Variational Auto-encoder: A Short Tutorial

Yong Jiang

Tencent AI Lab && ShanghaiTech

http://jiangyong.site

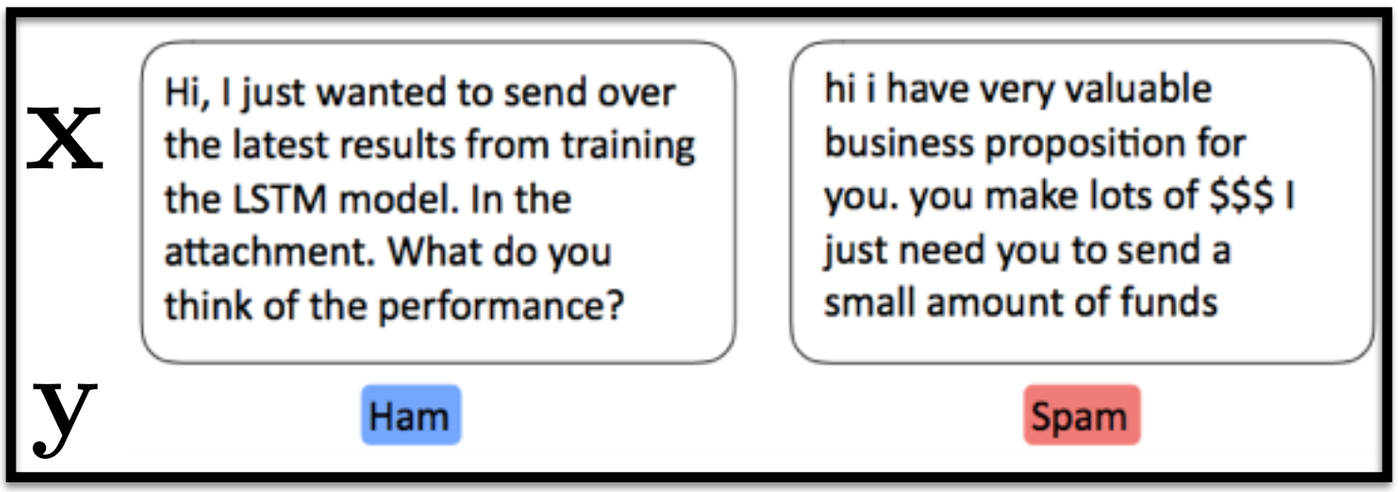
Survey

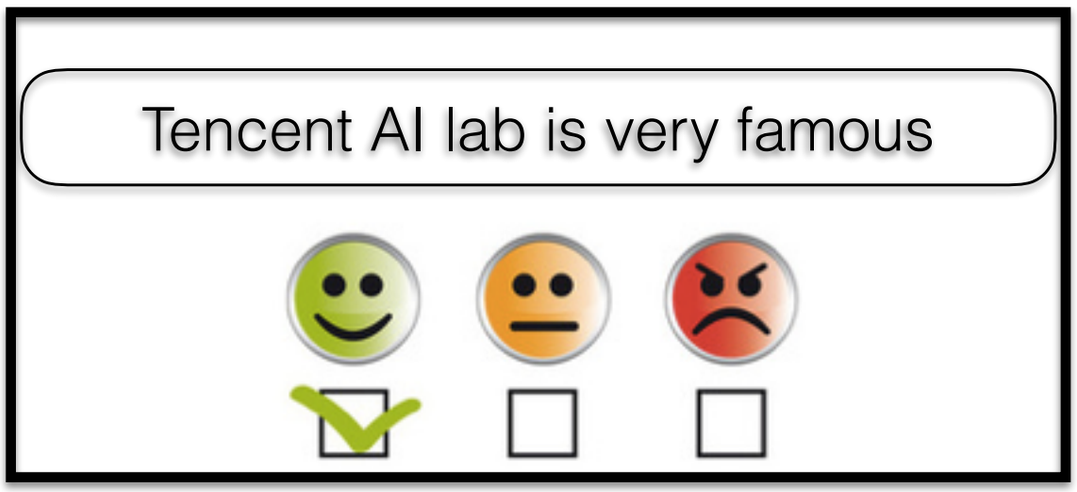
- Publish papers on VAE
- Write code on VAE
- Read papers on VAE

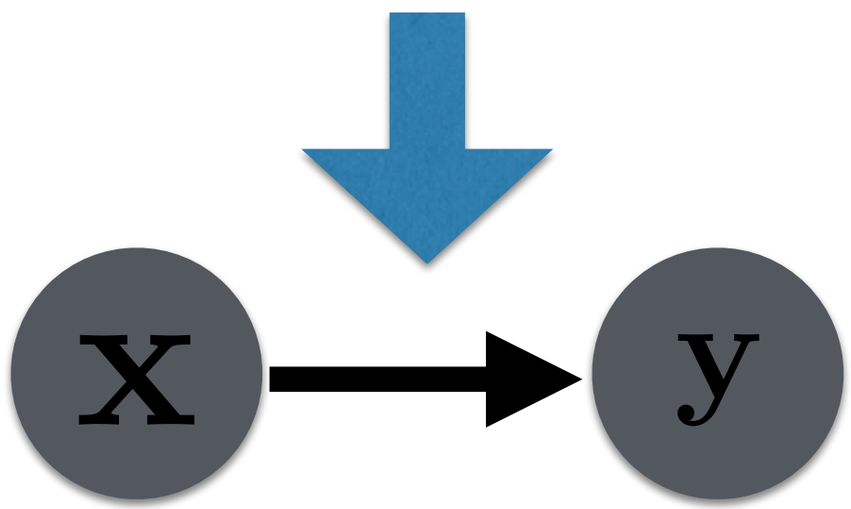
- Familiar with EM/variational inference/MCMC
- Familiar with latent variable modeling
- Heard of latent variable modeling

