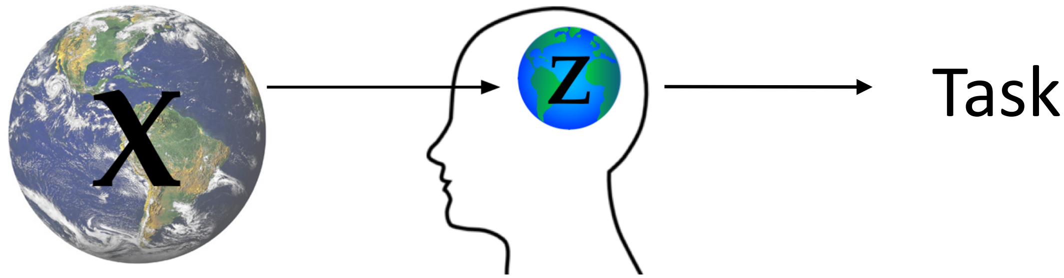
Beyond Shannon Applications in Machine Learning

Greg Ver Steeg

USC Universit
Southern Cali
Information Sciences I

Beyond Shannon Workshop, 2019



Z is a learned *representation** of X

$z = f_\theta($

E.g., parametrized by a
neural network

Outline:

- “Disentangling”
- “Fair” representation learning
- Lossy Compression in representation learning
  - Gaussian vs Echo compression
  - Higher order interactions

Slides with all paper links: http://bit.ly/ST

* Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspecti
transactions on pattern analysis and machine intelligence, 35(8): 1798–1828, 2013.

KL is King

In machine learning, Kullback-Leibler divergence and related quantities are dominant: *Maximum likelihood, cross entropy loss, variational inference, Mutual information*

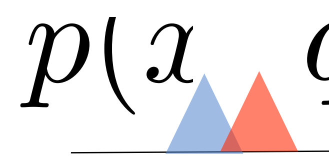
But there are practical issues for machine learning:

- Estimation is hard
- Often rely on loose bounds
- Usage is incorrect / inappropriate / poorly motivated
- Highly non-convex optimization
- Underflows lead to numerical errors with rare events, log 0 = - ∞

Other information measures in ML?

Smoother optimization (than KL divergence) with Wasserstein distance (earth mover distance for distributions)

- Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein GAN, https://arxiv.org/abs/1701.07875



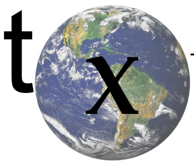
Other (in)dependence measures: Hilbert-Schmidt Independence Criterion (HSIC) / Maximum-Mean Discrepancy (MMD) / kernel trick

- Introduced: Gretton et al 2007
- Example usage: InfoVAE, arXiv:1706.02262

Almost never in ML: Renyi, info decomposition, synergy, intersection, common info

- Wyner common information: arXiv:1606.02307
- Synergy: arXiv:1710.03839
- Hierarchical decomposition of total correlation: arXiv:1410.7404

‘Disentangling” – encourage Z_j to mat
‘intuitive” factors of variation (without
abels)



, Tian Qi, et al.
ating sources of
tanglement in
tional
encoders." *NIPS* 2018

(a) Baldness (-6, 6) (b) Face width (0, 6) (c) Gender (-6, 6) (d) Mustache (-6, 0)

? Good old statistical
pendence

$$(Z) \equiv \mathbb{E}_p \left[\log \frac{p(z_1,}{\prod_j} \right.$$

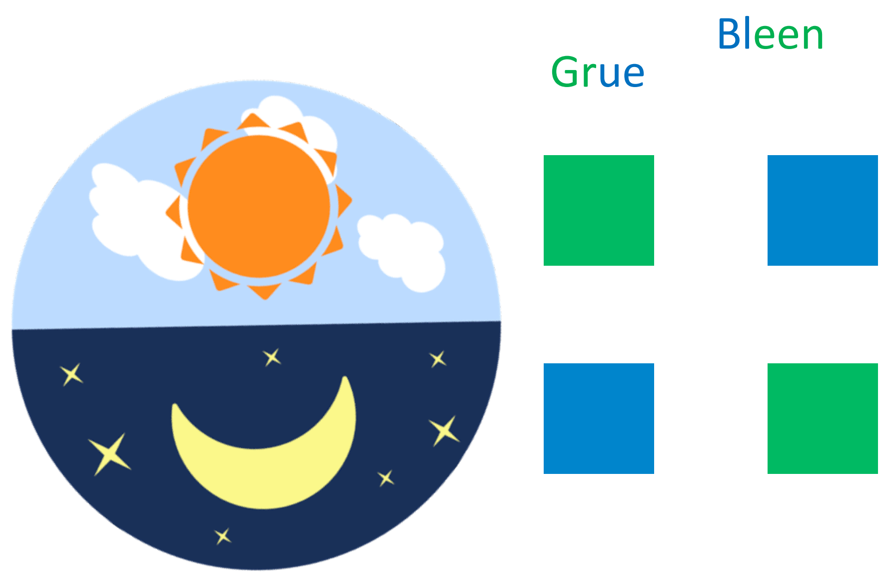
papers working on this angle… with dubious results, see review:

co Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Scholkopf, and OlivierBachem. Challenging com
ions in the unsupervised learning of disentangled representations. arXiv preprint arXiv:1811.12359, 2018.

The Grue Language doesn’t have words for “blue” or “green”

Grue = Green during the **day**, blue at *night*

Bleen = Blue during the **day**, green at *night*



Needlessly complicated?

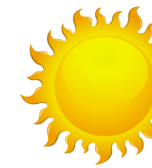
glish is needlessly complicated because:

