# Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Francesco Locatello [1 2] Stefan Bauer [2] Mario Lucic [3] Gunnar Rätsch [1] Sylvain Gelly [3] Bernhard Schölkopf [2]
Olivier Bachem [3]

1. ETH Zurich, Department for Computer Science
2. Max Planck Institute for Intelligent Systems
3. Google Research, Brain Team

ICML 2019

Hanqing Chao

# Representation Learning Problem Definition

## Step 1

- A multivariate latent random variable $\boldsymbol{z}$ is **sampled from** a distribution $P(\boldsymbol{z})$.
- Intuitively, $\boldsymbol{z}$ corresponds to **semantically meaningful factors** of variation of the observations (e.g., content + position of objects in an image).

## Step 2

- The observation $\mathbf{x}$ is **sampled from** the conditional distribution $P(\mathbf{x}|\boldsymbol{z})$.

## Goal

- Find useful transformations $r(\mathbf{x})$ of $\mathbf{x}$ that "make it easier to extract useful information when building classifiers or other predictors" (Bengio et al., 2013).

arxiv: 1811.12359

# Definition of Disentanglement *informal*

A change in a **single underlying factor** of variation $z_i$ should lead to a change in a **single factor** in the learned representation $r(\mathbf{x})$.

arxiv: 1811.12359

# Frequency VS Bayesian

Maximum Likelihood Estimation

$$p(x^{(1)}, \cdots, x^{(m)} | \theta)$$

# Frequency VS **Bayesian**

Maximum Posterior Estimation

$$p(\theta|x^{(1)},\cdots,x^{(m)}) = \frac{p(x^{(1)},\cdots,x^{(m)}|\theta)p(\theta)}{p(x^{(1)},\cdots,x^{(m)})}$$

On Test

$$p(x^{(m+1)}|x^{(1)},\cdots,x^{(m)}) = \int p(x^{(m+1)}|\theta)p(\theta|x^{(1)},\cdots,x^{(m)})d\theta$$

# VAE

- Assume a specific prior $P(\boldsymbol{z})$ on the latent space.
- Parameterize the conditional probability $P(\mathbf{x}|\boldsymbol{z})$ using a DNN. (Decoder)
- The distribution $P(\boldsymbol{z}|\mathbf{x})$ is approximated using a variational distribution $Q(\boldsymbol{z}|\mathbf{x})$ ($Q(\boldsymbol{z}|\mathrm{x}_i)$). (Encoder)

The representation for $r(\mathbf{x})$ is usually taken to be the **mean** of the approximate posterior distribution $Q(\boldsymbol{z}|\mathbf{x})$.

arxiv: 1811.12359

# Disentanglement

Try to enforce a factorized aggregated

posterior $\int_{\mathbf{x}} Q(\boldsymbol{z}|\mathbf{x})P(\mathbf{x})\, d\mathbf{x}$,

arxiv: 1811.12359

# IMPOSSIBLE

**Theorem 1.** *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\boldsymbol{u})}{\partial u_j} \neq 0$ almost everywhere for all $i$ and $j$ (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

arxiv: 1811.12359

# IMPOSSIBLE

- Assume we have $p(z)$ and some $P(x|z)$ defining a generative model.
- Consider any unsupervised disentanglement method and assume that it finds a representation $r(x)$ that is perfectly disentangled with respect to $z$ in the generative model.
- Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{z} = f(z)$ where $\hat{z}$ is completely entangled with respect to $z$ and thus also $r(x)$
- Furthermore, since f is deterministic and $p(z) = p(\hat{z})$ almost everywhere, both generative models have the same marginal distribution of the observations x by construction, i.e.

$$P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

The (unsupervised) disentanglement method only has access to observations x, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

arxiv: 1811.12359

# IMPOSSIBLE

- Assume we have $p(z)$ and some $P(x|z)$ defining a generative model.
- Consider any unsupervised disentanglement method and assume that it finds a representation $r(x)$ that is perfectly disentangled with respect to $z$ in the generative model.
- Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{z} = f(z)$ where $\hat{z}$ is completely entangled with respect to $z$ and thus also $r(x)$
- Furthermore, since f is deterministic and $p(z) = p(\hat{z})$ almost everywhere, both generative models have the same marginal distribution of the observations x by construction, i.e.

$$P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

The (unsupervised) disentanglement method only has access to observations x, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

arxiv: 1811.12359

# IMPOSSIBLE

- Assume we have $p(z)$ and some $P(x|z)$ defining a generative model.
- Consider any unsupervised disentanglement method and assume that it finds a representation $r(x)$ that is perfectly disentangled with respect to $z$ in the generative model.
- Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{z} = f(z)$ where $\hat{z}$ is completely entangled with respect to $z$ and thus also $r(x)$
- Furthermore, since f is deterministic and $p(z) = p(\hat{z})$ almost everywhere, both generative models have the same marginal distribution of the observations x by construction, i.e.

$$P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

The (unsupervised) disentanglement method only has access to observations x, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

arxiv: 1811.12359

# IMPOSSIBLE

- Assume we have $p(z)$ and some $P(x|z)$ defining a generative model.
- Consider any unsupervised disentanglement method and assume that it finds a representation $r(x)$ that is perfectly disentangled with respect to $z$ in the generative model.
- Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{z} = f(z)$ where $\hat{z}$ is completely entangled with respect to $z$ and thus also $r(x)$
- Furthermore, since f is deterministic and $p(z) = p(\hat{z})$ almost everywhere, both generative models have the same marginal distribution of the observations x by construction, i.e.

