Conjugate Energy-Based Models

Hao
Wu* [1]

Babak
Esmaeili* [1]

Michael
Wick [2]

Jean-Baptiste
Tristan [3]

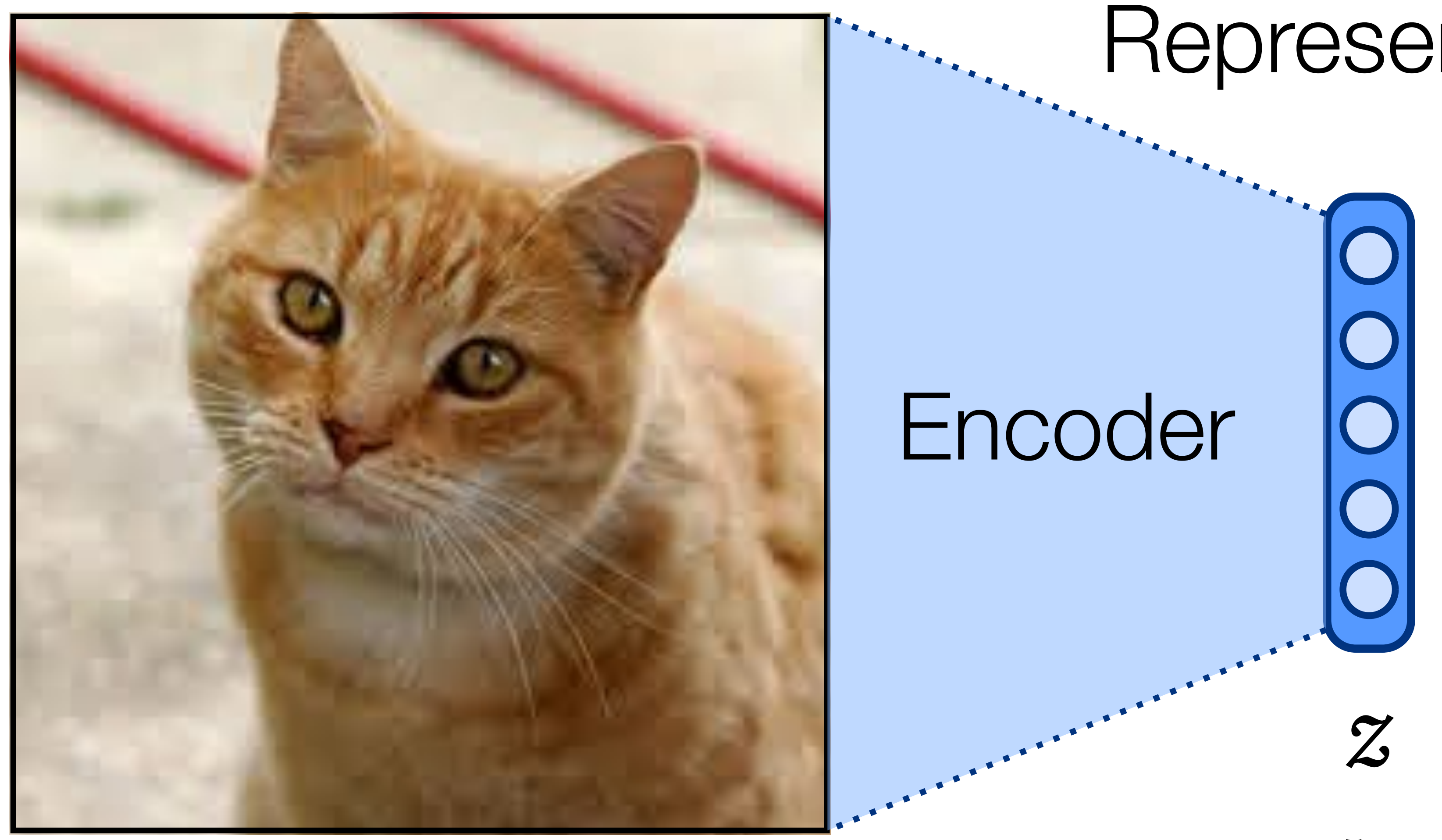
Jan-Willem
van de Meent [1]

