

Challenges for deep learning

Yoshua Bengio

U. Montreal

November 6th, 2013

ICONIP'2013

Plenary talk

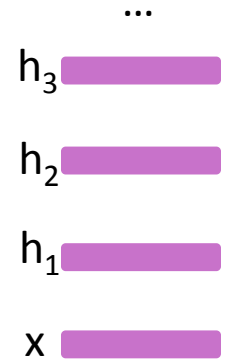


Ultimate Goals

- **AI**
- Needs **knowledge**
- Needs **learning**
- Needs **generalization**
- Needs ways to fight the curse of dimensionality
- Needs disentangling the underlying explanatory factors

Deep Representation Learning

Learn multiple levels of representation of increasing complexity/abstraction



- theory: exponential gain
- brains are deep
- cognition is compositional
- Better mixing (Bengio et al, ICML 2013)
- **They work! SOTA on industrial-scale AI tasks**
(object recognition, speech recognition, language modeling, music modeling)

Includes many approaches, not all neural, not just RBMs, many inspired by Fukushima 1980.

Google Image Search:

Different object types represented in the same space



Google:

S. Bengio, J.
Weston & N.
Usunier



(IJCAI 2011,
NIPS'2010,
JMLR 2010,
MLJ 2010)



$\Phi_I(\text{img})$

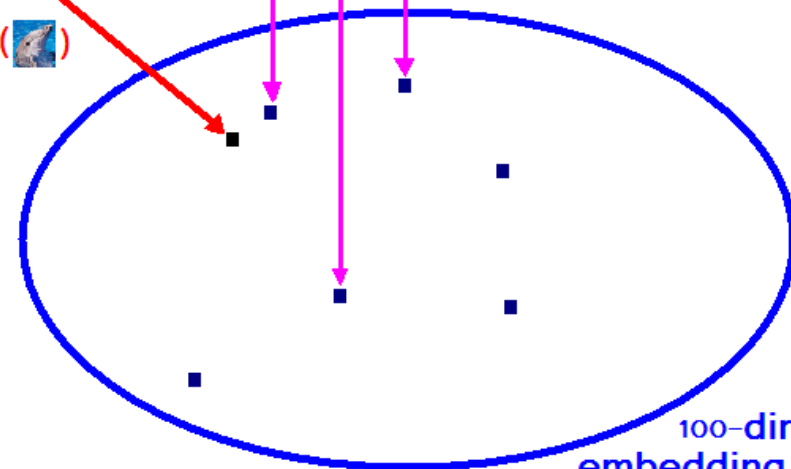
$\Phi_W(\text{DOLPHIN})$

DOLPHIN

OBAMA

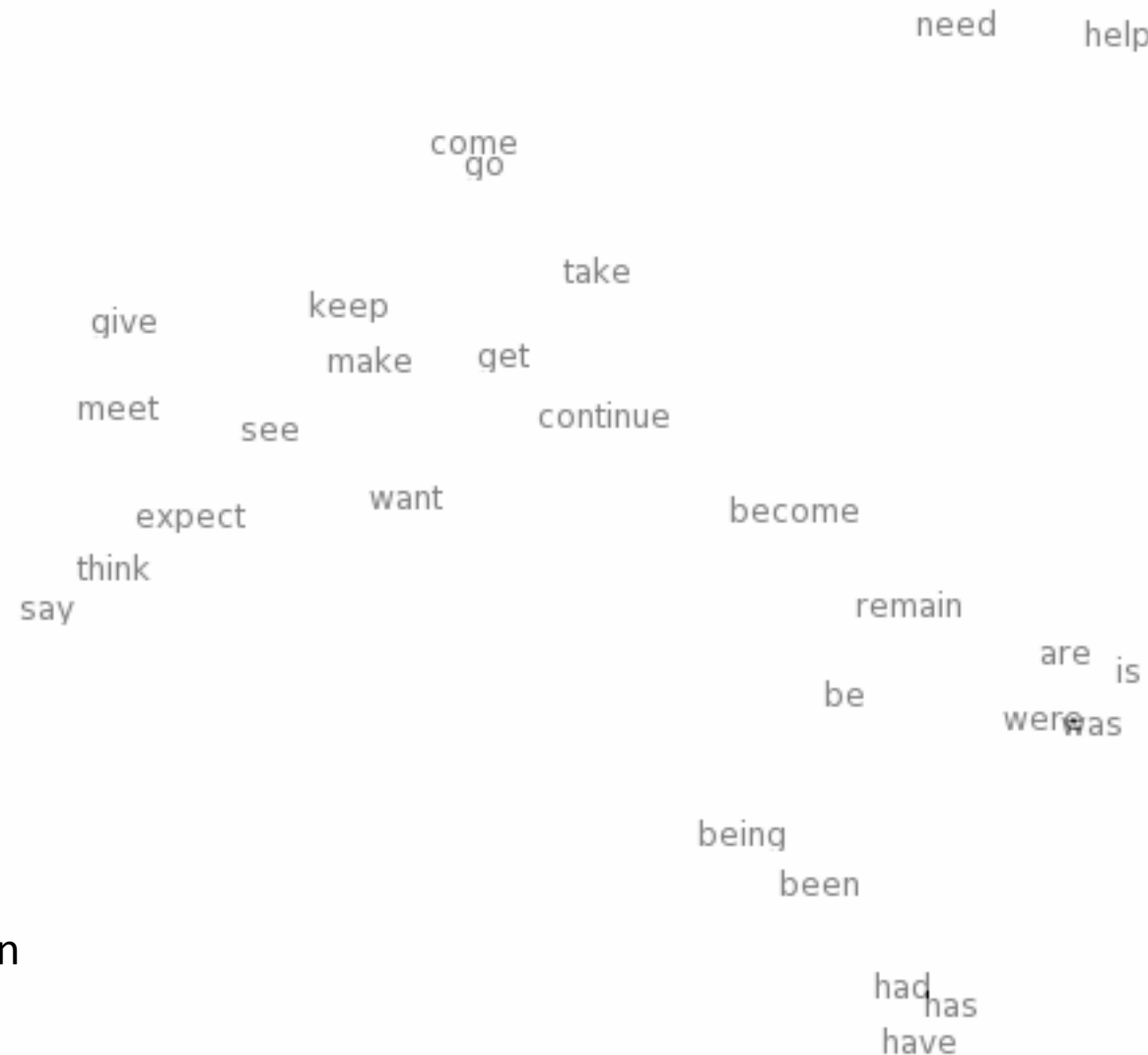
EIFFEL TOWER

.....



Learn $\Phi_I(\cdot)$ and $\Phi_W(\cdot)$ to optimize precision@k.

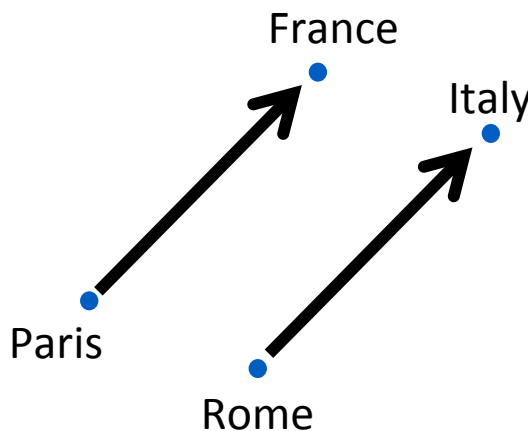
Following up on (Bengio et al NIPS'2000) Neural word embeddings



2-D visualization
by t-SNE

Analogical Representations for Free (Mikolov et al, ICLR 2013)

- Semantic relations appear as linear relationships in the space of learned representations
- King – Queen \approx Man – Woman
- Paris – France + Italy \approx Rome



