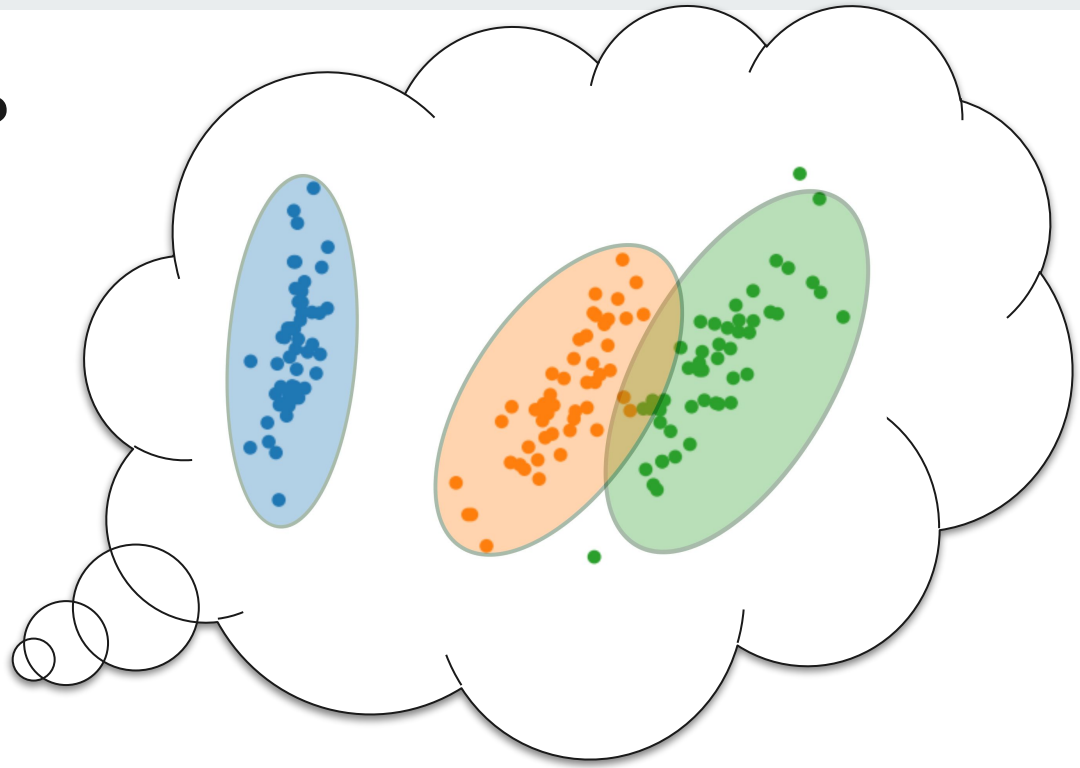
Bayesian Structured Representation Learning

Sharad Vikram, UCSD

What is structure?

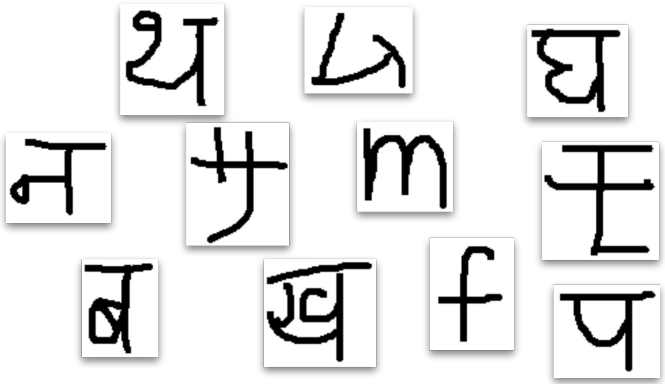
Fisher Iris

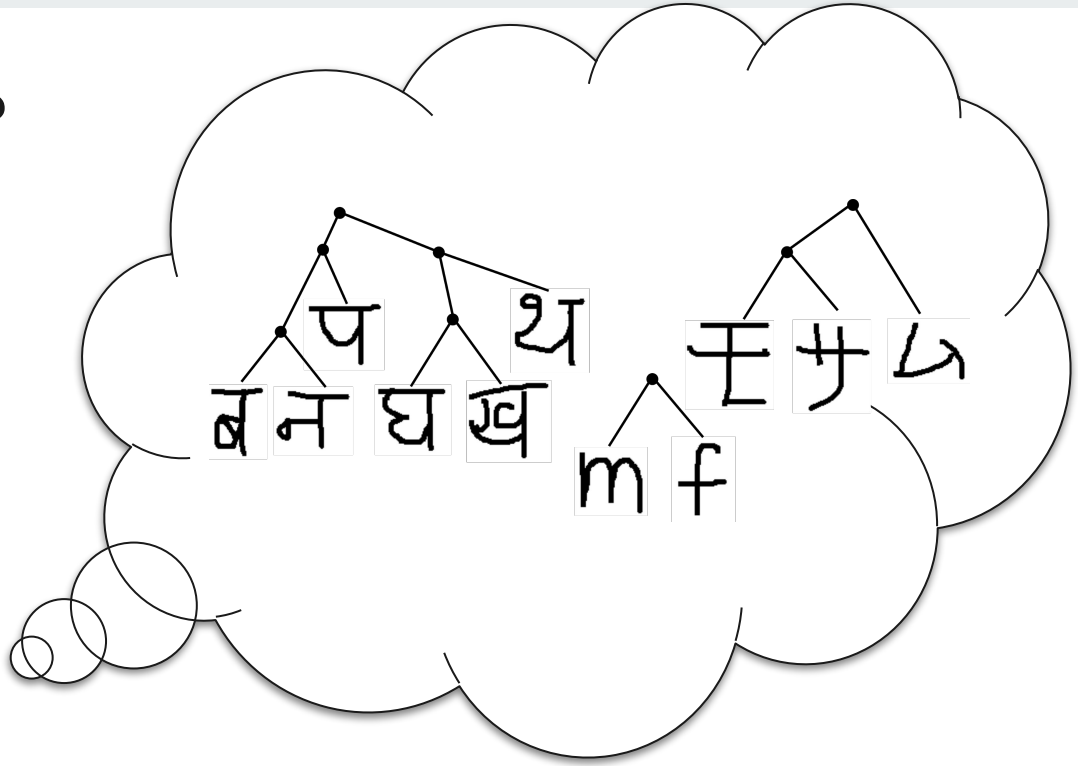
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1		1.4	0.2	0.0
1	4.9		1.4	0.2	0.0
2	4.7		1.3	0.2	0.0
3	4.6		1.5	0.2	0.0
4	5.0		1.4	0.2	0.0
5	5.4		1.7	0.4	0.0
6	4.6		1.4	0.3	0.0
7	5.0	3.4	1.5	0.2	0.0
8					0.0
9					0.0
10					0.0
11					0.0
12					0.0
13					0.0



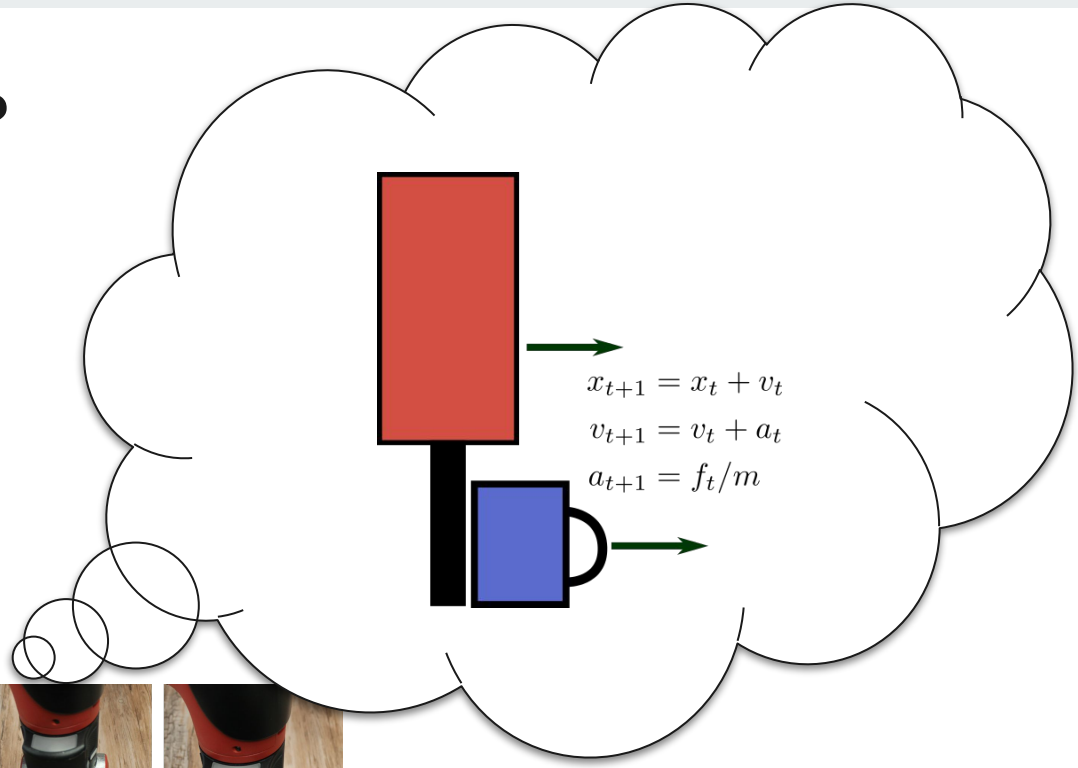
[1]

What is structure?





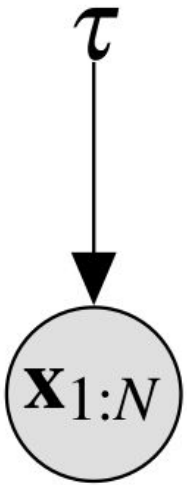
What is structure?



What is structure learning?

Organizing data in a way that is useful!

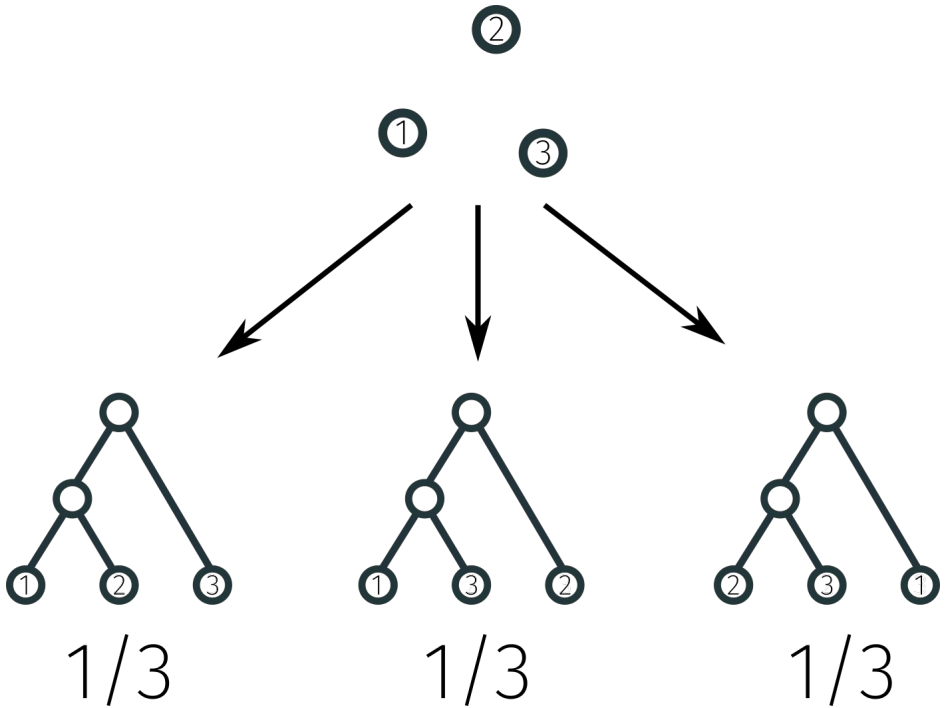
What is structure learning?