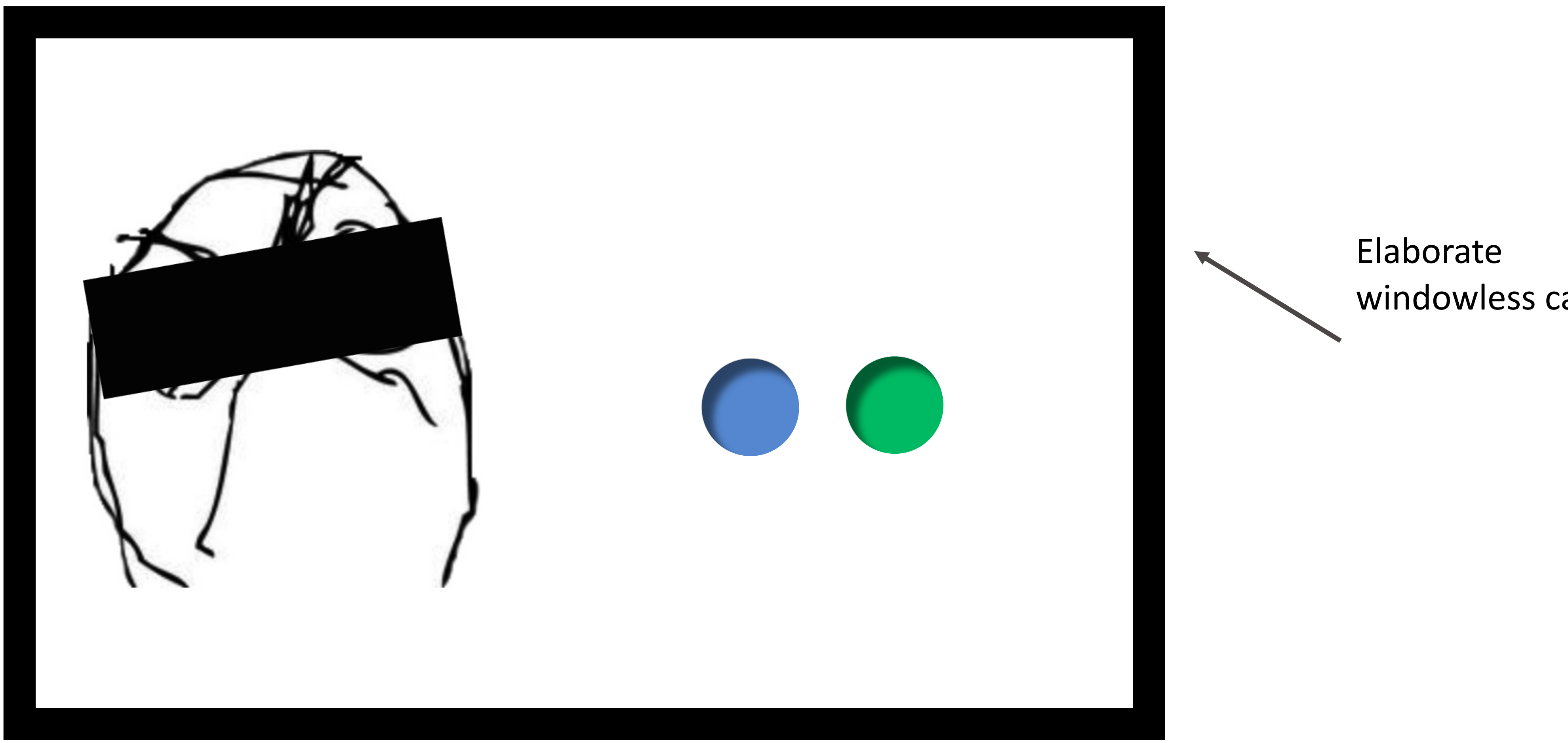
een = Grue during day and bleen at night
e = Bleen during day and grue at night

Grue langu
speake

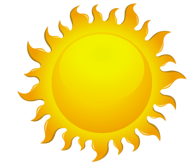
(Also, what is deal with "the" and "a"? Why you have these useless words?)

Push the bleen button and we let you go…



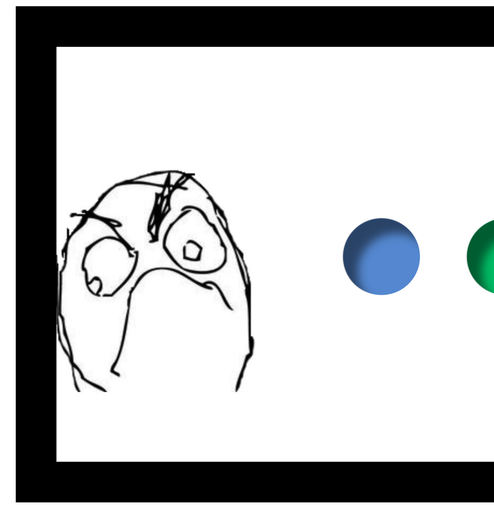


Grue language is synergistic



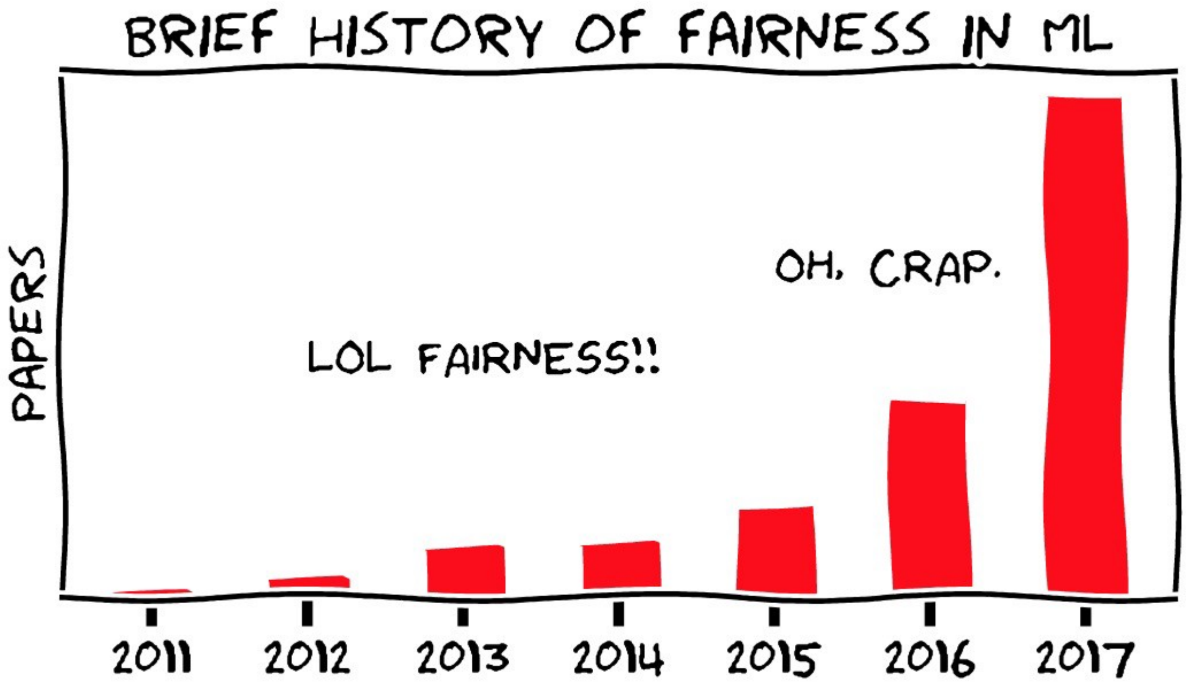
me and grue > time only *or* grue only
nplete info)   (no info)

