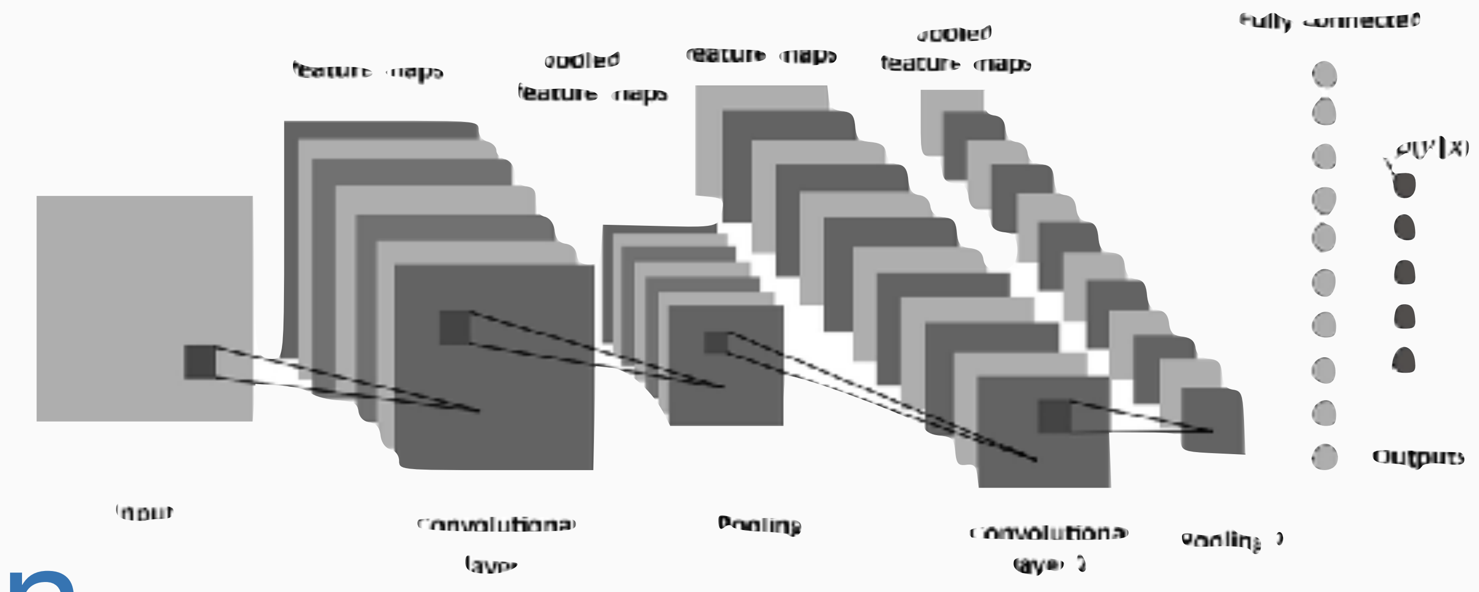
CS 103: Representation Learning, Information Theory and Control

Lecture 4, Feb 1, 2019

Seen last time

1. What is a nuisance for a task?
2. Invariance, equivariance, canonization
3. A linear transformation is group equivariant if and only if it is a group convolution
   - Building equivariant representations for translations, sets and graphs
4. Image canonization with equivariant reference frame detector
   - Applications to multi-object detection
5. Accurate reference frame detection: the SIFT descriptor
   - A sufficient statistic for visual inertial systems

