Controlling the Black Box: Learning Manipulable and Fair Representations

Richard Zemel

Vector Institute and University of Toronto

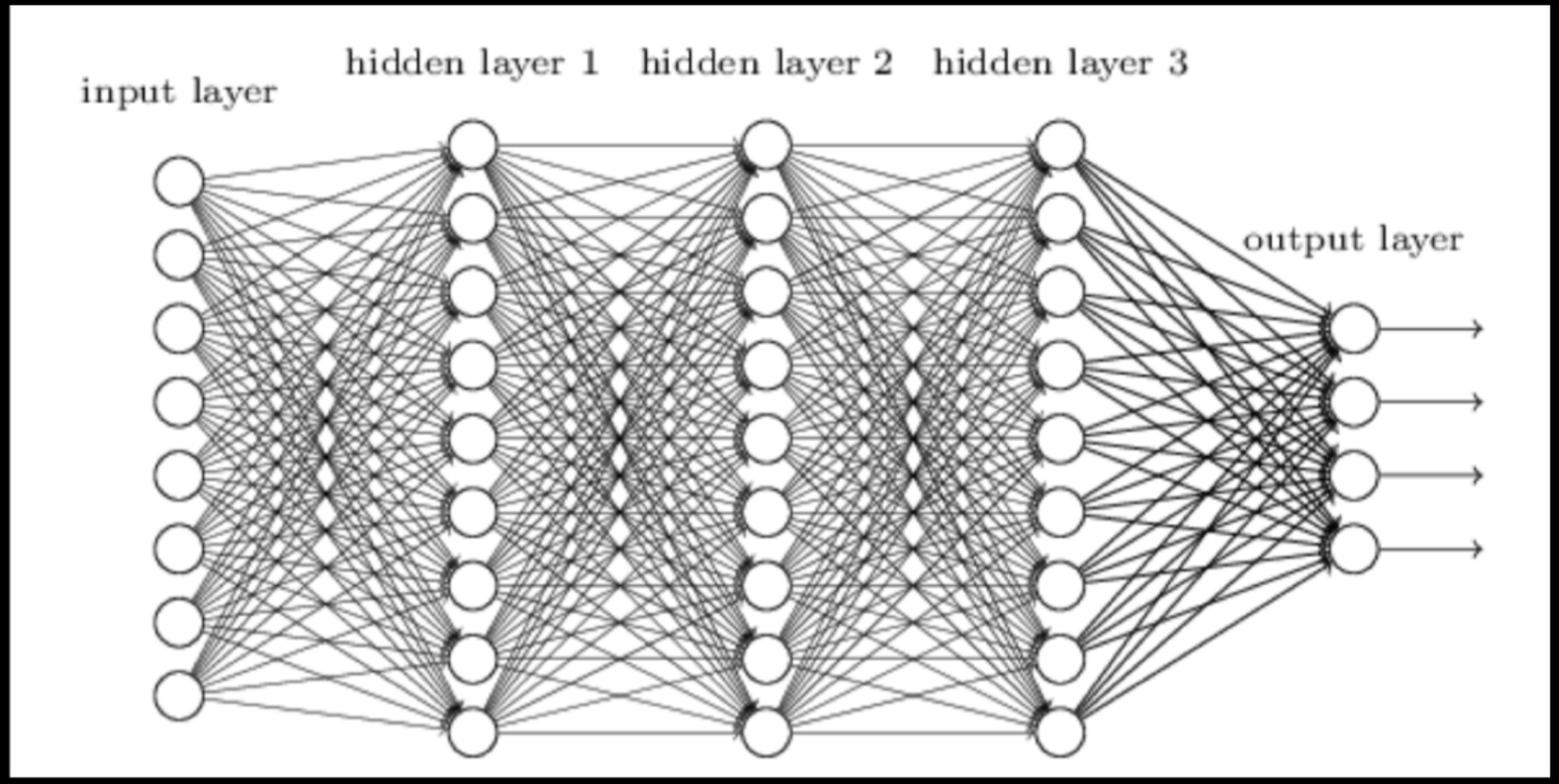
Deep Learning Defined

Deep Learning model: neural network that maps input to output via a series of simple transformations (layers)

Each layer composed of units

Simple computation determines unit's activity

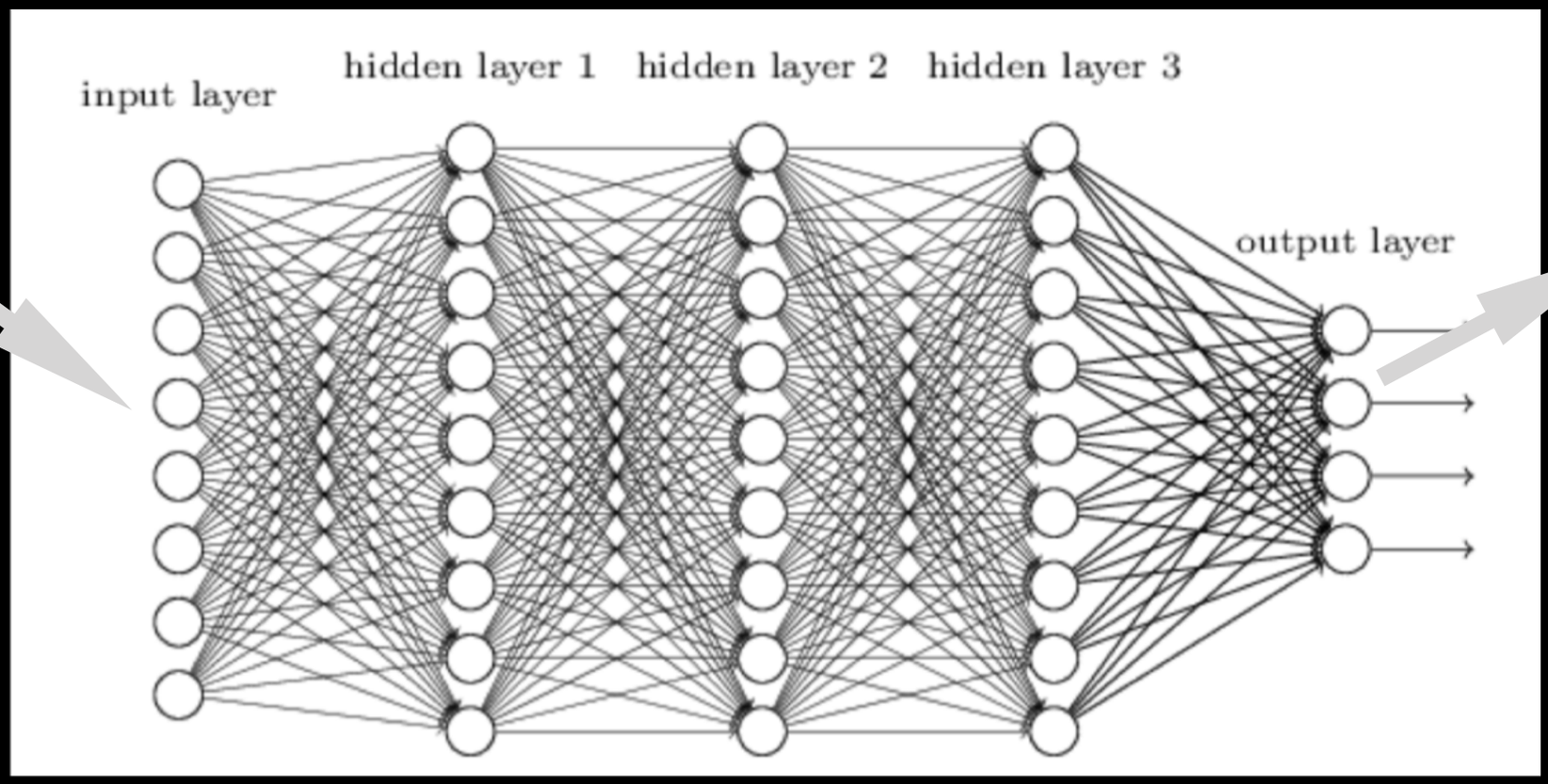
Key: values of weights



Deep Learning Defined

Deep Learning model: neural network that maps input to output via a series of simple transformations (layers)

Each layer of the network – a new representation of the input



“HIPPO”

Example:

- Image classification– each pixel an input dimension, output = cat, dog, etc.

Progress in Machine Learning

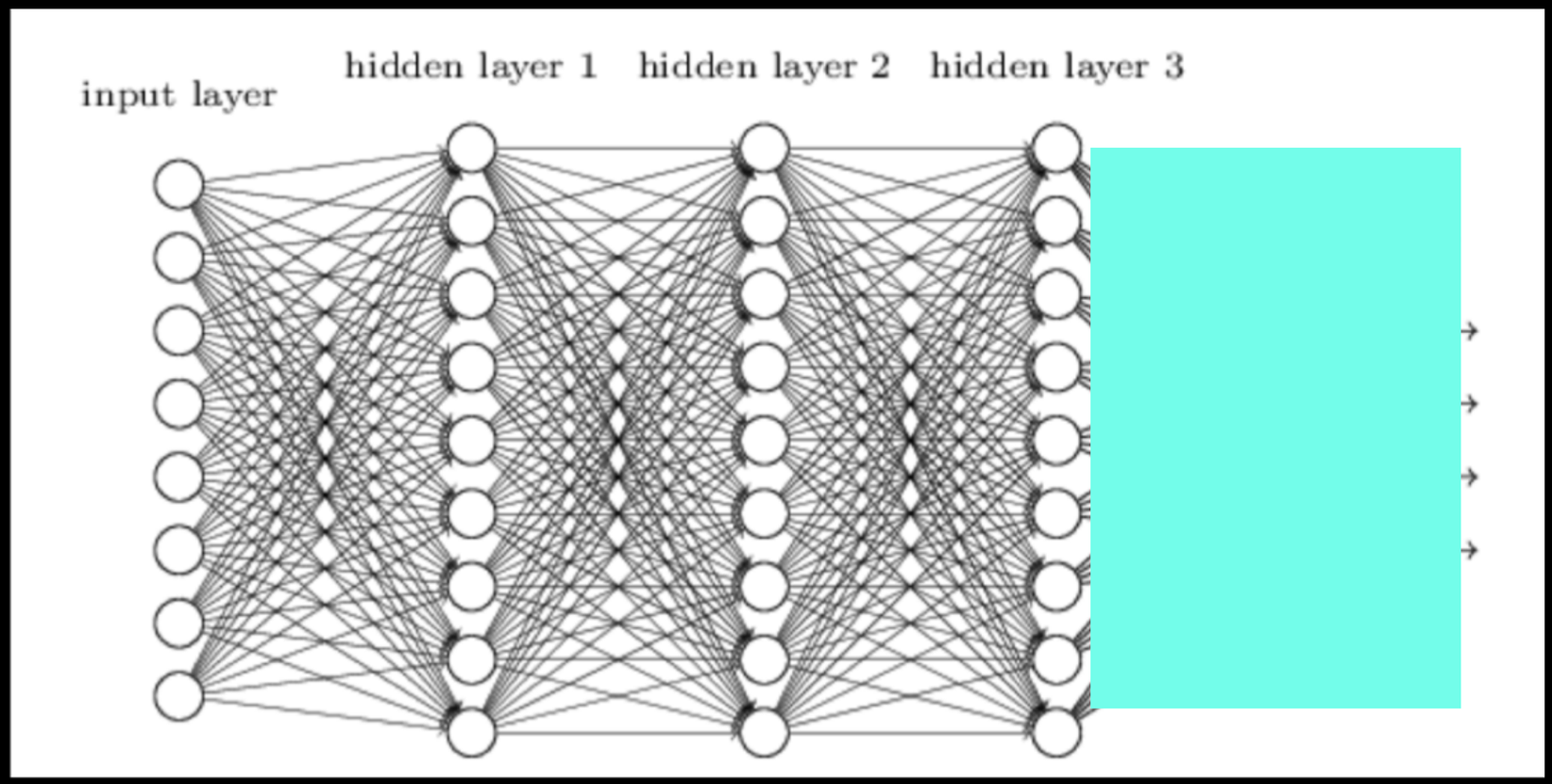
Supervised learning: Deep Learning’s biggest success stories

