Modeling Natural Images with Higher-Order Boltzmann Machines

Marc'Aurelio Ranzato

Department of Computer Science – Univ. of Toronto

ranzato@cs.toronto.edu

joint work with Geoffrey Hinton and Vlad Mnih

CIFAR summer school, 12 August 2010

How to model natural images?

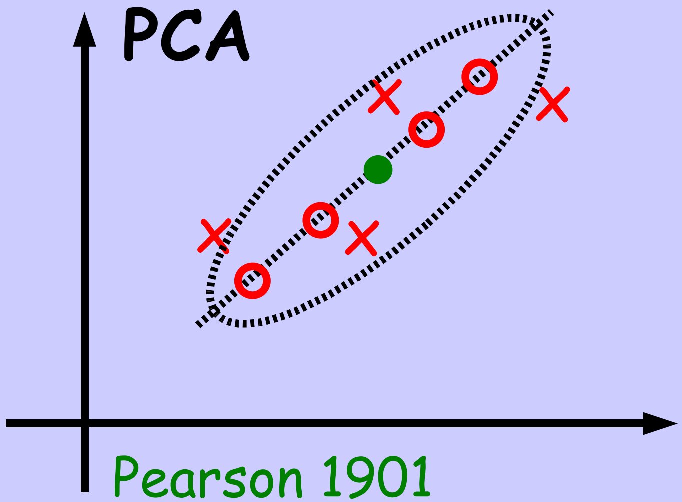
$p(v,h)$

v: pixels

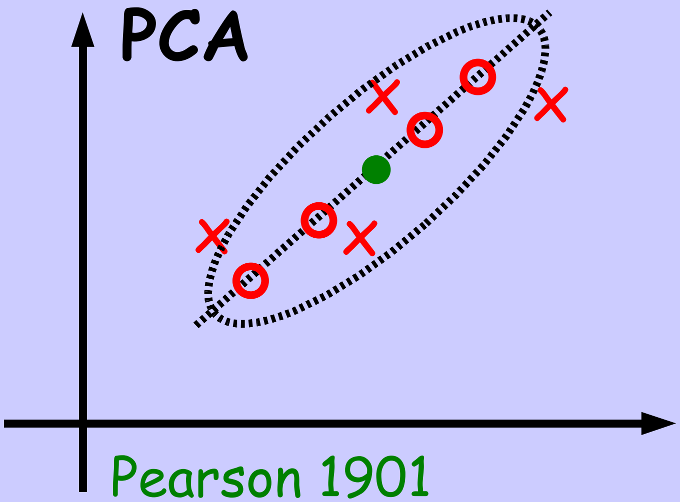
h: latent features

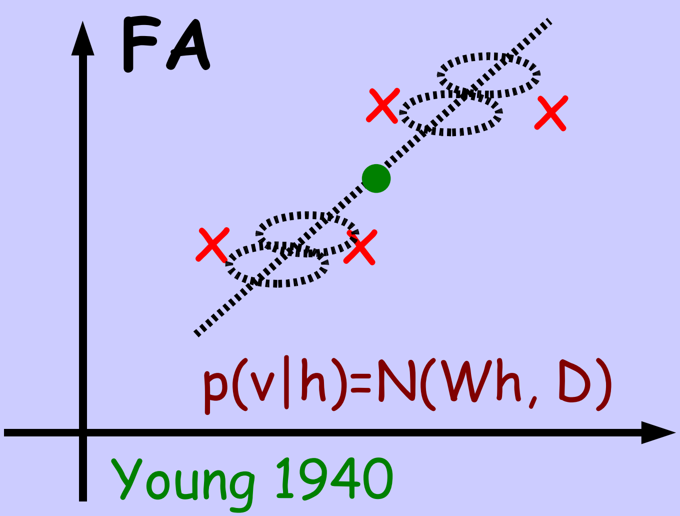
image space v

A generative perspective: p(v|h)

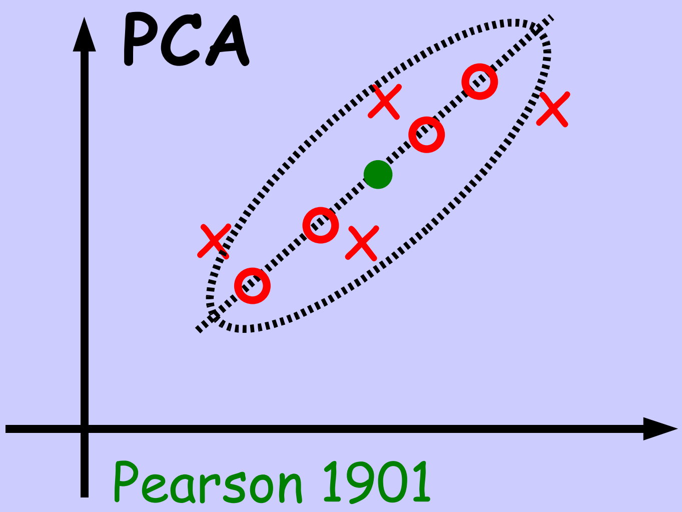


A generative perspective: $p(v|h)$

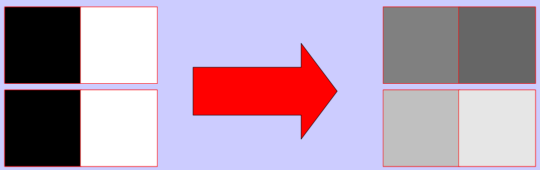




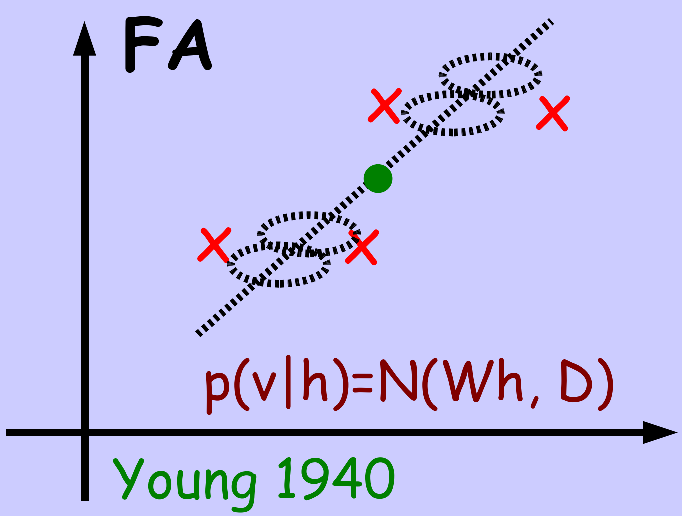
A generative perspective: p(v|h)



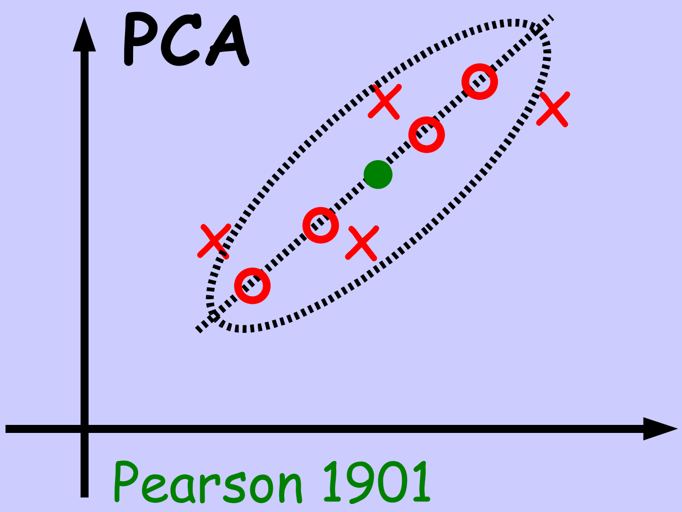
mean sample ~ p(v|h)

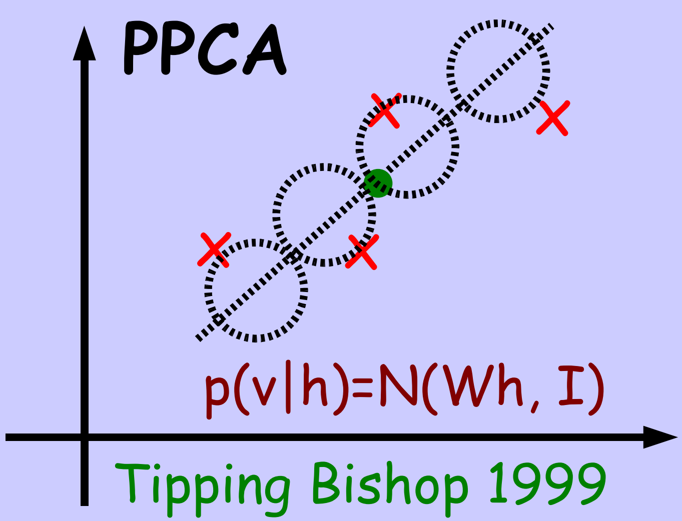


structure is lost!

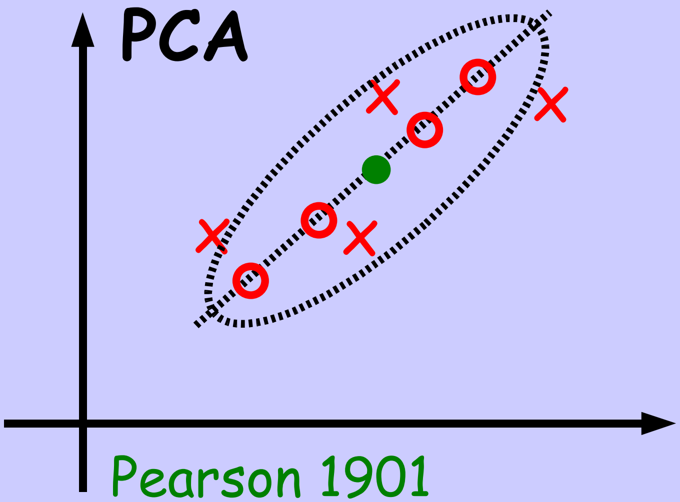


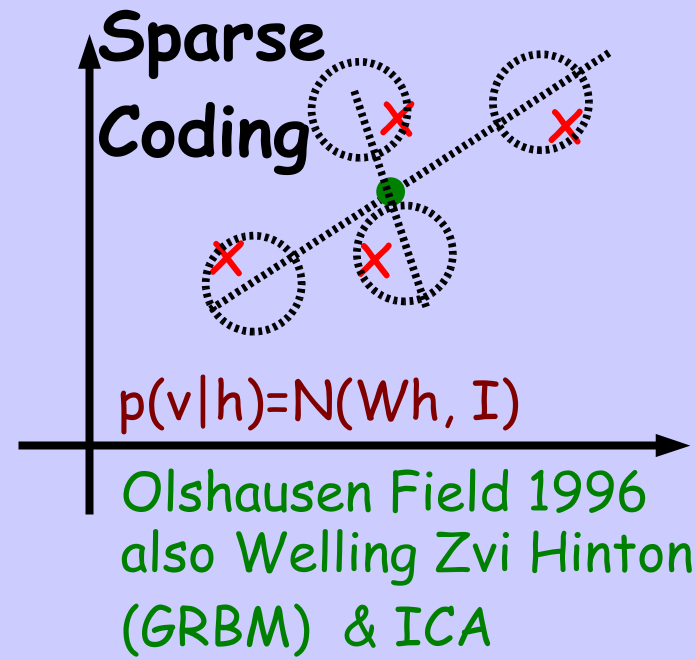
A generative perspective: p(v|h)

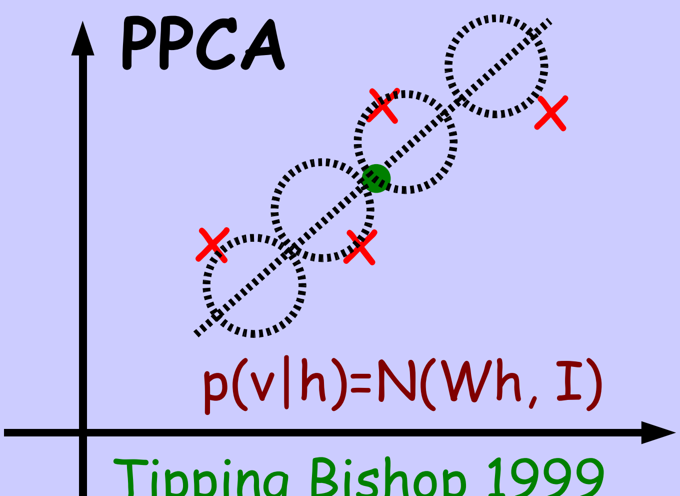




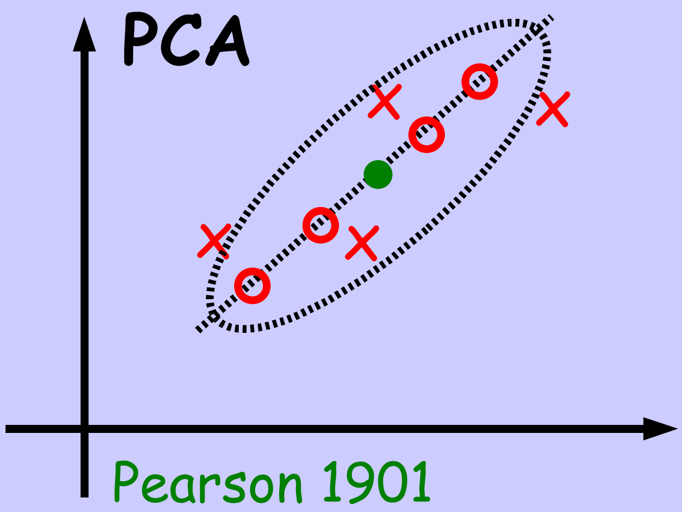
A generative perspective: p(v|h)

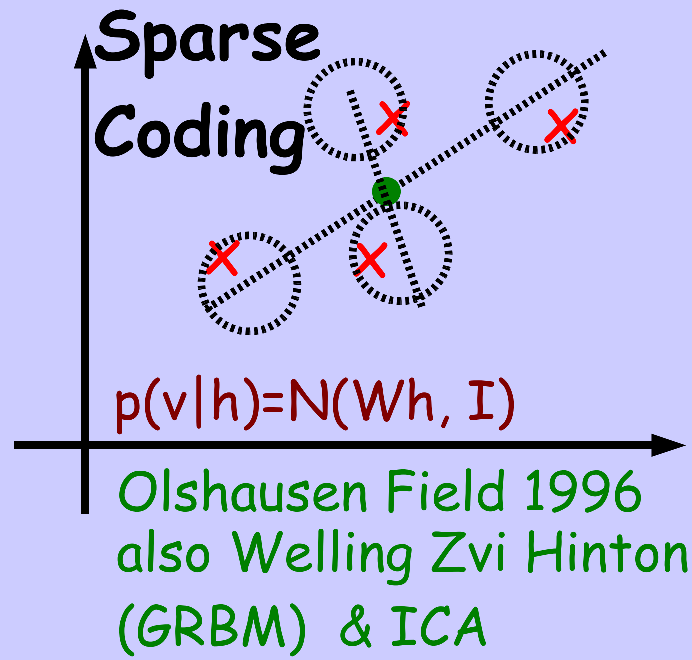


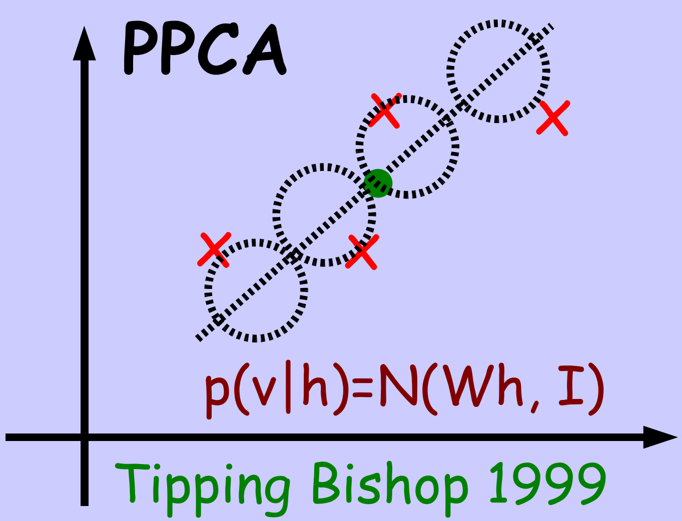


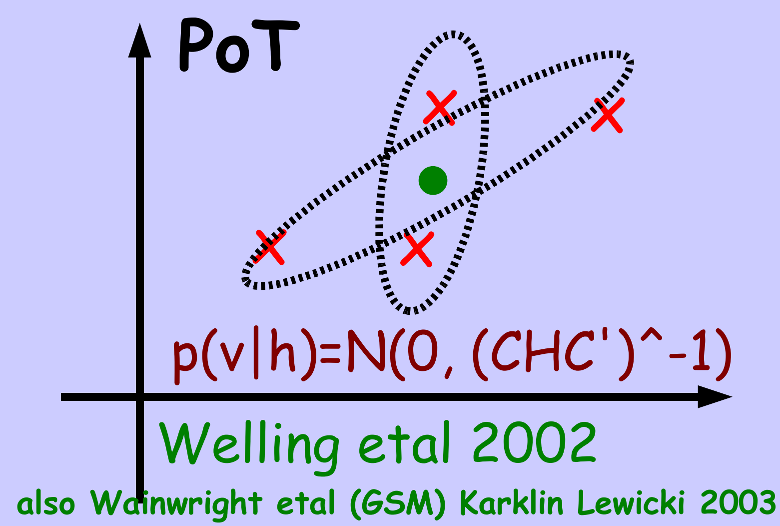


A generative perspective: p(v|h)

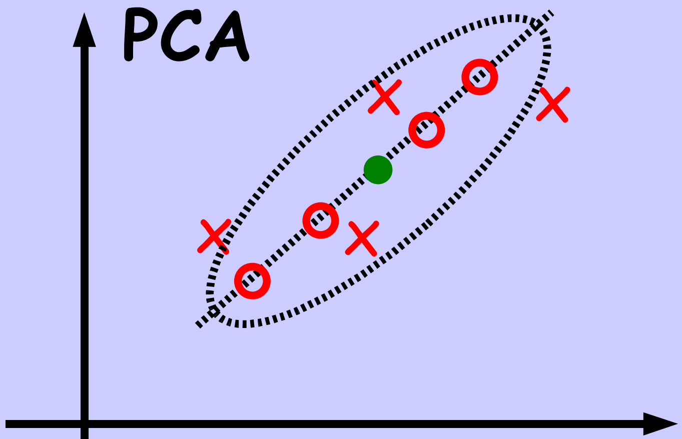




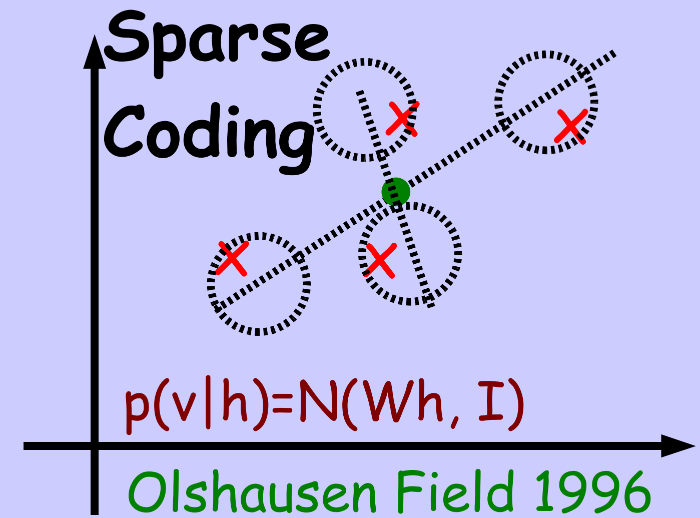




A generative perspective: p(v|h)

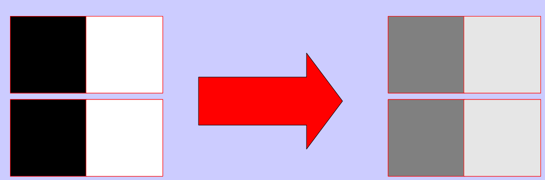


Pearson 1901

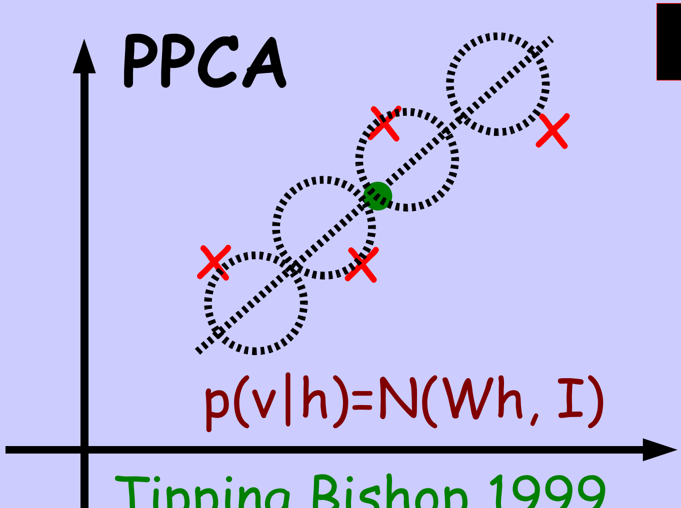


Olshausen Field 1996
also Welling Zvi Hinton (GRBM) & ICA

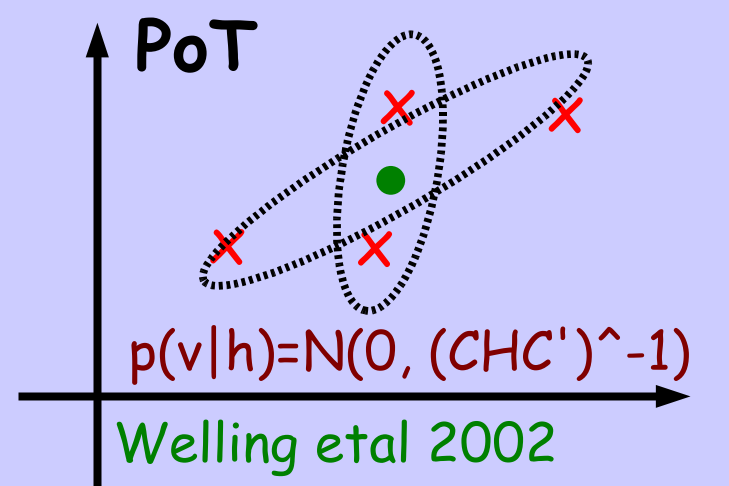
data point sample ~ p(v|h)



structure is preserved!



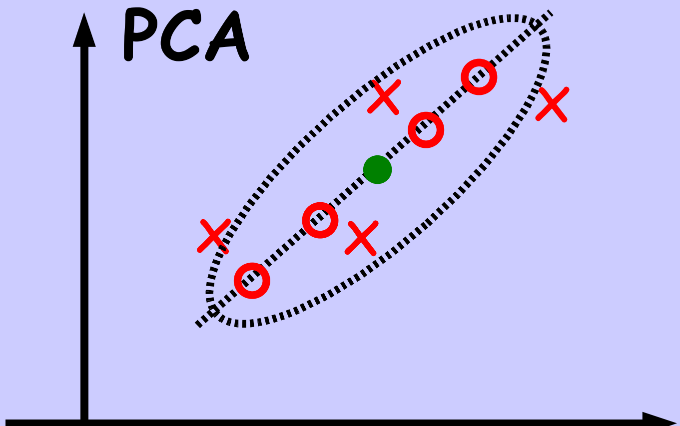
Tipping Bishop 1999



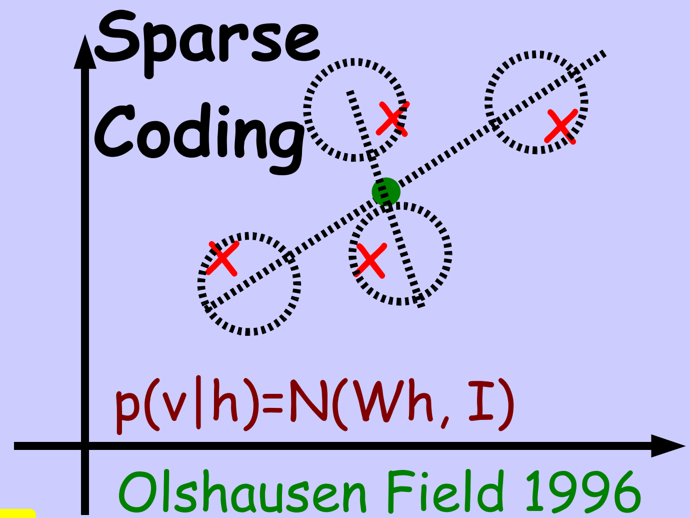
Welling etal 2002

also Wainwright etal (GSM) Karklin Lewicki 2003

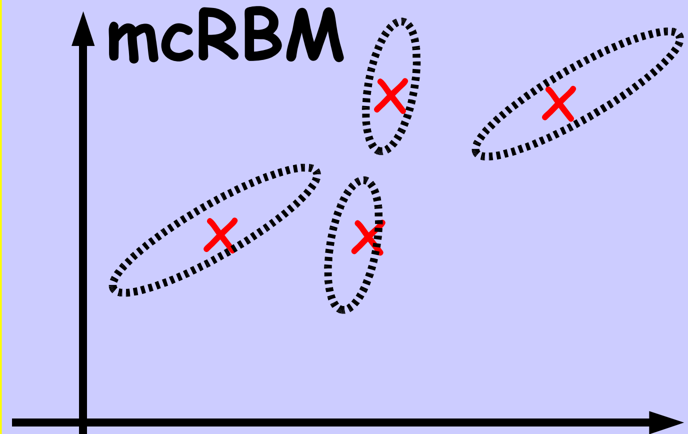
A generative perspective: p(v|h)



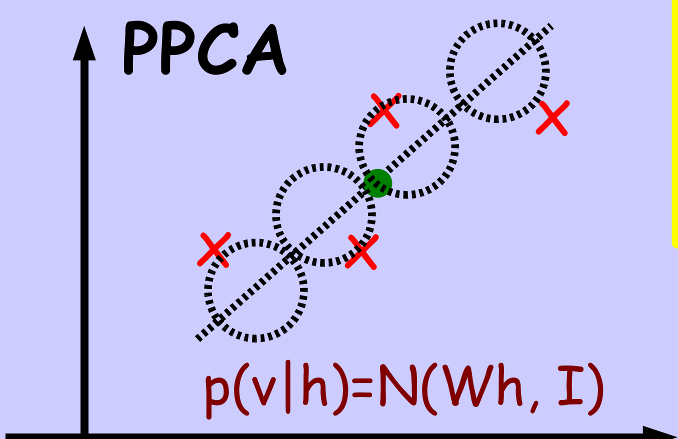
Pearson 1901



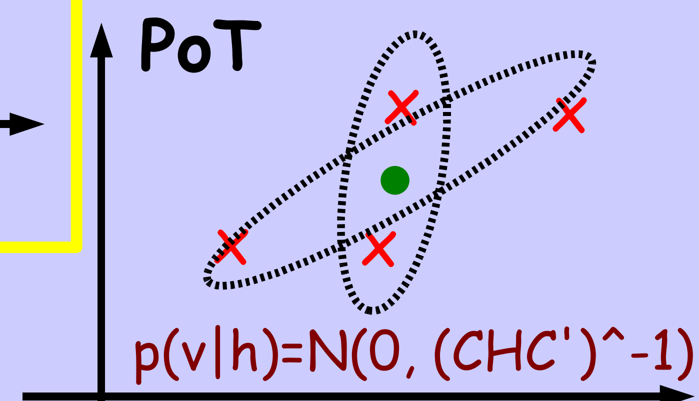
Olshausen Field 1996
also Welling Zvi Hinton
(GRBM) & ICA