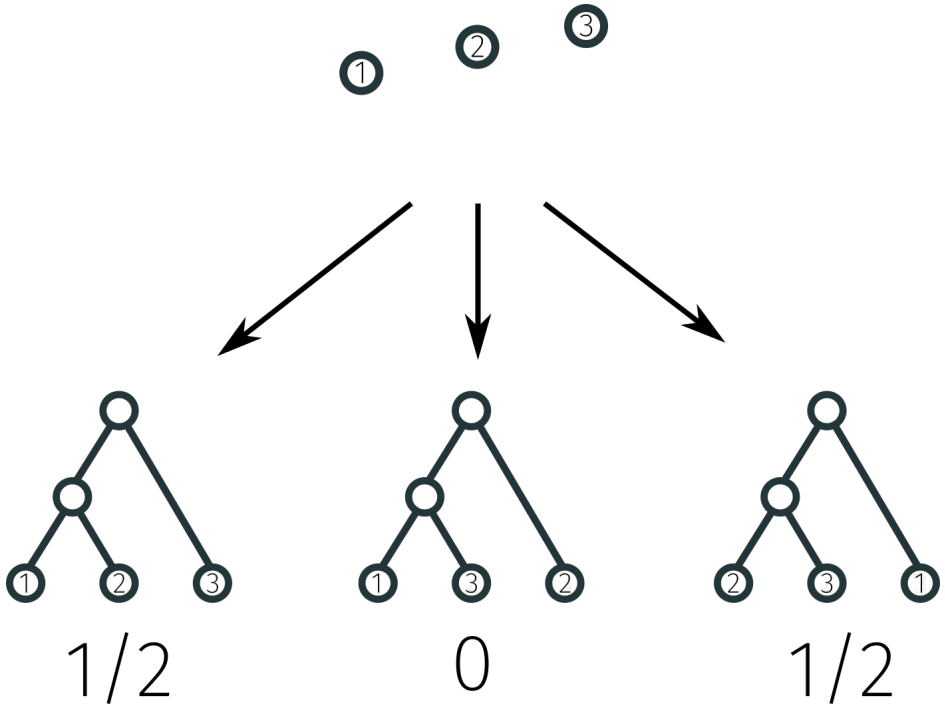
Goal: obtain $\tau^* = \arg\max_{\tau} p(\mathbf{x}_{1:N} \mid \tau)$

Problem: ambiguity in structure

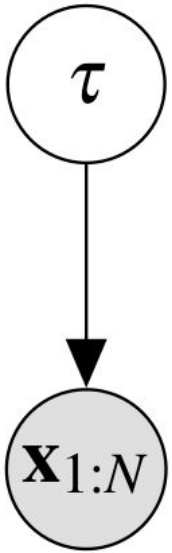
Problem: ambiguity in structure



Problem: ambiguity in structure

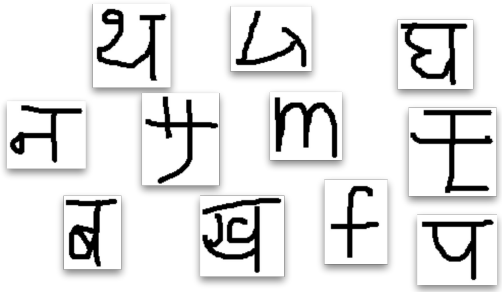


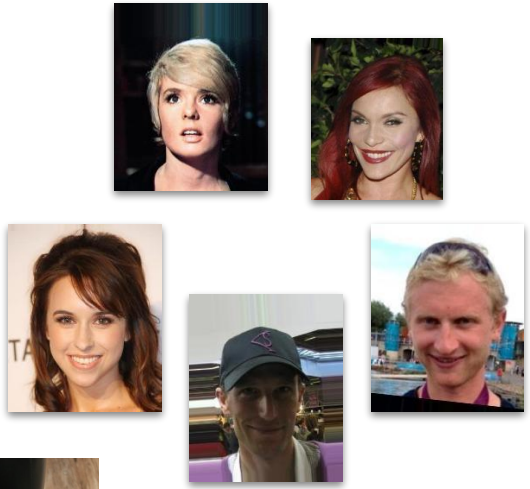
Bayesian structure learning



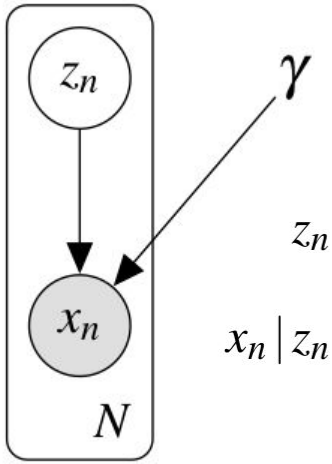
Goal: obtain $p(\tau \mid \mathbf{x}_{1:N})$

Analyzing complex, high-dimensional data





Variational autoencoder



$$z_n \sim N(0, I)$$

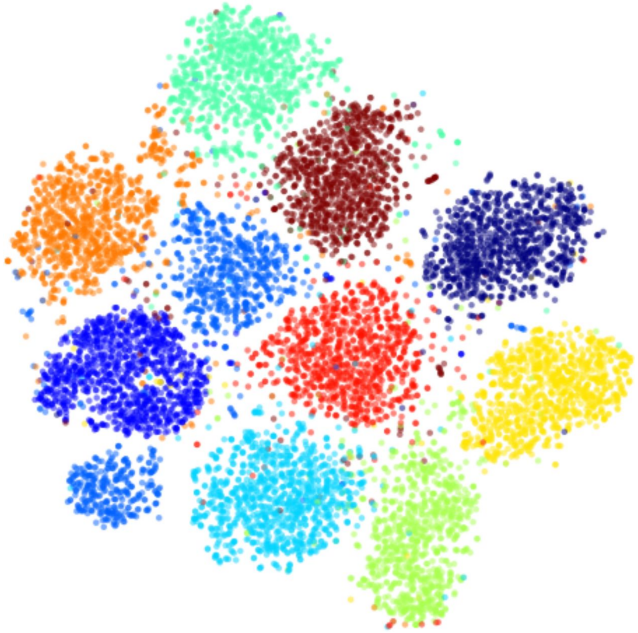
$$x_n \mid z_n \sim p_\gamma(x_n \mid z_n)$$

Use recognition network $q_\phi(z_n \mid x_n)$ to approximate $p(z_n \mid x_n)$

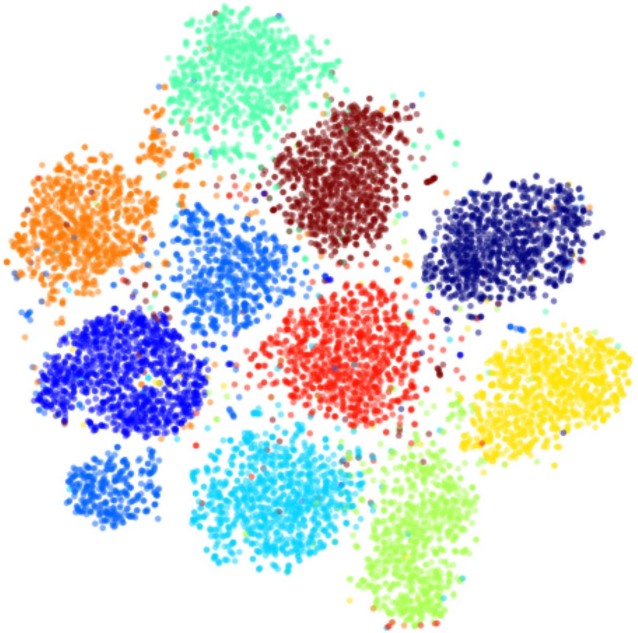
Learn neural network weights by optimizing ELBO:

$$L[q] \triangleq \mathbb{E}_q \left[\log \frac{p(\mathbf{z}_{1:N}) \prod_n p_\gamma(x_n \mid z_n)}{\prod_n q_\phi(z_n \mid x_n)} \right]$$

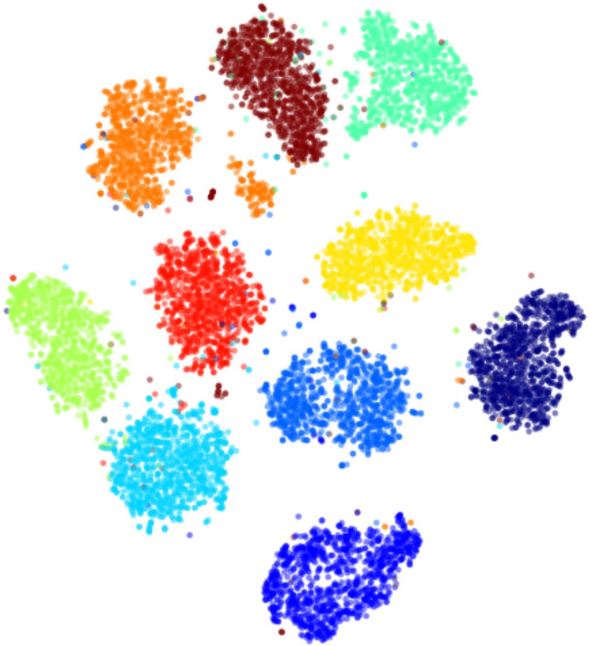
Learning a latent space with a VAE



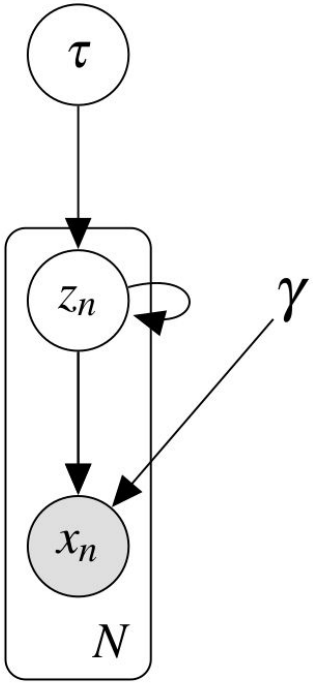
Opinionated priors for VAEs



vs.



