Composing graphical models and neural networks for structured representations and fast inference

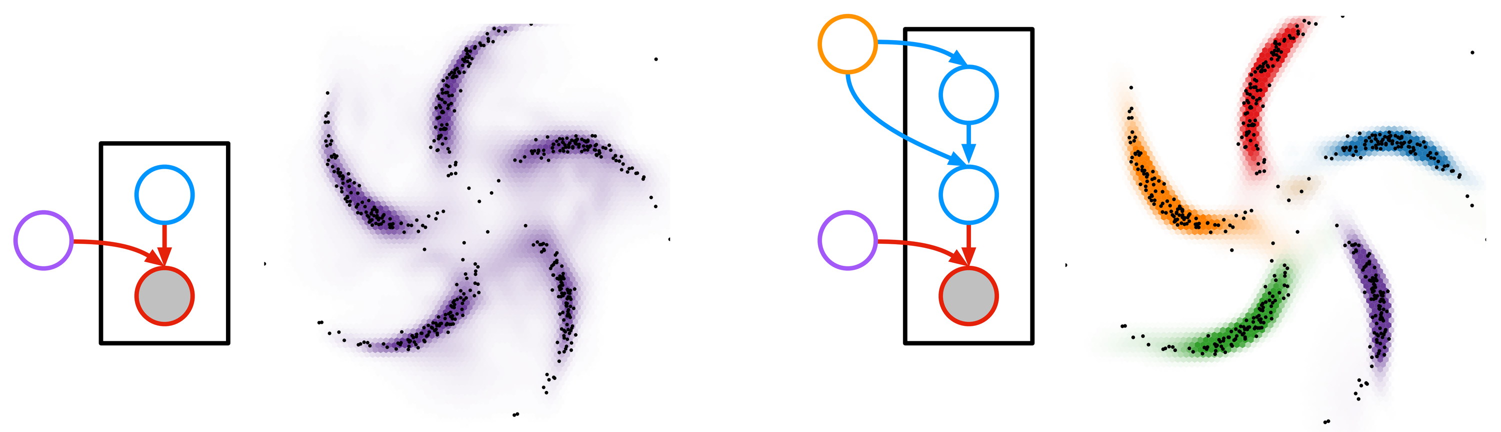
HARVARD INTELLIGENT PROBABILISTIC SYSTEMS
HARVARD MEDICAL SCHOOL
github.com/mattjj/svae

Matthew James Johnson mattjj@csail.mit.edu
David Duvenaud duvenaud@cs.toronto.edu
Alexander B. Wiltschko awiltsch@fas.harvard.edu
Sandeep R. Datta srdatta@hms.harvard.edu
Ryan P. Adams rpa@seas.harvard.edu

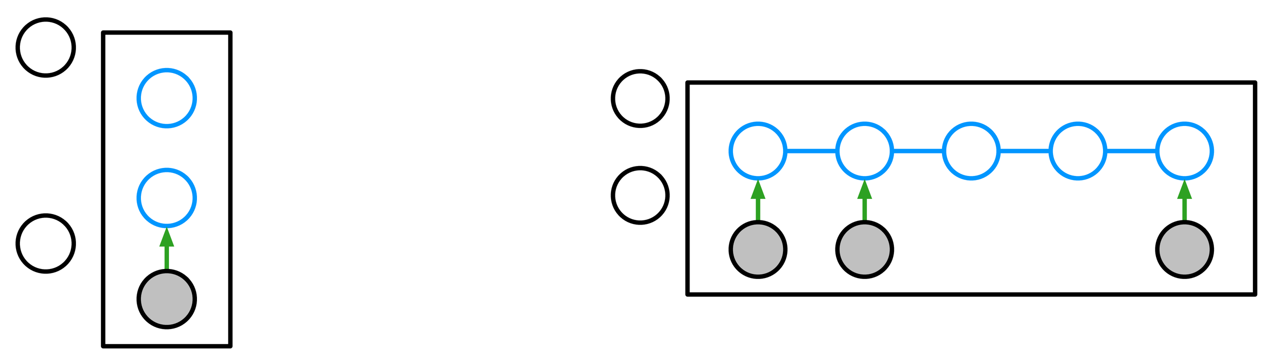
motivation

TL;DR variational autoencoders + latent graphical models

modeling idea use PGM priors to **organize** the latent space, along with neural net observation models for **flexible representations**



inference idea use PGMs to **synthesize information** from **recognition nets** instead of making a single inference net do everything