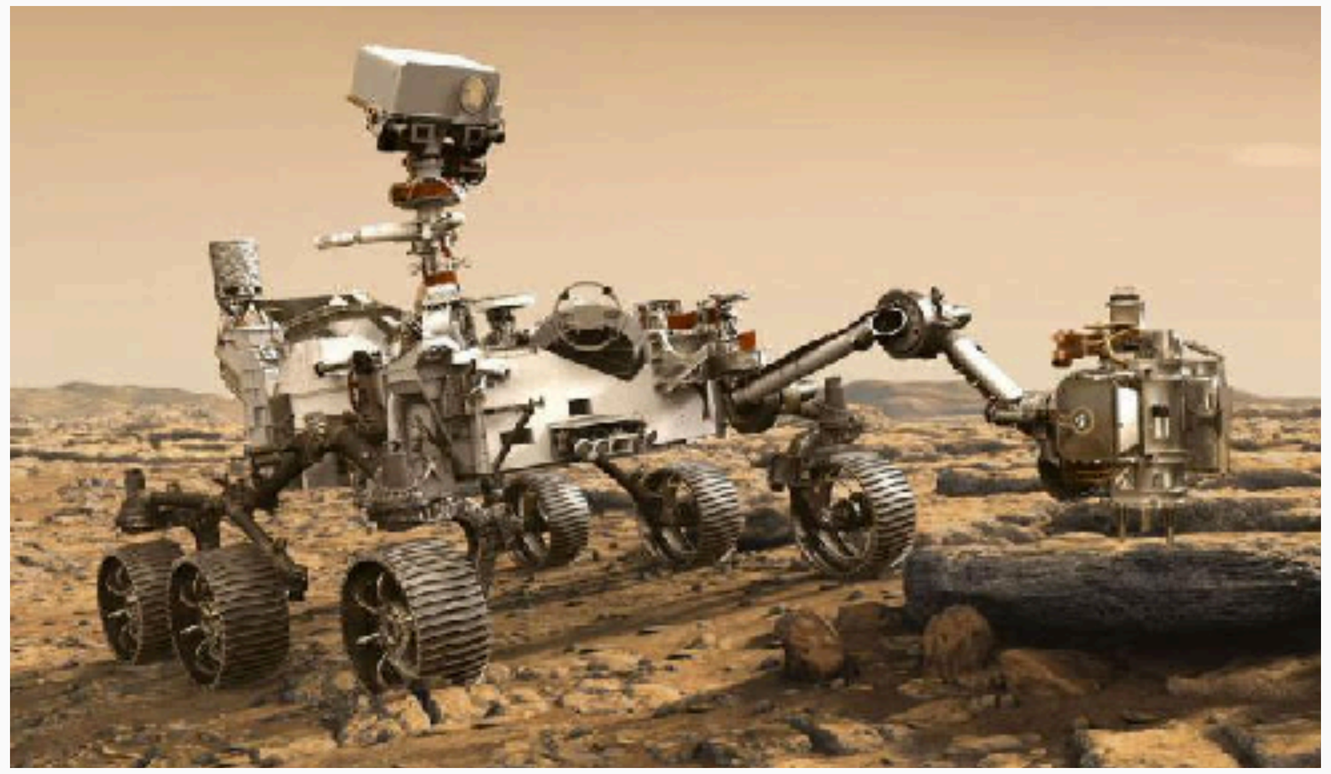
Where are we now



Cognition

Sensing

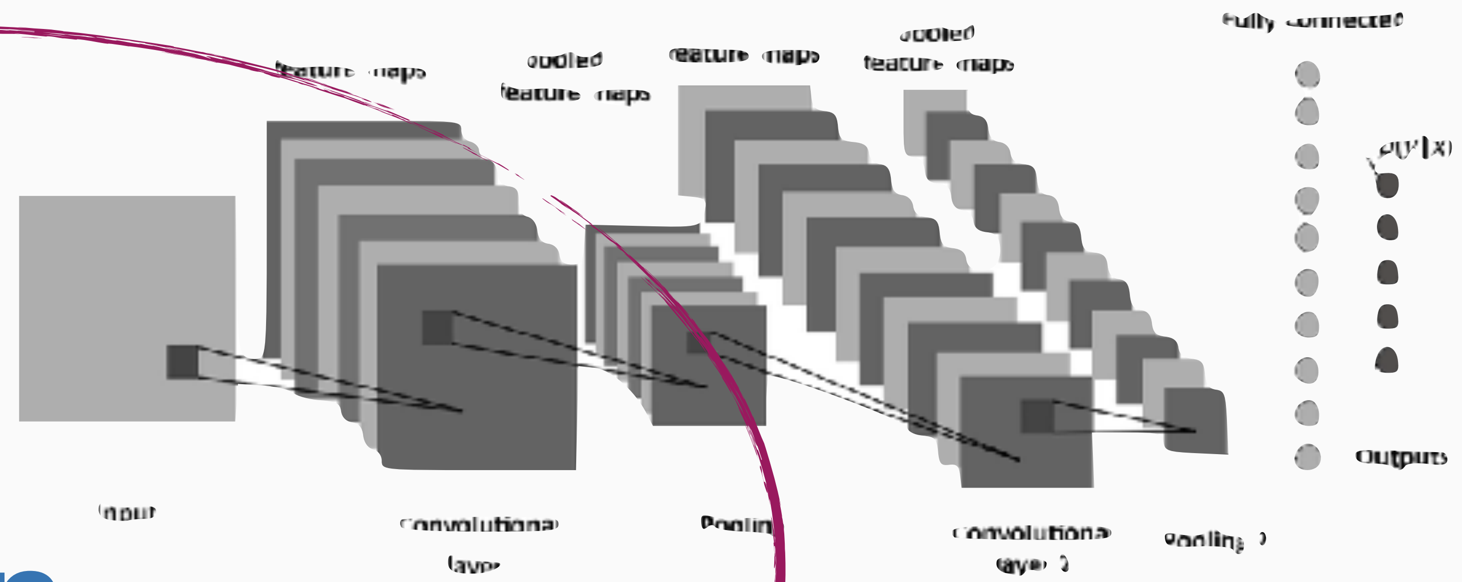
Action



Invariance to simple geometric nuisances, corner detectors, …

Where are we now