Deep Learning in the News



Yoshua Bengio. Image: C

WIRED
Researcher Dreams Up
Machines That Learn
Without Humans
06.27.13



The New York Times

Monday, June 25, 2012 Last Update: 11:50 PM ET

SUBSCRIPTION: 4 WEEKS

Follow Us 

The New York Times Scientists See Promise in Deep-Learning Programs

John Markoff
November 23, 2012

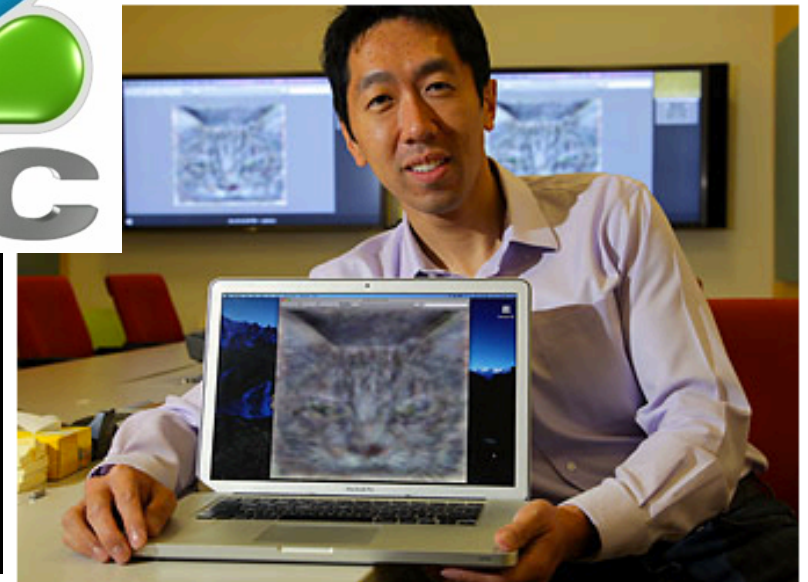
THE GLOBE AND MAIL
CANADA'S NATIONAL NEWSPAPER • FOUNDED 1844

Google taps U
of T professor
to teach
context to
computers
03.11.13



WIRED
The Man Behind the Google Brain: Andrew Ng
and the Quest for the New AI

BY DANIELA HERNANDEZ 05.07.13 6:30 AM



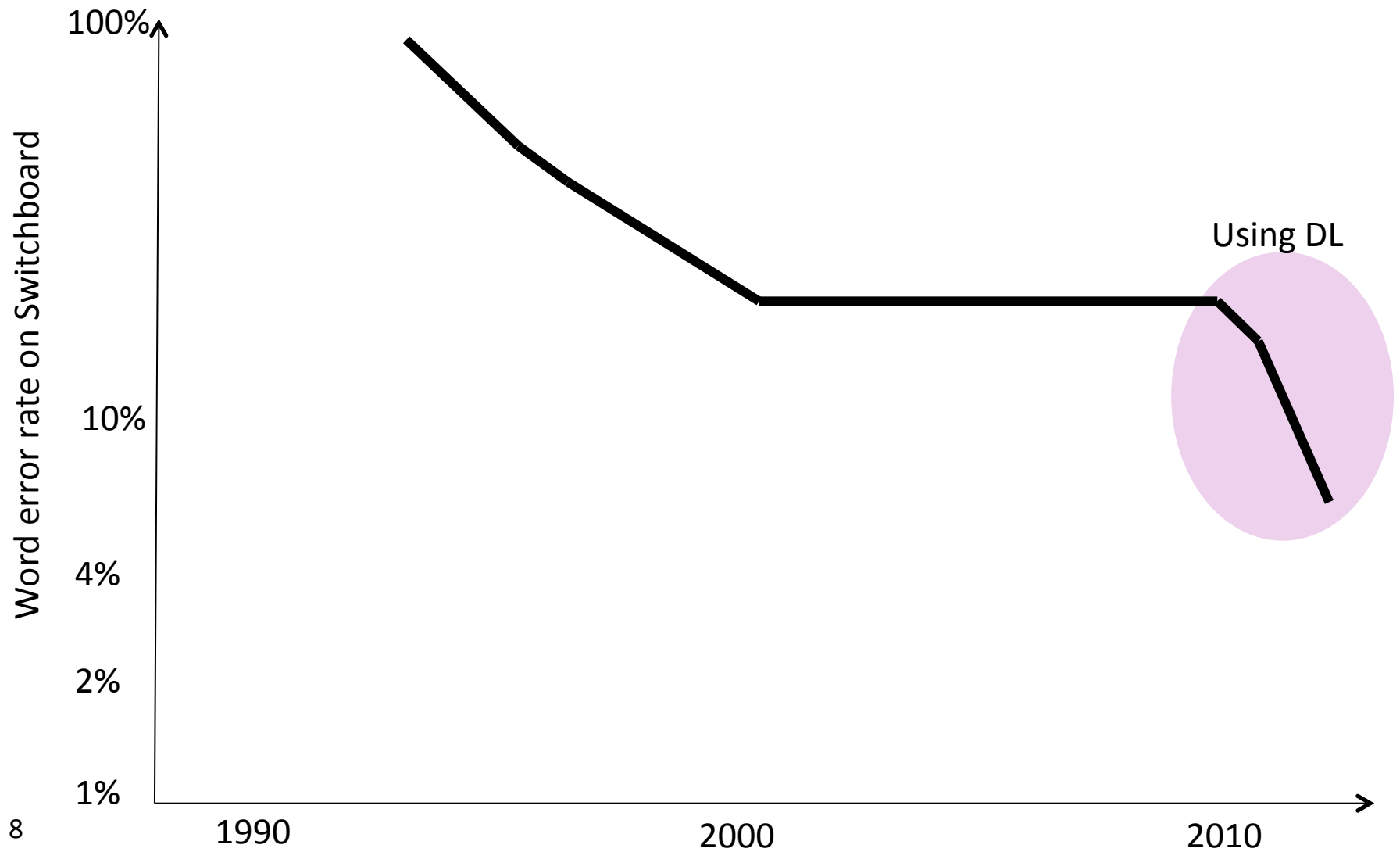
Jim Wilson/The New York Times

Despite Itself, a Simulated Brain Seeks Cats

By JOHN MARKOFF 12 minutes ago

A Google research team, led by Andrew Y. Ng, above, and Jeff Dean, created a neural network of 16,000 processors that reflected human obsession with Internet felines.

The dramatic impact of Deep Learning on Speech Recognition



10 BREAKTHROUGH TECHNOLOGIES 2013

Intr

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. →

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous. →

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child? →

Adv Man

Ske
prin
wor
mar
the
tech
jet p

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain

Smart Watches

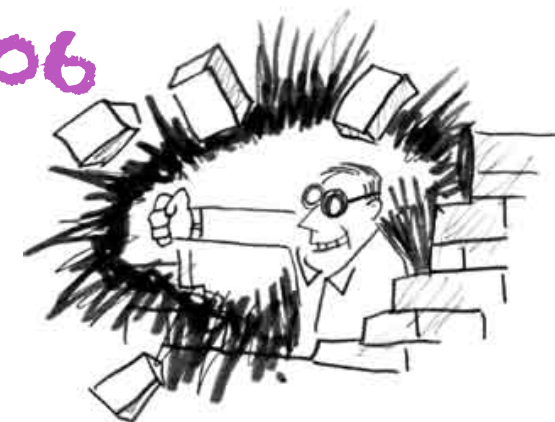
Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely

Big Pho

Coll
ana
from
pho

Major Breakthrough in 2006

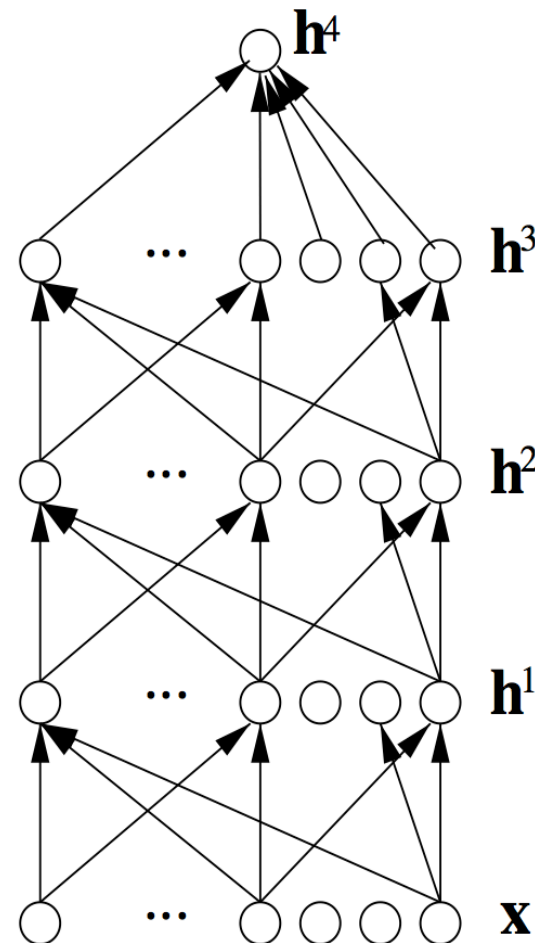


- Train deep nets by **layer-wise unsupervised pre-training**, whereas previous purely supervised attempts had failed
- Unsupervised feature learners:
 - RBMs
 - Auto-encoder variants
 - Sparse coding variants



Deep Supervised Neural Nets

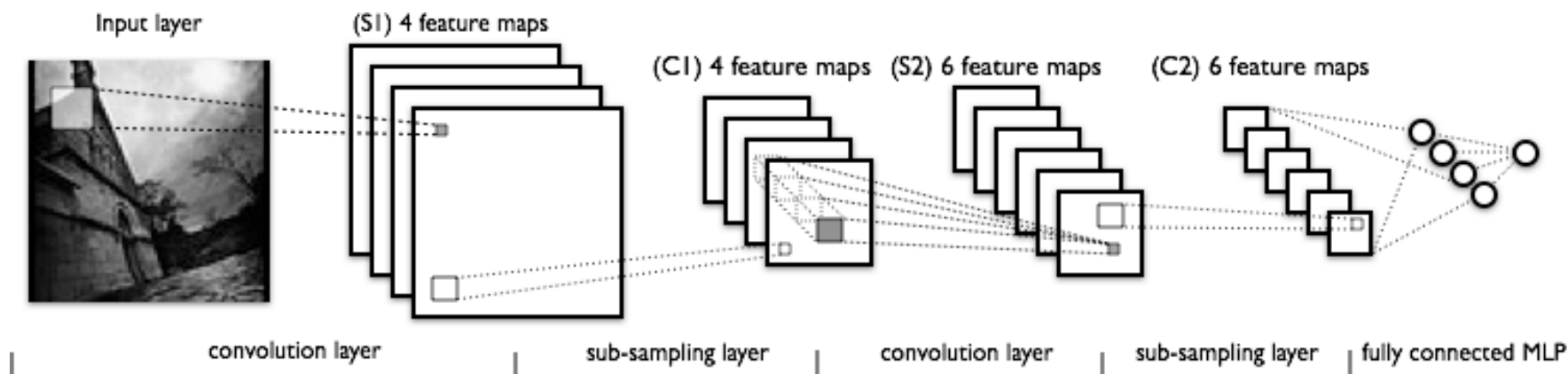
- Now can train them even without unsupervised pre-training:
better initialization and non-linearities (rectifiers, maxout), generalize well with large labeled sets and regularizers (dropout)
- **Unsupervised pre-training:**
rare classes, transfer, smaller labeled sets, or as extra regularizer.



Current State-of-the-Art for Pattern Recognition Applications

(Krizhevsky et al NIPS 2012; Goodfellow et al ICML 2013)

- Deep conv. nets
- Rectifiers or (even better) **maxout**
- Noise injection (denoising AE, dropout)
- Use unsupervised pre-training if small labeled set or rare classes
- **Deep** (5 to 8 layers) + good initialization
- Supervised fine-tuning

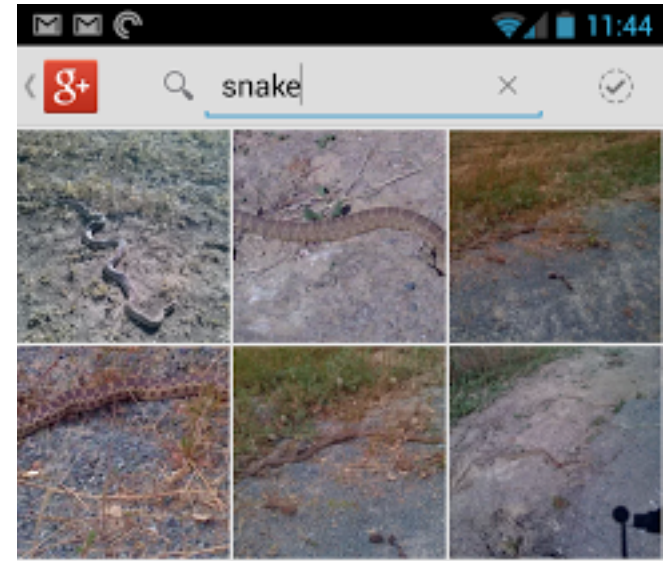


Industrial-scale object recognition

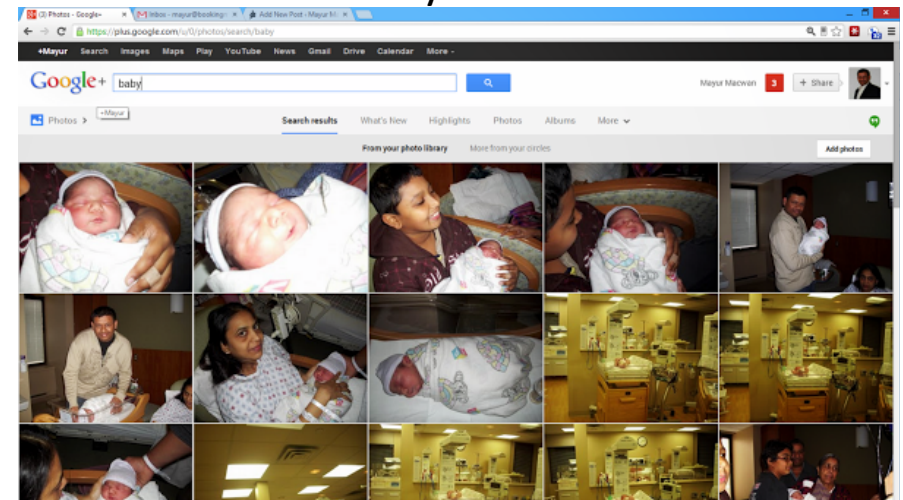
- Krizhevsky, Sutskever & Hinton NIPS 2012

	1 st choice	Top-5
2 nd best		27% err
Previous SOTA	45% err	26% err
Krizhevsky et al	37% err	15% err