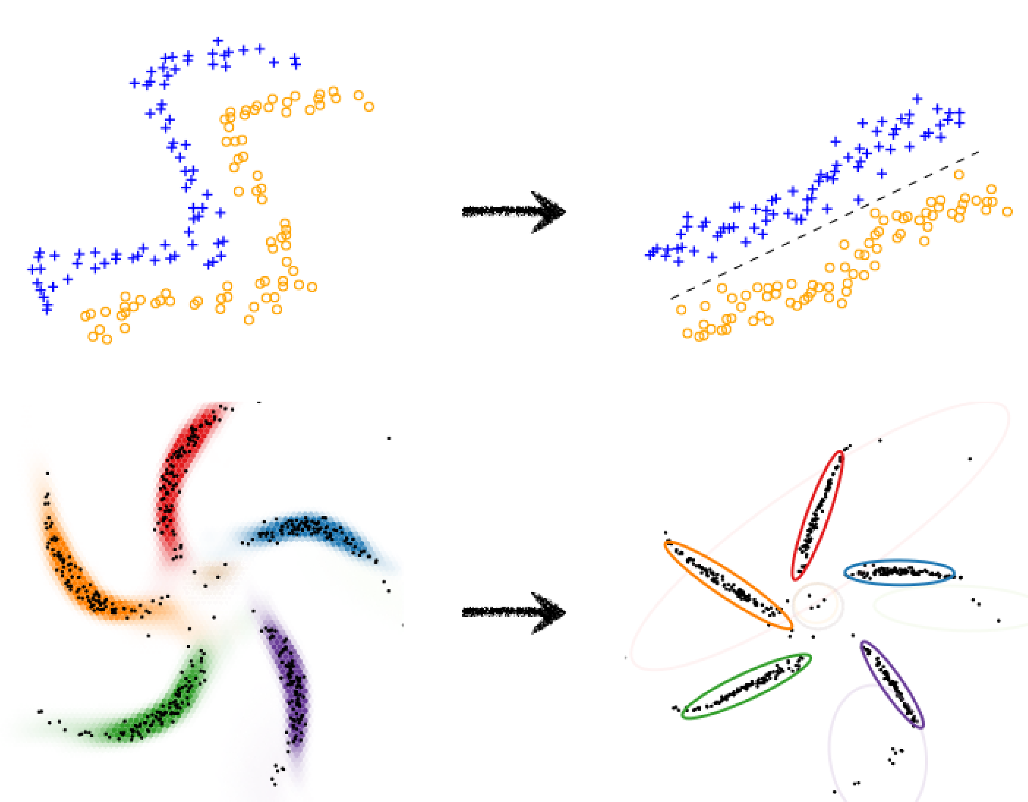


<u>probabilistic graphical models</u>	<u>deep neural networks</u>
+ structured representations	– neural net "goo"
+ priors and uncertainty	– difficult parameterization
+ data and computational efficiency within rigid model classes	– can require lots of data
– rigid assumptions may not fit	+ flexible, high capacity
– feature engineering	+ feature learning
– more flexible models can require slow top-down inference	+ recognition networks for fast bottom-up inference