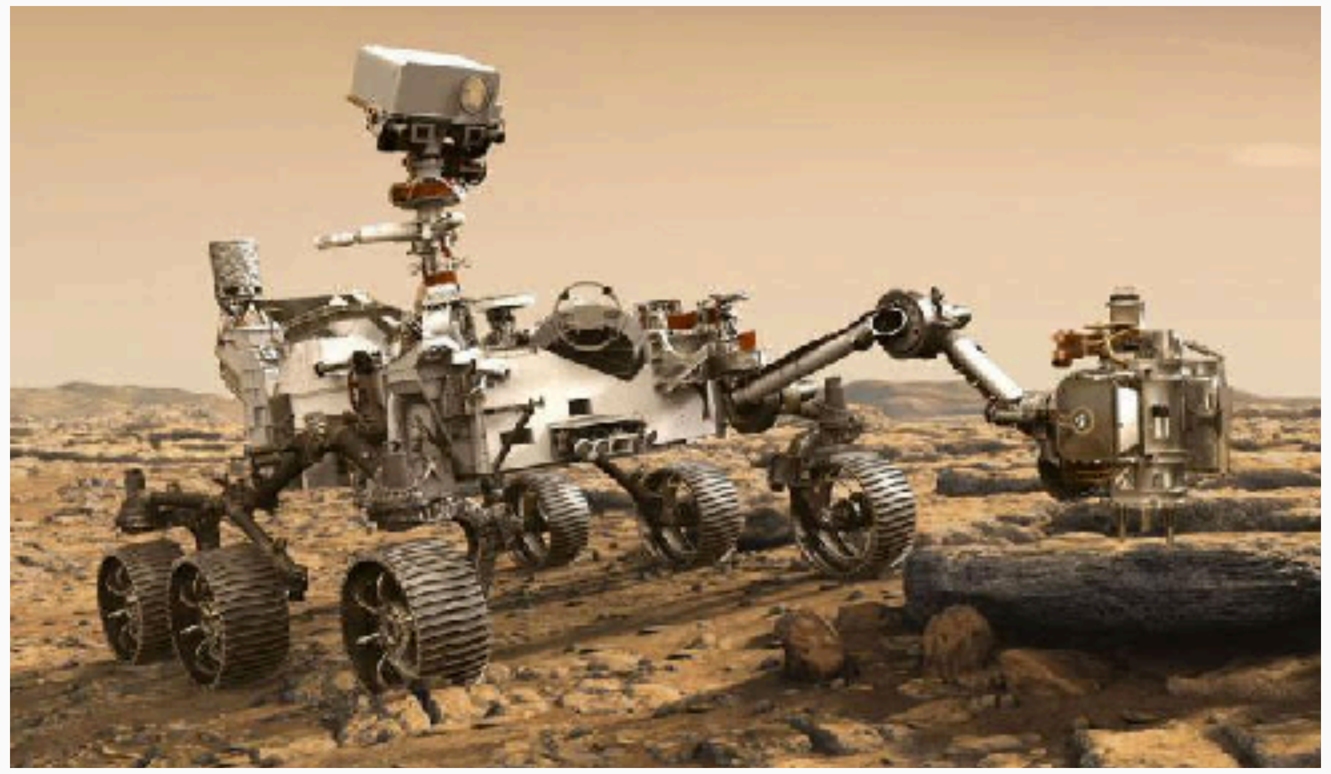
Invariance to complex nuisances,
classification, detection, …



Cognition

Sensing

Action



Compression without loss of *useful* information

Task Y = Is this the picture of a dog?

