Stat 212b: Topics in Deep Learning
Lecture 14

Joan Bruna
UC Berkeley

Review:"Vanishing Gradient" Problem

- The parameters of the RNN are trained by gradient descent by *unrolling* T steps of the recurrence:

$$Y_{t+1} = \rho(A_{Y,Y} Y_t + A_{Y,X} X_t)$$

- For the purpose of updating the parameters, the loss at time t is thus expressed in terms of T previous hidden states:

$$\ell(\hat{O}_t, O_t) = G(Y_t, Y_{t-1}, \ldots, Y_{t-T}, X_t, \ldots, X_{t-T})$$

$$\frac{\partial \ell(\hat{O}_t, O_t)}{\partial A_{Y,Y}} = \sum_{i \leq T} \frac{\partial G}{\partial Y_{t-i}} \frac{\partial Y_{t-i}}{\partial A_{Y,Y}}$$

$$= \sum_{i \leq T} \frac{\partial G}{\partial Y_t} \left(\prod_{j \leq i} \frac{\partial Y_{t-j}}{\partial Y_{t-j-1}} \right) \frac{\partial Y_{t-i}}{\partial A_{Y,Y}}$$

Review:RNN Perspectives and Open Questions

- Prediction Challenge: capture long-term dependencies with tractable models.

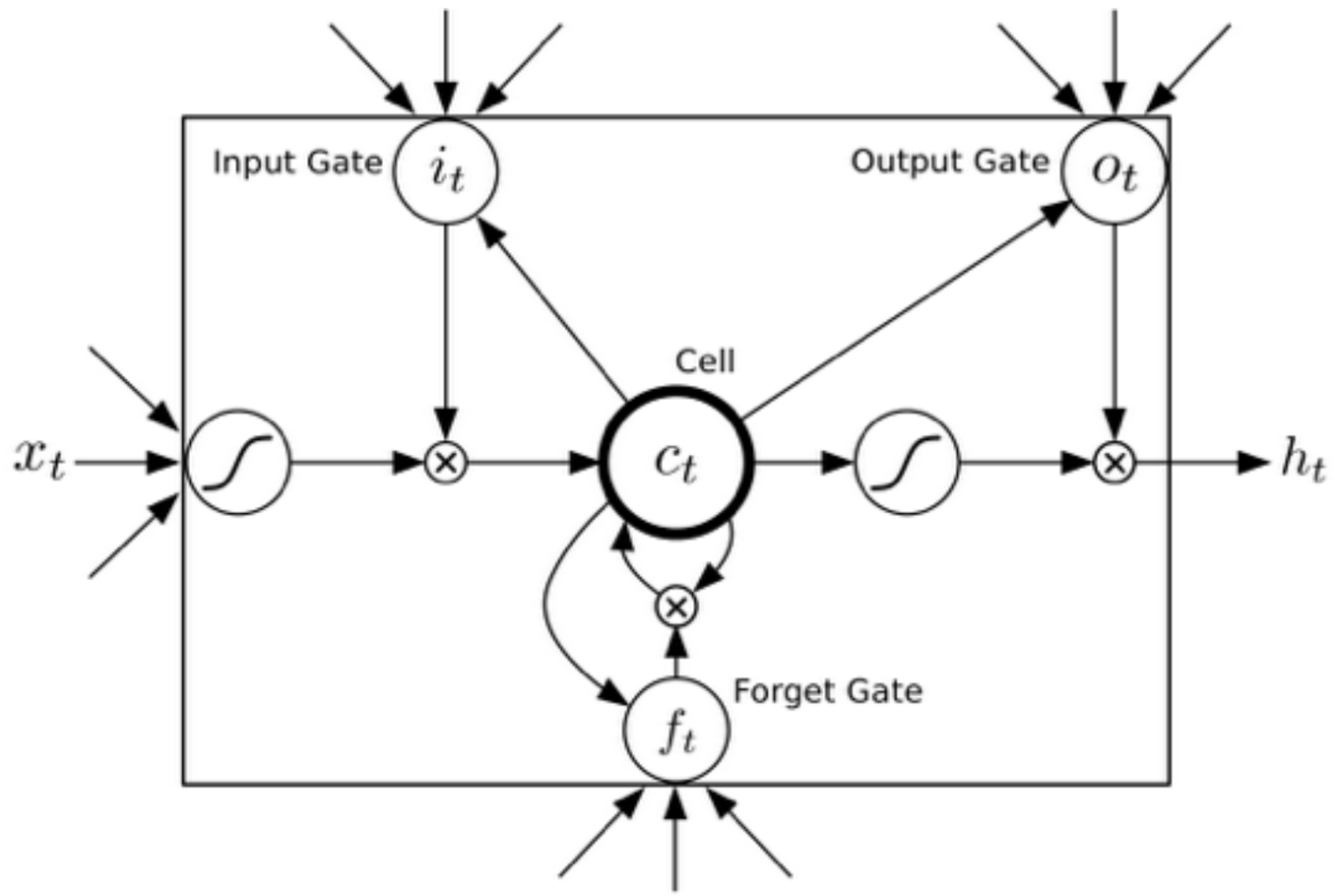
- Linear vs Non-linear state-space dynamics.
  - Can we trade-off higher dimensional linear dynamics with non-linear, lower-dimensional dynamics?
  - Role of gating and relationship with Residual Training. Optimization advantage or a more fundamental principle?

- Inference?

Review:Long Short Term Memory (LSTM)

[Hochreiter & Schmidhuber'97]

- A very popular and efficient alternative is to modify the transition operator using *gating mechanisms*:



- The *cell* is a memory that needs to be explicitly erased in order to disappear.
- What to store and when to write/erase is modeled with differentiable gates, trained with gradient descent.

Objective

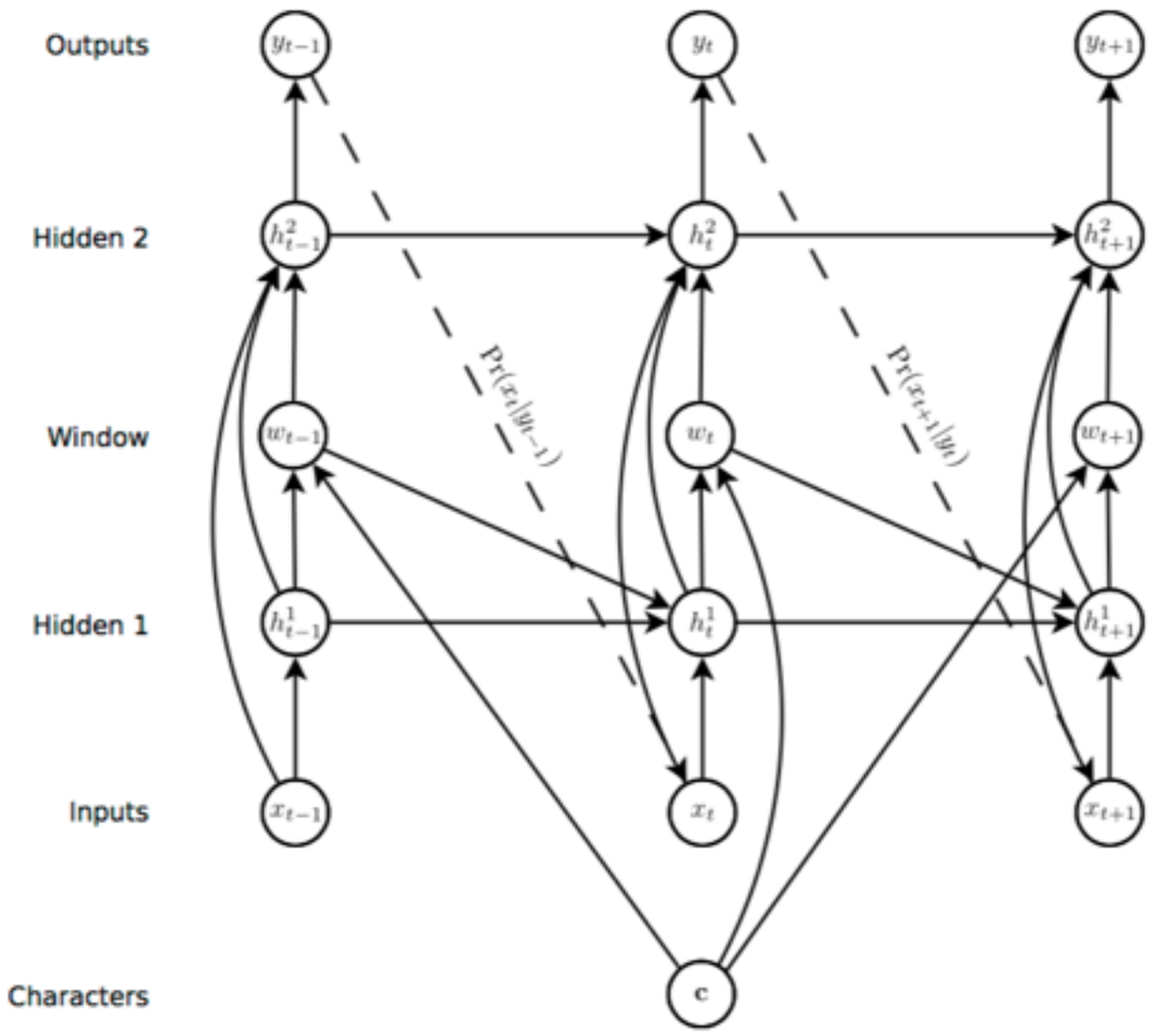
- Sequence Structured Prediction
  - Examples

- Unsupervised Learning
  - Graphical Models
  - Markov Random Fields
  - Introduction to Variational Inference

