# Learning the Structure of Deep, Sparse Graphical Models

**Hanna M. Wallach**

University of Massachusetts Amherst

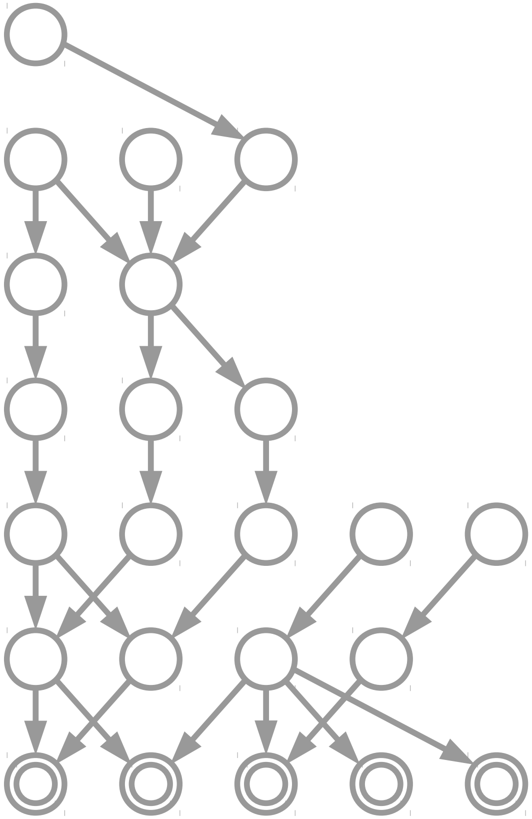
wallach@cs.umass.edu

Joint work with Ryan Prescott Adams & Zoubin Ghahramani

# Deep Belief Networks

> "Deep belief nets are probabilistic generative models that are composed of multiple layers of stochastic latent variables. The latent variables typically have binary values and are often called hidden units or feature detectors. [...] The lower layers receive top-down, directed connections from the layer above. The states of the units in the lowest layer represent a data vector."
>
> — Geoff Hinton ('09) Scholarpedia

# Network Structure

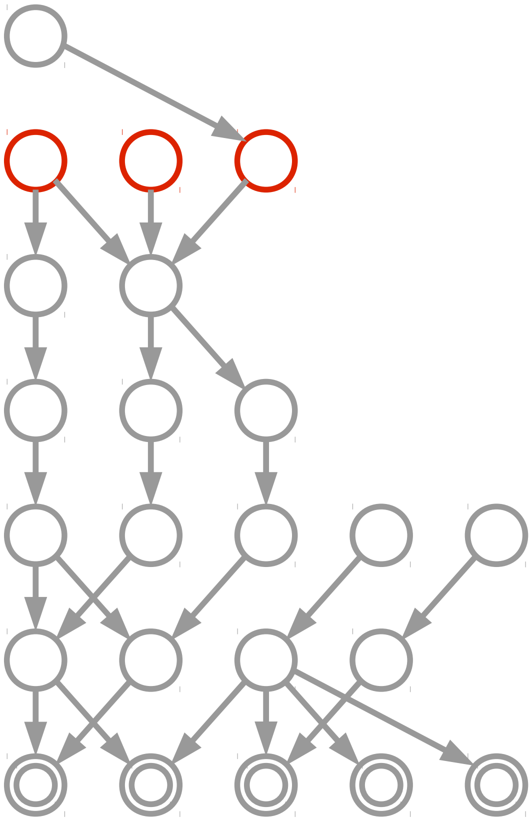


Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# Network Structure

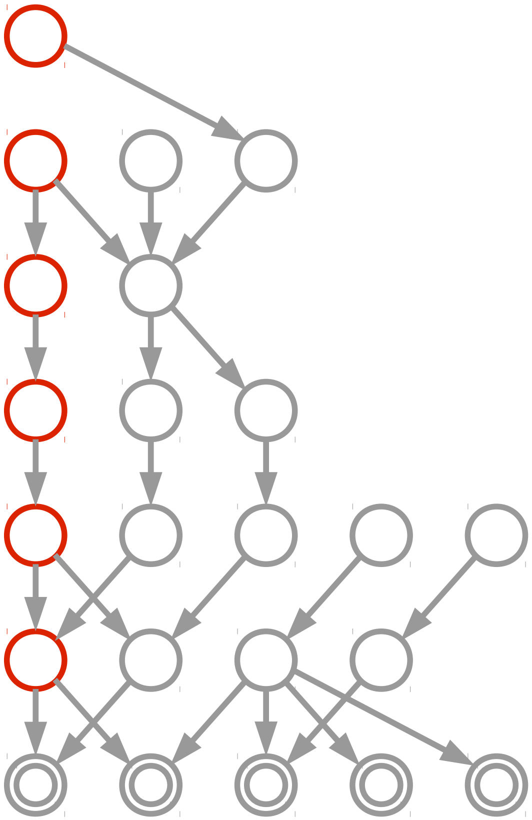


Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# Network Structure

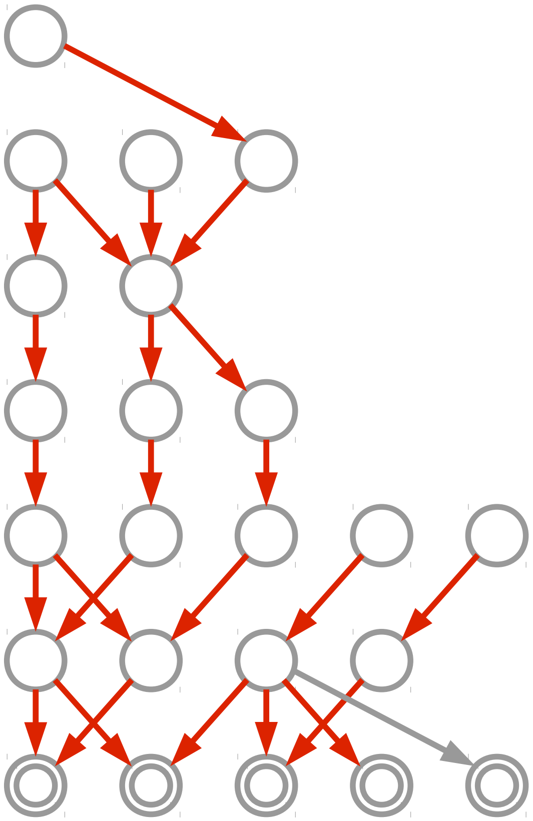


Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# Network Structure

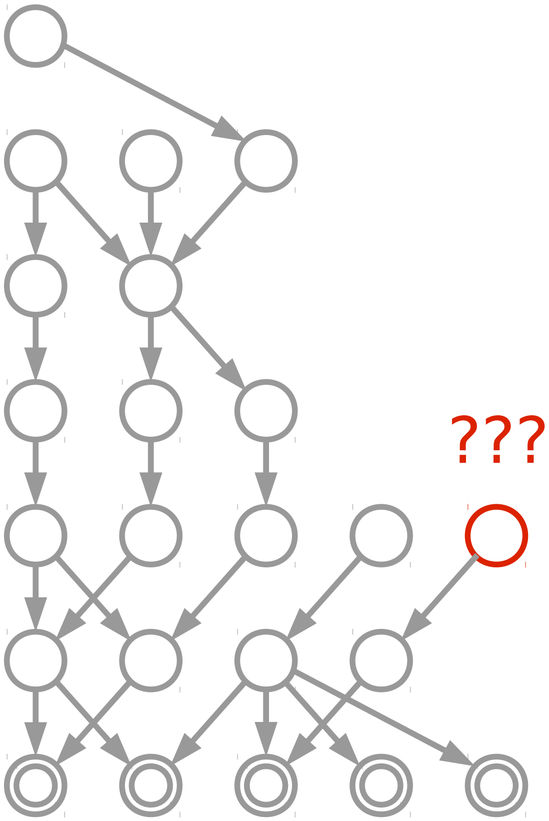


Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# Network Structure

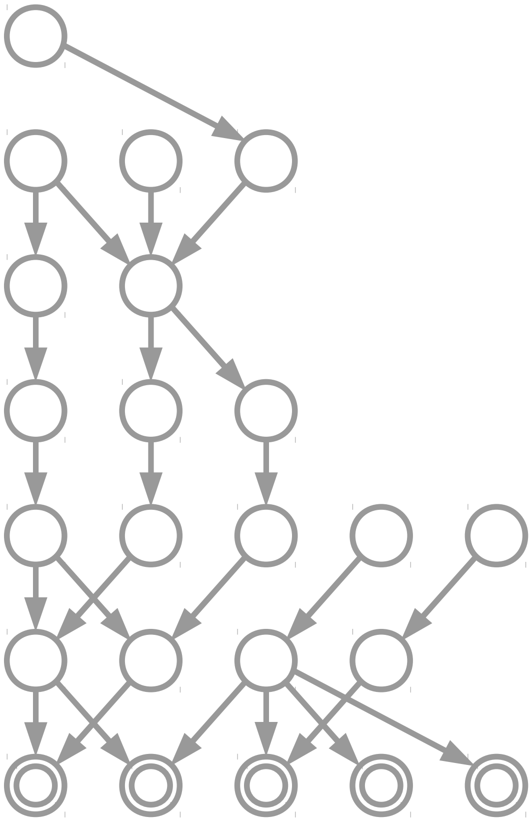


Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# Network Structure



Structural questions:

- # units in each hidden layer?
- # hidden layers?
- What network connectivity?
- What type(s) of unit behavior?