Important Tasks in NLP

- Text Classification
  - Ham or Spam
  - Sentiment Classification

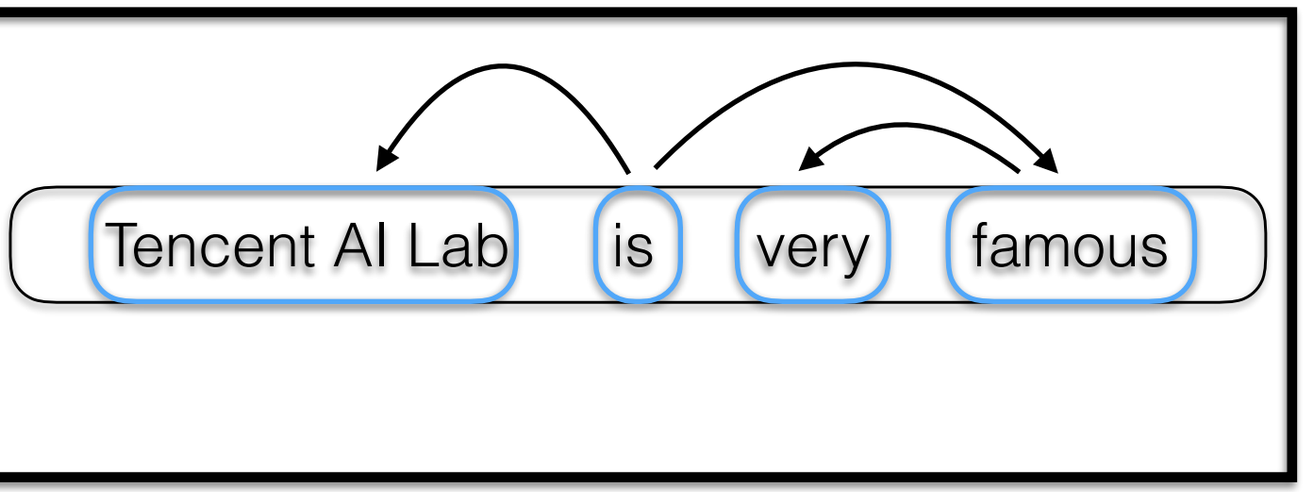


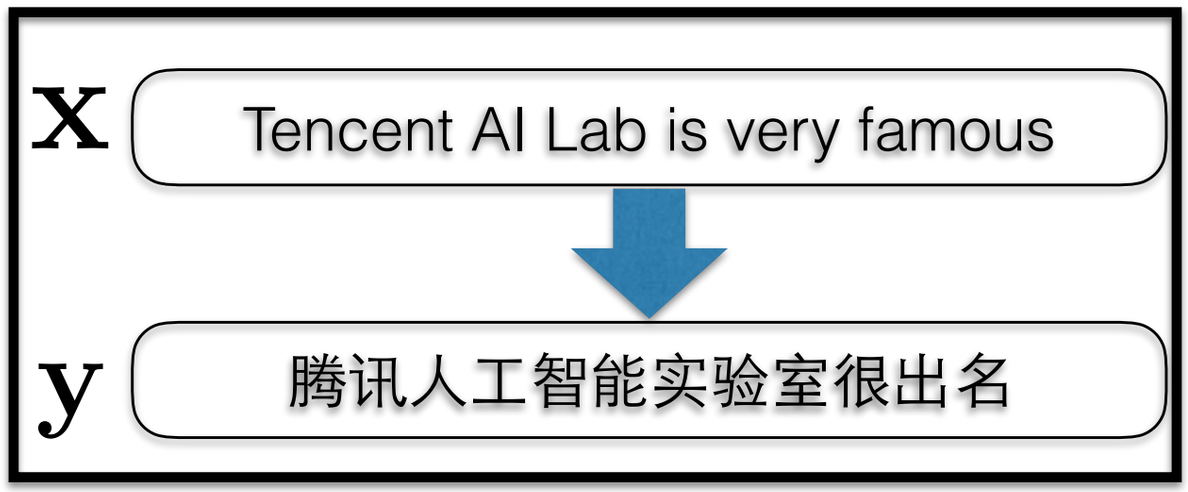


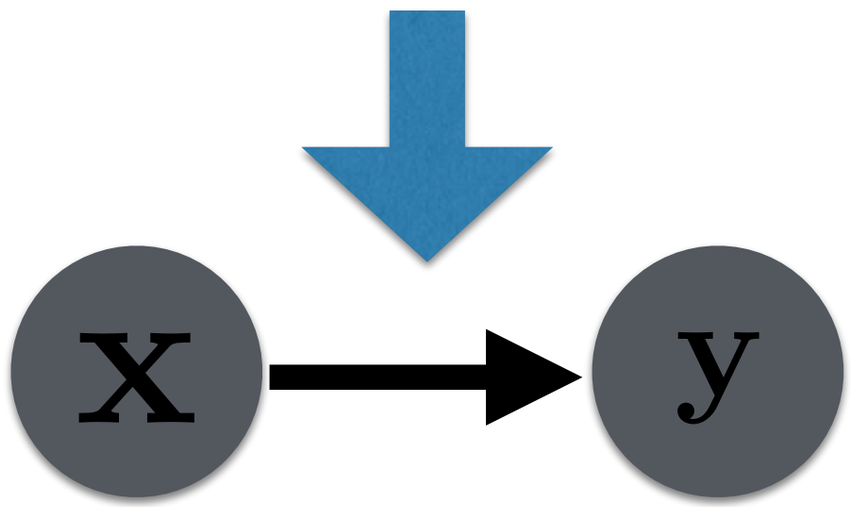


Important Tasks in NLP

- Structured Prediction
  - Parsing
  - Machine Translation

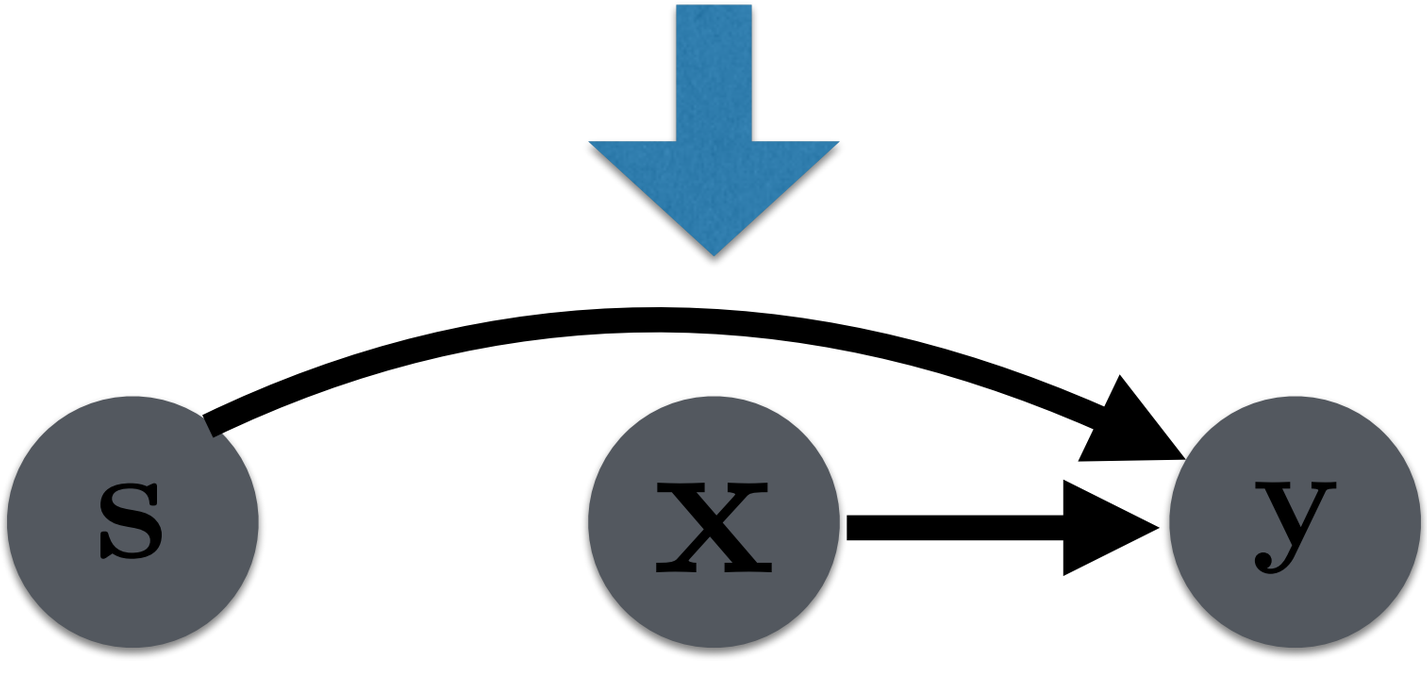






Important Tasks in NLP

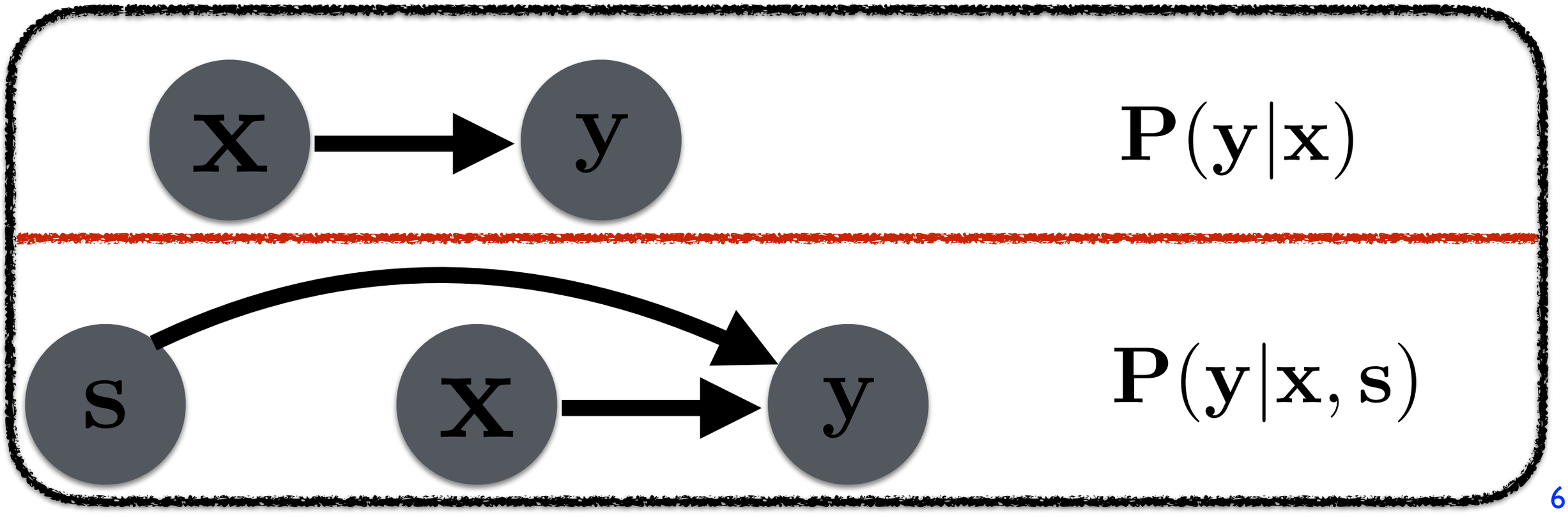
- Knowledge based Question Answering
  - Stock QA



Important Tasks in NLP

- Text Classification
- Structured Prediction
- Knowledge based QA

All random variables are observed in training data



However…

Life is not easy!

- A simple classification problem…

However…

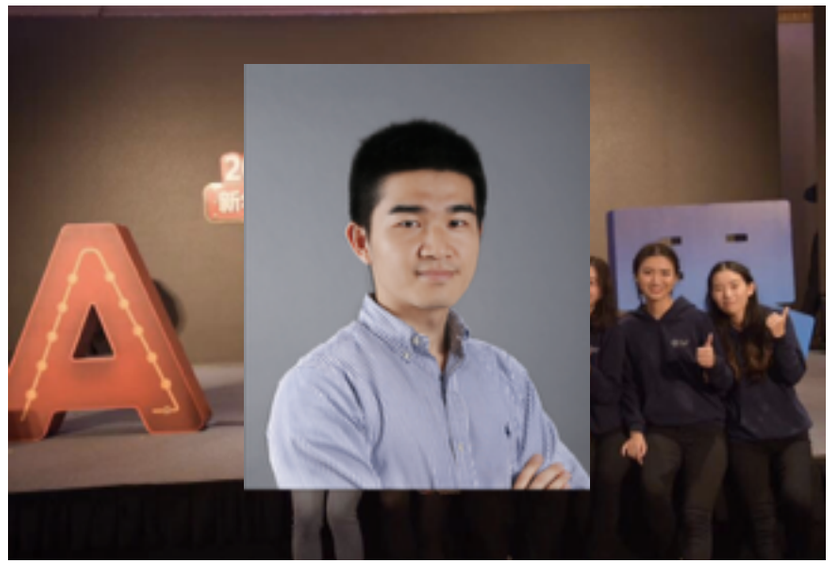
Life is not easy!

- A simple classification problem…

However…

Life is not easy!

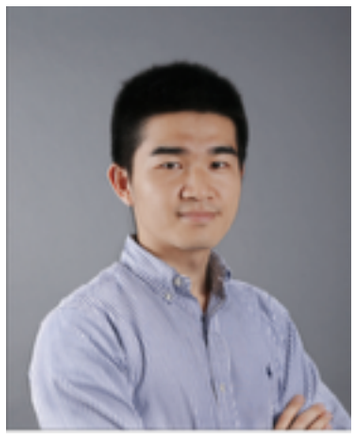
- A simple classification problem…



Chairman Intern Problem

Life is not easy!

- A simple classification problem…



Not Interview!

Interview!

Chairman Intern Problem

Life is not easy!

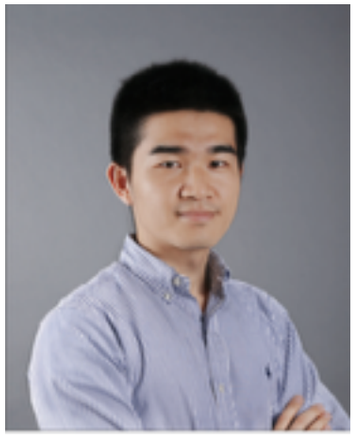
- A simple classification problem…

Feature	Value
Name	?
GPA	4.0
# Paper	10

Not Interview!

Feature	Value
Name	WZ
# Paper	100
Hobby	Fitness

Interview!



Chairman Intern Problem

Life is not easy!

- A simple classification problem…

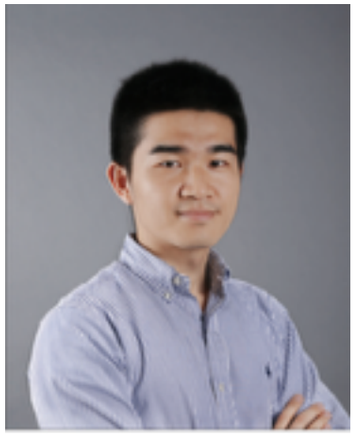
Feature	Value
Name	?
GPA	4.0
# Paper	1

Naive Bayes

Not Interview!

Feature	Value
Name	WZ
# Paper	100
Hobby	Fitness

Interview!

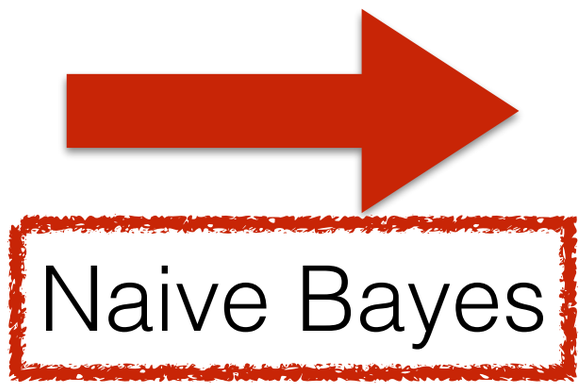


Chairman Intern Problem: Model I

Life is not easy!

- A simple classification problem…

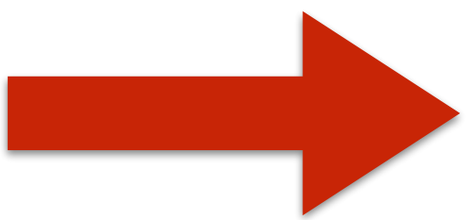
Feature	Value
# Papers	7



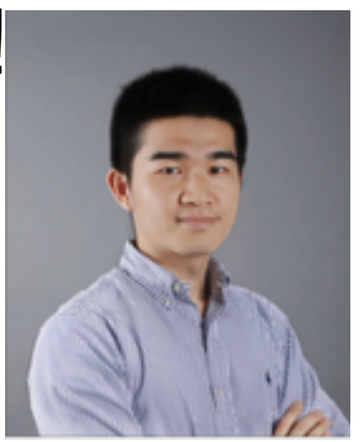
Naive Bayes

Interview!

Feature	Value
# Papers	1



Not Interview!



$$\mathbf{P}(\mathbf{y}, \mathbf{x_1})$$

Chairman Intern Problem: Model I

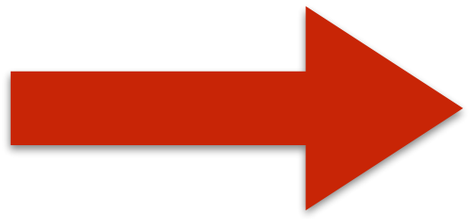
Life is not easy!

- A simple classification problem…

Feature	Value
Name	JY
# Papers	7
Genda	male

Interview!

Feature	Value
Name	MS
# Papers	1
Genda	female



Not Interview!

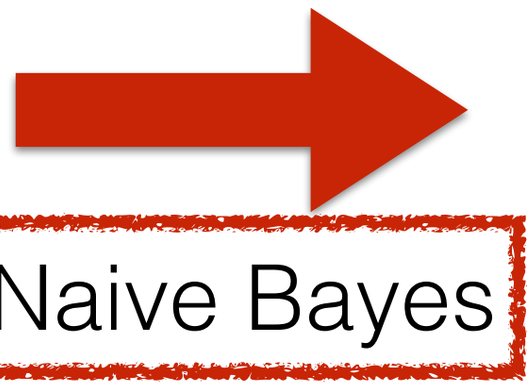
$$\mathbf{P}(\mathbf{y}, \mathbf{x_1})$$

Chairman Intern Problem: Model I

Life is not easy!

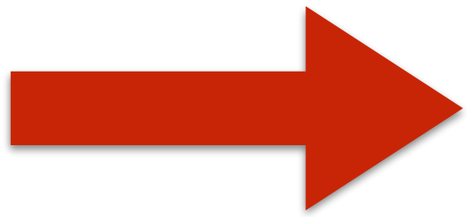
- A simple classification problem…

Feature	Value
Name	JY
# Papers	7
Genda	male



Interview!

Feature	Value
Name	MS
# Papers	1
Genda	female



Not Interview!

$$\mathbf{P}(\mathbf{y}, \mathbf{x_1})$$

Chairman Intern Problem: Model I

Life is not easy!

- A simple classification problem…

Feature	Value
Name	JY
# Papers	7
Genda	male

Naive Bayes

Feature	Value
Name	MS
# Papers	1
Genda	female

Interview!

Not Interview!

$$\mathbf{P}(\mathbf{y}, \mathbf{x_1})$$

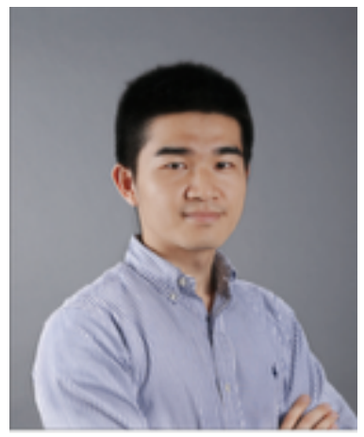
Chairman Intern Problem: Model 2

Life is not easy!

- A simple classification problem…

	Feature	Value
$\mathbf{x}_1$	Name	0/1
$\mathbf{x}_2$	Image	0/1
$\mathbf{x}_3$	Gender	0/1
$\mathbf{x}_4$	GPA	>3.0?
$\mathbf{x}_5$	# papers	>50?
	…	…

Interview ?



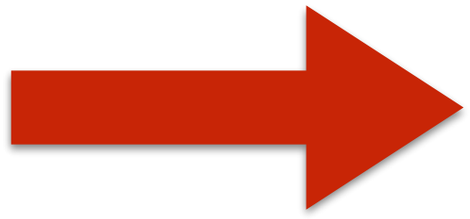
$$\mathbf{P(y, x_1, x_2, \ldots, x_n)}$$

Chairman Intern Problem: Model 2

Life is not easy!

- A simple classification problem…

	Feature	Value
$\mathbf{x}_1$	Name	MS
$\mathbf{x}_4$	# Paper	1
$\mathbf{x}_6$	Genda	female



Interview!

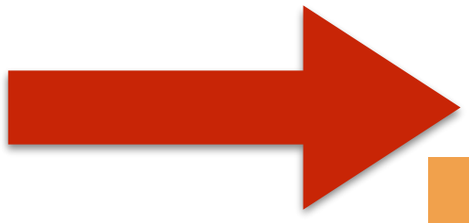
$$\mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_4}, \mathbf{x_6}) = \sum_{\mathbf{x_2}, \mathbf{x_3}, \mathbf{x_5}} \mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_6})$$

Chairman Intern Problem: Model 2

Life is not easy!