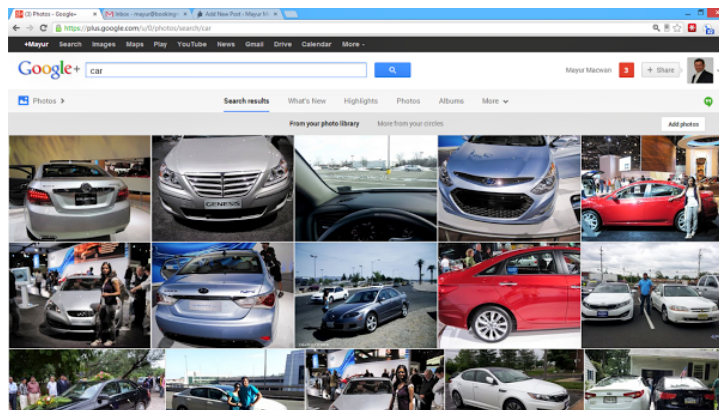
- Google incorporates DL in Google+ photo search, “A step across the semantic gap” (Google Research blog, June 12, 2013)
- Baidu now offers with similar services




baby



car



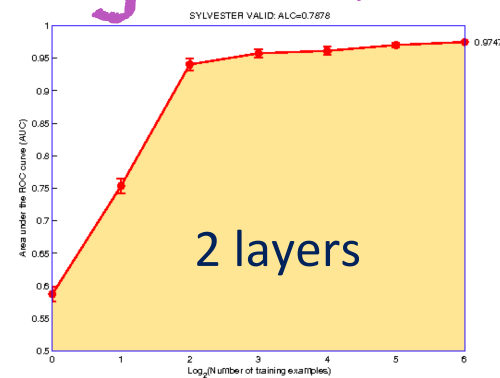
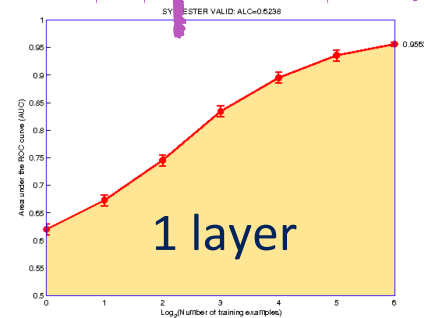
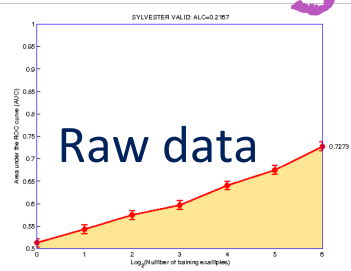
Deep Learning Tricks of the Trade

- Y. Bengio (2013), “Practical Recommendations for Gradient-Based Training of Deep Architectures”
 - Unsupervised pre-training
 - Stochastic gradient descent and setting learning rates
 - Main hyper-parameters
 - Learning rate schedule
 - Early stopping
 - Minibatches
 - Parameter initialization
 - Number of hidden units
 - L1 and L2 weight decay
 - Sparsity regularization
 - Debugging
 - How to efficiently search for hyper-parameter configurations

How do humans generalize from very few examples?

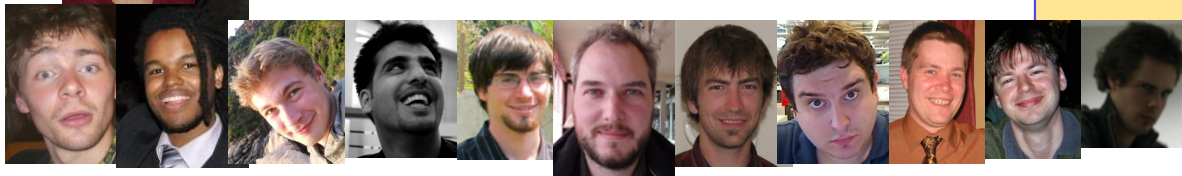
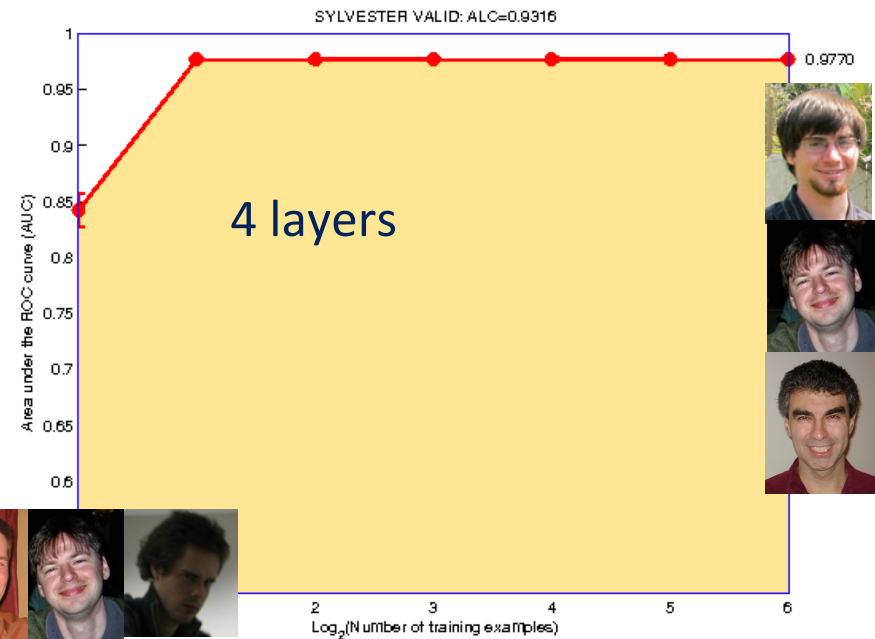
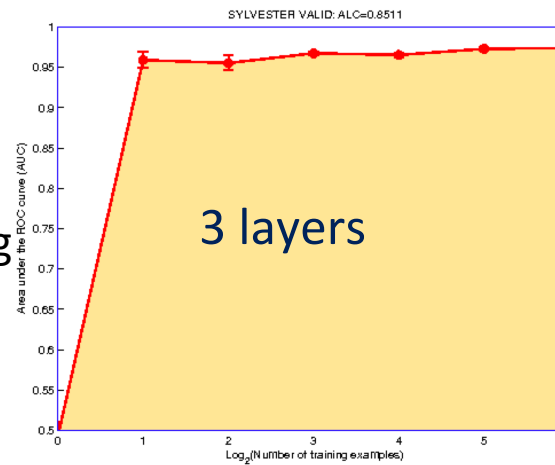
- They **transfer** knowledge from previous learning:
 - **Abstract** (i.e. deep) representations
 - Explanatory factors

Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Deep Learning 1st Place



NIPS'2011
Transfer
Learning
Challenge
Paper:
ICML'2012

ICML'2011
workshop on
Unsup. &
Transfer Learning



Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Disentangling Factors of Variation
- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Reasoning & One-Shot Learning of Facts

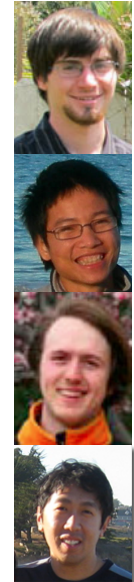
Challenge: Disentangling

- Invariant features
- Which invariances?
- Alternative: learning to disentangle factors
- Good disentangling →
avoid the curse of dimensionality



Emergence of Disentangling

- (Goodfellow et al. 2009)
- (Glorot et al. 2011)
- different features specialize on different aspects



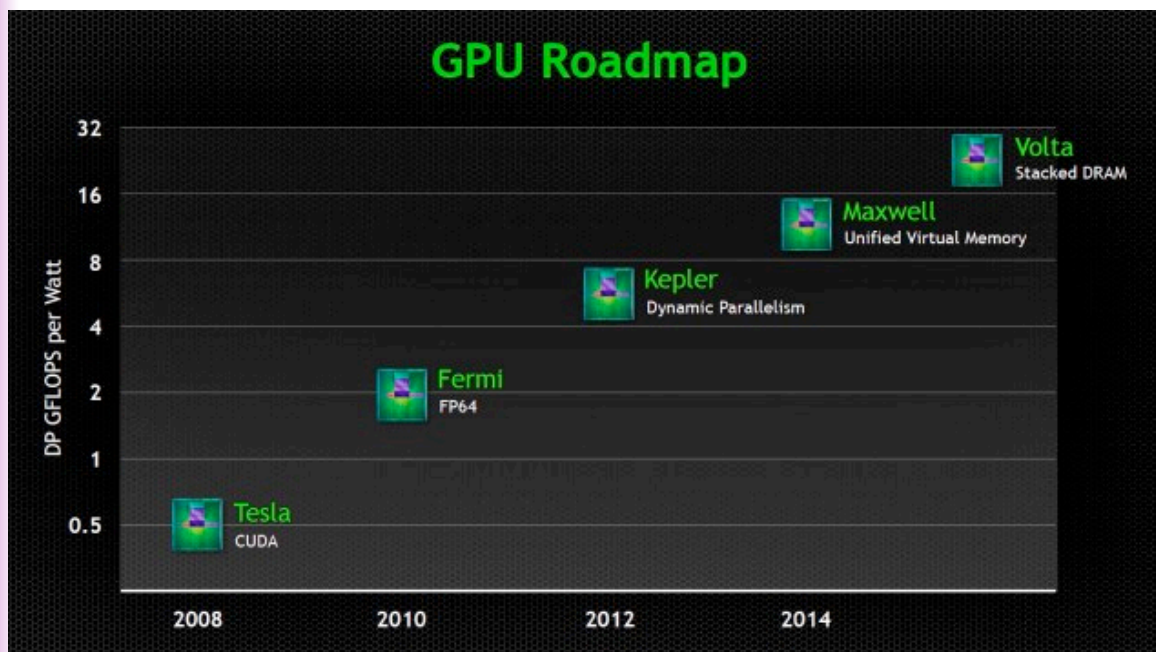
WHY?

Broad Priors as Hints to Disentangle the Factors of Variation

- **Multiple factors**: distributed representations
- Multiple levels of abstraction: **depth**
- **Semi-supervised** learning: Y is one of the factors explaining X
- **Multi-task** learning: different tasks share some factors
- **Manifold** hypothesis: probability mass concentration
- Natural **clustering**: class = manifold, well-separated manifolds
- Temporal and spatial **coherence**
- **Sparsity**: most factors irrelevant for particular X
- **Simplicity of factor dependencies** (in the right representation)

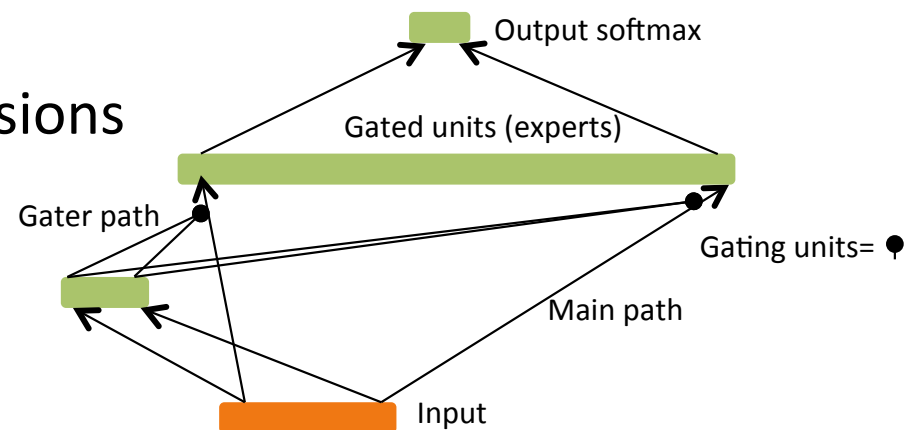
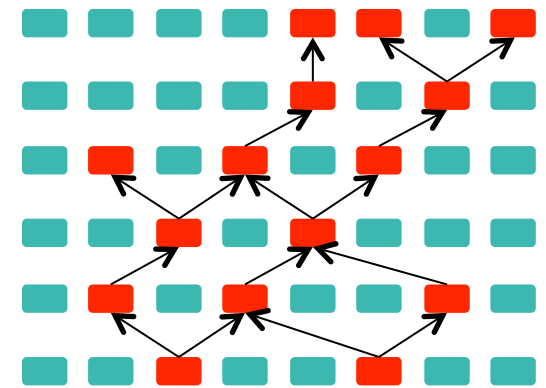
Challenge: Computational Scaling