⇒ **Goal:** learn the structure

# This Talk

A nonparametric Bayesian approach for learning the structure of a layered, directed, deep belief network.

[Geoff Hinton] + [Mike Jordan] [Yee Whye Teh] [Zoubin Ghahramani]

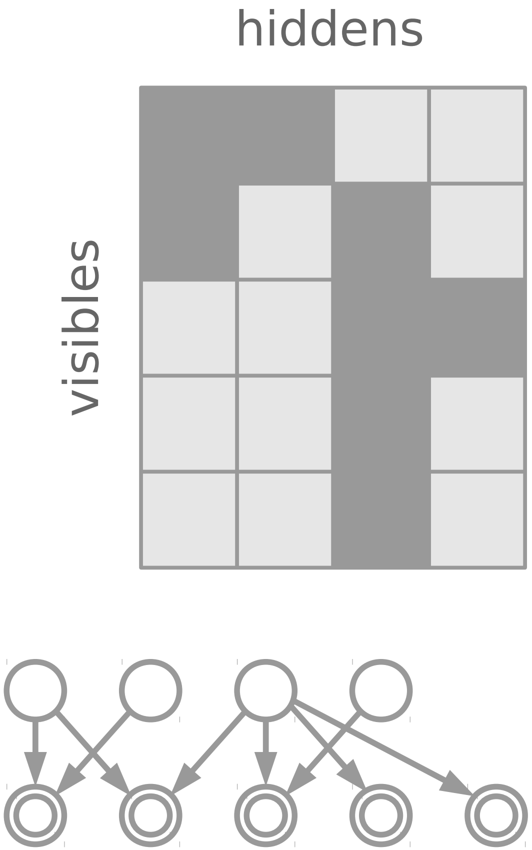
# Outline

- Background: finite single-layer networks
- Infinite belief networks:
  - Learning the number of hidden units in each layer
  - Learning the number of hidden layers
  - Learning the type(s) of unit behavior
- Experimental results

# Outline

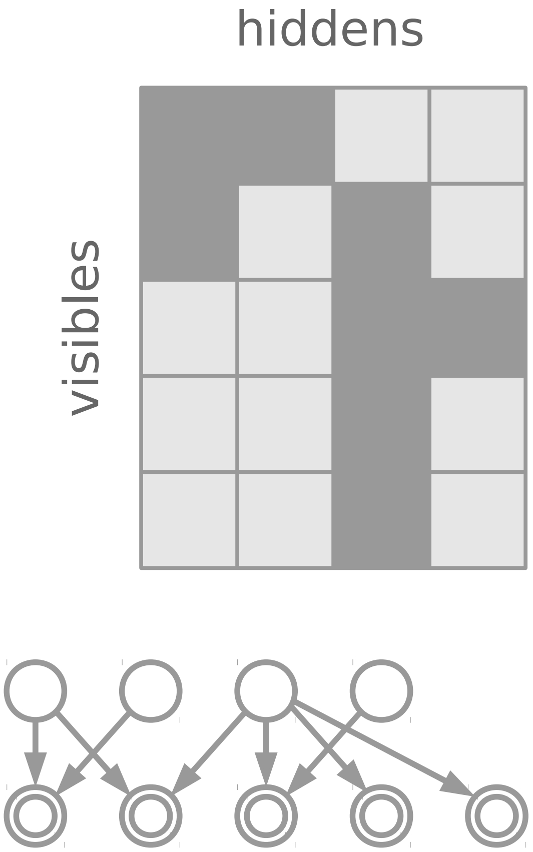
- Background: finite single-layer networks

# Finite Single-Layer Networks



- Use a binary matrix to represent the edge structure (connectivity) of a directed graph
- A prior distribution on binary matrices ⇒ a prior distribution on single-layered belief networks

# Finite Single-Layer Networks



- An infinite number of columns ⇒ an infinite number of hidden units
- Can we let these binary matrices to have an infinite number of columns?

⇒ **Yes:** Indian buffet process

# Outline

- Background: finite single-layer networks
- Infinite belief networks:
  - Learning the number of hidden units in each layer

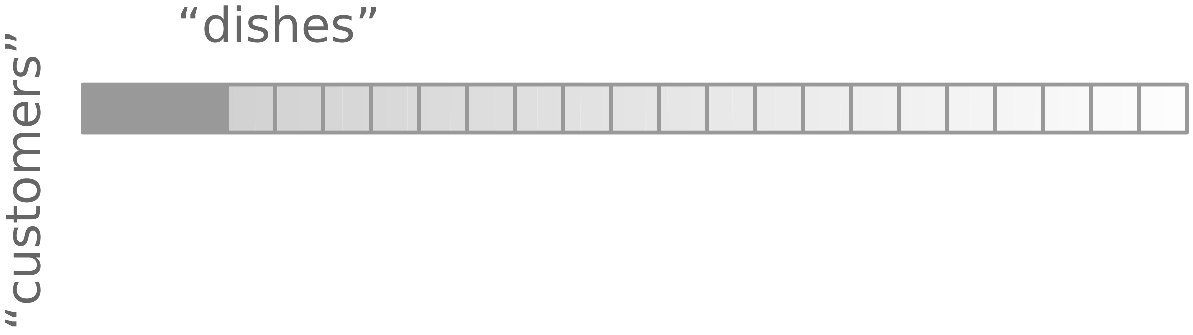
# Infinitely-Wide Layers

[Wood, Griffiths & Ghahramani, '06]

- Use an Indian buffet process (IBP) as a prior on binary matrices with **countably infinite columns**:
  - Unbounded number of hidden units
- Posterior inference determines the subset of hidden units responsible for the observations
- The IBP ensures that the matrices are extremely sparse: always a finite number of nonzero columns
  - Finite number of actively-used hidden units

# The Indian Buffet Process

[Ghahramani et al., '07]

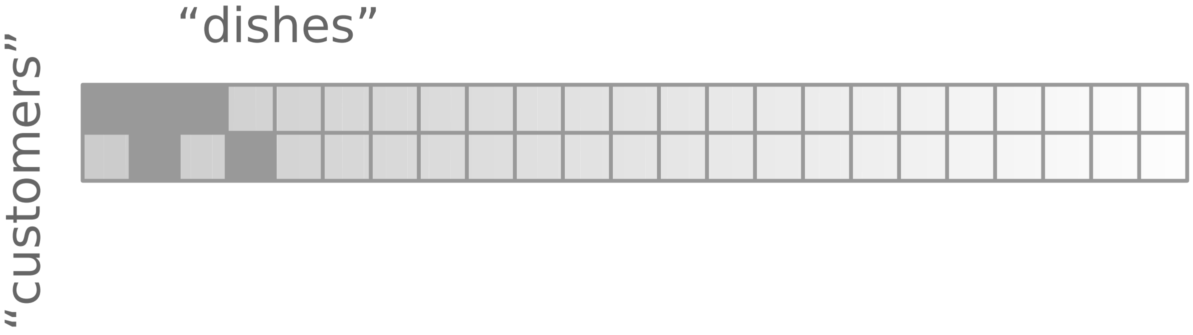


Parameters:
α and β

- First customer tries Poisson(α) dishes

# The Indian Buffet Process

[Ghahramani et al., '07]

