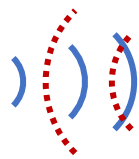
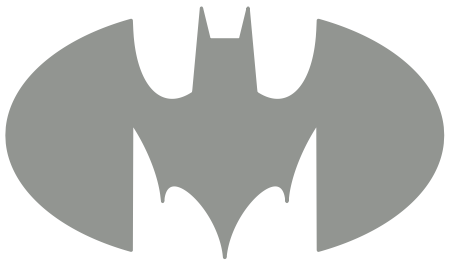


# Exact Information Bottlenecks for Arbitrary Distributions with Echo Noise

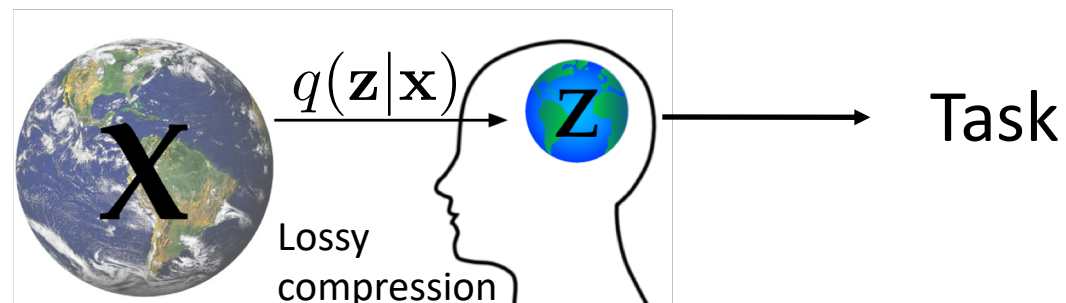


Rob Brekelmans, Daniel Moyer,  
**Greg Ver Steeg**, Aram Galstyan



Feb. 11, 2019  
Crossroads of Physics & ML

# Information bottleneck / rate-distortion



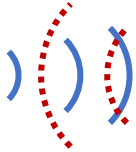
Given observed distribution  $q(X=x, T=t)$ , for inputs  $X$  and task variable  $T$ .  
The Lagrangian form:

$$\min_{q(\mathbf{z}|\mathbf{x})} \text{Distortion}(Z, T) + \beta \underbrace{I(Z; X)}_{\text{Mutual information / compression rate}} \quad \text{???$$

e.g. prediction error

Only for toy cases: low-d categorical variables and Gaussian noise/inputs

# Outline

- **Rate bounds:** mean-field variational bounds for information rate
- **Echo:** a noisy channel with exact rate 
- **Application:** variational auto-encoders as information bottlenecks
- **Results:** better likelihood and rate-distortion trade-offs
- **Controversy:** Do deep nets compress? Does this explain generalization?

Mean-field variational bound for  
information rate

# Information in a noisy channel

Input distribution  $q(\mathbf{x})$   $\xrightarrow{\text{Noisy channel}}$   $\mathbf{z}$   $q(\mathbf{z}|\mathbf{x})$   $q(\mathbf{z}, \mathbf{x}) = q(\mathbf{z}|\mathbf{x})q(\mathbf{x})$

Mutual information

$$\begin{aligned} I(Z; X) &= H(Z) - H(Z|X) \\ &= D_{KL} [q(\mathbf{z}|\mathbf{x}) || q(\mathbf{z})] \\ &= \mathbb{E}_q \log \frac{q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} \end{aligned}$$

# Problem

$$q(\mathbf{z}) = \int d\mathbf{x} q(\mathbf{z}|\mathbf{x})q(\mathbf{x})$$

High-d  
integral

Could be complex  
(images, audio,  
gene expression...)

# Mean-field variational approximation, $p(\mathbf{z})$

Mutual  
information

$$\begin{aligned} I(Z; X) &= \mathbb{E}_q \left[ \log \frac{q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} \right] \\ &= \mathbb{E}_q \left[ \log \frac{q(\mathbf{z}|\mathbf{x}) p(\mathbf{z})}{q(\mathbf{z}) p(\mathbf{z})} \right] \\ &= D_{KL} [q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})] - D_{KL} [q(\mathbf{z}) || p(\mathbf{z})] \quad (\text{Gibb's inequality}) \\ &\leq D_{KL} [q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})] \end{aligned}$$

$$p(\mathbf{z}) = \prod_i p(z_i) \quad (\text{A typical mean-field approximation})$$

# Mean-field example for Gaussian noise channel

$$z_i = x_i + s\varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, 1)$$

$$q(z_i|x_i) = \mathcal{N}(x_i, s^2)$$

An optimal variational distribution,  $p$ , would be:  $p(z_i) = \mathcal{N}(\mathbb{E}(x_i), \text{Var}(x_i) + s^2)$

$$I(Z; X) \leq \sum_i \frac{1}{2} \log \left( 1 + \frac{\text{Var}(x_i)}{s^2} \right)$$

Only tight if the *input is Gaussian*, and *each channel is independent*

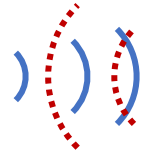


# Echo noise

By choosing a more flexible noise channel, we can exactly specify information rates for arbitrary inputs



# Echo noise: make the noise look like the signal



Why?

- For correlated Gaussian noise, optimal signal is correlated in same basis
- **Key property** for analytic mutual information under arbitrary input

How do we make the noise look like the signal?

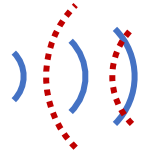
$$z = f(\mathbf{x}) + s\varepsilon$$

$$\varepsilon = f(\mathbf{x}'), \quad \mathbf{x}' \stackrel{iid}{\sim} q(\mathbf{x})$$

$$z = f(\mathbf{x}) + sf(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \stackrel{iid}{\sim} q(\mathbf{x})$$

$$\mathcal{E} \stackrel{d}{\neq} \mathcal{Z} \quad (\text{R.V.'s not equal in distribution})$$

# Echo noise: make the noise look like the signal



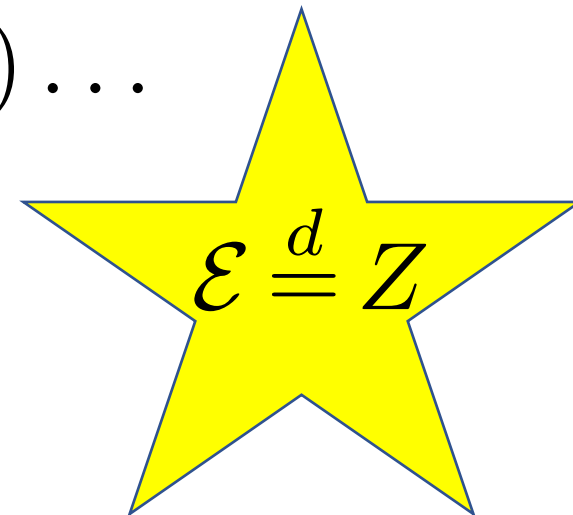
How do we make the noise look like the signal?

$$z = f(\mathbf{x}) + s\varepsilon$$

$$\varepsilon = f(\mathbf{x}^{(0)}) + sf(\mathbf{x}^{(1)}) + s^2 f(\mathbf{x}^{(2)}) \dots, \quad \mathbf{x}^{(\ell)} \stackrel{iid}{\sim} q(\mathbf{x})$$

$$z = f(\mathbf{x}) + s \left( f(\mathbf{x}^{(0)}) + sf(\mathbf{x}^{(1)}) + s^2 f(\mathbf{x}^{(2)}) \dots \right)$$

Multiply out and re-label iid samples....  
these are the same (in distribution)!

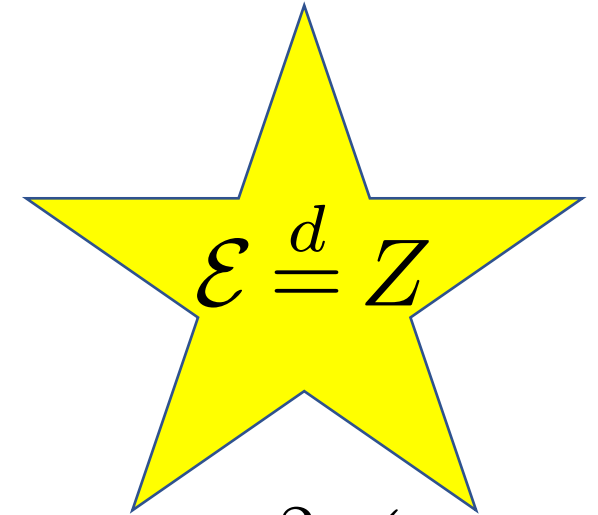


For vectors, with attenuation depending on  $\mathbf{x}$

$$\mathbf{z} = f(\mathbf{x}) + \underbrace{S(\mathbf{x})}_{\text{Attenuation matrix is a function of } \mathbf{x}} \varepsilon$$

Diagram: The word "vectors" is written above the equation. Three blue arrows point from "vectors" to  $\mathbf{z}$ ,  $f(\mathbf{x})$ , and  $S(\mathbf{x})$ . A blue bracket is drawn under  $S(\mathbf{x})$ .