Original X

X ~ 350KB

Compressed Z

Z ~ 5KB

Z is as useful as X to answer the question Y, but it is much smaller.

Image source https://en.wikipedia.org/wiki/File:Terrier_mixed-breed_dog.jpg

Compression without loss of *useful* information

Task Y = Is this the picture of a dog?

Z is as useful as X to answer the question Y, but it is much smaller.

Image source https://en.wikipedia.org/wiki/File:Terrier_mixed-breed_dog.jpg

The “classic” Information Bottleneck

Some notation

Cross-entropy: The standard loss function in machine learning

$$H_{q,p}(x) = \mathbb{E}_{x\sim q(x)}[-\log p(x)]$$

Kullback-Leibler divergence: “Distance” between two distribution (used in variational inference)

$$\begin{aligned}\mathrm{KL}(q(z)\|p(z)) &= \mathbb{E}_{z\sim q(z)}\left[\log \frac{q(z)}{p(z)}\right] \\ &= H_{q,p}(x) - H_q(x)\end{aligned}$$

Mutual Information: Expected divergence between the posterior *p(z|x)* and the prior p(z).

$$\begin{aligned}I(x;z) &= \mathbb{E}_{x\sim p(x)}[\mathrm{KL}(p(z|x)\|p(z))] \\ &= H_p(z) - H_p(z|x)\end{aligned}$$

The Information Bottleneck Lagrangian

Tishby et al., 1999

Given data x and a task y, find a representation z that is **useful** and **compressed**.

$$\begin{aligned} \text{minimize}_{p(z|x)} \quad & I(x;z) \\ \text{s.t.} \quad & H(y|z) = H(y|x) \end{aligned}$$

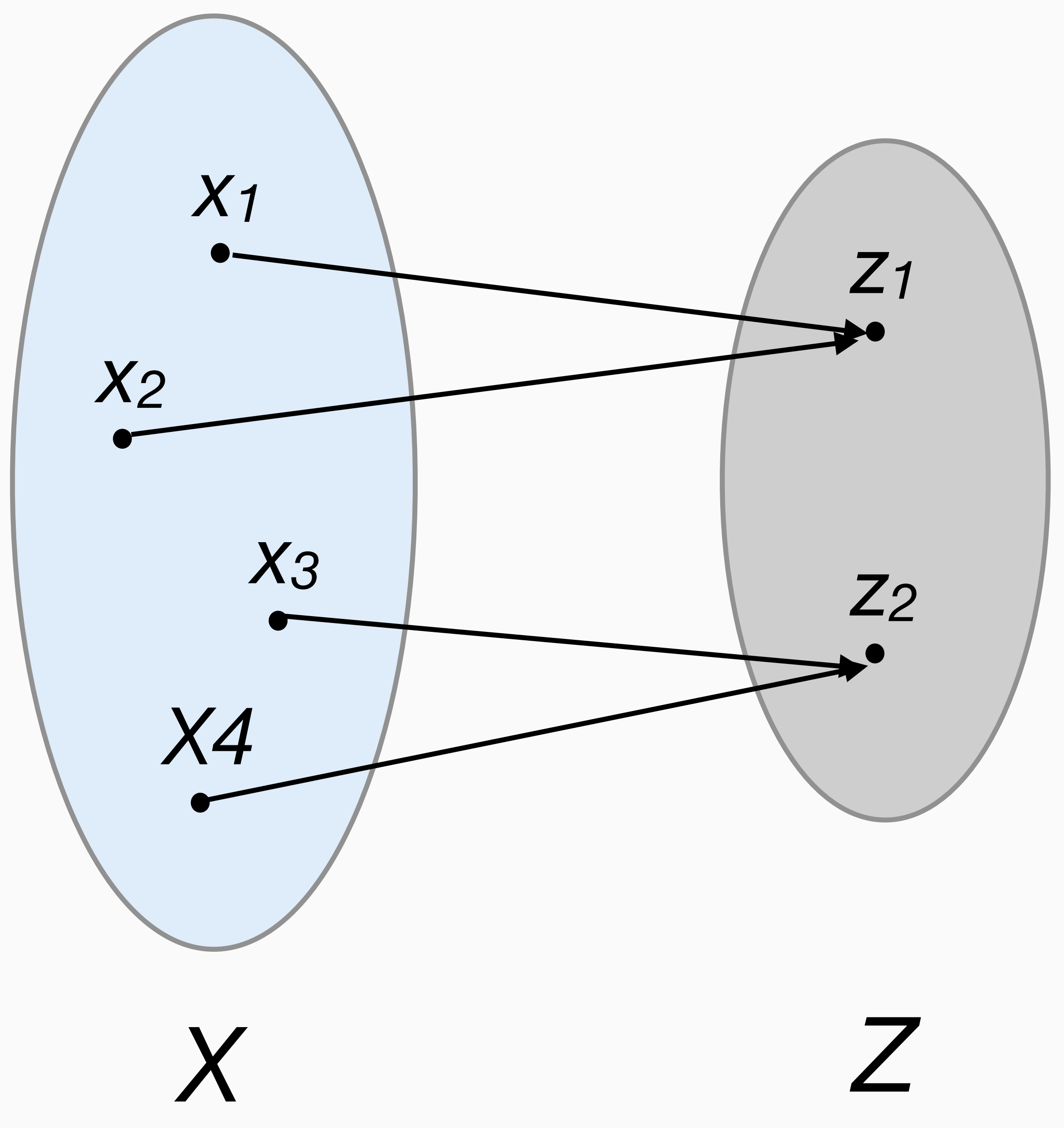
Consider the corresponding Lagrangian (the Information Bottleneck Lagrangian)