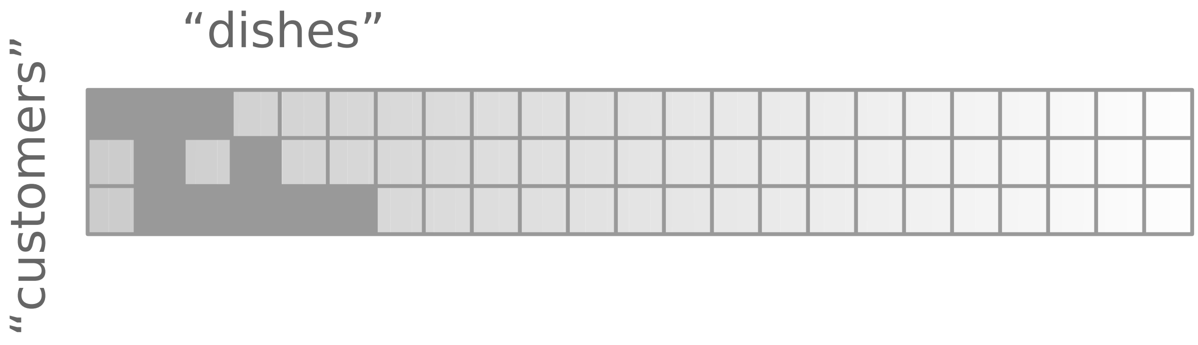


Parameters:
α and β

- First customer tries Poisson(α) dishes
- n^th^ customer tries:
  - Previously-tasted dish k with probability $n_k$ / (β + n - 1)
  - Poisson(αβ / (β + n – 1)) completely new dishes

# The Indian Buffet Process

[Ghahramani et al., '07]

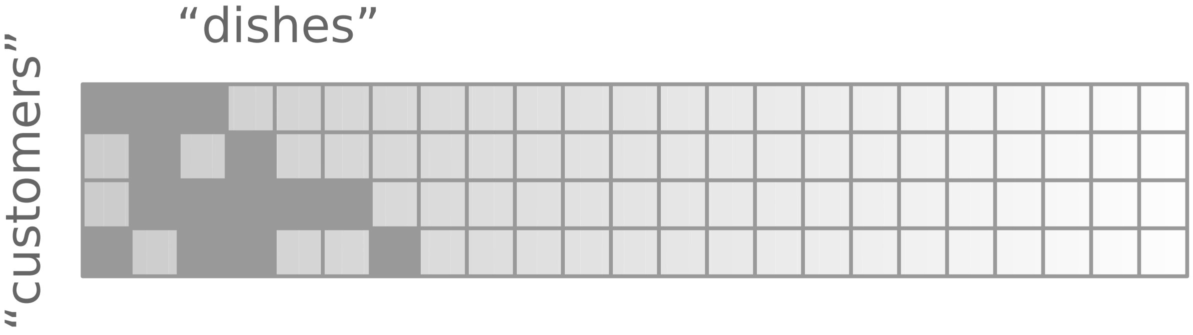


Parameters:
α and β

- First customer tries Poisson(α) dishes
- nth customer tries:
  - Previously-tasted dish k with probability $n_k$ / (β + n - 1)
  - Poisson(αβ / (β + n – 1)) completely new dishes

# The Indian Buffet Process

[Ghahramani et al., '07]



Parameters:
α and β

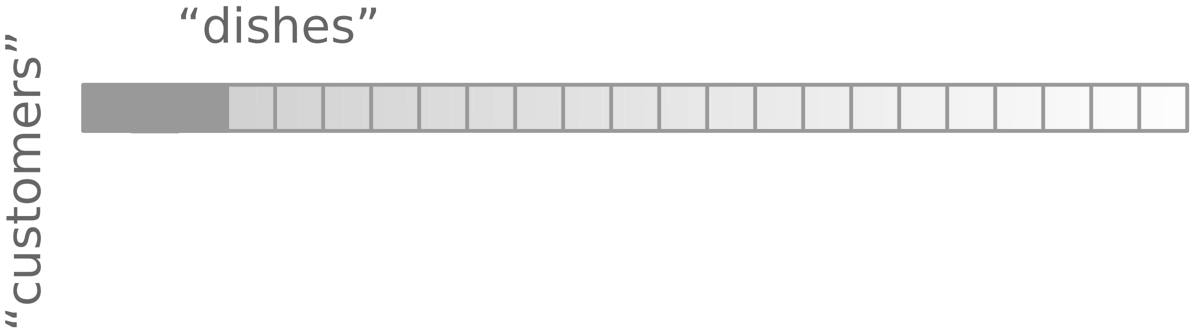
- First customer tries Poisson(α) dishes
- $n^{th}$ customer tries:
  - Previously-tasted dish k with probability $n_k$ / (β + n - 1)
  - Poisson(αβ / (β + n – 1)) completely new dishes

# Properties of the IBP

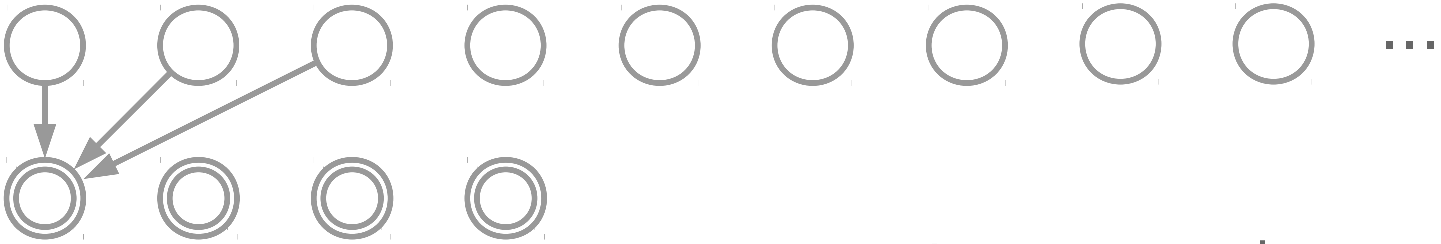
- For a finite number of customers, there will always be a finite number of dishes tasted
- Infinitely exchangeable rows and columns
- There is a related “stick-breaking” construction
- Popular for shared latent feature (hidden cause) models
- Latent features can be added/removed from a model without dimensionality-altering MCMC methods

# Single-Layer Belief Networks

[Wood, Griffiths & Ghahramani, '06]



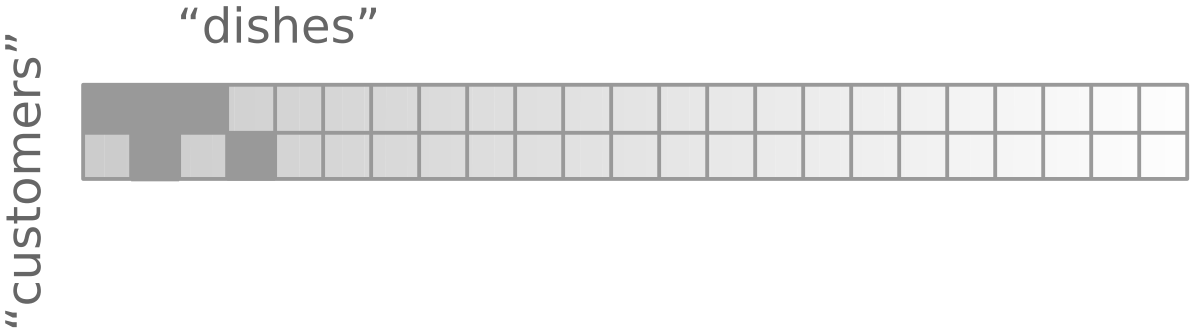
Parameters:
α and β



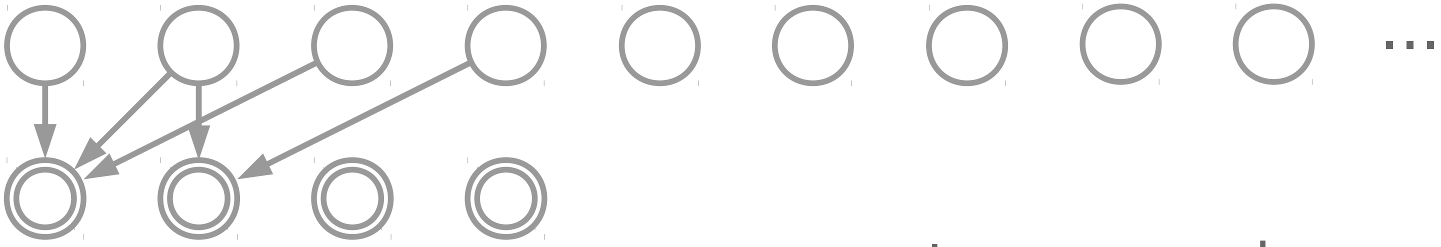
customers = observed units
dishes = hidden units