Ranzato Hinton 2010

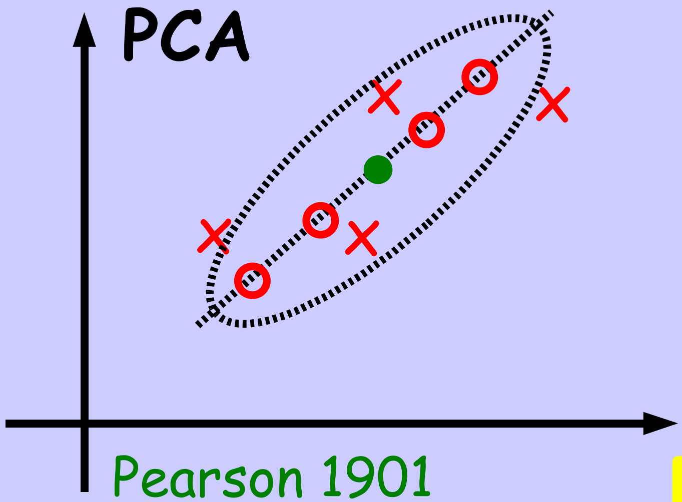


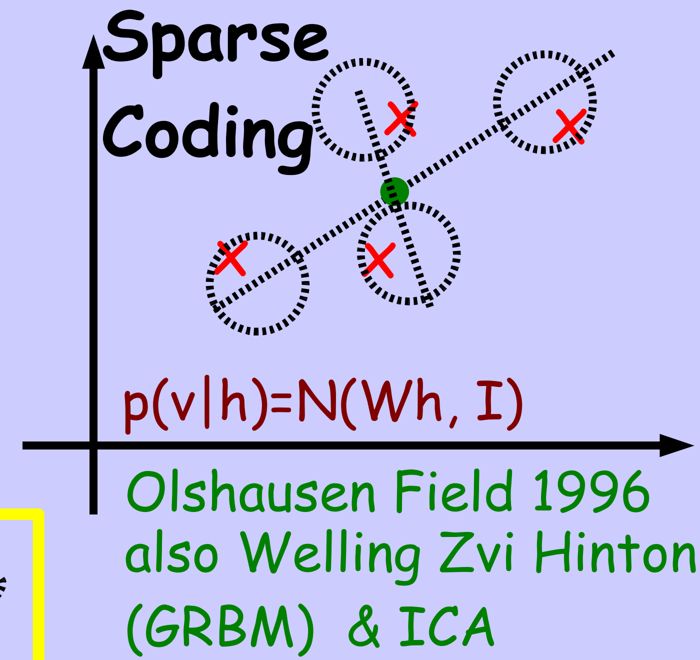
Tipping Bishop 1999

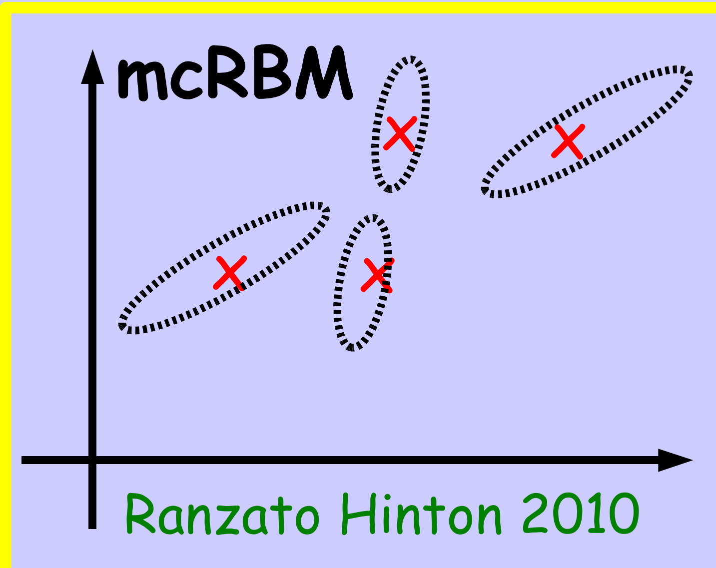


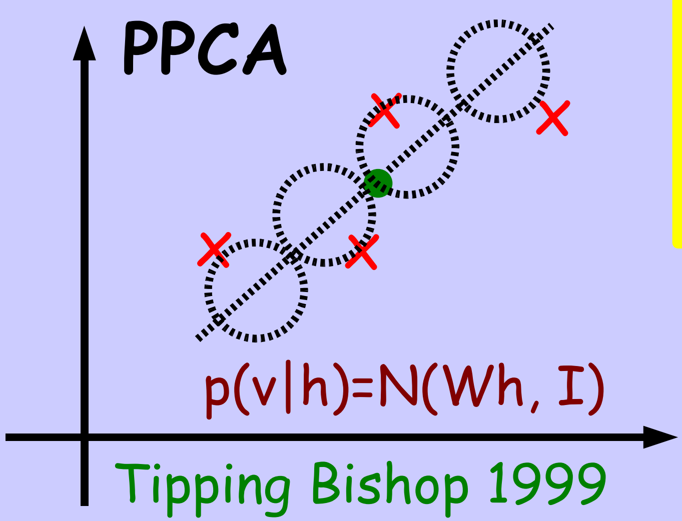
Welling etal 2002
also Wainwright etal (GSM) Karklin Lewicki 2003

A generative perspective: p(v|h)

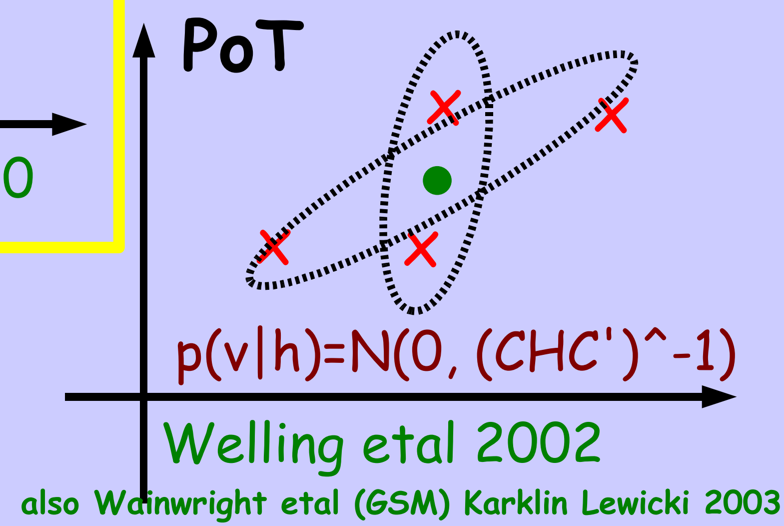




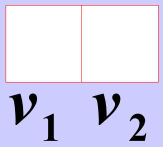




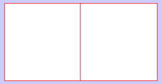
THIS IS NOT A MIXTURE OF GAUSSIANS!!

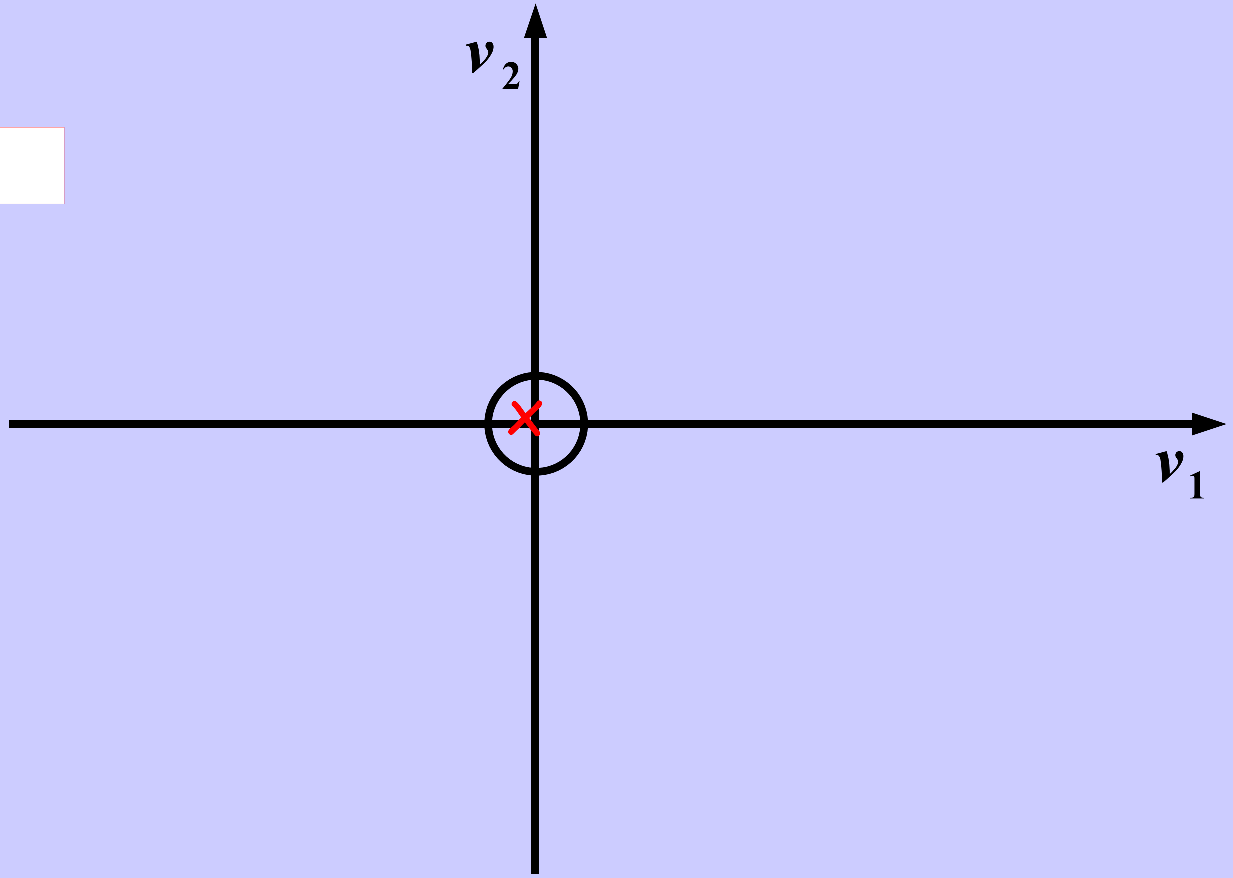


Why modeling covariance, p(v|h)

Example: image with two pixels 

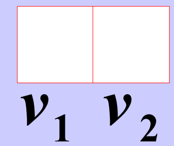
- assume we subtract off the mean from the image
- most of the times pixels take the same value

    - input 
    - p(v|h)



Why modeling covariance, p(v|h)

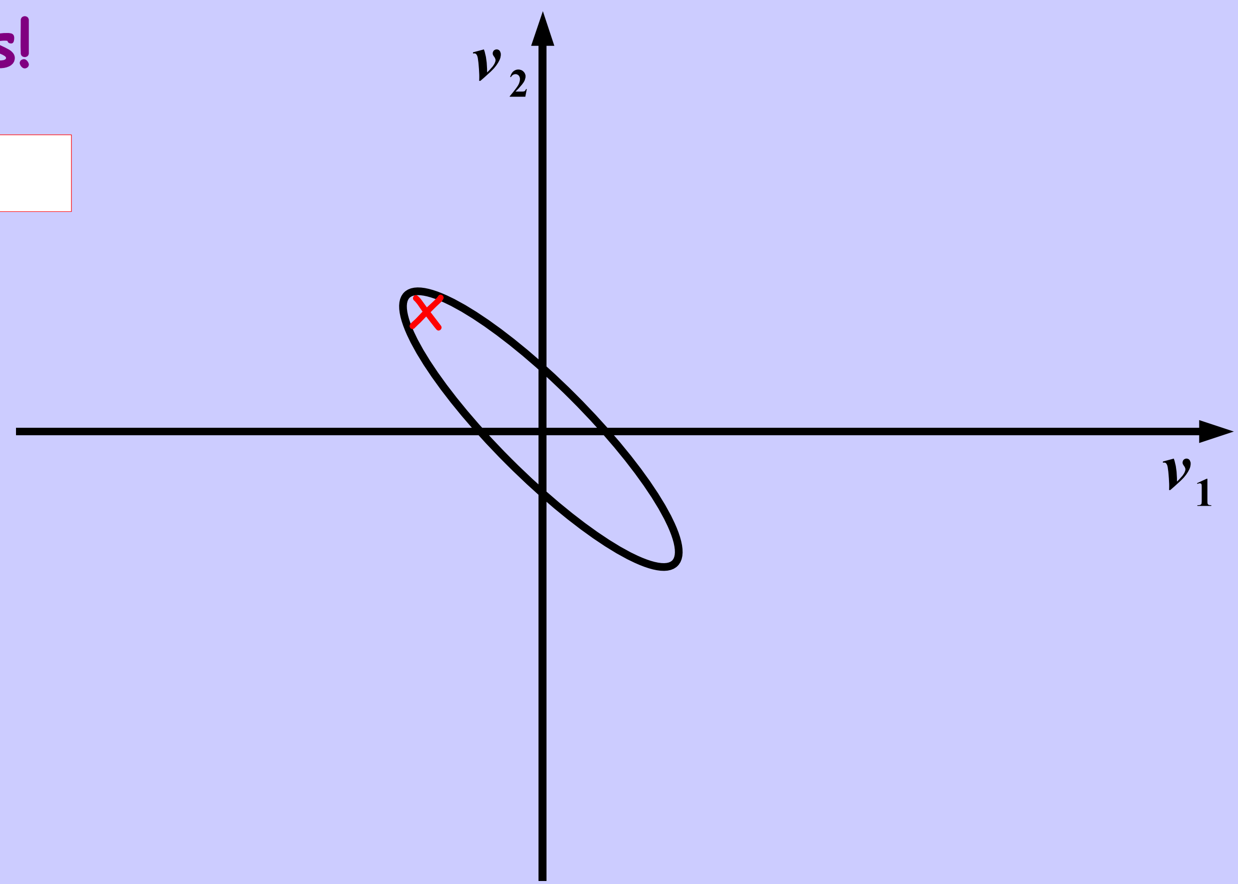
Example: image with two pixels



- assume we subtract off the mean from the image
- most of the times pixels take the same value
- with some exceptions!

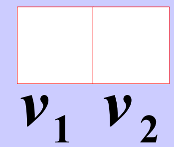
  - input 

  - p(v|h)

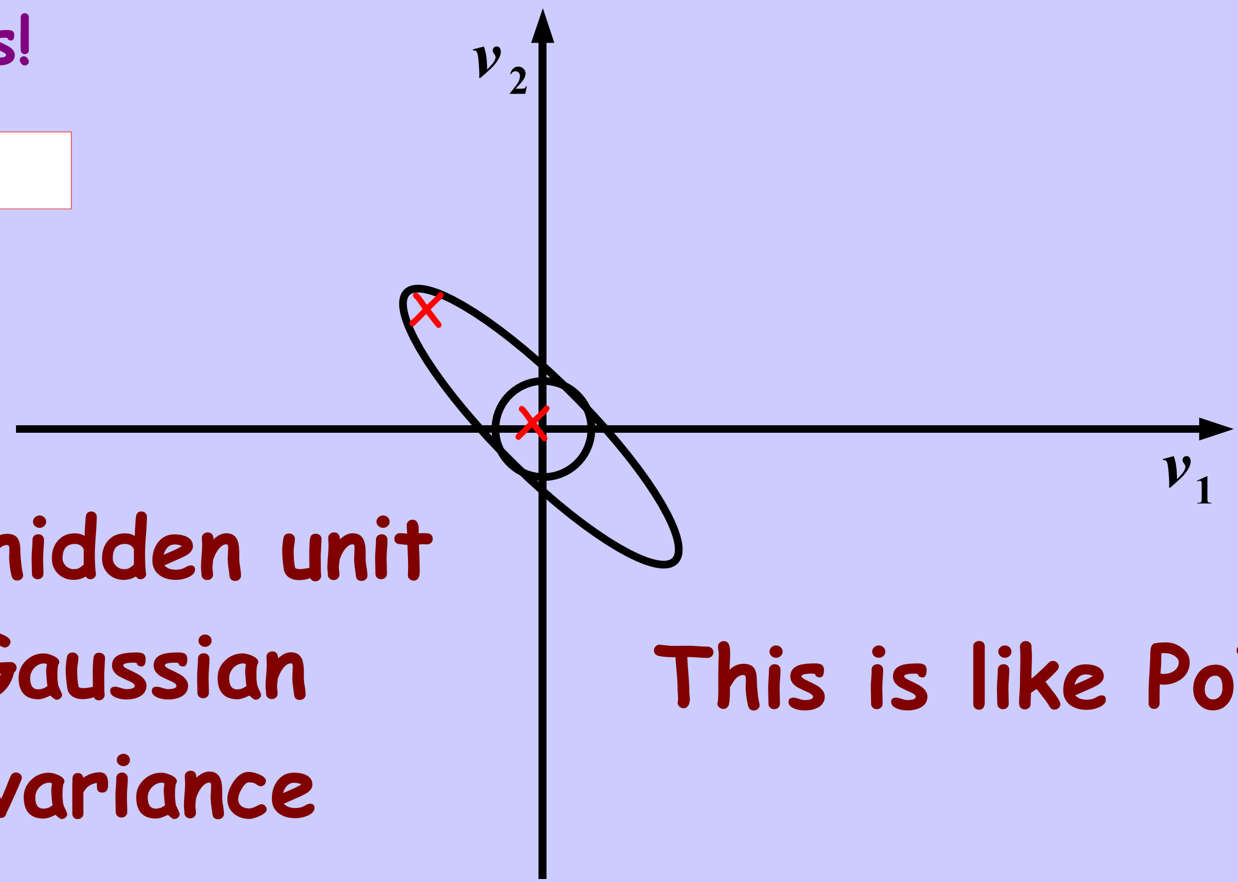


Why modeling covariance, p(v|h)

Example: image with two pixels



- assume we subtract off the mean from the image
- most of the times pixels take the same value
- with some exceptions!
  - input 
  - p(v|h)



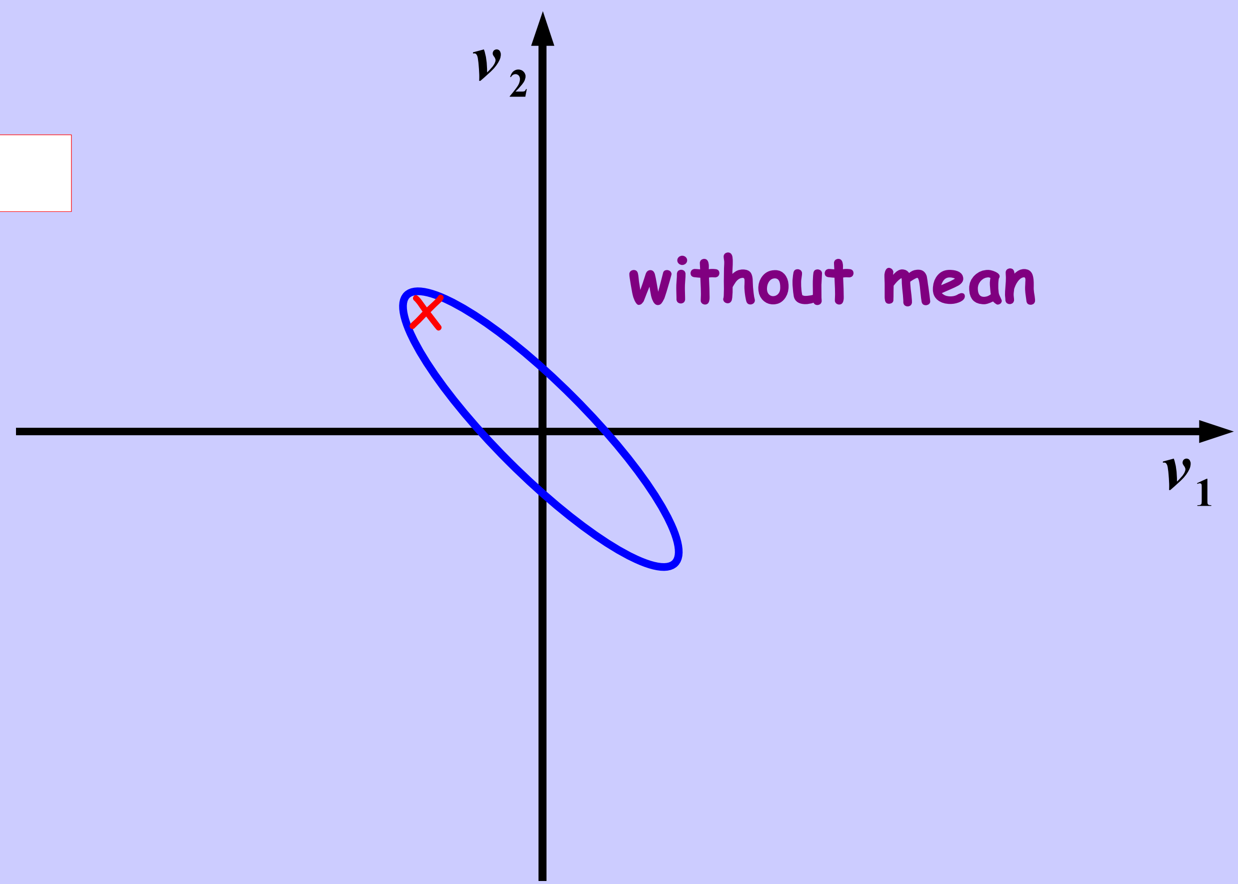
With one binary hidden unit we can pick the Gaussian with the right covariance

This is like PoT

Why modeling both mean and covariance, p(v|h)

We can produce an even more precise fit by modeling the mean.

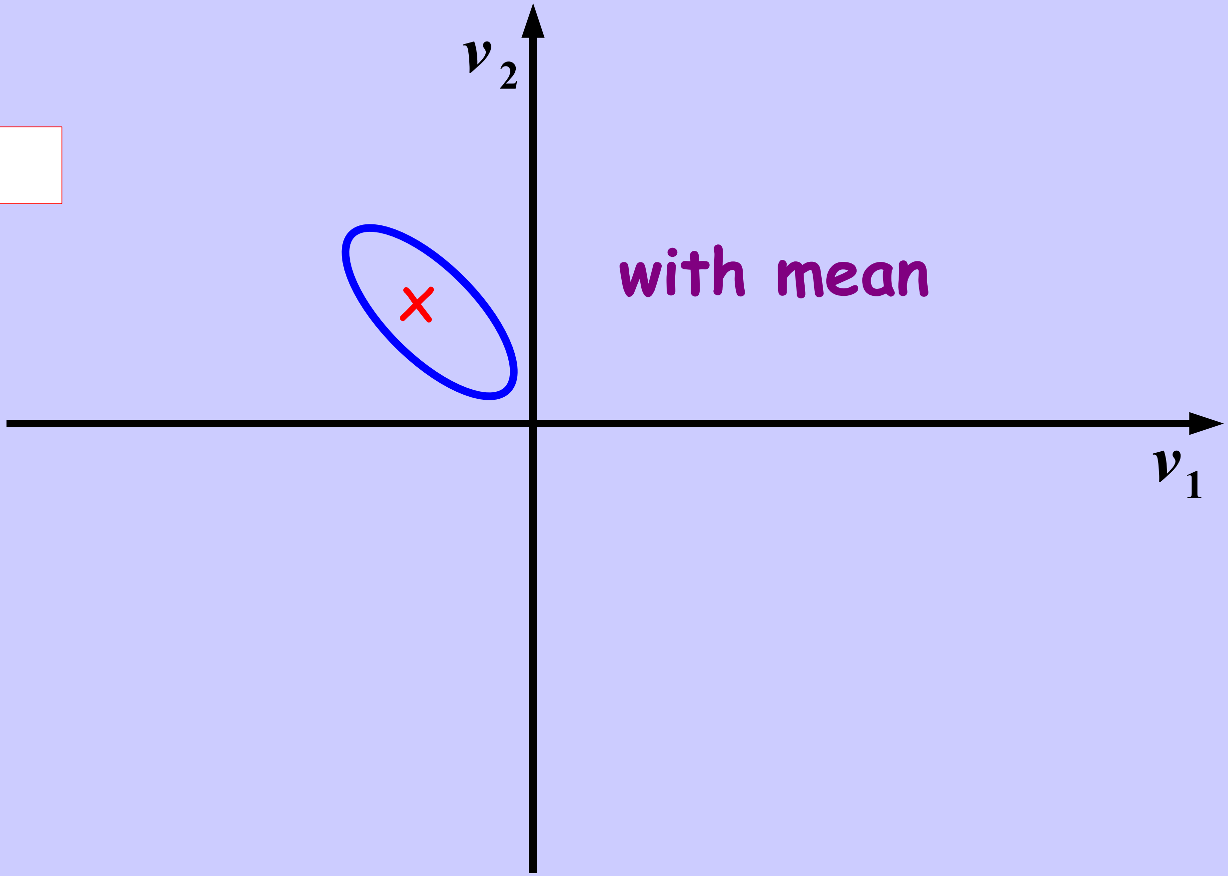
- input 
- p(v|h)



Why modeling both mean and covariance, p(v|h)

We can produce an even more precise fit by modeling the mean.

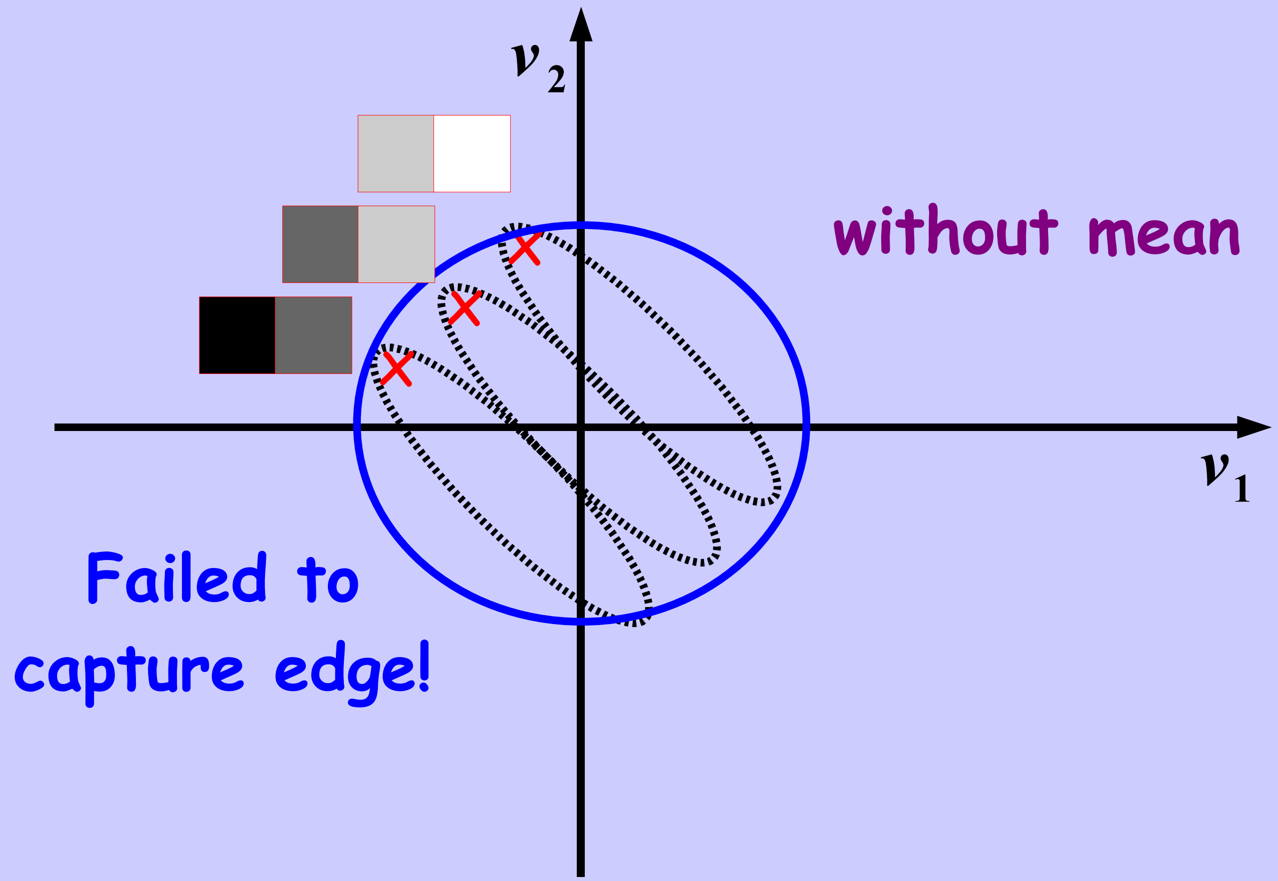
- input 
- p(v|h)



Why modeling both mean and covariance, p(v|h)

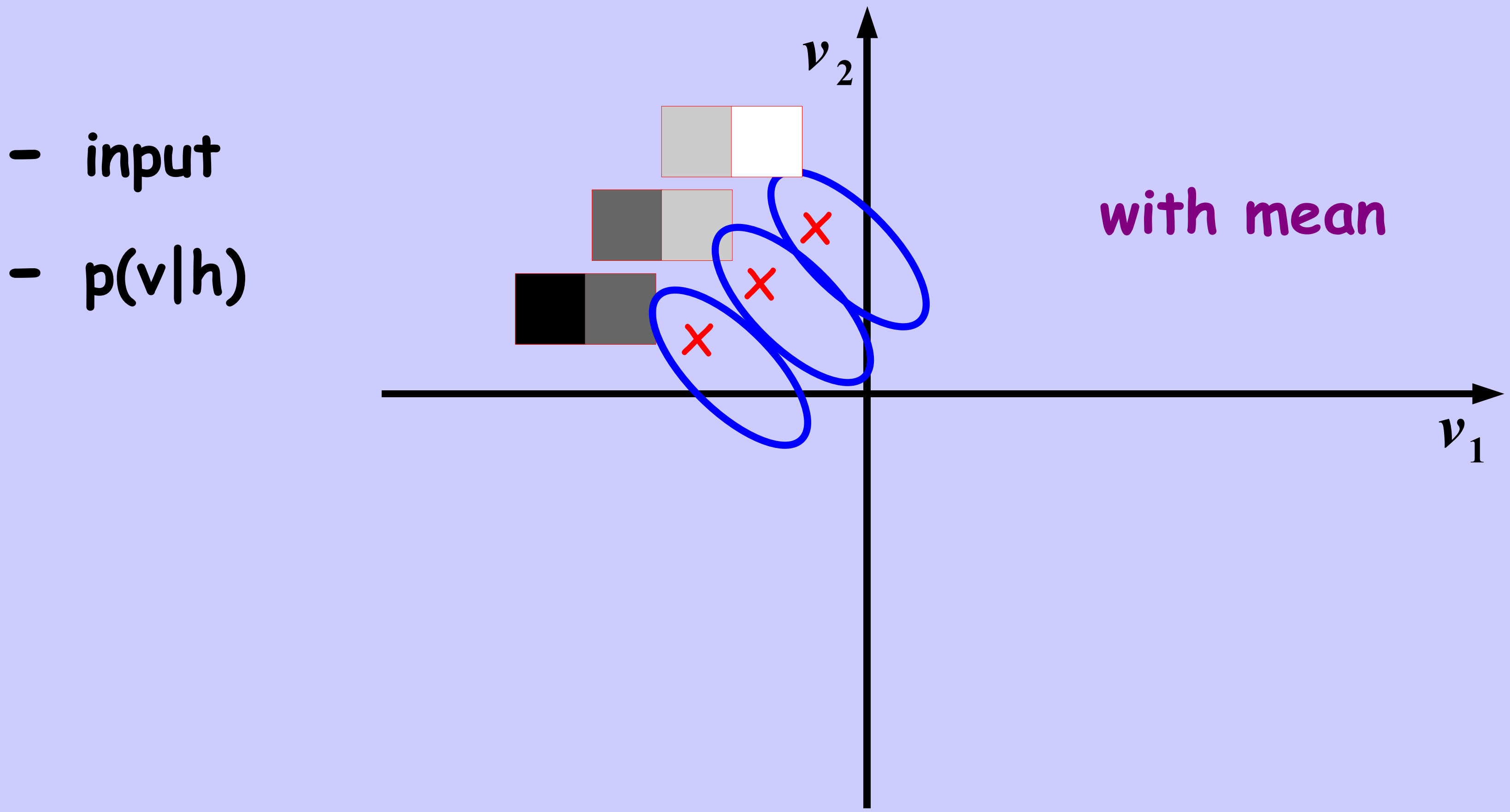
We can produce an even more precise fit by modeling the mean.
When data is not centered, this is even more dramatic.