predicting what you will see



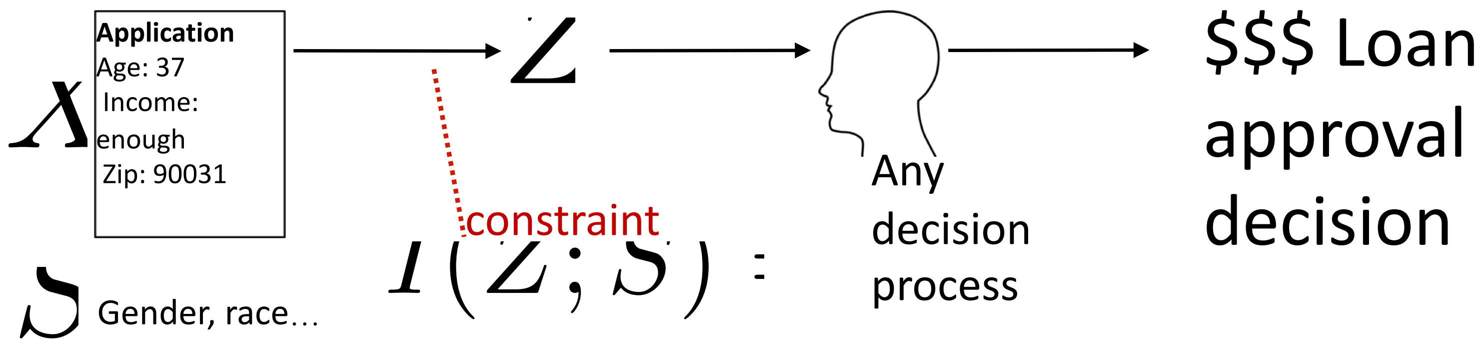
Is there hope for a plausible information-theoretic principle of "disentangling"? And one that can be optimized for learning?

Fair and Invariant Representations



https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

An information-theoretic notion of fairness



Example: your task is to approve loans

To be fair, you don’t want to discriminate based on protected variable

For historical reasons, S may be correlated with X

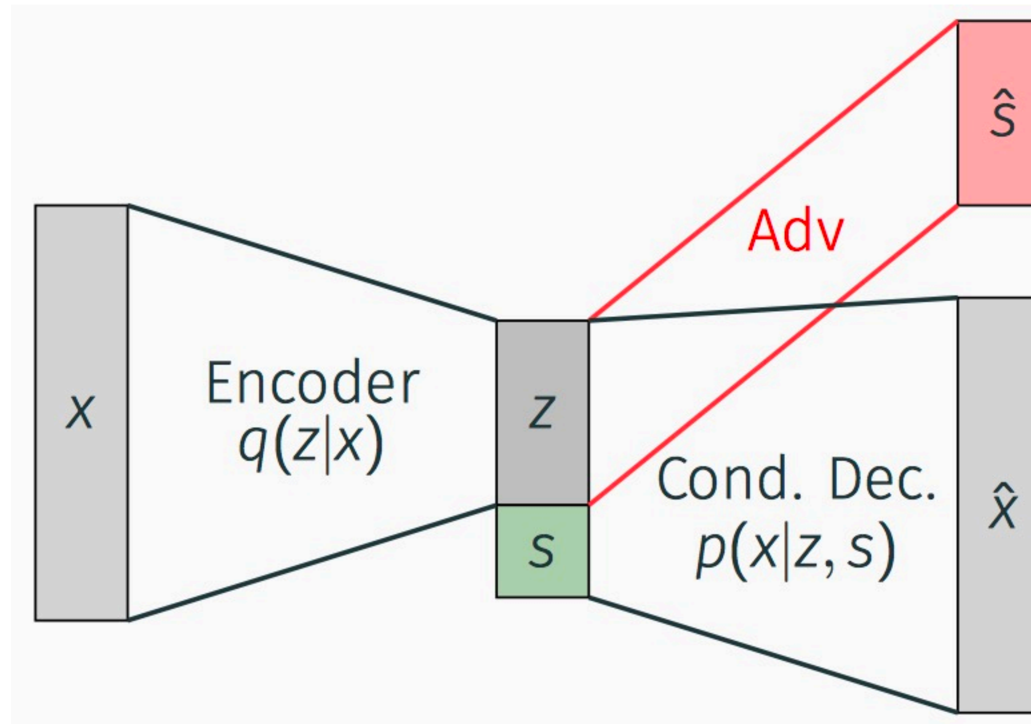
By removing information about S from the representation Z that we u
for decision making, we make it impossible to be biased against S

Adversarial approach

Make sure an adversary cannot reconstruct S from Z

This doesn't guarantee that I(Z;S) is small! Adversaries only provide *lower* bounds on mutual information.

(Nevertheless adversarial learning revolutionized some problems)

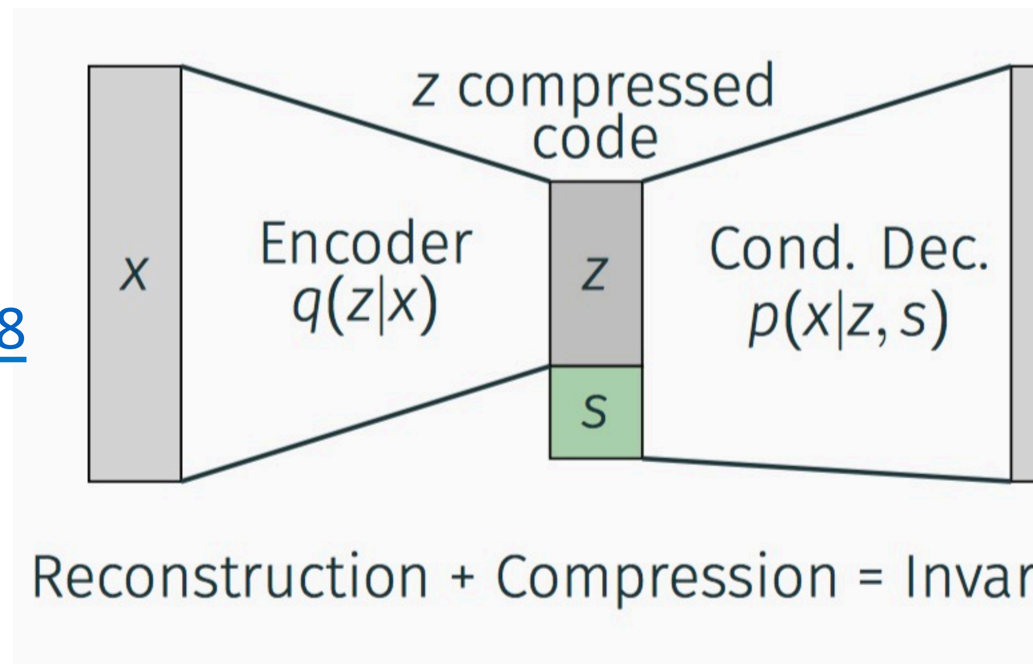


C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. variational fair autoencoder. *arXiv preprint arXiv:1511* 2015.
Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllal invariance through adversarial feature learning. Neur