# Single-Layer Belief Networks

[Wood, Griffiths & Ghahramani, '06]



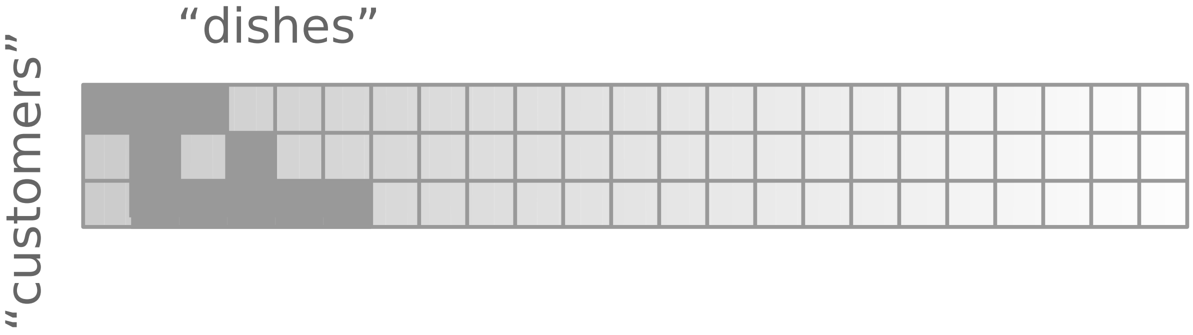
Parameters:
α and β



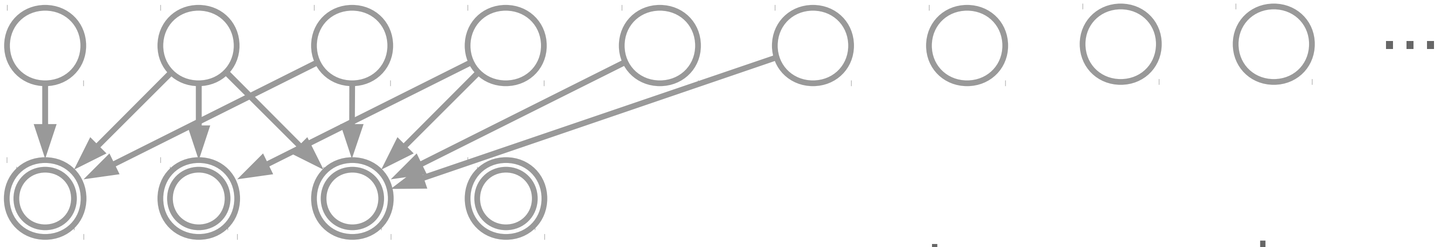
customers = observed units
dishes = hidden units

# Single-Layer Belief Networks

[Wood, Griffiths & Ghahramani, '06]



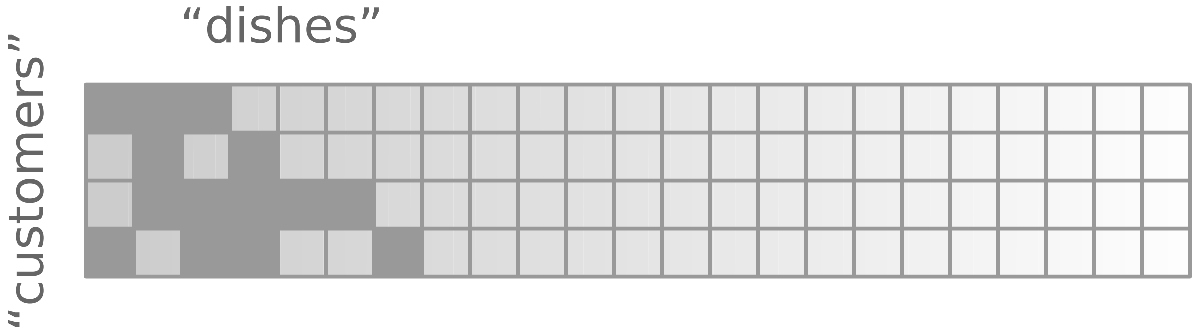
Parameters:
α and β



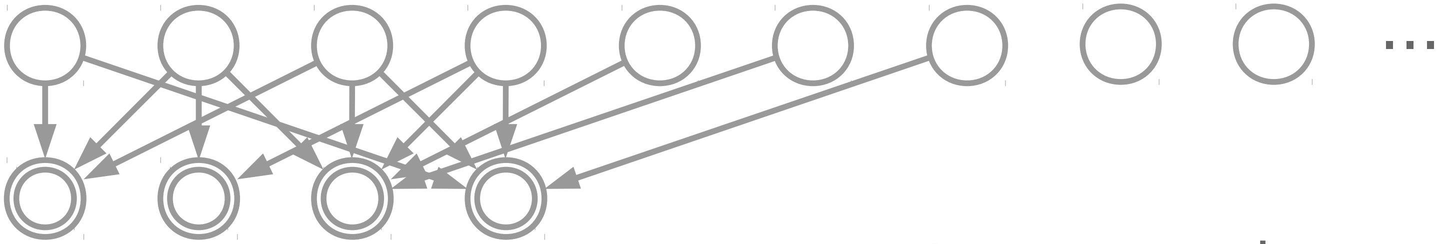
customers = observed units
dishes = hidden units

# Single-Layer Belief Networks

[Wood, Griffiths & Ghahramani, '06]



Parameters:
α and β



customers = observed units
dishes = hidden units

# Single-Layer ⇒ Multi-Layer?

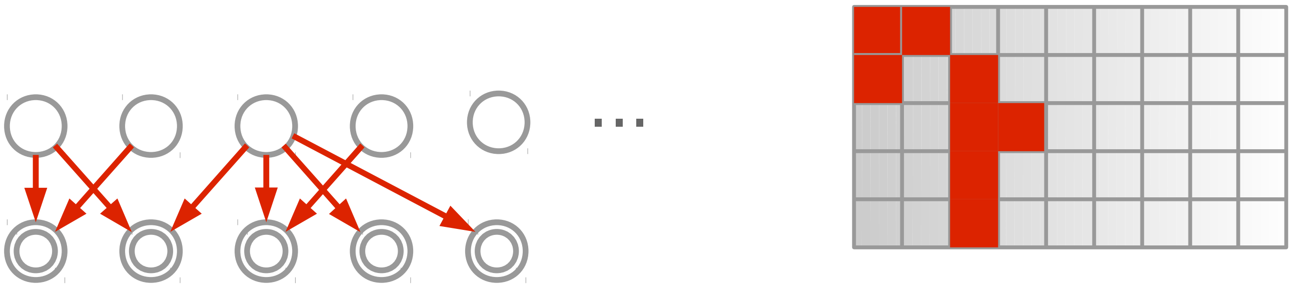
- Single-layer belief networks have limited utility:
  - Hidden units are independent a priori
- Deep networks = multiple hidden layers:
  - Hidden units are dependent a priori

⇒ **Goal:** extend the IBP in order to construct deep belief networks with unbounded width and depth