- Recent breakthroughs in speech, object recognition and NLP hinged on faster computing, GPUs, and large datasets
- A 100-fold speedup is possible without waiting another 10yrs?
 - Challenge of distributed training
 - Challenge of conditional computation



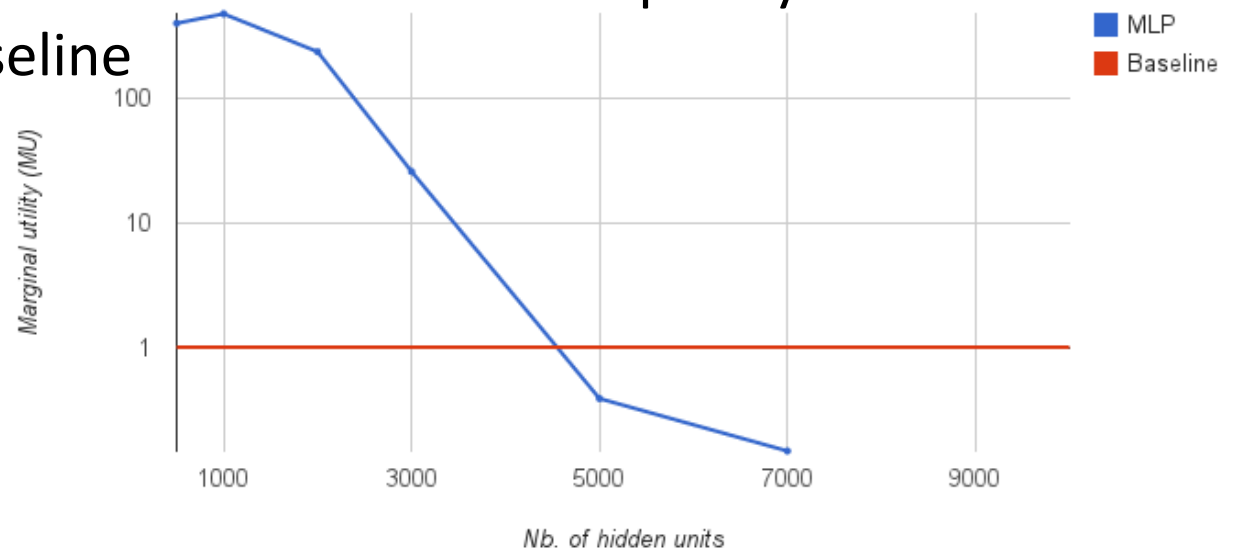
Conditional Computation: only visit a small fraction of parameters / example

- Deep nets vs decision trees
- sparse distributed gaters selecting combinatorial subsets of a deep net
- Challenges:
 - Back-prop through hard decisions
 - Architecture?



Challenge: Optimization & Underfitting

- On large datasets, major obstacle is underfitting
- **Marginal utility** of wider MLPs decreases quickly below memorization baseline



- Current limitations: local minima or ill-conditioning?
- Adaptive learning rates and stochastic 2nd order methods
- Conditional comp. & sparse gradients → better conditioning: when some gradients are 0, many cross-derivatives are also 0.

Challenge: Distributed Training

- Minibatches (too large = slow down)
- Large minibatches + 2nd order methods
- **Asynchronous SGD** (Bengio et al 2003, Le et al ICML 2012, Dean et al NIPS 2012)
 - Bottleneck: communicating weights between nodes
- New ideas:
 - Low-resolution sharing only where needed
 - Specialized conditional computation

Basic Challenge with Probabilistic Models: marginalization

- Joint and marginal likelihoods involve intractable sums over configurations of random variables (inputs x , latent h), e.g.

$$P(x) = \sum_h P(x, h)$$

$$P(x, h) = e^{-\text{energy}(x, h)} / Z$$

$$Z = \sum_{x, h} e^{-\text{energy}(x, h)}$$

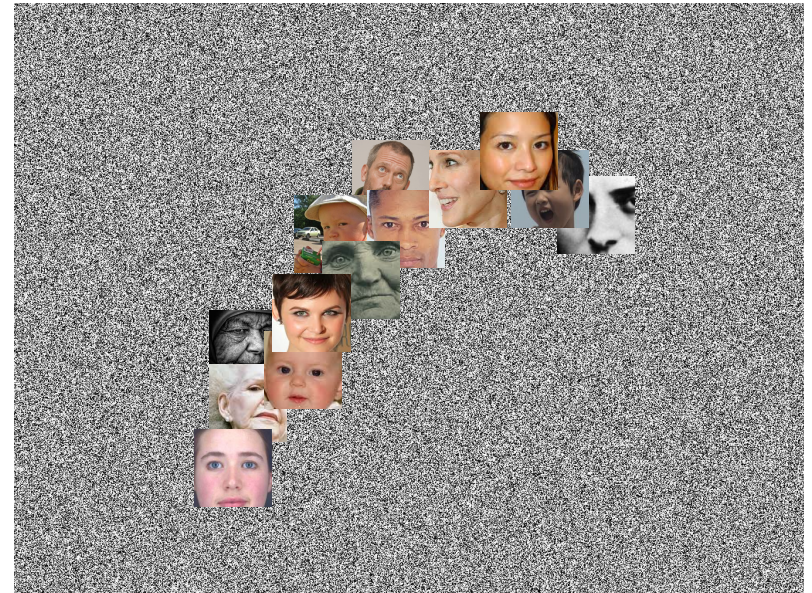
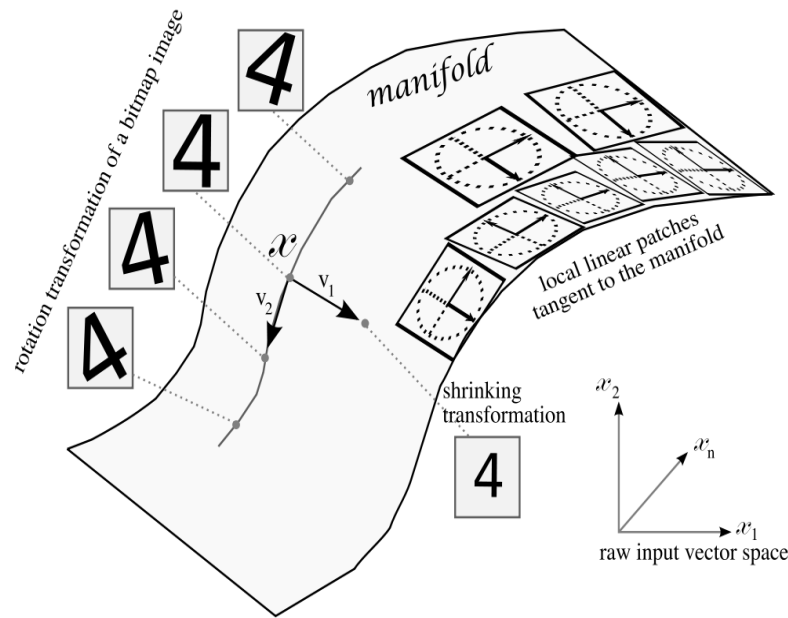
- MCMC methods approximate these sums

Two Fundamental Problems with Probabilistic Models with Many Random Variables

1. MCMC mixing between modes
(manifold hypothesis)
2. Many non-negligible modes
(both in posterior & joint distributions)

For AI Tasks: Manifold structure

- examples **concentrate** near a lower dimensional “manifold”
- Evidence: most input configurations are unlikely



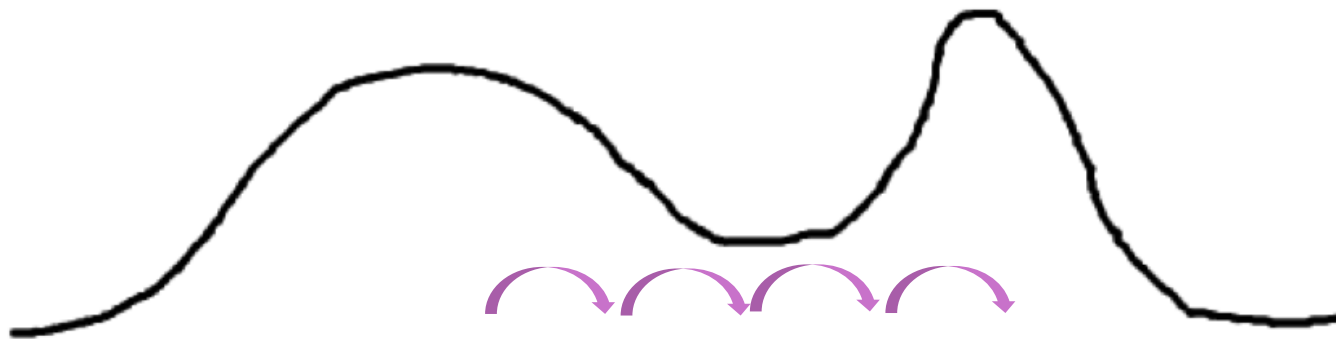
Mixing Between Well-Separated Modes is Fundamentally Hard

- MCMC steps local
- Chances of going from manifold A to manifold B = prob. accepting a long string of improbable moves = exponentially small