Handwitten Synthesis

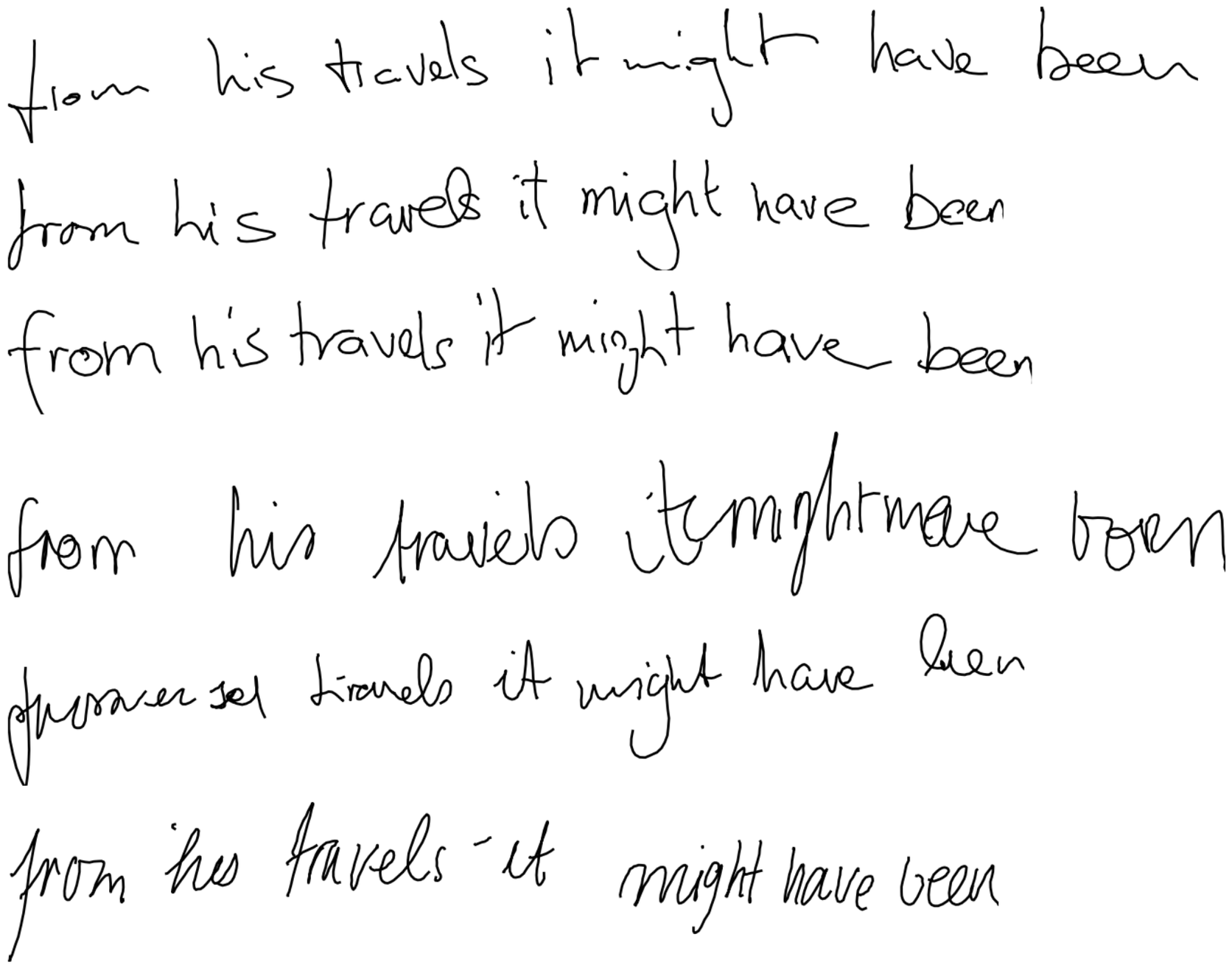
[A. Graves]

- Handwritten text is modeled with a mixture distribution over three-dimensional data (spatial coordinates and end-of-stroke)
- Three-layer LSTM network with approximately 3M parameters.



Handwitten Synthesis

[A. Graves,'13]



Sequence Structured Prediction

- Many tasks require a prediction from sequence to sequence:

Machine Translation

There is a light that never goes out

Il y a une lumière qui ne disparait jamais

Question Answering

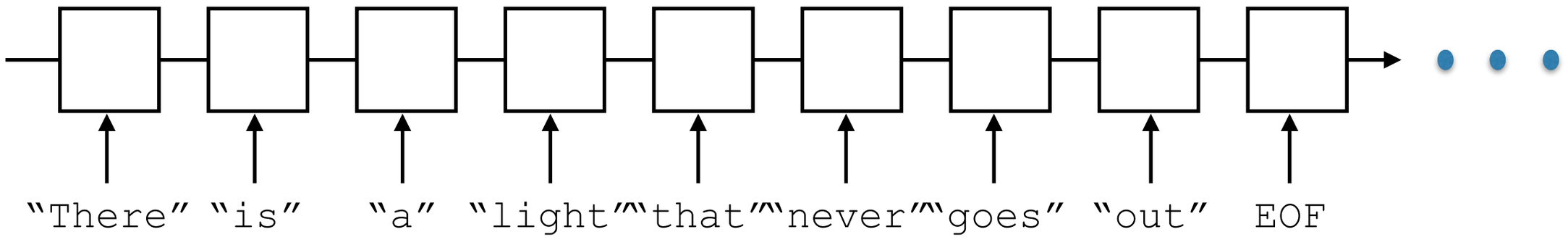
What is the best ramen place in the Bay Area?

Ramen Shop, in Rockridge

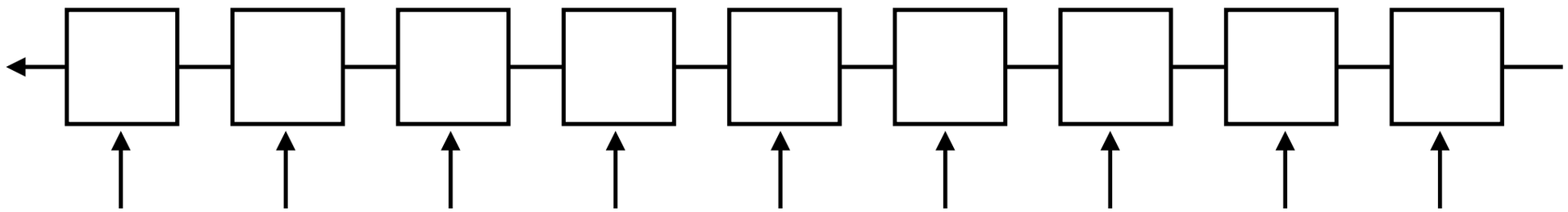
Sequence Structured Prediction

- Conditional model:
  - Input sequence is used to initialize the state of the output decoder.

input sequence X



$\phi(X)$

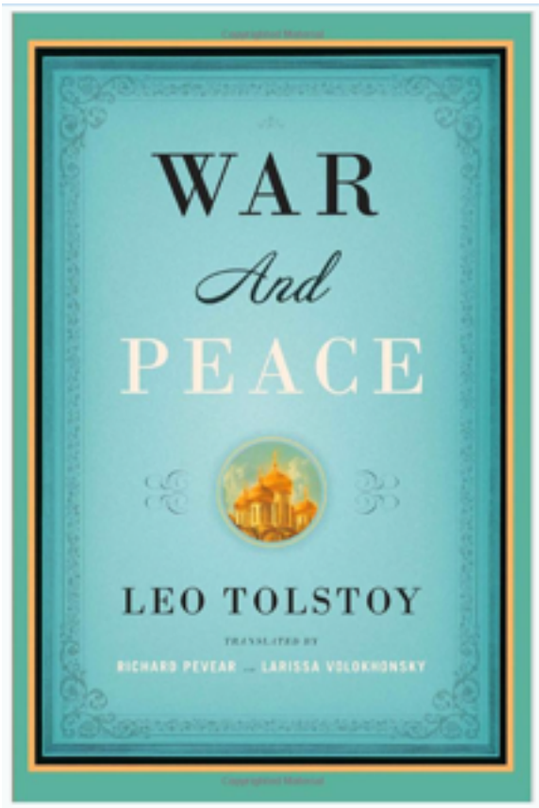


output sequence $\tilde{X}$

$$p(\tilde{X} \mid X) = \prod_{t=0}^{T} p(\tilde{X}_{t+1} \mid X, \tilde{X}_1, \ldots, \tilde{X}_t)$$

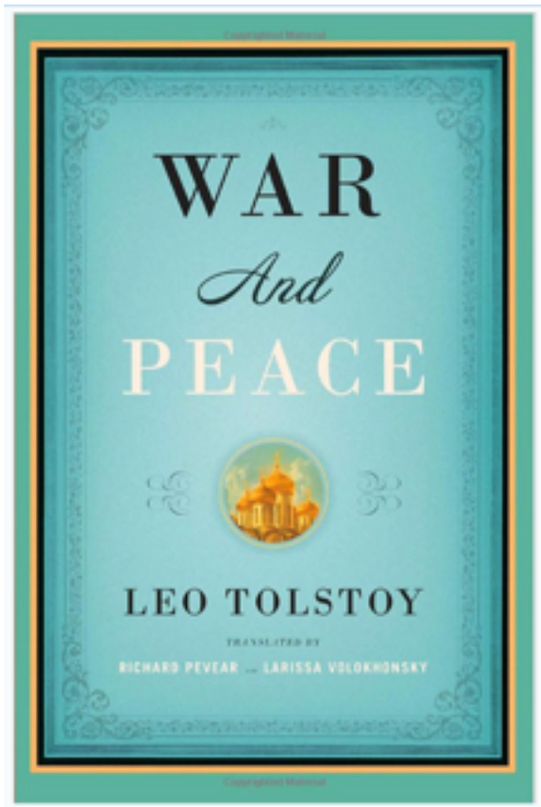
"Attention" Mechanisms

- Limits of sequence-to-sequence model.
  - All the information of the input sequence is contained in the vector $\phi(X)$
  - As the length of input increases, we require more information to perform the translation.



"Attention" Mechanisms

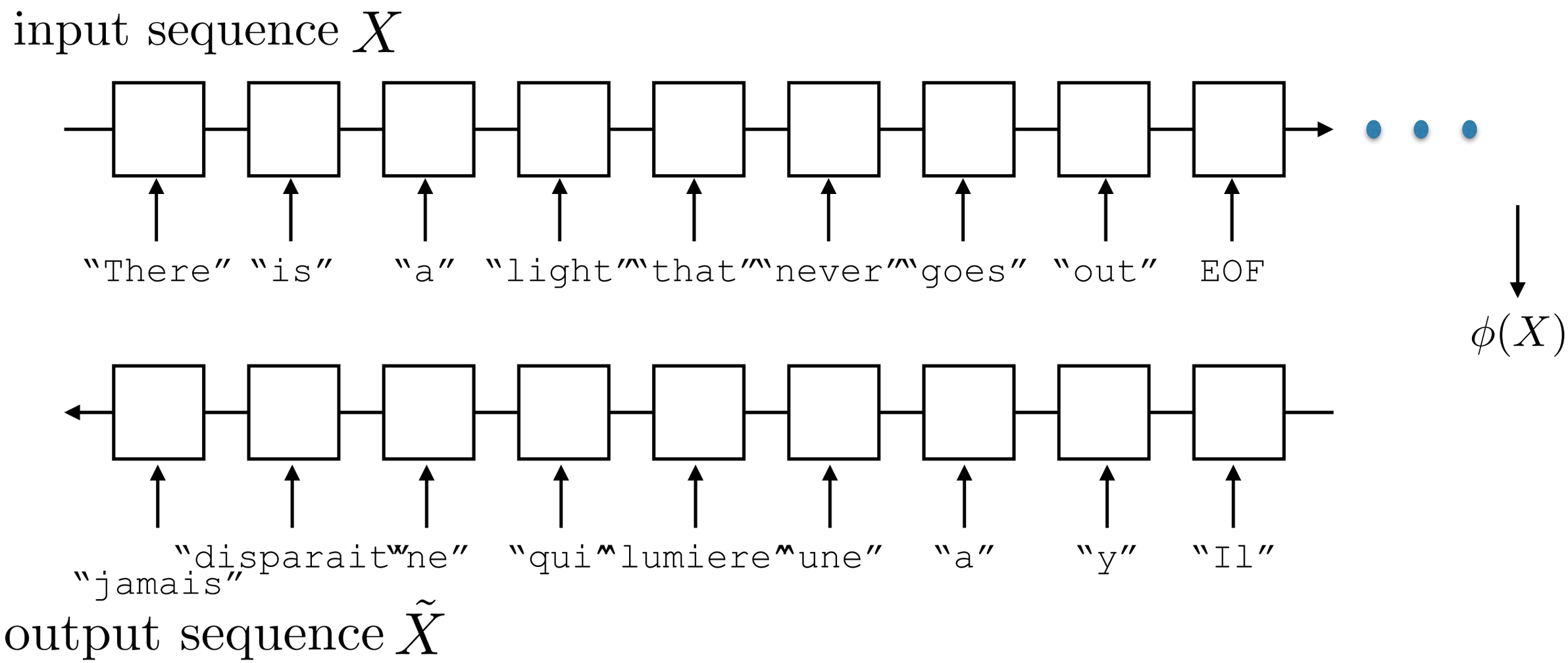
- Limits of sequence-to-sequence model.
  - All the information of the input sequence is contained in the vector $\phi(X)$
  - As the length of input increases, we require more information to perform the translation.