Attenuation matrix is a function of  $\mathbf{x}$

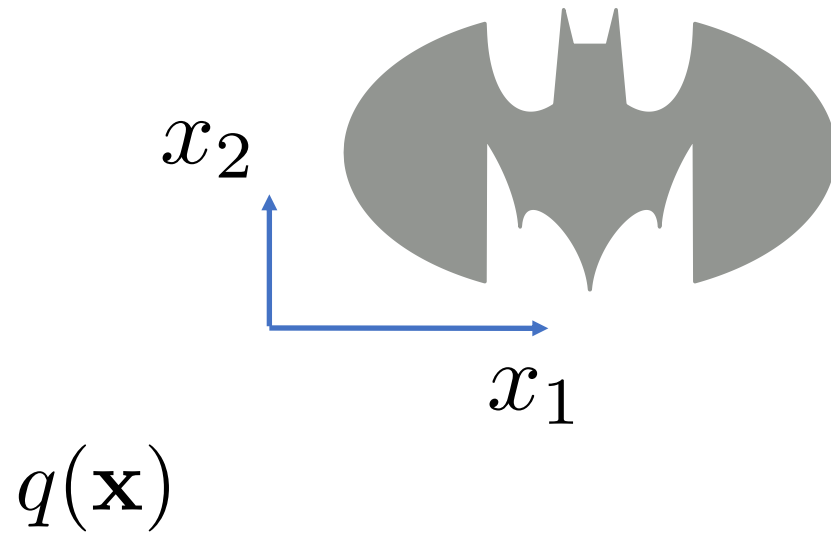


$$\varepsilon = f(\mathbf{x}^0) + S(\mathbf{x}^0) \left( f(\mathbf{x}^1) + S(\mathbf{x}^1) \left( f(\mathbf{x}^2) + S(\mathbf{x}^2) (\dots \right. \right.$$

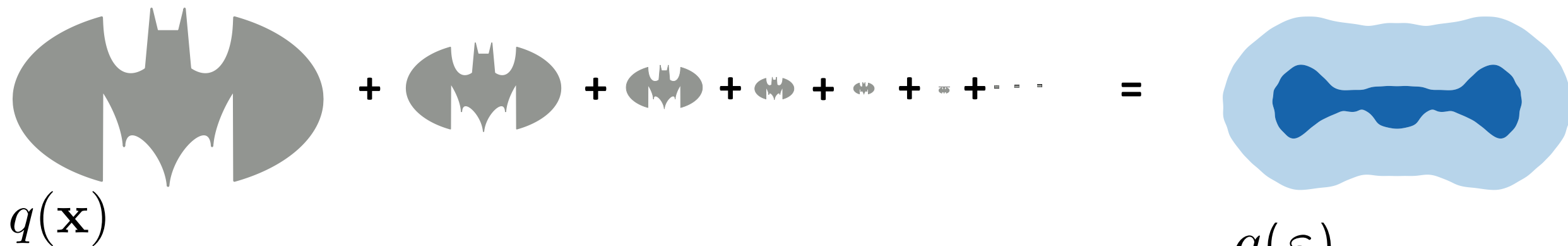
$$\varepsilon = \sum_{l=0}^{\infty} \left( \prod_{l'=1}^l S(\mathbf{x}^{l'}) \right) f(\mathbf{x}^l), \quad \mathbf{x}^{l'} \stackrel{iid}{\sim} q(\mathbf{x})$$

# Example: a non-Gaussian input distribution

A uniform distribution in  $\mathbb{R}^2$  with a shape that strikes fear into the heart of villains and Gaussians



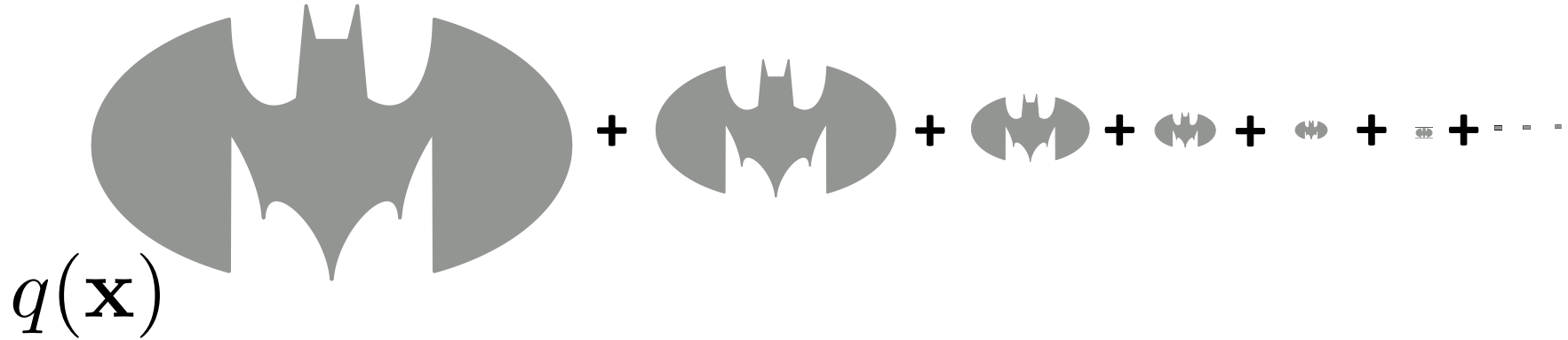
Echo example,  $z = x + s \varepsilon$ , with  $s = 1/2$



$$\varepsilon = \mathbf{x}^{(0)} + \frac{1}{2}\mathbf{x}^{(1)} + \frac{1}{4}\mathbf{x}^{(2)} + \frac{1}{8}\mathbf{x}^{(3)} + \dots$$



Sampling

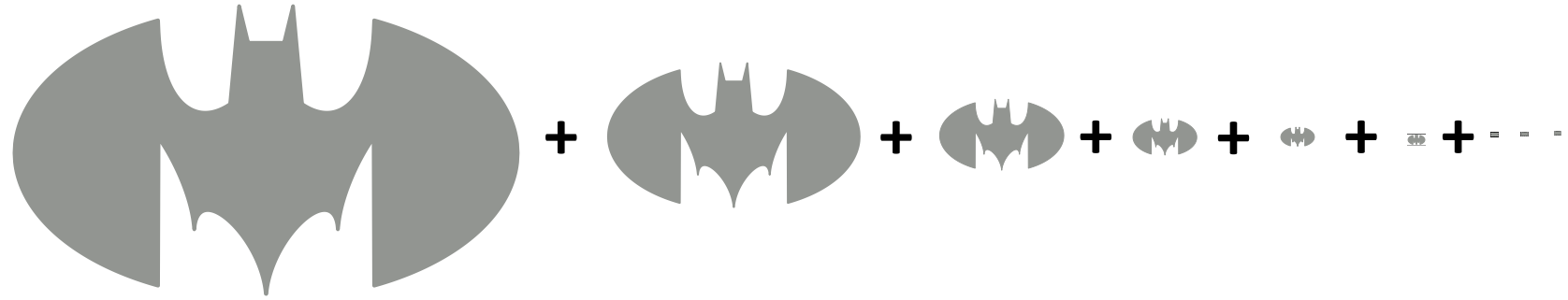


- The noise is data-driven. To sample noise, we just need iid samples from the input distribution
- For this sampling procedure to converge, we need  $s < 1$  (or spectral radius of  $S(x) < 1$  in vector case)
- We can sample noise to within machine precision by controlling number of terms in the series

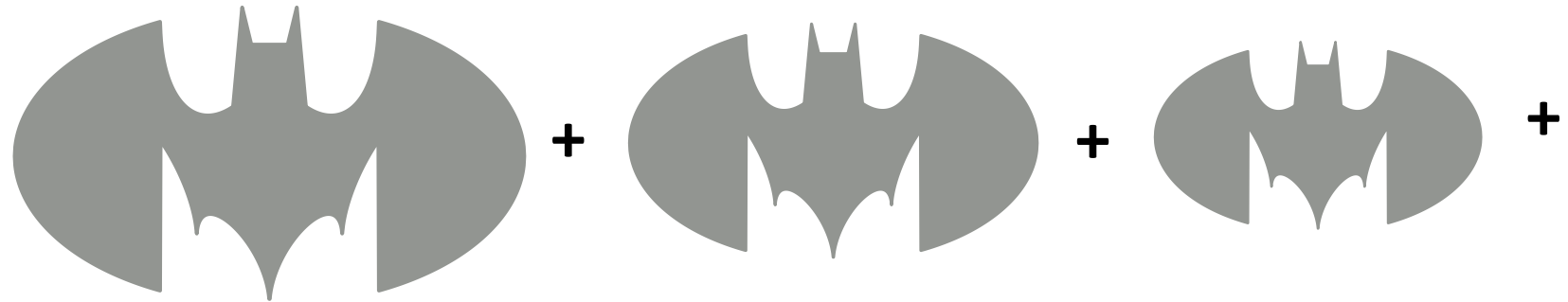
$$\varepsilon = \sum_{l=0}^{\infty} s^l f(\mathbf{x}^{(l)})$$

(caveat)

$s=1/2$



$s=4/5$



- We can sample noise to within machine precision by controlling number of terms in the series
- To get HIGH noise, or LOW mutual information, we need MANY terms in the series

$$\varepsilon = \sum_{l=0}^{\infty} s^l f(\mathbf{x}^{(l)})$$



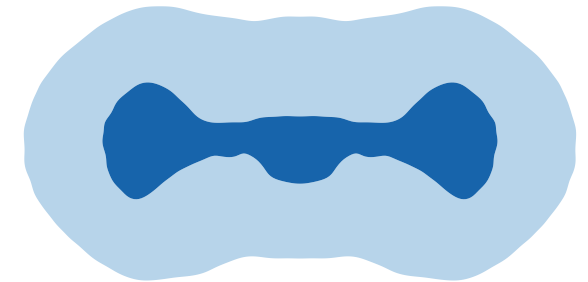
# Information in Echo Noise Channel

- Because they are the same in distribution, differential entropy,  $H$ , matches:

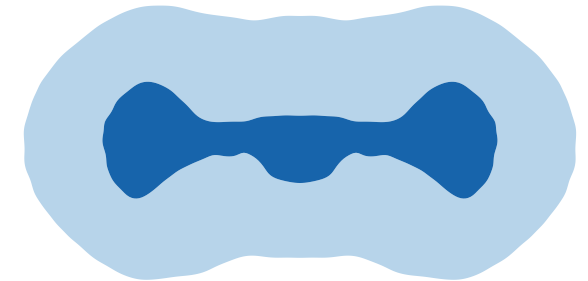
$$H(Z) = H(\mathcal{E})$$

- We don't know  $H()$ ... but we don't care, we are trying to get the **mutual information**,

$$I(\mathbf{Z};\mathbf{X}) = H(\mathbf{Z}) - \underline{H(\mathbf{Z}|\mathbf{X})}$$