Bayesian structured representation learning



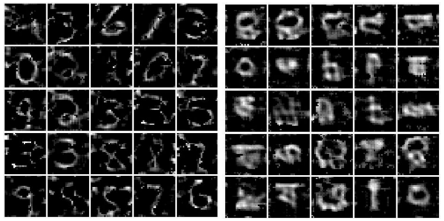
$$\tau \sim p(\tau)$$

$$\gamma \quad \mathbf{z}_{1:N} \mid \tau \sim p(\mathbf{z}_{1:N} \mid \tau)$$

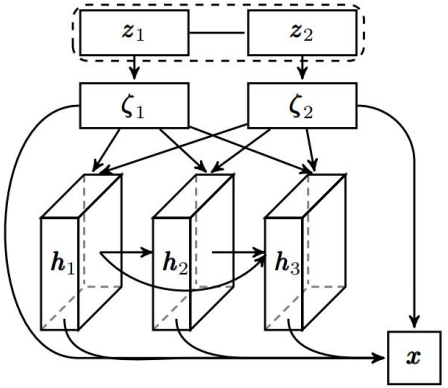
$$x_n \mid z_n \sim p_\gamma(x_n \mid z_n)$$

Embed data into a space that's organized in a particular way

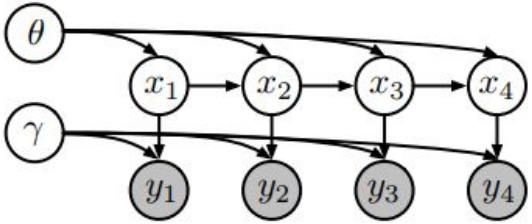
Examples



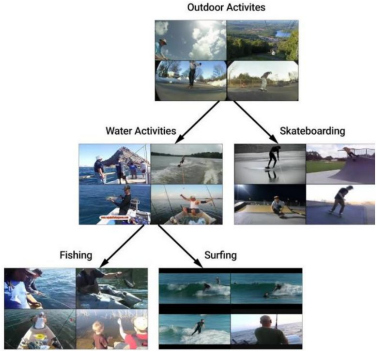
VampPrior [4]



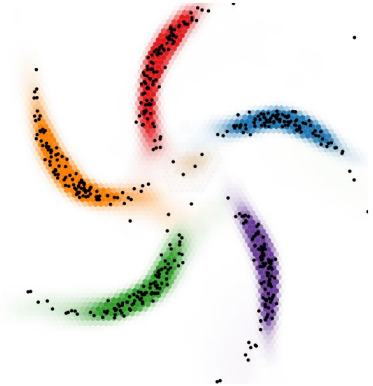
DVAE/DVAE++/DVAE# [5,6,7]



LDS SVAE/DKF/pFLDS
[9, 10, 11]

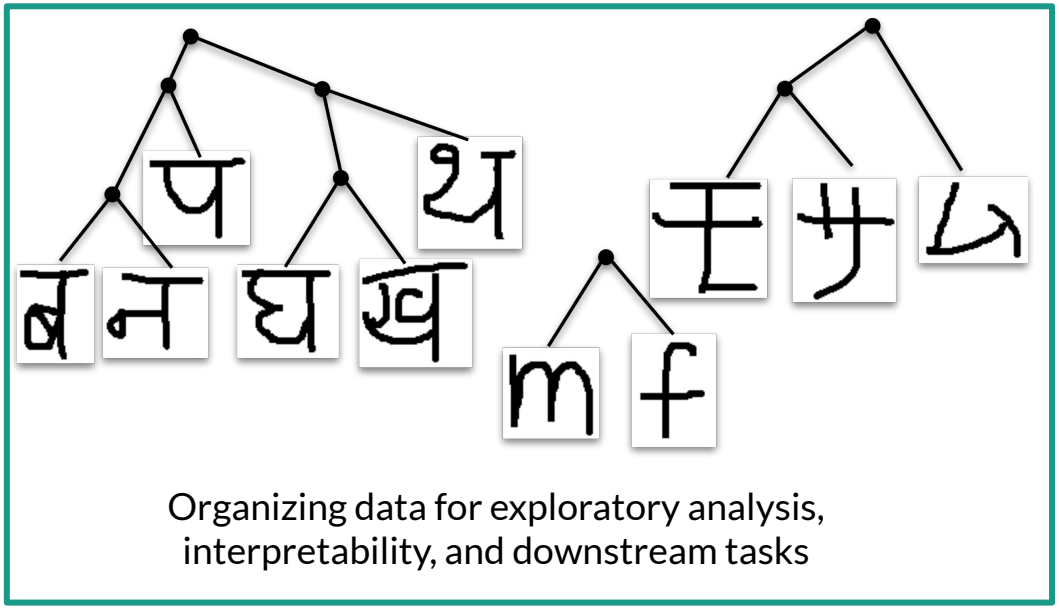


nCRP VAE [8]

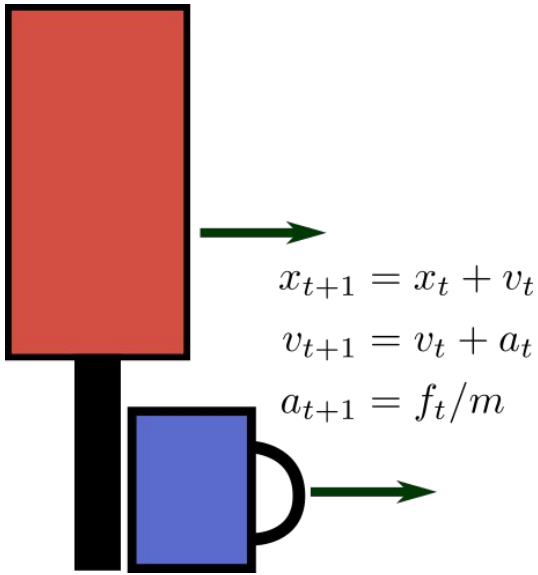


GMM SVAE [10]

Focus of this talk



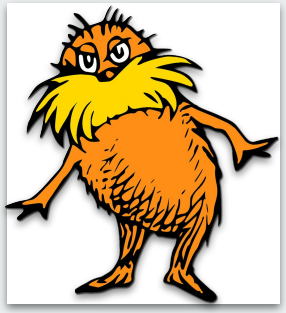
Organizing data for exploratory analysis, interpretability, and downstream tasks