- A simple classification problem…

	Feature	Value
$\mathbf{x}_1$	Name	MS
$\mathbf{x}_4$	# Paper	1
$\mathbf{x}_6$	Genda	female



Interview!

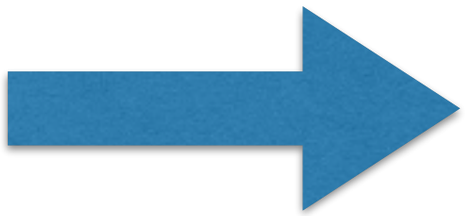
$$\mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_4}, \mathbf{x_6}) = \sum_{\mathbf{x_2}, \mathbf{x_3}, \mathbf{x_5}} \mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{x}_6)$$

Chairman Intern Problem: Model 3

Life can be harder!

- A simple classification problem…

	Feature	Value
$\mathbf{x}_1$	Name	0/1
$\mathbf{x}_2$	Image	**0~1**
$\mathbf{x}_3$	Gender	**0~1**
$\mathbf{x}_4$	GPA	#
$\mathbf{x}_5$	# papers	**0~200**
	…	…



Interview ?

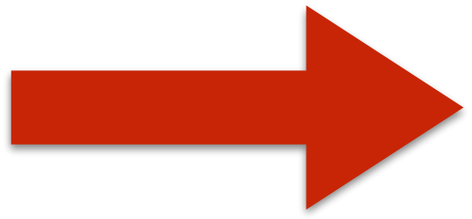
$$\mathbf{P(y, x_1, x_2, \ldots, x_n)}$$

Chairman Intern Problem: Model 3

Life can be harder!

- A simple classification problem…

	Feature	Value
$\mathbf{x}_1$	Name	MS
$\mathbf{x}_4$	# Paper	1
$\mathbf{x}_6$	Genda	female



Interview ?

$$\mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_4}, \mathbf{x_6}) = \int_{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5} \mathbf{P}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_6}) d\mathbf{x}_2 d\mathbf{x}_3 d\mathbf{x}_5$$

Latent Variable Models

- Missing Information
  - Expensive to label
  - Partially labeled
- Modeling Purpose
  - Capturing correlations of variables
  - Directly represent important factors.

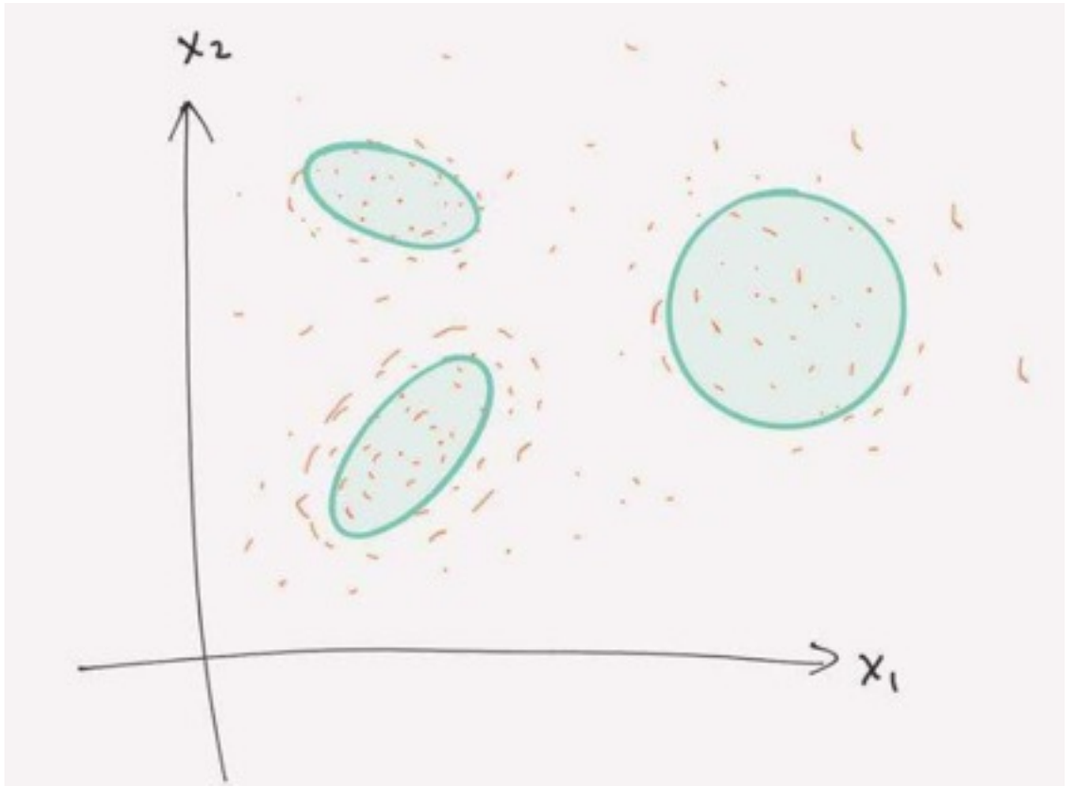
Learning and Inference in Latent Variable Graphical Models, Ph.D Thesis, 2016, Wei Ping, UCLA

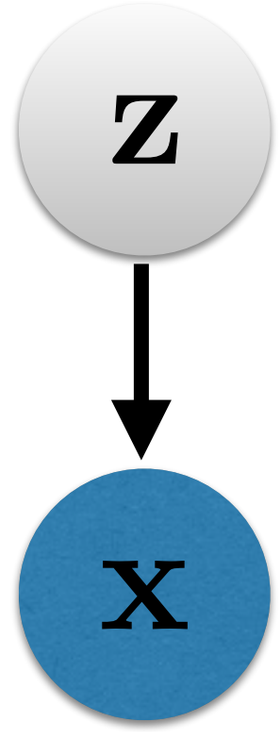
Latent Variable Modeling

- Latent variable models are everywhere in machine learning and natural language processing.
  - Mixture of Gaussian for Clustering
  - HMM for POS Induction
  - Latent Variable PCFG
  - Latent Tree Model for Classification

Latent Variable Modeling

- Unsupervised learning: Mixture of Gaussian

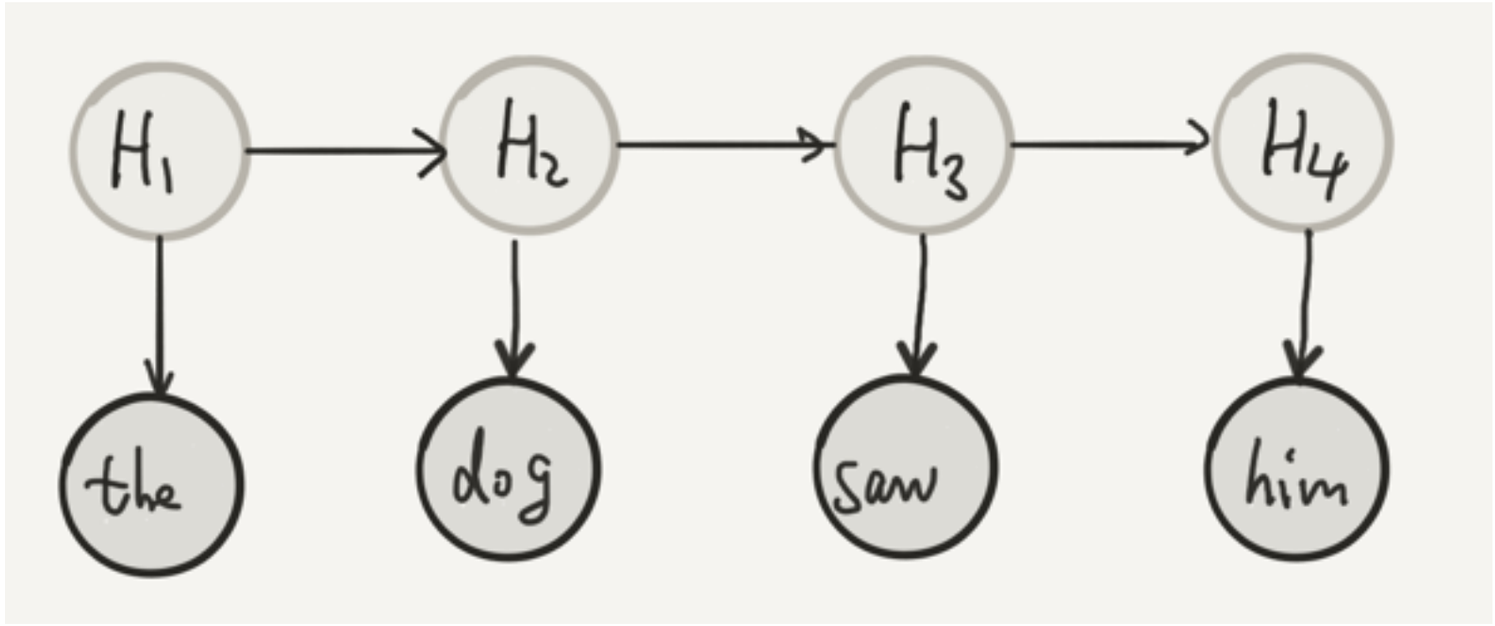


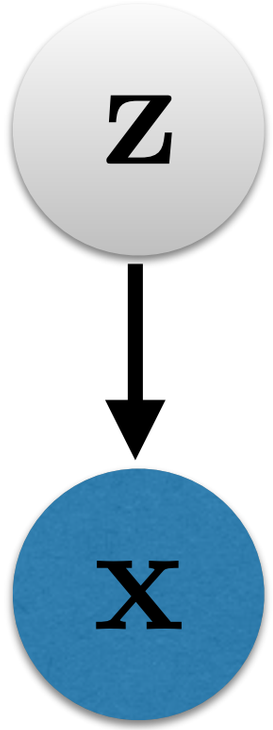


$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

Latent Variable Modeling

- Unsupervised learning: POS Induction

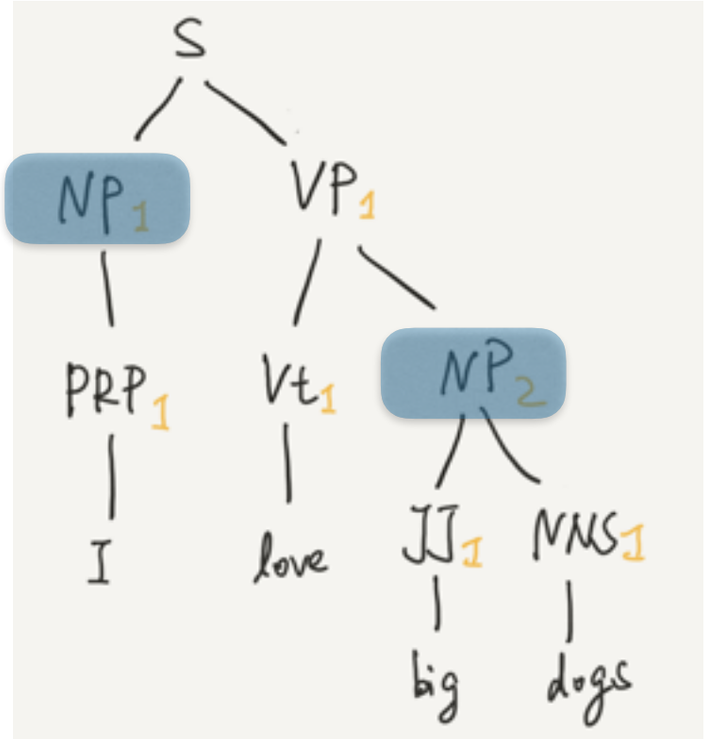




$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

Latent Variable Modeling

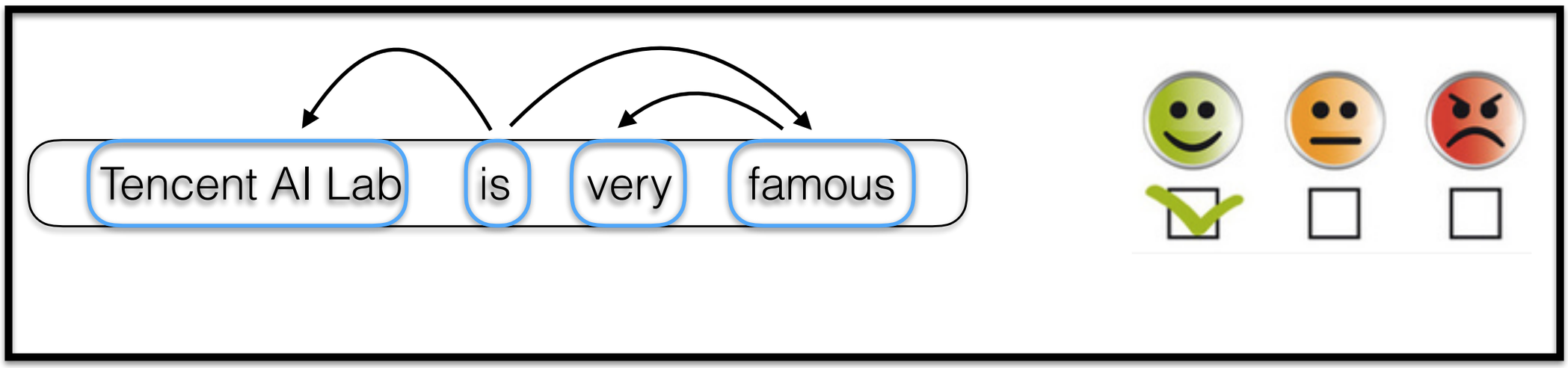
- Supervised learning: Latent Variable PCFG
  - latent states: sub type nonterminal
    - NP: NP(1)=subject NP(2)=object