- Recognizing objects in images
- Machine translation: English sentences –> French sentences
- Mapping sensors (images, LIDAR) to controls for driving

Most noteworthy advance: not performance on individual tasks, but in developing the intermediate representations

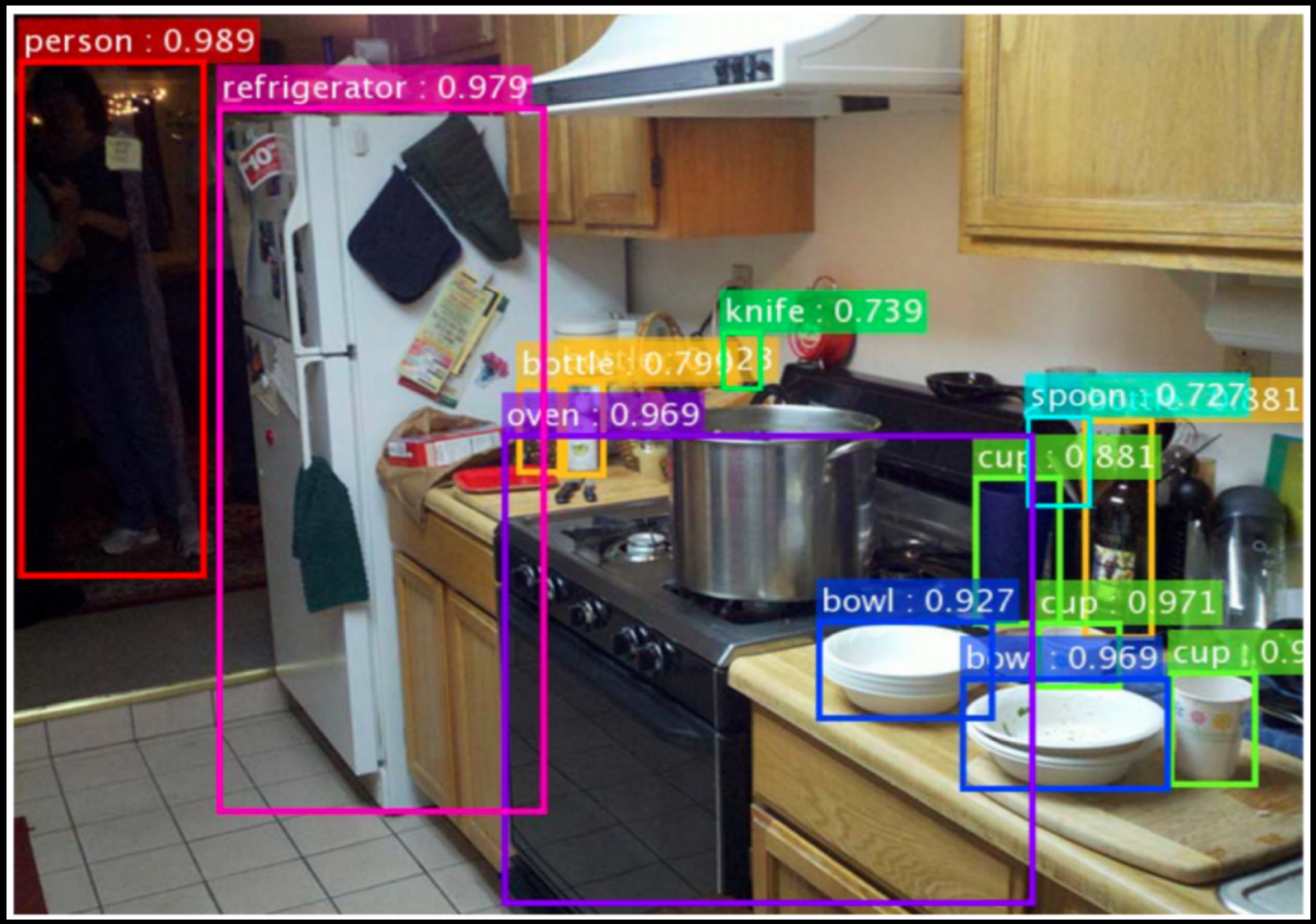
Multi-Purpose Image Representations

Most intriguing result: decapitate network -> penultimate layer representations useful for many other image tasks

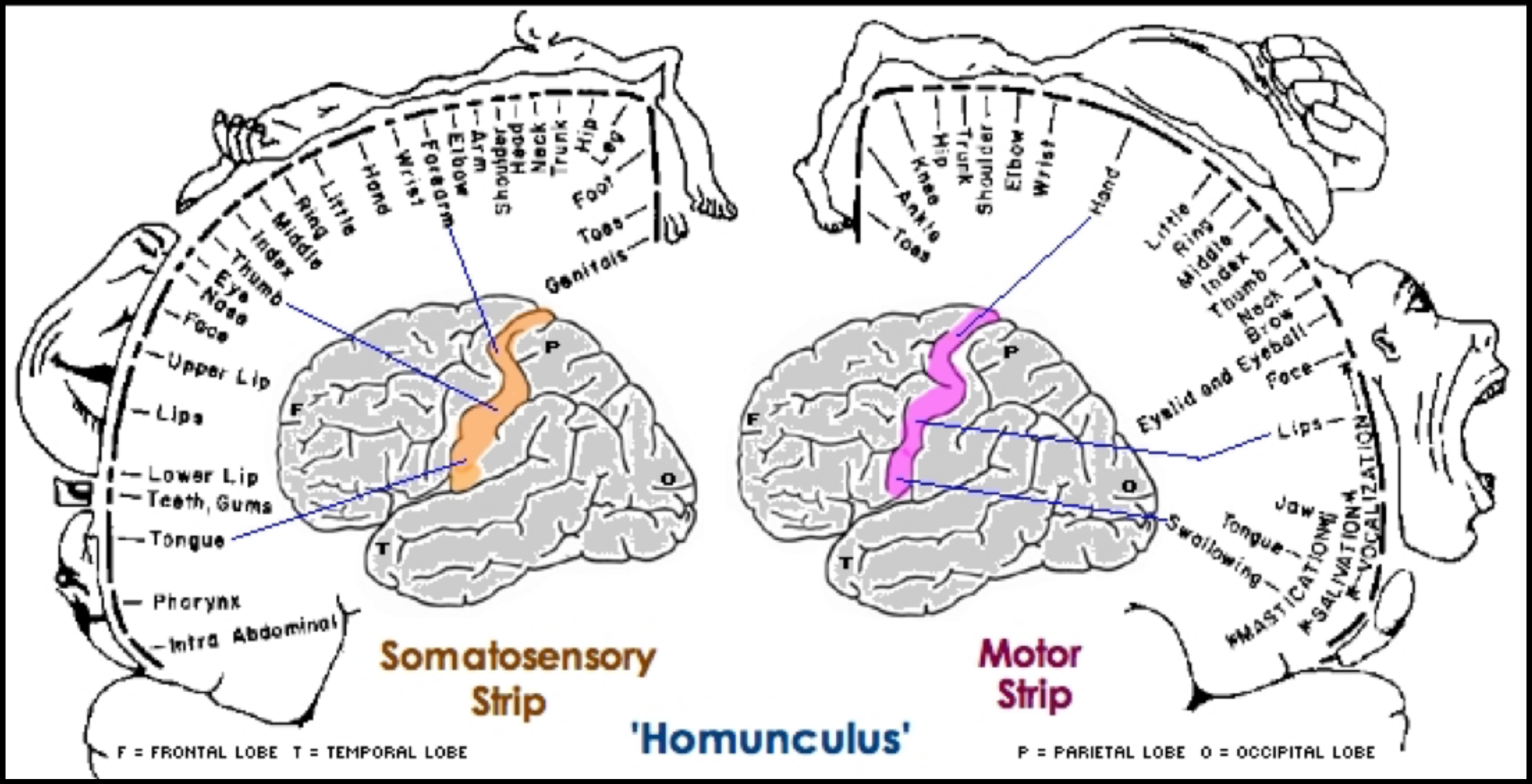


Multi-Purpose Image Representations

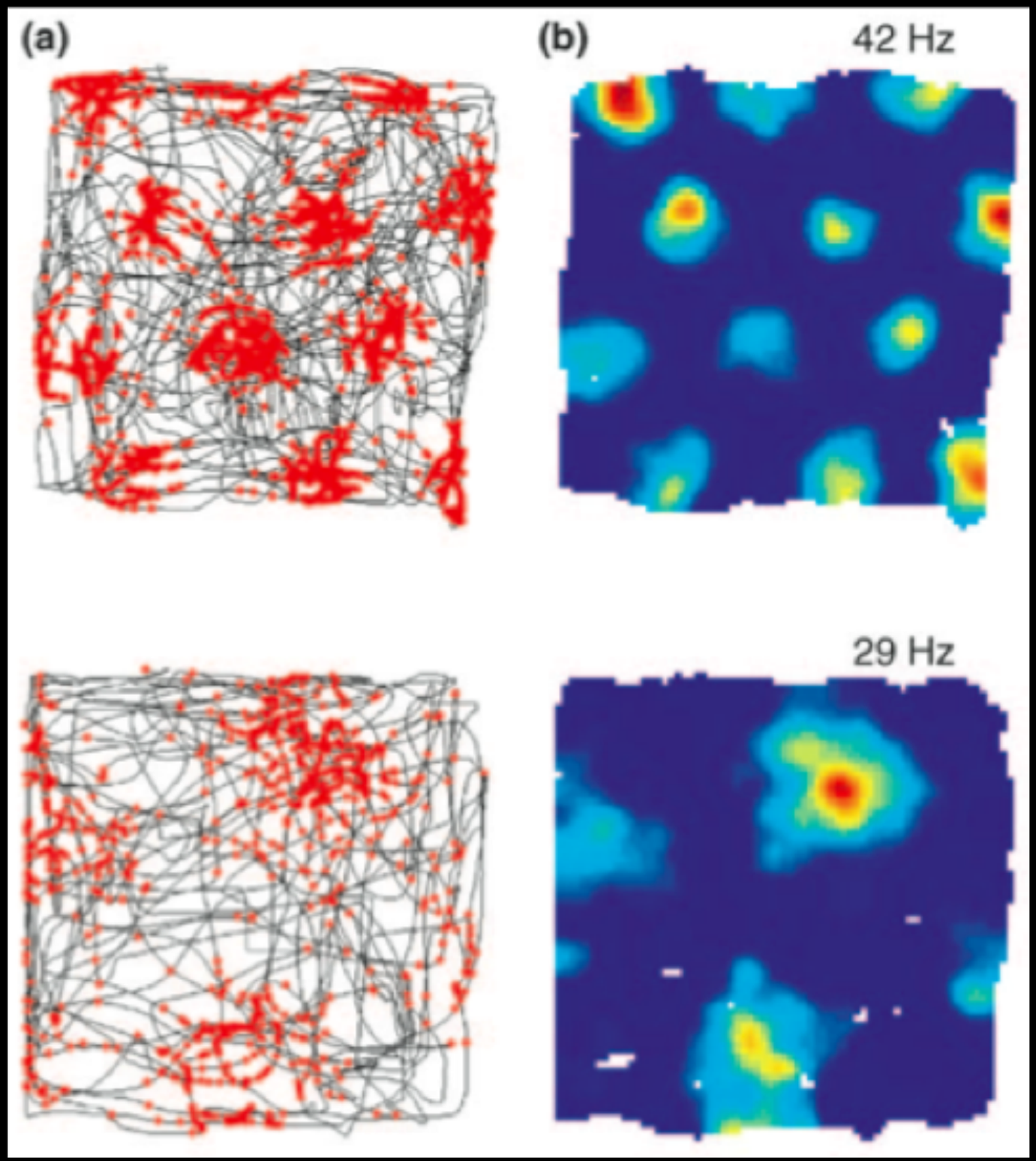
Most intriguing result: decapitate network -> penultimate layer representations useful for many other image tasks



It’s Representations All the Way Down



Generally Useful Neural Representations



Aim: Study & Develop Strong Representations

How to develop good representations, that can be useful in multiple tasks?