  - Although the ***global*** amount of information grows, the ***local*** amount of information required to translate does not. How to exploit it?

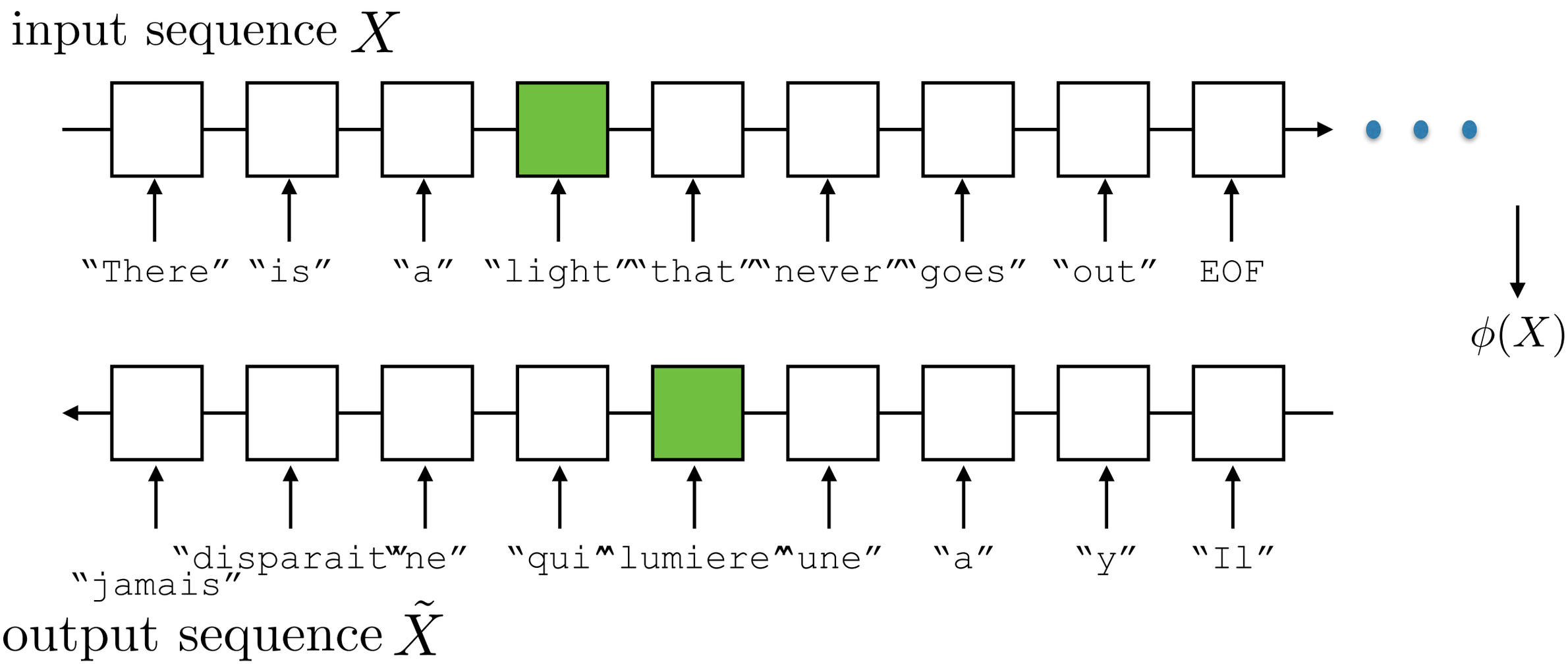
"Attention" Mechanisms

[Badhanu et al.'15]



"Attention" Mechanisms

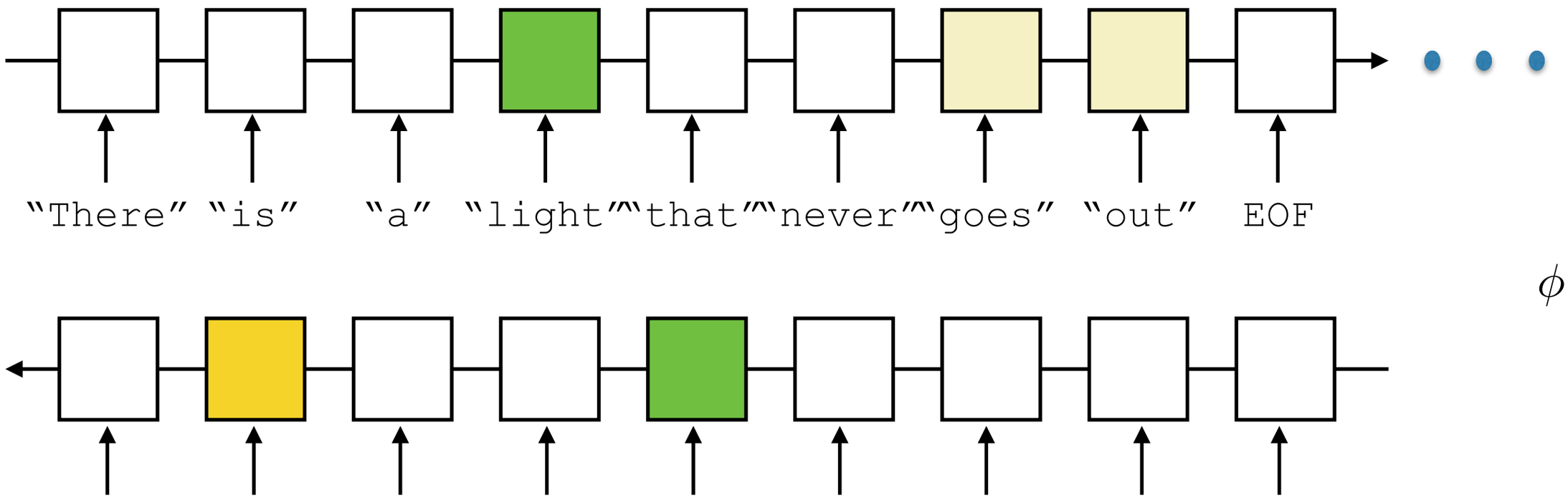
[Badhanu et al.'15]



"Attention" Mechanisms

[Badhanu et al.'15]

input sequence X



output sequence $\tilde{X}$

$$p(\tilde{X} \mid X) = \prod_{t=0}^{T} p(\tilde{X}_{t+1} \mid att(X, \tilde{X}_1, \ldots, \tilde{X}_t), \tilde{X}_1, \ldots, \tilde{X}_t)$$

"Attention" Mechanisms

[Badhanu et al.'15]

- Pros
  - Generalizes to larger input/output sequences.
- Challenges
  - Harder to train
  - How to address larger memories efficiently?
  - Learning where to look?

Machine Translation

[Badhanu et al., '15]

Source	An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.
Reference	Le privilège d'admission est le droit d'un médecin, en vertu de son statut de membre soignant d'un hôpital, d'admettre un patient dans un hôpital ou un centre médical afin d'y délivrer un diagnostic ou un traitement.
RNNenc-50	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.
RNNsearch-50	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Google Translate	Un privilège admettre est le droit d'un médecin d'admettre un patient dans un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, fondée sur sa situation en tant que travailleur de soins de santé dans un hôpital.

Image Captioning

[Vinyals et al'14, Karpathy et al '14, Donahue et al'14, Kiros et al'14,MSR'14]

- Sequence generation conditioned on visual features.

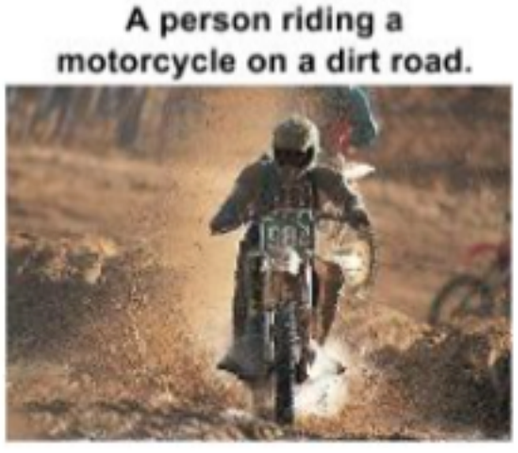


Image Captioning

[Vinyals et al'14, Karpathy et al '14, Donahue et al'14, Kiros et al'14,MSR'14]

- Sequence generation conditioned on visual features.

- Also, trained with visual attention:

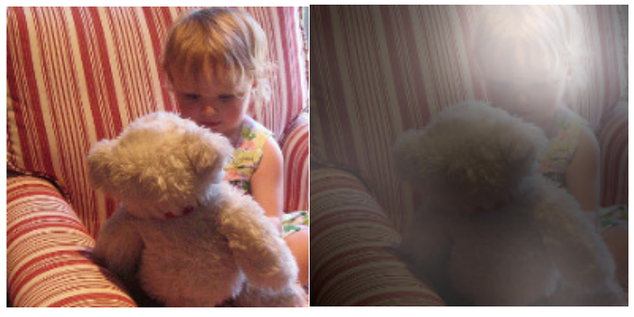
[Xu et al.'15]

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Unsupervised Learning

Unsupervised Learning

- Given high-dimensional data $X = (x_1, \ldots, x_n)$, we want to estimate a low-dimensional model characterizing the population.

- Why is this an important problem?

Unsupervised Learning

- Given high-dimensional data $X = (x_1, \ldots, x_n)$ we want to estimate a low-dimensional model characterizing the population.

- Why is this an important problem?
- It is an essential building block in most high-dimensional prediction tasks.
  - Inverse Problems (super-resolution, inpainting, denoising, etc.).
  - Structured Output Prediction (translation, Q&A, pose estimation, etc.)
  - "Disentangling" or Posterior Inference.
  - Learning with few labeled examples

Curse of Dimensionality

- *Challenge*: How to model $p(x)$, $x \in \mathbb{R}^N$ (or $x \in \Omega^N$) for large N ?

Curse of Dimensionality

- *Challenge*: How to model $p(x)$, $x \in \mathbb{R}^N$ (or $x \in \Omega^N$) for large N ?

- An existing hypothesis is that, although the ambient dimensionality is high, the *intrinsic* dimensionality of x is low.