Learning Accurate, Compact, and Interpretable Tree Annotation, ACL 2006

Latent Variable Modeling

- Supervised learning: Latent Variable Tree RNN
  - Syntax relation between words in a sentence



Generative Latent Variable Modeling

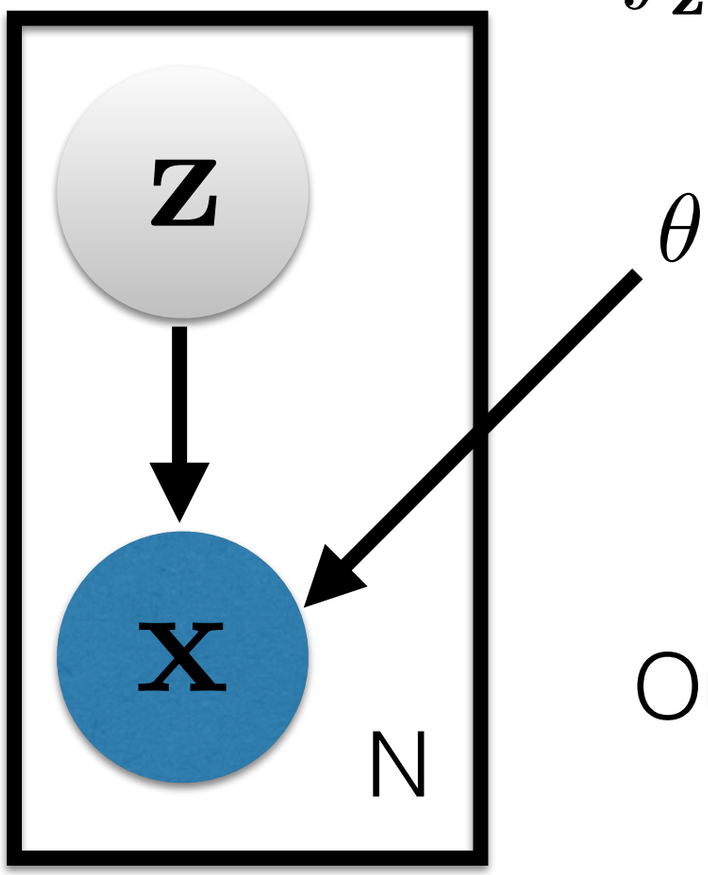
- Data generation process
  - Latent variable z: prior distribution is simple
  - A mapping from latent code z to observation x

$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} \mathbf{P}_\theta(\mathbf{x}|\mathbf{z}) \mathbf{P}(\mathbf{z}) d\mathbf{z}$$

Generative Latent Variable Modeling

- Data generation process
  - Latent variable z: prior distribution is simple
  - A mapping from latent code z to observation x

$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} \mathbf{P}_\theta(\mathbf{x}|\mathbf{z}) \mathbf{P}(\mathbf{z}) d\mathbf{z}$$



Example

$$\mathbf{P}(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{P}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mu, \sigma^2)$$

Output of a neural network

$$\mu = f_{\theta_1}(\mathbf{z})$$

$$\log \sigma^2 = f_{\theta_2}(\mathbf{z})$$

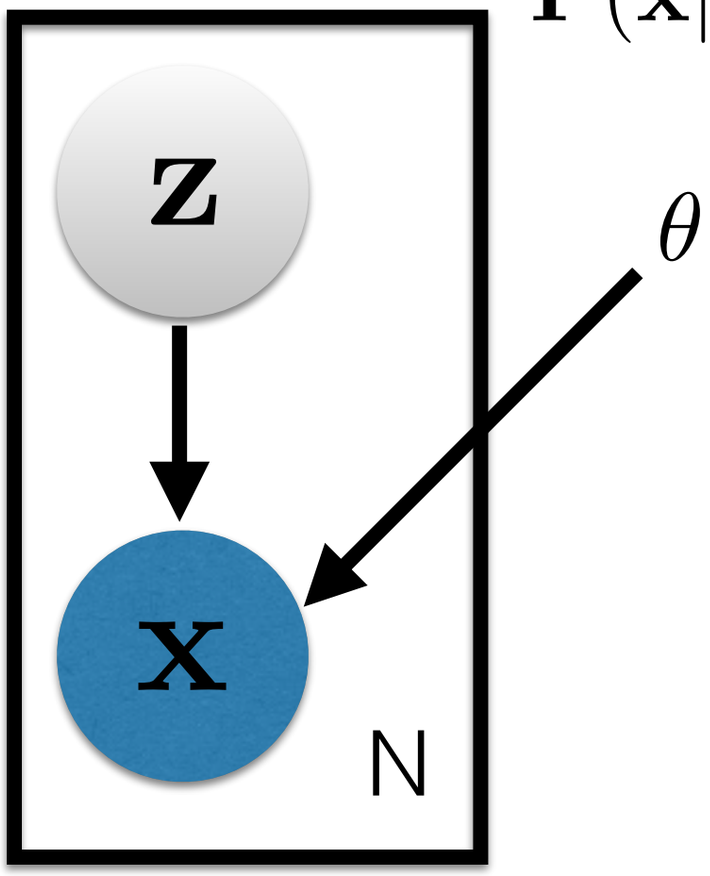
Generative Latent Variable Modeling

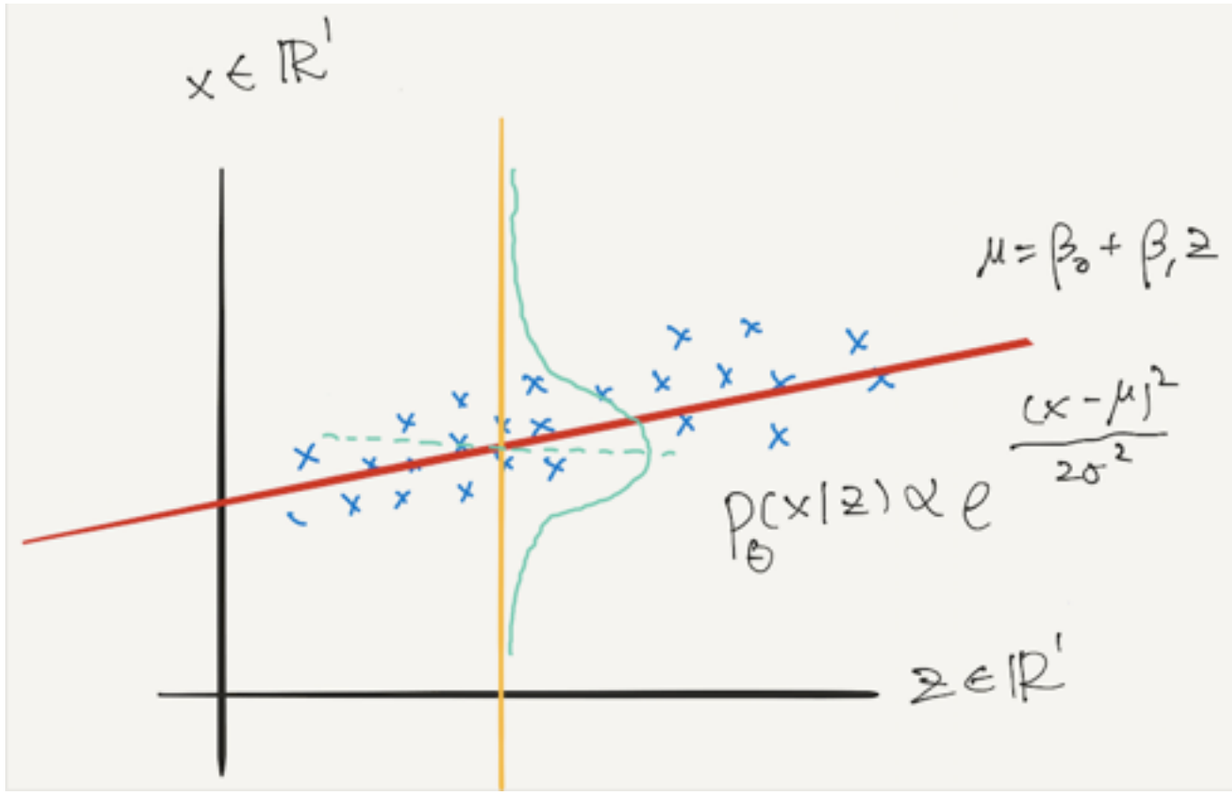
- 1D case: Linear Gaussian Model

$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} \mathbf{P}_\theta(\mathbf{x}|\mathbf{z}) \mathbf{P}(\mathbf{z}) d\mathbf{z}$$

$$\mathbf{P}(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{P}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mu, \sigma^2)$$





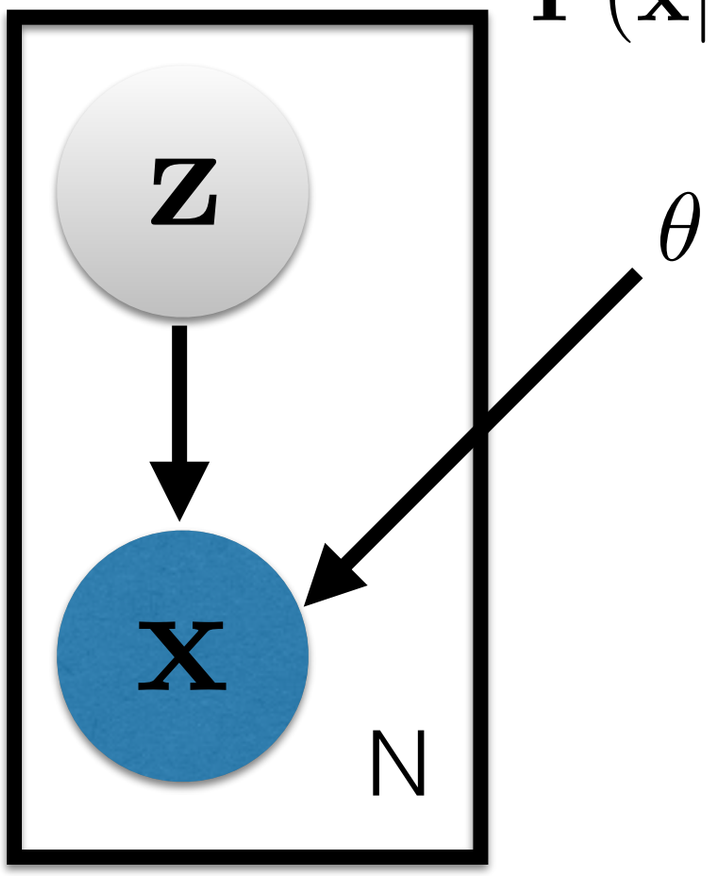
Generative Latent Variable Modeling

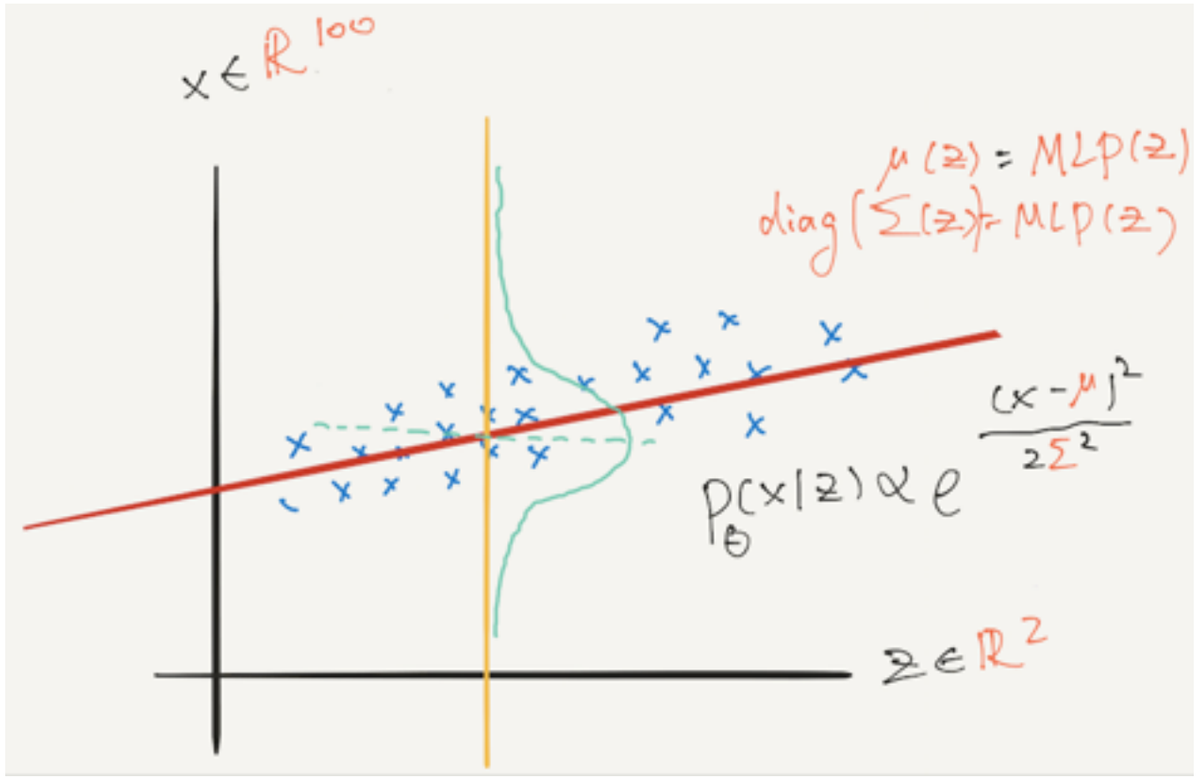
- 100D case

$$\mathbf{P}(\mathbf{x}) = \int_{\mathbf{z}} \mathbf{P}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} \mathbf{P}_\theta(\mathbf{x}|\mathbf{z}) \mathbf{P}(\mathbf{z}) d\mathbf{z}$$