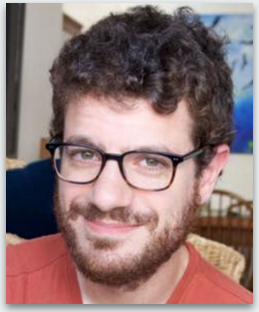


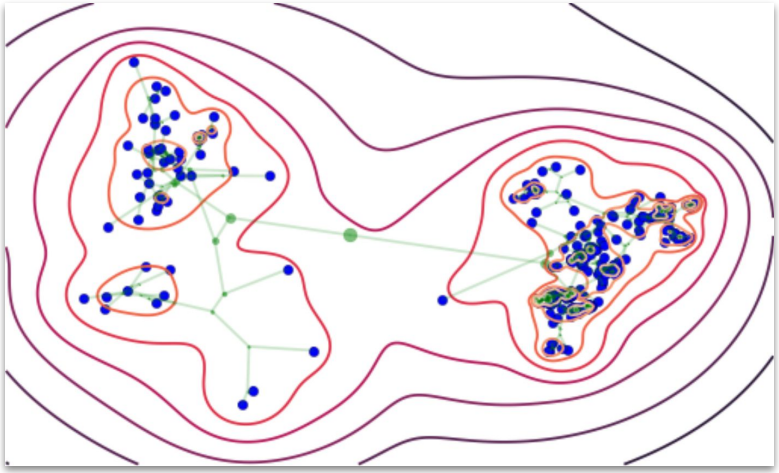
Organizing data for control

The LORACs Prior

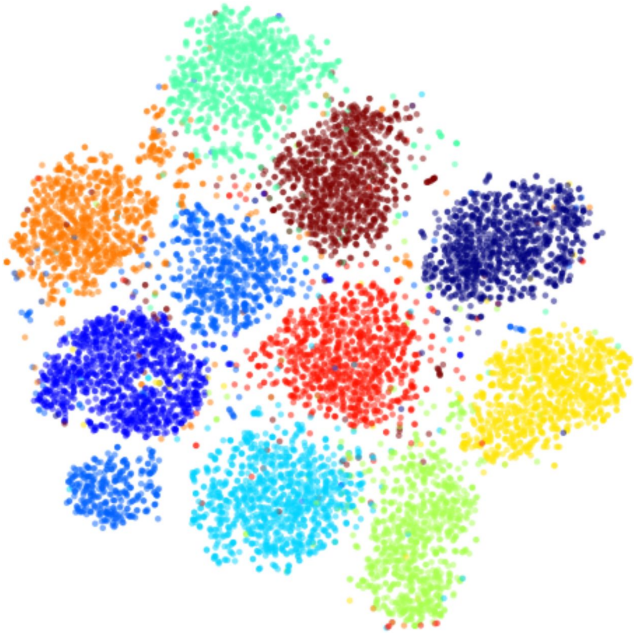
Letting the trees speak for the data

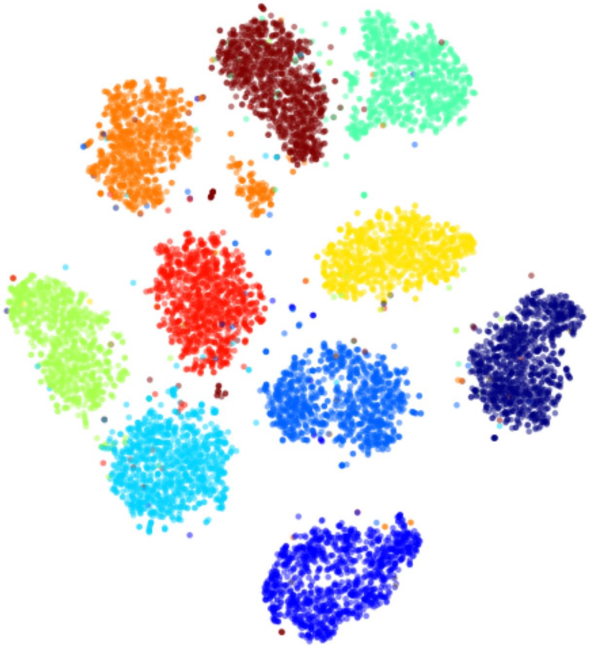




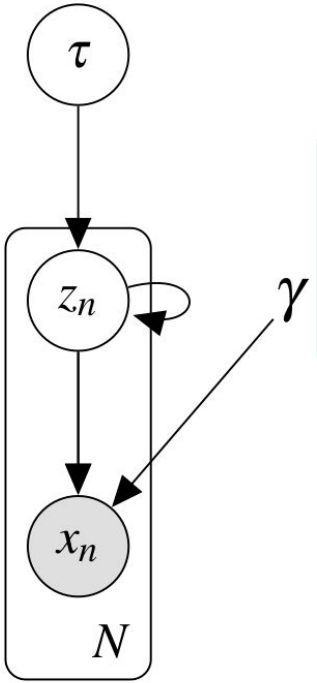


Choice of prior in VAEs





Deep Bayesian hierarchical clustering

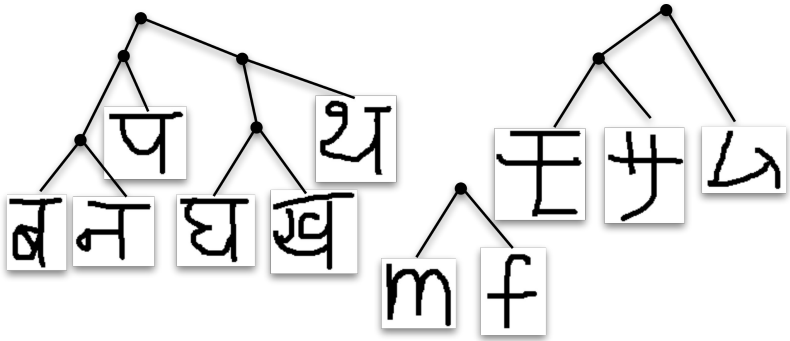


γ

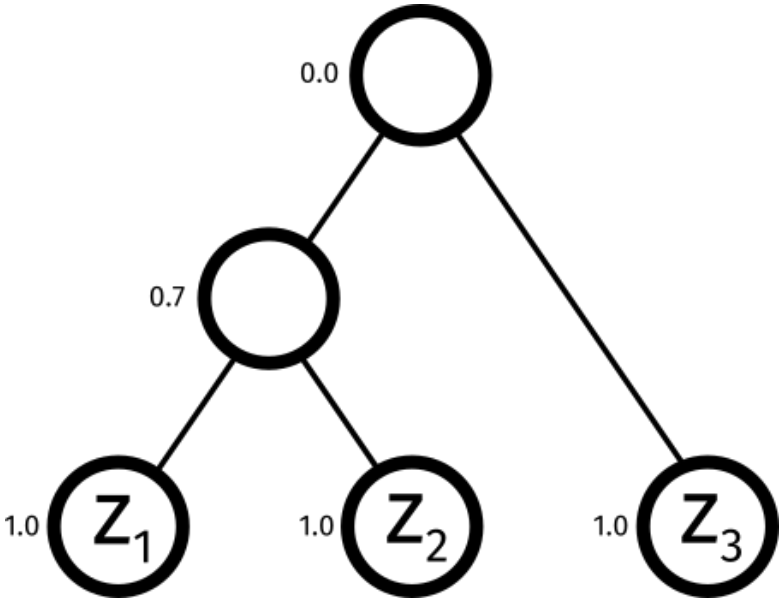
$$\tau \sim p(\tau)$$

$$\mathbf{z}_{1:N} \mid \tau \sim p(\mathbf{z}_{1:N} \mid \tau)$$

$$x_n \mid z_n \sim p_\gamma(x_n \mid z_n)$$



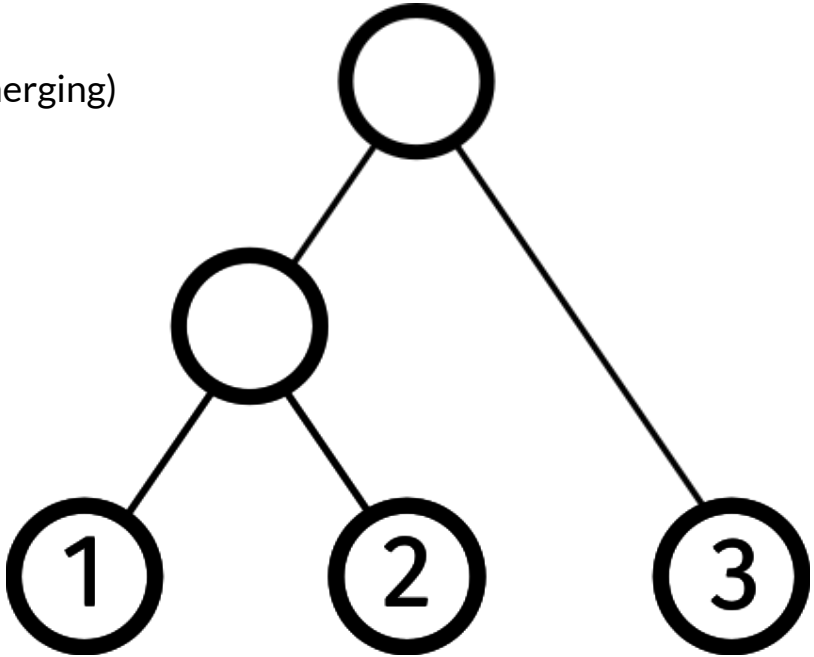
What is $p(\tau)$?



We are interested in distributions over *phylogenies*

Time-marginalized coalescent (TMC)

First we generate just the tree (random merging)



Time marginalized coalescent

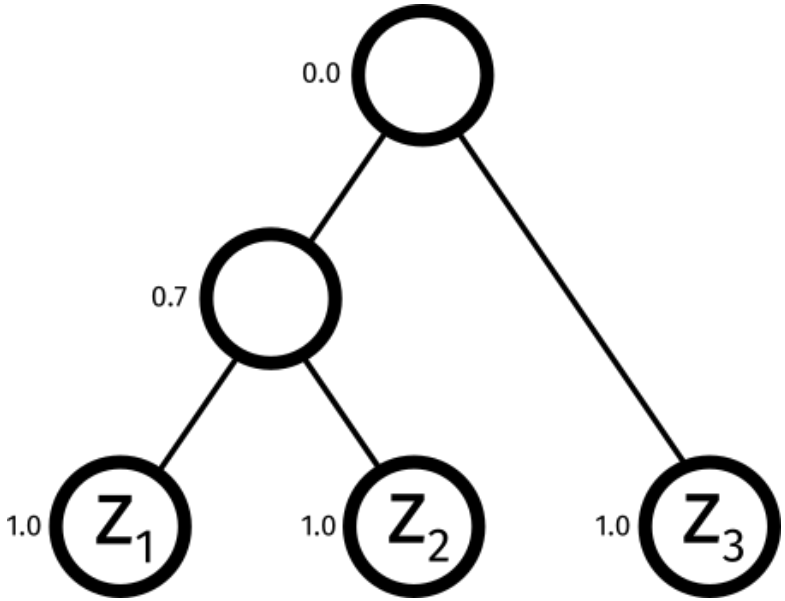
Then we generate times (stick breaking)

$$t_{\text{root}} = 0$$

$$t_{\text{leaves}} = 1$$

$$\beta_{\text{child}} \sim \text{Beta}(a, b)$$

$$t_{\text{child}} = t_{\text{parent}} + \beta_{\text{child}} * (1 - t_{\text{parent}})$$

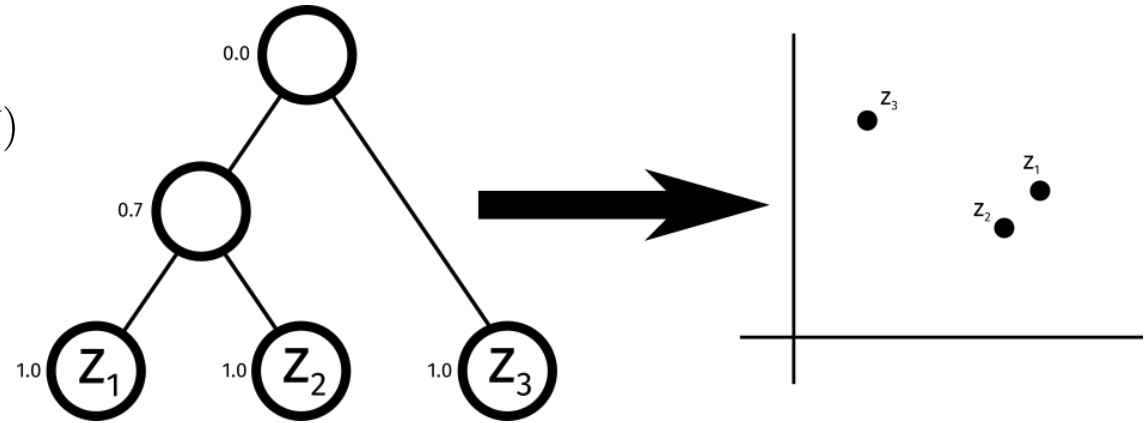


What is $p(\mathbf{z}_{1:N} \mid \tau)$?

We use a Brownian motion

$$z_{\text{root}} \sim N(0, I)$$

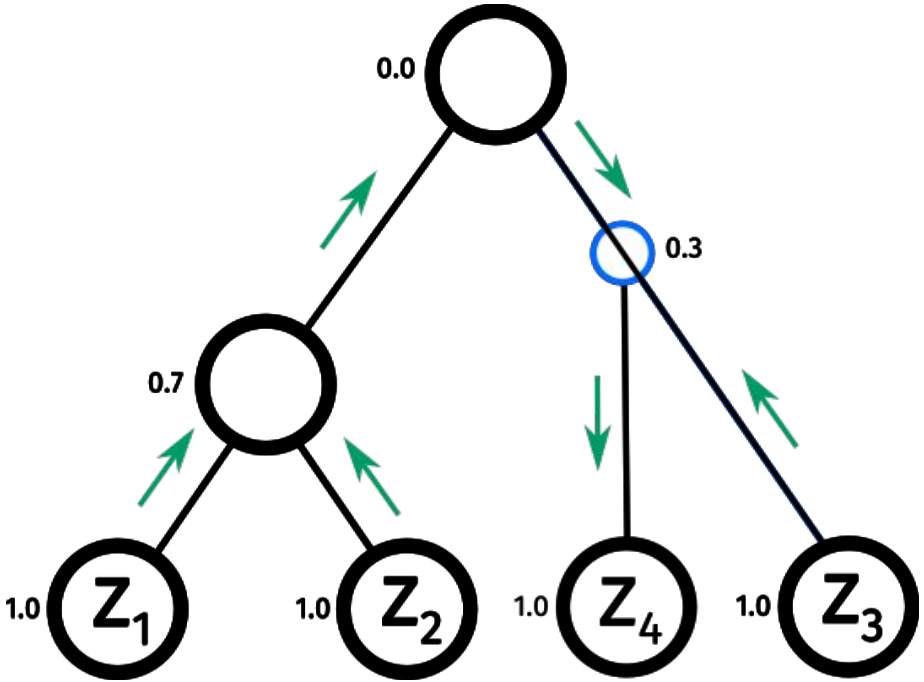
$$z_{\text{child}} \sim N(z_{\text{parent}}, (t_{\text{child}} - t_{\text{parent}})I)$$



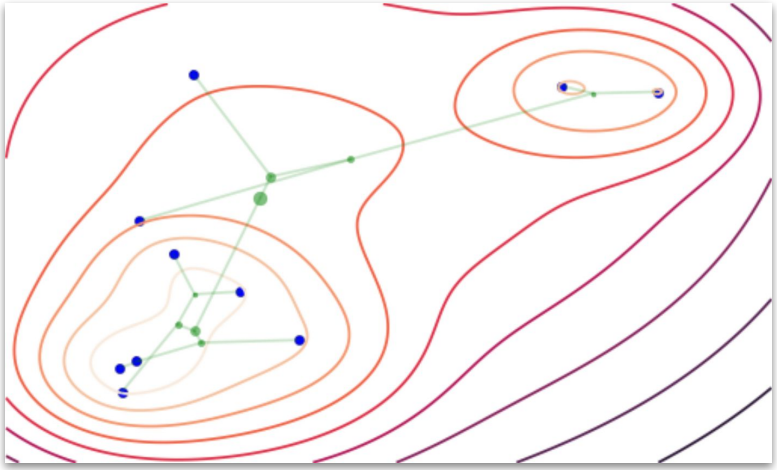
TMC as a density estimator

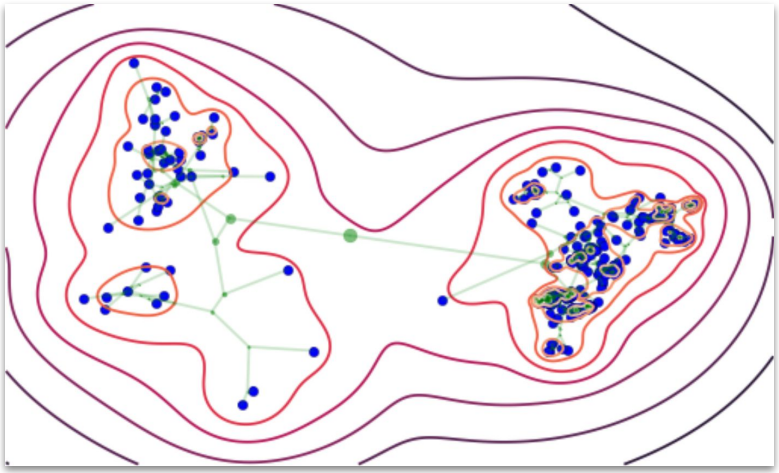
Posterior predictive density

$$p\left(z_{N+1} \mid \tau, \mathbf{z}_{1:N}\right)$$



TMC as a density estimator

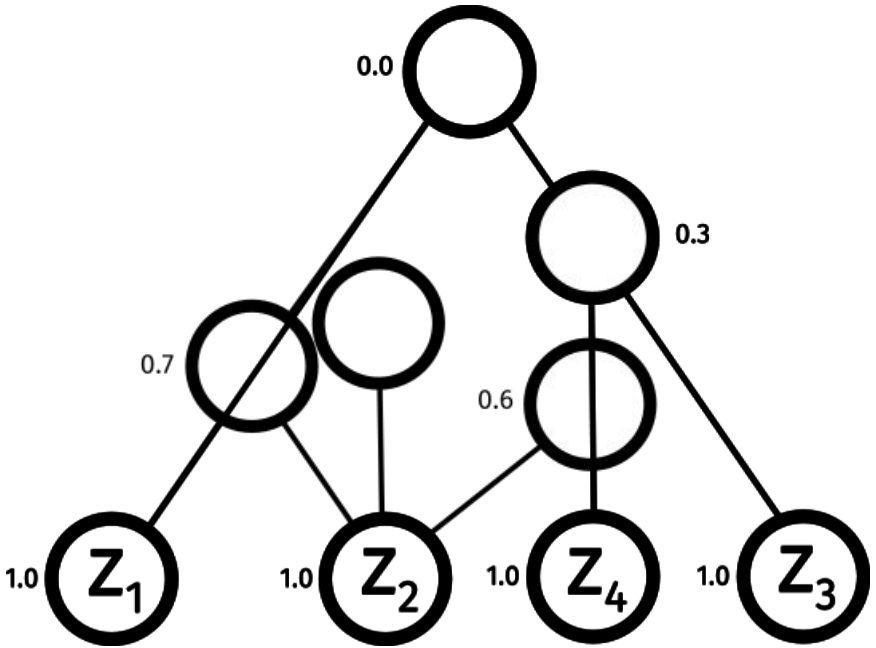




How to sample $p(\tau \mid \mathbf{z}_{1:N})$?