- input
- p(v|h)

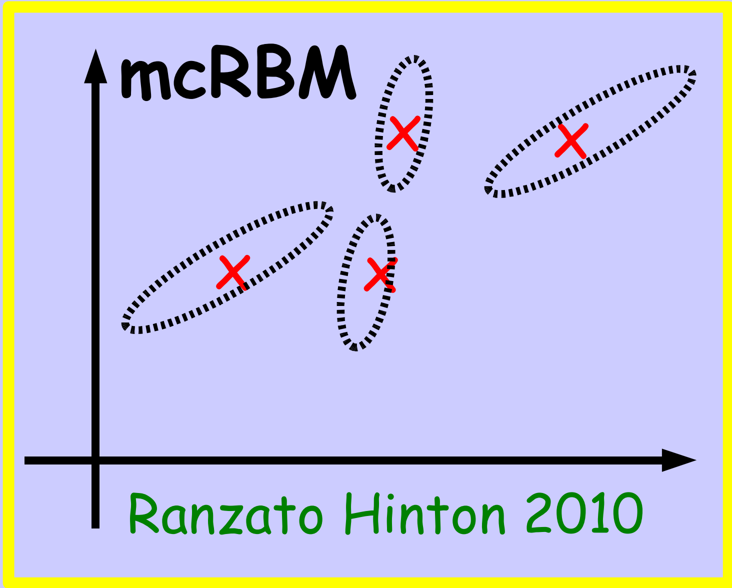


Why modeling both mean and covariance, p(v|h)

We can produce an even more precise fit by modeling the mean.
When data is not centered, this is even more dramatic.



p(v|h) hiddens determine an image-specific mean and an image-specific covariance.



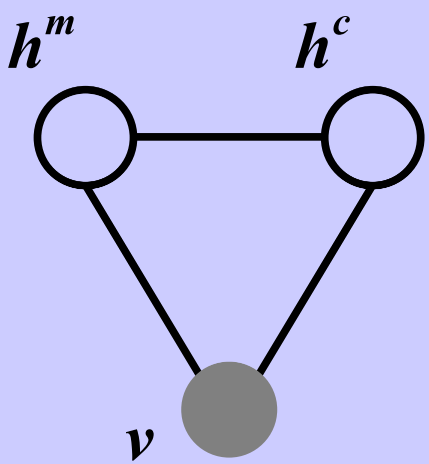
How to modulate mean and covariance using hidden units?

$$p(v, h^m, h^c) \propto \exp(-E(v, h^m, h^c))$$

- easy generation
- slow inference

$$p(v|h^m, h^c) = N(m(h^m), \Sigma(h^c))$$

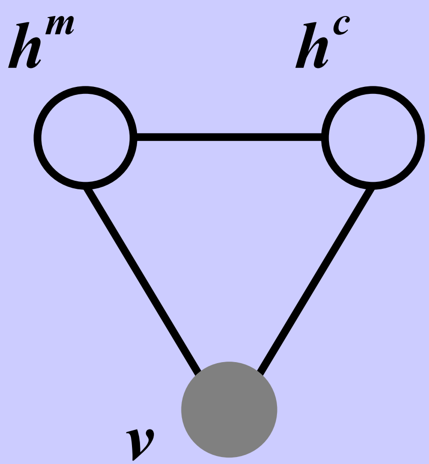
$$E = \frac{1}{2}(v - m)' \Sigma^{-1} (v - m)$$



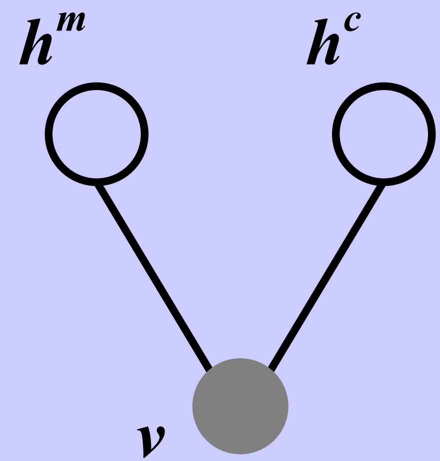
- easy generation
- slow inference

$$p(v|h^m, h^c) = N(m(h^m), \Sigma(h^c))$$

$$E = \frac{1}{2}(v - m)' \Sigma^{-1}(v - m)$$



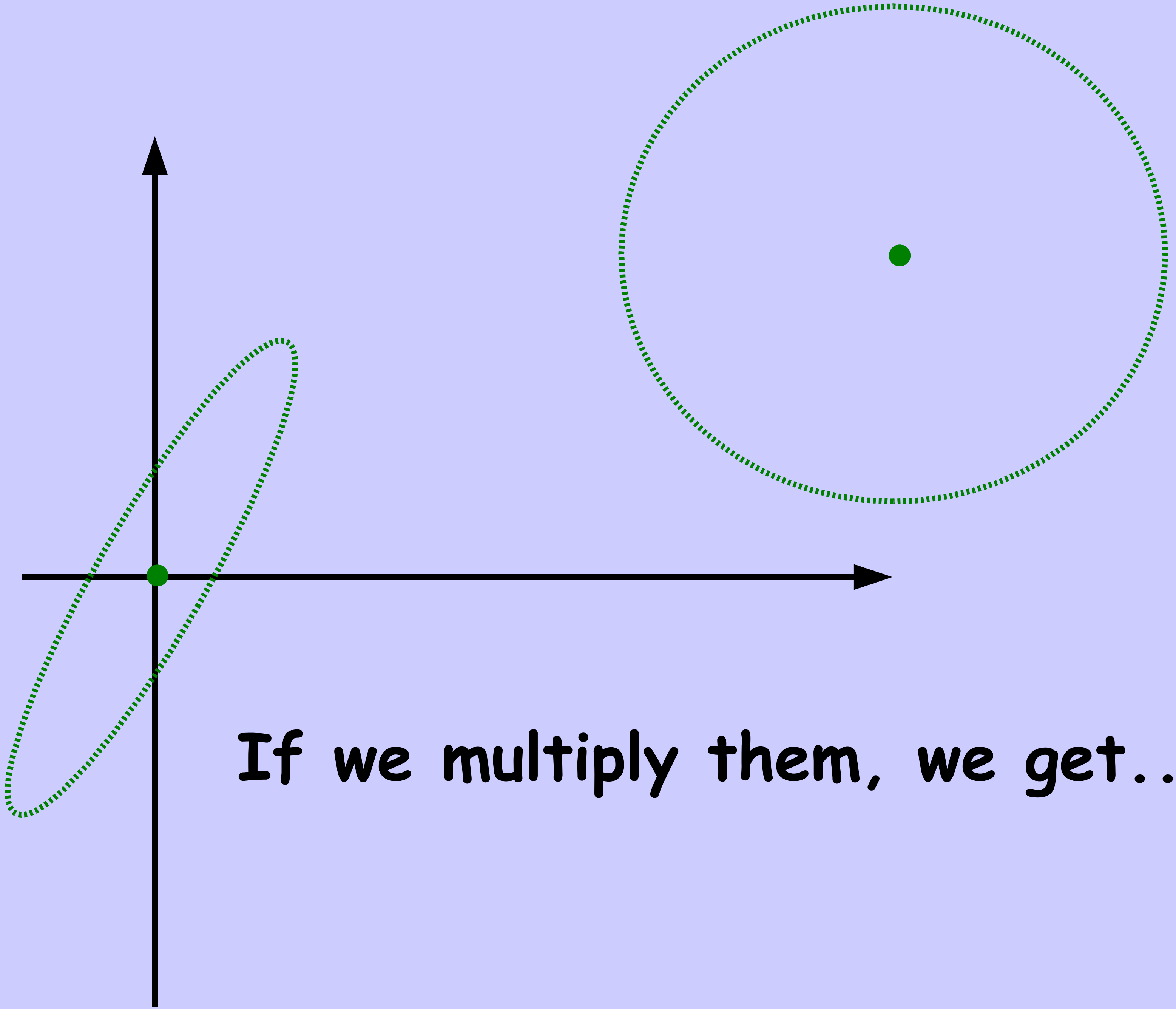
- less easy generation
- fast inference



$$E = \frac{1}{2} v' \Sigma^{-1} v - m v$$

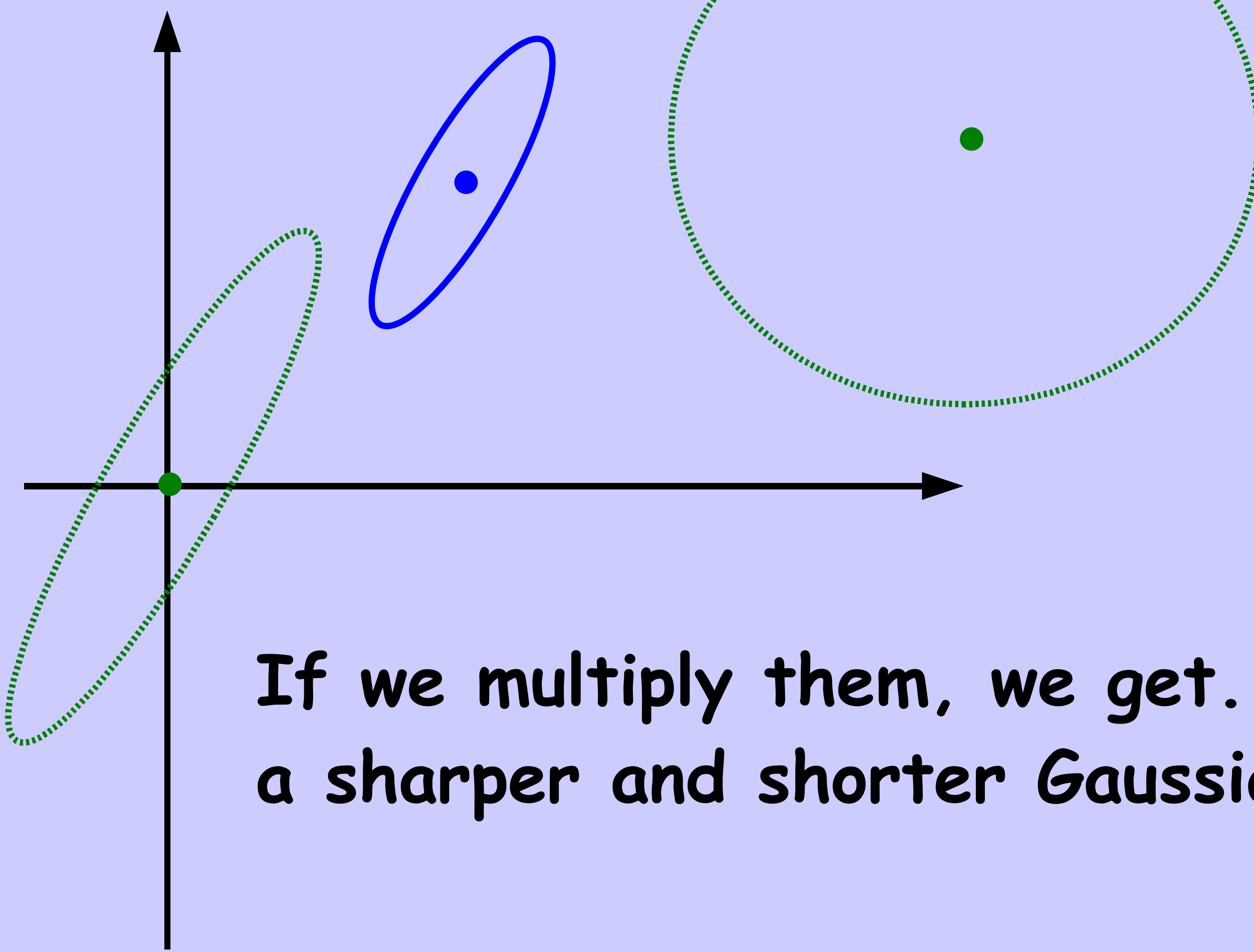
$$p(v|h^m, h^c) = N(\Sigma(h^c)^{-1} m(h^m), \Sigma(h^c))$$

Geometric interpretation



If we multiply them, we get...

Geometric interpretation

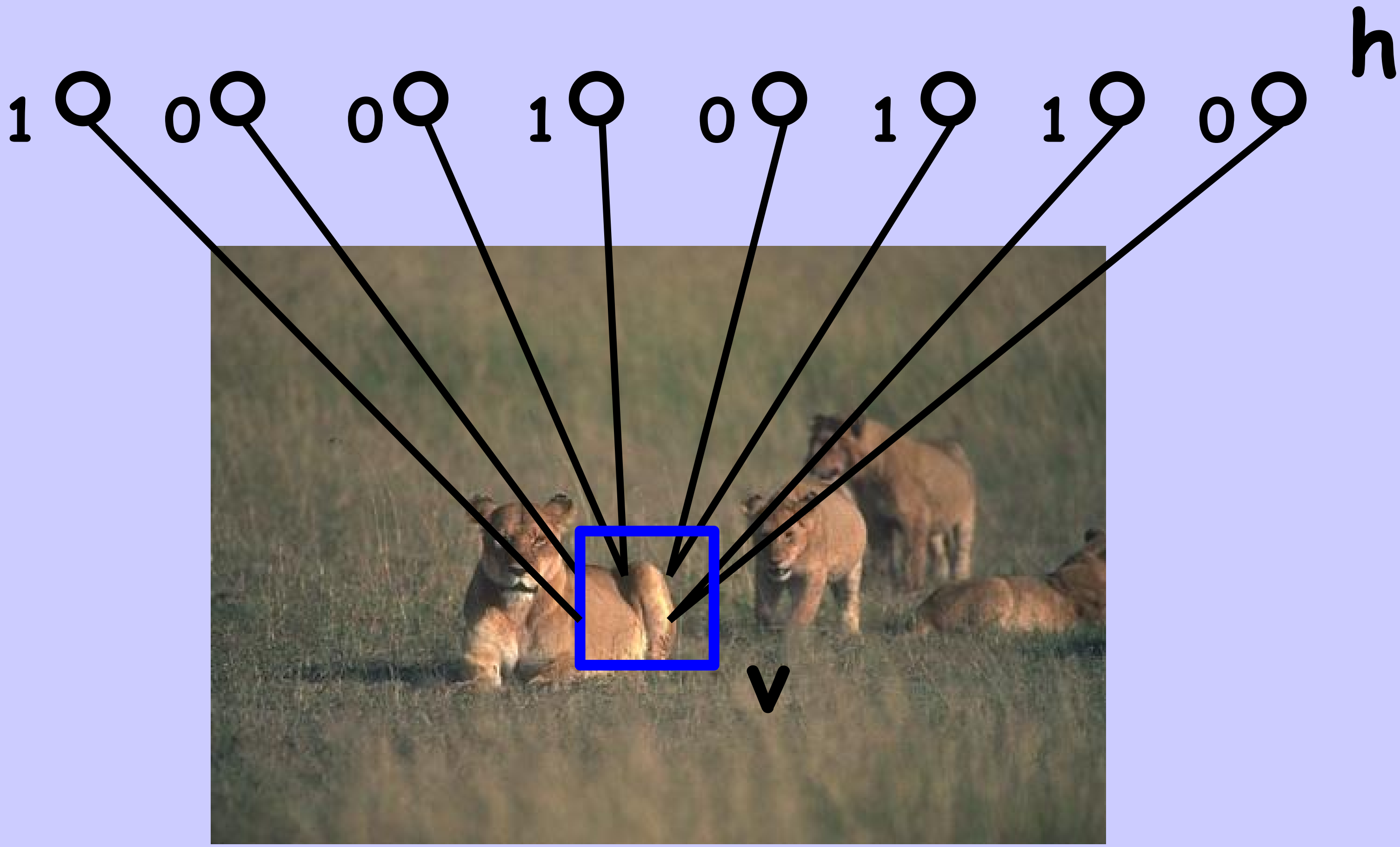


If we multiply them, we get...
a sharper and shorter Gaussian

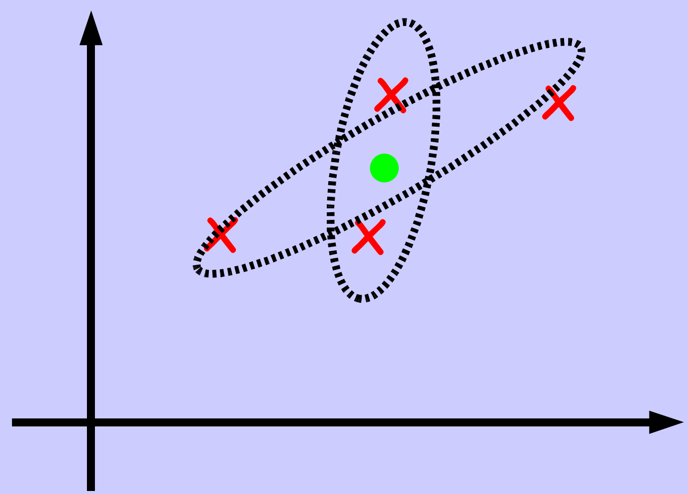
We start by modeling small image patches.

- v visibles
- h hiddens

p(v,h)



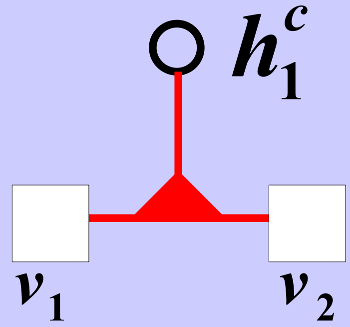
Modeling the covariance only (using binary hiddens): cRBM



$$E = \frac{1}{2} v' \Sigma^{-1} v$$

Ranzato Krizhevsky Hinton AISTATS 2010

gated MRF



$$E^c(v, h^c) = w_1 v_1 v_2 h_1^c + \ldots$$

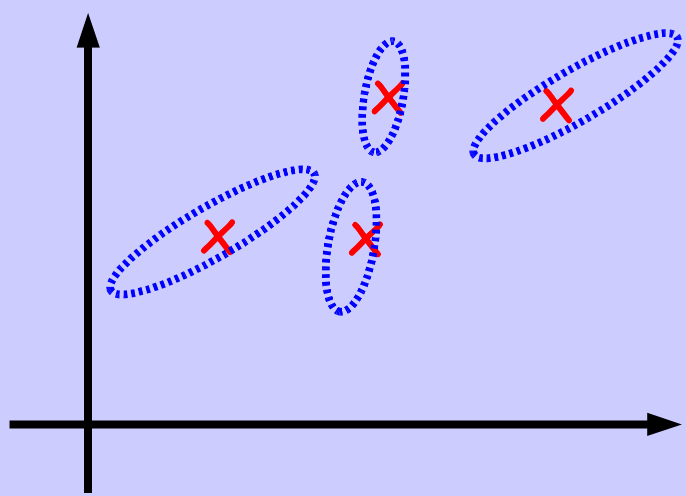
Interactions determined by state of latent variables

$$E^c(v, h^c) = \frac{1}{2} v' \Sigma^{-1} v$$

$$\Sigma^{-1} = C \, diag(P h^c) C' \qquad h_k^c \in \{0,1\}$$

So far we modeled the covariance only...
now we add also the mean

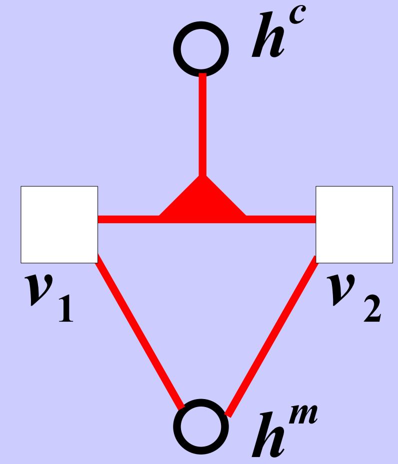
mcRBM



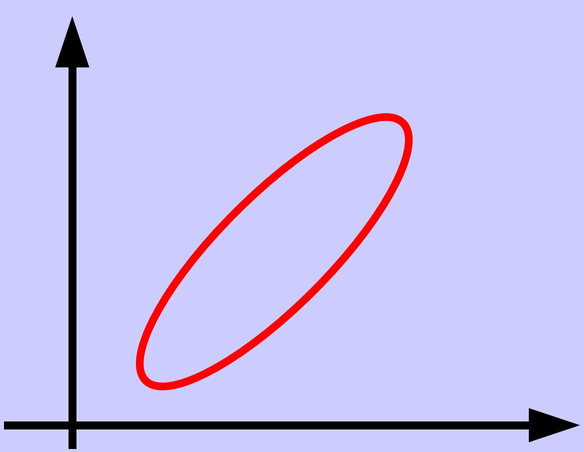
$$E = \frac{1}{2} v' \Sigma^{-1} v - m v$$

Ranzato Hinton CVPR 2010

$$E(v,h^c,h^m)=\frac{1}{2}v'\underbrace{\Sigma^{-1}}_{\text{cRBM}}v-\sum_{ij}W_{ij}v_i h_j^m$$



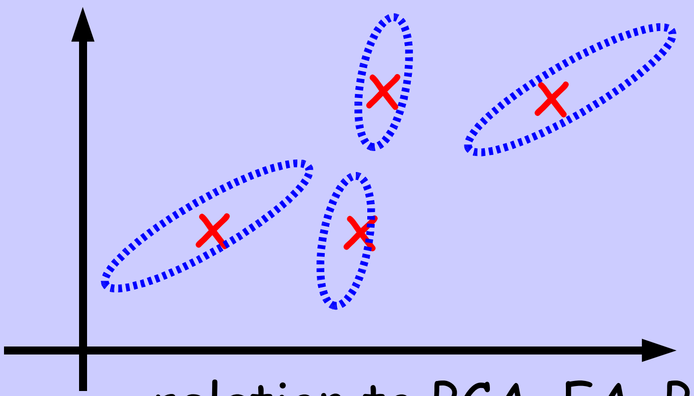
Conditional over visibles has non-zero mean that depends on both sets of hiddens:



$$p(v|h^c,h^m)=N(\Sigma(W h^m),\Sigma)$$

Interpreting mcRBM

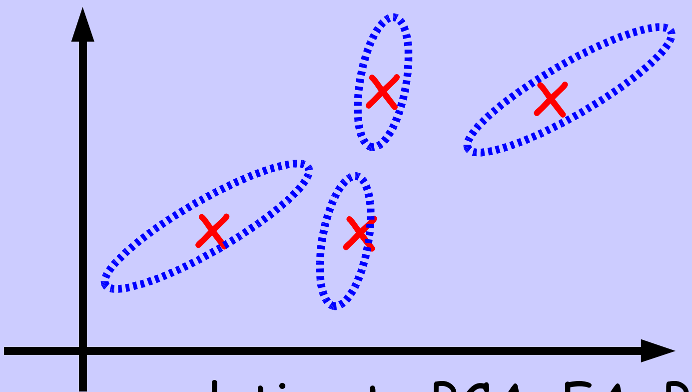
- Looking at $p(v|h)$



- relation to PCA, FA, PoT, etc.

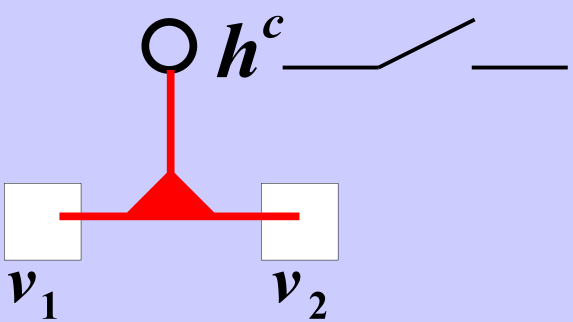
Interpreting mcRBM

- Looking at $p(v|h)$



- relation to PCA, FA, PoT, etc.

- Looking at hiddens

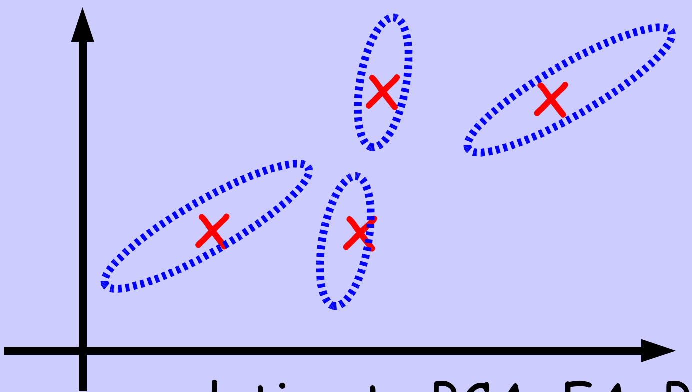


- relation to line process and PoT

Geman etal 84, Blake etal 87, Black etal 96

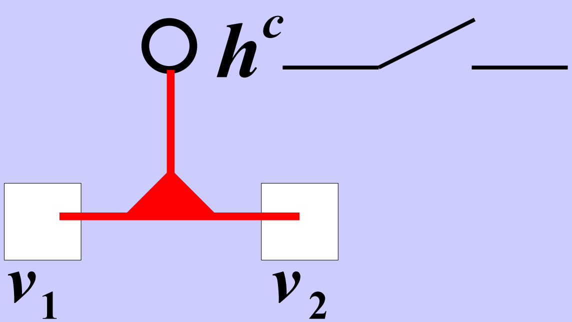
Interpreting mcRBM

- Looking at $p(v|h)$



- relation to PCA, FA, PoT, etc.