Key: Generative Model – system that can produce inputs

Motivation:

- Test that information not lost about input
- Generalize: generate related but novel inputs
- Construct density model of inputs – anomaly detection

Autoencoder

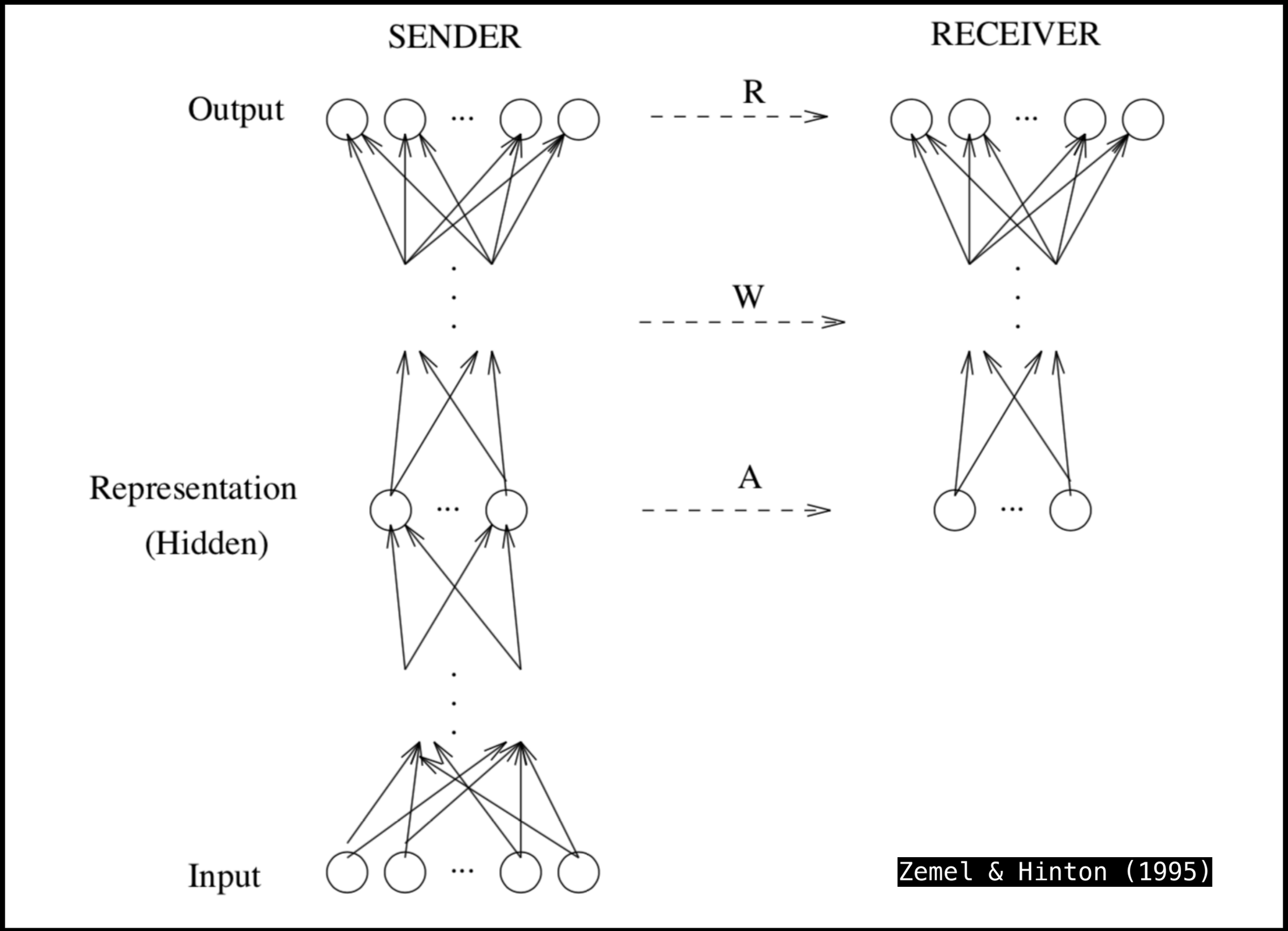
Original formulation of generative model

Main idea:

- Encoder maps input to a vector, via a multi-layer neural network
- Decoder maps vector to an estimate of the input

Objective used to optimize/learn the network parameters based on information theoretic formulation: bits required to recover original input given the estimate

Early Autoencoder Formulation



Zemel & Hinton (1995)

Outline

- Key aim: Learn strong representations via generative model

- Examples of approach, utility:

  1. Expose variables to control fabrication of new items
  2. Develop fair automated decision makers
  3. Build more robust classifiers

- Conclusions & current directions

Outline

- Key aim: Learn strong representations via generative model

- Examples of approach, utility:

  1. Expose variables to control fabrication of new items

  2. Develop fair automated decision makers

  3. Build more robust classifiers

- Conclusions & current directions

Conditional Subspace Autoencoder

Ideally can discover relevant underlying structure in purely unsupervised manner

However, ill-posed problem to find interpretable, manipulable representations

Assume access to additional labels, e.g., image tags

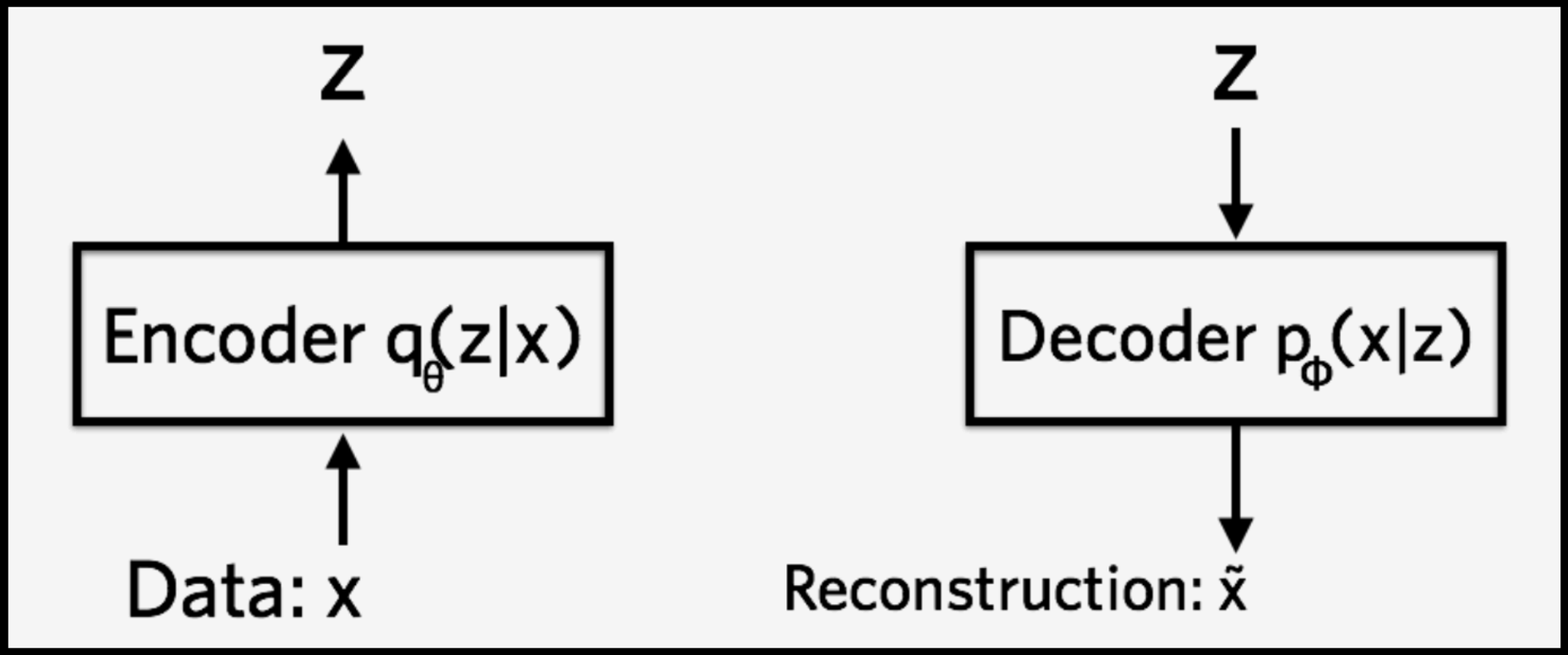
Goals:

1. Learn representations that uncover latent structure correlated with labels (conditional subspace)

2. Expose these representations – easy to interpret and manipulate when generating or modifying data

Background: Variational Autoencoder (VAE)

Re-formulation of autoencoders:
- Each input encoded into a distribution in latent space
- Output prediction obtained by sampling from distribution, mapping through decoder



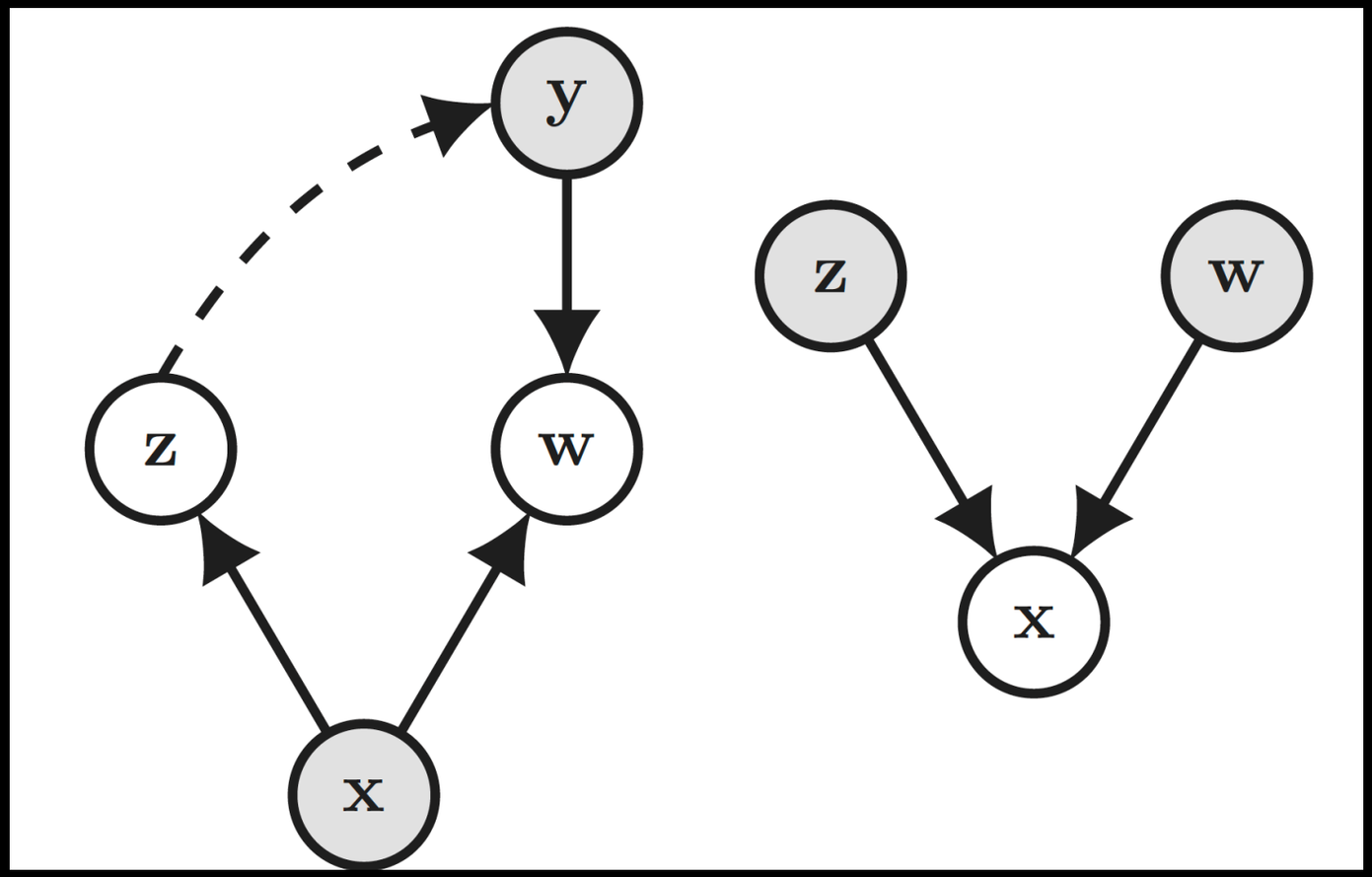
Allows maximum-likelihood based density modeling:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right)$$