We can compute $p(\tau, \mathbf{z}_{1:N})$efficiently
so we use Metropolis-Hastings.

This is called subtree-prune and regraft (SPR).



TMC as a generative model

Can model both discrete and continuous structure

Posterior trees are useful for interpretability

Density estimator whose complexity grows with data

Downside: restrictive likelihood assumption, can VAE fix this?

TMC VAE?

What happens if we try the TMC VAE? First we assume a mean-field variational family

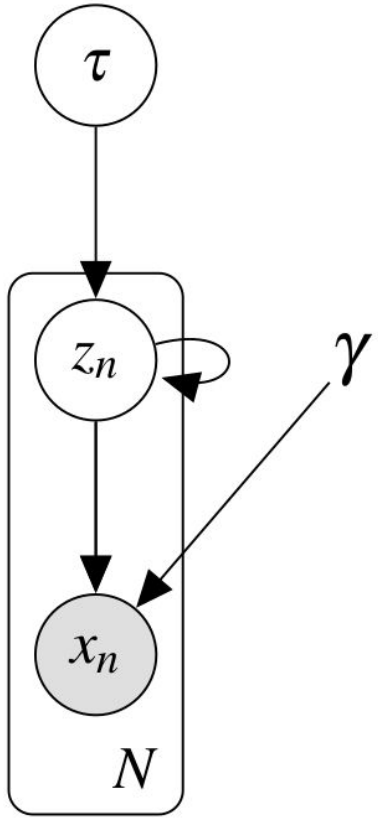
$$L[q] \triangleq \mathbb{E}_q \left[\log \frac{p(\tau)p(\mathbf{z}_{1:N} \mid \tau) \prod_{n=1}^{N} p_\gamma(x_n \mid z_n)}{q(\tau) \prod_{n=1}^{N} q_\phi(z_n | x_n)} \right]$$

Looking at new factor $q(\tau)$

$$q^*(\tau) \propto \exp\{\mathbb{E}_q [\log p(\tau, z_{1:N}, x_{1:N})]\}$$

$$\propto \exp\{\mathbb{E}_q [\log p(\tau) p(z_{1:N} | \tau)]\}$$

$$\propto \exp\{\log p(\tau) + \mathbb{E}_q [\log p(\mathbf{z}_{1:N} | \tau)]\}$$



TMC VAE?

$$L[q] \triangleq \mathbb{E}_q \left[\log \frac{p(\tau) p(\mathbf{z}_{1:N} \mid \tau) \prod_{n=1}^{N} p_\gamma(x_n \mid z_n)}{q(\tau) \prod_{n=1}^{N} q_\phi(z_n \mid x_n)} \right]$$

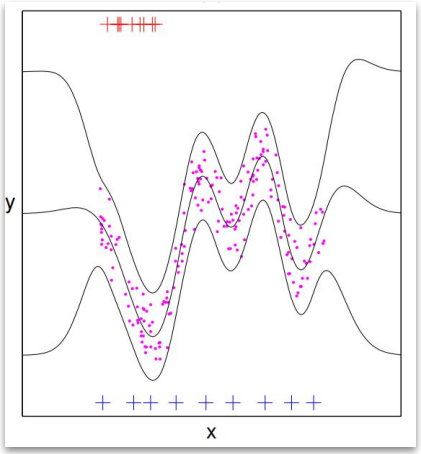
$$p(\mathbf{z}_{1:N} \mid \tau) \neq \prod_n p(z_n \mid \tau)$$

1. Can’t compute minibatch ELBO
2. MCMC over big trees can be slow

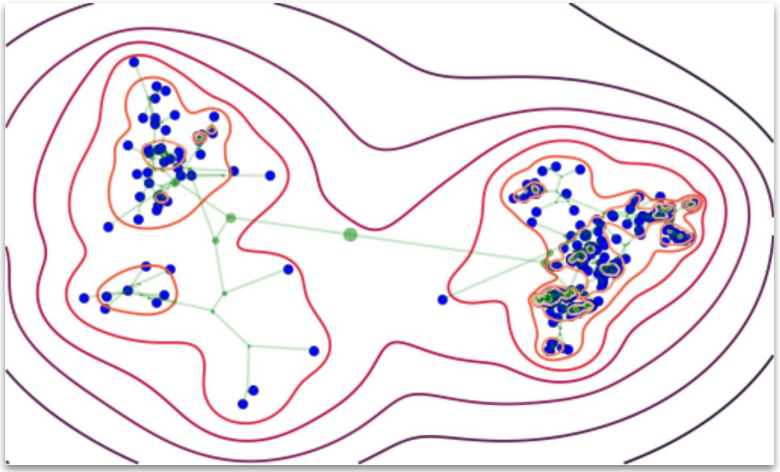
The LORACs Prior

LORACs: *Latent ORganization of Arboreal Clusters*

Core idea: **inducing point approximation**



Intuition: Sparse GPs



Inducing point approximation

Use TMC prior with **M** freely parametrized inducing points $\mathbf{s}_{1:M}$ as leaves

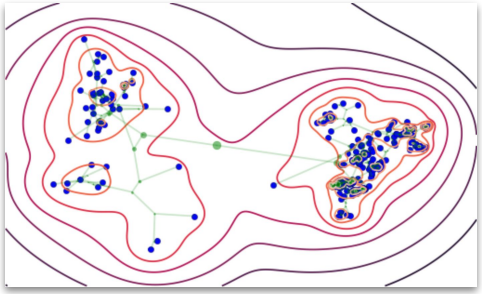
$$\tau \sim p(\tau \mid \mathbf{s}_{1:M})$$

Latent codes $\mathbf{z}_{1:N}$ are generated IID according to posterior predictive density

$$z_n \mid \tau \sim p(z_n = s_{M+1} \mid \tau, \mathbf{s}_{1:M})$$

Observed data are generated as in VAE

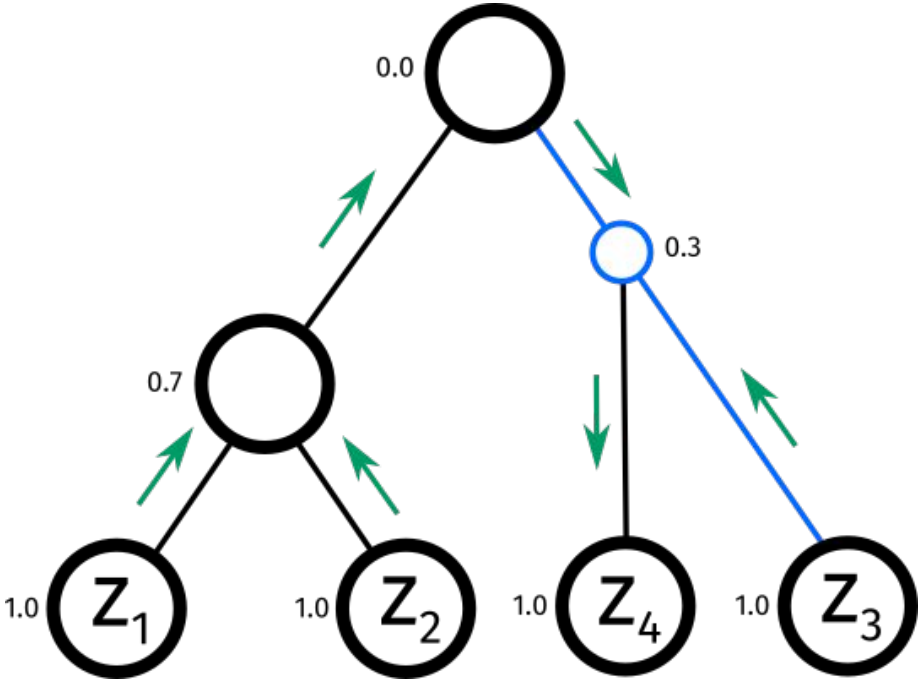
$$x_n \mid z_n \sim p_\gamma(x_n \mid z_n)$$



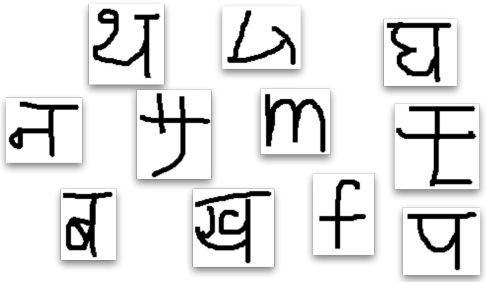
Inference in LORACs

Hybrid of variational inference with recognition networks and MCMC (Metropolis Hastings)

Inducing points and neural network weights learned with SGD on ELBO



Evaluation datasets



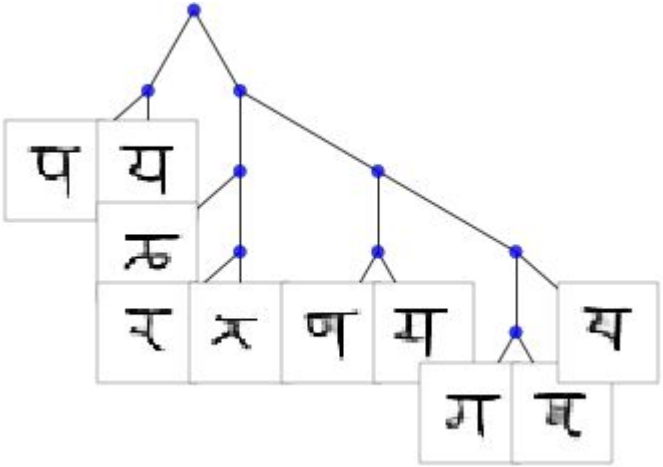
Visualizing inducing points

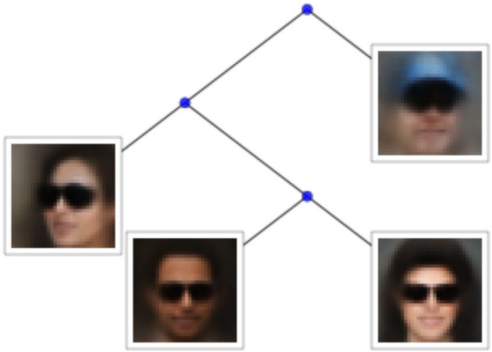
Omniglot (1000)

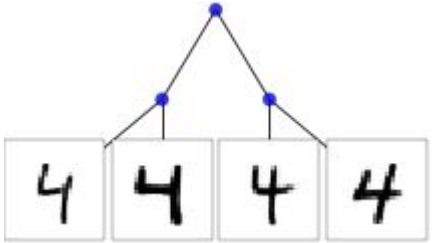
CelebA (500)

MNIST (200)