$$\mathcal{L} = H_{p,q}(y|z) + \beta I(z;x)$$

Trade-off between accuracy and compression governed by parameter β.

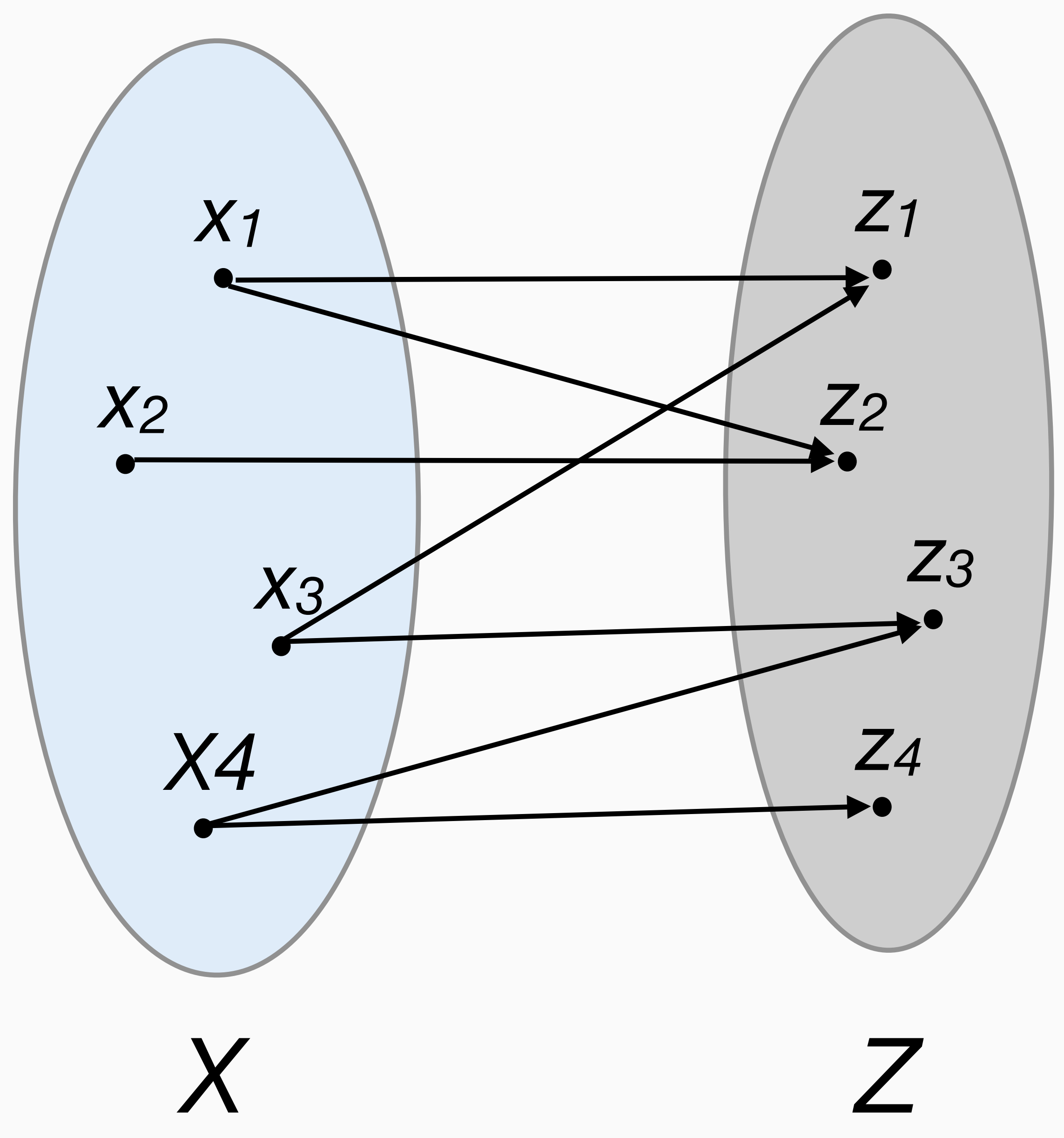
Compression in practice

Reduce the dimension



Examples: max-pooling, dimensionality reduction