Conditional Subspace VAE: Graphical Model

Assume access to additional relevant information $\mathbf{y}$

Form rich latent representation of $\mathbf{y}$



Encoder Decoder

CSVAE: Training Objective

$$\min_{\theta,\phi,\gamma} \beta_1 \mathscr{M}_1 + \beta_2 \mathscr{M}_2 \qquad \max_{\delta} \beta_3 \mathscr{N}$$

Extend standard VAE: model of observations x,y

$$\mathscr{M}_1 = \mathbb{E}_{\mathscr{D}(\mathbf{x},\mathbf{y})}\{-\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{w}|\mathbf{x},\mathbf{y})}\left[\log p_\theta\left(\mathbf{x}|\mathbf{w},\mathbf{z}\right)\right] + D_{KL}\left(q_\phi\left(\mathbf{w}|\mathbf{x},\mathbf{y}\right) \parallel p_\gamma\left(\mathbf{w}|\mathbf{y}\right)\right)$$
$$+D_{KL}\left(q_\phi\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right) \parallel p\left(\mathbf{z}\right)\right) - \log p\left(\mathbf{y}\right)\}$$

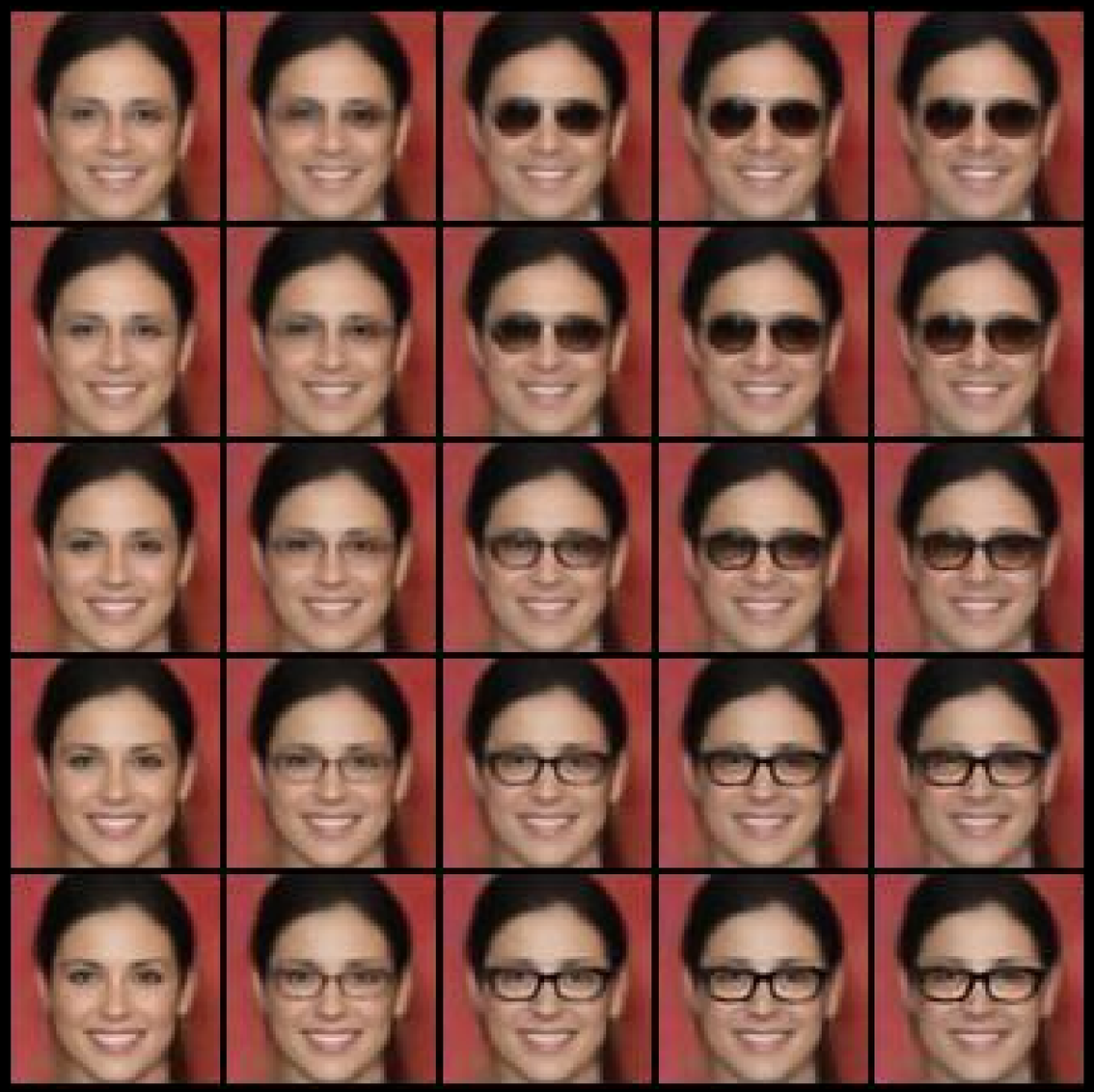
Adversary trying to maximize information about y in z

$$\mathscr{N} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})\mathscr{D}(\mathbf{x},\mathbf{y})}\left[q_\delta\left(\mathbf{y}|\mathbf{z}\right)\right]$$

Model fighting adversary: minimize information about y in z

$$\mathscr{M}_2 = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})\mathscr{D}(\mathbf{x})}\left[\int_Y q_\delta\left(\mathbf{y}|\mathbf{z}\right)\log q_\delta\left(\mathbf{y}|\mathbf{z}\right)d\mathbf{y}\right]$$

CSVAE: Exploring Latent Space



CSVAE: Joint Attribute Transfer

Richard Zemel

Vector Institute & University of Toronto

11/15/2019

CSVAE: Quantitative Evaluation

	Accuracy		
	TFD	CelebA-Glasses	CelebA-FacialHair
VAE	19.08%	25.03%	49.81%
CondVAE	62.97%	96.04%	88.93%
CondVAE-*info*	62.27%	95.16%	88.03%
CSVAE (ours)	**76.23%**	**99.59%**	**97.75%**

	target - changed	original - changed	target - original
VAE	75.8922	13.4122	91.2093
CondVAE	74.3354	18.3365	91.2093
CondVAE-*info*	74.3340	18.7964	91.2093
CSVAE (ours)	**71.0858**	28.1997	91.2093

(#3) Results

We were barely able to catch the breeze at the beach , and it felt as if someone stepped out of my mind . She was in love with him for the first time in months , so she had no intention of escaping . The sun had risen from the ocean , making her feel more alive than normal . She 's beautiful , but the truth is that I do n't know what to do . The sun was just starting to fade away , leaving people scattered around the Atlantic Ocean . I d seen the men in his life , who guided me at the beach once more .

(#3) Results

You re the only person on the beach right now
you know
I do n't think I will ever fall in love with you
and when the sea breeze hits me
I thought
Hey

Generating with Style

Generated story about image
Model: Romantic Novels

“He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder.

He wanted to strangle me, considering the beatiful boy I’d become wearing his boxers.”

Outline

- Key aim: Learn strong representations via generative model

- Examples of approach, utility:

  1. Expose variables to control fabrication of new items
  2. Develop fair automated decision makers
  3. Build more robust classifiers

- Conclusions & current directions

Fairness in Automated Decision Making

Algorithmic unfairness: Algorithms are pervasive, high-stakes, high-impact

Need more than just ”accuracy”

Fair Classification

Explosion of fairness research over last five years

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a sensitive attribute (race, gender, age, SES)

We want to learn a classifier that is:

- accurate
- fair with respect to A
-

Fair Representations

Classification: a tale of two parties

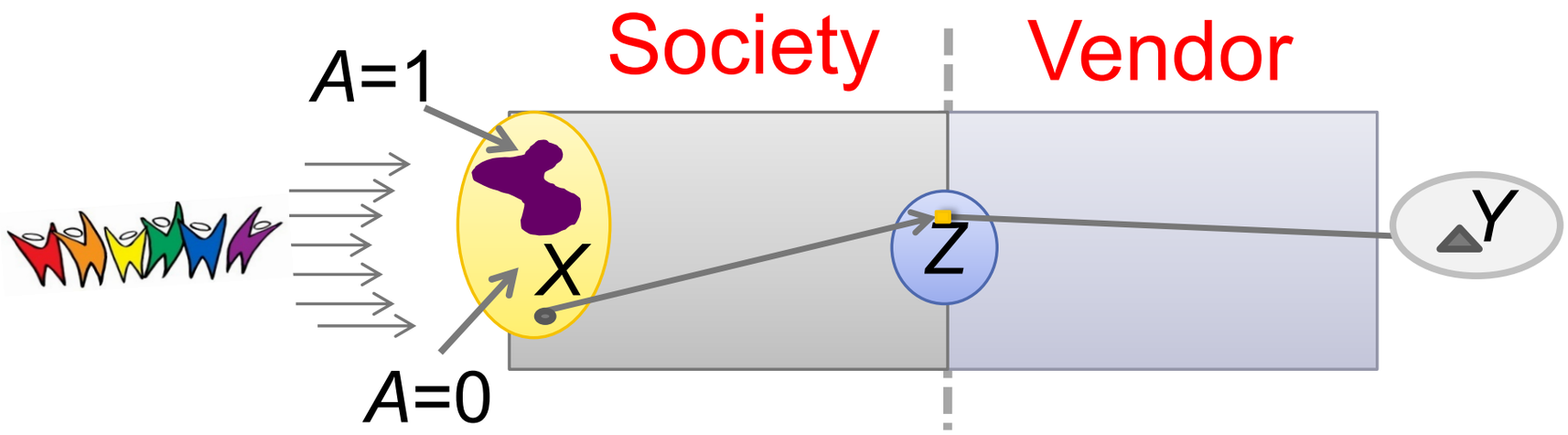
- Example: **targeted advertising**: Owner –> Vendor –> Prediction
-