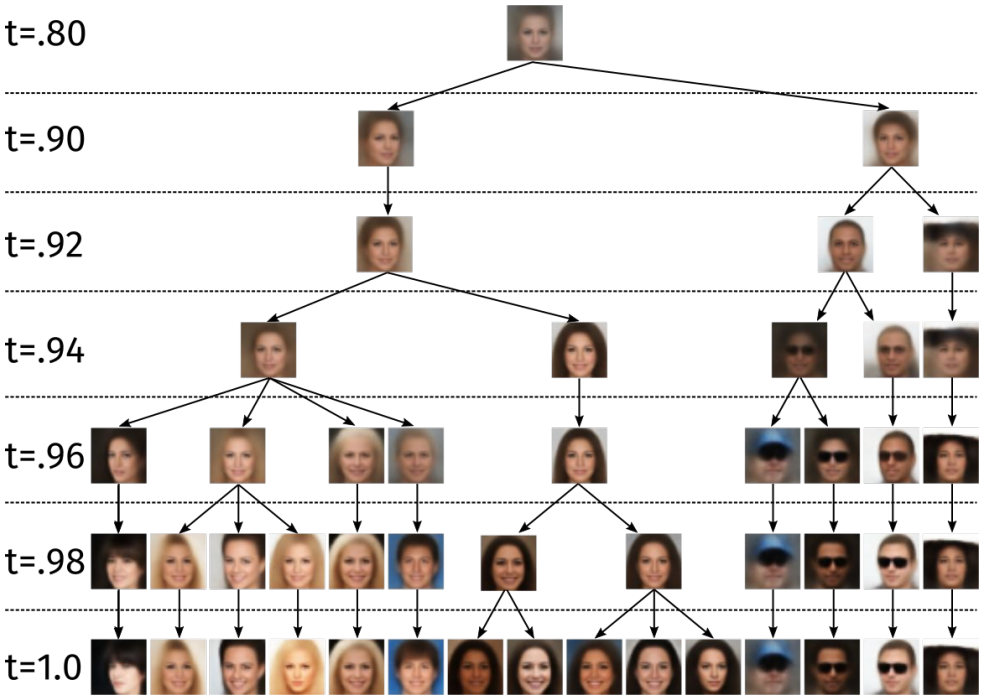
Learned subtrees



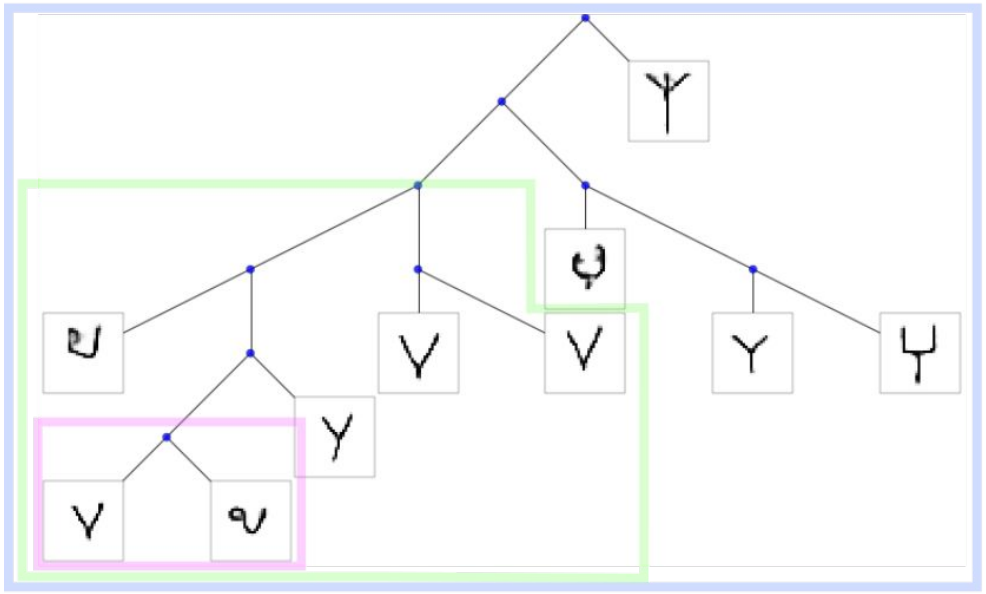


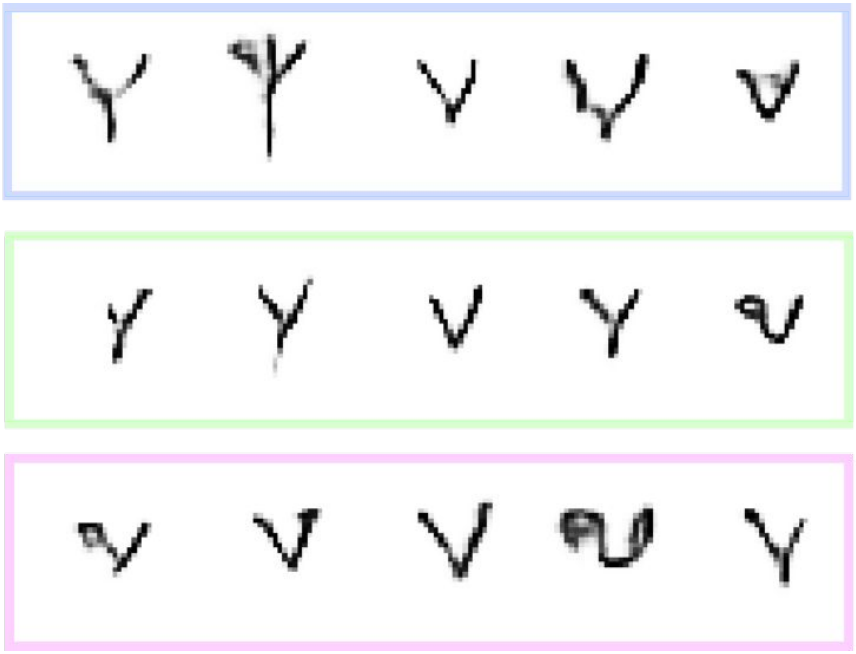


Evolutionary interpretation

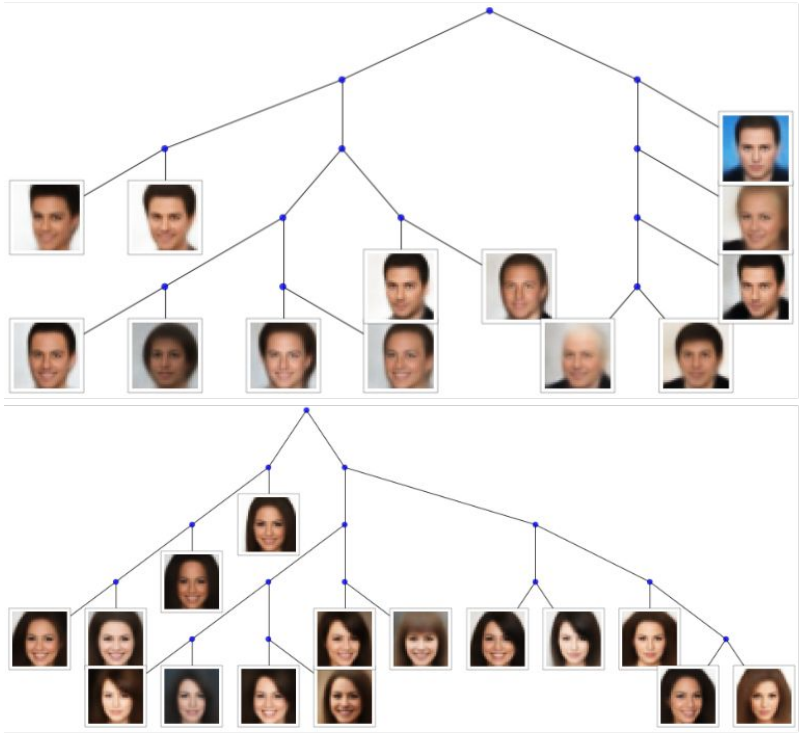


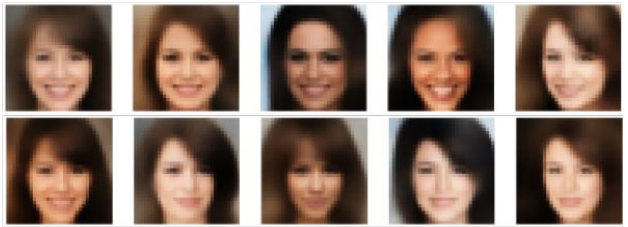
Conditional Sampling



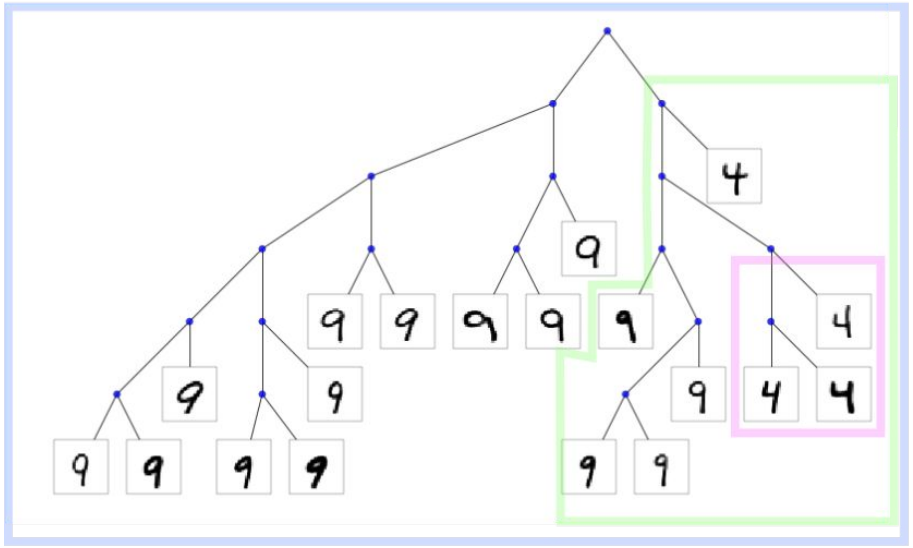


Conditional Sampling

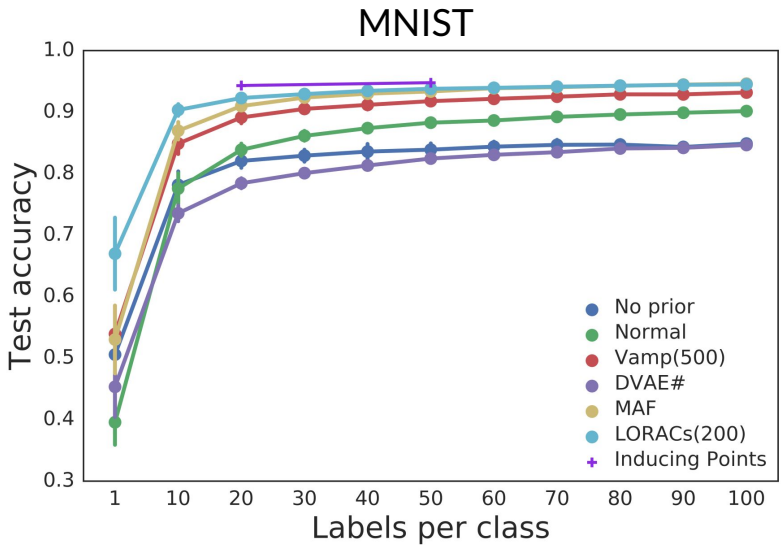


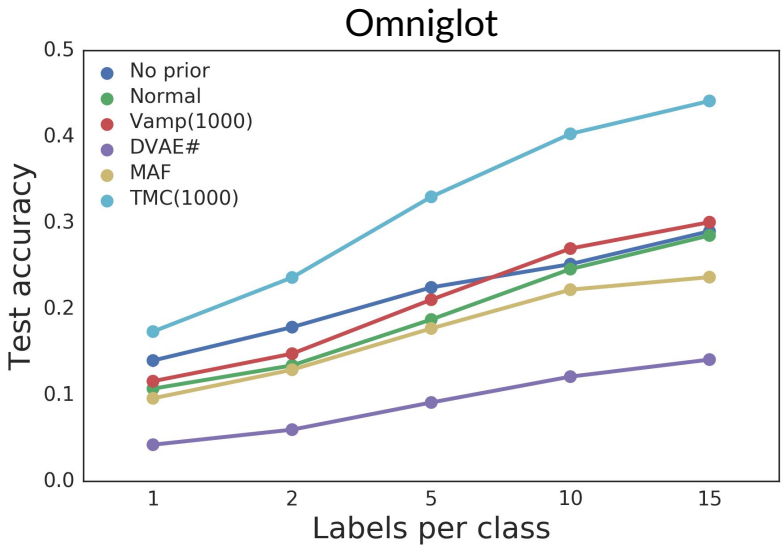


Conditional Sampling

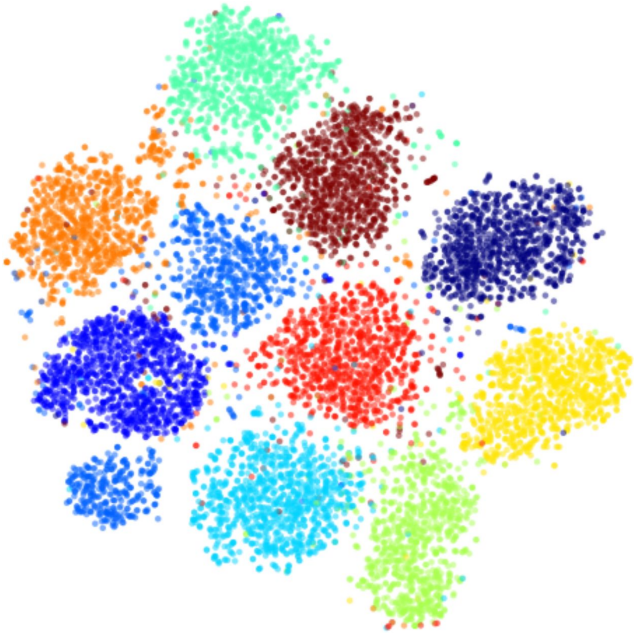


Few-shot learning

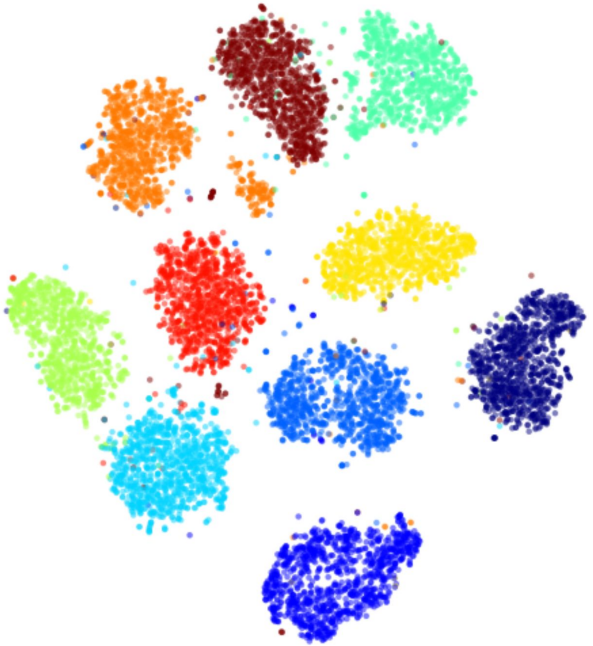




TSNE of latent spaces

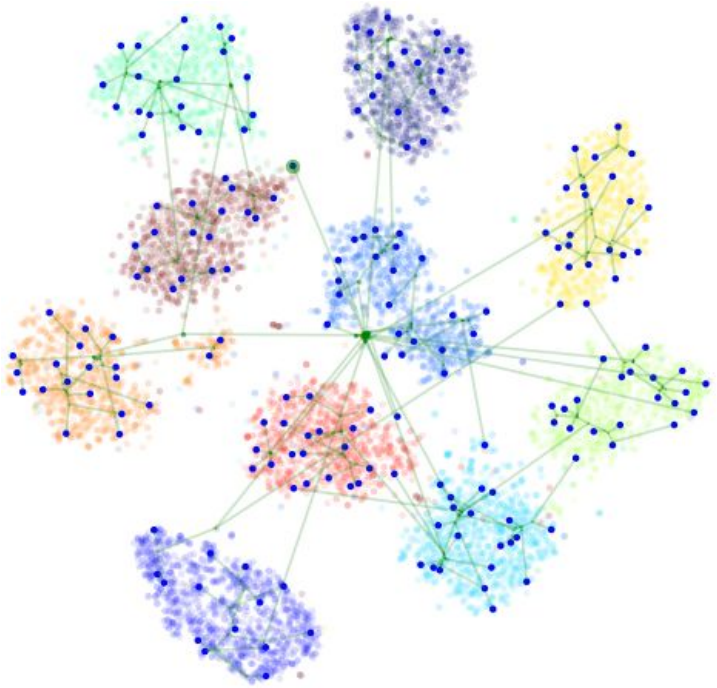


Normal prior



LORACs prior

TSNE of latent spaces



More quantitative results

Prior	MNIST	Omniglot
No prior	0.429	0.078
Normal	0.317	0.057
VAMP	0.502	0.063
DVAE#	0.490	0.024