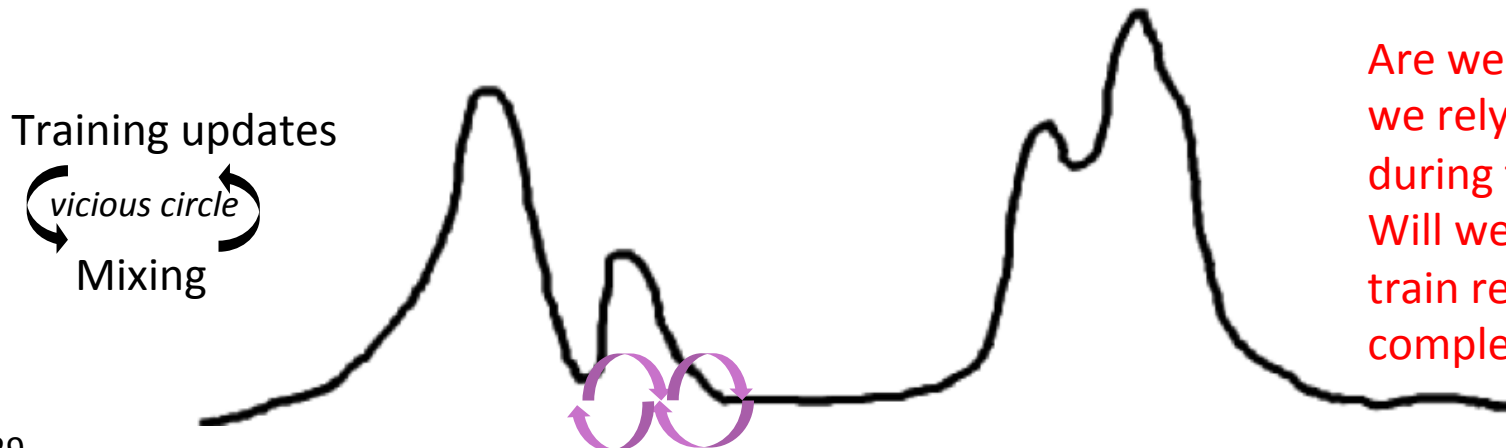


Mixing Between Modes: Vicious Circle Between Learning and MCMC Sampling

- Early during training, density smeared out, mode bumps overlap



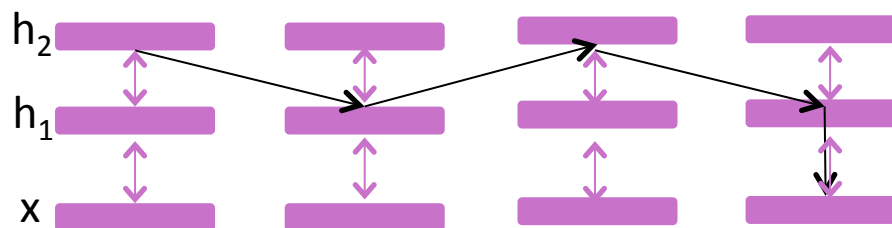
- Later on, hard to cross empty voids between modes



Poor Mixing: Depth to the Rescue

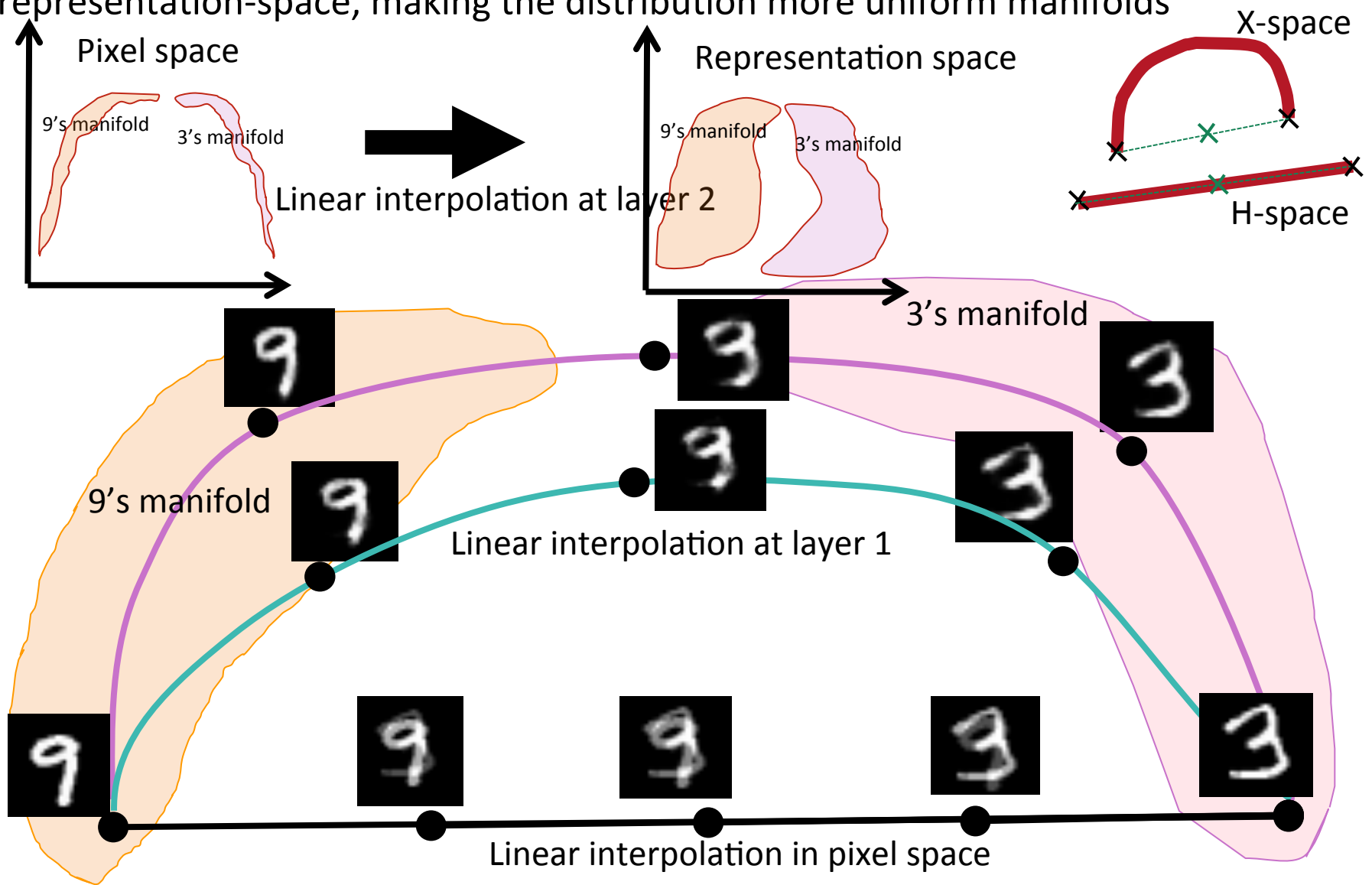
(Bengio et al ICML 2013)

- MCMC at top level visit more modes (classes) faster! **WHY?**



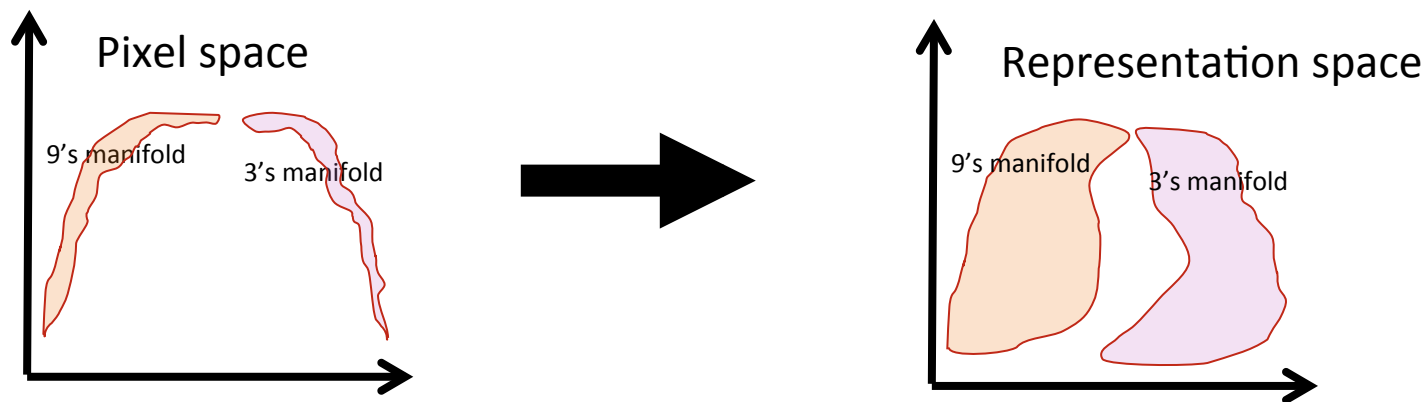
Space-Filling in Representation-Space

High-probability samples fill space between them when viewed in the learned representation-space, making the distribution more uniform manifolds



Poor Mixing: Depth to the Rescue

- **Deeper representations → abstractions → disentangling**
- E.g. reverse video bit, class bits in learned representations: easy to Gibbs sample between modes at abstract level
- Some units may directly indicate manifold
 - Easier mixing between modes



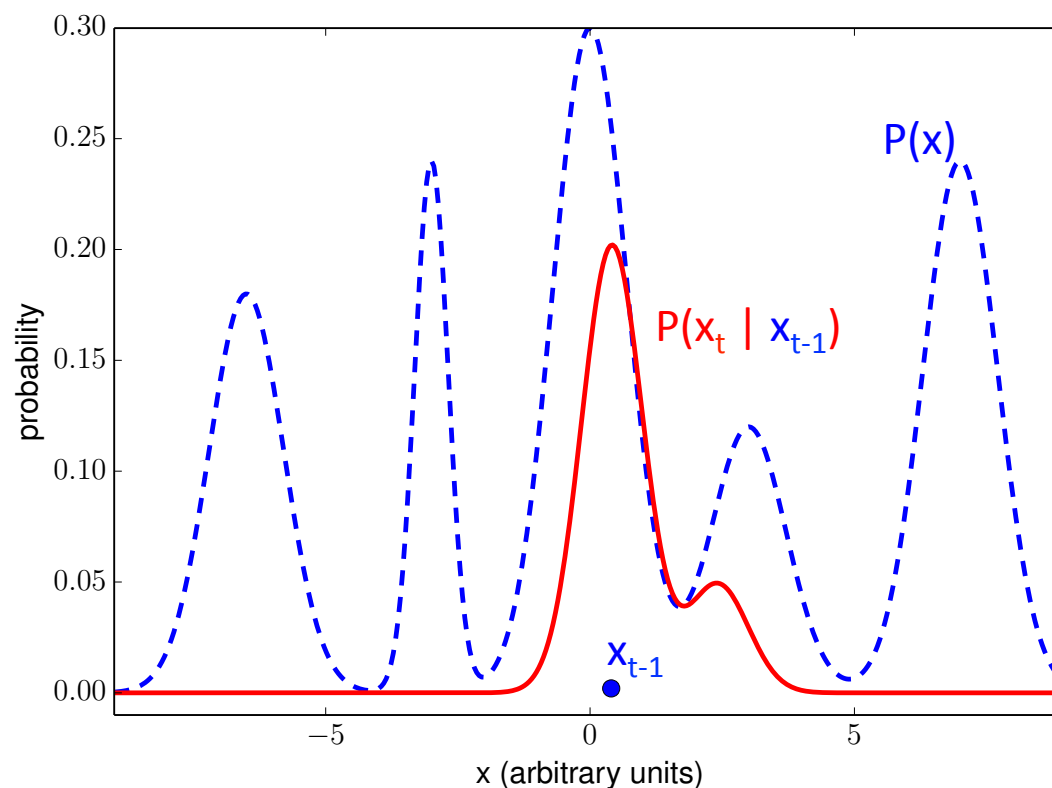
Potentially Huge Number of Modes in the Posterior $P(h|x)$

- **Human** hears foreign speech x , y =answer to question:
 - 10 word segments
 - 100 plausible candidates per word
 - 10^6 possible segmentations
 - Most configurations (999999/1000000) implausible
 - ➔ 10^{20} high-probability modes
- Humans probably don't consider all these in their mind
- **All known approximate inference scheme break down if the posterior has a huge number of modes** (fails MAP & MCMC) and not respecting a variational approximation (fails variational)



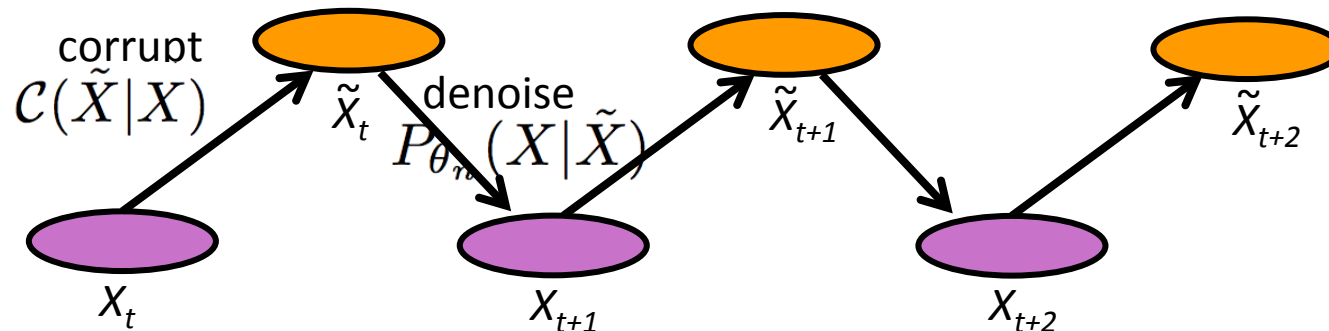
Instead of Learning $P(x)$ directly, Learn Markov chain operator $P(x_t | x_{t-1})$

- $P(x)$ may have many modes, making the normalization constant intractable, and MCMC approximations poor
- $P(x_t | x_{t-1})$ could be much simpler because most of the time a local move, might even be well approximated by unimodal



How to train the transition operator?