$q(\varepsilon)$

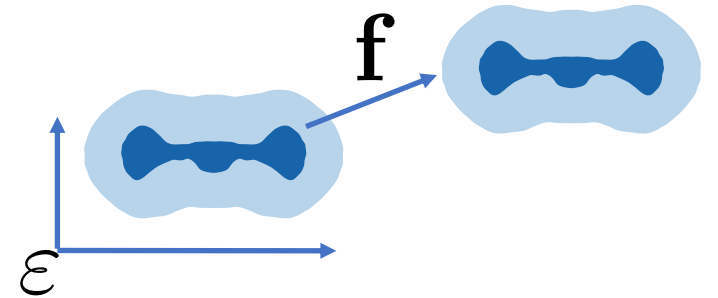


$q(\mathbf{z})$

# Sample-dependent noise scaling

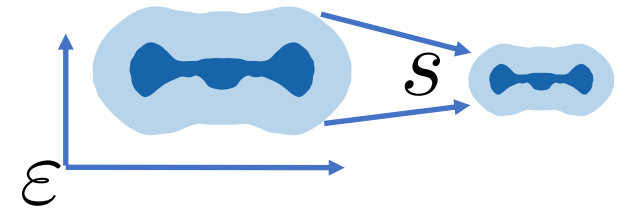
$$\mathbf{z} = f(\mathbf{x}) + S(\mathbf{x})\varepsilon$$

$$\begin{aligned} H(Z|X) &= H(f(X) + S(X)\mathcal{E} | X) \\ &= \mathbb{E}_{\mathbf{x}} H(f(\mathbf{x}) + S(\mathbf{x})\mathcal{E} | X = \mathbf{x}) \\ &= \mathbb{E}_{\mathbf{x}} H(S(\mathbf{x})\mathcal{E} | X = \mathbf{x}) \\ &= H(\mathcal{E}) + \mathbb{E} \log |\det S(\mathbf{x})| \end{aligned}$$



$$H(\mathcal{E}) = H(\mathcal{E} + \mathbf{f})$$

**Translation invariance**



$$H(s\mathcal{E}) = H(\mathcal{E}) + \log s$$

**Scale property**

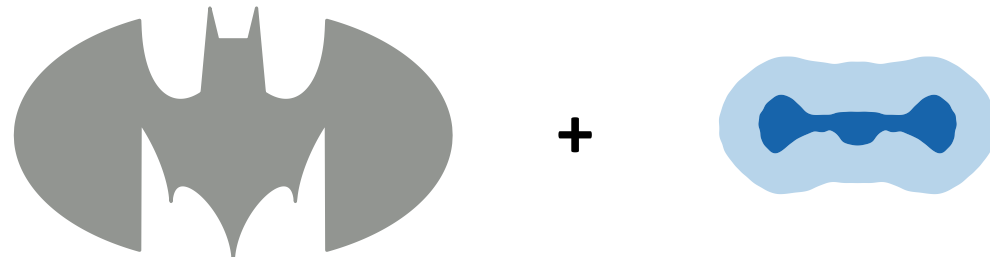
# Mutual information for the echo noise channel

$$H(Z|X) = H(\mathcal{E}) + \mathbb{E} \log |\det S(\mathbf{x})| \quad \text{Sample-dependent noise scaling}$$

$$H(Z) = H(\mathcal{E}) \quad \text{Self-similar noise (Echo)}$$

$$I(Z; X) = H(Z) - H(Z|X) \quad \text{Mutual information decomposed}$$

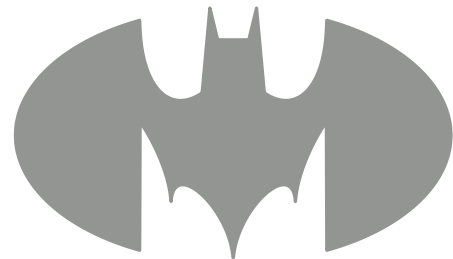
$$I(Z; X) = -\mathbb{E} \log |\det S(\mathbf{x})|$$



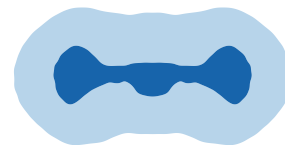
# Mutual information for the echo noise channel

- Works for any input (sampling noise requires samples of input)
- Set  $S(x) = s$  to get a simple, exact MI =  $-\log s$
- But  $S(x)$  is controllable (e.g. specify with a neural net) – a powerful way to get more flexible noise models

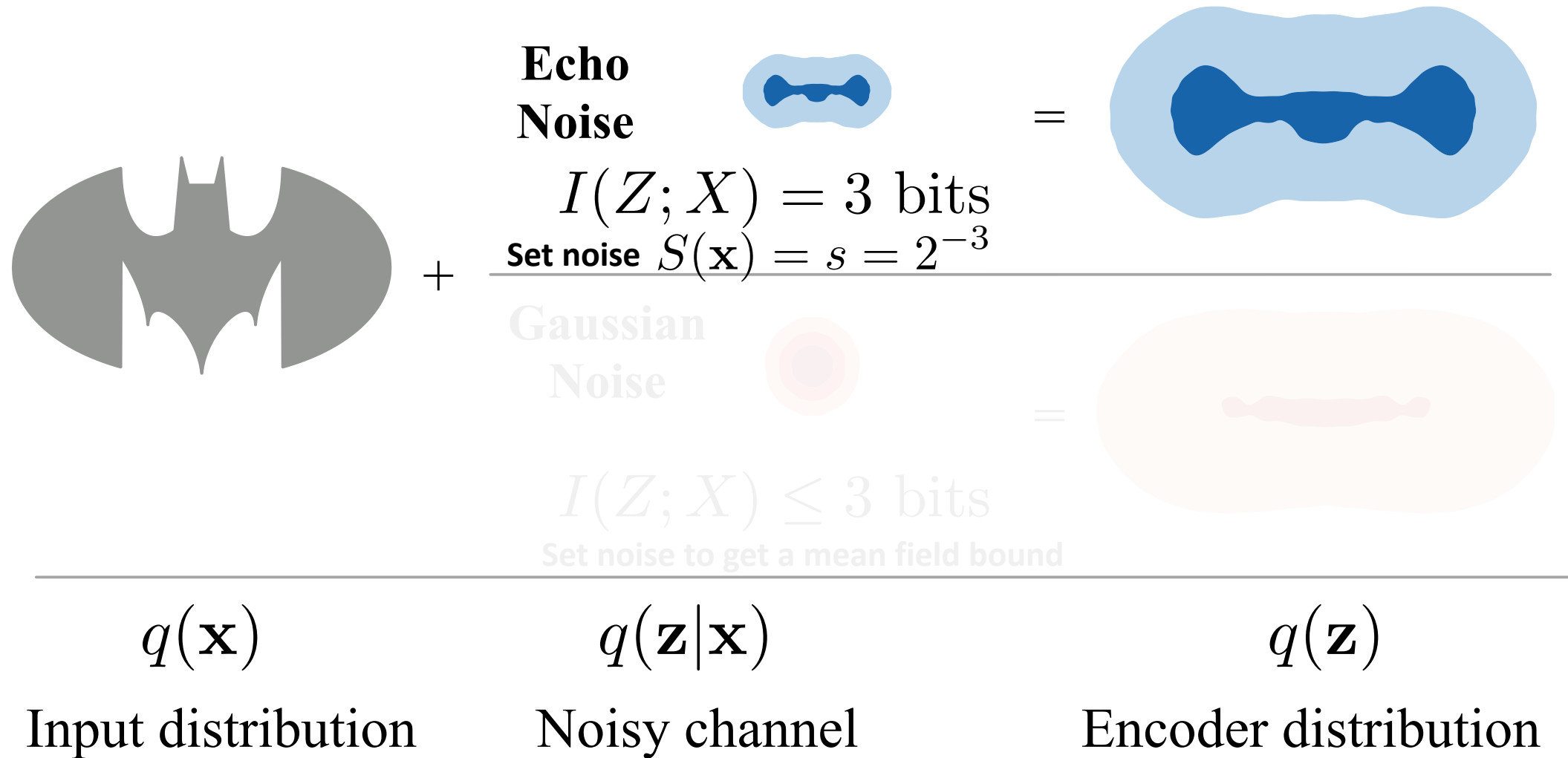
$$I(Z; X) = -\mathbb{E} \log |\det S(\mathbf{x})|$$



+



# Echo versus loose bounds



Application:

Variational Auto-Encoder (VAE) is  
an Information Bottleneck

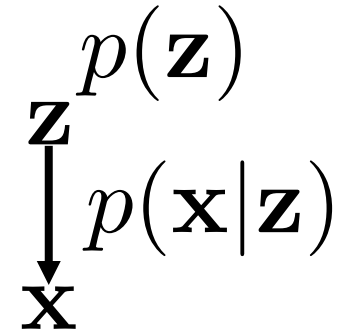
See nice discussion with connections to statistical physics: [arXiv:1803.08823](https://arxiv.org/abs/1803.08823)

# VAE perspective

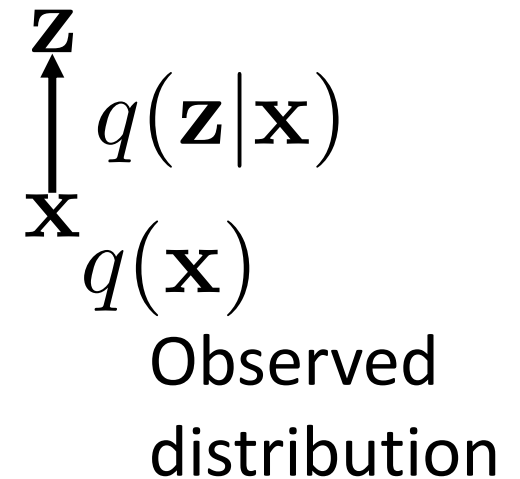
- Encoder and decoder parametrized by neural nets
- What's a good objective? Make the data likely under the generative model
- Problematic:

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

Decoder  
(generative  
model)



Encoder  
(variational  
approximation  
to  $p(\mathbf{z}|\mathbf{x})$ )



# Likelihood of data under generative model

$$\begin{aligned}\mathbb{E}_q \log p_\theta(\mathbf{x}) &\geq \mathbb{E}_q \log p_\theta(\mathbf{x}) - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_q \log p_\theta(\mathbf{x}|\mathbf{z}) - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\end{aligned}$$

Distortion / reconstruction loss ????