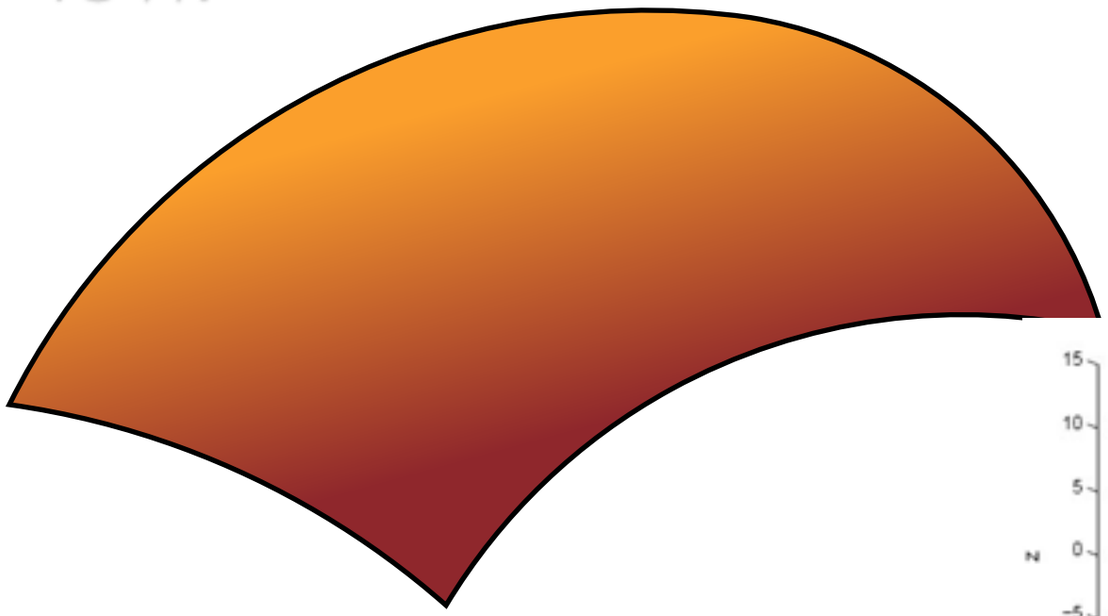
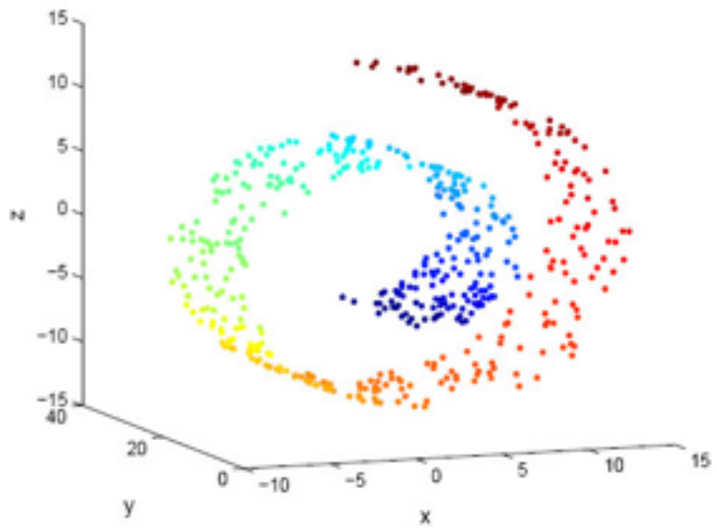
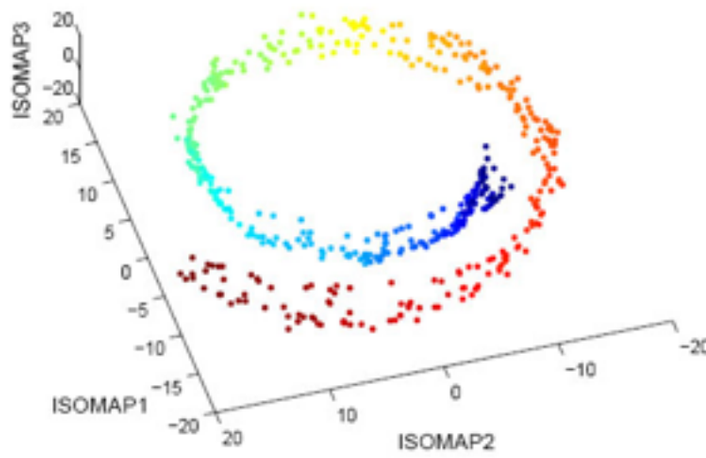


figure from Carter et al.



(a) Swiss Roll

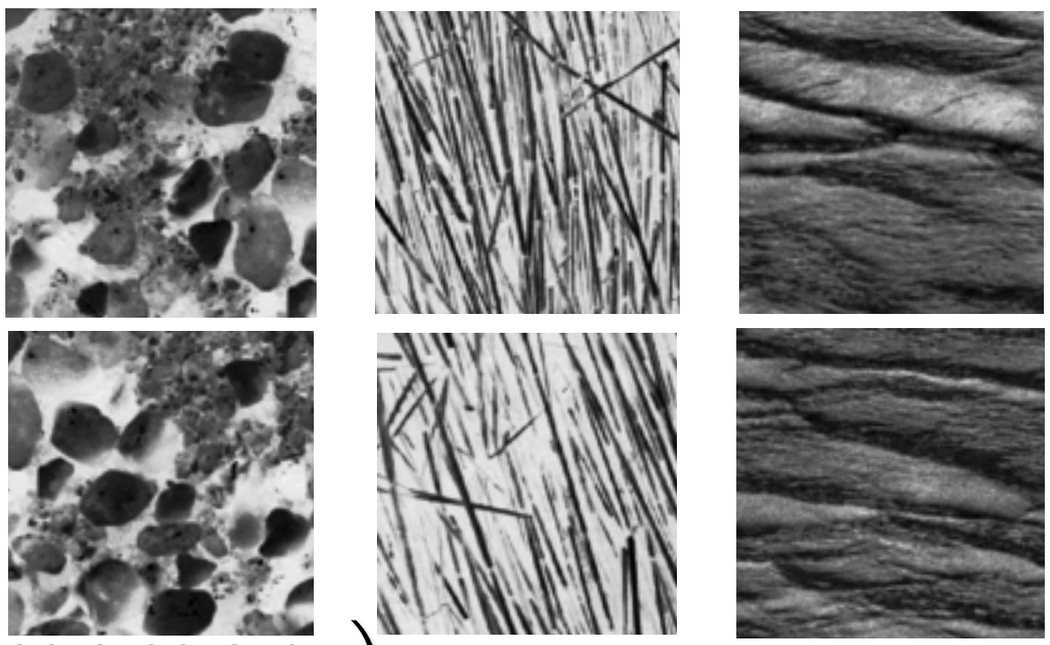


(b) Isomap embedding

Curse of Dimensionality

- However, many signals of interest *do have high intrinsic dimensionality:*
- Deformation structure is high-dimensional:

- Textures:



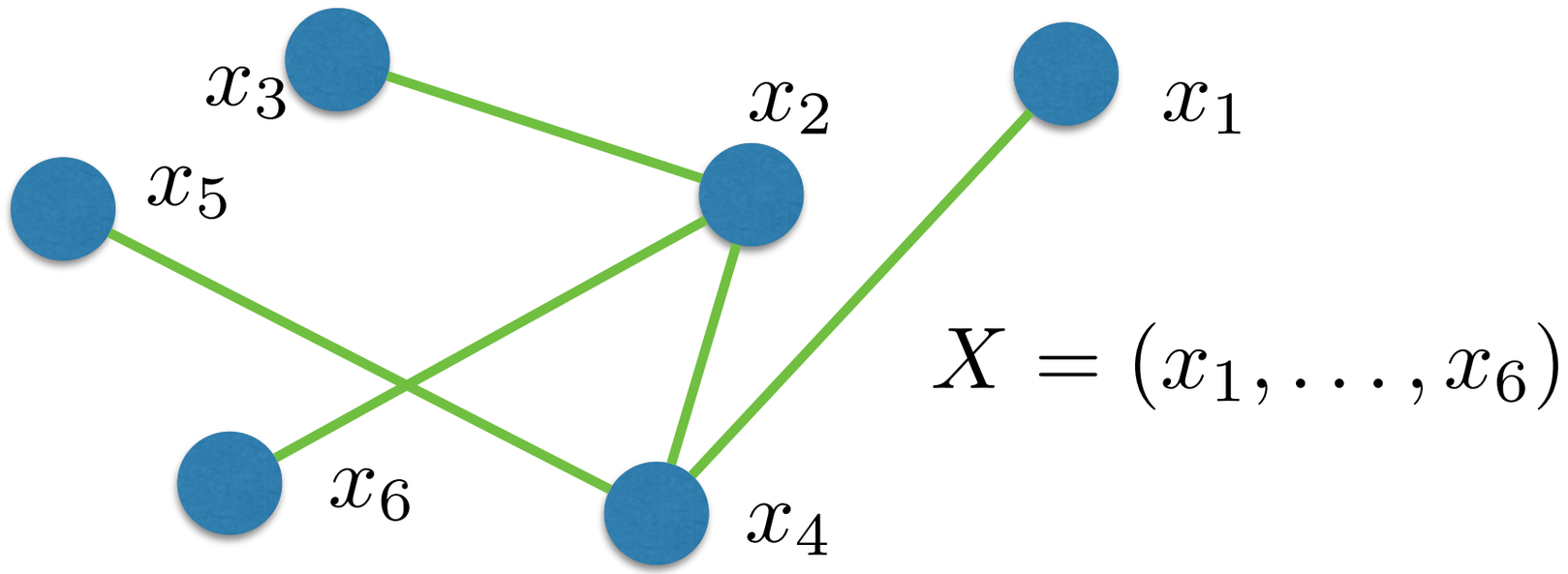
- Text (long memory)

Unsupervised Learning

- Latent Graphical Models
  - Variational Inference
  - Boltzmann Machines
- Autoencoders and dimensionality reduction.
  - Variational Autoencoders
- Measure Transportation
  - Generative Adversarial Networks.
- Gibbs Energy Models
  - Markov Random Fields
  - Maximum Entropy distributions
- How to evaluate high-dimensional generative models?
- From unsupervised to *self-supervised* models.