$$P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

The (unsupervised) disentanglement method only has access to observations x, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

arxiv: 1811.12359

# IMPOSSIBLE

We need **inductive biases** on **both** datasets and models

arxiv: 1811.12359

# Experiments

- 6 models
- 6 disentangle measurements
- 50 random initial
- 7 datasets
- 12,000 models
- https://github.com/google-research/disentanglement_lib

arxiv: 1811.12359

# Experiments

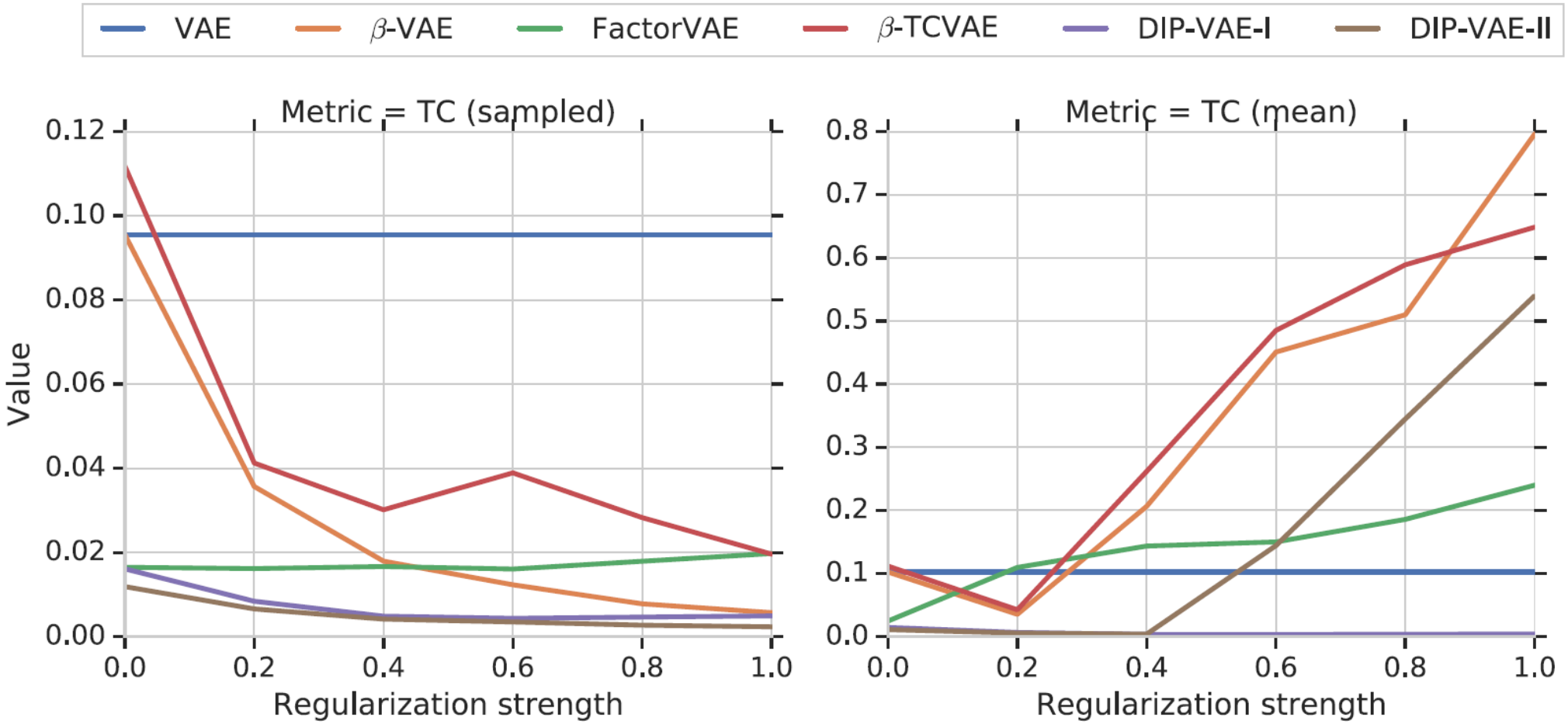
dSprites is a dataset of 2D shapes procedurally generated from 6 ground truth independent latent factors. These factors are *color*, *shape*, *scale*, *rotation*, *x* and *y* positions of a sprite.

All possible combinations of these latents are present exactly once, generating N = 737280 total images.

## Latent factor values

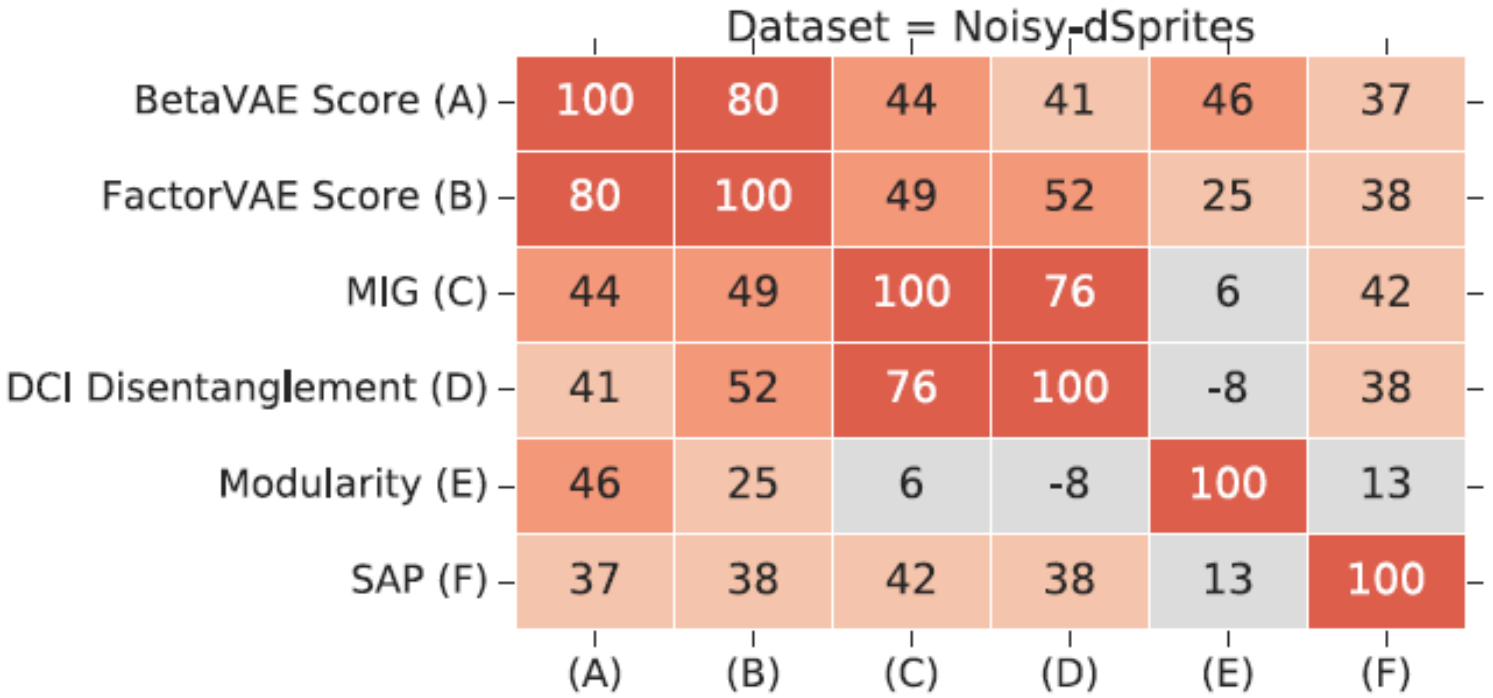
- Color: white
- Shape: square, ellipse, heart
- Scale: 6 values linearly spaced in [0.5, 1]
- Orientation: 40 values in [0, 2 pi]
- Position X: 32 values in [0, 1]
- Position Y: 32 values in [0, 1]