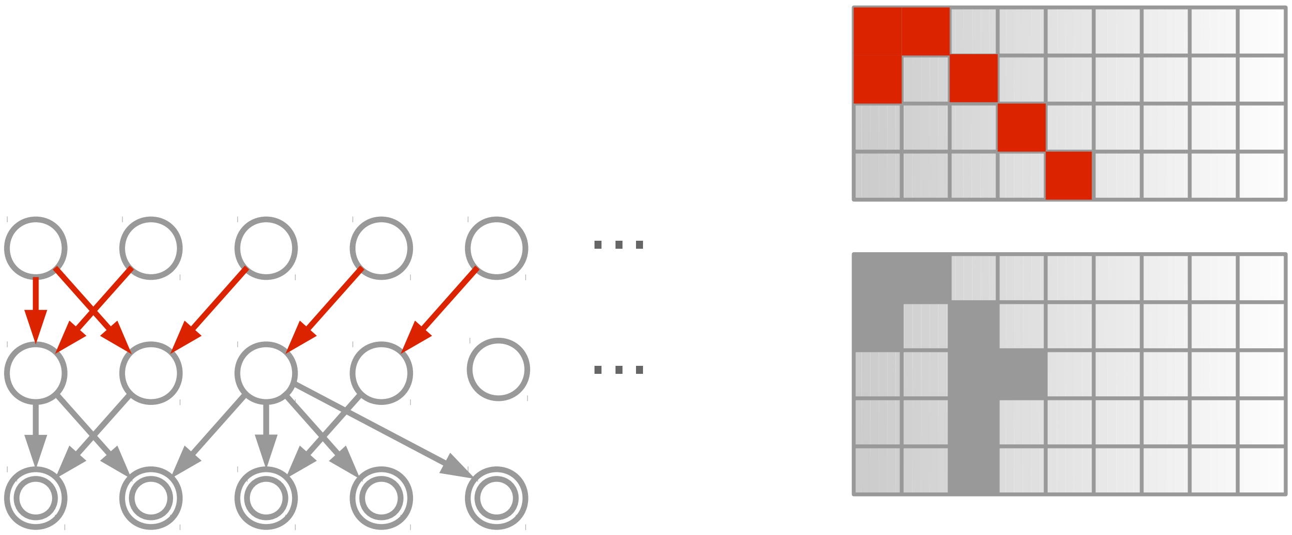
# Outline

- Background: finite single-layer networks
- Infinite belief networks:
    - Learning the number of hidden units in each layer
    - Learning the number of hidden layers

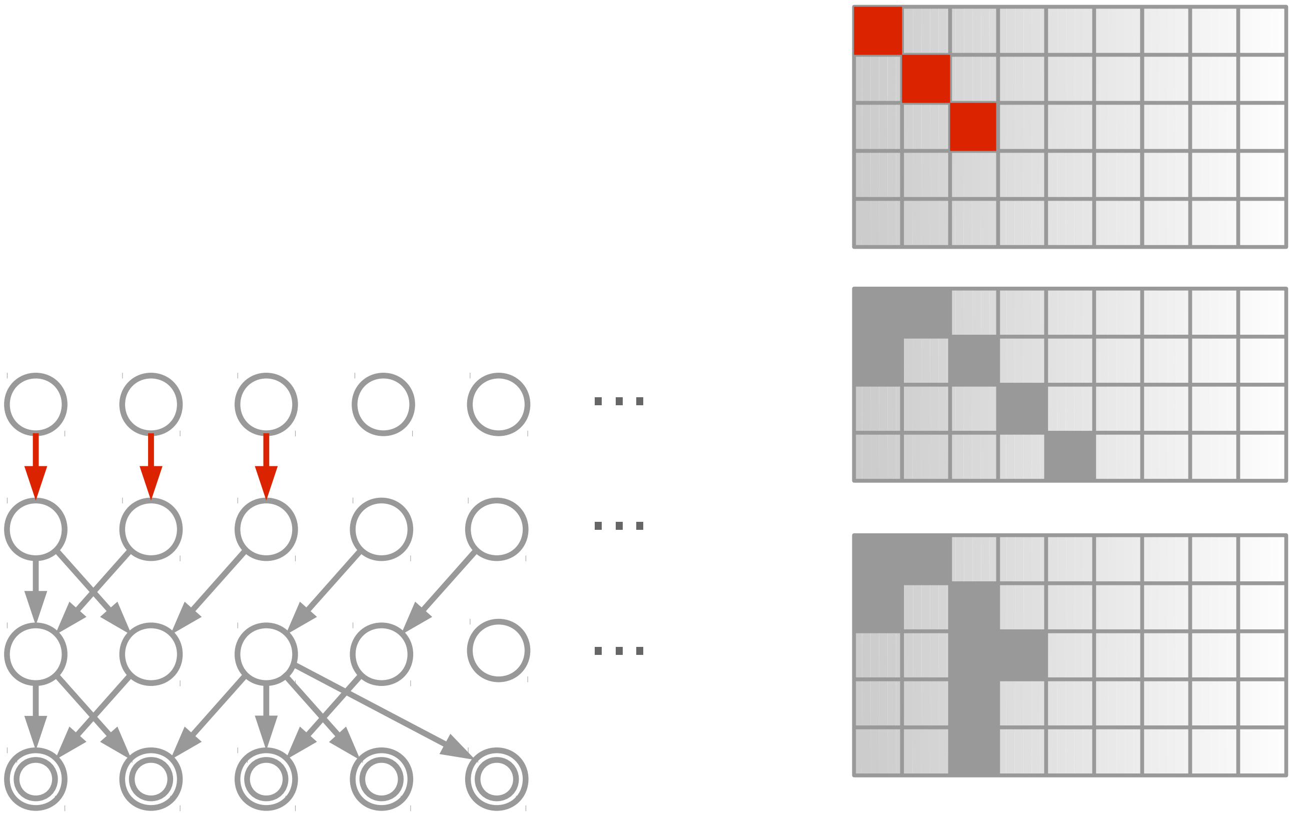
# Multi-Layer Belief Networks



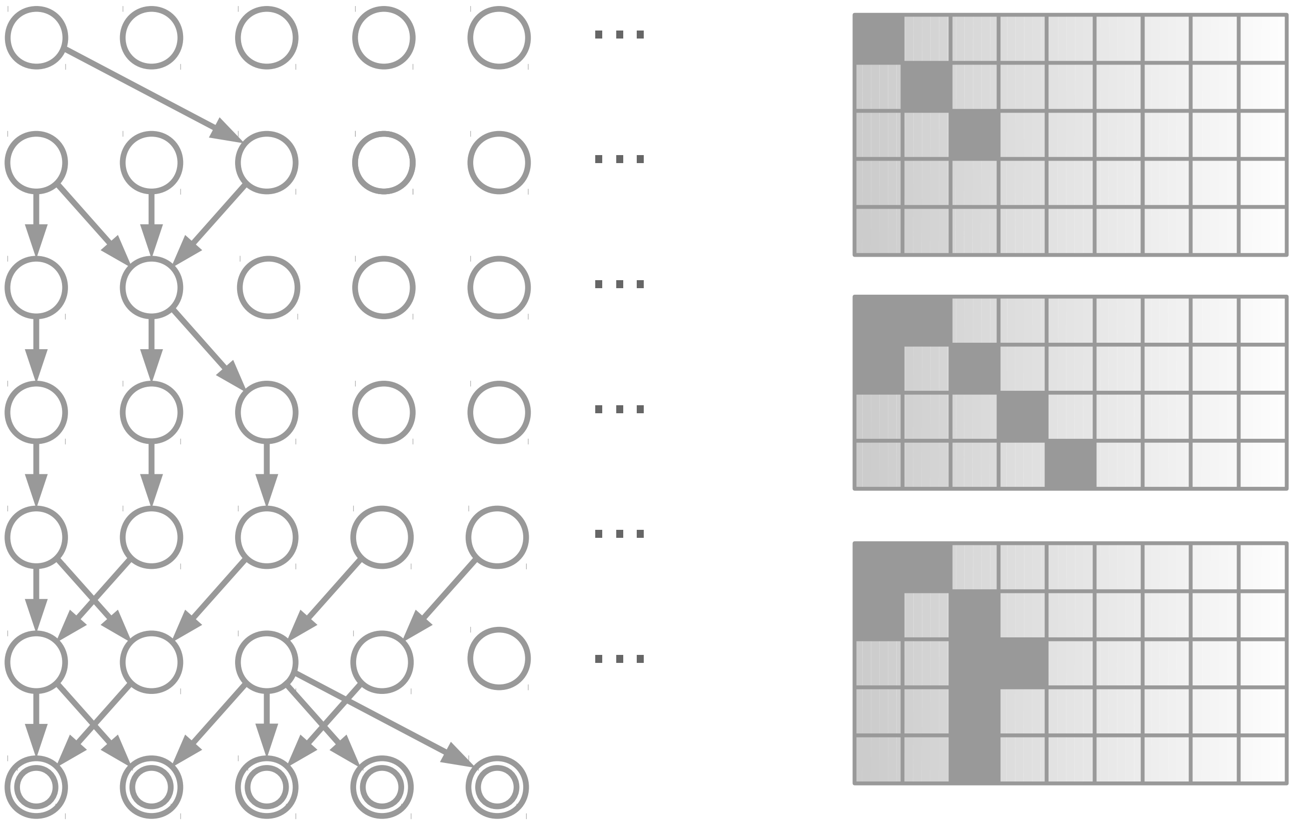
# Multi-Layer Belief Networks



# Multi-Layer Belief Networks



# Multi-Layer Belief Networks



# Layered Belief Nets

- We could use a finite number of IBPs
- But... what about using an infinite recursion:
  - Every “dish” is a “customer” in another restaurant
  - Unbounded number of layers, each of unbounded width
- Remarkably, we will always hit a layer with zero units!
  - Always stop at a finite but unbounded depth
- We don't have to make an ad hoc choice of depth