[1] Northeastern University, [2] Oracle Labs, [3] Boston College

Learning Representation with Energy-Based Models

Data

Latent Representation



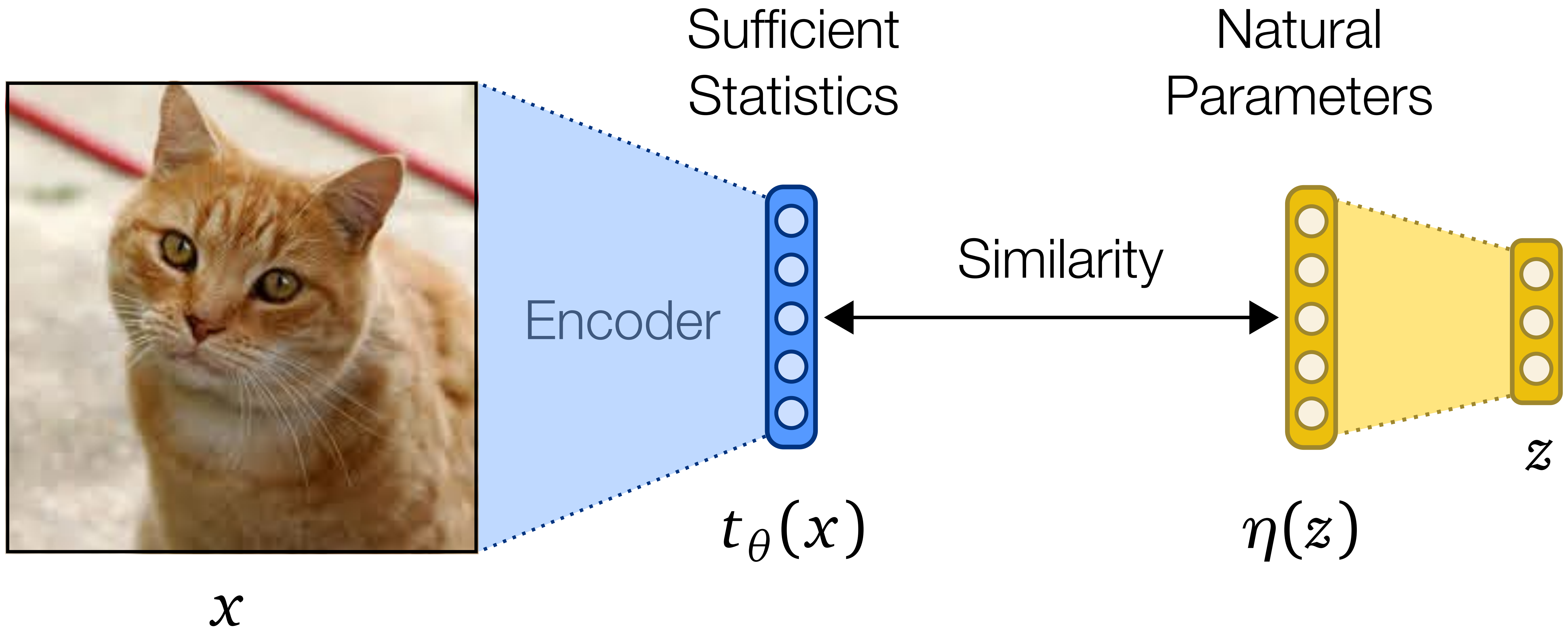
x

$$p_\theta(x,z) \propto \exp(-E_\theta(x,z))$$

$$z \sim p_\theta(z \mid x)$$

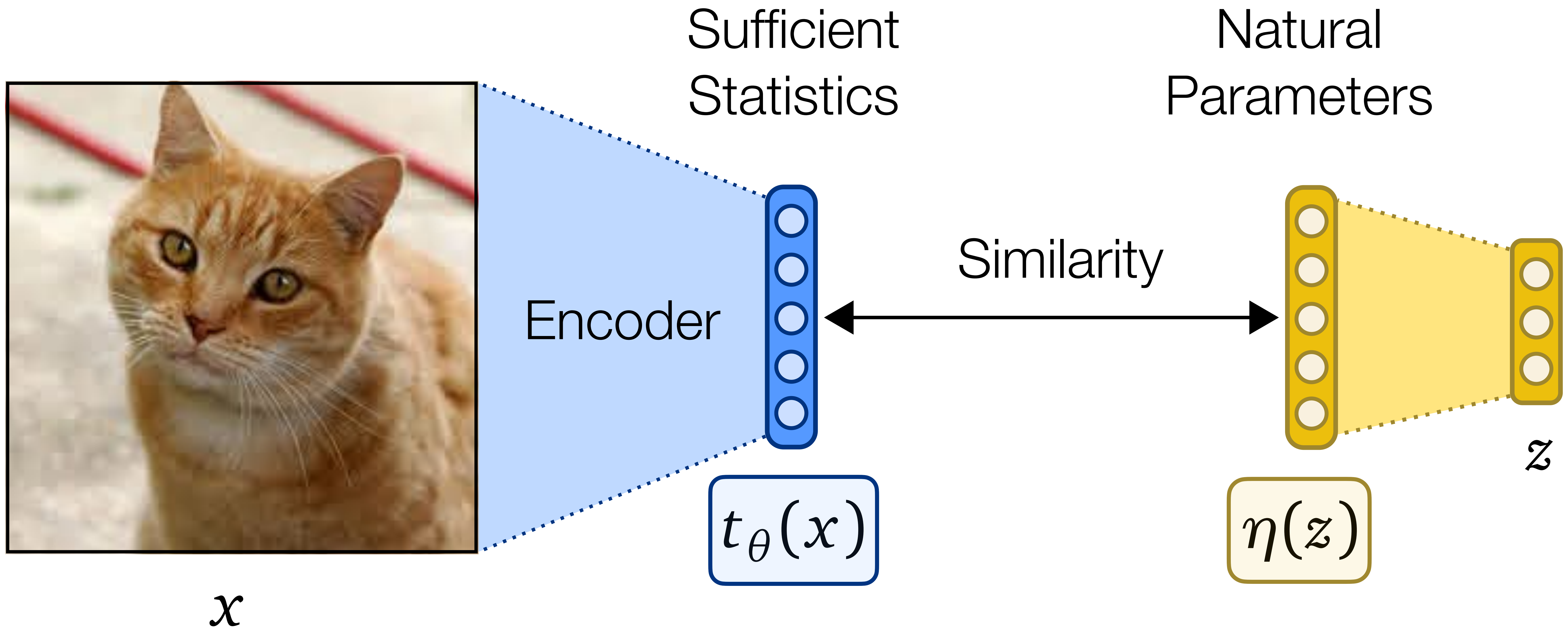
Downstream Tasks: Classification, Out-of-Domain Detection, …