# Can current methods enforce a uncorrelated aggregated posterior and representation?



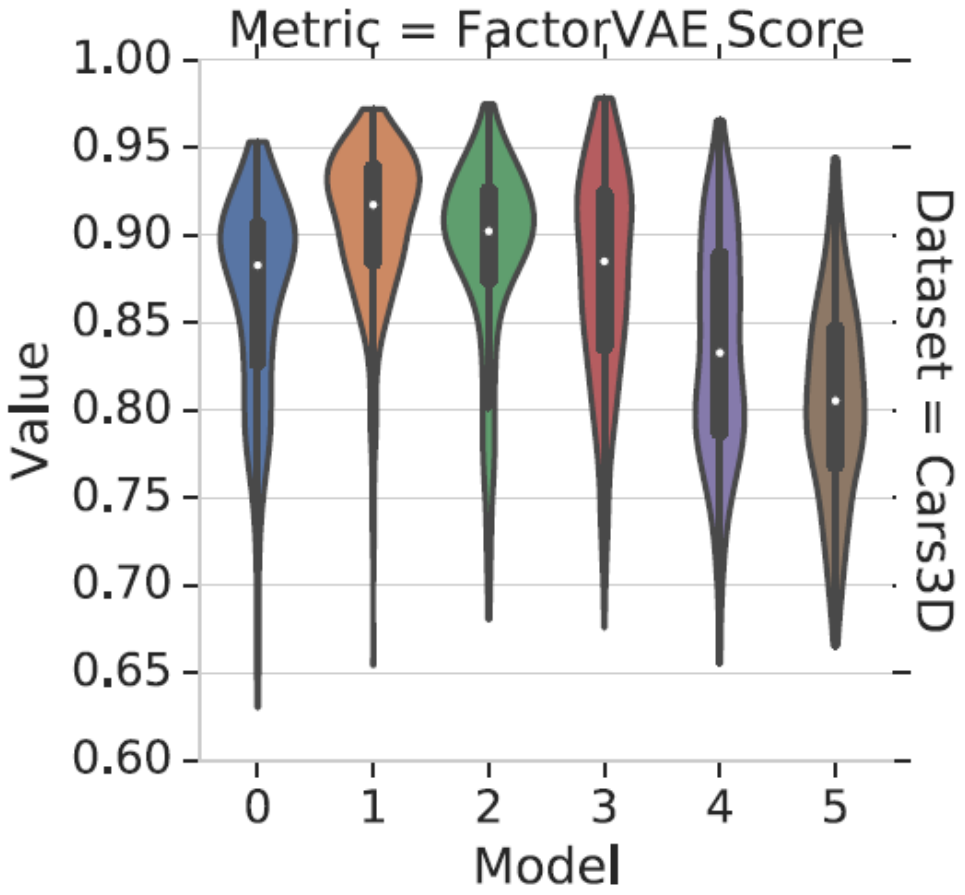

arxiv: 1811.12359

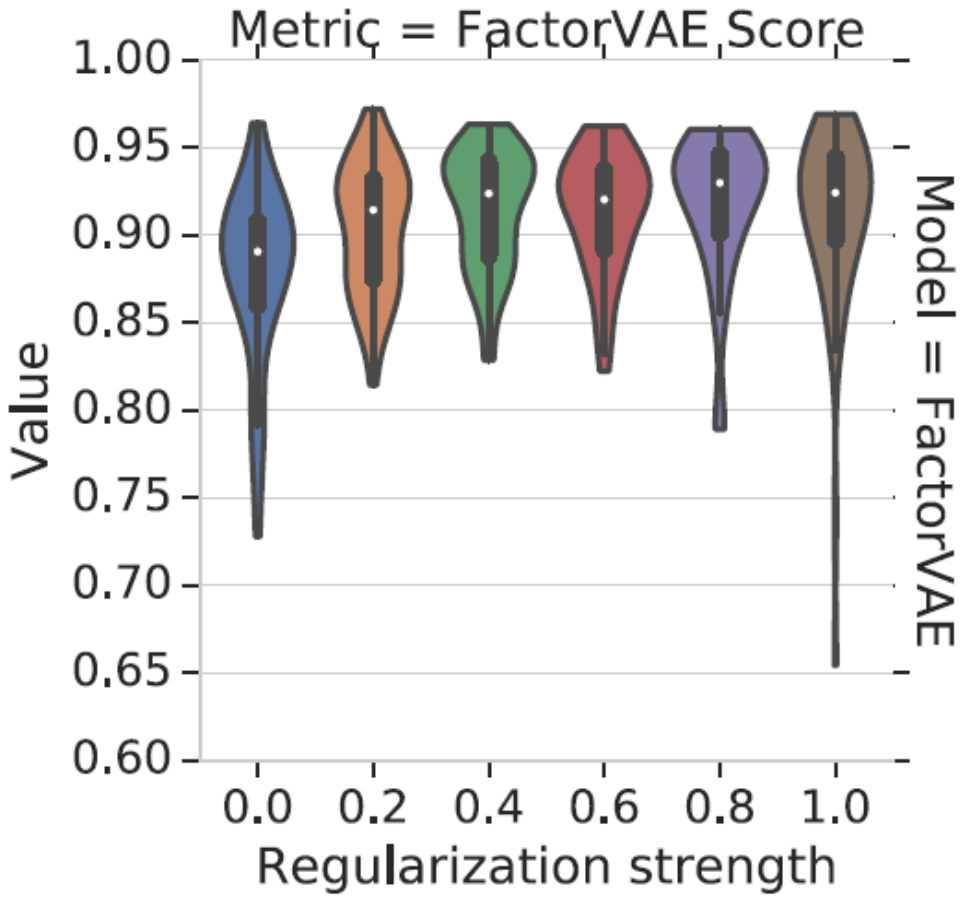
# How much do the disentanglement metrics agree?