A direct, information-theoretic approach

We derive an *upper* bound on I(Z;S) that can be tractably optimized

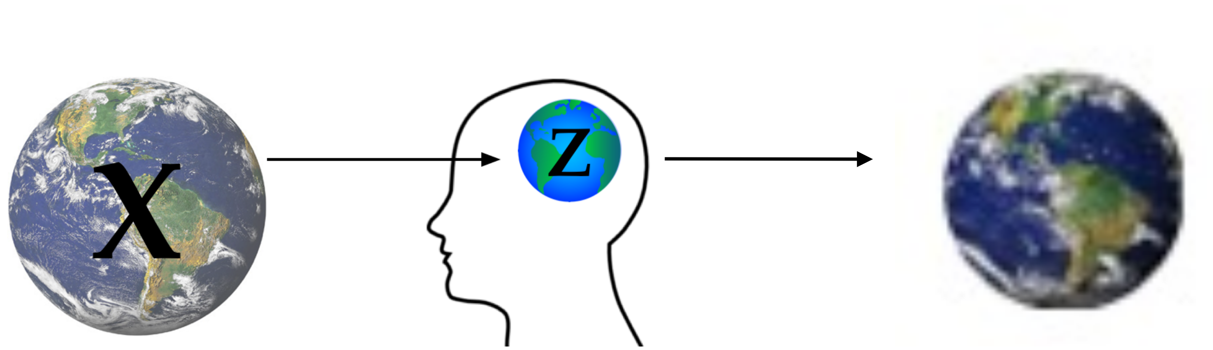
$$I(Z;S) \leq \underbrace{-\mathbb{E}_q \left[\log\right.}_{\text{Conditional reconstruction}} p(x\overbrace{|z,s)\left.\right]}^{\text{Compression}} + \overbrace{I_q(Z;}^{\text{Constant}}$$



Main idea in NIPS 2018, arXiv:1805.09458

Application to fMRI: arXiv:1904.05375

When and how to remove information?



Z is a compressed *representation* of X,
that should be useful for reconstruction

_ossy compression (and VAE's)

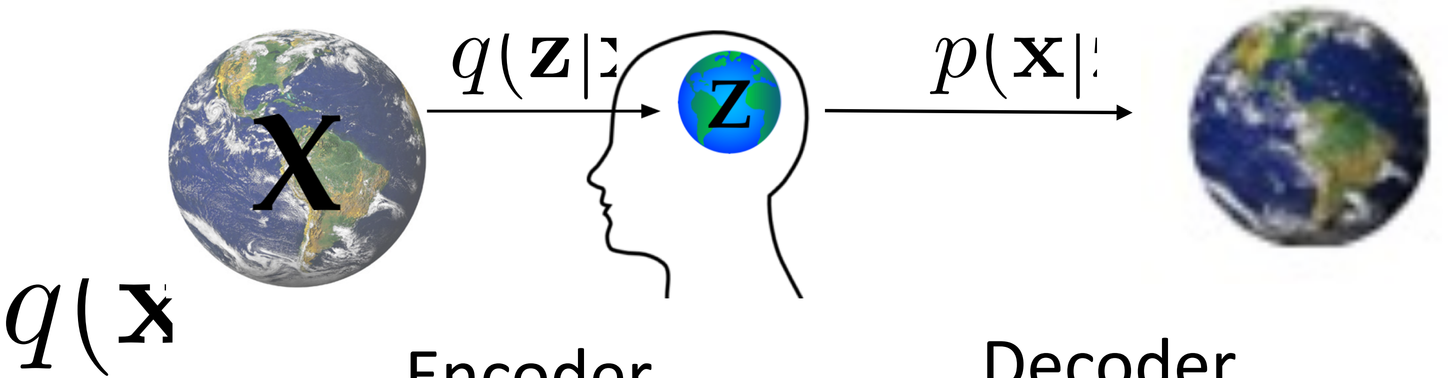
ariational auto-encoders (VAEs) are our motivating example. They have an interpretation in terms of enerative models as well as with rate-distortion, as discussed in several papers:

lemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In *nternational Conference on Machine Learning*, pp. 159–168, 2018.

ezende, D. J. and Viola, F. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

rekelmans, Moyer, Galstyan, and Ver Steeg. "Exact Rate-Distortion in Autoencoders via Echo oise." *arXiv preprint arXiv:1904.07199* (2019).

Variational Auto-Encoder (VAE)



Observed distribution

Encoder (variational approximation to p(z|x))

Decoder (generative model)

ncoder and decoder parametrized by neural nets

oal is to maximize likelihood of observations under generative model

Variational Evidence Lower BOund (ELBO)

$$\mathbb{E}_q \log p_\theta(\mathbf{x}) \geq \mathbb{E}_q \log p_\theta(\mathbf{x}) - KL[q_\phi$$

$$= \mathbb{E}_q \log p_\theta(\mathbf{x}|\mathbf{z}) - D_{KL}[$$

True for any p(z). The choice with the tightest bound is p(z)=q(z)

$$\mathbb{E}_{q(\mathbf{x})} \log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})$$

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

Distortion / reconstruction loss

Compression

Compression: bounds and a new exact approach with echo noise

nformation in a noisy channel

Noisy channel

$\mathbf{x} \longrightarrow \mathbf{z}$

Input distribution $q(\mathbf{x}$ $q(\mathbf{z}|\mathbf{x}$ $q(\mathbf{z}, \mathbf{x}) = q($

Mutual information $= D_{KL}[q(\mathbf{z}|$

Problem

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

High-d integral

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

Approximate as Gaussian for (MaxEnt) upper bound

$$z_i = x_i + \qquad \varepsilon_i \sim \mathcal{N}\,(($$

Approximate q(z) as Normal: $p(z_i) = \mathcal{N}\,(\mathbb{E}(x_i), \mathrm{Var}($

$$I(Z;X) \leq \sum_i \frac{1}{2} \log\left(1 + \right.$$

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

Echo noise: make the noise look like the signal

How do we make the noise look like the signal?

$$= f(\mathbf{x}) \,\text{-}$$

$$= f(\mathbf{x}^{(0)}) + s f(\mathbf{x}^{(1)}) + s^2 f(\mathbf{x}^{(2)}) \ldots$$