Conjugate Energy-Based Models



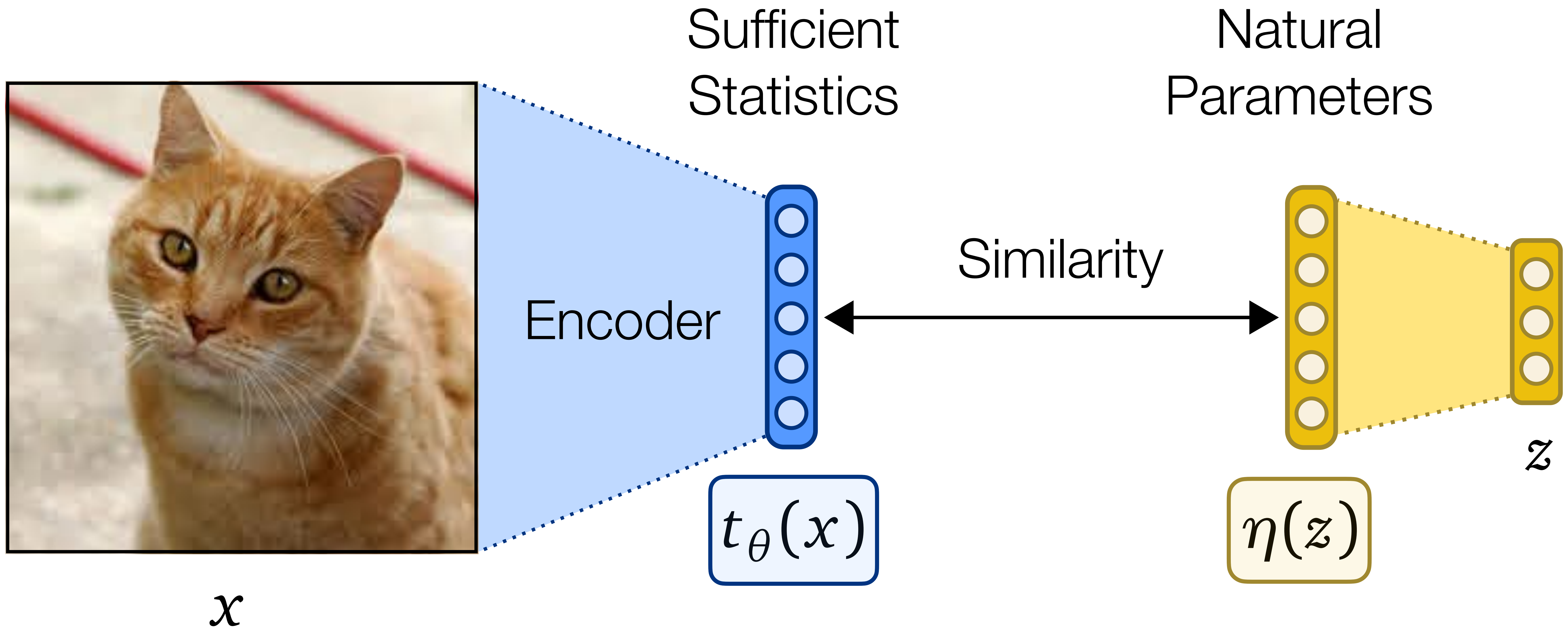
$$-E_\theta(x,z) = \underbrace{\langle\, t_\theta(x)\,,\ \eta(z)\,\rangle}_{\substack{\text{Similarity} \\ \text{(inner product)}}} + \underbrace{\langle\, \lambda\,,\ \eta(z)\,\rangle - A(\lambda)}_{\substack{\text{Bias: log p(z | λ)} \\ \text{(exponential family)}}}$$

Conjugate Energy-Based Models



$$
\begin{aligned}
-E_\theta(x, z) &= \langle t_\theta(x), \eta(z) \rangle + \langle \lambda, \eta(z) \rangle - A(\lambda) \\
&= \langle t_\theta(x) + \lambda, \eta(z) \rangle - A(\lambda) \\
&\quad + A(t_\theta(x) + \lambda) - A(t_\theta(x) + \lambda)
\end{aligned}
$$

Conjugate Energy-Based Models

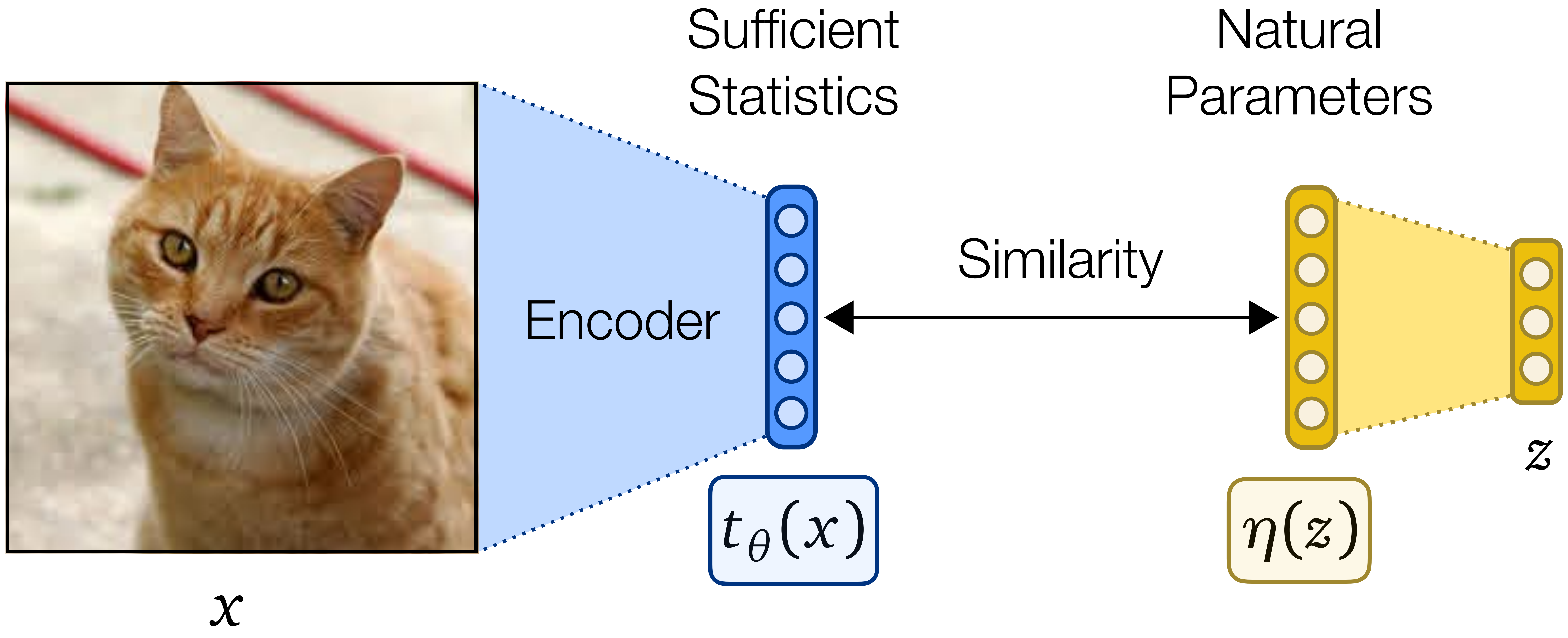


$$-E_\theta(x,z) = \langle t_\theta(x), \eta(z) \rangle + \langle \lambda, \eta(z) \rangle - A(\lambda)$$