arxiv: 1811.12359

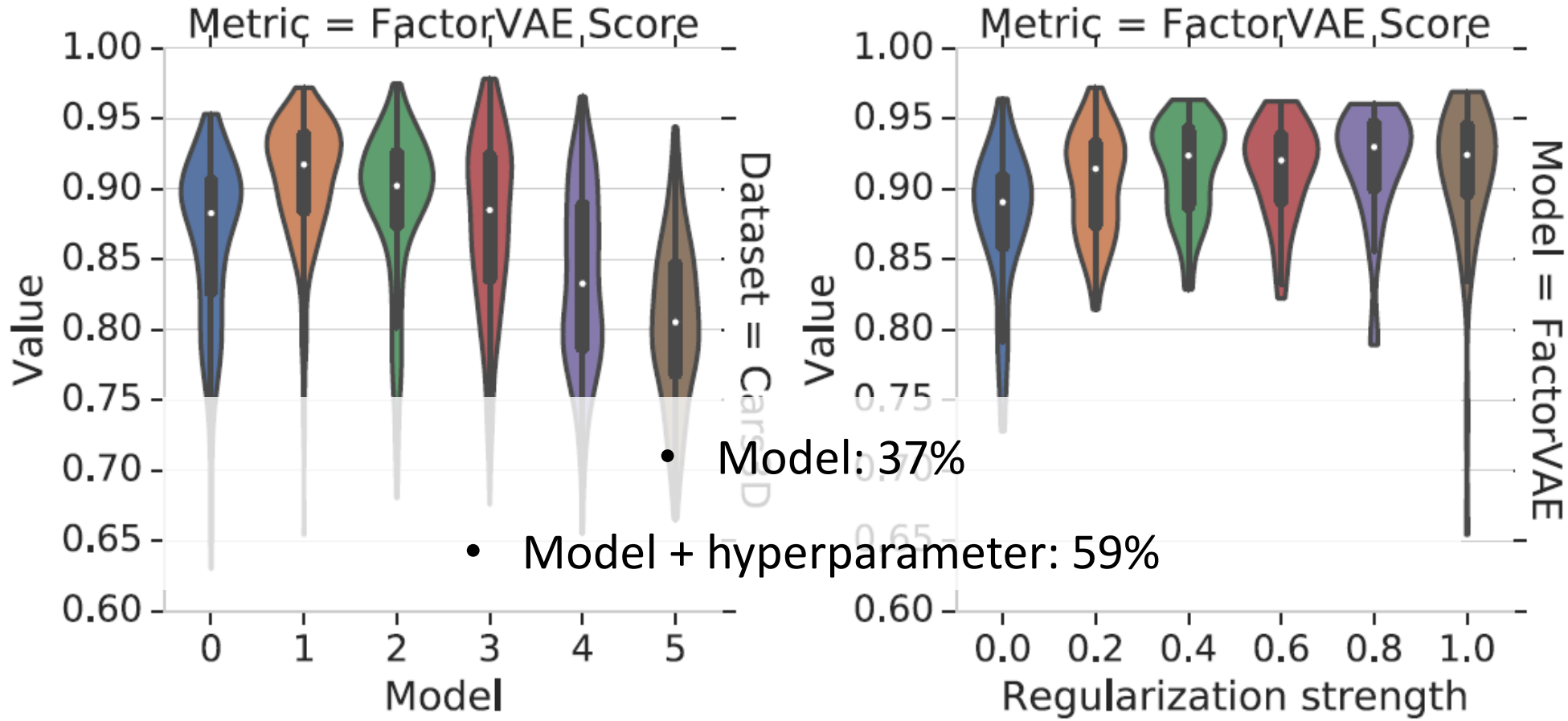
# How important are different models and hyperparameters for disentanglement?