Data Owner

Vendor

FAIRNESS THROUGH AWARENESS

Dwork, Hardt, Pitassi, Reingold, Zemel, 2012



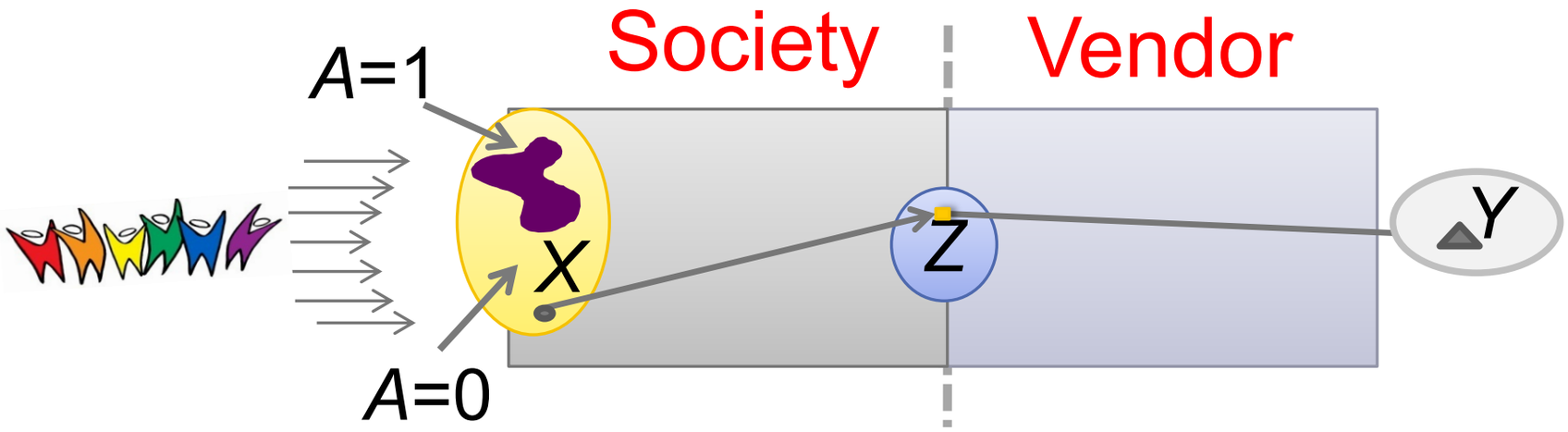
Goal: Assign individual *X* representation *Z by being aware of membership in group A*

(1). **Individual Fairness**: Treat similar individuals similarly

(2). **Group Fairness:** equalize two groups (A=1 = minority; A=0 is majority) at the level of outcomes (statistical parity)

FAIR REPRESENTATION LEARNING: FRAMEWORK

Zemel. Wu, Swersky, Pitassi, Dwork, 2013



Goal: Learn a mapping from X to distributions over representations *Z that is fair*

Aims for *Z*:

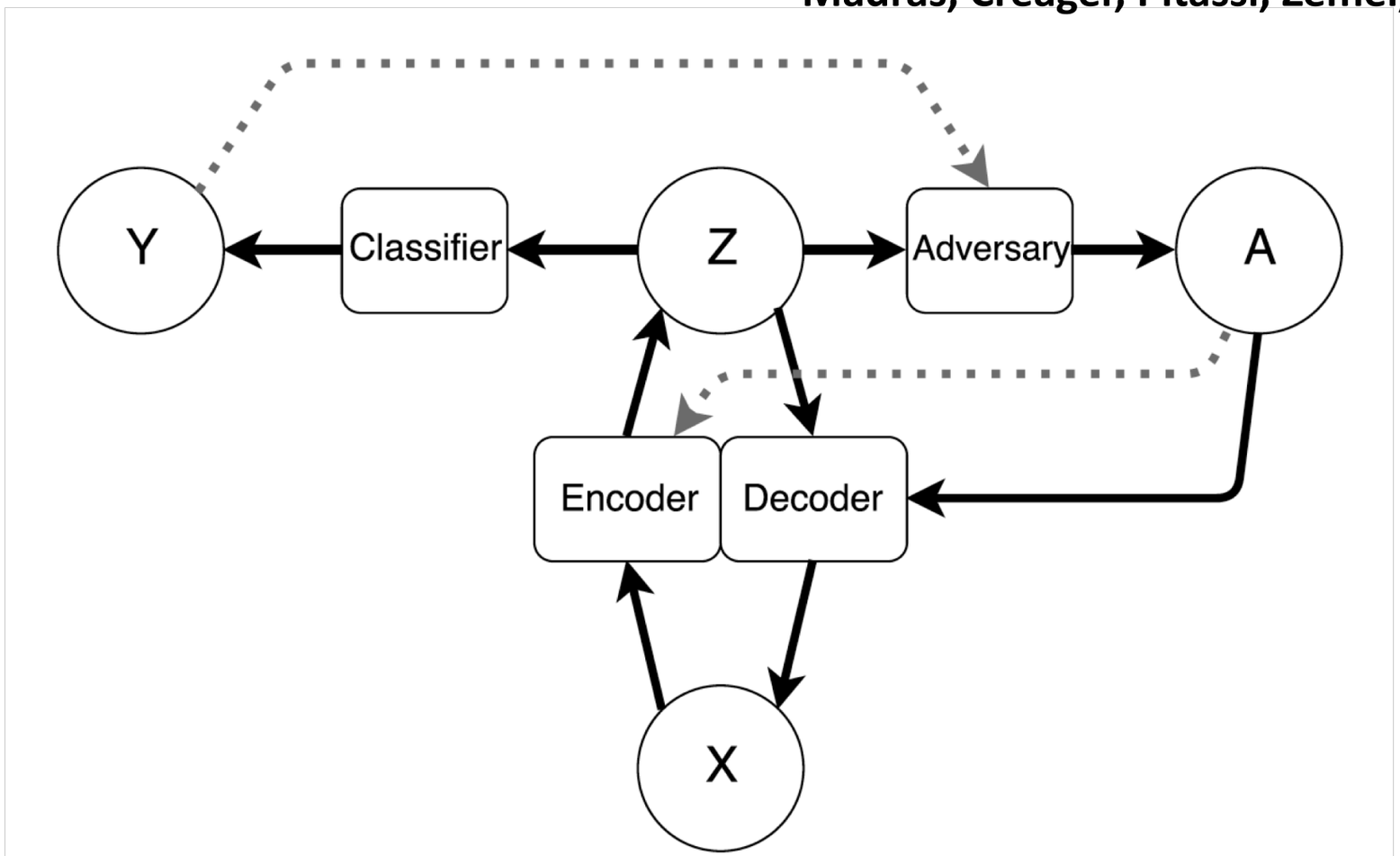
1. Lose information about A:

$$P[Z=k \mid A=1] = P[Z=k \mid A=0]$$

2. Retain information about X
3. Preserve information for classification so vendor can max utility [decisions Y = g(Z)]

LEARNING ADVERSARIALLY FAIR TRANSFERABLE REPRESENTATIONS

Madras, Creager, Pitassi, Zemel, 2018



- The classifier is indifferent vendor, forcing the encoder to make the representations useful

- The adversary is the malicious vendor, forcing the encoder to hide the sensitive attributes in the representations

Learning Flexibly Fair Representations

Important limitation: only considering single, fixed sensitive attribute

Typically several, and which ones apply to a single situation is not known a priori

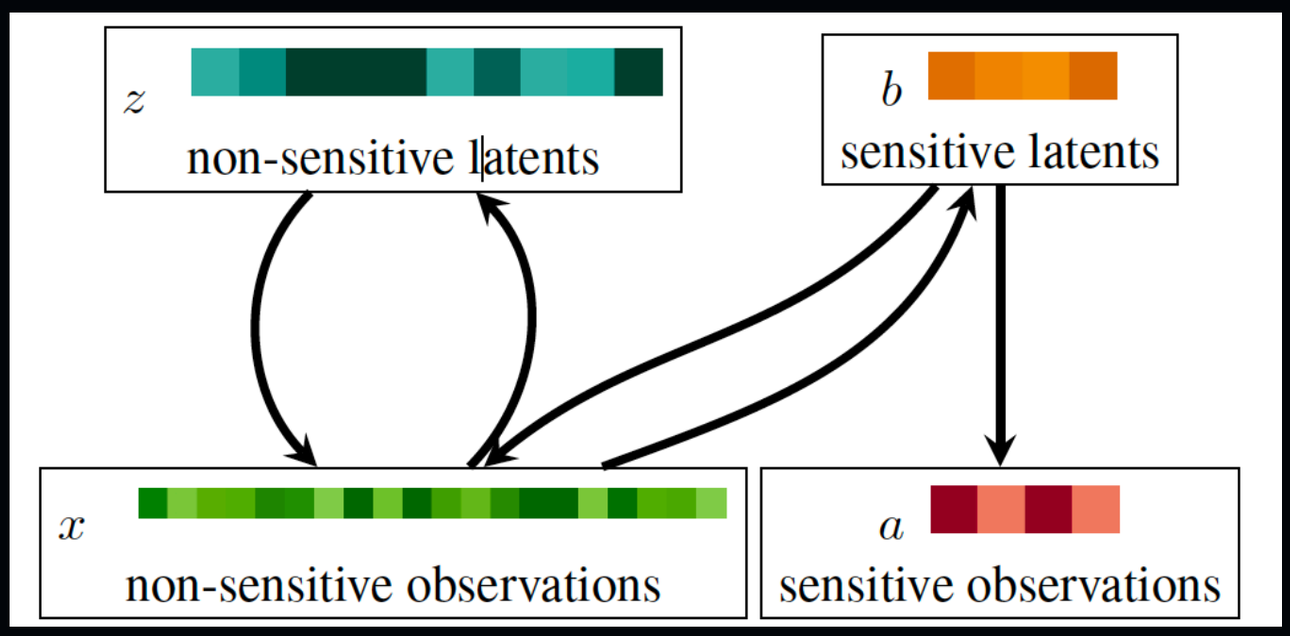
Aim: learn flexibly fair representations – can be adapted to a variety of protected groups and their intersections:

- *Simple:* easily adapt to different protected attributes
- *Compositional:* fair to conjunctions of variables (subgroup discrimination – fair to women, not black women over 60)
- *Transferrable:* same representation applies to several downstream tasks

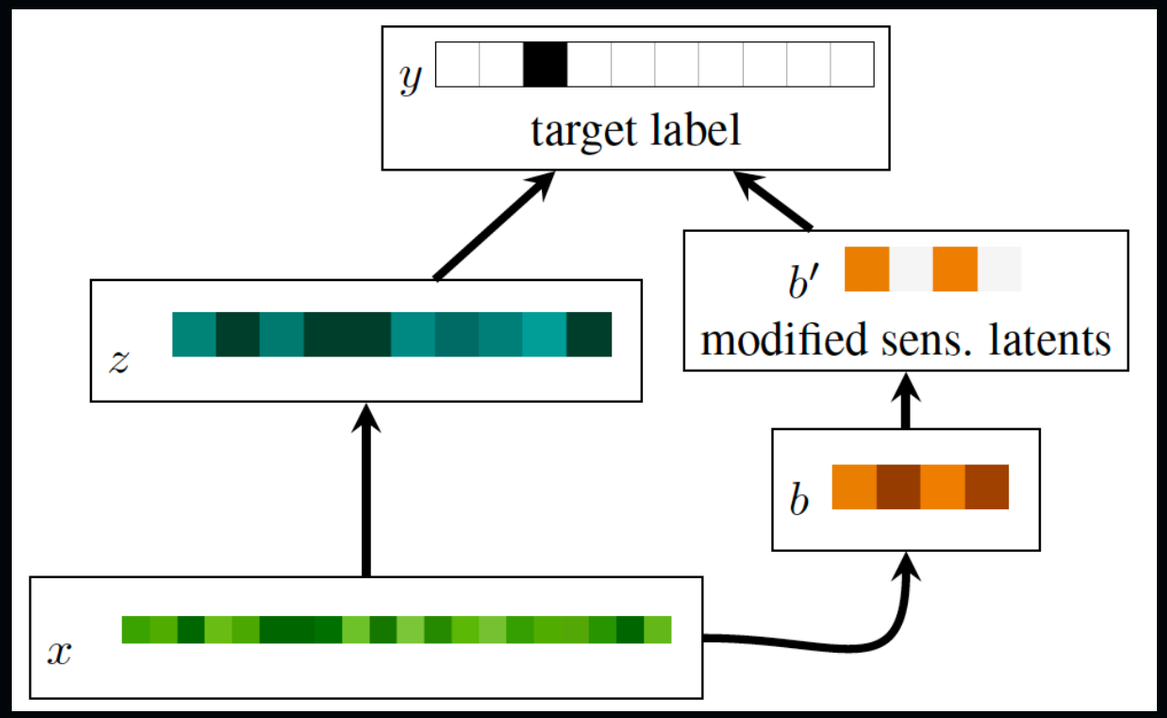
Flexibly Fair VAE (FFVAE)