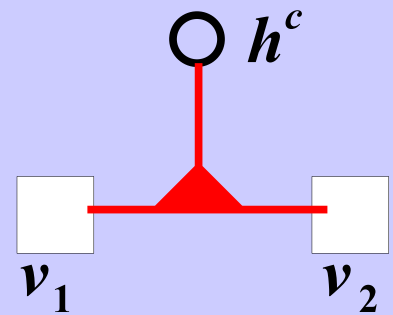
- Looking at hiddens



- relation to line process and PoT

Geman etal 84, Blake etal 87, Black etal 96

- Looking at $E(v, h)$



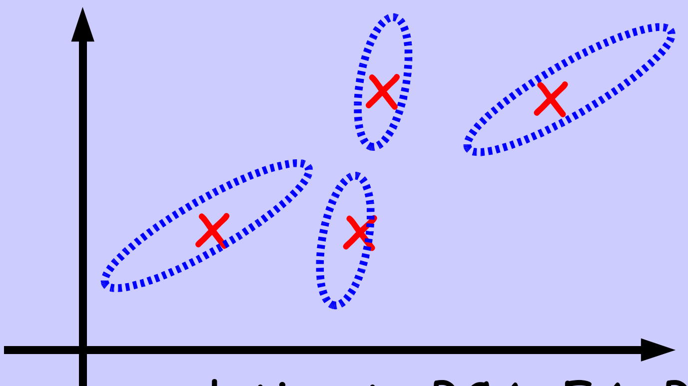
3rd order BM

- relation to conditional 3-way RBM

Memisevic et al 07

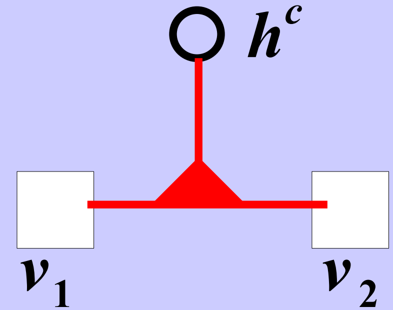
Interpreting mcRBM

- Looking at $p(v|h)$



- relation to PCA, FA, PoT, etc.

- Looking at $E(v,h)$

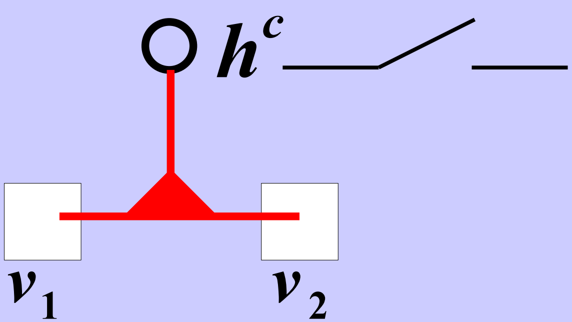


3rd order BM

- relation to conditional 3-way RBM

Memisevic et al 07

- Looking at hiddens

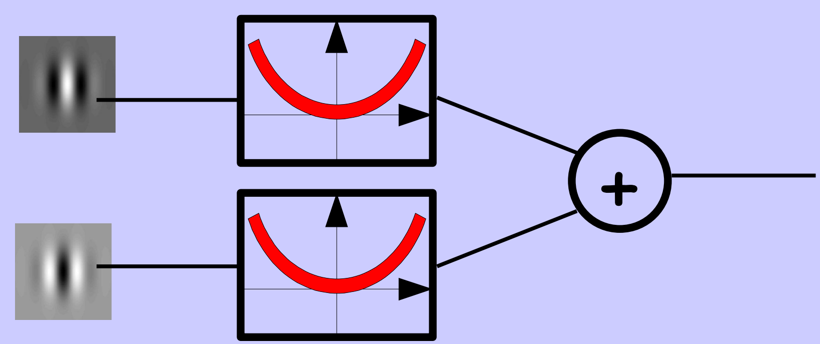


- relation to line process and PoT

Geman etal 84, Blake etal 87, Black etal 96

- Looking at $p(h|v)$

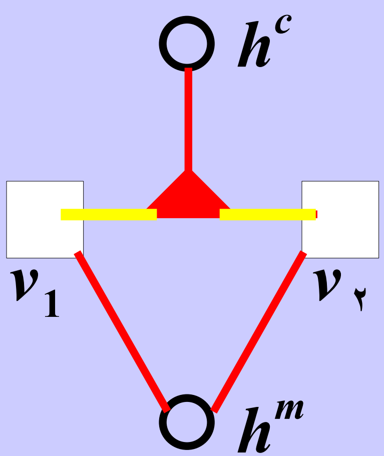
$$p(h_k^c=1|v)=\sigma(-\frac{1}{2}P_k(C'v)^2+b_k)$$



- relation to simple-complex cell model

LEARNING

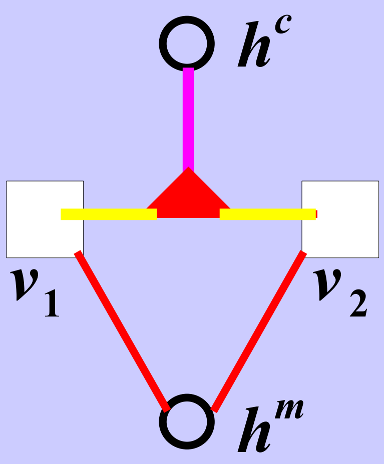
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



$$E(v, h^c, h^m) = \frac{1}{2}\sum_k h_k^c P_k (C\, v)^2 - \sum_j h_j^m W_j\, v$$

LEARNING

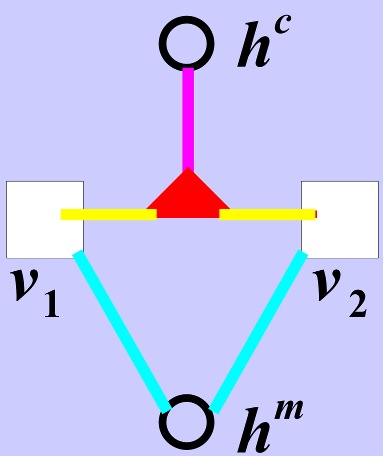
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



$$E(v, h^c, h^m) = \frac{1}{2}\sum_k h_k^c P_k (C v)^2 - \sum_j h_j^m W_j v$$

LEARNING

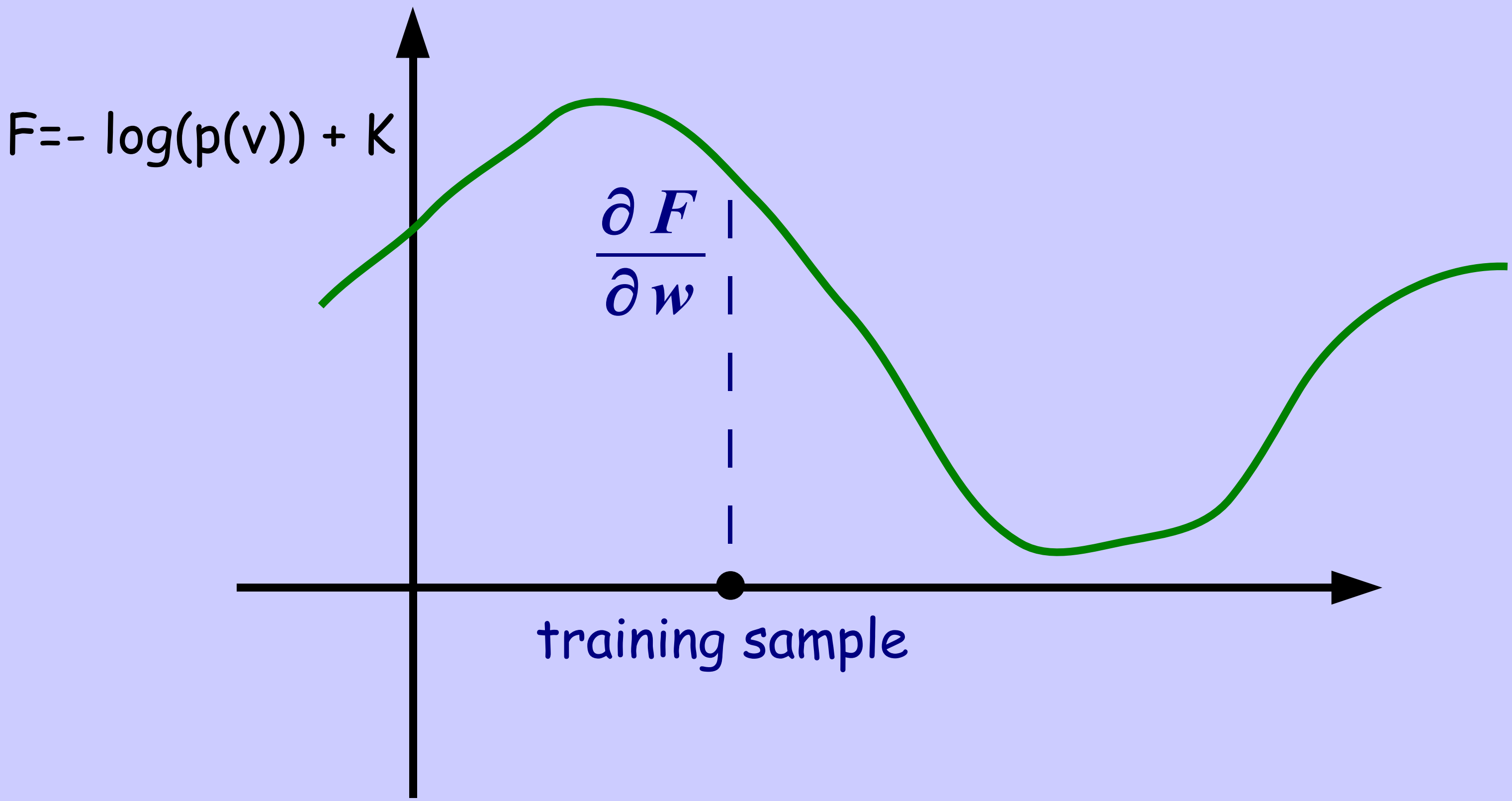
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



$$E(v, h^c, h^m) = \frac{1}{2}\sum_k h^c_k P_k (C v)^2 - \sum_j h^m_j W_j v$$

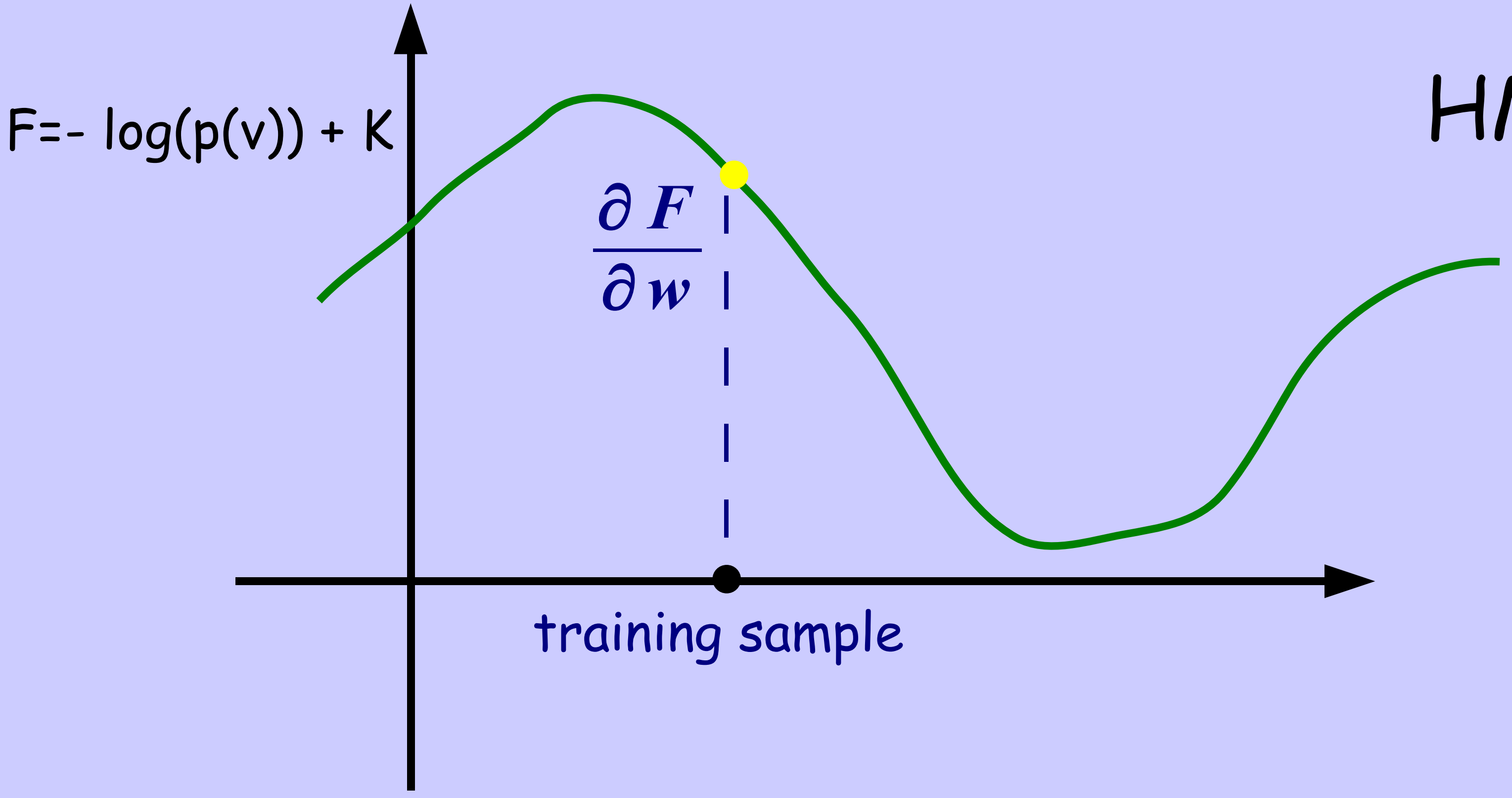
LEARNING

- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



LEARNING

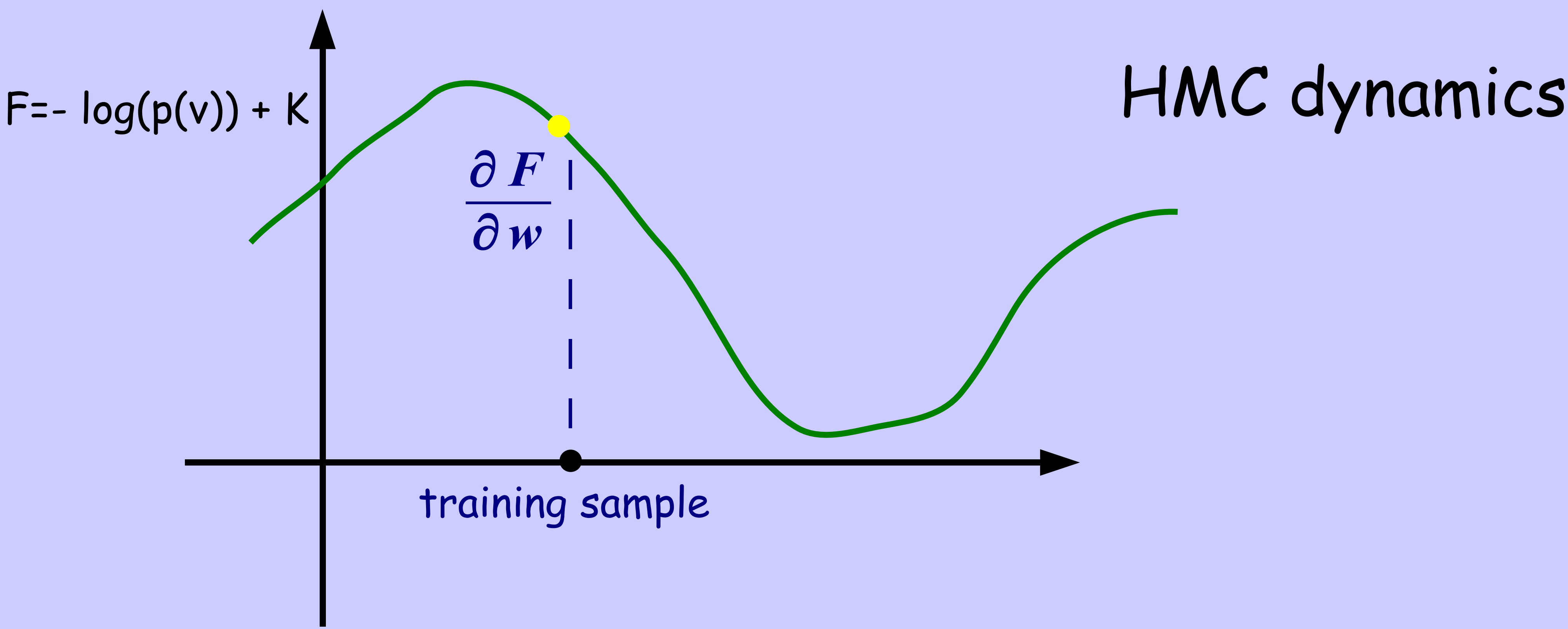
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



HMC dynamics

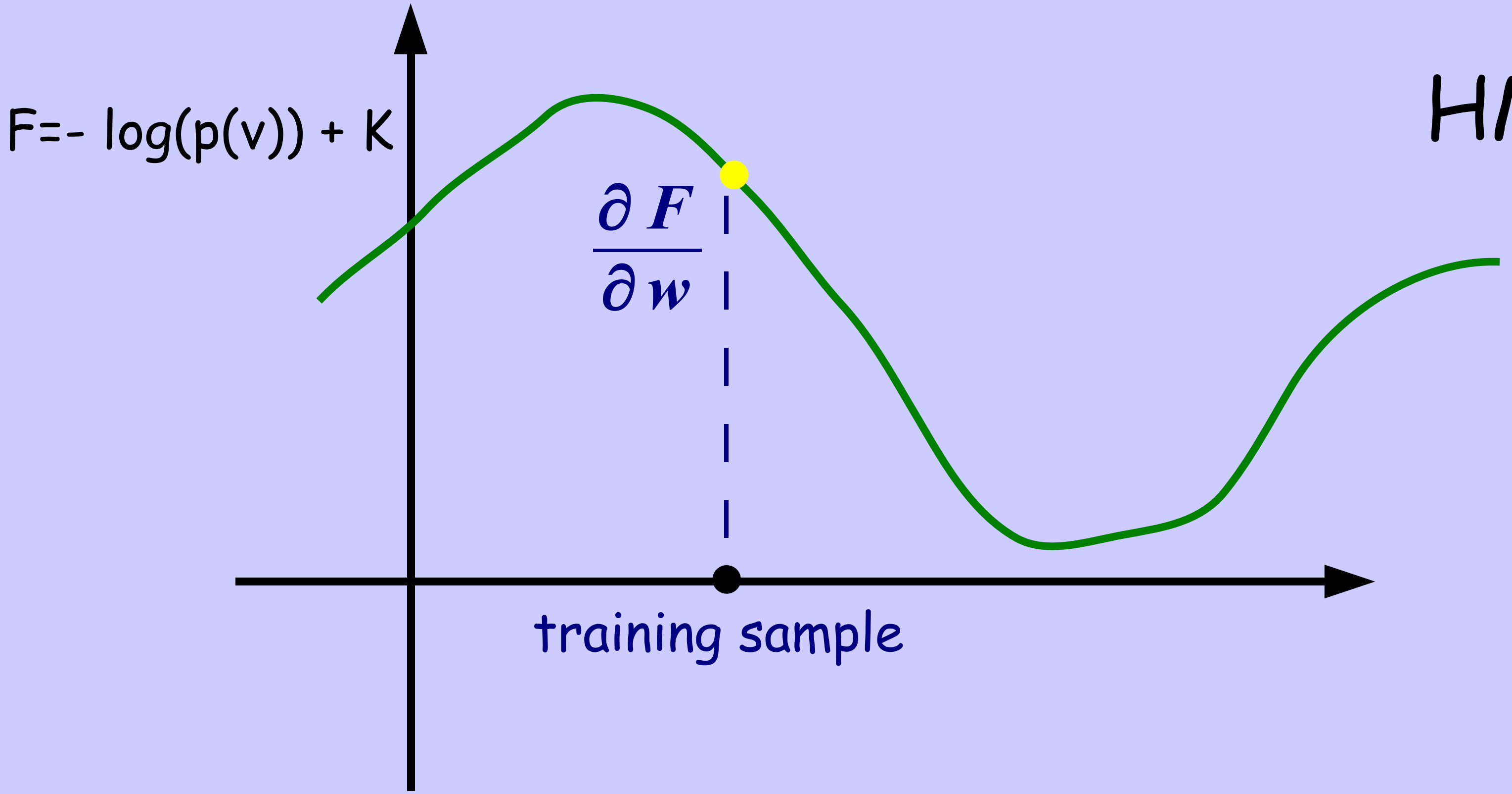
LEARNING

- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



LEARNING

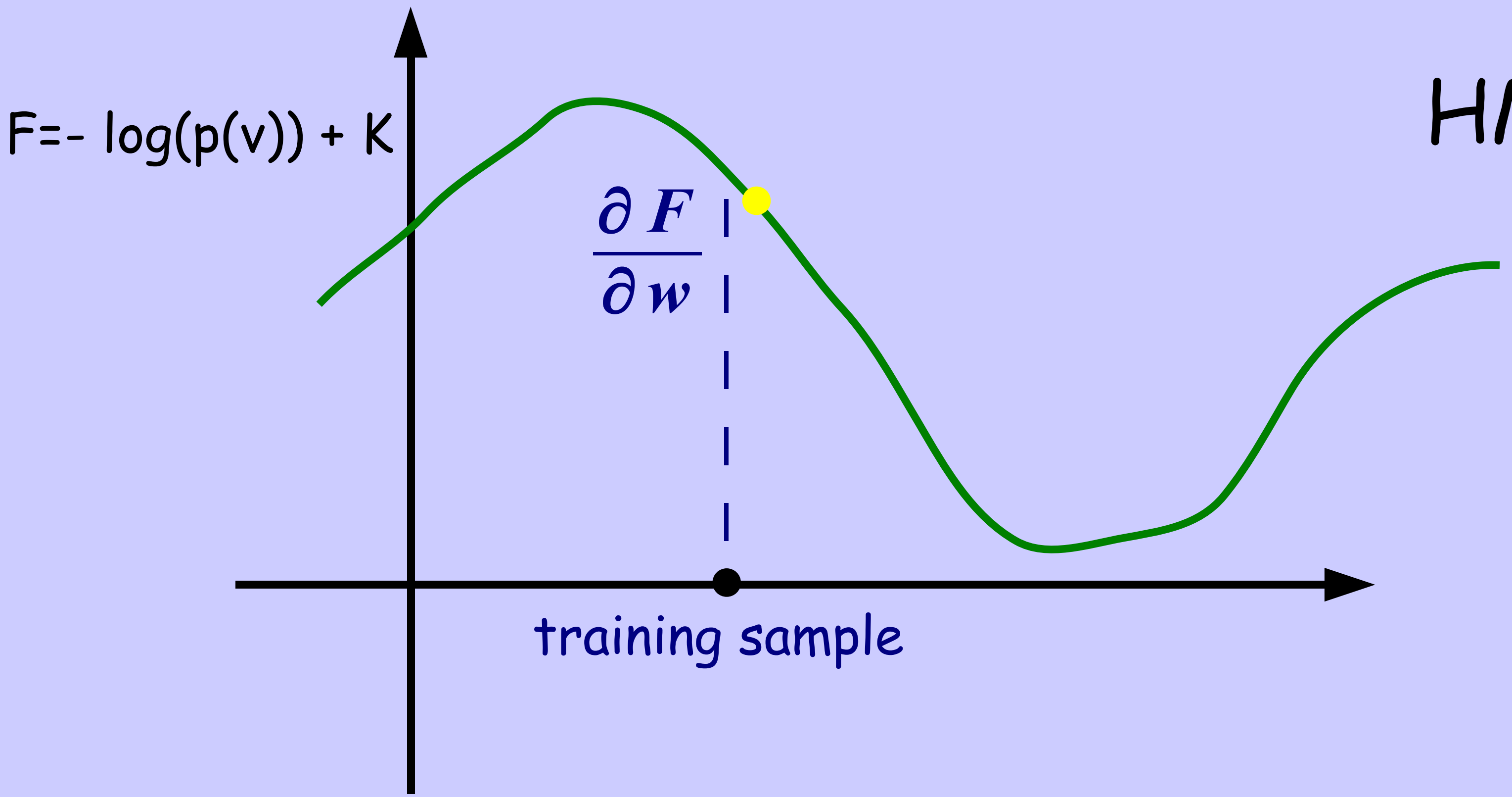
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



HMC dynamics

LEARNING

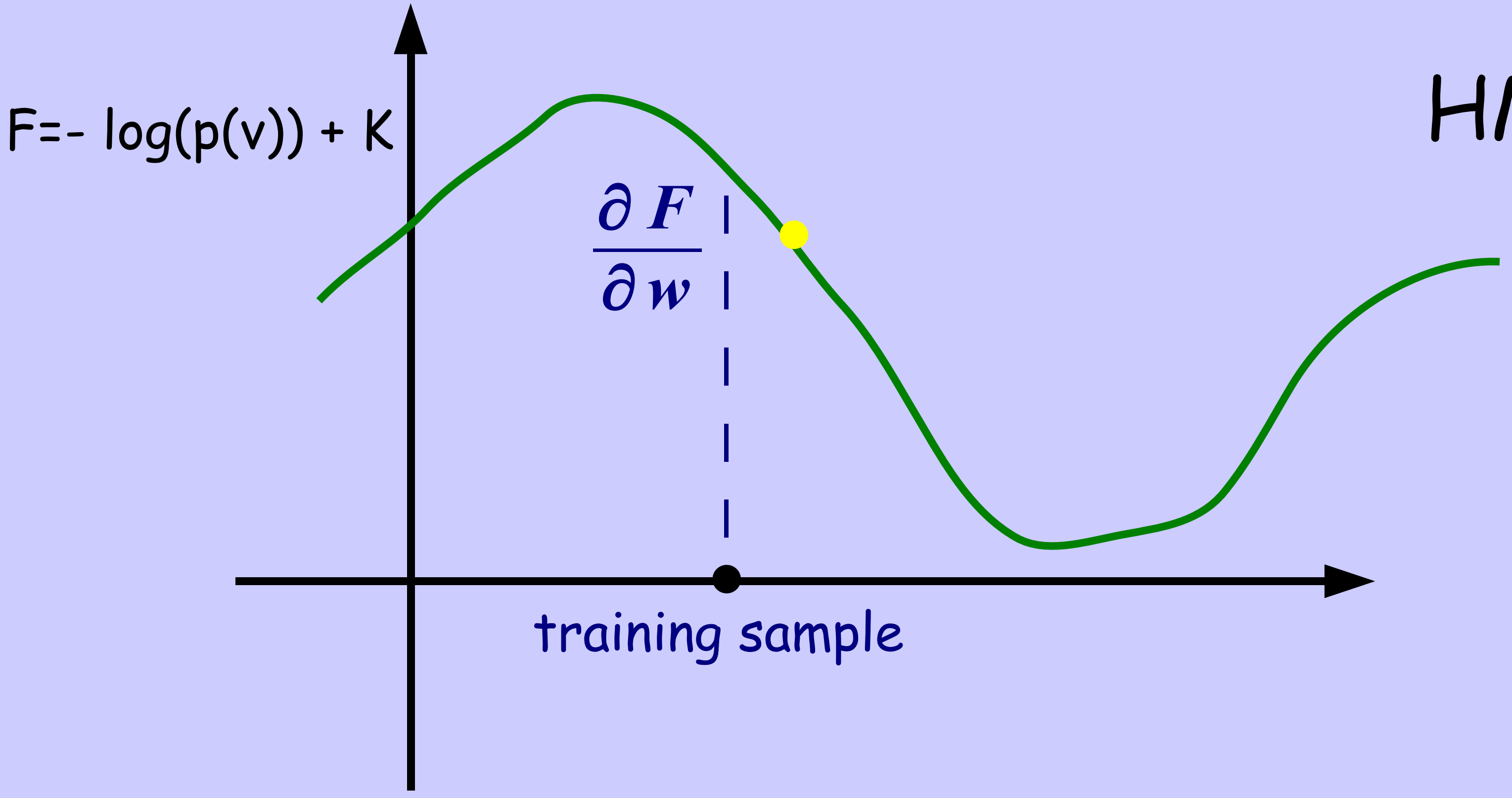
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