We already saw this term. It's the variational, mean-field bound for the information in the noisy channel,  $q$ .

$$I_q(Z; X) \leq D_{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$$

The best choice we can make is  $p(\mathbf{z}) = q(\mathbf{z})$ , leading to the tightest bound on likelihood



# Echo optimization problem

$$\mathbb{E}_{q(\mathbf{x})} \log p_{\theta}(\mathbf{x}) \geq \mathbb{E}_{q_{\phi}} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - I_{q_{\phi}}(X; Z)$$

$$\max_{\theta, \phi} \mathbb{E}_{q_{\phi}} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - \beta I_{q_{\phi}}(X; Z)$$

Distortion /  
reconstruction loss                      Compression

- Generalized info bottleneck, add a tunable trade-off (aka “beta-VAE”)
- Echo version:  $z = f(x) + \text{echo noise}$ . Then:

$$I(Z; X) = -\mathbb{E} \log |\det S(\mathbf{x})|$$

- Reconstruction term as usual for VAE

# Results

# Negative Log-Likelihood (lower is better)

Binary MNIST

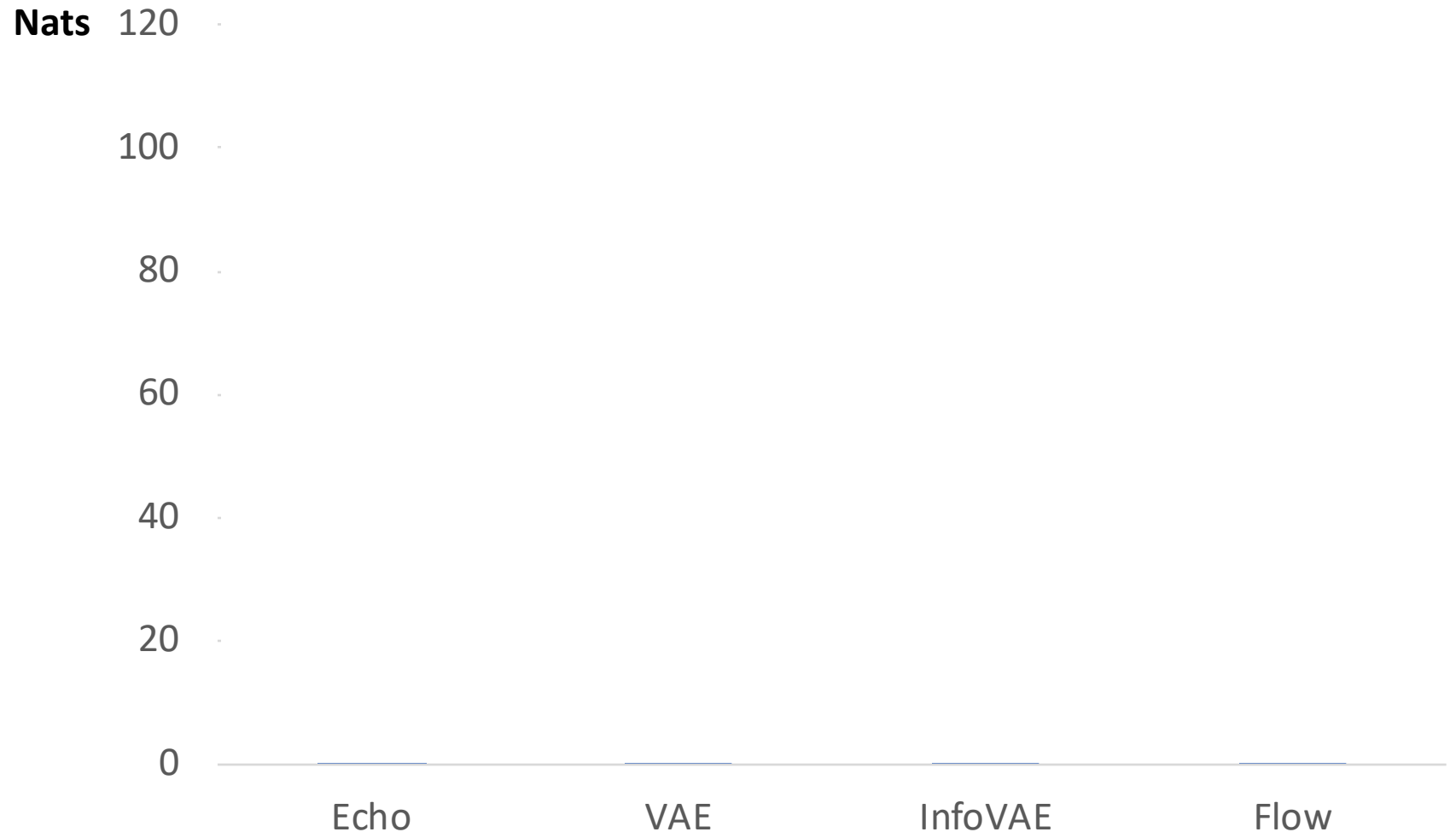


For reference:

Pure static = 543 nats/image

Constant = 0 nats/image

Optimal would be true  
entropy of data (unknown)