MAF	0.398	0.070
LORACs	**0.626**	**0.087**

Table 7.1: Averaged precision-recall AUC on MNIST/Omniglot test datasets

Prior	MNIST	Omniglot
Normal	-83.789	-89.722
MAF	**-80.121**	**-86.298**
Vamp	-83.0135	-87.604
LORACs	-83.401	-87.105

Table 7.2: MNIST/Omniglot test log-likelihoods

Future work

Applying LORACs prior to non-image data (text, audio)

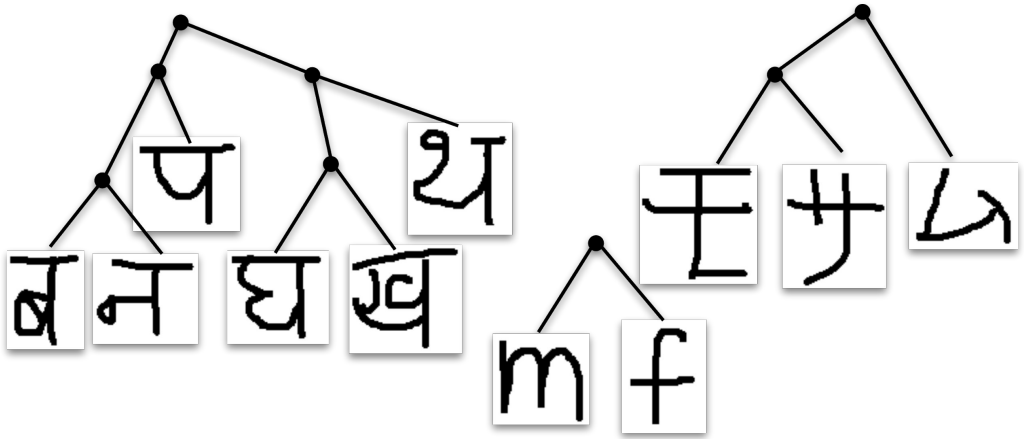
Incorporate user interaction to resolve ambiguous clusterings

Inducing point approximations for other BNP distributions (e.g. graphs)

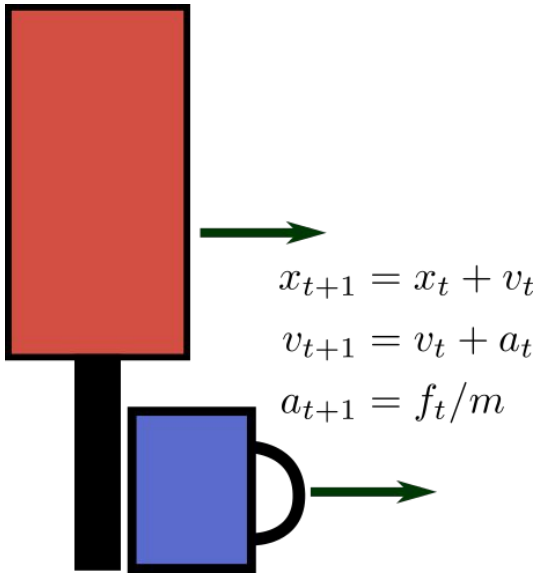
Hierarchical clusterings across axes/dimensions

Combining structured learning with normalizing flows

Focus of this talk



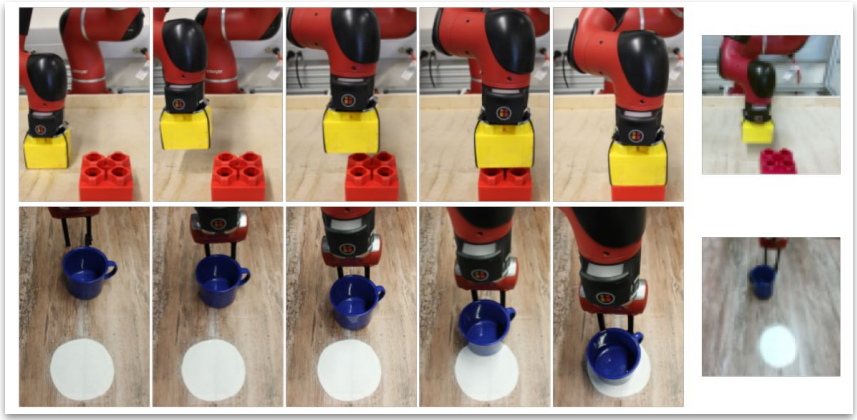
Organizing data for exploratory analysis, interpretability, and downstream tasks