HMC dynamics

LEARNING

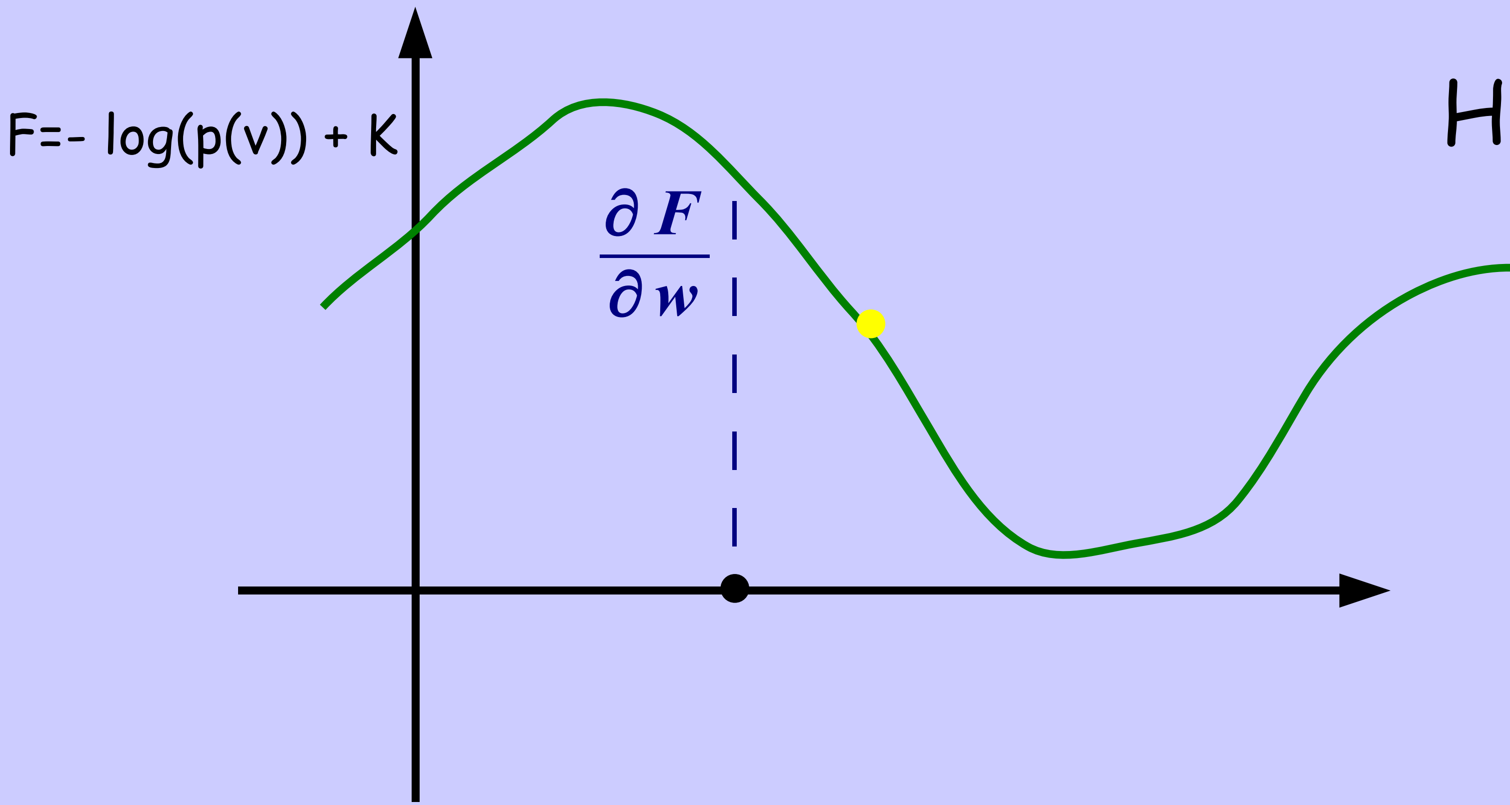
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



HMC dynamics

LEARNING

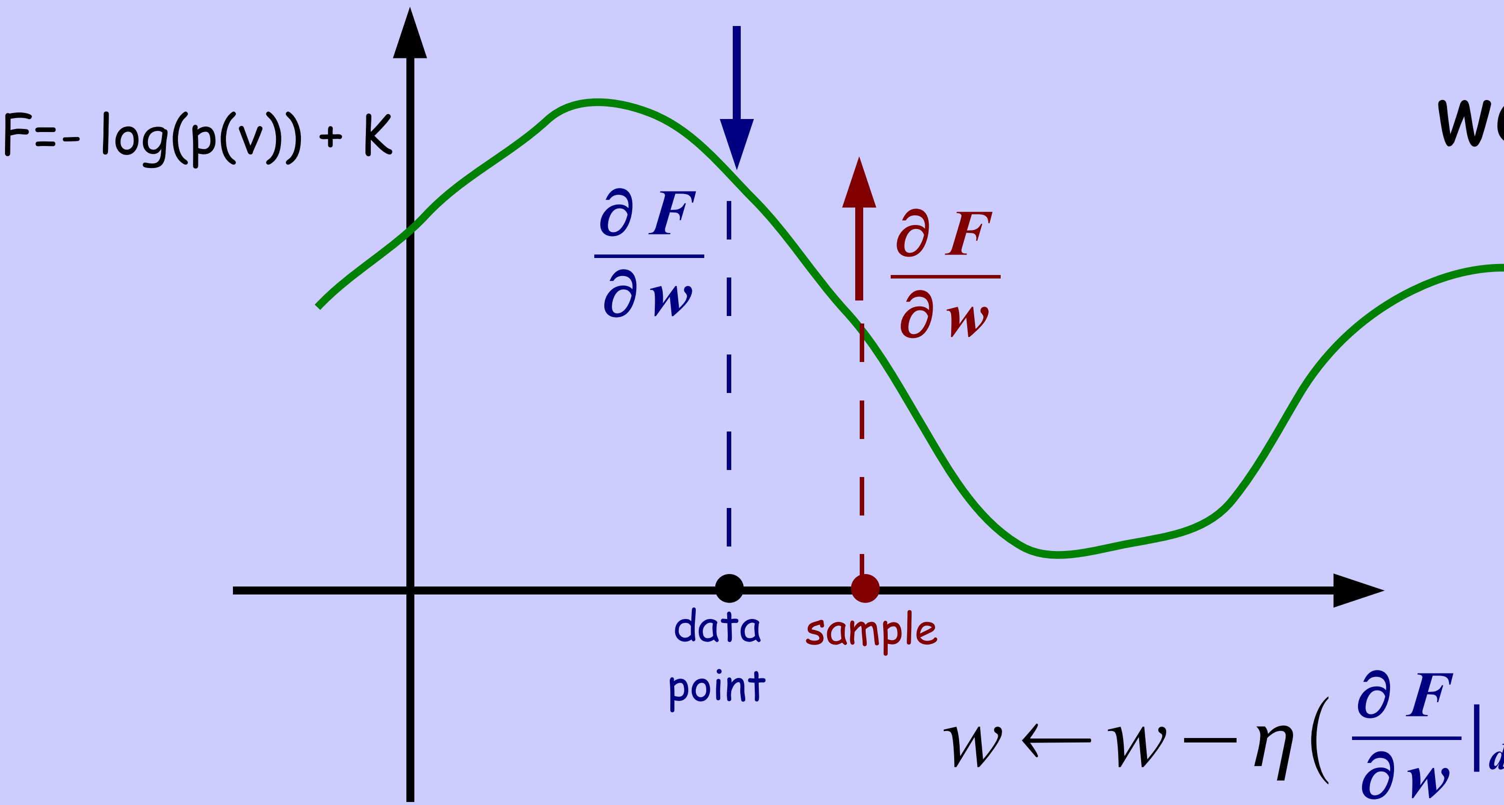
- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



HMC dynamics

LEARNING

- maximum likelihood
  - Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



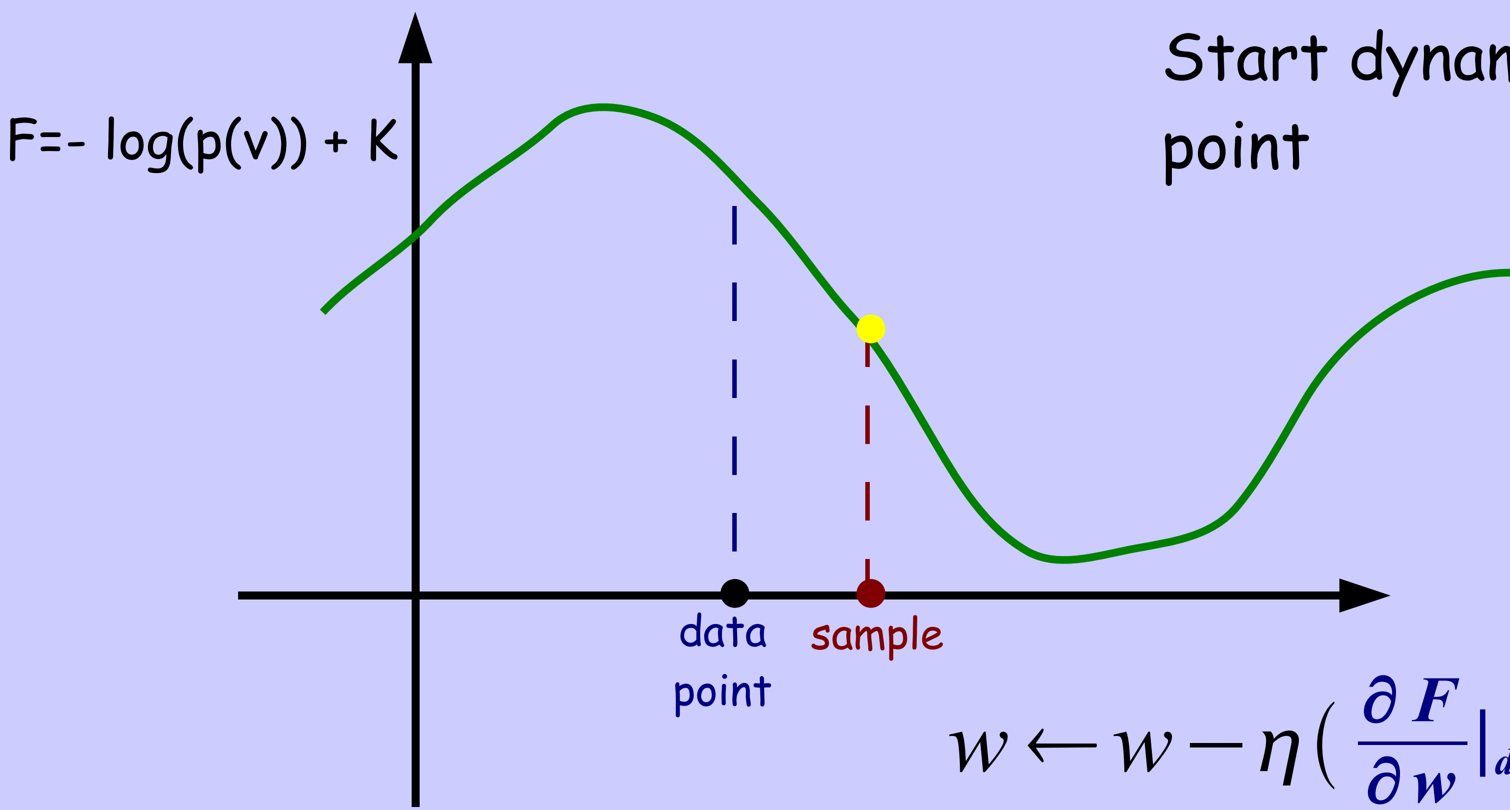
weight update

$$w \leftarrow w - \eta \left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample} \right)$$

LEARNING

- maximum likelihood
  - Persistent Contrastive Divergence
  - Hybrid Monte Carlo to draw samples

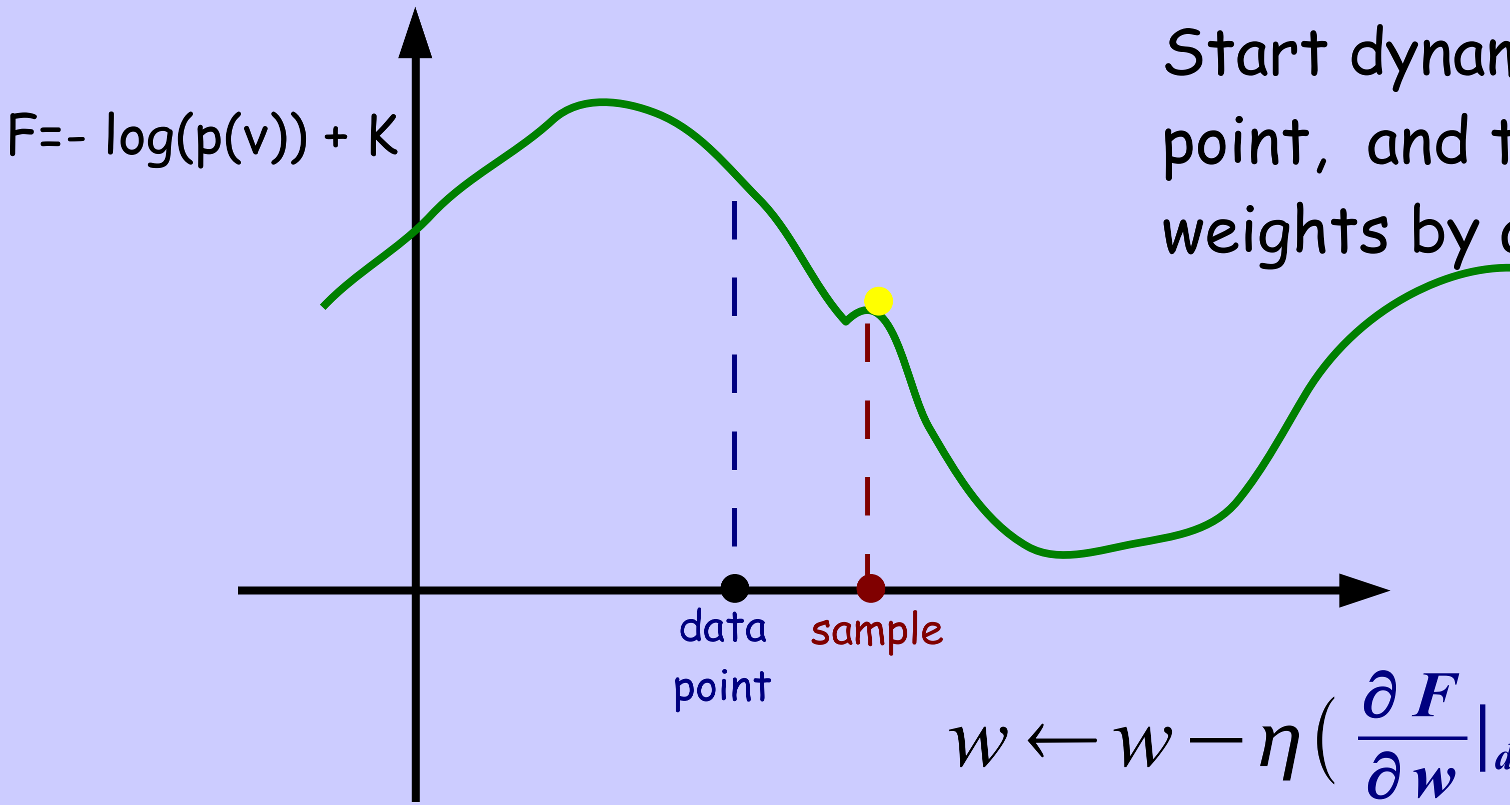
Start dynamics from previous point



$$w \leftarrow w - \eta\left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample} \right)$$

LEARNING

- maximum likelihood
  - Fast Persistent Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



Start dynamics from previous point, and temporarily perturb weights by a little bit

$$w \leftarrow w - \eta\left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample} \right)$$

Tieleman and Hinton ICML 2009

LEARNING

- maximum likelihood
  - Fast Persistent Contrastive Divergence

Initialize: $w, w_f, \eta < \eta_f$

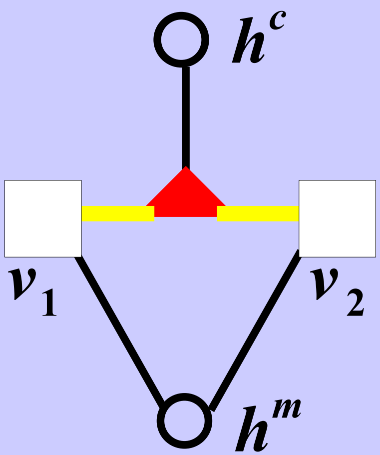
for each training data case do:

- get training sample: v^+
- compute derivatives: $g^+ = \partial F / \partial w|_{v^+}$
- draw sample: $v^- \leftarrow HMC(v^-; w + w_f)$
- compute derivatives: $g^- = \partial F / \partial (w + w_f)|_{v^-}$
- update true parameters: $w \leftarrow w - \eta(g^+ - g^-)$
- update fast weights: $w_f \leftarrow 0.95 w_f - \eta_f (g^+ - g^-)$

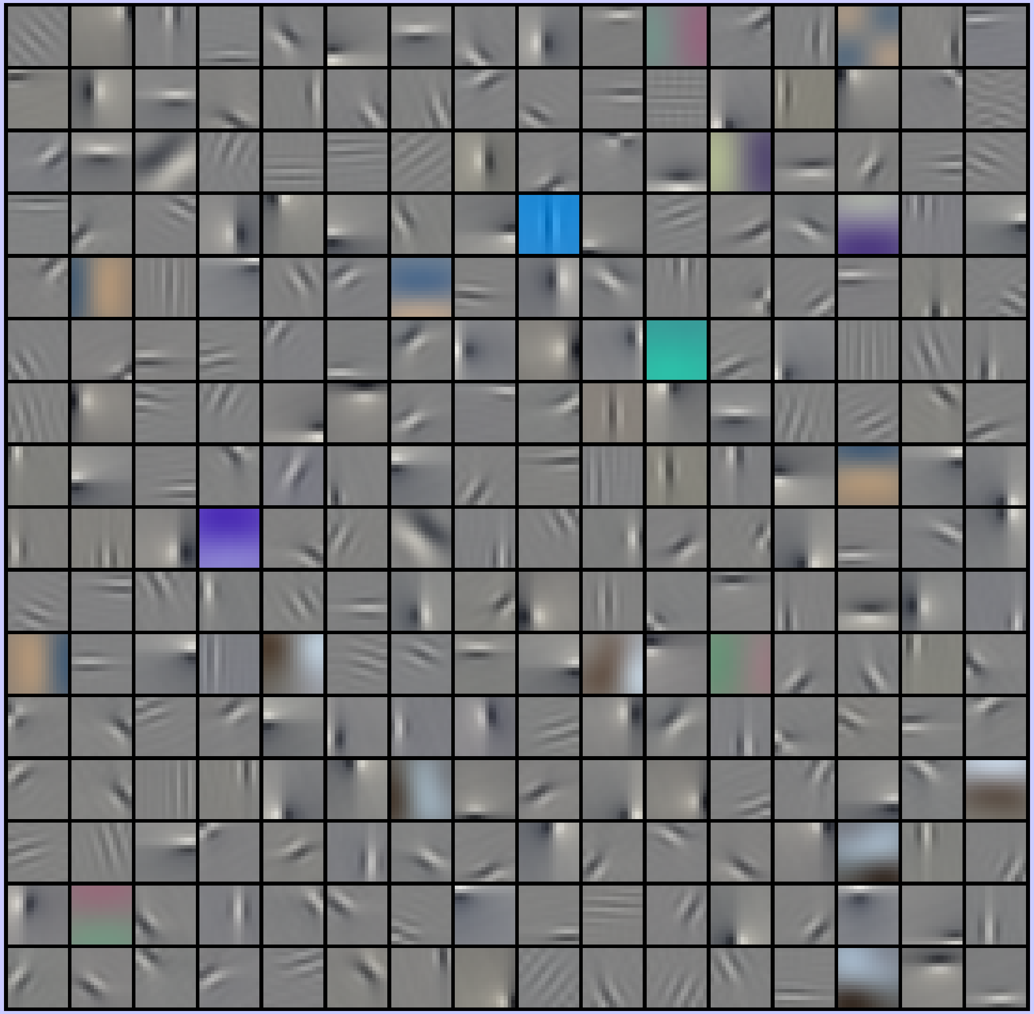
Tieleman and Hinton ICML 2009

Experiments

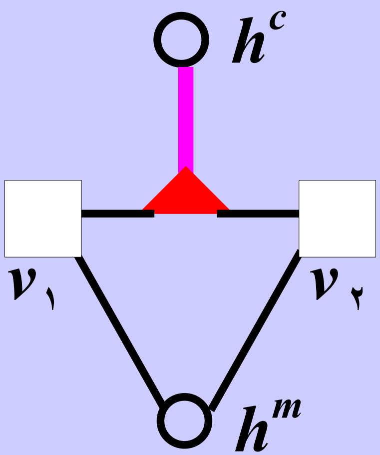
- Learn from 16x16 natural image patches
  - pre-processing: PCA whitening



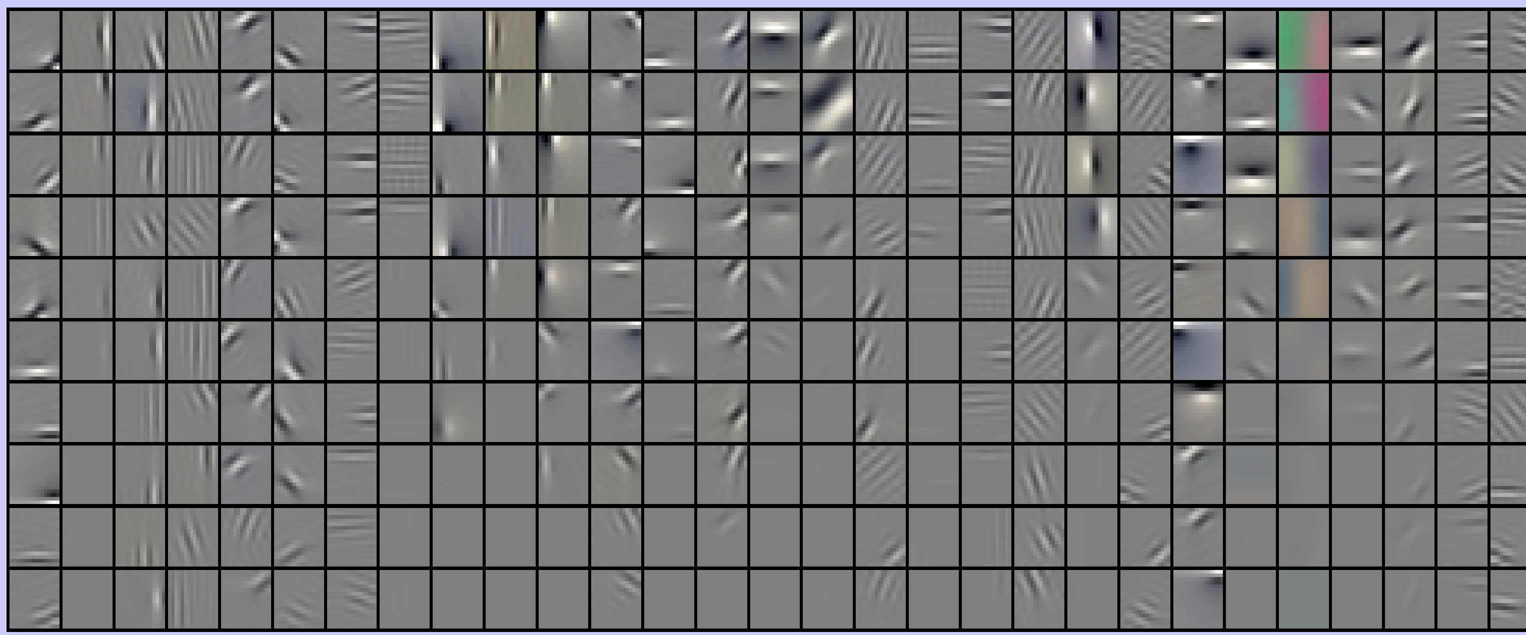
covariance filters



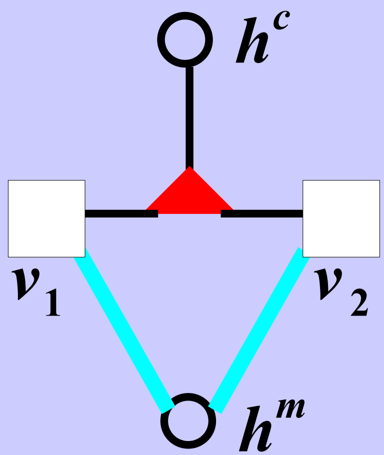
- Learn from 16x16 natural image patches
  - pre-processing: PCA whitening



grouping of covariance filters



- Learn from 16x16 natural image patches
  - pre-processing: PCA whitening

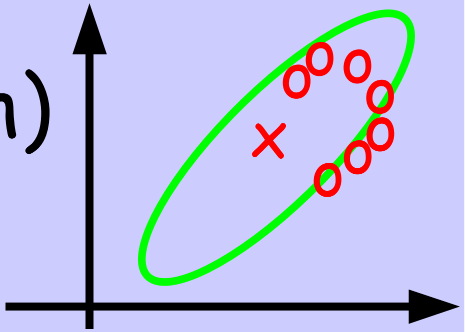


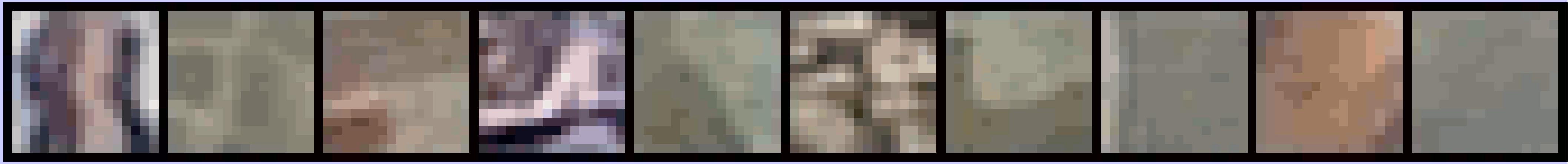
mean intensity filters

1) given image -> infer latent variables using p(h|v)

2) keeping latent variables fixed, sample from p(v|h)

random walk in input space sampling p(v|h)