$$= \langle t_\theta(x) + \lambda, \eta(z) \rangle - A(\lambda)$$

$$+ A(t_\theta(x)+\lambda) - A(t_\theta(x)+\lambda)$$

Conjugate Energy-Based Models

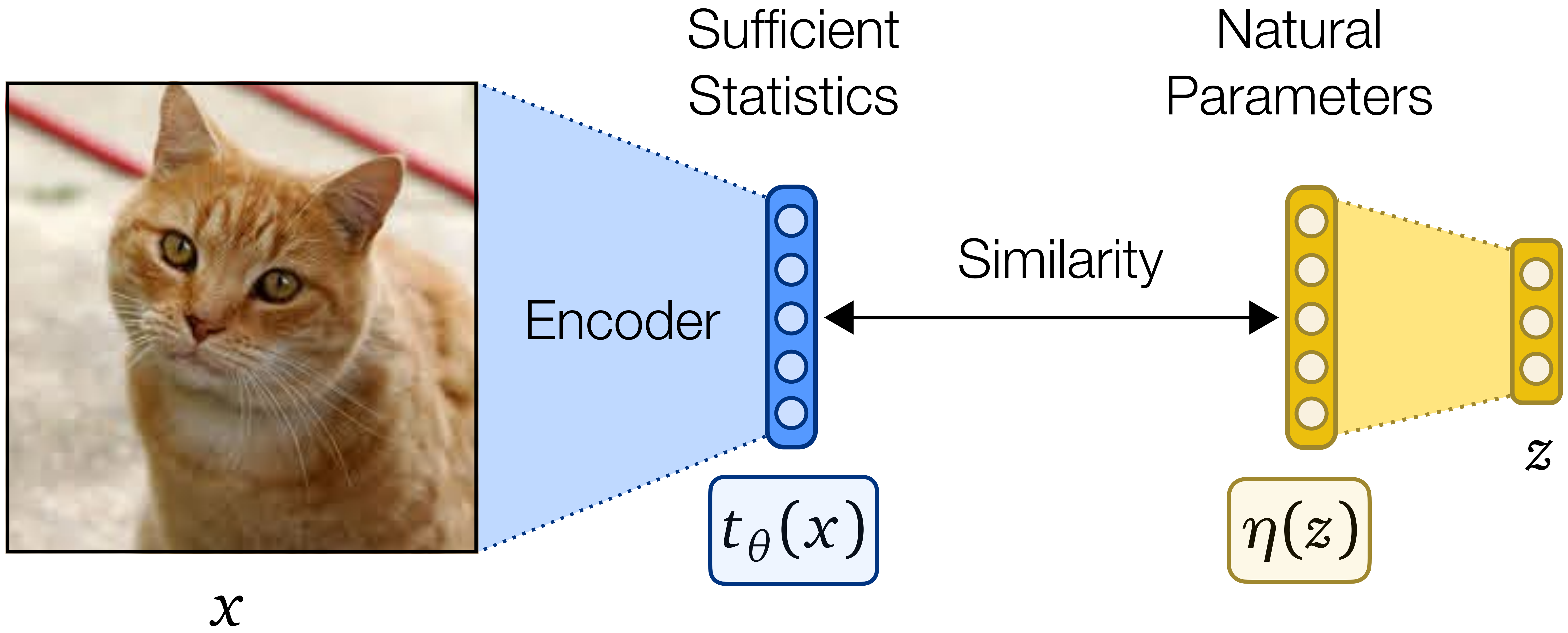


$$-E_\theta(x,z) = \langle \boxed{t_\theta(x)}, \boxed{\eta(z)} \rangle + \langle \boxed{\lambda}, \boxed{\eta(z)} \rangle - \boxed{A(\lambda)}$$

$$= \langle \boxed{t_\theta(x)} + \boxed{\lambda}, \boxed{\eta(z)} \rangle - \boxed{A(t_\theta(x)+\lambda)}$$