- One solution was recently discovered, based on the denoising auto-encoder research
- The transition operator is decomposed in two steps:
 - Corruption process $\mathcal{C}(\tilde{X}|X)$
 - Reconstruction (denoising) distribution $P_{\theta_n}(X|\tilde{X})$
- The parameters can be trained by maximum likelihood over the pairs \tilde{X}, X

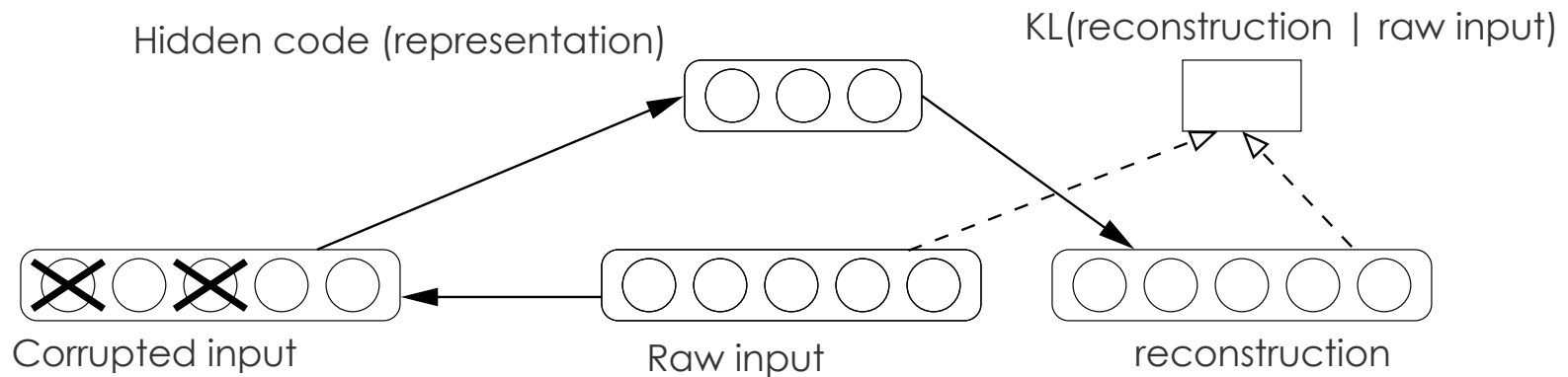


Denoising Auto-Encoder

(Vincent et al 2008)



- Corrupt the input during training only
- Train to reconstruct the uncorrupted input

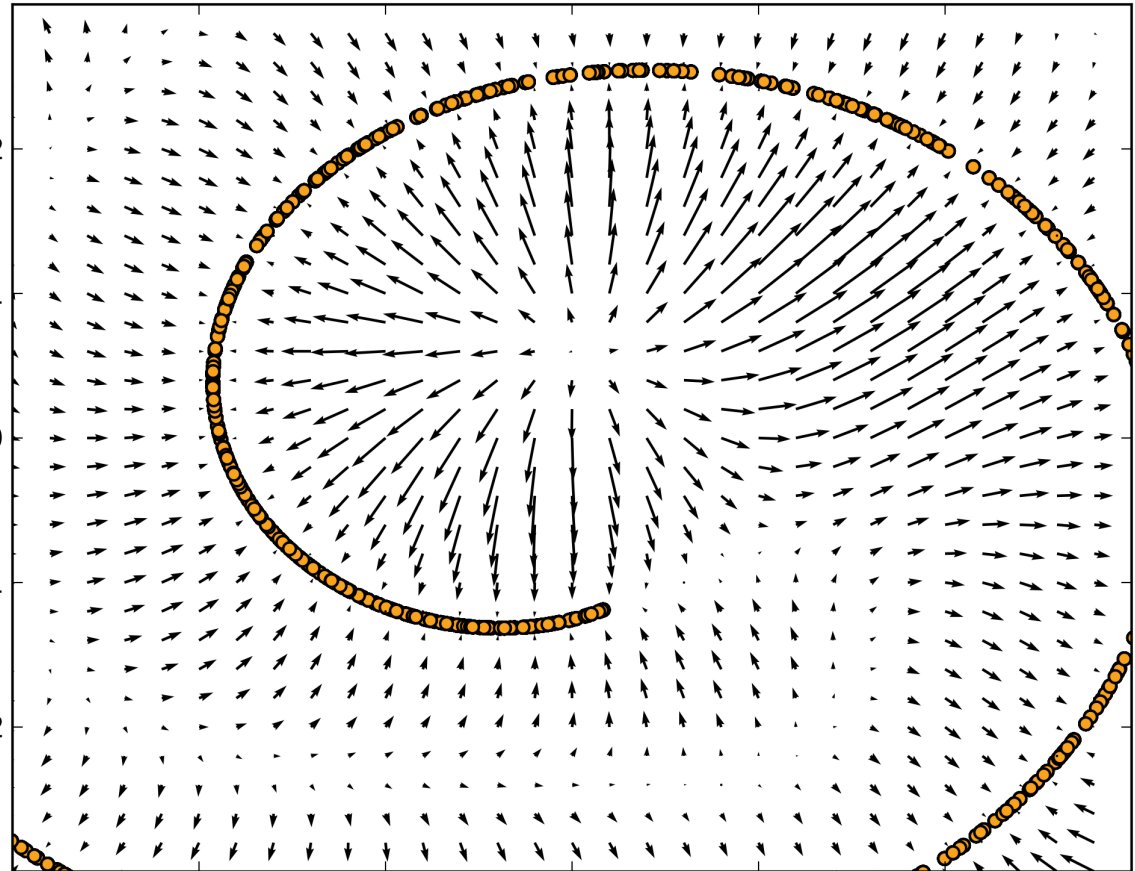


- Encoder & decoder: any parametrization
- As good or better than RBMs for unsupervised pre-training

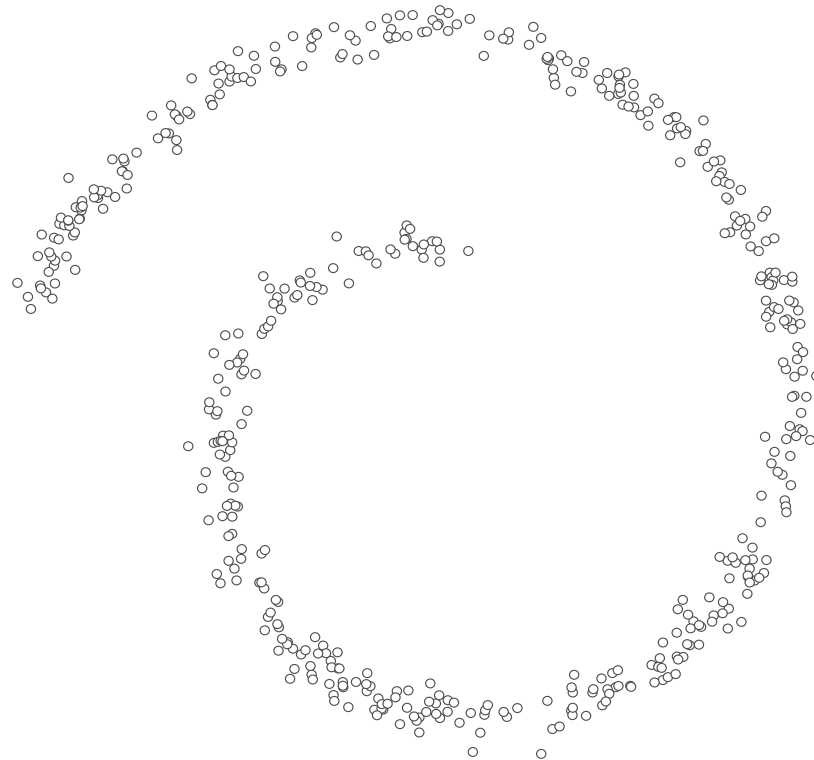
Learning a Vector Field that Estimates a Gradient Field

- Continuous inputs
- Gaussian corruption
- Squared error
- **Reconstruction(x)-x** estimates $d\log p(x)/dx$
- Zero reconstruction error could be either local min or local max of density

Reconstruction(x)-x

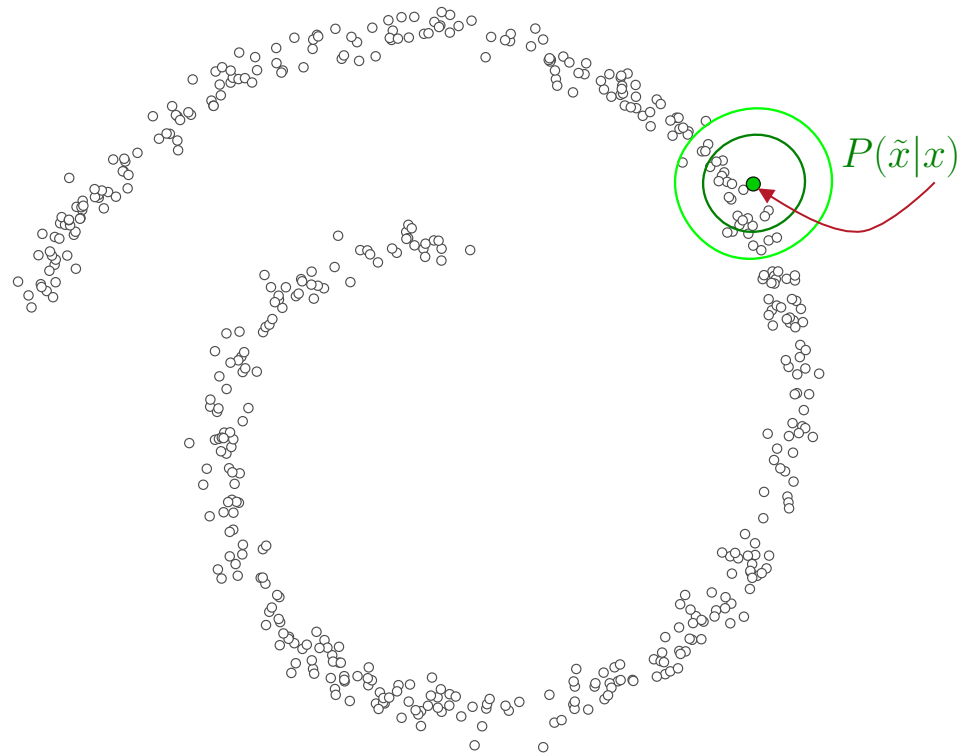


Learning with a simpler normalization constant, a nearly unimodal conditional distribution instead of a complicated multimodal one