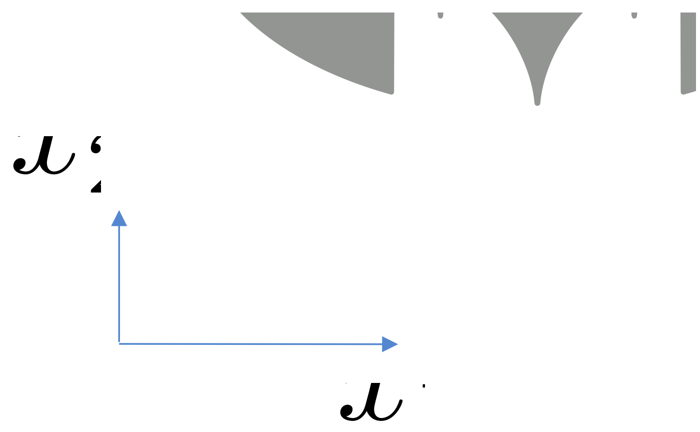
$$= f(\mathbf{x}) + s\Big(f(\mathbf{x}^{(0)}) + s f(\mathbf{x}^{(1)}) +$$

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

$$\mathcal{E} \stackrel{d}{=}$$

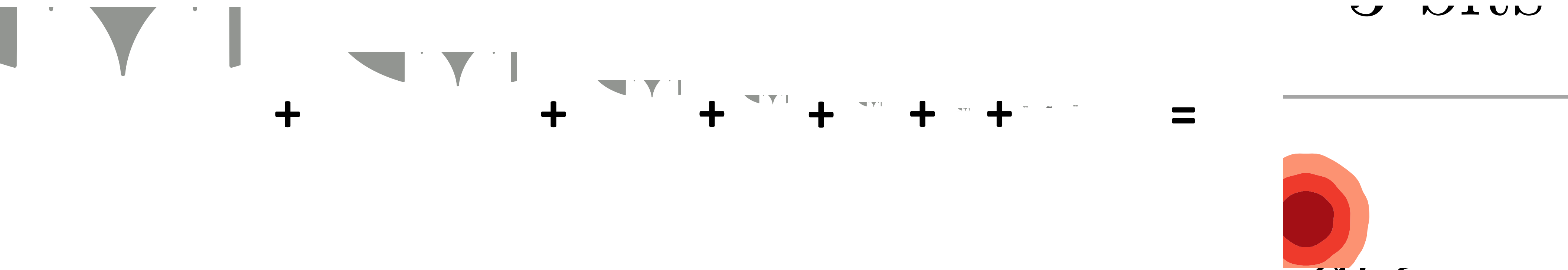
Example: a non-Gaussian input distribution

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

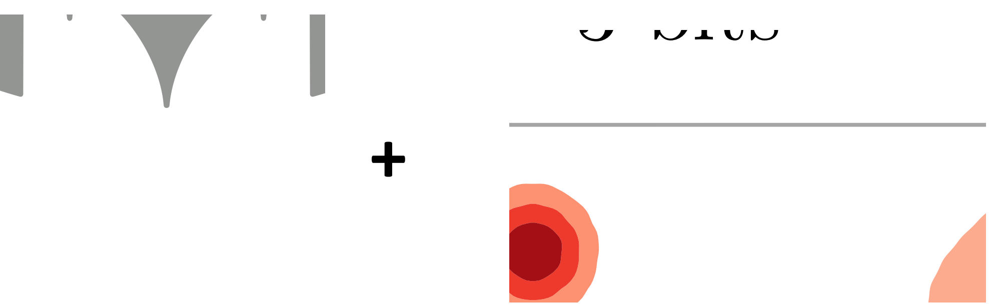


$q(\mathbf{x}$

Echo example, z=x+s ε, with s=1/2



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



$$\mathbf{z} = \mathbf{x} +$$

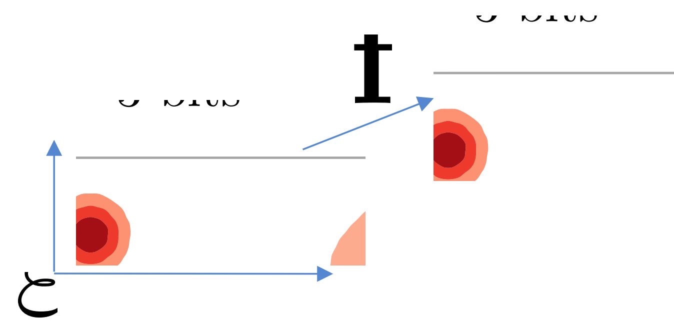
Sample-dependent noise scaling

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

$$= \mathbb{E}_{\mathbf{x}} H(f(\mathbf{x}) + S(\mathbf{x}$$

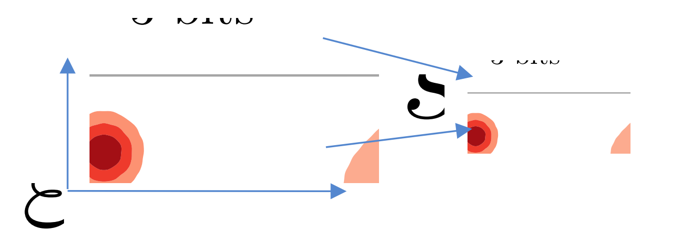
$$= \mathbb{E}_{\mathbf{x}} H(S(\mathbf{x})\mathcal{E} \mid X =$$

$$= H(\mathcal{E}) + \mathbb{E} \log |\mathrm{de}$$



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

Translation invarian



$$H(s\mathcal{E}) = H(\mathcal{E}) -$$

Scale property

lutual information for the echo noise
hannel

Sample-dependent noise s

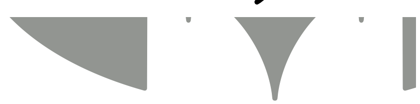
Self-similar noise (Echo)

Mutual information decom

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

$$Z;X) = H(Z) - H(Z|$$

$$(Z;X) = -\mathbb{E}\log|$$

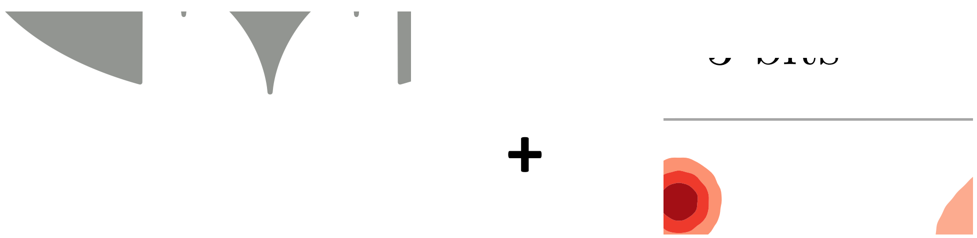


+

Mutual information for the echo noise hannel

- 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

$$(Z;X) = -\mathbb{E}\log|$$



Echo results: arXiv:1904.07199

- Exact mutual information, rather than bound
- Better log likelihood bounds
- Better rate-distortion trade-offs
- Simpler than other state-of-the-art methods that require a complicated "autoregressive flow" model to parametrize non-Gaussianity

Bigger question: How should we compress?

Simple noise (Gaussians, dropout) may not be optimal.