$$+ \boxed{A(t_\theta(x)+\lambda)} - \boxed{A(\lambda)}$$

Conjugate Energy-Based Models



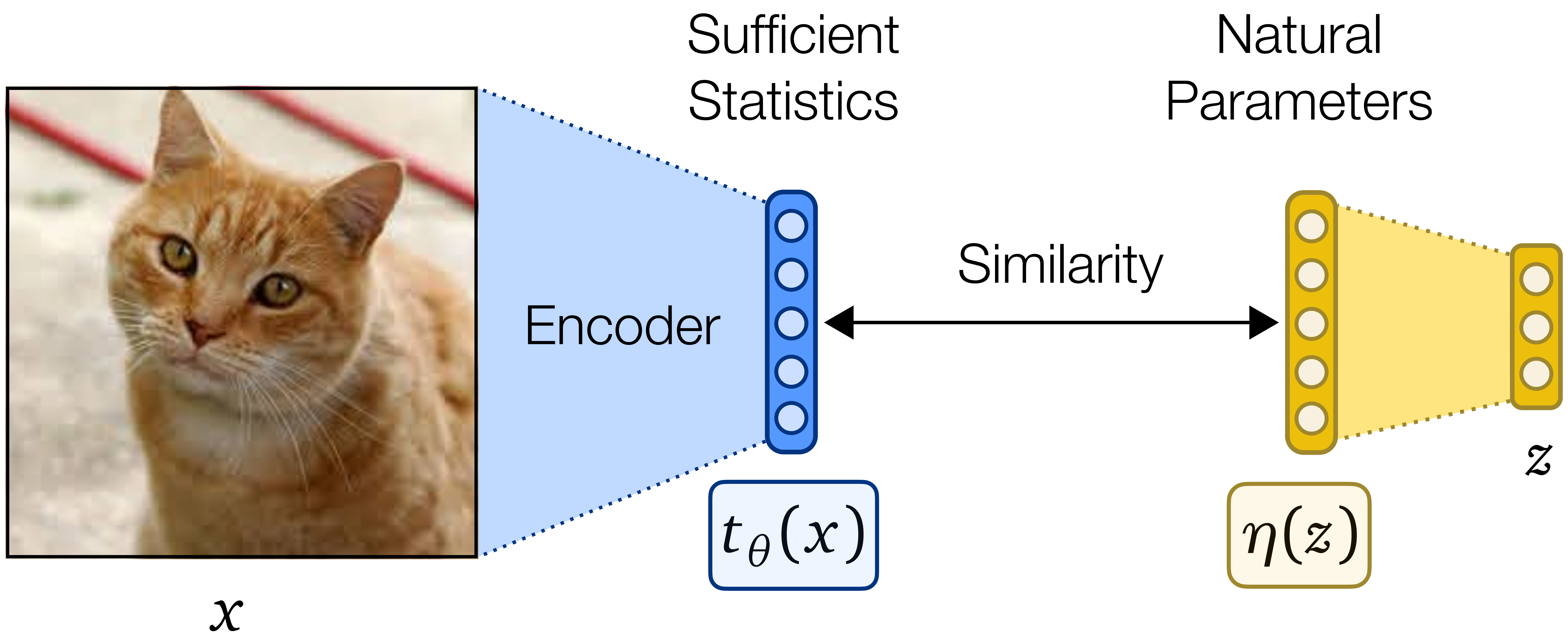
$$-E_\theta(x,z) = \langle \boxed{t_\theta(x)}, \boxed{\eta(z)} \rangle + \langle \boxed{\lambda}, \boxed{\eta(z)} \rangle - \boxed{A(\lambda)}$$

$$= \langle \boxed{t_\theta(x)} + \boxed{\lambda}, \boxed{\eta(z)} \rangle - \boxed{A(t_\theta(x)+\lambda)}$$

Conjugate Posterior
log p(z | $t_\theta(x)+\lambda$)

$$+ \boxed{A(t_\theta(x)+\lambda)} - \boxed{A(\lambda)}$$

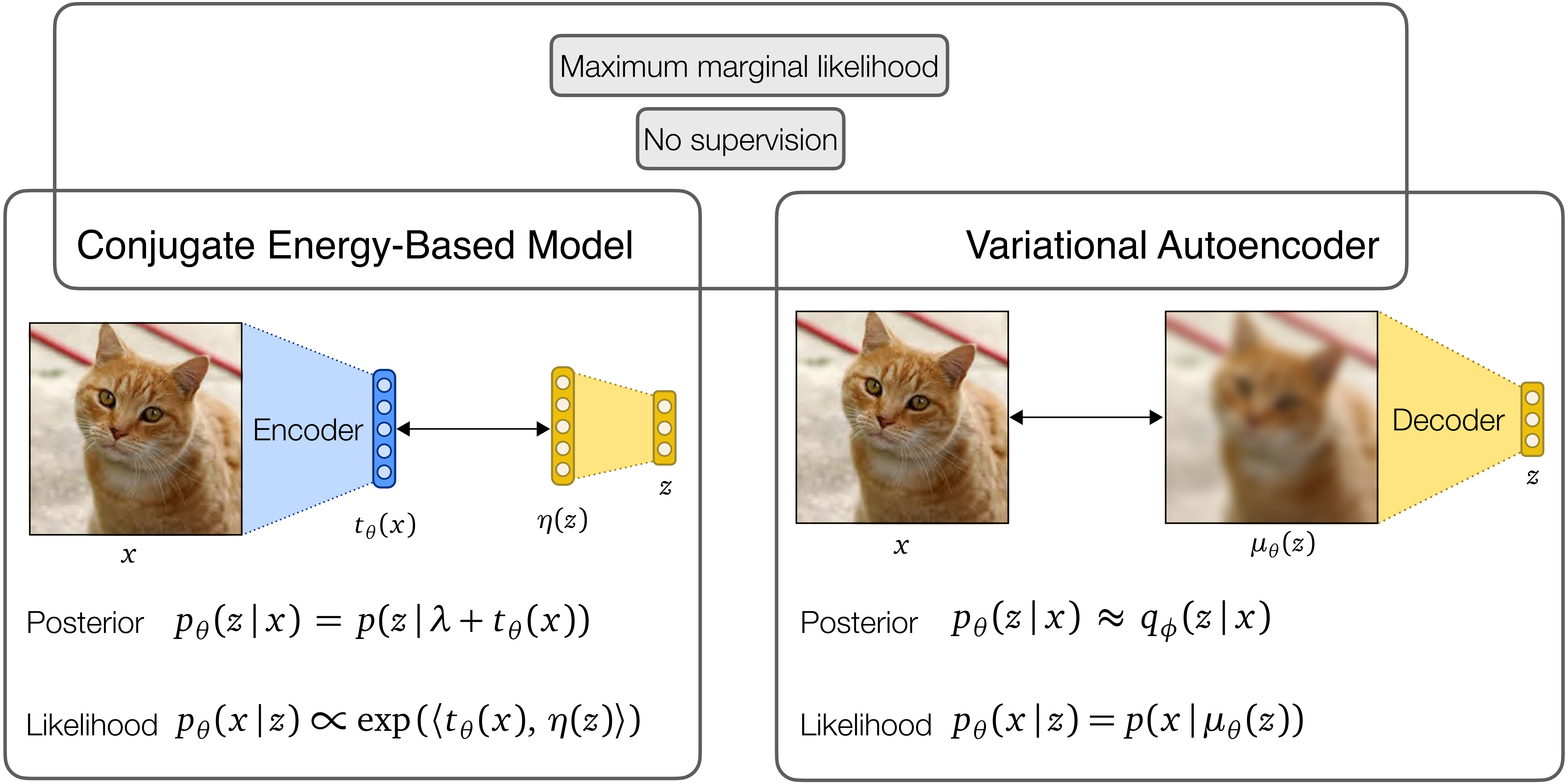
Conjugate Energy-Based Models



$$-E_\theta(x,z) = \langle t_\theta(x), \eta(z) \rangle + \langle \lambda, \eta(z) \rangle - A(\lambda)$$

$$= \langle t_\theta(x) + \lambda, \eta(z) \rangle - A(t_\theta(x) + \lambda)$$

$$+ A(t_\theta(x) + \lambda) - A(\lambda)$$ EBM (can train with standard methods)



