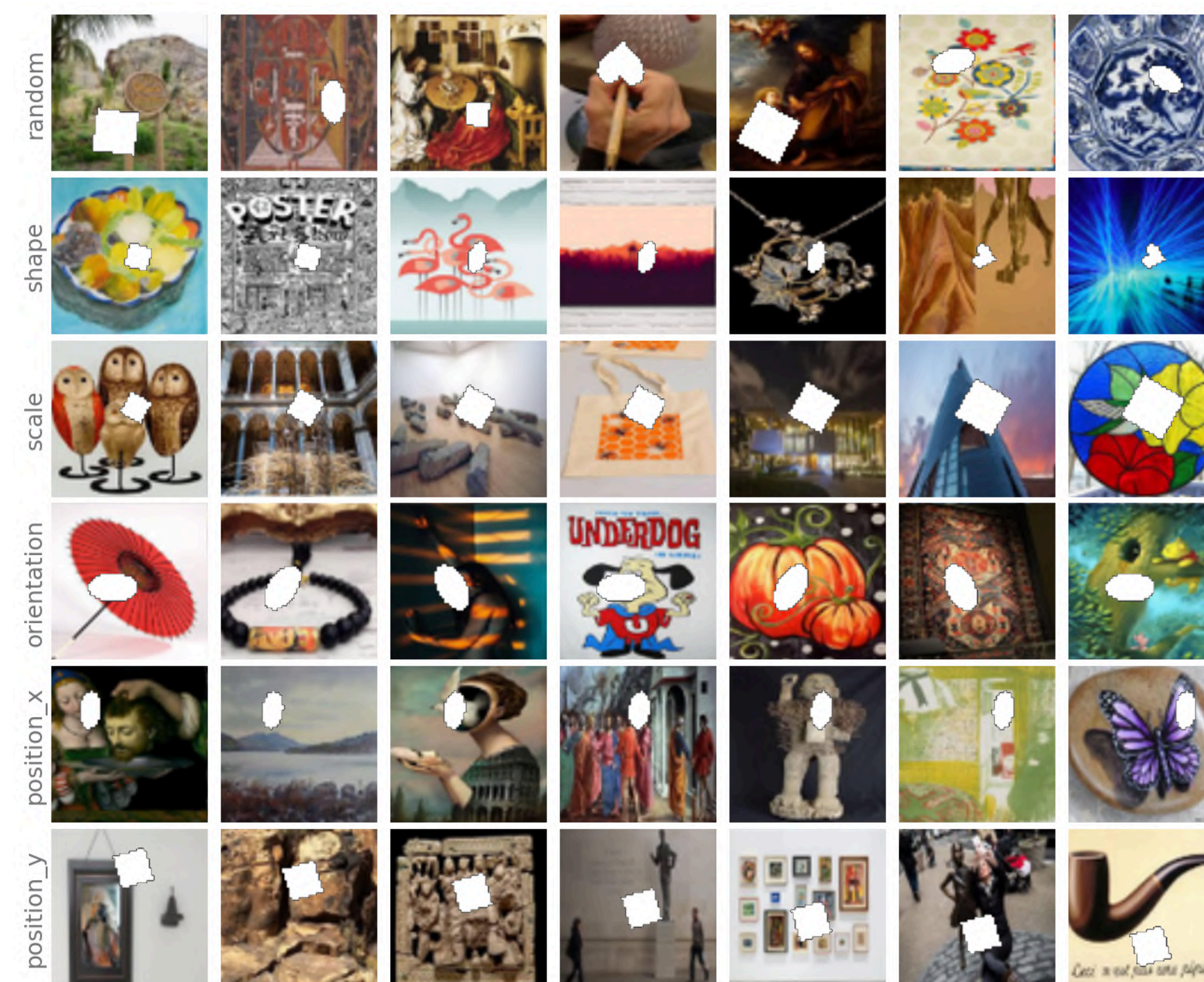
If You Have Extra Information Available, Use it!

Use available information!

- If you have **sequences** of frames, **or** **location** information. You can use this to better learn representations
- **Supervision:** E.g. Adaptive triplet

VAEs might not be the best solution for your problem.

- Consider contrastive learning, eg. SimCLR



∞. **The End**