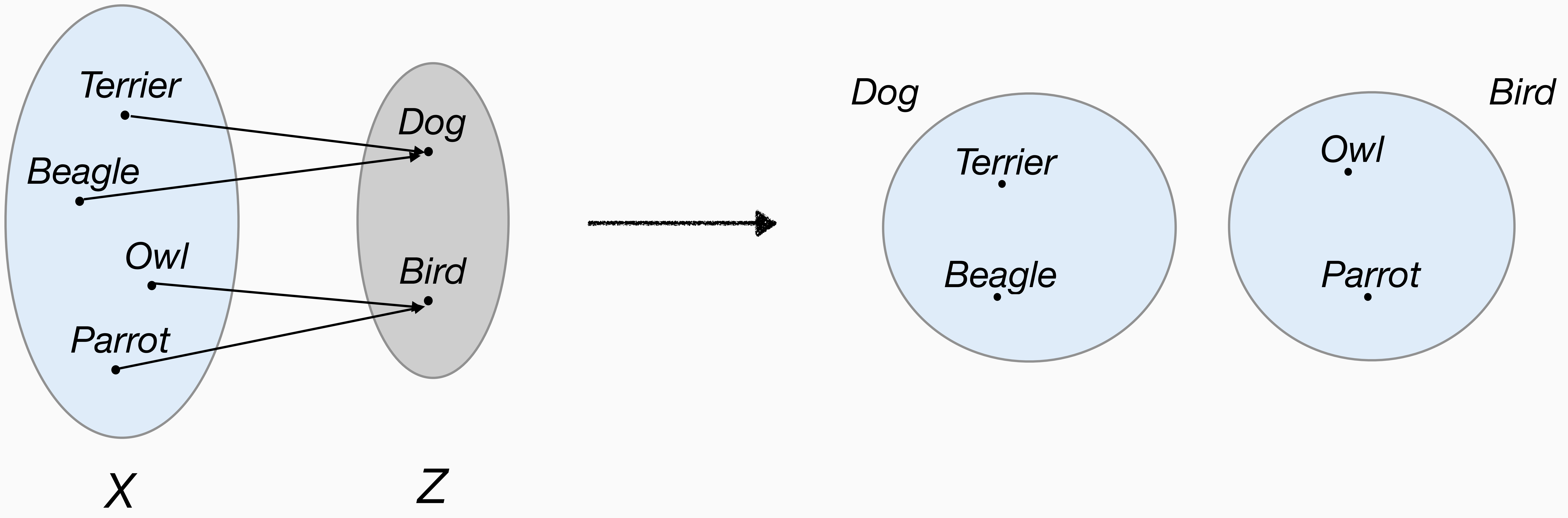
Increase dimension +
Inject noise in the map



Examples: Dropout, batch-normalization

Application to Clustering

An important application is task-based clustering, or summaries extraction.