Graphical Models

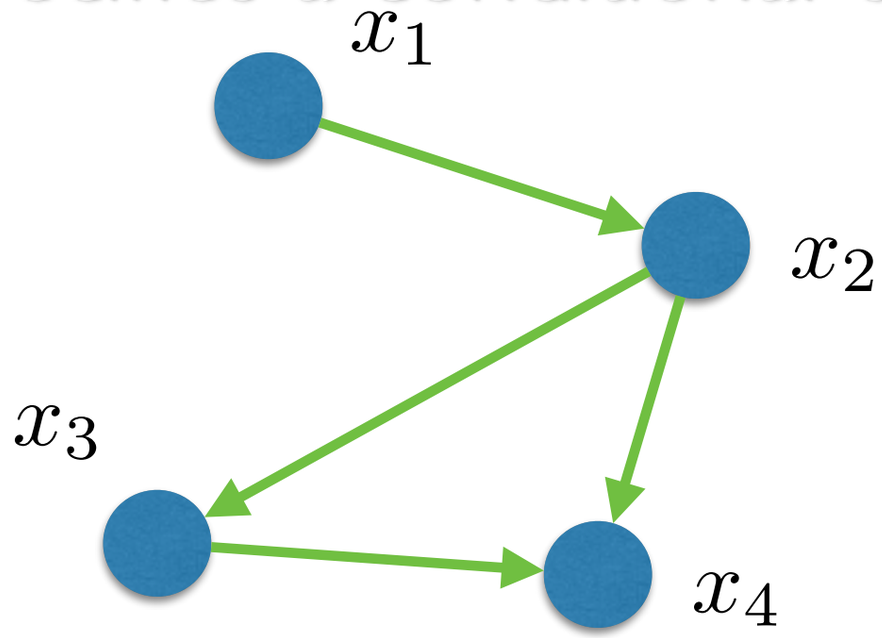
- A n-dimensional random vector is represented in a graph with n nodes.
  - Edges model statistical dependency between variables



- Two types of Graphical Models:
  - Directed Graphs: Use conditional distributions. Can express causal relationships.
  - Undirected Graphs: Energy based models.

Directed Graphical Models

- The direction specifies a conditional distribution:

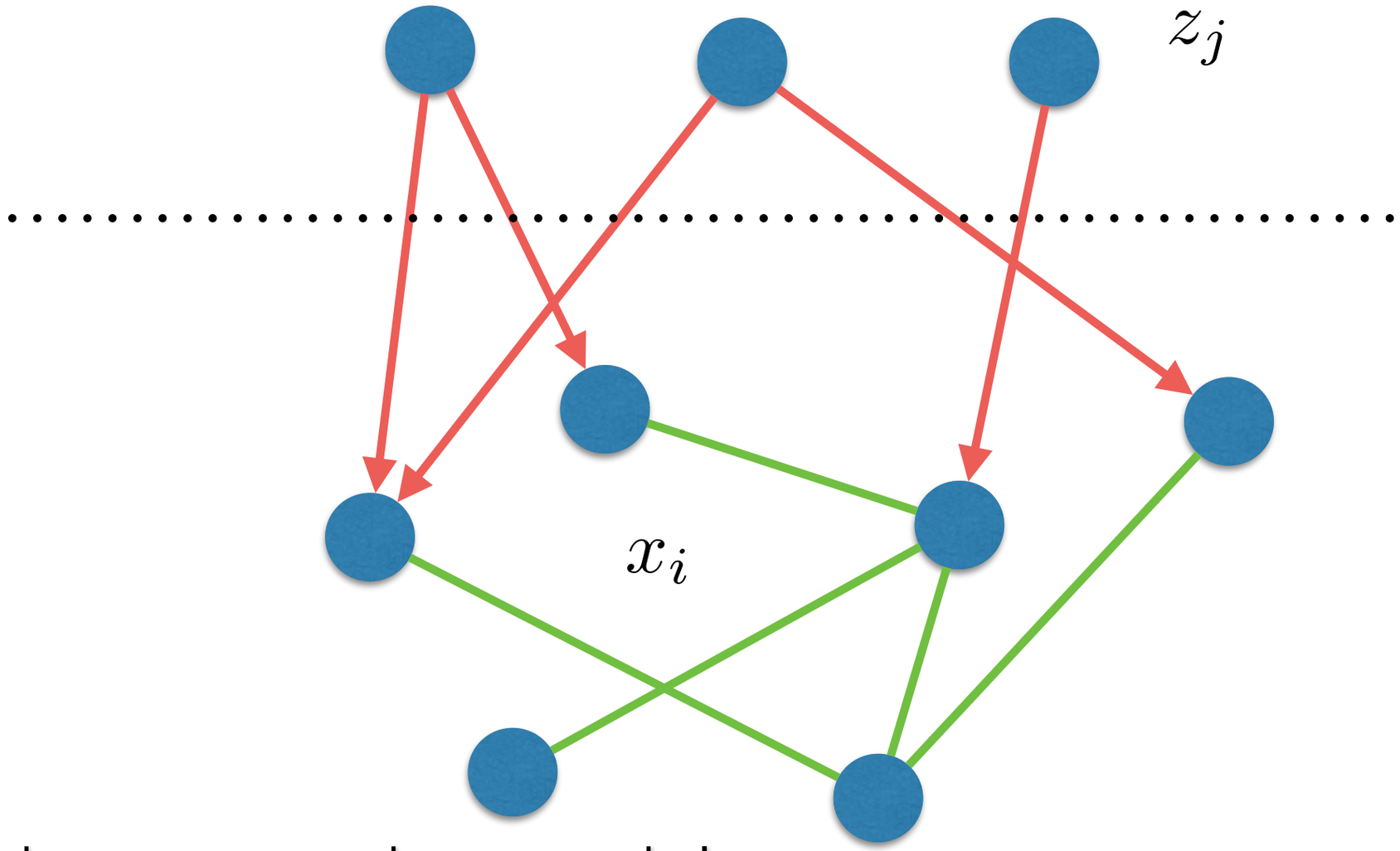


$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_2, x_3)$$

- It is well defined if the directed graph has no cycles (Directed Acyclic Graph).

Latent or Mixture Models

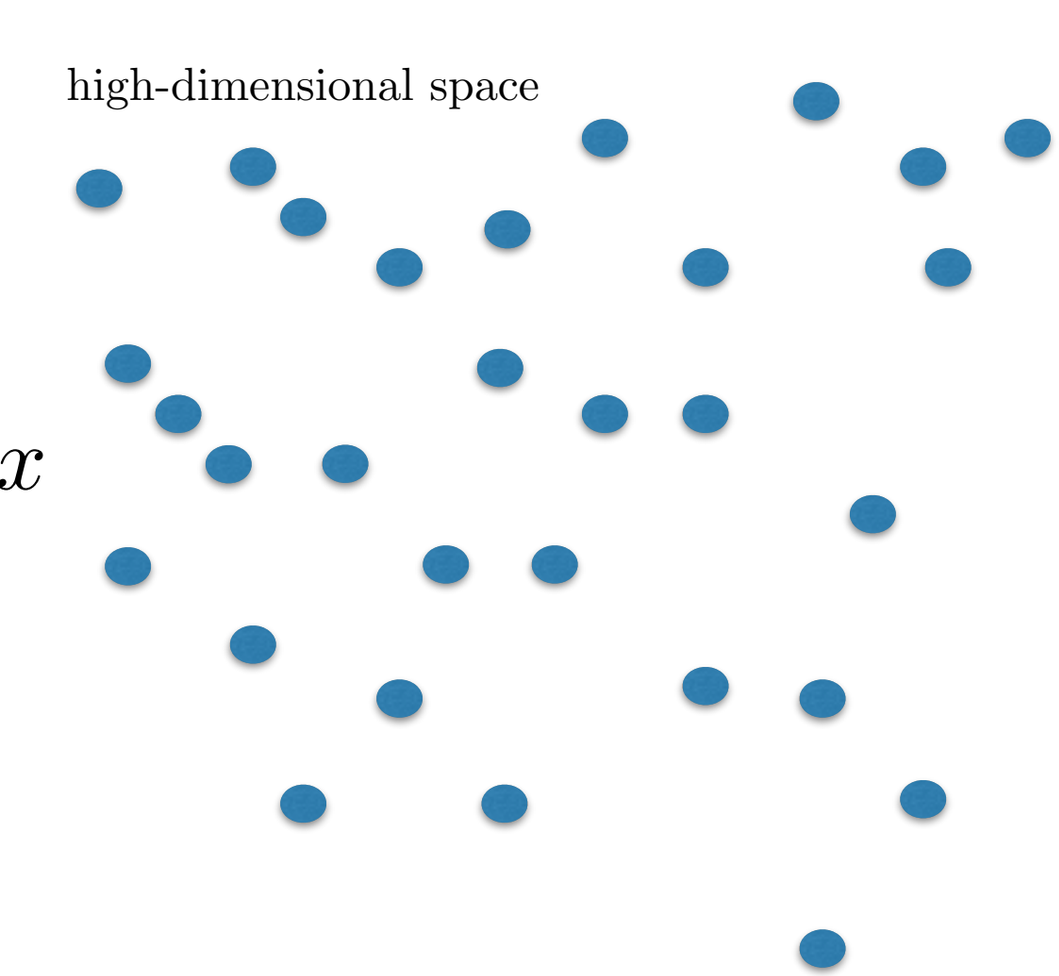
- A directed graph with bipartite structure, in which some variables are unobserved:

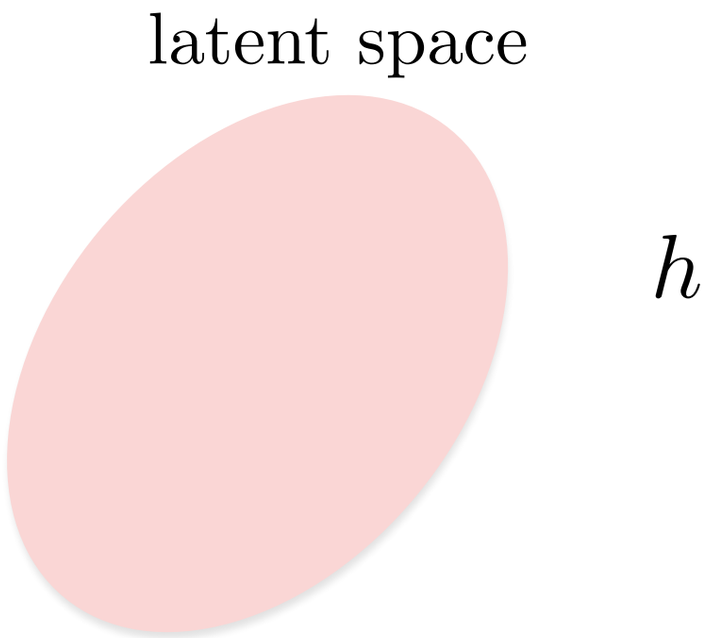


- Additive generative models
  - We can build complex generative models by *additively* combining simpler models.

Generative Models of Complex data

- Latent Graphical Models or *Mixtures*.