automatically learn representations in which structured PGMs fit well

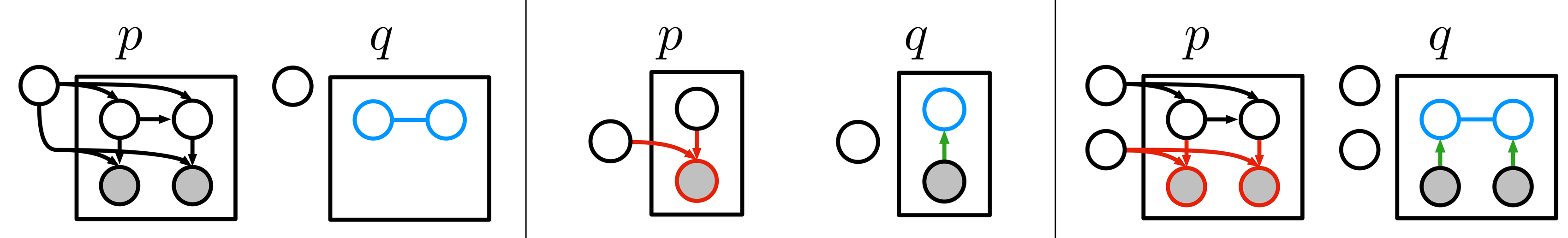
supervised learning

unsupervised learning

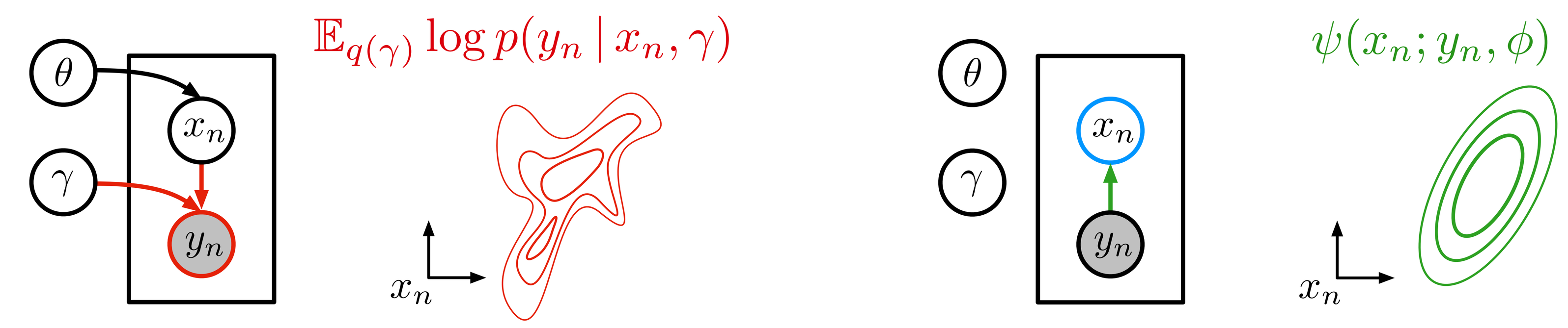


inference

natural gradient SVI	variational autoencoders	structured VAEs (this work)
$q^*(x) \triangleq \arg\max_{q(x)} \mathcal{L}[\,q(\theta)q(x)\,]$	$q^*(x) \triangleq \mathcal{N}(x\,\vert\,\mu(y;\phi), \Sigma(y;\phi))$	$q^*(x) \triangleq$?
+ optimal local factor	– suboptimal local factor	± optimal given conj. evidence
– expensive for general obs.	+ fast for general obs.	+ fast for general obs.
+ exploit conj. graph structure	– ϕ does all local inference	+ exploit conj. graph structure
+ arbitrary inference queries	– limited inference queries	+ arbitrary inference queries
+ natural gradients	– no natural gradients	+ some natural gradients



main idea learn to summarize complicated evidence with simple conjugate potentials (as in CRFs)