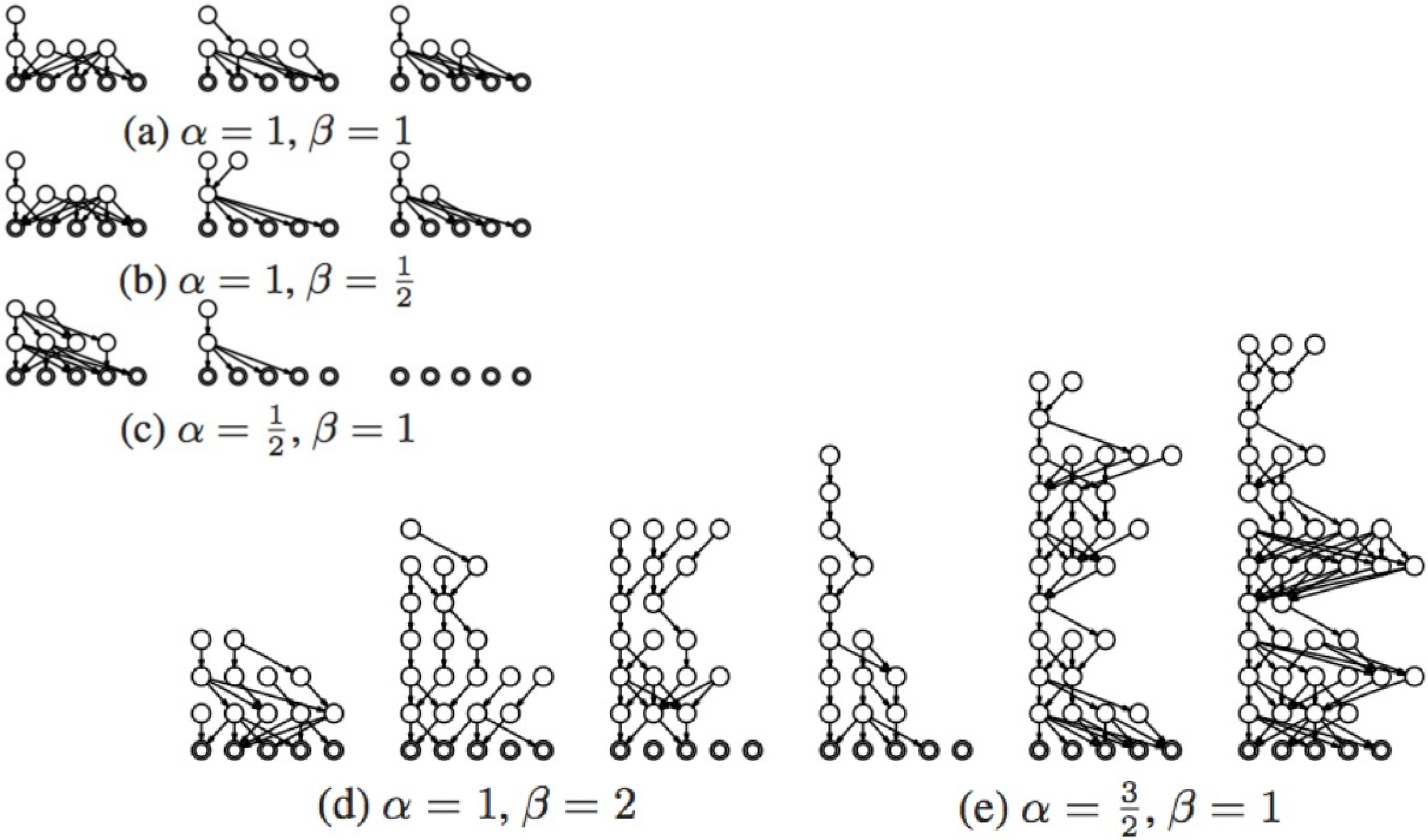
# The Cascading IBP (CIBP)

- A stochastic process which results in an infinite sequence of infinite binary matrices
  - Each matrix is exchangeable in both rows and columns
- How do we know the CIBP converges?
  - The number of dishes in one layer depends only on the number of customers in the previous layer
  - Can prove that this Markov chain reaches an absorbing state in finite time with probability one

# CIBP Properties

- For a unit in layer m+1:
  - Expected # of parents: $\alpha$
  - Expected # of children: $c(\beta, K_m) = 1 + (K_m - 1) / (1 + \beta)$
  - $\lim_{\beta \to 0} c(\beta, K_m) = K_m$ and $\lim_{\beta \to \infty} c(\beta, K_m) = 1$
- We do not want network properties to be constant at all depths, e.g., some levels should be sparser than others:
  - Each layer can have different IBP parameters $\alpha$ and $\beta$ so long as they are bounded from above

# Samples from the CIBP Prior



(a) $\alpha = 1, \beta = 1$

(b) $\alpha = 1, \beta = \frac{1}{2}$

(c) $\alpha = \frac{1}{2}, \beta = 1$

(d) $\alpha = 1, \beta = 2$

(e) $\alpha = \frac{3}{2}, \beta = 1$

# Outline

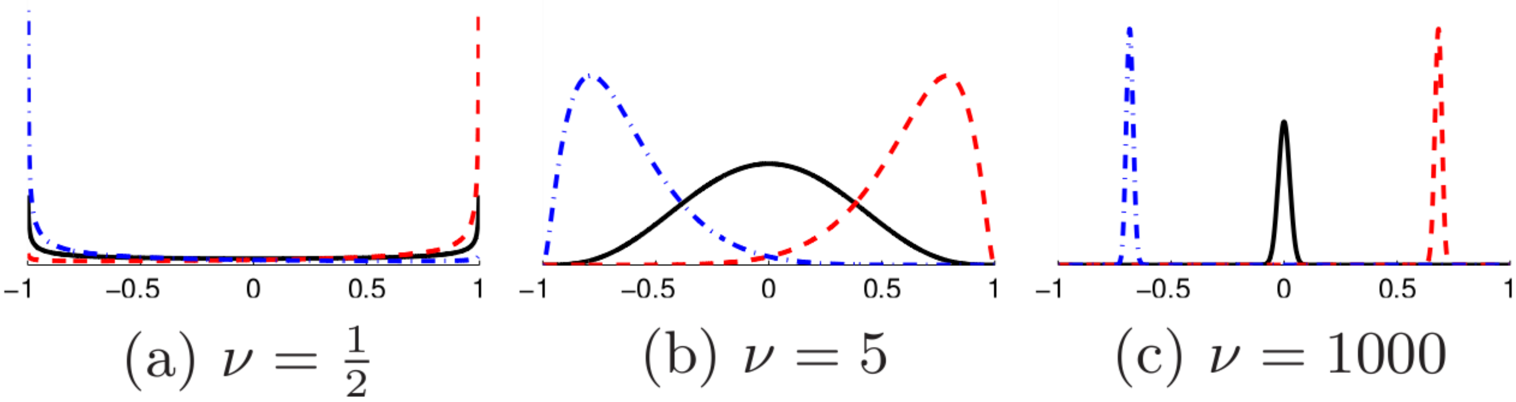
- Background: finite single-layer networks
- Infinite belief networks:
  - Learning the number of hidden units in each layer
  - Learning the number of hidden layers
  - Learning the type(s) of unit behavior

# Learning Unit Behavior

[Frey, '97; Frey & Hinton, '99]

- Unit activations are weighted linear sums with biases
- Nonlinear Gaussian belief network approach:

sigmoid(activation + Gaussian noise with precision ν)