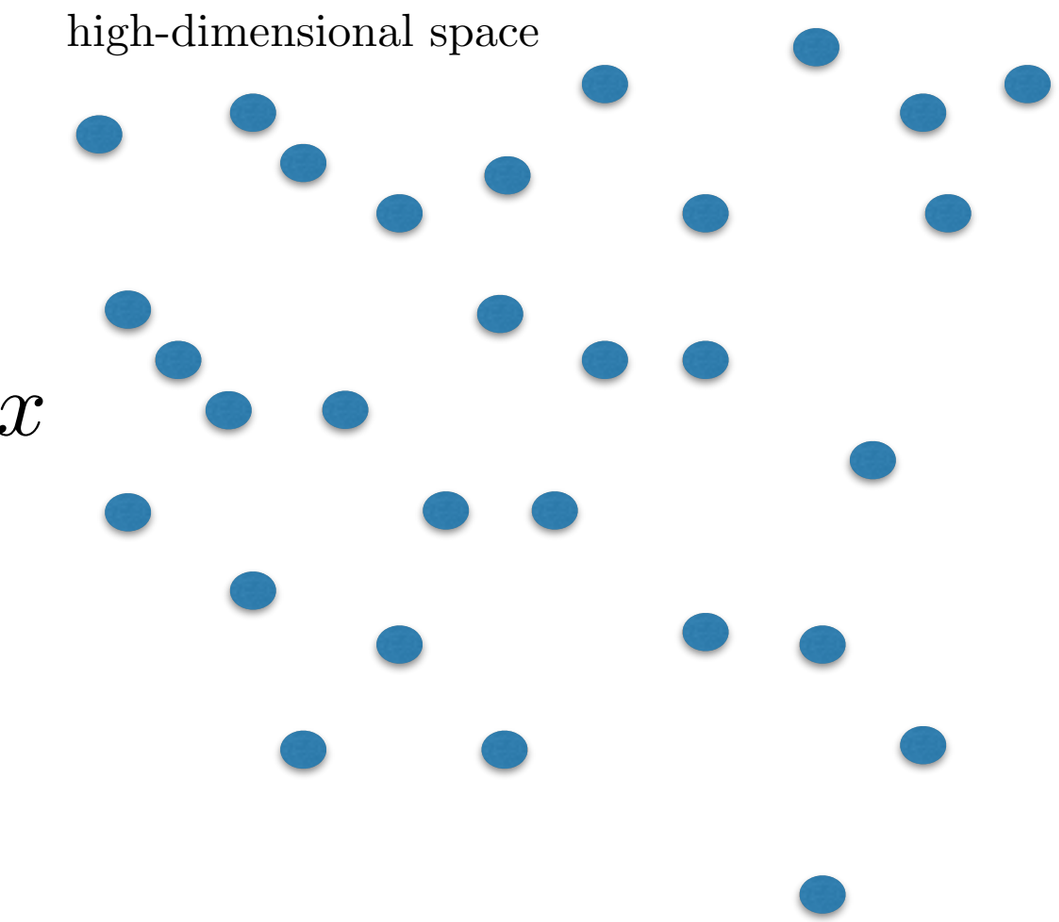


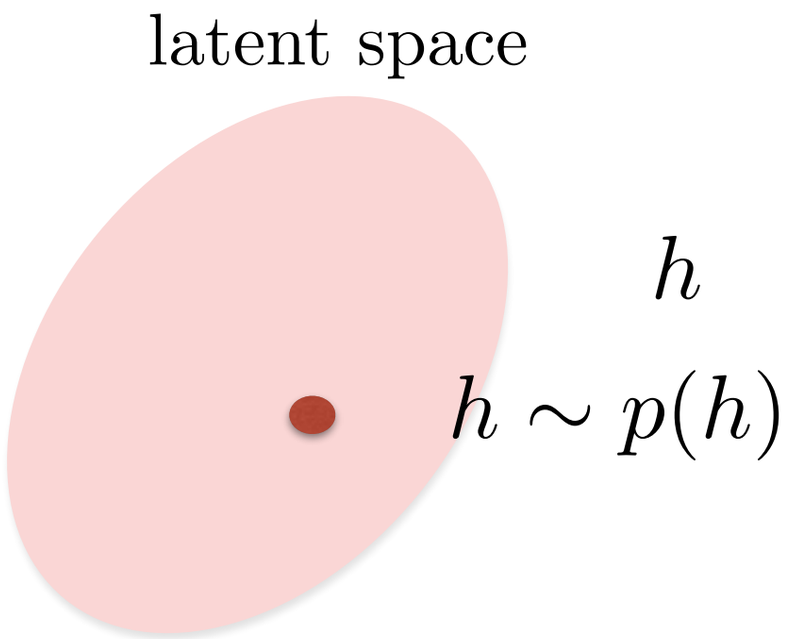


$$p(x) = \int p(x, h)dh = \int p(x \mid h)p(h)dh$$

Generative Models of Complex data

- Latent Graphical Models or *Mixtures*.

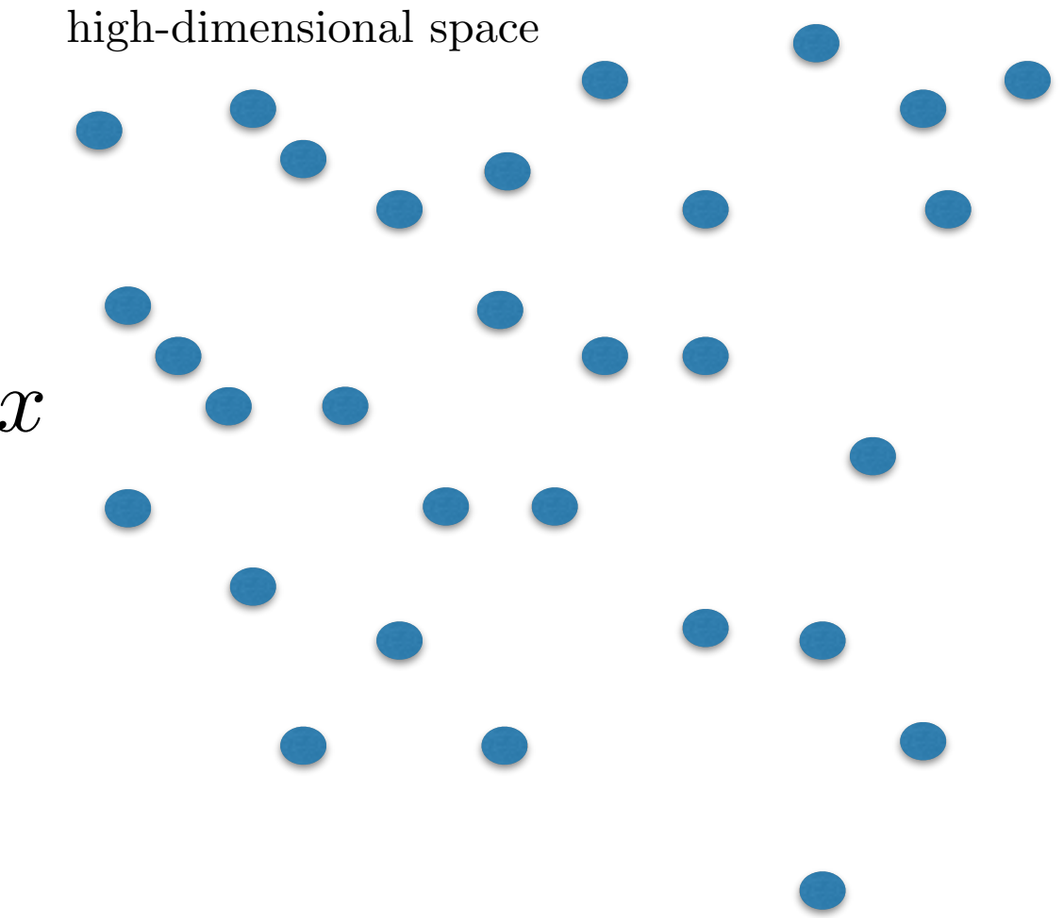


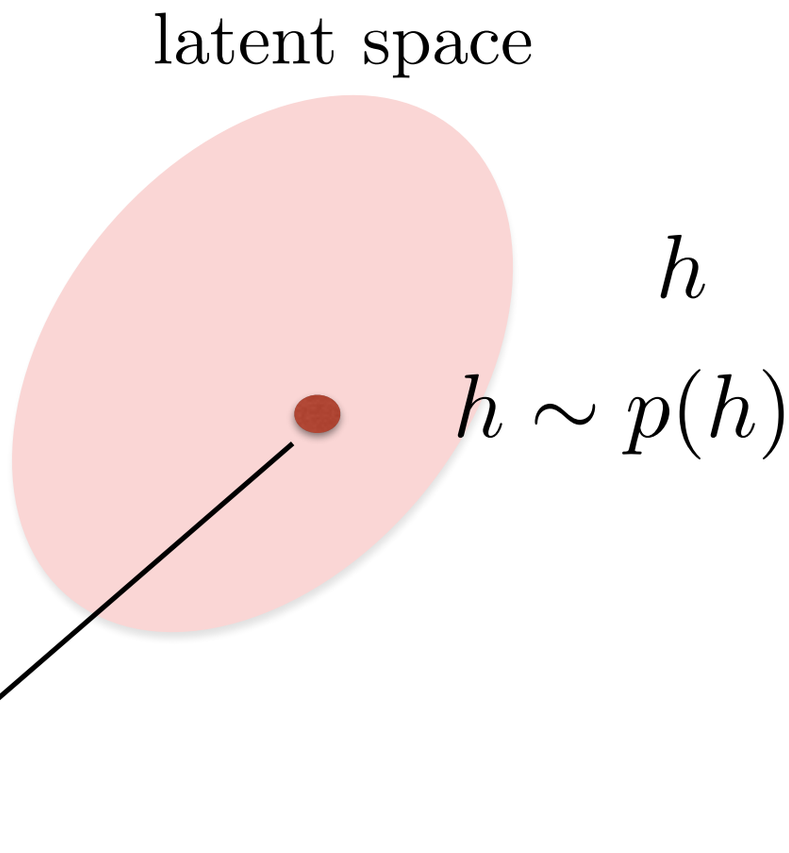


$$p(x) = \int p(x, h)dh = \int p(x \mid h)p(h)dh$$

Generative Models of Complex data

- Latent Graphical Models or *Mixtures*.