$$\mathbf{P}(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$$

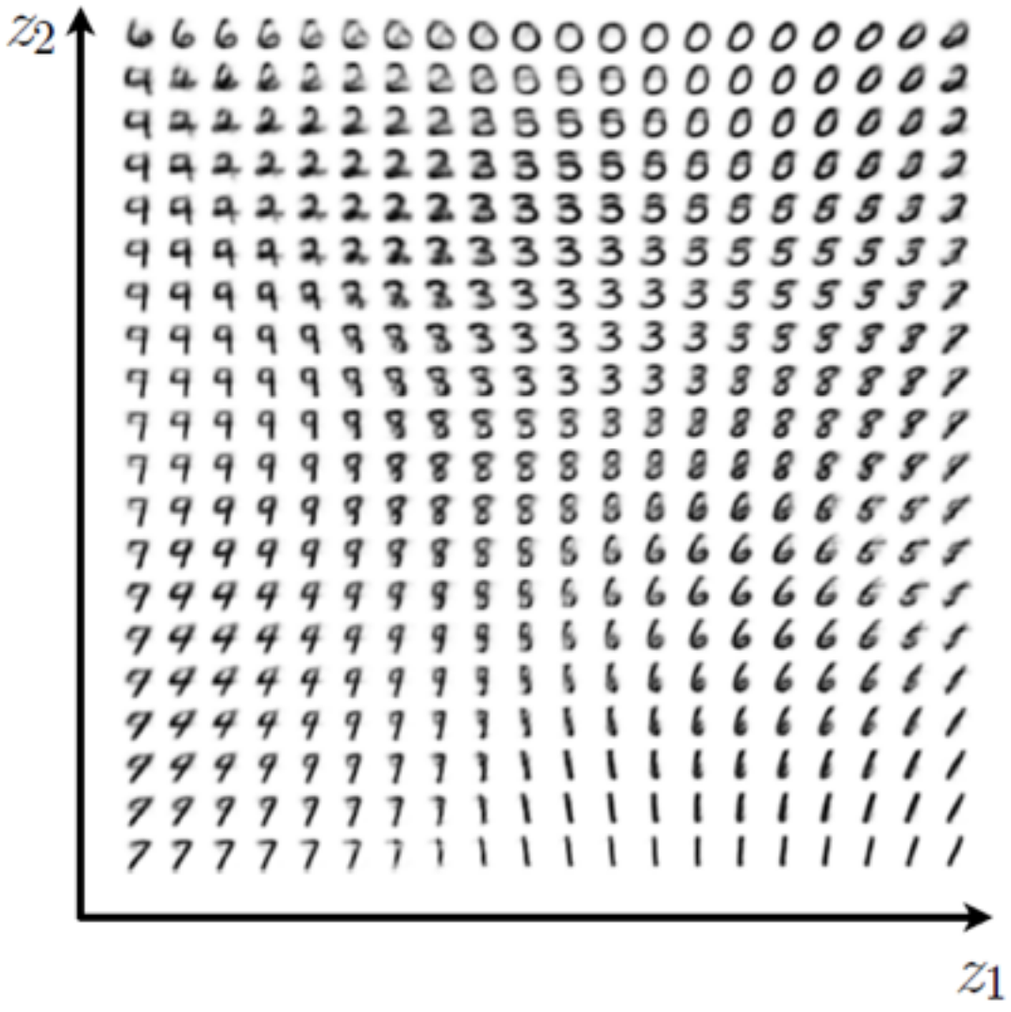
$$\mathbf{P}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mu, \sigma^2)$$



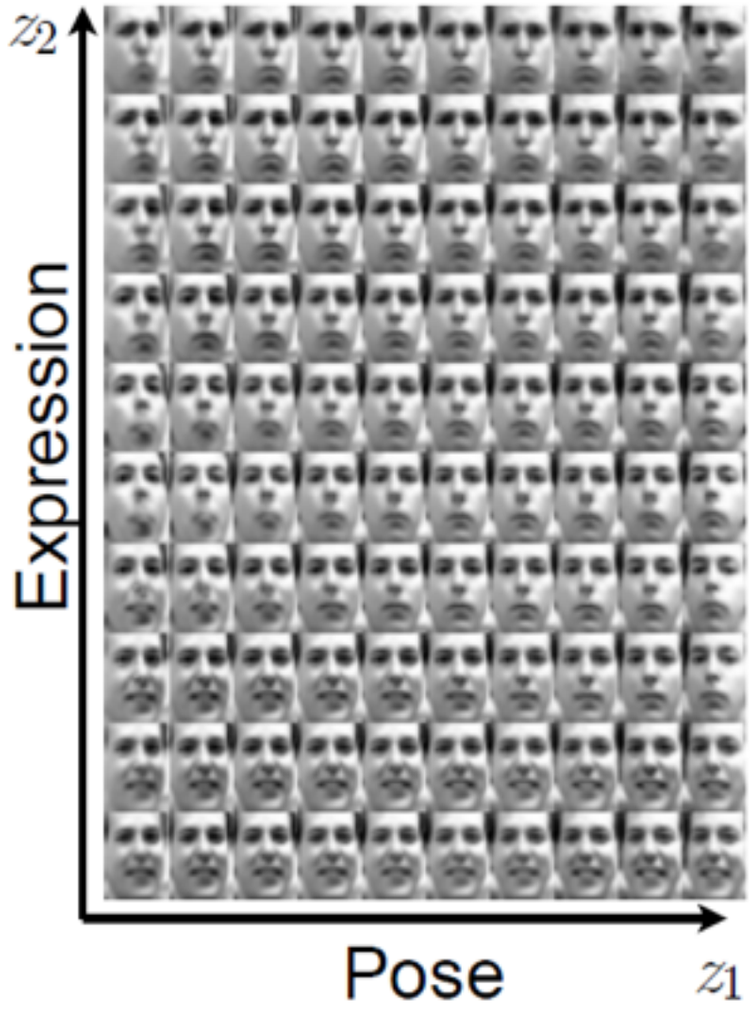


Latent Space of LVM

MNIST:



Frey Face dataset:



Slides from the talk at DLSS 2015 by Aaron Courville

How to Achieve it?

- Inference
  - Given an observation x, what is the probable z?
  - Compute the posterior P(z|x) (intractable)

How to Achieve it?

- Inference
  - Given an observation x, what is the probable z?
  - Compute the posterior P(z|x) (intractable)
- Learning
  - Given a set of training data
  - Estimate the parameters of P(x|z) (inefficient)
    - MCMC
    - Variational EM: mean field approximation
      - need to calculate some expectation terms exactly

Variational Autoencoder

- Inference
  - Introduce a parametric model Q(z|x) to approximate the true posterior Q(z|x)
  - Parameters of Q are learned with P
  - Use Q to perform inference
- Learning
  - Based on Maximum Likelihood Estimation (MLE).
  - Direct optimization is challenging: seeking alternative objective

VAE: Learning

- Learning
  - Based on Maximum Likelihood Estimation (MLE).
  - Direct optimization is challenging: seeking alternative objective

$$
\begin{aligned}
\log \mathbf{P}_\theta(\mathbf{x}) &= \log \int_{\mathbf{z}} \mathbf{P}_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\
&= \log \int_{\mathbf{z}} \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x}) \frac{\mathbf{P}_\theta(\mathbf{x}, \mathbf{z})}{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&\geqslant \int_{\mathbf{z}} \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x}) \log \frac{\mathbf{P}_\theta(\mathbf{x}, \mathbf{z})}{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\log \mathbf{P}_\theta(\mathbf{x}, \mathbf{z}) - \log \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})] = \mathcal{L}(\mathbf{x}; \theta, \phi)
\end{aligned}
$$

VAE: Learning

- VAE objective:
  - Q(z|x) is parametrised by another neural net
  - Interpreting VAE objective

$$\begin{aligned}\mathcal{L}(\mathbf{x};\theta,\phi) &= \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\log \mathbf{P}_\theta(\mathbf{x},\mathbf{z}) - \log \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})] \\ &= \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\log \mathbf{P}_\theta(\mathbf{x}|\mathbf{z}) + \log \mathbf{P}_\theta(\mathbf{z}) - \log \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})] \\ &= -\mathbf{KL}(\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})||\mathbf{P}_\theta(\mathbf{z})) + \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\log \mathbf{P}_\theta(\mathbf{x}|\mathbf{z})]\end{aligned}$$

regularisation reconstruction

VAE: Optimization Method I

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\log \mathbf{P}_\theta(\mathbf{x},\mathbf{z}) - \log \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})]$$
$$= \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[f_{\theta,\phi}(\mathbf{x},\mathbf{z})]$$

- Stochastic gradient descent

$$\nabla_\phi \mathcal{L}(\mathbf{x};\theta,\phi) = \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\phi f_{\theta,\phi}(\mathbf{x},\mathbf{z}) + f_{\theta,\phi}(\mathbf{x},\mathbf{z})\nabla_\phi \log \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})]$$
$$= \frac{1}{L}\sum_{i=1}^{L}[\nabla_\phi f_{\theta,\phi}(\mathbf{x},\mathbf{z}^{(i)}) + f_{\theta,\phi}(\mathbf{x},\mathbf{z}^{(i)})\nabla_\phi \log \mathbf{Q}_\phi(\mathbf{z}^{(i)}|\mathbf{x})]$$

$$\mathbf{z}^{(i)} \sim \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})$$

High variance for gradient estimation!

VAE: Optimization Method 2

- Reparameterization trick
  - Reparameterize z~Q(z|x) with a differentiable transformation of an auxiliary noise variable

$$\mathbf{z}^{(i)} \sim \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x}) \quad \Rightarrow \quad \mathbf{z}^{(i)} = g_\phi(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim q(\epsilon)$$

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[f_{\theta,\phi}(\mathbf{x},\mathbf{z})] \approx \frac{1}{L}\sum_{i=1}^{L} f_{\theta,\phi}(\mathbf{x}, g_\phi(\epsilon^{(i)}, \mathbf{x}))$$
$$\epsilon^{(i)} \sim q(\epsilon)$$

VAE: Optimization Method 2

- Reparameterization trick
  - Reparameterize z~Q(z|x) with a differentiable transformation of an auxiliary noise variable

$$\mathbf{z}^{(i)} \sim \mathbf{Q}_\phi(\mathbf{z}|\mathbf{x}) \quad \Rightarrow \quad \mathbf{z}^{(i)} = g_\phi(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim q(\epsilon)$$

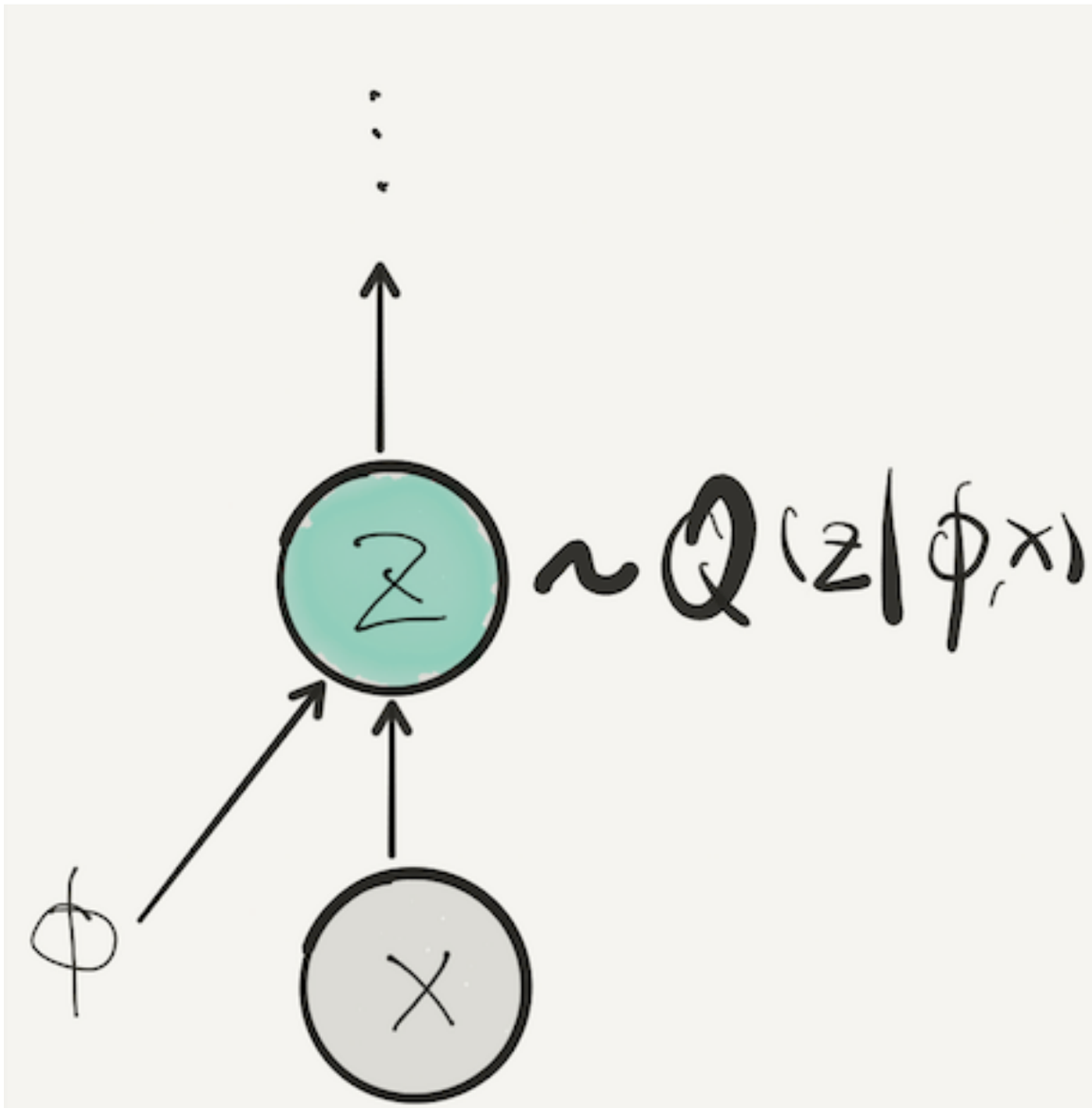
$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbf{E}_{\mathbf{Q}_\phi(\mathbf{z}|\mathbf{x})}[f_{\theta,\phi}(\mathbf{x},\mathbf{z})] \approx \frac{1}{L}\sum_{i=1}^{L} f_{\theta,\phi}(\mathbf{x}, g_\phi(\epsilon^{(i)}, \mathbf{x})) \qquad \epsilon^{(i)} \sim q(\epsilon)$$