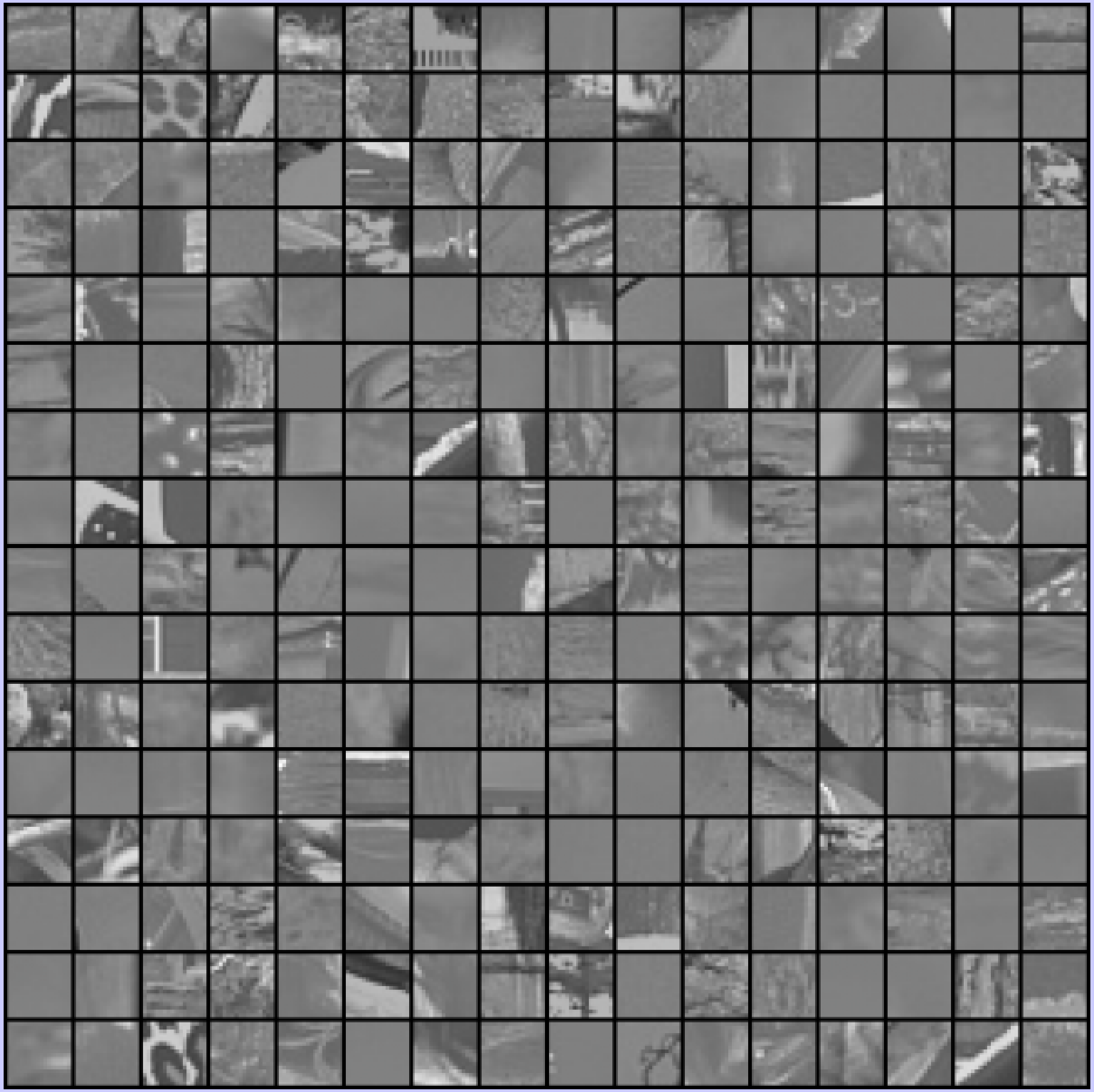




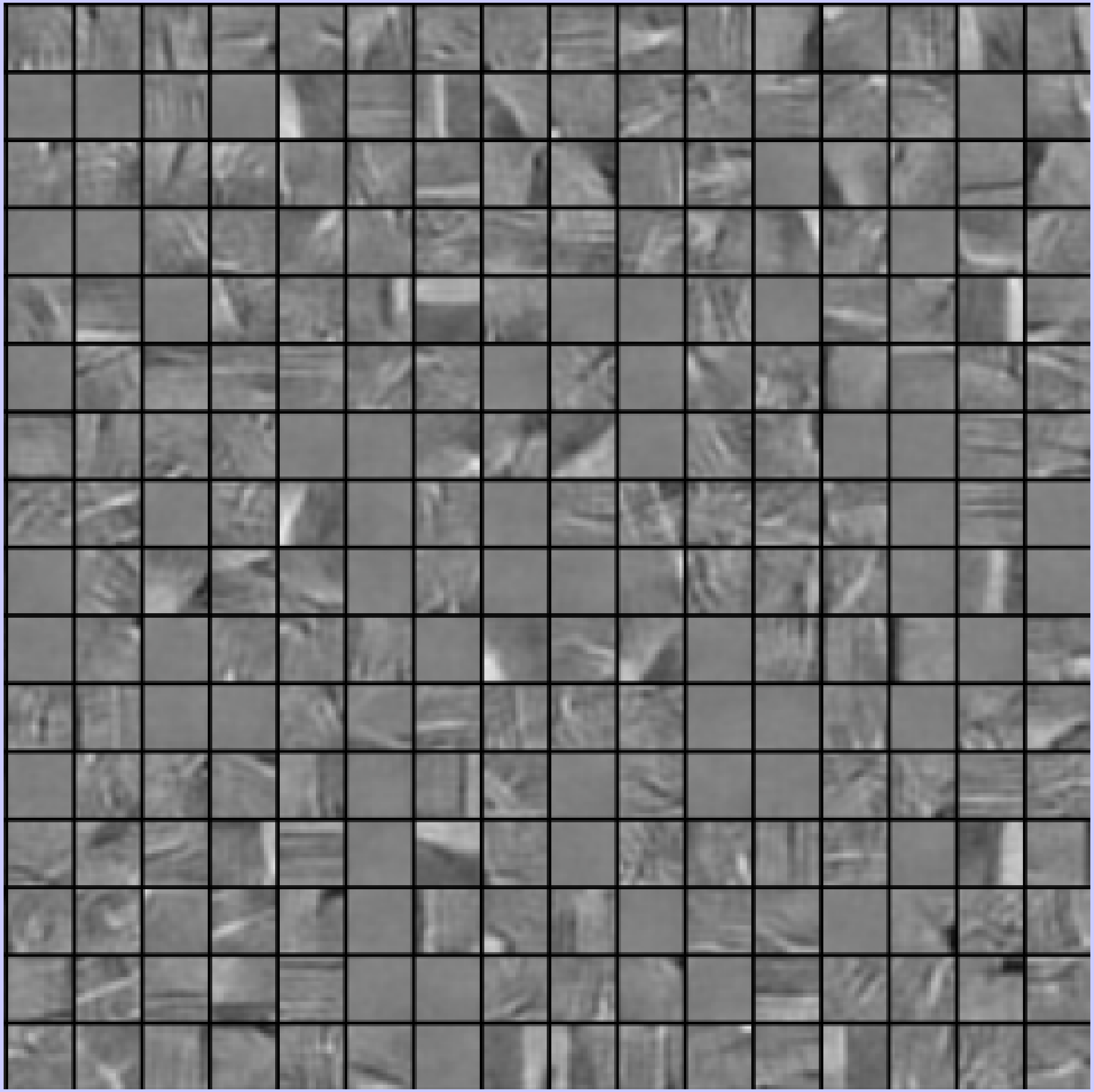
The latent configuration induces a whole subspace of images.

The latent representation learns to be robust to small distortions.

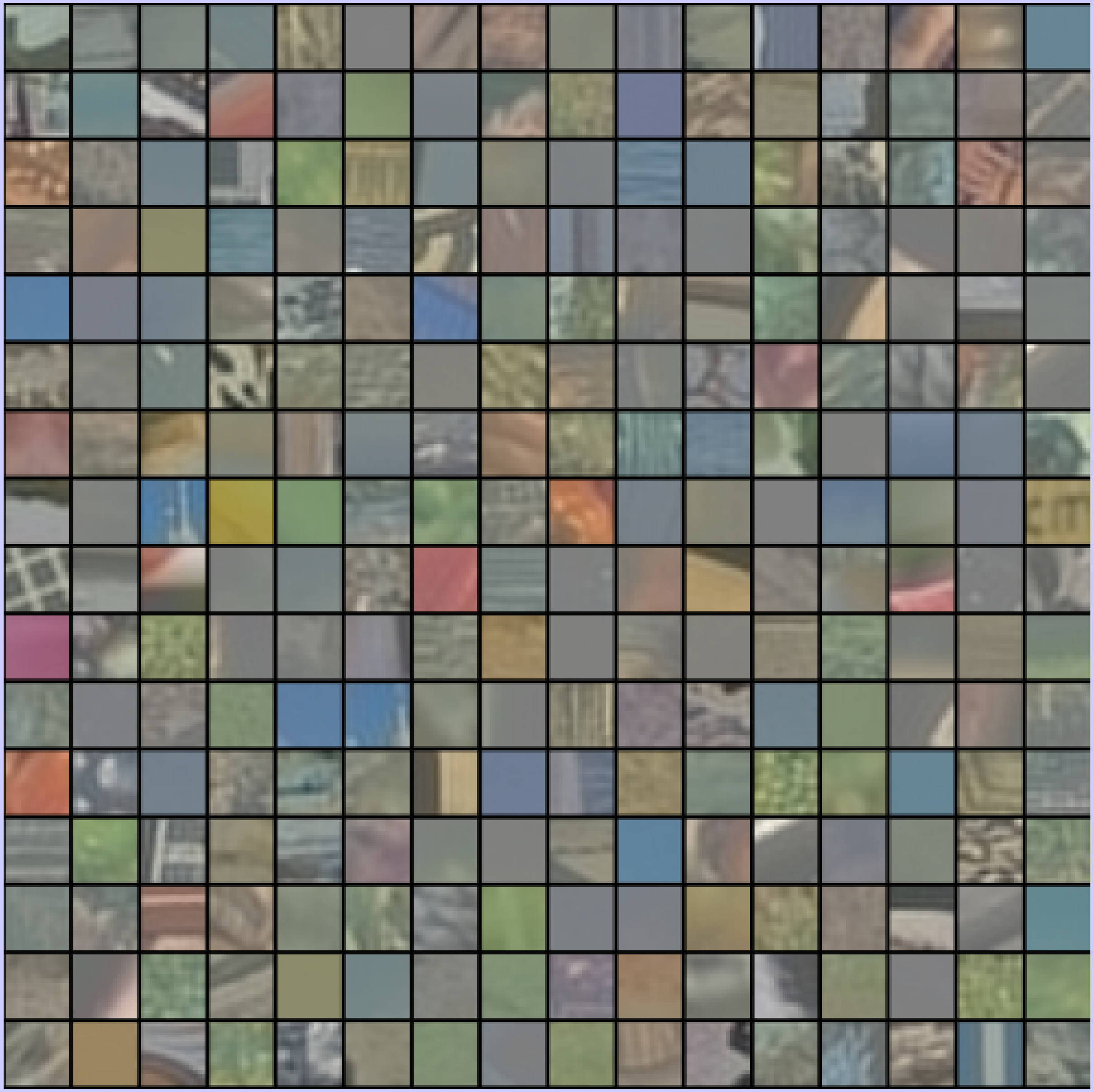
Example of image patches used during training



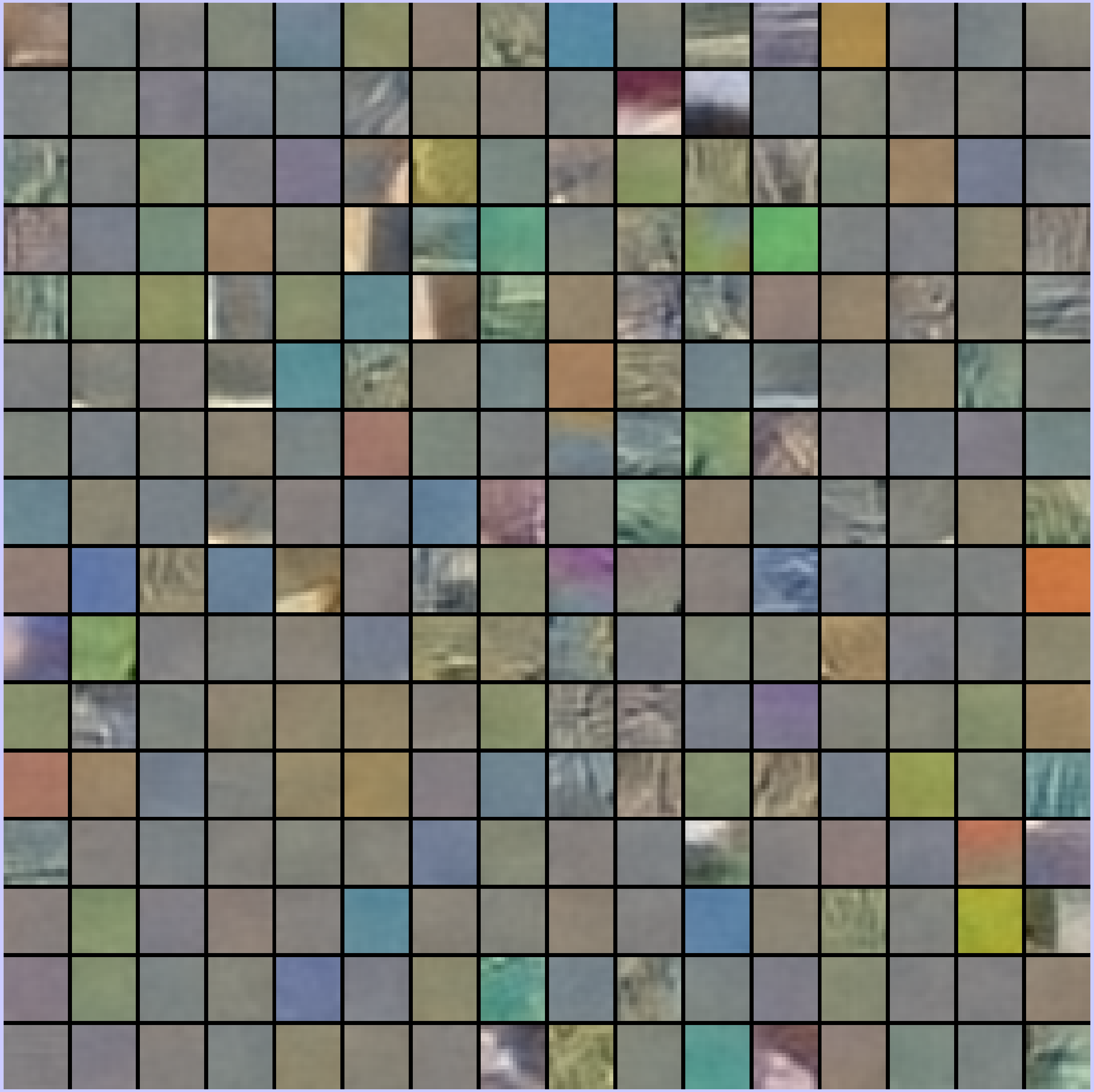
Samples drawn from the model (using HMC)



Example of image patches used during training

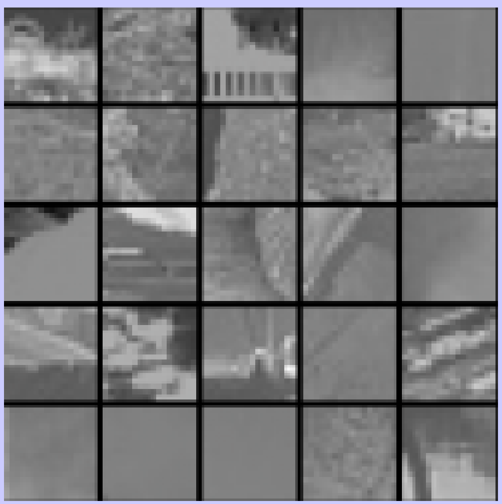


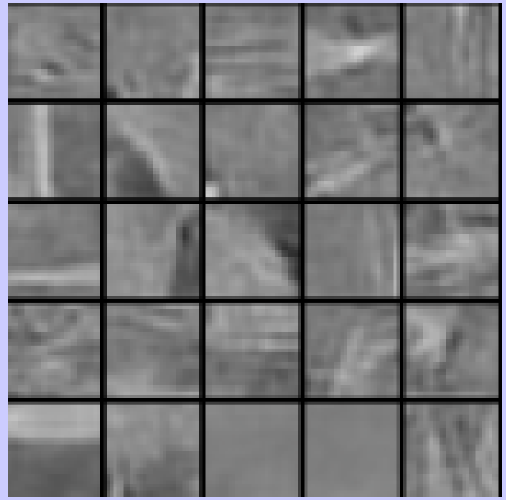
Samples drawn from the model (using HMC)



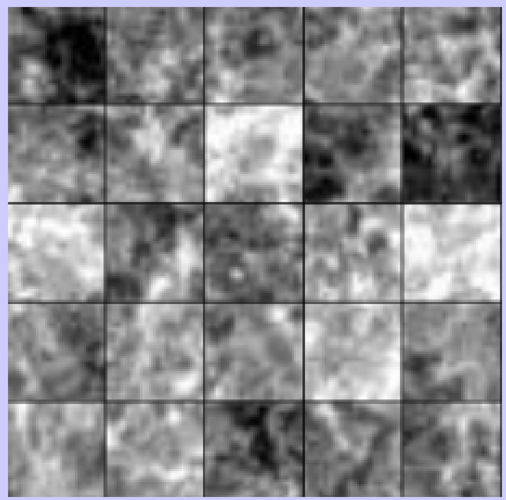
Comparison

Natural images



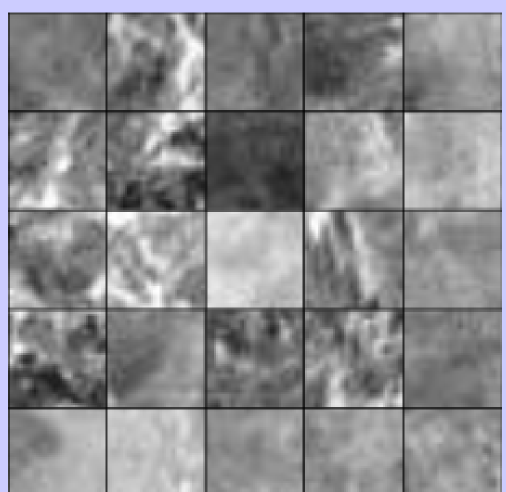


mcRBM



GRBM

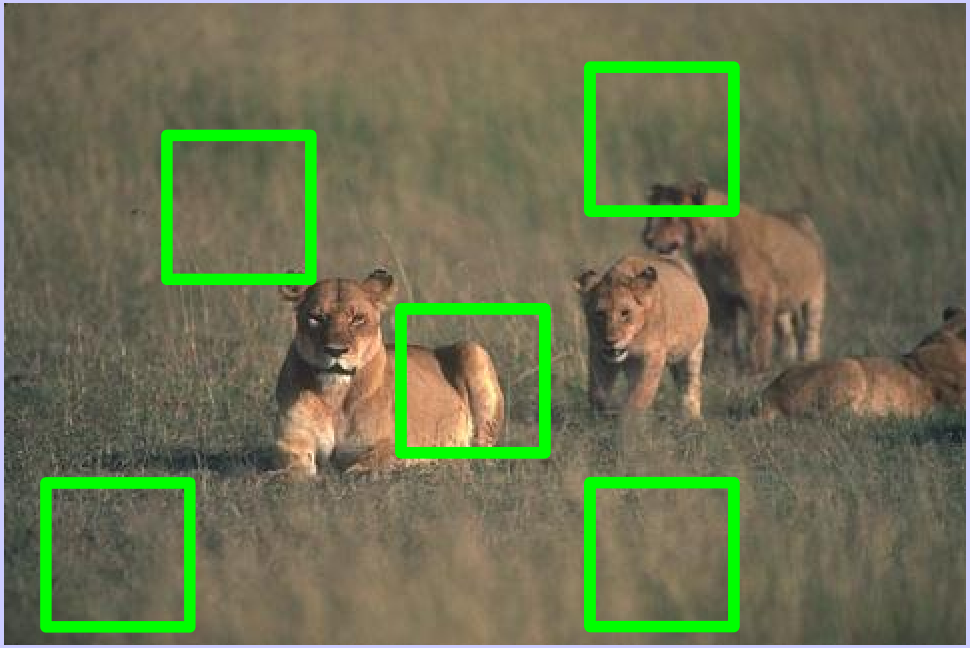
from Osindero and Hinton NIPS 2008



S-RBM + DBN

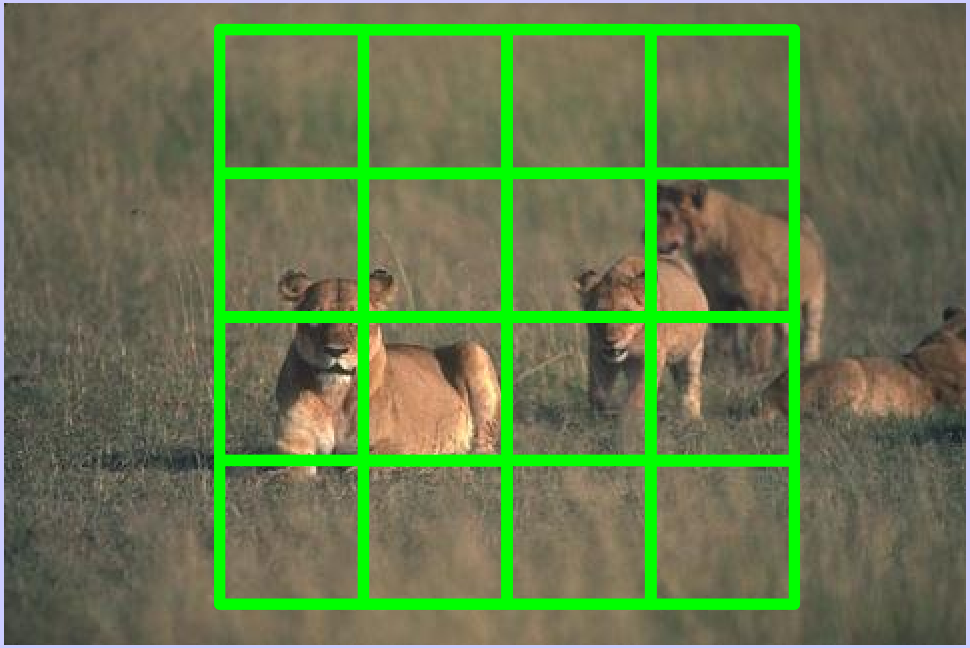
from Osindero and Hinton NIPS 2008

From patches to big images



Training by picking patches at random

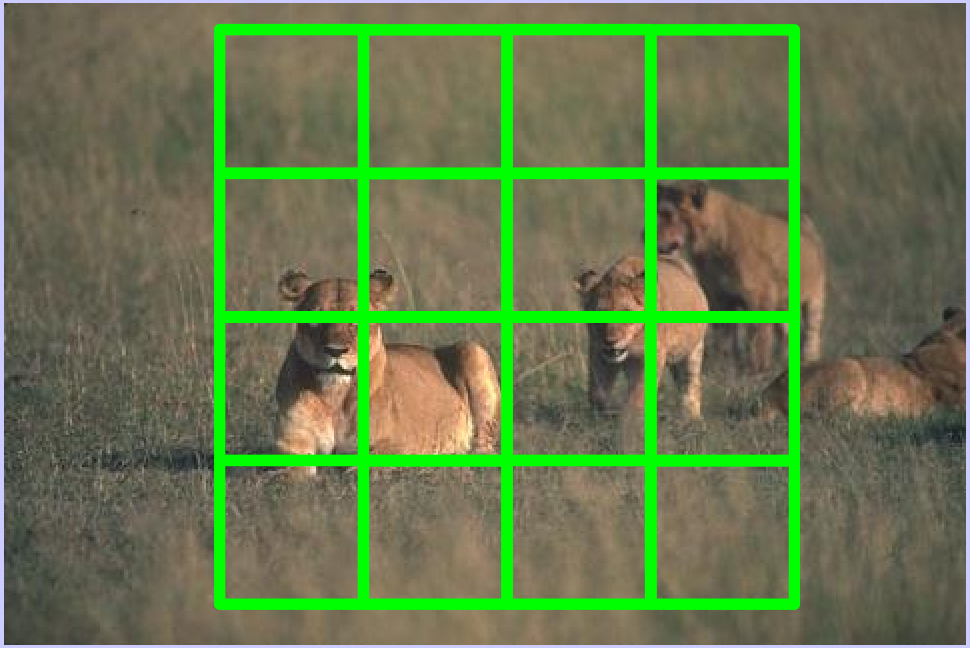
From patches to big images



But we could also take them from a grid

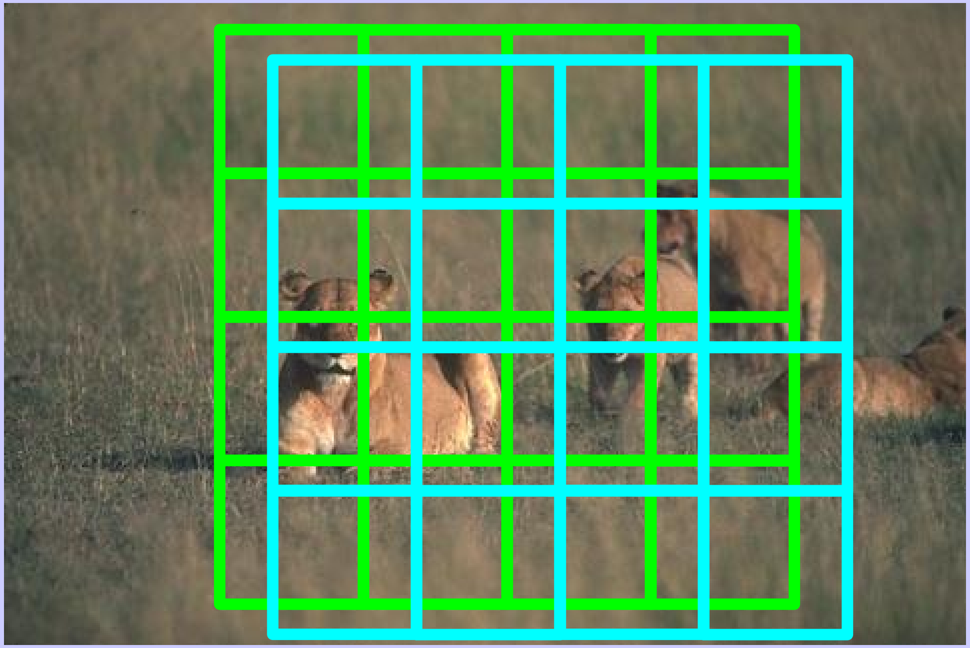
This is not a good way to extend the model to big images: block artifacts

From patches to big images



But a subset of filters applied to these patches and...