Variational Evidence Lower BOund (ELBO)

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

Distortion / reconstruction loss

Compression

Reconstruction and higher order interactions

Higher order dependence

Total Correlation / multivariate mutual information (Watanabe, 1960)

$$TC(X) = \sum_i H(X_i)$$

- Maximized when variables are redundant

"Cohesion" measure of order k

$$C^{(k)}(X) = \frac{}{\binom{n-1}{k-1}} \sum_{X_A \in \mathcal{E}_k} H(X$$

- Papers discussing: (Fujishige, 1978), Nihat Ay information geometry 2007/2011
- Maximizers correspond to error correcting codes: arXiv:1811.10839

ELBO for higher order interactions

Normal ELBO (can be derived from TC)

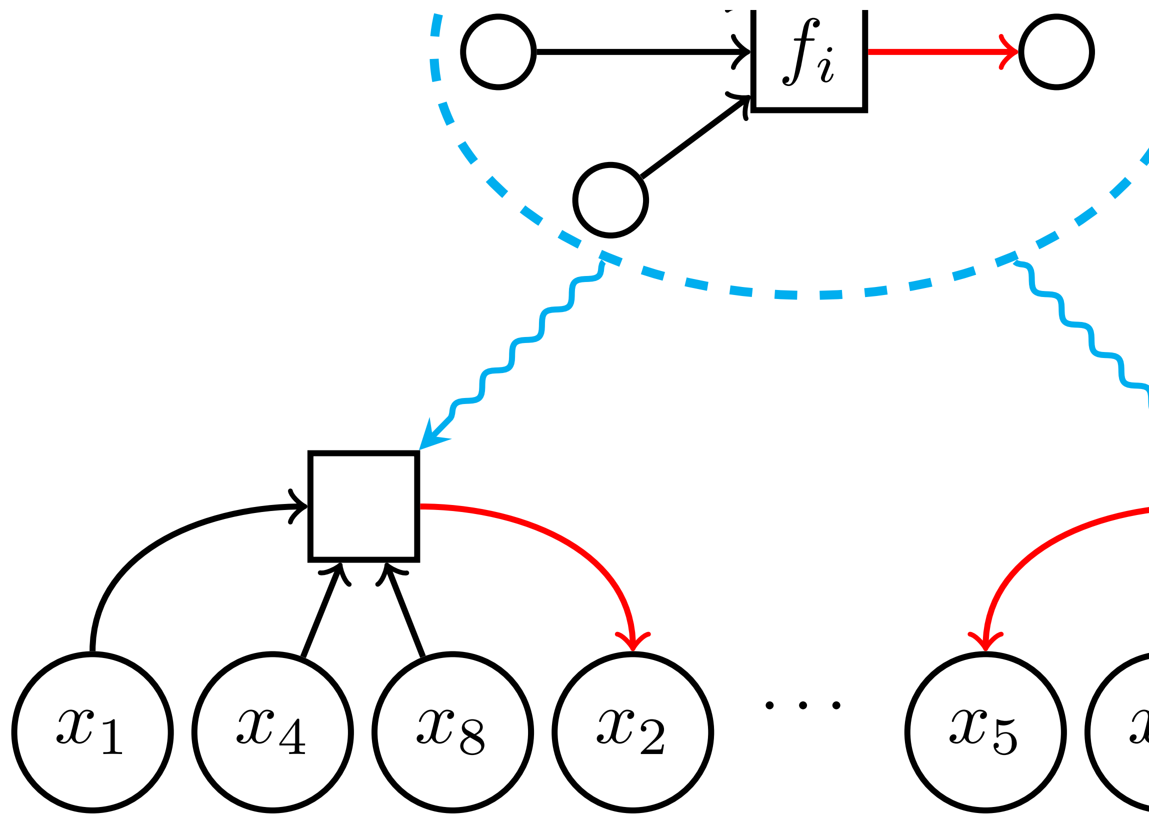
$$\mathbb{E} \log p(x) \leq \sum_i \mathbb{E} \log p(x_i | z)$$

New ELBO based on C^k

$$\mathbb{E} \log p(x) \geq \frac{}{\binom{n-1}{k-1}} \sum_{X_A \in \mathcal{E}_k} \mathbb{E} \log p(x$$

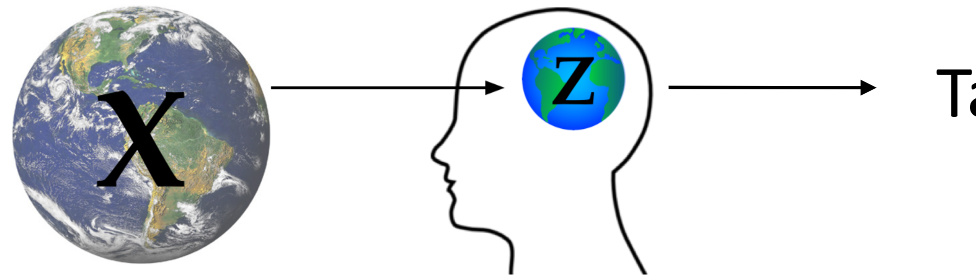
Can we use it to optimize VAEs to detect higher order interactions?
(preliminary results: maybe!)

Detect embedded synergies? (challenge idea?)



E.g.
J.C: *"fraudulent white noise"*
Kristian: information lost in high order correlations

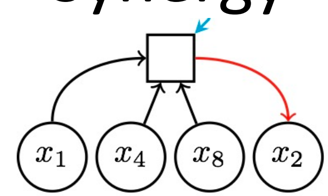
Conclusion/questions



Manipulating information in (high-d) representation learning

- Does it matter how we compress? (E.g. Gaussian vs echo)
- What do we want to reconstruct? (marginals versus higher order interactions)
- Fair / invariant representations − How to omit sensitive info?
- Is there an information principle to disentangle?
- Puzzles about the success of adversarial learning
- Do information bottlenecks improve generalization?