- Example
  - Univariate Gaussian: $\mathbf{z} \sim \mathbf{Q}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$

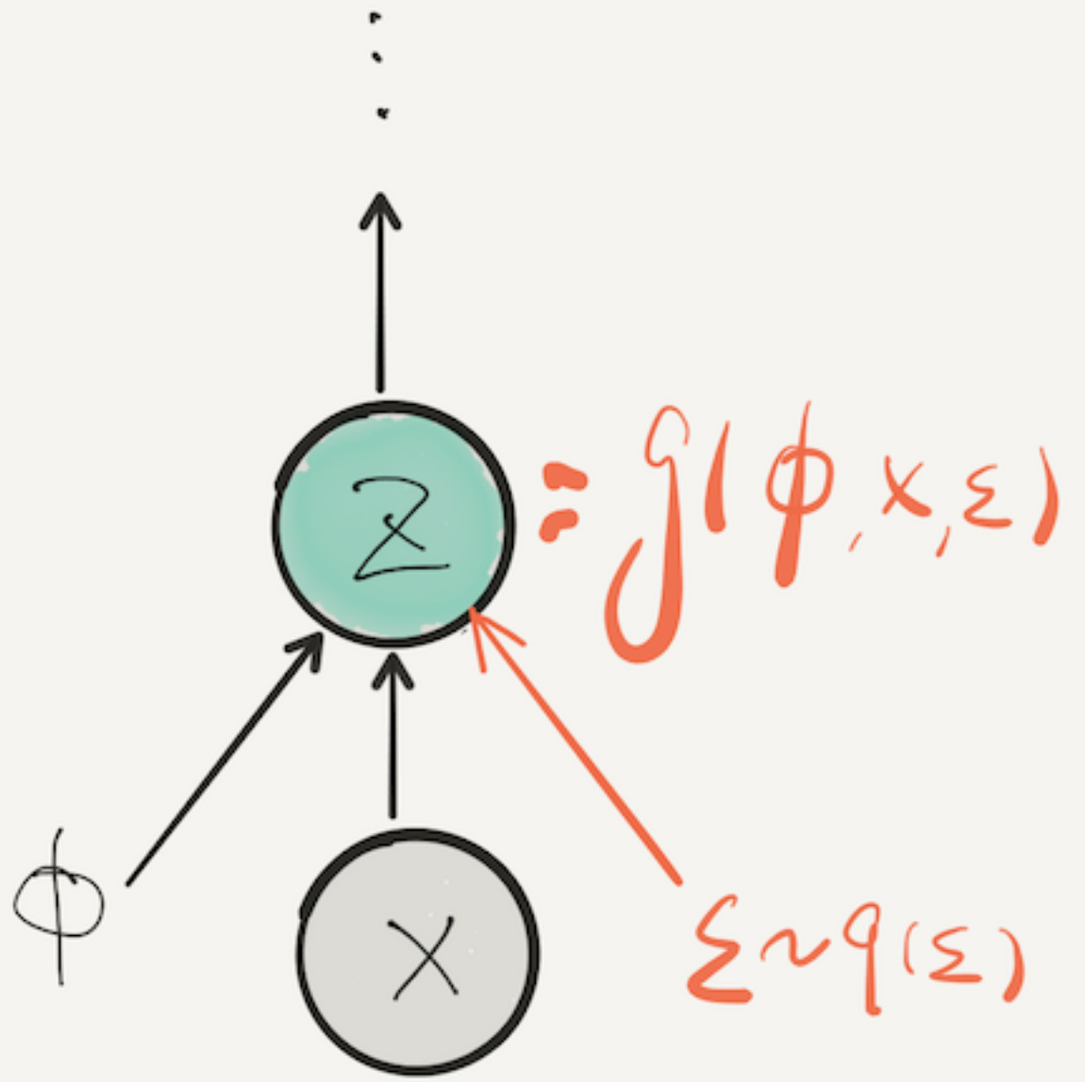
$$\mathbf{z} = \mu + \sigma\epsilon \qquad \epsilon \sim \mathcal{N}(0,1)$$

$$\mathbf{E}_{\mathbf{z}\sim\mathbf{Q}}[f(\mathbf{z})] = \mathbf{E}_\epsilon[f(\mu+\sigma\epsilon)] \approx \sum_{i=1}^{L} f(\mu + \sigma\epsilon^{(i)})$$

VAE: Network Interpretation

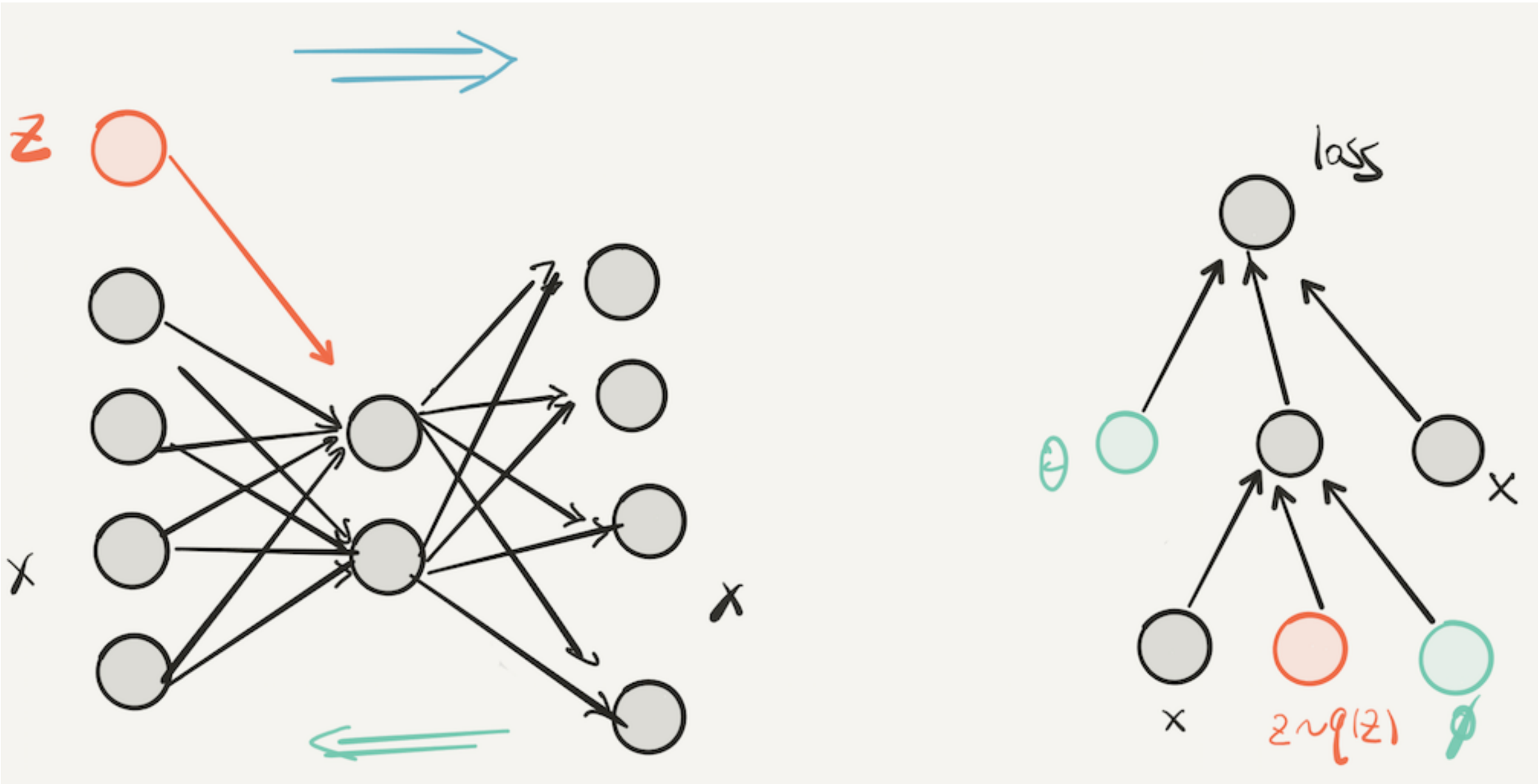


original form



reparameterised form

VAE: End-to-end Training with BP



VAE: Experiments

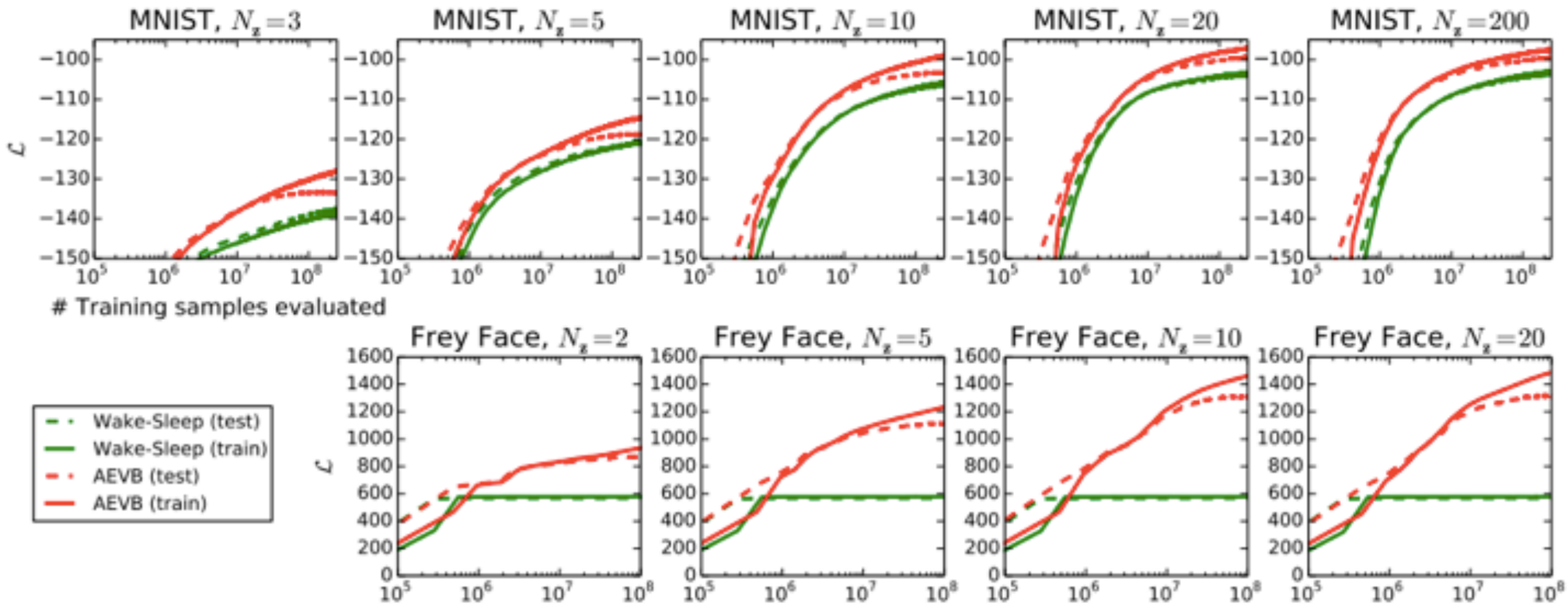
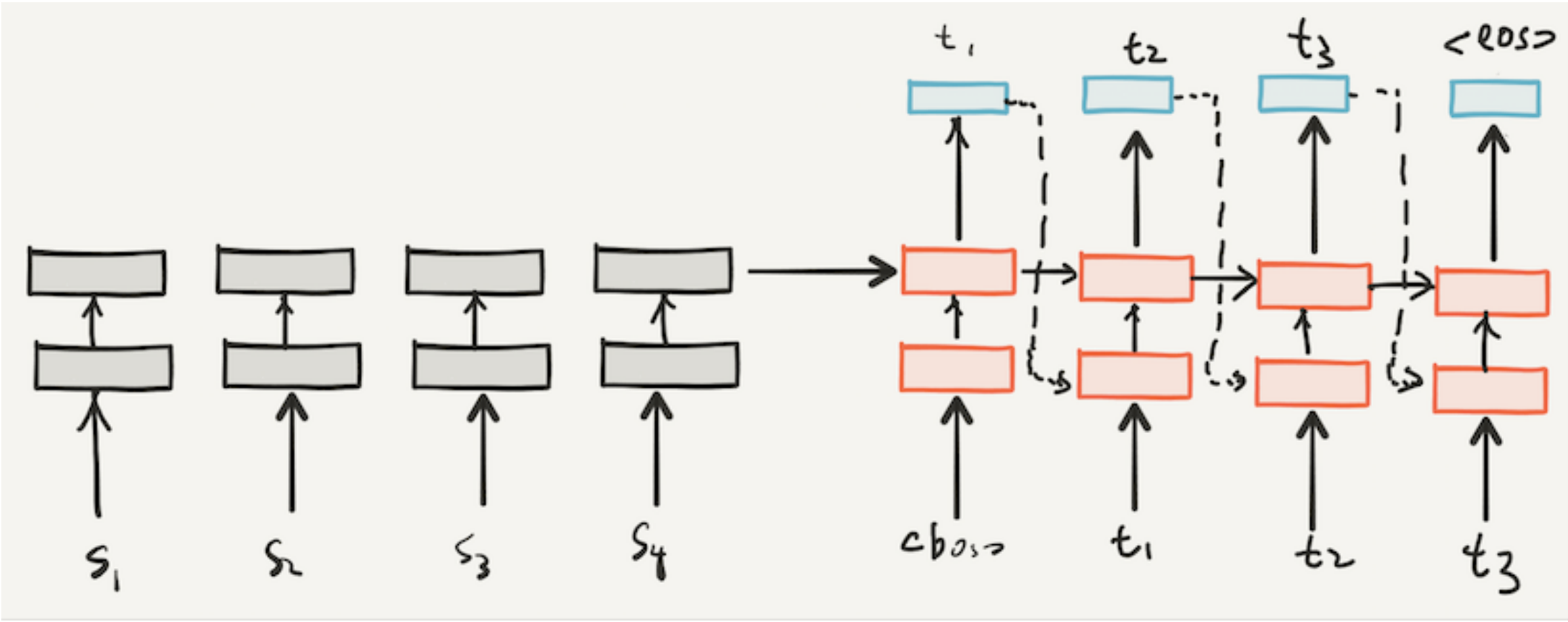