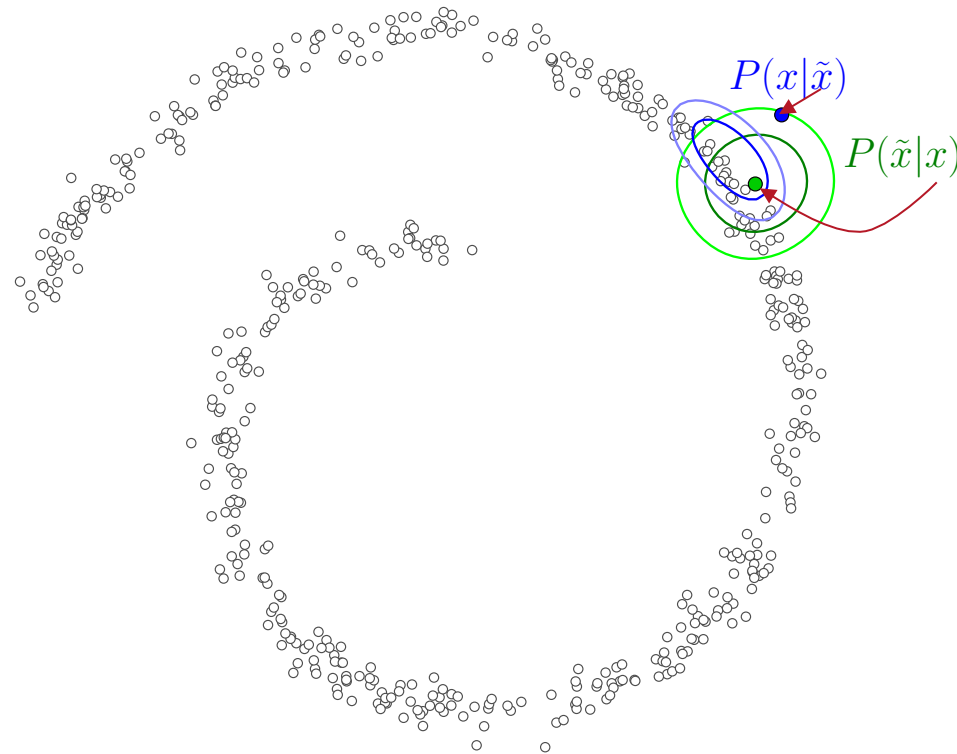
Thanks:
Jason Yosinski

Learning with a simpler normalization constant, a nearly unimodal conditional distribution instead of a complicated multimodal one



Thanks:
Jason Yosinski

Learning with a simpler normalization constant, a nearly unimodal conditional distribution instead of a complicated multimodal one

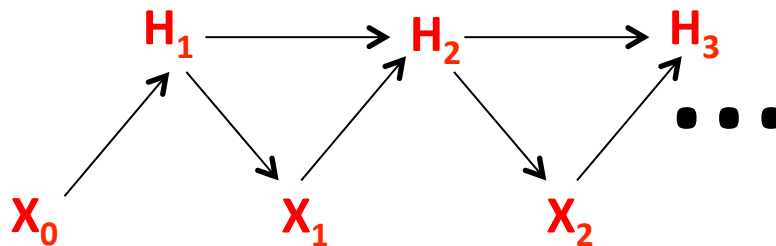


Thanks:
Jason Yosinski

Generative Stochastic Networks

- Generalizes the denoising auto-encoder training scheme
 - Introduce latent variables in the Markov chain (over X, H)
 - Instead of a fixed corruption process, have a deterministic function with parameters θ_1 and a noise source Z as input

$$H_{t+1} = f_{\theta_1}(X_t, Z_t, H_t)$$



$$H_{t+1} \sim P_{\theta_1}(H|H_t, X_t)$$

$$X_{t+1} \sim P_{\theta_2}(X|H_{t+1})$$

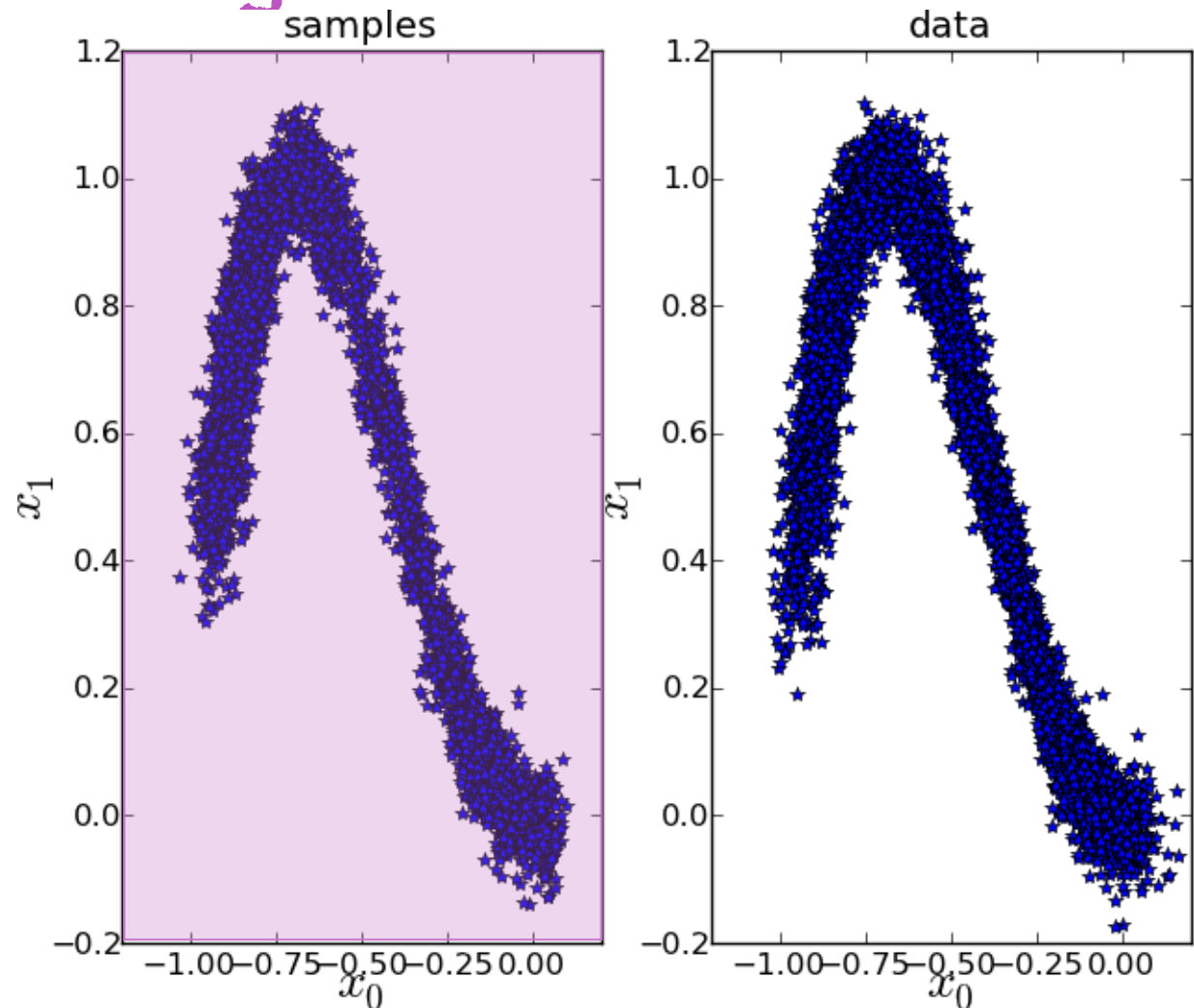
Consistent Estimator Theorem

Theorem:

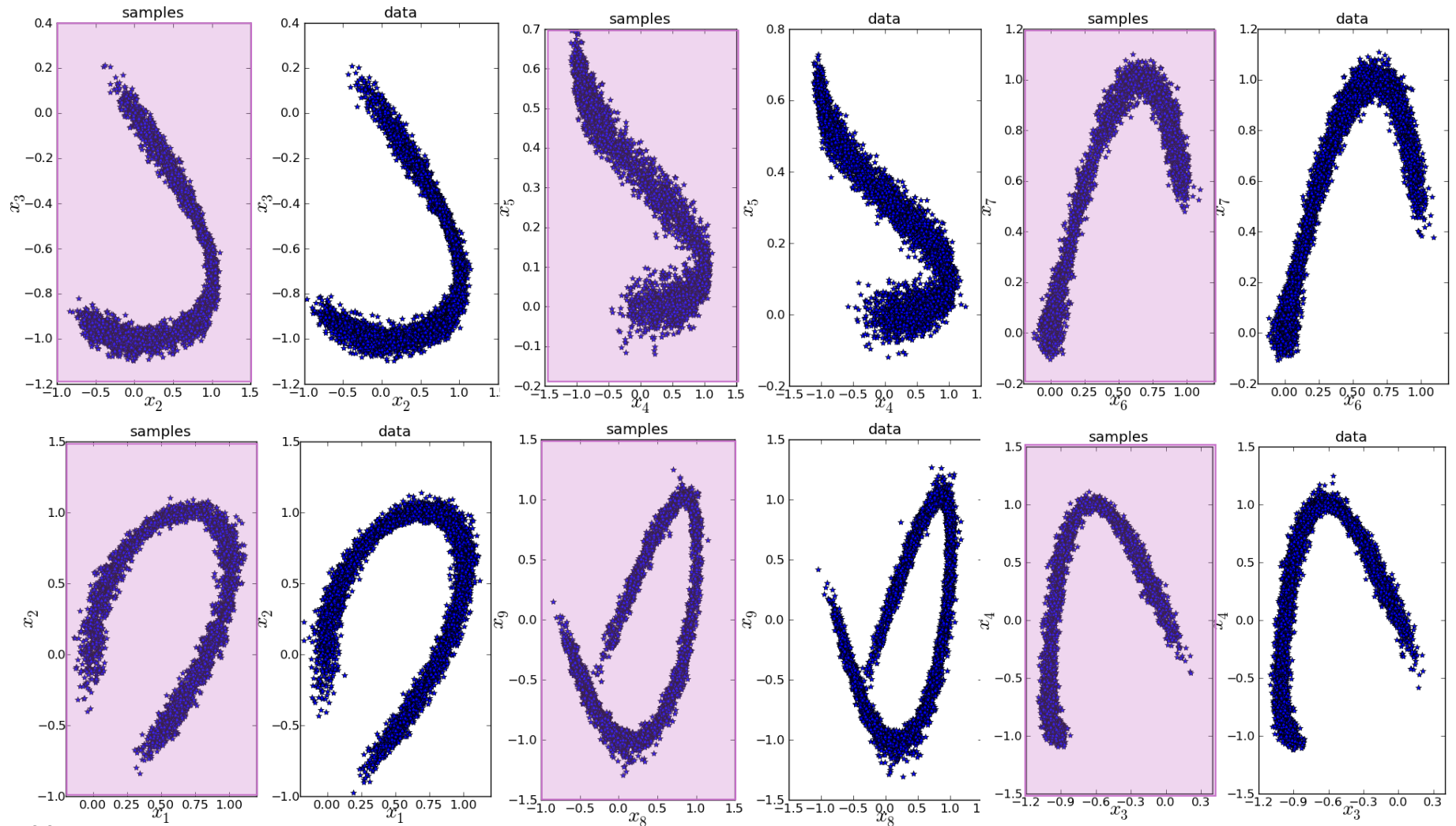
If the parametrization is rich enough to have $P(X|H)$ a consistent estimator and the Markov chain is ergodic, then maximizing the expected log of $P_{\theta_2}(X|f_{\theta_1}(X, Z_{t-1}, H_{t-1}))$ makes the stationary distribution of the Markov chain a consistent estimator of the true data generating distribution.