Unsupervised training: no Y, but are given full range of sensitive attributes, values A

Test time: specify subset of sensitive attributes and target Y



Training



Testing

FFVAE Objective

Aim for latent representations that are:

1. *Predictive* – dimension of latent code corresponds to single underlying factor: high MI(bi, ai)
2. *Disentangled* – posterior factorizes
3.

reconstruction *predictiveness*

$$L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)}[\log p(x|z,b) + \alpha \log p(a|b)]$$
$$-\gamma D_{KL}[q(z,b)||q(z)\prod_j q(b_j)] - D_{KL}\left[q(z,b|x)||p(z,b)\right]$$

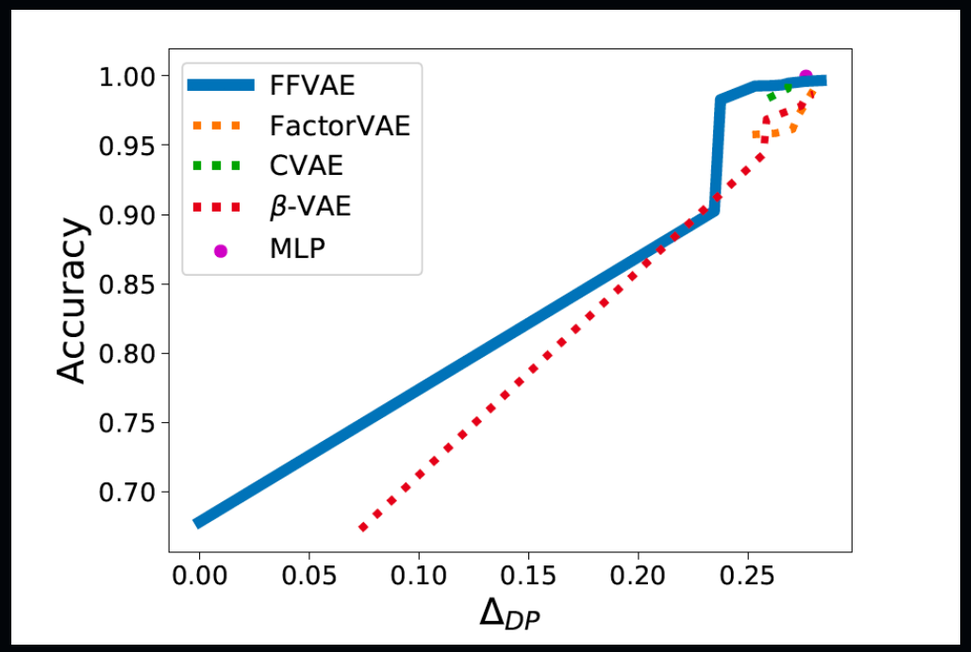
disentanglement *prior*

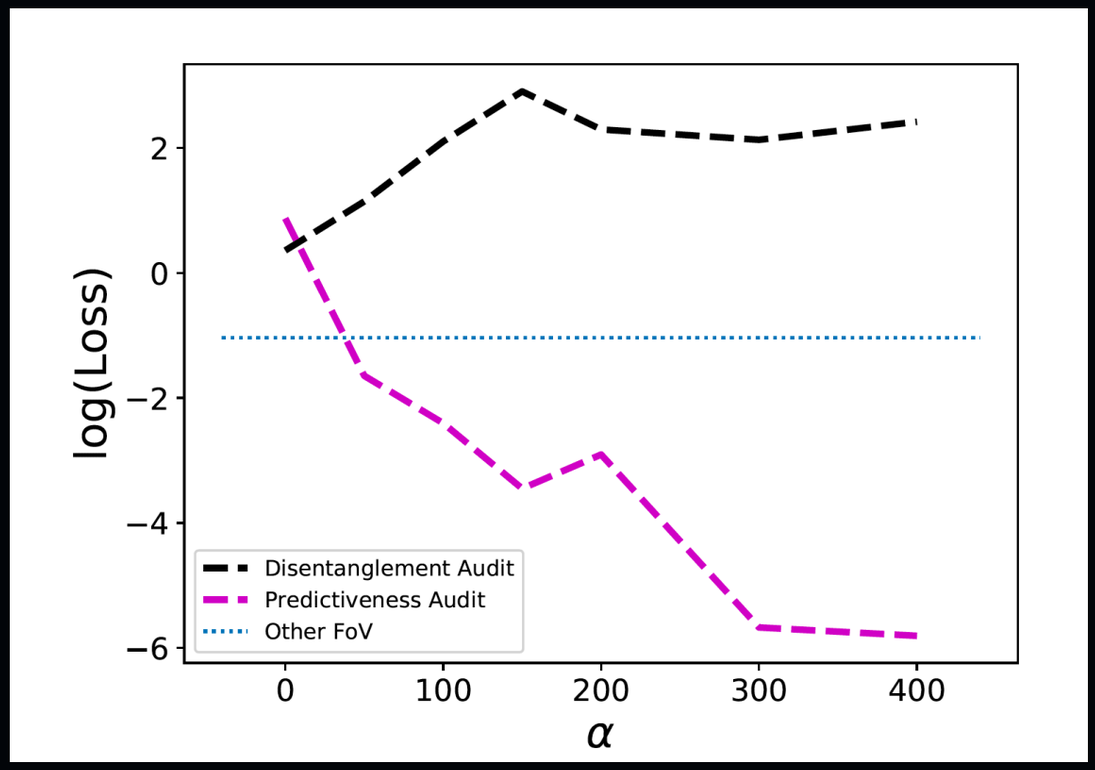
FFVAE Experiments

- **DSpritesUnfair**: vary in color, shape, scale, orientation, Xposition, Yposition; *Xposition* and *Shape* sensitive
  - Difficult due to positive correlation in sensitive attributes

- **Communities and Crime**: neighborhood statistics – sensitive attributes *racePctBlack, blackPerCap, pctNotSpeakEnglWell*; predict violentCrimesPerCap

- **CelebA**: *male, eyeglasses, chubby* as sensitive attributes; tested on pairwise combinations, predict heavy-makeup, then attractive





Outline

- Key aim: Learn strong representations via generative model

- Examples of approach, utility:

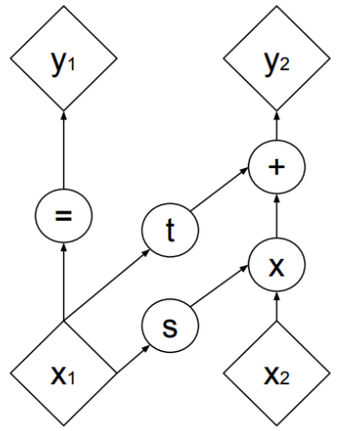
  1. Expose variables to control fabrication of new items
  2. Develop fair automated decision makers
  3. Build more robust classifiers

- Conclusions & current directions

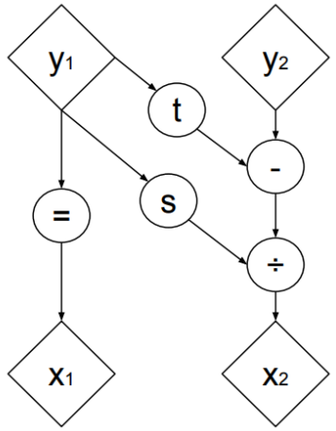
Invertible Deep Networks

- Fairly recent development in deep networks: invertible networks

- Invertible (aka Reversible, Bijective) Networks retain all information about input



(a) Forward propagation



(b) Inverse propagation

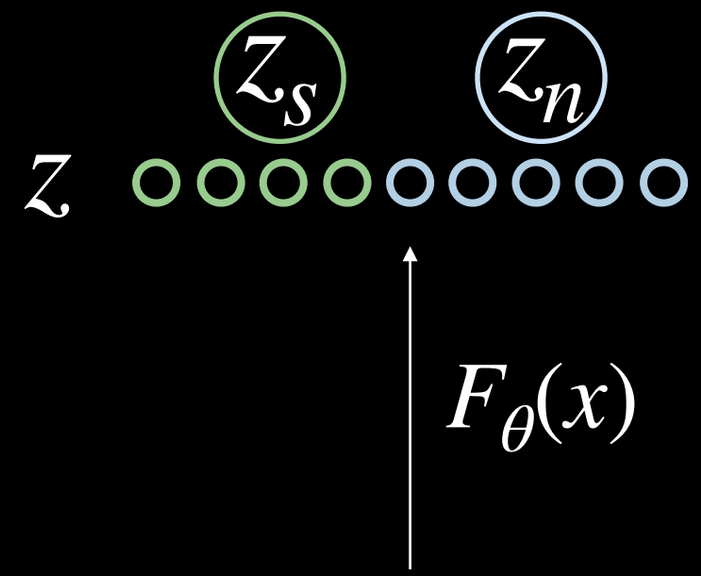
"Density Estimation using Real NVP", Dinh et al., 2017

- Invertible networks with tractable Jacobian determinant, inverse allow maximum-likelihood based density modeling:

$$\log p_x(x) = \log p_z(z) + \log |J_F(x)|$$

- Can also build strong invertible classifiers: loss of information is isolated to final linear layer

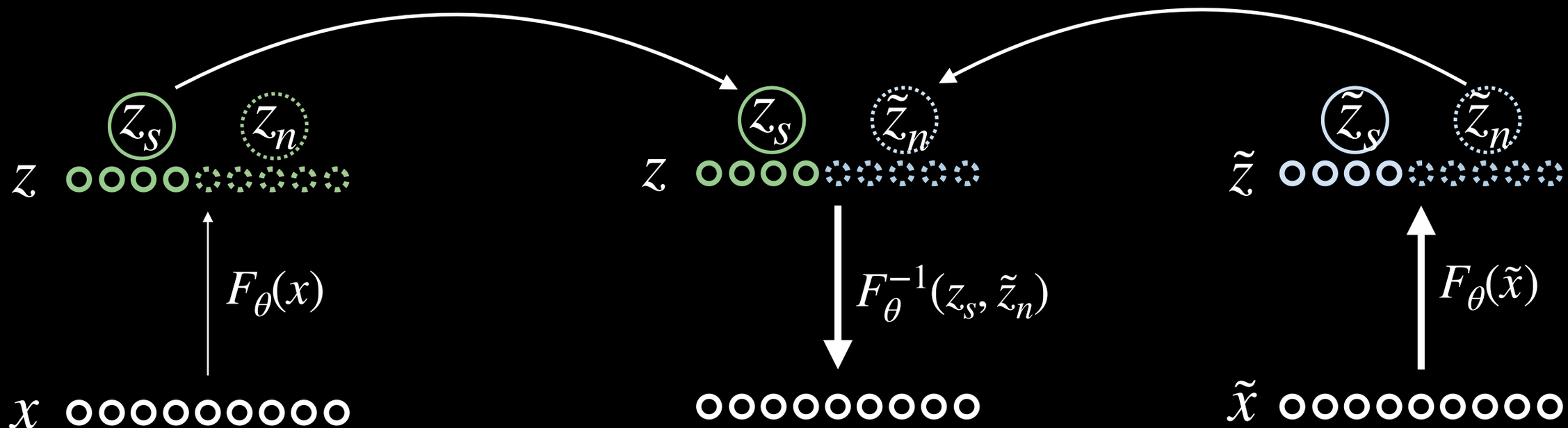
Accessing Decision-space of Invertible Classifiers