arxiv: 1811.12359

# How important are different models and hyperparameters for disentanglement?

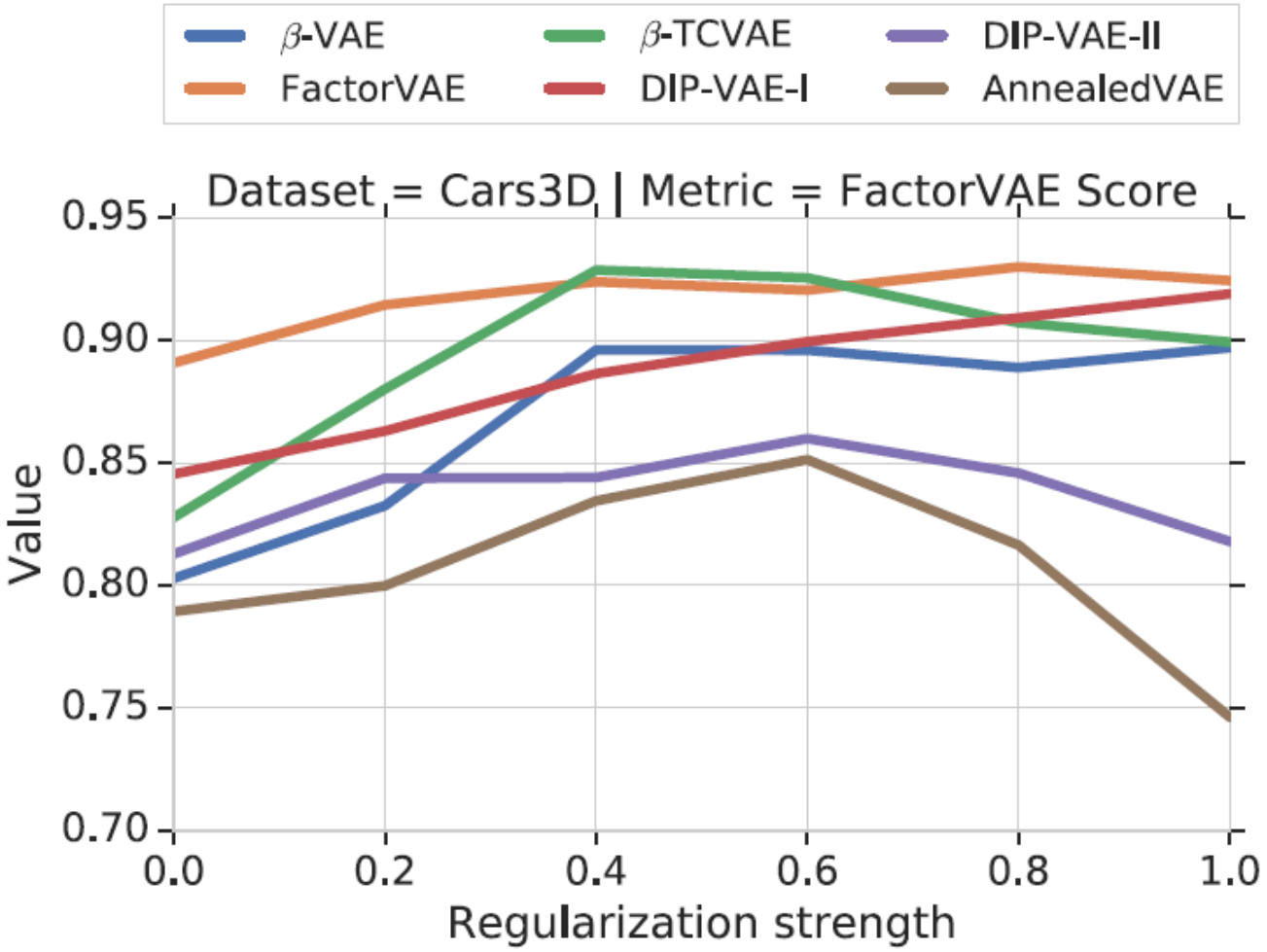



- Model: 37%
- Model + hyperparameter: 59%

arxiv: 1811.12359

# Are there reliable recipes for model selection?

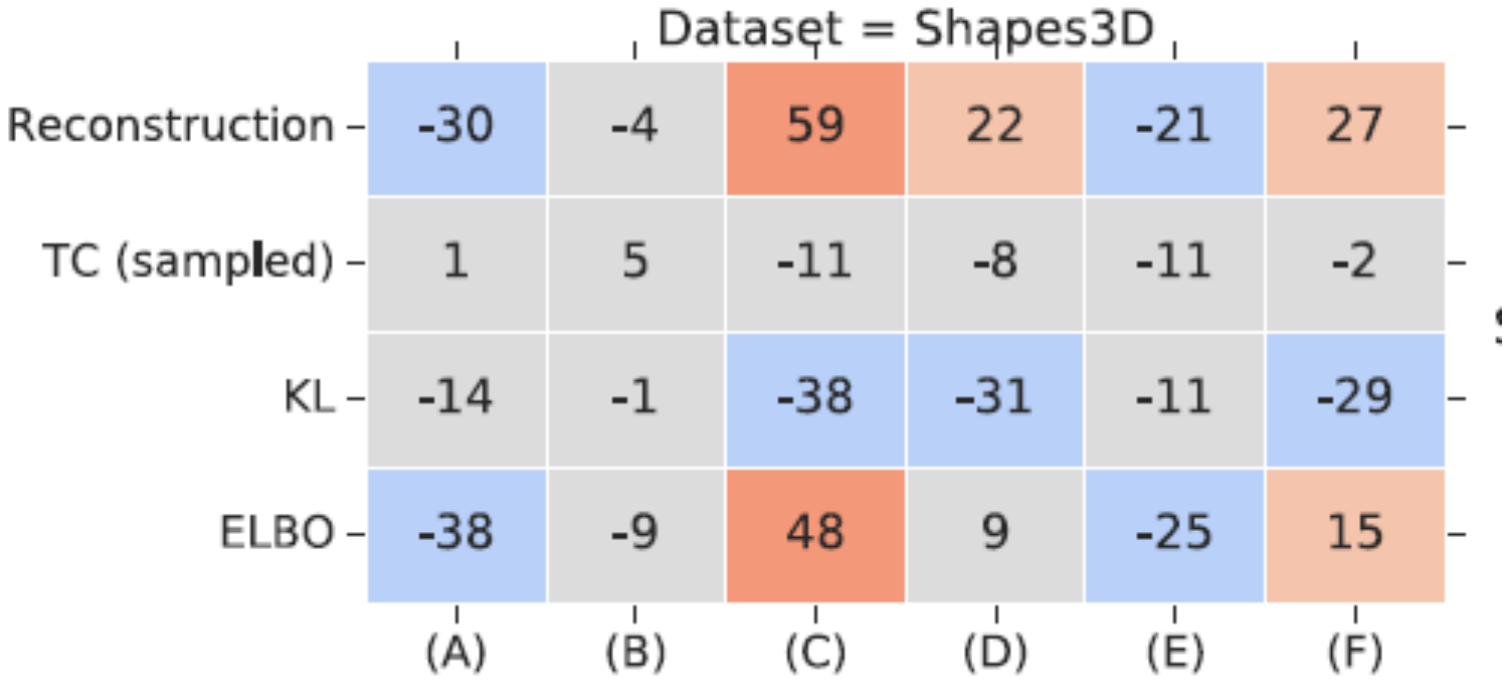
*General recipes for hyperparameter selection.*