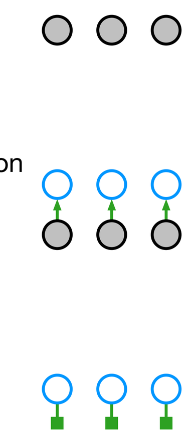


<u>step 1</u> compute local evidence with recognition networks

1(a) sample minibatch

1(b) apply recognition networks…

1(c) …to get PGM potentials

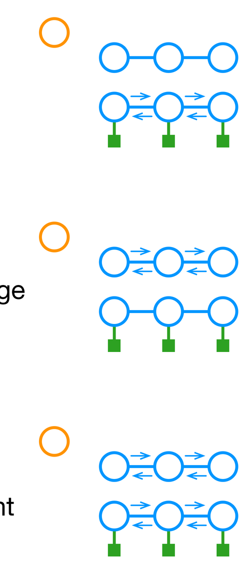


<u>step 2</u> run fast PGM inference

2(a) run local mean field

2(b) and message passing

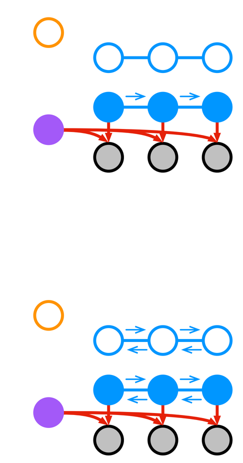
2(c) to fixed point



<u>step 3</u> compute unbiased ELBO gradients with respect to all parameters

3(a) sample and compute flat gradients w.r.t ϕ and γ

3(b) use already-computed values to get natural gradient w.r.t. η_θ



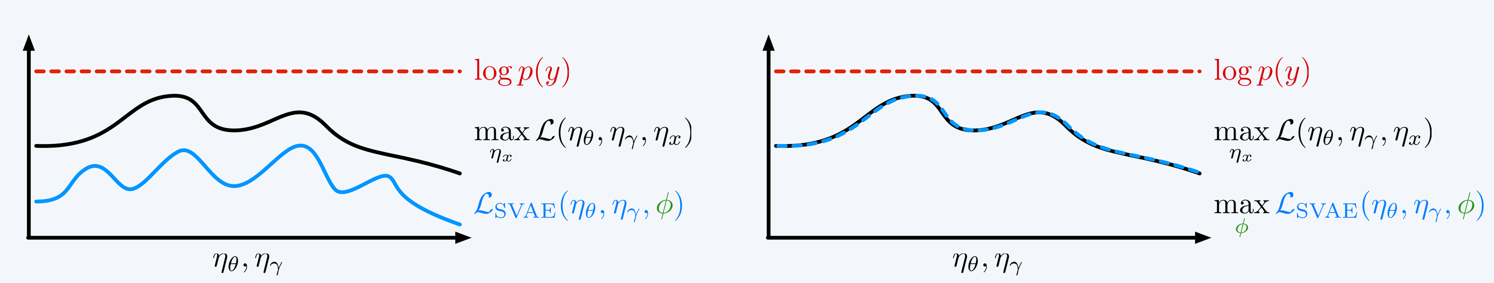
$$\mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x) \triangleq \mathbb{E}_{q(\theta)q(\gamma)q(x)}\left[\log \frac{p(\theta,\gamma,x)p(y\,\vert\,x,\gamma)}{q(\theta)q(\gamma)q(x)}\right]$$

$$\widehat{\mathcal{L}}(\eta_\theta, \eta_x, \phi) \triangleq \mathbb{E}_{q(\theta)q(\gamma)q(x)}\left[\log \frac{p(\theta,\gamma,x)\exp\{\psi(x;y,\phi)\}}{q(\theta)q(\gamma)q(x)}\right]$$

where $\phi(x; y, \phi)$ is a conjugate potential for $p(x\,\vert\,\theta)$

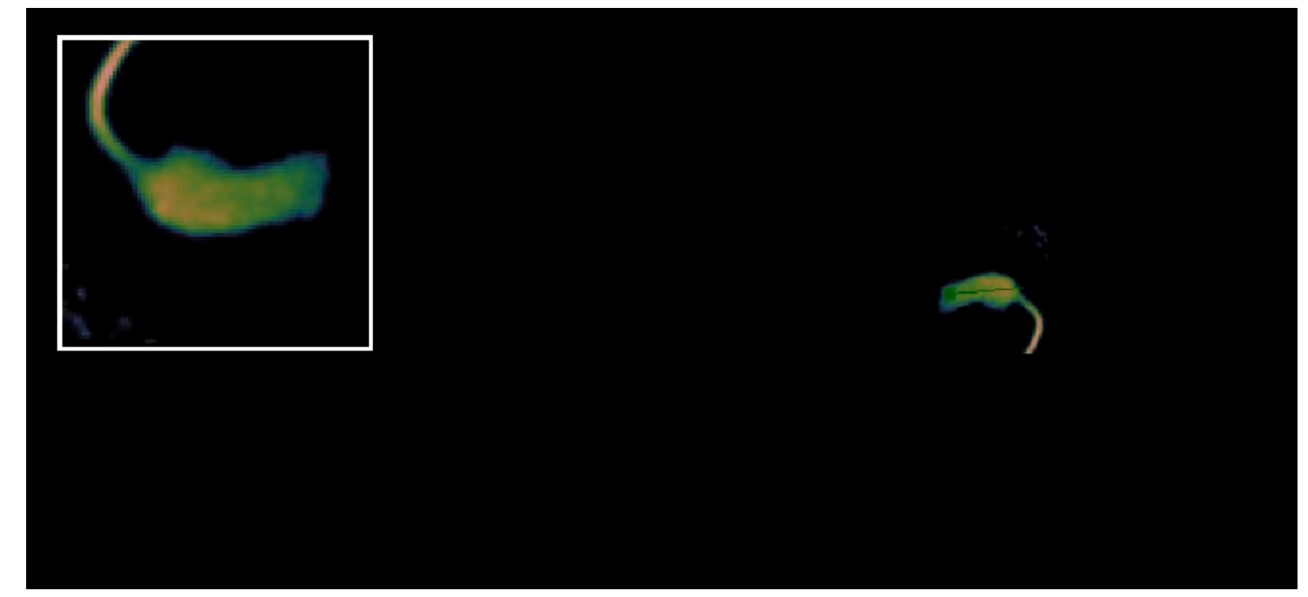
$$\eta_x^*(\eta_\theta, \phi) \triangleq \arg\max_{\eta_x} \widehat{\mathcal{L}}(\eta_\theta, \eta_x, \phi) \qquad \mathcal{L}_{\text{SVAE}}(\eta_\theta, \eta_\gamma, \phi) \triangleq \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x^*(\eta_\theta, \phi))$$