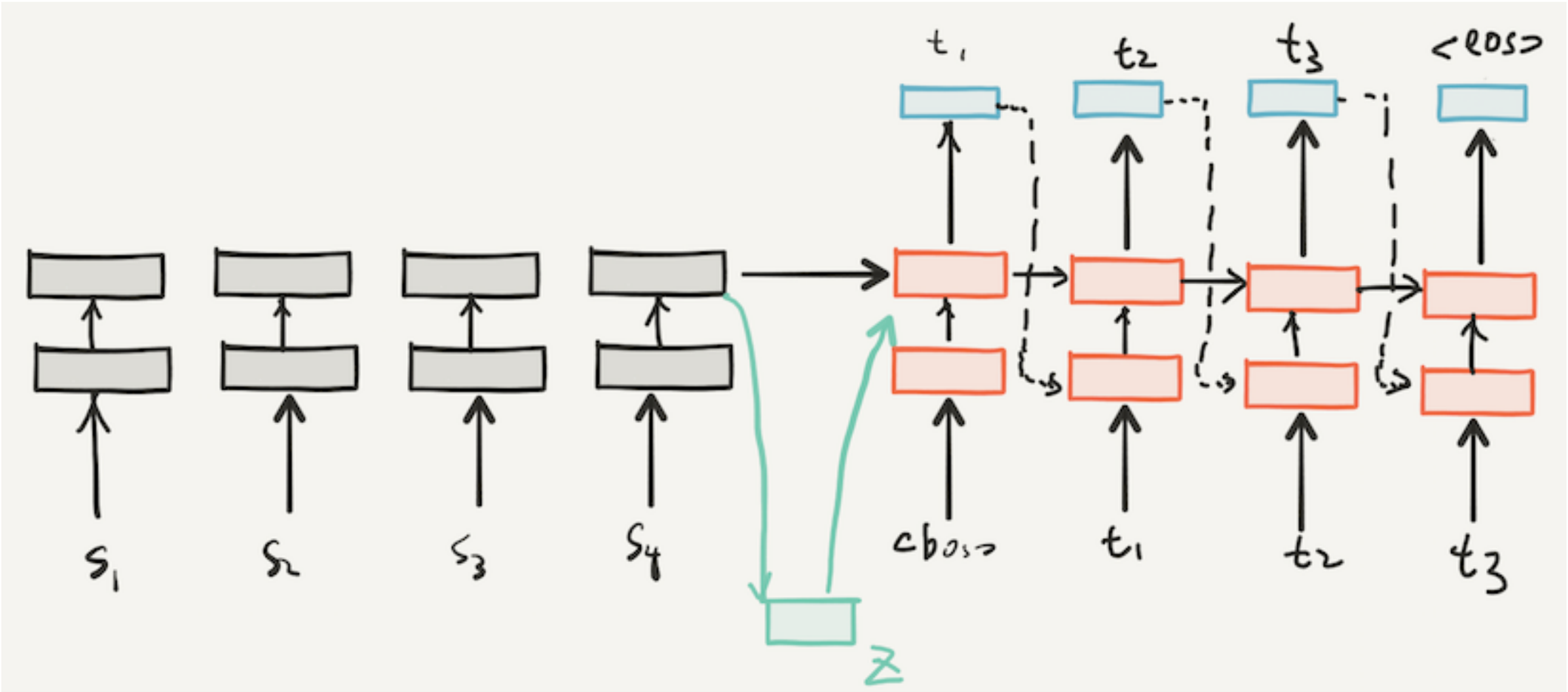
Figure 2: Comparison of our AEVB method to the wake-sleep algorithm, in terms of optimizing the lower bound, for different dimensionality of latent space ($N_{\mathbf{z}}$). Our method converged considerably faster and reached a better solution in all experiments. Interestingly enough, more latent variables does not result in more overfitting, which is explained by the regularizing effect of the lower bound. Vertical axis: the estimated average variational lower bound per datapoint. The estimator variance was small (< 1) and omitted. Horizontal axis: amount of training points evaluated. Computation took around 20-40 minutes per million training samples with a Intel Xeon CPU running at an effective 40 GFLOPS.

Auto-Encoding Variational Bayes, ICLR 2014

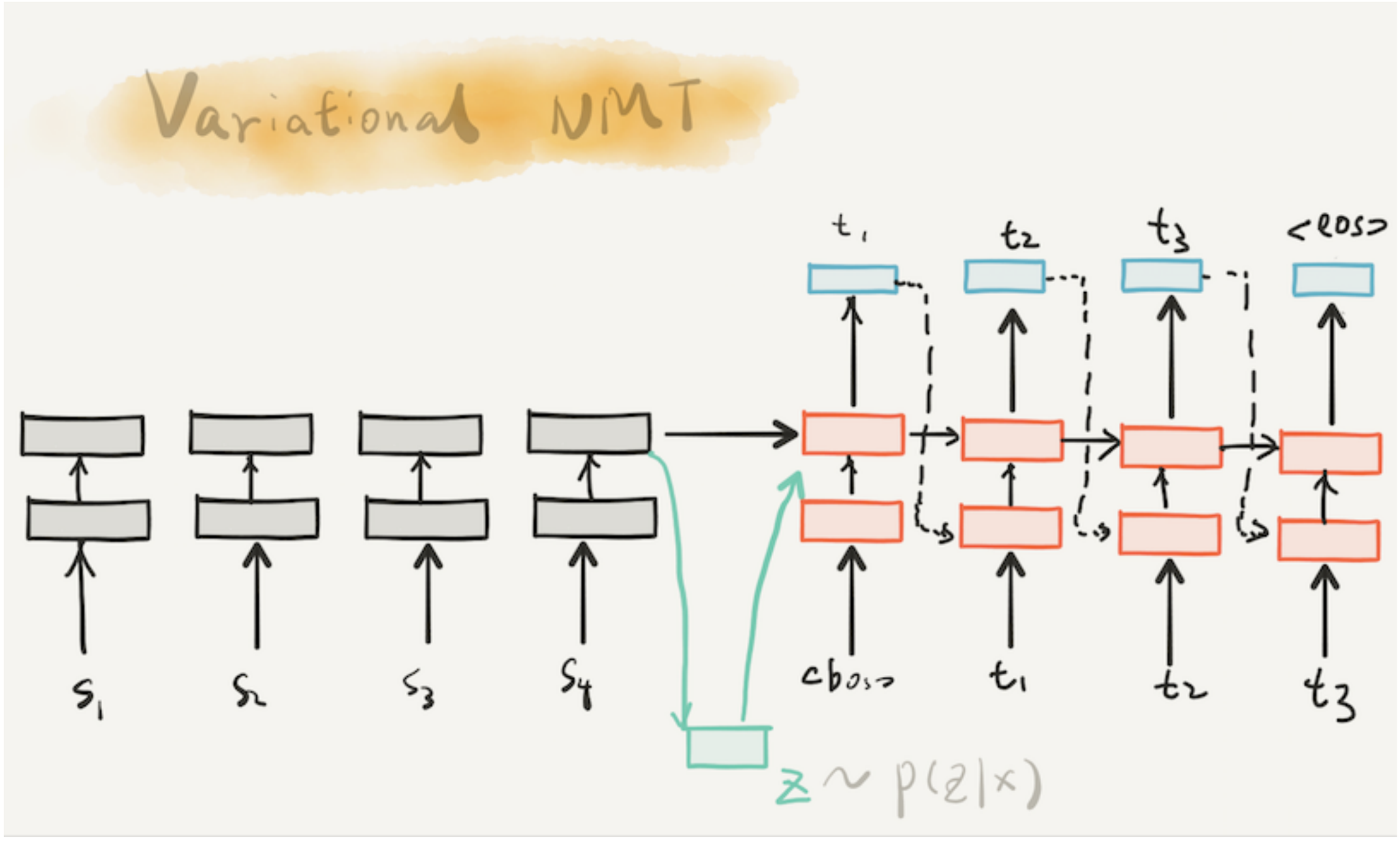
How to Write a paper on VAE-Seq?



How to Write a paper on VAE-Seq?

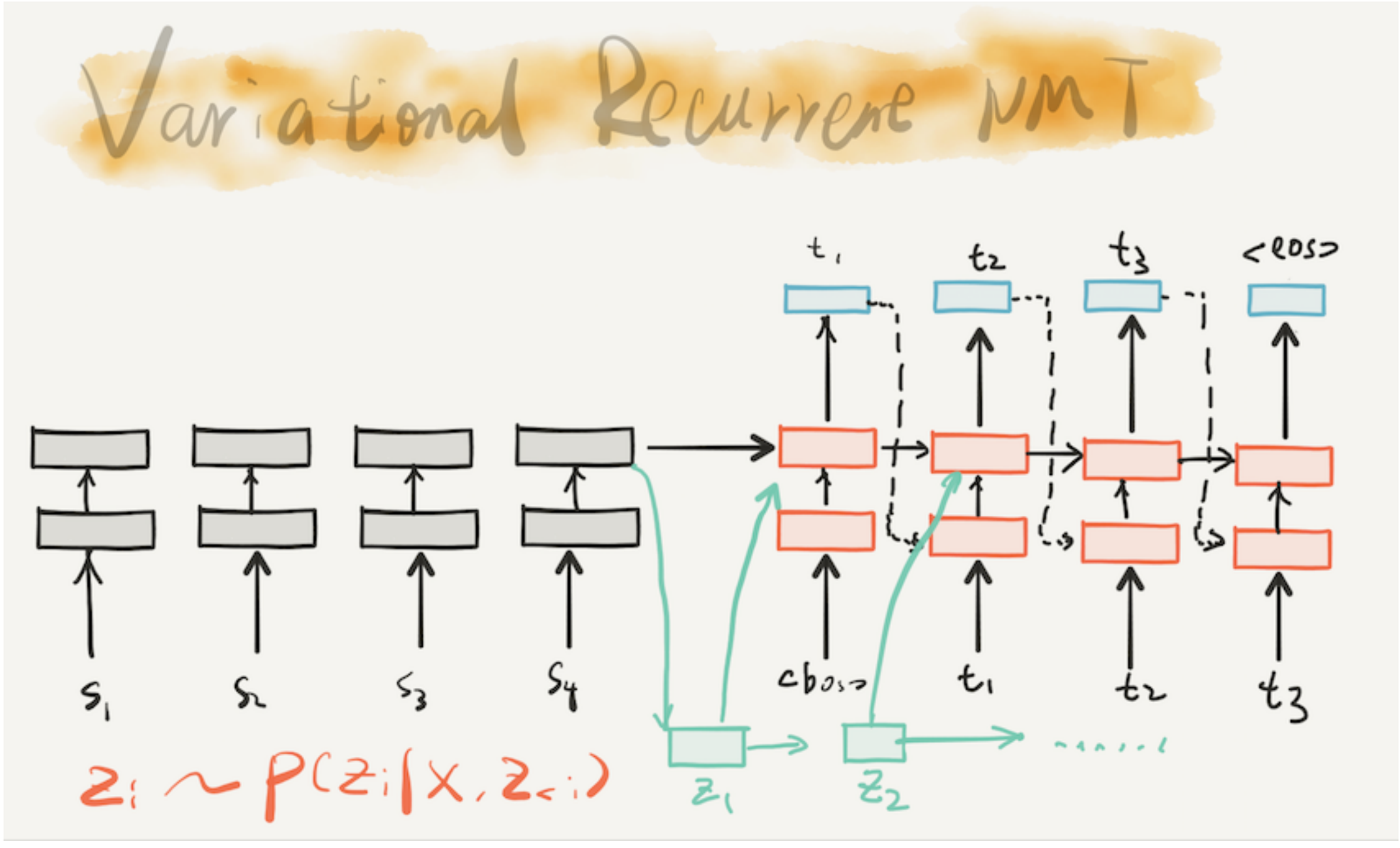


How to Write a paper on VAE-Seq?