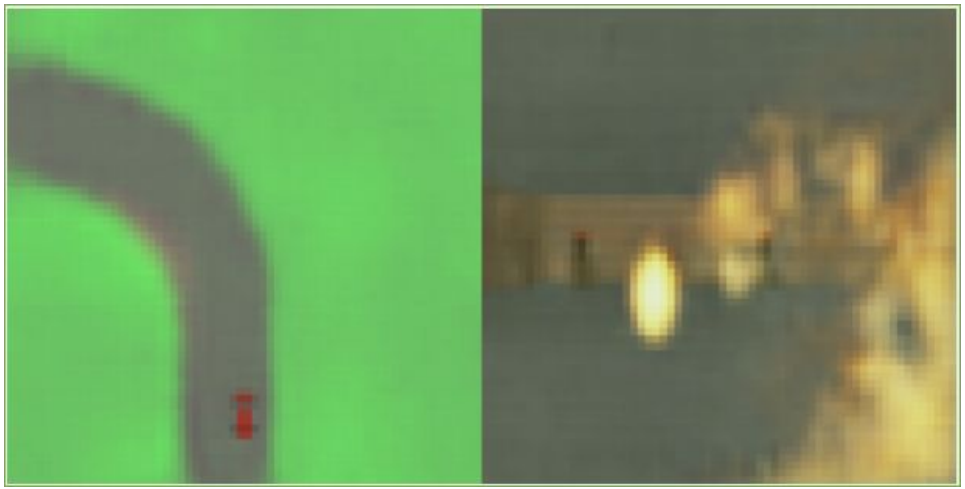
Organizing data for control

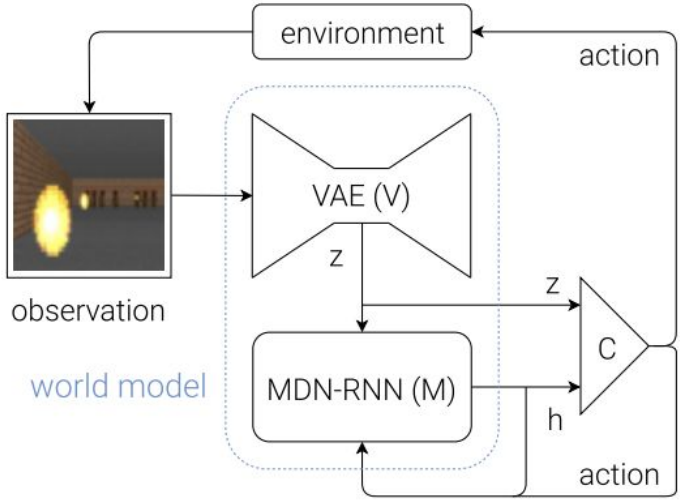
SOLAR

Deep Structured Representations for Model-Based Reinforcement Learning



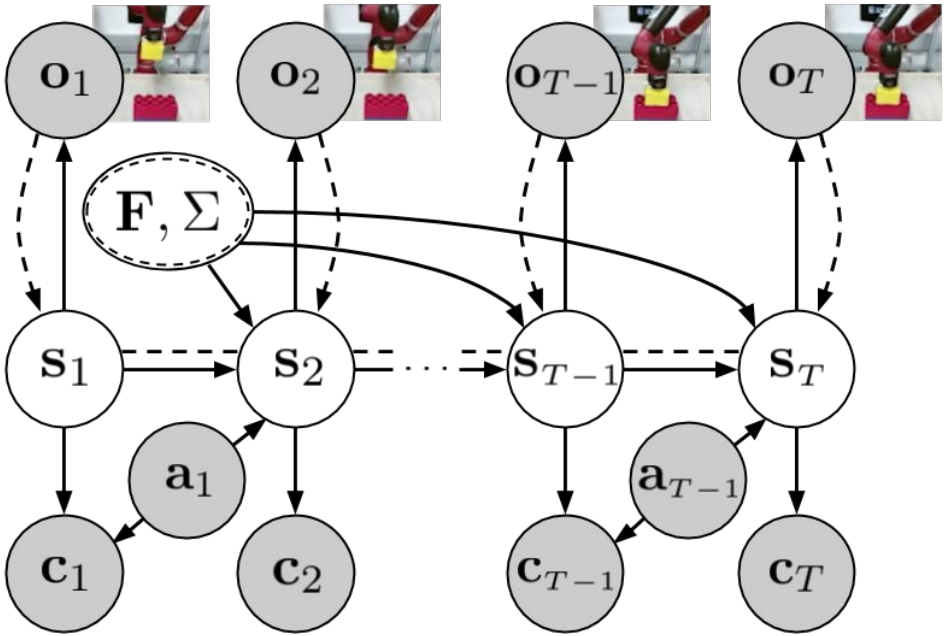
World models for RL





[12]

LDS SVAE for reinforcement learning



$$\mathbf{s}_1 \sim N(0, \mathbf{I}) \,,$$

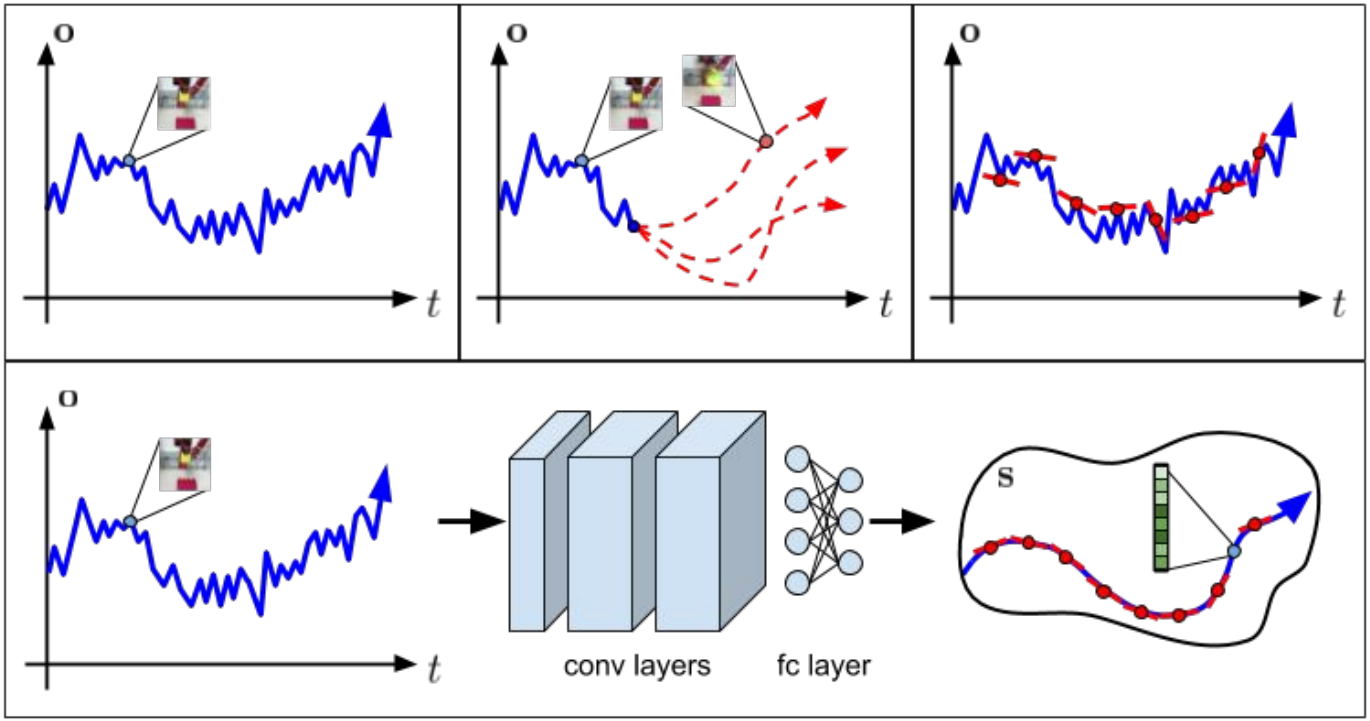
$$\mathbf{F}, \boldsymbol{\Sigma} \sim MNIW(\Psi, \nu, \mathbf{F}_0, \mathbf{V}) \,,$$

$$\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t, \mathbf{F}, \boldsymbol{\Sigma} \sim N\left(\mathbf{F}\begin{bmatrix}\mathbf{s}_t \\ \mathbf{a}_t\end{bmatrix}, \boldsymbol{\Sigma}\right) \,,$$

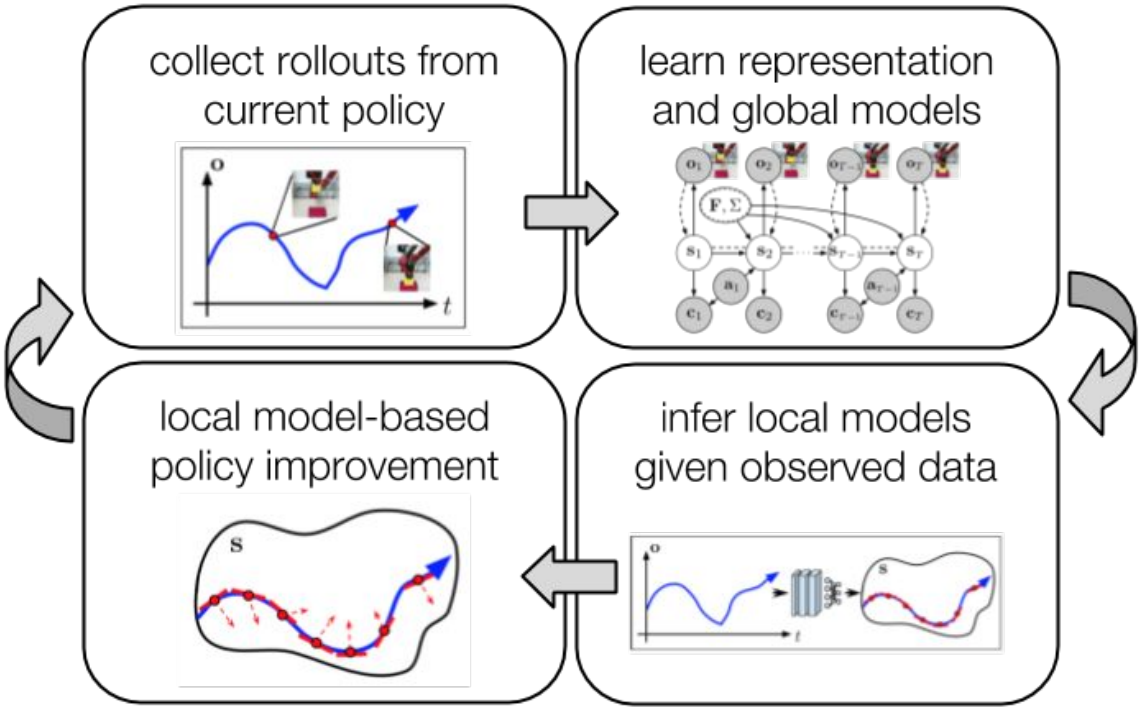
$$\mathbf{o}_t \mid \mathbf{s}_t \sim f_\gamma(\mathbf{s}_t) \,,$$

$$c_t \mid \mathbf{s}_t, \mathbf{a}_t \sim N\left(\hat{C}(\mathbf{s}_t, \mathbf{a}_t), 1\right) \,.$$

Local model based control



The SOLAR algorithm



Results


Final Policy - SOLAR
5x speed
Image Observations
True Reward
Predicted Reward
Trial 2 - Success
Model Reconstructions


Image Observations
VAE Ablation - Task 3
300 Training Episodes
Final Policy
5x speed
Model Reconstructions
Trial 2
Failure

Thank you!

Any questions?

My Publications

SOLAR: Deep Structured Latent Representations for Model-Based Reinforcement Learning
Marvin Zhang*, Sharad Vikram*, Laura Smith, Pieter Abbeel, Matthew J. Johnson, Sergey Levine
ICML 2019

How to pick the domain randomization parameters for sim-to-real transfer of reinforcement learning policies?
Quan Vuong, Sharad Vikram, Hao Su, Sicun Gao, Henrik I. Christensen
Preprint

The LORACs prior for VAEs: Letting the Trees Speak for the Data
Sharad Vikram, Matthew D. Hoffman, Matthew J. Johnson
AISTATS 2019

Neural Variational Message Passing
Sharad Vikram
Preprint

Estimating Reactions and Recommending Products with Generative Models of Reviews
Jianmo Ni, Zachary Lipton, Sharad Vikram, Julian McAuley
IJCNLP 2017

Interactive Bayesian Hierarchical Clustering
Sharad Vikram, Sanjoy Dasgupta
ICML 2016

Bibliography

[1] Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of eugenics 7.2 (1936): 179-188.
[2] Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction." Science 350.6266 (2015): 1332-1338.
[3] Liu, Ziwei, et al. "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.
[4] Tomczak, Jakub M., and Max Welling. "VAE with a VampPrior." arXiv preprint arXiv:1705.07120 (2017).
[5] Rolfe, Jason Tyler. "Discrete variational autoencoders." arXiv preprint arXiv:1609.02200 (2016).
[6] Vahdat, Arash, et al. "DVAE++: Discrete variational autoencoders with overlapping transformations." arXiv preprint arXiv:1802.04920 (2018).
[7] Vahdat, Arash, Evgeny Andriyash, and William Macready. "DVAE#: Discrete variational autoencoders with relaxed Boltzmann priors." Advances in Neural Information Processing Systems. 2018.
[8] Goyal, Prasoon, et al. "Nonparametric variational auto-encoders for hierarchical representation learning." Proceedings of the IEEE International Conference on Computer Vision. 2017.
[9] Gao, Yuanjun, et al. "Linear dynamical neural population models through nonlinear embeddings." Advances in neural information processing systems. 2016.
[10] Johnson, Matthew, et al. "Composing graphical models with neural networks for structured representations and fast inference." Advances in neural information processing systems. 2016.
[11] Rahul G Krishnan, Uri Shalit, and David Sontag. "Deep Kalman filters." arXiv preprint arXiv:1511.05121. 2015.
[12] Ha, David, and Jürgen Schmidhuber. "World models." arXiv preprint arXiv:1803.10122. 2018.
[13] Boyles, Levi, and Max Welling. "The time-marginalized coalescent prior for hierarchical clustering." Advances in Neural Information Processing Systems. 2012.
[14] LeCun, Yann, Corinna Cortes, and C. J. Burges. "MNIST handwritten digit database." AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist 2 (2010): 18.