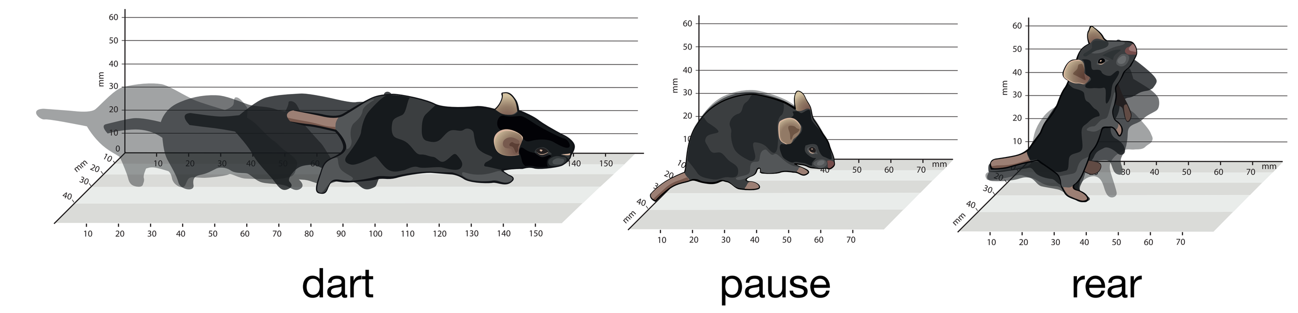
Proposition (log evidence lower bound)

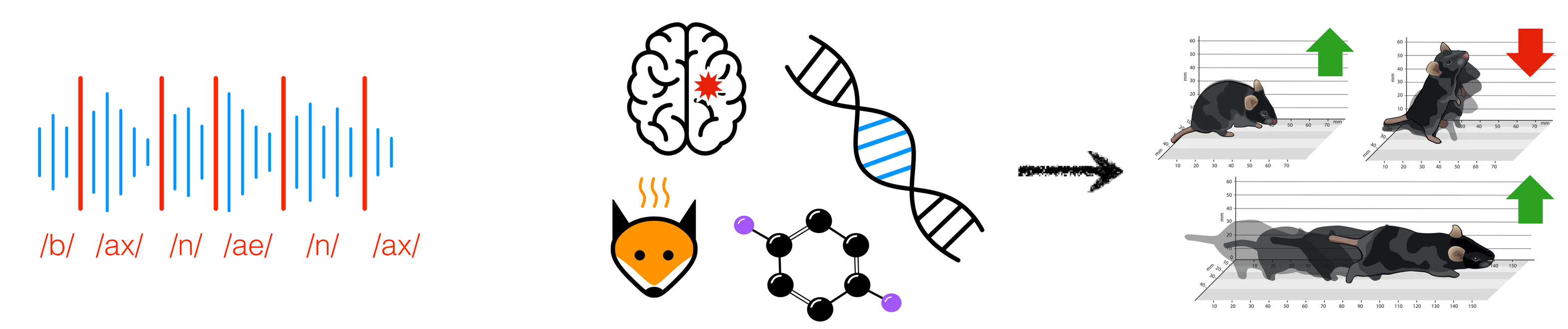


if $\exists\,\phi \in \mathbb{R}^m$ with $\psi(x; y, \phi) = \mathbb{E}_{q(\gamma)} \log p(y\,\vert\,x, \gamma)$

learning to parse mouse behavior from depth video







SVAE approach fit a latent switching linear dynamical system (SLDS) and a neural network image model for observations