arxiv: 1811.12359

# Are there reliable recipes for model selection?

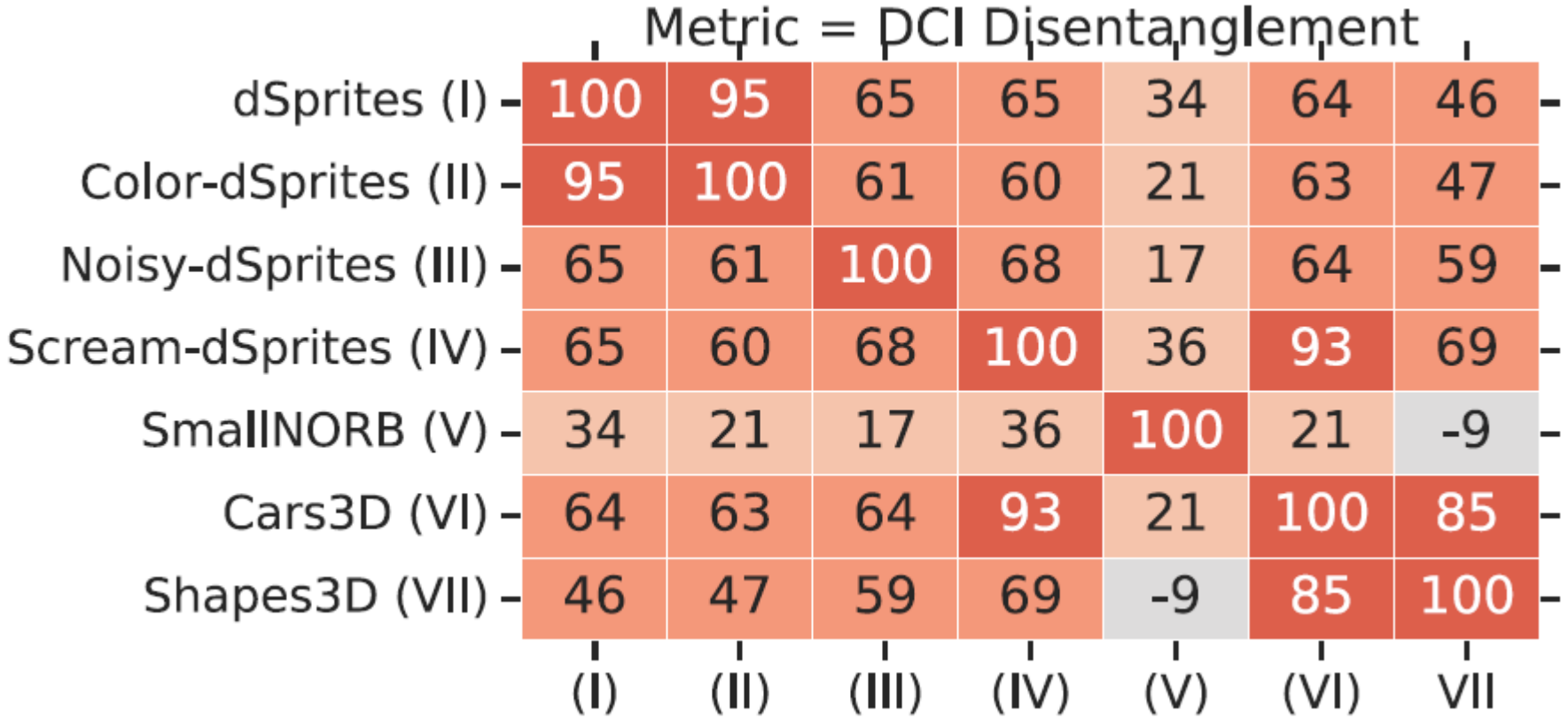
*Model selection based on unsupervised scores.*




arxiv: 1811.12359

# Are there reliable recipes for model selection?

*Hyperparameter selection based on transfer.*




arxiv: 1811.12359

# Are there reliable recipes for model selection?

- Unsupervised model selection remains an unsolved problem.

arxiv: 1811.12359

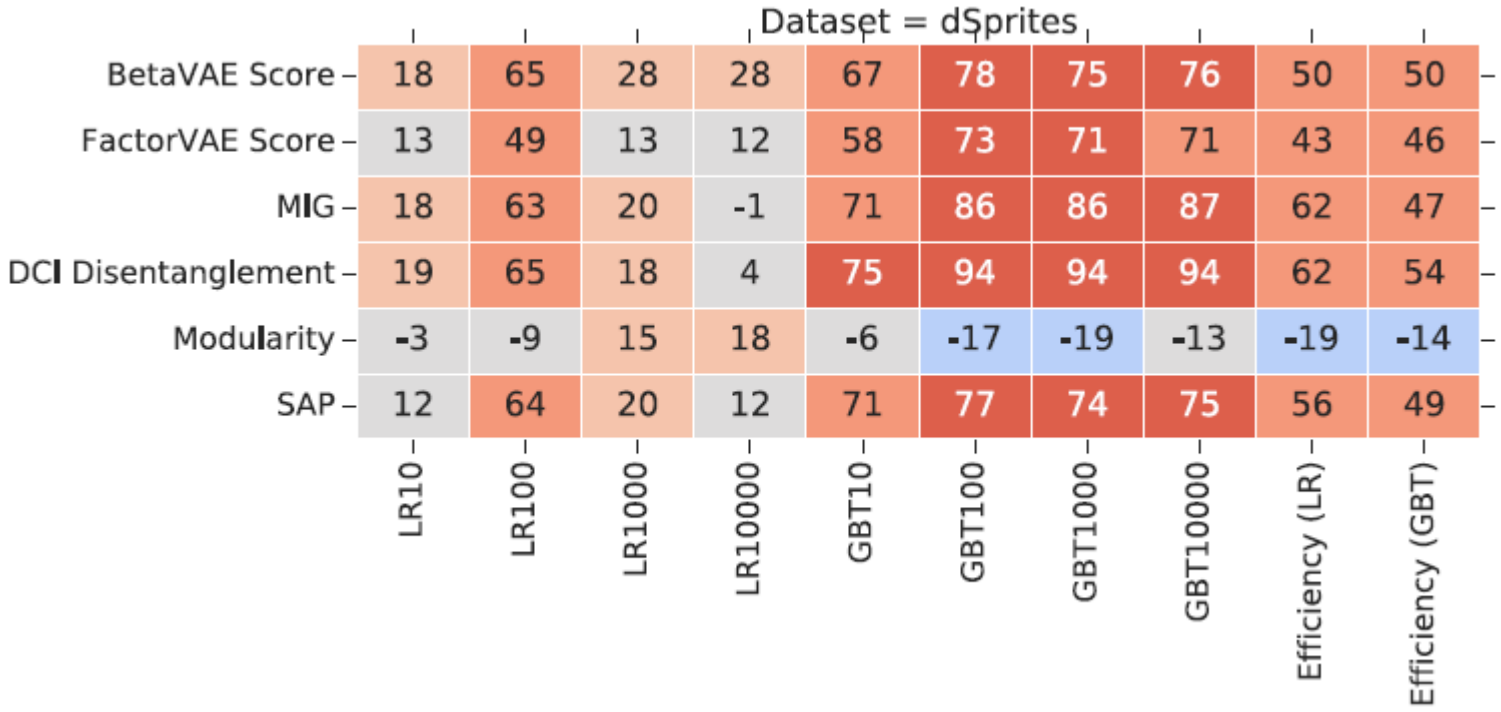
# Are these disentangled representations useful for downstream tasks in terms of the sample complexity of learning?