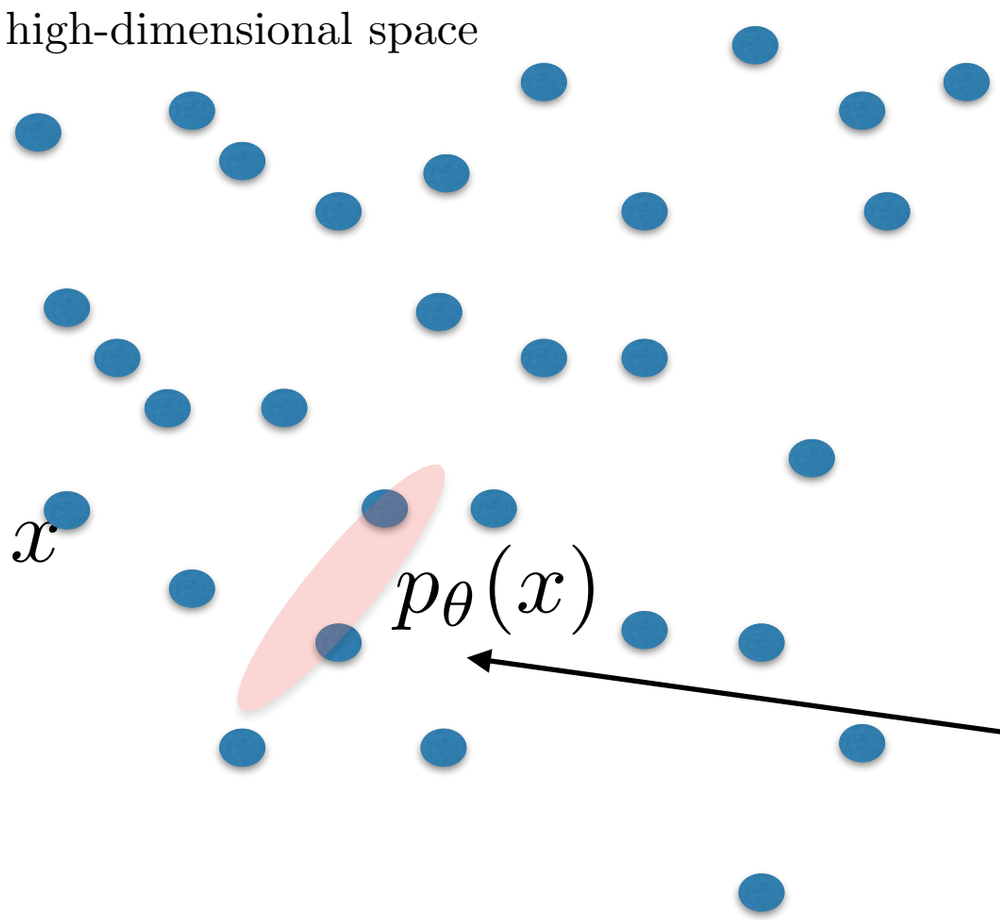


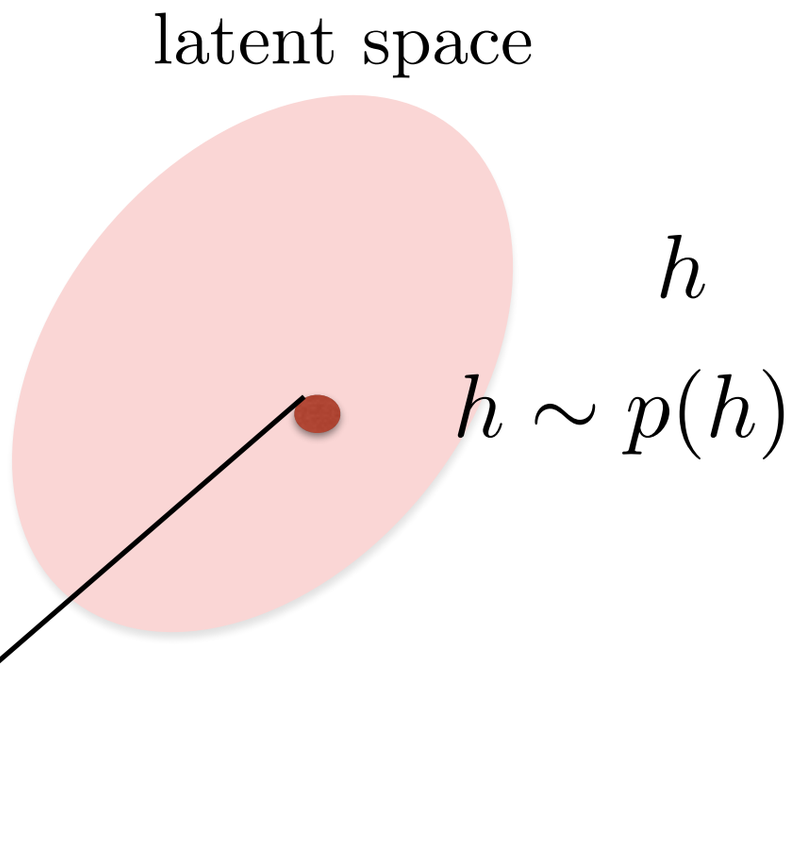
$$p(x \mid h) = p_{\Phi(h)}(x)$$

$$p(x) = \int p(x, h)dh = \int p(x \mid h)p(h)dh$$

Generative Models of Complex data

- Latent Graphical Models or *Mixtures*.



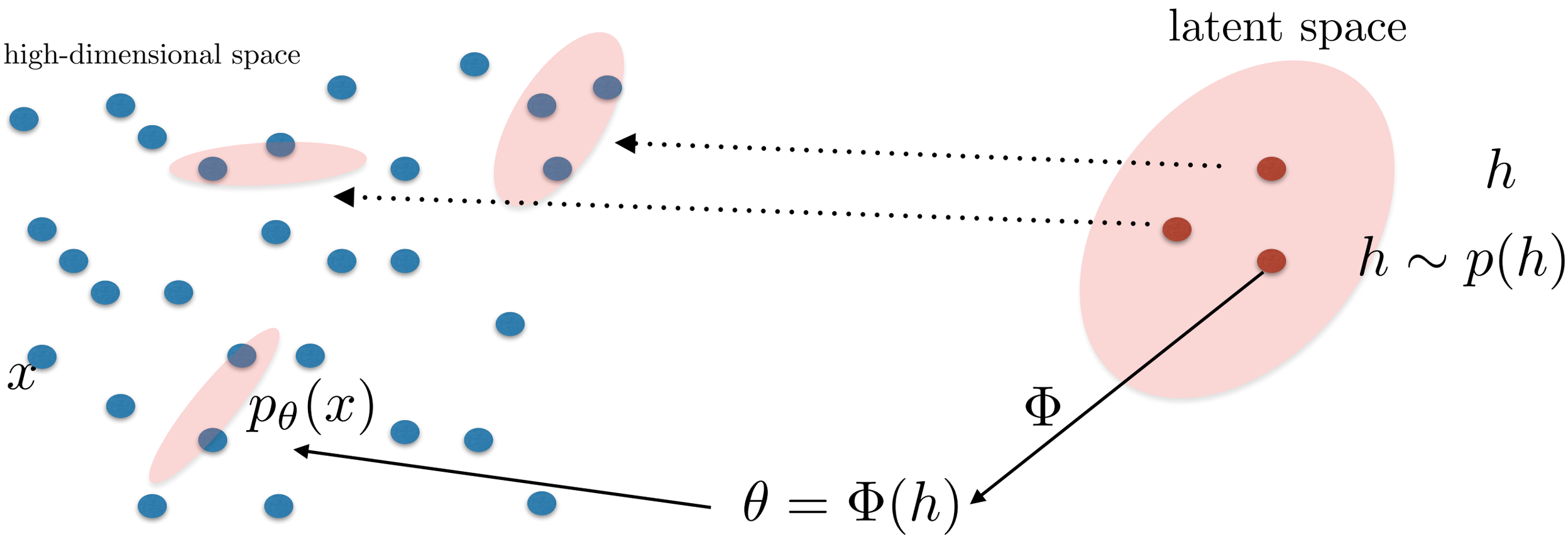


$$p(x \mid h) = p_{\Phi(h)}(x)$$

$$p(x) = \int p(x, h) dh = \int p(x \mid h) p(h) dh$$

Generative Models of Complex data

- Latent Graphical Models or *Mixtures*.



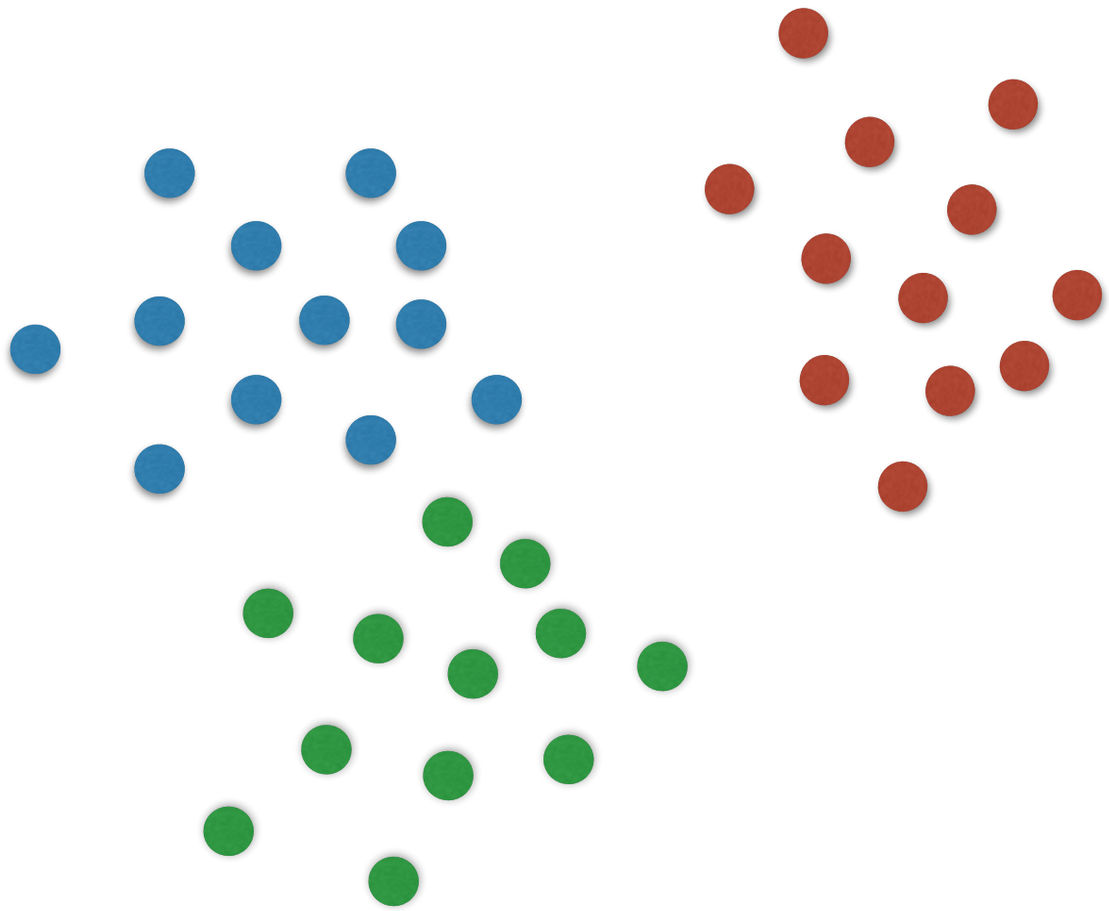
$$p(x \mid h) = p_{\Phi(h)}(x)$$

$$p(x) = \int p(x, h) dh = \int p(x \mid h) p(h) dh$$

RBM
DBN
DBM
VAE
...