Maximum marginal likelihood
No supervision
Conjugate Energy-Based Model
Variational Autoencoder
Encoder
x
$t_\theta(x)$
$\eta(z)$
z
Posterior $p_\theta(z \mid x) = p(z \mid \lambda + t_\theta(x))$
Likelihood $p_\theta(x \mid z) \propto \exp(\langle t_\theta(x), \eta(z) \rangle)$
Decoder
x
$\mu_\theta(z)$
z
Posterior $p_\theta(z \mid x) \approx q_\phi(z \mid x)$
Likelihood $p_\theta(x \mid z) = p(x \mid \mu_\theta(z))$

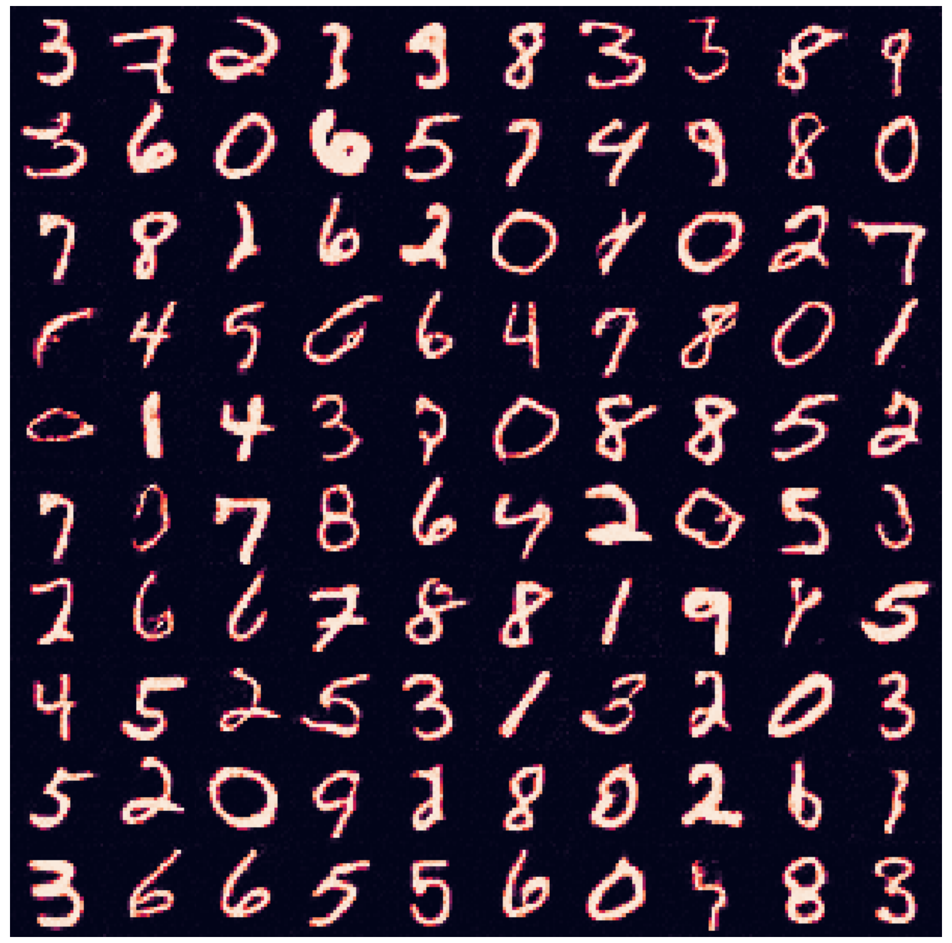
Experiments

CEBM: Gaussian inductive bias

GMM-CEBM: GMM inductive bias

Can CEBMs approximate the data distribution?