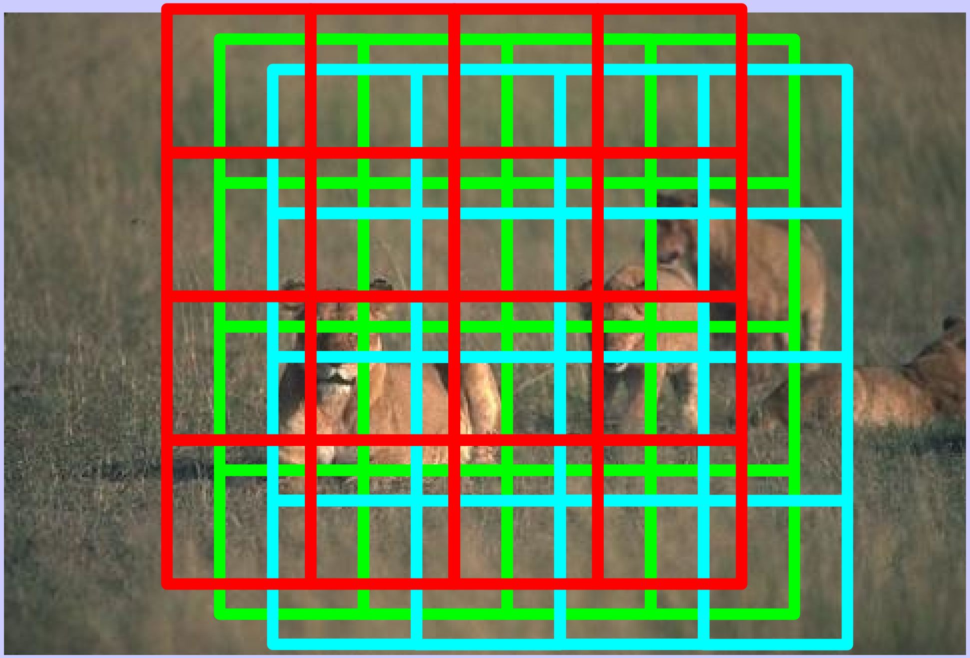
From patches to big images



But a subset of filters applied to these patches and...
other subsets applied to shifted grids

Gregor LeCun 2010, our paper in submission

From patches to big images



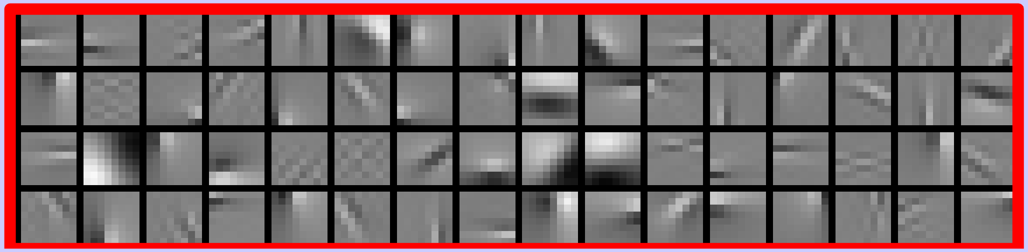
But a subset of filters applied to these patches and...
other subsets applied to shifted grids

no block artifacts & little redundancy

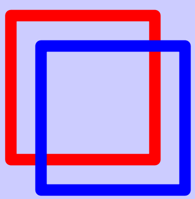
Gregor LeCun 2010, our paper in submission