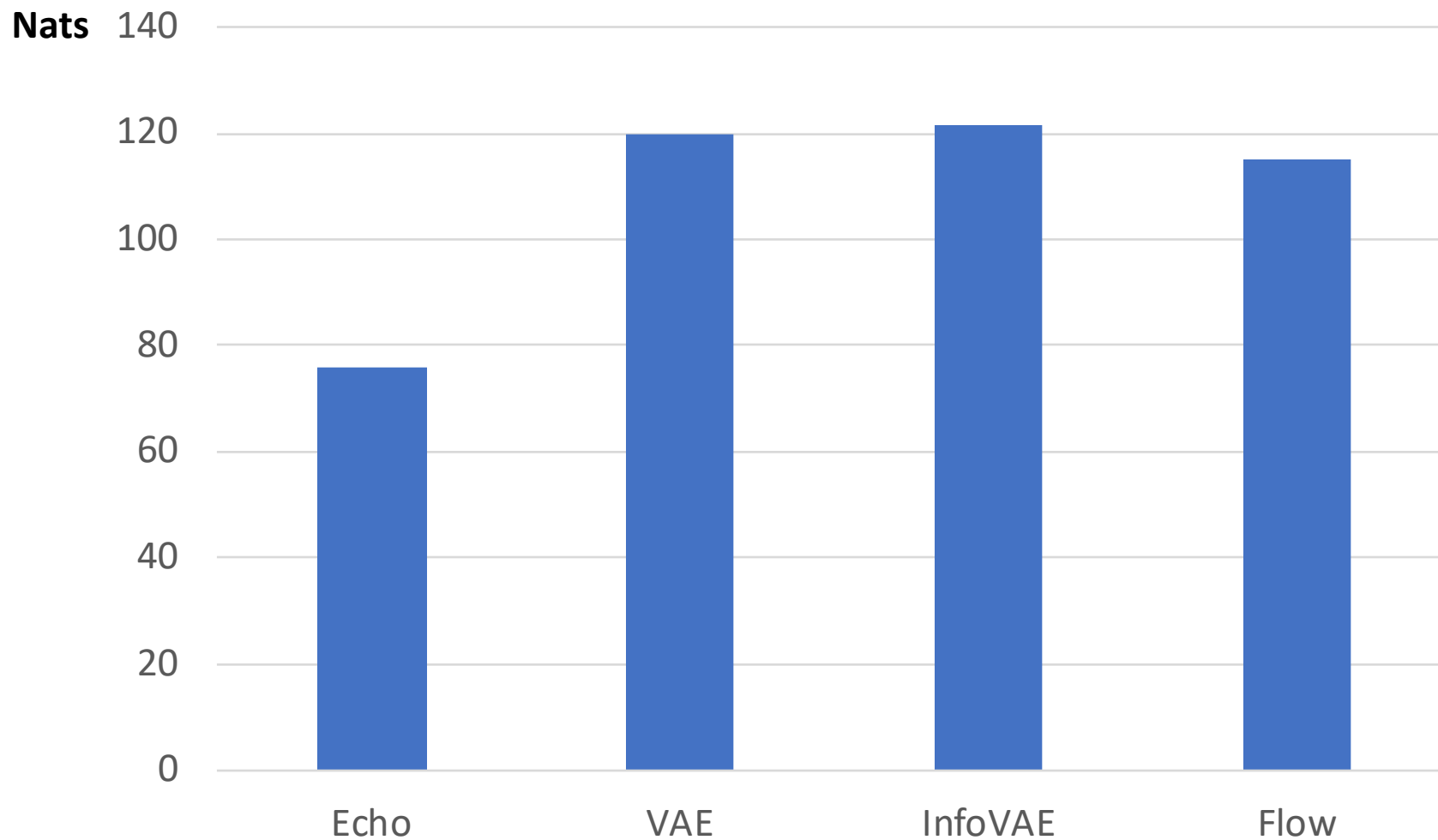




# Negative Log-Likelihood (lower is better)

Omniglot

5 7 7 7 7 7



# FashionMNIST



# Negative Log-Likelihood (lower is better)

Fashion MNIST



Nats

300

250

200

150

100

50

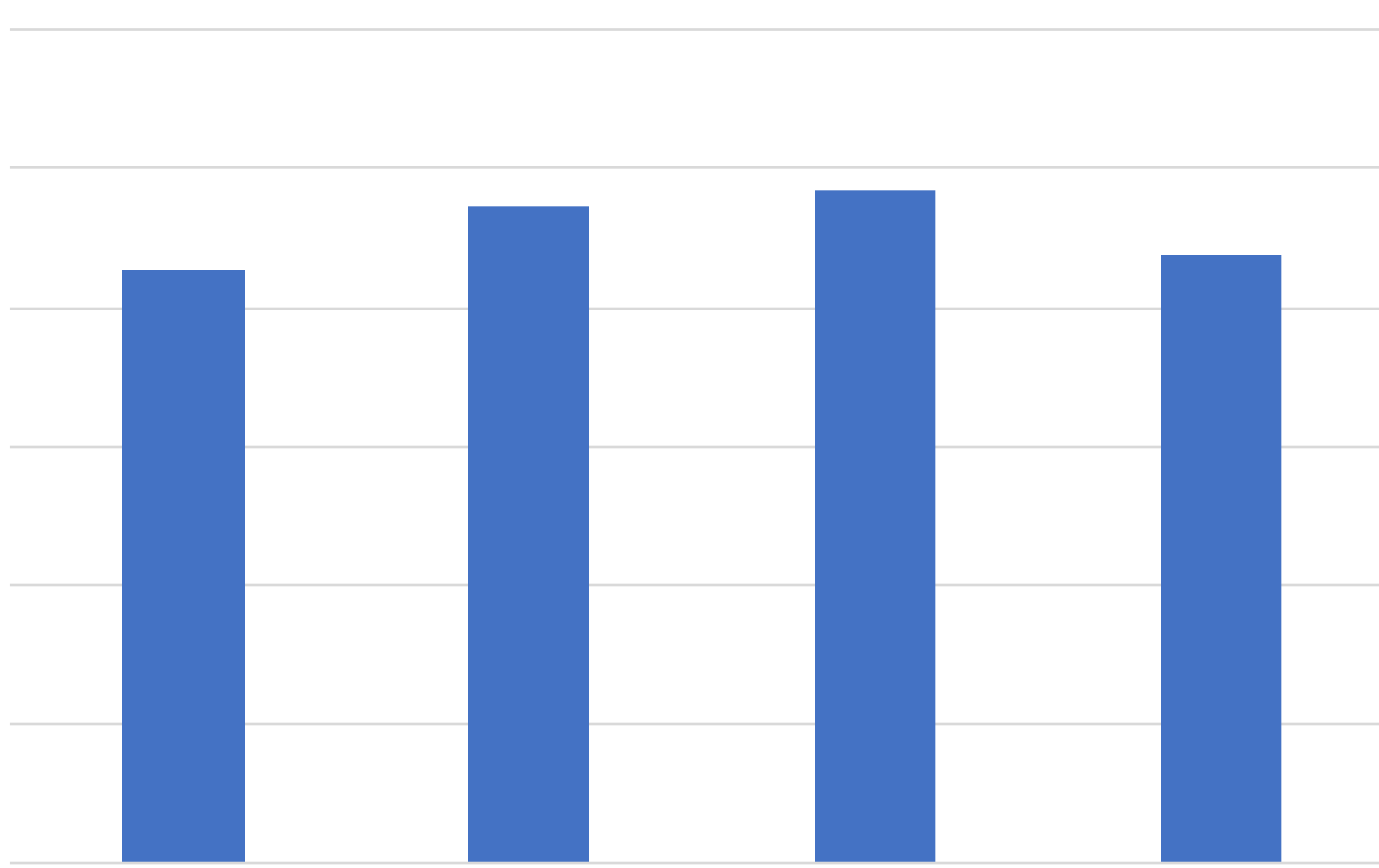
0

Echo

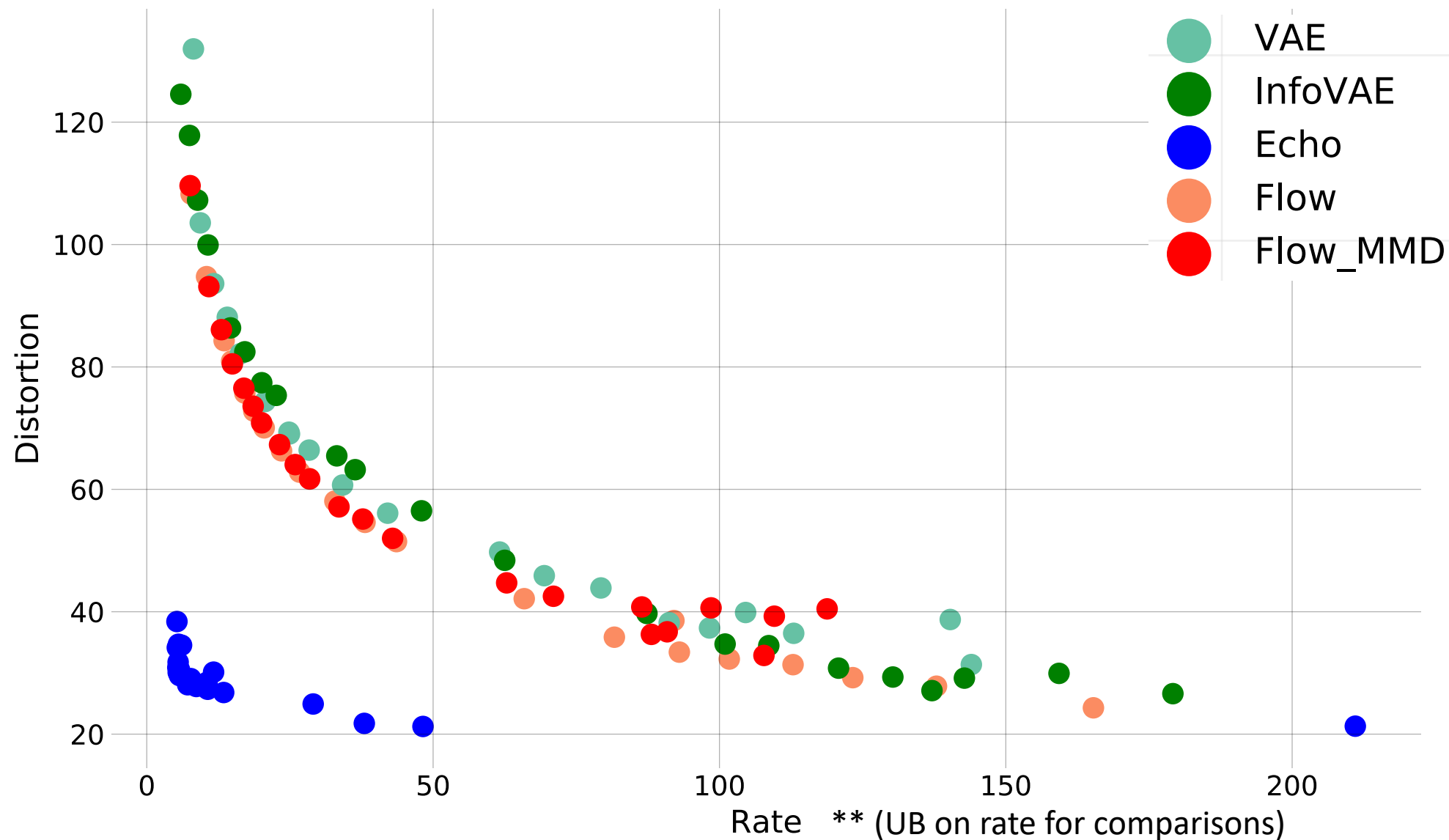
VAE

InfoVAE

Flow

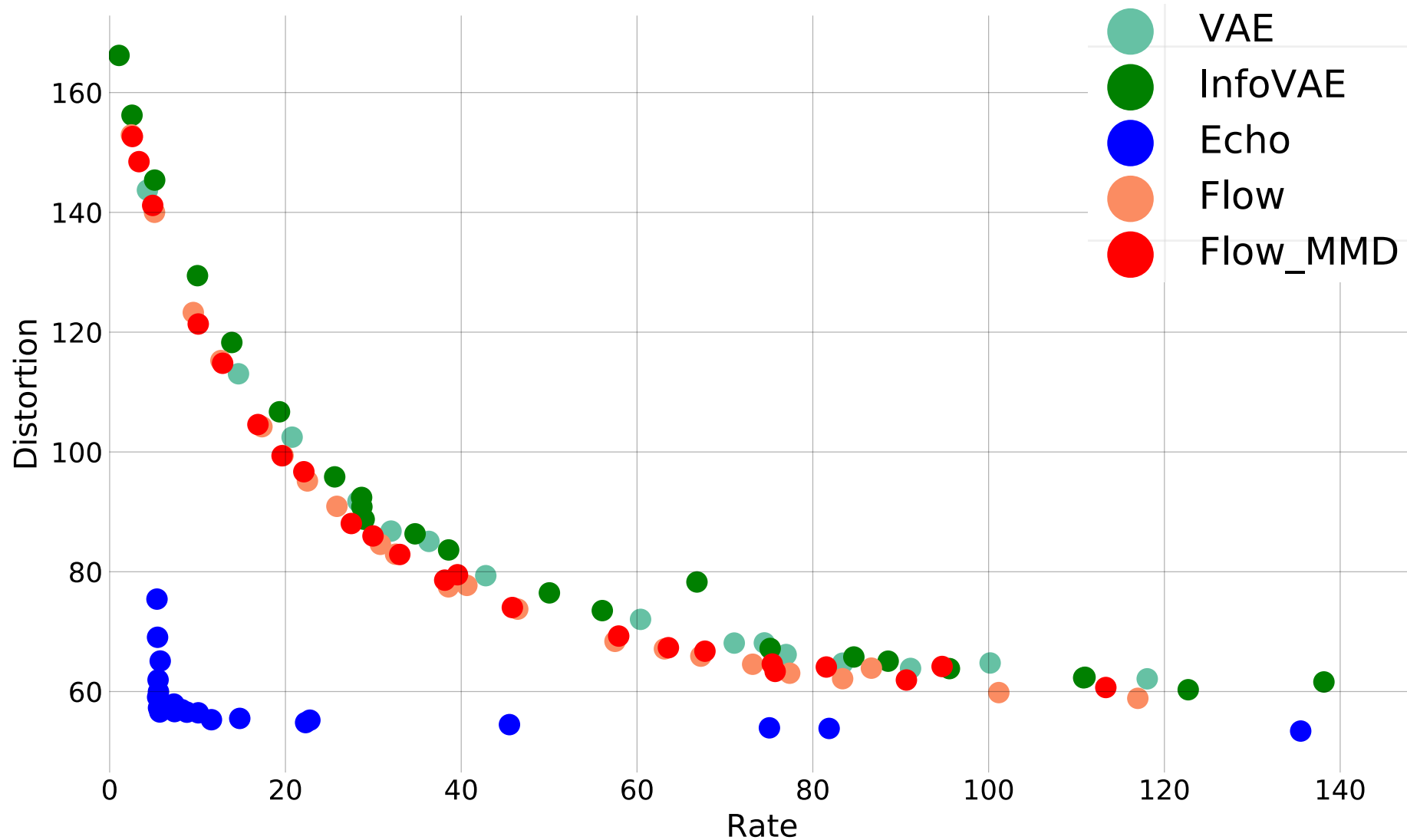


# Rate-Distortion – binary MNIST

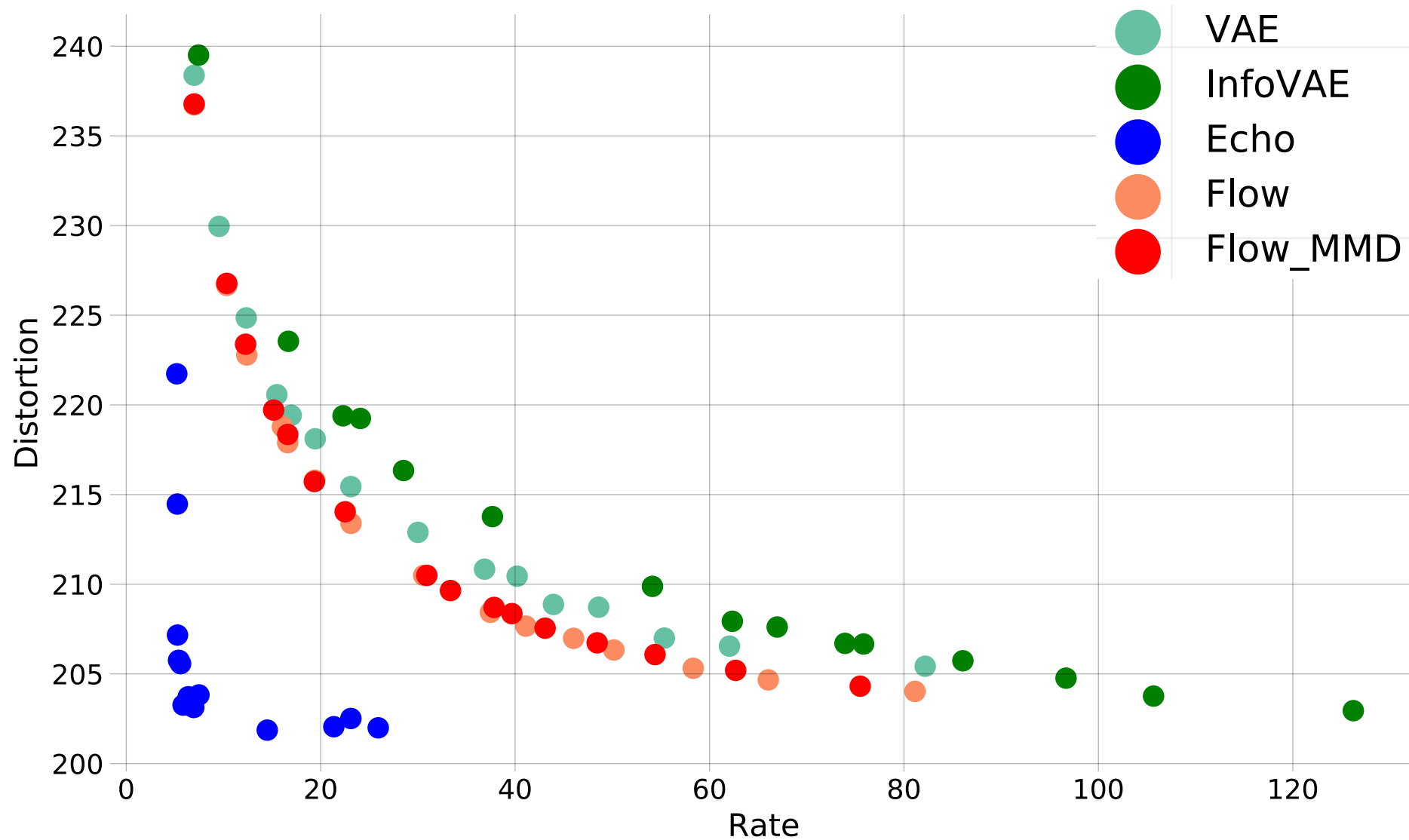
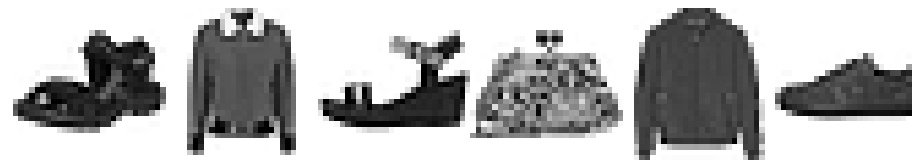




# Rate-Distortion - OmniGlott

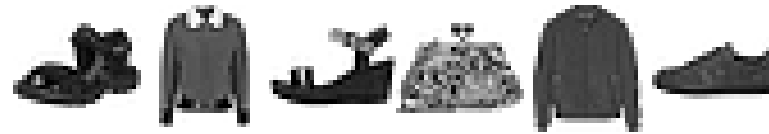


# Rate-Distortion – FMNIST



# Reconstruction example – Fashion MNIST

Data Sample



Low rate

High rate

$$\beta = 1.0$$

$$R=5.4 \quad D=205.8$$

$$\beta = 0.05$$

$$R=23.1 \quad D=202.5$$

Echo



$$\beta = 2.0$$

$$R=9.5 \quad D=230.0$$

$$\beta = 0.3$$

$$R=30.0 \quad D=212.9$$

VAE



# Compression / Generalization Controversy

# Opening the black box of Deep Neural Networks via Information

**Ravid Schwartz-Ziv**

*Edmond and Lilly Safra Center for Brain Sciences  
The Hebrew University of Jerusalem  
Jerusalem, 91904, Israel*

RAVID.ZIV@MAIL.HUJI.AC.IL

**Naftali Tishby\***

*School of Engineering and Computer Science  
and Edmond and Lilly Safra Center for Brain Sciences  
The Hebrew University of Jerusalem  
Jerusalem, 91904, Israel*

TISHBY@CS.HUJI.AC.IL

arXiv:1703.00810v3 [cs.LG] 29 Apr 2017

## Abstract

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work [Tishby and Zaslavsky (2015)] proposed to analyze DNNs in the *Information Plane*; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. They suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer.

In this work we follow up on this idea and demonstrate the effectiveness of the Information-Plane visualization of DNNs. Our main results are: (i) most of the training epochs in standard DL are spent on compression of the input to efficient representation and not on fitting the training labels. (ii) The representation compression phase begins when the training errors becomes small and the Stochastic Gradient Decent (SGD) epochs change from a fast drift to smaller training error into a stochastic relaxation, or random diffusion, constrained by the training error value. (iii) The converged layers lie on or very close to the Information Bottleneck (IB) theoretical bound, and the maps from the input to any hidden layer and from this hidden layer to the output satisfy the IB self-consistent equations. This generalization through noise mechanism is unique to Deep Neural Networks and absent in one layer networks. (iv) The training time is dramatically reduced when adding more hidden layers. Thus the main advantage of the hidden layers is computational. This can be explained by the reduced relaxation time, as this it scales super-linearly (exponentially for simple diffusion) with the information compression from the previous layer. (v) As we expect critical slowing down of the stochastic relaxation near phase transitions on the IB curve, we expect the hidden layers to converge to such critical points<sup>†</sup>

- Learning exhibits a “fitting phase” followed by a “compression phase”
- Compression leads to good generalization for deep nets