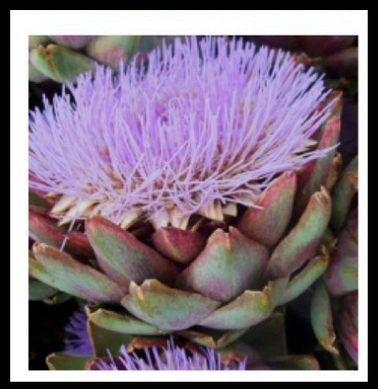


Simplified readout structure: subset of dimensions represents logits of classifier

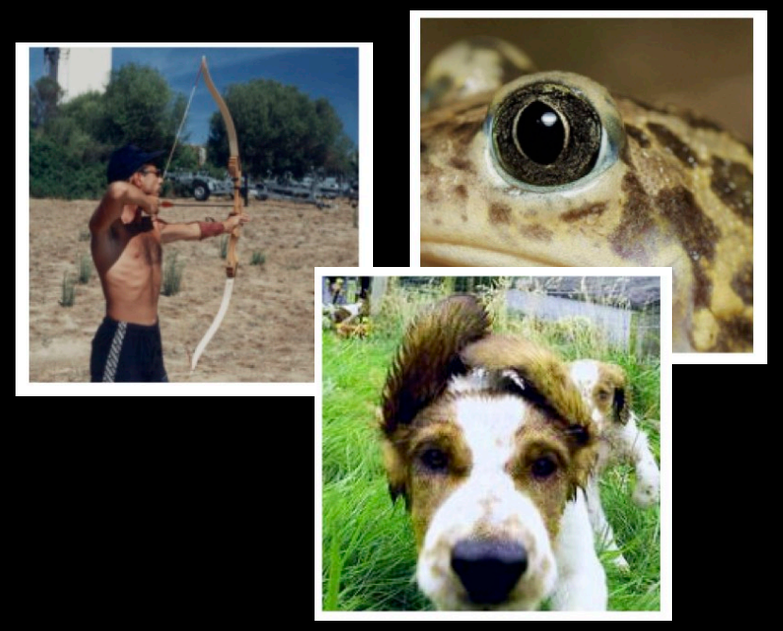
$$p(c_k \mid x) = \frac{\exp(l_k)}{\sum_j \exp(l_j)}$$

Investigating Pre-images





?



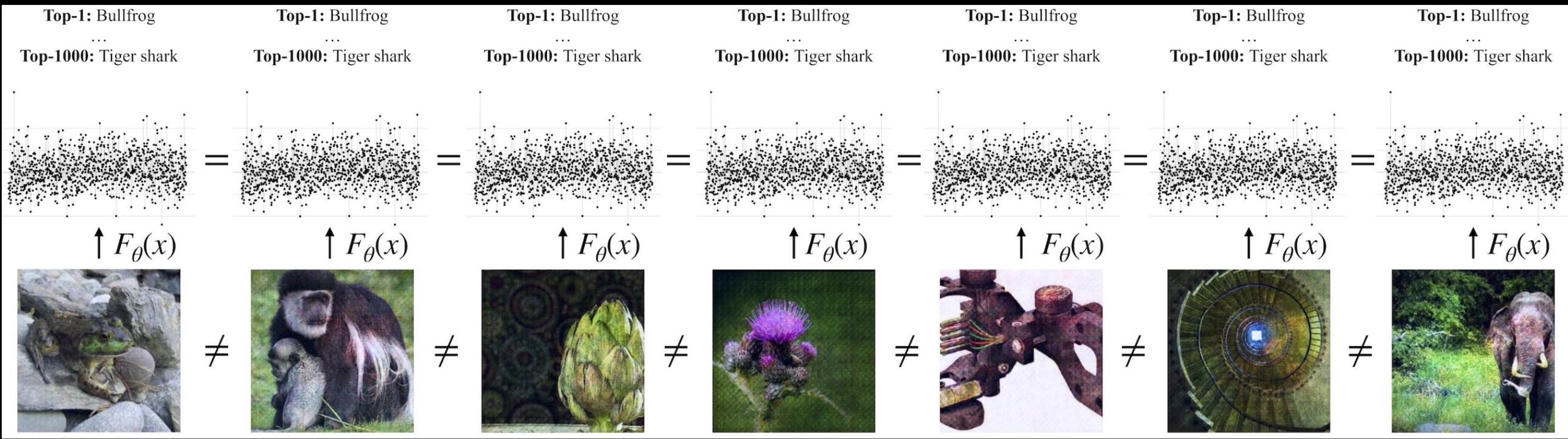
Result is an image which lies in pre-image of logits configuration given x

Excessive Invariance Across Tasks

Instead of learning about discriminative features for task, we have created analytical adversarial attack

Deep classifiers are *too invariant to task-relevant changes* on various problems

Sampling from Logit Pre-image



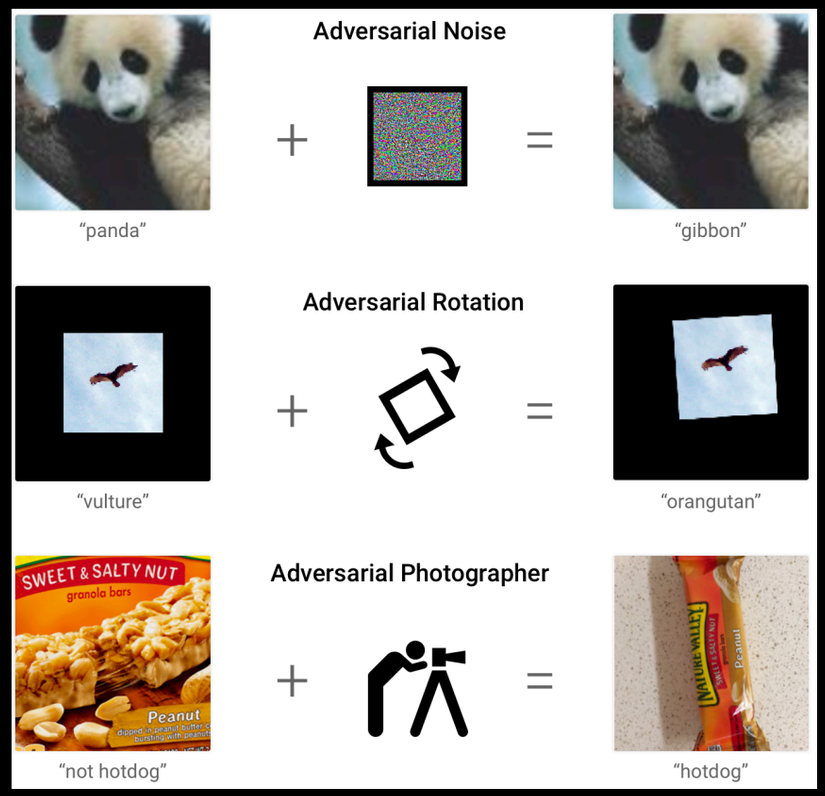
- Discarded part of signal dominates image content!
- Logits only capture a fraction of information relevant to humans

Step Back: Primer on Adversarial Examples

Distribution shift: distribution of test examples differs from training distribution.

- Core idea of adversarial example research: benchmark generalization under distribution shift

- Deep nets exhibit striking, unintuitive failures



Unrestricted Adversarial Examples, Brown et al., 2018

- Poor understanding of failures, few theoretical guarantees

- Invertible networks give new angle to formalize the problem!

Invariance-based Adversarial Examples

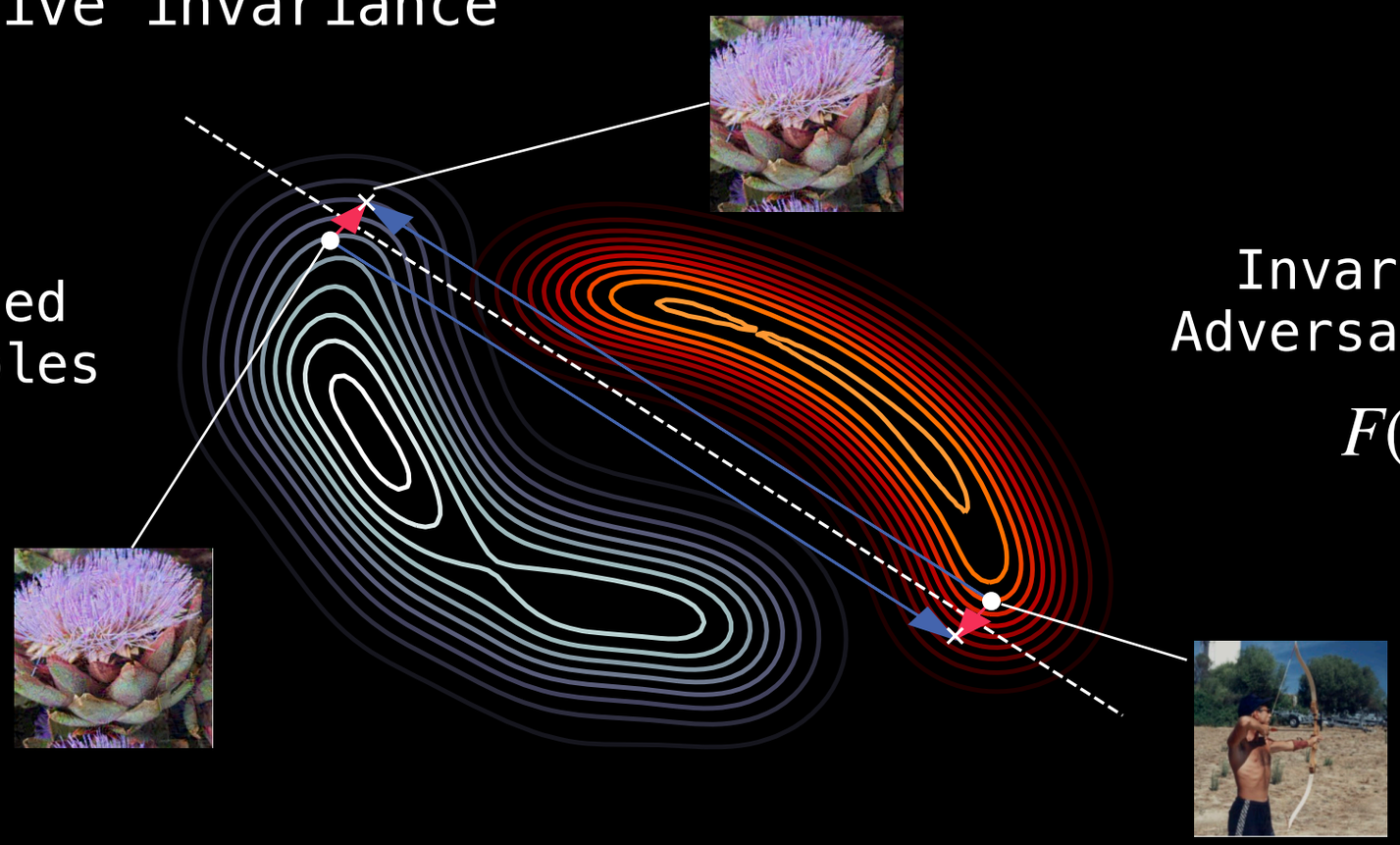
- Norm-bounded adversarial examples benchmark stability

- Invariance-based adversarial examples are complementary to perturbation-based: benchmark excessive invariance

(Classical) Perturbation-based Adversarial Examples

$$F(x) \neq F(x^*)$$

$$y = y^*$$



Invariance-based Adversarial Examples

$$F(x) = F(x^*)$$

$$y \neq y^*$$

Intuitively:

Which task-relevant changes can be applied to the input that do not change the prediction?