(a) $\nu = \frac{1}{2}$ (b) $\nu = 5$ (c) $\nu = 1000$

# Priors and Inference

- Layer-wise Gaussian priors on weights and biases
- Layer-wise Gamma priors on noise precisions
- Layer-wise parameters tied via global hyperparameters
- Markov chain Monte Carlo inference:
  - Parameter inference is easy given the network structure
  - Edges added/removed using Metropolis-Hastings

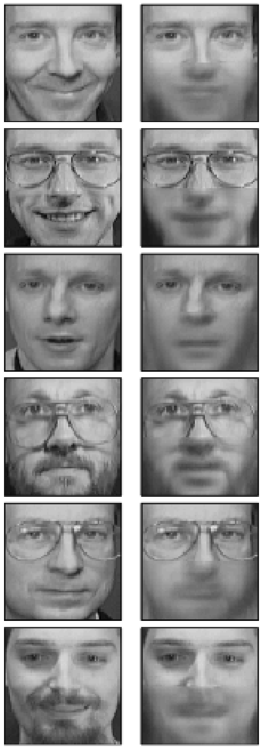
# Outline

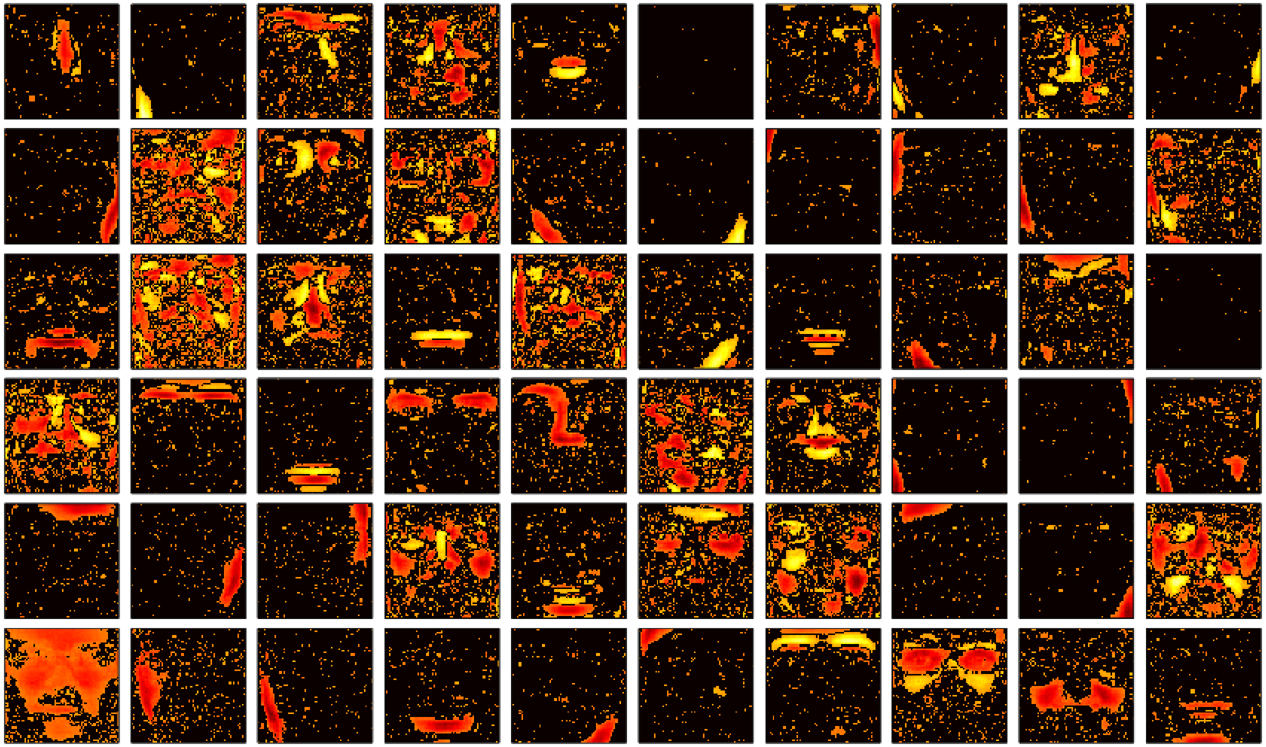
- Background: finite single-layer networks
- Infinite belief networks:
  - Learning the number of hidden units in each layer
  - Learning the number of hidden layers
  - Learning the type(s) of unit behavior
- Experimental results

# Experimental Results

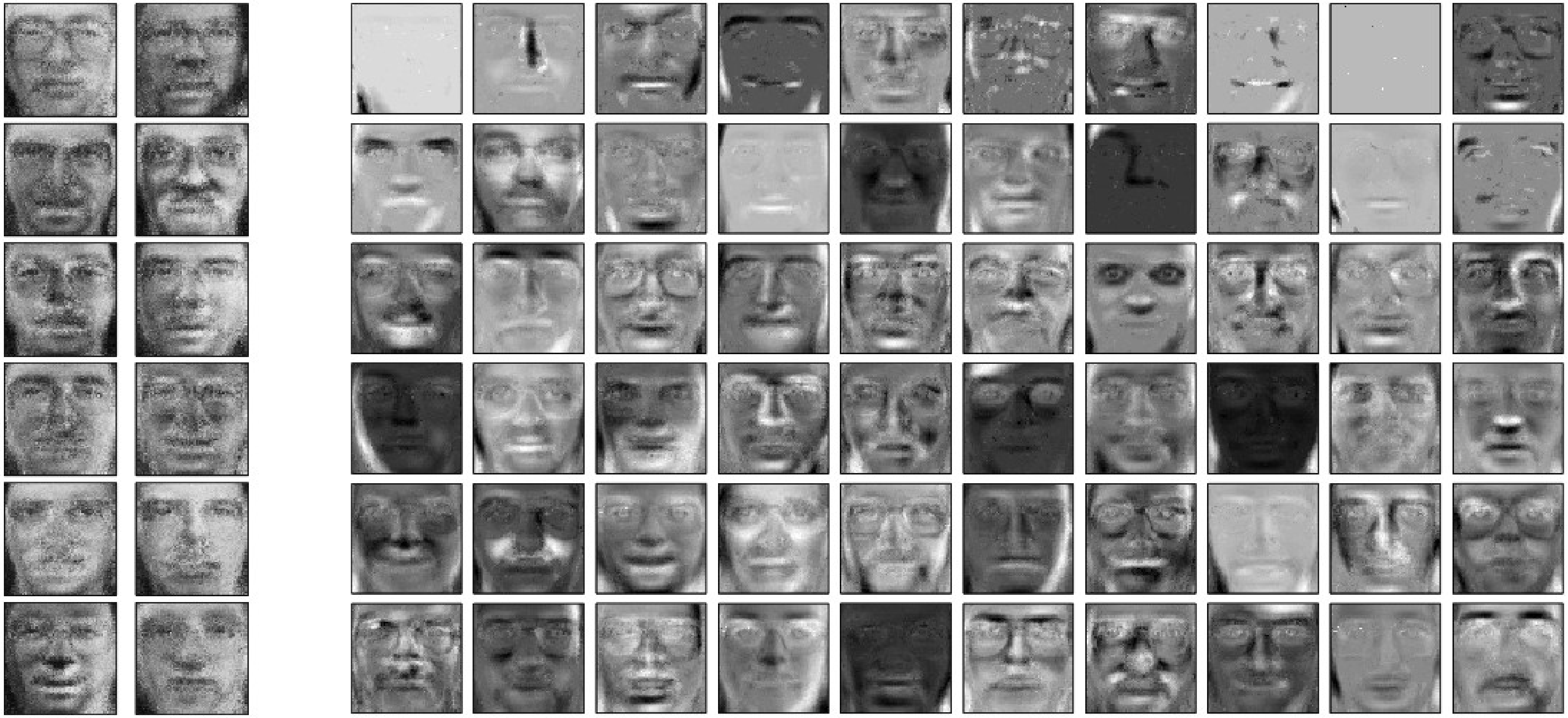
- **Olivetti Faces:** 350+50 images of 40 faces; 64x64
  - Inferred: ~3 hidden layers; 70 units per hidden layer
- **MNIST Digits:** 50+10 images of 10 digits; 28x28
  - Inferred: ~3 hidden layers; 120, 100, 70 units
- **Frey Faces:** 1865+100 images of Brendan Frey; 20x28
  - Inferred: ~3 hidden layers; 260, 120, 35 units