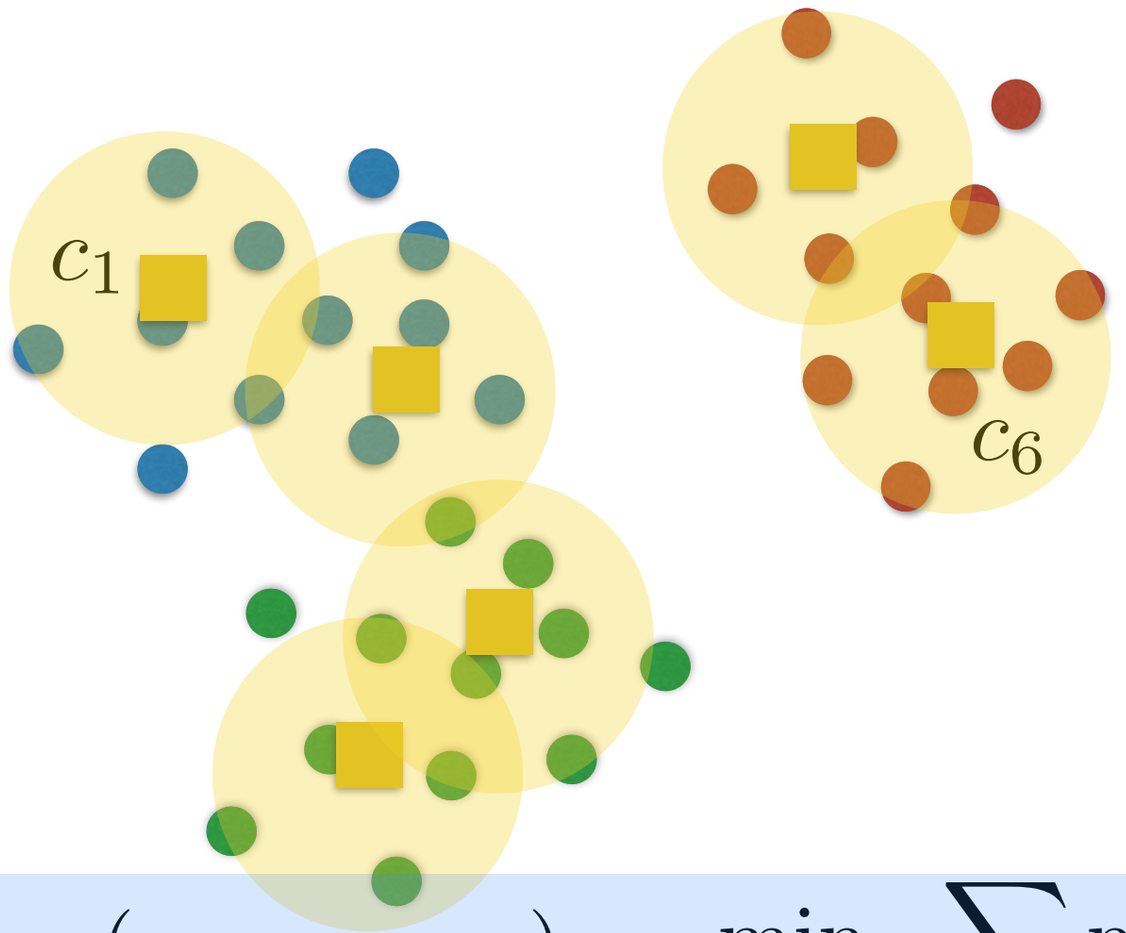
Latent Variables

- The simplest model is K-means clustering:



Latent Variables

- The simplest model is K-means clustering:



$K = 6$

Given data $X = (x_1, \ldots, x_n)$, $\min_{c_1,\ldots,c_K} \sum_{i \leq n} \min_j \|x_i - c_j\|^2$

Floyd Algorithm

- For each i, we define r_i a one-hot vector of length K encoding its cluster.
- Cost function is

$$E(c, r) = \sum_i \sum_k r_i(k) \|x_i - c_k\|^2$$

Floyd Algorithm

- For each i, we define r_i a one-hot vector of length K encoding its cluster.
- Cost function is

$$E(c, r) = \sum_i \sum_k r_i(k) \|x_i - c_k\|^2$$

- Fixing c, we optimize r as

$$r_i \leftarrow \arg\min_k \|x_i - c_k\|$$

Given assignments r, optimize E with respect to c:

$$c_k = \frac{\sum_i r_i(k) x_i}{\sum_i r_i(k)}$$

mean of all datapoints falling in cluster k

Floyd Algorithm

- This iterative algorithm converges towards a local optimum (each step decreases the cost).

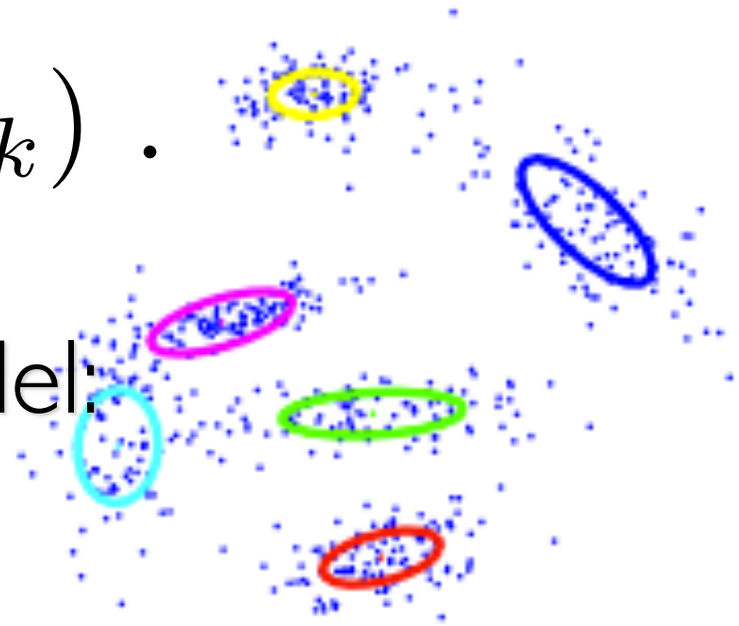
- It is in fact an instance of the Expectation-Maximization algorithm (EM).

- In that case, the discrete latent variables are the cluster assignments.

Gaussian Mixture Models (GMM)

- A generalization of K-Means is given by a Gaussian Mixture:

$$k \sim \text{Mult}(\pi) \ , \ x \sim \mathcal{N}(\mu_k, \Sigma_k) \ .$$



(figure from R.Salakhutdinov)

- This is also a discrete latent variable model:

$$z \in \{0,1\}^K \ , \ \sum_k z_k = 1 \ .$$

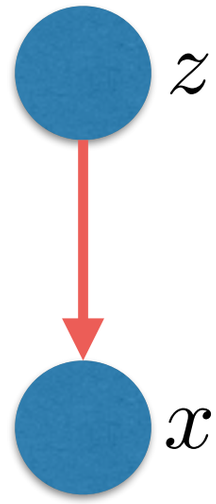
- The distribution of the latent variable is multinomial:

$$p(z_k = 1) = \pi_k \ , \ 0 \leq \pi_k \leq 1 \ , \ \sum_k \pi_k = 1 \ .$$

Gaussian Mixture Models (GMM)

- We can write

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad p(x \mid z_k = 1) = \mathcal{N}(x; \mu_k, \Sigma_k)$$

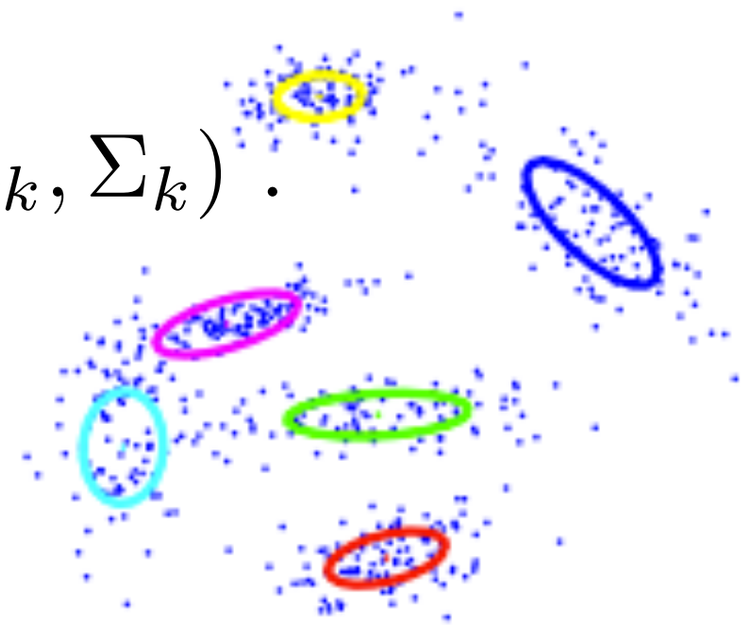


- Thus $p(x \mid z) = \prod_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k)^{z_k}$

- Joint and marginal distributions are given by

$$p(x, z) = p(x \mid z)p(z) \ ,$$

$$p(x) = \sum_z p(x, z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k) \ .$$



GMM and Posterior Inference

- What about the conditional p(z | x) ? Ie, given data, which mixture components are "responsible" ?

$$p(z_k = 1 \mid x) = \frac{p(z_k = 1, x)}{\sum_{k' \leq K} p(z_{k'} = 1, x)} = \frac{p(z_k = 1)p(x \mid z_k = 1)}{\sum_{k' \leq K} p(z_{k'} = 1)p(x \mid z_{k'} = 1)}$$

$$= \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x; \mu_{k'}, \Sigma_{k'})}$$

- The posterior probability that $z_k = 1$ is a weighted average of prior probabilities that depends upon the data.
- Q: How to estimate the parameters $\{\pi, \mu, \Sigma\}$?

Maximum Likelihood Estimation

- Given independent samples $X = \{x_1, \ldots, x_n\}$, the total log-likelihood is

$$E(\pi, \mu, \Sigma) = \log p(X \mid \pi, \mu, \Sigma) = \sum_{i \leq n} \log \left(\sum_k \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$$

- $$\frac{\partial E}{\partial \mu_k} = \sum_i \frac{\pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i; \mu_{k'}, \Sigma_{k'})} \Sigma_k^{-1}(x_i - \mu_k) \ .$$

$$\mu_k = \frac{1}{N_k} \sum_i p(z_{i,k} = 1 \mid x_i) x_i \ , \ N_k = \sum_i p(z_{i,k} = 1 \mid x_i) \ .$$

Thus the mean μ_k is the weighted average of datapoints, with weights given by the posterior probabilities of belonging to component k.

Maximum Likelihood Estimation

- Similarly

$$\frac{\partial E}{\partial \Sigma_k} = 0 \;\Rightarrow\; \Sigma_k = \frac{1}{N_k} \sum_i p(z_{i,k} = 1 \mid x_i)(x_i - \mu_k)(x_i - \mu_k)^T \;.$$

$$\frac{\partial E}{\partial \pi_k} = 0 \;\Rightarrow\; \pi_k = \frac{N_k}{n} \;.$$

- MLE parameters do not have closed-form solution
  - Parameters depend upon posterior probabilities $p(z_k = 1 \mid x)$, which themselves depend upon parameters.
- Iterative algorithm: Expectation-Maximization (EM):
  - E-step: Update posterior probabilities with parameters fixed.
  - M-step: Update parameters with posterior probabilities fixed.

The EM algorithm

- It is designed to find MLE solutions of latent variable models.
- In general, we have log-likelihoods of the form

$$\log p(X \mid \theta) = \log \left(\sum_Z p(X, Z \mid \theta) \right) \quad , \quad \begin{array}{l} \theta = \text{model parameters} \; . \\ Z = \text{latent variables} \end{array}$$

- Using current parameters θ_{old} , we compute the expected total likelihood of the model (E-step):

$$\mathcal{Q}(\theta, \theta_{old}) = \mathbb{E}_{Z \sim p(Z \mid X, \theta_{old})} \log p(X, Z \mid \theta)$$

- Then we update the parameters to maximize this likelihood:

$$\theta_{new} = \arg\max_{\theta} Q(\theta, \theta_{old}) \; .$$