GSN Experiments: validating the theorem in a continuous non-parametric setting

- Continuous data, X in R^{10} , Gaussian corruption
- Reconstruction distribution = Parzen (mixture of Gaussians) estimator
- 5000 training examples, 5000 samples
- Visualize a pair of dimensions

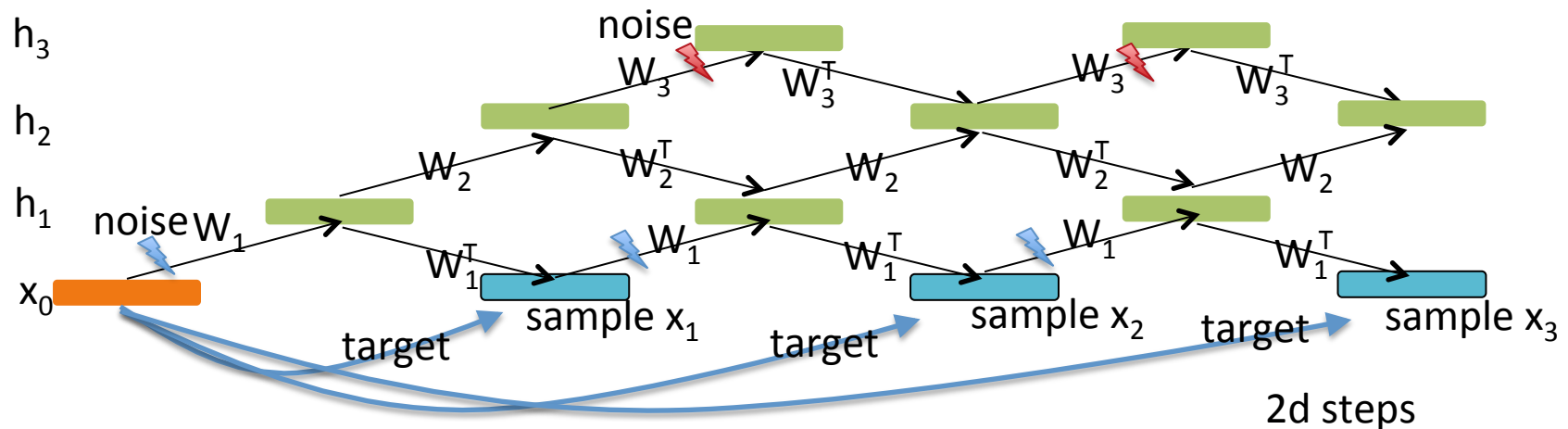


GSN Experiments: validating the theorem in a continuous non-parametric setting



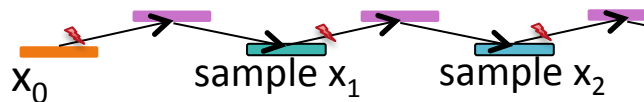
GSN Emulating a Deep Boltzmann Machine, TRAINED BY BACK-PROP!

- Noise injected in input and hidden layers
- Trained to max. reconstruction prob. of example at each step
- Structure inspired from the DBM Gibbs chain:

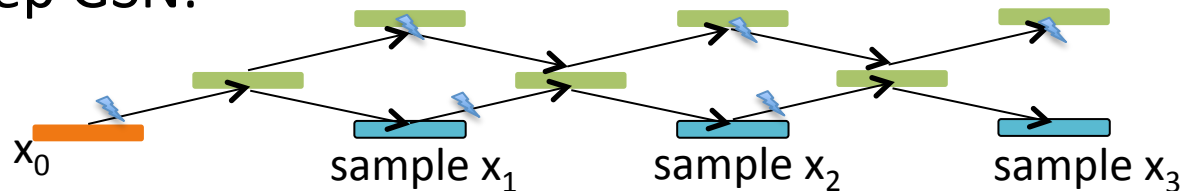


Experiments: Shallow vs Deep

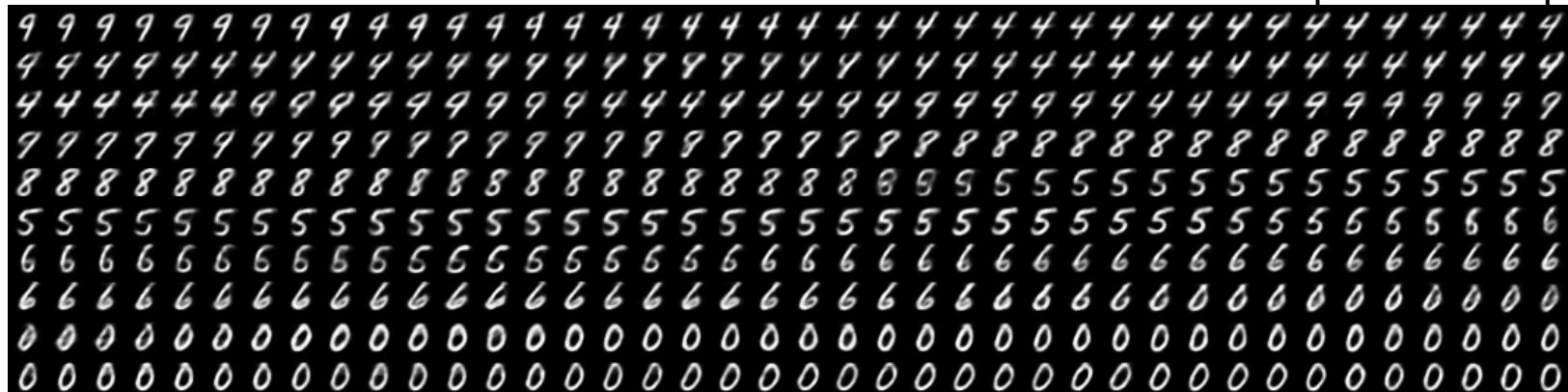
- Shallow (DAE), no recurrent path at higher levels, state=X only



- Deep GSN:



Better compromise between mixing and spurious samples



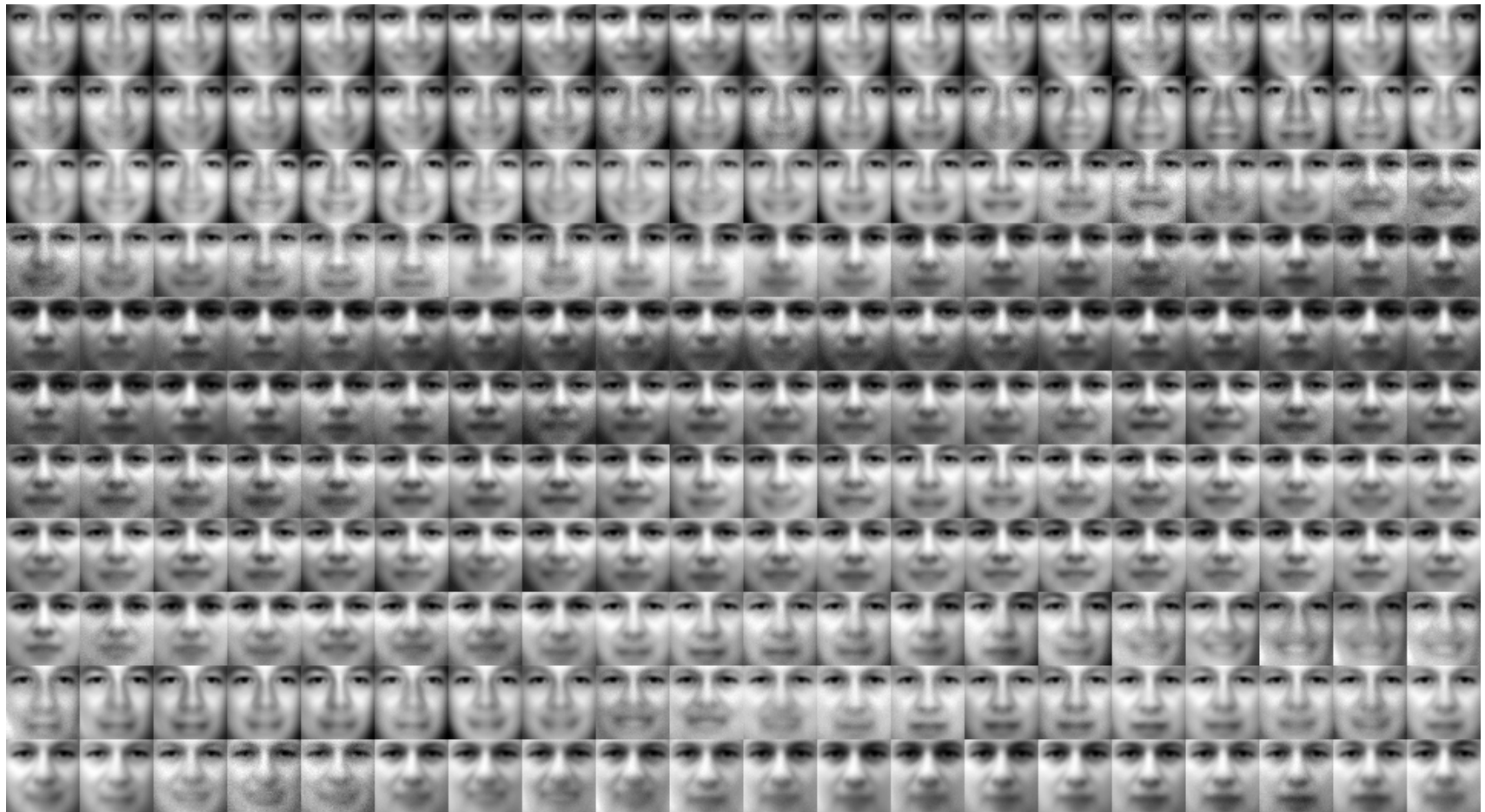
Experiments: Structured Conditionals

- Stochastically fill-in missing inputs, sampling from the chain that generates the conditional distribution of the missing inputs given the observed ones (notice the fast burn-in!)



Not Just MNIST: experiments on TFD

- 3 hidden layer model, consecutive samples:



Conclusions

- Several important challenges ahead for deep learning: computational scaling, numerical optimization, and marginalization, all important for the final goal of disentangling the underlying factors of variation
- **GSN: radically different approach to probabilistic unsupervised learning of generative models through learning a transition operator**
 - Can address mode mixing with depth (deep representation)
 - Avoid marginalization during training
- Consistent estimator
- Can be used to handle missing inputs or structured outputs
- Easy to train and sample from, hard to compute $P(x)$

LISA team: **Merci! Questions?**



LISA team: **Merci! Questions?**