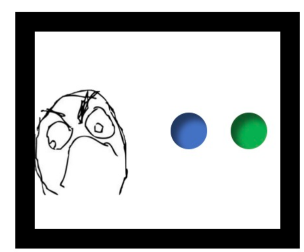
Synergy



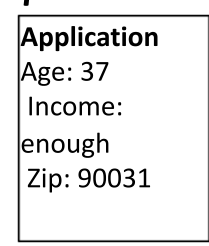
Practical issues:

- Appropriate measures*, that also have*:
- Efficient estimates or bounds, *and*
- Differentiable / smoothly optimizable

Grue



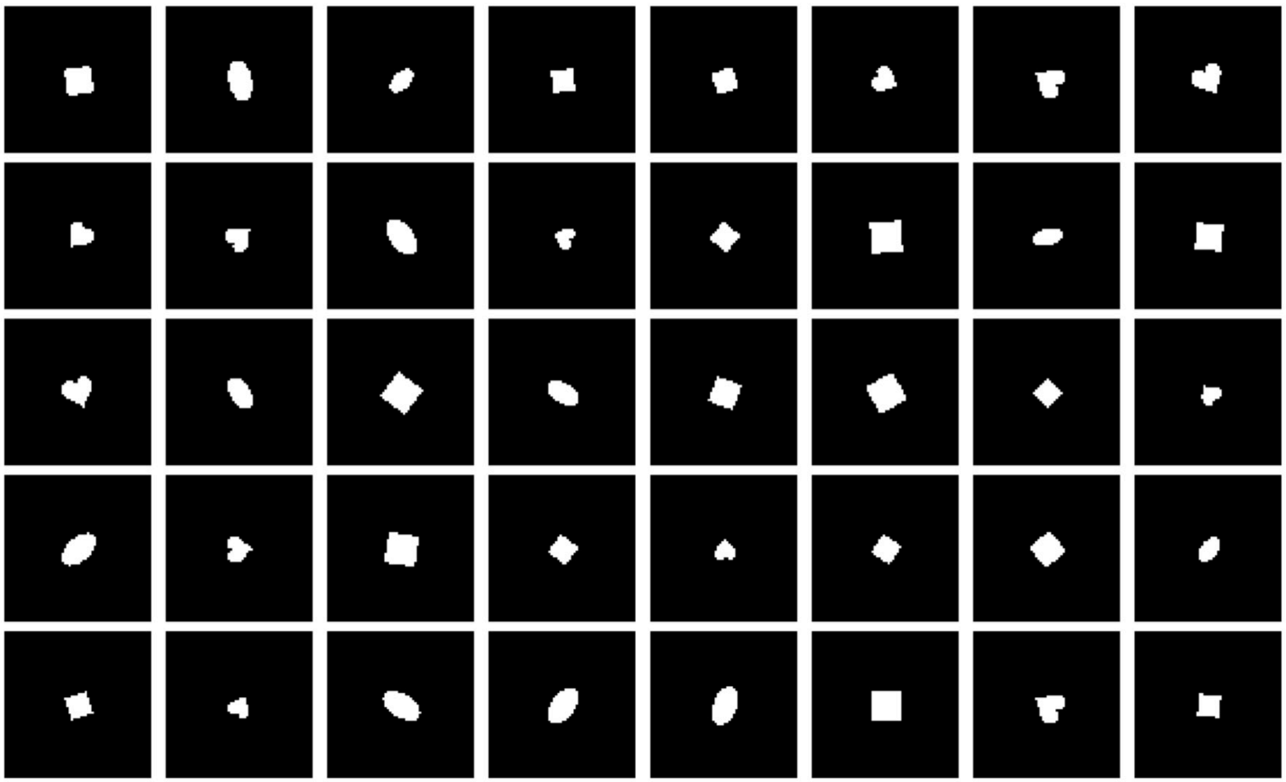
$$ Fair l



Contact: gregv@isi.edu
Get these slides: http://bit.ly/STEEG_BS

Disentangling toy data: dSprites

Suppose we observe the following dataset:



Factors:
- Scale
- Shape
- Orientation

This is part of a common artificial test set you'll see in the literature. Generally, we need a test set with known ground truth factors for evaluation

Adversarial learning and Jensen Shannon Divergence

Examples generated by StyleGAN arXiv:181

Figure 2. Uncurated set of images produced by our style-generator (config F) with the FFHQ dataset. Here we used a

- enerative Adversarial Networks are the ate-of-the-art way to generate realistic nages.
- nsen-Shannon Divergence (JSD) tells us ow easily we can distinguish two stributions: in this case "real images" and enerated images".
- stead of minimizing JSD, adversaries that y to distinguish provide a **lower bound** on D which is then **minimized**

Neural nets as information bottlenecks

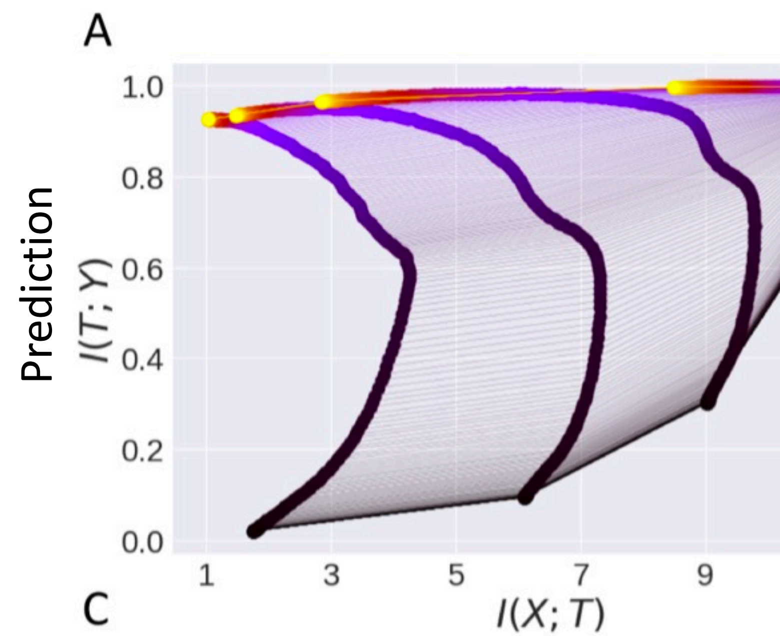
inputs) − T (representation) − Y (labels to predict)

$$\max_{T} I(T;Y) - \beta I$$

wartz-Ziv and N. Tishby. Opening the black
ox of deep neural networks via information.
Xiv:1703.00810, 2017.

N. Nets function as bottlenecks with a
"fitting phase" and "compression phase"

Compression aids generalization



Compression

xe et al. "On the information bottleneck
eory of deep learning", ICLR 2018.

Counter-examples for each of Tishby's claims

One problem for Tishby's analysis is estimating mutual information

For continuous random variables, if z=f(x), I(Z;X) $\rightarrow \infty$

Tishby discretizes to estimate the mutual information…

If you discretize differently (or apply invertible transformations with the same discretization), you get different results

Alternatives:

- Add noise to bound mutual information of continuous variables
- Other estimators?

Jse neural nets to estimate mutual nformation

Mixture of Gaussians

- Kolchinsky and Tracey, arXiv:1706.02419

Donsker-Varadhan (lower bound on KL/mutual information)

- Belghazi, Mohamed Ishmael, et al. "Mine: mutual information neural estimation." arXiv preprint arXiv:1801.04062 (2018)
- Poole, Ben, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and Georg Tucker. "On variational lower bounds of mutual information." In NeurIPS Workshop on Bayesian Deep Learning. 2018.

Impossible with few samples?

- McAllester, David, and Karl Statos. "Formal Limitations on the Measurement of Mutual Information." *arXiv preprint arXiv:1811.04251* (2018).