**Goal**

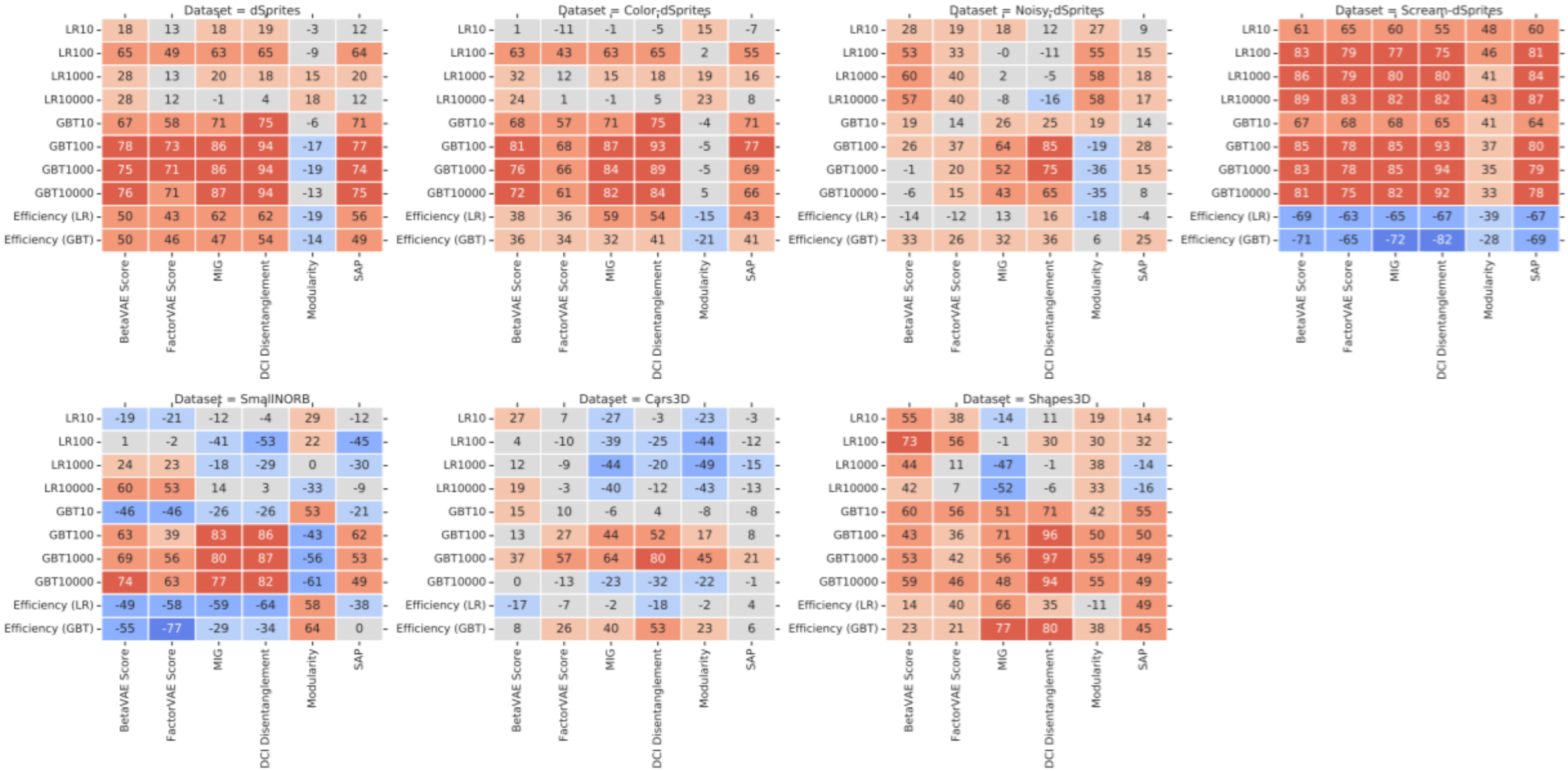
To **recover the true factors** of variations from the learned representation using either multi-class logistic regression (LR) or gradient boosted trees (GBT).

arxiv: 1811.12359

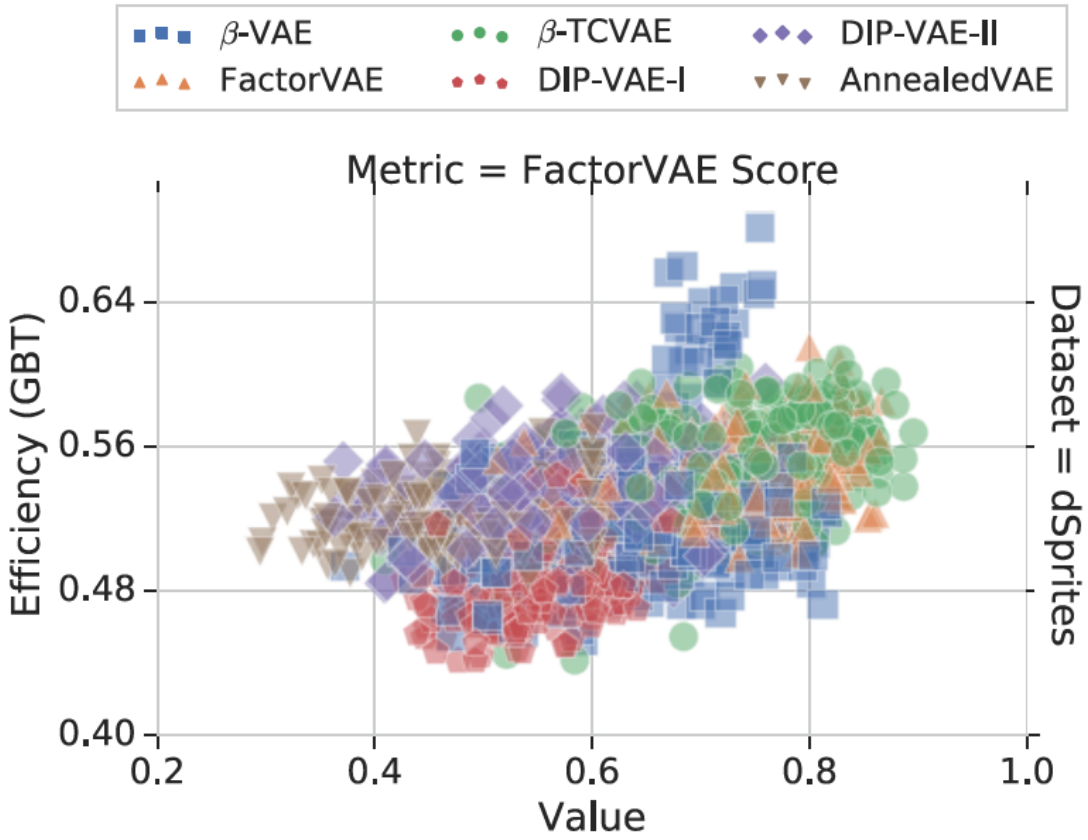
# Are these disentangled representations useful for downstream tasks in terms of the sample complexity of learning?