Learning on high-resolution images

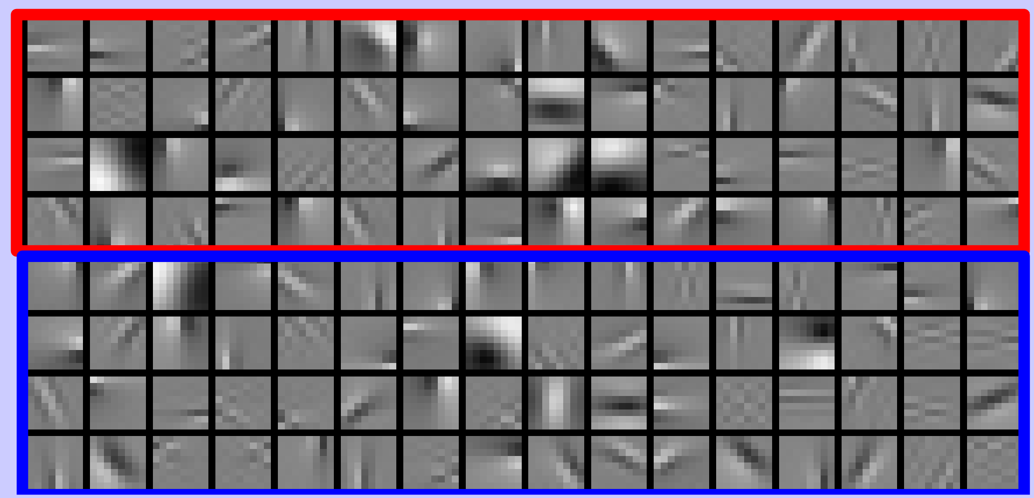
covariance filters



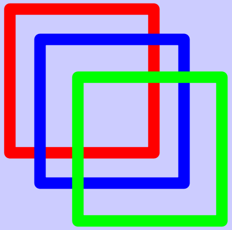
Learning on high-resolution images



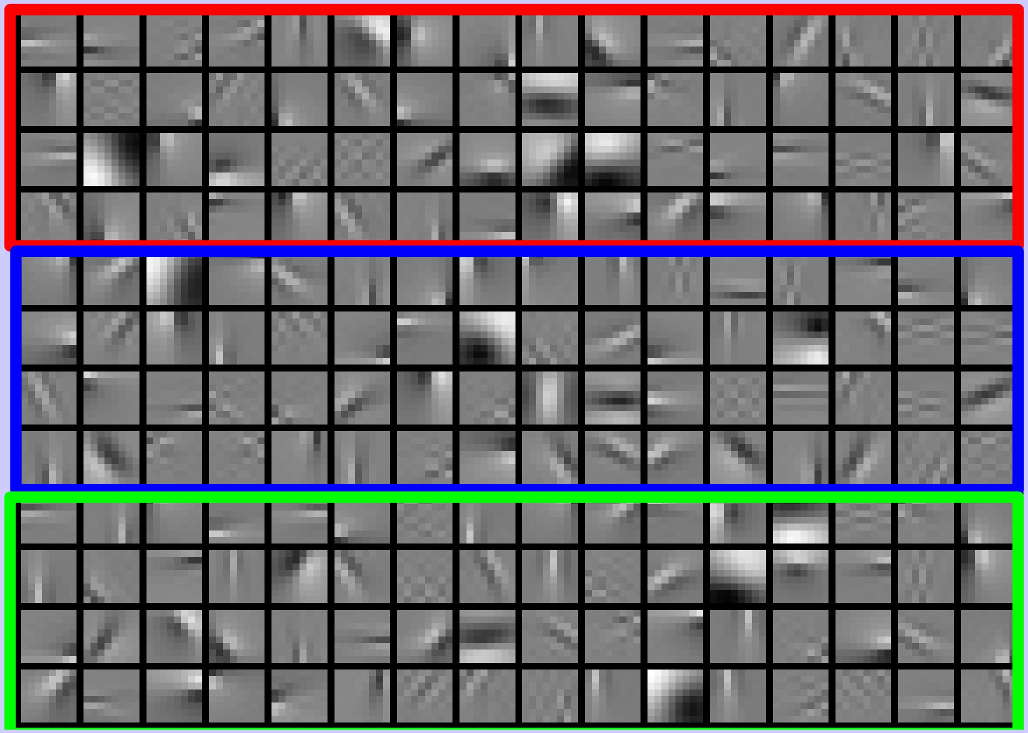
covariance filters



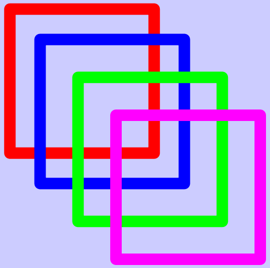
Learning on high-resolution images



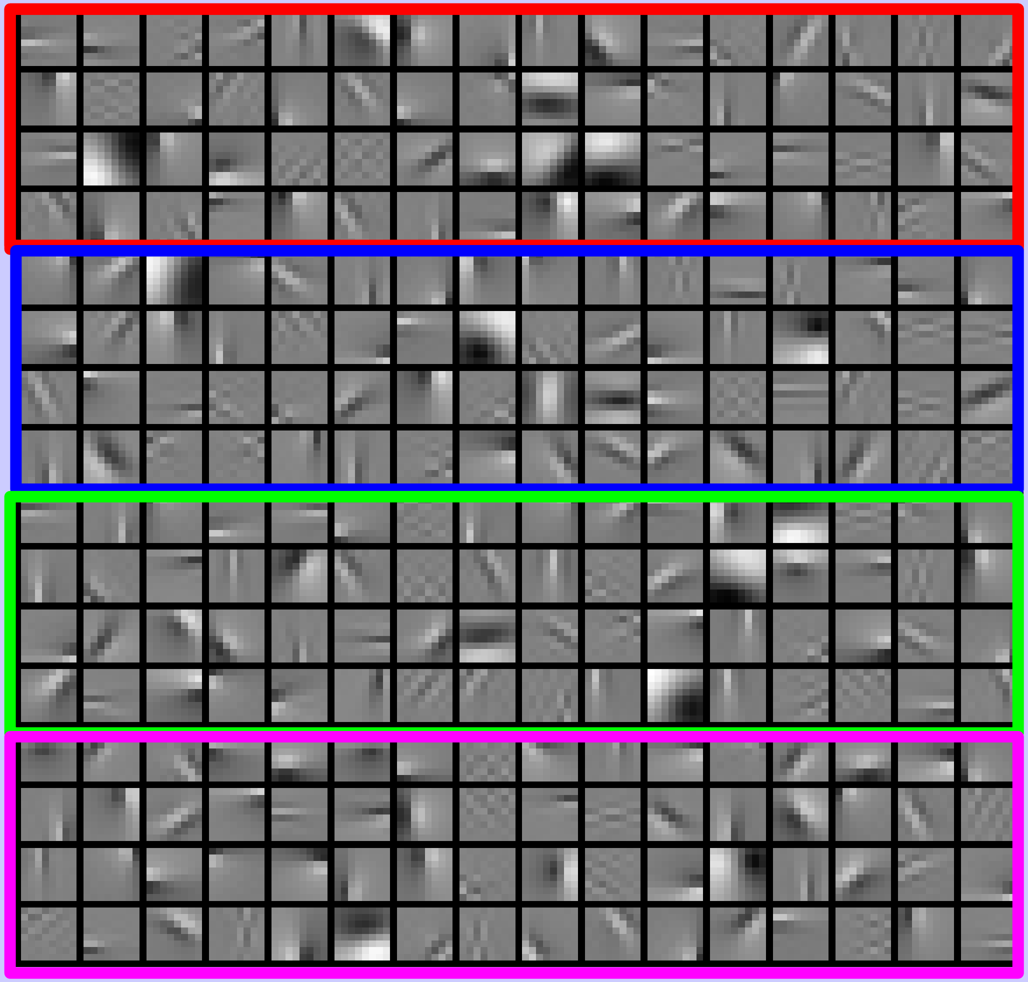
covariance filters



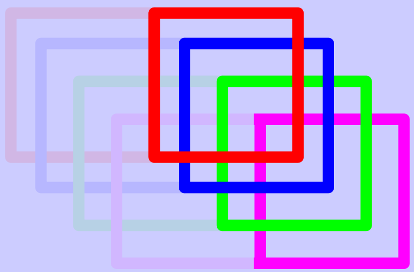
Learning on high-resolution images



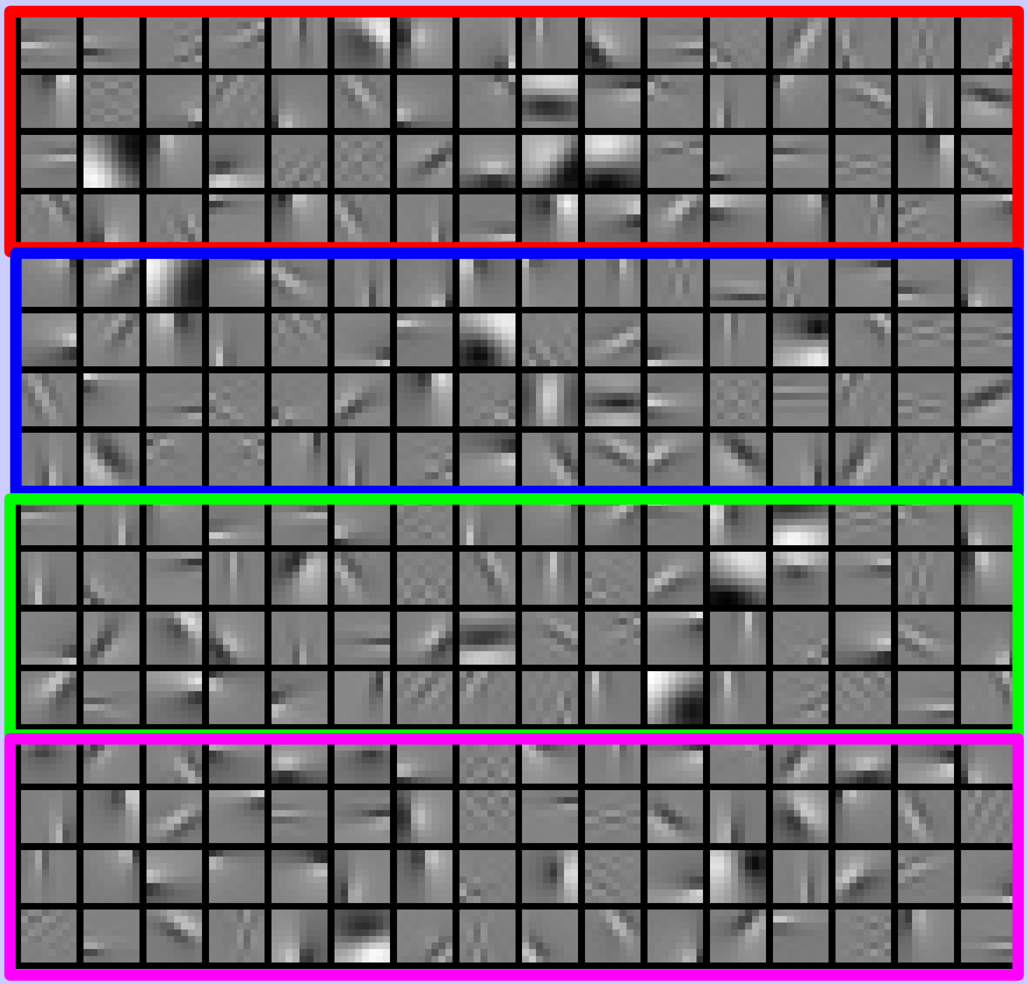
covariance filters



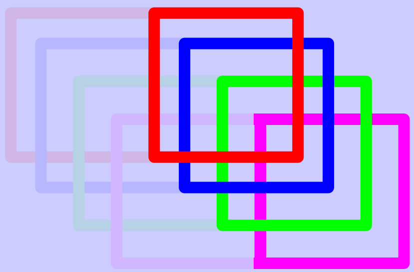
Learning on high-resolution images

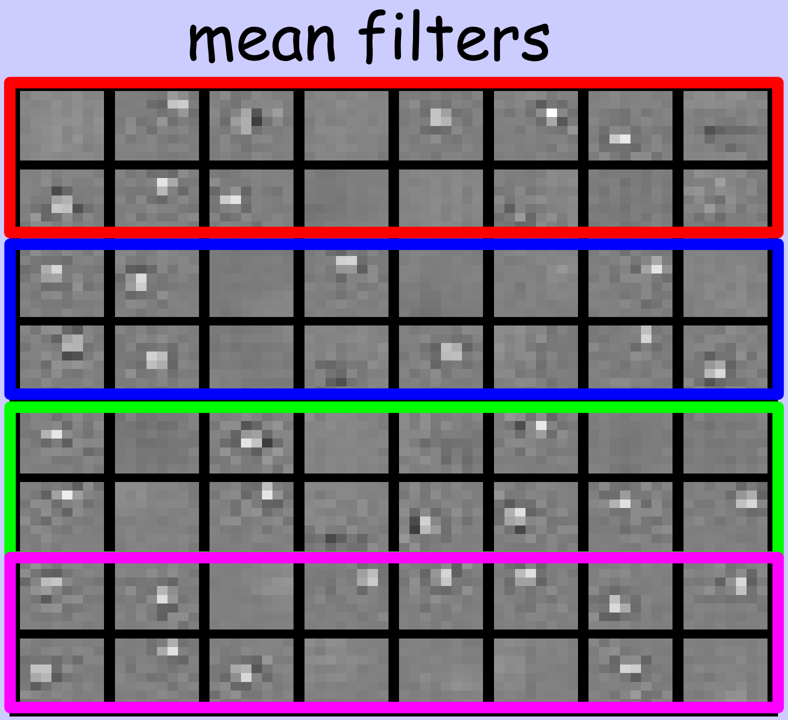


covariance filters



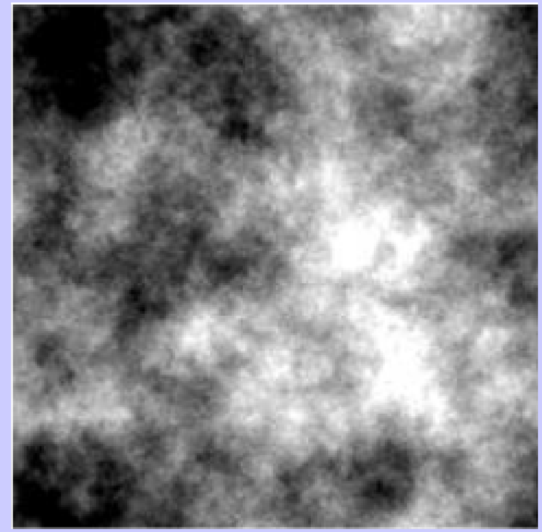
Learning on high-resolution images





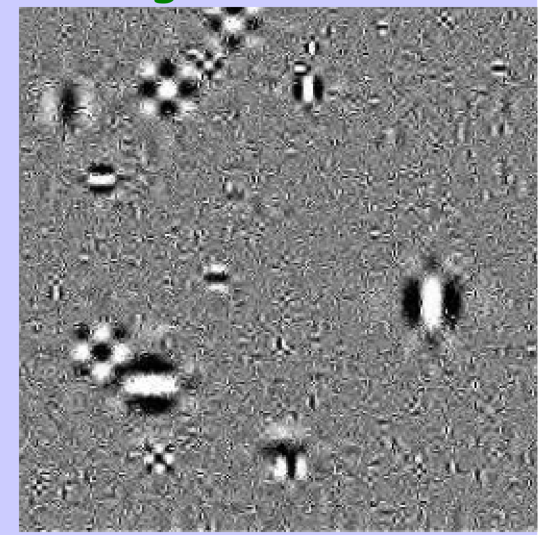
Sampling high resolution images

Gaussian model



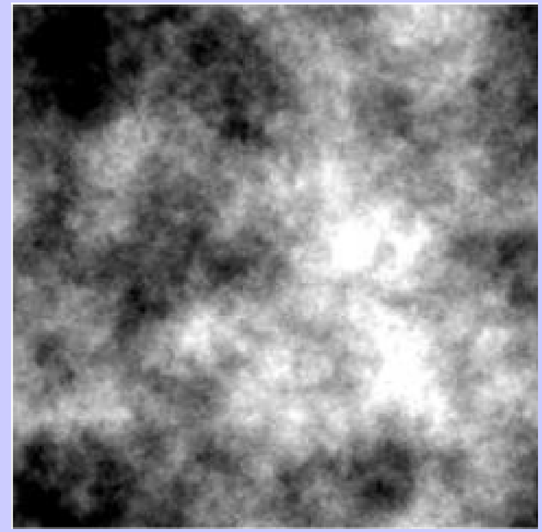
from Simoncelli 2005

marginal wavelet

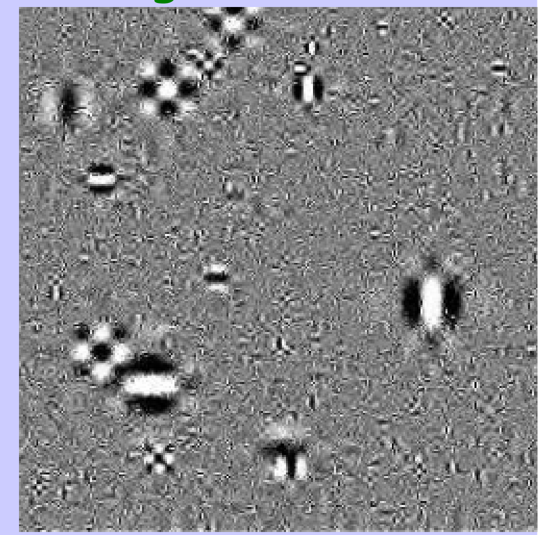


Sampling high resolution images

Gaussian model

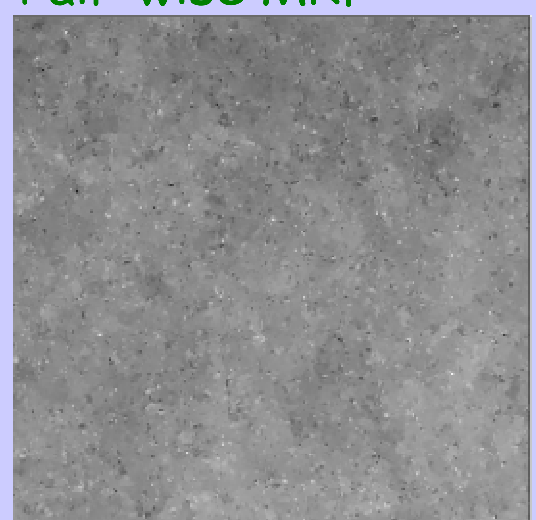


marginal wavelet



from Simoncelli 2005

Pair-wise MRF

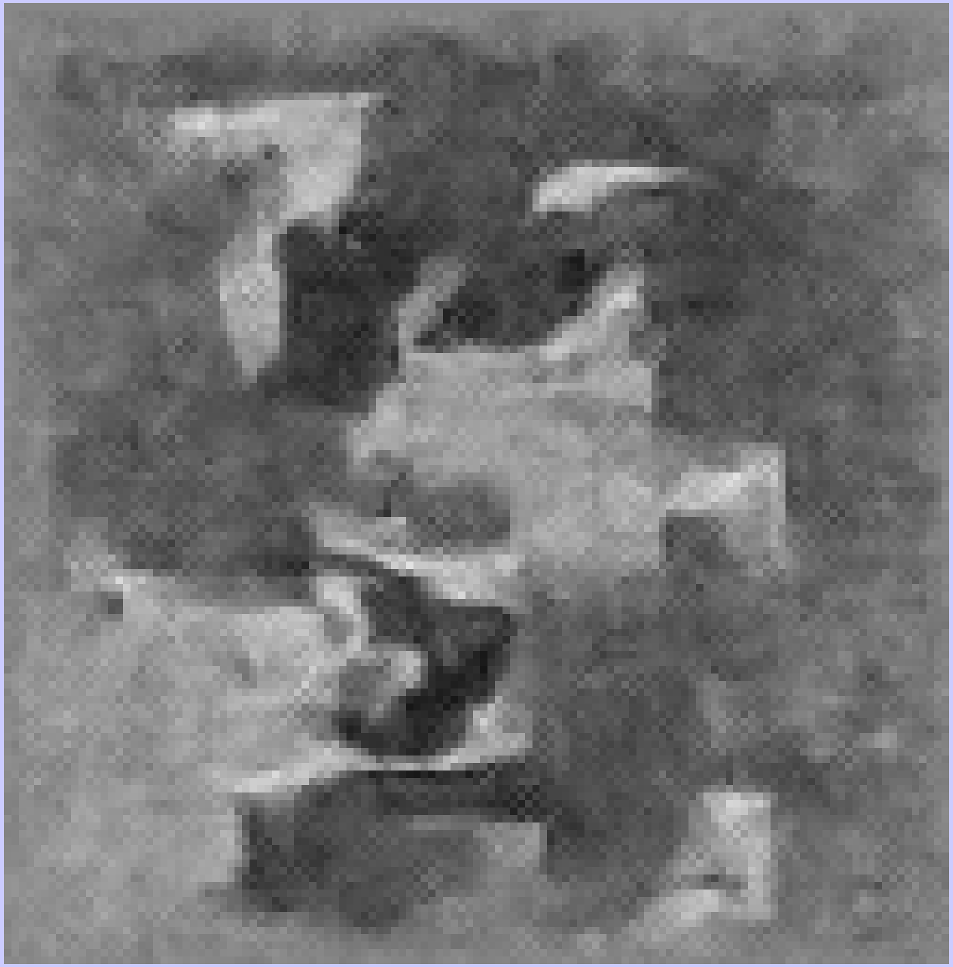


FoE

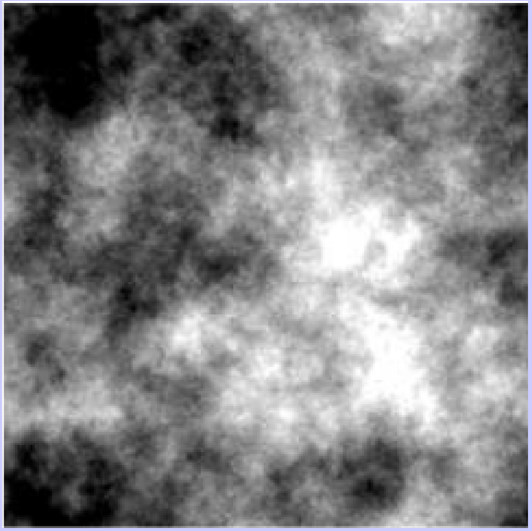
from Schmidt, Gao, Roth CVPR 2010

Sampling high resolution images

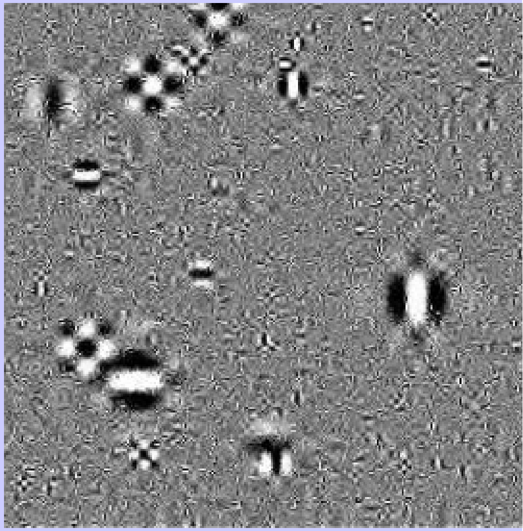
Mean Covariance Model



Gaussian model

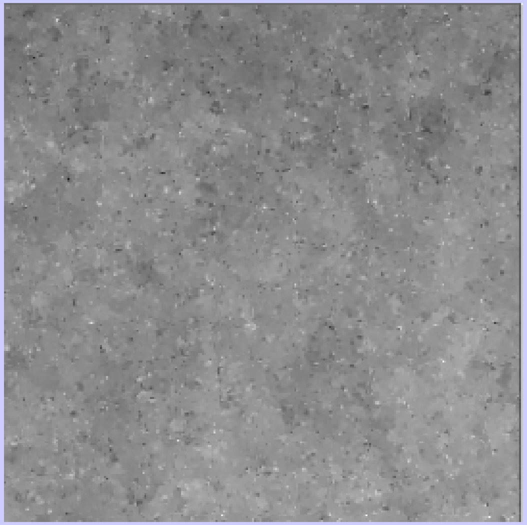


marginal wavelet



from Simoncelli 2005

Pair-wise MRF

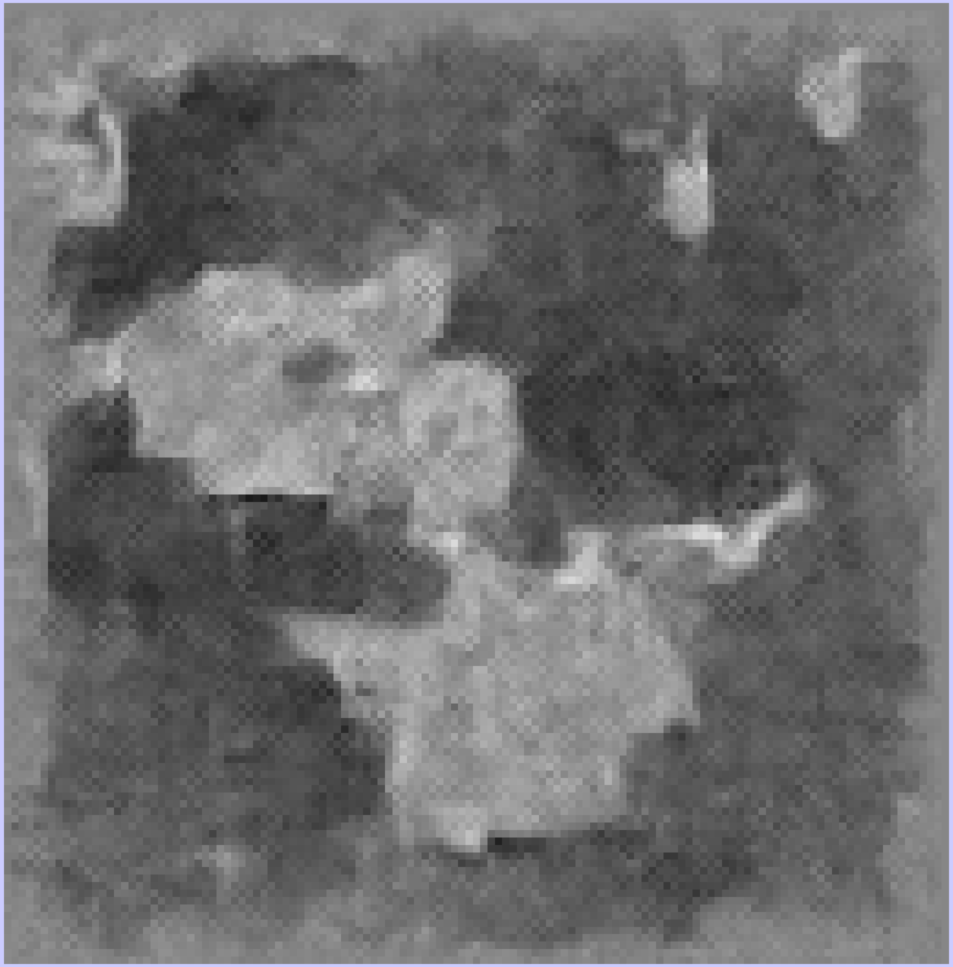


FoE

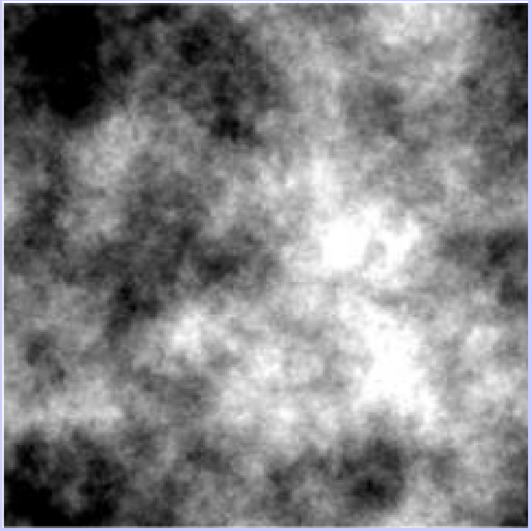
from Schmidt, Gao, Roth CVPR 2010

Sampling high resolution images

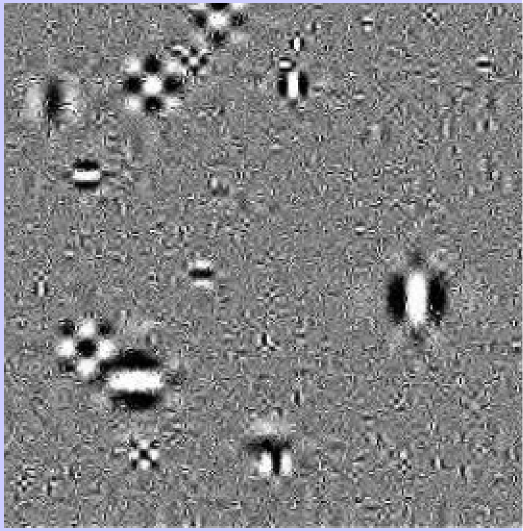
Mean Covariance Model



Gaussian model

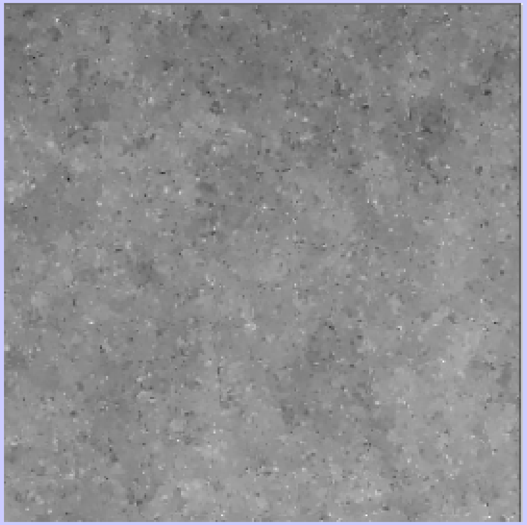


marginal wavelet



from Simoncelli 2005

Pair-wise MRF



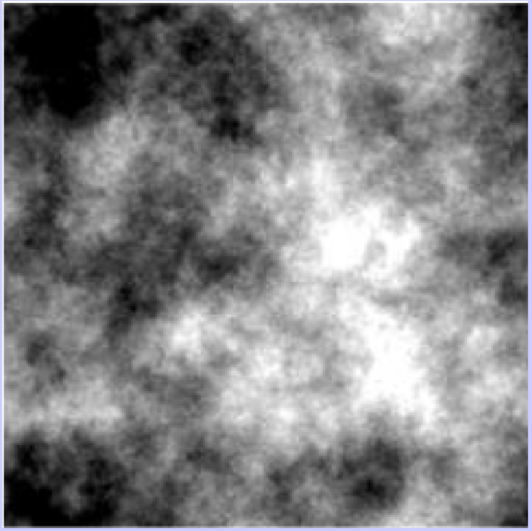
FoE

from Schmidt, Gao, Roth CVPR 2010

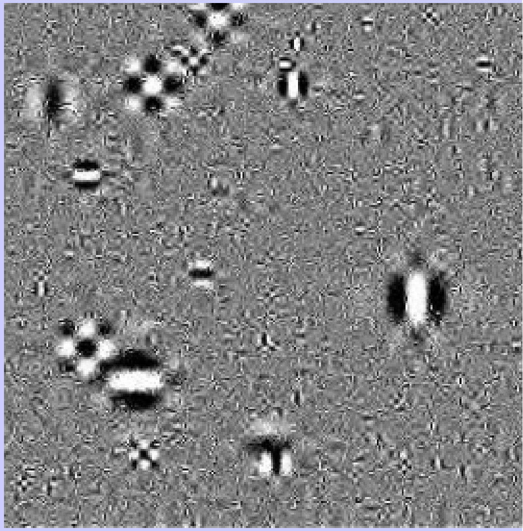
Sampling high resolution images

Mean Covariance Model

Gaussian model

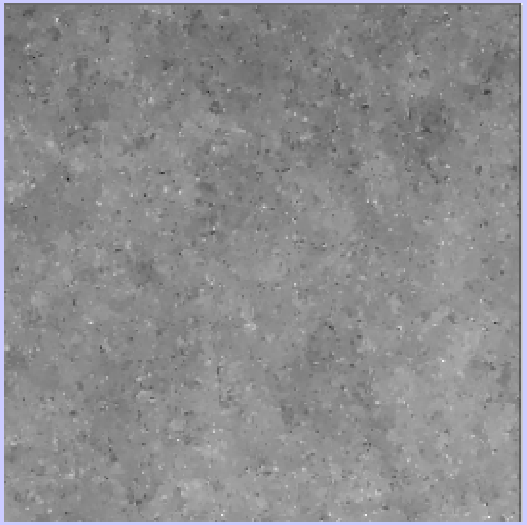


marginal wavelet



from Simoncelli 2005

Pair-wise MRF



FoE

from Schmidt, Gao, Roth CVPR 2010

Sampling high resolution images

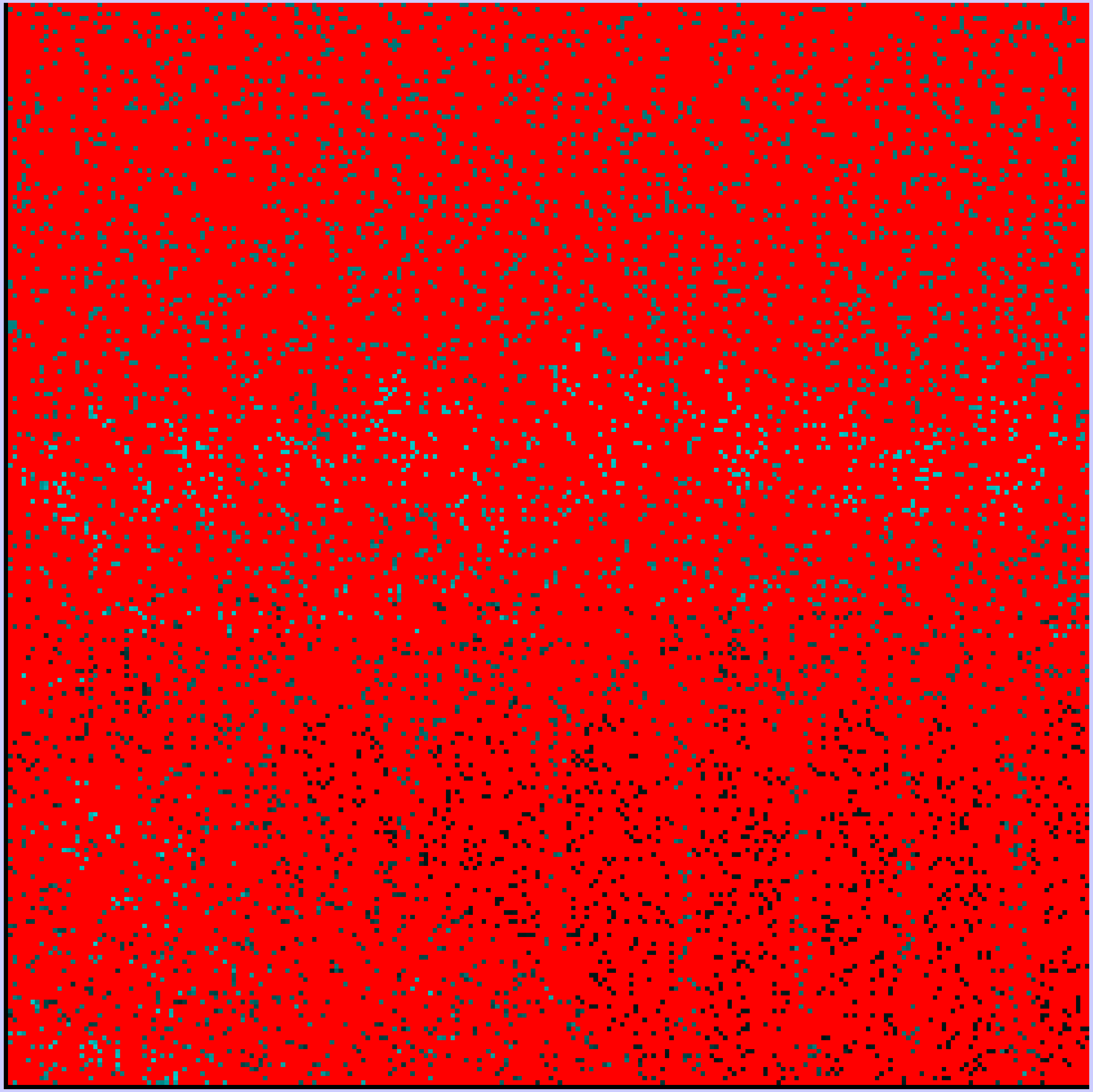
Sampling starting from natural image

Sampling high resolution images

Sampling starting from natural image

Inpainting: *guessing the 90% of the pixels*

masked input image

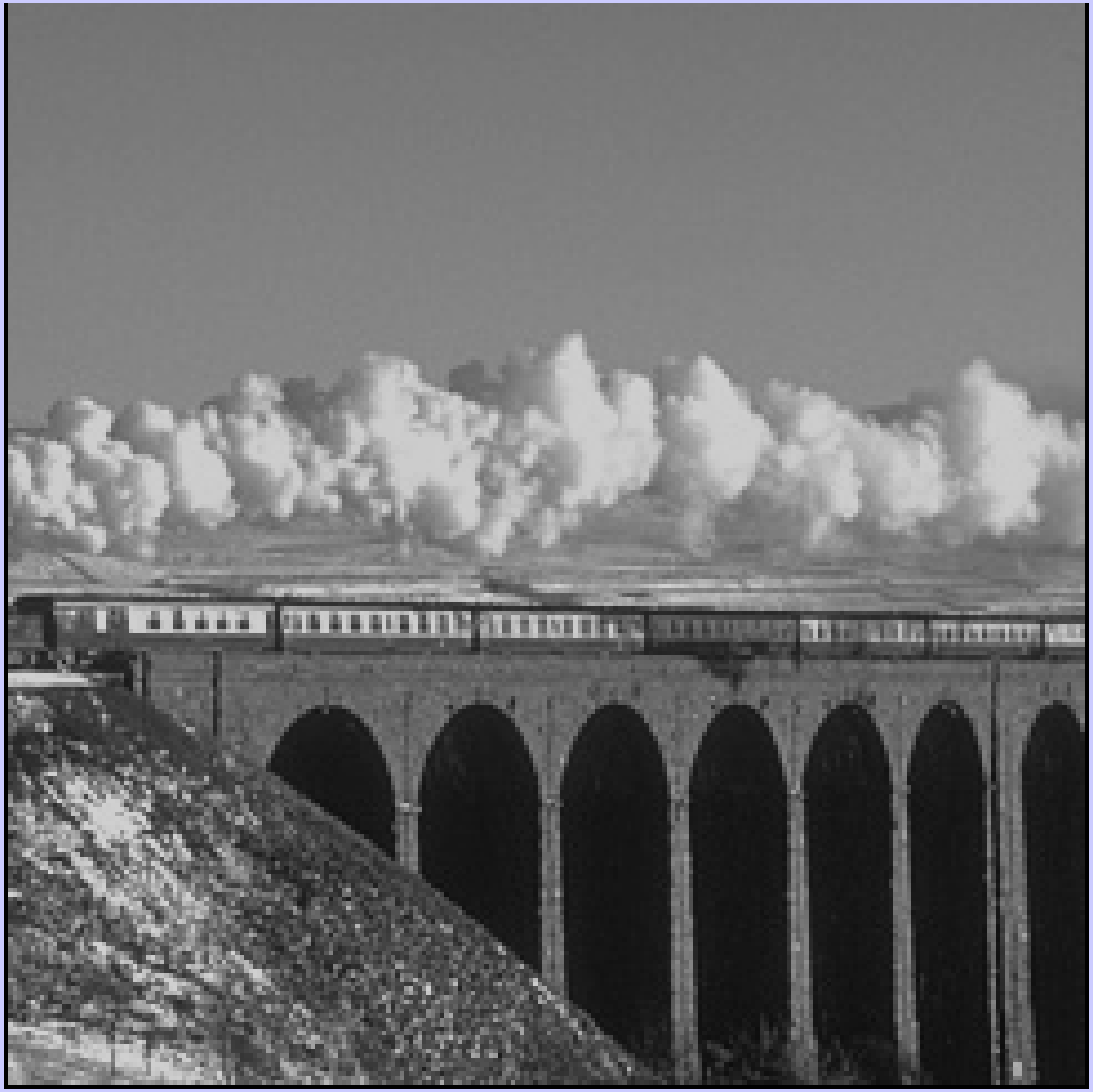


Inpainting: guessing the 90% of the pixels

MAP estimate missing pixels

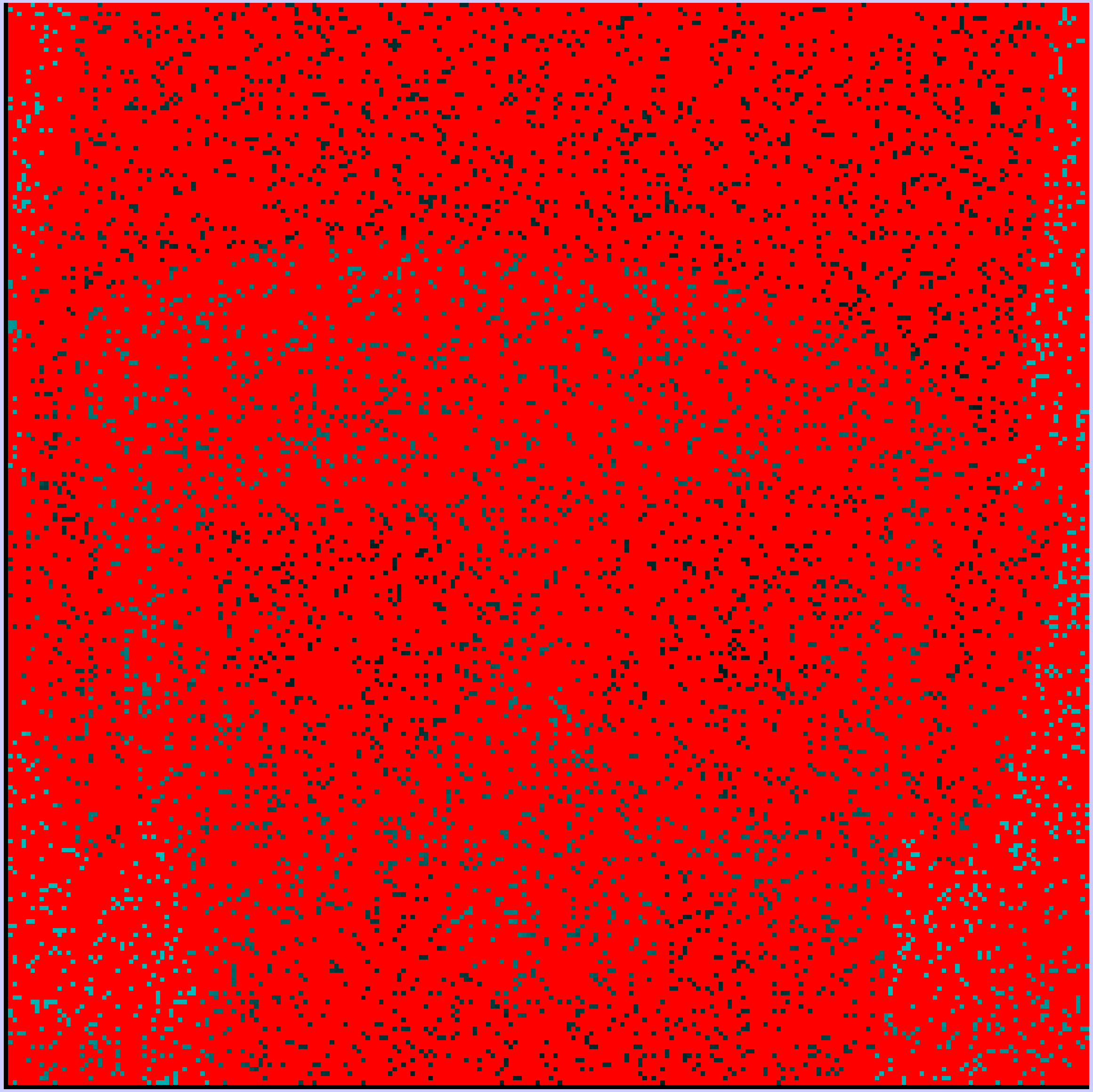
Inpainting: guessing the 90% of the pixels

true image



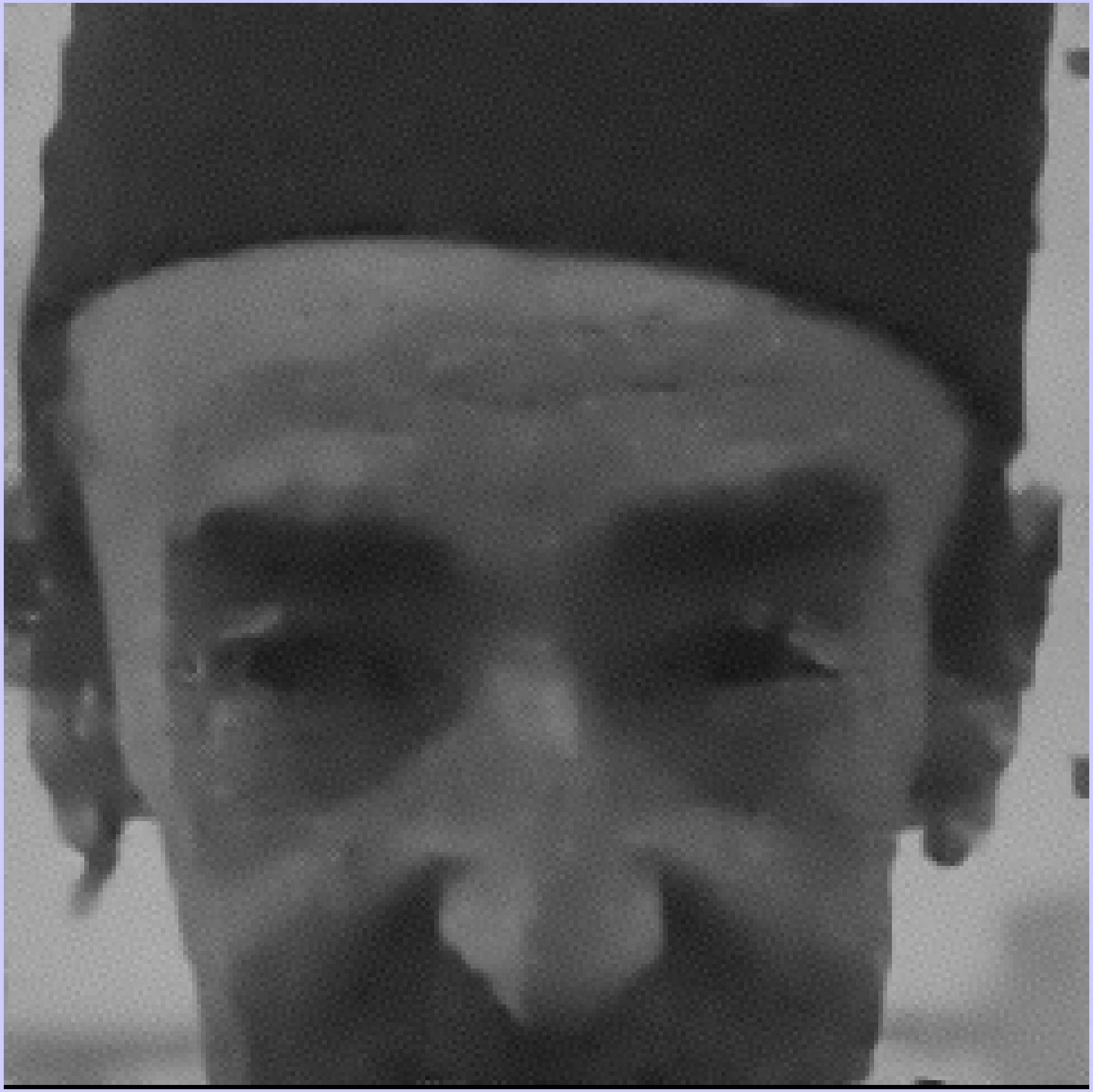
Inpainting: guessing the 90% of the pixels

masked input image



Inpainting: guessing the 90% of the pixels

MAP estimate missing pixels

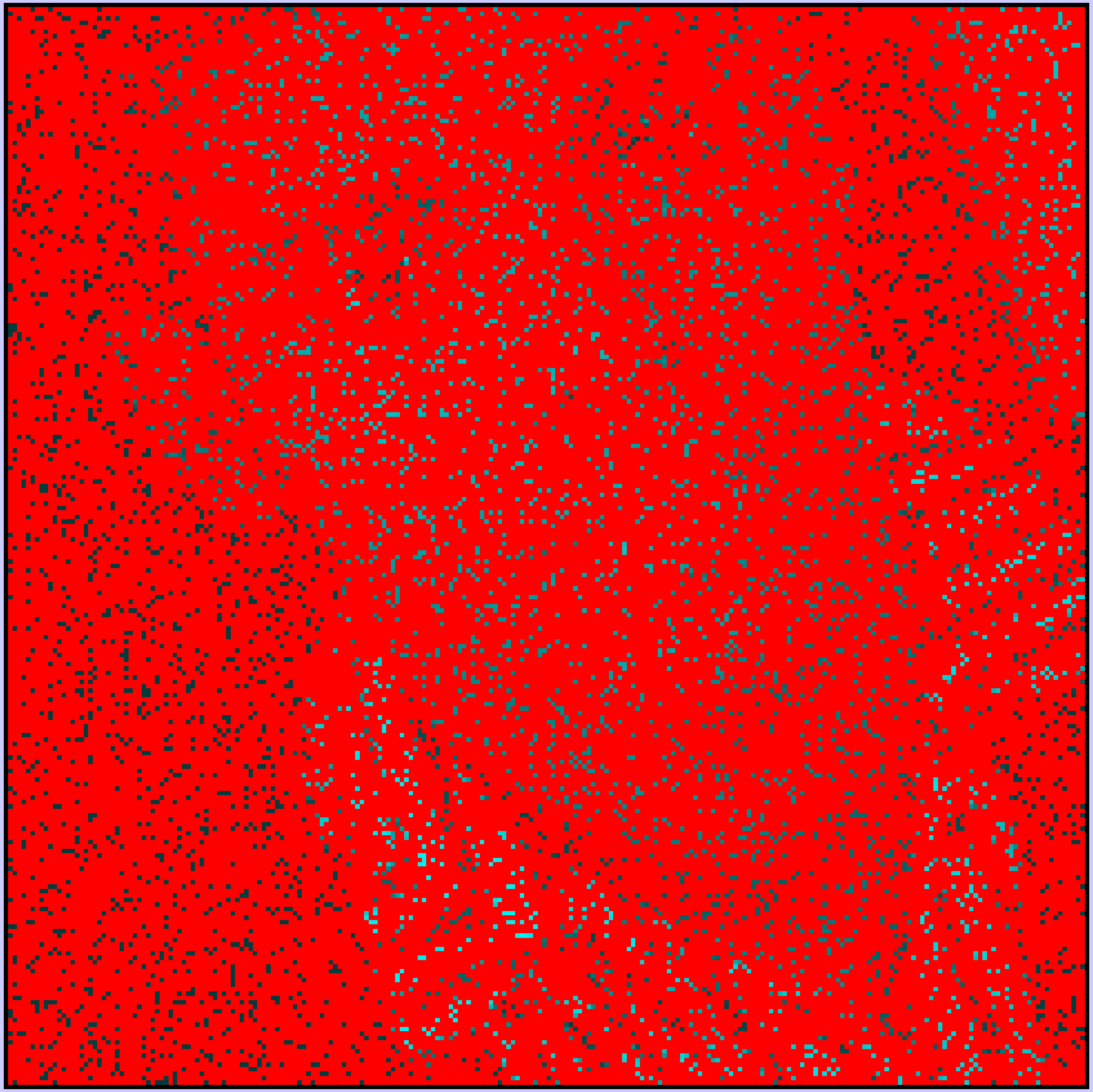


Inpainting: guessing the 90% of the pixels

true image

Inpainting: guessing the 90% of the pixels

masked input image

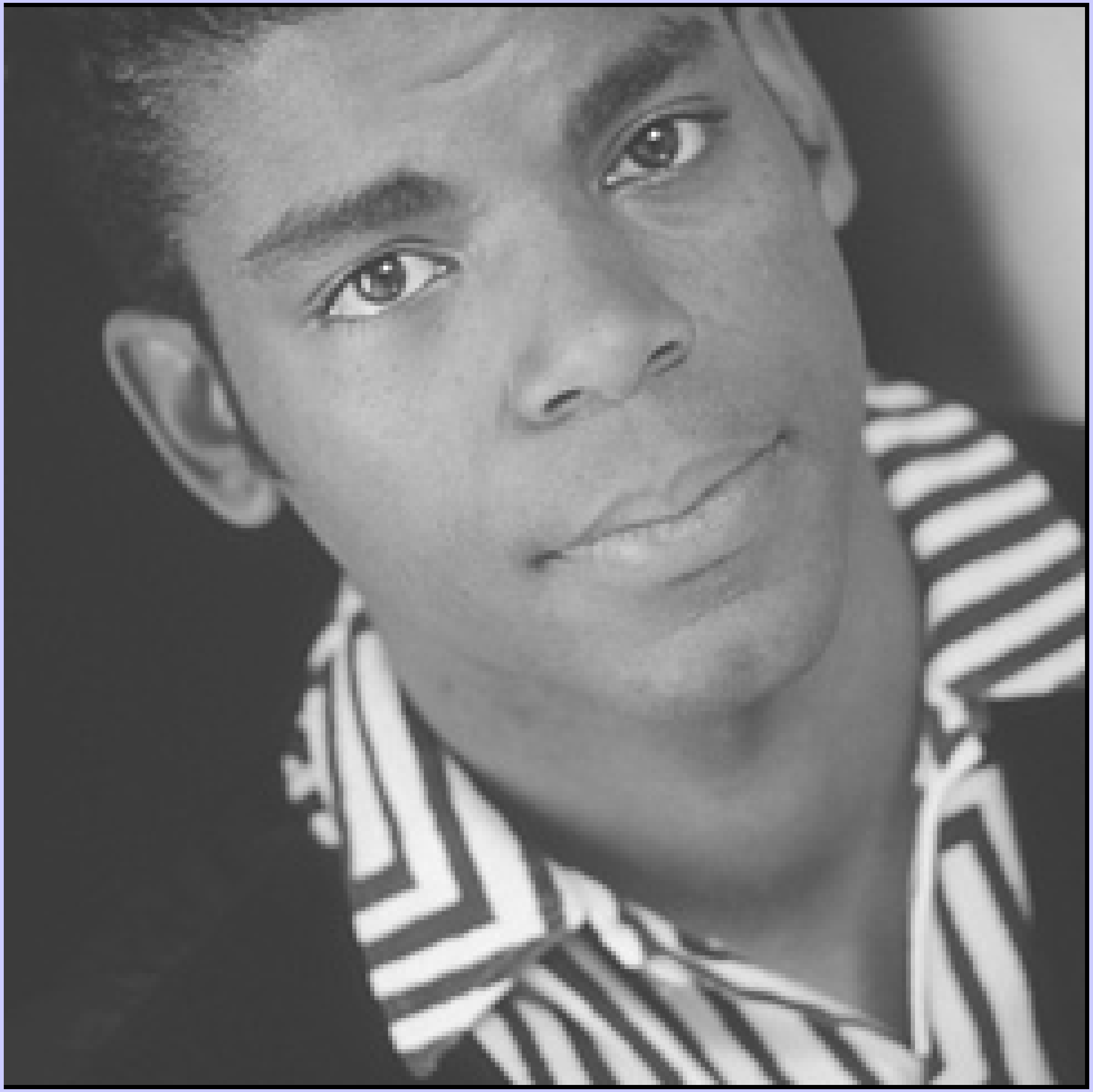


Inpainting: guessing the 90% of the pixels

MAP estimate missing pixels

Inpainting: guessing the 90% of the pixels

true image



Conclusion

- 3-way Boltzmann Machines
  - Joint model: fast inference, easy to interpret conditionals
  - It generates very realistic samples
  - Training is hard because of partition function
  - Future:
    - applications: segmentation, denoising
    - multi-scale
    - we'll make it DEEPER!!

THANK YOU

www.cs.toronto.edu/~ranzato