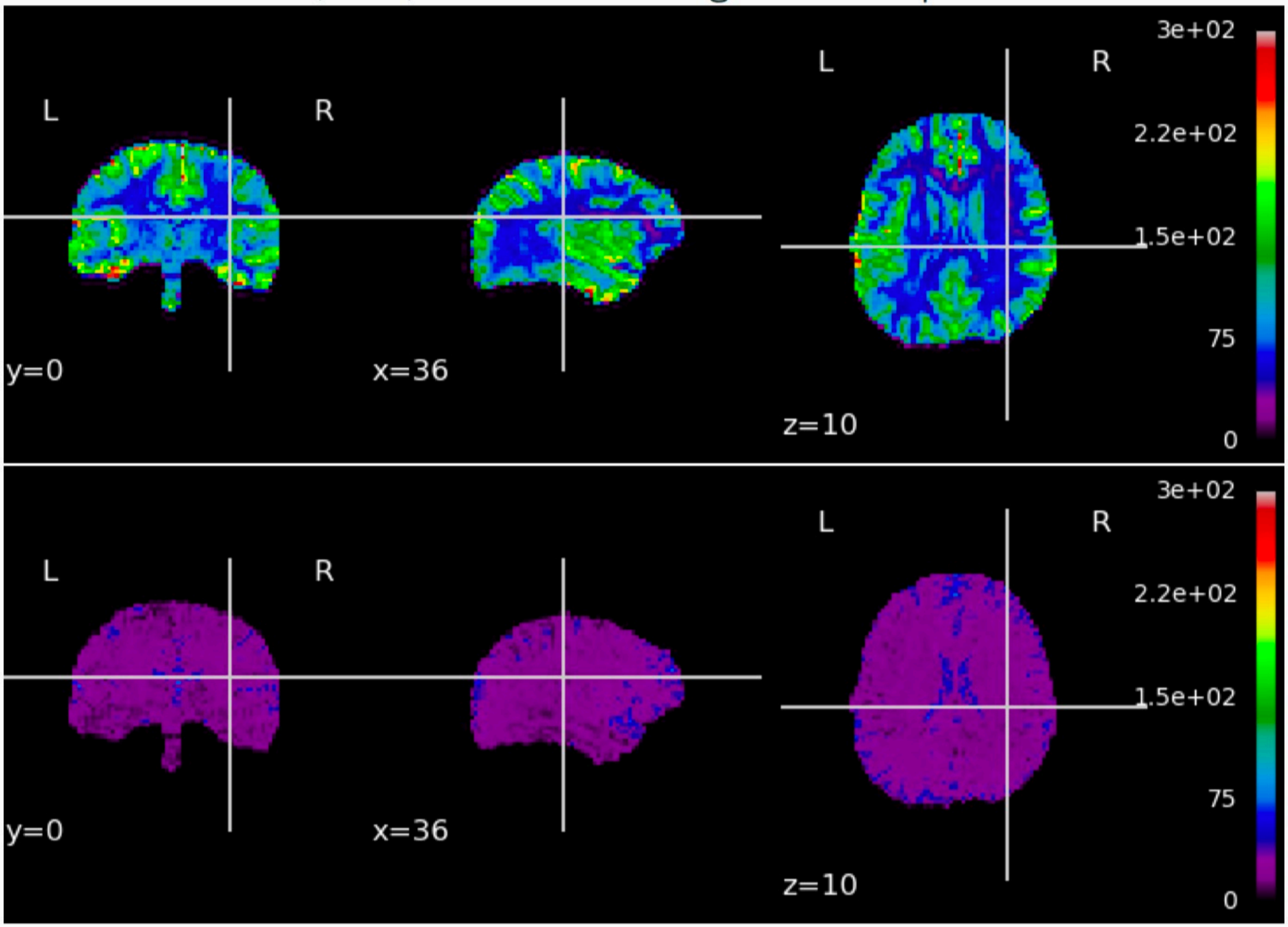
Other papers I forgot

The Information Complexity of Learning Tasks, their Structure and their Distance. Alessandro Achille ∗ & Giovanni Paolini † & Glen Mbeng ‡ & Stefano Soatto ∗

Information-Theoretic Lower Bounds on BayesRisk in Decentralized Estimation. Aolin Xu and Maxim Raginsky Senior Member, IEEE

Error Plots – Prisma 30 to Connectom 30

Mirzaalian et al. (2018) Baseline vs Single-site Proposed Method

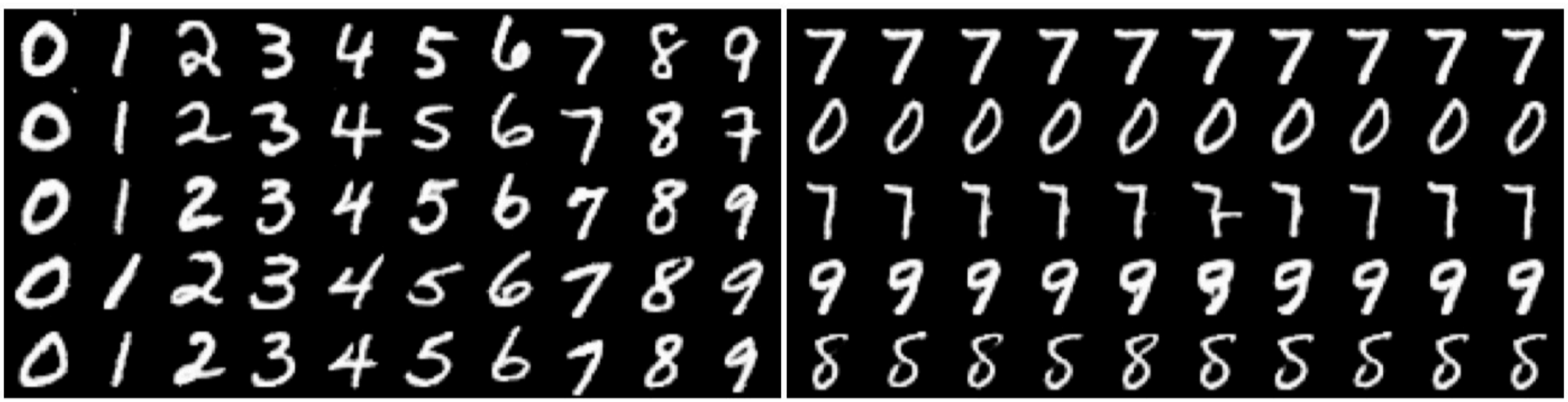


nterpretability through mutual information regularization

i Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel (2016) InfoGAN: nterpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Guide most information into a special neuron, c_1

Max Loss + $I(c_1; x)$



(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)

Another example of this phenomenon: arXiv:1802.05822