# ON THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

**Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani**

Harvard University

{asaxe,madvani}@fas.harvard.edu, {ybansal,dapello}@g.harvard.edu

**Artemy Kolchinsky, Brendan D. Tracey**

Santa Fe Institute

{artemyk,tracey.brendan}@gmail.com

**David D. Cox**

Harvard University

MIT-IBM Watson AI Lab

davidcox@fas.harvard.edu

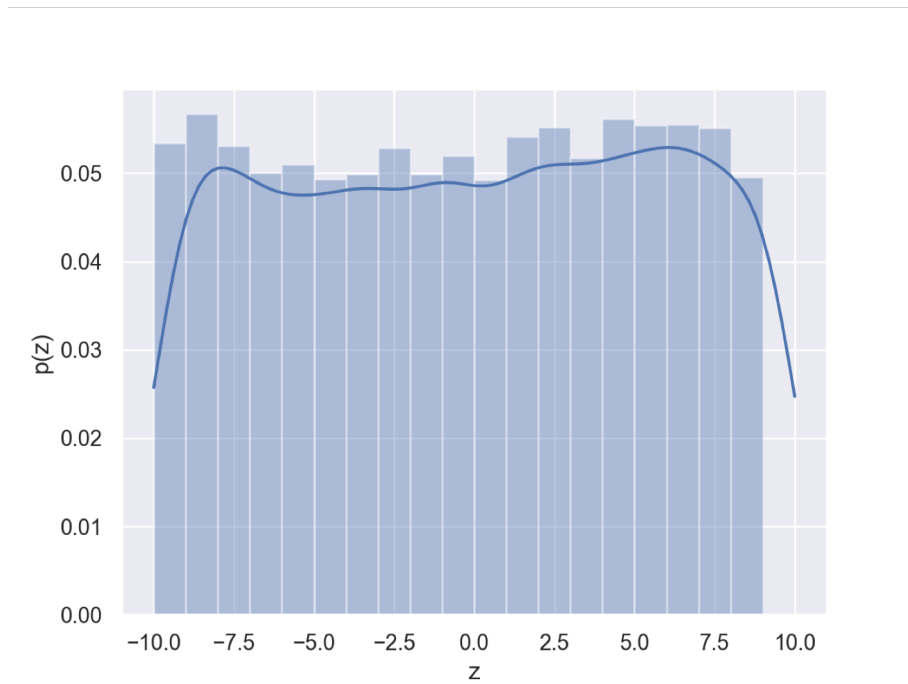
david.d.cox@ibm.com

## ABSTRACT

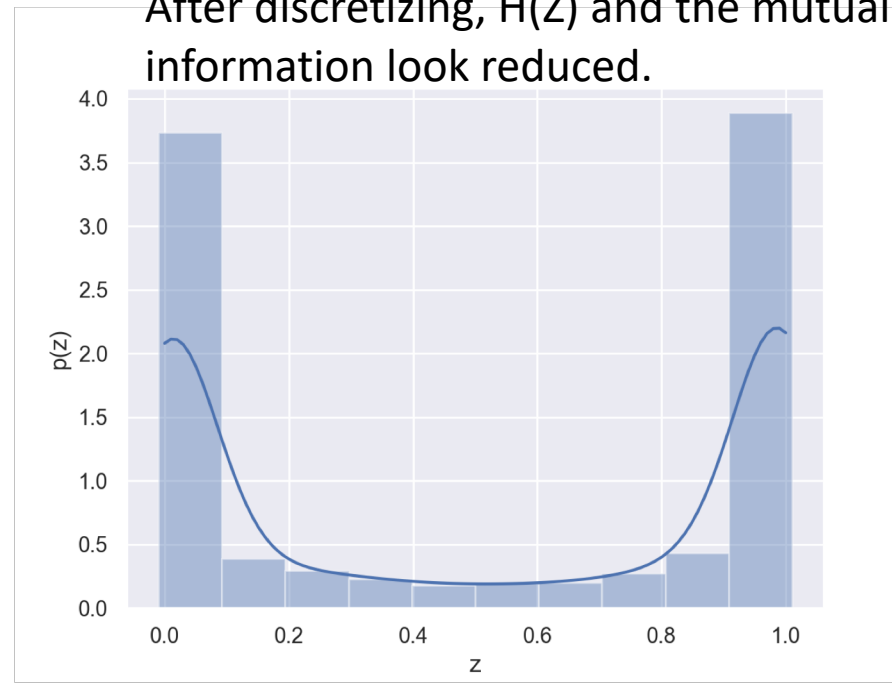
The practical successes of deep neural networks have not been matched by theoretical progress that satisfyingly explains their behavior. In this work, we study the information bottleneck (IB) theory of deep learning, which makes three specific claims: first, that deep networks undergo two distinct phases consisting of an initial fitting phase and a subsequent compression phase; second, that the compression phase is causally related to the excellent generalization performance of deep networks; and third, that the compression phase occurs due to the diffusion-like behavior of stochastic gradient descent. Here we show that none of these claims hold true in the general case. Through a combination of analytical results and simulation, we demonstrate that the information plane trajectory is predominantly a function of the neural nonlinearity employed: double-sided saturating nonlinearities like tanh yield a compression phase as neural activations enter the saturation regime, but linear activation functions and single-sided saturating nonlinearities like the widely used ReLU in fact do not. Moreover, we find that there is no evident causal connection between compression and generalization: networks that do not compress are still capable of generalization, and vice versa. Next, we show that the compression phase, when it exists, does not arise from stochasticity in training by demonstrating that we can replicate the IB findings using full batch gradient descent rather than stochastic gradient descent. Finally, we show that when an input domain consists of a subset of task-relevant and task-irrelevant information, hidden representations do compress the task-irrelevant information, although the overall information about the input may monotonically increase with training time.

Counter-point: observed compression is an artifact of the (discretized) information estimators. With different nonlinearities, different behavior emerges.

- Mutual information is invariant under invertible transformations.
- BUT, if you discretize a continuous signal, you might get different answers.



Invertible nonlinearity (sigmoid) applied.  
After discretizing,  $H(Z)$  and the mutual information look reduced.



# Entropy and mutual information in models of deep neural networks

Marylou Gabrié<sup>\*1</sup>, Andre Manoel<sup>2,3</sup>, Clément Luneau<sup>4</sup>, Jean Barbier<sup>4</sup>, Nicolas Macris<sup>4</sup>,  
Florent Krzakala<sup>1,5,6,7</sup>, Lenka Zdeborová<sup>3,5</sup>

<sup>1</sup>Laboratoire de Physique Statistique, École Normale Supérieure, PSL University

<sup>2</sup>Parietal Team, INRIA, CEA, Université Paris-Saclay

<sup>3</sup>Institut de Physique Théorique, CEA, CNRS, Université Paris-Saclay

<sup>4</sup>Laboratoire de Théorie des Communications, École Polytechnique Fédérale de Lausanne

<sup>5</sup>Department of Mathematics, Duke University, Durham NC

<sup>6</sup>Sorbonne Universités

<sup>7</sup>LightOn, Paris

arXiv:1805.09785v2 [cs.LG] 29 Oct 2018

## Abstract

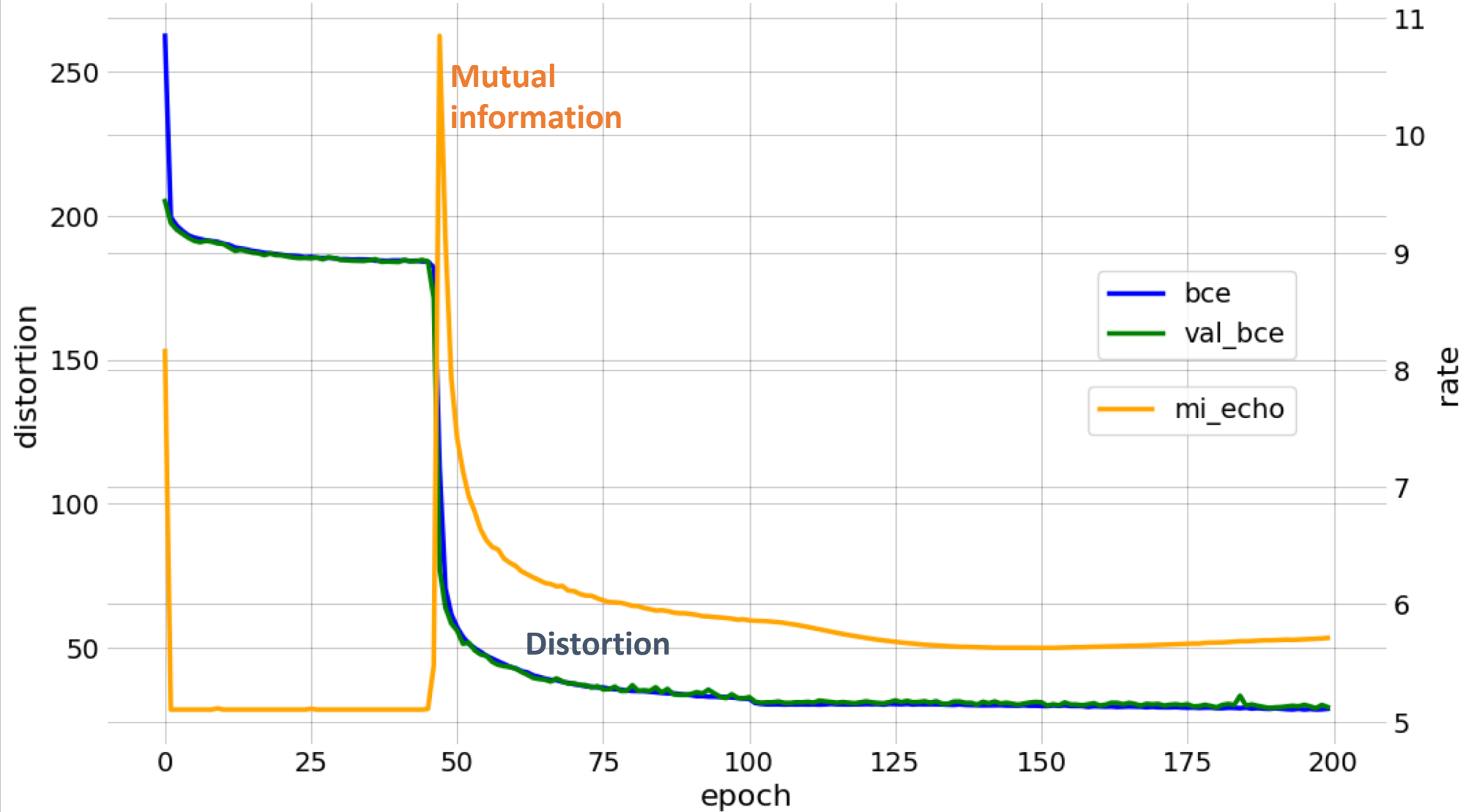
We examine a class of stochastic deep learning models with a tractable method to compute information-theoretic quantities. Our contributions are three-fold: (i) We show how entropies and mutual informations can be derived from heuristic statistical physics methods, under the assumption that weight matrices are independent and orthogonally-invariant. (ii) We extend particular cases in which this result is known to be rigorously exact by providing a proof for two-layers networks with Gaussian random weights, using the recently introduced adaptive interpolation method. (iii) We propose an experiment framework with generative models of synthetic datasets, on which we train deep neural networks with a weight constraint designed so that the assumption in (i) is verified during learning. We study the behavior of entropies and mutual informations throughout learning and conclude that, in the proposed setting, the relationship between compression and generalization remains elusive.