arxiv: 1811.12359


Dataset = dSprites
LR10 18 13 18 19 -3 12
LR100 65 49 63 65 -9 64
LR1000 28 13 20 18 15 20
LR10000 28 12 -1 4 18 12
GBT10 67 58 71 75 -6 71
GBT100 78 73 86 94 -17 77
GBT1000 75 71 86 94 -19 74
GBT10000 76 71 87 94 -13 75
Efficiency (LR) 50 43 62 62 -19 56
Efficiency (GBT) 50 46 47 54 -14 49
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = Color-dSprites
LR10 1 -11 -1 -5 15 -7
LR100 63 43 63 65 2 55
LR1000 32 12 15 18 19 16
LR10000 24 1 -1 5 23 8
GBT10 68 57 71 75 -4 71
GBT100 81 68 87 93 -5 77
GBT1000 76 66 84 89 -5 69
GBT10000 72 61 82 84 5 66
Efficiency (LR) 38 36 59 54 -15 43
Efficiency (GBT) 36 34 32 41 -21 41
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = Noisy-dSprites
LR10 28 19 18 12 27 9
LR100 53 33 -0 -11 55 15
LR1000 60 40 2 -5 58 18
LR10000 57 40 -8 -16 58 17
GBT10 19 14 26 25 19 14
GBT100 26 37 64 85 -19 28
GBT1000 -1 20 52 75 -36 15
GBT10000 -6 15 43 65 -35 8
Efficiency (LR) -14 -12 13 16 -18 -4
Efficiency (GBT) 33 26 32 36 6 25
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = Scream-dSprites
LR10 61 65 60 55 48 60
LR100 83 79 77 75 46 81
LR1000 86 79 80 80 41 84
LR10000 89 83 82 82 43 87
GBT10 67 68 68 65 41 64
GBT100 85 78 85 93 37 80
GBT1000 83 78 85 94 35 79
GBT10000 81 75 82 92 33 78
Efficiency (LR) -69 -63 -65 -67 -39 -67
Efficiency (GBT) -71 -65 -72 -82 -28 -69
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = SmallNORB
LR10 -19 -21 -12 -4 29 -12
LR100 1 -2 -41 -53 22 -45
LR1000 24 23 -18 -29 0 -30
LR10000 60 53 14 3 -33 -9
GBT10 -46 -46 -26 -26 53 -21
GBT100 63 39 83 86 -43 62
GBT1000 69 56 80 87 -56 53
GBT10000 74 63 77 82 -61 49
Efficiency (LR) -49 -58 -59 -64 58 -38
Efficiency (GBT) -55 -77 -29 -34 64 0
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = Cars3D
LR10 27 7 -27 -3 -23 -3
LR100 4 -10 -39 -25 -44 -12
LR1000 12 -9 -44 -20 -49 -15
LR10000 19 -3 -40 -12 -43 -13
GBT10 15 10 -6 4 -8 -8
GBT100 13 27 44 52 17 8
GBT1000 37 57 64 80 45 21
GBT10000 0 -13 -23 -32 -22 -1
Efficiency (LR) -17 -7 -2 -18 -2 4
Efficiency (GBT) 8 26 40 53 23 6
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP
Dataset = Shapes3D
LR10 55 38 -14 11 19 14
LR100 73 56 -1 30 30 32
LR1000 44 11 -47 -1 38 -14
LR10000 42 7 -52 -6 33 -16
GBT10 60 56 51 71 42 55
GBT100 43 36 71 96 50 50
GBT1000 53 42 56 97 55 49
GBT10000 59 46 48 94 55 49
Efficiency (LR) 14 40 66 35 -11 49
Efficiency (GBT) 23 21 77 80 38 45
BetaVAE Score
FactorVAE Score
MIG
DCI Disentanglement
Modularity
SAP

# Are these disentangled representations useful for downstream tasks in terms of the sample complexity of learning?




arxiv: 1811.12359

# Conclusion

- Inductive biases and implicit and explicit supervision. Deviates from the static, purely unsupervised setting.
- Concrete practical benefits of disentangled representations.
- Experimental setup and diversity of data sets. it is easy to draw spurious conclusions from experimental results if one only considers a subset of methods, metrics and data sets.

arxiv: 1811.12359

# Conclusion

- Inductive biases and implicit and explicit supervision. Deviates from the static, purely unsupervised setting.
- Concrete practical benefits of disentangled representations.
- Experimental setup and diversity of data sets. it is easy to draw spurious conclusions from experimental results if one only considers a subset of methods, metrics and data sets.

arxiv: 1811.12359

# Conclusion

- Inductive biases and implicit and explicit supervision. Deviates from the static, purely unsupervised setting.
- Concrete practical benefits of disentangled representations.
- Experimental setup and diversity of data sets. it is easy to draw spurious conclusions from experimental results if one only considers a subset of methods, metrics and data sets.

arxiv: 1811.12359

# Thanks！