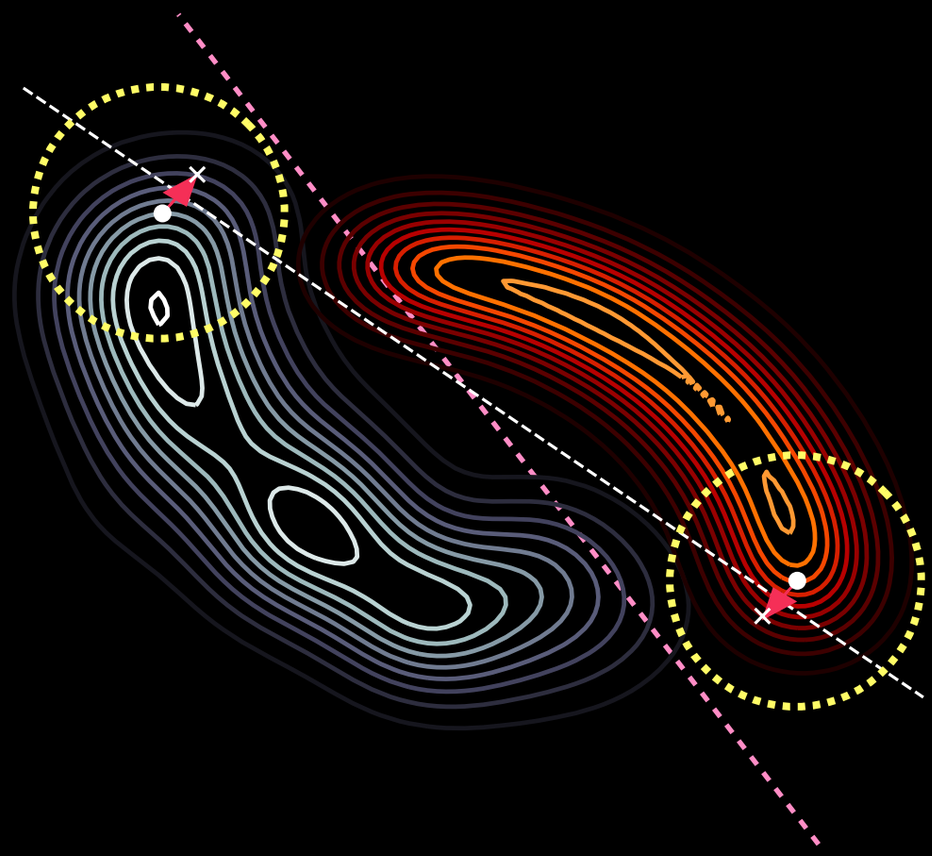
Relationship Non-trivial!

Solving perturbation robustness may increase invariance vulnerability

–> Excessive stability leads to new erroneous invariances

(Classical)
Perturbation-based
Adversarial Examples

$$F(x) \neq F(x^*)$$
$$y = y^*$$



Invariance-based
Adversarial Examples

$$F(x) = F(x^*)$$
$$y \neq y^*$$

Need to control perturbation and invariance robustness alongside accuracy for models to generalize well

But: How to control task-dependent invariance?

Insufficiency of Standard Objective

Information preservation allows us to "collect" invariant part of the signal:

$$\begin{aligned} I(y;x) &= I(y;F_\theta(x)) \\ &= I(y;z_s^\theta, z_n^\theta) \\ &= I(y;z_s^\theta) + I(y;z_n^\theta \,|\, z_s^\theta) \end{aligned}$$

Standard Cross-Entropy classification loss maximizes bound on mutual information:

$$\text{argmax}_\theta \; I(y;z_s^\theta)$$

No incentive to explain all task-relevant variability!

Enforce Information Separation

No incentive to explain all task-relevant variability!

A way out, maximize conditional mutual information:

$$\text{argmax}_\theta \; I(y; z_s^\theta \,|\, z_n^\theta)$$

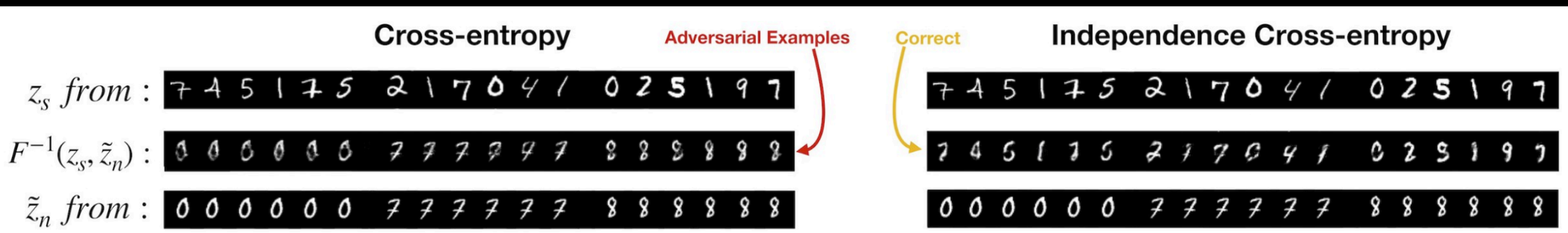
In practice, extend vanilla cross-entropy objective:

$$\min_\theta \mathscr{L}_{iCE}(\theta, \theta_{nc}) = \underbrace{\sum_{i=1}^{C} - y_i \log \tilde{F}_\theta^{z_s}(x)_i}_{=: \mathscr{L}_{sCE}(\theta)}$$

With independence term:

$$\min_\theta \max_{\theta_{nc}} \mathscr{L}_{iCE}(\theta, \theta_{nc}) = \underbrace{\sum_{i=1}^{C} - y_i \log \tilde{F}_\theta^{z_s}(x)_i}_{=: \mathscr{L}_{sCE}(\theta)} + \underbrace{\sum_{i=1}^{C} y_i \log D_{\theta_{nc}}(F_\theta^{z_n}(x))_i}_{=: \mathscr{L}_{nCE}(\theta, \theta_{nc})}$$

Independence Cross-entropy in Practice

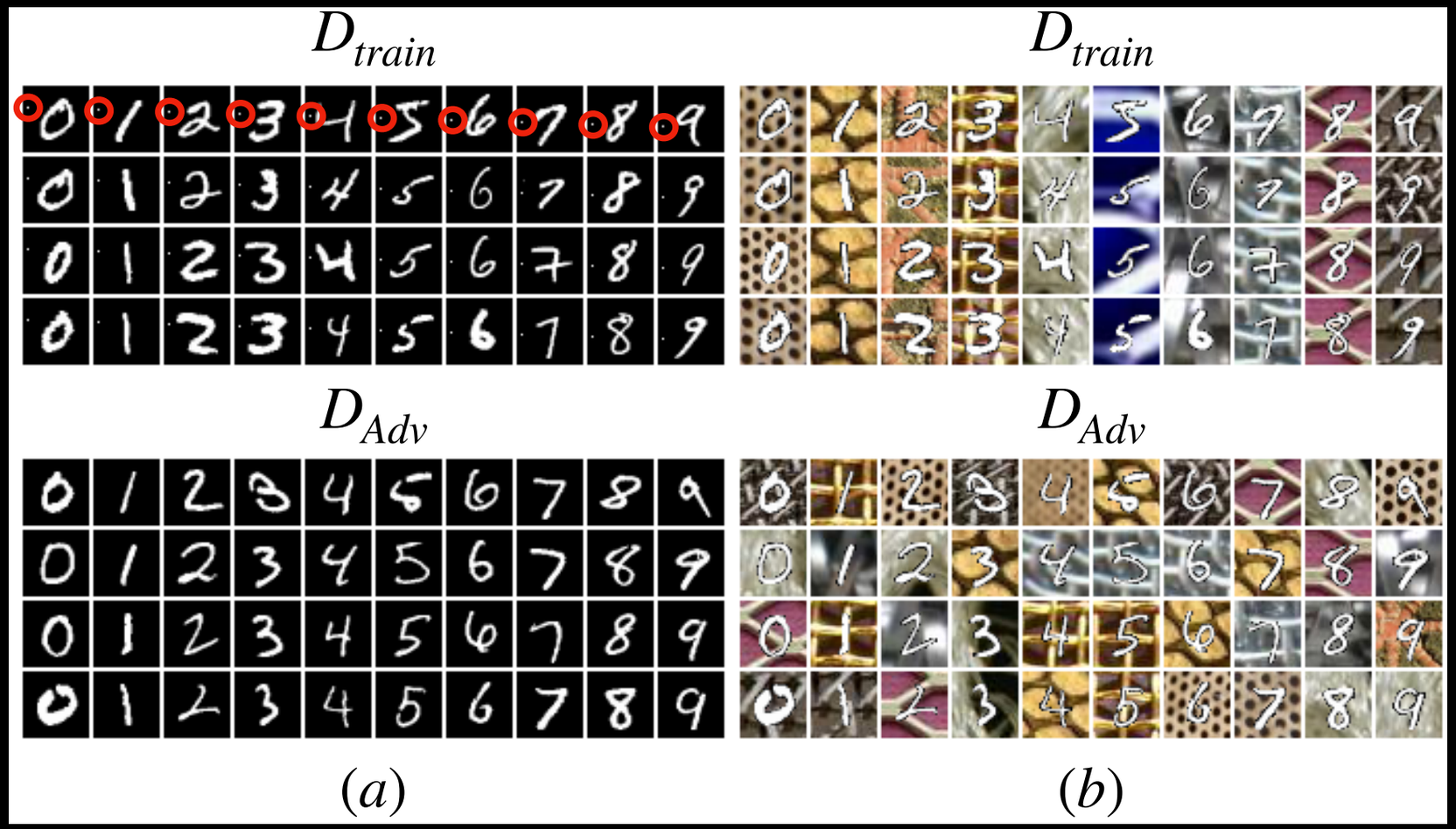


z_n governs nuisances, z_s the true task semantics

Transformations in z_n space now cannot change class identity

Not possible to create invariance-based adversarial examples

Breaking Classifier, one Pixel/Regularity at a Time



Binary

% Error	Dtrain	DAdv
CE ResNet	00.00	73.80
CE fiRevNet	00.00	57.09
iCE fiRevNet	00.02	34.73
Difference	**00.02**	**38.33**

Textured

% Error	Dtrain	DAdv
CE ResNet	00.00	87.83
CE fiRevNet	00.18	73.71
iCE fiRevNet	00.53	59.99
Difference	**00.53**	**27.84**

Conclusions

- Can learn manipulable subspace –> range of novel examples

- Methods enable control over information contained in learned representations

- Invertible models allow insight into learned classifiers:
  - —invariance needs to be controlled
  - —learn models that generalize well

- Underlying theme: information separation in representations

Current Directions

- Application, extension of techniques to holy grail of ML: semi-supervised learning (few labeled examples)

- Transfer learning: how do controlled representations enable transfer / learning of new tasks?

- White-box models: incorporate known variables, relations

- Uncertainty representations: can model know what it knows and does not know?

Thank You & Collaborators

Jake Snell
Jack Klys

Elliot Creager
Kevin Swersky
David Madras
Toni Pitassi

Jorn Jacobsen
Jens Behrmann
Matthias Bethge