# Olivetti: Reconstructions & Features

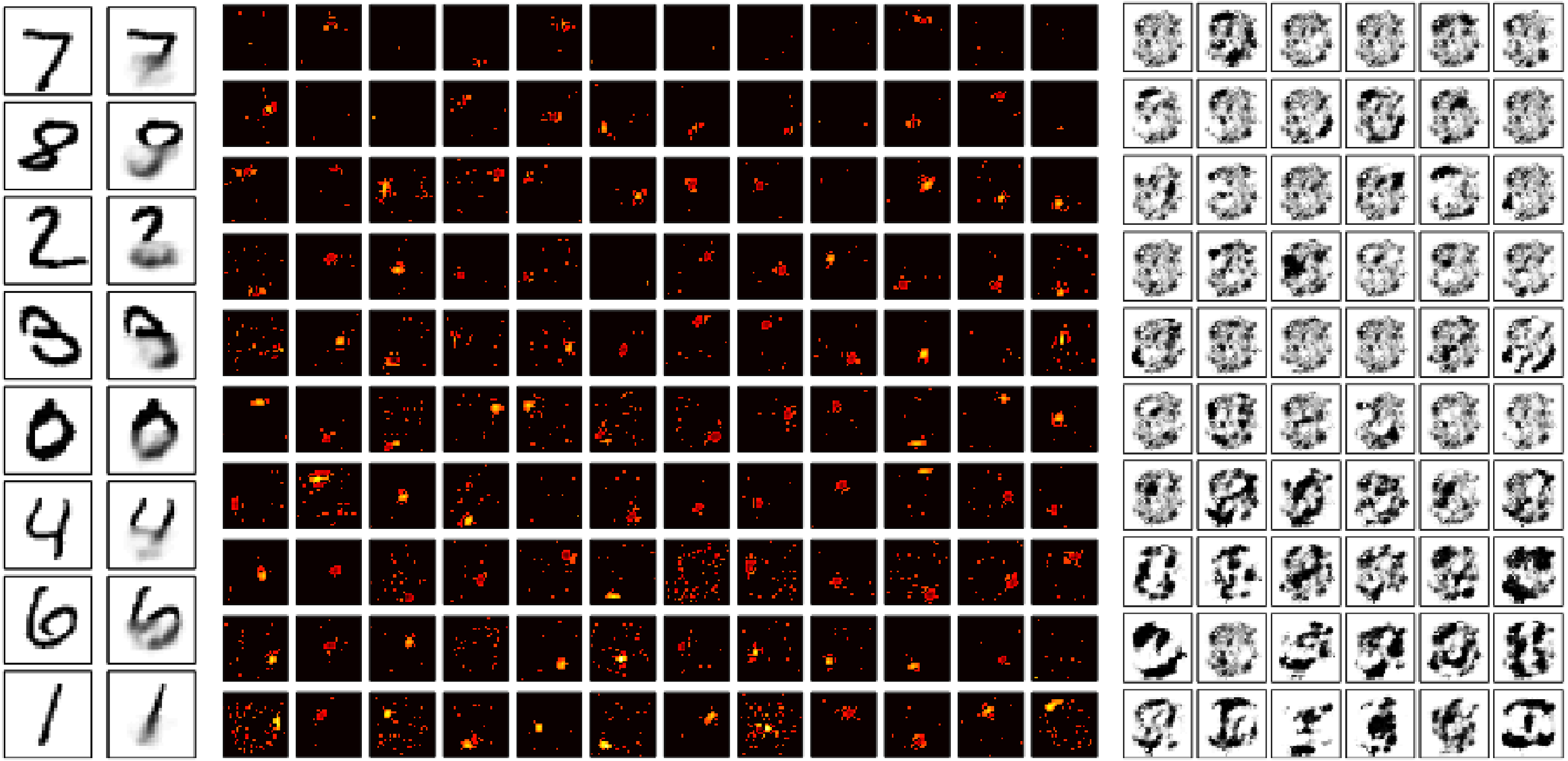




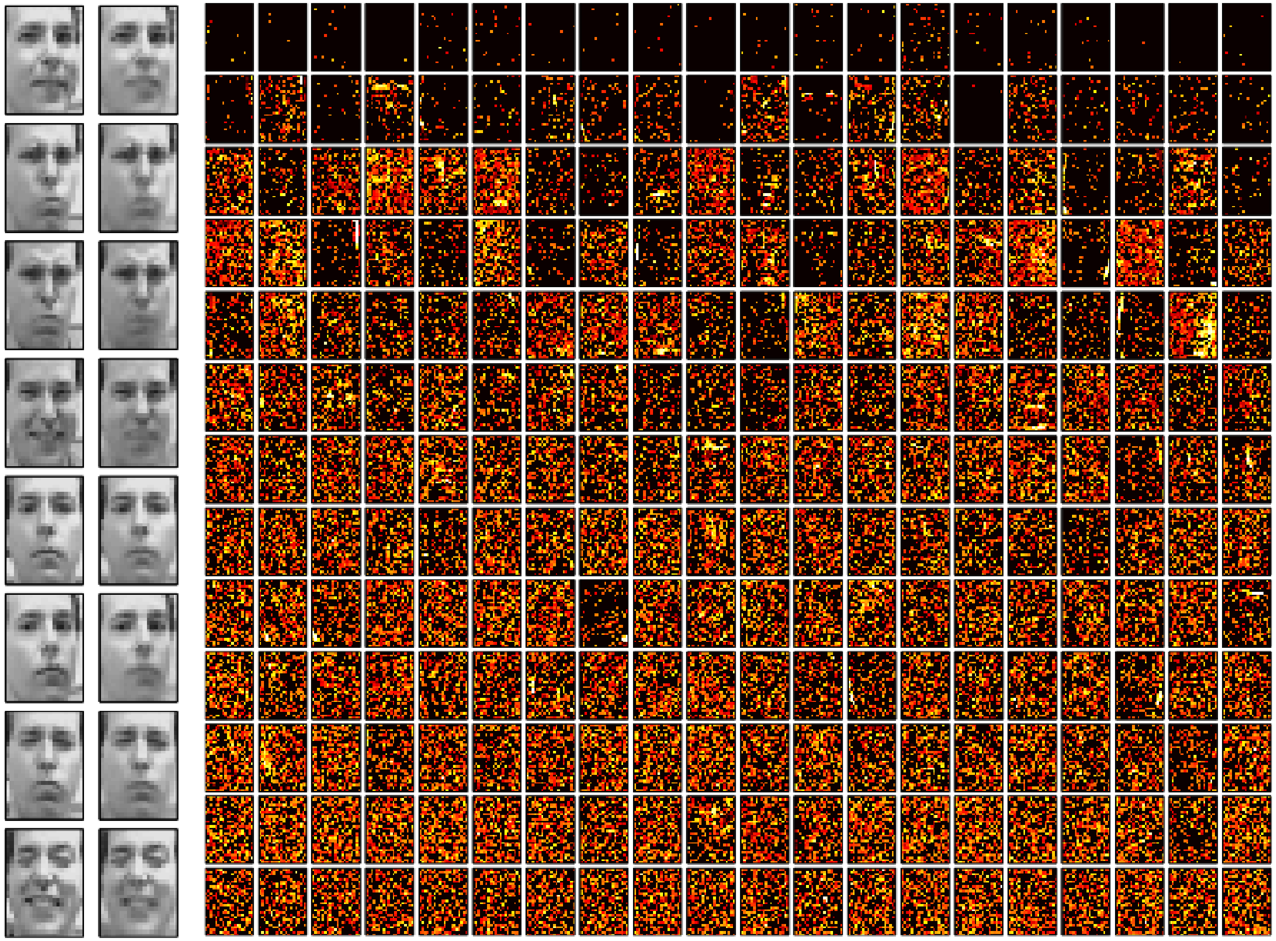
# Olivetti: Fantasies & Activations



# MNIST Digits



# Frey Faces



# Summary

- United deep belief networks & Bayesian nonparametrics
- Introduced the CIBP & proved convergence properties
- Addressed 3 issues with deep belief networks:
  - Number of units in each hidden layer
  - Number of hidden layers
  - Type(s) of hidden unit behavior

# Thanks!

Ryan Prescott Adams & Zoubin Ghahramani

Acknowledgements: Brendan Frey, David MacKay, Iain Murray, Radford Neal