See also Deterministic Information Bottleneck for hard-clustering vs soft-clustering.

Strouse and Schwab, *The Deterministic Information Bottleneck, 2016*

Information Bottleneck and Rate-Distortion

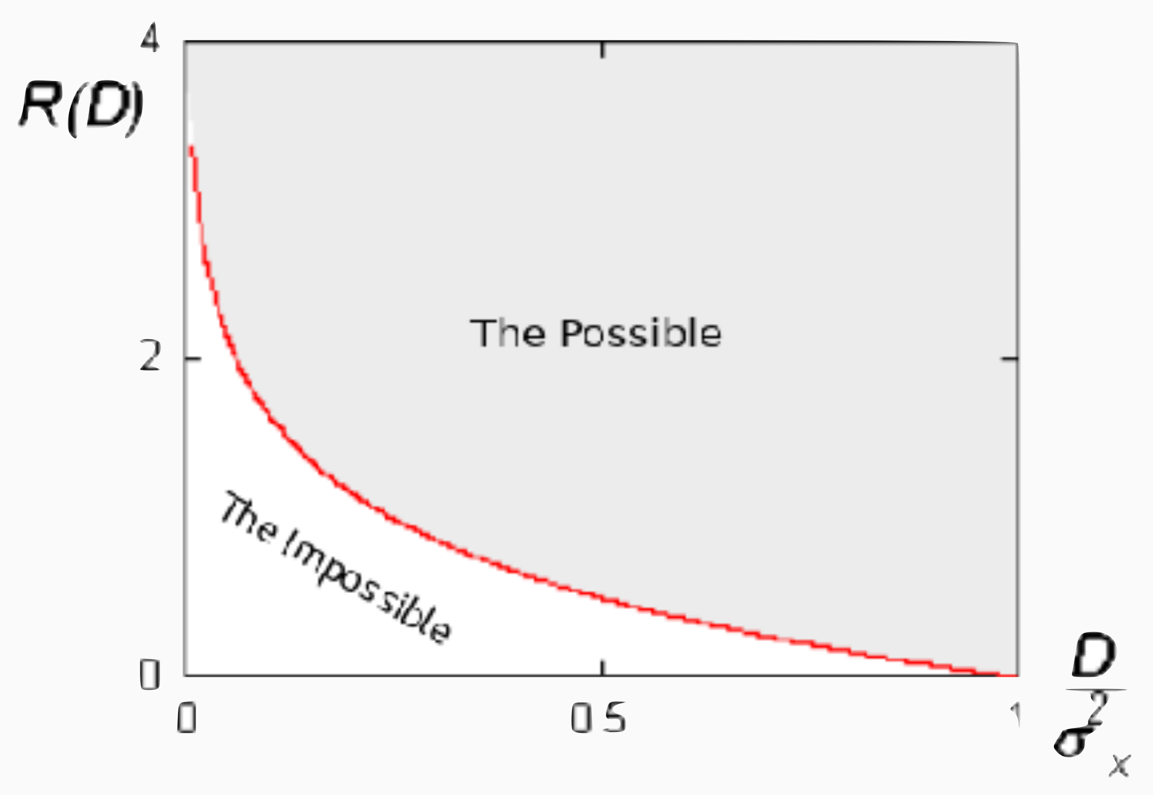
Rate-Distortion theory: What is the least distortion D obtainable with a given capacity R?

$$\min_{p(z|x)} \quad \mathbb{E}_{x,z}[d(x,z)]$$
$$\text{s.t} \quad I(z;x) \leq R$$

Equivalent to IB when *d(x, z)* is the information that *z* retains about *y*:

$$d(x,z) = KL(p(y|x) \| p(y|z))$$

Rate-distortion/IB curve:



Blahut-Arimoto algorithm

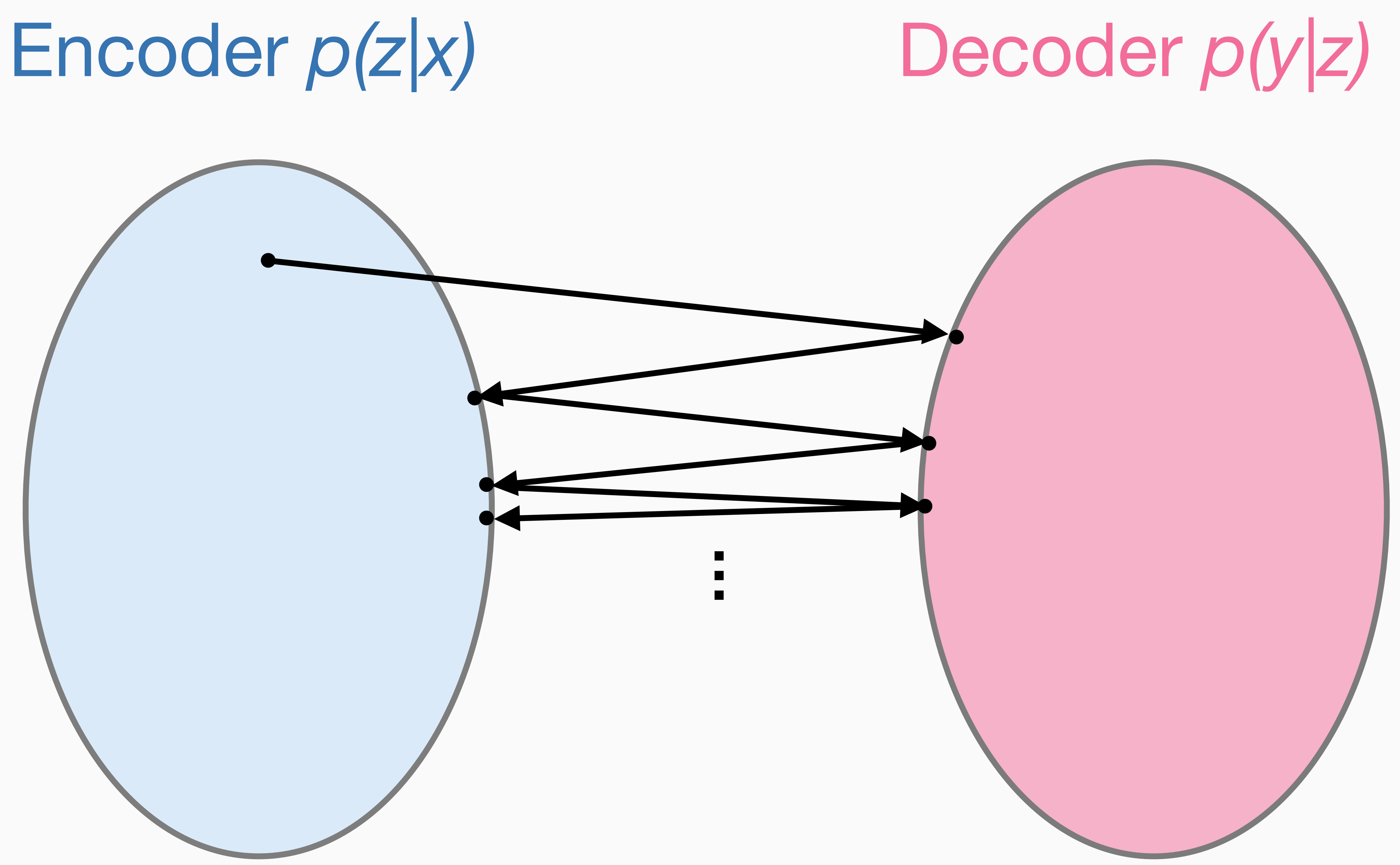
Blahut, 1972; Arimoto, 1972; Tishby et al., 1999

In general, no closed form solution. But we have the following iterative algorithm:

$$p_t(z|x) \leftarrow \frac{p_t(z)}{Z_t(x,\beta)} \exp(-1/\beta d(x,z))$$

$$p_{t+1}(z) \leftarrow \sum_x p(x) p_t(z|x)$$

$$p_{t+1}(y|z) \leftarrow \sum_y p(y|x) p_t(x|z)$$



But what happens if *p(z|x)* is too large, or parametrized in a non-convex way?