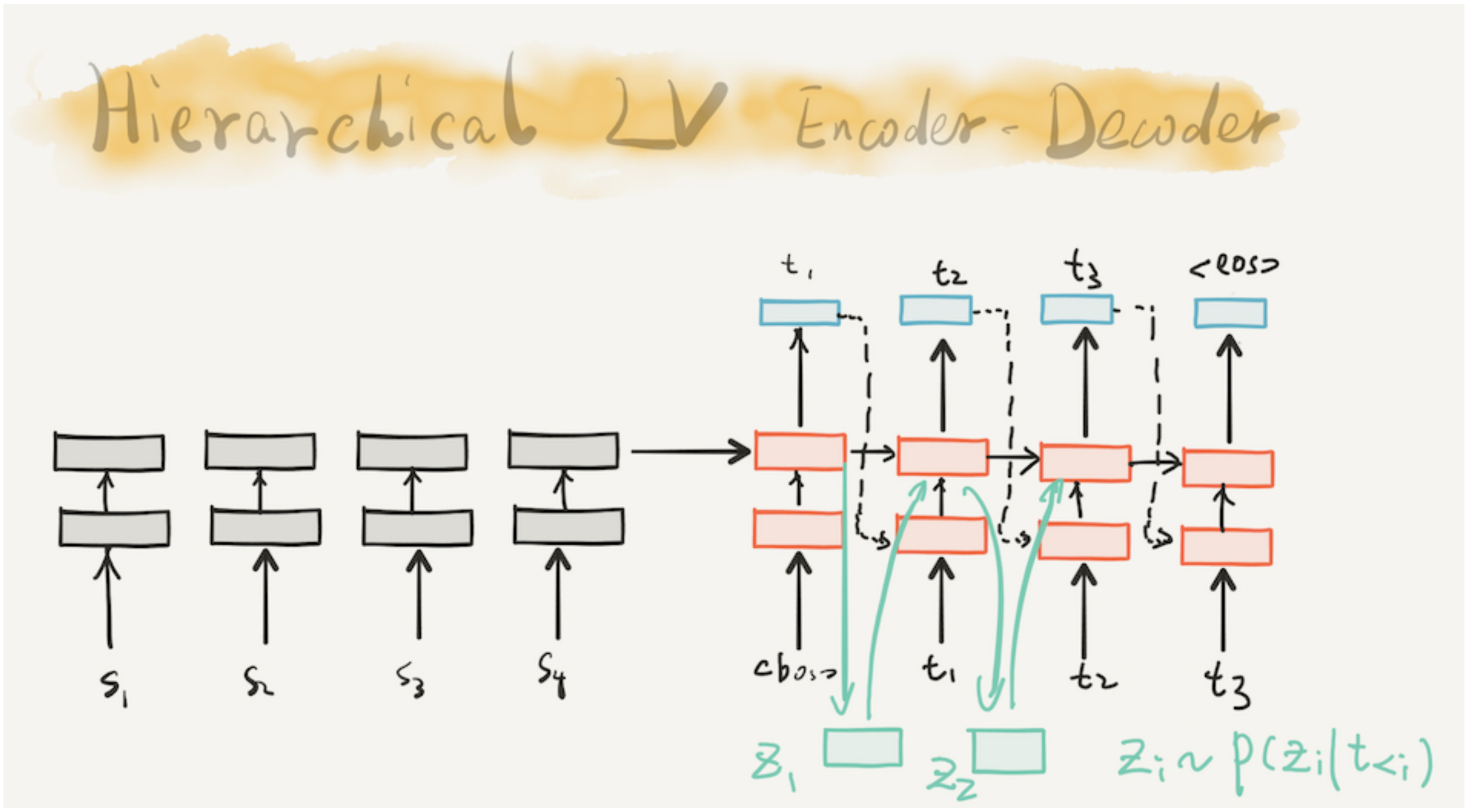
Variational Neural Machine Translation, EMNLP 2016

How to Write a paper on VAE-Seq?



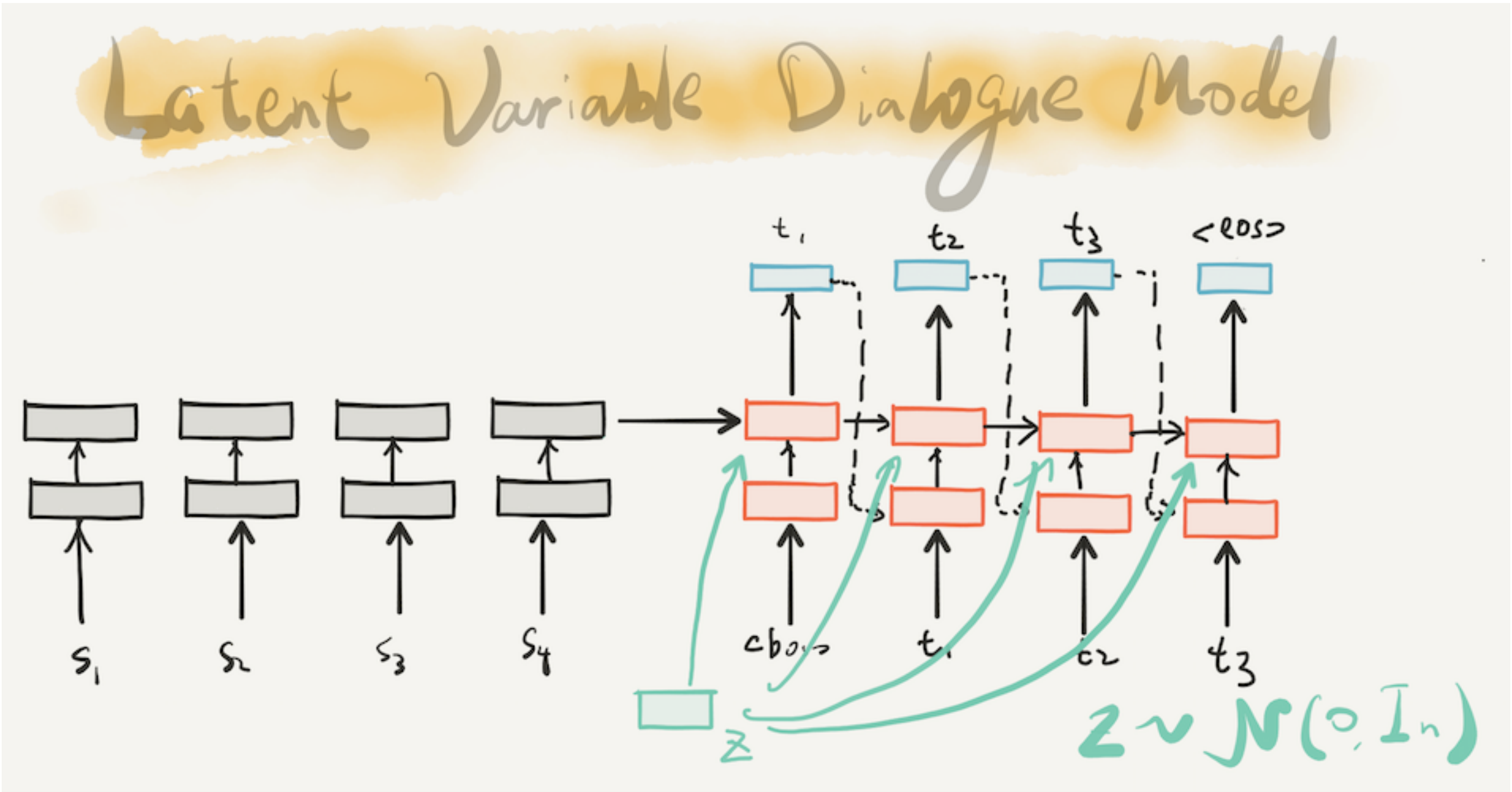
Variational Recurrent Neural Machine Translation, AAAI 2018

How to Write a paper on VAE-Seq?



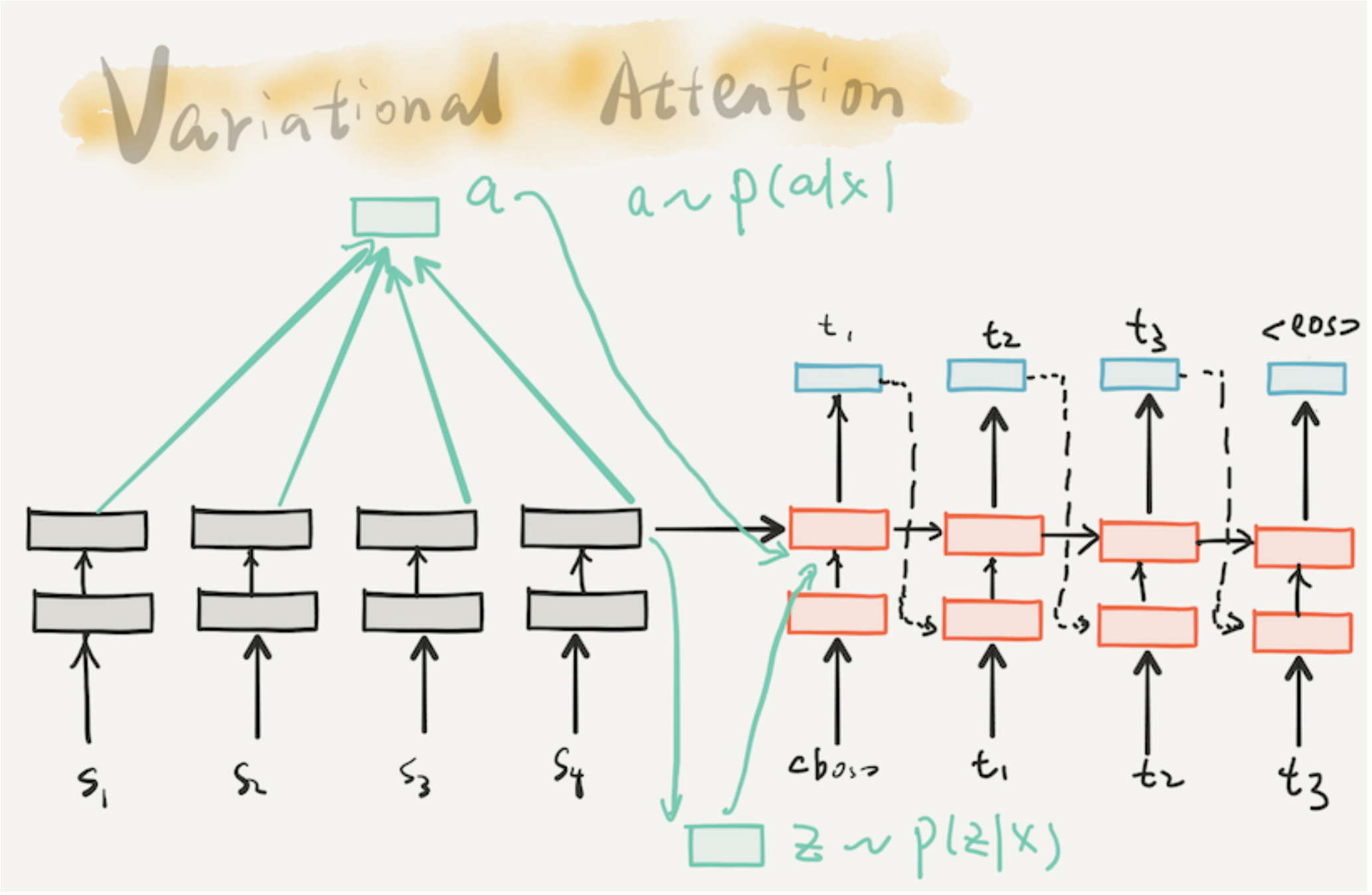
Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues, AAAI 2017

How to Write a paper on VAE-Seq?



Latent Variable Dialogue Models and their Diversity, EACL 2017

How to Write a paper on VAE-Seq?



Variational Attention for Sequence-to-Sequence Models, 2018

Take Home Message

- Chairman Intern Problem

Take Home Message

- Chairman Intern Problem
- Latent variable modeling is an importance task
  - both unsupervised and supervised learning

Take Home Message

- Chairman Intern Problem
- Latent variable modeling is an importance task
  - both unsupervised and supervised learning
- VAE is a framework to do LVM

Take Home Message

- Chairman Intern Problem
- Latent variable modeling is an importance task
  - both unsupervised and supervised learning
- VAE is a framework to do LVM
- VAE can be used in both continuous and discrete cases

Take Home Message

- Chairman Intern Problem
- Latent variable modeling is an importance task
  - both unsupervised and supervised learning
- VAE is a framework to do LVM
- VAE can be used in both continuous and discrete cases
- How to write a paper on VAE
  - Design deterministic nodes in the computational graph
  - Randomise the deterministic variable with stochastic.
    - Typically, a gaussian!!!

Take Home Message

- Chairman Intern Problem
- Latent variable modeling is an importance task
  - both unsupervised and supervised learning
- VAE is a framework to do LVM
- VAE can be used in both continuous and discrete cases
- How to write a paper on VAE
  - Design deterministic nodes in the computational graph
  - Randomise the deterministic variable with stochastic.
    - Typically, a gaussian!!!

Thanks!