MNIST



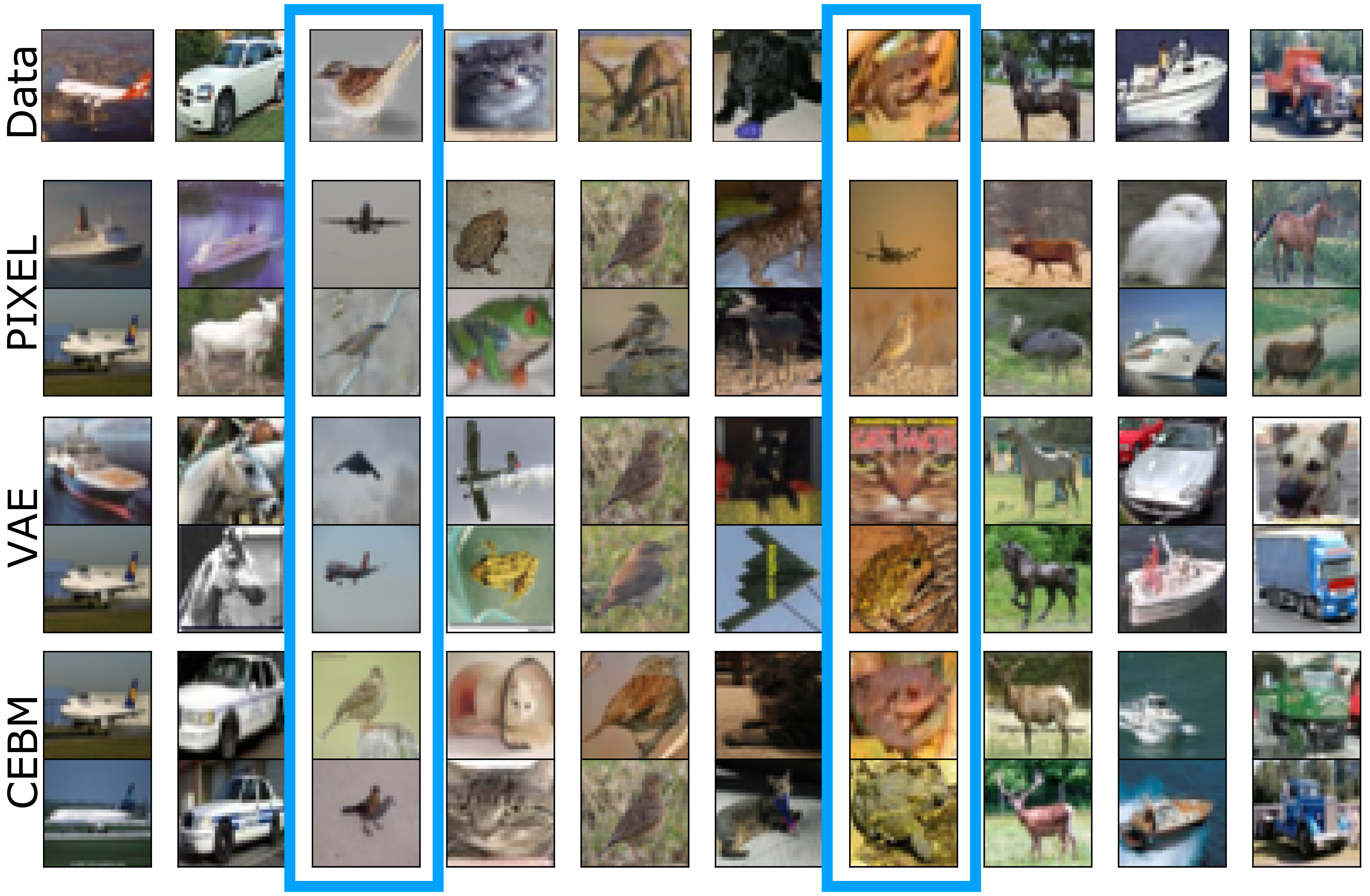
F-MNIST

CIFAR-10

SVHN

Alignment Between Learned Presentations and Class Labels

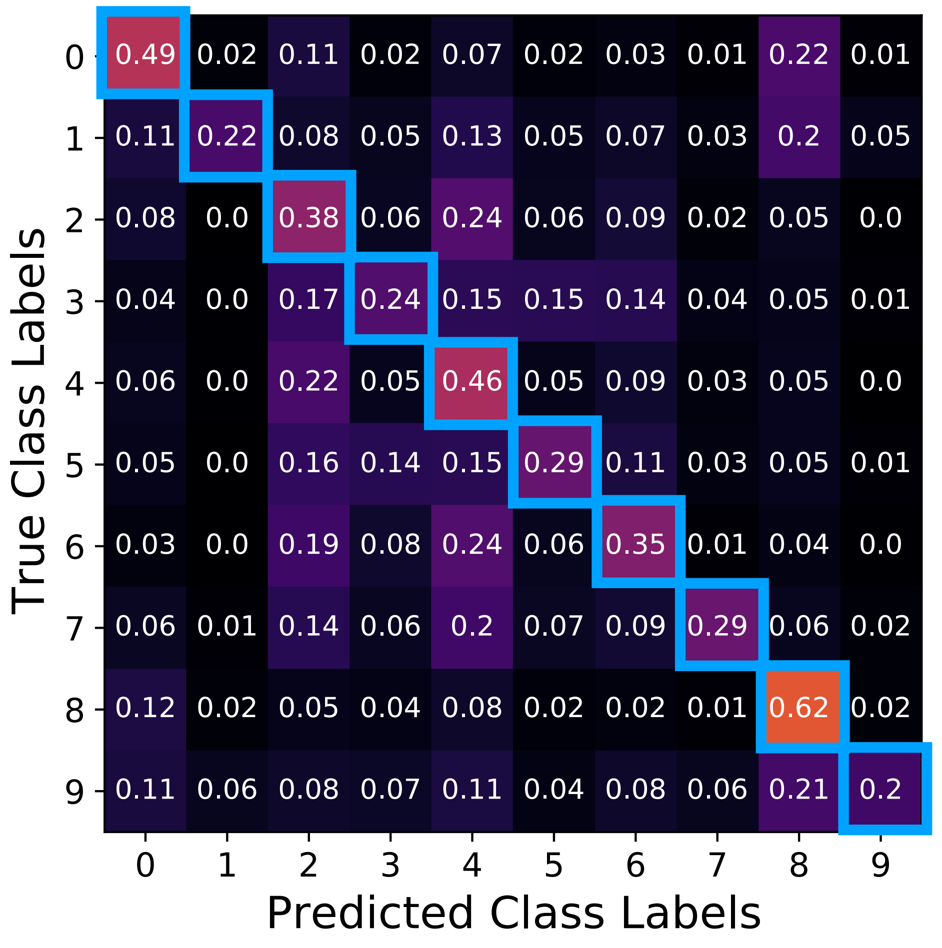
Nearest Neighbors



Confusion Matrices

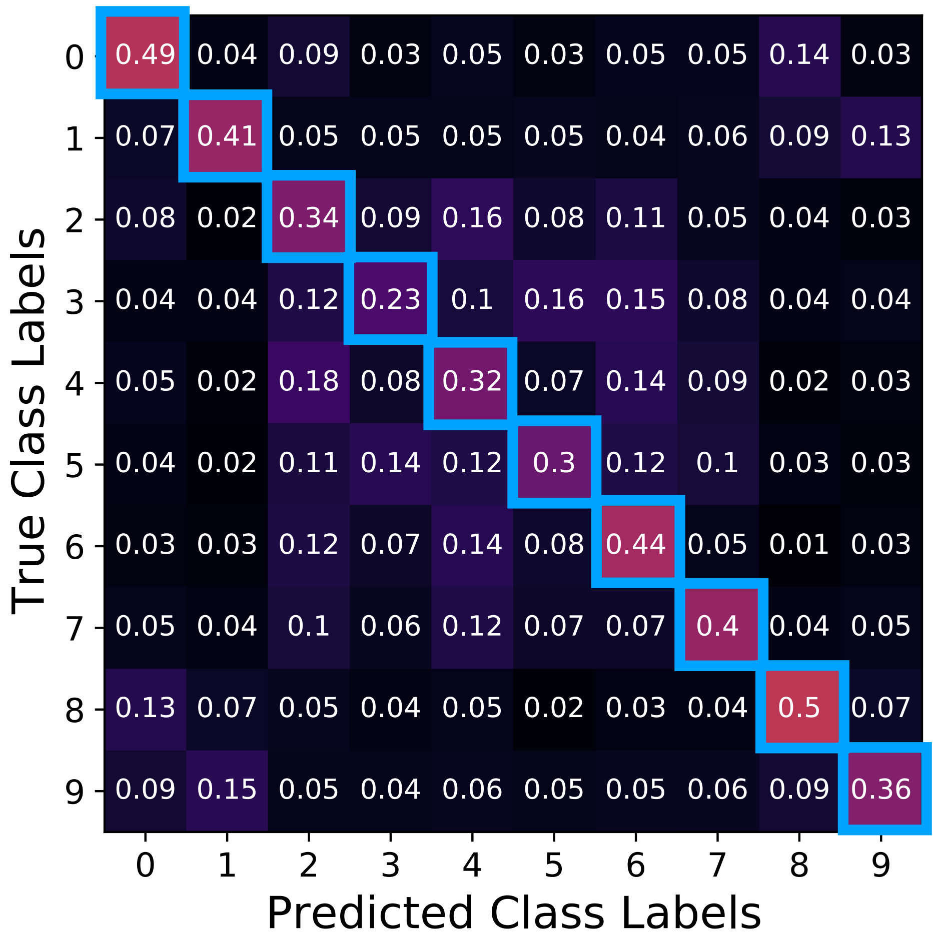
PIXEL

0.35



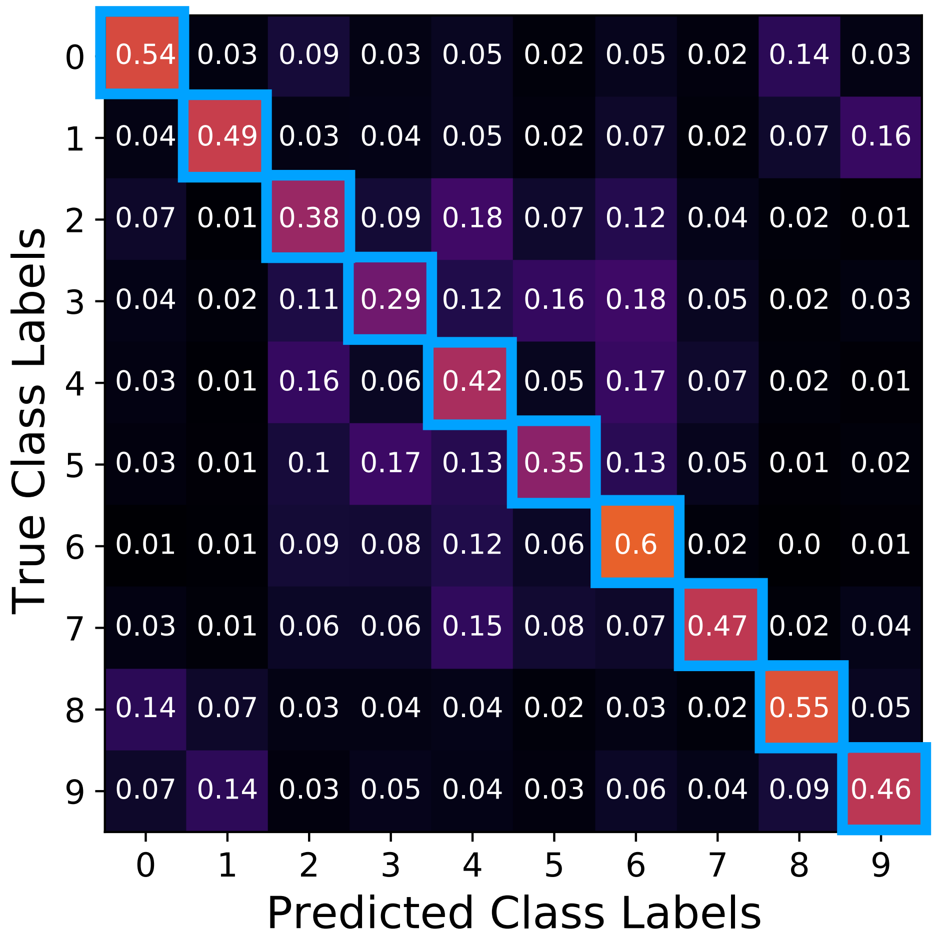
VAE

0.38



CEBM

0.45



Downstream Task: Few-label Classification

Models	MNIST 1	MNIST 10	MNIST 100	MNIST $full$	Fashion-MNIST 1	Fashion-MNIST 10	Fashion-MNIST 100	Fashion-MNIST $full$	CIFAR-10 1	CIFAR-10 10	CIFAR-10 100	CIFAR-10 $full$	SVHN 1	SVHN 10	SVHN 100	SVHN $full$
VAE	42	85	92	95	41	63	72	81	16	22	31	38	**13**	13	16	36
GMM-VAE	53	86	93	97	49	68	79	84	**19**	23	33	39	**13**	14	23	56
BIGAN	33	67	85	91	46	65	75	81	18	**30**	**43**	52	11	20	42	56
IGEBM	63	89	95	97	50	**70**	79	83	16	26	33	42	10	16	35	49
CEBM	**67**	89	95	97	**52**	**70**	77	83	**19**	**30**	42	**53**	12	**25**	**48**	**70**
GMM-CEBM	**67**	**91**	**97**	**98**	**52**	**70**	**80**	**85**	16	29	42	52	10	17	39	60

Thank you!