For certain random networks,  
mutual information can be  
analytically calculated.

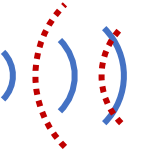
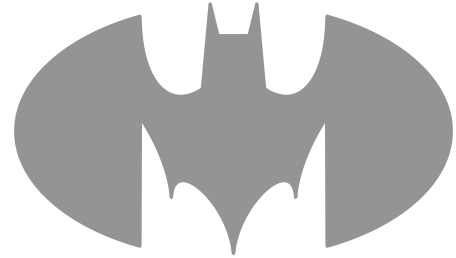
- Compression can occur even  
without nonlinearities
- No implications for  
generalization are observed.



Loss over Training Epochs



# Conclusion



- Echo is a more powerful noise model, with exact rate (instead of loose bounds)
- Useful wherever optimizing lossy compression is useful: (supervised) information bottleneck and VAE...
- Compression/generalization? At least we can nail down compression

arXiv preprint in next day or two, poster

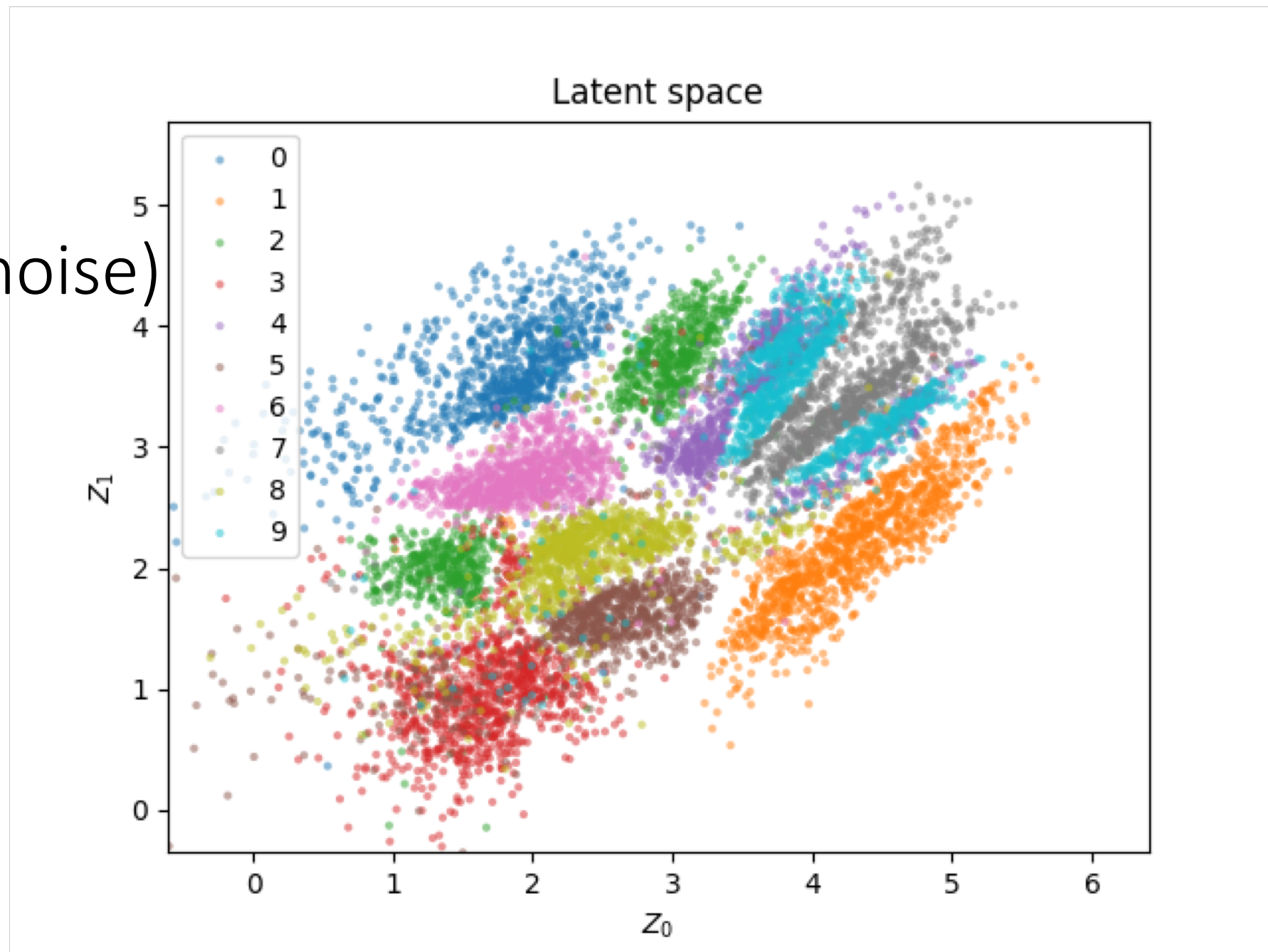
Me: [gregv@isi.edu](mailto:gregv@isi.edu) , Rob: [brekelma@usc.edu](mailto:brekelma@usc.edu)

**Thanks!**

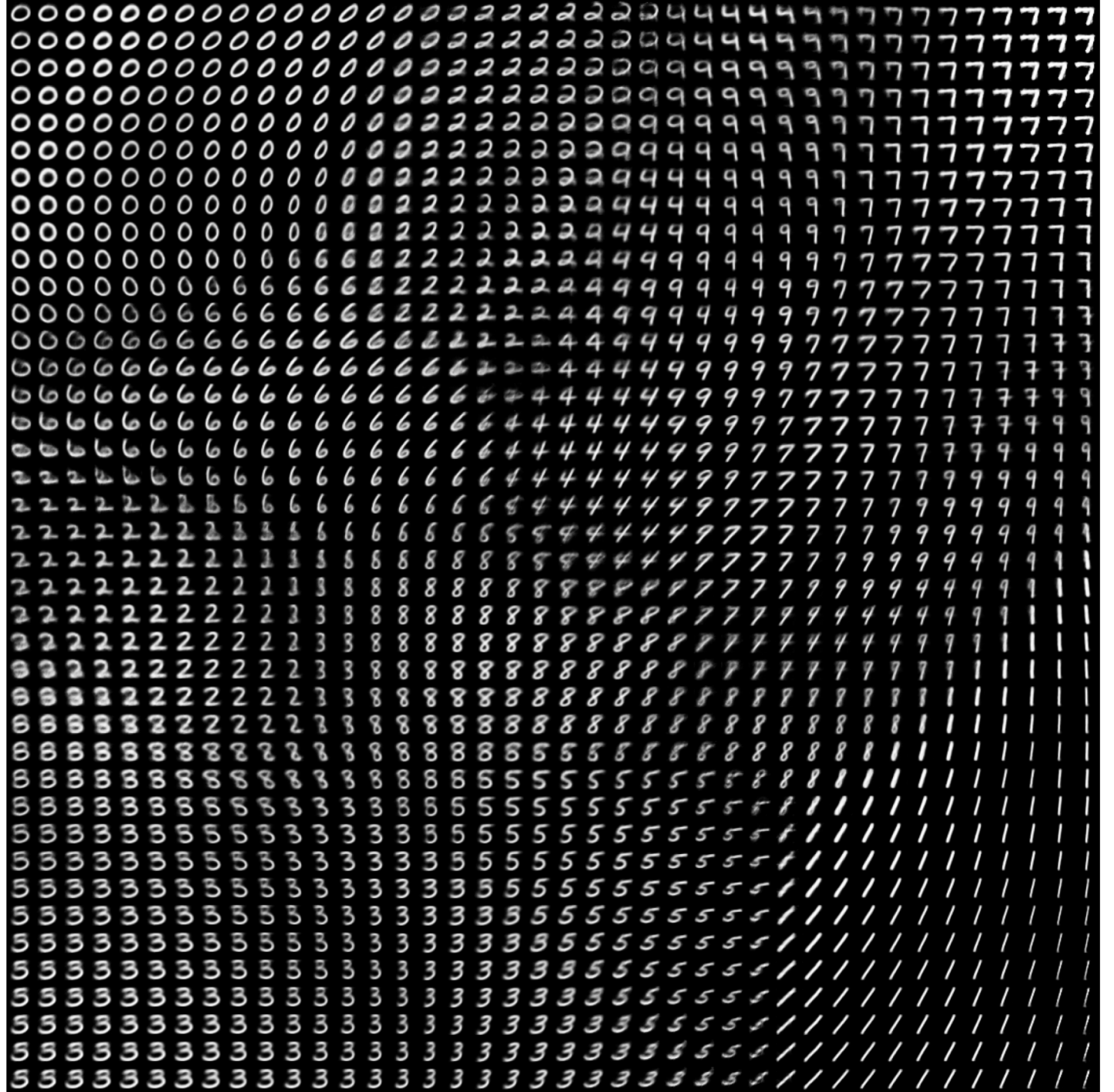
# Why?

- Strongly outperforms standard VAE for rate-distortion for three possible reasons:
  - Bound on likelihood is tighter, due to exact MI characterization
  - Noise model is stronger, more general
  - No assumptions of independence/Gaussianity required in noise or  $p(z)$

Echo MNIST  
(no sample  
dependent noise)



The manifold  
(no sample  
dependent noise)



# Sample dependent noise

